

CAPÍTULO 6

INTRODUÇÃO À BIOINFORMÁTICA

Fernanda C. Koyama

Gerente de Pesquisa e Desenvolvimento na OC Precision Medicine –
Grupo Oncoclínicas, São Paulo

Bruno B. de Souza

Bioinformata na OC Precision Medicine – Grupo Oncoclínicas, São Paulo

Introdução

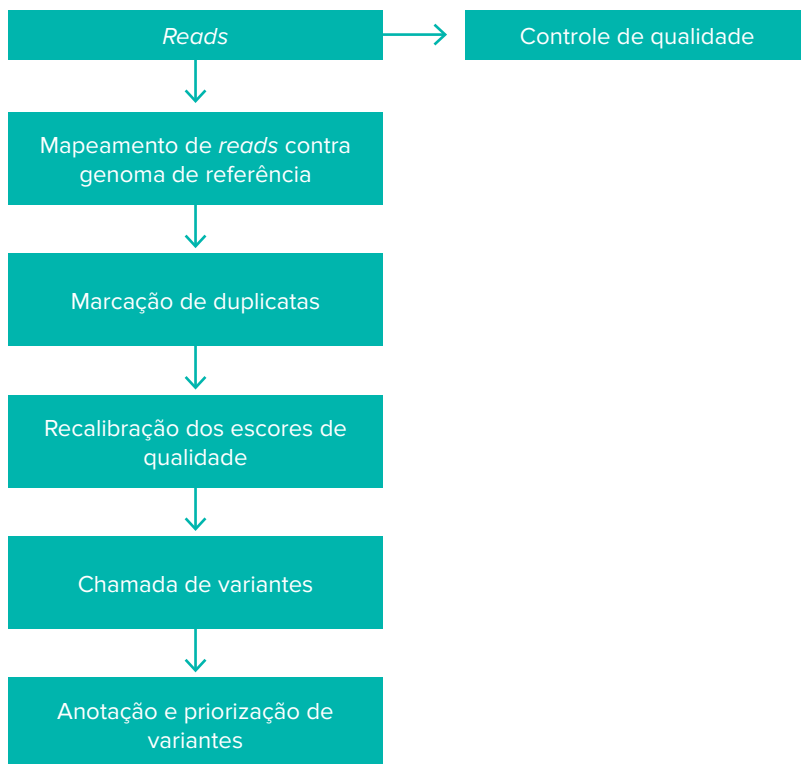
Os avanços tecnológicos das duas últimas décadas resultaram em uma queda dramática no custo do sequenciamento de ácidos nucleicos, ampliando o acesso a essas metodologias. No entanto, o sequenciamento é apenas o início do processo de conversão de uma amostra biológica em informação genética significativa. A análise de dados envolve o uso extensivo de vários métodos computacionais para converter dados brutos em informações de sequência e a aplicação de técnicas de bioinformática para a interpretação dessas informações. A enorme quantidade de dados gerados por tecnologias de sequenciamento de nova geração (NGS) deslocou a carga de trabalho da preparação de amostras para os processos de análise. Isso significa que, junto ao desenvolvimento de tecnologias de sequenciamento, o desenvolvimento simultâneo de soluções de informática apropriadas é urgentemente necessário para possibilitar a interpretação clínica de variantes genômicas¹.

A Bioinformática é um campo interdisciplinar em rápida evolução. Avanços em tecnologias moleculares, como sequenciamento, plataformas de *microarray* e tecnologias *high-throughput* resultaram em um excesso de dados, em resposta ao qual o desenvolvimento de ferramentas e repositórios de *softwares* de bioinformática avançou rapidamente, consumindo uma enorme quantidade de tempo (e, portanto, dinheiro) para sua criação e manutenção. Como consequência, a formação em Bioinformática teve que evoluir e crescer paralelamente à tecnologia, para atender à crescente demanda em diversas áreas dos conhecimentos, que vão da Biologia à Química, Matemática, Física e Ciência da Computação².

Especificamente para laboratórios de diagnóstico molecular, a bioinformática se tornou um importante componente, pois participa dos processos de geração, análise e interpretação de dados moleculares em testes genéticos. Devido à rápida adoção de testes por NGS, os provedores de serviços precisaram desenvolver *workflows* que atendessem a padrões regulatórios e de boas práticas e que fossem flexíveis quanto a frequentes atualizações em termos de metodologia e *softwares* para análise³.

As análises de bioinformática são desenhadas para, primeiramente, realizar a conversão de sinal em dados, dados em informação biológica e informação biológica em acionabilidade ou significância clínica. Esse processo pode ser dividido em três etapas principais: análise primária, secundária e terciária (figura 1)¹. Em resumo, a análise primária consiste em transformar o sinal gerado no instrumento em pequenas sequências de nucleotídeos⁴. A análise secundária envolve o mapeamento das pequenas sequências (ou *reads*) contra um genoma de referência (no caso de Oncologia Molecular, o genoma humano de referência) e, a partir desse mapeamento, determinar os pontos de variação genômica⁵. Por fim, a análise terciária consiste em acessar as características dessas variantes, impacto funcional e significância clínica, quando possível⁶.

