



CODIGO DA PROVA:

RP 001 / 008



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

FOLHA DE RESPOSTA

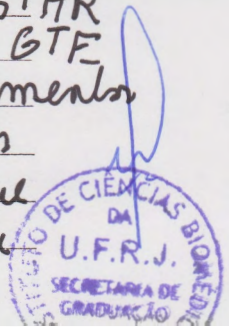
Importante: O código da prova só será colocado na entrega da prova ao fiscal. As provas serão escaneadas e enviadas aos membros da banca avaliadora sem o nome do candidato.

2. Controle da expressão gênica com ênfase em métodos computacionais

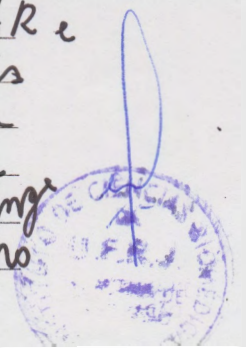
O controle da expressão gênica envolve diversos mecanismos complexos tal como metilações, modificações da cromatina que reflete na estrutura 3D, afinidade com fatores transcricionais que podem ser ativadores ou repressores e a velocidade da maquinaria de transcrição com a presença dos tipos de codon. Tal como raros, frequentes, ótimos e sub-ótimos. Para cada um desses processos existem técnicas experimentais de alto rendimento associadas, e por consequência, métodos estatísticos e computacionais aplicados para maximizar a detecção dos verdadeiros positivos e negativos enquanto minimizam os falsos positivos e negativos. Para entender o controle da expressão gênica é necessário fazer o experimento de RNA-seq e calcular a expressão diferencial dos genes (DEGs). Portanto inicia-se com o planejamento do experimento e a determinação dos contrastes, isto é, quais são as condições experimentais que os experimentos serão submetidos, o número de réplicas e os tratamentos escolhidos bem como os fatores a serem estudados e se necessária possíveis interações. Essa

etapa é necessária para ir orientar os demais experimentos de alto rendimento que serão integrados aos dados do RNA-seq. Nessa etapa de planejamento desta-se a linguagem R com os seus diversos pacotes que auxiliam no planejamento e no número de replicas para experimentos de alto rendimento, no repositório do R que é o Bioconductor constantemente é lançados esses tipos de pacote.

Para entender melhor esse controle da transcrição os experimentos tais como BS-seq, ChIP-seq e Hi-C podem ser integrados com os dados do RNA-seq para criar um perfil de como está ocorrendo a regulação da transcrição. Todos esses experimentos, incluindo o RNA-seq, inicia-se com o sequenciamento dos fragmentos (reads) de RNA ou DNA. A primeira etapa da análise de dados consiste em filtrar os fragmentos de baixa qualidade, normalmente abaixo de Q30 ou com bases não informativas (N). Em experimentos como o BS-seq que detecta sítios de metilação (ilhas CpG), ChIP-seq que detecta regiões no DNA onde se liga fatores transcricionais específicos e o Hi-C que detecta conformações da cromatina é necessário uma etapa de retirada de ruídos e enriquecimento de picos. Para a filtragem utiliza-se o FASTQC para os relatórios da qualidade do sequenciamento bem como o rendimento por biblioteca. Logo após aplica-se os programas que removem os fragmentos de baixa qualidade tal como o Trimmomatic, ou BBMAP/BBDUCT ou o Cutadapt. No RNA-seq após essa etapa, se o organismo for eucariota, utiliza-se programas que mapeiam no genoma levando em conta as funções de splicing tal como o programa STAR ou Hisat2 que usa o arquivo GTF3 ou GTF para determinar as coordenadas que os fragmentos foram mapeados. Nos outros experimentos numa etapa anterior ao mapeamento que é a exclusão de resíduos (noise), na verdade



