

Aula prática: Chamada de variantes. 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 ferramentas básicas de chamada de variantes e identificar bases variantes de um sequenciamento de segunda geração.

Ferramentas:

- 1- Linux.
- 2- WebServer.
- 3- bwa
- 4- samtools
- 5- mpileup
- 6- VarScan
- 7- SnpEff

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/
```

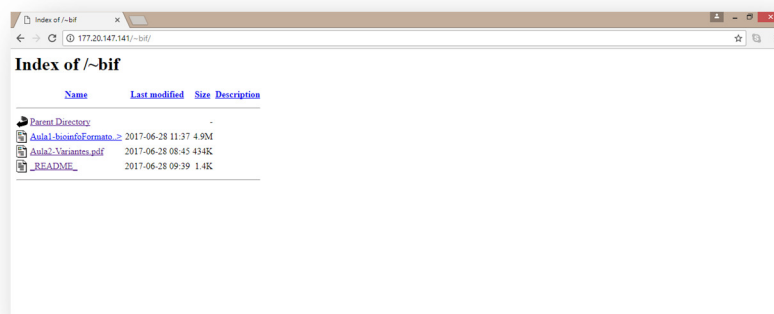
Servido WEB:

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 bwa
```

5) Entre no diretório criado:

```
cd bwa
```

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

```
pwd
```

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

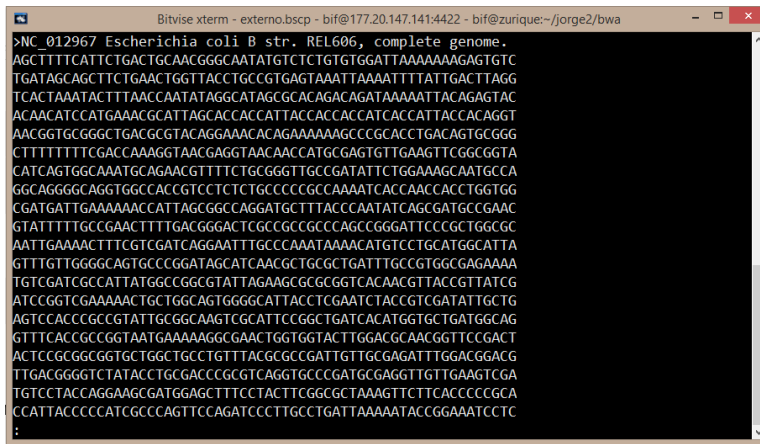
7) Crie links simbólicos para os arquivos:

```
ln -s /home/treinamento/NGS/NC_012967.1.fa .  
ln -s /home/treinamento/NGS/SRR5714077_1_s.1.fastq .  
ln -s /home/treinamento/NGS/SRR5714077_2_s.1.fastq .
```

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

```
less -S NC_012967.1.fa
```


*ps. para sair digite a letra q



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

```
less -S SRR5714077_1_s.1.fastq
```

*ps. para sair digite a letra q



*ps. mais informação do formato FASTQ em https://en.wikipedia.org/wiki/FASTQ_format.

Mapeamento:

10) Para realizar o mapeamento, primeiro temos que criar os indexs do genoma de referência, com os comandos:

```
bwa index -a is NC_012967.1.fa
```

```
samtools faidx NC_012967.1.fa
```

11) Agora vamos rodar BWA utilizando os arquivos gerados até aqui:

```
bwa bwasw -t 4 NC_012967.1.fa  
SRR5714077_1_s.1.fastq  
SRR5714077_2_s.1.fastq -f bwa.sam
```

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

Utilizando o Samtools (analisando o alinhamento):

Agora que temos o arquivo SAM vamos converter-lo para BAM e utilizar o Samtools para manipula-lo e extrair algumas estatísticas básicas.

*ps. informação formato .bam em: http://genome.sph.umich.edu/wiki/SAM_Format

12) Convertendo de SAM para BAM:

```
samtools view -b -S bwa.sam -o bwa.bam
```

13) Visualizando um arquivo BAM:

```
samtools view bwa.bam | less -S
```

14) Visualizando apenas as sequencias não mapeadas:

```
samtools view -f 4 bwa.bam | less -S
```

15) Visualizando apenas as sequencias mapeadas:

```
samtools view -F 4 bwa.bam | less -S
```

16) Quantificando as sequencias não mapeadas:

```
samtools view -c -f 4 bwa.bam
```

17) Quantificando as sequencias com qualidade MAPQ superior a 42:

```
samtools view -c -q 42 bwa.bam
```

Atividade, responda:

Quantas sequencias mapeadas para a referência?

Quantas sequencias mapeadas com qualidade superior a MAPQ 30?

Quantos pareamento corretos existem?

Em busca das variantes:

Agora vamos tentar identificar as variantes genômicas, para tanto temos que gerar o arquivo mpileup, mas antes temos que ordenar as sequencias do arquivo BAM e remover a amplificação de PCR.

