

Introduction to Galaxy

INF-BIO5121/INF-BIO9121

October 7. 2015, Oslo

Sveinung Gundersen

Every baby
knows the

scientific method!



- We are doing science, also on the computer!
- 4-5-6 is typically done on the computer anyway
- But the methods/software used in bioinformatics often give very varied results
- We should really think of computer analysis as part of the experiment, aiming for the same level of rigor and reproducibility!

My claim:

Bioinformaticians

(esp. those with a *biology background*)

are too fond of the command line!

The command-line approach to bioinformatics

- We want to run a tool, say Bowtie
- Try: “bowtie”
- “module load bowtie”
- Try: “bowtie” (Yes, it’s there)
- What were the options?
- “bowtie -h”

bowtie -h

Usage:

```
bowtie [options]* <ebwt> {-l <m1> -2 <m2> | --l2 <r> | <s>} [<hit>]
```

<m1> Comma-separated list of files containing upstream mates (or the sequences themselves, if -c is set) paired with mates in <m2>

<m2> Comma-separated list of files containing downstream mates (or the sequences themselves if -c is set) paired with mates in <m1>

<r> Comma-separated list of files containing Crossbow-style reads. Can be a mixture of paired and unpaired. Specify "-" for stdin.

<s> Comma-separated list of files containing unpaired reads, or the sequences themselves, if -c is set. Specify "-" for stdin.

<hit> File to write hits to (default: stdout)

Input:

-q query input files are FASTQ .fq/.fastq (default)

-f query input files are (multi-)FASTA .fa/.mfa

-r query input files are raw one-sequence-per-line

-c query sequences given on cmd line (as <mates>, <singles>)

-C reads and index are in colorspace

-Q/--quals <file> QV file(s) corresponding to CSFASTA inputs; use with -f -C

--Q1/--Q2 <file> same as -Q, but for mate files 1 and 2 respectively

-s/--skip <int> skip the first <int> reads/pairs in the input

-u/--qupto <int> stop after first <int> reads/pairs (excl. skipped reads)

-5/--trim5 <int> trim <int> bases from 5' (left) end of reads

-3/--trim3 <int> trim <int> bases from 3' (right) end of reads

--phred33-quals input quals are Phred+33 (default)

--phred64-quals input quals are Phred+64 (same as --solexa1.3-quals)

--solexa-quals input quals are from GA Pipeline ver. < 1.3

--solexa1.3-quals input quals are from GA Pipeline ver. >= 1.3

--integer-quals qualities are given as space-separated integers (not ASCII)

bowtie -h

Alignment:

- v <int> report end-to-end hits w/ <=v mismatches; ignore qualities
or
- n/--seedmms <int> max mismatches in seed (can be 0-3, default: -n 2)
- e/--maqerr <int> max sum of mismatch quals across alignment for -n (def: 70)
- l/--seedlen <int> seed length for -n (default: 28)
- nomaqround disable Maq-like quality rounding for -n (nearest 10 <= 30)
- I/--minins <int> minimum insert size for paired-end alignment (default: 0)
- X/--maxins <int> maximum insert size for paired-end alignment (default: 250)
- fr/--rf/--ff -l, -2 mates align fw/rev, rev/fw, fw/fw (default: --fr)
- nofw/--norc do not align to forward/reverse-complement reference strand
- maxbts <int> max # backtracks for -n 2/3 (default: 125,800 for --best)
- pairtries <int> max # attempts to find mate for anchor hit (default: 100)
- y/--tryhard try hard to find valid alignments, at the expense of speed
- chunkmbs <int> max megabytes of RAM for best-first search frames (def: 64)

Reporting:

- k <int> report up to <int> good alignments per read (default: 1)
- a/--all report all alignments per read (much slower than low -k)
- m <int> suppress all alignments if > <int> exist (def: no limit)
- M <int> like -m, but reports 1 random hit (MAPQ=0); requires --best
- best hits guaranteed best stratum; ties broken by quality
- strata hits in sub-optimal strata aren't reported (requires --best)

Output:

- t/--time print wall-clock time taken by search phases
- B/--offbase <int> leftmost ref offset = <int> in bowtie output (default: 0)
- quiet print nothing but the alignments
- refout write alignments to files refXXXXXX.map, 1 map per reference
- refidx refer to ref. seqs by 0-based index rather than name

bowtie -h

--al <fname> write aligned reads/pairs to file(s) <fname>
--un <fname> write unaligned reads/pairs to file(s) <fname>
--max <fname> write reads/pairs over -m limit to file(s) <fname>
--suppress <cols> suppresses given columns (comma-delim'ed) in default output
--fullref write entire ref name (default: only up to 1st space)

Colorspace:

--snpphred <int> Phred penalty for SNP when decoding colorspace (def: 30)
or
--snpfrac <dec> approx. fraction of SNP bases (e.g. 0.001); sets --snpphred
--col-cseq print aligned colorspace seqs as colors, not decoded bases
--col-cqual print original colorspace quals, not decoded quals
--col-keepends keep nucleotides at extreme ends of decoded alignment

SAM:

-S/--sam write hits in SAM format
--mapq <int> default mapping quality (MAPQ) to print for SAM alignments
--sam-nohead suppress header lines (starting with @) for SAM output
--sam-nosq suppress @SQ header lines for SAM output
--sam-RG <text> add <text> (usually "lab=value") to @RG line of SAM header

Performance:

-o/--offrate <int> override offrate of index; must be >= index's offrate
-p/--threads <int> number of alignment threads to launch (default: 1)
--mm use memory-mapped I/O for index; many 'bowtie's can share
--shmem use shared mem for index; many 'bowtie's can share

Other:

--seed <int> seed for random number generator
--verbose verbose output (for debugging)
--version print version information and quit
-h/--help print this usage message

At last....

- Call “bowtie /path/input.fastq ... (a bunch and of options and some some, and even more options)... > /path/to/bowtieLog.txt 2>&1 &”
- We get back to it next morning

Now isn't this good enough ?!

Log in to server.
Profile OK?

Confusing and
error-prone

- Call “bowtie /path/input.fastq ... (a bunch and of options and some some, and even more options)... > /path/to/bowtieLog.txt 2>&1 &”
- We get back to it next morning

How to keep this running when I log off? nohup? screen?

Will I remember?
Will it be ready then?

Where did I log to this time?

How was this again?

But I wanted to run it on the cluster!!

- How were those SLURM things again?...

Galaxy

- Developed at Penn State and Emory Universities, for over 10 years by a large development team
- Aims to be a framework for “supporting
 - Accessible
 - Reproducible
 - Transparent
- computational research in the life sciences” (*Goecks et. al., Genome Biology 2010*)

Accessible

- Users do not need to learn the command line
- Web-based solution, point-and-click
- Consistent look and feel
- Easy to upload your own datasets, or import datasets from established data warehouses

Reproducible

- Bioinformaticians gets surprised every time they need to redo/modify previous analyses
- But lab biologists already know the importance of reproducibility!
- They also know that even with a detailed lab journal, reproduction is a challenge
- The question is then how this manifests itself when doing analysis on a computer

What is *in silico* reproducibility?

- Basically the same issues as in the lab:
 - Materials -> Data sources
 - Experiment conditions -> Analysis parameters
 - Equipment (with model number) -> Programs (with version number)
- And the same challenges:
 - Are all relevant conditions described accurately?
 - Will the same materials and equipment be available?

What is the status of reproducibility in the literature?

- Less than half of selected microarray experiments published in Nature Genetics could be reproduced (*Ioannidis et al., Nat Genet 2009*)
- More than half [of surveyed papers] do not provide primary data and list neither the version nor the parameters used [for read mapping]
(*Nekrutenko and Taylor., Nat Rev Genet 2012*)

Why should you care?

(about making your analyses reproducible)

- Because it's the right thing to do!
- Journals are becoming aware of the issues
- Reviewers will value it
- But the main argument:
 - The one that's struggling with reproducing your analysis is often the future you

Galaxy supports reproducibility

- Automatically tracks *metadata* at every step
 - Which are the datasets?
 - What are the parameters?
 - Which tools, and which version of the tool?
 - What are the outputs
- Users can annotate the steps to capture the *intent* of the analysis!

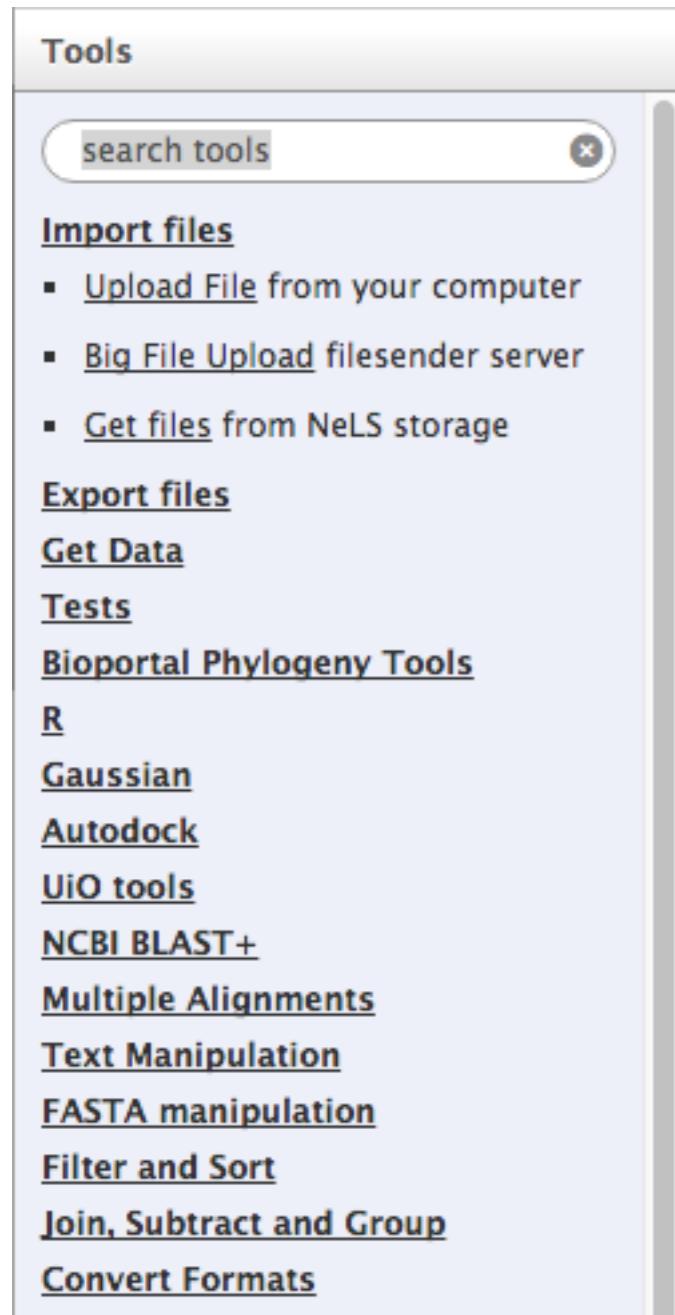
Galaxy supports reproducibility

- All jobs can be rerun later, by independent scientists
- Galaxy automatically captures data inputs, analysis parameters, program versions, and intermediate data
- Workflows capture common analysis sequences, *i.e.* typical experimental setups. Can be reused for other datasets and experiments

