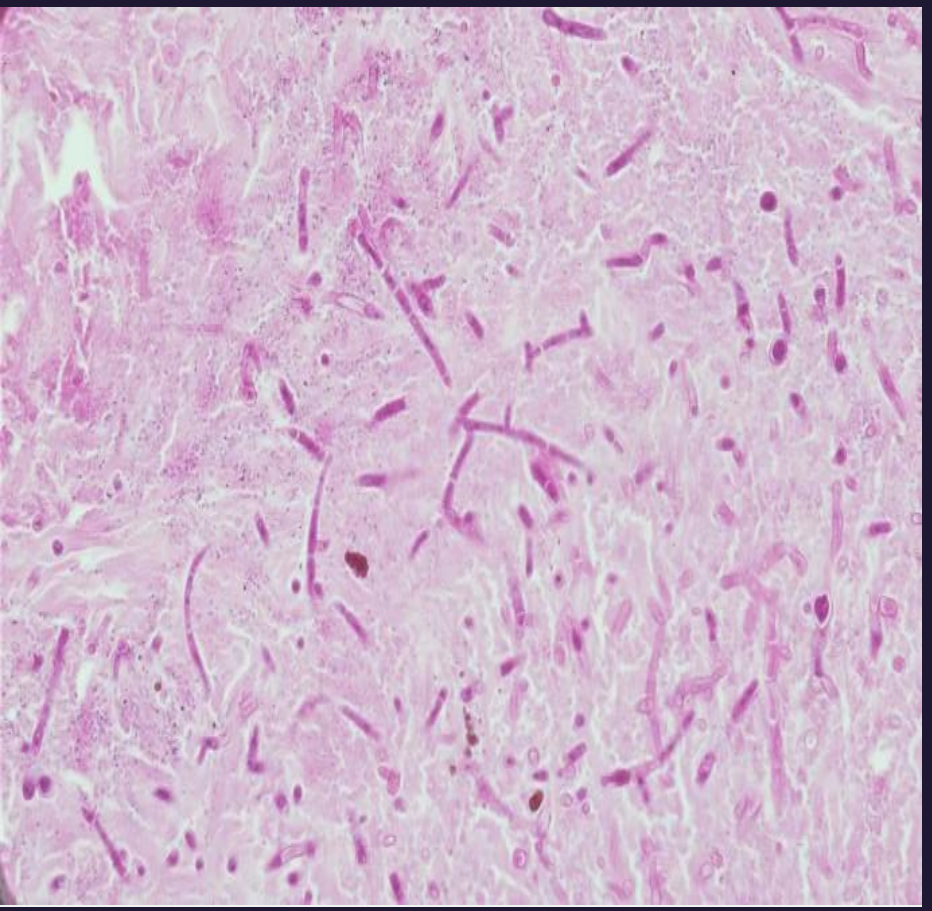


ITSoneDB V1.144 AND BioMaS@ITSoneWB: TWO ELIXIR-IT MAIN RESOURCES FOR AMPLICON BASED MYCOBIOME INVESTIGATION

Defazio G., Tangaro M.A., Pesole G., Fosso B.

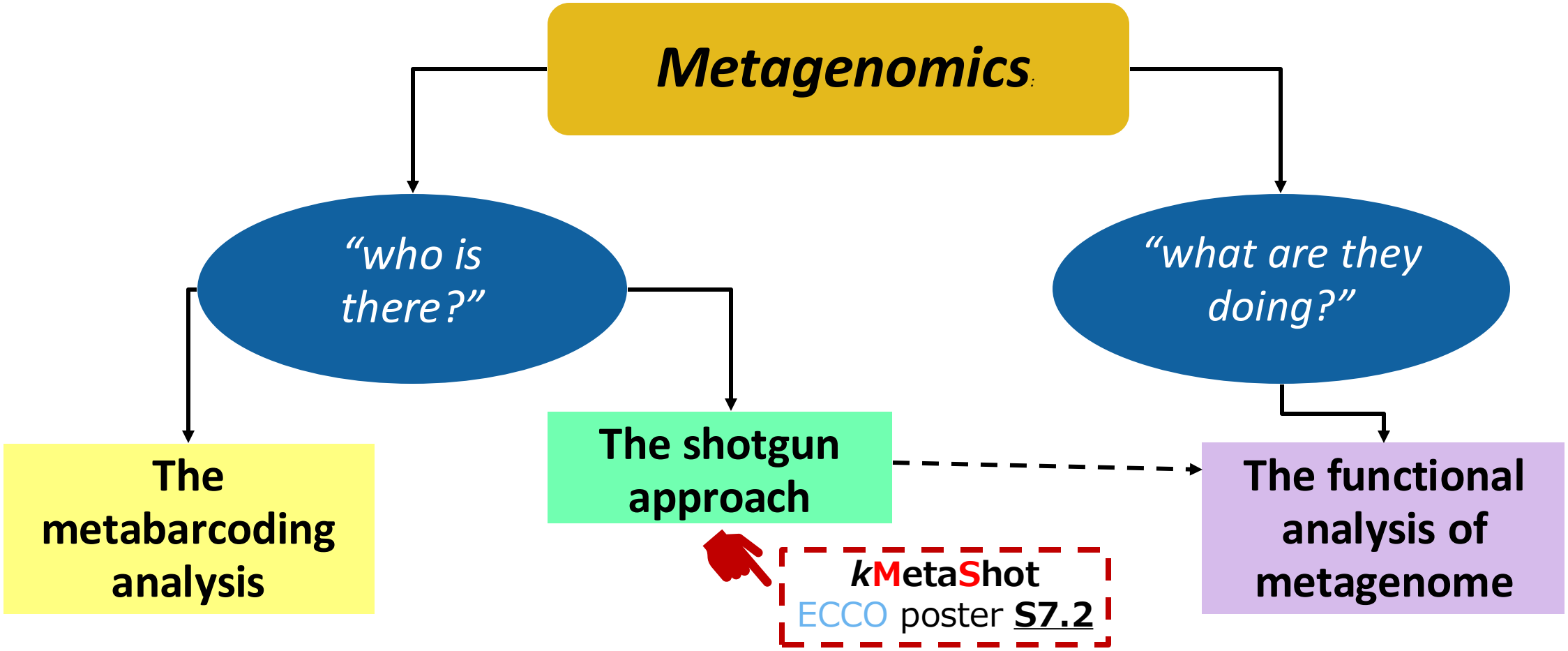
Giuseppe Defazio, Ph.D.
Department of Biosciences, Biotechnologies and Environment
University of Bari ''Aldo Moro''
giuseppe.defazio@uniba.it

