

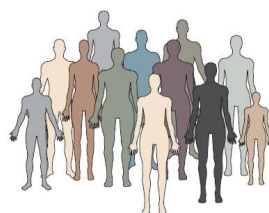
Sequenciação de Nova Geração

- Sequenciação do DNA
- NGS: formatos^{VCF} de ficheiro
- NGS: controlo de qualidade e pré-processamento de fastQ
- NGS: alinhamento de sequências



1

Genoma Humano



- Todos os seres humanos são semelhantes ao nível do DNA, sendo que mais de **99% do material genético é igual** entre diferentes pessoas.
- Os restantes 1% são muito importantes, visto serem a fração que nos diferencia entre si:
 - Determinam a cor dos nossos olhos, cabelo ou pele;
 - Também influenciam o risco de contrair uma determinada doença ou a resposta a um tratamento com um fármaco.

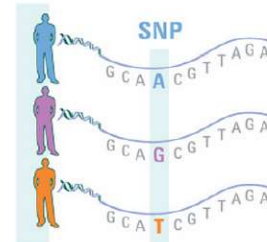


Laboratórios de Bioinformática

2

Polimorfismo de nucleotídeo único

- Polimorfismo de nucleotídeo único ou SNP é **uma variação de nucleótido** que ocorre numa **única posição da sequência de DNA** entre vários indivíduos.
- Se um SNP ocorrer **dentro de um gene**, então o gene é descrito como tendo **mais que um alelo**.
- Alguns SNP estão associados a **determinadas doenças**, por essa razão, **avaliações genéticas** que detetam a predisposição de **doenças em determinados pacientes** são efetuadas com o recurso de **identificação de SNPs**.



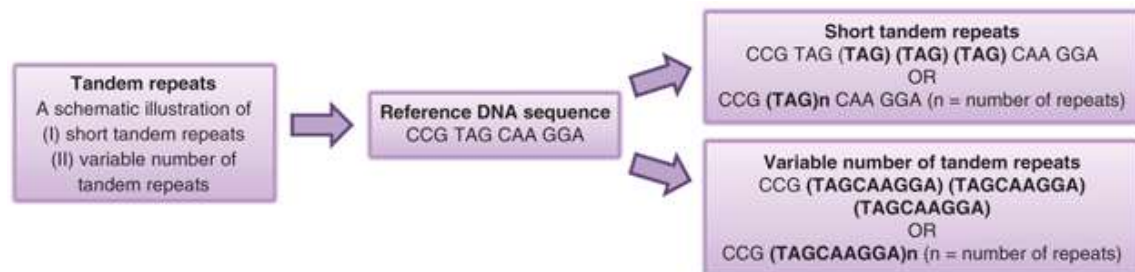
Tipos de alterações SNP



Adapted from C.Ku et al, " the discovery of human genetic variations and their use as disease markers: past, present and future" Journal of Human Genetics, 2017



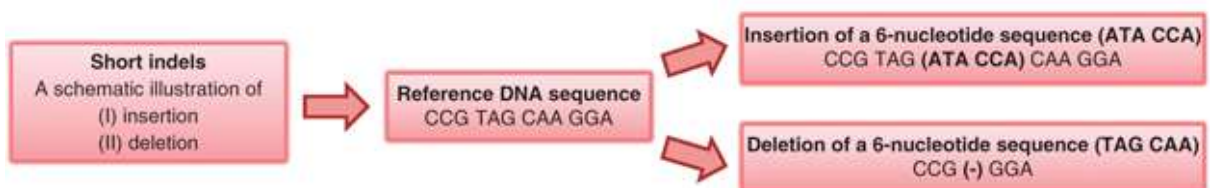
Repetições em tandem



Adapted from C.Ku et al, " the discovery of human genetic variations and their use as disease markers: past, present and future" Journal of Human Genetics, 2017



