



CODIGO DA PROVA: RP001/2016



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS
CONCURSO:

FOLHA DE RESPOSTA

Importante: O código da prova só será colocado na entrega da prova ao fiscal. As provas serão escaneadas e enviadas aos membros da banca avaliadora sem o nome do candidato.

7 - Bioinformática aplicada a área de epigenômica

A bioinformática associada a diferentes plataformas e abordagens de sequenciamento pode ser aplicada para compreender as alterações epigenéticas que podem ocorrer no genoma.

As alterações epigenéticas não alteram a sequência de nucleotídeos (bases nitrogenadas), mas provocam alterações químicas no DNA ou RNA, que podem alterar a expressão gênica. Como exemplo de alterações epigenéticas temos a metilação do DNA, assim como, a fosforilação, acetilação e metilação de histonas que podem provocar alterações na cromatina, e alterar a expressão gênica.

Para estudar as alterações epigenéticas, pode ser realizado o sequenciamento de nova-geração associado ao tratamento do DNA com bisulfito de sódio, sequenciamento baseado de BS-Seq. Por meio desse sequenciamento é possível identificar as citosinas que foram metiladas, uma vez que o bisulfito converte citosinas não metiladas em uracila.

Para realizar a análise de bioinformática de dados de BS-Seq, inicialmente os arquivos brutos do sequenciamento (reads) são analisados em relação a qualidade utilizando o programa FastQC, depois ~~as~~ as reads são ~~tr~~ filtradas

COM SOFTWARES COMO O BBDUK E TRIMMOMATIC, EM SEGUIDA ESSES PODEM SER ANALISADOS COM O SOFTWARE BISMARCK, QUE TEM COMO DEPENDÊNCIA OS SOFTWARES HISAT, BOWTIED, MINIMAP E SAMTOOLS.

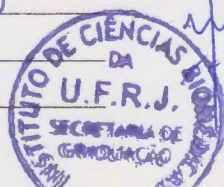
O BISMARCK É UM SOFTWARE QUE FOI DESENVOLVIDO COM O OBJETIVO DE ANALISAR DADOS DE BS-SEQ, E IDENTIFICAR A CITOSINAS METILADAS, EM RELAÇÃO AS CITOSINAS NÃO METILADAS. EM ASSOCIAÇÃO COM O BISMARCK PODE SER UTILIZADO O SOFTWARE, QUE É UMA BIBLIOTECA EM R, CHAMADO DE METHKIT. COM ESSE PACOTE EM R É POSSÍVEL GERAR AS VISUALIZAÇÕES DOS DADOS DE METILAÇÃO EM R, ALÉM DE REALIZAR ANÁLISE DE EXPRESSÃO DIFERENCIAL DE METILAÇÃO. ESSES DADOS PODEM SER INTEGRADOS COM DADOS DE RNA-SEQ PARA ENTENDER O PROCESSO DE METILAÇÃO NA EXPRESSÃO DOS GENES. NORMALMENTE A METILAÇÃO ESTÁ ASSOCIADA COM O SILENCIAMENTO GÊNICO.

OUTRO MÉTODO QUE PODE SER APLICADO PARA O ESTUDO DE ALTERAÇÕES EPIGENÉTICAS É O SEQUENCIAMENTO CHIP-SEQ, QUE É CONHECIDO COMO IMUNOPRECIPITAÇÃO DE CROMATINA. A PARTIR DESSE SEQUENCIAMENTO É POSSÍVEL ATRAVÉS DA UTILIZAÇÃO DE ANTICORPOS ESPECÍFICOS, ESTUDAR A ACETILAÇÃO DE HISTONAS E ALTERAÇÕES NA CROMATINA, A LIGAÇÃO NO DNA DE FATORES DE TRANSCRIÇÃO, ASSIM COMO, OUTRAS INTERAÇÕES DE PROTEÍNAS COM O DNA.

PARA ANALISAR DADOS DE CHIP-SEQ É POSSÍVEL UTILIZAR ~~OS~~ SOFTWARES COMO O MACS, QUE PODE IDENTIFICAR PICOS DE READS QUE MARCAM EM REGIÕES ESPECÍFICAS DO GENOMA, E A PARTIR DESSES PICOS É POSSÍVEL IDENTIFICAR O LOCAL NO GENOMA DE INTERAÇÃO DO DNA COM A PROTEÍNA.

