Novel bioinformatics tools to assess microbial diversity in life support systems

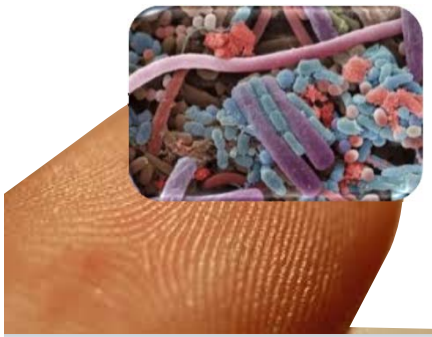
Mohamed Mysara, Natalie Leys, Pieter Monsieurs

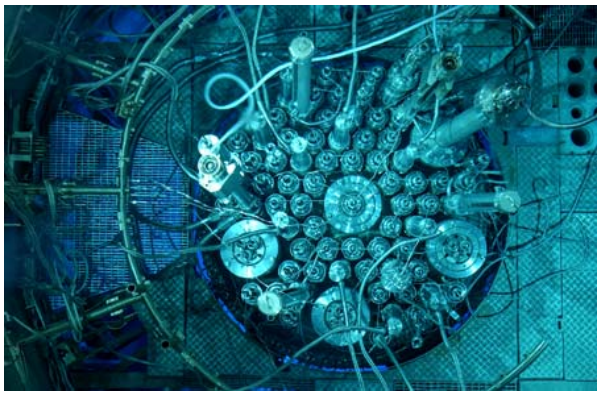
mahemd@sckcen.be; pmonsieu@sckcen.be

SCK•CEN
STUDIECENTRUM VOOR KERNENERGIE
CENTRE D'ETUDE DE L'ENERGIE NUCLEAIRE

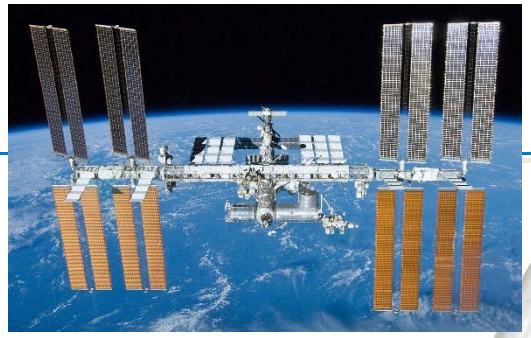
Bacterial communities

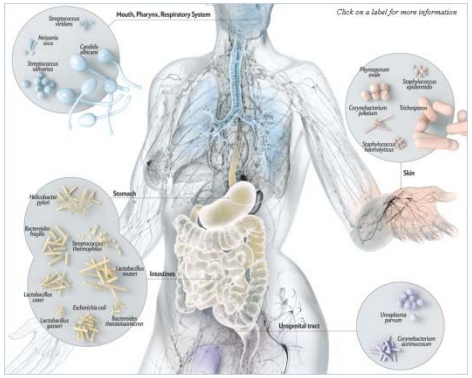
- Bacterial species live in communities, rather than individual species
- They interaction, depend and talk to each others
- These dynamic communities referred to as **microbiome**

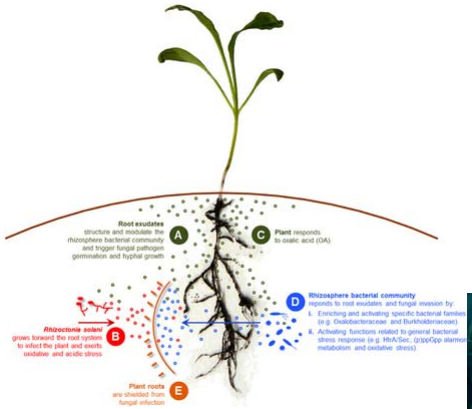


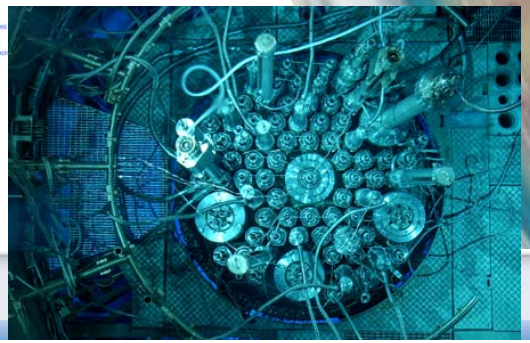


Microbiome studies





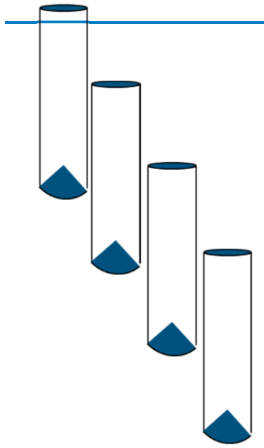




However, culture based approach exhibit various disadvantages:

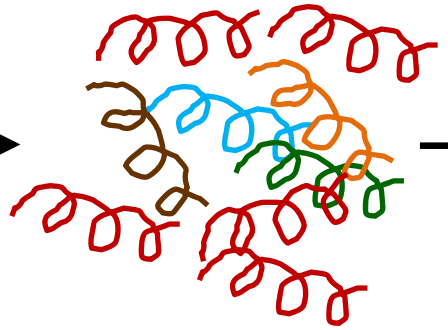
- Time and labor consuming
- Most of the species can not be cultured in lab conditions

From sample to data (16S rRNA gene sequencing)



Extract

DNA



Amplify one

common gene

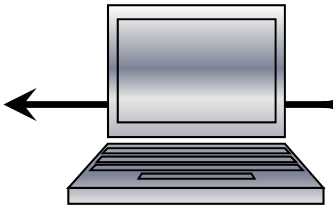
"PCR"

Illumina Miseq

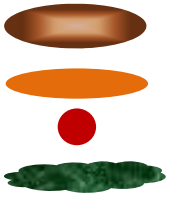
Determine DNA sequence

atcccgtactaccg
atcccgtagtaccg
atcccgtggtaccg
atcccgttataccg
atcccgtcctaccg
atcccgtactaccg
atcccgtagtaccg
atcccgtggtaccg
atcccgttataccg
atcccgtcctaccg

Thousands – millions of sequences



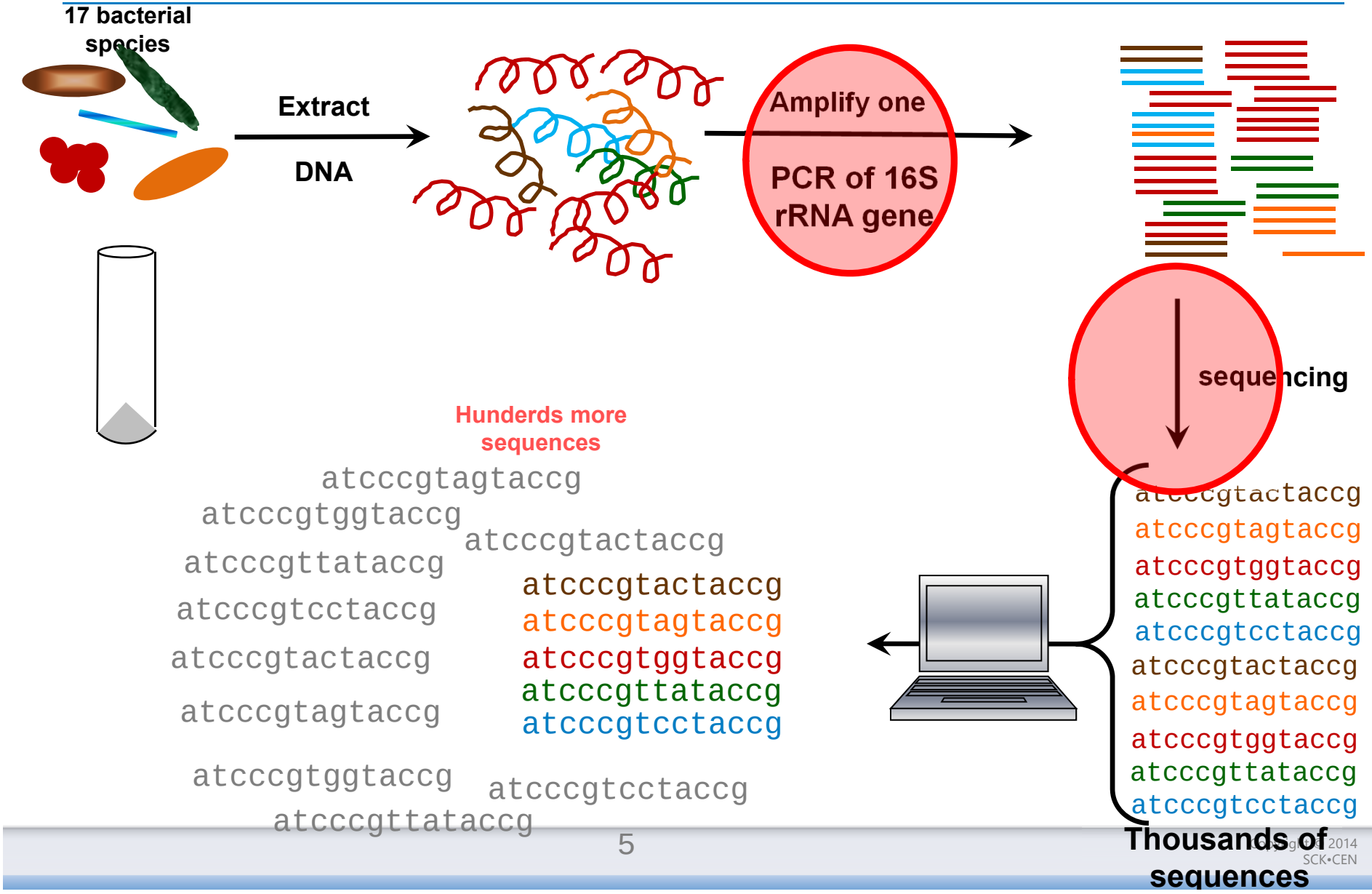
Pseudomonas =
Cupriavidus =
Acidovorax =
Arthrosipra =
Gracilibacter =



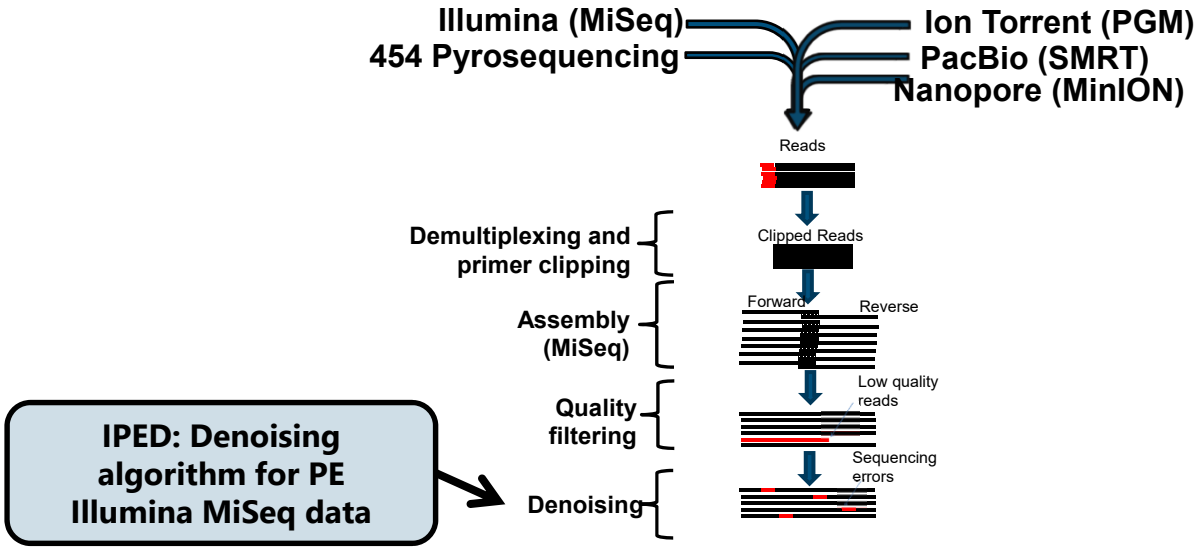
atcccgtactaccg
atcccgtagtaccg
atcccgtggtaccg
atcccgttataccg
atcccgtcctaccg

Each sequence is mapped to a species

Microbiome & 16S rRNA metagenomics



16S rRNA Metagenomics Analysis Pipeline



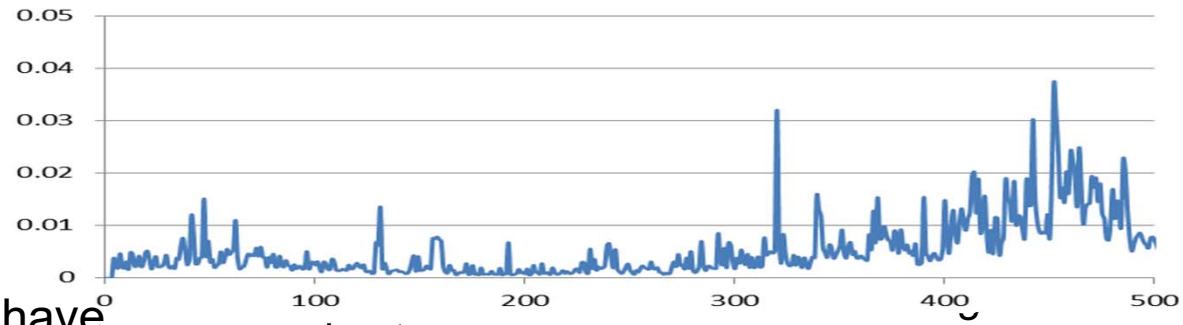
Chapter 2 & 3: Training artificial intelligent model to handle this problem

I) Selection of datasets where the ground truth is known (Mock dataset)

What does the machine report	A	G	-	A	G	T	C	G	A	T
What should the machine report	A	G	**C**	A	G	**G**	C	**-**	A	T
Status	**T**	**T**	**D**	**T**	**T**	**S**	**T**	**I**	**T**	**T**

II) Identification of features contributing to the sequencing errors

e.g. Position in the read



For this purpose, we have developed two artificial intelligence tools:

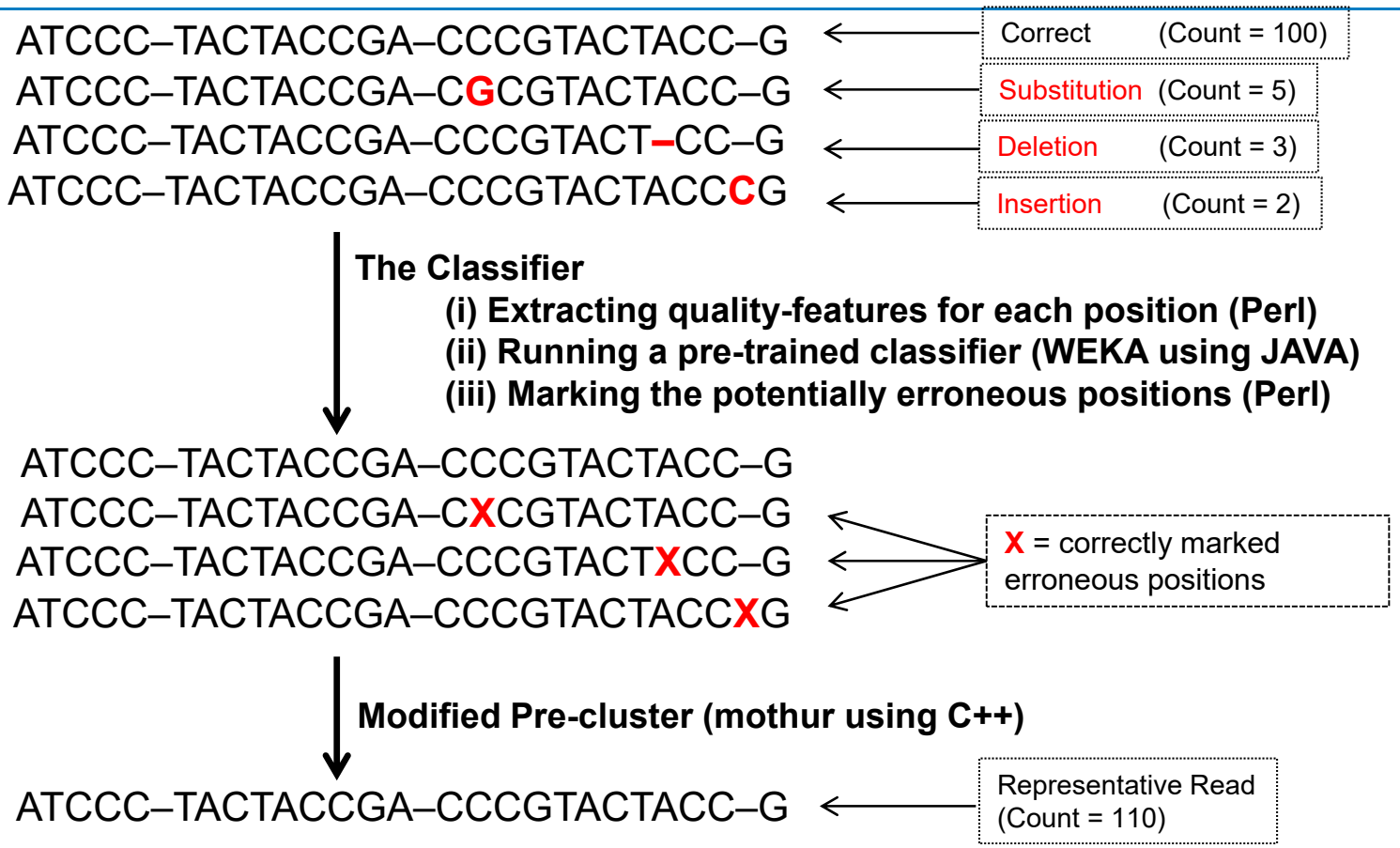
A) NoDe for 454

B) IPED for MiSeq

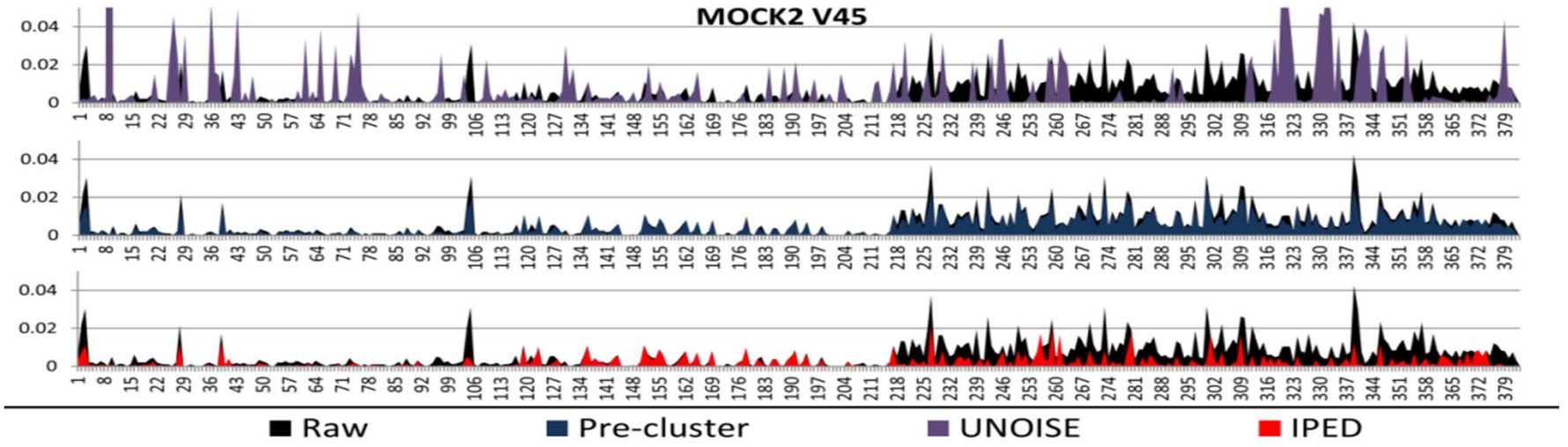
subset → Training → Testing → subset

IV) Selection of the best performing model

Denoising algorithms concept

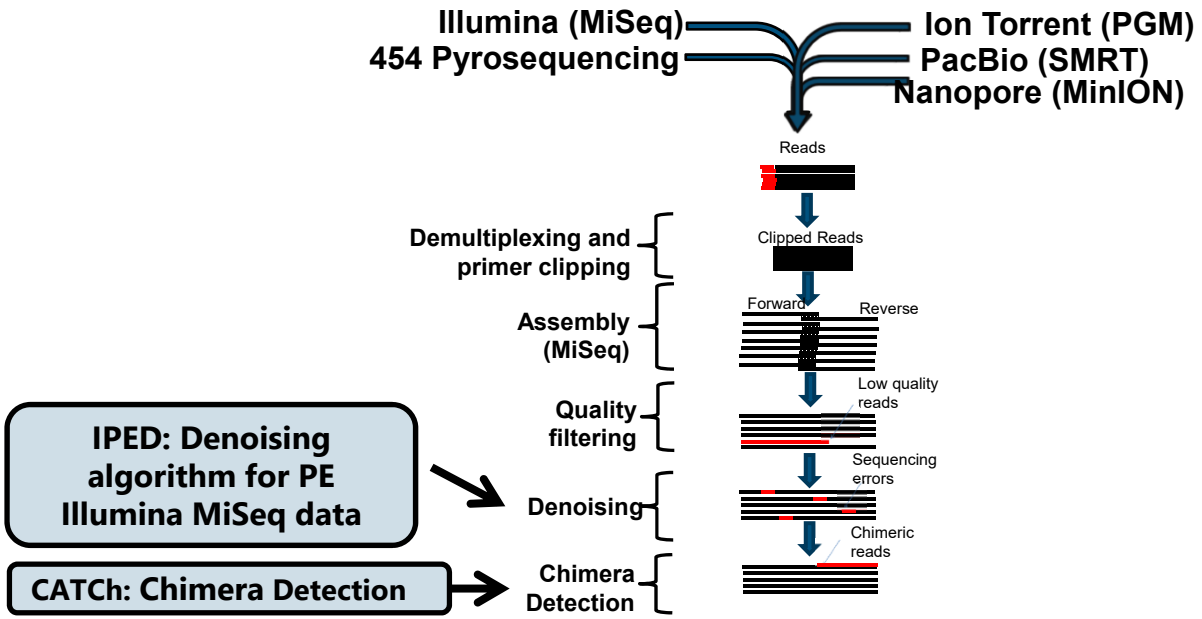


Chapter 3: IPED comparative analysis

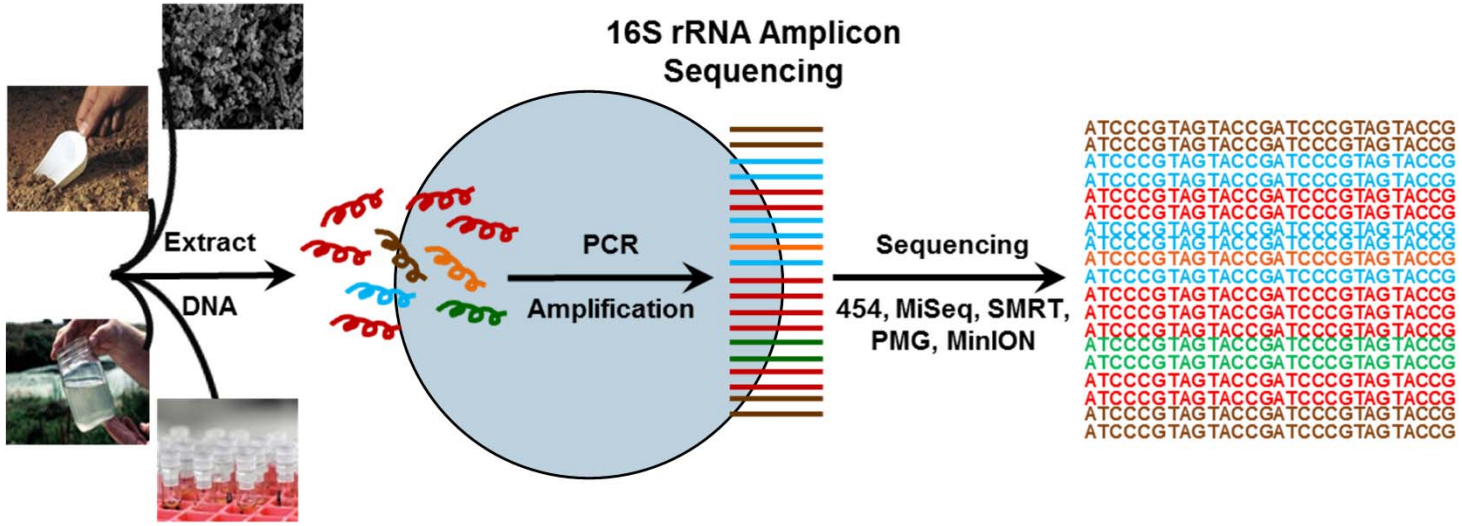


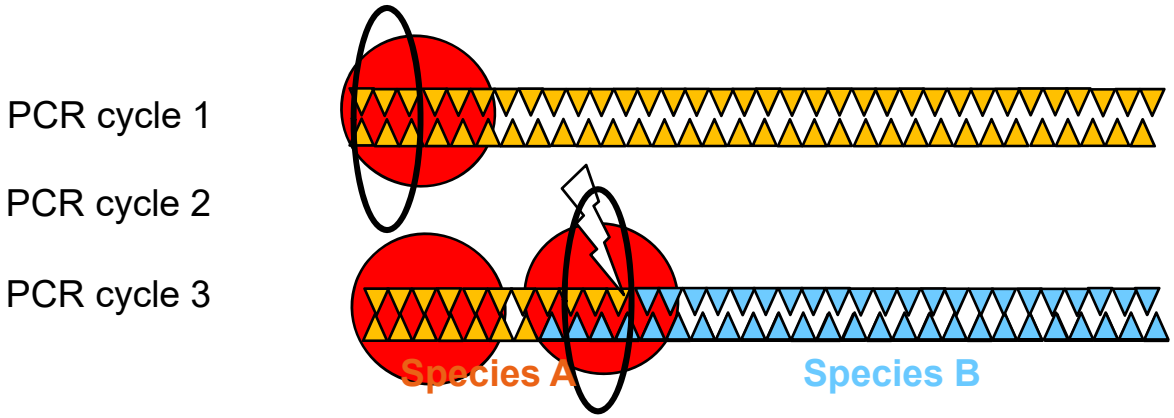
	Error	CPU Cost
UNOISE	0.18%	14 sec
Pre-cluster	0.18%	12 sec
IPED	**0.10%**	**70 sec**

