CLUSTERS NA LITERATURA CIENTÍFICA:

GENÉTICA DA ENDOMETRIOSE

SANTO ANDRÉ, SP

2023

CLUSTERS NA LITERATURA CIENTÍFICA:

GENÉTICA DA ENDOMETRIOSE

Trabalho de Conclusão de Curso apresentado na Universidade Federal do ABC, para conclusão do Bacharelado em Ciências Biológicas

SANTO ANDRÉ, SP

2023

FOLHA DE APROVAÇÃO

Hera Campeche Cruz

CLUSTERS NA LITERATURA CIENTÍFICA:

GENÉTICA DA ENDOMETRIOSE

Trabalho de Conclusão de Curso apresentado na Universidade Federal do ABC, para conclusão do Bacharelado em Ciências Biológicas

Aprovado em: 27/04/2023

BANCA EXAMINADORA:

_____

Prof(a). Dr(a). Danilo Trabuco do Amaral (orientador)

_____

Prof(a). Dr(a). Marcella Pecora Milazzotto

_____

Prof(a). Dr(a). Milca Rachel da Costa Ribeiro Lins

SANTO ANDRÉ, SP

2023

AGRADECIMENTOS

Gostaria de expressar meus sinceros agradecimentos a todas as pessoas e organizações que tornaram este projeto possível. Agradeço profundamente ao meu orientador, Danilo, que aceitou me orientar nesse projeto e esteve sempre disponível, auxiliando em todas as dúvidas e contribuindo para a construção de todo conhecimento de bioinformática aqui exposto. Agradeço também aos professores que me auxiliaram na escrita do projeto, professora Dra. Bianca Banco, da Faculdade de Medicina do ABC, especialista em Genética e Reprodução Humana, com pesquisa na área da Endometriose, e professor Dr. Jesus Mena-Chalco, da UFABC, especialista na área de Processamento de Linguagem Natural e Mineração de Grafos. Agradeço aos meus professores em geral, desde a infância, incentivando minha curiosidade e sede de aprender, me permitindo construir os conhecimentos que me trouxeram até aqui e que me levam a continuar nessa busca.

Agradeço a UFABC, por me proporcionar experiências incríveis, como a mobilidade acadêmica, na qual cursei Biotecnologia Medicinal e pude aprender mais ainda sobre bioinformática e aplicar nesse projeto. Agradeço à Escola Superior de Saúde, no Porto, Portugal, que me acolheu por um semestre e onde aprendi muito.

Agradeço aos meus pais, Chirles e Hans, e minha irmã, Hanna, por todo apoio aos meus estudos, auxílios diversos e incentivos. Agradeço aos meus padrinhos, Luciana e Glayson, por me incentivarem nos projetos e me alertarem sobre a endometriose, doença que me acomete assim como à minha madrinha, e darem toda força para que eu executasse essa pesquisa, a qual pode auxiliar na detecção da doença por um diagnóstico precoce. Agradeço a toda minha família que sempre me incentivou nos estudos, me deu suporte e me fez acreditar que era possível. Agradeço imensamente aos meus amigos, que me apoiaram nos momentos difíceis, me ajudaram em etapas do projeto com conhecimentos compartilhados, que deram força quando necessário e comemoraram juntos as minhas conquistas.

E por último, agradeço imensamente a Deus e aos meus guias espirituais, que através da fé, me deram força e iluminação para seguir os meus sonhos e executar esse projeto.

Obrigada!

RESUMO

Endometriose é uma inflamação crônica que representa uma das doenças ginecológicas benignas mais comuns e que afeta mulheres em idade reprodutiva. Caracteriza-se pela presença de tecido histologicamente similar ao endométrio fora da cavidade uterina, podendo causar dor pélvica crônica, dismenorreia, dispareunia e infertilidade. Essa condição impacta significativamente a qualidade de vida das mulheres e sua causa ainda não foi definida, mas existem várias teorias que relacionam aspectos clínicos, imunológicos, genéticos e epigenéticos. Sendo assim, esse projeto buscou analisar a literatura científica dos últimos dez anos sobre a genética da endometriose, desenvolvendo um algoritmo com o uso de algumas ferramentas de Processamento de Linguagem Natural (PLN) na linguagem Python, a fim de identificar os contextos onde se agrupam termos de interesse sobre a genética da endometriose. Entre os 711 artigos coletados e processados, foram encontrados, através da busca por expressão regular, um total de 1960 termos, dos quais 1302 foram identificados como genes humanos e anotados por comandos de rotina em *bash*, através da busca por similaridade com o banco de dados Genbank para genes humanos. Com a identificação desses, foi possível realizar a busca nos textos analisados, formando-se uma matriz de presença e ausência, e, posteriormente, a clusterização, através de um algoritmo *k-means*, que é amplamente utilizado para agrupar dados em clusters com base em suas semelhanças. Os termos foram agrupados desde k=2 até k=5, sendo k a quantidade de núcleos dos *clusters*; constatou-se que quando k=5, há melhor distribuição dos termos, e delimitam-se dois grupos bem concentrados, que podem ser genes envolvidos na mesma rota metabólica ou, então, que estejam relacionados a diferentes vias da doença, mas que apresentam padrão de expressão gênica similar. Apesar do estudo ter sido capaz de definir associações entre alguns genes de interesse na genética da endometriose, há alguns desafios futuros, como aprimorar os comandos para busca automatizada e anotação gênica, delimitar melhor os grupos e associar os termos com as vias metabólicas a qual pertencem pode auxiliar a compreender ainda melhor a doença e suas vias de ação, permitindo, eventualmente, um diagnóstico precoce.

Palavras-chave: endometriose; genética da endometriose; revisão de literatura; análise de *cluster*.

ABSTRACT

Endometriosis is a chronic inflammation that represents one of the most common benign gynecological diseases affecting women of reproductive age. It is characterized by the presence of tissue histologically similar to the endometrium outside the uterine cavity, which can cause chronic pelvic pain, dysmenorrhea, dyspareunia, and infertility. This condition significantly impacts women's quality of life, and its pathogenesis has not yet been established, although several theories relate to clinical, immunological, genetic, and epigenetic aspects. Therefore, this project aimed to analyze the scientific literature of the last ten years on the genetics of endometriosis, developing an algorithm using Natural Language Processing (NLP) tools in Python language, to identify contexts where terms of interest on the genetics of endometriosis are clustered. Among the 711 collected and processed articles, a total of 1960 terms were found through the created regular expression search, of which 1302 were identified as human genes and annotated by routine bash commands through similarity search with the Genbank database for human genes. With the identification of these, it was possible to search the analyzed texts, forming a presence-absence matrix, then, clustering was performed through the k-means algorithm, which is widely used to group data into clusters based on their similarities. The terms were grouped from k=2 to k=5, with k being the number of cluster cores, and it was found that when k=5, there is a better distribution of the terms, and two well-concentrated groups can be delimited, which may be genes involved in the same metabolic pathway or that are related to different pathways of the disease but have a similar gene expression pattern. Although the study was able to define associations between some genes of interest in the genetics of endometriosis, there are some future challenges, such as improving the commands for automated search and gene annotation, more precisely delimiting groups and associating terms with their respective metabolic pathways can further enhance our understanding of the disease and its mechanisms of action, allowing for early diagnosis.

Keywords: Endometriosis; Endometriosis Genetics; Literature Review; Cluster Analysis.

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

A endometriose é uma doença na qual o endométrio – tecido que reveste a camada interna do útero – é encontrado na cavidade abdominal, mais comumente no septo pélvico, peritoneal e retovaginal. Tal deslocamento anormal desencadeia um processo inflamatório crônico que pode causar dor pélvica crônica, dismenorreia, dispareunia e infertilidade, reduzindo drasticamente a qualidade de vida da mulher. A doença é estrogênio-dependente, aparecendo durante os anos reprodutivos, principalmente entre os 32 e 44 anos (ROWLANDS *et al.*, 2020), o que torna o diagnóstico e manejo da doença ainda mais difíceis (CHAPRON *et al.*, 2019).

Estima-se que a endometriose afete 1 em cada 10 mulheres em idade fértil, o que equivale a aproximadamente 176 milhões de mulheres em todo o mundo (WORLD ENDOMETRIOSIS RESEARCH FOUNDATION, 2015). No Brasil, segundo a Federação Brasileira das Associações de Ginecologia e Obstetrícia (FEBRASGO), cerca de 7 milhões de brasileiras – o que corresponde a cerca de 10% a 15% das mulheres brasileiras em idade reprodutiva – sofrem com os sintomas da endometriose (FEBRASGO, 2015). Ela afeta diversas áreas da vida da portadora, como atividades diárias, função sexual e relações pessoais, e é associada com depressão e fadiga, afetando a redução da produtividade no trabalho e causando prejuízos econômicos. Considerando esses aspectos, a endometriose deveria ser considerada uma questão de saúde pública (CHAPRON *et al.*, 2019).

Neste estudo, foram utilizadas literaturas científicas sobre a endometriose; por literatura científica, entende-se todo material intelectual produzido por estudiosos e pesquisadores, publicado, discutido e julgado por pares, entre outros pesquisadores da área, para que haja um consenso que produza confiabilidade. O formato das publicações são variados, desde artigos de periódicos (os quais constituem a maioria utilizada aqui), relatórios, livros, entre outros (MUELLER, 2000).

Para a análise desse material, utilizou-se a mineração de textos de forma multidisciplinar, envolvendo o Processamento de Linguagem Natural, Recuperação de Informação e Mineração de Dados (XAVIER *et al.* 2012), juntamente com conhecimentos em

Biologia e Biotecnologia na área Genética. Utilizando uma subárea da Inteligência Computacional, o KDT (*Knowledge Database Text*), foi possível extrair informações relevantes (no caso do estudo, os genes humanos) dos documentos não estruturados. Assim, cumpriu-se etapas de pré-processamento (de forma a melhorar a qualidade do texto), e a mineração em si, com a aplicação de técnicas como extração da informação, associação de documentos, classificação e clusterização, para então, no pós-processamento, analisar os resultados da mineração (XAVIER *et al.* 2012). A clusterização, como o próprio nome sugere, baseia-se na construção de grupos, que são conjuntos de dados que apresentem as mesmas propriedades, ou seja, façam parte de um grupo de elementos mais similares entre si do que a qualquer outro que não pertença ao grupo (LACHI *et al*. 2005).

Tendo em vista os fatos supracitados quanto à doença, seu impacto e a utilidade da mineração de textos, a construção de um *cluster* de genes encontrados na literatura científica sobre a endometriose se faz necessário. Assim, tal ferramenta tem como objetivo auxiliar na identificação de genes associados à doença, revelando possíveis conexões entre genes ainda não diretamente relacionados ao perfil genético da mulher portadora. Isso torna possível um diagnóstico precoce, em estágios iniciais da enfermidade, antes mesmo da idade fértil.

A tendência é que essas associações formadas nos agrupamentos sejam encontradas, na prática, em pacientes portadoras da doença, visto a confiabilidade apresentada pelas fontes originais usadas (literaturas científicas). Tais fontes, utilizam uma rigorosa metodologia científica para gerar o conhecimento descrito e têm resultados amplamente divulgados e discutidos entre diversos cientistas (MUELLER, 2000). Este tipo de *cluster* possibilita auxiliar e acelerar nos processos de tomadas de decisão de médicos e especialistas em casos de doenças específicas, como é o caso da endometriose, a qual foi foco deste estudo.

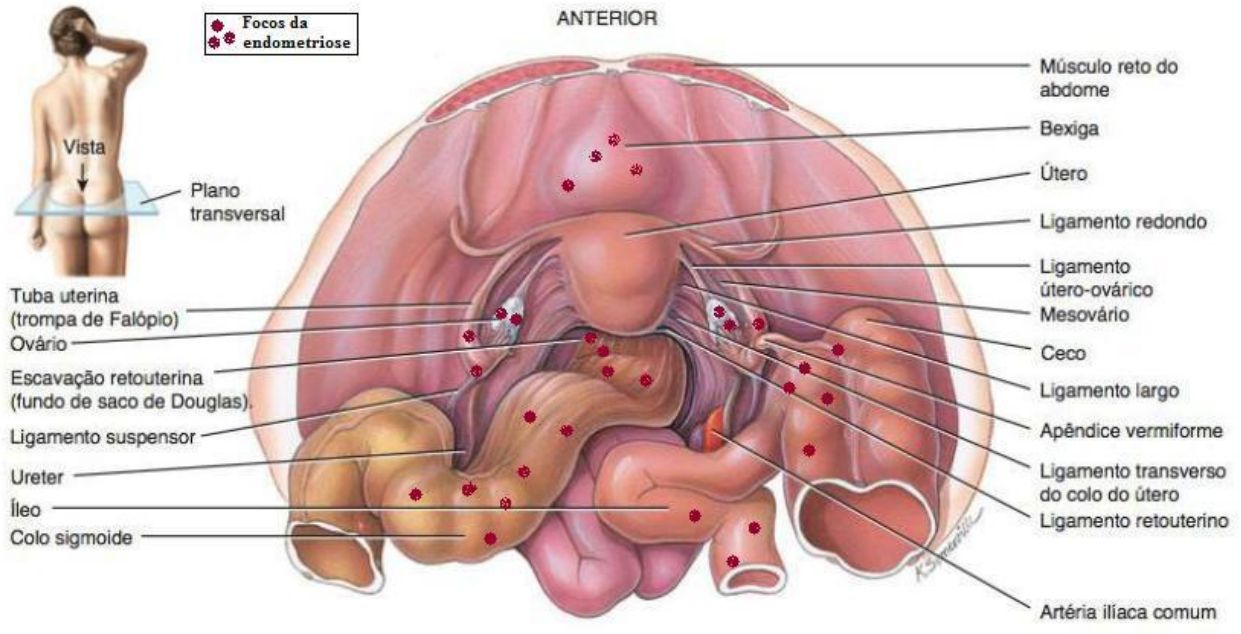
1.2 Revisão De Literatura

1.2.1 Endometriose

A endometriose é uma doença ginecológica crônica e estrogênio-dependente, na qual as portadoras apresentam tecido endometrial em locais extra uterinos (figura 1) (MARQUI, 2014). Processos inflamatórios que provocam dores intensas (dismenorreia e dispareunia) são

causados pela implantação do endométrio fora da cavidade uterina (QUEIROZ,2015). É uma doença heterogênea, que apresenta três fenótipos bem definidos; esses fenótipos são definidos a partir do local e forma das lesões, podendo ser superficiais e peritoneais (SUP), endometriomas ovarianos (OMA) ou de infiltração profunda (DIE). A primeira forma, é a mais leve, sendo lesões superficiais no peritônio, o tecido que circunda a cavidade pélvica; a segunda, cria massas de tecido ectópico endometrial (endometrioma) dentro do ovário, afetando diretamente a fertilidade, visto que atrapalha a liberação dos óvulos; já a terceira, forma mais grave da doença, são lesões que podem infiltrar tecidos musculares de órgãos que circundam o útero (como bexiga, intestino, uretra) ou o tecido peritoneal, como o ligamento útero-sacro. Nesta forma, ainda, podem acometer locais na cavidade que não estejam próximos ao útero, como o tecido umbilical, diafragmático ou pleural. Essas infiltrações são geralmente mais profundas do que 5mm abaixo da superfície dos tecidos, e raramente isolados, se apresentando normalmente como distribuições multifocais e podendo causar adesões que obstruem canais como o intestinal ou urinário. A Sociedade Americana de Medicina Reprodutiva (ASRM) classifica a Endometriose em 4 estágios, de acordo com a severidade das lesões, localização e avaliação cirúrgica do tamanho dessas, e ainda a ocorrência de adesões extensas. (CHAPRON *et al.*, 2019)

Figura 1: Possíveis focos da endometriose



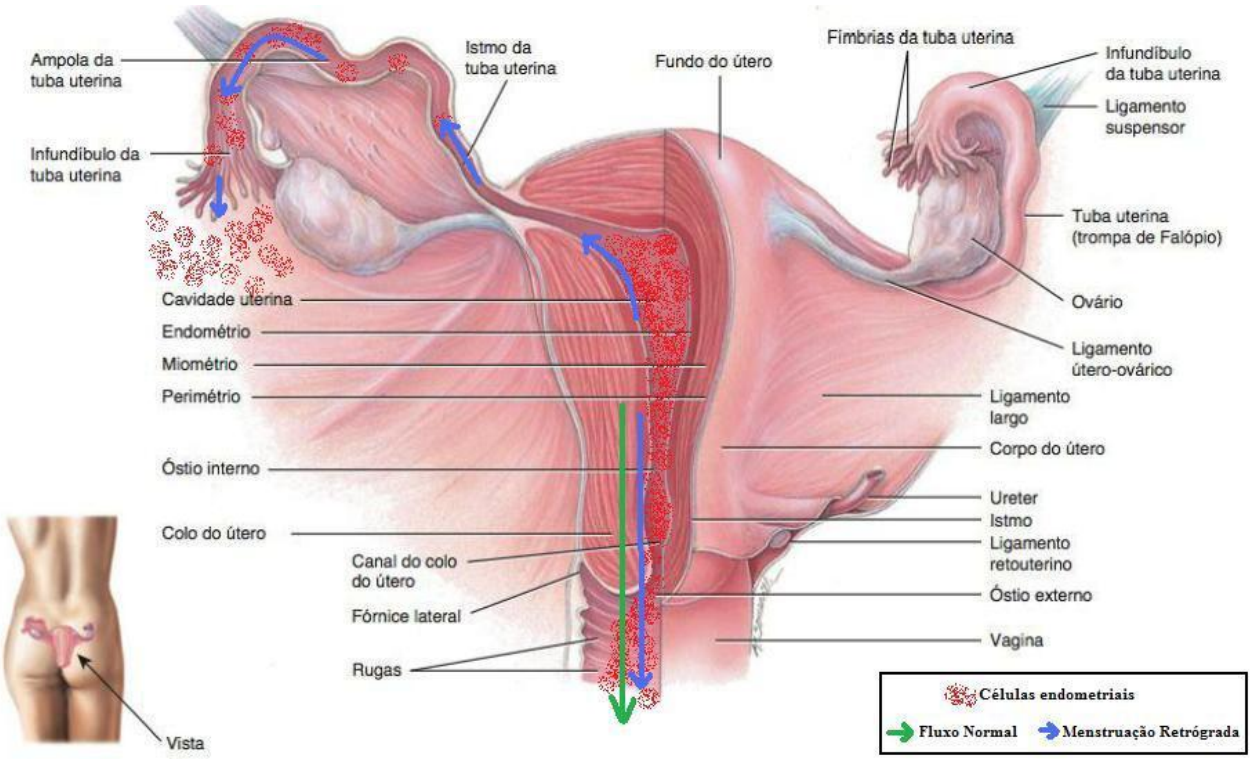
Fonte: Adaptado de (TORTORA & NIELSEN, 2013). Vista superior do corte transversal, apresentando os locais possíveis e mais comuns dos focos da endometriose.

