LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

CNR-IBIOM: Infrastructure, data and analysis resources

Fosso B. , Balech B., Tangaro M.A. and Pesole G.

LifeWatch ERIC thematic meeting on Genomics
26-28 February 2020 - Porto

Outlook

The research infrastructure

Bio-molecular data resources

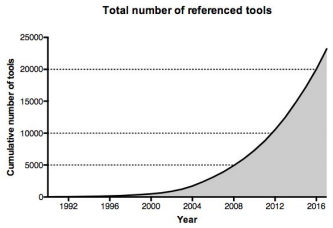
- ITSoneDB
- COXI-DB

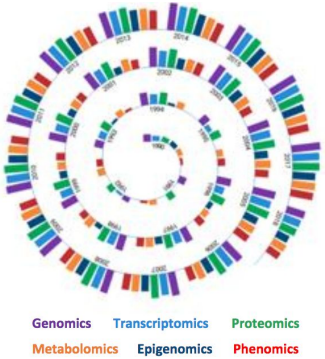
Analysis resources

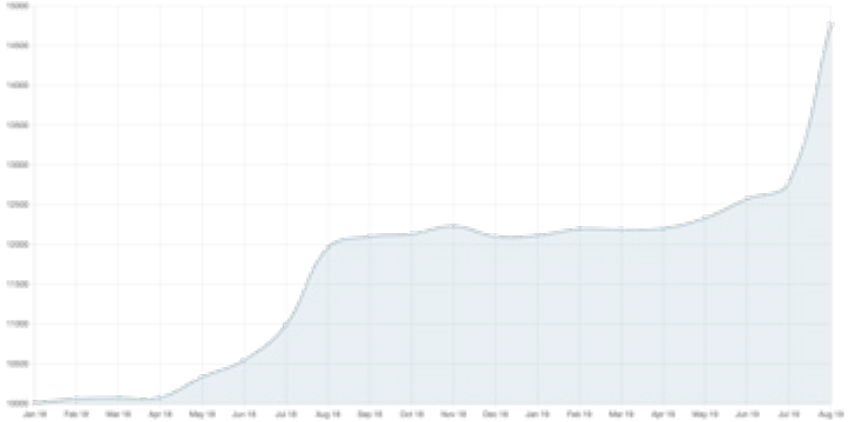
- MetaShot
- BioMaS

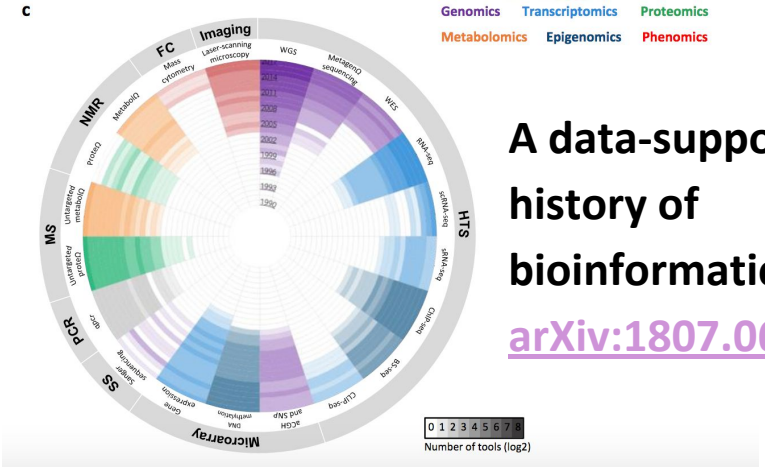
The research infrastructure

Bioinformatics software/services









A data-supported history of bioinformatics tools

arXiv:1807.06808

Over 14 thousands entries in Bio.tools

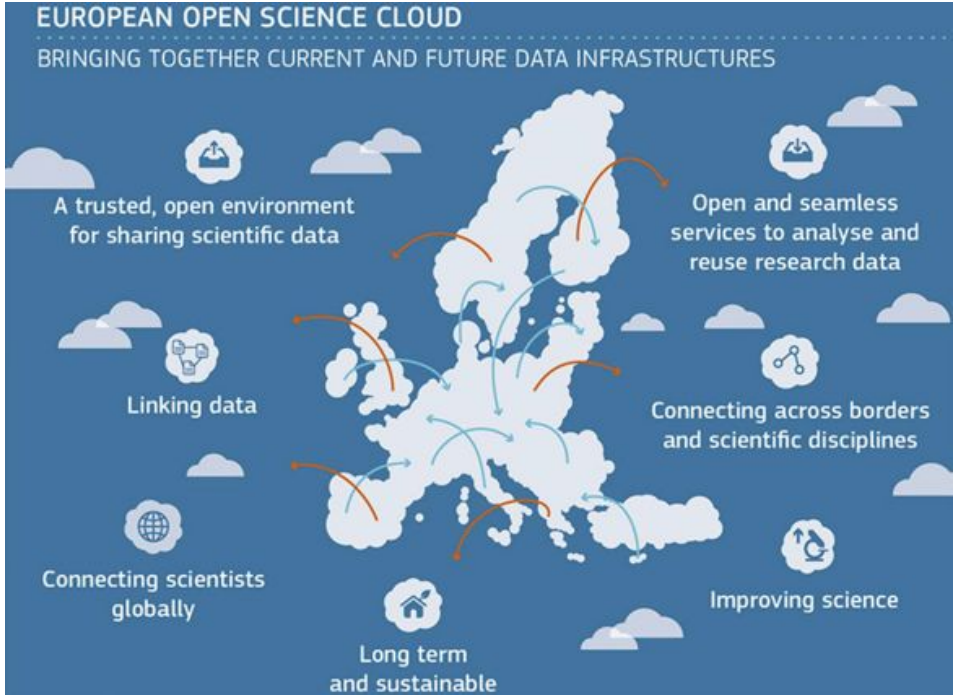
LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

Building an infrastructure for life scientists

Services for science are moving to the Cloud.

The **"European Open Science Cloud"** aims to create a trusted environment for hosting and processing research data to support EU science in its global leading role.

Develop Cloud services which can be exploited by life science community.



