

# Computational Tools and Resources for the Life Sciences

Gil Poiares-Oliveira, MSc

INESC-ID | BioData.pt | ELIXIR Portugal

XIII Bioinformatics Open Days • 2024-05-16

University of Minho School of Engineering • Braga, Portugal

# What is a research infrastructure?



Large Hadron Collider - CERN, Switzerland

Image: © 2014 CERN. Dominguez, Daniel



ELIXIR Nodes - ELIXIR, Europe

Image: © 2024 ELIXIR

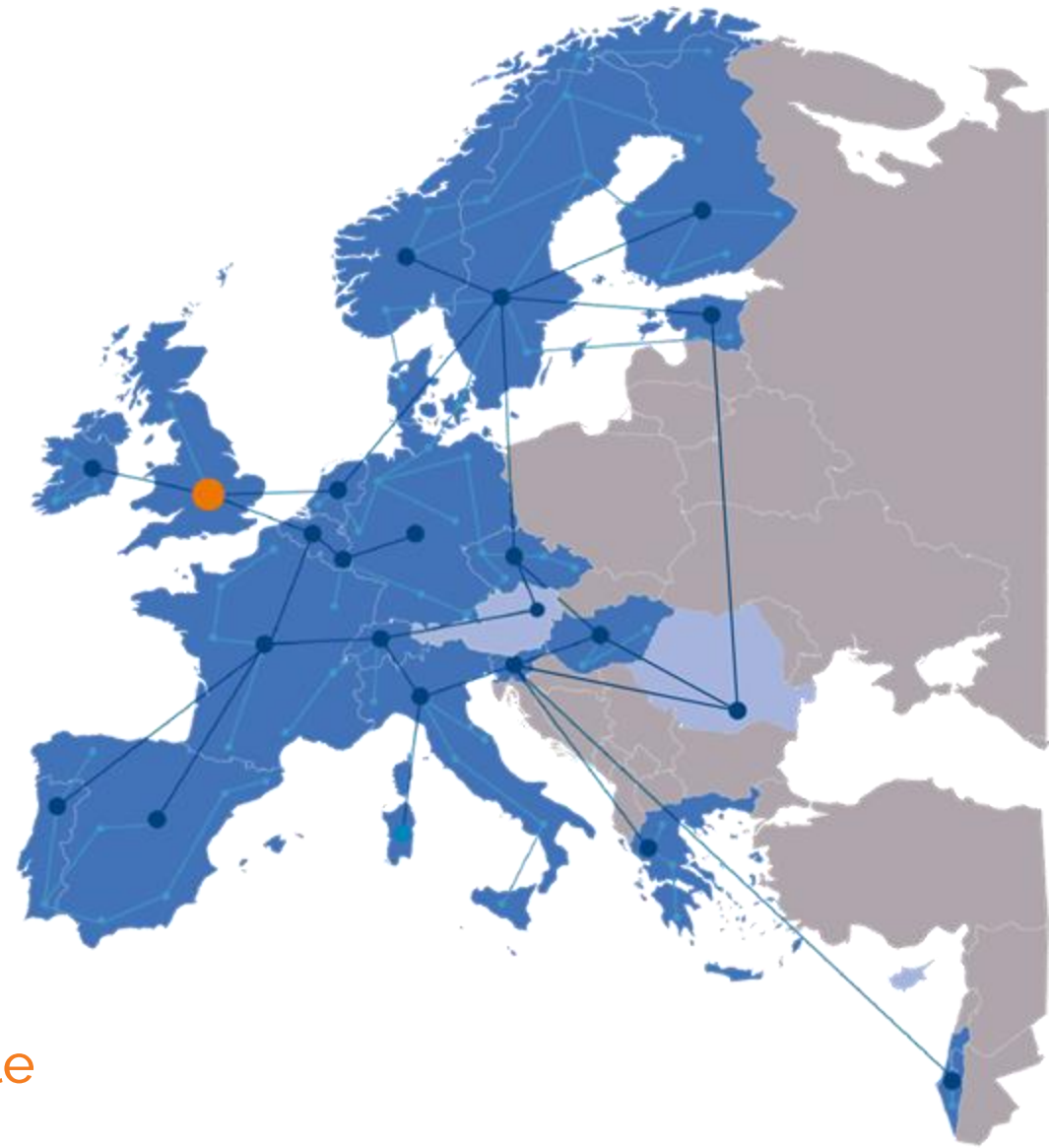
# ELIXIR Europe

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](http://elixir-europe.org)