16S rRNA Metagenomics Analysis Pipeline

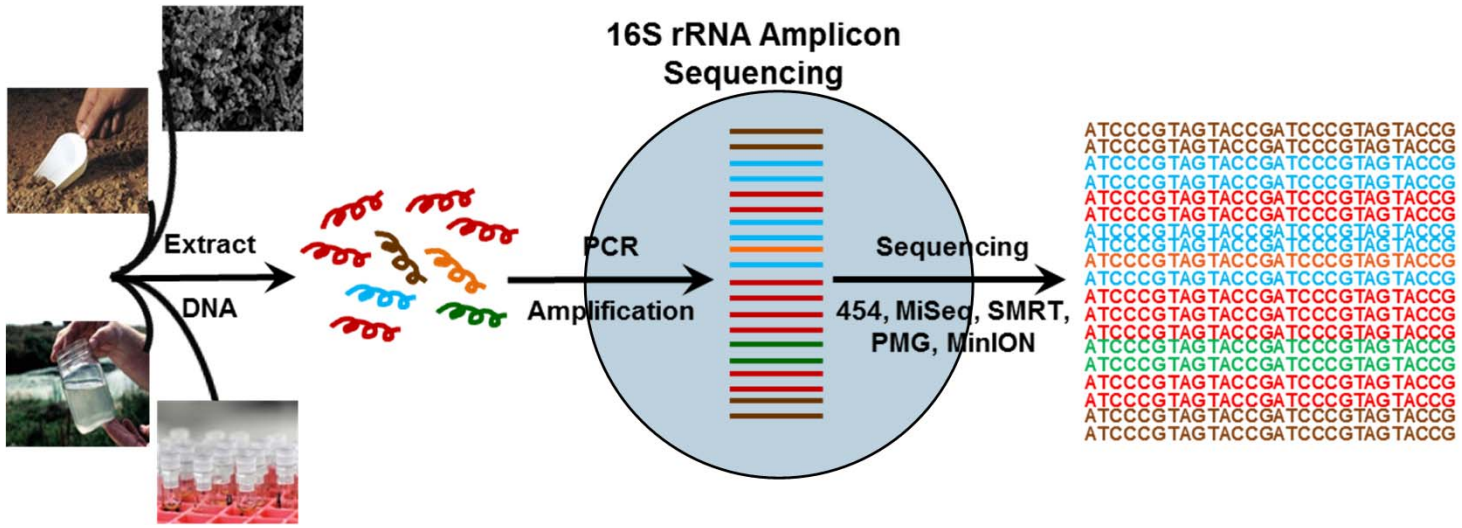


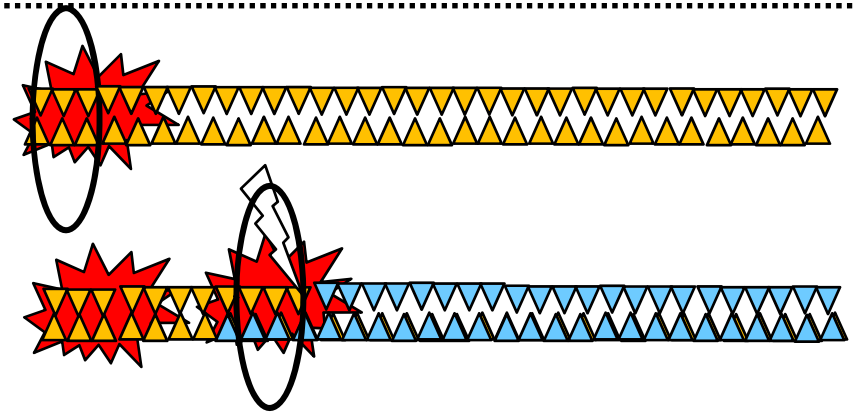
Chimeric problem





Chimeric problem





Sequence1
atcccgtactaccg

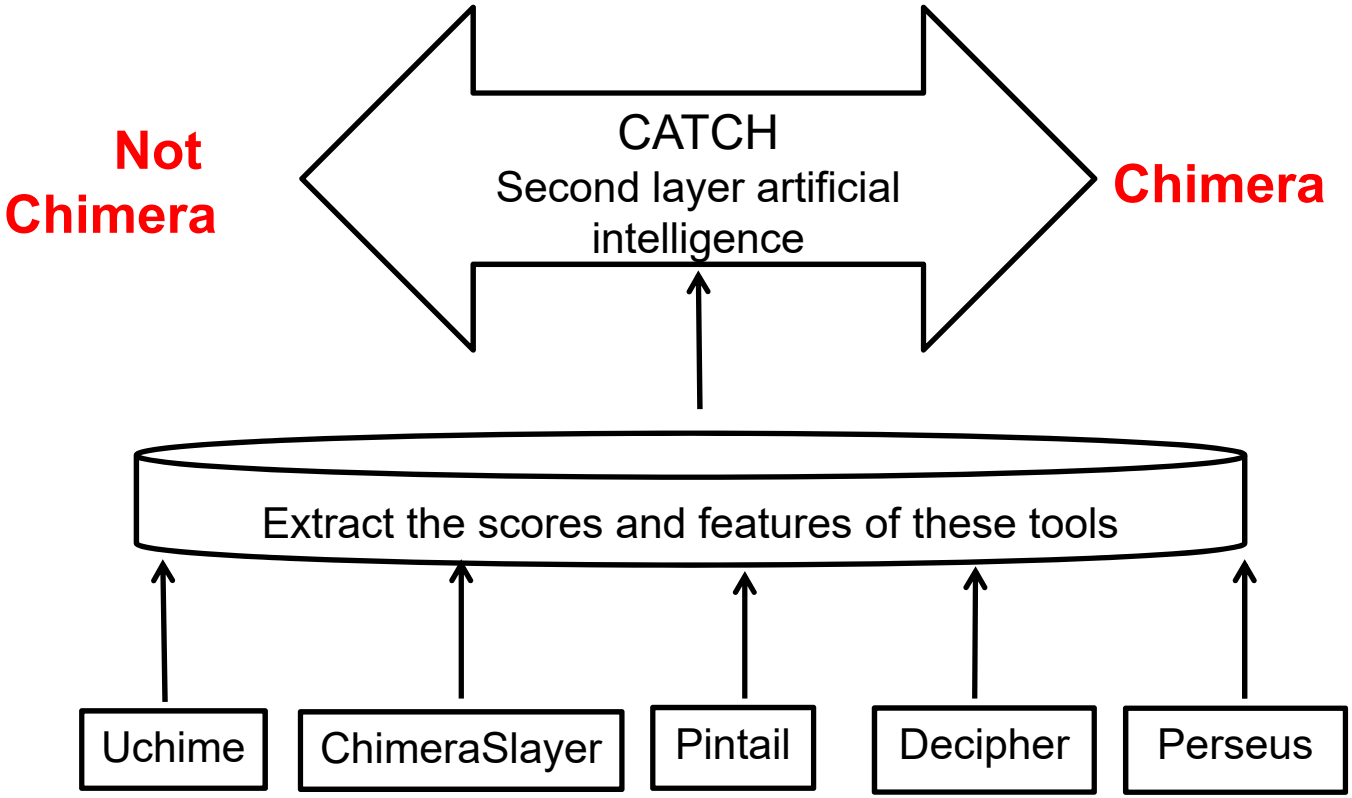
Sequence 2
tagctacgtacgat

Chimera

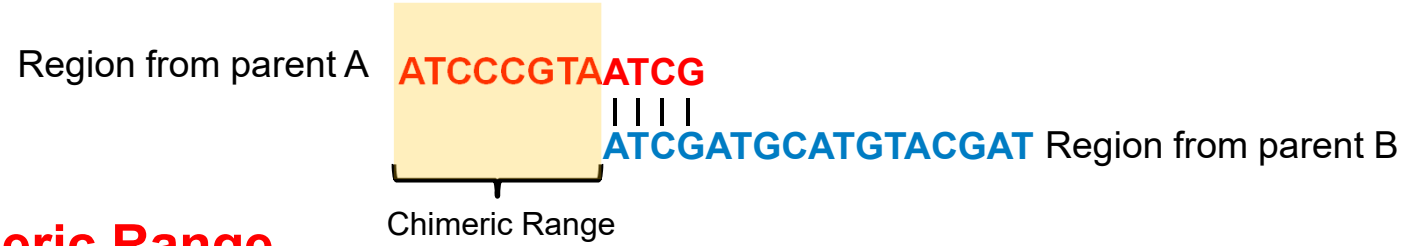
This hybrid sequence is not real, it could be mistaken as a **false** **NOVEL** species

Chimeric rate in Next generation sequencing run can reach up to 45% of the reads.

CATCh Training/Running

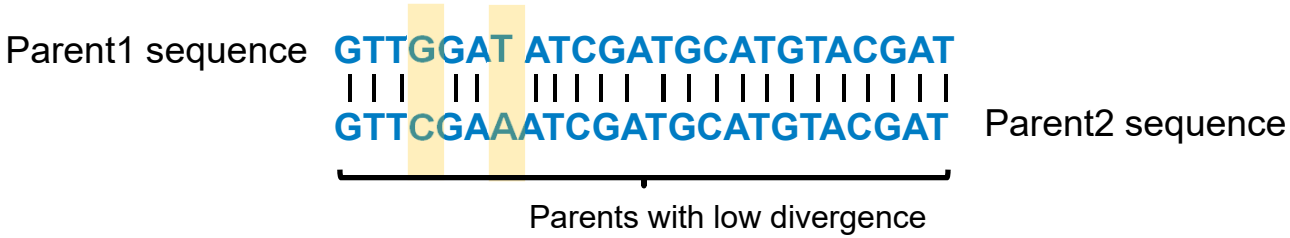


Chapter 4: Chimera Detection Challenges



Chimeric Range

Length added by the smaller parent



Divergence

A measure of the differences between parents

Bimera ATCCCGTAATGCATGTACGAT

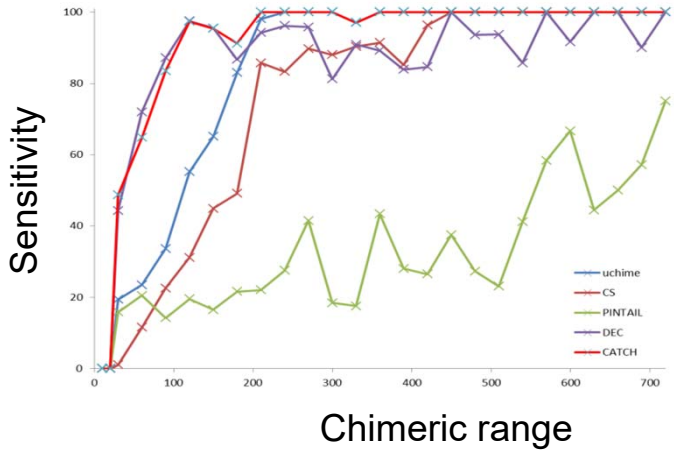
Trimera ATCCCGTA ATGCATGTTCTAGCTAGC

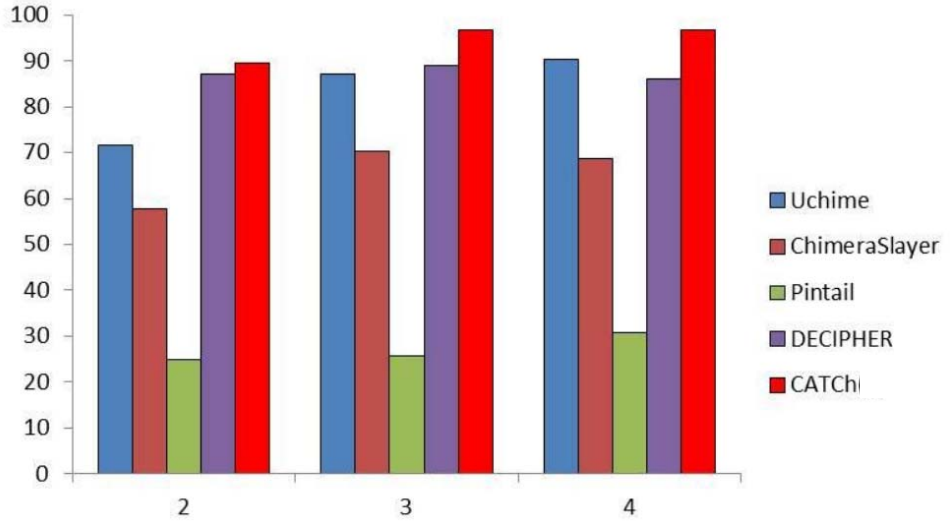
Tetramera ATCCGTA ATGCATCTAGCTAGCATGCAT

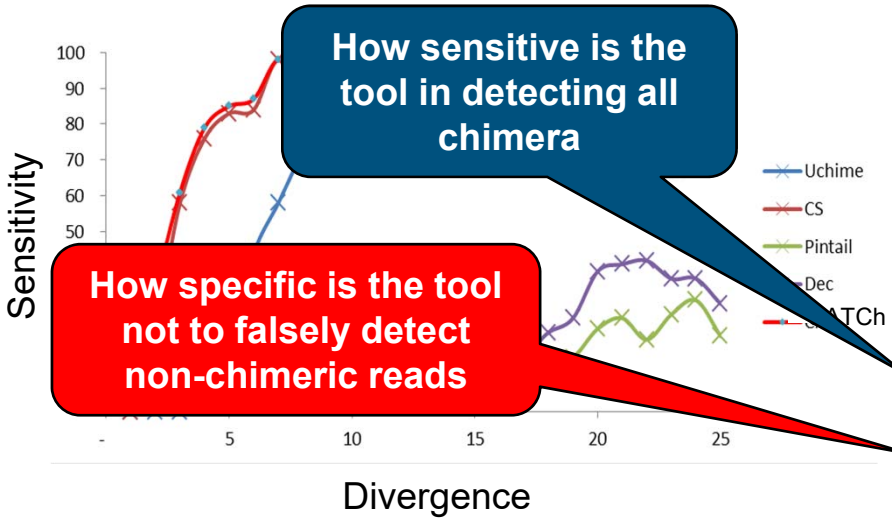
Number of parents

Number of parent read forming the chimeras

CATCh Comparative Analysis

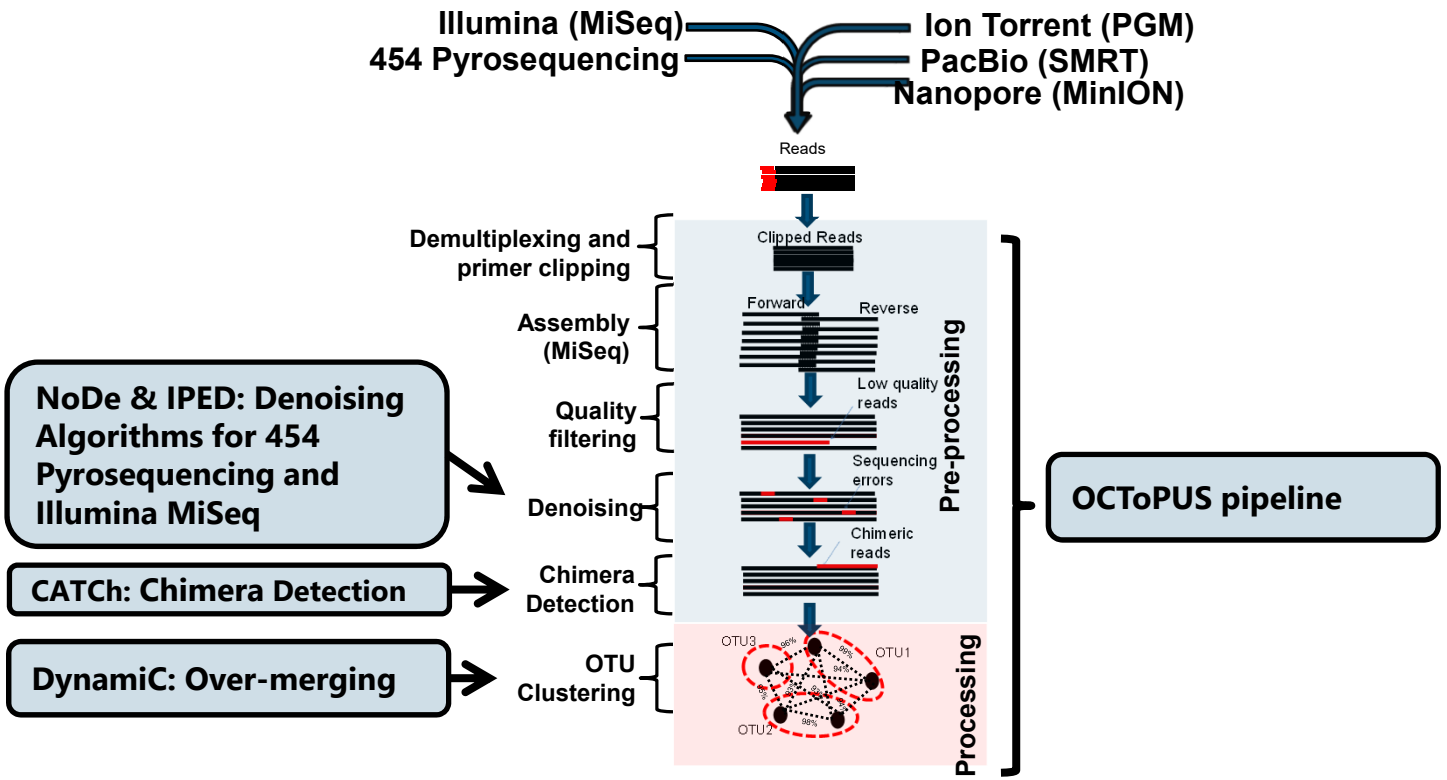




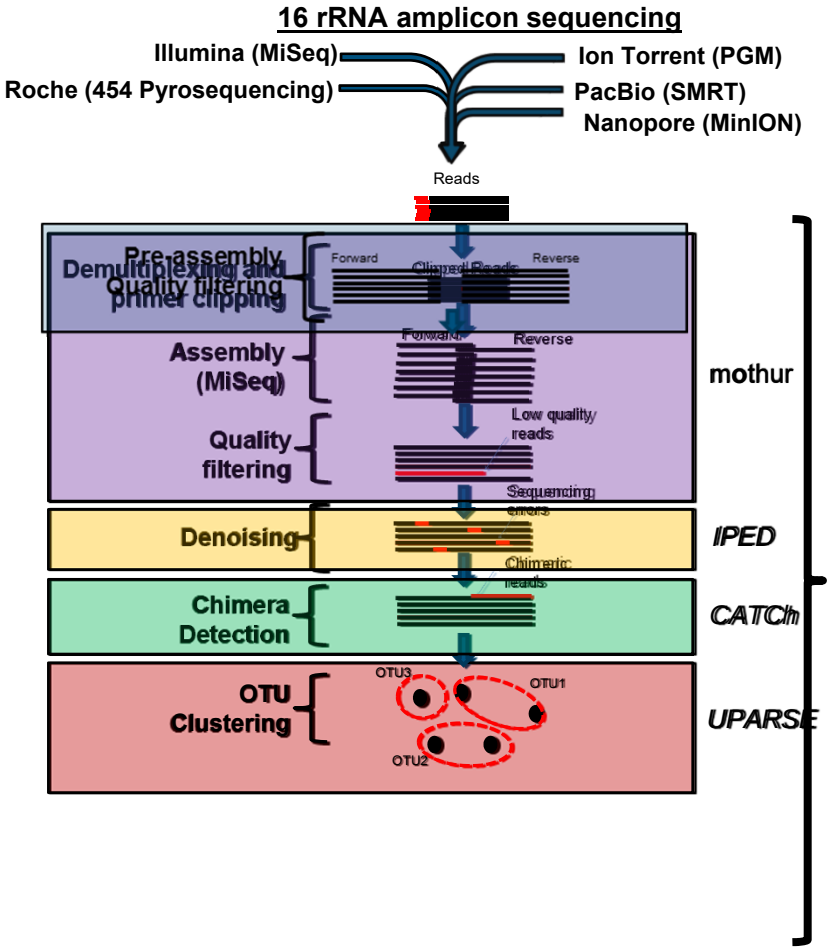


	Reference tools					*De novo* tools			
Evaluation	UCHIME	ChimeraSlayer	Pintail	DECIPHER	**CATCh**	UCHIME	ChimeraSlayer	Perseus	**CATCh**
Sensitivity	78	67	29	57	**85**	60	53	62	**70**
Specificity	97	98	75	97	**96**	97	96	96	**95**

