Galaxy

make it run your jobs for you

Martin Čech – Telč – 28.3.25

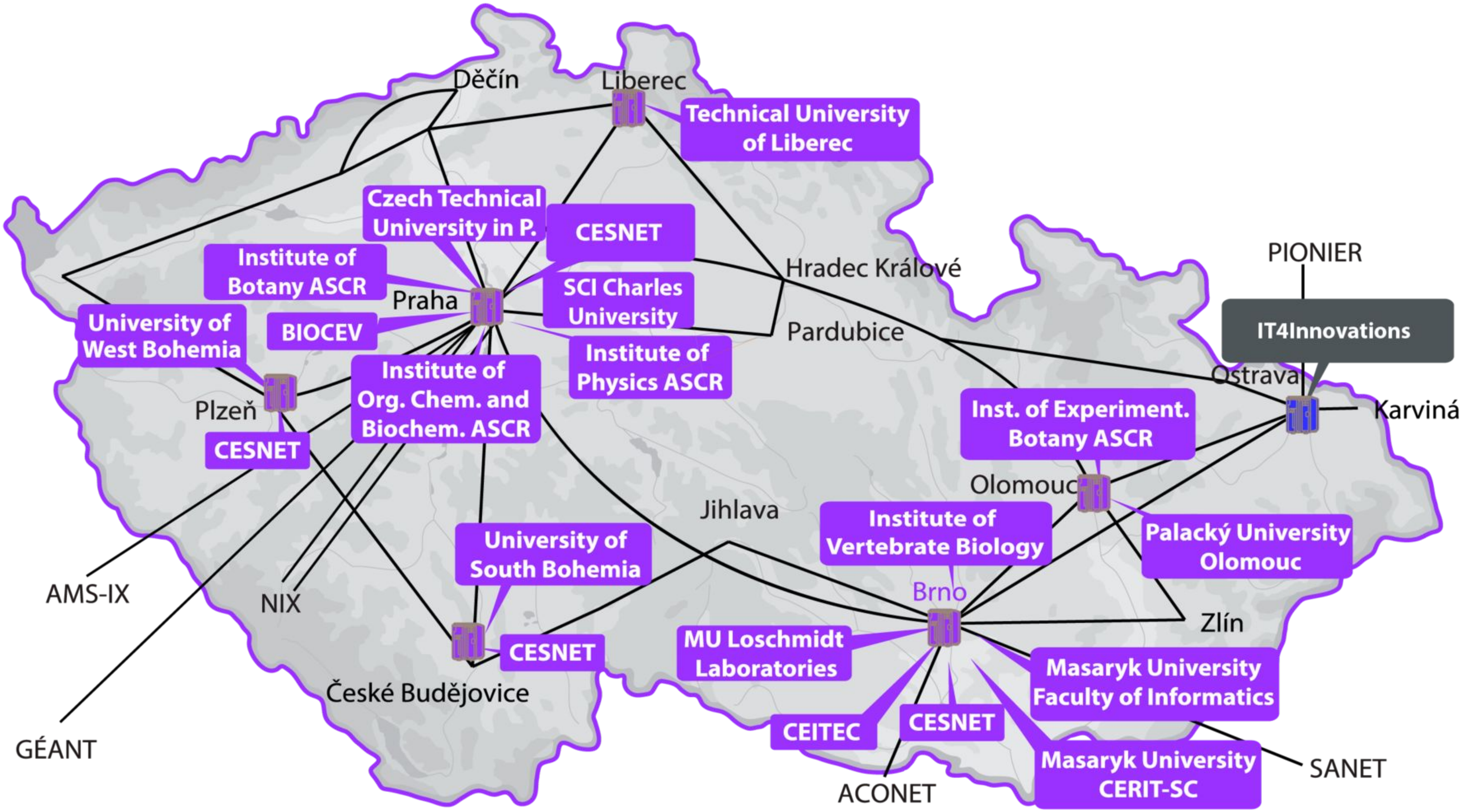
slides at: **ces.net/telc**

Outline

- Computing in Czechia with MetaCentrum
- Galaxy’s purpose & capabilities
- Galaxy in Czechia

Computing in Czechia

with MetaCentrum


Děčín
Liberec
Technical University of Liberec
Czech Technical University in P.
CESNET
Institute of Botany ASCR
Praha
SCI Charles University
Hradec Králové
University of West Bohemia
BIOCEV
Pardubice
Institute of Physics ASCR
Institute of Org. Chem. and Biochem. ASCR
PIONIER
IT4Innovations
Ostrava
Karviná
Plzeň
CESNET
Inst. of Experiment. Botany ASCR
Olomouc
Jihlava
Institute of Vertebrate Biology
Palacký University Olomouc
University of South Bohemia
AMS-IX
NIX
Brno
Zlín
CESNET
MU Loschmidt Laboratories
Masaryk University Faculty of Informatics
České Budějovice
CEITEC
CESNET
GÉANT
Masaryk University CERIT-SC
SANET
ACONET

cesnet metacentrum

- National Grid Infrastructure (NGI)
- Provider of computational resources and data storage
- Free (as in beer)
  - For employees and students in Czech Academia
  - But also for industry users (non-profit public research, upon individual request)

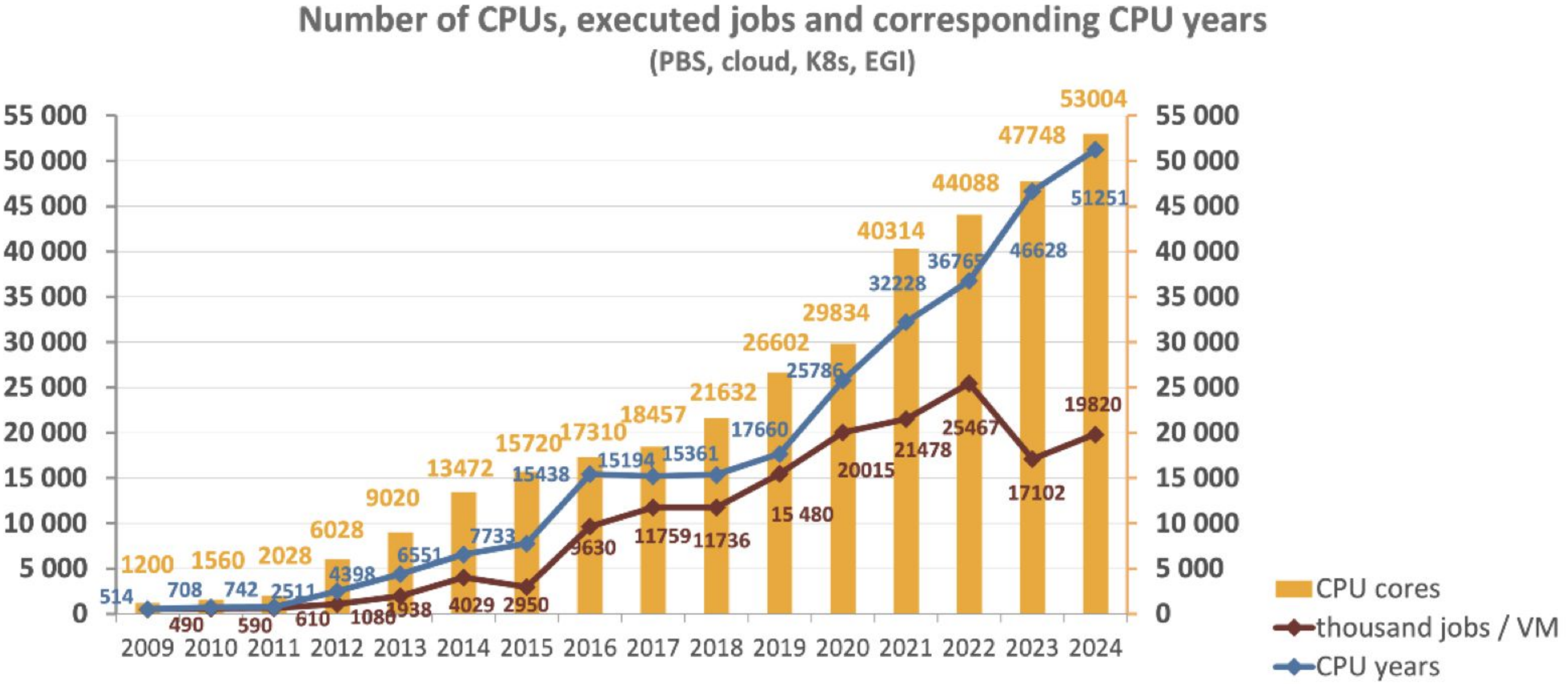
cesnet
metacentrum

Compute resources are pooled together by all partners (CESNET, universities, Czech Academy of Sciences) and…

- ... are centrally managed
- ... are shared among all users
- ... have privileged access for cluster owners
- ... are replaceable during an outage
- … include support for federated AAI
- ... are dedicated to grid HPC/HTC, containerised computing, cloud computing, data storage capacities

cesnet
metacentrum

stats



Number of CPUs, executed jobs and corresponding CPU years
(PBS, cloud, K8s, EGI)

MetaCentrum targets

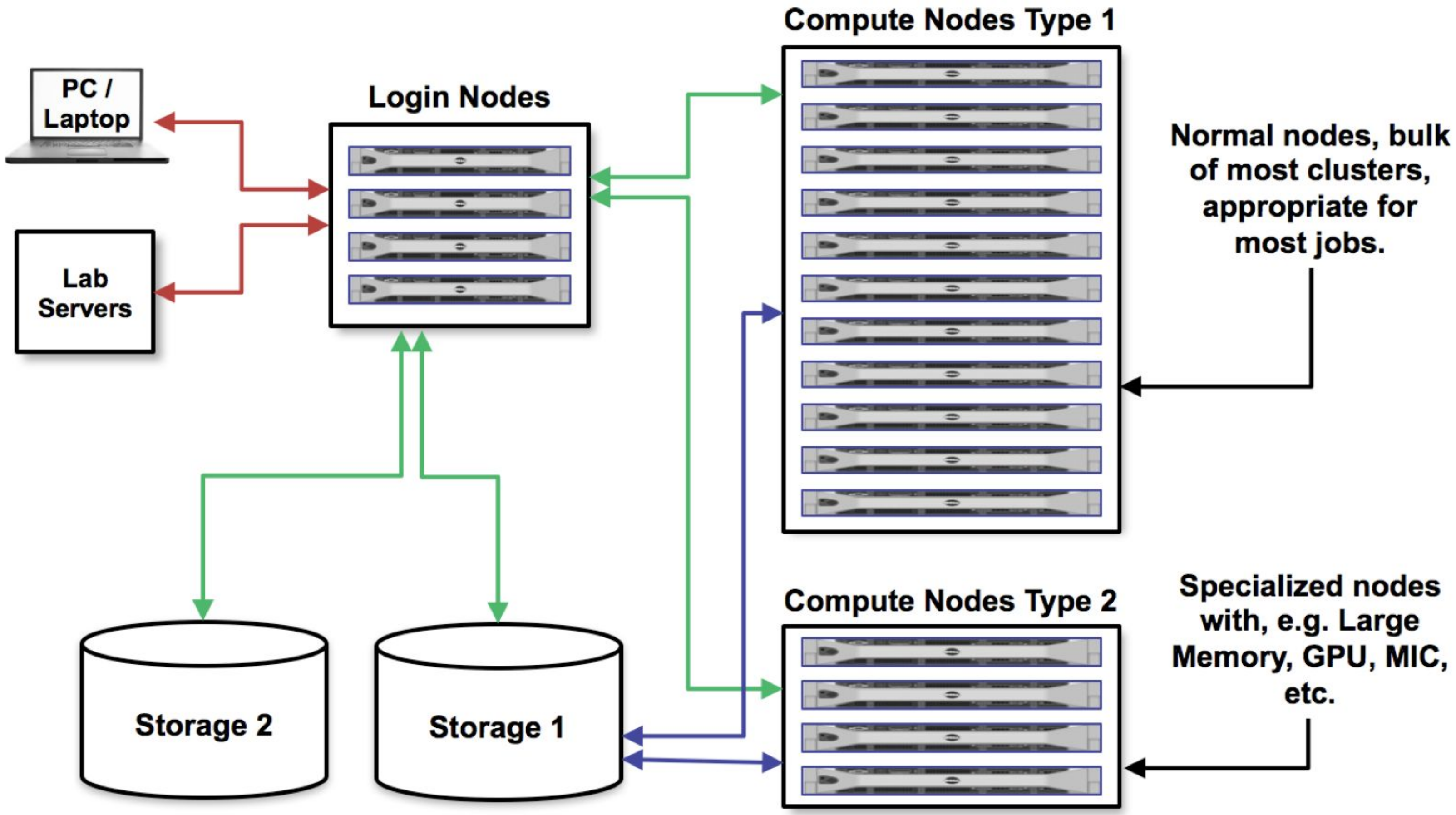
- individual users (access to resources)
- projects (cooperation, sharing data in a group)
- organisations (incorporate their resources under central management)

