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 À 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 BISSULFITO DE SÓDIO, SEQUENCIAMENTO NOMEADO DE BS-SEQ. POR MEIO DESSE SEQUENCIAMENTO É POSSÍVEL IDENTIFICAR AS CITOSINAS QUE FORAM METILADAS, UMA VEZ QUE O BISSULFITO 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 ANALISADAS EM RELAÇÃO A QUALIDADE UTILIZANDO O PROGRAMA FASTQC, DEPOIS ~~AS~~ A AS READS SÃO TRIMADAS

Comissão Organizadora do Concurso
Gabinete da Direção
UFRJ

com softwares como o ~~BSSUK~~ e trimmomatic, em seguida esses podem ser analisados com o software Bismark, que tem como dependência os softwares ~~Hista~~ Hisat, Bowtie, Miniapd e samtools.

O Bismark 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 Bismark pode ser utilizado o software, que é uma biblioteca em R, chamado de Methykit. 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 ~~ano~~ proteínas com o DNA.

Para analisar dados de ChIP-seq é possível utilizar ~~outro~~ softwares como o MACS, que irá identificar picos de reads que mapearam 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 dessa ~~outros~~ regiões de mapeamento.

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 ILUMINA NOVASEQ. A PARTIR DO SEQUENCIAMENTO DE ATAC-SEQ É POSSÍVEL ESTUDAR A ACESSIBILIDADE DA CROMATINA, E ASSOCIAR ESSES DADOS COM DADOS OBTIDOS DO BS-SEQ, CHIP-SEQ E RNA-SEQ PARA COMPREENDER AS ALTERAÇÕES EPIGENÔMICAS 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 EDGER.

OUTRA PLATAFORMA QUE ESTÁ SENDO UTILIZADA PARA O ESTUDO DE METILAÇÃO É O SEQUENCIAMENTO DE ~~EA~~ READS LONGAS NO NANION. POR MEIO DESSE SEQUENCIAMENTO, NO QUAL O DNA OU RNA É SEQUENCIADO DIRETAMENTE ATRAVÉS DA PASSAGEM EM UM PROTEÍNA DE POLO QUE ESTÁ SUBTIDA 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 fastqc, limadas com softwares como bbduk e trimmomatic, para remoção de adaptadores e bases de baixa qualidade, e em seguida analisados em programas específicos para cada tipo de sequenciamento, como: macs, peaks hunter para os dados de chip-seq, bismack e methykit, para os dados de bs-seq e macs2 e edger 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 da 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~~ genoma com softwares como o spades, realizar a anotação do genoma e comparar 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, samtools 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 VARIAÇÕES COMO SUBSTITUIÇÕES OU INDELS.

A PARTIR DE ESTUDOS DE GENÔMICA E COM A APLICAÇÃO DE FERRAMENTAS DE BIOINFORMÁTICA JÁ FOI POSSÍVEL IDENTIFICAR ALTERAÇÕES EM GENES QUE FORAM 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 ISOCITRATO DESIDROGENASE (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 BISSULFITO 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 BISMARK 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 DESSES DADOS INTEGRADOS COM DADOS DE RNA-SEQ, FOI POSSÍVEL ESTRATIFICAR OS TIPOS DE GLIOMAS DE ACORDO COM OS PERFIS DE EXPRESSÃO E METILAÇÃO, ASSIM COM IDENTIFICAR NOVOS ALVOS PARA A TERAPIA.

