2
Contexto Biológico

Neste capítulo abordaremos o contexto biológico para o entendimento deste trabalho. Serão abordados os aspectos gerais da genômica, expostos os processos do sequenciamento genético e as principais características dos métodos de sequenciamento mais utilizados.

2.1.
Genômica

Genômica é um ramo da bioquímica que estuda o genoma completo de um organismo e seus padrões genéticos existentes. O material genético dos organismos vivos constitui o genótipo e este é transmitido a diferentes gerações por meio da reprodução. Na maioria dos seres vivos as informações genéticas são armazenadas no cromossomo, que é composto por DNA (ácido desoxirribonucleico) e proteína, em alguns vírus esse material genético é o RNA (ácido ribonucleico).

O entendimento da composição molecular de cada organismo é de suma importância pois permite a comparação de genomas de diversos indivíduos da mesma espécie. Sendo assim, é possível correlacionar síndromes com mutações e variações genéticas. Entre espécies próximas é possível aprofundar o entendimento dos mecanismos de evolução.

O DNA é um polinucleotídeo composto por quatro nucleotídeos. Um nucleotídeo é formado por uma base nitrogenada, uma pentose e um grupo de fosfato. Esses quatro nucleotídeos se diferenciam pelas suas bases nitrogenadas: adenina (A), guanina (G), citosina (C) e timina (T). Como pode ser vista na Figura 1, a estrutura do DNA é conhecida como sendo uma dupla hélice, e cada fita de nucleotídeo é unida através de pontes de hidrogênio entre suas bases nitrogenadas.

O RNA é um polinucleotídeo assim como o DNA, porém sua estrutura possuiu apenas uma fita e não possuiu a base nitrogenada timina (T). Em seu lugar existe a base nitrogenada uracilo (U).

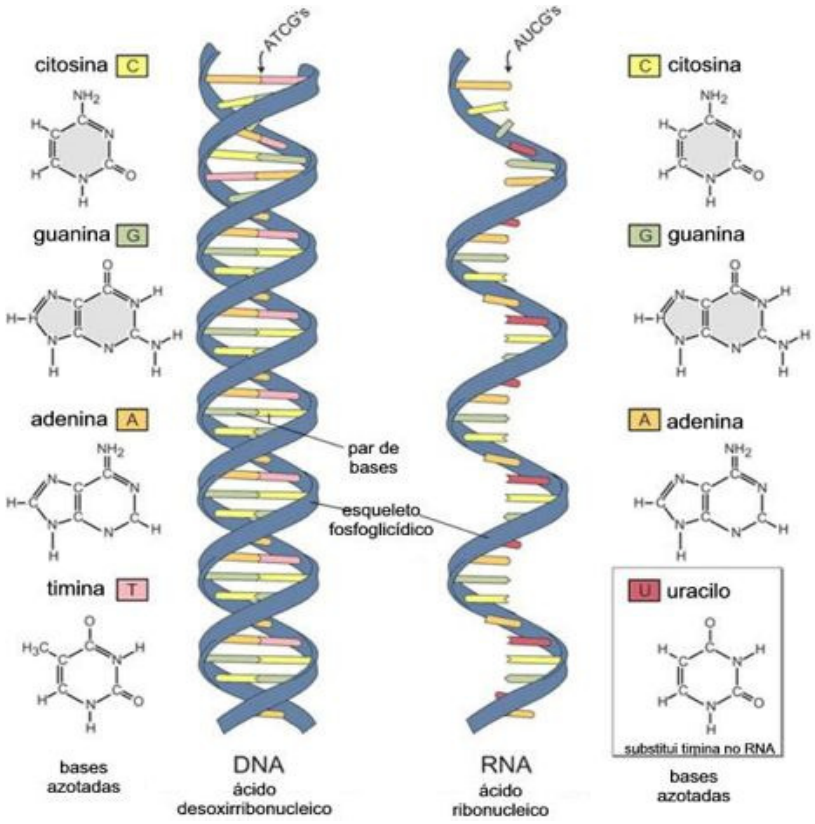

Figura 1 Representação do DNA e RNA

2.2. Sequenciamento

Com a tecnologia que temos atualmente, é inviável fazer a leitura de um genoma por completo, já que a sua cadeia de nucleotídeos pode possuir bilhões de pares de bases nitrogenadas. Portanto, para realizar a leitura o gene precisa ser lido por partes. Ou seja, o cromossomo é fragmentado, cada parte é lida e ao final é realizada a montagem dos fragmentos, obtendo desta forma o sequenciamento completo do genoma.

