ELIXIR Cloud and Compute and the Global Alliance for Genomics and Health (GA4GH)

ICRI EGI: Enabling global research with interoperable digital infrastructures

19 October 2022

elixir

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Jonathan Tedds (ELIXIR Hub)
ELIXIR Compute, Tools Platform & EOSC Coordinator

jonathan.tedds@elixir-europe.org

www.elixir-europe.org

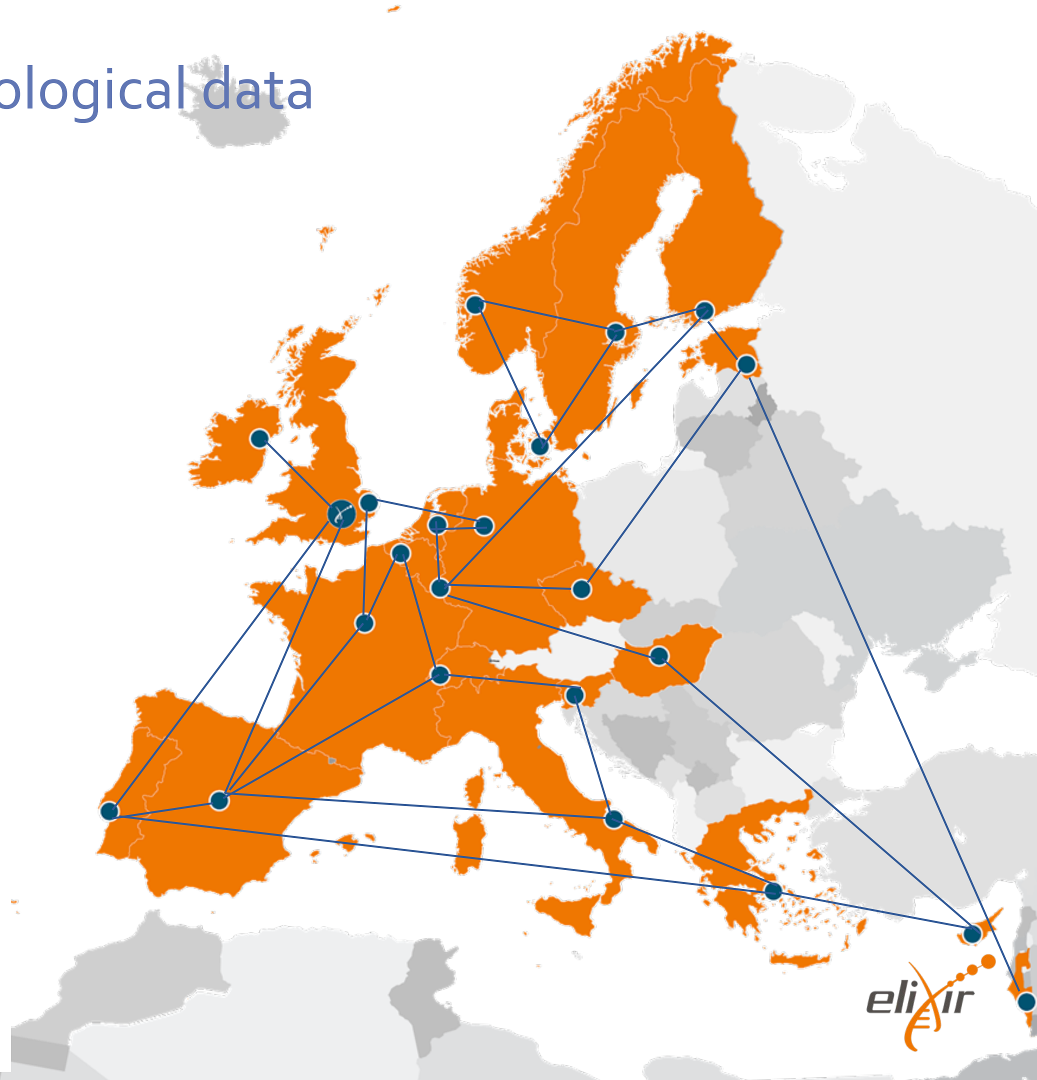
A sustainable infrastructure for biological data

ELIXIR Members

- Belgium
- Czech Republic
- Denmark
- EMBL
- Estonia
- Finland
- France
- Germany
- Greece
- Hungary
- Ireland
- Italy
- Luxembourg
- Israel
- Netherlands
- Norway
- Slovenia
- Portugal
- Spain
- Sweden
- Switzerland
- United Kingdom

