Aplicação de ferramentas
de bioinformática para

ANÁLISE DE EXPRESSÃO GÊNICA POR RNA-SEQ

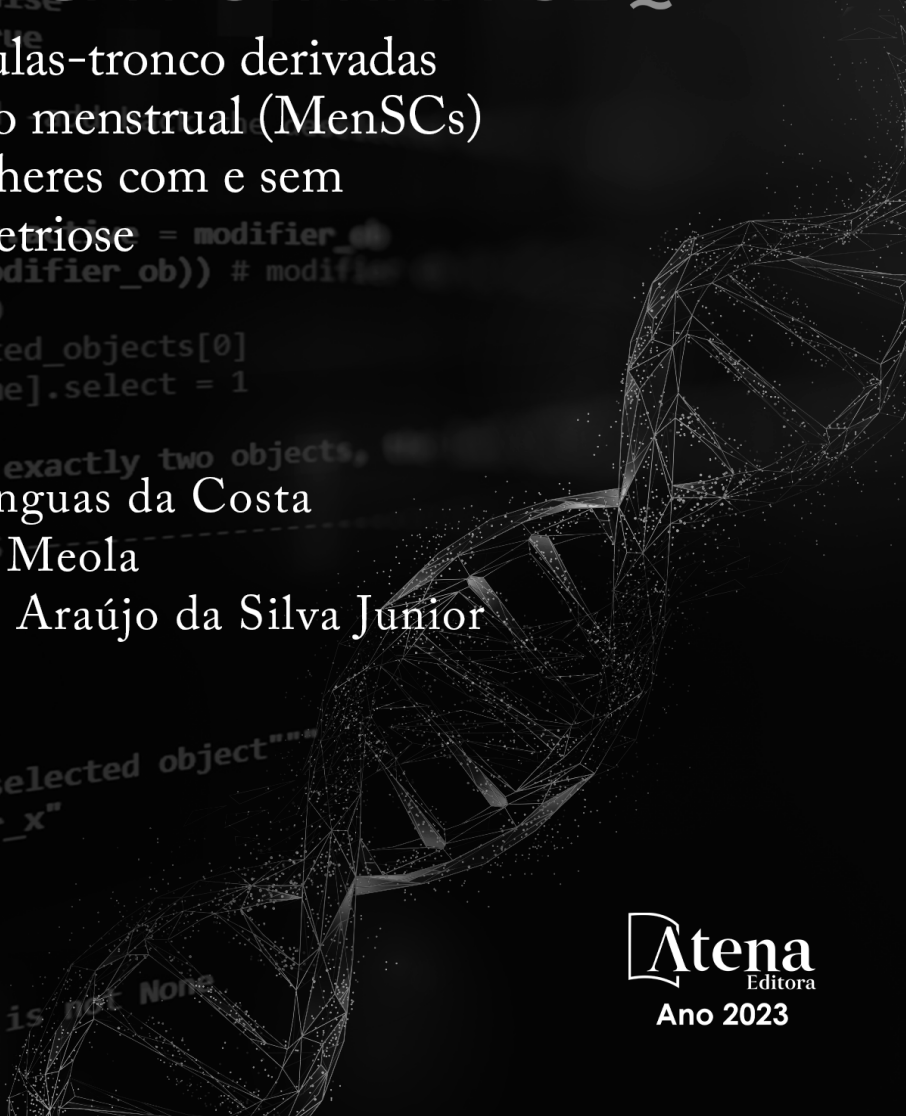
em células-tronco derivadas
de fluxo menstrual (MenSCs)
de mulheres com e sem
endometriose

Alef Janguas da Costa
Juliana Meola
Wilson Araújo da Silva Junior

Atena
Editora
Ano 2023

Aplicação de ferramentas de bioinformática para ANÁLISE DE EXPRESSÃO GÊNICA POR RNA-SEQ em células-tronco derivadas de fluxo menstrual (MenSCs) de mulheres com e sem endometriose

Alef Janguas da Costa
Juliana Meola
Wilson Araújo da Silva Junior



Atena Editora
Ano 2023

CIÊNCIAS BIOLÓGICAS E DA SAÚDE

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Camila Alves de Cremo
Ellen Andressa Kubisty
Luiza Alves Batista
Nataly Evilin Gayde
Thamires Camili Gayde
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

CIÊNCIAS BIOLÓGICAS E DA SAÚDE

Aplicação de ferramentas de Bioinformática para análise de expressão Gênica por RNA-seq de Células Tronco Mesenquimais Endometriais no fluxo menstrual (MenSCs) de mulheres com e sem endometriose

Diagramação: Letícia Alves Vitral
Correção: Flávia Roberta Barão
Indexação: Amanda Kelly da Costa Veiga
Revisão: Os autores
Autores: Alef Janguas da Costa
Juliana Meola
Wilson Araújo da Silva Junior

Dados Internacionais de Catalogação na Publicação (CIP)

C837 Costa, Alef Janguas da
Aplicação de ferramentas de Bioinformática para análise de expressão Gênica por RNA-seq de Células Tronco Mesenquimais Endometriais no fluxo menstrual (MenSCs) de mulheres com e sem endometriose / Alef Janguas da Costa, Juliana Meola, Wilson Araújo da Silva Junior. – Ponta Grossa - PR: Atena, 2023.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-258-1709-5
DOI: https://doi.org/10.22533/at.ed.095231809

1. Bioinformática. I. Costa, Alef Janguas da. II. Meola, Juliana. III. Silva Junior, Wilson Araújo da. IV. Título.
CDD 570.285

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao conteúdo publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que o texto publicado está completamente isento de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

“O que sabemos é uma gota; o que ignoramos é um oceano.

(Isaac Newton)

À minha família, amigos e todos que também me apoiaram no caminho.

AGRADECIMENTOS

Agradecimentos à minha família por estar sempre me apoiando, me bancando e aguentando todas as dificuldades pela qual passei durante esta graduação.

Aos meus amigos, que me ajudaram a estudar para disciplinas, a relaxar em momentos difíceis e a discutir assuntos importantes. A vivência na salinha 500B, idas ao cinema, almoços, bandejões, festas.

A minha orientadora que me deu essa oportunidade de aprender mais sobre bioinformática, aprender mais sobre essa condição e suas complicações. Foi uma ótima experiência que dificilmente obteria no curso ou em qualquer outra empresa fora da faculdade.

Aos professores que ensinaram bastante sobre persistência, superações, sonhos e realizações.

E agradecimentos ao centro estudantil da informática biomédica, à empresa júnior que estão sempre dispostos a fazer a diferença para quem está no curso.

RESUMO

Este estudo abrange a análise por bioinformática dos dados de sequenciamento de nova geração de transcritos (RNA-seq) em larga escala das células tronco mesenquimais obtidas do fluxo menstrual (MenSCs) de mulheres com e sem endometriose. Nesta análise avaliamos 2 diferentes métodos estatísticos do pacote EdgeR: Exato e *General Linear Model (GLM)* para encontrarmos genes diferencialmente expressos e definirmos a expressão gênica diferencial da endometriose. Os métodos estatísticos avaliados obtiveram resultados semelhantes. Neste estudo obtivemos um conjunto de genes e com o que eles estão associados segundo o *Database for Annotation, Visualization and Integrated Discovery* (DAVID). Entretanto, com esses dados não observamos um perfil de expressão genica diferencial entre os grupos estudados (Controle e Endometriose). Indicando que as células mesenquimais do fluxo menstrual de mulheres com e sem endometriose possuam diferença de expressão discreta. Sendo assim, este estudo caracteriza-se como um estudo piloto.

PALAVRAS-CHAVE: Protocolos clínicos; prática clínica baseada em evidências; Avaliação em saúde; Sistema de apoio à decisão; avaliação de evidências; endometriose; RNAseq; MenSC.

This study covers the bioinformatics analysis of large-scale next-generation transcript sequencing (RNA-seq) data from mesenchymal stem cells obtained from the menstrual flow (MenSCs) of women with and without endometriosis. In this analysis we evaluated two different statistical methods of the EdgeR package: Exact and General Linear Model (GLM) to find differentially expressed genes and define the differential gene expression of endometriosis. The statistical methods evaluated obtained similar results. In this study we obtained a set of genes and what they are involved according to the Database for Annotation, Visualization and Integrated Discovery (DAVID). However, with these data, we did not observe a differential gene expression profile between the studied groups (Control and Endometriosis). Indicating that the mesenchymal cells of the menstrual flow of women with and without endometriosis have a slight difference in expression. Therefore, this study is characterized as a pilot study.

KEYWORDS: Clinical protocols; clinical practice based on evidence; Health evaluation; System decision support; evidence assessment; endometriosis; RNAseq; MenSC.

ABSTRACT

SUMÁRIO

SUMÁRIO

SUMÁRIO

INTRODUÇÃO

ETIOPATOGENIA DA ENDOMETRIOSE

A endometriose é uma doença ginecológica estrogênio-dependente que afeta 6 a 10% das mulheres em idade reprodutiva (Ozkan *et al.*). Caracteriza-se por implantes de tecido endometrial (glândulas e/ou estroma) fora da cavidade uterina (tecido ectópico), onde se desenvolvem e formam lesões (Eskenazi e Warner). Esse tecido é encontrado mais frequentemente no peritônio pélvico e nos ovários, mas também pode situar-se em outros órgãos pélvicos, além de septo reto-vaginal, pleura, parede abdominal e, raramente, no cérebro. O quadro clínico é bastante diversificado, variando desde assintomático até dor pélvica crônica, dismenorréia, dispareunia, sangramento uterino e infertilidade (Bulun). Tanto por seu impacto na saúde física e psicológica, como pelo impacto sócio-econômico diante dos custos para o seu diagnóstico, tratamento e monitoramento, a endometriose tem sido considerada atualmente um problema de saúde pública (Signorile e Baldi).

As manifestações clínicas da endometriose e a presença do tecido ectópico são provavelmente o resultado da combinação de vários processos biológicos aberrantes, que incluem a menstruação retrógrada em mulheres com resposta imune imprópria e com predisposição para desenvolver os implantes ectópicos, que possivelmente estão expostos a um microambiente alterado (Bischoff e Simpson; Halme J Fau - Hammond *et al.*). A origem do endométrio ectópico tem sido objeto de muita investigação. Até o momento, a teoria mais aceita para a etiologia da endometriose é a de que haveria aderência de tecido endometrial decorrente de fluxo menstrual retrógrado, que carrega células com alterações funcionais capazes de permitir sua implantação e desenvolvimento ao atingir a cavidade peritoneal e órgãos adjacentes (Sampson).

As demais teorias sobre a etiologia da endometriose são: dos resquícios embrionários (células residuais de origem mülleriana seriam capazes de desenvolver lesões endometrióticas sob a influência de estrógeno) (Wood Russell); da disseminação linfovascular (células endometriais se disseminariam por meio de vasos linfáticos ou sanguíneos), o que explicaria o aparecimento de focos endometriais em sítios distantes da pelve, como cérebro, pulmões e linfonodos (Sasson e Taylor); da metaplasia celômica (sugere que o epitélio celômico poderia transformar-se em tecido semelhante ao endométrio) (Meyer, 1919). Apesar de muitas, nenhuma dessas prerrogativas sozinha é capaz de explicar por qual razão a doença se origina.

Mais recentemente, uma hipótese vem sendo sugerida para complementar as teorias sobre a etiopatogenia da endometriose: sugere-se a participação de células

tronco (CT) endometriais (denominadas *endometrial Mesenchymal Stem Cells*, ou *eMSC*) na origem das lesões endometrióticas. Algumas suposições são feitas como: 1) estas *eMSCs* molecularmente alteradas atingem a cavidade peritoneal através da menstruação retrógrada e se implantam no peritônio originando as lesões endometrióticas; 2) ou que estas *eMSCs* seriam molecularmente normais e poderiam implantar-se em um peritônio com receptividade aumentada, sugerindo que o microambiente peritoneal seja alterado nas mulheres com endometriose; 3) ou ainda, uma combinação das duas coisas, alteração molecular das células *eMSC* combinadas a um ambiente peritoneal também alterado e receptivo (Gargett e Masuda; Paula Gabriela Marin e Figueira, 2011).