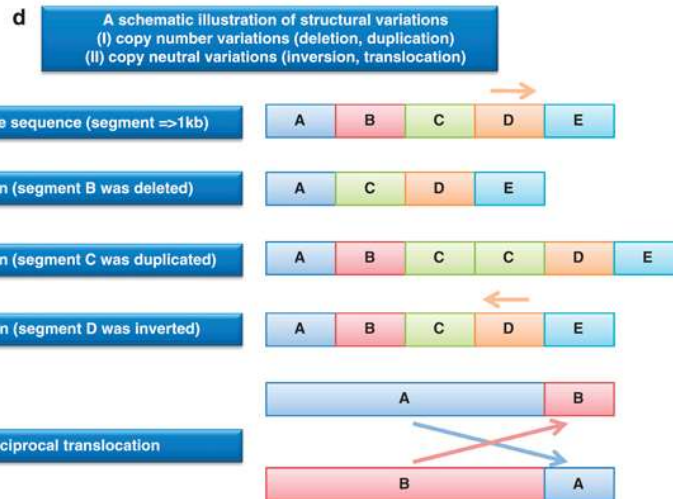
Inserções ou Deleções (short indels)



Adapted from C.Ku et al, " the discovery of human genetic variations and their use as disease markers: past, present and future" Journal of Human Genetics, 2017



Variações



Adapted from C.Ku et al, " the discovery of human genetic variations and their use as disease markers: past, present and future" Journal of Human Genetics, 2017



Laboratórios de Bioinformática

7

Sequenciação do DNA

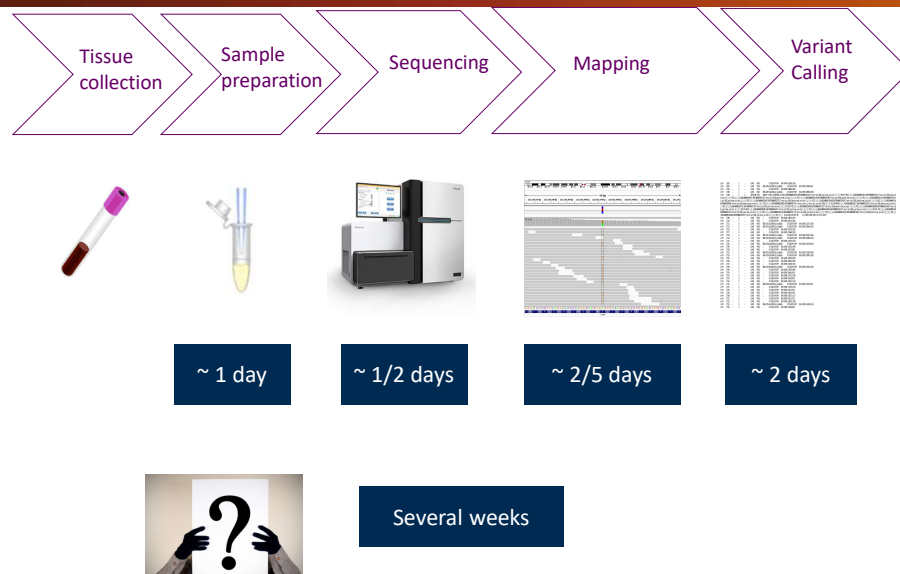
- As aplicações dos sequenciadores de nova geração para o DNA incluem:
 - Assemblagem de um novo genoma;
 - Descoberta de variações de nucleótidos únicos;
 - Detecção de variações de números de cópia de DNA;
 - Detecção de rearranjos estruturais no DNA.