ELIXIR Observers

- Cyprus



Coordinating services

Databases and Data Resources

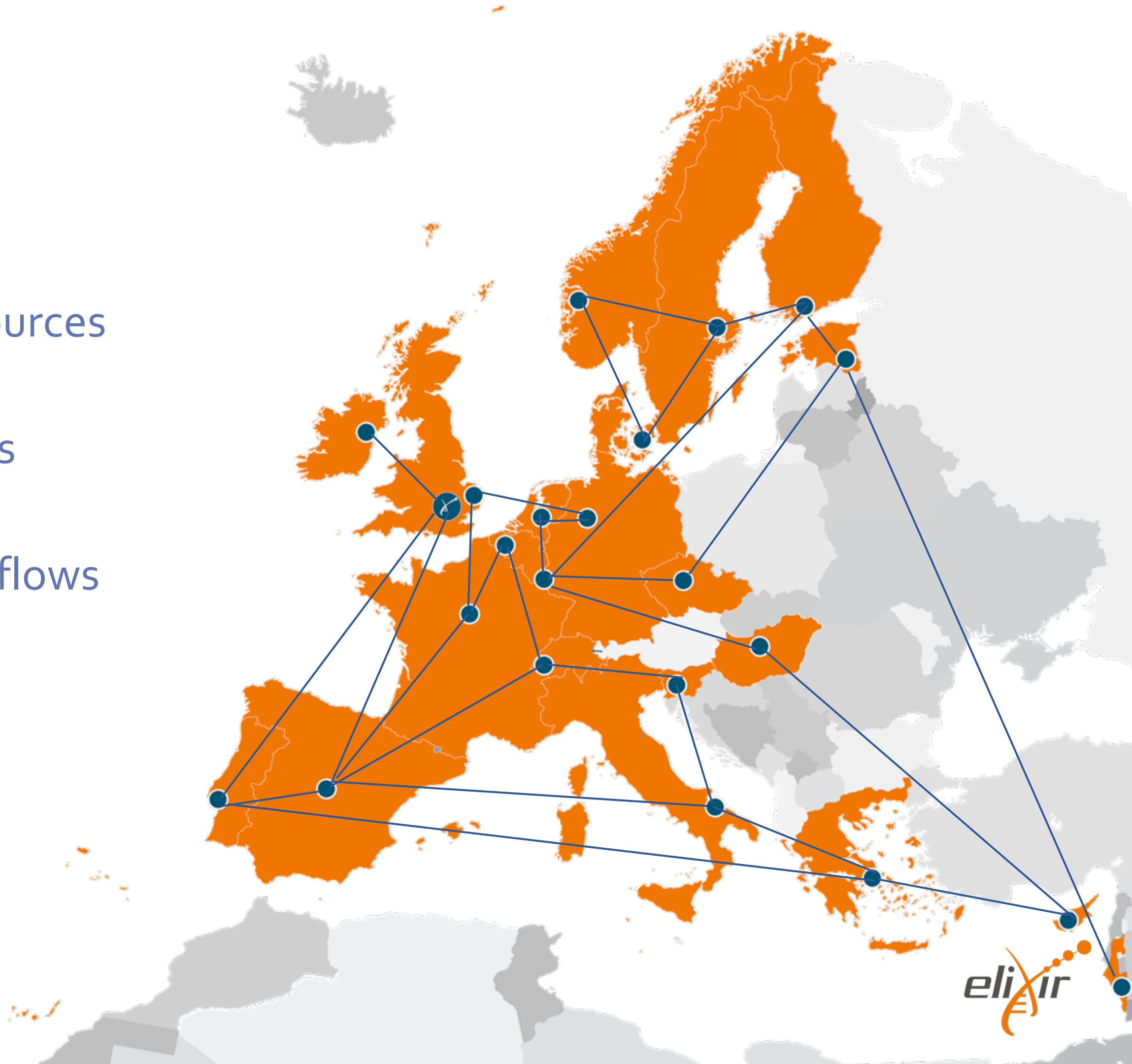
Interoperability Resources

Tools, software and workflows

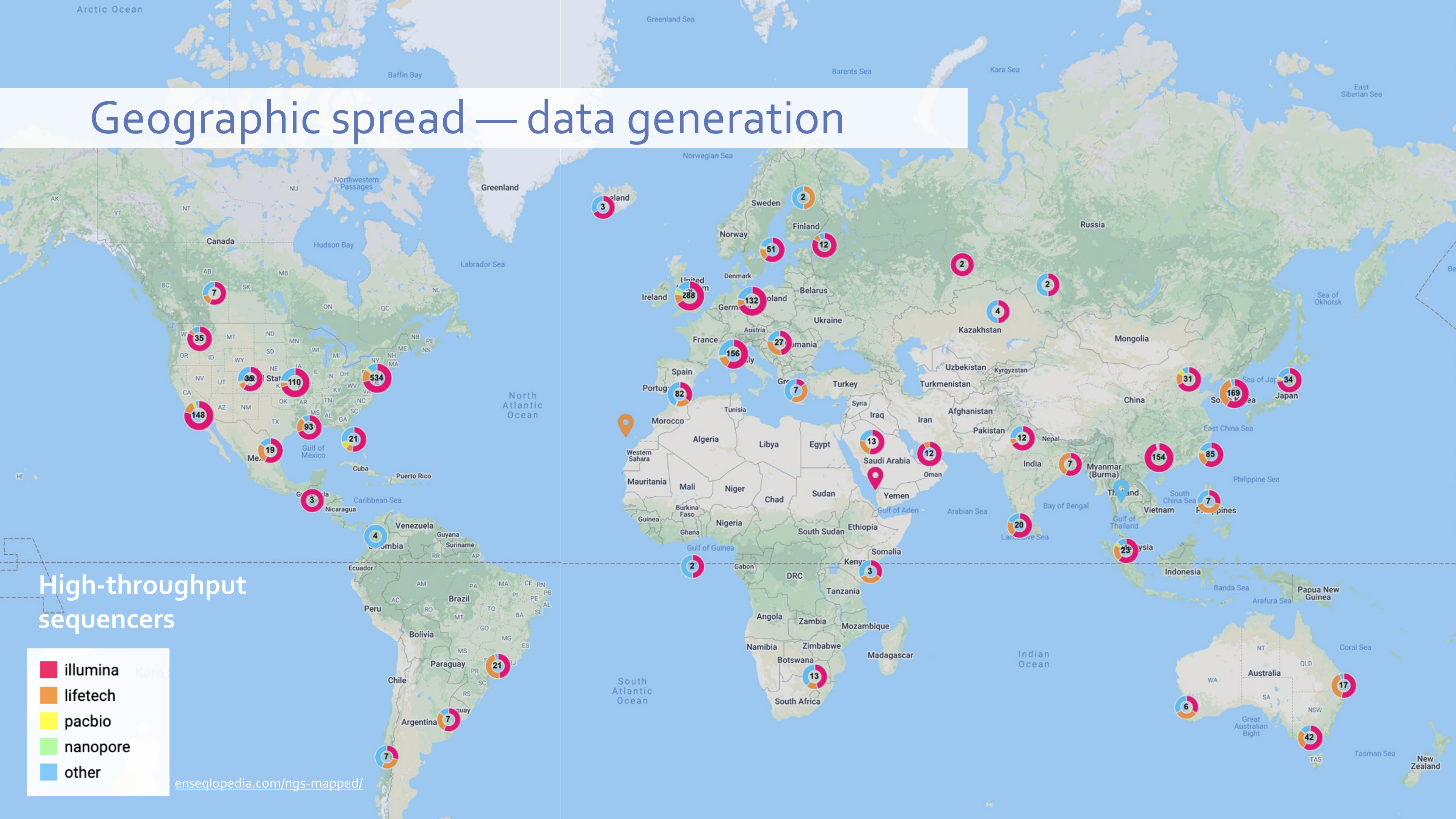
Compute Capabilities

Training Opportunities

Data management



Geographic spread — data generation



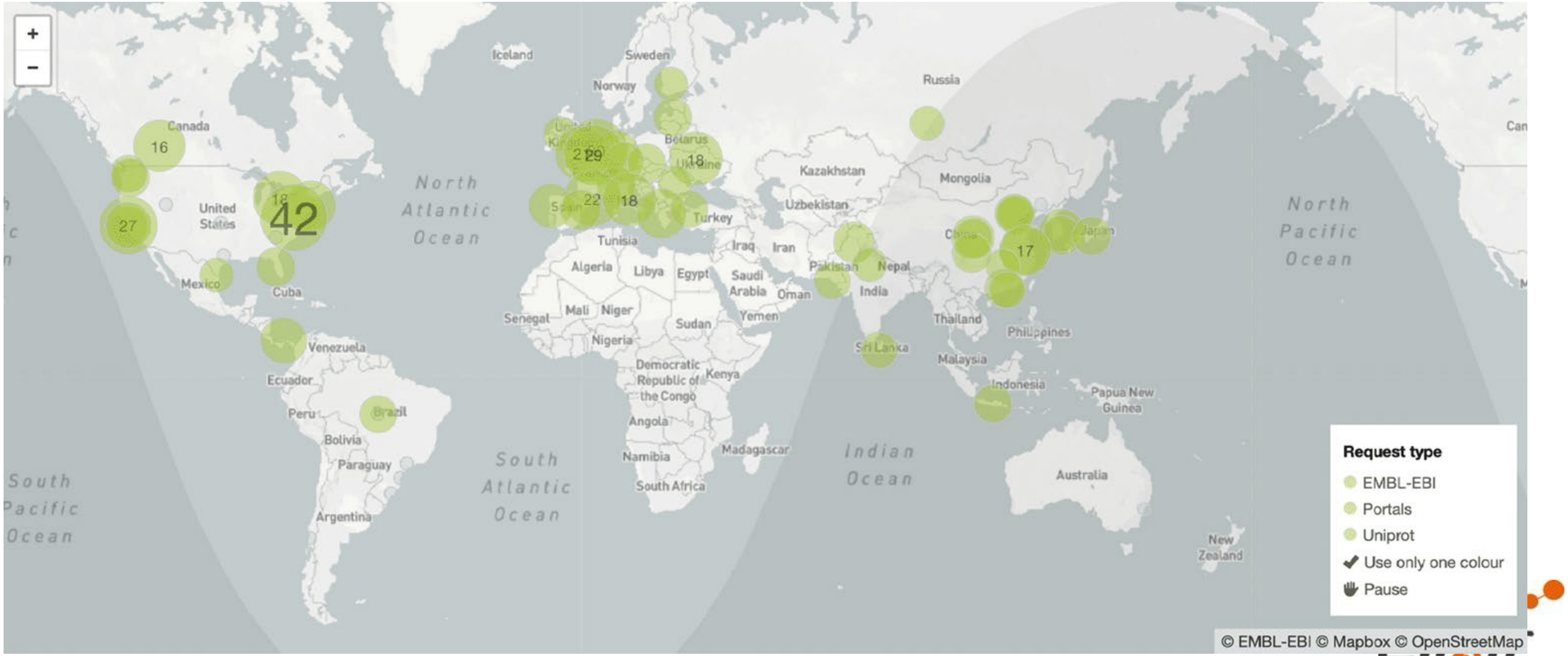
High-throughput sequencers

- 🟥 illumina
- 🟧 lifetech
- 🟨 pacbio
- 🟩 nanopore
- 🟦 other

enseqlopedia.com/ngs-mapped/

Geographic spread — usage of data

EMBL-EBI live data map



© EMBL-EBI © Mapbox © OpenStreetMap