IBIOM Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

IIB INFRASTRUTTURA ITALIANA DI BIOINFORMATICA | elixir ITALY

METAGENOMICS

DEFINITION: if genomics is the study of ”the complete set of the DNA molecules of an organism", metagenomics is literally "*beyond genomics*" ("meta" = "beyond") indeed, is the study of all microorganisms genomes present in an environmental or host sample.



METAGENOMICS

Shotgun

Metagenomics

Metabarcoding

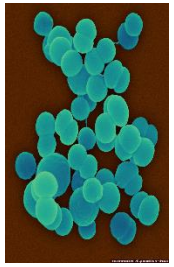
Reference: all available prokaryotic, eukaryotic and viral genomes

Reference: curated barcode collections ***SILVA, RDP UNITE, ITSoneDB***

The Metabarcoding Analysis

The amplification and sequencing of ‘barcode’ DNA regions

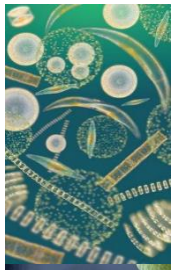
Prokaryota:



Eubacteria,
Archaebacteria

16S rRNA gene

Eukaryota:



Phytoplankton

18S rRNA gene

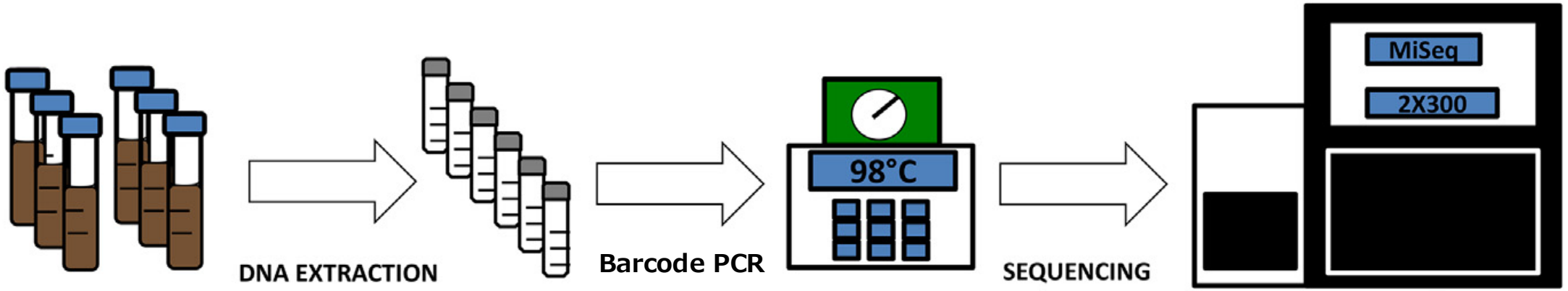
Fungi

ITS (Internal Transcribed Spacers)

Barcode features

- it is **ubiquitous** in the taxonomic research field;
- it is **variable** enough to allow the **discrimination** at the **deepest taxonomic levels**;
- it is **flanked by conserved regions** suited for the PCR primers design;
- it has a **size** compatible with currently available NGS technologies;

The Metabarcoding Analysis workflow



Raw data

Paired – end reads

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
```

Merging

Dereplication

Denoising

Chimera removal

Alignment vs. reference

Taxonomic assignment

Here we present…

ITSoneDB

release 1.144 (June 2024) is based on ENA April 2024 data

https://itsonedb.cloud.ba.infn.it/index.jsp

Galaxy COMMUNITY HUB

docker

BioMaS@ITSoneWB

Bioinformatic analysis of Metagenomic AmpliconS

