Capítulo 3. Plataformas de Sequenciamento de Nova Geração & Pré-processamento de dados

Luciano da Silva Pinto [1], Frederico Schmitt Kremer [1]

[1] *Laboratório de Bioinformática e Proteômica, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas*

Objetivos: Apresentar as principais características das diferentes plataformas de sequenciamento de DNA de nova geração.

Introdução ao sequenciamento de DNA

A partir da descoberta da estrutura do DNA, importantes avanços levaram a compreensão da complexidade e diversidade de genomas. Os primeiros métodos de sequenciamento direto do DNA só foram criados na década de 1970. Os conhecimentos existentes sobre a organização do gene e genoma eram baseados principalmente em estudos de genética reversa, na qual a sequência de aminoácidos do produto do gene de interesse é retrotraduzida em uma sequência de nucleotídeos com base nos códons apropriados. Considerando a característica degenerada do código genético, este processo pode ser complicado e os resultados não corresponderem a realidade.

Os dois primeiros métodos de sequenciamento de DNA foram os de Maxam-Gilbert, conhecido como método de clivagem química e o método de terminação de cadeia de Sanger, tendo este último dominado os trabalhos até meados dos anos 2000. Projetos de sequenciamento, incluindo principalmente o **Projeto Genoma Humano**, propiciaram o desenvolvimento de soluções tecnológicas mais avançadas tanto para a geração dos dados quanto para a análise destes. Estes avanços ajudaram a responder aos novos questionamentos que surgiram, mas as principais barreiras, que eram a produção limitada e os altos custos de sequenciamento, permaneciam.O lançamento da primeira

plataforma de sequenciamento de alto rendimento (eg: *high througput*), o Roche 454, em meados da década de 2000, propiciou uma redução de 50.000 vezes no custo do sequenciamento. A nova geração de sequenciadores de DNA (NGS) continuou a evoluir, aumento a capacidade por um fator de 100-1.000.

Embora seja um grande avanço na forma de se analisar genomas, essas novas abordagens tem suas limitações. À medida que novas tecnologias surgiram os problemas existentes foram exacerbados ou surgirem novos problemas. As novas plataformas apesar de fornecer grandes quantidades de dados possuem taxas de erro associadas mais elevados. Além disso, as leituras são geralmente mais curtas do que o do tradicional sequenciamento de Sanger, exigindo exame mais cuidadoso dos resultados. Cabe salientar que, devido ao grande número de sequencias gerado, a tecnologia de processamento dos dados também teve que evoluir, incluindo a capacidade computacional associada e software.

Em princípio, o conceito subjacente a essa tecnologia se assemelha com o mecanismo de eletroforese através de capilares, onde as bases de um pequeno fragmento de DNA podem ser identificadas sequencialmente a partir de sinais emitidos. No entanto, os métodos mais modernos ao invés de se limitarem a analisar pequenos fragmentos de DNA, passaram a avaliar milhões deles em uma única corrida. Com isso, esse avanço tecnológico permitiu que fosse realizado um sequenciamento mais eficiente, com uma maior cobertura incluindo genomas inteiros através de uma única reação. É importante destacar que o método de terminação em cadeia não deixou de ser utilizado, mas está caindo em desuso com o passar dos anos.

Next-Generation Sequencing - Sequenciadores de nova geração

Método do Pirosequenciamento (ROCHE 454)

Com a necessidade de desenvolver metodologias mais eficientes para o processo de sequenciamento de DNA, indústrias farmacêuticas, como a Roche e a Applied Biosystems, iniciaram uma corrida para o desenvolvimento de uma nova geração de

sequenciadores. Um destes sequenciadores foi desenvolvido pela Roche e denomina-se 454, utilizando um método conhecido como *pirossequenciamento*.

O pirosequenciamento baseia-se nos seguintes passos:

1. Fragmentação do DNA molde.

2. Ligação de sequências *adaptadoras* nas extremidades de cada fragmento gerado, onde são utilizados dois adaptadores diferentes, um para cada extremidade da fita (5' e 3'). Esse processo pode ser visualizado na figura 1.

3. Separação das fitas de DNA duplo (dsDNA) em uma fita de DNA simples (ssDNA).

4. Ligação das sequencias adaptadoras das sequencias em ***nanoesferas***, denominadas *beads*. Estas *nanoesferas* possuem sequências que pareiam com as sequência adaptadoras, sendo que, estatisticamente, espera-se que apenas uma sequência se ligue a cada *bead*. Um modelo de nanoesfera pode ser visualizado na figura 2.

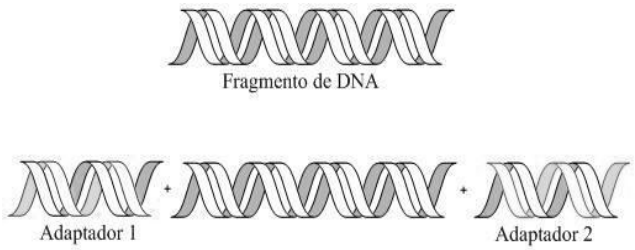

Figura 1. Representação dos adaptadores ligados a cada extremidade das fitas simples de DNA

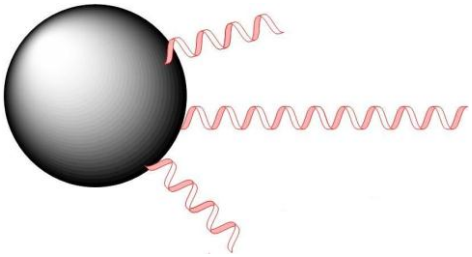

Figura 2. Representação de uma bead (preto) e de seus adaptadores (vermelho).

