



# Meta-pipe analysis pipeline

Espen Mikal Robertsen, UiT – The Arctic University of Norway  
Tromsø, 2018

# Outline

- Motivation / Why Meta-pipe
- Steps prior to bioinformatic analysis
- Overview of Meta-pipe
- Interfaces

# Motivation behind META-pipe

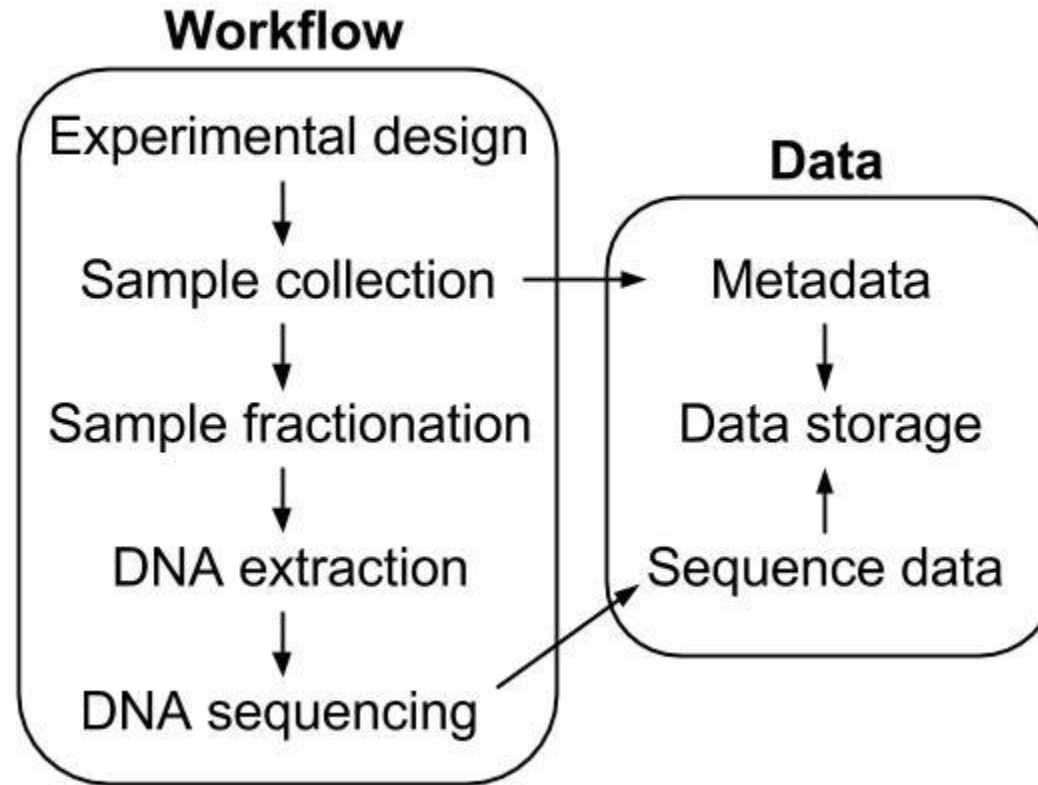
- Lack of metagenomics resources for the marine domain
- “Lack” of metagenomics resources for analysis of full-length genes
- User simplicity
- Resource availability

Mineta, Katsuhiko, Gojobori, Takashi 2016 Gene, ISSN: 1879-0038, Vol: 576, Issue: 2 Pt 1, Page: 724-8

