FIOCRUZ

Concurso Público Fiocruz 2023

Pesquisador em Saúde Pública

Prova Discursiva

PE29

Biologia de Sistemas com foco em Genômica e Transcriptômica de Células Individuais

Espelho de Resposta

Pontuação de cada Questão Discursiva conforme Anexo II do Edital nº 3, de acordo com a Unidade detentora da vaga.

Espera-se que o candidato, no desenvolvimento do tema, tenha feito considerações técnicas adequadas sobre os seguintes pontos:

Questão 01

1) A) As leituras brutas geradas devem ser rigorosamente pré-processadas, com objetivo de remover regiões com bases de baixa qualidade. Esse rigor deve-se à necessidade da montagem de um transcritoma de referência, e que esse transcritoma não contenha erros de sequenciamento. B) Para a montagem do transcritoma, pode-se utilizar qualquer montador que use grafos de Brujin, como rnaSpades, Trinity, Velvet Oases. C) A quantificação pode ser feita através do mapeamento com algum mapeador que não precise identificar exons e introns - como STAR, HiSat, Bowtie2, BWA – uma vez que o trancritoma montado reflete o mRNA já sem introns. Outra forma de quantificar é através de K-mer usando ferramentas como Salmon. D) Por tratar-se de um trabalho descritivo, as réplicas biológicas não têm muito impacto, uma vez que o objetivo não é fazer inferências. Por ser descritivo, e demandar uma comparação intra-individual, é necessária uma normalização com o cálculo de TPM, FPKM, RPKM, ou abordagem semelhante para acessar a quantidade da expressão dos genes em relação a outros genes dentro da mesma amostra. E) A contextualização biológica pode ser feita com bancos de dados secundários de funções, como o Gene Ontlology, Kegg Pathways, Panther...
2) A) As leituras brutas geradas devem ser pré-processadas, com objetivo de remover regiões com bases de baixa qualidade. Não é necessário um rigor alto, uma vez que o próprio mapeamento à referência já funciona como filtro de qualidade. No caso de RNAseq provenientes de célula única é necessária uma etapa da identificação de UMI para separar sequências provenientes de células únicas. Tal identificação pode ser feita por ferramentas como Cell Ranger (10x Genomics), ferramentas para Drop-Seq ou programas customizados para classificar as sequências pelo seu "código de barras". B) Não há necessidade de montagem do transcritoma, uma vez que o camundongo é um organismo modelo com genoma sequenciado e bem anotado que pode ser usado como referência para mapeamento ou quantificação. C) A quantificação pode ser feita através do mapeamento com algum alinhador *splicing aware* – capaz de identificar os introns durante o mapeamento ao genomas referência - como STAR, HiSat, BBMap. Outra forma de quantificar é através de K-mer usando ferramentas como Salmon, usando o transcritoma (CDS) como referência. Nos casos de RNAseq de células únicas, é necessária a união de células do mesmo tecido usando uma estratégia conhecida por "pseudobulk", que reconstitui o transcritoma do tecido a partir das células únicas. D) Por tratar-se de um trabalho comparativo, as réplicas biológicas são

essenciais para identificar a variação da expressão de um gene explicadas pela variação biológica, e não pela condição experimental. Por se tratar de análise comparativa do mesmo gene em organismos diferentes, não há a necessidade de normalização, e ferramentas baseadas em distribuição binomial negativas – como edgeR, limma e DESeq2 – usam as contagens absolutas para os testes estatísticos. E) A contextualização biológica pode ser feita com bancos de dados secundários de funções, como o Gene Ontlology, Kegg Pathways, Panther, e complementada com análises de *Gene Set Enrichment Analysis*.

Questão 02

A análise de variantes a partir de sequenciamento genômico tem como objetivo identificar mutações genéticas que podem ser associadas a fenótipos.

A) Para acessar tal informação biológica, é essencial o controle de qualidade das sequências. Somente leituras com bases de alta qualidade devem seguir nas análises. A remoção de leituras e regiões de baixa qualidade garante que uma variante de nucleotídeo não seja erroneamente atribuída a um artefato de erro de sequenciamento.
B) As sequências de alta qualidade deve ser precisamente mapeadas a um genoma de referência. É essencial que a ferramenta escolhida tenha uma precisão alta, mesmo que implique em um alto número de sequências não mapeadas. Nesse caso, a especificidade é mais importante que a sensibilidade, uma vez que um mapeamento espúrio pode levar à predição de uma variante genética erroneamente.
C) As sequências mapeadas são comparadas à referência para a determinação da existência de uma mutação real, e exclusão da hipótese de uma troca de bases espúria. Tal decisão é feita a partir do número de observações que suportem a existência da mutação. Por isso, a profundidade de sequenciamento alta é essencial, pois quanto mais observações evidenciando um evento, maior a probabilidade de ele ser biologicamente real e reprodutível nas proporções observadas, permitindo inclusive a classificação de heterozigose e homozigose.
D) Algumas mutações têm impacto biológico e funcional. As mutações em regiões codificantes de proteínas podem ter seu impacto predito através da identificação da alteração de aminoácidos na proteína codificada, ou pela inserção de um códon de parada, ou mudança na janela de leitura dos códons. Ferramentas como SIFT e PolyPhen identificam qual o impacto, e podem anotar a mutação com informações sobre a possível patogenicidade. Mutações sinônimas tendem as não ter impacto, enquanto mutações associadas surgimento de códon de parada e mudança da janela de leitura tendem a ser classificadas como patogênicas. Bancos de dados como ClinVar, gnomAD, e 1000 genomes trazem dados clínicos e populacionais para a anotação das mutações.
E) Muitas mutações fora de regiões codificantes ainda podem ter papel de alteração de fenótipos. Mutações em regiões regulatórias e intergênicas podem ter impacto, porém esse não pode ser predito pelas ferramentas anteriormente mencionadas, uma vez que não há alteração direta na produção de uma proteína, por exemplo. O impacto dessas mutações pode ser identificado através dos GWAS. Os estudos de associação genômica têm como objetivo identificar mutações que têm um papel indireto na determinação de um fenótipo. Como o papel é indireto, e depende de outros fatores, é necessário um número muito grande de observações para a inferência estatística da associação, fazendo com que o experimento demande o sequenciamento de milhares de amostras, tornando o experimento difícil e financeiramente custoso.