EVIDÊNCIAS DE CÉLULAS TRONCO NO ENDOMÉTRIO E A ENDOMETRIOSE

O endométrio humano é altamente regenerativo e está sujeito a mais de 400 ciclos de crescimento, diferenciação e descamação durante a vida reprodutiva da mulher (Jabbour *et al.*). Estrutural e funcionalmente, é dividido em: 1) camada funcional (superior), formada por epitélio glandular e estroma rico em células, sendo capaz de sofrer alterações morfológicas e bioquímicas cíclicas em resposta aos hormônios ovarianos; 2) camada basal (inferior), composta por glândulas e estroma denso e serve como compartimento germinal para regeneração da camada funcional em todo ciclo menstrual. Durante ciclo menstrual, a camada funcional e uma pequena porção da camada basal descamam (Okulicz *et al.*). Acredita-se que a camada basal abriga maior número de células tronco que a camada funcional (Gargett e Masuda; Spencer *et al.*) e, baseado na dinâmica do remodelamento endometrial durante o ciclo menstrual e na gravidez, tem-se sugerido que células tronco adultas realizam um papel proeminente na manutenção e funcionamento do endométrio (J, 2008). Em 2004, a primeira evidência da capacidade clonogênica das células endometriais humanas foi demonstrada pela existência de populações de células progenitoras epiteliais (0,22±0,07%) e estromais (1,25±0,18%) (Chan, Schwab Ke Fau - Gargett*, et al.*)

As observações combinadas de que as camadas basal e funcional do endométrio contêm células tronco (Masuda *et al.*, 2010); que as lesões endometrióticas têm origem clonal (Wu *et al.*); e que mulheres com endometriose têm maior volume de fluxo menstrual (Halme J Fau - Hammond *et al.*) e maior prevalência de fragmentos descamados da camada basal no fluxo menstrual em relação a mulheres saudáveis (Leyendecker *et al.*) permitem inferir que os implantes ectópicos são iniciados por *eMSCs* presentes no fluxo menstrual retrógrado (Hwang *et al.*; Macer e Taylor). Em 2011, estudos revelaram achados importantes: células com propriedades iguais às de *eMSCs* foram identificadas em lesões ectópicas, tanto em tecido fresco, como em cultura de células (Chan, Ng Eh Fau - Yeung*, et al.*; Kao *et al.*). Observou-se que as *eMSC* derivadas de cultura de células estromais

ectópicas têm maior potencial proliferativo, migratório e invasivo que as células isoladas do endométrio eutópico dessas mesmas pacientes (Kao *et al.*). Além disso, evidencias sugerem que as *eMSCs* descamam preferencialmente em mulheres com endometriose (Gargett; Leyendecker *et al.*).

MENSCS E ENDOMETRIOSE

As células tronco obtidas do fluxo menstrual (*Menstrual Mesenchymal Stem Cells - MenSCs*) exigem procedimentos de baixo custo para obtê-las, são abundantes, de fácil acesso, expandem-se facilmente quando em cultura, e têm potencial de diferenciação em diversas linhagens celulares, que incluem músculo cardíaco e esquelético, e linhagens neuronais (Cui et al., 2007; Meng et al., 2007; Hida et al., 2008; Musina et al., 2008; Patel et al., 2008; Zhong et al., 2009; Khanmohammadi et al., 2012).

A descamação de *eMSCs* durante a menstruação sugere que estas células possam ter uma importante função no início das lesões endometrióticas (Sasson e Taylor, 2008; Gargett e Masuda, 2010; Deane et al., 2013). Gargett e colaboradores (2011) apresentaram dados preliminares de que *eMSCs* parecem descamar preferencialmente no fluxo menstrual e fluido peritoneal de mulheres com endometriose, sugerindo uma função chave no início do desenvolvimento das lesões endometrióticas. Apesar destes estudos, os dados na literatura sobre o isolamento, características moleculares e quantidades das *MenSCs* de pacientes com endometriose comparadas a pacientes saudáveis são escassos.

Além disso, não dispomos de dados consistentes na literatura sobre a quantidade e qualidade dos transcritos produzidas pelas *MenSCs* de pacientes com endometriose comparadas a pacientes saudáveis. Sabe-se que o endométrio eutópico de mulheres com endometriose é uma fonte experimental única e bem estabelecida para investigação de mecanismos moleculares de disfunções reprodutivas e que permite identificar possíveis marcadores específicos para a doença (Kao et al., 2003). Os endométrios eutópico e ectópico de mulheres com endometriose compartilham alterações que não são encontradas no endométrio de mulheres sem endometriose, o que corrobora a ideia de que este endométrio alterado, ao cair na cavidade peritoneal, tem um potencial inicial de desenvolver a doença (Sharpe-Timms, 2001).

INVESTIGAÇÕES DO "OMICS" NA ENDOMETRIOSE E NGS (*NEXT GENERATION SEQUENCE*)

OMICS" significa o estudo global de algo, assim quando este estudo é feito nos genes dizemos genoma, nos RNAs - transcriptoma, nas proteínas - proteoma e nos metabólitos - metaboloma (Figura 1). Para um maior entendimento da fisiopatologia de uma determinada

doença este tipo de investigação global vem sendo muito utilizada (Nair *et al.*, 2004).

Assim devido à complexa fisiopatologia da endometriose, ela é uma doença alvo de interesse para as metodologias “omics” (Siristatidis), e vem sendo largamente estudada em análises globais do genoma (Taylor *et al.*), transcriptoma (Wren *et al.*) e perfil proteico (Poliness *et al.*). Entretanto, o entendimento da origem da doença é "ainda um sonho distante".

A tecnologia de sequenciamento *high-throughput* fornece uma poderosa ferramenta para a análise do transcriptoma (RNA-seq) e trazem grandes vantagens sobre métodos convencionais de *screening*, como os *microarrays*. Embora *microarray* forneça uma avaliação mais rápida de transcritos, ela apresenta limitações como baixa sensibilidade em detectar transcritos raros e presença de falsos positivos devido às hibridações cruzadas entre as sequências com alta homologia. As metodologias de sequenciamento para quantificar níveis diferencialmente expressos de RNAm tem custos reduzidos, maior espectro dos transcritos e, consequentemente, aumentada capacidade de detectar transcritos raros, isoformas raras de *splincing* alternativos, sequencias não codificantes e quantificação direta da abundância dos transcritos (Morozova et al., 2009; Tariq et al., 2011).

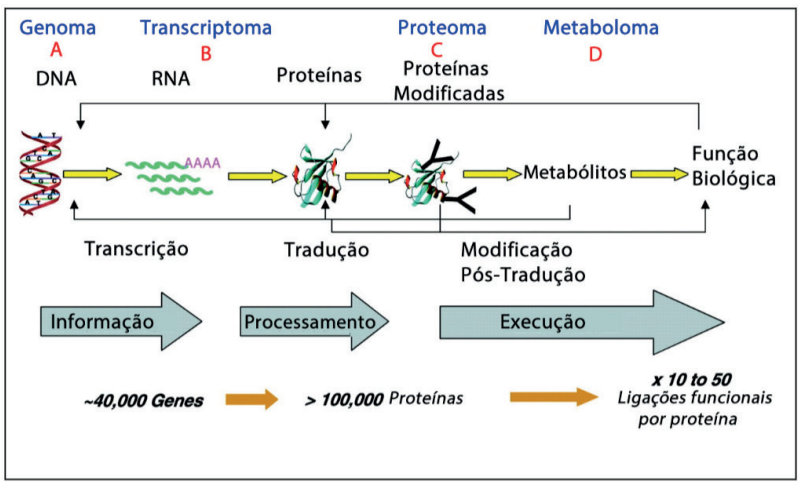


Figura 1: (A) Genoma: é o código genético contido no DNA;(B) Transcriptoma: é conjunto de RNAs carregam a informação genética; (C) Proteoma: é a tradução da informação genética em proteínas; (D) Metaboloma: metabolitos proteicos com função biológica.
Fonte: (Nair et al., 2004).

Sendo assim, a metodologia de sequenciamento de nova geração (NGS) tem sido considerada o "padrão ouro" para quantificação de expressão gênica do transcriptoma completo (Everaert *et al.*). Em tal abordagem, diferentemente dos microarrays, não é necessário ter o conhecimento do conteúdo do transcriptoma estudado, obtendo uma visão sem viés dos conjuntos de transcritos de uma amostra. O método de trabalho de RNA-seq

é dividido em dois tipos. No primeiro, são utilizadas técnicas que alinham os *reads* com um genoma de referência e logo após mapeiam esses *reads*. O segundo tipo são técnicas de pseudoalinhamentos que quebram os *reads* em k-*mers* antes de atribuí-los aos transcritos, resultando em um ganho de velocidade comparado com aos do primeiro tipo (Everaert *et al.*).

Não existem estudos disponíveis que fazem uso de metodologias de *screening* de transcritos diferencialmente expressos em eMSCs obtidas do fluxo menstrual de mulheres com endometriose. Além disso, visto a importante participação do fluxo menstrual para o desenvolvimento e estabelecimento da endometriose, questiona-se a necessidade de estudos comparativos entre os transcritos de eMSC oriundos do fluxo menstrual de mulheres com e sem a doença, a fim de que vias gênicas possivelmente alteradas e novas isoformas para a endometriose, sejam identificadas.

Nesse contexto, o uso de ferramentas de bioinformática tem um papel fundamental na análise do transcriptoma identificado no RNA-seq (Yao Shen et. al., 2017).

FERRAMENTAS ESTATÍSTICAS ACOPLADAS À BIOINFORMÁTICA

Na etapa de cálculo estatístico para identificar genes diferencialmente expressos (DGE) utiliza-se o software R (Team, 2008), que é um sistema para computação estatística e gráficos. É composto de uma linguagem computacional mais um ambiente estatístico de operação com gráficos, acesso a certas funções do sistema e capacidade de rodar comandos armazenados em arquivos (*script*) (Team, 2008).

Bioconductor é um projeto de desenvolvimento de software aberto baseado em R, que provê pacotes para análise e manipulação de dados genômicos (Huber *et al.*, 2015). Um desses pacotes, o EdgeR é utilizado para normalização e expressão diferencial de dados brutos resultantes de análises de RNA-seq (Robinson *et al.*).

O método estatístico no qual o EdgeR foi baseado assumindo uma distribuição binomial negativa dos dados, ou seja, uma distribuição de probabilidades discreta, incluindo estimativas bayesianas, testes exatos, modelos lineares generalizados (GLM) e teste de c^2 (Qui-quadrado). Esse modelo estatístico busca diferenciar a variação biológica da variação técnica, estimando de forma confiável a diferença de expressão gênica entre os grupos estudados em condições com poucas réplicas biológicas (Paul *et al.*, 2017).

Uma abordagem comum que estes métodos estatísticos utilizam é a de testar a hipótese nula de que o valor logarítmico do fold change (Log FC) entre controle e tratamento para expressão gênica é de exatamente zero, significando que aquele gene não foi afetado pelo tratamento. E o objetivo final dessa análise é produzir uma tabela com a lista de genes

analisados passando por vários ajustes, classificados pelo p-valor, chamado também de probabilidade de significância sobre a hipótese nula (Love Mi Fau - Huber *et al.*).

Assim a proposta deste TCC, é fazendo uso de ferramentas de bioinformática (análise *in silico*) definir o perfil diferencial de transcritos das células eMSC obtidas do fluxo menstrual de mulheres com e sem endometriose e relacionar os genes diferencialmente expressos com vias moleculares e funções biológicas que possam estar relacionadas com a etiopatogenia da endometriose.

OBJETIVOS

OBJETIVO GERAL

O objetivo deste trabalho consiste em identificar *in silico* alterações no nível de expressão gênica de células-tronco mesenquimais de endométrio descamadas no fluxo menstrual de mulheres saudáveis e de pacientes com endometriose que poderiam ser relacionados com o desenvolvimento da endometriose. Aplicando metodologias diferentes das previamente utilizadas em Penariol et al., 2022.

OBJETIVOS ESPECÍFICOS

- Identificar os genes diferencialmente expressos (DGE) entre as amostras do fluxo menstrual de controle e com endometriose;
- Comparar os dados do RNAseq com dois métodos de análise do EdgeR (Exato e GLM);
- Definir, por bioinformática, vias de sinalização alterados no estudo comparativo.

MATERIAL E MÉTODOS

Este trabalho de conclusão de curso (TCC) fez parte de um projeto de Auxilio Jovem Pesquisador apoiado pela FAPESP (processo 2013/22431-3) e com aprovação do Comitê de Ética em Pesquisa desta instituição (parecer número CEP/HCFMRP 193.005). Este subprojeto teve início depois que os dados de RNAseq foram gerados.

Brevemente abaixo encontra-se o fluxograma do trabalho que precedeu este subprojeto de TCC e em seguida descrevemos os critérios de inclusão das pacientes incluídas no estudo, para que seja contextualizado de onde os dados foram gerados. Este estudo é uma caso-controle composto por 10 pacientes saudáveis e 10 pacientes diagnosticadas com endometriose.