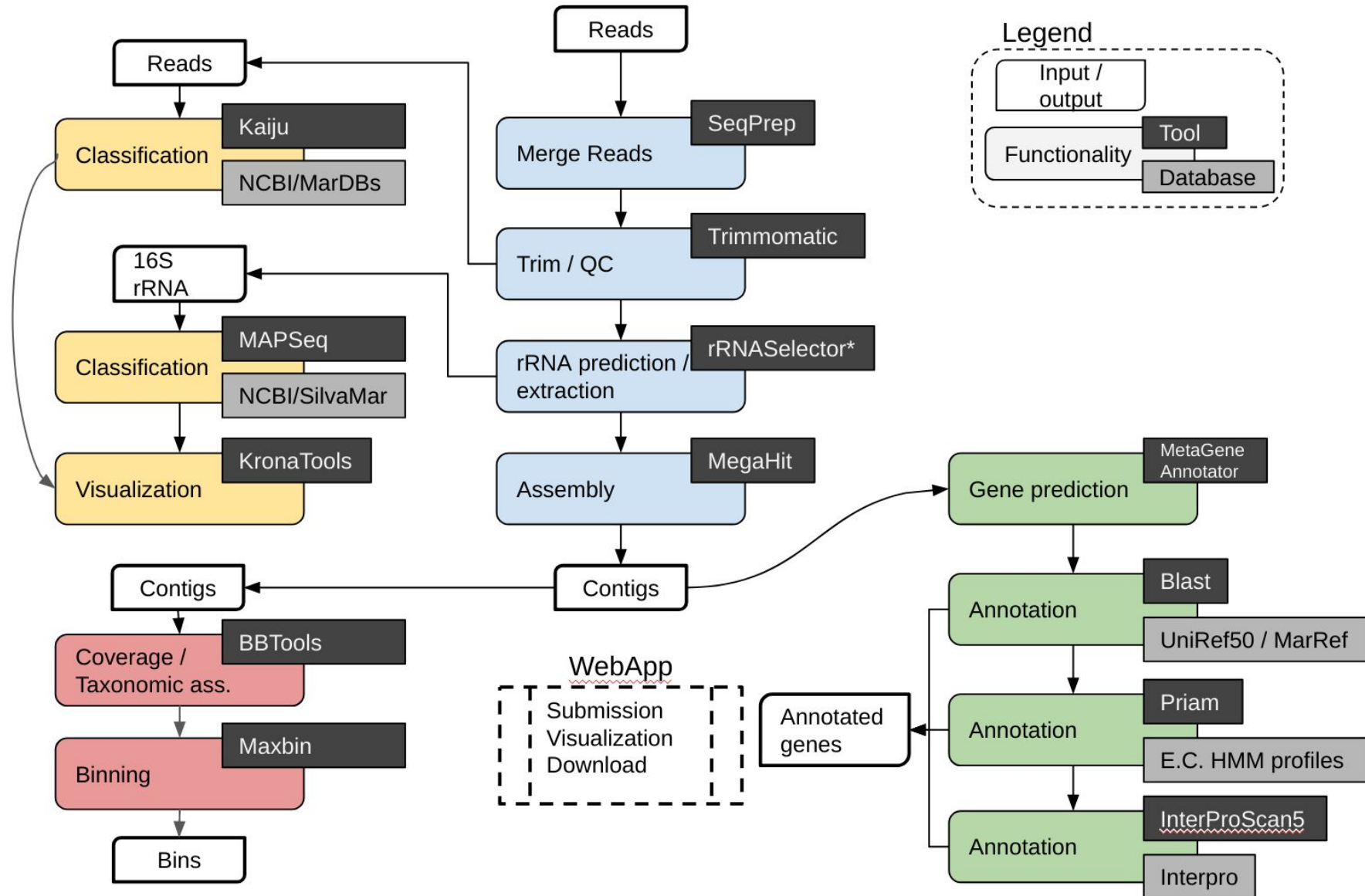
Tromsø, May, 2018



# Steps prior to bioinformatic analysis



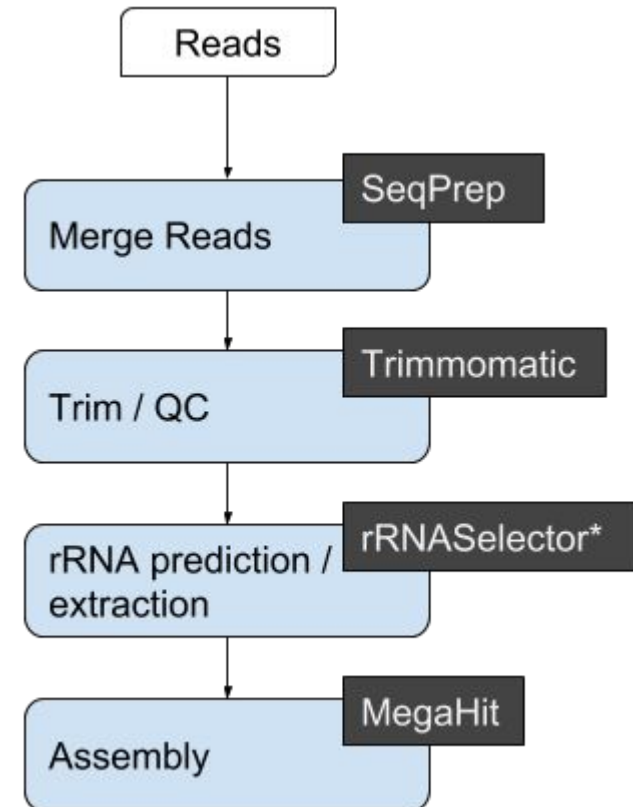
# Biological Overview of Meta-pipe



# Biological Overview

## Preprocessing

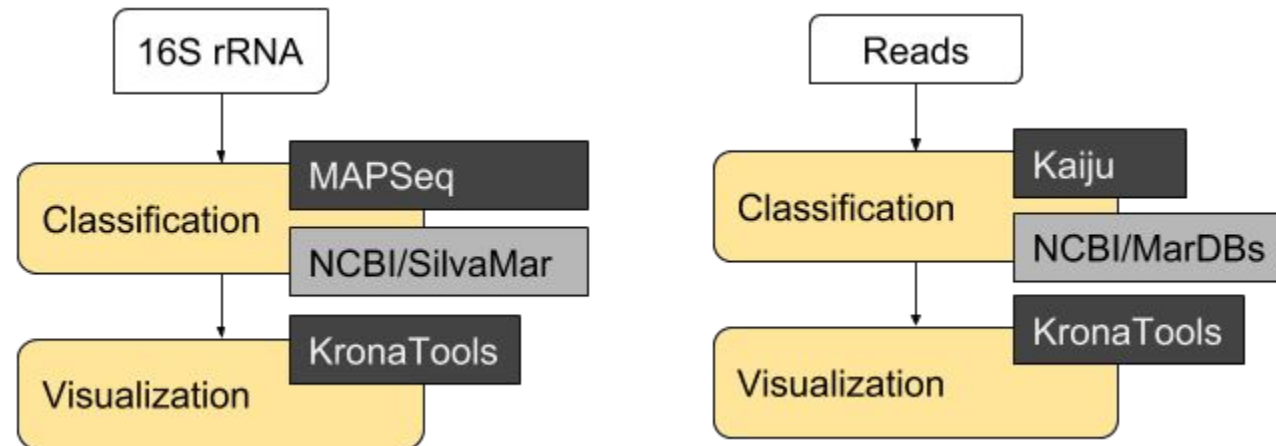
- SeqPrep – Merging of Illumina reads
- Trimmomatic – QC of reads
- (DeconSeq) – Filter unwanted sequences with reference
