

UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Marine metagenomic infrastructure services as driver for research and industrial innovation

Lars Ailo Bongo
Center for Bioinformatics,
University of Tromsø – The Arctic University of Norway



WP6 Organization

- WP leaders:
 - Nils P. Willassen (University of Tromsø, Norway)
 - Rob Finn (EMBL-EBI)
- Participants:
 - Calouste Gulbenkian Foundation (Portugal)
 - CCMAR – Center for Marine Sciences (Portugal)
 - CNRS - The National Center for Scientific Research (France)
 - CNR - National Research Council (Italy)



Outline

- Use case overview
 - Background
 - Pilot action
 - Use case
- Meta-pipe
 - Norwegian e-Infrastructure for Life Sciences (NeLS)
 - Galaxy/ supercomputer/ NeLS integration
- Technical workflow in use case

Metagenomics – Environmental Samples Analysis



Marine Metagenomics



- Marine genomics and metagenomics are rapidly expanding
- Need customized data, tools, pipelines, and analysis services for the marine domain



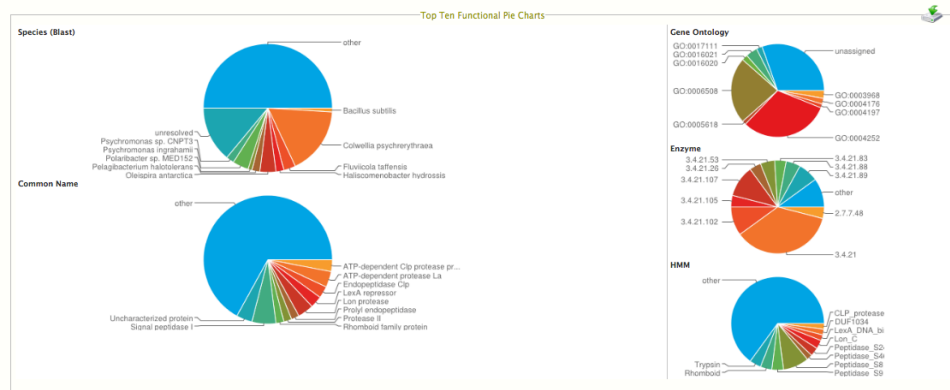
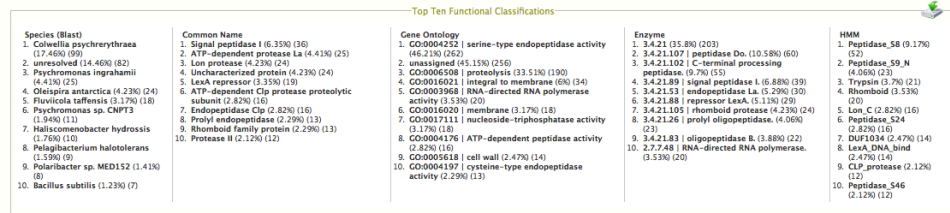
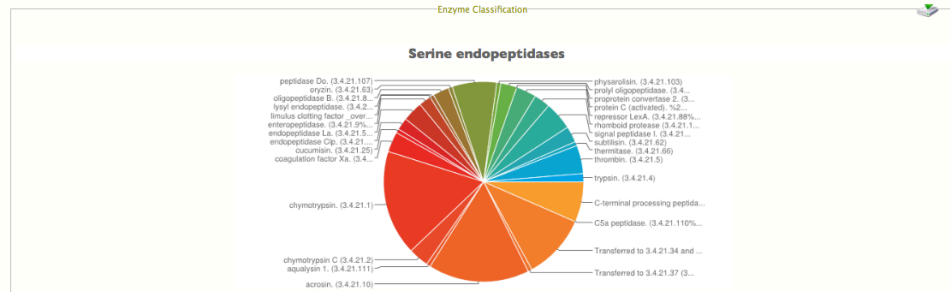
Cold Adapted Enzymes



[QUICK](#) [SEARCH](#) [NEW PROJECT](#) [PROJECTS](#) [LIST](#) [DASHBOARD](#) [LOG OUT](#)
[Projects](#) [View Project](#) [Browse Dataset](#)

Browse Enzymes Sea_Urchin (Sea_Urchin)

