Computational Tools and Resources for the Life Sciences

Gil Poiares-Oliveira

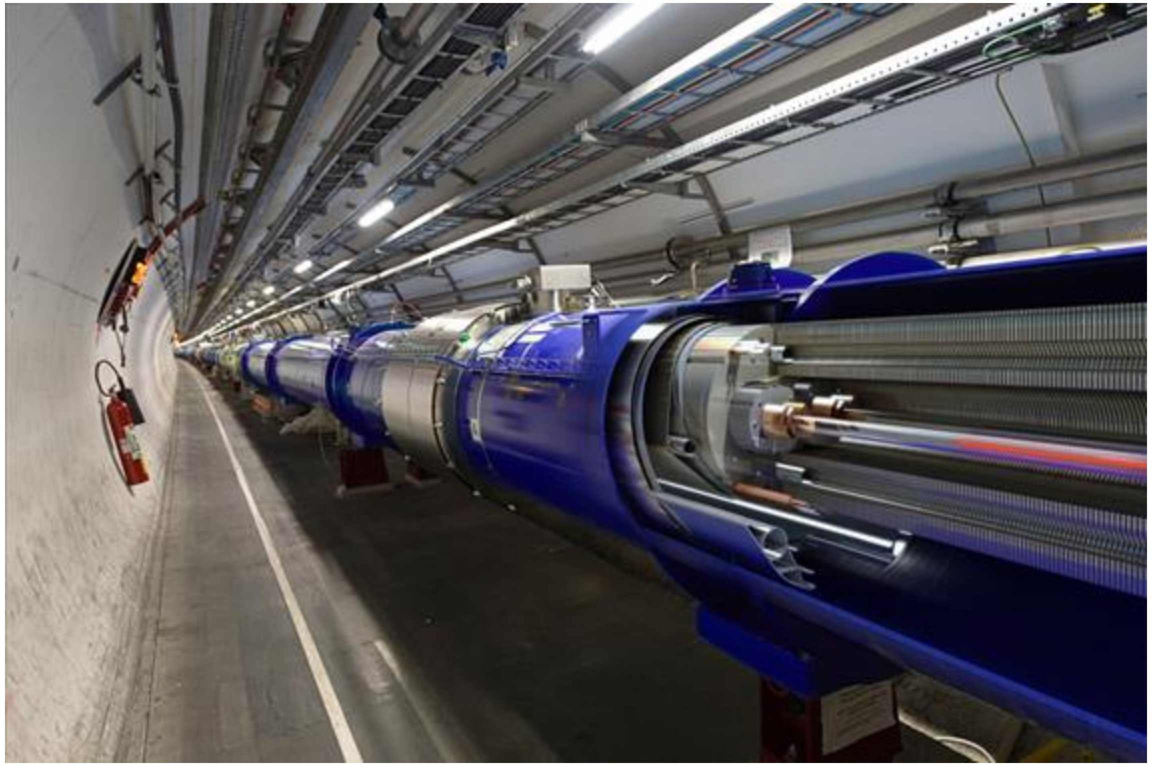
INESC-ID | BioData.pt | ELIXIR Portugal

ITQB NOVA PhD Student Meeting • 2024-05-16
ITQB NOVA • Oeiras, Portugal

Learning outcomes

- Understand the role of ELIXIR and BioData.pt in the European bioinformatics landscape
- Identify ELIXIR and BioData.pt services and projects like FEGA, Beacon and TeSS
- Perform basic genetic and phenoclinic queries using Beacon and the Beacon Network
- Create a bioinformatics analysis pipeline using Galaxy workflows
- Use ELIXIR TeSS to discover new training courses and materials

What is a research infrastructure?