# 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



## Computation

servers, tool  
hosting



## Mentorship

1-on-1  
sessions

# Our associates

NOVA MEDICAL SCHOOL

itp 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

MM Instituto de Medicina  
Molecular | João Lobo  
Antunes

FACULDADE DE  
MEDICINA  
LISBOA

1290  
UNIVERSIDADE D  
COIMBRA

Ciências  
ULisboa



Universidade do Minho

NOVA  
NOVA UNIVERSITY  
LISBON

universidade  
de zzeiro

TÉCNICO  
LISBOA

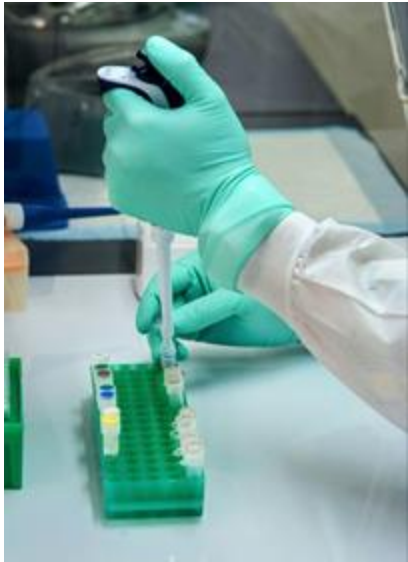
► If you're affiliated with one of our associates, you can benefit from our services



