

# CURSO PRÁTICO DE BIOINFORMÁTICA:

Aula 2 – Linux – Cont.

Renan Paulo Martin

17/07/19

# Configurar Host SSH

- Criar o arquivo
  - `~/.ssh/config`
  - Incluir as linhas abaixo substituindo o `USER` pelo usuário (primeiro nome)

`Host bioinfo`

`HostName 172.22.169.15`

`User USER`

`ControlMaster auto`

`ControlPath ~/.ssh/control:%h:%p:%r`

`ssh bioinfo`

# Script

- Utilizado para automatizar uma tarefa
- Conjunto de instruções
- Execução de linha a linha
- Existem linguagens específicas de Script
- Normalmente são interpretadas
- Exemplos
  - Python
  - Perl
  - Java Script
  - Bash

# Bash Script

- Interpretador de comandos Bash
  - Já foi utilizado na aula anterior
- Um comando por linha
  - “\” barra invertida para pular a linha e continuar o mesmo comando
- Utilizar o comando complete com seus argumentos
- No final da execução há o encerramento da aplicação

# Comandos Bash Script

- sleep 10
  - Interrompe a execução pelo tempo especificado
- read gene
  - Lê uma variável
- &
  - Executa o comando em segundo plano
  - continua a execução do Script
- exit
  - Encerra o programa
- #
  - Tag usada para que o interpretador ignore a partir desse ponto

# Primeiro Programa

- Escreva um Script solicitando ao usuário que digite seu nome em seguida exiba uma mensagem de boas vindas ao curso de Bioinformática.

# Primeiro Programa

- Escreva um Script solicitando ao usuário que digite seu nome em seguida exiba uma mensagem de boas vindas ao curso de Bioinformática.

```
#!/bin/bash
```

```
echo "Digite seu nome:"  
read nome  
echo "Olá $nome seja bem vindo ao curso de  
Bioinformática"
```

# Segundo Programa

- Calcular área do triângulo
  - Recebe valor da base
  - Recebe valor da altura
  - Calcula base \* altura / 2
  - Exibe o resultado

```
#!/bin/bash
```

```
echo "Informe a base do triângulo:"
```

```
read base
```

```
echo "Informe a altura do triângulo:"
```

```
read altura
```

```
area=$((base * altura / 2))
```

```
echo $area
```

```
#!/bin/bash
```

```
echo "Informe a base do triângulo:"
```

```
read base
```

```
echo "Informe a altura do triângulo:"
```

```
read altura
```

```
area=$((base * altura / 2))
```

```
echo $area
```

```
#!/bin/bash
```

```
echo "Informe a base do triângulo:"
```

```
read base
```

```
echo "Informe a altura do triângulo:"
```

```
read altura
```

```
area=$((base * altura / 2))
```

```
echo $area
```

```
#!/bin/bash

echo "Informe a base do triângulo:"
read base
echo "Informe a altura do triângulo:"
red read altura
area=$((base * altura / 2))
echo $area
```

```
#!/bin/bash

echo "Informe a base do triângulo:"
read base
echo "Informe a altura do triângulo:"
read altura
area=$((base * altura / 2))
echo $area
```

```
#!/bin/bash

echo "Informe a base do triângulo:"
read base

echo "Informe a altura do triângulo:"
read altura

area=$((base * altura / 2))

echo $area
```

# Execução

- Os programas somente serão executados enquanto a sessão for mantida
- Caso a sessão seja encerrada o programa será encerrado no ponto em que houve o encerramento
- Não é possível retomar o programa no ponto em que foi encerrado
- As variáveis somente existirão durante a execução do programa
- Existem programas que criam sessões virtuais permanentes

# Screen

- Utilizado para a criação de multiplas instâncias de terminal
- screen
  - Inicia uma nova instância
- Ctrl + a + d
  - Retorna para a sessão principal
- screen –ls
  - Lista todas as sessões abertas
- screen –r
  - Retorna para uma instância virtual
- exit
  - Encerra uma sessão

# Terceiro Programa

```
#!/bin/bash
```

```
echo "Olá vou cochilar por 10 segundos"  
sleep 10  
echo "Agora vou cochilar por 100 segundos"  
sleep 100  
echo "Agora serão 200 segundos"  
sleep 200  
echo "Cansei, deixa eu dormir mais 500 segundos"  
sleep 500  
echo "Pronto, agora já posso dormir"
```

# CURSO PRÁTICO DE BIOINFORMÁTICA:

Aula 3 - Bioinformática

Renan Paulo Martin

17/07/19

# Bioinformática

- Refere-se ao emprego da informática na resolução de problemas biológicos
- Bigdata
- Grande aplicação para os estudos ômicos
- Área em ascenção

## Tecnologia

## Conhecimento

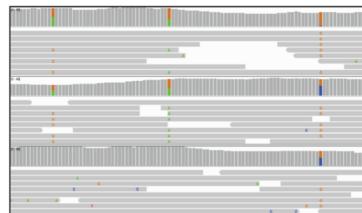
### Laboratório



Preparação  
de amostra      Sequenciamento



### Informática



Mapeamento e  
Descoberta de variantes

### Anotações



**OMIM®**

Literatura      Base de  
dados

### Interpretações

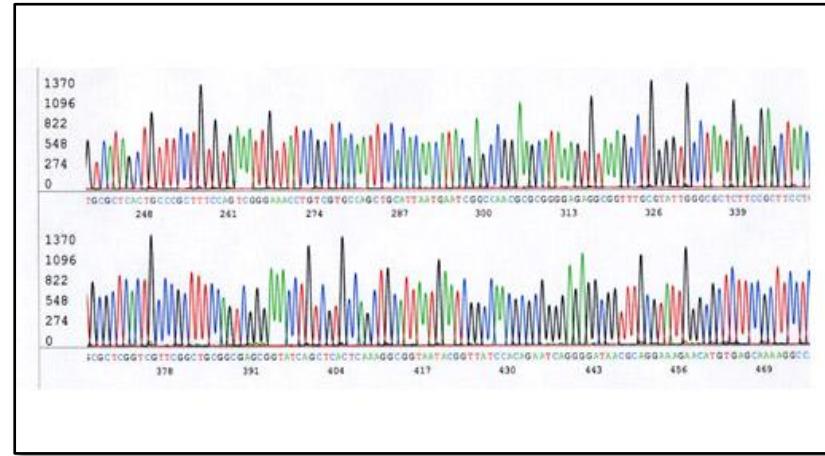


Análise e report

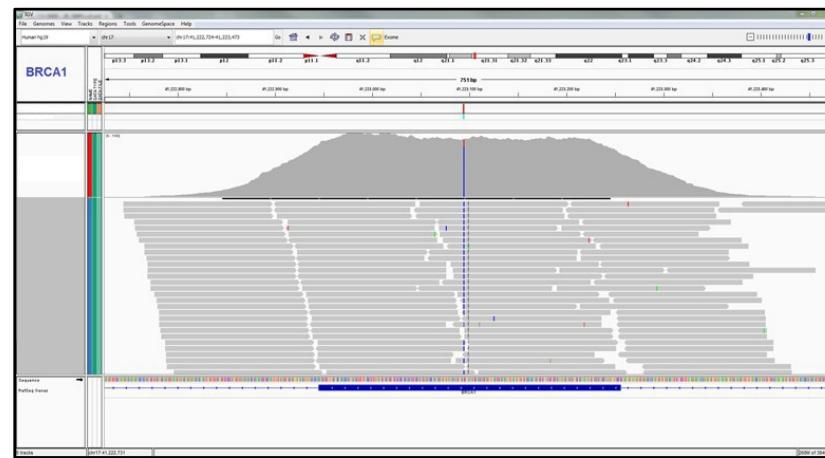
# *Histórico da Tecnologia de Sequenciamento*