MetaCentrum offers

- immediate access to HW resources
- various application tools (commercial, free, open source)
- CPU/GPU resources, GUI applications and access, cloud services

Great starting point is: https://docs.metacentrum.cz/

MetaCentrum batch job work setup



Normal nodes, bulk of most clusters, appropriate for most jobs.

Specialized nodes with, e.g. Large Memory, GPU, MIC, etc.

MetaCentrum batch job work setup

qsub Manual Page

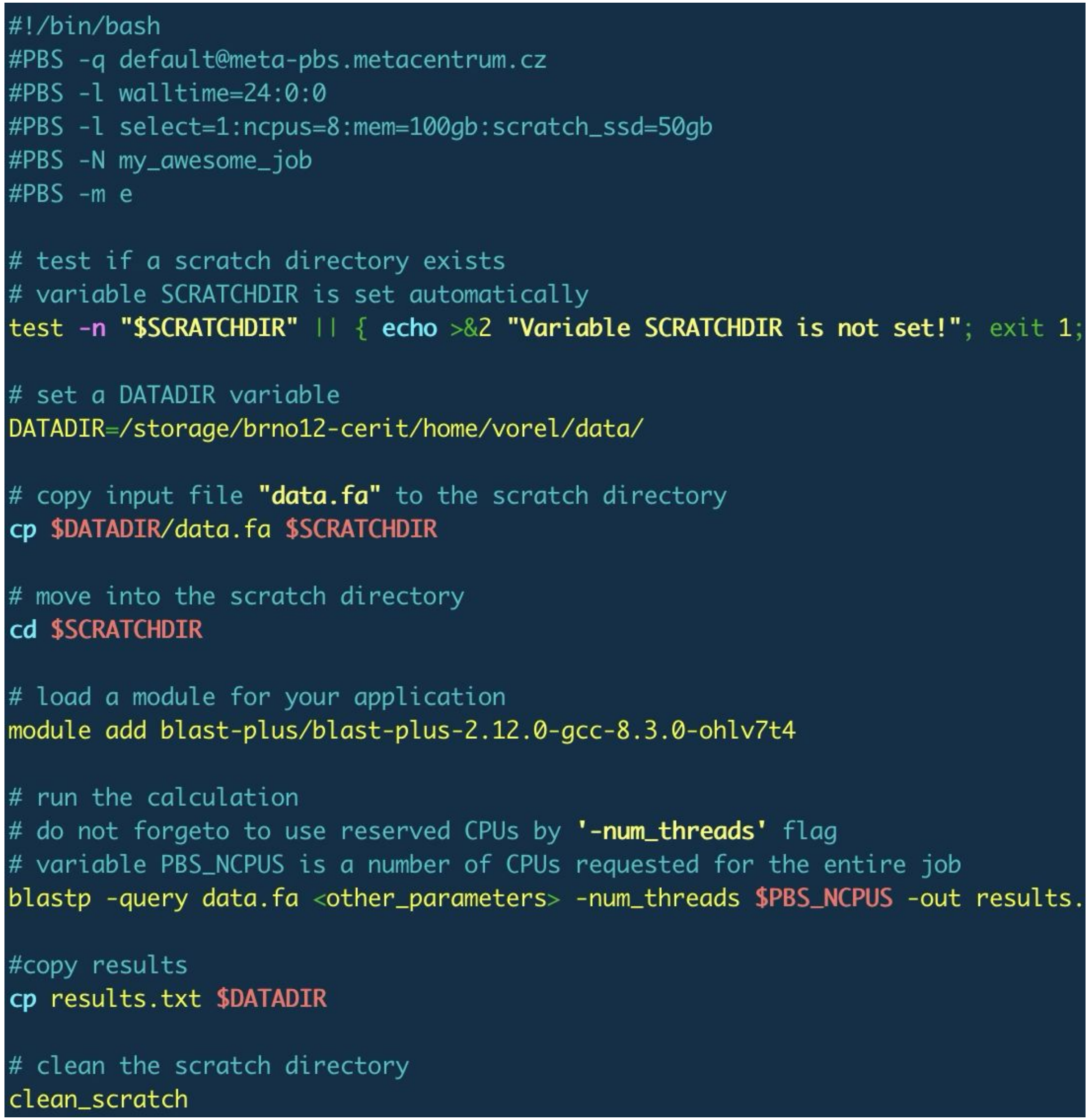
NAME

qsub - submit job

DESCRIPTION

To create a job is to submit an executable script to a batch server. The batch server will be the default server unless the -q option is specified. Typically, the script is a shell script which will be executed by a command shell such as sh or csh.

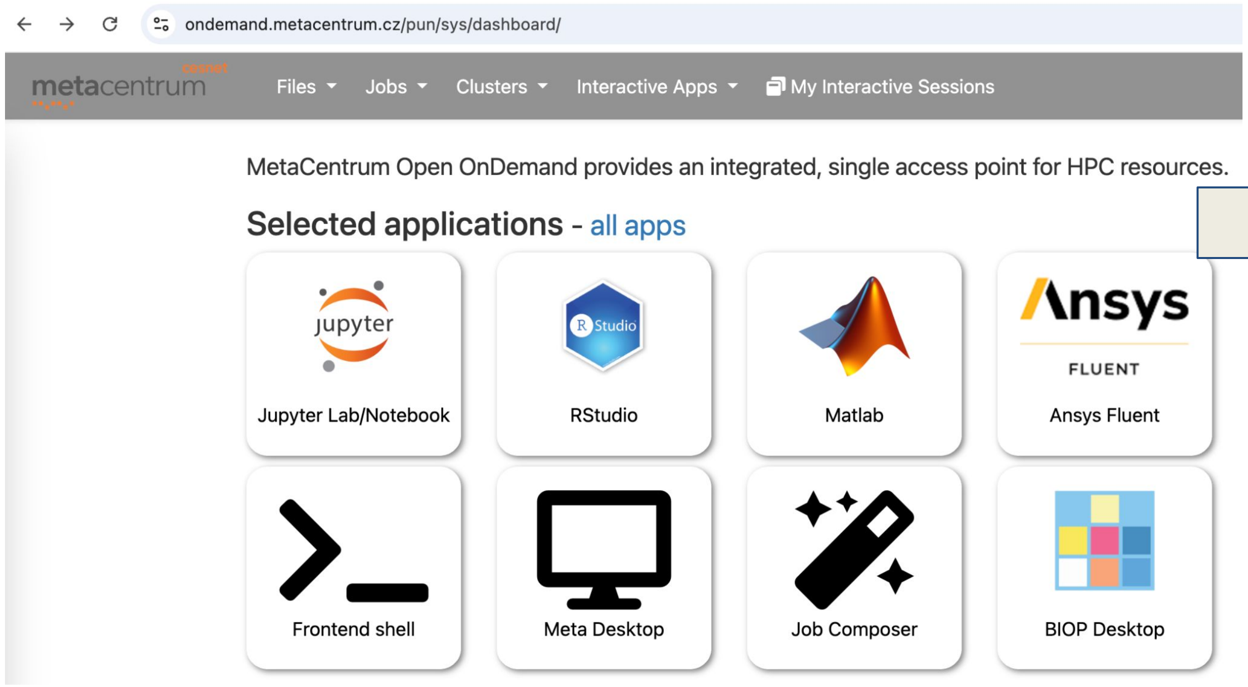
Options on the qsub command allow the specification of attributes which affect the behavior of the job.



```
#!/bin/bash
#PBS -q default@meta-pbs.metacentrum.cz
#PBS -l walltime=24:0:0
#PBS -l select=1:ncpus=8:mem=100gb:scratch_ssd=50gb
#PBS -N my_awesome_job
#PBS -m e

# test if a scratch directory exists
# variable SCRATCHDIR is set automatically
test -n "$SCRATCHDIR" || { echo >&2 "Variable SCRATCHDIR is not set!"; exit 1;

# set a DATADIR variable
DATADIR=/storage/brno12-cerit/home/vorel/data/

# copy input file "data.fa" to the scratch directory
cp $DATADIR/data.fa $SCRATCHDIR

# move into the scratch directory
cd $SCRATCHDIR

# load a module for your application
module add blast-plus/blast-plus-2.12.0-gcc-8.3.0-ohlv7t4

# run the calculation
# do not forgeto to use reserved CPUs by '-num_threads' flag
# variable PBS_NCPUS is a number of CPUs requested for the entire job
blastp -query data.fa <other_parameters> -num_threads $PBS_NCPUS -out results.

#copy results
cp results.txt $DATADIR

# clean the scratch directory
clean_scratch
```

