Data Analysis

Speakers: Siri Kallhovd (UiB),
Espen Robertsen (UiT), and
Erik Hjerde (UiT)
Moderator: Alexandra Gade (NCMM, UiO)

elixir NORWAY

15 June 2022

Siri Kallhovd (UiB)

Computing Resources

✓ Learning Objectives

In this session, we will learn:

- what High-Performance Computing (HPC) is
- how to get access to HPC systems provided by Sigma2 for Norwegian research
- how to find other local computing resources available at your university

NRIS

HPC Systems

Sigma2 owns the HPC Systems Saga, Fram and Betzy and NRIS operates the systems.

NRIS is a collaboration of the four BOTT universities and Sigma2 to pool competencies, resources and services.

Apply for resources: https://www.sigma2.no/high-performance-computing

Saga

System: HPE Apollo 2000

Number of cores: 16064

Number of nodes: 364 + 8 bigmem + 8 GPU nodes

Max floating point performance, double: 0.65 Petaflop/s

Total memory: 97.5 TiB

Total disk capacity: 5.3 PB (++)

Network: FDR IB (56 Gbit)

File System: BeeGFS, HPE 4150

NRIS

elixir

Software modules

Need for an isolated environment on a shared HPC resource

(Scientific) software should not be installed in the global PATH

Need some way of bringing software in and out of your environment

Many different solutions to this:

- Lmod (modules) + EasyBuild (installation)
- Virtualenv + pip
- Conda
- Singularity containers

We use Lmod + EasyBuild for the scientific software stack on all our HPC machines (Fram, Saga and Betzy)

Users can use any of the above to manage their own software stack on our machines

EasyBuild

EasyBuild can be used as wrapper around different ways of installing software

- Targeting HPC systems
- Automation: less prone for human errors
- Reproducibility: rebuild software stack, (usually) portable to other systems
- Performance: (usually) build from source for particular hardware
- Very pedantic with versioning and dependencies

EasyBuild is tightly connected to the module system and automatically generates module files after installation

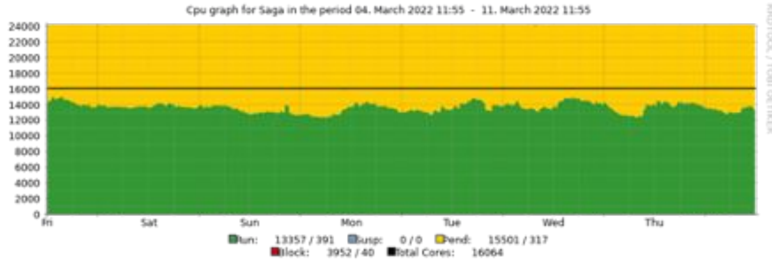
Slurm jobs optimize the usage of HPC systems

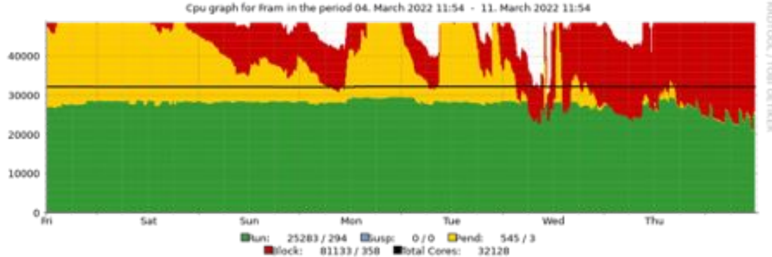
Job scripts specify:

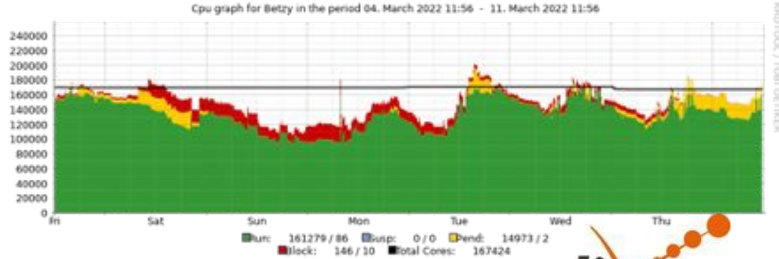
memory: Asking for too much can mean that you block idle CPUs and get charged for them

cores: Asking for too few can lead to underused nodes or longer run time

time: Asking for too little cuts the job







General local computing resources

NREC (Norwegian Research and Education Cloud, nrec.no)

-operated by UiB and UiO, the platform is built on OpenStack, users can create virtual machines

Fox: local HPC machine at UiO

Idun: local HPC machine at NTNU

Galaxy Introduction to Galaxy

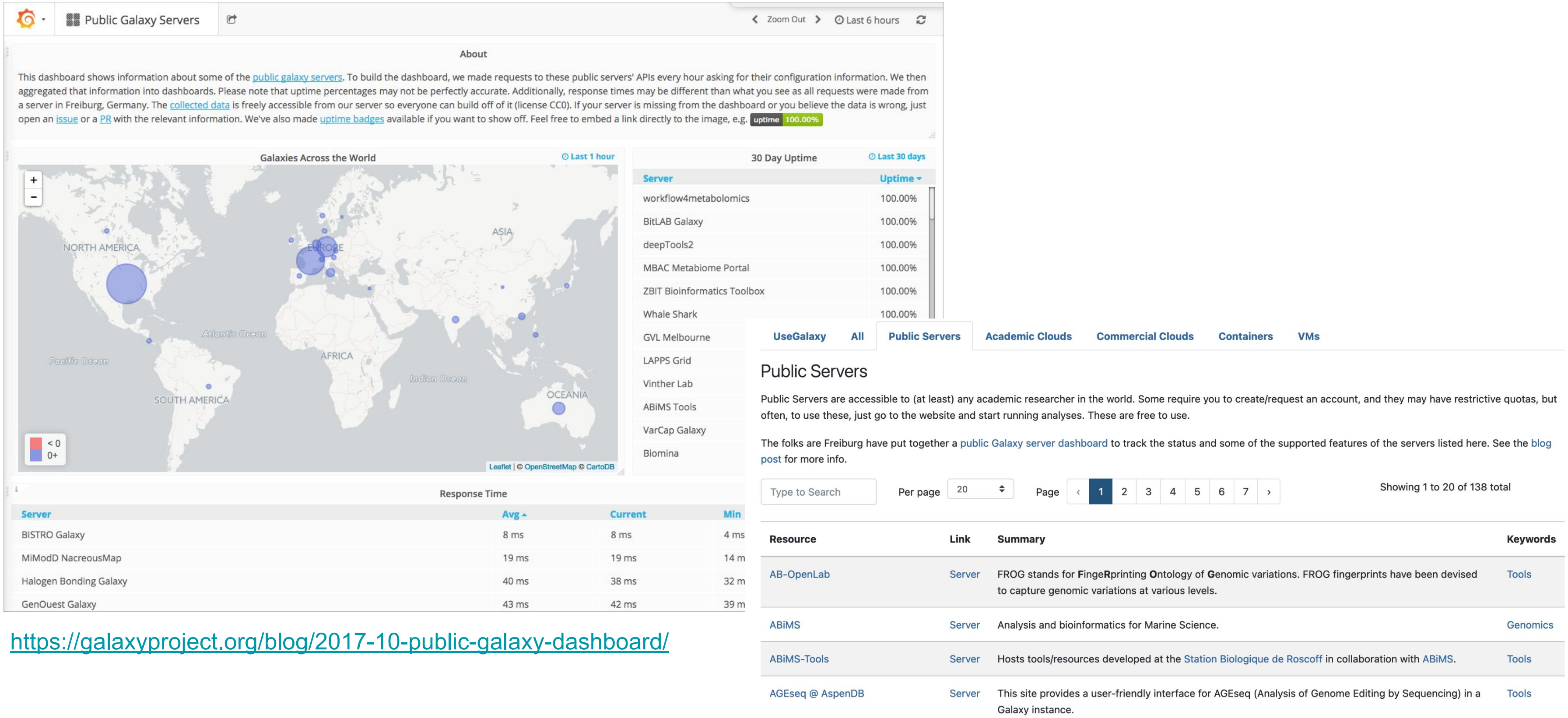
Erik Hjerde (UiT)

✓ Learning Objectives

At the end of this session, you will be able to:

- start exploring the Norwegian Galaxy instance: usegalaxy.no
- test 2000 analysis tools
- disseminate information to other users about this service

Galaxy platform - browser accessible workbench for scientific computing



About

This dashboard shows information about some of the public galaxy servers. To build the dashboard, we made requests to these public servers' APIs every hour asking for their configuration information. We then aggregated that information into dashboards. Please note that uptime percentages may not be perfectly accurate. Additionally, response times may be different than what you see as all requests were made from a server in Freiburg, Germany. The collected data is freely accessible from our server so everyone can build off of it (license CC0). If your server is missing from the dashboard or you believe the data is wrong, just open an issue or a PR with the relevant information. We've also made uptime badges available if you want to show off. Feel free to embed a link directly to the image, e.g. uptime 100.00%

30 Day Uptime

Server	Uptime
workflow4metabolomics	100.00%
BitLAB Galaxy	100.00%
deepTools2	100.00%
MBAC Metabiome Portal	100.00%
ZBIT Bioinformatics Toolbox	100.00%
Whale Shark	100.00%
GVL Melbourne	
LAPPS Grid	
Vinther Lab	
ABiMS Tools	
VarCap Galaxy	
Biomina	

Response Time

Server	Avg	Current	Min
BISTRO Galaxy	8 ms	8 ms	4 ms
MiModD NacreousMap	19 ms	19 ms	14 m
Halogen Bonding Galaxy	40 ms	38 ms	32 m
GenOuest Galaxy	43 ms	42 ms	39 m

Public Servers

Public Servers are accessible to (at least) any academic researcher in the world. Some require you to create/request an account, and they may have restrictive quotas, but often, to use these, just go to the website and start running analyses. These are free to use.

The folks are Freiburg have put together a public Galaxy server dashboard to track the status and some of the supported features of the servers listed here. See the blog post for more info.

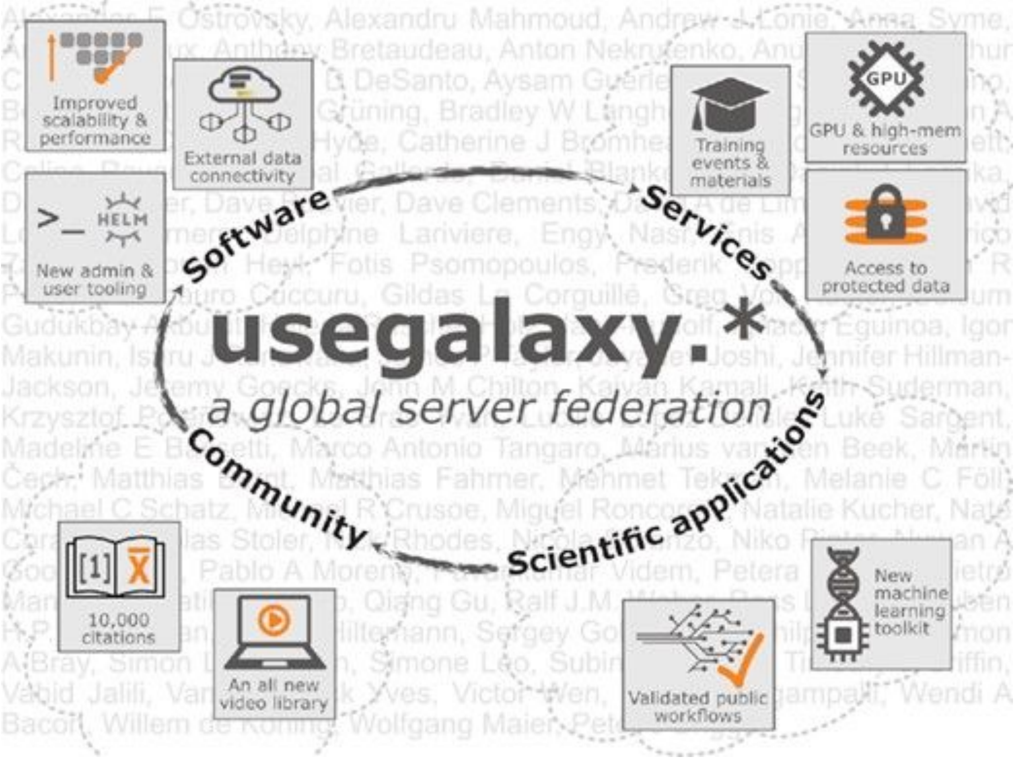
Type to Search   Per page 20   Page ‹ 1 2 3 4 5 6 7 ›   Showing 1 to 20 of 138 total

Resource	Link	Summary	Keywords
AB-OpenLab	Server	FROG stands for **F**inge**R**printing **O**ntology of **G**enomic variations. FROG fingerprints have been devised to capture genomic variations at various levels.	Tools
ABiMS	Server	Analysis and bioinformatics for Marine Science.	Genomics
ABiMS-Tools	Server	Hosts tools/resources developed at the Station Biologique de Roscoff in collaboration with ABiMS.	Tools
AGEseq @ AspenDB	Server	This site provides a user-friendly interface for AGEseq (Analysis of Genome Editing by Sequencing) in a Galaxy instance.	Tools

https://galaxyproject.org/blog/2017-10-public-galaxy-dashboard/

usegalaxy.* – Federation of free public Galaxy servers

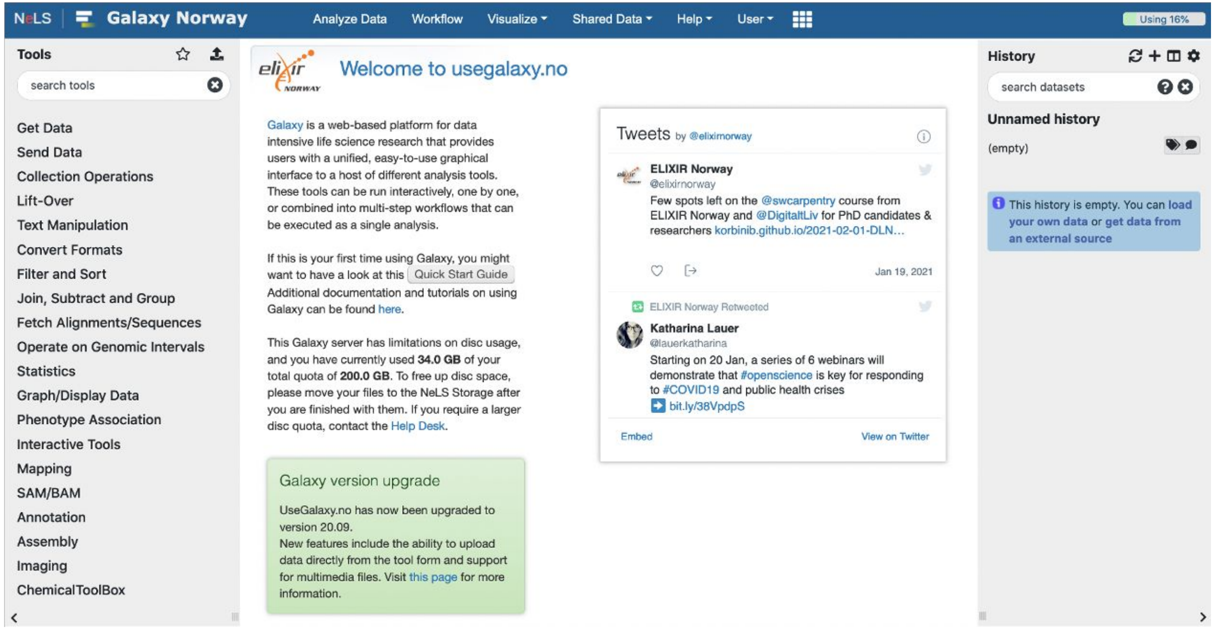
- Common core set of tools and reference genomes
- Open to anyone to use
- Backed by significant computational resources
- Support from multiple national infrastructure providers
- Galaxy Training Network



https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac247/6572001

usegalaxy.no – The Norwegian national Galaxy server

- Provide ~2000 bioinformatic tools and workflows
- Open to all Norwegian user and collaborators
- Directly connected to the NeLS storage
- Backed by significant computational resources
- End-user training and support through ELIXIR Norway support desk



