UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Campus Duque de Caxias Professor Geraldo Cidade

FABIO SILVA ROGERIO

PRODUÇÃO DE UMA FERRAMENTA COMPUTACIONAL PARA ANÁLISE E PROCESSAMENTO DE DADOS DE EXPRESSÃO GÊNICA EM LARGA ESCALA

Duque de Caxias- RJ

Janeiro de 2021

Fabio Silva Rogerio

PRODUÇÃO DE UMA FERRAMENTA COMPUTACIONAL PARA ANÁLISE E PROCESSAMENTO DE DADOS DE EXPRESSÃO GÊNICA EM LARGA ESCALA

Monografia apresentada ao curso de Ciências Biológicas: Biotecnologia da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários à obtenção de grau de Bacharel em Ciências Biológicas: Biotecnologia.

Orientador: Prof. Dr. Francisco José Pereira Lopes

Co-orientador: Prof. Dr. Franklin De Lima Marquezino

Duque de Caxias- RJ
Janeiro de 2021

AGRADECIMENTOS

Agradeço primeiramente aos meus orientadores e Professores com título de PhD Francisco José Pereira Lopes e Franklin de Lima Marquezino da Universidade Federal do Rio de Janeiro (UFRJ) Campus Duque de Caxias que tornaram possível a elaboração e desenvolvimento deste projeto, como também a construção desta monografia. Expresso minha gratidão também ao revisor deste trabalho, o aluno de doutorado em Modelagem Computacional do Laboratório Nacional de Computação Científica (LNCC) Claudio Daniel Tenório de Barros, por ter me dado um direcionamento pontual quanto ao conteúdo do documento e pela dedicação na revisão do texto. Também sou grato às Professoras e PhD Camila Silva de Magalhães da UFRJ Campus Duque de Caxias e Eliana Saul Furquim Werneck Abdelhay do Instituto Nacional do Câncer (INCA) que se dispuseram a avaliar o meu trabalho. Por fim, agradeço a todos os profissionais da UFRJ Campus Duque de Caxias ou o antigo Pólo de Xerém que contribuíram ou participaram da minha formação, o que permitiu que eu chegasse até a qualificação que tenho. Sou grandemente grato às coordenações envolvidas no gerenciamento dos alunos de Ciências Biológicas: Biotecnologia, especialmente à Coordenação de Desenvolvimento Educacional e Suporte Acadêmico (CODESA) e a Comissão de Orientação e Acompanhamento Acadêmico (COAA) do Campus de Duque de Caxias que por meio das quais não teria sido possível eu estar concluindo o curso.

RESUMO

Produção de uma ferramenta computacional para análise e processamento de dados de expressão gênica em larga escala é um projeto desenvolvido com o objetivo de fazer uma análise em larga escala e ao mesmo tempo uma análise refinada de dados de expressão gênica relacionados ao câncer de mama. Num primeiro estágio, estamos focando na expressão gênica de *Factor nuclear kappa B 3* (*NF-kB3)* durante a transição epitélio-mesenquimal. A via do *Factor nuclear kappa B* (*NF-kB*) tem sido relatada como uma via importante para diversos processos celulares como crescimento da célula, apoptose e resposta imune. A desregulação desta via tem sido historicamente associada a diversas formas de câncer, especialmente ao câncer de mama. Resultados recentes têm reforçado esse papel, indicando que essa via pode desempenhar papel crucial na metástase, que é responsável pela disseminação da doença para outras regiões do organismo, estando diretamente associada aos piores prognósticos da doença. Pires e colaboradores (2017) demonstraram recentemente que o gene para a expressão de *NF-kB3* ativa os genes *Twist* family BHLH transcription factor 1 (*TWIST1*), *Snail family transcriptional repressor 2* (*SNAIL2*) e *Smad interacting protein 1* (*SIP1*), que estão diretamente envolvidos na *Epithelial to mesenchymal transition* (*EMT*) ou transição epitélio-mesenquimal. Nessa transição, a célula cancerígena se desdiferencia, desprende-se de seu tecido original podendo cair na corrente sanguínea e popular outros tecidos, gerando dessa forma a metástase. O presente trabalho visa contribuir para uma melhor compreensão da regulação do gene *NF-kB3* no câncer de mama a partir da análise de dados metagenômicos de bancos de dados de linhagens celulares e de pacientes. Para esse fim, buscamos a produção dessa ferramenta computacional por meio da linguagem `Python` para o tratamento dos dados recolhidos, visando a análise de redes de sinalização gênica.

A ferramenta foi desenvolvida com o uso de orientação a objetos e diferentes módulos do `Python`. Por meio dela, produzimos gráficos para analisar os níveis de expressão gênica de 19 genes de interesse que estão ou podem estar relacionados a via do *NF-kB3, entre 1116 amostras de 1092 pacientes*. Sendo assim produzidos 627 gráficos estáticos em `PNG` e mais 209 gráficos em HTML, o que esses gráficos representam ou como estão relacionados será discutido no presente trabalho.

ABSTRACT

Production of a computational tool for the analysis and processing of gene expression data on a large scale is a project developed with the objective of carrying out a large-scale analysis and at the same time a refined analysis of gene expression data related to breast cancer. In a first stage, we are focusing on the gene expression of Nuclear factor kappa B 3 (NF-kB3) during the epithelium-mesenchymal transition. The Nuclear factor kappa B (NF-kB) pathway has been reported as an important pathway for several cellular processes such as cell growth, apoptosis and immune response. Deregulation of this pathway has historically been associated with several forms of cancer, especially breast cancer. Recent results have reinforced this role, indicating that this pathway can play a crucial role in metastasis, which is responsible for the spread of the disease to other regions of the body, being directly associated with the worst prognosis of the disease. Pires and colleagues (2017) recently demonstrated that the gene for NF-kB3 expression activates the genes Twist family BHLH transcription factor 1 (TWIST1), Snail family transcriptional repressor 2 (SNAIL2) and Smad interacting protein 1 (SIP1), which are directly involved in the Epithelial to mesenchymal transition (EMT) or epithelial-mesenchymal transition. In this transition, the cancer cell de-differentiates, detaches from its original tissue and can fall into the bloodstream and popularize other tissues, thus generating metastasis. The present work aims to contribute to a better understanding of the regulation of the NF-kB3 gene in breast cancer from the analysis of metagenomic data from databases of cell lines and patients. To that end, we seek to produce this computational tool through the Python language for the treatment of the collected data, aiming at the analysis of gene signaling networks.

The tool was developed using object orientation and different Python modules. Through it, we produce graphs to analyze the levels of gene expression of 19 genes of interest that are or may be related to the NF-kB3 pathway, among 1116 samples from 1092 patients. Thus, 627 static graphics are produced in PNG and another 209 graphics in HTML, what these graphics represent or how they are related will be discussed in the present work.

SUMÁRIO

1 INTRODUÇÃO

1.1. O câncer no mundo

O câncer tem sido o principal problema de saúde pública no mundo e se enquadra entre as quatro principais causas de morte prematura no globo. Atualmente, um óbito antes dos 70 anos é considerado como morte prematura na maioria dos países. A ocorrência de câncer no mundo tem aumentado, isso se deve, dentre outros fatores, ao envelhecimento da população global e ao aumento da exposição a fatores que contribuem para ocorrência do câncer devido ao desenvolvimento socioeconômico. Outra mudança que vêm ocorrendo é a diminuição da incidência de tipos de cânceres causados por infecções ou por más condições de vida à medida que o país se desenvolve. No entanto, se observa nestes países o aumento da incidência de outros tipos de câncer associados ao estilo de vida urbano trazido com o desenvolvimento socioeconômico (sedentarismo, alimentação inadequada, consumo de produtos que contêm compostos cancerígenos, dentre outros) (*BRAY et al., 2018*).

Em 2018, o tipo de câncer com maior incidência de novos casos no mundo foi o câncer de pulmão (2,1 milhões), seguido pelo câncer de mama (2,1 milhões) e pelo câncer de cólon e reto (1,8 milhões). Juntos, esses três tipos de câncer são a causa de um terço das mortes por câncer no mundo. A estimativa feita para o Brasil é de que para cada ano do triênio 2020-2022 ocorrerão 625 mil casos novos de câncer. Se forem excluídos os casos de câncer de pele não melanoma, serão 450 mil novos casos. O câncer que terá maior incidência no Brasil será o câncer de pele não melanoma (177 mil), seguido pelo câncer de mama (66 mil), próstata (66 mil), cólon e reto (41 mil), pulmão (30 mil) e estômago (21 mil) (*BRAY et al., 2018*).

O câncer de mama é o maior em causas de morte entre as mulheres no mundo (15%). Com quase todos os casos ocorrendo em mulheres, somente menos de 1% dos casos de câncer de mama acomete homens, e o câncer de mama corresponde a cerca de um a cada quatro casos de câncer diagnosticados entre as mulheres no mundo (24,2%) (*BRAY et al. 2018; FERLAY et al. 2018; LAUTRUP et al. 2017; MIRBAGHERI et al., 2020*).

1.2. O câncer de mama

O câncer de mama é uma doença que ocorre devido à multiplicação desordenada das células da mama, sendo geradas células anormais e levando à formação de um tumor. O tipo histológico mais comum para o câncer de mama feminino é o carcinoma de células epiteliais, e os carcinomas mais frequentes são os ductais ou lobulares, podendo ser invasivo ou *in situ* (local de origem). Há diferentes fatores que podem levar ao surgimento desta neoplasia em mulheres, mas o principal deles é a idade acima dos 50 anos (*BRAY et al. 2018; FERLAY et al., 2018*).

O câncer de mama tem sido descrito como uma doença altamente heterogênea por sua histologia, epidemiologia e propriedades moleculares. Por isso, há vários tipos de câncer de mama, com diferentes formas de desenvolvimento, alguns exibindo crescimento acelerado e outros lento. São seis os subtipos moleculares identificados para este câncer: *normal breast-like* ou *normal-like*, *luminal A* e *luminal B*, *basal-like*, *claudin-low*, e *Human Epidermal Growth Factor Receptor-type 2*/Erb-B2 Receptor Tyrosine Kinase 2 *overexpressing* (*HER2/ERB2 overexpressing*) (*TESTA et al., 2020*).

Segundo Testa e colaboradores (2020) parte dos pesquisadores agruparam e classificaram os cânceres de mama de acordo com a expressão dos importantes marcadores funcionais do *Estrogen* receptor (*ER*) ou receptor de estrogênio, do *Progesterone* receptor (*PR*) ou receptor de progesterona e do receptor da proteína *HER2*, permitindo a identificação de subtipos de tumor com diferentes desfechos. Os hormônios de estrogênio e progesterona são importantes para o crescimento das células de câncer de mama. As células de câncer de mama que se enquadram no grupo de *receptor hormonal positivo* são menos agressivas em relação às células negativas para esses receptores, pois possuem um crescimento mais lento. Se um subtipo de câncer de mama é classificado como *receptor hormonal positivo* para *ER* recebe a notação de ER+ e se classificado como *receptor hormonal negativo* para *ER* recebe a notação de ER-, da mesma forma para o PR que o subtipo de câncer de mama pode receber as notações de PR+ ou PR-, e para a proteína *HER2* o subtipo pode ser classificado como HER2+ ou HER2-. Esses marcadores também podem ser usados para caracterizar adicionalmente os subtipos moleculares: *luminal A* que é definido como ER + e / ou PR +, HER2-; o subtipo *luminal B* é definido como ER + e / ou PR +, HER2 +; e, o subtipo *HER2/ERB2 overexpressing* é definido como ER-, PR-, HER2 +. O subtipo *luminal A* é caracterizado como ER + e PR +, HER2-, tem geralmente baixas taxas de proliferação, um baixo índice de Ki67 (um marcador de

proliferação) e um bom prognóstico. Os cânceres de mama *luminal B* podem ser subdivididos em HER2- e HER2 +: os tumores HER2- são geralmente ER + (expressão mais baixa do que em tumores do subtipo *luminal A*), têm altas taxas de proliferação, alto índice de Ki67 e apresentam um prognóstico intermediário. Enquanto os *luminais B* HER2 + são geralmente ER +, PR +, têm um índice Ki67 alto e um prognóstico intermediário. O subtipo *HER2/ERB2 overexpressing* representa um grupo de cânceres de mama agressivos que estão associados a um mau prognóstico o que será descrito a seguir (*GIROUX et al. 2017; TESTA et al. 2020; WEIGELT et al., 2010*).

Há o grupo de células de câncer de mama denominadas de triplo negativo, pois não expressam os receptores hormonais citados nem o receptor para *HER2*, sendo esse grupo ER-, PR- e HER2-. Os subtipos que se enquadram neste grupo são: *normal breast-like*, *basal-like* e o *claudin-low*. Sendo o mais agressivo, esse grupo representa de 10% a 15% dos casos de câncer de mama (*AZIM et al. 2019; TESTA et al., 2020*). Os subtipos podem ser identificados por meio de técnicas moleculares visando verificar se há ou não receptores de estrogênio, progesterona e a não expressão de *HER2* já que normalmente é pouco expresso pelas células. Segundo Testa e colaboradores (2020) os cânceres do tipo *basal-like* formam um grupo heterogêneo de cânceres de mama, que provavelmente surgem de células progenitoras diferentes daquelas envolvidas em outros cânceres de mama. O câncer de mama do subtipo *claudin-low* é um grupo peculiar de cânceres de mama agressivos que além de ser caracterizado pela expressão negativa de *ER*, *PR* e *HER2*, são também caracterizados pela aquisição de metaplasia mesenquimal / sarcomatóide e / ou escamosa do epitélio maligno da mama. De acordo com Giroux e Rustgi (2017) a metaplasia seria a substituição de um tipo de célula somática diferenciada por outro tipo de célula somática diferenciada no mesmo tecido (*GIROUX et al. 2017; TESTA et al. 2020; WEIGELT et al., 2010*).

Quando as células neoplásicas se desprendem do tumor primário e se espalham para tecidos vizinhos e para outros órgãos, dá-se o processo de metástase. A metástase se dá nas seguintes etapas: a célula cancerígena cresce de forma invasiva; desprende-se do tumor primário; degrada a membrana basal; migra para o sistema circulatório; evita o ataque imunológico; escapa dos capilares (extravasação); invade e se prolifera em órgãos distantes. As metástases são responsáveis por quase todas as mortes causadas por câncer. Cerca de 90% das mortes são causadas por metástases de tumores sólidos. E quando a metástase ocorre, o tempo médio de tratamento costuma durar em torno de 5 anos, e o tempo de sobrevida entre os

pacientes varia. O tratamento é feito à base de quimioterapia ou radioterapia (*BRAY et al. 2018*; *PIRES et al. 2017*; *TESTA et al., 2020*).

1.3. Tecnologia do RNA-seq

As amostras utilizadas para as análises do estudo que será abordado mais à frente foram obtidas por meio de *RNA-seq* (sequenciamento de *RNA*) que é uma técnica que pode examinar a quantidade e as sequências de *RNA* em uma amostra usando *Next-generation sequencing* (*NGS*) ou sequenciamento de próxima geração. *NGS* é uma tecnologia usada para determinar sequências de *DNA* ou *RNA* para estudos de variações genéticas associadas às doenças ou com outros fenômenos biológicos. Essa tecnologia analisa o transcriptoma dos padrões de expressão de genes codificados em uma amostra de *RNA*. O *RNA-seq* permite investigar e descobrir o transcriptoma, o conteúdo celular total dos *RNAs*, incluindo *mRNA*, *rRNA* e *tRNA*. A compreensão do transcriptoma é fundamental se o objetivo de uma análise é conectar as informações de um genoma com a expressão de proteínas funcionais. O *RNA-seq* pode informar quais genes estão ativados em uma célula, qual é seu nível de expressão e em que momentos eles são ativados ou desativados. Isso permite um entendimento mais profundo da biologia de uma célula e a avaliação de mudanças que podem indicar doenças. Algumas das técnicas mais populares que usam *RNA-seq* são perfis transcricionais, identificação *Single Nucleotide Polymorphism* (*SNP*), edição de *RNA* e análise de expressão diferencial de genes. Podendo-se fornecer informações vitais sobre a função dos genes. Por exemplo, o transcriptoma pode destacar todos os tecidos nos quais um gene de função desconhecida é expresso, o que pode indicar qual é o seu papel (*HAN et al. 2015; OZSOLAK et al. 2011; WANG et al., 2009*).