Os principais sintomas da endometriose são a dor pélvica, a dificuldade em engravidar e a presença de massa pélvica em mulheres na fase reprodutiva, de forma isolada ou em associações. A exploração dos sintomas serve de base para o diagnóstico clínico, assim como o exame ginecológico e a identificação de fatores de risco, que favoreçam a teoria da endometriose como um distúrbio dependente da ação estrogênica e possivelmente secundária ao refluxo menstrual para a cavidade peritoneal (ÁVILA, CARNEIRO E FILOGONIO, 2015).

A patogênese da endometriose não está estabelecida definitivamente; no entanto, há algumas hipóteses predominantes (SCHENKEN, 2011):

- A teoria da implantação sugere que células endometriais sejam transportadas da parte interna do útero pelas trompas de falópio durante a menstruação (“menstruação retrógrada”), assim conseguindo acesso às estruturas pélvicas e implantando-se (figura 2).

Figura 2: Menstruação Retrógrada



Fonte: Adaptado de (TORTORA & NIELSEN, 2013). Vista posterior do útero e dos órgãos associados, demonstrando o fluxo normal da menstruação e o fluxo da menstruação retrógrada.

- A endometriose localizada em regiões não pélvicas pode ser explicada pela disseminação de tecido ou células endometriais através de vasos sanguíneos e linfáticos.
- A teoria da metaplasia celômica propõe que a cavidade celômica (peritoneal) contém células capazes de diferenciação ou indiferenciadas, que podem se diferenciar como tecido endometrial. Essa teoria é baseada em estudos embriológicos, os quais demonstraram que os órgãos pélvicos derivam-se de células que revestem a cavidade celômica.

A implantação dos fragmentos endometriais pode resultar de uma programação ontogênica endometrial com falhas, o que pode ser consequência de exposições neonatais ou ainda intrauterinas a certas características maternais. Entre estas características se destacam desde fatores ginecológicos (endometriose, miomas e fibroses) até hábitos maternos durante a

amamentação, nascimento prematuro, pré-eclâmpsia na gravidez ou ainda amamentação por produtos (leites não maternos). Outro fator que aumenta o risco de endometriose são as condições que aumentam o fluxo menstrual, primeira menarca precoce e períodos menstruais muito longos e intensos. (CHAPRON *et al*, 2019).

No entanto, nenhuma das teorias conseguiu explicar todos os locais de implantação e sintomas da doença. Isso levou pesquisadores a buscarem novas teorias que, de forma isolada ou em conjunto com as hipóteses já propostas, possam explicar melhor a etiologia da endometriose (PABALAN *et al*, 2017). A teoria da menstruação retrógrada, a mais aceita atualmente, considera um aspecto importante na patogênese da doença: a distribuição das lesões, que no geral são assimétricas, e isso pode ser explicado pela anatomia pélvico-abdominal e o fluxo peritoneal da menstruação em sentido horário, além do efeito da gravidade no fluxo menstrual. Como evidência que reforça esse fluxo, pode-se observar que as lesões são encontradas em sua maioria no lado esquerdo e no compartimento posterior e aquelas localizadas no tórax e abdômen são frequentemente mais encontradas do lado direito (CHAPRON *et al*, 2019).

A endometriose pleural pode ser explicada pela combinação da menstruação retrógrada, citada acima, o fluxo horário do fluido peritoneal e a passagem transdiafragmática de tecidos endometriais através dos poros do diafragma. Com as teorias citadas anteriormente, como a metaplasia, podem ser explicadas as lesões que aparecem em locais não comuns, fora da cavidade abdominal, como no cérebro ou pulmões. Entretanto, apesar das evidências apontarem para o uso dessas teorias combinadas, há um fator que ainda não se explica através delas: o mecanismo de aderência do tecido endometrial aos outros tecidos. O que se considera atualmente é a junção de diversos fatores, como inflamações, imunidade desregulada, hormônios, fatores ambientais e genéticos ou epigenéticos (CHAPRON *et al*, 2019).

É provável que fatores genéticos influenciem a suscetibilidade de um indivíduo à doença (SCHEKEN, 2011). Muitas pesquisas vêm sendo desenvolvidas na área genética e molecular, visando identificar os genes e, consequentemente, proteínas expressos pelo endométrio que possam estar ligados à sua receptividade. Apesar da teoria da menstruação retrógrada ser a mais aceita atualmente, como esse também é um evento comum em mulheres sem endometriose, faz-se necessária a presença de características moleculares para a implantação e progressão da implantação ectópica (QUEIROZ, 2015). Segundo

VERCELINNI *et al.* (2013), “A doença possui uma etiologia genética complexa que requer a interação de inúmeras variantes genéticas e fatores ambientais.”

1.2.2. Análise de *cluster*

Uma importante atividade humana é a classificação de objetos similares em grupos; a classificação sempre exerceu uma função significante na ciência e é utilizada desde a infância pelos seres humanos. Nesse sentido, classificar e agrupar são comportamentos inerentes e busca-se cada vez mais formas de realizá-los. Além disso, possibilitam compreender padrões discretos existentes entre objetos agrupados (MOBBS *et al.,* 2021). Uma análise de *cluster* é uma forma de classificar objetos em grupos, sendo uma técnica estatística multivariada, baseada em valores de inúmeras variáveis preditoras (MEHLER, SHAROFF E SANTINI, 2010).

Na utilização de técnicas de clusterização, há alguns passos que podem ser seguidos, a partir de definições. A primeira definição é sobre a representação dos dados de entrada, a forma que esses serão representados para serem agrupados; em seguida, deve-se definir uma medida que seja adequada para a aproximação entre os dados e qual será a técnica de clusterização a ser aplicada para a construção dos clusters. O passo seguinte é a definição de uma abstração dos dados, que deve permitir uma representação simples e compacta do conjunto de dados; do ponto de vista humano, ser simples é ser uma representação fácil de ser compreendida e intuitiva, enquanto do ponto de vista da máquina, deve possibilitar um processamento eficiente posteriormente; esse passo é opcional, dependendo da aplicação desenvolvida. O último passo é a avaliação do resultado da clusterização, a análise do agrupamento; geralmente, baseia-se em algum critério especificado subjetivamente, o que pode ser um problema, mas é geralmente utilizado para concluir o quão bom foi o agrupamento de dados obtido (LACHI, 2005).

“Análise de *cluster* é a arte de encontrar grupos em dados” e descobrir tais grupos é a função principal da análise de *cluster*; apesar disso, não existe uma definição geral para tal termo (KAUFMAN E ROUSSEEUW, 2009). É complexa a obtenção de um significado único para o termo, visto que depende de julgamentos de valores pela pessoa que executa o

projeto/pesquisa/análise (VALLI, 2012). Neste trabalho, denominar-se-á *cluster* os grupos de genes relacionados à endometriose, encontrados na literatura científica.

Como visto anteriormente, o principal objetivo da análise de *cluster* é o encontro de agrupamentos naturais de indivíduos (objetos, pontos, elementos, espécies, unidades, etc.). De um ponto de vista formal, esta análise visa a alocação de indivíduos em grupos de elementos mutuamente exclusivos, semelhantes, ou seja, os mais parecidos entre si. Ainda, é possível observar o inter-relacionamento entre as variáveis de estudo (VALLI, 2012).

Uma análise de *cluster* pode ser feita para realizar exploração de dados, redução de dados, geração de hipótese e predição baseada em grupos. Na exploração de dados, se inicia com uma análise de cluster que será usada para constatar uma ideia de agrupamento e aprimorá-la, podendo ser geradas novas hipóteses e descartados aqueles dados não relevantes para o estudo. Para que uma análise de cluster tenha êxito, ela deve ter permitido a estruturação e hierarquização de grupos de dados e o aprimoramento de agrupamentos nos mesmos (VALLI, 2012).

1.2.3. Processamento de Linguagem Natural (PLN)

A linguagem natural pode ser definida como uma linguagem usada cotidianamente para comunicação entre humanos e está em constante evolução. O Processamento de Linguagem Natural utiliza conceitos linguísticos, chamados de *Part-of-Speech*, e estruturas gramaticais para "descobrir quem faz o quê, a quem, quando, onde, como e porquê" (Robertson, 1946 *apud* BARBOSA *et al*., 2017), tratando o texto como uma sequência de caracteres, mas considerando a estrutura hierárquica da linguagem.

Segundo Indurkhya e Damerau (2010), o objetivo do PLN é extrair significados mais completos e representações de textos escritos em linguagem natural, ou seja, linguagem usada por humanos no dia-a-dia. É uma área da computação que trata o texto como uma simples sequência de caracteres. Várias atividades de PLN envolvem correspondência de padrões (BARBOSA *et al*., 2017), como a busca direcionada que será utilizada neste trabalho.

> *"O PLN também lida com situações mais complexas, como anáforas e ambiguidades. Isso se dá através de várias representações de conhecimento, como léxicos de palavras e seus significados; propriedades e regras gramaticais da linguagem; tesauros de*

sinônimos ou abreviações; e ontologias de entidades e ações [Indurkhya and Damerau, 2010].”

O *Natural Language Toolkit* é um pacote de ferramentas criado em 2001, na Universidade da Pensilvânia, como parte do curso de linguística computacional do Departamento de Computação e Ciência da Informação. O NLTK pode ser usado para criar programas de PLN em Python, visto que fornece implementações padrão para cada tarefa, que podem ser combinadas para resolver problemas complexos, interfaces padrão para executar tarefas, análise sintática e classificação de texto, e classes básicas para representar dados relevantes para processamento de linguagem natural (BIRD *et al.,* 2009). Nas Ciências Biológicas e áreas afim, o PLN vêm sendo amplamente empregado nas últimas décadas, nas mais diferentes frentes, desde de reconhecimento de padrão de distribuição de espécies e gráficos até o reconhecimento dos padrões da variante omicron da COVID-19 no organismo humano (CLEGG; SHEPHERD, 2007; MÓZSIK *et al.*, 2021; JIWANI; GUPTA; WHIG, 2022). Assim, explorá-la neste trabalho para identificação de agrupamentos é uma vertente das diversas funções desta linguagem.

1.2.5. Linguagens de fonte livre: Python

O uso de PLN depende de aplicações de linguagens computacionais de uso cotidiano e amplo dentro das diferentes áreas do conhecimento (ZHANG *et al.*, 2020). Assim, existem diversas linguagens de fonte livre na atualidade sendo empregados para estudos biológicos e biomédicos associadas ao PLN, como Perl, Ruby e Python (SERT *et al.,* 2022). A última vem sendo amplamente aplicada devido a facilidade associadas a tempo de latência, velocidade de processamento e recursos de armazenamento. Assim, empregar esse tipo de linguagem em estudos de agrupamento de literatura vem mostrando alta eficiência dentro das Ciências Biológicas.

Python é uma linguagem orientada a objetos dinamicamente tipados; todas as variáveis são objetos e não precisam ser declaradas para ser utilizadas (BARBOSA *et al*., 2017). É uma ferramenta simples, apesar de potente, para processamento de dados linguísticos (BIRD *et al.,* 2009). Nas últimas décadas, tornou-se uma das linguagens mais aplicadas em diferentes áreas, devido sua facilidade de aprendizado e uso, número de pacotes e funcionalidades e agilidade

de processamento e armazenamento (FOURMENT; GILLINGS, 2008; LV *et al.,* 2021). Devido a estas vantagens, ela é amplamente empregada em estudos na área da biologia e bioinformática, com várias versatilidades e pacotes utilizados no processo de *clustering,* tais como o *fastcluster* e o *Coclust* (MULLNER, 2013; ROLE; MORBIEU; NADIF, 2019). Dessa maneira, tornou-se a ferramenta fundamental aplicada neste estudo, o primeiro a realizar uma busca sistemática e produzir agrupamentos para a genética da endometriose.

1.3 Justificativa

A patogênese e o desenvolvimento da endometriose ainda não estão bem definidos, e acredita-se que fatores genéticos e ambientais estejam envolvidos nesses processos. Devido a isso, os estudos na área genética têm crescido, mas faz-se necessário um olhar interdisciplinar para lidar com a complexidade da doença (ROGERS; MONTGOMERY, 2015).

Neste trabalho, a interdisciplinaridade se dará pelo uso da linguagem Python na busca por *clusters*, relacionando linguística, computação e biologia. Como uma linguagem orientada a objetos, o Python permite que dados e métodos sejam armazenados e reutilizados em um momento posterior (BIRD *et al.,* 2009). Além disso, um fator muito importante a ser levado em conta para a relevância desse projeto é a análise ampla que será efetuada, visto que diversos trabalhos, com metodologias, populações e classificação variadas da doença serão utilizados como dados. Essas variações refletem a complexidade da doença e diferentes desenhos de estudo. Até o momento, esta é uma das primeiras tentativas a realizar busca sistemática por *cluster* para o termo genética da endometriose.

2 OBJETIVOS

2.1 Objetivo Geral

Esse projeto visa analisar publicações sobre a endometriose, buscando entender como os fatores genéticos são associados à doença na literatura científica dos anos 2009 a 2019 (acessada no Portal PubMed, em inglês), através da busca por *clusters* na literatura, em relação aos genes citados, para que possam ser vistas correlações genéticas ainda não percebidas.

2.2 Objetivos Específicos

- Aplicar ferramentas de PLN em Python para identificar contextos onde se agrupam determinadas palavras de interesse na genética da endometriose
- Avaliar, quantitativamente, os diferentes genes recuperados na literatura
- Listar e agrupar os genes associados à endometriose na literatura científica

3 METODOLOGIA

Para o desenvolvimento do projeto foram necessários um computador com acesso à internet, acesso ao Portal PubMed e o Python versão 3. O projeto se dividiu em etapas: i) seleção de artigos, ii) aplicação das ferramentas de processamento de linguagem natural, iii) análise do material, iii) levantamento de genes anotados e associados e iv) interpretação dos resultados obtidos.

A busca dos artigos foi realizada no Portal PubMed, através da busca avançada, com palavras chave diretamente relacionadas ao tema central – genética da endometriose. Então, foi feita a coleta dos artigos listados. Em seguida, foi utilizado um código escrito em Python (Apêndice I), com ferramentas do pacote NLTK, para comparar a similaridade entre os artigos, construindo uma matriz termo-documento.

Para encontrar nos textos os possíveis genes humanos, utilizou-se uma expressão regular (RegEx), criada pela autora (Apêndice I), baseadas nas normas de nomenclatura para genes humanos. “Expressões regulares são sequências de caracteres que definem um padrão de busca.” (BARBOSA *et al.* 2017), e servem para buscar objetos de interesse em textos que apresentem algum tipo de padrão.

Embora ainda não existam normas que consigam resolver todas questões de nomeação de genes, uma regra básica é o uso de letras maiúsculas (SPLENDORE, 2005). Então, para a criação da expressão regular, isso foi levado em conta, e também a possibilidade de existirem números e sinais como o hífen nos nomes. Logo, foi criada a expressão que busca por uma letra maiúscula, que pode ser seguida por uma outra letra, um número ou um hífen, e assim por diante. A partir da consolidação da expressão regular que melhor abrangeu os possíveis genes, foi desenvolvido um programa em Python para a busca utilizando a RegEx, que

abrangeu não só os objetos de estudo (possíveis genes) como a frequência de aparição dos mesmos.

Utilizando esses resultados tabelados, pode-se definir um número de frequência mínimo que se considerou relevante para o estudo e, assim, foi possível realizar a anotação gênica. Para identificar e classificar as funções dos genes recuperados na busca, foi realizada a anotação dos genes utilizando o banco de dados "Genes" presente no NCBI ("*National Center for Biotechnology Information*"). Para tal, foi feita uma busca local de similaridade de termos, na qual foi realizado o download do banco de dados contendo às informação dos dados depositados e suas descrições (https://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Human_Data.gene_info.gz; download realizado na data de 19/01/2023). As comparações foram feitas por comandos de rotina em *bash* (Apêndice II) e analisados em planilhas, gerando o dicionário final de genes (Anexo I).

A partir do código python criado e discutido acima (Apêndice I), criou-se um novo programa (Apêndice III), o qual realiza a busca dos termos do dicionário nos artigos explorados, descrevendo a ausência (0) ou presença (1) do termo. Com isso, uma nova tabela de *input* para o processo de clusterização dos termos genéticos foi gerada. Esta entrada foi implementada em um método de clusterização de texto não supervisionada (Apêndice IV), o qual busca agrupar textos em grupos que mais fazem sentido (AGRAWAL; GUPTA, 2014), testando diferentes valores de k (de 2 a 5). Como resultado do método, foram geradas planilhas com os agrupamentos de genes obtidos, sendo possível verificar quais genes estão presentes em cada *cluster*.

