Computational Inference of Microbial Genotype-Phenotype Relationships

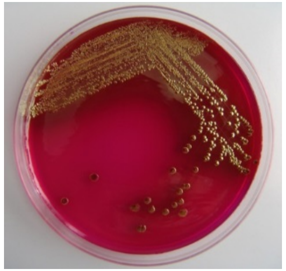
Prof. Alice C. McHardy
Computational Biology of Infection Research
Helmholtz Centre for Infection Research

Towards personalized molecular diagnostics and therapy for infectious diseases

Inferring Genotype-Phenotype and Genotype-Environment associations from microbial omics & biomedical data

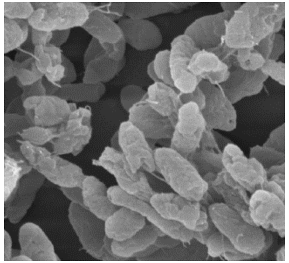
- May indicate biological functions & mechanisms
- Molecular markers
- Support for diagnostic and therapeutic decisions and prognostics

Characterizing antimicrobial resistances



Test growth conditions and resistances (takes time, other organisms found..)

Pure culture (n. a. cultivable)



M. Rohde, HZI

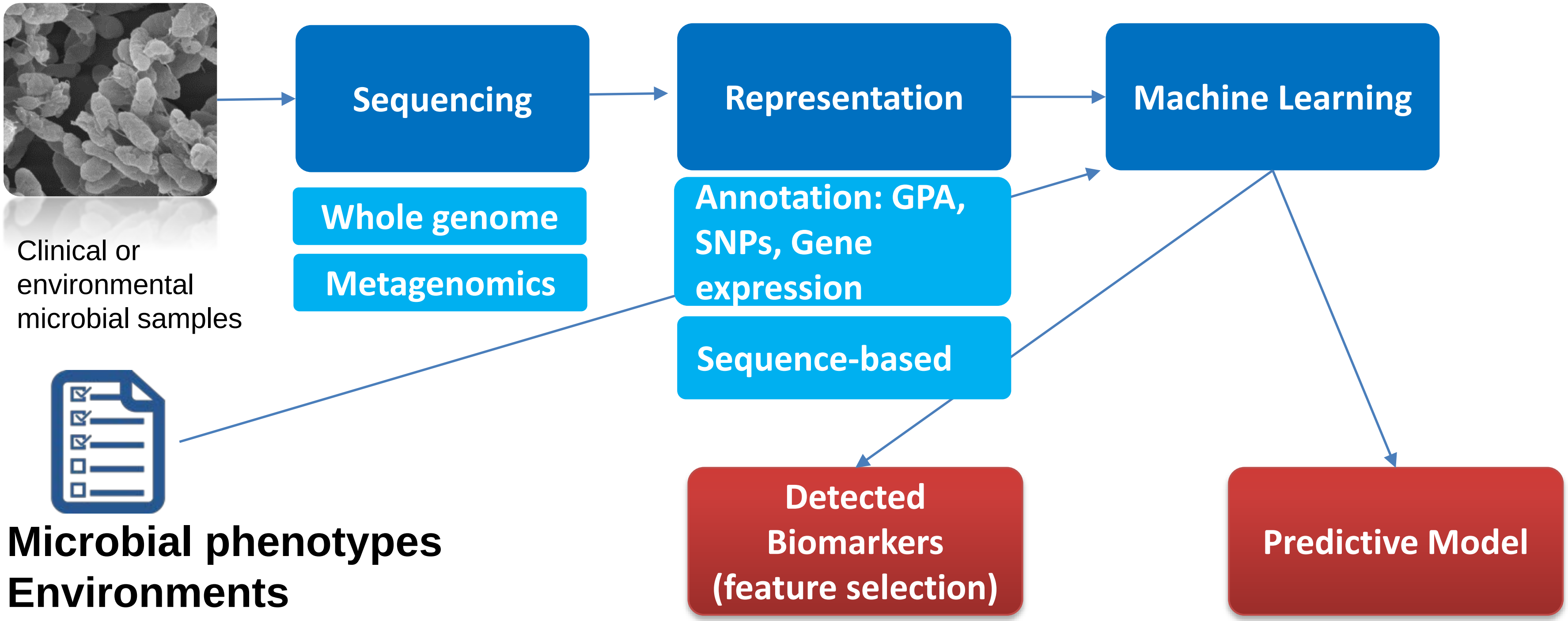
Phenotypes

Metagenome sequencing & bioinformatics

```
[...]sdATGACGATTCCGAAACACGAGCCGCGCGAGGTCTTCGATCGGGCGA
TCGAGCATACGCGGGCGCTTTCCC
CGGCCACTTTCGATCAGTGGTTCGGGGGAGTTCAGTTCGATGATCTGACCG
ACGGCGTCCTGACGCTGCG
CGAGCAGCAGCGCGCCATAGCGCGCCAGCACGTCGCGGGCCCGCGCCATGC
GCCTCGCCGCGTCGTCCTC
GGATCCCGTGAAGATGAGCACGAGCGCCGTGACCTCGCGCCGCTCGC

CCTCGACGGGGGTGCGCTCGTTGGCCCCGTTCGGCTCCTCTTGGAGCACGG
CGCTCGCCTGGATCTCGGC

GCTCTTCTCCTCGTGGAACGGCTCCAGGAAATCGGCGAGGTCGTTCGCCCC
GAACCGCTCGGCCGTCGCG
TAGAAGAAGCCGAGGAGCGCCTCGTGCAGCCGCCCGGCGTCGGGGAACCG
ATCCTCCGGGTTCTTCGCGA
GCATCGTGCTCACGACGTCGGCGAGCGACCGGGGCACGTCGGGGCGCACG
AGCGCGAGCGGCGGATACTC

GCTCGCCTGCAGCCGCCGCATGATCTCGAACGCCGTCGGCGCCGCGAACGG
GTTCGACCCGGCGAGCATC
GATGGTCGTGTTCCGCGGGATGAGCTTGAAGCTCACGCCGCCGCCGGTCTC
GACGCCGAGCGACAGCGGC

GTCACGTCGAGGAGCAGCACCTCGTCCACCTGCCCGGAGAGCGCGGCGCCC
TGCAGCGCAGCGCCGACCG
CGACGACCTCGTCGGGGTTGACGCCCTTGTTCGGCTCCCGGCCGAAGAACT
CGCGCACCGCCGCCTGCAC
CGCGGGCATCCGGGTCATGCCGCCGACGAGGACCACCGTGTTGACGGCGG
AGACGGGGAGCTTGGCGTCC
CCGAGCGTCGCGCGACAGACGTCGATGGTCCGCCGGATGAGGCCCTCGCAG
AGCATCTCGAGCTCGTTGC
GCCGCATCGTCCGCTCGAGGTGGAGCGGCCCTCCGCCCGGGCCGACGGCGA
TCAACGCCATCTTCATCTC
```

