

UiT

THE ARCTIC  
UNIVERSITY  
OF NORWAY

# METApipeline – Metagenomics Analysis Pipeline

Lars Ailo Bongo

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

<http://sfb.cs.uit.no>



Photo: Jo Jørem 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.



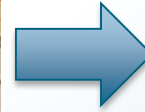
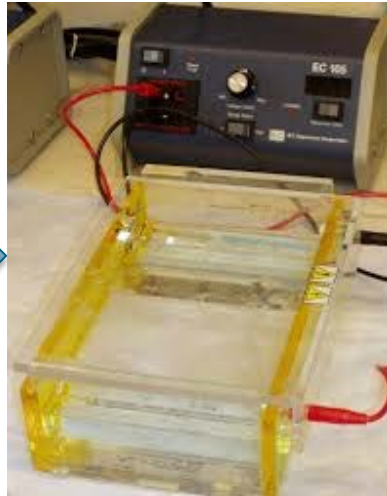
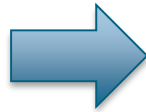
# Outline

---

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

# Biology - a computational science

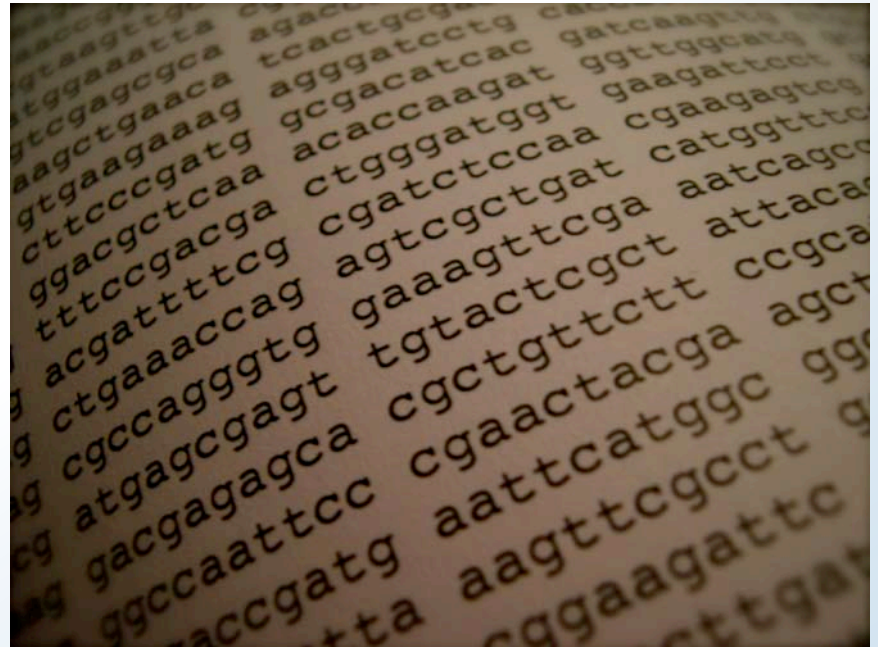
---



# Bioinformatics

---

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



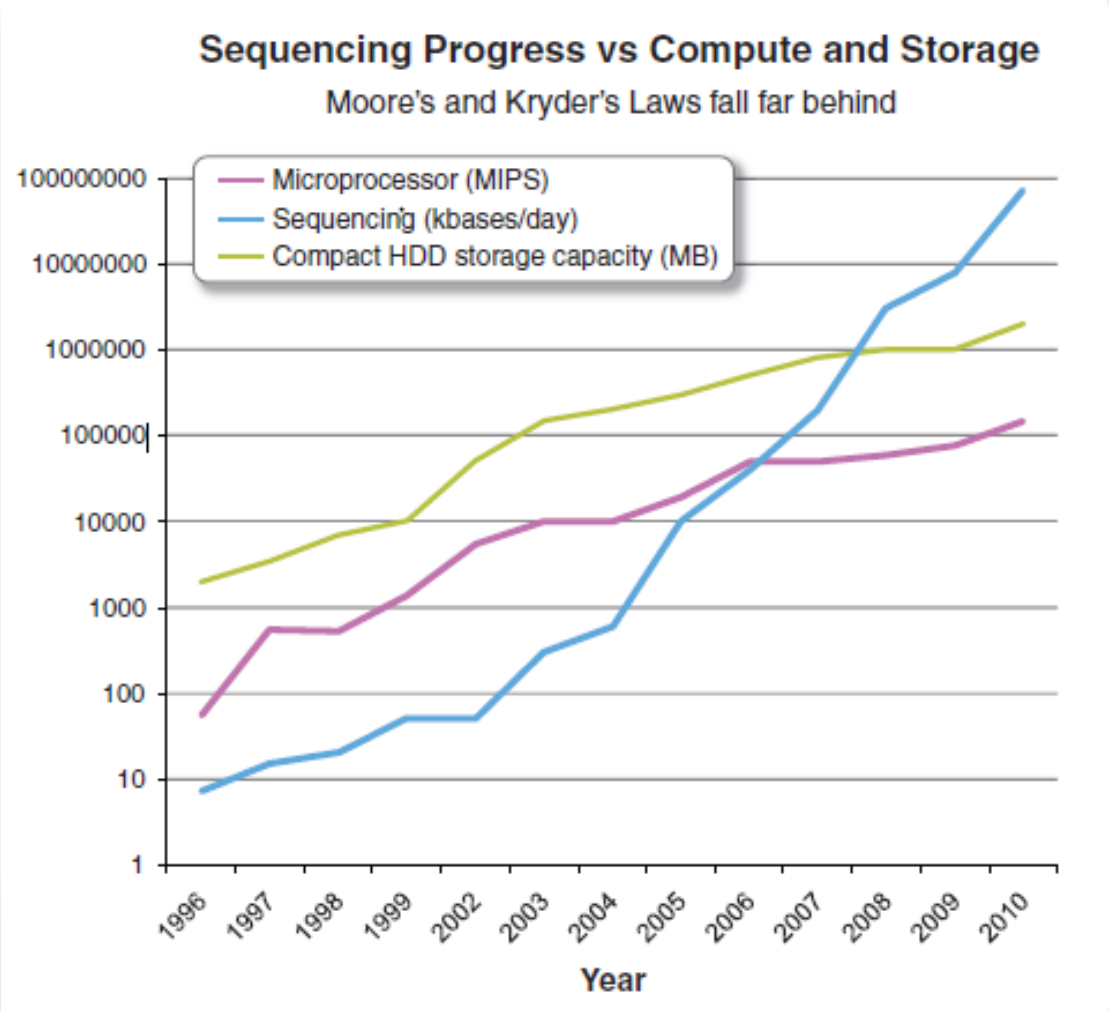
# Bioinformatics– a supercomputing science

---



Stallo Supercomputer, University of Tromsø

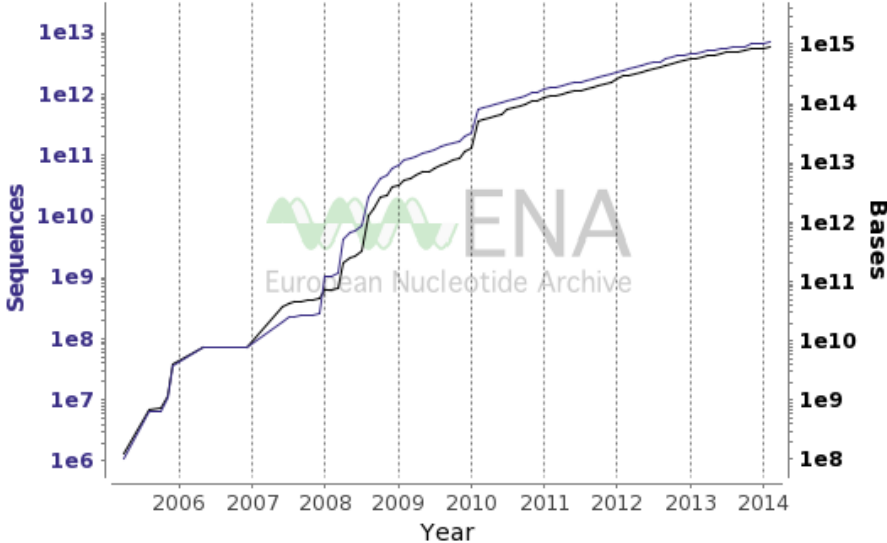
# Data Deluge



# Data Deluge (2)

## Reads growth

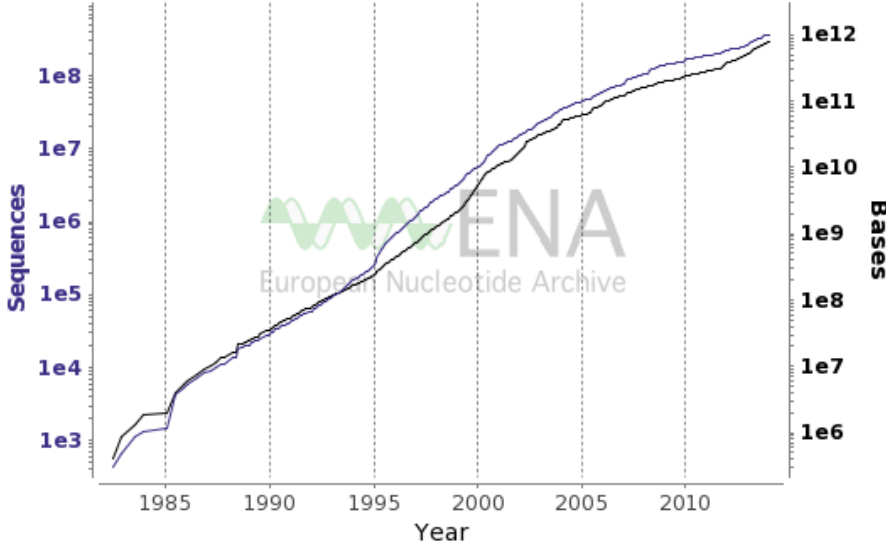
Sequence Read Archive (SRA) Growth  
17-Feb-2014