Laboratórios de Bioinformática

8

Procedimento para uma análise de variantes

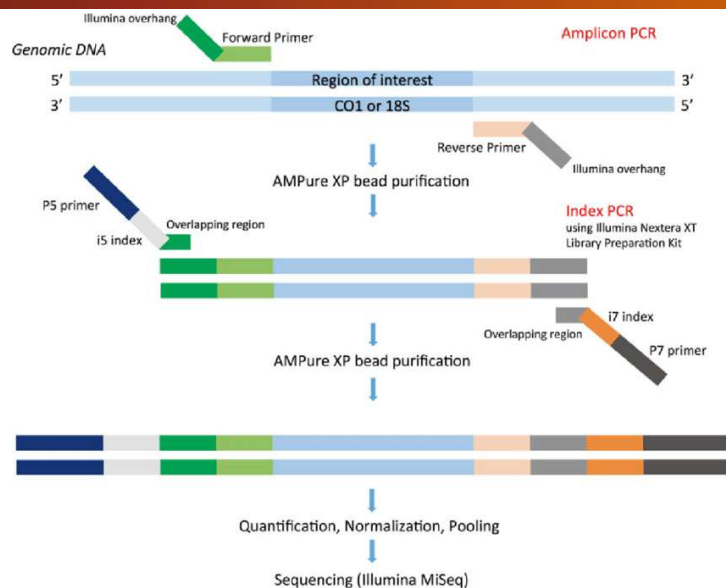


Laboratórios de Bioinformática

9

Preparação das amostras com primers

- Os sequenciadores requerem primers que permitem a iniciação do PCR bem como a identificação das sequências obtidas das amostras.
- O existem diversas bibliotecas de preparação com primers (por exemplo o Illumina Nextera XT kit ou o Illumina TruSeq PE V3 kit)

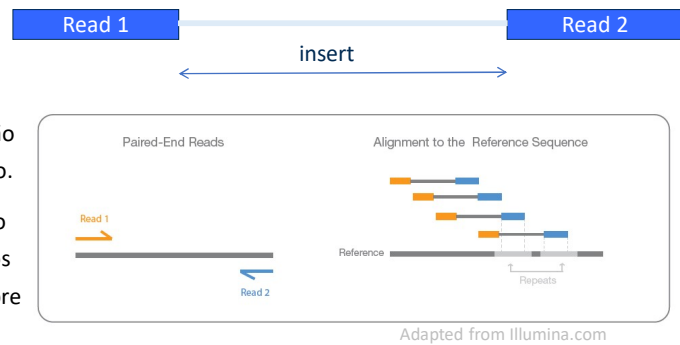


Laboratórios de Bioinformática

10

Reads de seqüências (conceitos)

- *Insert* – O fragmento de DNA utilizado para a sequenciação.
- *Read* – A secção do *insert* que foi sequenciada
- *Single Read (SR)* – Um procedimento de sequenciação no qual o *insert* é sequenciado somente num sentido.
- *Paired End (PE)* – Um procedimento de sequenciação no qual o *insert* é sequenciado em ambos os sentidos (a quantidade de sequencias lidas a tratar será sempre o dobro).
- *Insert size* – A distância média em nucleótidos entre os pares de reads



11

NGS: formatos de ficheiro

- Equipamento de NGS pode exportar dados nos seguintes formatos:
FASTA, FASTQ, SAM/BAM
- Além destes, também existem formatos comuns utilizados na deteção das variações mencionadas anteriormente:
BED, GFF, GTP, VCF

Ficheiro zip com os datasets dos slides:

<https://nextcloud.bio.di.uminho.pt/s/C46mZjoE3oStqsT>



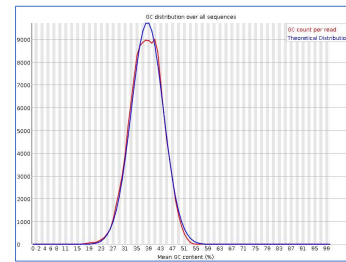
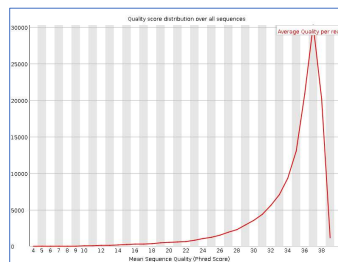
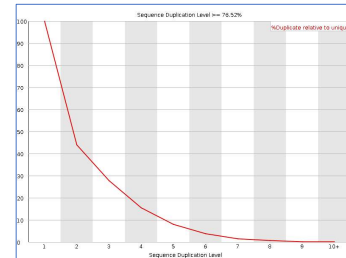
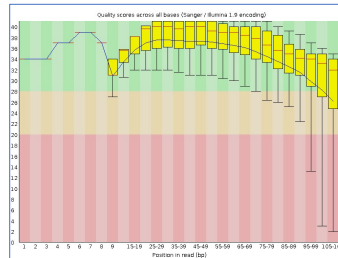
12

