UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE CIÊNCIAS MÉDICAS

WÉLLITON DE SOUZA

ANÁLISE E DESENVOLVIMENTO DE PROTOCOLOS EM BIOINFORMÁTICA PARA ESTUDOS DE EPIGENÉTICA

CAMPINAS

2016

WÉLLITON DE SOUZA

ANÁLISE E DESENVOLVIMENTO DE PROTOCOLOS EM BIOINFORMÁTICA PARA ESTUDOS DE EPIGENÉTICA

Dissertação apresentada à Faculdade de Ciências Médicas da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciências.

ORIENTADORA: ISCIA TERESINHA LOPES CENDES

ESTE EXEMPLAR CORRESPONDE À VERSÃO
FINAL DA DISSERTAÇÃO DEFENDIDA PELO
ALUNO WÉLLITON DE SOUZA, E ORIENTADO PELA
PROFA. DRA. ISCIA TERESINHA LOPES CENDES.

CAMPINAS

2016

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2013/24801-2

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Ciências Médicas
Maristella Soares dos Santos - CRB 8/8402

So89a

Souza, Wélliton de, 1990-
Análise e desenvolvimento de protocolos em bioinformática para estudos de epigenética / Wélliton de Souza. – Campinas, SP : [s.n.], 2016.

Orientador: Iscia Teresinha Lopes Cendes.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Ciências Médicas.

1. Repressão epigenética. 2. Metilação de DNA. 3. Biologia computacional. 4. Epilepsia. I. Lopes-Cendes, Íscia Teresinha,1964-. II. Universidade Estadual de Campinas. Faculdade de Ciências Médicas. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Analysis and development of bioinformatics protocols for epigenetics studies
Palavras-chave em inglês:
Epigenetic repression
DNA methylation
Computational biology
Epilepsy
Área de concentração: Fisiopatologia Médica
Titulação: Mestre em Ciências
Banca examinadora:
Iscia Teresinha Lopes Cendes [Orientador]
Francisco Pereira Lobo
Samara Flamini Kiihl
Data de defesa: 26-01-2016
Programa de Pós-Graduação: Fisiopatologia Médica

BANCA EXAMINADORA DA DEFESA DE MESTRADO

WÉLLITON DE SOUZA

Orientador (a) PROF(A). DR(A). ISCIA TERESINHA LOPES CENDES

MEMBROS:

1. PROF(A). DR(A). ISCIA TERESINHA LOPES CENDES

2. PROF(A). DR(A). FRANCISCO PEREIRA LOBO

3. PROF(A). DR(A). SAMARA FLAMINI KIIHL

Programa de Pós-Graduação em Fisiopatologia Médica da Faculdade de Ciências Médicas da Universidade Estadual de Campinas.

A ata de defesa com as respectivas assinaturas dos membros da banca examinadora encontra-se no processo de vida acadêmica do aluno.

Data: 26 de janeiro 2016

DEDICATÓRIA

Dedico este trabalho ao meu avô,

José Ribeiro de Souza.

AGRADECIMENTOS

À minha orientadora, Profa. Dra. Iscia Teresinha Lopes-Cendes, pelos ensinamentos teóricos e práticos, pela confiança e também pelas oportunidades oferecidas ao longo do Mestrado.

Ao meu coorientador, Prof. Dr. Benilton de Sá Carvalho, pela amizade e lições ensinadas, dentro e fora do ambiente de pesquisa.

À Dr. Cristiane de Souza Rocha, pelo convite de conhecer o Laboratório de Biologia Computacional e Bioestatística que resultou na minha colaboração inicial com o grupo de pesquisa.

Ao Murilo Guimarães Borges pela amizade e cooperação nos trabalhos realizados. Agradeço também pela ajuda durante os primeiros meses na Unicamp.

Aos membros do grupo de pesquisa, Alexandre Hilário Berenguer de Matos, Amanda Donatti, Amanda Morato do Canto, Andre Schwambach Vieira, Camila Vieira Soler, Cláudia Vianna Maurer Morelli, Danyella Barbosa Dogini, Fábio Rossi Torres, Joana Prota, Kátia Brumatti Gonçalves, Marcella Bergamini de Baptista, Marina Coelho Gonsales, Mônica Paiva Quast, Patrícia Aline Oliveira Ribeiro, Patrícia Gonçalves Barbalho, Renato Oliveira dos Santos, Rodrigo Secolin, Simoni Helena Avansini, Vanessa Simão de Almeida e Vanessa Zago, pelo tempo compartilhado e cooperação nos trabalhos realizados.

Aos estagiários Lucas Cendes e Rodrigo Carmo, pelas contribuições para o Laboratório de Biologia Computacional e Bioestatística. Obrigado pela oportunidade de ensina-los e aprender com vocês.

À Faculdade de Ciências Médicas da Unicamp, pela infraestrutura e apoio durante o Mestrado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, pela bolsa de Mestrado e auxílios concedidos durante esses anos.

Aproveito a oportunidade para agradecer meu orientador de Iniciação Científica, Prof. Dr. Mauricio Bacci Junior, e a todos as pessoas que conheci no Centro de Estudos de Insetos Sociais da UNESP de Rio Claro. Um agradecimento especial à Milene Ferro e ao Erik Aceiro Antonio, por ajudarem a dar meus primeiros

passos na área de Bioinformática. Agradeço também ao meu amigo Diego Henrique Negretto e ao meu professor preferido, Orlando Saraiva do Nascimento Júnior.

Ao meu pai Darcy Ribeiro de Souza, à minha mãe Cleonice Luzia Faustino de Souza, e à minha irmã Ariadne de Souza, pelo suporte, amor e carinho. Agradeço também a todos os meus familiares.

Aos meus amigos Felipe Furlan, Guilherme Gasparini, Johann Girnos e ao meu primo Raul Sanches Faustino, pela amizade e companheirismo. Aos integrantes do moto clube Os Vira Latas do Asfalto, por compartilhar o prazer de viajar longas viagens de moto.

A todas as pessoas que participaram, contribuindo para realização deste trabalho, direta ou indiretamente, meu agradecimento.

RESUMO

As variações que ocorrem durante a divisão celular e modificam a expressão gênica sem alterar a sequência de DNA associada àquela região são denominadas modificações epigenéticas. Elas estão envolvidas no silenciamento de genes, diferenciação de tecidos e, eventualmente, nos mecanismos moleculares responsáveis por uma série de fenótipos. A metilação de DNA, uma modificação epigenética que altera a cromatina, tem sido estudada extensivamente em vários processos biológicos normais e patológicos. Com o avanço dos métodos e tecnologias de sequenciamento de DNA, hoje é possível realizar análises para determinação de perfis de metilação e suas associações com fenótipos de interesse. Esta dissertação adiciona a perspectiva analítica da epigenética associada aos estudos de modelos animais de epilepsia. Foram definidos protocolos em bioinformática para processamento de dados de metilação do DNA. Os perfis de metilação foram determinados e associados aos fenótipos de epilepsia dos animais estudados, o que também permitiu a identificação de oportunidades para o desenvolvimento de novas ferramentas e automatizações dos protocolos definidos.

Epilepsia. Epigênese Genética. Metilação de DNA. Biologia Computacional.

ABSTRACT

Epigenetic changes are those that occur during cellular division, modifying gene expression without changes on the DNA sequence. They are involved in gene silencing, tissue differentiation and in molecular mechanisms of a number of phenotypes. DNA methylation, an epigenetic modification that changes chromatin, has been studied extensively in various normal and pathologic biological processes. With the advancement of methods and technologies for DNA sequencing, it is now possible to perform analysis to determine methylation profiles and their association with phenotypes of interest. This dissertation adds the analytical perspective of epigenetics related to studies of animal models of epilepsy. Protocols were defined in bioinformatics processing of DNA methylation data. Methylation profiles were determined and associated with epilepsy phenotypes of animals studied, which also allowed the identification of opportunities for the development of new tools and automation of defined protocols.

Epilepsy. Epigenesis, Genetic. DNA Methylation. Computational Biology.

SUMÁRIO

INTRODUÇÃO

Epigenética

Epigenética é o estudo de mudanças herdáveis na função de um gene que não podem ser explicadas pelas mudanças na sequência de DNA (1). As alterações epigenéticas influenciam no fenótipo através da regulação da expressão gênica, herdada entre gerações de células (mitose) ou entre gerações de indivíduos (meiose). A herança mitótica da epigenética está envolvida na diferenciação celular e no destino da célula, enquanto que a herança meiótica da epigenética é causada pela reprogramação incompleta no embrião que resulta na propagação da informação epigenética dos pais para os filhos (2). Alguns exemplos de alterações epigenéticas são modificações nas proteínas histonas e metilação do DNA.

A metilação do DNA é a única modificação epigenética que afeta diretamente o DNA. Nela, um grupo metil é adicionado a uma base de citosina. A alteração não afeta a forma como a citosina é transcrita em RNA mensageiro mas promove localmente a compactação da cromatina, afetando o fator de ligação de transcrição (3). Nos vertebrados, a forma mais comum de metilação do DNA é a 5-metilcitosina (5mC), que afeta de 70% a 80% dos dinucleotídeos CpG, que são citosinas seguidas diretamente de uma guanina, no genoma humano (4). Quando localizado em regiões promotoras de genes, a metilação do DNA é normalmente uma marca repressiva, isto é, induz redução da expressão gênica (5).

Após a divisão celular por mitose, as fitas de DNA filhas não contêm informações de metilação. A enzima DNMT1 tem o papel de propagar a informação do perfil de metilação, ou seja, tem o papel de adicionar grupos metil nos dinucleotídeos CpG das fitas das células filhas de acordo com a informação de metilação das fitas da célula mãe. A cada divisão a informação de metilação será copiada pela DNMT1 nas novas fitas filhas, dessa forma, a metilação é considerada mitoticamente herdável.

Ilhas CpG são regiões ricas em bases GC e possuem densidade relativa alta de CpG. Essas ilhas estão frequentemente localizadas nas regiões promotoras dos genes. Regiões promotoras definem o início dos genes, é

nessa região que indutores dos mecanismos de transcrição se ligam ao DNA. As ilhas CpG tendem a ser protegidas de metilação, ou seja, a metilação não ocorre naturalmente nas ilhas CpG, entretanto a metilação de ilhas CpG pode resultar no silenciamento do gene. Além disso, um dinucleotídeo CpG pode ser ligado por uma proteína de domínio de ligação a metil, MeCP1 e MeCP2 por exemplo. Essas proteínas de ligação têm domínio de repressão transcricional e podem recrutar outros mecanismos que condensam a cromatina.

As alterações epigenéticas podem ter um papel relevante em doenças neurológicas com herança complexa, como epilepsia e acidente vascular cerebral (6). A identificação de perfis de metilação do DNA têm implicações importantes na compreensão das causas de diferenças de expressão em regiões específicas do transcriptoma em contextos de desenvolvimento específicos e também como alterações epigenéticas associam-se com doenças e padrões aberrantes de expressão (5).

Epilepsia

Epilepsia é um conjunto de doenças do cérebro caracterizadas por predisposição aumentada em gerar crises epilépticas. A crise epiléptica é decorrente de atividade neuronal anormal gerando um conjunto de sinais e/ou sintomas limitados no tempo (7). Do ponto de vista prático, o diagnóstico de epilepsia é feito por qualquer uma das seguintes condições: i) pelo menos duas crises não provocadas (sem um fator ambiental desencadeando, tal como febre, infecção ou trauma) ocorridas com uma diferença de tempo maior que 24 horas entre elas; ii) uma crise não provocada, mas acompanhada de uma alta probabilidade de ocorrência de novas crises (8).

A esclerose hipocampal (ES) é uma anormalidade histopatológica comumente encontrada em adultos com epilepsia do lobo temporal (ELT) resistente ao tratamento com medicações. Caracteriza-se por apresentar padrões específicos de perda de células nas regiões do hipocampo (região medial do cérebro humano), e é uma lesão conhecida por ter potencial epileptogênico, ou seja, ser capaz de gerar crises (9).

Uma área de grande interesse cientifico é o estudo das características moleculares de tecidos anormais gerados pelos processos

patológicos. No caso das doenças do sistema nervoso central as amostras de tecidos da região do cérebro de pacientes são difíceis de obter e as vezes são obtidas em quantidade insuficiente para estudos mais complexos. Portanto, é necessário buscar alternativas para estudar os mecanismos básicos relacionados as doenças neurológicas com o uso de modelos animais.

Existem dois modelos animais clássicos para o estudo da ELT: animais induzidos através de estímulos elétricos restritos e consecutivos na via perfurante (10) e animais induzidos por injeção de pilocarpina (11). Ambos os modelos apresentaram ES clássica, semelhante ao observado em pacientes com ELT.

Recentemente, modelos animais de epilepsias foram utilizados em um estudo de epigenética para identificar diferenças no perfil de metilação do DNA entre ratos (*Rattus norvegicus*) controles e ratos tratados com pilocarpina para indução de crises epilépticas (12). Os pesquisadores demonstraram inicialmente que havia uma diferença significativa na metilação de determinados genes nos animais que se tornaram epiléticos. Além disso, demonstraram que a administração precoce de uma dieta especial, dieta cetogênica, que tem ação conhecida como entiepiléptica, estava associada a mudanças na metilação do DNA desses animais, associando o padrão de metilação do DNA com a presença de epilepsia nesses animais (11).

Estudos epigenéticos quantificam características biológicas que são dinâmicas ao longo da vida e essas características são suscetíveis a estímulos do ambiente externo. A natureza dinâmica das marcas epigenéticas exigem experimentos bem detalhados e análises capazes de considerar fatores não biológicos que podem afetar os resultados desses estudos.