Large Hadron Collider - CERN, Switzerland

Image: © 2014 CERN. Dominguez, Daniel

ELIXIR Nodes - ELIXIR, Europe

Image: © 2024 ELIXIR

BioData.pt

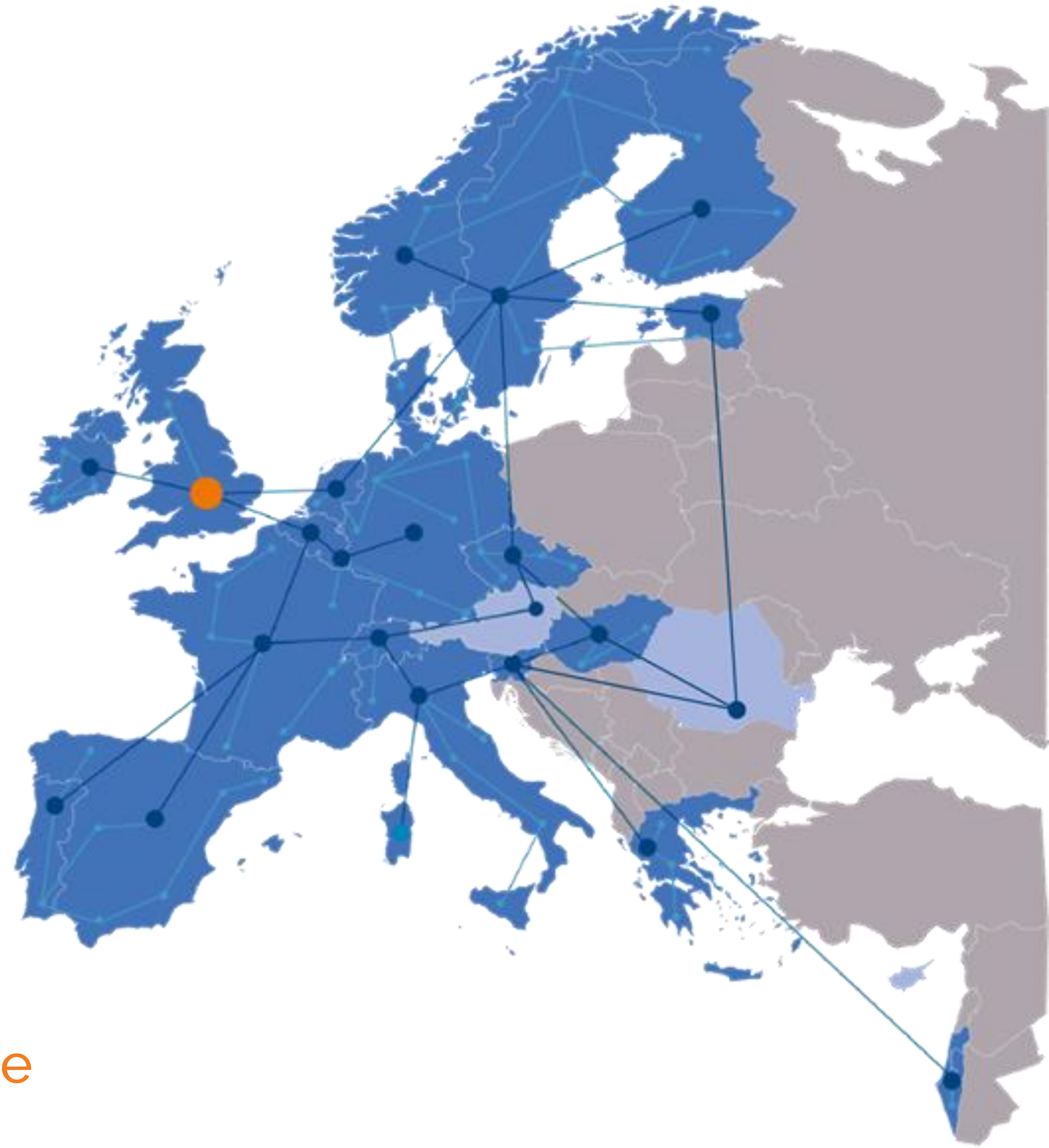
elixir PORTUGAL

ELIXIR Europe

ELIXIR is an intergovernmental organisation that brings together life science resources such as

- databases
- software tools
- training resources
- interoperability resources
- compute resources
- data management support

The goal of ELIXIR is to coordinate bioinformatics resources from across Europe so they form a single infrastructure.



elixir-europe.org

BioData.pt elixir PORTUGAL

Meet BioData.pt

- Portuguese distributed infrastructure for life & health data, and home of ELIXIR Portugal

- We offer **training**, **tools**, and **repositories** to the life sciences community and **connect** them with an **international network**

BioData.pt Services

Training

Bioinformatics and RDM

Bioinformatics analysis

RNA-Seq, DE, multiomics

Computing

servers, tool hosting

Mentorship

1-on-1 sessions

Our associates

NOVA MEDICAL SCHOOL

itqb nova

inesc id lisboa

CCMAR Centro de Ciências do Mar

UCIBIO @REQUIMTE

iBET Instituto de Biologia Experimental e Tecnológica

GULBENKIAN CIÊNCIA

cebal CENTRO DE BIOTECNOLOGIA AGRÍCOLA E AGRO-ALIMENTAR DO ALENTEJO

ciimar Centro Interdisciplinar de Investigação Marinha e Ambiental

iMM Instituto de Medicina Molecular | João Lobo Antunes

FACULDADE DE MEDICINA LISBOA

1 2 9 0 UNIVERSIDADE D COIMBRA

Ciências ULisboa

Universidade do Minho

NOVA NOVA UNIVERSITY LISBON

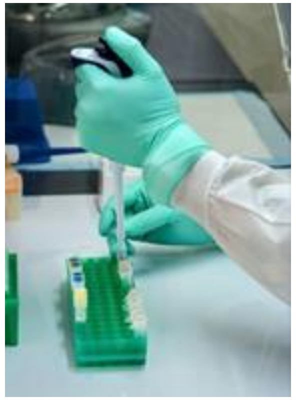
universidade de aveiro

TÉCNICO LISBOA

- If you're affiliated with one of our associates, you can benefit from our services

How does a patient DNA sample get into a computer?