NGS: Controlo de qualidade de ficheiros FASTQ

FastQC Report

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✗ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✗ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ✓ [Adapter Content](#)



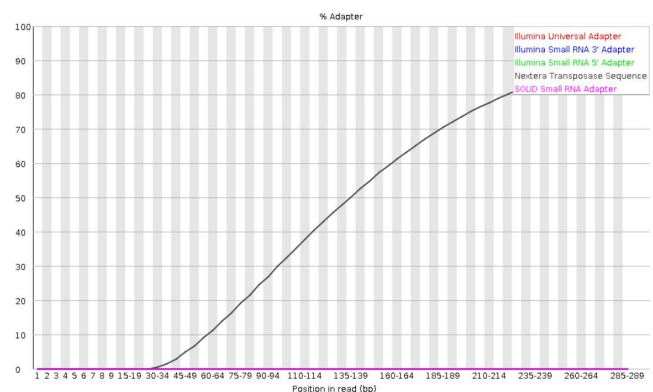
Laboratórios de Bioinformática

15

NGS: Pré-processamento de ficheiros FASTQ

- Por vezes as reads do ficheiro FASTQ contêm os primers de sequenciação e necessitam de ser removidos.
- Para este processo existe softwares como o trimmomatic, fastxtools, cutadapt, bduk, sickle, etc.

✗ Adapter Content



Laboratórios de Bioinformática

16

NGS: Pré-processamento de ficheiros FASTQ

```

rrodrigues@MancoComEsteroides:~/assembly_taster$ docker run --rm -v ~/assembly_taster:/data
-it quay.io/biocontainers/trimmomatic:0.35--6 bash
bash-4.2# cd /data
bash-4.2# ls
README.txt          mt_barcode.txt      top_1.fq
glimmer_to_gbk.perl mt_reads.fastq      top_2.fq
human_mitochondrial.gbk top.bam
bash-4.2# ls /usr/local/share/trimmomatic-0.35-6/adapters
NexteraPE-PE.fa  TruSeq2-SE.fa  TruSeq3-PE.fa
TruSeq2-PE.fa    TruSeq3-PE-2.fa TruSeq3-SE.fa
bash-4.2#

```

O Trimmomatic disponibiliza as sequências dos primers mais utilizados! Estes ficheiros podem ser utilizados no parâmetro do ILLUMINACLIP

Comando docker:

```
docker run --rm -v ~/assembly_taster:/data -it quay.io/biocontainers/trimmomatic:0.35--6 bash
```

- http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf
- <https://thesequencingcenter.com/knowledge-base/trimming-illumina-adapter-sequences/>



Laboratórios de Bioinformática

17

NGS: Pré-processamento de ficheiros FASTQ

Exemplo comando Trimmomatic para SE:

```
trimmomatic SE -phred33 input.fq.gz output.fq.gz ILLUMINACLIP:/usr/local/share/trimmomatic-0.35-6/adapters/TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 AVGQUAL:28
```

Exemplo comando Trimmomatic para PE:

```
trimmomatic PE input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz
output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz
ILLUMINACLIP:/usr/local/share/trimmomatic-0.35-6/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3
TRAILING:3 MINLEN:36 AVGQUAL:28
```



