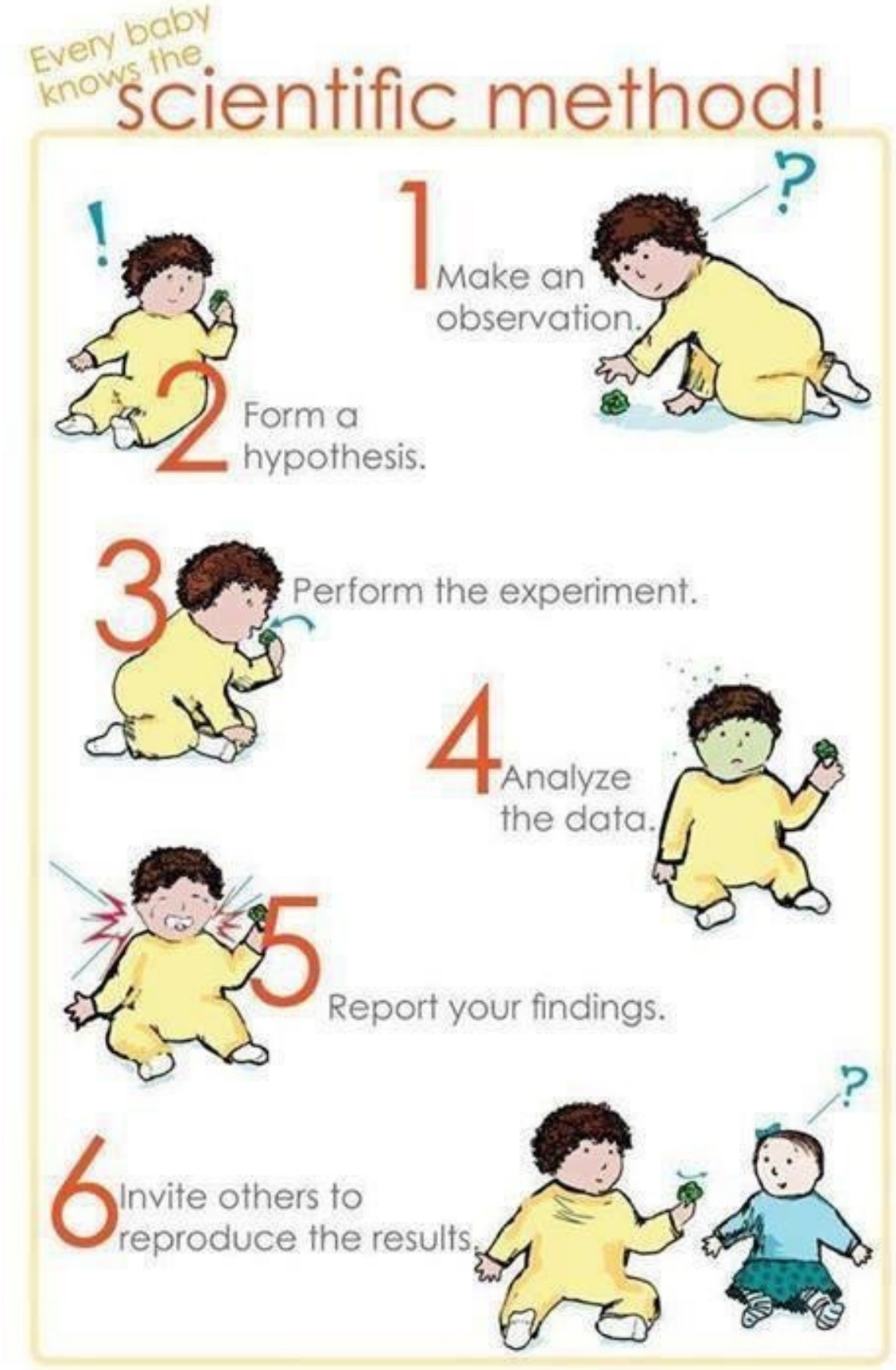
Introduction to Galaxy

INF-BIO5121/INF-BIO9121
October 7. 2015, Oslo

Sveinung Gundersen



by Tiffany Ard, Nerdy Baby artwork, https://www.facebook.com/NerdyBabyLLC

- 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> {-1 <m1> -2 <m2> | --12 <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     -1, -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 refXXXXX.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       supppress header lines (starting with @) for SAM output
  --sam-nosq         supppress @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 was this again?

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?

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!

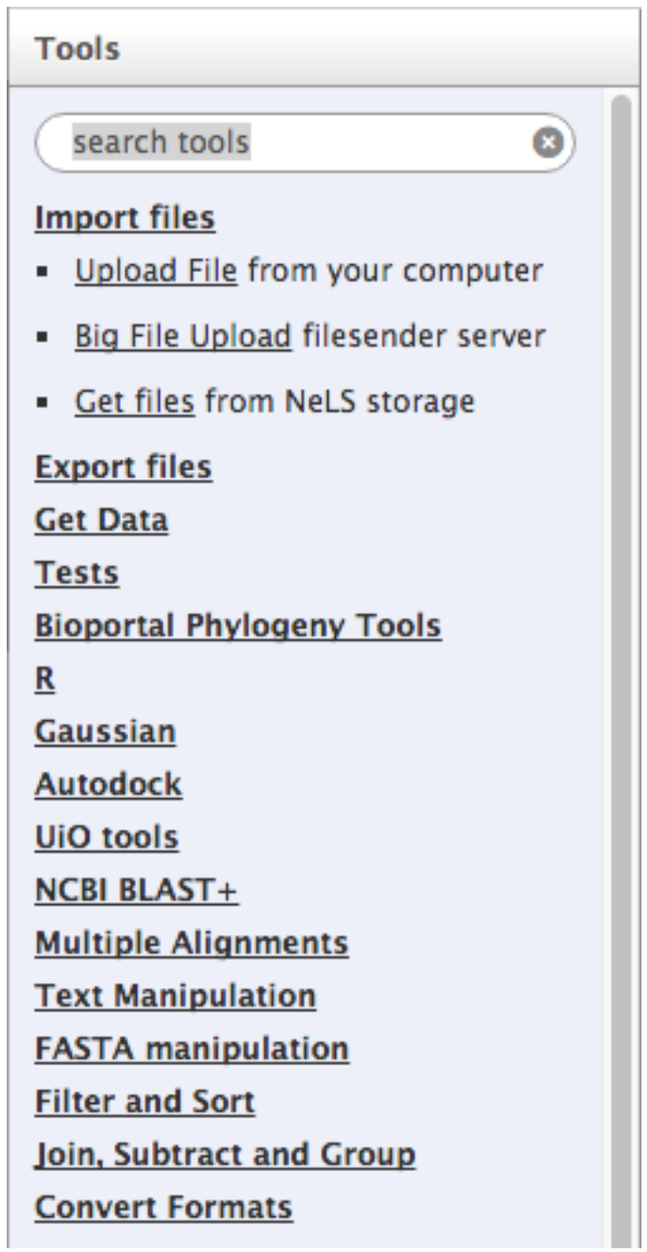
Galaxy supports reproducibility

- 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

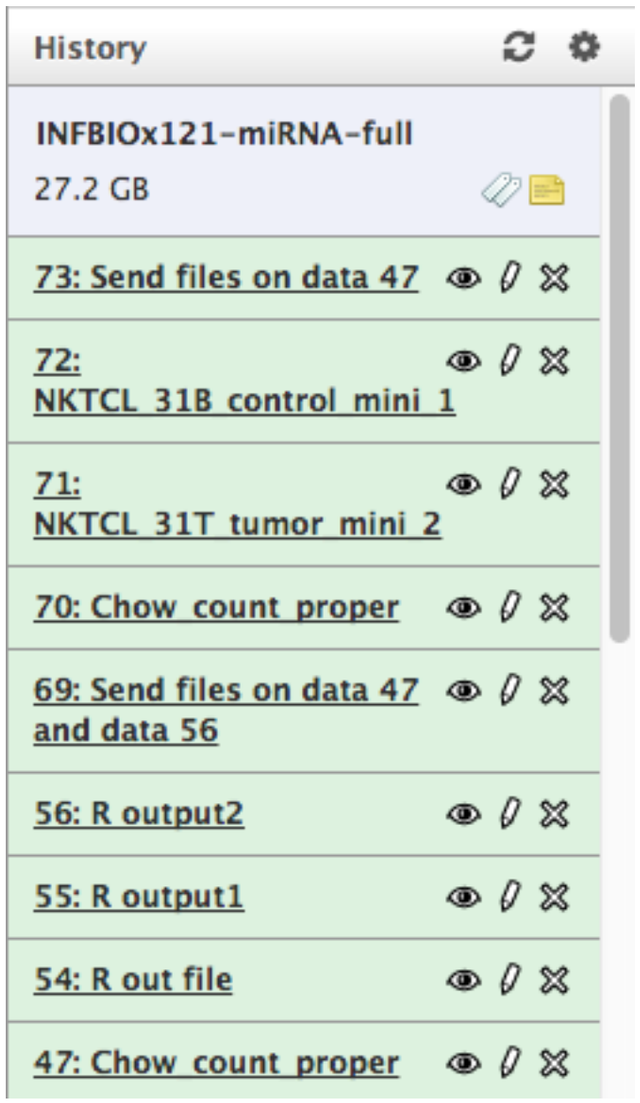
- "Enabling users to share and communicate their experimental results and outputs in a meaningful way" *(Goecks et. al., Genome Biology 2010)*
- Everything can be shared: Datasets, histories (i.e. experimental logbook), tools, workflows
- Provides public repositories
- Galaxy Pages are web-based documents 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



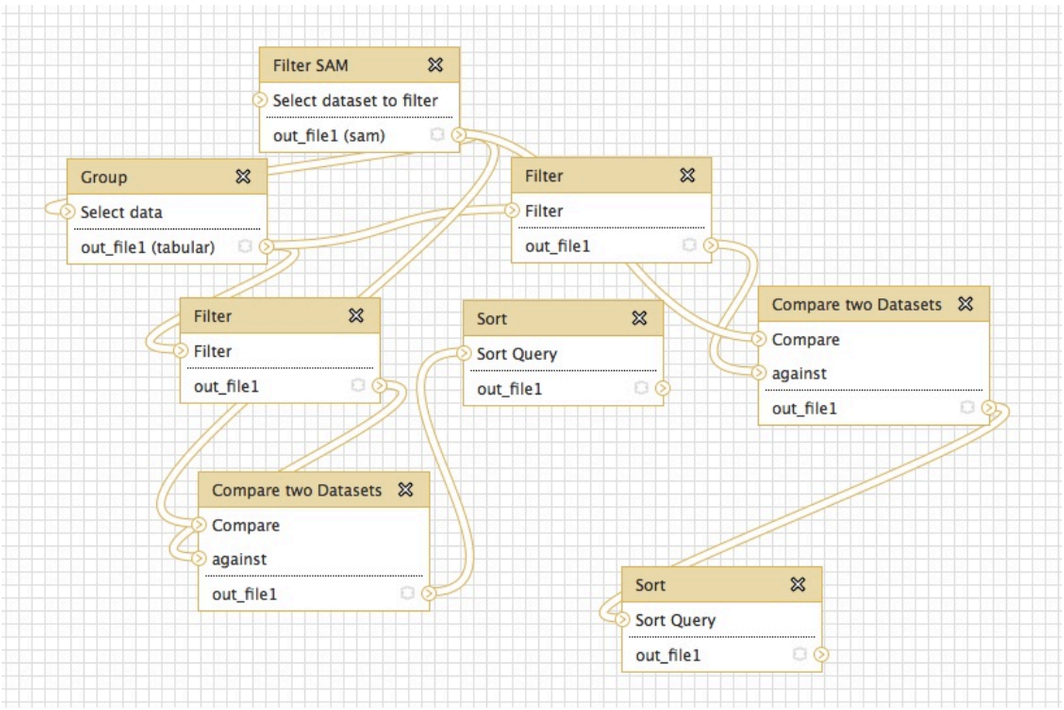
- The history is a log of you analysis
- Contains all:
  - Input datasets
  - Intermediate datasets
  - Analysis output

    (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!