- rRNASelector – Removal/Extraction of rRNA sequences
- MEGAHIT – Assembly



# Biological Overview

## Taxonomic Classification

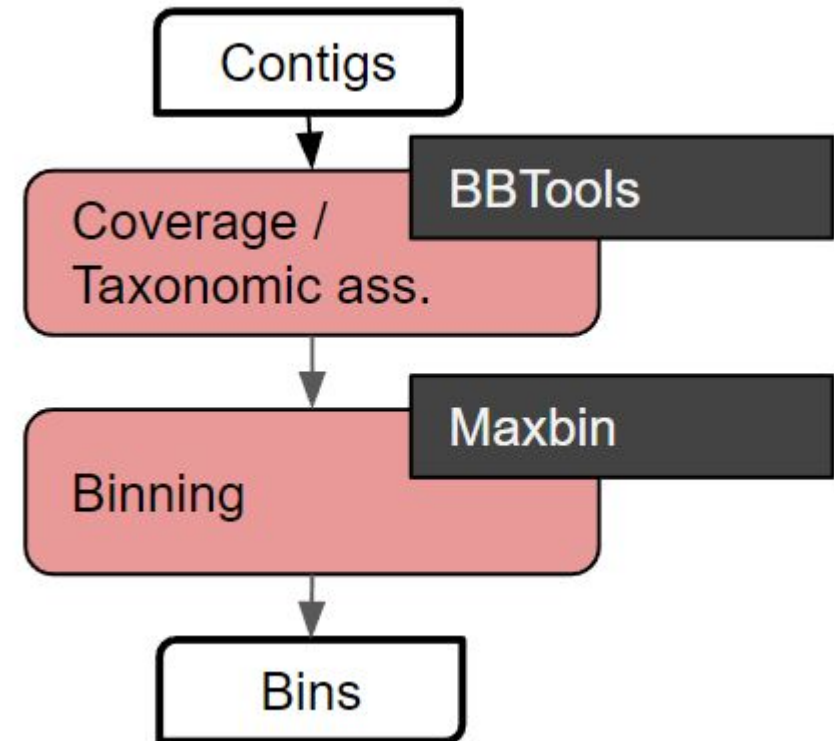
- Kaiju - Classification of reads using protein reference databases
- Mapseq - High throughput rRNA sequence classification
- KronaTools – Visualization of output



