Tutorial Unix e linha de comandos

Ministério da Saúde

Instituto_Nacional de Saúde
Doutor Ricardo Jorge

Análise computacional e bioinformática de variantes em doença genética

10 a 13 Out. 2023
nas instalações do Instituto Ricardo Jorge, em Lisboa

Este curso, de natureza teórico-prática, dá a conhecer as várias etapas envolvidas na análise de variantes de linha germinativa associadas a doença genética, em paralelo com a análise prática de casos reais.

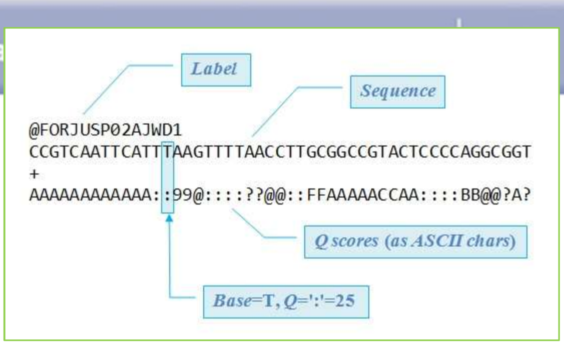
Destinatários: Profissionais de saúde, investigadores e estudantes de mestrado ou doutoramento, que estejam envolvidos em atividades de diagnóstico ou investigação no contexto de estudo de variantes de linha germinativa associadas a doença genética

Formadores: Luís Vieira, José Ferrão, Hugo Martiniano e Daniel Sobral

Coordenação: Luís Vieira

Instituto_Nacional de Saúde
Doutor Ricardo Jorge

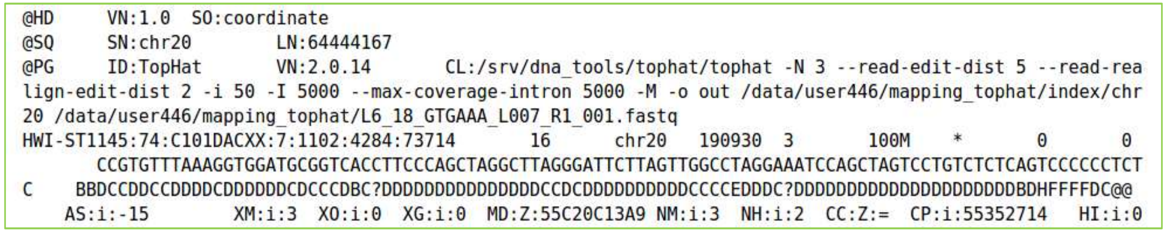
basecalling



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

.fastq

mapping

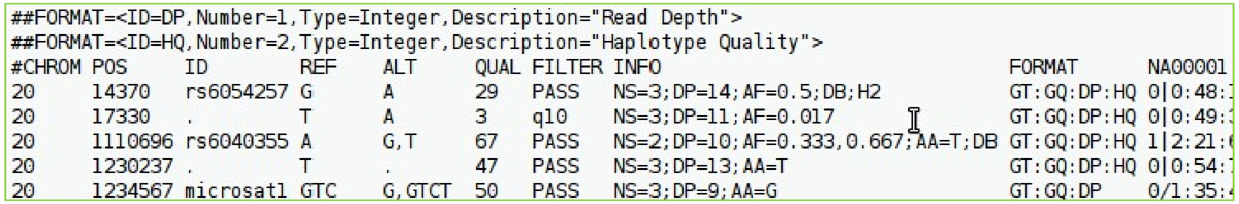


```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714    HI:i:0
```

.sam/.bam/.cram

pre variant calling (BQSR, MarkDup)

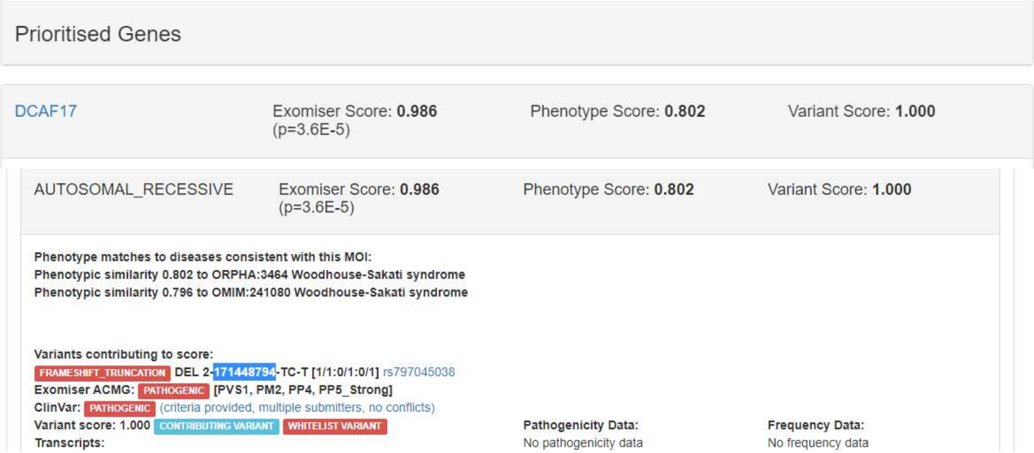
variant calling



```
##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
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3
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
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4
```

.vcf

variant annotation/ priorization



Prioritised Genes

DCAF17	Exomiser Score: **0.986** (p=3.6E-5)	Phenotype Score: **0.802**	Variant Score: **1.000**
AUTOSOMAL_RECESSIVE	Exomiser Score: **0.986** (p=3.6E-5)	Phenotype Score: **0.802**	Variant Score: **1.000**

Phenotype matches to diseases consistent with this MOI:
Phenotypic similarity 0.802 to ORPHA:3464 Woodhouse-Sakati syndrome
Phenotypic similarity 0.796 to OMIM:241080 Woodhouse-Sakati syndrome

Variants contributing to score:
FRAMESHIFT_TRUNCATION **DEL 2-171448794-TC-T [1/1:0/1:0/1]** rs797045038
Exomiser ACMG: PATHOGENIC **[PVS1, PM2, PP4, PP5_Strong]**
ClinVar: PATHOGENIC (criteria provided, multiple submitters, no conflicts)
Variant score: 1.000 CONTRIBUTING VARIANT WHITELIST VARIANT
Transcripts:

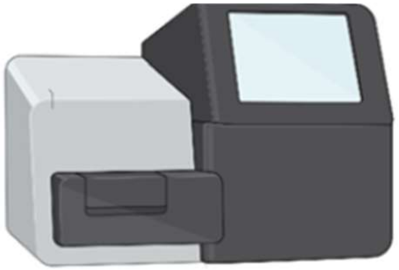
Pathogenicity Data:
No pathogenicity data