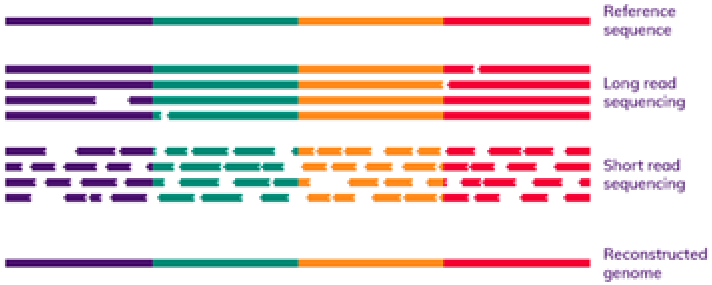
Collection



DNA isolation & library prep

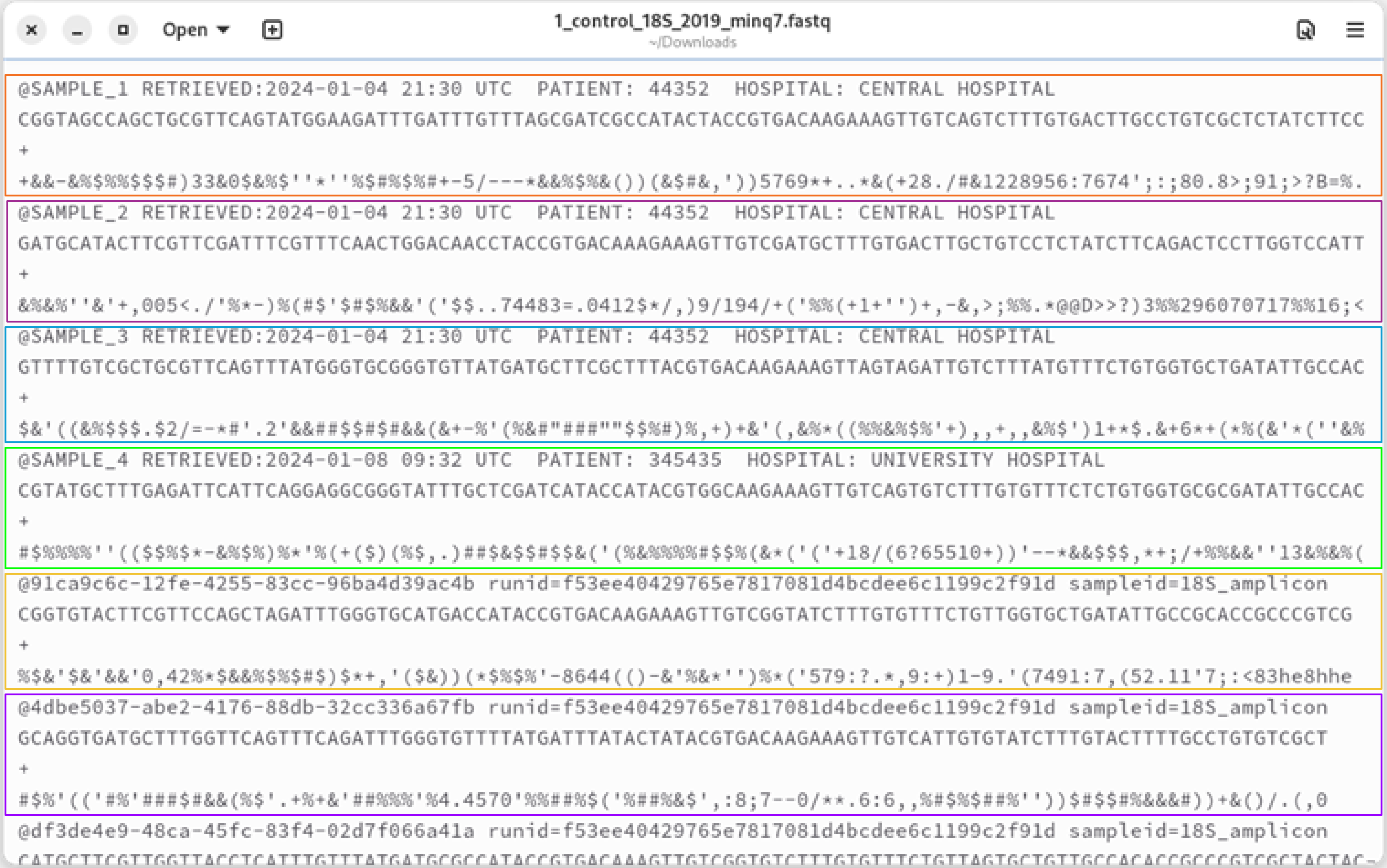
Sequencing

FASTQ



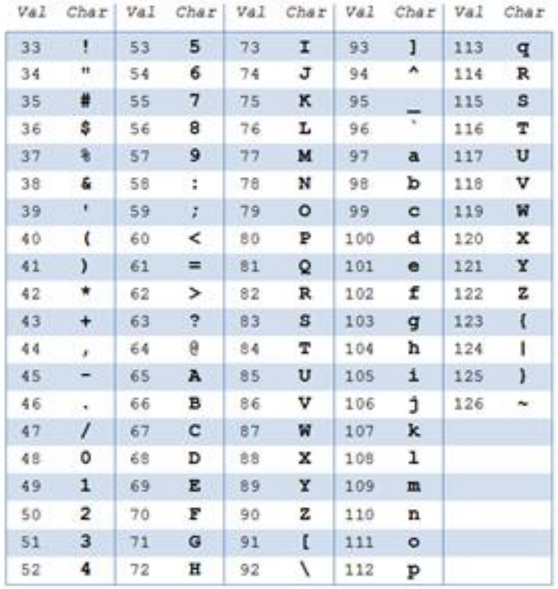
Assembly

Example FASTQ file



```
@SAMPLE_1 RETRIEVED:2024-01-04 21:30 UTC  PATIENT: 44352  HOSPITAL: CENTRAL HOSPITAL
CGGTAGCCAGCTGCGTTCAGTATGGAAGATTTGATTTGTTTAGCGATCGCCATACTACCGTGACAAGAAAGTTGTCAGTCTTTGTGACTTGCCTGTCGCTCTATCTTCC
+
+&&-&%$%%$$$#)33&0$&%$''*''%$#%$%#+-5/---*&&%$%&())(&$#&,'))5769*+..*&(+28./#&1228956:7674';:;80.8>;91;>?B=%.
@SAMPLE_2 RETRIEVED:2024-01-04 21:30 UTC  PATIENT: 44352  HOSPITAL: CENTRAL HOSPITAL
GATGCATACTTCGTTCGATTTCGTTTCAACTGGACAACCTACCGTGACAAAGAAAGTTGTCGATGCTTTGTGACTTGCTGTCCTCTATCTTCAGACTCCTTGGTCCATT
+
&%&%''&'+,005<./'%*-)%(#$'$#$%&&'('$$..74483=.0412$*/,)9/194/+('%%(+1+'')+,-&,>;%%.*@@D>>?)3%%296070717%%16;<
@SAMPLE_3 RETRIEVED:2024-01-04 21:30 UTC  PATIENT: 44352  HOSPITAL: CENTRAL HOSPITAL
GTTTTGTCGCTGCGTTCAGTTTATGGGTGCGGGTGTTATGATGCTTCGCTTTACGTGACAAGAAAGTTAGTAGATTGTCTTTATGTTTCTGTGGTGCTGATATTGCCAC
+
$&'((&%$$$.$2/=-*#'.2'&&##$$#$#&&(&+-%'(%&#"###""$$%#)%,+)+&'(,&%*((%%&%$%'+),,+,,&%$')1+*$.&+6**(*%(&'*(''&%
