MetaShot: a complete workflow for the characterization of human microbiome from shotgun data

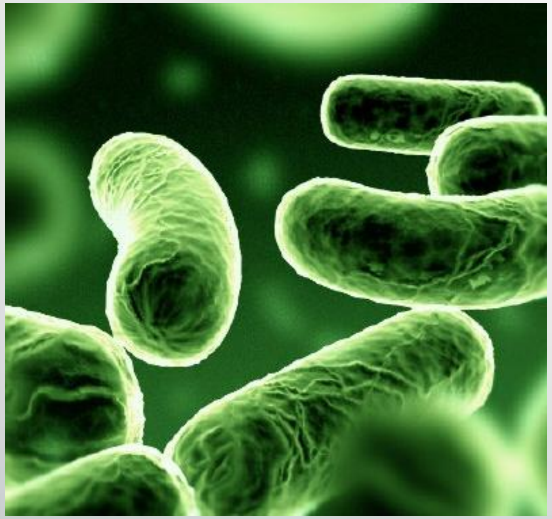
Joint NETTAB and Integrative Bioinformatics meeting 2015

Bari, October 14th

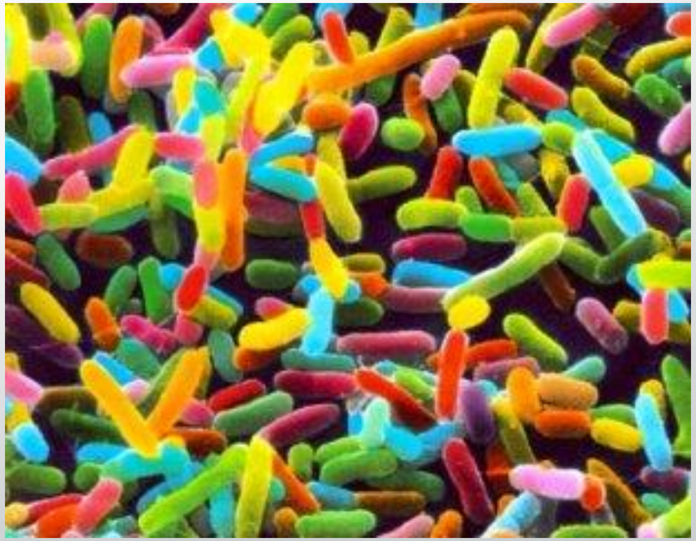
Bruno Fosso, Ph.D. – IBBE-CNR

METAGENOMICS

Metagenomics is an innovative methodological approach that allows to unveil the composition and function of mixed microbial communities in any environmental niche. Indeed, the **Biodiversity** of each environment, i.e. all living organisms (mostly microbial) can be fully represented by their genetic material.



