UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE BIOTECNOLOGIA
CURSO DE GRADUAÇÃO EM BIOTECNOLOGIA

RONIEL VÍCTOR ALVES SOARES

ANÁLISE E APLICAÇÃO DA FERRAMENTA DE BIOINFORMÁTICA GEOCANCERPROGNOSTICDATASETSRETRIEVER EM DADOS GENÔMICOS GENE EXPRESSION OMNIBUS (GEO) PARA PROGNÓSTICOS DE CÂNCER

PATOS DE MINAS – MG
NOVEMBRO DE 2023

RONIEL VÍCTOR ALVES SOARES

ANÁLISE E APLICAÇÃO DA FERRAMENTA DE BIOINFORMÁTICA GEOCANCERPROGNOSTICDATASETSRETRIEVER EM DADOS GENÔMICOS GENE EXPRESSION OMNIBUS (GEO) PARA PROGNÓSTICOS DE CÂNCER

Monografia apresentada ao Instituto de Biotecnologia da Universidade Federal de Uberlândia como requisito final para a obtenção do título de Bacharel em Biotecnologia.

Prof. Dr. Laurence Rodrigues do Amaral

PATOS DE MINAS – MG

NOVEMBRO DE 2023

RONIEL VÍCTOR ALVES SOARES

Análise E Aplicação Da Ferramenta De Bioinformática Geocancerprognosticdatasetsretriever Em Dados Genômicos Gene Expression Omnibus (Geo) Para Prognósticos De Câncer

Monografia apresentada ao Instituto de Biotecnologia da Universidade Federal de Uberlândia como requisito final para a obtenção do título de Bacharel em Biotecnologia.

Banca Examinadora:

_____

Prof. Dr. Laurence Rodrigues do Amaral - UFU

Presidente

_____

MSc. Tamires Caixeta Alves - UFU

Membro

_____

MSc. Carlos Bruno de Araújo - UFU

Membro

Patos de Minas, MG, 23 de novembro de 2023

AGRADECIMENTOS

Agradeço primeiramente a Deus, que deu essa oportunidade de ser um dos únicos da minha família a cursar um ensino superior de qualidade.

Agradeço aos meus pais e meu irmão pelo constante apoio e sempre ter me levantado nos momentos mais difíceis durantes esses anos.

Agradeço aos amigos que ganhei durante esse período dentro da universidade, amigos que vou levar para a vida, amigos aqueles que realmente estão do seu lado quando você precisa e comemoram com suas conquistas.

Agradeço aos professores que passaram suas experiencias e conhecimentos da melhor forma possível, cada um em seu momento marcante nessa minha trajetória. E principalmente ao Prof. Laurence, que depositou sua confiança e me fez crescer profissionalmente, aquele que tenho como referência para vida dentro da universidade e pessoal.

Agradeço especialmente aos amigos que ganhei no estágio, além de colegas de trabalho, pessoas que me fizeram crescer em muitos sentidos e fizeram a diferença, pessoas que vou levar para a vida.

Agradeço também aos colegas que de alguma maneira me ajudou, orientou, deram forças para que hoje eu conquiste um dos meus objetivos.

"O melhor presente Deus me deu, a vida
me ensinou a lutar pelo que é meu"
.: Charlie Brown Jr.

RESUMO

O câncer está presente no cotidiano de diversas pessoas, afetando diretamente nos estilos de vida deles, no qual é necessário a produção de dados e pesquisas na área oncológica. Este trabalho pode ser utilizado como um material de apoio trazendo informações necessárias ou acessórias para o processamento de dados e a aplicação de uma ferramenta de bioinformática impactando na pesquisa oncológica. O objetivo do trabalho foi descrever a aplicação da ferramenta de bioinformática geoCancerPrognosticDatasetsRetriever disponível gratuitamente em linguagem de programação Perl para identificação de conjuntos de dados prognósticos de câncer fornecidos pelo repositório público de dados genômicos Gene Expression Omnibus (GEO). Na ferramenta analisada, pode-se utilizar diferentes comandos para diferentes prognósticos de câncer, com o objetivo de otimizar a busca no repositório. A ferramenta tem como função agrupar os dados de prognósticos de câncer e fornecer de maneira rápida todos os conjuntos de dados publicados no repositório. Sendo assim, o tempo gasto utilizando a ferramenta é significativamente menor quando comparado ao agrupamento manual diretamente no site do repositório GEO. Assim podemos utilizar facilmente os dados prognósticos de expressão gênica agrupados pela ferramenta e fornecidos pelo repositório, oferecendo uma maior usabilidade para o pesquisador, que nem sempre é um especialista em bioinformática.

Palavras-chave: Ferramenta de Bioinformática. Prognósticos. Câncer.

ABSTRACT

Cancer is present in the daily lives of many people, directly affecting their lifestyles, which requires the production of data and research in the oncological area. This work can be used as supporting material providing necessary or accessory information for data processing and the application of a bioinformatics tool impacting oncology research. The objective of the work was to describe the application of the geoCancerPrognosticDatasetsRetriever bioinformatics tool freely available in the Perl programming language to identify cancer prognostic datasets provided by the public genomic data repository Gene Expression Omnibus (GEO). In the analyzed tool, different commands can be used for different cancer prognoses, with the aim of optimizing the search in the repository. The tool's function is to group cancer prognosis data and quickly provide all datasets published in the repository. Therefore, the time spent using the tool is significantly less when compared to manual grouping directly on the GEO repository website. This way we can easily use the prognostic gene expression data grouped by the tool and provided by the repository, offering greater usability for the researcher, who is not always an expert in bioinformatics.

Keywords: Bioinformatics Tool. Prognostic. Cancer.