Biomarker discovery / phenotype prediction

Genotype-Phenotype / Environment Associations



Clinical or environmental microbial samples

Sequencing

Whole genome

Metagenomics

Representation

Annotation: GPA, SNPs, Gene expression

Sequence-based

Machine Learning

Detected Biomarkers (feature selection)

Predictive Model

Microbial phenotypes
Environments

Cultivation conditions
Antibiotic resistances
Host disease condition
…..

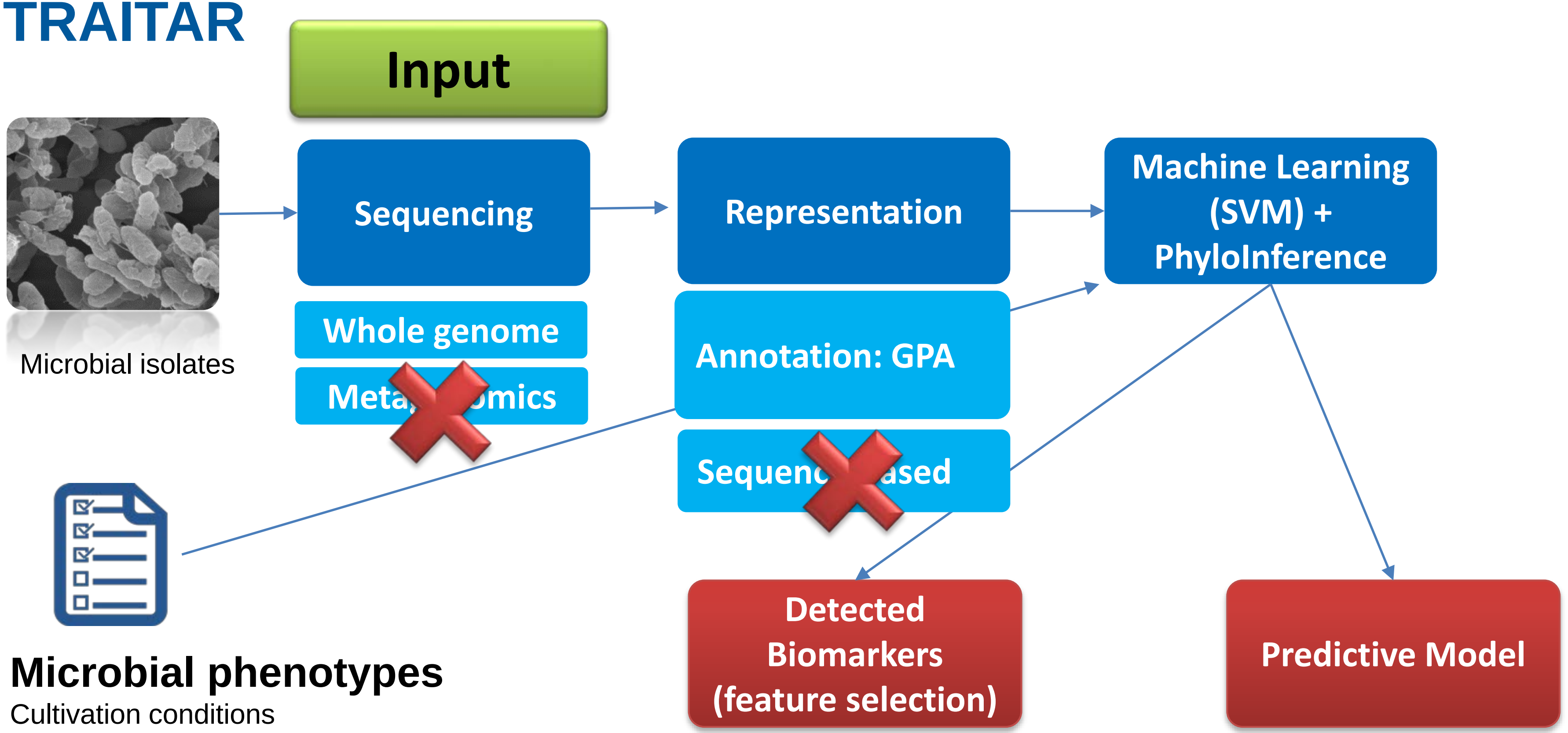
Characterization
Designing diagnostics

Clinical diagnosis and treatment

Method overview

Software	Setting	Input	Features	Predictive model	Biomarker detection	Tree inference
TRAITAR	Microbial genomics	Sequences	Gene family presence or absence (GPA)	✓	✓	×
Seq2Geno2Pheno	Microbial genomics / transcriptomics	Sequence and gene expression levels	Sequences (SNPs, GPA) and expression levels	✓	✓	✓
MicroPheno	16S rRNA amplicon data	Sequences	K-mers	✓	×	×
DiTaxa	16S rRNA amplicon data	Sequences	Variable length subsequences	✓	✓	×

TRAITAR



Input

Microbial isolates

Sequencing

Whole genome

~~Metagenomics~~

Representation

Annotation: GPA

~~Sequence based~~

Machine Learning (SVM) + PhyloInference

Detected Biomarkers (feature selection)

Predictive Model

Microbial phenotypes

Cultivation conditions
Antibiotic resistances
..