Single microbial species

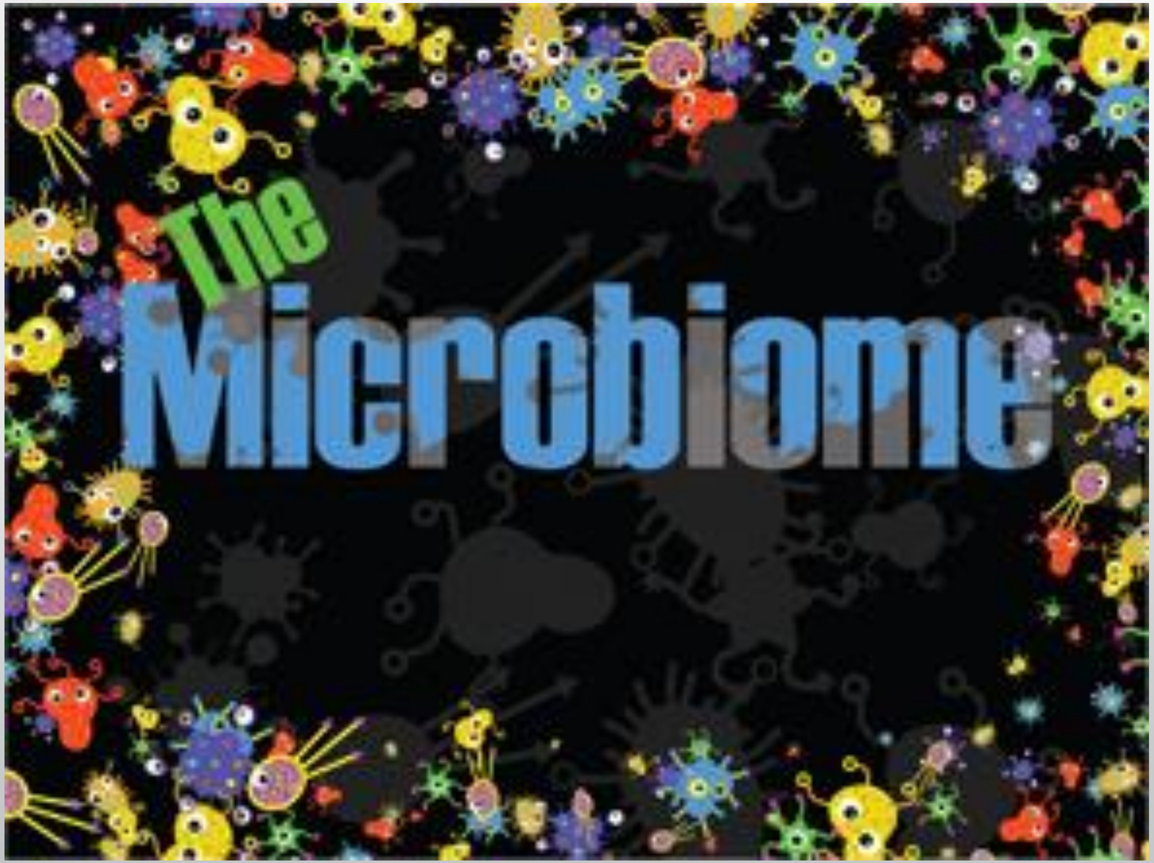


Mixed microbial community
living into a specific environment

Kunin et al., 2008; Wooley et al., 2010

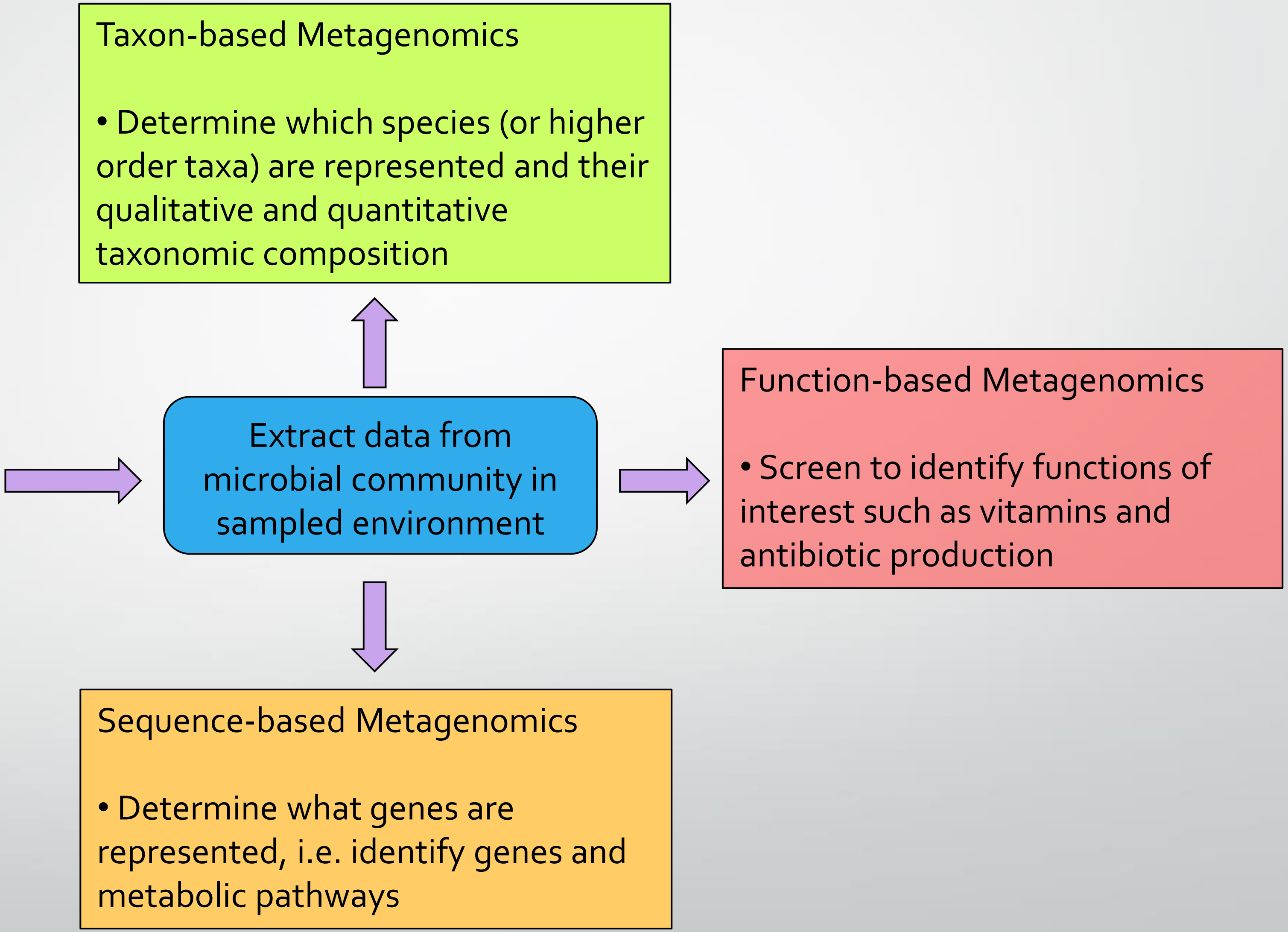
METAGENOMICS

The **metagenome** consists in the ensemble of the genetic material extracted from all microbial species (i.e. the **Microbiome**) living in a given environmental sample, including those which could not be isolated and cultivated in the lab (≥ 99%).



THE METAGENOMIC ANALYSIS

NETTAB & IB 2015



Extract data from microbial community in sampled environment

Taxon-based Metagenomics

- Determine which species (or higher order taxa) are represented and their qualitative and quantitative taxonomic composition

Function-based Metagenomics

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

Sequence-based Metagenomics

- Determine what genes are represented, i.e. identify genes and metabolic pathways

DIFFERENT APPROACHES FOR METAGENOMICS

- Target-oriented metagenomics (amplicon-based)

  Massive parallel sequencing of a specific target region (e.g. 16S rRNA or ITS) from amplicons obtained by using universal primers specific for a given (the larger as possible) taxonomic group.

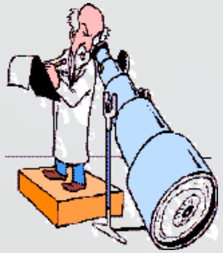
- Shotgun metagenomics

  Shotgun sequencing of total DNA (or RNA) extracted from environmental samples.

The first approach is particularly suitable for specific taxonomic groups, for which universal conserved primers are available which are able to amplify a targeted genome region in a large number of species (e.g. Bacteria).
RNA deep-sequencing (Metatranscriptomics) may provide the overall global expression profile of the microbial community in the environmental sample (i.e. which genes are expressed and how much)

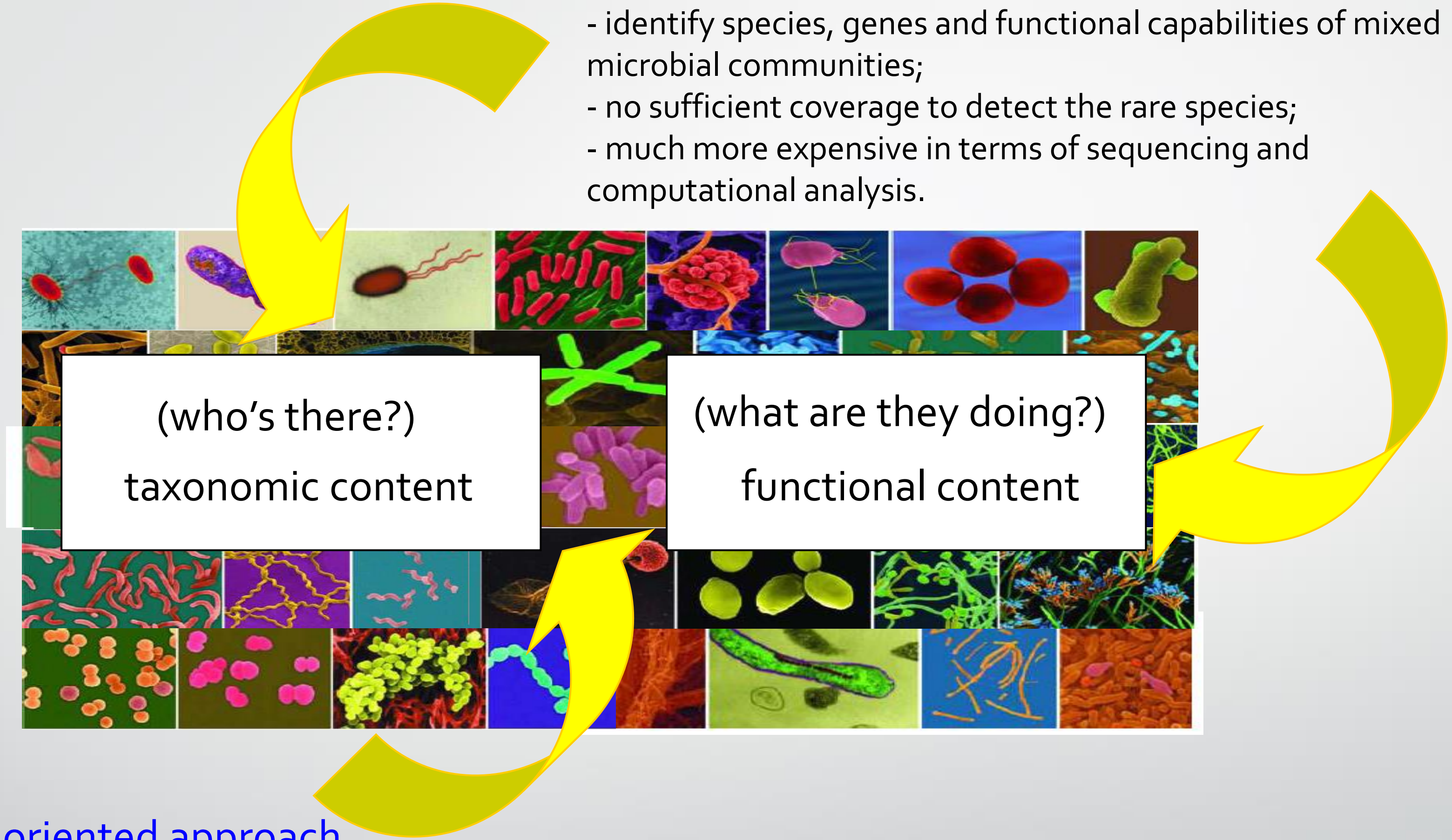
NETTAB & IB 2015

NGS FOR METAGENOMICS



Random shotgun NGS approach

- identify species, genes and functional capabilities of mixed microbial communities;
- no sufficient coverage to detect the rare species;
- much more expensive in terms of sequencing and computational analysis.



