Aula prática: Sequence Quality Control. Explorando dados de NGS.

06/07/2017

Professor: Jorge Estefano Santana de Souza, jorge@imd.ufrn.br;

Monitores: Danilo Lopes Martins, danilolmartins@gmail.com;
Luan Pereira, luanpereira00@outlook.com.

Objetivos:

Utilizar as ferramentas básicas de analise de qualidade para obter um perfil inicial da qualidade do sequenciamento.

Ferramentas:

1- Linux.
2- WebServer.
3- fastq_screen
4- fastqc
5- samstat
6- DynamicTrim.pl
7- trim_galore
8- cutadapt

Comandos Básicos:

Durante a execução dos tutoriais necessitaremos saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

http://wiki.ubuntu-br.org/ComandosBasicos

Login Servidor:

Inicialmente vamos fazer o login no servidor, abra um terminal no linux e digite:

```
ssh -p 4422 bif@10.7.5.38
```

Irá pedir uma senha, digite:

```
bif0003
```

*ps. não aparece a digitação, o teclado não quebrou não!

Regras para login no servidor:

Interno à UFRN:

ssh -p 4422 bif@10.7.5.38
Senha: bif0003

Externo à UFRN:

ssh -p 4422 bif@177.20.147.141
Senha: bif0003

Dados brutos (raw data):

Durante a execução dos tutoriais necessitaremos de alguns dados iniciais, em via de regra estarão disponíveis no diretório:

```
/home/treinamento/NGS/
```

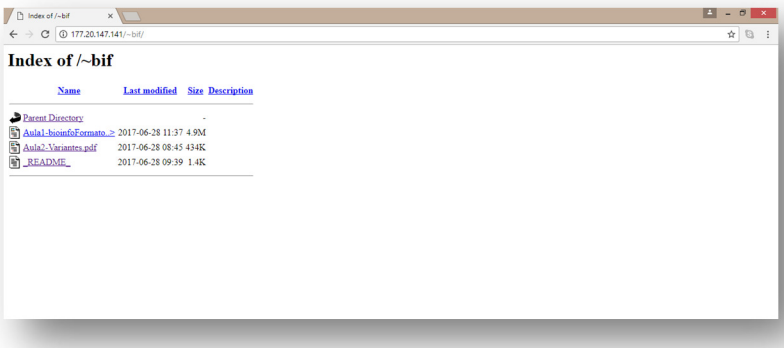
Servido WEB:

Como os trabalhos realizados no servidor são de difícil visualização, iremos necessitar de uma área web para facilitar nossa tarefa, todos os arquivos copiados para o diretório:

```
/home/bif/public_html/
```

Estarão disponíveis via navegador web em:

```
http://177.20.147.141/~bif/
```



Iniciando o Workflow:

1) Vamos começar pelo básico, certifique-se de que a pasta atual é:

```
/home/bif
```

Para isso digite o comando:

```
pwd
```

2) Crie um diretório contendo o seu nome, digite o comando:

```
mkdir SeuNome
```

3) Entre no diretório criado:

```
cd SeuNome
```

4) Crie um diretório chamado bwa:

```
mkdir qual
```

5) Entre no diretório criado:

```
cd qual
```

6) certifique-se de que a pasta atual é a correta:

```
pwd
```

O diretório atual deve ser: /home/bif/SeuNome/qual

7) Inicialmente necessitaremos de arquivos no formato fastq, crie os links:

```
ln -s /home/treinamento/NGS/ERR844339.fastq  .

ln -s /home/treinamento/NGS/10_S5_R1_001.fastq .

ln -s /home/treinamento/NGS/polipo.fastq .
```

8) Agora vamos ver como um arquivo fastq é. comando:

```
less -S ERR844339.fastq
```

*ps. para sair digite a letra q

Fastq_screen:

9) Agora vamos procurar contaminações com o programa fastq_screen. comando:

```
fastq_screen --nohits --subset 0 ERR844339.fastq
--outdir .
```

