CODIGO DA PROVA: RP001/008

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 genica com ênfase em métodos computacionais

O controle da expressão gênica envolve diversos mecanismos complexos tal como metilação, modificação 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 codons. Tal como raros, frequentes, ótimos e sub-ótimos. Para cada um desses processos existem técnicas experimentais de alto rendimento associadas, e por consequencia, metodos 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 genica é necessario 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 necessário possiveis interações. Essa

Comissão Organizadora do Concurso
Gabinete da Direção

etapa é necessária pois irá orientar os demais experimentos de alto rendimento que será integrado aos dados do RNA-Seq. Nessa etapa de planejamento destaca-se a linguagem R com os seus diversos pacotes que auxiliam no planejamento e no número de réplicas para experimentos de alto rendimento, no repositório do R que é o Bioconductor constantemente é lançado 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 esta 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 analise 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 conformação 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/BBDuck ou o Cutadapt. No RNA-seq após essa etapa, se o organismo for eucariota, utiliza-se programas que mapeam no genoma levando em conta as junções de splicing tal como o programa STAR ou Hisat2 que usa o arquivo GFF3 ou GTF para determinar as coordenadas que os fragmentos foram mapeados. Nos outros experimentos uma 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 representa 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 Bismark 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 posição 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/ Tratamento. Na abordagem do Hi-C que estuda a estrutura 3D da cromatina a filtragem dos ruídos bem 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 erro-padrão do log2 fold change, destaca-se os pacotes do R DESeq 2, edgeR, limma-voom dentre outros. A etapa posterior que é a analise funcional usando o gene ontology (GO), KEGG pathways (vias de sinalização e metabólicas) e Reactome permite entender quais fenomenos biologicos esses genes podem estar atuando. A integração desses diferentes resultados podem ser feitos usando redes (grafos) onde os nós representam os genes, regiões metiladas, sitios de fatores transcricionais ou estruturas 3D da cromatina (manifold) e as arestas a localização desses elementos nas regiões promotoras ou mais proximas dos genes. Nessas redes com diferentes tipos de nós e arestas pode-se usar o Cytoscape ou Gephi que são programas gratuitos e altamente flexiveis para integrar esse diversos elementos na rede permitindo por meio da anotação funcional e das interações das redes uma inferencia mais aprofundada e completa sobre como é o controle da expressão gênica dada as condições ambientais ou fisiológicas.

- 7. Bioinformatica aplicada a analise epigenomica

O silenciamento gênico bem como a modulação da expressão gênica decorrentes de estímulos ambientais e fisiologicos são muito importantes para a formação de orgãos nos organismos mais complexos. A metilação é uma das principais prima de silenciamento gênicos e o metiloma apresenta como o perfil global de metilação do DNA dada uma condição fisiologica. Para estudar o metiloma surgiu experimentos de alto-rendimento que sequencia todas as regiões metiladas do DNA. Essa tecnica busca encontrar ilhas ou sitios CpG metilados em amostras com diferentes replicas biologicas em mais de um tratamento. Essa tecnica que é o BS-seq (bissulfito sequenciamento) baseia-se no sequenciamento dessas regiões com diferentes perfil de metilação. As técnicas de bioinformatica é aplicada nos arquivos do sequenciamento representado por pragmentos no formato FASTA. Esses pragmentos apesar da alta qualidade que os sequenciadores geram, tal como o Illumina Nextseq 1000 ou 2000, ainda contém pragmentos (reads) de baixa qualidade ou bases não informativas por isso é necessário aplicar programas que gerem relatórios sobre o sequenciamento tal como o FASTQC que é gratuito e livremente disponivel. Nesse relatórios possiveis adaptadores também são detectados e podem ser removidos na próxima etapa de filtragem. Programas como o BBduck, 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 ruidos são detectados e removidos, os sitios de metilação CpG são encontrados por meio do enriquecimento de reads que suportam essas regiões em relação ao "background" do fundo. A saida do Bismark que são as coordenadas