16S rRNA Metagenomics Analysis Pipeline

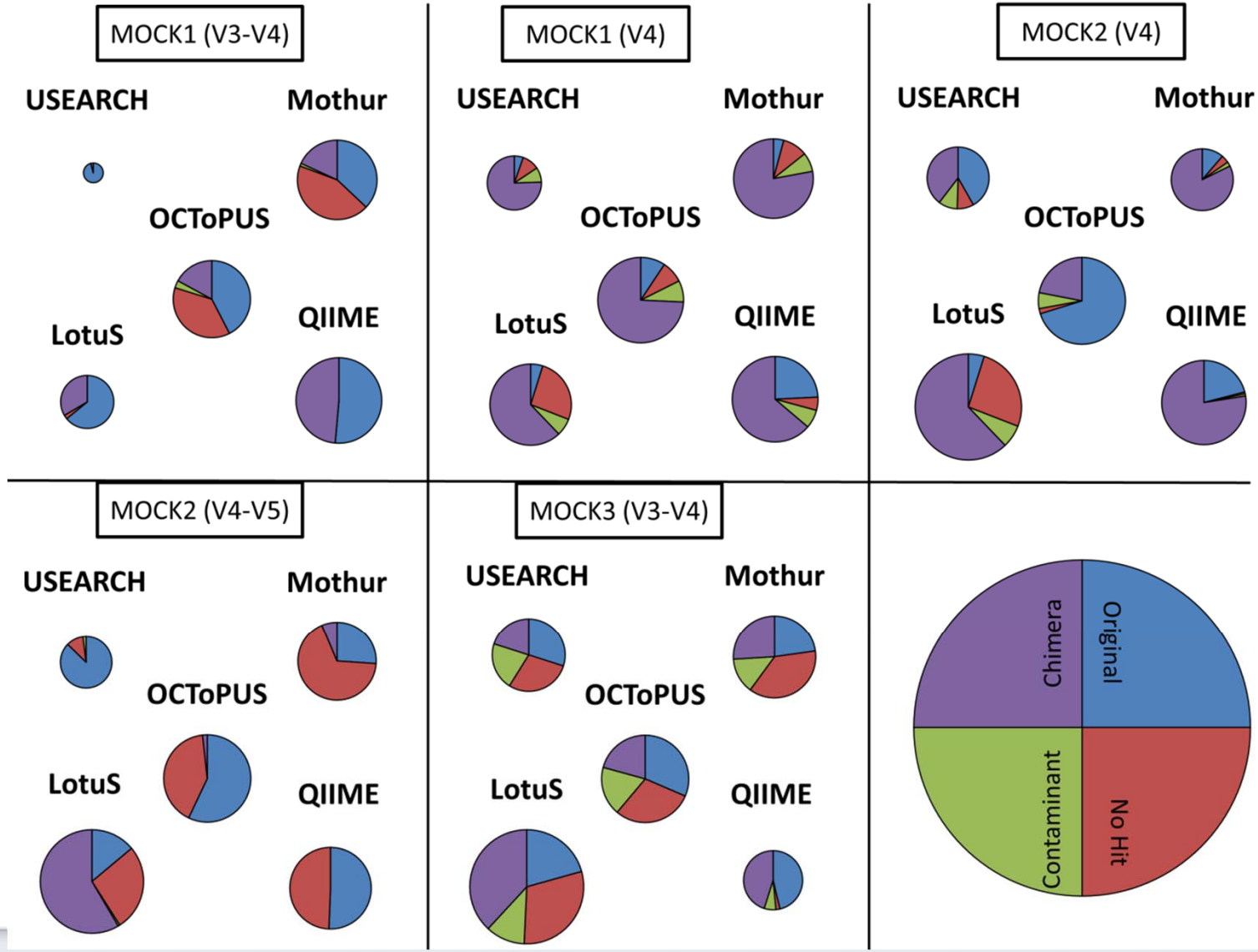


16S rRNA Metagenomics Analysis Pipeline

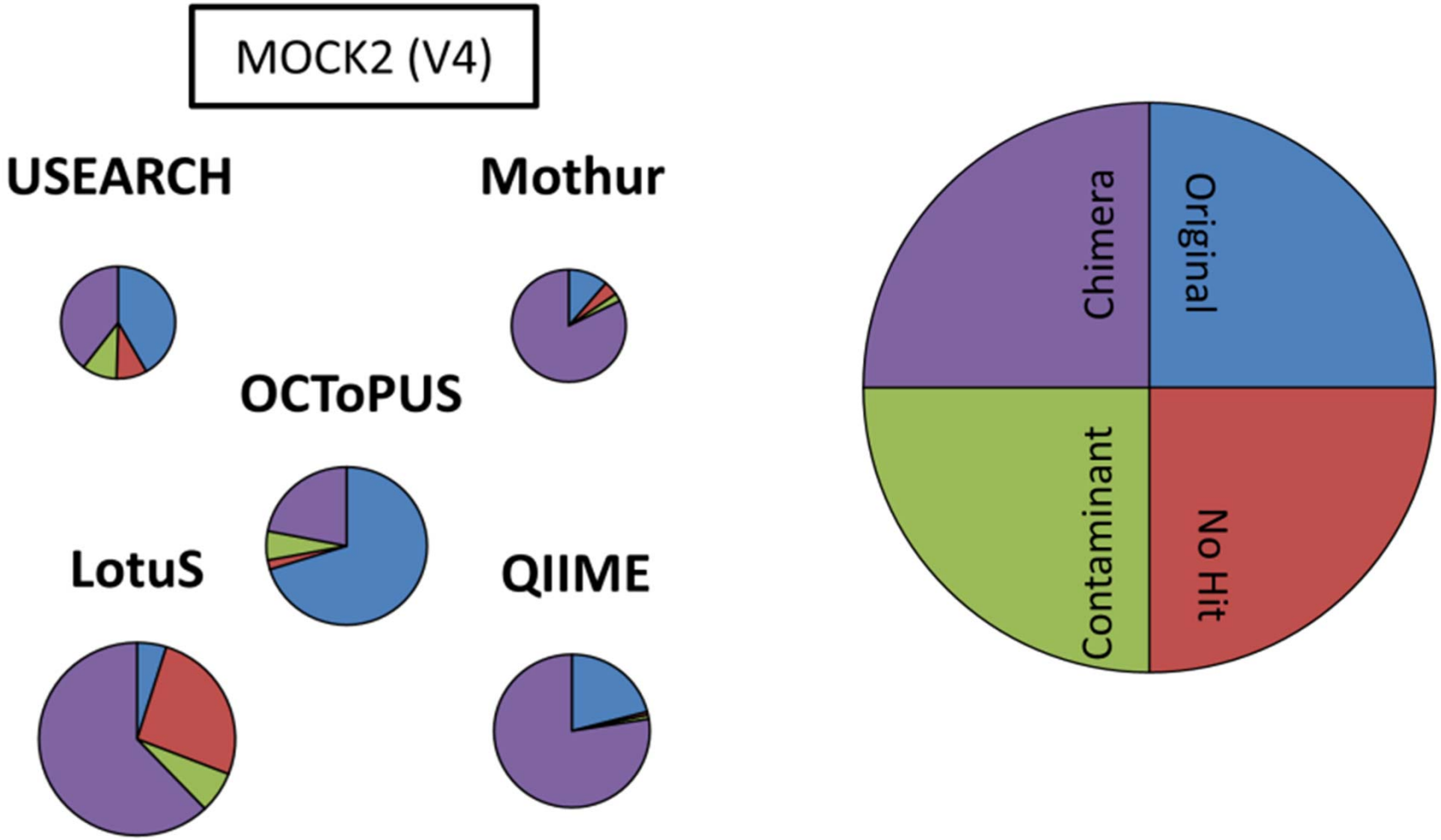


***O**ptimized **C**AT**C**h, m**o**thur, I**P**ED, **U**PARSE and **S**PAdes (**OCToPUS**)*

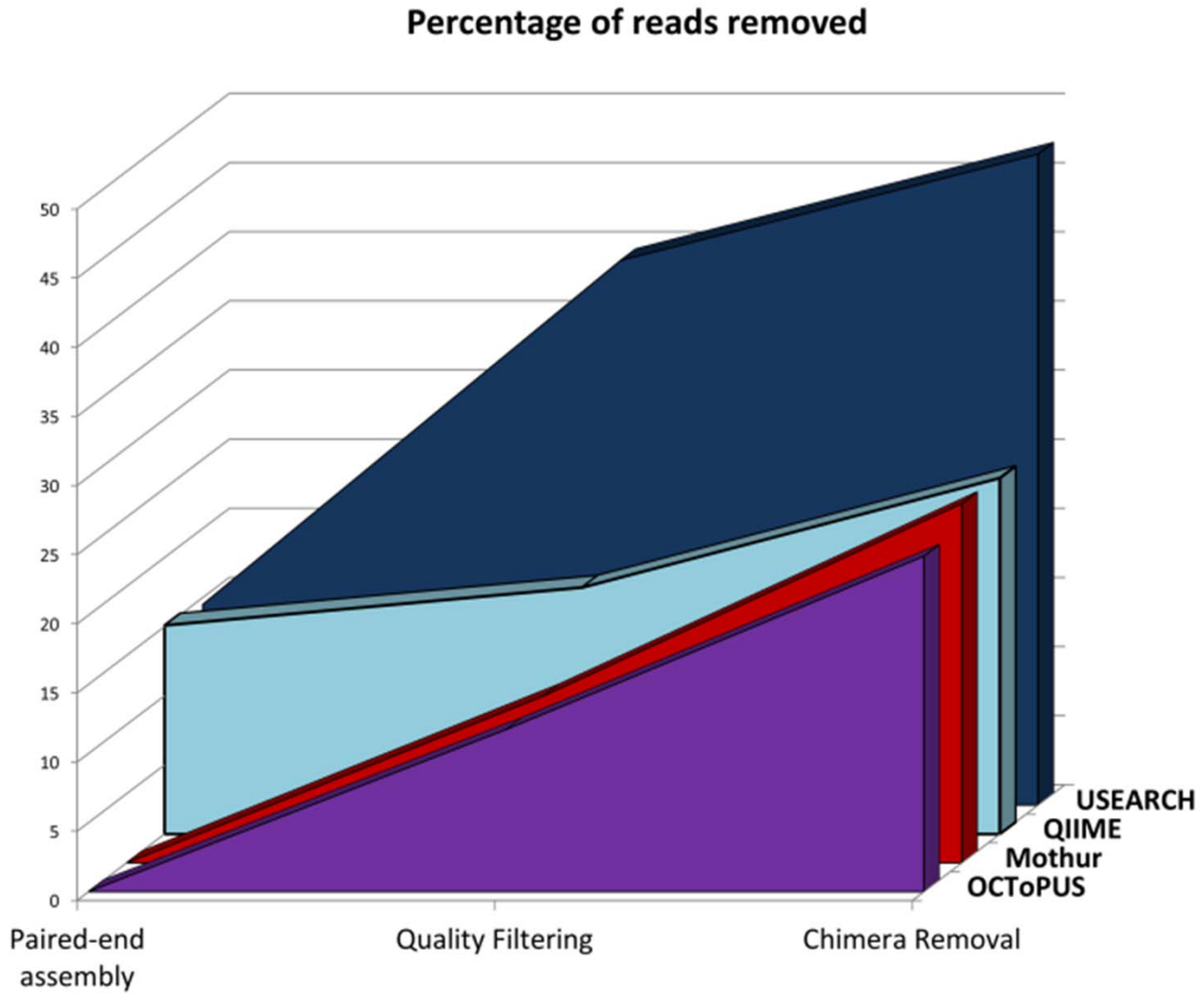
Benchmark - accuarcy



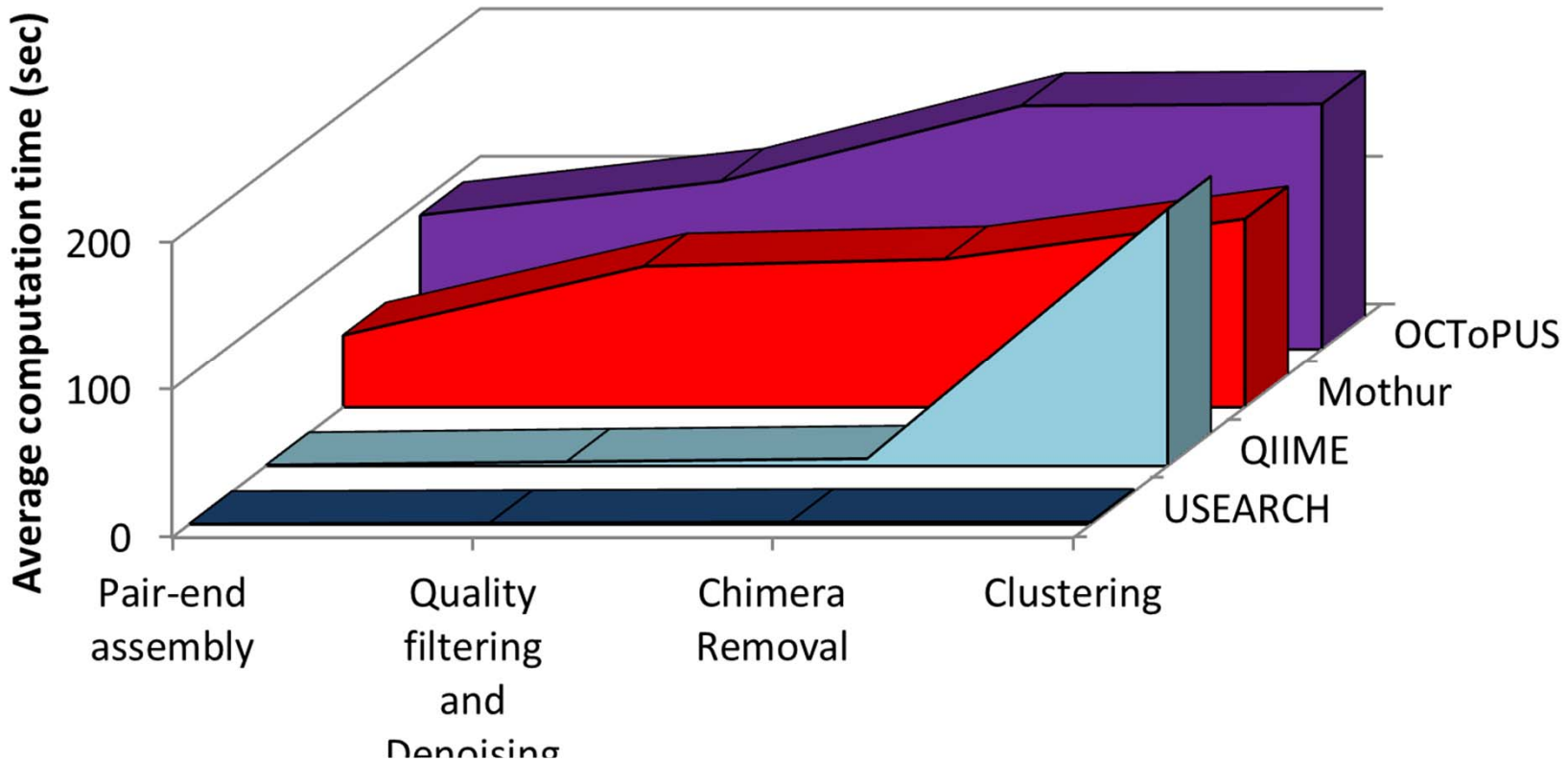
Benchmark



Benchmark – data retrieval



Benchmark – Computational cost



de range of applications

Investigations on radio-protective properties of edible cyanobacterium in a mouse model

