Thais Talarico Hosokawa

IMPLEMENTAÇÃO DE MÉTODOS DE ANÁLISE POR PROGRAMAÇÃO VISUAL EM DADOS DE CÂNCER DE COLO DE ÚTERO DO *THE CANCER GENOME ATLAS*

Dissertação apresentada ao Programa de Pós-Graduação da Fundação Pio XII – Hospital de Câncer de Barretos para obtenção do Título de Mestre em Ciências da Saúde

Área de Concentração: Oncologia

Orientador: Prof. Dra. Adriane Feijó Evangelista
Co-orientador: Prof. Dr. Rui Manuel Reis
Linha de pesquisa: Oncologia molecular e patologia tumoral

Barretos, SP
2019

Thais Talarico Hosokawa

IMPLEMENTAÇÃO DE MÉTODOS DE ANÁLISE POR PROGRAMAÇÃO VISUAL EM DADOS DE CÂNCER DE COLO DE ÚTERO DO *THE CANCER GENOME ATLAS*

Dissertação apresentada ao Programa de Pós-Graduação da Fundação Pio XII – Hospital de Câncer de Barretos para obtenção do Título de Mestre em Ciências da Saúde

Área de Concentração: Oncologia

Orientador: Prof. Dra. Adriane Feijó Evangelista
Co-orientador: Prof. Dr. Rui Manuel Reis
Linha de pesquisa: Oncologia molecular e patologia tumoral

Barretos, SP
2019

H825i Hosokawa, Thais Talarico.

Implementação de métodos de análise por programação visual em dados de câncer de colo de útero Do The Cancer Genome Atlas. / Thais Talarico Hosokawa. - Barretos, SP 2019.
93 f. : il.
Orientadora: Adriane Feijó Evangelista.
Coorientador: Dr. Rui Manuel Reis.

Dissertação (Mestrado em Ciências da Saúde) – Fundação Pio XII – Hospital de Câncer de Barretos, 2019.

1. Neoplasias do colo do útero. 2. Metilação de DNA. 3. Exoma. 4. Expressão gênica. 5. Biologia computacional. 6. Genômica. I. Autor. II. Evangelista, Adriane Feijó. III. Reis, Rui Manuel. IV. Título.

CDD 616.994

FICHA CATALOGRÁFICA
Preparada por Martins Fideles dos Santos Neto CRB 8/9570
Biblioteca da Fundação Pio XII – Hospital de Câncer de Barretos

Esta dissertação foi elaborada e está apresentada de acordo com as normas da Pós-Graduação do Hospital de Câncer de Barretos – Fundação Pio XII, baseando-se no Regimento do Programa de Pós-Graduação em Oncologia e no Manual de Apresentação de Dissertações e Teses do Hospital de Câncer de Barretos. Os pesquisadores declaram ainda que este trabalho foi realizado em concordância com o Código de Boas Práticas Científicas (FAPESP), não havendo nada em seu conteúdo que possa ser considerado como plágio, fabricação ou falsificação de dados. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade dos autores e não necessariamente refletem a visão da Fundação Pio XII – Hospital de Câncer de Barretos. Os pesquisadores declaram não ter qualquer conflito de interesse relacionado a este estudo.

Esse trabalho é dedicado à minha família e aos meus amigos,
que souberam entender a dedicação necessária para obtê-lo.
Dedico também ao meu pai, que mesmo tendo partido tão cedo,
deixou o bom exemplo marcado nas minhas ações.
Ainda dedico a qualquer pessoa que ousar sonhar e realizar.

AGRADECIMENTOS

À *Profª. Dra. Adriane Feijó Evangelista* por toda a orientação, dedicação e compartilhamento de sua vivência. Pelos bons exemplos acadêmicos, pelas ideias inovadoras e por acreditar no meu trabalho.

Ao *Prof. Dr. Rui Manuel Reis*, pela experiência e objetividade. Por estar sempre por perto, mesmo tendo tantas funções a serem desempenhadas o tempo todo.

Ao meu marido, *Fernando Takashi Hosokawa*, que me incentivou a seguir a carreira acadêmica, mesmo sabendo que isso significaria menos tempo para ele. À *minha família*, que é sempre o meu suporte.

Aos membros da banca de qualificação, *Prof. Dr. Celso Teixeira Mendes Junior* e *Prof. Dr. José Humberto Tavares Guerreiro Fregnani*, pelas sugestões e críticas ao longo da elaboração desta dissertação, trazendo grandes contribuições ao trabalho.

À *Dra. Luciane Sussuchi da Silva* e ao *Murilo Machado*, que uniram esforços nas instalações e desenvolvimento, sempre com excelente humor. Ao Gabriel Andrade, da Tecnologia da Informação do Hospital de câncer de Barretos, que muito contribuiu.

À Dra.Mariana Bisarro dos Reis, por tantas vezes oferecer seu conhecimento para as análises epigenéticas. Tenho muito a te agradecer.

À *Profª. Dra. Fabiana de Lima Vazquez*, sempre disposta academicamente e como amiga.

Ao Dr.Vinicius Duval da Silva, patologista que realizou prontamente a revisão das lâminas.

Aos funcionários do Instituto de Ensino e Pesquisa, membros da diretoria e do Instituto social, que muito contribuíram no processo de obtenção dos dados necessários para esse

trabalho: Henrique Prata, Henrique Moraes Prata, Daniela Girardi, Marcela Marchioreto de Oliveira, Carolina Sgorlon Jorgetto e Joyce Silva Pinto.

Ao Núcleo de Epidemiologia e Bioestatística, por ter membros sempre disponíveis, academicamente e como amigos.

Ao Registro Hospitalar de Câncer do Hospital de Câncer de Barretos pela atualização de segmento das participantes, às funcionárias Lívia Duarte e Sara Almeida.

À Milene Gírio Marques e Martins Fideles dos Santos Neto da biblioteca científica pela prontidão em atender.

À Comissão de Pós-graduação da Fundação Pio XII – Hospital de Câncer de Barretos, especialmente a funcionária Mariana Evangelista. Também às ex-funcionárias Silvana Rodrigues Guitarrari e Brenda Honda Morais.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

TCGA	*The Cancer Genome Atlas* (O Atlas do Genoma do Câncer)
IARC	International Agency for Research on Cancer (Agência Internacional para pesquisa em Câncer)
HPV	*Human Papiloma Virus* (Vírus do papiloma humano)
DNA	*Deoxyribonucleic acid* (Ácido desoxirribonucleico)
RNA	*Ribonucleic acid* (Ácido ribonucleico)
NCI	*National Cancer Institute* (Instituto Nacional de Câncer dos Estados Unidos)
NIH	*National Institute of Health* (Agência Nacional de Saúde dos Estados Unidos)
RNAseq	*Ribonucleic acid sequencing* (sequenciamento de ribonucleico)
HCB	Hospital de Câncer de Barretos
API	*Application program interface*
UCSC	*University of California Santa Cruz*
HTTP	Hypertext transfer protocol
DMP	*Differentially methylated probes* (sondas diferencialmente metiladas)
DMR	*Differentially methylated regions* (regiões diferencialmente metiladas)

LISTA DE SÍMBOLOS

PB	Peta bytes
%	Porcentagem

RESUMO

Talarico T. Implementação de métodos de análise por programação visual em dados de câncer de colo de útero do *The Cancer Genome Atlas*. Dissertação (Mestrado) Barretos: Hospital de Câncer de Barretos; 2019.

JUSTIFICATIVA: O câncer de colo de útero é o quarto câncer mais comum em mulheres, com estimativa de 570.000 novos casos em 2018 e apesar dos importantes avanços nas formas de prevenção, constitui ainda um problema de saúde pública que deve ser alvo de novos estudos, especialmente em casos avançados detectados tardiamente. Mesmo com a classificação por métodos de citologia e da possibilidade de testar para HPV como parte dos programas de rastreamento, existe uma busca pela caracterização molecular que possibilite melhores formas de diagnóstico, bem como de prognóstico e tratamento. Os consórcios internacionais de pesquisa do câncer têm utilizado técnicas moleculares em larga-escala, gerando uma grande quantidade de informação. Um desses consórcios, o TCGA, revolucionou o entendimento biológico do câncer por meio da análise integrada de milhares de tumores. Para o máximo aproveitamento da grande quantidade de dados gerados faz se necessário o uso de complexas ferramentas de bioinformática que integrem e analisem os resultados e uma das principais barreiras atuais é a elevada complexidade das análises e a necessidade de profissionais com aprofundado conhecimento de bioinformática. Dessa forma, existe a necessidade do desenvolvimento de ambientes simplificados no qual pesquisadores sem conhecimento aprofundado de programação computacional possam gerar suas hipóteses e analisar seus resultados.

OBJETIVO: Este estudo tem como objetivo a implementação da plataforma *Galaxy*, um ambiente que permite o uso de métodos da programação visual, para análise de dados de câncer de colo de útero do TCGA, com foco nas amostras da população brasileira.

MATERIAIS E MÉTODOS: Consiste na implementação local do *Galaxy* como plataforma de análise de dados genômicos. A instalação local foi necessária por diversos motivos, incluindo se tratar de dados sensíveis. Para análise genômica, foram escolhidos três tipos de dados genômicos – exoma, RNAseq e metilação – de câncer de colo de útero do projeto TCGA, com foco nas amostras de 54 participantes da população brasileira, além dos dados clínicos. Foram implementados workflows, que são uma forma de visualização da análise completa, que permite reuso de modo reprodutível. Para obtenção dos dados, que são controlados pela

agência que financiou o TCGA, precisamos seguir os passos para obtenção e, tão logo obtivemos, prosseguimos para as análises, visando a identificação de biomarcadores epigenéticos. Dessa forma, a análise do metiloma teve como foco a comparação dos pacientes com câncer de colo de útero da população brasileira com um grupo controle, compreendendo a identificação de differentially methylated probes (DMPs), differentially methylated regions (DMRs), processos biológicos associados com a metilação (GSEA) e o interactoma das principais alterações com foco na região próxima ao sítio de início da transcrição.

RESULTADOS: A plataforma *Galaxy* foi implementada em um servidor local do Hospital de Amor, e os workflows puderam ser construídos nesse processo. Parte das ferramentas tiveram que ser adequadas por meio de um ambiente interativo, principalmente as associadas com o metiloma que foi o foco principal do estudo. Também foi possível a obtenção dos dados brutos para realização das análises. Em relação as análises com foco epigenético, após os filtros de qualidade restaram 396.035 probes, das quais 24.172 encontraram-se diferencialmente metiladas e puderam separar claramente os pacientes dos controles, assim como um subgrupo de amostras HPV-negativas e carcinoma mucinoso. Dessas, destacam-se as probes do gene PITX2 e IKZF1, em concordância com a literatura. Das DMRs, destacam-se os genes CALCA, EDNRB, RAB3C e GALR1 com várias probes em uma mesma região de enhancer. Dentre os processos biológicos, destacam-se os processos associados com alterações da histona H3K27me3 e do fator de transcrição SP1. Finalmente, os genes LNX1, SMAD2 e TNFRSF10A parecem constituir importantes nós nas redes de interactoma identificadas, com importantes papéis na regulação de diversos genes no contexto do câncer de colo de útero, identificados na população brasileira.

CONCLUSÃO: A implementação de uma plataforma para análise de dados por programação visual em um centro de pesquisa é uma tarefa complexa, que exige estrutura computacional e pessoal especializado. No entanto é também necessária haja vista que o *Galaxy* se provou uma plataforma muito útil nas análises desse trabalho e também na disponibilização de ferramentas para análise genômica no diagnóstico, proporcionando uma forma facilitada, confiável e reprodutível de análise para pesquisadores sem formação aprofundada de bioinformática. A implementação de uma plataforma para análise de dados por programação visual se provou uma plataforma muito útil nas análises desse trabalho, proporcionando uma forma facilitada, confiável e reprodutível de análise para pesquisadores sem formação aprofundada de bioinformática. As análises com foco no metiloma apresentaram novos

biomarcadores com foco na população brasileira que fornecem novos achados para compreensão dessa doença.

PALAVRAS-CHAVE: Neoplasias do Colo do Útero, metilação de DNA, exoma, biologia computacional, genômica

ABSTRACT

Talarico T. Implementação de métodos de análise por programação visual em dados de câncer de colo de útero do *The Cancer Genome Atlas*. Dissertação (Mestrado) Barretos: Hospital de Câncer de Barretos; 2019.

BACKGROUND: Cervical cancer is the fourth most common cancer in women, with estimates of 570,000 new cases in 2018 and despite of important advances in prevention, still represents an important health problem that must be the target of new studies, especially the advanced late detected cases. Even with the classification by cytology and the possibility of HPV-testing as part of screening programs, there is a demand for molecular characterization to achieve better diagnosis ways, as well as prognosis and treatment. The international research consortia in cancer have been using high-throughput sequencing techniques, generating a huge amount of data. One of them, the TCGA, revolutionized the knowledge on cancer biology by the integrative analysis of thousands of tumors. For the maximum use of the great amount of data generated it is necessary the use of complex bioinformatics' tools that can integrate and analyze the result, and one of the main barriers is the high complexity of the analysis and the requirement of professionals with in-deep knowledge in bioinformatics. Thus, there is a demand for the development of simplified environments where researchers without an in-deep knowledge of computing programming can generate their hypothesis and analyze their results.

PURPOSE: The aim of this study is the implementation of the platform Galaxy, an environment that allows the use of visual programming, to analyze the cervical cancer data of the TCGA, focusing on the Brazilian population samples.

MATERIALS AND METHODS: Consists of implementing Galaxy locally as a platform for processing genomic data analysis. The local implementation was necessary for many reasons, one of which is the need to process sensitive data. For the analysis were chosen three genomic datatypes – exome, RNAseq and methylation – of cervical cancer of the TCGA Project, focusing on the 54 participant samples of the TCGA project, as well as their clinical data. We designed flowcharts to plan the analysis of the three datatypes using Bizagi Modeler. We then customized the installed platform, implementing tools for planned analysis, obtaining the workflows that are a visualization of the complete analyze, allowing the reuse in a reproducible manner. In order to obtain the data permission download, which is controlled

by the TCGA's funder agency, it was necessary to follow the steps and as soon as acquired, we proceed the analysis, aiming epigenetic biomarkers. Thus, the methylome analysis had as the aim a comparison between cervical cancer patients and a control group, comprising the identification of differentially methylated probes (DMPs), differentially methylated regions (DMRs), associated biological processes with the methylation (GSEA) and the interactome of the main alterations focused on the region near to the transcription start site.

RESULTS: The platform Galaxy was implemented in a local server of Barretos Cancer Hospital and the workflows could be developed in this process. Part of the tools had to be used by the interactive environment, especially the ones related to the methylome, the focus of this study. Also, was possible to obtain the raw data to perform the analysis. Regarding to the analysis focused on epigenetics, after the quality filters were used there were 396,035 probes, of which 24,172 were differentially methylated and clearly separated patients and controls, as well as a subgroup of samples HPV negative and mucinous carcinoma. Of this, we can highlight the probes of gene PITX2 and IKZF1, in accordance with the literature. Of the differentially methylated regions, we can highlight the genes CALCA, EDNRB, RAB3C e GALR1 with many probes in a single enhancer region. Regarding the biological processes, stand out the associated with histone alteration H3K27me3 and the transcription factor SP1. Finally, the genes LNX1, SMAD2 e TNFRSF10A appear to constitute important nodes on the identified interactoma networks, with important roles on the regulation of a variety of genes on de cervical cancer context, identified on the Brazilian population.

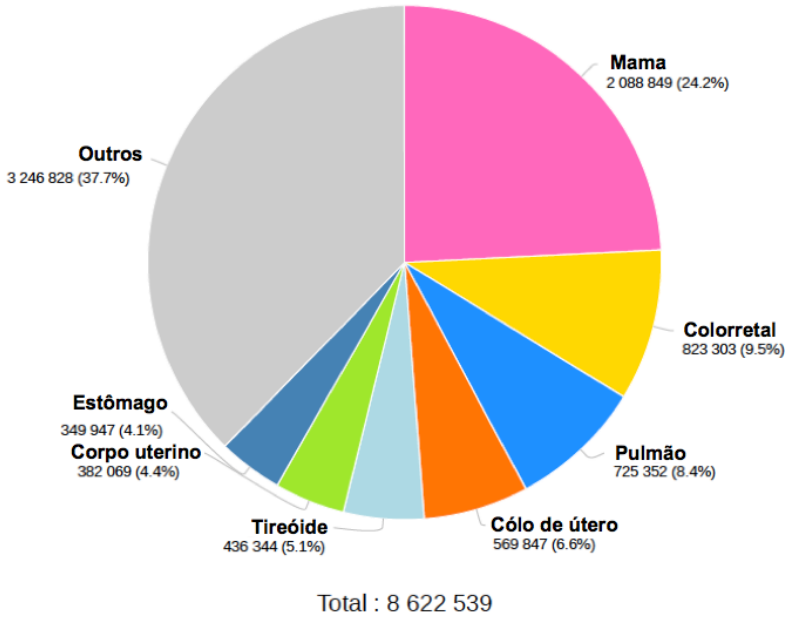
CONCLUSION: The implementation of a data Analysis platform by visual programming in a Research center is a complex task that demands computational structure and specialized staff. However, it's also necessary given that Galaxy has proven to be a very useful platform on the study analysis as well as on offering tools for genomic analysis on diagnosis department, providing a facilitated, reliable and reproducible form of analysis for researchers without in-deep knowledge of bioinformatics. The methylome analysis presented new biomarkers focused on the Brazilian population, providing new findings to understand this disease.

KEYWORDS: Uterine Cervical Neoplasms, DNA Methylation, exome

1 INTRODUÇÃO

1.1 Incidência estimada, mortalidade e prevalência do câncer de colo de útero

O câncer de colo de útero é o quarto câncer mais comum em mulheres, com estimativa de 570.000 novos casos em 2018 e a quarta causa de morte por câncer em mulheres. A maioria dos casos, aproximadamente 85%, ocorre em regiões menos desenvolvidas, representando quase 12% dos cânceres em mulheres[1]. No Brasil, é o terceiro câncer mais comum em mulheres segundo o INCA, com incidência estimada de 16.370 casos para o ano-biênio de 2018/2019[2]. Na Figura 1 encontra-se representada a incidência mundial dos principais tipos de câncer em mulheres, segundo estimativas da Agência Internacional para pesquisa em Câncer (International Agency for Research on Cancer - IARC)[3].

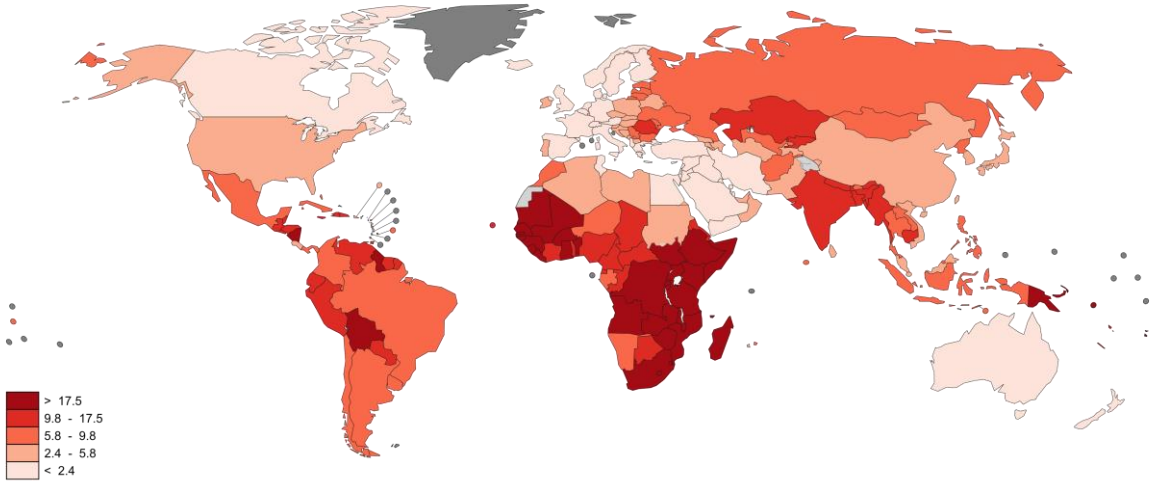


Fonte: Adaptado de *Population Fact Sheets*[3].

Figura 1 - Incidência dos principais tipos de câncer em mulheres no mundo segundo a IARC.

Em relação a mortalidade por câncer de colo de útero em 2018, foram estimados mais de 311.000 casos no mundo, sendo a segunda causa de morte por câncer em mulheres.

A grande maioria dessas mortes, cerca de 90%, ocorrem em países de baixa e média renda, e poderia ser reduzida com uma abordagem que combinasse prevenção, diagnóstico precoce, rastreamento efetivo e programas de tratamento[4]. A Figura 2 mostra a estimativa da mortalidade de câncer de colo de útero no mundo.



Fonte: Retirado de *Globocan*[3].

Figura 2 - Estimativa mundial da mortalidade por câncer de colo de útero (por 100.000 casos).

