UiO : University of Oslo

How to run HPC jobs through a web-page : *Lifeportal*

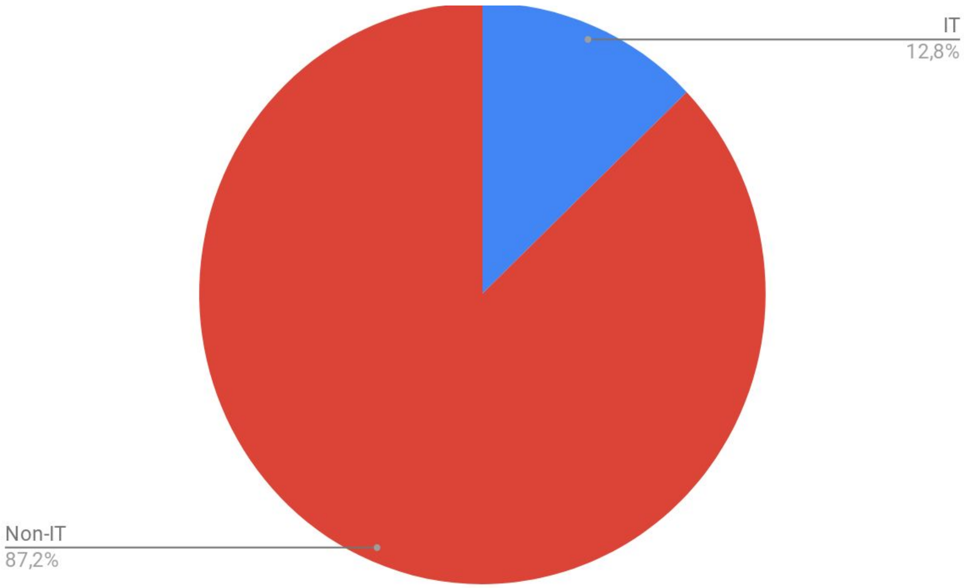
Nikolay Vazov & Sabry Razick

Department for Research Infrastructure Services, University of Oslo

Who are our HPC users ?

- Analysis of 2138 support requests (about 90% non IT)

71	math.uio.no
71	student.matnat.uio.no
80	kjemi.uio.no
83	nhm.uio.no
119	fys.uio.no
139	geo.uio.no
273	medisin.uio.no
274	ifi.uio.no
537	ibv.uio.no



What do they want to achieve

- Mainly analysis of (research) data using existing software

- In varying degree
    - Build and use own software
    - Simulations
    - Statistics
    - Visualization

- Not interested or do not have time to become HPC experts.

- Not all are (computer) technical experts either
    - e.g. May not know whether the software they are using is designed for parallel processing or not

Lifeportal - code base

- Galaxy* based

  Galaxy is an open source, web-based platform for data intensive biomedical research. Now used in other fields as well

- Several in-house (UiO) customizations to meet the requirements of security (*authentication* and *authorisation*) and *resource allocation*

*The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

Lifeportal - prerequisites and access

- A decent web browser adequate to get started
    - Terminal or ssh connection not used (yet api exist as well)
- FEIDE login (via Dataporten)
- Facebook, Twitter & LinkedIn (via Dataporten)
- All Dataporten-supported IdPs can be implemented on demand

Lifeportal - a simple way to the HPC

- *Web front-end* to submit jobs to an HPC cluster

  hides the complexity of direct interacting with the cluster - jobs are submitted through a web-form with text fields, checkboxes, dropdown menus, etc.