1977 – “Sanger”



1986 – Automated DNA Sequencing



2005 – Next Generation DNA Sequencing

# *Comparação das tecnologias*

## **Next Gen**

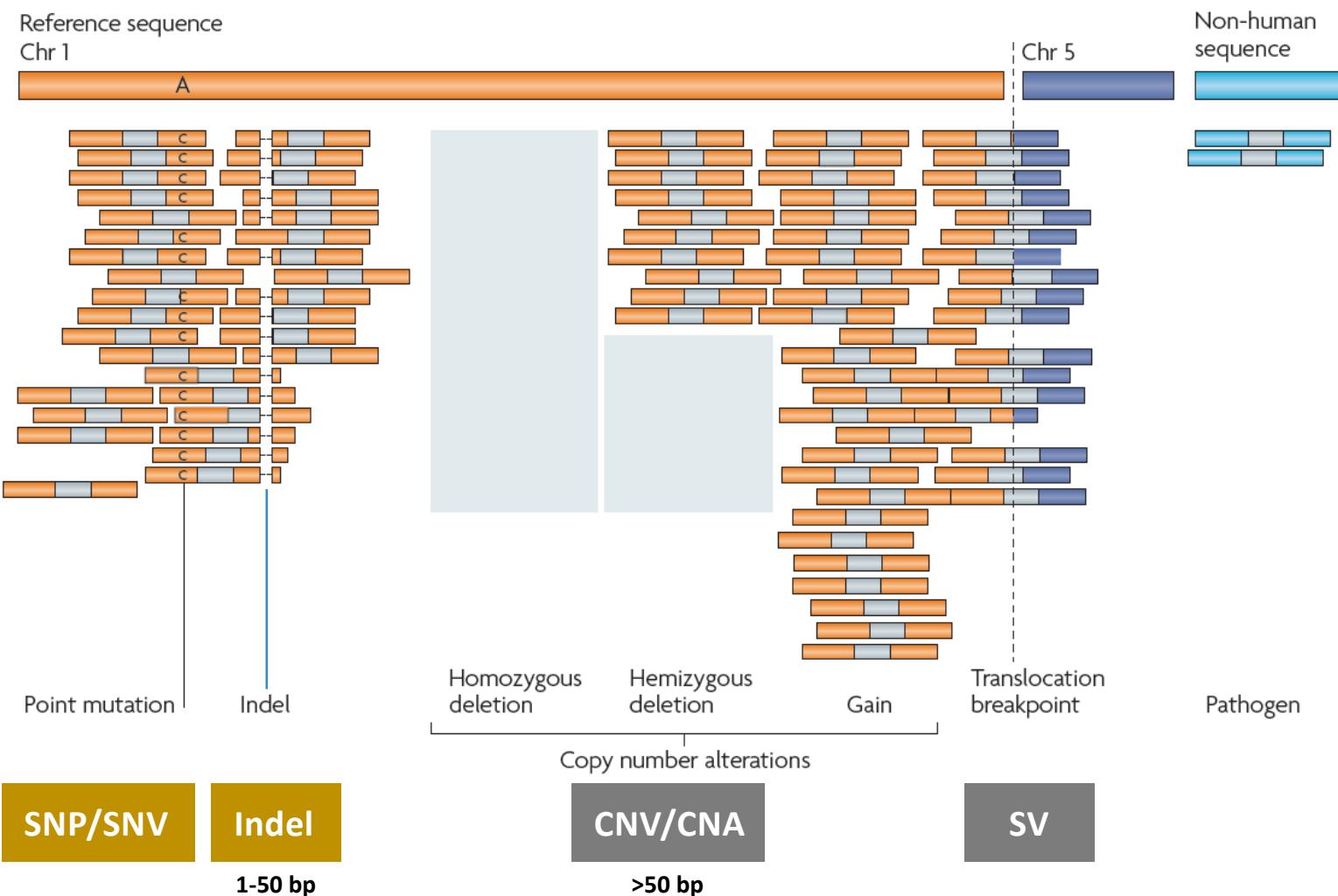
- High coverage, low cost\*
- Fast; High throughput
- Short reads: 100-150 bp
- Sequence by synthesis
- Massively parallel
- Detect SNVs / Indels

## **Sanger**

- Low coverage, high cost\*
- Slow; Low throughput
- Average read: 800-1000 bp
- Chain termination
- Serial processing
- Repeat expansions

\* Large targets (e.g. gene panels, exomes, whole genomes)