Frequency Data:
No frequency data

Instituto_Nacional de Saúde 1899 www.insa.pt

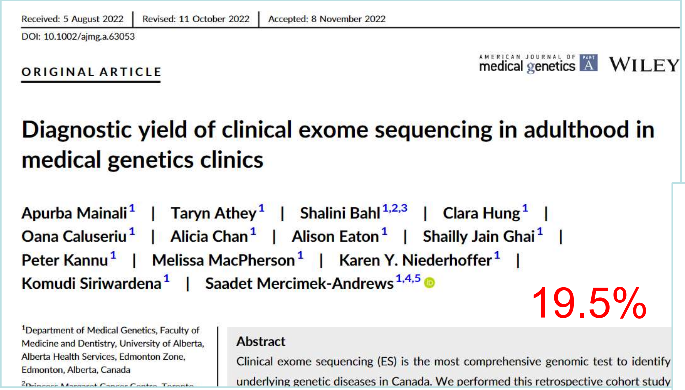
Clinical exome sequencing

Departamento Genética Humana



Instituto_Nacional de Saúde
Doutor Ricardo Jorge

- Exome sequencing:
  - diagnosis of genetic disorders
  - discovery of new Mendelian-disease genes
  - **Clinical exome sequencing (CES)** - genes associated to clinical phenotypes



Received: 5 August 2022 | Revised: 11 October 2022 | Accepted: 8 November 2022
DOI: 10.1002/ajmg.a.63053

ORIGINAL ARTICLE

AMERICAN JOURNAL OF medical genetics PART A WILEY

Diagnostic yield of clinical exome sequencing in adulthood in medical genetics clinics

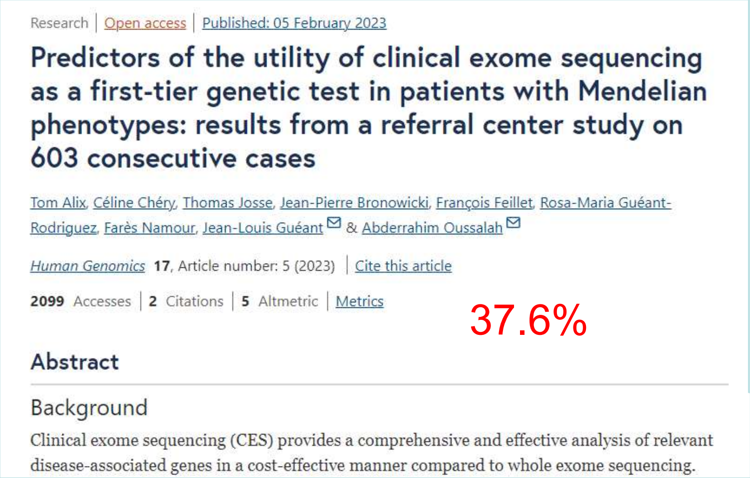
Apurba Mainali[1] | Taryn Athey[1] | Shalini Bahl[1,2,3] | Clara Hung[1] | Oana Caluseriu[1] | Alicia Chan[1] | Alison Eaton[1] | Shailly Jain Ghai[1] | Peter Kannu[1] | Melissa MacPherson[1] | Karen Y. Niederhoffer[1] | Komudi Siriwardena[1] | Saadet Mercimek-Andrews[1,4,5]

[1]Department of Medical Genetics, Faculty of Medicine and Dentistry, University of Alberta, Alberta Health Services, Edmonton Zone, Edmonton, Alberta, Canada
[2]Princess Margaret Cancer Centre, Toronto

19.5%

Abstract

Clinical exome sequencing (ES) is the most comprehensive genomic test to identify underlying genetic diseases in Canada. We performed this retrospective cohort study



Research | Open access | Published: 05 February 2023

Predictors of the utility of clinical exome sequencing as a first-tier genetic test in patients with Mendelian phenotypes: results from a referral center study on 603 consecutive cases

Tom Alix, Céline Chéry, Thomas Josse, Jean-Pierre Bronowicki, François Feillet, Rosa-Maria Guéant-Rodriguez, Farès Namour, Jean-Louis Guéant ✉ & Abderrahim Oussalah ✉

Human Genomics **17**, Article number: 5 (2023) | Cite this article

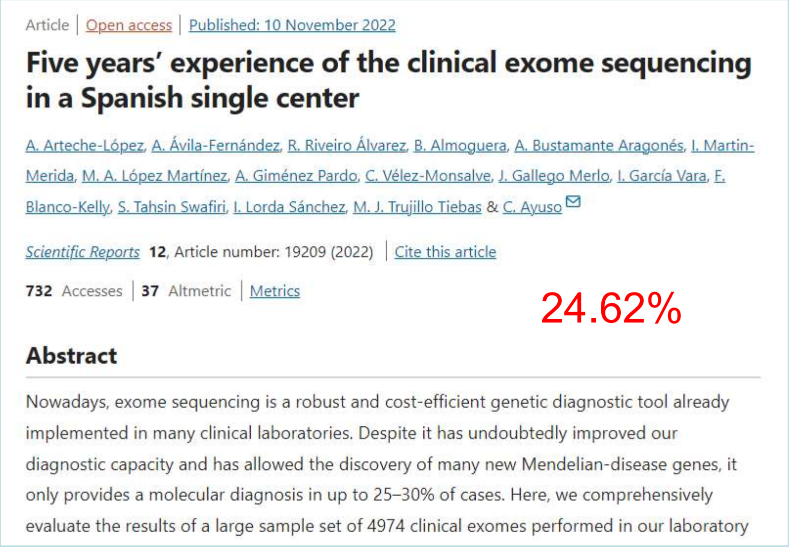
2099 Accesses | **2** Citations | **5** Altmetric | Metrics

37.6%

Abstract

Background

Clinical exome sequencing (CES) provides a comprehensive and effective analysis of relevant disease-associated genes in a cost-effective manner compared to whole exome sequencing.