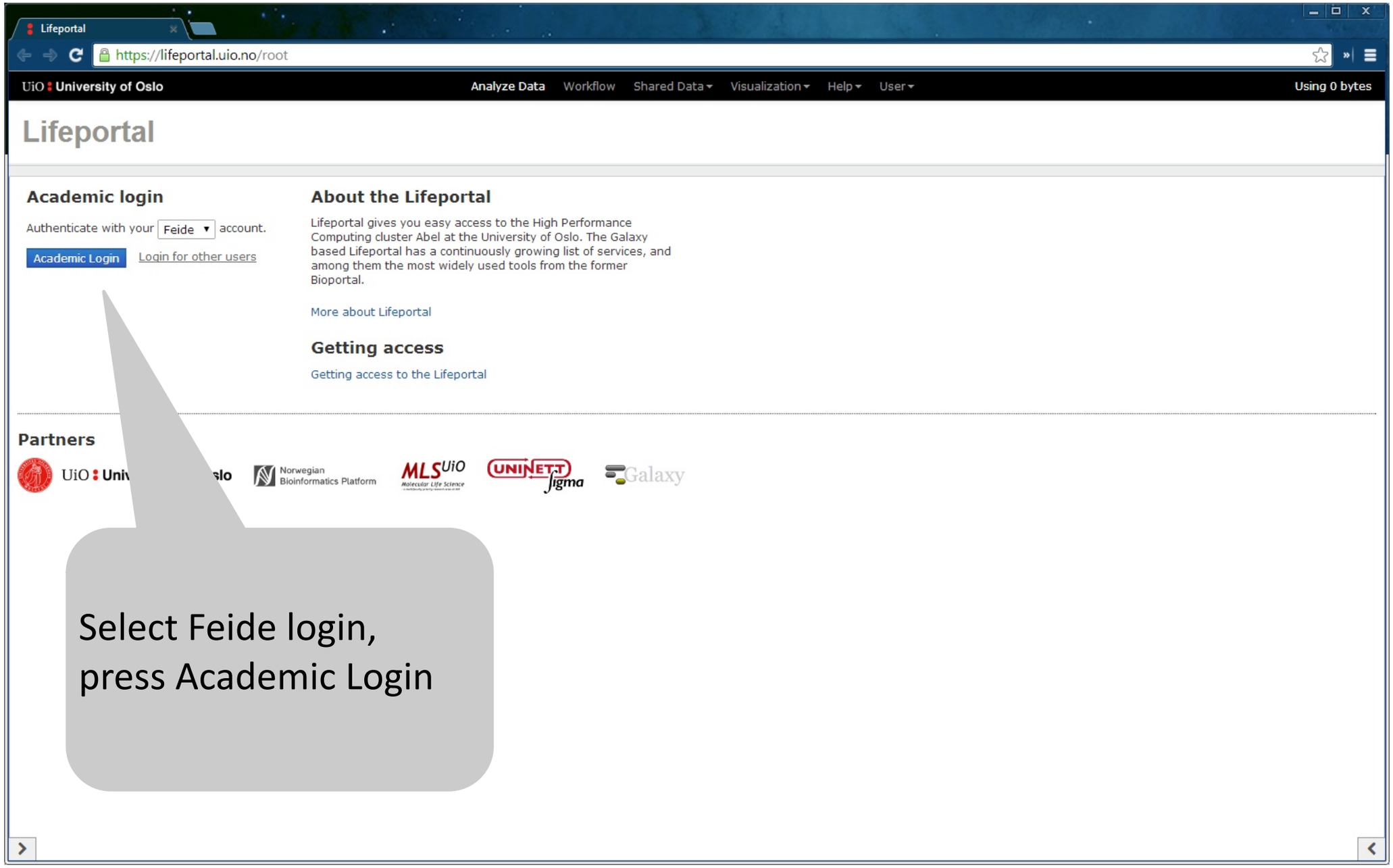
Basic concepts: workflow



- 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

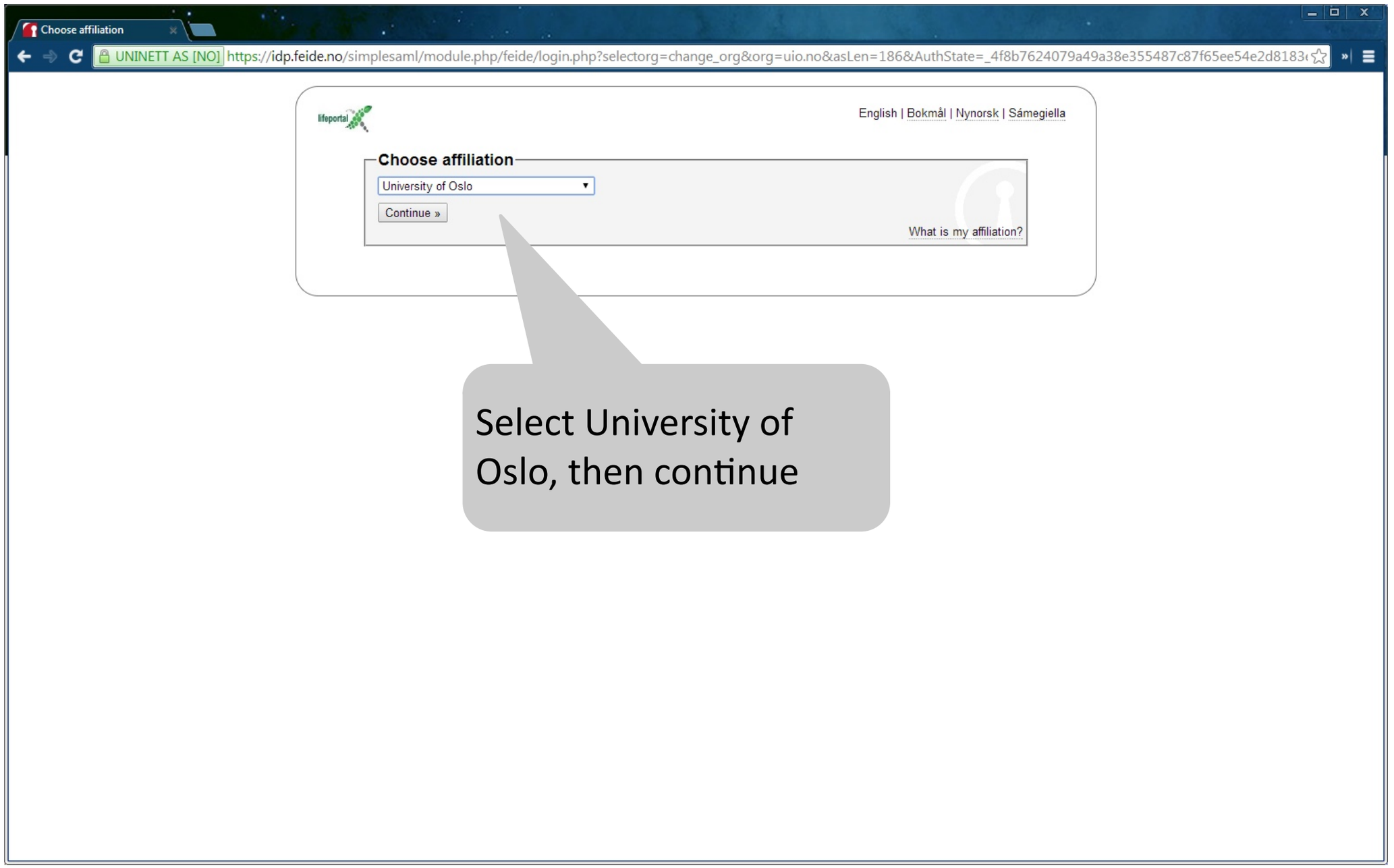
- Galaxy installation at UiO, running on the Abel cluster
- Contains hundreds of tools, from Phylogeny tools to High Throughput Sequencing analysis
- Available for all FEIDE users (all university users and several colleges)
- Other users may apply for access

