Teste de Qualidade

de um sequenciamento

Clovis Ferreira dos Reis
PhD Bioinformática

Autor: Clovis F. Reis
E-mail: cfreis230@gmail.com

Objetivo: Fornecer noções básicas sobre testes de qualidade a serem realizados em uma amostra logo após o seu sequenciamento.

Sumário

1 Introdução

Todo sequenciamento é um processo complexo e bastante sensível, onde um sofisticado equipamento procura identificar a ordem correta em que nucleotídeos aparecem em uma amostra de DNA ou cDNA, tudo isto ocorrendo em uma escala molecular. Neste processo, erros na correta identificação de bases podem vir a acontecer pelas mais diversas razões, desde a simples contaminação e erros na preparação de amostras e corridas, falha na remoção de adaptadores, até erros intrínsecos do próprio equipamento de sequenciamento.

É importante que, antes de se iniciar qualquer outro trabalho com as amostras sequenciadas, estas passem por um processo de controle de qualidade onde tais incorreções podem ser detectadas e, muitas vezes, corrigidas. Assim, antes de mais nada é preciso compreender como é possível avaliar tal qualidade.

1.1 O formato FASTQ

Após o sequenciamento, os dados precisam ser armazenados em disco e, para isto, utilizam-se arquivos de formato padronizado, normalmente um arquivo do tipo FASTQ. O formato do arquivo FASTQ é baseado no formato FASTA, criado como padrão de saída de um antigo programa de comparação de sequências de proteínas e nucleotídeos de mesmo nome [1]. Nele letras eram utilizadas para representar sequências nucleotídicas ou peptídicas em um código de letra única, ao qual o formato FASTQ adicionou informações de qualidade.

No Quadro 1 temos um exemplo de como uma única sequência (ou *read*) é representada em um arquivo FASTQ, composto por quatro linhas distintas.

- A primeira linha contém o identificador da sequência, iniciado pelo símbolo "@". Ela contém informações diversas sobre o sequenciamento, como equipamento usado, número da corrida, trilha, identificação do cluster, dentre outras. As informações contidas nesta linha podem variar bastante, dependendo do equipamento e versão do software utilizado. Para maiores informações, consulte a documentação do fabricante do sequenciador ou a documentação da Illumina [2];
- A segunda linha contém a sequência de bases de nucleotídeos propriamente dita, identificadas pelas letras ***A,C,G*** e ***T***. Um ***N*** normalmente assinala posições onde não foi possível realizar o *base call* e a base não foi identificada.

  No caso do exemplo do Quadro 1, cada *read* possui um comprimento de 100 pares de bases (bp);
- A terceira linha inicia-se com o sinal "+" e pode conter uma repetição do código existente na primeira linha. Nas versões mais modernas dos sequenciadores Illumina esta linha contém apenas o sinal "+";
- É na quarta linha que se encontram as informações de qualidade codificadas em um único caractere ASCII para cada posição. Desta forma, a linha 4 do nosso exemplo também possuirá 100 escores de qualidade, um para cada base identificada na linha 1, assinalados em ***Phred+33***.

Quadro 1 - Exemplo de sequência FASTQ

```
@HWUSI-EAS209_0003:2:1:999:20161#0/1
NCAATAAAAACAGGTCCTTCGGCTACAATCAGTGTGCATTAGTGTGCATTAATACAATCTTCTCGTTTTTTCTCGTTTTTGATTAAAGTCCACGTTTTCT
+HWUSI-EAS209_0003:2:1:999:20161#0/1
BTRQQYVRLLfefffefffcddddddddddeceeeffffffefffefffcddda]aYaaaf`fffdceeeeceeeffffffdddddaa^aZfcfeeffced
```

1.2 O indicador de qualidade Q

O *Phred quality score* (Q) [3] é um indicador de qualidade baseado na probabilidade de erro na identificação de uma base em uma determinada posição do *read*, e pode ser definido pela equação 1

$$Q = -10\ log_{10}P \tag{1}$$

onde P é a probabilidade de erro na chamada de uma determinada base.

Na prática, podemos dizer que ele define a acurácia na chamada da base. Uma acurácia de 99% tem a probabilidade de ocorrência de um erro em cada 100 leituras (0.1), logo, aplicando-se a fórmula, tem-se um índice de qualidade Q igual à 10. Uma acurácia de 99,9% terá um erro em cada 1000 chamadas (0.01) e um Q de 20, e assim sucessivamente.

Como Q é um valor numérico representado por vários dígitos optou-se por codificá-lo de modo que cada Q ocupasse o espaço de apenas um caractere no arquivo FASTQ, de forma a haver uma correspondência exata entre as posições das bases e seus escores de qualidade. Assim somou-se 33 a cada valor de Q [i] e, utilizando-se a tabela de caracteres ASCII, converteu-se cada número em uma letra ou símbolo. A tabela 1 mostra os 50 primeiros valores de Q e os respectivos valores de P, caracteres e códigos ASCII.

Q	P	ASCII	Q	P	ASCII	Q	P	ASCII	Q	P	ASCII	Q	P	ASCII
0	1.00000	! (33)	10	0.10000	+ (43)	20	0.01000	5 (53)	30	0.00100	? (63)	40	0.00010	I (73)
1	0.79433	” (34)	11	0.07943	, (44)	21	0.00794	6 (54)	31	0.00079	@ (64)	41	0.00008	J (74)
2	0.63096	# (35)	12	0.06310	- (45)	22	0.00631	7 (55)	32	0.00063	A (65)	42	0.00006	K (75)
3	0.50119	$ (36)	13	0.05012	. (46)	23	0.00501	8 (56)	33	0.00050	B (66)	43	0.00005	L (76)
4	0.39811	% (37)	14	0.03981	/ (47)	24	0.00398	9 (57)	34	0.00040	C (67)	44	0.00004	M (77)
5	0.31623	& (38)	15	0.03162	0 (48)	25	0.00316	: (58)	35	0.00032	D (68)	45	0.00003	N (78)
6	0.25119	’ (39)	16	0.02512	1 (49)	26	0.00251	; (59)	36	0.00025	E (69)	46	0.00003	O (79)
7	0.19953	((40)	17	0.01995	2 (50)	27	0.00200	< (60)	37	0.00020	F (70)	47	0.00002	P (80)
8	0.15849	) (41)	18	0.01585	3 (51)	28	0.00158	= (61)	38	0.00016	G (71)	48	0.00002	Q (81)
9	0.12589	* (42)	19	0.01259	4 (52)	29	0.00126	> (62)	39	0.00013	H (72)	49	0.00001	R (82)

Tabela 1: Alguns valores de Q, P e seus respectivos códigos ASCII

Analisando a sequência contida no exemplo do Quadro 1, podemos ver que ela possui qualidade de chamada de bases muito boa, já que o menor escore para uma base identificada é um “L” [ii] ($Q = 43$), que corresponde a uma acurácia aproximada de 0.00005, ou uma base incorreta a cada 20.000 bases chamadas.

Quando utilizamos dados provenientes de NGS, não é viável realizar uma checagem manual dos dados de qualidade, uma vez que o volume de informação gerada é enorme. Além disso, tais dados de qualidade estimam apenas os erros inerentes à dificuldade que o sequenciador teve em identificar corretamente uma determinada base no momento do sequenciamento, não avaliando possíveis erros cometidos na preparação das amostras ou em outra fase do processo. Assim, o uso de software especializado é fundamental para a realização de um controle de qualidade efetivo.

O aplicativo que veremos a seguir é capaz de aferir a qualidade da leitura das amostras feita pelo sequenciador, bem como identificar outros problemas que também tem potencial para comprometer nossa análise final. Seu nome é **FastQC**.

2 Instalação dos softwares

Os procedimentos de instalação abaixo foram testados na distribuição Fedora Workstation versão 31 do Linux. Para outras distribuições talvez haja a necessidade de adaptação de alguns dos caminhos, especialmente no caso do MultiQC. Não abordaremos aqui a instalação envolvendo outros sistemas operacionais pelas razões já descritas em um pequeno artigo meu [4].