Article | Open access | Published: 10 November 2022

Five years' experience of the clinical exome sequencing in a Spanish single center

A. Arteche-López, A. Ávila-Fernández, R. Riveiro Álvarez, B. Almoguera, A. Bustamante Aragonés, I. Martin-Merida, M. A. López Martínez, A. Giménez Pardo, C. Vélez-Monsalve, J. Gallego Merlo, I. García Vara, F. Blanco-Kelly, S. Tahsin Swafiri, I. Lorda Sánchez, M. J. Trujillo Tiebas & C. Ayuso ✉

Scientific Reports **12**, Article number: 19209 (2022) | Cite this article

732 Accesses | **37** Altmetric | Metrics

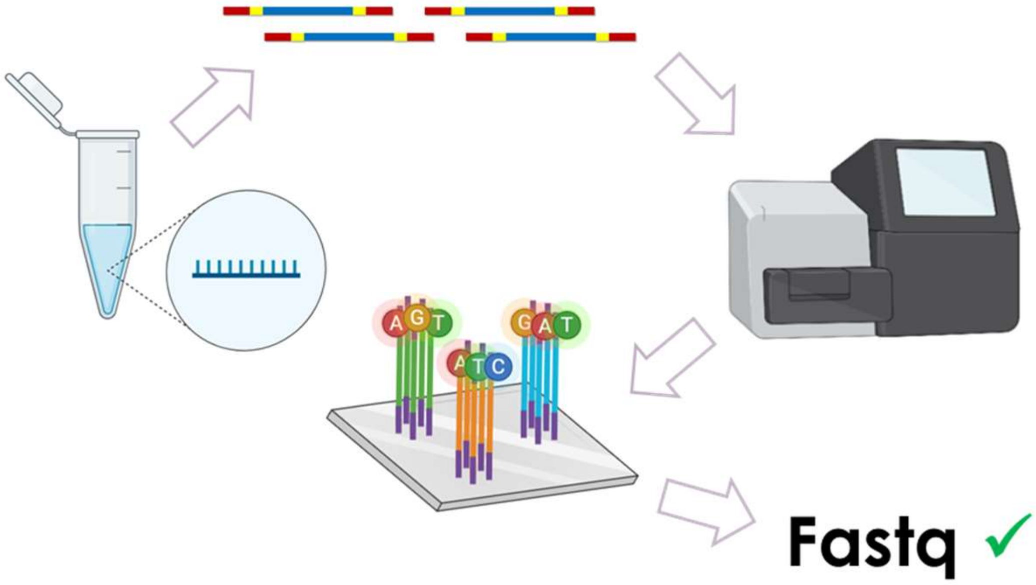
24.62%

Abstract

Nowadays, exome sequencing is a robust and cost-efficient genetic diagnostic tool already implemented in many clinical laboratories. Despite it has undoubtedly improved our diagnostic capacity and has allowed the discovery of many new Mendelian-disease genes, it only provides a molecular diagnosis in up to 25–30% of cases. Here, we comprehensively evaluate the results of a large sample set of 4974 clinical exomes performed in our laboratory

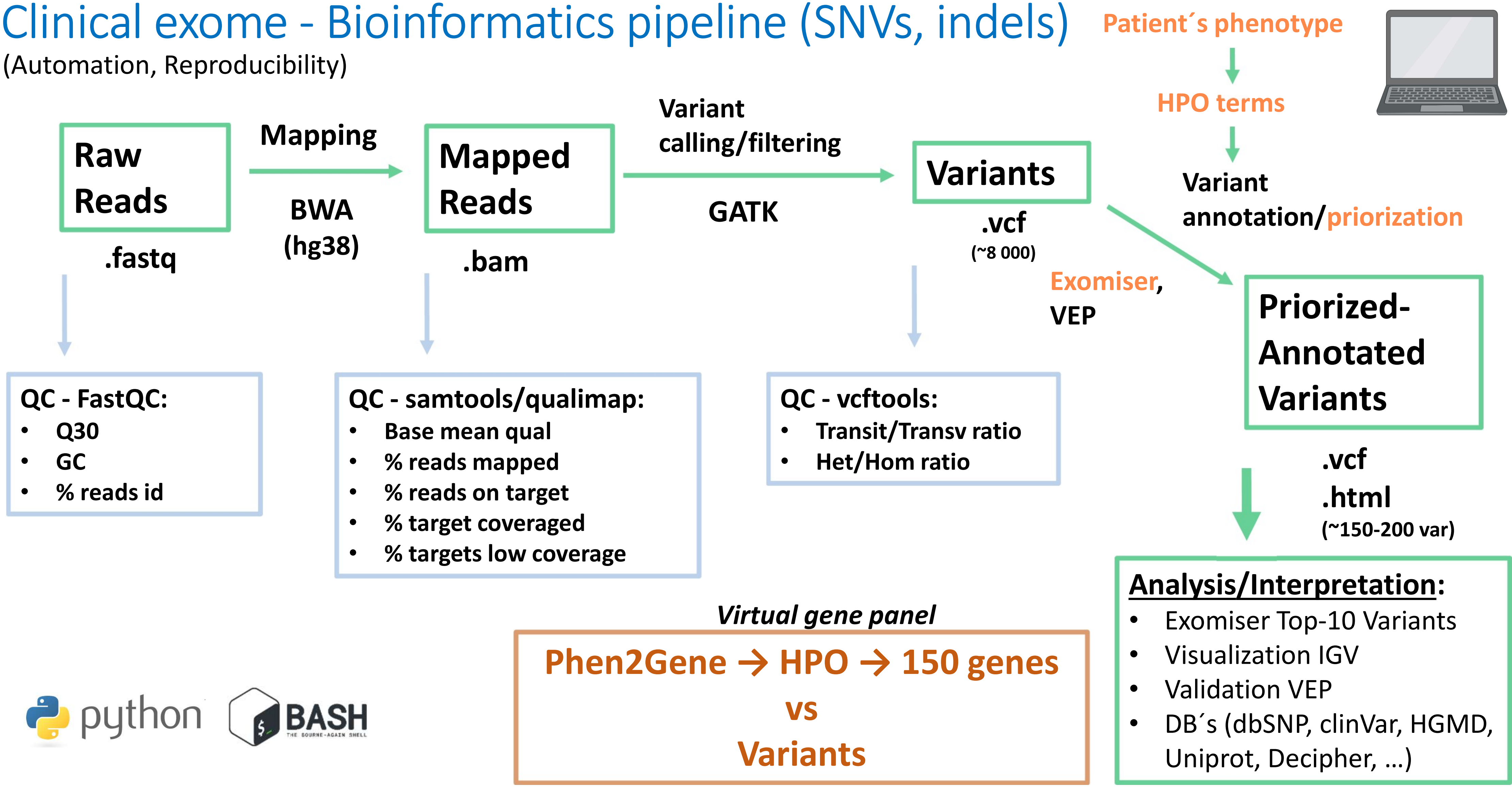
Clinical exome - Experimental procedure

- Library:
  - TruSight One sequencing panel (4 800 genes; ~62 000 targets)
- Sequencing:
  - MiSeq/NextSeq
  - Paired-end, 2x150pb