Estudos de associações epigenéticas

Estudos de associação em epigenomas completos (EWAS) examinam o estado epigenético de vários *loci* para um conjunto de indivíduos e avaliam se algum desses *loci* estão associados com o fenótipo de interesse (13). Em geral, esses estudos comparam o perfil de uma determinada marca epigenética, principalmente a metilação do DNA, de um tecido específico, entre amostras controle e amostras com a presença da doença ou outra

característica de interesse para o estudo. A Figura 1 apresenta o passo-a-passo para realização de um estudo de associações epigenéticas.

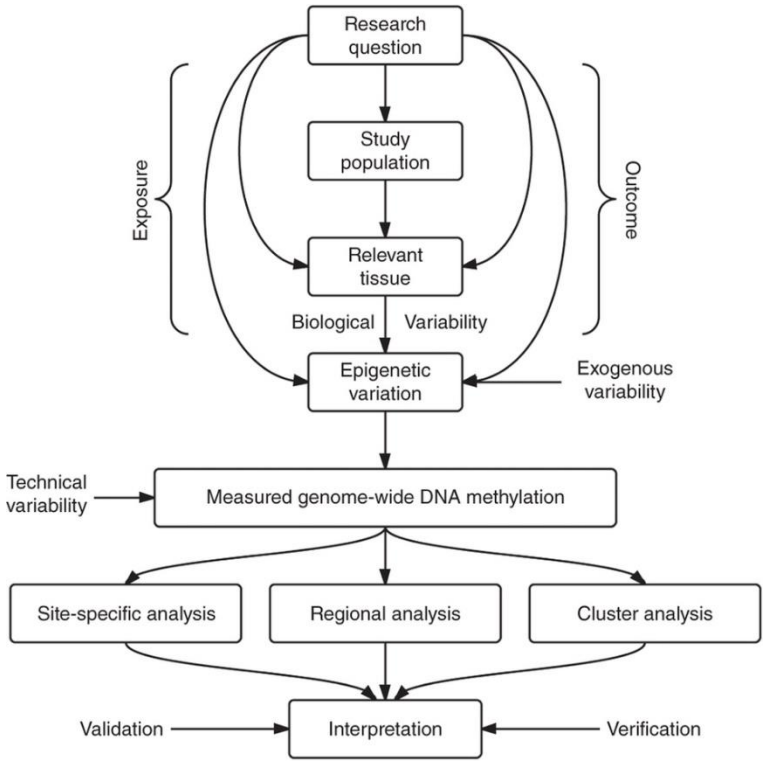


Figura 1 - Passo-a-passo de um EWAS. A exposição (drogas, envelhecimento) ou o fenótipo (doença) definirão a pergunta da pesquisa, a população do estudo, o tecido relevante e a variabilidade biológica presente no perfil de marcas epigenéticas. A utilização de protocolo e tecnologia para detecção de metilação do DNA induzirá variabilidade técnica nos dados gerados. Os dados poderão ser analisados região por região, por regiões especificas ou análise de agrupamentos. Análise região por região é realizar testes univariados de associação para identificar regiões onde a metilação de citosinas varia com a exposição ou fenótipo, seguido de ajustes de múltiplos testes como o FDR. Análise de regiões específicas é identificar regiões diferencialmente metiladas (DMR) considerando os níveis de metilação adjacentes, o que aumenta o poder estatístico. Isso é possível porque a metilação do DNA é encontrada frequentemente em agrupamentos de dinucleotídeos CpG. Análise de agrupamentos é utilizar métodos de agrupamento não supervisionada para agrupar dinucleotídeos CpG com nível de metilação próximos ao longo do genoma. Essa abordagem diminui o volume de dados e contribui para testes estatísticos de associação o perfil de metilação do DNA e fenótipos de interesse. A última etapa do estudo é interpretar os resultados das análises considerando a validação desses resultados com novos experimentos. Retirado de (13).

O mapeamento da metilação do DNA é feito pela combinação de técnicas moleculares que evidenciam marcas epigenéticas, e tecnologias de quantificação, como o sequenciamento de nova geração (NGS) e microarranjo. Isso é necessário porque durante a amplificação por reação em cadeia da polimerase (PCR) as informações de metilação do DNA são perdidas. No entanto, já existem novas tecnologias de sequenciamento capazes de identificar diversas modificações de bases na resolução de uma única molécula. Os métodos para mapeamento do perfil de metilação do DNA são utilizados como ferramentas para identificar regiões candidatas diferencialmente metiladas quando dois ou mais grupos de amostras são comparados entre si (14,15).

Os métodos de mapeamento do perfil de metilação do DNA criam amplas oportunidades para a pesquisa epigenética, mas eles também colocam desafios substanciais em termos de processamento de dados, análise estatística e interpretação biológica das diferenças observadas (2). A principal técnica de quantificação de metilação do DNA é o tratamento de moléculas de DNA com bissulfito de sódio seguido de sequenciamento de alto rendimento.

Sequenciamento por bissulfito

A técnica molecular de conversão química do DNA por bissulfito pode ser utilizada para evidenciar citosinas metiladas. Nesse tratamento, o bissulfito de sódio converte as **citosinas não metiladas** em **uracilas**, mantendo as citosinas metiladas inalteradas. Após a conversão da amostra de DNA, o próximo passo é a amplificação por PCR, que converte as uracilas em timinas. O tratamento químico do DNA por bissulfito permite a quantificação da metilação de todo o genoma através do sequenciamento de alto rendimento (WGBS). Esse método oferece ampla cobertura genômica, quantificação precisa com resolução por base e boa reprodutibilidade (14). Entretanto seu custo é alto, e conversão química por bissulfito é incapaz de discriminar entre 5-metilcitosinas (5mC) e 5-hidroximetilcitosina (5hmC).

Para diminuir o custo de estudos de epigenéticos, foram desenvolvidos métodos alternativos ao tratamento por bissulfito capazes de

enriquecer fragmentos de DNA que possuem dinucleotídeos CpG metilados, seguido de sequenciamento de alto rendimento.

Sequenciamento por enriquecimento

Os métodos baseados em enriquecimento utilizam anticorpos específicos para reconhecer a metilação do DNA, proteínas domínio de ligação a metil, ou ainda enzimas de restrição para enriquecer os fragmentos de DNA metilado (ou não metilado). O material genético enriquecido é quantificado através do sequenciamento de alto desempenho. O método tem um custo de sequenciamento por amostra considerado relativamente baixo, é eficiente para cobrir todo o genoma e é capaz de distinguir entre 5mC e 5hmC (14). Entretanto, este método resulta em menor poder estatístico para discriminar diferenças em regiões genômicas pobres de CpGs. Além disso, esses métodos são altamente suscetíveis a vieses induzidos durante o experimento. Qualquer flutuação na cobertura do sequenciamento irá afetar diretamente o nível de metilação obtido em na região específica do DNA. É também necessário corrigir diferenças de densidade de CpG de cada região genômica estudada.

Nos métodos de enriquecimento é utilizado um anticorpo específico ou uma proteína de domínio de ligação a metil (MBD) para selecionar e capturar fragmentos de sequências, previamente fragmentados por sonicação, que possuam citosinas metiladas; aqui, os fragmentos restantes são descartados. A partir da PCR, os fragmentos selecionados são enriquecidos e, posteriormente, quantifica-se o nível de enriquecimento dos fragmentos a partir de equipamentos de sequenciamento de alto rendimento. O método MeDIP-seq (16) e MethylCap-seq (17) utilizam anticorpo e MBD respectivamente. Os dois métodos produzem picos de metilação com forma, tamanho e localização genômica similares. Isso sugere que os dois métodos enriquecem fragmentos de DNA similares (14).

Há, entretanto, um viés nos resultados desses dois métodos: o enriquecimento dos fragmentos é dependente da densidade de CpG da região genômica. Isso significa que contagens baixas de fragmentos podem indicar tanto a falta de CpGs naquela região, quanto a falta de metilação de DNA na presença de CpGs (14). Por isso, é necessário aplicar uma correção estatística

que envolve a contagem de fragmentos alinhados e a densidade de CpGs para cada região genômica melhorando os resultados de níveis de metilação do DNA (15).

A resolução do nível de metilação reportado é dependente do tamanho dos fragmentos de DNA enriquecidos. Essa limitação não é um problema em estudos que buscam avaliar o nível de metilação em regiões genômicas como ilhas CpGs e regiões promotoras de genes.

Outra limitação é, dado que o número de *reads* alinhados em uma região genômica indica apenas o nível *relativo* de metilação nessa região. O nível relativo de metilação não pode ser utilizado para comparar diferentes amostras porque cada amostra pode ter um tamanho de biblioteca de sequenciamento diferente. É necessário utilizar métodos estatísticos para normalizar o nível relativo de metilação levando em conta o tamanho das bibliotecas de sequenciamento de todas as amostras. Essa limitação pode tornar-se um problema em estudos onde as amostras possuem baixos níveis de cobertura de sequenciamento.

Além dos problemas relacionados com métodos baseados em captura de DNA metilado, há dois problemas específicos identificados na técnica MethylCap-seq que podem comprometer os resultados da análise (18). Durante a etapa de captura dos fragmentos metilado podem ocorrer falhas resultando no sequenciamento de fragmentos não metilados. Isso pode acarretar na ausência ou inconsistência do enriquecimento de uma amostra. Outro problema é o limite do poder estatístico nos testes de metilação diferencial devido a bibliotecas de sequenciamento com baixa cobertura de CpG e baixa complexidade dos fragmentos.

Embora os estudos epigenéticos podem querer responder a perguntas semelhantes, o uso de diferentes métodos de quantificação de DNA exige protocolos em bioinformáticas específicos para cada tipo de dados, sendo compostos por uma coleção de programas de computador.

A metilação do DNA é uma marca binária, ou seja, o grupo metil está ou não presente em uma determinada base, mas os protocolos atuais utilizam milhões de cópias de DNA para quantificar o perfil de metilação. Dessa forma,

os sinais epigenéticos são considerados medidas quantitativas que representam a porcentagem de moléculas de DNA que estão metiladas em uma determinada região.

Bioinformática

Independente da tecnologia utilizada para mapeamento do perfil de metilação do DNA, a grande quantidade de informações gerada e a natureza dinâmica dos sinais epigenéticos, que são variáveis ao longo da vida, tornaram os métodos estatísticos para interpretação desses dados ainda mais complexos (16).

Análise de dados de WGBS

Os protocolos de processamento e análise de dados de WGBS e relacionados foram revisados recentemente (19,20). Os dados brutos devem passar pelo controle de qualidade (QC) para avaliação da qualidade do sequenciamento e a porcentagem de citosinas ao longo dos ciclos, pode ser utilizado os programas como FastQC e Rqc. Espera-se que a porcentagem de citosinas esteja abaixo de 5% ao longo de todos os fragmentos de sequencias (20). É importante que a qualidade média dos fragmentos seja alta porque quanto menor a qualidade das bases menor a taxa no alinhamento. Outro fator que afeta a taxa de alinhamento é a contaminação de adaptadores que deverão ser removidos utilizando ferramentas adequadas como o Trimmomatic (21) e Trim Galore!.

Para medir o nível de metilação do DNA, é necessário identificar o estado de metilação de cada base, que pode estar metilada ou não. Para tanto, é realizado o alinhamento das sequências de DNA tratadas com bissulfito a um genoma de referência. Quando uma base C de uma sequência tratada sobrepõe uma base C na referência, infere-se que a base C da referência é, no mínimo, metilada em uma molécula da amostra. Para obter maior precisão nos resultados, é calculado o nível absoluto de metilação por base, ou seja, a proporção de bases metiladas que alinharam na mesma posição no genoma de referência. Devido à conversão de bases durante as etapas de tratamento por bissulfito e sequenciamento, os métodos computacionais de alinhamento por

referência devem considerar bases ambíguas, pois uma base T pode ser literalmente timina ou uma citosina convertida em uracila.

Atualmente, existem dois métodos de alinhamento por referência para sequências tratadas por bissulfito. Ambos removem a penalidade na pontuação associada ao alinhamento de bases C e T das sequências de entrada (*reads*) com bases C no genoma de referência. Estes métodos são conhecidos como alinhamento *Wild-card* e *Three-letter* (19). A Figura 2 apresenta os dois métodos de alinhamento por referência para determinação do nível de metilação.