O próximo passo após ocorrer a ligação do DNA às beads, é a amplificação. No caso do pirosequenciamento, essa amplificação será realizada em uma solução oleosa emulsificada, denominada PCR em emulsão. Desse modo, as *beads* serão internalizadas por gotículas de óleo e a única fita de DNA ligada a nanoesfera dará origem a várias outras fitas, que também se ligarão a essas nanoesferas. Desta forma, cada *bead* ficará coberta por milhares de cópias de uma única sequencia de DNA.

Realizado o processo de amplificação, a solução é transferida para uma placa contendo milhões de *nanopoços,* que possuem diâmetros suficientes para apenas uma *bead*. Após isso, os reagentes da reação de sequenciamento são adicionados, sendo os reagentes utilizados, neste caso, as enzimas *DNA polimerase*, *ATP sulfurilase*, *luciferase* e *apirase* e o substrato *luciferina*.

O sequenciador adicionará um nucleotídeo (dNTP) por vez, que será utilizado no lugar do ATP pelo fato de ser o substrato da enzima luciferase. A incorporação de um nucleotídeo liberará um *pirofosfato,* que será convertido em ATP pela ação da enzima ATP sulfurilase, na presença de dATPαS. O ATP gerado será utilizado pela enzima *luciferase* para converter a *luciferina* em *oxiluciferina*, um processo que liberará luz. Caso um nucleotídeo adicionado pelo seqüenciador não seja incorporado, a enzima aspirase o degradará, garantindo a limpeza do sistema e permitindo que um novo nucleotídeo seja adicionado à placa. Nas figuras 2 e 3 podem ser vistas ilustrações deste processo.

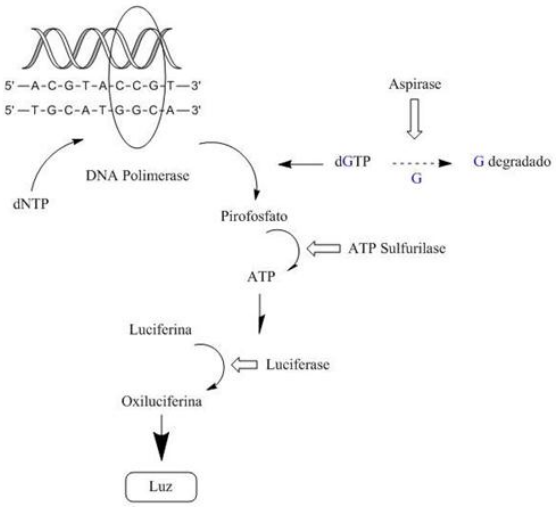

Figura. 5. Esquema do fluxo entre enzimas e substratos durante uma reação de pirosequenciamento.

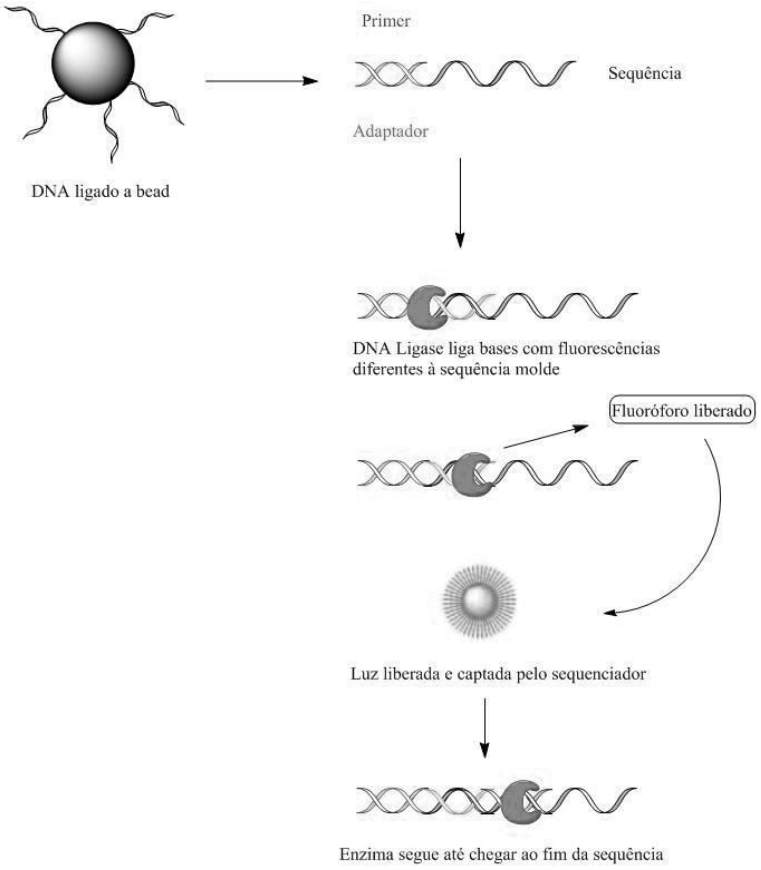

Figura 3. Esquema simplificado representando o Método de Pirosequenciamento.