— Sequences (7.0 trillions) — Bases (884.8 trillions)

## Assembled/annotated sequence growth

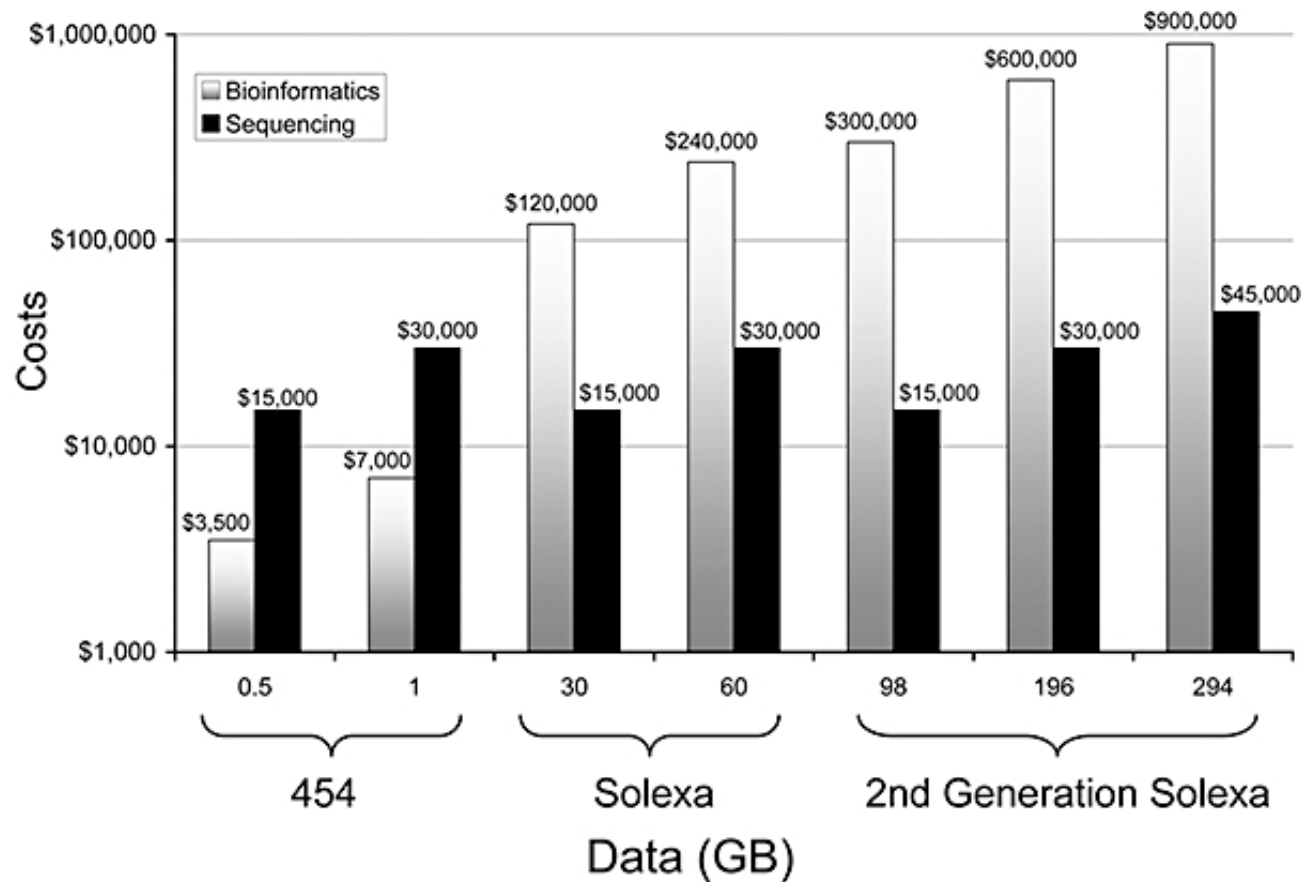
EMBL-Bank Growth  
17-Feb-2014



— Sequences (369.0 millions) — Bases (757.5 billions)



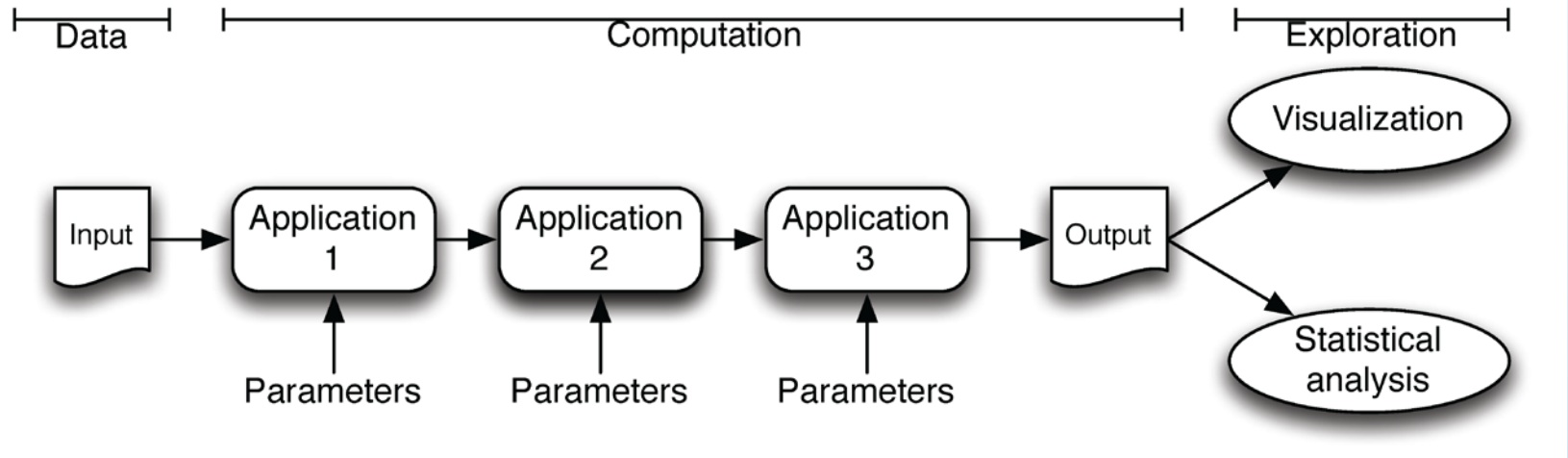
# Cost of Sequence Based Screening (on Amazon EC2 in 2010)



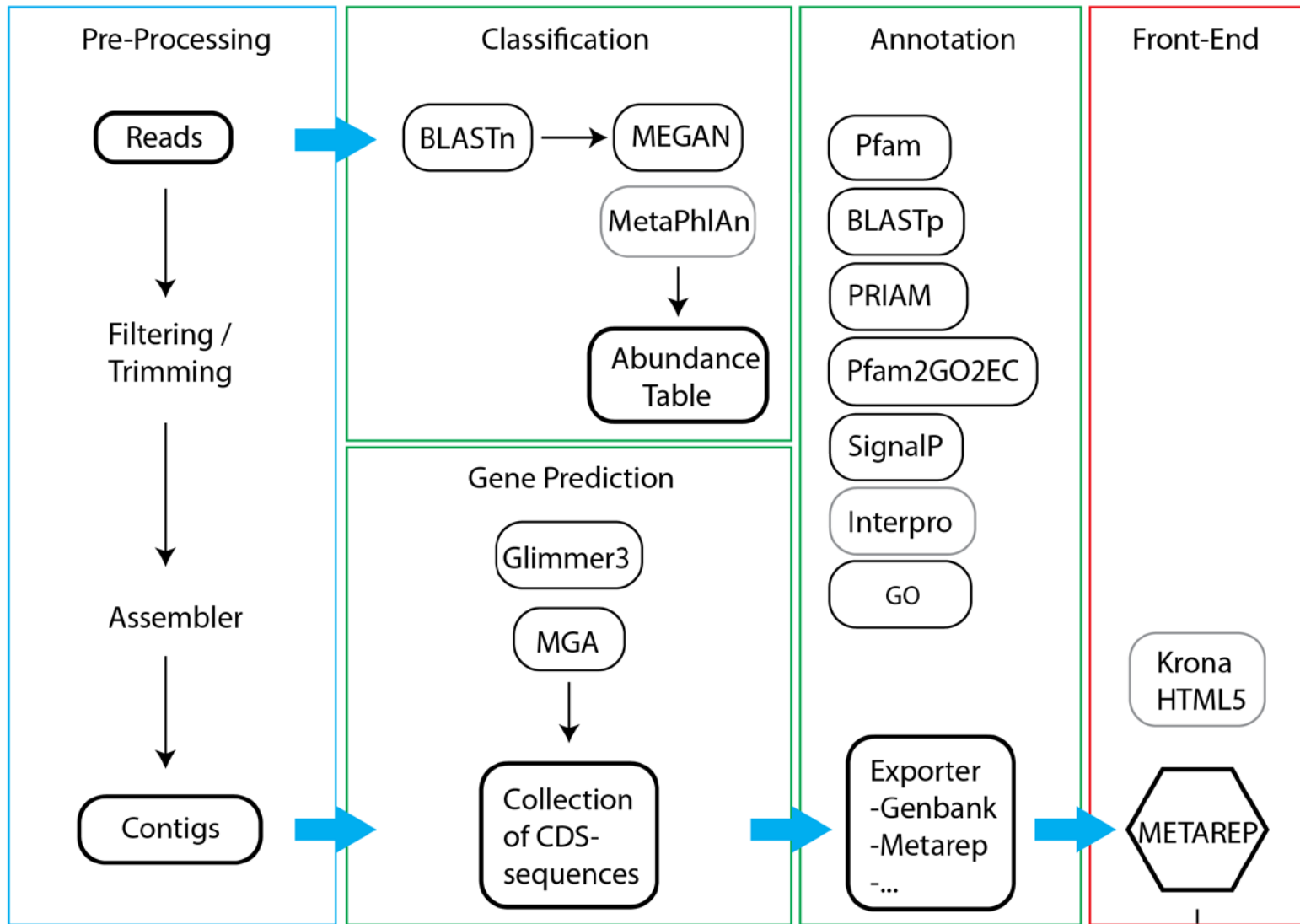
Heidelberg, KB, Gilbert, JA and Joint, I (2010) Marine genomics: at the interface of marine microbial ecology and biodiscovery. *Microb Biotechnol.* 3(5): 531–543