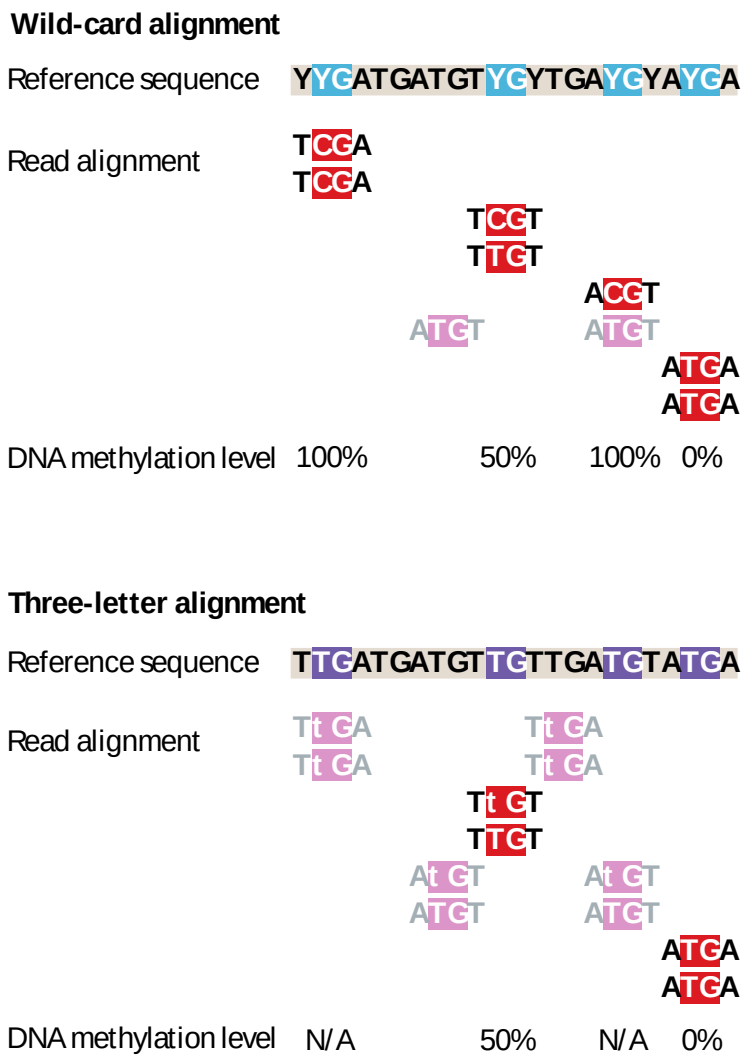


Figura 2 - Em *Wild-card alignment*, o alinhador substitui as letras C na sequência de referência pela letra Y (curinga), que alinha com as letras C e T nas sequências de entrada. Em *Three-letter aligment*, o alinhador substitui as letras C na sequência de referência pela letra T (maiúscula) e pela letra t (minúscula) nas

sequências de entrada. Ambos os métodos descartam sequências que alinharam em mais de um local. Para cada CpG, é calculado o nível de metilação, o percentual de C e T é determinado entre todos os *reads* que foram alinhados em cada C na sequência de referência. Extraído de (19).

No método *Three-letter*, uma etapa de pré-processamento é realizada antes do alinhamento. Nessa etapa, todas as citosinas são convertidas para timinas tantos nos fragmentos sequenciados, quanto no genoma de referência. Esse método tem como vantagem a possibilidade de utilizar alinhadores conhecidos, como o Bowtie (22). Um exemplo de alinhador *Three-letter* é o programa Bismark (23). Esse programa converte os fragmentos e o genoma de referência em um alfabeto de apenas três letras para depois executar o alinhador Bowtie.

O método Wild-card procura lidar com bases ambíguas convertendo as citosinas da sequência de referência para pirimidinas, o que permite que citosinas e timinas das amostras sequenciadas sejam alinhadas a estas posições (24).

Os alinhadores *Wild-card* podem atingir coberturas maiores do genoma ao custo de introduzir tendência de aumento dos níveis de metilação do DNA. Alinhadores *Three-letter* minuem a complexidade das sequencias removendo as bases C, de tal modo que uma porcentagem maior de sequencias são rejeitadas devido a alinhamentos múltiplos. Dessa forma, os alinhadores *Wild-card* reportam coberturas de alinhamento mais altas que alinhadores *Three-letter* ao custo de introduzir viés na tabela de frequência de metilação (19). Métodos estatísticos como o pacote bsseq (25), filtram os CpGs de acordo com a cobertura de alinhamento. Em resultados de alinhadores *Three-letter*, isso pode representar uma quantidade menor de dinucleotídeos CpG testados pelo método estatístico.

Após o alinhamento dos fragmentos de sequência contra o genoma de referência é importante reavaliar a qualidade dos fragmentos alinhados utilizando métricas específicas para dados de WGBS, como por exemplo o gráfico M bias (26).

A tabela de frequência de metilação contém a posição genômica de cada base, a informação de sentido da fita (para dados *paired-end*), a quantidade de *reads* que alinharam sobre aquela posição (cobertura), e a frequência de bases C que foram alinhadas naquela posição (frequência de metilação). O pacote *methylKit* (27) lê os arquivos de alinhamento e gera uma tabela de frequência de metilação. Por padrão, o pacote mantém apenas posições genômicas onde a cobertura é de pelo menos 10 *reads* e todas as bases que alinhara nessa posição tenha pelo menos qualidade 20 na escala PHRED (28). A ferramenta quantifica tanto a metilação de citosinas no contexto de CpG quanto no contexto CHH (onde H pode ser A, T ou C). O pacote methylKit carrega os dados a partir dessa tabela de frequência de metilação.

A ferramenta Bismark possui um programa utilitário para gerar a tabela de frequência de metilação a partir dos dados de alinhamento, entretanto a tabela gerada não possui informação de sentido da fita. O pacote estatístico *bsseq* (25) carrega os dados a partir dessa tabela de frequência de metilação.

Análise de dados de MethylCap-seq

Os dados brutos devem ser avaliados por ferramentas de controle de qualidade, principalmente após o alinhamento dos fragmentos em busca de problemas de sequenciamento ou enriquecimento do material genético.

Os dados gerados por métodos baseados em enriquecimento de DNA metilado seguido por sequenciamento de alto rendimento devem ser alinhados contra o genoma de referência da mesma espécie que as amostras são derivadas. É importante avaliar a qualidade dos fragmentos sequenciados antes do alinhamento e também verificar o grau de saturação de cada amostra e a cobertura obtida após o alinhamento.

Os fragmentos duplicados, que alinharam exatamente na mesma localização genômica, podem ser réplicas de PCR ou representar abundância de DNA metilado em uma mesma região. Para controlar possíveis artefatos de PCR, é recomendado remover os fragmentos duplicados. O programa MEDIPS (29) identifica e substitui os fragmentos sequenciados com a mesma posição

genômica e orientação de fita por uma representante contabilizando apenas uma vez.

O tamanho dos fragmentos de DNA metilado capturados pelo método MethylCap-seq pode variar, por exemplo 150-200 pares de bases (30) e 200-300 pares de bases (12). Entretanto, o tamanho dos fragmentos reportados pelo sequenciador pode ser menor que o fragmento capturado pela técnica MethylCap-seq, como por exemplo 32 pares de bases (12). Dessa forma é necessário estender virtualmente o tamanho do fragmento de acordo com o tamanho médio dos fragmentos capturados.

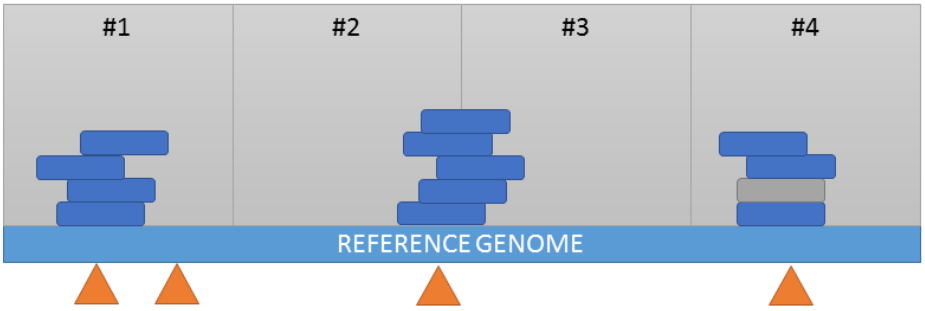
São necessários controles de qualidade específicos para dados MethylCap-seq que destacam os possíveis problemas associados aos métodos de enriquecimento de DNA metilado seguido de sequenciamento de alto desempenho. Foram definidos três métodos de controle de qualidade para dados de enriquecimento (29): (i) enriquecimento de CpG; (ii) análise de curva de saturação; e (iii) cobertura de CpG.

O enriquecimento de CpG é a frequência de dinucleotídeos CpG observada na amostra sequenciada comparada com a frequência esperada. Isso pode indicar falhas no início da etapa de captura de fragmentos. A pontuação de enriquecimento é calculada dividindo a frequência relativa de CpGs interrogados pelos fragmentos de sequência alinhados pela frequência relativa de CpG no genoma. O parâmetro de corte para o enriquecimento de CpG é definido em 0,4 segundo (18).

A análise de curva de saturação é uma estimação do coeficiente de correlação de Pearson da complexidade da biblioteca de sequenciamento e da reprodutibilidade. A análise de saturação ajuda a verificar se o conjunto de *reads* mapeados é suficiente para gerar um perfil de cobertura reprodutível e saturada do genoma de referência. Amostras com saturação baixa (menor que 0.5) sugerem dificuldades em reproduzir o perfil de metilação se a biblioteca for sequenciada novamente.

A cobertura de CpG é a fração de dinucleotídeos CpG no genoma de referência sequenciados de pelo menos 5 vezes.

Após o alinhamento e controle de qualidade específico para dados de sequenciamento por enriquecimento, mede-se o sinal de metilação ao longo do genoma. O pacote MEDIPS implementa o método de janelas genômicas (16,29). Nesse método o genoma de referência é dividido em janelas adjacentes. O nível de metilação é dado pela contagem de *reads* que alinharam dentro de uma janela genômica. O tamanho da janela genômica determina a resolução computacional desse método. Tem sido utilizado janelas com tamanho fixo de 500 pares de bases (18,30). Utilizar janelas menores pode diminuir o poder estatístico devido à baixa contagem de *reads* por janela. Quanto menor o tamanho da janela maior o requisito de memória principal e maior a tabela de resultado da contagem. O método de janelas genômicas pode contabilizar erroneamente *reads* a uma janela destituída de dinucleotídeos CpG. A Figura 3 exemplifica o método de janelas genômicas.



Bin ID	Number of reads	Number of CpGs
#1	4	2
#2	5	1
#3	5	0
#4	3	1

Figura 3 – Método de janelas genômicas. O genoma de referência é dividido em janelas adjacentes de tamanho fixo (representadas pelas caixas cinzas #1, #2, #3 e #4). Os fragmentos que alinharam dentro dessas janelas são contados gerando a tabela de contagem com as colunas número da janela, quantidade de fragmentos e quantidade de CpGs. Os fragmentos que alinharam na região de duas janelas são contados duas vezes. Os fragmentos que alinharam na mesma posição genômica são contados apenas como um (fragmento cinza na janela #4).

Na contagem dos *reads* que alinharam dentro de janelas genômicas, é comum janelas que não possuem dinucleotídeos CpG reportarem contagens (31). Esse artefato ocorre porque o fragmento alinhado, que foi virtualmente estendido, pode estar presente em mais de uma janela genômica e dessa forma é contado em janelas adjacente que podem não ter CpGs. Outro fator são fragmentos não metilados que foram sequenciados devido algum problema durante a etapa de captura de fragmentos. Uma maneira de resolver este artefato é remover janelas sem dinucleotídeos CpG do cálculo de cobertura diferencial. A Figura 4 apresenta o total de janelas com contagens de fragmentos, mas sem contagem de CpGs.

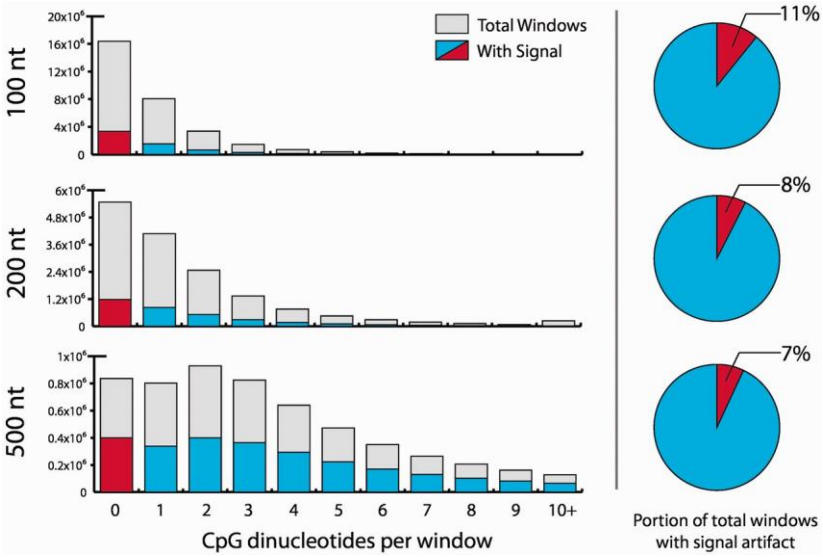


Figura 4 - Total de janelas e contagem de fragmentos ao longo de janelas com diferentes quantidades de dinucleotídeos CpG. Fragmentos contados em janelas que não possuem CpGs podem representar até 11 % do total de janelas com contagens. Retirado de (31).

Definiu-se um protocolo de análise para dados de MethylCap-seq (18). Nesse protocolo, os fragmentos alinhados ao genoma de referência são virtualmente estendidos para o tamanho médio dos fragmentos enriquecidos. O tamanho dos fragmentos é determinado pelo analisador de fragmentos da BioAnalyzer. Os fragmentos estendidos alinhados são contados em janelas genômicas de 500 pares de bases ao longo do genoma. A distribuição da tabela de contagem é normalizada pelo total de *reads* alinhados. Os valores das contagens são convertidos para *reads* por milhões (RPM). As janelas

genômicas são agrupadas de acordo com as informações de localização genômica de regiões de interesse, como por exemplo ilhas CpG, costa (das ilhas) CpG e regiões promotoras. A contagem normalizada dessas janelas é somada para cada *locus* associado a região de interesse. As regiões diferencialmente metiladas é dado pelo teste estatístico Wilcoxon entre grupos de amostras. O teste estatístico é realizado para cada *locus* e depois o p-valor reportado é corrigido para múltiplos testes (FDR). O valor mínimo de corte do p-valor ajustado é 0,05.

