CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

NEXT GENERATION SEQUENCING

Análise de Dados de Sequenciadores de Segunda Geração

Prof. Dr. JORGE ESTEFANO SANTANA DE SOUZA

E-mail: jorge@imd.ufrn.br

Chamada de variantes

BioME Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

2Bio Instituto de Bioinformática e Biotecnologia

IMD Instituto Metrópole Digital | PPG BIOINFO UFRN | UFRN Universidade Federal do Rio Grande do Norte

Objetivo:

Utilizar ferramentas básicas de chamada de variantes e identificar bases variantes de um sequenciamento.

Comandos Básicos de Linux:

Para trabalhar com nossos dados, vamos precisar saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

http://wiki.ubuntubr.org/ComandosBasicos

Ferramentas:

1- Linux.
2- WebServer.

3- bwa
4- samtools
5- mpileup
6- VarScan
7- SnpEff

Inicial:

Login maquina local:
Login:
Senha:

Login no server:

ssh -p 4422 bif@177.20.147.141
Senha: bif0003

Inicial:

Pasta com dados iniciais:

/home/treinamento/NGS/

Pasta servidor WEB:

/home/bif/public_html/

Preparando os dados iniciais:

```
mkdir seu_nome

cd seu_nome

mkdir bwa

cd bwa/

cp  /home/treinamento/NGS/NC_012967.1.fa .

bwa index -a is NC_012967.1.fa

samtools faidx NC_012967.1.fa

bwa bwasw -t 4  NC_012967.1.fa
                /home/treinamento/NGS/SRR5714077_1_s.1.fastq
                /home/treinamento/NGS/SRR5714077_2_s.1.fastq    -f bwa.sam
```

Perguntas:

Dê uma olhada no primeiro alinhamento:

Quantas sequencias mapeadas para a referência?
Quantas sequencias mapeadas com qualidade?
Quantos pareamento corretos existem?

SAMtools

It is more important tools after the alignment process, it can process aligned sequence reads, and manipulate them with ease.

For example:

→ convert between the two most common file formats (SAM and BAM),

→ sort and index files (for speedy retrieval later),

→ extract specific genomic regions of interest.

→ It also enables quality checking of reads,

→ automatic identification of genomic variants.

Converting SAM to BAM

To convert from SAM to BAM, use the SAMtools `view` command:

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

- `-b` : indicates that the output is BAM.
- `-S` : indicates that the input is SAM.
- `-o` : specifies the name of the output file.

BAM files are stored in a compressed, binary format, and cannot be viewed directly. However, you can use the same `view` command to display all alignments. For example, running:

```
samtools  view  bwa.bam  | more
```

will display all your reads in the unix `more` paginated style.

You can also use `view` to only display reads which match your specific filtering criteria. For example:

```
samtools  view  -f 4 bwa.bam  | more
```

- `-f INT` : extracts only those reads which match the specified SAM flag. In this case, we filter for only those reads with flag value of 4 = read fails to map to the reference genome.

https://broadinstitute.github.io/picard/explain-flags.html

or:

```
samtools  view  -F 4 bwa.bam  | more
```

- `-F INT` : removes reads which match the specified SAM flag.

You can also try out the `-c` option, which does not output the reads, but rather outputs the number of reads which match your criteria. For example:

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

indicates that xx of our artificial reads failed to align to the reference genome.

Finally, you can use the `-q` parameter to indicate a minimal quality mapping filter. For example:

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

outputs the total number of aligned reads that have a mapping quality score of 42 or higher.

Vamos gerar um Mpileup:

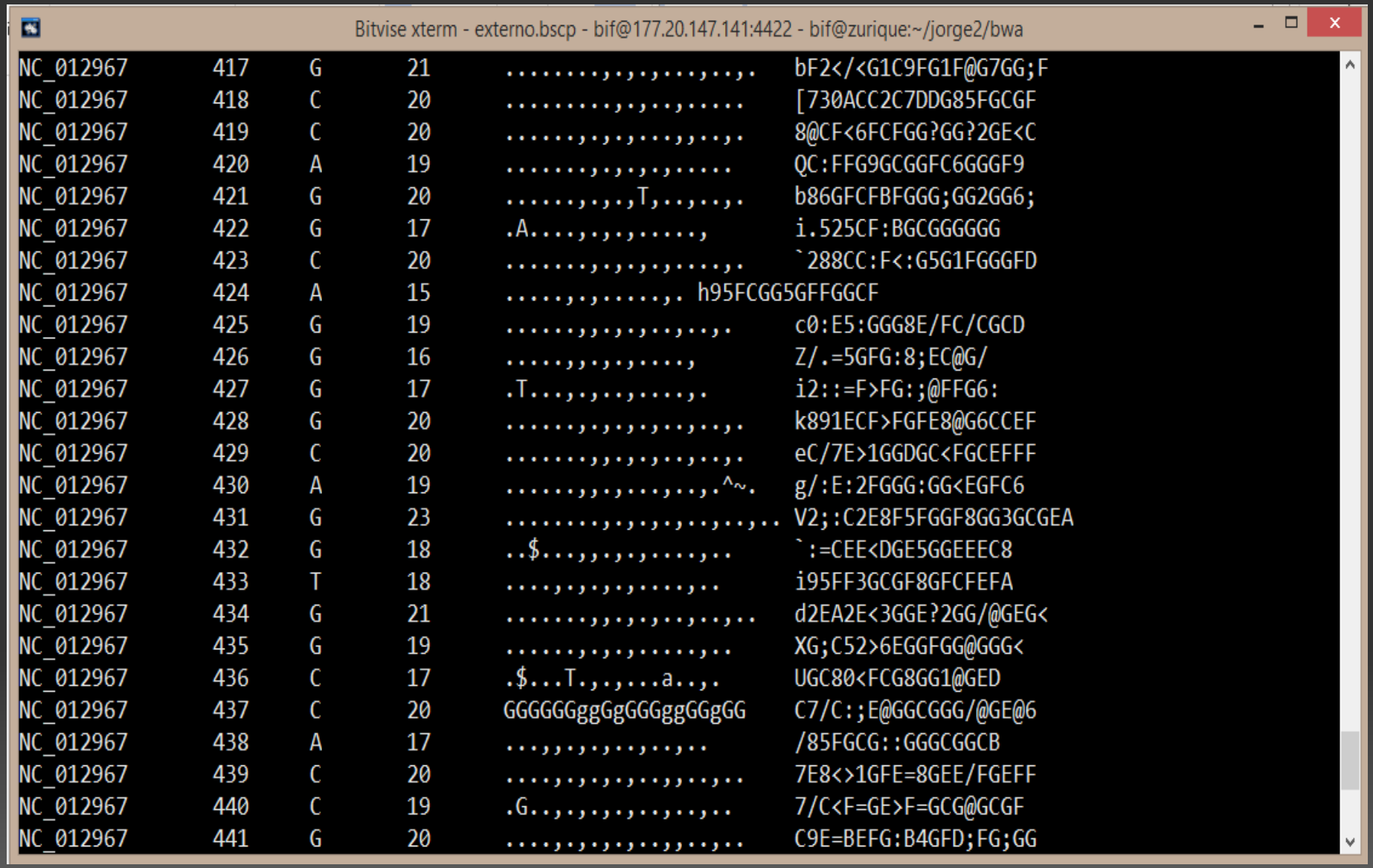
```
samtools view -bS  bwa.sam  -o bwa.bam

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

samtools rmdup bwa.sorted.bam bwa.rmd.bam

samtools mpileup -f NC_012967.1.fa bwa.rmd.bam | less

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


Bitvise xterm - externo.bscp - 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 .....,.,.,T,.,.,.. b86GFCFBFGGG;GG2GG6;
NC_012967 422 G 17 .A...,.,.,.....,, i.525CF:BGCGGGGGG
NC_012967 423 C 20 .....,.,.,.,....,. `288CC:F<:G5G1FGGGFD
NC_012967 424 A 15 .....,.....,. h95FCGG5GFFGGCF
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@G6CCEF
NC_012967 429 C 20 .....,,.,.,.,.,.,. eC/7E>1GGDGC<FGCEFFF
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.,.,...a..,. UGC80<FCG8GG1@GED
NC_012967 437 C 20 GGGGGGggGgGGGggGGgGG C7/C:;E@GGCGGG/@GE@6
NC_012967 438 A 17 ...,.,.,.,.,.,.,.. /85FGCG::GGGCGGCB
NC_012967 439 C 20 ....,.,.,.,.,.,.,.,.. 7E8<>1GFE=8GEE/FGEFF
NC_012967 440 C 19 .G..,.,.,.,.,.,.,.. 7/C<F=GE>F=GCG@GCGF
NC_012967 441 G 20 ....,.,.,.,.,.,.,.,.. C9E=BEFG:B4GFD;FG;GG

Marking PCR duplicates

Map Reads

Picard MarkDuplicates:

```
java -Xmx4g -Djava.io.tmpdir=/tmp
-jar picard/MarkDuplicates.jar
INPUT=input.bam
OUTPUT=input.marked.bam
METRICS_FILE=metrics
CREATE_INDEX=true
VALIDATION_STRINGENCY=LENIENT
```

SamTools rmdup

```
samtools rmdup input.bam out.rmdup.bam
```

Perguntas:

1- O que você observa sobre a saída?

2- Olhando para os dados, qual é a profundidade da cobertura para sua amostra?

3- A cobertura é consistente em todo o genoma? Isso varia? Há lugares onde ela varia mais?

4- Você pode detectar qualquer potencial SNP?

NEXT

Vamos fazer a chamada de variantes:

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

VCF files

Determine Variants

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
#CHROM POS     ID        REF    ALT    QUAL FILTER  INFO                              FORMAT      NA00001            NA00002
20     14370   rs6054257 G      A       29   PASS    NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:AD    0|0:48:102:51,51   1|0:48:102:51,51
20     17330   .         T      A       3    q10     NS=3;DP=11;AF=0.017               GT:GQ:DP:AD    0|0:49:108:58,50   0|1:3:68:65,3
20     1110696 rs6040355 A      G,T     67   PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:AD    1|2:21:50:23,27    2|1:2:20:18,2
20     1230237 .         T      .       47   PASS    NS=3;DP=13;AA=T                   GT:GQ:DP:AD    0|0:54:106:56,50   0|0:48:102:51,51
```

VCF files

Determine Variants

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">

 FORMAT         NA00001            NA00002
 GT:GQ:DP:AD    0|0:48:102:51,51   1|0:48:102:51,51
 GT:GQ:DP:AD    0|0:49:108:58,50   0|1:3:68:65,3
 GT:GQ:DP:AD    1|2:21:50:23,27    2|1:2:20:18,2
 GT:GQ:DP:AD    0|0:54:106:56,50   0|0:48:102:51,51
```

NEXT

Run snpEff.

Annotate variants

Download Database:

```
snpEff  download   Escherichia_coli_B_REL606_uid58803
```

Annotate SNP:

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

Tag in VCF file:

```
Effect ( Effefct_Impact | Codon_Change | Amino_Acid_change | Gene Name | Gene BioType |
                                Coding |  Transcript |   Rank [ | ERRORS | WARNINGS ] )
```

```
NON_SYNONYMOUS_CODING(MODERATE|MISSENSE|Aag/Gag|K1222E|2240|PDE4DIP||CODING|NM_001198832.1|29|1)
```

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

Effect	Note & Example	Impact
Classic		
CDS	The variant hits a CDS.	MODIFIER
CHROMOSOME_LARGE DELETION	A large parte (over 1%) of the chromosome was deleted.	HIGH
CODON_CHANGE	One or many codons are changed e.g.: An MNP of size multiple of 3	MODERATE
CODON_INSERTION	One or many codons are inserted e.g.: An insert multiple of three in a codon boundary	MODERATE
CODON_CHANGE_PLUS CODON_INSERTION	One codon is changed and one or many codons are inserted e.g.: An insert of size multiple of three, not at codon boundary	MODERATE
CODON_DELETION	One or many codons are deleted e.g.: A deletion multiple of three at codon boundary	MODERATE
CODON_CHANGE_PLUS CODON_DELETION	One codon is changed and one or more codons are deleted e.g.: A deletion of size multiple of three, not at codon boundary	MODERATE
DOWNSTREAM	Downstream of a gene (default length: 5K bases)	MODIFIER
EXON	The vairant hits an exon.	MODIFIER
EXON_DELETED	A deletion removes the whole exon.	HIGH
FRAME_SHIFT	Insertion or deletion causes a frame shift e.g.: An indel size is not multple of 3	HIGH
GENE	The variant hits a gene.	MODIFIER
INTERGENIC	The variant is in an intergenic region	MODIFIER
INTERGENIC_CONSERVED	The variant is in a highly conserved intergenic region	MODIFIER
INTRAGENIC	The variant hits a gene, but no transcripts within the gene	MODIFIER
INTRON	Variant hits and intron. Technically, hits no exon in the transcript.	MODIFIER

CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

Obrigado.

E-mail: jorge@imd.ufrn.br

Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

2Bio Instituto de Bioinformática e Biotecnologia