1. O que é um arquivo do tipo FASTQ e como posso verificar se um FASTQ é válido?

Um arquivo *fastq* diferente do arquivo fasta, tem o início da sequência um "@" e segue com alguns dados que contém informações diversas sobre o sequenciamento, como equipamento usado, número da corrida, trilha, identificação do cluster, dentre outras. A segunda linha contém as sequencias de nucleotídeos a terceira linha há uma repetição da primeira linha e geralmente uma substituição do "@" pelo sinal de + e pôr fim a quarta linha encontram-se os valores da qualidade, que estão representados pelas notas *phred* de qualidade. Para testar um arquivo *fastq* pode ser feita de várias maneiras, pode-se inicialmente avaliar o arquivo *gzipped* extraindo e/ou abrindo os arquivos quando os arquivos são pequenos, porém geralmente os arquivos são muito grandes, nesse caso cabe a utilização de softwares com *fastqc* para a avaliação de todo o arquivos, como quantidade de Ns, valor médio de qualidade das bases, comprimento dos *reads*, regiões super-representadas, presença de adaptadores (importante) etc. softwares como o *multiQC* podem gerar gráficos e reports de todas as leituras feitas pelo *fastqc*.

2. Quais são as etapas mais comuns de um pipeline de bioinformática de NGS para DNAseq e quais ferramentas podem ser usadas em cada etapa?

Inicialmente é feita o download dos dados, armazenamento dos mesmos e backup, em seguida é feita uma análise de qualidade dos dados brutos, para isso podem ser utilizados o *fastqc* que irá gerar um resumo de várias métricas, que podem ser apresentados por reports via *multiQC* (caso seja necessário). O *fastqc* gera gráficos como por exemplo qualidade, tamanho médio dos *reads* e presença de adaptadores dentre outras avaliações. Dentre as avaliações citadas podemos destacar a presença de adaptadores, regiões ligadas ao DNA alvo, que devem ser removidas, para isso podemos usar alguns softwares como o *trimmomatic*, o pacote do *fatxtoolkit* e o *sickle*, esse último faz o processo em paralelo e permite também a filtragem dos *reads* por qualidade e tamanho. Podem também ser aproveitados informações do sequenciador que gerou os dados bem como uso de vários threads para aceleração do processo. Em seguida, porém podem ser feitos no mesmo passo anterior é o processo de "trimagem" a filtragem dos *reads* por qualidade e tamanho a depender do nível de qualidade e complexidade dos dados é necessária uma filtragem mais rigorosa quanto a qualidade do *base call*, por exemplo SNPs e INDELs. Para o processo de trimagem podemos usar novamente o *trimommatic*, *fastxtoolkit* e o *sickle*. O ideal é avaliar novamente os *reads* após os processos de

“clipagem”, *clipper* e “trimagem” *trimmer* para checar se ainda existe algum resíduo de adaptador, sequências de baixa qualidade, Ns etc.

Com os dados filtrados podem ser feitas as montagens por duas estratégias *denovo*, aonde não existem genomas de espécies ou gêneros próximos nesse caso pode-se optar por softwares que façam grafos afim de resolver da melhor maneira possível gargalos nas forquilhas de replicação, um dos mais usados é o *spades*, mas podemos usar *velvet*, *soapdenovo*, entre outros. Ao final de uma montagem *denovo* é importante avaliação da montagem, para esse fim podemos usar o software *quast* que reporta parâmetros importantes como número de *contigs* (ou *scaffolds*), tamanho de *contigs* (ou *scaffolds*), valor de N50, genes preditos já que conta com o *genemark* implementado, e as nova versão pode ser feita a busca por genes conservados pela ferramenta BUSCO que retorna uma porcentagem desses genes, *housekeeping*, encontrados na montagem. Vale salientar que um mesmo montador pode usar *kmers* (palavras chaves, para resolver bolhas de replicação) diferentes cabe ao analista identificar o melhor tamanho. Caso haja necessidade o podem ainda serem feitas etapas como predição de genes, o *genemark* pode ser utilizado tanto para genomas eucariotos como para procariotos, e o *funannotate* pode ser usado para eucariotos. Em seguida o processo de anotação desses genes pode ser feito pelo *prokka*, *rast* (procariotos) e *funannotate* e *interproscan* (eucarioto), etapas como essa dependem da máxima atenção já que o processo de anotação e predição muitas vezes são feitas pela leitura de star/stop códon, que podem gerar erros nos tamanhos dos genes preditos esses erros então podem passar para o processo de anotação, para evitar problemas assim podemos criar bancos de dados, ou alinha sequenciar e visualizar no *artemis*. Outras etapas que ainda podem ser feitas é a metagenômica caso seja específica a condição de cada uma das amostas.