Demo: lifeportal.uio.no

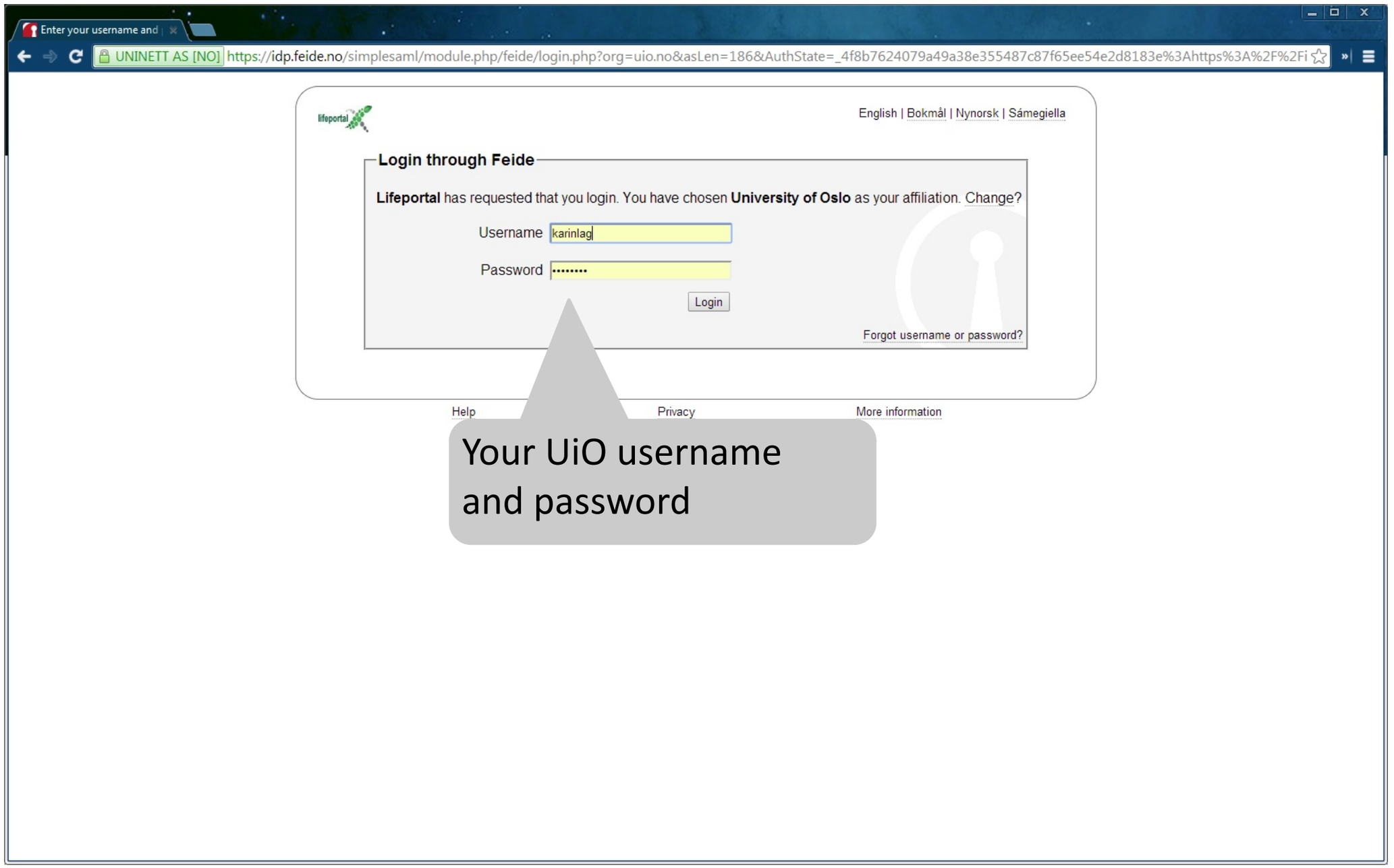


Copied from presentation by Karin Lagesen

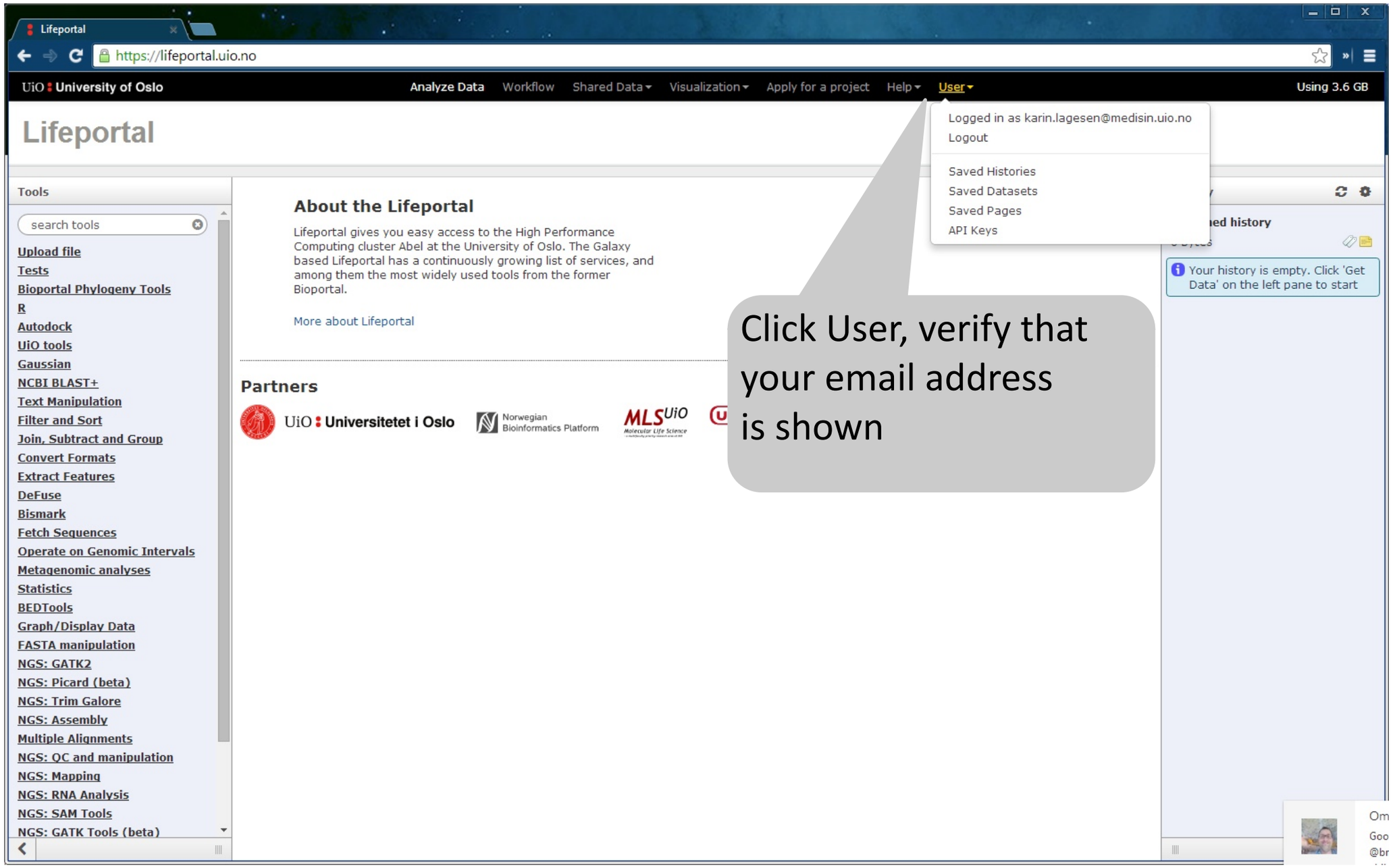
Select your institution



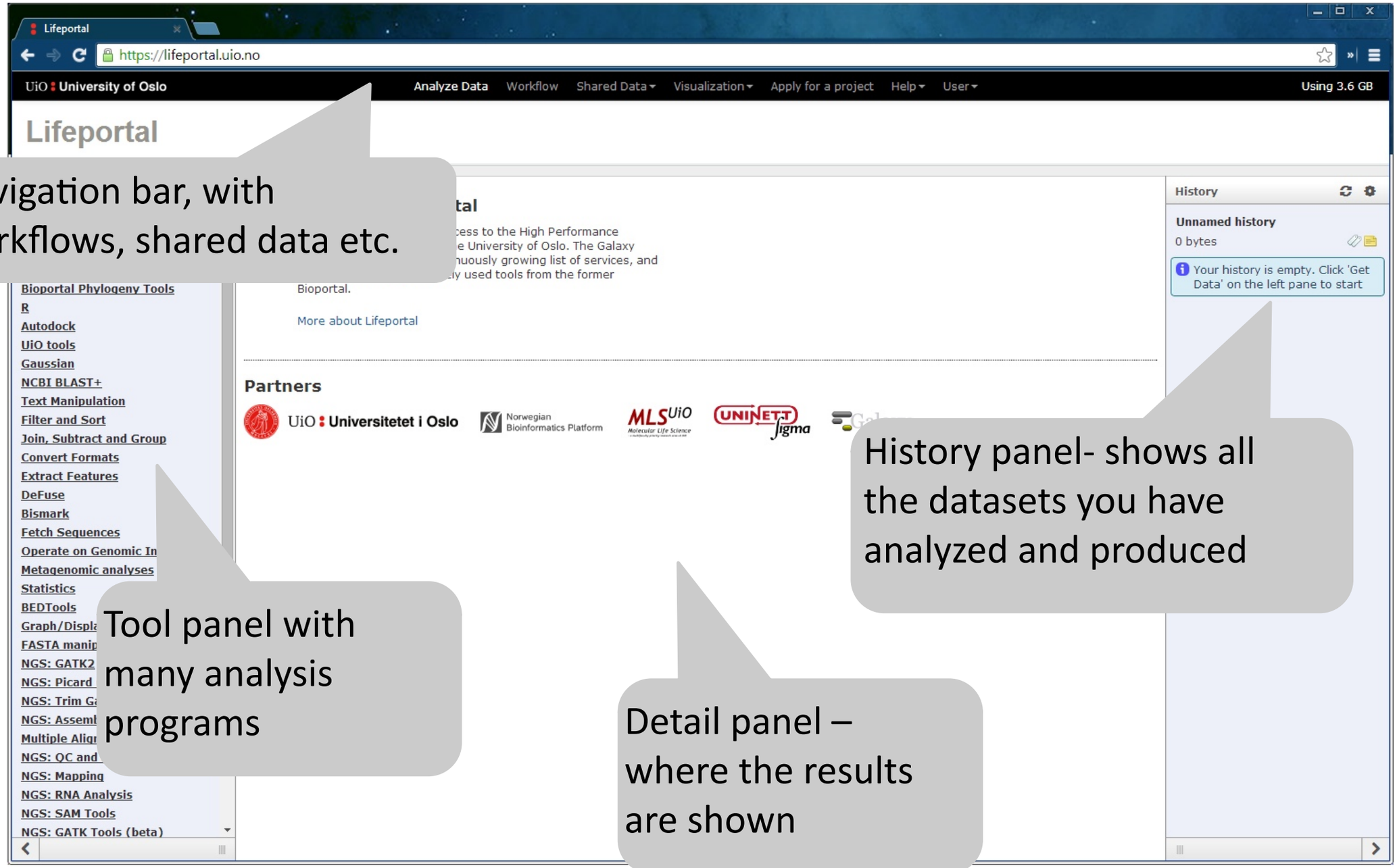
Use UiO username/password



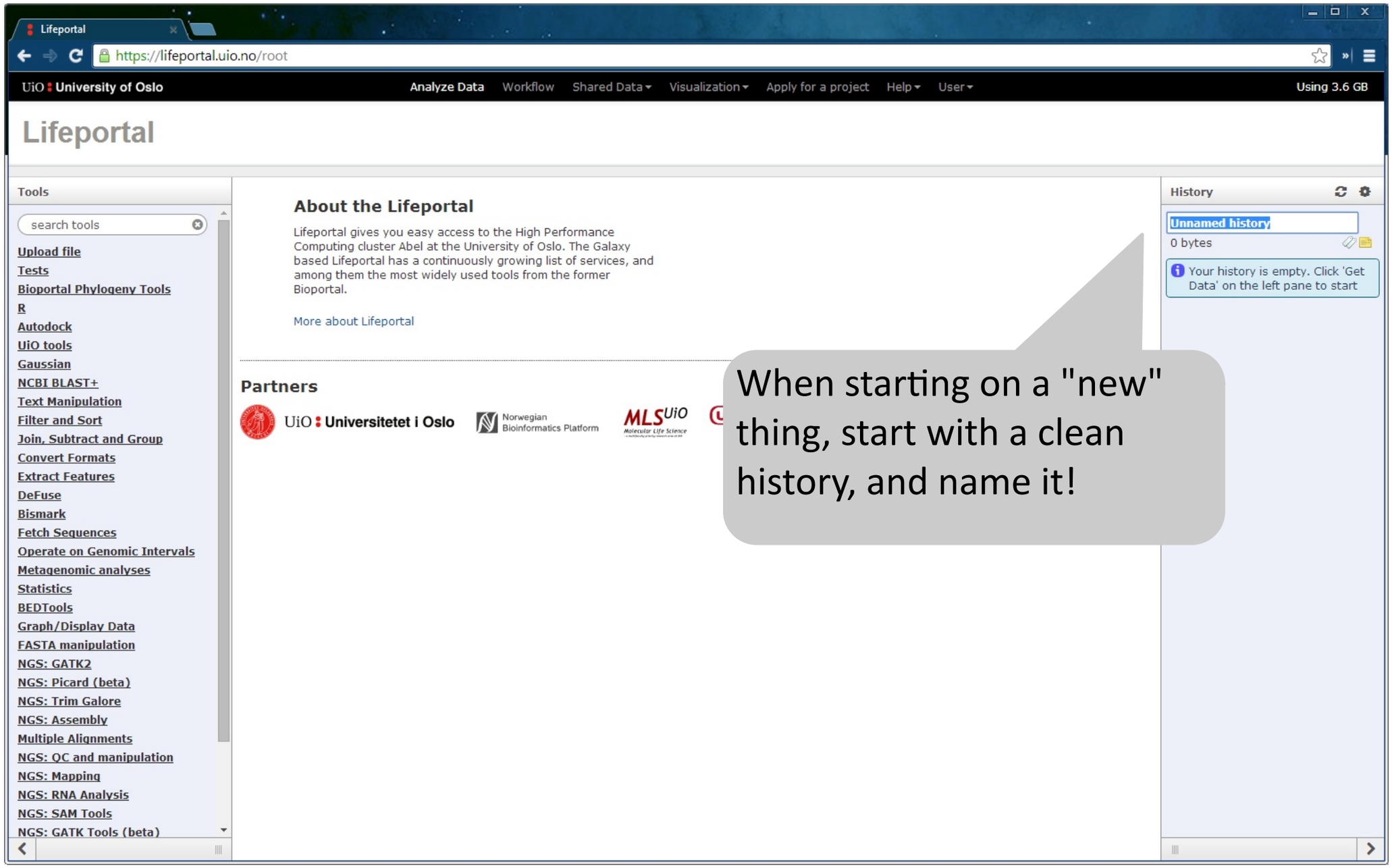
Verify login information



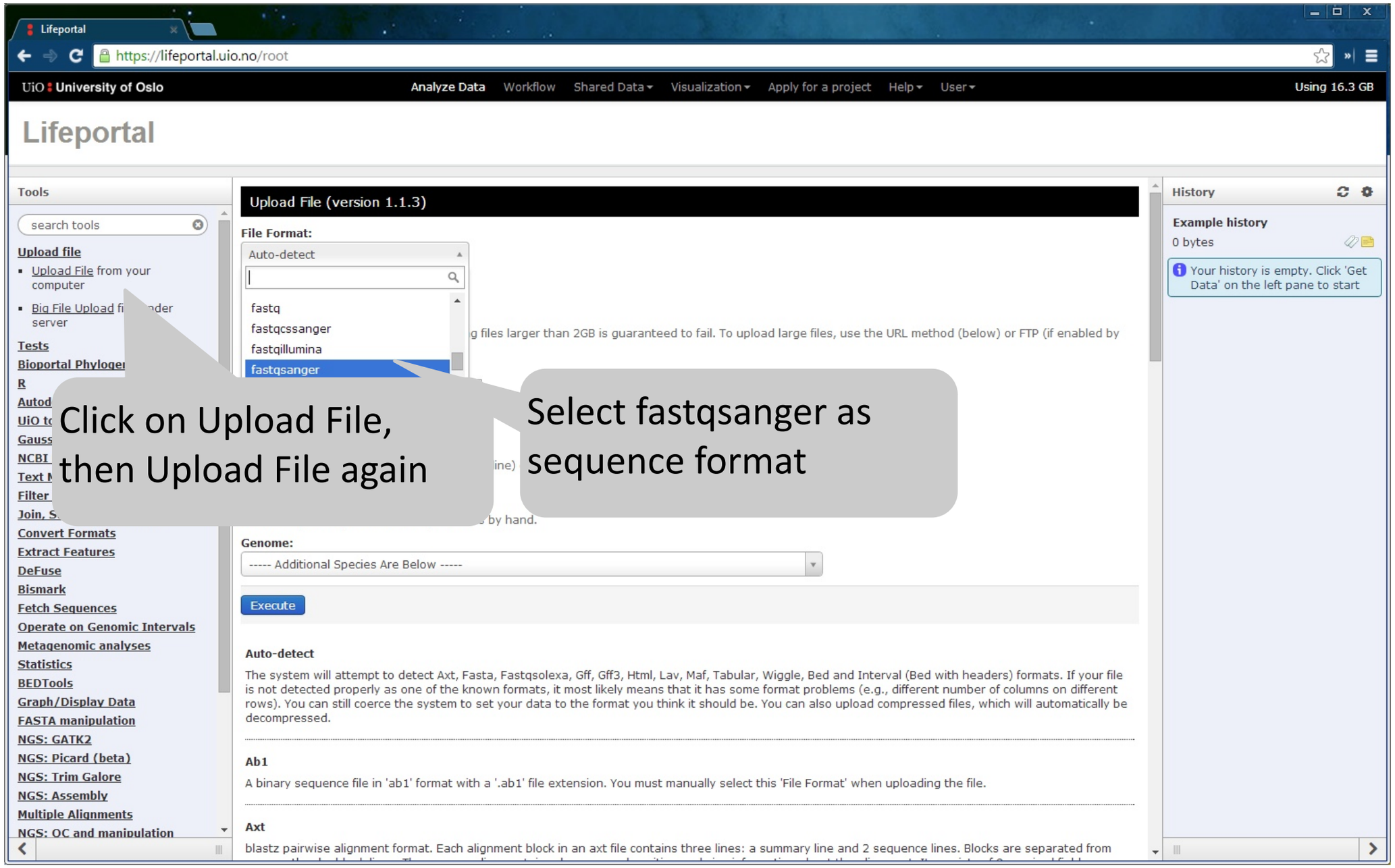
Page orientation



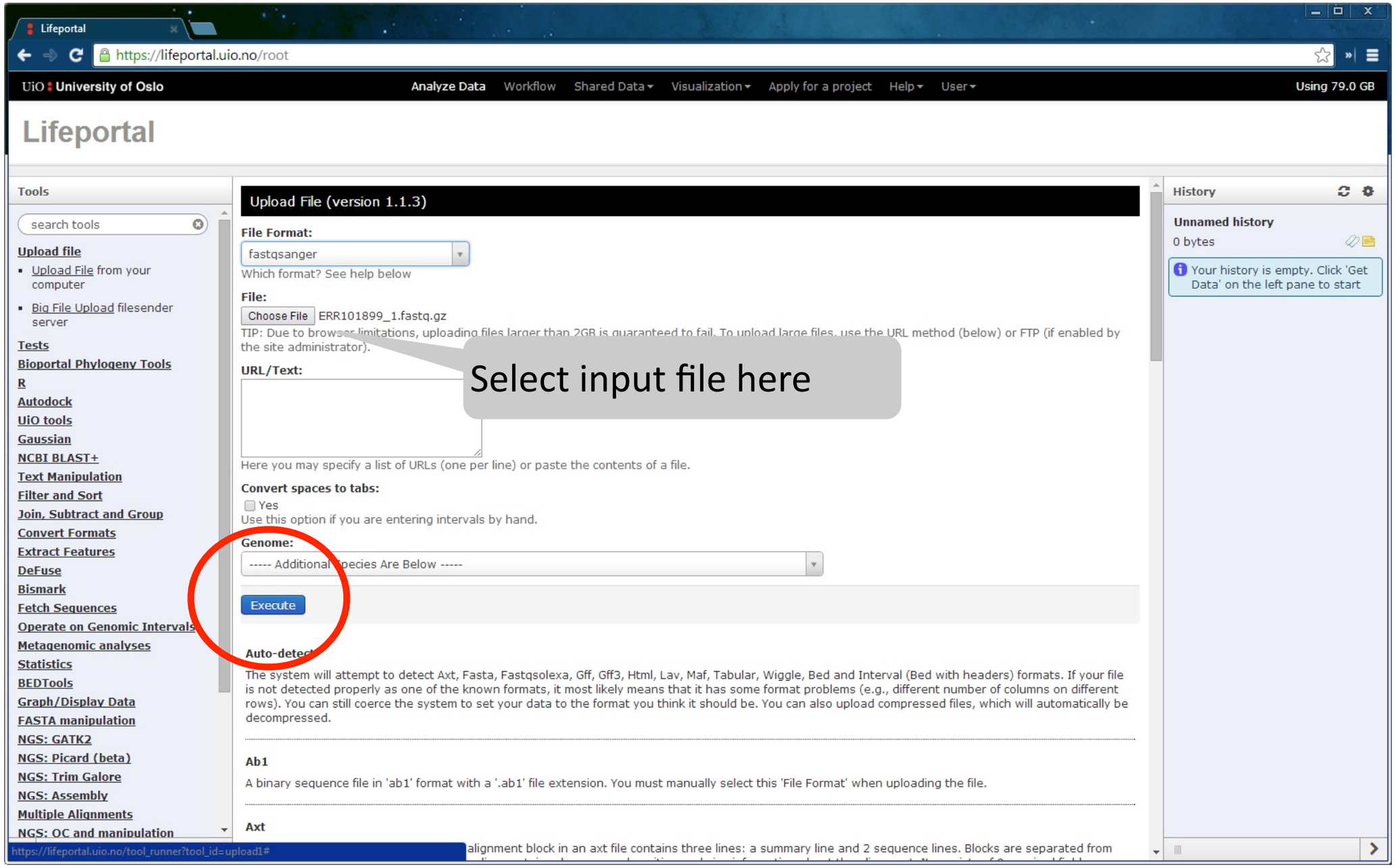
Create a new history