# Tipos de Variantes



# *Custo do Genoma Humano*

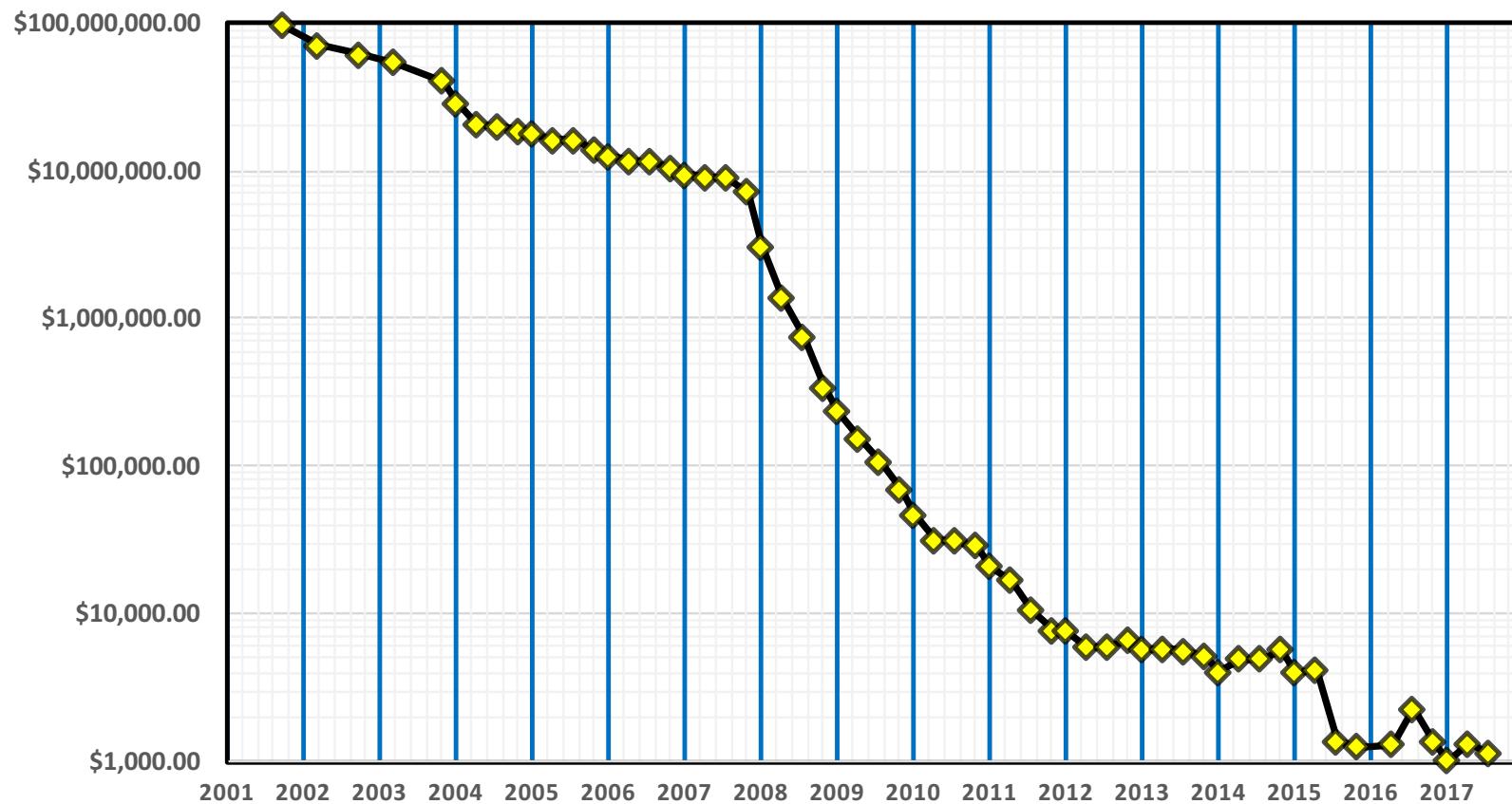
## **Human Genome Project**

- 1 genoma
- \$3 bilhões USD
- 13 anos de duração
- 3.3 Gb de data

## **Next Gen Sequencing**

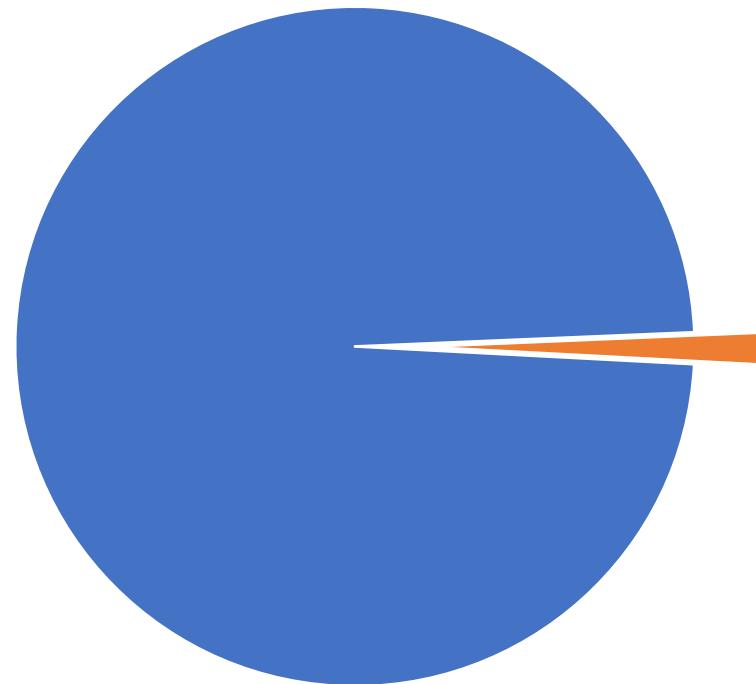
- >1 genoma
- \$1,000 USD
- 1-2 dias/corrida
- >1 Tb de data/corrida!

# Custo por genoma



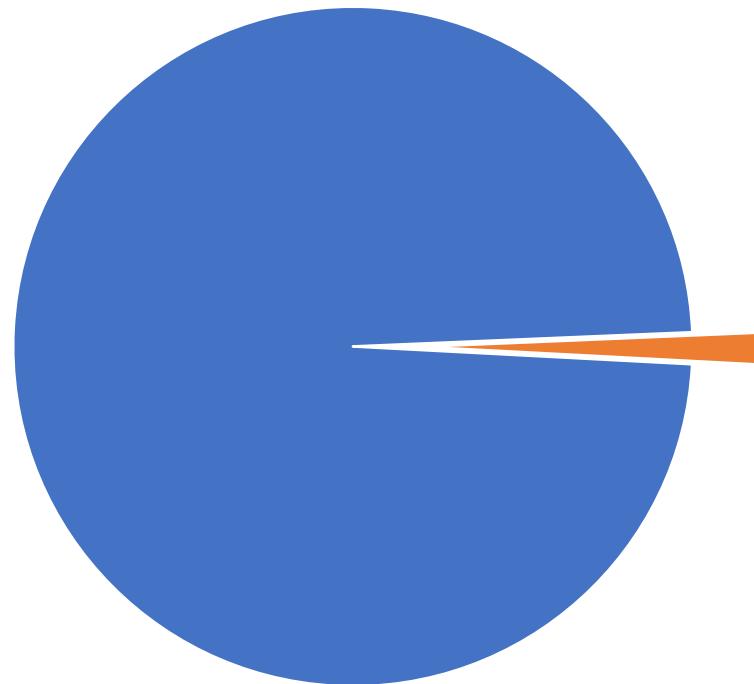
Wetterstrand. - NHGRI Genome Sequencing Program (GSP) . 2019. [www.genome.gov/sequencingcostsdata](http://www.genome.gov/sequencingcostsdata)

# Exoma vs Genoma



■ Não Codificante ■ Codificante

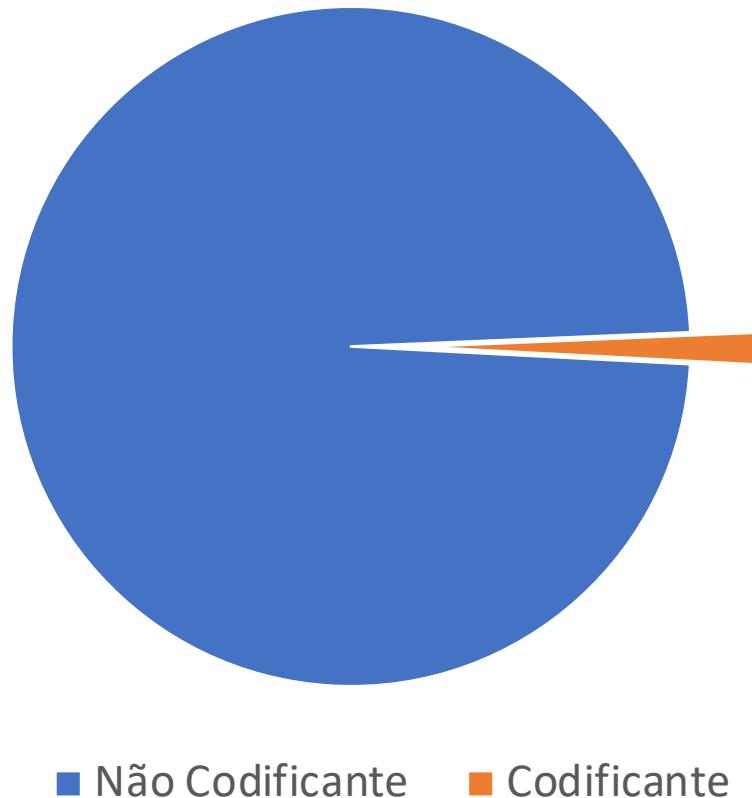
# Exoma vs Genoma



Genoma Completo

≥ 3 bilhões de bases  
região codificante e  
região não codificante

# Exoma vs Genoma



Genoma Completo

≥ 3 bilhões de bases  
região codificante e  
região não codificante

Exoma Completo

1,5% do Genoma  
região codificante  
abra 85% das  
mutações associadas a  
doenças