Clinical exome - Bioinformatics pipeline (SNVs, indels)

(Automation, Reproducibility)



Patient´s phenotype

HPO terms

Raw Reads

.fastq

Mapping

BWA (hg38)

Mapped Reads

.bam

Variant calling/filtering

GATK

Variants

.vcf

(~8 000)

Variant annotation/priorization

Exomiser, VEP

Priorized-Annotated Variants

.vcf

.html

(~150-200 var)

QC - FastQC:
- Q30
- GC
- % reads id

QC - samtools/qualimap:
- Base mean qual
- % reads mapped
- % reads on target
- % target coveraged
- % targets low coverage

QC - vcftools:
- Transit/Transv ratio
- Het/Hom ratio

Virtual gene panel

Phen2Gene → HPO → 150 genes

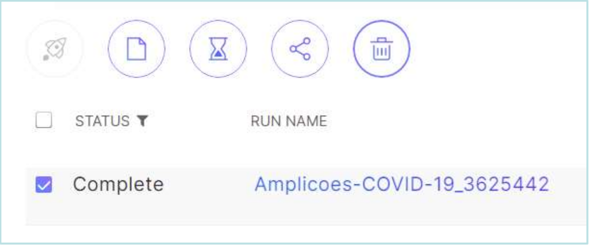
vs

Variants

Analysis/Interpretation:
- Exomiser Top-10 Variants
- Visualization IGV
- Validation VEP
- DB´s (dbSNP, clinVar, HGMD, Uniprot, Decipher, …)

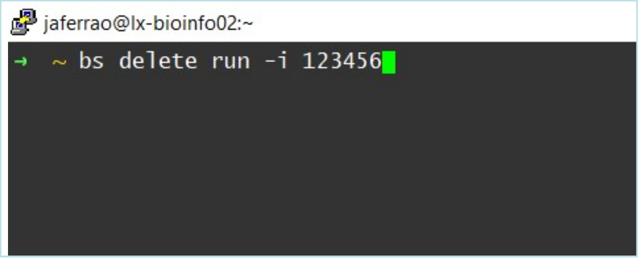
python BASH

Graphical user interface (GUI) vs Command-line interface (CLI)



□	STATUS	RUN NAME
☑	Complete	Amplicoes-COVID-19_3625442

Interação meios visuais



```
jaferrao@lx-bioinfo02:~
→  ~ bs delete run -i 123456
```

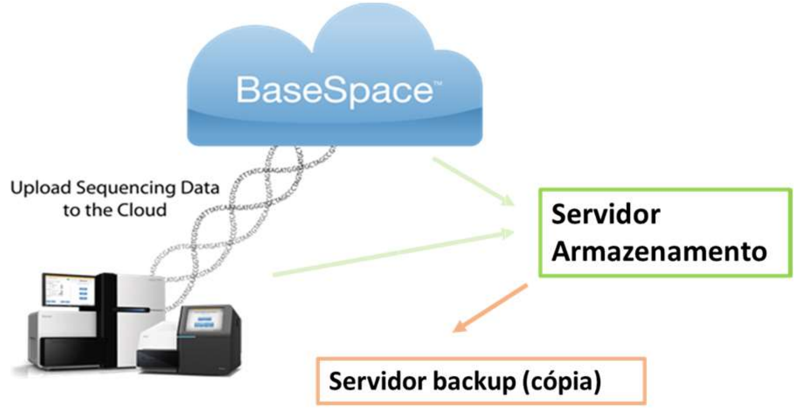
Interação comandos de texto

Windows vs Unix

Ferramentas específicas
Grandes datasets/Rec. Inform.
Servidores/clusters
Automatização
Reprodutibilidade

Automatização de procedimentos

- **Gestão automatizada armazenamento dados em bruto NGS**



- **Automatização controlo de qualidade NGS (InterOp, FastQC)**

MultiQC

Análise da qualidade da sequenciação

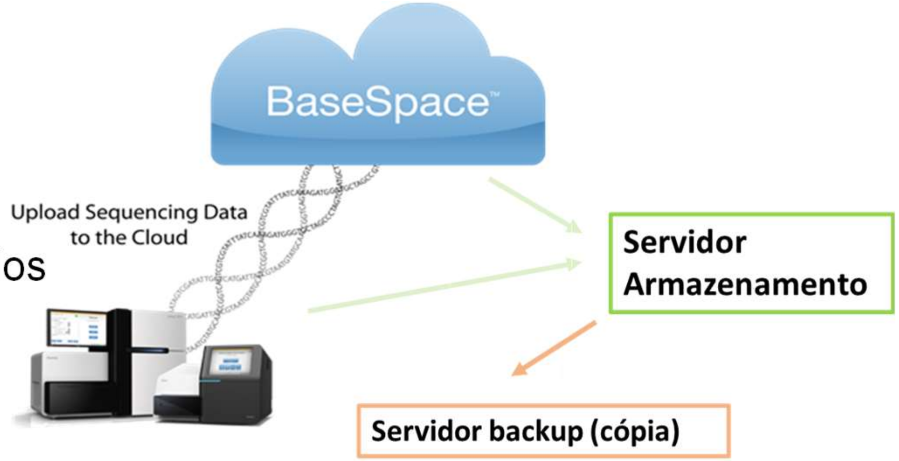
Dep. de Genética Humana - Unidade de Tecnologia e Inovação

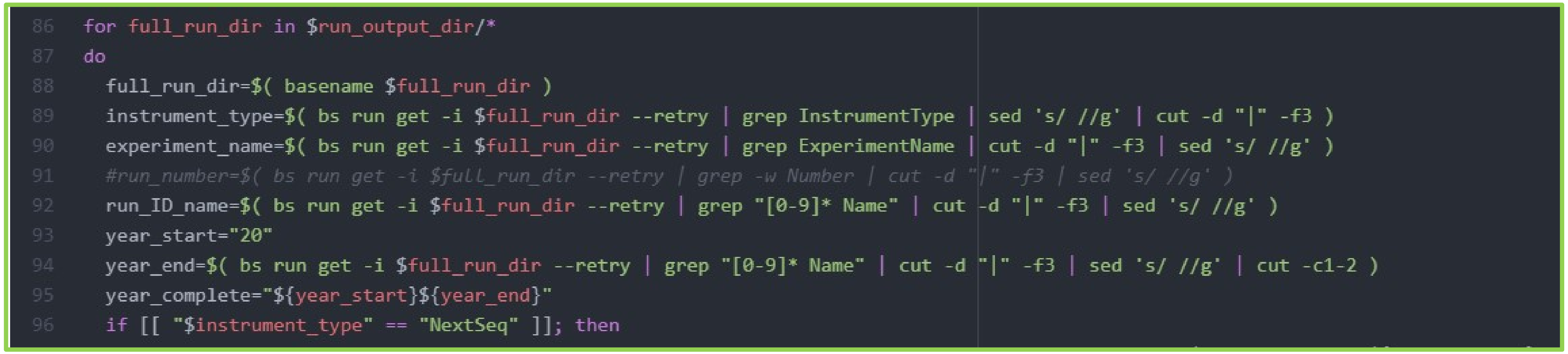
A análise primária foi efectuada usando os programas Interop e FastQC

