UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

METApipe – Metagenomics Analysis Pipeline

Lars Ailo Bongo
Dept. of Computer Science & Center for Bioinformatics,
University of Tromsø, Norway

http://sfb.cs.uit.no

SfB Center for Bioinformatics

elixir Data for Life

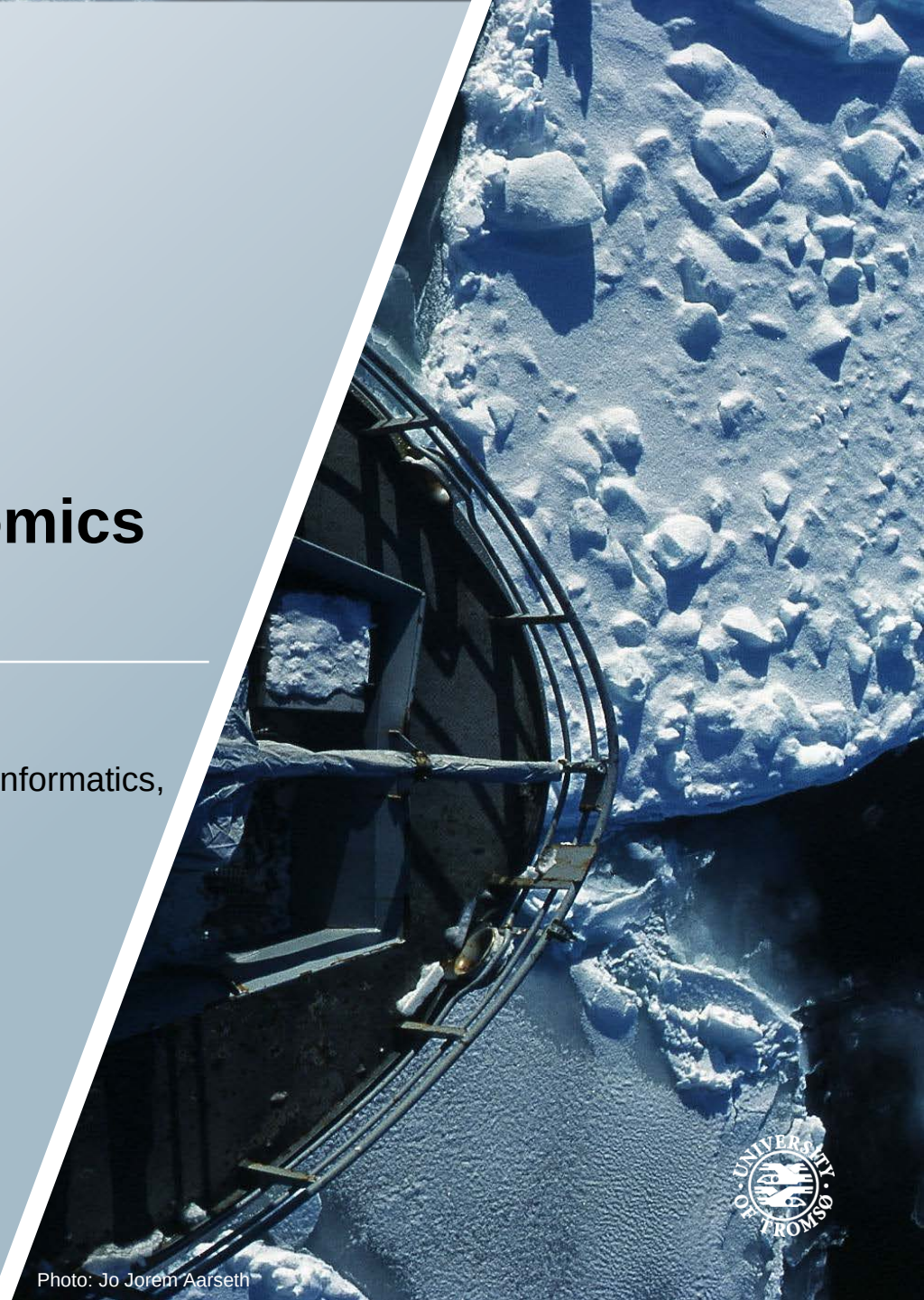


Photo: Jo Jorem Aarseth

On Top of the High North

Our position, on top of the High North, reflects both a geographical fact and an ambition.

We are the northernmost university in the world, at 69° North.

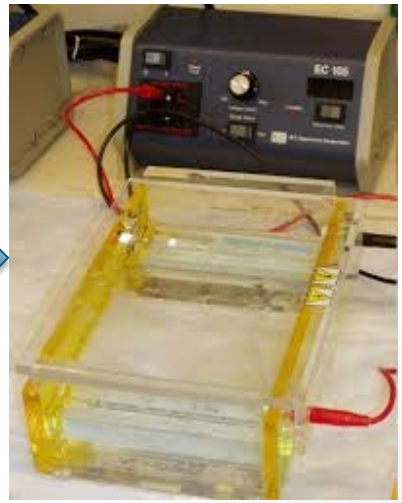
Our location on the edge of the Arctic also implies a mission, as the Arctic is of increasing global importance.

Our ambition is to be on top of all things north. Because if it affects the north, it affects the world.

Outline

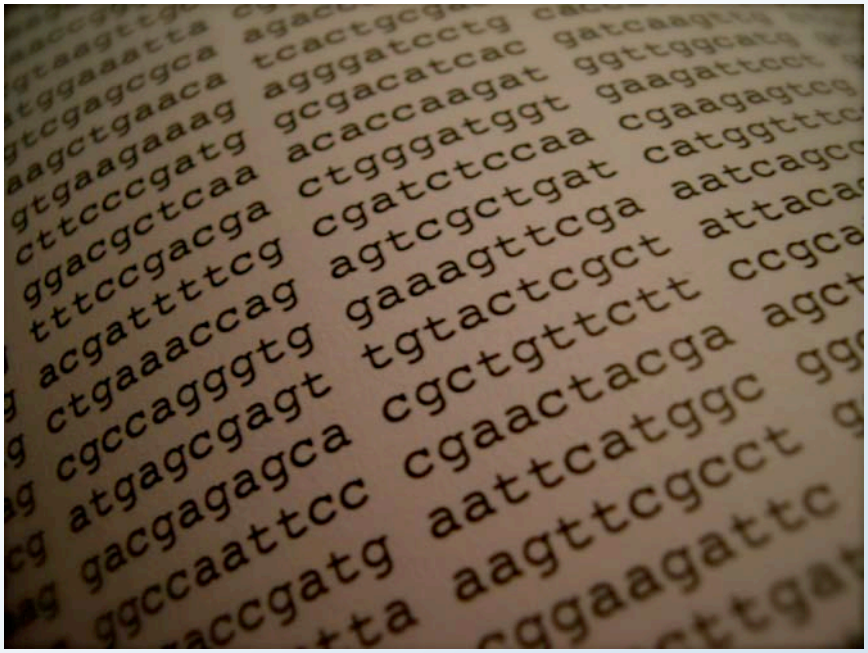
- Context:
  - Biological data processing
  - Challenges
  - Infrastructure
  - Metagenomics
- METApipe:
  - Overview
  - Deployment
- Our research
- Relevance to NESUS/ WG6

Biology - a computational science

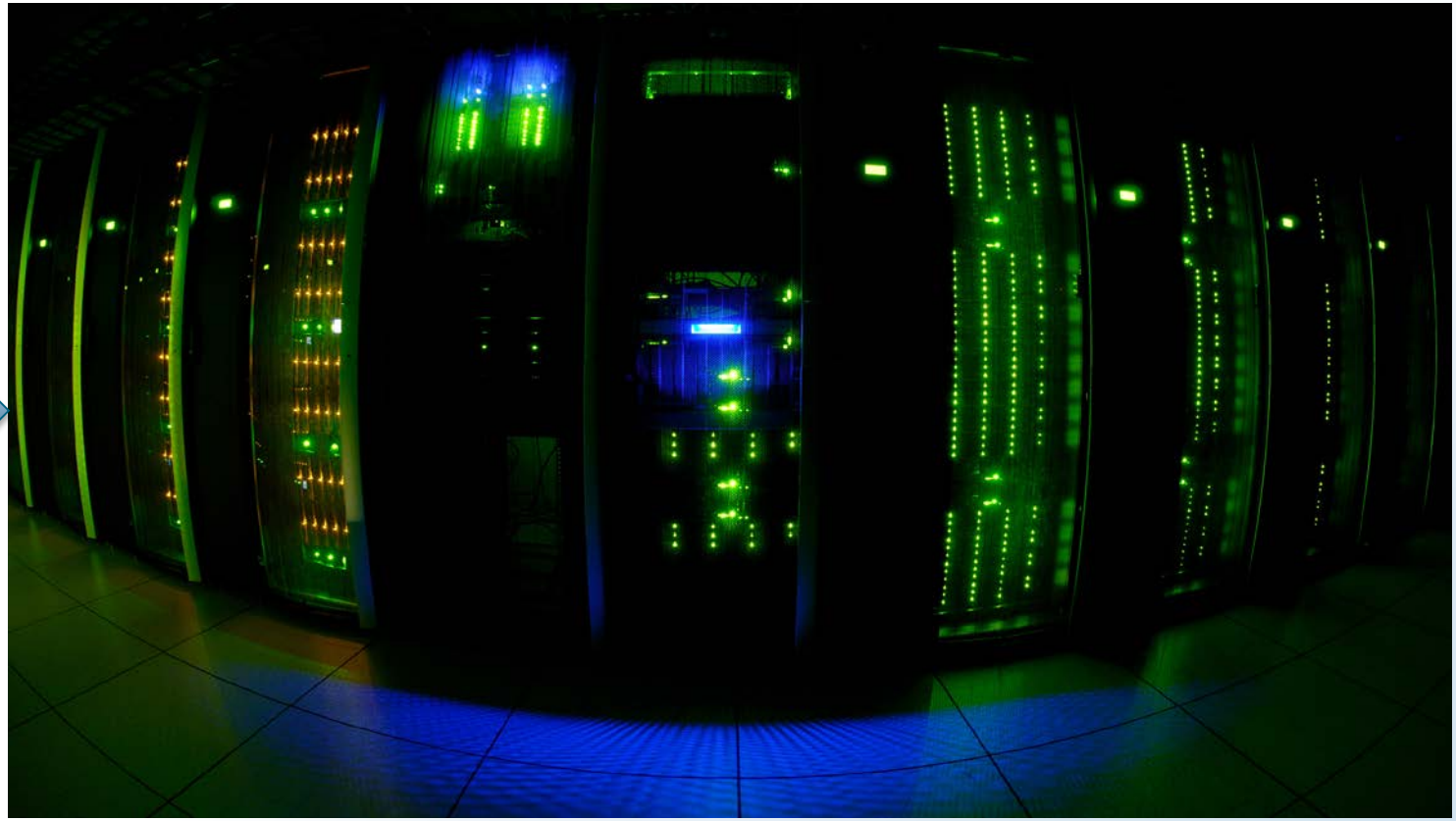


Bioinformatics

- The science of integrating large amounts of biological data
- Interdisciplinary: biology, computer science, statistics, etc.
- At the heart of modern biology