@SAMPLE_4 RETRIEVED:2024-01-08 09:32 UTC  PATIENT: 345435  HOSPITAL: UNIVERSITY HOSPITAL
CGTATGCTTTGAGATTCATTCAGGAGGCGGGTATTTGCTCGATCATACCATACGTGGCAAGAAAGTTGTCAGTGTCTTTGTGTTTCTCTGTGGTGCGCGATATTGCCAC
+
#$%%%%''(($$%$*-&%$%)%*'%(+($)(%$,.)##$&$$#$$&('(%&%%%%#$$%(&*('('+18/(6?65510+))'--*&&$$$,**;/+%%&&''13&%&%(
@91ca9c6c-12fe-4255-83cc-96ba4d39ac4b runid=f53ee40429765e7817081d4bcdee6c1199c2f91d sampleid=18S_amplicon
CGGTGTACTTCGTTCCAGCTAGATTTGGGTGCATGACCATACCGTGACAAGAAAGTTGTCGGTATCTTTGTGTTTCTGTTGGTGCTGATATTGCCGCACCGCCCGTCG
+
%$&'$&'&&'0,42%*$&&%$%$#$)$*+,'($&))(*$%$%'-8644(()-&'%&*'')%*('579:?.*,9:+)1-9.'(7491:7,(52.11'7;:<83he8hhe
@4dbe5037-abe2-4176-88db-32cc336a67fb runid=f53ee40429765e7817081d4bcdee6c1199c2f91d sampleid=18S_amplicon
GCAGGTGATGCTTTGGTTCAGTTTCAGATTTGGGTGTTTTATGATTTATACTATACGTGACAAGAAAGTTGTCATTGTGTATCTTTGTACTTTTGCCTGTGTCGCT
+
#$%'(('#%'###$#&&(%$'.+%+&'##%%%'%4.4570'%%##%$('%##%&$',:8;7--0/**.6:6,,%#$%$##%''))$#$$#%&&&#))+&()/.(,0
@df3de4e9-48ca-45fc-83f4-02d7f066a41a runid=f53ee40429765e7817081d4bcdee6c1199c2f91d sampleid=18S_amplicon
CATGCTTCGTTGGTTACCTCATTTGTTTATGATGCGCCATACCGTGACAAAGTTGTCGGTGTCTTTGTGTTTCTGTTAGTGCTGTTGCCACACCGCCCGTCGCTACTAC-
```