(who's there?)
taxonomic content

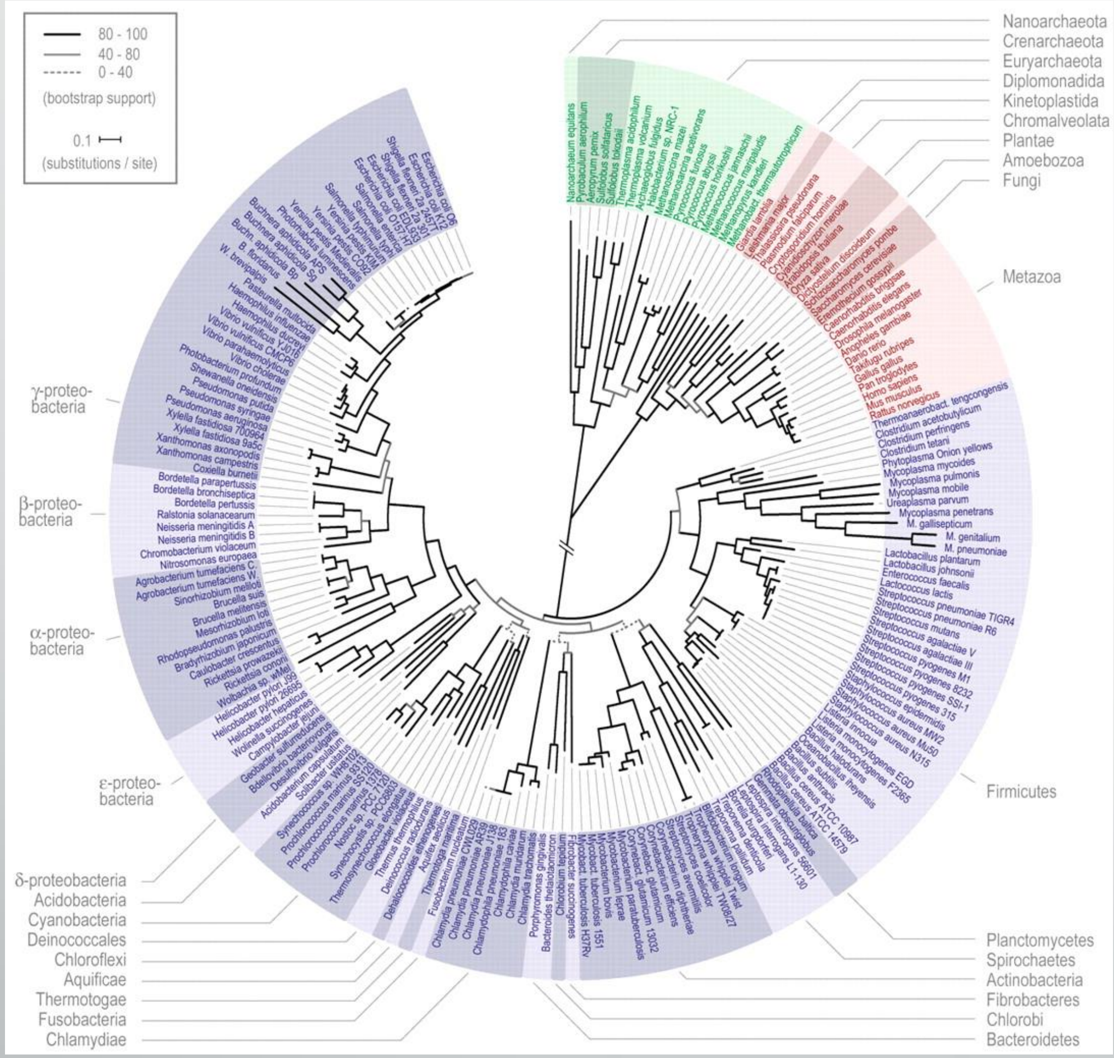
(what are they doing?)
functional content

Target-oriented approach

- 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.

NGS-BASED METAGENOMICS: A DEEP INSIGHT IN EVOLUTION

NETTAB & IB 2015



The large-scale exploration of metagenomic data gives us the extraordinary and unprecedented possibility to unravel the taxonomic complexity of **ALL** living beings (not only those that can be cultured in the lab, <5% of the total) and … more to gain a comprehensive overview of the products of evolution and selection in different environments and conditions.

New genes and functions can be discovered to foster a large variety of biotechnological processes and applications.

HUMAN MICROBIOME

What it means to be Human?

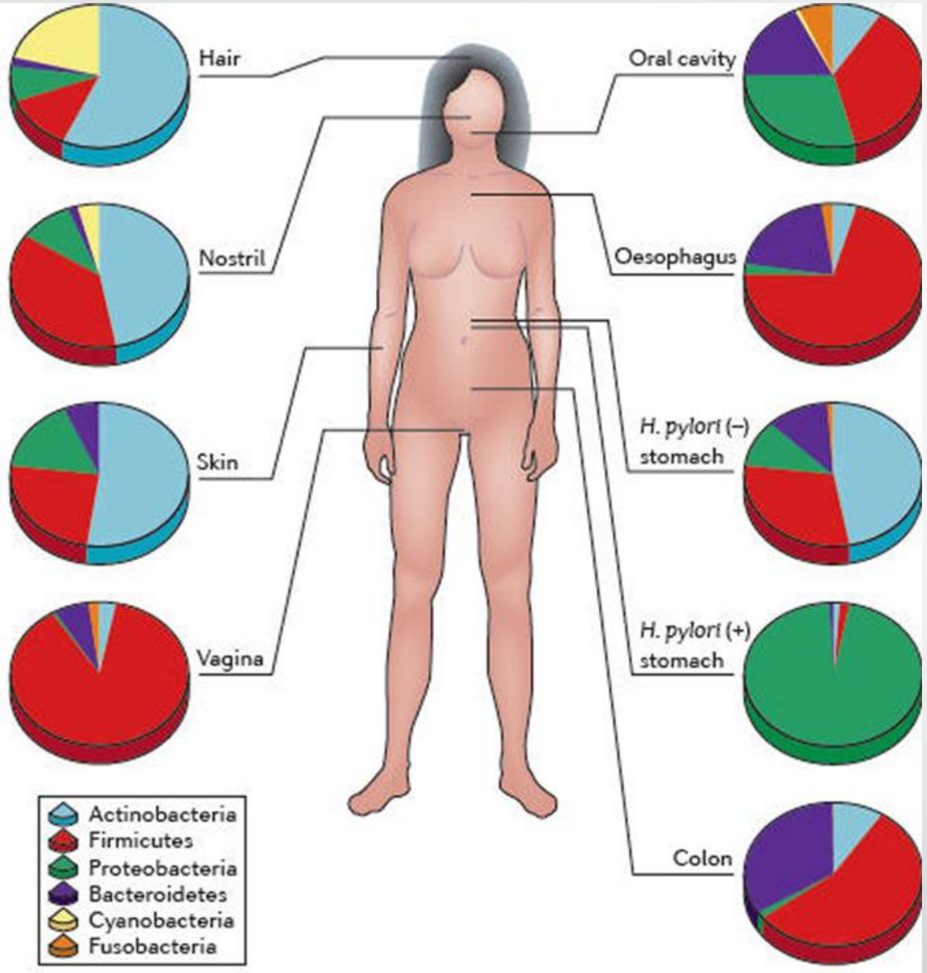
- Our body is essentially sterile during gestation;

- Starting from birth it is colonized by a tremendous diversity of bacteria, archaea, fungi, and viruses.

- The advent of Metagenomics and NGS technologies allowed the investigation of the complex relationship between the human body and its microbial communities.