OnDemand user experience

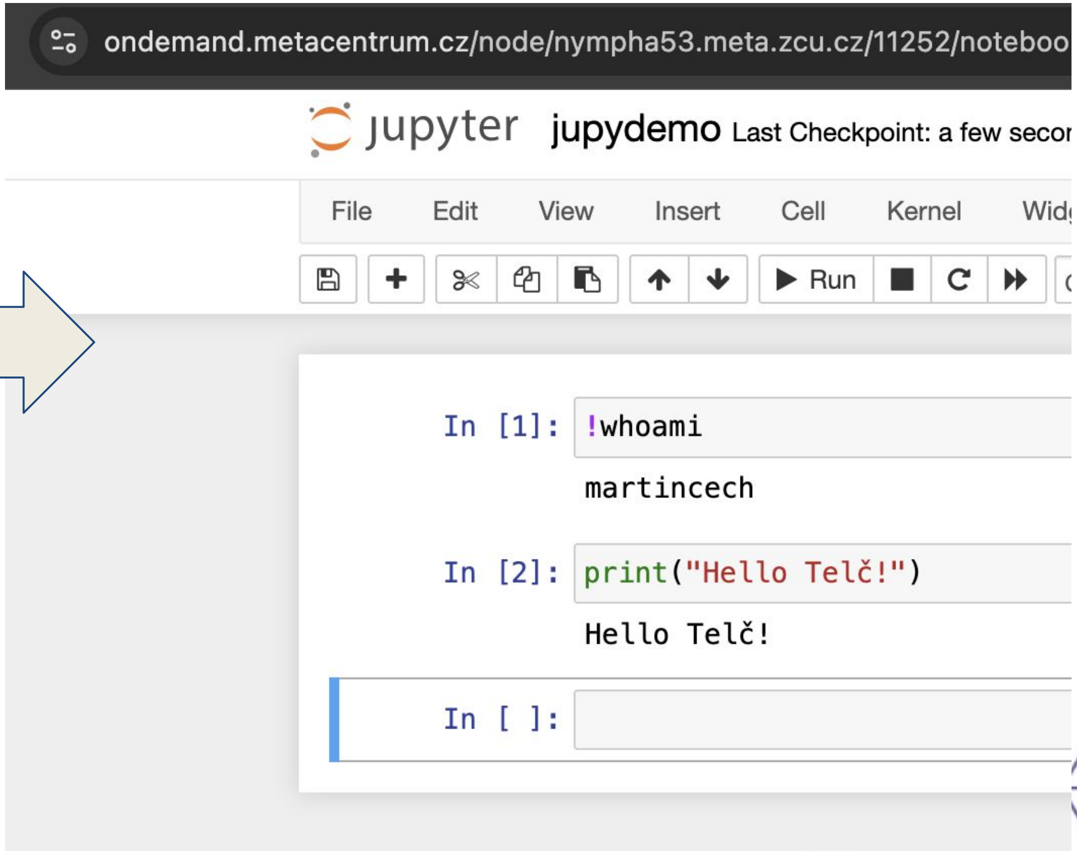
- I need 64 cores, 1 TB RAM, 4 hi-end GPU for my Matlab/Jupyter/RStudio/other calculation
- Wow, this is expensive…
- But I need it 6 hours per month only
- Go to https://ondemand.metacentrum.cz/
- Fill a formwith what you need for how long
- Profit (I mean *research*)



ondemand.metacentrum.cz/pun/sys/dashboard/

MetaCentrum Open OnDemand provides an integrated, single access point for HPC resources.

Selected applications - all apps

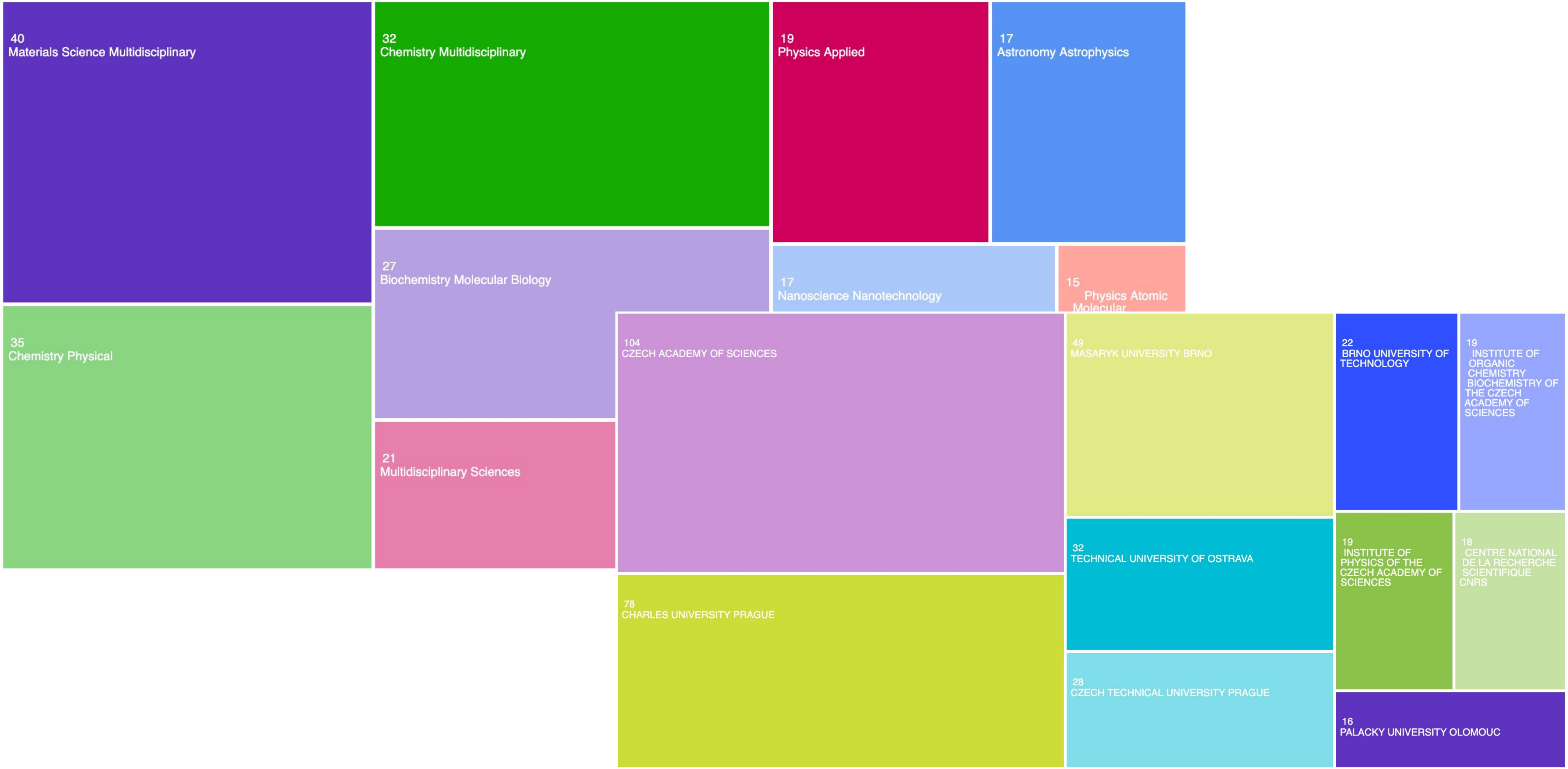


ondemand.metacentrum.cz/node/nympha53.meta.zcu.cz/11252/noteboo

jupydemo Last Checkpoint: a few secor

```
In [1]: !whoami
        martincech

In [2]: print("Hello Telč!")
        Hello Telč!

In [ ]:
```

MetaCentrum supported pubs in 2024



IT4Innovations – supercomputers at your service

- Three times a year an open access grant competition
- But also Fast Track Access for smaller projects (4 months)

Computational resources distributed every 4 months (node hours):

- Barbora CPU: 500,000 n/h, GPU: 30,000 n/h, FAT: 3,400 n/h
- Karolina CPU: 1,012,000 n/h Karolina GPU: 83,000 n/h, FAT: 1,200 n/h

details: https://www.it4i.cz/en/for-users/open-access-competition

Galaxy

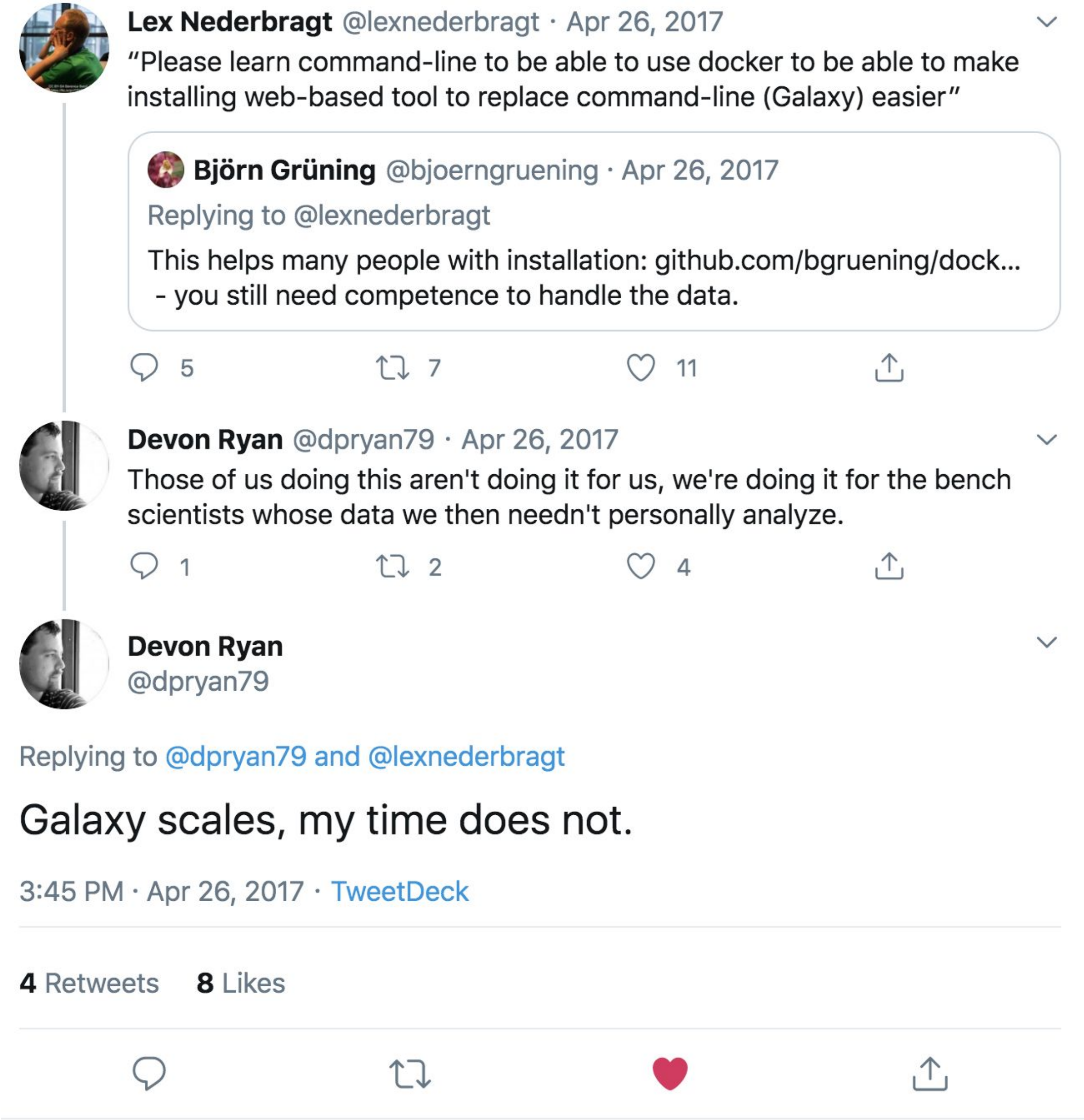
Galaxy’s purpose & capabilities

simplified interface to software and compute infrastructure

Galaxy - key purpose

- Analyze data using thousands of tools without installation and maintenance
- Learn or publish methods and techniques
- Create and share workflows, using an expressive graphical interface
- Distribute one’s own tools, locally or globally
- Expose compute and storage infrastructure for easy access by users
- Born from biology but domain-agnostic
  - Natural Language Processing, Sociology, Ecology, Astronomy, Weather Modelling

Capability - support scale



Lex Nederbragt @lexnederbragt · Apr 26, 2017

“Please learn command-line to be able to use docker to be able to make installing web-based tool to replace command-line (Galaxy) easier”

> **Björn Grüning** @bjoerngruening · Apr 26, 2017
>
> Replying to @lexnederbragt
>
> This helps many people with installation: github.com/bgruening/dock...
> - you still need competence to handle the data.