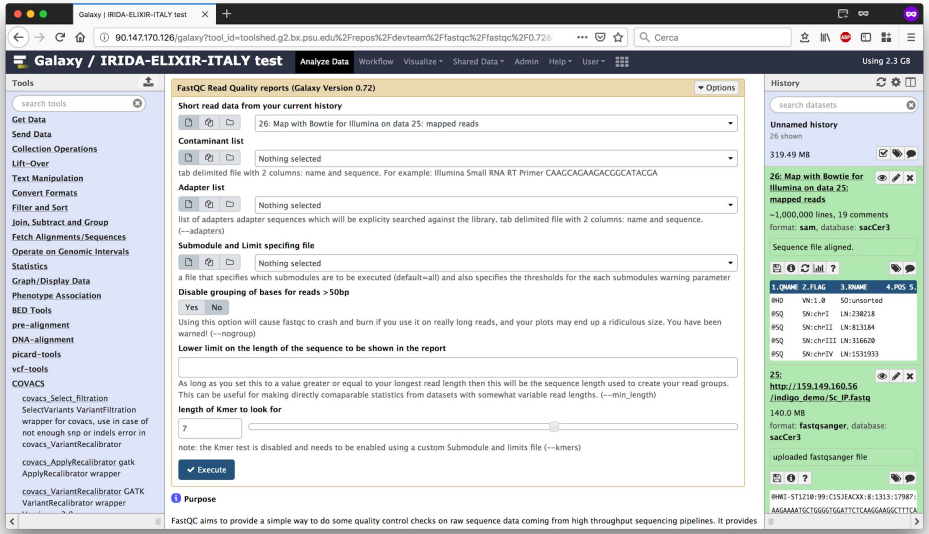
LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

Galaxy

Galaxy

Galaxy (galaxyproject.org) is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data.

Through a coherent work environment and an **user-friendly web interface** it organizes data, tools and workflows providing reproducibility, transparency and data sharing functionalities to users.

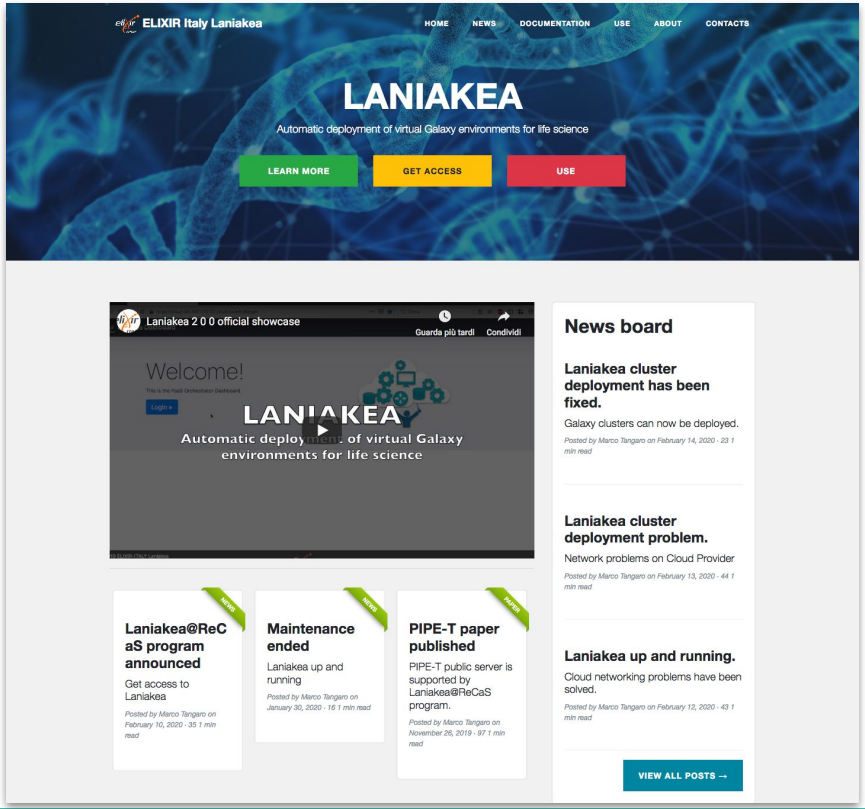


- **Need IT infrastructure (proportional to workload).**
- **Need IT expertise**
- **Need Galaxy admin expertise**

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

Laniakea: Galaxy "on-demand" platform

Laniakea



https://laniakea-elixir-it.github.io

LANIAKEA is a cloud Galaxy instance provider, based on INDIGO-DataCloud software catalogue. Its architecture automates the creation of Galaxy-based virtualized environments exploiting the software catalogue provided by the INDIGO-DataCloud project.

- **No need of local IT infrastructure / expertise**
- **Data privacy and security (encrypted data volumes)**
- **Full control over Galaxy instance(s)**
- **Fully customisable**