```
docker pull ibiomcnr/biomas2
```

Who we are…

CNR.BiOmics
BIG DATA FOR BETTER LIFE

CNR Milano:
~1200 CPU cores and
5 PB storage (mirror)
CNR Milano:
5 PB storage (backup)

UNIVERSITÀ DEGLI STUDI DI MILANO
UNIVERSITÀ DEGLI STUDI DI PADOVA

UniMi:
INDACO update
UniBo:
IT infrastructure update
UniPd:
IT infrastructure update

elixir Next GenIT ITALY
Consolidation of the Italian Infrastructure for Omics Data and Bioinformatics

CNR (Bari) and University of Bari
Sequencing facility

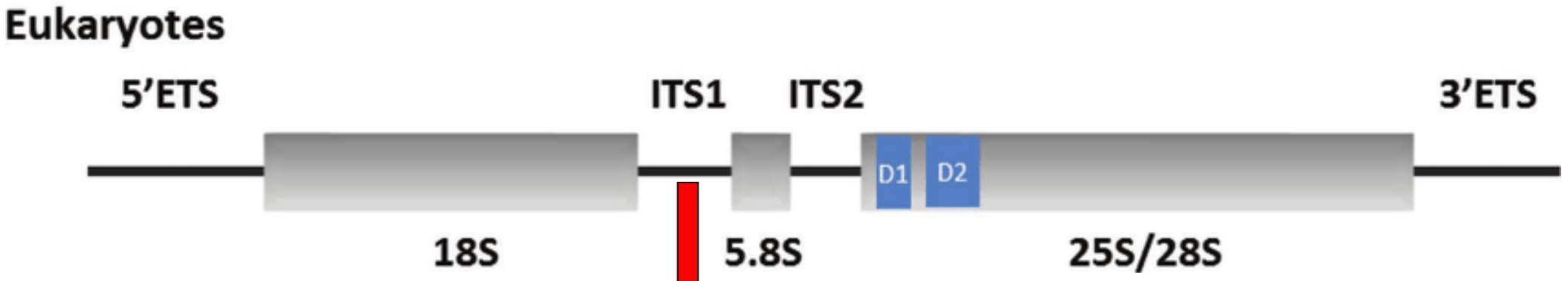
RNA &GENETHERAPY
ReCaS BARI

CNR Bari:
12.320 CPU cores, 10 GPUs
7,2 PB storage; 25Gb network
INFN Bari:
4.192CPU cores
2,1 PB storage; 10Gb network

UniBa (Physics Dept.)
4000 CPU cores and 16 GPUs
5.5 PB Cloud Storage
2 PB Posix Storage
20 PB Tape Library

Uniba (Physics Dept.)
5 PB Storage HPC/HTC
250 CPU cores and 8 GPUs
1.5 PB Storage Cloud

The ITSoneDB version 1.144



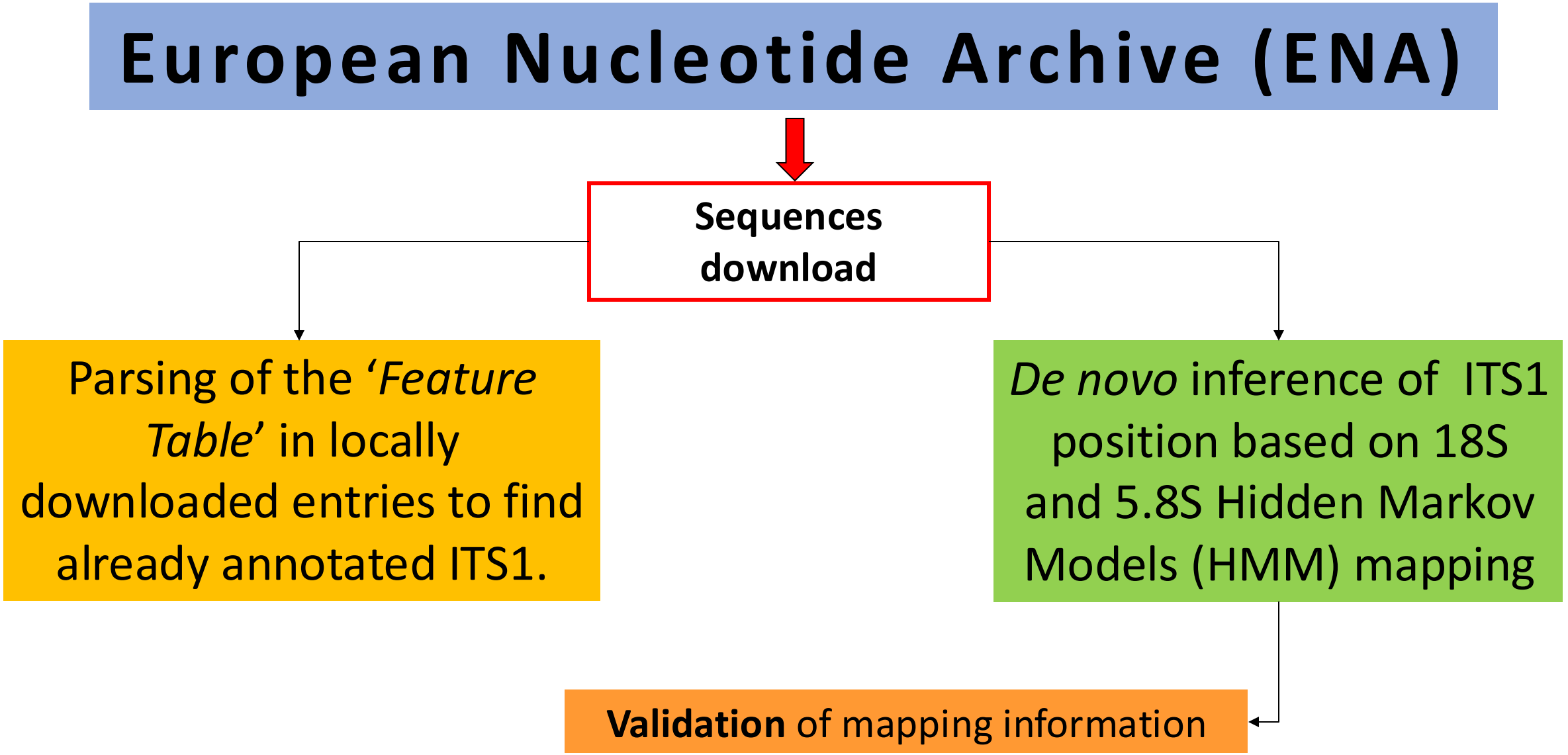
ITSoneDB

The database collects ITS1 sequences of eukaryotic microorganisms

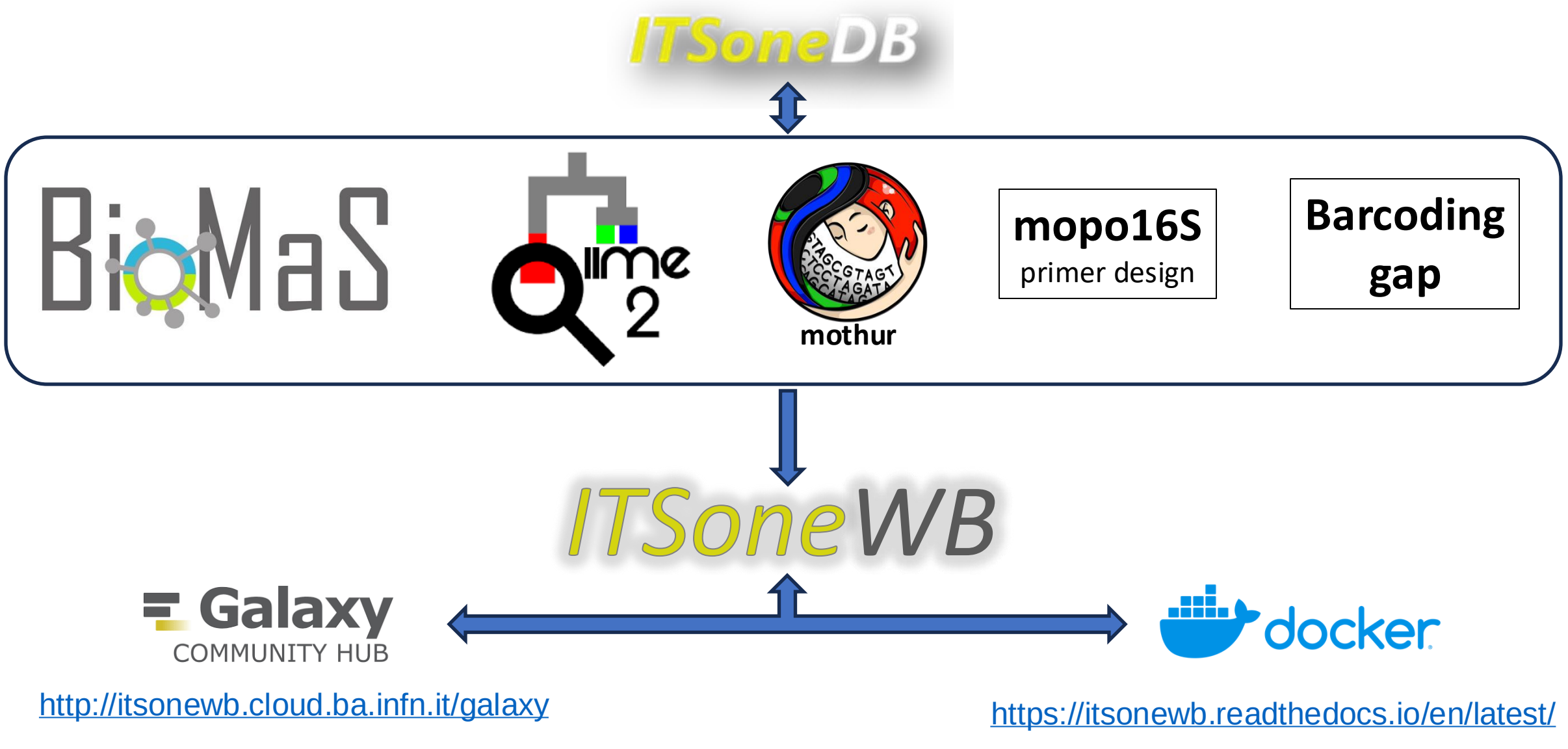
The Internal Transcribed Spacer 1 (ITS1) is the most reliable *'barcode'* for fungal communities survey.