HUMAN MICROBIOME

NETTAB & IB 2015



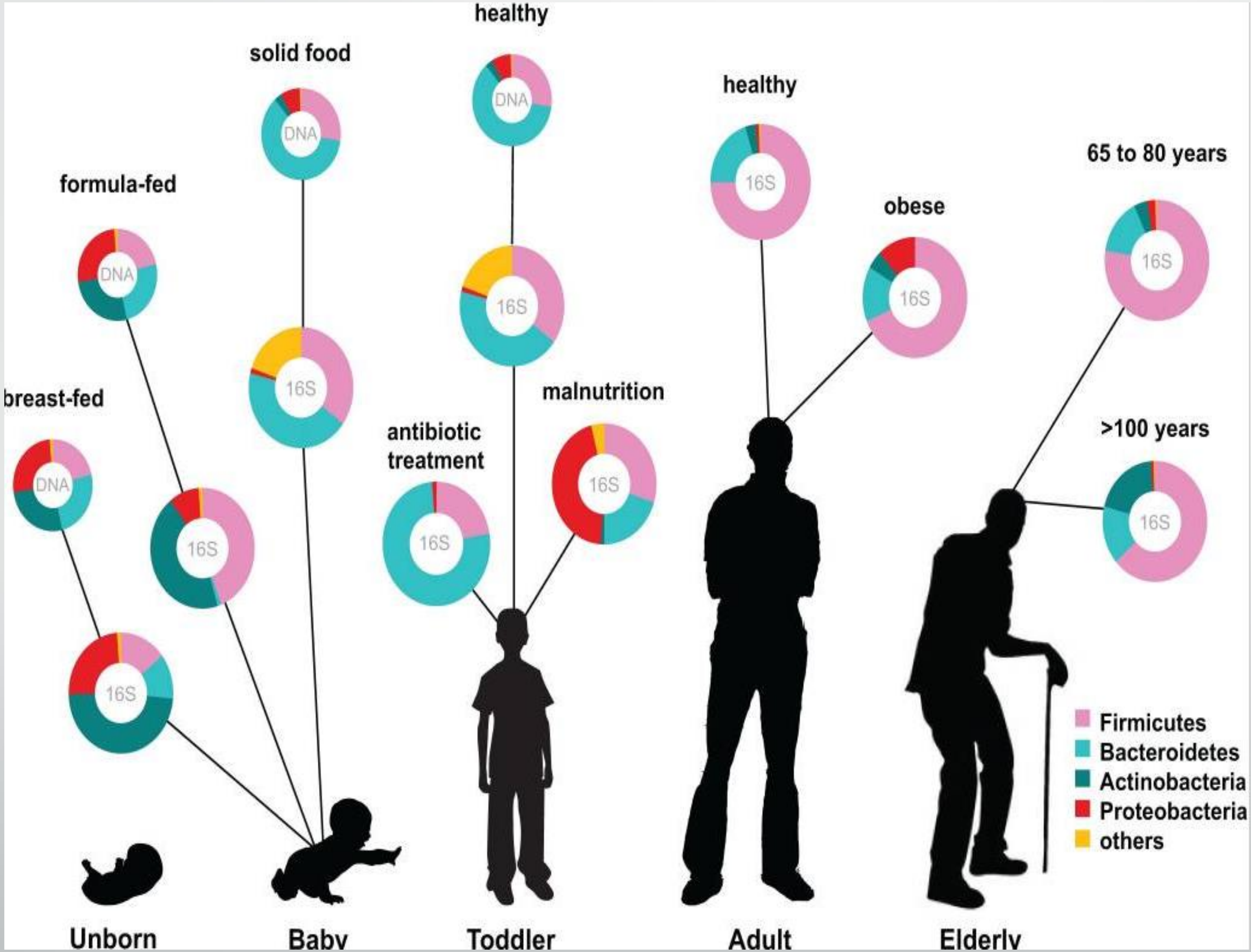
Humans: Meta-organisms

10-fold greater numbers of microbial than human cells with a biomass >1 Kg

Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012 Mar 13;13(4):260-70. doi: 10.1038/nrg3182. Review. PubMed PMID: 22411464; PubMed Central PMCID: PMC3418802.

HUMAN MICROBIOME

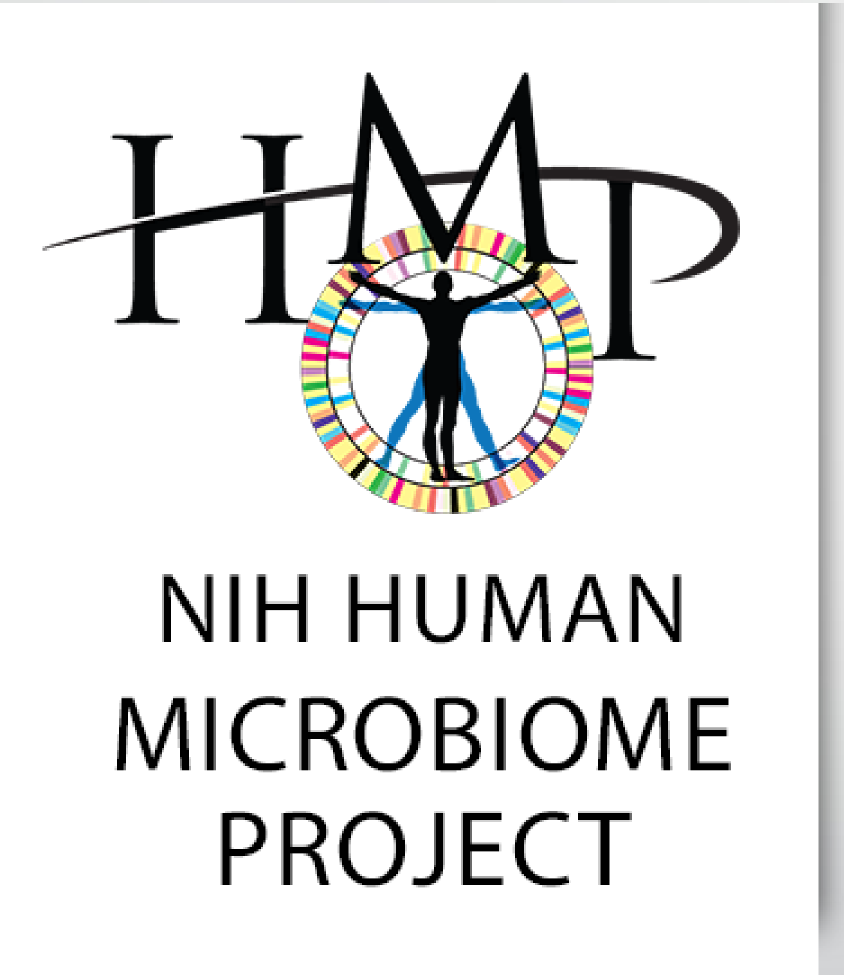
NETTAB & IB 2015



Ottman N, Smidt H, de Vos WM, Belzer C. The function of our microbiota: who is out there and what do they do? Front Cell Infect Microbiol. 2012;2:104.

NETTAB & IB 2015

HUMAN MICROBIOME PROJECT



- Development of a reference set of 3,000 isolate microbial genome sequences;
- Initial 16S & mWGS metagenomic studies to generate an estimate of the complexity of the microbial community at each body site;
- Demonstration projects to determine the relationship between disease and changes in the human microbiome;
- Development of new tools and technologies for computational analysis, establishment of a data analysis and coordinating center (DACC), and resource repositories;
- Examination of the ethical, legal and social implications (ELSI) to be considered in the study and application of the metagenomic analysis of the human microbiota;

http://hmpdacc.org

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.

- The analysis of data requires advanced skills for establishing and running complex workflow including many steps.

LifeWatch

MoBiLab

Laboratorio di Biodiversità Molecolare

elixir

In this scenario we developed three new bioinformatic resources aimed to support molecular researches in advanced analyses of NGS metagenomic data:

- ***BioMaS*** *(Bioinformatic analysis of Metagenomic ampliconS)*, a comprehensive pipeline for the taxonomic analyses of meta-barcode NGS dataset;
- ***ITSoneDB***, a curated collection of taxonomically annotated ITS1 sequences suitable for metagenomic studies of fungal communities;
- ***MetaShot (Metagenomics Shotgun)***, a pipeline for the taxonomic characterization of shotgun NGS metagenomic data, particularly oriented towards the study of human and other host microbiome.