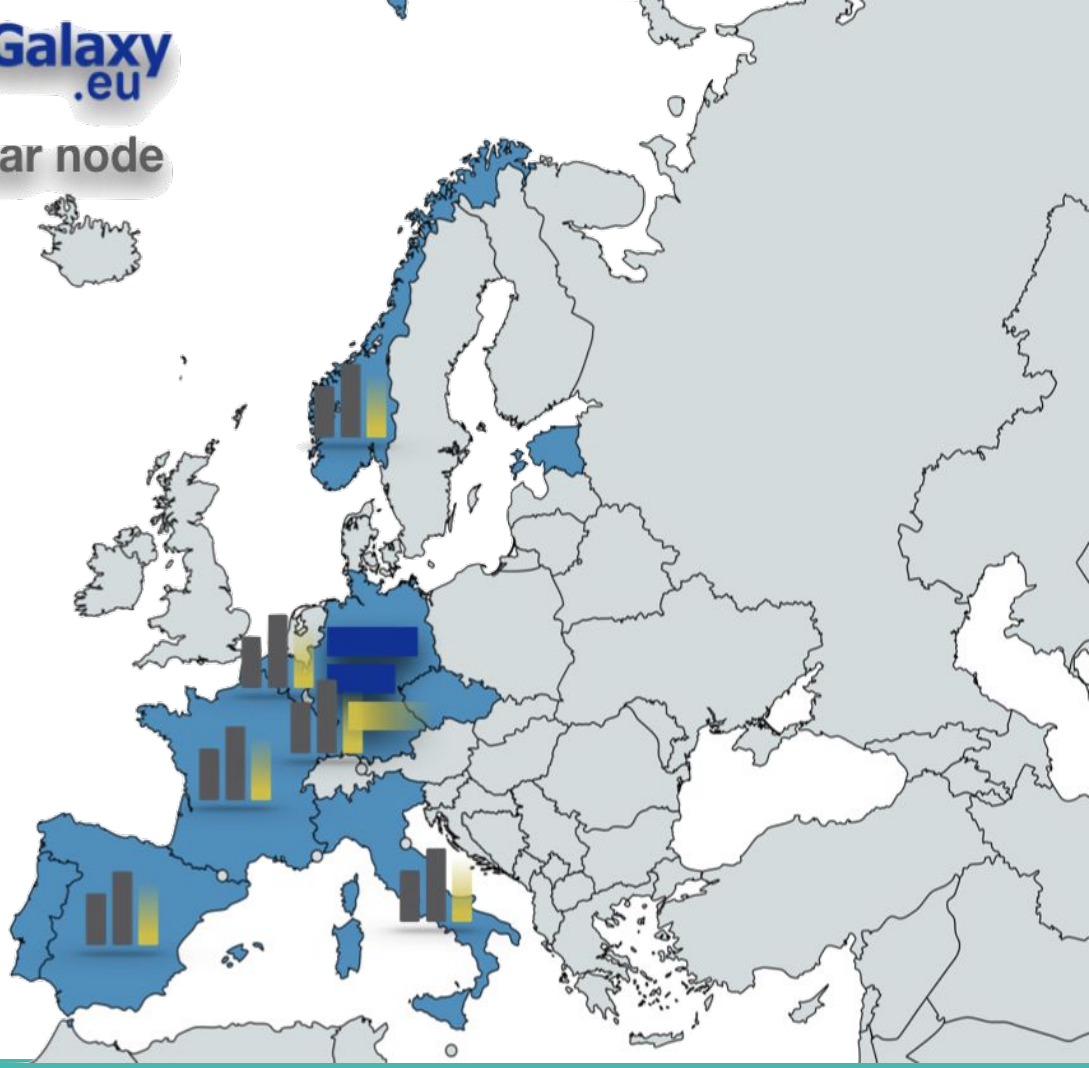
(*) The Laniakea Supercluster (Laniakea; also called Local Supercluster or Local SCl or sometimes Lenakaeia) is the galaxy supercluster that is home to the Milky Way and approximately 100,000 other nearby galaxies [Wikipedia].

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

useGalaxy.eu
Pulsar node

The Pulsar Network

The most innovative computing centers across Europe are currently interested to share their remote computation power to support the European Galaxy server UseGalaxy.eu load.



The funding model: ELIXIR and H2020 projects

ELIXIR European Research Infrastructure for biological data which primary objective is to support research in the field of "life sciences" and their translational activities to medicine, environment , biotechnological industries and society.

[2015-2017]]Develop an open source computing and data platform, targeted at multi-disciplinary scientific communities, provisioned over public and private e-infrastructures.

[2019-2022] EOSC-Life brings together biological and medical RIs to create an open collaborative space for digital biology. It aims to publish FAIR life science data resources for cloud use creating an ecosystem of innovative tools in EOSC and enabling groundbreaking data-driven research in Europe by connecting life scientists to EOSC.

[2019-2022] EOSC-Pillar will support the implementation of the European Open Science Cloud by leveraging national initiatives of the EU Member States and thematic initiatives developed by research communities working in national and European collaborations.

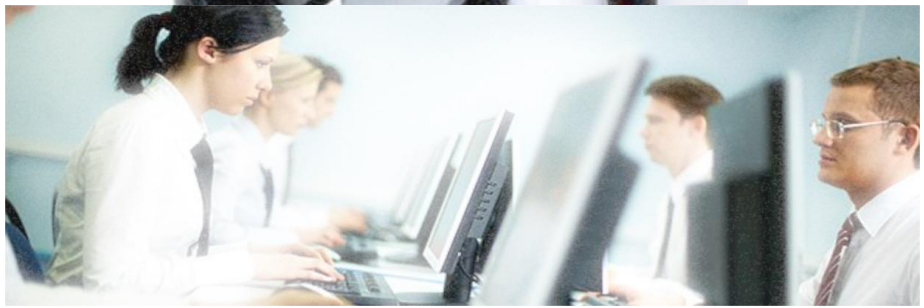
Empowering the infrastructure: PON R&I 2014-2020 Avviso 424/2018 Azione II. 1

Bio-molecular data resources

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

MoBiLab

Laboratorio di Biodiversità Molecolare



Molecular Biodiversity Laboratory (**MoBiLab)** has a fully operative platforms based on:

- NGS technologies
- Data storage resources
- Computational analysis

In addition to their support to LifeWatch, the services hosted by MoBiLab contribute to the Italian node of the European infrastructure **ELIXIR**.