BASH
THE BOURNE-AGAIN SHELL

Gestão automatizada armazenamento dados em bruto NGS

- Centenas Gigabytes dados por semana
- Gestão automatizada/programada semanal
- Transfere ficheiros corridas NGS para servidor armazenamento dados
- Guarda pasta com designação/formato específico
- Envia alertas por email





```
for full_run_dir in $run_output_dir/*
do
  full_run_dir=$( basename $full_run_dir )
  instrument_type=$( bs run get -i $full_run_dir --retry | grep InstrumentType | sed 's/ //g' | cut -d "|" -f3 )
  experiment_name=$( bs run get -i $full_run_dir --retry | grep ExperimentName | cut -d "|" -f3 | sed 's/ //g' )
  #run_number=$( bs run get -i $full_run_dir --retry | grep -w Number | cut -d "|" -f3 | sed 's/ //g' )
  run_ID_name=$( bs run get -i $full_run_dir --retry | grep "[0-9]* Name" | cut -d "|" -f3 | sed 's/ //g' )
  year_start="20"
  year_end=$( bs run get -i $full_run_dir --retry | grep "[0-9]* Name" | cut -d "|" -f3 | sed 's/ //g' | cut -c1-2 )
  year_complete="${year_start}${year_end}"
  if [[ "$instrument_type" == "NextSeq" ]]; then
```

Automatização controlo de qualidade NGS

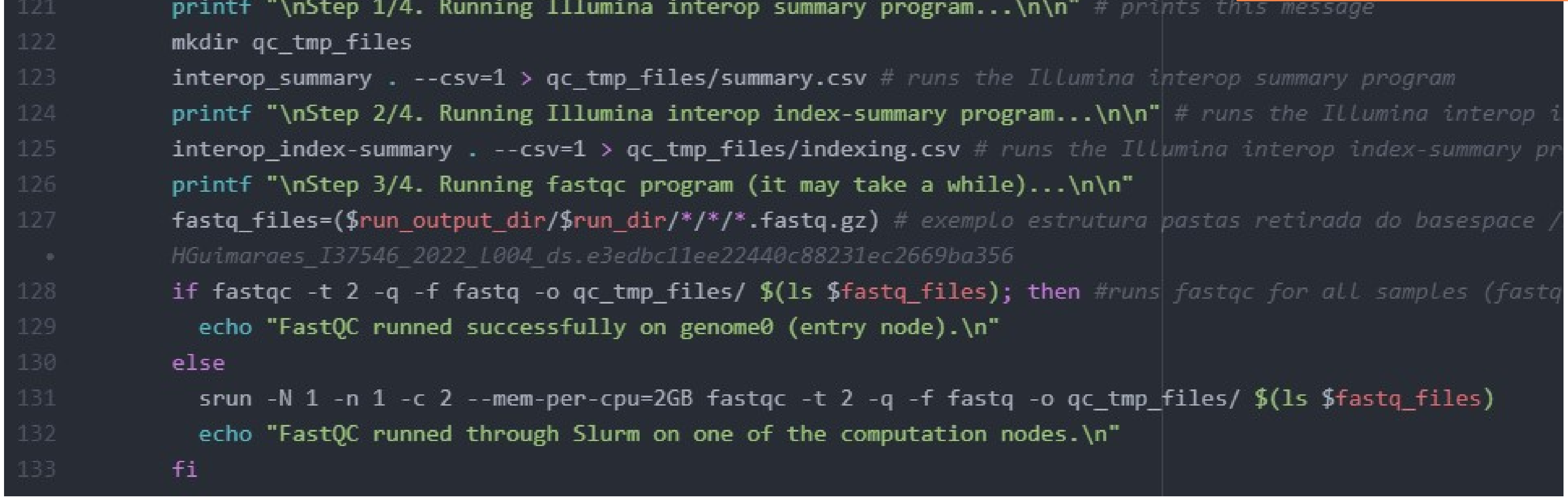
- Corre os programas de QC Illumina: interop summary e index-summary
- Corre o programa FastQC; Corre o MultiQC para gerar relatório
- Envia por email o relatório MultiQC (*.html)