Santamaria et al. 2018

The ITSoneDB population pipeline



The ITSone WorkBench (ITSoneWB)

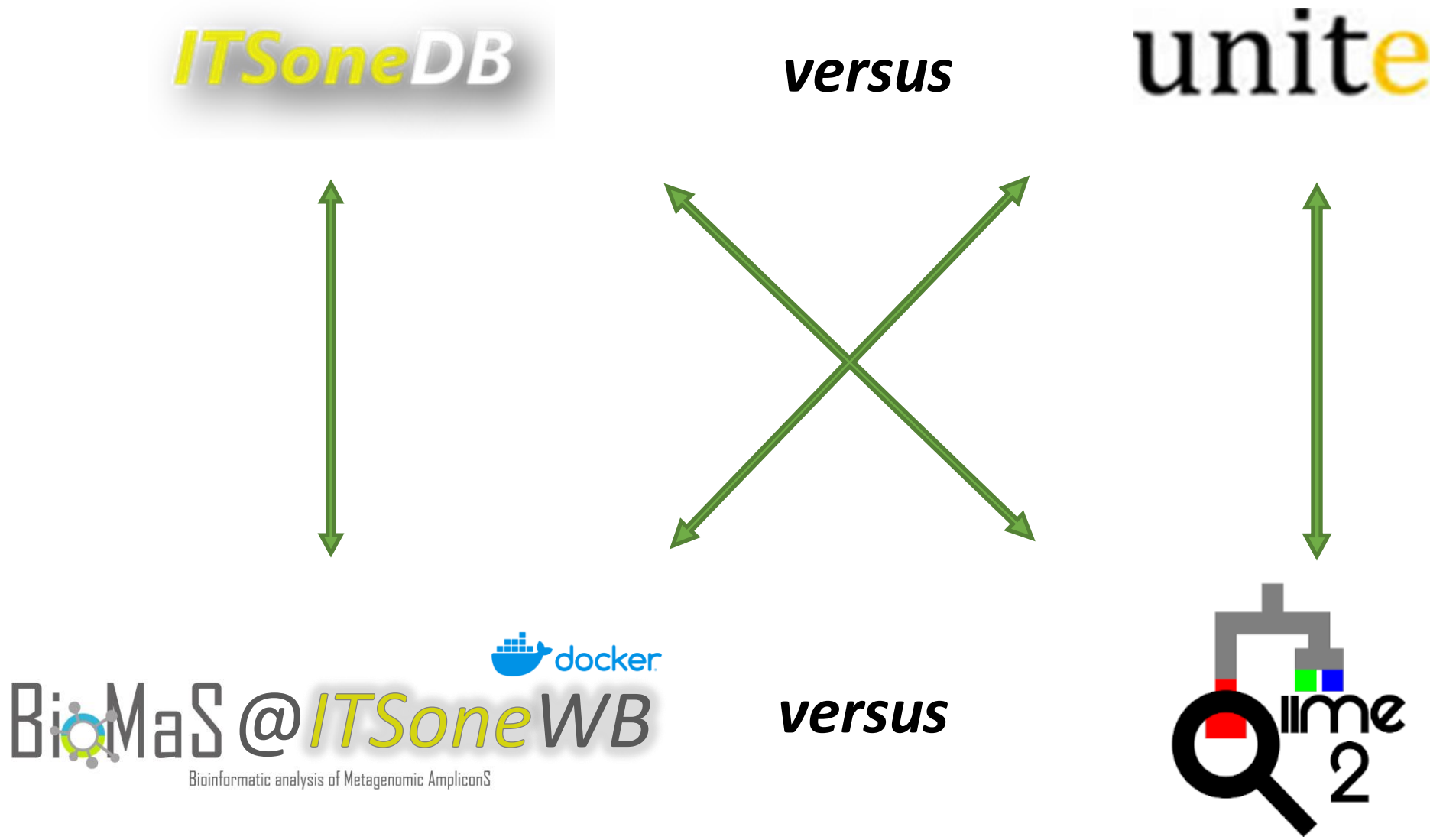


http://itsonewb.cloud.ba.infn.it/galaxy

https://itsonewb.readthedocs.io/en/latest/

Tangaro M.A, Defazio G. et al. 2021

Benchmark of ITSoneDB 1.144 and BioMaS@ITSoneWB docker

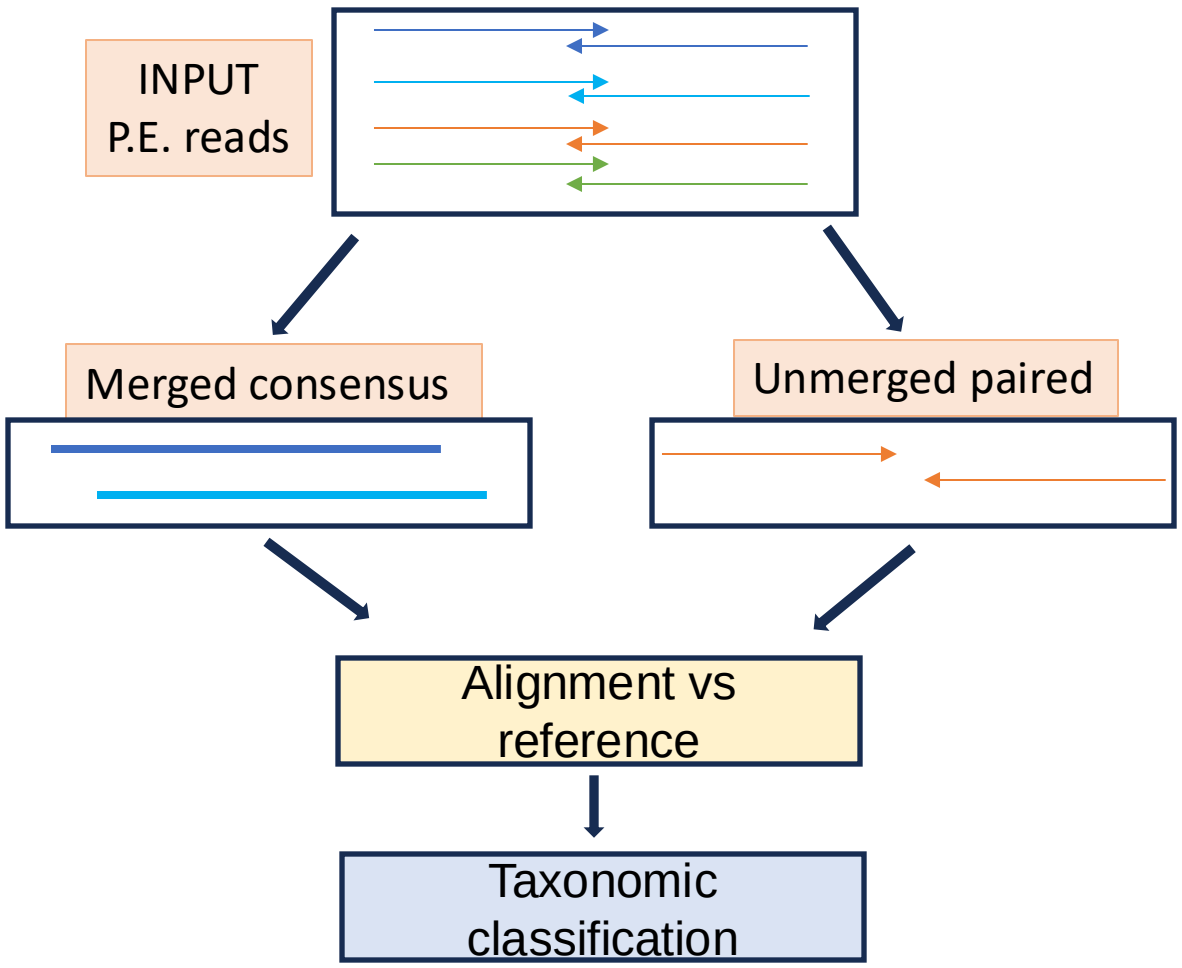


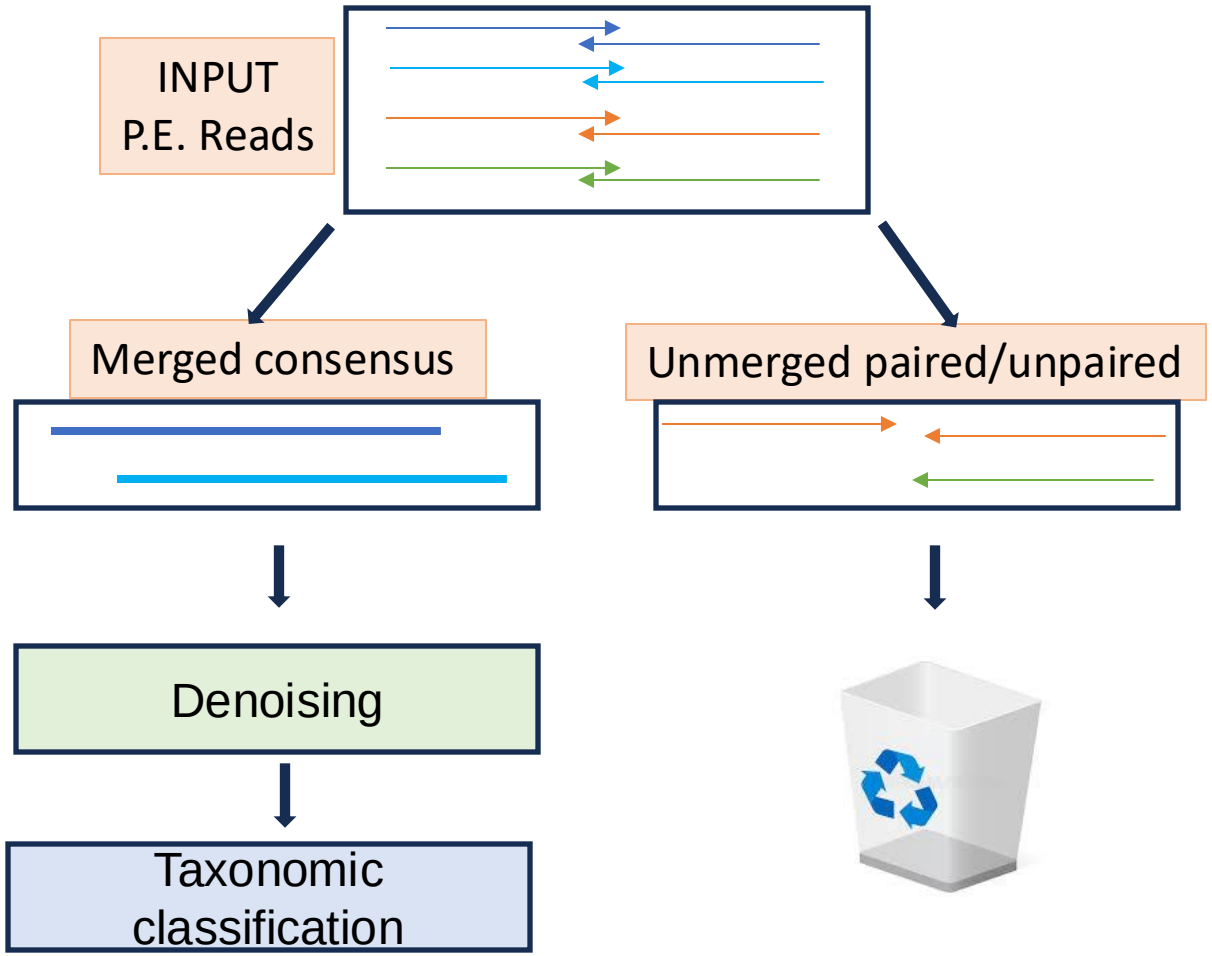
BioMaS vs. QIIME2 pros and cons

BioMaS @ITSoneWB

Bioinformatic analysis of Metagenomic AmpliconS

versus





Benchmark of ITSoneDB 1.144 and BioMaS@ITSoneWB docker

SAMPLE 1

- Real Mock community
- MiSeq platform
- Paired End 2 x 250
- P.E. reads 412,262

SAMPLE 2

- Sawdust sample of *Picea jezoensis var. hondoensis* deadwood
- MiSeq platform
- Run: DRR144721 Bioproject: PRJDB7194
- Paired End 2 x 300
- P.E. reads 290,585

SAMPLE 1

Mock community composition

Percentage	Organism Name	ATCC code
10	*Aspergillus fumigatus*	ATCC MYA-4609D-5
10	*Cryptococcus neoformans var. grubli*	ATCC 208821D-2
10	*Trichophyton interdigitale*	ATCC 9533D-5
10	*Penicillium chrysogenum*	ATCC 10106D-5