SUMÁRIO

INTRODUÇÃO

Existem hoje muitos tipos e subtipos de câncer descritos em trabalhos relevantes podendo afetar no estilo de vida dos seres humanos de tal maneira a apresentar sinais ou sintomas prejudicando a sua saúde. As causas podem ser devido a hábitos alimentares, estilos de vida, radiações ou outros agentes carcinogênicos, infecções ou inflamações, questões hereditárias entre outras (WANG et al, 2018). O câncer se descreve basicamente por um crescimento anormal das células com grande potencial para se espalhar para outras partes importantes do corpo humano como por exemplo os órgãos do indivíduo (WANG et al, 2018). Sendo que o câncer ainda é uma das principais causas de morte e a cada ano que se passa essa taxa aumenta, necessitando de tratamentos mais eficazes e conhecimentos específicos para o desenvolvimento desses tratamentos (WANG et al, 2018). Assim, evidenciando um dos objetivos das pesquisas oncológicas em diferentes áreas do conhecimento envolvendo as descobertas, desenvolvimento e aplicação dos tratamentos.

Hoje, grande parte das análises realizadas com dados biológicos são realizadas por ferramentas que foram desenvolvidas para um ambiente web. Em práticas e análises laboratoriais geralmente os pesquisadores fazem e refazem um mesmo processo de maneira mais demorada quando realizadas manualmente, até mesmo tediosas e propensa a erros. Sendo fundamental a criação e desenvolvimento de métodos automatizados e programados para melhorar a eficiência em muitos aspectos (VOS et al, 2011). Um dos problemas encontrados nessas ferramentas está ligada ao baixo volume de dados que podem ser processados ao mesmo tempo, um limitado alcance de análises, coleta de dados e principalmente o considerável tempo utilizado na tarefa, mesmo com a disponibilidade de banco de dados. Dessa forma, é importante o desenvolvimento de *scripts* para que um maior número de análises possa ser realizado de modo automatizado, aumentando também sua robustez e precisão, otimizando o tempo.

Com o desenvolvimento da linguagem de programação, destaca-se a linguagem de programação Perl onde originalmente foi utilizada para facilitar a manipulação de textos, delegando tarefas em um sistema web, redes, programas dentro de um sistema operacional ou plataforma entre outros (https://perldoc.perl.org/perlintro). A linguagem de programação Perl é usada em diversos recursos com mais de 30 anos de desenvolvimento e sendo capaz de ser executada em mais de 100 plataformas, adequada para processos rápidos ou em larga escala (https://www.perl.org/about.html). Essa linguagem tem o objetivo de automatizar processamentos de dados de forma eficiente, facilitada e de maneira completa

(https://perldoc.perl.org/perlintro). Resultando em um processo em que o seu valor pode ser agregado a outras áreas de conhecimento sendo usada como ferramenta de amplo alcance, como podemos observar nas aplicações em análises biológicas tradicionais, tais como as realizadas de forma manualmente em bancada, análises de dados e apresentação de resultados.

Diante de um objetivo no qual a busca de informações robustas que comprovem as descobertas feitas por pesquisadores e usadas como informações acessórias. É fundamental e de extrema importância de que essas informações cheguem até eles de maneira rápida, reduzindo o tempo de bancada e de uma pesquisa em andamento, assim o *Gene Expression Omnibus* (GEO) vem com esse propósito.

GEO é um repositório público de maneira funcional no qual fornecem diversas informações de dados genômicos e prognósticos de trabalhos disponíveis gratuitamente online e de forma atualizada, contando com um grande conjunto de dados disponíveis (CLOUGH, BARRETT, 2016). Dentre um dos objetivos do repositório GEO destaca-se o fornecimento de mecanismos a serem usados por pesquisadores para a consulta de expressões gênicas no qual possuem interesse. Porém devido ao grande volume de conjuntos de dados disponíveis, a pesquisa e o tratamento desses dados se tornam uma tarefa que demanda tempo e muitas das vezes tempo precioso no qual o pesquisador não pode dedicar, principalmente quando se trata de pesquisas oncológicas. Para que exista uma maneira eficiente de prover a procura por conjuntos de dados de expressão gênica de prognósticos dos tipos diferentes de câncer disponíveis no repositório GEO, foi desenvolvido o software geoCancerPrognosticDatasetsRetriever (ALAMEER, CHICCO, 2022). O software foi escrito utilizando a linguagem de programação Perl, por Abbas Alameer (*Kuwait University*) em colaboração com Davide Chicco (*University of Toronto*) em 2021. O software está disponível gratuitamente no repositório de módulos desenvolvidos utilizando a linguagem de programação Perl, CPAN que em outubro de 2023 possuía 214.026 módulos em linguagem de programação Perl, em 44.488 distribuições, escritos por 14.425 autores distribuídos em um único servidor (https://www.cpan.org/) em pesquisa via MetaCPAN (https://metacpan.org/).

O software geoCancerPrognosticDatasetsRetriever é uma ferramenta de Bioinformática para a recuperação dos conjuntos de dados de prognósticos de câncer em trabalhos fornecidos pelo repositório GEO (https://metacpan.org/pod/App::geoCancerPrognosticDatasetsRetriever). No qual possui como entrada na linha de comando o nome do tipo do câncer de interesse, busca por ele na plataforma GEO e retorna informações necessárias dos códigos de acesso no GEO dos conjuntos de dados dos prognósticos. Assim, destaca-se a robustez do aplicativo em buscar

e agrupar as informações dos conjuntos de dados disponíveis no repositório ao contrário de uma busca um por um de maneira manual. O software recupera vários conjuntos de dados em pouco tempo impactando nas pesquisas oncológicas e otimizando o tempo utilizado.