MultiQC

Análise da qualidade da sequenciação

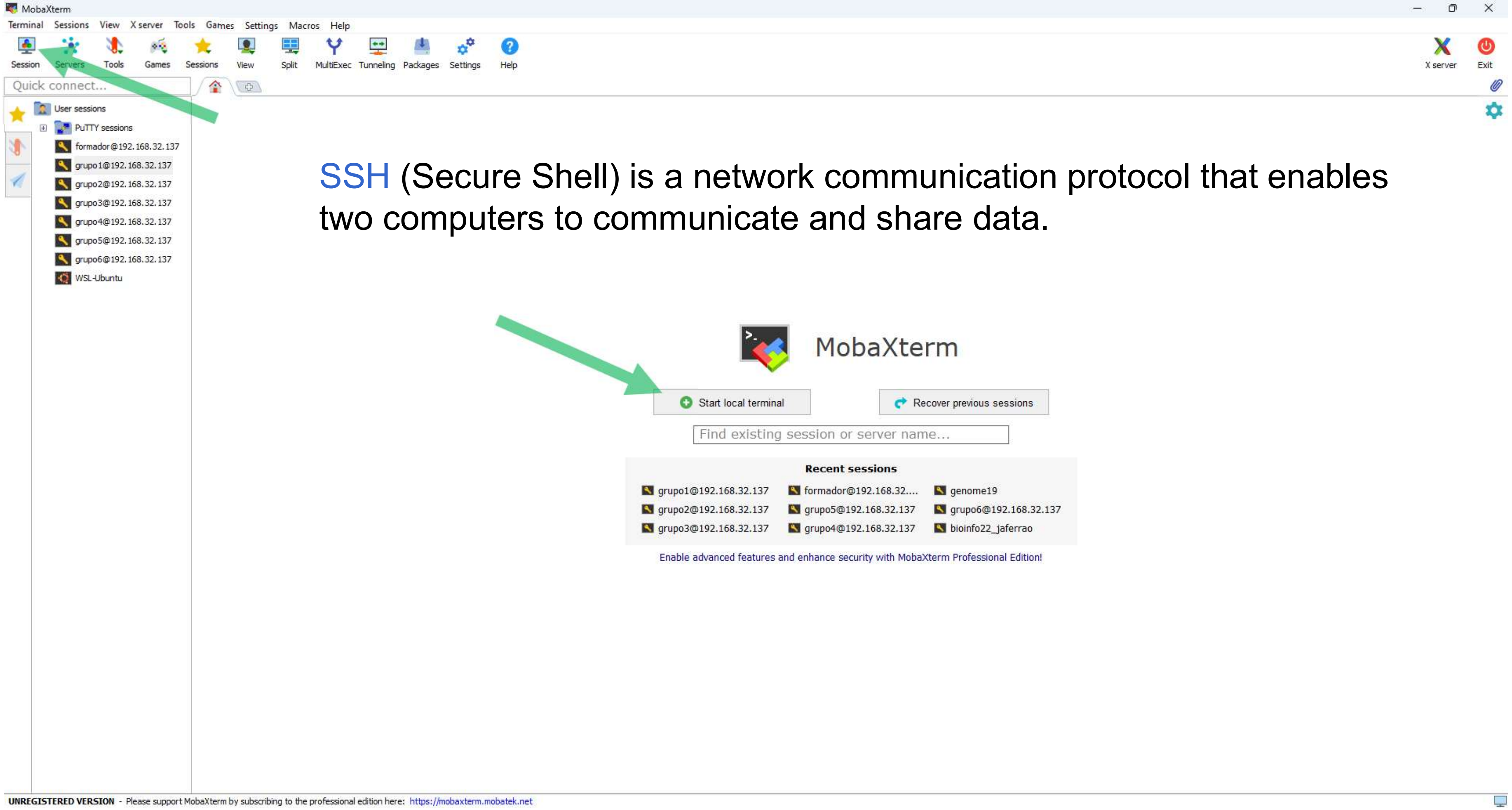
Dep. de Genética Humana - Unidade de Tecnologia e Inovação

A análise primária foi efectuada usando os programas Interop e FastQC



```
printf "\nStep 1/4. Running Illumina interop summary program...\n\n" # prints this message
mkdir qc_tmp_files
interop_summary . --csv=1 > qc_tmp_files/summary.csv # runs the Illumina interop summary program
printf "\nStep 2/4. Running Illumina interop index-summary program...\n\n" # runs the Illumina interop i
interop_index-summary . --csv=1 > qc_tmp_files/indexing.csv # runs the Illumina interop index-summary pr
printf "\nStep 3/4. Running fastqc program (it may take a while)...\n\n"
fastq_files=($run_output_dir/$run_dir/*/*/*.fastq.gz) # exemplo estrutura pastas retirada do basespace /
HGuimaraes_I37546_2022_L004_ds.e3edbc11ee22440c88231ec2669ba356
if fastqc -t 2 -q -f fastq -o qc_tmp_files/ $(ls $fastq_files); then #runs fastqc for all samples (fastq
  echo "FastQC runned successfully on genome0 (entry node).\n"
else
  srun -N 1 -n 1 -c 2 --mem-per-cpu=2GB fastqc -t 2 -q -f fastq -o qc_tmp_files/ $(ls $fastq_files)
  echo "FastQC runned through Slurm on one of the computation nodes.\n"
fi
```

Connection to the Unix server



SSH (Secure Shell) is a network communication protocol that enables two computers to communicate and share data.

MobaXterm
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
X server Exit
Quick connect...
Start local terminal
Recover previous sessions
Find existing session or server name...

User sessions
- PuTTY sessions
- formador@192.168.32.137
- grupo1@192.168.32.137
- grupo2@192.168.32.137
- grupo3@192.168.32.137
- grupo4@192.168.32.137
- grupo5@192.168.32.137
- grupo6@192.168.32.137
- WSL-Ubuntu