Mostrando a efetividade de programas de prevenção podemos citar os Estados Unidos, que já tiveram esse tipo tumoral como a causa de morte mais comum em mulheres, mas ao longo dos últimos 40 anos essa taxa diminuiu mais de 50%, sendo a razão principal dessa mudança o aumento dos testes de Papanicolau[5]. Esse procedimento de rastreamento pode encontrar alterações no colo do útero antes do desenvolvimento do câncer, ou seja, em estágios mais curáveis. Esse tipo tumoral tende a ser encontrado, na maioria, em mulheres antes dos 50 anos e raramente em mulheres com menos de 20 anos[6]. Em mulheres que realizam regularmente tais exames preventivos, a detecção de câncer de colo de útero é menos frequente[5]. Como importante agente causal, o vírus do papiloma humano (HPV) é considerado causa primária de cânceres uterino-cervicais[7].

A prevalência global de HPV em câncer de colo de útero em 22 países é de 99,7%[6,7]. Para controle da doença, existem programas de prevenção primária e secundária. A prevenção primária visa diminuir o risco de contágio pelo HPV. Os programas de rastreamento estão na categoria de prevenção secundária e seu papel é o de detectar precocemente lesões pré-

cancerígenas[5]. Dentro desse contexto, o uso de preservativo trata-se de uma forma de prevenção primária, mas a principal forma é a vacina contra o HPV[6,7].

Os programas de vacinação são oferecidos em alguns países. No Brasil, o Ministério da Saúde implementou em 2014 no calendário vacinal feminino, a vacina tetravalente contra o HPV e em 2017 para indivíduos do sexo masculino. O grupo etário é de 9 a 14 anos dado que ela é mais eficaz antes do início da vida sexual, dado seu caráter preventivo. A meta é que seja vacinada 80% da população-alvo para que a incidência desse câncer diminua nas próximas décadas no país[6].

Apesar dos importantes avanços nas formas de prevenção, até que uma grande mudança no cenário mundial ocorra, o câncer de colo de útero ainda é um problema de saúde pública que deve ser alvo de novos estudos, especialmente em casos avançados detectados tardiamente.

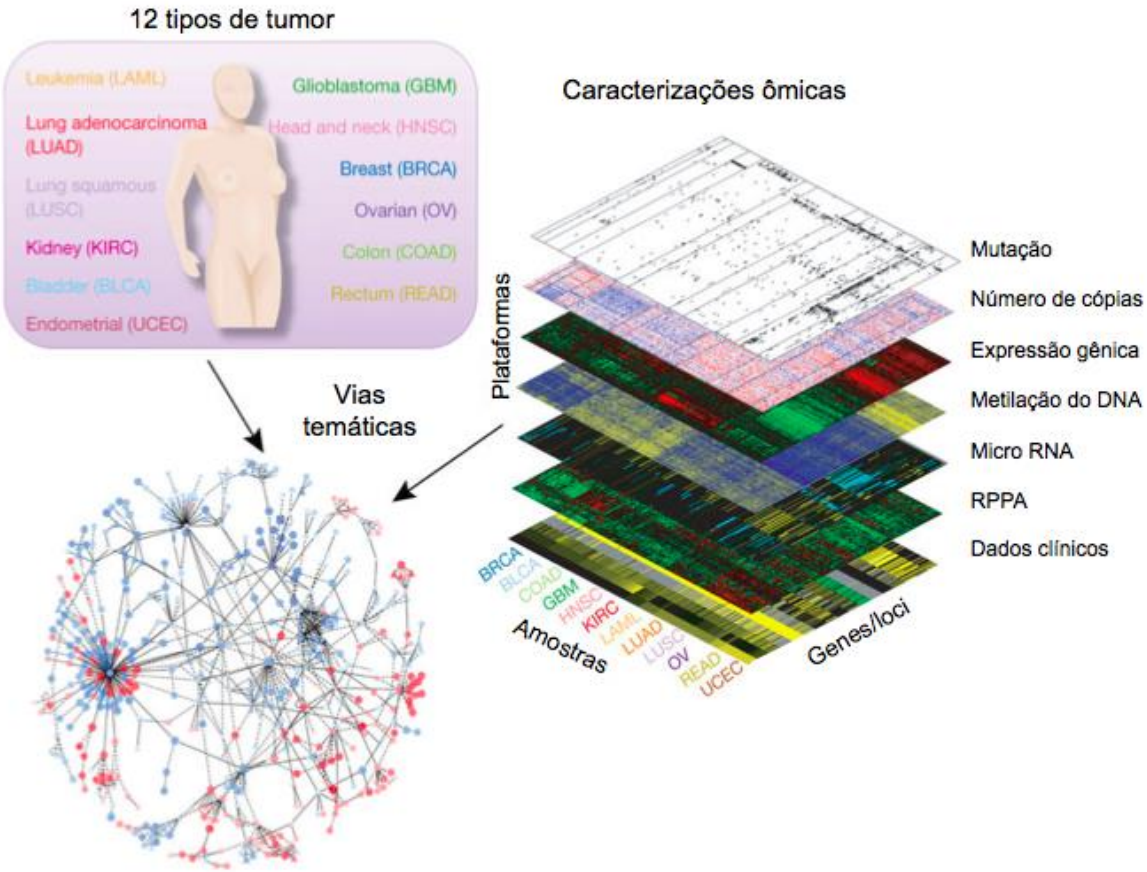
1.2 Classificação do câncer de colo de útero e a importância de consórcios internacionais para a caracterização da doença

O câncer de colo de útero caracteriza-se por ser precedido por uma longa fase de doença pré-invasiva, chama de neoplasia intraepitelial cervical (NIC)[8]. A categorização dessa neoplasia é mensurada em graus, sendo I, II e III dependendo da proporção da espessura do epitélio que apresenta células maduras e diferenciadas[6]. Os graus mais avançados de NIC (II e III) têm maior probabilidade de progressão do câncer[9]. Todos os tipos de NIC são consideradas lesões precursoras, e são classificadas em ordem de gravidade porque a maioria das lesões NIC I regridem entre 12 a 24 meses ou não progridem em NIC II ou III [5]. A lesão precursora que se origina do epitélio colunar é denominada de adenocarcinoma *in situ* (AIS). As lesões intraepiteliais podem ainda ser classificadas como baixo grau (L-SIL) ou NIC1 e alto grau (H-SIL) ou NICII/NICIII[8,10]. A nomenclatura dos exames citopatológicos utilizada no Brasil foi baseada no Sistema Bethesda (2001) e, para os exames histopatológicos, é utilizada a nomenclatura de Richart[9], a classificação de neoplasias intraepiteliais cervicais citada acima. Apesar das possibilidades de classificação por citologia e da possibilidade de testar para HPV como parte dos programas de rastreamento, existe uma busca pela caracterização molecular que possibilite melhores formas de diagnóstico, bem como de prognóstico e tratamento.

Grandes avanços têm sido obtidos quanto ao entendimento das bases moleculares do câncer de colo de útero[11,12,13]. Isso é proporcionado pelos consórcios internacionais em câncer, com foco no genoma destes, que têm utilizado técnicas moleculares em larga-escala, gerando uma grande quantidade de informação. Grande parte dessa informação encontra-se disponível publicamente, permitindo que novas abordagens de análise sejam aplicadas[14].

Muitos dados desses projetos já estão sendo reanalisados[14]. A ideia central da reanálise é permitir que pesquisadores de diferentes áreas possam gerar novos resultados a partir da grande quantidade de dados gerados. No contexto da genômica do câncer, encontram-se principalmente os bancos de dados *Oncomine*[15]*, Gene Expression Omnibus* (GEO)[16], *Tumorscape*[17] e, mais recentemente, o *The Cancer Genome Atlas* (TCGA)[18] e o *International Cancer Genome Consortium (ICGC)*[19].

O TCGA reuniu dados de milhares de pacientes com tumores primários ocorrendo em diferentes sítios do corpo. No projeto todo, foram 6 diferentes tipos de dados gerados (exoma, número de cópias, expressão gênica, metilação, microRNA e RPPA) a partir de análises diferentes do material tumoral em comparação com o sangue[20]. As camadas dos diferentes tipos de dados gerados são representadas na Figura 3.



Fonte: Adaptado de [20]

Figura 3 - Representação das estratégias de análise do projeto TCGA, no qual diferentes tipos tumorais são avaliados e caracterizados por meio de diferentes conjuntos de dados ou "camadas" para uma compreensão integrada dos mesmos.

Análises integradas e complexas desse tipo fornecem grandes possibilidades de conhecimento da biologia tumoral, assim como novos métodos de classificação, a busca de novos biomarcadores e um entendimento mais aprofundado dos mecanismos moleculares envolvidos. No entanto, exigem grande poder computacional, armazenamento e profundo expertise em bioinformática[21].

1.3 Dados moleculares de câncer de colo de útero fornecidos pelo The Cancer Genome Atlas (TCGA)

Os estudos sobre do câncer têm mostrado que se trata de uma doença molecular complexa, onde um vasto conjunto de alterações moleculares levam à transformação de uma célula normal e uma neoplasia agressiva, através da aquisição de proliferação aumentada,

inibição de apoptose, indução de angiogênese, invasão e metastização, alteração de metabolismo, escape imunológico, instabilidade genética e imortalização celular[22]. Dentre os diferentes eventos moleculares considerados críticos para o processo de carcinogênese, ou também conhecidos como *drivers*, estes podem ser agrupados em duas classes principais em relação ao aumento ou diminuição da atividade do produto gênico: os proto-oncogenes (cujas mutações, fusões ou amplificações ocasionam o ganho de função, gerando os oncogenes) e os genes supressores tumorais (cujas mutações, deleções, ou mecanismos epigenéticos leva à perda de função que contribuir para o desenvolvimento tumoral) [23].

As alterações que levam ao desenvolvimento do câncer podem ser identificadas atualmente, de forma sistemática, por análise genômica. Dentre as estratégias disponíveis, pode-se realizar o sequenciamento completo do genoma tumoral, o sequenciamento da região codificante dos 21 mil genes humanos que codificam proteínas (o exoma). O avanço destas análises, tem permitido a distinção de mutações responsáveis pela patogênese da doença (mutações *drivers*) de outras mutações (*passengers*). A análise do genoma completo permite a identificação de todos as regiões que contenham mutações, deleções ou duplicações, e rearranjos estruturais complexos do seu genoma. Já o exoma tem como foco mutações que alterem a sequência de aminoácidos do produto. Além disso pode-se monitorar alterações epigenéticas, alterações nos níveis de expressão gênica por análise de mRNAs, microRNA e alterações proteicas (RPPA), e nessas abordagens, em geral, se compara células cancerosas e normais (como controle) idealmente sendo as não cancerosas do mesmo tecido e do mesmo paciente (Figura 3)[23].

Essa integração de dados moleculares, juntamente com a história clínica do paciente, fornece novas informações que têm contribuído para um melhor diagnóstico, prognóstico e resposta terapêutica oncológica em geral, e de colo uterino em particular[24].

O TCGA iniciou suas atividades em 2009, financiado pelo Instituto Nacional de Câncer (do inglês *National Cancer Institute* – NCI) e pelo Instituto Nacional de Pesquisa em Genoma Humano (do inglês *National Human Genome Research Institute* – NHGRI), ambos parte da Agência Nacional de Saúde dos Estados Unidos (do inglês *National Institute of Health* - NIH). O projeto iniciou como piloto e o objetivo era sequenciar 20 tumores. Essa parte inicial custou 50 milhões de dólares e teve duração de 3 anos. Após a parte inicial, o projeto contou com o investimento de mais de 175 milhões de dólares de fundos de investimentos (*American*

Recovery and Reinvestment Act Funds)[25], gerando aproximadamente 2,5 *petabytes* de dados de 11.000 amostras tumorais[18].

Em termos comparativos, 1 *peta byte* seria o equivalente a 20 milhões de armários de arquivo de 4 gavetas cheios de textos ou, se fossem fotos, 1,5 petabytes podem armazenar 10 bilhões de fotos, por exemplo[26]. O desafio para análise desses dados tem impulsionado a melhoria da infraestrutura, geração de novas ferramentas bioinformáticas, assim como novos métodos de análise. A integração de diversos tipos de análises (com foco no DNA, RNA, metilação, microRNA e proteínas) em larga-escala fornece uma visão das bases moleculares do câncer, considerando sua característica complexa, caracterizada por uma diversidade de alterações genéticas e epigenéticas[14].

Os dados fornecidos pelo TCGA em colo de útero apresentam exomas de 192 amostras (54 de pacientes brasileiras) pareadas com sangue, em um total de 43.324 mutações somáticas, incluindo 24.551 *missense*, 2.470 *nonsense*, 9.260 sinônimas, 5.841 não-codificantes, 535 em sítios de *splice*, 74 *nonstop*, 475 *frameshift indels* e 118 *in-frame indels*. Dentre as amostras analisadas, onze apresentaram frequências de mutações *outliers* (> 600/amostra), sendo classificados como hipermutadas[11]. A análise realizada no citado trabalho para identificação de genes *drivers*, utilizando o algoritmo *MutSig2CV* identificou 14 genes com *qvalue* <=0,1. Dentre os drivers identificados, destacam-se as mutações somáticas nos genes *SHKBP1*, *ERBB3*, *CASP8*, *HLA-A* e *TGFBR2* como inéditas nesse tipo tumoral, assim como mutações conhecidas em *PIK3CA*, *EP300*, *FBXW7*, *HLA-B*, *PTEN*, *NFE2L2*, *ARID1A*, *KRAS* e *MAPK1*[11,12].

Tal como previamente descrito, o TCGA além de análise de mutações (exoma), fornece informações sobre outros tipos de dados moleculares. A figura 4 mostra a visualização, de forma integrativa, de:

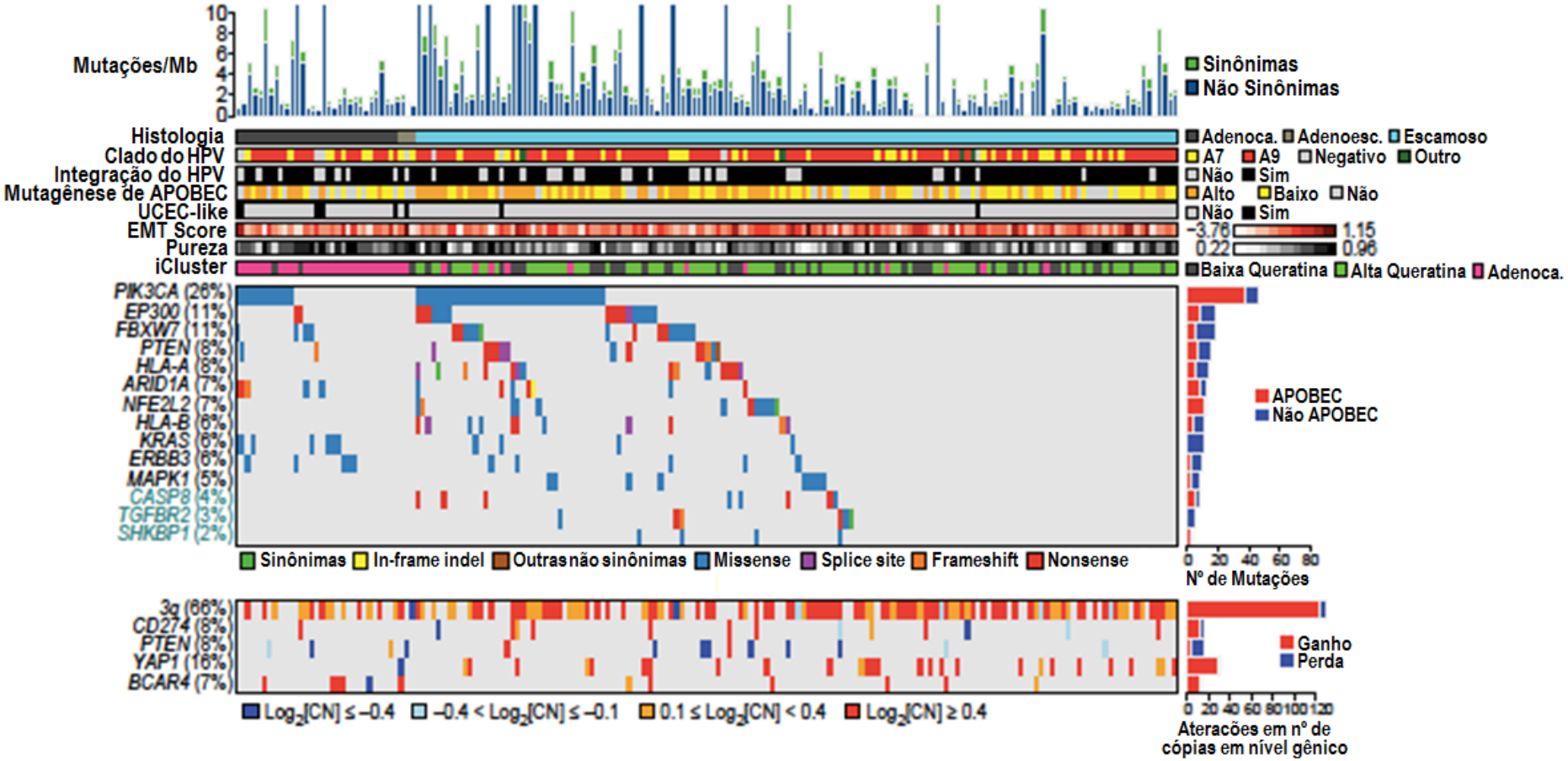
A) Número de mutações por megabase do genoma codificante. Ou seja, quantas mutações estão representadas em uma janela de 1.000.000 pares de base.

B) Dados clínicos (histologia, tipo e integração do HPV no DNA), dados moleculares (padrão de hipermutação, padrão similar a outro tipo de tumor (*UCEC-like*, do inglês *Uterine corpus endometrial*)), classificações (status de transição epitélio-mesenquimal e pureza da amostra) e agrupamento molecular de vários tipos moleculares por um algoritmo integrativo (iCluster).

C) Frequência de mutação dos 14 genes *drivers* identificados e categorizados conforme sua classificação molecular (mutação sinônima, inserção/deleção *in-frame*, *missense*, de sitio de

splice, *frameshift*, *nonsense* e outras mutações não sinônimas). Encontram-se destacados os genes drivers identificados de forma inédita pelo TCGA, assim como o número de mutações associadas ao padrão da enzima APOBEC, uma enzima da família das citidinas deaminases que convertem citosinas em uracilas durante o processo de edição de RNA, conferindo um padrão especifico de mutação. Acredita-se que essas enzimas estejam associadas à hipermutação em alguns casos, um mecanismo conhecido como *kataegis*[27].

D) Alterações no número de cópias do genoma dos genes identificados por um algoritmo específico que permite a identificação de regiões mais frequentes de alterações de número de cópias do DNA (GISTIC2.0 - Version 2.0.22). Encontra-se representado em cores as faixas por perdas e ganhos em escala logarítmica de base 2, que são: menor ou igual a -0,4 (perda de duas cópias do DNA), entre -0,4 e -0,1 (perda de uma cópia do DNA), entre 0,1 e 0,4 (ganho de uma cópia do DNA) e maior ou igual a 0,4 (ganho de duas ou mais cópias do DNA). A frequência global de ganhos e perdas encontra-se também destacada no painel à direita.



Fonte: Adaptado de [11].

Figura 4 - Alterações somáticas em câncer de colo de útero e associações com características moleculares. As amostras são representadas em colunas e ordenadas por taxa de mutação (A), características clínicas e moleculares (B), genes significativamente mutados conhecidos como *drivers* (C) e alterações de número de cópias do DNA (D).

Além dos novos achados moleculares e integrativos, os dados fornecidos pelo TCGA permitiram a identificação de grupos especialmente relacionados com o padrão de expressão por mRNAs:

A) Grupo escamoso com alta expressão de genes membros da família de queratinas (identificado como *keratin-high*),

B) Grupo escamoso com baixa expressão de genes de queratina (identificado como *keratin-low*);

C) Grupo com predominância de adenocarcinoma (identificado como adenocarcinoma).

Os grupos em questão foram obtidos por análise integrativa de diversos conjuntos de dados (alteração do número de cópias, metilação, mRNA e miRNA), utilizando a ferramenta *iCluster*. Os agrupamentos consensos identificados foram associados com o padrão de expressão de transcritos da família de queratinas. A relevância clínica desses achados não foi determinada. No entanto, a caracterização de subgrupos dentro de um tipo tumoral é de importância para a compreensão do comportamento do mesmo, bem como dos mecanismos moleculares associados[12].

Além disso, é possível correlacionar a ocorrência de regiões amplificadas por análise de número de cópias do DNA com sítios de integração de vírus. Ojesina e colaboradores encontraram principalmente HPV16, HPV18 e HPV52 integrados ao DNA das células tumorais. Finalmente, análises de dados de expressão mostraram que alterações nos genes *MYC*, *ERBB2*, *GLI2*, *TNIK*, *NR4A2*, *PROX1*, *EIF2C2*, *FAM179B*, e *SERPINB4* podem estar associadas com regiões de perdas e ganhos do DNA, assim como sítios de inserção de vírus[12].

Dentro desse contexto, o Hospital de Câncer de Barretos (HCB) participou com um total de 54 amostras de câncer de colo de útero para o projeto TCGA. Dessa forma, podemos caracterizar as amostras da população brasileira dentro desse consórcio.