Também a importante salientar sobre a utilização de *scripts* devido a sua facilidade de automatizar processos, sendo acessíveis para todos os usuários (HERNÁNDEZ et al, 2018). Dessa forma, resulta-se na obtenção de dados cada vez mais confiáveis e com a otimização do tempo. Além disso, o uso de *scripts* ou módulos vai disponibilizar automatizações nas análises e a diminuição do trabalho manual e diminuindo o gasto de tempo de forma considerável. Outra contribuição do uso de alternativas computacionais, principalmente da área da Bioinformática, é a diminuição dos erros possíveis, buscando em repositórios de dados, *scripts* e módulos funcionais impactando nas diversas pesquisas, sempre otimizando o tempo e automatizando as buscas de dados cada vez mais confiáveis.

OBJETIVOS

Objetivo Geral

Descrever a aplicação da ferramenta de Bioinformática geoCancerPrognosticDatasetsRetriever fornecido pelo um repositório de módulos MetaCPAN buscando por trabalhos publicados no repositório *Gene Expression Omnibus* (GEO) detalhando prognósticos de câncer de maneira automatizada resultando em dados acessórios para a pesquisa oncológica.

Objetivos Específicos

- Instalar e executar a ferramenta de Bioinformática geoCancerPrognosticDatasetsRetriever, disponível gratuitamente pelo repositório de módulos CPAN via MetaCPAN;
- Utilizar o repositório de dados genômicos do *Gene Expression Omnibus* (GEO) como fornecedor na pesquisa oncológica;
- Descrever a utilização da ferramenta em linguagem de programação Perl para gerar dados e informações necessárias ou acessórias na utilização da área biotecnológica;
- Fornecer as orientações necessárias para buscar dados acessórios com foco em pesquisas oncológicas provendo funções específicas de cada tipo de câncer para a confirmações das informações produzidas em estudos;

- Gerar material de apoio a ser utilizado em pesquisa oncológica, com mais precisão de dados, confiabilidade dos resultados otimizando o tempo.

MATERIAL E MÉTODOS

Especificações

Foram utilizados para atingir os objetivos descritos algumas ferramentas de Bioinformática. Uma delas foi a ferramenta disponível gratuitamente no repositório de módulos CPAN via MetaCPAN, chamada geoCancerPrognosticDatasetsRetriever e o repositório de dados genômicos *Gene Expression Omnibus* (GEO) disponível online gratuitamente, utilizando linha de comando via terminal Ubuntu devidamente instalado em um computador pessoal com configuração Intel Core i5-4440 CPU @ 3.10GHz, 12,0 GB RAM, utilizando o sistema operacional Windows 10 com a taxa de download de aproximadamente 350,0 Mbps via cabo.

Ferramenta geoCancerPrognosticDatasetsRetriever

A ferramenta foi desenvolvida em linguagem de programação Perl podendo ser utilizada no sistema operacional Windows, Linux ou MacOS. No presente trabalho foi utilizado no sistema operacional Windows no qual é necessário a pré-instalação do subsistema Linux, *Windows Subsystem for Linux* (WSL) (ALAMEER, CHICCO, 2022).

Processo de Instalação e Comandos

Durante os trabalhos foram destacados diversos requisitos de pré-instalações não descritas no trabalho supracitado. Porém, esses requisitos são necessários para a utilização da ferramenta no sistema operacional Windows. A instalação do WSL pode ser feita via Prompt de Comando sendo executado como administrador com o comando:

```
$ wsl --install
```

Em seguida foi solicitado a reiniciação do sistema. Assim que iniciado a plataforma instalada Ubuntu resultou em um erro descrito como “Please enable the Virtual Machine

Platform Windows feature and ensure virtualization is enabled in the BIOS.” resolvido através habilitação da virtualização pela BIOS da máquina.

Logo após a devida instalação do WSL, foi iniciado o terminal Ubuntu via administrador, sendo necessário a criação de um usuário e senha para disponibilidade da plataforma seguido pelos comandos descritos no trabalho, então utilizado alguns comandos em sequência:

```
$ sudo su
$ sudo apt-get install -y libfile-homedir-perl
$ sudo apt update
$ sudo apt -y install cpanminus
$ perl -MCPAN -e 'install "LWP::Simple"'
$ sudo apt-get install -y libnet-ssleay-perl
```

Após as instalações dos pré-requisitos na máquina, foi utilizado um último comando que se trata propriamente da instalação da ferramenta:

```
$cpanm App::geoCancerPrognosticDatasetsRetriever
```

(https://github.com/AbbasAlameer/geoCancerPrognosticDatasetsRetriever).

Execução da Ferramenta e Observações relevantes

Com a ferramenta e seus pré-requisitos devidamente instalados, o próximo passo é a execução da ferramenta. Para isso, é preciso abrir o terminal Ubuntu via administrador e executado o seguinte comando, usando como parâmetros as plataformas “GPL570 GPL97 GPL96” estabelecidas no trabalho (ALAMEER, CHICCO, 2022). GPL570: Affymetrix Human Genome U133 Plus 2.0 Array. GPL97: Affymetrix Human Genome U133B Array. GPL96: Affymetrix Human Genome U133A Array.
(https://www.ncbi.nlm.nih.gov/geo/browse/?view=platforms).

```
$ geoCancerPrognosticDatasetsRetriever -d “nome câncer” -p “GPL570 GPL97 GPL96” -f “/nome_files/” -k
```

(ALAMEER, CHICCO, 2022).

No presente trabalho foi utilizado o comando para 34 tipos de câncer, pesquisados diretamente da plataforma Google, onde foi substituído no comando acima o “nome câncer” pelo tipo do câncer.

Através do total de prognósticos encontrados para cada tipo de câncer foi realizado uma busca pelo código do prognóstico GEO diretamente no banco de dados e diretamente nos arquivos gerados, no qual foi encontrado as informações dos trabalhos de prognósticos publicados no repositório.