Outro protocolo para quantificação do nível de metilação em dados de MethylCap-seq foi implementado na ferramenta PrEMeR-CG (31). A ferramenta reporta valores de metilação em resolução de base (alta resolução), assim como ocorre em dados de bissulfito seguido de sequenciamento. Segundo os autores os resultados são favoráveis quando comparados com dados de bissulfito e tem maior poder preditivo que métodos baseados em janelas. Para obter o nível de metilação para cada dinucleotídeo CpG em dados de captura, a ferramenta depende do perfil do tamanho dos fragmentos capturados. O kit *Agilent Bioanalyzer High Sensitivity DNA* permite analisar fragmentos de DNA produzindo um histograma do tamanho dos fragmentos para cada amostra. Esse histograma é utilizado pelo programa PrEMeR-CG para predizer o nível de metilação em alta resolução. Essa ferramenta é inviável para analisar dados públicos de metilação porque a única informação disponibilizada são arquivos contendo os fragmentos sequenciados.

Estatística para estudos genômicos

A abordagem mais comum para análises comparativas em estudos genéticos é testar a hipótese nula de que o log-razão entre grupos de interesse (em geral, grupos tratado e controle) de uma medida genômica observada (como expressão) seja zero. Divergências desta hipótese sugerem efeito diferencial entre os grupos. O objetivo da análise diferencial é produzir uma lista de itens candidatos a efeitos diferenciais e o modo mais comum para a obtenção de tal lista é ordenar os itens testados de acordo com seus p-valores corrigidos para múltiplos testes (32).

A metilação diferencial pode ser feita através de estratégias diferentes que incluem aquelas implementadas nos pacotes edgeR (33) e DESeq2 (32). Os métodos edgeR e DESeq2 são recomendados para estudos que possuem réplicas biológicas. Nesse caso os dados são normalizados pelo tamanho da biblioteca (quantidade de reads). O edgeR provê a normalização e a modelagem para a análise de metilação diferencial entre os grupos a partir da tabela de contagem de *reads*. O DESeq2 é um novo método utilizado para expressão diferencial de RNA-seq, embora não tenha sido utilizado para dados de métodos baseados em enriquecimento de DNA metilado.

Correção para múltiplos testes

O teste de hipótese é um método de inferência estatística baseado na análise de uma amostra aleatória. Este teste requer a definição de duas hipóteses: nula e alternativa. A hipótese nula descreve o *status quo*, i.e., a igualdade de tratamentos ou ausência de efeito diferencial, conforme supracitado. A hipótese alternativa assume a existência de efeito diferencial entre os tratamentos considerados. O teste de hipóteses permite a tomada de decisão (rejeitar ou não a hipótese nula) baseada na evidência provida pelo conjunto de dados em questão. Em termos práticos, pesquisadores pré-definem um valor de referência (representado por α e habitualmente fixado em 5%) e verificam se o valor de significância do teste (p-valor) é menor que α, situação na qual rejeita-se a hipótese nula. Se o p-valor for maior que α, opta-se por não rejeitar a hipótese nula.

Ao realizar um teste de hipóteses, o analista pode cometer dois erros: A) rejeitar a hipótese nula quando ela era a opção correta; B) não rejeitar a hipótese nula quando ela é a opção incorreta. O primeiro erro acima citado é denominado Erro Tipo I (falso positivo); enquanto o segundo, Erro Tipo II (falso negativo).

Estudos genômicos habitualmente consistem na realização de um teste de hipóteses para cada região (genômica) de interesse. Por exemplo, em estudos de expressão em humanos comparando grupos tratado a controle, testa-se a existência de expressão diferencial em aproximadamente 20.000 genes. Para cada gene, realiza-se um teste de hipóteses, totalizando-se

aproximadamente 20.000 testes de hipóteses. A frequência de erros (tipos I e II) aumenta de acordo com a multiplicidade de testes, fazendo com que as listas de itens candidatos possuam quantidades elevadas de falsos-positivos. Faz-se necessária, então, a aplicação de métodos que permitam controlar o número de erros na lista de candidatos a efeitos diferenciais. Estes métodos *ad-hoc* são conhecidos como métodos de correção de múltiplos testes.

O método de Bonferroni é uma estratégia bastante conservadora para correção de múltiplos testes. Com ele, se foram realizados *n* testes de hipóteses, então o procedimento permite a rejeição da hipótese nula apenas para aqueles testes cujos p-valores sejam inferiores a α/n (note que anteriormente rejeitavam-se as hipóteses cujos p-valores eram inferiores a α). Desta maneira, as listas de itens candidatos acabam significativamente reduzidas e, em algumas vezes, tais listas possuem comprimento nulo.

O método Benjamini–Hochberg (BH) ou false discovery rate (FDR) também controla a proporção de hipóteses nulas incorretamente rejeitadas (erro tipo I) utilizando a seguinte estratégia:

1. Para um α pré-estabelecido, encontra-se o maior *k* para o qual: $p_k \leq \frac{k}{n}\alpha$;
2. Rejeitam-se todas as hipóteses nulas H1, H2, ..., Hk.

O método BH é escalável permitindo ser utilizado em estudos com diferentes quantidades de variáveis. Isso é possível porque o Q-valor é proporcional a quantidade total de descobertas. Por exemplo se o Q-valor é definido como 5% então um estudo com 100 descobertas é aceitável ter 5 falsos positivos, assim como também é aceitável ter 50 falsos positivos em um estudo com 1000 descobertas. O pacote estatístico DESeq2 (32), utilizado em protocolos de análise para identificação de genes candidatos a expressão diferencial corrige os p-valores utilizando o método BH por padrão.

O p-valor ajustado deve ser utilizado apenas como guia em estudos seguintes para validação dos genes candidatos a expressão diferencial como por exemplo experimentos de bancada.

Pesquisa reprodutível

Uma pesquisa científica deve ser reprodutível e replicável. A reprodutibilidade á a habilidade de recalcular os resultados de uma análise a partir de um conjunto de dados observado e conhecimento do protocolo de análise utilizado. A replicabilidade é a chance que um experimento independente visando a mesma pergunta científica irá produzir resultados consistentes (34).

Para que uma pesquisa seja reprodutível é necessário que os dados brutos do experimento estejam disponíveis; o código estatístico e a documentação para reproduzir a análise também estejam disponíveis; e deve ser realizado uma análise correta dos dados. Por outro lado, alguns problemas podem afetar a reprodutibilidade e a replicabilidade dos dados como por exemplo variáveis omitidas, desenho do estudo ruim e dados faltantes. É necessário também identificar métodos estatísticos, protocolos de análise e programas de computador que incrementam a replicabilidade e reprodutibilidade da pesquisa. A Figura 5 apresenta o espectro da reprodutibilidade da pesquisa.

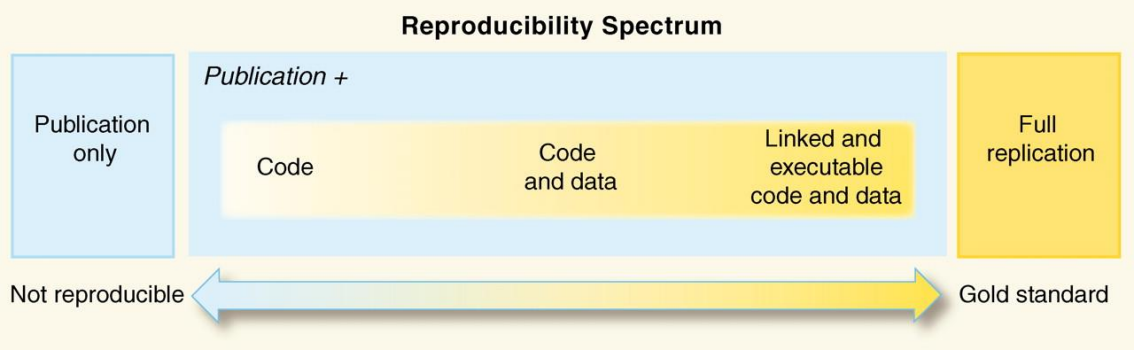


Figura 5 – Espectro da reprodutibilidade da pesquisa. A publicação de um artigo científico torna-se mais reprodutivo com a adesão (como material suplementar) de artefatos da pesquisa, como o código da análise, os dados analisados e seus resultados, e ainda a disponibilidade pública de todos esses artefatos. Retirado de (35).

Estão disponíveis ferramentas computacionais que simplificam o processo de distribuir análises reprodutíveis, como por exemplo o IPython e o knitr. O knitr é um pacote de software projetado para gerar relatórios dinâmicos com o programa estatístico R. Tem a mesma utilidade que os

programas Latex e Sweave com o diferencial de ser capaz de gerar relatórios no formato web HTML e utilizar a linguagem de marcação Markdown simplificando o desenvolvimento de relatórios de análise de dados.

OBJETIVOS

Objetivo principal

Analisar protocolos e métodos computacionais em bioinformática para estudos de associação dos perfis de metilação de modelos animais de epilepsia de lobo temporal mesial (ELTM).

Objetivos específicos

- Gerar conhecimento técnico-científico em protocolos de análise para determinação dos perfis de metilação de modelos animais de ELMT.
- Avaliar o desempenho e precisão dos métodos computacionais utilizados para estudos de associação do epigenoma.
- Propor possíveis automatizações dos protocolos em bioinformática para estudos de associação ao longo do epigenoma.

MATERIAL E MÉTODOS

Dados públicos

Para o desenvolvimento e teste da ferramenta *methylCap*, foi utilizado um conjunto de dados públicos referente ao estudo epigenético com modelos animais de epilepsia (12). Desse conjunto de dados, foram utilizadas amostras de modelos animais administrados com pilocarpina (n = 4) e animais controle (n = 5). O DNA das amostras foi fragmentado a um tamanho mediano de 200-300 pares de bases por meio de sonicação. O enriquecimento foi feito pela captura do DNA metilado usando o protocolo *MethylMiner*, método MethylCap-seq. As bibliotecas de DNA foram sequenciadas utilizando o *Illumina Genome Analyzer IIx*, que produziu sequências com 36 pares de bases de comprimento, fita única.

Dados inéditos

Dados inéditos foram gerados pelo nosso grupo de pesquisa no âmbito da investigação biológica de modelos animais de epilepsia. Esses dados são resultados de um ensaio piloto. Foram utilizados dados de WGBS provenientes de modelos animais de epilepsia e animais controles. Os dados consistem em duas amostras de animais controle e duas amostras de modelos animais de epilepsias induzidos pelo método da pilocarpina (11). Todas as amostras de DNA foram tratadas com bissulfito utilizando o protocolo *Illumina TruSeq DNA Methylation* e então sequenciados utilizando o equipamento *Illumina HiSeq 2500*.

Ambiente computacional

As tarefas de análise e seleção dos programas de bioinformática, assim como o desenvolvimento e execução dos protocolos de bioinformática, foram realizadas nos servidores do Laboratório de Biologia Computacional e Bioestatística. Os servidores são equipados com sistema operacional GNU/Linux Ubuntu Server 14.04. Todos os servidores possuem no mínimo 8 unidades de processamento e mais de 100 GB (gigabytes) de memória principal, além de cerca de um total de 5 TB (terabytes) de espaço em disco disponível.

Os protocolos em bioinformática foram desenvolvidos na linguagem de programação e ambiente estatístico R. Diversos pacotes do projeto Bioconductor foram utilizados para auxiliar o desenvolvimento e analisar os dados. Foi utilizado o ambiente de desenvolvimento de desenvolvimento integrado RStudio para análise e o serviço GitHub para gerenciamento de códigos.

Formato dos protocolos

Os protocolos em bioinformática desenvolvidos foram escritos em um formato reprodutível e replicável. Esse formato é obtido a partir da utilização de várias ferramentas para criação de relatórios dinâmicos. Os arquivos-fonte foram escritos na linguagem de marcação Markdown com trechos de códigos em R. Os arquivos-fonte foram processados pelo pacote knitr. O resultado corresponde a arquivos nos formatos PDF e HTML que contêm informações sobre documentação, códigos e resultados dos códigos executados, incluindo textos, tabelas e figuras. O documento final é: A) reprodutível, podendo ser executado novamente em outro ambiente computacional ou em um outro momento, gerando exatamente os mesmos resultados; e B) replicável, podendo ser utilizado como molde para a análise de outros conjuntos de dados cujos protocolos de geração e objetivos analíticos sejam concordantes com os definidos para os nossos estudos.

RESULTADOS E DISCUSSÃO

Software para processamento de dados de MethylCap-seq

Desenvolveu-se o programa *methylCap* para processamento e análise de dados gerados pelo método baseado em enriquecimento de DNA metilado MethylCap-seq. A ferramenta foi implementada na linguagem de programação R e estruturada como um pacote do projeto Bioconductor (36). Nós optamos por utilizar as soluções existentes para cada tarefa de processamento e análise, e adequar essas soluções para dados de MethylCap-seq.

O pacote MEDIPS é utilizado internamente pela ferramenta para dividir o genoma referência em janelas adjacentes com tamanho fixo. O pacote DESeq2 é utilizado na normalização da contagem, metilação diferencial entre dois grupos de amostras e correção dos p-valores para múltiplos testes. Nós utilizamos o pacote DESeq2 porque temos acesso à contagem normalizada, estatística não oferecida pelo edgeR. A contagem normalizada é utilizada para calcular o nível de metilação nas regiões genômicas de interesse e calcular a distribuição da metilação global das amostras. A ferramenta também utiliza a contagem normalizada para a gerar o gráfico *heatmap* da análise de agrupamento não supervisionada.

Além do gráfico de *heatmap*, o pacote methylCap também disponibiliza os gráficos *Volcano Plot* e *MA Plot* importantes para averiguação entre a metilação diferencial das janelas genômicas e o nível de significância desses resultados. O gráfico de distribuição da metilação global é um indicativo visual das diferenças entre amostra de mesmo grupo e de grupos diferentes.

