Introduction to metagenomics

Nils Peder Willassen

Centre for bioinformatics

UiT The Arctic University of Norway

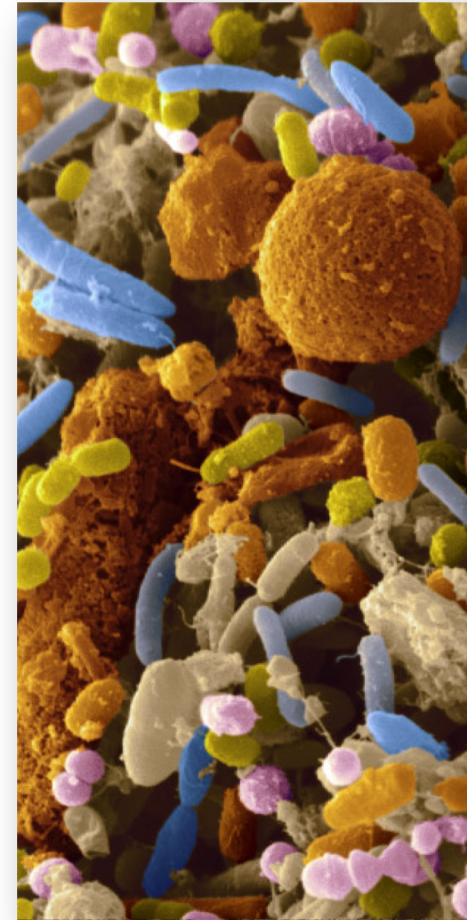*ELIXIR Norway*

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

Marine Metagenomics

Outline

- Background

- Applications

- Steps in metagenomics

- Main challenges

Marine Metagenomics

Definition

Metagenomics is the study of genetic material recovered directly from environmental samples
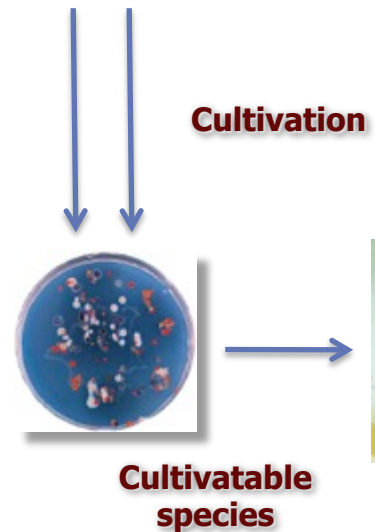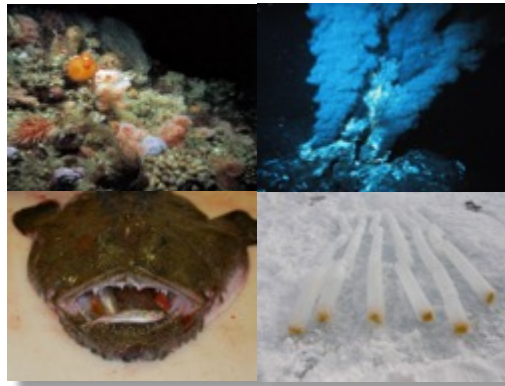
Meta- meaning "after", or "beyond"

Aim often to explore the relation between the microbes and habitats

© IMGM Laboratories

Marine Metagenomics Workshop, Nov 26-30, 2018, Tromsø, Norway

Marine Metagenomics

Omics

Direct isolation of DNA from the environments

Metagenomics

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

Meta´omics

Isolation of RNA, protein & metabolites

Cultivation

Transcriptomics, proteomics, metabolomics: response of organism/community to varying environmental perturbations

Metatranscriptomics
Metaproteomics
Metabolomics

Multi´omics
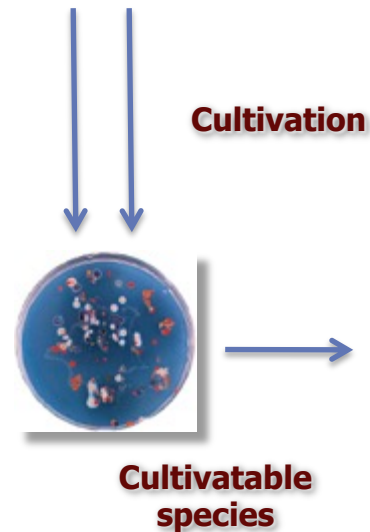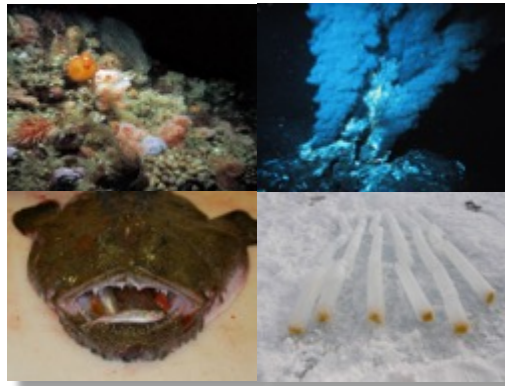
Isolation of DNA

Traditional cultivation methods and traditional genomics can at best access 1% (human >35%)

Cultivatable species

Genomics

Marine Metagenomics

Omics

Direct isolation of DNA from the environments

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

Metagenomics

Cultivation

Isolation of RNA, protein & metabolites

Transcriptomics, proteomics, metabolomics: response of organism/community to varying environmental perturbations

Meta´omics

Metagenomics assembled genome (MAG)

Single amplified genome (SAG)

Multi´omics

Isolation of DNA

Traditional cultivation methods and traditional genomics can at best access 1% (human >35%)