Um *workflow* para a realização deste trabalho de fragmentação, leitura e montagem é apresentado na Figura 2 e descrito através dos passos abaixo:

A. Fragmentação e leitura do genoma gerando como saída uma biblioteca em arquivo texto com as informações dos fragmentos.
B. A biblioteca é processada por programas de montagem de fragmentos, onde os fragmentos são ligados através de suas sobreposições.
C. São realizadas diversas operações para a remoção de erros de leitura e erros de montagem dos fragmentos.

D. Conjuntos de fragmentos que formam uma sobreposição entre si são unificados, criando um fragmento maior que os fragmentos de entrada, formando os *contigs*.

E. Continuando o processo de montagem do genoma completo, os *contigs* são alinhados de forma a se identificar os espaços vazios entre eles. Essa identificação ocorre através da análise das sobreposições dos *contigs*. Esta sequência de *contigs* alinhados e com as marcações dos espaços vazios são chamados de *scaffolds*.

F. Através da análise das sobreposições dos muitos *scaffolds* é feita a remoção dos espaços vazios, obtendo desta forma o mapeamento completo do genoma.

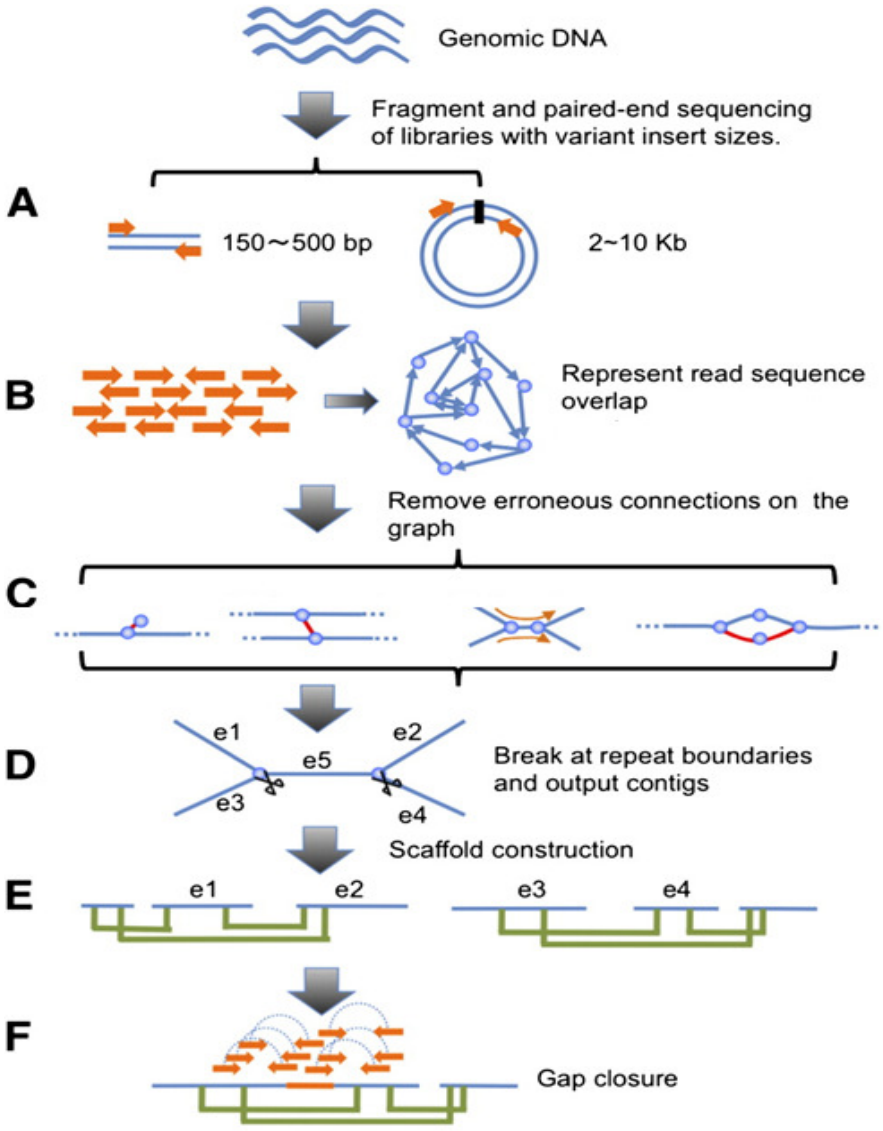

Figura 2 Exemplo de workflow do processo completo de sequenciamento de DNA e RNA[9]

2.3. Cobertura

Cobertura é a média de quantas vezes cada base do genoma foi sequenciada, ou seja, quantas vezes ela foi fragmentada e lida durante o processo de sequenciamento. A quantidade de cobertura requerida está relacionada principalmente a três fatores.

O primeiro fator se dá pela característica do processo de fragmentação. Por se tratar de um processo químico, este é realizado de forma aleatória, não sendo possível controlar quais partes do genoma foram ou não fragmentadas. Portanto, para buscar uma leitura completa do genoma, a fragmentação e a leitura dos fragmentos do cromossomo deve ser realizada diversas vezes.

O segundo fator está ligado ao tamanho do fragmento gerado pelo método de sequenciamento. Como o processo de fragmentação é aleatório, quanto menor o tamanho do fragmento, maior é a probabilidade de os fragmentos terem sobreposição e conter áreas não fragmentados do genoma. Por isso quanto menor o fragmento, maior deve ser a cobertura.

