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

Bioinformatics Open Days 2024

Universidade do Minho
Escola de Engenharia

What is a research infrastructure?

Large Hadron Collider - CERN, Switzerland

Image: © 2014 CERN. Dominguez, Daniel

ELIXIR Nodes - ELIXIR, Europe

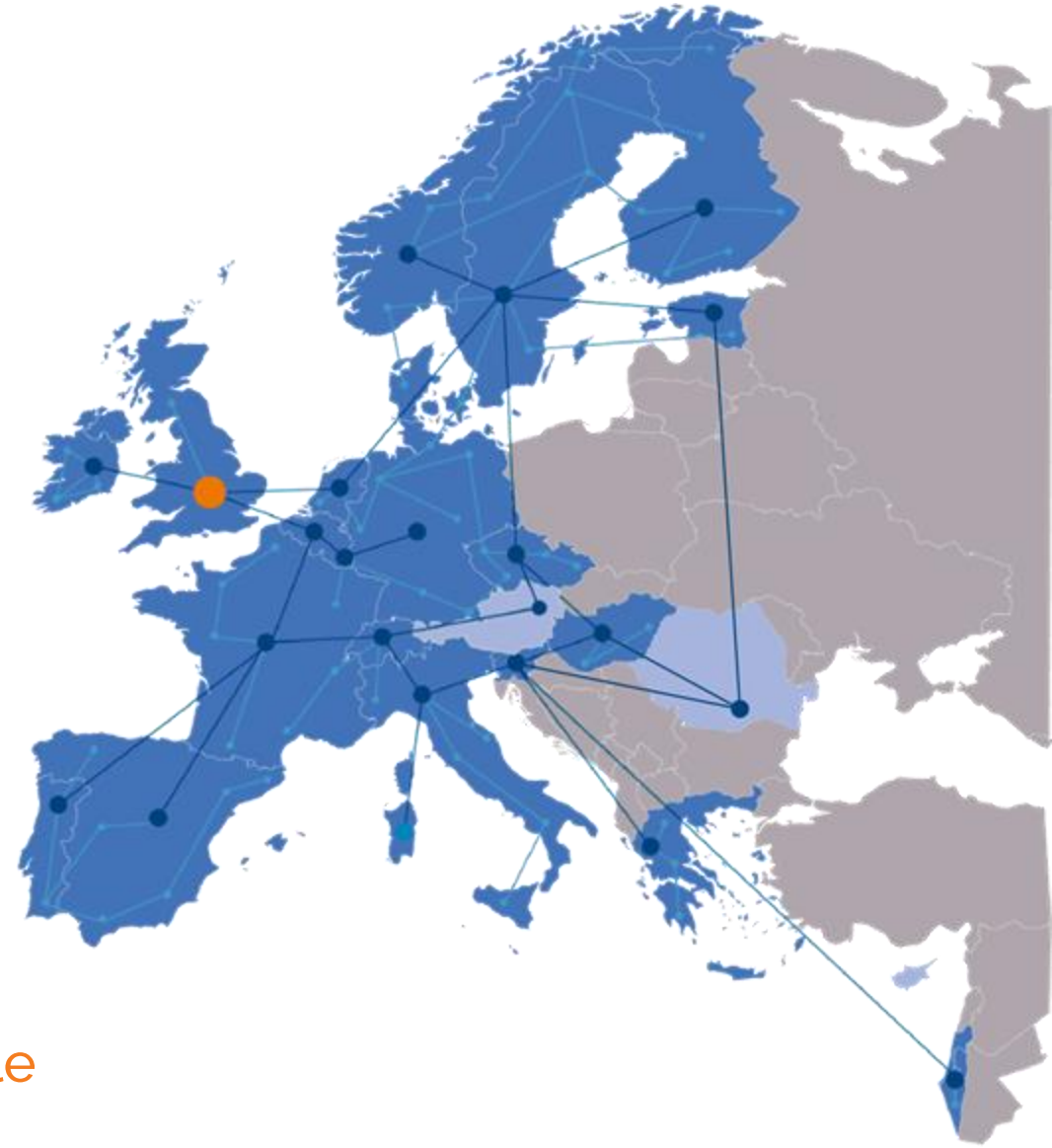
Image: © 2024 ELIXIR

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

Computation

servers, tool hosting

Mentorship

1-on-1 sessions

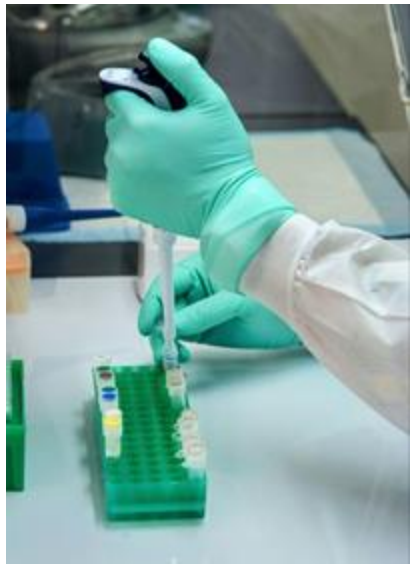
BioData.pt | elixir PORTUGAL

Our associates

- 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?