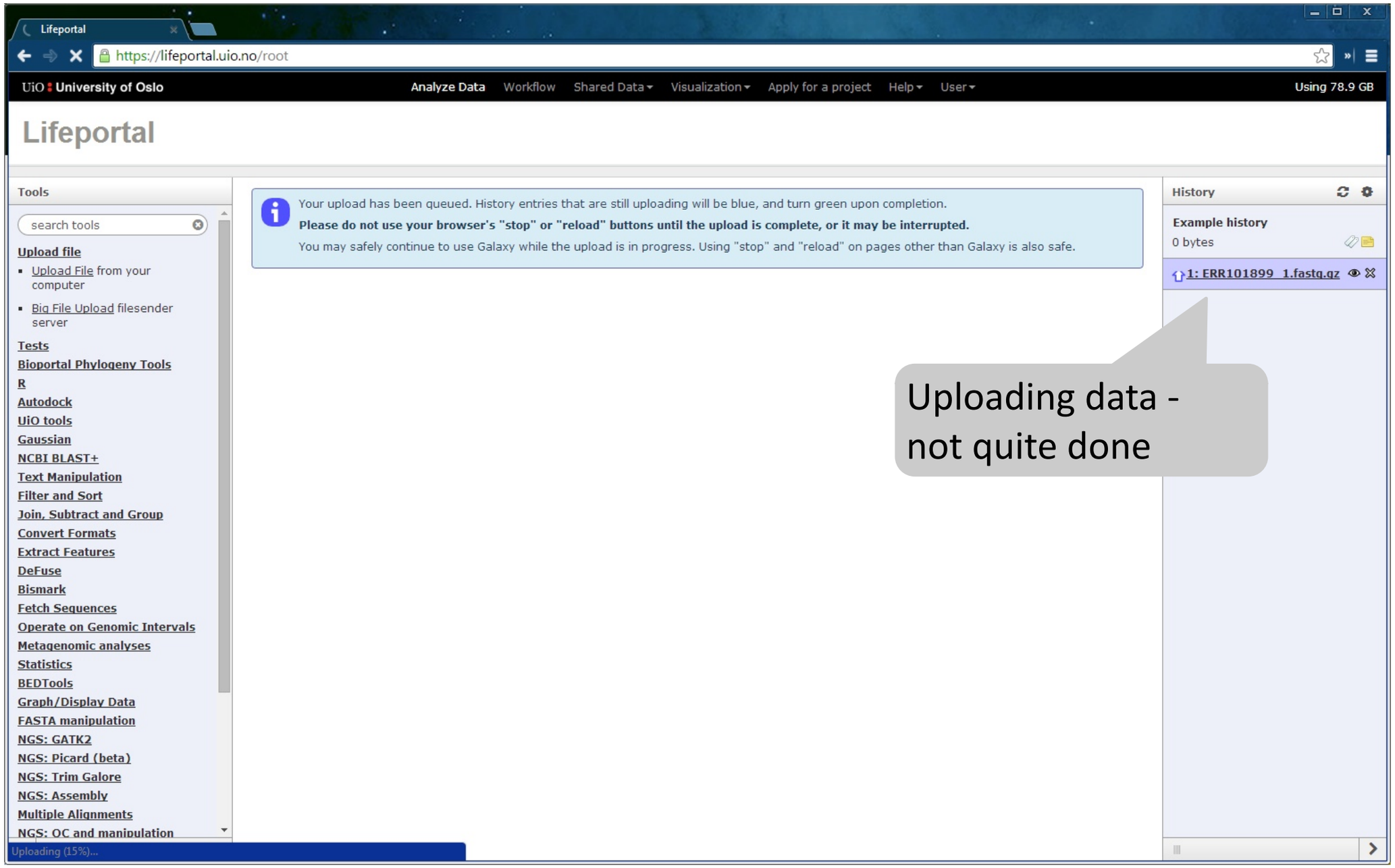
Getting data: uploading



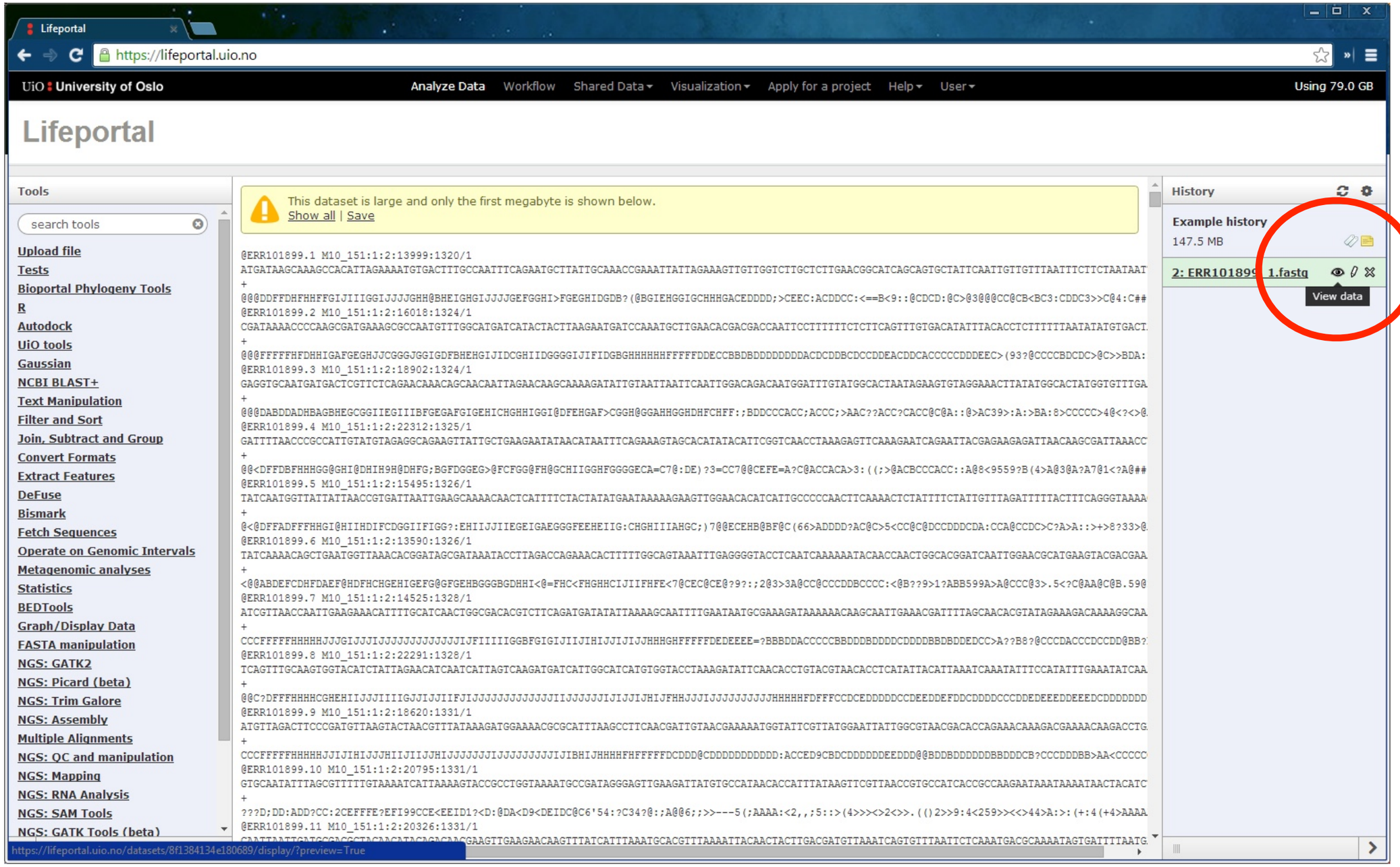
Uploading data



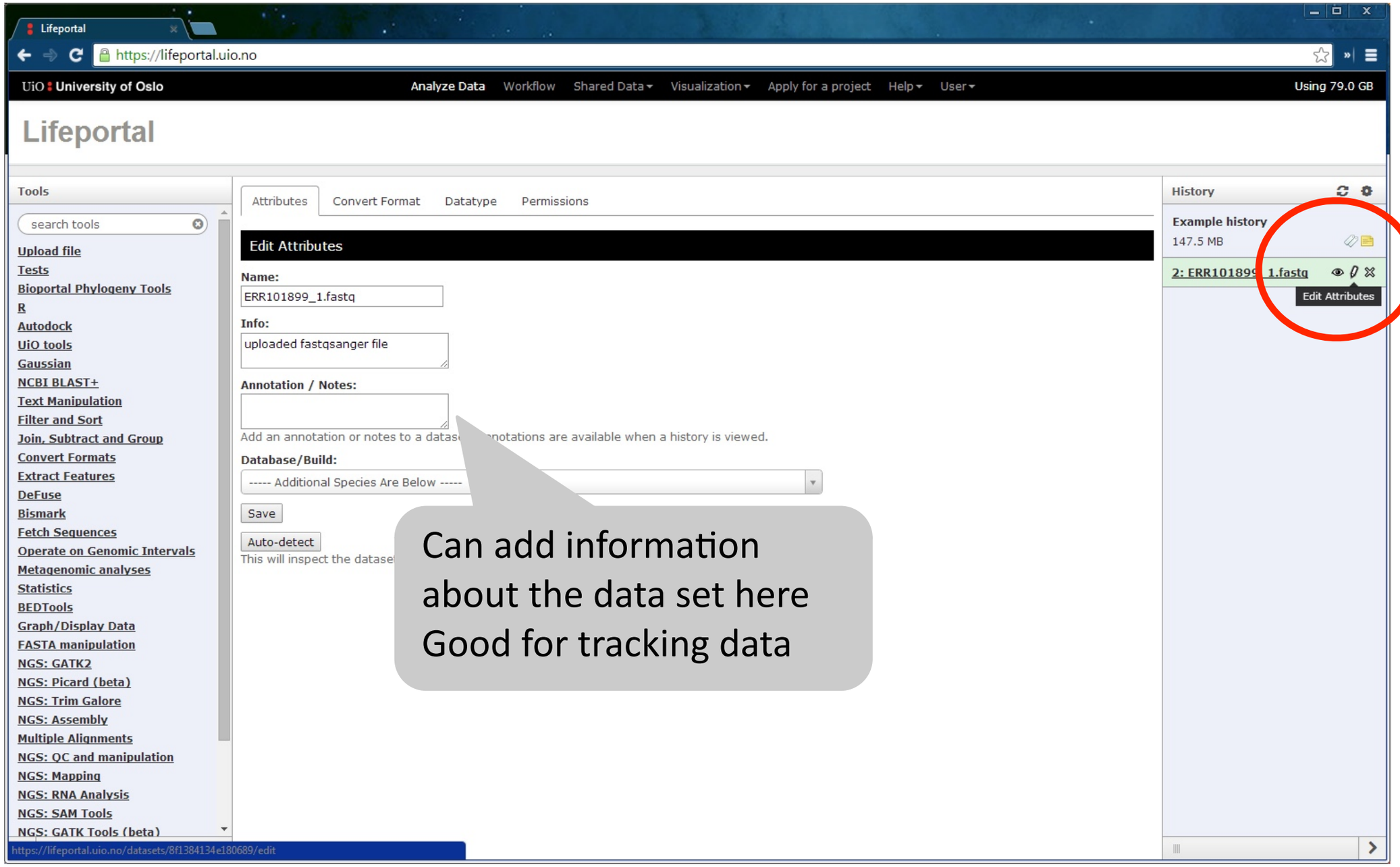
Uploaded data



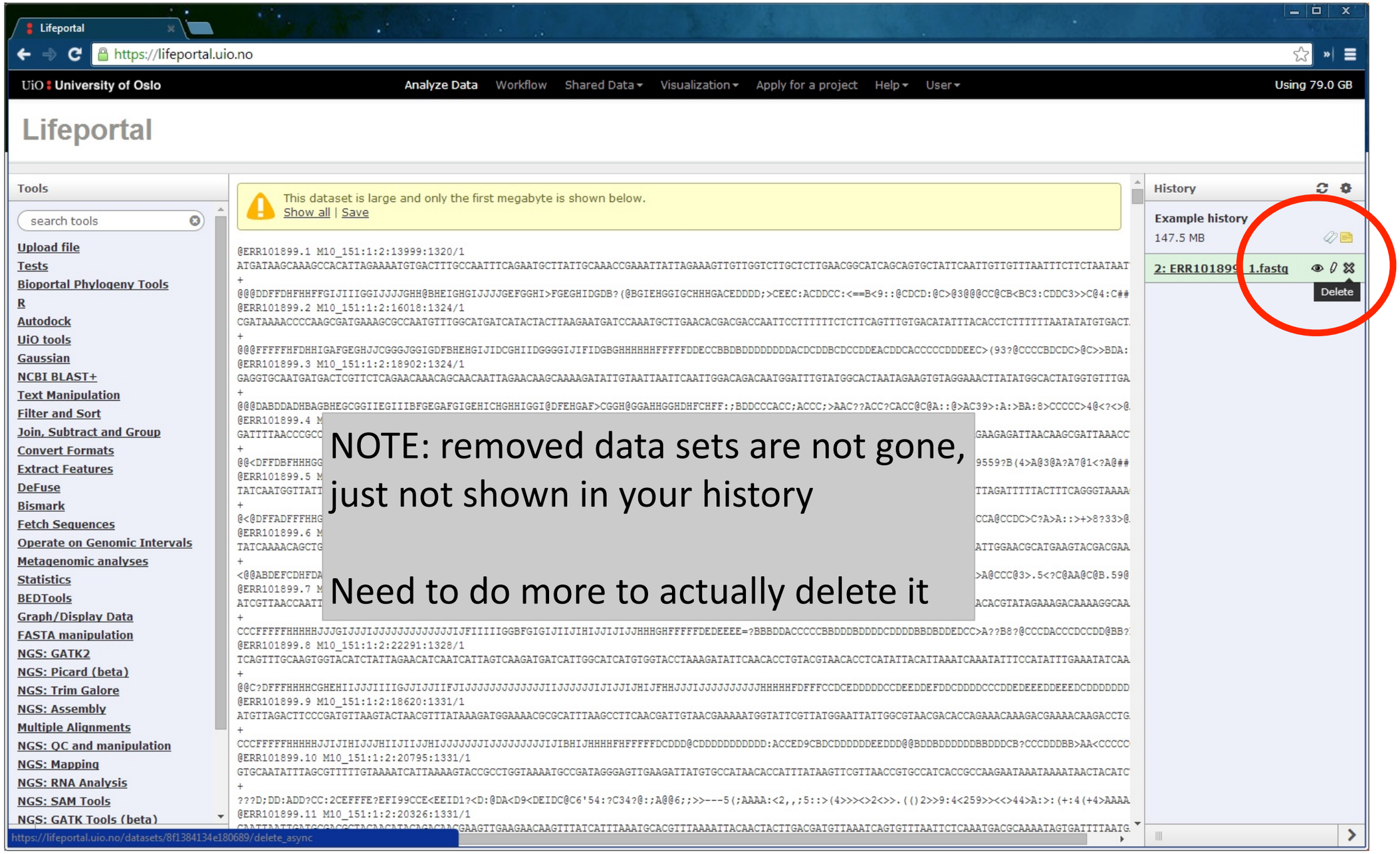
Look at data - eye symbol



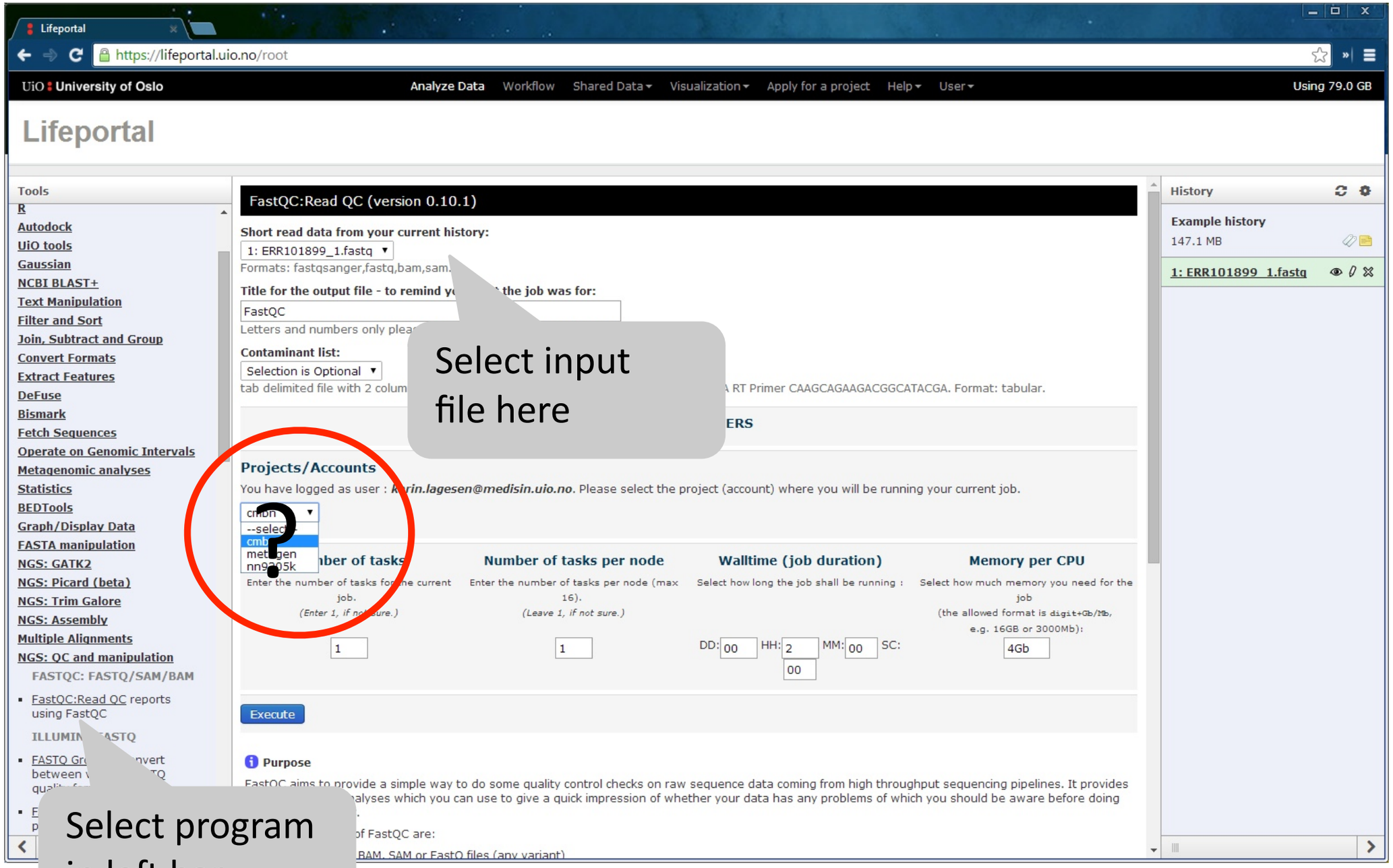
Data annotation - pen symbol



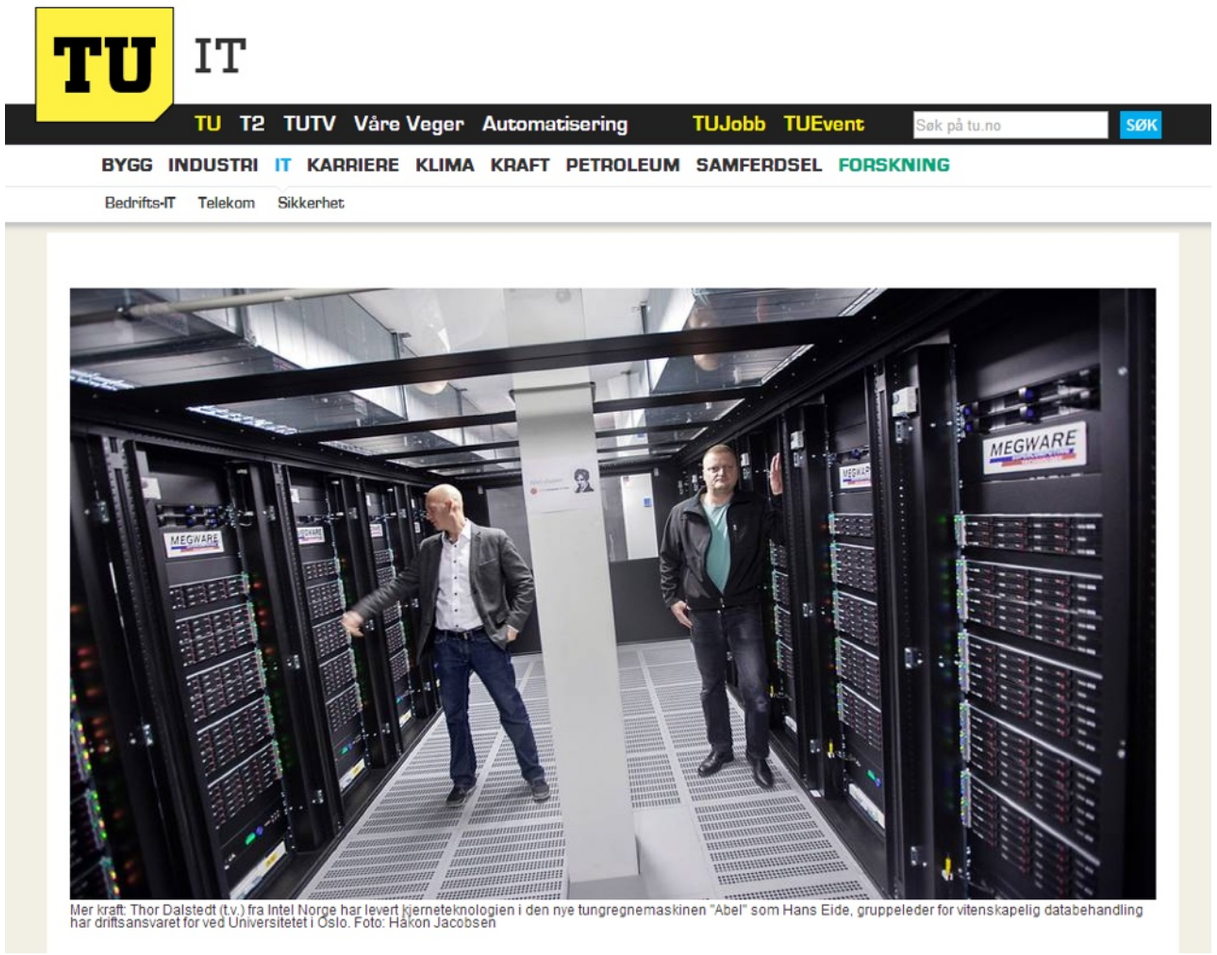
Removing data set - X



Analyzing data

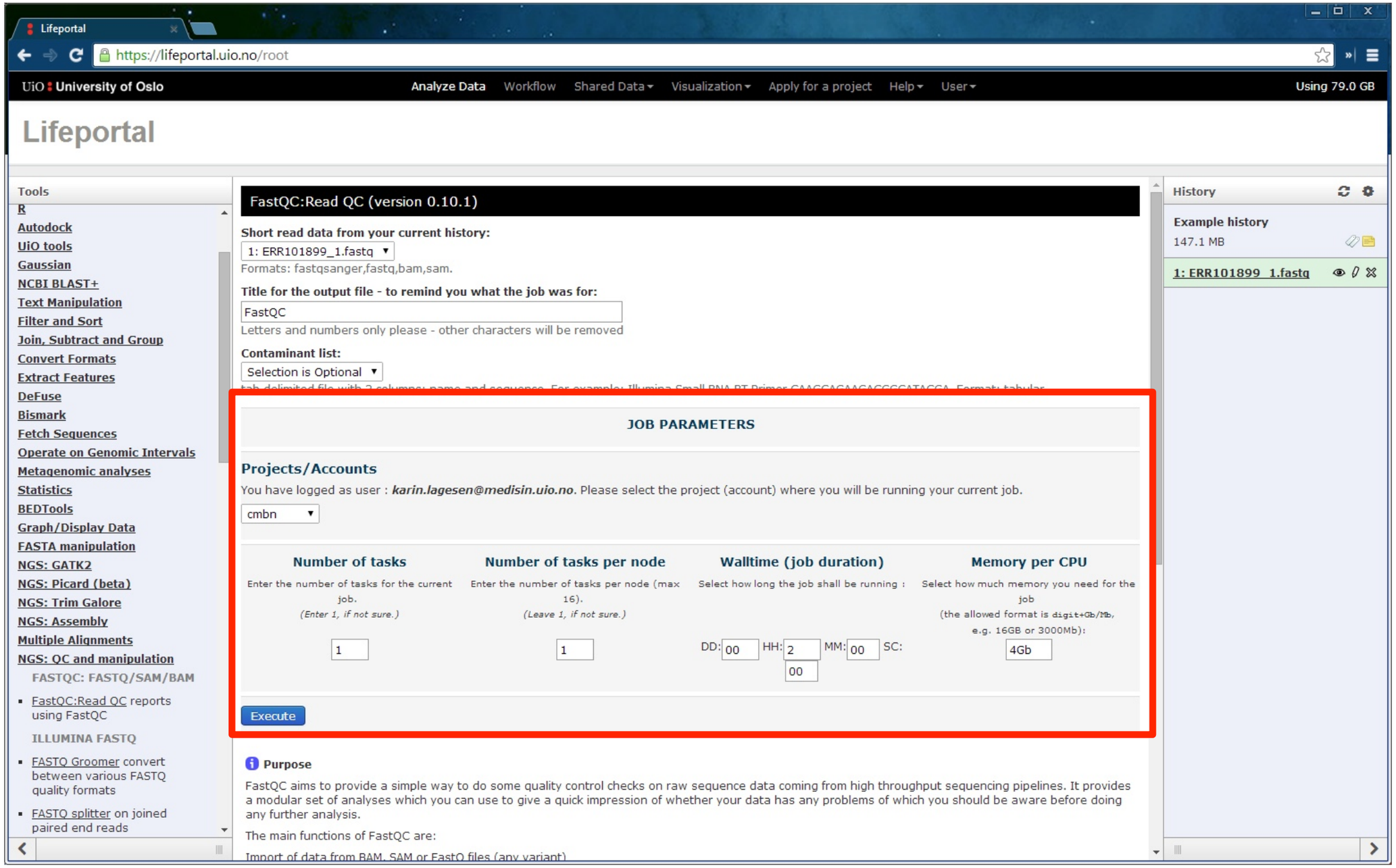