# Biological Overview

## Binning

- BMap - Coverage calculation
- MaxBin - Binning of contigs
- BBSketch - Taxonomic assignment of bins



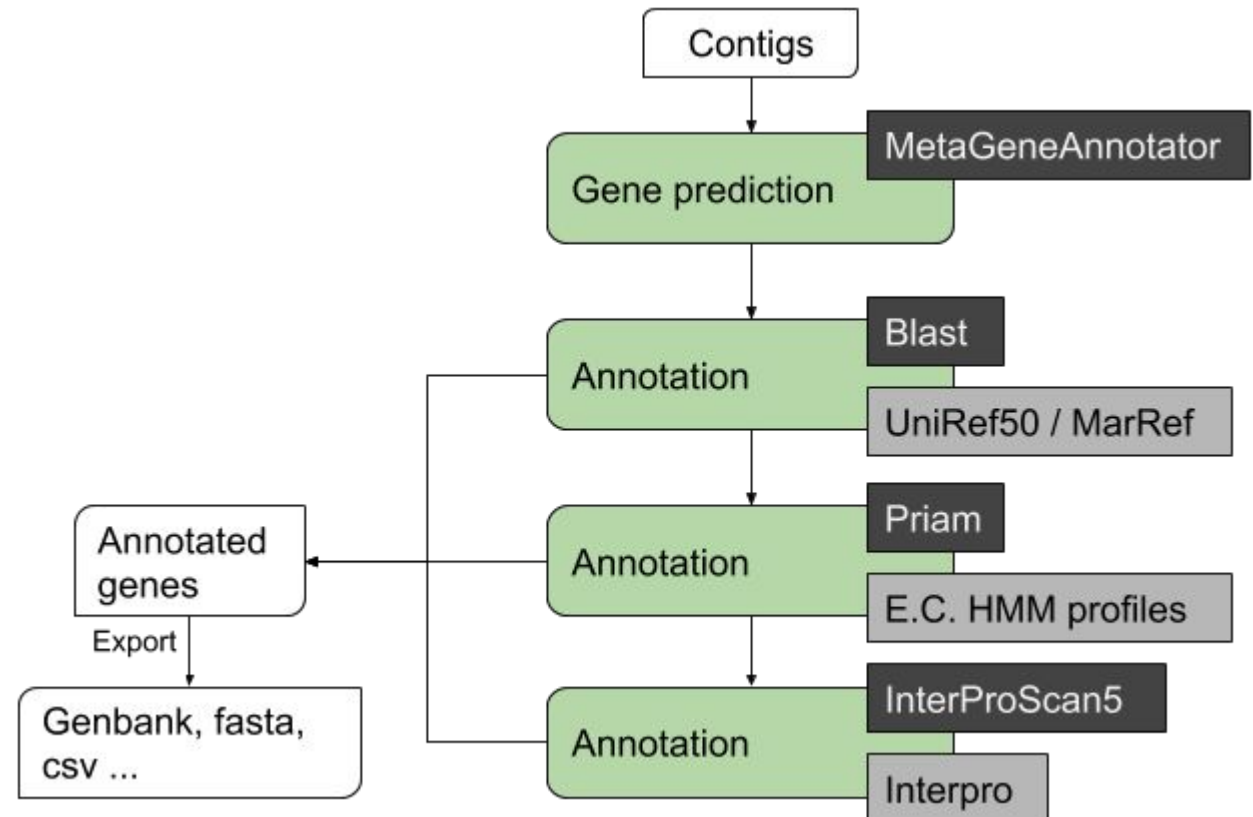




# Biological Overview

## Functional Analysis

- PRIAM – Annotation of EC-numbers with RPS-blast
- Uniref50 – Clustered UniProt using Blast
- InterPro – Collection of 14 databases queried with InterProScan



# Biological Overview

## Functional Analysis

Artemis Entry Edit: export (1).gb

File Entries Select View Goto Edit Create Run Graph Display

Entry:  export (1).gb

Nothing selected

Artemis Release 16.0.0  
1. Standard

wellcome trust  
sanger  
institute  
pathogen genomics group

fasta\_record

4800 5600 6400 7200 8000 8800 9600 10400 11200 12000

k99\_1114 gene 5 k99\_1114 gene 6 k99\_1114 gene 7 k99\_1114 gene 8 k99\_1114 gene 9 k99\_1114 gene 10 k99\_1114 gene 11 k99\_1114 gene 12 k99\_1114 gene 13 k99\_1114 gene 14 k99\_1114 gene 15