A atual implementação do pacote methylCap foi estruturada para manter os dados de análise unificados. Esses dados são armazenados em uma estrutura comum utilizada por todas as funcionalidades disponibilizadas pelo pacote. Essa decisão de projeto implica na facilidade de uso da ferramenta e também habilita o processamento parcial e seletivo das etapas com suporte a recuperação de erros.

Após o processamento dos dados, a ferramenta desenvolvida gera um relatório contendo todas as informações de entrada, como configurações e arquivos utilizados, e resultados de todas as etapas. O relatório pode ser gerado como um arquivo no formato web HTML ou no formato PDF. A partir do documento gerado, é possível reproduzir a análise de bioinformática em outro ambiente de pesquisa.

Análise dos dados de MethylCap-seq

O protocolo de análise para dados de MethylCap-seq é dividido em três etapas, pré-processamento, processamento e análise. A etapa de pré-processamento é a obtenção do conjunto de dados pela internet, ou seja, essa etapa é necessária apenas para dados públicos. A etapa de processamento consiste no controle de qualidade, alinhamento dos fragmentos sequenciados contra o genoma de referência e a avaliação da qualidade específica para dados de enriquecimento. A etapa de análise é apresentada na Figura 6.

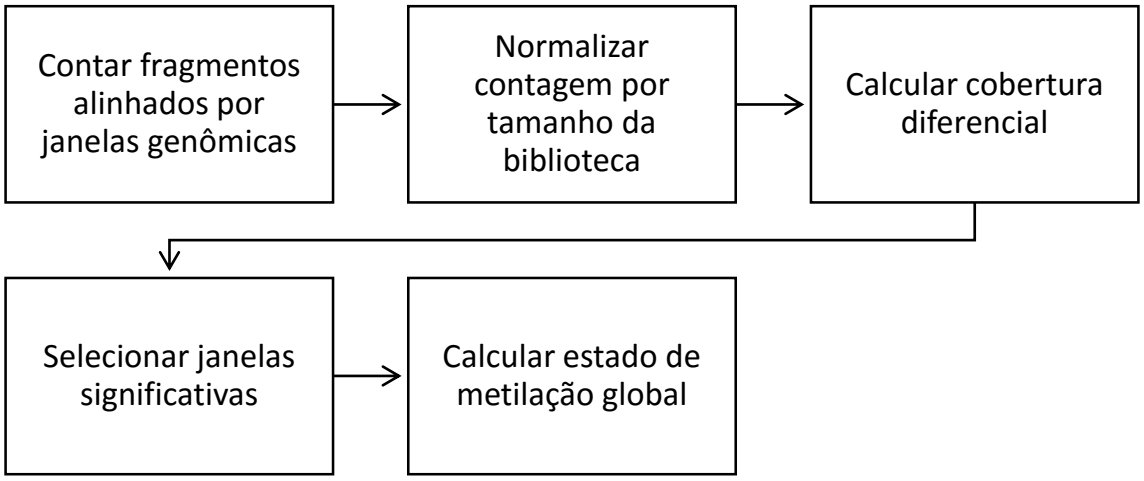


Figura 6 - Etapas da análise de dados de MethylCap-seq. Baseado em (12,18,29).

Pré-processamento

Foi utilizado um conjunto de dados públicos referentes a um estudo de associação entre o perfil de metilação de DNA e expressão gênica de

modelos animais de epilepsia (12). Obtivemos os identificadores dos arquivos referentes às amostras contendo os fragmentos de sequências a partir da consulta no serviço *Gene Expression Omnibus* (GEO). Os arquivos são transferidos do banco de dados públicos *Sequence Read Archive* (SRA), e posteriormente convertidos para arquivos no formato FASTQ (28).

Processamento

O controle de qualidade dos dados de sequenciamento foi realizado por meio da ferramenta Rqc. A ferramenta processou uma amostra aleatória de um milhão de fragmentos de cada arquivo FASTQ. A Figura 7 apresenta um gráfico sobre a distribuição da qualidade média dos *reads* de cada amostra, colorido por grupo.

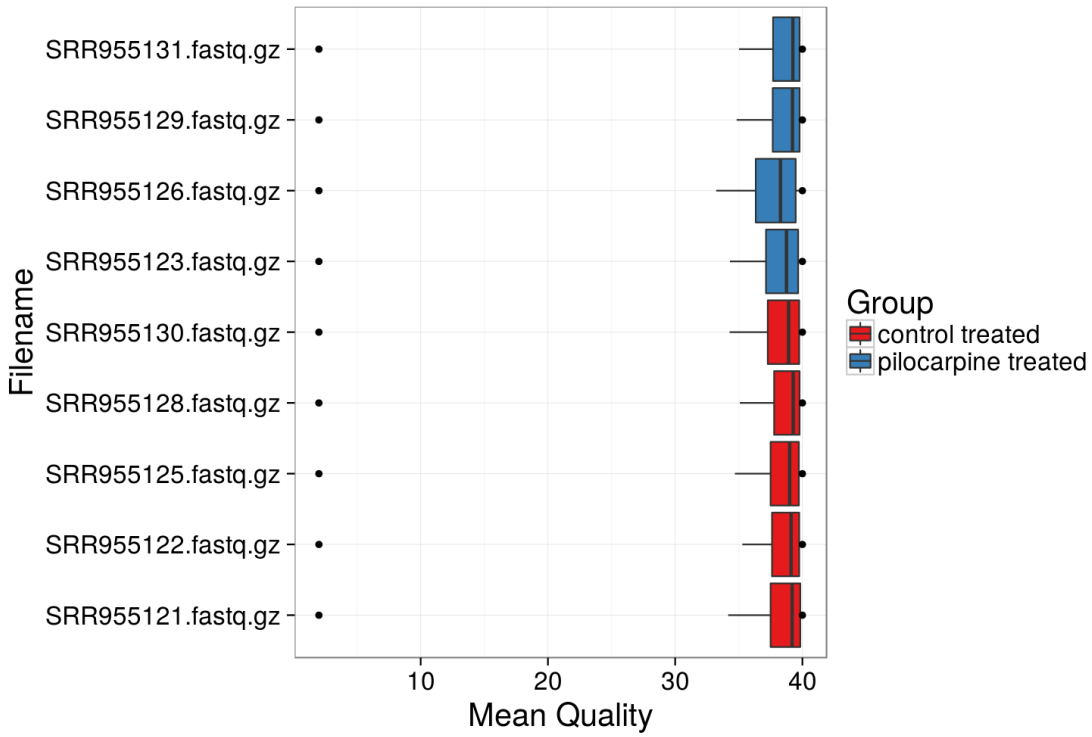


Figura 7 - Distribuição da qualidade média dos *reads* por amostra. As amostras estão divididas por grupos. Os pontos no gráfico representam os valores mínimo e máximo encontrados. Os dados utilizados possuem qualidade média satisfatória onde 90% dos reads das amostras possuem qualidade média acima de 30 na escala PHRED.

Foi utilizado o programa Bowtie versão 1 para alinhamento dos fragmentos de sequências contra o genoma de referência *Rattus norvegicus*,

versão 5. Os parâmetros do Bowtie utilizados para o alinhamento estão de acordo com (18,37).

Foram aplicadas três análises de controle de qualidade específicas para dados específicos de enriquecimento: enriquecimento de CpG, cobertura e análise de saturação. As amostras estão 4 vezes mais enriquecidas em relação ao genoma de referência. A cobertura dos *reads* mapeados no genoma de referência é de até 72 vezes embora a maior proporção está entre 1 e 8 vezes. Os gráficos da análise de saturação são apresentados na Figura 8.

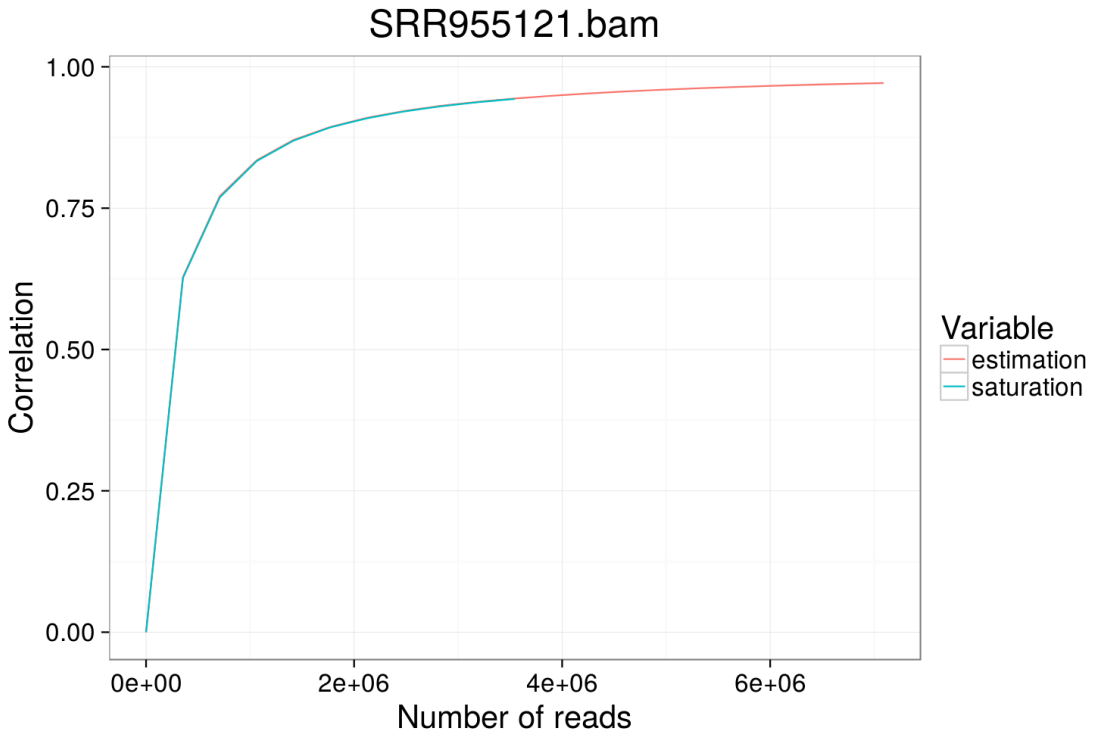


Figura 8 - Gráfico da análise de saturação. Todas as amostras possuem um número de *reads* suficiente para a análise de acordo com a estimativa.

Análise

A análise dos dados de enriquecimento de DNA metilado foi realizada com a utilização da ferramenta *methylCap*. Os arquivos de alinhamento são carregados no ambiente estatístico R pelo pacote. As amostras são definidas em dois conjuntos, que representam os grupos controle e administração por pilocarpina.

O tamanho da janela foi definido como 500, de acordo com (18). O genoma é dividido em regiões de 500 pares de bases, utilizadas para contar a

quantidade de fragmentos mapeados dentro de cada região. Também é definido a extensão virtual dos *reads* para 150 pares de bases, essa configuração está associada ao protocolo de fragmentação e ao equipamento sequenciador.

O resultado do processamento realizado pelo software methylCap é a tabela das janelas genômicas e seus respectivos valores do teste estatístico. Os dois principais valores são p-valor ajustado (FDR) e log razão na base 2 da diferença no nível de metilação entre grupos de amostras (*log2 fold change*). O p-valor ajustado é resultado da correção do p-valor para múltiplos testes (método BH). Quanto menor o valor do p-valor ajustado menor as chances de que as diferenças observadas serem ao acaso. A log razão é a ordem de magnitude do grupo tratado em relação ao grupo controle. Dessa forma valores negativos representam hipometilação e valores positivos representam hipermetilação.

A tabela é filtrada por um valor mínimo de corte (*cutoff*) para o p-valor ajustado. Foi definido o *cutoff* em 0,1 para considerarmos que uma região genômica pode estar potencialmente diferente no grupo tratado em relação ao controle. Esse valor serve de guia para definir novas hipóteses sobre as regiões selecionadas para serem validadas. A tabela filtrada é utilizada para gerar gráficos que apresentam diferentes perspectivas dos dados analisados. Esses gráficos são importantes para a interpretação dos resultados e são descritos abaixo.

O gráfico *volcano plot* mostra os p-valores ajustados contra a log razão das diferenças no nível de metilação entre dois grupos de amostras. O gráfico contribui para identificação visual rápida das janelas genômicas que apresentam grandes diferenças na metilação e que também são estatisticamente significativas. Cada ponto nesse gráfico representa uma janela genômica. O eixo y é o valor negativo do log na base 10 dos p-valores ajustados. Isso faz com que as janelas mais significativas apareçam no topo do gráfico. No eixo x as janelas são dispostas pela log razão na base 2 das diferenças nos níveis de metilação. O valor 0 no eixo x indica não há diferença nos níveis de metilação entre grupos de amostras. As janelas genômicas com

as maiores diferenças no nível de metilação são dispostas nos extremos do eixo x no gráfico. O gráfico *volcano plot* é apresentado na Figura 9.