FLUXOGRAMA DO PROJETO JOVEM PESQUISADOR

O fluxograma do experimento representado na figura 2 contém todos os passos realizados até o momento em que se inicia as análises de bioinformática. Brevemente, consistiu de: 1) seleção de pacientes e coleta da amostra de fluxo menstrual; 2) isolamento das células mesenquimais atendendo os critérios mínimos que definem células estromais mesenquimais multipotentes (aderência a plástico, expressão de antígenos de superfície específicos por citometria de fluxo e potencial de diferenciação em condróciotos, osteócitos e adipócitos) (Dominici *et al.*); 3) Teste de clonogenicidade (CFU - *colony-forming unit*); 4) Isolamento do RNA total das células em cultura na passagem 3; 5) Avaliação de qualidade do RNA por bioanalyzer (Agilent Genomics); 6) A preparação das bibliotecas de cDNA para sequenciamento foram realizadas com o Kit TruSeq®Stranded Total RNA with Ribo-Zero Gold Sample Preparation (Illumina, part#15031048). O primeiro passo da preparação consistiu na remoção do RNA ribossomal (RNAr) usando biotinilação e *beads* magnéticas, neste kit depleda-se o RNAr citoplasmático e mitocondrial, e após purificação, o RNA é fragmentado em pedaços pequenos usando cátions divalentes sob elevadas temperaturas. Os fragmentos de RNA clivados são sintetizados na primeira fita de cDNA (scDNA) usando a transcriptase reversa e *primers* randômicos, seguido da síntese da segunda fita usando a DNA polimerase I e a RNase H. Nestes fragmentos de cDNA, então, são adicionados uma base única de adenina (A) na porção 3' e subsequente ligação aos adaptadores. Os produtos são purificados e enriquecidos por PCR gerando a biblioteca final. As bibliotecas são validadas, normalizadas e misturadas (*pooling*); 7) Foram realizadas 3 corridas de sequenciamento *paired-end* contendo 6 amostras cada (3 controles e 3 endometrioses) distribuídas em 4 *lanes* no equipamento NextSeq 500 da Illumina com os reagentes NextSeq 500/550 Kits v2 (Illumina, FC-404-2004).

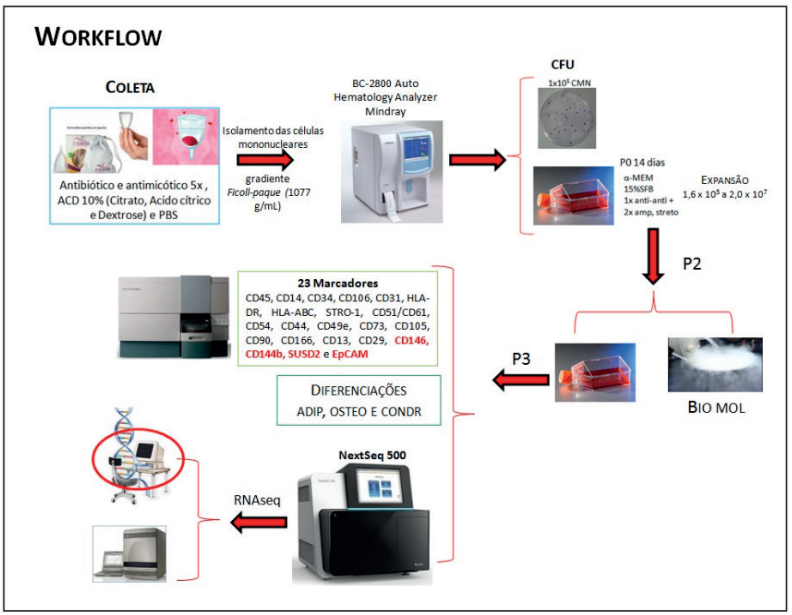


Figura 2: Fluxograma da metodologia realizada. Destacado com um círculo vermelho está o passo em que este trabalho foi iniciado.

CRITÉRIOS DE ELEGIBILIDADE E FLUXOGRAMA DAS PACIENTES

Os critérios adotados para elegibilidade das pacientes foram:

- *Grupo Endometriose*:

Mulheres com diagnóstico videolaparoscopico de endometriose estádio III ou IV segundo os critérios definidos pela *American Society for Reproductive Medicine* (1997). O diagnóstico deve ter sido feito no mínimo a 1 ano antes da coleta.

- *Grupo controle*

Pacientes férteis (com pelo menos 2 filhos vivos) sem história de aborto recorrente, sem diagnóstico clínico e videolaparoscópico de endometriose a no máximo 2 anos antes da coleta, que foram submetidas à videolaparoscopia para laqueadura tubária;

- *Critérios de Elegibilidade Comuns aos dois grupos*

Não obesas (índice de massa corporal - IMC menor ou igual a 30 Kg/m²); com ciclos menstruais regulares (intervalos de 24 a 32 dias ± 3 dias; 2 a 7 dias de duração); idade entre 18 e 40 anos; ausência de doenças sistêmicas tais como: *Diabetes mellitus* ou outras endocrinopatias, doença cardiovascular, lupus eritematoso sistêmico e outras doenças reumatológicas; não ter hábitos tabagistas ou alcoólicos; sem uso de qualquer terapia hormonal há pelo menos 3 meses antes da coleta.

Tratou-se de um estudo observacional caso-controle entre células-tronco mesenquimais obtidas de fluxo menstrual de mulheres com e sem endometriose.

No período de novembro de 2014 a dezembro de 2016 foram analisados 1215 prontuários, sendo 1131 prontuários de mulheres atendidas no Ambulatório de Esterilidade (AEST) do Serviço de Reprodução Humana Assistida do Hospital das Clínicas da FMRP-USP para recrutar o grupo caso e 84 prontuários de mulheres submetidas à videolaparoscopia para laqueadura tubária no Centro de Referência da Saúde da Mulher (MATER) para recrutar o grupo controle.

Do total de prontuários analisados, 54 pacientes apresentavam os critérios de elegibilidade (descritos acima), sendo 20 para o grupo caso e 34 para o grupo controle. Das 20 pacientes do grupo caso selecionadas, efetivamente 17 foram incluídas e coletadas, 7 foram excluídas por contaminação e assim restaram 10 que tiveram as células isoladas, caracterizadas e destas 9 foram sequenciadas. Do grupo controle foram efetivamente incluídas e coletadas 21 pacientes, sendo que 11 foram excluídas por contaminação em cultura, assim 10 tiveram as células isoladas, caracterizadas e destas 9 foram sequenciadas (Figura 3).

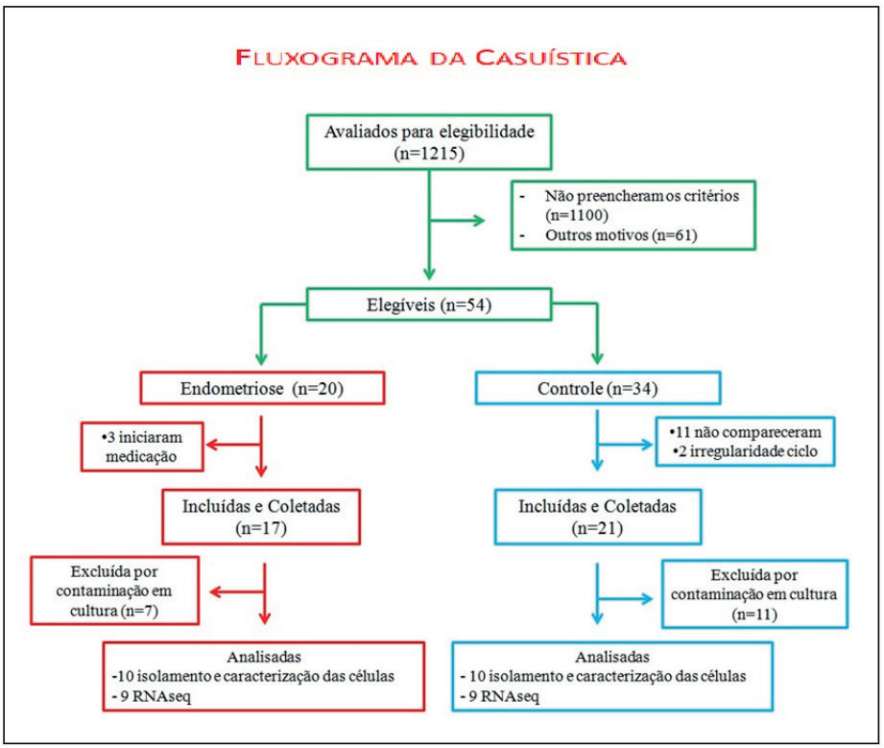


Figura 3: Fluxograma da Casuística. Estudo compreende o período de Nov/2014 aDez/2017.

ANÁLISES POR BIOINFORMÁTICA

Para o pré-processamento de dados e da análise em RNA-Seq deve-se seguir os passos descritos por Dillies e colaboradores em 2012: 1) As sequencias curtas (*short reads*) provenientes do sequenciamento são pré processadas, a fim de remover os adaptadores e as sequências com baixa qualidade, sendo em seguida mapeadas em um genoma de referência ou a um genoma alinhado; 2) O nível de expressão é estimado para cada transcrito (por exemplo, para cada loco); 3) Os dados são normalizados; 4) Uma análise estatística é usada para identificar os transcritos diferencialmente expressos (DEG).

Nossa análise seguiu o padrão de qualidade e fluxo de trabalho apresentado a seguir (Figura 4). Toda a análise foi realizada no sistema operacional CentOS Linux release 7.2.1511.

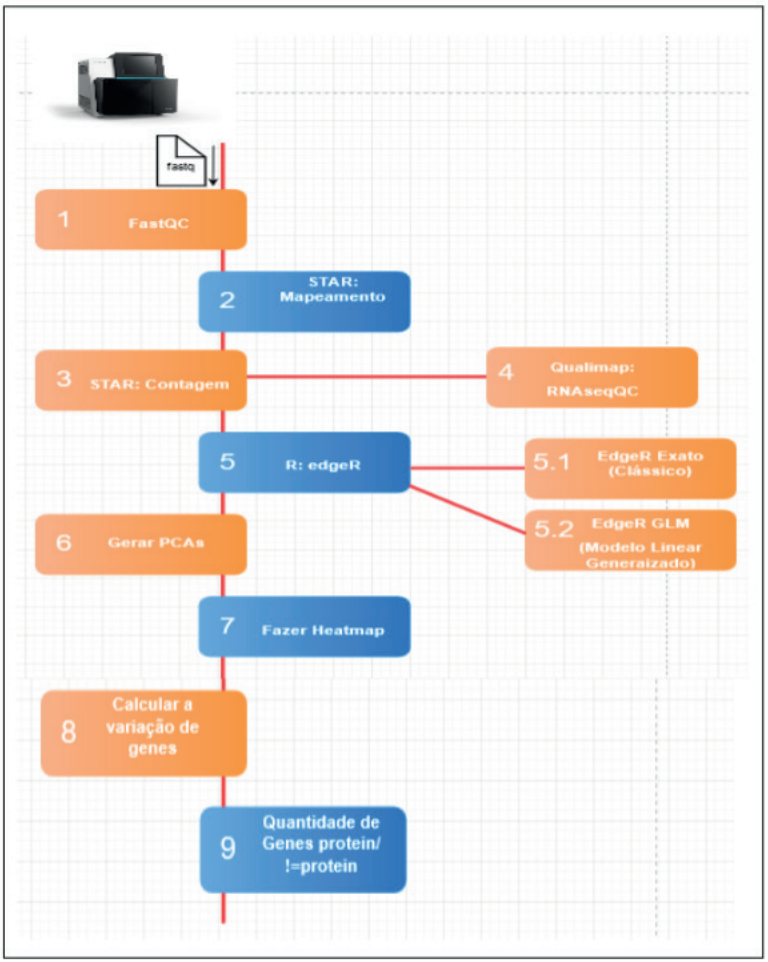


Figura 4: Fluxograma de trabalho com as ferramentas de bioinformática utilizadas nas análises de expressão gênica.

1. FastQC v0.11.5 foi utilizado para verificar a qualidade dos *reads* (sequências) vindos do sequenciamento da Illumina, verificando a qualidade de cada base da sequência; o total de sequências que possui naquela amostra; conteúdo GC; a presença de adaptadores; e duplicações de *reads* (Andrews; S., 2010).