Em cada nanopoço está ocorrendo milhares de reações ao mesmo tempo, sendo que uma câmera CCD é responsável por filmar a placa e enviar todos os dados para um computador. Chegando ao computador, poderá ser realizada a identificação da sequência de cada nanopoço, através da análise dos picos luminosos gerados pela adição de cada nucleotídeo.

O pirosequenciamento é considerado um método mais eficiente que o método de Sanger se considerarmos que a cobertura da sequencia é muito maior, sendo uma determinada base lida várias vezes. Entretanto, o método tem como inconveniente a necessidade de utilizar um número elevado de reagentes.

Plataforma Illumina/Solexa

Essa plataforma realiza seu sequenciamento, assim como no método de Sanger, pela síntese através da enzima DNA polimerase e de nucleotídeos terminadores marcados

com diferentes fluoróforos. A diferença inovadora consiste no fato da execução do método de clonagem dos fragmentos in vitro ser realizada em uma plataforma sólida de vidro (PCR de fase sólida).

Primeiramente, realiza-se a fragmentação de forma aleatória e, esses fragmentos serão, posteriormente, ligados a adaptadores localizados nas duas extremidades. A adesão das moléculas de DNA fita simples ocorre por afinidade ao suporte sólido, local onde se encontram os oligonucleotídeos complementares a esses adaptadores. Na fase de anelamento, o adaptador da extremidade da molécula que se encontra livre, se une ao seu oligonucleotideo complementar no suporte e forma uma estrutura em ponte, logo dá-se início a PCR, fazendo uso da extremidade 3' livre do oligonucleotídeo como primer. Já durante a etapa de desnaturação, essa ponte é desfeita através do uso de altas temperaturas. Posteriormente, a etapa de anelamento é repetida, dando origem a novas estruturas em ponte, iniciando um novo ciclo de amplificação. É importante destacar que nesta plataforma podem ser gerados fragmentos tanto paired-end quanto single-read para a preparação de bibliotecas de DNA.

No momento em que ocorrer uma quantidade suficiente desses ciclos, serão obtidos clusters de moléculas idênticas ligadas ao suporte e, através da incorporação de nucleotídeos terminadores marcados e da excitação a laser, é gerado sinal, que será captado por um aparelho que realizará a leitura e a interpretação dos possíveis nucleotídeos componentes da cadeia. O procedimento de leitura é feito de forma sequencial, o que permite a montagem da sequência completa de cada um dos clusters gerados.

Método Applied Biosystems SOLiD™ 4

O método empregado para o sequenciamento através do SOLiD, também faz uso de fragmentos de DNA ligados a beads, no entanto possui alguns diferenciais, como:

1. O uso da enzima ligase ao invés da polimerase;
2. A leitura de duas bases a cada sinal luminoso;

3. O fato de não utilizar nanopoços, fazendo com que as beads sejam depositadas aleatoriamente sobre uma lamina de vidro.

A reação de sequenciamento através deste método envolve algumas etapas, que serão abordadas de forma simplificada neste capítulo. Uma das etapas consiste na construção de bibliotecas, onde ocorre a adição de primers (chamados P1 e P2) em cada uma das extremidades de cada fragmento. Posteriormente, durante o sequenciamento, ocorre a adição de um primer complementar a P1 mais um pool equivalente de probes (ou fragmentos) que se estenderá por toda a extensão do template. Após o anelamento das probes, tem-se a etapa de amplificação, onde a enzima DNA ligase, a partir de um primer universal, atua unindo as sondas, onde as duas bases específicas pareiam com o fragmento de DNA ligado a bead (sequência molde). No momento da ligação através da enzima ligase, o fluoróforo é liberado emitindo um sinal luminoso, que, posteriormente, será capturado pelo sistema óptico do sequenciador. As três bases degeneradas permanecem ao lado das duas bases específicas e, assim, a enzima ligase segue atuando até chegar ao final da sequência molde.

Após a fixação pela ligase, ocorre uma espécie de lavagem onde todas as probes que não se aderiram são removidas. A partir desse momento, um novo primer universal contendo uma base a menos é ligada à sequencia molde e todo processo de amplificação é repetido. Essa troca ocorre cinco vezes, pois este passo é necessário para que todas as bases da sequência molde sejam lidas.

Para a obtenção dos resultados, tem-se que cada sinal fluorescente liberado com o processo de ligação representará a leitura de duas bases para cada cor e, no final desse processo, um arquivo com as cores lidas (csfasta) e um arquivo de qualidade Phred são gerados, contendo as informações de cada di-base. Nas imagem a seguir é possível visualizar o processo.