no genoma dos sitios CpG, o suporte de reads por cada região e sitio e a significancia esta. Tertica por sitios CpG detectados. A próxima. etapa envolve a detecção de regiões diferencialmente metiladas (DMR) que pode-se usar o pacote do sistema R DMR-Reg. Nesse pacote é necessario o arquivo do desenho experimental que consta as amostras relacionadas as réplicas e aos. Tratamentos que serão comparados. juntamente com os arquivos provenientes do Bismark é calculada a estatistica onde o número de sitios metilados são comparados entre os tratamentos. Regiões hipermetiladas ou hipometiladas são detectadas levando com o contraste estabelecido e a "significancia estatistica 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 proximos. A localização dos DMRs é importante pois podem estar em regiões promotoras, internos ao gene (entre o +1 da transcrição e a região de terminação) ou mais distante do gene, geralmente nos meio de regiões intergenicas. Para a anotação funcional pode-se usar o gene Ontology (GO) que classifica os genes associados as DMRs em processos biologicos, funções moleculares e componente celular, tambem o KEGG pathways nos genes associados a vias metabolicas, Reactome que é focado em vias de sinalização e metabolicos. Além de analises de redes usando banco de dados de interação proteina-proteina (PPI) como o BIOGRID e STRING, sendo esse ultimo tambem contendo interações genéticas como por exemplo por fatores transcionais. Para analisar e visualizar essas redes podemos usar o Cytoscape ou Gephi.

Outra tecnica aplicada a epigenomica é o Hi-C que é sequenciamento de regiões especifica onde a conformação da cromatina é estudada e comparada nos diferentes condiçõ

ou tratamentos determinando regiões de hetero e eucromatina, bem como estruturas 3D de grampos e fractais e se alguma composição pode está enriquecida em certos condições ou Tratamento. Para análises esses dados usa o 3D-pipeline e Hi-C framework que filtra os ruídos e fazem o enriquecimento, detecção e classificação dos região 3D da cromatina. Também detecta-se sítios de enzimas de restrições que são usados para calcular a estrutura 3D associada. Tal como os outros experimentos de alto-rendimento usa-se uma estatística para calcular a significancia da presença das estrutura 3D em diferentes Tratamentos e a quais regiões gênicos 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 biologicas

~~1- Impacto da bioinformática~~

9 - Aplicações da bioinformática nos estudos das patologias humanas

Uma das principais técnicas para estudos patologias humanas associada a genética é o sequenciamento de exomas. Nesse sequenciam os regiões transcritos do DNA visando a detecção de variações únicas em nucleotídeos (SNV) que podem estar associados a certas patologia. 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 ~~GENBANK~~, db SNP e Polyphen entre diversos outros. Nesses bancos busca associar 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 doenças genéticas raras é possível estudos e melhor associar. Para esse tipo de análise pode-se usar o SNIPEFF e SNIPSIFF que são pipelines que calcula e detecta essas variantes. Nessa abordagem quanto maior número de indivíduos sequenciados melhor a inferência de associação entre uma variante e a doença.

Técnicas como RNA-seq, Proteoma e metabolômica e a suas integrações em um sistema permite inferir como o metabolismo de uma patologia esta funcionando em comparação com uma condição saudável. Os níveis da expressão genica, o de abundancia de proteínas e de metabolitos permitem entender como a patologia evolui e se desenvolve. Nessa abordagem analise de redes e a abordagem da biologia integrativa permite em uma rede complexa usando o Cytoscape ou Gephi analisar e compor diversas redes em condição de saúde e as suas modificações nas doenças. Ferramentas em R ou Python com a netanálysis e bibliotecas network detectam possíveis hubs e conexões

presentes no tecido doente e não no saudável e também os contrários. A tirdogia de sistema bem como a transcriptomica Single-Cell permite entender o perfil por celula em tecidos doentes e saudaveis bem como as tecnicas de estatistica com normalizaçao, reduçao de dimenssionalidade e agrupamento entende o perfil por tipo de celula.

Tambem vale destacar tecnicas de aprendizados de maquina, aprendizado profundo e modelos generativos nas analises de perfis moleculares em experimento de alto-rendimento. Metodos não-supervisionados como os citados em Single-Cell e os supervisionados nas chamadas de variantes do exoma. Bibliotecas em Pythom como o Scikit-learn e o PYTORCH esta sendo amplamente usada nos metodos de bioinformatica.