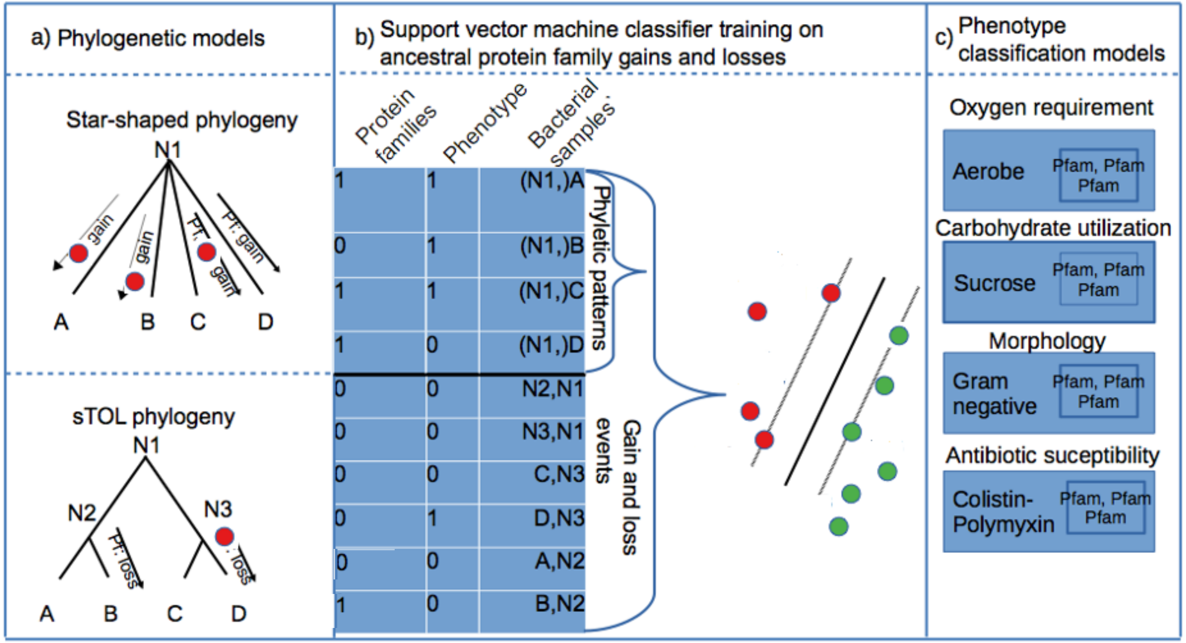
Characterization
Designing diagnostics

Clinical diagnosis and treatment

HELMHOLTZ
CENTRE FOR INFECTION RESEARCH

TRAITAR

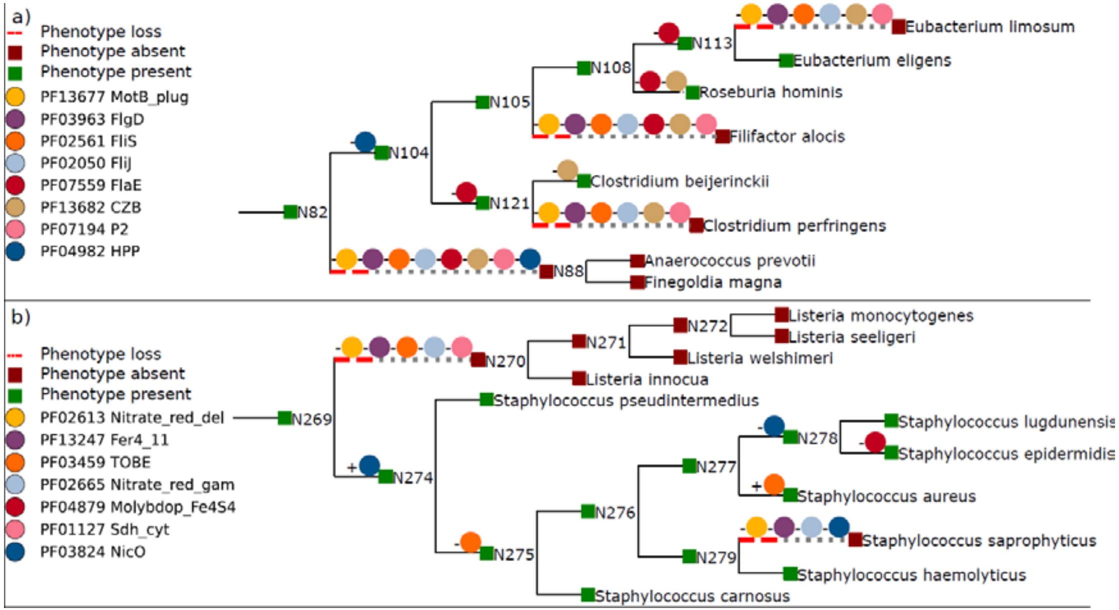
Machine learning combined with evolutionary modelling for predicting microbial phenotypes



Protein families	Phenotype	Bacterial samples
1	1	(N1,)A
0	1	(N1,)B
1	1	(N1,)C
1	0	(N1,)D
0	0	N2,N1
0	0	N3,N1
0	0	C,N3
0	1	D,N3
0	0	A,N2
1	0	B,N2