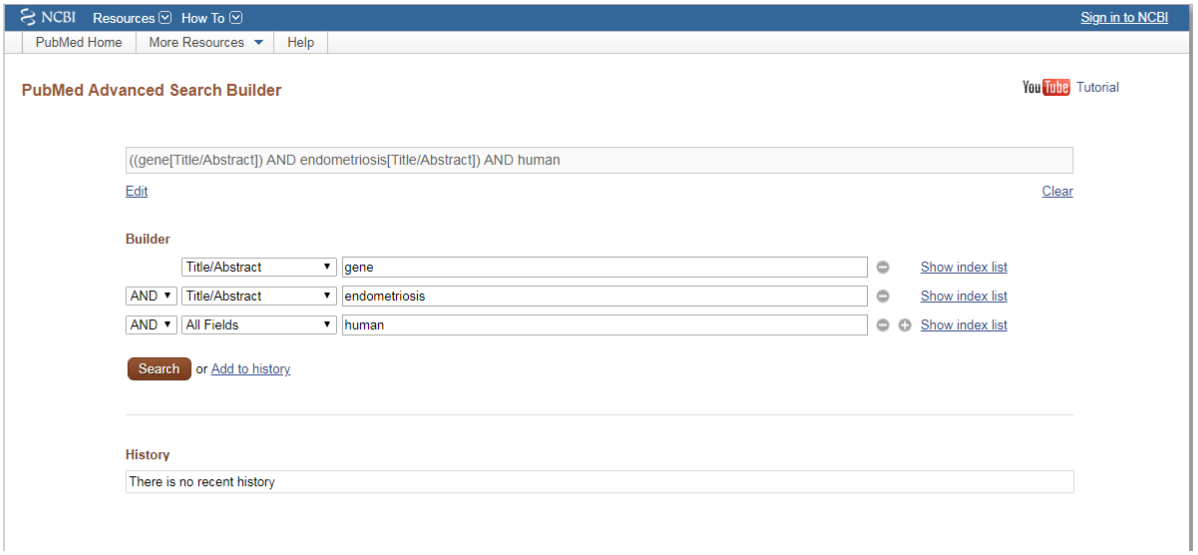
4 RESULTADOS

4.1. Levantamento de bibliografia

A definição dos termos utilizados para a pesquisa foi realizada com a colaboração da Prof(a). Dr(a). Bianca Bianco (Faculdade de Medicina do ABC), especialista na área. Os termos de busca utilizados foram "*gene*" e "*endometriosis*" na categoria "*Title/Abstract*" e "*human*" na categoria "*All fields*". Além disso, foi utilizado um filtro de data de publicação de

dez anos. A busca pelos artigos foi realizada no dia 07/01/2019, e foram encontrados 777 artigos que se enquadram nos parâmetros.

Figura 3: Parâmetros de busca no Portal PubMed



Fonte: Captura de tela da pesquisa na base de dados PubMed, realizada pela autora.

Com base nos resultados obtidos, foi criada uma lista com todos os artigos, autores e códigos de busca. A partir dessa lista, então, os artigos foram sendo acessados manualmente. Alguns desses artigos se encontravam disponíveis na própria plataforma PMC do PubMed, mas a maioria deles foi encontrada em outras plataformas. Também foi utilizada a plataforma ResearchGate, por meio da qual artigos não disponíveis gratuitamente foram solicitados aos autores. Os detalhes da coleta dos artigos podem ser encontrados na Tabela 1.

Tabela 1: Análise da coleta dos artigos

Artigos encontrados na busca	777
Artigos não disponíveis	13
Artigos encontrados apenas em outras línguas	34
Artigos solicitados e não recebidos	19
Artigos solicitados e recebidos	4
Total de artigos incluídos no estudo	711

Fonte: elaborado pelo autora

Para esse projeto, foram selecionados apenas os artigos encontrados em inglês, e obtidos gratuitamente. Dos artigos encontrados apenas em outras línguas, segue a tabela 2, que contém dados sobre a quantidade de artigos listados para cada língua; pode-se observar que a grande maioria dos artigos não encontrados na língua inglesa foram encontrados somente na língua chinesa.

Tabela 2: Artigos em outras línguas

Língua do artigo	Quantidade
Chines	24
Russo	2
Português	2
Alemão	2
Polonês	2
Finlandês	1
Húngaro	1
Total	**34**

Fonte: elaborado pela autora

4.2. Identificação dos termos genéticos associados à endometriose

Dos 777 artigos listados inicialmente, obteve-se para conversão um total de 711 artigos. Como o número de artigos obtidos ainda foi muito grande tornou-se inviável para o projeto a leitura de cada um deles. Sendo assim, todos eles foram convertidos para formato texto (.txt), de forma que pudessem ser lidos em Python. Essa conversão foi realizada por meio de um programa disponível gratuitamente, chamado PDFMate PDF Converter (disponível em: https://www.pdfmate.com/pdf-converter-free.html; acessado em 05 de junho de 2022).

Na conversão dos artigos, alguns deles não obtiveram uma conversão adequada, por dois fatores. O primeiro deles é o texto do PDF estar em formato de imagem, o que impede o programa de selecionar o texto; o segundo é a edição do texto no arquivo estar dividida em

duas colunas, o que dificulta o programa a selecionar as frases na ordem certa. No primeiro momento, julgou-se de baixa relevância realizar essa seleção, já que a influência desses erros na comparação entre artigos e termos não ocasionam vieses à análise.

O código de programação escrito em Python (Apêndice I) fez a "leitura" dos artigos, construindo um vocabulário com todas as palavras presentes nos mesmos. Dos 711 arquivos de texto (formato ".txt"), obteve-se um vocabulário de 121439 palavras. Para encontrar os genes e/ou proteínas ligados à endometriose, foi necessária a definição de um dicionário de palavras de importância, com a finalidade de reduzir o vocabulário amplo.

Dado a grande quantidade de palavras nos artigos, optou-se pela utilização de uma expressão regular (RegEx), criada para buscar diretamente os possíveis genes citados nos artigos e a frequência dos mesmos. Com a aplicação da RegEx, foi obtido um retorno de 16063 termos, planilhados e com frequências entre 1 e 14015 aparições do termo nos artigos. Considerando que a intenção do estudo é a visualização de relações entre genes ainda não reconhecidas na literatura científica, definiu-se uma ocorrência acima de 10 vezes como relevante. Então, a lista foi reduzida para 3087 termos, que foram processados manualmente, através da busca em plataformas de dados genéticos, reduzindo-se para 1960 termos. Na última etapa de processamento, foi realizada uma busca local de similaridade de termos, através de comparações feitas por comandos de rotina em *bash*, a partir de informações obtidas através do download do banco de dados genéticos.

4.3. Anotação gênica dos termos levantados

A partir dos 1960 genes levantados dentro dos trabalhos avaliados, foram anotados, com base em dados gênicos e bioquímicos para humanos, 1302 genes (~66,4% dos termos), os quais fazem parte do dicionário criado neste trabalho (Apêndice III). Entre eles, um total de 8 genes (~0,6%) foram recuperados como "não caracterizados" ou "proteínas hipotéticas" e necessitam de mais trabalhos bioquímicos para sua definição. Os demais supostos genes não anotados apresentam homólogos conhecidos em outras espécies, como ratos, coelhos, fungos, bactérias, entre outros. Por não tratarem-se de organismos de interesse para este trabalho, tais produtos gênicos e termos não foram explorados aqui.

4.4. Clusterização dos termos

A clusterização teve como primeira etapa a construção de uma tabela de presença (1) ou ausência (0), que possuía os genes do dicionário nas colunas e os artigos nas linhas. O gene “ESC” (anotado como *tescalcin*) foi o que apresentou maior frequência entre os genes do dicionário, seguido dos genes “ER” e “VEGF” (Tabela 03).

Tabela 03: Dicionário de genes com frequências de aparecimentos nos textos.

Termo	Freq.	Anotação	Tipo de termo
ESC	2986	tescalcin	protein-coding
ER	2760	serpin family A member 3	protein-coding
VEGF	2061	vascular endothelial growth factor D	protein-coding
OR	1964	adenosine A1 receptor	protein-coding
ARID1	1679	AT-rich interaction domain 1A	protein-coding
CC	1580	ATP binding cassette subfamily C member 6	protein-coding
CA	1485	ATP binding cassette subfamily A member 1	protein-coding
IV	1203	HIVEP zinc finger 1	protein-coding
PR	1202	ADP-ribosylarginine hydrolase	protein-coding
EM	1175	transmembrane protein 258	protein-coding
HOXA1	1129	homeobox A1	protein-coding
CYP19	921	cytochrome P450 family 19 subfamily A member 1	protein-coding

Fonte: Trecho do Output do Apêndice III, elaborado pela autora.

Com base na análise dos dados de clusterização na tabela de presença/ausência, foi identificado que o gene "IV" (anotado como *HIVEP zinc finger 1*) foi o gene mais citado em todos os artigos, com um total de 171 ocorrências (conforme Tabela 04). Além disso, o artigo de número 537 intitulado "*Endometrial biology during trophoblast invasion*" foi o que apresentou o maior número de genes diferentes citados, com um total de 128 genes. Importante ressaltar que 19 dos 711 artigos não mencionaram nenhum dos genes do dicionário final. Ainda, dos 1302 genes totais, 317 foram encontrados ocorrendo 1 vez nos artigos, e todos os genes tiveram ocorrência em algum texto.

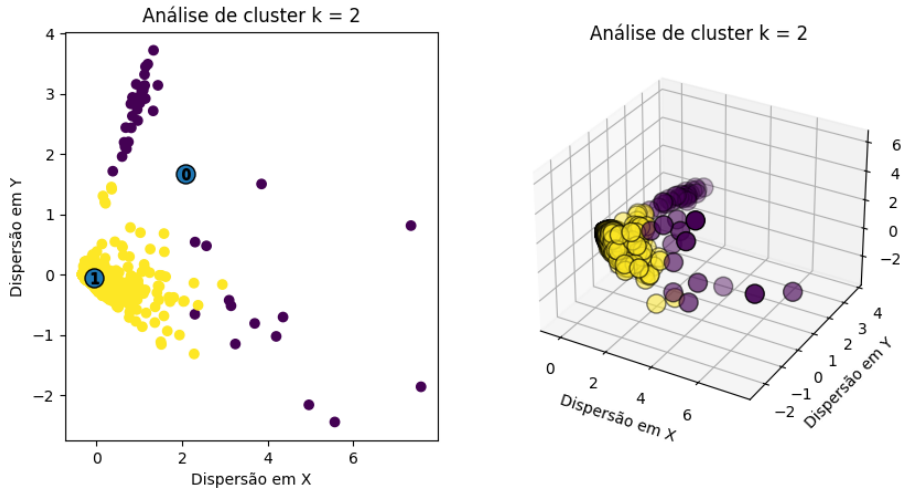
Tabela 04: Genes *versus* frequência de artigos

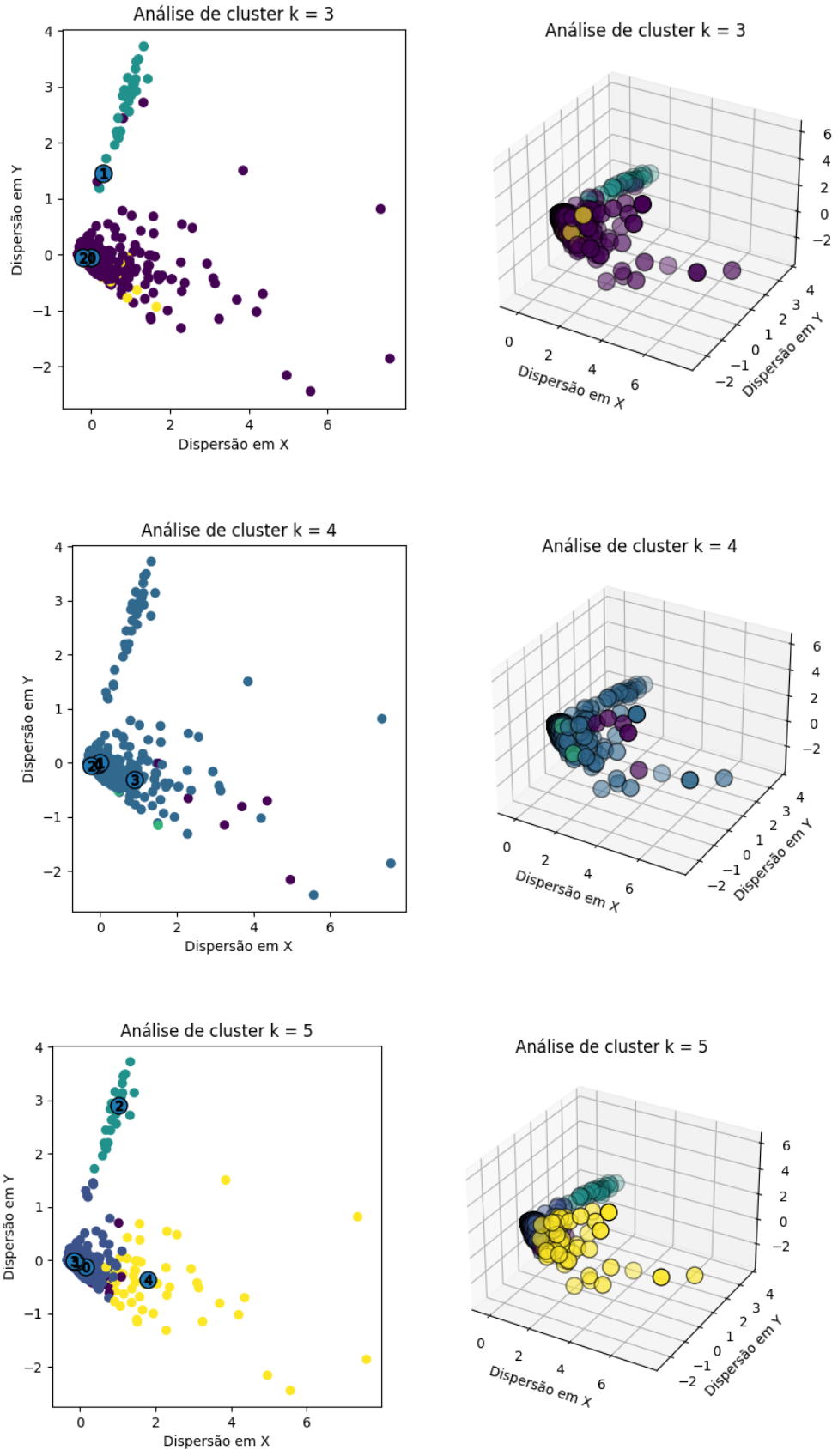
Gene	Qntd. de artigos
IV	171
AND	169
OR	122
ABI	114
VEGF	114
GAPDH	111
CC	99
ER	85
PR	84
MAPK	77

Fonte: elaborado pela autora, obtida a partir do apêndice II.

Utilizando a tabela de presença/ausência de termos genéticos por artigo como dado de entrada, foi possível realizar uma análise de clusterização não-supervisionada. Ela teve como base o número de vezes que estes termos aparecem (frequência) nos artigos, estando os resultados apresentados abaixo.

Figura 04: Clusters dos genes associados à endometriose encontrados na literatura científica





Fonte: Output do Apêndice IV, elaborado pela autora.

A maioria dos *clusters* (figura 04) apresenta grande número de termos genéticos agrupados, o que não oferece uma resposta específica para um grupo de genes de interesse. No primeiro gráfico, é possível perceber um agrupamento que se apresenta na cor roxa, de genes com menor correlação, que começa a ser subdividido a medida que o k (número de agrupamentos) aumenta. O cluster de número 1, na cor amarela no primeiro gráfico, se mantém, independente dos agrupamentos, desde k=2 até k=5; entretanto, com k=3 e k=4, ele incorpora genes que eram mais distantes em k=2, enquanto com k=5, ele se torna mais concentrado. Isso significa que ele agrupou um número menor de genes, os quais podem ter um grande interesse para estudos genômicos, bioquímicos e moleculares da endometriose.

5 DISCUSSÃO

A endometriose é uma doença inflamatória crônica, estrogênio-dependente, que afeta mulheres nos anos reprodutivos (ROWLANDS *et al.*, 2020) e ainda possui diagnóstico e manejo muito difíceis (CHAPRON *et al.*, 2019). Afeta cerca de 176 milhões de mulheres em todo o mundo (WORLD ENDOMETRIOSIS RESEARCH FOUNDATION, 2015) e cerca de 7 milhões de brasileiras, correspondendo entre 10% e 15% das mulheres brasileiras em idade reprodutiva (FEBRASGO, 2015). Devido aos sintomas, como dor pélvica crônica, dispareunia e infertilidade, impacta negativamente diversas áreas da vida da portadora, desde as atividades diárias, função sexual e relações pessoais, e deve ser considerada uma questão de saúde pública (CHAPRON *et al.*, 2019).

Entretanto, é possível notar que a genética da endometriose é pouco estudada e tem pouca visibilidade. Neste trabalho, foram coletados artigos entre os períodos de 2009 a 2019, somando-se 10 anos, em um dos maiores portais de publicações médicas mundiais, e obtidos apenas 777 artigos, o que resulta em um média de aproximadamente 78 artigos/ano. É interessante observar que a quantidade de artigos encontrados em língua portuguesa foi baixa, totalizando apenas 2 artigos, sendo que o Brasil é um país com grande incidência da doença, como mencionado previamente. A língua que mais foram encontrados artigos, depois do inglês, foi o Chinês, revelando uma crescente preocupação com a doença também nesse país.

Para a execução da análise proposta aqui, foi necessário estabelecer um fluxo de atividades em etapas. A primeira etapa foi a definição dos termos de busca, e a coleta dos

artigos; a partir daí, se inicia a mineração dos dados. Como o material obtido foi em PDF, um formato que não é processado pelas linguagem de programação, foi necessário realizar uma conversão para o formato texto; então, enfim foi possível iniciar a análise proposta, começando pela obtenção do vocabulário dos artigos. Nessa etapa, verificou-se que devido a amplitude desse, o ideal seria a criação de um dicionário, que pudesse direcionar as buscas. Então, frente ao objetivo do trabalho, de encontrar clusters de genes ligados à endometriose, foi utilizado o recurso de busca direta por Expressão Regular, que retornou os dados dos possíveis genes. A clusterização dos genes, etapa final do trabalho, envolveu a busca de frequência dos termos nos textos, definido para um valor mínimo de 10 vezes, a eliminação de sequências muito grandes de aminoácidos e busca pelos termos individualmente no GenBank (banco de dados genéticos). Na lista final, que compõe o dicionário de genes, há 1302 genes humanos que foram citados nos artigos relacionados à genética da endometriose. A partir disso, os genes foram anotados por meio de comandos de rotina em *bash*, comparando a lista obtida na etapa anterior e dados do banco de dados "Genes" do NCBI, realizando uma busca local de similaridade de termos, e posteriormente processados manualmente para atualização ou eliminação daqueles que não eram genes codificadores de proteína; a parte manual consistiu em buscar cada termo encontrado no *Genbank* com a condição de ser um gene humano. Com estes dados processados, foi possível utilizar um algoritmo de clusterização destes termos em *clusters* de genes relacionados à endometriose (Fig. 04).