The Abel computer cluster



- 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



Job options

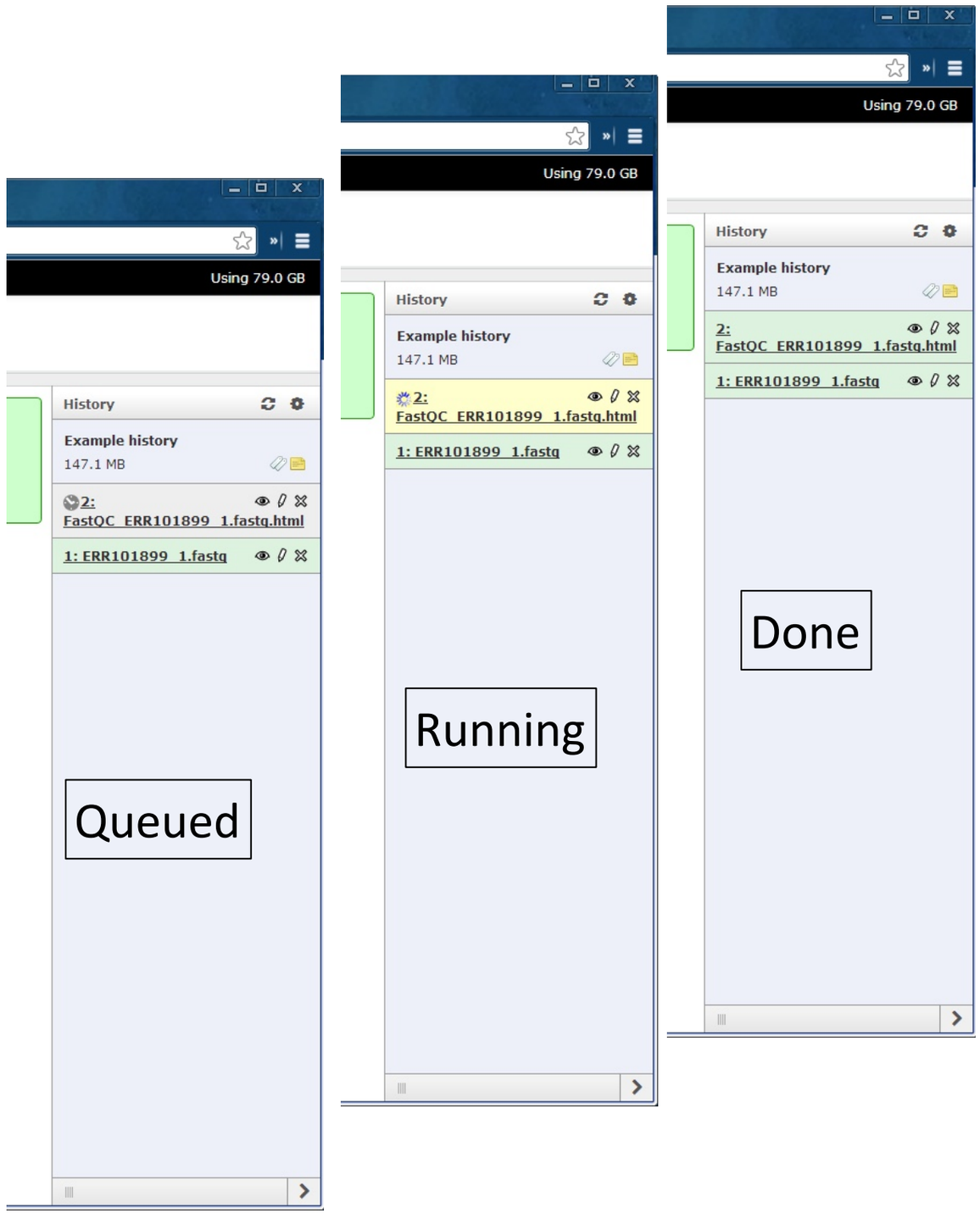
- # tasks = # cores you want to use
- # tasks per node:
  - One node has 16 cores, sometimes programs run faster if all cores are in the same node
- Wall time: guesstimated runtime.
  - Note: jobs exceeding that will be killed!
- Memory per cpu: each CPU has 4 GB of memory - just leave this option