Segundo Zelli e colaboradores (2020) a tecnologia *NGS* é capaz de sequenciar massivamente milhões de leituras de DNA, permitindo uma caracterização precisa do "estado" de múltiplos genes, usando uma quantidade muito baixa de ácidos nucléicos com considerável redução de tempo e custo. A análise da paisagem molecular dos tumores pode fornecer informações de utilidade clínica em termos de diagnóstico, prognóstico e previsão de resposta à terapia. Além disso, a tecnologia *NGS* fornece um método muito poderoso para a identificação e descoberta de novos genes responsáveis pela suscetibilidade ao câncer, com a possibilidade de aconselhar os pacientes e suas famílias em relação ao rastreamento, vigilância e opções de redução de risco (*HAN et al. 2015; OZSOLAK et al. 2011; WANG et al., 2009*).

1.4. A via de sinalização do *NF-kB* e a transição epitélio-mesenquimal

A família do complexo proteico *Factor nuclear kappa B* (*NF-kB*) tem sido descrita como importante na regulação de diferentes processos biológicos, como proliferação celular, resposta imune e inflamação. Fazem parte dessa família cinco subunidades: p50 (*NF-κB1*), p52 (*NF-κB2*), p65 (*RelA*), *c-Rel* (*Rel*) e *RelB*, que se associam para formar homo- e heterodímeros funcionais. O complexo *NF-κB* é geralmente inativo e localizado no citoplasma enquanto ligado às proteínas inibidoras de *Inhibitor of NF-kB* (*IκB*). O complexo *NF-κB* é liberado de seu inibidor quando a proteína *IκB* é fosforilada pelo complexo de quinase *IκB* (*IKK*), o que leva à ubiquitinação de *IκB* e subsequente degradação pelo proteassoma 26S. Então, o NF-κB é translocado para o núcleo e ativa a transcrição de diversos genes pela ligação ao DNA alvo específico da sequência, que é conhecido como sítios κB (50-GGGRNYYYCC-30, onde R: purina, Y: pirimidina e N: qualquer nucleotídeo) (*DE LUCA et al. 2020; PIRES et al., 2017*).

Pires e colaboradores (2017) descreveram recentemente que o *NF-kB* ativa fatores de transcrição envolvidos na *Epithelial to mesenchymal transition* (*EMT*) ou transição epitélio-mesenquimal, que desempenha papel fundamental na metástase de células de câncer de mama. Durante a *EMT*, as células de tumor perdem suas características epiteliais, com a mudança na expressão de caderinas e integrinas e a inibição da expressão de queratinas, adquirindo um fenótipo mesenquimal e levando a um comportamento migratório. Essa transição induz a resistência à apoptose após a perda da adesão celular, sendo este processo de apoptose conhecido como **anoiquia**. A interação da adesão celular e a manutenção da arquitetura tecidual normal são mediadas por várias moléculas, dentre elas as caderinas, proteínas transmembranas e cálcio-dependentes. As caderinas mais estudadas, conhecidas como clássicas, são as E (epitelial), P (placentária) e N (neural) (*PIRES et al., 2017*).

Pires e colaboradores (2017) utilizaram duas estratégias para reduzir a atividade do *NF-kB* e verificar seu papel na regulação de fatores de transcrição relacionados à *EMT*, que tem sido associada com a agressividade e com o potencial de metástases em carcinomas. A redução dessa atividade foi obtida pelo uso do inibidor *dehydroxymethylepoxyquinomicin* (*DHMEQ*) ou desidroximetilpoxiquinomicina, que é um derivado do antibiótico epoxiquinomicina que se liga diretamente ao *NF-kB/p65* ou *Factor Nuclear kappa B* 3 (*NF-kB3*) e reprime especificamente sua translocação nuclear e sua atividade de ligação ao *DNA*. A redução na transcrição do *NF-kB* foi obtida pelo uso *de siRNA*. Os *siRNA* (*RNA* de interferência curto) são sequências de *RNA* sintéticas que fazem parte do grupo dos *RNA* de interferência (*RNAi*) que participam do

mecanismo biológico usado como "proteção" contra invasões externas. Na teoria, ele pode silenciar quaisquer genes relacionados a doenças de uma maneira específica para a sequência (*DANA al et. 2017*; *HU al et., 2020*). Foram utilizadas linhagens com diversos níveis de agressividade: *MDA-MB-231 (ATCC HTB-26), HCC-1954 (ATCC CRL-2338)* e *MCF-7 (ATCC HTB-22)*. Sendo as três linhagens respectivamente do subtipo triplo-negativo, *HER2* e *luminal A*. Este bloqueio levou a uma redução da invasividade e do potencial de migração das linhagens de *MDA-MB-231* e *HCC-1954*, como também levou a baixa regulação do *SLUG* ou *Snail family transcriptional repressor 2* (*SNAIL2*), do *Twist* family BHLH transcription factor 1 (*TWIST1*), do *Smad interacting protein 1* (*SIP1*), do *Matrix Metallopeptidase 11* (*MMP11*) e da transcrição da N-caderina. Também levou a uma alta regulação da transcrição de E-caderinas. Segundo Hu e colaboradores (2019) o *SNAIL2* é um repressor da E-caderina durante o desenvolvimento do câncer, mas grande parte dos mecanismos de sua via ainda são desconhecidos. Na linhagem *MCF-7* foi feito o bloqueio com *DHMEQ*, mas não se observou uma mudança significativa na sua agressividade por ser uma linhagem do tipo *luminal A* que é menos agressiva (*HU et al., 2019, PIRES et al., 2017*).

Ferramentas de bioinformática foram utilizadas para identificar sítios de ligação do *NF-kB* ao longo dos promotores dos genes *Snail family transcriptional repressor* (*SNAIL*), *SLUG* (*SNAIL2*), *SIP1* e *TWIST1*. Por meio das técnicas de imunoprecipitação da cromatina e ensaio da luciferase foram confirmadas as ligações *NF-kB/p65* nas regiões promotoras de *TWIST1*, *SLUG* e *SIP1*. Esse estudo mostrou, portanto, que o *NF-kB* regula diretamente a transcrição dos genes envolvidos na *EMT-TF* em câncer de mama (*PIRES et al., 2017*).

1.5. Contexto dos bancos de dados no mundo e motivações para o presente trabalho

Segundo Jensen e colaboradores (2017) o recente aumento do volume de sequenciamento de alto rendimento de genomas e transcriptomas de câncer produziu um grande problema de dados que atrapalha muitos pesquisadores e oncologistas na coleta de conhecimento a respeito desses dados, da natureza dos casos de tumor maligno e da relação entre os perfis genômicos do tumor e a resposta ao tratamento. Por isso, diferentes plataformas como a *National Institute of Health Genomic Data Commons* (*NIH GDC*) (https://portal.gdc.cancer.gov/), o Cancer Genome Atlas (TCGA; https://cancergenome.nih.gov/) e o International Cancer Genomics Consortium (https://icgc.org/) surgiram e têm feito um vasto trabalho de coleta e organização dos dados a

respeito de casos clínicos de câncer. Para caracterizar os genomas de milhares de casos de câncer dentre todo vasto conjunto de histologias, esses semelhantes programas reuniram a experiência e os recursos de instituições de pesquisa, centros de câncer e empresas privadas em esforços científicos de equipes de referência. A maciça coleta de dados permitiu que as realizações de estudos de casos de câncer fossem feitas de forma mais adequada, o que elevou o nível das análises estatísticas para identificar variações gênicas nos casos de neoplasia (*JENSEN et al. 2017; MIRBAGHERI et al., 2020*).

De acordo com Mirbagheri e colaboradores (2020) falta compatibilidade entre os termos usados nos dados a respeito do câncer entre o meio científico e o meio médico, por serem muitos dados e serem oriundos de diversas fontes, o que dificulta o desenvolvimento ou a aplicação de novos tratamentos para os diferentes tipos de câncer, como o câncer de mama. Todavia, os *softwares* têm reduzido essas incompatibilidades, padronizado os dados coletados a respeito dos casos de câncer, gerado dados estatísticos diversos, análises em larga escala e disponibilizado os dados coletados e gerados para que tanto pesquisadores quanto médicos tenham acesso e desenvolvam soluções diferenciadas ou mais assertivas do que as que já existem para os casos de câncer. Sendo assim, o *Genomic Data Commons* (*GDC*) tem democratizado o acesso aos dados genômicos do câncer e promovido o compartilhamento desses dados para levar a abordagens médicas precisas para o diagnóstico e tratamento do câncer (*JENSEN et al. 2017; MIRBAGHERI et al., 2020*).

A iniciativa de colaboração entre o nosso grupo (Biologia do Desenvolvimento e Sistemas Dinâmicos) com a Professora e Doutora Eliana Abdelhay do Instituto Nacional do Câncer (INCA), juntamente com os resultados do trabalho de Pires e colaboradores (2017), assim como a disponibilidade de dados por meio da plataforma NIH GDC foi o que nos estimulou a iniciar o presente trabalho relacionado à via do *NF-kB* se utilizando de dados genômicos de pacientes com câncer de mama disponibilizados por essa plataforma, com a proposta inicial de desenvolver de uma ferramenta em `Python` que faça uma busca refinada já que, a princípio, o foco é a via do *NF-kB3* e outras vias de expressão gênica que podem estar ou estão relacionadas com a expressão dele.

1.6 O banco de dados do National Cancer Institute

O *NIH GDC* (https://portal.gdc.cancer.gov/) é um banco de dados que visa armazenar, analisar e compartilhar dados genômicos e clínicos de pacientes com câncer. Segundo Jensen e colaboradores (2017), o GDC armazena dados genômicos brutos, e permite uma reanálise contínua à medida que os métodos de computação e as anotações do genoma melhoram. Essa plataforma utiliza de mapeamentos de bioinformática compartilhados para facilitar comparações de estudos cruzados e de análises integradas de vários tipos de dados. Também mantém os dados clínicos em uma organização harmonizada, altamente estruturada e extensa, tendo como alvo manter o armazenamento a longo prazo dos dados genômicos do câncer para permitir o acesso gratuito aos pesquisadores e o compartilhamento desses dados, para levar a uma oncologia mais precisa.

A plataforma GDC é interativa com imagens e gráficos dinâmicos, o que facilita o acesso aos dados disponibilizados. Para uma visualização interativa dos dados em geral dos tipos de câncer, a página inicial possui as imagens de duas silhuetas quase sobrepostas, de um homem e de uma mulher como apresentado na **Figura 1** abaixo. Cada silhueta possui os órgãos onde há alguma incidência de câncer destacados. Os órgãos para os quais existem dados disponibilizados estão destacados em ambas as silhuetas. Ao localizar o cursor do mouse sobre a imagem de um órgão, surge a visualização para o número de casos registrados na plataforma e número de casos relacionados.

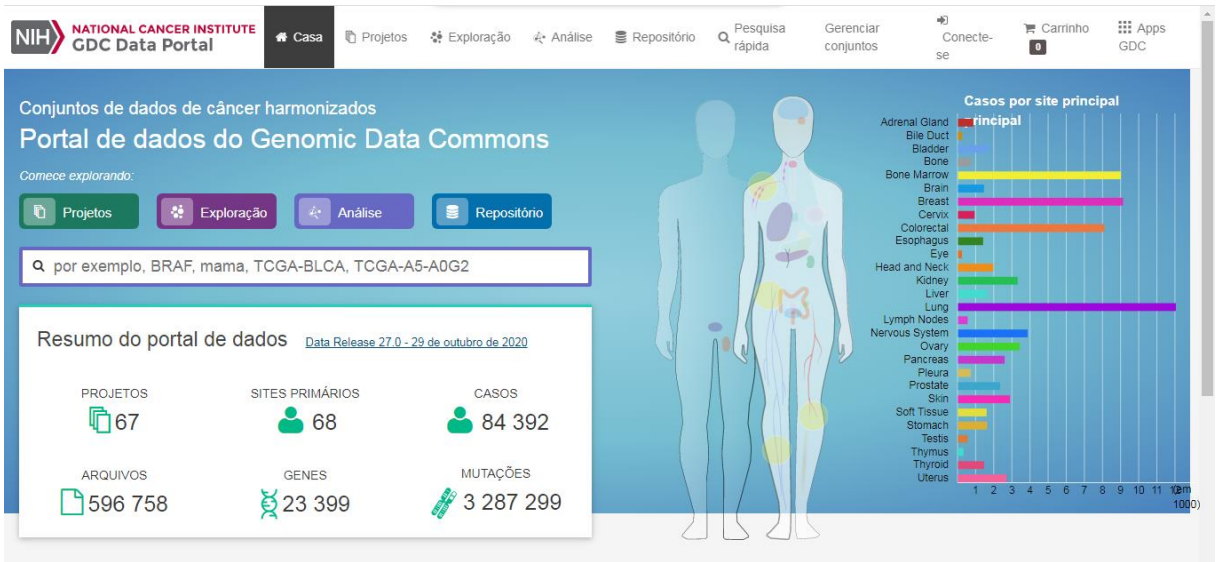

Figura 1 - *Esta é a imagem da página de entrada do Genomic Data Commons (GDC) Data Portal. A página possui imagens dinâmicas de silhuetas humanas, os botões para acessar as páginas: Projetos (Projects), Exploração (Exploration), Análises (Analysis), Repositório (Repository). A página possui o resumo dos dados*

gerais do portal e uma caixa de entrada para que os usuários possam fazer buscas no site. ***Fonte:*** <*https://portal.gdc.cancer.gov/*>.

Na página inicial do GDC há quatro botões: o botão *Projetos*, *Exploração*, *Análise* e *Repositório* (A **Figura 2** mostra detalhadamente os botões). O botão *Projetos* dá acesso aos dados selecionados de acordo com os projetos desenvolvidos. O botão *Exploração* permite o acesso aos dados relacionados aos diferentes tipos de câncer de pacientes. Na página de *Exploração* o usuário pode filtrar as informações, por exemplo, se quiser acessar apenas os dados relacionados ao câncer de próstata basta clicar no botão *prostate gland*. O botão *Análise* leva uma página onde há três opções de dados de análise são estes: *Operações de conjuntos* exibe os dados gerados a partir da interseção dos conjuntos de casos de câncer; *Análises de Dados Clínicos* exibe análises estatísticas básicas de acordo com o conjunto de dados selecionado, como a relação de casos de câncer por gênero ou por idade; *Comparação de coorte* exibe análises de sobrevivência de seus conjuntos de casos. As análises estatísticas da página da página Análises seriem úteis para análises gerais, mas, no momento, não seriam úteis para as análises relacionadas à via de síntese do *NF-kB* que envolve um conjunto de genes e que está relacionada à metástase do câncer de mama. Por fim, o botão *Repositório* leva a uma página que possui diversos dados como variações de sequências de nucleotídeos e sobre quantas cópias há destas variações. Essa página possui o mesmo mecanismo de filtro da página *Exploração*.