1. @ sample ID
2. Nucleotide seq.
3. +
4. Quality score



Val	Char	Val	Char	Val	Char	Val	Char	Val	Char
33	!	53	5	73	I	93	]	113	q
34	"	54	6	74	J	94	^	114	R
35	#	55	7	75	K	95	_	115	S
36	$	56	8	76	L	96	`	116	T
37	%	57	9	77	M	97	a	117	U
38	&	58	:	78	N	98	b	118	V
39	'	59	;	79	O	99	c	119	W
40	(	60	<	80	P	100	d	120	X
41	)	61	=	81	Q	101	e	121	Y
42	*	62	>	82	R	102	f	122	Z
43	+	63	?	83	S	103	g	123	{
44	,	64	@	84	T	104	h	124	|
45	-	65	A	85	U	105	i	125	}
46	.	66	B	86	V	106	j	126	~
47	/	67	C	87	W	107	k		
48	0	68	D	88	X	108	l		
49	1	69	E	89	Y	109	m		
50	2	70	F	90	Z	110	n		
51	3	71	G	91	[	111	o		
52	4	72	H	92	\	112	p		

BioData.pt

elixir PORTUGAL

1st Challenge: Getting the data

What we know:

- First recorded symptoms in Bergen, Norway
- Some carriers are asymptomatic, genetic factors may have influence
- University of Bergen has patients' sequencing data
- We are researchers at ITQB NOVA

How can we help from here?

BioData.pt

elixir PORTUGAL

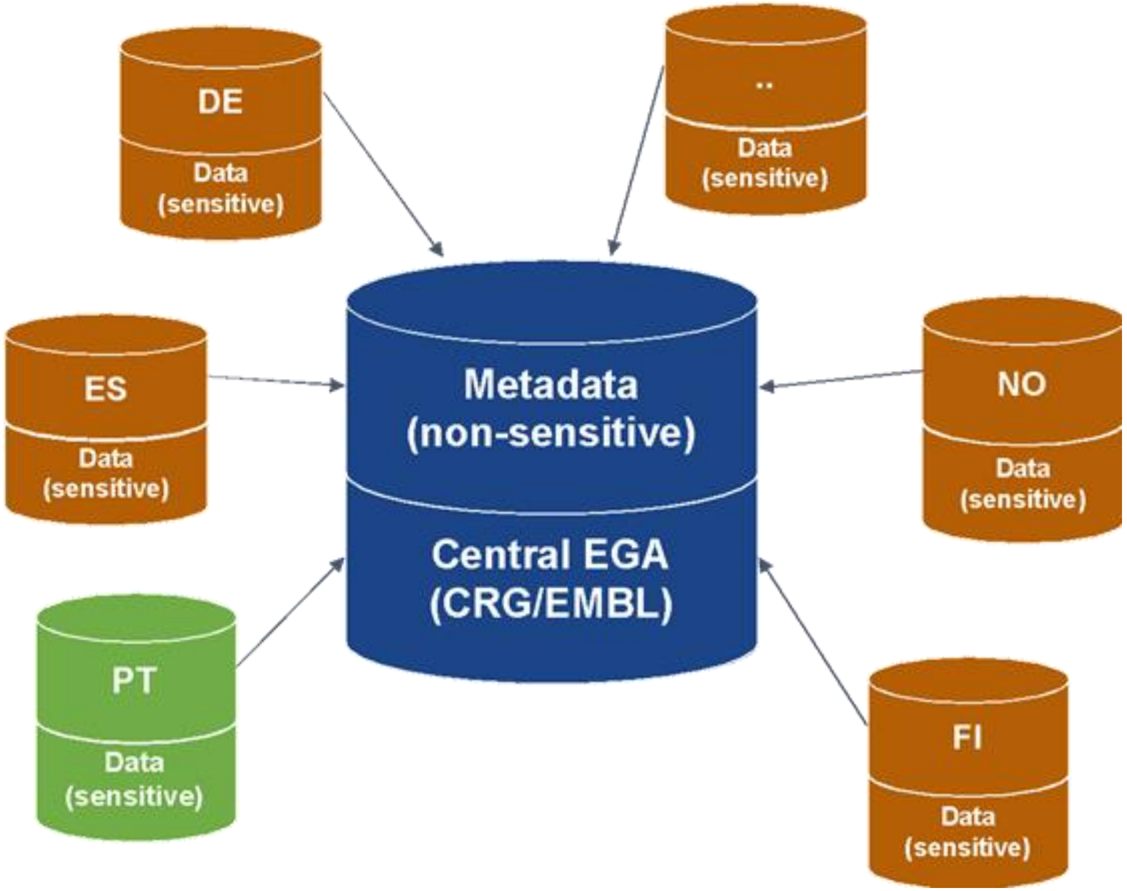
Federated EGA, a resource for discovery and access of human data across national borders
https://www.youtube.com/watch?v=bRFHyI6hnNk

Federated European Genome-phenome Archive Portugal

fegaportugal.biodata.pt

User journey

1. User **registration** in EGA
2. User approval as **Submitter**
3. Creation of Data Access Committee (**DAC**) and associated Data Access Policies.
4. **Upload** of a real genome file into the local server.
5. Submission of **metadata** by user.
6. **Request** to access to the genome, DAC approval and rejection
7. Genome **download** and decryption.