2. Mapeamos as sequências com um genoma de referência usando a ferramenta STAR - *Spliced Transcripts Alignment to a Reference* (Dobin *et al*.) versão 2.5.2a, que é utilizado para mapear o transcriptoma com maior agilidade e apresenta alta precisão e sensitividade para analisar o *splicing* alternativo quando comparado a outros alinhadores. Usamos os o genoma de referência *Genome Reference Consortium human* 38 (GRCh38) (https://www.ncbi.nlm.nih.gov/grc/human), a anotação gênica *release* 85 do Ensembl e os parâmetros "*outFilterMultimapNmax 1*" , "*alignIntronMin 20*" que permitem a seleção das sequências mapeadas com apenas um único hit.

3. Realizamos a contagem de quantos reads mapeados temos por gene utilizando o STAR, com o modo *unstranded* e foi removido resquícios de RNA ribossomal que haviam sido identificados pelo mapeamento dentro ambiente R estatístico. Tabela contendo todos os genes ribossomais no arquivo "ribossomal.csv" contido na pasta de anexo, com o link desta pasta nos anexos complementares.

4. Para avaliarmos os dados de mapeamento e explorarmos melhor as sequências de que foram mapeadas em regiões intrônicas não anotadas (*NoFeature*), e garantirmos que não houve contaminação nas nossas amostras, utilizamos o QualiMap versão 2.2.1 RNA-seq QC (Okonechnikov, 2015) com parâmetros padrão.

5. Utilizamos o R versão 3.3.1 juntamente com o pacote edgeR versão 3.16.5 do bioconductor para encontrarmos os genes diferencialmente expressos. Utilizamos 2 estatísticas possíveis deste pacote:

    5.1. A estatística exata na qual é recomendado se utilizar quando queremos comparar somente 2 grupos, como é o nosso caso, controle e endometriose, só que sem remoção do *batch effect* de corrida de sequenciamento.

    5.2. A estatística do Modelo Linear Generalizado (GLM) que é recomendada para mais grupos, e onde há remoção de batch dentro do desenho experimental.

6. Com o resultados do edgeR fizemos as análises de principais componentes (PCA) que busca combinações lineares dos componentes principais (PCs) que podem efetivamente representar os efeitos das medidas originais. Desta análise geramos três gráficos, um com todas as amostras, um com somente o grupo controle e um somente com grupo endometriose.

7. Fizemos *heatmap* com o pacote padrão do R chamado stats, com os genes de interesse selecionados.

8. No ambiente do R, calculamos o desvio padrão e a média de expressão dos genes utilizando todas as amostras e logo em seguida foi feito o mesmo por grupo controle e grupo endometriose.

9. No ambiente do R, calculamos a quantidade de genes totais que codificam proteínas, e o grupo de genes com outro tipo de anotação no Ensembl.

As informações de todos os pacotes e versões carregados no ambiente do R se encontram em anexo complementar (Anexo 1).

QUALIDADE DOS ARQUIVOS FASTQ

Os dados de qualidade para cada amostra obtidos no FastQC estão descritos na tabela 1. Para a análise de RNA-seq *pair-end* a quantidade de *reads* ideal para que fosse atingida uma boa cobertura das amostras e dos seus transcritos raros seria em média acima de 60 milhões de *reads* (Sheng Q Fau - Vickers *et al*.). Em nossas análises obtivemos em média aproximadamente 62 milhões de leituras, variando de 44 milhões à 86 milhões de sequências. Para o conteúdo de GC total por amostra espera-se cerca de 40 a 60 por cento como ideal (Biostars, 2012), sendo que em nossas amostras sequenciadas foi em média 50 por cento. Além disso, o tamanho dos fragmentos obtidos entre 35-151 bases.

Os níveis aceitáveis quanto ao controle de qualidade das bases das sequências variam de 20 a 36 e, quanto maior for esse valor melhor é qualidade daquela base em específico. Nos nossos dados as bases se encontraram 90 por cento dentro dos valores de 30 a 36 para os dois grupos estudados (controle e endometriose), como pode ser visto nas figuras 5 e 6, respectivamente. A cor de fundo dos gráficos representa o nível de qualidade da leitura das bases; verde: boa, laranja: razoável e vermelho: ruim.

Nome da Amostra	Total de sequências	Pair-end	Porcentagem de GC	Tamanho das Sequências
C10	123700992	61850496	52	35-151
C17	136609990	68304995	53	35-151
C22	155349594	77674797	49	35-151
C29	137634140	68817070	50	35-151
C31	122299216	61149608	52	35-151
C32	120439368	60219684	55	35-151
C34	89969044	44984522	52	35-151
C35	115808730	57904365	54	35-151
C38	159142520	79571260	51	35-151
E2	121175152	60587576	49	35-151
E3	92100346	46050173	56	35-151
E4	118422270	59211135	54	35-151
E7	136779650	68389825	48	35-151
E8	173986868	86993434	54	35-151
E11	111873658	55936829	50	35-151

E12	106348706	53174353	47	35-151
E13	113236056	56618028	55	35-151
E27	115282624	57641312	51	35-151

Tabela 1: Resumo das informações do resultado obtido do FastQC de todas as amostras em uma tabela.

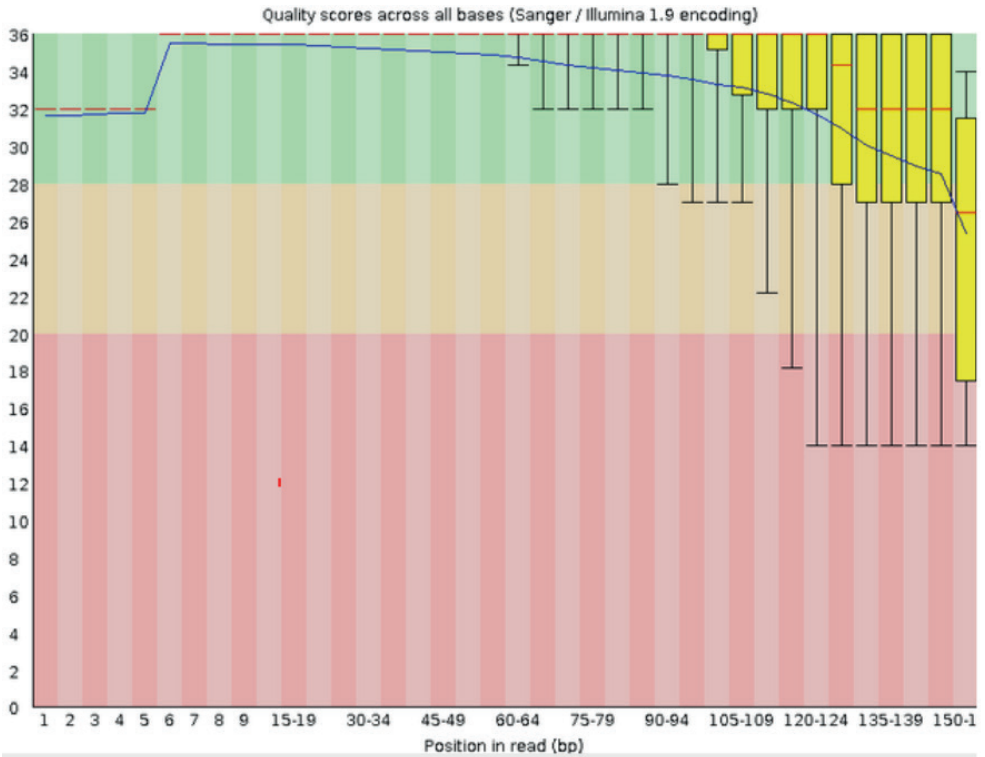


Figura 5: Exemplo do gráfico de qualidade para cada base de uma read para uma amostra grupo controle.

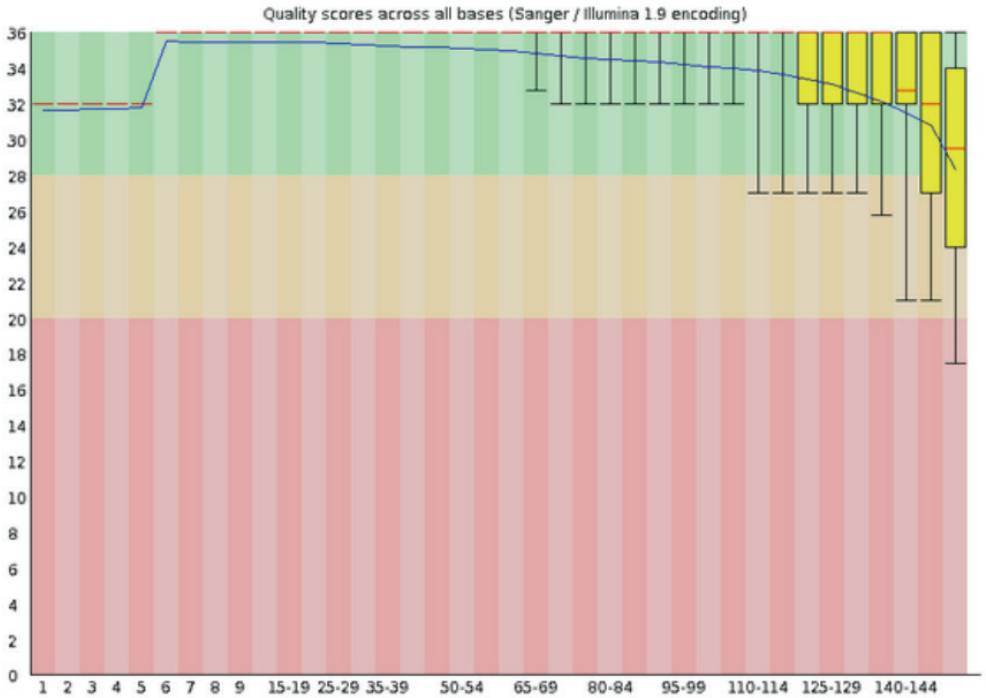


Figura 6: Exemplo de uma amostra do gráfico de qualidade de cada base do grupo endometriose.

MAPEAMENTO E CONTAGEM DAS SEQUÊNCIAS

Os dados de mapeamento e contagem para cada amostra obtidos no STAR estão descritos na tabela 2. Para o resultado do mapeamento do STAR, é esperado que a porcentagem do total de sequências mapeadas sejam em torno de 80 por cento (Seqanswers, 2013) Nossas amostras variaram de 76,73 a 86,12 por cento (média de 80 por cento) de reads mapeadas com o genoma de referência, como pode ser observado na tabela 2.

A quantidade de sequências identificadas como NoFeatures e ambíguas deve ser menor que 20 por cento do total mapeado. Entretanto, seis amostras (C22, C38, E2, E7, E11, E12) das amostras analisadas tiveram quantidade de NoFeatures acima do padrão aceitável (Tarazona *et al*., 2011), variando de 24 a 31 por cento.

A quantidade de *reads* ambíguas para nossas amostras estão em torno de 13,77 por cento.

O mapeamento e contagem com STAR identificou 58051, desses 16.435 genes únicos foram analisados nas nossas amostras, sendo esses genes identificados com o código *Ensembl* único. A tabela completa com todas as contagens se encontra no arquivo

"countsTotal.csv", dentro da pasta de anexo de arquivos disponibilizada nos anexos complementares (anexo 2).

Amostra	Sequências Totais Input	Sequências Mapeados	Porcentagem do total que foi mapeado	NoFeatures	Ambíguas
C10	61850496	50690158	81,96%	7992574	7528932
C17	68304995	55566336	81,35%	6322014	8473367
C22	77674797	64271165	82,74%	15581494	7346639
C29	68817070	58184202	84,55%	9924870	8785913
C31	61149608	51461281	84,16%	6161126	7903522
C32	60219684	50074683	83,15%	3275181	9006390
C34	44984522	38475771	85,53%	5708871	5816135
C35	57904365	44428020	76,73%	6695536	6891337
C38	79571260	65617404	82,46%	16458528	8328055
E2	60587576	52178604	86,12%	15487459	4966138
E3	46050173	37929335	82,37%	1869581	7676990
E4	59211135	48002697	81,07%	7200912	4405698
E7	68389825	58942413	86,19%	18809170	5894988
E8	86993434	71964386	82,72%	5685850	10848668
E11	55936829	46768022	83,61%	11564307	5004363
E12	53174353	45495919	85,56%	16033683	4093697
E13	56618028	47200493	83,37%	3985620	7334664
E27	57641312	47557561	82,51%	9133434	7291621

Tabela 2: Resumo dos Resultados de mapeamento e contagem do STAR.