# Biological Data Processing

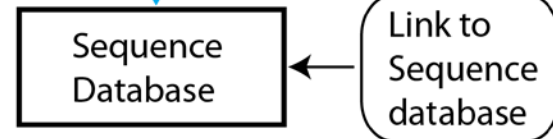
---



# METApipe Runtime System

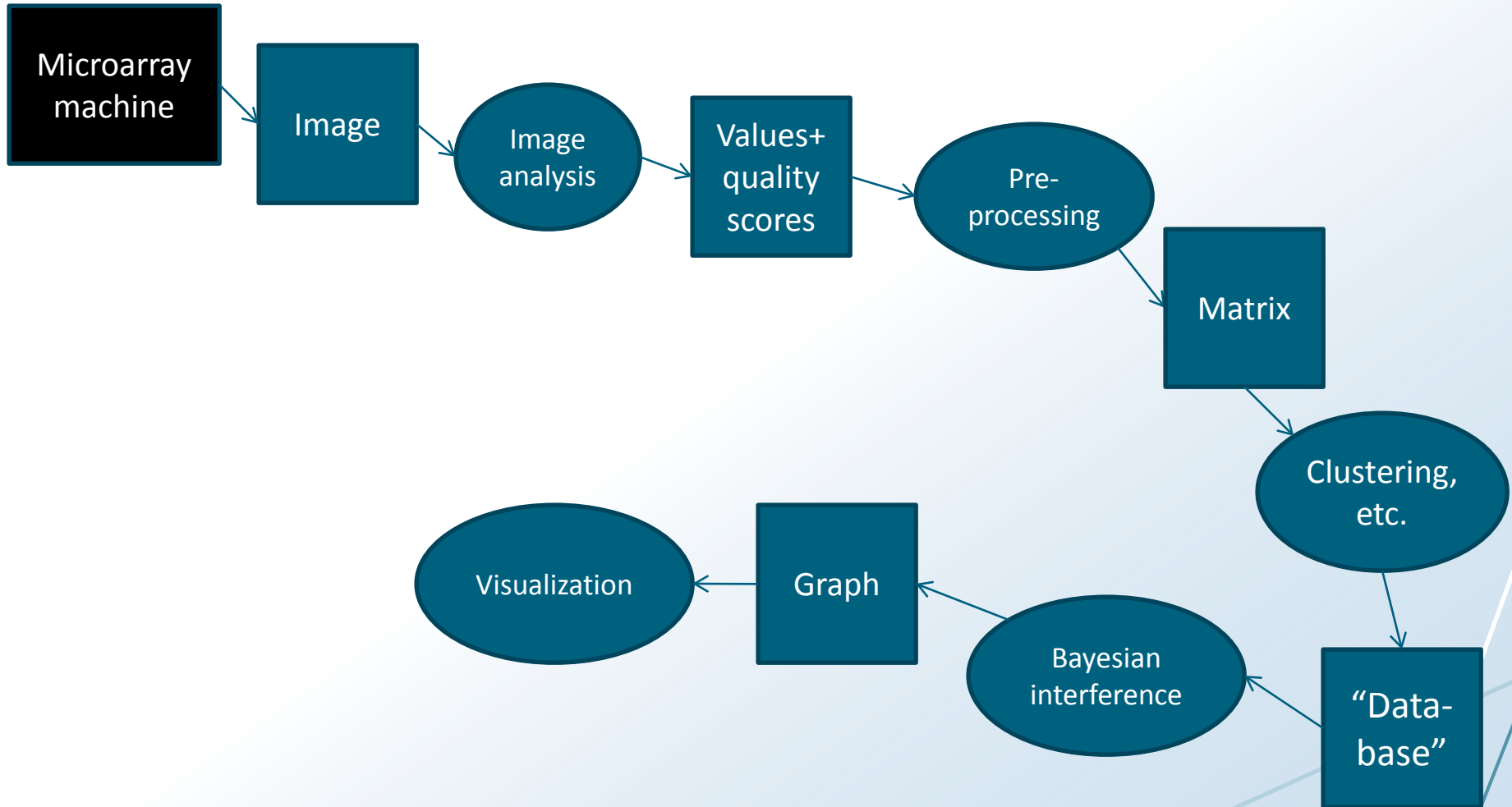


- Modules included
- Scheduled / Work in progress

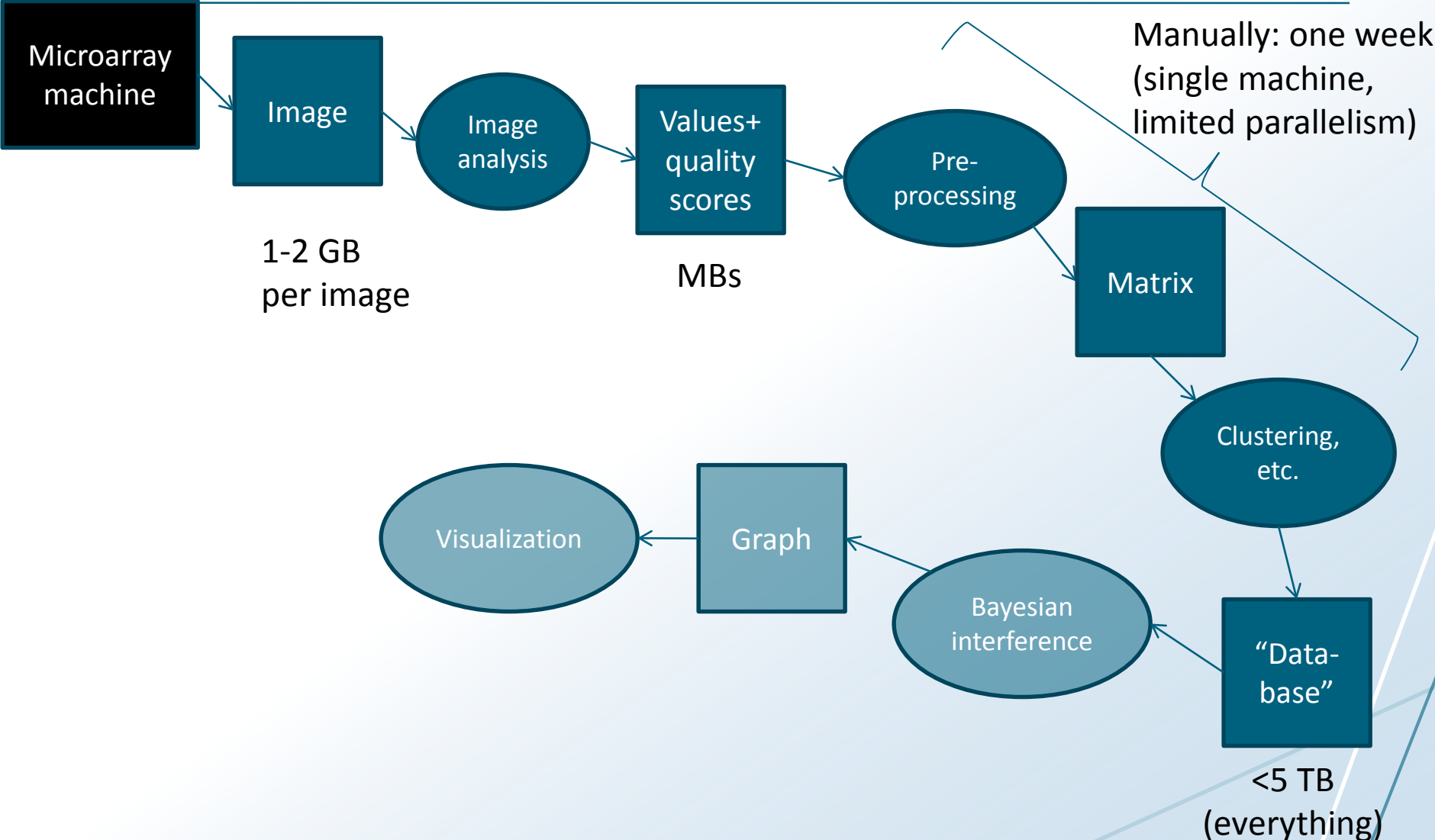


# Microarray Pipeline

---

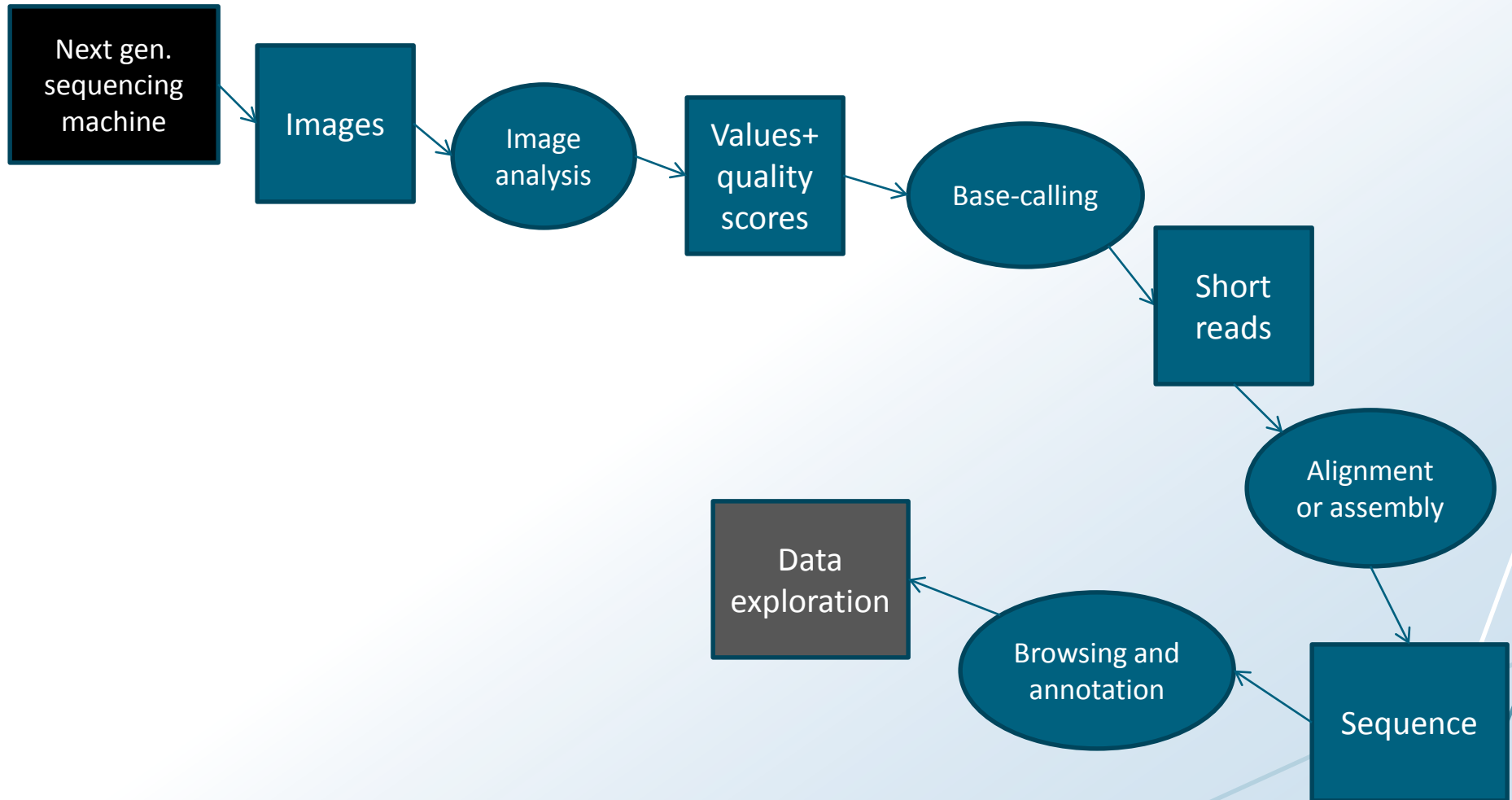


# Microarray Pipeline

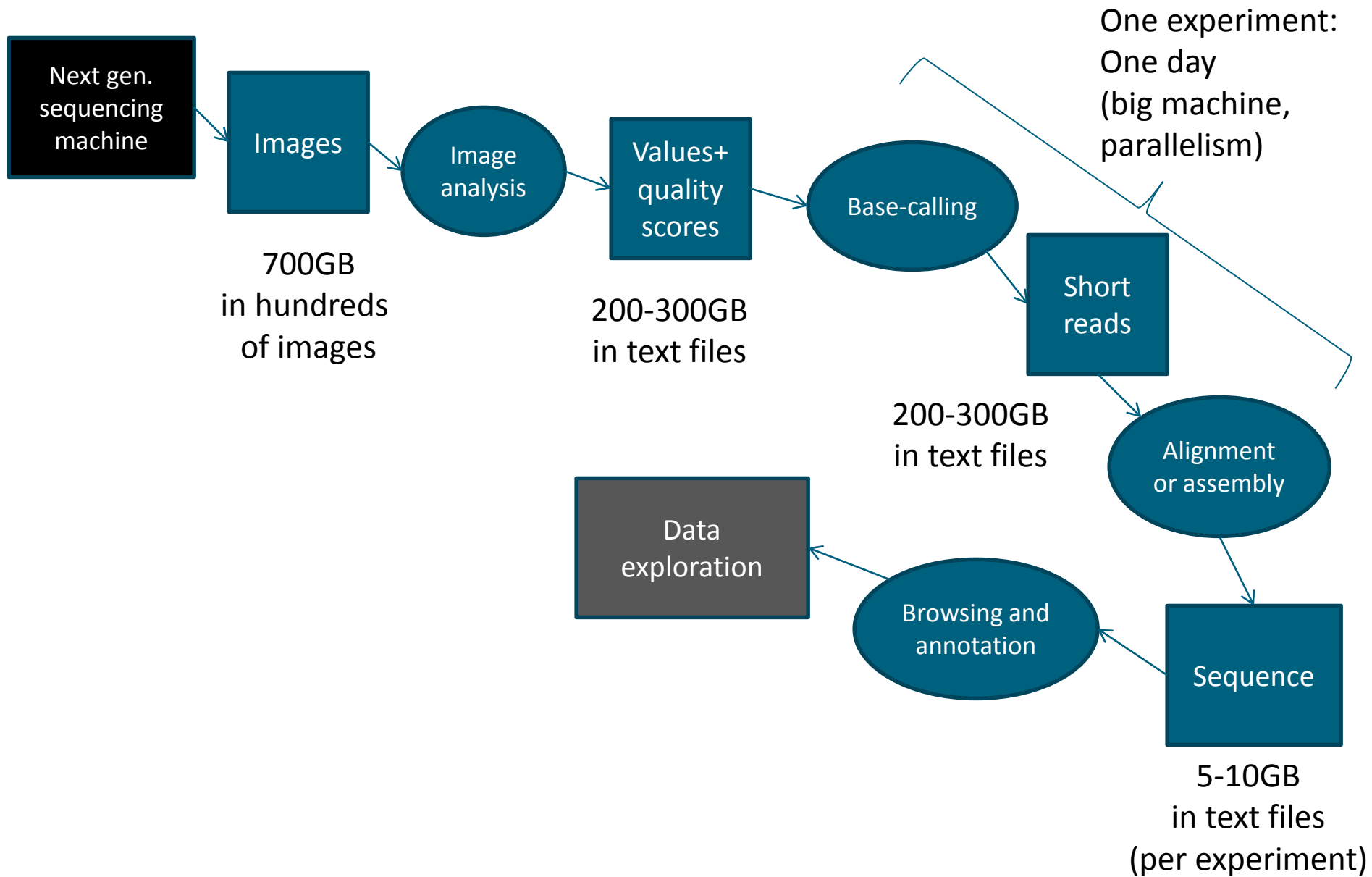


# Next-generation sequencing pipeline

---



# Next-generation sequencing pipeline (per experiment)



# Galaxy

The screenshot displays the Galaxy web interface for editing a workflow titled "Taxonomic Classification". The workflow canvas contains several interconnected tools:

- Input dataset**: Provides the initial data input.
- Filter FASTQ**: Takes a FASTQ file as input and produces a filtered FASTQ file.
- FASTQ to FASTA**: Converts the filtered FASTQ file into a FASTA format.
- Predict 16S rRNA Reads**: Takes a FASTA file as input and predicts 16S rRNA reads.
- Megablast**: Takes a FASTA file as input and performs BLAST-like searches, producing XML output.
- LCAClassifier**: Takes a BLAST XML file as input and generates taxonomic classification outputs, including composition, tree, and assignments.
- Krona**: Takes a tabular file as input and generates hierarchical Krona charts (HTML output).

The **Krona** tool is highlighted with a blue border. The right-hand sidebar provides details for this tool:

- Tool:** Krona
- Version:** 1.0.0
- Input tab file:** Data input 'input' (tabular)
- Edit Step Actions:** Includes a "Rename Dataset" dropdown and a "Create" button.
- Edit Step Attributes:** Includes an "Annotation / Notes" field for adding workflow-specific information.
- Krona Description:** A text box explaining that Krona allows hierarchical data to be explored with zoomable HTML5 pie charts and that it supports various bioinformatics tools and raw data formats.

The left sidebar lists various tool categories such as "Import and Export Data", "Genome annotation", and "Transcriptomics". The top navigation bar includes options for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Admin", "Help", and "User".





**ELIXIR**



*European Life Sciences Infrastructure for Biological Information*  
[www.elixir-europe.org](http://www.elixir-europe.org)