OUTROS PROGRAMAS QUE PODEM SER UTILIZADOS É O PEAKS E HOMER PARA REALIZAR A ANOTAÇÃO DESSAS REGIÕES DE MARCAMENTO.

OUTRO SEQUENCIAMENTO DE NGS QUE TEM SIDO UTILIZADO PARA ENTENDER ALTERAÇÕES NA CROMATINA QUE PODEM ESTAR ASSOCIADAS A

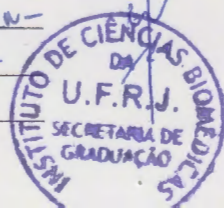


Alterações epigenéticas é o ATAC-SEQ, que é um sequenciamento de nova-geração, que inicialmente no preparo de biblioteca o DNA é exposto a presença de uma transposase (Tns) que irá clivar as regiões de cromatina aberta, em seguida o DNA clivado é preparado para o sequenciamento em plataformas de sequenciamento, como Illumina NovaSeq. A partir do sequenciamento de ATAC-SEQ é possível estudar a acessibilidade da cromatina, e associar esses dados com dados oriundos do BS-SEQ, CHIP-SEQ e RNA-SEQ para compreender as alterações epigenéticas e o impacto dessas na expressão gênica.

A análise do ~~BS~~ ATAC-SEQ pode ser realizada com o MACS2, que é um software que vai identificar os picos onde as reads foram mapeadas no genoma de referência. Dados de mapeamento e contagem das reads de ATAC-SEQ podem ser analisados no programa em R e RER.

Outra plataforma que está sendo utilizada para o estudo de metilação é o sequenciamento de ~~as~~ reads longas no MinION. Por meio desse sequenciamento, no qual o DNA ou RNA é sequenciado diretamente através da passagem em um poro de poro que está suscitado a uma corrente elétrica constante, é possível detectar alterações na corrente, que irão ser associadas a cada base (A, T, G e C) e a partir do sinal elétrico, também é possível identificar, por exemplo, citosinas metiladas. Programas como Guppy e Dorado podem ser utilizados para realizar o basecalling, e em associação com outras ferramentas computacionais, identificar bases metiladas.

Por fim, para a análise de dados de epigenômica, de forma geral, as reads do sequenciamento de nova geração precisam ser analisadas em relação a qualidade, utilizando



PROGRAMAS COMO PASTOR, TRIMMOMATIC, PARA REMOÇÃO DE ADAPTAÇÕES E BASES DE BAIXA QUALIDADE, E EM SEQUIDA ANÁLISADOS EM PROGRAMAS ESPECÍFICOS PARA CADA TIPO DE SEQUENCIAMENTO, COMO: MACS, PEAKS HONER PARA OS DADOS DE CHIP-SEQ, BISMARK E METHYKIT, PARA OS DADOS DE BS-SEQ E MACS2 E EDGE2 PARA OS DADOS DE ATAC-SEQ.

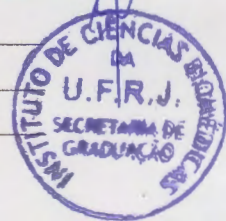
O ideal para uma análise mais abrangente é que os dados de diferentes abordagens sejam integrados para compreender as alterações epigenéticas e o impacto delas no genoma e na expressão gênica.

9 - APLICAÇÃO DA BIOINFORMÁTICA NO ESTUDO DAS PATOLOGIAS HUMANAS.

No estudo das patologias humanas a bioinformática pode ser utilizada para compreender como mutações em genes podem acarretar/causar doenças, não apenas mutações, como número de cópias de genes, alterações na expressão gênica, alterações epigenéticas, entre outros fatores.

Para identificar mutações que podem estar relacionadas com doenças, como o câncer por exemplo, pode-se sequenciar o genoma completo ou apenas realizar o sequenciamento do exoma. A partir dos dados de sequenciamento é possível realizar a montagem do genoma com softwares como o SPADIS, realizar a anotação do genoma e compará-lo com outros genomas, por exemplo, através de alinhamentos, para identificar as mutações.