CPU quotas

- Quotas calculated as # CPU hours
- All have 200 hrs to use

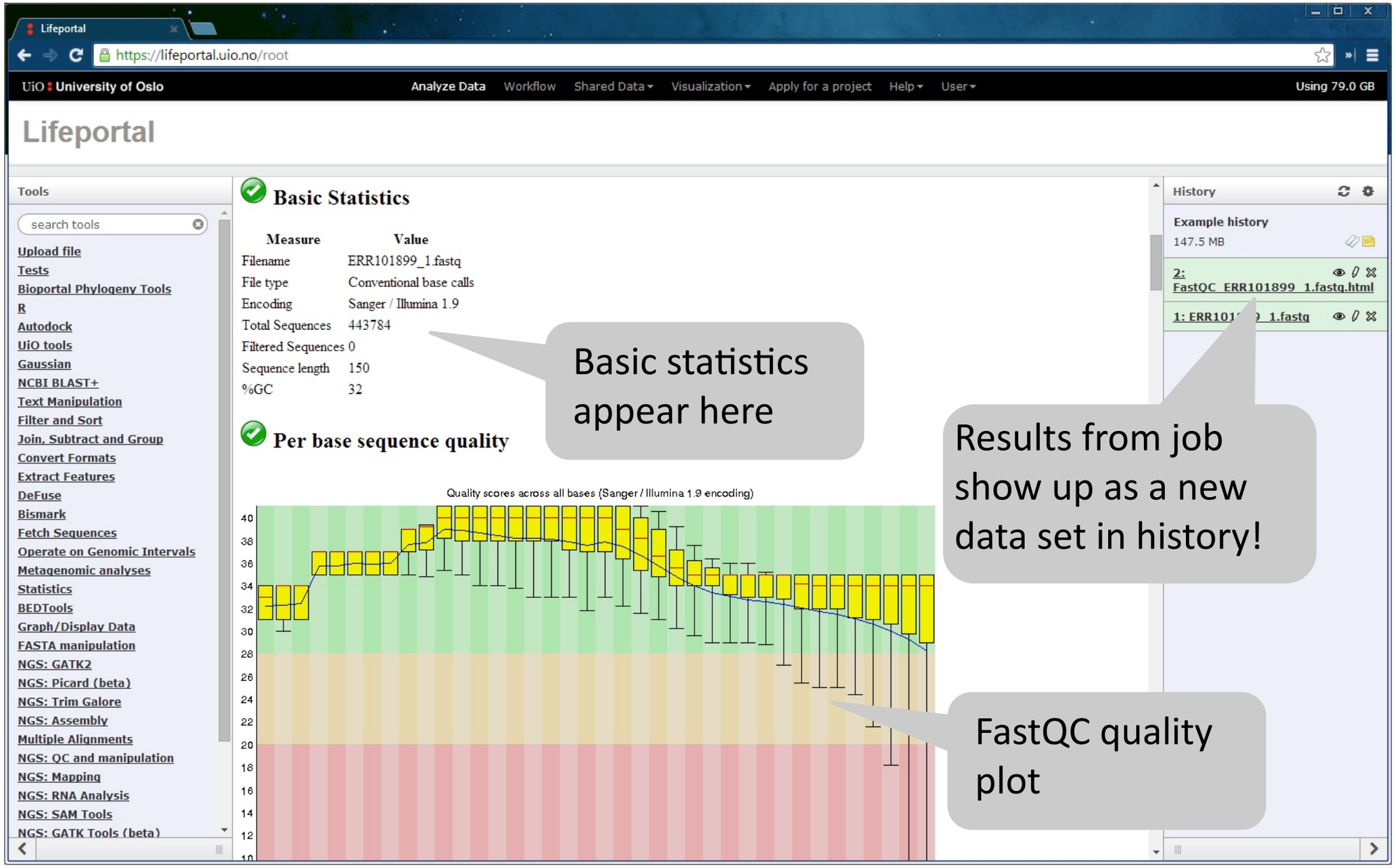
- Big projects should apply for their own quotas

Running job status

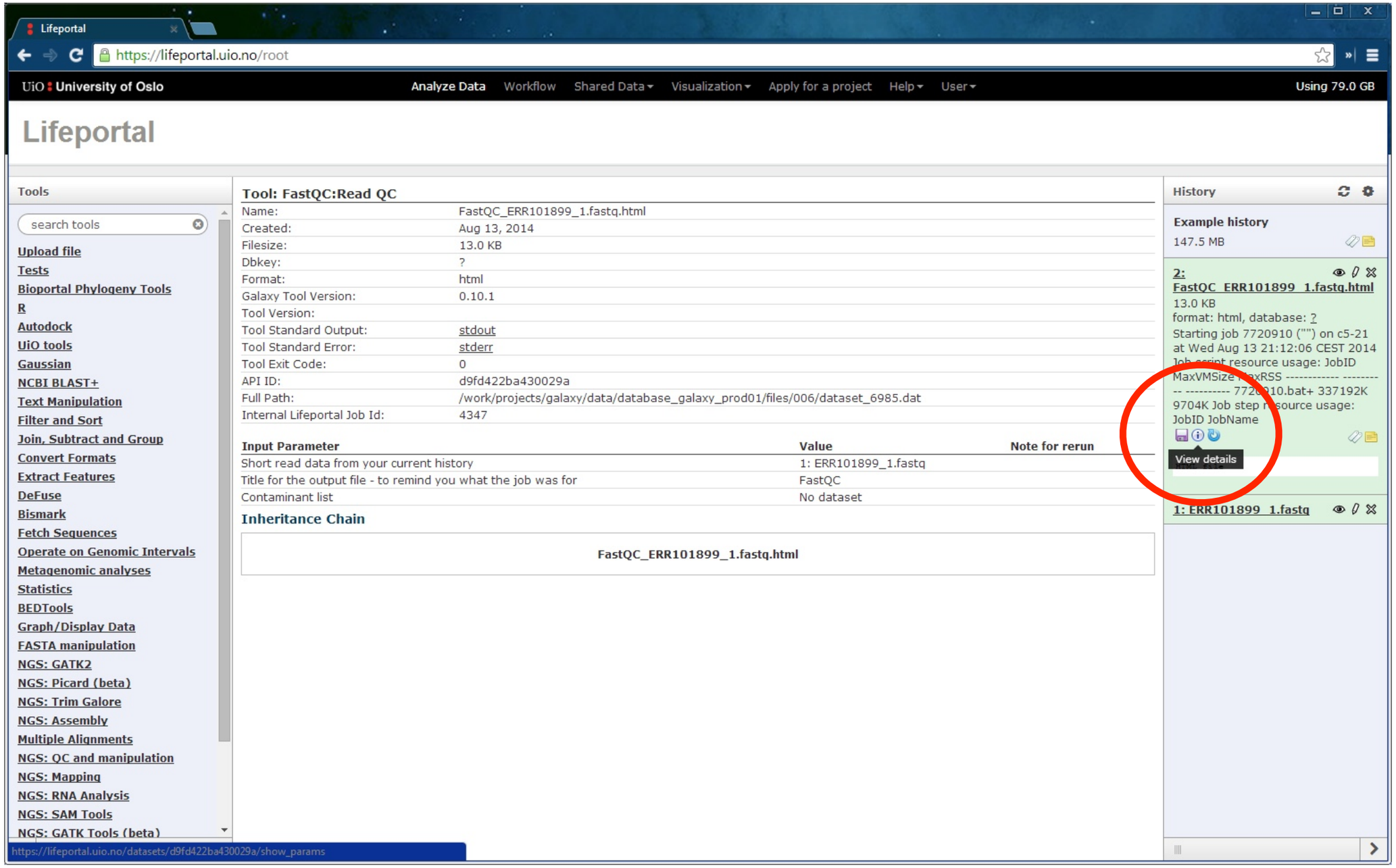


- Colors show the status of the job
- Purple: data uploading
- Gray: analysis queued
- Yellow: running
- Green: done
- Red: error has occured

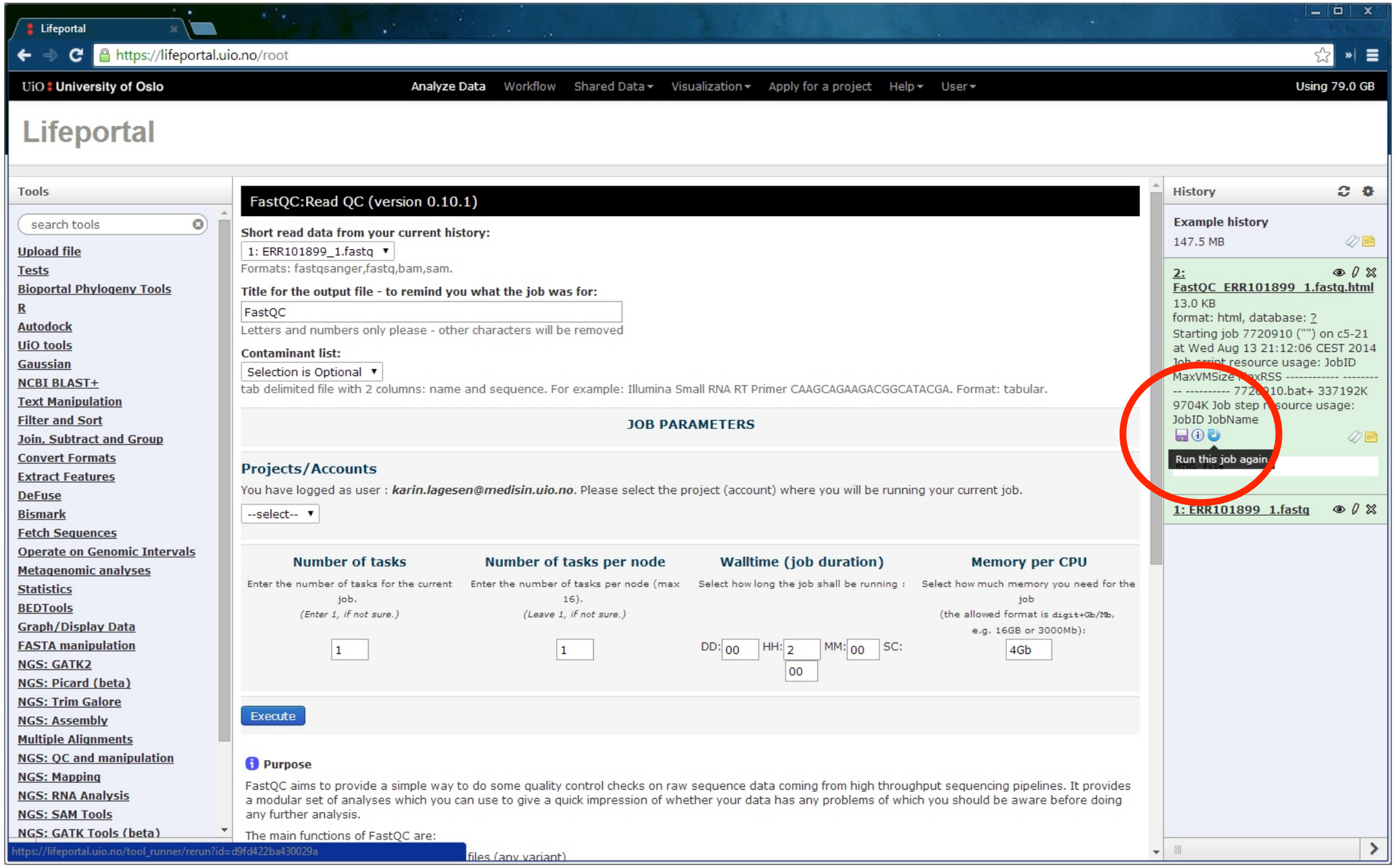
Results show up as new data set!