1.4 Programação visual

Além dos avanços por meio dos achados moleculares, a grande quantidade de dados oriundos das “ômicas” tem criado a necessidade de utilização de bancos de dados e da programação computacional. Tal situação representa um problema, visto que necessita de uma complexa e custosa infraestrutura de computação e existe uma deficiência de profissionais com experiência computacional nessa área. Em paralelo, tem surgido a

representação de linguagens computacionais de forma visual, por meio de *workflows.* Pode-se definir *workflows* como fluxos de dados estruturados e resumidos que ajudam os usuários a construir uma série de passos de maneira organizada[21].

Assim, pode-se aplicar diversas linguagens de programação em um ambiente intuitivo, por meio de nós pré-computados. O método de análise em formato de fluxo de trabalho ou programação visual tem sido usado em diversas áreas da ciência, com novas aplicações na bioinformática[28]. Para esse projeto, foi selecionada uma ferramenta de grande destaque para análise genômica, o *Galaxy*[29].

1.5 *Galaxy*

O projeto *Galaxy* começou em 2005 com o objetivo principal de permitir que pessoas da área de biológicas, sem especialização em programação e administração de sistemas, desenvolvam análises computacionais através de interface web[30]. O Galaxy foi desenvolvido como um conjunto de componentes de software separados que trabalham juntos para executar tarefas. O componente central coordena a ação, executa a tarefa e mantém registro dos históricos de atividades do usuário enquanto a interface do usuário e operação/ferramenta/biblioteca de saída são implementados separadamente. Toda a comunicação com outros sites (*UCSC table browser*, dentre outros) é feita pelo componente central. Os benefícios desse arranjo incluem extensibilidade (facilidade de adição de novas ferramentas e interfaces) e divisão conveniente de trabalho e expertise de programadores. A interface de usuário se comunica com o componente central via requisição de HTTP (Web). O componente central também fornece um API (do inglês *application program interface*, uma interface de comunicação de aplicativos) para coordenar as tarefas executadas em diferentes websites, permitindo o gerenciamento de diversas interfaces de usuários[31]. O componente central e bibliotecas operacionais são escritas em C e construídos nos padrões do grupo de bioinformática do UCSC (*University of California Santa Cruz*), para facilitar a comunicação com o table browser da citada Universidade. As comunicações por requisição de HTTP facilitam a comunicação entre os sistemas via API.

O projeto é composto por diversos componentes: um servidor público (acessível em https://usegalaxy.org) e o instalador local do software. O site está disponível desde 2007 e

suporta múltiplos usuários e diversas tarefas por mês. Além disso, trata-se de uma ferramenta altamente customizável e se integra com uma variedade de ambientes computacionais, podendo ser usados em computadores portáteis, clusteres e computação em nuvem. O *Galaxy ToolShed* facilita o compartilhamento de ferramentas entre instalações centralizando-as e permitindo que desenvolvedores carreguem nesse ambiente as suas configurações de ferramentas[32].

Desse modo, diversas ferramentas de análise de bioinformática estão disponíveis gratuitamente no *Galaxy*, facilitando o processo de análise de dados[33]. É uma plataforma que também permite a integração de diversas ferramentas para geração de *workflows* reprodutíveis[33,34]. Também pode ser instalado localmente ou em servidor, o que permite a instalação de ferramentas pré-computadas disponíveis em uma biblioteca conhecida como *ToolShed*[35], além da criação de *workflows* próprios. Os *Workflows* correspondem a integração de diversas ferramentas em etapas consecutivas de análise que permite a replicação exata da mesma, na qual altera-se somente o dado importado[31]. Conta ainda com uma customização que pode ser feita com a integração do Jupyter notebook, que permite o uso de diversas linguagens de programação dentro do ambiente seguro e acessível via internet do *Galaxy*, permitindo refinamento das análises, caso necessário[36].

Trata-se assim de uma plataforma customizável que pode ser aplicada a diferentes tipos de análises complexas de bioinformática, facilitando as mesmas e a sua reprodutibilidade permite que seja aplicada a vários estudos.

2 JUSTIFICATIVA

O consórcio TCGA revolucionou o entendimento biológico do câncer através da análise integrada de milhares de tumores[20]. Para o máximo aproveitamento da grande quantidade de dados gerados faz se necessário o uso de complexas ferramentas de bioinformática que integrem e analisem os resultados. Uma das principais barreiras atuais é a elevada complexidade das análises e a necessidade de profissionais com aprofundado conhecimento de bioinformática.

Dessa forma, existe a necessidade do desenvolvimento de ambientes simplificados onde pesquisadores sem conhecimento aprofundado de programação computacional possam gerar suas hipóteses e analisar seus resultados. Para esse trabalho, a análise epigenética foi escolhida como um exemplo de análise porque além de sua relevância biológica, produz como arquivos brutos pequenos, possibilitando uma análise desde o início sem demandar tanto processamento computacional.

3 OBJETIVOS

1.5 Objetivo geral

Esse estudo tem por objetivo a implementação de métodos da programação visual (plataforma Galaxy) para análise dos dados do TCGA, com foco nas amostras de pacientes brasileiras.

1.6 Objetivos específicos

1 Instalação e adequação das ferramentas para geração de *workf*lows por programação visual para análise de dados em larga escala oriundos do projeto TCGA;

2 Solicitação e análise de dados do TCGA de sequenciamento de DNA, metilação e expressão gênica (RNAseq);

3 Identificação de biomarcadores epigenéticos de câncer de colo de útero.

4 MATERIAIS E MÉTODOS

4.1 Desenho do estudo, população e amostragem

Trata-se de um estudo observacional com coleta de dados retrospectiva, utilizando dados do banco de dados público do consórcio internacional chamado *The Cancer Genome Atlas* (TCGA).

Serão verificados os desfechos das participantes de pesquisa oriundas do Hospital de Câncer de Barretos. No portal do TCGA é possível obter os dados pré-processados. No entanto, realizamos a análise a partir dos dados brutos, visando montar um fluxograma de análise em etapas para todos os passos. A atualização do seguimento das pacientes (*status* da paciente) foi realizada utilizando dados do Registro Hospitalar de Câncer do Hospital de Câncer de Barretos.

A população do estudo inclui pacientes com câncer de colo de útero incluídas no estudo do consórcio internacional TCGA, sendo a amostra composta por dados de 54 mulheres do TCGA oriundas da população brasileira fornecidas pelo Hospital de Câncer de Barretos. Além disso 3 amostras controles e vinte pacientes com câncer de endométrio, adenocarcinoma do tipo endometrióide com estádio II também foram incluídas para comparação e controle de qualidade.

4.2 Critérios de Inclusão e aspectos éticos

Para cumprir os critérios de inclusão a participante precisa ter dados disponíveis no banco de dados do consórcio TCGA.

O presente protocolo de pesquisa foi desenvolvido conforme as normas da Resolução CNS 466/12 do Conselho Nacional de Saúde/Ministério da Saúde (CAAE 60545416.7.0000.5437) e foi aprovado pelo Comitê de Ética e Pesquisa (CEP) da Fundação Pio XII HCB sob o número 1261/2016.

Riscos ao participante: Apesar de trabalhar com dados de sequenciamento genético, o objetivo é analisar dados de mutações somáticas e não há risco de encontrar, acidentalmente,

qualquer alteração genética que denote necessidade de aconselhamento genético. Desse modo, o risco é mínimo e representado apenas pela quebra de sigilo dos dados do participante, dados estes que serão protegidos pelo pesquisador.

Benefícios ao participante: a pesquisa não trará nenhum benefício imediato ao participante de pesquisa. Os possíveis benefícios estão relacionados as informações que essa pesquisa poderá trazer a outras pacientes no futuro.

4.3 Obtenção dos dados brutos do TCGA

Os dados do TCGA passaram por diferentes etapas de controle de qualidade. Devido aos critérios altamente restritivos, algumas amostras foram consideradas inadequadas para o projeto, e por isso foram excluídas do projeto (conhecemos apenas o número final de amostras que foram incluídas no projeto, as 307 amostras de câncer de colo de útero, dentre as quais as 54 brasileiras). Mesmo após aceito no consórcio, os dados disponíveis para um participante podem não ser completos. Um exemplo do que pode ocorrer é ter disponíveis os dados de exoma e metiloma pois o DNA estava de boa qualidade e não ter RNAseq, pois foi detectada degradação do RNA dessa amostra, dentre outras variações[11]. Dessa forma, o primeiro passo para analisar dados desse consórcio de um conjunto de amostras de interesse é avaliar a disponibilidade de cada tipo de dado das mesmas.

O processo de requisição para obtenção dos dados brutos do TCGA consiste inicialmente em realizar o cadastro da instituição (no caso, o Hospital de Câncer de Barretos) no eRA Commons[37], uma interface *online* da Agência Nacional de Saúde Norte Americana. Todos os envolvidos na solicitação dos dados devem estar cadastrados. Com esse cadastro, o pesquisador responsável faz a requisição de acesso de dados no dbGAP[38,39]. Após isso, é necessário que o responsável pela instituição acesse o dbGAP com o login do eRA Commons e aceite a requisição de acesso de dados feita pelo pesquisador responsável. Esse processo é feito não apenas para o TCGA como para outros projetos que foram financiados pelo NIH e cada projeto tem um comitê que avalia vários quesitos[40].

Os arquivos possuem um identificador único, o UUID (do inglês *universally unique identifier*)[41]. Apesar de diversas possibilidades de filtros fornecidas dentro dessa interface, é recomendado o filtro por UUID. A plataforma online do TCGA permite a realização desses

filtros de forma simples e também disponibiliza uma ferramenta de *download* de dados, a GDC *Data Transfer Tool*.

4.4 Análise de dados utilizando o Galaxy

A plataforma *Galaxy* permite a criação de *workflows* de análise. Dessa forma, cada etapa de processamento é uma unidade isolada que contém uma ferramenta de bioinformática já disponível ou comandos em uma linguagem de programação que pode ser inserido localmente. O arquivo de saída de cada unidade é a entrada da próxima, ou seja, os passos são conectados. Esse processo permite a visualização global de cada processo de análise e o controle da versão de cada ferramenta, garantindo a reprodutibilidade. A Figura 5 mostra um exemplo de comandos que são executados automaticamente em um passo de análise no *Galaxy,* ou seja, o resultado de todo um processo de implementação. Se a ferramenta fosse utilizada regularmente, ela precisaria ser instalada no servidor usando comandos de instalação e posteriormente, comandos de análise. Ao incluir a ferramenta no *Galaxy* o usuário final não precisa fazer a parte de instalar e programar a ferramenta. Além disso, o *Galaxy* salva todos os metadados, garantindo reprodutibilidade.

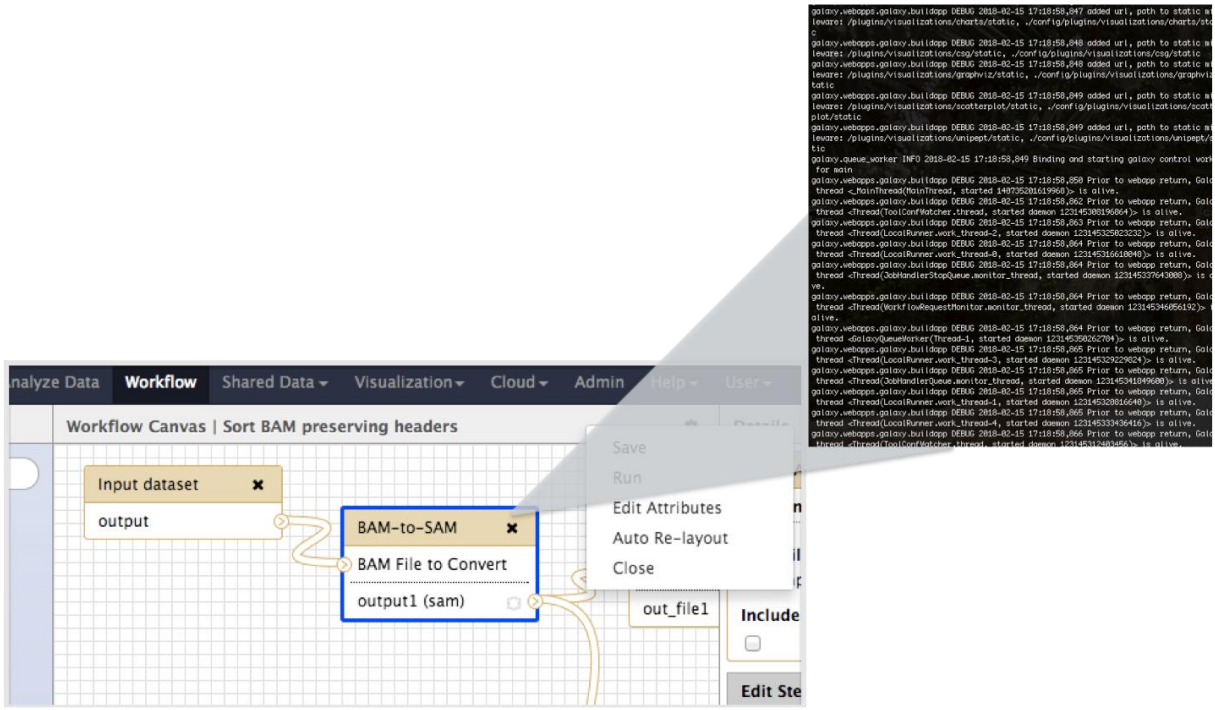


Figura 5 - Representação esquemática dos comandos implícitos em uma unidade de análise dentro de um *workflow*, na plataforma *Galaxy*.

Desse modo, os passos dos *workflows* diferem das etapas de análise (algoritmos), na qual esses últimos representam um processo inicial de organização. Dessa forma, foram criados algoritmos utilizando o software Bizagi modeler da Microsoft[42], para melhor compreensão e planejamento de cada etapa a ser implementada para construção dos *workflows*.

4.5 Análise de dados de exoma

O algoritmo para análise de dados do exoma encontra-se disponível na Figura 6. Os dados brutos do exoma encontram-se em formato textual (FASTQ) ou em formato binário, previamente alinhado com o genoma de referência (BAM). Dessa forma, a análise deve permitir dupla entrada. Inicialmente, para a importação do FASTQ, deve-se considerar o tipo de FASTQ utilizado, que pode variar de acordo com a plataforma de sequenciamento. Dessa forma, deve ser realizada a conversão por meio da ferramenta *FASTQ groomer* (para converter em formato Sanger & Illumina 1.8+ que é o formato utilizado no presente estudo). Posteriormente deve ser efetuado o controle de qualidade dos dados. Após essa etapa, devem ser retirados os adaptadores (um processo conhecido como trimagem) e os dados também devem ser filtrados por qualidade). O próximo passo inclui o alinhamento das *reads* sequenciadas e filtradas com o genoma de referência utilizando o alinhador *Burrows-Wheeler Aligner* (bwa), seguindo os seguintes parâmetros: seleção de GRCh37 como genoma de referência, modificação dos parâmetros para detecção da biblioteca como *paired-end* e variação dos parâmetros básicos dos algoritmos para teste de performance. O processo de avaliação de SNVs (*Single Nucleotide Variations*) e *Indels* no Galaxy será realizada conforme os seguintes passos: função *Variant Detection and Analysis* para chamada de variantes, seleção do genoma de referência para GRCh37, função *VCF manipulation SNPSift* para aplicação de filtros e função *Annotate Variants* que inclui a anotação das variantes por meio de diversos bancos de dados.

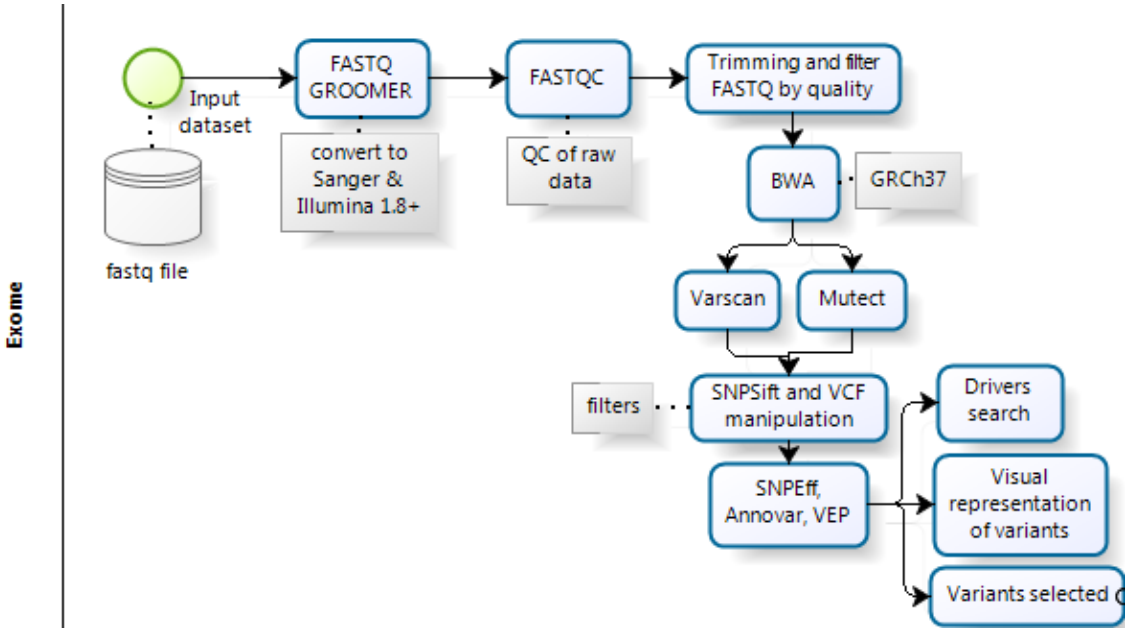


Figura 6 - Algoritmo de análise de exoma para criação do *workflow* usando a plataforma *Galaxy*.

4.6 Análise de expressão diferencial de dados de RNAseq

O input de dados de RNAseq é similar ao de exoma, sendo tanto FASTQ (arquivo texto contendo as *reads* sequenciadas) quanto BAM (arquivo binário, dados alinhados com o genoma de referência) (Figura 7). Dessa forma, o *workflow* deve permitir a entrada de dados brutos e realizar o alinhamento, assim como permitir a entrada de dados previamente alinhados. A etapa de conversão de FASTQ para formato Sanger & Illumina 1.8+ deve ser realizada (*FASTQ Groomer*), assim como a remoção de adaptadores e bases de baixa qualidade (filtragem), conforme Figura 7. A diferença principal é o alinhador, no caso o Tophat2 é adequado para dados de RNAseq. Os principais pacotes para análise estatística diferencial de RNAseq (egdeR e DESeq) não se encontram disponíveis atualmente no *Galaxy*, dessa forma, precisam ser criados, por meio de um processo conhecido como empacotamento, no qual scripts em R precisam ser incluídos como um nó do *workflow*, utilizando a ferramenta Planemo[43,44].

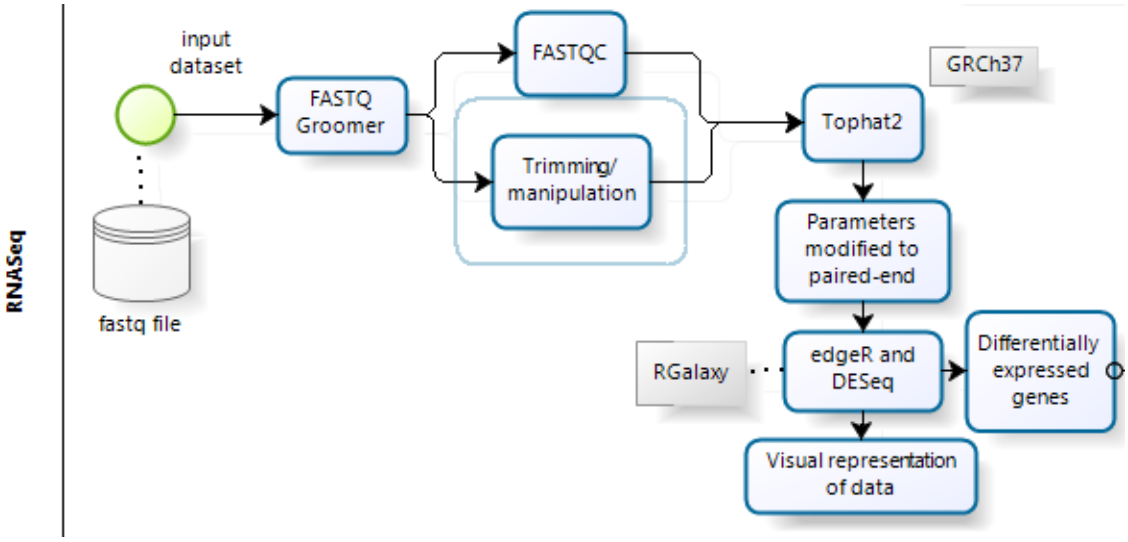


Figura 7 - Algoritmo para análise de RNAseq para criação do *workflow* na plataforma *Galaxy.*

4.7 Análise de dados de metilação

A plataforma *Illumina Infinium HM450 array* (Illumina, San Diego, CA, USA), que inclui *probes* de mais de 480.000 sítios CpG, foi utilizada para geração de dados de metilação do TCGA das amostras selecionadas. A plataforma utiliza dois tipos de ensaios: Infinium I e II os quais utilizam dois tipos de nucleotídeos marcados: nucleotídeos Adenina, marcados em vermelho (*Cy5*), são incorporados em sítios não-metilados enquanto nucleotídeos Guanina, marcados em verde (*Cy3*), são incorporados em sítios metilados. No ensaio *Infinium* I, duas sondas são utilizadas para interrogar o locus CpG, uma sonda para CpG metilado e uma para CpG não metilado. Nesse modelo, a região 3´ da sonda é desenhada para parear com a base citosina (*design* metilado) ou parear coma base timina resultado da conversão por bissulfito e amplificação (*design* não metilado). No ensaio *Infinium* II apenas uma sonda por locus CpG é utilizada, sendo a região 3´ desta complementar a base anterior ao sítio de interesse (C pareado a G) e a extensão resulta na adição de uma guanina ou adenina marcada complementar a citosina (*design* metilado) ou timina (*design* não metilado)[45].