- Filter:
- Help:
- Browse Enzymes**
- Oxidoreductases (1,---) (level 1) [9,350 hits]
 - Transferases (2,---) (level 1) [113,723 hits]
 - Hydrolases (3,---) (level 1) [112,300 hits]
 - Acting on ester bonds (3.1,--) (level 2) [3,248 hits]
 - Acting on carbon-phosphorus bonds (3.11,--) (level 2) [9 hits]
 - Glycosylases (3.2,--) (level 2) [1,404 hits]
 - Acting on other bonds (3.3,--) (level 2) [60 hits]
 - Acting on peptide bonds (peptidases) (3.4,--) (level 2) [2,179 hits]
 - Aminopeptidases (3.4.11,--) (level 3) [310 hits]
 - Dipeptidases (3.4.13,--) (level 3) [100 hits]
 - Dipeptidyl-peptidases and tripeptidyl-peptidases (3.4.14,--) (level 3) [51 hits]
 - Peptidyl-dipeptidases (3.4.15,--) (level 3) [38 hits]
 - Serine-type carboxypeptidases (3.4.16,--) (level 3) [97 hits]
 - Metallo-carboxypeptidases (3.4.17,--) (level 3) [67 hits]
 - Cysteine-type carboxypeptidases (3.4.18,--) (level 3) [1 hits]
 - Omega peptidases (3.4.19,--) (level 3) [55 hits]
 - Serine endopeptidases (3.4.21,--) (level 3) [567 hits]**
 - chymotrypsin (3.4.21.1) (level 4) [144 hits]
 - acrosin (3.4.21.10) (level 4) [142 hits]
 - C-terminal processing peptidase (3.4.21.102) (level 4) [55 hits]
 - physaralysin (3.4.21.103) (level 4) [3 hits]
 - rhomboid protease (3.4.21.105) (level 4) [24 hits]
 - peptidase Do (3.4.21.107) (level 4) [60 hits]
 - Transferred to 3.4.21.37 (3.4.21.11) (level 4) [5 hits]
 - Csa peptidase (3.4.21.110) (level 4) [1 hits]
 - aqualysin 1 (3.4.21.111) (level 4) [4 hits]
 - chymotrypsin C (3.4.21.2) (level 4) [29 hits]
 - cucumisin (3.4.21.25) (level 4) [6 hits]
 - prolyl oligopeptidase (3.4.21.26) (level 4) [23 hits]
 - trypsin (3.4.21.4) (level 4) [11 hits]
 - thrombin (3.4.21.5) (level 4) [37 hits]
 - lysin endopeptidase (3.4.21.50) (level 4) [7 hits]
 - endopeptidase La (3.4.21.53) (level 4) [30 hits]
 - coagulation factor Xa (3.4.21.6) (level 4) [28 hits]
 - subtilisin (3.4.21.62) (level 4) [20 hits]
 - oryzin (3.4.21.63) (level 4) [3 hits]
 - thermitase (3.4.21.66) (level 4) [4 hits]
 - protein C (activated) (3.4.21.89) (level 4) [1 hits]
 - Transferred to 3.4.21.34 and 3.4.21.35 (3.4.21.8) (level 4) [92 hits]
 - oligopeptidase B (3.4.21.83) (level 4) [22 hits]
 - limulus clotting factor_ overbar_ B (3.4.21.85) (level 4) [2 hits]
 - repressor LexA (3.4.21.88) (level 4) [29 hits]
 - signal peptidase I (3.4.21.89) (level 4) [39 hits]
 - entropепtidase (3.4.21.99) (level 4) [18 hits]
 - endopeptidase Ctp (3.4.21.92) (level 4) [17 hits]
 - proprotein convertase 2 (3.4.21.94) (level 4) [1 hits]
 - Cysteine endopeptidases (3.4.22,--) (level 3) [109 hits]
 - Aspartic endopeptidases (3.4.23,--) (level 3) [113 hits]
 - Metalloendopeptidases (3.4.24,--) (level 3) [436 hits]
 - Threonine endopeptidases (3.4.25,--) (level 3) [37 hits]
 - Acting on carbon-nitrogen bonds, other than peptide bonds (3.5,--) (level 2) [1,220 hits]
 - Acting on acid anhydrides (3.6,--) (level 2) [3,576 hits]
 - Acting on carbon-carbon bonds (3.7,--) (level 2) [40 hits]
 - Acting on halide bonds (3.8,--) (level 2) [22 hits]
 - Lyases (4,---) (level 1) [3,363 hits]
 - Isomerases (5,---) (level 1) [2,458 hits]
 - Ligases (6,---) (level 1) [3,269 hits]



website <http://metarep.cs.ucl.ac.uk/metarep>
 source code <http://github.com/JCVI/METAREP>
 blog <http://blogs.jcvi.org/tag/metarep>
 contact metarep-support@jcvi.org