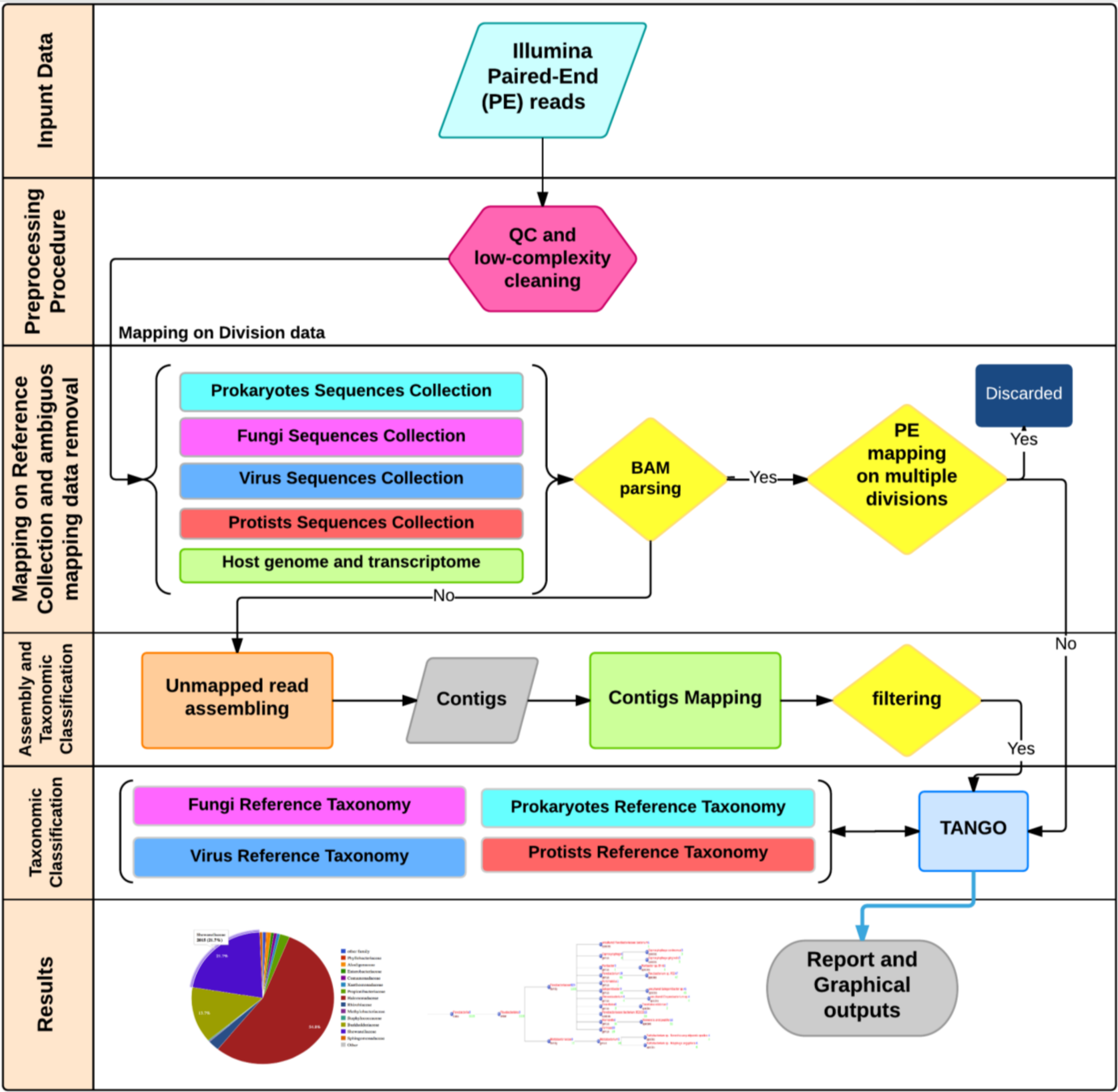
METASHOT

METASHOT is an automated pipeline designed for the identification of 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.

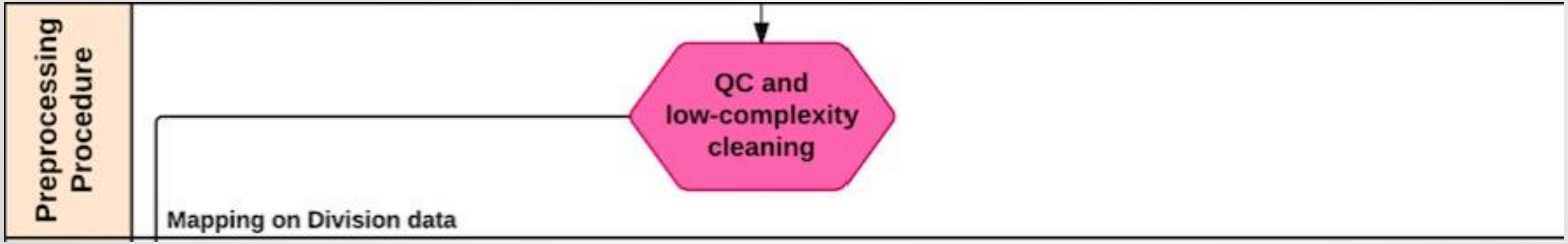
NETTAB & IB 2015

METASHOT



Inpunt Data
Illumina Paired-End (PE) reads
Preprocessing Procedure
QC and low-complexity cleaning
Mapping on Division data
Mapping on Reference Collection and ambiguos mapping data removal
Prokaryotes Sequences Collection
Fungi Sequences Collection
Virus Sequences Collection
Protists Sequences Collection
Host genome and transcriptome
BAM parsing
Yes
PE mapping on multiple divisions
Yes
Discarded
No
No
Assembly and Taxonomic Classification
Unmapped read assembling
Contigs
Contigs Mapping
filtering
Yes
Taxonomic Classification
Fungi Reference Taxonomy
Virus Reference Taxonomy
Prokaryotes Reference Taxonomy
Protists Reference Taxonomy
TANGO
Results
Report and Graphical outputs

META*SHOT*

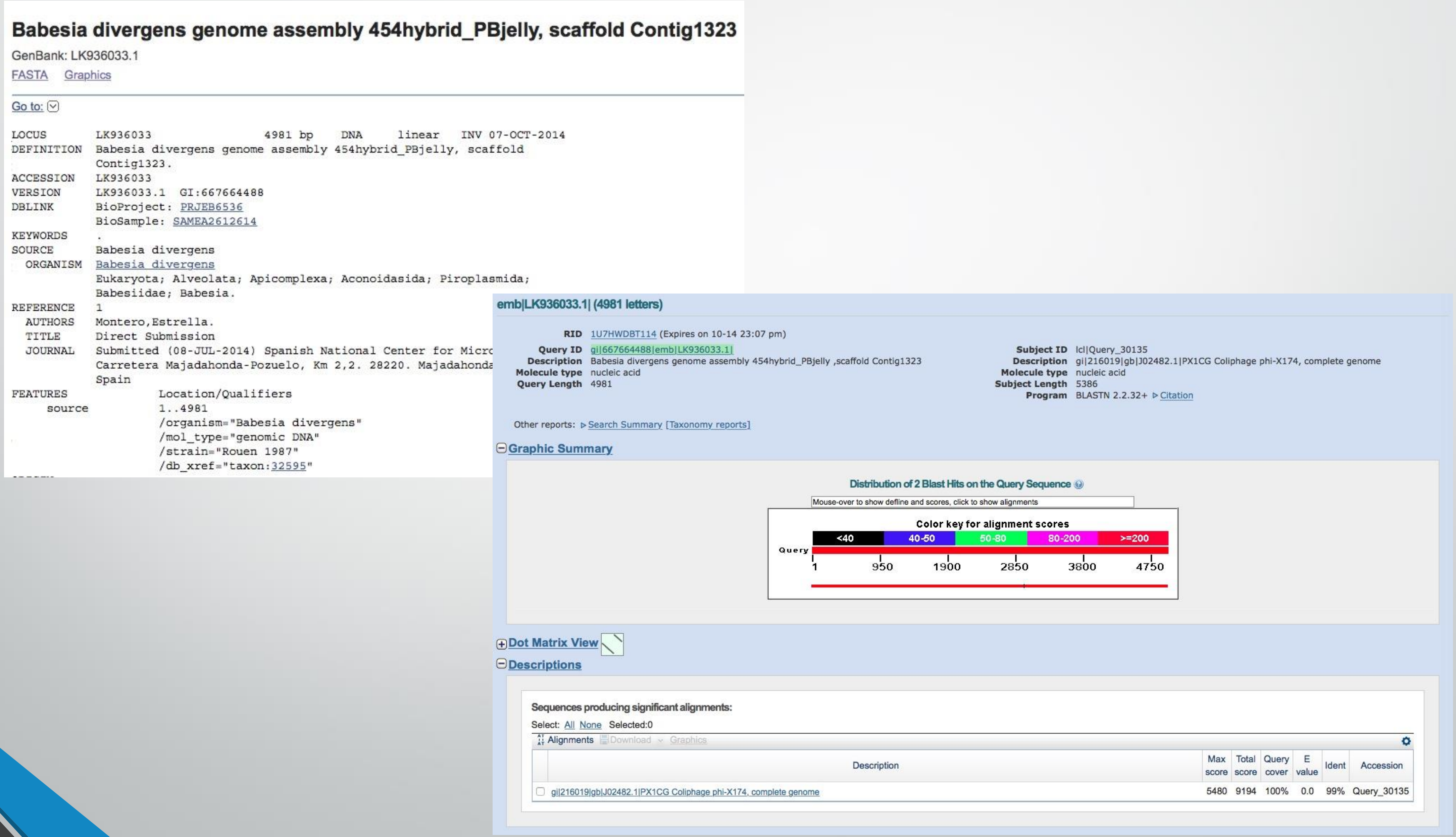


Pre-processing Procedure:

- Low quality region removal
- Low complexity region removal
- Short reads (≤ 50nt) removal

Is it enough?