Devon Ryan @dpryan79 · Apr 26, 2017

Those of us doing this aren't doing it for us, we're doing it for the bench scientists whose data we then needn't personally analyze.

Devon Ryan
@dpryan79

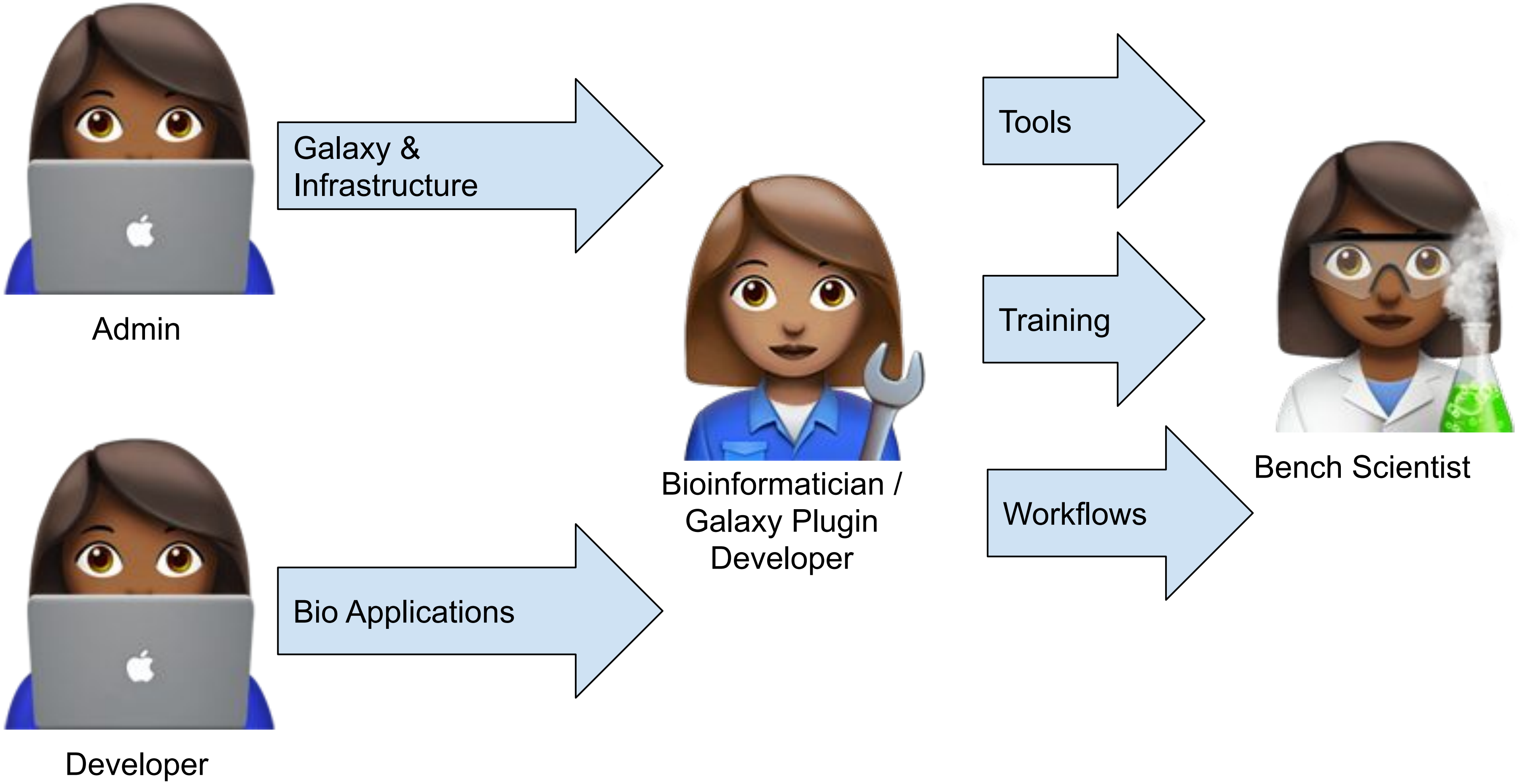
Replying to @dpryan79 and @lexnederbragt

Galaxy scales, my time does not.

3:45 PM · Apr 26, 2017 · TweetDeck

4 Retweets **8** Likes

Capability - support specializations



Galaxy & Infrastructure

Admin

Bio Applications

Developer

Bioinformatician / Galaxy Plugin Developer

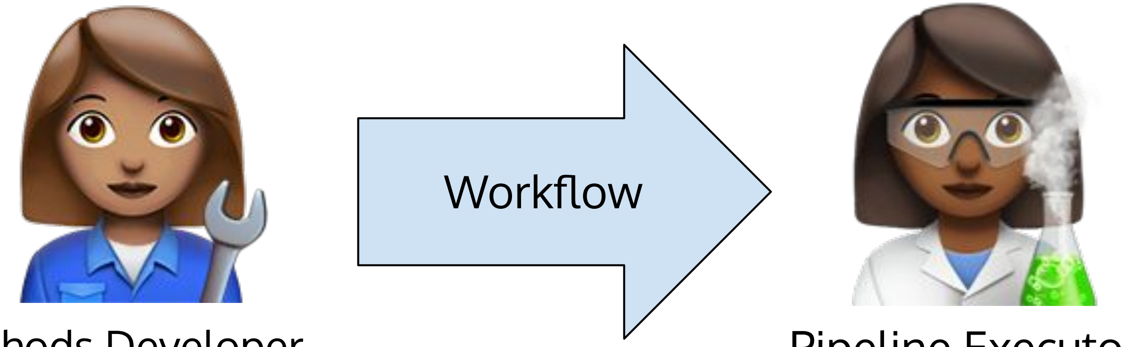
Tools

Training

Workflows

Bench Scientist

Capability - hide underlying complexity



Methods Developer

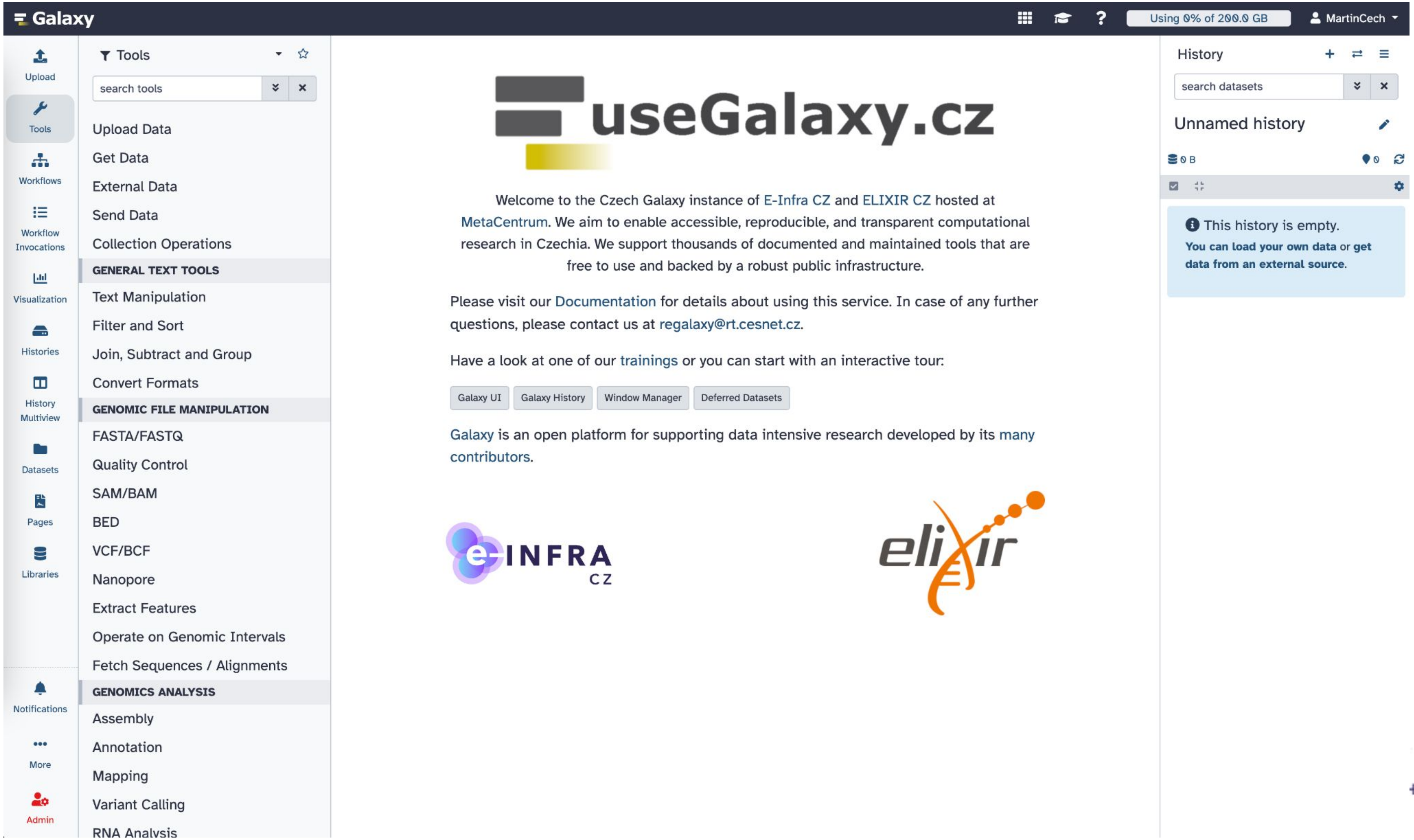
Workflow

Pipeline Executor

Galaxy UI enables method developers without knowledge of scripting, etc...

No need to understand command-lines, etc.. Embedded visualizations, etc..

How does it look like



Welcome to the Czech Galaxy instance of E-Infra CZ and ELIXIR CZ hosted at MetaCentrum. We aim to enable accessible, reproducible, and transparent computational research in Czechia. We support thousands of documented and maintained tools that are free to use and backed by a robust public infrastructure.