A partir de dados do sequenciamento do genoma, também é possível realizar montagens por referência com a integração de softwares como MINIMAP2, SAMBOL2 e BCFtools. A partir do arquivo BAM é possível realizar uma chamada de variantes como com o



GATK ou BCFTOOLS e IDENTIFICAR AS REGIÕES ONDE HÁ VARIAGÕES COMO SUBSTITUIÇÕES OU INDEL.

A partir de estudos de GENÔMICA e COM A APLICAÇÃO DE FERRAMENTAS DE BIOINFORMÁTICA JÁ FOI POSSÍVEL IDENTIFICAR ALTERAÇÕES EM GENES QUE FOCAM ASSOCIADOS COM O DESENVOLVIMENTO DE ALGUNS TIPOS DE CÂNCER, COMO MAMA, TUMORES CEREBRAIS e ~~DOENÇAS~~ DOENÇAS NEURODEGENERATIVAS COM ALZHAIMER. POR EXEMPLO, NO CÂNCER DE MAMA FORAM IDENTIFICADAS MUTAÇÕES EM GENES COMO TP53, BRCA1, BRCA2, NO CASO DE TUMORES CEREBRAIS, EM GENES CODIFICADORES DE ISOCITATO DESSIDROGENASE (IDH), E EM RELAÇÃO AO ALZHAIMER, NOS GENES CODIFICADORES DE APOE.

ALTERAÇÕES EPIGENÔMICAS TAMBÉM JÁ FORAM ASSOCIADAS COM O DESENVOLVIMENTO E PROGRESSÃO DE ALGUMAS DOENÇAS. A PARTIR DE DADOS DE BS-SEQ, QUE É O SEQUENCIAMENTO DO DNA TRATADO COM BISULFETO PARA A IDENTIFICAÇÃO DE METILAÇÕES, TAMBÉM PODEM SER UTILIZADO PARA COMPREENDER AS PATOLOGIAS HUMANAS. OS DADOS DE BS-SEQ PODEM SER ANALISADOS COM SOFTWARES COMO O BISMARCK E METHYKIT (EM R).

EM UM ESTUDO DE GLIOMAS, O BS-SEQ FOI APLICADO PARA COMPREENDER OS GLIOMAS DE BAIXO GRAU E ALTO GRAU, E A PARTIR NESTES DADOS INTEGRADOS COM DADOS DE RNA-SEQ, FOI POSSÍVEL ESTABELECIER OS TIPOS DE GLIOMAS DE ACORDO COM OS PERFIS DE EXPRESSÃO E METILAÇÃO, ASSIM COMO IDENTIFICAR NOVOS ALVOS PARA A TERAPIA.

ANÁLISES DE EXPRESSÃO GÊNICA TAMBÉM PODEM SER APLICADAS PARA A ~~COM~~ COMPREENSÃO DE PATOLOGIAS HUMANAS, COM POR EXEMPLO, DOENÇAS NEURODEGENERATIVAS COM PARKINSON, ELA E ALZHAIMER. A PARTIR DO SEQUENCIAMENTO DO RNA E UTILIZANDO PROGRAMAS COMO BOWTIE2, HTSEQ E DESEQ É POSSÍVEL IDENTIFICAR GENES DIFERENCIALMENTE EXPRESOS NO CONTEXTO DA SAÚDE



E DA DOENÇA.

A partir de tabelas de contagens geradas pelo HTSeq é possível gerar redes de co-expressão utilizando a análise de correlação ponderada com a biblioteca WGCNA, que foi desenvolvida em R. Também é possível através de aprendizado profundo utilizando redes neurais criar redes de regulação gênica (GRNs) na qual pode-se estabelecer relações de causalidade, e identificar se um gene está regulando o outro positivamente ou negativamente.

Um grande exemplo da aplicação da bioinformática para o estudo de patologias humanas é o The Cancer Genome Atlas (TCGA) que integra dados de genômica, RNA-seq, BS-seq, ATAC-seq, entre outros dados para compreender os diferentes tipos de câncer através das análises de bioinformática e desenvolvimento de novos métodos computacionais.

