CODIGO DA PROVA: RP001-0021

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.

9 - Aplicação da Bioinformática no estudo de patologias humanas

Os avanços em metodologias de sequenciamento que obtivemos nas últimas décadas têm aprofundado nosso conhecimento sobre diferentes patologias humanas. Para isso, foram necessários avanços também na bioinformática, uma vez que temos hoje um aumento significativo no número de dados disponíveis. Esse conjunto de produção e análise de informações biológicas tem nos permitido desenvolver técnicas de terapia e tratamentos cada vez mais robustos e personalizáveis.

Nas décadas de 80 e 90, o estudo de doenças genéticas era feito majoritariamente em doenças raras e monogênicas. Para isso, utilizavam-se técnicas de ligação genética e sequenciamento de baixo rendimento, como Sanger. A doença de Huntington, por exemplo, precisou do envolvimento de mais de 100 cientistas e mais de 10 anos para que se achasse o gene e a mutação responsável pela doença. Hoje, através das tecnologias de Next Generation Sequencing (NGS), podemos, em alguns passos, sequenciar um painel de genes, todos os genes (whole Exome Sequencing - WES) ou até mesmo todo o genoma humano (whole genome Sequencing - WGS). Utilizando dados de WES somos capazes de identificar milhares de variantes gênicas de uma só vez. Essa técnica é muito utilizada utilizada no estudo de diferentes tipos de câncer, uma vez que é possí-

Comissão Organizadora do Concurso
Gabinete da Direção

vel comparar amostras tumorais com amostras normais de um mesmo paciente. Através de algoritmos computacionais pode-se então obter quais mutações somáticas foram geradas pela doença. Além disso é possível criar um "assinatura gênica" para o tipo específico de câncer. Por outro lado, a técnica de WES só permite a visualização de variantes em 2 a 5% do genoma. Para uma visão mais geral, utiliza-se a técnica WGS. Com ela, é possível detectar também variações estruturais, como inserções, deleções, translocações e inversões, além de informações sobre a variação no número de cópias para cada gene. Além de regiões gênicas, é possível também obter informações sobre variantes presentes em regiões de microssatélites, que são regiões com altas taxas de mutações. Quando essas regiões não são regoladas é possível detectar, através de WGS, a chamada Microsatelite Instability (MSI). A MSI é um sinal de que a maquinaria de reparo chamada Mismatch Repair (MMR) não está ativa e, por conseqüência, no processo de replicação são acumuladas mutações. A MSI está associada a diferentes tipos de cânceres e também a síndrome de Lynch, que é uma predisposição genética ao câncer colorretal.

Além de variações a nível de DNA é importante também detectar variação nos níveis de produção de RNAs. Para isso, utiliza-se a técnica de RNA sequencing (RNA-Seq). Através dela, é possível obter informações sobre o nível de expressão de diferentes genes, o que é essencial já que em diferentes tipos de doenças ocorre a desregulação dos níveis transcricionais. No RNA-Seq, leituras são geradas a partir de fragmentos de transcritos. Essas leituras são mapeadas ao genoma de referência e são posteriormente quantificadas e normalizadas. Através de algoritmos e ferramentas computacionais como DESeq2 e EdgeR é possível detectar variações positivas ou negativas nos níveis

de expressão gênica entre duas amostras. Isso é muito utilizado, por exemplo, em estudos que buscam identificar quais genes são regulados diferencialmente em tecidos tumorais em relação a tecidos normais. Uma vez detectados genes diferencialmente expressos, pode-se obter quais funções celulares estão sendo mais afetadas. Para isso o grupo de genes identificados é submetido a uma análise funcional, onde se aplicam métodos estatísticos em conjunto com bases de dados como Gene Ontology e MSigDB, as quais possuem informações sobre a associação entre cada gene e suas funções. O RNA-seq ainda é capaz de detectar e quantificar a presença de long non-coding RNAs (lncRNAs) o quais já foram implicados em diferentes doenças como infarto do miocárdio, diabetes e câncer de mama.