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide. Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09. New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

Tweets by @elixirnorway

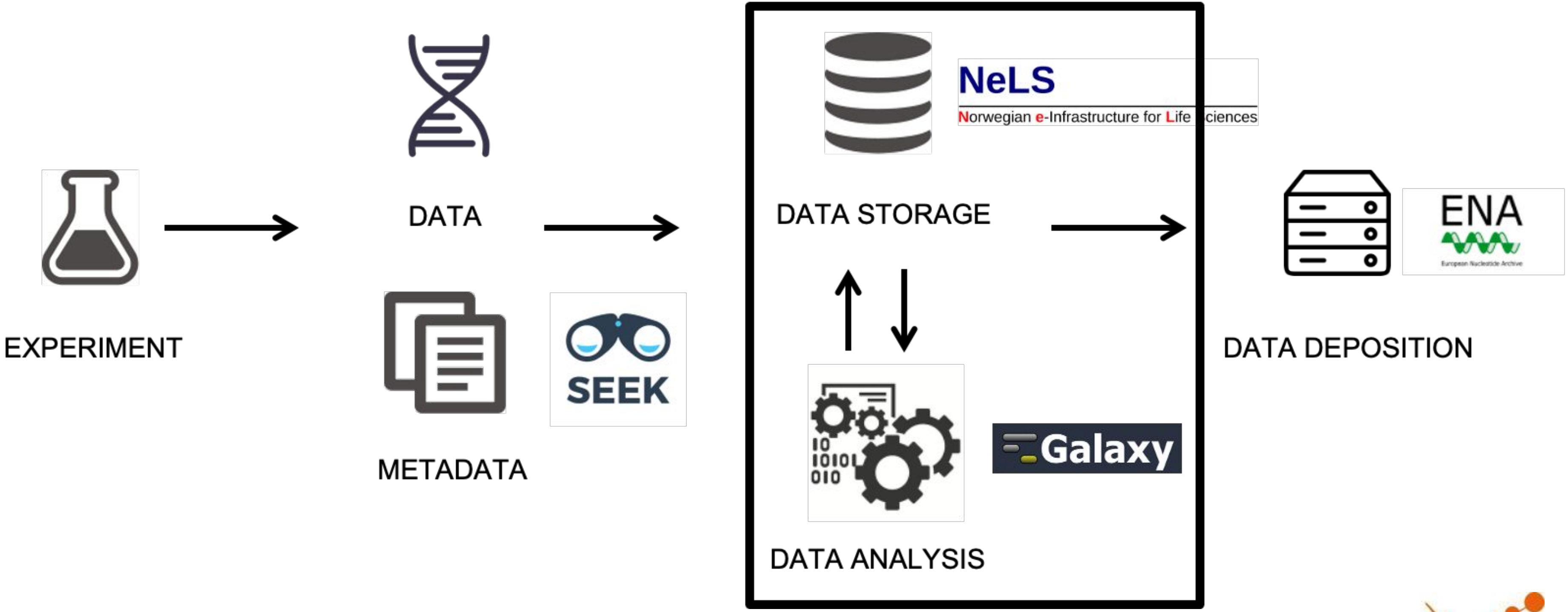
ELIXIR Norway @elixirnorway
Few spots left on the @swcarpentry course from ELIXIR Norway and @DigitaltLiv for PhD candidates & researchers korbinib.github.io/2021-02-01-DLN...
Jan 19, 2021

ELIXIR Norway Retweeted
Katharina Lauer @lauerkatharina
Starting on 20 Jan, a series of 6 webinars will demonstrate that #openscience is key for responding to #COVID19 and public health crises
bit.ly/38VpdpS

Embed
View on Twitter

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

ELIXIR Norway - analysis connected with storage



EXPERIMENT

DATA

METADATA

DATA STORAGE

DATA ANALYSIS

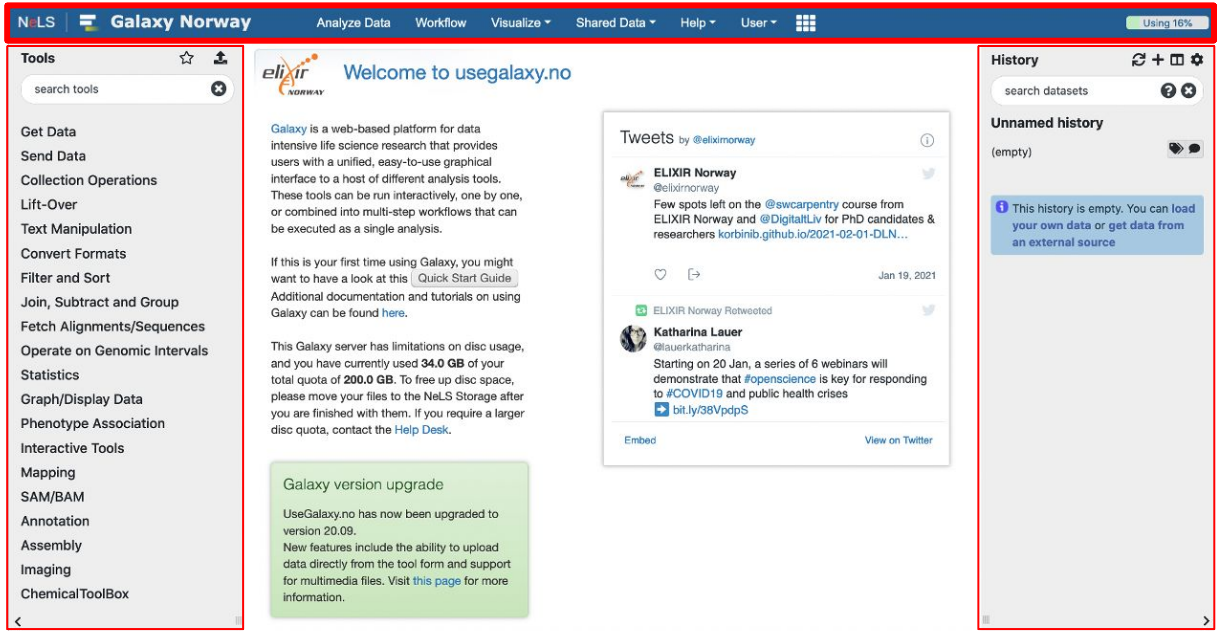
DATA DEPOSITION

NeLS
Norwegian e-Infrastructure for Life ciences
SEEK
Galaxy
ENA
European Nucleotide Archive

usegalaxy.no – The basics

- Important features:
- Tool menu with ~2000 tools sorted in sections
- Current disk usage (default is 200 GB total personal disk space)
- Server alerts
- Quick start guide
- Contact support
- Q&A forum

Main menu



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide. Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09.
New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

Tweets by @elixirnorway

ELIXIR Norway @elixirnorway
Few spots left on the @swcarpentry course from ELIXIR Norway and @DigitaltLiv for PhD candidates & researchers korbinib.github.io/2021-02-01-DLN...
Jan 19, 2021

ELIXIR Norway Retweeted
Katharina Lauer @lauerkatharina
Starting on 20 Jan, a series of 6 webinars will demonstrate that #openscience is key for responding to #COVID19 and public health crises
bit.ly/38VpdpS

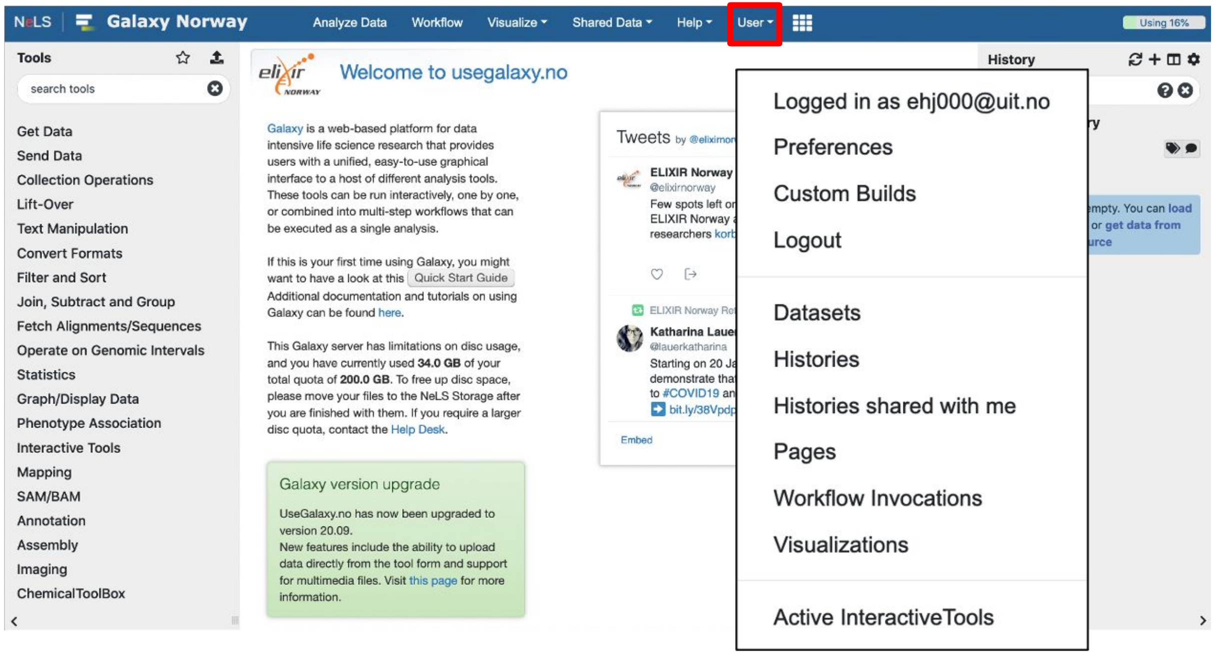
Embed View on Twitter

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

Tool menu Main window History

Your account and saved data

- Important features:
- Tool menu with ~2000 tools sorted in sections
- Current disk usage (default is 200 GB total personal disk space)
- Server alerts
- Quick start guide
- Contact support
- Q&A forum



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09.
New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

Tweets by @elixirnor

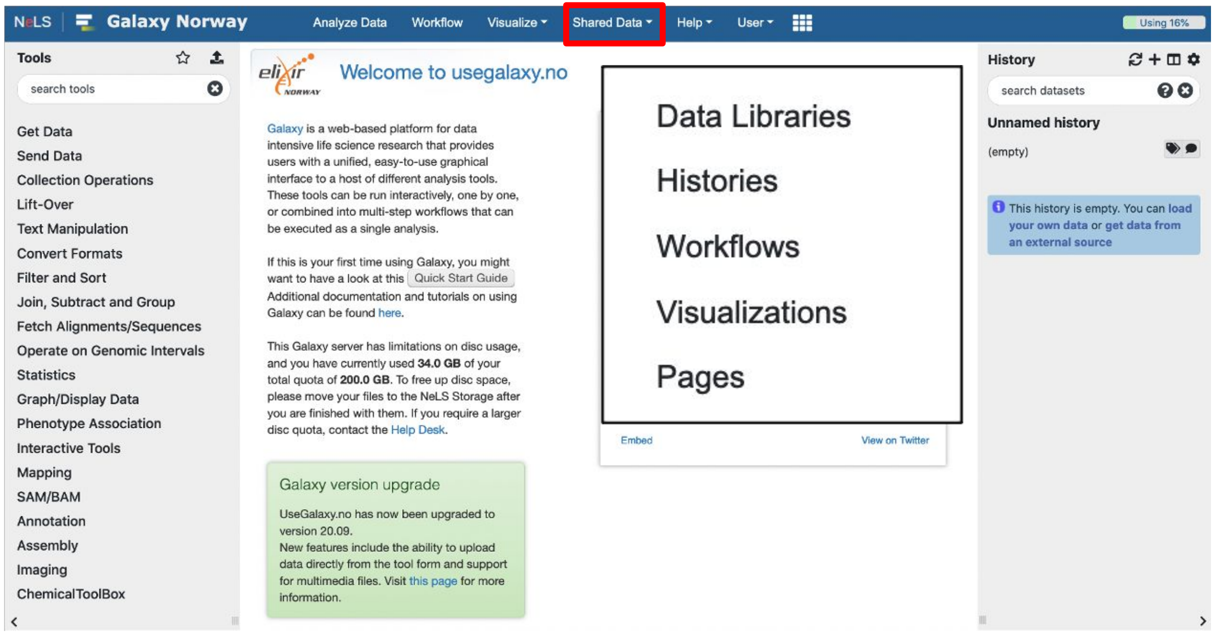
ELIXIR Norway
@elixirnorway
Few spots left or
ELIXIR Norway a
researchers korb

ELIXIR Norway Ret
Katharina Laue
@lauerkatharina
Starting on 20 Ja
demonstrate tha
to #COVID19 an
bit.ly/38Vpdp