www.ebi.ac.uk/web/livemap/

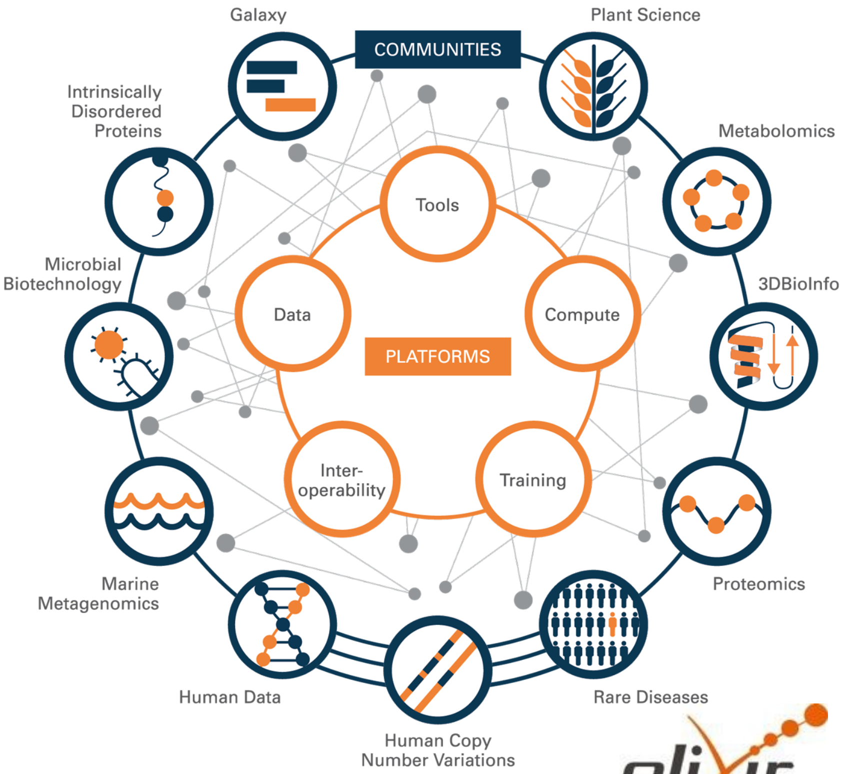
ELIXIR and the Compute Platform

ELIXIR Compute Platform

- Identity and access management
- Data Integration for Compute
- ELIXIR Hybrid Cloud Ecosystem
- Deploying Reproducible Containers and Workflows across Cloud Environments

ELIXIR Tools Platform

- Packaging, Containerisation and Deployment



Global Alliance for Genomics & Health

Enabling responsible genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

650+ member institutions in 50+ countries

"Collaborate on standards, compete on implementations!"