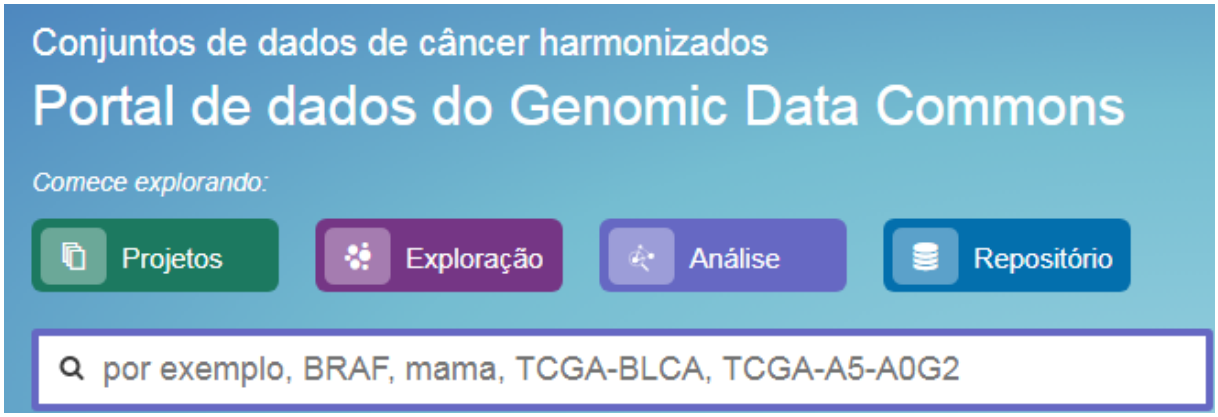

Figura 2 - *Imagem ampliada dos botões: Projetos (Projects), Exploração (Exploration), Análises (Analysis), Repositório (Repository). Abaixo dos botões se encontra a caixa de entrada ampliada.* ***Fonte:*** <*https://portal.gdc.cancer.gov/*>.

Na página inicial do GDC o usuário pode fazer uma busca direta por meio de uma caixa de entrada abaixo dos botões representada pela **Figura 3** abaixo. Além disso, esta página

fornece informações gerais sobre os seus dados, como os números de arquivos contidos em sua plataforma e o número de genes já descritos nela.

Figura 3 - *Imagem ampliada do Resumo do portal de dados com o resumo dos dados gerais da plataforma.* ***Fonte:*** <*https://portal.gdc.cancer.gov/*>.

Os arquivos utilizados para a análise da ferramenta foram obtidos por meio da página Repositório (*Repository*) representada na **Figura 4** a seguir.

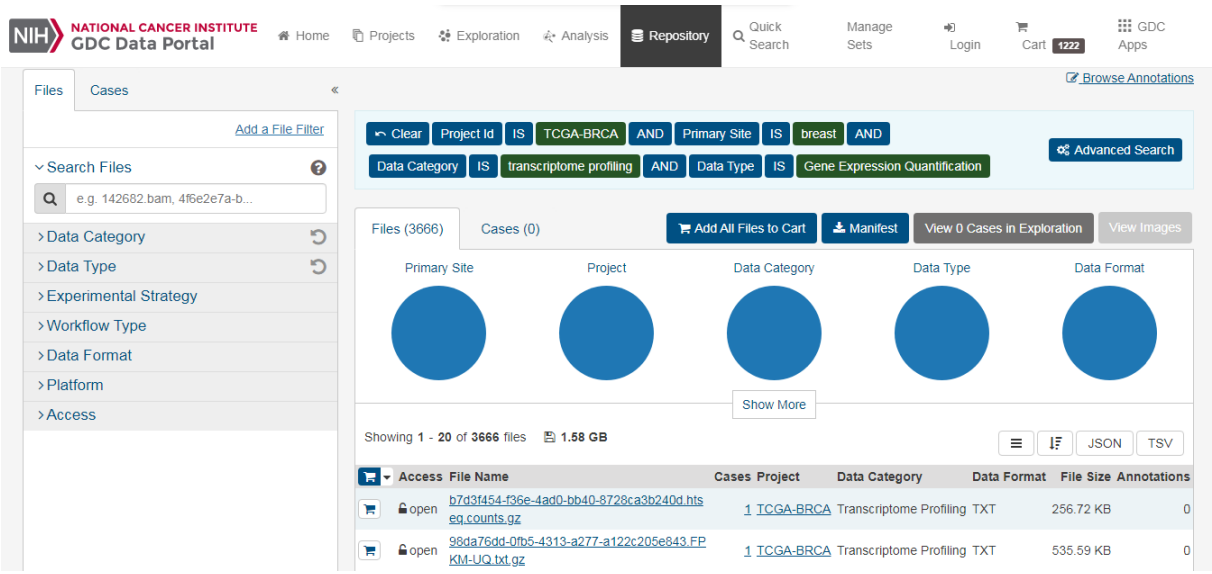

Figura 4 - *Imagem da página Repository da plataforma Genomic Data Commons (GDC) Data Portal.* **Fonte:** <*https://portal.gdc.cancer.gov/repository*>.

1.7 A linguagem Python

Com a disponibilidade dos dados de expressão gênica relacionados ao câncer de mama da plataforma *GDC*, a ferramenta apresentada neste trabalho foi desenvolvida se utilizando da linguagem de programação `Python` (https://www.python.org/) na versão 3.8.3, com o uso de orientação a objetos. O `Python` é uma linguagem de programação de alto nível, sendo, portanto, mais próxima da linguagem humana. Desse modo, o programador precisa escrever menos linhas de comando para criar um programa em comparação com linguagens de baixo nível, que são mais próximas dos comandos executados pelos computadores. Além disso, `Python` é uma linguagem interpretada, isto é, lê cada linha do código por vez, traduz para a linguagem de máquina e executa os comandos. Diferente de uma linguagem compilada, como a linguagem C que lê o código completo de uma vez, traduz para a linguagem de máquina e executa as instruções. (Anaconda Software Distribution, 2021).

Costuma-se dizer que o `Python` possui um conjunto de "baterias inclusas", que são as várias ferramentas e módulos de suporte disponíveis, o que torna o `Python` uma linguagem de uso geral. Em especial, o Python é uma linguagem muito útil para o desenvolvimento de um programa que trabalha com banco de dados, como é o caso do presente trabalho. Por suas características, o `Python` foi escolhido como linguagem para o desenvolvimento da ferramenta em questão.

Nossa ferramenta foi feita utilizando-se de orientação a objeto, um paradigma de programação baseado na composição e interação entre diversas unidades chamadas de objetos. Um objeto é uma entidade de estruturação de um programa formada por dados, em formas de atributos ou propriedades, e por códigos, em forma de procedimentos ou métodos, e é tratado como instância de classes. A classe define quais propriedades e métodos um objeto pertencente a ela pode ter, sendo, portanto, uma estrutura de comandos responsável por especificar diferentes tipos de dados trabalhados em um programa. A orientação a objetos permite escrever programas complexos de modo mais legível e facilita a reutilização de código em projetos futuros (Anaconda Software Distribution, 2021).

Atributos e métodos seriam, respectivamente, variáveis e funções associadas a uma classe. Enquanto uma função é uma estrutura criada por meio da declaração de um nome por meio do comando `def`, podendo ser criado um conjunto de comandos para executar uma tarefa, e por meio da função a execução de um conjunto de passos pode ser realizada em qualquer parte de um código sem que se repita a estrutura e, sim o nome da função (chamada da função). A

variável é um espaço de memória do computador criado por meio de uma declaração e com um endereçamento. No caso de um método construtor, ele é criado por meio da função `__init__`, esta função é chamada sempre que um objeto é criado, por meio dela são criados os atributos da classe e a definição de parâmetros como dados de entrada para um objeto. Assim, é criada uma estrutura onde se aplica o conceito de encapsulamento onde o conjunto de comandos que gera o objeto não é acessado diretamente, mas sim por meio de métodos de acesso. Os dados tipados são classificados por tipo, como o tipo de dado `string` que é um tipo de carácter imutável que pode ser acessado por meio de um endereçamento do Python, é reconhecido pelo interpretador do `Python` como um texto. Outro exemplo de tipo de dado seriam os caracteres do tipo `int` que são interpretados pelo programa como um número inteiro (Anaconda Software Distribution 2021; Python Software Foundation**,** 2020).

2 OBJETIVOS

2.1 Objetivo geral

Desenvolver uma ferramenta computacional para análise em larga escala da expressão gênica de genes envolvidos no câncer de mama.

2.2 Objetivos específicos

Aprimorar os conhecimentos de computação adquiridos durante a graduação, aplicando-os para o desenvolvimento de uma ferramenta computacional na linguagem `Python`.

Desenvolver um breve conjunto de passos para a extração de dados do *Portal GDC* do *NIH*, dos Estados Unidos, visando produzir um documento de referência para futuras obtenções de dados da plataforma.

Desenvolver uma ferramenta em `Python` que receba e leia os dados de expressão gênica obtidos, os identifique e os agrupe pelo método de clusterização, e gere os gráficos para as futuras análises a respeito da via do *NF-kB* (Essas análises não foram feitas até a realização do presente documento).

Realizar uma análise preliminar dos resultados em cima de alguns dos gráficos obtidos.

3 MATERIAIS E MÉTODOS

3.1 Os dados experimentais utilizados

No *Portal GDC* (https://portal.gdc.cancer.gov/), por meio das categorias representadas pelas imagens circulares, apresentadas mais detalhadamente na **Figura 5** abaixo, foi feita a seleção dos dados de interesse. De *Primary Site* foi selecionada a categoria *breast* (Câncer de mama), de *Project* (Projeto) foi selecionada a categoria *TCGA-BRCA*, de *Data Category* (Categoria de dado) foram selecionados os arquivos com o perfil *Transcriptome Profiling* (Perfil Transcriptoma), de *Data type* (Tipo de dado) foram selecionados os tipos dados *Gene Expression Quantification*. Em *Data Format* a única opção que há para se obter os arquivos é a `txt`.

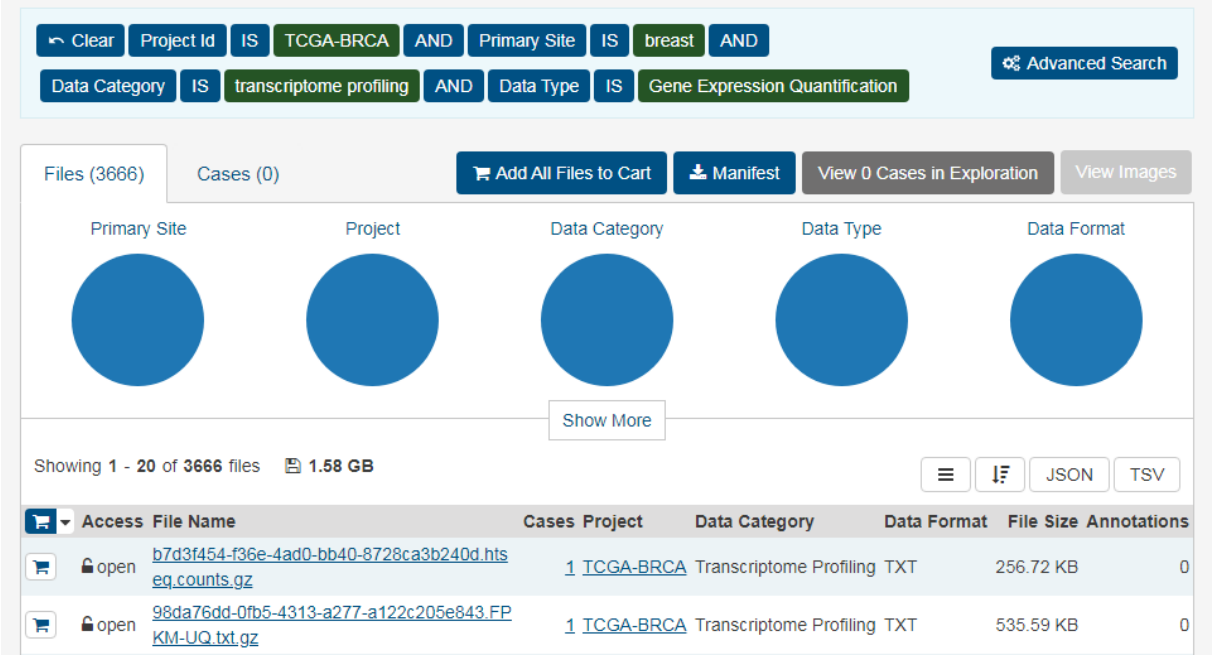

Figura 5 - *Imagem ampliada da página Repository destacando as imagens circulares com as categorias selecionadas.* **Fonte:** <*https://portal.gdc.cancer.gov/repository*>.

Em *Files* na aba *Experimental Strategy*, foi selecionada como estratégia experimental a categoria *RNA-Seq*. A aba *Workflow Type* (Tipo de Fluxo de Trabalho) permite que o usuário escolha o formato que deseja baixar o conjunto de arquivos, sendo o formato escolhido o *HTSeq – FPKM* para os dados utilizados neste projeto. No botão *Access File Name*, todos os arquivos são mantidos selecionados, ao se clicar no botão *Add All Files to Cart* os arquivos de interesse são guardados na página *Cart* mostrada na **Figura 6** abaixo. No canto superior da página

Repository, como exibido na **Figura 4** na **seção 1.6**, há o botão *cart* que leva a sua respectiva página, lá se pode visualizar e baixar os arquivos selecionados.

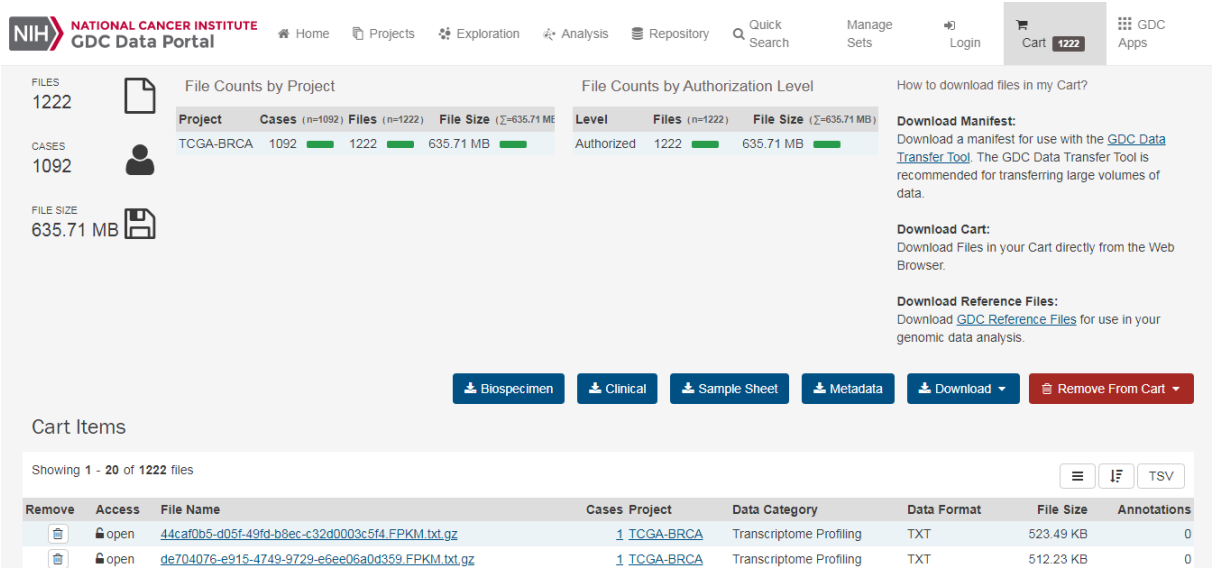

Figura 6 - *Imagem da página Cart com os dados e arquivos disponíveis para acesso e download.* **Fonte:** <*https://portal.gdc.cancer.gov/cart*>.

Por meio do botão *Download* foram baixados compactados no formato `zip` onde cada arquivo se encontra compactado em `gz`. Por intermédio do botão *Sample Sheet* se obtém a planilha da correlação entre as amostras dos pacientes e suas respectivas pastas. Os botões se encontram mais detalhados na **Figura 7** a seguir.