Como passo inicial, é interessante criar diretórios específicos para o download e instalação de programas de bioinformática e você pode fazê-lo utilizando os comandos abaixo[iii]:

[i] Dando origem ao nome Phred+33

[ii] As letras minusculas e símbolos que aparecem como Q *score* possuem valores de qualidade superior a 49 e foram omitidos na Tabela 1

[iii] Um *shell script* contendo todos os comandos para instalação, download de amostras e suas análises aqui descritos encontra-se disponível no site da no Github `https://github.com/cfreis/LacenPB/raw/master/TesteQualidade/testeQual.sh`

Quadro 2 - Criação de diretórios (se for o caso)

```
01 [user]$  cd
02 [user]$  mkdir -p $HOME/bin/BioInfoTools/Downloads
```

2.1 FastQC (v0.11.9)

O FastQC é uma ferramenta que realiza algumas das principais verificações de controle de qualidade em dados de sequenciamento NGS, podendo trabalhar com arquivos do tipo BAM, SAM ou FASTQ. Ao final da análise ele fornece, de forma gráfica, um panorama geral dos problemas encontrados.

Como ele é uma aplicação java, antes de testá-lo tenha certeza que seu sistema possui um *Java Runtime Environment* (JRE) instalado. Consulte a documentação de sua distribuição Linux para maiores informações sobre a instalação de um JRE.

A seguir você poderá instalar o FastQC, seguindo os procedimentos descritos abaixo:

Quadro 3 - Download e instalação do FastQC

```
01 [user]$  cd $HOME/bin/BioInfoTools/Downloads
02 [user]$  wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.9.zip
03 [user]$  unzip -d ../ fastqc_v0.11.9.zip
04 [user]$  chmod 755 ../FastQC/fastqc
05 [user]$  export PATH=$PATH:$HOME/bin/BioInfoTools/FastQC/
06 [user]$  fastqc --version
FastQC v0.11.9      [a]
```

[a] Para a exportação permanente do caminho da ferramenta execute:
echo 'export PATH=$PATH:$HOME/bin/BioInfoTools/FastQC/' >> $HOME/.bashrc

Maiores informações sobre o produto e manual do FastQC podem ser encontradas em [5].

2.2 MultiQC (versão 1.8)

Para fins de verificação de qualidade, o MultiQC não é obrigatório, já que ele utiliza as informações e análises realizadas pelo FastQC. Porém ele pode ser bastante útil quando processando diversas amostras, já que ele consolida e sumariza informações de diversos relatórios FastQC em uma única saída.

Para a utilização da ferramenta você já deve ter instalado em seu computador o ***Python 3.7***. A forma mais simples de instalar o MultiQC é utilizando o gerenciador de pacotes ***pip*** que também já deve estar pré-instalado em seu sistema. Verifique o procedimento para instalação do ***python, pip*** e demais pacotes necessários na documentação da sua distribuição.

Quadro 4 - Instalação do MultiQC

```
01 [user]$  pip install multiqc --user
02 [user]$  export PATH=$PATH:$HOME/.local/bin
03 [user]$  multiqc --version
multiqc, version 1.8      [a]
```

[a] Verifique na documentação da sua distribuição o local padrão de instalação do *pip* quando utilizando a opção - *-user* e ajuste a linha 2 caso necessário

O manual do MultiQC e informações sobre outras formas de instalação podem ser encontradas em [6].

3 Download de amostras

Para a realização deste minicurso, utilizaremos quatro amostras *toy*, sendo uma amostra de genoma de boa qualidade, uma de genoma de qualidade ruim, uma de RNA-Seq contaminada por dímeros de adaptadores e uma processada por RRBS, todos do tipo *fastq.gz*[iv]. Utilizaremos para download um *wget* simples (Quadro 5).

Quadro 5 - Download dos arquivos de sequenciamento

```
01 [user]$  mkdir -p $HOME/analise/amostras
02 [user]$  cd $HOME/analise/amostras
03 [user]$  wget https://github.com/cfreis/LacenPB/raw/master/TesteQualidade/data/good.fastq.gz
04 [user]$  wget https://github.com/cfreis/LacenPB/raw/master/TesteQualidade/data/bad.fastq.gz
05 [user]$  wget https://github.com/cfreis/LacenPB/raw/master/TesteQualidade/data/RNA-Seq.fastq.gz
06 [user]$  wget https://github.com/cfreis/LacenPB/raw/master/TesteQualidade/data/RRBS.fastq.gz
```

4 Utilizando os softwares

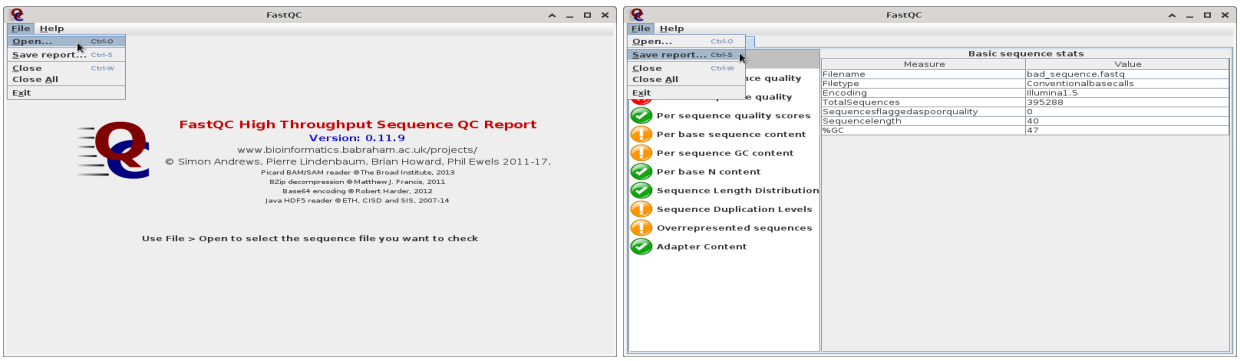
4.1 FastQC

Utilizar o FastQC é muito simples e pode ser feito de dois modos distintos: no modo gráfico e no modo linha de comando. Quando o número de amostras for bastante reduzido, pode-se utilizá-lo no modo gráfico. Para acessá-lo, realize a chamada do programa como mostrado no Quadro 6.

Quadro 6 - Chamada do FastQC no modo gráfico

```
01 [user]$  fastqc
```

Após a exibição da tela inicial (Figura 1a), simplesmente abra o arquivo a ser analisado. O processo inicia-se automaticamente e os resultados são exibidos (Figura 1b). Neste modo de operação os resultados não são salvos por padrão.



(a) Tela inicial e abertura de um arquivo

(b) Exibindo e salvando os resultados da análise

Figura 1: Utilização do FastQC em modo gráfico

Porém, quando se realiza o processamento de diversas amostras, utilizar a interface gráfica não é eficiente. Mesmo porque, em alguns casos, o modo gráfico pode nem estar disponível [4]. Neste caso utiliza-se a linha de comando, o que também não é complicado e nos permite um número maior de opções. Pode-se especificar, por

[iv] Os arquivos *.fastq.gz* são arquivos FASTQ compactados.

exemplo, a utilização do formato *casava* [v], processar saída dos sequenciadores do tipo *nanopore*, especificar listas de adaptadores e contaminantes que deverão ser considerados na análise, dentre outras. Uma opção que se destaca é a escolha do número de *threads* a serem utilizadas, que permitirá a análise de diversas amostras simultaneamente. Para maiores detalhes sobre opções disponíveis e sua utilização, consulte a documentação do FastQC [5].

Desta forma, para realizar o processamento de todas as amostras executamos os comandos constantes no Quadro 7

Quadro 7 - Utilizando o FastQC em linha de comando

```
01 [user]$  mkdir -p $HOME/analise/resultados
02 [user]$  cd $HOME/analise/amostras
03 [user]$  fastqc -t 4 -o ../resultados/ $(ls *.fastq *.fastq.gz)
Started analysis of bad.fastq.gz
Approx 5% complete for bad.fastq.gz
…
Started analysis of good.fastq.gz
Approx 5% complete for good.fastq.gz
…
Started analysis of RNA-Seq.fastq.gz
…
04 [user]$  ls ../resultados/

bad_fastqc.html             good_fastqc.zip             RRBS_fastqc.html
bad_fastqc.zip              RNA-Seq_fastqc.html         RRBS_fastqc.zip
good_fastqc.html            RNA-Seq_fastqc.zip
```

O comando na linha 01 cria um diretório para resultados e o comando da linha 03 realiza a chamada do FastQC:

- A opção *-t 4* inicia quatro *threads* simultâneas, cada uma processando uma amostra distinta. Este número é limitado apenas à quantidade de núcleos e memória [vi] que o seu computador possui;
- A opção *-o ../resultados/* encaminha todos os resultados para o diretório especificado; e
- O comando *$(ls *.fastq *.fastq.gz)* nos poupa o trabalho de escrever cada nome de arquivo individualmente, listando todos os arquivos *.fastq* e *.fastq.gz* disponíveis no diretório.