A intensidade de fluorescência de acordo com a base incorporada (nucleotídeos Adenina, marcados em vermelho (*Cy5*), são incorporados em sítios não-metilados enquanto nucleotídeos Guanina, marcados em verde (*Cy3*), são incorporados em sítios metilados) é medida com o escaneamento das lâminas e os dados brutos são no formato ".idat"[45,46]. Dentre os arquivos de saída, encontram-se os *Beta-values*, valores que são utilizados de forma geral

para visualização de agrupamento em formato de heatmaps. Como critério para filtro nos bancos de dados de referência, foi utilizado deltaBeta maior ou igual a 0.3 e que tivesse evidência de atividade de *enhancer*. Antes da montagem do *workflow* de análise, as etapas da análise de metilação foram montadas em um fluxo, conforme Figura 8.

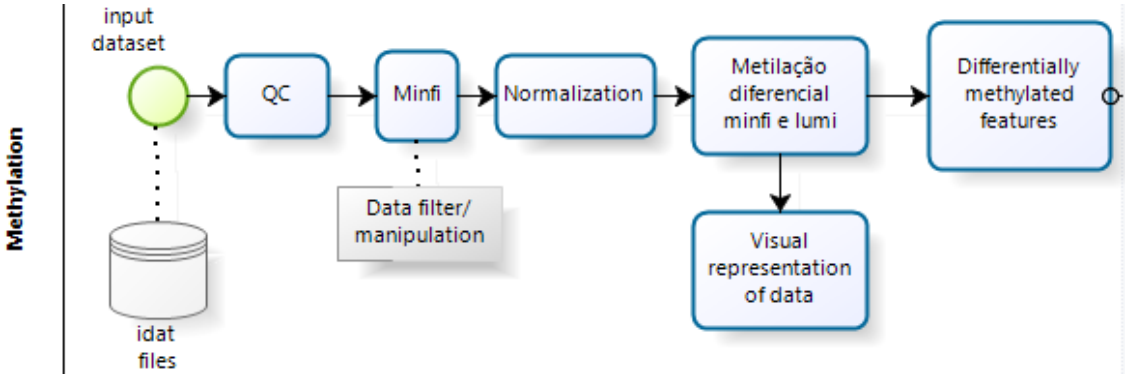


Figura 8 - Algoritmo para análise de metilação para criação do *workflow* na plataforma *Galaxy*

Existe disponível no *Galaxy toolshed*, o repositório de ferramentas do *Galaxy* que permite a instalação de forma facilitada, uma ferramenta em única etapa de *workflow* contendo todas as fases de análise de forma simplificada. No entanto, de acordo com as boas práticas, é recomendado que o processo seja separado em diferentes etapas de análise para melhor controle, ajuste de parâmetros e reprodutibilidade. Para implantar uma ferramenta no *Galaxy* é necessário realizar uma preparação que é chamada de empacotamento e recentemente, foram empacotadas novas funções do pacote minfi por um grupo da Universidade de Bradford[48]. Elas ainda não foram publicadas, mas foram disponibilizadas por meio de colaboração. As funções empregadas no novo *workflow* são resumidas (Anexo C). O processo inclui a conversão em vários formatos de arquivos, que se encontram resumidos na Figura 10. A versão resumida do processo de conversão (Anexo D).

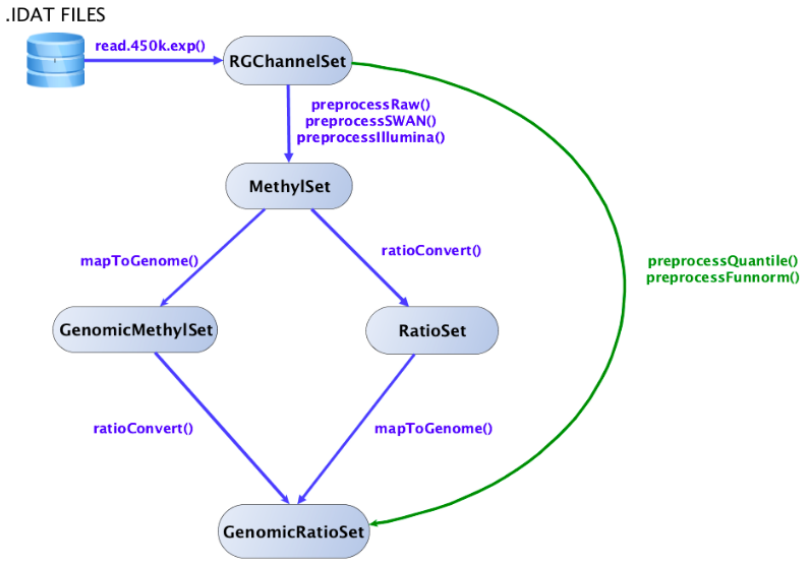


Figura 9 - Objetos gerados e processos de conversão dos mesmo para a realização da análise de metilação.

4.8 Obtenção de dados clínicos

Os dados clínicos do TCGA estão disponíveis no portal GDC (http://portal.gdc.cancer.gov) e foram exportados utilizando o pacote do RTCGA (https://www.bioconductor.org/rtcga). Os dados foram organizados utilizando o software REDCap[47] para posterior seleção e análise estatística, conforme Figura 11.

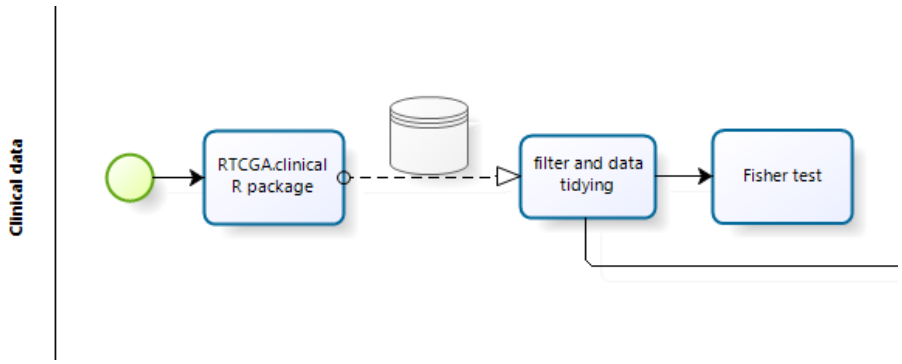


Figura 10 - Algoritmo para exportação dos dados clínicos e consistência.

4.9 Análise estatística dos dados

Foram criados nós contendo funções do R de pacotes específicos para geração de curvas ROC para avaliação de sensibilidade e especificidade de possíveis biomarcadores (identificados a partir de análises prévias) e comparação dos dados moleculares com características clínico patológicas (metástase linfonodal, estadio, radioterapia e prognóstico). Foram incluídas também funções que permitem a comparação entre as alterações obtidas pela população brasileira com os resultados do TCGA global de câncer de colo de útero. É importante ressaltar que não há necessidade de repetir a análise global do TCGA, visto que os dados suplementares da publicação já contemplam as informações necessárias para análises comparativas, ou seja, tabelas com valores quantitativos das alterações encontradas e características clínicas gerais. Os pacotes do R a serem utilizados serão ROCR, stats e survival. Os dados serão analisados pelos testes Exato de Fisher. O nível de significância a ser utilizado é de 5%. A figura 11 apresenta uma visão geral do algoritmo de análise, usando o Bizagi, conforme citado no item 4.4. Ela representa a metodologia proposta antes do início das análises.

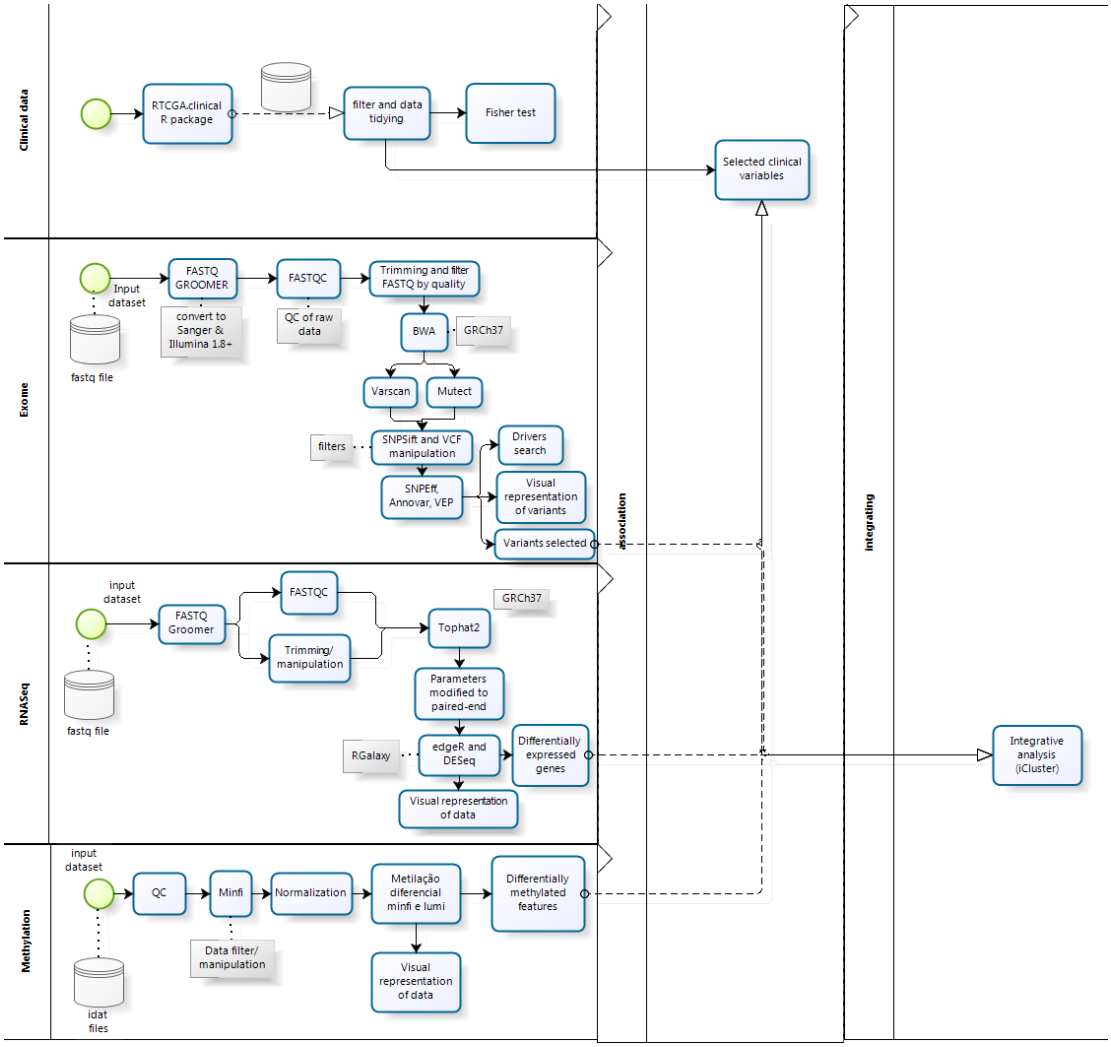


Figura 11 - Algoritmo contendo todas as etapas de análise a serem implementadas na plataforma Galaxy.

4.10 Figuras adicionais

Foram criados nós contendo funções do R de pacotes específicos para criação de gráficos (ComplexHeatmaps e circlize) utilizando a ferramenta Planemo.

5 RESULTADOS

5.1 Implementação do *Galaxy*: Visão geral dos dados do TCGA de colo de útero disponíveis publicamente

Como descrito anteriormente, os dados para determinado participante de pesquisa do TCGA, pode não estar completo pela qualidade de determinado material biológico utilizado na obtenção do dado, dentre outros fatores. A Tabela 1 descreve os dados que atualmente são de livre acesso para cada tipo de análise proposta. O levantamento foi feito filtrando por TCGA-CESC (nome abreviado do projeto de câncer de colo de útero). Alguns dados estão públicos apenas para cultura de células de colo de útero (identificados como CCLE-CESC, outro projeto de cultura de células).

Tabela 1. Arquivos do TCGA disponíveis no portal GDC. Fonte: Retirado do portal de dados do TCGA(50).

Tipo de análise	Número de arquivos	Total de amostras	Projeto	Tipo de arquivo	Tamanho do arquivo	Categoria dos dados
Metilação	624	307	TCGA-CESC	idat	8,10MB	Dado bruto
Exoma	6	305	TCGA-CESC	maf	8-94MB	Dado analisado
RNAseq	24/ 2	24/ 307	CCLE-CESC/ TCGA-CESC	Bam e txt	9-16GB	Dado parcialmente bruto e dado analisado

Após a aprovação para obtenção dos dados controlados (Anexo B), os dados disponíveis da população brasileira foram baixados, conforme relatório obtido pelo REDCap (Research Eletronic Data Capture)[48] e encontram-se apresentados na Tabela 2. Além disso foram baixados também dados dos 3 controles de colo de útero (tecido normal) e dos casos de câncer de endométrio para serem usados no controle de qualidade.

Tabela 2. Dados da população brasileira de câncer de colo de útero do TCGA

Tipo de dado	Formato	N	Min	Max	Média	Total	Mediana
Exoma	BAM	54	18,94 GB	48,35 GB	27,41 GB	1480,25 GB	25,79 GB
Metilação	IDAT	54	8,51 MB	8,51 MB	8,51 MB	920,5 MB	8,51 MB
RNAseq	BAM	53	2,84 GB	10,07 GB	6,12 GB	324,52 GB	6,11 GB

Os dados clínicos do TCGA, como ocorre na maioria dos consórcios genômicos onde o foco é o sequenciamento e compreensão de tais dados, apesar de haver uma grande quantidade de variáveis a serem coletadas pelos centros participantes geralmente se apresentam com baixa completude. No entanto, algumas covariáveis interessantes podem ser apresentadas, separando os grupos HPV positivo e negativo das 54 pacientes que são o foco desse trabalho, conforme pode ser observado na tabela 3.

Tabela 3. Caracterização das 54 pacientes brasileiras quanto ao tipo histológico, raça e Índice de massa corporal entre os grupos HPV positivo e negativo.

		Status de HPV	
		Negativo Frequência absoluta ou média	Positivo Frequência absoluta ou média
Tipo histológico	Carcinoma escamocelular	2	40
	Adenocarcinoma mucinoso endocervical	4	8
Raça	Asiática	1	0
	Branca	3	29
	Negra/afrodescendente	0	4
	Ignorado	2	15
IMC		28,57	25,99

5.2 *Workflow* para análise de dados de exoma

Para a importação do arquivo original em formato fastq, são necessárias duas entradas, uma vez que o sequenciamento do exoma é *pair-ended,* ou seja, sequenciado nas duas direções. Dessa forma, conforme mostrado na Figura 12, a primeira etapa contempla a importação de dois arquivos em formato fastq.

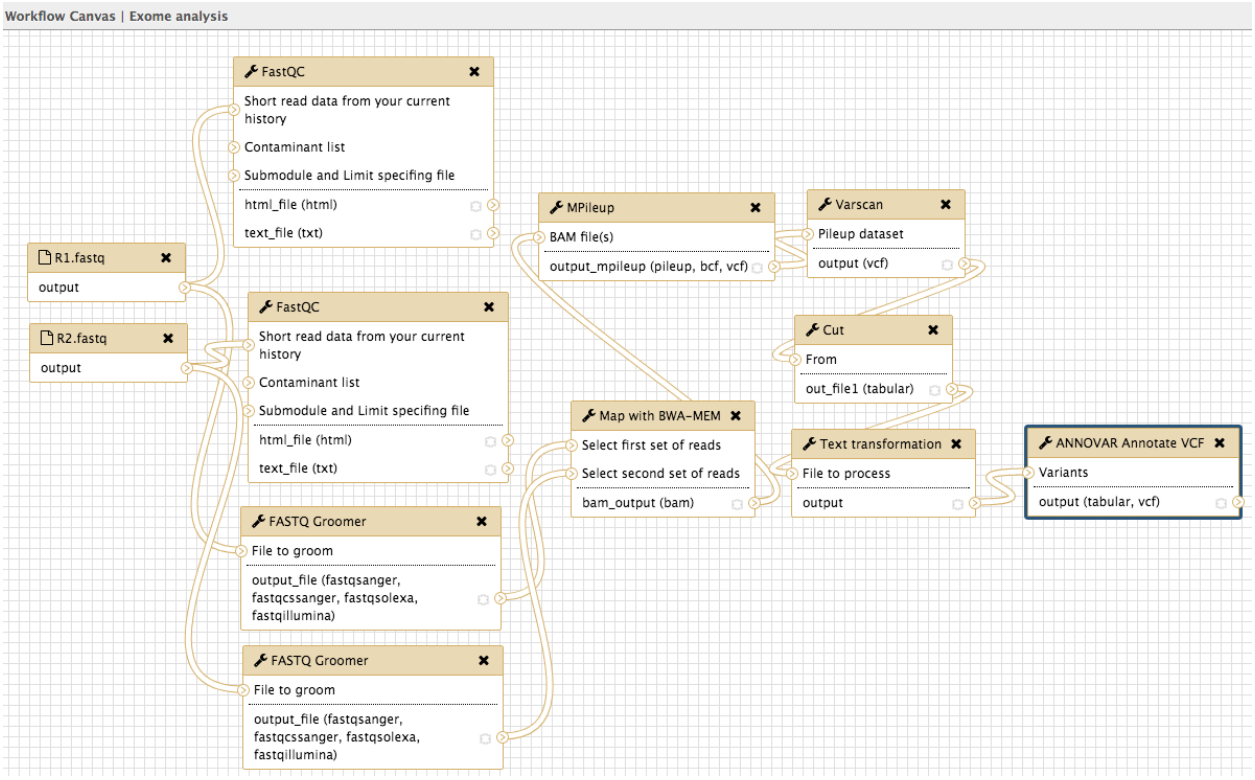


Figura 12 - *Workflow* para análise de dados do exoma de câncer de colo de útero.

De forma sequencial, os arquivos fastq passam pela etapa de controle de qualidade (FASTQC). Essa etapa foi montada de forma a ser possível avaliar a lista de contaminantes e exportar um relatório nos formatos html e txt. Uma segunda saída a partir da caixa de entrada do arquivo fastq permite uma conversão do mesmo para que possa ser realizada a etapa seguinte, de alinhamento com o genoma de referência. Essa etapa de conversão é conhecida como *FASTQ Groomer*. É uma etapa importante pois os arquivos fastqs podem ter diferentes formatos dependendo do tipo de sequenciador - e assegura que o arquivo esteja em formato *fastq illumina 1.8+*.

Em seguida, ambos arquivos são importados na caixa *Map with BWA-MEM,* onde são unificados e alinhados com o genoma de referência na versão GRCh37. O arquivo de saída é único, em formato bam e foi convertido em um arquivo intermediário (mpileup) para que possa ser realizado o processo de chamada de variantes.

O processo de identificação de SNPs (do inglês *Single Nucleotide Polymorfirm*, ou polimorfismo de base única) e *indels* foi configurado para executar a ferramenta Varscan e o arquivo de saída é um arquivo em formato vcf que se trata de um arquivo texto que necessita ser submetido em processos de conversão (caixas *cut* e *text transformation*) para que possa

passar pela etapa de anotação das variantes (caixa *ANNOVAR Annotate VCF*). A etapa de anotação permite caracterizar a variante por integração com bancos de dados disponíveis pela ferramenta Annovar. Essa etapa permite a exportação do arquivo em formato maf (o mesmo que se encontra disponível pelo portal do TCGA).

O presente *workflow* permite a análise de qualquer dado bruto em formato fastq, e pode ser aplicado em qualquer exoma, com ampla aplicação em nosso centro de pesquisa. Ao montar o algoritmo de análise, o planejamento foi analisar o dado bruto, anterior ao alinhamento. No entanto, os dados brutos do TCGA disponibilizados apenas com a aprovação estão no formato BAM (já alinhados – o FASTQ não é disponibilizado). De todo modo, o *workflow* obtido permite partir de um dado ainda não alinhado. Se necessário analisar dados mais processados, como é o caso do BAM, é possível fazer a entrada de dados em outro ponto, como foi feito nesse trabalho.