Por último, há um fator relacionado à geração de erros durante o processo de sequenciamento. Estes erros podem ocorrer por falhas durante a leitura dos fragmentos ou por variações genéticas da amostra lida. Quando não existe um genoma de referência para fazer a validação das leituras, a detecção de erros é mais complexa, exigindo uma cobertura ainda maior para identificar os fragmentos errados devido a sua baixa repetição na biblioteca gerada.

2.4. Métodos de sequenciamento

Foram desenvolvidos diversos métodos para a fase de sequenciamento, o mais tradicional é o Sanger, desenvolvido na década de 70. Porém, nos últimos anos, uma nova geração de sequenciamento de DNA (NGS: Next-Generation Sequencing) tem sido desenvolvida com ótimos resultados em relação ao custo e velocidade de sequenciamento, como Illumina, Roche 454, Proton e SOLid [1].

2.4.1.
Método Tradicional (Sanger)

Este método de sequenciamento de fragmentos foi o mais utilizado desde a década de 70. Também conhecido como terminação de cadeia, a leitura de cada nucleotídeo do fragmento é realizada através da adição de nucleotídeos modificados (didesoxiribonucleotídeos), os quais permitem determinar a sequência da cadeia do fragmento de DNA. Esta leitura é realizada durante a síntese destas cadeias a partir do fragmento, pois os nucleotídeos modificados estão marcados radioativamente numa extremidade e assim se diferem entre si por um nucleotídeo. Após, é realizada a separação das cadeias truncadas através de eletroforese, podendo-se assim estabelecer a sequência de nucleotídeos do fragmento de DNA original.

O tamanho gerado de cada fragmento fica em torno de 400 a 900 pares de bases (bp) e a cobertura requerida para este método é na ordem de 8 e 10 vezes.

2.4.2.
Next-Generation Sequencing (NGS)

A característica que tornou a NGS altamente popular está relacionada ao seu custo. Quando o genoma humano foi sequenciado pela primeira vez na íntegra, em 2003, isso foi considerado como um feito raro que seria poucas vezes realizado. O custo total do projeto ficou em torno dos 3 bilhões de dólares e os primeiros genomas realizados custaram centenas de milhões de dólares cada [10].