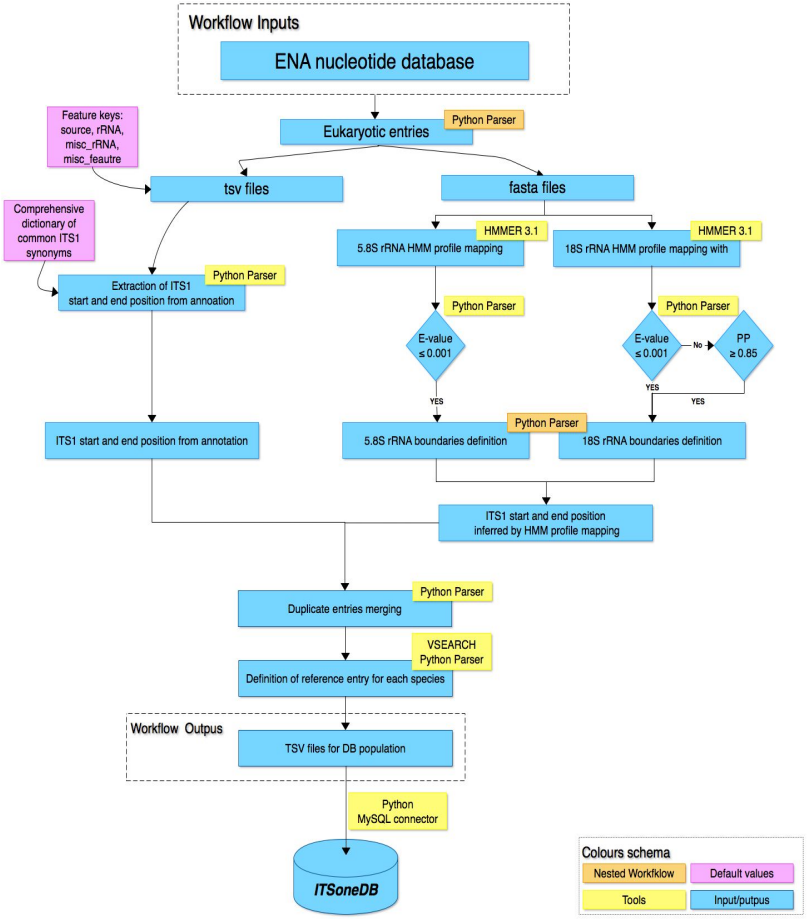
Bio-Molecular Databases

Curated specialized reference databases

- ✔ ***ITSoneDB:*** **RNA Internal Transcribed Spacer 1 (*ITS1*)** database
- ✔ ***COXI-DB: Cytochrome Oxidase subunit-I*** database

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

ITSoneDB



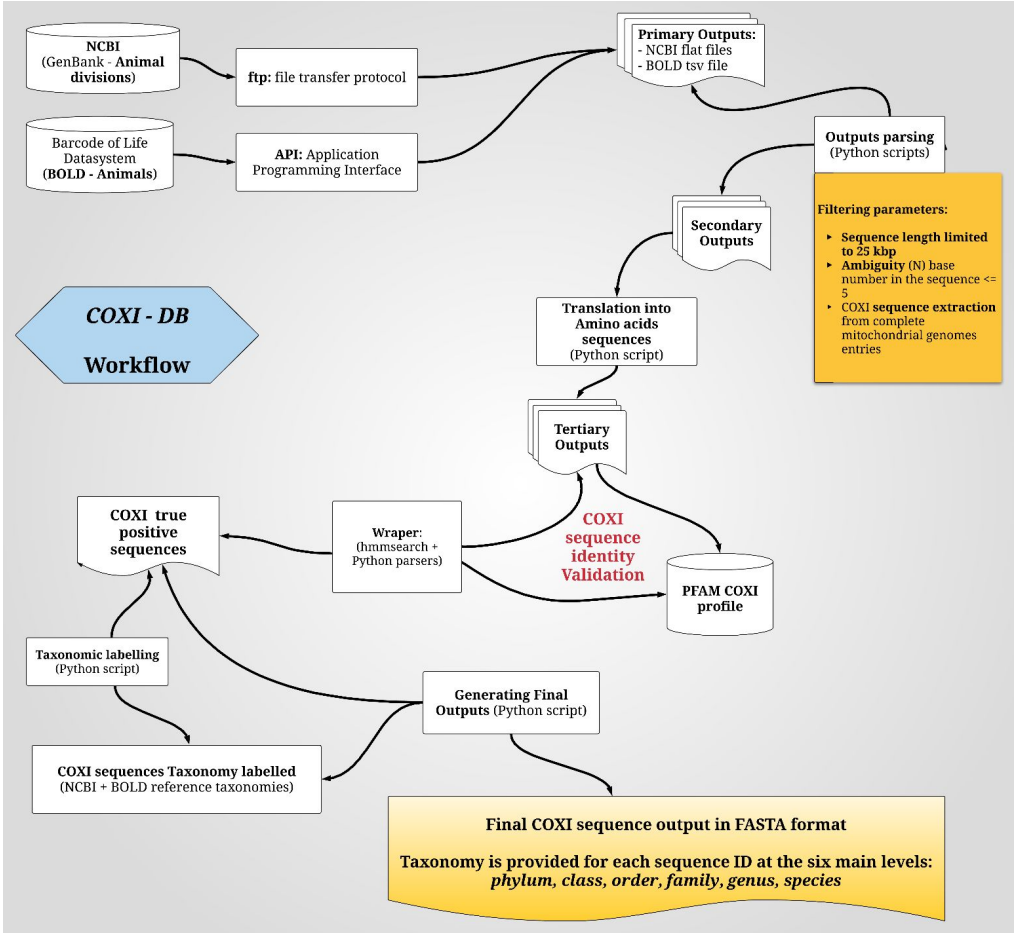
- A comprehensive **collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1)** sequences
- Hidden Markov Model (**HMM**) profiles of ITS1 flanking - 18S and 5.8S rRNA genes – are mapped on all the eukaryotic sequences in order to refine boundaries location of this region
- More than **1.1 million** curated sequences

Availability: http://itsonedb.cloud.ba.infn.it/

Santamaria, M., *et al.* (2018), *NAR*, **46**, D127-D132.