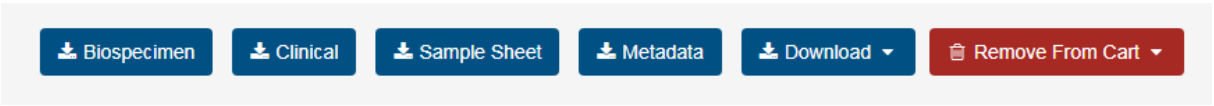

Figura 7 - *Imagem ampliada dos botões da página Cart que permitem o download dos dados selecionados na plataforma Genomic Data Commons (GDC).* **Fonte:** <*https://portal.gdc.cancer.gov/cart*>.

Há três opções de formato para se baixar os dados que são o *HTSeq - Counts*, *HTSeq - FPKM* e *HTSeq - FPKM-UQ* como mostra a **Figura 8** abaixo. O *HTSeq* é um pacote `Python` que calcula o número de leituras mapeadas para cada gene. A primeira etapa na geração de valores de expressão gênica a partir de um alinhamento de *RNA-Seq* no *GDC* é gerar uma contagem das leituras mapeadas para cada gene. Essas contagens são realizadas usando *HTSeq* e são calculadas no nível do gene. Os arquivos *HTSeq-Count* estão disponíveis em um formato

delimitado por tabulação com uma coluna de ID de gene Ensembl e uma coluna de leitura mapeada para cada gene. Esses arquivos são então processados posteriormente com scripts customizados para gerar valores FPKM e FPKM-UQ (*Genomic Data Commons Data Portal, 2021*). O *FPKM* é definido como Fragmentos por quilobase de transcrição por milhão de leituras mapeadas é um método de normalização de nível de expressão simples. O FPKM normaliza a contagem de leituras com base no comprimento do gene e no número total de leituras mapeadas. Enquanto o *FPKM-UQ* é definido como Fragmentos por quilobase de transcrição por milhão de leituras mapeadas quartil superior (FPKM-UQ) é um método de normalização de expressão baseado em RNA-Seq. O FPKM-UQ é baseado em uma versão modificada do método de normalização FPKM (*Genomic Data Commons Data Portal, 2021*).

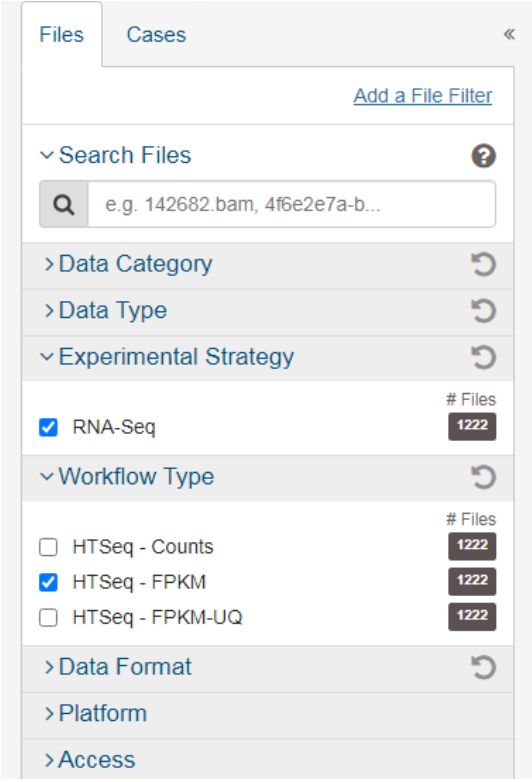

Figura 8 - *Imagem ampliada da página Repository da janela de opções para formatação dos dados a serem baixados. Sendo o RNA-seq escolhido como Estratégia Experimental (Experimental Strategy), e o HTSeq - FPKM escolhido como Tipo de Fluxo de Dados (Workflow Type).* **Fonte:** <https://portal.gdc.cancer.gov/repository>.

3.2 A estrutura dos dados experimentais

Como primeiro dado de entrada para a ferramenta foi utilizado um conjunto de 1164 arquivos, cada arquivo referente às expressões gênicas de uma das amostras de um dos 1092 casos dos pacientes com câncer de mama, todos em formato `txt`. Cada arquivo possui um código e está dentro de uma pasta com um código de identificação, e todas as pastas estão contidas dentro de uma única pasta. Cada arquivo possui duas colunas de dados, a primeira a esquerda com o código de expressão do gene e a segunda à direita possui os valores das expressões dos genes, sendo um total de 60483 expressões de genes. Um exemplo de arquivo de amostra está representado na **Figura 9** abaixo.

	Código do gene	Expressão do gene
0	ENSG00000242268.2	0.018671
1	ENSG00000270112.3	0.002591
2	ENSG00000167578.15	5.013642
3	ENSG00000273842.1	0.000000
4	ENSG00000078237.5	3.972959

Figura 9 - *Imagem com as cinco primeiras linhas de um dos arquivos de uma das amostras de pacientes. O arquivo se encontra exibido em formato de data frame. Na coluna* ***Código do gene*** *se encontram os códigos dos genes expressos, enquanto na coluna* ***Expressão do gene*** *se encontram os valores de níveis de expressão do gene.*

O segundo conjunto de arquivos utilizado foi o das 11 planilhas no formato `xlsx` contendo as amostras dos pacientes, todas as planilhas se encontram dentro de uma única pasta. São cinco pares de planilhas, cada par referente a um subtipo de câncer de mama, os quais são abordados na **seção 1.2**: *luminal A*; *luminal B*; *HER2 overexpressing*; *basal-like*; *normal-like*. Para cada um dos cinco grupos as amostras foram subdivididas em dois subgrupos, um de amostras do tecido com a neoplasia e o outro com as amostras de tecidos adjacentes a tecidos com células cancerígenas. A 11ª planilha é a junção dos códigos de todas as amostras de tecidos subjacentes normais. A planilha possui duas colunas, a primeira é a *Sample ID* que possui o código da amostra, e a segunda coluna é a *SUBTYPE* com o subtipo da amostra. No caso da imagem abaixo as amostras são referentes ao subtipo *Basal* de algum tecido subjacente ao tecido com neoplasia. A **Figura 10** abaixo representa uma das 11 planilhas utilizadas, a que se encontra na imagem é referente às amostras de subtipo de câncer de mama *Basal* e ao grupo de amostras normal (De tecido adjacente ao tecido com a neoplasia).

	A	B
1	Sample ID	SUBTYPE
2	TCGA-A7-A0CE-11A	BRCA_BASAL
3	TCGA-A7-A13E-11A	BRCA_BASAL
4	TCGA-BH-A0B3-11B	BRCA_BASAL
5	TCGA-BH-A0BW-11A	BRCA_BASAL
6	TCGA-BH-A0E0-11A	BRCA_BASAL
7	TCGA-BH-A18Q-11A	BRCA_BASAL
8	TCGA-BH-A18V-11A	BRCA_BASAL
9	TCGA-BH-A1F0-11B	BRCA_BASAL
10	TCGA-BH-A1F6-11B	BRCA_BASAL
11	TCGA-BH-A1FC-11A	BRCA_BASAL
12	TCGA-E2-A158-11A	BRCA_BASAL
13	TCGA-E2-A1LH-11A	BRCA_BASAL
14	TCGA-E2-A1LS-11A	BRCA_BASAL
15	TCGA-E9-A1N9-11A	BRCA_BASAL
16	TCGA-E9-A1ND-11A	BRCA_BASAL
17	TCGA-GI-A2C9-11A	BRCA_BASAL

Figura 10 - *Planilha com os códigos das amostras e com o subtipo de câncer. Esta imagem se trata do grupo de amostras do subtipo Basal de tecido normal (adjacente a um tecido com câncer). A Coluna* ***Sample ID*** *possui os códigos das amostras, e a coluna* ***SUBTYPE*** *possui a classificação de subtipo de câncer de mama.*

O terceiro dado de entrada utilizado foi um arquivo no formato `tsv` contendo basicamente a relação entre os códigos das amostras com os códigos dos seus respectivos arquivos contendo os códigos e valores das expressões dos genes (A imagem do arquivo se encontra representada na **Figura 11**). Na primeira coluna se encontram os código dos arquivos arquivos com as expressões dos genes, na segunda coluna os nomes dos arquivos, na terceira coluna estão descritas as categorias de dados (no caso transcriptoma), na quarta coluna o tipo de dado (quantificação e expressão gênica), na quinta coluna o tipo de projeto, na sexta coluna se encontram os códigos dos casos que ocorreram, na sétima coluna estão os códigos das amostras, e na última coluna se encontra a descrição do tipo de amostra de tumor.

	File ID	File Name	Data Category	Data Type	Project ID	Case ID	Sample ID	Sample Type
0	cadfedcc-2742-42ad-9fd3-733d01086392	dea7fd8b-f6c6-4208-861f-0da8c0808074.FPKM.txt.gz	Transcriptome Profiling	Gene Expression Quantification	TCGA-BRCA	TCGA-GM-A5PX	TCGA-GM-A5PX-01A	Primary Tumor
1	3a1703b1-969e-4806-9206-ddeadecc1e5b	fd3c752c-4a40-4ae9-bfb0-b0315eca3d1d.FPKM.txt.gz	Transcriptome Profiling	Gene Expression Quantification	TCGA-BRCA	TCGA-AN-A0XP	TCGA-AN-A0XP-01A	Primary Tumor
2	47a8e538-7928-4f6c-9e92-7b574961785f	c7169e06-bce9-4524-ab91-84c24205814c.FPKM.txt.gz	Transcriptome Profiling	Gene Expression Quantification	TCGA-BRCA	TCGA-A7-A4SF	TCGA-A7-A4SF-01A	Primary Tumor
3	7e454408-5662-4268-b605-681df38ff5bf	d9b4ed1e-39fe-4378-a9f8-f5f74a284c57.FPKM.txt.gz	Transcriptome Profiling	Gene Expression Quantification	TCGA-BRCA	TCGA-A2-A0T0	TCGA-A2-A0T0-01A	Primary Tumor
4	a479e796-26c0-474d-af84-ad603ff421a5	4c05f1ea-cd90-4a8c-b445-fb60850144d1.FPKM.txt.gz	Transcriptome Profiling	Gene Expression Quantification	TCGA-BRCA	TCGA-BH-A0HU	TCGA-BH-A0HU-01A	Primary Tumor

Figura 11 - *Tabela em formato data frame com a relação entre os códigos dos arquivos, nome dos arquivos e o código das amostras.*

3.3 As bibliotecas do Python empregadas

Para o desenvolvimento da ferramenta em `Python`, foram utilizados comandos das bibliotecas `os`, `zipfile`, `gzip`, `pandas`, `matplotlib.pyplot` e `plotly.express`. Da biblioteca `os`, foi utilizado o método `mkdir` que permite a criação de uma nova pasta no diretório desejado que é dado como parâmetro ao método.

A biblioteca `os` é um módulo do `Python` que implementa funções que podem acessar diretórios ou caminhos, podendo receber como parâmetros de caminho valores do tipo `string` ou `byte`. As bibliotecas `zipfile` e `gzip` são módulos com funções relacionadas com a descompressão de arquivos, respectivamente para arquivos do tipo `zip` e `gz`. Os módulos relacionados à descompressão foram muito importantes para se permitir a extração de arquivos comprimidos, pois os dados utilizados para análise proposta foram extraídos comprimidos de seus bancos de dados. O módulo os foi importante para se poder direcionar os arquivos extraídos para os diretórios desejados, para acessar diretórios ou caminhos existentes, também foi importante para direcionar o armazenamento dos resultados gerados (Python Software Foundation, 2020).

A ferramenta foi desenvolvida usando a plataforma Anaconda (https://anaconda.com), uma distribuição da linguagem Python direcionada à computação científica, contendo diversos módulos extras instalados para aplicações em processamento e análise de dados, além da geração de gráficos em alta qualidade, e facilitando o desenvolvimento da ferramenta. Todavia, para se utilizar as bibliotecas `matplotlib.pyplot` e `plotly.express` foram necessárias suas instalações separadamente. Ambos os módulos implementam funções que permitem a geração de gráficos. O módulo `matplotlib` (https://matplotlib.org/) permite a criação de gráficos 2D com comandos simples, podendo-se criar gráficos de diferentes tipos como barra, histograma, pizza, exponencial, dentre outros. O módulo `plotly` permite a criação de forma simples de gráficos dinâmicos. O `plotly` (https://plotly.com/python/) foi importante para a criação dos gráficos de dispersão, principalmente os dinâmicos em `HTML`. Com o `matplotlib`, poderiam ter sido criados gráficos dinâmicos de dispersão, mas teria que se utilizar uma quantidade de comandos consideravelmente maior do que para se criar um gráfico dinâmico por meio do `plotly`.

A biblioteca mais importante para o desenvolvimento desta ferramenta foi o `pandas` (https://pandas.pydata.org/), pois este módulo traz um conjunto de comandos que permitem a manipulação, filtragem, busca e classificação de dados. Por meio dos comandos desta

biblioteca, foi possível um acesso mais rápido e assertivo do que se tivesse sido feito o algoritmo para análise de um conjunto de dados sem a utilização do módulo pandas. Por meio da classe `DataFrame`, pode-se acessar arquivos em formato `xlsx`, `txt`, `csv` e `tsv`, organizá-los (quando necessário) em linhas e colunas, retornando-os como estruturas de dados do tipo `dataframe` que podem ser manipuladas. O módulo `pandas` permitiu que fossem calculados a média e o desvio padrão de cada grupo de amostras fornecidos como dados de entrada para o programa, que serão importantes para as análises futuras dos resultados gerados pela ferramenta (Python Software Foundation, 2020).

Da biblioteca padrão do `Python`, foram utilizados comandos para criação e manipulação de dicionários. Os dicionários são uma estrutura de dados onde se pode acessar um valor específico por meio de uma chave. Com os comandos do `pandas` juntamente com os relacionados a dicionários, pode-se criar novas estruturas do tipo `dataframe` com dados provenientes de diferentes arquivos ao mesmo tempo, pois para se criar um `dataframe` com a combinação de dados de diferentes arquivos basta transformar o conjunto em dicionário e em seguida transformar o dicionário em `dataframe`.

3.4 Descrição da ferramenta

Para construir a ferramenta, foi necessária a importação de alguns módulos do `Python` (https://www.python.org/) como explicado na **seção 3.3**, sendo os seguintes: `os`; `zipfile`; `gzip`; `pandas` como `pd`; `matplotlib.pyplot` como `plt`; `plotly.express` como `px`. Do módulo os foi importado o método `mkdir`. O algoritmo da ferramenta se encontra no **APÊNDICE** no final do trabalho. Para a descrição da ferramenta destacamos os nomes da função e dos métodos criados em negrito e no formato Courier.

Criamos a função **`Busca`** que utiliza da técnica de busca sequencial**.** Uma busca sequencial ocorre quando o conjunto de valores onde será feita a busca estão relacionados entre si, estando cada elemento em uma posição do conjunto, mas não estando estes elementos em uma ordem numérica ou alfabética, então a busca é feita acessando-se as posições. Essa função recebe como dado de entrada o código de expressão de gene dado pelo usuário, e através de ferramentas do `pandas` verifica se há um equivalente na lista contendo todos os códigos de expressão de gene relacionados às amostras dos pacientes com câncer de mama, pois os códigos podem conter caracteres numéricos a mais que são separados por ponto e que servem para indicar a versão da sequência de gene presente. Por exemplo, se usuário digitar o código

ENSG00000173039 para gerar os dados do *NF-kB3*, a função **Busca** irá verificar a lista com os códigos dos genes e vai retornar como dado de saída o código compatível que seria ENSG00000173039.17. Todavia, o contrário não é possível, se o usuário digitar o código de uma versão diferente da contida na lista como ENSG00000173039.19 em que os últimos caracteres depois do ponto são 19 e não 17, a função vai retornar None, ou seja, o código não será encontrado na lista embora se refira ao mesmo gene. Assim, se o código para o gene tiver uma pequena variação, mas sendo de um mesmo gene o programa vai ignorar a pequena variação no código. A chamada dessa função é feita assim que o usuário digita o código da expressão de gene de interesse.