T # D \* P T G V P R \* K S + F W P H C L V L S P L A P L P O K K S A L A F L Q L T H R L N R W T P I T K S P V S T S T W P T R C A N R C R P S  
H K I D R R E F H D E K V S S G R T A + F E R L W R R C R R K N Q L W R F C N L P T V \* I D G R Q # P N R R F R H R P G Q R A V Q T D A G R V  
I R L L T D G S S T H K K L L V L L A A L L S E A F I G A A A A F K I L S F G V S A T V P P F E S Y D A N N Q T A G E D I D L A N A L G T Q M G A E  
ACATAAGATTGACCGACGGAGTTCCACGATGAAAAGTITAGTCTGGCCACTGCTTGTCTTTCGCTTGGCCCGCTGCCGACAGAAAATCACTTGGCGTTCTGCAACTTACCCACCGTTGAATGATGGACGCCAATAACCAATCGCCGTTTCGACATCGACCTGGCCACCGGCTGTGCAACAGATGCAGCCSAGT  
TGTATTCTAACTGGCTGCCCTCAAGTGTACTTTTTCAATCAAGACCGCGTGACGAATCAAGAAAGCGGAAACCGCGCGAAGCGGCTCTTTTTAGTGGAAACCGCAAGACGTTGAATGGGTGGCAAACTTAGCTACCTGGGTTATTGGTTTAGCGGCCAAAGCTGTAGCTGGACCGGTTGGCGGACACGTTTGTCTACGTCGGGCTCA  
M L N V S P L E V I F F N T R A A S S L E K A K P A A A A S F I L K P T E A V # G N S D I S A L L W I A P K S M S R A L A S H L C I C A S H  
Y S Q G V P T G R H F L # N Q G C Q K T R E G K A G S G C F F D A K A N R C S V W R K F R H V G I V L D G T E V D V Q G V R Q A F L H L G L  
C L I S R R S N W S S F T L E P R V A # N K R R Q R R Q R L F F \* S Q R K Q L K G V T Q I S P R W Y G F R R N R C R G P W R A T C V S A P R T

source	1	108378
fasta_record	1	41673
CDS	30	761
CDS	891	2363
CDS	2360	3085
CDS	3155	4285
CDS	4328	4816
CDS	4874	5674
CDS	5716	6869
CDS	6890	7858
CDS	6890	7858
CDS	7816	9028
CDS	7816	9028
CDS	9377	9856

- Arginine ABC transporter
- Osmosensitive K+ channel histidine kinase KdpD
- Response regulator receiver domain protein
- 23S rRNA (uracil(747)-C(5))-methyltransferase RlmC
- Inner membrane protein YbjO
- Putrescine transport system permease protein PotI
- Putrescine transport system permease protein PotH
- Putrescine transport ATP-binding protein PotG
- Putrescine transport ATP-binding protein PotS
- Putrescine-binding periplasmic protein
- Putrescine-binding periplasmic protein
- Uncharacterized protein ybjN