Cultivatable species

Genomics

Marine Metagenomics

Applications

Host-microbe interaction - Improve yield

Monitoring and optimizing wastewater treatment

Bioprospecting – hunting for novel enzymes and compounds

Function and ecology of the oceans

Development of functional feed

Disease surveillance and treatment

Monitoring pollution

Conversion of biomass/rest raw materials

excelerate
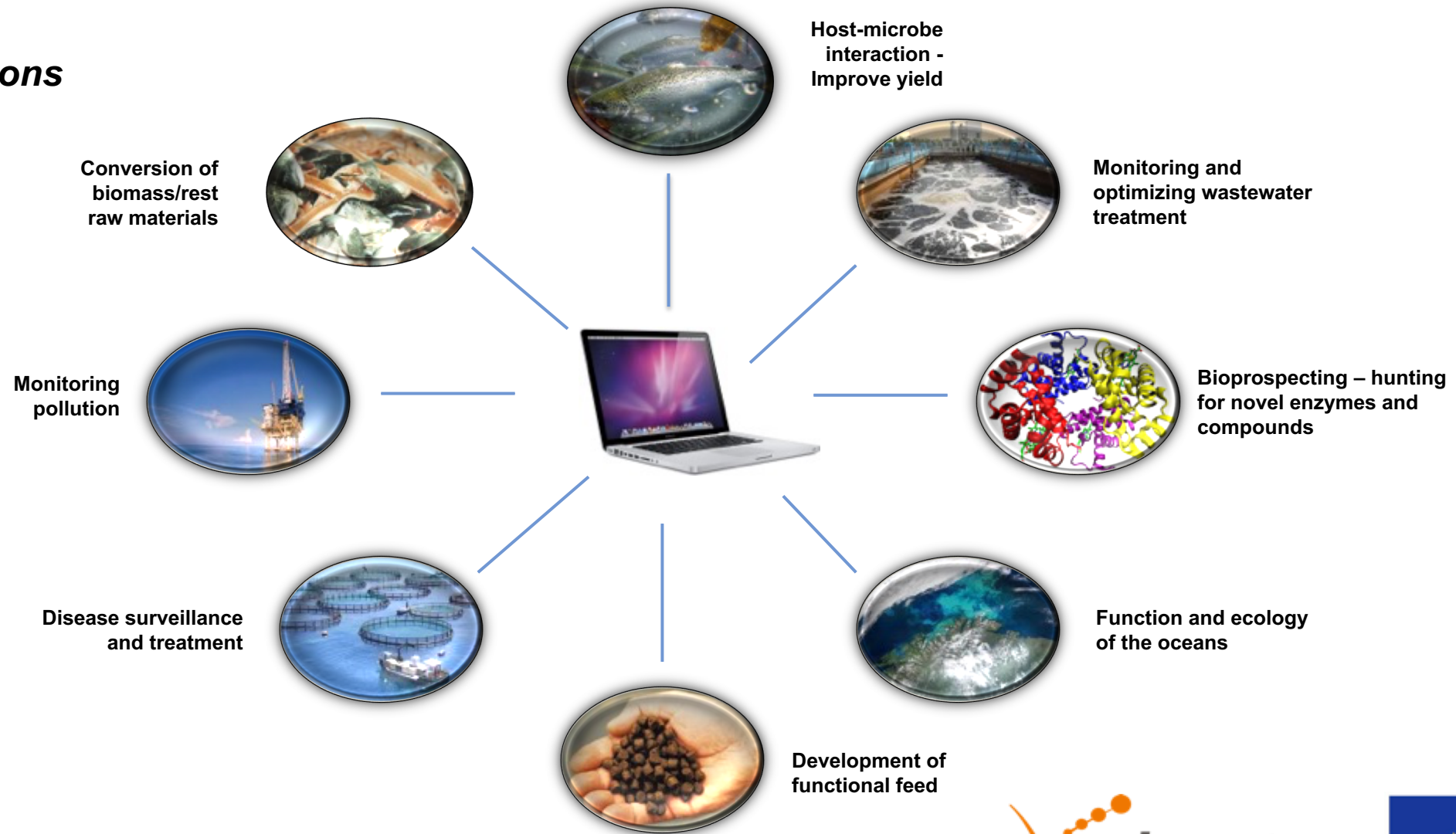
Marine Metagenomics

Big picture

Plan → Sampling → DNA isolation → Data generation (sequencing) → Data processing and annotation → Analysis → Archiving
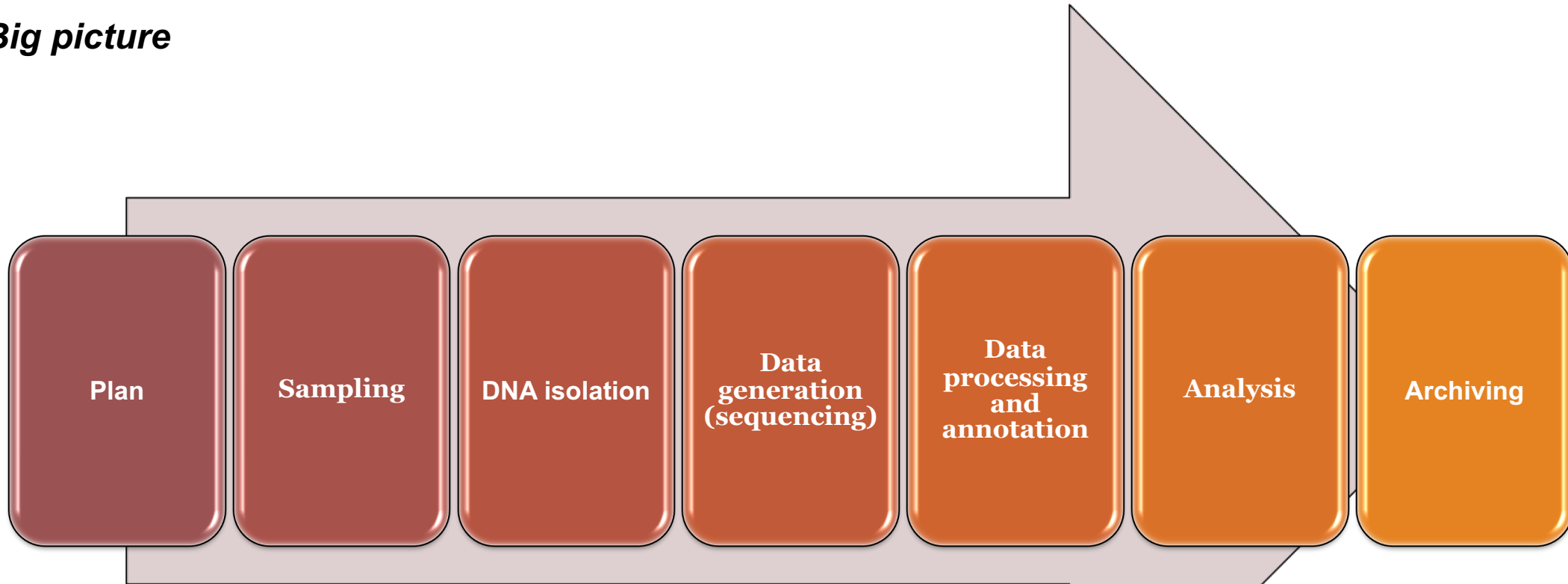
Marine Metagenomics

Steps in metagenomics - plan

Project
plan

Aim (golds) of
the project

"The plan"

Experimental
design (DoE)

Data Management
Plan (DMP)

Marine Metagenomics

Steps in metagenomics - sampling

What, were, when and how to sample

Required equipment

Enrichment e.g. filtering

Biological replicates

"The plan"

Sampling

How to transport and store the samples

Knowledge to diversity

Marine Metagenomics

Steps in metagenomics - sampling

Marine environment

Bacteria (archaea/eubacteria)
Species
$->10^9$ (billion)?
Genome size
0,6 - 12 Mb (5 Mb)
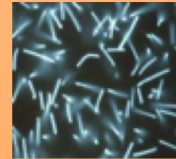
Virus (bacteriophages)
species:
$-> 10^6$ (million)?
genome size
3 200 bp - 1.2 Mb (50 Kbp)