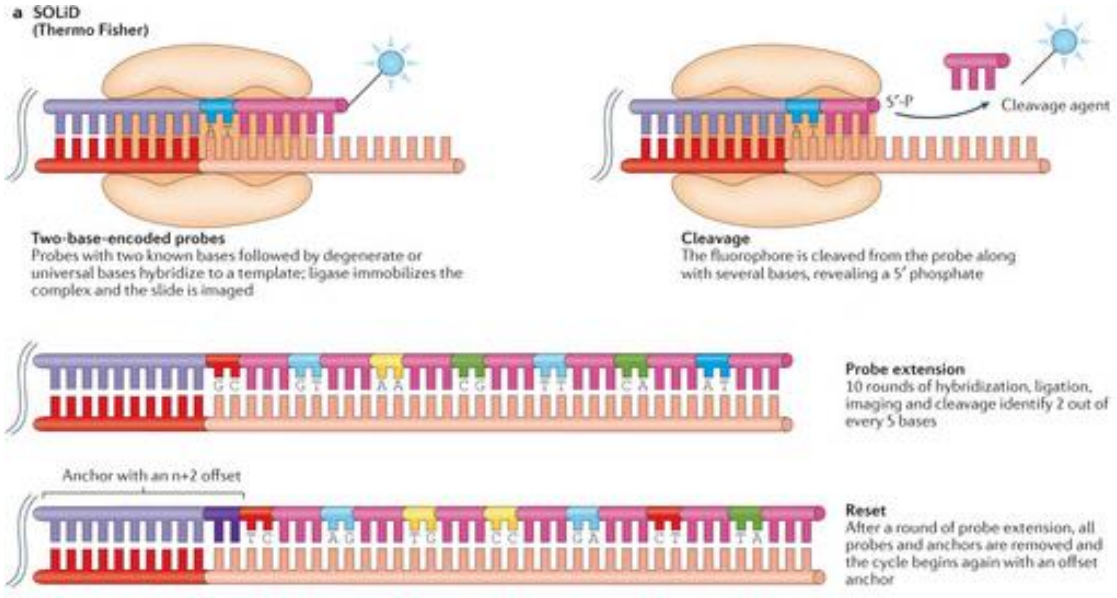

Figura 5. Sequenciamento por SOLiD™

Método Life Technologies Ion Torrent™

A plataforma de Sequenciamento Ion Torrent desenvolvida pela *Life Technologies* (Figura 6) possui uma abordagem de seqüenciamento de DNA diferenciada, visto que a identificação das bases se dá unicamente por pH, e não por ddNTPs ou reações luminosas, como era o caso das tecnologias vistas até então.

Figura 6. Imagem do sequenciador Ion Torrent PGM.

Nessa plataforma, a amostra de interesse é colocada em um pequeno chip (Figura 7), que contém em sua superfície nanoporos providos de medidores de pH, em nanoescala.

Após ser colocado no chip, o DNA sofre uma reação irreversível de ligação a esses nanoporos.

Figura 7. Um chip de sequenciamento da plataformas Ion Torrent PGM.

Com o DNA ligado é iniciado o processo de sequenciamento. Da mesma forma que no pirosequenciamento, no IonTorrent as bases serão adicionadas uma de cada vez e um ciclo será repetido milhares de vezes. Caso a base correta seja adicionada, a enzima DNA polimerase agirá e um íon H+ será liberado, provocando um aumento do pH no nanoporo. Esse aumento será detectado e, dessa forma, a sequência presente em cada nanoporo será determinada.

A plataforma Ion é interessante, pois faz uso de um método simples e relativamente barato, por esse motivo, ela vem sendo cada vez mais empregada em alguns países da Europa e nos EUA, em procedimentos de sequenciamento de genomas, em clínicas, com o objetivo de análises genéticas para auxiliar o diagnóstico médico.

Pacific Bioscience - PacBio (SMRT®)

PacBio é um sistema baseado em uma nova tecnologia de sequenciamento de molécula única em tempo real – Single-Molecule, Real Time Technology Sequencing (SMRT®). Essa tecnologia possibilita a observação em tempo real da síntese de DNA através da enzima DNA polimerase, garantindo a esse método uma precisão superior a 99,999%, independentemente do conteúdo CG. Além disso, permite que esse método possua uma faixa de leitura superior a qualquer outra tecnologia, com média de aproximadamente 8000 pares de bases e que todo o processo seja feito de forma rápida e simples.

O uso da tecnologia SMRT, permite que o sequenciamento ocorra em células, cada uma contendo 150.000 “poços” Zero-Mode Waveguides (ZMWs). Cada um desses ZMWs é iluminado na parte inferior por um laser capaz de detectar a adição de bases, por uma enzima, marcadas com fluorescência à um DNA molde.

Para o procedimento de preparo das amostras, a biblioteca do DNA molde é composta por fragmentos de DNA dupla fita conectados à adaptadores em suas terminações. Esses fragmentos com adaptadores são chamados de ***SMRTbells™***. Estes adaptadores serão capazes de transformar os fragmentos de DNA de fita dupla em moldes circulares, nos quais a enzima polimerase vai continuar a funcionar até que se inative ou até que ocorra o fim do período de observação. É realizada uma corrida com múltiplas passagens em torno desse molde circular, permitindo a condensação numa sequência consenso de maior precisão.