Ao executar o comando da linha 04 notamos que, para cada arquivo analisado, duas saídas são criadas. O primeiro deles é um arquivo *.html* contendo as informações da análise em formato de página *web*. Já o segundo é um arquivo compactado contendo as mesmas informações do arquivo *.html*, mas sob a forma de tabelas e sumários em modo texto, além de todas as figuras geradas no processamento.

4.2 MultiQC

O MultiQC é um software destinado a consolidar resultados de análise de outras ferramentas. Desta forma, ele não realiza qualquer análise de qualidade *per se*, tampouco realiza a chamada destes softwares.

Sua grande vantagem aparece quando é necessário analisar um número elevado de relatórios de qualidade de amostras similares. Utilizando o FastQC obteremos um relatório de qualidade para cada amostra processada. Isso pode se transformar em um grande inconveniente caso este número seja significativamente elevado. A utilidade do MultiQC reside exatamente em transformar estes N relatórios em apenas um, permitindo a rápida identificação de amostras de baixa qualidade.

Sua utilização é extremamente simples: basta rodar o programa, indicando o caminho onde se encontram os relatórios gerados pelo FastQC. Se os resultados já estiverem todos organizados em um diretório específico, como

[v] O formato *Casava* é utilizado como saída de alguns softwares da Illumina e é, basicamente, um FASTQ normal, onde os dados de uma amostra podem ser divididos em vários arquivos.

[vi] Segundo a documentação do FastQC, serão alocados 250MB de memória para cada *thread* adicional.

é o nosso caso, é só navegar até ele e executar o comando de dentro deste diretório, conforme descrito no Quadro 8.

Quadro 8 - Executando o MultiQC

```
01 [user]$  cd $HOME/analise/resultados/
02 [user]$  mkdir MultiQC
03 [user]$  multiqc -o MultiQC ./
04 [user]$  cd MultiQC
```

A opção *-o MultiQC* indica que os resultados serão armazenados no diretório MultiQC e o *./* indica que serão processados todos os relatórios de amostras existentes no diretório atual. Podemos determinar que MultiQC ignore determinadas amostras pela opção - *-ignore*.Também podemos passar uma lista de relatórios a serem processados com a opção - *-file-list*. Para maiores informações sobre estas e outras opções disponíveis, consulte a documentação online do produto [6].

Examinando o diretório *MultiQC* vemos o arquivo *multiqc_report.html*, contendo o relatório da análise de todas as amostras, e o diretório *multiqc_data* contendo informações complementares sobre o processamento.

5 Interpretação dos resultados

Para poder interpretar corretamente os resultados das análises faz-se necessário revisar alguns conceitos sobre o funcionamento do sequenciamento. A Figura 2 ilustra, de forma simplificada e genérica, os passos realizados durante o preparo e sequenciamento das amostras.

Após a extração do DNA (Figura 2a), este é amplificado em um processo que visa criar inúmeras cópias idênticas de uma mesma sequência de DNA (Figura 2b). Esta etapa é fundamental para que se consiga uma cobertura adequada após o alinhamento [8, 9], assegurando que uma mesma posição do genoma seja sequenciada diversas vezes, dando confiabilidade ao resultado final.

Em sistemas NGS, normalmente será necessária a fragmentação do DNA, que pode ocorrer de forma mecânica ou enzimática (Figura 2c). Este procedimento é necessário para adequar o tamanho dos fragmentos de DNA à capacidade do sequenciador. Assim, cada cópia da amostra é transformada em centenas de milhares de fragmentos cortados de forma aleatória e independente[vii].

A seguir, os fragmentos passam por um processo de *tagmentação* (do inglês *tagmentation*, ou inserir *tags* – Figura 2d), onde serão adicionados adaptadores adequados à tecnologia a ser utilizada no sequenciamento. Estes adaptadores nada mais são que sequências específicas de DNA, compostas por umas poucas bases, que serão anexadas às extremidades de cada fragmento, permitindo que eles se liguem a superfícies preparadas para capturá-los, como *microbeads* da tecnologia Ion Torrent ou *nanowells* nas *flow cells* da Illumina. Neste momento também podem ser inseridos *bar codes* de identificação de amostras, já que algumas tecnologias permitem o sequenciamento de diversas amostras diferentes em uma mesma corrida, barateando todo o processo.

Finalmente, estas amostras são inseridas em um sequenciador que irá fazer a chamada das bases e armazenar os resultados em um arquivo FASTQ (Figura 2e). Os vídeos [10, 11] podem dar uma boa ideia de como isto ocorre.

E é com base nestas informações que realizamos a interpretação dos resultados dos softwares de controle de qualidade.

5.1 Resultados do FastQC

Os relatórios de análise do FastQC são divididos em onze seções. Aqui iremos discorrer sobre seus pontos mais significativos. Maiores informações podem ser obtidas na documentação do FastQC [5].

[vii] Mas se os fragmentos são todos de tamanho diferente, por que as sequências no arquivo FASTQ tem todas o mesmo tamanho? Primeiro, porque os protocolos de laboratório, quando seguidos corretamente, produzem fragmentos de tamanho médio previsível, que pode ser testado em equipamentos especiais. Segundo, o sequenciador não sequencia, necessariamente, todo o fragmento. Ele realiza um número determinado de ciclos que identificarão uma quantidade específica de bases, escolhida em função da cobertura desejada. Normalmente, um número maior de ciclos de chamada de bases melhora a cobertura média de um sequenciamento, mas também torna o processo mais caro.