RESULTADOS E DISCUSSÃO

Análises dos Dados

Primeiramente foi observado em pesquisas a publicação recente sobre o desenvolvimento da ferramenta geoCancerPrognosticDatasetsRetriever disponível gratuitamente para utilização. Nela foi observada a aplicação para pesquisa oncológica atuando no agrupamento de dados de prognósticos fornecidos pelo repositório GEO, gerando dados acessórios para pesquisa científica de diferentes tipos de câncer.

O repositório GEO atualmente possui mais de 6 milhões de amostras publicadas em seu banco de dados com cerca de 25 mil plataformas descritas. Com o foco em prognósticos de câncer foi usado apenas três parâmetros de plataformas “GPL570 GPL97 GPL96” estabelecidos pelo trabalho anterior (ALAMEER, CHICCO, 2022) para o uso da ferramenta geoCancerPrognosticDatasetsRetriever. A ferramenta faz o agrupamento desses dados publicados no repositório, retornando os seus códigos de prognósticos e os arquivos relacionados aos tipos de câncer pesquisados, conforme citado foi usado o comando no terminal Ubuntu com 34 tipos de câncer e cada um deles gerou resultados diferentes conforme detalhados na tabela abaixo:

App :: geoCancerPrognosticDatasetsRetriever
Results by parameter: "GPL570 GPL97 GPL96"

Nº	Cancer List	Total Number Prognostic Datasets Found	Prognostic Datasets Name Found	Size on Disk (GB)
1	bile duct	0	..	0,18
2	bladder	2	GSE31684; GSE5287	1,11
3	bone	0	..	0,58
4	brain	9	GSE43378; GSE34771; GSE36245; GSE33331; GSE23554; GSE16581; GSE13041; GSE4271; GSE4412	7,82
5	breast	33	GSE157284; GSE158309; GSE135565; GSE124648; GSE124647; GSE71053; GSE58984; GSE63205; GSE88770; GSE58812; GSE61304; GSE48390; GSE45255; GSE43502; GSE27120; GSE31448; GSE20711; GSE31519; GSE26639; GSE25066; GSE25065; GSE25055; GSE17705; GSE21653; GSE19615; GSE16391; GSE12945; GSE11121; GSE9195; GSE7390; GSE5327; GSE6532; GSE4922	43,90
6	cervical	0	..	1,45
7	colorectal	10	GSE161158; GSE143985; GSE92921; GSE72969; GSE72968; GSE38832; GSE39582; GSE17537; GSE17536; GSE12945	2,82
8	esophageal	1	GSE42363	1,20
9	extragastrointestinal stromal	6	GSE58984; GSE56315; GSE56313; GSE63885; GSE27120; GSE7788	3,45
10	eye	1	GSE22138	1,00
11	gallbladder	0	..	0,08
12	GIST	0	..	3,71
13	head	2	GSE27020; GSE42363	4,55
14	head and neck	2	GSE27020; GSE42363	4,55
15	kidney	2	GSE33371; GSE22541	8,80

16	leukemia	12	GSE107951; GSE94801; GSE58445; GSE47018; GSE49896; GSE43176; GSE37642; GSE22762; GSE11877; GSE12945; GSE12417; GSE8970	40,40
17	liver	2	GSE40873; GSE33371	14,20
18	lung	17	GSE157011; GSE157010; GSE157009; GSE68465; GSE19722; GSE50081; GSE30219; GSE37745; GSE31210; GSE28571; GSE29013; GSE14814; GSE19188; GSE10072; GSE8894; GSE3593; GSE3141	24,10
19	lymphoma	13	GSE110376; GSE107951; GSE58445; GSE56315; GSE56313; GSE48097; GSE34771; GSE22771; GSE22470; GSE7788; GSE16131; GSE12945; GSE10846	23,50
20	mesothelioma	0	..	0,70
21	multiple myeloma	3	GSE82307; GSE24080; GSE9782	13,70
22	myeloma	3	GSE82307; GSE24080; GSE9782	13,30
23	nasopharyngeal	0	..	0,00
24	neck	2	GSE27020; GSE42363	4,55
25	neuroendocrine	0	..	7,10
26	ovarian	8	GSE65986; GSE63885; GSE44104; GSE30161; GSE32062; GSE23554; GSE26712; GSE14764	12,20
27	pancreatic	1	GSE32676	6,71
28	prostate	1	GSE25136	8,26
29	skin	0	..	1,04
30	soft tissue sarcoma	2	GSE17674; GSE17618	3,45
31	stomach	2	GSE38749; GSE17187	6,53
32	testicular	0	..	3,09
33	thyroid	0	..	1,30
34	uterine	2	GSE65986; GSE29436	5,87
..	**TOTAL**	**136**	**..**	**275,20**

Tabela 1: Listagem dos diferentes tipos de câncer executados na ferramenta geoCancerPrognosticDatasetsRetriever com seus códigos de prognósticos e tamanhos em gigabytes. Fonte: Próprio autor.

Conforme observado na Tabela 1, a ferramenta foi utilizada para 34 tipos de câncer no qual foi gerado 136 prognósticos totais como saída de dados. Cada tipo de câncer possui seu número de prognóstico, seguido dos seus códigos e consequentemente do seu tamanho em disco ocupado em download, cerca de 275,20 GB ocupadas em disco na máquina.