Marine metagenomics pilot – towards domain specific service

- October 2014 – September 2015
- University of Tromsø and EMBL-EBI
- Start to harmonize our pipelines:
 - EBI Metagenomics Portal: generic metagenomics analysis pipeline
 - UiT META-pipe: marine metagenomics pipeline
- User community for marine metagenomics analysis in ELIXIR

Pilot Deliverables

- **Harmonize existing metagenomics pipelines to ensure interoperability**
- Assess new functionally specialized databases to enhance or enrich pipeline output
- Investigate the use of other approaches for taxonomic assignment, expanding beyond prokaryotic assignments
- **Explore and prototype with EBI Embassy Cloud**
- Report on gap analysis related to establishment of reference genomes for the marine environment
- Collaboration Workshop on Marine Informatics (16-17 March 2015)
- Report about the needs for specific investment in connection to services for the marine sector

Use case: Marine metagenomic infrastructure services as driver for research and industrial innovation

- Objectives:
 - Development and implementation of selected **standards** for the marine domain
 - Development and implementation of **databases** specific for the marine metagenomics
 - Evaluation and implementation of **tools and pipelines** for metagenomics analysis
 - Development of a **search engine** for interrogation of marine metagenomics datasets and establish training workshops for end users

Task 6.1: Data standards for marine domain

- Data format conventions and standards
- Reporting standards
- Validation tools

Task 6.2: Marine specific data resources

- Build high quality marine specific reference databases
 - Sequences from ENA, UniProt, and other datasets, including...
 - ...TaraOceans and Ocean Sampling Day
 - Collaboration with EMBRC and MIRRI
- TaraOceans: 15TB storage + 200 CPU years to calculate

Task 6.3: Gold standards for metagenomics analysis

- Evaluation and implementation of new tools and pipelines
 - Environmental applications
 - Results publically available
- Available through:
 - UiT META-pipe
 - EBI Metagenomics Portal (EBI MGP)
 - EMBL Embassy Cloud
 - ...

Task 6.3: Gold standards for metagenomics analysis

- Web based search engine for interrogation of marine metagenomics results available from EBI MGP
 - Using existing web services
 - Discovery of data through metadata, taxonomic and functional fields
- Expanded comparison tool
 - User selected datasets
 - Functional and taxonomic comparison
 - Federated search over META-pipe and MGP results
 - Identification of common trends and/or differences

Task 6.4: Training workshops for end user

- Collaboration with ELIXIR Training Programme WP
- Exploit provided data, tools, pipelines, and compute infrastructures

META-Pipe

- Our marine metagenomics data analysis pipeline
- Elixir-NO delivery to Elixir
- Integrated with national infrastructure platforms
- Run on local supercomputer
- Deployed as national service