META**SHOT**



Babesia divergens genome assembly 454hybrid_PBjelly, scaffold Contig1323

GenBank: LK936033.1

FASTA Graphics

Go to:

```
LOCUS       LK936033                4981 bp    DNA     linear   INV 07-OCT-2014
DEFINITION  Babesia divergens genome assembly 454hybrid_PBjelly, scaffold
            Contig1323.
ACCESSION   LK936033
VERSION     LK936033.1  GI:667664488
DBLINK      BioProject: PRJEB6536
            BioSample: SAMEA2612614
KEYWORDS    .
SOURCE      Babesia divergens
  ORGANISM  Babesia divergens
            Eukaryota; Alveolata; Apicomplexa; Aconoidasida; Piroplasmida;
            Babesiidae; Babesia.
REFERENCE   1
  AUTHORS   Montero,Estrella.
  TITLE     Direct Submission
  JOURNAL   Submitted (08-JUL-2014) Spanish National Center for Micro
            Carretera Majadahonda-Pozuelo, Km 2,2. 28220. Majadahonda
            Spain
FEATURES             Location/Qualifiers
     source          1..4981
                     /organism="Babesia divergens"
                     /mol_type="genomic DNA"
                     /strain="Rouen 1987"
                     /db_xref="taxon:32595"
```

emb|LK936033.1| (4981 letters)

RID	1U7HWDBT114 (Expires on 10-14 23:07 pm)		
Query ID	gi|667664488|emb|LK936033.1|	**Subject ID**	lcl|Query_30135
Description	Babesia divergens genome assembly 454hybrid_PBjelly ,scaffold Contig1323	**Description**	gi|216019|gb|J02482.1|PX1CG Coliphage phi-X174, complete genome
Molecule type	nucleic acid	**Molecule type**	nucleic acid
Query Length	4981	**Subject Length**	5386
		Program	BLASTN 2.2.32+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports]

⊖ Graphic Summary

Distribution of 2 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

⊕ Dot Matrix View

⊖ Descriptions

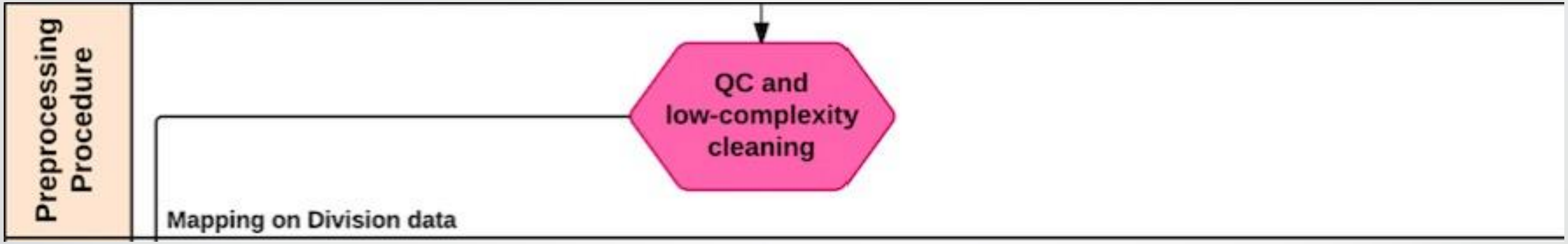
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	gi|216019|gb|J02482.1|PX1CG Coliphage phi-X174, complete genome	5480	9194	100%	0.0	99%	Query_30135

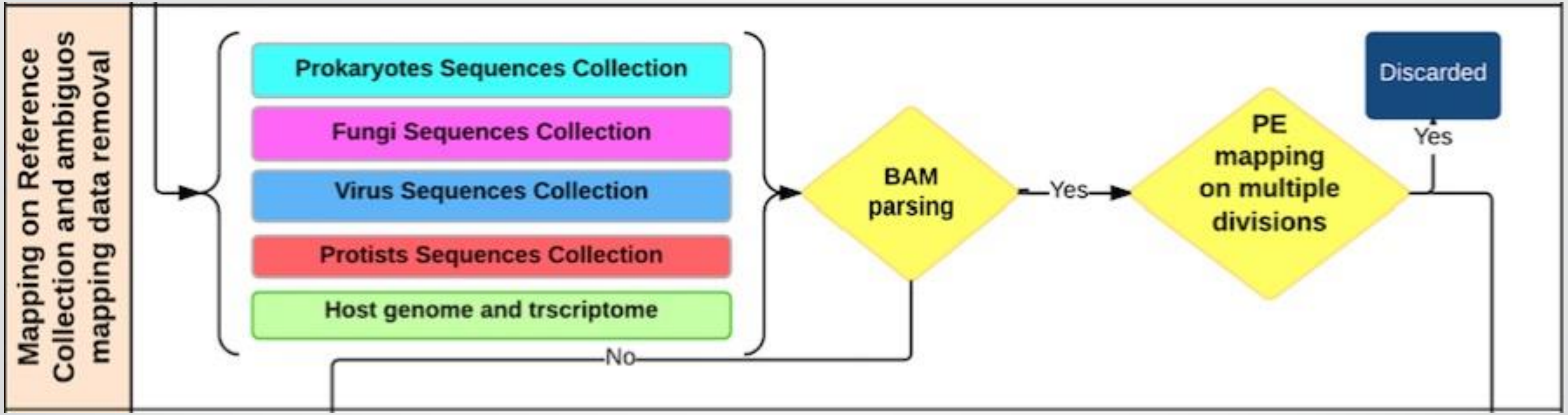
*META**SHOT***



Pre-processing Procedure:

- Low quality region removal
- Low complexity region removal
- Short reads (≤ 50nt) removal
- Phix removal

*META**SHOT***


Mapping on Reference Collection and ambiguos mapping data removal
Prokaryotes Sequences Collection
Fungi Sequences Collection
Virus Sequences Collection
Protists Sequences Collection
Host genome and trscriptome
BAM parsing
Yes
No
PE mapping on multiple divisions
Yes
Discarded

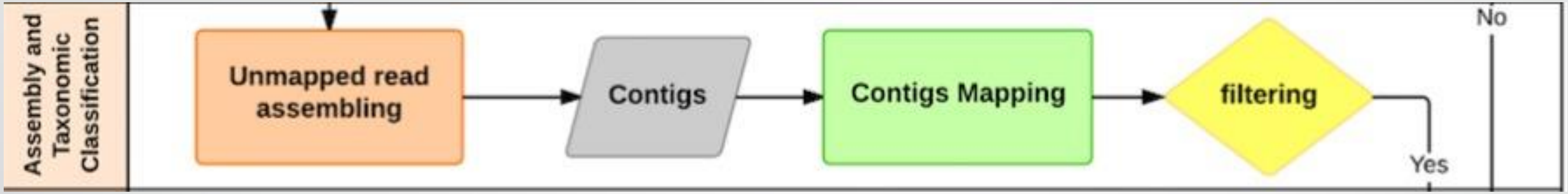
Denoised data are mapped against the reference collections.