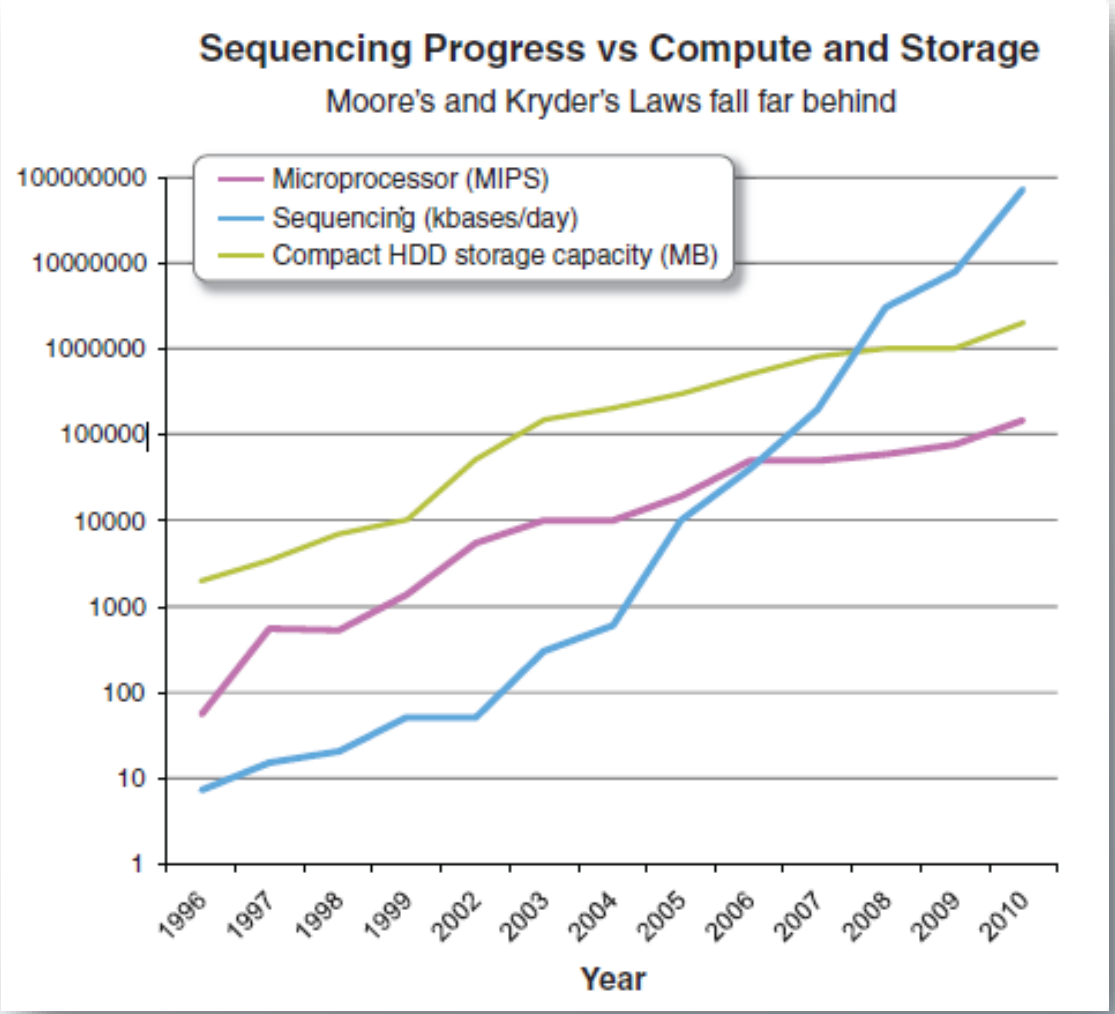


Bioinformatics– a supercomputing science



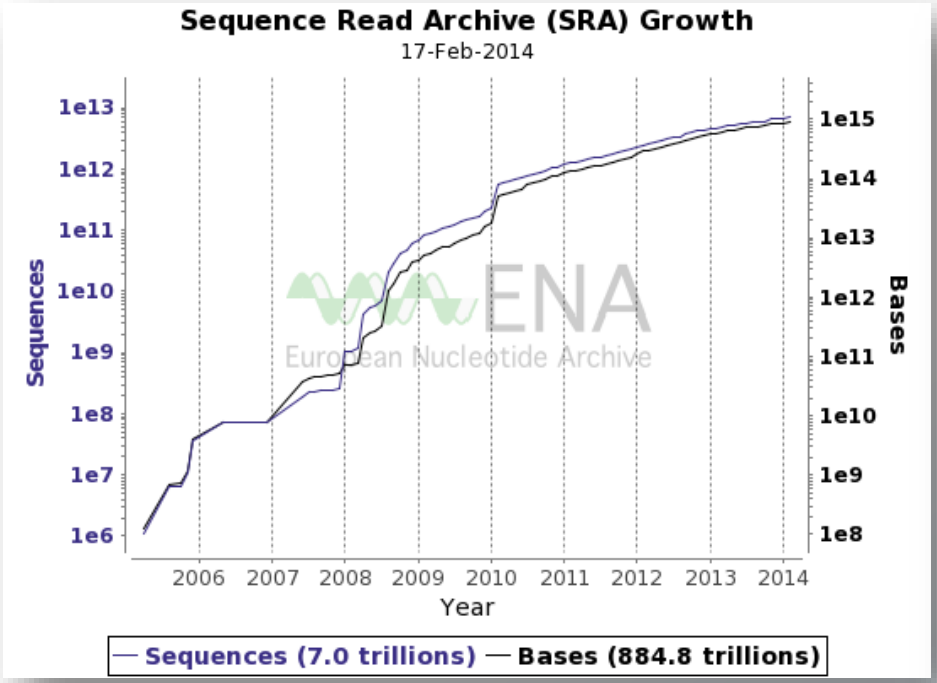
Stallo Supercomputer, University of Tromsø

Data Deluge

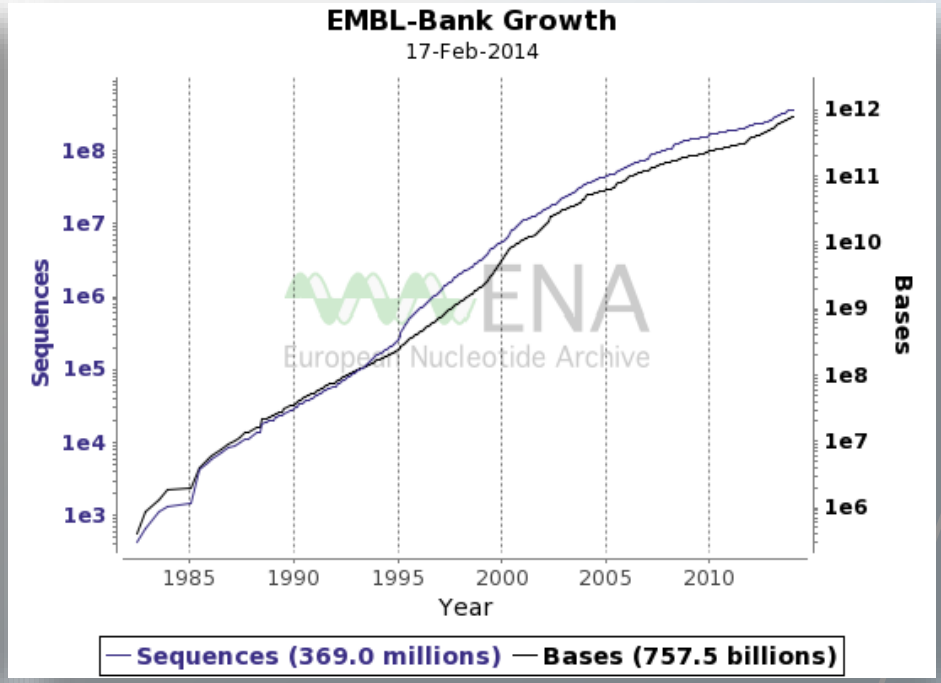

Sequencing Progress vs Compute and Storage
Moore's and Kryder's Laws fall far behind
Microprocessor (MIPS)
Sequencing (kbases/day)
Compact HDD storage capacity (MB)
100000000
10000000
1000000
100000
10000
1000
100
10
1
1996
1997
1998
1999
2002
2003
2004
2005
2006
2007
2008
2009
2010
Year

Data Deluge (2)