BioData.pt | elixir PORTUGAL

Beacon Project

- A Beacon allows an anonymous queries on genomic datasets.

  e.g. Check if a genome contains a genome with a given base at a particular coordinate

  beacon-project.io

  🇵🇹 beacon.biodata.pt

elixir beacon network

- Query all ELIXIR's beacons

  beacon-network.elixir-europe.org

Challenge 1

1. How many people in the Portuguese Beacon have an adenine instead of guanine at position 69848 of chromosome 1?
2. How many people have asthma?
3. How many of them are women?
4. How many of the female asthma patients are of African ethnicity?
5. Which country's beacon has a patient with a mitochondrial mutation at the 10th nucleotide which switches a thymine to a cytosine

Step 2: Analysing the data

Here's what a sample pipeline looks like:

1. **Trim** adapters and low quality reads
2. Perform **quality control** on the data
3. **Map reads** against reference genome
4. **Analyse** according to biological question

   e.g. variant calling - use variant detector to compare against reference genome

Tools:

e.g. Trimmomatic, seqtk_trimq

e.g. FastQC

e.g. STAR, HISAT2, Bowtie2, TopHat2

e.g. FreeBayes

Minimum requirements:

- Trimmomatic
- FASTQC
- STAR
- HISAT2
- Bowtie2
- TopHat2
- FreeBays
- *That cool tool you've just discovered on WoS*

1. A powerful enough computer…

- Basic Unix command line knowledge
- Introductory programming skills*
- Familiarity with Python/R*
- Knowledge of Pandas/matplotlib/…*

*additionally, depending on the tool

or

- A bioinformatician

or

Galaxy
PROJECT

Galaxy Project

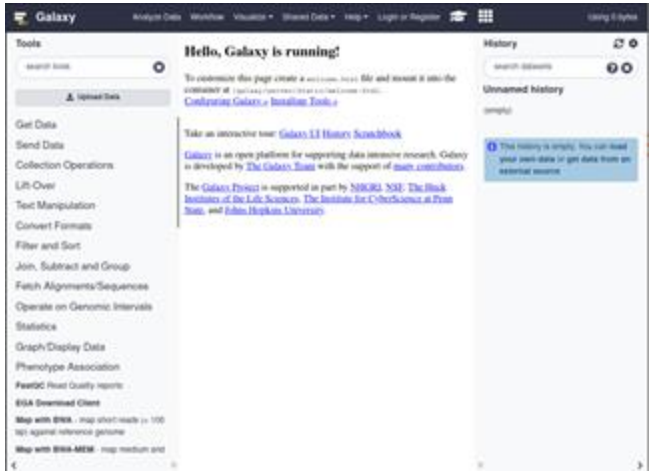
Galaxy PROJECT

- Galaxy is a free, open-source system for analyzing data, authoring workflows, training and education, publishing tools, managing infrastructure, and more.

galaxyproject.org

Galaxy & the Pulsar Network

Self-hosted



Hello, Galaxy is running!

To customize this page create a welcome.html file and mount it into the container at /galaxy/server/static/welcome.html.
Configuring Galaxy » Installing Tools »

Take an interactive tour: Galaxy UI History Scratchbook

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

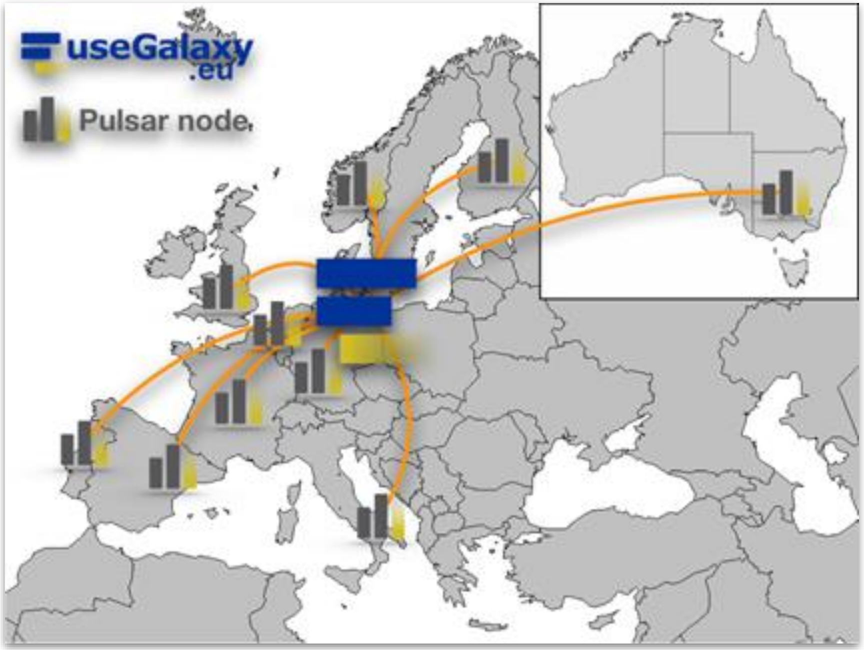
The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