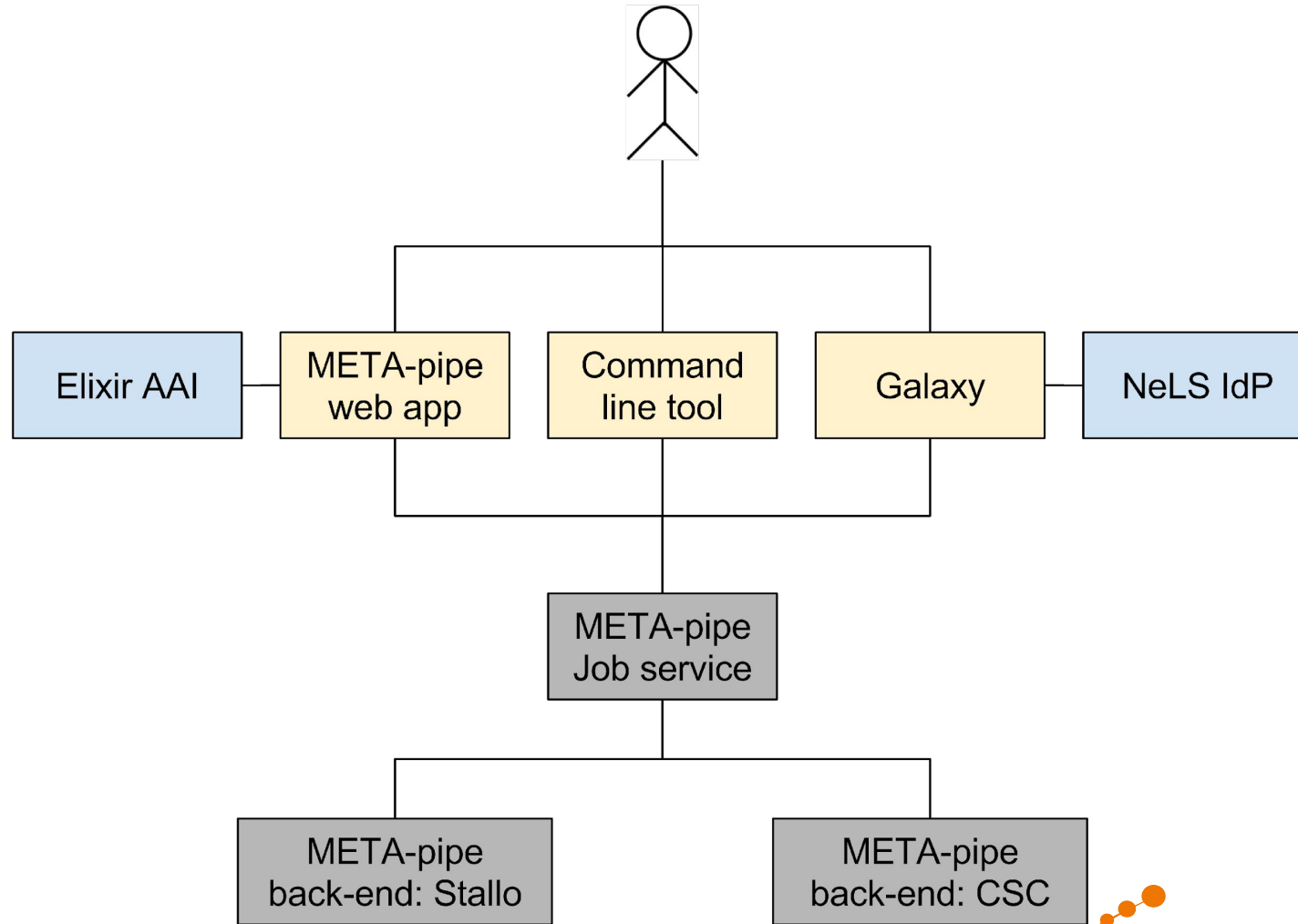
# Inclusion and processing of MarDBs

- MarRef integrated with Meta-pipe
- MarDB scheduled for integration with Meta-pipe
- Meta-pipe currently processing samples for MarCat



**MARINE  
REFERENCE  
DATABASES**

# User Access



# User Access Galaxy with NeLS-idp

The screenshot displays the Galaxy web interface with the following components:

- Header:** Galaxy / uit, navigation tabs (Analyze Data, Workflow, Shared Data, Visualization, Help, User), and a memory usage indicator (Using 88.3 GB).
- Tools Panel (Left):** A sidebar with a search bar and categories: Get Data, Send Data, Text Manipulation, Filter and Sort, Join, Subtract and Group, Metagenomics, Statistics, Meta-pipe (with 'Meta-pipe 2.0 (beta)' selected), UiT, FASTA manipulation, NGS: QC and manipulation, NGS: Mapping, NGS: SAM Tools, NGS: GATK Tools (beta), Transcriptomics, NGS: Picard, Assembly, and Annotation.
- Tool Configuration (Center):** 'Meta-pipe 2.0 (beta) (version 0.1.0)'.
  - input1:** 7: Meta-pipe 2.0 (beta) on data 1
  - Cutoff for assembled contigs:** 500 (with a note: 'Remove any contigs that are shorter than this length')
  - Database search tools:** Select All, Unselect All, Blast+/Uniref50, InterProScan5, PRIAM.
  - Enable developer options (not supported):** no (recommended) (with a note: 'This is for internal development use only.')
  - Execute** button.
- History Panel (Right):** Shows a job titled '7: Meta-pipe 2.0 (beta) on data 1' with 95 sequences in fasta format. It includes a progress bar and a list of log messages:
  - 16:28:47.458 [main] INFO n.u.m.c.client.StorageClientImpl - Uploading dataset (105.958984375 KiB) to https://storage.metapipe.uit.no/system\_
  - 21:05:02.940 [main] INFO n.u.m.c.client.StorageClientImpl - Downloadin

# User Access

## Webapp with ELIXIR-AAI



Welcome to the Marine Metagenomics Portal (MMP).  
We provide data resources and analysis services targeted for the marine domain.  
MMP is developed with support from the ELIXIR-EXCELERATE grant and ELIXIR Norway.