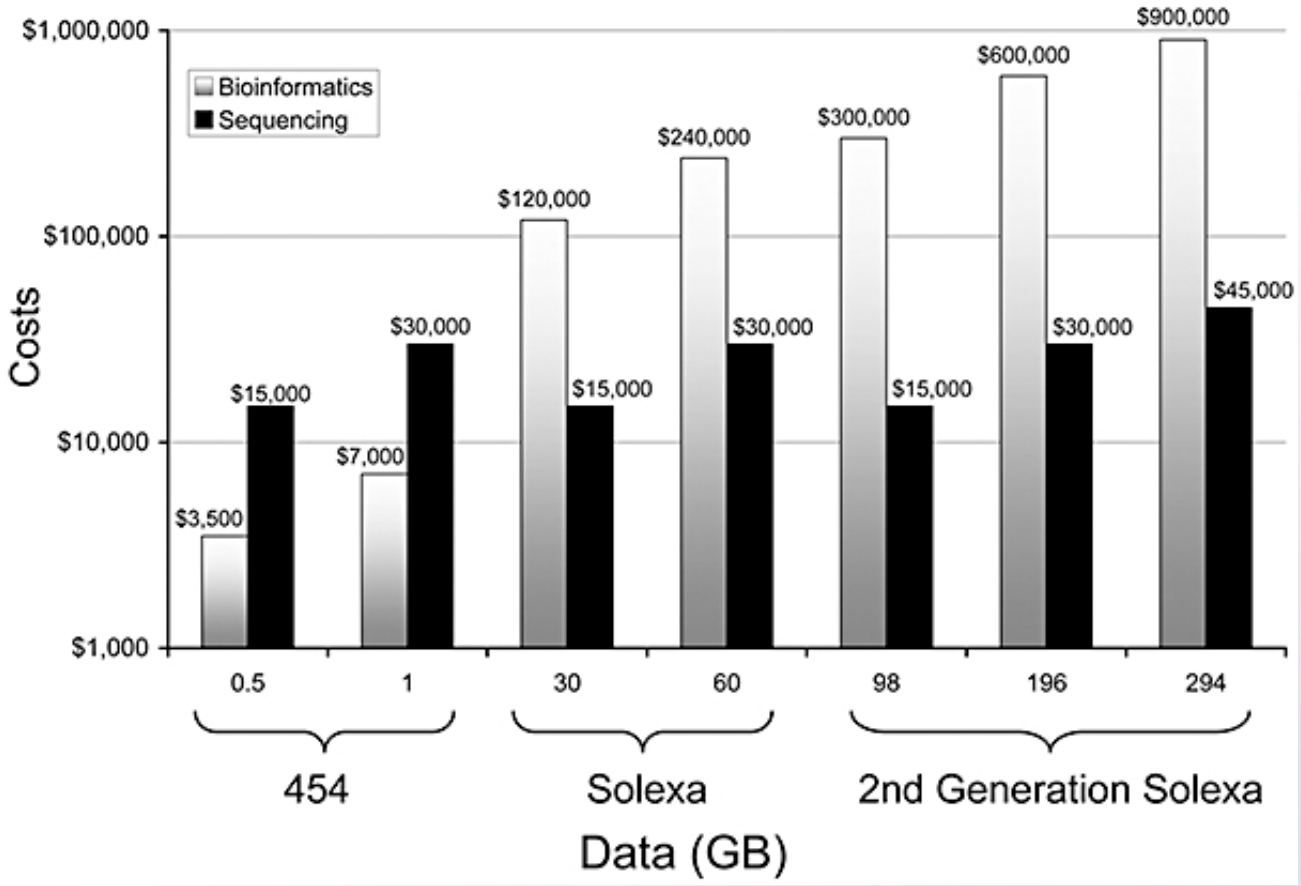
Reads growth



Assembled/annotated sequence growth

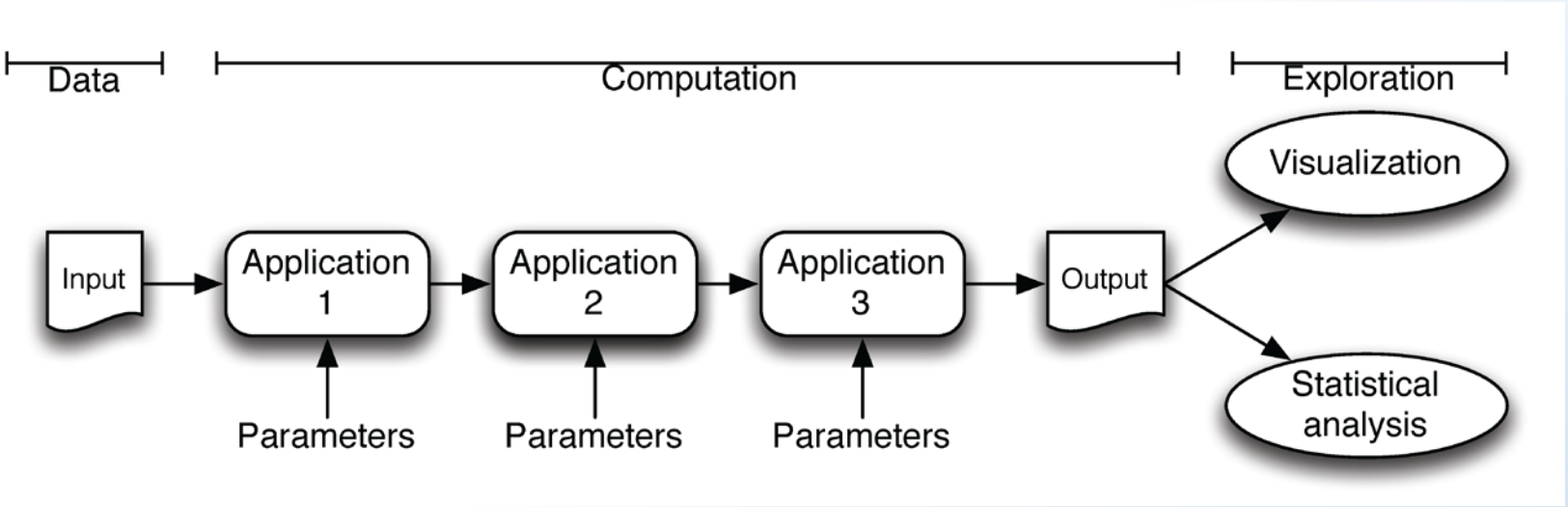


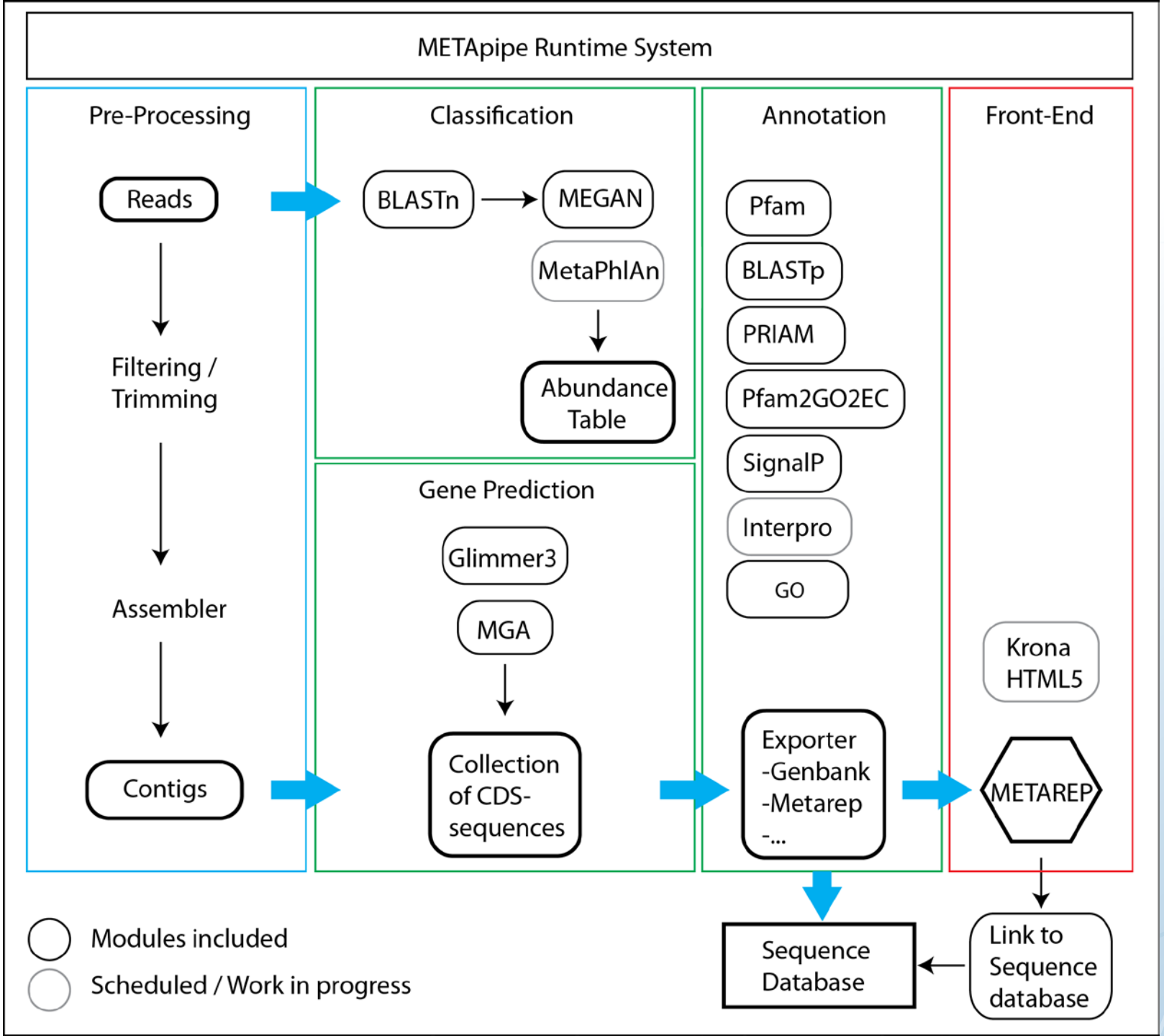
Cost of Sequence Based Screening (on Amazon EC2 in 2010)