Apesar dessas novas tecnologias de sequenciamento possuírem vantagens como: possibilidade de leitura direta em uma única molécula de DNA, além de poderem fornecer uma visão mais clara da organização do genoma e de seu conteúdo genético, elas possuem como desvantagem estarem suscetíveis a altas taxas de erro. Esse processo pode ser, muitas vezes, decorrendo da leitura errônea de um nucleotídeo (A, T, C ou G) durante a análise de uma cadeia de DNA. Por esse motivo, pesquisas estão sendo realizadas visando a otimização desse método de sequenciamento e sua possível utilização futura.

Pré-processamento de dados de NGS

Apesar de cada plataforma de sequenciamento ter suas próprias características quanto ao método que é utilizada para a identificação das bases, o tamanho dos fragmentos que são sequenciamento e os tipos de bibliotecas são são suportados, a forma como os seus resultados são apresentados muitas vezes são similares. Desta forma, existem determinados formatos de arquivos de proprósito geral que podem ser utilizadas para armazenar informações provinentes de diferentes plataformas, por exemplo.

Formatos de arquivo

Um **formato de arquivo** é uma forma estrutura e padronizada de armazenar informações, pode este ser na forma de **texto** (entendível por humanos, normalmente com tamanho maior por ser pouco compactado) ou **binário** (entendível por máquinas, normalmente menor por passar com compactação). Para a representação de dados de sequenciamento, os principais arquivos são: FASTA, FASTA + QUAL, SFF, csFASTA (+ QUAL), FASTQ e BAM.

FASTA: O formato FASTA foi inicialmente proposto como padrão para o pacote de alinamento de sequencias FASTA, "antecessor espiritual" do atual pacote BLAST. Cada sequenciamento armezenada em um arquivo FASTA possui dos campos: cabeçalho e sequencia. O cabeça é a primeira linha de qualquer sequenciamento, e deve começar com o símbolo ">", seguido de uma **identificação única** e uma **descrição**. Todas as demais linhas após o cabeçalho serão interpretadas como referentes à sequência até que um novo cabeçalho seja identificado. Um exemplo de conteúdo de arquivo FASTA é:

Arquivo FASTA

```
>Seq0001 fragmento sequenciado 25/08/2016
GCATCTAGCAGCCGCGCGATCACGATCGCTATCACGATCAGCTACGATCGATATATAC
```

FASTA + QUAL: O formato FASTA armazena apenas a sequencia de um fragmento ou mais fragmentos sequenciados, mas muitas vezes ao trabalhar com dados de sequenciamento é importante se ter também a informação da **qualidade** de cada base identificada, que indica o seu grau de confiabilidade. Desta forma, no começo dos projetos de sequenciamento um segundo formato de arquivo foi criado para armazenar este tipo de informação, tendo este a extensão ".qual".

Cada sequencia em um arquivo ".fasta" é acompanhada de seus respectivos valores de qualidade no arquivo ".qual", sendo usados os mesmos cabeçalhos e a mesma ordem dos fragmentos. Entretanto, ao invés de possuir um campo referente à sequencias, o ".qual" possui valores de qualidade no formato Phred (próxima sessão), com tamanho fixo de 2 dígitos, separados por espaços simples.

Arquivo FASTA

```
>Seq0001 fragmento sequenciado 25/08/2016
GCATCTAGCAGCCGCGCGATCACGATCGCTATCACGATCAGCTACGATCGATATATAC
```

Arquivo QUAL

```
>Seq0001 fragmento sequenciado 25/08/2016
20 20 20 20 20 20 21 19 18 20 20 20 20 20 20 21 19 18 20 20 20
20 20 20 21 19 18 20 20 20 20 20 20 21 19 18 20 20 20 20 20 20
21 19 18 20 20 20 20 20 20 21 19 18 20 20 20 20
```

SFF: O formato SFF é um formato binário utilizado como padrão pelas plataformas de sequenciamento da família Roche 454 e nas primeiras versões do Ion Torrent PGM. Além das informações a cada base identificada e seus respectivos valores de qualidade, o formato SFF também armazena diversos metadados que podem ser utilizados pelos programas de análise para a geração de resultado mais otimizado, o que inclui a minimização do efeito negativo de possíveis artefatos de sequenciamento.

csFASTA (+QUAL): O formato csFASTA (color-space FASTA) é utilizados pelos sequenciadores da família ABI SOLiD como formato de saída padrão, mas hoje é poucos software oferecem suporte nativo para ele, sendo necessário muitas vezes se converter este para outro formato mais usual (ex: FASTQ). A estrutura de um arquivo csFASTA é similar a um arquivo FASTA comum, mas a sequencia é representada como uma sequência de códigos de cores (1, 2, 3 ou 4), que indicam diferentes conversões entre bases em relação a base anterior, sendo este formato está intimamente relacionada a forma como o SOLiD realiza o sequenciamento. Para que seja possível se identificar a sequencia de cada fragmento, a primeira base identificada nas sequencias de um

csFASTA é sempre constante, sendo a ultima base do adaptador adiciono durante construção da biblioteca.