A análise de cluster na biologia é uma técnica útil para agrupar conjuntos de dados biológicos em clusters ou grupos com base em suas semelhanças. Isso permite uma melhor compreensão da estrutura e função dos sistemas biológicos e pode ajudar a identificar padrões e relações significativas entre os dados. A análise de cluster pode ser usada em várias áreas da biologia, como genômica, proteômica, transcriptômica e metagenômica, para auxiliar na classificação de espécies, no estudo de vias metabólicas, na identificação de genes ou proteínas relacionados a doenças e na análise de comunidades microbianas (VALLI, 2012).

Dentro das áreas da biologia e saúde, para estudos populacionais relacionados à temáticas como doenças e genética. Por exemplo, CARDOSO (2015), em seu trabalho, utiliza a clusterização para delinear agrupamentos geográficos de mucopolissacaridose tipo I no Brasil, através de um levantamento de dados e recenseamento de isolados populacionais com

alta prevalência de doenças genéticas; como auxílio à gestão do Sistema Único de Saúde (SUS), TANAKA *et al*. (2015) utilizaram dados deste sistemas para agrupar os municípios em clusters com base nos indicadores selecionados, permitindo a identificação de pontos fortes e fracos em relação à gestão dos serviços de saúde em cada grupo.

Para o presente trabalho, foi verificado, a partir do dicionário criado, quais genes estavam presentes em cada artigo e construiu-se uma matriz de presença/ausência dos genes por artigo e, posteriormente, sua clusterização. Pode-se perceber uma falha no processo de anotação genômica dentro do banco de dados utilizado; o gene que foi, teoricamente, o mais citado entre os artigos, tem sigla "ESC" e foi anotado como *tescalcin*; entretanto, ao buscar pelo nome no próprio GenBank, pode-se notar que a sigla correta do gene anotado é "TESC" e nenhuma aliase retornou o nome do termo buscado; o mesmo foi verificado para o gene que apareceu em mais artigos ("IV", anotado como *HIVEP zinc finger 1*, e de sigla correta "HIVEP1"). Portanto, constatou-se que no banco de dados, quando realizada a busca automatizada, recupera-se a anotação gênica de termos parciais, o que ocasiona possível viés na busca. Este viés parece ser recorrente e deverá ser melhor investigado e corrigido nas futuras versões dos algoritmos criados ou em novos programas aqui desenvolvidos.

Considerando o viés encontrado, o gene mais frequentemente recuperado foi o "ER", uma aliase encontrada para o gene de nome oficial *estrogen receptor 1* (ESR1), que codifica um receptor de estrogênio e um fator de transcrição ligante-ativado; a proteína codificada por este gene regula a transcrição de muitos genes induzidos por estrogênio, envolvidos em diversas funções como crescimento, metabolismo, desenvolvimento sexual, funções reprodutivas, entre outras (GENBANK, 2023). O próximo termo mais encontrado, "VEGF", representa uma aliase do gene VEGFA, de nome oficial "*vascular endothelial growth factor A*"; VEGF é ainda uma família de isoformas moleculares, codificadas por um único gene (FÁTIMA *et al*, 2010). O VEGFA é um gene que codifica uma proteína de ligação à heparina e desempenha papel na proliferação e migração de células vasculares endoteliais, sendo essencial para a angiogênese fisiológica e patológica; sua interrupção em camundongos resultou na formação anormal de vasos sanguíneos embrionários, levando a letalidade durante o desenvolvimento fetal. Além disso, a expressão do VEGFA está positivamente regulada em diversos tipos de tumores e está relacionada com o estágio e a progressão do tumor (NCBI Gene, 2023). A expressão do VEGF é estimulada pela ação de alguns hormônios, como

estradiol, hormônio luteinizante (LH), progesterona, entre outros, e também é regulada por citocinas como o fator de necrose tumoral, crescimento tumoral, crescimento epidermal, interleucinas e fator de crescimento fibroblástico básico. Possui importância crítica com a angiogênese (formação de novos vasos sanguíneos através de outro pré-existentes, num processo de migração e proliferação de células), promovendo o crescimento das células endoteliais derivadas de artérias e veias (FÁTIMA *et al*, 2010).

O VEGF é um gene altamente polimórfico, sendo um marcador essencial em diversas doenças e tem um papel crítico no desenvolvimento da angiogênese, estando envolvido necessariamente na criação de novos vasos sanguíneos, através de vários mecanismos (RASHID *et al*, 2019). Estudos têm associado os polimorfismos do VEGF à ocorrência da endometriose, uma condição que apresenta características similares a neoplasias, como a necessidade de formação de novos vasos sanguíneos para a implantação e desenvolvimento do tecido endometrial, bem como a presença de genes relacionados à angiogênese (BRUNO *et al*, 2018). Ele atua na progressão da doença, visto que influencia as células endoteliais vasculares, desencadeia a proliferação, sobrevivência e migração de células endoteliais, criação de novos vasos sanguíneos e crescente penetrabilidade vascular. Em portadoras da doença, diversos estudos indicaram o aumento dos níveis de VEGF mRNA e suas proteínas, o que corrobora um fator fundamental para a angiogênese relacionada à endometriose (RASHID et al, 2019).

O artigo que mais citou genes foi o "*Endometrial biology during trophoblast invasion*", o qual identifica genes envolvidos na conexão embrio-endometrial, para compreender patologias como aborto espontâneo e endometriose. No desenvolvimento embrionário, o endométrio sofre uma séria de mudanças fisiológicas, bioquímicas e também morfológicas, denominadas coletivamente de decidualização endometrial. Em pacientes com endometriose, as taxas de crescimento folicular são baixas, assim como a capacidade funcional dos folículos pré-ovulatórios e taxas de fertilização, havendo ainda anormalidades no desenvolvimento embriogênico pré-implantação e alterações na função lútea. Após a concepção, há ainda dificuldades na implantação por possíveis defeitos na comunicação cruzada embrião-endometrial, através da qual o embrião se comunica com o tecido endometrial do útero para preparar o ambiente ideal para a implantação, através de sinais

moleculares que interagem e influenciam a expressão de genes e as funções celulares em ambos os tecidos (SALKER *et al*, 2012).

Diversos genes foram analisados no artigo citado, e percebeu-se que a atuação do gene HOXA10 é essencial para a receptividade endometrial. Sua expressão é induzida ainda mais pelos estímulos embriogênicos durante a decidualização, sendo que diversos estudos demonstraram que esse gene regula ainda a expressão de marcadores essenciais para o processo, como IGFBP1 e prolactina. Além disso, inúmeros genes, imunomoduladores, proteases, moléculas sinalizadoras e relacionadas à citoesqueleto e adesão, assim como moléculas transportadoras se expressam de forma atenuada ou até oposta em portadoras de endometriose. Isso impede que o repertório molecular endometrial requerido para a correta implantação seja recrutado, corroborando que a baixa taxa de implantação em mulheres com endometriose seja não só devido a questões fisiológicas, mas toda uma alteração na comunicação embrião-endometrial. Assim, o artigo cumpre seu objetivo de identificar os genes envolvidos nesse processo (SALKER *et al*, 2012).

Para a clusterização, o algoritmo escolhido foi o *k-means*, amplamente utilizado para agrupar dados em clusters com base em suas semelhanças. O objetivo do algoritmo é encontrar os centros de cluster, chamados de centróides, e alocar os pontos restantes ao redor desses pontos centrais (VALLI, 2012). Dessa forma, a distância entre os genes e a proximidade desses pode ser investigada. Foram selecionadas análises desde k=2 até k=5, onde verificou-se, pois os testes realizados com valores de k maiores que 5 resultaram em poucas alterações comparadas ao k=5. Ao analisar os resultados obtidos na clusterização, pode-se identificar alguns grupos, nos quais determinados termos aparecem associados (por exemplo, *clusters* 0 e 2 do agrupamento k=5; Fig. 04). Tais associações podem estar envolvidas com genes de mesma rota metabólica ou, então, que estejam relacionados a diferentes vias da doença, mas que apresentam padrão de expressão gênica similar.

6 CONCLUSÃO

Neste trabalho, foi realizado pela primeira vez o agrupamento de termos genéticos baseado na literatura científica disponível sobre a endometriose. Esta doença, como discutido aqui, atinge um grande número de mulheres e preocupa as autoridades da saúde. Assim,

proporcionar o aumento de conhecimento nesta área por métodos da biologia computacional e aprendizado de máquina são fundamentais. Os programas aqui construídos em Python, através da aplicação de ferramentas em PLN, foram capazes de realizar associações entre alguns genes de interesse na genética da endometriose, criando um script funcional para tal, avaliando quantitativamente aqueles presentes e citados em cada artigo, listando e agrupando-os. Com isso, os métodos de clusterização mostraram-se úteis no processo de relacionar termos distintos da genética da endometriose.

Embora os objetivos do trabalho tenham sido alcançados ao mostrar associações entre os termos gênicos e os conhecimentos já disponíveis na literatura, ainda são necessários melhores delineamentos destas buscas e agrupamentos. Por exemplo, aprimorar os comandos para busca automatizada e anotação gênica, delimitar de forma mais precisa os grupos e associar os termos às vias metabólicas às quais pertencem pode auxiliar a compreender ainda melhor a doença e suas vias de ação. Logo, futuros trabalhos que realizem este tipo de consociação tornarão mais simples e fácil a compreensão genética, o diagnóstico e a pesquisa dentro da área da endometriose.

7 REFERÊNCIAS

1. AGRAWAL, A., GUPTA, U. (2014). **Extraction based approach for text summarization using k-means clustering**. *International Journal of Scientific and Research Publications*, *4*(11), 1-4.
2. ÁVILA, I; CARNEIRO, M. M., FILOGONIO, I. D. S. e colaboradores. **Manual de Endometriose.** FEBRASGO (2015).
3. BARBOSA *et al.* **Introdução ao Processamento de Linguagem Natural usando Python**. III Escola Regional de Informática do Piauí. Livro Anais - Artigos e Minicursos, v. 1, n. 1, p. 336-360, jun, 2017.
4. BARBOSA, J.L.N. VIEIRA, J.P.A. SANTOS, R.L.S. MAGALHÃES JR, G.V.M, MUNIZ, M.S. MOURA, R.S. **Introdução ao Processamento de Linguagem Natural usando Python.** III Escola Regional de Informática do Piauí. Livro Anais - Artigos e Minicursos, v. 1, n. 1, p. 336-360, jun, 2017. www.eripi.com.br/2017 - ISBN: 978-85-7669-395-6

5. BIRD, S., KLEIN&E., LOPER, E. **Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit**, 2009.
6. BRUNO, L. T. PRATA-LIMA, M. F. RUIZ-CINTRA, M. T. MARQUI, A. B. T. **Investigation of VEGF gene polymorphism rs35569394 in endometriosis.** LABORATORY MEDICINE, Original article • J. Bras. Patol. Med. Lab. 54 (6) • Nov-Dec 2018 • https://doi.org/10.5935/1676-2444.20180057
7. CARDOSO, G.C. **Identificação de "CLUSTERS" de doenças genéticas em populações isoladas do Brasil**. LUME, Repositório Digital UFRGS. 2015. Disponível em <Identificação de "CLUSTERS" de doenças genéticas em populações isoladas do Brasil (ufrgs.br)>.
8. CHAPRON, C., MARCELLIN, L., BORGHESE, B., & SANTULLI, P.. **Rethinking mechanisms, diagnosis and management of endometriosis.** Nature Reviews Endocrinology. 2019. doi:10.1038/s41574-019-0245-z.
9. CLEGG, A. B., SHEPHERD, A. J. **Benchmarking natural-language parsers for biological applications using dependency graphs.** *BMC bioinformatics*, vol. *8, n.* 1, 1-17, 2007.
10. FÁTIMA, L. A. PAPA, P. C. **Fator De Crescimento Do Endotélio Vascular (VEGF): Regulação Transcricional E Pós-Transcricional.** Setor de Anatomia, Departamento de Cirurgia, Faculdade de Medicina Veterinária e Zootecnia, USP Recebido 19out09 / Aceito 14jan10 / Publicação inicial 15abr10
11. FOURMENT, M., GILLINGS, M. R. **A comparison of common programming languages used in bioinformatics.** *BMC bioinformatics*, vol. *9*, 1-9, 2008.
12. FUNG, J. N., ROGERS, P. A. W., MONTGOMERY, G. W. **Identifying the Biological Basis of GWAS Hits for Endometriosis**. BIOLOGY OF REPRODUCTION (2015) 92(4):87, 1–12
13. GENBANK. **Gene ID 2099**. 2023. <https://www.ncbi.nlm.nih.gov/gene/2099>
14. INDURKHYA, N. & DAMERAU, F. J. **Handbook of Natural Language Processing**. Chapman & Hall/CRC, 2nd edition; 2010.
15. JIWANI, N., GUPTA, K., WHIG, P. **Analysis of the Potential Impact of Omicron Crises Using NLTK (Natural Language Toolkit).** In *Proceedings of Third Doctoral*

Symposium on Computational Intelligence: DoSCI 2022 (pp. 445-454). Singapore: Springer Nature Singapore.

16. KAUFMAN, L.& ROUSSEEUW, J. P. **Finding Groups in Data: An Introduction to Cluster Analysis.** John Wiley& Sons, Inc., Hoboken, New Jersey, 2005.
17. LACHI, R. L. ROCHA, H. V. **Aspectos básicos de clustering: conceitos e técnicas.** Technical Report - IC-05-003 - Relatório Técnico. Núcleo de Informática Aplicada à Educação (Nied) Instituto de Computação – Universidade Estadual de Campinas. Fev/2005.
18. LV, Z., CUI, F., ZOU, Q., ZHANG, L., XU, L. **Anticancer peptides prediction with deep representation learning features**. *Briefings in bioinformatics*, vol. *22, n.* 5, bbab008, 2021.
19. MARQUI, A. B. T. **Endometriose: do diagnóstico ao tratamento**. RevEnferm Atenção Saúde [Online]. jul/dez 2014; 3(2):97-105
20. MEHLER, A., SHAROFF, S., SANTINI, M. **Genres on the Web: Computational Models and Empirical Studies**. *Volume 42 de Text, Speech andLanguage Technology*, 2010.
21. MOBBS, D., WISE, T., SUTHANA, N., GUZMÁN, N., KRRIGESKORTE, N., LEIBO, J. Z. **Promises and challenges of human computational ethology.** Neuron, vol. 109, n. 14, 2224-2238, 2021.
22. MÓZSIK, L., POHL, C., MEYER, V., BOVENBERG, R. A., NYGåRD, Y., DRIESSEN, A. J. **Modular synthetic biology toolkit for filamentous fungi.** *ACS Synthetic Biology*, vol. *10, n.* 11, 2850-2861, 2021.
23. MUELLER, S. P. M. **A ciência, o sistema de comunicação científica e a literatura científica.** Fontes de informação para pesquisadores e profissionais. Bernadete Santos Campello, Beatriz Valadares Cendón, Jeannette Marguerite Kremer/Organizadoras. - Belo Horizonte: Ed. UFMG, 2000.
24. MULLNER, D. **Fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python.** *Journal of Statistical Software*, vol. *53*, 1-18, 2013.
25. NCBI Gene. (2023). **VEGFA vascular endothelial growth factor A [Homo sapiens (human)].** Acesso em 05 de Abril de 2023, disponível em <VEGFA vascular endothelial growth factor A [Homo sapiens (human)] - Gene - NCBI (nih.gov)>.

26. PABALAN N, JARJANAZI H, CHRISTOFOLINI DM, BIANCO B, BARBOSA CP. **Association of the protein tyrosine phosphatase non-receptor 22 polymorphism (PTPN22) with endometriosis: a meta-analysis**. Einstein (Sao Paulo). 2017;15(1):105-111
27. QUEIROZ, A. M. **Aspectos genéticos e moleculares da endometriose** [trabalho de conclusão de curso]. Brasília: Centro Universitário de Brasília – UniCEUB, Faculdade de Ciências da Educação e Saúde;2015.
28. RAHMIOGLU, N. NYHOLT, D. R. MORRIS, A. P. MISSMER, S. A. MONTGOMERY, G. W. ZONDERVAN, K. T. **Genetic variants underlying risk of endometriosis: insights from meta-analysis of eight genome-wide association and replication datasets.** Human Reproduction Update, Vol.20, No.5 pp. 702–716, 2014
29. RASHIDI, B.H., SARHANGI, N., AMINIMOGHADDAM, S. *et al.* **Association of vascular endothelial growth factor (VEGF) Gene polymorphisms and expression with the risk of endometriosis: a case–control study.** *Mol Biol Rep* 46, 3445–3450 (2019). https://doi.org/10.1007/s11033-019-04807-6
30. ROLE, F., MORBIEU, S., NADIF, M. **Coclust: a python package for co-clustering.** *Journal of Statistical Software*, vol. *88*, 1-29, 2019.
31. SALKER, M. S., NAUTIYAL, J., STEEL, J. H., WEBSTER, Z., ŠUĆUROVIĆ, S., NICOU, M., SINGH, Y., LUCAS, E. S., MURAKAMI, K., CHAN, Y.-W., JAMES, S., ABDALLAH, Y., KUMARENDRAN, B., HACETTEPE, M., BARRATT, C., BECKER, C. M., BROSENS, J. J., & DEY, S. K. (2012). **Endometrial biology during trophoblast invasion.** Frontiers in bioscience (Scholar edition), S4, 1151–1171. https://doi.org/10.2741/s330
32. SCHENKEN, R. S. **Pathogenesis, clinical features, and diagnosis of endometriosis**.2011
33. SERT, O. C., ÖZYER, S. T., BESTEPE, D., ÖZYER, T. **Temptracker: a service oriented temporal natural language processing based tool for document data characterization and social network analysis.** Int. Arab J. Inf. Technol., vol. 19, n.3, 342-352, 2022.