# Exoma vs Genoma

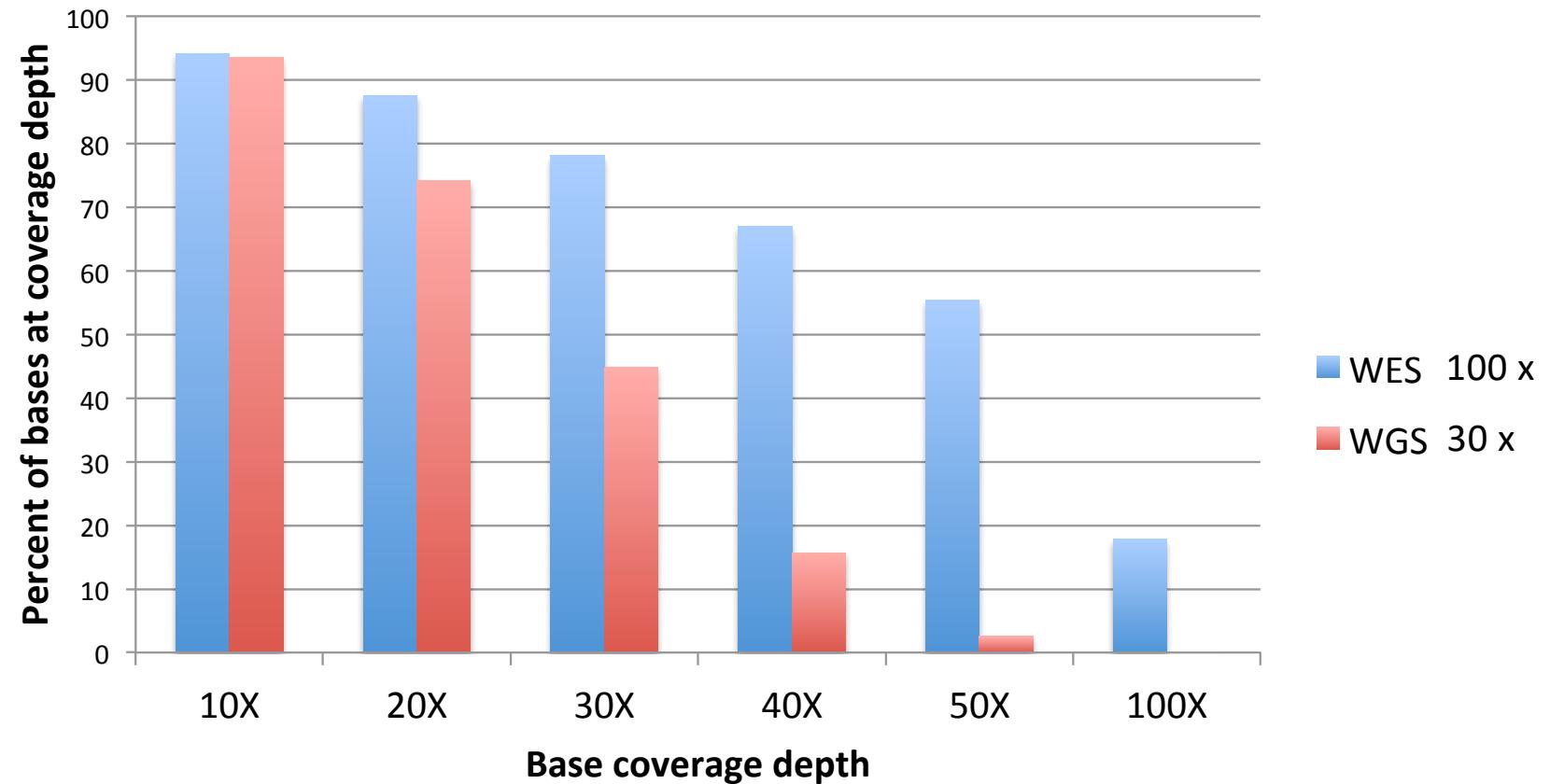
## Genoma

- Genoma completo é preparado
- PCR-free
- Custo maior
- Exons e introns
- Cobertura quase completa

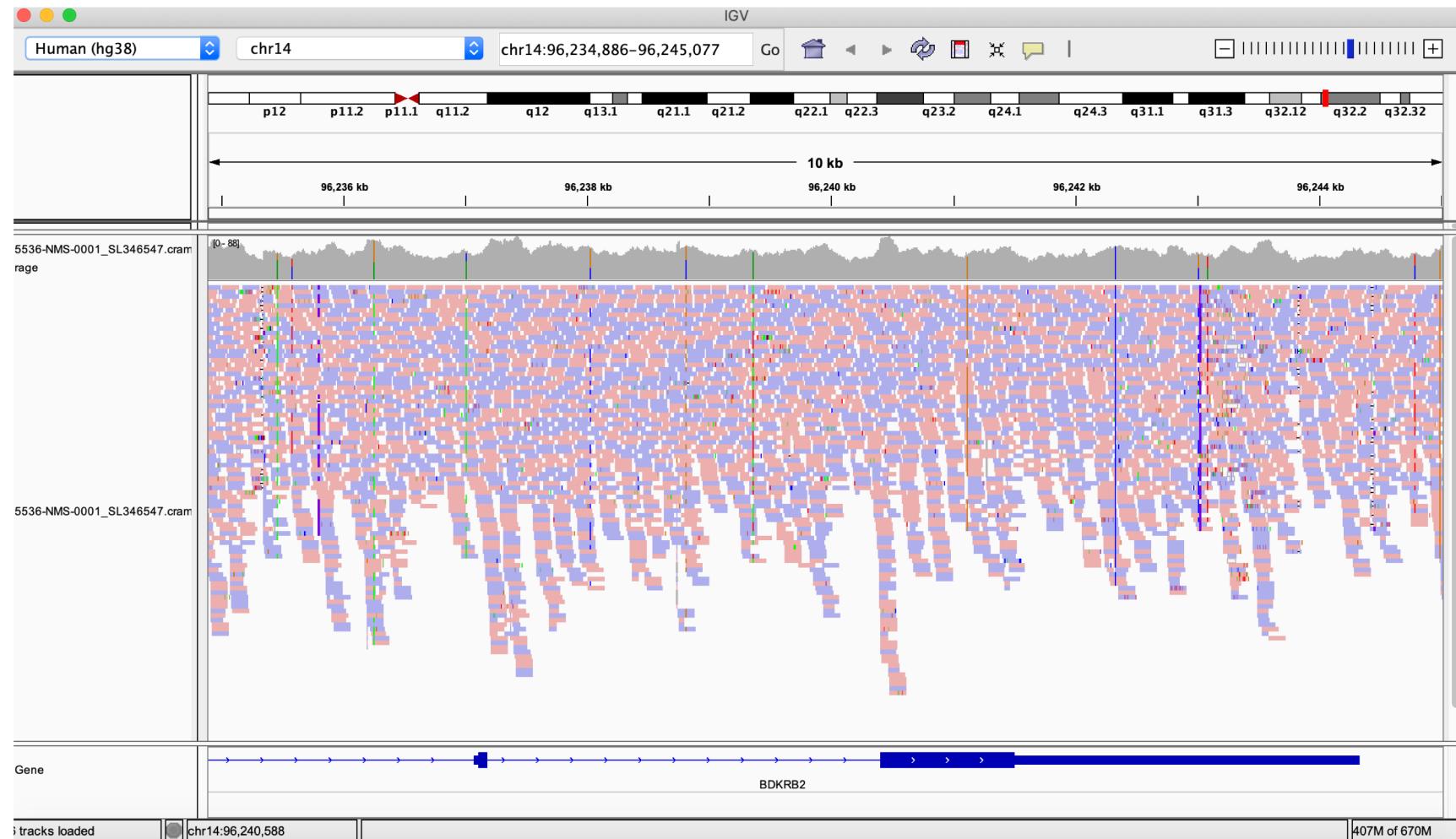
## Exoma

- Captura de regiões alvo
- Requer PCR
- Custo menor
- Somente exons (regiões flankeadoras)
- Cobertura somente nas regiões alvo