Weimann *et al.*, mSystems 2016

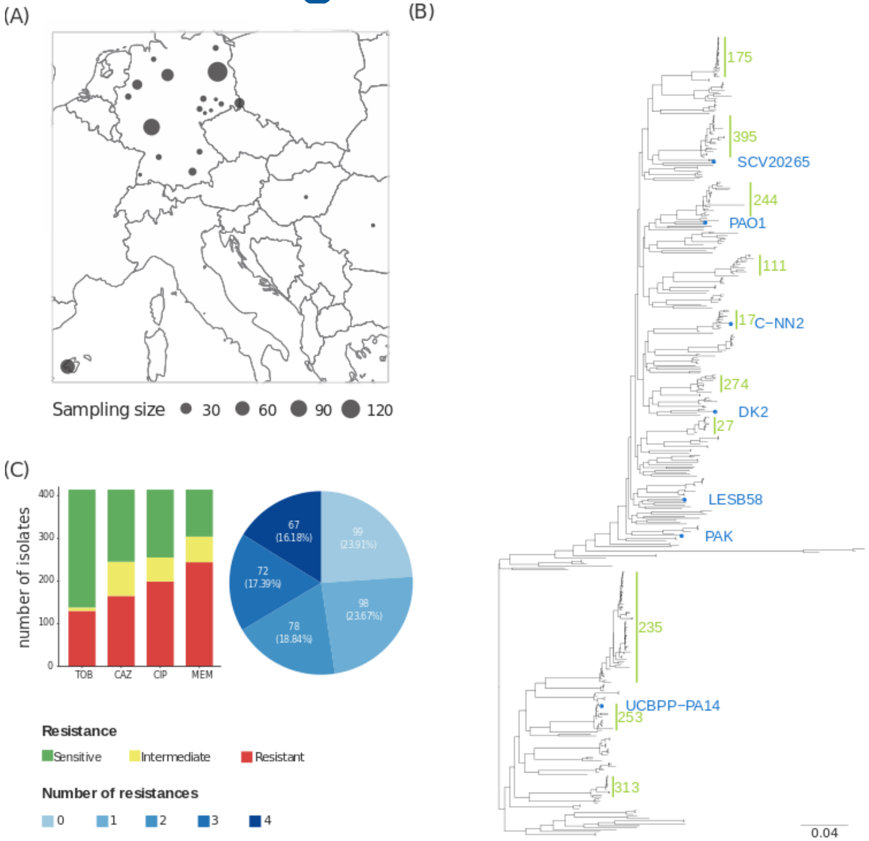
Key protein families for selected phenotypes



Motility

Nitrate reduction

Predicting antimicrobial resistances for *P. aeruginosa*

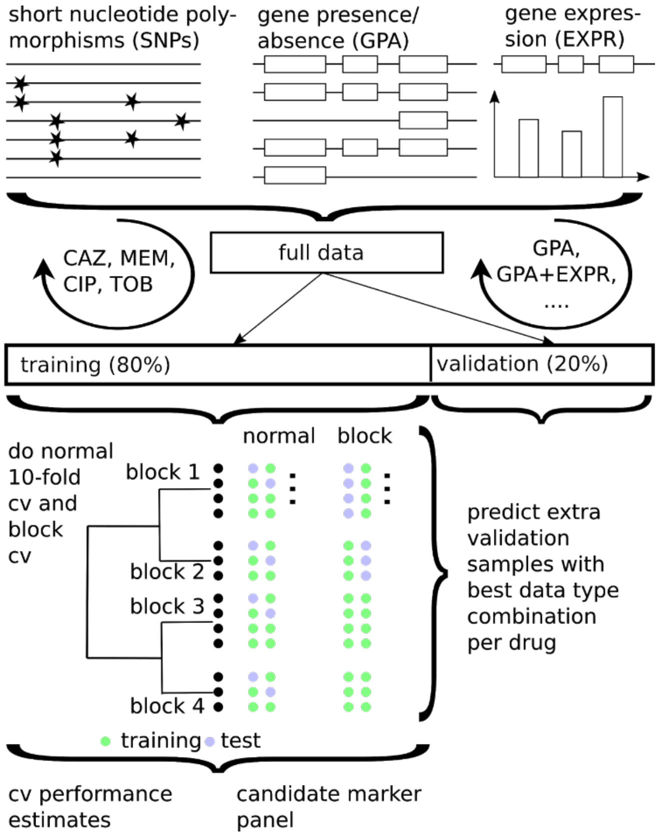


S. Häussler,
HZI

- Bacterial pathogen with multiple AMRs, causing complications e.g. in cystic fibrosis
- 414 clinical isolates
- 4 common antibiotics
- Genome, transcriptome

{Khaledi, Weimann} *et al*., EMBO Mol. Medicine, 2020

Seq2Geno2Pheno



short nucleotide polymorphisms (SNPs)

gene presence/ absence (GPA)

gene expression (EXPR)

full data

CAZ, MEM, CIP, TOB

GPA, GPA+EXPR, ...

training (80%)

validation (20%)