- Holistic approach bringing together all stakeholders: patients, healthcare professionals, industry, academia, lawmakers...
- Developing modular, functional and interoperable community standards & guidelines driven by [24 Driver Projects]
- Organized in Work Streams (e.g., Cloud, Data Security, Regulatory & Ethics)
- ELIXIR is a Strategic Partner of GA4GH

How we work

Represent ELIXIR stakeholders in GA4GH & **promote** GA4GH standards within ELIXIR

Prototype real-world use cases with ELIXIR stakeholders, **develop** PoCs & **deploy** at ELIXIR nodes

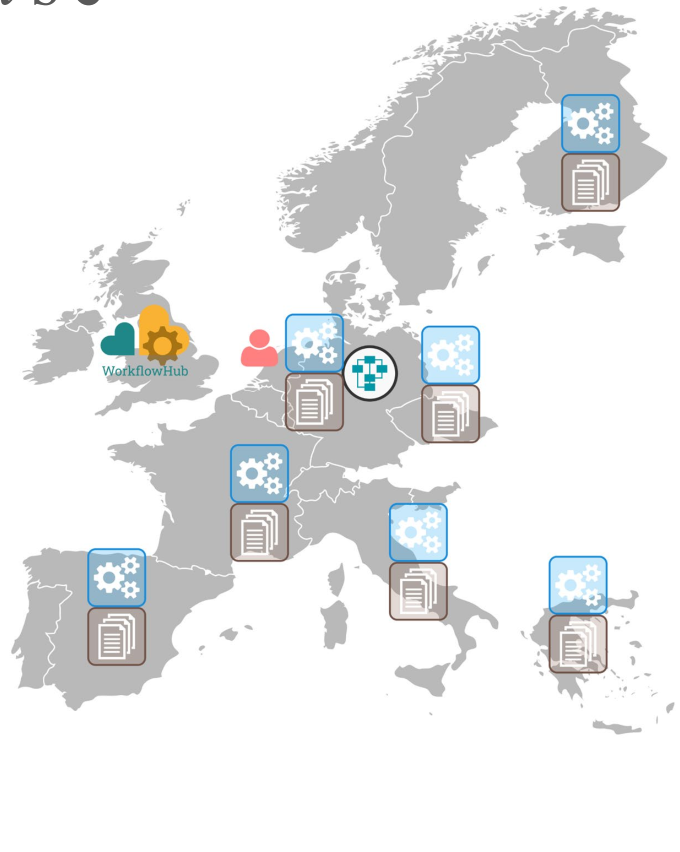
Consult on integrating GA4GH standards into existing solutions and provide **technical support**

Interoperability testing with third party GA4GH-powered solutions

An example use case

?

1. Process sensitive data across multiple ELIXIR nodes (workflow)
2. Collect processed data ( )
3. Centrally analyze at host institution (workflow)



How to federate clouds?

Adoption of

➢GA4GH standards

➢Common governance & security guidelines

Core services

to further harmonize behavior & facilitate integration

Reference implementations & auxiliary services

to bridge gaps & facilitate uptake

GA4GH Cloud and Access Management APIs

Passport (AAI)

Grant access to data & compute

TRS: Tool Registry Service API

Access workflows and container images

DRS: Data Repository Service API

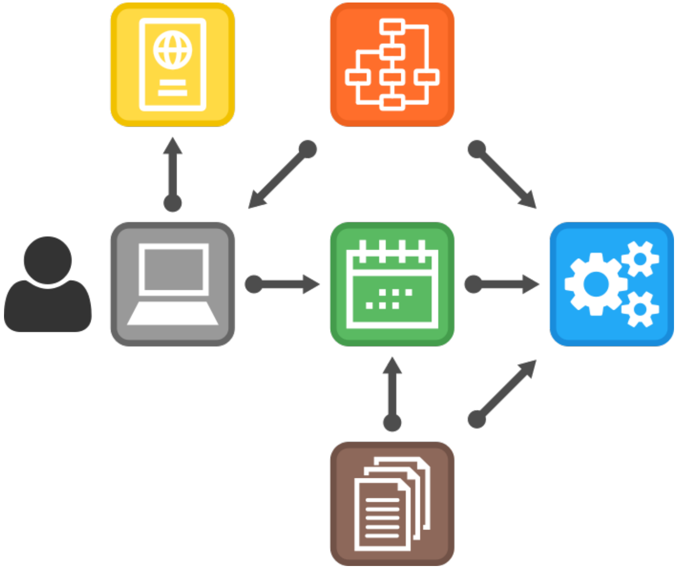
Access to data sets

WES: Workflow Execution Service API

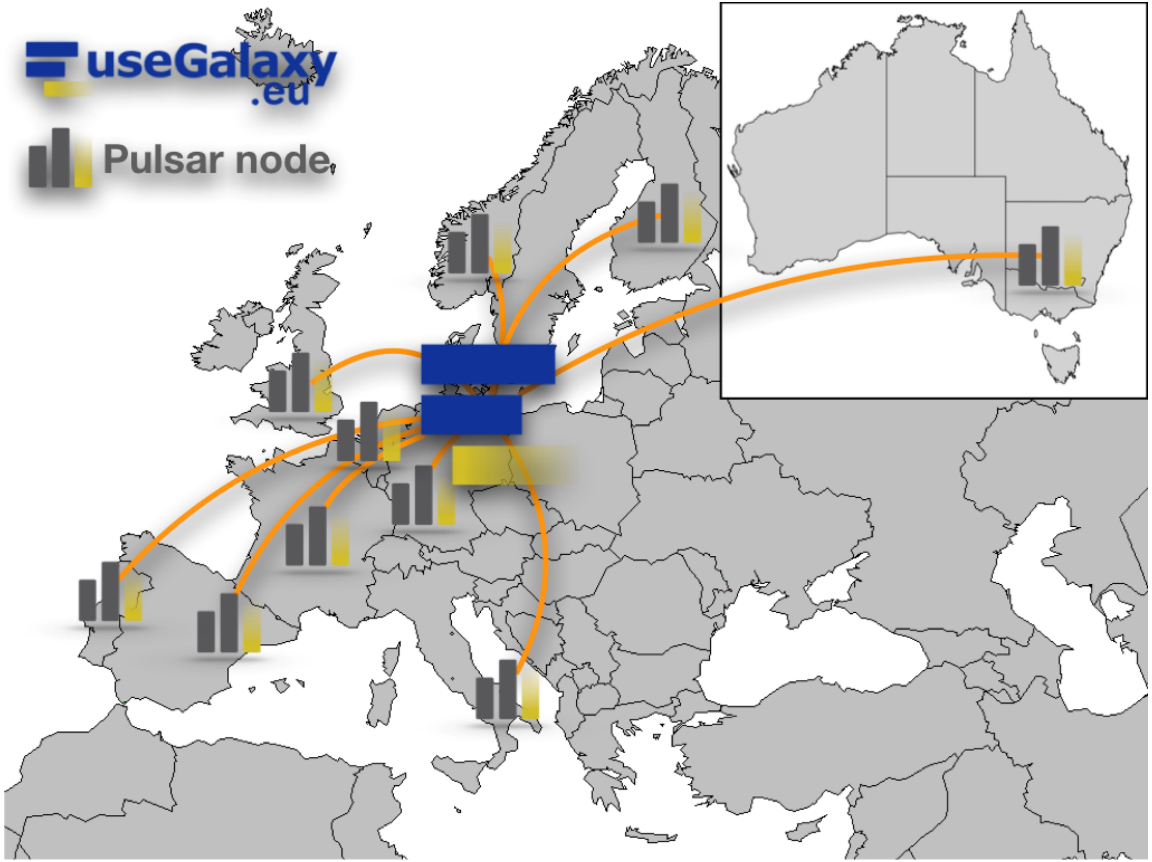
Interpret workflows & schedule task execution