Please visit our Documentation for details about using this service. In case of any further questions, please contact us at regalaxy@rt.cesnet.cz.

Have a look at one of our trainings or you can start with an interactive tour:

Galaxy is an open platform for supporting data intensive research developed by its many contributors.

This history is empty.

You can load your own data or get data from an external source.

A little Galaxy demo

if wifi holds

Galaxy Show case: VGP

VGP
VERTEBRATE
GENOMES
PROJECT

A PROJECT OF THE G10K CONSORTIUM

https://galaxyproject.org/projects/vgp

VGP plan:

- Publish genome assemblies of 12 species every week and accelerate
- Until all the 75,923 extant vertebrate species genomes are known
- which is expected within a decade
- =~ petabytes of assembly data (ignoring raw data)
- *Reference point: Assembly of human genome took 10+ years and $3 billion dollars*

Galaxy Show case: VGP

VGP VERTEBRATE GENOMES PROJECT

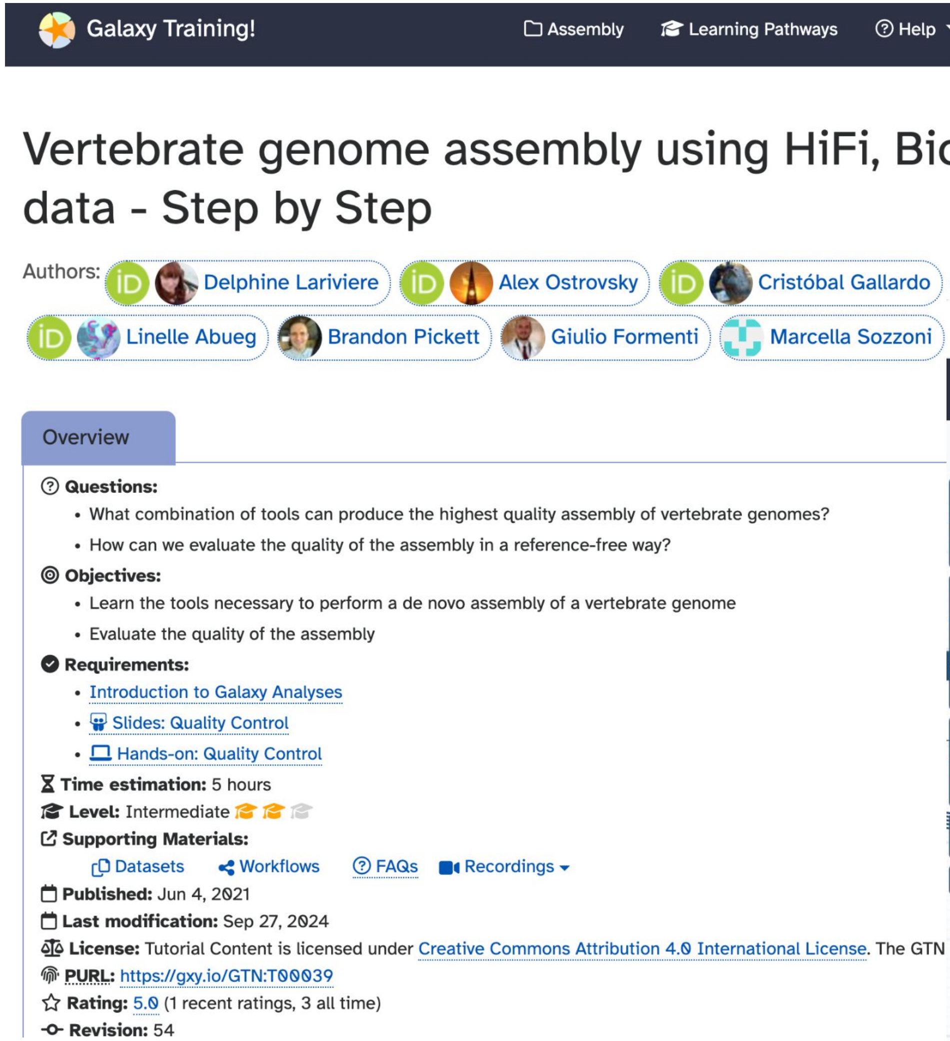
A PROJECT OF THE G10K CONSORTIUM

https://galaxyproject.org/projects/vgp

Galaxy provides:

- Integration of the Genome Ark on public Galaxy servers.
- A Galaxy platform with toolkits specifically tailored for Genome assembly
- Workflows available using the most up-to-date VGP pipelines.
- A list of publicly-available histories for each assembly completed on Galaxy as they are generated.
- Extensive training materials used to spread the load and accelerate

Using Galaxy helps make the assemblies reproducible and enables scaling



Galaxy Training! | Assembly | Learning Pathways | Help | Settings | Search Tutorials

Vertebrate genome assembly using HiFi, Bionano and Hi-C data - Step by Step

Authors: Delphine Lariviere, Alex Ostrovsky, Cristóbal Gallardo, Anna Syme, Linelle Abueg, Brandon Pickett, Giulio Formenti, Marcella Sozzoni

Overview

Questions:

- What combination of tools can produce the highest quality assembly of vertebrate genomes?
- How can we evaluate the quality of the assembly in a reference-free way?

Objectives:

- Learn the tools necessary to perform a de novo assembly of a vertebrate genome
- Evaluate the quality of the assembly

Requirements:

- Introduction to Galaxy Analyses
- Slides: Quality Control
- Hands-on: Quality Control

Time estimation: 5 hours

Level: Intermediate

Supporting Materials:

Datasets | Workflows | FAQs | Recordings

Published: Jun 4, 2021

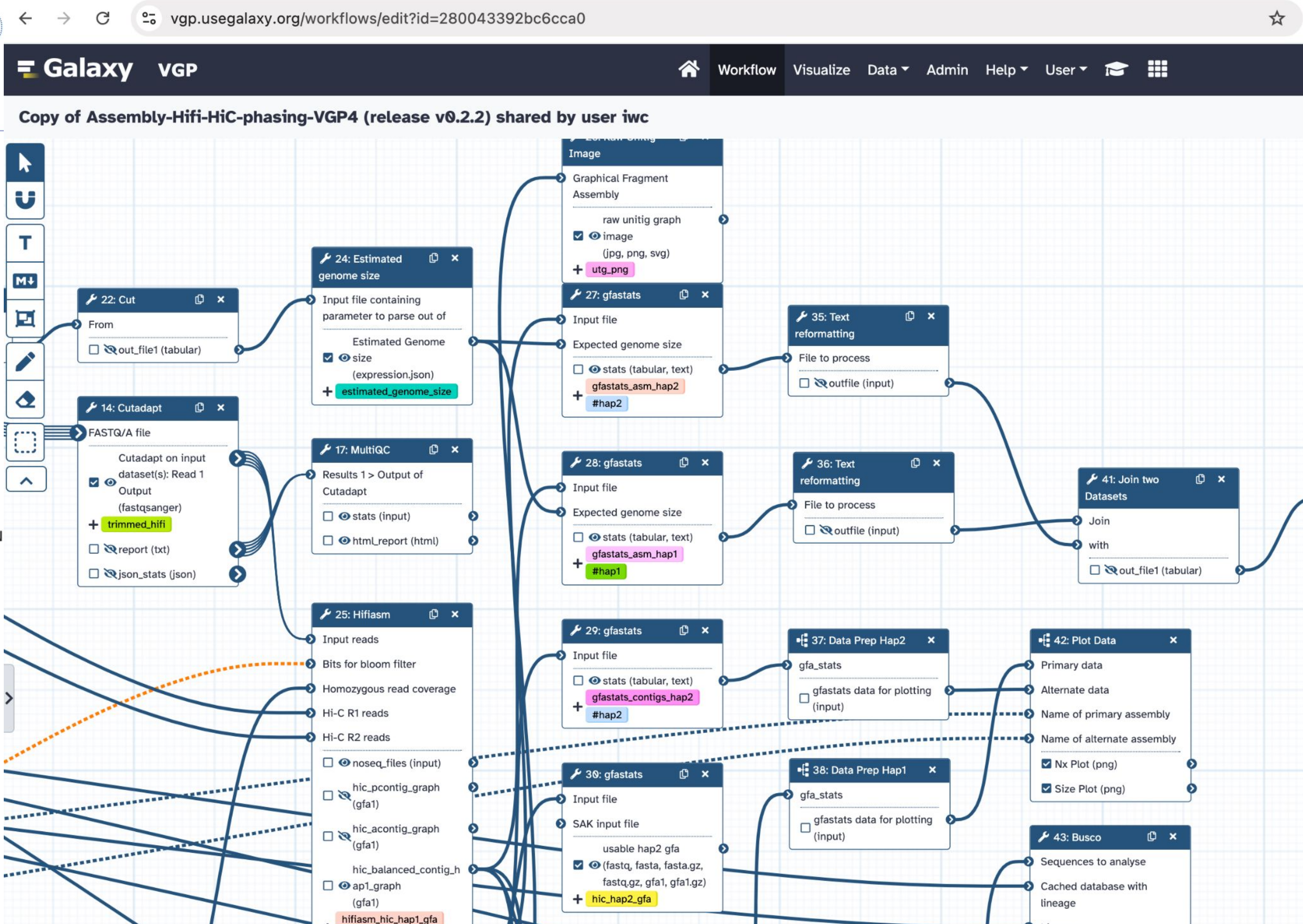
Last modification: Sep 27, 2024

License: Tutorial Content is licensed under Creative Commons Attribution 4.0 International License. The GTN

PURL: https://gxy.io/GTN:T00039

Rating: 5.0 (1 recent ratings, 3 all time)