Transparent

- “Enabling users to share and communicate their experimental results and outputs in a meaningful way” (Goecks et. al., *Genome Biology* 2010)
- Everything can be shared: Datasets, histories (i.e. experimental logbook), tools, workflows
- Provides public repositories
- Galaxy Pages are web-based documents for publishing results. Every level of detail can be accessed by readers

Basic concepts: tools



- Typically command-line programs or scripts wrapped as web tools
- Hundreds of tools available

Basic concepts: history and datasets

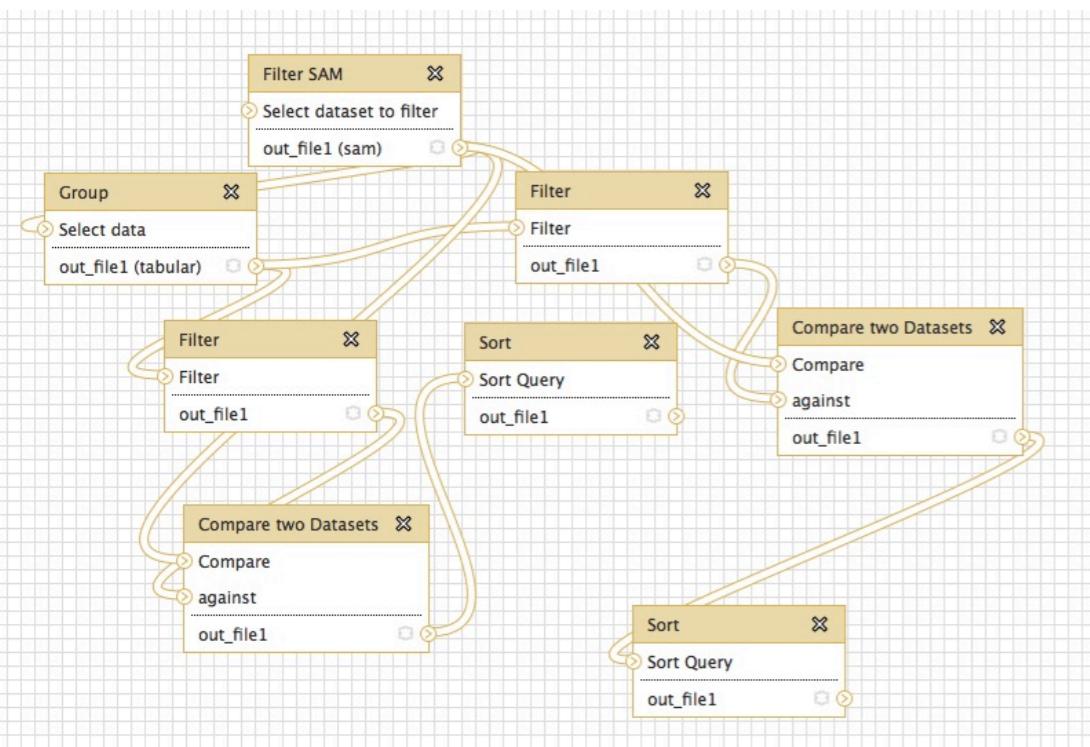
History	
INFBIOnx121-miRNA-full	27.2 GB
73: Send files on data 47	0 X
72: NKTCL 31B control mini 1	0 X
71: NKTCL 31T tumor mini 2	0 X
70: Chow count proper	0 X
69: Send files on data 47 and data 56	0 X
56: R output2	0 X
55: R output1	0 X
54: R out file	0 X
47: Chow count proper	0 X

- The history is a log of your analysis
- Contains all:
 - Input datasets
 - Intermediate datasets
 - Analysis output
 - (which in Galaxy terminology also are datasets)
- Every tool takes input from the history and generates output to the history
- One may create a history for each investigation
- Histories may be shared with others, privately or publicly

Basic concepts: datatype

- All datasets (history elements) have a datatype
- This is basically a text string specifying what format the data is in
- Examples:
 - “fastq”, “fastqsanger”, “fastqillumina”, “bed”...
 - All tools take datasets of specific datatypes as input, and produces datasets of (other) specific datatypes
 - Changing datatype does not change the data!

Basic concepts: workflow



- Tools can be arranged into workflows
- Output and input data are routed with lines
- Other parameters of the tools may be specified
- Conceptual representation (without data)
- Workflows can be shared



- Galaxy installation at UiO, running on the Abel cluster
- Contains hundreds of tools, from Phylogeny tools to High Throughput Sequencing analysis
- Available for all FEIDE users (all university users and several colleges)
- Other users may apply for access

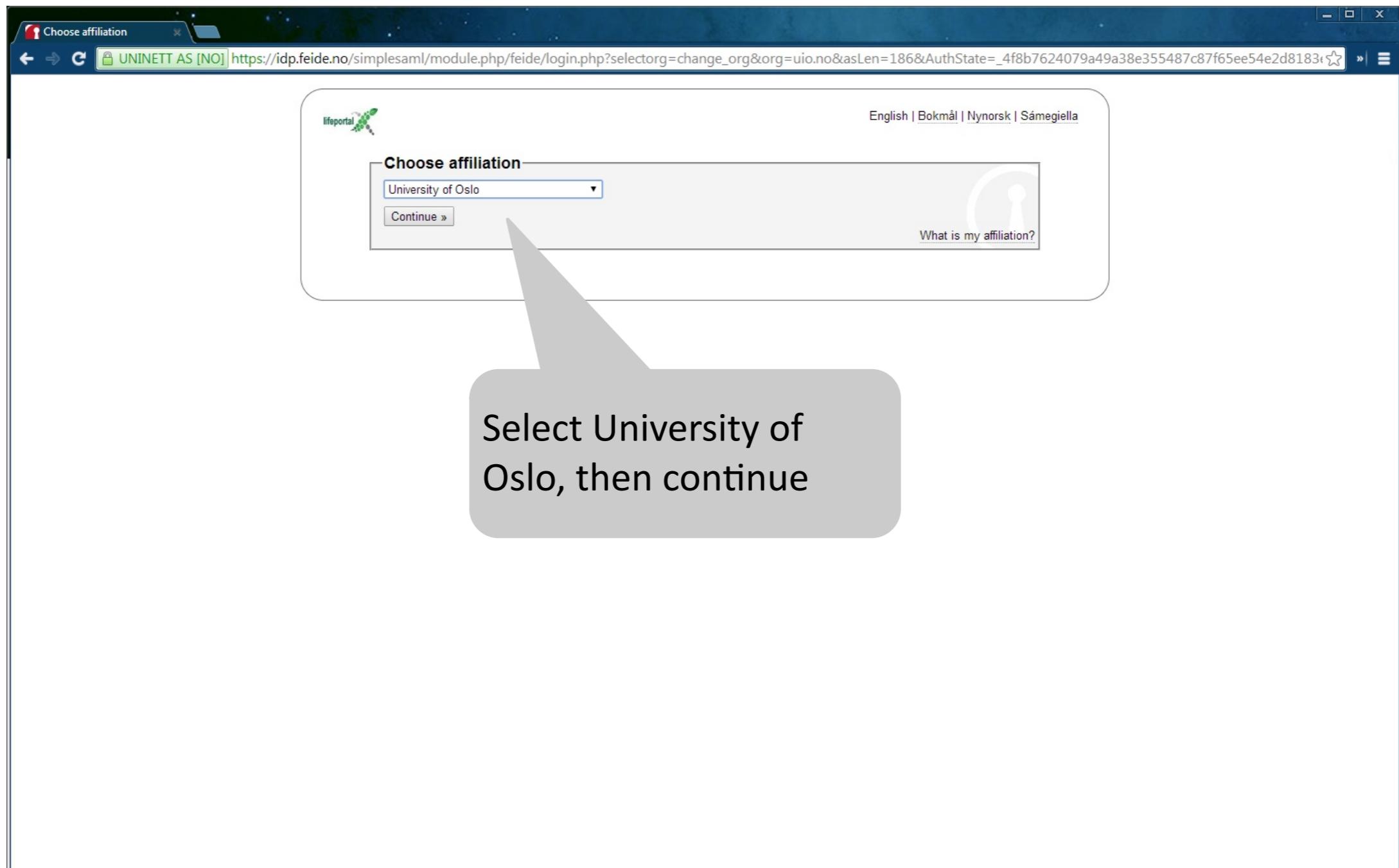
Demo: lifeportal.uio.no

The screenshot shows the Lifeportal homepage in a web browser. The URL in the address bar is <https://lifeportal.uio.no/root>. The page has a dark blue header with the University of Oslo logo and navigation links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. Below the header, there's a main content area with sections for Academic login, About the Lifeportal, and Getting access. A large grey callout bubble in the bottom-left corner contains the text "Select Feide login, press Academic Login". At the bottom of the page, there's a Partners section with logos for UiO, Norwegian Bioinformatics Platform, MLSUiO, UNINETT sigma, and Galaxy.