# ELIXIR's mission

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



society

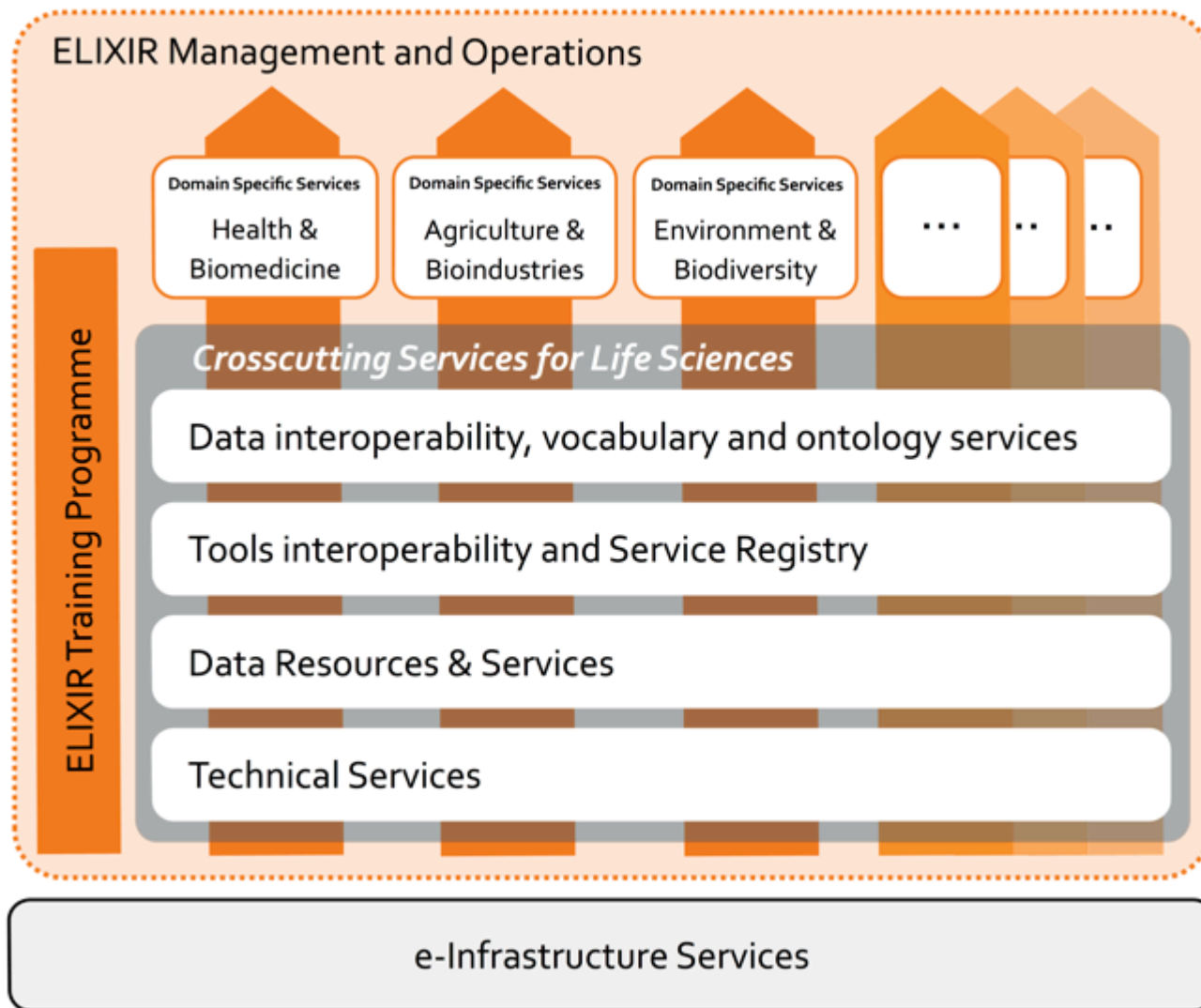
bioindustries

environment

biomedicine

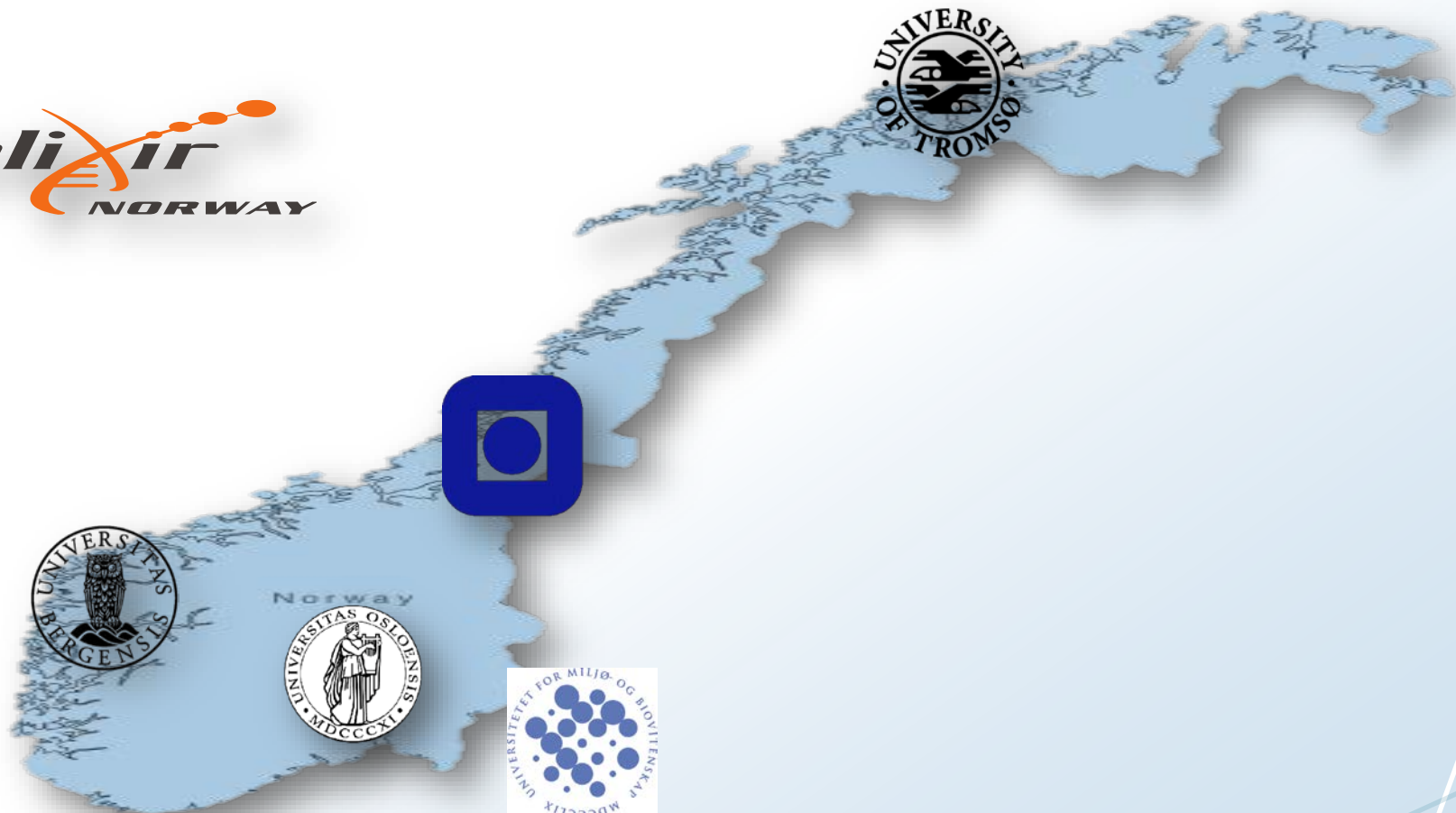


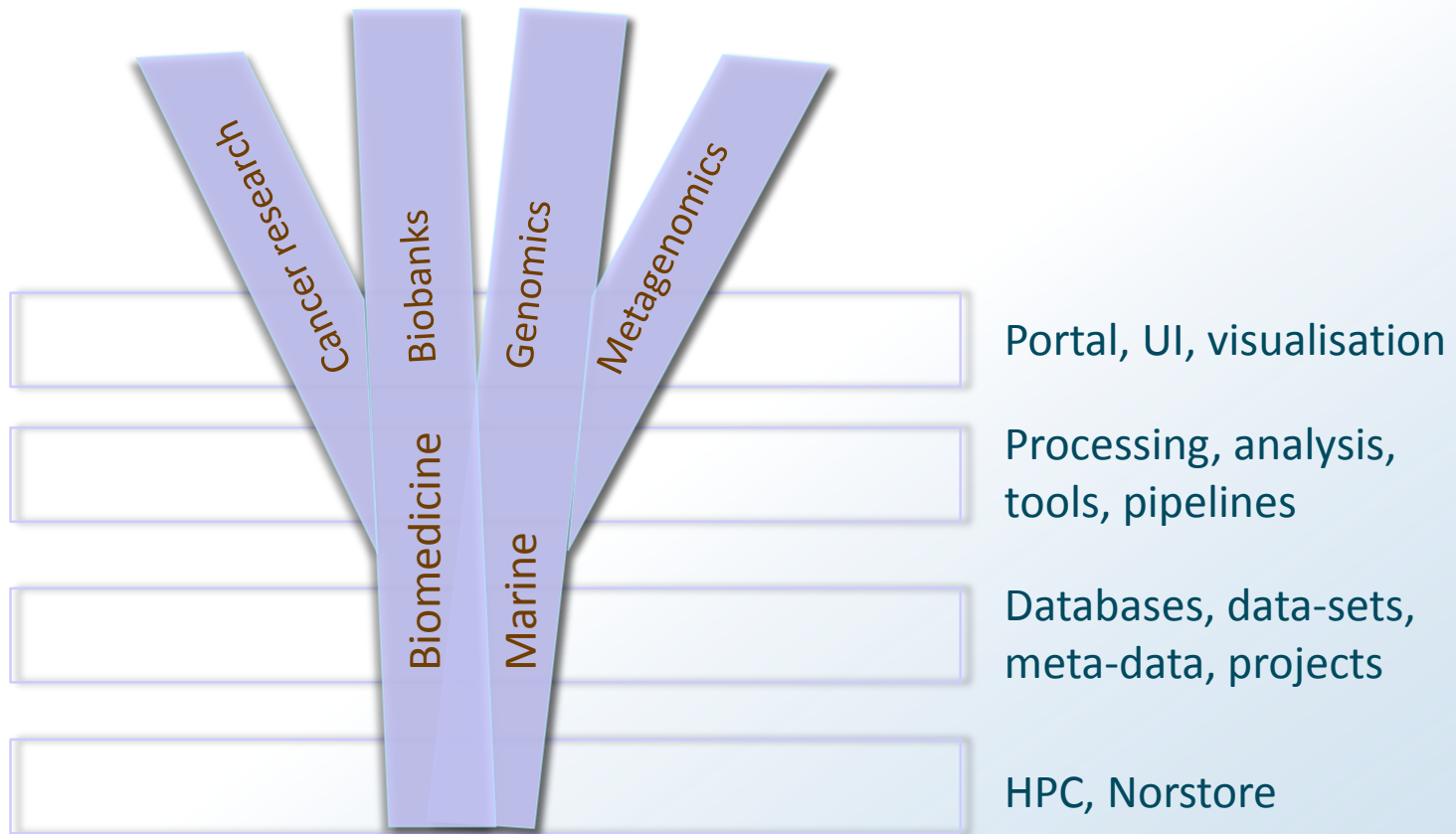
# ELIXIR Programme



# ELIXIR.NO

---





## Tromsø node - Tasks

---

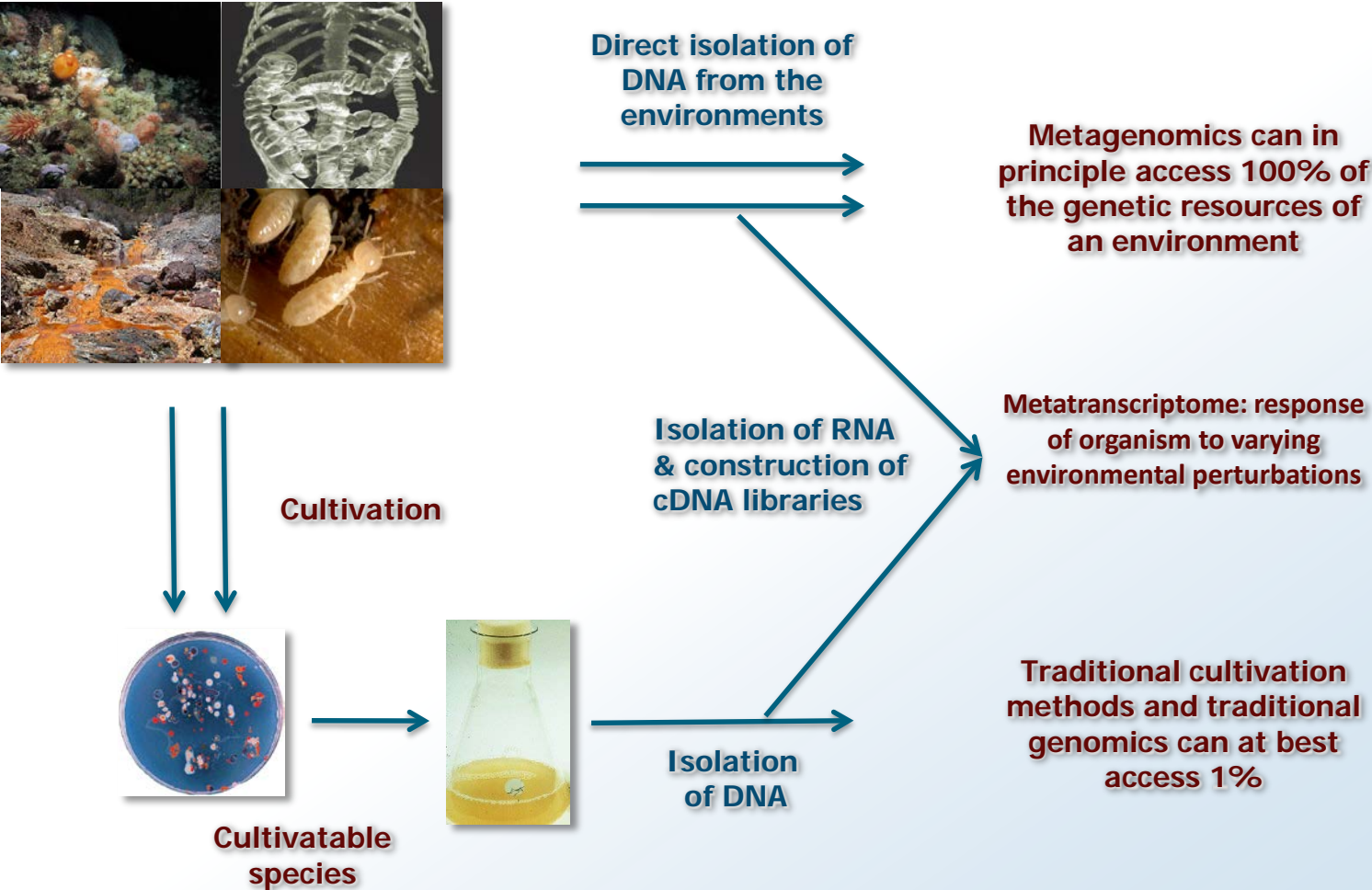