10	*Fusarium keratoplasticum*	ATCC 36031D-5
10	*Candida albicans*	ATCC 10231D-5
10	*Nakaseomyces glabratus* (alias *Candida glabrata*)	ATCC 2001D-5
10	*Malassezia globosa*	ATCC MYA-4612D-5
10	*Saccharomyces cerevisiae*	ATCC 201390D-5
10	*Cutaneotrichosporon dermatis*	ATCC 204094D-5

The ITSoneDB + BioMaS@ITSoneWB mock results at genus level

Selecting Items > than 0.01 %

SAMPLE 1

Sensitivity 90%
Precision 90%

	PE reads nr	PE reads %
Total PE reads	412,262	100
Merged	137,897	33.45
Classified	379,600	92.08
Correctly classified	322,395	78.2

Taxon Name	Total Assigned	Percent	
Cutaneotrichosporon	63,070	16.61	TP
Fusarium	56,327	14.83	TP
Cryptococcus	50,598	13.32	TP
Trichophyton	50,359	13.26	TP
Saccharomyces	36,081	9.50	TP
Nakaseomyces	25,701	6.77	TP
Aspergillus	17,624	4.64	TP
Candida	13,316	3.50	TP
Penicillium	9,319	2.45	TP
Trichosporon	7,397	1.94	FP (*)
Malassezia	0	0	FN

* *Trichosporon* is the basionym of Cutaneotrichosporon

The ITSoneDB + QIIME2 mock results at genus level

Selecting Items > than 0.01 %

SAMPLE 1

Taxon Name	Percentage	
Fusarium	50.09	TP
Cryptococcus	11.68	TP
Saccharomycetales_genus	11.36	FP
Cutaneotrichosporon	8.37	TP