Arquivo csFASTA

```
>Seq0001 fragmento sequenciado 25/08/2016
G112434142324424242142434142341423414234142
>Seq0002 fragmento sequenciado 25/08/2016
G112424341424224241424341243424124241424241
```

Os arquivos csFASTA, da mesma forma que os em formato FASTA convencional, podem também ser acompanhados de um arquivo QUAL.

FASTQ: O formato FASTQ foi desenvolvido com o objetivo de se unificar as informações de sequencia e qualidade em um único arquivo de texto, e ao mesmo tempo diminuir o espaço em disco ocupado por estas informações, sendo proposto (e posteriormente adotado como) como o formato padrão *de facto* para dados de leituras de sequenciamento de nova geração. Para reduzir o espaço ocupado pelos valores de qualidade, estes são representados sequencialmente sem a adição de espaços (em contraste com o formato QUAL), e condificados em caracteres do alfabeto ASCII de forma que apenas um caracter seja utilizada para cada valor de qualidade.

Da mesma forma que um arquivo FASTA, os arquivos FASTQ também possuem seus cabeçalhos, que iniciam sempre com um símbolo "@" para indicar uma nova sequência, que é sempre acompanhada de dados de identificação do fragmento em questão. Após o cabeçalho, as linhas seguinte são utilizada para indicar a sequencia propriamente dita, apesar de raramente se utilizar mais de uma linha para este tipo de informação. Após a sequencia, uma nova linha iniciada com um símbolo "+" é utilizada para indicar o começo dos dados de qualidade, podendo esta ser acompanhada ou não dos mesmos dados de cabeçalhos utilizadas para a sequencia. Por fim, os dados de qualidade das bases são apresentados codificados em caracteres ASCII, podendo esta codificação ser utilizando o intervalo 33 – 126 ("Phred 33"), 59 – 126 ("Phred Solexa") ou 64 – 126 ("Phred 64"), de acordo com a plataforma de sequenciamento.

Formato FASTQ

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

BAM: O formato BAM não foi desenvolvido com o propósito de se armazenar leituras de sequenciamento, mas sim dados de alinhamento. Entretanto, em suas versões mais atuais as plataformas IonTorrent tem utilizado este formato como alternativa ao formato SFF. Sendo um formato binário derivado da compactação de arquivos no formato SAM, o BAM possibilita o armazenamento de todos os dados representaveis em formato FASTQ além de meta-informações sobre a corrida do sequenciamento.

Phred score

O algoritmo Phred foi desenvolvida ainda no começo da era do sequenciamento de Sanger automática, na década de 90, para se converter dados brutos de sequenciamento (cromatogramas), e suas respectivas sequencias. Este processo, denominado ***base calling***, e varia de acordo com a plataforma de sequenciamento em questão. Além da sequência propriamente dita, o Phred também calcula um grau de confiabilidade para cada base identificada, com o base na qualidade do sinal que foi captada, sendo este valor denominado Phred score. Apesar de cada plataforma de sequenciamento de nova geração ter suas próprias abordagens de base calling, o nome *Phred score*, e seu cálculo base foram mantidos.

Tendo-se um uma probabilidade de erro (***P***) para uma determinada base, o Phred score (***Q***) é calculado através da equação:

$$Q = -10 \log_{10} P$$

A relação entre probabilidade de erro e Phred score está exemplificada abaixo.

Phred Score (Q)	Confiabilidade da base (%)	Probabilidade de erro (%)
10	90	10
20	99	1
30	99,9	0,1
40*	99,99	0,001
50	99,999	0,0001
60	99,9999	0,00001

* = Valores acima de 40 não são esperados em dados brutos de sequenciamento, sendo obtidos quase que unicamente após a sobreposição de leituras com alta cobertura, qualidadede leituras.

Conversão de arquivos

Conversão de arquivos BAM para FASTQ: Os arquivos BAM gerados pelas plataformas da família Ion Torrent podem ser convertidos para o formato FASTQ com a ferramenta `bamToFastq` disponível no pacote `bedtools`. É possível se instalar este pacote em um sistema operacional Linux Ubuntu através do gerenciar APT com uso da seguinte linha de comando:

```
$ sudo apt-get install bedtools
```

Para a realizer o processo de conversão, basta se informar o nome do arquivo de entrada, através do argumento `-i`, e o nome do arquivo de saída, através do arguimento `-fq`. No caso de arquivos contendo leituras pareadas (paired-end ou mate-pair) é possível também se utilizar um terceiro argumento, `-fq2`, para indicar o segundo arquivo onde serão salvas as leituras irmãs.

Connversão de leituras simples (*single-end*)

```
$ sudo bamToFastq -i leituras.bam -fq leituras.fastq
```

Connversão de leituras pareadas (*paired-end* ou *mate-pairs*)

```
$ bamToFastq -i leituras.bam -fq leituras_1.fastq -fq2
leituras_2.fastq
```

Conversão de arquivos FASTQ para FASTA: Em algumas situações pode ser necessário se analisar arquivos de sequenciamento com ferramentas que não oferecem suporte ao formato FASTQ, mas sim ao formato FASTA. É possível se fazer a conversão entre estes formatos através da ferramenta `fastq_to_fasta` disponível no pacote `fastx-toolkit`.