É importante salientar o número diferente da quantidade de prognósticos obtidos para cada tipo de câncer, onde alguns possuem uma discrepância considerável com números extremos de 0 prognósticos como é o caso de vários na lista e de 33 prognósticos como é o caso do câncer de mama. Essa diferença se dá pelas publicações no repositório GEO diretamente ligada as pesquisas sobre aquele determinado câncer no qual alguns possuem interesse maior na comunidade científica do que outros e consequentemente possuem mais dados publicados à disposição, cabíveis de alterações já que o repositório sempre recebe novas submissões. Diante dos arquivos gerados pela ferramenta foi destacado em relação ao tamanho ocupado em disco de cada câncer, mesmo com a quantidade de prognósticos zero. Isso se dá pelo fato de que a ferramenta correlaciona o nome do câncer pesquisado com as plataformas que no presente trabalho foi usado três como parâmetro “GPL570 GPL97 GPL96”. Dessa maneira, a ferramenta correlaciona a palavra-chave “nome do câncer” com a plataforma selecionada de entrada para gerar os prognósticos corretos onde dentro dos arquivos também possuem outros códigos que não foram destacados como prognósticos e isso se dá justamente pelo fato da correlação citada acima, consequentemente resultando no tamanho em GB de cada um. Esses parâmetros de entrada também podem ser modelados de acordo com o interesse e conhecimento do pesquisador.

Destacando um exemplo do câncer de bexiga que resultou em dois prognósticos “GSE31684 e GSE5287”, pesquisando diretamente pelo código gerado no repositório GEO (https://www.ncbi.nlm.nih.gov/geo/) obtivemos as seguintes figuras:

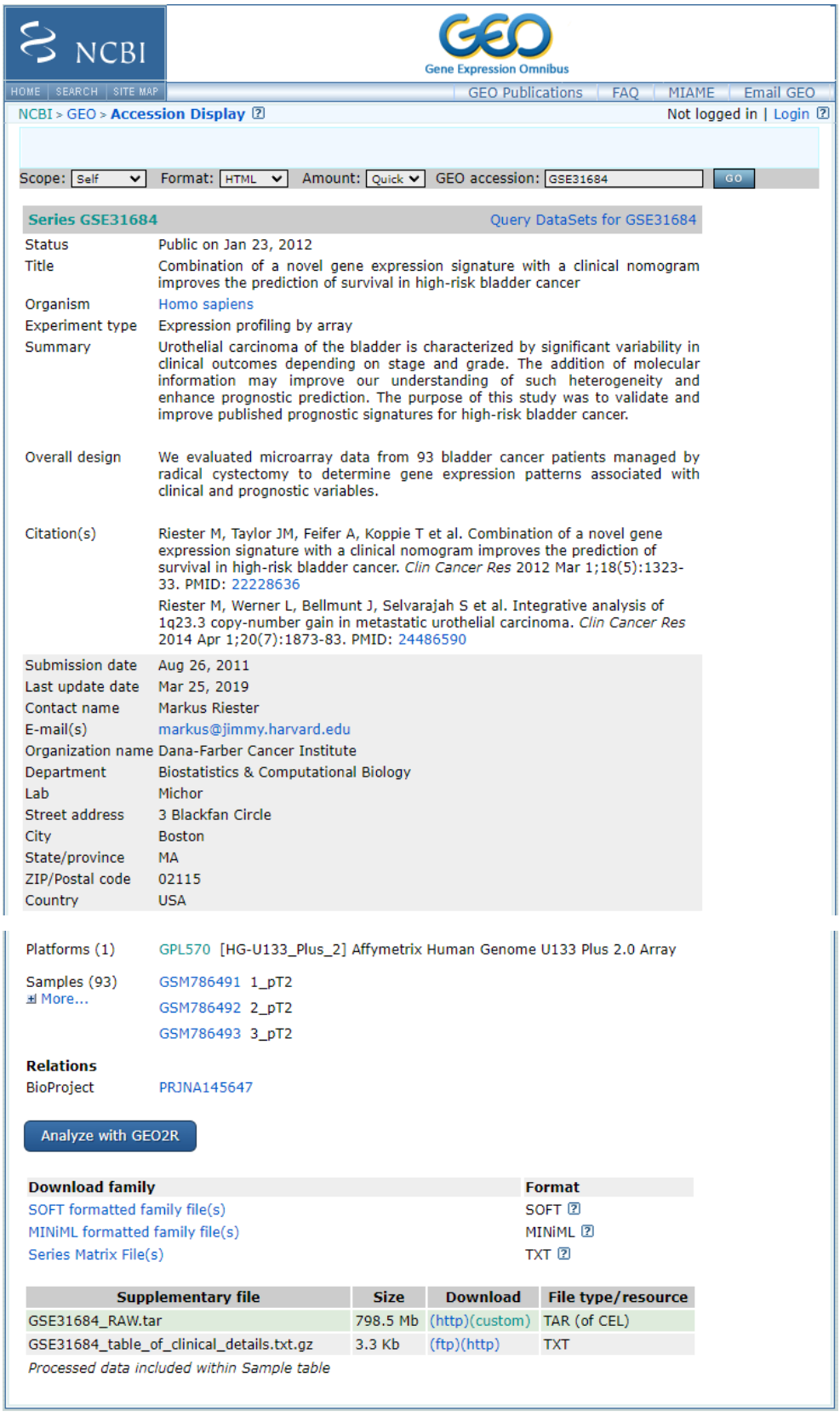


Figura 1: *Print* da página do resultado da pesquisa do prognóstico “GSE31684” câncer de bexiga (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31684).