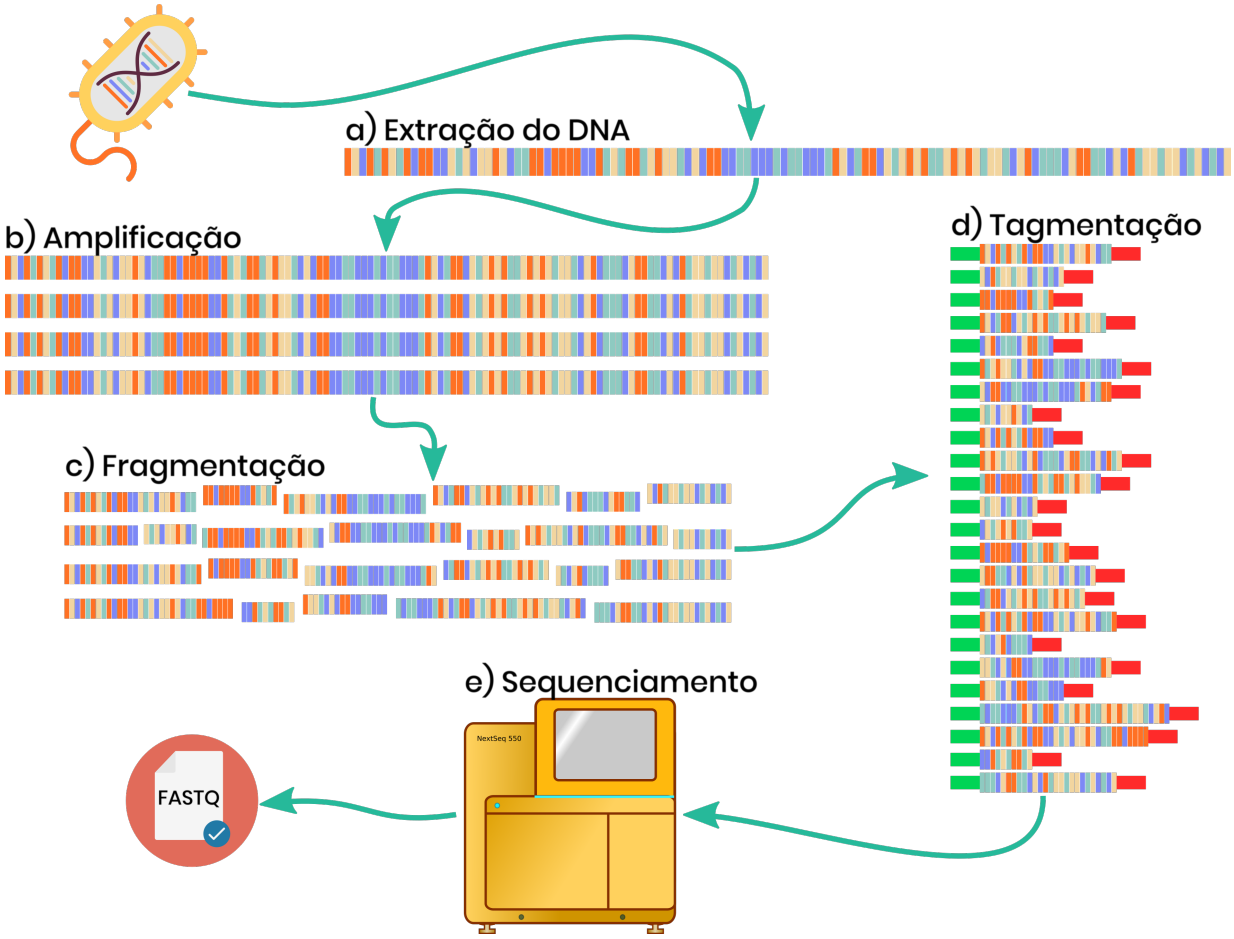


Figura 2: Modelo esquemático do preparo e sequenciamento de amostras

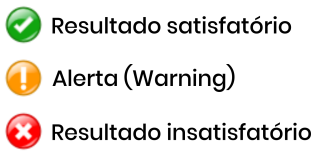


Figura 3: Ícones de classificação de resultado.

O lado esquerdo da página do relatório traz um sumário das estatísticas realizadas contendo o item analisado precedido de um ícone que indica se o teste passou, falhou ou se ocorreu alguma condição de alerta, conforme pode ser visto na Figura 3. Porém, cabe aqui ressaltar que, em alguns casos, apesar do FastQC identificar uma condição de erro ou alerta, pode se tratar de uma situação normal e esperada para o tipo de análise que está sendo realizada, já que o software utiliza parâmetros fixos para realizar tal classificação. Cabe ao bioinformata verificar, com base em todos os parâmetros do experimento, se o teste é realmente válido ou não.

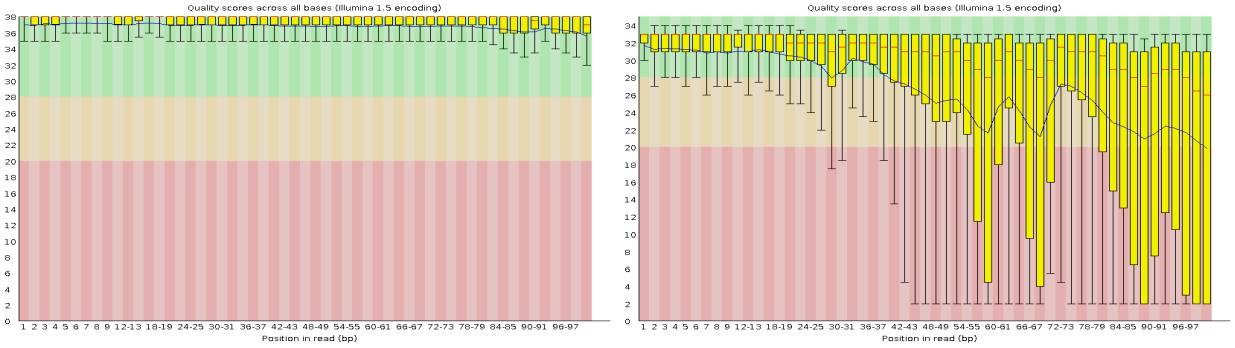
5.1.1 Basic Statistics

Esta seção traz informações sobre o arquivo processado, onde podemos encontrar o nome do arquivo *.fastq*, quantidade de sequências, tamanho do *reads*, etc.

5.1.2 Per base sequence quality

O primeiro gráfico a ser exibido no relatório do FastQC é o de qualidade por base. Nele o eixo dos X representa a posição ocupada pelas bp e o eixo dos Y o escore de qualidade Q, conforme pode ser visto na Figura 4.

Como ambos os arquivos representados na Figura 4 possuem 100bp de comprimento, o eixo dos X possui 100 posições. Em *reads* maiores, estas posições podem aparecer agrupadas em *bins*. O eixo dos Y possui três regiões com cores distintas:



(a) Boa qualidade (*good.fastq.gz*) (b) Má qualidade (*bad.fastq.gz*)

Figura 4: Gráfico de qualidade das sequências por base

Vermelha – escores com baixa qualidade: $0 \leq Q < 20$

Amarela – escores de qualidade razoável: $20 \leq Q < 28$

Verde – escores de boa qualidade: $Q \geq 28$

As barras verticais são *boxplots* da distribuição dos valores de qualidade das bases que ocupam as mesmas posições no *read*. O traço vermelho representa a mediana e a caixa amarela representa o intervalo entre 25 e 75% da distribuição. Os bigodes inferior e superior do boxplot indicam o percentual de 10 e 90% da distribuição, respectivamente. A linha azul representa o valor médio de qualidade das bases na posição. Desta forma, um *boxplot* pequeno representa posições onde a qualidade foi mais consistente em todos os *reads* e *boxplots* inseridos em posições elevadas do eixo Y indicam leituras de melhor qualidade. Logo, o gráfico ideal é aquele onde os *boxplots* são pequenos e em posições elevadas.

Em sequenciadores Illumina é normal a qualidade começar elevada nas primeiras bases e sofrer uma queda gradual nas posições mais à direita do gráfico. Isto se dá porque as leituras da parte final do *read* sofrem influência dos erros ocorridos na leitura das bases de menor posição, pois esses vão se acumulando no decorrer do processo. Normalmente, um maior número de ciclos de chamada de bases implica em uma maior quantidade de erros acumulados e uma menor qualidade na leitura das bases das últimas posições.

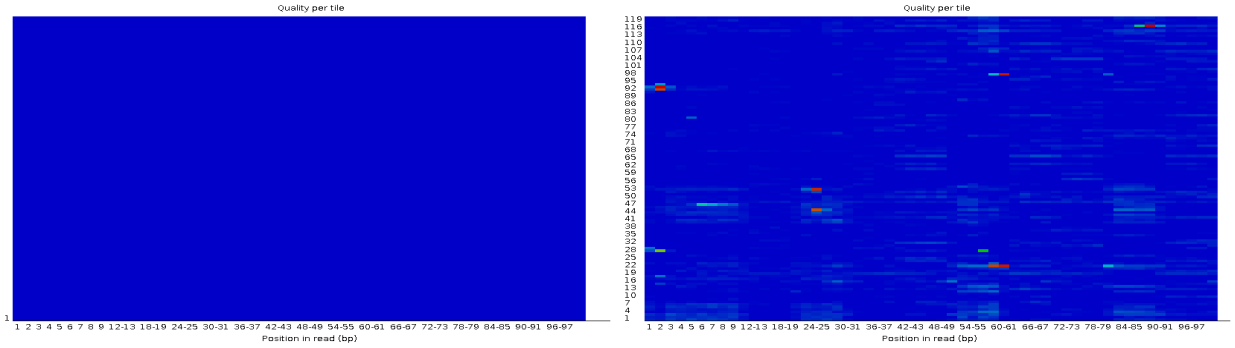
Assim, analisando o relatório apresentado, notamos que a amostra da Figura 4a possui uma qualidade bem superior à da amostra da Figura 4b. Na primeira, as distribuições de qualidade de todas as bases encontram-se na faixa verde, enquanto na segunda os escores de qualidade são baixos e bastante dispersos.