A ferramenta implementa uma classe que encapsula projetos para analisar dados de pacientes e dos subtipos de câncer estudados, e esta foi nomeada **Projeto_Laboratorio**. Sendo os métodos dessa classe nomeados como: **Extrai_zip; Extrai_gz; Dados; Graficos.** Cada objeto desta classe é uma instância cujos atributos são definidos pelo método construtor. Dentro do método construtor foram criados atributos para serem utilizados ao longo do código que são os seguintes: diretorio; subtipos; amostra; amostras; paciente; lista; lista_paciente; dicionario; lista_subtipo; gene. O atributo diretorio armazena o caminho para diretório onde se encontram as pastas dos arquivos dos pacientes, o atributo subtipos armazena o caminho para a pasta em que se encontram as planilhas contendo a distribuição dos pacientes entre os subtipos de câncer de mama, o atributo amostra guarda o caminho para o diretório da planilha contendo a relação entre os pacientes e suas amostras, o atributo amostras guarda o valor do tipo booleano None o qual permite que o atributo permaneça vazio, então, quando o método **Graficos** é chamado esse atributo é preenchido por atribuição com o arquivo do tipo dataframe criado com os dados da planilha com a relação entre os pacientes e suas amostras. O atributo paciente guarda o código do paciente quando o método **Dados** é chamado, o atributo lista armazena uma lista que é preenchida com os caminhos dos diretórios onde se localizam os arquivos das expressões de gene de cada paciente quando o método **Extrair_gz** é chamado, o atributo lista_paciente é uma lista vazia que quando o método **Extrair_gz** é chamado armazena os códigos dos pacientes, o atributo dicionario armazena o valor do tipo inteiro 0 que na chamada do método **Extrair_gz** é substituído por atribuição por um dicionário criado através dos valores dos atributos lista_paciente e lista. O atributo lista_subtipo armazena uma lista vazia, mas na chamada do método **Graficos** a lista

recebe o caminho do endereço de cada uma das planilhas dos subtipos de câncer de mama, o que é diferente do atributo `subtipos` que só armazena o caminho para a pasta que contém essas planilhas. Por fim, o atributo `gene` armazena um valor do tipo `string` vazio que é substituído pelo código da expressão de um gene específico quando é feita a chamada do método **`Graficos`** assim que o usuário escolhe pela **opção "2"** no código principal que será abordado mais adiante.

O método construtor foi feito para receber três parâmetros. O primeiro parâmetro é o endereço no computador da pasta contendo as subpastas das amostras dos pacientes. O segundo parâmetro é o endereço da pasta contendo os arquivos dos subtipos de câncer. O terceiro e último parâmetro é para receber o diretório do arquivo contendo os códigos das amostras de todos os pacientes e suas relações com o primeiro conjunto de pastas.

Para extrair arquivos de compressões do tipo `zip` e `gz` foram implementados os métodos **`Extrai_zip`** e **`Extrai_gz`**, respectivamente. Ambos os métodos verificam se há algum arquivo comprimido dentre o conjunto de subpastas dos pacientes, realizam a descompressão e os métodos extraem os arquivos para as suas próprias pastas. Além disso, o método **`Extrai_gz`** cria um dicionário que tem como chave o código do paciente, e como valor o endereço do arquivo do paciente. Por fim, retorna o atributo `dicionario` criado no método construtor contendo o dicionário.

No código principal foi criado um menu contendo três opções. A **opção "1"** permite que sejam produzidos os gráficos de histograma, histograma com a normalização de Gauss, de dispersão estático e de dispersão dinâmico em `HTML` de um único gene digitado pelo usuário, no caso um conjunto de dados para cada um dos onze subgrupos. A **opção "2"** gera os mesmo tipos de gráficos produzidos através da **opção "1"**, mas para às 19 expressões de gene de interesse, que são os genes mais relacionados a expressão do *NF-kB*, incluindo as expressões de gene relatadas na pesquisa: *SNAIL*, *SLUG (SNAIL2): snail family transcriptional repressor 2*, *SIP1* (*ZEB2*) e *TWIST1*. A **opção "3"** permite que o usuário visualize a expressão de um gene específico de um paciente determinado, a partir dos dados de entrada. Por exemplo, se o usuário digitar o nome do arquivo da amostra do paciente como `257b1c80-203b-4cf0-9841-5a1242299270.FPKM` e em seguida digitar o código da expressão de um gene como `ENSG00000173039` ou `ENSG00000173039.17` referente ao *NF-kB3*, será retornado o valor correspondente à `25.528131` que é a contagem normalizada de leituras de expressão gênica com base no comprimento do gene e no número total de leituras de expressão gênica mapeadas.

As expressões dos genes de interesse foram organizadas em um arquivo `xlsx`. Na **Figura 12** abaixo está a imagem do arquivo com 3 colunas, tendo sido utilizadas somente as duas primeiras colunas. A primeira possui os 19 códigos dos genes de interesse, enquanto a segunda possui o nome abreviado do que o gene expressa. A terceira coluna não foi utilizada, mas contém os nomes por extenso ou a descrição do que é expresso por cada gene.

	A	B	C
1	**Ensembl**	**Gene Synonyms**	**Descrption**
2	ENSG00000173039	NFKB3, RELA, p65	RELA proto-oncogene, NF-kB subunit
3	ENSG00000109320	NFKB1, KBF1, NF-kappaB, NFKB-p50, NFkappaB, p105, p50	Nuclear factor kappa B subunit 1
4	ENSG00000077150	NFKB2, LYT-10, NF-kB2, p105, p49/p100, p52	Nuclear factor kappa B subunit 2
5	ENSG00000104856	REL-B	RELB proto-oncogene, NF-kB subunit
6	ENSG00000162924	c-Rel, I-Rel	REL proto-oncogene, NF-kB subunit
7	ENSG00000122691	TWIST1	Twist family bHLH transcription factor 1
8	ENSG00000019549	SNAIL2, SLUG, SLUGH, SLUGH1	Snail family transcriptional repressor 2
9	ENSG00000124216	SNAIL, SNAIL1, SLUGH2, SNA, SNAH	Snail family transcriptional repressor 1
10	ENSG00000169554	SIP-1, SIP1, ZFHX1B, KIAA0569	Zinc finger E-box binding homeobox 2
11	ENSG00000141736	HER2, HER-2, NEU, CD340, NGL	Erb-b2 receptor tyrosine kinase 2
12	ENSG00000091831	ESR1, ER-alpha, ESR, Era, NR3A1	Estrogen receptor 1
13	ENSG00000140009	ESR2, ER-beta, Erb, NR3A2	Estrogen receptor 2
14	ENSG00000082175	PGR, NR3C3, PR	Progesterone receptor
15	ENSG00000120659	RANKL, CD254, ODF, OPGL, TRANCE	TNF superfamily member 11
16	ENSG00000232810	TNF, TNFA, DIF, TNF-alpha, TNFSF2	Tumor necrosis factor
17	ENSG00000100906	IKBA, IkappaBalpha, MAD-3, NFKBI, NFKBIA	NFKB inhibitor alpha
18	ENSG00000146232	IKBE, NFKBIE	NFKB inhibitor epsilon
19	ENSG00000104365	IKK-beta, IKK2, IKKB, NFKBIKB, IKBKB	Inhibitor of nuclear factor kappa B kinase subunit beta
20	ENSG00000109321	AREG, AR, SDGF, AREGB, CRDGF	AREG amphiregulin [Homo sapiens (human)]

Figura 12 - *Planilha com os 19 genes de interesse relacionados com os níveis de expressão do NF-kB. A coluna Ensembl possui o código dos genes, a coluna Gene Synonyms possui o nome das expressões dos genes e seus sinônimos abreviados. A coluna Description possui a descrição do que é expresso pelos genes.*

O método **`Dados`** foi criado para que seja feita a busca sequencial da expressão de um gene específico de um paciente determinado. Este método só é chamado se o usuário, por meio do código principal, solicitar como alternativa a **opção "3"**. Então, o próprio método cria os espaços de memória, sendo variáveis locais, que irão receber o nome do arquivo do paciente e o código da expressão de gene específica, como mostrado no exemplo relacionado à **opção "3"** na **Página 28**. Por intermédio do dicionário é feita a busca, e é retornado para o usuário o valor da expressão do gene. Não sendo uma busca tão complexa já que os valores são acessados por intermédio de suas respectivas chaves no dicionário que seriam os caminhos para os diretórios que armazenam os arquivos com as expressões gênicas dos pacientes relacionadas ao câncer de mama, e os valores são os respectivos nomes dos arquivos de cada paciente.

O último método criado até o momento foi o **`Graficos`** que permite a criação de gráficos de histograma, histograma com o método de Gauss, dispersão e dispersão em `HTML`. Se espera analisar com os gráficos produzidos se ouve uma variação significativa no nível de expressão dos genes de interesse entre os pacientes com câncer de mama, se essa possível

variação significativa está relacionada com a expressão do *NF-kB3* que segundo Pires e colaboradores influencia na ocorrência da metástase do câncer de mama, e verificar qual o estado dos pacientes onde ocorreram as maiores expressões para cada gene de interesse e a relação dessas expressões com o estado atual deles. Esta parte do código distribui os dados das amostras de acordo com um dos 5 subtipos de câncer alvos da pesquisa e mencionados na **seção 3.2**. Também as amostras são distribuídas por sua classificação como tecido com células cancerígenas ou como tecido adjacente sem neoplasia, sendo que para cada subtipo abordado há amostras de pacientes para as duas categorias totalizando 10 grupos. Assim, foram gerados 10 grupos de gráficos, além desses, foi gerado um 11º grupo de gráficos por meio da junção de todas as amostras de tecidos normais adjacentes aos tecidos com células de tumor, na **Figura 13** abaixo se encontra a tabela representando a divisão dos grupos (no caso para a expressão do gene *NF-kB3).*

Grupo de amostras	Média (μ)	Desvio padrão (σ)
SubType_ID_Normal_tissue_BASAL	17,969773	3,40662
SubType_ID_Tumor_BASAL	21,381058	6,815397
SubType_ID_Normal_tissue_HER2	16,460214	3,209475
SubType_ID_Tumor_HER2	18,345157	5,323197
SubType_ID_Normal_tissue_LUMA	18,0471	4,380292
SubType_ID_Tumor_LUMA	19,221365	4,706811
SubType_ID_Normal_tissue_LUMB	18,751385	3,152292
SubType_ID_Tumor_LUMB	18,397078	5,344799
SubType_ID_Normal_tissue_NORMAL	21,689123	0,311115
SubType_ID_Tumor_NORMAL	20,327852	4,160568
SubType_ID_Normals_tissue	18,112497	3,925887

Figura 13 - *Tabela relacionada com a expressão de NF-kB3, contém os dados dos grupos de amostras dos pacientes referentes a essa expressão de gene. A coluna* ***Grupo de amostras*** *possui o nome fornecido para cada grupo, a coluna* ***Média (μ)*** *possui a média de cada um e a coluna* ***Desvio padrão (σ)*** *possui os valores de desvio padrão dos grupos.*

O método verifica se seus dois parâmetros de entrada foram alterados. Ambos os parâmetros possuem como argumento prévio a palavra de propriedade do Python `None`, quando o usuário digita a **opção "2"** o primeiro parâmetro recebe o código do gene de interesse do usuário e o segundo parâmetro recebe o nome do gene. Então, utilizando-se dos dados de entrada fornecidos ao objeto no código principal, todas as expressões relacionadas ao gene dado como parâmetro para o método **`Graficos`** são selecionadas e agrupadas pelo método de clusterização de acordo com os grupos de amostras apresentados pelas 11 tabelas.

Clusterização ou **classificação não-supervisionada** seria método computacional de agrupamento de dados de acordo com características em comum entre eles, como no caso dos dados das amostras dos pacientes que são agrupadas de acordo com sua identificação prévia dentro de um dos subtipos de câncer de mama. Para a expressão do gene ser selecionada de cada amostra, cada um dos 1164 arquivos é acessado por meio do sistema de códigos na tabela `tsv` (tabela mostrada na **Figura 11** na **seção 3.2**) com a correlação dos códigos das amostras com os códigos dos arquivos. Os gráficos foram gerados e salvos automaticamente no formato png em uma pasta criada pelo método, sendo o diretório para criação da pasta já pré-determinado no código. Os gráficos de histogramas, com e sem a normalização de Gauss, foram criados com os comandos da biblioteca `matplotlib`, enquanto os gráficos de dispersão dinâmicos e estáticos foram criados por meio de comandos da biblioteca `plotly`. Além disso, o método calculou a média e o desvio padrão para cada grupo de amostras para cada gene.

4 RESULTADOS

4.1 Gráficos gerados

Foram criados um total de 627 gráficos estáticos em formato `PNG` relacionados aos 19 genes de interesse e mais 209 gráficos dinâmicos em `HTML` com a mesma relação, todos serão utilizados para análises futuras que não foram realizadas previamente para este trabalho. Todavia, apenas 6 gráficos serão apresentados e discutidos no presente documento. Os 19 genes de interesse para este trabalho se encontram descritos na **Figura 12** na **seção 3.4**. Em cada gráfico de histograma com ou sem a normalização de Gauss foi impresso como legenda o respectivo desvio padrão e a média para facilitar as futuras análises, ambos os valores foram utilizados para se fazer a distribuição normal para cada gráfico de histograma, sendo criados os gráficos de histograma com a normalização de Gauss, o cálculo da conversão dos valores para uma distribuição normalizada será apresentado mais à frente nesta seção.

As 1164 amostras dos 1092 pacientes estão distribuídas em 11 grupos como descrito na **seção 3.2**. Sendo estes grupos divididos em duas categorias, no caso, a categoria de amostras de tecido normal extraídas de tecidos adjacentes aos tecidos com o tumor e a categoria das amostras de tecido extraídas de tecidos com a neoplasia na mama. Em ambas as categorias as amostras de tecido estão distribuídas entre cinco dos subtipos de câncer como descrito na **seção 3.2** As características de cada subtipo de câncer de mama se encontram descritas na **seção 1.2**. Assim, totalizando 10 grupos de amostras. O 11º grupo se refere a junção das amostras dos 5 grupos das amostras de tecido normal retiradas de tecidos adjacentes tecidos com o câncer de mama por se terem poucas amostras nesta categoria em comparação com cada grupo de amostras de tecido com tumor.

Cada amostra possui as quantificações para 60483 expressões de genes feitas pelo método *FPKM* (Fragmentos por quilobase de transcrição por milhão de leituras mapeadas) que foi descrito na **seção 3.1**. Dentre os genes quantificados se encontram os genes de interesse no momento. Os gráficos foram produzidos para cada um dos 19 genes de interesse, sendo que as amostras estão distribuídas dentre os 10 grupo de acordo com o subtipo de câncer de mama e do estado do tecido de origem como descrito no parágrafo anterior, tendo sido criado também o 11° grupo. Para cada um dos 11 grupos de cada gene foram criados 3 gráficos estáticos (histograma, histograma com a normalização de Gauss e dispersão) e 1 dispersão dinâmico, ou seja, foram criados 33 gráficos estáticos em `PNG` e 11 dinâmicos em `HTML` por gene, somando-

se os gráficos de todos os genes de interesse a quantidade vai para 627 gráficos estáticos e 209 gráficos dinâmicos produzidos, totalizando 836 gráficos.