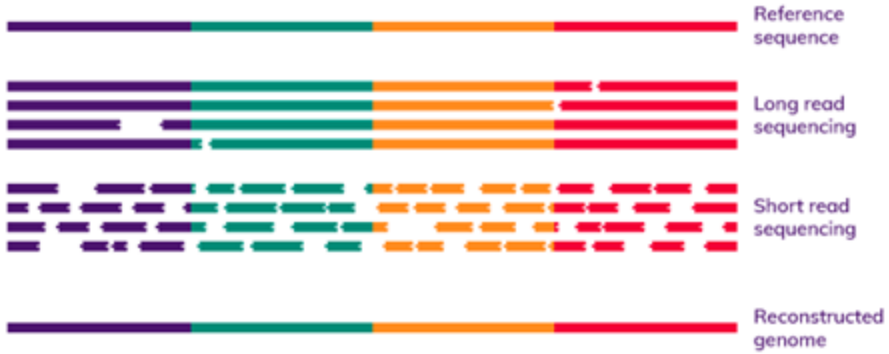
Collection



DNA isolation & library prep

Sequencing

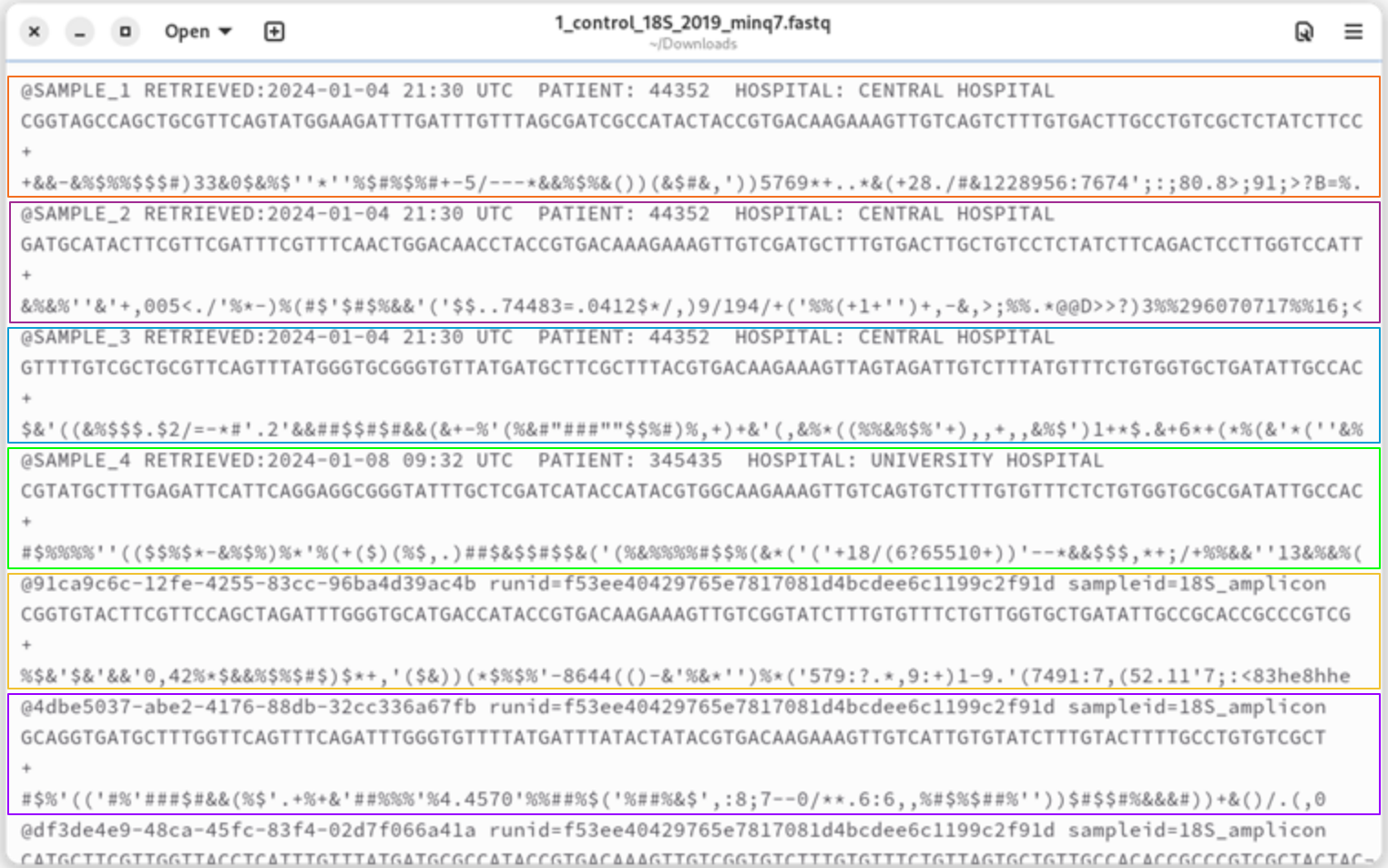
FASTQ



Assembly

BioData.pt | elixir PORTUGAL

Example FASTQ file