This history is empty. You can load your own data or get data from an external source

Your institute's server

Galaxy software

Pulsar Network

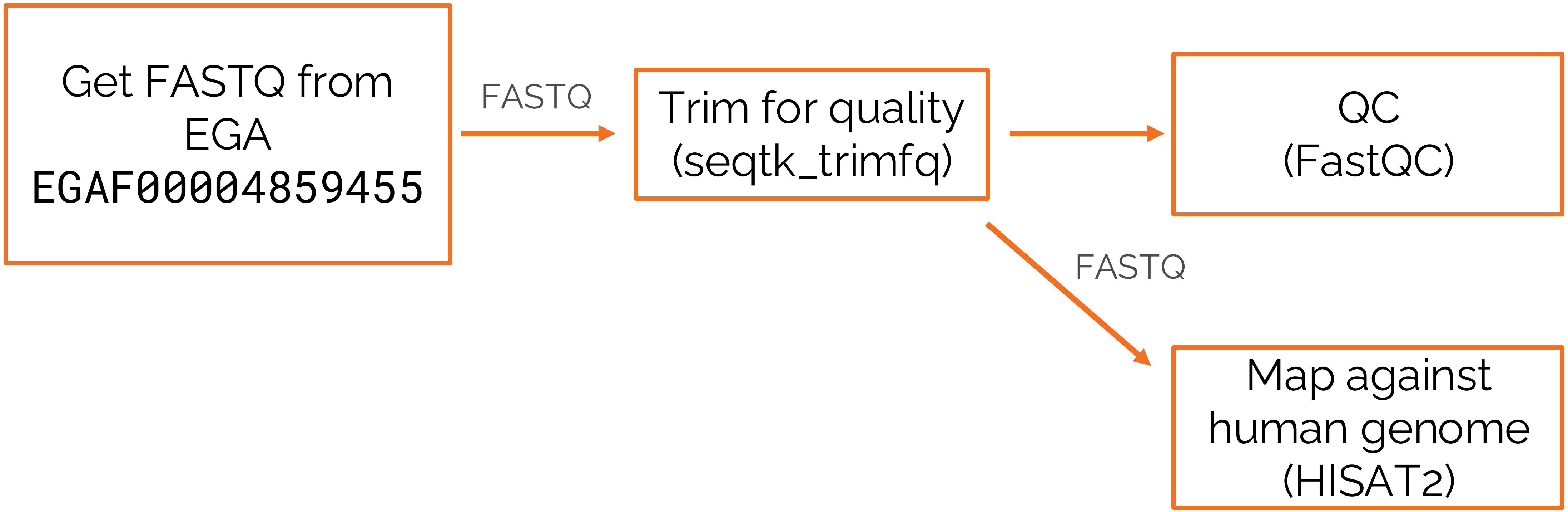


usegalaxy.eu

BioData.pt

elixir PORTUGAL

Challenge 2: build a workflow in Galaxy



Get FASTQ from EGA
`EGAF00004859455`

FASTQ

Trim for quality (seqtk_trimfq)

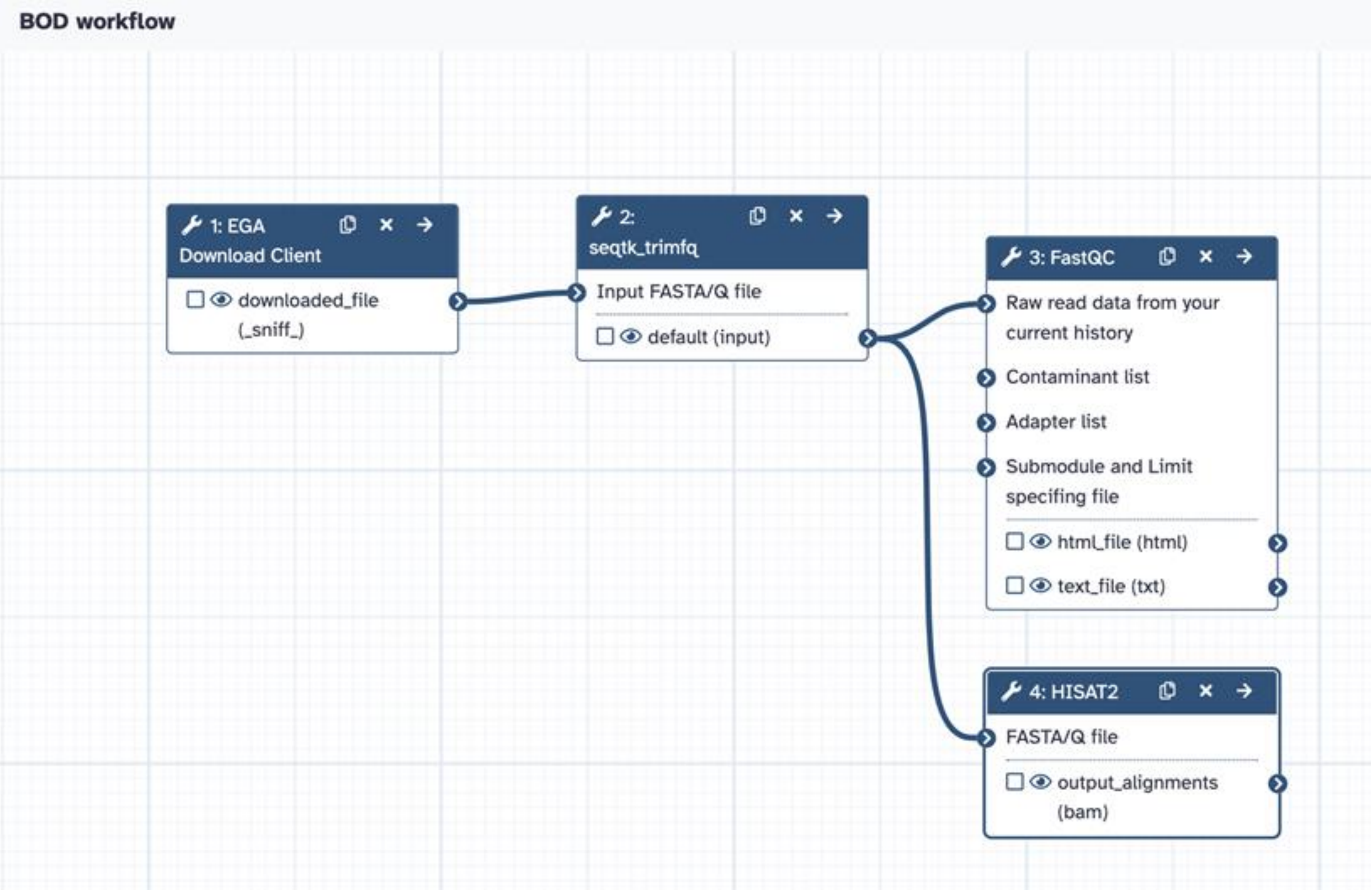
QC (FastQC)