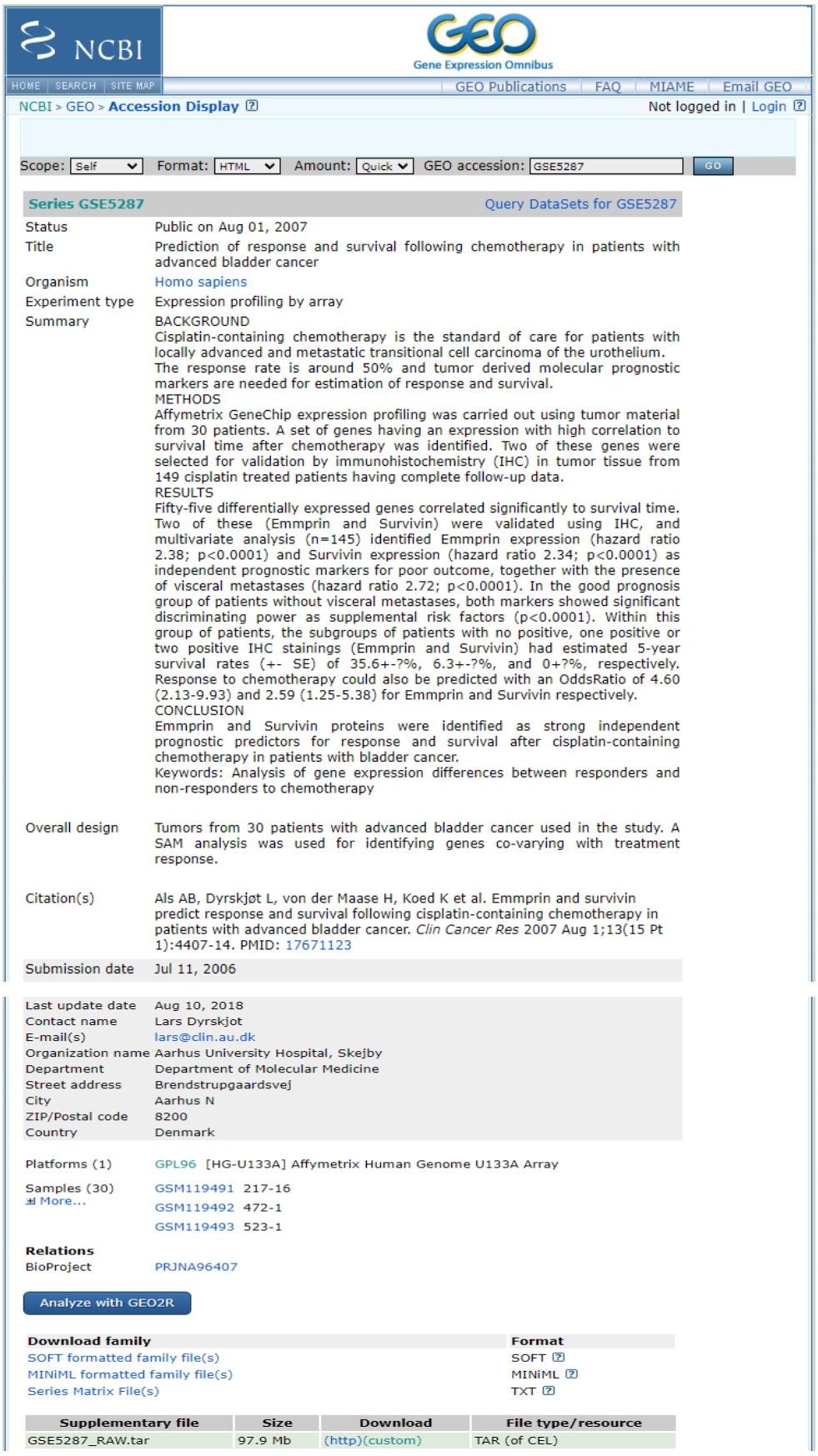


Figura 2: *Print* da página do resultado da pesquisa do prognóstico “GSE5287” câncer de bexiga (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5287).

Nas figuras 1 e 2 obteve-se o resultado da pesquisa pelo código resultado do câncer de bexiga diretamente no site do repositório GEO, no qual é descrito o trabalho publicado de acordo com os parâmetros solicitados. Descrevem a pesquisa sobre os tratamentos realizados em diversos pacientes, autores, datas de submissão e atualização, contatos para com os autores da pesquisa, cita também a plataforma de parâmetro usada que foi correlacionada pela ferramenta geoCancerPrognosticDatasetsRetriever e os arquivos texto que podem ser feitos o download.

Análise Geral dos Arquivos

Os arquivos gerados pela execução da ferramenta são salvos separadamente por pastas de cada tipo de câncer e consequentemente dentro delas os seus prognósticos. Selecionando os arquivos corretos pelo código de cada prognóstico de câncer e se aprofundando no arquivo texto contendo informações dos trabalhos publicados sobre os prognósticos. Foi observado que além das informações relevantes dos trabalhos que foram publicados no repositório também se obtinham informações irrelevantes, que até o momento que foram gerados pela execução da ferramenta, informações tipo “loop”. Por sua vez estão relacionadas pelo considerável tamanho em disco dos arquivos.

Vale citar também a importância da escolha do banco de dados fornecedor para a pesquisa oncológica, conforme o *Gene Expression Omnibus* (GEO) que por sua vez possui uma quantidade considerável de amostras, experimentos e perfis de expressão gênica com facilidade no acesso e busca, sendo um repositório público funcional para a pesquisa oncológica.

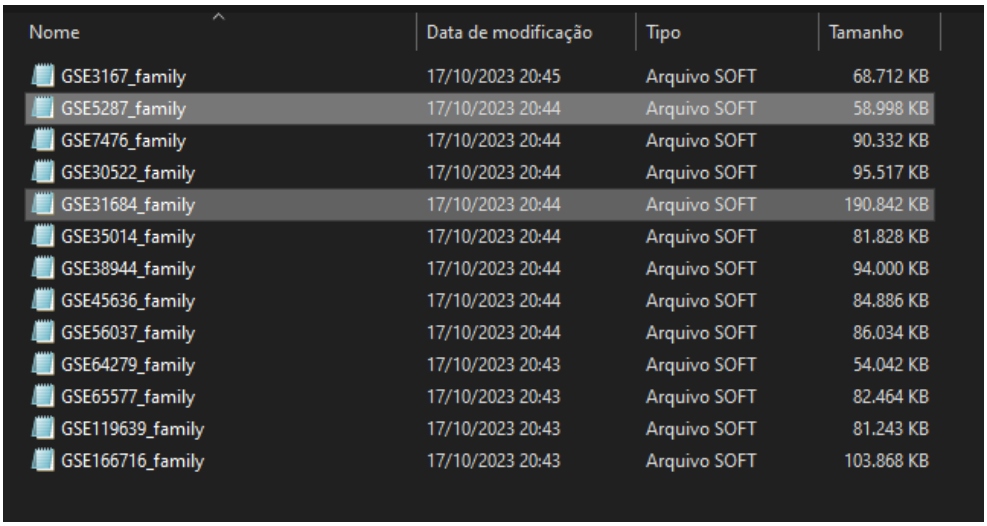
O processo manual de observação dentro dos arquivos de texto demanda um tempo considerável e consequentemente cabíveis de erros, recomendado então usar os códigos dos prognósticos gerados apenas para pesquisar diretamente no repositório GEO que por sua vez temos acesso para os possíveis trabalhos, não descartando definitivamente o uso dos arquivos salvos na máquina.