Outra técnica similar ao RNA-seq, porém com foco em estudar pequenos RNAs não codificantes é chamada de small RNA Sequencing (small RNA-Seq). Existem 3 tipos de pequenos RNAs não codificantes, são eles os microRNAs (miRNAs), os piwi-interacting RNAs (piRNAs) e os small interfering RNAs (siRNAs). A principal análise que diferencia essas diferentes categorias é a frequência de tamanho de leituras presentes nos dados. Isso porque miRNAs possuem entre 21 e 22 nucleotídeos, bem como siRNAs, e piRNAs possuem entre 24 e 30 nucleotídeos. Além disso são utilizadas outras análises envolvendo composição nucleotídica e local de mapeamento no genoma. Os miRNAs se associam com proteínas da família AGO e são capazes de regular a expressão de diferentes genes. Eles estão envolvidos em diferentes tipos de doenças, tais como doenças neurológicas (Alzheimer, Parkinson, Huntington), autoimunes (lúpus) e diferentes tipos de câncer. Já os piRNAs se associam a proteínas da família PIWI e são classicamente associados a regulação da expressão de elementos transponíveis. Porém encontram-se também ligados a diferentes patogias com artrite reuma-

tóide, leucemia e linfoma de células B. Os siRNAs, por outro lado, não estão comumente associados a causa de doenças, mas sim ao tratamento. (A[illegible]) Os siRNAs também se associam com proteínas da família AGO e regulam a expressão gênica, porém com mecanismos diferentes dos miRNAs. Apesar de desafiadora, a terapia para diferentes doenças utilizando siRNAs e nanocarreadores tem se desenvolvido e pode talvez ser útil para o tratamento de diferentes patologias.

Diversas outras técnicas de sequenciamento e análise existem e são capazes de visualizar os fatores envolvidos nas patologias humanas por diferentes ângulos. A bioinformática tem papel central no progresso e desenvolvimento de soluções que busquem melhores diagnósticos, terapias e tratamentos. (f[illegible]o)

7 - Bioinformática aplicada a análise epigenômica

Desde a definição cunhada por Waddington, em 1940, o termo "Epigenética" (f[illegible]) teve seu significado reescrito diversas vezes ao longo da história. Uma das definições mais atuais é a de Cavalli e Heard (2019) que diz que epigenética é o "estudo dos mecanismos e moléculas envolvidos na propagação de estados alternativos de atividade gênica sem que haja alteração na sequência de DNA!! São 3 os mecanismos epigenéticos mais estudados: alteração na estrutura da cromatina, modificações de histonas e metilação do DNA. Os avanços obtidos em técnicas de Next Generation Sequencing (NGS) nos permitem hoje gerar e analisar, através de métodos de bioinformática, (a[illegible]) dados referentes aos diferentes mecanismos epigenéticos

A cromatina é o conjunto de DNA mais proteínas, sendo as mais presentes chamadas de histonas. Dentro do núcleo, a cromatina pode se encontrar de uma for-

mais linearizada (eucromatina) ou empacotada (heterocromatina). Para que ocorra a expressão de um gene, é importante que sua região promotora esteja acessível para a ligação de fatores de transcrição e da RNA polimerase. A formação de heterocromatina nessas regiões é, portanto, um mecanismo epigenético capaz de inativar a expressão gênica. Para identificar regiões acessíveis presentes no genoma, uma das técnicas de sequenciamento mais modernas é chamada de Assay for Tagmentation of Accessible Chromatin followed by Sequencing (ATAC-seq). Nessa técnica se utiliza uma enzima transposase chamada de Tn5, a qual possui ligada a ela adaptadores para sequenciamento NGS. Uma vez adicionada na amostra de DNA a Tn5 é capaz de se ligar a regiões acessíveis do genoma e cliválas. As sequências divadas se ligam aos adaptadores e podem então ser amplificadas para o posterior sequenciamento. O sequenciamento gera milhões de leituras que são armazenadas em arquivos digitais. O processamento dessas leituras através de métodos computacionais é capaz de gerar uma mapa genômico preciso de todas as regiões de acessibilidade da cromatina.