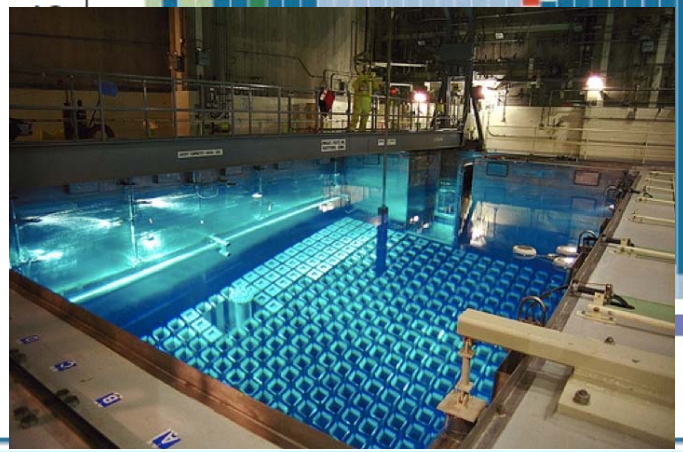
Wannes Van Beeck [1,2], Mieke Verslegers[1], Sarah Baatout[1], Mohamed Mysara[1], Pieter Monsieurs[1], Natalie Leys[1], Sarah Lebeer[2], Felice Mastroleo[1]
[1] Belgian Nuclear Research Centre, SCK•CEN, Mol, Belgium - [2]University of Antwerp, Antwerp, Belgium
E-mail: wannes.vanbeeck@uantwerpen.be ; felice.mastroleo@sckcen.be

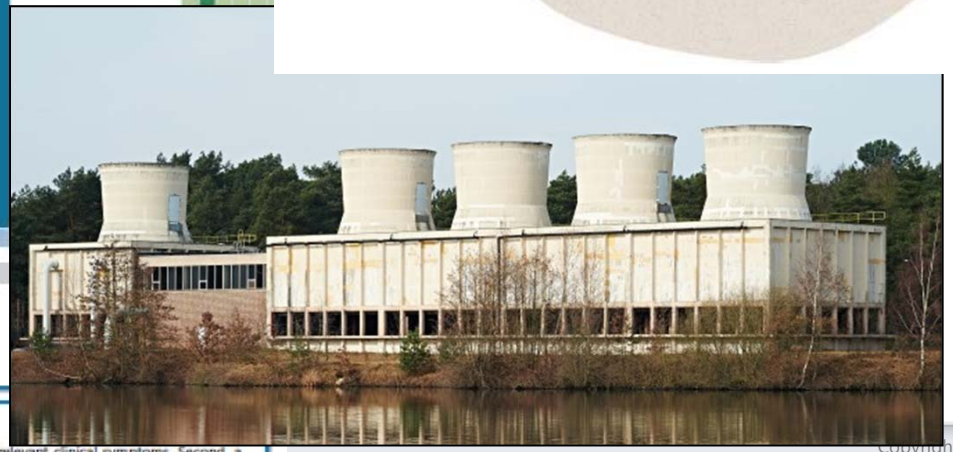
Introduction

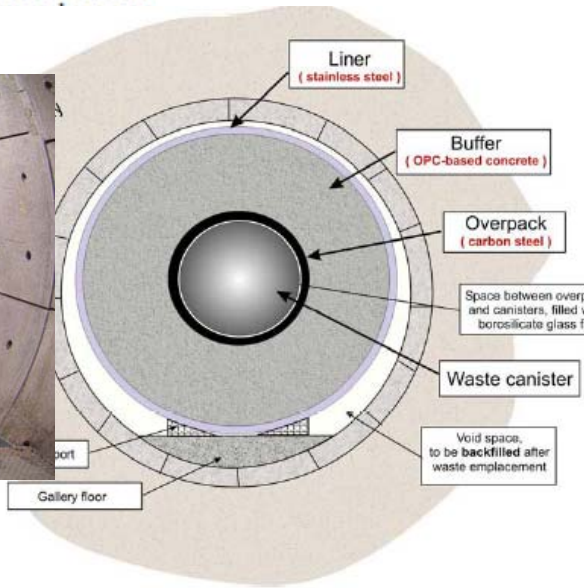
Pelvic radiotherapy is a commonly used treatment to treat specific types of cancer (e.g. colon cancer). After exposure to ionizing radiation, the intestine is always affected. The intestinal epithelium is very sensitive to the increased oxidative stress after exposure to ionizing radiation (Riley, 1994), resulting in a loss

Chitinophagaceae
Comamonadaceae
Chitinophagaceae

Acidobacteria subdivision 3 (incertae sedis)



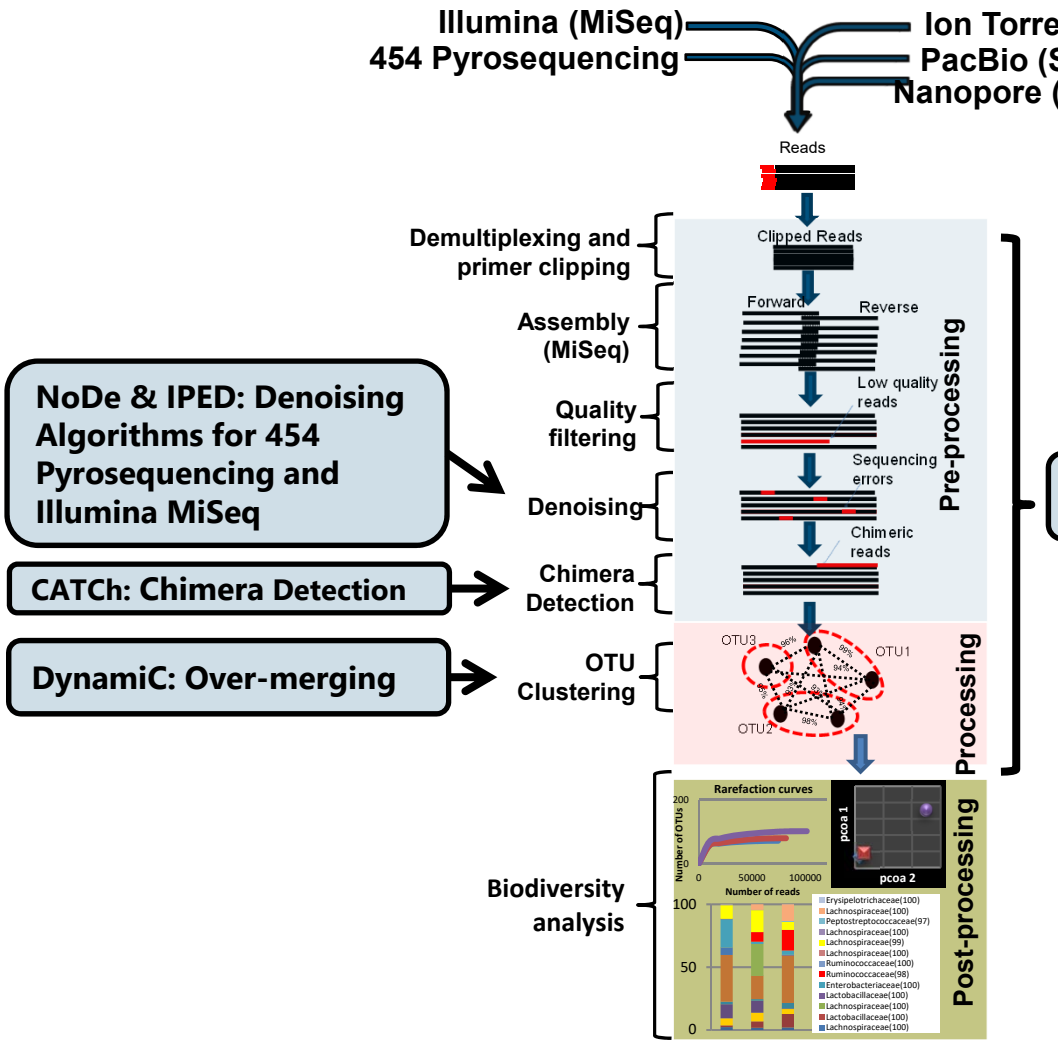




Conclusion

In future experiments: first the dose of the irradiation has to be optimized for our mouse model in order to exhibit relevant clinical symptoms. Second, a more specific examination of apoptosis and examination of oxidative stress will be performed. And third, a larger subset of inflammation markers will also be used. In addition, the formulation and dose of *Arthrospira* sp. as food supplement will also be further optimized.

Summary



Publications

CATCh: Mysara et al., AEM (2015)
NoDe: Mysara et al. BMC Bioinformatics (2015)
iPED: Mysara et al. BMC Bioinformatics (2016)
DynamiC: Mysara et al. GigaScience (2017)
OCToPUS: Mysara et al. FEMS Ecology (2017)

https://github.com/M-Mysara/OCToPUS

Future work – Extend expertise

- New long-read sequencing technologies (NanoPore, PacBio)
- Complex mock communities (n > 200)
- Shotgun metagenomics

Future work

- New long-read sequencing technologies
  - NanoPore
  - PacBio

- Challenging 16S rRNA amplicon sequencing pipelines
  - Complex mock communities (n > 200 strains)
  - Strain / subspecies variation detection

- Shotgun metagenomics
  - Taxonomy versus metabolic potential

Acknowledgements

Dr. Mohamed Mysara
Ruben Props
Dr. Natalie Leys

Prof. Dr. Peter Vandamme
Prof. Dr. Yvan Saeys

Prof. Dr. Jeroen Raes

Prof. Dr. Daniel Charlier

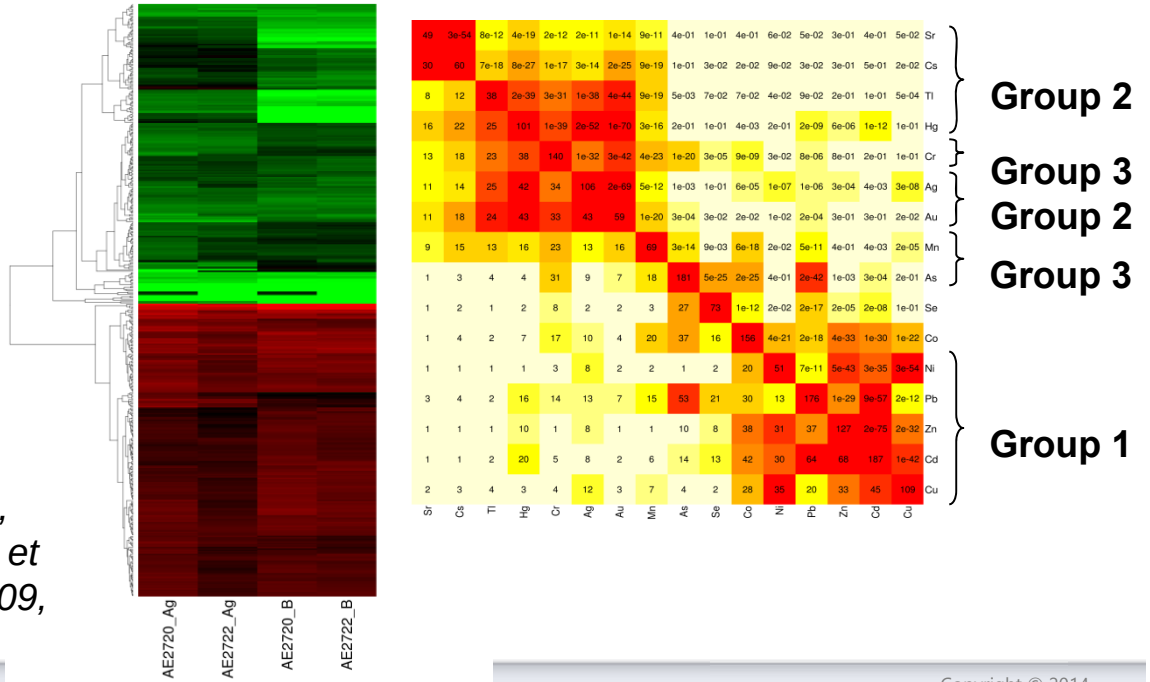
SCK•CEN
Studiecentrum voor Kernenergie
Centre d'Etude de l'Energie Nucléaire
Belgian Nuclear Research Centre

Stichting van Openbaar Nut
Fondation d'Utilité Publique
Foundation of Public Utility

Registered Office: Avenue Herrmann-Debrouxlaan 40 – BE-1160 BRUSSELS
Operational Office: Boeretang 200 – BE-2400 MOL

Microarrays

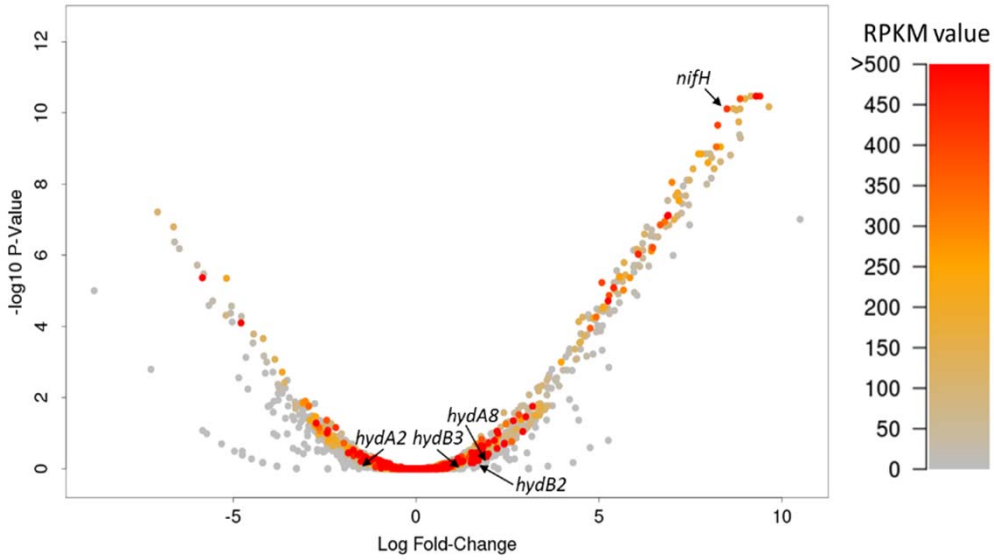
- Types
  - Two-color arrays: *Rhodospirillum rubrum, Cupriavidus metallidurans*
  - Affymetrix: *Pseudomonas aeruginosa*
  - Nimblegen: *Arthrospira sp.* PCC8005
  - Agilent: *Cupriavidus metallidurans*
- In-house facilities
  - Two-color microarrays
  - Affymetrix arrays
- Analysis
  - MIC: BioConductor
  - RDB: Partek



Monsieurs et al., 2011, Van Houdt et al. 2012, Crabbe et al. 2011, Crabbe et al. 2010, Pycke et al, 2010, Badri et al. 2014, Mastroleo et al. 2009, …

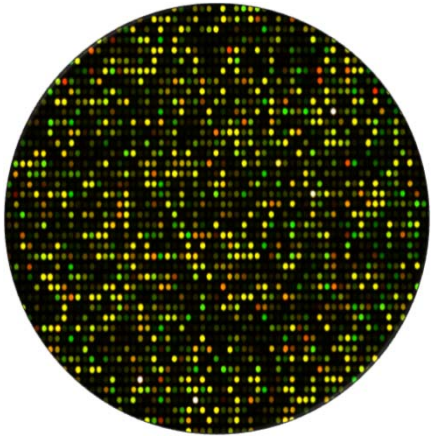
Transcriptomics via RNA-seq

- Application area
  - MIC
    - *Arthrospira* sp. PCC8005
    - *Clostridium bytiricum*
    - *Pseudomonas aeruginosa*
  - BIS
    - To be started
- Analysis
  - BWA / Bowtie
  - EdgeR