Revision: 54



vgp.usegalaxy.org/workflows/edit?id=280043392bc6cca0

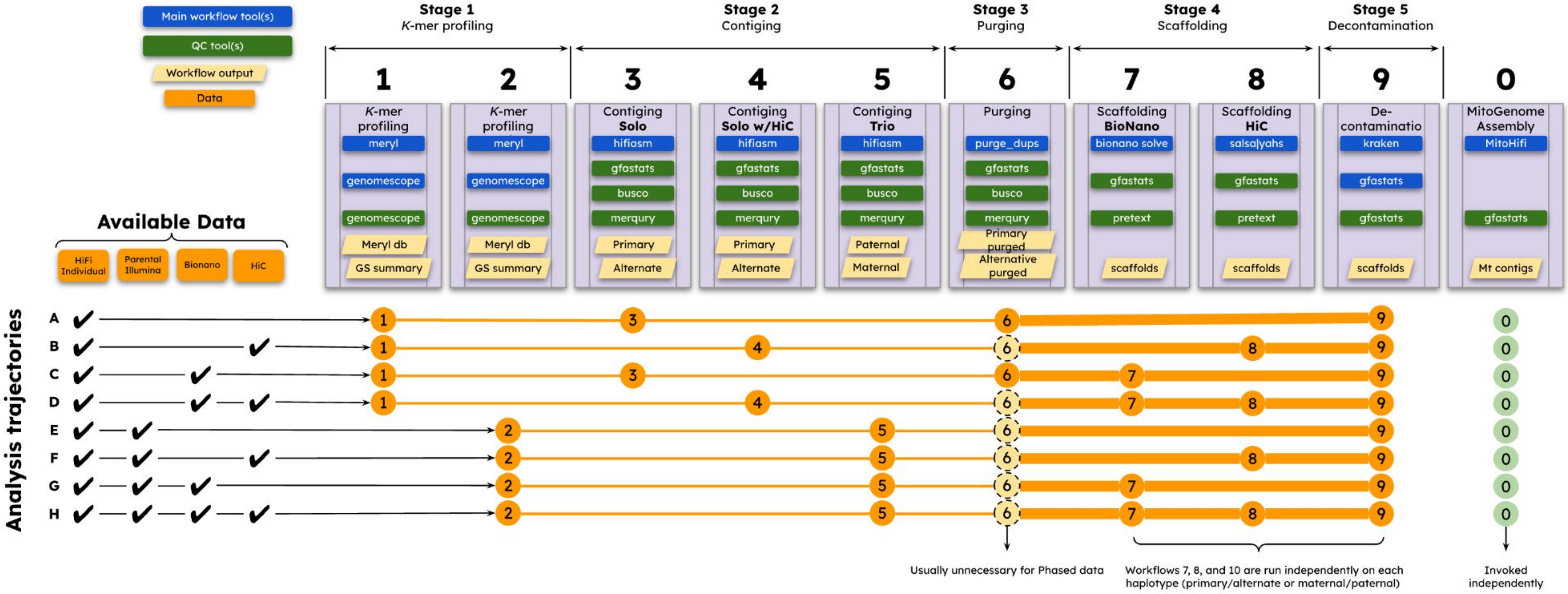
Copy of Assembly-Hifi-HiC-phasing-VGP4 (release v0.2.2) shared by user iwc

bit.ly/vgp-training

Workflows Overview

Eight analysis trajectories are possible depending on the combination of input data. Decision on invocation of workflow 6 is based on the analysis of QC output of workflows 3, 4, or 5 (see below). Thicker lines connecting workflows 7, 8, and 9 represent the fact that these workflows are invoked separately for each phased assembly (once for maternal [or hap1] and once for paternal [or hap2]). **Solo** = data is only available for the sample whose genome is being assembled. In this case, you can make either a pseudohaplotype assembly, or a HiC-phased assembly if you have HiC data from the same individual.

Trio = parental information is available in the form of Illumina reads from each parent of the F1 being assembled.



Main workflow tool(s)

QC tool(s)

Workflow output

Data

Stage 1 *K*-mer profiling

Stage 2 Contiging

Stage 3 Purging

Stage 4 Scaffolding

Stage 5 Decontamination

1 *K*-mer profiling: meryl, genomescope, genomescope, Meryl db, GS summary

2 *K*-mer profiling: meryl, genomescope, genomescope, Meryl db, GS summary

3 Contiging **Solo**: hifiasm, gfastats, busco, merqury, Primary, Alternate

4 Contiging **Solo w/HiC**: hifiasm, gfastats, busco, merqury, Primary, Alternate

5 Contiging **Trio**: hifiasm, gfastats, busco, merqury, Paternal, Maternal

6 Purging: purge_dups, gfastats, busco, merqury, Primary purged, Alternative purged

7 Scaffolding **BioNano**: bionano solve, gfastats, pretext, scaffolds

8 Scaffolding **HiC**: salsa|yahs, gfastats, pretext, scaffolds

9 De-contaminatio: kraken, gfastats, gfastats, scaffolds

0 MitoGenome Assembly: MitoHifi, gfastats, Mt contigs

Available Data

Analysis trajectories

	HiFi Individual	Parental Illumina	Bionano	HiC	1	2	3	4	5	6	7	8	9	0
A	✓				1		3			6			9	0
B	✓			✓	1			4		6		8	9	0
C	✓		✓		1		3			6	7		9	0
D	✓		✓	✓	1			4		6	7	8	9	0
E	✓	✓				2			5	6			9	0
F	✓	✓		✓		2			5	6		8	9	0
G	✓	✓	✓			2			5	6	7		9	0
H	✓	✓	✓	✓		2			5	6	7	8	9	0

Usually unnecessary for Phased data

Workflows 7, 8, and 10 are run independently on each haplotype (primary/alternate or maternal/paternal)

Invoked independently

A peak into VGP workflows

if wifi holds

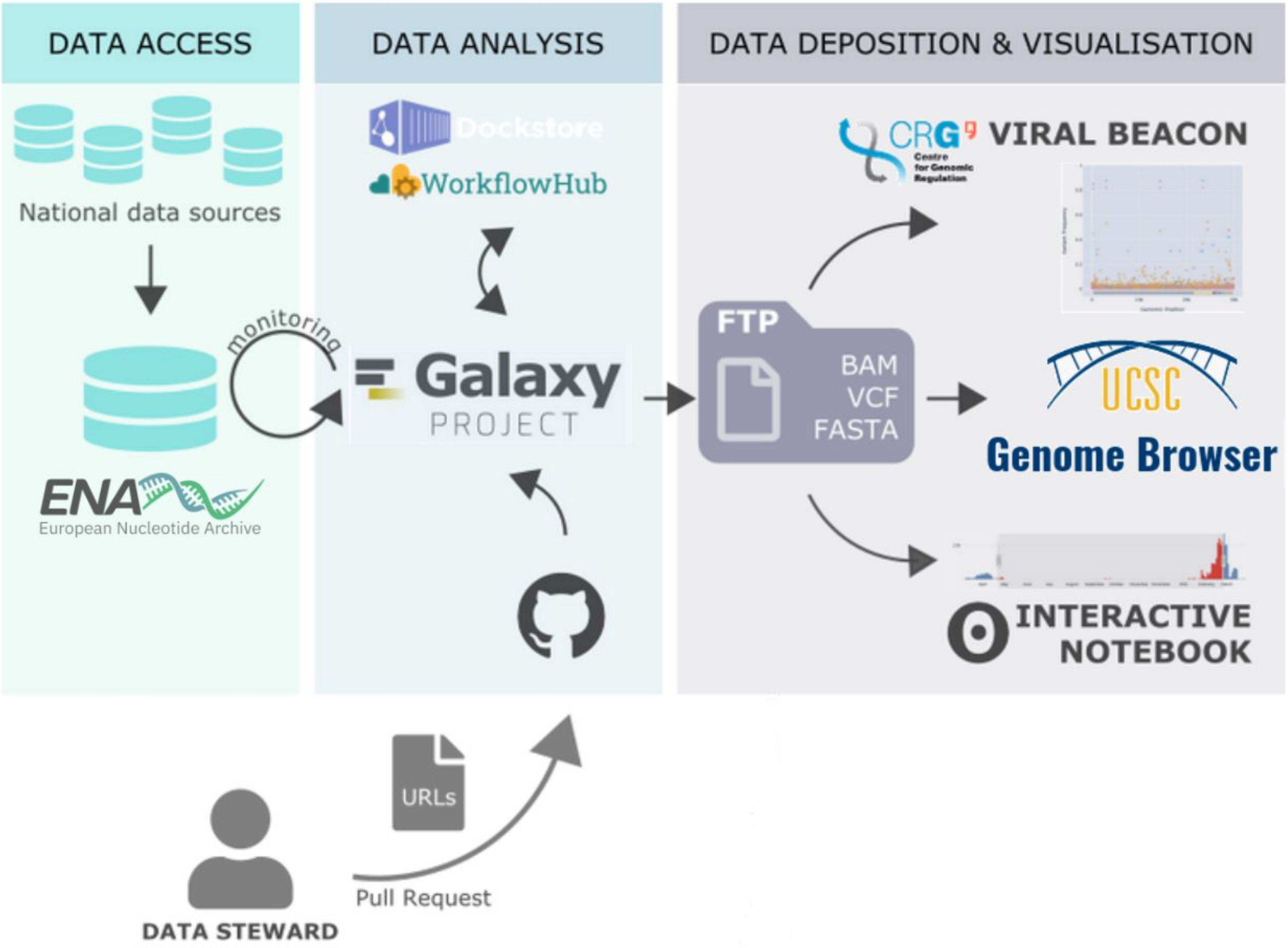
Galaxy Show case: SARS-CoV-2

galaxyproject/
SARS-CoV-2

Four Goals:

- Continuously analysis of within-host sequence variants in high quality public read-level datasets
- Maintenance of curated workflows for the analysis of SARS-CoV-2 sequence data
- Development of continuously updated analysis page and dashboard summarizing latest insights from the variant.
- Providing access to all results in raw and aggregated form for immediate use.