Mapping data filtering:

- Query coverage (≥ 70%)
- Identity percentage (≥97%)

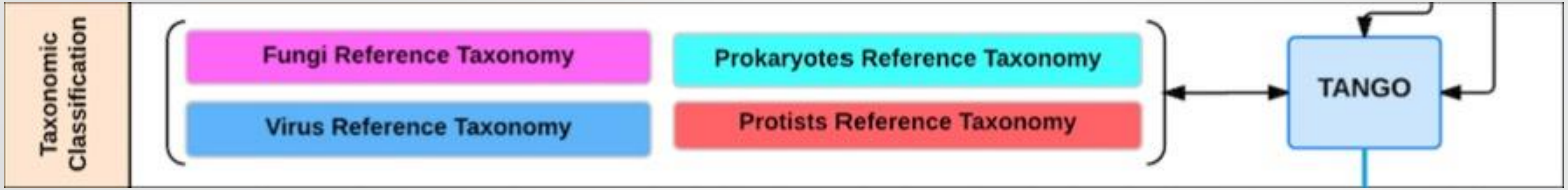
PE reads mapping on two or more divisions are discarded

METASHOT



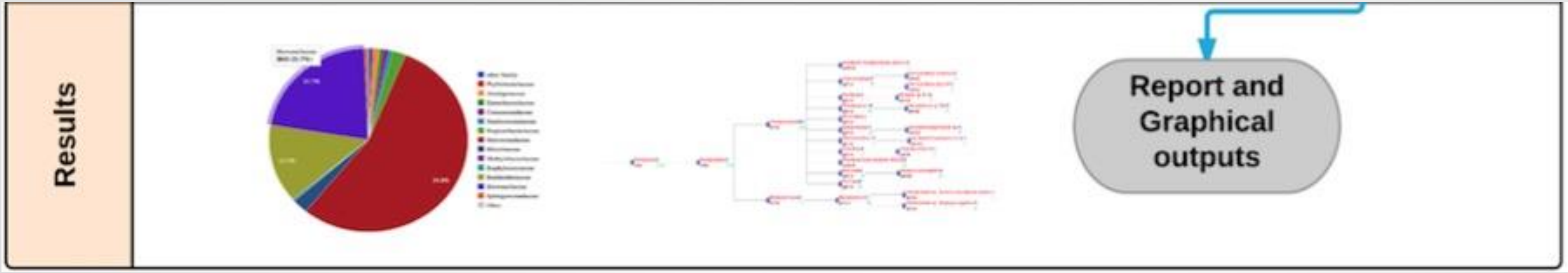
- All the unmapped PE reads are assembled
- Resulting contigs are mapped on reference collections

METASHOT



- PE reads and contigs mapping on only one reference collections are taxonomically classified by using the NCBI taxonomy.
- The taxonomic assignment obtained are stored in a NHX tree

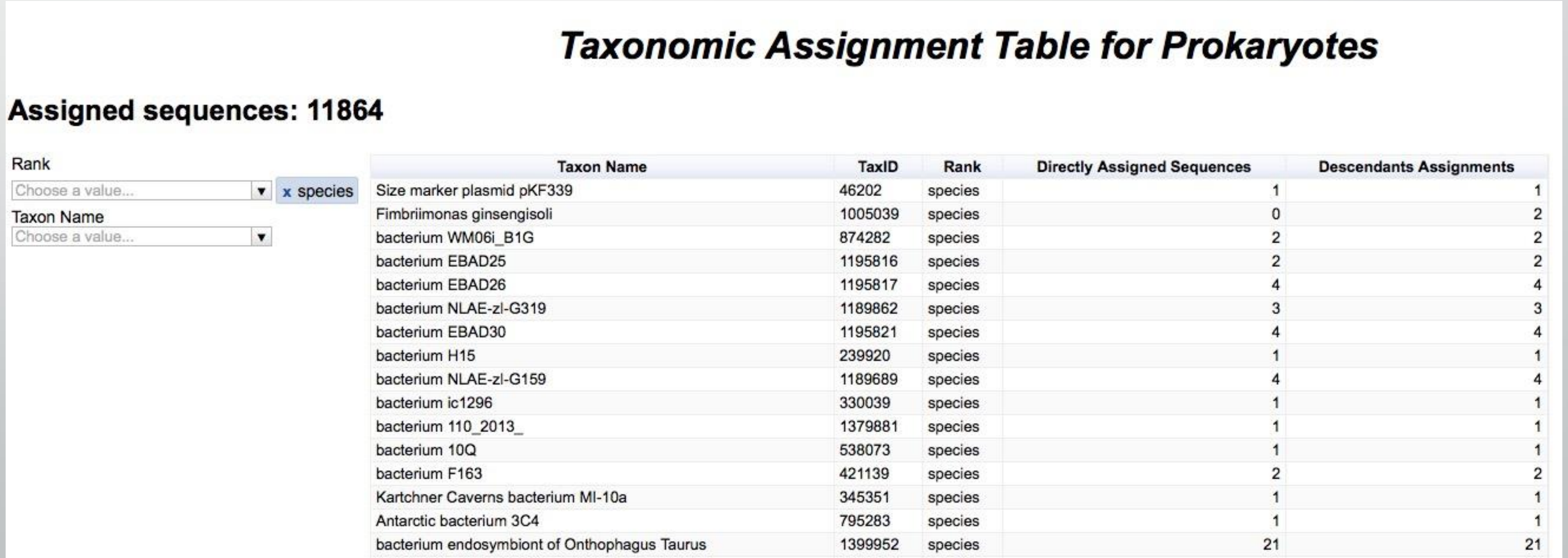
METASHOT



Results

Report and Graphical outputs

- High-resolution tree
- CSV file
- HTML interactive table



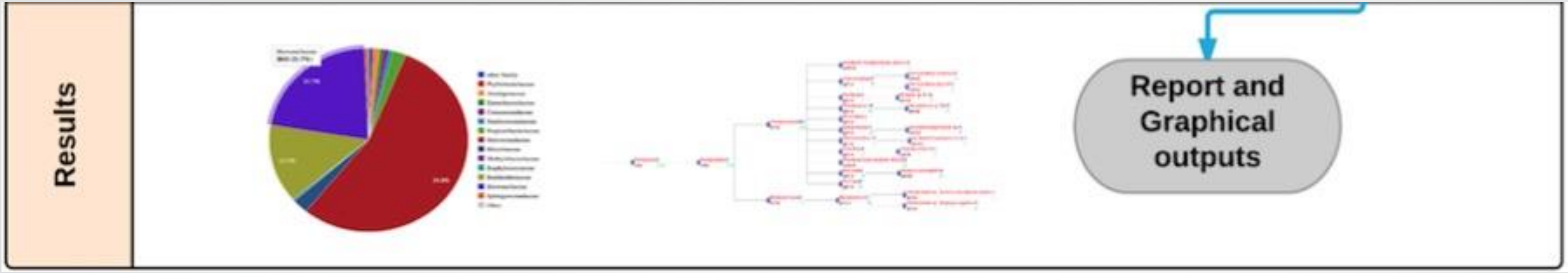
Taxonomic Assignment Table for Prokaryotes