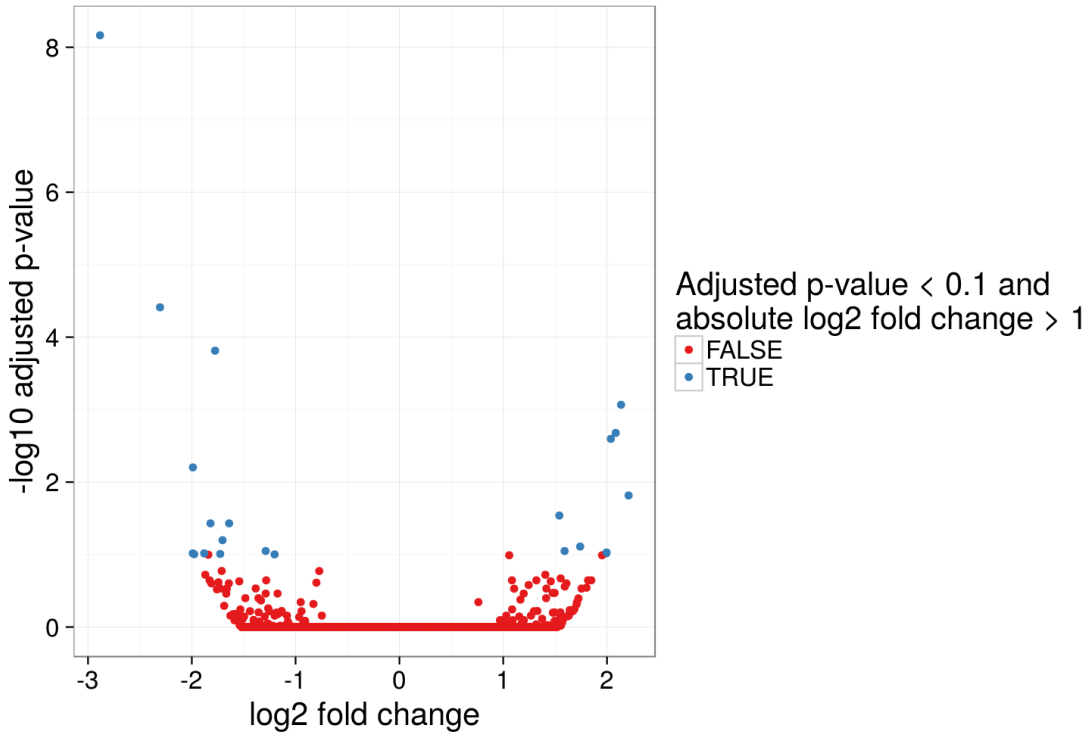


Figura 9 - Gráfico *volcano plo*t destaca as janelas genômicas com p-valores ajustados menores que o valor mínimo e que também possuem diferenças na metilação de pelo menos duas vezes maior ou menor. O eixo y é o valor negativo na base 10 dos p-valores ajustados e o eixo x é a log razão na base 2.

O gráfico *Manhattan Plot* oferece uma visão geral das janelas genômicas ao longo de todo o genoma. Nesse gráfico é possível observar que a grande maioria das janelas não possuem diferenças significativas entre os grupos de amostras. O gráfico *Manhattan Plot* é apresentado na Figura 10.

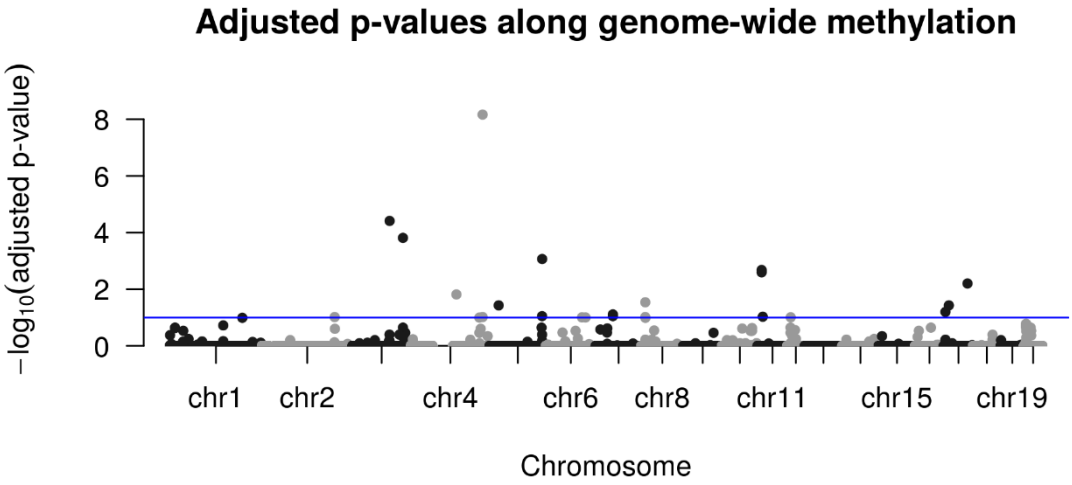


Figura 10 – Gráfico *Manhtattan plot* exibe p-valores ajustados ao longo do genoma. A linha horizontal azul representa o valor de *cutoff* definido em 0,1.

O gráfico *MA Plot* mostra a relação entre a log razão da contagem normalizada (eixo x) e a log razão da diferença na metilação dos grupos (eixo y). Cada ponto vermelho representa uma janela genômica com p-valor ajustado menor que o *cutoff*. A Figura 11 apresenta o gráfico *MA Plot*.

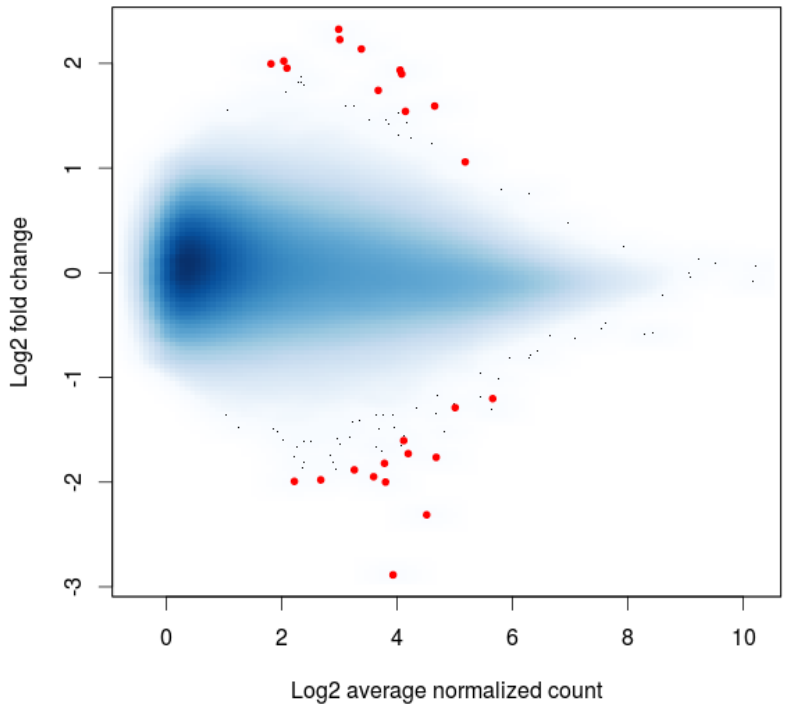


Figura 11 - Os pontos em vermelho representam as janelas diferencialmente significativas.

A visualização gráfica da análise de agrupamentos não supervisionada pode ser feita pelo gráfico composto de *heatmap* e dendrograma. O eixo y são as janelas genômicas significativas de acordo com o *cutoff* para o p-valor ajustado. O eixo X são as amostras onde a etiqueta do eixo descreve o nome da amostra e o grupo. A cor gradiente representa a intensidade relativa do nível de metilação. A intensidade relativa é calculada a partir dos valores normalizados da contagem de *reads*. A cor vermelha é baixa metilação, amarela é média, e branco são os valores mais altos de metilação. O dendrograma agrupa as janelas genômicas (eixo y) e as amostras (eixo x) pela distância Euclidiana. É esperado que as amostras do mesmo grupo se juntem no mais alto cluster. O gráfico de agrupamento é apresentado na Figura 12.

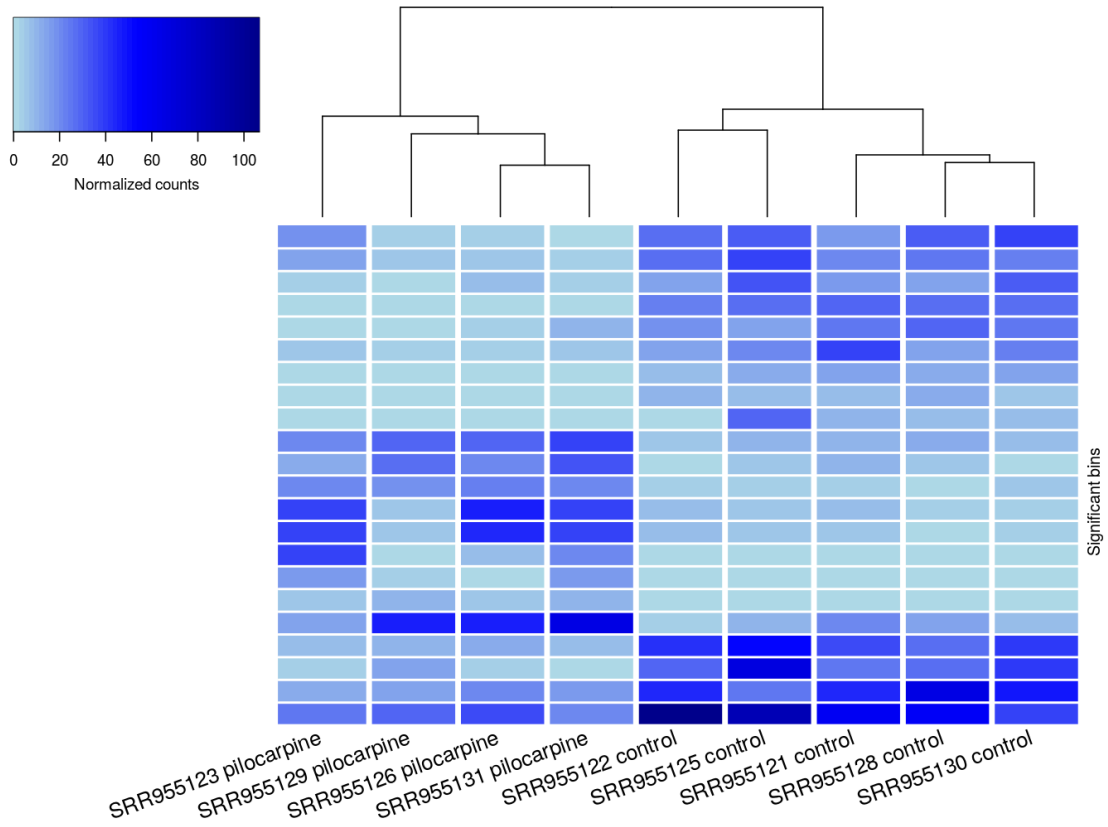


Figura 12 – Gráfico composto de heatmap e dendrograma.

A pontuação global da metilação é calculada pela média da log razão da diferença no nível de metilação do grupo de amostras tratadas contra o grupo de amostras controle. Os valores são divididos em hipermetilação (log-razão maior que 0) e hipometilação. As janelas genômicas são selecionados

por diferentes pontos de corte de p-valores atualizados, partindo dos menores valores de p-valor ajustado até a p-valor ajustado definido. Para cada seleção de janelas de genômicas é calculado a pontuação global de metilação. Isso significa que as pontuações de metilação globais são valores cumulativos. Estes valores são usados para criar um gráfico de sobrevivência (Figura 13).

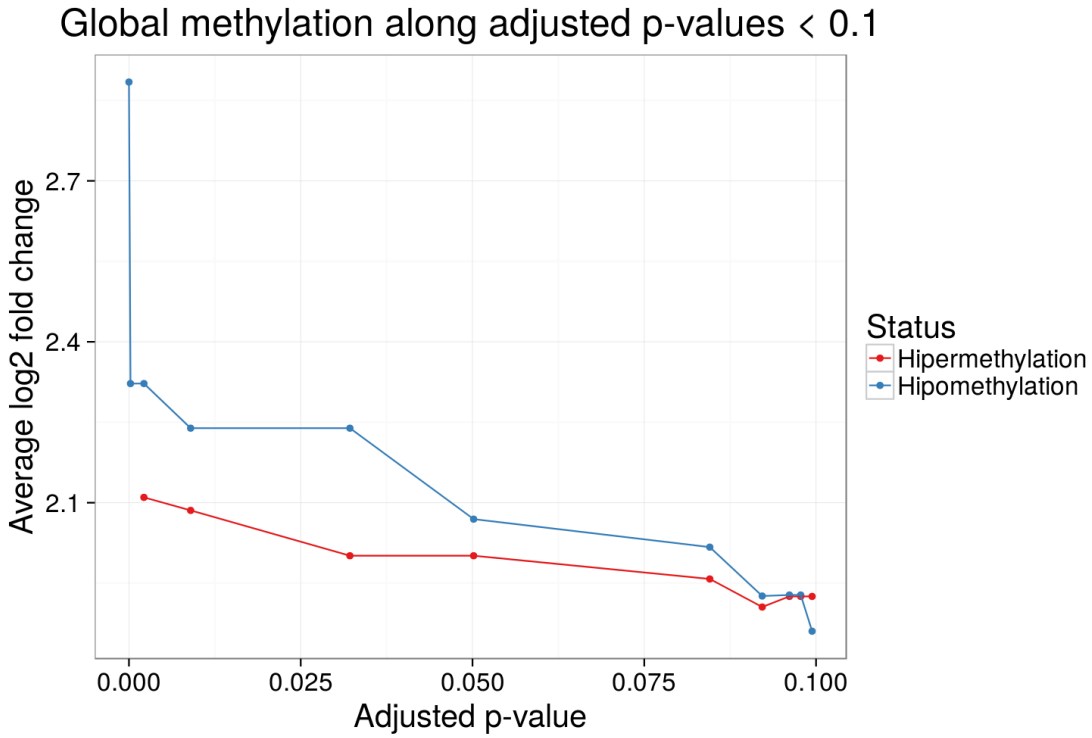


Figura 13 – Gráfico de sobrevivência: o valor médio da log-razão da diferença na metilação entre os grupos de amostra varia de acordo com o valor máximo definido para o p-valor ajustado.