TES: Task Execution Service API

Execute tasks



Technological Implementations:-Galaxy Network Pulsar



UNI FREIBURG

de.NBI

EOSC-Life

GenOuest

Universitas Bergensis

EOSC-Pillar
Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

diamond

cesnet

EOSC NORDIC

CSC

IBIOM
Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies

TÉCNICO LISBOA

VIB

ReCaS BARI
Università degli Studi di Bari Aldo Moro
INFN

Vlaams Supercomputer Centrum
Vlaanderen is supercomputing

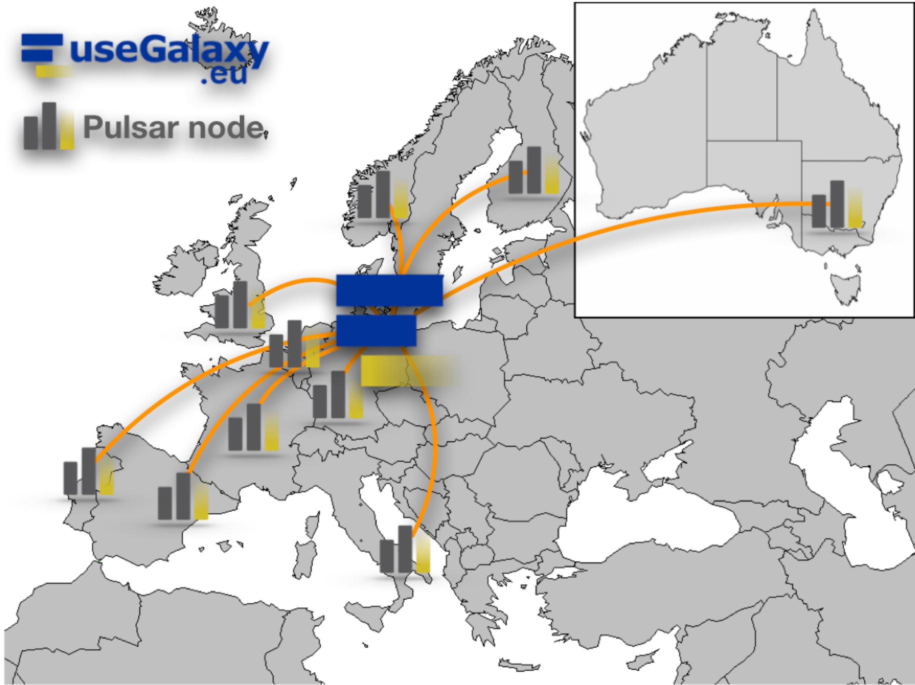
Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

https://pulsar-network.readthedocs.io

Galaxy Community becomes an EOSC Infrastructure…

- deferred data / remote data
- distributed storage
- scheduling taking storage into account

HORIZON-INFRA-2021-EOSC-01 – EuroScienceGateway just kicked off ☺



Workflow registry
https://workflowhub.eu

Workflows

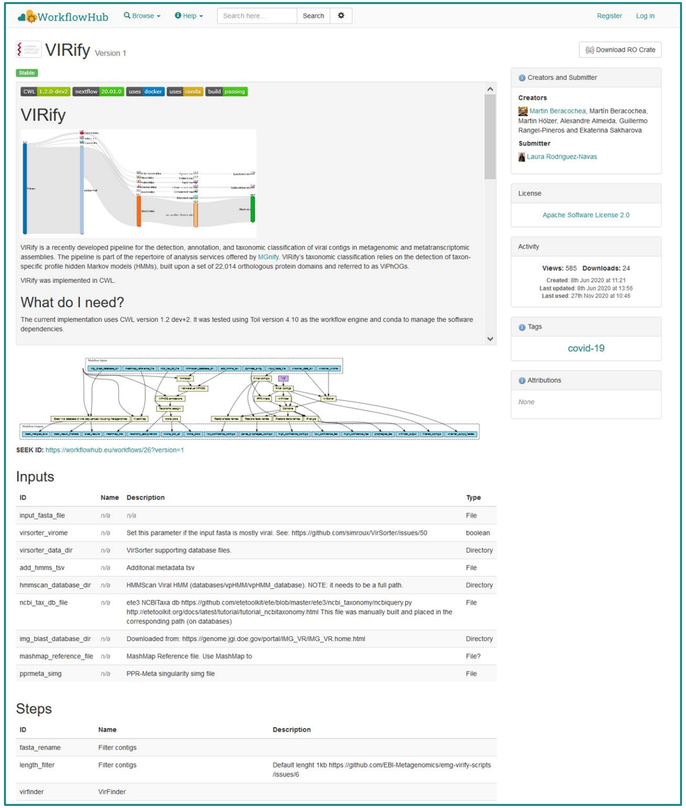
- Any kind of system
- May remain in home repository
- Linked with data, documentation
- Can have DOIs

WorkflowHub

- Links with other system, import/export (GA4GH TRS, RO-Crate)
- Metadata standards
- Github integration

Mixed depth of support for Workflow Management Systems

- Lift metadata from different systems
- Coupled to execution platforms



VIRify Version 1

Creators and Submitter

Creators

Martin Beracochea, Martín Beracochea, Martin Hölzer, Alexandre Almeida, Guillermo Rangel-Pineros and Ekaterina Sakharova

Submitter

Laura Rodriguez-Navas

License

Apache Software License 2.0

Activity

Views: 585 **Downloads:** 24

Created: 8th Jun 2020 at 11:21
Last updated: 8th Jun 2020 at 13:56
Last used: 27th Nov 2020 at 10:46

Tags

covid-19

Attributions

None

VIRify

VIRify is a recently developed pipeline for the detection, annotation, and taxonomic classification of viral contigs in metagenomic and metatranscriptomic assemblies. The pipeline is part of the repertoire of analysis services offered by MGnify. VIRify's taxonomic classification relies on the detection of taxon-specific profile hidden Markov models (HMMs), built upon a set of 22,014 orthologous protein domains and referred to as ViPhOGs.

VIRify was implemented in CWL.

What do I need?

The current implementation uses CWL version 1.2 dev+2. It was tested using Toil version 4.10 as the workflow engine and conda to manage the software dependencies.