Na montagem por referência usamos uma sequência de organismo próximo, mesma espécie ou gênero (à depender). Inicialmente indexamos a referência, usando o *bwa* (index) ou o bowtie(2), em seguida podem fazer o alinhamento dos reads ao genoma de referência usando o bwa ou bowtie(2) (aln ou mem depende do tamanho dos genomas/fragmento), convertemos os arquivos para o formato sam em seguida para o formato bam, podemos fazer pelo bwa ou pelo samtools. Pelo samtools sort geramos os arquivos sorted e indexamos, removemos de duplicatas, indexamos novamente para fazermos a chamada de variants. Com os arquivos bam podemos gerar os arquivos fasta do genoma montado (consenso) usando *samtools* e *bcftools*. Para chamada de variantes

usamos o *picard* para preparar os grupos e gerar o arquivo *dict* da referência, indexamos e assim os softwares gatk, freeebays, samtools entre outros podem fazer a chamada de variantes. Ao final da chamada de variantes é importante a depender do caso um critério maior quanto a cobertura da base (qualidade da *base call*) variantes raras podem estar subrepresentadas pela baixa cobertura da região, nesse caso usamos o bcftools para filtrar a variantes pela qualidade e qualidade média. Caso seja necessário o uso de mais de um software para chamada de base como gatk e freebays (mais utilizados), podemos comprimir e tabular os arquivos vcf afim de mesclar as duas ou mais "montagens". As variantes são então anotadas pelo *snpEFF*, e caso seja necessário a comparação das condições das mutações versus fenótipo (genótipo vs fenótipo) podemos realiazar um teste de desequilíbrio de ligação quando mais simples, ou usamos o *plink* para realizar um teste de associação, em paralelo inteligência artificial podem auxiliar no resultados.

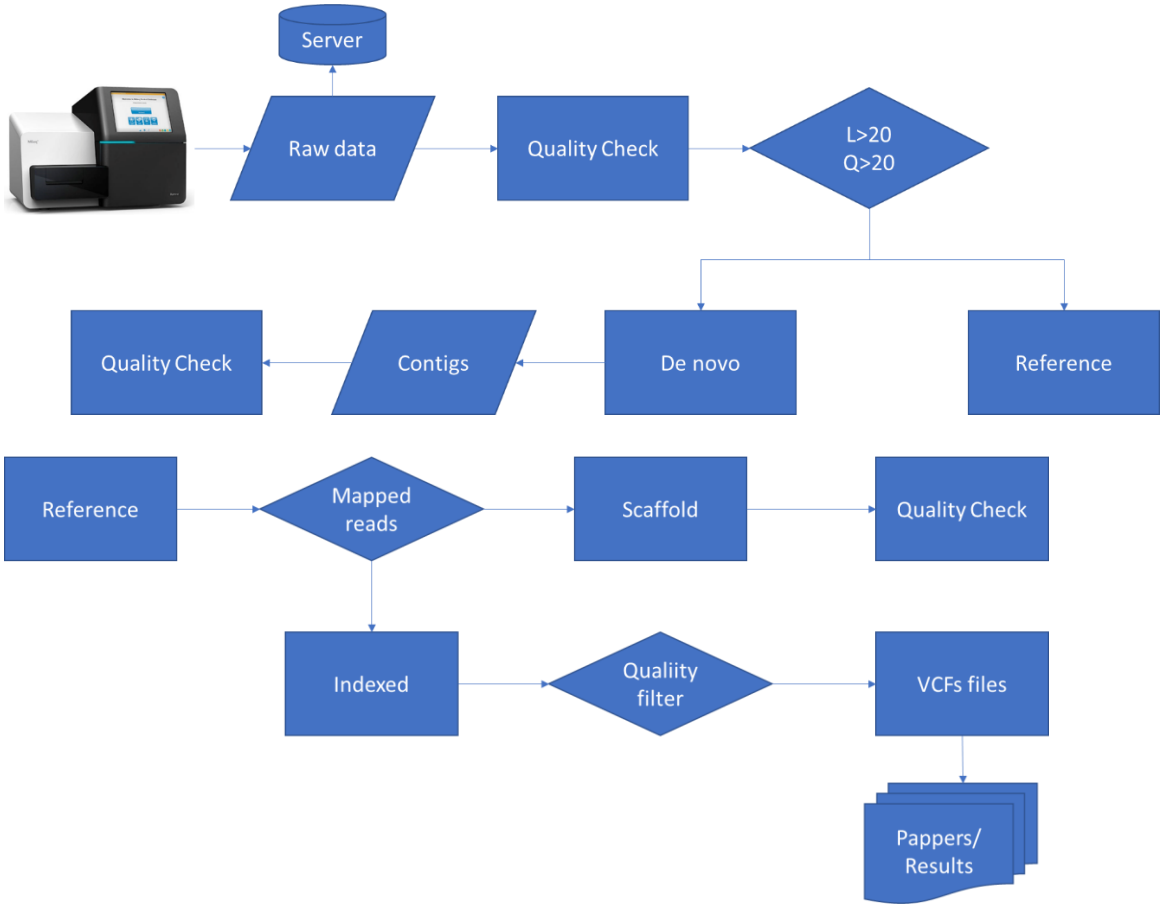
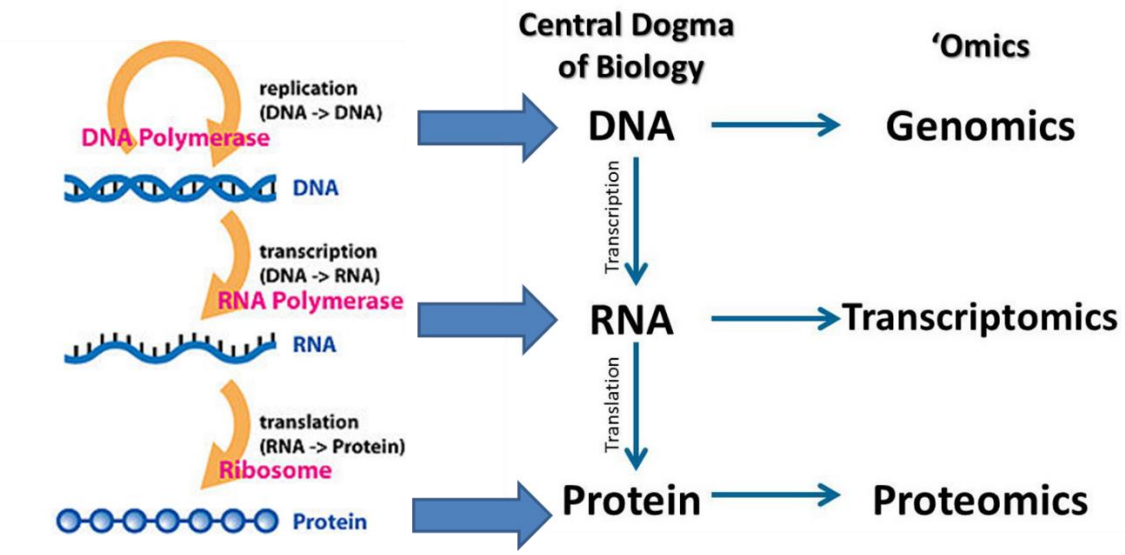