Figura 1. Diagrama ilustrando os principais componentes de um *pipeline* de bioinformática



Cada uma dessas etapas requer o uso de bancos de dados, algoritmos, *softwares* e conhecimentos específicos para sua execução. A principal limitação para a ampliação do uso de NGS em laboratórios clínicos é a interpretação de dados de sequenciamento, que ainda é um campo nascente em termos de desenvolvimento de algoritmos, ferramentas analíticas apropriadas e bases de evidências para as associações genótipo-fenótipo. Embora as etapas envolvidas na interpretação e na aplicação dos resultados variem de acordo com o ambiente clínico específico, o propósito do teste e a questão clínica, é provável que haja pontos em comum tanto no *pipeline* de análise básica quanto nas ferramentas utilizadas. Da mesma forma, embora alguns detalhes possam mudar com a introdução de tecnologias de sequenciamento de terceira geração, como aquelas que envolvem a detecção de moléculas em tempo real, o desafio da análise de dados para tecnologias de sequenciamento *high-throughput* permanecerá essencialmente o mesmo¹.

Neste capítulo, descreveremos primeiramente os conceitos básicos e os formatos de arquivos utilizados em bioinformática para, então, nos aprofundarmos nos principais processos das análises primária, secundária e terciária, comentando as principais ferramentas utilizadas em cada passo, assim como as bases de dados mais importantes no cenário atual que auxiliam na interpretação de dados na área da Oncologia. Também mostraremos de forma ilustrativa alguns exemplos de comandos básicos de algumas ferramentas. É importante salientar que o objetivo deste capítulo não é fornecer um treinamento em bioinformática, apenas apresentar uma visão global em como os *softwares* funcionam e como podem ser aplicadas na área de Oncologia.

Principais conceitos e formatos de arquivos

Cobertura

A cobertura média define-se como o número médio de vezes que cada nucleotídeo foi sequenciado, baseado em um certo número de *reads* de determinado comprimento e a suposição de que as *reads* possuem distribuição aleatória. A cobertura real representa a quantidade média de leituras em cada região do genoma, baseando-se no mapeamento⁷.

Genoma de referência

O genoma de referência é uma sequência de DNA de alta qualidade gerada a partir de um *pool* de indivíduos, que pode servir como uma referência comum para a compreensão da base genética de diversas doenças. Especificamente, a sequência de referência é a base para bancos de dados e algoritmos de bioinformática usados para definir regiões-alvo, realizar estudos de ampla associação genômica (*genome-wide association studies* – GWAS) ou medir a conservação entre espécies, por exemplo.

Formato FASTA

Arquivos FASTA são arquivos que podem conter sequências de DNA, RNA ou proteínas, sendo que tanto nucleotídeos e aminoácidos são representados com códigos de letra única. Um arquivo FASTA (figura 2) pode conter uma ou mais sequências. Um arquivo com várias sequências é chamado de arquivo multi-FASTA. A primeira linha de cada nova entrada começa com um sinal de maior (>), que é seguido por uma descrição de linha única. Essa linha de título pode ter qualquer comprimento, incluindo simplesmente o sinal de maior seguido por nenhum caractere adicional. As linhas subsequentes contêm a sequência. Recomenda-se que as linhas da sequência tenham menos de 80 caracteres⁸.

Figura 2. Exemplo de formato FASTA

```
>Sequencia 1
GGCAGATTCCCCCTAGACCCGCCGACCATGGTCAGGCATGCCCTCTCTATCGCTGGGCAAGCCAGAGGGT
ATTAACAGTCTGGAGGCTGGCGGGGAGGCCAGCTGAGTCTCTGAGCAGAGCCAGCGCAGCCACCGAGACAC
ATGAGAGCCCTCAGACTCTCTGGCCCTATTGGCCCTGGCCGCACTTTGCATCGCTGGCCAGGAGGTGAGTGGCC
CAGCTCCCTCAGGCCGATTCAGTGGGGGCTGAGAGGAGGAGCACCATGGCCACCTCTTCTCAGCCCTTTG
GCTGGCAGTCCCTTTCAGTCTAACACCTTTTGCAGGCTCAATCAATTCAGTCCCTCTGCTCTGAGAGG
GAGAGGAGGAGAGAGCAGCTGCCCGAGACGAGGAGGAGAGGATGAGGGCCCTGGGATGAGCTGGGGTGAAC
```

Formato FastQ

Dentre os tipos de arquivos de saída das técnicas de sequenciamento, o formato FASTQ (figura 3) emergiu como um formato que permite a troca de dados entre ferramentas. Ele nada mais é do que uma extensão do formato FASTA, possuindo a característica de conter escores de qualidade associados a cada nucleotídeo de uma sequência.

Figura 3. Formato FASTQ

```
@sequencia_1
TACGTGATCGATGAGAGATCGCTAGA.
+
986B;C<@>;=;;519||.166A21
```

O formato FASTQ possui quatro linhas. A primeira linha começa com o sinal de '@', que contém informações da corrida, sequenciador, adaptadores etc. É de formato livre e sem limite de tamanho.

Na segunda linha, vem a sequência de nucleotídeos em si (ATCG), também em formato de letras únicas (IUPAC). A terceira linha começa com um sinal de '+', sendo opcionalmente precedida da mesma informação localizada no cabeçalho. A quarta linha contém os escores de qualidade para as bases contidas na linha 2. Esta deve conter o mesmo número de símbolos quanto o de letras na sequência⁹.

Veja esse conteúdo na íntegra
acessando a OC Academia:
<http://ocacademia.com>