Embed

Shared data

- Data shared by other users or ELIXIR-NO with all users of usegalaxy.no
- E.g. workflows and complete histories
- You can import shared data to you user
- Instructions how to use ELIXIR-NO supported workflows are also here



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

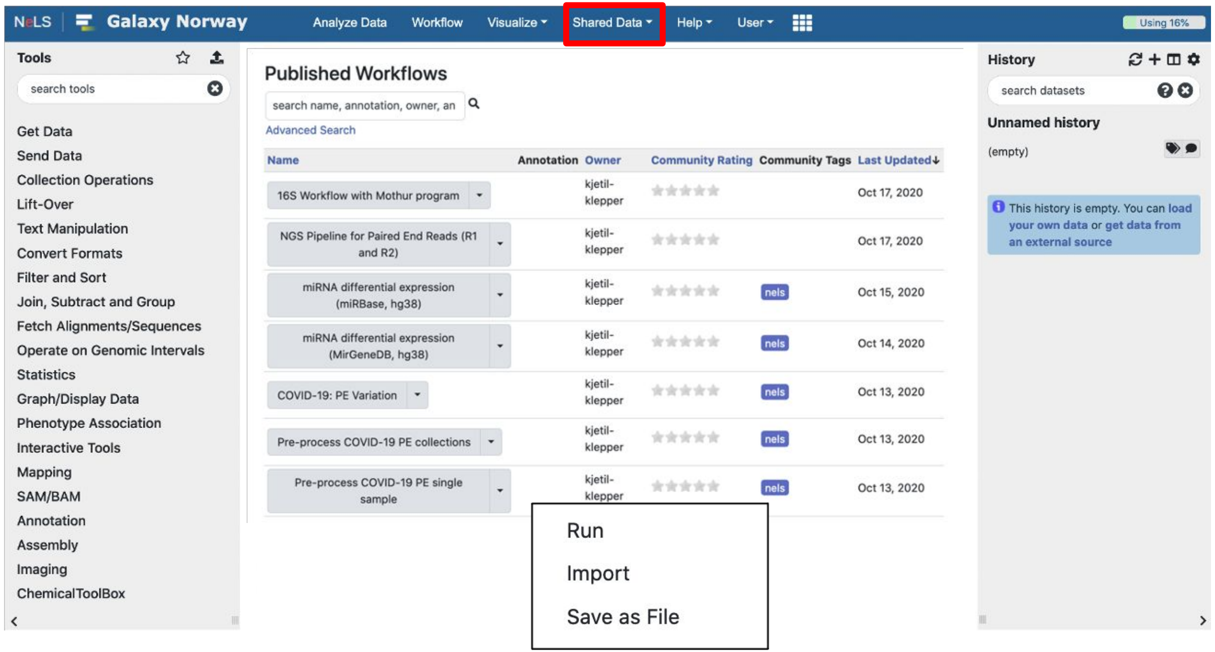
Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09.
New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

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

Shared data: workflows

- List of all workflows that are shared with all usegalaxy.no users
- You can import shared data to your user
- By selecting any workflow you can run data analysis, import into your user or save it on another computer



Published Workflows

search name, annotation, owner, an

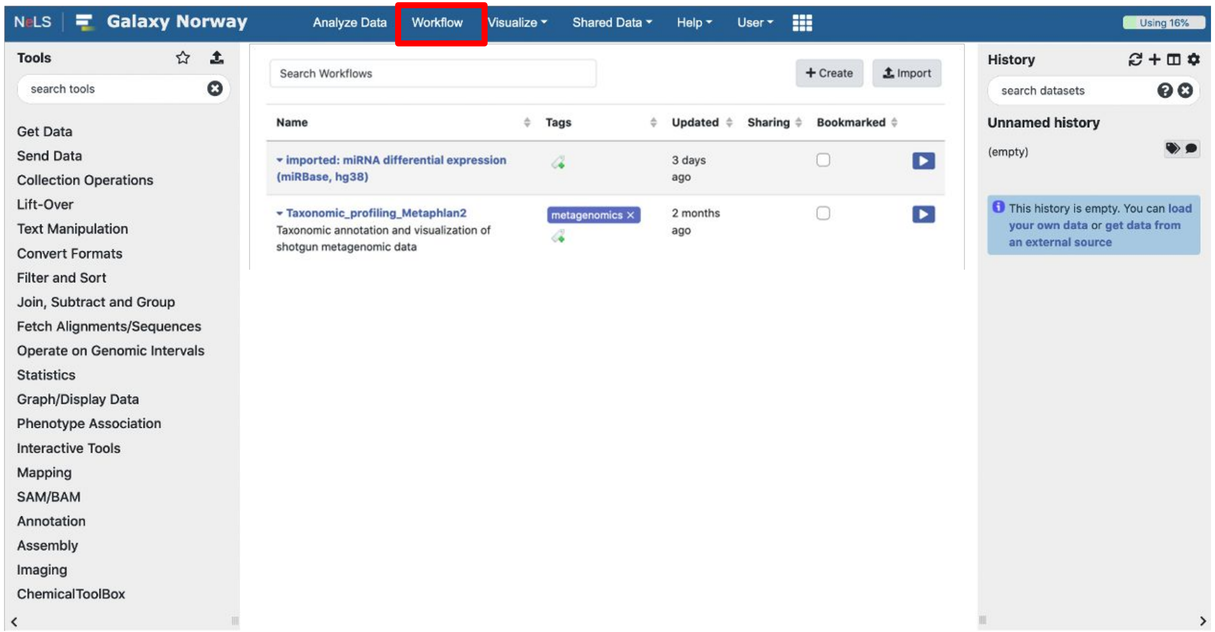
Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
16S Workflow with Mothur program		kjetil-klepper	☆☆☆☆☆		Oct 17, 2020
NGS Pipeline for Paired End Reads (R1 and R2)		kjetil-klepper	☆☆☆☆☆		Oct 17, 2020
miRNA differential expression (miRBase, hg38)		kjetil-klepper	☆☆☆☆☆	nels	Oct 15, 2020
miRNA differential expression (MirGeneDB, hg38)		kjetil-klepper	☆☆☆☆☆	nels	Oct 14, 2020
COVID-19: PE Variation		kjetil-klepper	☆☆☆☆☆	nels	Oct 13, 2020
Pre-process COVID-19 PE collections		kjetil-klepper	☆☆☆☆☆	nels	Oct 13, 2020
Pre-process COVID-19 PE single sample		kjetil-klepper	☆☆☆☆☆	nels	Oct 13, 2020

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

Workflows

- Your workflows. These are the imported or the workflows you have made
- You can create new workflows here



Name	Tags	Updated	Sharing	Bookmarked
imported: miRNA differential expression (miRBase, hg38)		3 days ago		☐
Taxonomic_profiling_Metaphlan2 Taxonomic annotation and visualization of shotgun metagenomic data	metagenomics ×	2 months ago		☐

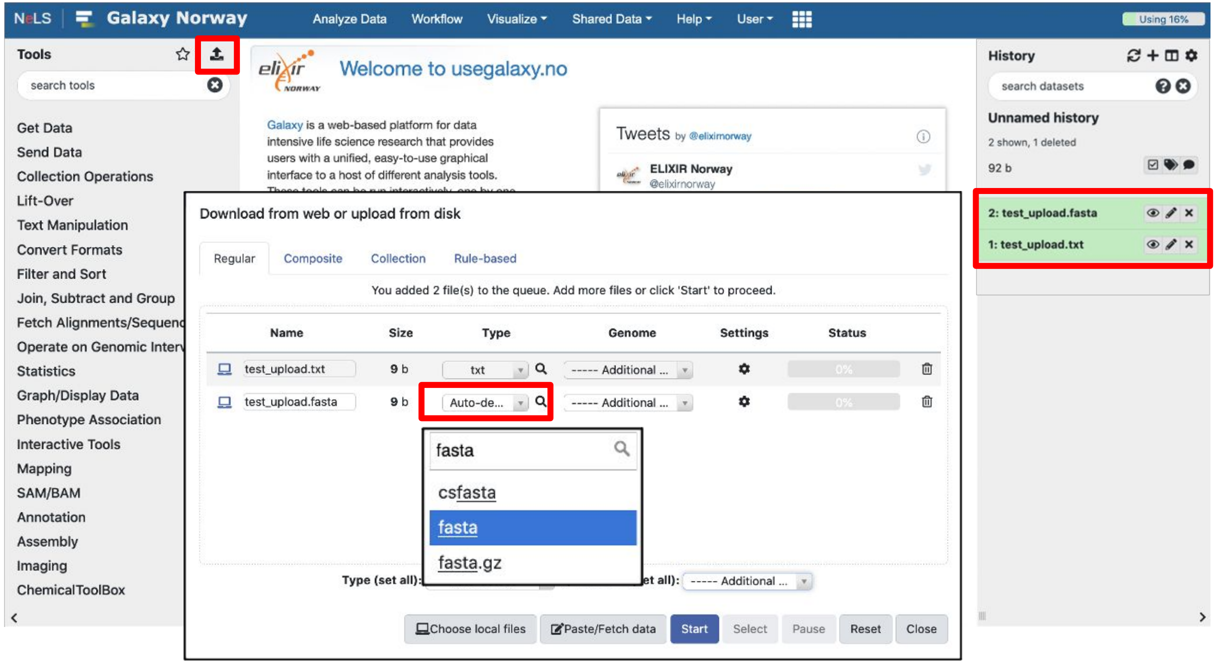
Unnamed history

(empty)

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

Import data from your computer or from the web

- Drag and drop, or browse and select file
- Alternatively, paste url for data available on the web
- Specify data type if you know (e.g. Fasta)
- The two imported files will appear as two datasets in your history



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools.

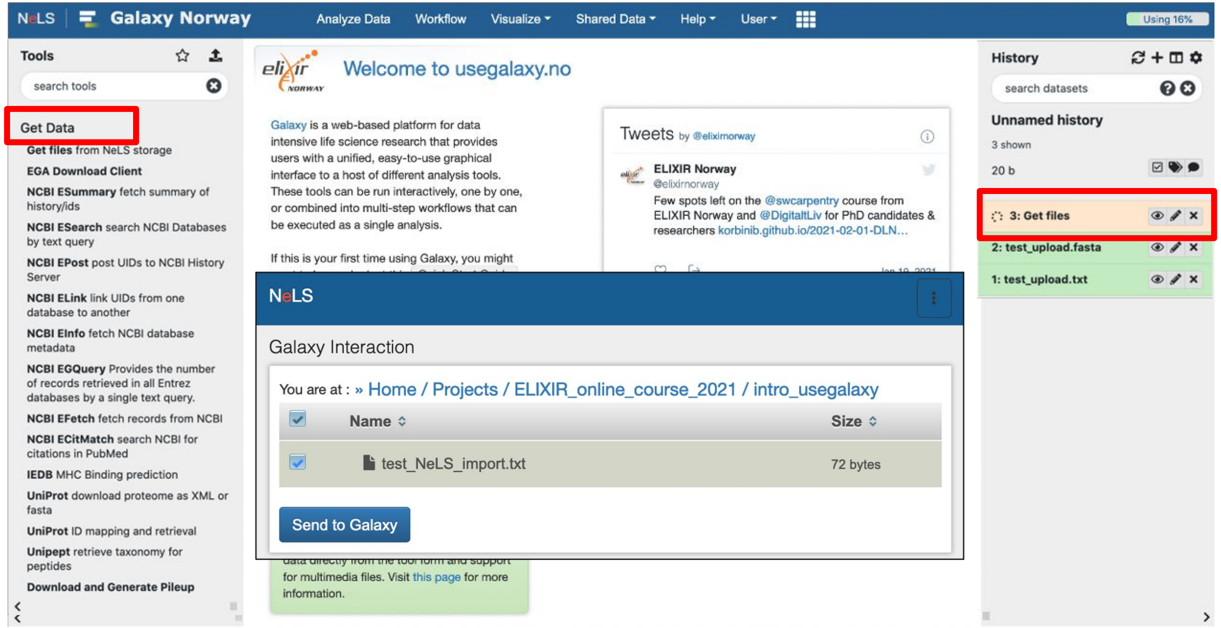
Download from web or upload from disk

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
test_upload.txt	9 b	txt	----- Additional ...		0%
test_upload.fasta	9 b	Auto-de...	----- Additional ...		0%

Import data from NeLS

- Import data from Personal or Project folders in NeLS
- Redirect to the NeLS portal (require login)
- Files are selectable
- Imported data from NeLS will appear in your history
- Note: the yellow colour of files as this jobs are being processed (green = job complete)



Get files from NeLS storage
EGA Download Client
NCBI ESummary fetch summary of history/ids
NCBI ESearch search NCBI Databases by text query
NCBI EPost post UIDs to NCBI History Server
NCBI ELink link UIDs from one database to another
NCBI EInfo fetch NCBI database metadata
NCBI EGQuery Provides the number of records retrieved in all Entrez databases by a single text query.
NCBI EFetch fetch records from NCBI
NCBI ECitMatch search NCBI for citations in PubMed
IEDB MHC Binding prediction
UniProt download proteome as XML or fasta
UniProt ID mapping and retrieval
Unipept retrieve taxonomy for peptides
Download and Generate Pileup

Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might

Tweets by @elixirnorway

ELIXIR Norway @elixirnorway
Few spots left on the @swcarpentry course from ELIXIR Norway and @DigitaltLiv for PhD candidates & researchers korbinib.github.io/2021-02-01-DLN...

for multimedia files. Visit this page for more information.