BETA
NeLS

Norwegian e-Infrastructure for Life Sciences

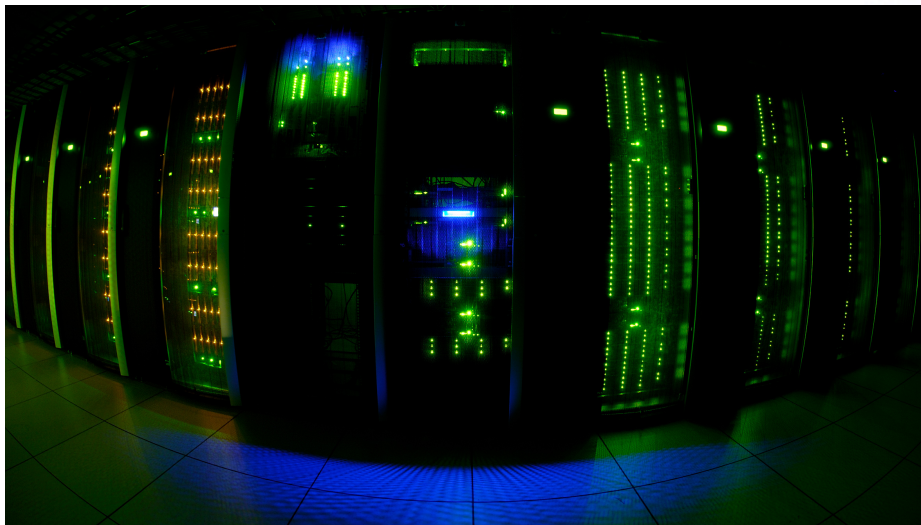


NeLS

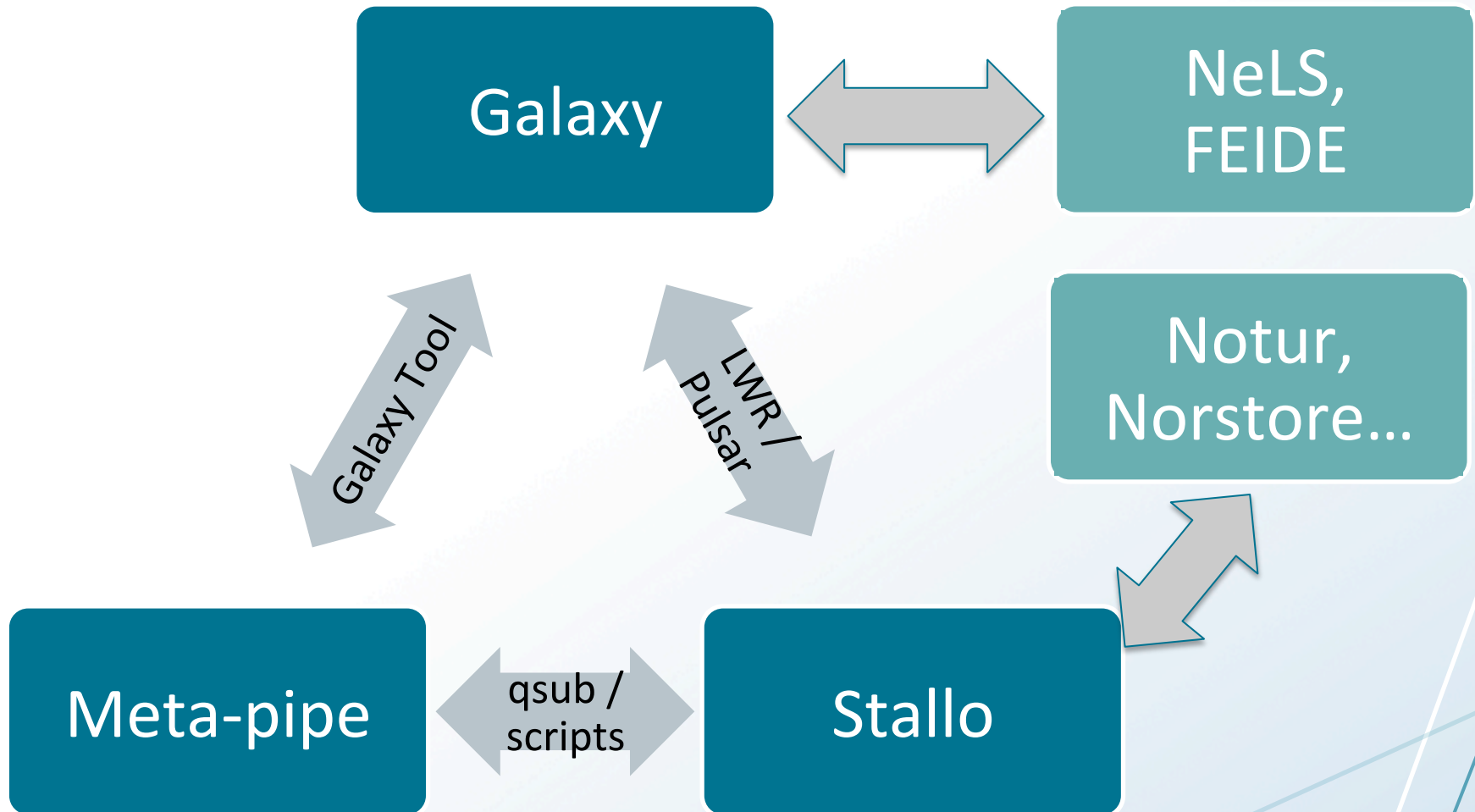
- Galaxy as common GUI
- Federated login
 - Meta-pipe is accessible to all *FEIDE* users
 - Non-*FEIDE* users are special case
- Data sharing
 - META-pipe input and output data can be imported/exported to NeLS storage using Galaxy or scp
 - 40TB (soon 300TB) temporal storage for NeLS projects
 - Permanent storage in NorStore coordinated by StoreBioInfo

Stallo Supercomputer (UiT)

- 750 nodes, 14.000 cores
- 12.8 TB DRAM
- 2.1 PB of total disk capacity (1PB shared temporary storage)
- Centralized storage via Infiniband
- PBS/Torque



Integration



Summary – Who are the stakeholders?