Foram obtidas, na utilização da ferramenta, duas observações importantes: uma delas ao iniciar o terminal Ubuntu, ao se solicitar o comando $ sudo su no início e em seguida o comando para execução da ferramenta. Dessa forma, os arquivos gerados ficarão salvos no *root* do computador utilizado. Isso traz problemas pois não é possível se ter acesso aos arquivos gerados, devido a política de segurança do sistema operacional Ubuntu. Para se resolver esse problema, basta ao iniciar o terminal, não é preciso solicitar o comando $ sudo su e sim

diretamente o comando da execução da ferramenta, sendo os arquivos gerados salvos em \\wsl.localhost\Ubuntu\home\"UsersLocal", no qual temos acesso pela máquina. A outra observação foi encontrada ao realizar testes de instalação e execução em outras máquinas que possuem conexão sem fio de internet como por exemplo notebooks, utilizando o mesmo sistema operacional usado anteriormente. Nesse cenário, a execução da ferramenta não foi concluída corretamente fazendo com que alguns arquivos sejam corrompidos devido baixa taxa de conexão, sugerindo que execute somente com conexão via cabo na máquina.

Após executado a ferramenta pela linha de comando no terminal Ubuntu e concluído o download dos arquivos, é possível ter acesso aos conjuntos de dados gerados e seus determinados códigos de acesso. Dentre os arquivos é possível encontrar o total de prognósticos gerados para cada tipo de câncer, onde alguns deles possuem mais de um prognóstico, seus códigos de acesso direto ao GEO e seus arquivos de dados. Foi importante observar os tamanhos consideravelmente grandes ocupados no disco de armazenamento em gigabytes de cada pasta dos arquivos gerados, pois cada um possui tempo de download e tamanhos diferentes.

Seguindo com o mesmo exemplo de câncer, o de bexiga, assim que executada a ferramenta, ela gera a pasta “bladder_files”, onde nela é dividida em um arquivo de texto contendo as fontes das amostras por onde a ferramenta fez o agrupamento, por todo o caminho e no final ela reporta o “Total prognostic datasets found” que nesse exemplo são no total de dois “[1] GSE31684; [2] GSE5287” conforme a figura 4 e uma outra pasta contendo todos os arquivos de texto .SOFT conforme a figura 3 gerados pelo agrupamento da ferramenta inclusive os reportados dos prognósticos. Também estão disponíveis para download no repositório GEO quando buscados diretamente pelo código conforme realizado nas figuras 1 e 2. Os arquivos SOFT podem ser abertos como arquivos texto e são neles que existem as informações relevantes e irrelevantes de “loop” citadas. São esses arquivos que possuem um tamanho no disco considerável e ao serem abertos a máquina precisam de um determinado tempo para processar e abrir para leitura ou buscar pelas informações na linha de comando Ubuntu, podendo assim serem um alvo a ser considerável para estudos futuros.



Nome	Data de modificação	Tipo	Tamanho
GSE3167_family	17/10/2023 20:45	Arquivo SOFT	68.712 KB
GSE5287_family	17/10/2023 20:44	Arquivo SOFT	58.998 KB
GSE7476_family	17/10/2023 20:44	Arquivo SOFT	90.332 KB
GSE30522_family	17/10/2023 20:44	Arquivo SOFT	95.517 KB
GSE31684_family	17/10/2023 20:44	Arquivo SOFT	190.842 KB
GSE35014_family	17/10/2023 20:44	Arquivo SOFT	81.828 KB
GSE38944_family	17/10/2023 20:44	Arquivo SOFT	94.000 KB
GSE45636_family	17/10/2023 20:44	Arquivo SOFT	84.886 KB
GSE56037_family	17/10/2023 20:44	Arquivo SOFT	86.034 KB
GSE64279_family	17/10/2023 20:43	Arquivo SOFT	54.042 KB
GSE65577_family	17/10/2023 20:43	Arquivo SOFT	82.464 KB
GSE119639_family	17/10/2023 20:43	Arquivo SOFT	81.243 KB
GSE166716_family	17/10/2023 20:43	Arquivo SOFT	103.868 KB

Figura 3: *Print* dos arquivos salvos na máquina sobre o exemplo do câncer de bexiga. Fonte: Próprio autor.