Em 2006, os custos de sequenciamento já tinham declinado substancialmente. Porém, seu declínio estava bem relacionado a lei de Moore [11], caso permanecesse nessa linha de declínio, no ano de 2014 o custo para sequenciar um genoma humano estaria em torno de 1 milhão de dólares. Entretanto, com a forte chegada da NGS ao mercado comercial, o custo para realizar o sequenciamento caiu drasticamente, para menos de 10 mil dólares, como pode ser observado no gráfico da Figura 3.

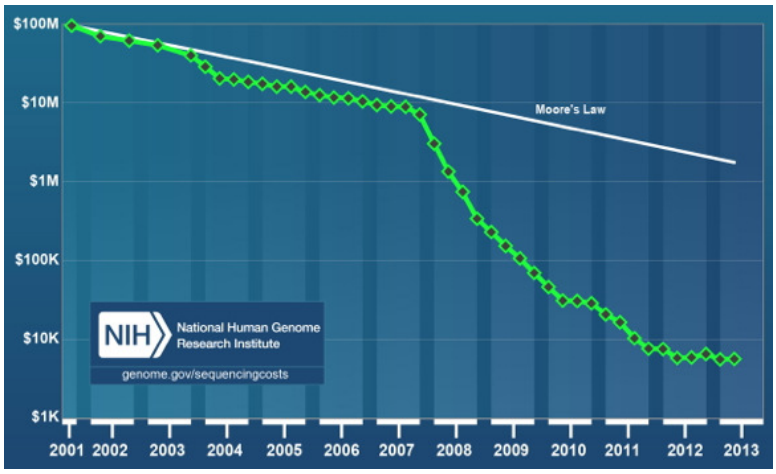

Figura 3 Custos por sequenciamento do genoma até meados de 2013[10]

2.4.2.1. Ilumina (Solexa)

Com base em suas pesquisas acadêmicas em Cambridge, Balasubramanian e Klenerman em 1998 criaram a empresa Solexa para continuar seus trabalhos na área de sequenciamento, em 2007 a empresa foi comprada pela Ilumina.

Este método de sequenciamento funciona em três etapas básicas: amplificação, sequenciamento e análise. O processo inicia com o DNA purificado, que é cortado em pedaços menores e são incluídos adaptadores de dados, índices e outros tipos de modificações moleculares que atuam como pontos de referência durante o processo. O DNA modificado é colocado em um chip onde a amplificação e o sequenciamento é realizado. O fundo do chip possui centenas de milhares de pequenos fragmentos sintéticos (oligonucleótidos), eles são fixados no chip e são capazes de se unir os fragmentos de DNA que têm as sequências complementares. Uma vez que os fragmentos foram ligados, a fase de geração de cluster se inicia. Esta garante que sejam geradas cerca de mil cópias de cada fragmento de DNA. Em seguida, os iniciadores e nucleótidos modificados são inseridos no chip. Estes nucleotídeos têm reversíveis bloqueadores 3' que forçam os iniciadores a adicionar em apenas um nucleótido de cada vez, bem como os marcadores fluorescentes. Depois de cada rodada de síntese, a câmera tira uma foto do chip e um computador determina que base foi acrescentada pelo comprimento de onda da etiqueta fluorescente e grava-a para cada ponto no chip. Após cada rodada, as moléculas não incorporadas são lavadas. Um passo de desproteção química é então utilizado na

remoção do grupo de bloqueio do terminal 3' e o corante num único passo. O processo continua até que a molécula de DNA completa é sequenciada.

O tamanho gerado de cada fragmento fica em torno de 50 a 300bp exigindo desta forma uma cobertura maior que 50 vezes, aumentando de acordo com o nível de qualidade esperada[12].

2.5. Conclusão

Neste capítulo foram apresentados os aspectos gerais da genômica e a necessidade de o sequenciamento ser realizado em fragmentos. Além disso foi exposta a relação existente entre a qualidade e completude do mapeamento do genoma com a quantidade de cobertura requerida. Ademais, foram introduzidas algumas técnicas de sequenciamento relevantes.

No próximo capítulo serão abordados os conceitos da montagem de fragmentos e o problema computacional que envolve este processo. Que de igual forma é necessário para contextualizar esta pesquisa.