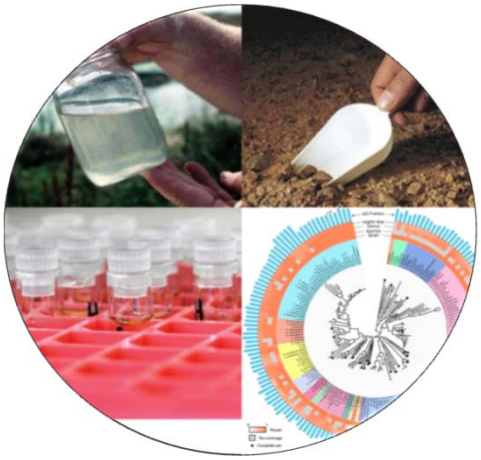


Overview

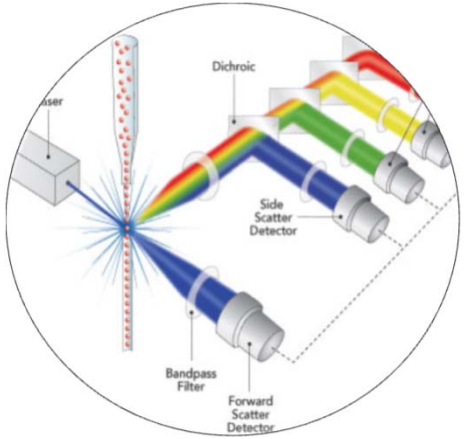
Genomics



Transcriptomics



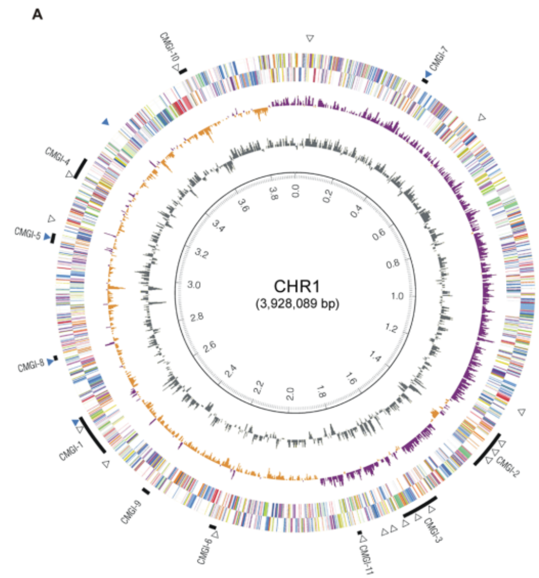
Metagenomics



Flow cytometry

Genome assembly

- *De novo* genome assembly
  - Based on 454 pyrosequencing data
  - Based on Illumina HiSeq data
  - Annotation ➔ GenoScope - MaGe
    - CmetScope
    - ArthroScope
- Applications
  - MIC: Metal resistant bacteria
  - BIS: *Lemna minor* (~ 400 Mbp genome)

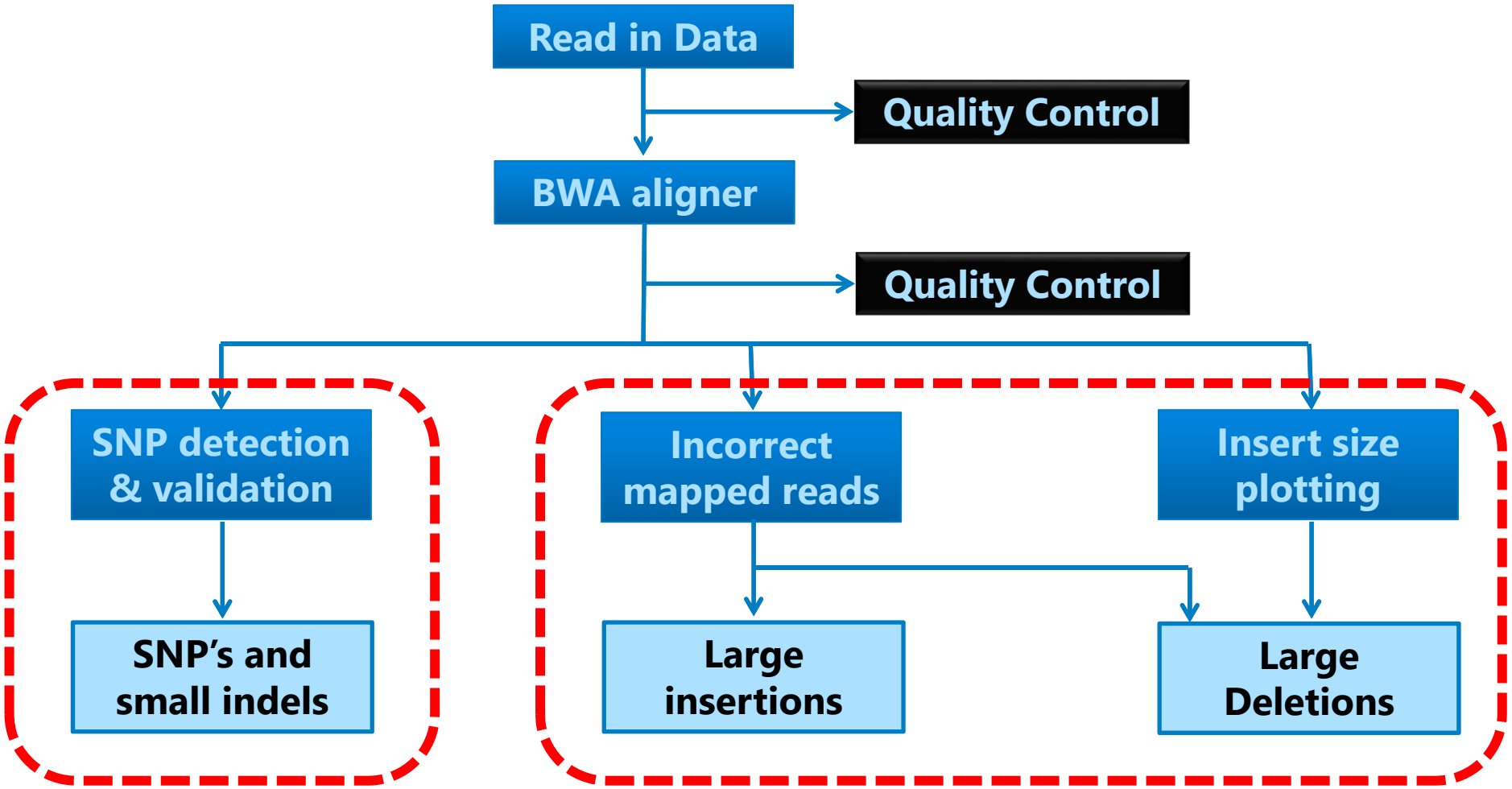


Platform	Insert Size	Read Length	No. Reads	No. Nucleotides	Genome Coverage
Illumina HiSeq 2000	200 bp	2*100 bp	207.985.822	40.730.561.447	100X

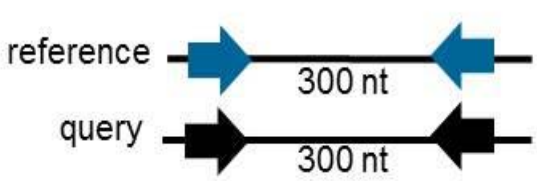
Statistics	SOAPdenovo2[2]	CLC Bio	CLC Bio
Input	preprocessed data	preprocessed data	processed reads + flash data
K-mer size	63	53	53
No. Scaffolds	108607	116254	117403
Max scaffold size	[illegible]	[illegible]	[illegible]
genome length	401	388	410
N50	10194	8059	8423

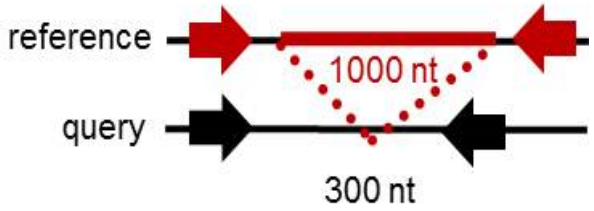
Janssen et al., 2010, Monsieurs et al., 2013, Monsieurs et al., 2014a, Monsieurs et al., 2014b

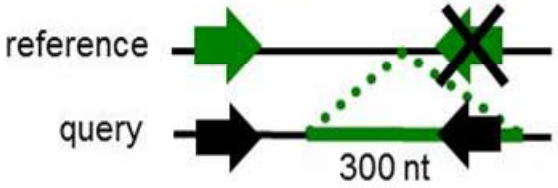
Bacterial resequencing - workflow



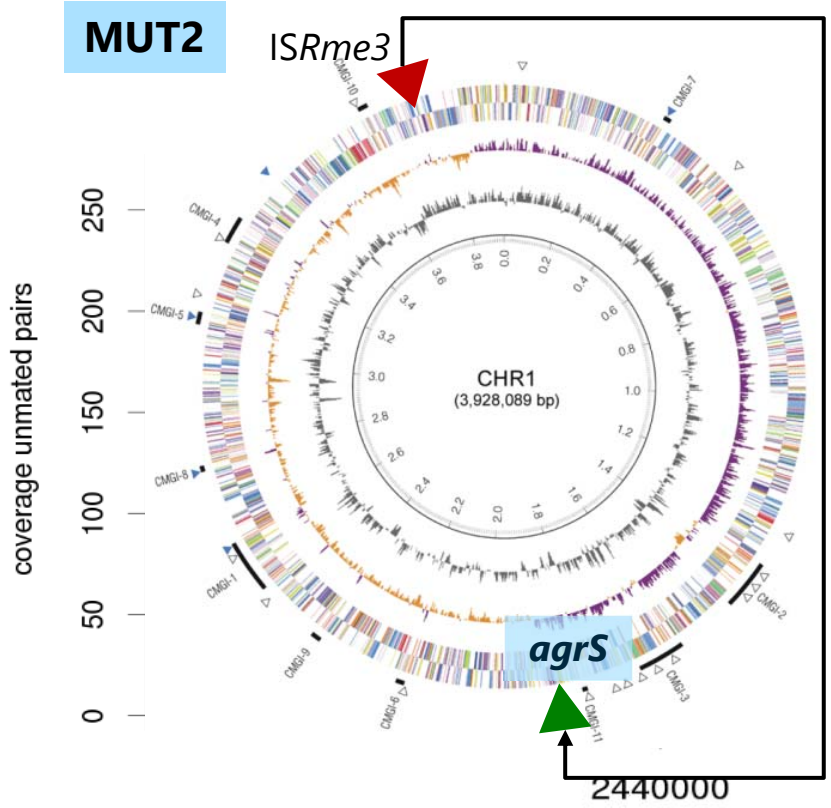
BWA / Samtools / GATK genomic suite



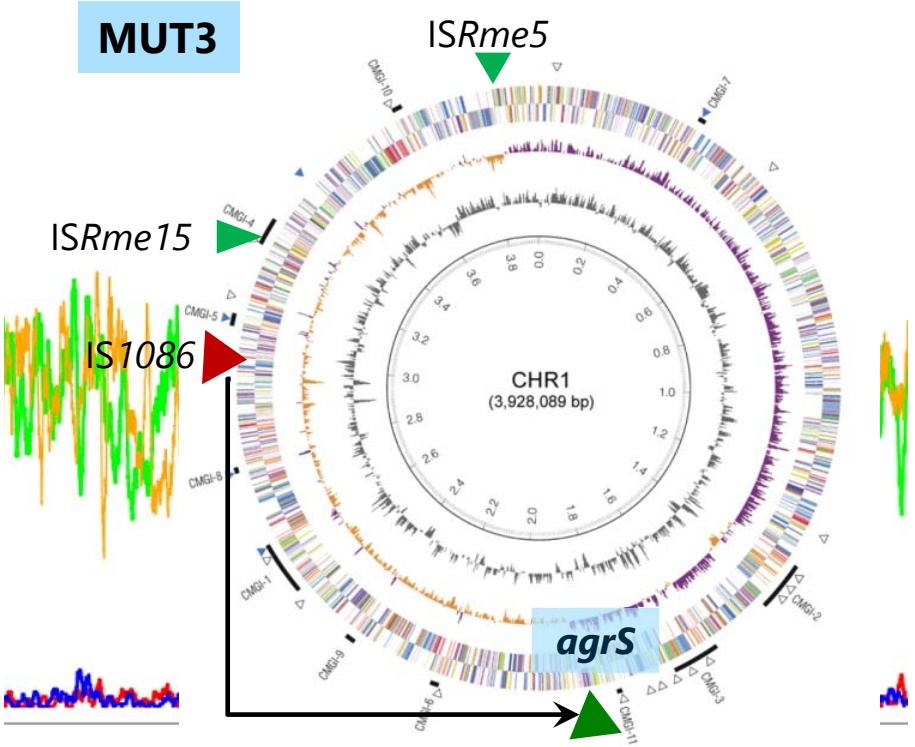




MUT2



MUT3

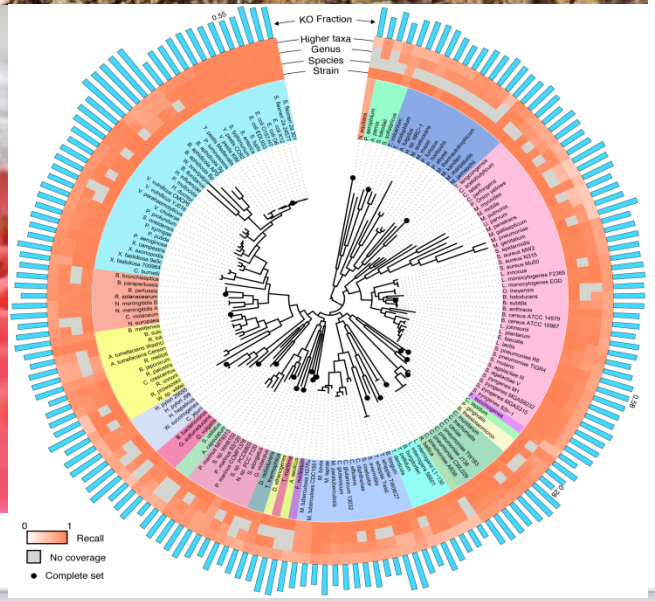


Facilities

- Bio-informatics server – calculations
  - 24 processors – 3.2 GHz
  - 96 Gb memory
  - 6.0 Tb hard drive (RAID5 configuration)
- Data server - storage
  - 1 Tb hard drive in RAID10 configuration
  - Automatic backup to EqualLogic tapes
  - Storage of all raw microarrray and sequencing data
- FERMI cluster
  - **Details to be added**

Slide title

- To be added?: Slides Roel on Affymetrix and Partek



16S rRNA metagenomics algorithms

(GIGA)n SCIENCE

Issues | Advance Articles | Publish ▾ | Alerts | About ▾

Article Contents

Abstract

Supplementary data

ACCEPTED MANUSCRIPT

From reads to operational taxonomic units: an ensemble processing pipeline for MiSeq amplicon sequencing data

Mohamed Mysara; Mercy Njima; Natalie Leys; Jeroen Raes; Pieter Monsieurs

Gigascience giw017. **DOI:** https://doi.org/10.1093/gigascience/giw017
Published: 18 January 2017

Views ▾ | PDF | Cite | Share ▾ | Tools ▾

Abstract

Introduction: The development of high-throughput sequencing technologies has provided microbial ecologists with an efficient approach to assess bacterial diversity at an unseen depth, particularly with the recent advances in the Illumina MiSeq sequencing platform. However, analysing such high-throughput data is posing important computational challenges, requiring specialized bioinformatics solutions at different stages during the processing pipeline, such as