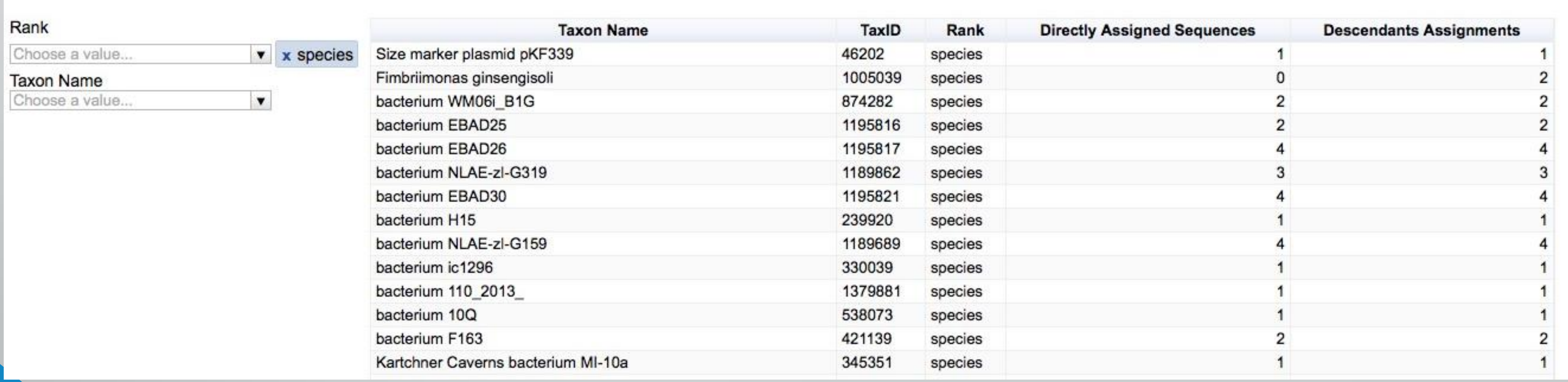
Assigned sequences: 11864

Taxon Name	TaxID	Rank	Directly Assigned Sequences	Descendants Assignments
Size marker plasmid pKF339	46202	species	1	1
Fimbriimonas ginsengisoli	1005039	species	0	2
bacterium WM06i_B1G	874282	species	2	2
bacterium EBAD25	1195816	species	2	2
bacterium EBAD26	1195817	species	4	4
bacterium NLAE-zl-G319	1189862	species	3	3
bacterium EBAD30	1195821	species	4	4
bacterium H15	239920	species	1	1
bacterium NLAE-zl-G159	1189689	species	4	4
bacterium ic1296	330039	species	1	1
bacterium 110_2013_	1379881	species	1	1
bacterium 10Q	538073	species	1	1
bacterium F163	421139	species	2	2
Kartchner Caverns bacterium MI-10a	345351	species	1	1
Antarctic bacterium 3C4	795283	species	1	1
bacterium endosymbiont of Onthophagus Taurus	1399952	species	21	21

NETTAB & IB 2015

METASHOT

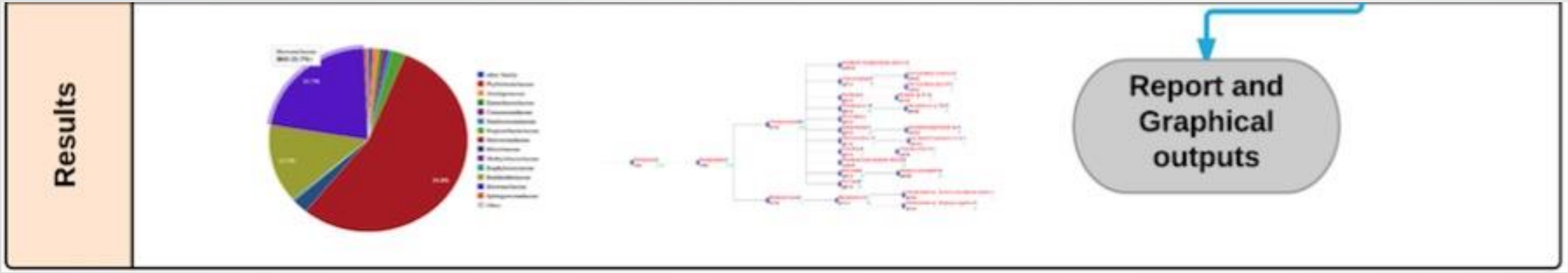




Taxon Name	TaxID	Rank	Directly Assigned Sequences	Descendants Assignments
Size marker plasmid pKF339	46202	species	1	1
Fimbriimonas ginsengisoli	1005039	species	0	2
bacterium WM06i_B1G	874282	species	2	2
bacterium EBAD25	1195816	species	2	2
bacterium EBAD26	1195817	species	4	4
bacterium NLAE-zl-G319	1189862	species	3	3
bacterium EBAD30	1195821	species	4	4
bacterium H15	239920	species	1	1
bacterium NLAE-zl-G159	1189689	species	4	4
bacterium ic1296	330039	species	1	1
bacterium 110_2013_	1379881	species	1	1
bacterium 10Q	538073	species	1	1
bacterium F163	421139	species	2	2
Kartchner Caverns bacterium MI-10a	345351	species	1	1

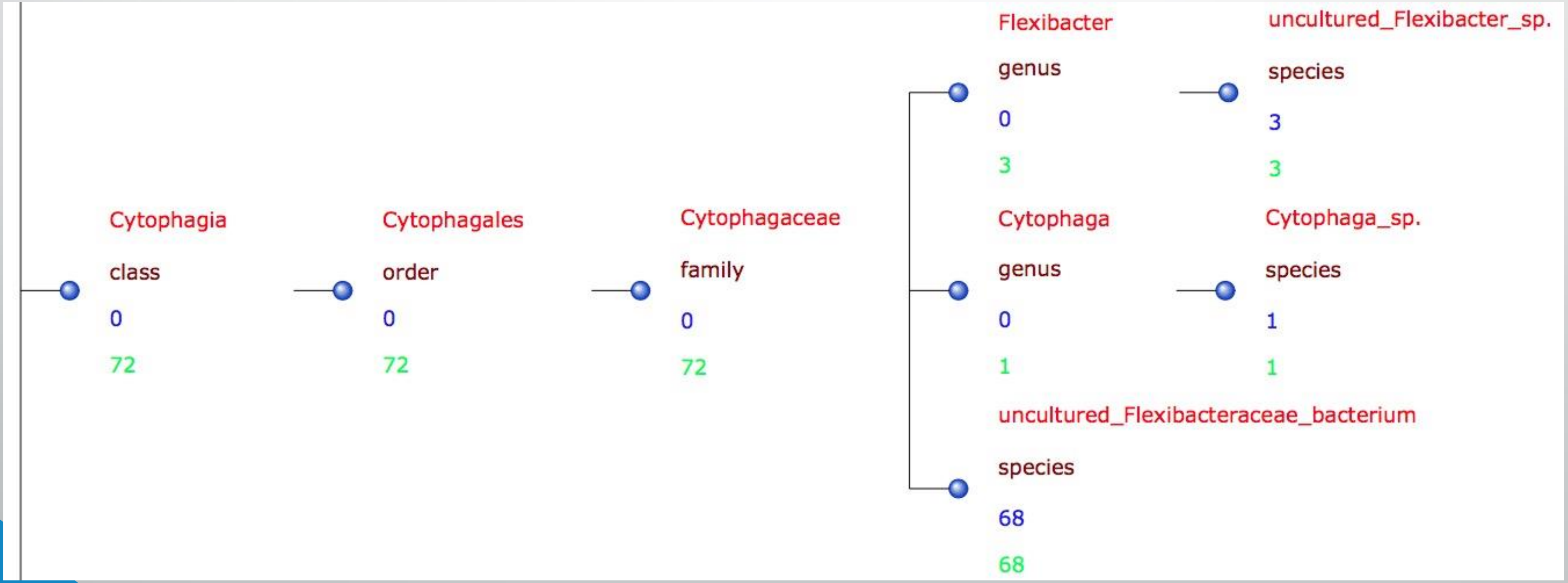
METASHOT

NETTAB & IB 2015



Results

Report and Graphical outputs



NETTAB & IB 2015

CASE STUDY

METASHOT have been applied to the investigation of a uterine cervix sample.

Pre-Precessing (PP) data

Sample	PE reads	PP Pass	% PP Pass
DNA	528,034,456	512,253,714	97.01%
RNA	61,318,866	59,303,563	96.71%

Taxonomic Analysis data

Sample	Human	Prokaryotes	Virus	Fungi	Protists
DNA	501,609,424	2,541	25,211	91	71
RNA	52,312,428	7,200	14,253	41	14

CASE STUDY

Both for DNA and RNA data about the 98% of viral assignments regards the HPV serotype 31

↓

The presence of the HPV (Human Papilloma Virus) serotype 31 has been confirmed by PCR analysis.

The same data have been analysed by using Kraken.
It was unable to identify the presence of HPV serotype 31.

CONCLUTIONS

- ***MetaShot*** is an effective pipeline for the characterization of host-associated microbiome
- It performs all the required steps for NGS data taxonomic analysis
- It will be released as a stand-alone package embedded in a Ubuntu-based virtual machine

Acknowledgments

IBBE-CNR and University of Bari, Bari, Italy

M Santamaria

D Lovero

G Pesole

Department of Oncological Surgery, Gynecologic Oncology Unit, "Regina Elena" National Cancer Institute, Roma, Italy

G Corrado

E Vizza

Dept. of Cell Biology and Neurosciences, Italian National Institute of Health, Rome, Italy

N Passaro

M Crescenzi

KEEP
CALM
AND
THANKS FOR
YOUR ATTENTION