Os resultados apontam que, dado o valor máximo do p-valor ajustado para selecionar as janelas genômica pode alterar os resultados da pontuação global de metilação. Os gráficos seguintes mostram o gráfico de análise de componentes principais (PCA) para todas as janelas genômicas (Figura 14) e o gráfico de PCA calculado com apenas as janelas genômicas selecionadas de acordo com o valor de *cutoff* definido (Figura 15).

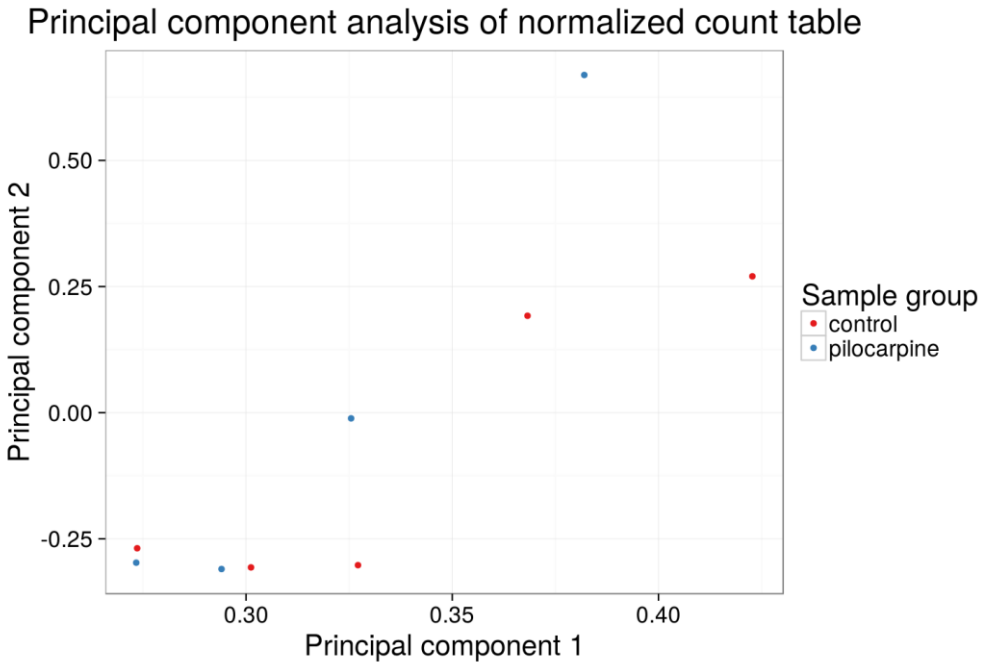


Figura 14 – PCA calculado a partir de todas as janelas genômicas. As amostras de grupos diferentes estão próximas nas duas dimensões com maior variabilidade.

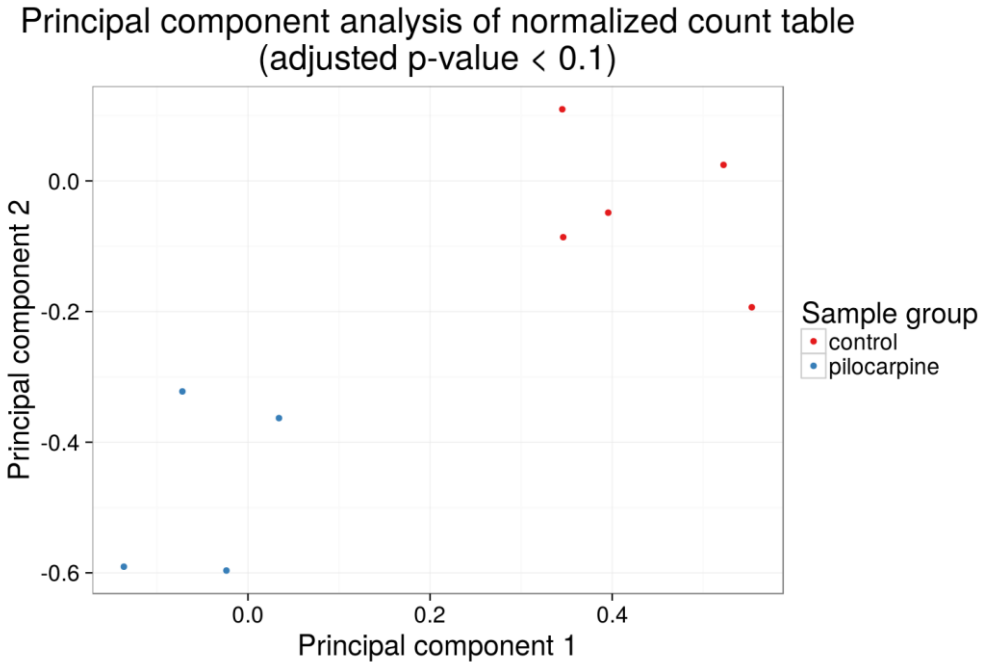


Figura 15 – PCA calculado com apenas as janelas selecionadas. As amostras de mesmo grupo estão próximas umas das outras e afastadas de amostras do outro grupo.

Análise dos dados de WGBS

O protocolo de análise para dados de WGBS é dividido nas etapas de processamento e análise. Essas etapas consistem no controle de qualidade dos fragmentos sequenciados, remoção de adaptadores de sequência, alinhamento contra o genoma de referência, cálculo da tabela de porcentagem de metilação do DNA e identificação de regiões diferencialmente metiladas. A Figura 16 apresenta todas as etapas do protocolo para dados de WGBS.

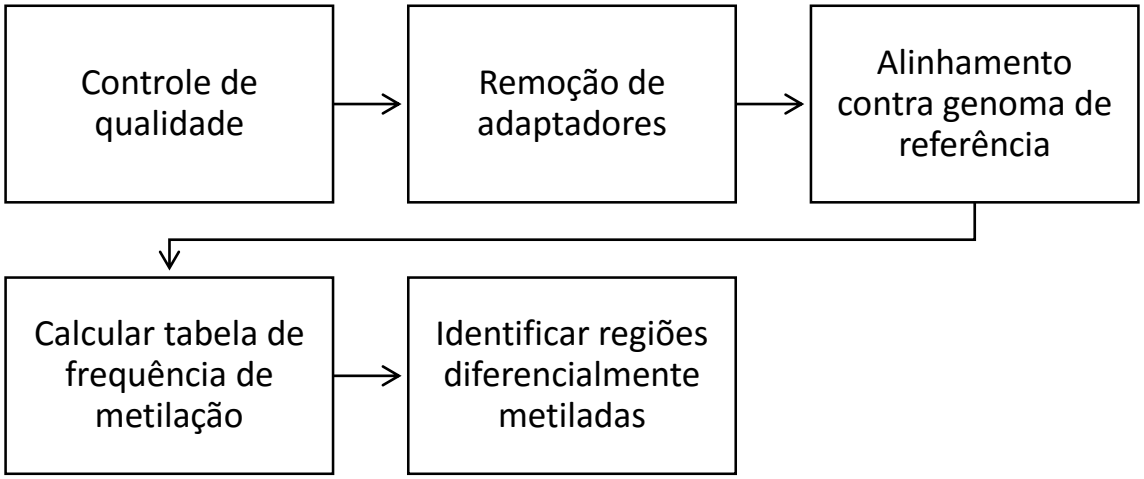


Figura 16 - Etapas do protocolo de análise para dados de WGBS.

Processamento

Foi utilizado um conjunto de dados inicial, desenvolvido em nosso laboratório, de WGBS. A qualidade dos dados foi avaliada pelas ferramentas Rqc e FastQC. Nessa etapa foi identificado viés na porcentagem de citosinas metiladas (reportadas como base C pelo sequenciador) no início e no fim dos ciclos de sequenciamento. É esperado menos de 5% de citosinas metiladas ao longo de todos os ciclos (20). A Figura 17 apresenta o viés identificado em todas as amostras, aqui é apresentado apenas uma amostra de animal controle.

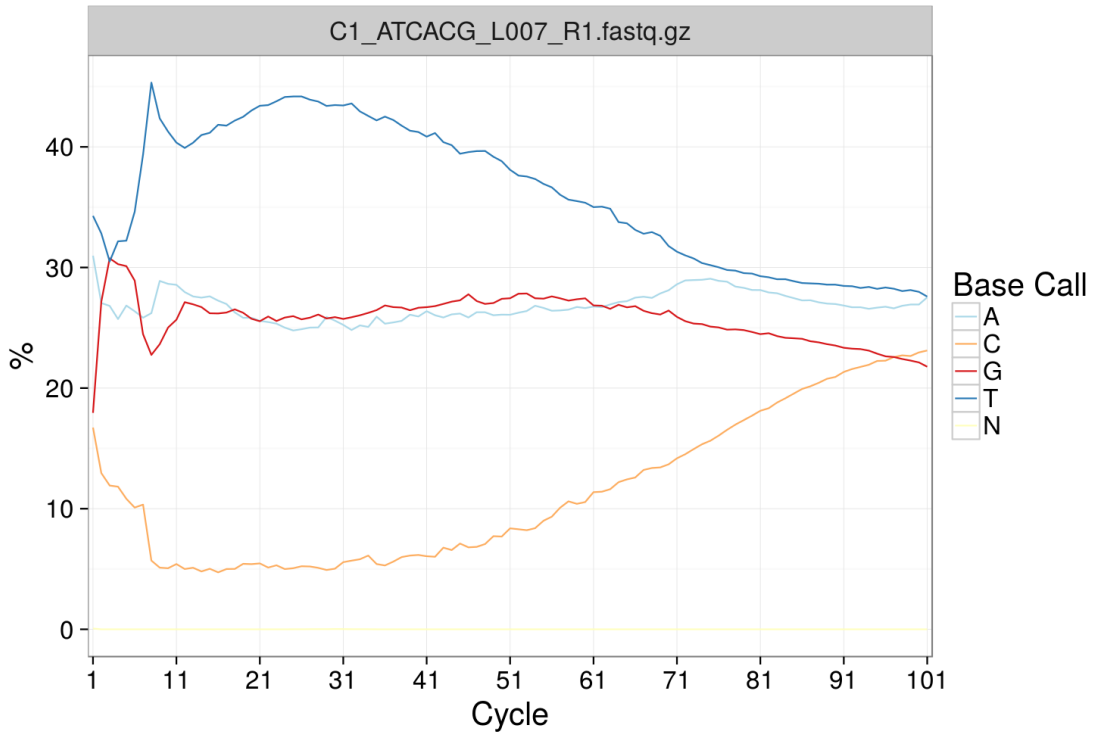


Figura 17 – Porcentagem de bases reportadas ao longo dos ciclos de sequenciamento. A amostra tem viés de proporção de bases C.

Para resolver o problema foi executado o programa Trim Galore!, para remover as sequencias de adaptadores, usados durante a etapa de preparação da biblioteca de sequenciamento, dos fragmentos de sequência. De um total de 238.917.318 fragmentos sequenciados, aproximadamente 98% foram mantidos após o filtro. A Figura 18 apresenta o gráfico de controle de qualidade após a edição e filtragem dos fragmentos.

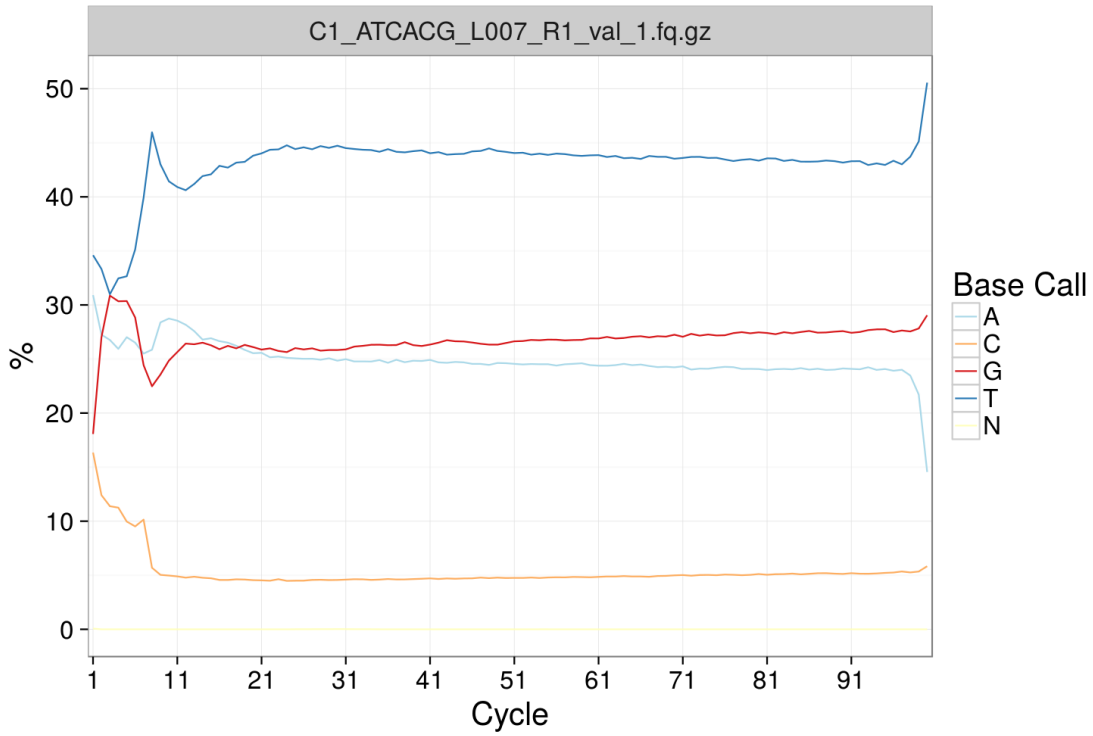


Figura 18 - Porcentagem de bases reportadas ao longo dos ciclos de sequenciamento. Após a filtragem dos fragmentos o viés de citosinas metiladas diminuiu, embora ainda exista picos no início dos fragmentos.