do normal 10-fold cv and block cv

normal

block

block 1

block 2

block 3

block 4

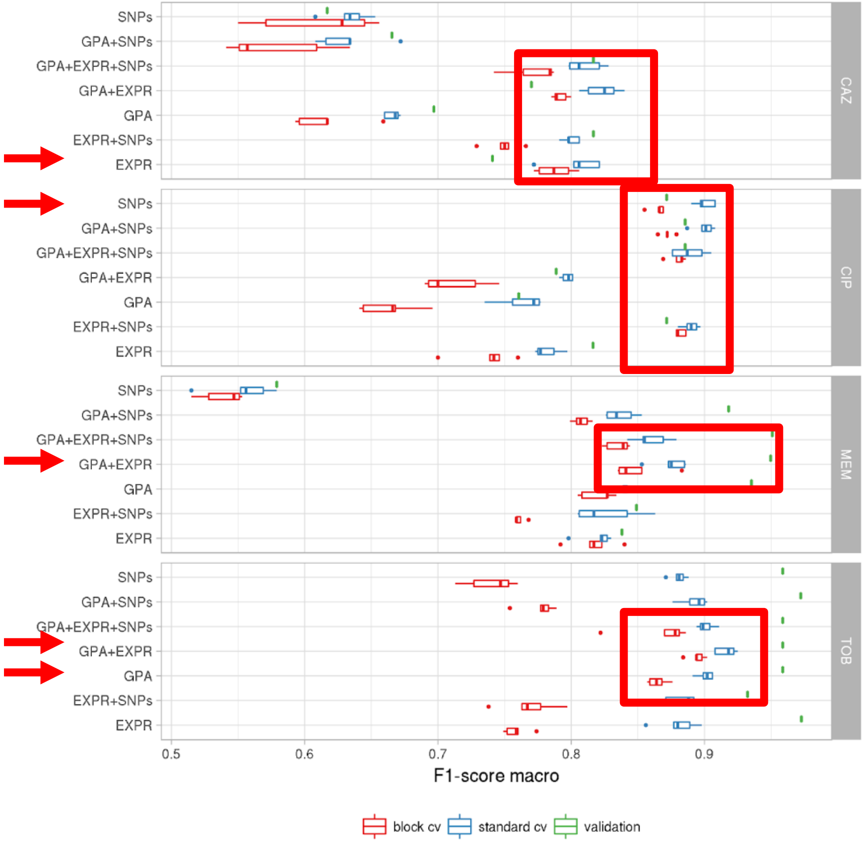
● training ● test

predict extra validation samples with best data type combination per drug

cv performance estimates

candidate marker panel

Predicting sensitivity / resistance with high recall/precision



Ceftazidime

Ciprofloxacin

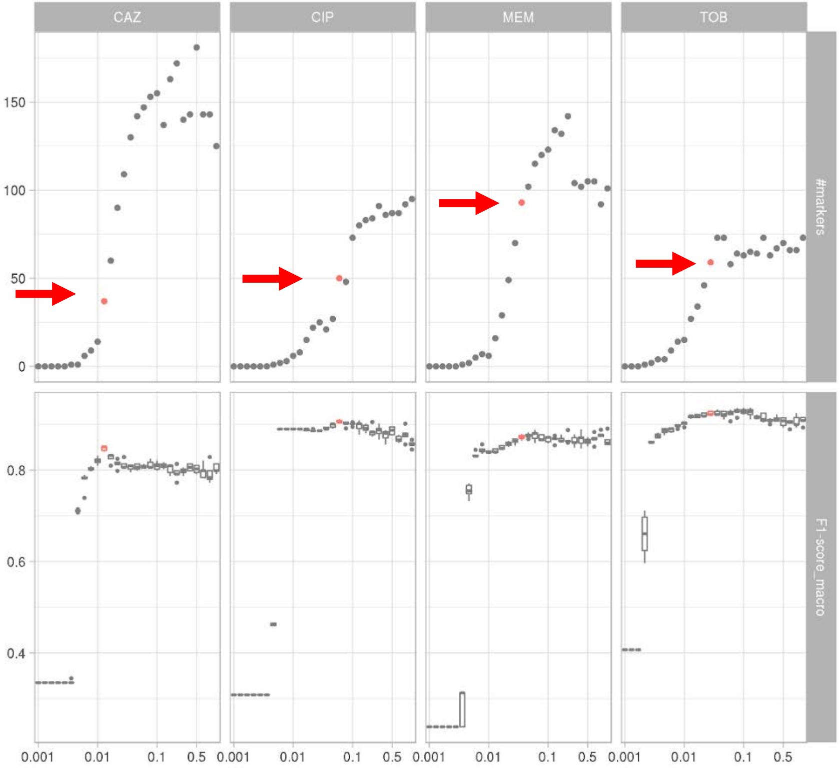
Meropenem

Tobramycin

HELMHOLTZ
CENTRE FOR
INFECTION RESEARCH

.. From few molecular markers

Model with fewest features within 1 std. dev. to best performing model



# markers

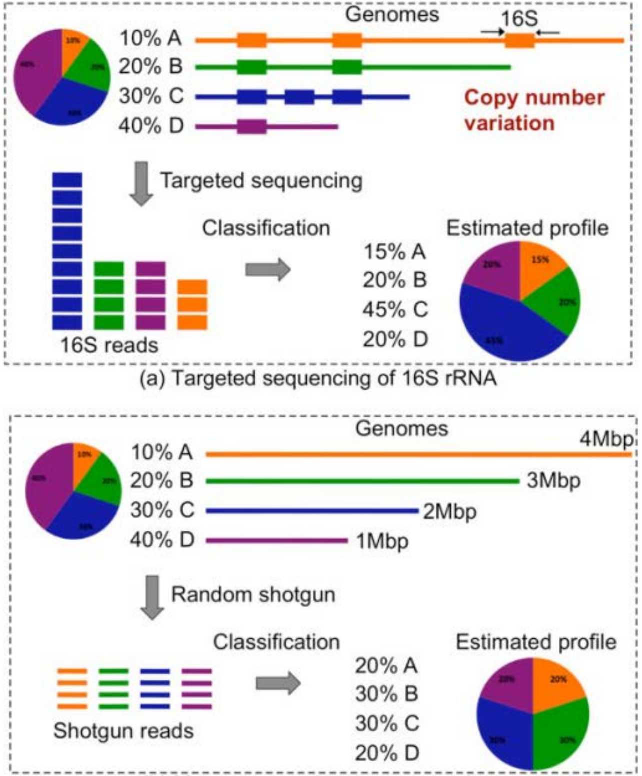
Macro F1-score

SVM C parameter

Meta-omics - Studying microbial communities by sequencing

Studying microbial communities - the basics

- Who is there? Taxonomic profiling
  - by marker gene (rRNA, ITS regions)
  - shotgun metagenome sequencing