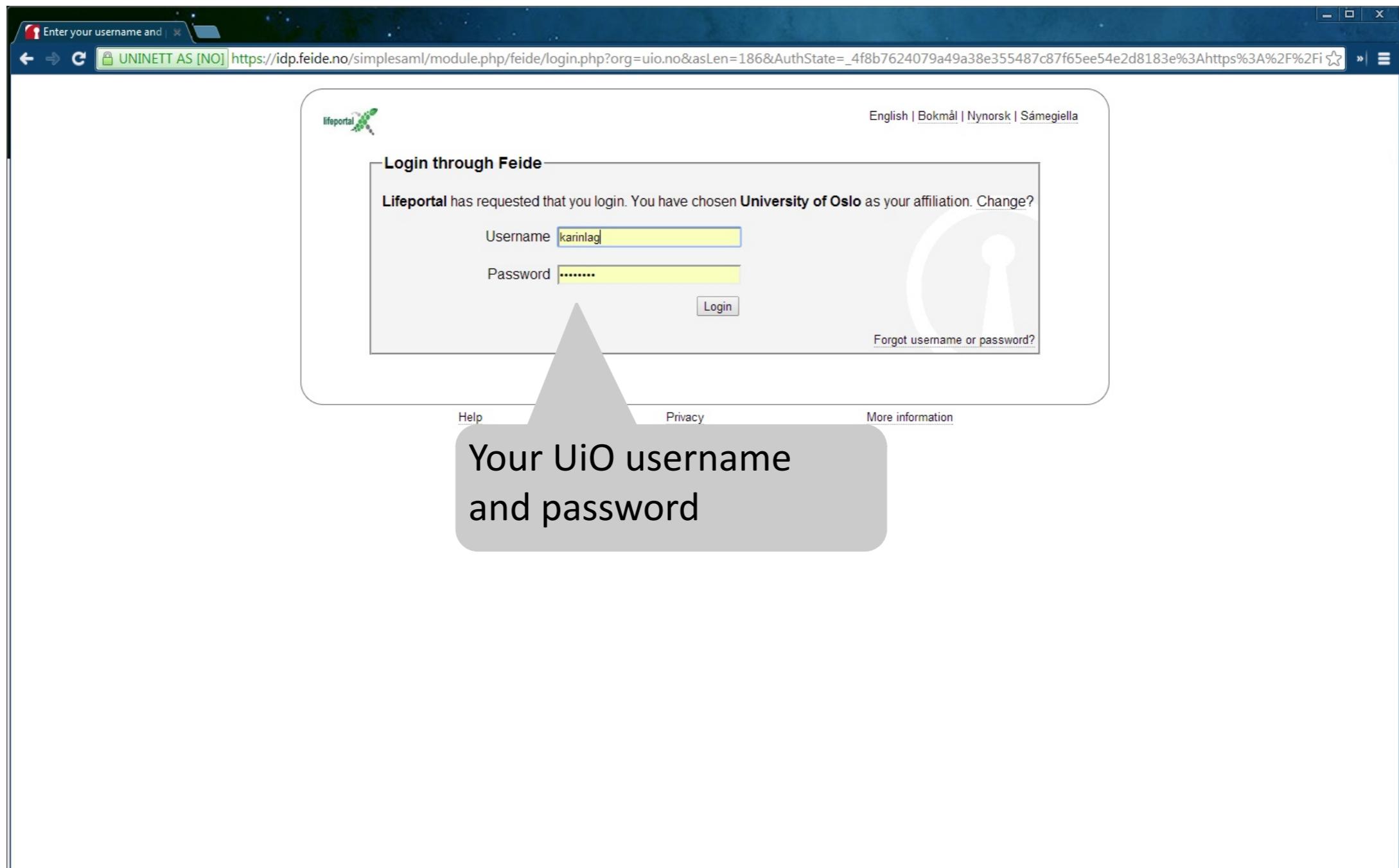
Select Feide login,
press Academic Login

Copied from presentation by Karin Lagesen

Select your institution



Use UiO username/password



Verify login information

The screenshot shows the Lifeportal web interface at <https://lifeportal.uio.no>. The page title is "Lifeportal". The top navigation bar includes "Analyze Data", "Workflow", "Shared Data", "Visualization", "Apply for a project", "Help", and "User". A dropdown menu under "User" shows the logged-in email address "karin.lagesen@medisin.uio.no" and options for "Logout", "Saved Histories", "Saved Datasets", "Saved Pages", and "API Keys". A large gray callout bubble points to the "User" menu with the text: "Click User, verify that your email address is shown". On the left, a sidebar lists various tools such as "Upload file", "Tests", "Bioportal Phylogeny Tools", "R", "Autodock", "UiO tools", "Gaussian", "NCBI BLAST+", "Text Manipulation", "Filter and Sort", "Join, Subtract and Group", "Convert Formats", "Extract Features", "DeFuse", "Bismark", "Fetch Sequences", "Operate on Genomic Intervals", "Metagenomic analyses", "Statistics", "BEDTools", "Graph/Display Data", "FASTA manipulation", "NGS: GATK2", "NGS: Picard (beta)", "NGS: Trim Galore", "NGS: Assembly", "Multiple Alignments", "NGS: QC and manipulation", "NGS: Mapping", "NGS: RNA Analysis", "NGS: SAM Tools", and "NGS: GATK Tools (beta)". The main content area includes sections for "About the Lifeportal" and "Partners".

Click User, verify that
your email address
is shown

Page orientation

The screenshot shows the Lifeportal web interface with the following components:

- Navigation bar:** Located at the top, featuring links for Analyze Data, Workflow, Shared Data, Visualization, Apply for a project, Help, and User.
- Tool panel:** On the left side, titled "Bioportal Phylogeny Tools", listing various analysis programs such as R, Autodock, and NCBI BLAST+.
- Detail panel:** The central area containing the main content, including a "Partners" section with logos for UiO, Norwegian Bioinformatics Platform, MLS, UNINETT sigma, and Galaxy.
- History panel:** On the right side, showing an empty history with the message "Your history is empty. Click 'Get Data' on the left pane to start".

Annotations with callout bubbles explain the purpose of each panel:

- The Navigation bar is annotated with the text "Navigation bar, with workflows, shared data etc."
- The Tool panel is annotated with the text "Tool panel with many analysis programs".
- The Detail panel is annotated with the text "Detail panel – where the results are shown".
- The History panel is annotated with the text "History panel- shows all the datasets you have analyzed and produced".

Create a new history

The screenshot shows the Lifeportal web interface at <https://lifeportal.uio.no/root>. The interface includes a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, Apply for a project, Help, and User. A message at the top right indicates "Using 3.6 GB". The main content area features a sidebar titled "Tools" listing various bioinformatics tools, and a central panel with sections for "About the Lifeportal" and "Partners". On the right, a "History" panel displays an "Unnamed history" entry with 0 bytes, accompanied by a note: "Your history is empty. Click 'Get Data' on the left pane to start". A large gray callout bubble points from the text in the "History" panel to the following quote.

When starting on a "new" thing, start with a clean history, and name it!

Getting data: uploading

Lifeportal

UiO University of Oslo Analyze Data Workflow Shared Data Visualization Apply for a project Help User Using 16.3 GB

Tools

Upload file

- Upload File from your computer
- Big File Upload from under server

Tests

Bioportal Phylogeny R Autodownload Uio tools Gauss NCBI Text Miner Filter Join, S Convert Formats Extract Features DeFuse Bismark Fetch Sequences Operate on Genomic Intervals Meta genomic analyses Statistics BEDTools Graph/Display Data FASTA manipulation NGS: GATK2 NGS: Picard (beta) NGS: Trim Galore NGS: Assembly Multiple Alignments NGS: OC and manipulation

Upload File (version 1.1.3)

File Format:

- Auto-detect
- fastq
- fastqcssanger
- fastqillumina
- fastqsanger

g files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by your administrator).

Genome: ----- Additional Species Are Below -----

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will automatically be decompressed.

Ab1

A binary sequence file in 'ab1' format with a '.ab1' file extension. You must manually select this 'File Format' when uploading the file.

Axt

blastz pairwise alignment format. Each alignment block in an axt file contains three lines: a summary line and 2 sequence lines. Blocks are separated from

History

Example history 0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Click on Upload File, then Upload File again

Select fastqsanger as sequence format

Uploading data

The screenshot shows the Lifeportal web interface at <https://lifeportal.uio.no/root>. The main title is "Lifeportal". On the left, there's a sidebar with "Tools" and a list of bioinformatics tools. A red circle highlights the "Execute" button in the central "Upload File (version 1.1.3)" form. A callout bubble points to the "Choose File" input field with the text "Select input file here". The "File Format" dropdown is set to "fastqsanger". The "URL/Text" section contains a placeholder "ERR101899_1.fastq.gz". The "History" panel shows an "Unnamed history" entry with 0 bytes.

Tools

- search tools
- Upload file
 - Upload File from your computer
 - Big File Upload filesender server
- Tests
 - Bioportal Phylogeny Tools
 - R
 - Autodock
 - UiO tools
 - Gaussian
 - NCBI BLAST+
 - Text Manipulation
 - Filter and Sort
 - Join, Subtract and Group
 - Convert Formats
 - Extract Features
 - DeFuse
 - Bismark
 - Fetch Sequences
 - Operate on Genomic Intervals
 - Meta genomic analyses
 - Statistics
 - BEDTools
 - Graph/Display Data
 - FASTA manipulation
 - NGS: GATK2
 - NGS: Picard (beta)
 - NGS: Trim Galore
 - NGS: Assembly
 - Multiple Alignments
 - NGS: OC and manipulation

Upload File (version 1.1.3)

File Format:

fastqsanger

Which format? See help below

File:

Choose File ERR101899_1.fastq.gz

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Select input file here

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Convert spaces to tabs:

Yes

Use this option if you are entering intervals by hand.

Genome:

----- Additional Species Are Below -----

Execute

Auto-detect

The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your file is not detected properly as one of the known formats, it most likely means that it has some format problems (e.g., different number of columns on different rows). You can still coerce the system to set your data to the format you think it should be. You can also upload compressed files, which will automatically be decompressed.

Ab1

A binary sequence file in 'ab1' format with a '.ab1' file extension. You must manually select this 'File Format' when uploading the file.

Axt

alignment block in an axt file contains three lines: a summary line and 2 sequence lines. Blocks are separated from

History

Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

Using 79.0 GB

Uploaded data

The screenshot shows the Lifeportal web interface at <https://lifeportal.uio.no/root>. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Apply for a project, Help, User, and a status indicator showing "Using 78.9 GB". The main content area displays a "Tools" sidebar with categories like Bioportal Phylogeny Tools, R, Autodock, etc., and a "History" panel showing an uploaded file named "1: ERR101899_1.fasta.gz". A central message box informs the user that their upload has been queued and advises against using browser stop or reload buttons during the process. A large gray speech bubble on the right contains the text "Uploading data - not quite done". At the bottom, a progress bar indicates "Uploading (15%)".

Your upload has been queued. History entries that are still uploading will be blue, and turn green upon completion.

Please do not use your browser's "stop" or "reload" buttons until the upload is complete, or it may be interrupted.

You may safely continue to use Galaxy while the upload is in progress. Using "stop" and "reload" on pages other than Galaxy is also safe.

Uploading data -
not quite done

Uploading (15%)...

Look at data - eye symbol

The screenshot shows the Lifeportal web interface. In the center, there is a large yellow warning box stating: "This dataset is large and only the first megabyte is shown below." Below this message, a portion of a FASTA file is displayed, starting with the header @ERR101899.1 M10_151:1:2:13999:1320/1 and containing several lines of sequence data. To the left, a sidebar lists various bioinformatics tools such as Bioportal Phylogeny Tools, R, Autodock, and many others. On the right, a "History" panel shows a recent entry: "2: ERR101899_1.fasta" (147.5 MB). This entry is circled in red, and a "View data" button is visible below it. The browser address bar at the top shows the URL <https://lifeportal.uio.no/datasets/8f1384134e180689/display/?preview=True>.

Data annotation - pen symbol

The screenshot shows the Lifeportal web interface for managing datasets. On the left, a sidebar lists various bioinformatics tools. The main area displays the 'Edit Attributes' form for a dataset named 'ERR101899_1.fastq'. The form includes fields for 'Name', 'Info', 'Annotation / Notes', 'Database/Build', and buttons for 'Save' and 'Auto-detect'. A large callout bubble points to the 'Annotation / Notes' field with the text: 'Can add information about the data set here Good for tracking data'. On the right, a 'History' panel shows a list of datasets, with the entry '2: ERR101899_1.fastq' highlighted and circled in red.