Galaxy Interaction

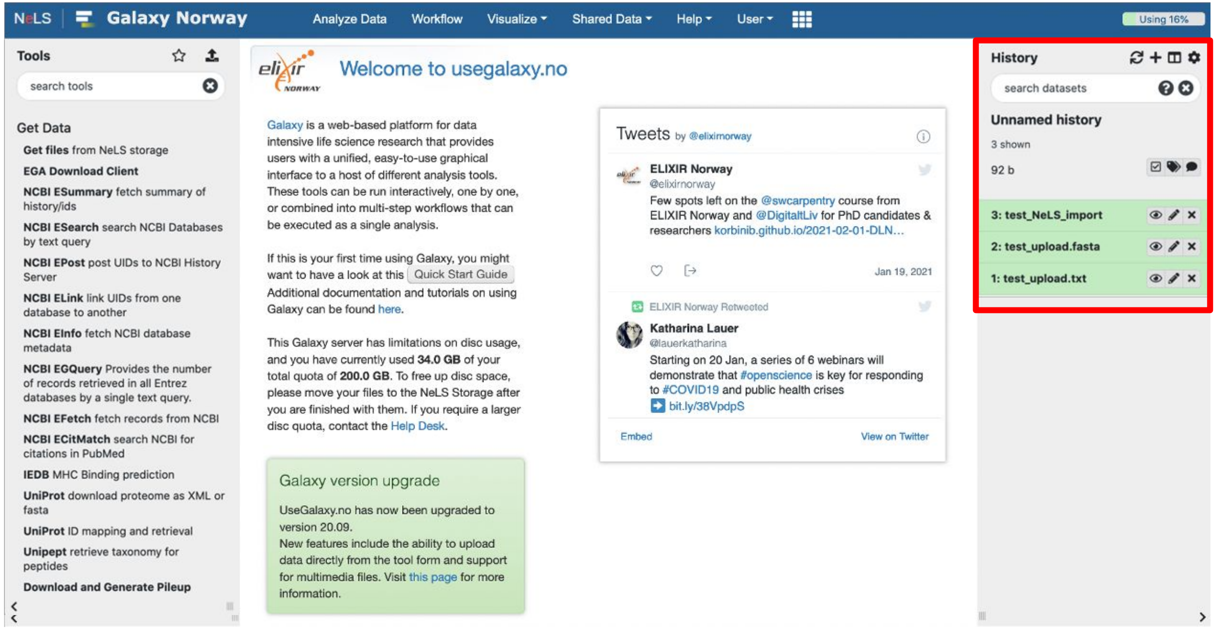
You are at : » Home / Projects / ELIXIR_online_course_2021 / intro_usegalaxy

☑	Name	Size
☑	test_NeLS_import.txt	72 bytes

Unnamed history
3 shown
20 b
3: Get files
2: test_upload.fasta
1: test_upload.txt

Use galaxy histories to organise data

- The current history is ”your current work space”
- The history panel displays datasets in the order in which they were created
- You can make as many histories as you want and switch between them
- Typically, you can have one history for each project or analysis



Get Data

Get files from NeLS storage

EGA Download Client

NCBI ESummary fetch summary of history/ids

NCBI ESearch search NCBI Databases by text query

NCBI EPost post UIDs to NCBI History Server

NCBI ELink link UIDs from one database to another

NCBI EInfo fetch NCBI database metadata

NCBI EGQuery Provides the number of records retrieved in all Entrez databases by a single text query.

NCBI EFetch fetch records from NCBI

NCBI ECitMatch search NCBI for citations in PubMed

IEDB MHC Binding prediction

UniProt download proteome as XML or fasta

UniProt ID mapping and retrieval

Unipept retrieve taxonomy for peptides

Download and Generate Pileup

Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09.
New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

Tweets by @elixirnorway

ELIXIR Norway
@elixirnorway
Few spots left on the @swcarpentry course from ELIXIR Norway and @DigitalLiv for PhD candidates & researchers korbinib.github.io/2021-02-01-DLN…

Jan 19, 2021

ELIXIR Norway Retweeted

Katharina Lauer
@lauerkatharina
Starting on 20 Jan, a series of 6 webinars will demonstrate that #openscience is key for responding to #COVID19 and public health crises
bit.ly/38VpdpS

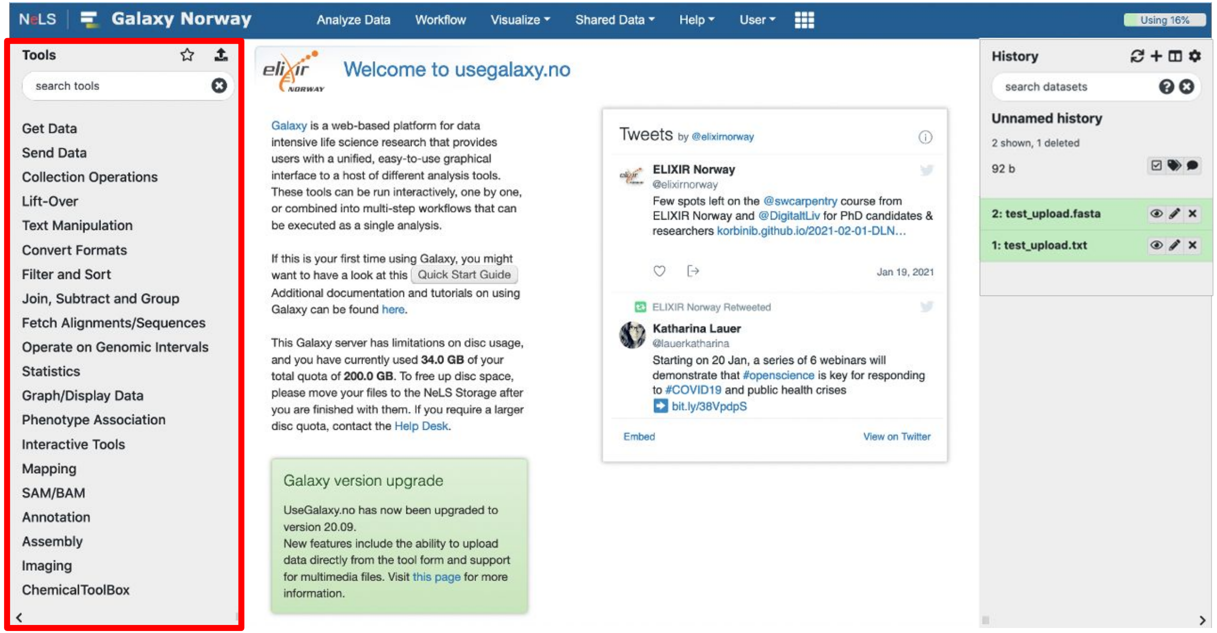
Embed
View on Twitter

Unnamed history
3 shown
92 b

3: test_NeLS_import
2: test_upload.fasta
1: test_upload.txt

Galaxy tools

- Tools are available from the Tool menu
- Organised under sub-menus
- Possible to browse and search by name
- You can make your own list of favourite tools



Welcome to usegalaxy.no

Galaxy is a web-based platform for data intensive life science research that provides users with a unified, easy-to-use graphical interface to a host of different analysis tools. These tools can be run interactively, one by one, or combined into multi-step workflows that can be executed as a single analysis.

If this is your first time using Galaxy, you might want to have a look at this Quick Start Guide Additional documentation and tutorials on using Galaxy can be found here.

This Galaxy server has limitations on disc usage, and you have currently used **34.0 GB** of your total quota of **200.0 GB**. To free up disc space, please move your files to the NeLS Storage after you are finished with them. If you require a larger disc quota, contact the Help Desk.

Galaxy version upgrade

UseGalaxy.no has now been upgraded to version 20.09.
New features include the ability to upload data directly from the tool form and support for multimedia files. Visit this page for more information.

Tweets by @elixirnorway

ELIXIR Norway
@elixirnorway
Few spots left on the @swcarpentry course from ELIXIR Norway and @DigitalLiv for PhD candidates & researchers korbinib.github.io/2021-02-01-DLN...

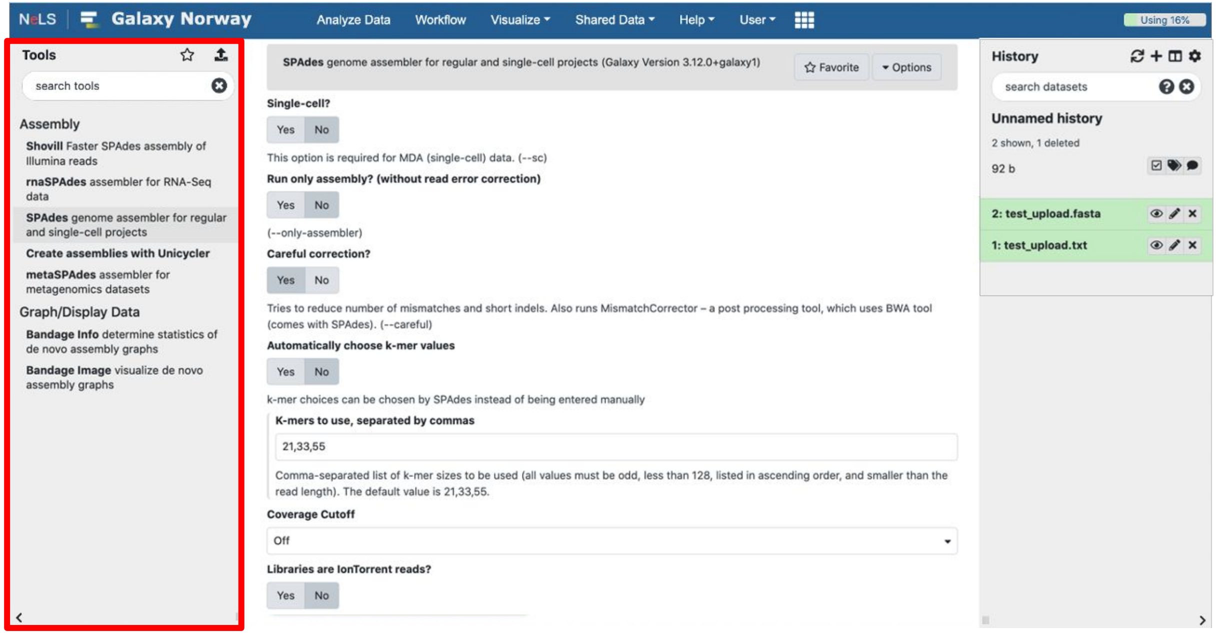
Jan 19, 2021

ELIXIR Norway Retweeted
Katharina Lauer
@lauerkatharina
Starting on 20 Jan, a series of 6 webinars will demonstrate that #openscience is key for responding to #COVID19 and public health crises
bit.ly/38VpdpS

Embed View on Twitter

Galaxy tools = command line tools

- Command line tools are wrapped into Galaxy so they become accessible with a GUI
- Some Galaxy tools may have reduced the number of optional parameter settings for the tool
- Example here is the assembly tool called SPAdes



Assembly

Shovill Faster SPAdes assembly of Illumina reads

rnaSPAdes assembler for RNA-Seq data

SPAdes genome assembler for regular and single-cell projects

Create assemblies with Unicycler

metaSPAdes assembler for metagenomics datasets

Graph/Display Data

Bandage Info determine statistics of de novo assembly graphs

Bandage Image visualize de novo assembly graphs

SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.12.0+galaxy1)

Single-cell?

Yes No

This option is required for MDA (single-cell) data. (--sc)

Run only assembly? (without read error correction)

Yes No

(--only-assembler)

Careful correction?

Yes No

Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector – a post processing tool, which uses BWA tool (comes with SPAdes). (--careful)

Automatically choose k-mer values

Yes No

k-mer choices can be chosen by SPAdes instead of being entered manually

K-mers to use, separated by commas

21,33,55

Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.

Coverage Cutoff

Off

Libraries are IonTorrent reads?