(a) Targeted sequencing of 16S rRNA

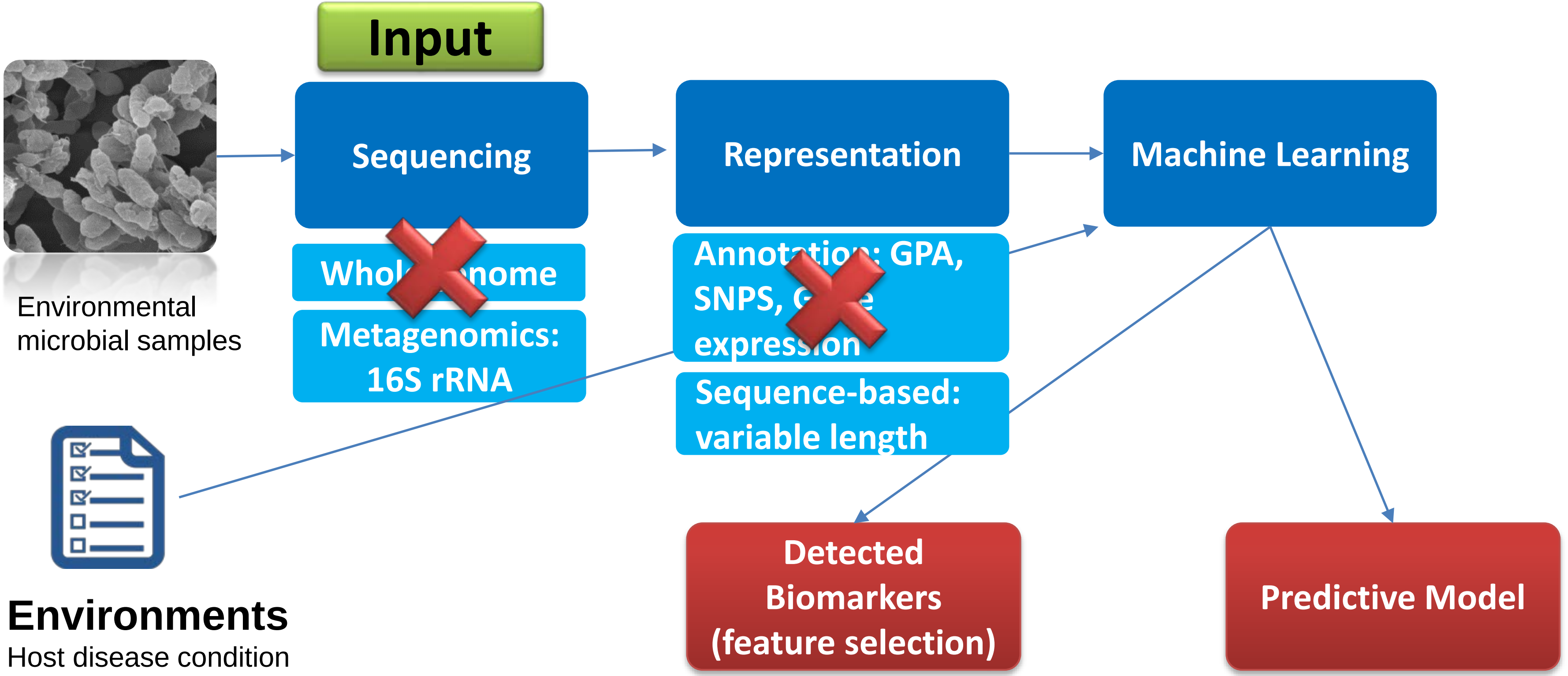
(b) Metagenome shotgun sequencing

Source: Liu *et al.* (2011)

OTU clustering

- After sequencing, 16S rRNA data are usually clustered into groups of closely related sequences, referred to as Operational Taxonomic Units (OTUs)
    - Computationally expensive, as needs sequence alignment
    - Taxonomically inconsistent
    - Sequence similarities between OTUs are ignored

DiTaxa



Input

Environmental microbial samples

Sequencing

~~Whole genome~~

Metagenomics: 16S rRNA

Representation

~~Annotation: GPA, SNPS, Gene expression~~

Sequence-based: variable length

Machine Learning

Environments

Host disease condition

Detected Biomarkers (feature selection)

Predictive Model

Characterization
Designing diagnostics

Clinical diagnosis and treatment

Asgari *et al*., Bioinformatics 2019

DiTaxa: biomarker detection from 16S rRNA

- Inference of variable length features using Nucleotide-Pair Encoding (NPE)
- Better than OTUs in detecting differential taxa for host disease phenotypes, and for host disease phenotype prediction

w. M. Mofrad
UC Berkeley

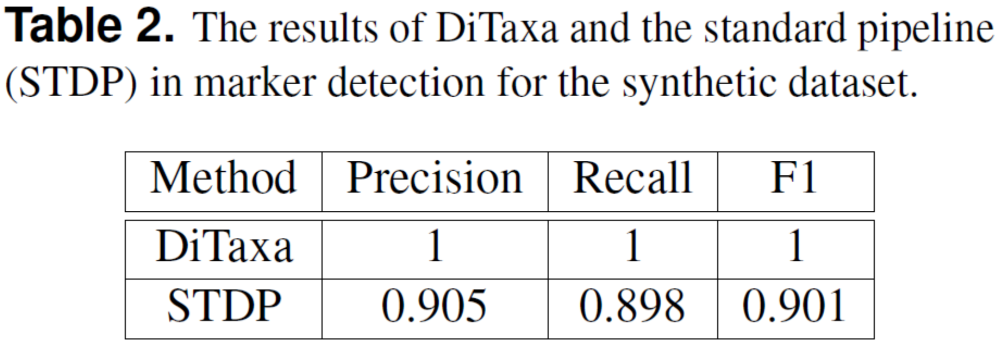


Table 2. The results of DiTaxa and the standard pipeline (STDP) in marker detection for the synthetic dataset.