# How does a patient DNA sample get into a computer?



Collection



DNA isolation  
& library prep



Sequencing

FASTQ



Assembly

# Example FASTQ file

```
1_control_18S_2019_minq7.fastq
~/Downloads

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

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		



# 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 CEB-UMinho in Braga

**How can we help from here?**

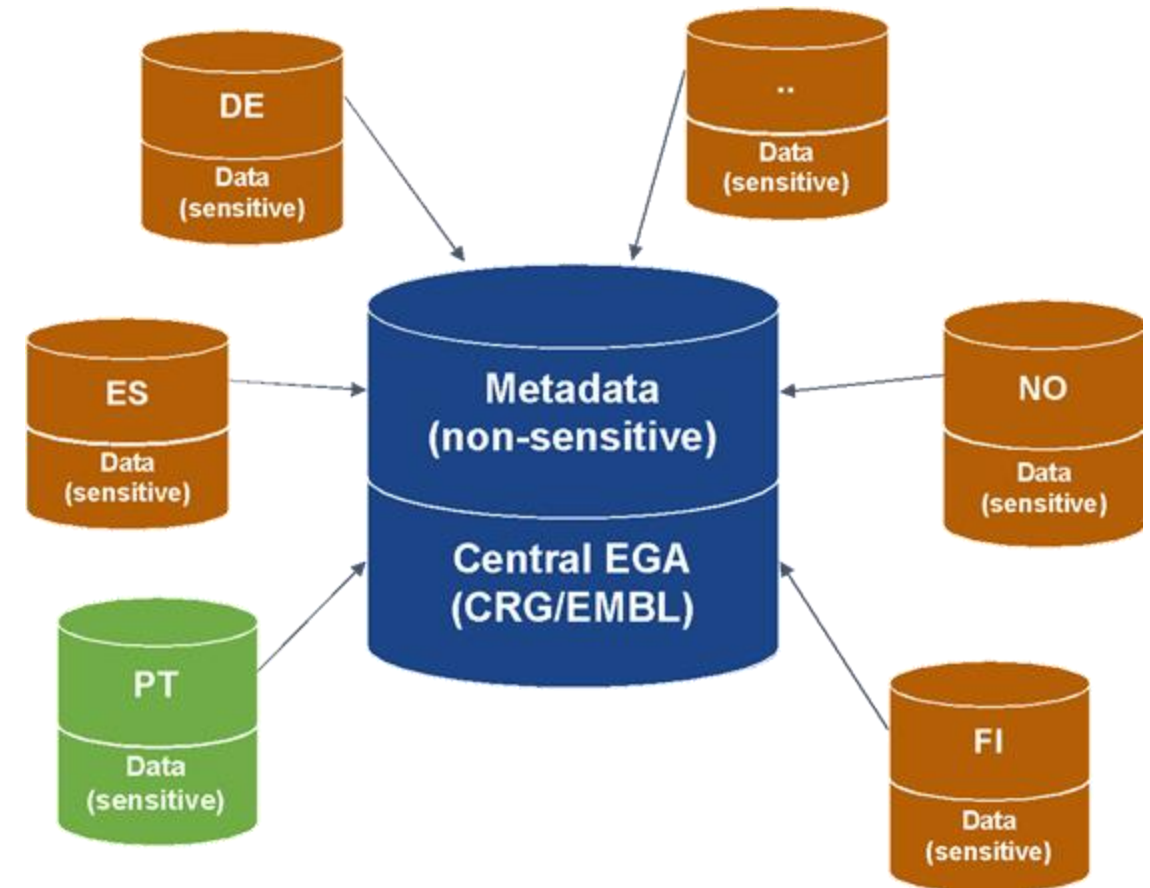


# **Federated EGA, a resource for discovery and access of human data across national borders**

<https://www.youtube.com/watch?v=bRFHyl6hnNk>

## 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.





# 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](https://beacon-project.io)



[beacon.biodata.pt](https://beacon.biodata.pt)



- ▶ Query all ELIXIR's beacons  
[beacon-network.elixir-europe.org](https://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 of them are women?
3. How many of the female patients are of African ethnicity?
4. 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:

## 1. A powerful enough computer...

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

- 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

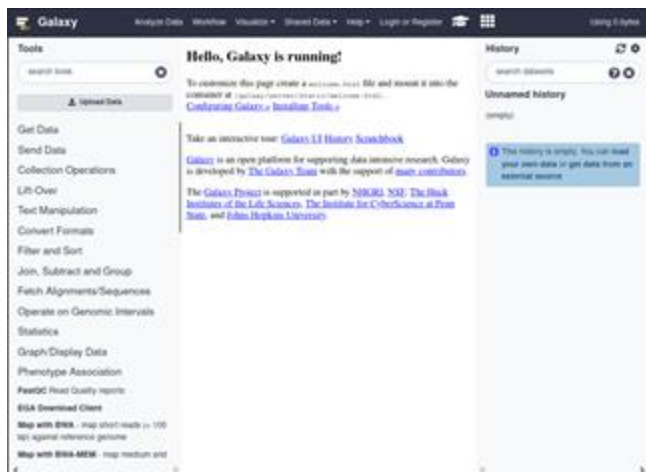
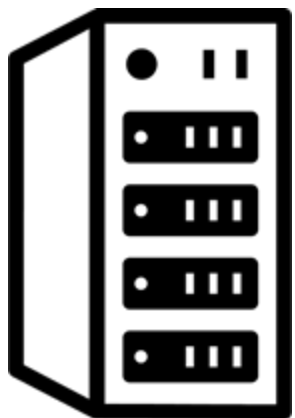


- ▶ Galaxy is a free, open-source system for analyzing data, authoring workflows, training and education, publishing tools, managing infrastructure, and more.