Yes No

Unnamed history

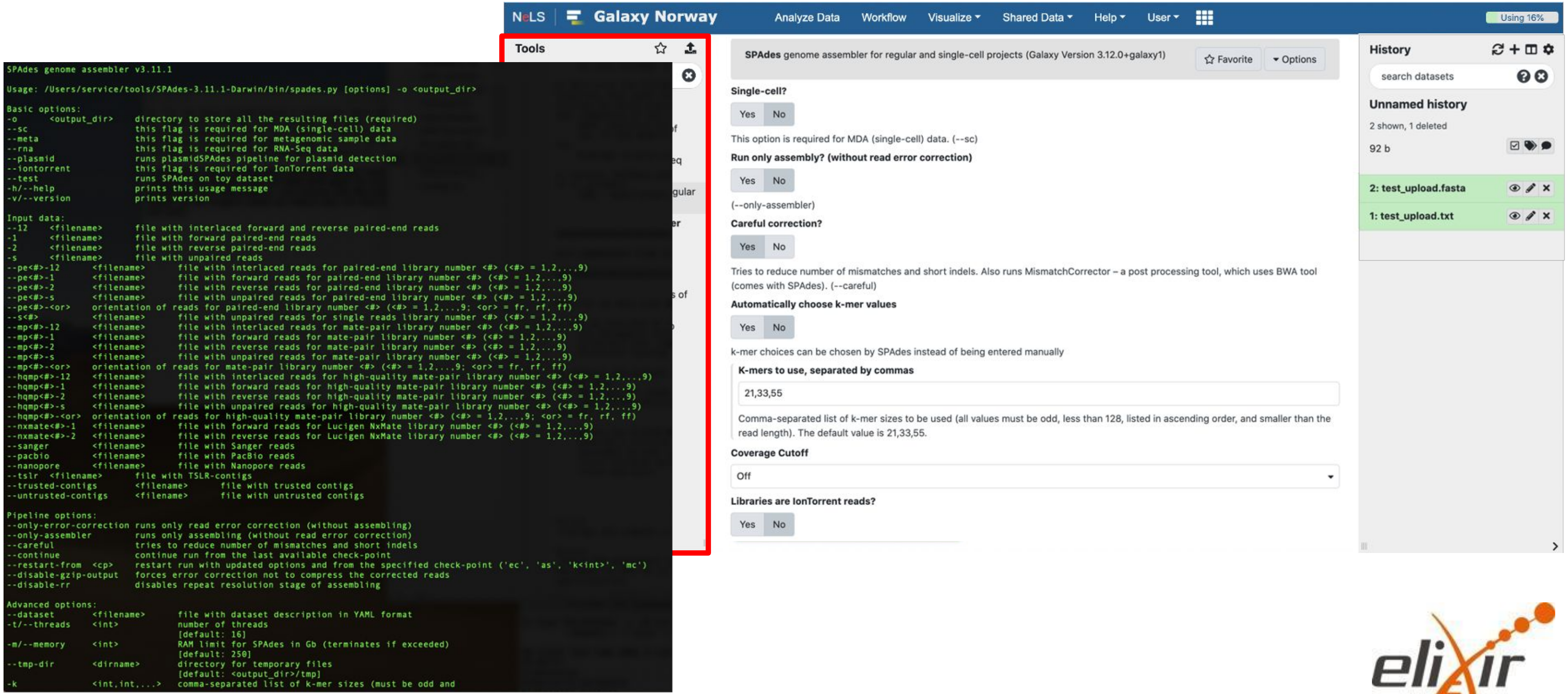
2 shown, 1 deleted

92 b

2: test_upload.fasta

1: test_upload.txt

Galaxy tools = command line tools



```
SPAdes genome assembler v3.11.1

Usage: /Users/service/tools/SPAdes-3.11.1-Darwin/bin/spades.py [options] -o <output_dir>

Basic options:
-o      <output_dir>    directory to store all the resulting files (required)
--sc                    this flag is required for MDA (single-cell) data
--meta                  this flag is required for metagenomic sample data
--rna                   this flag is required for RNA-Seq data
--plasmid               runs plasmidSPAdes pipeline for plasmid detection
--iontorrent            this flag is required for IonTorrent data
--test                  runs SPAdes on toy dataset
-h/--help               prints this usage message
-v/--version            prints version

Input data:
--12    <filename>      file with interlaced forward and reverse paired-end reads
-1      <filename>      file with forward paired-end reads
-2      <filename>      file with reverse paired-end reads
-s      <filename>      file with unpaired reads
--pe<#>-12      <filename>      file with interlaced reads for paired-end library number <#> (<#> = 1,2,...,9)
--pe<#>-1       <filename>      file with forward reads for paired-end library number <#> (<#> = 1,2,...,9)
--pe<#>-2       <filename>      file with reverse reads for paired-end library number <#> (<#> = 1,2,...,9)
--pe<#>-s       <filename>      file with unpaired reads for paired-end library number <#> (<#> = 1,2,...,9)
--pe<#>-<or>    orientation of reads for paired-end library number <#> (<#> = 1,2,...,9; <or> = fr, rf, ff)
--s<#>          <filename>      file with unpaired reads for single reads library number <#> (<#> = 1,2,...,9)
--mp<#>-12      <filename>      file with interlaced reads for mate-pair library number <#> (<#> = 1,2,...,9)
--mp<#>-1       <filename>      file with forward reads for mate-pair library number <#> (<#> = 1,2,...,9)
--mp<#>-2       <filename>      file with reverse reads for mate-pair library number <#> (<#> = 1,2,...,9)
--mp<#>-s       <filename>      file with unpaired reads for mate-pair library number <#> (<#> = 1,2,...,9)
--mp<#>-<or>    orientation of reads for mate-pair library number <#> (<#> = 1,2,...,9; <or> = fr, rf, ff)
--hqmp<#>-12    <filename>      file with interlaced reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmp<#>-1     <filename>      file with forward reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmp<#>-2     <filename>      file with reverse reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmp<#>-s     <filename>      file with unpaired reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9)
--hqmp<#>-<or>  orientation of reads for high-quality mate-pair library number <#> (<#> = 1,2,...,9; <or> = fr, rf, ff)
--nxmate<#>-1   <filename>      file with forward reads for Lucigen NxMate library number <#> (<#> = 1,2,...,9)
--nxmate<#>-2   <filename>      file with reverse reads for Lucigen NxMate library number <#> (<#> = 1,2,...,9)
--sanger        <filename>      file with Sanger reads
--pacbio        <filename>      file with PacBio reads
--nanopore      <filename>      file with Nanopore reads
--tslr  <filename>      file with TSLR-contigs
--trusted-contigs       <filename>      file with trusted contigs
--untrusted-contigs     <filename>      file with untrusted contigs

Pipeline options:
--only-error-correction runs only read error correction (without assembling)
--only-assembler        runs only assembling (without read error correction)
--careful               tries to reduce number of mismatches and short indels
--continue              continue run from the last available check-point
--restart-from  <cp>    restart run with updated options and from the specified check-point ('ec', 'as', 'k<int>', 'mc')
--disable-gzip-output   forces error correction not to compress the corrected reads
--disable-rr            disables repeat resolution stage of assembling

Advanced options:
--dataset       <filename>      file with dataset description in YAML format
-t/--threads    <int>           number of threads
                                [default: 16]
-m/--memory     <int>           RAM limit for SPAdes in Gb (terminates if exceeded)
                                [default: 250]
--tmp-dir       <dirname>       directory for temporary files
                                [default: <output_dir>/tmp]
-k              <int,int,...>   comma-separated list of k-mer sizes (must be odd and
```

SPAdes genome assembler for regular and single-cell projects (Galaxy Version 3.12.0+galaxy1)

Single-cell?

Yes [✓ No]

This option is required for MDA (single-cell) data. (--sc)

Run only assembly? (without read error correction)

Yes [✓ No]

(--only-assembler)

Careful correction?

[✓ Yes] No

Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector – a post processing tool, which uses BWA tool (comes with SPAdes). (--careful)

Automatically choose k-mer values

Yes [✓ No]

k-mer choices can be chosen by SPAdes instead of being entered manually

K-mers to use, separated by commas

21,33,55

Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.

Coverage Cutoff

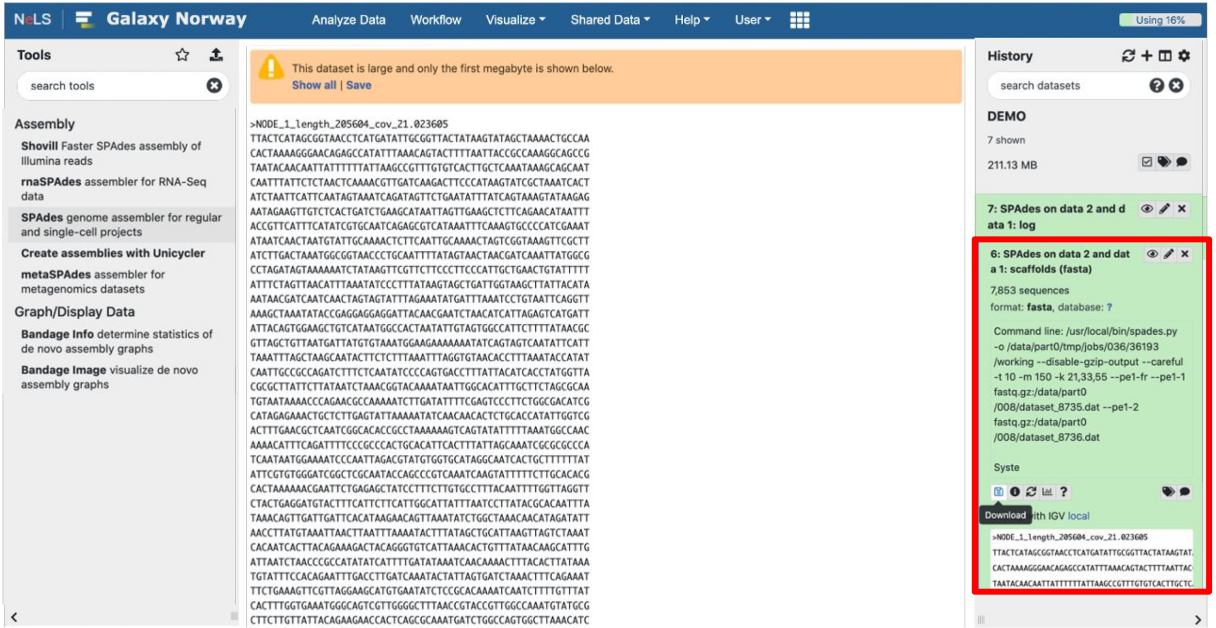
Off

Libraries are IonTorrent reads?

Yes [✓ No]

Tool output

- It is possible to preview the output (result), view it in the main window or download the dataset
- You can also copy the dataset over to another history


NeLS
Galaxy Norway
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 16%
Tools
search tools
Assembly
Shovill Faster SPAdes assembly of Illumina reads
rnaSPAdes assembler for RNA-Seq data
SPAdes genome assembler for regular and single-cell projects
Create assemblies with Unicycler
metaSPAdes assembler for metagenomics datasets
Graph/Display Data
Bandage Info determine statistics of de novo assembly graphs
Bandage Image visualize de novo assembly graphs
History
search datasets
Download

This dataset is large and only the first megabyte is shown below.
Show all | Save

```
>NODE_1_length_205604_cov_21.023605
TTACTCATAGCGGTAACCTCATGATATTGCGGTTACTATAAGTATAGCTAAAACTGCCAA
CACTAAAAGGGAACAGAGCCATATTTAAACAGTACTTTTAATTACCGCCAAAGGCAGCCG
TAATACAACAATTATTTTTTATTAAGCCGTTTGTGTCACTTGCTCAAATAAAGCAGCAAT
CAATTTATTCTCTAACTCAAAACGTTGATCAAGACTTCCCATAAGTATCGCTAAATCACT
ATCTAATTCATTCAATAGTAAATCAGATAGTTCTGAATATTTATCAGTAAAGTATAAGAG
AATAGAAGTTGTCTCACTGATCTGAAGCATAATTAGTTGAAGCTCTTCAGAACATAATTT
ACCGTTCATTTCATATCGTGCAATCAGAGCGTCATAAATTTCAAAGTGCCCCATCGAAAT
ATAATCAACTAATGTATTGCAAAACTCTTCAATTGCAAAACTAGTCGGTAAAGTTCGCTT
ATCTTGACTAAATGGCGGTAACCCTGCAATTTTATAGTAACTAACGATCAAATTATGGCG
CCTAGATAGTAAAAAATCTATAAGTTCGTTCTTCCCTTCCCATTGCTGAACTGTATTTTT
ATTTCTAGTTAACATTTAAATATCCCTTTATAAGTAGCTGATTGGTAAGCTTATTACATA
AATAACGATCAATCAACTAGTAGTATTTAGAAATATGATTTAAATCCTGTAATTCAGGTT
AAAGCTAAATATACCGAGGAGGAGGATTACAACGAATCTAACATCATTAGAGTCATGATT
ATTACAGTGGAAGCTGTCATAATGGCCACTAATATTGTAGTGGCCATTCTTTTATAACGC
GTTAGCTGTTAATGATTATGTGTAAATGGAAGAAAAAAATATCAGTAGTCAATATTCATT
TAAATTTAGCTAAGCAATACTTCTCTTTAAATTTAGGTGTAACACCTTTAAATACCATAT
CAATTGCCGCCAGATCTTTCTCAATATCCCCAGTGACCTTTATTACATCACCTATGGTTA
CGCGCTTATTCTTATAATCTAAACGGTACAAAATAATTGGCACATTTGCTTCTAGCGCAA
TGTAATAAAACCCAGAACGCCAAAAATCTTGATATTTTCGAGTCCCTTCTGGCGACATCG
CATAGAGAAACTGCTCTTGAGTATTAAAAATATCAACAACACTCTGCACCATATTGGTCG
ACTTTGAACGCTCAATCGGCACACCGCCTAAAAAAGTCAGTATATTTTTAAATGGCCAAC
AAAACATTTCAGATTTTCCCGCCCACTGCACATTCACTTTATTAGCAAATCGCGCGCCCA
TCAATAATGGAAAATCCCAATTAGACGTATGTGGTGCATAGGCAATCACTGCTTTTTTAT
ATTCGTGTGGGATCGGCTCGCAATACCAGCCCGTCAAATCAAGTATTTTTCTTGCACACG
CACTAAAAAACGAATTCTGAGAGCTATCCTTTCTTGTGCCTTTACAATTTTGGTTAGGTT
CTACTGAGGATGTACTTTCATTCTTCATTGGCATTATTTAATCCTTATACGCACAATTTA
TAAACAGTTGATTGATTCACATAAGAACAGTTAAATATCTGGCTAAACAACATAGATATT
AACCTTATGTAAATTAACTTAATTTAAAATACTTTATAGCTGCATTAAGTTAGTCTAAAT
CACAATCACTTACAGAAAGACTACAGGGTGTCATTAAACACTGTTTATAACAAGCATTTG
ATTAATCTAACCCGCCATATATCATTTTGATATAAATCAACAAAACTTTACACTTATAAA
TGTATTTCCACAGAATTTGACCTTGATCAAATACTATTAGTGATCTAAACTTTCAGAAAT
TTCTGAAAGTTCGTTAGGAAGCATGTGAATATCTCCGCACAAAATCAATCTTTTGTTTAT
CACTTTGGTGAAATGGGCAGTCGTTGGGGCTTTAACCGTACCGTTGGCCAAATGTATGCG
CTTCTTGTTATTACAGAAGAACCACTCAGCGCAAATGATCTGGCCAGTGGCTTAAACATC
```

DEMO

7 shown

211.13 MB

7: SPAdes on data 2 and data 1: log

6: SPAdes on data 2 and data 1: scaffolds (fasta)

7,853 sequences

format: **fasta**, database: ?

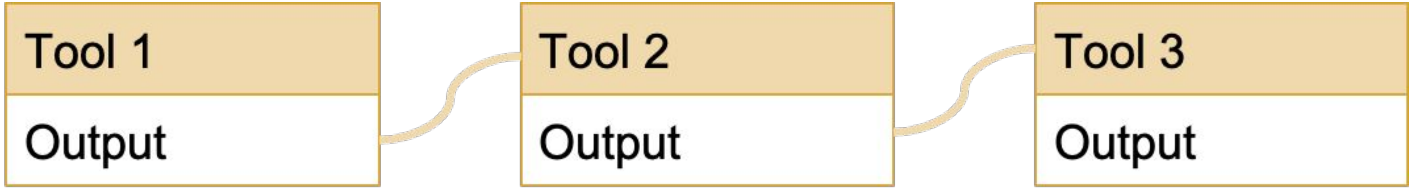
```
Command line: /usr/local/bin/spades.py
-o /data/part0/tmp/jobs/036/36193
/working --disable-gzip-output --careful
-t 10 -m 150 -k 21,33,55 --pe1-fr --pe1-1
fastq.gz:/data/part0
/008/dataset_8735.dat --pe1-2
fastq.gz:/data/part0
/008/dataset_8736.dat

Syste
```

ith IGV local

```
>NODE_1_length_205604_cov_21.023605
TTACTCATAGCGGTAACCTCATGATATTGCGGTTACTATAAGTAT
CACTAAAAGGGAACAGAGCCATATTTAAACAGTACTTTTAATTAC
TAATACAACAATTATTTTTTATTAAGCCGTTTGTGTCACTTGCTC
```