Laboratórios de Bioinformática

18

NGS: Alinhamento de sequencias – SAM/BAM

- SAM/BAM (.sam, .bam) é abreviação para *sequence alignment map*. Este formato de ficheiro contem colunas separadas por tabs, no qual detalha informação de *reads* contra um genoma/sequência de referencia.
SAM apresenta os dados em formato de texto, enquanto que o BAM é a versão binária com otimização para executar indexação.
- O cabeçalho destes ficheiros começa com um símbolo '@'.
- Cada coluna consiste em:

Descrição detalhada de cada coluna em:
<http://bioinformatics.cvr.ac.uk/blog/java-cigar-parser-for-sam-format/>

Format:

1. **QNAME** Query template/pair NAME
2. **FLAG** bitwise FLAG
3. **RNAME** Reference sequence NAME
4. **POS** 1-based leftmost POSition/coordinate of clipped sequence
5. **MAPQ** MAPping Quality (Phred-scaled)
6. **CIGAR** extended CIGAR string
7. **MRNM** Mate Reference sequence NaMe ('=' if same as RNAME)
8. **MPOS** 1-based Mate POSition
9. **LEN** inferred Template LENgth (insert size)
10. **SEQ** query SEQUENCE on the same strand as the reference
11. **QUAL** query QUALity (ASCII-33 gives the Phred base quality)
12. **OPT** variable OPTional fields in the format TAG:VTYPE:VALUE



Laboratórios de Bioinformática

19

NGS: Alinhamento de sequências – SAM/BAM

- Samtools é um pacote de ferramentas para a manipulação de ficheiros SAM e BAM.

```

File Edit View Search Terminal Help
bioinformatics@bioinformatics-VirtualBox:~$ docker run -it biocontainers/samtools:v1.7.0_cv4 bash
bioinformatics@eb16e56b84:/data$ samtools --help

Program: samtools (Tools for alignments in the SAM format)
Version: 1.7 (using HTSLib 1.7)

Usage: samtools <command> [options]

Commands:
  -- Indexing
    dlist      create a sequence dictionary file
    faidx      index/extract FASTA
    index      index alignment
  -- Editing
    calmd      recalculate MD/NM tags and '=' bases
    fixmate    fix mate information
    reheader   replace BAM header
    targetcut  cut fosmid regions (for fosmid pool only)
    addreplacrg adds or replaces RG tags
    markdup    mark duplicates
  -- File operations
    cat        shuffle and group alignments by name
    collate    concatenate BAMs
    merge      merge sorted alignments
    mpileup    multi-way pileup
    sort       sort alignment file
    split      splits a file by read group
    quickcheck quickly check if SAM/BAM/CRAM file appears intact
    fastq      converts a BAM to a FASTQ
    fasta      converts a BAM to a FASTA
  -- Statistics
    bedcov     read depth per BED region
    depth      compute the depth
    flagstat   simple stats
    lsstats    BAM index stats
    phase      phase heterozygotes
    stats      generate stats (former bancheck)
  -- Viewing
    flags      explain BAM flags
    tview      text alignment viewer
    view       SAM<->BAM conversion
    dmpsd      convert padded BAM to unpadded BAM
  bioinformatics@eb16e56b84:/data$

```

Bioinformatics linux contains samtools docker:

```
docker run -it biocontainers/samtools:v1.7.0_cv4 bash
```

Usage: samtools <command> [options]

Command: view	SAM<->BAM conversion
sort	sort alignment file
mpileup	multi-way pileup
depth	compute the depth
faidx	index/extract FASTA
tview	text alignment viewer
index	index alignment
idxstats	BAM index stats (r595 or later)
fixmate	fix mate information
flagstat	simple stats
calmd	recalculate MD/NM tags and '=' bases
merge	merge sorted alignments
rmDup	remove PCR duplicates
reheader	replace BAM header
cat	concatenate BAMs
bedcov	read depth per BED region
targetcut	cut fosmid regions (for fosmid pool only)
phase	phase heterozygotes
bamshuf	shuffle and group alignments by name