- Build and implement workflows for genomics and metagenomics
- Service towards users
- **Special focus on marine**

# METApipeline - overview

---

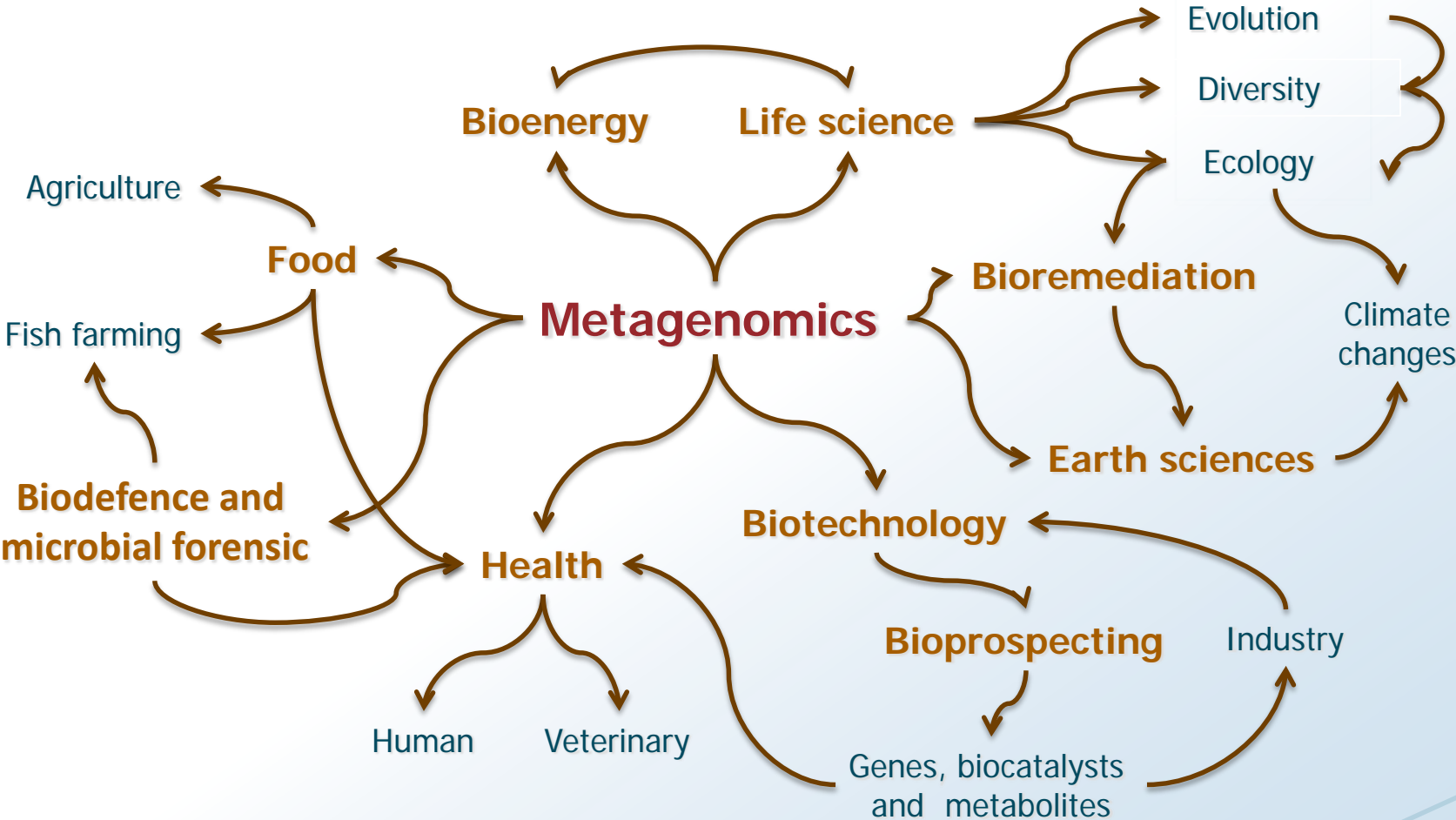
- Metagenomics analysis pipeline
- Pipeline combines
  - Standard bioinformatics tools
  - Custom made tools
- Interactive data exploration
- Deployed at Center for Bioinformatics