Os gráficos de histograma ou de distribuição de frequências em `PNG` foram criados para se observar as ocorrências dos diferentes níveis de expressão de um mesmo gene dentro de um grupo de amostras, onde a altura ou o eixo Y dos gráficos seriam o número de amostras com altura máxima de 90 amostras por coluna, distribuídas ao longo do eixo X onde está a escala com as expressões do gene. Como exemplo, a **Figura 15** na **seção 4.2** que apresenta um dos gráficos produzidos se refere à expressão do *NF-kB3* da categoria de amostras de tumor do subtipo *luminal A*. Os gráficos de histograma com a normalização de Gauss foram criados com o mesmo formato e quase os mesmos propósitos que os gráficos de histograma descritos anteriormente, a diferença é que a escala do eixo X foi padronizada para uma distribuição normal onde o ponto médio ou média (μ) dentre os valores de cada grupo é igual a 0, sendo os pontos da escala pertencentes aos reais e estando contidos no intervalo ($-\infty,+\infty$). A **Figura 14** na **seção 4.2** mostra um dos gráficos de histograma com a normalização de Gauss gerados, este gráfico está relacionado ao mesmo grupo do gráfico da **Figura 15**, a diferença se encontra na escala do eixo X com a distribuição normal. A distribuição normal torna os gráficos de diferentes grupos ou de diferentes expressões de genes mais comparáveis entre si, também permite o cálculo da probabilidade da ocorrência de cada intervalo de valores de expressão de um gene dentro da área abaixo da curva do gráfico.

Para a normalização de Gauss cada valor de expressão de gene foi aplicado na equação de distribuição normal dada como: $Z = \frac{X - \mu}{\sigma}$, sendo ($Z$) a distribuição normal, (X) a amostra que neste caso seria o valor para expressão de um gene, (μ) a média entre os níveis de expressão de genes do grupo de amostras e (σ) o desvio padrão do mesmo grupo. Sendo a variância (Var): **Var = $(\sigma)^2$** (Variância é igual ao desvio padrão elevado ao quadrado). A **Figura 13** na **seção 3.4** mostra as médias e desvios padrões dos 11 grupos para *NF-Kb3*.

Os gráficos de dispersão foram criados com o propósito de se visualizar a distribuição de cada amostra individualmente dentro do grupo, e se observar como está a distribuição dos pontos de amostras dentro das faixas de valores de expressão gênica no eixo Y dos gráficos para cada grupo. Os gráficos foram criados com a relação entre os níveis de expressão gênica no eixo Y e os códigos das amostras no eixo X, como mostrado na **Figura 16** na **seção 4.2** que se refere à expressão de *NF-kB3* da categoria de amostras de tumor de tecidos com o câncer de mama do subtipo *luminal A*. Foram produzidos dois tipos de gráficos de dispersão: estático e

dinâmico. No caso, dois gráficos iguais para cada grupo de cada gene, mas com formatos diferentes. Observamos que os gráficos de dispersão estáticos não são os ideais para análise quando há muitos pontos com valores próximos e sobrepostos, já que a visualização e distinção dos pontos fica limitada, mesmo assim, foram produzidos pois são mais adequados para colocar em um documento estático, tendo sido criados em formato `PNG`. Os gráficos de dispersão dinâmicos em `HTML` foram produzidos para se poder acessar individualmente cada ponto e identificar o código do paciente e nível de expressão gênica no gráfico, pois esse tipo de gráfico diferencia cada ponto e permite à amplificação da imagem para a visualização, mesmo após o gráfico ter sido criado. As vantagens dos gráficos de dispersão em `HTML` irão favorecer as análises, pois precisaremos identificar as amostras nos gráficos para se verificar, por exemplo, como se encontram atualmente os pacientes doadores das amostras que tiveram uma expressão anormal para um conjunto de genes.

4.2 Gráficos escolhidos para o presente documento

Selecionamos somente 6 gráficos para apresentar e discutir no presente trabalho, como mencionado no início da **seção 4.1**. Sendo 3 gráficos relacionados à expressão de *NF-kB3* ou *p65* e 3 gráficos relacionados à expressão de *SNAIL2* (*SLUG*). Os gráficos de dispersão dinâmicos de ambos os grupos serão citados na **seção 4.3**, mas não estão presentes no trabalho, pois o anexo deles não se mostrou viável em um documento estático. Os gráficos referentes ao *NF-kB3* e ao *SNAIL2* foram escolhidos, primeiramente como exemplos, secundariamente porque segundo Pires e colaboradores (2017), como explicado na **seção 1.4**, a expressão de *NF-kB3* durante a transição epitélio-mesenquimal influência na expressão do *SNAIL2 (SLUG)*, *TWIST1* e *SIP1*, via que está envolvida na metástase do câncer de mama, assim, dentre os três escolhemos os gráficos referentes à expressão de *SLUG* aleatoriamente. Os grupos escolhidos foram das categorias de amostras normais e de amostras com o tumor do subtipo de câncer de mama *luminal A*, pois o subtipo *luminal A* é o que concentra a maior proporção de amostras de pacientes em meio aos dados.

Os gráficos relacionados ao *NF-kB3* são do grupo de amostras de tecidos com câncer de mama do subtipo *luminal A* nomeados de *SubType ID Tumor LUMA – NFKB3*, sendo 1 gráfico de histograma estático sem a normalização, um de histograma estático com a normalização e um de dispersão estático. Os 3 gráficos relacionados ao *SNAIL2* são do mesmo grupo de amostras e foram nomeados como *SubType ID Tumor LUMA – SNAIL2*, sendo os mesmos tipos de gráficos do grupo *SubType ID Tumor LUMA – NFKB3*. Os gráficos de histograma sem e

com a normalização e os de dispersão estáticos do grupo *SubType ID Tumor LUMA – NFKB3* se encontram dentre as imagens abaixo: **Figura 14**; **Figura 15**; **Figura 16**. Os gráficos referentes ao grupo *SubType ID Tumor LUMA – SNAIL2* se encontram nas imagens abaixo nomeadas de: **Figura 17**; **Figura 18**; **Figura 19**.

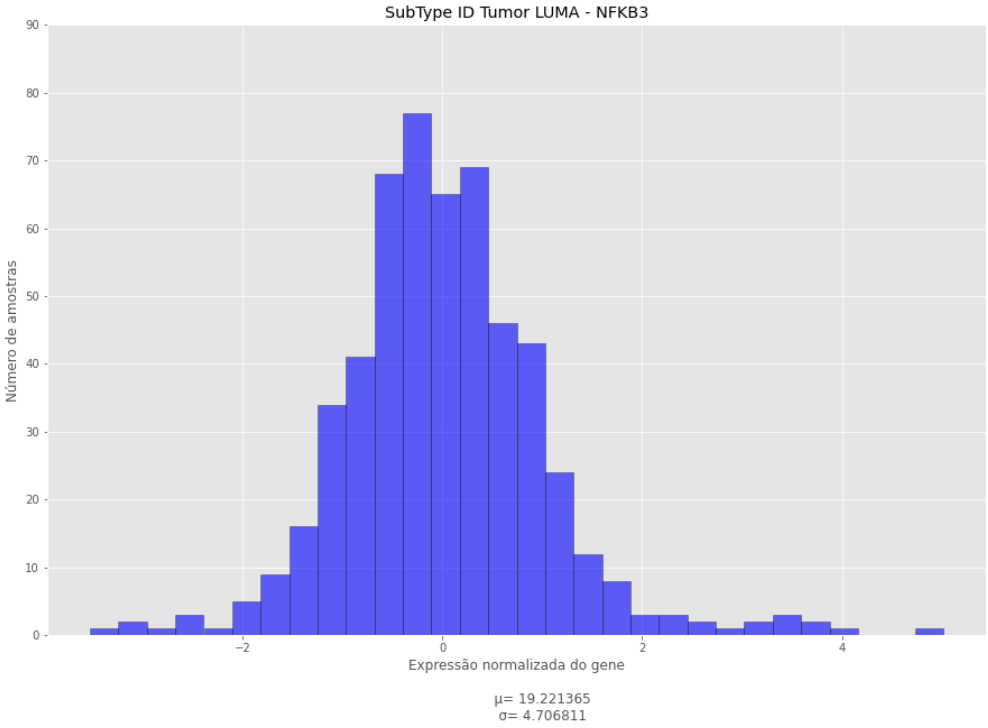

Figura 14: *Gráfico de histograma com a normalização de Gauss da expressão de NF-kB3 do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Número de amostras*** *se refere ao número de amostras dos pacientes distribuídas entre as colunas em relação aos pontos do eixo horizontal, e o eixo na horizontal* ***Expressão normalizada do gene*** *se refere a uma escala de valores normalizados da expressão gênica das amostras.* **μ***: média entre os níveis de expressão de gene das amostras.* **σ**: desvio padrão.

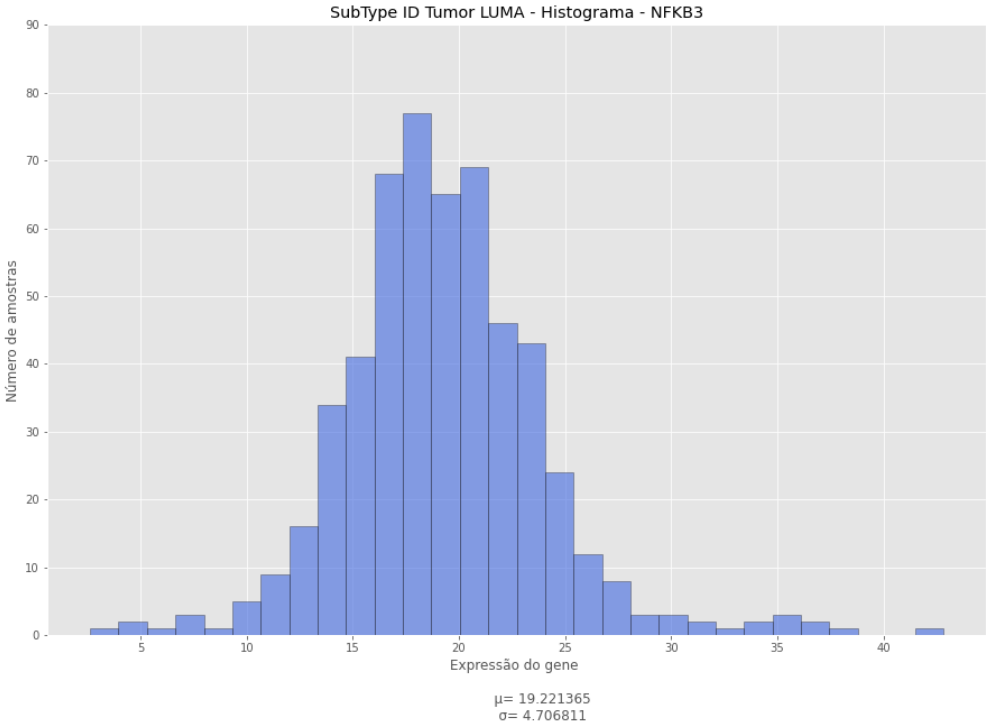

Figura 15: *Gráfico de histograma da expressão de NF-kB3 do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Número de amostras*** *se refere ao número de amostras de pacientes distribuídas entre as colunas em relação aos pontos do eixo horizontal, e o eixo horizontal* ***Expressão do gene*** *se refere a uma escala de valores da expressão gênica do grupo de amostras.* ***μ****: média entre os níveis de expressão de gene das amostras.* **σ**: desvio padrão.

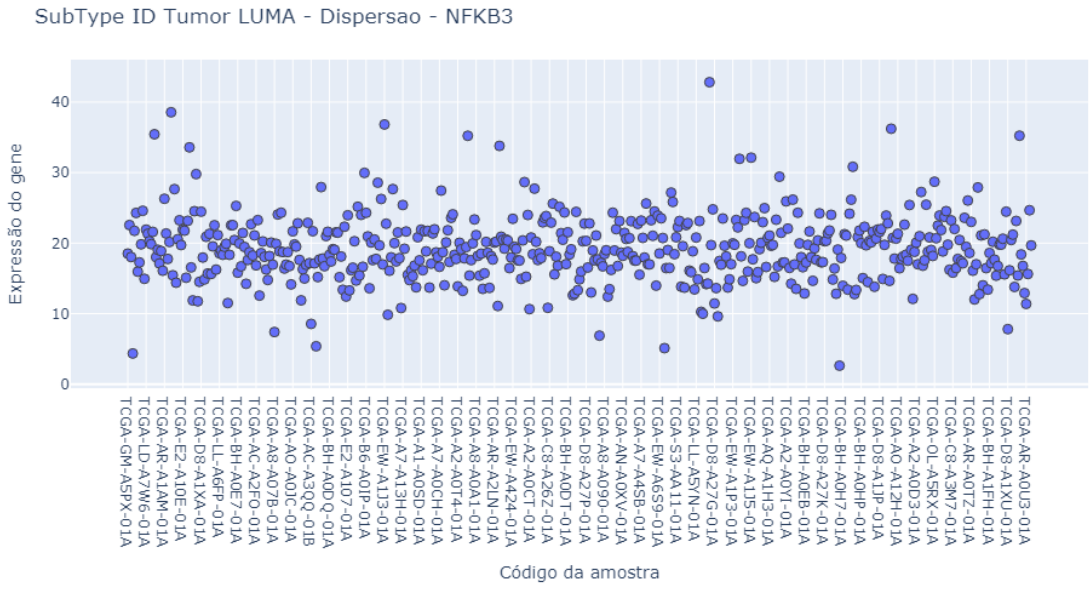

Figura 16: *Gráfico de dispersão da expressão de NF-kB3 do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Expressão do gene*** *se refere aos valores de expressão de gene das amostras, e o eixo* ***Código da amostra*** *se refere ao código das amostras dos pacientes desse grupo.*

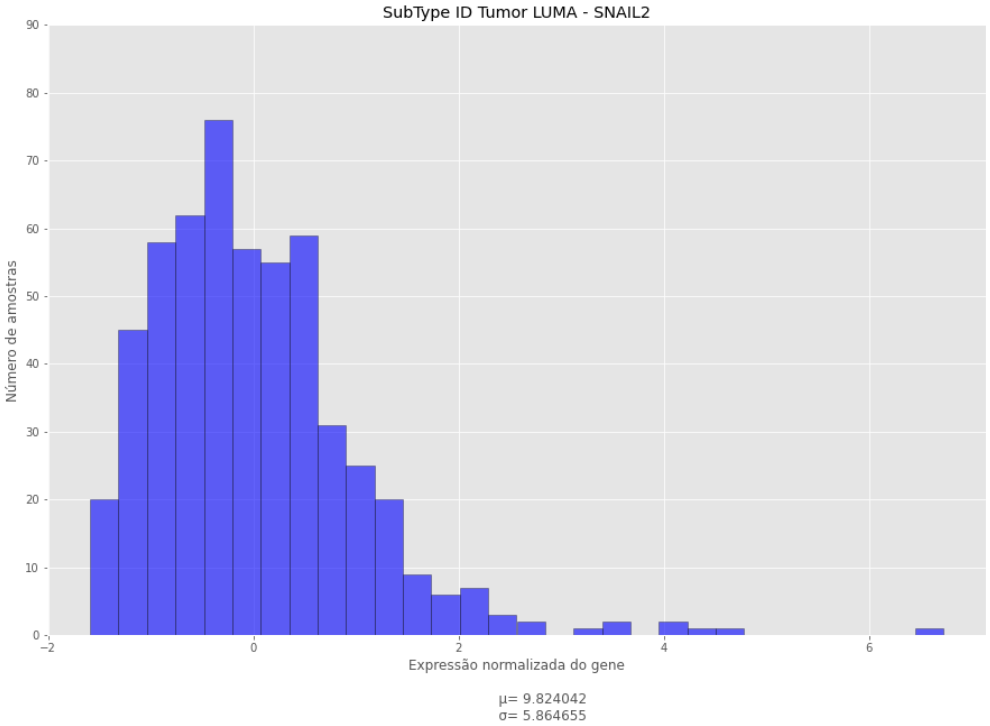

Figura 17: *Gráfico de histograma com a normalização de Gauss da expressão de SNAIL2 (SLUG) do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Número de amostras*** *se refere ao número de amostras dos pacientes distribuídas entre as colunas em relação aos pontos do eixo horizontal, e o eixo na horizontal* ***Expressão normalizada do gene*** *se refere a uma escala de valores normalizados da expressão gênica das amostras.* ***μ****: média entre os níveis de expressão de gene das amostras.* **σ**: desvio padrão.