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

COXI-DB
on-going



- A **collection of Animals' Cytochrome Oxidase subunit-I (COXI)** sequences
- More than **5 million** curated sequences
- Sequences **validated** against PFAM (COXI profile)
- Updated records' Taxonomy according to **NCBI taxonomy** (*whenever possible)*

METAGENOMICS

Function-based Metagenomics

- Screen to identify functions of interest such as vitamins and antibiotic production

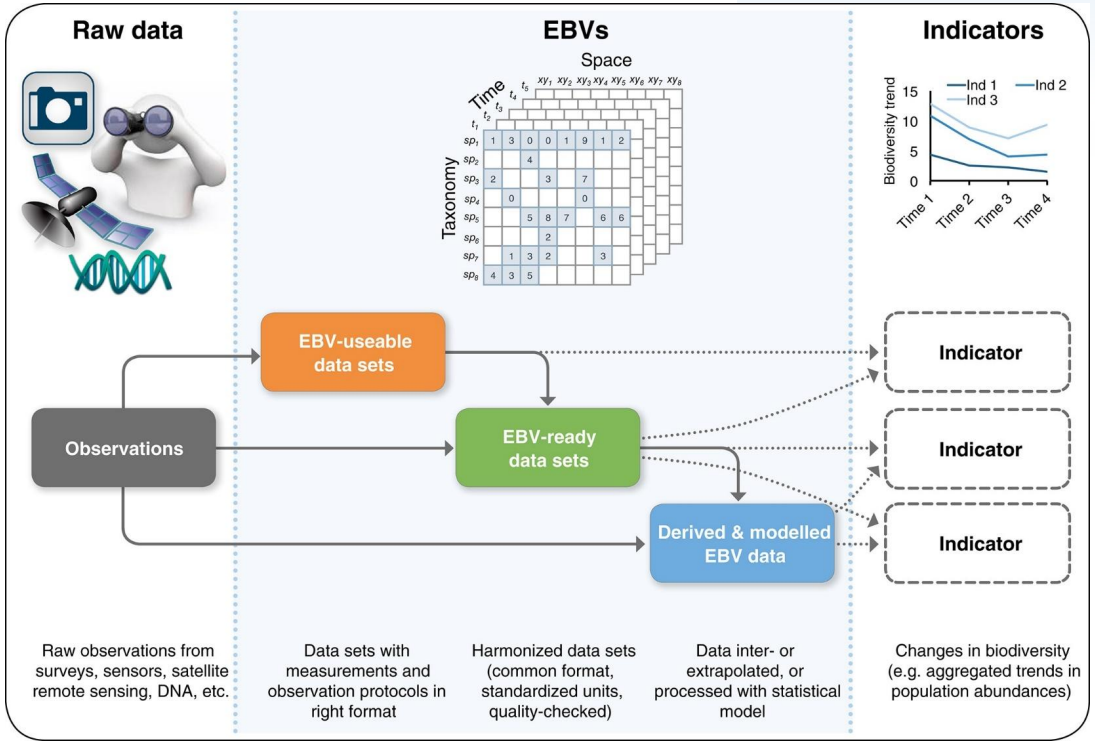
Extract data from microbial community in sampled environment

Sequence-based Metagenomics

- Determine the taxonomic composition of microbial communities;
- Determine what genes are represented, i.e. identify genes and metabolic pathways

Essential Biodiversity Variables (EBVs)

Globis-B



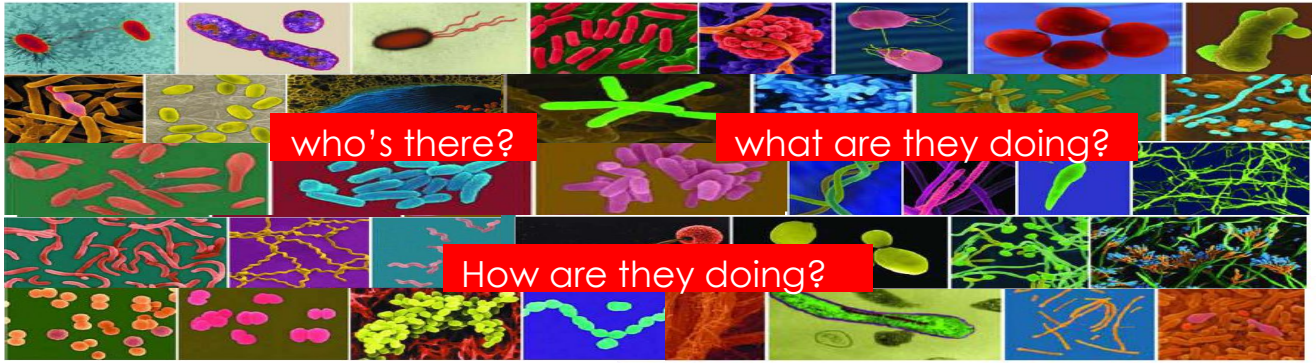
Kissling WD *et al.*, 2018

- Global Biodiversity **Monitoring**
- Studying, reporting and managing **Biodiversity change**
- **Harmonization** & **standardization** of biodiversity data

Analysis resources

Shotgun Metagenomics

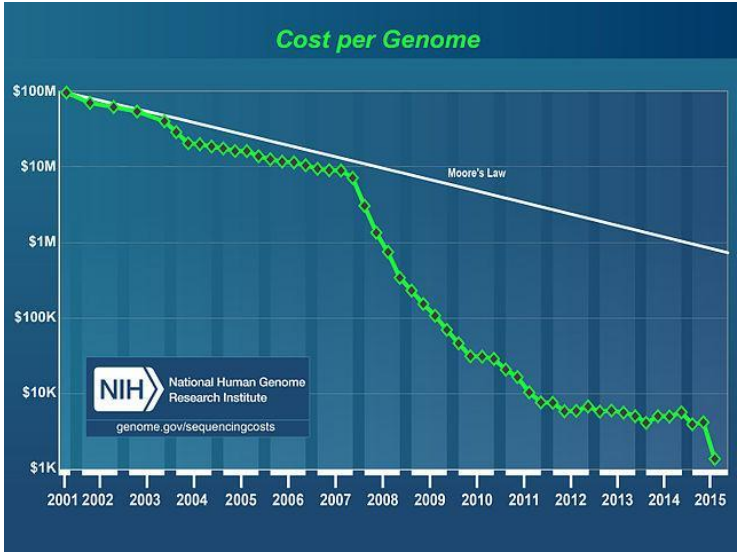
- identify species, genes and functional capabilities of mixed microbial communities;
- much more expensive in terms of sequencing and computational analysis.



