FIOCRUZ

# Concurso Público Fiocruz 2023

## Pesquisador em Saúde Pública

## Prova Discursiva

## PE31

## Virologia Evolutiva e Computacional

**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**

Pontos fundamentais

- Como funcionam as diferentes metodologias (PCR em tempo real, sequenciamento de Sanger, sequenciamento NGS de reads curtas e longas).
- Relação entre custo, tempo de processamento e geração de resultados de cada método e sua adequação para laboratórios de análises clínicas e de pesquisa fundamental.
- rtPCR básico versus alelo-específico (ou multiplex) para diferentes variantes.
- Vantagem do NGS de reads curtas para estudo de variações intra-hospedeiro.
- Vantagem do NGS de reads longas para determinação de fase de variações nucleotídicas virais.
- Vantagem dos métodos de acordo com o número de amostras analisadas e tempo de geração dos resultados.

A PCR em tempo real (rtPCR) é sem dúvida a metodologia diagnóstica mais custo-efetiva. A rtPCR simples detecta somente a presença do vírus em amostras de swabs naso-orofaríngeos e de outros fluidos e secreções pela detecção de alvos específicos (um ou dois genes diferentes do vírus). Com o surgimento e estabelecimento de outros variantes virais, PCRs alelo-específicos permitiram a identificação de variantes específicos, posicionando-se os primers em variações nucleotídicas típicas de cada variante, de forma multiplexada (ou seja, tendo primers específicos para cada variante simultaneamente). A rtPCR diagnostica a infecção e pode até identificar, em escala mais grosseira, alguns variantes. No entanto, variantes mais específicas com assinaturas em outras áreas do genoma não serão identificadas. Além disso, a rtPCR está susceptível a resultados falso-negativos decorrentes de variações nas sequências em que os primers da PCR anelam, um evento raro mas possível, e já descrito inclusive para SARS-CoV-2. A rtPCR, embora útil e custo-efetiva para a detecção do vírus, não gera a sequência do vírus (nem parcial e muito menos total), e desta forma não é útil para estudos filogenéticos, filodinâmicos e de evolução viral. Em um cenário de saúde pública em meio a uma pandemia, a rtPCR é a mais indicada para laboratórios diagnósticos. Vale ressaltar, no entanto, que métodos filogenéticos e filogeográficos atualmente empregados no estudo da dispersão viral localizada e internacional, pilares da vigilância genômica de novas epidemias, requer técnicas de sequenciamento viral para a obtenção de informação filogenética qualificada.

Neste novo cenário da vigilância em saúde pública, o sequenciamento parcial ou total do genoma viral se torna essencial. Através das sequências do vírus, podemos avaliar as relações entre genomas virais de forma comparativa, e inferir relações evolutivas e filogenéticas entre eles. Isso permite traçar as trajetórias de transmissão do vírus ao longo do espaço e, se aplicado um modelo de relógio molecular, também sua trajetória temporal. Essas metodologias permitem em última análise inferir os padrões de dispersão global de um agente viral, o que tem sido feito de forma massiva para o SARS-CoV-2.

Embora a metodologia de sequenciamento por Sanger (com base na incorporação de didesoxi-nucleotídeos que interrompem a síntese da fita de DNA) possa ser utilizada para o sequenciamento do genoma completo do SARS-CoV-2, este processo é dispendioso e demanda muito tempo, e por isso ele é mais usado para o sequenciamento de regiões parciais do vírus. Já os sequenciamentos do tipo de nova geração (NGS, do inglês “next-generation sequencing”) são atualmente muito menos custosos e, desta forma, mais utilizados para sequenciar o genoma completo do SARS-CoV-2. Para a identificação de variantes específicas do vírus, o sequenciamento de Sanger precisa incluir regiões que definam tais variantes (assinaturas), ao passo que as metodologias NGS de genoma total necessariamente incluirão tais assinaturas. Para fins de rastreio filogenético / filodinâmico, o sequenciamento do genoma completo do vírus é notadamente importante, muitas vezes essencial, já que o SARS-CoV-2 possui poucas variações nucleotídicas fixadas em genomas de vírus proximamente relacionados (por exemplo, em um cluster de transmissão local), possuindo poucos sítios informativos filogeneticamente ao longo do genoma que permitam uma distinção confiável entre vírus proximamente relacionados.