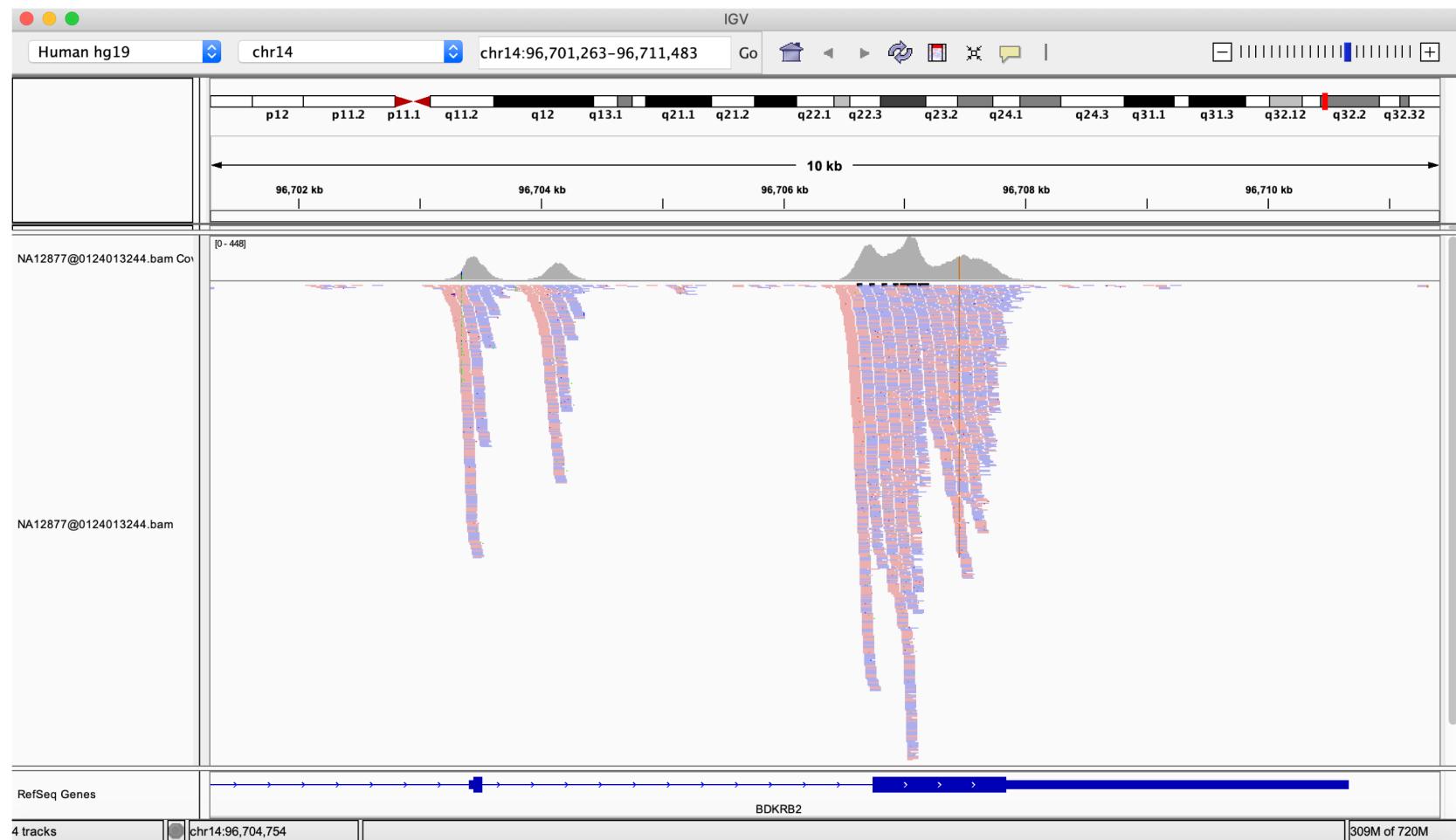
# Métrica típica



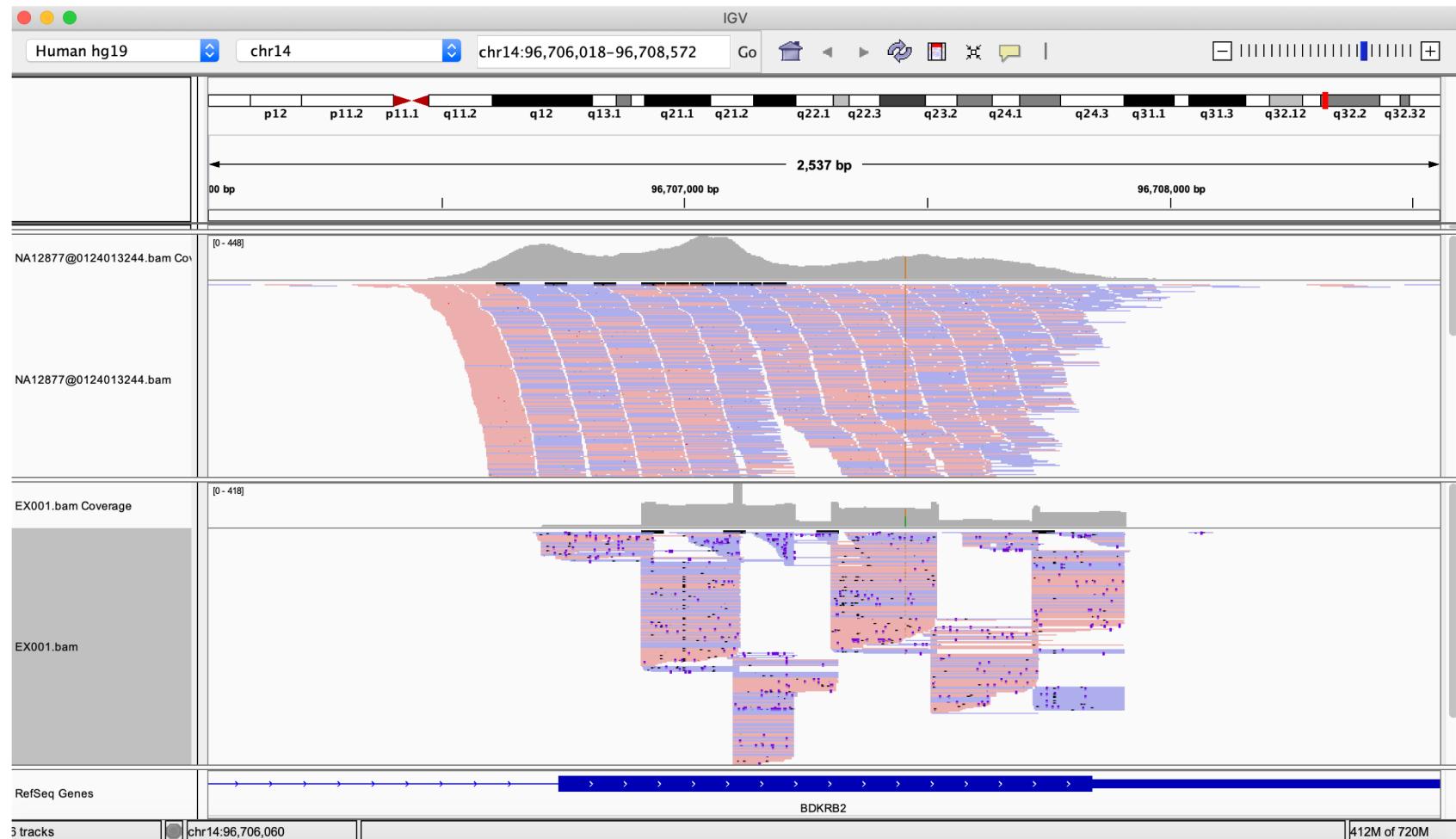
# WGS



# WES



# ILLUMINA vs ION TORRENT



# FASTQ

- Arquivo de texto contendo as bases determinadas no sequenciamento e suas respectivas qualidades;
- Formato utilizado como entrada para geração do arquivo SAM/BAM.