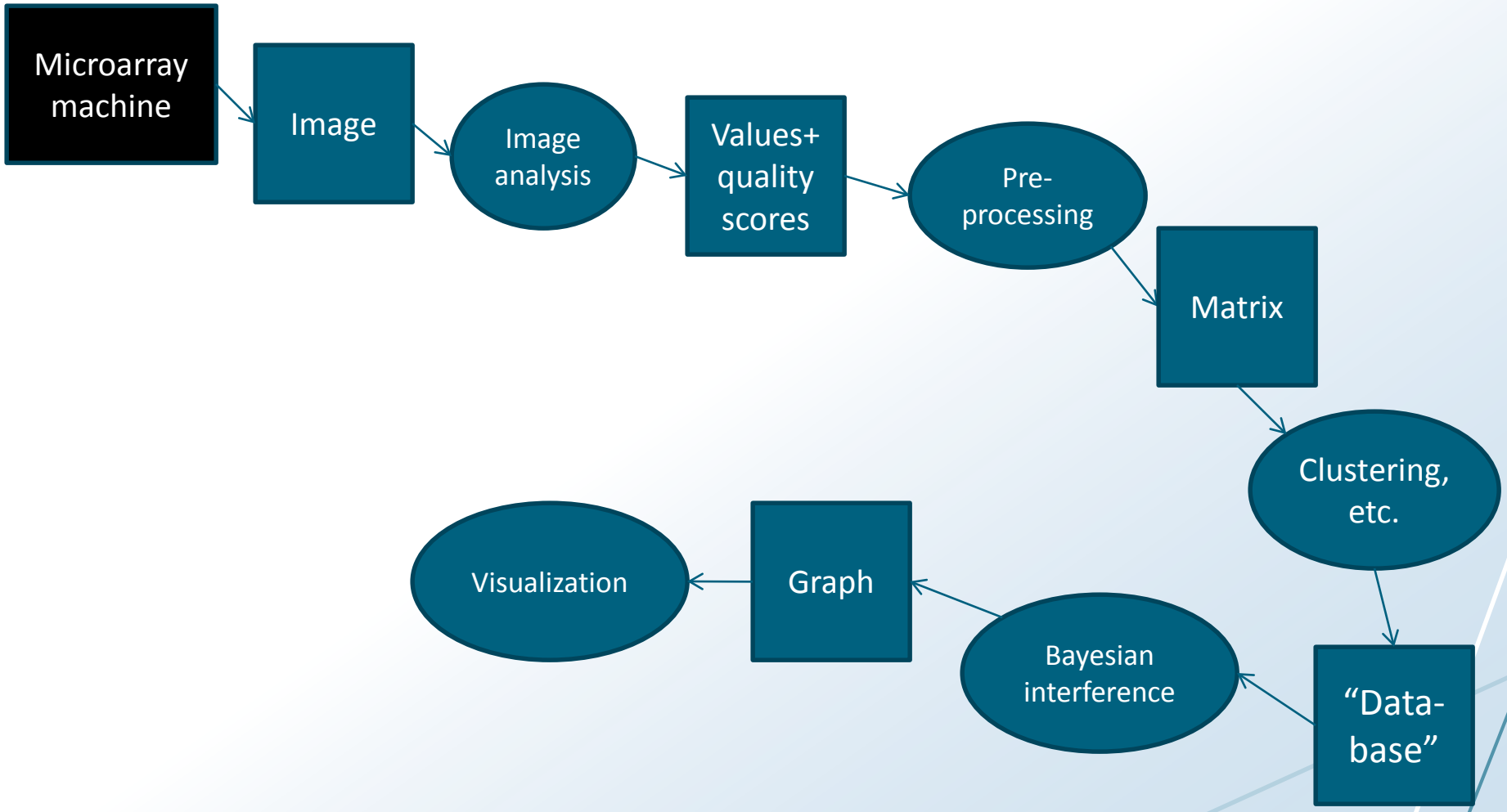
Heidelberg, KB, Gilbert, JA and Joint, I (2010) Marine genomics: at the interface of marine microbial ecology and biodiscovery. Microb Biotechnol. 3(5): 531–543

Biological Data Processing

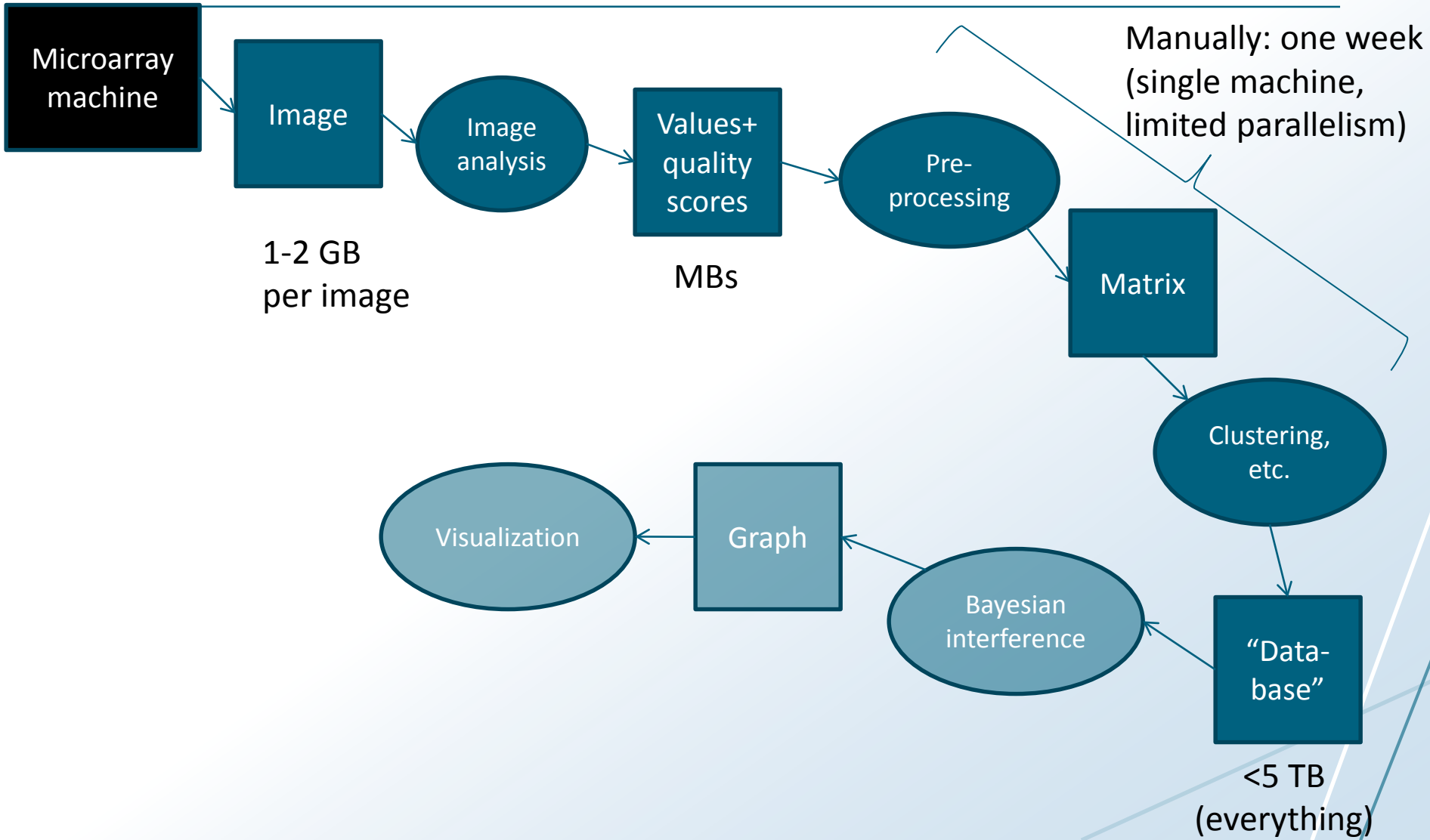



METApipe Runtime System
Pre-Processing
Reads
Filtering / Trimming
Assembler
Contigs
Classification
BLASTn
MEGAN
MetaPhlAn
Abundance Table
Gene Prediction
Glimmer3
MGA
Collection of CDS-sequences
Annotation
Pfam
BLASTp
PRIAM
Pfam2GO2EC
SignalP
Interpro
GO
Exporter -Genbank -Metarep -...
Front-End
Krona HTML5
METAREP
Sequence Database
Link to Sequence database
Modules included
Scheduled / Work in progress