Dentre as principais metodologias de sequenciamento de NGS utilizadas para o sequenciamento do SARS-CoV-2 se encontram o sequenciamento por reads curtas ou por reads longas. No primeiro caso, pequenos fragmentos ao longo de todo o genoma viral são amplificados por PCR de forma multiplexada (em dois diferentes pools de reações) e uma sequência-consenso viral completa é montada in silico a partir dos pequenos fragmentos sobrepostos. Esta metodologia é representada majoritariamente pela empresa Illumina no Brasil. No segundo caso, fitas únicas são lidas e reads únicas são geradas, sendo a sequência-consenso gerada a partir do consenso destas fitas. No Brasil, a tecnologia nanopore (Oxford) domina este tipo de metodologia. Cada uma destas duas metodologias possui vantagens e limitações. Em estudos de evolução viral intra-hospedeiro, onde se queira estudar o surgimento ou desaparecimento de variações nucleotídicas em uma ou várias amostras ao longo do tempo dentro de um indivíduo infectado, o método de pequenas reads é mais apropriado, pois possui menores taxas de erro de leitura dos nucleotídeos sequenciados. Ele permite a visualização e análise de todas as variações na população de reads para cada posição nucleotídica da sequência do genoma viral, podendo-se calcular a frequência de tais variações. O limite das frequências toma como base a profundidade do sequenciamento em cada posição, e é uma função da própria carga viral da amostra. Este tipo de análise não pode ser feita com metodologias de reads longas como o nanopore, que possui taxas de erro de leitura de nucleotídeos mais altas, gerando uma quantidade grande de variações nucleotídicas de baixa frequência que são na verdade falso-positivas. Já a metodologia de reads longas da Pacific Biosciences não possui tal limitação, mas seu custo (equipamento e corrida) é superior e pouco usado no Brasil com propósito de sequenciamento de vírus. Por outro lado, a determinação de fase (ou colinearidade) entre duas variações nucleotídicas encontradas, ou seja, se duas dadas variações se encontram na mesma molécula do genoma viral, só pode ser alcançada com o uso de métodos de reads longas, onde as duas variações são lidas a partir de uma mesma fita de DNA durante o sequenciamento. Somente variações muito próximas na sequência (aquelas que se encontrem dentro de uma mesma read curta) podem ter sua fase determinada pelo sequenciamento de reads curtas, o que limita profundamente este tipo de análise pelo método.

As plataformas de NGS de reads longas podem ser menos ou mais custosas (comparando-se por exemplo a tecnologia Oxford nanopore com a da Pacific Biosciences), mas os rendimentos de cada corrida são proporcionalmente menores com o menor custo. A tecnologia de reads curtas da Illumina possui um custo e um desempenho intermediário (utilizando-se por exemplo a plataforma MiSeq). A escolha destas diferentes tecnologias também deve levar em consideração a demanda de sequências a serem geradas, dado que plataformas com alto desempenho requerem o acúmulo de muitas amostras para executar a reação de sequenciamento, enquanto plataformas com menor “throughput” podem rodar menos sequências, mas de forma mais frequente.

**Questão 02**

- Conceito de “cluster” em inferência filogenética e como comprovar a existência de clusters (uso de sequências próximas, de localidade e data próximas, e de bases de dados).
- Limitações das análises de cluster (subamostragem, vírus com taxa evolutiva lenta).
- Datação de transmissão, relógio molecular, ancestral comum mais recente.
- Limitações de relógio molecular e datação (vírus com taxa evolutiva lenta, erros de sequenciamento e geração de consenso).
- Filogeografia para dispersão espaço-temporal de vírus.
- Limitação de amostragem na filogeografia / filodinâmica: subamostragem, superamostragem, falta de metadados.