Laboratórios de Bioinformática

20

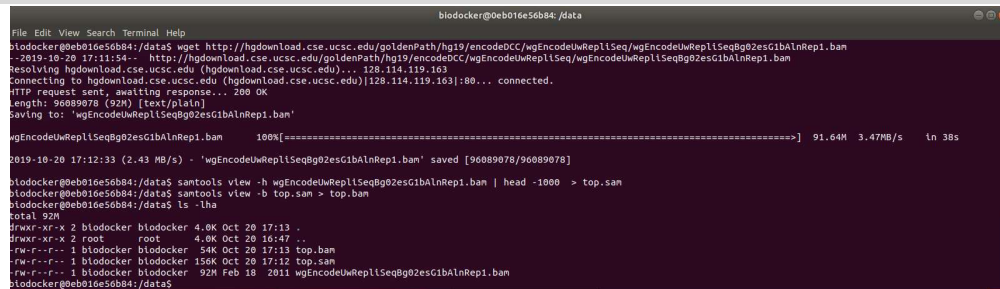
NGS: Alinhamento de sequências – SAM/BAM

- Conversão de um ficheiro SAM para BAM pode ser efetuada com:

```
$ wget
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam

$ samtools view -h wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam | head -1000 > top.sam

$ samtools view -b top.sam > top.bam
```



```
blodocker@0eb016e56b84:/data
blodocker@0eb016e56b84:/data$ wget http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam
--2019-10-20 17:11:54-- http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam
Resolving hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)... 128.114.119.163
Connecting to hgdownload.cse.ucsc.edu (hgdownload.cse.ucsc.edu)|128.114.119.163|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 96089078 (92M) [text/plain]
Saving to: 'wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam'

wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam 100%[=====] 91.64M 3.47MB/s in 38s
2019-10-20 17:12:33 (2.43 MB/s) - 'wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam' saved [96089078/96089078]

blodocker@0eb016e56b84:/data$ samtools view -h wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam | head -1000 > top.sam
blodocker@0eb016e56b84:/data$ samtools view -b top.sam > top.bam
blodocker@0eb016e56b84:/data$ ls -lha
total 92M
drwxr-xr-x 2 blodocker blodocker 4.0K Oct 20 17:13 .
drwxr-xr-x 2 root root 4.0K Oct 20 16:47 ..
-rw-r--r-- 1 blodocker blodocker 54K Oct 20 17:13 top.bam
-rw-r--r-- 1 blodocker blodocker 156K Oct 20 17:12 top.sam
-rw-r--r-- 1 blodocker blodocker 92M Feb 18 2011 wgEncodeUwRepliSeqBg02esG1bAlnRep1.bam
blodocker@0eb016e56b84:/data$
```

Laboratórios de Bioinformática

21

NGS: Mapear reads para o genoma (chr21)

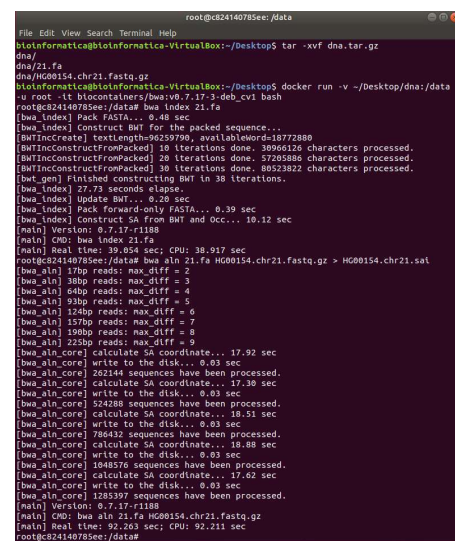
- Neste caso de estudo pretende-se utilizar um dataset de reads (HG00154.chr21) do projeto 1000 Genomes.
- Para efetuar o alinhamento dos reads deve-se criar um index com o algoritmo *Burrows–Wheeler transform* para a sequência de referência.
- Posteriormente, deve-se gerar as coordenadas genómicas para todos os reads (ficheiro .sai).