AVALIAÇÃO DO MAPEAMENTO PELO QUALIMAP RNA-SEQQC

Explorando mais o resultado da quantidade alterada de NoFeatures encontrados no STAR e seguindo o exemplo de um resultado em humanos disponibilizado pelo Qualimap RNA-seq QC(Qualimap, 2015). O esperado para amostras humanas mapeadas seria por volta de 76 por cento em região exônica, 17 por cento em região intrônica e de 6,27 por cento em região intragênica, podendo variar em regiões de sobreposição. Na tabela 3, nossas sequências mapeadas variam nas seguintes regiões: exônicas, com 56,39 à 87,17 por cento, em intrônicas, de 11,57 à 41,29 por cento, em intragênica de 1,17 à 2,61 por cento e em sobreposição com 2,32 à 4,64 por cento. Em destaque temos as amostras que apresentavam quantidade alterada de *NoFeatures*, e podemos observar que essas amostras

possuem aumento de quantidade de sequências em regiões intrônicas, representando um aumento de 11,57 à 26,16 por cento para 30,11 à 41,29 por cento.

Essas regiões intrônicas estão em análise com o software *Cufflinks* versão 2.2.1 (Trapnell *et al.*, 2010).

Amostra	Exonic	Introgenic	Intergenic	Intronic/intergenic overlapping exon:
C10	67,826,489 / 75.39%	20,407,367 / 22.68%	1,729,251 / 1.92%	3,990,392 / 4.44%
C17	79,372,870 / 80.74%	17,267,549 / 17.57%	1,665,282 / 1.69%	4,176,964 / 4.25%
C22	79,401,816 / 67.39%	35,475,278 / 30.11%	2,939,073 / 2.49%	4,587,381 / 3.89%
C29	76,129,017 / 73.94%	24,860,515 / 24.15%	1,965,363 / 1.91%	4,516,831 / 4.39%
C31	72,709,838 / 79.78%	16,873,567 / 18.51%	1,556,395 / 1.71%	4,026,267 / 4.42%
C32	73,883,131 / 85.76%	11,257,041 / 13.07%	1,008,107 / 1.17%	3,812,495 / 4.43%
C34	52,569,626 / 77.23%	14,301,881 / 21.01%	1,197,998 / 1.76%	2,728,986 / 4.01%
C35	60,388,079 / 77.11%	16,356,523 / 20.89%	1,567,821 / 2%	3,078,410 / 3.93%
C38	78,883,628 / 66.54%	36,572,992 / 30.85%	3,097,980 / 2.61%	4,379,937 / 3.69%
E2	60,916,415 / 62.58%	33,920,405 / 34.85%	2,501,663 / 2.57%	3,485,717 / 3.58%
E3	55,469,417 / 87.17%	7,360,982 / 11.57%	805,897 / 1.27%	2,990,074 / 4.7%
E4	70,689,542 / 78.39%	17,969,565 / 19.93%	1,519,489 / 1.68%	3,283,885 / 3.64%
E7	65,657,073 / 60.18%	41,116,608 / 37.69%	2,326,577 / 2.13%	3,680,753 / 3.37%
E8	107,994,299 / 84.27%	18,409,648 / 14.37%	1,742,696 / 1.36%	5,829,439 / 4.55%
E11	57,880,445 / 67.05%	26,543,776 / 30.75%	1,897,788 / 2.2%	3,350,160 / 3.88%
E12	47,955,594 / 56.39%	35,112,413 / 41.29%	1,972,636 / 2.32%	3,058,000 / 3.6%
E13	69,590,037 / 83.38%	12,728,356 / 15.25%	1,140,579 / 1.37%	3,870,380 / 4.64%
E27	59,950,526 / 71.69%	21,875,014 / 26.16%	1,798,667 / 2.15%	3,452,359 / 4.13%

Tabela 3: Resumo dos resultados do Qualimap RNA-seq QC.

ANÁLISE DOS GENES DIFERENCIALMENTE EXPRESSOS UTILIZANDO O EDGER E SEUS MÉTODOS ESTATÍSTICOS

Um grande problema na utilização da inferência Bayesiana pura aplicada a experimentos de RNA-seq está relacionada, novamente, ao baixo conhecimento não só sobre a função de um gene específico em uma condição determinada, mas também ao pouco conhecimento sobre a relação na função de vários genes analisados simultaneamente (The Bayesian Choice, 2nd Ed. (Book Review) (Brief Article)).

Uma possível solução para a estimação da distribuição a priori a ser utilizada em

RNA-Seq é o método Bayesiano Empírico. Neste método, os parâmetros da distribuição a priori são determinados a partir dos próprios dados gerados pelo experimento. Por isso, esse método não é considerado propriamente um método Bayesiano. Atualmente, o número de métodos Bayesianos desenvolvidos para RNA-Seq é muito grande e continua crescendo. Na prática de RNA-Seq, os métodos Bayesianos têm sido utilizados para melhorar a estimação da dispersão dos dados com poucas réplicas biológicas. O edgeR utiliza em ambos exato quanto GLM o método bayesiano empírico (Robinson *et al.*).

Nossas contagens foram corrigidas pelo método TMM do edgeR (Trimmed Mean of M values) (Robinson *et al.*), que corrigiu as possíveis alterações de contagem para cada gene. Nesta técnica calcula-se um fator de correção baseado na média ponderada da variação dos genes sem expressão diferenciada entre as amostras.

EdgeR Exato

Utilizando o teste exato com distribuição binomial negativa, sem remoção de batch, determinamos a expressão diferencial e identificamos através da função *topTags()* dez genes diferencialmente expressos de acordo com o grau de significância e a Taxa de falsa descoberta (FDR). Está função analisa os valores de Fold Change, contagem por milhão, FDR e p-valor para identificar os genes diferencialmente expressos (Robinson *et al.*). Foi encontrado dois genes com o FDR menores que 0.05 e outros oito genes com valores de FDR maiores que 0.05. A tabela com esses dez genes, log dos valores de Fold Change, valores de Pvalor, valores de FDR, log dos valores de contagem por milhão (CPM) e o símbolo gênico deles encontra-se na tabela 4.

Código Ensembl	logFC	logCPM	PValue	FDR	Símbolo do Gene
ENSG00000170627	-5.80339	-2.03017	3.77E-06	0.035471	*GTSF1*
ENSG00000104332	-3.28318	4.849795	4.25E-06	0.035471	*SFRP1*
ENSG00000240583	3.327514	3.194212	1.14E-05	0.063677	*AQP1*
ENSG00000162631	-2.19899	2.334147	2.24E-05	0.078422	*NTNG1*
ENSG00000121898	2.81745	-0.55362	2.35E-05	0.078422	*CPXM2*
ENSG00000227076	1.213101	0.039441	5.12E-05	0.139382	*AL158166.1*
ENSG00000261026	2.710645	0.896529	6.43E-05	0.139382	*AC105046.1*
ENSG00000119508	2.43304	1.313314	6.67E-05	0.139382	*NR4A3*
ENSG00000183098	3.189453	1.679274	9.55E-05	0.160489	GPC6
ENSG00000073756	1.987815	0.256483	0.000105	0.160489	*PTGS2*

Tabela 4: Genes identificados através da função topTags() do edgeR Exato.

EdgeR GLM

Logo em seguida, foi utilizado o teste binomial negativo GLM, com a remoção de batch, para determinar a expressão diferencial, utilizando a função *topTags*() e valores de FDR menores que 0.05, não encontramos nenhum gene (tabela 5).

	logFC	logCPM	LR	PValue	FDR
ENSG00000170627	-5.59835	-2.03037	18.36241	1.83E-05	0.241809
ENSG00000230076	-5.2326	3.53997	17.41372	3.01E-05	0.241809
ENSG00000214146	3.334094	-1.58392	16.68454	4.41E-05	0.241809
ENSG00000227076	1.201172	0.039464	15.93892	6.54E-05	0.268792
ENSG00000162631	-2.03876	2.334146	15.38161	8.78E-05	0.288727
ENSG00000114315	1.944855	5.097257	15.03302	0.000106	0.289383
ENSG00000230202	-2.33267	3.4978	14.26583	0.000159	0.33791
ENSG00000119508	2.296553	1.313339	14.19871	0.000164	0.33791
ENSG00000153234	1.888838	2.511367	13.66964	0.000218	0.365075
ENSG00000261026	2.565177	0.89655	13.63395	0.000222	0.365075

Tabela 5: Genes identificados através da função topTags() do edgeR GLM.

Para a análise de expressão gênica diferencial entre controle e endometriose, o número de genes significativos foi determinado pelo FDR, que corrige o p-valor para múltiplas hipóteses pelo método de taxa de falsas descobertas.

Desta forma, com FDR de um por cento não obtivemos nenhum DEG, obteve-se dois genes com FDR de cinco por cento e oito genes com catorze por cento utilizando o método Exato do edgeR.

valor do FDR	EdgeR Exato	EdgeR GLM
0,01	0	0
0,05	2	0
0,14	8	0

Tabela 6 : valores de FDR e quantidade de genes achados por metodo estatístico do EdgeR.

Como obtivemos apenas dois genes com p-valor ajustado (FDR) menor que 0,05 (Tabela 6) e é descrito que um p-valor não ajustado de 5 por cento é capaz de descobrir verdadeiros positivos e minimizar os falsos (Zhang et al.), investigamos o p-valor dos genes de referência e geramos os gráficos nas figuras 7 e 8 para avaliar o nível de expressão (logaritmo de “fold-change”, logFC), sendo que cada ponto representa um gene e destacados em vermelho são os genes significativamente expressos (Figura 7 e 8).

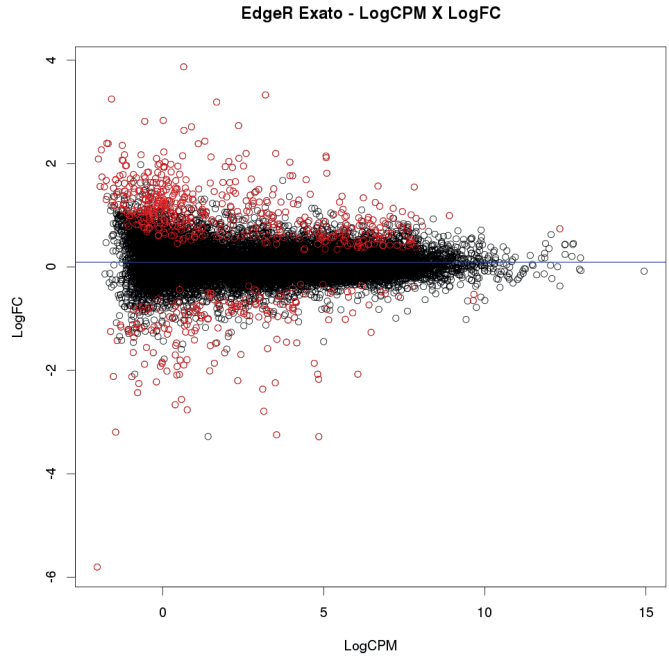


Figura 7: Gráfico com os valores de logFC e logCPM de cada gene identificado no metodo exato do edgeR, sendo os genes em vermelho com os valores de p-valor menores que 0.05.

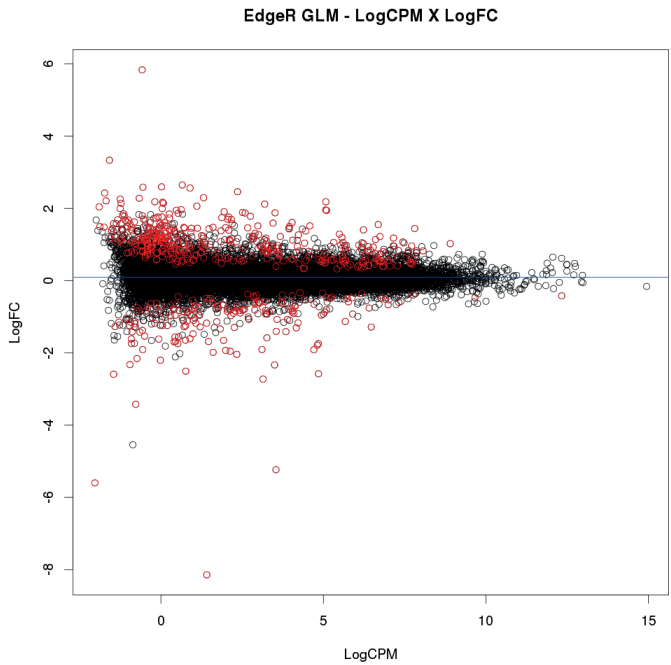


Figura 8: Gráfico com os valores de logFC e logCPM de cada gene identificado no metodo GLM do edgeR, sendo os genes em vermelho com os valores de p-valor menores que 0.05.