- Simple extension from traditional FASTA format.
- Each block has 4 elements ( in 4 lines):
  - Sequence Name (read name, group, etc.)
  - Sequence
  - + (optional: Sequence name again)
  - Associated quality score.
- Example record:

@EAS54_6_R1_2_1_413_324	Identifier
CCCTTCTTGTCTTCAGCGTTCTCC	Sequence
+	
;;3;;;;;;7;;;;;88	Base Qualities (ASCII 33 + Phred scaled Q)

Official specification in <http://maq.sourceforge.net/fastq.shtml>

# SAM/BAM /CRAM

- Arquivo de texto (SAM – mapa de alinhamento de sequência - do inglês Sequence alignment map) ou binário (BAM – SAM binário - do inglês binary SAM) contendo as bases sequenciadas alinhadas ao genoma de referência.
- Formato utilizado como entrada para a realização da chamada de variantes para a geração do arquivo VCF.

```

@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:247249719
@SQ SN:chr2 LN:242951149
[cut for clarity]
@SQ SN:chr9 LN:140273252
@SQ SN:chr10 LN:135374737
@SQ SN:chr11 LN:134452384
[cut for clarity]
@SQ SN:chr22 LN:49691432
@SQ SN:chrX LN:154913754
@SQ SN:chrY LN:57772954
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI
@PG ID:BWA VN:0.5.7 CL:tk
@PG ID:GATK PrintReads VN:1.0.2864

```

20FUKAAXX100202:1:1:12730:189900 163 chrM 1 60 101M = 282 381

GATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCATTGGTA...[more bases]

?BA@A>BBBBACBAC@BBCBBCBC@BC@CAC@:BBCBBCACAACBABCBCAB...[more quals]

RG:Z:20FUK.1 NM:i:1 AM:i:37 MD:Z:72G28 MQ:i:60 PG:Z:BWA UQ:i:33

Required: Standard header

Essential: contigs of aligned reference sequence. Should be in karyotypic order.

Essential: read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

Useful: Data processing tools applied to the reads

Official specification in <http://samtools.sourceforge.net/SAM1.pdf>

# VCF/BCF

- Arquivo de texto (VCF – formato de chamada de variantes - do inglês variant call format) ou binário (BCF – VCF binário - do inglês binary VCF) contendo uma lista das variantes encontradas no sequenciamento.
- Formato utilizado como entrada para anotação das variantes encontradas na geração de tabelas que serão utilizadas para interpretação do sequenciamento.

```

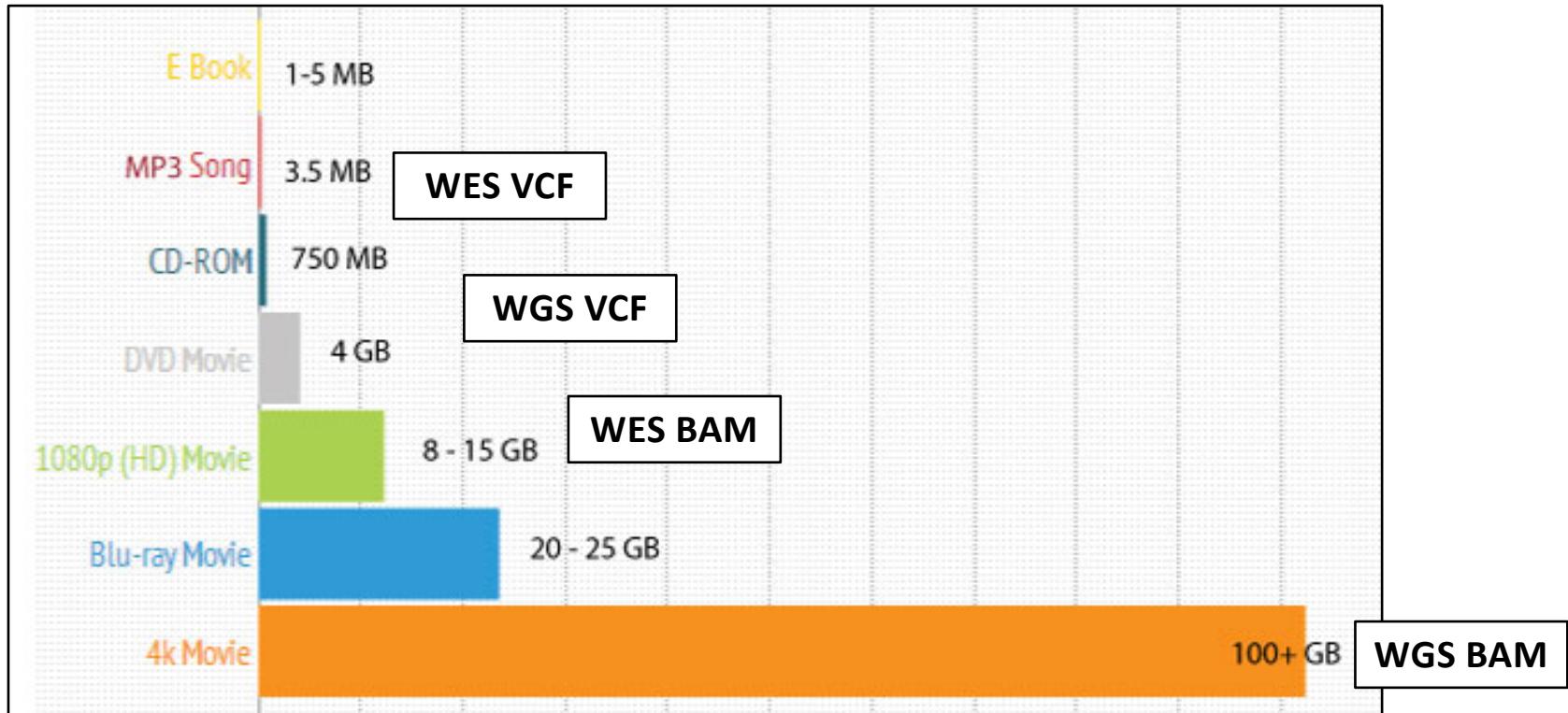
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,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=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:..
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3