SEEK ID: https://workflowhub.eu/workflows/26?version=1

Inputs

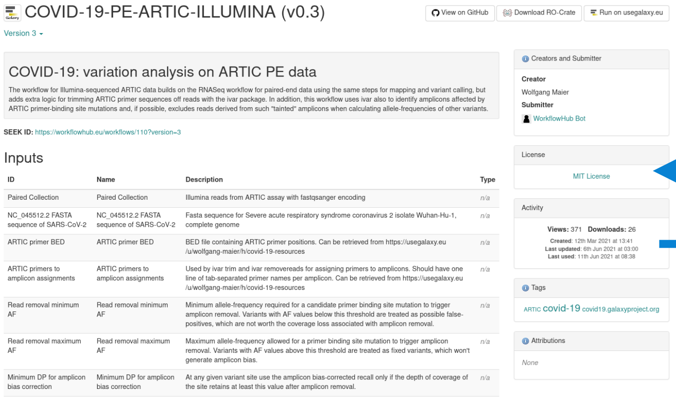
ID	Name	Description	Type
input_fasta_file	*n/a*	*n/a*	File
virsorter_virome	*n/a*	Set this parameter if the input fasta is mostly viral. See: https://github.com/simroux/VirSorter/issues/50	boolean
virsorter_data_dir	*n/a*	VirSorter supporting database files.	Directory
add_hmms_tsv	*n/a*	Additonal metadata tsv	File
hmmscan_database_dir	*n/a*	HMMScan Viral HMM (databases/vpHMM/vpHMM_database). NOTE: it needs to be a full path.	Directory
ncbi_tax_db_file	*n/a*	ete3 NCBITaxa db https://github.com/etetoolkit/ete/blob/master/ete3/ncbi_taxonomy/ncbiquery.py http://etetoolkit.org/docs/latest/tutorial/tutorial_ncbitaxonomy.html This file was manually built and placed in the corresponding path (on databases)	File
img_blast_database_dir	*n/a*	Downloaded from: https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html	Directory
mashmap_reference_file	*n/a*	MashMap Reference file. Use MashMap to	File?
pprmeta_simg	*n/a*	PPR-Meta singularity simg file	File

Steps

ID	Name	Description
fasta_rename	Filter contigs	
length_filter	Filter contigs	Default lenght 1kb https://github.com/EBI-Metagenomics/emg-virify-scripts /issues/6
virfinder	VirFinder	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

Coupled to execution environments



COVID-19-PE-ARTIC-ILLUMINA (v0.3)

View on GitHub | Download RO-Crate | Run on usegalaxy.eu

Version 3

COVID-19: variation analysis on ARTIC PE data

The workflow for Illumina-sequenced ARTIC data builds on the RNASeq workflow for paired-end data using the same steps for mapping and variant calling, but adds extra logic for trimming ARTIC primer sequences off reads with the ivar package. In addition, this workflow uses ivar also to identify amplicons affected by ARTIC primer-binding site mutations and, if possible, excludes reads derived from such "tainted" amplicons when calculating allele-frequencies of other variants.

SEEK ID: https://workflowhub.eu/workflows/110?version=3

Inputs

ID	Name	Description	Type
Paired Collection	Paired Collection	Illumina reads from ARTIC assay with fastqsanger encoding	*n/a*
NC_045512.2 FASTA sequence of SARS-CoV-2	NC_045512.2 FASTA sequence of SARS-CoV-2	Fasta sequence for Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome	*n/a*
ARTIC primer BED	ARTIC primer BED	BED file containing ARTIC primer positions. Can be retrieved from https://usegalaxy.eu /u/wolfgang-maier/h/covid-19-resources	*n/a*
ARTIC primers to amplicon assignments	ARTIC primers to amplicon assignments	Used by ivar trim and ivar removereads for assigning primers to amplicons. Should have one line of tab-separated primer names per amplicon. Can be retrieved from https://usegalaxy.eu /u/wolfgang-maier/h/covid-19-resources	*n/a*
Read removal minimum AF	Read removal minimum AF	Minimum allele-frequency required for a candidate primer binding site mutation to trigger amplicon removal. Variants with AF values below this threshold are treated as possible false-positives, which are not worth the coverage loss associated with amplicon removal.	*n/a*
Read removal maximum AF	Read removal maximum AF	Maximum allele-frequency allowed for a primer binding site mutation to trigger amplicon removal. Variants with AF values above this threshold are treated as fixed variants, which won't generate amplicon bias.	*n/a*
Minimum DP for amplicon bias correction	Minimum DP for amplicon bias correction	At any given variant site use the amplicon bias-corrected recall only if the depth of coverage of the site retains at least this value after amplicon removal.	*n/a*

Creators and Submitter

Creator

Wolfgang Maier

Submitter

WorkflowHub Bot

License

MIT License

Activity

Views: 371 **Downloads:** 26

Created: 12th Mar 2021 at 13:41
Last updated: 6th Jun 2021 at 03:00
Last used: 11th Jun 2021 at 08:38

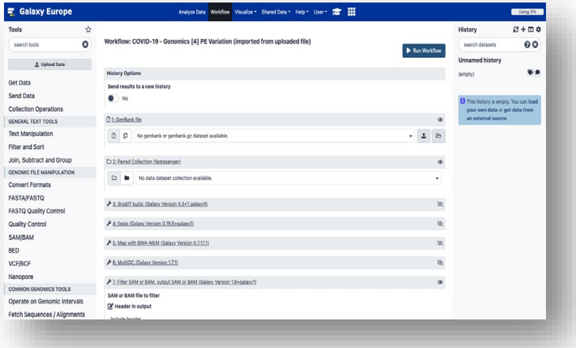
Tags

ARTIC covid-19 covid19.galaxyproject.org

Attributions

None

Run on usegalaxy.eu



Workflow: COVID-19 - Genomics [4] PE Variation (imported from uploaded file)

History Options

Send results to a new history

No

1: GenBank file

No genbank or genbank.gz dataset available.

2: Paired Collection (fastqsanger)

No data dataset collection available.