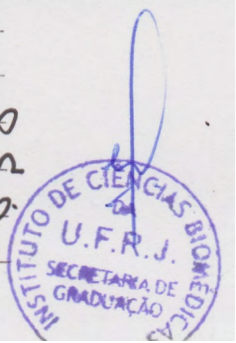
ruídos. No BS-seq destaca-se o programa de
 acesso livre e código aberto o Bismarck, nesse
 pipeline que recebe como entrada arquivos
 FASTQ, remove os ruídos e detecta os picos
 que representam os sítios de metilação CpG,
 juntamente com as coordenadas no genoma,
 isto é, a posição que esses sítios assumem em
 relação ao gene mais próximo e em regiões
 promotoras. Na etapa posterior dessa técnica,
 usando os arquivos provenientes do Bismarck
 usa-se a chamada de regiões diferencialmente
 metilada usando o programa do sistema R
 o DMR-seq, este em formato de pacote, detecta
 regiões hipermetiladas ou hipometiladas no
 genoma em relação a posições dos genes. No
 Chip-seq também tem a etapa de retirada
 de ruídos usando o programa livre e de código
 aberto MACS 2 que também faz o enriquecimento
 de picos que são as regiões onde os fatores
 transcricionais, retornando essas coordenadas
 em relação aos genes mais próximos. Para as
 estatísticas em relação aos contrastes estudados
 também utiliza o sistema R como o exemplo
 do pacote chipseq que calcula fold-change, p-
 valor ajustados dos picos entre as condições /
 tratamentos. Na abordagem do Hi-C que estuda
 a estrutura 3D da cromatina a filtragem dos
 ruídos tem como a detecção das estruturas 3D
 associadas, nível de compactação, sítios de
 enzimas de restrição destaca-se o programa
 Hi-C framework and pipeline presente no
 github, sendo gratuito e código aberto. Diante
 desses resultados pode-se integrar com o RNA-seq
 com os genes diferencialmente expressos que
 são detectados após o mapeamento pelo STAR e
 usando um contador como o FeatureCounts
 ou HTSeq aplicar no programa que calcula
 a estatística de significância de cada gene
 para cada contraste usando o log2 Fold Change
 e p-valor ajustado dentre outras medidas como



estimação da abundância gênica e uso-padrão do \log_2 Fold Change, destaca-se os pacotes do R DESeq2, edgeR, limma-voom dentre outros. A etapa posterior que é a análise funcional usando o gene ontology (GO), KEGG pathways (vias de sinalização e metabólicas) e Reactome permite entender quais fenômenos biológicos esses genes podem estar atuando. A integração desses diferentes resultados podem ser feitos usando redes (grafos) onde os nós representam os genes, regiões metiladas, sítios de fatores transcricionais ou estruturas 3D de cromatina (manifold) e as arestas a localização desses elementos nas regiões promotoras ou mais próximas dos genes. Nessas redes com diferentes tipos de nós e arestas pode-se usar o Cytoscape ou Gephi que são programas gratuitos e altamente flexíveis para integrar esse diversos elementos na rede permitindo por meio da anotação funcional e das interações das redes uma inferência mais aprofundada e completa sobre como é o controle da expressão gênica dada as condições ambientais ou fisiológicas.

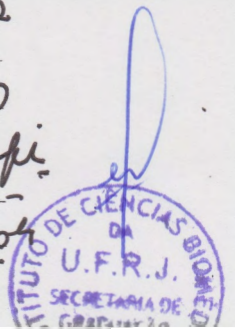
- 9. Bioinformática aplicada a análise epigenômica

O silenciamento gênico tem como a modulação da expressão gênica decorrentes de estímulos ambientais e fisiológicos são muito importantes para a formação de órgãos nos organismos mais complexos. A metilação é uma das principais forma de silenciamento gênico e o metiloma apresenta como o perfil global de metilação do DNA dada uma condição fisiológica. Para estudar o metiloma, surgiu experimentos de alto-rendimento que sequencia todos as regiões metiladas do DNA. Essa técnica busca encontrar ilhas ou sítios CpG metilados em amostras com diferentes replicas biológicas em mais de um tratamento. Essa técnica que é o BS-seq (bisulfito sequenciamento) baseia-se no sequenciamento dessas regiões com diferentes perfis de metilação. As técnicas de bioinformática é aplicada nos arquivos do sequenciamento representado por fragmentos no formato FASTQ. Esses fragmentos apesar da alta qualidade que os sequenciadores geram, tal como o Illumina Nextseq 2000 ou 2000, ainda contém fragmentos (reads) de baixa qualidade ou bases não informativas por isso é necessário aplicar programas que geram relatórios sobre o sequenciamento tal como o FASTQC que é gratuito e livremente disponível. Nesse relatório possuem adaptadores também são detectados e podem ser removidos na próxima etapa de filtragem. Programas como o BBduk, Cutadapt e Trimmomatic auxiliam nessa remoção de reads de baixa qualidade ($< Q30$), N e adaptadores. Na próxima etapa pode-se usar o programa Bismark que é um pipeline onde os ruídos são detectados e removidos, os sítios de metilação CpG são encontrados por meio do enriquecimento de reads que suportam essas regiões em relação ao "background" do fundo. A saída do Bismark que são as coordenadas