Protist (eukaryote microbes)
Species
Estimate >> 300 000?
Genome size
>>35 Mb - 215 Gb (10 000 Mb)
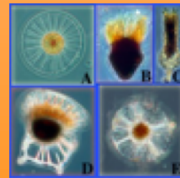
Fungi (yeast/mold)
Species
estimate >150 000?
Genome size
> 10Mb - 1,2 Gb (100 Mb)
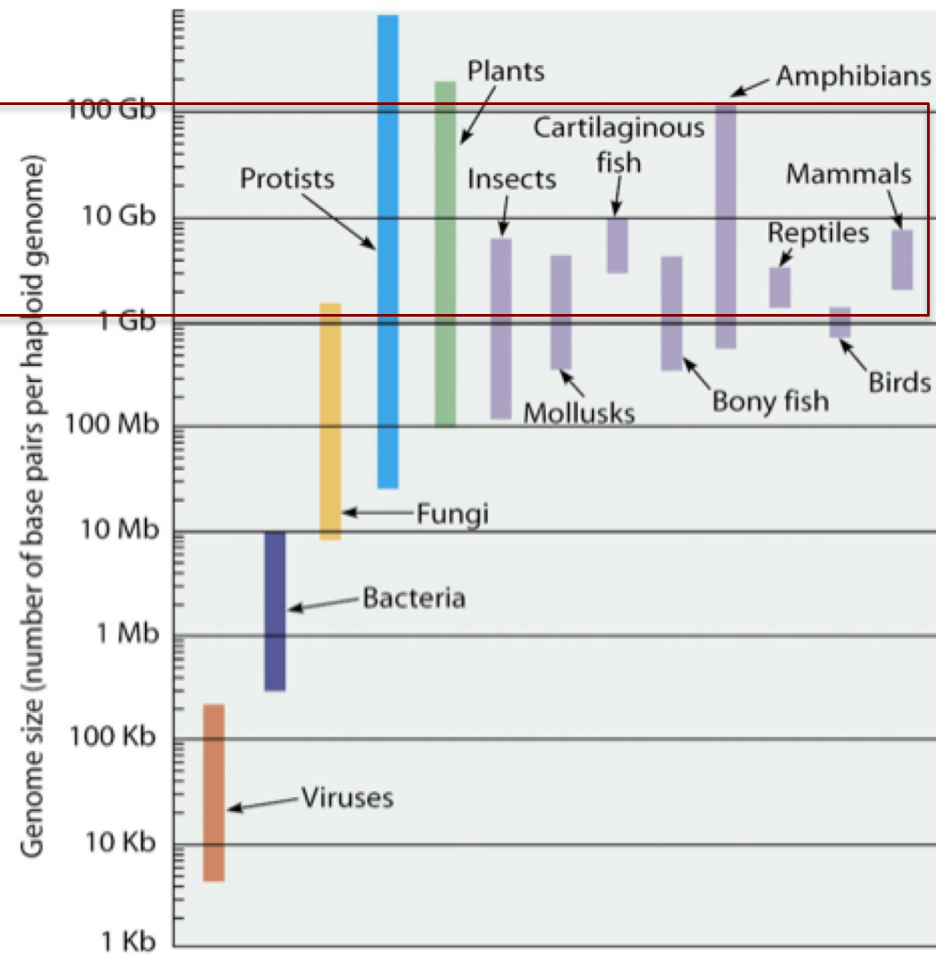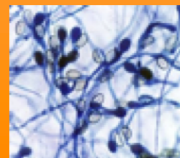
Marine Metagenomics

Steps in metagenomics – DNA isolation

How to lyse cells and isolate DNA

Removal of major contaminants

Lysis of microorganisms
- **Gram negative/positive**
- **Fungi**
- **Virus**

"The plan"

Sampling

DNA isolation

DNA isolation methods
- **Negative extraction control (buffer)**
- **Positive extraction control (mock sample)**
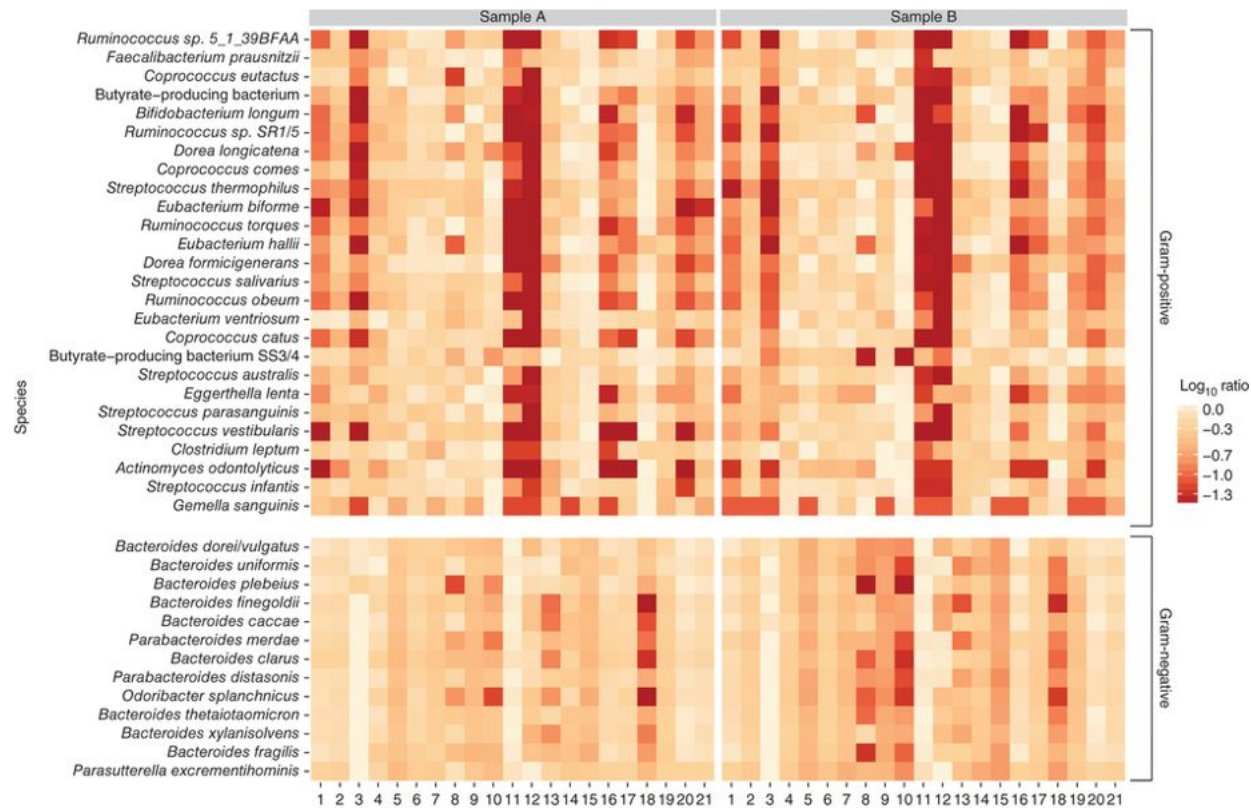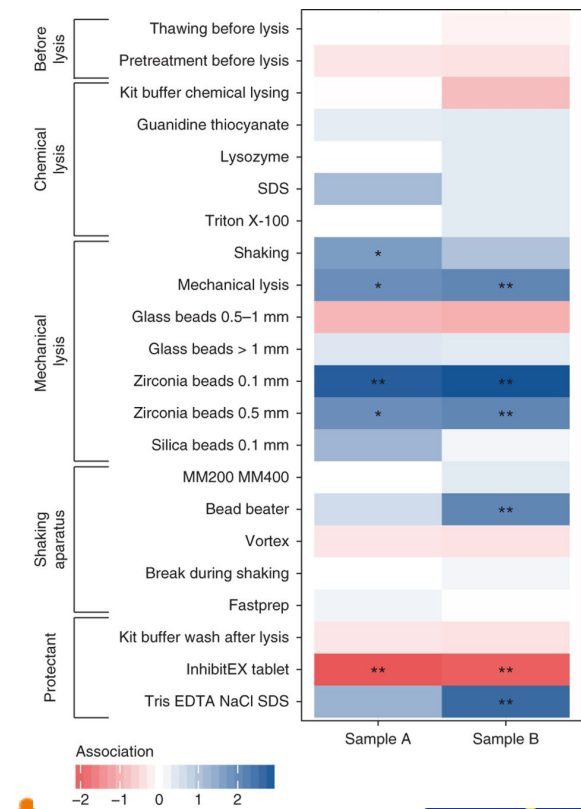
Quantity and quality of DNA