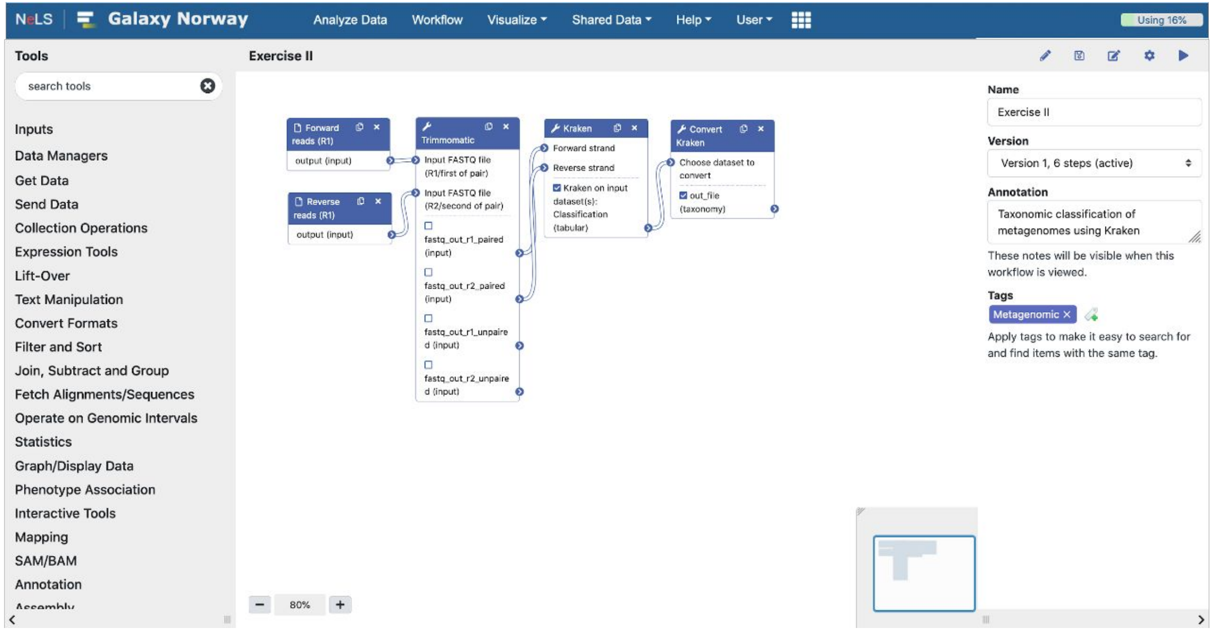
Galaxy workflows

- A workflow in Galaxy is basically a string of tools, where the output from one tool becomes the input for the next



Galaxy workflows

- The “nodules” indicate which output file from one acts as input for the next tool
- Each workflow has a name and version
- Additional text that describe the workflow and tags can be added



Name: Exercise II

Version: Version 1, 6 steps (active)

Annotation: Taxonomic classification of metagenomes using Kraken

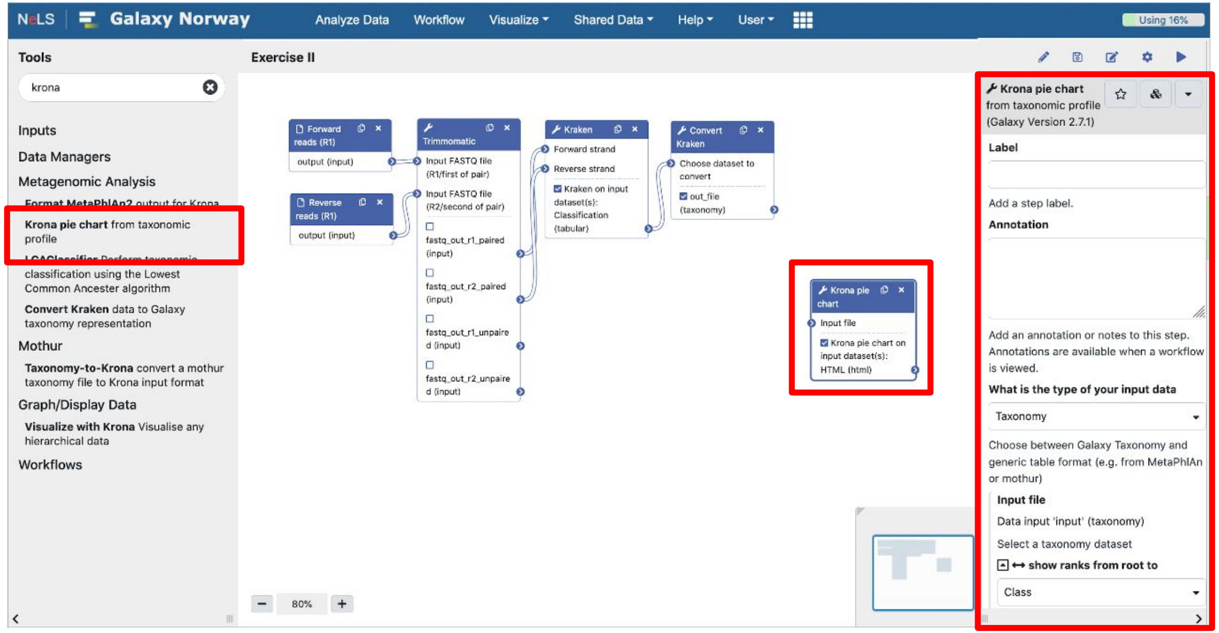
These notes will be visible when this workflow is viewed.

Tags: Metagenomic

Apply tags to make it easy to search for and find items with the same tag.

Galaxy workflows

- New tools can be added by clicking on the tool in the Tool menu
- The tool will appear in the workflow editor
- Tool parameter settings can be pre-set or made up to the user to set when running the workflow


NeLS
Galaxy Norway
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 16%
Tools
Exercise II
Forward reads (R1)
output (input)
Reverse reads (R1)
output (input)
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_r1_paired (input)
fastq_out_r2_paired (input)
fastq_out_r1_unpaired (input)
fastq_out_r2_unpaired (input)
Kraken
Forward strand
Reverse strand
Kraken on input dataset(s): Classification (tabular)
Convert Kraken
Choose dataset to convert
out_file (taxonomy)
Krona pie chart
Input file
Krona pie chart on input dataset(s): HTML (html)
80%

krona

Inputs

Data Managers

Metagenomic Analysis

Format MetaPhlAn2 output for Krona

Krona pie chart from taxonomic profile

LCAClassifier Perform taxonomic classification using the Lowest Common Ancester algorithm

Convert Kraken data to Galaxy taxonomy representation

Mothur

Taxonomy-to-Krona convert a mothur taxonomy file to Krona input format

Graph/Display Data

Visualize with Krona Visualise any hierarchical data

Workflows

Krona pie chart from taxonomic profile (Galaxy Version 2.7.1)

Label

Add a step label.

Annotation

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

What is the type of your input data

Taxonomy

Choose between Galaxy Taxonomy and generic table format (e.g. from MetaPhlAn or mothur)

Input file

Data input 'input' (taxonomy)

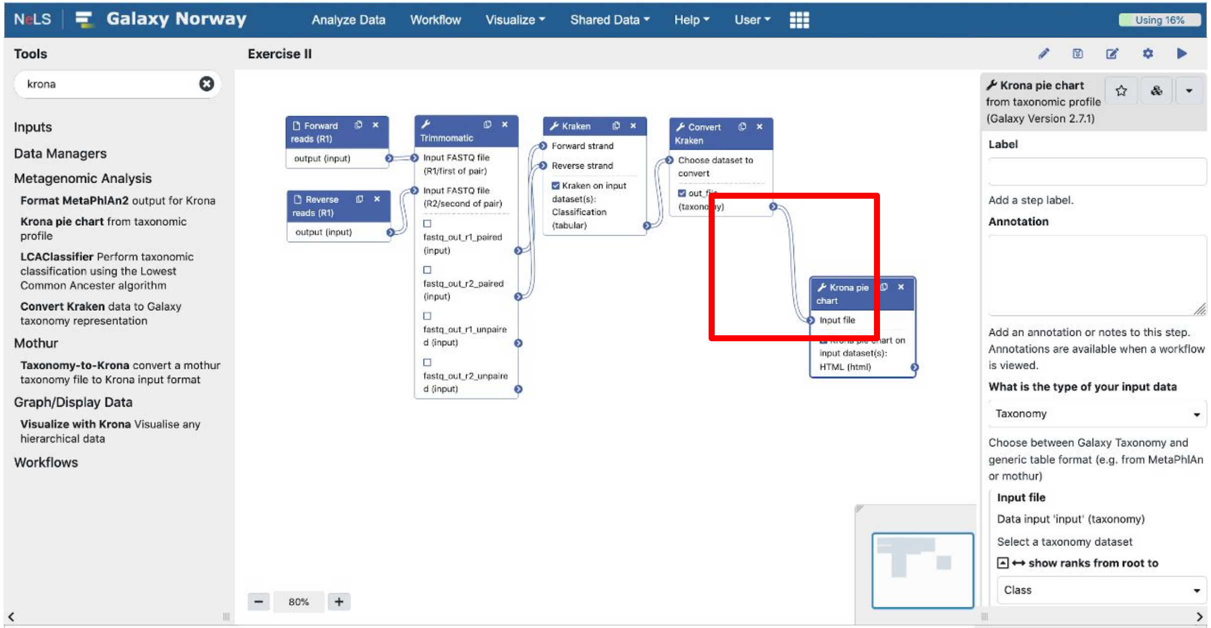
Select a taxonomy dataset

⊡ ↔ **show ranks from root to**

Class

Galaxy workflows

- The output from another tools can be connected as an input for the new tool



NeLS
Galaxy Norway
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 16%
Tools
Exercise II
Forward reads (R1)
output (input)
Reverse reads (R1)
output (input)
Trimmomatic
Input FASTQ file (R1/first of pair)
Input FASTQ file (R2/second of pair)
fastq_out_r1_paired (input)
fastq_out_r2_paired (input)
fastq_out_r1_unpaired (input)
fastq_out_r2_unpaired (input)
Kraken
Forward strand
Reverse strand
Kraken on input dataset(s): Classification (tabular)
Convert Kraken
Choose dataset to convert
out_file (taxonomy)
Krona pie chart
Input file
Krona pie chart on input dataset(s): HTML (html)
80%

krona

Inputs

Data Managers

Metagenomic Analysis

Format MetaPhlAn2 output for Krona

Krona pie chart from taxonomic profile

LCAClassifier Perform taxonomic classification using the Lowest Common Ancester algorithm

Convert Kraken data to Galaxy taxonomy representation

Mothur

Taxonomy-to-Krona convert a mothur taxonomy file to Krona input format

Graph/Display Data

Visualize with Krona Visualise any hierarchical data

Workflows

Krona pie chart from taxonomic profile (Galaxy Version 2.7.1)

Label

Add a step label.

Annotation

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

What is the type of your input data

Taxonomy

Choose between Galaxy Taxonomy and generic table format (e.g. from MetaPhlAn or mothur)

Input file

Data input 'input' (taxonomy)

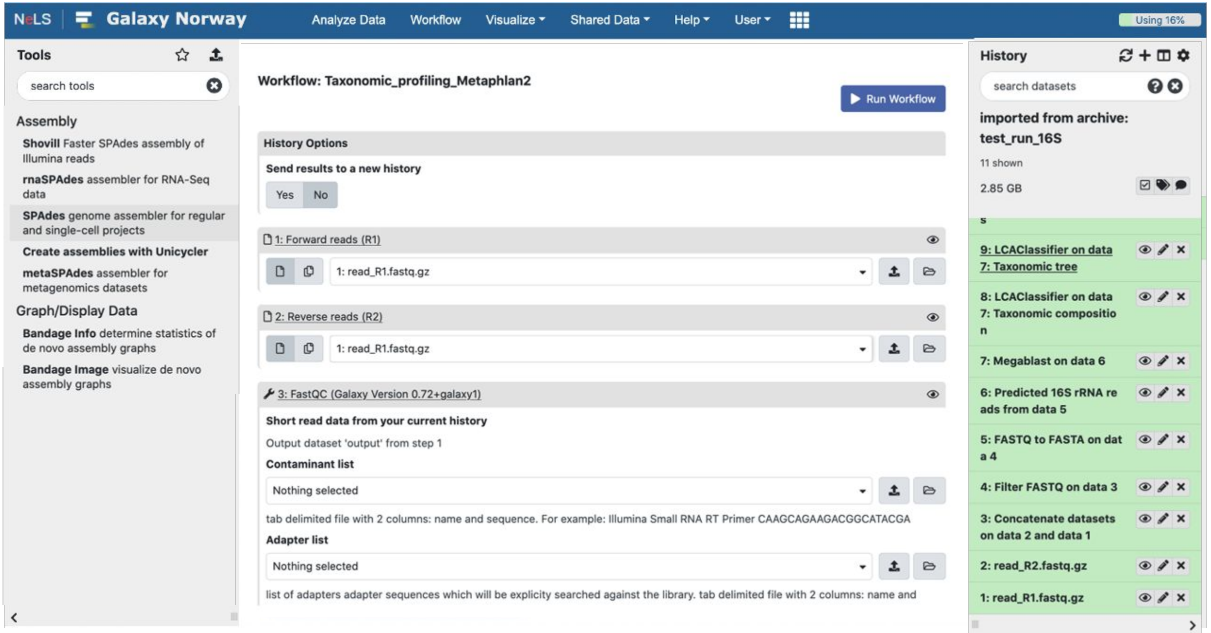
Select a taxonomy dataset

show ranks from root to

Class

Galaxy workflows

- Select input files
- Tools in the workflow will run successively
- You can view and download the result files



Assembly

Shovill Faster SPAdes assembly of Illumina reads

rnaSPAdes assembler for RNA-Seq data

SPAdes genome assembler for regular and single-cell projects

Create assemblies with Unicycler

metaSPAdes assembler for metagenomics datasets

Graph/Display Data

Bandage Info determine statistics of de novo assembly graphs

Bandage Image visualize de novo assembly graphs

Workflow: Taxonomic_profiling_Metaphlan2

History Options

Send results to a new history

Yes No

1: Forward reads (R1)

1: read_R1.fastq.gz

2: Reverse reads (R2)

1: read_R1.fastq.gz

3: FastQC (Galaxy Version 0.72+galaxy1)