Galaxy Show case: SARS-CoV-2



DATA ACCESS

DATA ANALYSIS

DATA DEPOSITION & VISUALISATION

National data sources

monitoring

ENA

European Nucleotide Archive

Dockstore

WorkflowHub

Galaxy PROJECT

FTP

BAM

VCF

FASTA

CRG Centre for Genomic Regulation

VIRAL BEACON

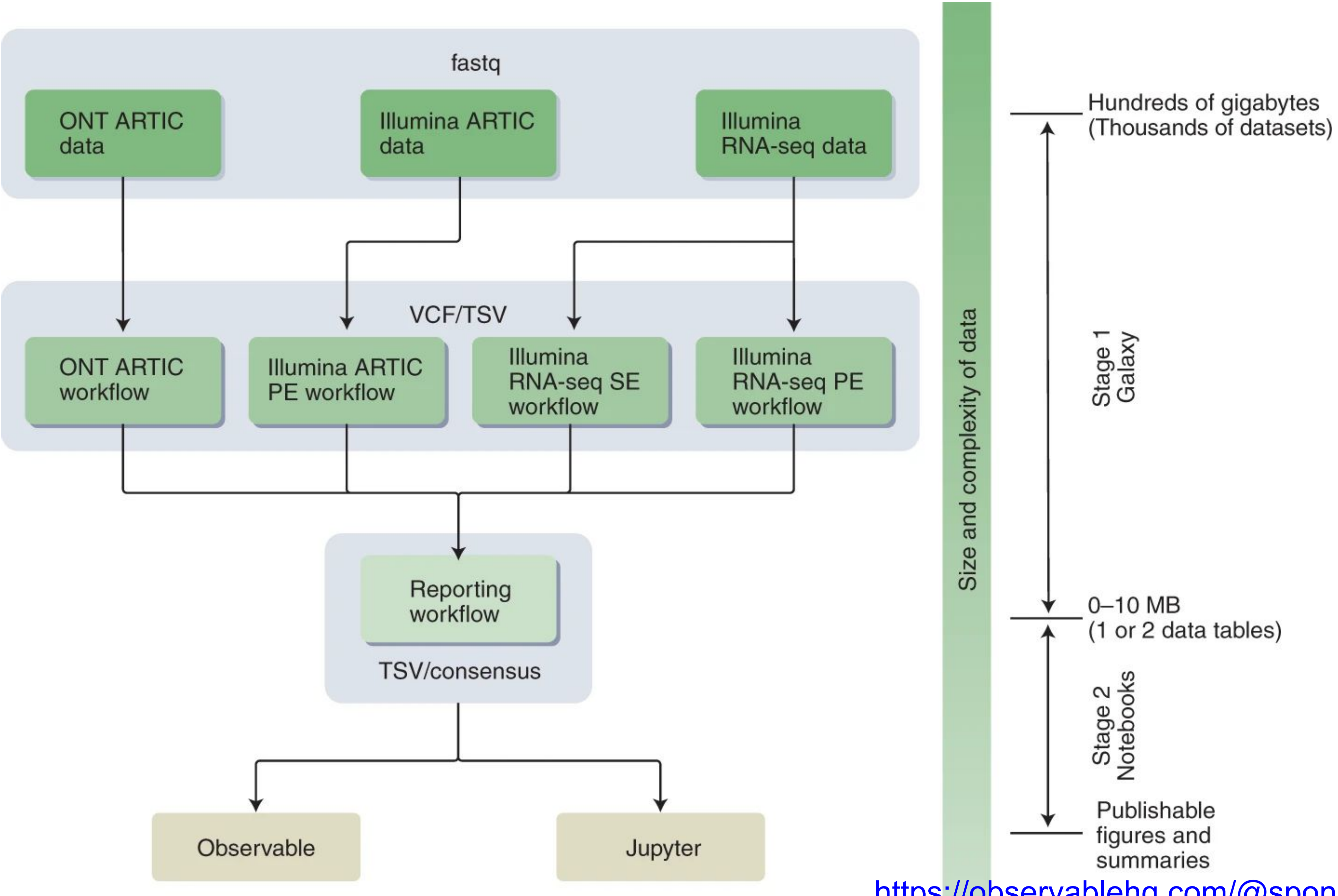
UCSC Genome Browser

INTERACTIVE NOTEBOOK

URLs

Pull Request

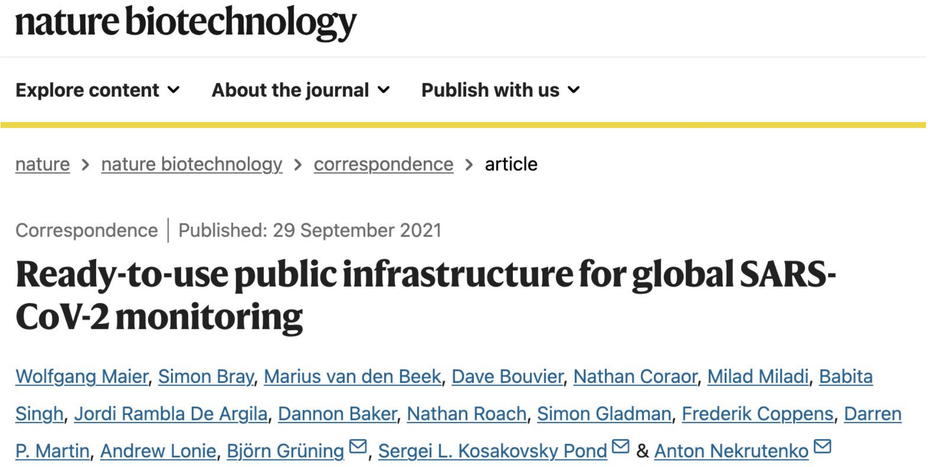
DATA STEWARD


fastq
ONT ARTIC data
Illumina ARTIC data
Illumina RNA-seq data
VCF/TSV
ONT ARTIC workflow
Illumina ARTIC PE workflow
Illumina RNA-seq SE workflow
Illumina RNA-seq PE workflow
Reporting workflow
TSV/consensus
Observable
Jupyter
Size and complexity of data
Hundreds of gigabytes (Thousands of datasets)
Stage 1 Galaxy
0–10 MB (1 or 2 data tables)
Stage 2 Notebooks
Publishable figures and summaries

https://observablehq.com/@spond/intrahost-dashboard

Galaxy Show case: SARS-CoV-2

galaxyproject/
SARS-CoV-2



nature biotechnology

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature biotechnology › correspondence › article

Correspondence | Published: 29 September 2021

Ready-to-use public infrastructure for global SARS-CoV-2 monitoring

Wolfgang Maier, Simon Bray, Marius van den Beek, Dave Bouvier, Nathan Coraor, Milad Miladi, Babita Singh, Jordi Rambla De Argila, Dannon Baker, Nathan Roach, Simon Gladman, Frederik Coppens, Darren P. Martin, Andrew Lonie, Björn Grüning ✉, Sergei L. Kosakovsky Pond ✉ & Anton Nekrutenko ✉

Everything implemented in unprecedented levels of open science

Repository with bot processing requests:
https://github.com/usegalaxy-eu/sars-cov-2-processing-requests

This project eventually analyzed nearly 500,000 public SARS-CoV-2 sequencing datasets until 2022.

Galaxy in Czechia

everybody can run Galaxy

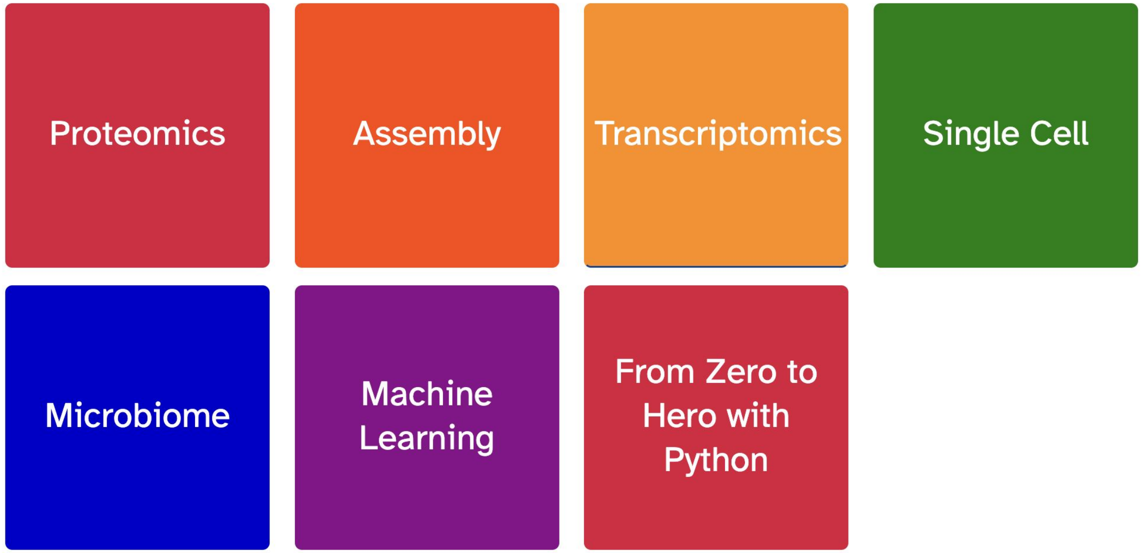
Galaxies we run

- RepeatExplorer Galaxy
  - specialized instance for graph-based clustering and characterization of repetitive sequences in NGS
  - run Institute of Plant Molecular Biology in Ceske Budejovice
  - one of **ELIXIR CZ** services
- UMSA Galaxy
  - Untargeted Mass Spectrometry Analysis
  - operated on behalf of **MUNI RECETOX**
  - publish tools, workflows, and other research results
  - provide pipeline as a service for collaborators
- UseGalaxy.cz
  - all purpose instance (and we'll support your tools)
  - aims to make the resources of National Grid Infrastructure more accessible

Galaxy Training Academy 2025

12-16 May 2025 (online)

- 5-day Global Online and Asynchronous learning event
- choose your own path from 400 trainings
- slides, hands-on, videos, screencasts
- daily sync times, continuous support on slack
- years of experience running these, they're really good
- everything lives on https://training.galaxyproject.org



Proteomics

Assembly

Transcriptomics

Single Cell

Microbiome

Machine Learning

From Zero to Hero with Python

Coming up

Events:

22 April 2025	**Galaxy Imaging Hackathon 2025** (Freiburg, GER + online)
12-16 May 2025	**Galaxy Training Academy 2025** (online)
27-29 May 2025	**12th Repeat Explorer Workshop** (Ceske Budejovice, CZ)
23-27 June 2025	**Galaxy and Bioconductor Community Conference** (NY, USA)
1-3 October 2025	**European Galaxy Days** (Freiburg, GER)

Projects:

- National Repository Platform for Research Data 2024-2028
  - imagine it as a national Zenodo instance per research area
  - Galaxy will provide integration with the repositories
- EuroScienceGateway
  - Pooling of research compute resources between EU countries with Galaxy Project.

Take home message

- after PhD every project you work on will be bigger
- your time does not scale
- Excel, Matlab, and homegrown approaches will slow you down

- MetaCentrum gives you **the iron**
- Galaxy gives you **the utility**
- Galaxy Training Network gives you **the mastery**

- https://docs.metacentrum.cz
- https://usegalaxy.cz

p.s. If you take only one thing from here be it training.galaxyproject.org

Credits and thank yous

- Delphine Lariviere – Galaxy, VGP
- Wolfgang Maier – Freiburg, SARS-CoV-2
- Nate Coraor – Galaxy
- John Chilton – Galaxy
- Jirka Vorel – CESNET
- Aleš Křenek – CESNET/MUNI

& The Galaxy Community

Thank You

Questions, please?