- Users:
 - Local, national, European, and international (accounting)
 - Academic and commercial (licenses)
- Service providers:
 - Pipelines hosted at EBI and UiT
 - Search engine hosted at EBI?
 - Marine resources databases hosted at?
- Resource providers:
 - EBI and UiT has resources for executing pipelines and searches
 - ...storage resources in next slide
- Other stakeholders:
 - Other big EU projects: EMBRC and MIRRI

Summary – Where is the data?

- (Raw input data at individual labs)
- Example 1, EBI MGP:
 - Submit data to ENA
 - Analysis run at EBI
 - Results stored at EBI
- Example 2, META-pipe (Norwegian users):
 - Live data stored at Supercomputer
 - Temporary storage on NeLS storage systems
 - Permanent storage on national storage system
- Future:
 - Archival services and databases (ENA, UniProt)
 - Mirroring of MGP and META-pipe data for federated search?
 - Marine resources database at?
 - Search engine data?
 - Embassy cloud?

Summary – How big is the data?

What are our computational needs?

- Estimate to use at least 0.5PB of disk
- We are optimizing our pipelines
- META-pipe:
 - UiT Supercomputer group not worried about resource usage
 - Pipeline scales with regards #users

Summary – How do we plan to manage data?

- Examples:
 - EBI MGP
 - META-pipe
 - Pilot project (harmonize results)

Summary – How do our users use the data?

- EBI MGP is a web application
- META-pipe is integrated with Galaxy
 - Will implement a web application interface for META-pipe
- Search engine will likely be a web application
- Also_ REST APIs and direct access to files

Summary

- Marine metagenomics
 - Pilot action
 - Use case
- Meta-pipe
 - Deployed as national service
 - Galaxy GUI
 - Integrated with local supercomputer
 - National storage resources (NeLS)
 - Federated login (NeLS)

Ongoing Work and Challenges

- Increase user base
 - Storage and compute requirements?
 - Where to store data for European users?
 - EBI MGP usage logs can be used to predict requirements
 - Pipelines and infrastructure scalable wrt number of users
 - HPC group at UiT not worried wrt available resources
- Reduce response time
 - What are the requirements and promises to users?
 - Improved algorithms, optimized data structures, dedicated cluster nodes?, Hadoop type of cluster?
- Single sign-in for all users
 - Accounting? Who is paying?
 - Academic vs. commercial users (licences, etc)?
 - Work in progress in NeLS
 - Need to decide and implement policies

Ongoing Work and Challenges

- Improved fault tolerance for 24/7 service
- Stallo cloud
 - OpenStack
 - Well suited for interactive jobs
- Large meta-database management
- Avoid Stallo bottlenecks
 - Scheduler
 - NAS/ Centralized storage
 - CPU (for some tools)
- Non-Galaxy web-interface
 - Remove many potential failures
 - Easier to predict resource usage
- Harmonization with EBI MGP

Use Case Deliverables

- Specific marine databases made publicly available
- Report on comprehensive metagenomic data standards environment
- Report describing a set of tools, pipelines and search engine for interrogation of marine metagenomic

Galaxy / uit x
https://galaxy-uit.bioinfo.no

Analyze Data Workflow Shared Data Visualization Admin Help User Using 3.9 GB

Tools

search tools

[Get Data](#)
[Send Data](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Metagenomics](#)
[Statistics](#)
[Meta-pipe](#)
[FASTA manipulation](#)
[NGS: QC and manipulation](#)
[Transcriptomics](#)
[NGS: Picard](#)

[Workflows](#)
▪ [All workflows](#)

Meta-pipe (version 1.0)

Genome fasta file:
1: MVIS.fas
Fasta sequence has to be in nucleotide format.

Number of CPUs to use::
128
Define the number of CPUs that are used in parallel for running the pipeline.

Gene prediction tool:
MetaGeneAnnotator
This defines the tool used for the initial prediction of CDSs

Select database search tools::

 Blastp
 Blastn
 Fasta32 (protein)
 Fasta32 (nucleotide)
 Pfam (Hmmer)
 Priam (Hmmer)
The selected tools will be run on each predicted CDS. Keep in mind the run-time of each tool (see help below for further information)

Select database taxa (optional):
Bacteria
Fungi
Archaea
Invertebrates
Tip: Reduce the run-time by selecting only a sub-set of taxonomical divisions of the Swiss-Prot and Uniprot databases.

Execute SignalP:
No
Includes signal-peptide prediction in annotation

Exported file format::
Extended EMBL (recommended)

History

Unnamed history
4.9 MB

1: MVIS.fas

Oppdatert figur (EBI vs. M-P)