excelerate

Marine Metagenomics

Steps in metagenomics – DNA isolation

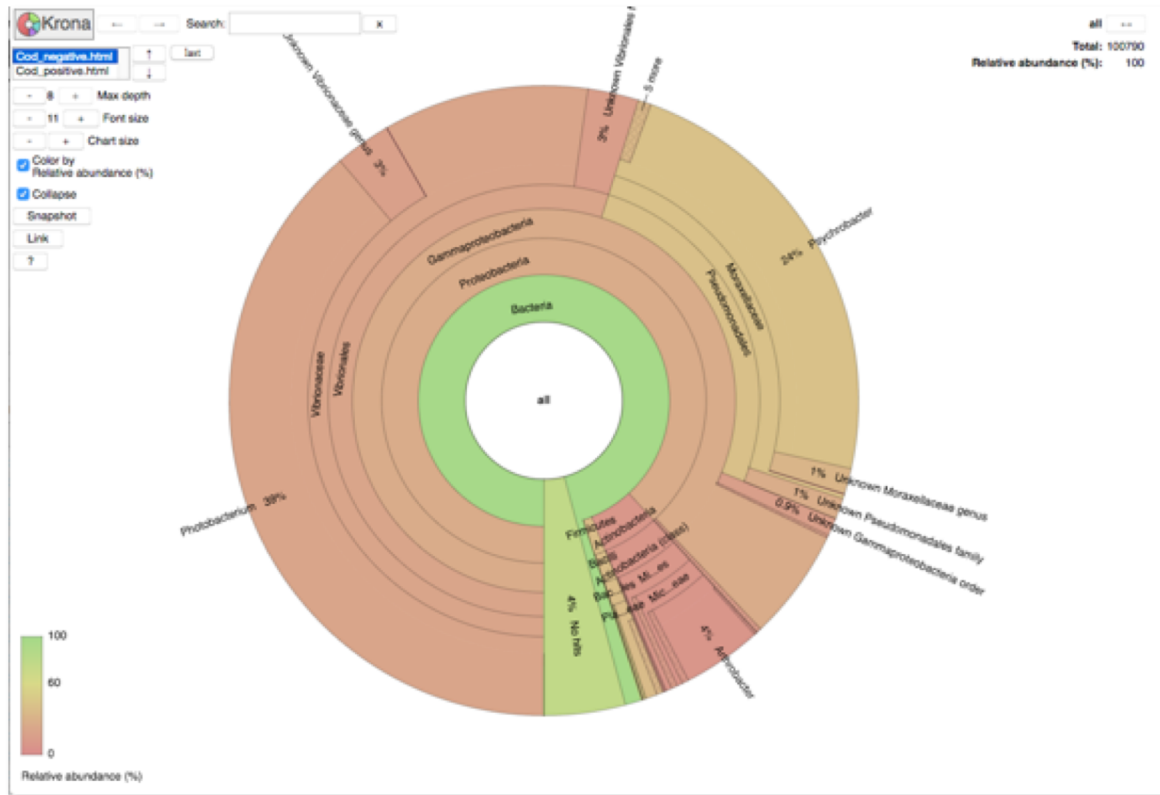**Species-specific abundances variation across 21 isolation methods**

Effects of protocol manipulations on sample composition.

Marine Metagenomics

Steps in metagenomics – DNA isolation

Marine Metagenomics

Steps in metagenomics – data genertaion

"The plan"

Sampling

DNA isolation
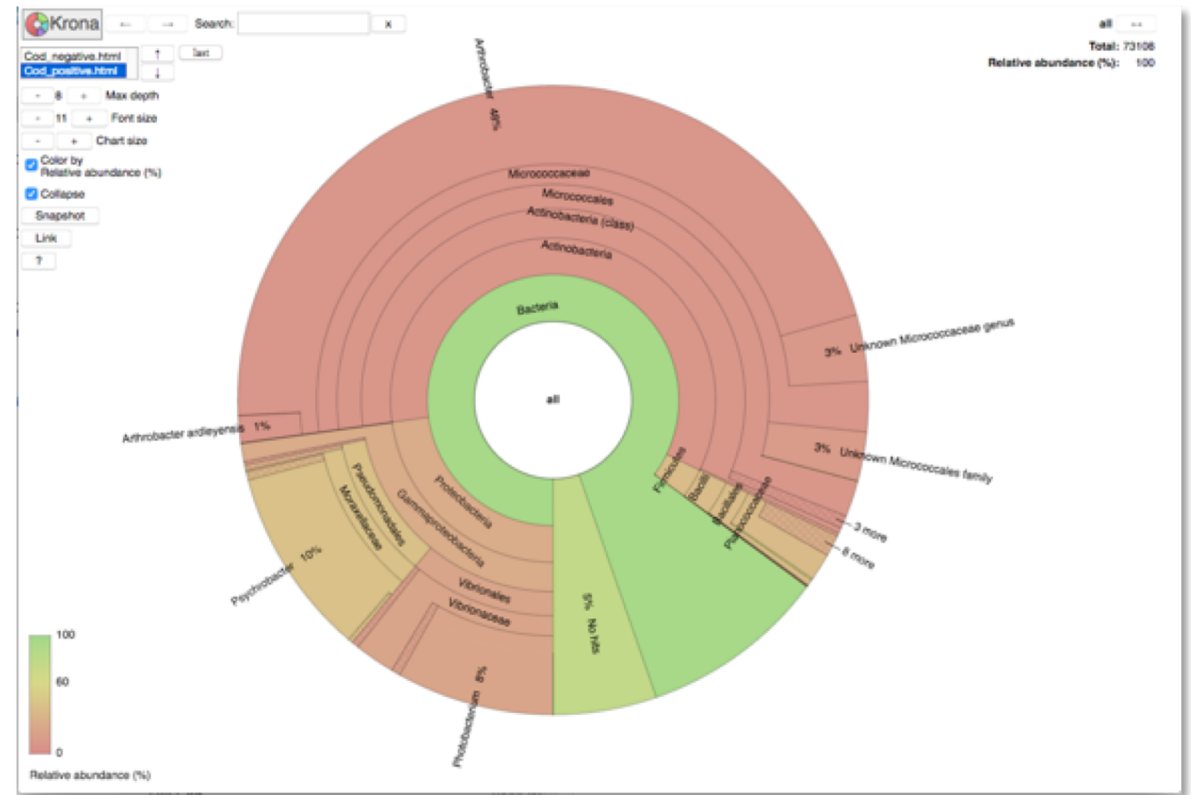
Sequencing

Which sequencing technology to use

Illumina, Ion Torrent
PacBio, Oxford Nanopores

Sequencing mode:
Single reads/ Paired end/ Read length

Technical replicates

Sequencing depth

Price
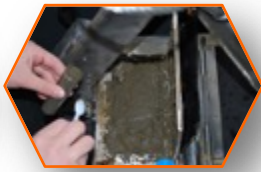
Marine Metagenomics

Steps in metagenomics – data generation

Sequence DNA from 1 gram marine sediment (5 Tb) with MiSeq (1 x coverage)