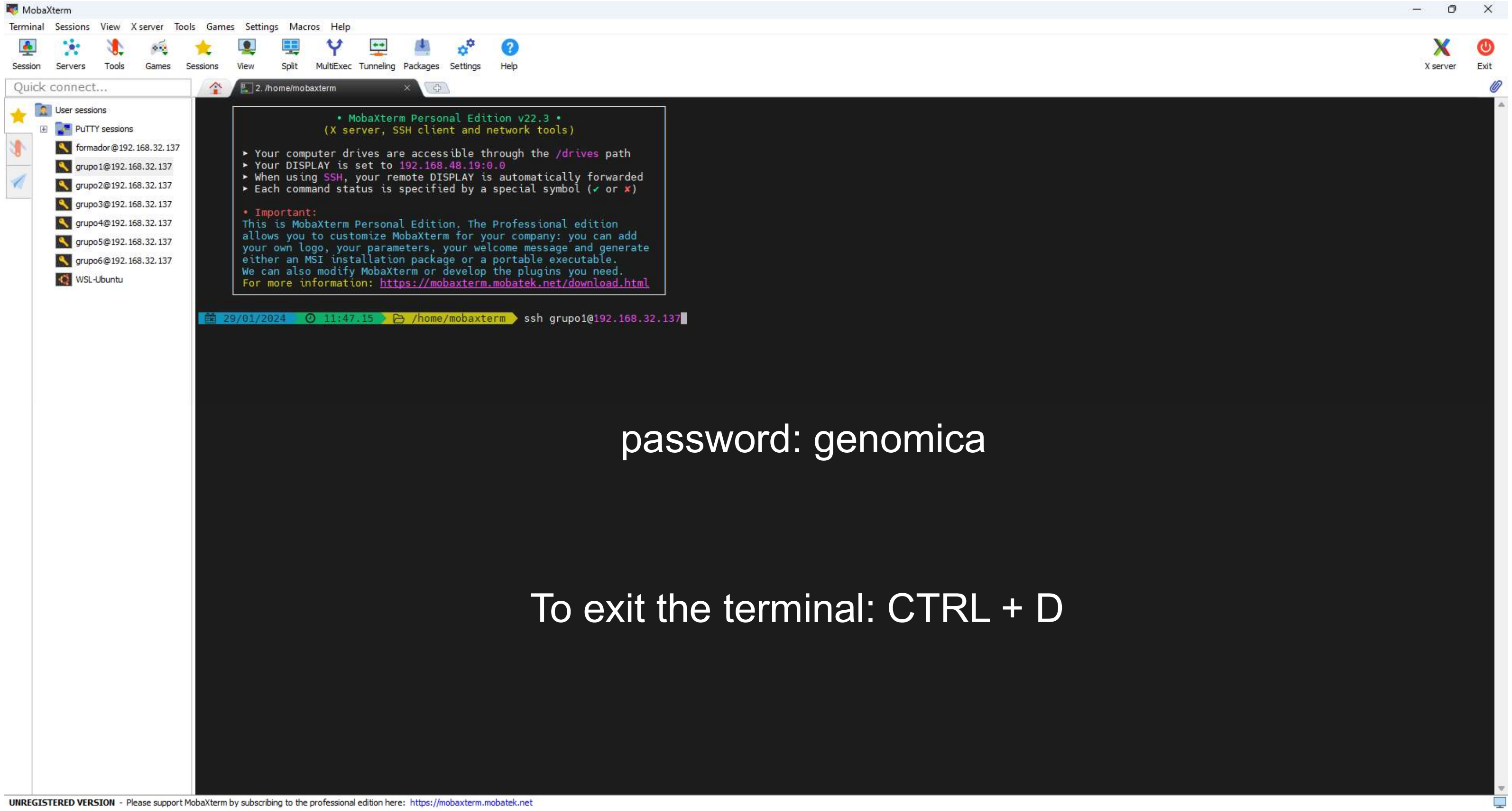
Recent sessions

grupo1@192.168.32.137	formador@192.168.32....	genome19
grupo2@192.168.32.137	grupo5@192.168.32.137	grupo6@192.168.32.137
grupo3@192.168.32.137	grupo4@192.168.32.137	bioinfo22_jaferrao

Enable advanced features and enhance security with MobaXterm Professional Edition!

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

Connection from local terminal:
ssh grupo1@192.168.32.137 [ENTER]



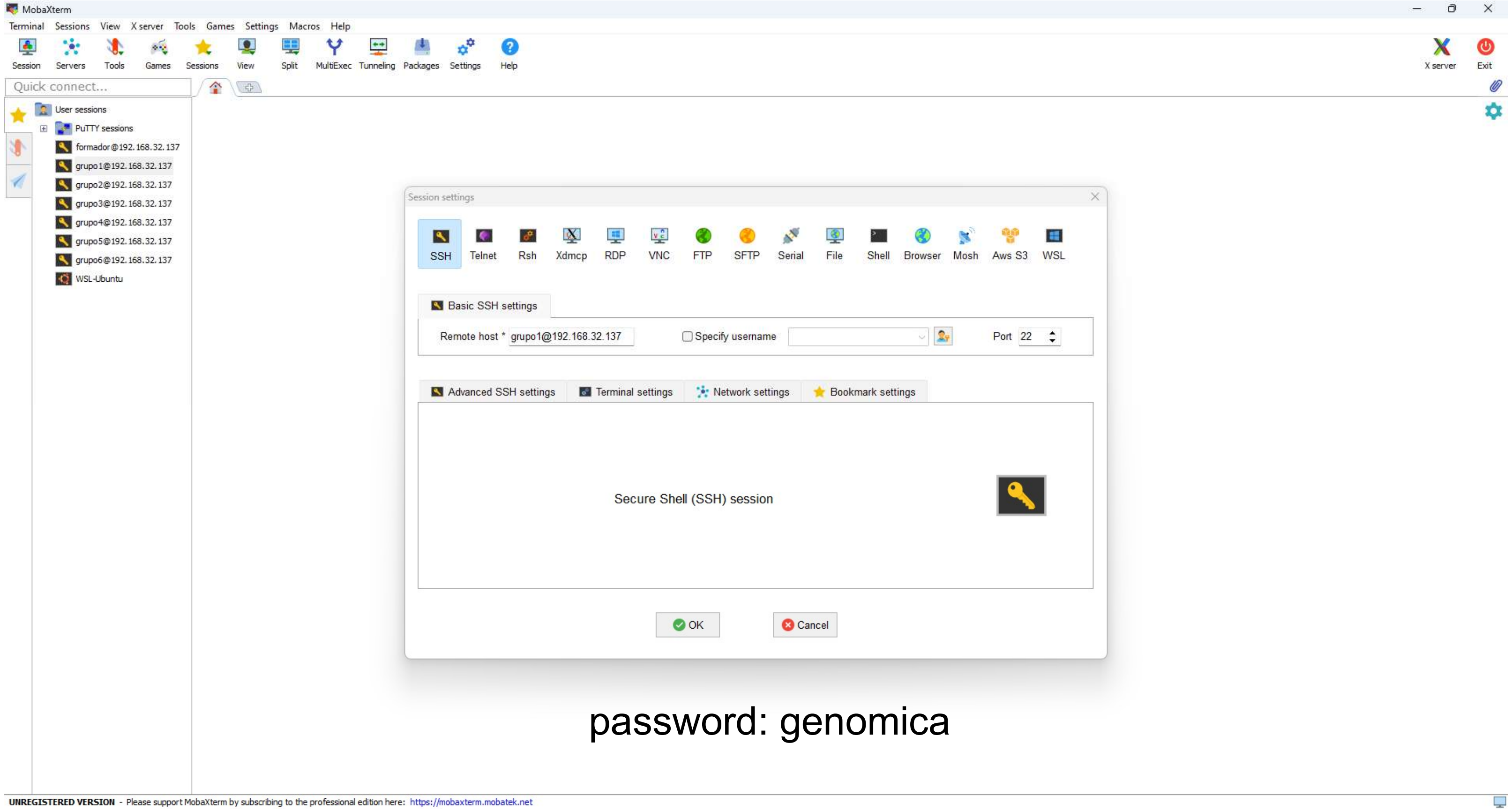
```
                 • MobaXterm Personal Edition v22.3 •
               (X server, SSH client and network tools)

 ▸ Your computer drives are accessible through the /drives path
 ▸ Your DISPLAY is set to 192.168.48.19:0.0
 ▸ When using SSH, your remote DISPLAY is automatically forwarded
 ▸ Each command status is specified by a special symbol (✓ or ✗)

 • Important:
 This is MobaXterm Personal Edition. The Professional edition
 allows you to customize MobaXterm for your company: you can add
 your own logo, your parameters, your welcome message and generate
 either an MSI installation package or a portable executable.
 We can also modify MobaXterm or develop the plugins you need.
 For more information: https://mobaxterm.mobatek.net/download.html

 29/01/2024  11:47.15  /home/mobaxterm  ssh grupo1@192.168.32.137
```

