FIOCRUZ

Concurso Público Fiocruz 2023

Pesquisador em Saúde Pública

Prova Discursiva

PE69

Bioinformática aplicada a doenças infecciosas

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

a) Geração de dados de leituras longas usando tecnologias como Oxford Nanopore ou PacBio; ou leituras híbridas usando a combinação das sequências longas com sequências curtas geradas por Illumina, ThermoFisher Ion, ou BGI-Seq. Não se deve usar leituras curtas, pois essas são ineficazes nas montagens de regiões repetitivas. A estratégia de montagem deve ser a "*de novo",* uma vez que o objetivo é identificar diferença entre a nova espécie e os membros da sua família. A montagem por referência identificaria apenas o que é semelhante, descartando as diferenças genômicas. A predição funcional pode ser feita com pipelines como Augustus, Braker e Maker, ao usar predição *ab initio* de ORFs, seguida por buscas de evidências evolutivas a partir dos genomas de outras espécies da família.

b) Geração de dados de leituras curtas com o objetivo de ter profundidade e menor custo. Montagem a partir de referência, uma vez que diminui o risco de erros de montagem, e permite a comparação com as estruturas genômicas o SARS-CoV-2 referência, que é o base para a caracterização de variantes. Predição gênica por referência, com o objetivo de identificar os genes com mutações e localizar as variações que caracterizam as variantes do vírus.

c) Geração de dados de leituras curtas usando Illumina ou BGI-Seq, que são tecnologias que geram uma grande profundidade de sequenciamento. Montagem *de novo* utilizando algoritmo de grafos *de Brujin*. Predição de genes *ab initio* buscando por todas as ORFs possíveis para identificar todo o potencial funcional do organismo, e uso de ferramenta para a identificação do 16s rRNA, ou de gene marcador para a caracterização taxonômica.

Questão 02

O sequenciamento shotgun fragmenta aleatoriamente todo o DNA contido no ambiente, com isso ele pode identificar organismos de todos os táxons, embora a classificação taxonômica seja limitada a poucas leituras com potencial de distinção taxonômica. O sequenciamento aleatório gera leituras geradas a partir das porções funcionais do genoma, o que permite montar, identificar genes e caracterizar o potencial funcional. Pipelines para análise envolvem a remoção de regiões de baixa qualidade das leituras, usando ferramentas como Trimmomatic e Prinseq; montagem do metagenoma usando ferramentas específicas como MetaSpades e MegaHit; predição de genes *ab initio* como GeneMark para eucariotos e procariotos, ou prodigal para Bactérias e Archaea. Os genes preditos são funcionalmente caracterizados usando ferramentas como Diamond, BLAST, HMMer, e bancos de dados secundários como Pfam, KEGG, COG.

O sequenciamento do gene do 16s rRNA acessa somente uma região do gene que tem um potencial de caracterização taxonômica muito forte, porém não acessa outras regiões do genoma com informação funcional. Embora a caracterização taxonômica seja precisa, ela é limitada aos organismos que possuem tal gene: Bactéria e Archaea. Por sequenciar somente um gene usado como marcador taxonômico, não é possível acessar com precisão a composição funcional do ecossistema. Pipelines envolvem a remoção de regiões com baixa qualidade das leituras, a junção de pares de leituras para a reconstrução do amplicon, a identificação de Unidades Taxonômicas Operacionais (OTU) ou Variantes de Sequência do Amplicon (ASVs), e a classificação taxonômica dessas OTUs ou ASVs. Ferramentas como dada2 e qiime2 fazem a montagem dos amplicons, e bancos dados como Silva, RDP e GreenGenes fornecem dados para a caracterização taxonômica.