Can add information
about the data set here
Good for tracking data

Attributes Convert Format Datatype Permissions

Edit Attributes

Name: ERR101899_1.fastq

Info: uploaded fastqsanger file

Annotation / Notes:

Add an annotation or notes to a dataset. Annotations are available when a history is viewed.

Database/Build: ----- Additional Species Are Below -----

Save Auto-detect This will inspect the dataset

History

Example history 147.5 MB

2: ERR101899_1.fastq 0 × Edit Attributes

UIO University of Oslo Analyze Data Workflow Shared Data Visualization Apply for a project Help User Using 79.0 GB

Lifeportal

Tools search tools Upload file Tests Bioportal Phylogeny Tools R Autodock UIO tools Gaussian NCBI BLAST+ Text Manipulation Filter and Sort Join, Subtract and Group Convert Formats Extract Features DeFuse Bismark Fetch Sequences Operate on Genomic Intervals Metagenomic analyses Statistics BEDTools Graph/Display Data FASTA manipulation NGS: GATK2 NGS: Picard (beta) NGS: Trim Galore NGS: Assembly Multiple Alignments NGS: QC and manipulation NGS: Mapping NGS: RNA Analysis NGS: SAM Tools NGS: GATK Tools (beta)

https://lifeportal.uio.no/datasets/8f1384134e180689/edit

Removing data set - X

The screenshot shows the Lifeportal web interface. On the left, a sidebar lists various tools under categories like 'Upload file', 'Tests', 'Bioportal Phylogeny Tools', and 'NGS: GATK2'. In the center, a large text area displays a sequence of error messages from a dataset named 'ERR101899.1'. A yellow warning box at the top states: 'This dataset is large and only the first megabyte is shown below.' Below this, a modal window contains the text: 'NOTE: removed data sets are not gone, just not shown in your history' and 'Need to do more to actually delete it'. On the right, a 'History' panel shows a list of datasets, with one entry circled in red: '2: ERR101899.1.fasta' (147.5 MB). The 'Delete' button for this entry is highlighted.

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

NOTE: removed data sets are not gone,
just not shown in your history

Need to do more to actually delete it

History

Example history
147.5 MB

2: ERR101899.1.fasta

Delete

https://lifeportal.uio.no/datasets/8f1384134e180689/delete_async

Analyzing data

Lifeportal

UiO University of Oslo Analyze Data Workflow Shared Data Visualization Apply for a project Help User Using 79.0 GB

FastQC:Read QC (version 0.10.1)

Short read data from your current history:
1: ERR101899_1.fastq

Title for the output file - to remind you what the job was for:
FastQC

Letters and numbers only please.

Contaminant list:
Selection is Optional

ART Primer CAAGCAGAAGACGGCATACTGA. Format: tabular.

ERS

Projects/Accounts

You have logged as user : Karin.lagesen@medisin.uio.no. Please select the project (account) where you will be running your current job.

cmb
--selected
cmb
metagen
nn905k

Number of tasks per node

Enter the number of tasks for the current job.
(Enter 1, if not sure.)

1

Number of tasks per node (max 16).

1

Walltime (job duration)

Select how long the job shall be running :

DD: 00 HH: 2 MM: 00 SC: 00

Memory per CPU

Select how much memory you need for the job
(the allowed format is digit+Gb/Mb, e.g. 16GB or 3000Mb):

4Gb

Execute

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing anything else with it.

The main types of FastQC are:

BAM, SAM or FastQ files (any variant)

History

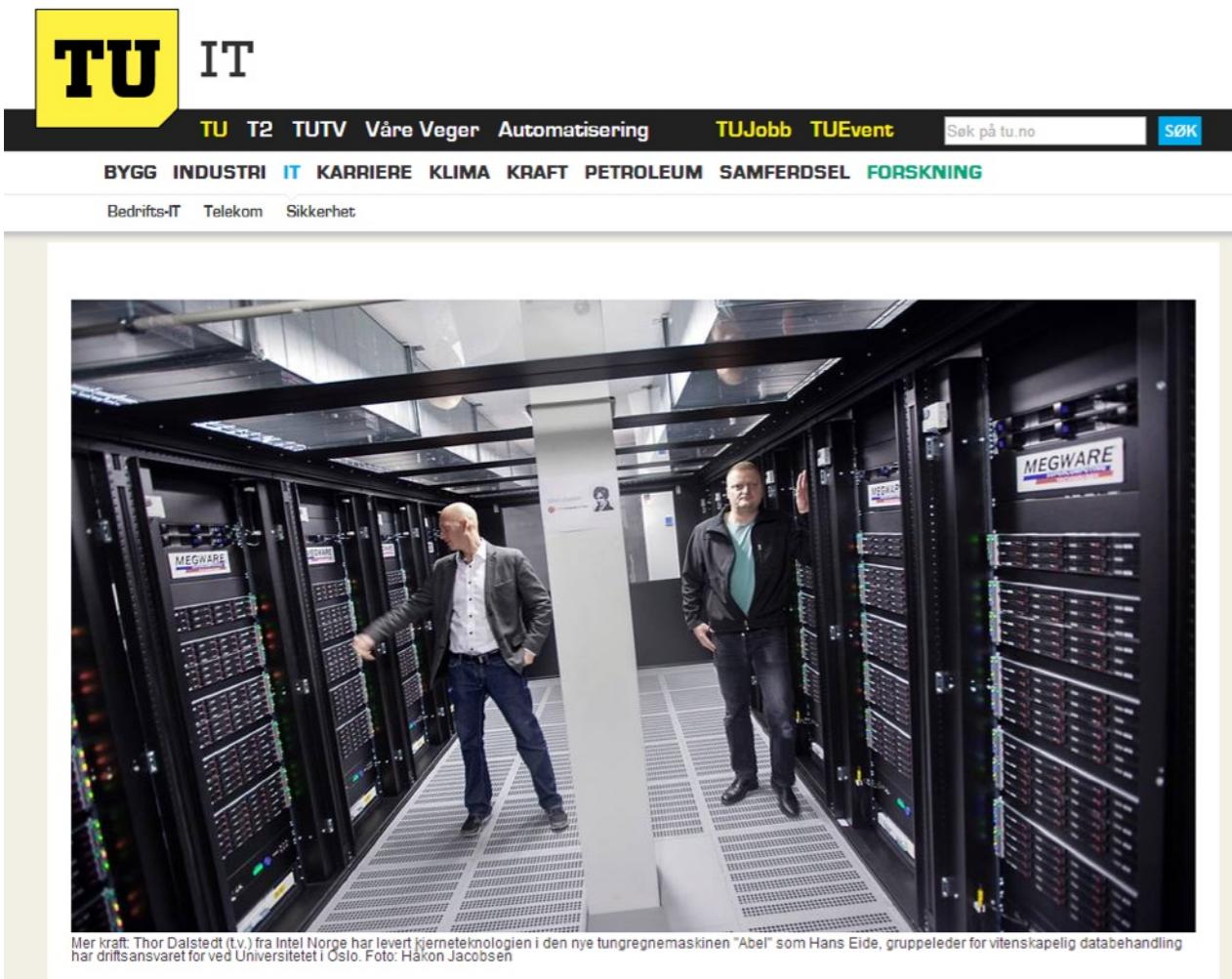
Example history
147.1 MB

1: ERR101899_1.fastq

Select input file here

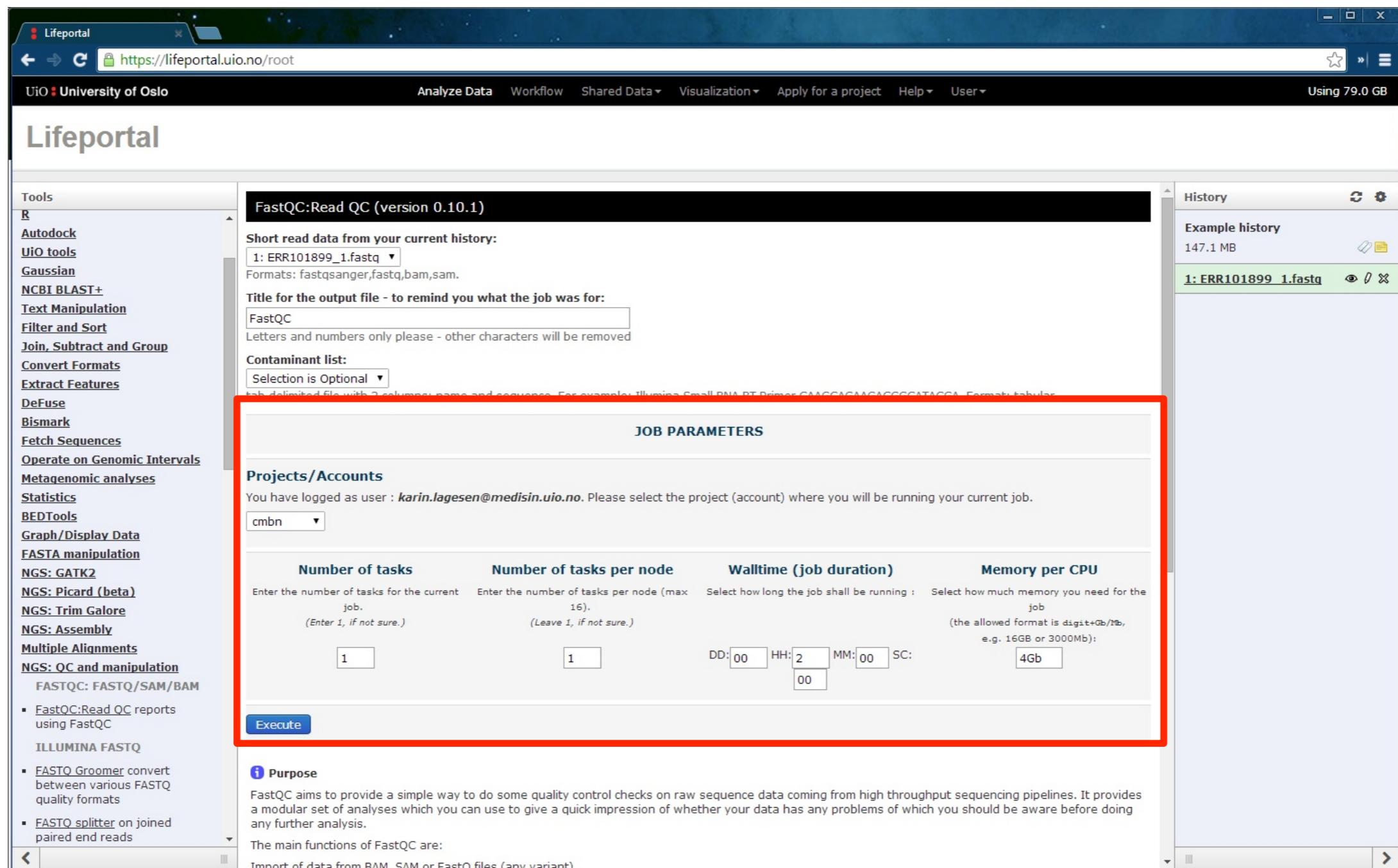
Select program in left bar

The Abel computer cluster



- Lifeportal runs on the Abel computer cluster
- > 10 000 cores!
- > 40 TB memory!
- Lifeportal submits jobs to the Abel cluster
- Can use several cores for a single job

Choose job options



The screenshot shows the Lifeportal web interface for the **FastQC:Read QC (version 0.10.1)** tool. The left sidebar lists various tools under the **Tools** category, including R, Autodock, UIo tools, Gaussian, NCBI BLAST+, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, DeFuse, Bismark, Fetch Sequences, Operate on Genomic Intervals, Metagenomic analyses, Statistics, BEDTools, Graph/Display Data, FASTA manipulation, NGS: GATK2, NGS: Picard (beta), NGS: Trim Galore, NGS: Assembly, Multiple Alignments, and NGS: QC and manipulation.