BLADDER-CANCER_2023-010-17_h2043 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM181nnn/GSM181000/
326.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180999/
327.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180998/
328.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180997/
329.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180996/
330.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180995/
331.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180994/
332.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180993/
333.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180992/
334.
Platform: GPL570 Series: GSE7476
Organism: Homo sapiens
FTP download: GEO (CEL) ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM180nnn/GSM180991/
================================================================================
========
Total prognostic datasets found: 2
[1] GSE31684
[2] GSE5287
```

Ln 1, Col 1 | 100% | Unix (LF) | UTF-8

Figura 4: *Print* do arquivo salvo na máquina, fornecendo o total e os códigos de prognósticos encontrados do exemplo câncer de bexiga. Fonte: Próprio autor.

Perspectivas para o Trabalho

Conforme citado, os arquivos gerados possuem um tamanho considerável dentro do armazenamento em disco na máquina. As informações agrupadas pela ferramenta dentro dos arquivos texto é de difícil compreendimento devido algumas das informações estarem embaralhadas ao meio de outras informações irrelevantes até o momento em forma de “loop”. Para que esses arquivos sejam abertos para leitura a máquina necessita de um determinado tempo para processar, assim surgindo uma oportunidade de estudos e pesquisas com foco nessas informações e arquivos agrupados.

O ponto de partida de uma possível pesquisa é justamente trabalhar em cima dos dados gerados pela ferramenta com o objetivo de processar e filtrar por meio de palavras chaves os dados para que seja mais viável e possível o uso aplicado dessas informações. Também a serem usados as informações dos trabalhos já realizados e publicados no repositório que podem ser pesquisados pelos prognósticos gerados pela ferramenta diretamente no GEO, a serem usadas na pesquisa oncológica e de trabalhos relacionados por pesquisadores da área.

CONCLUSÃO

A pesquisa oncológica não pode parar e necessita cada vez mais de investimentos, avanços tecnológicos, desenvolvimento e tratamentos. É de suma importância possuir opções e caminhos para se obter as informações acessórias e apoio conforme a possibilidade gerada no presente trabalho.

A ferramenta geoCancerPrognosticDatasetsRetriever tem com o objetivo facilitar os trabalhos e estudos na área de atuação, servindo como apoio e fornecendo a confiabilidade necessária para os trabalhos científicos.

A Bioinformática é uma opção que na maioria das vezes pode ser usada para facilitar a execução de trabalhos científicos e conforme descrito, a ferramenta geoCancerPrognosticDatasetsRetriever consegue agrupar os dados do repositório GEO fornecendo as informações acessórias e relevantes para a pesquisa oncológica, destacando possíveis elementos biológicos que podem ser utilizados como prognóstico de câncer.

Por sua vez o tamanho de armazenamento dos arquivos gerados pode se tornar também um ponto de investigação e aplicação de trabalho, podendo servir como ponto de partida, filtrando e coletando as informações.

REFERÊNCIAS

HERNÁNDEZ Y, BERNSTEIN R, PAGAN P, VARGAS L, MCCAIG W, RAMRATTAN G, AKTHER S, LARRACUENTE A, DI L, VIEIRA FG, QIU WG. **BPWRAPPER: BIOPERL-BASED SEQUENCE AND TREE UTILITIES FOR RAPID PROTOTYPING OF BIOINFORMATICS PIPELINES**. BMC BIOINFORMATICS. 2018 MAR 2;19(1):76. DOI: 10.1186/S12859-018-2074-9. PMID: 29499649; PMCID: PMC5833151.

THE COMPREHENSIVE PERL ARCHIVE NETWORK - WWW.CPAN.ORG. DISPONÍVEL EM: <HTTPS://WWW.CPAN.ORG/>. ACESSO EM: 28 SET. 2023

THE PERL PROGRAMMING LANGUAGE - WWW.PERL.ORG. DISPONÍVEL EM: <HTTPS://WWW.PERL.ORG/>. ACESSO EM: 28 SET. 2023.

VOS RA, CARAVAS J, HARTMANN K, JENSEN MA, MILLER C. **BIO::PHYLO-PHYLOINFORMATIC ANALYSIS USING PERL**. BMC BIOINFORMATICS. 2011 FEB 27;12:63. DOI: 10.1186/1471-2105-12-63. PMID: 21352572; PMCID: PMC3056726.

GENE EXPRESSION OMNIBUS - WWW.NCBI.NLM.NIH.GOV/GEO/. DISPONÍVEL EM: <HTTPS://WWW.NCBI.NLM.NIH.GOV/GEO/>. ACESSO EM: 02 OUT. 2023.

APP :: GEO CANCER PROGNOSTIC DATASETS RETRIEVER - WWW.METACPAN.ORG/POD/APP::GEOCANCERPROGNOSTICDATASETSRETRIEVER. DISPONÍVEL EM: < HTTPS://METACPAN.ORG/POD/APP::GEOCANCERPROGNOSTICDATASETSRETRIEVER>. ACESSO EM: 02 OUT. 2023.

CLOUGH E, BARRETT T. **THE GENE EXPRESSION OMNIBUS DATABASE**. METHODS MOL BIOL. 2016;1418:93-110. DOI: 10.1007/978-1-4939-3578-9_5. PMID: 27008011; PMCID: PMC4944384.

ALAMEER A, CHICCO D. **GEOCANCERPROGNOSTICDATASETSRETRIEVER: A BIOINFORMATICS TOOL TO EASILY IDENTIFY CANCER PROGNOSTIC DATASETS ON GENE EXPRESSION OMNIBUS (GEO)**. BIOINFORMATICS, VOLUME 38, ISSUE 6, MARCH 2022, PAGES 1761–1763.

ABBAS ALAMEER / GEO CANCER PROGNOSTIC DATASETS RETRIEVER - WWW.GITHUB.COM/ABBASALAMEER/GEOCANCERPROGNOSTICDATASETSRETRIEVER. DISPONÍVEL EM: < HTTPS://GITHUB.COM/ABBASALAMEER/GEOCANCERPROGNOSTICDATASETSRETRIEVER>. ACESSO EM: 16 OUT. 2023.

WANG J J, LEI K F, HAN F. **TUMOR MICROENVIRONMENT: RECENT ADVANCES IN VARIOUS CANCER TREATMENTS**. EUROPEAN REVIEW FOR MEDICAL AND PHARMACOLOGICAL SCIENCES, V. 22, N. 12, P. 3855–3864, 1 JUN. 2018.