Short read data from your current history

Output dataset 'output' from step 1

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

Nothing selected

list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and

imported from archive: test_run_16S

11 shown

2.85 GB

9: LCAClassifier on data 7: Taxonomic tree

8: LCAClassifier on data 7: Taxonomic composition

7: Megablast on data 6

6: Predicted 16S rRNA reads from data 5

5: FASTQ to FASTA on data 4

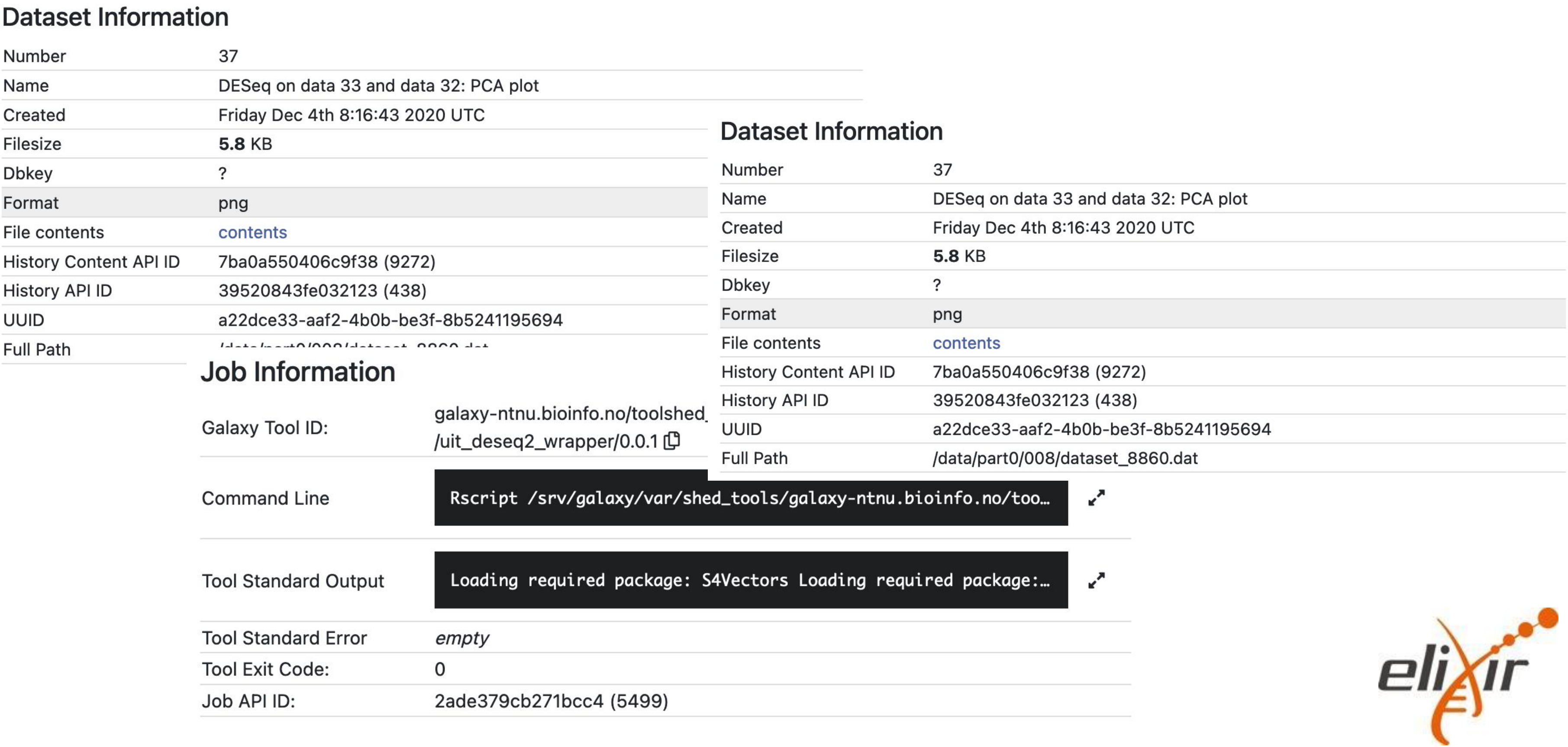
4: Filter FASTQ on data 3

3: Concatenate datasets on data 2 and data 1

2: read_R2.fastq.gz

1: read_R1.fastq.gz

Reproducibility & Transparency



Dataset Information

Number	37
Name	DESeq on data 33 and data 32: PCA plot
Created	Friday Dec 4th 8:16:43 2020 UTC
Filesize	**5.8** KB
Dbkey	?
Format	png
File contents	contents
History Content API ID	7ba0a550406c9f38 (9272)
History API ID	39520843fe032123 (438)
UUID	a22dce33-aaf2-4b0b-be3f-8b5241195694
Full Path	/data/part0/008/dataset_8860.dat

Dataset Information

Number	37
Name	DESeq on data 33 and data 32: PCA plot
Created	Friday Dec 4th 8:16:43 2020 UTC
Filesize	**5.8** KB
Dbkey	?
Format	png
File contents	contents
History Content API ID	7ba0a550406c9f38 (9272)
History API ID	39520843fe032123 (438)
UUID	a22dce33-aaf2-4b0b-be3f-8b5241195694
Full Path	/data/part0/008/dataset_8860.dat

Job Information

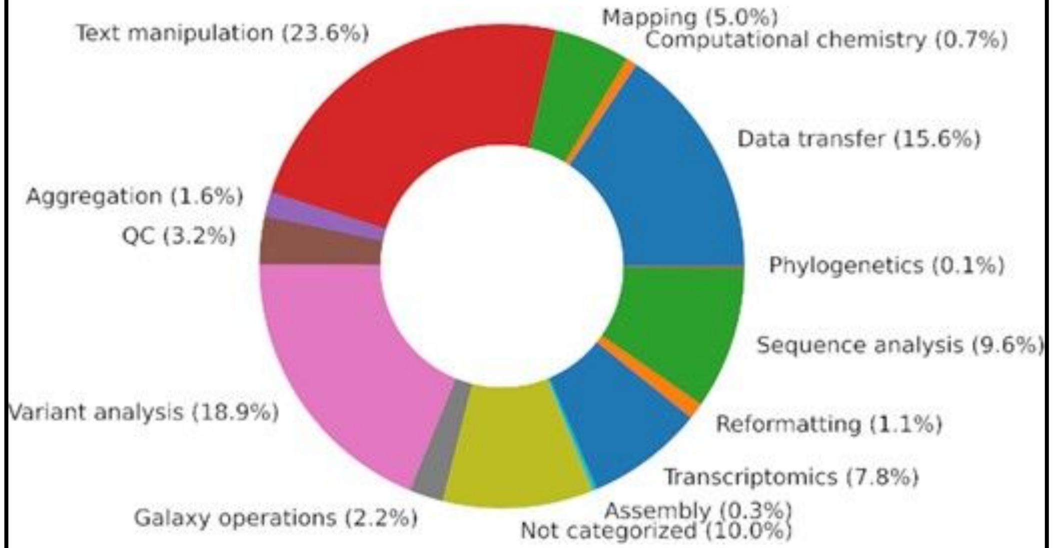
Galaxy Tool ID:	galaxy-ntnu.bioinfo.no/toolshed_ /uit_deseq2_wrapper/0.0.1
Command Line	`Rscript /srv/galaxy/var/shed_tools/galaxy-ntnu.bioinfo.no/too…`
Tool Standard Output	`Loading required package: S4Vectors Loading required package:…`
Tool Standard Error	*empty*
Tool Exit Code:	0
Job API ID:	2ade379cb271bcc4 (5499)

Why use usegalaxy.no?

- Easy to use
- Connected directly with storage (NeLS)
- Backed by significant computational resources
- Training and support

- Transparent research
- Reproducible research
- Enable easy sharing of data and tools

Tool categories executed by usegalaxy users



Nucleic Acids Res, gkac247, https://doi.org/10.1093/nar/gkac247

Quiz questions

Introduction to Galaxy:

1. usegalaxy.no is a web platform for:
   a. Storing life science data
   b. Analysing life science data
   c. Developing bioinformatic tools
   d. Generating data management plans

Useful Resources

- usegalaxy.no
- usegalaxy.eu

Espen Robertsen (UiT)

Workflow Management Systems

✓ Learning Objectives

At the end of this session, you will be able to:

- Conveniently use tonnes of ready-made Nextflow pipelines to analyse your data in a FAIR manner
- Utilize auxiliary resources such as Biocontainers, Workflowhub and nf-core to find containers and workflows enabling FAIR analyses

What are workflow management systems (WfMS)?

- “Standards for describing computational data-analysis workflows”
- Streamline routine processes for optimal efficiency
- Handles input / output / execution of tools in a DAG-like manner
- Comes in various colors and flavours

nextflow

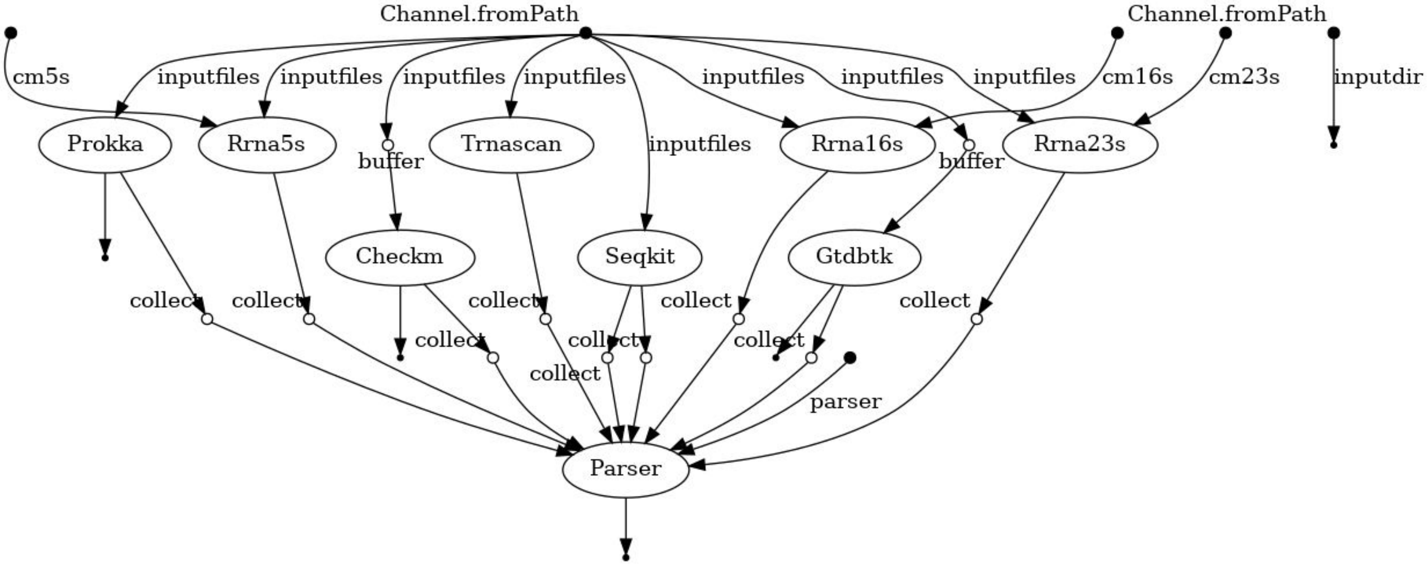
WorkflowHub

COMMON
WORKFLOW
LANGUAGE

snakemake

What are workflow management systems (WfMS)?

- Directed acyclic graphs (DAG) can be used to describe workflows visually



("Autoget" nextflow backend processing pipeline @ UiT)

WfMS’ enhances the principles of “FAIR”

- WfMS’s provide an analysis description in their own right
- Enables portability and reproducibility in different computational environments
- Versioning and container integration ensures uniform data analysis

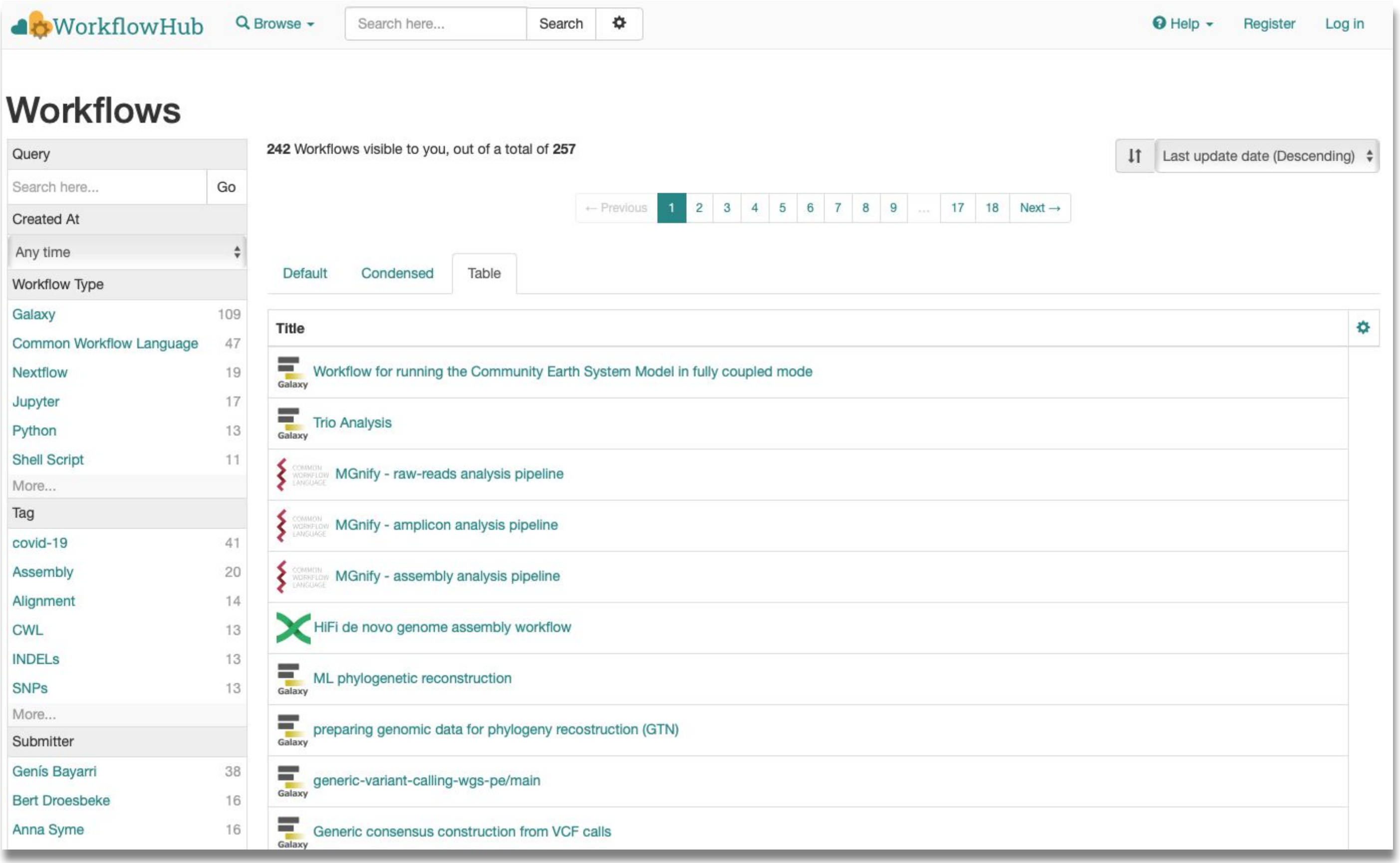
Most WfMS integrates with containers, which makes Biocontainers.pro very useful!