5.3 *Workflow* para análise de dados de RNAseq

Para análise dos dados de RNA, a primeira etapa é a importação dos dados para o Galaxy e a seguir a checagem da qualidade com a ferramenta FastQC. Do mesmo modo que na análise de exoma, a entrada do arquivo é feita no formato fastq, permitindo uma conversão do mesmo para que possa ser realizada a etapa seguinte, de alinhamento com o genoma de referência (*FASTQ Groomer)*. A saída passa por uma ferramenta chamada Trimmomatic, que faz ajustes na qualidade do arquivo, verificada anteriormente. A seguir é feito o alinhamento com o genoma de referência usando TopHat. São contadas as regiões diferencialmente expressas usando a referência, conforme representado na Figura 13.

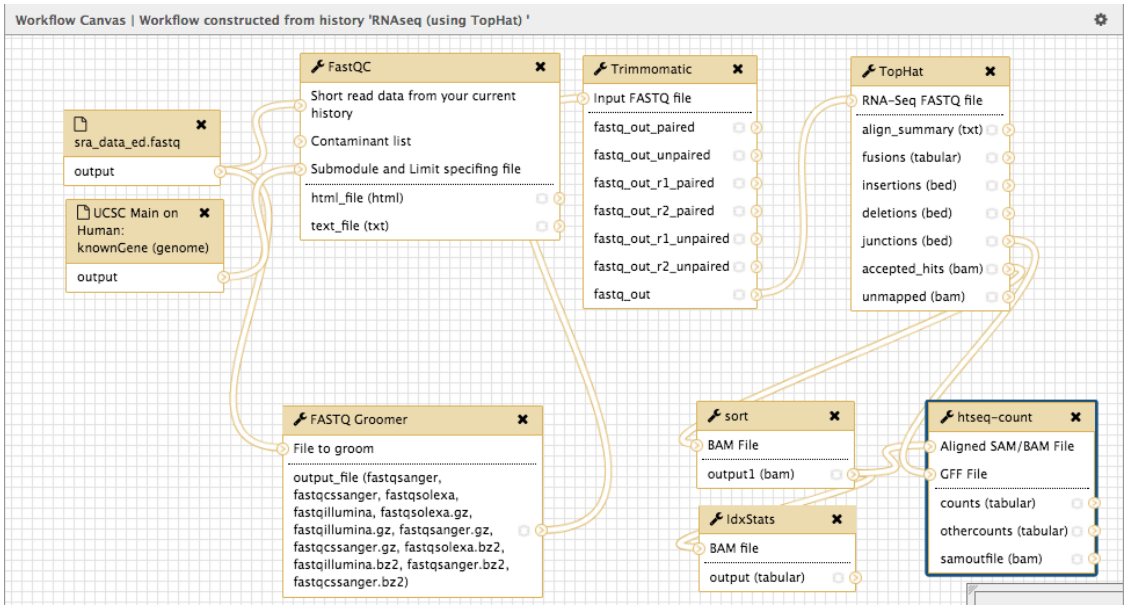


Figura 13 - *Workflow* para análise de dados de expressão gênica de câncer de colo de útero.

5.4 *Workflow* para análise de dados de metilação

O *workflow* da análise de dados da metilação encontra-se representado na Figura 14. Para cada amostra, são importados dois arquivos de dados brutos em formato idat. Em seguida, foram decididas as características para comparação, representada como Treatment/Wildtype. Em seguida, são gerados gráficos de qualidade (qc_report.pdf), gráficos de análise mds (mds_plot.pdf) e dois arquivos textos associados à probes diferencialmente expressas (do inglês, *differential methylated probes* - dmps.csv) e às regiões diferencialmente expressas (do inglês *differential methylated regions* - dmrs.csv). As DMPs correspondem a uma posição específica do genoma, enquanto uma DMR pode conter múltiplas e consecutivas DMPs.

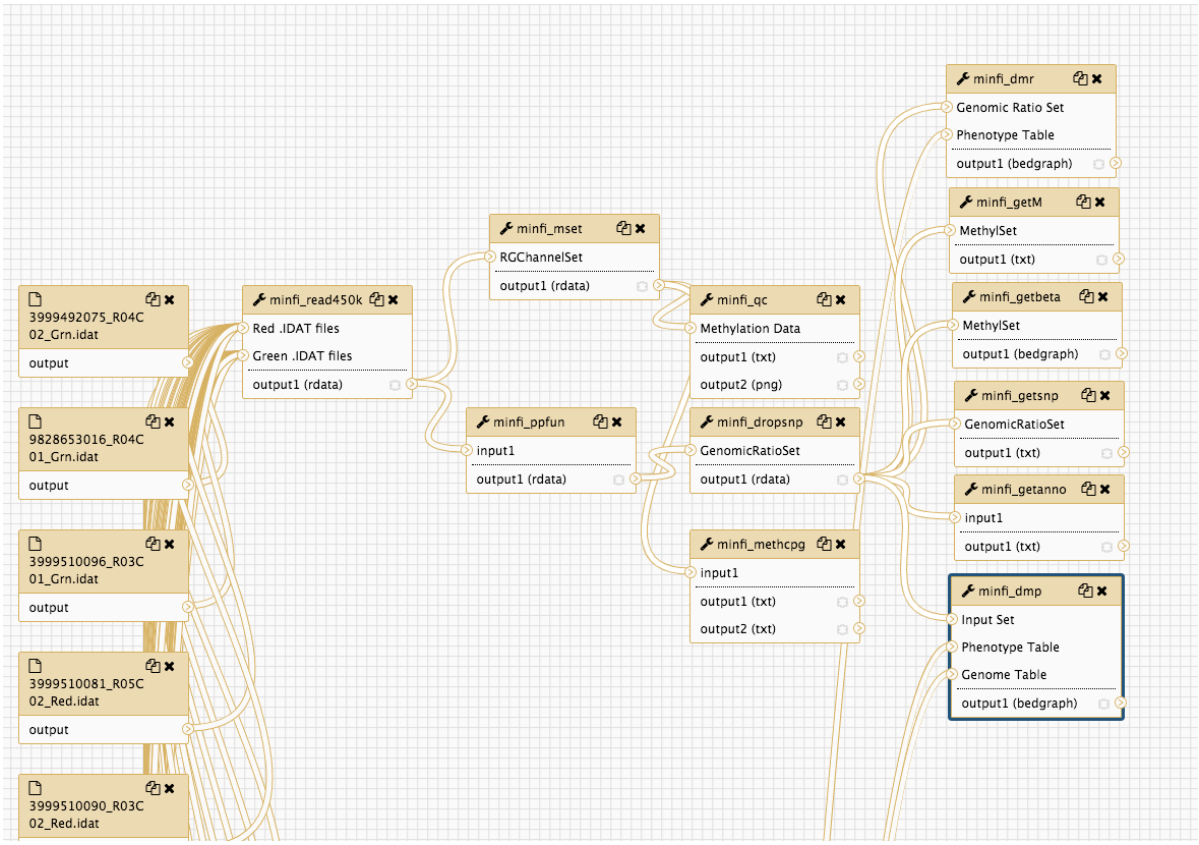


Figura 14. *Workflow* para análise de dados de metilação de câncer de colo de útero, com as etapas desmembradas após instalação das novas ferramentas. A importação dos dados, por serem muitos arquivos importados juntos, é mostrada parcialmente.

5.5 Análises epigenéticas

As 54 amostras da população brasileira selecionadas para esse projeto apresentaram todos os dados disponíveis, exceto dados de RNAseq para somente uma das amostras. Para as outras amostras, foi possível obter todos os dados de exoma, RNAseq e metilação.

No projeto TCGA, além das amostras de câncer de colo de útero, haviam três que eram tecido normal. Essas foram utilizadas como controle nesse estudo (essas amostras não são da população brasileira, mas foram utilizadas por serem os únicos controles normais disponíveis do mesmo tecido). Além disso, vinte pacientes com câncer de endométrio (também do TCGA), adenocarcinoma do tipo endometrióide com estádio II, foram consideradas como grupo externo para controle de qualidade, visando determinar possíveis contaminações e/ou erros na classificação. Essas amostras foram utilizadas somente para esse fim, e a seleção foi

realizada uma vez que alguns casos de colo uterino poderiam ter sido erroneamente diagnosticados como adenocarcinomas de endométrio.

Em relação ao controle de qualidade, todas as amostras passaram nos critérios avaliados. Tal fato é esperado, considerando os critérios restritivos de qualidade do TCGA. O gráfico MDS (do inglês, *multidimensional scalling*) é a principal imagem de visualização de redução de dimensionalidade e similaridade entre as amostras, sendo amplamente utilizado para análise de dados de metilação. De forma geral, as 1000 *probes* com maior variabilidade entre as amostras analisadas são incluídas no gráfico, conforme representado na Figura 15, 16 e 17, evidenciando a separação entre as amostras normais e tumorais, amostras tumorais HPV-negativas e positivas, assim como todas as amostras do presente estudo comparadas com a casuística de endométrio, respectivamente. Uma amostra controle misturou com a dos pacientes e foi retirada das análises posteriores.

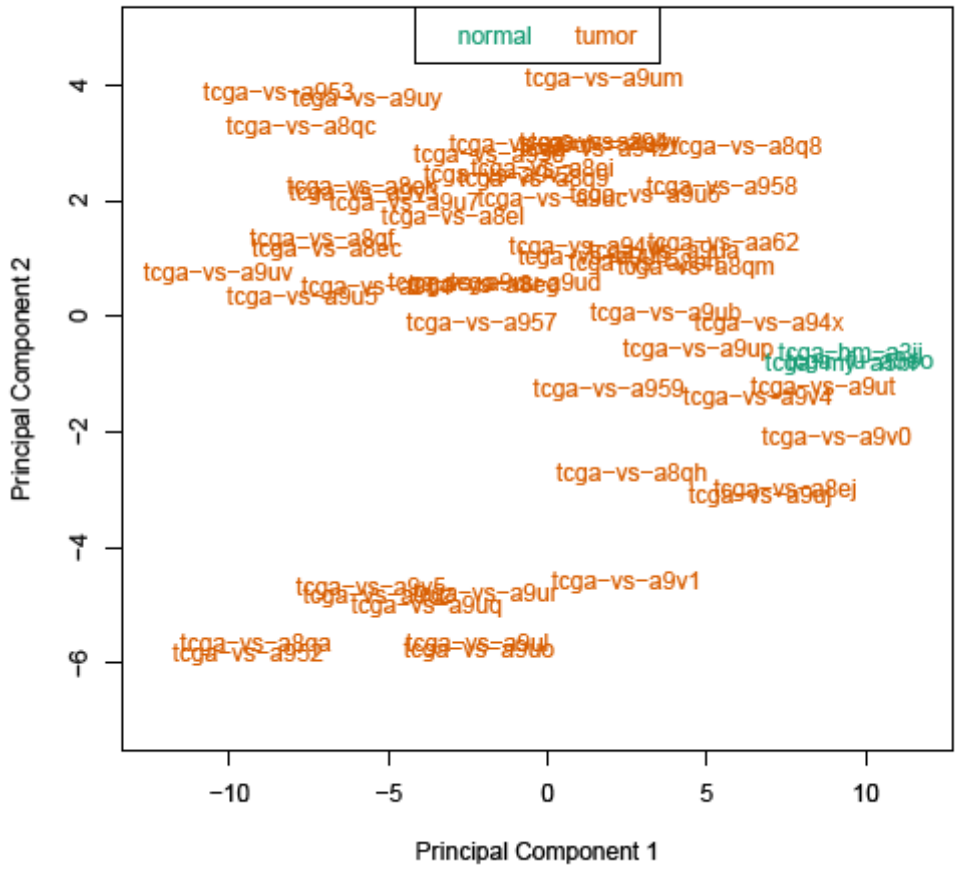


Figura 15 - *MDS-plot* das 54 amostras da população brasileira de câncer de colo de útero no TCGA. Em laranja encontram-se amostras tumorais e em verde amostras normais.

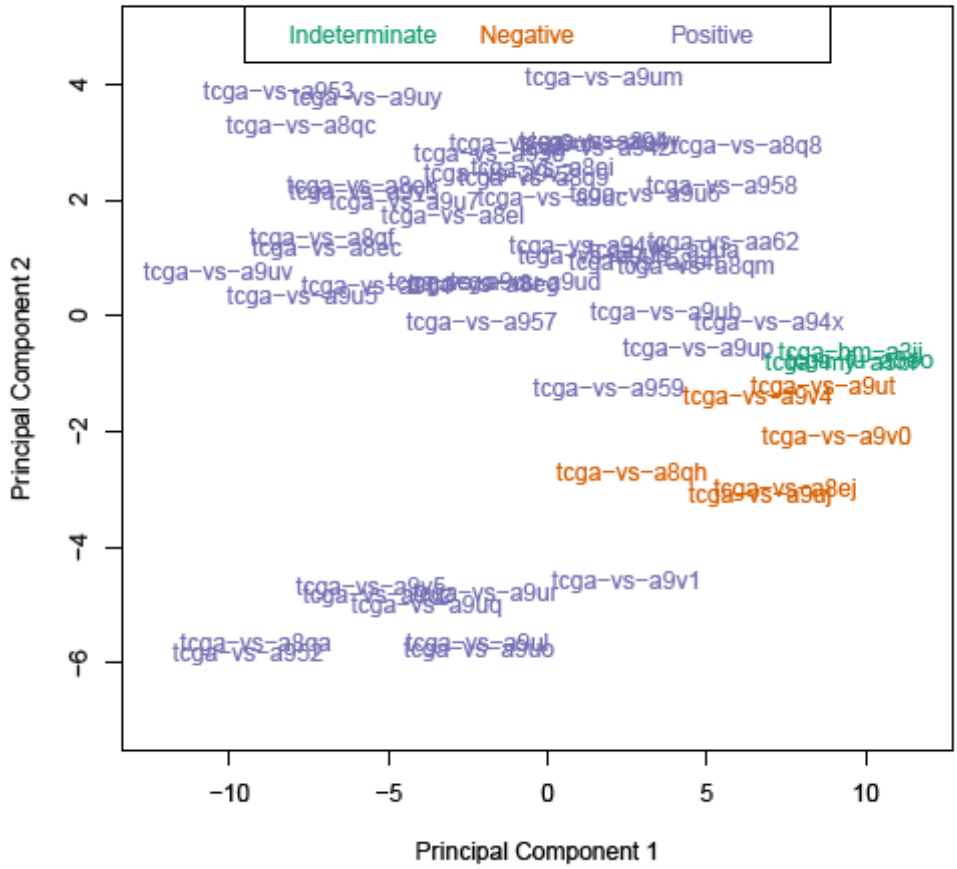


Figura 16 - *MDS-plot* das 54 amostras da população brasileira de câncer de colo de útero no TCGA. Em verde encontram-se amostras HPV-positivas e em roxo amostras HPV-negativas.

A similaridade entre as amostras observadas na Figura 16 mostra um indicativo de separação entre as amostras HPV-positivas e negativas, assim como a formação de dois grandes grupos entre as amostras HPV-positivas.

Considerando os dados obtidos, e como um critério de qualidade adicional, foram incluídas três amostras normais (controles) e vinte amostras de câncer de endométrio, conforme representado na Figura 17.

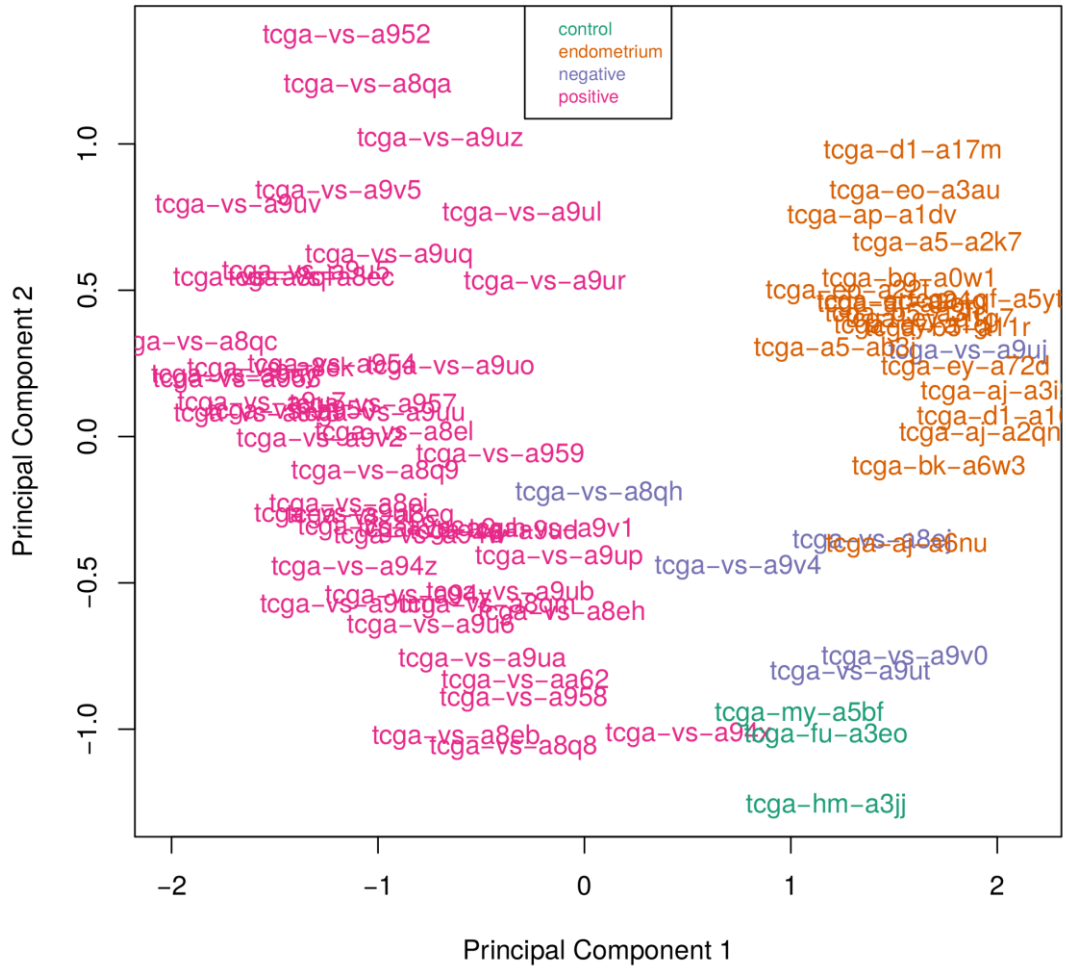


Figura 17 - *MDS-plot* das 54 amostras da população brasileira de câncer de colo de útero no TCGA, 3 controles e 20 amostras de câncer de endométrio. Em rosa encontram-se amostras HPV-positivas, em azul as amostras HPV-negativas, em verde os controles normais e em laranja as amostras com câncer de endométrio.

A figura 18 representa o agrupamento hierárquico das 24.172 *probes* diferencialmente metiladas entre controles e das 54 amostras da população brasileira de câncer de colo de útero no TCGA. É possível observar que apesar da tendência de separação entre pacientes com e sem HPV pelo *MDS-plot,* os mesmos encontram-se misturados no heatmap.

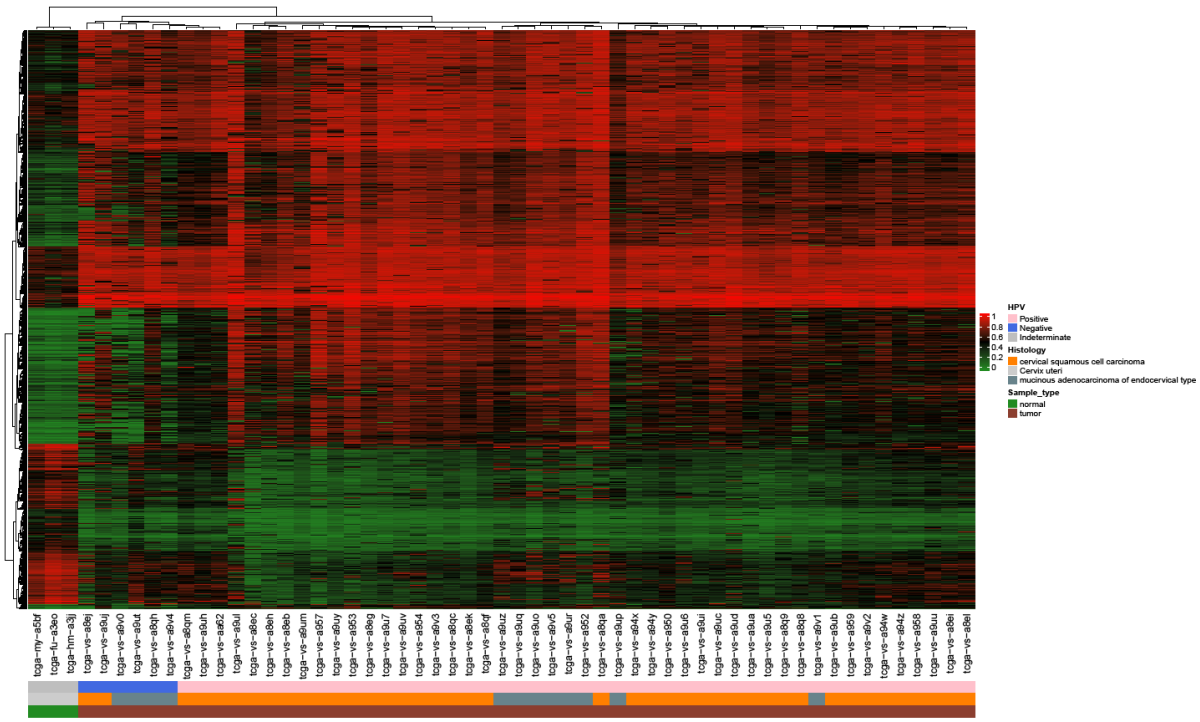


Figura 18 - Heatmap representativo das *probes* diferencialmente metiladas entre controles e pacientes. Na escala 0 (verde) a 1 (vermelho) de *Beta-values*.