MiSeq

- *output 15 Gb/run (2 x 300 bp approx. 67 hours)*

- *# of runs: 5 Tb/15 Gb => 333 runs*

- *# of hours: 333 x 67 = 22333 hours = 2.5 years*

- *Cost: 333 x 3.500 Euro/run => 1 165 000 Euro*

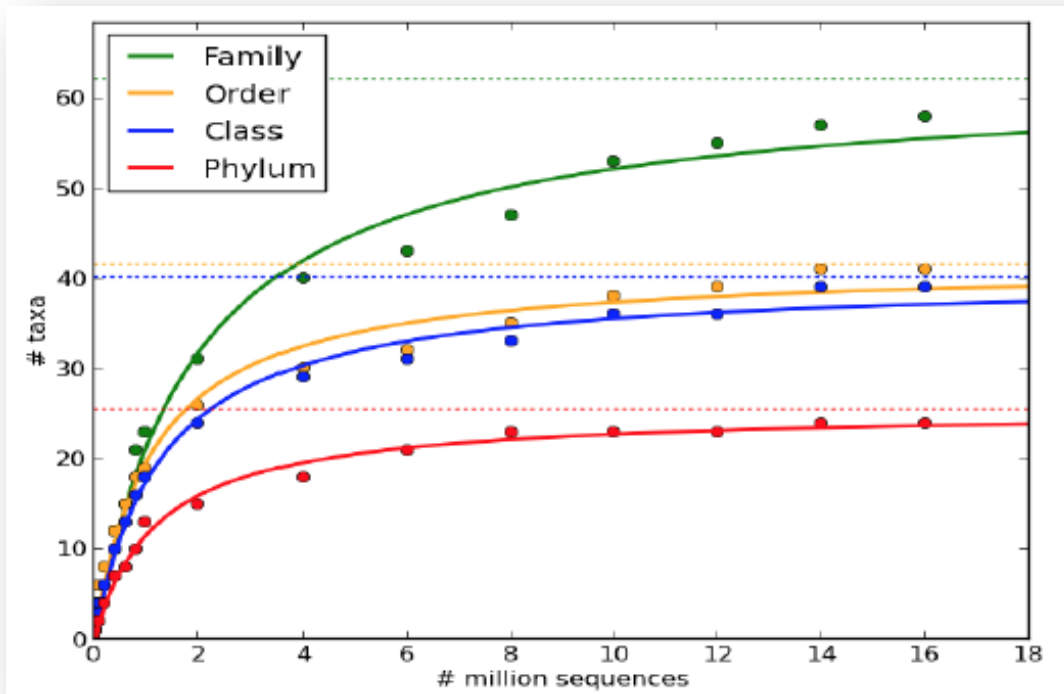- *100 x coverage => 250 years and 116.5 million Euro*

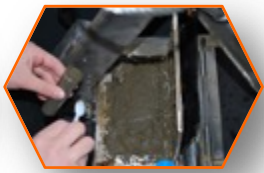# Marine Metagenomics

Steps in metagenomics – sequencing depth

Rarefaction analysis

	Phyla	Class	Order	Family	Genus
Observed taxa	23	39	41	55	102
Estimated taxa	25	40	41	62	145

Marine Metagenomics

Steps in metagenomics - data processing

"The plan" **Sampling** **DNA isolation** **Sequencing** **Processing & annotation**

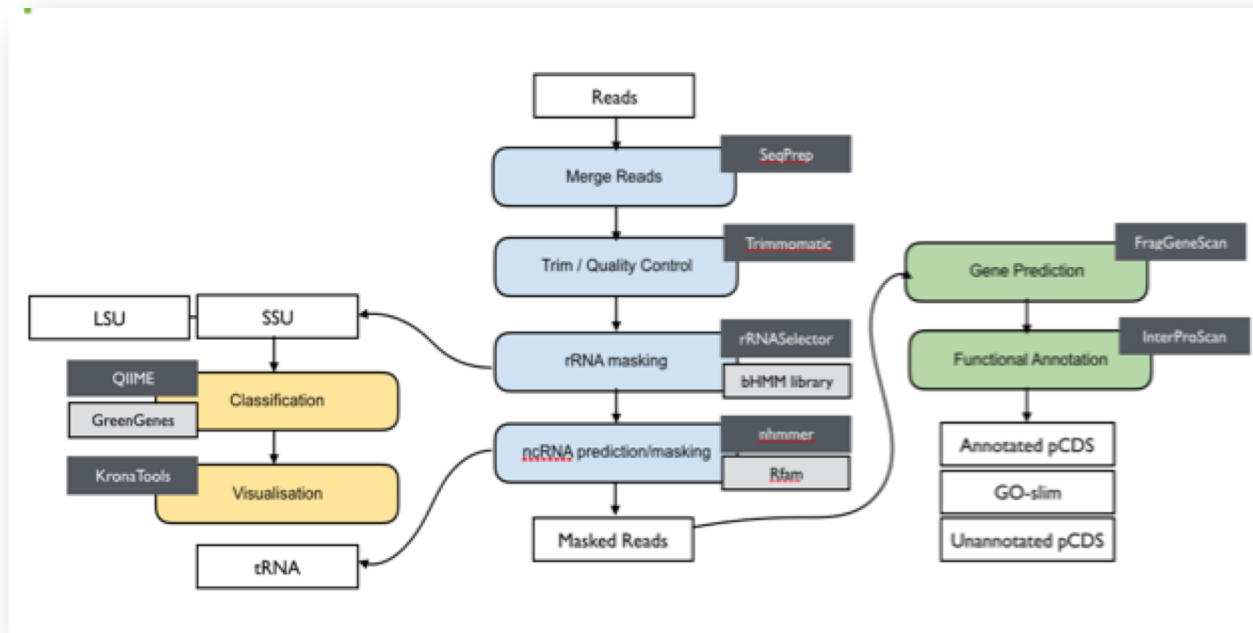*Transfer and storage of raw sequence data*