3: SnpEff build: (Galaxy Version 4.3+T.galaxy4)

4: fastp (Galaxy Version 0.19.5+galaxy1)

5: Map with BWA-MEM (Galaxy Version 0.7.17.1)

6: MultiQC (Galaxy Version 1.7.1)

7: Filter SAM or BAM, output SAM or BAM (Galaxy Version 1.8+galaxy1)

SAM or BAM file to filter

Header in output

Include header

Unnamed history

(empty)

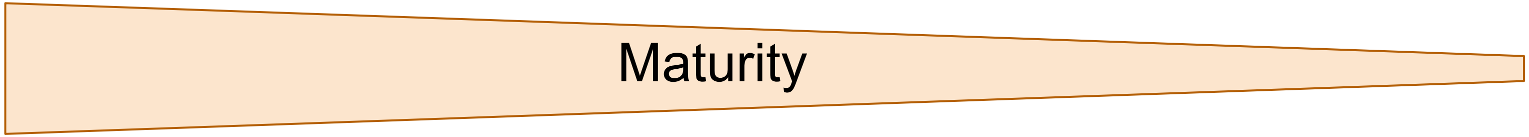
This history is empty. You can load your own data or get data from an external source

COMMON WORKFLOW LANGUAGE
SAPPORO WES
DDBJ
DNA Data Bank of Japan

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824087

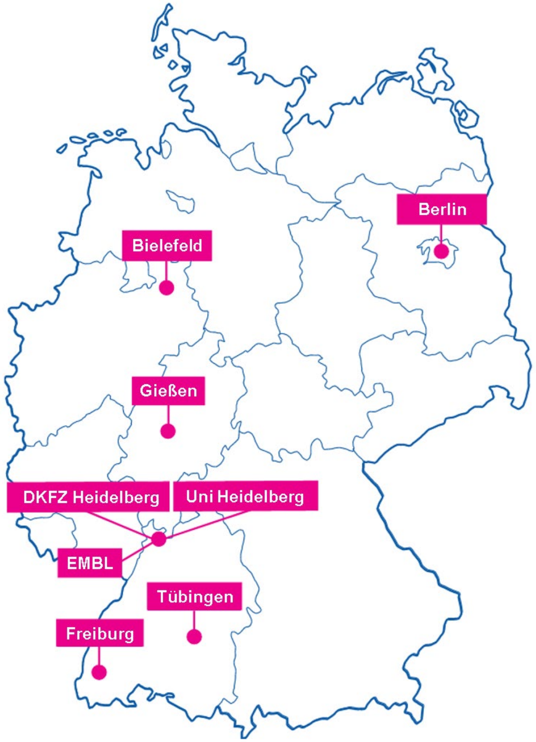
ELIXIR::GA4GH enabled Services & Solutions

Galaxy PROJECT	WorkflowHub	ELIXIR::GA4GH Cloud
Web-based platform for reproducible computational analysis	Registry for describing, sharing and publishing scientific computational workflows	Federated, interoperable network of workflow engines and compute nodes based on GA4GH standards
ELIXIR Community	EOSC-Life resource	GA4GH Driver Project
APIs & (third-party) GUIs	API & GUI	APIs & third-party GUIs



Maturity

National Implementations: de.NBI Cloud Federation



https://cloud.denbi.de

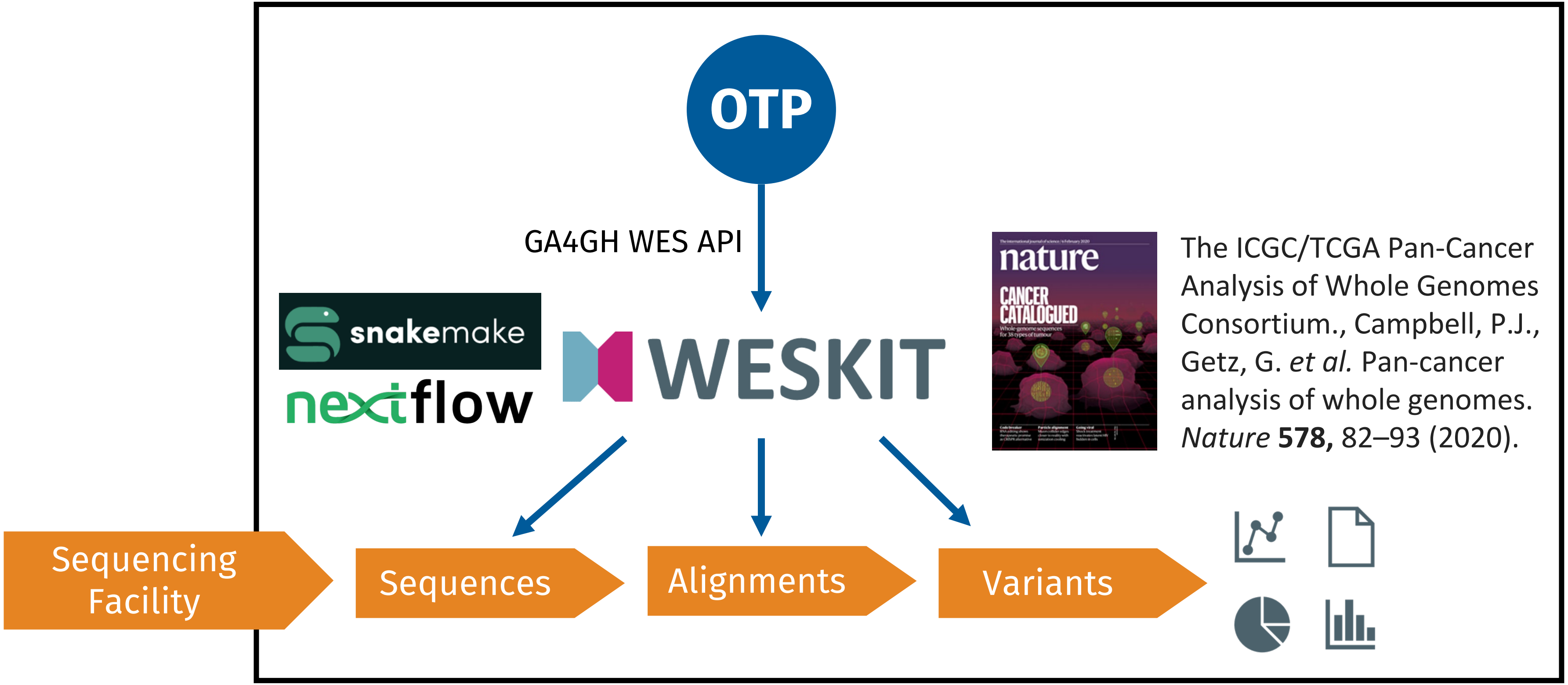
- Fully academic cloud federation
- Established 2016
- Provides storage and computing resources for the life sciences community
- Free of charge for academic use
- Federation is maintained by the eight German cloud centers located in Berlin, Bielefeld, Freiburg, Gießen, Heidelberg and Tübingen (+ FZ Jülich in 2022/2023)
- de.NBI Cloud offers a solution to enable integrative analyses, the efficient use of data in research, and computational capacities for bioinformatics training.