Na figura 19, o que podemos observar são regiões gênicas, relação com ilhas CPGs e presença de *enhancer*, separadas nas categorias: todas as sondas, DMPs e DMRs.

Figura 19. *Donut plots* com as regiões gênicas, relação com ilhas CPG e presença de *enhancer*, separadas em categorias: todas as sondas, DMPs e DMRs.

Sabe-se que a análise funcional dos genes metilados é altamente enviesada[22]. Dessa forma, realizamos uma análise funcional baseada em *gene-sets* previamente conhecidos por estarem relacionados com o processo de metilação. A Figura 20 mostra os dez processos mais significativos em escala de -10*log10. O p valor de 0.05 possui um valor de 13 nessa escala, mostrando a alta significância dos resultados encontrados.

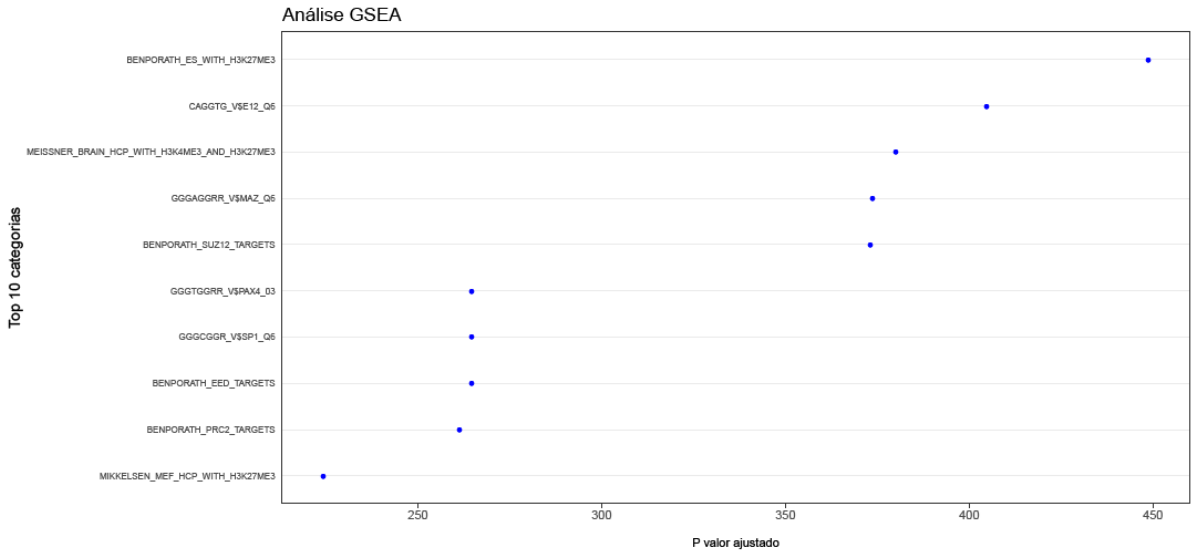


Figura 20 - Análise de enriquecimento funcional utilizando gene-sets. O p valor encontra-se em escala de -10*log10.

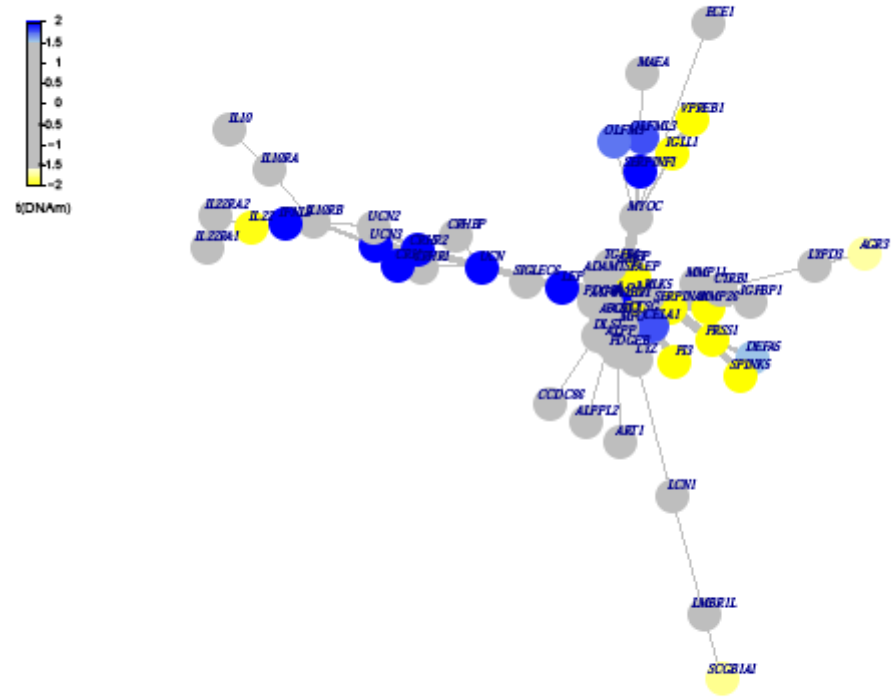
A definição oficial de cada um dos *gene-sets* encontra-se na Tabela 4.

Tabela 4. Descrição das categorias encontradas a partir dos *gene-sets* associados com os processos de metilação. Também se encontra representado o número total de genes do *gene-set* (nList), o número de genes com DMPs (nOVLAP), e os p valores encontrados.

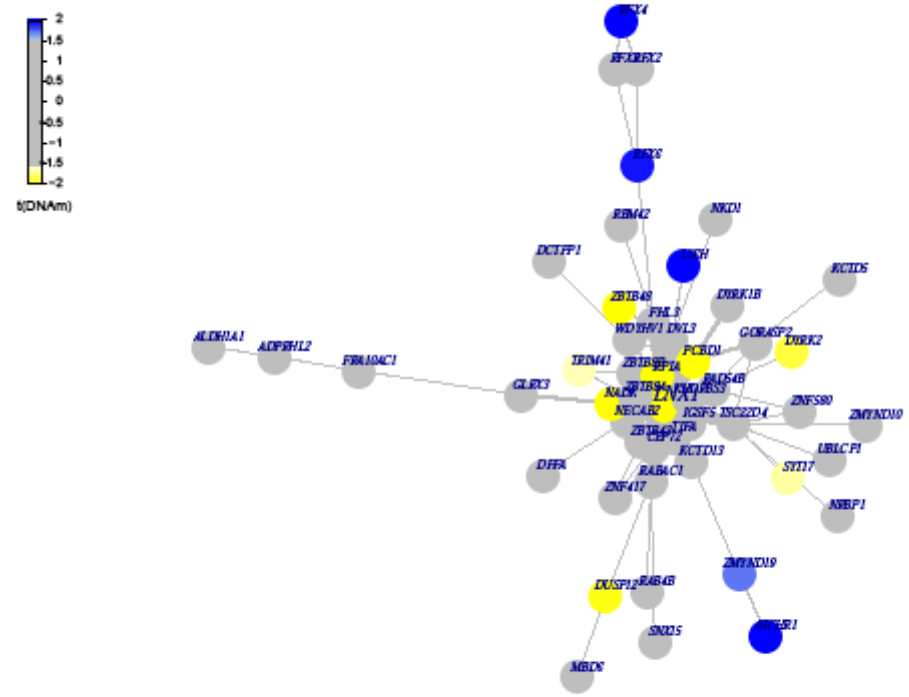
Gene_List	Descrição	nList	nOVLAP	P.value	adjPval
BENPORATH_ES_WITH_H3K27ME3	Set 'H3K27 bound': genes posessing the trimethylated H3K27 (H3K27me3) mark in their promoters in human embryonic stem cells, as identified by ChIP on chip.	1118	781	4,34E-41	3,61E-37
BENPORATH_SUZ12_TARGETS	Set 'Suz12 targets': genes identified by ChIP on chip as targets of the Polycomb protein SUZ12 [GeneID=23512] in human embryonic stem cells.	1038	691	1,60E-35	6,66E-32
CAGGTG_V$E12_Q6	Genes with 3'UTR containing motif CAGGTG which matches annotation for TCF3: Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	2485	1519	2,09E-33	5,80E-29
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3	Genes with high-CpG-density promoters (HCP) bearing histone H3 dimethylation at K4 (H3K4me2) and trimethylation at K27 (H3K27me3) in brain.	1069	742	2,98E-23	6,20E-20
BENPORATH_EED_TARGETS	Set 'Eed targets': genes identified by ChIP on chip as targets of the Polycomb protein EED [GeneID=8726] in human embryonic stem cells.	1062	697	4,71E-22	7,85E-19
GGGAGGRR_V$MAZ_Q6	Genes having at least one occurence of the highly conserved motif M24 GGGAGGRR sites. The motif matches transcription factor binding site V$MAZ_Q6 (v7.4 TRANSFAC).	2274	1352	1,21E-21	1,68E-19
BENPORATH_PRC2_TARGETS	Set 'PRC2 targets': Polycomb Repression Complex 2 (PRC) targets; identified by ChIP on chip on human embryonic stem cells as genes that: posess the trimethylated H3K27 mark in their promoters and are bound by SUZ12 [GeneID=23512] and EED [GeneID=8726] Polycomb proteins.	652	455	5,92E-20	7,05E-18
CTTTGT_V$LEF1_Q2	Genes with 3'UTR containing motif CTTTGT which matches annotation for LEF1: lymphoid enhancer-binding factor 1	1972	1166	7,88E-18	6,57E-14
CAGCTG_V$AP4_Q5	Genes with 3'UTR containing motif CAGCTG which matches annotation for REPIN1: replication initiator 1	1524	926	1,11E-17	1,16E-14
MIKKELSEN_MEF_HCP_WITH_H3K27ME3	Genes with high-CpG-density promoters (HCP) bearing histone H3 trimethylation mark at K27 (H3K27me3) in MEF cells (embryonic fibroblast).	590	431	6,63E-17	6,14E-16

Finalmente foi realizada uma análise de interactoma, na qual é comparada a ocorrência de DMPs e DMRs, priorizando regiões ao redor do sitio de início da transcrição, do inglês *transcription start site (TSS),* conforme mostrado na Figura 21.

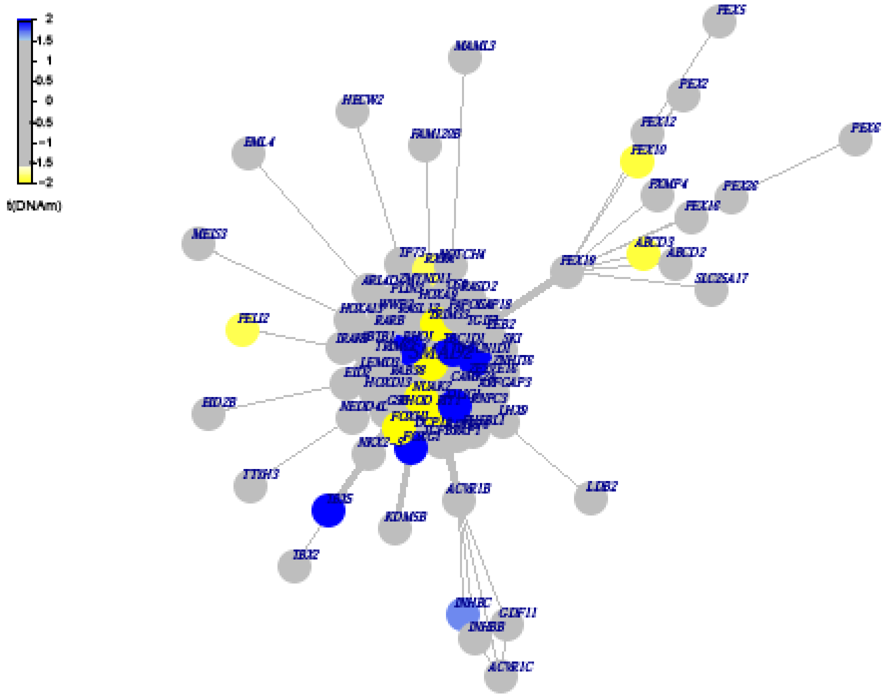
A



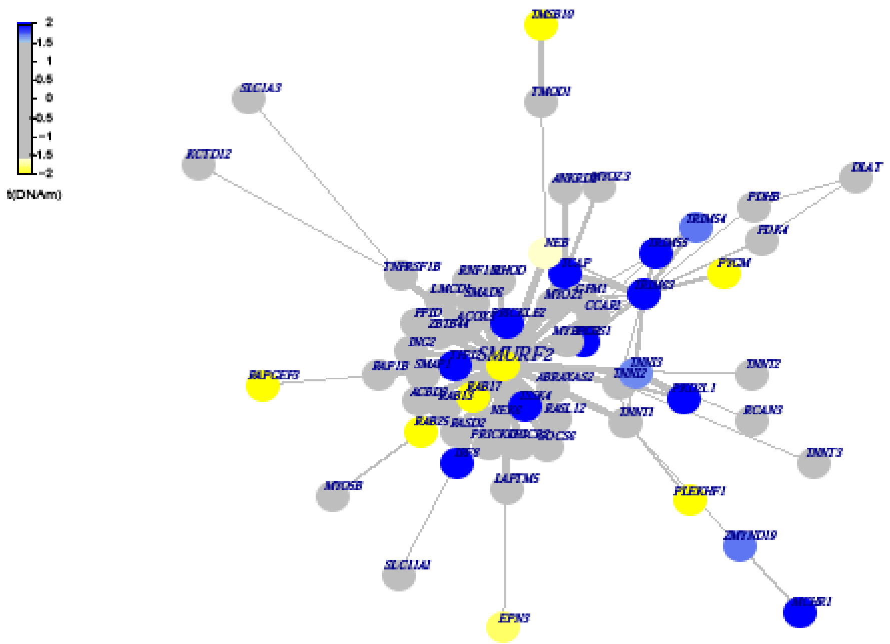
B



C


t(DNAm)
PEX5
MAML3
HECW2
PEX2
PEX12
PEX6
EML4
FAM120B
PEX10
PXMP4
PEX26
PEX16
MEIS3
ABCD3
ABCD2
PEX19
SLC25A17
FELI2
SKI
HID2B
LH39
TTBH3
ACVR1B
LDB2
KDM5B
INHBC
GDF11
INHBB
ACVR1C

D


t(DNAm)
TMSB10
SLC1A3
TMOD1
KCTD12
DLAT
PDHB
PDK4
NEB
SMURF2
TNNT2
RCAN3
TNNT3
TNNT1
PLEKHF1
ZMYND19
MYO5B
LAPTM5
SLC11A1
EFN3

E

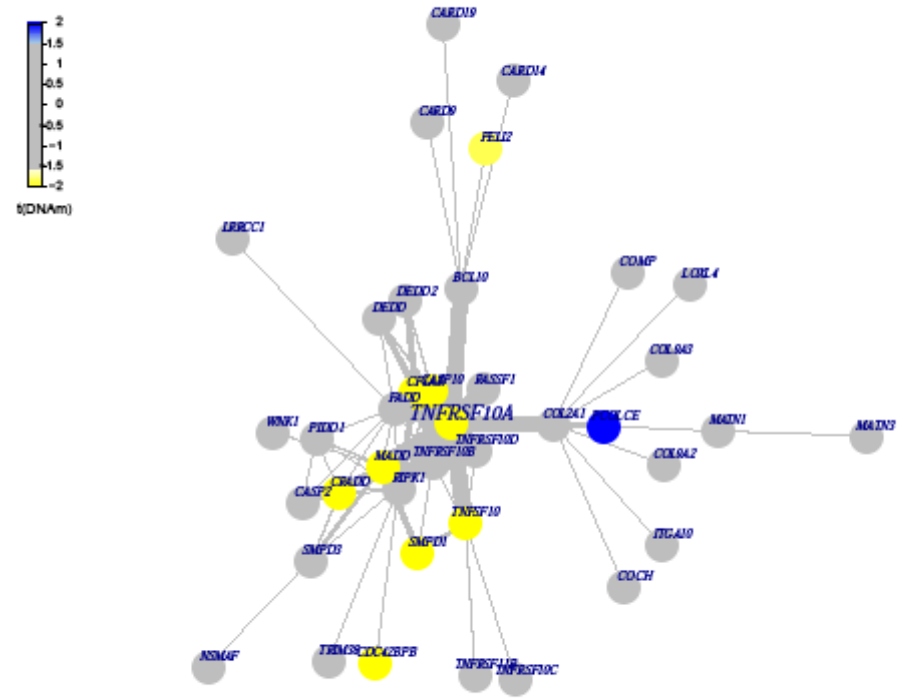


Figura 21 - Interactoma dos principais genes que apresentam DMPs e DMRs próximos ao TSS. Foram selecionados apenas módulos que apresentam genes previamente descritos em câncer de colo de útero. Em A, encontra-se o módulo do gene *A2M.* Em B do *LNX1.* Em C do *SMAD2.* Em D do *SMURF2.* E em F do gene *TNFRSF10A*. Em azul e amarelo encontram-se as escalas de metilação dos mesmos.

6 DISCUSSÃO

O câncer de colo de útero apresenta alta incidência e mortalidade relacionada ao câncer em mulheres no mundo apesar dos avanços dos métodos de prevenção e rastreamento[1]. Trata-se de uma doença multicausal, na qual o vírus do papiloma humano (HPV) encontra-se como principal agente causal[7]. Estudos recentes de caracterização molecular do câncer de colo de útero têm mostrado novas alterações estruturais e aumento da expressão de genes alvo específicos em decorrência da infecção por HPV de alto risco (com foco principalmente em HPV 16, 18 e 52)[11,12]. No entanto, sabe-se que a infecção por HPV não é suficiente para o desenvolvimento e a progressão da doença em questão[1].

Novos achados a partir das análises moleculares oriundas do TCGA levantaram novas perspectivas acerca da discriminação entre subgrupos, podendo ter implicações na prática clínica. Além da diferença entre pacientes HPV positivos e negativos, especialmente no que diz respeito ao perfil mutacional, foram identificados outros subgrupos moleculares. Foram descritas diferenças entre pacientes de acordo com a expressão de queratina (*keratin-high* e *keratin-low*). Explorou-se também diferenças nos padrões de mutagênese da enzima APOBEC nos pacientes com câncer de colo de útero. A implicação desses achados é evidente, especialmente se forem consideradas as novas amplificações dos alvos PD-L1/2 descritos pela primeira vez nesse tipo tumoral, e a modulação do lncRNA *BCAR4* que foi associada com a resposta ao fármaco lapatinib[11].

Dentro desse contexto, o presente estudo iniciou-se pelo interesse em uma reanálise dos dados do TCGA de colo de útero (identificado como TCGA-CESC) com foco nas amostras da população brasileira do Hospital de Câncer de Barretos (atual Hospital de Amor). A reanálise de dados de consórcios internacionais tem sido estimulada no meio científico em decorrência do grande volume de dados produzido em sequenciamentos genômicos e outros métodos de alta performance[14]. Sabe-se que uma análise desse tipo de dado dificilmente conseguirá esgotar as conclusões científicas possíveis. É importante ressaltar que só para o TCGA-CESC encontram-se disponíveis para análise os dados de exoma, transcriptoma e metiloma no total de 8.575 arquivos contendo 10,85 TB [52]discriminados por tipo de dado e em relação à população brasileira nas Tabelas 1 e 2, respectivamente. Além disso, análises

desse tipo permitem o desenvolvimento de toda uma estrutura computacional em uma instituição, uma das principais razões que motivou o presente estudo.

6.1 Solicitação dos dados do TCGA/ Instalação e Adequação das Ferramentas de Análise do Galaxy

A reanálise dessa quantidade de dados é uma tarefa desafiadora por demandar infraestrutura computacional e pessoal especializado. Para tal, a primeira etapa consistiu em implementar uma plataforma de análise que pudesse ser acessível para os alunos com formação na área biológica e sem conhecimento aprofundado em bioinformática. Existem atualmente diversas plataformas conhecidas por realizar a chamada programação visual para análises bioinformáticas[28]. Destas, duas plataformas tinham estrutura suficiente para analisar os dados do TCGA. São elas o *Knime* e o *Galaxy*[29,31]. No entanto, o *Galaxy* foi a plataforma escolhida por diversas razões: por possuir uma ferramenta adicional que permite criar os próprios pacotes - conhecida como *Planemo*, permitir trabalhar com *docker containers* aumentando a reprodutibilidade das análises, ser *open-source,* possuir uma comunidade ativa e encontrar-se disponível em diversos servidores no mundo todo[31-36].