### MARINE REFERENCE DATABASES

Marine reference databases are richly annotated and manually curated contextual and sequence databases. [MarRef](#) contains completely sequenced marine prokaryotic genomes. [MarDB](#) includes all sequenced marine prokaryotic genomes regardless of level of completeness. [MarCat](#) is a catalogue of marine genes and proteins derived from metagenomics samples.

Browse



### META-PIPE

META-pipe is a complete workflow for the analysis of marine metagenomic data. It provides assembly of high-throughput sequence data, functional annotation of predicted genes, and taxonomic profiling. META-pipe is not released as an ELIXIR service yet. For now you may use the [NeLS META-pipe service](#).

Run



### MMP BLAST


MMP BLAST provides BLAST search on all genes and protein coding sequences from the marine reference databases ([MarRef](#), [MarDB](#) and [MarCat](#)).

BLAST




# User Access


## Webapp with ELIXIR-AAI

 Meta-pipe [About](#) [Contact](#) [Meta-pipe](#) | Espen Mikal Robertsen ▾

### Start Meta-pipe

Select dataset

  
Upload new

  
Choose existing

Select parameters

**Executor**

**Cutoff**

Start Meta-pipe

© UiT - The Arctic University of Norway · [Home](#) · [Privacy](#) · [Not Found](#) | Viewport: 1920x1020



# User Access

## Webapp with ELIXIR-AAI

Select parameters

Quality control and assembly

**Cutoff**  
500

Remove non-complete genes

**Minimum contig length**  
1000

Taxonomic classification

MapSeq    Silva    Kaiju    Default (NR)  
 Silva Mar    MarDB

Functional assignment

Uniref50 / Blast+  
 Interpro scan5  
 Priam  
 MarRef / Blast+

Output

Create a Genbank file containing all the contigs and annotations merged together as a single entry

Start Meta-pipe


Enable dev tools

*ate*









Tromsø, May, 2018

# Output

 Meta-pipe [About](#) [Contact](#) [Meta-pipe](#) | [Espen Mikal Robertsen](#) ▾

[Run Meta-pipe](#) [Running jobs](#) [See results](#)

Result files

-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/genes.prot.fasta](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/genes.prot.fasta)
-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/output.j](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/output.j)
-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/entriesMerged.j](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/entriesMerged.j)
-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/export.gb](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/export.gb)
-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/annotations.j](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/annotations.j)
-  [https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so\\_T6CXQVEMH2/genes.nuc.fasta](https://storage.metapipe.uit.no/a8af066bfe/outputs/test-executor/so_T6CXQVEMH2/genes.nuc.fasta)

© UiT - The Arctic University of Norway · [Home](#) · [Privacy](#) · [Not Found](#) | Viewport: 1441x692



# Output

## Visualization of contigs in detail

The screenshot displays the Artemis genome browser interface. At the top, the window title is "Artemis Entry Edit: export (1).gb". Below the menu bar, the entry name is "export (1).gb" and it is noted that "Nothing selected". The main area shows a genomic map with contigs labeled "k99\_1114 gene 4" through "k99\_1114 gene 15". A detailed protein sequence is shown below the map, with amino acid residues highlighted in blue. The sequence starts with "T # D \* P T G V P R \* K S + F W P H C L V L S P L A P L P O K K S A L A F L Q L T H R L N R W T P I T K S P V S T S T W P T R C A N R C R P S" and ends with "C L I S R R S N W S S F T L E P R V A # N K R R Q R R Q R L F F \* S Q R K Q L K G V T Q I S P R W Y G F R R N R C R G P W R A T C V S A P R T".