Which tools or pipelines to use

Reference databases

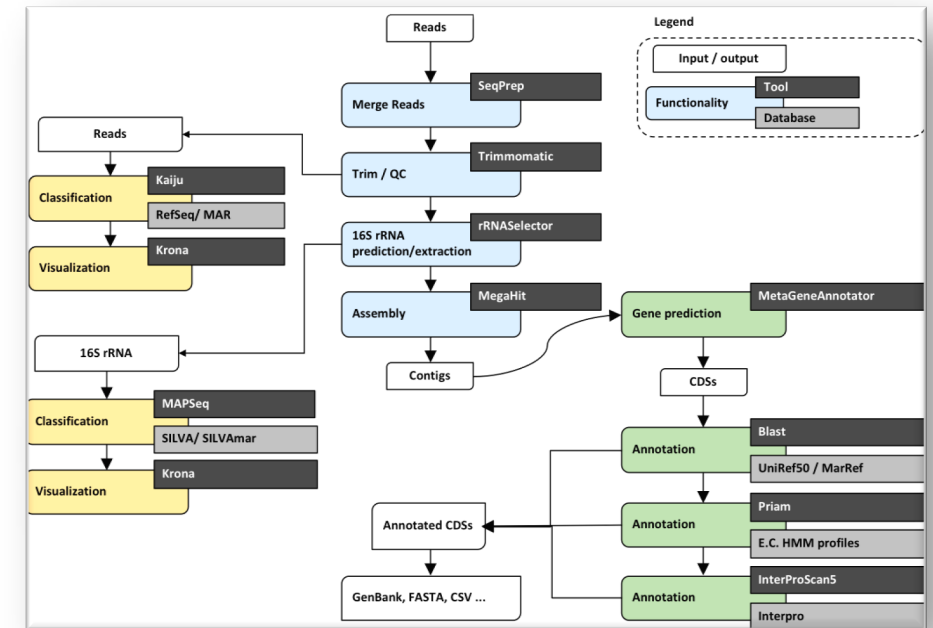*High-performance computing and storage of processed data*

Availability and scalability

Marine Metagenomics

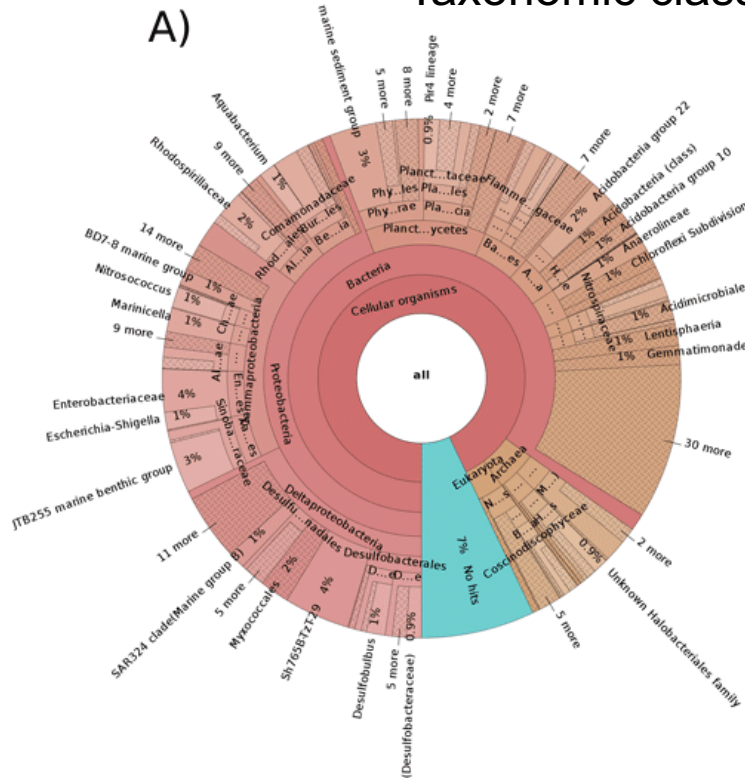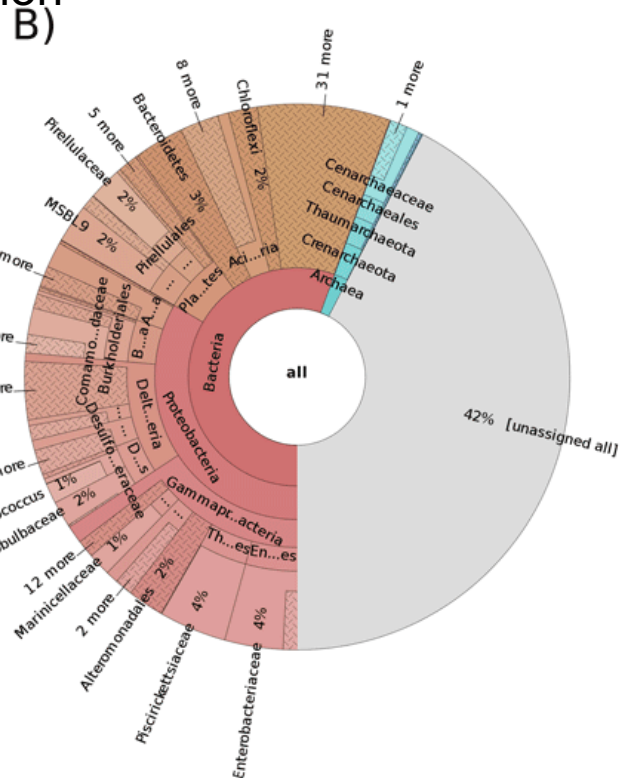**Steps in metagenomics**