Trichophyton	7.37	TP
Malassezia	5.41	TP
Aspergillus	3.39	TP
Penicillium	2.29	TP
Saccharomyces	0	FN
Nakaseomyces	0	FN
Candida	0	FN

Sensitivity 70%
Precision 87.5%

	PE reads nr	PE reads %
input	412,262	100
filtered	412,238	99.99
denoised	410,603	99.99
merged	318,858	77.34
non-chimeric	79,855	19.37
Correctly classified	38,011	9.22

The UNITE + BioMaS@ITSoneWB mock results at genus level

SAMPLE 1

The combination of UNITE + BioMaS@ITSoneWB produces no classification for any reads pair

The UNITE + QIIME2 mock results at genus level

Selecting Items > than 0.01 %

SAMPLE 1

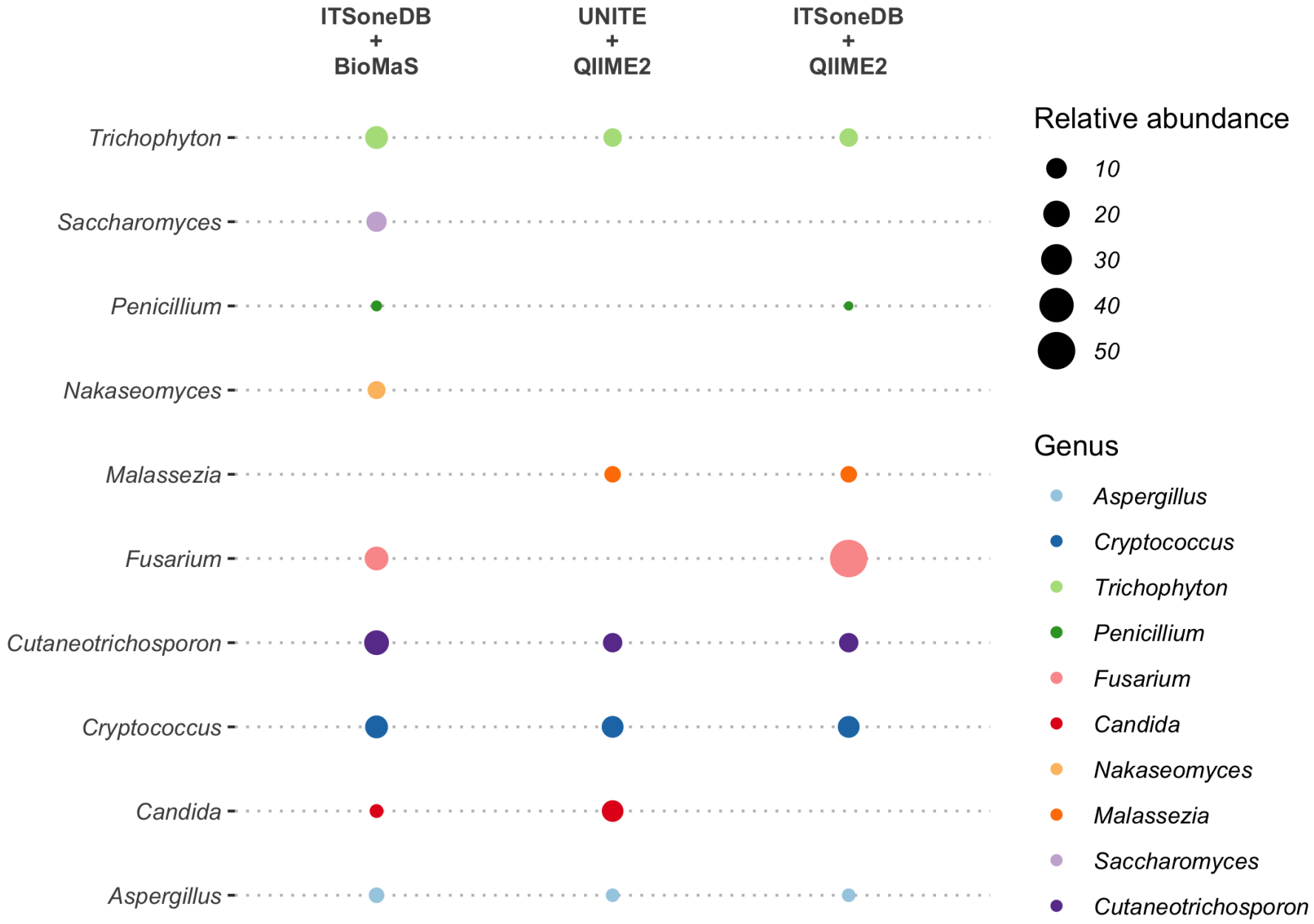
Taxon Name	Percentage	
Fungi_genus	52.39	FP
Cryptococcus	11.68	TP
Candida	11.36	TP
Cutaneotrichosporon	8.37	TP
Trichophyton	7.37	TP
Malassezia	5.41	TP
Aspergillus	3.39	TP
Saccharomyces	0	FN
Nakaseomyces	0	FN
Penicillium	0	FN
Fusarium	0	FN

Sensitivity 60%
Precision 85.7%

	PE reads nr	PE reads %
input	412,262	100
filtered	412,238	99.99
denoised	410,603	99.99
merged	318,858	77.34
non-chimeric	79,855	19.37
Correctly classified	38,011	9.22

Observed abundance of expected taxa

SAMPLE 1