34. SPLENDORE, A. **Para que existem as regras de nomenclatura genética?** Educacional • Rev. Bras. Hematol. Hemoter. 27 (2) • Jun 2005 • https://doi.org/10.1590/S1516-84842005000200020
35. TANAKA, O. Y., JUNIOR, M. D., CRISTO, E. B., SPEDO, S. M., & PINTO, N. R. DA S. (2015). **Uso da análise de clusters como ferramenta de apoio à gestão no SUS.** Saúde e Sociedade, 24(1), 23-32. https://doi.org/10.1590/S0104-12902015000100003
36. TORTORA, G. J. & NIELSEN, M. T. **Princípios de Anatomia Humana**. Rio de Janeiro: Guanabara Koogan, 2013.
37. VALLI, M. **Análise de Cluster**. Augusto Guzzo Revista Acadêmica, São Paulo, n. 4, p. 77-87, aug. 2012. ISSN 2316-3852. Disponível em: <http://fics.edu.br/index.php/augusto_guzzo/article/view/107>. Acesso em: 06 mar. 2023. doi: https://doi.org/10.22287/ag.v0i4.107.
38. VERCELLINI, P., VIGANÒ, P., SOMIGLIANA, E., FEDELE, L. **Endometriosis: pathogenesis and treatment.** *Nat. Rev. Endocrinol.* 10, 261–275; 2014.
39. World Endometriosis Society e World Endometriosis Research Foundation, **Facts about endometriosis**. Publicado em Setembro de 2015.

APÊNDICE I

Código de busca por termos gênicos

```
import os
import re

## identificamos a lista de artigos
files = []

# r=root, d=directories, f = files
for r, d, f in os.walk("./TXT/"):
    for file in f:
        if '.txt' in file:
            files.append(os.path.join(r, file))


regex = r"[A-Z][0-9]?[A-Z]-?[A-Z]?[0-9]?[A-Z]?-?[A-Z]?[0-9]?"
frequencies = dict([])

# lendo os artigos
for f in files:
    print("lendo o artigo:", f)

    # leitura do documento
    documento = open(f, 'r')
    conteudo  = documento.read()

    # identificacao de genes
    genes= re.findall(regex, conteudo)

    # quantidade de vezes no documento
    for g in genes:
        #g = g.lower()
        if g not in frequencies:
            frequencies[g] = 0
        frequencies[g] += 1

for g in frequencies:
  print ("{}\t{}".format(g, frequencies[g]))
```

APÊNDICE II

Loop usando o comando *for* para realização da busca "circular" dos termos gênicos identificados. Em cada etapa, fez-se a busca pelo termo específico (*"$Line"*) dentro do banco de dados (*All_Data.txt*), retornando um arquivo para cada termo. Na segunda etapa foi realizada a concatenação dos arquivos, com a inserção de um arquivo por linha, gerando a planilha final (*planilha_final*).

Etapa 1

```
#!/bin/bash
File="Gene_Dict.txt"
Lines=$(cat $File)
for Line in $Lines
do
     mkdir "$Line"
     cd "$Line"
     search="$Line"
     awk   -vcol="$3"   -vsearch="$search"   '$3   ~   search'
../All_Data.txt > a
     awk 'FNR <= 1' a > "$Line"
     awk '{print FILENAME (NF?",":"") $0}' "$Line" > file.txt
     cd ..
done
```

Etapa 2

```
cat */file.txt > planilha_final.txt
```

APÊNDICE III

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
import os
import re
file = open('dicionario.txt')
genes = file.readlines()
## identificamos a lista de artigos
files = []
#r=root, d=directories, f = files
for r, d, f in os.walk("./TXT/"):
        for file in f:
                if '.txt' in file:
                        files.append(os.path.join(r, file))
        # lendo os artigos
for f in files:
        lista = []
        print("lendo o artigo:", f)

        # leitura do documento
        documento = open(f, 'r')
        conteudo  = documento.read()

        for i in range(len(genes)):
            B = genes[i]
            C = B.strip()
            A = (" "+C+" ")
            if A in conteudo:
                print("Encontrado gene ",A,"em ",f)

        for i in range(len(genes)):
            B = genes[i]
            C = B.strip()
            A = (" "+C+" ")
            if A in conteudo:
                print(A," 1")
            else:
                print(A," 0")
```

APÊNDICE IV

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
import os
import pandas as pd
import numpy as np
from sklearn.cluster import MiniBatchKMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import nltk
from sklearn.metrics import silhouette_samples,
silhouette_score, v_measure_score
from sklearn.datasets import load_files
import re
from unidecode import unidecode
from mpl_toolkits import mplot3d
import sys
import numpy
numpy.set_printoptions(threshold=sys.maxsize)

file = pd.read_csv('genes.csv')

for cluster in range(2,11):
    cls = MiniBatchKMeans(n_clusters=cluster, random_state=0)
    cls.fit(file)

fig = plt.figure(figsize=plt.figaspect(0.5))

ax = fig.add_subplot(1, 2, 1)
```

```
#Visualização gráfica 2D
pca = PCA(n_components=2, random_state= 0)
reduced_features = pca.fit_transform(file)
reduced_cluster_centers = pca.transform(cls.cluster_centers_)

 #Plota gráfico 2D
ax.scatter(reduced_features[:,0], reduced_features[:,1],
c=cls.predict(file))
ax.scatter(reduced_cluster_centers[:, 0],
reduced_cluster_centers[:,1], marker='o', s=150,
edgecolor='k')

#Plota números nos clusters
for i, c in enumerate(reduced_cluster_centers):
     ax.scatter(c[0], c[1], marker='$%d$' % i, alpha=1,s=50,
edgecolor='k')

plt.title("Análise de cluster k = %d" % cluster)
plt.xlabel('Dispersão em X')
plt.ylabel('Dispersão em Y')

ax = fig.add_subplot(1, 2, 2,projection="3d")

plt.title("Análise de cluster k = %d" % cluster)
plt.xlabel('Dispersão em X')
plt.ylabel('Dispersão em Y')

    #converte dados para 3D
pca = PCA(n_components=3, random_state=0)
reduced_features = pca.fit_transform(file)
```

```
ax.scatter3D(reduced_features[:,0], reduced_features[:,1],
reduced_features[:,2], marker='o', s=150, edgecolor='k',
c=cls.predict(file))
reduced_cluster_centers = pca.transform(cls.cluster_centers_)

    #Salva arquivo de imagem
plt.savefig("/mnt/c/Users/Pichau/Desktop/grafico_cluster_k=%d"
% cluster)

print(cls.predict(file))
```

ANEXO I

Planilha contendo os genes associados com o termo "genética de endometriose", frequência com que estes termos apareceram e sua anotação.

Termo	Freq.	Anotação
ESC	2986	tescalcin
ER	2760	serpin family A member 3
VEGF	2061	vascular endothelial growth factor D
OR	1964	adenosine A1 receptor
ARID1	1679	AT-rich interaction domain 1A
CC	1580	ATP binding cassette subfamily C member 6
CA	1485	ATP binding cassette subfamily A member 1
IV	1203	HIVEP zinc finger 1
PR	1202	ADP-ribosylarginine hydrolase
EM	1175	transmembrane protein 258
HOXA1	1129	homeobox A1
CYP19	921	cytochrome P450 family 19 subfamily A member 1
LC	903	activated leukocyte cell adhesion molecule
SE	877	serpin family A member 3
JA	842	jagged canonical Notch ligand 1
NM	836	dynamin 1
AA	819	arylacetamide deacetylase
MA	818	basic helix-loop-helix ARNT like 1
SD	815	arylsulfatase D

Termo	Freq.	Anotação
NS	810	asparagine synthetase (glutamine-hydrolyzing)
SC	796	erythroblast membrane associated protein (Scianna blood group)
RH	743	ADP-ribosylarginine hydrolase
MAPK	722	mitogen-activated protein kinase 14
MD	707	adenosylmethionine decarboxylase 1
MS	702	membrane spanning 4-domains A1
GO	683	golgin A1
JM	682	JMJD7-PLA2G4B readthrough
GSTM1	671	glutathione S-transferase mu 1
AM	653	angio associated migratory cell protein
CT	645	actin alpha 1, skeletal muscle
CG	644	glycoprotein hormones, alpha polypeptide
PF	624	ATP synthase peripheral stalk subunit F6
JL	617	recombination signal binding protein for immunoglobulin kappa J region like
SA	602	arylsulfatase A
PGR	583	progesterone receptor
AG	568	aspartylglucosaminidase
RA	546	adenosine A1 receptor
RN	546	Rho family GTPase 3
HS	545	alpha 2-HS glycoprotein

Termo	Freq.	Anotação
SH	539	shroom family member 2
GAPDH	534	glyceraldehyde-3-phosphate dehydrogenase
JC	531	DnaJ heat shock protein family (Hsp40) member C4
ESR1	527	estrogen receptor 1
TNF	526	TNF receptor superfamily member 17
ECSC	508	endothelial cell surface expressed chemotaxis and apoptosis regulator
ES	507	carboxylesterase 1
MSC	503	musculin
PA	501	poly(ADP-ribose) polymerase 1
SP	501	cysteine rich secretory protein 1
PL	499	perilipin 2
RT	480	adenine phosphoribosyltransferase
TP53	480	tumor protein p53
GA	477	aspartylglucosaminidase
PIK3CA	477	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
ABC	476	ATP binding cassette subfamily A member 1
AKT	469	AKT serine/threonine kinase 1
NK	463	ankyrin 1
ESF	446	ESF1 nucleolar pre-rRNA processing protein homolog
IP	438	agouti signaling protein
MJ	437	JMJD7-PLA2G4B readthrough

Termo	Freq.	Anotação
TF	434	activating transcription factor 1
CM	433	chymase 1
ERK	426	ceramide kinase
FSH	424	follicle stimulating hormone subunit beta
AMP	423	angio associated migratory cell protein
CYP17	423	cytochrome P450 family 17 subfamily A member 1
SW	422	opsin 1, short wave sensitive
SEM	421	semaphorin 3F
AL	416	aminolevulinate dehydratase
SM	416	acetylserotonin O-methyltransferase
ML	402	calmodulin like 3
UK	401	component of inhibitor of nuclear factor kappa B kinase complex
AR	395	alanyl-tRNA synthetase 1
PE	395	alanyl aminopeptidase, membrane
KLF11	393	KLF transcription factor 11
KL	392	kallikrein related peptidase 3
SJ	389	FtsJ RNA 2'-O-methyltransferase 1
GSTT1	387	glutathione S-transferase theta 1
EMT	377	phosphatidylethanolamine N-methyltransferase
BMI	376	BMI1 proto-oncogene, polycomb ring finger
DS	376	acyl-CoA dehydrogenase short chain

Termo	Freq.	Anotação
WNT4	373	Wnt family member 4
LA	372	aminolevulinate dehydratase
SK	371	C-terminal Src kinase
NGF	369	nerve growth factor
DNMT3	368	DNA methyltransferase 3 alpha
LD	365	aldehyde dehydrogenase 1 family member A1
BA	362	4-aminobutyrate aminotransferase
TM	362	ATM serine/threonine kinase
UTR	357	utrophin
PG	353	alkaline phosphatase, germ cell
DE	351	cell death inducing DFFA like effector a
DA	349	arylacetamide deacetylase
LS	347	lipoic acid synthetase
FOXO1	346	forkhead box O1
ESR2	345	estrogen receptor 2
WT	345	WT1 transcription factor
HSD17	342	hydroxysteroid 17-beta dehydrogenase 10
NA	340	N-acetyltransferase 1
EGF	339	heparin binding EGF like growth factor
DM	338	acyl-CoA dehydrogenase medium chain
LE	332	TLE family member 5, transcriptional modulator

Termo	Freq.	Anotação
IL	327	interleukin 1 alpha
AT	325	N-acetyltransferase 1
PCB	322	pterin-4 alpha-carbinolamine dehydratase 1
TC	317	actin alpha cardiac muscle 1
MMP9	316	matrix metallopeptidase 9
MMP	309	matrix metallopeptidase 1
LH	307	folate hydrolase 1
RJ	304	protein tyrosine phosphatase receptor type J
AJ	302	DnaJ heat shock protein family (Hsp40) member B2
EGFR	301	epidermal growth factor receptor
NC	290	NIMA related kinase 9
JS	288	junctional sarcoplasmic reticulum protein 1
RM	287	cholinergic receptor muscarinic 1
CJ	285	IQ motif containing J
TA	283	actin alpha 1, skeletal muscle
NF	280	TNF receptor superfamily member 17
TS	280	steroid sulfatase
MET	278	MET proto-oncogene, receptor tyrosine kinase
DRD2	277	dopamine receptor D2
FJ	276	four-jointed box kinase 1
ID	274	BH3 interacting domain death agonist

Termo	Freq.	Anotação
JH	274	junctional cadherin complex regulator
CE	273	CEA cell adhesion molecule 1
CYP1A1	270	cytochrome P450 family 1 subfamily A member 1
AND	266	zinc finger AN1-type containing 5
EMS	266	EMSY transcriptional repressor, BRCA2 interacting
DL	263	acyl-CoA dehydrogenase long chain
AC	262	arylacetamide deacetylase
PM	260	carboxypeptidase M
SS	259	adenylosuccinate synthase 2
GC	254	crystallin gamma C
STAT3	254	signal transducer and activator of transcription 3
AE	253	AE binding protein 1
CXCL1	252	C-X-C motif chemokine ligand 1
GE	252	advanced glycosylation end-product specific receptor
IKK	252	WAP, follistatin/kazal, immunoglobulin, kunitz and netrin domain containing 1
CR	249	acrosin
LM	249	BLM RecQ like helicase
CYP2C1	248	cytochrome P450 family 2 subfamily C member 19
NR	245	apelin receptor
MH	244	anti-Mullerian hormone

Termo	Freq.	Anotação
CDKN2	241	cyclin dependent kinase inhibitor 2A
CS	240	amyloid P component, serum
KS	240	CDC28 protein kinase regulatory subunit 1B
FCRL3	239	Fc receptor like 3
FSHR	237	follicle stimulating hormone receptor
RC	237	baculoviral IAP repeat containing 2
GREB1	236	growth regulating estrogen receptor binding 1
WNT	236	Wnt family member 1
CDH1	235	cadherin 1
BRCA1	234	BRCA1 DNA repair associated
BRAF	233	B-Raf proto-oncogene, serine/threonine kinase
TIMP1	232	TIMP metallopeptidase inhibitor 1
DP	231	ADP-ribosylarginine hydrolase
YH	230	mutY DNA glycosylase
DC	230	doublecortin
SL	229	adenylosuccinate lyase
CH	228	acetylcholinesterase (Cartwright blood group)
HE	228	acetylcholinesterase (Cartwright blood group)
KC	228	dyskerin pseudouridine synthase 1
RP	228	serpin family A member 3

Termo	Freq.	Anotação
DUSP2	227	dual specificity phosphatase 2
HM	227	betaine--homocysteine S-methyltransferase
MG	225	high mobility group box 1
VEZT	224	vezatin, adherens junctions transmembrane protein
IGFBP1	222	insulin like growth factor binding protein 1
TRPV1	222	transient receptor potential cation channel subfamily V member 1
JE	221	YjeF N-terminal domain containing 3
RL	221	activin A receptor like type 1
JR	219	Jrk helix-turn-helix protein
GR	218	G protein-coupled receptor kinase 2
JD	217	JMJD7-PLA2G4B readthrough
MT	216	aminomethyltransferase
PJ	215	recombination signal binding protein for immunoglobulin kappa J region
AS	213	acid sensing ion channel subunit 2
HGF	211	hepatocyte growth factor
SY	211	synaptonemal complex protein 1
ECM	210	extracellular matrix protein 2
EBP	208	AE binding protein 1
EL	208	amelogenin X-linked
HC	208	adenosylhomocysteinase

Termo	Freq.	Anotação
CP	206	acid phosphatase 1
HOX	205	paired like homeobox 2A
CD82	204	CD82 molecule
CXCR4	203	C-X-C motif chemokine receptor 4
KG	203	diacylglycerol kinase gamma
IN	201	serpin family A member 3
ET	200	guided entry of tail-anchored proteins factor 3, ATPase
GW	200	phosphatidylinositol glycan anchor biosynthesis class W
PRL	200	prolactin releasing hormone receptor
TCF21	200	transcription factor 21
NR5A1	199	nuclear receptor subfamily 5 group A member 1
RS	199	alanyl-tRNA synthetase 1
HDAC	198	histone deacetylase 1
MAF	198	MAF bZIP transcription factor
TL	198	actin like 6A
GM	197	protein phosphatase 1 regulatory subunit 3A
IS	196	cysteine rich secretory protein 1
NAL	196	G protein subunit alpha L
AK	195	adenylate kinase 1
GS	194	crystallin gamma S
ROS	194	protein S
VOL	194	ovo like transcriptional repressor

Termo	Freq.	Anotação
		1
NME1	192	NME/NM23 nucleoside diphosphate kinase 1
BD	191	3-hydroxybutyrate dehydrogenase 1
KA	191	cholecystokinin A receptor
KM	191	creatine kinase, M-type
BC	190	ATP binding cassette subfamily A member 1
MO	190	flavin containing dimethylaniline monoxygenase 1
AD	187	arylacetamide deacetylase
CD14	187	CD14 molecule
DEG	187	delta 4-desaturase, sphingolipid 1
KH	183	ketohexokinase
IM	182	tripartite motif containing 23
RB	182	adenosine deaminase RNA specific B1
ACE	179	angiotensin I converting enzyme
AB	178	4-aminobutyrate aminotransferase
ABI	178	RAB interacting factor
MF	178	autocrine motility factor receptor
VEGFA	178	vascular endothelial growth factor A
CB	177	ATP binding cassette subfamily B member 7
JK	177	JNK1/MAPK8 associated membrane protein
EU	176	neuraminidase 1