EBI metagenomics portal

META-pipe

Marine Metagenomics

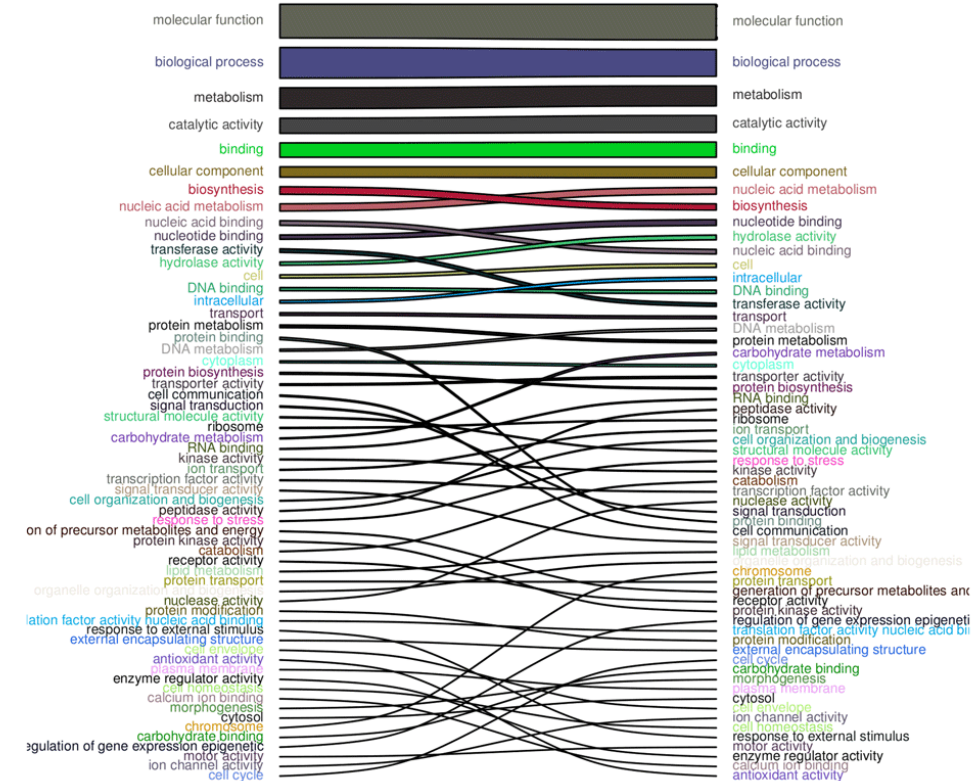## Steps in metagenomics

Taxonomic classification

META-pipe

EBI metagenomics portal
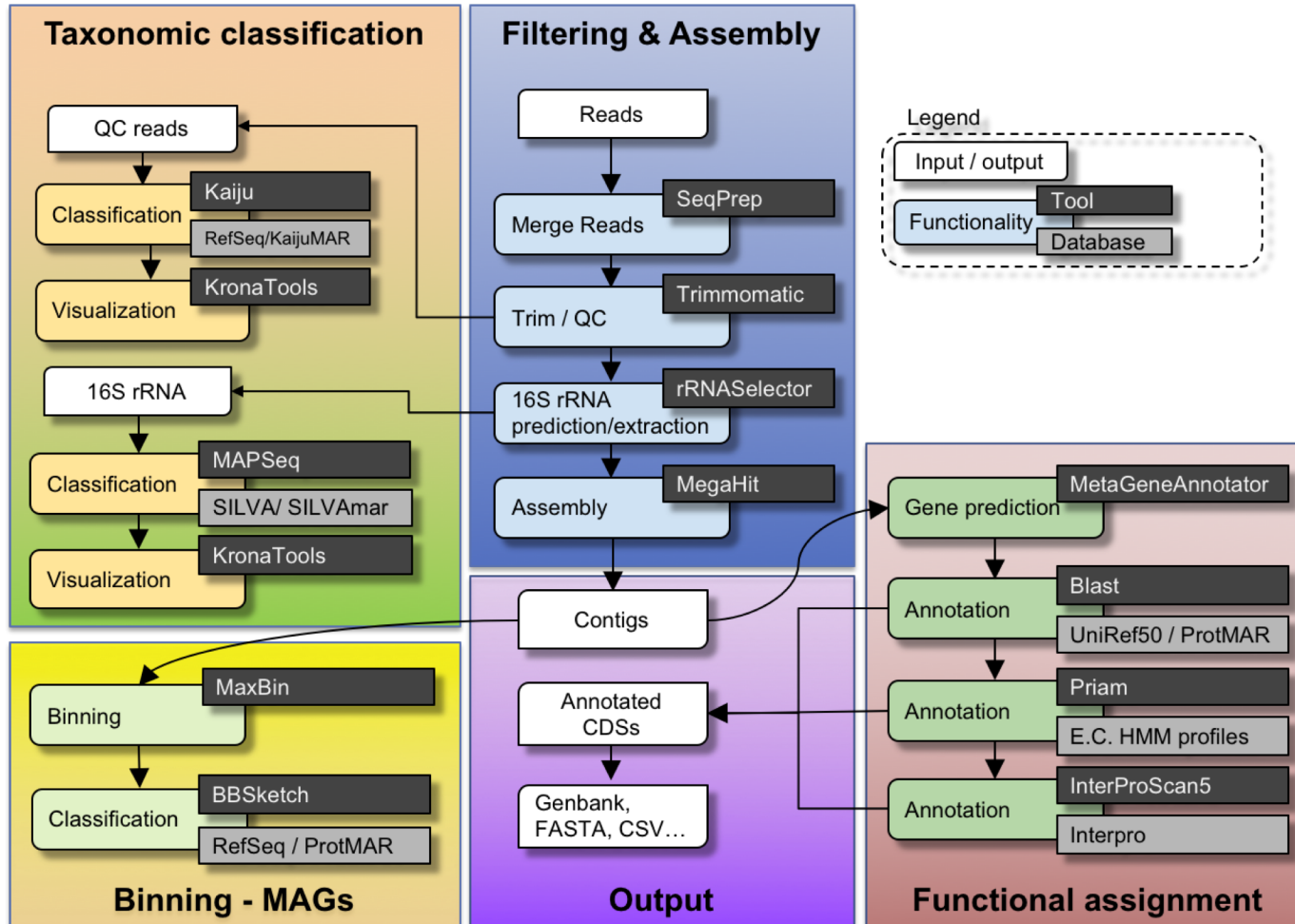
Functional assignment (GO-slim)

META-pipe v3

META-pipe

13 different tools
20 different databases

Average Tara Ocean sample:
    2 x 14 Gb.gzip
    44 Gb sample

HPC:
$\Rightarrow$ 256 Gb RAM, 400 cores

Analyze time:
$\Rightarrow$ approx. 36 hours

Marine Metagenomics

Steps in metagenomics
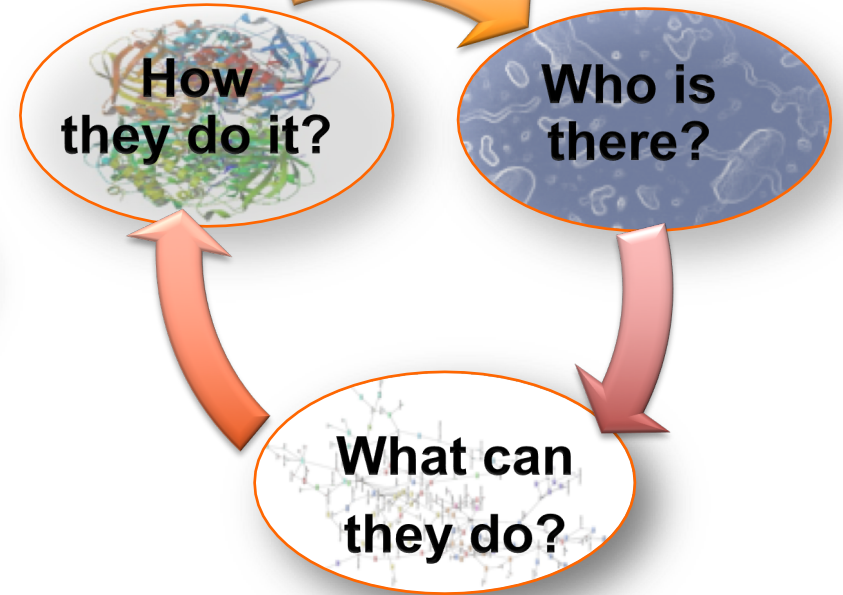

"The plan"   Sampling   DNA isolation   Sequencing   Processing & annotation

Taxonomic and/or functional analysis

Tools and data resources available

How they do it?

Who is there?