5.1.3 Per tile sequence quality

Uma *flow cell* Illumina é dividida em *lanes* e cada lane é, por sua vez, subdividida nas *tiles* que contém as *nanowells*[12] onde os clusteres de sequenciamento irão se formar. Assim esta seção somente será exibida quando o sequenciamento for feito em alguns modelos de sequenciadores Illumina que registram o identificador da *tile*. De forma semelhante ao gráfico de qualidade por posição da base aqui também são exibidas informações a respeito do desvio em relação à qualidade média dentro de uma *tile*. O eixo dos X indica a posição no *read* e o eixo dos Y a posição da *tile*. Cores quentes indicam que os fragmentos que estavam naquela *tile* tiveram uma qualidade média pior que a média das demais, conforme ilustrado na Figura 5. Quanto menos *tiles* com cores diferentes do azul, melhor.

5.1.4 Per sequence quality scores

Este gráfico mostra a curva de distribuição da qualidade média dos *reads*. O eixo dos X representa o escore Q médio de cada *read* e o eixo dos Y a sua frequência. O melhor cenário é aquele onde todos os *reads* possuam a qualidade média máxima, representada por apenas um pico na extremidade direita do eixo X. A Figura 6 mostra o gráfico das quatro amostras de exemplo, ordenadas da maior qualidade para a menor.



(a) Boa qualidade (*good.fastq.gz*) (b) Má qualidade (*bad.fastq.gz*)

Figura 5: Gráfico de qualidade das sequências por *tile*

Na figura 6a vemos que praticamente não existem *reads* de qualidade entre 0 e 29, representado no gráfico como uma linha horizontal em 0. A partir de $Q \cong 30$ a linha vermelha começa a se afastar do eixo dos X, havendo uma grande concentração de sequências entre os escores 36 e 37. A Figura 6b mostra que a frequência de *reads* começa a crescer um pouco mais cedo, em $Q \cong 18$, havendo um grande pico entre os escores 38 e 39. De forma semelhante, a Figura 6c também mostra um aumento prematuro e maior da frequência de *reads* em baixas qualidades. Já na Figura 6d vemos que a linha do gráfico já começa a se elevar em índices de qualidade bastante baixos ($Q = 9$) e que o escore de qualidade máximo ficou em 33, o menor dentre as quatro análises.

5.1.5 Per base sequence content

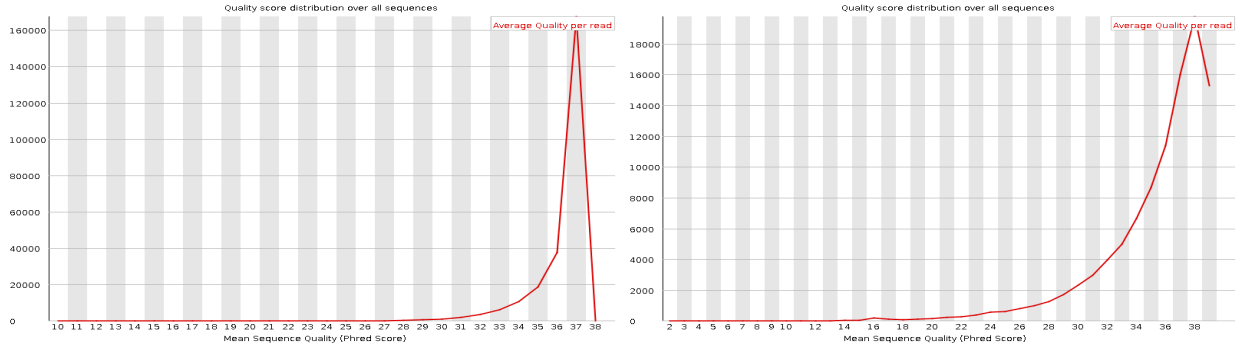
Esta seção mostra o percentual de cada um dos tipo de base encontrado nas diferentes posições da sequência. O eixo dos X indica a posição dentro do *read* e o eixo dos Y, o percentual de cada base, identificadas por linhas coloridas. O resultado esperado é uma distribuição uniforme, representada por uma linha horizontal, e proporcional ao originalmente encontrado no organismo estudado, como visto na Figura 7a. Porém, em alguns casos, pode-se encontrar uma distribuição diferente, principalmente quando for esperada uma grande repetição de fragmentos curtos, como pode ser o caso em um estudo envolvendo microRNA. Assim, os resultados desta seção devem ser observados com cautela e analisados cuidadosamente, sendo apenas um indicador de possível problema no sequenciamento.

5.1.6 Per sequence GC content

Aqui é comparada uma distribuição normal estimada de frequências das bases GC (linha azul na Figura 8) com a distribuição observada nas sequências (linha vermelha). O eixo dos X indica os valores percentuais médios do conteúdo GC dos *reads* e o eixo dos Y a frequência destes *reads*. Isto significa que em 0 aparecerá a contagem dos *reads* que não possuem nenhuma base GC em sua composição e em 100% a frequência daqueles compostos apenas por bases GC. Em tese, é desejável uma coincidência entre a linha azul e vermelha. Porém, esta informação não deve ser tomada de forma absoluta, já que resultados diferentes podem ser consistentes com o tipo de experimento realizado.

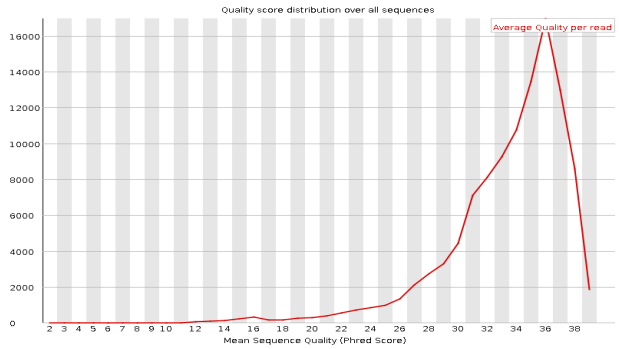
5.1.7 Per base N content

Esta seção mostra o percentual de bases não identificadas (N) por posição do *read*. O eixo dos X indica a posição da sequência e o eixo dos Y o percentual de N, considerando-se todos os *reads* listados no arquivo *.fastq*. Obviamente, o ideal é encontrar-se um número muito baixo de N em todas as posições do *read*, semelhante ao que pode ser visto na Figura 9a. Já na Figura 9b, nota-se uma elevada ocorrência de N ($\cong 5\%$) próximo às posições 60 e 67 dos *reads*.

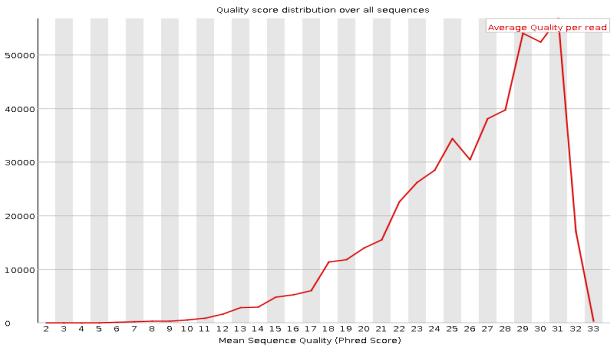


(a) Boa qualidade (*good.fastq.gz*)

(b) Qualidade razoável (*RNA-Seq.fastq.gz*)



(c) Baixa qualidade (*RRBS.fastq.gz*)



(d) Má qualidade (*bad.fastq.gz*)

Figura 6: Gráfico de frequência de qualidade média de *reads*

5.1.8 Sequence Length Distribution

Alguns sequenciadores produzem *reads* de tamanho variável enquanto outros os produzem com tamanho fixo. O gráfico desta seção mostra a distribuição de tamanho de *reads* observada na amostra. Segundo o manual do FastQC [5], o software marcará com um aviso de erro sempre que detectar algum *read* com tamanho igual a zero e emitirá um *warning* quando detectar sequencias de tamanho variado.

Obviamente, dependendo da tecnologia utilizada no sequenciamento, isto pode não ser um problema, como é o caso do exemplo da Figura 10b, cujas sequências foram obtidas a partir de um sequenciador de *reads* longos da PacBio [13].