password: genomica

To exit the terminal: CTRL + D

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

Connection using a session



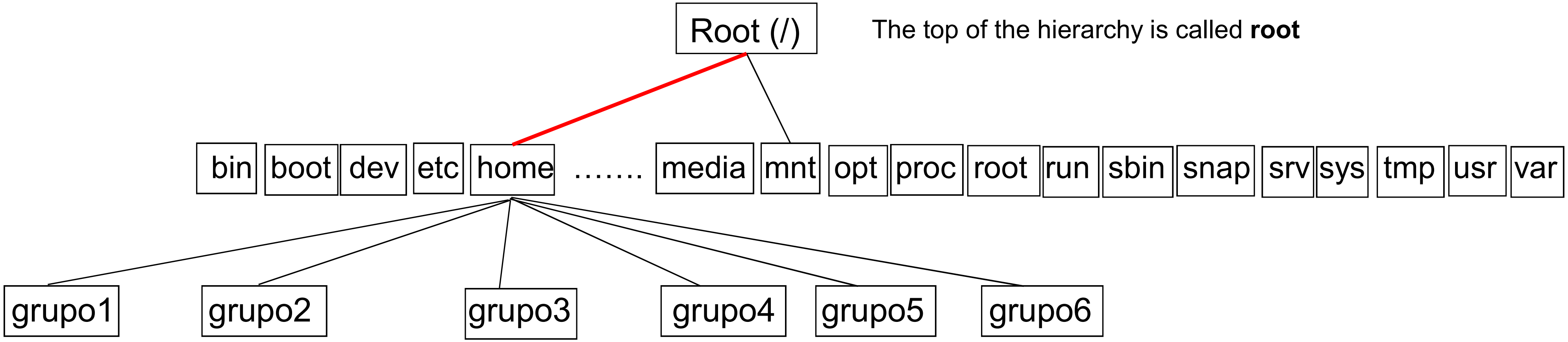
MobaXterm
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
X server Exit
Quick connect...
User sessions
PuTTY sessions
formador@192.168.32.137
grupo1@192.168.32.137
grupo2@192.168.32.137
grupo3@192.168.32.137
grupo4@192.168.32.137
grupo5@192.168.32.137
grupo6@192.168.32.137
WSL-Ubuntu
Session settings
SSH Telnet Rsh Xdmcp RDP VNC FTP SFTP Serial File Shell Browser Mosh Aws S3 WSL
Advanced SSH settings
Terminal settings
Network settings
Bookmark settings
Secure Shell (SSH) session
OK
Cancel
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

Basic SSH settings

Remote host * grupo1@192.168.32.137 ☐ Specify username Port 22

password: genomica

File-system on the Unix server



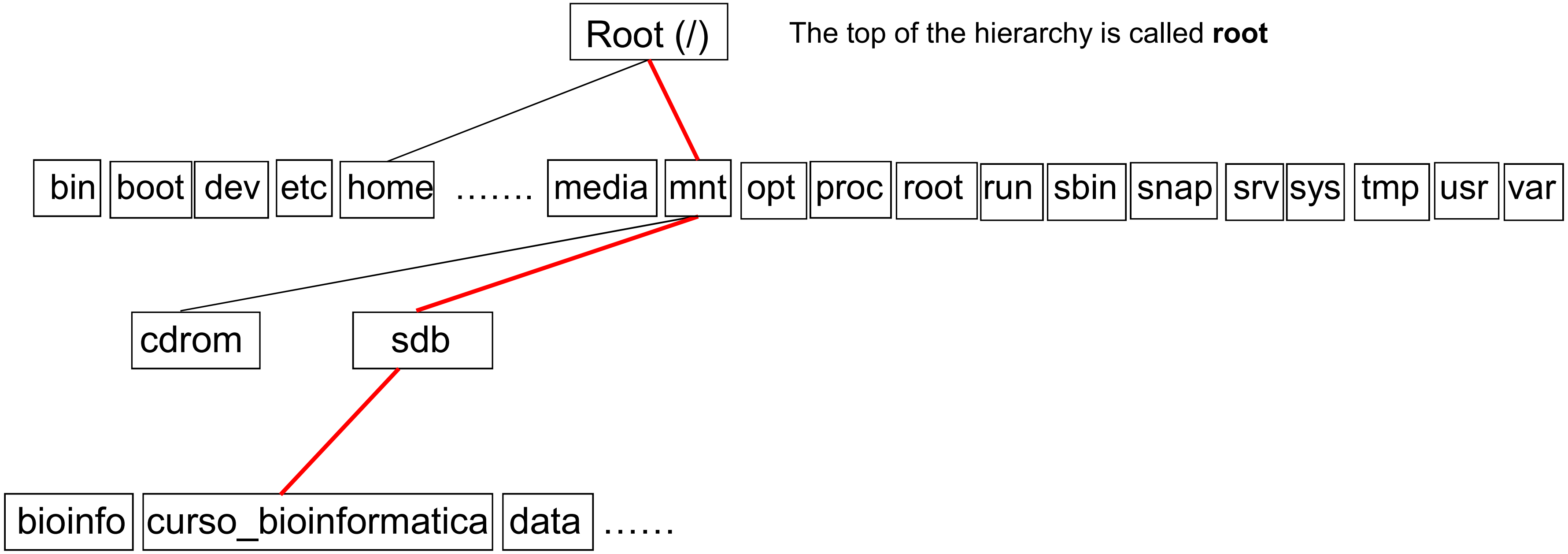
Root (/)

The top of the hierarchy is called **root**

bin boot dev etc home ……. media mnt opt proc root run sbin snap srv sys tmp usr var

grupo1 grupo2 grupo3 grupo4 grupo5 grupo6

File-system on the Unix server



Root (/)

The top of the hierarchy is called **root**

bin boot dev etc home ……. media mnt opt proc root run sbin snap srv sys tmp usr var

cdrom sdb

bioinfo curso_bioinformatica data ……

Git and GitHub

- Version control:
  - helps developers track and manage changes to code
- Colaboration
- CLI

- user-friendly interface (GUI)
- public code repository for free
- popular open-source projects

Tutorial linha de comandos Unix

- https://github.com/krother/bash_tutorial (clone this repository on home dir)

```
cd ~
git clone https://github.com/krother/bash_tutorial.git
```

- Extra: https://ubuntu.com/tutorials/command-line-for-beginners#1-overview

Conda

- Conda provides package, dependency, and environment management for any language.

- Conda allows users to install different versions of binary software packages and any required libraries appropriate for their computing platform. Also, it allows users to switch between package versions and download and install updates from a software repository.

- A popular Conda channel for bioinformatics software is *Bioconda*, which provides multiple software distributions for computational biology.

Conda

CONDA®

- conda env list
- conda activate curso_amb
- conda list
- conda deactivate
- conda activate curso_amb_vep