Termo	Freq.	Anotação
BL	175	ABL proto-oncogene 1, non-receptor tyrosine kinase
FC	175	Fc alpha receptor
MIF	172	macrophage migration inhibitory factor
MW	172	DM1 locus, WD repeat containing
LP	171	alkaline phosphatase, intestinal
DN	170	brain derived neurotrophic factor
MMP2	170	matrix metallopeptidase 2
NJ	168	potassium inwardly rectifying channel subfamily J member 1
TOR	168	torsin family 1 member A
YY	168	peptide YY
CCA	168	fibrillin 2
CO2	166	aconitase 2
EJ	166	polycystin family receptor for egg jelly
KLF10	165	KLF transcription factor 10
RR	165	arrestin 3
IGFBP	164	insulin like growth factor binding protein 1
CD44	163	CD44 molecule (Indian blood group)
DD	162	adducin 1
AH	160	adenosylhomocysteinase
BPA	159	complement component 4 binding protein alpha
MLH1	158	mutL homolog 1

Termo	Freq.	Anotação
RE	158	autoimmune regulator
TGF	158	prostaglandin F receptor
BM	157	basic helix-loop-helix ARNT like 1
TH	157	tyrosine hydroxylase
GATA2	156	GATA binding protein 2
PD	155	adenosine monophosphate deaminase 1
EP	154	alanyl aminopeptidase, membrane
ESS	152	ess-2 splicing factor homolog
JB	152	gap junction protein beta 1
IC	151	acid sensing ion channel subunit 2
TSA	151	cathepsin A
DI	150	Rho GDP dissociation inhibitor alpha
LPS	150	colipase
MK	150	calcium/calmodulin dependent protein kinase IV
ON	149	one cut homeobox 1
BS	148	Bardet-Biedl syndrome 1
MUC1	148	mucin 1, cell surface associated
RO	148	shroom family member 2
CA12	147	carbonic anhydrase 12
ED	146	metallophosphoesterase domain containing 2
NSC	146	MANSC domain containing 1
AW	145	pro-apoptotic WT1 regulator
MII	145	migration and invasion inhibitory

Termo	Freq.	Anotação
		protein
EDC	142	enhancer of mRNA decapping 4
CD	141	ATP binding cassette subfamily D member 1
CY	141	aminoacylase 1
DT	141	nudix hydrolase 2
MB	140	ameloblastin
DNMT1	139	DNA methyltransferase 1
TG	139	actin gamma 1
CFL1	137	cofilin 1
SR	137	calcium sensing receptor
YM	137	crystallin mu
NPY	136	neuropeptide Y
TED	136	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 1
CTNNB1	135	catenin beta 1
FOXP3	135	forkhead box P3
RG	135	arginase 1
ZZ	135	tafazzin, phospholipid-lysophospholipid transacylase
ART	134	ADP-ribosyltransferase 1
CREB3	134	cAMP responsive element binding protein 3
DR6	134	WD repeat domain 6
KE	134	Kell metallo-endopeptidase (Kell blood group)

Termo	Freq.	Anotação
LR	134	RNA polymerase III subunit D
RERG	134	RERG like
CD10	133	CD101 molecule
EST	133	mesoderm specific transcript
TRPA1	133	transient receptor potential cation channel subfamily A member 1
FA	133	FA complementation group A
CAG	132	SHH signaling and ciliogenesis regulator SDCCAG8
AMH	131	anti-Mullerian hormone
CRIPT	131	CXXC repeat containing interactor of PDZ3 domain
PP	131	alkaline phosphatase, placental
PPAR	131	peroxisome proliferator activated receptor alpha
CREB	130	cAMP responsive element binding protein 1
GL	130	amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase
PTPN2	130	protein tyrosine phosphatase non-receptor type 2
DF	130	complement factor D
IL6	129	interleukin 6
XY	129	FXYD domain containing ion transport regulator 2
YS	129	bystin like
DNMT	128	DNA methyltransferase 1
PB	128	amyloid beta precursor protein

Termo	Freq.	Anotação
		binding family A member 1
RUNX3	128	RUNX family transcription factor 3
STS	128	steroid sulfatase
MSE	127	enolase 3
DG	127	adhesion G protein-coupled receptor B1
PAR2	127	lysophosphatidic acid receptor 2
SST	127	somatostatin
KLF9	126	KLF transcription factor 9
BR	125	ABR activator of RhoGEF and GTPase
IL1A	125	interleukin 1 alpha
SWI	124	zinc finger SWIM-type containing 8
FM	123	afamin
CCL2	121	C-C motif chemokine ligand 2
PS	121	VPS51 subunit of GARP complex
RASSF1	121	Ras association domain family member 1
UL	121	glutamate-ammonia ligase
AQP5	120	aquaporin 5
JT	120	jumping translocation breakpoint
MYC	120	MYC proto-oncogene, bHLH transcription factor
WT1	120	WT1 transcription factor
COMT	119	catechol-O-methyltransferase
SDC1	119	syndecan 1
DHT	118	dehydrogenase E1 and transketolase domain containing 1

Termo	Freq.	Anotação
ERE	118	arginine-glutamic acid dipeptide repeats
FOXL2	118	forkhead box L2
JAZF1	117	JAZF zinc finger 1
CCNE1	116	cyclin E1
CDKN1	116	cyclin dependent kinase inhibitor 1A
SB	116	acyl-CoA dehydrogenase short/branched chain
GATA6	115	GATA binding protein 6
GHRH	115	growth hormone releasing hormone
HSD1	115	hydroxysteroid 17-beta dehydrogenase 10
MPA	115	inositol monophosphatase 1
SFRP2	115	secreted frizzled related protein 2
HDAC1	113	histone deacetylase 1
HB	112	branched chain keto acid dehydrogenase E1 subunit beta
YJ	112	YJU2 splicing factor homolog
NOS	111	anosmin 1
HH	110	enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase
WC	110	CWC27 spliceosome associated cyclophilin
ANG	109	angiogenin
DH	109	alcohol dehydrogenase 1A (class I), alpha polypeptide
NIH	109	cornichon family AMPA receptor

Termo	Freq.	Anotação
		auxiliary protein 1
XRCC1	109	X-ray repair cross complementing 1
CAS	108	calpastatin
CNV	108	potassium voltage-gated channel modifier subfamily V member 1
FOXD3	108	forkhead box D3
IA	108	X-linked inhibitor of apoptosis
LIMK1	108	LIM domain kinase 1
PGP	108	pyroglutamyl-peptidase I
TJ	108	tight junction protein 1
CAM	107	activated leukocyte cell adhesion molecule
IVD	107	isovaleryl-CoA dehydrogenase
SIRT1	107	sirtuin 1
SLC22	107	solute carrier family 22 member 18
CV	106	activin A receptor type 1
DK	106	adenosine kinase
KIR	106	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 1
SNF	106	SNF8 subunit of ESCRT-II
HA	105	branched chain keto acid dehydrogenase E1 subunit alpha
ITGB1	105	integrin subunit beta 1
IR	104	autoimmune regulator
LL	104	helicase, lymphoid specific
OC	104	amine oxidase copper containing 1

Termo	Freq.	Anotação
PI	104	serpin family A member 3
HT	103	huntingtin
ILK	103	integrin linked kinase
IPA	103	lipase A, lysosomal acid type
KB	103	cholecystokinin B receptor
MAP	103	mitogen-activated protein kinase kinase kinase 8
ST	103	steroid sulfatase
MUC4	102	mucin 4, cell surface associated
NR4A	102	nuclear receptor subfamily 4 group A member 1
SG	102	alpha 2-HS glycoprotein
TE	102	testis expressed 28
AQP2	101	aquaporin 2
AT1	101	N-acetyltransferase 1
HF	101	dihydrofolate reductase
RD	101	BRCA1 associated RING domain 1
VDR	101	vitamin D receptor
CGG	99	CGG triplet repeat binding protein 1
GJ	99	gap junction protein alpha 1
HSD3B2	98	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2
LIF	98	cobalamin binding intrinsic factor
ND	98	Rho family GTPase 3
RK	98	G protein-coupled receptor kinase

Termo	Freq.	Anotação
		2
US	98	dual specificity phosphatase 1
TCL	97	ras homolog family member J
PH	97	amphiphysin
HP	96	deoxyhypusine synthase
MN	96	formin like 1
VA	96	NOVA alternative splicing regulator 1
YC	96	cytochrome c1
BRCA2	95	BRCA2 DNA repair associated
CF	95	ATP binding cassette subfamily F member 1
SF	95	arylsulfatase F
TERT	95	telomerase reverse transcriptase
FIG	94	FIG4 phosphoinositide 5-phosphatase
FL	94	cofilin 1
KP	94	karyopherin subunit alpha 1
OM	94	shroom family member 2
PLA2G2	94	phospholipase A2 group IIA
HLA-G	93	major histocompatibility complex, class I, G
RAS	93	HRas proto-oncogene, GTPase
SDS	93	serine dehydratase
VCAN	93	versican
BAS	93	Beta-adrenergic stimulation, response to

Termo	Freq.	Anotação
BAD	92	BCL2 associated agonist of cell death
MI	92	BMI1 proto-oncogene, polycomb ring finger
BCL2	91	BCL2 apoptosis regulator
GD	91	Rho GDP dissociation inhibitor alpha
GST	91	glutathione S-transferase alpha 1
RI	91	cysteine rich secretory protein 1
VIP	91	vasoactive intestinal peptide
YWHAZ	91	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein zeta
ALX	90	ALX homeobox 3
KK	90	MKKS centrosomal shuttling protein
SPARC	90	secreted protein acidic and cysteine rich
EOS	90	IKAROS family zinc finger 4
HN	89	chimerin 1
KN	89	cyclin dependent kinase inhibitor 1A
PCNA	89	proliferating cell nuclear antigen
VE	89	vascular endothelial growth factor D
CREB1	88	cAMP responsive element binding protein 1
YL	88	cylicin 1
CDC42	87	cell division cycle 42

Termo	Freq.	Anotação
CYP3A4	87	cytochrome P450 family 3 subfamily A member 4
FN1	87	fibronectin 1
PGRMC	87	progesterone receptor membrane component 2
ADC	87	antizyme inhibitor 2
BW	86	neuropeptides B and W receptor 1
MY	86	amylase alpha 1A
CAT	85	acetyl-CoA acetyltransferase 1
CD4	85	CD4 molecule
HY	85	hyaluronidase 1
NFKB1	85	nuclear factor kappa B subunit 1
SI	85	acid sensing ion channel subunit 2
DY	84	dynein cytoplasmic 1 heavy chain 1
HL	84	biphenyl hydrolase like
OS	84	oxidative stress induced growth inhibitor family member 2
QC	84	complement C1q C chain
RFRP	84	ADP ribosylation factor related protein 1
ACT	83	actin alpha 1, skeletal muscle
FOXO	83	forkhead box O1
HD	83	chromodomain helicase DNA binding protein 1
LDL	83	low density lipoprotein receptor class A domain containing 4
SMAD3	83	SMAD family member 3

Termo	Freq.	Anotação
IAP	82	X-linked inhibitor of apoptosis
IGF	82	insulin like growth factor 1
TRI	82	tripartite motif containing 23
ACTB	81	actin beta
EG	81	amphiregulin
LB	81	albumin
NP	81	alanyl aminopeptidase, membrane
PCA	81	hippocalcin
TCT	81	T cell leukemia translocation altered
DKK1	80	dickkopf WNT signaling pathway inhibitor 1
IF	80	allograft inflammatory factor 1
NF1	80	potassium voltage-gated channel modifier subfamily F member 1
ONE	80	one cut homeobox 1
CYP1B1	79	cytochrome P450 family 1 subfamily B member 1
FKBP4	79	FKBP prolyl isomerase 4
HR	79	aryl hydrocarbon receptor
EGR1	78	early growth response 1
EPHB4	78	EPH receptor B4
MUC2	78	mucin 2, oligomeric mucus/gel-forming
PPI	78	peptidylprolyl isomerase A
RHR	78	corticotropin releasing hormone receptor 1

Termo	Freq.	Anotação
HSD	77	hydroxysteroid 17-beta dehydrogenase 10
KY	77	kynurenine aminotransferase 1
SHP	77	sedoheptulokinase
SOD1	77	superoxide dismutase 1
STX	77	syntaxin 2
SV	77	cathepsin V
AGE	76	advanced glycosylation end-product specific receptor
BMAL1	76	basic helix-loop-helix ARNT like 1
GATA	76	GATA binding protein 1
GH	76	growth hormone 1
MTT	76	microsomal triglyceride transfer protein
NAT2	76	N-acetyltransferase 2
CD15	75	CD151 molecule (Raph blood group)
PDGF	75	platelet derived growth factor subunit A
GTP	74	GTP binding protein 6 (putative)
NTP	74	ectonucleoside triphosphate diphosphohydrolase 1
SOX9	74	SRY-box transcription factor 9
WI	74	twist family bHLH transcription factor 1
BDNF	73	brain derived neurotrophic factor
HIF1A	73	hypoxia inducible factor 1 subunit alpha
HO	73	ras homolog family member A

Termo	Freq.	Anotação
IGF1R	73	insulin like growth factor 1 receptor
KR	73	aldo-keto reductase family 1 member B
SUZ12	73	SUZ12 polycomb repressive complex 2 subunit
TLR	73	toll like receptor 1
WA	73	von Willebrand factor A domain containing 5A
XRCC3	73	X-ray repair cross complementing 3
AT2	72	N-acetyltransferase 2
FOS	72	Fos proto-oncogene, AP-1 transcription factor subunit
NFKB	72	nuclear factor kappa B subunit 1
NFKBI	72	NFKB inhibitor alpha
PTGS2	72	prostaglandin-endoperoxide synthase 2
TGFB1	72	transforming growth factor beta 1
BB	71	amyloid beta precursor protein binding family B member 1
HES	71	hes family bHLH transcription factor 1
LK	71	ALK receptor tyrosine kinase
SAS	71	SAM and SH3 domain containing 1
SMAD4	71	SMAD family member 4
FXR	70	FMR1 autosomal homolog 1
IGF2	70	insulin like growth factor 2
FGFR2	69	fibroblast growth factor receptor 2

Termo	Freq.	Anotação
FMR1	69	fragile X messenger ribonucleoprotein 1
NRP1	69	neuropilin 1
OMA	69	OMA1 zinc metallopeptidase
WF	69	twinfilin actin binding protein 1
CDX1	68	caudal type homeobox 1
DES	68	desmin
IGF1	68	insulin like growth factor 1
ODN	68	podocan like 1
PDF	68	peptide deformylase, mitochondrial
PDGFB	68	platelet derived growth factor subunit B
ROC	68	protein C, inactivator of coagulation factors Va and VIIIa
AMHR2	67	anti-Mullerian hormone receptor type 2
FMO3	67	flavin containing dimethylaniline monoxygenase 3
GB	67	chorionic gonadotropin subunit beta 3
HCG	67	luteinizing hormone/choriogonadotropin receptor
HMGA2	67	high mobility group AT-hook 2
FF	66	DNA fragmentation factor subunit alpha
KIR2DS5	66	killer cell immunoglobulin like receptor, two Ig domains and short cytoplasmic tail 5
SULT1	66	sulfotransferase family 1E member

Termo	Freq.	Anotação
		1
TNFRS	66	TNF receptor superfamily member 17
FGF2	65	fibroblast growth factor 2
HLA-D	65	major histocompatibility complex, class II, DM alpha
KO	65	G-patch domain and KOW motifs
SN	65	asparagine synthetase (glutamine-hydrolyzing)
TMA	65	prothymosin alpha
TOF	65	otoferlin
BMP2	64	bone morphogenetic protein 2
DAB	64	DAB adaptor protein 1
IRS2	64	insulin receptor substrate 2
RSA	64	arylsulfatase A
SERPI	64	serpin family A member 3
XRCC4	64	X-ray repair cross complementing 4
CK	63	branched chain keto acid dehydrogenase E1 subunit alpha
FSHB	63	follicle stimulating hormone subunit beta
IRS	63	insulin receptor substrate 1
PK	63	mitogen-activated protein kinase 14
CD68	62	CD68 molecule
CO	62	aconitase 1
COL1A1	62	collagen type I alpha 1 chain
RASSF2	62	Ras association domain family

Termo	Freq.	Anotação
		member 2
XX	62	death domain associated protein
CB1	61	nucleobindin 1
ESN	61	sestrin 1
EZH2	61	enhancer of zeste 2 polycomb repressive complex 2 subunit
INE	61	serpin family E member 1
PHF1	61	PHD finger protein 1
SF1	61	TNF receptor superfamily member 17
CCL5	60	C-C motif chemokine ligand 5
HG	60	Rho GTPase activating protein 1
LV	60	biliverdin reductase A
LY	60	ATP citrate lyase
MRI	60	methylthioribose-1-phosphate isomerase 1
PLT	60	phospholipid transfer protein
SMA	60	SMAD family member 1
CCT	60	FLVCR heme transporter 2
CCR2	59	C-C motif chemokine receptor 2
CHD5	59	chromodomain helicase DNA binding protein 5
COC	59	cochlin
COX2	59	cytochrome c oxidase subunit II
CPP	59	calcineurin like phosphoesterase domain containing 1
CYP11	59	cytochrome P450 family 11 subfamily A member 1

Termo	Freq.	Anotação
GAG	59	G antigen 1
KLE	59	prickle planar cell polarity protein 3
HDAC2	58	histone deacetylase 2
IB	58	Rho GDP dissociation inhibitor beta
NR4A1	58	nuclear receptor subfamily 4 group A member 1
OPN	58	opsin 1, short wave sensitive
PKA	58	inositol-trisphosphate 3-kinase A
SOD2	58	superoxide dismutase 2
CTG	57	actin gamma 1
LW	57	opsin 1, long wave sensitive
ROCK1	57	Rho associated coiled-coil containing protein kinase 1
SOD	57	superoxide dismutase 1
THE	57	thymocyte selection associated family member 2
ACA	56	acetyl-CoA acyltransferase 1
BRCA	56	BRCA1 DNA repair associated
BT	56	betacellulin
EI	56	eukaryotic translation initiation factor 2D
GAA	56	alpha glucosidase
PHB	56	EPH receptor B1
VL	56	acyl-CoA dehydrogenase very long chain
AGR2	55	anterior gradient 2, protein