# Metagenomics



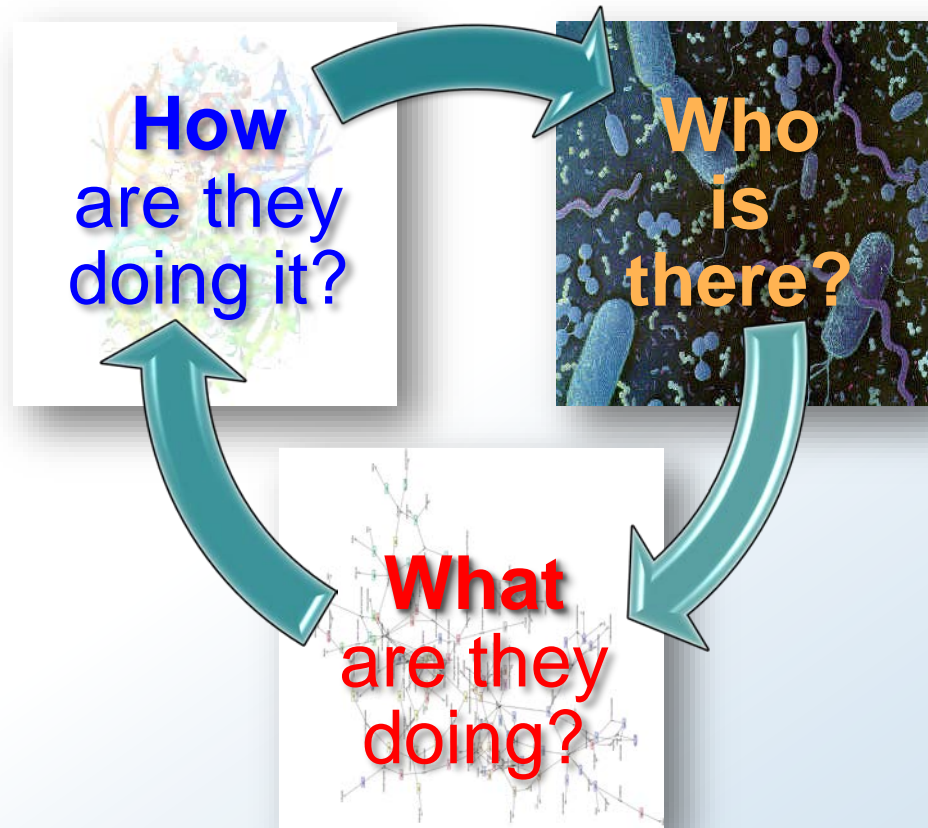


# Metagenomics



# Metagenomics

---





# Sea urchin – novel enzymes

JCVI Metagenomics Reports

SEARCH NEW PROJECT PROJECTS LIST DASHBOARD LOG OUT

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) [3,268 hits]
  - Acting on carbon-phosphorus bonds (3.1.1.-.-) (level 2) [9 hits]
  - Glycosylases (3.2.-.-) (level 2) [1,404 hits]
  - Acting on ether 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) [148 hits]
      - acrosin (3.4.21.10) (level 4) [142 hits]
      - C-terminal processing peptidase (3.4.21.102) (level 4) [55 hits]
      - phyllosain (3.4.21.103) (level 4) [3 hits]
      - rhomoid 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 (3.4.21.111) (level 4) [5 hits]
      - CSa peptidase (3.4.21.110) (level 4) [1 hits]
      - aqueylase 1 (3.4.21.113) (level 4) [4 hits]
      - chymotrypsin C (3.4.21.2) (level 4) [29 hits]
      - cucumisain (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 (3.4.21.9) (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]

**Enzyme Classification**

**Serine endopeptidases**

peptidase Do (3.4.21.107)  
oryzin (3.4.21.63)  
oligopeptidase B (3.4.21.83)  
lysin endopeptidase (3.4.21.50)  
limulus clotting factor\_ overbar\_ B (3.4.21.85)  
endopeptidase La (3.4.21.53)  
cucumisain (3.4.21.25)  
coagulation factor Xa (3.4.21.6)

phyllosain (3.4.21.103)  
prolyl oligopeptidase (3.4.21.26)  
proprotein convertase 2 (3.4.21.94)  
repressor LexA (3.4.21.88)  
rhomoid protease (3.4.21.105)  
signal peptidase I (3.4.21.89)  
subtilisin (3.4.21.62)  
thermitase (3.4.21.66)  
thrombin (3.4.21.5)

trypsin (3.4.21.4)  
C-terminal processing peptidase (3.4.21.102)  
CSa peptidase (3.4.21.110)  
Transferred to 3.4.21.34 and 3.4.21.35 (3.4.21.8)  
Transferred to 3.4.21.37 (3.4.21.111)  
chymotrypsin (3.4.21.1)  
chymotrypsin C (3.4.21.2)  
aqueylase 1 (3.4.21.113)  
acrosin (3.4.21.10)

**Top Ten Functional Classifications**

| Species (Blast)                                    | Common Name                               | Gene Ontology  | Enzyme   | HMM                            |
|--|---|--|--|--------------------------------|
| 1. <i>Colwellia psychroerythraea</i> (17,450) (59) | 1. Signal peptidase I (6,350) (30)        | 1. GO:0004252   serine-type endopeptidase activity (46,230) (262)  | 1. 3.4.21.105   rhomoid protease Do (10,580) (60)            | 1. Peptidase_S8 (9,170) (52)   |
| 2. <i>unresolved</i> (14,400) (82)                 | 2. ATP-dependent protease La (4,430) (25) | 2. unassigned (45,159) (256)                                       | 3. 3.4.21.102   C-terminal processing peptidase (9,700) (55) | 2. Peptidase_S9_N (4,050) (23) |
| 3. <i>Psychromonas ingrahamii</i> (4,416) (25)     | 3. Lon protease (4,230) (24)              | 3. GO:0005008   proteolysis (33,530) (190)                         | 4. 3.4.21.89   signal peptidase I (6,880) (39)               | 3. Trypsin (3,790) (21)        |
| 4. <i>Oleispira antarctica</i> (2,230) (24)        | 4. Uncharacterized protein (4,230) (24)   | 4. GO:0016021   integral to membrane (630) (34)                    | 5. 3.4.21.53   endopeptidase La (5,290) (30)                 | 4. Rhomoid (3,530) (20)        |
| 5. <i>Fluviicola taftensis</i> (3,170) (18)        | 5. LexA repressor (3,350) (19)            | 5. GO:0013968   RNA-directed RNA polymerase activity (5,530) (20)  | 6. 3.4.21.88   repressor LexA (5,130) (24)                   | 5. Lon_C (2,820) (16)          |
| 6. <i>Psychromonas</i> sp. CNPT3 (1,940) (11)      | 7. Endopeptidase Ctp (2,820) (16)         | 6. GO:0016020   membrane (3,170) (18)                              | 7. 3.4.21.105   rhomoid protease (4,230) (24)                | 6. Peptidase_S24 (2,820) (16)  |
| 7. <i>Haliomenobacter hydroxilis</i> (1,760) (10)  | 8. Prolyl endopeptidase (2,290) (13)      | 7. GO:0017111   nucleoside-triphosphatase activity (3,170) (18)    | 8. 3.4.21.26   prolyl oligopeptidase (4,000) (23)            | 7. DUF1034 (2,470) (14)        |
| 8. <i>Pelagibacterium halotolerans</i> (1,590) (9) | 9. Rhomoid family protein (2,290) (13)    | 8. GO:0004376   ATP-dependent peptidase activity (2,820) (16)      | 9. 3.4.21.83   peptidase B (3,880) (22)                      | 8. LexA_DNA_bind (2,470) (14)  |
| 9. <i>Polaribacter</i> sp. MED152 (1,430) (8)      | 10. Protease II (2,120) (12)              | 9. CO:0005618   cell wall (2,470) (14)                             | 10. 2.7.7.48   RNA-directed RNA polymerase (3,530) (20)      | 9. CLP_protease (2,120) (12)   |
| 10. <i>Bacillus subtilis</i> (1,230) (7)           |   | 10. GO:0004197   cysteine-type endopeptidase activity (2,290) (13) |  | 10. Peptidase_S46 (2,120) (12) |

**Top Ten Functional Pie Charts**

**Species (Blast)**

**Common Name**

**Gene Ontology**

**Enzyme**

**HMM**

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@jcv.org](mailto:metarep-support@jcv.org)

# MabCent – hunting for novel enzymes

## Browse Kegg Pathways (EC) MabcentResults (MabCent)

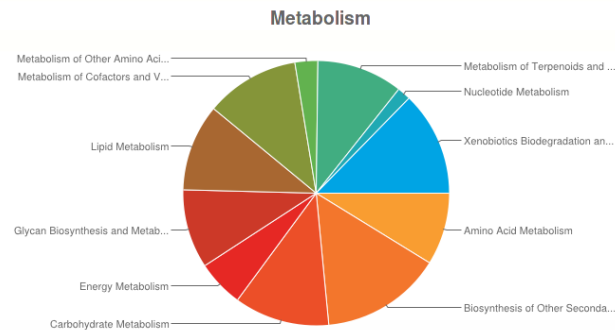
Filter

Help \*:\* Filter

### Browse Kegg Pathways (EC)

- Metabolism [841 hits]**
  - Carbohydrate Metabolism [594 hits]
  - Energy Metabolism [293 hits]
  - Lipid Metabolism [550 hits]
  - Nucleotide Metabolism [86 hits]
  - Amino Acid Metabolism [450 hits]
  - Metabolism of Other Amino Acids [147 hits]
  - Glycan Biosynthesis and Metabolism [499 hits]
  - Metabolism of Cofactors and Vitamins [591 hits]
  - Metabolism of Terpenoids and Polyketides [539 hits]
  - Biosynthesis of Other Secondary Metabolites [756 hits]
  - Xenobiotics Biodegradation and Metabolism [655 hits]
- Genetic Information Processing [5 hits]
  - Translation [5 hits]