Method	Precision	Recall	F1
DiTaxa	1	1	1
STDP	0.905	0.898	0.901

Microbial markers for periodontal disease

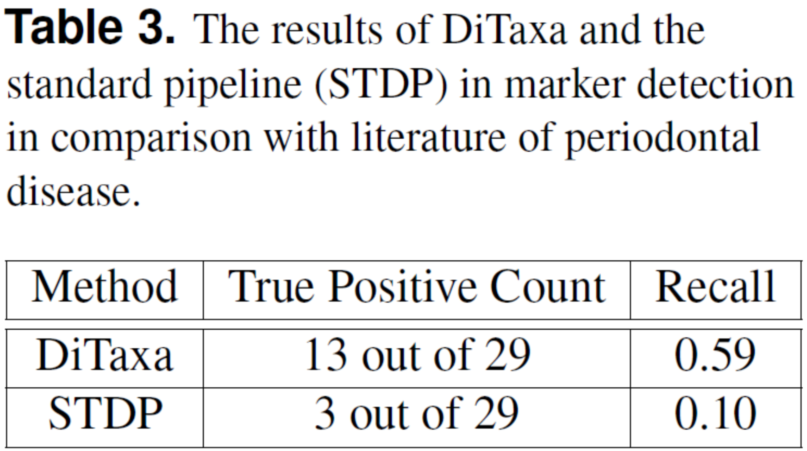
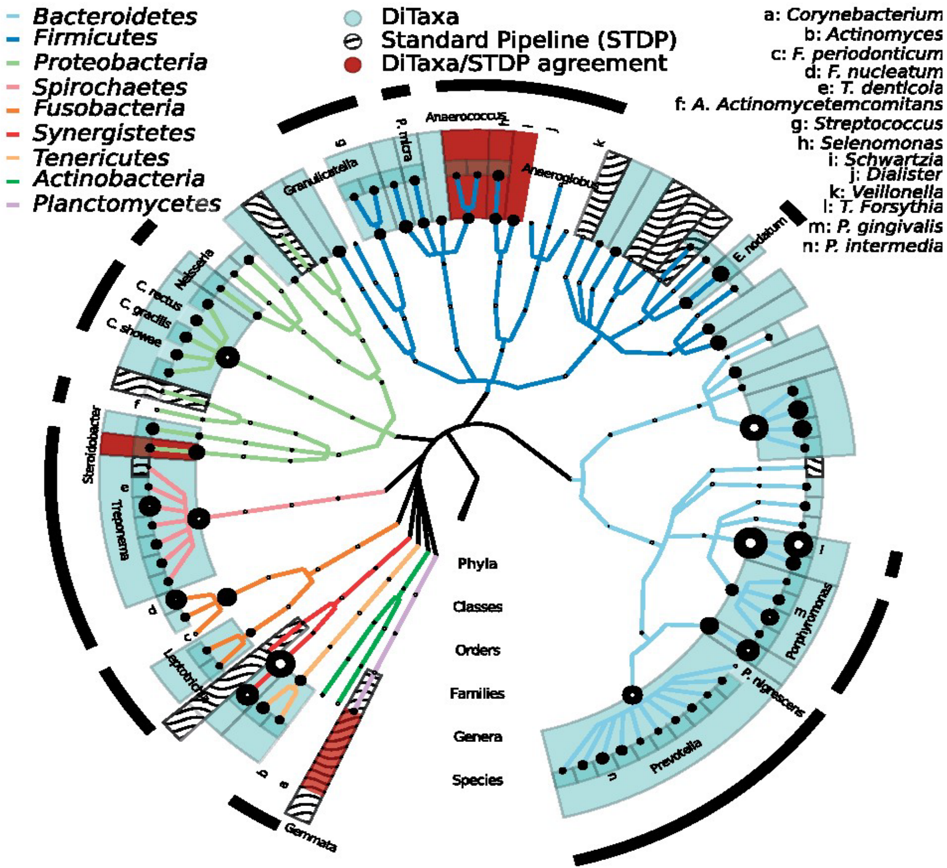
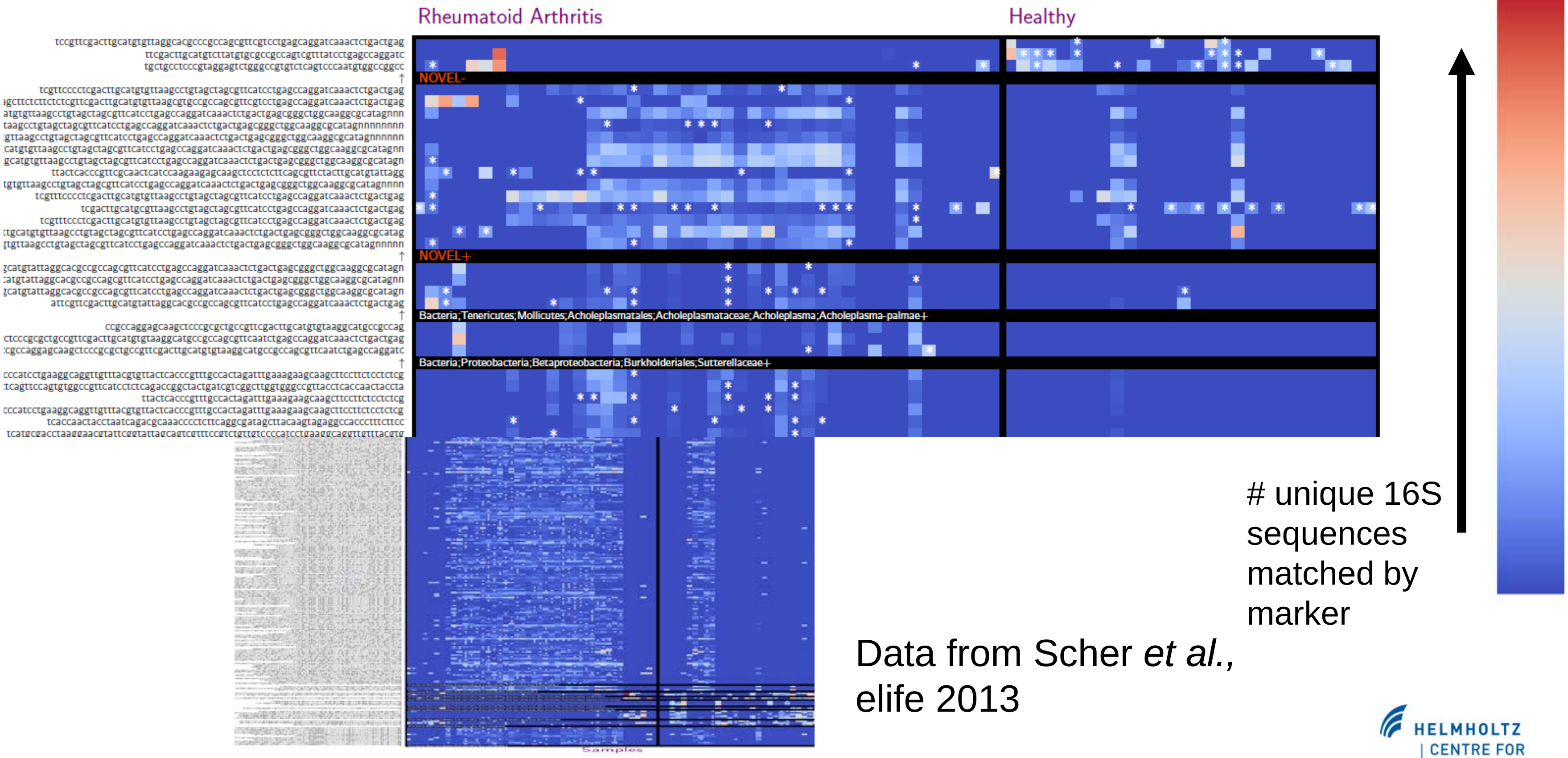