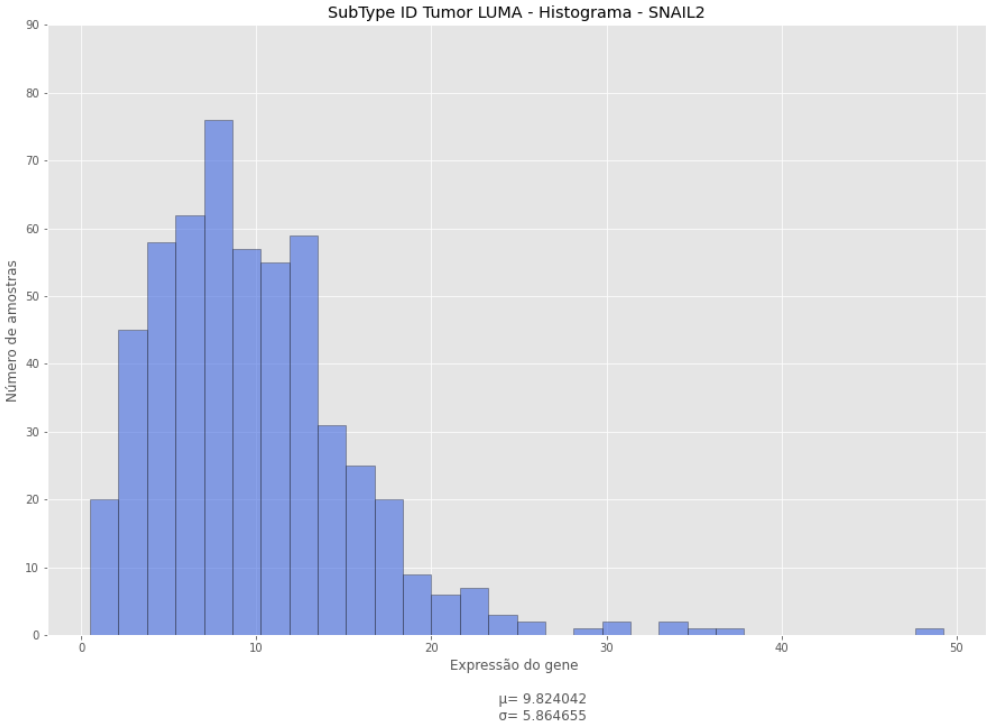

Figura 18: *Gráfico de histograma da expressão de SNAIL2 (SLUG) do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Número de amostras*** *se refere ao número de amostras de pacientes distribuídas entre as colunas em relação aos pontos do eixo horizontal, e o eixo horizontal* ***Expressão do gene*** *se refere a uma escala de valores da expressão gênica do grupo de amostras.* ***μ****: média entre os níveis de expressão de gene das amostras.* **σ**: desvio padrão.

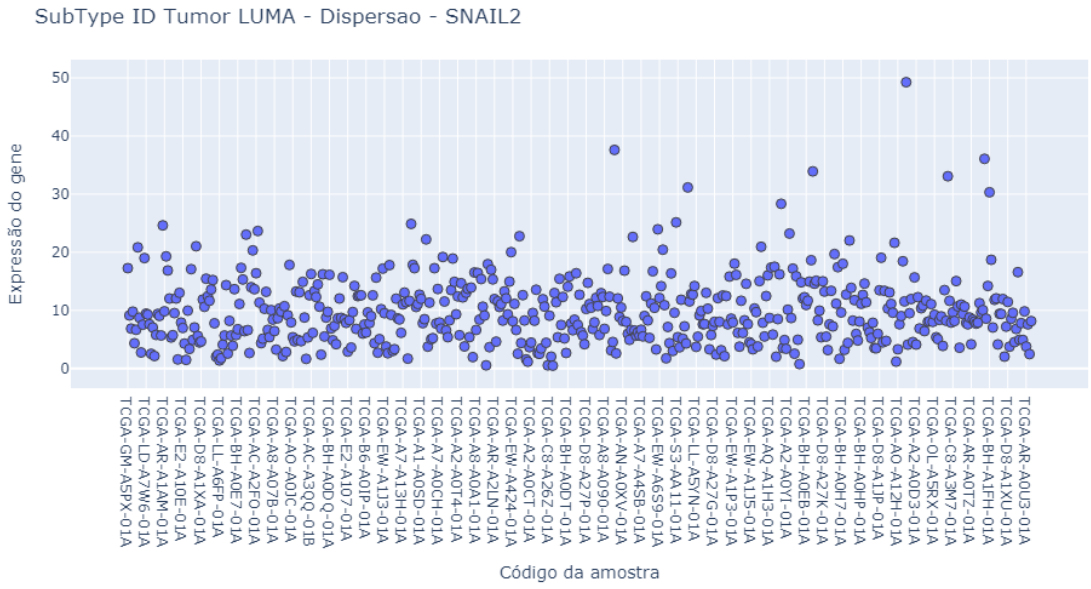

Figura 19: *Gráfico de dispersão da expressão de SNAIL2 (SLUG) do grupo de amostras de tumor do subtipo de câncer de mama luminal A. O eixo* ***Expressão do gene*** *se refere aos valores de expressão de gene das amostras, e o eixo* ***Código da amostra*** *se refere ao código das amostras dos pacientes desse grupo.*

4.3 Análises preliminares dos resultados

Como primeiras análises iremos verificar o estado dos pacientes dos quais as amostras apresentaram níveis de expressão gênica para um ou um conjunto de genes dos que nos interessa e que podem tem correlação com o trabalho de Pires e colaboradores (2017), selecionaremos as amostras com o nível de expressão gênica acima da metade do valor do desvio padrão (σ) no gráfico de histograma normalizado, se não, o valores mais próximos da metade do valor do desvio padrão. Escolheremos essa faixa de valor para tentar abranger os possíveis casos anormais de expressão gênica. Primeiro, vamos identificar as colunas no gráfico de histograma com a normalização de Gauss que cumprem esse requisito citado, em seguida iremos verificar qual o intervalo de valor no gráfico de histograma não normalizado onde as colunas se encontram, e por fim verificaremos no gráfico de dispersão (de preferência no dinâmico) quais os códigos ou o código de amostra estão no intervalo de valor de expressão gênica determinada. Sendo as amostras identificadas, iremos verificar à quais pacientes pertencem, assim, por meio dos dados da plataforma *Genomic Data Commons* (GDC) (https://portal.gdc.cancer.gov/) verificaremos o estado dos pacientes. Nesse procedimento, correlacionaremos as amostras selecionadas dentre os diferentes genes expressos de interesse, para avaliar se essas amostras são as mesmas, e observaremos a correlação das expressões gênicas entre si. Também calcularemos a variância (Var) para verificar se as expressões anormais têm diferença significativa em relação as expressões de gene das demais amostras em um determinado grupo. O cálculo da variância é descrito na **seção 4.1**.

Abaixo se encontram os gráficos estáticos dos grupos nomeados de *SubType ID Tumor LUMA – NFKB3* e *SubType ID Tumor LUMA – SNAIL2*, como exemplo e se baseando nos passos do parágrafo anterior discutiremos esses dados. Para facilitar a discussão iremos chamar o grupo *SubType ID Tumor LUMA – NFKB3* de A e o grupo *SubType ID Tumor LUMA – SNAIL2* de B. Os grupos A e B se referem ao mesmo grupo de amostras, mas estão relacionados com expressões de genes diferentes, assim, ambos possuem 543 amostras. Como evidenciado pela **Figura 14** da **seção 4.2**, um gráfico de histograma com a normalização, a maioria das amostras do grupo A está concentrada no intervalo de [-2,2] no eixo **Expressão normalizada do gene** do gráfico, ou segundo a **Figura 15** da **seção 4.2**, um gráfico de histograma sem a normalização, a maioria dos amostras teve expressão para *NF-kB3* no intervalo de [10,30], sendo a média entre as expressões das amostras do grupo A μ = 19,221365 e o desvio padrão σ = 4,706811. Como descrito no parágrafo anterior, da **Figura 14** selecionaremos as colunas de amostras que tiverem expressão acima da metade do desvio padrão que nesse caso seria

aproximadamente 2, o que irá corresponder a 8 colunas de amostras selecionadas. Então no gráfico da **Figura 15** iremos identificar qual o intervalo de expressão gênica dessas colunas para o *NF-kB3*, o intervalo identificado seria [30,50). No gráfico de dispersão estático da **Figura 16** da **seção 4.2** e no de dispersão dinâmico iremos identificar quais amostras tiveram expressão para *NF-kB*3 dentro do intervalo selecionado, são no total 15 amostras que tiveram essa expressão. Em seguida, verificaremos qual a expressão de *SNAIL2* dessas 15 amostras por meio dos gráficos do grupo B que se encontram na **seção 4.2**, e observaremos qual a situação dos pacientes que doaram as amostras. Também faremos o caminho inverso que será a seleção de amostras a partir dos gráficos do grupo B. Fazendo o caminho a partir do grupo B, seriam aproximadamente 15 amostras selecionadas também, com μ = 9,824042 e o desvio padrão σ = 5,864655. Por fim, iremos verificar se houve diferença significativa na expressão gênica das amostras selecionas em relação aos respectivos grupos através da variância.

A influência da expressão do *NF-kB/p65* sobre a expressão do *SNAIL2 (SLUG)*, assim como sobre a do *TWIST1* e do *SIP1* já foi demostrada pelo trabalho de Pires e colaboradores (2017), com as evidências de diferentes experimentos os quais comprovam que a inibição de *NF-kB3* leva a baixa regulação de outros três genes abordados no estudo deles. No caso do *NF-kB/p65* em relação ao *SNAIL2*, sua expressão não inibida ao longo do desenvolvimento do câncer de mama leva à uma alta regulação de *SNAIL2* que é um repressor de E-caderinas como mencionado na **seção 1.4**, a repressão de E-caderinas contribui para que as células do tumor da mama percam as características epiteliais, adquiram características mesenquimais e desenvolvam um comportamento migratório, o que leva a metástase, ou seja, o que se esperaria nessa análise preliminar é que as amostras de paciente que tiveram maior expressão de *NF-kB3* dentro do intervalo selecionado, tenham tido uma expressão mais elevada de *SNAIL2* em relação à média de expressão gênica entre as amostras do grupo B, e que pelo menos parte significativa dos casos clínicos de pacientes fossem com metástase do câncer de mama. O que ainda faremos para este projeto, a princípio, será verificar qual a correlação do *NF-kB3* com os 18 genes de interesse (incluindo o *SNAIL2*, *TWIST1* e *SIP1*), e a correlação das expressões dos 19 genes entre si, e ainda o estado dos pacientes onde se tem expressões gênicas incomuns ou acima de um intervalo como descrito nos parágrafos anteriores.

5 CONCLUSÃO

5.1 Conclusões do trabalho

A ferramenta foi criada nos levando a aprimorar os conhecimentos a respeito da linguagem `Python`, já que tivemos que aplicar métodos e conceitos da linguagem que vão além do curso de graduação de Ciências Biológicas: Biotecnologia. Como o estudo, uso e aplicações das ferramentas do `pandas`, `matplotlib` e `plotly`. O algoritmo da ferramenta se encontra no **APÊNDICE**.

Como primeira etapa da análise dos dados, coletamos os arquivos com os níveis de expressão dos genes das amostras dos pacientes, dentre eles a expressão gênica de *NF-kB3* que desempenha um papel fundamental na transição epitélio mesenquimal ou *ETM*. Utilizando-se dos dados coletados, o algoritmo foi capaz de fazer a leitura automática dos arquivos, a identificação dos dados se utilizando do método de busca sequencial e de gerar os gráficos para a análise comparativa dos níveis de expressão gênica de acordo com o subtipo de câncer de mama no qual é classificado o paciente.

A ferramenta foi capaz de produzir gráficos em formato `PNG` e `HTML` que se mostraram viáveis para as futuras análises, o que foi demonstrado com as análises preliminares na **seção 4.3**, sendo gráficos de histograma sem e com a normalização de Gauss e gráficos de dispersão estáticos e dinâmicos. A ferramenta também gerou os valores de média e desvio padrão de cada grupo para cada um dos 19 genes diferentes, o que irá contribuir com o trabalho de Pires e colaboradores (2017) em colaboração com a Professora e Doutora Eliana Abdelhay do INCA.

5.2 Perspectivas futuras

Diversas etapas serão realizadas a partir de agora. Dentre elas, uma análise de variância para verificar quais grupos ou amostras dentro de cada grupo mostram diferenças significativas nos níveis de expressão dos genes de interesse.

Futuramente, serão feitas análises estatísticas, como o cálculo de frequência da ocorrência de valores de expressão gênica dentro de intervalos determinados. Também serão avaliadas as amostras acima de uma faixa de valor em relação ao desvio padrão para cada grupo de cada um dos genes de relevância. Essa análise será feita para se verificar a relação da

expressão de cada um dos genes destacados, principalmente a correlação deles com a expressão do NF-kB.

A ferramenta será refinada e melhorada para que tenha uma interação mais eficiente com o usuário, e será modificada para que sejam produzidos outros dados relevantes para as análises relacionadas ao câncer de mama. Pretendemos implementar na ferramenta técnicas de aprendizado de máquina e reconhecimento de padrões, como técnicas de caracterização e classificação de dados.

REFERÊNCIAS

Anaconda Software Distribution. Computer software. Vers. 2-2.4.0. Anaconda, Jan. 2021. Web. <https://anaconda.com>.

AZIM, Hamdy A.; GHOSN, Marwan; OUALLA, Karima; KASSEM, Loay. Personalized treatment in metastatic triple-negative breast cancer: the outlook in 2020. **The Breast Journal**, Giza, Egypt, v. 26, n. 1, p. 69-80, 23 dez. 2019. Wiley. http://dx.doi.org/10.1111/tbj.13713.

BRAY, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA**: a cancer journal for clinicians, Hoboken, v. 68, n. 6, p. 394-424, Nov. 2018.

CHAVEZ, Kathryn J.; GARIMELLA, Sireesha V.; LIPKOWITZ, Stanley. Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. **Breast Disease**, [S.L.], v. 32, n. 1-2, p. 35-48, 15 mar. 2011. IOS Press. http://dx.doi.org/10.3233/bd-2010-0307.

COSTA-SILVA, Juliana; DOMINGUES, Douglas; LOPES, Fabricio Martins. RNA-Seq differential expression analysis: an extended review and a software tool. **Plos One**, Cornélio Procópio, Pr, Brazil, v. 12, n. 12, p. 1-18, 21 dez. 2017. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0190152.

DANA, Hassan; CHALBATANI, Ghanbar Mahmoodi; MAHMOODZADEH, Habibollah; KARIMLOO, Rezvan; REZAIEAN, Omid; MORADZADEH, Amirreza; MEHMANDOOST, Narges; MOAZZEN, Fateme; MAZRAEH, Ali; MARMARI, Vahid. Molecular Mechanisms and Biological Functions of siRNA. **Int J Biomed Sci**, [*s. l*], v. 2, n. 13, p. 48-57, jun. 2017. PMCID: PMC5542916.

FERLAY, J. et al. (ed.). **Cancer today**. Lyon, France: International Agency for Research on Cancer, 2018. (IARC CAncerBase, n. 15). Available at: https://publications.iarc.fr/Databases/Iarc-Cancerbases/Cancer-Today-Powered-By-GLOBOCAN-2018--2018. Access in: 9 Sep. 2019.

Genomic Data Commons Data Portal. GDC, Jan. 2021. Web <https://portal.gdc.cancer.gov/>.