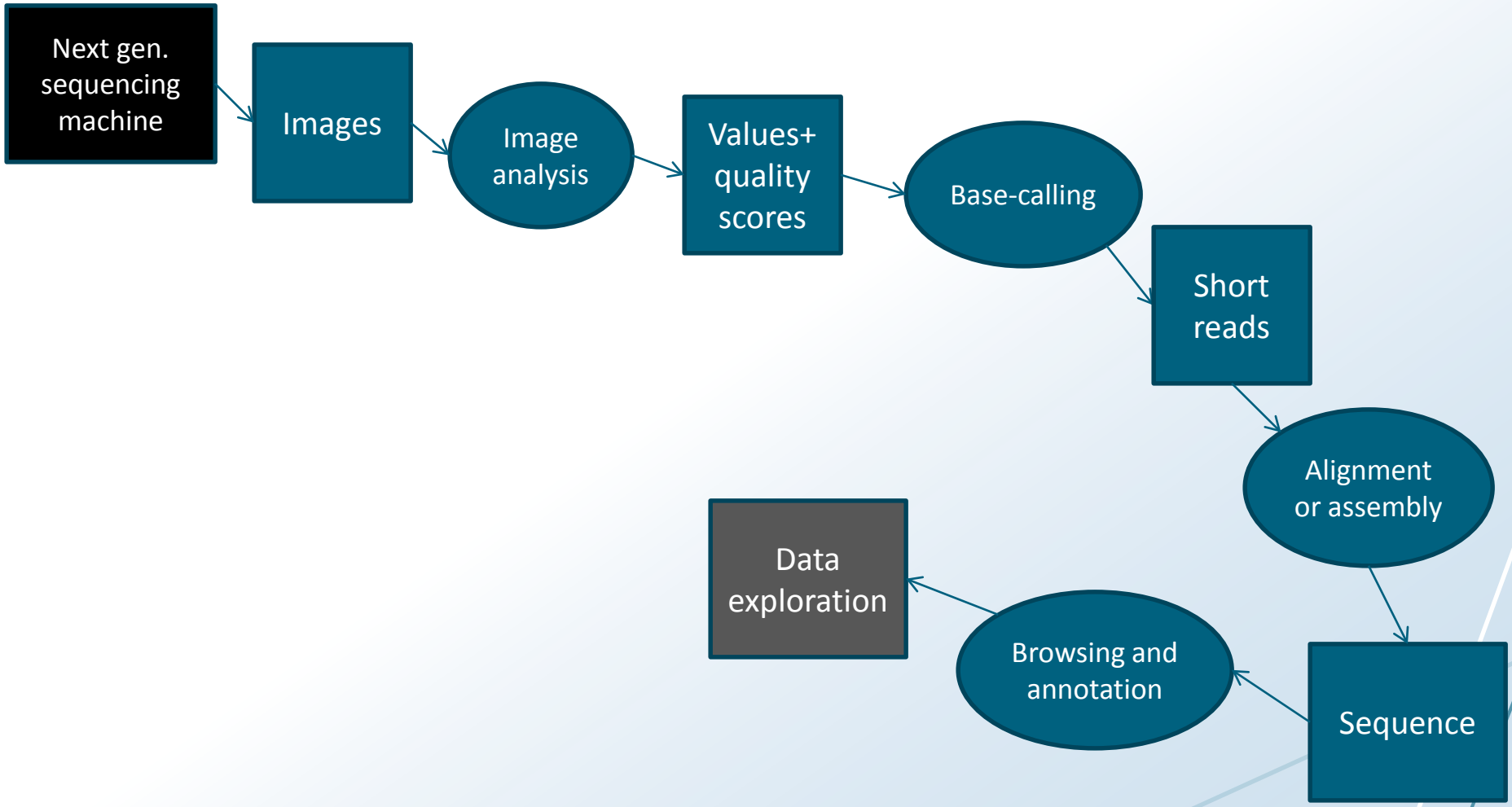
Microarray Pipeline



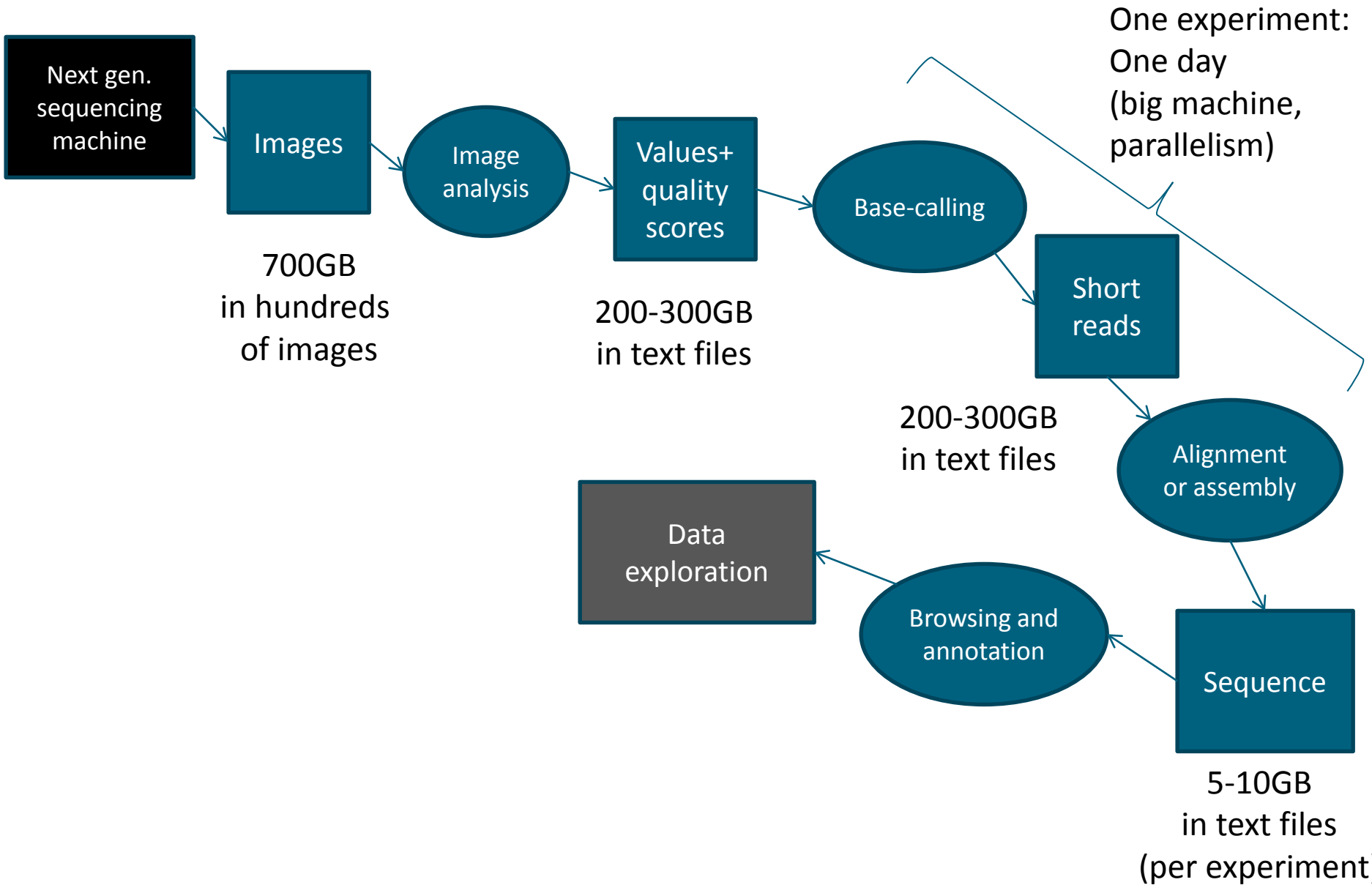
Microarray Pipeline



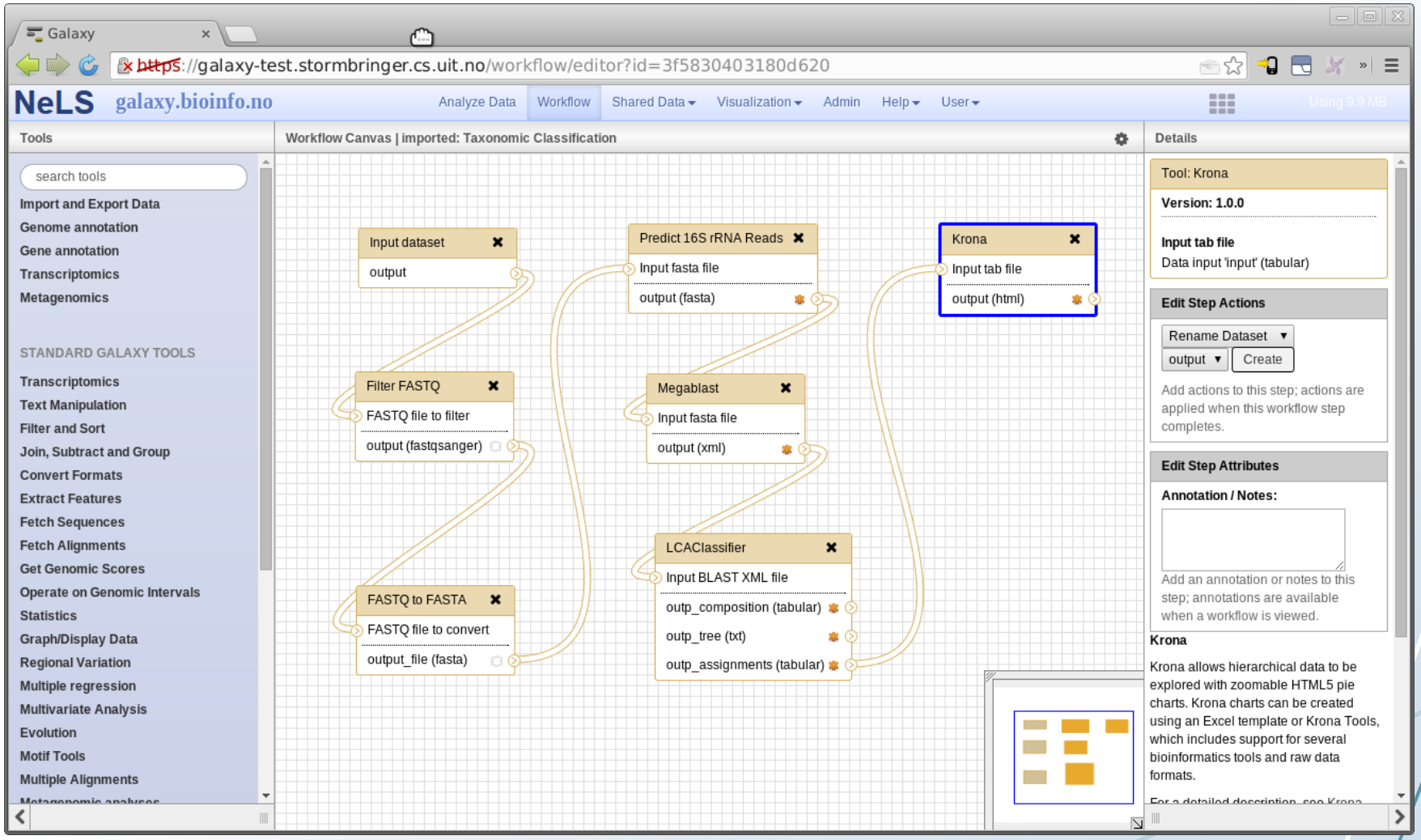
Next-generation sequencing pipeline



Next-generation sequencing pipeline (per experiment)