[galaxyproject.org](http://galaxyproject.org)

# Galaxy & the Pulsar Network

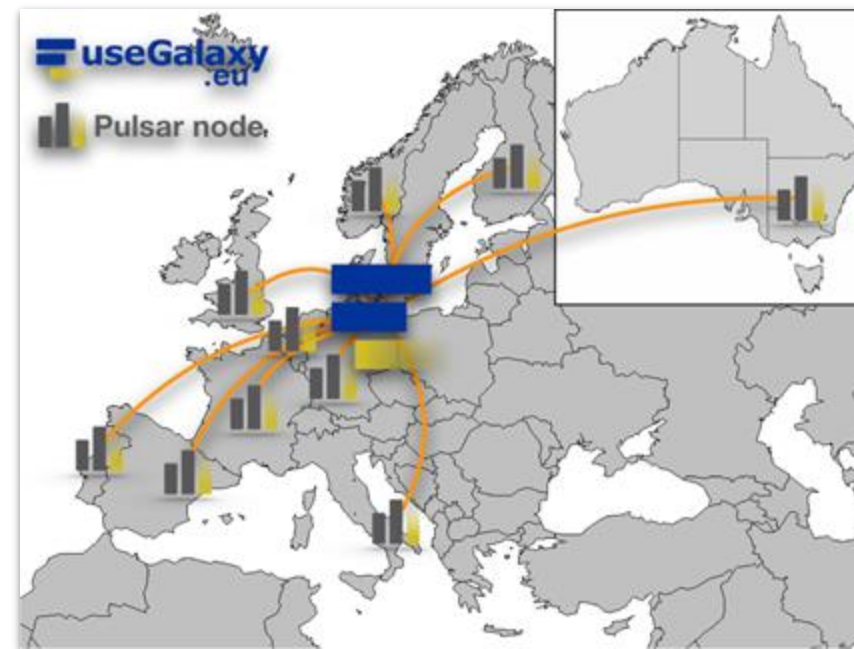
Self-hosted



Your institute's  
server

Galaxy software

Pulsar Network



usegalaxy.eu

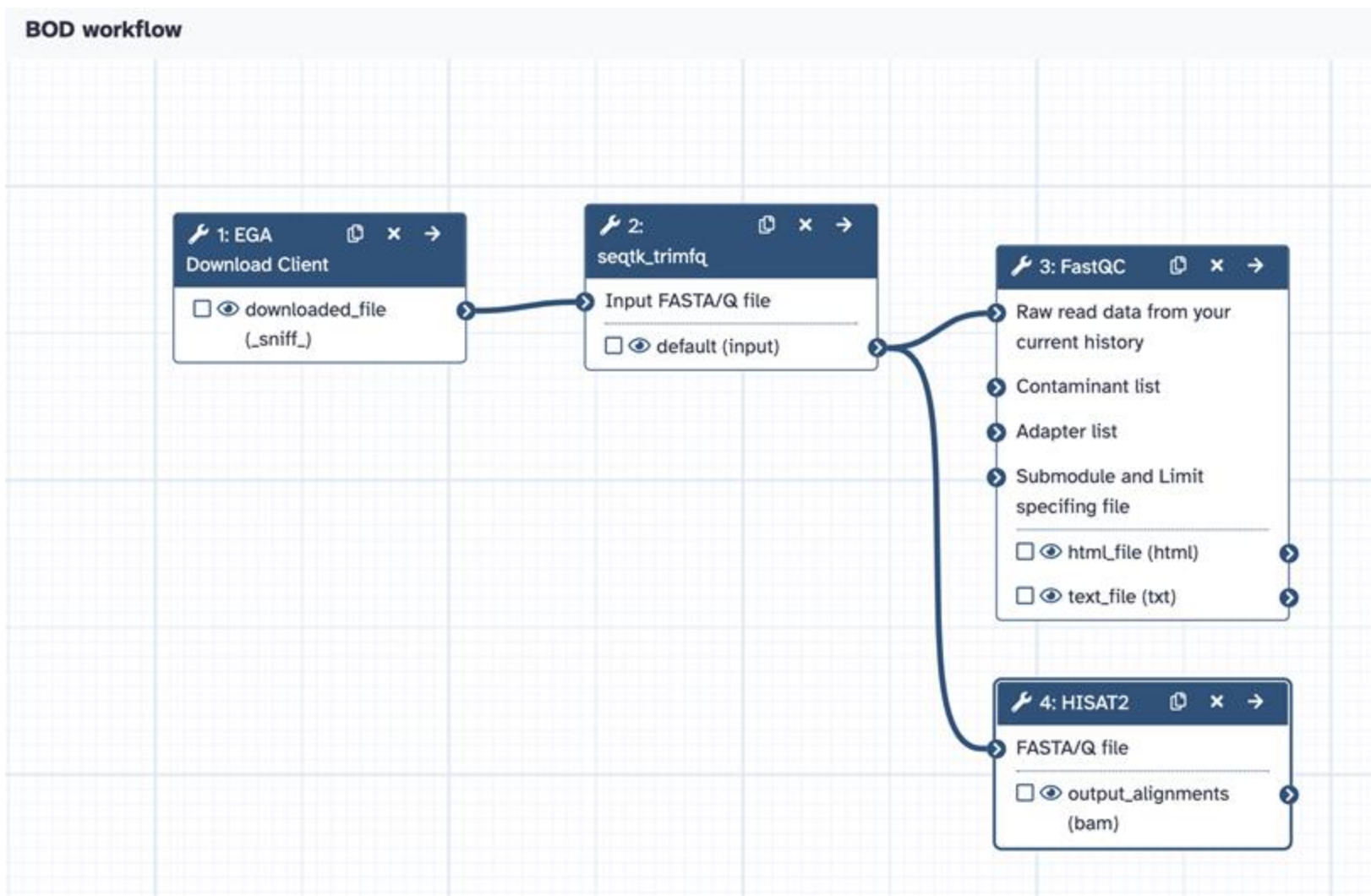
# Challenge 2: build a workflow in Galaxy





# Compare with mine

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



# Step 3: Learning more



- ▶ Find training courses and materials from ELIXIR Nodes and institutes

[tess.elixir-europe.org](https://tess.elixir-europe.org)

# Challenge 3: Diving into TeSS

I, Gil Poiares-Oliveira was one of the trainers of a course taught somewhere in Portugal. Starting from TeSS, can you tell me...

- What was it called?
- Where was it? What was the venue?
- Is the course targeted at Master's students?
- Module 2 of the course was "Data Collection" and it featured an hands-on activity. What's the name of the tool we used?

# 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



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

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](https://github.com/BioData-PT/computational-tools-resources)

**Gil Poiares-Oliveira, MSc**

**[gpo@biodata.pt](mailto:gpo@biodata.pt)**

INESC-ID | BioData.pt | ELIXIR Portugal



**BioData.pt**  
*Living data*

[biodata.pt](https://biodata.pt) | [info@biodata.pt](mailto:info@biodata.pt) | (+351) 937 990 500