Comando docker:

```
docker run -it biocontainers/bwa:v0.7.17-3-deb_cv1 bash
```

```
# Build an index for the sequence of chr 21
$ bwa index 21.fa
```

```
# align reads to reference sequence, getting genomic coordinates
$ bwa aln 21.fa HG00154.chr21.fastq.gz > HG00154.chr21.sai
```



```
root@cd24140785ee:/data
bioinformatics@bioinformatics-VirtualBox:~/Desktop$ tar -xvf dna.tar.gz
dna/
dna/21.fa
dna/HG00154.chr21.fastq.gz
bioinformatics@bioinformatics-VirtualBox:~/Desktop$ docker run -v ~/Desktop/dna:/data
-u root -lt biocontainers/bwa:v0.7.17-3-deb_cv1 bash
root@cd24140785ee:/data# bwa index 21.fa
[bwa_index] Pack FASTA... 0.46 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncrConstructFromPacked] 18 iterations done, 38060128 characters processed.
[BWTIncrConstructFromPacked] 20 iterations done, 57205886 characters processed.
[BWTIncrConstructFromPacked] 30 iterations done, 86523822 characters processed.
[Done] Finished constructing BWT in 30 iterations.
[bwa_index] 27.73 seconds elapsed.
[bwa_index] Update BWT... 0.20 sec
[bwa_index] Pack forward-only FASTA... 0.39 sec
[bwa_index] Construct SA from BWT and Occ... 10.12 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index 21.fa
[main] Real time: 39.054 sec; CPU: 38.917 sec
root@cd24140785ee:/data# bwa aln 21.fa HG00154.chr21.fastq.gz > HG00154.chr21.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 30bp reads: max_diff = 3
[bwa_aln] 60bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 223bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 17.92 sec
[bwa_aln_core] write to the disk... 0.03 sec
[bwa_aln_core] 262144 sequences have been processed.
[bwa_aln_core] calculate SA coordinate... 17.30 sec
[bwa_aln_core] write to the disk... 0.03 sec
[bwa_aln_core] 525288 sequences have been processed.
[bwa_aln_core] calculate SA coordinate... 18.51 sec
[bwa_aln_core] write to the disk... 0.03 sec
[bwa_aln_core] 786432 sequences have been processed.
[bwa_aln_core] calculate SA coordinate... 18.88 sec
[bwa_aln_core] write to the disk... 0.03 sec
[bwa_aln_core] 1048576 sequences have been processed.
[bwa_aln_core] calculate SA coordinate... 17.62 sec
[bwa_aln_core] write to the disk... 0.03 sec
[bwa_aln_core] 1285397 sequences have been processed.
[main] Version: 0.7.17-r1188
[main] Cpu: bwa aln 21.fa HG00154.chr21.fastq.gz
[main] Real time: 92.263 sec; CPU: 92.211 sec
root@cd24140785ee:/data#
```