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 CEB-UMinho in Braga

How can we help from here?

BioData.pt | elixir PORTUGAL

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

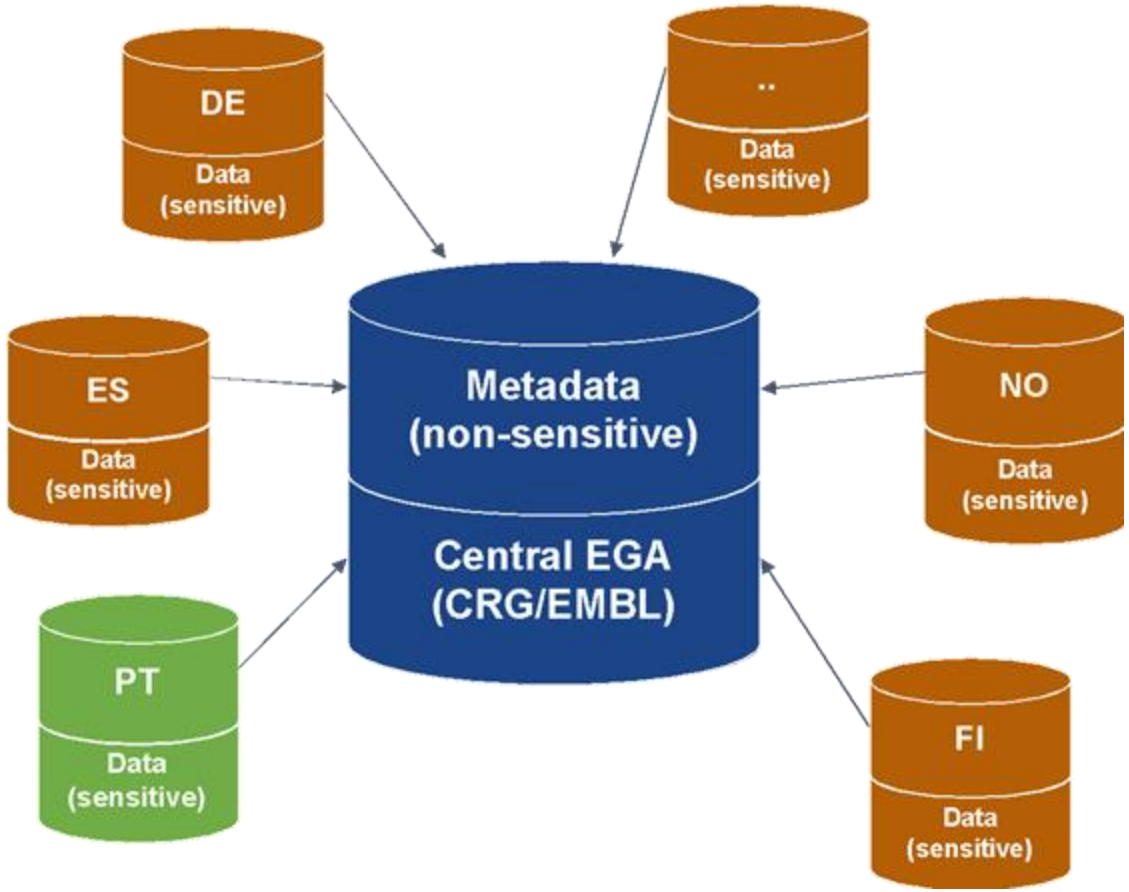
https://www.youtube.com/watch?v=bRFHyI6hnNk