cesnet

elixir CZECH REPUBLIC

e-INFRA CZ

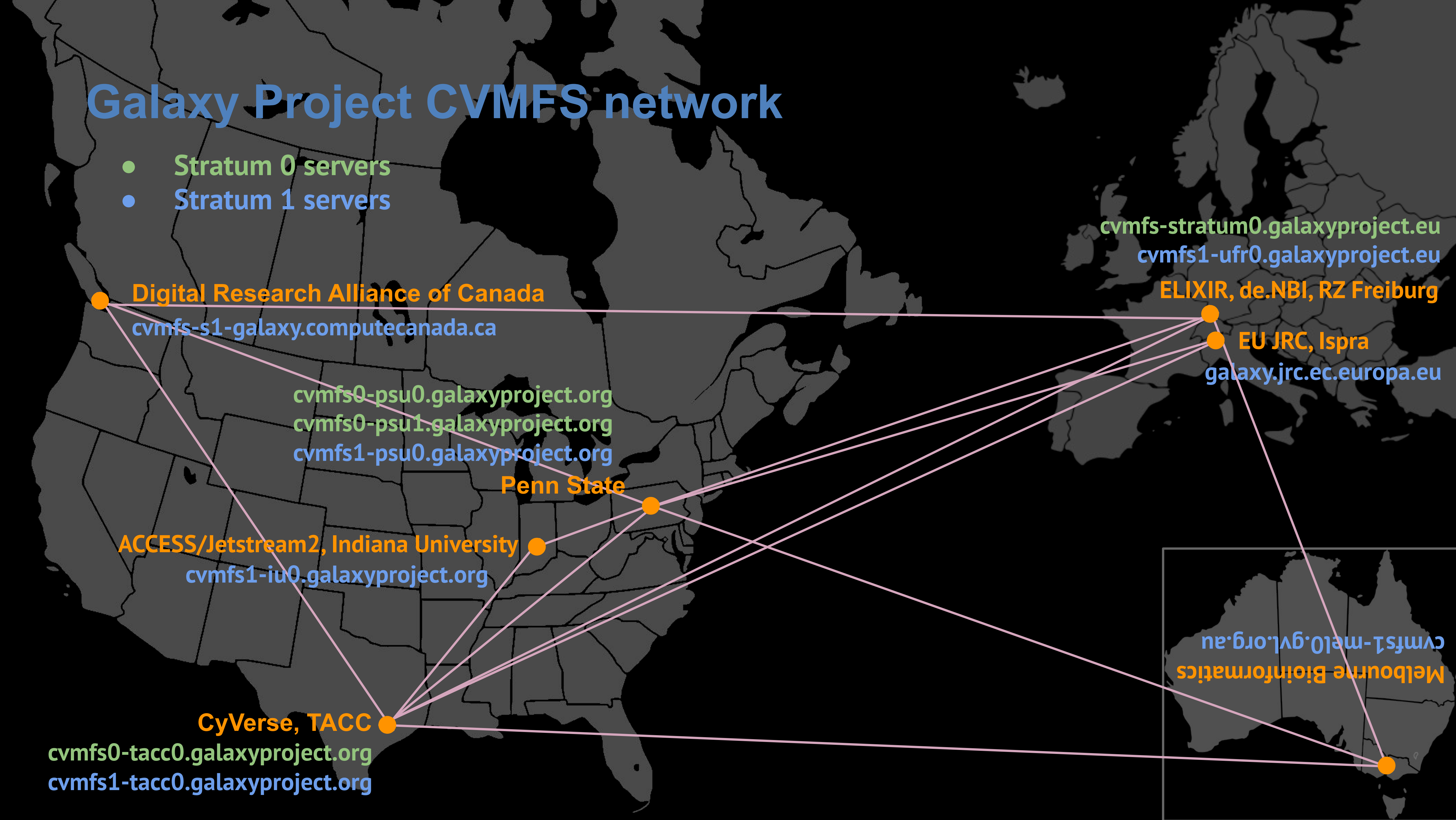
Martin Čech – Telč – 28.3.25

@martenson

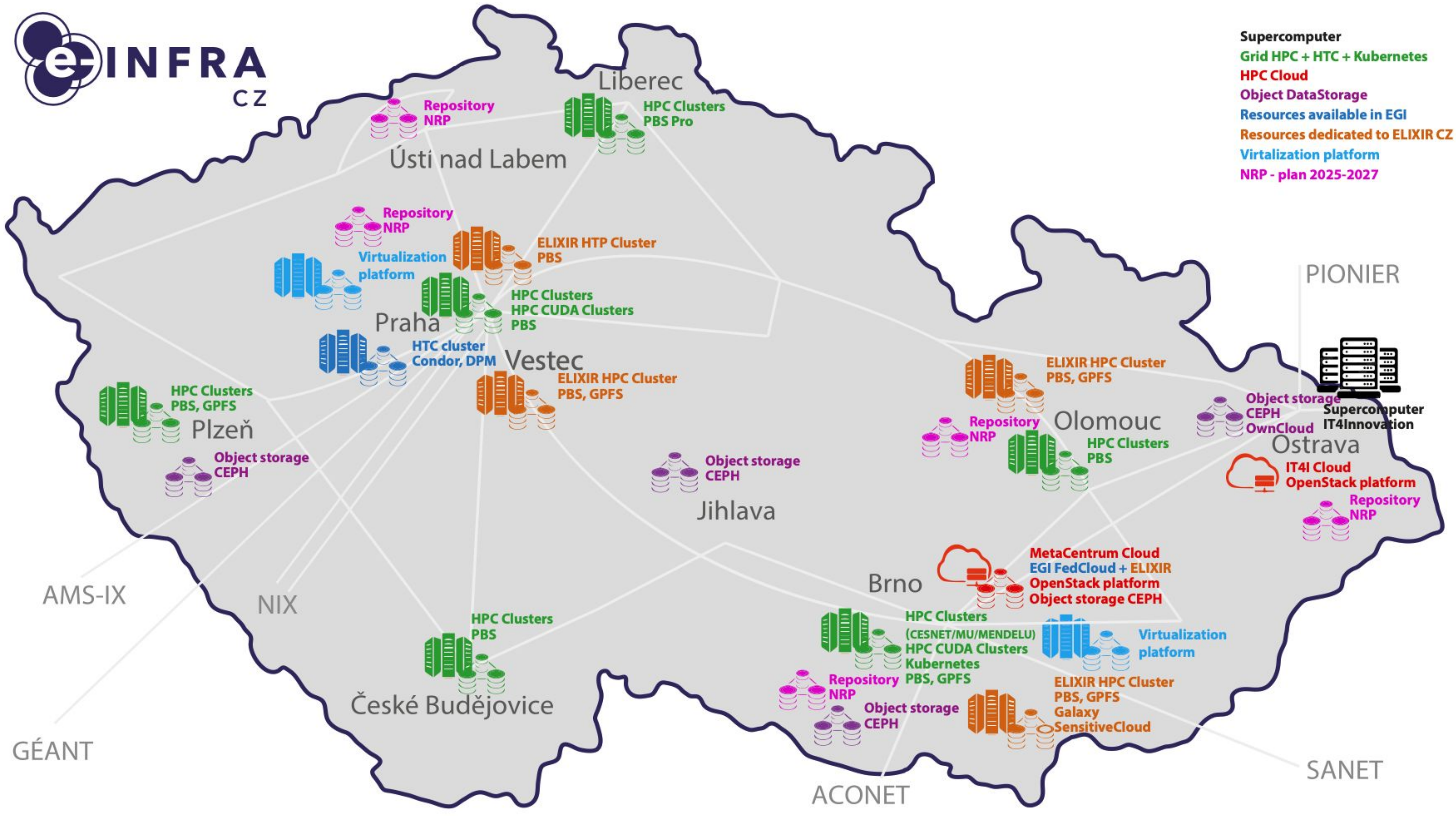
slides at: **ces.net/telc**

Galaxy Project CVMFS network

- ● Stratum 0 servers
- ● Stratum 1 servers


Digital Research Alliance of Canada
cvmfs-s1-galaxy.computecanada.ca
cvmfs0-psu0.galaxyproject.org
cvmfs0-psu1.galaxyproject.org
cvmfs1-psu0.galaxyproject.org
Penn State
ACCESS/Jetstream2, Indiana University
cvmfs1-iu0.galaxyproject.org
CyVerse, TACC
cvmfs0-tacc0.galaxyproject.org
cvmfs1-tacc0.galaxyproject.org
cvmfs-stratum0.galaxyproject.eu
cvmfs1-ufr0.galaxyproject.eu
ELIXIR, de.NBI, RZ Freiburg
EU JRC, Ispra
galaxy.jrc.ec.europa.eu
Melbourne Bioinformatics
cvmfs1-mel0.gvl.org.au

e-INFRA CZ


Supercomputer
Grid HPC + HTC + Kubernetes
HPC Cloud
Object DataStorage
Resources available in EGI
Resources dedicated to ELIXIR CZ
Virtalization platform
NRP - plan 2025-2027
Repository NRP
Liberec
HPC Clusters PBS Pro
Ústí nad Labem
Repository NRP
ELIXIR HTP Cluster PBS
Virtualization platform
HPC Clusters HPC CUDA Clusters PBS
Praha
HTC cluster Condor, DPM
Vestec
ELIXIR HPC Cluster PBS, GPFS
HPC Clusters PBS, GPFS
Plzeň
Object storage CEPH
Object storage CEPH
Jihlava
PIONIER
ELIXIR HPC Cluster PBS, GPFS
Repository NRP
Olomouc
HPC Clusters PBS
Object storage CEPH OwnCloud
Supercomputer IT4Innovation
Ostrava
IT4I Cloud OpenStack platform
Repository NRP
MetaCentrum Cloud
EGI FedCloud + ELIXIR
OpenStack platform
Object storage CEPH
Brno
HPC Clusters (CESNET/MU/MENDELU) HPC CUDA Clusters Kubernetes PBS, GPFS
Virtualization platform
Repository NRP
Object storage CEPH
ELIXIR HPC Cluster PBS, GPFS Galaxy SensitiveCloud
HPC Clusters PBS
České Budějovice
AMS-IX
NIX
GÉANT
ACONET
SANET

GTN

Galaxy Training Academy this week (~3k ppl)

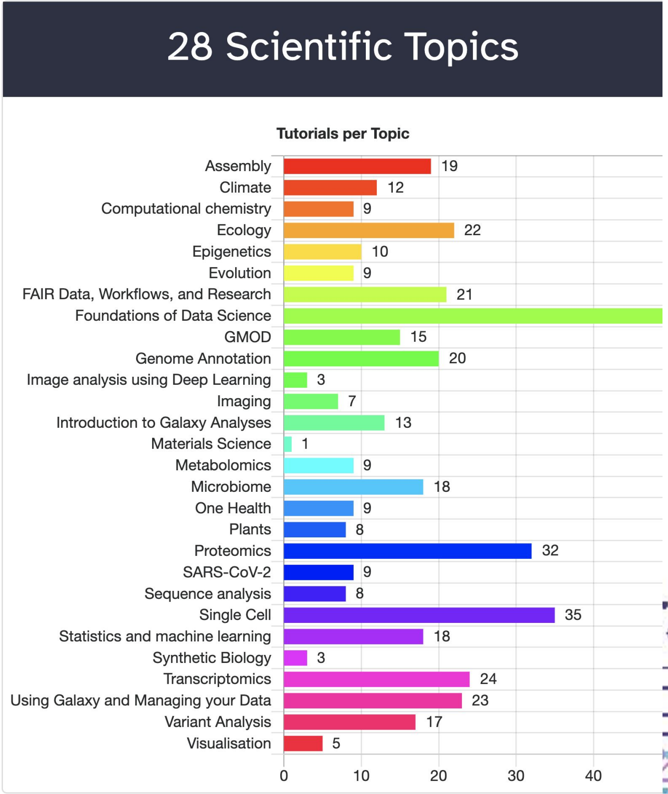
Per public stats (training.galaxyproject.org/training-material/s

419 trainings

400 contributors

55k unique visitors per month

Couples with Zenodo for input data management


28 Scientific Topics
Tutorials per Topic
Assembly 19
Climate 12
Computational chemistry 9
Ecology 22
Epigenetics 10
Evolution 9
FAIR Data, Workflows, and Research 21
Foundations of Data Science
GMOD 15
Genome Annotation 20
Image analysis using Deep Learning 3
Imaging 7
Introduction to Galaxy Analyses 13
Materials Science 1
Metabolomics 9
Microbiome 18
One Health 9
Plants 8
Proteomics 32
SARS-CoV-2 9
Sequence analysis 8
Single Cell 35
Statistics and machine learning 18
Synthetic Biology 3
Transcriptomics 24
Using Galaxy and Managing your Data 23
Variant Analysis 17
Visualisation 5