O pacote `fastx-toolkit` pode ser instalado através do gerenciador do pacotes APT, com a linha de comandos:

```
$ sudo apt-get install fastx-toolkit
```

O programa `fastq_to_fasta` recebe através do argumento `-i` o nome do arquivo FASTQ que será convertido e através do argumento `-o` o nome do arquivo que será gerado.

```
$ fastq_to_fasta -i leituras.fastq -o leituras.fasta
```

Análise de qualidade

Para se avaliar a distribuição dos valores de qualidade nas leituras que foram sequenciadas é possível se utilizar o programa FastQC, que é particularmente útil quando os dados a serem processados foram obtidos por meio das plataformas Illumina. O programa por ser obtido no site www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Obs: Para executar o FastQC é necessário se ter o Java instalado. Caso você não tenha, utilizado o comando:

```
$ sudo apt-get install default-jre
```

Após baixar o FastQC e descompactar, basta entrar na pasta e digitar:

```
$ perl fastqc
```

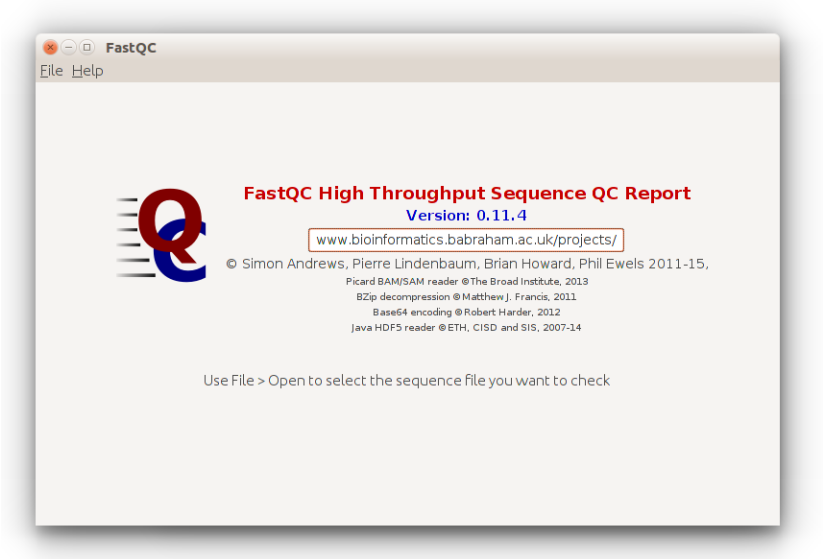

Figura 9. Tela inicial do FastQC.

Para carregar os arquivos de leitura, vá em `File / Open`, e selecione os arquivos de interesse.

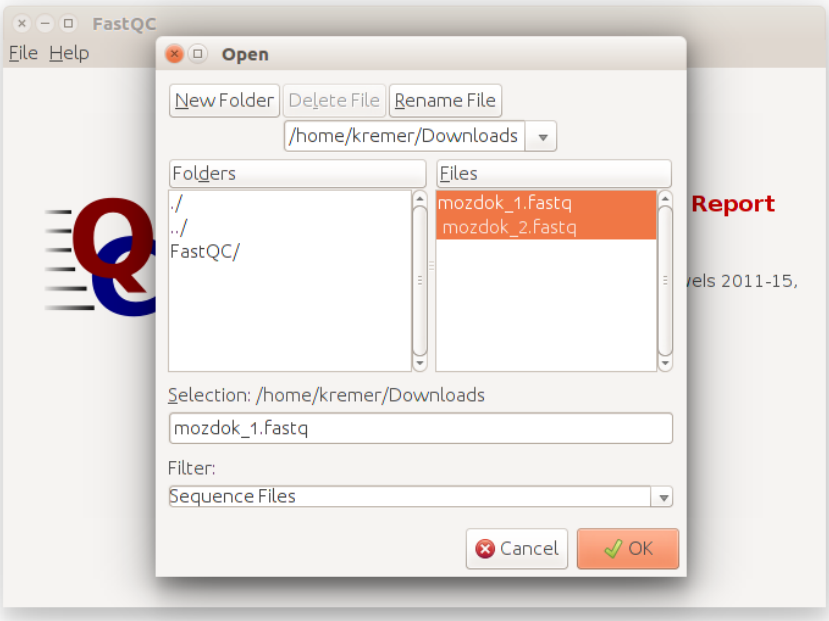

Figura 10. Selecionando os arquivos FASTQ no FastQC.

Para cada arquivo carregado o FastQC gerá diferentes análise, incluindo uma mapa de distribuição de qualidade Phred ao longo das leituras. O padrão de decaimento da qualidade nas regiões mais próximas do 3' observado na Figura 11 é característica de leituras obtidos a partir de plataformas da Illumina. Nesta imagem, um *threshold* de Phred 20 é usado para separar as regiões de baixa e alta confiabilidade.