ANÁLISES DE EXPRESSÃO GÊNICA TAMBÉM PODEM SER APLICADAS PARA A ~~DOE~~ COMPREENSÃO DE PATOLOGIAS HUMANAS, COM POR EXEMPLO, DOENÇAS NEURODEGENERATIVAS COM PARKISON, ELA E ALZHAIMER. A PARTIR DO SEQUENCIAMENTO DO RNA E UTILIZANDO PROGRAMAS COMO Bowtie2, HTSEQ E DESEQ2 É POSSÍVEL IDENTIFICAR GENES DIFERENCIALMENTE EXPRESSOS 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 CASUALIDADE, 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 ARCHIEVE (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 AS PATOLOGIAS HUMANAS. A PARTIR DOS INSIGHTS OBTIDOS, PODEM SER DESENVOLVIDAS NOVA TERAPIAS, PLATAFORMAS ~~DIA~~ DE DIAGNÓSTICO E MELHORAR O PROGNÓSTICO DO PACIENTE E DESFECHO CLÍNICO.

TAMBÉM PODEM SER REALIZADOS ESTUDOS DE MODELAGEM DE PROTEÍNAS RELACIONADAS AS DOENÇAS PARA ENTENDER O IMPACTO DE MUTAÇÕES NA ESTRUTURA TRIDIMENSIONAL DA PROTEÍNA, ~~utilizar~~ UTILIZANDO POR EXEMPLO O ALPHAFOLD, QUE É UM PROGRAMA QUE UTILIZA REDES NEURAIS PROFUNDAS PARA

MODELAR A ESTRUTURA DE PROTEÍNAS. EM SEGUIDA TAMBÉM É POSSÍVEL REALIZAR ANÁLISES DE DOCKING MOLECULAR PARA ~~INT~~ 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 JÁ FORAM GERADAS E AS MODIFICAÇÕES PÓS-TRADUCIONAIS NO CONTEXTO DA SAÚDE E DA DOENÇA. ATUALMENTE AINDA HÁ UMA LIMITAÇÃO EM RELAÇÃO AOS PROGRAMAS DE BIOINFORMÁTICA ~~UTI~~ UTILIZADOS PARA ANÁLISE DE DADOS DE ESPECTROMETRIA DE MASSAS EM LARGA ESCALA. MAS ~~PROGR~~ 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 TODO 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 TRANSCRICIONAL, NO QUAL ESTÃO ENVOLVIDOS OS FATORES DE TRANSCRIÇÃO (TFs) QUE SE LIGAM EM REGIÕES REGULATÓRIAS, 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 ~~POSSUI~~ 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), ~~ANO~~ 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, PFAM, 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 DADO DE RNA-SE É POSSÍVEL ENTENDER OS GENES DIFERENCIALMENTE EXPRESSOS, ASSIM ~~COMO~~ COMO IDENTIFICAR ESSA FORMA E SPLICING ALTERNATIVO.

A PARTIR DAS BIBLIOTECAS DE RNA-SE PODE-SE UTILIZAR PROGRAMAS COM TOPHAT, HTSEQ E DESEQ2 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 descobertas falsas (FDR), revolvendo assim a chance de encontrar dados positivos em múltiplas 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 aprendizado de máquina e redes neurais profundas.

Com o DESEQ2 é possível identificar genes diferencialmente expressos e gerar visualizações como MA plot, PCA, volcano plot, heatmap, 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 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, que realiza a análise por homologia, para identificar novos pequenos RNAs pode ser utilizado o miRDeep2, e o miRanda ou TargetScan 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 mRNAs é o programa CPC, que identifica o o potencial codificante de RNAs longos não ~~tno~~ codificantes. Esses dados podem ser utilizados para construção de redes de co-expressão e GRNs, juntamente com outros dados de RNA-seq, para verificar a atuação desses elementos no controle da expressão gênica.

Para análise de isoformas pode ser utilizado o software Kallisto, que realiza pseudo-alinhamento, em conjunto com o programa 3DRNA-seq. Para estudo de splicing-alternativo pode ser utilizado o ~~star~~ STAR e ExpSplicing.

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 irão fornecer dados sobre metilação, interação DNA-proteína e ~~te~~ ac acessibilidade de cromatina. Esses dados podem ser analisados com programas como Bismark, methykit, MACS, peaks, Homer e edger. Em conjunto, esses dado forneceram grandes insights sobre o controle da expressão gênica.