WESKIT

GA4GH WES implementation for Cancer Genomics
https://gitlab.com/one-touch-pipeline/weskit

Features

- GA4GH WES implementation
- Execution of Snakemake and Nextflow workflows
- Focus: stability and high data throughput
- Developed and used at BIH, DKFZ & Sanger
- HPC and cloud deployment
- OIDC support
- Developed in cooperation with ELIXIR compute platform



OTP

GA4GH WES API

snakemake

nextflow

WESKIT

Sequencing Facility

Sequences

Alignments

Variants

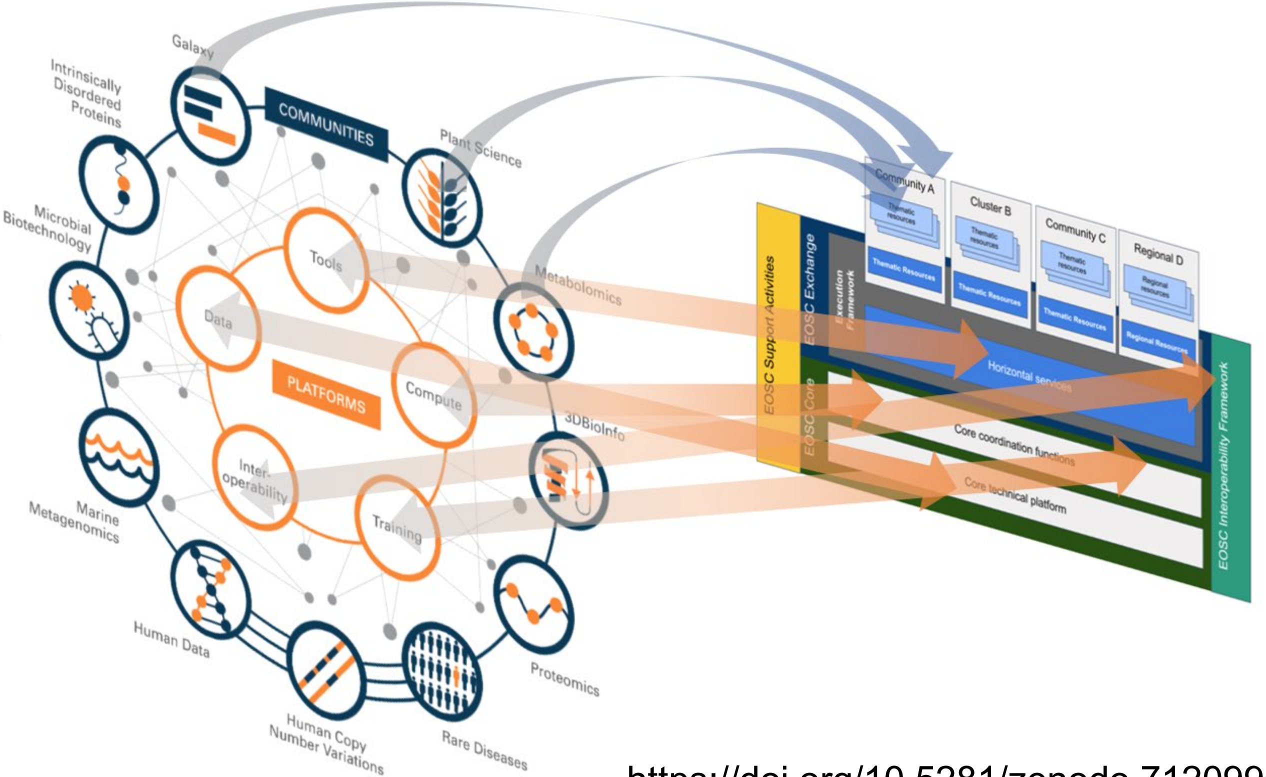
The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., Campbell, P.J., Getz, G. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (2020).

ELIXIR's EOSC Strategy: ELIXIR is an existing Infrastructure - EOSC is a delivery partner

ELIXIR's communities represent domain-specific expertise and resources.

ELIXIR's Commissioned Services represent an ecosystem which meets many of the same goals as EOSC.

We will co-develop EOSC, influence components where we have expertise and adopting technologies when they have demonstrated utility.

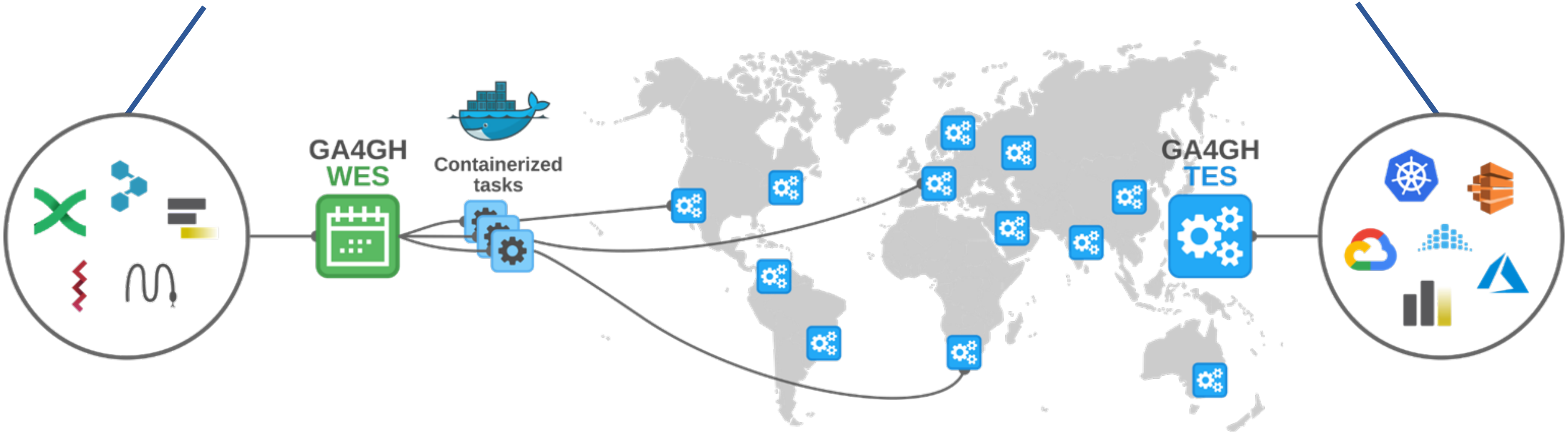


https://doi.org/10.5281/zenodo.7120996

ELIXIR is enabling federated analytics through the use of GA4GH Cloud API specifications

Different workflow languages

Different compute backends



"Send the compute to where the data is!"

Thanks

Contact: jonathan.tedds@elixir-europe.org

www.elixir-europe.org