Bancos de dados como o TCGA, Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), que armazenam dados brutos, assim como bancos de dados como o UniProt, Ensembl, Interpro, podem ser utilizados para identificar mutações em genes que podem levar ao desenvolvimento de doenças, identificar genes diferencialmente expressos, vias metabólicas, vias de sinalização celular, entre outros dados que podem ser relacionados às patologias humanas. A partir dos insights obtidos, podem ser desenvolvidas novas terapias, plataformas de diagnóstico e melhorar o prognóstico do paciente e desfecho clínico.

Também podem ser realizados estudos de modelagem de proteínas relacionadas às doenças para entender o impacto de mutações na estrutura tridimensional da proteína, utilizando por exemplo o AlphaFold, que é um programa que utiliza redes neurais profundos para



MODELAR A ESTRUTURA DE PROTEÍNAS. EM SEQUÊNCIA TAMBÉM É POSSÍVEL REALIZAR ANÁLISES DE DOCKING-MOLECULAR PARA ENTENDER A INTERAÇÃO DE PROTEÍNAS E LIGANTES E DESENVOLVER NOVAS TERAPIAS.

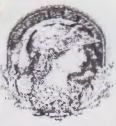
TAMBÉM É IMPORTANTE DESTACAR NO ENTENDIMENTO DAS PATOLOGIAS HUMANAS ESTUDOS DE PROTEÔMICA, COM POR EXEMPLO PROTEÔMICA BOTTOM-UP PARA ENTENDER O CONJUNTO DE PROTEÍNAS QUE SÃO GERADAS E AS MODIFICAÇÕES PÓS-TRADUÇIONAIS NO CONTEXTO DA SAÚDE E DA DOENÇA. ATUALMENTE AINDA HÁ UMA LIMITAÇÃO EM RELAÇÃO AOS PROGRAMAS DE BIOINFORMÁTICA ~~VAZ~~ UTILIZADOS PARA ANÁLISE DE DADOS DE ESPECTROMETRIA DE MASSAS EM LARGA ESCALA. MAS PROGRAMAS COMO BLASTP, DIAMOND E HMMER PODEM SER UTILIZADOS EM CONJUNTO COM BANCOS DE DADOS DO UNIPROT E NCBI, PARA ANALISAR ESSE TIPO DE DADO.

d - CONTROLE DA EXPRESSÃO GÊNICA COM ÊNFASE EM MÉTODOS COMPUTACIONAIS.

EM RELAÇÃO AO CONTROLE DA EXPRESSÃO GÊNICA BASEADA EM MÉTODOS COMPUTACIONAIS, ESSA É UMA ÁREA NA QUAL A PARTIR DE DADOS DE SEQUENCIAMENTO DE NOVA-GERAÇÃO (NGS), ASSOCIADOS A ENSAIOS DE PROTEÔMICA E ANÁLISES DE BIOINFORMÁTICA, É POSSÍVEL COMPREENDER TUDO O FLUXO DE INFORMAÇÃO DO DNA ATÉ A PROTEÍNA.

O CONTROLE DA EXPRESSÃO GÊNICA OCORRE EM DIFERENTES ETAPAS, E ENVOLVE DIFERENTES PROCESSOS CELULARES, COMO POR EXEMPLO, O CONTROLE TRANSCRIPCIONAL, NO QUAL ESTÃO ENVOLVIDOS OS FATORES DE TRANSCRIÇÃO (TF) QUE SE LIGAM EM REGIÕES REGULADORIAS, COMO AS REGIÕES PROMOTORAS, TAMBÉM HÁ PARTICIPAÇÃO NESSE PROCESSO DE ENHANCERS, LIGAÇÃO DE CO-FATORES, ENTRE OUTROS PROCESSOS. NO CONTEXTO DA EXPRESSÃO GÊNICA, TAMBÉM OCORREM ETAPAS APÓS A TRANSCRIÇÃO, COMO O SPLICING





Alternativo, que pode gerar diferentes formas de uma proteína a partir de um único gene + nesse processo também participa os pequenos RNAs não codificantes que existe possuem entre 20 e 30 nucleotídeos e os RNAs longos não codificantes.