No total foram encontrados 16.435 genes, sendo 655 (EdgeR exao) e 542 (EdgeR GLM) genes preditos pelo p-valor 0,05. Desses genes, 490 foram encontrados com ambos métodos do edgeR(Figura 9). O que sinaliza que, para os nossos dados, não obtivemos tanta diferença entre um método e outro.

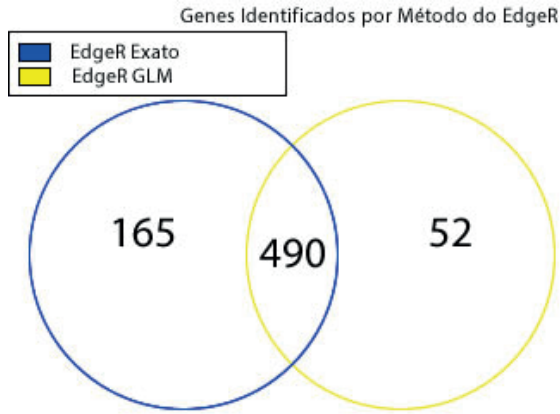


Figura 9: Diagrama de Veen demonstrando quantos genes cada método encontrou e sua intersecção com o p-valor menor que 0,05.

Nos gráficos (Figuras 7 e 8) podemos verificar que a maior parte dos dados foram positivamente expressos (linha azul , logFC~=0), sendo 475 (EdgeR Exato) e 381 (EdgeR GLM), valores obtidos da tabela 7.

Valor do P-Valor	EdgeR Exato	EdgeR GLM
0,01	172	140
0,05	655	542
0,14	1651	1514

Tabela 7: Número de genes encontrados pelo EdgeR exato e EdgeR GLM.

Os dados com os genes menores que 0,05 de ambos métodos " e os dados da intersecção entre os dois métodos estão disponíveis no link do anexo complementar 2 com os nomes "Exact_pvalue_05.csv" e "GLM_pvalue_05.csv e "intersect_pvalue_05.csv".

A ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

É uma abordagem de redução de dimensão clássica que constrói combinações lineares de expressões gênicas, chamadas componentes principais (PCs). Ele busca combinações lineares dos PCs que podem efetivamente representar os efeitos das medidas originais. Os PCs são ortogonais um do outro e podem ter dimensões muito menores do que as medidas originais (Jolliffe, 2014).

Na Figura 10 estão representados os dados de todas as amostras (Controle e Endometriose), observando-se o não agrupamento dos grupos sendo bem divergentes. O que é característico de variabilidade amostral.

A partir desta análise de PCA resolvemos analisar os componentes principais de cada grupo separadamente (Figuras 11 e 12), e observamos que dentro dos grupos individuais existe uma grande variabilidade.

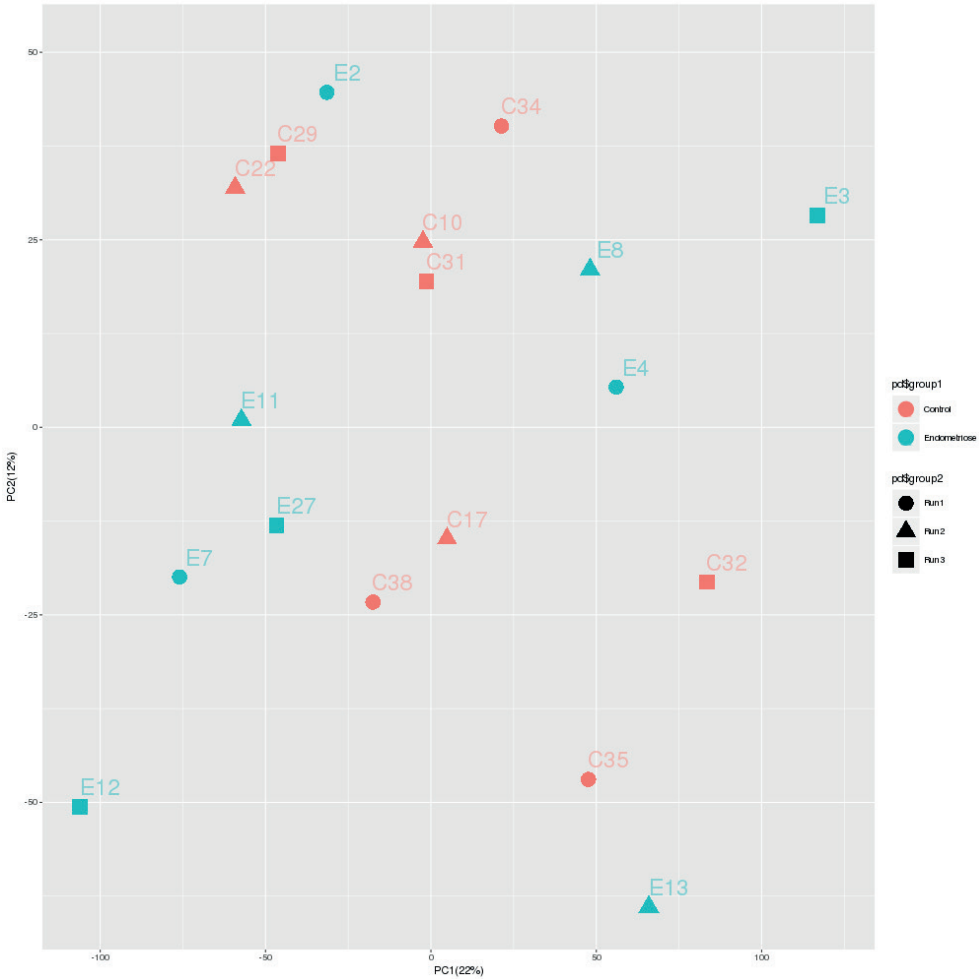


Figura 10: Figura com todos os dados de expressão de todas amostras. Utilizando os dois primeiros principais componentes, (PC1 com 22 por cento de variabilidade e PC2 com 12 por cento de variabilidade) .

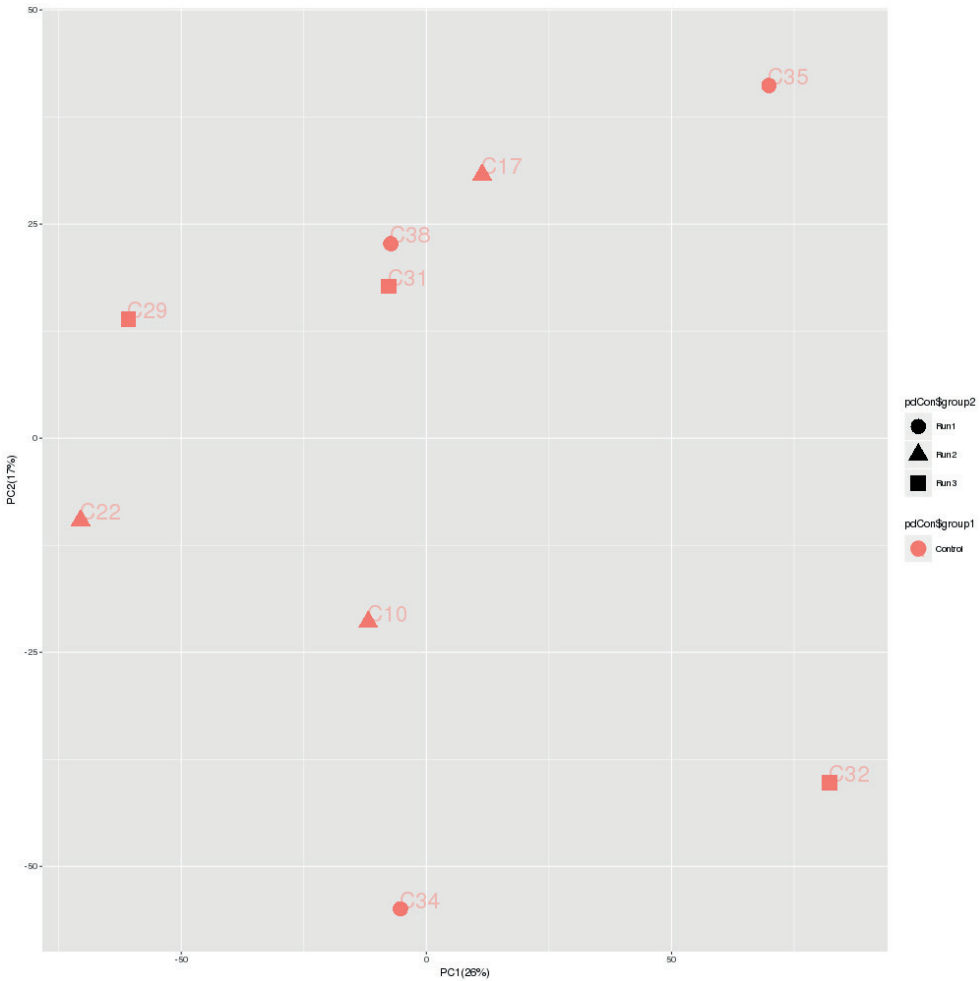


Figura 11: Gráfico com apenas os dados de expressão das amostras controle e utilizando os dois primeiros principais componentes (PC1 com 26 por cento de variabilidade e PC2 com 17 por cento de variabilidade).

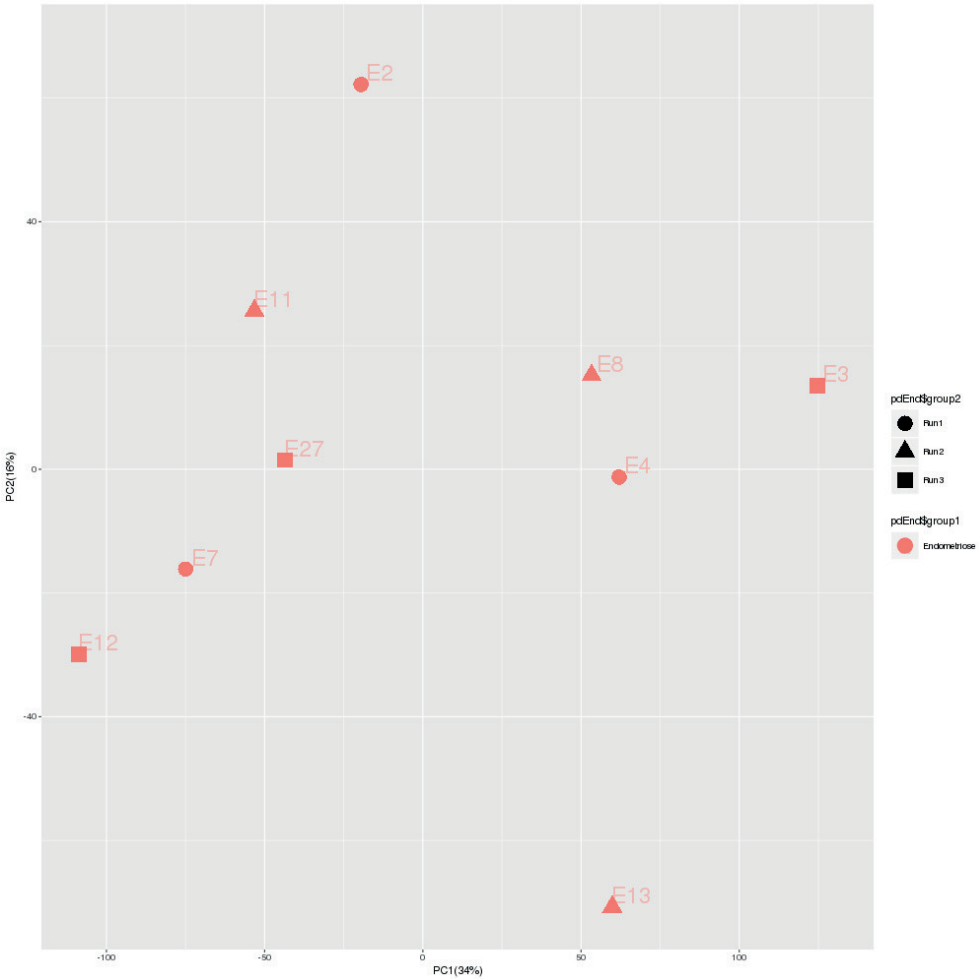


Figura 12: Gráfico com apenas os dados de expressão das amostras Endometriose e utilizando os dois primeiros principais componentes (PC1 com 34 por cento de variabilidade e PC2 com 16 por cento de variabilidade).