Os fragmentos filtrados foram alinhados contra o genoma de referência (*Rattus norvegicus*, versão 5) pelo programa Bismark. Após o alinhamento foi gerado o gráfico *M-bias Plot* para as amostras. Esse gráfico apresenta possíveis problemas na cobertura de dinucleotídeos CpG. A Figura 19 apresenta o gráfico *M-bias Plot* da amostra controle.

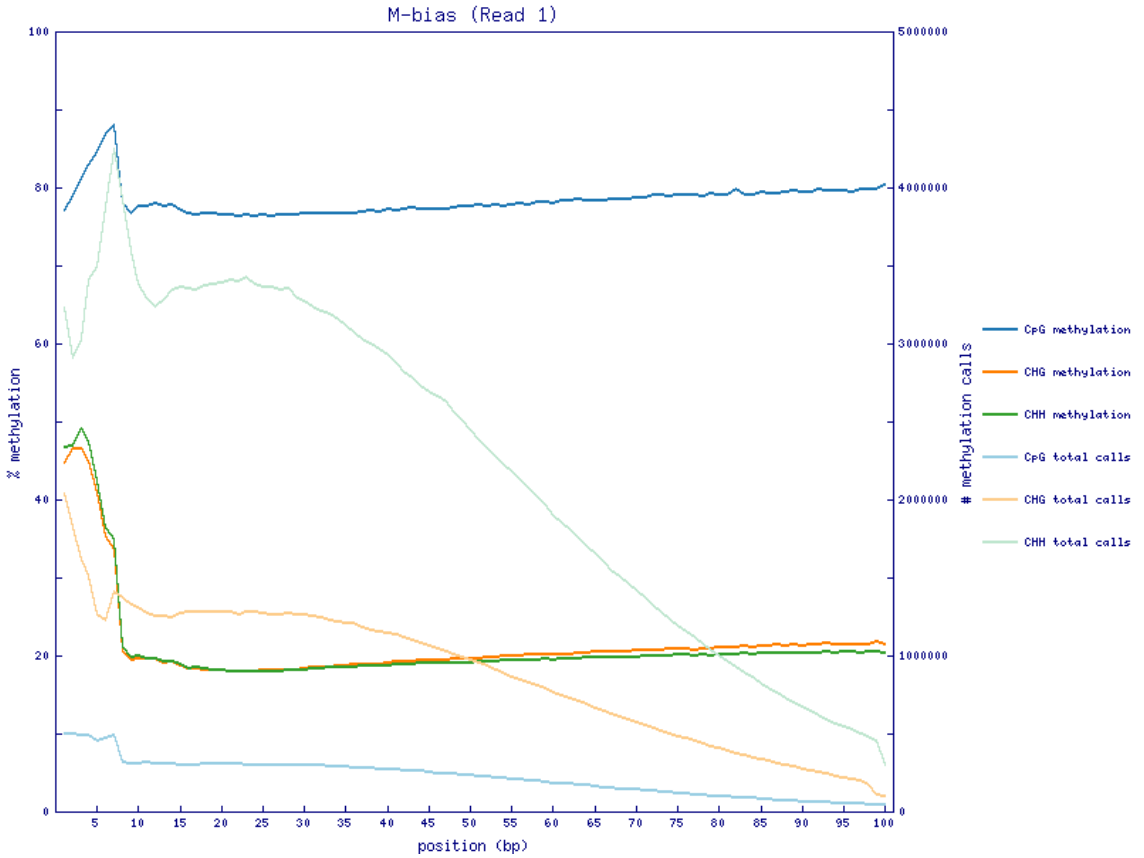


Figura 19 – Gráfico *M-bias Plot* aponta que o viés observado no controle de qualidade anterior foi mantido afetando a proporção de metilação nos dinucleotídeos CpG.

Análise

Foi utilizado o pacote estatístico *bsseq* para identificar DMRs nos dados de WGBS. A ferramenta encontrou 18 DMRs com os parâmetros padrão. A Figura 20 apresenta uma DMR encontrada.

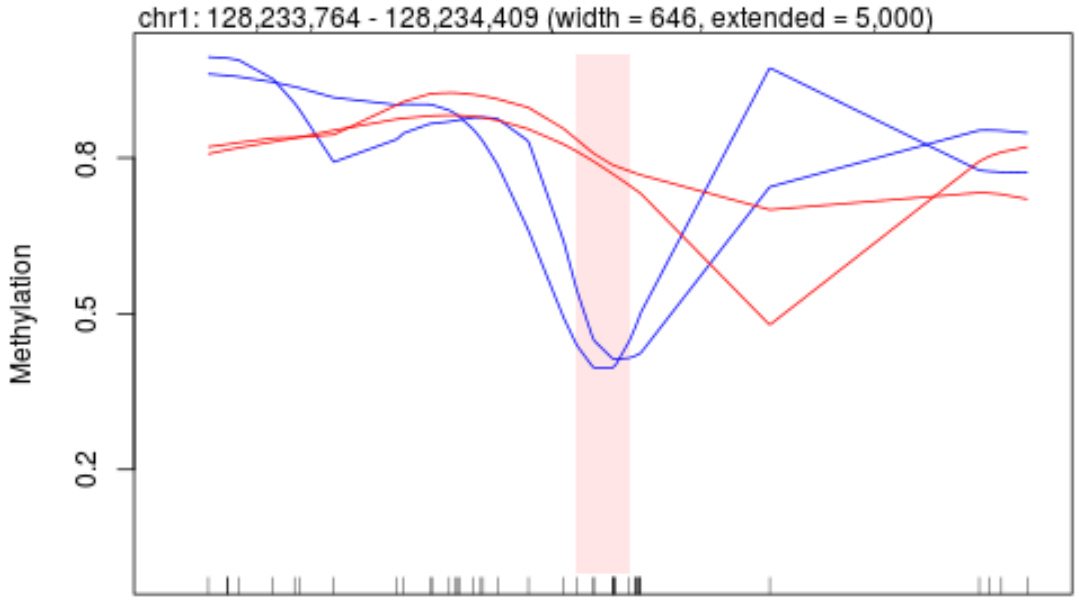


Figura 20 – Região diferencialmente metilada encontrada no cromossomo 1. Essa DMR tem 646 pares de bases. As linhas vermelhas representam amostras tratadas com pilocarpinas e as linhas azuis representam as amostras controle. Nessa região as amostras de pilocarpina estão hipermetiladas em relação as amostras controle.

MethylCap-seq é um método descrito recentemente para estudos epigenéticos utilizando sequenciamento de alto rendimento. Por essa razão, falta protocolos bem descritos para processamento e análise de dados de MethylCap-seq. Foi utilizado ferramentas disponíveis para dados de sequenciamento por enriquecimento e adaptamos para funcionar com dados de captura de DNA metilado. Foi observado que a identificação de regiões diferencialmente metiladas nos dados de MethylCap-seq é afetada negativamente pelo viés existente dos dados que está relacionado com a própria técnica de metilação de DNA.

Diferente dos métodos de metilação diferencial utilizados, como regiões conhecidas e predição de metilação em resolução de base, foi realizado comparações em janelas individuais a partir da contagem normalizada por tamanho da biblioteca. Dessa forma não foi necessário normalizar a contagem pela quantidade de dinucleotídeos CpG. A metilação global é então calculada pelos resultados estatísticos obtidos na etapa de

metilação diferencial. O protocolo definido foi implementado como um pacote de software chamado methylCap.

A análise de dados de WGBS mostrou-se melhor definida na literatura do que o a análise de dados de MethylCap-seq. Isso ocorre porque a técnica de tratamento por bissulfito seguido de sequenciamento de alto rendimento apresentam resultados melhores. Entretanto foi observado a necessidade de padronizar os formatos de arquivos de porcentagem de metilação porque diferentes programas esperam arquivos de formatos específicos que são gerados de forma diferentes. Isso pode induzir a erros na porcentagem de metilação na resolução de base e afetar os resultados finais.

Todos os protocolos foram desenvolvidos em um formato que promove a pesquisa reprodutível das análises realizadas. A ferramenta desenvolvida, methylCap, gera como resultado final um relatório reprodutível.

CONCLUSÕES

Foram avaliadas e selecionadas ferramentas disponíveis para processamento de dados de metilação do DNA. A partir da seleção dessas ferramentas, foram definidos dois protocolos em bioinformática para analisar dados de estudos de associação dos perfis de metilação de modelos animais de epilepsia. Um protocolo foi definido para dados de WGBS e outro protocolo para dados de MethylCap-seq.

O protocolo para dados de WGBS foi utilizado em um estudo sobre mapeamento de do perfil de metilação em modelos animais de epilepsia. O protocolo para dados de MethylCap-seq foi implementado como uma ferramenta chamada *methylCap*, permitindo a automatização do processamento. A ferramenta foi utilizada para analisar os dados públicos de modelos animais de epilepsia.

REFERÊNCIAS

1. Russo V, Martienssen R, Riggs A. Epigenetic mechanisms of gene regulation. Cold Spring Harbor monograph series. 1996. 692 p.
2. Bock C, Lengauer T. Computational epigenetics. Bioinformatics. 2008 Jan 1;24(1):1–10.
3. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002 Jan 1;16(1):6–21.
4. Ehrlich M, Gama-Sosa M a., Huang L-H, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. 1982;10(8):2709–21.
5. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. Nature Publishing Group; 2010 Mar 3;11(3):191.
6. Levenson JM, Sweatt JD. Epigenetic mechanisms in memory formation. Nat Rev Neurosci. 2005 Feb 14;6(2):108–18.
7. Fisher RS, Boas W van E, Blume W, Elger C, Genton P, Lee P, et al. Epileptic Seizures and Epilepsy: Definitions Proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). Epilepsia. 2005 Apr;46(4):470–2.
8. Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE Official Report: A practical clinical definition of epilepsy. Epilepsia. 2014 Apr;55(4):475–82.
9. Blümcke I, Thom M, Aronica E, Armstrong DD, Bartolomei F, Bernasconi A, et al. International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: A Task Force report from the ILAE Commission on Diagnostic Methods. Epilepsia. 2013 Jul;54(7):1315–29.
10. Norwood BA, Bumanglag A V., Osculati F, Sbarbati A, Marzola P, Nicolato E, et al. Classic hippocampal sclerosis and hippocampal-onset epilepsy produced by a single “cryptic” episode of focal hippocampal excitation in awake rats. J Comp Neurol. 2010 May 20;518(16):3381–407.
11. Navarro Mora G, Bramanti P, Osculati F, Chakir A, Nicolato E, Marzola P, et al. Does Pilocarpine-Induced Epilepsy in Adult Rats Require Status epilepticus? PLoS One. 2009 Jun 2;4(6):e5759.
12. Kobow K, Kaspi A, Harikrishnan KN, Kiese K, Ziemann M, Khurana I, et al. Deep sequencing reveals increased DNA methylation in chronic rat epilepsy. Acta Neuropathol. 2013;126:741–56.
13. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013 Sep 27;10(10):949–55.
14. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping

technologies. Nat Biotechnol. 2010;28(10):1106–14.

15. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2010 Oct;28(10):1097–105.

16. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol. 2008;26(7):779–85.

17. Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. Whole-genome DNA methylation profiling using MethylCap-seq. Methods. 2010;52(3):232–6.

18. Rodriguez BAT, Frankhouser D, Murphy M, Trimarchi M, Tam H, Curfman J, et al. Methods for high-throughput MethylCap-Seq data analysis. BMC Genomics. BioMed Central Ltd; 2012;13 Suppl 6(Suppl 6):S14.

19. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012 Sep 18;13(10):705–19.

20. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. Nat Methods. 2012 Jan 30;9(2):145–51.

21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

22. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;1–10.

23. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2.

24. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10:232.

25. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. BioMed Central Ltd; 2012;13(10):R83.

26. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, et al. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. Nature Publishing Group; 2011;43(8):768–75.

27. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. BioMed Central Ltd; 2012;13(10):R87.

28. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 Apr 1;38(6):1767–71.

29. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics. 2014 Jan 15;30(2):284–6.

30. Yan P, Frankhouser D, Murphy M, Tam H-H, Rodriguez B, Curfman J, et al. Genome-wide methylation profiling in decitabine-treated patients with acute myeloid leukemia. Blood. 2012 Sep 20;120(12):2466–74.

31. Frankhouser DE, Murphy M, Blachly JS, Park J, Zoller MW, Ganbat J-O, et al. PrEMeR-CG: inferring nucleotide level DNA methylation values from MethylCap-seq data. Bioinformatics. 2014 Dec 15;30(24):3567–74.

32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

34. Leek JT, Peng RD. Opinion: Reproducible research can still be wrong: Adopting a prevention approach: Fig. 1. Proc Natl Acad Sci. 2015 Feb 10;112(6):1645–6.

35. Peng RD. Reproducible Research in Computational Science. Science (80-). BioMed Central Ltd; 2011 Dec 2;334(6060):1226–7.

36. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004 Jan;5(10):R80.

37. Trimarchi MP, Murphy M, Frankhouser D, Rodriguez B a T, Curfman J, Marcucci G, et al. Enrichment-based DNA methylation analysis using next-generation sequencing: sample exclusion, estimating changes in global methylation, and the contribution of replicate lanes. BMC Genomics. 2012;13 Suppl 8(Suppl 8):S6.