What can they do?
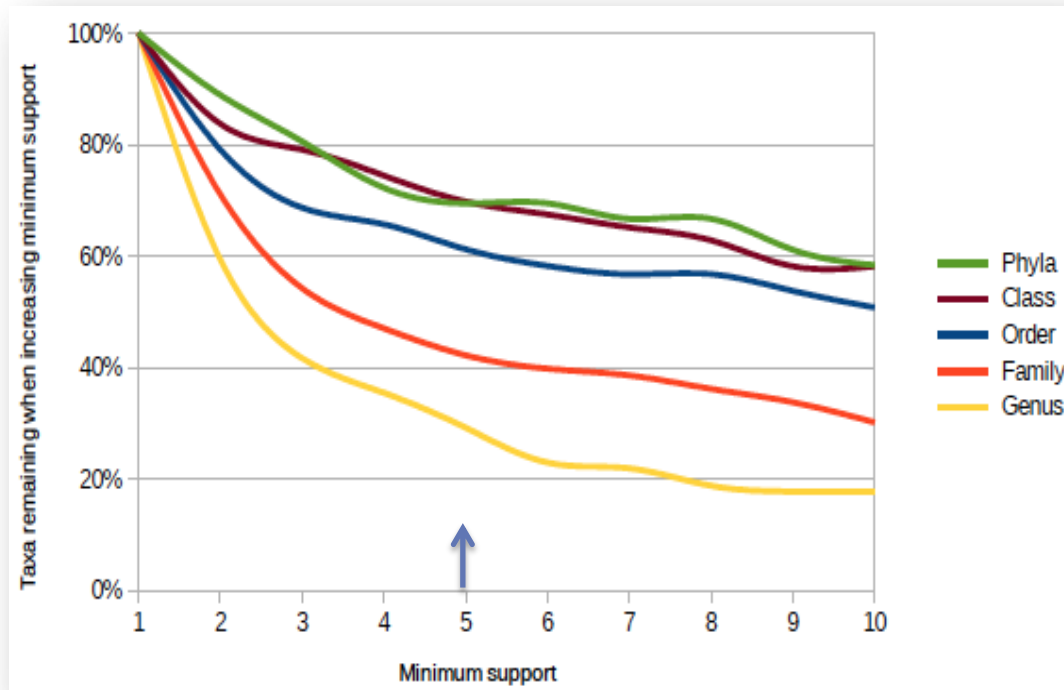
Analyse results

Depend heavily on tools and databases

Marine Metagenomics

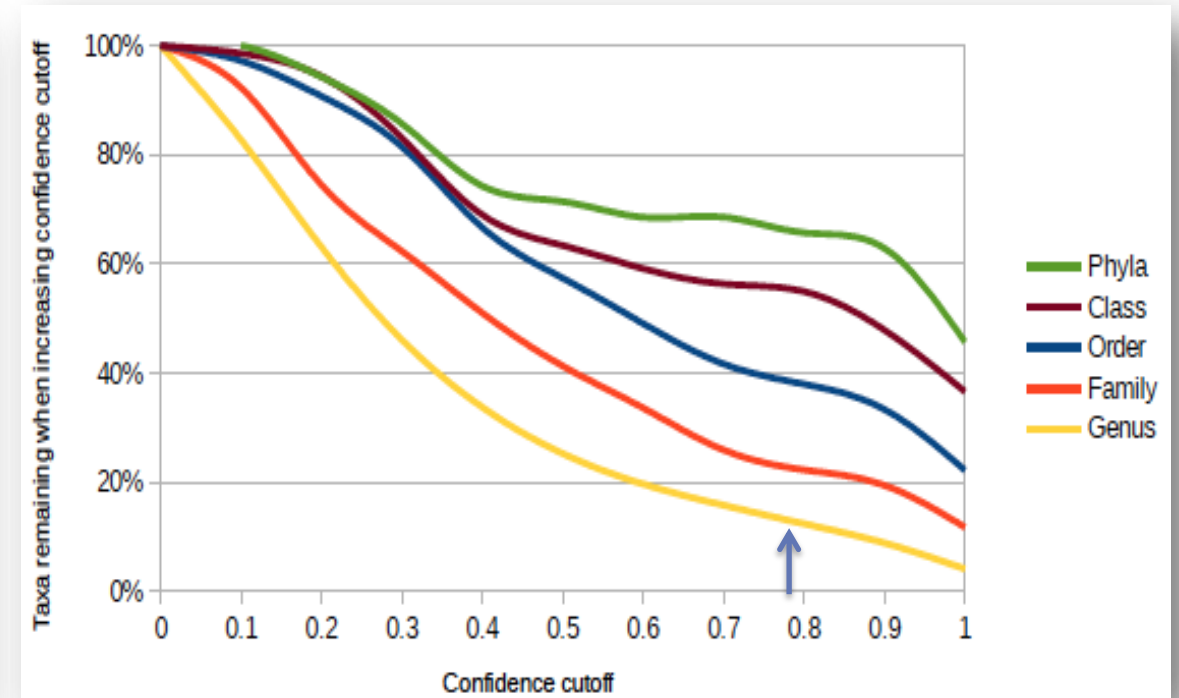**Steps in metagenomics**

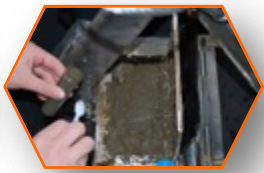MEGAN4 - NCBI taxonomy

QIIME - RDP Classifier
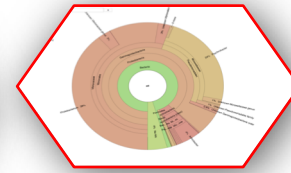
Marine Metagenomics

Steps in metagenomics

Publication of results

Deposit sequences into public archives

"The plan" **Sampling** **DNA isolation** **Sequencing** **Processing & annotation** **Analyse results** **Archiving**

FAIR – Findable, Accessible, Interoperable & reusable
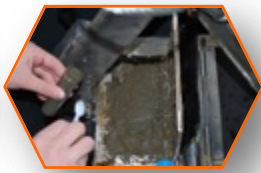
MIMS (Minimum Information about a Metagenome Sequence)

excelerate

Marine Metagenomics

Steps in metagenomics

| "The plan" | Sampling | DNA isolation | Sequencing | Processing & annotation | Analyse results | Archiving |

Project plan — *Experimental design* / *Data management plan* — *Standard Operational Procedures (SOPs)* — *Provenance data (metadata)* — *FAIR – Findable, Accessible, Interoperable & reusable* — *"Gold" standard/ best practise of annotation and analysis*
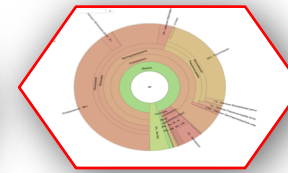
Marine Metagenomics

Main challenges

- Number and size of marine projects increases

- Scalable tools and pipelines => need HPC

- Short sequence reads are hard to annotate

- Few hits in reference databases => need domain specific databases

- Results depend heavily on tools, underlying algorithms and reference databases

- Description of data (metadata) often lacking

Thanks

www.elixir-europe.org/excelerate