Galaxy


Galaxy
https://galaxy-test.stormbringer.cs.uit.no/workflow/editor?id=3f5830403180d620
NeLS
galaxy.bioinfo.no
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Tools
search tools
Import and Export Data
Genome annotation
Gene annotation
Transcriptomics
Metagenomics
STANDARD GALAXY TOOLS
Transcriptomics
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Workflow Canvas | imported: Taxonomic Classification
Input dataset
output
Predict 16S rRNA Reads
Input fasta file
output (fasta)
Krona
Input tab file
output (html)
Filter FASTQ
FASTQ file to filter
output (fastqsanger)
Megablast
Input fasta file
output (xml)
LCAClassifier
Input BLAST XML file
outp_composition (tabular)
outp_tree (txt)
outp_assignments (tabular)
FASTQ to FASTA
FASTQ file to convert
output_file (fasta)
Details
Tool: Krona
Version: 1.0.0
Input tab file
Data input 'input' (tabular)
Edit Step Actions
Rename Dataset
output
Create
Add actions to this step; actions are applied when this workflow step completes.
Edit Step Attributes
Annotation / Notes:
Add an annotation or notes to this step; annotations are available when a workflow is viewed.
Krona
Krona allows hierarchical data to be explored with zoomable HTML5 pie charts. Krona charts can be created using an Excel template or Krona Tools, which includes support for several bioinformatics tools and raw data formats.

ELIXIR

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

ELIXIR's mission

To build a sustainable European infrastructure for biological information, supporting life science research and its translation to:

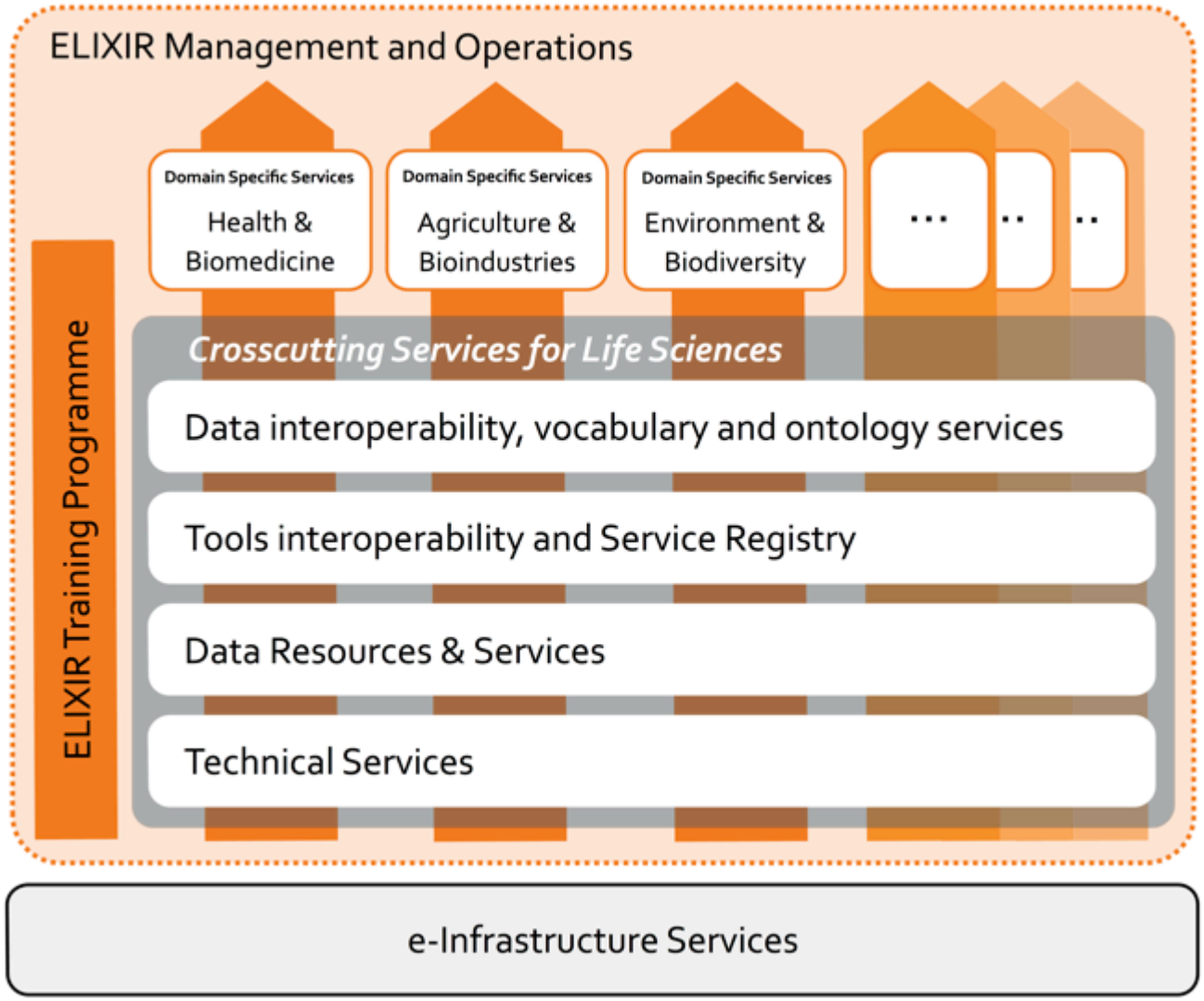
biomedicine

environment

bioindustries

society

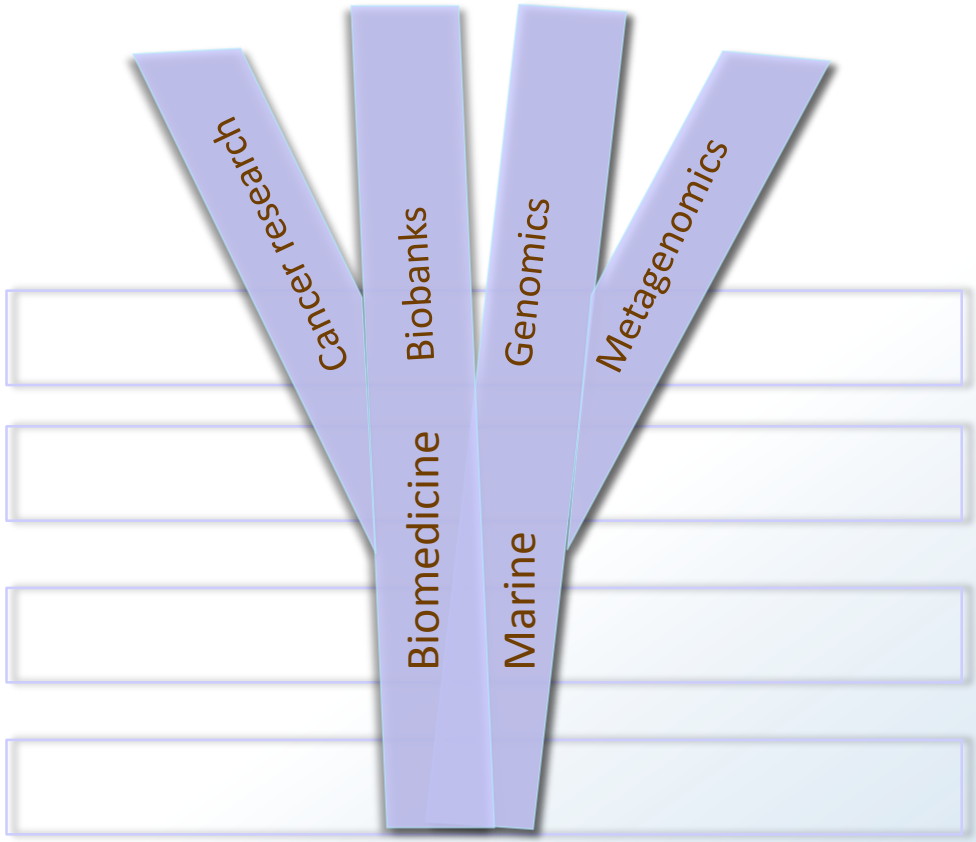
ELIXIR Programme



ELIXIR.NO


elixir
NORWAY
UNIVERSITY OF TROMSØ
Norway
UNIVERSITAS BERGENSIS
UNIVERSITAS OSLOENSIS
MDCCCXI
UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
MDCCCLIX

ELIXIR *Norway*



Portal, UI, visualisation

Processing, analysis, tools, pipelines

Databases, data-sets, meta-data, projects

HPC, Norstore

Tromsø node - Tasks

- Build and implement workflows for genomics and metagenomics
- Service towards users
  - **Special focus on marine**

METApipe - overview

- Metagenomics analysis pipeline
- Pipeline combines
  - Standard bioinformatics tools
  - Custom made tools
- Interactive data exploration
- Deployed at Center for Bioninformatics

Metagenomics

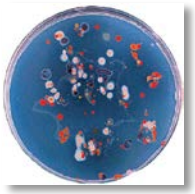
Direct isolation of DNA from the environments