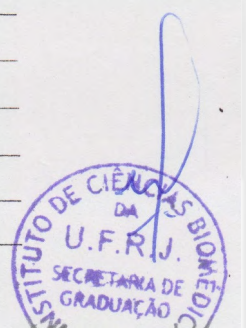


no genoma dos sítios CpG, o suporte de reads por cada região e sítio e a significância estatística por sítios CpG detectados. A próxima etapa envolve a detecção de regiões diferencialmente metiladas (DMR) que pode-se usar o pacote do sistema R DMR-seq. Nesse pacote é necessário o arquivo do desenho experimental que consta as amostras relacionadas as réplicas e aos tratamentos que serão comparados. juntamente com os arquivos morfométricos do Biomark é calculada a estatística onde o número de sítios metilados são comparados entre os tratamentos. Regiões hipermetiladas ou hipometiladas são detectados de acordo com o contraste estabelecido e a significância estatística pode ser calculada e o "cutoff" pode ser estabelecido. Esses DMRs são aplicados em programas de anotação funcional tal como o MONK que é livre permitindo a associação entre os DMRs e os genes próximos. A localização dos DMRs é importante pois podem estar em regiões promotoras, internas ao gene (entre o +1 da transcrição e a região de terminação) ou mais distante do gene, geralmente no meio de regiões intergenéricas. Para a anotação funcional pode-se usar o Gene Ontology (GO) que classifica os genes associados aos DMRs em processos biológicos, funções moleculares e componente celular, também o KEGG pathways nos genes associados a vias metabólicas, Reactome que é focado em vias de sinalização e metabólicas. Além de análises de redes usando bancos de dados de interações proteína-proteína (PPI) como o BIOGRID e STRING, sendo esse último também constando interações genéticas como por exemplo por fatores transcricionais. Para analisar e visualizar essas redes podemos usar o Cytoscape ou Gephi.

Outra técnica aplicada a epigenômica é o Hi-C que é sequenciamento de regiões específicas onde a conformação da cromatina é estudada e comparada nos diferentes condições.



ou tratamentos determinando regiões de hetero e eucromatina, bem como estruturas 3D de grupos e fractais e se alguma composição pode está enriquecida em certas condições ou tratamentos. Para analisar esses dados usa o 3D-pipeline e Hi-C fromellort que filtra os ruídos e fazem o enriquecimento, detecção e classificação das regiões 3D da cromatina. Também detecta-se sítios de enzimas de restrição que são usados para calcular a estrutura 3D associada. Tal como os outros experimentos de alto-rendimento usa-se uma estatística ^{para} calcular a significância da presença das estruturas 3D em diferentes tratamentos e a quais regiões gênicas estes estão associados. Esses dados podem ser associados a dados de RNA-seq e BS-seq e chip-seq para permitir associar a estrutura 3D à expressão gênica e o impacto que a estrutura da cromatina pode ter em determinadas condições biológicas



~~1. Impacto da bioinformática~~

9. Aplicações da bioinformática no estudo das patologias humanas

Uma das principais técnicas para estudar patologias humanas associada a genética é o sequenciamento de exomas. Nesse sequenciamento as regiões transcritas do DNA visando a detecção de variações únicas em nucleotídeos (SNV) que podem estar associadas a certas patologias. Essa técnica envolve sequenciamento de diversos indivíduos em uma população e essas informações armazenadas em grandes bancos de dados que podem ser consultados tal como o ~~SNPdb~~, dbSNP e PolyPhen entre diversos outros. Nesses bancos busca associações os SNVs aos fenótipos determinando a frequência desses na população. O enfoque nessas frequências é para os "alelos" raros, isto é, o de baixa frequência na população visto que em bancos genéticos raros é possível estudar e melhor associar. Para esse tipo de análise pode-se usar o SNIPEFF e SNIPSIFF que são pipelines que calcula e detecta essas variações. Nessa abordagem quanto maior número de indivíduos sequenciados melhor a inferência de associações entre uma variante e a doença.

Técnicas como RNA-seq, Proteômica e Metabolômica e a suas integrações em um sistema permite investigar como o metabolismo de uma patologia está funcionando em comparação com uma condição saudável. Os níveis da expressão genica, o de abundância de proteínas e de metabólitos permitem entender como a patologia evolui e se desenvolve. Nessa abordagem análise de redes e a abordagem da biologia integrativa permite em uma rede complexa usando o Cytoscape ou Gephi analisar e comparar diversas redes em condições de saúde e as suas modificações nas doenças. Ferramentas em R ou Python com a metanálises e bibliotecas network detectam possíveis hubs e conexões





presentes no tecido doente e não no saudável e também os contrários. A biologia de sistema, bem como a transcriptômica single-cell permite entender o perfil por célula em tecidos doentes e saudáveis, bem como as técnicas de estatística com normalizações, redução de dimensionalidade e agrupamentos entende o perfil por tipo de célula.

Também vale destacar técnicas de aprendizado de máquina, aprendizados profundos e modelos generativos nas análises de perfis moleculares em experimentos de alto rendimento. Métodos não-supervisionados como os citados em single-cell e os supervisionados nas chamadas de variantes do exoma. Bibliotecas em Python como o Scikit-learn e o PYTORCH está sendo amplamente usada nos métodos de bioinformática.