Em relação à questão de acesso aos dados do TCGA, existe uma política altamente restritiva de acesso aos dados brutos, o que inclui o cadastro no eRA Commons para solicitação de aprovação da instituição, que só é contemplado após a comprovação de que a mesma possui uma estrutura de segurança de dados compatível com o exigido. Além disso, cada processo de aprovação tem o prazo de validade por 1 ano, no qual os dados devem ser devidamente apagados, segundo protocolos pré-definidos[53]. É possível realizar a análise à partir dos dados normalizados, como é observado em plataformas de fácil acesso, como por exemplo o cBioPortal[56,57]. No entanto, o objetivo do presente trabalho era trabalhar com *workflows* de análise completos, e para tal, é necessário obter dados brutos. Uma vez implementado um *workflow* de análise completo, se os dados estiverem mais processados, pode-se partir de um ponto de entrada posterior no mesmo *workflow*.

Dessa forma, devido às políticas de acesso a dados brutos do TCGA, não é possível analisá-lo em servidores públicos do *Galaxy*, sendo necessária uma instalação local. Apesar de

existirem mais de 80 servidores públicos do Galaxy em diferentes lugares do mundo, a instalação local foi necessária porque os dados a serem processados seguem normas de segurança específicas. Além disso, a análise desses dados requer grande capacidade de processamento e armazenamento e de ferramentas que não estavam disponíveis em tais servidores, por isso é necessária também a permissão de administrador dentro do servidor, o que só é possível com uma instalação local.

Após a primeira instalação bastante trabalhosa do *Galaxy* ter sido concluída, o Hospital de Amor sofreu um ataque por *hackers* em junho de 2017 e impactou na mesma pois os servidores tiveram que ser formatados. Na ocasião os dados ainda não haviam sido baixados porque a aprovação do TCGA só ocorreu em janeiro de 2018. Como uma solução temporária, foi realizada uma colaboração com um membro mantenedor do *Galaxy*, o Dr. Bjorn Gruening, que disponibilizou o servidor da Universidade de Freiburg, para criação dos primeiros *workflows* deste trabalho. É importante ressaltar que o Galaxy possui a opção de exportação dos *workflows* prontos de um servidor para outro, o que permitiu que os mesmos pudessem ser exportados posteriormente para o nosso servidor após a nova instalação. Dos três tipos de dados (exoma, transcriptoma e metiloma), apenas os dados brutos de metiloma eram menores (Tabela 1 e 2) e não apresentavam as restrições mais rígidas do TCGA. Dessa forma, foi possível avançar com os *workflows* e análises para esse tipo de dado, que por essa razão passou a ser utilizado como foco principal de análise do presente trabalho.

Foram necessárias cinco solicitações ao TCGA no portal do eRA Commons para obtermos a aprovação, cujo tempo total para avaliação foi de aproximadamente 1 ano. Mesmo com suporte da equipe de TI do hospital, foi necessário mais de 1 mês para completar o *download* das 54 pacientes para os dados controlados de exoma e RNAseq, o que novamente justifica o foco de análise nos dados epigenéticos no presente trabalho.

Além da instalação, a adequação das ferramentas também faz parte dos objetivos do trabalho, uma vez que é necessário ter uma visão crítica acerca dos resultados obtidos a partir das ferramentas previamente disponíveis (nós) para a criação dos *workflows*. É possível que existam *bugs* ou mesmo ausência de parâmetros importantes que precisam ser implementados adicionalmente, como toda ferramenta de bioinformática. Utilizando como foco as análises de metiloma, foram observadas diferenças entre as análises com as ferramentas do pacote minfi disponíveis no *Galaxy* com o minfi original do Bioconductor.

Dessa forma, foi realizada uma customização da instalação desse pacote, por meio do *Interactive Environment*[36], um ambiente dentro do *Galaxy* que permite o uso de diversas linguagens de programação, como o R, permitindo o uso de pacotes do Bioconductor e, assim, ajustes nos nós originais usados para o *workflow* do metiloma.

Ainda em relação aos ajustes necessários para as análises de metilação, é amplamente conhecido que as análises funcionais clássicas, que incluem enriquecimento de vias e processos moleculares do *Gene Ontology*, são altamente enviesadas, ou seja, fornecem resultados aparentemente randômicos e de difícil validação[52]. Dessa forma, existe uma alternativa, conhecida como *Gene Set Enrichment Analysis* (GSEA) em que são incluídos *gene sets* específicos de metilação. Além disso, análises de interactoma também têm apresentado resultados promissores[53]. Tais análises encontram-se atualmente implementadas no pacote ChAMP, que pôde ser acoplada ao *workflow* modificado do Galaxy. Atualmente foi estabelecida uma colaboração do Hospital de Amor com a Universidade de Bradford que inicialmente criou esses nós, e futuramente esses ajustes poderão ser disponibilizados para a comunidade do *Galaxy*.

6.2 Visão Geral e Análise Epigenética das Amostras da População Brasileira de Câncer de Colo de Útero do TCGA

Dentre as características gerais da presente amostra (n = 54), é importante ressaltar as principais diferenças em relação a casuística completa de pacientes com câncer de colo de útero do consórcio. A mesma é composta por somente 17 casos de adenocarcinoma mucinoso (6% do total), sendo 12 destes oriundos da casuística brasileira (22% dos brasileiros). Ainda, em relação ao status de HPV, a casuística completa apresenta somente 22 casos negativos (7% do total), com total de 6 (11% dos casos brasileiros) na presente coorte de estudo[54].

Dentre as possíveis explicações para a ocorrência de amostras de colo de útero HPV negativas, encontra-se uma possível confusão com câncer de endométrio. É relatado que uma das características dos pacientes com câncer de endométrio é a idade avançada ao diagnóstico e IMC aumentado. Em revisão dos dados clínicos dessas pacientes foi constatado que essas participantes possuíam idade ao diagnóstico e IMC elevados (média de 64.4 e 28.4,

respectivamente). Em vista dessas evidências, para esclarecer essa questão, foi realizada a revisão dos 6 casos HPV negativos por um segundo patologista experiente. Além disso, também foi feita uma avaliação adicional por meio da comparação das amostras de câncer de colo de útero com 20 amostras de câncer de endométrio de alto grau do TCGA, mostrando que existe uma grande diferença entre os grupos (Figura 17). Além disso, houve um agrupamento das amostras de HPV negativas com as amostras controles (Figura 19), uma evidência de que as mesmas podem realmente não ser infectadas.

Para a avaliação de agrupamentos por MDS, foram também utilizadas amostras controles. A respeito da questão das mesmas, na publicação original do TCGA foram utilizados 120 controles, selecionados randomicamente, para realização das análises[11]. Essas amostras, de forma geral, são oriundas da borda do tecido e encontram-se disponíveis publicamente as informações acerca do fragmento utilizado, e até mesmo a imagem da lâmina utilizada para confirmação das avaliações patológicas[54]. No entanto, após uma avaliação cautelosa desses 120 controles, foi observado que existiam amostras de indivíduos do sexo masculino e de outros tecidos além de colo de útero. É importante ressaltar que inúmeros estudos abordam a importância de critérios de seleção de um grupo controle para confiabilidade dos resultados. Até mesmo um estudo recente do TCGA mostra que os agrupamentos são altamente influenciados pela célula de origem do tecido em tumores sólidos. Dessa forma, no presente estudo foram selecionadas somente as amostras controle de colo de útero normal (n = 3). Apesar do número baixo de amostras controles, considerando a homogeneidade das mesmas (Figura 15 e 16), foi possível obter padrões biologicamente relevantes e com p valores significativos. Além disso, uma das amostras controle corresponde a um paciente HPV positivo que agrupou separadamente com as outras amostras do grupo controle e mais próximo das HPV negativas (Figura 19) confirmando a pureza dos controles.

Nas presentes análises, a partir do grupo controle conforme descrito, o perfil de metilação das pacientes da população brasileira foi caracterizado por meio da análise supervisionada caso vs controle. Essa análise foi realizada considerando as *differentially methylated probes* (DMPs) e as *differentially methylated regions* (DMRs), conforme mostrado nas Tabelas 4 e 5 e na figura 20. Em uma comparação global das DMPs e DMRs em relação à todas as sondas após o filtro de qualidade (390.035 sondas) é possível observar diferenças nos critérios de comparação de região gênica, relação com ilha CpG e presença de *enhancer*

(Figura 19). Em relação a região gênica, a análise de DMPs não apresentou muita diferença em relação a proporção total do número de sondas, apresentando inclusive um número maior de IGRs. Já as DMRs identificadas compreendem um número menor de sondas tanto na IGR quanto no corpo de gene, além de apresentar um enriquecimento nas regiões TSS1500, TSS200 (sítios a até 200 ou 1500 bases do sítio de início de transcrição) e no primeiro exon. Além disso, as análises com foco em DMPs e DMRs apresentam resultados biológicos interessantes, porém parecem não se complementarem. A hipermetilação de promotores geralmente está associada a silenciamento gênico, enquanto a metilação no corpo do gene pode estar associada à maior expressão gênica[58].

Em relação, também não foram observadas diferenças nas proporções entre as DMPs identificadas e o número total de sondas. Já na análise de DMRs, o número de sondas em regiões de ilhas tem um aumento de mais de 30%. No entanto, a análise de DMRs não apresentou diferenças na proporção de *enhancers*, o que é aumentado na análise de DMPs. As ilhas CpGs são elementos responsáveis que suportam a iniciação transcricional e são associados com promotores em diversos genes, assim como *enhancers* ativos[59]. Dentro desse contexto, os dados visando as proporções parecem contraditórios nesse aspecto.

Além disso, os genes identificados nas 10 principais DMRs são como DMPs, mas diversas DMPs relevantes com a doença não são identificadas em DMRs. Um trabalho recente por Li *et al*. aplicou um método de análise não-supervisionada nas amostras de câncer de colo de útero do TCGA, com foco principal em dados de metiloma, encontrando novos subgrupos que se correlacionaram com o *status* histológico. Dos genes encontrados nesse trabalho, muitos deles encontram-se diferencialmente metilados nas nossas análises de DMPs mas não em DMRs.

Desses, o gene *PITX2* encontra-se hipometilado na presente análise de DMPs, com 23 sondas com delta beta menor que 0.2 nas regiões TSS1500, 5'UTR e corpo de gene. Esse gene codifica um membro da família homeobox (*RIEG/PITX*) que atua como fator de transcrição[60]. A hipometilação do mesmo encontra-se associado com outros tipos de canceres, em especial o colorretal[61]. Foi recentemente identificado em uma assinatura de metilação contendo 12 genes como biomarcadores de diagnóstico precoce de câncer de colo de útero[62].

Outro gene encontrado no trabalho de Li *et al*. foi o *IKZF1*, que apresentou 4 sondas hipometiladas na região TSS1500 e corpo do gene. Esse gene codifica um fator de transcrição que pertence à família dos zinc-fingers, e estão associados com o remodelamento da cromatina e considerado um gene com papel no sistema imunológico como um regulador crítico do desenvolvimento de linfócitos[63]. IKZF1 também está associado com a regulação da expressão de *MYC* em leucemia [64] e a regulação de H3K9me3 em câncer de fígado[65].

Em relação a análise de DMRs, das 10 regiões mais significativas, 4 correspondem a regiões intergênicas. Das 6 restantes, todas encontram-se hipometiladas em relação aos controles, com mais de 10 sondas com dela beta maior que 0.2. Dentre essas, foram identificados 4 genes em região de *enhancer*. São eles *CALCA, EDNRB, RAB3C* e *GALR1.* Tais genes possuem poucos trabalhos associados com câncer de colo de útero ou metilação, com resultados contraditórios[66,67,68].

Considerando as análises funcionais resultantes do GSEA, é importante notar que dos 10 principais processos possuem ao menos 400 genes diferencialmente metilados cada, associados a sítios específicos que se localizam próximos do início da transcrição. Dentre as principais categorias identificadas nessa análise, duas correspondem à metilação da histona H3 lisina 27 (H3K27me3), que possui vários estudos confirmando seu papel na carcinogênese[69] e cujo status de metilação é especialmente associado a infecção por HPV em câncer de colo de útero[70]. Já o fator de transcrição SP1 parece estar relacionado a radiosensibilidade[71] e também a processos relacionados indiretamente ao HPV[72]. Os demais processos não possuem muitos relatos na literatura, no entanto, levando-se em consideração o número de genes associados, é possível que existam processos interessantes, apesar da pouca descrição. Mais estudos são necessários para comprovação desses achados.

Em relação à análise do interactoma, a mesma possui um enfoque diferente por buscar por diversos modelos matemáticos possíveis *hotspots* funcionais em um contexto de interação proteína-proteína[53]. Dos 5 principais genes identificados como principais nós são os genes *A2M*, *LNX1*, *SMAD2*, *SMURF2* e *TNFRSF10A*. Destes, os genes *A2M* e *SMURF2* possuem poucas sondas e com resultados contraditórios, ou seja, tanto hipo quanto hipermetilados e com valores mais baixos de delta beta.

Dos 3 principais genes oriundos dessa análise, o gene *LNX1* encontra-se hipometilado, com 10 sondas com delta beta menor que 0.2 em regiões de TSS200, TSS1500 e 5'UTR. Este codifica uma proteína ligada a membrana que está envolvida na transdução de sinal e nas interações proteicas com potencial papel na tumorigênese[73]. Apesar de não existirem trabalhos associados com a metilação desse gene em câncer de colo de útero, sabe-se que a estrutura dos domínios proteicos de LNX1 permite o reconhecimento de vários alvos, conforme identificado em um estudo pela ligação com pelo menos 11 proteínas contendo domínios quinase (DAPK1, PAK1, CDK2, PLK3, AURKB, AURKC, MAPKAPK3, DYRK3, EPHA8, EPHB3 e TRIB1)[74].

O gene *SMAD2* encontra-se hipermetilado em relação aos controles, com 3 sondas com delta beta maior que 0.2, todas em regiões de TSS1500. Esse gene codifica proteínas que atuam como transdutores de sinal e moduladores transcricionais em múltiplas vias de sinalização. Regulam processos tais como proliferação celular, apoptose e diferenciação[75]. A expressão desse gene tem sido associada com a pior sobrevida em câncer de colo de útero[76]. Alterações no perfil de metilação desse gene tem sido associada à migração e invasão em câncer de mama[77] e processos gerais de diferenciação pluripotente[78]. Não existem publicações atuais associando a hipermetilação desse gene com câncer de colo de útero.

Outro gene hipermetilado em câncer de colo de útero em relação aos controles normais, identificado pela análise de interactoma foi o *TNFRSF10A*. Tal gene codifica uma proteína membro da superfamília dos *TNF-receptor,* que é ativada por um ligante associado a apoptose[79]. Esse gene foi identificado como um marcador de metilação de alta frequência em câncer de colo de útero, com potencial de detecção precoce do mesmo[80].

Finalmente, na presente análise, não houve nenhuma associação estatisticamente significante entre os níveis de metilação e o *status* de sobrevida global desta série de pacientes. Porém, esta análise quanto ao desfecho dos pacientes apresenta um pequeno número de pacientes, o que pode afetar os resultados. No entanto, os achados contribuem para o melhor conhecimento do papel da metilação de DNA em câncer de colo de útero na população brasileira.

7 CONCLUSÃO

Diante dos resultados do nosso trabalho concluímos que a implementação de uma plataforma para análise de dados por programação visual em um centro de pesquisa é uma tarefa complexa, que exige estrutura computacional e pessoal especializado, fato esse que motivou outro grande projeto na aquisição de novos fundos para nova estrutura. Apesar de complexa, é também necessária haja vista que o *Galaxy* se provou uma plataforma muito útil nas análises desse trabalho e também na disponibilização de ferramentas para análise genômica no diagnóstico, proporcionando uma forma facilitada, confiável e reprodutível de análise para pessoas com formação biológica.

Em relação à aquisição e análise de dados do TCGA, o processo pode ser bastante moroso, mas uma vez realizado, traz muitas possibilidades de análise com dados de alta qualidade, com análises que trouxeram importantes resultados.

REFERÊNCIAS

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov;68(6):394–424.

2. CÂNCER - Tipo - Colo do Útero Instituto Nacional de Câncer. [Internet]. Rio de Janeiro: INCA; 2018 [cited 2018 Aug 24]; Available from: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/ colo_utero/definicao.

3. GLOBOCAN Cancer Fact Sheets: Cervical cancer [Internet] Lyon: IARC; 2017 [cited 2019 Jan 21]; Available from: http://gco.iarc.fr/today/data/factsheets/cancers/23-Cervix-uteri-fact-sheet.pdf

4. World Health Organization [Internet] Geneva: WHO; 2019 [cited 2019 Jan 21]. Available from: https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/

5. Can Cervical Cancer Be Prevented? American cancer society. [Internet]. [cited 2017 Dec 20]. Available from: https://www.cancer.org/cancer/cervical-cancer/prevention-and-early-detection/can-cervical-cancer-be-prevented.html

6. Ações e Programas no Brasil - Controle do Câncer do Colo do Útero Instituto Nacional de Câncer. [Internet]. Rio de Janeiro: INCA, 2017 [cited 2017 Dec 20]; Available from: http://www2.inca.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_nacional_controle_cancer_colo_utero/prevencao

7. Kim HS, Kim TJ, Lee IH, Hong SR. Associations between sexually transmitted infections, high-risk human papillomavirus infection, and abnormal cervical Pap smear results in OB/GYN outpatients. J Gynecol Oncol. 2016 Sep;27(5):e49.

8. Lees BF, Erickson BK, Huh WK. Cervical cancer screening: evidence behind the guidelines. Am J Obstet Gynecol. 2016 Apr;214(4):438–43.

9. Diretrizes para o rastreamento_cancer_colo_utero.pdf Instituto Nacional de Câncer. [Internet]. Rio de Janeiro: INCA; 2018. [cited 2017 Dec 20]. Available from: http://bvsms.saude.gov.br/bvs/publicacoes/inca/rastreamento_cancer_colo_utero.pdf

10. Davey DD. Cervical cytology classification and the Bethesda System. Cancer J. 2003 Oct;9(5):327–34.

11. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. Nature. 2017 16;543(7645):378–84.

12. Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, et al. Landscape of genomic alterations in cervical carcinomas. Nature. 2014 Feb 20;506(7488):371–5.

13. Chung TKH, Van Hummelen P, Chan PKS, Cheung TH, Yim SF, Yu MY, et al. Genomic aberrations in cervical adenocarcinomas in Hong Kong Chinese women. Int J Cancer. 2015 Aug 15;137(4):776–83.

14. Blogger PG. Cancer Genomics: Data, Data and more Data [Internet]. Speaking of Medicine. 2015 [cited 2017 Dec 20]. Available from: http://blogs.plos.org/speakingofmedicine/2015/04/17/interview-francis-ouellette/

15. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia. 2007 Feb;9(2):166–80.

16. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009 Jan;37(Database issue):D885-890.

17. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb 18;463(7283):899–905.

18. The Cancer Genome Atlas Home Page [Internet]. The Cancer Genome Atlas - National Cancer Institute. [cited 2016 Sep 19]. Available from: http://cancergenome.nih.gov/.

19. International Cancer Genome Consortium [Internet]. [cited 2016 Sep 19]. Available from: https://icgc.org/.

20. The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764.

21. de la Garza L, Veit J, Szolek A, Röttig M, Aiche S, Gesing S, et al. From the desktop to the grid: scalable bioinformatics via workflow conversion. BMC Bioinformatics. 2016;17:127.

22. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646–74

23. Alberts, B. et al Câncer. In: Alberts, B. et al, editors. Biologia Molecular da Célula. Porto Alegre: Artmed, 2017.

24. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011 Mar;17(3):297–303.

25. Quick Facts [Internet]. The Cancer Genome Atlas - National Cancer Institute. [Internet] [cited 2018 Feb 20];Available from: https://cancergenome.nih.gov/newsevents/forthemedia/quickfacts

26. Data Size Matters [Infographic] - Blog [Internet]. 2013 [cited 2017 Nov 13]. Available from: https://datascience.berkeley.edu/big-data-infographic/

27. Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, et al. An APOBEC Cytidine Deaminase Mutagenesis Pattern is Widespread in Human Cancers. Nature genetics [Internet]. 2013 Sep [cited 2018 Mar 23];45(9):970. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3789062/

28. Aiche S, Sachsenberg T, Kenar E, Walzer M, Wiswedel B, Kristl T, et al. Workflows for automated downstream data analysis and visualization in large-scale computational mass spectrometry. Proteomics. 2015 Apr;15(8):1443–7.

29. KNIME | Open for Innovation [Internet]. [cited 2016 Sep 20]. Available from: https://www.knime.org/

30. Tool Shed [Internet]. [cited 2017 Nov 13]. Available from: https://toolshed.genouest.org/

31. Blankenberg D, Hillman-Jackson J. Analysis of next-generation sequencing data using Galaxy. Methods Mol Biol. 2014;1150:21–43.

32. Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.

33. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: A platform for interactive large-scale genome analysis. Genome Res [Internet]. 2005 Oct [cited 2017 Dec12];15(10):1451–5.Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240089/

34. The Galaxy Project: Online bioinformatics analysis for everyone [Internet]. [cited 2016 Sep 20]. Available from: https://galaxyproject.org/

35. Tool Shed [Internet]. [cited 2016 Sep 20]. Available from: https://toolshed.genouest.org/

36. Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers.http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005425(accessed 19 Feb2018).

37. Commons Login [Internet]. [cited 2018 Jul 31]. Available from: https://public.era.nih.gov/commons/public/login.

38. dbGaP: Authorized Access: dbGaP Authorized Access [Internet]. [cited 2018 Mar 1]. Available from: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login

39. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007 Oct;39(10):1181–6.