### Pathway Classification



## Browse Kegg Pathways (EC) MabcentResults (MabCent)

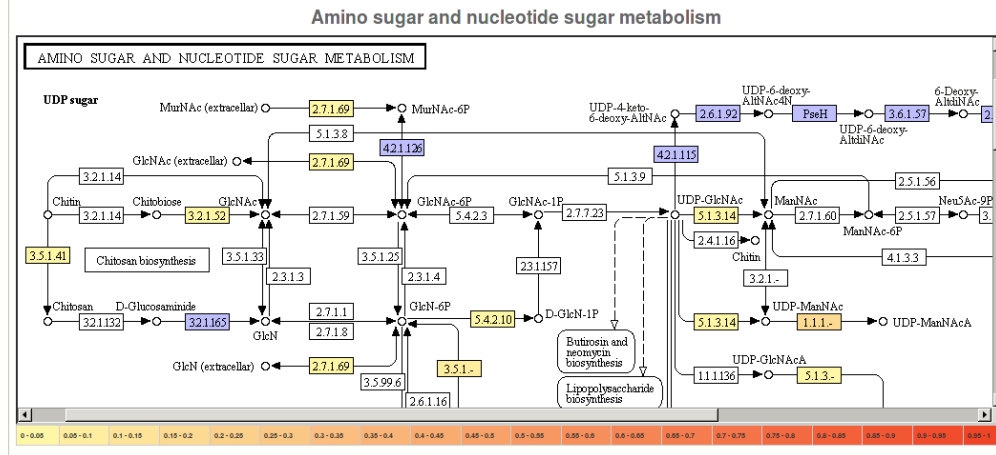
Filter

Help \*:\* Filter

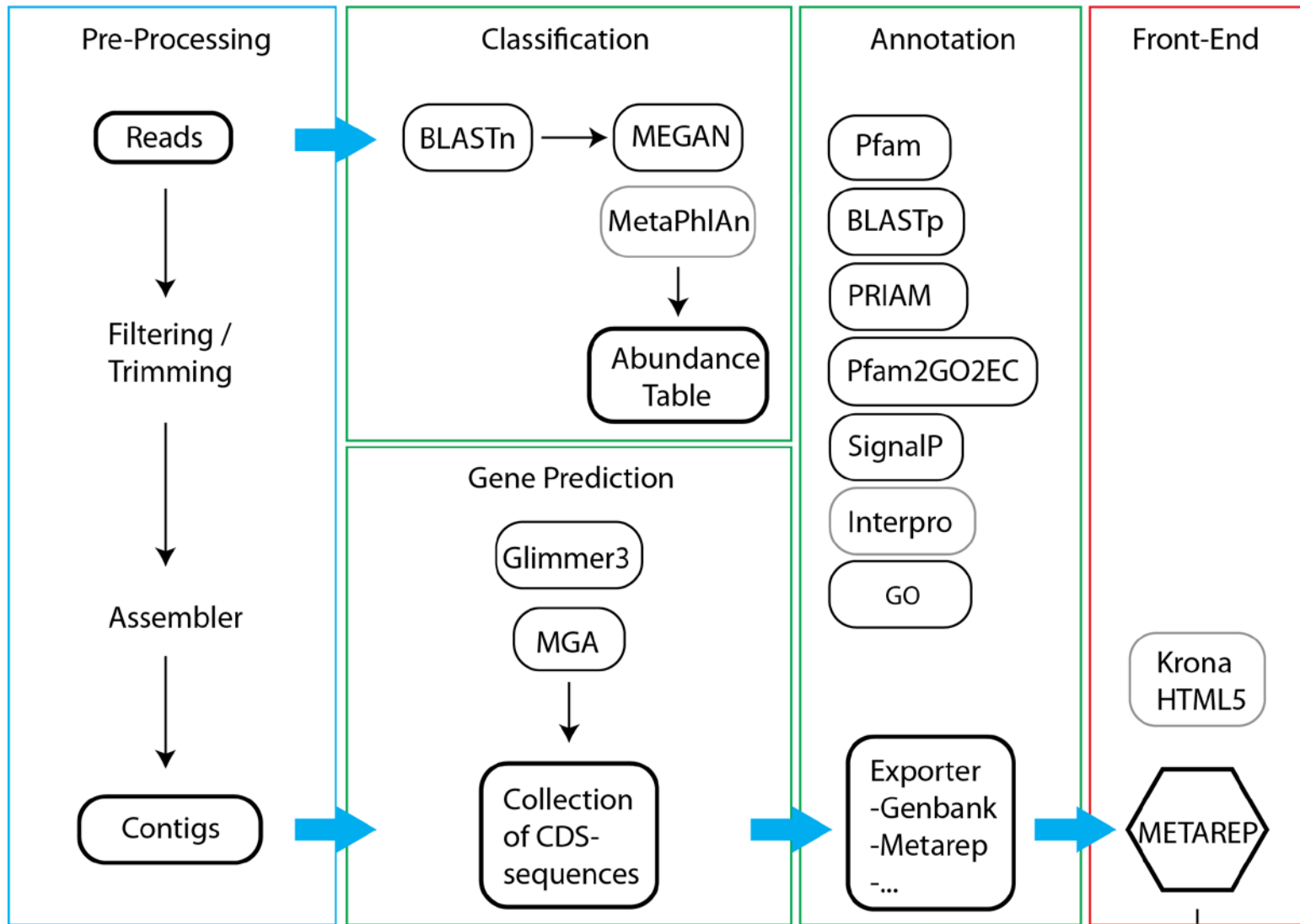
### Browse Kegg Pathways (EC)

- Metabolism [841 hits]**
  - Carbohydrate Metabolism [594 hits]
    - Glycolysis / Gluconeogenesis (pathway) [38 hits]
    - Citrate cycle (TCA cycle) (pathway) [12 hits]
    - Pentose phosphate pathway (pathway) [34 hits]
    - Pentose and glucuronate interconversions (pathway) [58 hits]
    - Fructose and mannose metabolism (pathway) [99 hits]
    - Galactose metabolism (pathway) [9 hits]
    - Ascorbate and aldarate metabolism (pathway) [226 hits]
    - Starch and sucrose metabolism (pathway) [12 hits]
  - Amino sugar and nucleotide sugar metabolism (pathway) [83 hits]**
    - Oxidoreductases (1.-.-.-) (enzyme) [279 hits]
    - With NAD+ or NADP+ as acceptor (1.-1.-.-) (enzyme) [83 hits]
    - GDP-mannose 6-dehydrogenase. (1.1.1.132) (enzyme) [0 hits]
    - GDP-6-deoxy-D-talose 4-dehydrogenase. (1.1.1.135) (enzyme) [0 hits]
    - UDP-N-acetylglucosamine 6-dehydrogenase. (1.1.1.136) (enzyme) [0 hits]
    - UDP-N-acetylmuramate dehydrogenase. (1.1.1.158) (enzyme) [0 hits]
    - GDP-4-dehydro-D-thiamine reductase. (1.1.1.187) (enzyme) [0 hits]
    - UDP-glucose 6-dehydrogenase. (1.1.1.22) (enzyme) [0 hits]
    - GDP-L-fucose synthase. (1.1.1.271) (enzyme) [0 hits]
    - GDP-4-dehydro-6-deoxy-D-mannose reductase. (1.1.1.281) (enzyme) [0 hits]

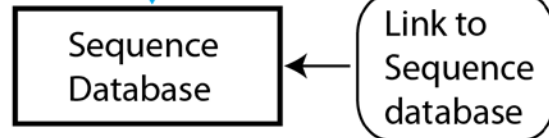
### Pathway Classification



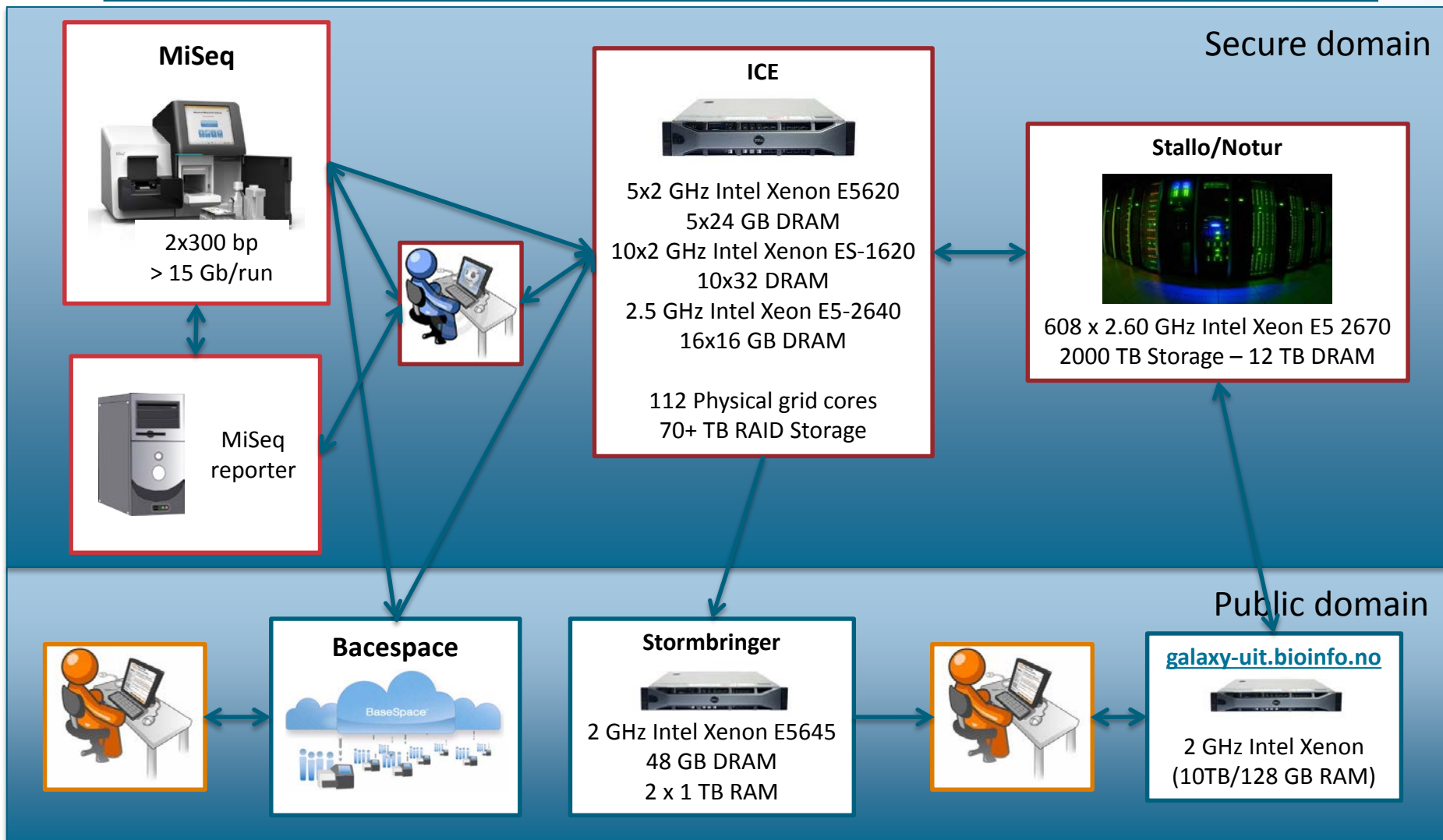
# METApipe Runtime System



- Modules included
- Scheduled / Work in progress



# Infrastructure @ UIT



# 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

---

## **METApipeline: 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