Metagenomics can in principle access 100 % of the genetic resources of an environment

Isolation of RNA & construction of cDNA libraries

Metatranscriptome: response of organism to varying environmental perturbations

Cultivation

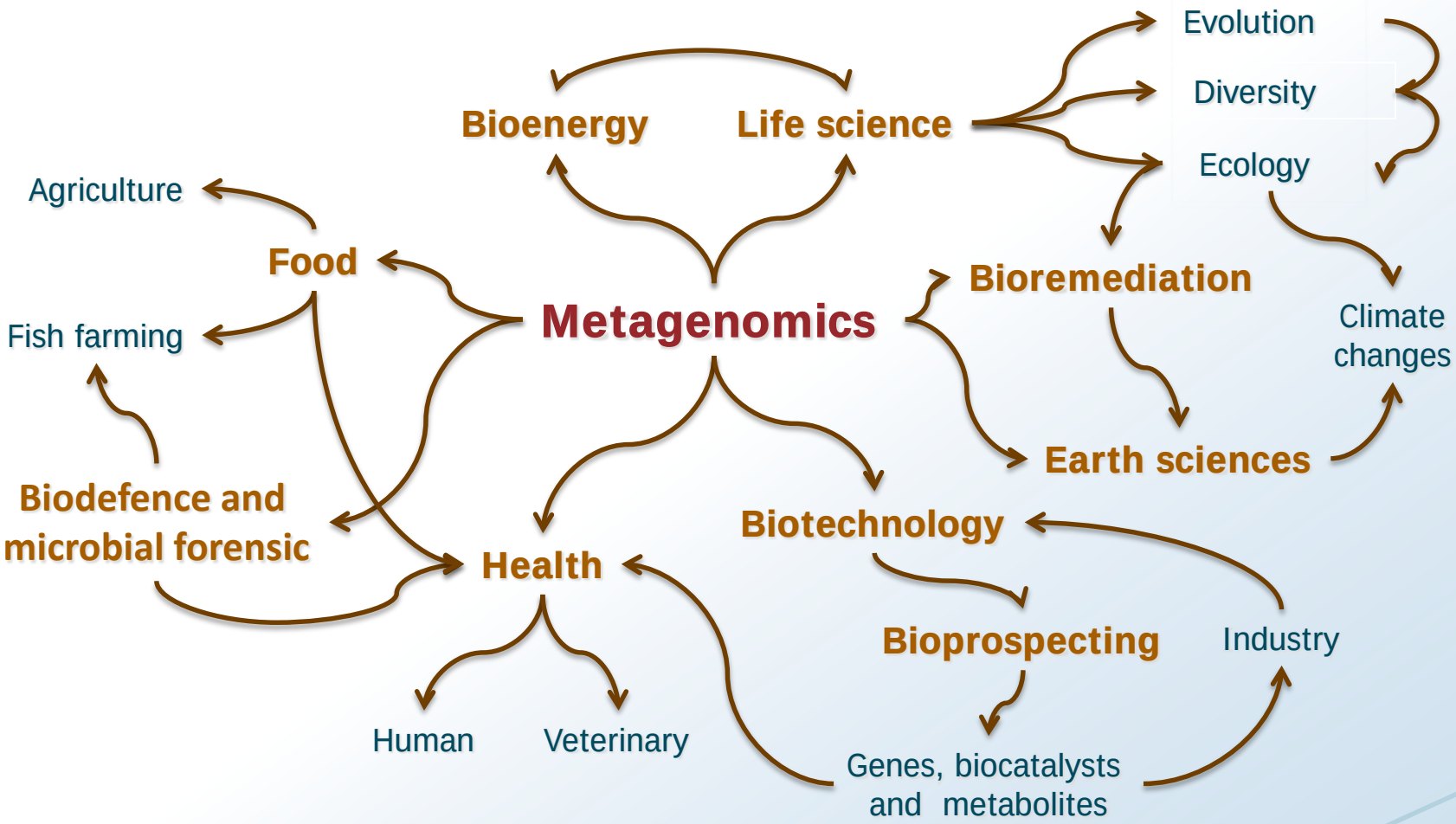


Cultivatable species

Isolation of DNA

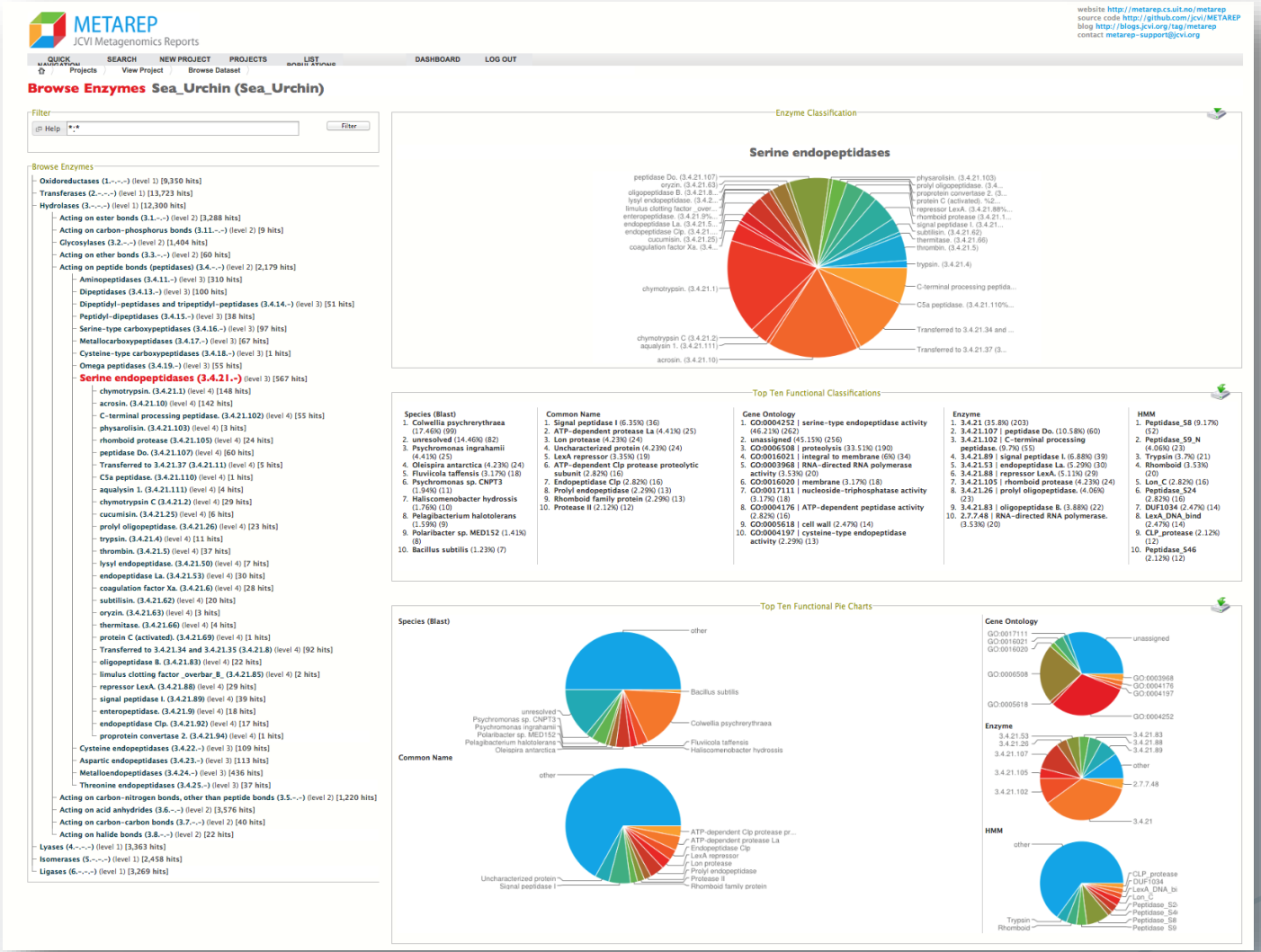
Traditional cultivation methods and traditional genomics can at best access 1 %