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

BioData.pt | elixir PORTUGAL

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

   Tools:

   e.g. Trimmomatic, seqtk_trimq

2. Perform **quality control** on the data

   e.g. FastQC

3. **Map reads** against reference genome

   e.g. STAR, HISAT2, Bowtie2, TopHat2

4. **Analyse** according to biological question

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

   e.g. FreeBayes

Minimum requirements:

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

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

Galaxy PROJECT

BioData.pt | elixir PORTUGAL

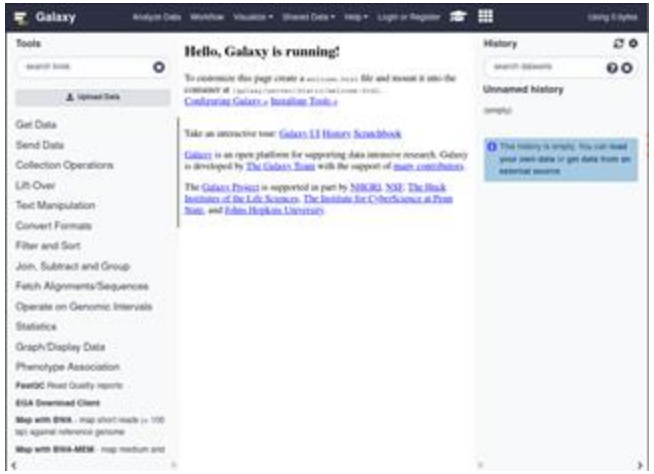
Galaxy Project

- 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

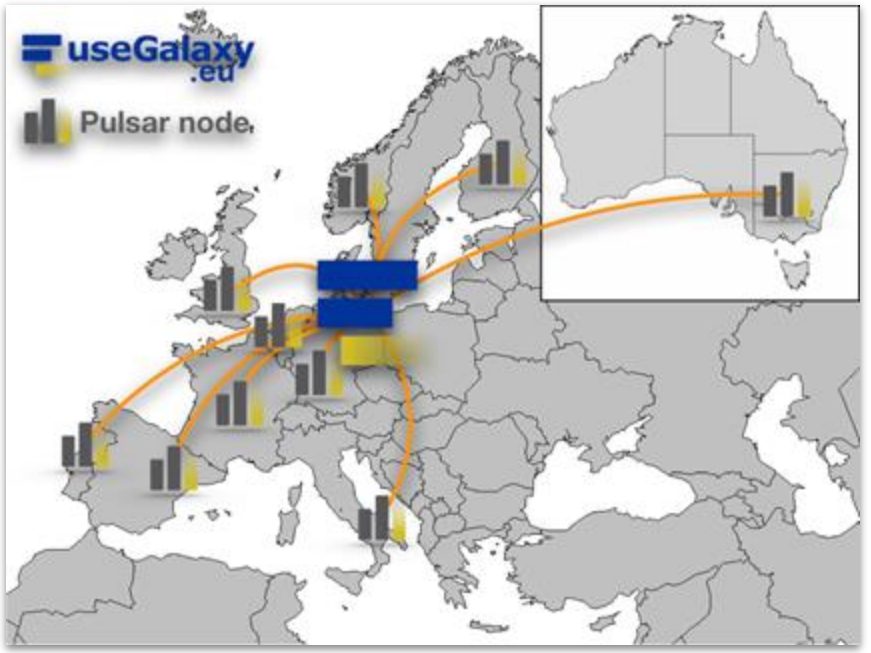