O alinhamento de sequências genômicas virais e a inferência de suas relações filogenéticas podem ser utilizados para gerar hipóteses sobre as rotas de transmissão do patógeno. Neste sentido, sequências virais de portadores do vírus que tenham sido expostos à mesma fonte infectante normalmente se agruparão filogeneticamente num “cluster”. Uma confiabilidade maior deste agrupamento, ou seja, rejeitando a hipótese nula de agrupamento destas amostras ao acaso, pode ser adquirida incluindo-se na inferência filogenética sequências de casos de fora do “cluster” em questão e adicionalmente, sequências de referência de bases de dados globais mais próximas das sequências do “cluster” (obtidas, por exemplo, através da recuperação destas sequências com o uso da ferramenta Blast do NCBI usando as sequências do “cluster” como “query”). A adição de sequências temporal e geograficamente compatíveis com as sequências do “cluster” à inferência filogenética também auxilia na rejeição da hipótese nula, caso aquelas não se agrupem no “cluster”. Já a mistura de sequências-referência ou de outra época ou local geográfico tendem a rejeitar a hipótese da existência do referido “cluster” com uma origem de exposição única.

O método de análise de “clusters” possui limitações. O agrupamento entre duas sequências não pode ser utilizado, por exemplo, para confirmar a transmissão direto do vírus entre dois portadores ou a transmissão de um único doador para vários recipientes, dado que a existência de outros indivíduos ou fontes de exposição não amostrados (sem sequência disponível) nunca pode ser descartada. Isto pode mesmo ser verdadeiro quando dois determinados vírus são idênticos (sem nenhuma substituição nucleotídica visível entre eles), no caso de vírus que possuem uma taxa de evolução muito lenta (mais lenta do que a transmissão entre dois hospedeiros), como é o caso por exemplo do SARS-CoV-2).

A taxa de substituição nucleotídica de um vírus pode ser avaliada quando há diversidade genética suficiente entre os diferentes vírus de uma linhagem ou “cluster”. Se além disso tivermos disponível as datas de amostragem de alguns destes vírus, podemos associar esta informação com a diversidade genética das sequências (aplicando um relógio molecular) e desta forma estimar a sequência ancestral comum mais recente entre os vírus comparados, obtendo assim a informação de quando no tempo aquela linhagem ou “cluster” começou a circular naquela população de hospedeiros. Este ancestral comum a todas as sequências derivadas pode assim ser inferido sem ser amostrado, identificando a circulação do vírus mesmo antes da ocorrência dos primeiros casos clínicos no caso de uma infecção viral patogênica.

A aplicação do relógio molecular para datação da circulação de vírus por inferência filogenética é limitada pela razão entre a taxa de substituição e a taxa de transmissão do vírus. No caso de a primeira ser menor do que a segunda, a data da transmissão não poderá ser inferida, pois as sequências genômicas poderão ser idênticas, sem diversidade genética alguma entre elas. Para vírus com baixa diversidade genética acumulada, erros de sequenciamento ou na geração de sequências-consenso podem ser confundidos com a diversidade real, e podem afetar as estimativas das taxas de evolução e dos tempos de divergência entre duas ou mais sequências.

O uso de sequências genômicas em análises filogeográficas e seu posicionamento na árvore podem informar se eventos de transmissão foram locais ou importados, particularmente se a análise tiver a associação de metadados, como informações de viagens, local de moradia etc. dos indivíduos

amostrados. A incorporação das datas de amostragem também permite a reconstrução do deslocamento espaço-temporal do vírus, informando onde e quando eventos de deslocamento de vírus podem ter ocorrido. No entanto, tais análises requerem cuidado especial na sua interpretação no que concerne a extensão e magnitude da amostragem. A subamostragem de determinados locais geográficos poderá influenciar forte e erroneamente as conclusões das análises filogeográficas, por exemplo levando a uma subestimação do número de introduções do vírus a partir de outras localidades (ou, reciprocamente, a uma superestimação indevida da transmissão comunitária ou local). A ausência de metadados como informações de viagem também limita as análises filogeográficas tornando impossível discernir entre a transmissão direta do vírus entre dois locais distintos ou o envolvimento de um local intermediário no qual informações genômicas não estão disponíveis. Por fim, a superamostragem de determinados locais tendência a posicioná-los como locais doadores (fontes) do vírus. Neste caso, equilibrar as amostragens, reduzindo a quantidade de sequências utilizadas destes locais, pode auxiliar na obtenção de conclusões mais robustas.