5.1.9 Sequence Duplication Levels

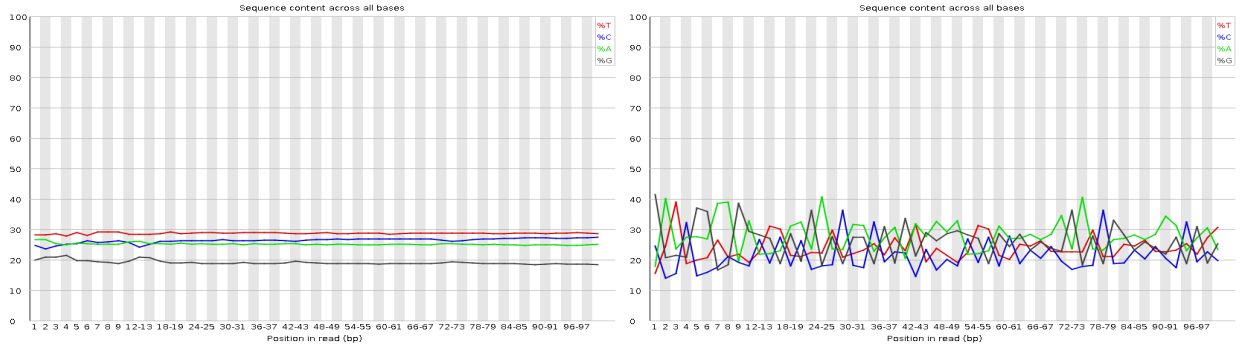
Esta seção exibe a quantidade de sequências duplicadas encontradas na amostra (Figura 11). O eixo dos X indica a soma das diversas frequências de *reads* duplicados, divididos por faixas e em uma escala não linear.O eixo dos Y indica o percentual de *reads* duplicados em relação ao total. A linha azul exibe os resultados efetivamente encontrados e a linha vermelha representa a simulação de um cenário onde as sequências repetidas são removidas. O título do gráfico também exibe qual seria o total de *reads* aproveitáveis caso a remoção ocorra.

Analisando a Figura 11a notamos que existe um grande número de sequências únicas, representadas no lado esquerdo do gráfico como um pico de aproximadamente 100% do total de *reads*. O título do gráfico nos diz que restariam 100% de *reads* caso as repetições fossem removidas.

Já no caso da Figura 11b, observamos que o pico inicial contém algo em torno de 70% de amostras únicas e níveis elevados de repetição, contendo mais de 1.000 tipos de *reads* repetidos, representado cerca de 20% do total

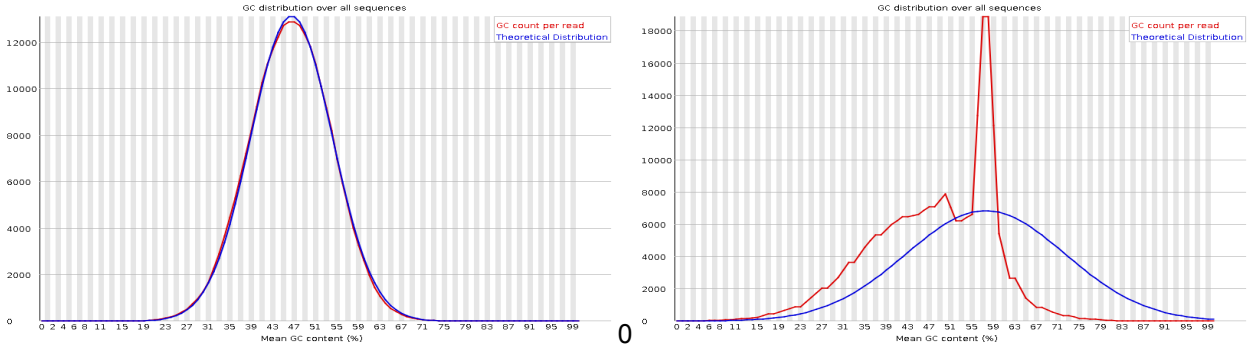
[viii] Ou não! Um resultado assim merece uma análise mais específica, onde devem ser considerados os parâmetros do experimento. Porém tais dados não estão disponíveis para esta amostra.

[ix] Mais uma vez este resultado precisaria ser melhor analisado à luz de outras informações não disponíveis neste exemplo.



(a) Boa qualidade (*good.fastq.gz*) (b) Má qualidade[viii] (*RNA-Seq.fastqc.gz*)

Figura 7: Gráfico de qualidade das sequências por tipo de base



(a) Boa qualidade (*good.fastq.gz*) (b) Exemplo de má qualidade[ix]

Figura 8: Gráfico de frequência de conteúdo GC

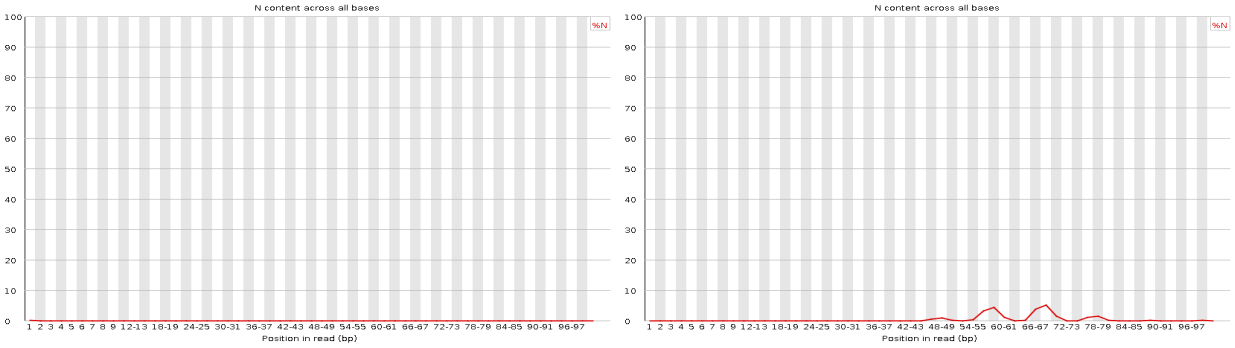
da amostra. A remoção das sequências repetitivas causaria a perda de mais de 28% do total de *reads* da amostra.

Mais uma vez, este é um indicador que não pode ser considerado de forma absoluta. Quando lidamos com genomas ou exomas, conforme visto na Figura 2, uma série de moléculas de DNA, grandes e idênticas, serão fragmentadas de forma aleatória, gerando uma variedade muito grande de *reads*. Neste cenário espera-se uma repetição muito pequena de sequências, já que a probabilidade de fragmentação idêntica também é pequena. Porém, quando sequenciando RNA-Seq ou outras formas de RNA curtos, é esperada a repetição de sequências, pela própria característica do estudo. Outro exemplo onde a multiplicidade de *reads* idênticos é esperada ocorre quando realizamos um sequenciamento com base em *amplicon*, já que a tendência de gerar reads duplicados no processo é maior. Assim, os resultados desta seção devem ser sempre analisados à luz do experimento realizado, bem como da forma como as bibliotecas foram preparadas.

5.1.10 Overrepresented sequences e Adapter Content

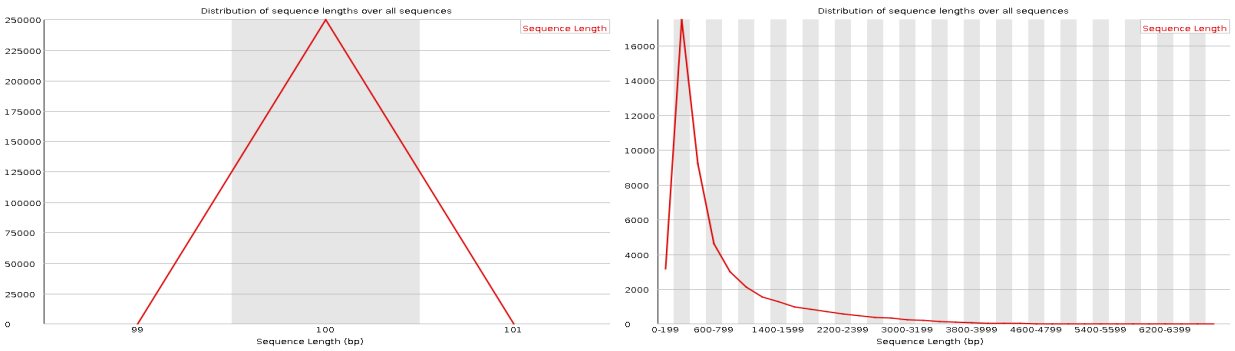
Se a seção anterior nos dizia sobre a análise de repetições de forma quantitativa, estas duas últimas seções dizem respeito à análise qualitativa das repetições, identificando quais foram as sequências que alcançaram altas frequências de ocorrência, tentando identificar a sua natureza.