40. GaP FAQ Archive. National Center for Biotechnology Information (US); 2009.

41. GDC Data Model Components | NCI Genomic Data Commons [Internet]. [cited 2018 Jul 31]. Available from: https://gdc.cancer.gov/developers/gdc-data-model/gdc-data-model-components

42. Bizagi [Internet]. Microsoft Store. [cited 2018 Mar 1]. Available from: https://www.microsoft.com/en-us/store/p/bizagi.

43. R: The R Project for Statistical Computing [Internet]. [cited 2018 Mar 1]. Available from: https://www.r-project.org/

44. Quick Start Planemo 0.49.0.dev0 documentation [Internet]. [cited 2018 Mar 1]. Available from: http://planemo.readthedocs.io/en/latest/readme.html

45. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct;98(4):288–95.

46. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014 Oct 23;159(3):676–90.

47. kpbioteam. Devel repository for minfi [Internet]. 2018 [cited 2018 Aug 13]. Available from: https://github.com/kpbioteam/minfi

48. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform [Internet]. 2009 Apr [cited 2018 Mar 1];42(2):377–81. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2700030/

49. Repository [Internet]. [cited 2018 Jan 10]. Available from: https://portal.gdc.cancer.gov/repository

50. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol [Internet]. 2014 [cited 2018 Aug 13];15(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4283580/

51. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol. 2012 Feb;41(1):200–9.

52. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics [Internet]. 2013 Aug 1 [cited 2019 Jan 28];29(15):1851–7. Available from: https://academic.oup.com/bioinformatics/article/29/15/1851/265573 .

53. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. 2014; 30: 428–430.

54. Projects [Internet].[cited 2019 Jan 26]. Available from: https://portal.gdc.cancer.gov/repository.

55. Documentation [internet] | NCI Genomic data commons. https://gdc.cancer.gov/documentation[https://docs.gdc.cancer.gov/Data/PDF/Data_UG.pdf] (acessed 26 Jan2019).

56. Cerami et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. May 2012 2; 401.

57. Gao et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 6, pl1 (2013).

58. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012; 13: 484–492.

59. Bell JSK, Vertino PM. Orphan CpG islands define a novel class of highly active enhancers. Epigenetics. 2017; 12: 449–464.

60. Gene [Internet]. [cited 2019 Jan 26]. Available from: https://www.ncbi.nlm.nih.gov/gene/5308.

61. Semaan A, Uhl B, Branchi V, Lingohr P, Bootz F, Kristiansen G et al. Significance of PITX2 Promoter Methylation in Colorectal Carcinoma Prognosis. Clin Colorectal Cancer. 2018; 17: e385–e393.

62. Bhat S, Kabekkodu SP, Varghese VK, Chakrabarty S, Mallya SP, Rotti H et al. Aberrant gene-specific DNA methylation signature analysis in cervical cancer. Tumour Biol. 2017; 39: 1010428317694573.

63. Gene [Internet]. [cited 2019 Jan 26]. Available from: https://www.ncbi.nlm.nih.gov/gene/10320.

64. Ge Z, Guo X, Li J, Hartman M, Kawasawa YI, Dovat S et al. Clinical significance of high c-MYC and low MYCBP2 expression and their association with Ikaros dysfunction in adult acute lymphoblastic leukemia.Oncotarget. 2015; 6: 42300–42311.

65. Huo Q, Ge C, Tian H, Sun J, Cui M, Li H et al. Dysfunction of IKZF1/MYC/MDIG axis contributes to liver cancer progression through regulating H3K9me3/p21 activity. Cell Death Dis. 2017; 8: e2766.

66. Lin H, Ma Y, Wei Y, Shang H. Genome-wide analysis of aberrant gene expression and methylation profiles reveals susceptibility genes and underlying mechanism of cervical cancer.Eur J Obstet Gynecol Reprod Biol. 2016; 207: 147–152.

67. Wisman GBA, Nijhuis ER, Hoque MO, Reesink-Peters N, Koning AJ, Volders HH et al. Assessment of gene promoter hypermethylation for detection of cervical neoplasia. Int J Cancer. 2006; 119: 1908–1914.

68. Kori M, Yalcin Arga K. Potential biomarkers and therapeutic targets in cervical cancer: Insights from the meta-analysis of transcriptomics data within network biomedicine perspective.PLoS ONE. 2018; 13: e0200717.

69. Takeshima H, Wakabayashi M, Hattori N, Yamashita S, Ushijima T. Identification of coexistence of DNA methylation and H3K27me3 specifically in cancer cells as a promising target for epigenetic therapy.Carcinogenesis. 2015; 36: 192–201.

70. Gameiro SF, Kolendowski B, Zhang A, Barrett JW, Nichols AC, Torchia J et al. Human papillomavirus dysregulates the cellular apparatus controlling the methylation status of H3K27 in different human cancers to consistently alter gene expression regardless of tissue of origin. Oncotarget. 2017; 8: 72564–72576.

71. Deng Y-R, Jiang H-P, Wu L-F, Chen W, Lin D, Guo S-Q. [Role of specificity protein 1 in modulating radiosensitivity of cervical cancer cell lines]. Nan Fang Yi Ke Da Xue Xue Bao. 2016; 36: 1226–1230.

72. Zhang J, Li S, Yan Q, Chen X, Yang Y, Liu X et al. Interferon-β induced microRNA-129-5p down-regulates HPV-18 E6 and E7 viral gene expression by targeting SP1 in cervical cancer cells. PLoS ONE. 2013; 8: e81366.

73. Gene [Internet]. [cited 2019 Jan 26]. Available from: https://www.ncbi.nlm.nih.gov/gene/84708.

74. Wolting CD, Griffiths EK, Sarao R, Prevost BC, Wybenga-Groot LE, McGlade CJ. Biochemical and Computational Analysis Of LNX1 Interacting Proteins. PLOS ONE. 2011; 6: e26248.

75. Gene [Internet]. [cited 2019 Jan 26]. Available from: https://www.ncbi.nlm.nih.gov/gene/4087.

76. Kloth JN, Kenter GG, Spijker HS, Uljee S, Corver WE, Jordanova ES et al. Expression of Smad2 and Smad4 in cervical cancer: absent nuclear Smad4 expression correlates with poor survival. Mod Pathol. 2008; 21: 866–875.

77. Lu Y, Wang L, Li H, Li Y, Ruan Y, Lin D et al. SMAD2 Inactivation Inhibits CLDN6 Methylation to Suppress Migration and Invasion of Breast Cancer Cells. Int J Mol Sci. 2017; 18. doi:10.3390/ijms18091863.

78. Wang L, Xu X, Cao Y, Li Z, Cheng H, Zhu G et al. Activin/Smad2-induced Histone H3 Lys-27 Trimethylation (H3K27me3) Reduction Is Crucial to Initiate Mesendoderm Differentiation of Human Embryonic Stem Cells. J Biol Chem. 2017; 292: 1339–1350.

79. Gene [Internet]. [cited 2019 Jan 26]. Available from: https://www.ncbi.nlm.nih.gov/gene/8797.

80. Wentzensen N, Sherman ME, Schiffman M, Wang SS. Utility of Methylation Markers in Cervical Cancer Early Detection: Appraisal of the State-of-the-Science. Gynecol Oncol. 2009; 112: 293–299

ANEXOS

Anexo A – Carta de aprovação do CEP

HOSPITAL DO CÂNCER DE BARRETOS / FUNDAÇÃO PIO XII

PARECER CONSUBSTANCIADO DO CEP

DADOS DA EMENDA

Título da Pesquisa: Aplicação de métodos de análise por programação visual em dados de câncer de colo de útero do The Cancer Genome Atlas

Pesquisador: Adriane Feijó Evangelista

Área Temática:

Versão: 2

CAAE: 60545416.7.0000.5437

Instituição Proponente: Fundação Pio XII

Patrocinador Principal: Fundação Pio XII
Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 1.894.025

Apresentação do Projeto:

O pesquisador contextualiza que a alta prevalência de infecção por papilomavírus humano (HPV), o principal agente etiológico do câncer cervical, especialmente em países em desenvolvimento, têm incentivado a implementação de programas de prevenção, monitoramento e conscientização sobre a doença. Em especial nos programas de rastreamento, a aplicação de vacinas profiláticas e a introdução de testes de DNA de HPV têm representado avanços no diagnóstico, tratamento e prevenção.

Hipótese:

Dados em larga escala de carcinoma cervical do TCGA podem ser analisados por meio de workflow userfriendly e reprodutíveis.

Metodologia Proposta:

4.3 Análise molecular dos dados

4.3.1 Controle de qualidade de Dados de NGS

4.3.2 Alinhamento com genoma de referencia

4.3.3 Busca de SNPs e Indels em dados de Sequenciamento

Endereço: Rua Antenor Duarte Vilela, 1331
Bairro: Dr. Paulo Prata **CEP:** 14.784-400
UF: SP **Município:** BARRETOS
Telefone: (17)3321-0347 **Fax:** (17)3321-6600 **E-mail:** cep@hcancerbarretos.com.br

4.3.4. Análise de expressão diferencial de dados de RNA-seq
4.3.5 Análise de dados de metilação
4.3.6 Análise geral de dados
4.3.7 Figuras adicionais
(detalhes sobre a metodologia de análise de dados estão abaixo na metodologia de análise de dados)

Objetivo da Pesquisa:
Objetivo Primário:
Uso de métodos da programação visual para análise dos dados do TCGA, com foco nas amostras da população brasileira disponizada pelo Hospital de Cancer de Barretos.
Objetivo Secundário:
- Geração de workflows por programação visual para análise de dados em larga escala oriundos do projeto TCGA;
- Análise dos dados de sequenciamento de DNA, metilação e expressão genica (RNA-seq);
- Identificação de biomarcadores que permitam a caracterização de amostras de
alto risco para câncer cervical.
- Comparação dos dados moleculares com características clínico patológicas (metástase linfonodal, estadio, radioterapia e prognóstico).

Avaliação dos Riscos e Benefícios:
Riscos ao participante: Apesar de trabalhar com dados de sequenciamento genético, o objetivo é analisar dados de mutações somáticas e não há risco de encontrar, acidentalmente, qualquer alteração genética que denote necessidade de aconselhamento genético. Desse modo, o risco é minimo e representado apenas pela quebra de sigilo dos dados do participante, dados estes que serão protegidos pelo pesquisador. A única atualização de dados será feita com dados de follow-up (status da paciente) utilizando dados do Registro Hospitalar de Câncer do Hospital de Câncer de Barretos (sabemos quais casos foram enviados por nossa instituição para o consórcio TCGA, perfazendo um total de
54 casos). Entretanto, os pesquisadores garantem o sigilo das participantes de pesquisa.
Benefícios ao participante: A pesquisa não trará nenhum benefício imediato ao participante de pesquisa. Os possíveis benefícios estão relacionados as informações que essa pesquisa trará a outros pacientes.

Comentários e Considerações sobre a Pesquisa:

Emenda referente a alteração em um termo do título.

No título do projeto, houve alteração de "cervical" para "colo de útero". Esse termo foi para o meio da frase, ficando assim: "Aplicação de métodos de análise por programação visual em dados de câncer de colo de útero do The Cancer Genome Atlas"

Considerações sobre os Termos de apresentação obrigatória:

Justificativa para dispensa de TCLE coerente:

"As participantes da pesquisa foram incluídas no mundo todo, sendo 54 amostras brasileiras. Tais participantes já concordaram que tivessem seus dados e materiais coletados e que esses dados constassem em bancos de dados públicos. Tais informações são disponibilizadas dessa forma para permitir que outras pessoas explorem esses dados, dessa maneira, peço dispensa porque estarei utilizando apenas dados que já obtiveram consentimento para serem disponibilizados".

Recomendações:

Sem recomendações

Conclusões ou Pendências e Lista de Inadequações:

Emenda adequada as legislações vigentes em ética em pesquisas envolvendo seres humanos

Considerações Finais a critério do CEP:

O Comitê de Ética em Pesquisa da Fundação Pio XII - Hospital de Câncer de Barretos analisou o(s) seguinte(s) documento(s) do projeto 1261/2016, e:

- Aprovou a emenda ao estudo, submetido em 05/01/2017;

Após análise do(s) documento(s) supracitado(s), o Comitê faz a seguinte recomendação:

(x) O Estudo deve Continuar;

( ) O Estudo dever ser Interrompido;

( ) O Estudo está Finalizado;

( ) Solicita-se Esclarecimento

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_850854_E1.pdf	05/01/2017 14:13:28		Aceito

Outros	Formulario_para_Emenda.docx	05/01/2017 14:07:32	Thais Talarico Hosokawa	Aceito
Outros	carta_emenda.pdf	05/01/2017 08:44:50	Thais Talarico Hosokawa	Aceito
Projeto Detalhado / Brochura Investigador	Projeto_Thais_V1.pdf	05/01/2017 08:44:00	Thais Talarico Hosokawa	Aceito
Folha de Rosto	folha_rosto_v2.pdf	05/01/2017 08:41:45	Thais Talarico Hosokawa	Aceito
Declaração de Pesquisadores	fonte_de_financiamento.pdf	30/09/2016 15:32:48	Thais Talarico Hosokawa	Aceito
Outros	cadastro_de_projeto.pdf	30/09/2016 15:32:31	Thais Talarico Hosokawa	Aceito
Declaração de Pesquisadores	declaracao_responsabilidade.pdf	28/09/2016 08:36:03	Thais Talarico Hosokawa	Aceito
Outros	mabin.pdf	28/09/2016 08:33:32	Thais Talarico Hosokawa	Aceito
Declaração de Instituição e Infraestrutura	declaracao_ciencia_estudo.pdf	28/09/2016 08:32:27	Thais Talarico Hosokawa	Aceito

Situação do Parecer:
Aprovado

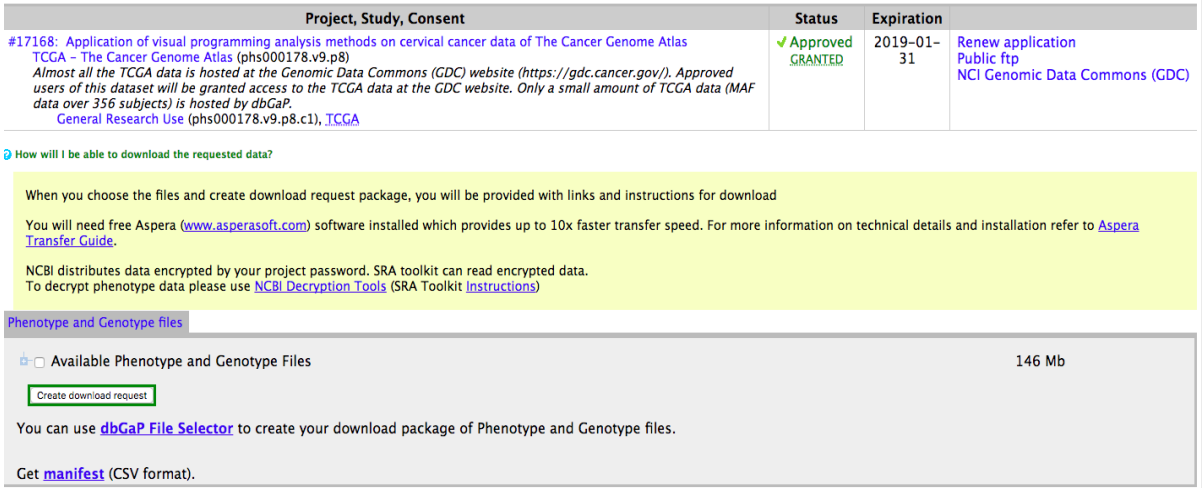
Necessita Apreciação da CONEP:
Não

BARRETOS, 19 de Janeiro de 2017

Assinado por:
Cleyton Zanardo de Oliveira
(Coordenador)

Anexo B – Processo de obtenção dos dados

Em 31 de janeiro de 2018 obtivemos aprovação para download dos dados controlados do TCGA, conforme figura abaixo. Foram necessárias cinco solicitações antes da aprovação.



Project, Study, Consent	Status	Expiration	
#17168: Application of visual programming analysis methods on cervical cancer data of The Cancer Genome Atlas TCGA - The Cancer Genome Atlas (phs000178.v9.p8) *Almost all the TCGA data is hosted at the Genomic Data Commons (GDC) website (https://gdc.cancer.gov/). Approved users of this dataset will be granted access to the TCGA data at the GDC website. Only a small amount of TCGA data (MAF data over 356 subjects) is hosted by dbGaP.* General Research Use (phs000178.v9.p8.c1), TCGA	✓ Approved GRANTED	2019-01-31	Renew application Public ftp NCI Genomic Data Commons (GDC)

How will I be able to download the requested data?

When you choose the files and create download request package, you will be provided with links and instructions for download

You will need free Aspera (www.asperasoft.com) software installed which provides up to 10x faster transfer speed. For more information on technical details and installation refer to Aspera Transfer Guide.

NCBI distributes data encrypted by your project password. SRA toolkit can read encrypted data.
To decrypt phenotype data please use NCBI Decryption Tools (SRA Toolkit Instructions)

Phenotype and Genotype files

Available Phenotype and Genotype Files 146 Mb

Create download request

You can use dbGaP File Selector to create your download package of Phenotype and Genotype files.

Get manifest (CSV format).

Captura de tela do site do dbGAP, após entrada na área de acesso autorizado(40).

Anexo C - Funções utilizadas no *workflow* elaborado para análise de metilação.

Ferramenta	Função
minfi_read450k	Importa os arquivos brutos de metilação, em formato. idat, e produz um objeto RGChannel (da classe RGSet). É um objeto inicial de análise do minfi que contém intensidades brutas dos canais verde e vermelho da leitura. Esse objeto contém também as intensidades de probes controle.
minfi_mset	Gera objetos em formato MethylSet, que contém sinais de metilado e não-metilado.
minfi_qc	Gera um gráfico simples que usa log da intensidade da mediana nos canais metilado (M) e não-metilado (U). Quando essas duas medianas são plotadas uma contra a outra, foi observado que boas amostras se agrupam enquanto que amostras que falharam tendem a se separar e ter intensidade mediana mais baixa. Para obter os sinais metilado e não metilado, precisamos converter o RGChannelSet a um objeto que contenha os sinais metilado e não-metilado. Essa ferramenta então recebe um objeto RGChannelSet e converte de intensidades verde e vermelha a metilado e não-metilado, de acordo com o desenho especial da probe 450k e retorna os sinais convertidos em um novo objeto da classe MethylSet. Não faz nenhuma normalização.
minfi_pppfun	Implementa o algoritmo de normalização funcional desenvolvido por Fortin e colaboradores[49]. Esse algoritmo usa as probes de controle interno presentes no *array* para inferir variação técnica entre *arrays.* É particularmente útil em estudos comparando condições com conhecidas diferenças em larga escala, como estudos câncer/normal ou entre tecidos. Tem sido mostrado que para esses tipos de estudo, normalização functional tem uma melhor performance que outras abordagens existentes. Por padrão, a função aplica a função de pré-processamento Noob, que é o primeiro passo para remoção de background e usa os dois primeiros components principais das robes de controle para inferior a variação não desejada.
minfi_dropsnp	Permite remover as sondas de SNPs em ilhas CpG. Esse é um passo recomendado de análise.
minfi_dmr	Usa uma função do pacote minfi chamada bumphunter, uma versão do algoritmo bump hunting[52] adaptado para o array 450k. Ao invés de procurar associação entre localizações genômicas específicas de um fenótipo de interesse, a função procura regiões genômicas que são diferencialmente metiladas entre duas condições.
minfi_getm	Retorna uma matriz com os valores de M. O *M-value* é a intensidade de *probes* metiladas em relação à não-metiladas em escala logarítmica de base 2. É o principal valor utilizado para análises estatísticas de dados de metiloma.
minfi_getbeta	Retorna uma matriz com os valores de beta. O *Beta-value* é a razão de intensidade da probe metilada em relação à intensidade global (soma das intensidades das probes metiladas e não-metiladas). É utilizado em geral para visualização, como por exemplo em formato de *heatmaps*.
minfi_getsnp	Retorna o cromossomo e a posição de cada SNP.
minfi_methcpg	Retorna matriz de intensidade de metilado e não-metilado de um objeto MethylSet.
minfi_getanno	Retorna a anotação como data frame com cada linha correspondendo a sítio de metilação.

Anexo D - Ferramentas utilizadas para a análise de metilação com os respectivos arquivos de entrada e saída.

Ferramenta	Arquivo de entrada	Arquivo de saída
minfi_read450k	idat	RGChannelSet
minfi_mset	RGChannelSet	MethylSet
minfi_qc	RGChannelSet	Report (html)
minfi_ppfun	RGChannelSet	GenomicRatioSet
minfi_dropsnp	GenomicMethylset ou GenomicRatioSet	GenomicRatioSet (sem SNPs)
minfi_dmr	GenomicRatioSet e tabela de fenótipo.txt	bedgraph
minfi_getM	GenomicRatioSet	txt
minfi_getbeta	GenomicRatioSet	bedgraph
minfi_getsnp	GenomicRatioSet	txt
minfi_methcpg	Methylset	2 txt (MpGs metilados e não metilados)
minfi_getanno	GenomicRatioSet	txt