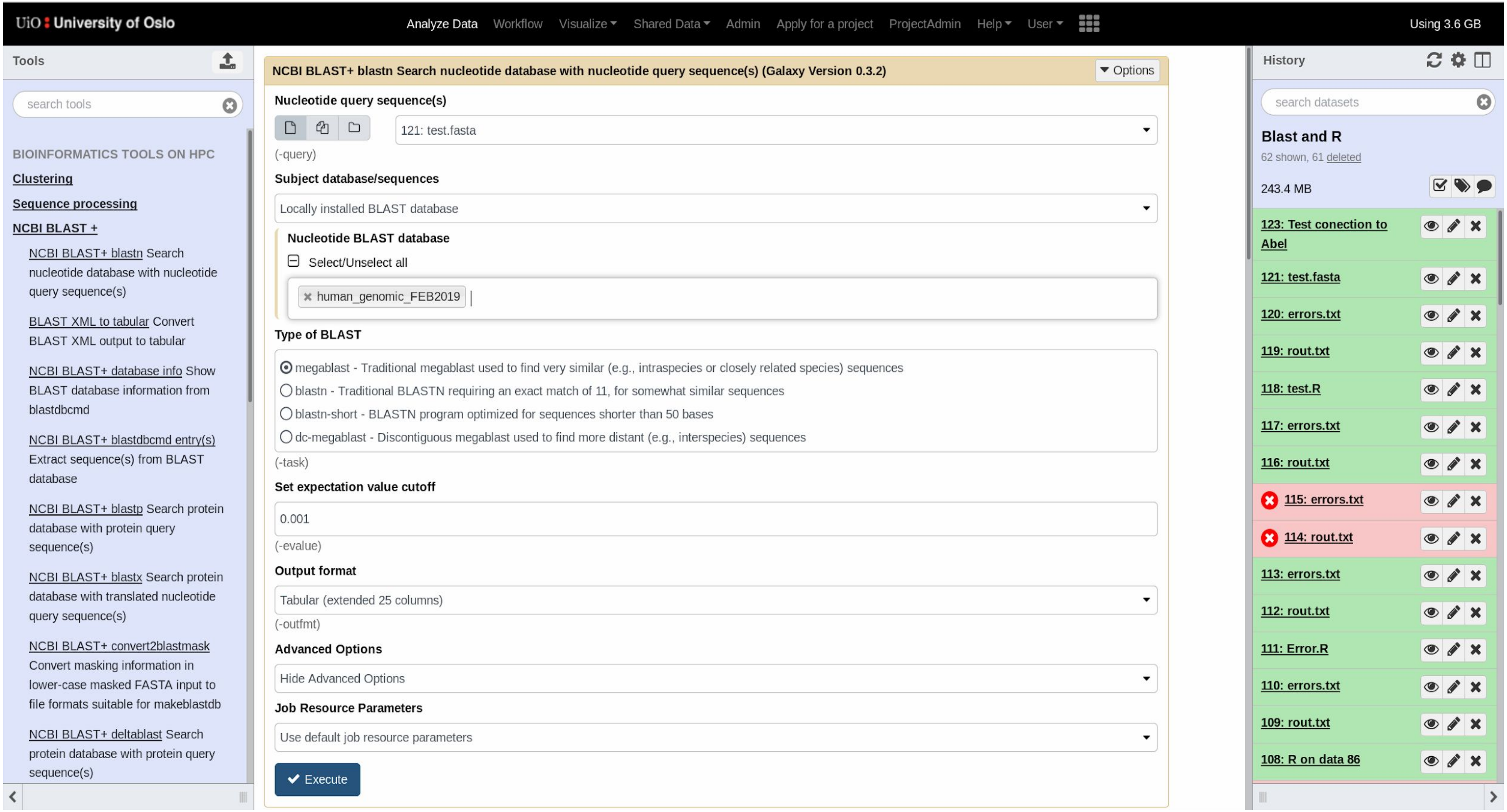
  - tool parameters (tool specific)
  - job parameters (queueing system specific)

- Provides *live job status*

Lifeportal - main features

Quick job start : everything the user needs is in a single visible workspace:

- tools list
- tool menus in a web-form
- history panel : contains all the data necessary to run a job: the uploads, the outputs, parameter files, etc



BIOINFORMATICS TOOLS ON HPC

Clustering

Sequence processing

NCBI BLAST +

- NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
- BLAST XML to tabular Convert BLAST XML output to tabular
- NCBI BLAST+ database info Show BLAST database information from blastdbcmd
- NCBI BLAST+ blastdbcmd entry(s) Extract sequence(s) from BLAST database
- NCBI BLAST+ blastp Search protein database with protein query sequence(s)
- NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
- NCBI BLAST+ convert2blastmask Convert masking information in lower-case masked FASTA input to file formats suitable for makeblastdb
- NCBI BLAST+ deltablast Search protein database with protein query sequence(s)

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 0.3.2)

Nucleotide query sequence(s)

121: test.fasta

(-query)

Subject database/sequences

Locally installed BLAST database

Nucleotide BLAST database

⊟ Select/Unselect all

× human_genomic_FEB2019

Type of BLAST

● megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences

○ blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences

○ blastn-short - BLASTN program optimized for sequences shorter than 50 bases

○ dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

0.001

(-evalue)

Output format

Tabular (extended 25 columns)

(-outfmt)

Advanced Options

Hide Advanced Options

Job Resource Parameters

Use default job resource parameters

Blast and R

62 shown, 61 deleted

243.4 MB

- 123: Test conection to Abel
- 121: test.fasta
- 120: errors.txt
- 119: rout.txt
- 118: test.R
- 117: errors.txt
- 116: rout.txt
- 115: errors.txt
- 114: rout.txt
- 113: errors.txt
- 112: rout.txt
- 111: Error.R
- 110: errors.txt
- 109: rout.txt
- 108: R on data 86

Lifeportal - main features *bis*

- Supports *multiple versions* of tools for backward compatibility and reusability of all analyses
- Supports *grouping of users into projects* and sub-projects with flexible resource allocation
- Possibility to create *workflows* (pipelines) from a history
- Efficient *error report system* - sends error logs directly to the user support

Lifeportal - main features *ter*

- Rich and flexible *sharing mechanisms* between users, groups and external users:
  - data
  - job routines
  - workflows
  - entire histories

- *Reusability* mechanisms:
  - allow the users to rerun the same jobs with minor modifications (data & parameter variations)
  - allow the reviewers to reproduce the results in submitted publications

Lifeportal - some data related features

- Data upload and direct data acquisition from data sources
- Connected to national data storage services (NeLS)
- Reference genomes and indices already included.

Thank you

- https://lifeportal.uio.no/
- lifeportal-help@usit.uio.no