Background: The development of high-throughput sequencing technologies has revolutionized the field of microbial ecology via the sequencing of phylogenetic marker genes (e.g. 16S rRNA gene amplicon sequencing). Denoising, the removal of sequencing errors, is an important step in preprocessing amplicon sequencing data. The increasing popularity of the Illumina MiSeq platform for these applications requires the development of appropriate

robiology Ecology, 94, 2017, fix029

/femsec/fix029
cess Publication Date: 0 2017
rticle

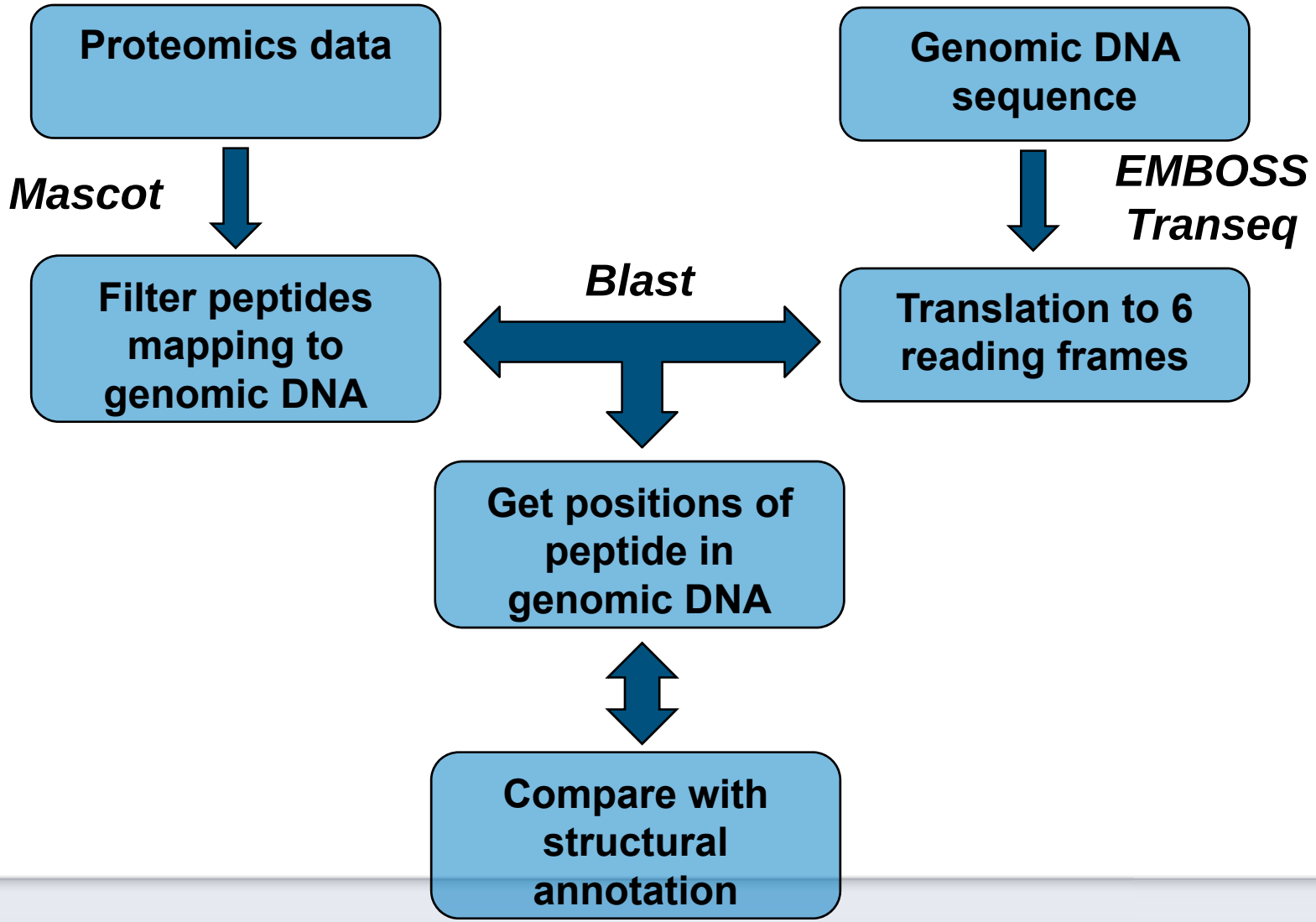
nal taxonomic units

ıben Props[6],
Nico Boon[6], Jeroen Raes[3,4]

), 2400 Mol, Belgium, [2]Department of
ım, [3]VIB lab for Bioinformatics and
obiology and Immunology, REGA institute,
ogy, Ghent University, Belgium and
iversity, Belgium

e-based and *de novo* CATCh) were developed by integrat-
werful method. When comparing our classifiers with exist-
ormance of our ensemble method was observed on a wide
nd Illumina MiSeq data sets. Since our algorithm combines
proach produces more robust results when challenged with
e chimeric range, and various numbers of parents. Addi-
ng pipeline has a beneficial effect on the quality of the clus-

Proteogenomics pipeline

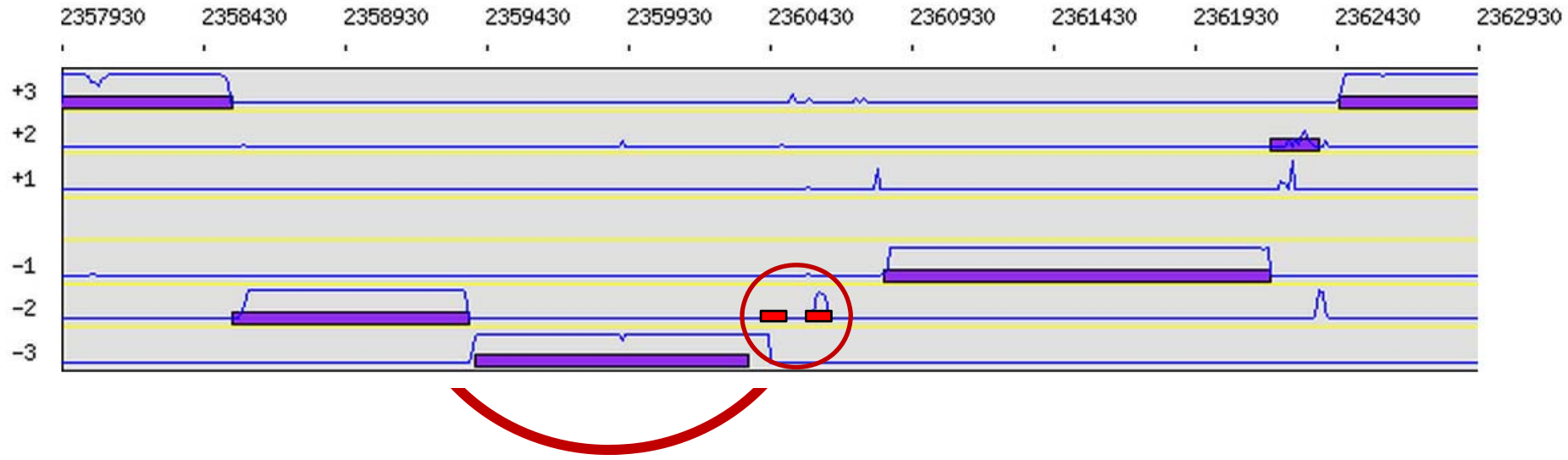


Rhodospirillum rubrum ATCC 11170 - chromosome Rru_A NC_007643

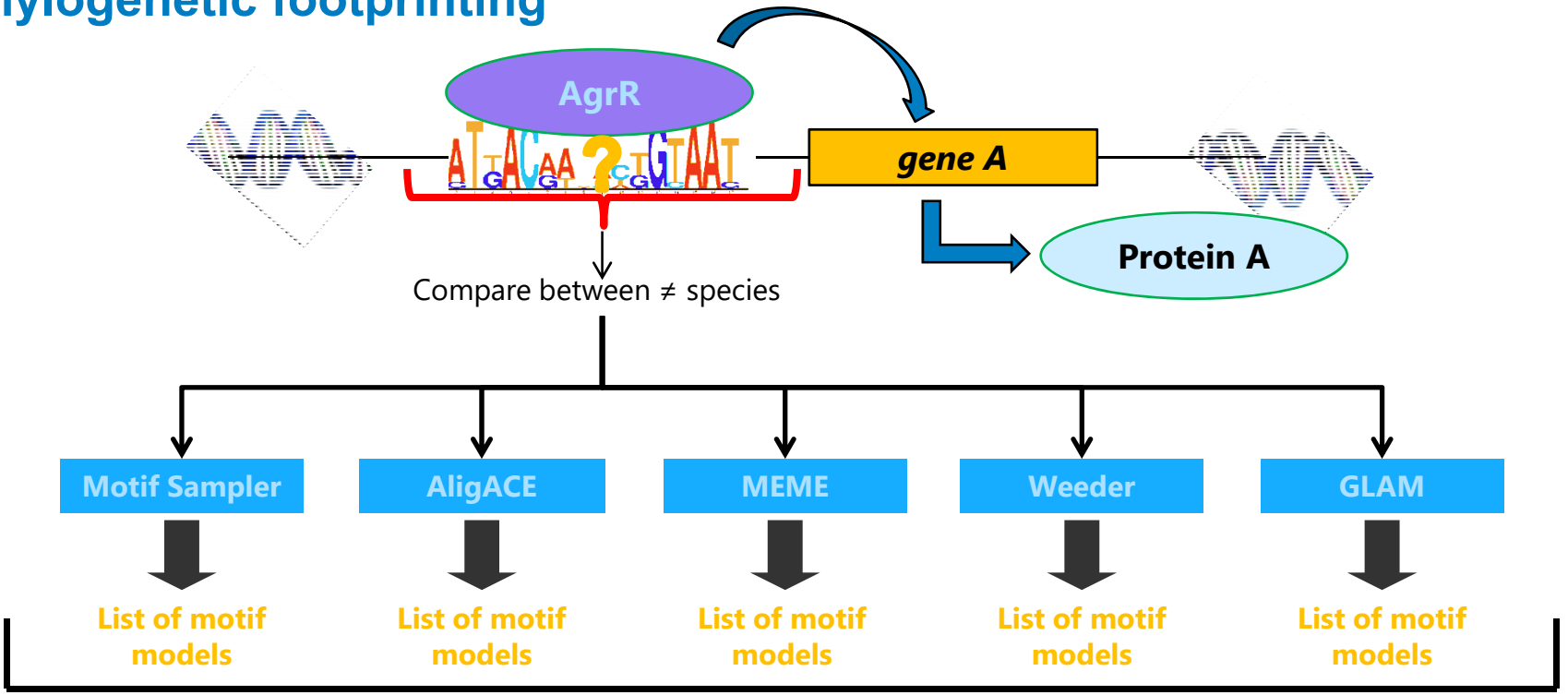
Rhodospirillum rubrum ATCC 11170 - chromosome Rru_A NC_007643

2357930 -- 2362930

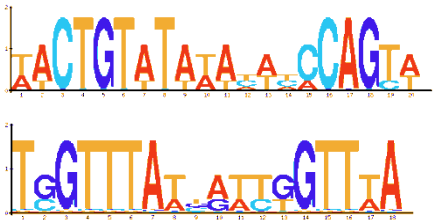
(sequence length : 4352825 bases)



Phylogenetic footprinting

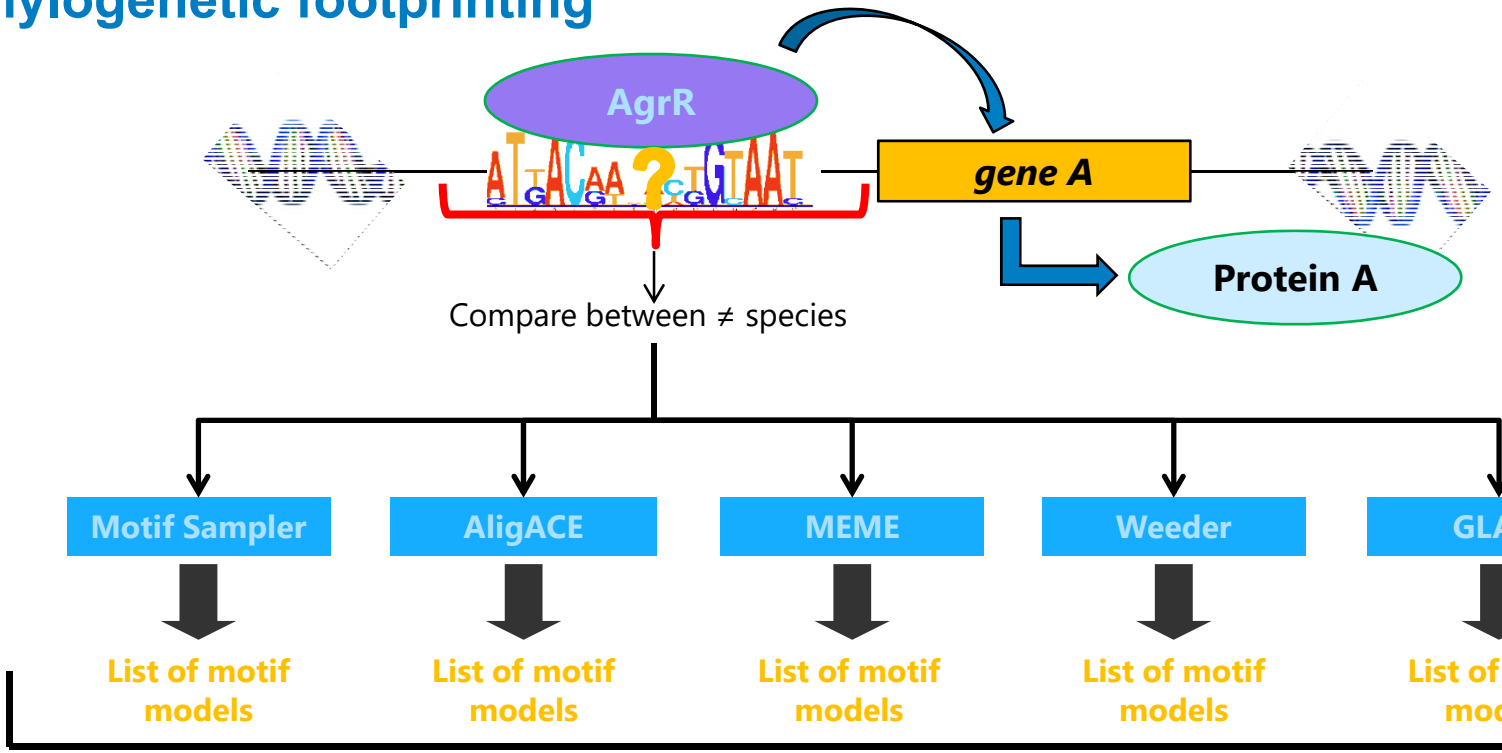


Cluster motif models:
Markov clustering (MCL) of output motif models based on the Pearson Correlation Coefficient