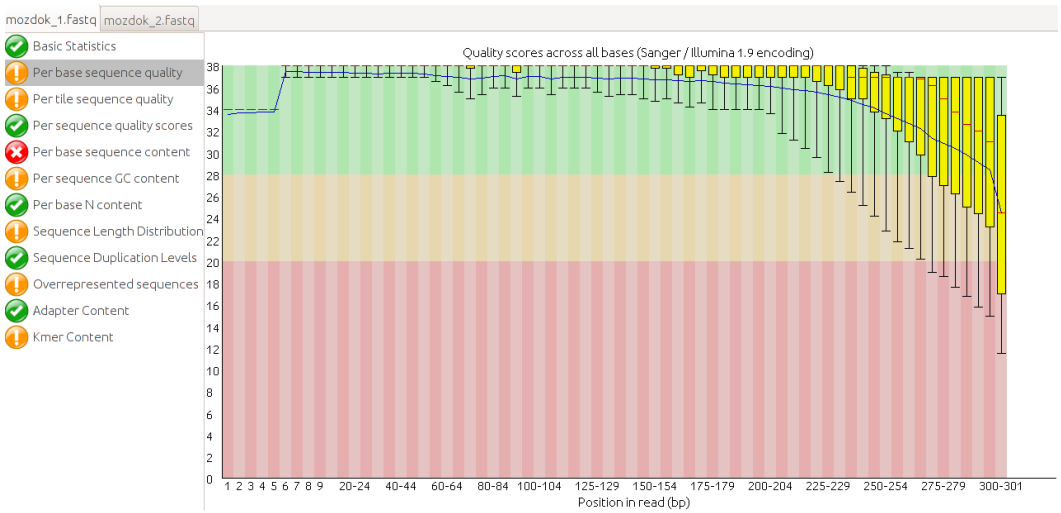

Figura 11. Representação da distribuição de qualidade Phred gerada pelo FastQC.

Filtragem de leituras com o pacote Fastx-toolkit (plataformas: Todas)

É possível se remover leituras com valores de Phred score baixo para que estas não interfiram na qualidade das análises posteriores. O pacote `fastx-toolkit` possui a ferramenta `fastq_quality_filter`, que serve para remover leituras que não apresentam uma porcentagem mínima de bases com valor de Phred maior que um determinado limiar. O arquivo de entrada é definido pelo argumento `-i`, o de saída pelo argumento `-o`, o limiar de qualidade pelo argumento `-q` e a porcentagem mínima de bases com esta pontuação ou mais pelo argumento `-p`. Exemplo:

```
$ fastq_quality_filter -i reads.fastq -o reads.fastq -p 90 -q 20
```

Neste caso, apenas as leituras que apresentam pelo menos 90% das bases com qualidade igual ou superior a Phred 20 são mantidas.

Filtragem / trimmagem de leituras com o Trimmomatic (plataformas: Illumina)

O Trimmomatic é uma ferramenta de desenvolvida especificamente para a manipulação de dados da Illumina, que inclui uma série de pequenas funções de filtragem, trimmagem (remoção de bases da extremidade) e remoção de adaptadores que podem ser integradas e executadas ao mesmo tempo. O programa também possui suporte para leituras *paired-end*, muito utilizadas nesta plataforma, o que facilita em muito algumas análises. Para instalar esta ferramenta basta se utilizar a linha de comando:

```
$ sudo apt-get install trimmomatic
```

Um ferramenta bastante útil do Trimmomatic é o método de trimagem baseado e *sliding window*, que remove regiões (de um tamanho estabelecido pelo usuário) na extremidade 3' das leituras quando a qualidade média destas está abaixo de um *threshold* estabelecido pelo usuário. Considerando dados *single-end*, é possível executar este algoritmo da seguinte forma:

```
$ TrimmomaticSE -phred33 mozdok.fastq Mozdok.out.fastq \
SLIDINGWINDOW:5:15
```

Neste caso, regiões de tamanho 5 com qualidade média menor que Phred 15 são removidas. Este mesmo algoritmo pode ser aplicado a leituras *paired-end* da seguinte forma:

```
$ TrimmomaticPE -phred33 mozdok_1.fastq mozdok_2.fastq \
mozdok_1.out.fastq mozdok_1.unpaired.fastq mozdok_2.out.fastq \
mozdok_2.unpaired.fastq SLIDINGWINDOW:5:15
```

Nesta linha de comandos, o –phred33 indica o formato de codificação do arquivo FASTQ; os 2 primeiros caminhos de arquivo ("mozdok_1.fastq" e "mozdok_2.fastq"), indicam os arquivos de entrada; os arquivos com final ".out.fastq" indica os arquivos finais de filtragem, enquanto os com o final "unpaired.fastq" consiste naquelas leituras sujas respectivas leituras-irmãs não foram matidas após a filtragem, sendo muitas vezes chamadas "leituras órfãs".

Além da trimagem baseada em *sliding window*, o Trimmomatic também oferece várias outras ações para filtragem das leituras, como por exemplo:

- Remover nucleotídeos com qualidade menor que um determindo *threshold* (Ex: 15) a partir da porção 5': `LEADING:15`.

- Remover nucleotídeos com qualidade menor que um determindo *threshold* (Ex: 15) a partir da porção 3': `TRAILING:15`.

- Remover leituras com tamanho menor que um determinado threshold (Ex: 30 pb): `MINLEN:30`.