Estas informações serão úteis para a identificação das causas que levaram a um inesperado índice de super-representação de algumas sequências, como no caso de uma contaminação por dímeros de adaptadores.



(a) Boa qualidade (*good.fastq.gz*) (b) Má qualidade (*bad.fastq.gz*)

Figura 9: Gráfico de frequência de N por posição no *read*



(a) Sequenciador com tamanho de *reads* constante (*good.fastq.gz*) (b) Sequenciador com tamanho de *reads* variável (*pacbio*)

Figura 10: Gráfico de distribuição de comprimentos de *reads*

5.2 Resultados do MultiQC

Como o MultiQC apenas sumariza os resultados obtidos pelo FastQC, suas seções são bastante parecidas e a análise dos resultados é basicamente a mesma. O que muda é apenas como a informação pode ser visualizada, havendo aqui diversas opções disponíveis, dependendo do caso.

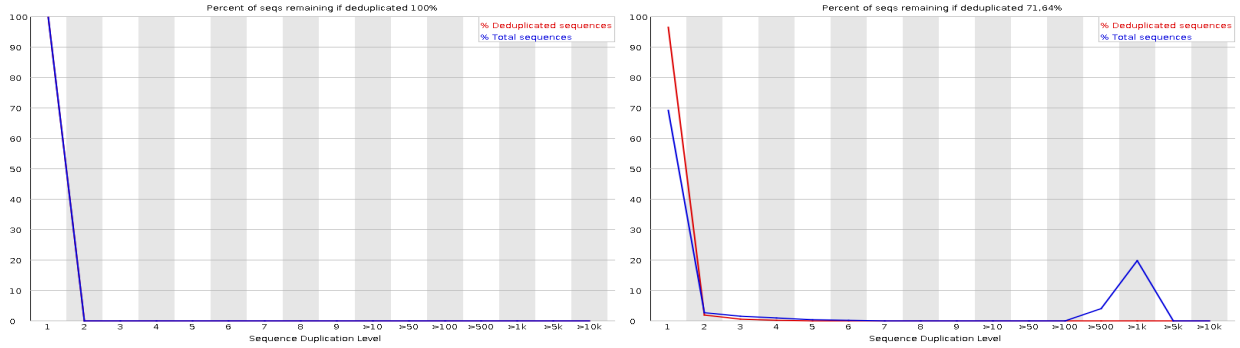
Assim, veremos apenas algumas peculiaridades com relação à forma com que o dado é exibido, já que a análise do conteúdo já foi abordada anteriormente. Maiores informações podem ser obtidas no site do produto [6].

5.2.1 General Statistics

Como o próprio nome já afirma, esta seção exibe um sumário das estatísticas de todas as amostras analisadas, sendo possível selecionar as colunas a serem exibidas e criar gráficos de correlação entre duas colunas selecionadas (botão *Plot*). A Figura 12 mostra o layout com todas as colunas disponíveis. Colocando-se o mouse sobre o identificador da coluna, obtém-se sua descrição.

5.2.2 Sequence Counts

Esta seção une dados de diversas partes do relatório do FasctQC. Traz informações sobre a quantidade total de *reads* que cada amostra continha e quantos foram considerados como duplicados em números absolutos e percentuais. O gráfico pode ser exibido tanto no modo percentual, quanto no modo de contagem de reads, como é visto na Figura 13.



(a) *Reads* com baixo índice de duplicação (*good.fastq.gz*) (b) *Reads* com alto índice de duplicação(*bad.fastq.gz*)

Figura 11: Gráfico de distribuição de níveis de duplicação de *reads*

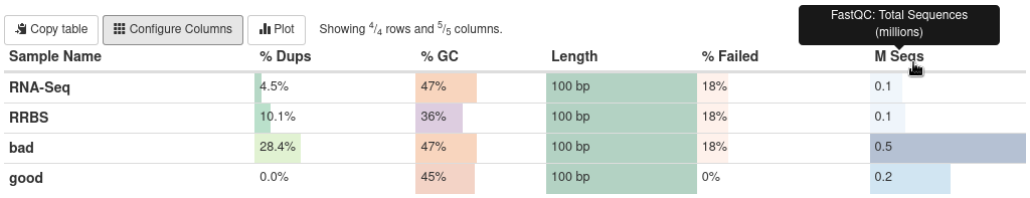


Figura 12: Estatísticas gerais do MultiQC

5.2.3 Sequence Quality Histograms

O gráfico contido nesta seção é muito similar ao exibido na seção *Per base sequence quality* do relatório FastQC. Como principal diferença, aqui é exibida apenas a linha de média de qualidade dos *reads* encontrados em cada posição. Cada linha do gráfico corresponde a uma amostra diferente, e para identificá-la basta colocar o ponteiro do mouse sobre ela, conforme pode ser visto na Figura 14.

5.2.4 Per Base Sequence Content

Aqui o gráfico é equivalente ao gráfico de mesmo nome do relatório do FastQC. Porém ele é exibido sob a forma de heatmap, onde cada linha corresponde a uma amostra e cada coluna corresponde a uma posição ou grupo de posições dentro dos *reads*. As cores traduzem o percentual de participação de cada base na posição. Deslocando-se o ponteiro do mouse sobre as *tiles*, podemos observar na parte superior do gráfico os dados relativos ao nome da amostra, a posição do bp e os percentuais de frequência dos quatro tipos de base (Figura 15). Um clique sobre a linha abre um gráfico bidimensional idêntico ao do FastQC.

5.2.5 Overrepresented sequences

Nesta seção é exibida sob a forma gráfica a mesma informação que no FastQC era exibida sob a forma de lista, em seção homônima. A barra azul claro observada na Figura 16 exibe o percentual de ocorrência da sequência super-representada de maior índice na amostra, e a barra cinza o somatório das demais. Aqui temos um exemplo claro de como os relatórios do FastQC e MultiQC trabalham em conjunto: no MultiQC podemos identificar rapidamente amostras que possuem alto percentual de *overrepresented sequences*, mas para poder analisar o seu conteúdo, devemos lançar mão do relatório do FastQC.

5.2.6 Status Checks

Esta seção contém apenas o *overview* de todas as seções do relatório do FastQC, identificando as amostras e o status do teste (Figura 17), representados pelas cores dos ícones constantes da Figura 3.

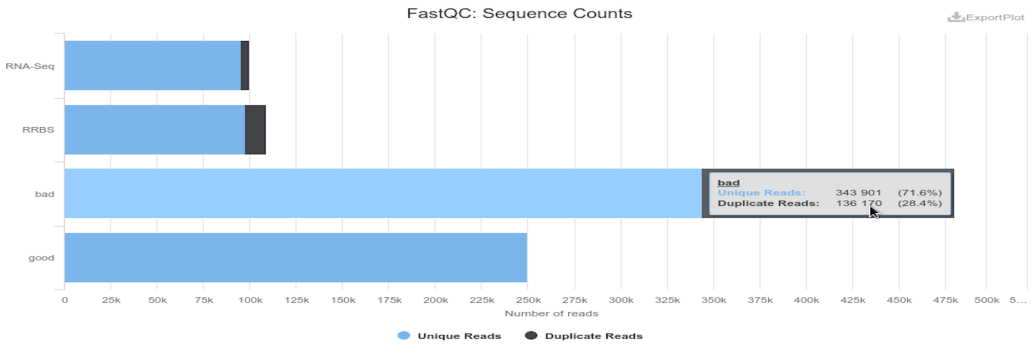


Figura 13: Contagem de sequências

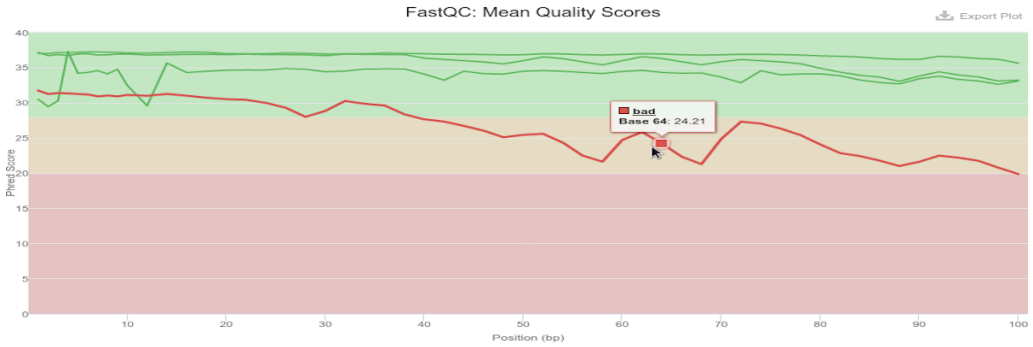


Figura 14: Histograma do valor médio da qualidade das sequências

5.2.7 Outras seções

As demais seções do relatório do MultiQC contém, basicamente, os mesmos gráficos das seções correspondentes do FastQC, onde se veem linhas distintas representando cada uma das amostras analisadas (Figura 18).