As histonas são proteínas ligados ao DNA e podem ser de cinco tipos diferentes denominados H1, H2, H3, H4 e H5. As histonas H2A, H2B, H3 e H4 formam o nucleossomo que é envolvido pelo DNA e é importante para a própria compactação da cromatina. As histonas H1 e H5 são chamadas de histonas ligantes e são responsáveis pela manutenção da ligação do DNA no nucleossomo. As histonas possuem caudas N terminais de aminoácidos que podem sofrer modificações, tais como metilação, acetilação, fosforilação e ubiquitinação. Essas modificações estão associadas com estados de ativação e repressão da expressão gênica. Por exemplo, a acetilação da lisina 27 da histona H3 (H3K27ac) se encontra em regiões promotoras e de enhacers e está associada a genes ativos. Enquanto

isso a trimetilação da lisina 27 da histona H3 (H3K27me3) está associada (~~[illegible]~~) a inativação da expressão gênica. As técnicas mais modernas para o estudo das modificações de Histona são chamadas de Chromatin Immunoprecipitation Sequencing (~~[illegible]~~) (ChIP-Seq) e Cleavage Under Target and Tagmentation (CUT&Tag). No ChIP-seq são utilizados anticorpos contra a modificação que se deseja estudar e depois é realizado o processo de imunoprecipitação. O DNA ligado a histona que possuia a modificação é purificado e preparado para o sequenciamento via NGS. A técnica de CUT&Tag é considerada mais atual e utiliza uma estratégia similar ao ATAC-Seq. Nesse caso, utilizando anticorpos, a Tn5 ligada com (~~[illegible]~~) sequências adaptadoras reconhece a modificação e cliva o DNA ao redor da histona. As sequências clivados se ligam as sequências adaptadoras e já podem ser submetidas a amplificação e sequenciamento. Os métodos de análise dos dados produzidos pelas duas técnicas é bem similar ao do ATAC-Seq e também permite mapear nesse caso a presença da modificação de histona em todo o genoma.

A metilação do DNA é o processo ao qual é adicionado um grupo metil a um nucleotídeo, comumente uma citosina que precede uma guanina (CpG). A metilação do DNA está associado a inativação da expressão gênica. Em diferentes tipos de cânceres é possível, por exemplo, detectar a hipermetilação do DNA em regiões promotoras de genes considerados supressores de tumor. A técnica mais moderna para o estudo do estado de metilação do DNA é chamada de Whole Genome Bisulfite Sequencing (WGBS). Nela, o DNA é tratado com um composto chamado bisulfite capaz de converter citosinas não metiladas em Uracila. O DNA é então sequenciado e, no processo de análise dos dados, (~~[illegible]~~) é possível detectar as citosinas que estavam metiladas através do mapeamento das leituras ao genoma.

Os estudos (~~[illegible]~~) epigenômicos avançaram muito

nos últimos anos graças a modernização dos métodos de sequenciamento e de análise. Através da bioinformática é possível processar, analisar e integrar esses diferentes conjuntos de dados e ampliar os nossos conhecimento acerca dos diferentes processos moleculares envolvidos na epigenética.

2- Controle da expressão gênica com ênfase em métodos computacionais

O controle da expressão gênica pode se dá em diferentes partes do processo de transcrição e envolven diferentes fatores como: fatores de transcrição, cofatores, enhancers e RNAs não codificantes. Além disso, diferentes mecanismos epigenéticos podem estar envolvidos. Com o avanço das abordagens de Next Generation Sequencing (NGS) diferentes métodos e ferramentas computacionais se desenvolveram buscando aprofundar nosso conhecimento biológico através do processamento de grandes volumes de dados.