HEATMAPS

Para visualizarmos melhor os resultados obtidos com o EdgeR e os PCAs, fizemos os *heatmaps* com os genes encontrados tanto no *topTags*() de ambos métodos, quanto no filtro de p-valor menor que 0.05.

Construímos assim quatro *heatmaps*, dois para o método exato e dois para o método GLM.

O primeiro *heatmap* mostra os 10 genes gerados pelo *topTags*()(Figura 13), onde observamos mistura de amostras na hora da clusterização. O segundo *heatmap* contendo todos os genes com p-valor menor que 0.05 do método exato (Figura 14), observamos que

houve uma melhor separação das amostras , mas com uma das amostras de endometriose no meio do grupo controle.

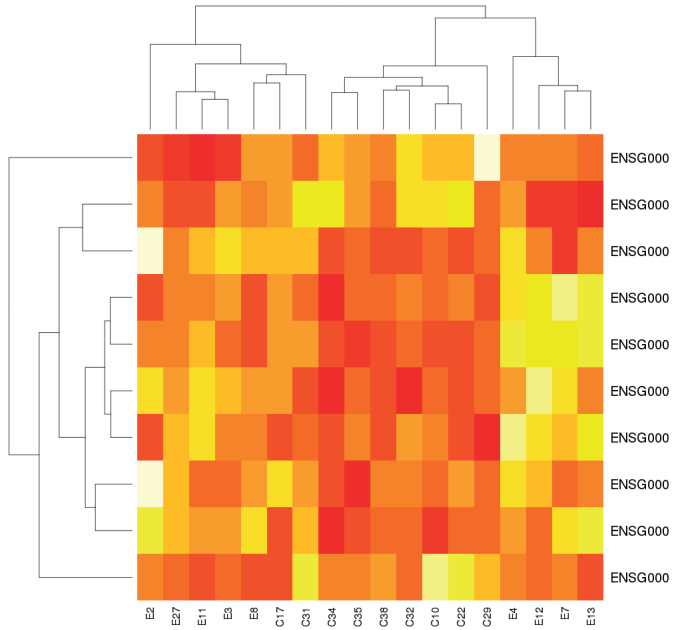


Figura 13: *Heatmap* dos genes gerados pelo *topTags*() do método exato.

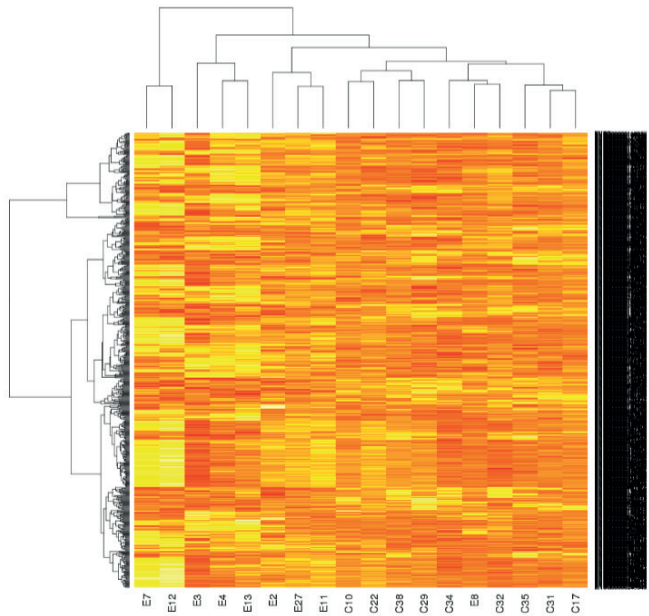


Figura 14: *Heatmap* dos genes com p-valor menores que 0.05 do método exato.

O terceiro *heatmap* mostra os 10 genes gerados pelo *topTags*() (Figura 15) do método GLM, onde observamos mistura das amostras na hora da clusterização, sendo dois controles no grupo endometriose e duas endometriose no grupo controle. O quarto *heatmap* contendo todos os genes com p-valor menor que 0.05 do método GLM (Figura 16), observamos que houve uma boa separação das amostras, so havendo a troca de uma amostra controle com uma endometriose.

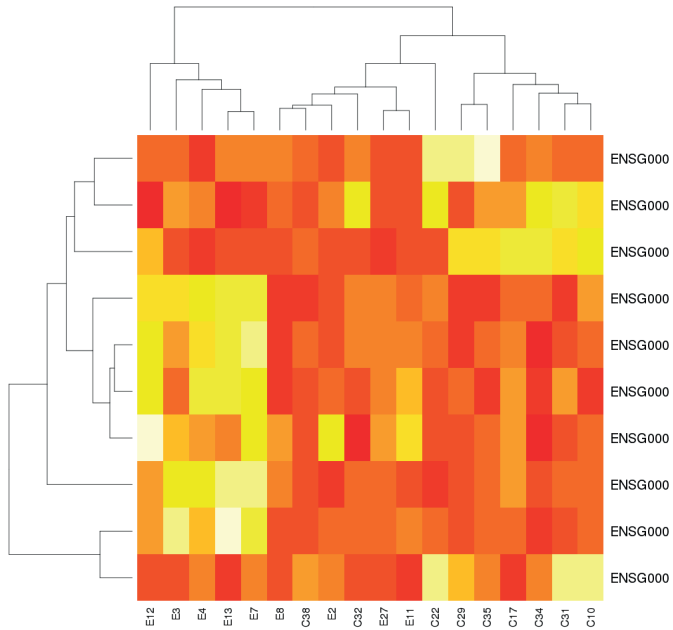


Figura 15: *Heatmap* dos genes gerados pelo *topTags*() do método GLM.

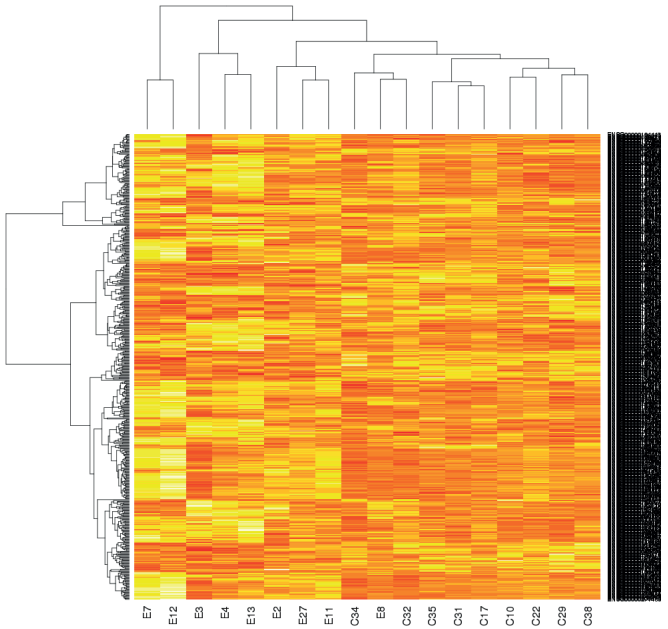


Figura 16: *Heatmap* dos genes com p-valor menores que 0.05 do método GLM.

Mas em todos os *heatmap*s, não encontramos um perfil gênico para os grupos utilizando os genes selecionados.

Utilizando os genes gerados pelo *topTags*(), de ambos os métodos, colocamos no DAVID (Dennis *et al., 2003*), e obtivemos as informações com o que os genes estão relacionado. No caso do exato temos 8 genes e no caso do GLM temos 6, obtivemos também genes que o DAVID não gerou resultado. (Figuras 17 e 18)

ENSEMBL_GENE_ID	Gene Name	Related Genes	Species
ENSG00000073756	prostaglandin-endoperoxide synthase 2(PTGS2)	RG	Homo sapiens
ENSG00000104332	secreted frizzled related protein 1(SFRP1)	RG	Homo sapiens
ENSG00000119508	nuclear receptor subfamily 4 group A member 3(NR4A3)	RG	Homo sapiens
ENSG00000121898	carboxypeptidase X, M14 family member 2(CPXM2)	RG	Homo sapiens
ENSG00000162631	netrin G1(NTNG1)	RG	Homo sapiens
ENSG00000170627	gametocyte specific factor 1(GTSF1)	RG	Homo sapiens
ENSG00000183098	glypican 6(GPC6)	RG	Homo sapiens
ENSG00000240583	aquaporin 1 (Colton blood group)(AQP1)	RG	Homo sapiens
ENSG00000261026			
ENSG00000227076			

Figura 17: Resultado obtido do DAVID utilizando os 10 genes obtidos pela função topTags() na análise EdgeR exata.

ENSEMBL_GENE_ID	Gene Name	Related Genes	Species
ENSG00000114315	hes family bHLH transcription factor 1(HES1)	RG	Homo sapiens
ENSG00000119508	nuclear receptor subfamily 4 group A member 3(NR4A3)	RG	Homo sapiens
ENSG00000153234	nuclear receptor subfamily 4 group A member 2(NR4A2)	RG	Homo sapiens
ENSG00000162631	netrin G1(NTNG1)	RG	Homo sapiens
ENSG00000170627	gametocyte specific factor 1(GTSF1)	RG	Homo sapiens
ENSG00000214146	uncharacterized LOC647323(LOC647323)	RG	Homo sapiens
ENSG00000230202			
ENSG00000230076			
ENSG00000227076			
ENSG00000261026			

Figura 18: Resultado obtido do DAVID utilizando os 10 genes obtidos pela função topTags() na análise EdgeR GLM.

DESVIO PADRÃO E MÉDIA DE EXPRESSÃO GÊNICA

Calculamos o desvio padrão e média da expressão dos genes com todas as amostras (Tabela 8), depois calculamos por grupo (Tabelas 9 e 10). Com esta análise afirmamos que realmente havia uma alta variação dentro de cada grupo, o que pode ser uma das explicações de não obtermos genes significativamente relevantes (FDR menor que 0.05).

	desvio padrao	media
ENSG00000227232	0.454207553	0.657466897
ENSG00000279457	0.396546848	1.982163203
ENSG00000225972	2.057062886	-1.037261116
ENSG00000225630	0.749339833	0.256392612
ENSG00000237973	0.961878531	4.949649863
ENSG00000229344	0.469368308	1.629656609
ENSG00000248527	0.405137336	6.014908758
ENSG00000198744	0.569419116	2.384608658
ENSG00000228327	0.371940215	0.542687099
ENSG00000237491	1.009472719	0.543246844
ENSG00000228794	0.487841099	2.967007173
ENSG00000225880	0.407506712	-0.322271414
ENSG00000223764	1.707435399	0.692883582
ENSG00000187634	1.714717917	2.707813428
ENSG00000188976	0.68509813	6.007445109
ENSG00000187961	0.498587294	3.142540749

ENSG00000187583	0.816502395	0.085737678
ENSG00000188290	0.504596414	2.425944446
ENSG00000187608	0.619332995	5.565802278

Tabela 8: Informação de desvio padrão e média de todos os genes encontrados em todas as amostras.

	desvio padrao	media
ENSG00000227232	0.611531256	0.58944334
ENSG00000279457	0.520082583	1.975093175
ENSG00000225972	2.119126353	-1.187997171
ENSG00000225630	0.545109942	0.038217782
ENSG00000237973	1.165461754	4.909902845
ENSG00000229344	0.546682162	1.532171279
ENSG00000248527	0.459438613	5.960122018
ENSG00000198744	0.676467231	2.463807238
ENSG00000228327	0.467403235	0.459343927
ENSG00000237491	0.698440627	0.505233385
ENSG00000228794	0.312028522	2.829144676
ENSG00000225880	0.43766408	-0.475717759
ENSG00000223764	1.462333709	0.583273424
ENSG00000187634	1.382296561	2.597743667
ENSG00000188976	0.460169545	6.012130383
ENSG00000187961	0.407734148	3.155584868
ENSG00000187583	0.404735284	0.173120005
ENSG00000188290	0.486149463	2.311900145
ENSG00000187608	0.605503127	5.624339989

Tabela 9: Informações de desvio padrão e media de todos os genes encontrados nas amostras do grupo controle.

	desvio padrao	media
ENSG00000227232	0.232412088	0.725490455
ENSG00000279457	0.252104386	1.989233232
ENSG00000225972	2.109534153	-0.886525061
ENSG00000225630	0.888236296	0.474567442

ENSG00000237973	0.77731345	4.989396881
ENSG00000229344	0.384587481	1.727141939
ENSG00000248527	0.361873108	6.069695497
ENSG00000198744	0.46613858	2.305410077
ENSG00000228327	0.244697801	0.62603027
ENSG00000237491	1.293978023	0.581260303
ENSG00000228794	0.604649519	3.10486967
ENSG00000225880	0.329170627	-0.168825069
ENSG00000223764	2.007395326	0.80249374
ENSG00000187634	2.07606453	2.817883189
ENSG00000188976	0.886331442	6.002759835
ENSG00000187961	0.601350372	3.12949663
ENSG00000187583	1.111618609	-0.00164465
ENSG00000188290	0.524840207	2.539988746
ENSG00000187608	0.663889567	5.507264568

Tabela 10: Informações de desvio padrão e media dos genes encontrados nas amostras do grupo endometriose.