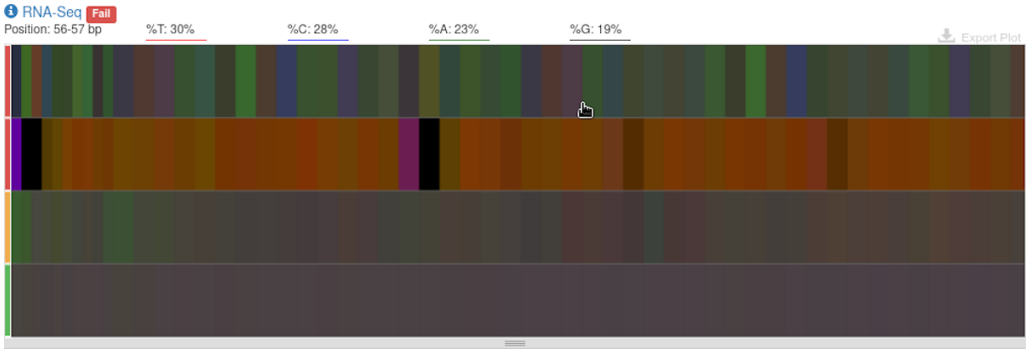


Figura 15: Percentual de bases por posição na sequência

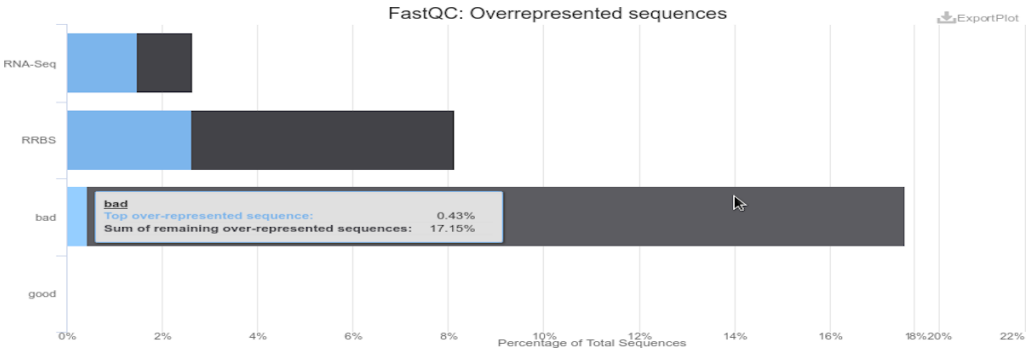


Figura 16: Gráfico de sequências super-representadas

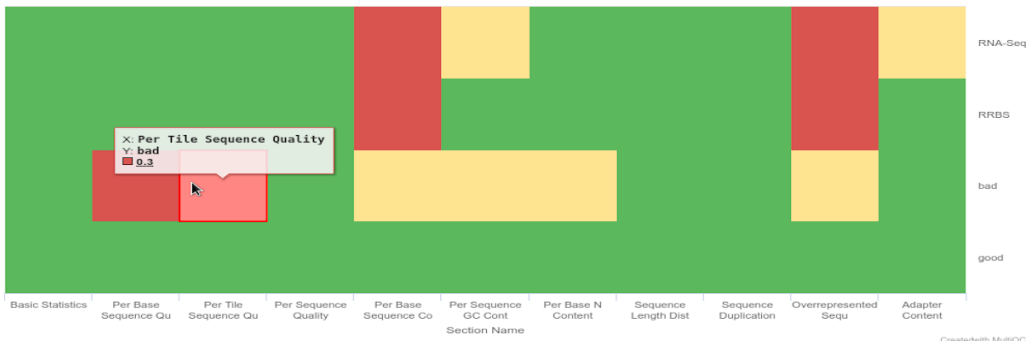


Figura 17: Sumário de status dos testes do FastQC

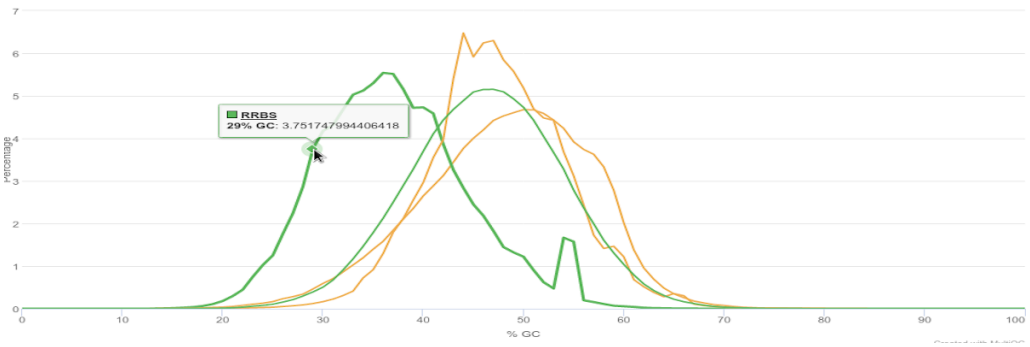


Figura 18: Gráfico de conteúdo de GC para múltiplas amostras.

6 Próximos passos

Depois de analisar a qualidade das sequências de nossas amostras, será necessário tomar medidas para corrigir os problemas encontrados, que podem, de uma maneira geral, serem resumidos a três procedimentos:

Trimagem – consiste na remoção das bases contidas nas extremidades dos *reads*, normalmente para corrigir a queda de qualidade natural naquela região, ou então, remoção de porções das sequências que coincidem com o padrão de algum adaptador;

Filtragem – consiste na remoção de sequências completas que apresentaram baixa qualidade em todo o seu corpo ou excessiva duplicidade;

Masking – implica na substituição de uma base que tenha apresentado baixa qualidade por um N.

Isto, porém, é assunto para outro microcurso.

Referências

[1] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85(8):2444–2448, Apr 1988.

[2] Illumina. Fastq files. https://help.basespace.illumina.com/articles/descriptive/fastq-files/. [Online; acessado 23/01/2020].

[3] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, 8(3):186–194, Mar 1998.

[4] C. F. Reis. Windows ou linux: uma escolha nada difícil. http://www.dunabioinfo.com/pt/blog/pqlinux. [Online; acessado 23/01/2020].

[5] FastQC. Fastqc documentation. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Online; acessado 23/01/2020].

[6] MultiQC. Using multiqc. https://multiqc.info/docs/. [Online; acessado 23/01/2020].

[7] SVGRepo. SVG Vectors – CC BY 4.0. https://www.svgrepo.com/svg/10665/bacteria. [Online; acessado 23/01/2020].

[8] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, Apr 1988.

[9] Illumina. Estimating sequencing coverage. https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf. [Online; acessado 23/01/2020].

[10] Ion Torrent. Ion torrent™ next-gen sequencing technology. https://youtu.be/WYBzbxIfuKs. [Online; acessado 23/01/2020].

[11] Illumina. Sequencing technology video. https://youtu.be/fCd6B5HRaZ8. [Online; acessado 23/01/2020].

[12] Ion Torrent. More data, reduced costs, and faster runs. https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/patterned-flow-cells.html. [Online; acessado 23/01/2020].

[13] Anthony Rhoads and Kin Au. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13, 11 2015.