Para estudar o controle é preciso primeiramente quantificar a expressão gênica. Para essa finalidade se utiliza o método de RNA-Seq. Nela, o RNA é fragmentado, convertido em cDNA e as pontas dos fragmentos são sequenciadas. Esse processo gera milhões de leituras que são armazenada em grandes arquivos digitais chamados de Fastq. O método computacional de análise desses dados envolve primeiramente o controle de qualidade das leituras presentes no Fastq. Como durante o processo são utilizadas sequências adaptadoras para o sequenciamento, é preciso retirá-las das leituras. Além disso, as extremidades das leituras geradas por NGS tendem a ter uma pior qualidade, sendo necessário também retirá-las, sendo esse processo chamado de trimming. As ferramentas computacionais que realizam esses processos são várias, mas as mais comuns são Trimmomatic e o Cutadapt.

Depois do trimming pode-se avaliar as qualidades das leituras através de diferentes métricas presentes no software FastQC. As leituras de qualidade são então mapeadas ao genoma de referência utilizando mapeadores como o STAR, que mapeia as (ed) leituras levando em consideração regiões de splicing. Isso é necessário uma vez que uma leitura pode vir de uma região do transcrito referente a região entre dois exons. Uma vez mapeadas as leituras é avaliado em quais genes essas leituras (ed) foram mapeadas. Essa avaliação leva a quantificação e, posteriormente à normalização. Para a contagem de leituras em genes utiliza-se softwares como featureCounts, parte do pacote Subread, ou HTSeq. A normalização pode ser feita de diferentes e existe alguns debates dentro da comunidade científica sobre métodos são mais adequados. Os métodos de normalização mais comuns são o Reads Per Kilobase Million (RPKM) e o Transcripts Per Million (TPM), os quais podem ser obtidos através de ferramentas com RSEM e Salmon. Para analisar o controle da expressão gênica em diferentes condições, faz-se uma análise de expressão diferencial. Nela, são aplicados métodos estatísticos para determinar quais genes foram regulados positiva ou negativamente. As ferramentas mais utilizadas para essa tarefa estão presentes como pacotes (ed) presentes na linguagem de programação R.

A regulação da expressão gênica pode se dar através da atuação de fatores de transcrição. Sendo assim, determinar as regiões de ligação de diferentes fatores de transcrição é essencial. Para isso, utiliza-se técnicas com ChIP-seq e CUT&Tag com anticorpos específicos. Os dados obtidos por essas duas técnicas podem ser (ed) pré-processados de maneira similar aos dados de RNA-seq. Porém o mapeamento é feito através de ferramentas como Bowtie (versão 2) e BWA. Uma vez mapeados é preciso (ed) os locais de acúmulo de leituras, uma vez que eles

estão relacionados aos locais de ligação da proteína ao DNA. Para isso, utiliza-se a ferramenta MACS2, (...) a qual quantifica o sinal (leituras) buscando por regiões de acúmulo (picos) e compara esse sinal com o ruído de leituras espúrias mapeadas em locais aleatórios do genoma, determinando assim onde estavam ligados o fator de transcrição. Essa metodologia computacional pode ser utilizada para outras técnicas de sequenciamento como o ATAC-seq, porém o MACS2, nesse caso, determinará os locais de acessibilidade da cromatina no [redacted]. Quando se deseja fazer análise de ligação diferencial, (...) buscando regiões que tiveram a ligação do fator alterada entre duas condições utiliza-se a ferramenta CSAW. É possível ainda realizar a detecção de que genes o fator está regulando utilizando a ferramenta ChIPSeeker.

Outras metodologias de sequenciamento podem ser utilizadas para avaliar o controle da expressão gênica, tais como WBGS, que avalia os níveis de metilação no genoma, o qual tem implicações diretas na regulação gênica. A técnica de smallRNA-seq pode auxiliar a detectar pequenas RNAs não codificantes capazes de regular a expressão gênica em níveis transcricionais e pós-transcricionais. O estudo e integração de diferentes dados ômicos é essencial para os diferentes trabalhos, aprofundando novo conhecimento molecular sobre o controle da expressão gênica.