```

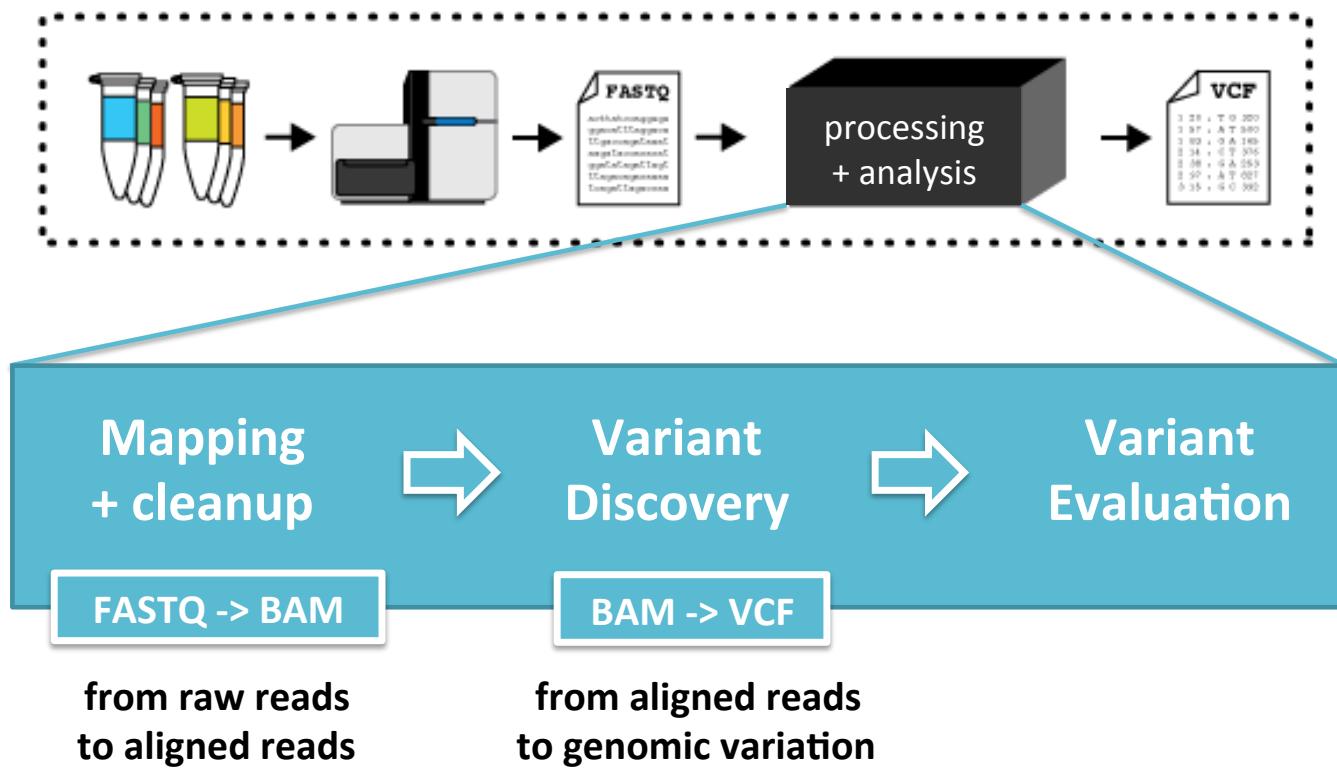
# TXT/CSV/XLS

- Arquivo de texto (TXT – texto; CSV – valores separados por vírgula - do inglês comma-separated values; e TSV – valores separados por tabulação - do inglês tab-separated values) ou planilha (XLS – formato de planilhas do programa Microsoft Excel) contendo as anotações das variantes encontradas no sequenciamento.
- Formato utilizado na análise e interpretação do sequenciamento, bem como para criação dos laudos e relatórios desse exame genético.

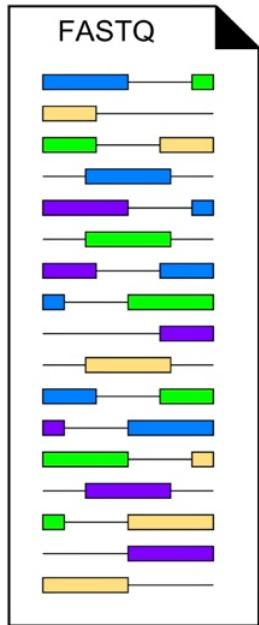
# Data File Sizes



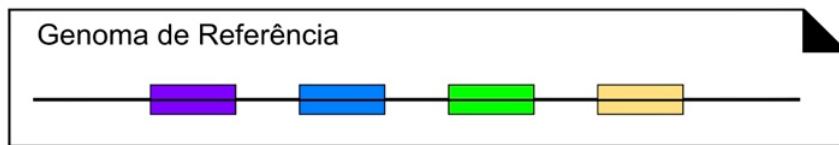
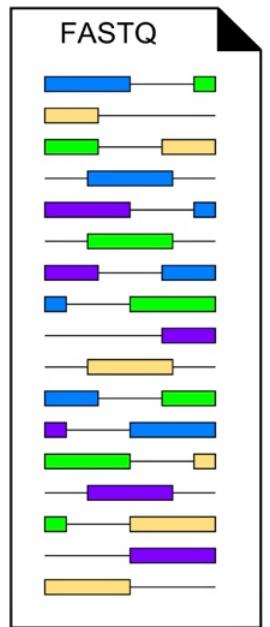
<http://filecatalyst.com/todays-media-file-sizes-whats-average/>