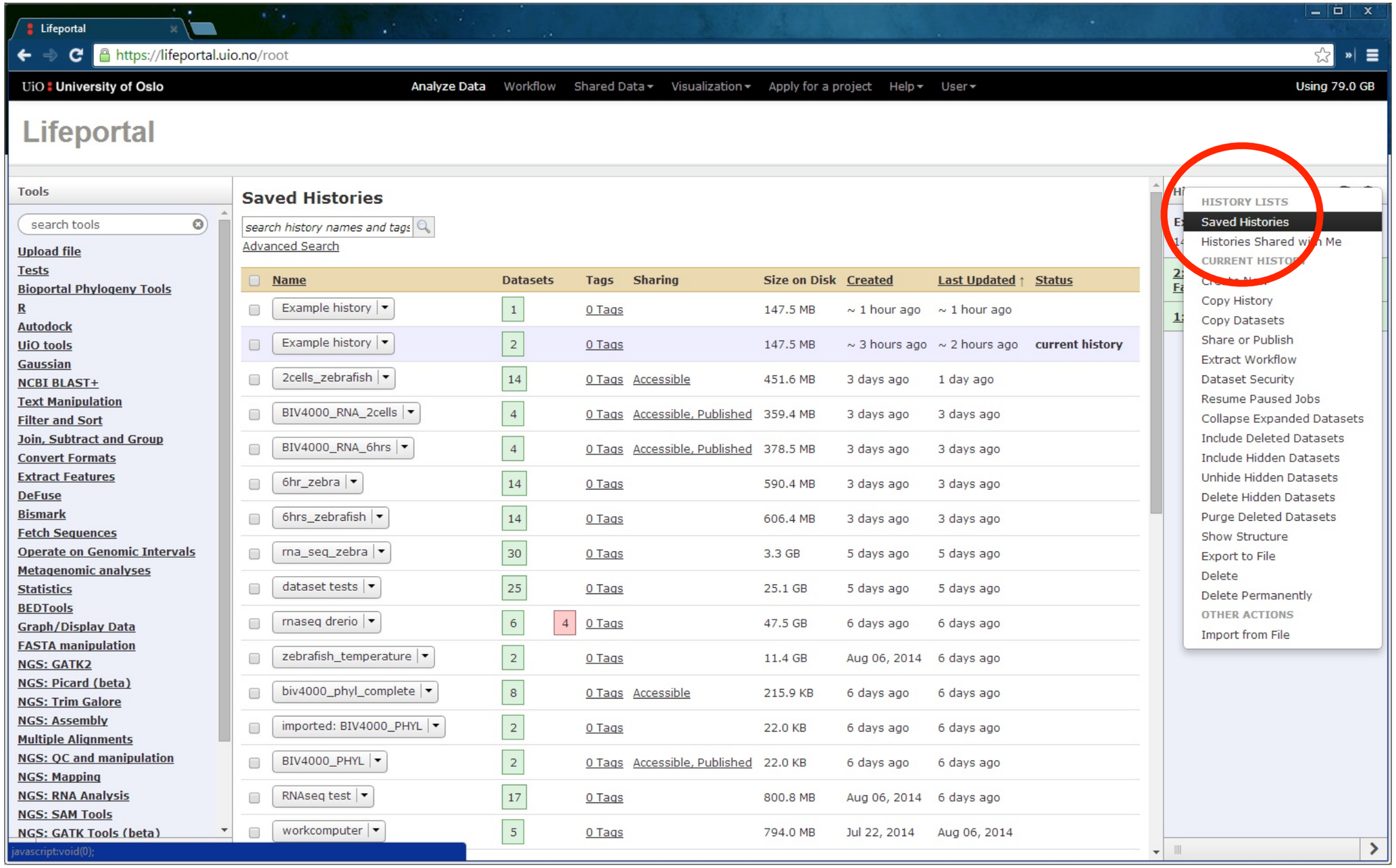
Data sets know how they were made



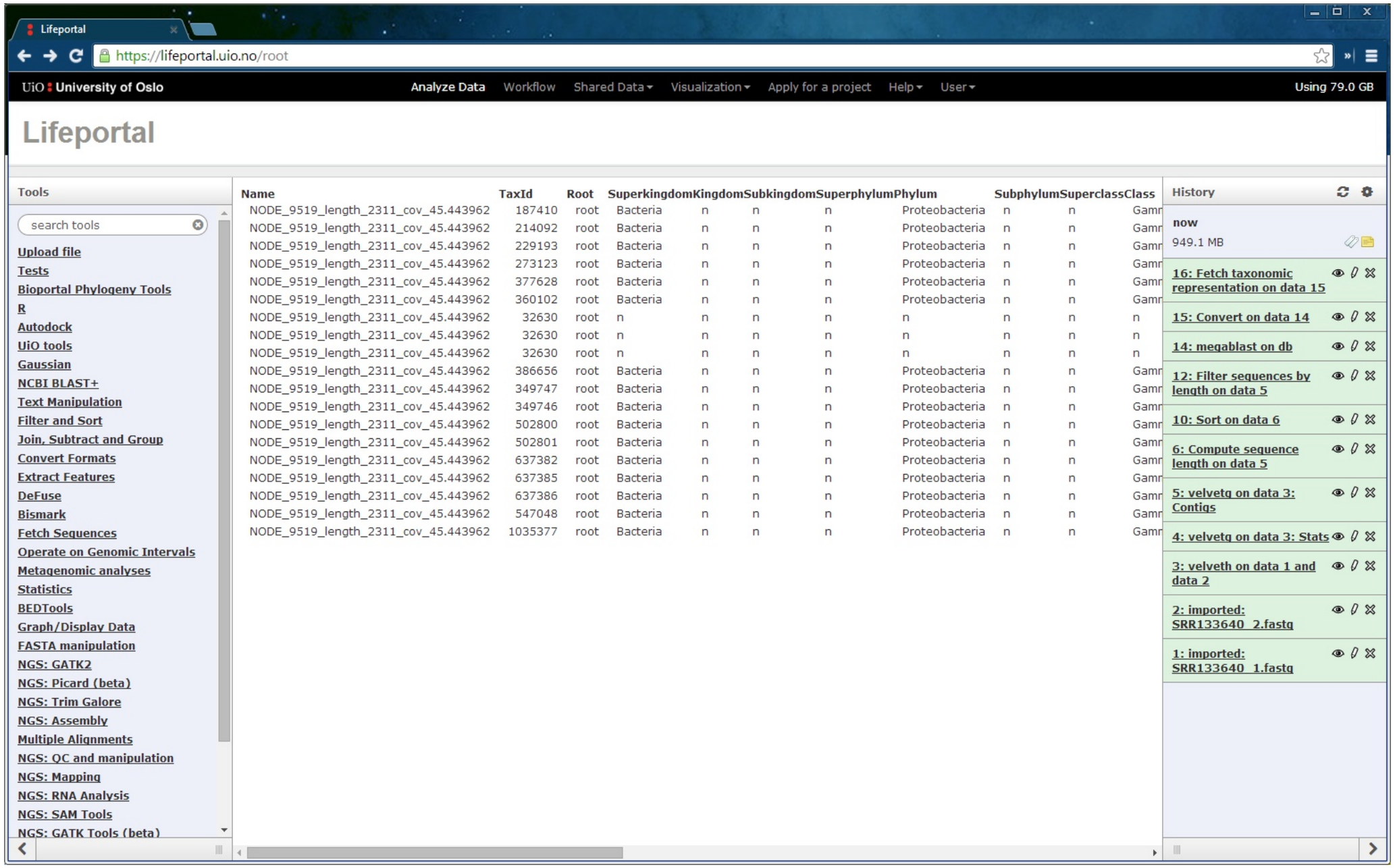
Can easily run analyses again



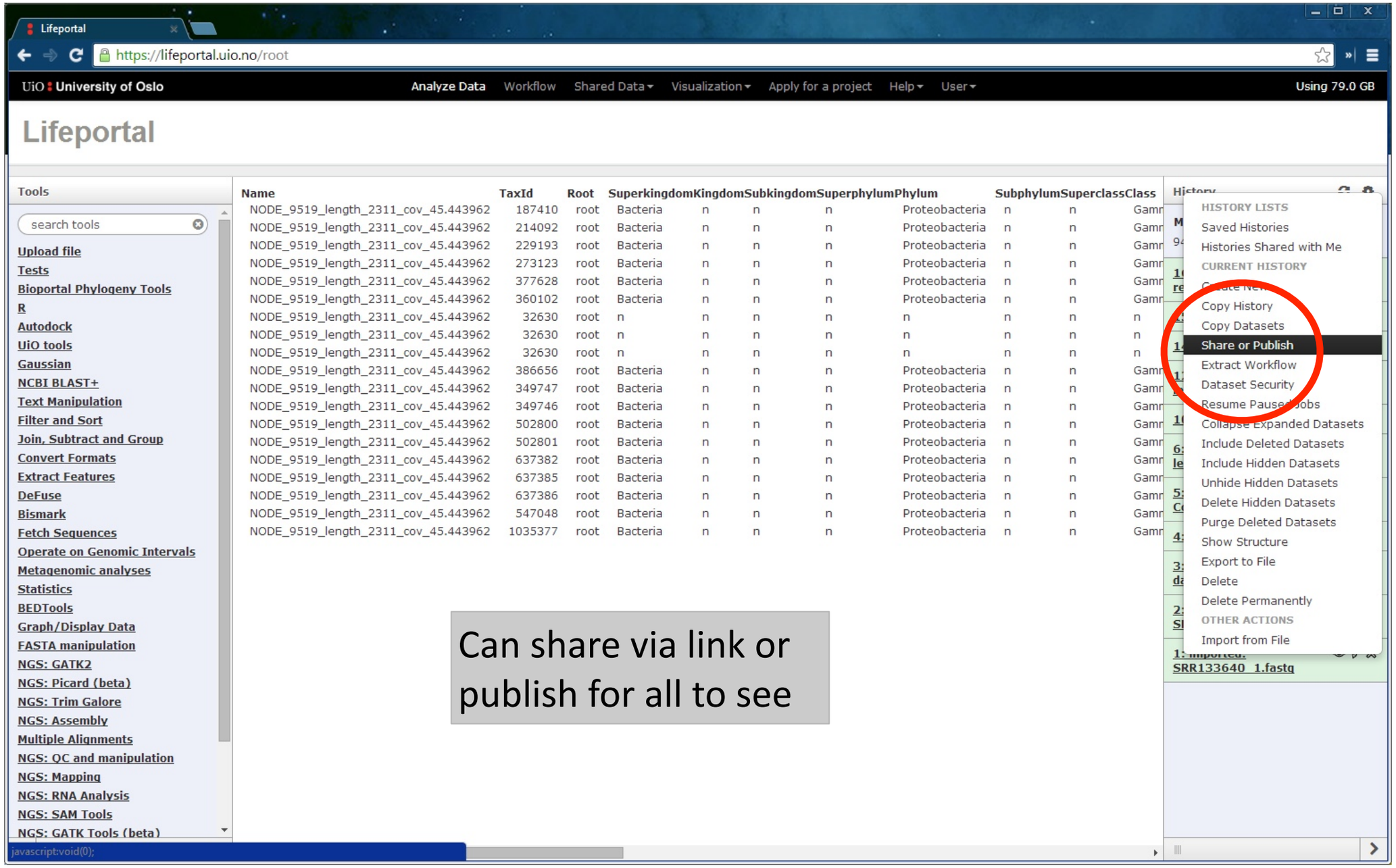
What did I do again....?