The main panel displays the configuration for the **FastQC:Read QC** tool. It includes fields for **Short read data from your current history:** (1: ERR101899_1.fastq), **Title for the output file - to remind you what the job was for:** (FastQC), **Contaminant list:** (Selection is Optional), and a note about tab-delimited files. The **JOB PARAMETERS** section is highlighted with a red box and contains fields for **Number of tasks** (1), **Number of tasks per node** (1), **Walltime (job duration)** (DD:00 HH:20 MM:00 SC:00), and **Memory per CPU** (4Gb). Below this, there is a **Execute** button and a **Purpose** section describing the tool's function.

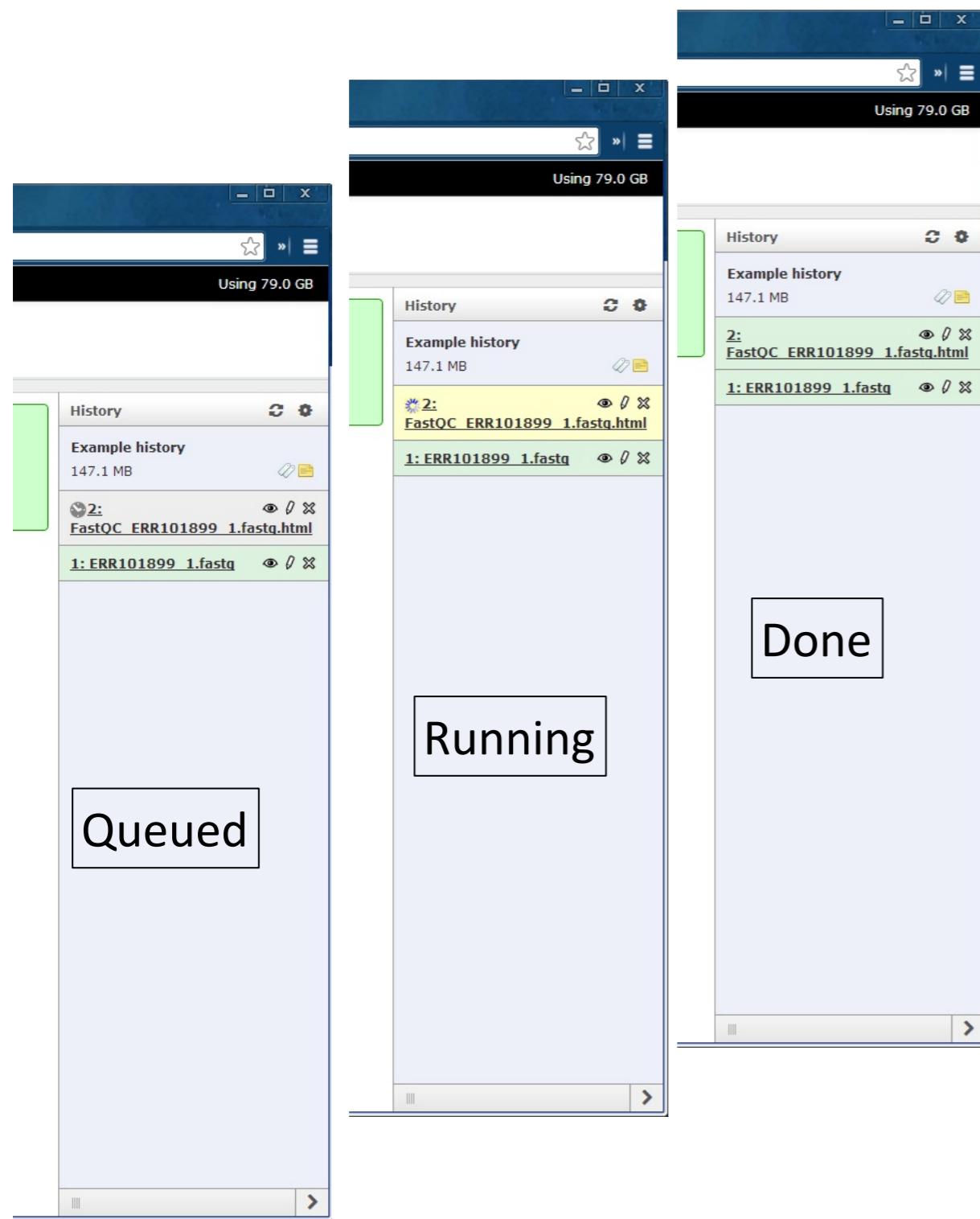
Job options

- # tasks = # cores you want to use
- # tasks per node:
 - One node has 16 cores, sometimes programs run faster if all cores are in the same node
- Wall time: guesstimated runtime.
 - Note: jobs exceeding that will be killed!
- Memory per cpu: each CPU has 4 GB of memory – just leave this option

CPU quotas

- Quotas calculated as # CPU hours
- All have 200 hrs to use
- Big projects should apply for their own quotas

Running job status



- Colors show the status of the job
- Purple: data uploading
- Gray: analysis queued
- Yellow: running
- Green: done
- Red: error has occurred

Results show up as new data set!

Screenshot of the Lifeportal interface showing results from a FastQC job.

The interface includes a sidebar with various tools, a main content area with "Basic Statistics" and "Per base sequence quality" sections, and a "History" panel on the right.

Basic statistics appear here: A callout points to the "Basic Statistics" section, which displays the following data:

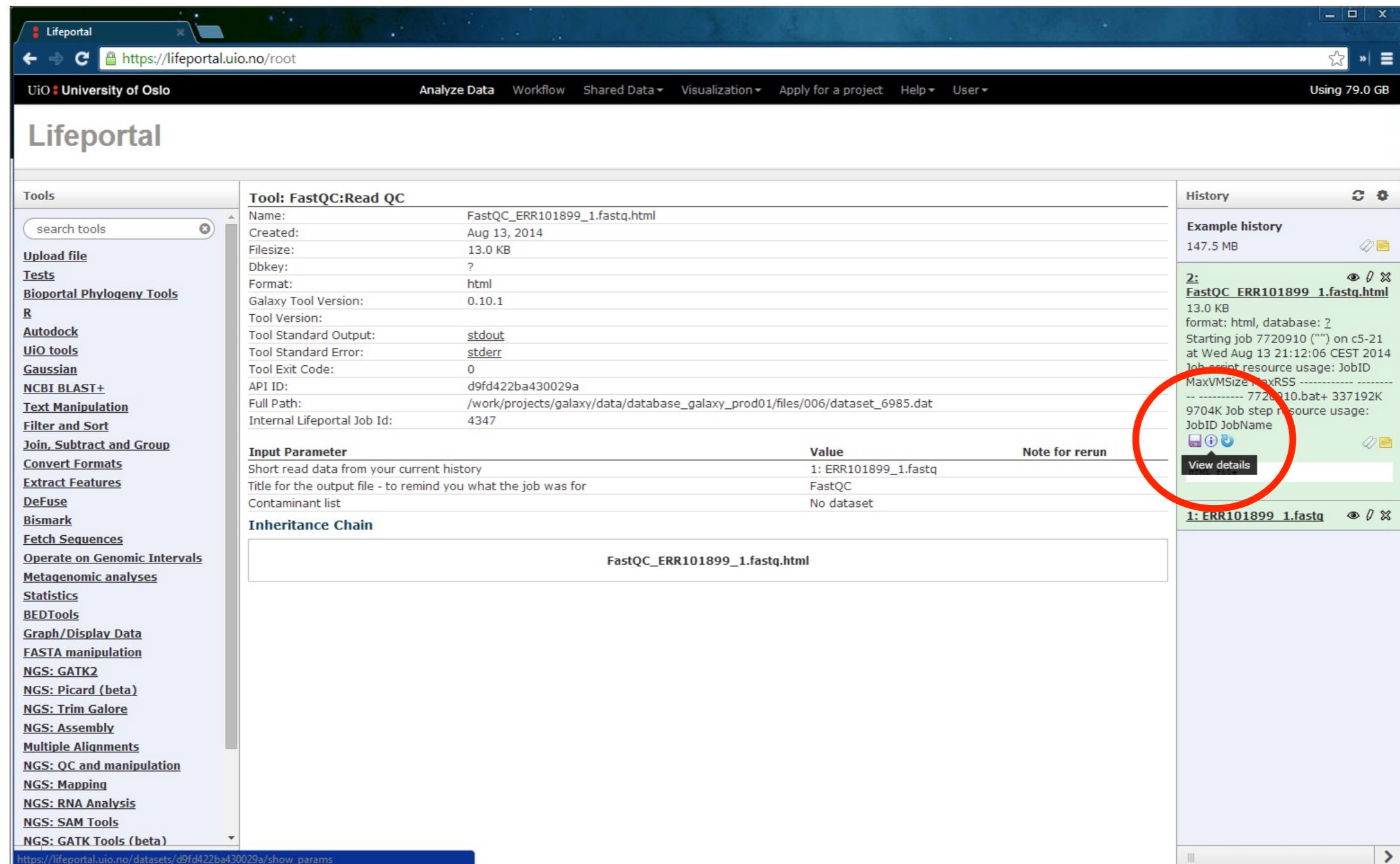
Measure	Value
Filename	ERR101899_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	443784
Filtered Sequences	0
Sequence length	150
%GC	32

Results from job show up as a new data set in history! A callout points to the "History" panel, which shows two entries:

- 2: FastQC ERR101899_1.fastq.html
- 1: ERR101899_1.fastq

FastQC quality plot A callout points to the "Per base sequence quality" section, which displays a stacked bar chart showing quality scores across all bases (Sanger / Illumina 1.9 encoding).