Além de todos os processos mencionados acima, o controle da expressão gênica pode ocorrer por meio de alterações epigenéticas que ocorrem no DNA, como a metilação em regiões de CpG, modificações de cromatina e alteração alteração de histonas, como metilação, acetilação e fosforilação.

Para identificar os fatores de transcrição em um organismo pode ser realizado o sequenciamento de genoma completo, realizar a montagem do genoma (SPADES), anotação (MAKER, SNAP, GENEMARK, AUGUSTUS), identificar os TFs por meio do BLAST, HMMER utilizando bancos de dados do próprio programa ou bancos de dados externos com Uniprot e NCBI.

A partir do genoma montado e anotado podemos utilizar banco de dados como Uniprot, PPRM, NCBI, ENSEMBL, e programas como o HMMER para identificar os fatores de transcrição, assim como, a partir de ferramentas computacionais identificar regiões cis-regulatórias, onde esses TFs podem se ligar para realizar o controle da expressão.

A partir de dados de RNA-se é possível entender os genes diferencialmente expressos, assim como identificar isoformas e splicing alternativo.

A partir das bibliotecas de RNA-se pode-se utilizar programas como TOPHAT, HISAT e DESEQ para realizar a análise de expressão diferencial.



Com a tabela de contagem do HTSEQ é possível utilizar programas como o DESEQ2 que aplica um modelo de regressão binomial negativo para normalizar as contagens com base no tamanho das bibliotecas e outros fatores de confusão. Além disso o DESEQ2 aplica o modelo de BENJAMINI-HOCHBERG PARA AJUSTAR OS VALORES P E DIMINUIR AS TAXAS DE DETECÇÃO FALSA (FDR), REQUERENDO ASSIM A CHANCE DE ENCONTRAR PAÍOS POSITIVOS EM MÚLTIPLAS ~~COMPARAÇÕES~~ COMPARAÇÕES.

A partir da tabela de contagem é possível gerar redes de co-expressão com o pacote em R WGCNA e redes de regulação gênica (GRNs) utilizando o aprendizado de máquina e redes neurais profundas.

Com o DESEQ2 é possível identificar genes diferencialmente expressos e gerar visualizações como NA plot, PCA, volcano plot, heatmaps, boxplots, entre outros gráficos de visualização. A partir desses dados e gráficos é possível compreender os genes diferencialmente expressos, que estão sendo co-expressos (análise WGCNA) e por meio das ~~em~~ redes de regulação gênica (GRNs), entender como um gene pode estar regulando o outro.

Para a análise de pequenos RNAs (20-30 nt) é possível a partir do sequenciamento de pequenos RNAs, identificá-los utilizando o software ~~mirBASE~~ mirBASE, que realiza a análise por homologia, para identificar novos RNAs e pequenos RNAs pode ser utilizado o miRDEEP2, e o MIRANDA ou ~~tags~~ target scan para identificar os RNAs mensageiros que os microRNAs e siRNAs podem estar se ligando. Para a análise de RNAs longos não-codificantes, podem ser utilizados os softwares PLEK, que diferencia



RNAs longos não codificantes de mRNA
é o programa CPC, que identifica o
o potencial codificante de RNAs longos não
~~no~~ codificantes. Esses dados podem ser uti-
lizados para construção de redes de co-
expressão e GRNs, juntamente com outros
dados de RNA-seq, para verificar a atuação
desse elemento no controle da expressão
gênica.

Para análise de isoformas pode ser
utilizado o software RALISTO, que realiza
pseudo-alinhamento, em conjunto com o
programa 3 RNA-seq. Para estudo
de splicing alternativo pode ser utilizado
o STAR e Cuffdiff.

Além de tudo que foi mencionado,
o controle da expressão gênica pode ser
estudado através de diferentes abordagens
de sequenciamento como BS-seq, CHIP-seq
e ATAC-seq que ~~se~~ não fornece dados
sobre metilação, interação DNA-proteína
e ~~de~~ acessibilidade de cromatina.

Esses dados podem ser analisados com
programas como Bismark, MethyKit, MACS,
PEAKS, HOMER e COBIC. Em conjunto,
esses dados fornecem grandes insights
sobre o controle da expressão gênica.