18) Ordenando as sequencias do arquivo BAM:

```
samtools sort bwa.bam -o bwa.sort.bam
```

19) Removendo a amplificação de PCR:

```
samtools rmdup bwa.sort.bam bwa.rmd.bam
```

20) Gerando o arquivo mpileup:

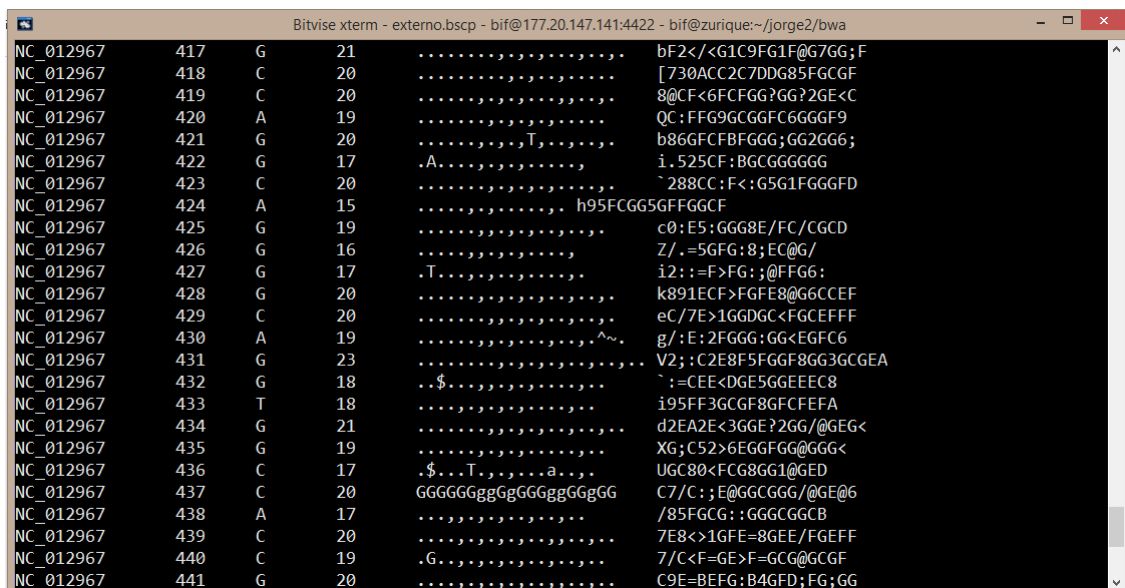
```
samtools mpileup -f  
NC_012967.1.fa bwa.rmd.bam > ecoli.mpileup
```

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

21) Agora vamos ver como um arquivo mpileup é. comando:

```
less -S ecoli.mpileup
```

*ps. para sair digite a letra q



```
Bitwise xterm - externo.bsc - bif@177.20.147.141:4422 - bif@zurique:~/jorge2/bwa
NC_012967 417 G 21 ..... bF2</G1C9FG1F@G7GG;F
NC_012967 418 C 20 ..... [730ACC2C7DDG85FGCGF
NC_012967 419 C 20 ..... 8@CF<6FCFGG?GG?2GE<C
NC_012967 420 A 19 ..... QC:FFG9GCGGFC6GGGF9
NC_012967 421 G 20 ..... b8GGFCFBFGGG;GG2GGG;
NC_012967 422 G 17 .A..... i.525CF:BGCGGGGG
NC_012967 423 C 20 ..... ^288CC:F<:G5G1FGGGFD
NC_012967 424 A 15 ..... h95FCG5GFFGGCF
NC_012967 425 G 19 ..... c0:E5:GGG8E/FC/CGCD
NC_012967 426 G 16 ..... Z/. =5GFG:8;EC@G/
NC_012967 427 G 17 .T..... i2: :=F>FG:;@FFG6:
NC_012967 428 G 20 ..... k891ECF>FGFE8@G6CCCE
NC_012967 429 C 20 ..... eC/7E>1GGDGC<FGCEFF
NC_012967 430 A 19 ..... g/:E:2FGGG:GG<EGFC6
NC_012967 431 G 23 ..... V2;:C2E8F5FGGF8GG3GCGEA
NC_012967 432 G 18 ..$...... ^:=CEE<DGE5GGEEEC8
NC_012967 433 T 18 ..... i95FF3GCGF8GFCFEFA
NC_012967 434 G 21 ..... d2EA2E<3GGE?2GG/@GEG<
NC_012967 435 G 19 ..... XG;C52>6EGGFGG@GGG<
NC_012967 436 C 17 .$....T..... UGC80<FCG8GG1@GED
NC_012967 437 C 20 GGGGGGGgGgGGGgGgGGG C7/C:;E@GGCGGG/@GE@6
NC_012967 438 A 17 ..... /85FGCG: :GGCGGCB
NC_012967 439 C 20 ..... 7E8<>1GFE=8GEE/FGEFF
NC_012967 440 C 19 .G..... 7/C<F=GE>F=GCC@GCCF
NC_012967 441 G 20 ..... C9E=BEFG:B4GFD;FG;GG
```

22) Agora fazer a chamada de variantes usando o programa VarScan:

```
varscan mpileup2snp ecoli.mpileup  
--output-vcf --strand-filter 0 > ecoli.vcf
```

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

Anotação das variantes:

O arquivo VCF (variant call format), contém todas as variantes (de base única, de inserção e de deleção), no entanto nessa versão inicial não estão anotadas todas as informações relevantes para extrair o significado biológico de cada variante, para tanto devemos executar o processo de anotação de variantes.

23) Agora fazer a anotação das variantes usando o programa SnpEff:

```
snpEff eff Escherichia_coli_B_REL606_uid58803  
ecoli.vcf > ecoli.eff.vcf
```

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

Agora temos o arquivo que contém todas as variantes e as informações relevantes para extrair o significado biológico de cada variante.

Referências:

- 1- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, 25:1754-60. [PMID: [19451168](#)]
- 2- Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics*, Epub. [PMID: 20080505]
- 3- Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9. [PMID: 19505943]
- 4- Li H A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

- 5- VarScan 1: Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* (Oxford, England), 25 (17), 2283-5 PMID: 19542151
- 6- VarScan 2: Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111
URL: <http://varscan.sourceforge.net>
- 7- A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. *Fly* (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672