Metagenomics

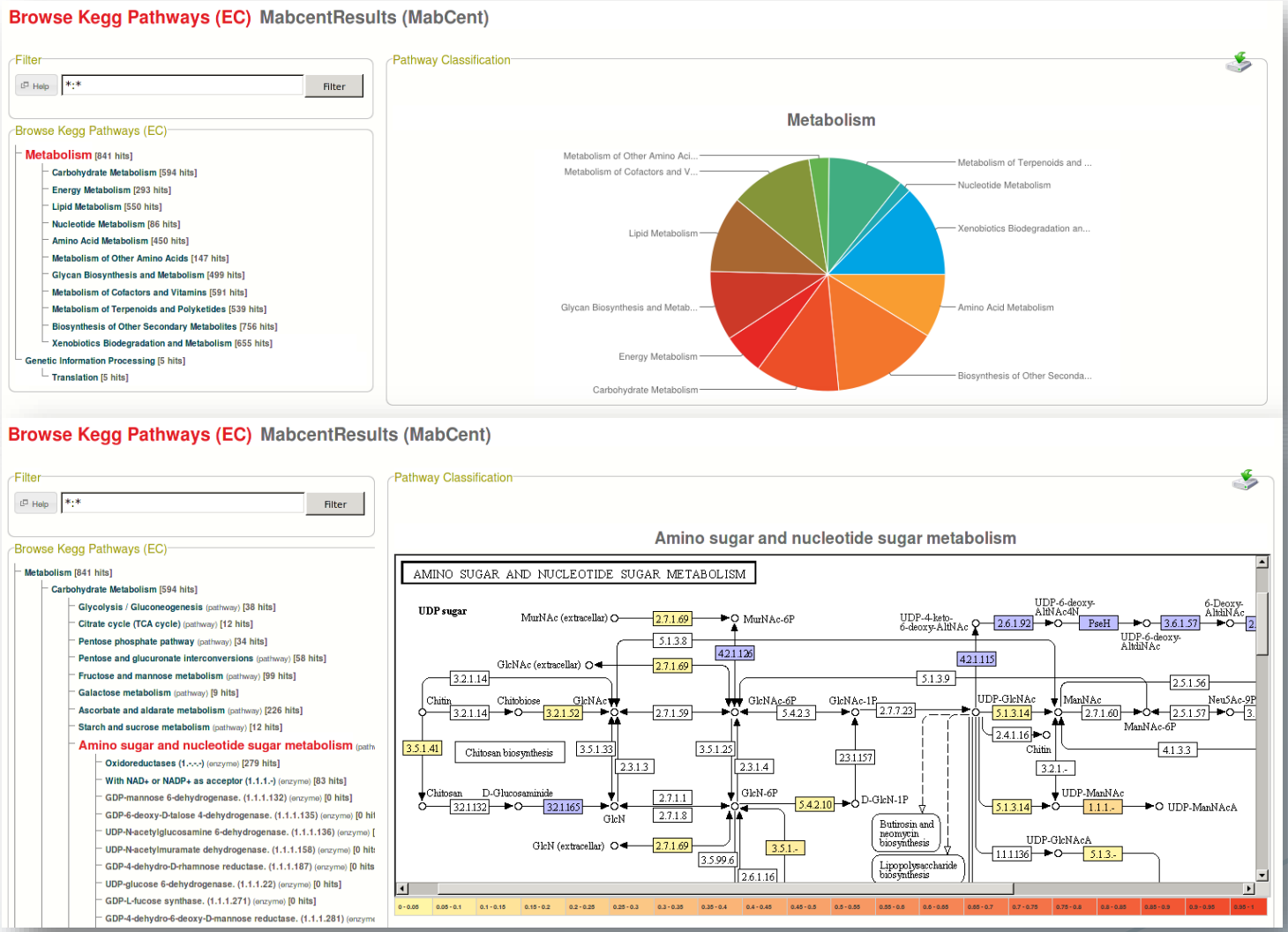


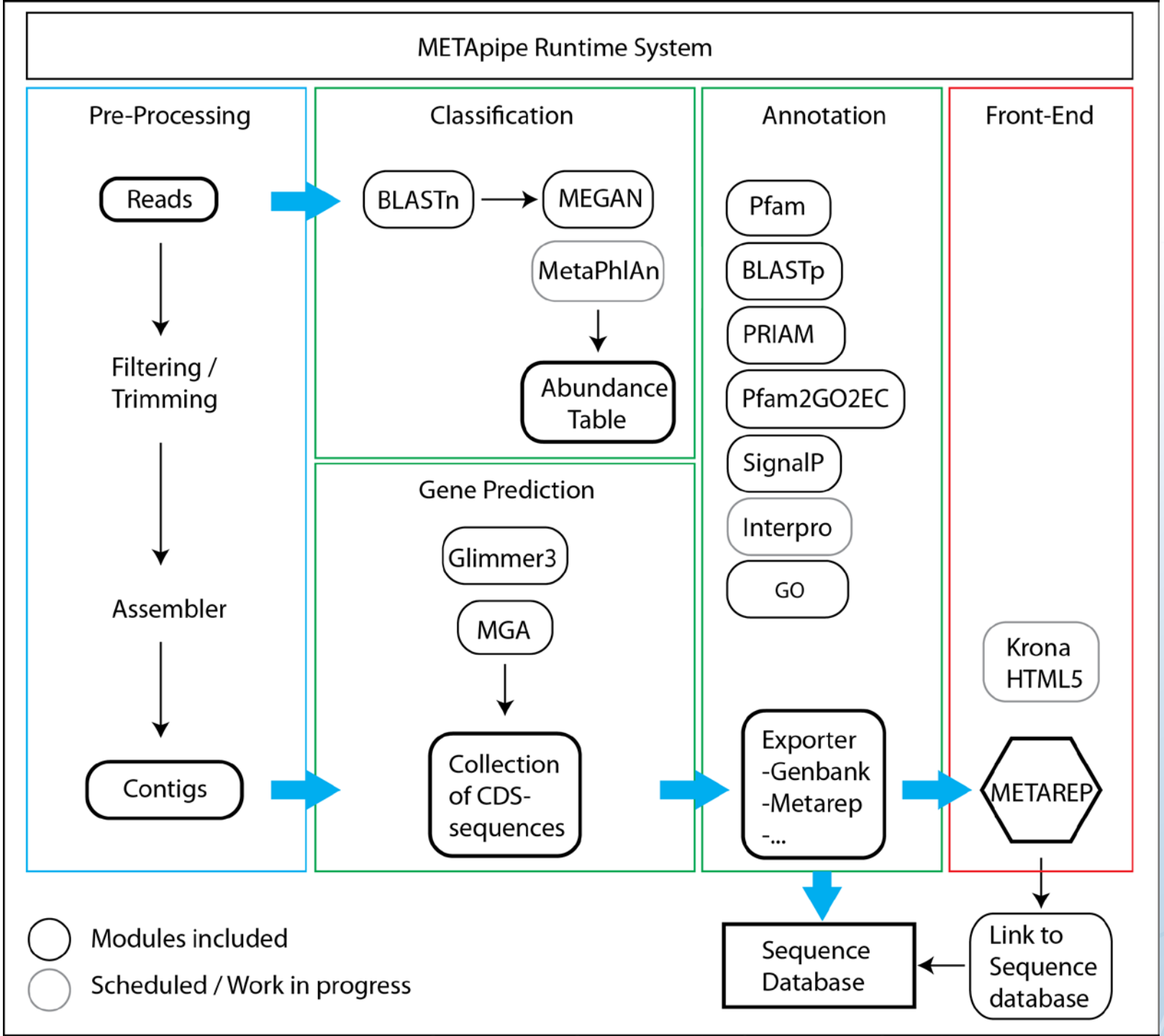
Metagenomics

Sea urchin – novel enzymes



MabCent – hunting for novel enzymes




METApipe Runtime System
Pre-Processing
Reads
Filtering / Trimming
Assembler
Contigs
Classification
BLASTn
MEGAN
MetaPhlAn
Abundance Table
Gene Prediction
Glimmer3
MGA
Collection of CDS-sequences
Annotation
Pfam
BLASTp
PRIAM
Pfam2GO2EC
SignalP
Interpro
GO
Exporter -Genbank -Metarep -...
Front-End
Krona HTML5
METAREP
Sequence Database
Link to Sequence database
Modules included
Scheduled / Work in progress

Infrastructure @ UIT

Secure domain

MiSeq

2x300 bp
> 15 Gb/run

ICE

5x2 GHz Intel Xenon E5620
5x24 GB DRAM
10x2 GHz Intel Xenon ES-1620
10x32 DRAM
2.5 GHz Intel Xeon E5-2640
16x16 GB DRAM

112 Physical grid cores
70+ TB RAID Storage

Stallo/Notur

608 x 2.60 GHz Intel Xeon E5 2670
2000 TB Storage – 12 TB DRAM

MiSeq
reporter

Public domain

Bacespace

Stormbringer

2 GHz Intel Xenon E5645
48 GB DRAM
2 x 1 TB RAM

galaxy-uit.bioinfo.no

2 GHz Intel Xenon
(10TB/128 GB RAM)

My Research Activities

- Biological Data Processing Systems Lab
- Build and experimentally evaluate infrastructure systems for next-generation bioinformatics applications
- Approach:
  - Utilize data-intensive computing systems
  - Provide new services
  - Unmodified analysis tools
  - Integrate with data analysis frameworks
- http://bdps.cs.uit.no

Results

- Troilkatt: scalable processing
  - Integrate all gene expression datasets in NCBI GEO
  - Built on Hadoop
- GeStore: incremental updates
  - For unmodified pipeline tools
  - Terabyte meta-database management
- Mario: interactive iterative processing
  - Tune pipeline parameters
- Kvik: data exploration for NOWAC postgenome biobank
  - Interactive visualizations
- Spark-SPELL:
  - Interactive scalable search
  - Built on Spark
- …

Concluding remarks

- Ultrascale computing system?
  - Lots of small jobs
  - Big data
  - Big projects (NOWAC, 1000 Genomes, …)
- Sustainability important?
  - Need programmability, data management, resilience, scalability
- Holistic view important?
  - ELIXIR will build big ecosystem
- System software?
  - Big data management and scalability
  - Novel services
- Redesign and/or reprogram applications?
  - Yes, for certain big problems (e.g. next-generation sequencing)
  - No, too many specialized tools and pipelines

Concluding remarks (2)

- Algorithms, applications, and services amenable to ultrascale systems?
  - Selected algorithms/ applications (BLAST, next-gen sequencing)
  - Services for the rest
- Impact of application requirements?
  - Interactivity
  - Flexibility (select tools and parameters)
- Key application?
  - Life sciences emerging domain in supercomputing
- Computational patterns?
  - Wide variety of tools and patterns
  - Mix of data, memory, computation intensive
  - Many are designed for a single computer

Acknowledgments

METApipe: University of Tromsø
- Nils Peder Willassen
- Peik Haugen
- Edvard Pedersen
- Espen Mikal Robertsen
- Tim Kahlke
- Erik Hjerde
- Erik Kjærner-Semb
- Inge Alexander Raknes
- Jon Ivar Kristiansen

troilkatt: Princeton University
- Olga Troyanskaya
- Kai Li
- Aaron Wang
- Chris Park
- Casey Greene
- Yuanfang Guan
- Alicja Tadych (Princeton)

NOWAC: University of Tromsø:
- Eiliv Lund
- Karina Standahl Olsen
- Mie Jareid
- Nicolle Mode
- Bjørn Fjukstad
- Einar Holsbø
- Giacomo Tartari