Data sets know how they were made



The screenshot shows the Lifeportal web interface at <https://lifeportal.uio.no/root>. The main content area displays a tool history for a "FastQC:Read QC" job. The job details include:

- Name: FastQC_ERR101899_1.fastq.html
- Created: Aug 13, 2014
- Filesize: 13.0 KB
- Dbkey: ?
- Format: html
- Galaxy Tool Version: 0.10.1
- Tool Version: 2
- Tool Standard Output: stdout
- Tool Standard Error: stderr
- Tool Exit Code: 0
- API ID: d9fd422ba430029a
- Full Path: /work/projects/galaxy/data/database_galaxy_prod01/files/006/dataset_6985.dat
- Internal Lifeportal Job Id: 4347

The "Input Parameter" section shows:

	Value	Note for rerun
Short read data from your current history	1: ERR101899_1.fastq	
Title for the output file - to remind you what the job was for	FastQC	
Contaminant list	No dataset	

The "Inheritance Chain" section lists:

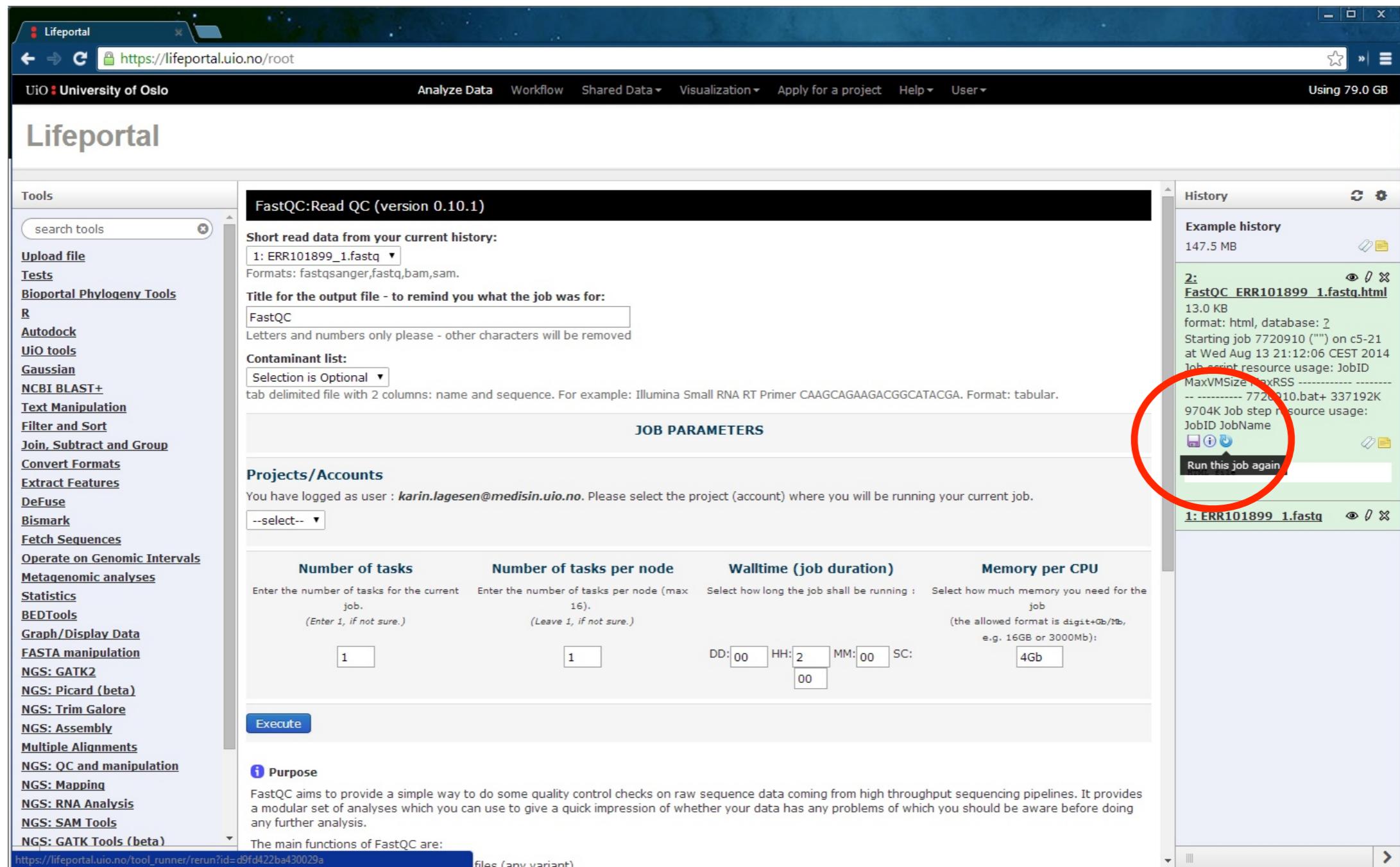
- FastQC_ERR101899_1.fastq.html

The right side of the screen shows a "History" panel with a list of previous jobs, one of which is circled in red. The circled job is "FastQC_ERR101899_1.fastq.html" with the following details:

- Size: 147.5 MB
- Format: html, database: 2
- Starting job 7720910 ("") on c5-21 at Wed Aug 13 21:12:06 CEST 2014
- Job script resource usage: JobID
- MaxVMSize, maxRSS
- 7720910.bat+ 337192K
- 9704K Job step resource usage: JobID JobName

A "View details" button is visible next to the circled job.

Can easily run analyses again



The screenshot shows the Lifeportal web interface for running bioinformatics analyses. On the left, a sidebar lists various tools categorized under 'Tools'. The main area is focused on the 'FastQC:Read QC (version 0.10.1)' tool. The configuration form includes fields for 'Short read data from your current history' (containing '1: ERR101899_1.fastq'), 'Title for the output file', 'Contaminant list', and 'JOB PARAMETERS' (with 'Number of tasks' set to 1, 'Number of tasks per node' set to 1, 'Walltime (job duration)' set to 00:20:00, and 'Memory per CPU' set to 4Gb). Below the parameters is an 'Execute' button. To the right, a 'History' panel displays a list of previous jobs. The job 'FastQC_ERR101899_1.fastq.html' is highlighted in green and has a red circle around its 'Run this job again' link.

What did I do again...?

Screenshot of the Lifeportal interface showing the 'Saved Histories' section. A red circle highlights the 'Saved Histories' option in the context menu.

Tools

- Upload file
- Tests
- Bioportal Phylogeny Tools
- R
- Autodock
- UiO tools
- Gaussian
- NCBI BLAST+
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- DeFuse
- Bismark
- Fetch Sequences
- Operate on Genomic Intervals
- Metagenomic analyses
- Statistics
- BEDTools
- Graph/Display Data
- FASTA manipulation
- NGS: GATK2
- NGS: Picard (beta)
- NGS: Trim Galore
- NGS: Assembly
- Multiple Alignments
- NGS: QC and manipulation
- NGS: Mapping
- NGS: RNA Analysis
- NGS: SAM Tools
- NGS: GATK Tools (beta)

Saved Histories

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
Example history ▾	1	0 Tags		147.5 MB	~ 1 hour ago	~ 1 hour ago	
Example history ▾	2	0 Tags		147.5 MB	~ 3 hours ago	~ 2 hours ago	current history
2cells_zebrafish ▾	14	0 Tags	Accessible	451.6 MB	3 days ago	1 day ago	
BIV4000_RNA_2cells ▾	4	0 Tags	Accessible, Published	359.4 MB	3 days ago	3 days ago	
BIV4000_RNA_6hrs ▾	4	0 Tags	Accessible, Published	378.5 MB	3 days ago	3 days ago	
6hr_zebra ▾	14	0 Tags		590.4 MB	3 days ago	3 days ago	
6hrs_zebrafish ▾	14	0 Tags		606.4 MB	3 days ago	3 days ago	
rna_seq_zebra ▾	30	0 Tags		3.3 GB	5 days ago	5 days ago	
dataset tests ▾	25	0 Tags		25.1 GB	5 days ago	5 days ago	
rnaseq drerio ▾	6	4	0 Tags	47.5 GB	6 days ago	6 days ago	
zebrafish_temperature ▾	2	0 Tags		11.4 GB	Aug 06, 2014	6 days ago	
biv4000_phyl_complete ▾	8	0 Tags	Accessible	215.9 KB	6 days ago	6 days ago	
imported: BIV4000_PHYL ▾	2	0 Tags		22.0 KB	6 days ago	6 days ago	
BIV4000_PHYL ▾	2	0 Tags	Accessible, Published	22.0 KB	6 days ago	6 days ago	
RNAseq test ▾	17	0 Tags		800.8 MB	Aug 06, 2014	6 days ago	
workcomputer ▾	5	0 Tags		794.0 MB	Jul 22, 2014	Aug 06, 2014	

HISTORY LISTS

- Saved Histories**
- Histories Shared with Me
- CURRENT HISTORY
- Create New History
- Copy History
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Include Deleted Datasets
- Include Hidden Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently
- OTHER ACTIONS
- Import from File

Can look at old analyses

Screenshot of the Lifeportal web interface showing a history of analyses.