Figura 1. As figuras acima representam de pipeline (fluxograma) de montagem de genomas e transcriptoma. Fonte própria.

3. O que é o dogma central da biologia e como a bioinformática pode ser aplicada em cada um dos processos?



O dogma central da biologia molecular compreende basicamente todos os processos desde o genoma (sequência de nucleotídeos), transcrição (produção de mRNA), tradução (conversão de RNA e proteínas), metabolômica (metabólitos secundários), efetômica, entre outras. Com o advento da biologia molecular é possível gerar dados para entendimento de todos os processos supracitados, e a bioinformática pode ser aplicada a cada um desses processos desde que haja dados para análise, por exemplo o sequenciamento genômico, pode ser analisado de modo a gerar a sequência de nucleotídeos, genes, metabólitos secundários, mutações etc, porém não podem ser analisados genes diferencialmente expressos como é o caso da transcriptoma, que depende de uma biblioteca de RNAseq, mas pode ser analisa pela bioinformática da mesma maneira, desde identificação de genes abundantes diferencialmente expressos, mutações etc. cada técnica utilizada na biologia molecular pode ser aplicada a uma ou mais técnicas de bioinformática cabe ao analista entender a demanda e a complexidade do problema.

4. Considerando que o alinhamento de sequências biológicas é uma das atividades mais recorrentes e importantes na área de bioinformática. Comente sobre os algoritmos computacionais de alinhamentos mais utilizados e em que cenários uns são mais indicados do que outros?

O alinhamento de sequências consiste no processo de comparar duas ou mais sequências (de nucleotídeos ou aminoácidos) de forma a se observar seu nível de similaridade; Quanto aos tipos podemos ter o simples: aquele realizado entre sequências de DNA ou proteínas, desde que duas a duas, o múltiplo: aquele realizado entre mais de duas sequências de DNA ou proteínas, o global: as sequências são alinhadas de ponta a ponta, o local, aonde pedaços das sequências é que são comparados, o heurístico: produz um resultado o mais próximo possível do resultado ótimo, mas, principalmente, produz um resultado de maneira muito veloz e o ótimo: produz o melhor resultado computacionalmente possível.

Para busca em bancos de dados usamos alinhadores heurísticos e local como exemplo podemos citar o blast. Para alinhamento de proteínas e sequencias de DNA (multifasta) usamos alinhadores heurísticos e global. Para o alinhamento de um grande número de fragmentos usamos alinhadores global ou local, ótimo ou heurístico, ade pender da complexidade dos dados e o máximo de erros ou tempo que necessitar a análise. Deve-se considerar que o alinhamento é normalmente baseado em um algoritmo *greedy* que tentará ao máximo executar de melhor maneira e mais rápida o encaixe de sequencias sobrepostas a uma dada referência.