22

NGS: Gerar o ficheiro BAM com alinhamentos

```
# Criar o ficheiro sam a partir dos reads
mapeados
$ bwa samse -f HG00154.chr21.aln.sam 21.fa
HG00154.chr21.sai HG00154.chr21.fastq.gz
# Converter o sam em bam
$ samtools view -Sb HG00154.chr21.aln.sam >
HG00154.chr21.aln.bam
```

```
root@6724487c2505:/data
File Edit View Search Terminal Help
bioinformatica@bioinformatica-VirtualBox:~/Desktop$ docker run -v ~/Desktop/dna:/data
-u root -it biocontainers/samtools:v1.7.0_cv4 bash
root@6724487c2505:/data# samtools view -Sb HG00154.chr21.aln.sam > HG00154.chr21.aln.
bam
root@6724487c2505:/data# ls -lha
total 640M
-rwxr-xr-x 2 root root 4.0K Oct 20 22:19 ..
-rwxr-xr-x 2 root root 4.0K Oct 20 22:18 .
-rw-r--r-- 1 root root 47M May 17 2016 21.fa
-rw-r--r-- 1 root root 281 Oct 20 22:08 21.fa.amb
-rw-r--r-- 1 root root 40 Oct 20 22:08 21.fa.ann
-rw-r--r-- 1 root root 46M Oct 20 22:08 21.fa.bwt
-rw-r--r-- 1 root root 12M Oct 20 22:08 21.fa.pac
-rw-r--r-- 1 root root 23M Oct 20 22:09 21.fa.sa
-rw-r--r-- 1 root root 95M Oct 20 22:19 HG00154.chr21.aln.bam
-rw-r--r-- 1 root root 309M Oct 20 22:14 HG00154.chr21.aln.sam
-rw-r--r-- 1 root root 76M May 17 2016 HG00154.chr21.fastq.gz
-rw-r--r-- 1 root root 33M Oct 20 22:11 HG00154.chr21.sai
root@6724487c2505:/data#
```

Comando docker:
docker run -it biocontainers/samtools:v1.7.0_cv4 bash

```
root@c824140785ee:/data
File Edit View Search Terminal Help
root@c824140785ee:/data# bwa samse -f HG00154.chr21.aln.sam 21.fa
HG00154.chr21.sai HG00154.chr21.fastq.gz
[bwa_aln_core] convert to sequence coordinate... 1.06 sec
[bwa_aln_core] refine gapped alignments... 0.18 sec
[bwa_aln_core] print alignments... 0.60 sec
[bwa_aln_core] 262144 sequences have been processed.
[bwa_aln_core] convert to sequence coordinate... 1.18 sec
[bwa_aln_core] refine gapped alignments... 0.17 sec
[bwa_aln_core] print alignments... 0.61 sec
[bwa_aln_core] 524288 sequences have been processed.
[bwa_aln_core] convert to sequence coordinate... 1.07 sec
[bwa_aln_core] refine gapped alignments... 0.19 sec
[bwa_aln_core] print alignments... 0.62 sec
[bwa_aln_core] 786432 sequences have been processed.
[bwa_aln_core] convert to sequence coordinate... 1.03 sec
[bwa_aln_core] refine gapped alignments... 0.16 sec
[bwa_aln_core] print alignments... 0.58 sec
[bwa_aln_core] 1048576 sequences have been processed.
[bwa_aln_core] convert to sequence coordinate... 0.90 sec
[bwa_aln_core] refine gapped alignments... 0.14 sec
[bwa_aln_core] print alignments... 0.52 sec
[bwa_aln_core] 1285397 sequences have been processed.
[main] Version: 0.7.17-r1188
[main] CMD: bwa samse -f HG00154.chr21.aln.sam 21.fa HG00154.chr21
.sai HG00154.chr21.fastq.gz
[main] Real time: 11.277 sec; CPU: 10.943 sec
root@c824140785ee:/data#
```

23

NGS: Ordenar os reads para gerar estatísticas

```
# Ordenar os reads por coordenadas genómicas
$ samtools sort HG00154.chr21.aln.bam -o HG00154.chr21.aln.sort.bam

# Criar um índice com o ficheiro bam
$ samtools index HG00154.chr21.aln.sort.bam

# Obter as estatísticas do alinhamento
$ samtools flagstat HG00154.chr21.aln.sort.bam
```

```
root@6724487c2505:/data
File Edit View Search Terminal Help
root@6724487c2505:/data# samtools sort HG00154.chr21.aln.bam -o HG00154.chr21.aln.sort.bam
root@6724487c2505:/data# samtools index HG00154.chr21.aln.sort.bam
root@6724487c2505:/data# samtools flagstat HG00154.chr21.aln.sort.bam
1285397 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
1143664 + 0 mapped (88.97% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
root@6724487c2505:/data#
```

24