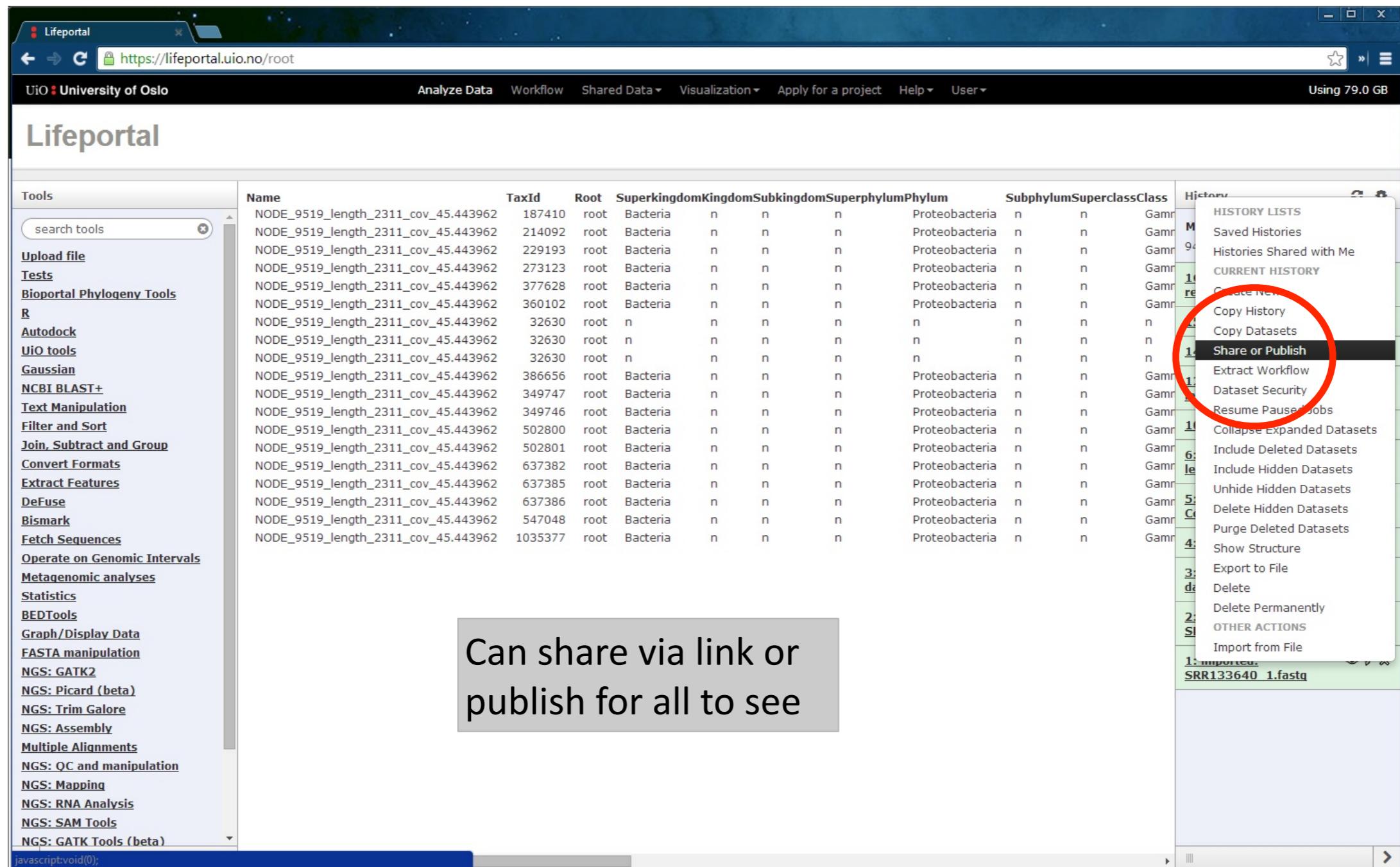
The interface includes a navigation bar with tabs for Analyze Data, Workflow, Shared Data, Visualization, Apply for a project, Help, and User, and a status message "Using 79.0 GB".

The main area displays a table of analysis steps:

Name	TaxId	Root	Superkingdom	Kingdom	Subkingdom	Superphylum	Phylum	Subphylum	Superclass	Class	History
NODE_9519_length_2311_cov_45.443962	187410	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	now
NODE_9519_length_2311_cov_45.443962	214092	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	949.1 MB
NODE_9519_length_2311_cov_45.443962	229193	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	
NODE_9519_length_2311_cov_45.443962	273123	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	
NODE_9519_length_2311_cov_45.443962	377628	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	16: Fetch taxonomic representation on data 15
NODE_9519_length_2311_cov_45.443962	360102	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	15: Convert on data 14
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n	14: megablast on db
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n	12: Filter sequences by length on data 5
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n	10: Sort on data 6
NODE_9519_length_2311_cov_45.443962	386656	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	6: Compute sequence length on data 5
NODE_9519_length_2311_cov_45.443962	349747	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	5: velvetg on data 3: Contigs
NODE_9519_length_2311_cov_45.443962	349746	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	4: velvetg on data 3: Stats
NODE_9519_length_2311_cov_45.443962	502800	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	3: velveth on data 1 and data 2
NODE_9519_length_2311_cov_45.443962	502801	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	2: imported: SRR133640_2.fastq
NODE_9519_length_2311_cov_45.443962	637382	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	1: imported: SRR133640_1.fastq
NODE_9519_length_2311_cov_45.443962	637385	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	
NODE_9519_length_2311_cov_45.443962	637386	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	
NODE_9519_length_2311_cov_45.443962	547048	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	
NODE_9519_length_2311_cov_45.443962	1035377	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam	

The left sidebar lists various tools and services available on the platform, such as Bioportal Phylogeny Tools, R, Autodock, and various NGS and bioinformatics tools.

Share or publish histories



The screenshot shows the Lifeportal web interface. On the left, there's a sidebar with a list of tools and analysis types. The main area displays a table of dataset metadata. On the right, a vertical menu is open under the 'History' heading. The 'Share or Publish' option is highlighted with a red circle. A callout box in the bottom center says: "Can share via link or publish for all to see".

Name	TaxId	Root	Superkingdom	Kingdom	Subkingdom	Superphylum	Phylum	Subphylum	Superclass	Class
NODE_9519_length_2311_cov_45.443962	187410	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	214092	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	229193	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	273123	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	377628	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	360102	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n
NODE_9519_length_2311_cov_45.443962	32630	root	n	n	n	n	n	n	n	n
NODE_9519_length_2311_cov_45.443962	386656	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	349747	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	349746	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	502800	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	502801	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	637382	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	637385	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	637386	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	547048	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam
NODE_9519_length_2311_cov_45.443962	1035377	root	Bacteria	n	n	n	Proteobacteria	n	n	Gam

Published histories open to all

Published Histories

search name, annotation, owner, and tags 

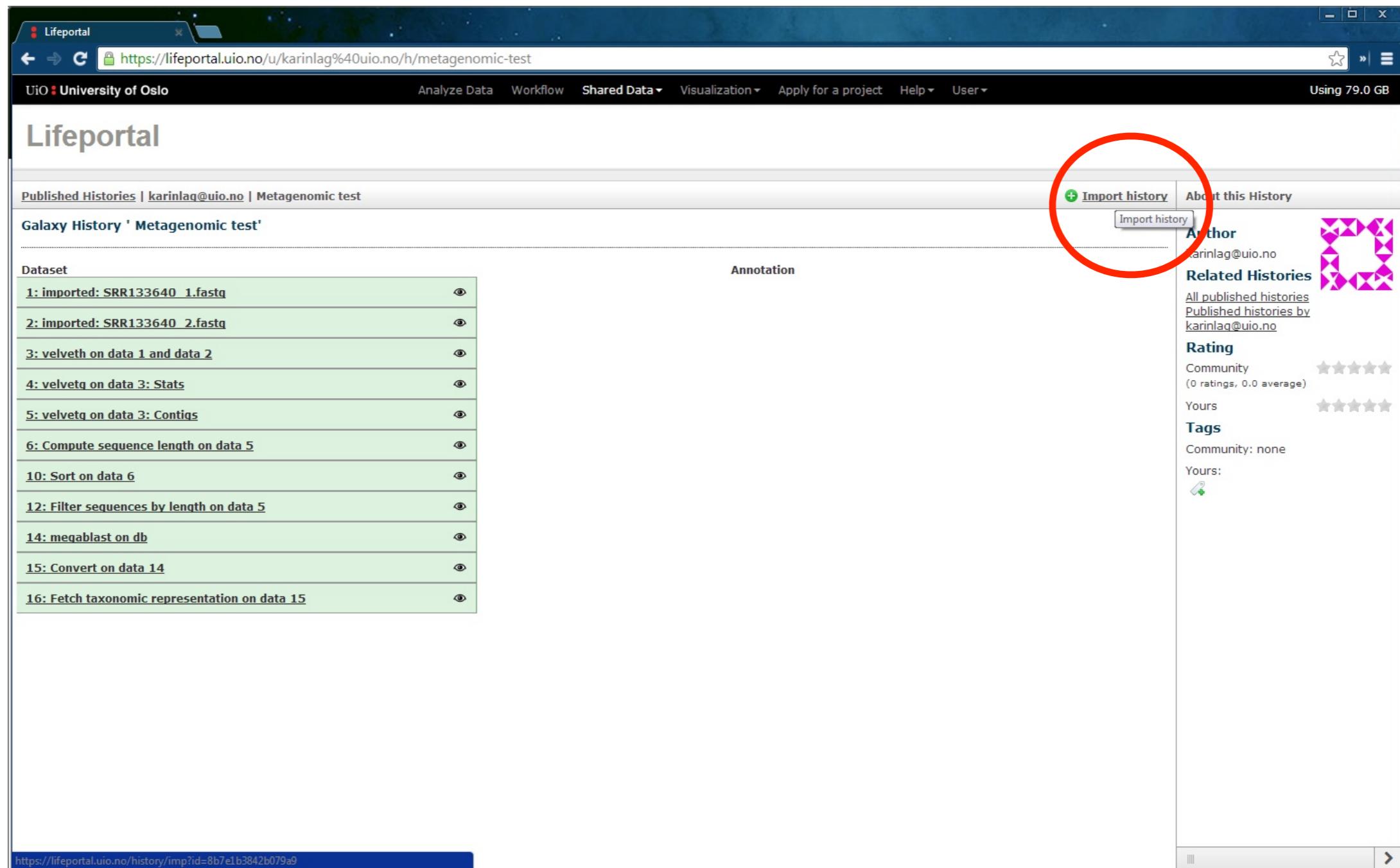
[Advanced Search](#)

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated ↓
Metagenomic test		karinlag@uio.no	★★★★★		less than a minute ago
BIV4000 RNA_2cells		karinlag@uio.no	★★★★★		3 days ago
BIV4000 RNA_6hrs		karinlag@uio.no	★★★★★		3 days ago
BIV4000 PHYL		karinlag@uio.no	★★★★★		6 days ago
RFK_Jan22_bartonella		katerim@uio.no	★★★★★		Jul 01, 2014
lifeportal basics		katerim@uio.no	★★★★★		Apr 08, 2014
hanshake		katerim@uio.no	★★★★★		Nov 15, 2013
LifePortal demo (cheating..)		sveinugu@uio.no	★★★★★		Oct 09, 2013
LifePortal demo 1		sveinugu@uio.no	★★★★★		Oct 09, 2013

NOTE: others can not only look at published histories, they can also copy data sets from it!

Practical way to share data!

Importing shared histories



The screenshot shows the Lifeportal web interface for a user named 'karinlag@ui.no'. The URL in the address bar is <https://lifeportal.uio.no/u/karinlag%40ui.no/h/metagenomic-test>. The main content area displays a 'Galaxy History' titled 'Metagenomic test'. On the left, there is a list of steps in the history:

- 1: imported: SRR133640_1.fastq
- 2: imported: SRR133640_2.fastq
- 3: velvet on data 1 and data 2
- 4: velvetq on data 3: Stats
- 5: velvetq on data 3: Contigs
- 6: Compute sequence length on data 5
- 10: Sort on data 6
- 12: Filter sequences by length on data 5
- 14: megablast on db
- 15: Convert on data 14
- 16: Fetch taxonomic representation on data 15

On the right side of the history, there is an 'Annotation' section and a sidebar with various history-related links and metrics. A red circle highlights the 'Import history' button in the top right corner of the main content area.