Meta-Barcoding

- High sensitivity in species resolution and identification;
- Less expensive in terms of sequencing and computational analysis;
- Universal conditions of PCR;
- Specialized reference database (e.g. RDP for 16S, ITSoneDB for ITS1)
- may be biased due to the different efficiency of marker amplification in the different species;
- No functional information.

BIOINFORMATICS



- The critical bottleneck for NGS based projects is "Bioinformatics". The huge amount of sequence data generated by NGS platforms requires adequate computational infrastructures and bioinformatic resources for storage, retrieval and analysis of the data.

<u>The analysis of data requires advanced skills for establishing and running complex workflow including many steps.</u>

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

METASHOT

METASHOT is an automated pipeline designed for the identification of the microbial component in genomic (DNA-Seq) and transcriptomic (RNA-Seq) data.

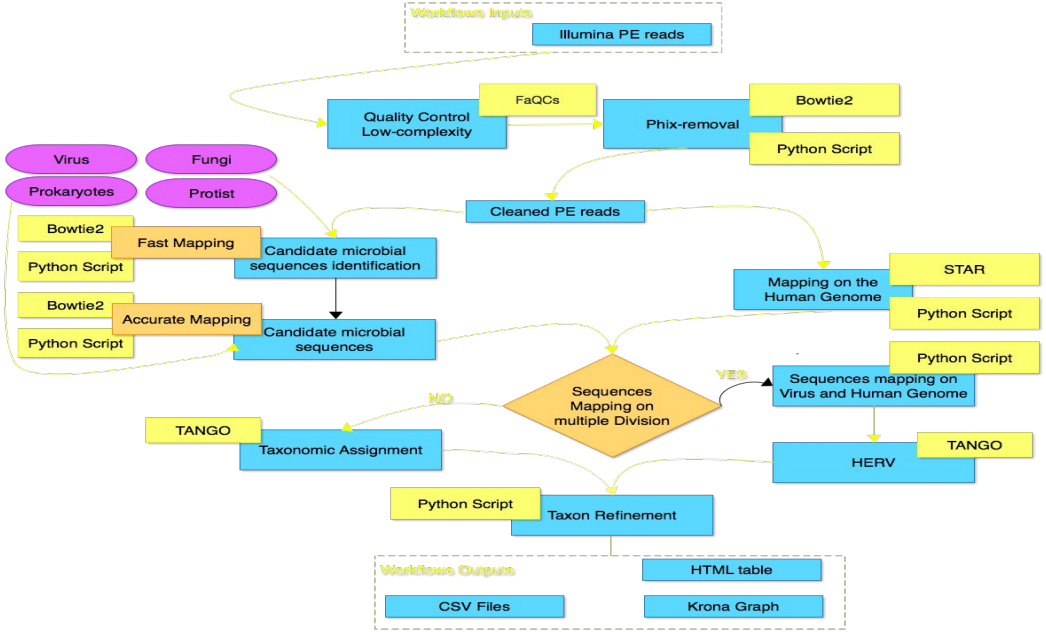
Third party tools and *ad hoc* developed Python and BASH scripts are integrated to manage, analyze and taxonomically assign Illumina PE data.

Bioinformatics, 2017, 1–3
doi: 10.1093/bioinformatics/btx036
Advance Access Publication Date: 27 January 2017
Applications Note
OXFORD

Sequence analysis

MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data

B. Fosso[1], M. Santamaria[1], M. D'Antonio[2], D. Lovero[1], G. Corrado[3], E. Vizza[3], N. Passaro[4], A.R. Garbuglia[5], M.R. Capobianchi[5], M. Crescenzi[4], G. Valiente[6] and G. Pesole[1,7,*]

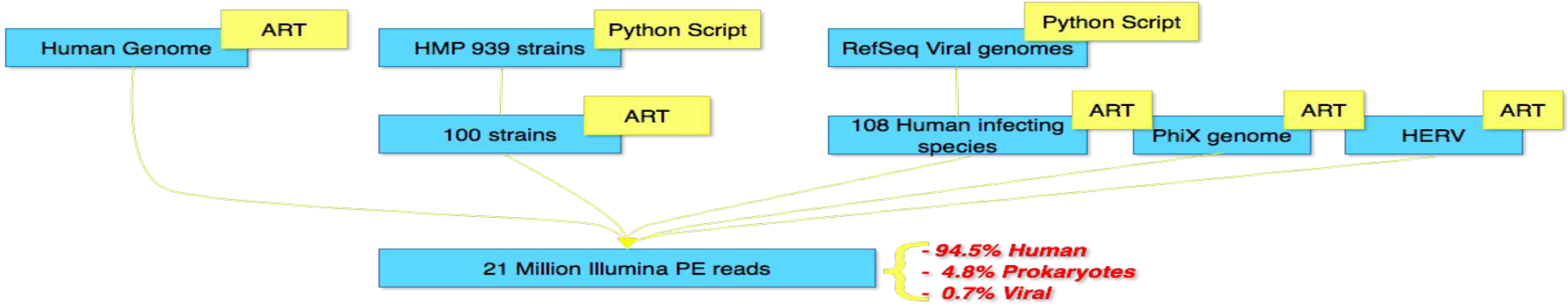


BENCHMARK

MetaShot has been benchmarked against ***Kraken*** and ***MetaPhlAn2***.

The Benchmark has been performed by using:

1. An in-silico generated human microbiome



2. A mock community consisting of 4 bacterial and 9 viral species (SRR3458569).

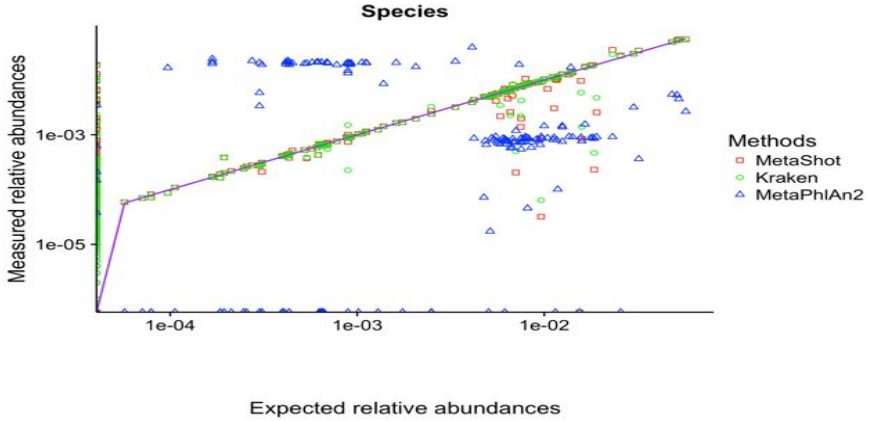
Article | OPEN

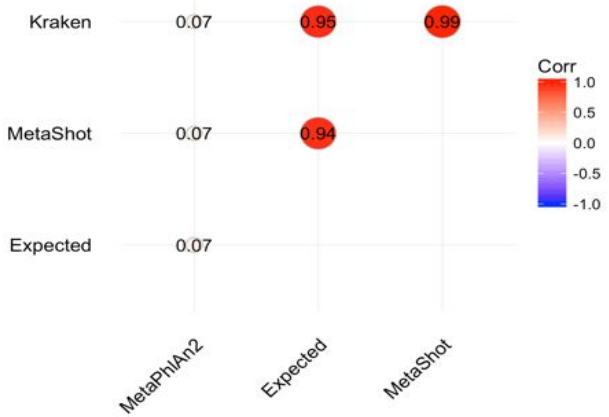
Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis

Nádia Conceição-Neto, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, Rob Lavigne, Piet Maes, Marc Van Ranst, Elisabeth Heylen & Jelle Matthijnssens

BENCHMARK

	Human (host)			Prokaryotes			Viruses		
	KR	MS	MP*	KR	MS	MP	KR	MS	MP
Precision	99.85	100.0	0	35.67	98.13	98.00	94.95	98.30	80.93
Recall	100.0	99.97	0	35.16	84.52	87.31	92.77	98.19	79.32
F-rate	99.92	100.0	0	35.36	86.79	90.72	92.82	98.07	79.93
Unclas	0.00	1.04	99.99	55.28	2.44	94.50	4.25	3.94	30.74





MetaShot outperforms ***Kraken*** and ***MetaPhlAn2*** in terms of the overall accuracy of reads assignment for the Prokaryotes and Viruses at the Family, Genus and Species levels.

Meta- Barcoding

SAMPLES COLLECTION

↓

METAGENOME EXTRACTION

↓

BARCODE AMPLIFICATION

↓

SEQUENCING

↓

BIOINFORMATIC ANALYSIS

Hypervariable regions of 16S rRNA

ITS1 of the ribosomal gene cluster

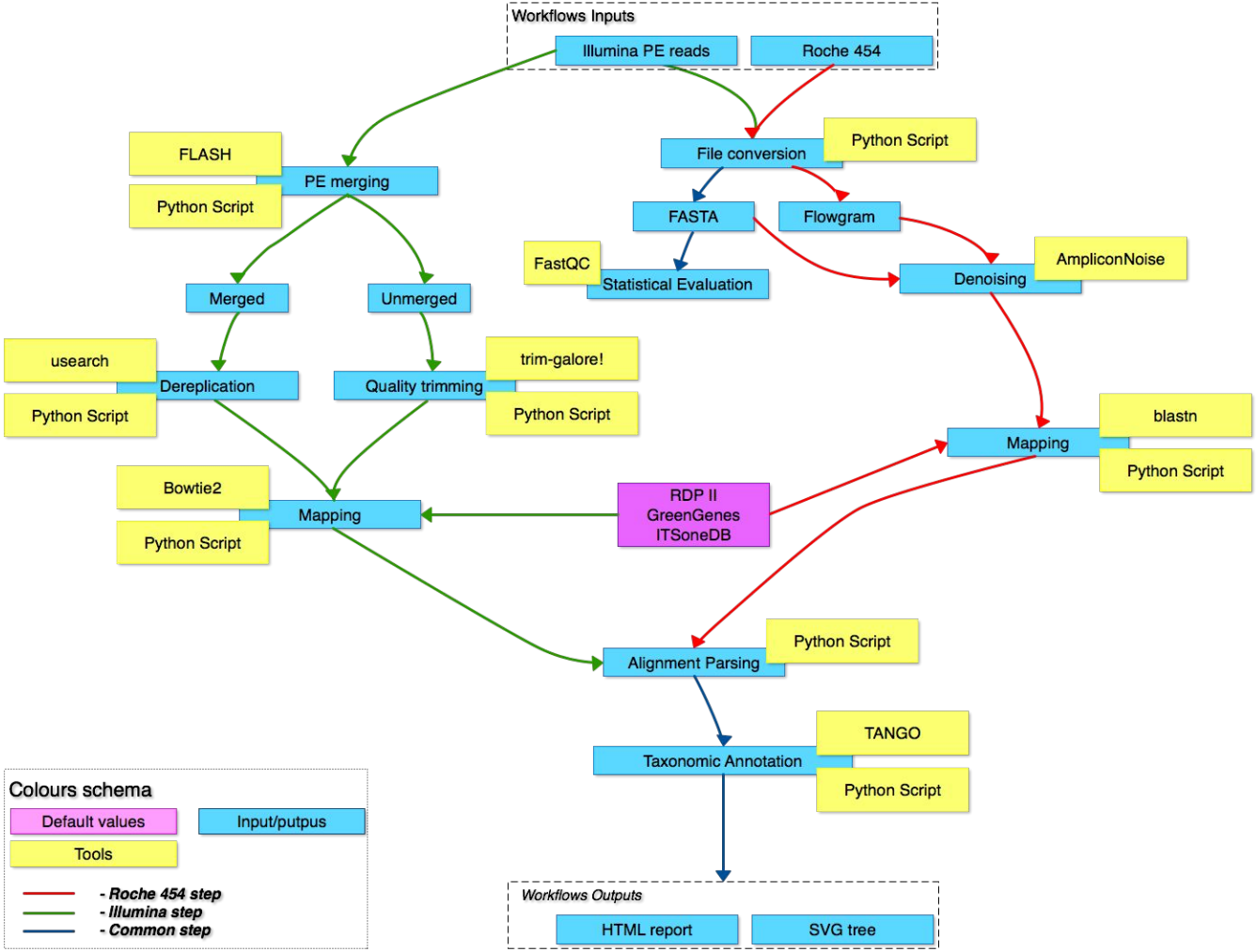
Taxonomic Genomic Markers

- Ubiquity in taxonomic range of interest (e.g. Bacteria, Fungi, etc)
- Reliable discrimination capacity at species level
- Hyper-variable regions flanked by highly conserved ones in taxonomic range under study
- Dimension fitting with the sequencing platforms read length