*ps. o comando deve ser digitado em apenas uma linha

10) Podemos escolher os bancos de contaminantes, para tanto necessitamos de um arquivo **.conf**, copie o exemplo para o diretório corrente:

```
cp /home/treinamento/NGS/fastq_screen.conf  .
```

11) apague o resultado anterior::

```
rm ERR844339_*
```

12) Edite o arquivo fastq_screen.conf e execute o comando:

```
fastq_screen --nohits --subset 0  --conf fastq_screen.conf
ERR844339.fastq --outdir .
```

*ps. o comando deve ser digitado em apenas uma linha

13) Para melhor visualização do resultado vamos copia-los para nossa área web:

Primeiro crie um diretório:

```
mkdir /home/bif/public_html/SeuNome
```

Depois copie:

```
cp ERR844339_* /home/bif/public_html/SeuNome
```

Agora entre no navegador web e visualize no endereço:

```
http://177.20.147.141/~bif/SeuNome/
```

Fastqutils:

14) Algumas estatísticas básicas podem ser obtidas com o programa fastqutils:

```
fastqutils stats  ERR844339.fastq   | more
```

```
fastqutils stats  ERR844339_no_hits.fastq   | more
```

FastQC:

15) Para ter uma estatística mais ampla do resultado do sequenciamento utilizamos o programa FastQC:

```
fastqc ERR844339.fastq -o  .
```

16) O resultado dele é um HTML, melhor visualizado na área web. Copie-o para lá:

```
cp ERR844339_fastqc.* /home/bif/public_html/SeuNome/
```

17) Repita o processo para os demais arquivos **.fastq** e visualize as diferenças:

```
fastqc 10_S5_R1_001.fastq  -o .

fastqc polipo.fastq  -o .

cp *_fastqc.* /home/bif/public_html/SeuNome/
```

SAMstat:

18) O programa FasqQC é um programa pré-alinhamento, para avaliar as sequencias pós-alinhadas podemos utilizar o programa samstat:

Antes vamos necessitar de um arquivo já alinhado (arquivo BAM), crie os links:

```
ln -s  /home/treinamento/NGS/ERR844339.bam  .

ln -s  /home/treinamento/NGS/polipo.bam  .
```

Rode o samstat:

```
samstat  ERR844339.bam

samstat  polipo.bam
```

Não se esqueça de copiar o resultado para área web:

```
cp *.samstat.* /home/bif/public_html/SeuNome/
```

*veja o resultado em: `http://177.20.147.141/~bif/SeuNome/`

Trim:

Após o uso dos programas FasqQC e Samstat talvez seja necessário aplicar algum processo de limpeza, isso irar melhorar o resultados finais e diminuir a taxa de erro de analises posteriores.

19) Podemos usar o DynamicTrim para trimar as pontas das sequencias por qualidade de bases:

```
DynamicTrim.pl -h 20 -i ERR844339.fastq
```

20) Podemos usar o cutadapt para remover adaptadores ou sequencias contaminantes conhecidas.

```
cutadapt -a TGGAATTCTCGG 10_S5_R1_001.fastq >
10_S5_R1_001.ct.fastq
```

*ps. o comando deve ser digitado em apenas uma linha

21) O programa trim_galore, quando temos um sequenciamento illumina e não sabemos os adaptadores usados no processo de sequenciamento

```
trim_galore  10_S5_R1_001.fastq
```

Vale ressaltar que após cada processo de limpeza devemos refazer a analise de qualidade novamente, e assim verificar a sua efetiva melhora.

Referências:

1- Lassmann et al. (2010) "SAMStat: monitoring biases in next generation sequencing data." Bioinformatics doi:10.1093/bioinformatics/btq614 [PMID: 21088025]

2- fastq_screen: https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

3- fastqc: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

4- trim_galore: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

5- Cutadapt removes adapter sequences from high-throughput sequencing reads. MARTIN, Marcel. EMBnet.journal, [S.l.], v. 17, n. 1, p. pp. 10-12, may. 2011. ISSN 2226-6089. Available at: <http://journal.embnet.org/index.php/embnetjournal/article/view/200>. Date accessed: 08 Jul. 2017. doi:http://dx.doi.org/10.14806/ej.17.1.200.