Galaxy: other tutorials

- For more tutorials and exercises, check out:

<http://wiki.g2.bx.psu.edu/Learn>

- Article with step-for-step examples/protocols making use of Galaxy in different scenarios:

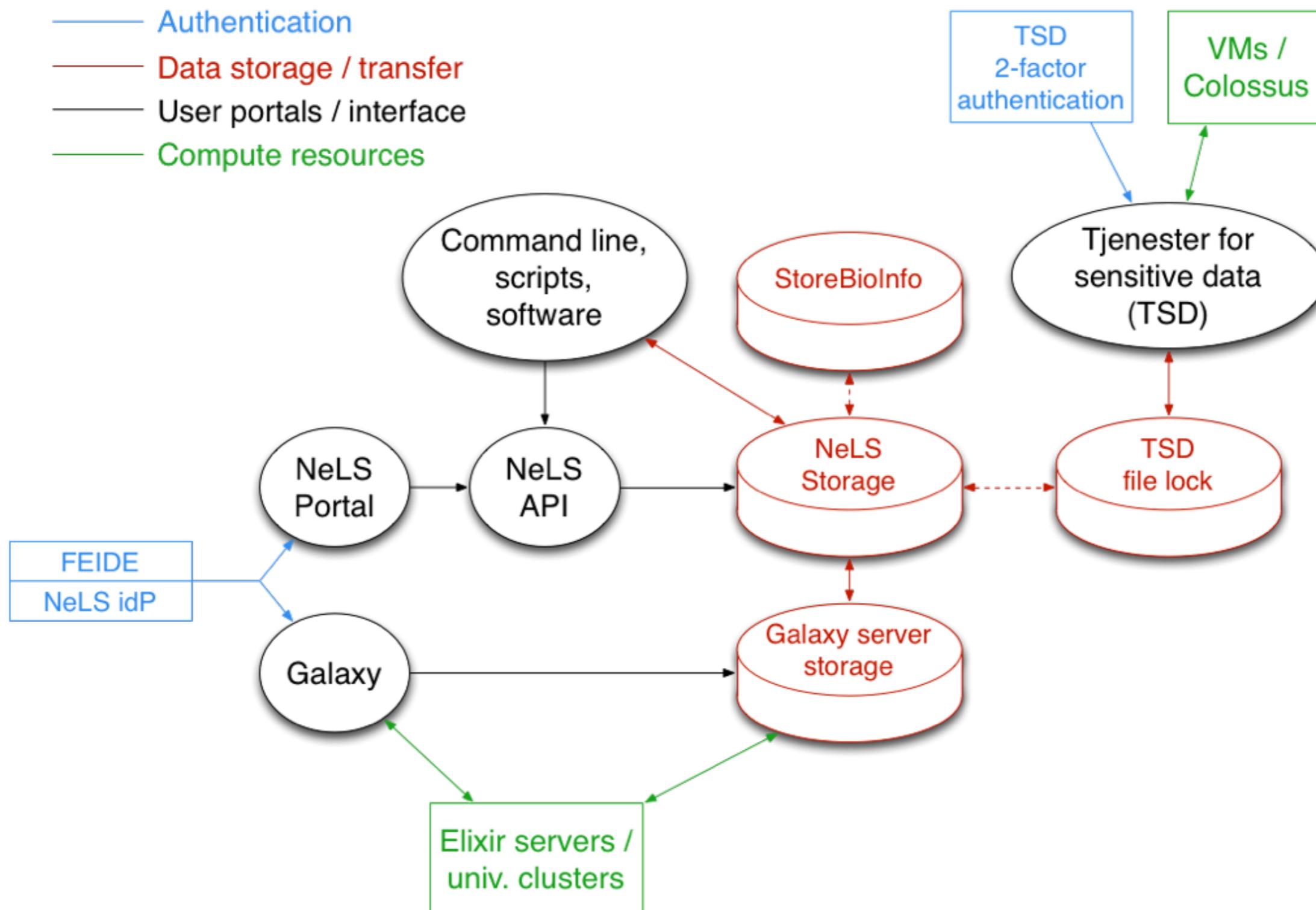
Blankenberg, D., et al., Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology, Jan 2010, Chapter 19.

ELIXIR.NO

- National project for building and maintaining e-Infrastructure for Life Scientists
- Part of ELIXIR Europe, cross-european collaboration on Life Science e-Infrastructure
- <http://www.bioinfo.no/elixir>
- Main product:
 - NeLS (Norwegian e-Infrastructure for Life Sciences)

- <http://nels.bioinfo.no>
- Central user authentication
- Connecting 6 Norwegian Galaxy installations:
LifePortal, galaxy-uio.bioinfo.no, galaxy-ntnu.bioinfo.no...
- Also connected to Tjenester for Sensitive Data (TSD)
- Possible to upload/download data from command line (using ssh), and Python/Java code
- Project storage for sharing between project members
- Long-time archiving of data using StoreBioInfo
- Sequencing providers will provide data using NeLS

Overview of National e-Infrastructure for Life Science



Demo: get files from NeLS storage

UiO : University of Oslo

Analyze Data Workflow Shared Data ▾ Visualization ▾ Apply for a project Help ▾ User ▾

Using 60.0 GB

Lifeportal

Tools

search tools

Import files

- Upload File from your computer
- Big File Upload filesender server
- Get files from NeLS storage

Export files

Get Data

Tests

Bioportal Phylogeny Tools

R

Gaussian

Autodock

UiO tools

NCBI BLAST+

Multiple Alignments

Text Manipulation

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Metagenomic analyses

Statistics

DeFuse

Bismark

NGS: Mapping

NGS: QC and manipulation

NGS: SAM Tools

NGS: RNA Analysis

BETA
NeLS
Norwegian e-Infrastructure for Life Sciences

Home

Logged in as: Sveinung Gundersen (NeLS ID : 12)

My Data

You are at : /Personal /RNA-seq

	Name	Size	Modified
<input type="checkbox"/>	1.chr19.fastq	134.1 MB	March 10, 2015
<input type="checkbox"/>	2.chr19.fastq	123.2 MB	March 10, 2015
<input type="checkbox"/>	3.chr19.fastq	124.9 MB	March 10, 2015
<input type="checkbox"/>	4.chr19.fastq	166.9 MB	March 10, 2015
<input type="checkbox"/>	mouse-rnaseq-chr19.tar	132.2 MB	March 10, 2015
<input type="checkbox"/>	mouse-rnaseq-chr19.tar.gz	87.3 MB	March 10, 2015

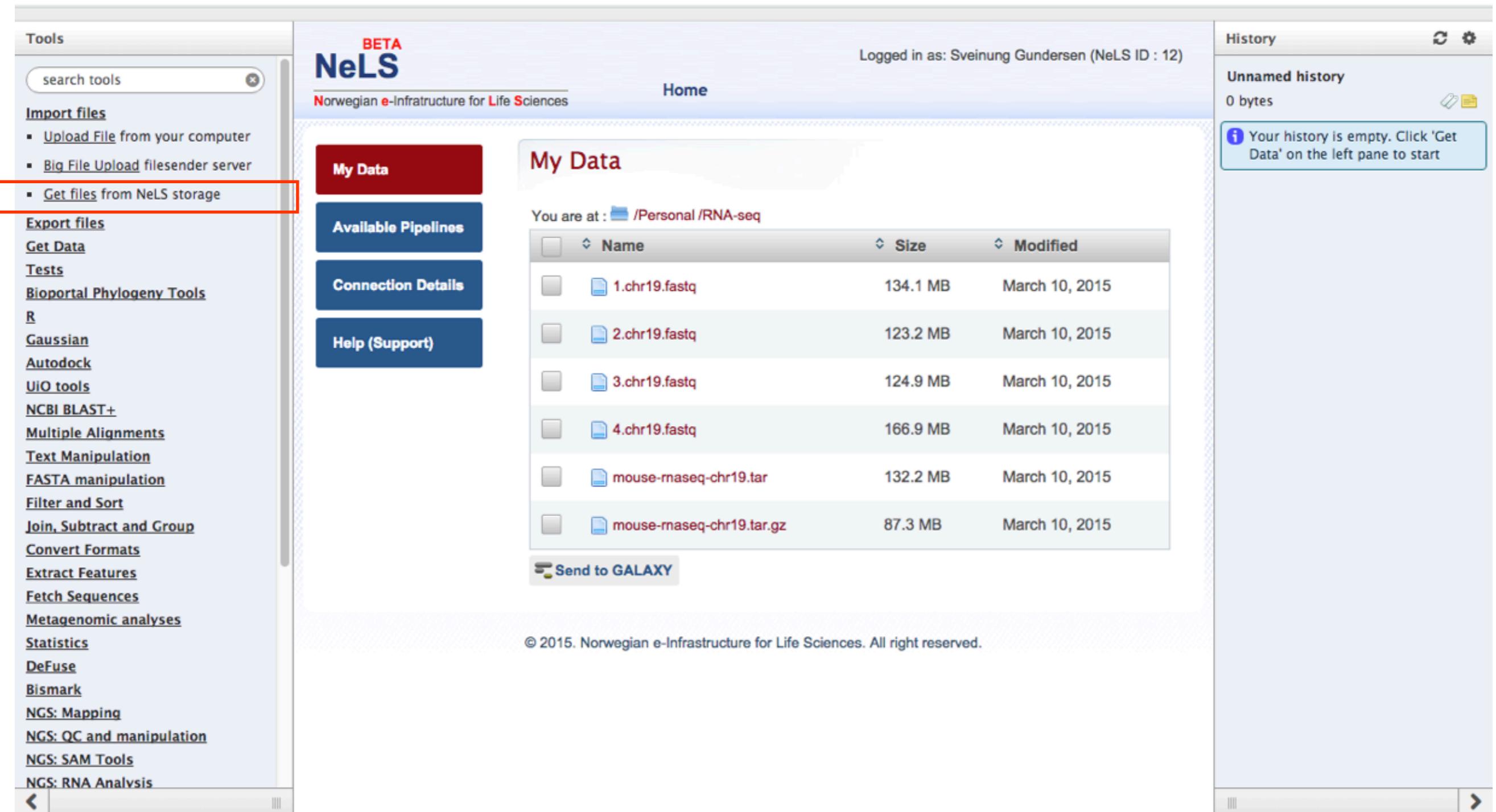
Send to GALAXY

History

Unnamed history

0 bytes

i Your history is empty. Click 'Get Data' on the left pane to start



Getting support from

UiO : University of Oslo

and

Lifeportal



- **lifeportal-help@usit.uio.no** (LifePortal, USIT):
 - Installing tools, technical issues (“Why doesn’t this work?”)
- **contact.bioinfo.no** (ELIXIR.NO):
 - Bioinformatics analysis
 - “Can you help me analyze my data?” (Parameters, workflows, best practices...)
 - “Can you analyze my data?” (Small projects for free, payment for larger projects)
 - User support on NeLS