The sawdust sample of *Picea jezoensis var. hondoensis* deadwood result comparison for taxonomic classification

SAMPLE 2

Input P.E. reads 290,585

BioMaS@ITSoneWB docker 498 genera found

Name	Reads	Percentage
Brachysporium	18,323	7.568
Curvichaeta	15,320	6.328
Hyphoderma	12,744	5.264
Acrodontium	10,870	4.490
Aniptodera	10,674	4.409
Leptodontidium	6,901	2.850
Ascocoryne	4,183	1.728

Unclassified P.E. reads 11

Brachysporium nigrum **wood** Saugatucket River (Lamore & Goos 1978)

Curvichaeta curvispora New Zealand **wood** (Reblova 2004)

Hyphoderma obtusiforme The **Corticiaceae** of North Europe (Erikss 1975)

Acrodontium **Fagus sylvatica decadey leaf** (de Hoog 1972)

Aniptodera from **Wood** in Freshwater Habitats (Shearer 1989)

Leptodontidium **root**-associated fungi (Melin 1922)

Ascocoryne **deadwood** associated fungi (Leonhardt et al. 2019)

QIIME2

Feature ID	ITSoneDB + QIIME2	UNITE + QIIME2	Reads
ASV1	Unassigned	k_Fungi	10
ASV2	g_Rhizophydium	k_Fungi	18

Discarded P.E. reads 290,557

Rhizophydium decomposing fungi

Conclusion

Nakaseomyces is appreciable exclusively with *ITSoneDB 1.144 + BioMaS@ITSoneWB docker* combination because of *Nakaseomyces grabratus* ITS1 sequence 862 nt long.

The highest Sensitivity and Precision for mock community are obtained with *ITSoneDB 1.144 + BioMaS@ITSoneWB docker.*

In both mock community and sawdust sample *ITSoneDB 1.144 + BioMaS@ITSoneWB docker* catches the highest diversity.

In real samples genera observed with *ITSoneDB 1.144 + BioMaS@ITSoneWB docker* are coherent with literature findings.

Conclusion

- Despite the large application in Prokaryotes metabarcoding analysis, Denoising and ASV inference applicability to ITS1 surveys is limited due barcode length variability (100-1000 nt) in eukaryotes.

- BioMaS relying on sequence mapping and direct classification, as higher sensitivity and accuracy for fungal/eukaryotes ITS1-based surveys.

- ITSoneDB shows a higher reliability and classifier tools adaptability than UNITE

- This pilot analysis should be confirmed by using a higher number of samples

IIB INFRASTRUTTURA ITALIANA DI BIOINFORMATICA

elixir ITALY

IBIOM Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

Thank you for your kind attention

Contacts: bruno.fosso@uniba.it
graziano.pesole@uniba.it