- 10.6K tools, 45.5K versions, 225.3K containers and packages
- Dockerfile recipes / Conda recipes to automatically build containers in BioContainers.
- Biocontainers Registry is a hosted registry of all BioContainers images that are ready to be used

BioContainers Flow
SOFTWARE
</>
CONTAINER
WORKFLOW

BIOCONDA®

Workflowhub covers you workflow needs and then some!



Workflows

242 Workflows visible to you, out of a total of 257

Last update date (Descending)

Query	
Search here...	Go
Created At	
Any time	

Workflow Type	
Galaxy	109
Common Workflow Language	47
Nextflow	19
Jupyter	17
Python	13
Shell Script	11
More...	

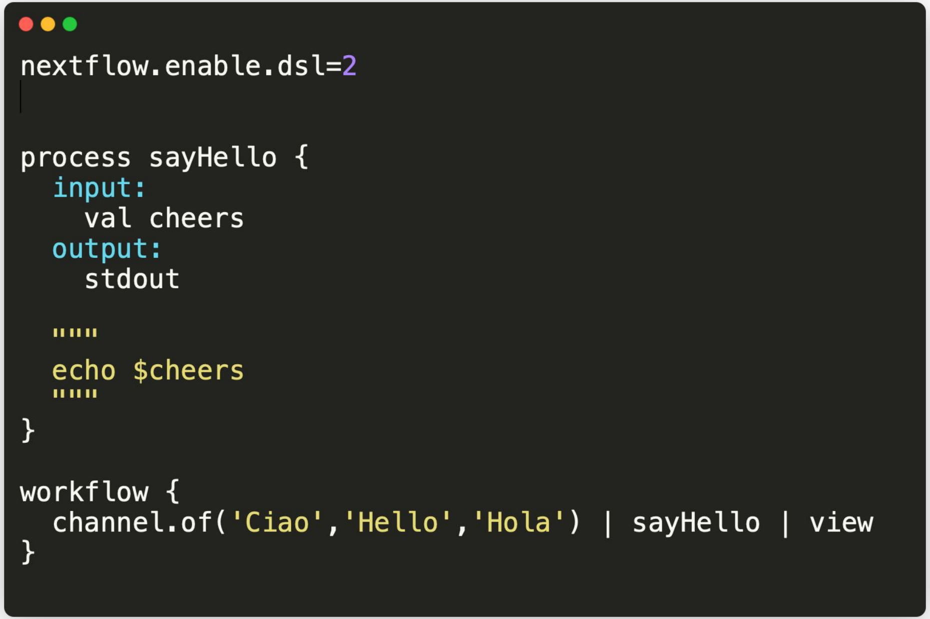
Tag	
covid-19	41
Assembly	20
Alignment	14
CWL	13
INDELs	13
SNPs	13
More...	

Submitter	
Genís Bayarri	38
Bert Droesbeke	16
Anna Syme	16

Title
Workflow for running the Community Earth System Model in fully coupled mode
Trio Analysis
MGnify - raw-reads analysis pipeline
MGnify - amplicon analysis pipeline
MGnify - assembly analysis pipeline
HiFi de novo genome assembly workflow
ML phylogenetic reconstruction
preparing genomic data for phylogeny recostruction (GTN)
generic-variant-calling-wgs-pe/main
Generic consensus construction from VCF calls

Nextflow.io

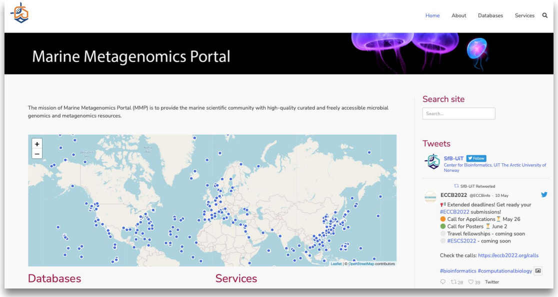
- Supports Bash, Python, R …
- Dataflow handled by Groovy (Java super-set, “python for Java”)
- Reproducibility
- Lots of “premium” features for free
- https://www.nextflow.io/docs/latest/



```
nextflow.enable.dsl=2


process sayHello {
  input:
    val cheers
  output:
    stdout

  """
  echo $cheers
  """
}

workflow {
  channel.of('Ciao','Hello','Hola') | sayHello | view
}
```

Some examples of Nextflow usage in ELIXIR

- Mar-databases backend processing
  - Results integrated with and available at portal pages
  - https://gitlab.com/uit-sfb/autoget-nextflow
- FHI-pipeline utilizing TSD
  - Desensitization pipeline for Norwegian covid19-samples affected by GDPR
  - https://gitlab.com/uit-sfb/fhi-desensitize
- Metapipe
  - Metagenomic analysis pipeline developed at UiT (ELIXIR)
  - https://gitlab.com/uit-sfb/metapipe
  - https://mmp2.sfb.uit.no/metapipe/



Home About Databases Services

Marine Metagenomics Portal

The mission of Marine Metagenomics Portal (MMP) is to provide the marine scientific community with high-quality curated and freely accessible microbial genomics and metagenomics resources.

Leaflet | © OpenStreetMap contributors

Databases

Services

Search site

Search...

Tweets

SfB-UiT Follow
Center for Bioinformatics, UiT The Arctic University of Norway

SfB-UiT Retweeted

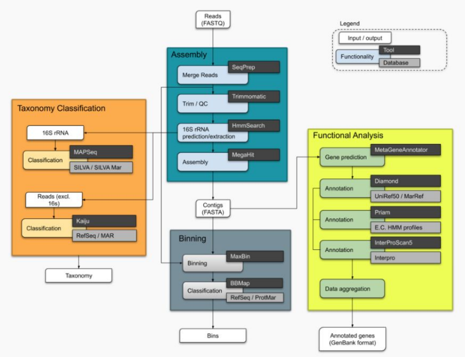
ECCB2022 @ECCBinfo · 10 May

📢 Extended deadlines! Get ready your #ECCB2022 submissions!
🟠 Call for Applications ⏳ May 26
🟢 Call for Posters ⏳ June 2
⚪ Travel fellowships - coming soon
⚪ #ESCS2022 - coming soon

Check the calls: https://eccb2022.org/calls

#bioinformatics #computationalbiology

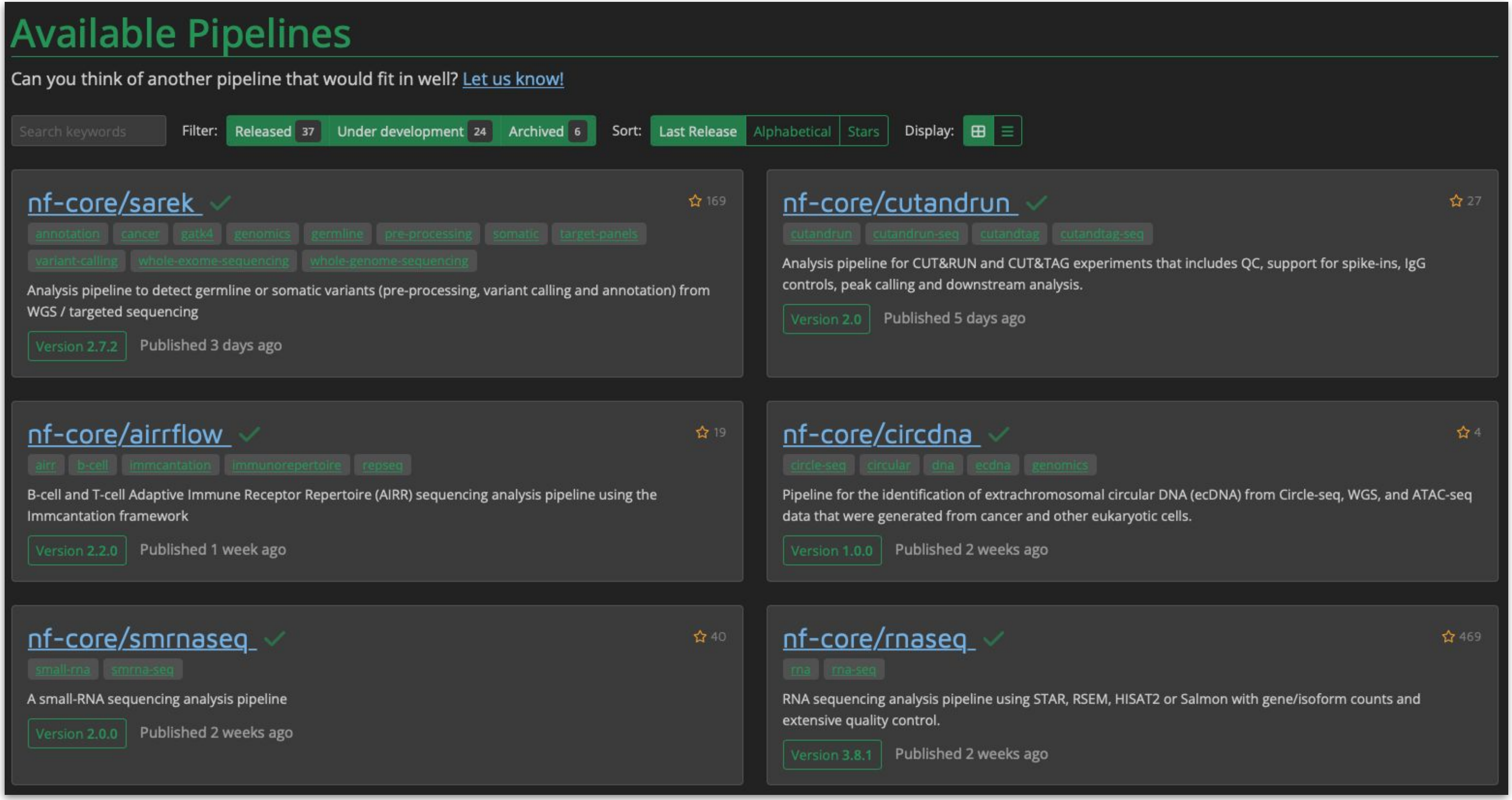
28 39 Twitter



nf-core is a great resource!

- Out-of-the-box high quality, production-ready, curated analysis pipelines built using Nextflow. (66 as of Jun, 2022)
- Maintained and validated releases ensure reproducibility.
- nf-core-tools for convenience (list, launch, create templates etc.)
- Excellent documentation!

nf-core is a great resource!



Available Pipelines

Can you think of another pipeline that would fit in well? Let us know!

Search keywords
Filter: Released 37 | Under development 24 | Archived 6
Sort: Last Release | Alphabetical | Stars
Display:

nf-core/sarek ✓

☆ 169

annotation · cancer · gatk4 · genomics · germline · pre-processing · somatic · target-panels · variant-calling · whole-exome-sequencing · whole-genome-sequencing

Analysis pipeline to detect germline or somatic variants (pre-processing, variant calling and annotation) from WGS / targeted sequencing

Version 2.7.2 Published 3 days ago

nf-core/cutandrun ✓

☆ 27

cutandrun · cutandrun-seq · cutandtag · cutandtag-seq

Analysis pipeline for CUT&RUN and CUT&TAG experiments that includes QC, support for spike-ins, IgG controls, peak calling and downstream analysis.

Version 2.0 Published 5 days ago

nf-core/airrflow ✓

☆ 19

airr · b-cell · immcantation · immunorepertoire · repseq

B-cell and T-cell Adaptive Immune Receptor Repertoire (AIRR) sequencing analysis pipeline using the Immcantation framework

Version 2.2.0 Published 1 week ago

nf-core/circdna ✓

☆ 4

circle-seq · circular · dna · ecdna · genomics

Pipeline for the identification of extrachromosomal circular DNA (ecDNA) from Circle-seq, WGS, and ATAC-seq data that were generated from cancer and other eukaryotic cells.

Version 1.0.0 Published 2 weeks ago

nf-core/smrnaseq ✓

☆ 40

small-rna · smrna-seq

A small-RNA sequencing analysis pipeline

Version 2.0.0 Published 2 weeks ago

nf-core/rnaseq ✓

☆ 469

rna · rna-seq

RNA sequencing analysis pipeline using STAR, RSEM, HISAT2 or Salmon with gene/isoform counts and extensive quality control.

Version 3.8.1 Published 2 weeks ago

Nextflow tower

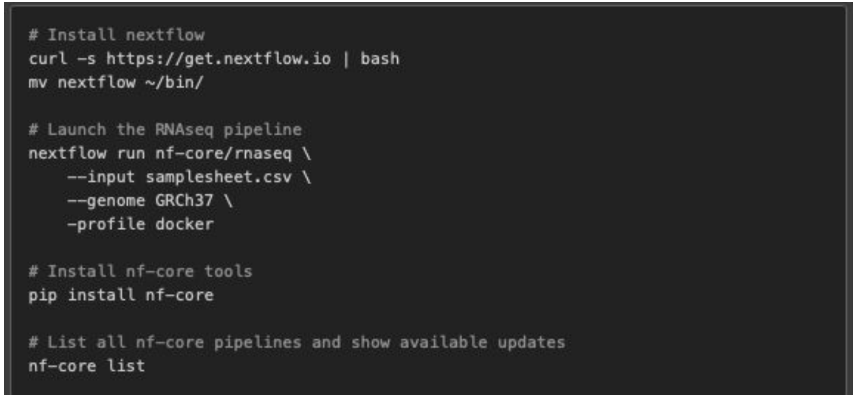
- “Centralized command post”; monitoring, logging & observability of distributed workflows conveniently available in your browser
- Simplifies the deployment of pipelines on any cloud, cluster or laptop.
- Quite expensive, but some essential features are free

Enough talk! Let’s play around with nextflow!

If we have time :)

- Quick tour / usage of nf-tower
- Launch nextflow pipeline -with-tower

https://gitlab.com/uit-sfb/autoget-nextflow



```
# Install nextflow
curl -s https://get.nextflow.io | bash
mv nextflow ~/bin/

# Launch the RNAseq pipeline
nextflow run nf-core/rnaseq \
    --input samplesheet.csv \
    --genome GRCh37 \
    -profile docker

# Install nf-core tools
pip install nf-core

# List all nf-core pipelines and show available updates
nf-core list
```

Useful Resources

- Nextflow, https://nextflow.io
- nf-core, https://nf-co.re
- nf-tower, https://tower.nf
- Biocontainers, https://biocontainers.pro
- Marine Metagenomics Portal, https://mmp2.sfb.uit.no
- Metapipe, https://mmp2.sfb.uit.no/metapipe/

Thank you!

🌐 elixir-norway.org

🐦 @elixirnorway

✉ contact@bioinfo.no

elixir NORWAY