GIROUX, Veronique; RUSTGI, Anil K.. Metaplasia: tissue injury adaptation and a precursor to the dysplasia‐cancer sequence. **Nature Reviews Cancer**, Philadelphia, v. 17, n. 10, p. 594-604, 1 set. 2017. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrc.2017.68.

HAN, Yixing; GAO, Shouguo; MUEGGE, Kathrin; ZHANG, Wei; ZHOU, Bing. Advanced Applications of RNA Sequencing and Challenges. **Bioinformatics And Biology Insights**, [S.L.], v. 91, p. 29-46, jan. 2015. SAGE Publications. http://dx.doi.org/10.4137/bbi.s28991.

HU, Bo; ZHONG, Liping; WENG, Yuhua; PENG, Ling; HUANG, Yuanyu; ZHAO, Yongxiang; LIANG, Xing-Jie. Therapeutic siRNA: state of the art. **Signal Transduction And Targeted Therapy**, [S.L.], v. 5, n. 1, p. 101-101, 19 jun. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41392-020-0207-x.

HU, Yue; ZHENG, Yayuan; DAI, Mingrui; WU, Jiaxin; YU, Bin; ZHANG, Haihong; KONG, Wei; WU, Hui; YU, Xianghui. Snail2 induced E-cadherin suppression and metastasis in lung carcinoma facilitated by G9a and HDACs. **Cell Adhesion & Migration**, [S.L.], v. 13, n. 1, p. 284-291, 1 jan. 2019. Informa UK Limited. http://dx.doi.org/10.1080/19336918.2019.1638689.

HUNTINGTON MEDICINA REPRODUTIVA (Brasil). **NGS – Sequenciamento de Nova Geração**. 2021. Disponível em: https://www.huntington.com.br/tratamentos/tecnicas-complementares/ngs-next-generation-sequencing/. Acesso em: 01 maio 2021.

JENSEN, Mark A.; FERRETTI, Vincent; GROSSMAN, Robert L.; STAUDT, Louis M.. The NCI Genomic Data Commons as an engine for precision medicine. **Blood**, [S.L.], v. 130, n. 4, p. 453-459, 27 jul. 2017. American Society of Hematology. http://dx.doi.org/10.1182/blood-2017-03-735654.

LAUTRUP, Marianne D.; THORUP, Signe S.; JENSEN, Vibeke; BOKMAND, Susanne; HAUGAARD, Karen; HOEJRIS, Inger; JYLLING, Anne-Marie B.; JOERNSGAARD, Hjoerdis; LELKAITIS, Giedrius; OLDENBURG, Mette H.. Male breast cancer: a nation-wide population-based comparison with female breast cancer. **Acta Oncologica**, [Vejle], v. 57, n. 5, p. 613-621, 23 dez. 2017. Informa UK Limited. http://dx.doi.org/10.1080/0284186x.2017.1418088.

LUCA, Francesco de. Regulatory role of NF-κB in growth plate chondrogenesis and its functional interaction with Growth Hormone. **Molecular And Cellular Endocrinology**, Kansas, United States, v. 514, p. 1-22, ago. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.mce.2020.110916.

Matplotlib. Computer software. Vers. 3.4.1. Marplotlib, Nov. 2020. Web <https://matplotlib.org/>.

MIRBAGHERI, Esmat; AHMADI, Maryam; SALMANIAN, Soraya. Common data elements of breast cancer for research databases: a systematic review. **Journal Of Family Medicine And Primary Care.** Tehran, p. 1296-1301. mar. 2020.

OZSOLAK, Fatih; MILOS, Patrice M.. RNA sequencing: advances, challenges and opportunities. **Nature Reviews Genetics**, [S.L.], v. 12, n. 2, p. 87-98, 30 dez. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg2934.

PIRES, Bruno; SILVA, Rafael; FERREIRA, Gerson; ABDELHAY, Eliana. NF-kappaB: two sides of the same coin. **Genes**, Rio de Janeiro, Rj, Brazil, v. 9, n. 1, p. 1-24, 9 jan. 2018. MDPI AG. http://dx.doi.org/10.3390/genes9010024.

PIRES, Bruno R. B.; MENCALHA, Andre L.; FERREIRA, Gerson M.; SOUZA, Waldemir F. de; MORGADO-DÍAZ, José A.; MAIA, Amanda M.; CORRêA, Stephany; ABDELHAY, Eliana S. F. W.. NF-kappaB Is Involved in the Regulation of EMT Genes in Breast Cancer Cells. **Plos One**, Rio de Janeiro, Rj, Brazil, v. 12, n. 1, p. 1-20, 20 jan. 2017. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0169622.

Pandas. Computer software. Vers. 1.1.4. Pandas, Nov. 2020. Web <https://www.python.org/>.

Plotly. Computer software. Plotly, Nov. 2020. Web <https://plotly.com/python/>.

Python Software Foundation. Computer software. Vers. 3.8.3. Python, Nov. 2020. Web <https://www.python.org/>.

SILVA, Jesse Lopes da; NUNES, Natalia Cristina Cardoso; IZETTI, Patricia; MESQUITA, Guilherme Gomes de; MELO, Andreia Cristina de. Triple negative breast cancer: a thorough review of biomarkers. **Critical Reviews In Oncology/hematology**, Brazilian National Cancer Institute (Inca), Brazil, v. 145, p. 102855-102855, jan. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.critrevonc.2019.102855.

TESTA, Ugo; CASTELLI, Germana; PELOSI, Elvira. Breast Cancer: a molecularly heterogenous disease needing subtype-specific treatments. **Medical Sciences**, Rome, Italy, v. 8, n. 1, p. 1-103, 23 mar. 2020. MDPI AG. http://dx.doi.org/10.3390/medsci8010018.

WANG, Zhong; GERSTEIN, Mark; SNYDER, Michael. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, [S.L.], v. 10, n. 1, p. 57-63, jan. 2009. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg2484.

WEIGELT, Britta; GEYER, Felipe C.; REIS-FILHO, Jorge S.. Histological types of breast cancer: how special are they?. **Molecular Oncology**, London, v. 4, n. 3, p. 192-208, 18 abr. 2010. Wiley. http://dx.doi.org/10.1016/j.molonc.2010.04.004.

ZELLI, Veronica; COMPAGNONI, Chiara; CANNITA, Katia; CAPELLI, Roberta; CAPALBO, Carlo; NOLFI, Mauro di Vito; ALESSE, Edoardo; ZAZZERONI, Francesca; TESSITORE, Alessandra. Applications of Next Generation Sequencing to the Analysis of Familial Breast/Ovarian Cancer. **High-Throughput**, Rome, Italy, v. 9, n. 1, p. 1-16, 10 jan. 2020. MDPI AG. http://dx.doi.org/10.3390/ht9010001.

APÊNDICE

#Projeto expressão gênica

```
import os
from os import mkdir
import zipfile
import gzip
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px

def Busca (dataframe,coluna,gene):
    novo_gene=None
    tamanho=len(dataframe)
    for i in range(tamanho):
        encontra=dataframe[coluna][i].find(gene)
        if(encontra!=-1):
            novo_gene=dataframe[coluna][i]
    return novo_gene

class Projeto_Laboratorio:
    def __init__(self,diretorio,subtipos,amostras):
        self.diretorio=diretorio
        self.subtipos=subtipos
        self.amostra=amostras
        self.amostras=None
        self.paciente=""
```

```
    self.lista=[]
    self.lista_paciente=[]
    self.dicionario=0
    self.lista_subtipo=[]
    self.gene=""

def Extrai_zip (self):
    for subdir,dirs,files in os.walk(self.diretorio):
        for file in files:
            if(file[-3:]=="zip"):
                arq=zipfile.ZipFile(self.diretorio+"/"+self.arquivo)
                arq.extractall(self.diretorio)
                arq.close()

def Extrair_gz (self):
    for subdir,dirs,files in os.walk(self.diretorio):
        tam=len(files)
        for file in files:
            #print(file)
            if(tam>1)and(file[-8:]=="FPKM.txt"):
                self.lista.append(subdir+"/"+file)
                self.lista_paciente.append(file)
            if(tam==1)and(file[-2:]=="gz"):
                arq=gzip.open(subdir+"/"+file,"rb")
                ler=arq.read()
                arq.close()
                arq2=open(subdir+"/"+file[:-3],"wb")
                arq2.write(ler)
                arq2.close()
```

```
                self.lista.append(subdir+"/"+file)
                self.lista_paciente.append(file)
        self.dicionario=dict(zip(self.lista_paciente,self.lista)
        return self.dicionario

    def Dados (self):
        self.paciente=input("Digite o código do paciente:")
        self.paciente=self.paciente+".txt"
        dados=pd.read_csv(self.dicionario[self.paciente],sep="\t",header=None)
        dados.columns=["Codigo do gene","Expressao do gene"]
        dataf=pd.DataFrame(dados)
        dataf.describe()
        opcao=input("""Digite o código do gene desejado para se visualizar sua expressão, ou "sair" para encerrar a busca:""")

        while(opcao!="sair"):

            busca=Busca(dataf,"Codigo do gene",opcao)

            expressao_gene=dataf.loc[dataf["Codigo do gene"]==busca]
            print(expressao_gene)

            opcao=input("""Digite o código do gene desejado para se visualizar sua expressão, ou "sair" para encerrar a busca:""")

    def Graficos (self,gene=None,nome=None):
        self.lista_subtipo=[]
        if(gene!=None):
            self.gene=gene
        else:
```

```
    self.gene=input("Digite o código da molécula que será analisada:")


for subdir,dirs,files in os.walk(self.subtipos):
    for file in files:
        if(file[-4:]=="xlsx"):
            self.lista_subtipo.append(file)
#print(self.lista_subtipo)
amostra=pd.read_table(self.amostra)
self.amostras=pd.DataFrame(amostra)
#print(amostras)
tamanho=len(self.amostras)
for i in self.lista_subtipo:
    subtipo=pd.read_excel(self.subtipos+"/"+i)
    subtipos=pd.DataFrame(subtipo)
    #print(subtipos)
    dataframe=pd.DataFrame()#columns=[gene]
    dataframes=pd.DataFrame()
    nv_diretorio="C:/Users/Fabio/Pictures/Imagens_Projeto_Laboratorio/Teste/"
    sinal1=False
    sinal2=False
    if(nome!=None):
        sinal2=os.path.isdir(nv_diretorio+nome)
        nv_diretorio=nv_diretorio+nome
    else:
        sinal1=os.path.isdir(nv_diretorio+self.gene)
        nv_diretorio=nv_diretorio+self.gene


    if(nome==None)and(sinal1==False):
        os.mkdir(nv_diretorio)
```

```
elif(nome!=None)and(sinal2==False):
    os.mkdir(nv_diretorio)

for j in range(tamanho):
    tipo=subtipos.loc[subtipos["Sample ID"]==self.amostras["Sample ID"][j]]

    tamanho_tipo=len(tipo)
    if(tamanho_tipo==1)and((self.amostras["File Name"][j][:-3] in self.dicionario)==True):
        #print(tipo)

        dado=pd.read_csv(self.dicionario[self.amostras["File Name"][j][:-3]],sep="\t",header=None)
        dado.columns=["Codigo do gene","Expressao do gene"]
        df=pd.DataFrame(dado)

        busca=Busca(df,"Codigo do gene",self.gene)
        expressao=df.loc[dado["Codigo do gene"]==busca]

        nome_grupo=i[:-5]#+" - "+busca
        nome_novo=nome_grupo.replace("_"," ")
        dtf=pd.DataFrame(expressao["Expressao do gene"])
        dtf.columns=["Expressao do gene"]
        dataframe=dataframe.append(dtf)

        dtfs=pd.DataFrame(data=[self.amostras["Sample ID"][j]],columns=["Codigo do paciente"])
        dataframes=dataframes.append(dtfs)

local=nv_diretorio+"/"+nome_grupo
sinal3=os.path.isdir(local)
if(sinal3==False):
```

```
    os.mkdir(local)


x=dataframes.to_dict("list")
#print(x)
y=dataframe.to_dict("list")
#print(y)
x.update(y)
#print(x)


novo_df=pd.DataFrame(x)


print(" ")
print(nome_novo)
print(" ")


descricao=dataframe.describe()
print(descricao)


media=descricao["Expressao do gene"]["mean"]
sigma=descricao["Expressao do gene"]["std"]


plt.style.use("ggplot")
plt.figure(figsize=(15,10))
plt.hist(y["Expressao do gene"],bins=30,ec="k",alpha=.6,color="royalblue")
if(nome==None):
    plt.title(nome_novo+" - Histograma - "+self.gene)
else:
    plt.title(nome_novo+" - Histograma - "+nome)
plt.xlabel("""Expressão do gene
```

```
μ= %f
σ= %f"""%(media,sigma))
plt.ylabel("Número de amostras")
plt.ylim(0,90)
plt.savefig(local+"/"+nome_grupo+" - Histograma.png",format="png")
plt.show()
plt.close()

gaussiana=[]
for e in y["Expressao do gene"]:
    z=(e-media)/sigma
    gaussiana.append(z)

plt.style.use("ggplot")
plt.figure(figsize=(15,10))
plt.hist(gaussiana,bins=30,ec="k",alpha=.6,color="blue")
if(nome==None):
    plt.title(nome_novo+" - "+self.gene)
else:
    plt.title(nome_novo+" - "+nome)
plt.xlabel("""Expressão normalizada do gene

μ= %f
σ= %f"""%(media,sigma))
plt.ylabel("Número de amostras")
plt.ylim(0,90)
plt.savefig(local+"/"+nome_grupo+" - Histograma (Gaussiana padronizada).png",format="png")
plt.show()
```

```
        plt.close()


        fig=px.scatter(novo_df,x="Codigo do paciente",y="Expressao do gene",log_x=False,width=1000)
        fig.update_traces(marker=dict(size=8,line=dict(width=1)),selector=dict(mode="markers"))
        if(nome==None):
            fig.update_layout(title=nome_novo+" - Dispersao - "+self.gene)
        else:
            fig.update_layout(title=nome_novo+" - Dispersao - "+nome)
        fig.update_xaxes(title="Código da amostra")
        fig.update_yaxes(title="Expressão do gene")
        fig.show()


        if not os.path.exists("C:/Users/Fabio/Pictures"):
            os.mkdir("C:/Users/Fabio/Pictures")
        fig.write_image(local+"/"+nome_grupo+" - Dispersao.png")
        fig.write_html(local+"/"+nome_grupo+" - Dispersao.html")
```

#Programa principal

```
diretorio="C:/Users/Fabio/Documents/Estudo_Projeto_de_Laboratório/Projeto/Novo/Nova/Data/1.mRNA/gdc_download"
subtipos="C:/Users/Fabio/Documents/Estudo_Projeto_de_Laboratório/Projeto/Novo/Nova/Subtipos_de_Tumor"
amostras="C:/Users/Fabio/Documents/Estudo_Projeto_de_Laboratório/Projeto/Novo/Nova/gdc_sample_sheet.2019-07-24.tsv"

projeto=Projeto_Laboratorio(diretorio,subtipos,amostras)
projeto.Extrai_zip()
projeto.Extrair_gz()

opcao=input("""
```

```
Digite umas das opções abaixo:


1 - Se deseja produzir os gráficos de um único gene;

2 - Se deseja produzir gráficos de um conjunto de genes;

3 - Se só deseja visualizar os genes ou um gene de um paciente especifico.


Opcão:

""")


if(opcao=="3"):
    projeto.Dados()
elif(opcao=="1"):
    projeto.Graficos()
elif(opcao=="2"):

interesse=pd.read_excel("C:/Users/Fabio/Documents/Estudo_Projeto_de_Laboratório/Projeto/Novo/Nova/Lista_
de_genes_de_interesse.xlsx")
    interes=pd.DataFrame(interesse)
    tam=len(interes)
    for i in range(tam):
        nome=interes["Gene Synonyms"][i].split(",")
        projeto.Graficos(interes["Ensembl"][i],nome[0])
```