Can look at old analyses

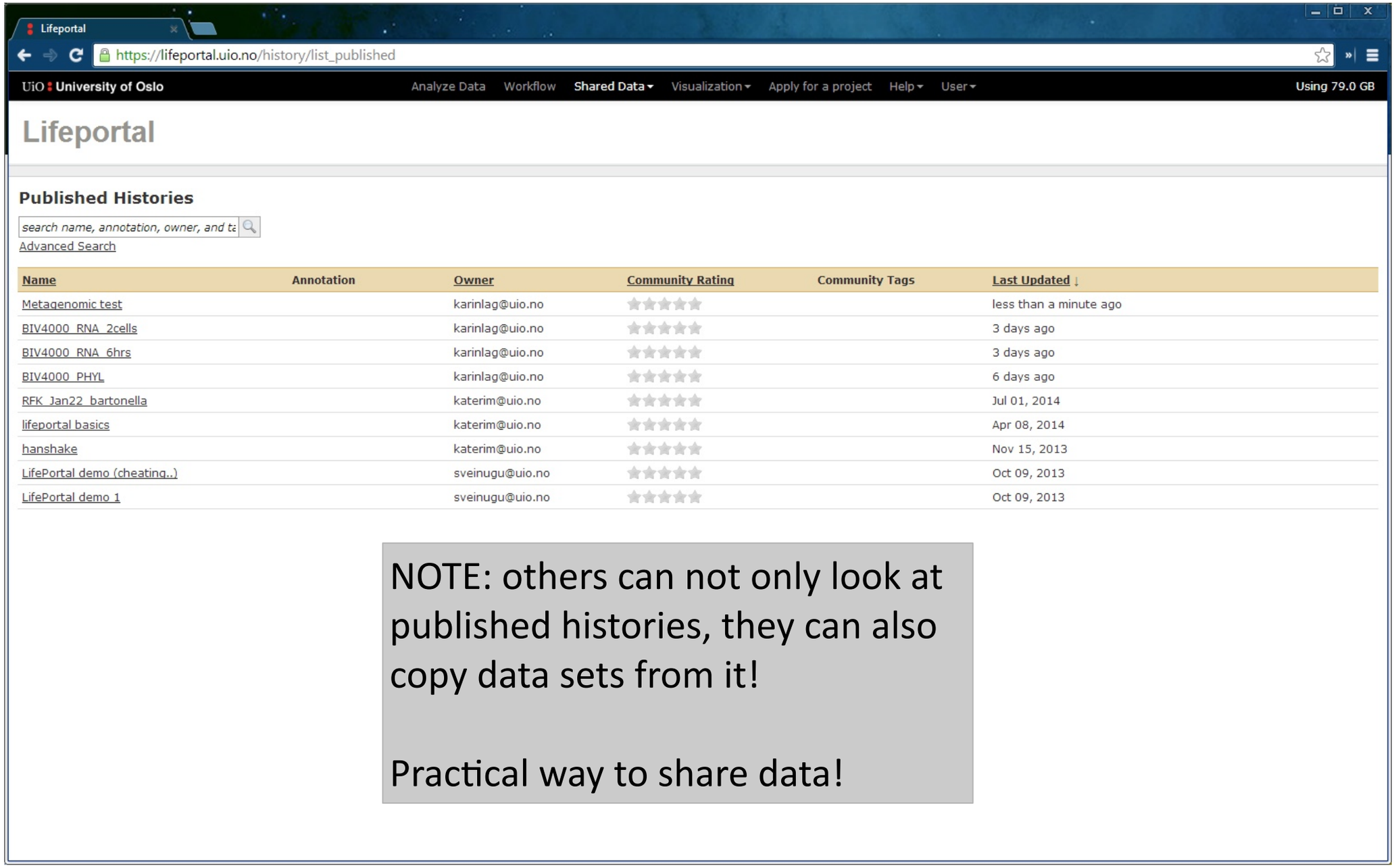


Share or publish histories

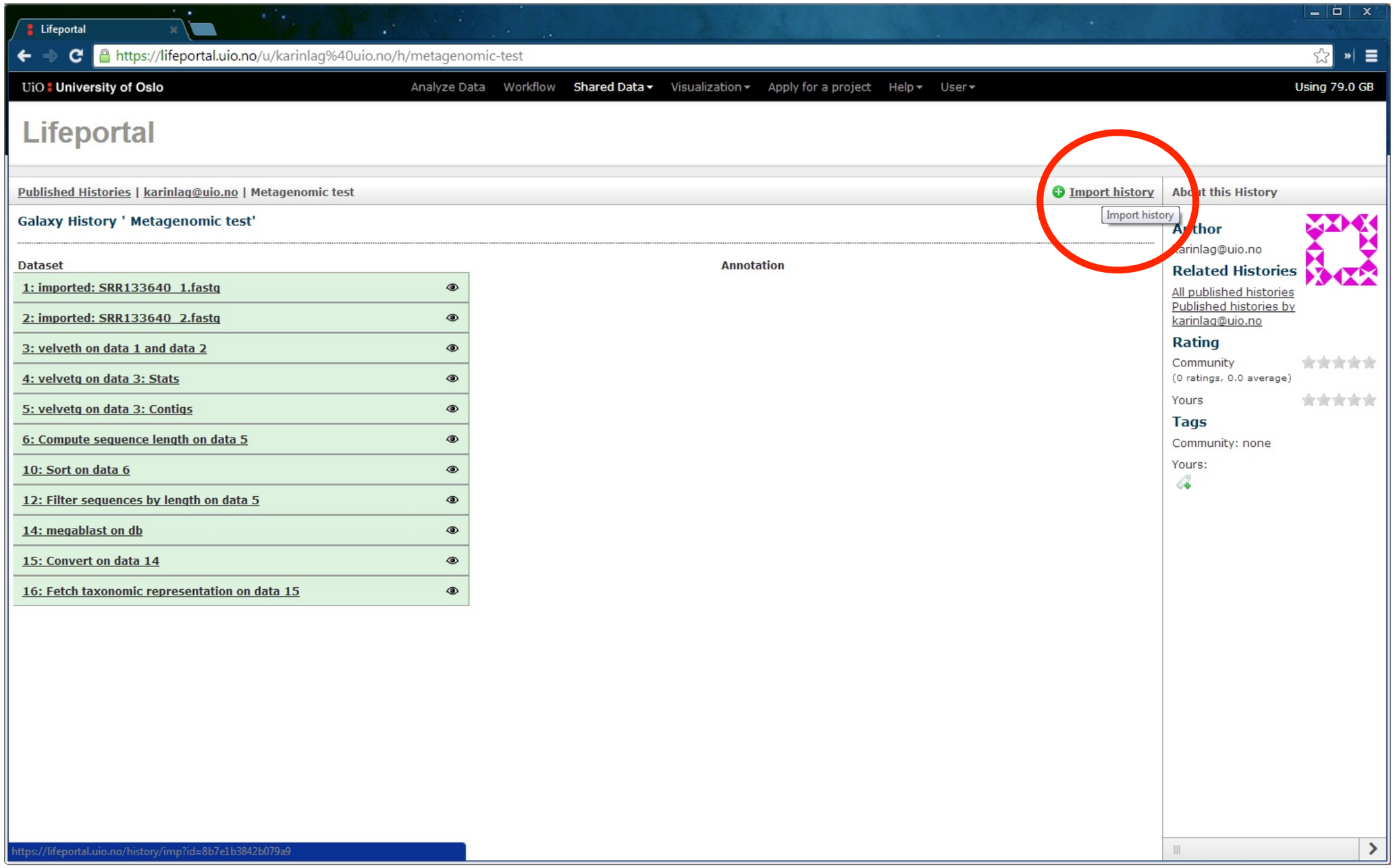


Can share via link or publish for all to see

Published histories open to all



Importing shared histories



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)

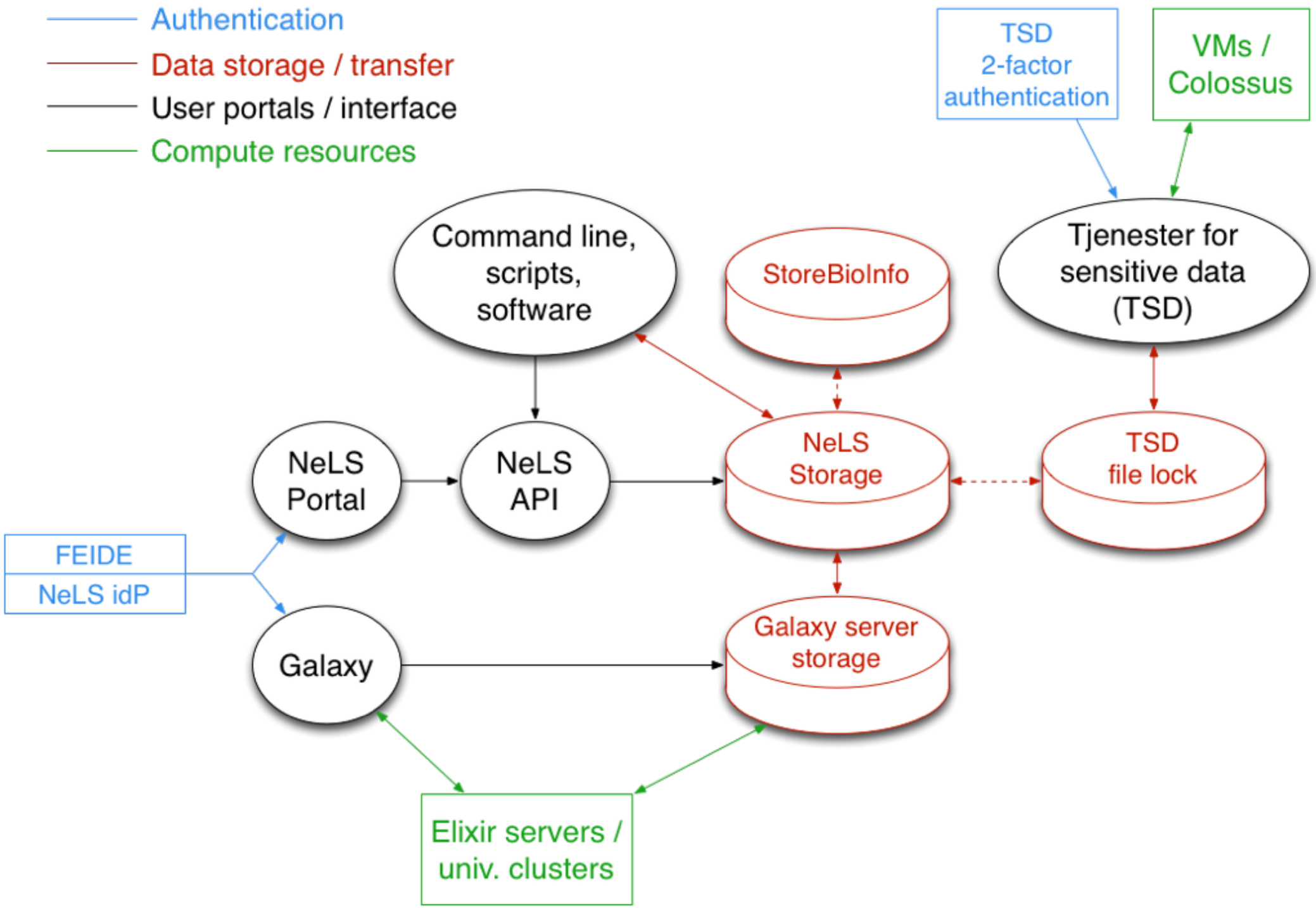
BETA

NeLS

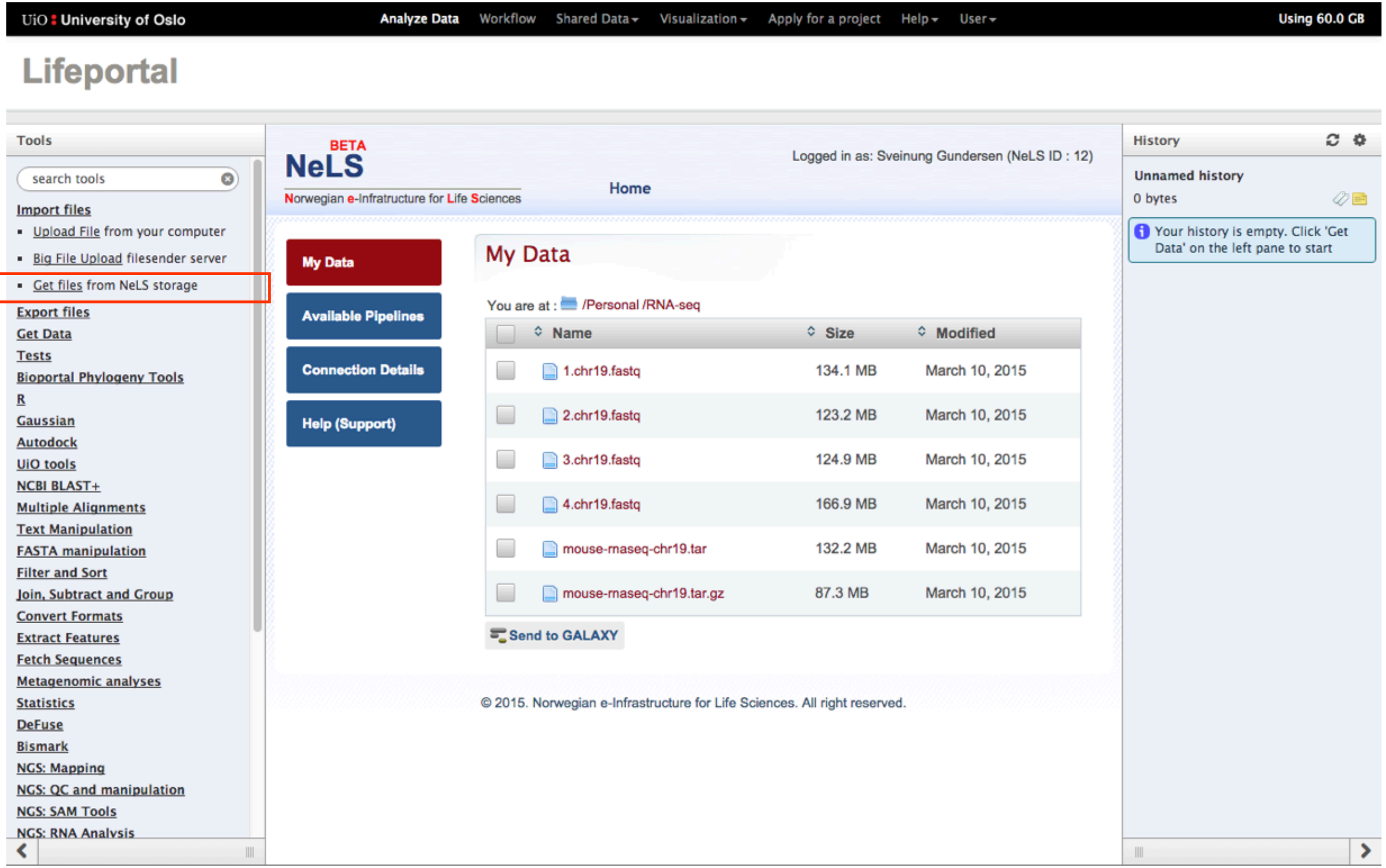
Norwegian e-Infratructure for Life Sciences

- 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



Getting support from

UiO : University of Oslo

and

Lifeportal

elixir NORWAY

- 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