Self-hosted



Your institute's server

Galaxy software

Pulsar Network



usegalaxy.eu

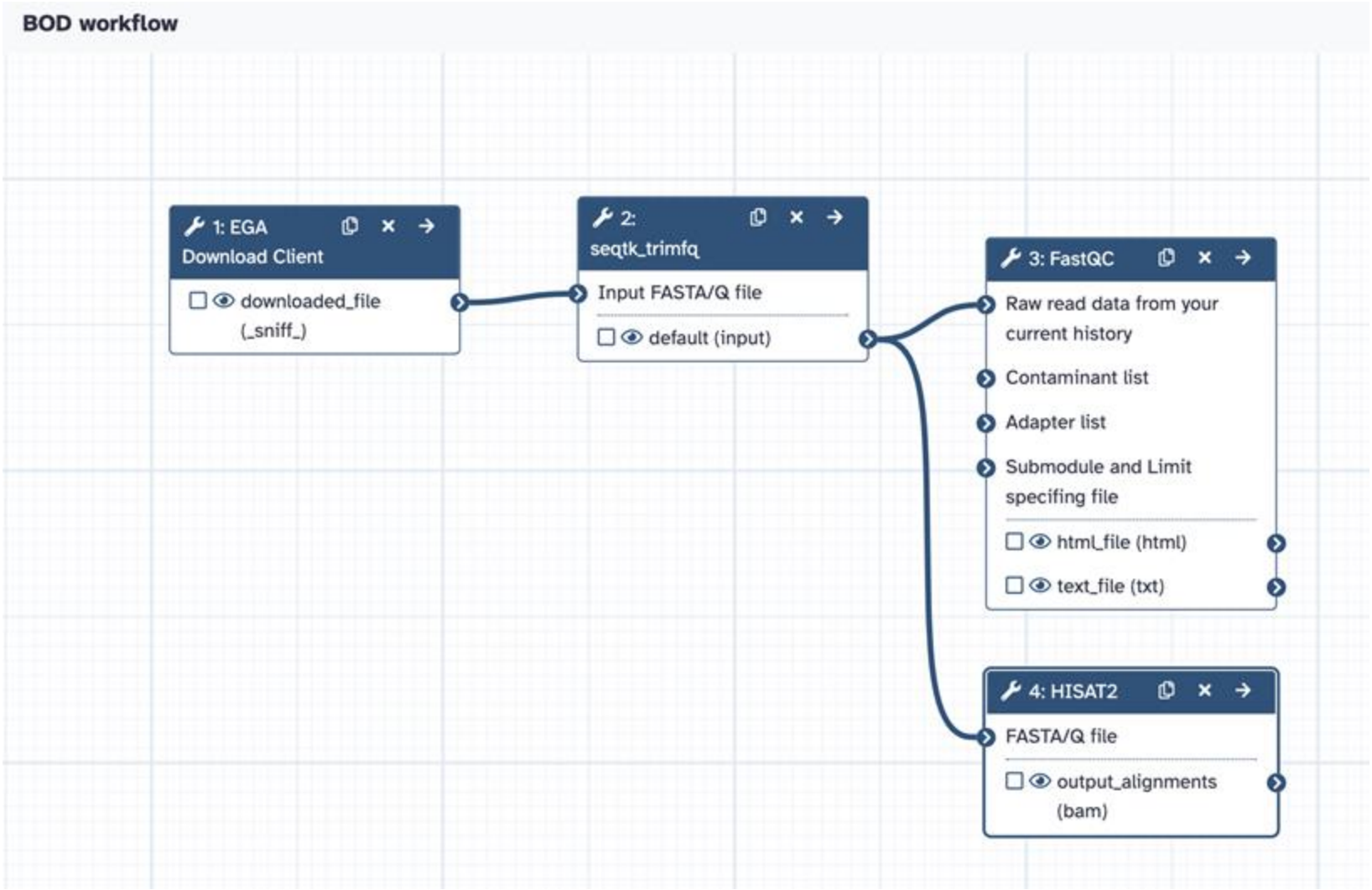
BioData.pt

elixir PORTUGAL

Challenge 2: build a workflow in Galaxy

Compare with mine

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



BioData.pt | elixir PORTUGAL

Step 3: Learning more

- Find training courses and materials from ELIXIR Nodes and institutes

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 o 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.

BioData.pt | ELIXIR PORTUGAL

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 :)

BioData.pt | elixir PORTUGAL

Thanks!

This presentation will 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

biodata.pt | info@biodata.pt | (+351) 937 990 500