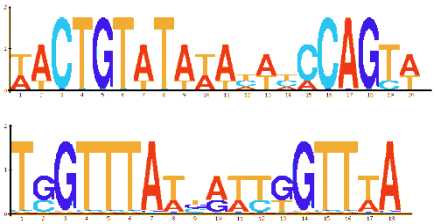


Other pipelines

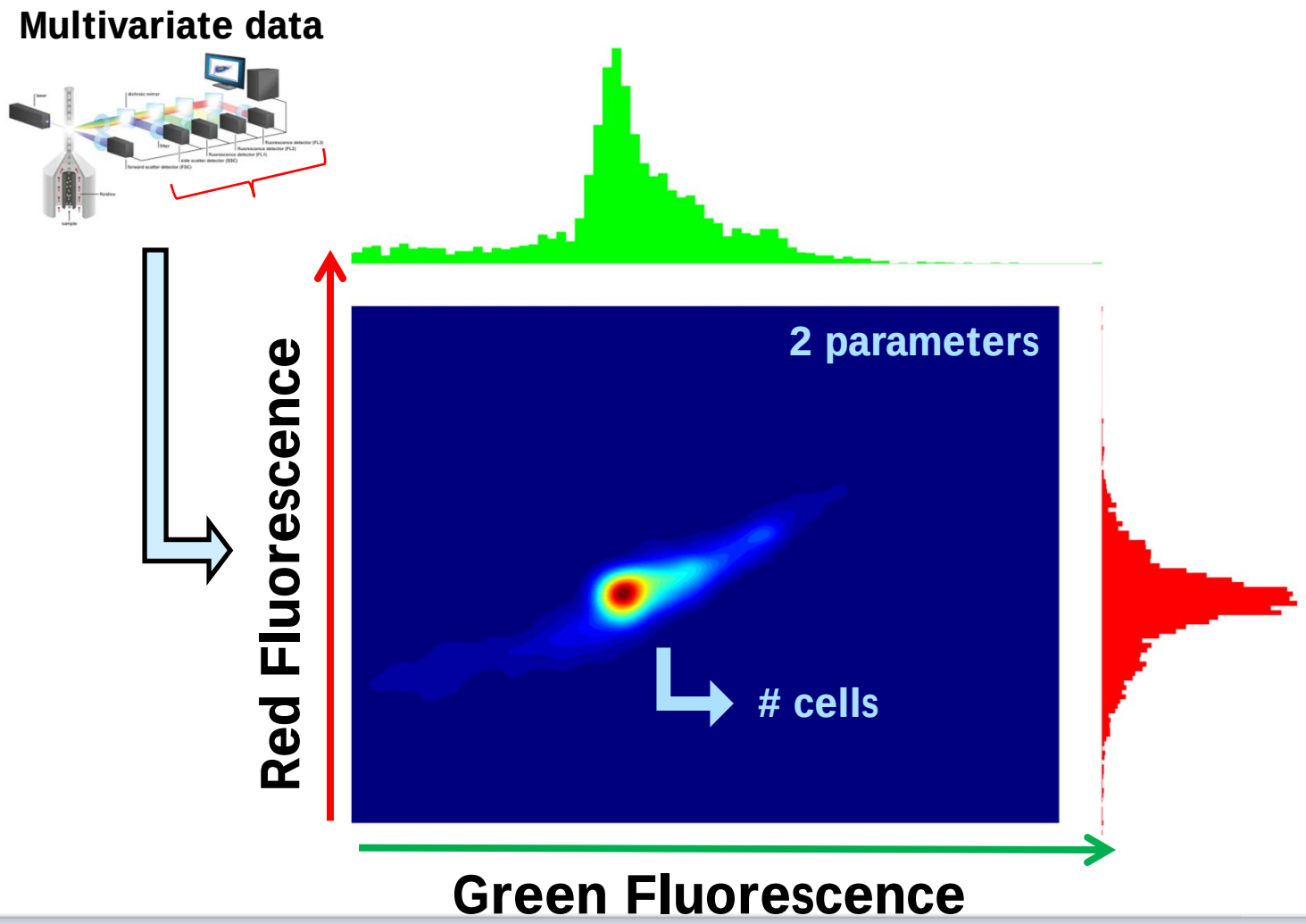
Phylogenetic footprinting



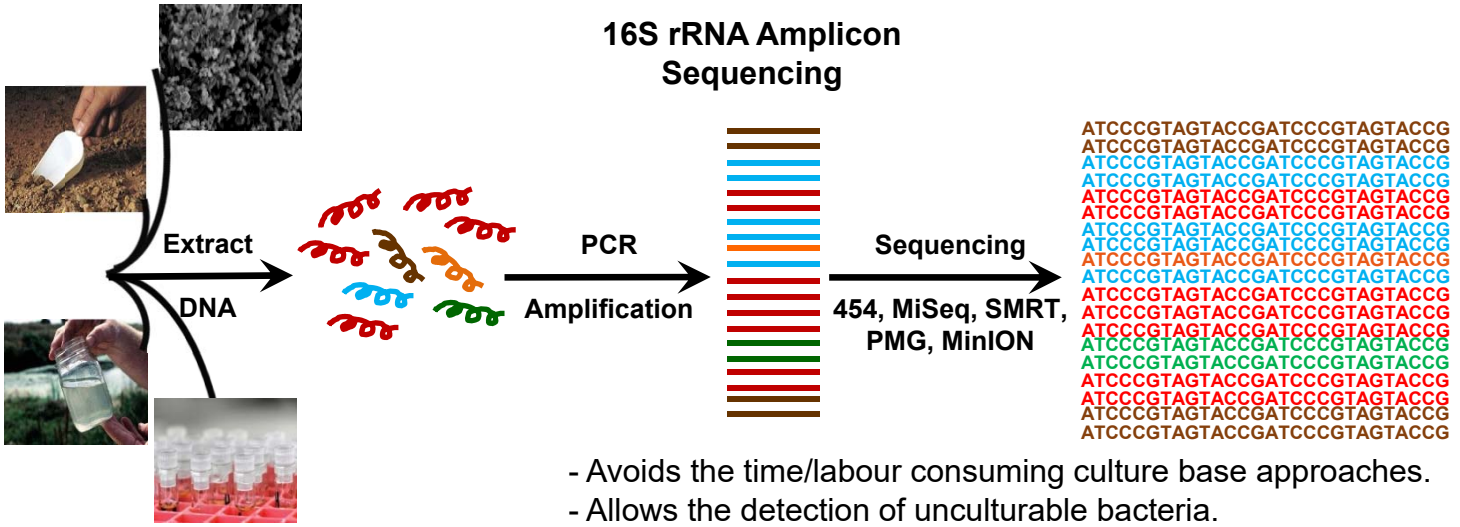
Cluster motif models:
Markov clustering (MCL) of output motif models based on the Pearson Correlation Coefficient



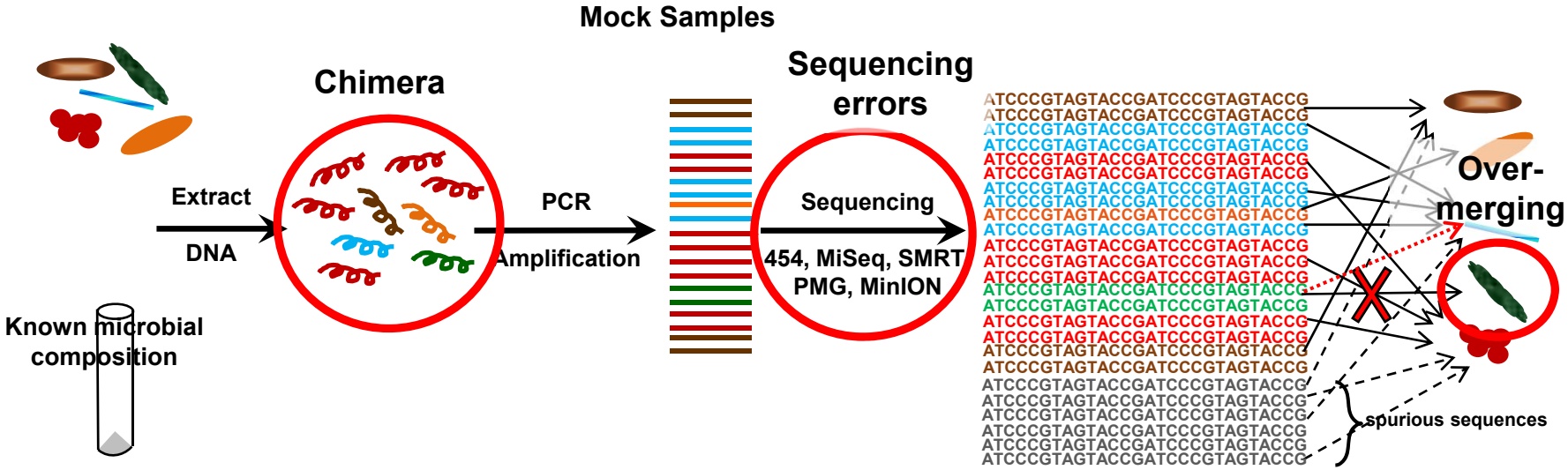
Phenotypic Biomarkers



Introduction: Amplicon sequencing and data analysis



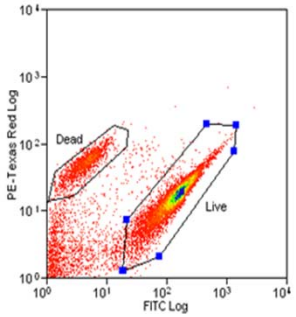
- Avoids the time/labour consuming culture base approaches.
- Allows the detection of unculturable bacteria.
- Can Target a wide range of bacterial species.



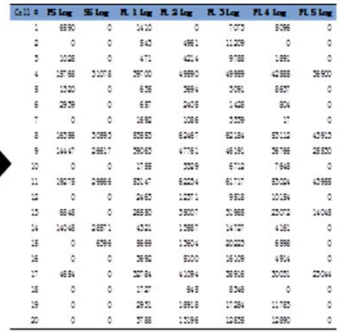
Flow Cytometry

A

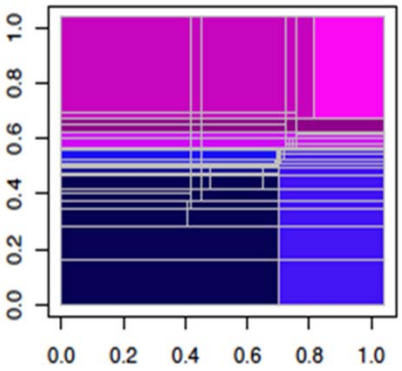
Flow cytometric analysis



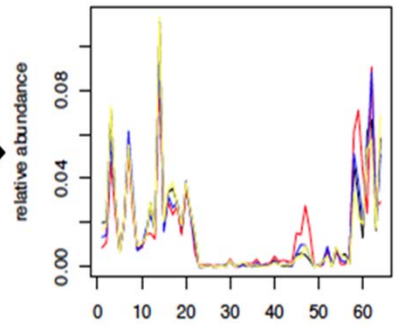
Data extraction

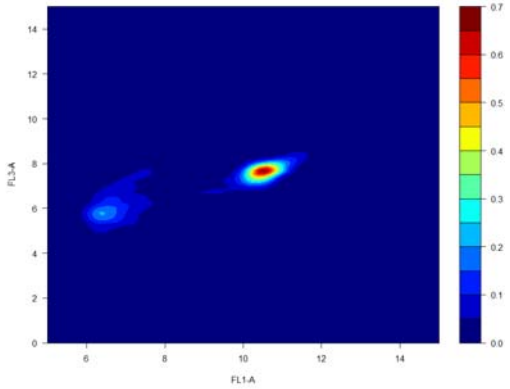


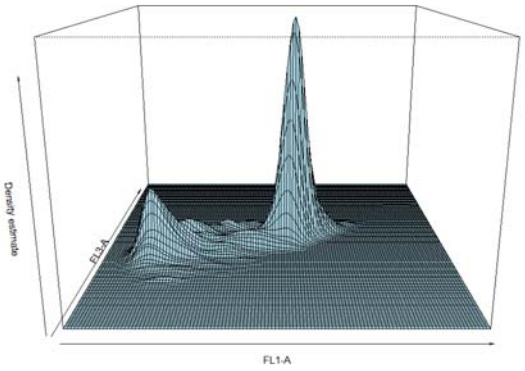
Create Fingerprint model

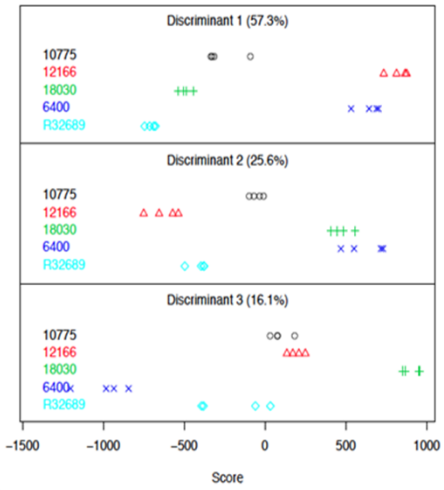


Create Fingerprints

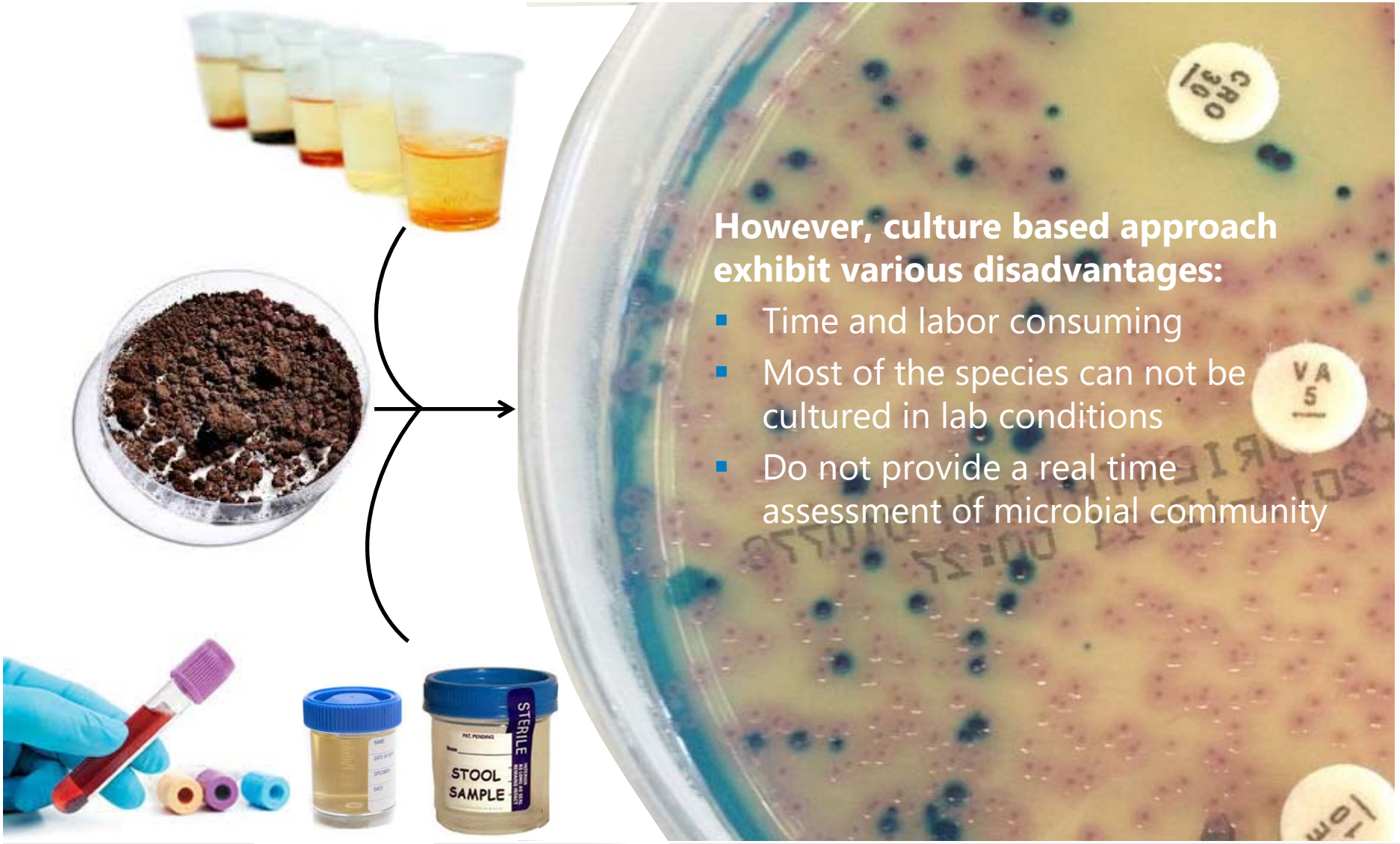








Culture-based approach



However, culture based approach exhibit various disadvantages:

- Time and labor consuming
- Most of the species can not be cultured in lab conditions
- Do not provide a real time assessment of microbial community

Flow cytometry – test case