LifeWatch
E-Science European Infrastructure for Biodiversity and Ecosystem Research

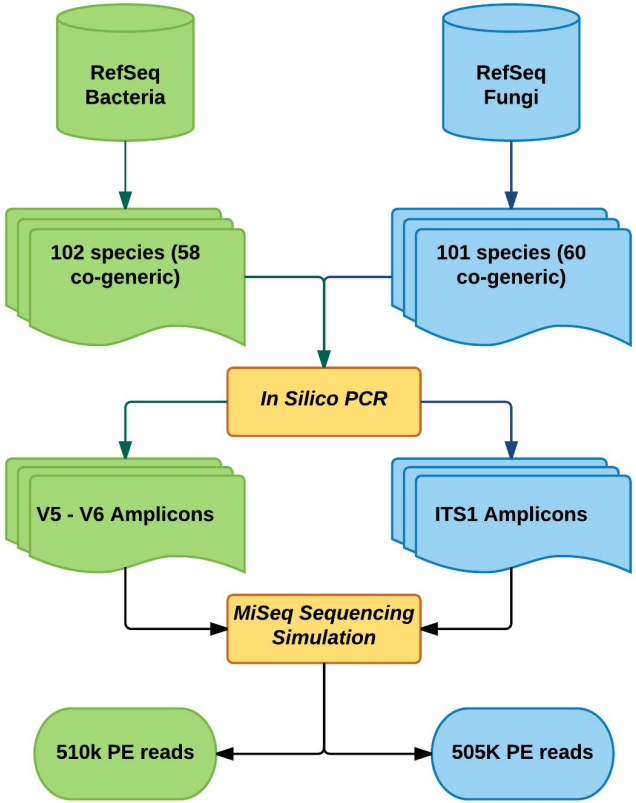
BioMaS

Bioinformatic analysis of Metagenomic AmpliconS

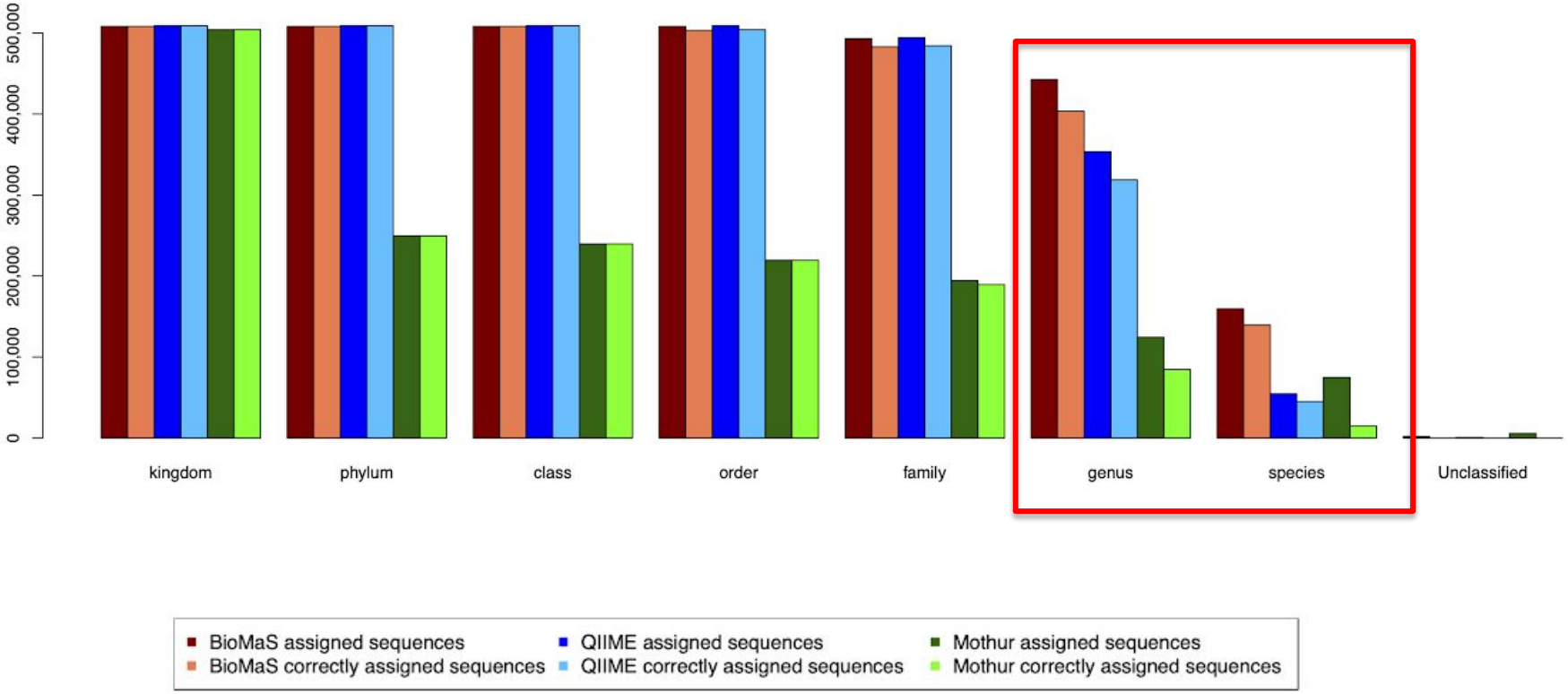


Benchmark

BioMaS classification performance has been compared to QIIME and Mothur



Benchmark: Bacteria



Benchmark: Fungi