Termo	Freq.	Anotação
		disulphide isomerase family member
MUC17	55	mucin 17, cell surface associated
MVD	55	mevalonate diphosphate decarboxylase
FCS	54	fucose kinase
GY	54	glycogenin 1
MSI	54	musashi RNA binding protein 1
PCE	54	interaction protein for cytohesin exchange factors 1
PTX3	54	pentraxin 3
RAF	54	A-Raf proto-oncogene, serine/threonine kinase
TBS	54	chitobiase
TIMP	54	TIMP metallopeptidase inhibitor 1
CCND1	53	cyclin D1
CD40	53	CD40 molecule
GSR	53	glutathione-disulfide reductase
GSTP1	53	glutathione S-transferase pi 1
MD1	53	adenosylmethionine decarboxylase 1
OA	53	aldolase, fructose-bisphosphate A
RPL19	53	ribosomal protein L19
STAT	53	signal transducer and activator of transcription 1
UA	53	alpha-L-iduronidase
ESR	52	estrogen receptor 1
HSD11	52	hydroxysteroid 11-beta dehydrogenase 1

Termo	Freq.	Anotação
IDO1	52	indoleamine 2,3-dioxygenase 1
RELA	52	RELA proto-oncogene, NF-kB subunit
RP11	52	poly(ADP-ribose) polymerase family member 11
SMARC	52	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 1
TTG	52	PTTG1 interacting protein
USF2	52	upstream transcription factor 2, c-fos interacting
WR	52	neuropeptides B and W receptor 1
ARNT	51	aryl hydrocarbon receptor nuclear translocator
HK	51	choline kinase alpha
IG	51	Rho GDP dissociation inhibitor gamma
KDR	51	kinase insert domain receptor
LF	51	KLF transcription factor 9
TIMP2	51	TIMP metallopeptidase inhibitor 2
UV	51	SUV39H1 histone lysine methyltransferase
VC	51	ARVCF delta catenin family member
CALD1	50	caldesmon 1
COX	50	acyl-CoA oxidase 1
CST	50	cystatin SN
DDC	50	dopa decarboxylase
DMBT1	50	deleted in malignant brain tumors 1

Termo	Freq.	Anotação
GLUT1	50	protein O-glucosyltransferase 1
HI	50	zinc finger HIT-type containing 2
OCC	50	ciliary rootlet coiled-coil, rootletin
QR	50	aquarius intron-binding spliceosomal factor
ALDH3	49	aldehyde dehydrogenase 3 family member A1
APEX	49	apurinic/apyrimidinic endodeoxyribonuclease 1
CTC	49	actin alpha cardiac muscle 1
IFNG	49	interferon gamma
JI	49	membrane anchored junction protein
KD	49	branched chain keto acid dehydrogenase E1 subunit alpha
TYK2	49	tyrosine kinase 2
WW	49	WW domain containing E3 ubiquitin protein ligase 1
YF	49	myogenic factor 5
BLM	48	BLM RecQ like helicase
CTT	48	cortactin
EP2	48	HIVEP zinc finger 2
GGC	48	gamma-glutamyl carboxylase
LG	48	Fas ligand
MMP1	48	matrix metallopeptidase 1
NL	48	formin like 1
WL	48	Wnt ligand secretion mediator

Termo	Freq.	Anotação
AAA	47	aladin WD repeat nucleoporin
CAA	47	acetyl-CoA acyltransferase 1
FOXA2	47	forkhead box A2
KF	47	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
MMP3	47	matrix metallopeptidase 3
ST2	47	bone marrow stromal cell antigen 2
ACC	46	transforming acidic coiled-coil containing protein 1
AGT	46	angiotensinogen
CARM1	46	coactivator associated arginine methyltransferase 1
CD24	46	CD247 molecule
FKBP5	46	FKBP prolyl isomerase 5
FR	46	autocrine motility factor receptor
GGA	46	golgi associated, gamma adaptin ear containing, ARF binding protein 2
HIF1	46	hypoxia inducible factor 1 subunit alpha
KL1	46	cyclin dependent kinase like 1
NI	46	BCL2 interacting protein 1
PROK1	46	prokineticin 1
RIN	46	BMP/retinoic acid inducible neural specific 1
SHC1	46	SHC adaptor protein 1
TEM	46	POTE ankyrin domain family member M

Termo	Freq.	Anotação
XL	46	AXL receptor tyrosine kinase
YT	46	acetylcholinesterase (Cartwright blood group)
AKT2	45	AKT serine/threonine kinase 2
B2M	45	beta-2-microglobulin
CXCR2	45	C-X-C motif chemokine receptor 2
EMX2	45	empty spiracles homeobox 2
FGF1	45	fibroblast growth factor 1
GAC	45	gamma-glutamylamine cyclotransferase
HSD2	45	FCH and double SH3 domains 2
KLF	45	KLF transcription factor 9
LILRB1	45	leukocyte immunoglobulin like receptor B1
MAT2A	45	methionine adenosyltransferase 2A
PN	45	opsin 1, short wave sensitive
PRB	45	proline rich protein BstNI subfamily 1
PTB	45	polypyrimidine tract binding protein 1
RIPA	45	GRIP1 associated protein 1
TGA	45	integrin subunit alpha 6
APC	44	APC regulator of WNT signaling pathway
CD36	44	CD36 molecule
DV	44	acyl-CoA dehydrogenase very long chain
EAC	44	CEA cell adhesion molecule 1

Termo	Freq.	Anotação
FZ	44	frizzled class receptor 2
HLA-C	44	major histocompatibility complex, class I, C
MIP	44	major intrinsic protein of lens fiber
MV	44	mevalonate diphosphate decarboxylase
IFN	43	interferon, type 1, cluster
ALDH1	43	aldehyde dehydrogenase 1 family member A1
APOE	43	apolipoprotein E
BGN	43	biglycan
CRABP2	43	cellular retinoic acid binding protein 2
FAM	43	transcription factor A, mitochondrial
FGF	43	fibroblast growth factor 1
IK	43	BCL2 interacting killer
LHCGR	43	luteinizing hormone/choriogonadotropin receptor
PIK3R1	43	phosphoinositide-3-kinase regulatory subunit 1
SIN3A	43	SIN3 transcription regulator family member A
UPP	43	uridine phosphorylase 1
BAX	42	BCL2 associated X, apoptosis regulator
CYP	42	acylphosphatase 1
JO	42	Josephin domain containing 1

Termo	Freq.	Anotação
NET	42	neuroepithelial cell transforming 1
PGE	42	Rap guanine nucleotide exchange factor 1
THB	42	thrombomodulin
TNFSF1	42	TNF superfamily member 11
VAS	42	vasodilator stimulated phosphoprotein
YK	42	receptor like tyrosine kinase
AGA	41	aspartylglucosaminidase
BG	41	alpha-1-B glycoprotein
BH	41	betaine--homocysteine S-methyltransferase
CTLA4	41	cytotoxic T-lymphocyte associated protein 4
GGT	41	gamma-glutamyltransferase 1
HMGCR	41	3-hydroxy-3-methylglutaryl-CoA reductase
INK4	41	serine peptidase inhibitor Kazal type 4
MARCH	41	membrane associated ring-CH-type finger 6
PDCD6	41	programmed cell death 6 interacting protein
AMIGO2	40	adhesion molecule with Ig like domain 2
CDK6	40	cyclin dependent kinase 6
HAS2	40	hyaluronan synthase 2
HPRT	40	hypoxanthine phosphoribosyltransferase 1

Termo	Freq.	Anotação
LATS1	40	large tumor suppressor kinase 1
STAT4	40	signal transducer and activator of transcription 4
TNFA	40	TNF alpha induced protein 1
GCT	40	glutaminyl-peptide cyclotransferase
ATC	39	nuclear factor of activated T cells 1
CPT	39	carnitine palmitoyltransferase 1A
LILRB2	39	leukocyte immunoglobulin like receptor B2
FH	38	zinc finger homeobox 3
HGS	38	hepatocyte growth factor-regulated tyrosine kinase substrate
LYP	38	lysophospholipase 1
RHOC	38	ras homolog family member C
SRS	38	serine and arginine rich splicing factor 1
ZNF21	38	zinc finger protein 213
KIR2DL4	37	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 4
PRA	37	protein tyrosine phosphatase receptor type A
AREG	36	amphiregulin
EK	36	checkpoint kinase 1
HES1	36	hes family bHLH transcription factor 1
ID4	36	inhibitor of DNA binding 4

Termo	Freq.	Anotação
JAK2	36	Janus kinase 2
NOS3	36	nitric oxide synthase 3
TAGLN	36	transgelin
CD16	35	CD164 molecule
ERBB2	35	erb-b2 receptor tyrosine kinase 2
FGFR1	35	fibroblast growth factor receptor 1
GPR14	35	G protein-coupled receptor 143
MAPK1	35	mitogen-activated protein kinase 14
RB1	35	adenosine deaminase RNA specific B1
STAR	35	steroidogenic acute regulatory protein
WD	35	DM1 locus, WD repeat containing
AS1	34	5'-aminolevulinate synthase 1
DOR	34	adenosine A1 receptor
EP4	34	centrosomal protein 43
GATA4	34	GATA binding protein 4
GDF9	34	growth differentiation factor 9
HBEGF	34	heparin binding EGF like growth factor
HPRT1	34	hypoxanthine phosphoribosyltransferase 1
ID2	34	glutamate ionotropic receptor delta type subunit 2
PDE	34	phosphodiesterase 1A
TN	34	actinin alpha 4

Termo	Freq.	Anotação
ACTA2	33	actin alpha 2, smooth muscle
APEX1	33	apurinic/apyrimidinic endodeoxyribonuclease 1
BK	33	inhibitor of nuclear factor kappa B kinase subunit beta
IL1B	33	interleukin 1 beta
LAMC1	33	laminin subunit gamma 1
NN	33	calponin 1
ADAM1	32	ADAM metallopeptidase domain 10
ALCAM	32	activated leukocyte cell adhesion molecule
BMP15	32	bone morphogenetic protein 15
BMP8B	32	bone morphogenetic protein 8b
E2F1	32	E2F transcription factor 1
JUN	32	Jun proto-oncogene, AP-1 transcription factor subunit
LIM	32	lens intrinsic membrane protein 2
LT	32	clathrin light chain A
NGFR	32	nerve growth factor receptor
TCF7L2	32	transcription factor 7 like 2
YES	32	YES proto-oncogene 1, Src family tyrosine kinase
AZF	32	azoospermia factor 1
CETP	31	cholesteryl ester transfer protein
EAPP	31	E2F associated phosphoprotein
NOBOX	31	NOBOX oogenesis homeobox
SP1	31	cysteine rich secretory protein 1

Termo	Freq.	Anotação
SRC	31	SRC proto-oncogene, non-receptor tyrosine kinase
STRIN	31	nitric oxide synthase trafficking
TAT	31	signal transducer and activator of transcription 1
TFAP2	31	transcription factor AP-2 alpha
XM	31	forkhead box M1
CD34	30	CD34 molecule
CFH	30	complement factor H
CYP26	30	cytochrome P450 family 26 subfamily A member 1
DBP	30	D-box binding PAR bZIP transcription factor
DNER	30	delta/notch like EGF repeat containing
ERA	30	ES cell expressed Ras
HSD3B1	30	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1
LEF1	30	lymphoid enhancer binding factor 1
LOXL1	30	lysyl oxidase like 1
MED12	30	mediator complex subunit 12
OB	30	aldolase, fructose-bisphosphate B
RPS26	30	ribosomal protein S26
SOX2	30	SRY-box transcription factor 2
SPL	30	extra spindle pole bodies like 1, separase
AHSG	29	alpha 2-HS glycoprotein

Termo	Freq.	Anotação
BF	29	beaded filament structural protein 1
CD9	29	CD9 molecule
CGA	29	glycoprotein hormones, alpha polypeptide
CHUK	29	component of inhibitor of nuclear factor kappa B kinase complex
FGF18	29	fibroblast growth factor 18
KAL1	29	CDK5 regulatory subunit associated protein 1 like 1
SPP1	29	secreted phosphoprotein 1
TIMP3	29	TIMP metallopeptidase inhibitor 3
VP	29	arginine vasopressin
BNC2	28	basonuclin 2
FE	28	EGF containing fibulin extracellular matrix protein 1
FTO	28	FTO alpha-ketoglutarate dependent dioxygenase
GALT	28	galactose-1-phosphate uridylyltransferase
HNF1	28	HNF1 homeobox A
IL1	28	interleukin 1 alpha
ITGB2	28	integrin subunit beta 2
TCF3	28	transcription factor 3
AAT	27	bile acid-CoA:amino acid N-acyltransferase
FGF9	27	fibroblast growth factor 9
GK	27	diacylglycerol kinase alpha

Termo	Freq.	Anotação
HF1	27	PHD finger protein 1
KIR2DL1	27	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 1
LI	27	perilipin 2
MAX	27	MYC associated factor X
MTHFR	27	methylenetetrahydrofolate reductase
PRELP	27	proline and arginine rich end leucine rich repeat protein
SRY	27	sex determining region Y
VN	27	vanin 2
XIAP	27	X-linked inhibitor of apoptosis
APOBEC3	26	apolipoprotein B mRNA editing enzyme catalytic subunit 3B
AURKA	26	aurora kinase A
CACNA2D1	26	calcium voltage-gated channel auxiliary subunit alpha2delta 1
CK7	26	dedicator of cytokinesis 7
DRA	26	adrenoceptor alpha 1D
FOXO3	26	forkhead box O3
GAT	26	GATA binding protein 1
HOXA9	26	homeobox A9
HTRA1	26	HtrA serine peptidase 1
HTT	26	huntingtin
IL1R1	26	interleukin 1 receptor type 1
LHR	26	prolactin releasing hormone receptor

Termo	Freq.	Anotação
NFE2L3	26	NFE2 like bZIP transcription factor 3
PDZ	26	PDZ domain containing 1
PEMT	26	phosphatidylethanolamine N-methyltransferase
PRDX6	26	peroxiredoxin 6
SFRP4	26	secreted frizzled related protein 4
SMR	26	oncostatin M receptor
VCAM1	26	vascular cell adhesion molecule 1
ADAMTS5	25	ADAM metallopeptidase with thrombospondin type 1 motif 5
CCL21	25	C-C motif chemokine ligand 21
CTCF	25	CCCTC-binding factor
GTA	25	glycoprotein alpha-galactosyltransferase 1 (inactive)
HMGA1	25	high mobility group AT-hook 1
HSP90AA1	25	heat shock protein 90 alpha family class A member 1
ICAM1	25	intercellular adhesion molecule 1
IHH	25	Indian hedgehog signaling molecule
MMP7	25	matrix metallopeptidase 7
SMOC2	25	SPARC related modular calcium binding 2
THBS1	25	thrombospondin 1
AO	24	amine oxidase copper containing 1
AQP	24	aquaporin 8

Termo	Freq.	Anotação
CTA	24	actin alpha 1, skeletal muscle
CTR	24	chymotrypsinogen B1
DHS	24	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic
E2F	24	E2F transcription factor 1
ENDO	24	endonuclease G
FIGLA	24	folliculogenesis specific bHLH transcription factor
NRAS	24	NRAS proto-oncogene, GTPase
PAR	24	poly(ADP-ribose) polymerase 1
PARP1	24	poly(ADP-ribose) polymerase 1
PMP	24	peripheral myelin protein 2
RAC1	24	Rac family small GTPase 1
RAR	24	retinoic acid receptor alpha
TWIST	24	twist family bHLH transcription factor 1
IGKC	23	immunoglobulin kappa constant
AP1	23	adenylate cyclase activating polypeptide 1
CAD	23	acyl-CoA dehydrogenase long chain
CCDC2	23	coiled-coil domain containing 28A
CD55	23	CD55 molecule (Cromer blood group)
CD8	23	CD8 subunit alpha
CNP	23	2',3'-cyclic nucleotide 3' phosphodiesterase
CX3CL1	23	C-X3-C motif chemokine ligand 1
ECE	23	endothelin converting enzyme 1

Termo	Freq.	Anotação
FAS	23	Fas cell surface death receptor
IL11	23	interleukin 11
ITGAV	23	integrin subunit alpha V
ITGB3	23	integrin subunit beta 3
ITGB7	23	integrin subunit beta 7
LEP	23	leptin
LIN28	23	lin-28 homolog A
PDGFR	23	platelet derived growth factor receptor alpha
TCF	23	transcription factor 4
CCL17	22	C-C motif chemokine ligand 17
DCT	22	dCMP deaminase
EPH	22	EPH receptor A2
FOXA1	22	forkhead box A1
GAP	22	Rho GTPase activating protein 1
GPX4	22	glutathione peroxidase 4
GSTA1	22	glutathione S-transferase alpha 1
INSR	22	insulin receptor
NPAS2	22	neuronal PAS domain protein 2
NR4A2	22	nuclear receptor subfamily 4 group A member 2
PPIA	22	peptidylprolyl isomerase A
VHL	22	von Hippel-Lindau tumor suppressor
HOXA	21	homeobox A cluster
CDC6	21	cell division cycle 6
EP3	21	E1A binding protein p300

Termo	Freq.	Anotação
FS	21	beaded filament structural protein 1
GNRHR	21	gonadotropin releasing hormone receptor
HCC	21	holocytochrome c synthase
INHA	21	inhibin subunit alpha
KLF4	21	KLF transcription factor 4
NANOG	21	Nanog homeobox
PAEP	21	progestagen associated endometrial protein
ROCK2	21	Rho associated coiled-coil containing protein kinase 2
TGFB2	21	transforming growth factor beta 2
TGFBR	21	transforming growth factor beta receptor 1
TO	21	atonal bHLH transcription factor 1
TSC2	21	TSC22 domain family member 3
WNT5A	21	Wnt family member 5A
ZEB1	21	zinc finger E-box binding homeobox 1
ANGPT	20	angiopoietin 1
CD63	20	CD63 molecule
CLDN1	20	claudin 11
DAZL	20	deleted in azoospermia like
DLX5	20	distal-less homeobox 5
DUSP1	20	dual specificity phosphatase 1
FABP5	20	fatty acid binding protein 5