Artemis Release 16.0.0  
1. Standard

source	1	108378
fasta_record	1	41673
CDS	30	761
CDS	891	2363
CDS	2360	3085
CDS	3155	4285
CDS	4328	4816
CDS	4874	5674
CDS	5716	6869
CDS	6890	7858
CDS	6890	7858
CDS	7816	9028
CDS	7816	9028
CDS	9377	9856

# Output

## Visualization prototype

Overview **Sequence Distribution** Taxonomic Classification Functional Assignment Geographic Context To MarCat Downloads

Histogram **Kernel density estimation**

View histogram for:

Reads **Contigs** Genes, complete Genes, truncated

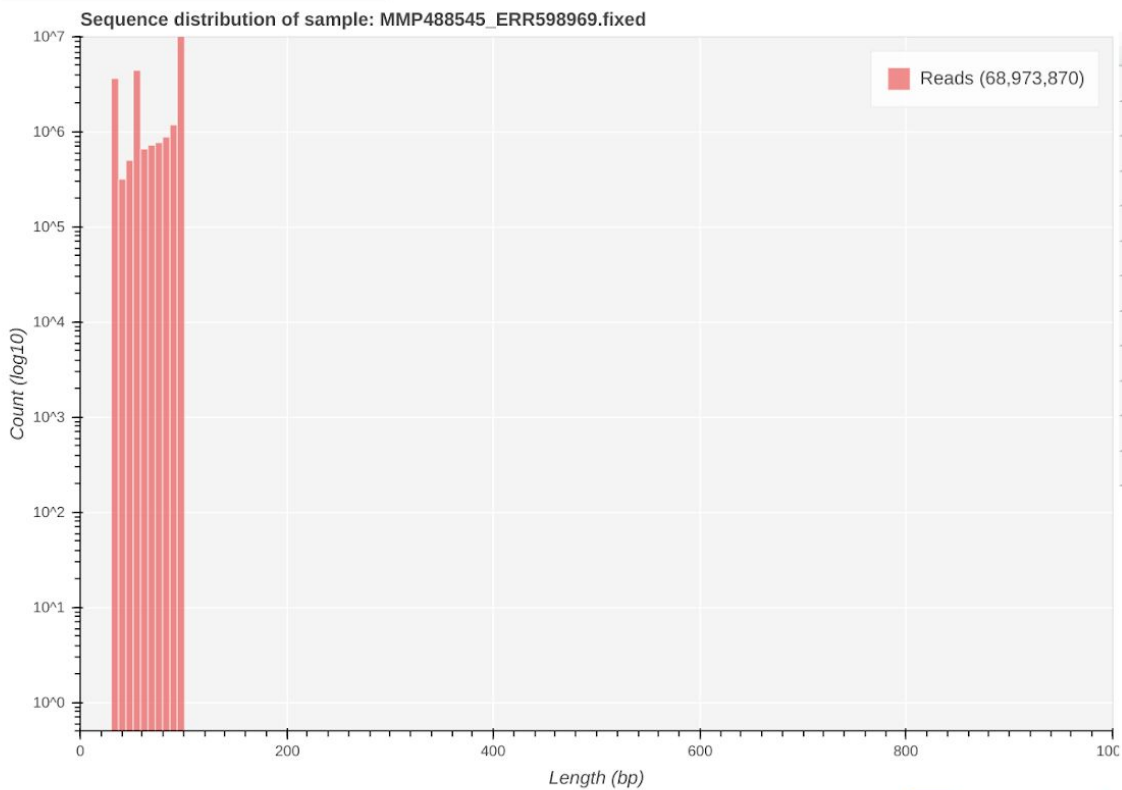
Y axis type

log

X scale: 0

View KD-estimation for:

Reads **Contigs** Genes, complete Genes, truncated



#	Metric	Reads	Contigs	Genes
0	Count (#)	68,973,870	162,298	62,591
1	Length (bp)	6,374,838,210	150,772,424	36,358,786
2	Over 100 bp	52,685,311	162,298	62,591
3	Over 500 bp	0	161,652	27,753
4	Over 1000 bp	0	39,001	8,392
5	Over 5000 bp	0	843	30
6	Over 10000 bp	0	177	0
7	Largest (bp)	101	166,928	8,936
8	Smallest (bp)	30	500	119
9	Average length (bp)	92	929	581
10	Median (bp)	101	708	452
11	N50	101	923	764

# Thank you!

## The Center for Bioinformatics team (SfB)

Alexandr Agafonov

Juan Fu

Espen Robertsen

Espen Åberg

Erik Hjerde

Inge Alexander Raknes

Lars Ailo Bongo

Nils Peder Willassen

Terje Klemetsen



UiT / THE ARCTIC UNIVERSITY  
OF NORWAY

Tromsø, May, 2018