SEPARAÇÃO DE GENES POR ANOTAÇÃO DO ENSEMBL

Dos 16435 genes identificados, separamos em dois grupos, no qual o primeiro seria o grupo com genes que possuem a anotação do Ensembl como *"gene protein"* e o segundo seria o grupo que contêm outro tipo de anotação no Ensembl. Identificamos 12873 genes no primeiro grupo e 3562 no segundo (Tabela 12). A anotação dos genes do segundo grupo com todos os tipos possíveis está no arquivo "noProteinType.csv" no anexo complementar 2.

	Gene Protein	Outro tipo de anotação
Quantidade de Genes	12873	3562

Tabela 11: Quantidade de gene Protein e quantidade de gene com outro tipo de anotação.

CONCLUSÃO

1. Com os dados observados de diferentes métodos estatísticos utilizados pelo EdgeR, pelas análises de componentes principais, *heatmap*s e cálculo dos desvios padrões e médias por genes, podemos indicar que com esses dados não observamos um perfil de expressão genica diferencial entre os grupos estudados
2. Os dois métodos de análise utilizados neste estudo apresentarem resultados semelhantes.
3. Os 10 genes encontrados pela função *topTags*() pra cada método.

LIMITAÇÕES DO ESTUDO

Este estudo caracteriza-se como um estudo piloto. É possível que as diferenças na expressão entre as MenSCs de mulheres com e sem endometriose sejam discretas não havendo um grande número de genes alterados. Ressaltamos, que os dados brutos utilizados aqui foram previamente publicados (PRJNA884641) e avaliados por metodologias diferentes da apresentada aqui. Assim, as abordagens de *trimagem* e montagem do genoma resultaram em desfechos diferentes.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDREWS, S. FastQC A Quality Control tool for High Throughput Sequence Data. **http://www.bioinformatics.babraham.ac.uk/projects/fastqc/**, Disponível em: < http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ >.

BIOSTARS. **What Is The Gc-Content Across Different Human Chromosomes?** 2012.

BISCHOFF, F.; SIMPSON, J. L. Genetic basis of endometriosis. n. 0077-8923 (Print),

Penariol LBC, Thomé CH, Tozetti PA, Paier CRK, Buono FO, Peronni KC, Orellana MD, Covas DT, Moraes MEA, Silva WA Jr, Rosa-E-Silva JC, Ferriani RA, Faça VM, Poli-Neto OB, Tiezzi DG, Meola J. What Do the Transcriptome and Proteome of Menstrual Blood-Derived Mesenchymal Stem Cells Tell Us about Endometriosis? Int J Mol Sci. 2022 Sep 29;23(19):11515. doi: 10.3390/ijms231911515

BULUN, S. E. Endometriosis. n. 1533-4406 (Electronic),

CHAN, R. W.; NG EH FAU - YEUNG, W. S. B.; YEUNG, W. S. Identification of cells with colony-forming activity, self-renewal capacity, and multipotency in ovarian endometriosis. n. 1525-2191 (Electronic),

CHAN, R. W.; SCHWAB KE FAU - GARGETT, C. E.; GARGETT, C. E. Clonogenicity of human endometrial epithelial and stromal cells. n. 0006-3363 (Print),

DENNIS, G., JR. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. n. 1474-760X (Electronic),

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics,** v. 29, n. 1, p. 15-21, ISSN 1367-4803.

DOMINICI, M. et al. Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. n. 1465-3249 (Print),

ESKENAZI, B.; WARNER, M. L. EPIDEMIOLOGY OF ENDOMETRIOSIS. **Obstetrics and Gynecology Clinics of North America,** v. 24, n. 2, p. 235-258, ISSN 0889-8545.

EVERAERT, C. et al. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. n. 2045-2322 (Electronic),

GARGETT, C. E. Uterine stem cells: what is the evidence? , n. 1355-4786 (Print),

GARGETT, C. E.; MASUDA, H. Adult stem cells in the endometrium. **Molecular human reproduction,** v. 16, n. 11, p. 818,

HALME J FAU - HAMMOND, M. G. et al. Retrograde menstruation in healthy women and in patients with endometriosis. n. 0029-7844 (Print),

HUBER, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. **Nat Meth,** v. 12, n. 2, p. 115-121, 2015. ISSN 1548-7091. Disponível em: <http://dx.doi.org/10.1038/nmeth.3252>. Disponível em: < 10.1038/nmeth.3252 >.

HWANG, J. H. et al. Identification of biomarkers for endometriosis in eutopic endometrial cells from patients with endometriosis using a proteomics approach. n. 1791-3004 (Electronic),

J, T. **Uterine stem cells.** RUEDA BR, P. J. StemBook [Internet]: bridge (MA): Harvard Stem Cell Institute 2008.

JABBOUR, H. N. et al. Endocrine regulation of menstruation. n. 0163-769X (Print),

JOLLIFFE, I. Principal Component Analysis. In: (Ed.). **Wiley StatsRef: Statistics Reference Online**: John Wiley & Sons, Ltd, 2014. ISBN 9781118445112.

KAO, A. P. et al. Comparative study of human eutopic and ectopic endometrial mesenchymal stem cells and the development of an in vivo endometriotic invasion model. n. 1556-5653 (Electronic),

LEYENDECKER, G. et al. Endometriosis results from the dislocation of basal endometrium. n. 0268-1161 (Print),

LOVE MI FAU - HUBER, W.; HUBER W FAU - ANDERS, S.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. n. 1474-760X (Electronic),

MACER, M. L.; TAYLOR, H. S. Endometriosis and infertility: a review of the pathogenesis and treatment of endometriosis-associated infertility. n. 1558-0474 (Electronic),

MASUDA, H. et al. Stem Cell-Like Properties of the Endometrial Side Population: Implication in Endometrial Regeneration. **PLOS ONE,** v. 5, n. 4, p. e10387, 2010. Disponível em: < https://doi.org/10.1371/journal.pone.0010387 >.

NAIR, K. S. et al. Proteomic research: potential opportunities for clinical and physiological investigators. **American Journal of Physiology - Endocrinology And Metabolism,** v. 286, n. 6, p. E863-E874, 2004.

OKONECHNIKOV, K., CONESA, A., & GARCÍA-ALCALDE, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics, btv566 2015.

OKULICZ, W. C.; ACE CI FAU - SCARRELL, R.; SCARRELL, R. Zonal changes in proliferation in the rhesus endometrium during the late secretory phase and menses. n. 0037-9727 (Print),

OZKAN, S.; MURK W FAU - ARICI, A.; ARICI, A. Endometriosis and infertility: epidemiology and evidence-based treatments. n. 0077-8923 (Print),

PAUL, J. A. et al. Transcriptomic analysis of gene signatures associated with sickle pain. v. 4, p. 170051, 2017. Disponível em: < http://dx.doi.org/10.1038/sdata.2017.51 >.Disponível em: < 10.1038/sdata.2017.51 >.

PAULA GABRIELA MARIN, F.; FIGUEIRA, P. G. M. **Stem cells in endometrium and their role in the pathogenesis of endometriosis**. ABRÃO, M. S.;GRACIELA, K.*, et al*. New York: New York. 1221**:** 10-17 p. 2011.

POLINESS, A. E. et al. Proteomic approaches in endometriosis research. n. 1615-9853 (Print),

QUALIMAP. **Qualimap Report: RNA Seq QC : kidney, human** 2015.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics,** v. 26, n. 1, p. 139-140, ISSN 1367-4803.

S., A. **FastQC: a quality control tool for high throughput sequence data.** 2010.

SAMPSON, J. A. Peritoneal endometriosis due to the menstrual dissemination of endometrial tissue into the peritoneal cavity. **American Journal of Obstetrics & Gynecology,** v. 14, n. 4, p. 422-469, 2017/06/19 ISSN 0002-9378. Disponível em: < http://dx.doi.org/10.1016/S0002-9378(15)30003-X >.

SASSON, I. E.; TAYLOR, H. S. Stem cells and the pathogenesis of endometriosis. n. 0077-8923 (Print),

SEQANSWERS. **Typical alignment mapping percentage with genome?** 2013.

SHENG Q FAU - VICKERS, K. et al. Multi-perspective quality control of Illumina RNA sequencing data analysis. LID - elw035 [pii]. n. 2041-2657 (Electronic),

SIGNORILE, P. G.; BALDI, A. Endometriosis: new concepts in the pathogenesis. n. 1878-5875 (Electronic),

SIRISTATIDIS, C. S. What have the 'omics done for endometriosis? , n. 1643-3750 (Electronic),

SPENCER, T. E. et al. Comparative developmental biology of the mammalian uterus. n. 0070-2153 (Print),

TARAZONA, S. et al. Differential expression in RNA-seq: A matter of depth. **Genome Research,** v. 21, n. 12, p. 2213-2223, 2011. ISSN 1088-9051-1549-5469. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227109/ >.

TAYLOR, R. N.; LUNDEEN SG FAU - GIUDICE, L. C.; GIUDICE, L. C. Emerging role of genomics in endometriosis research. n. 0015-0282 (Print),

TEAM, R. D. C. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria 2008.

The Bayesian Choice, 2nd Ed.(Book Review)(Brief Article). 40**:** 235 p.

TRAPNELL, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nat Biotech,** v. 28, n. 5, p. 511-515, 2010. ISSN 1087-0156. Disponível em: < http://dx.doi.org/10.1038/nbt.1621 >.Disponível em: < http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1621.html#supplementary-information >.

WOOD RUSSELL, W. Aberrant portions of the müllerian duct found in an ovary. **American Journal of Obstetrics and Gynecology,** v. 134, n. 2, p. 225-226, ISSN 0002-9378.

WREN, J. D.; WU Y FAU - GUO, S.-W.; GUO, S. W. A system-wide analysis of differentially expressed genes in ectopic and eutopic endometrium. n. 0268-1161 (Print),

WU, Y. et al. Resolution of clonal origins for endometriotic lesions using laser capture microdissection and the human androgen receptor (HUMARA) assay. n. 0015-0282 (Print),

ZHANG, Z. H. et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. n. 1932-6203 (Electronic),

ANEXO COMPLEMENTAR 1

```
> sessionInfo()
R version 3.3.1 (2016-06-21)
Platform: x86_64-redhat-linux-gnu (64-bit)
Running under: CentOS Linux 7 (Core)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] RColorBrewer_1.1-2  pheatmap_1.0.8      NMF_0.20.6
 [4] Biobase_2.34.0      BiocGenerics_0.20.0 cluster_2.0.5
 [7] rngtools_1.2.4      pkgmaker_0.22       registry_0.3
[10] ggplot2_2.2.1       edgeR_3.16.5        limma_3.30.13

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9         magrittr_1.5        splines_3.3.1       doParallel_1.0.10
 [5] munsell_0.4.3       colorspace_1.3-2    xtable_1.8-2        gridBase_0.4-7
 [9] lattice_0.20-35     rlang_0.1.1         foreach_1.4.3       stringr_1.2.0
[13] plyr_1.8.4          tools_3.3.1         grid_3.3.1          gtable_0.2.0
[17] iterators_1.0.8     lazyeval_0.2.0      digest_0.6.12       tibble_1.3.3
[21] reshape2_1.4.2      codetools_0.2-15    labeling_0.3        stringi_1.1.2
[25] scales_0.4.1        locfit_1.5-9.1
```

ANEXO COMPLEMENTAR 2

Pasta de arquivos disponíveis no link:

https://drive.google.com/drive/folders/1Jssquq2u4vRKpyi5SlWtmoUSm92-ROrP?usp=sharing

Aplicação de ferramentas de bioinformática para ANÁLISE DE EXPRESSÃO GÊNICA POR RNA-SEQ em células-tronco derivadas de fluxo menstrual (MenSCs) de mulheres com e sem endometriose

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2023

Aplicação de ferramentas
de bioinformática para

ANÁLISE DE EXPRESSÃO GÊNICA POR RNA-SEQ

em células-tronco derivadas
de fluxo menstrual (MenSCs)
de mulheres com e sem
endometriose

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Atena Editora
Ano 2023