# Pipeline DNaseq



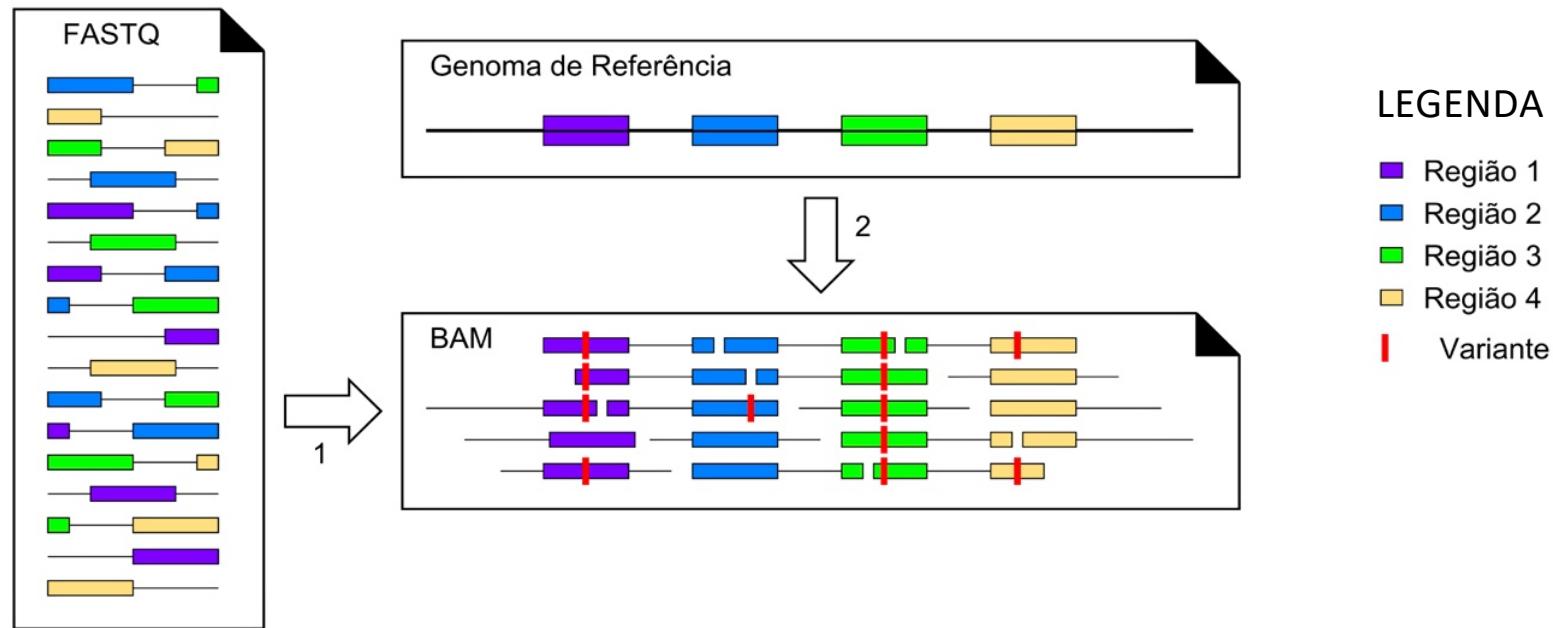
# Pipeline DNaseq



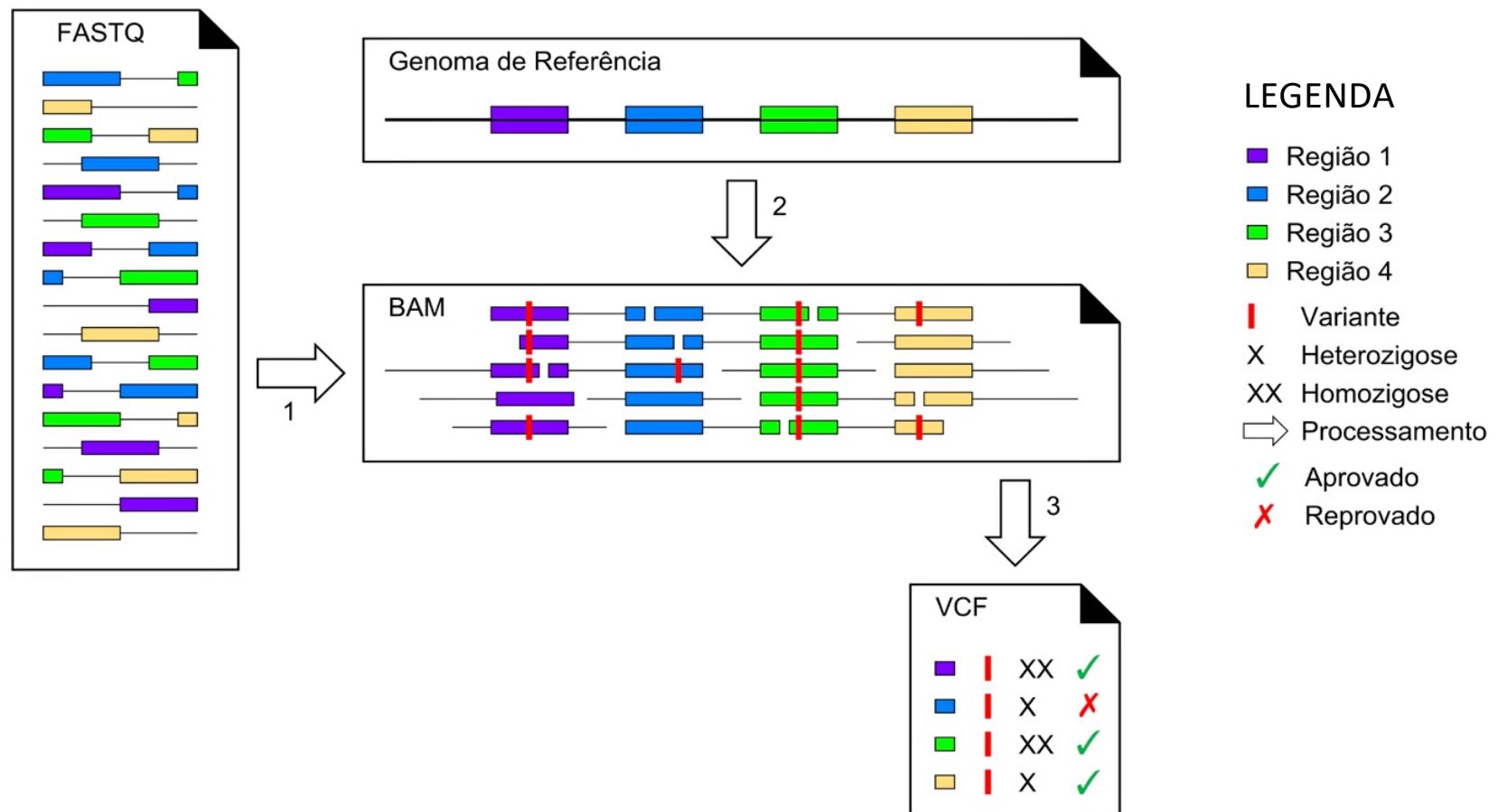
## LEGENDA

- Região 1
- Região 2
- Região 3
- Região 4

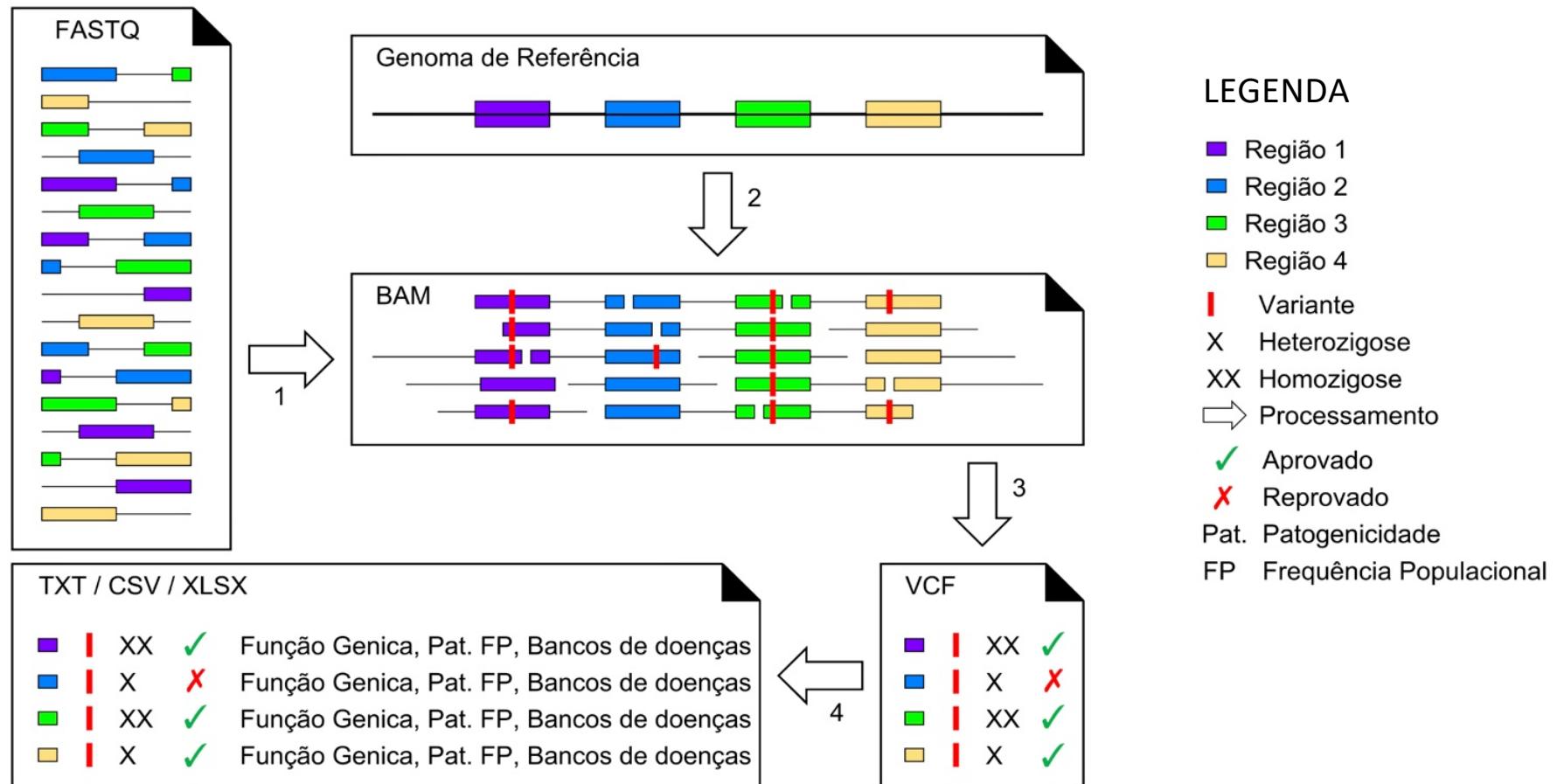
# Pipeline DNaseq



# Pipeline DNASeq



# Pipeline DNASeq



# Genoma de Referência

- Existem diferentes versões
  - HG19 ou GRCh37 (2009)
  - HG38 ou GRCh38 (2013)
- HG38 é o mais recente, possui mais GAPS, porém menos “N”
- HG19 é atualmente o mais usado, porém a maioria dos grupos de estudo e das bases de dados estão migrando para a nova versão

# Genoma de Referência

<http://hgdownload.cse.ucsc.edu/downloads.html#human>

- LiftOver para converter as coordenadas de uma versão para outra