Termo	Freq.	Anotação
HAT	20	choline O-acetyltransferase
ISH	20	cytokine inducible SH2 containing protein
NK1	20	ankyrin 1
NTN4	20	netrin 4
PMAIP1	20	phorbol-12-myristate-13-acetate-induced protein 1
PPARG	20	peroxisome proliferator activated receptor gamma
PPM1D	20	protein phosphatase, Mg2+/Mn2+ dependent 1D
RE2	20	microtubule associated protein RP/EB family member 2
AHR	19	aryl hydrocarbon receptor
ARHGAP11B	19	Rho GTPase activating protein 11B
CAL	19	calbindin 1
CBS	19	cystathionine beta-synthase
CCNB1	19	cyclin B1
END	19	endonuclease G
ERBB3	19	erb-b2 receptor tyrosine kinase 3
ERR	19	ERBB receptor feedback inhibitor 1
IGFBP3	19	insulin like growth factor binding protein 3
IL10	19	interleukin 10
IL6ST	19	interleukin 6 cytokine family signal transducer
LOX	19	arachidonate 12-lipoxygenase, 12S type

Termo	Freq.	Anotação
MEST	19	mesoderm specific transcript
PDK1	19	pyruvate dehydrogenase kinase 1
POMC	19	proopiomelanocortin
PPT	19	palmitoyl-protein thioesterase 1
RXRA	19	retinoid X receptor alpha
VAV3	19	vav guanine nucleotide exchange factor 3
ZEB2	19	zinc finger E-box binding homeobox 2
FD	19	elongator acetyltransferase complex subunit 1
ATF6	18	activating transcription factor 6 beta
CIN	18	calicin
CXCL2	18	C-X-C motif chemokine ligand 2
EDD	18	NEDD4 E3 ubiquitin protein ligase
FP	18	alpha fetoprotein
FX	18	FXYD domain containing ion transport regulator 2
G6PD	18	glucose-6-phosphate dehydrogenase
GALP	18	galanin like peptide
GAPD	18	glyceraldehyde-3-phosphate dehydrogenase
GCG	18	glucagon
HOXD1	18	homeobox D1
IGFBP5	18	insulin like growth factor binding protein 5
KIR3DL1	18	killer cell immunoglobulin like

Termo	Freq.	Anotação
		receptor, three Ig domains and long cytoplasmic tail 1
KISS1	18	KiSS-1 metastasis suppressor
KLF15	18	KLF transcription factor 15
NCOA1	18	nuclear receptor coactivator 1
OGG1	18	8-oxoguanine DNA glycosylase
PLA2	18	phospholipase A2 group IB
RUNX1	18	RUNX family transcription factor 1
RUNX2	18	RUNX family transcription factor 2
SCN9A	18	sodium voltage-gated channel alpha subunit 9
TGFB	18	transforming growth factor beta 1
TLR4	18	toll like receptor 4
UR	18	urocortin 2
AME	17	amelogenin X-linked
ASE	17	deoxyribonuclease 1
BRG1	17	gamma-aminobutyric acid type A receptor subunit gamma1
CALB1	17	calbindin 1
CBP	17	nuclear cap binding protein subunit 1
CCND2	17	cyclin D2
CFB	17	complement factor B
CXCL8	17	C-X-C motif chemokine ligand 8
HDL	17	high density lipoprotein binding protein
HOXC6	17	homeobox C6

Termo	Freq.	Anotação
HSC	17	HscB mitochondrial iron-sulfur cluster cochaperone
IL13	17	interleukin 13
LDHA	17	lactate dehydrogenase A
LTF	17	lactotransferrin
MAOA	17	monoamine oxidase A
MICA	17	microtubule associated monooxygenase, calponin and LIM domain containing 2
MSH2	17	mutS homolog 2
MTS	17	ADAM metallopeptidase with thrombospondin type 1 motif 4
PCC	17	propionyl-CoA carboxylase subunit alpha
PI3	17	peptidase inhibitor 3
RRAS	17	RAS related
SET	17	SET nuclear proto-oncogene
TPC	17	surfactant protein C
TYMP	17	thymidine phosphorylase
AGTR1	16	angiotensin II receptor type 1
ANA	16	aralkylamine N-acetyltransferase
APOA1	16	apolipoprotein A1
ARC	16	archain 1
BI	16	bridging integrator 1
CFL	16	cofilin 1
CITED2	16	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2

Termo	Freq.	Anotação
CML	16	opioid binding protein/cell adhesion molecule like
CRAD	16	CASP2 and RIPK1 domain containing adaptor with death domain
CREBB	16	CREB binding protein
CSF	16	colony stimulating factor 1
DICER	16	dicer 1, ribonuclease III
DIO2	16	iodothyronine deiodinase 2
FOXC1	16	forkhead box C1
FY	16	FYN binding protein 1
FYN	16	FYN proto-oncogene, Src family tyrosine kinase
GUSB	16	glucuronidase beta
IL15	16	interleukin 15
IL18	16	interleukin 18
IRE1	16	spire type actin nucleation factor 1
JUNB	16	JunB proto-oncogene, AP-1 transcription factor subunit
LHB	16	luteinizing hormone subunit beta
MAPK3	16	mitogen-activated protein kinase 3
MAR	16	microtubule affinity regulating kinase 2
MC4R	16	melanocortin 4 receptor
MD2	16	proteasome 26S subunit ubiquitin receptor, non-ATPase 2
MDM2	16	MDM2 proto-oncogene
MIM	16	small integral membrane protein 11

Termo	Freq.	Anotação
NRIP1	16	nuclear receptor interacting protein 1
PDCD4	16	programmed cell death 4
PTPN1	16	protein tyrosine phosphatase non-receptor type 1
REST	16	RE1 silencing transcription factor
SMC	16	sperm mitochondria associated cysteine rich protein
STC2	16	stanniocalcin 2
TAG	16	cancer/testis antigen 1B
TAGLN2	16	transgelin 2
YAP	16	adenylate cyclase activating polypeptide 1
ATA	15	GATA binding protein 1
BMP4	15	bone morphogenetic protein 4
CDK4	15	cyclin dependent kinase 4
CRF	15	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1
DM1	15	PR/SET domain 1
E2F4	15	E2F transcription factor 4
HAND2	15	heart and neural crest derivatives expressed 2
HNRNP	15	heterogeneous nuclear ribonucleoprotein A1
JAG1	15	jagged canonical Notch ligand 1
KIT	15	KIT proto-oncogene, receptor tyrosine kinase
PDE4	15	phosphodiesterase 4A

Termo	Freq.	Anotação
PGF	15	placental growth factor
PGK1	15	phosphoglycerate kinase 1
PLAU	15	plasminogen activator, urokinase
SAGE	15	sarcoma antigen 1
SFRP1	15	secreted frizzled related protein 1
STAT1	15	signal transducer and activator of transcription 1
TNFR	15	TNF receptor superfamily member 17
YWHAE	15	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon
B4GALT1	14	beta-1,4-galactosyltransferase 1
CCN1	14	cellular communication network factor 1
CDC2	14	cell division cycle 20
DAZ	14	deleted in azoospermia 1
DGCR8	14	DGCR8 microprocessor complex subunit
DS1	14	CDP-diacylglycerol synthase 1
ERBB	14	erb-b2 receptor tyrosine kinase 2
FUT4	14	fucosyltransferase 4
GPI	14	glucose-6-phosphate isomerase
GSTM	14	glutathione S-transferase mu 1
HNF4A	14	hepatocyte nuclear factor 4 alpha
INHBA	14	inhibin subunit beta A
MITF	14	melanocyte inducing transcription

Termo	Freq.	Anotação
		factor
MMP11	14	matrix metallopeptidase 11
MYH11	14	myosin heavy chain 11
NES	14	nestin
NR5A	14	nuclear receptor subfamily 5 group A member 2
PAK1	14	p21 (RAC1) activated kinase 1
PRLR	14	prolactin receptor
PRMT8	14	protein arginine methyltransferase 8
PTEN	14	phosphatase and tensin homolog
PTGIS	14	prostaglandin I2 synthase
ROX	14	prospero homeobox 1
STAT5	14	signal transducer and activator of transcription 5A
STAT6	14	signal transducer and activator of transcription 6
TAC1	14	tachykinin precursor 1
TFF1	14	trefoil factor 1
TLR3	14	toll like receptor 3
TRP	14	transient receptor potential cation channel subfamily M member 1
TSG	14	cathepsin G
UI	14	ubiquitin interaction motif containing 1
WNT2	14	Wnt family member 2
XO	14	decapping exoribonuclease

Termo	Freq.	Anotação
XRCC5	14	X-ray repair cross complementing 5
XT	14	alanine--glyoxylate aminotransferase
ACN	13	calcium voltage-gated channel subunit alpha1 A
ARF	13	ADP ribosylation factor 1
AXL	13	AXL receptor tyrosine kinase
BAK	13	BCL2 antagonist/killer 1
BIM	13	transmembrane BAX inhibitor motif containing 6
BMPR1	13	bone morphogenetic protein receptor type 1A
CA11	13	carbonic anhydrase 11
CEBPA	13	CCAAT enhancer binding protein alpha
CLOCK	13	clock circadian regulator
CLU	13	clusterin
COL1A2	13	collagen type I alpha 2 chain
CXCR1	13	C-X-C motif chemokine receptor 1
DKK3	13	dickkopf WNT signaling pathway inhibitor 3
FGD6	13	FYVE, RhoGEF and PH domain containing 6
GPX3	13	glutathione peroxidase 3
ICR	13	loricrin cornified envelope precursor protein
ID1	13	glutamate ionotropic receptor delta type subunit 1
IL1RA	13	interleukin 1 receptor accessory

Termo	Freq.	Anotação
		protein
ITGA2	13	integrin subunit alpha 2
ITIH4	13	inter-alpha-trypsin inhibitor heavy chain 4
KIR3DS1	13	killer cell immunoglobulin like receptor, three Ig domains and short cytoplasmic tail 1
MEP	13	meprin A subunit alpha
NOD	13	nodal growth differentiation factor
PARP	13	poly(ADP-ribose) polymerase 1
PAX2	13	paired box 2
PC1	13	glucose-6-phosphatase catalytic subunit 1
PDE8	13	phosphodiesterase 8A
PDPK1	13	3-phosphoinositide dependent protein kinase 1
PER2	13	period circadian regulator 2
PKM	13	pyruvate kinase M1/2
PLCB1	13	phospholipase C beta 1
RNH1	13	ribonuclease/angiogenin inhibitor 1
SCN11	13	sodium voltage-gated channel alpha subunit 11
SEMA3	13	semaphorin 3F
SLC	13	solute carrier family 25 member 4
SMS	13	spermine synthase
SRB	13	methionine sulfoxide reductase B2

Termo	Freq.	Anotação
STR	13	somatostatin receptor 1
TRIM2	13	tripartite motif containing 23
VIM	13	vimentin
XBP1	13	syntaxin binding protein 1
ARHI	13	DIRAS family GTPase 3
TSC1	13	TSC complex subunit 1
ACTN4	12	actinin alpha 4
AQP1	12	aquaporin 1 (Colton blood group)
CHD7	12	chromodomain helicase DNA binding protein 7
CNC	12	cyclin C
CYP2C	12	cytochrome P450 family 2 subfamily C member 19
DMD	12	dystrophin
DUSP6	12	dual specificity phosphatase 6
DVL1	12	dishevelled segment polarity protein 1
EPC1	12	enhancer of polycomb homolog 1
F1A	12	ATP synthase F1 subunit alpha
FAC	12	HGF activator
FGFRL1	12	fibroblast growth factor receptor like 1
GATA3	12	GATA binding protein 3
HSPA1	12	heat shock protein family A (Hsp70) member 1A
IGFBP2	12	insulin like growth factor binding protein 2

Termo	Freq.	Anotação
LAM	12	laminin subunit alpha 2
LAMA1	12	laminin subunit alpha 1
LAMA5	12	laminin subunit alpha 5
LARGE	12	LARGE xylosyl- and glucuronyltransferase 1
LSM	12	LSM6 homolog, U6 small nuclear RNA and mRNA degradation associated
MMP10	12	matrix metallopeptidase 10
MSH6	12	mutS homolog 6
OAT	12	ornithine aminotransferase
OCA2	12	OCA2 melanosomal transmembrane protein
PAX8	12	paired box 8
PDHB	12	pyruvate dehydrogenase E1 subunit beta
PIWI	12	piwi like RNA-mediated gene silencing 1
PKB	12	inositol-trisphosphate 3-kinase B
PRR	12	paired related homeobox 1
RAMP1	12	receptor activity modifying protein 1
RPL32	12	ribosomal protein L32
SCF	12	sec1 family domain containing 1
SHBG	12	sex hormone binding globulin
SOS1	12	SOS Ras/Rac guanine nucleotide exchange factor 1
TBX15	12	T-box transcription factor 15
TES	12	testis associated actin

Termo	Freq.	Anotação
		remodelling kinase 1
TFA	12	electron transfer flavoprotein subunit alpha
THADA	12	THADA armadillo repeat containing
TLR2	12	toll like receptor 2
TP73	12	tumor protein p73
VH	12	von Hippel-Lindau tumor suppressor
AMN	11	amnion associated transmembrane protein
ATF4	11	activating transcription factor 4
BHMT	11	betaine--homocysteine S-methyltransferase
C4BPA	11	complement component 4 binding protein alpha
CASP3	11	caspase 3
CEL	11	carboxyl ester lipase
COL3A1	11	collagen type III alpha 1 chain
CRY2	11	cryptochrome circadian regulator 2
CYP21	11	cytochrome P450 family 21 subfamily A member 2
DIRAS3	11	DIRAS family GTPase 3
DR5	11	WD repeat domain 5
EDNRA	11	endothelin receptor type A
ELK1	11	ETS transcription factor ELK1
ENO1	11	enolase 1
ERRFI1	11	ERBB receptor feedback inhibitor 1
FGF8	11	fibroblast growth factor 8

Termo	Freq.	Anotação
GJA1	11	gap junction protein alpha 1
IRS1	11	insulin receptor substrate 1
ITGB5	11	integrin subunit beta 5
KRT18	11	keratin 18
LIPC	11	lipase C, hepatic type
LKB1	11	kallikrein B1
MASP1	11	MBL associated serine protease 1
MGP	11	matrix Gla protein
MMP26	11	matrix metallopeptidase 26
MYL9	11	myosin light chain 9
NFATC2	11	nuclear factor of activated T cells 2
PDHA1	11	pyruvate dehydrogenase E1 subunit alpha 1
PER3	11	period circadian regulator 3
POLR2	11	RNA polymerase II subunit A
PPL	11	inositol polyphosphate phosphatase like 1
PPP1R1	11	protein phosphatase 1 regulatory subunit 12A
PRDM1	11	PR/SET domain 1
PROKR2	11	prokineticin receptor 2
PSEN2	11	presenilin 2
PTGER4	11	prostaglandin E receptor 4
RCC	11	regulator of chromosome condensation 1
RHOB	11	ras homolog family member B

Termo	Freq.	Anotação
SCARB1	11	scavenger receptor class B member 1
SMAD2	11	SMAD family member 2
SOX4	11	SRY-box transcription factor 4
SPOCK2	11	SPARC (osteonectin), cwcv and kazal like domains proteoglycan 2
TACR3	11	tachykinin receptor 3
TLR1	11	toll like receptor 1
MLR	11	nuclear receptor subfamily 3 group C member 2
ADAM	10	ADAM metallopeptidase domain 8
ADAM2	10	ADAM metallopeptidase domain 2
ADM	10	acyl-CoA dehydrogenase medium chain
AFP	10	alpha fetoprotein
ALDOC	10	aldolase, fructose-bisphosphate C
APOA2	10	apolipoprotein A2
CALR	10	calreticulin
CCR1	10	C-C motif chemokine receptor 1
CFD	10	complement factor D
CXCL9	10	C-X-C motif chemokine ligand 9
CYP2E1	10	cytochrome P450 family 2 subfamily E member 1
DC13	10	coiled-coil domain containing 136
DROSHA	10	drosha ribonuclease III
EPCAM	10	epithelial cell adhesion molecule
FASTK	10	Fas activated serine/threonine

Termo	Freq.	Anotação
		kinase
FGF7	10	fibroblast growth factor 7
FZD2	10	frizzled class receptor 2
GNA13	10	G protein subunit alpha 13
GNRH1	10	gonadotropin releasing hormone 1
GPER	10	G protein-coupled estrogen receptor 1
HIS	10	shisa family member 5
HSPB1	10	heat shock protein family B (small) member 1
IGSF2	10	immunoglobin superfamily member 21
IL16	10	interleukin 16
IL36	10	interleukin 36 receptor antagonist
IL6R	10	interleukin 6 receptor
MAP3K4	10	mitogen-activated protein kinase kinase kinase 4
MAP4K4	10	mitogen-activated protein kinase kinase kinase kinase 4
MGMT	10	O-6-methylguanine-DNA methyltransferase
MIB1	10	MIB E3 ubiquitin protein ligase 1
MME	10	membrane metalloendopeptidase
MMP14	10	matrix metallopeptidase 14
MTHFD1	10	methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate synthetase 1
MTR	10	5-methyltetrahydrofolate-homocysteine methyltransferase

Termo	Freq.	Anotação
MYH9	10	myosin heavy chain 9
ND2	10	cyclin D2
NME1-	10	NME1-NME2 readthrough
NR2C1	10	nuclear receptor subfamily 2 group C member 1
P2RX3	10	purinergic receptor P2X 3
PDGFC	10	platelet derived growth factor C
PRUNE2	10	prune homolog 2 with BCH domain
PTCH2	10	patched 2
PTGR	10	prostaglandin reductase 1
RAF1	10	Raf-1 proto-oncogene, serine/threonine kinase
RBBP7	10	RB binding protein 7, chromatin remodeling factor
RCS	10	BORCS8-MEF2B readthrough
RECK	10	reversion inducing cysteine rich protein with kazal motifs
SDHA	10	succinate dehydrogenase complex flavoprotein subunit A
SKP2	10	S-phase kinase associated protein 2
SLPI	10	secretory leukocyte peptidase inhibitor
SOHLH1	10	spermatogenesis and oogenesis specific basic helix-loop-helix 1
SRF	10	serum response factor
TMEM1	10	transmembrane protein 187
WNT1	10	Wnt family member 1

Termo	Freq.	Anotação
XR	10	ferredoxin reductase
MX	10	MX dynamin like GTPase 1