FASTQ

Map against human genome (HISAT2)

Compare with mine

https://usegalaxy.eu/u/gilpo/w/bod-workflow



BOD workflow

1: EGA Download Client
- □ downloaded_file (_sniff_)

2: seqtk_trimfq
- Input FASTA/Q file
- □ default (input)

3: FastQC
- Raw read data from your current history
- Contaminant list
- Adapter list
- Submodule and Limit specifing file
- □ html_file (html)
- □ text_file (txt)

4: HISAT2
- FASTA/Q file
- □ output_alignments (bam)

Step 3: Learning more

- Find training courses and materials from ELIXIR Nodes and institutes

tess.elixir-europe.org

Challenge 3: Diving into TeSS

Think about some bioinformatics/data science skill or tool that you've been meaning to learn

phylogenomics

Bash

Nextflow

RNA-Seq

Docker

multiomics

Git

PCA

R

differential expression

metagenomics

disease modelling

Python

Data Management Plans

data visualisation

machine learning

metabolic modelling

protein structure

FAIR principles

Linux

protein function prediction

Challenge 3: Diving into TeSS

Try to answer the following questions about your chosen topic using TeSS

1. How many training events are scheduled on that topic?
2. How many training events, in total are registered on that topic?
3. How many materials could you find on that topic?
4. How many materials, on all topics, are from the Portugal node?
5. How many materials, on all topics, are targeted at PhD Students?

There's many more...

bio.tools

bio.tools helps you find and select bionformatics software and connect it in workflows.

BioContainers

Search a repository of containerised software that you can build into workflows.

WorkflowHub

A registry for sharing and publishing scientific computational workflows.

FAIRsharing.org

FAIRsharing.org allows you to search for databases and data policies by aspects such as domain, species and country.

TeSS

Search for training courses, webinars, training materials and workflows in TeSS, ELIXIR's training portal.

Overview of good data management practices

The Research Data Management Kit (RDMkit) guides you through the whole data management life cycle and includes advice specific to your domain, your role and your country.

Step-by-step instructions

The FAIR Cookbook contains step-by-step recipes to accomplish specific data management tasks and to make your data FAIR (Findable, Accessible, Interoperable, Reusable).

Data management plan wizard

The Data Stewardship Wizard (DSW) is an online tool that guides researchers and data stewards through their data management planning.

LS Login

LS LOGIN

The Life Science Login enables researchers to use their home organisation credentials or community or other identities o sign in and access data and services they need.

Bioconda

BIOCONDA

Bioconda lets you install thousands of software packages related to biomedical research using the conda package manager.

Take home messages

- ELIXIR is the European life sciences data research infrastructure
- BioData.pt is the national life and health data research infrastructure, and hosts the Portuguese ELIXIR node
- Both provide tools and services to support life sciences researchers, e.g. FEGA, Beacon, Galaxy, TeSS
- You can count on us to help you navigate through the complexities of biological data :)

Thanks!

This presentation will be available at
github.com/BioData-PT/computational-tools-resources

Gil Poiares-Oliveira, MSc

gpo@biodata.pt

INESC-ID | BioData.pt | ELIXIR Portugal

BioData.pt
Living data