Table 3. The results of DiTaxa and the standard pipeline (STDP) in marker detection in comparison with literature of periodontal disease.

Method	True Positive Count	Recall
DiTaxa	13 out of 29	0.59
STDP	3 out of 29	0.10



Healthy versus new onset RA

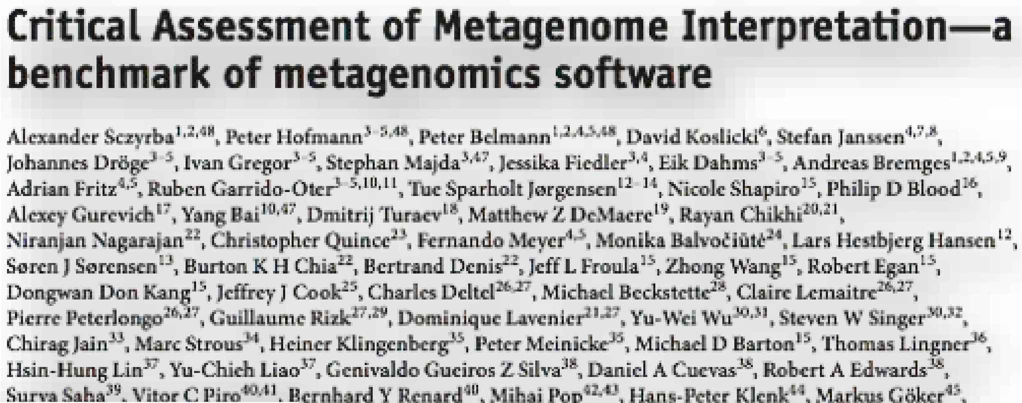


# unique 16S sequences matched by marker

Data from Scher *et al.*, elife 2013

Critical Assessment of Metagenome Interpretation

Towards a **comprehensive** and **objective** evaluation of computational metagenomics software

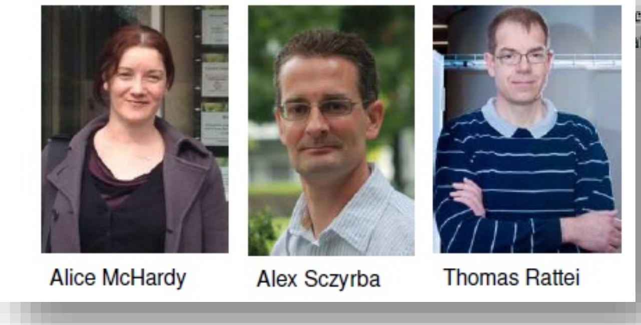


Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba[1,2,48], Peter Hofmann[3–5,48], Peter Belmann[1,2,4,5,48], David Koslicki[6], Stefan Janssen[4,7,8], Johannes Dröge[3–5], Ivan Gregor[3–5], Stephan Majda[3,47], Jessika Fiedler[3,4], Eik Dahms[3–5], Andreas Bremges[1,2,4,5,9], Adrian Fritz[4,5], Ruben Garrido-Oter[3–5,10,11], Tue Sparholt Jørgensen[12–14], Nicole Shapiro[15], Philip D Blood[16], Alexey Gurevich[17], Yang Bai[10,47], Dmitrij Turaev[18], Matthew Z DeMaere[19], Rayan Chikhi[20,21], Niranjan Nagarajan[22], Christopher Quince[23], Fernando Meyer[4,5], Monika Balvočiūtė[24], Lars Hestbjerg Hansen[12], Søren J Sørensen[13], Burton K H Chia[22], Bertrand Denis[22], Jeff L Froula[15], Zhong Wang[15], Robert Egan[15], Dongwan Don Kang[15], Jeffrey J Cook[25], Charles Deltel[26,27], Michael Beckstette[28], Claire Lemaitre[26,27], Pierre Peterlongo[26,27], Guillaume Rizk[27,29], Dominique Lavenier[21,27], Yu-Wei Wu[30,31], Steven W Singer[30,32], Chirag Jain[33], Marc Strous[34], Heiner Klingenberg[35], Peter Meinicke[35], Michael D Barton[15], Thomas Lingner[36], Hsin-Hung Lin[37], Yu-Chieh Liao[37], Genivaldo Gueiros Z Silva[38], Daniel A Cuevas[38], Robert A Edwards[38], Surya Saha[39], Vitor C Piro[40,41], Bernhard Y Renard[40], Mihai Pop[42,43], Hans-Peter Klenk[44], Markus Göker[45],

e[15], Julia A Vorholt[46], Paul Schulze-Lefert[10,11], Edward M Rubin[15],

ttei[18] & Alice C McHardy[3–5,11]



Alice McHardy

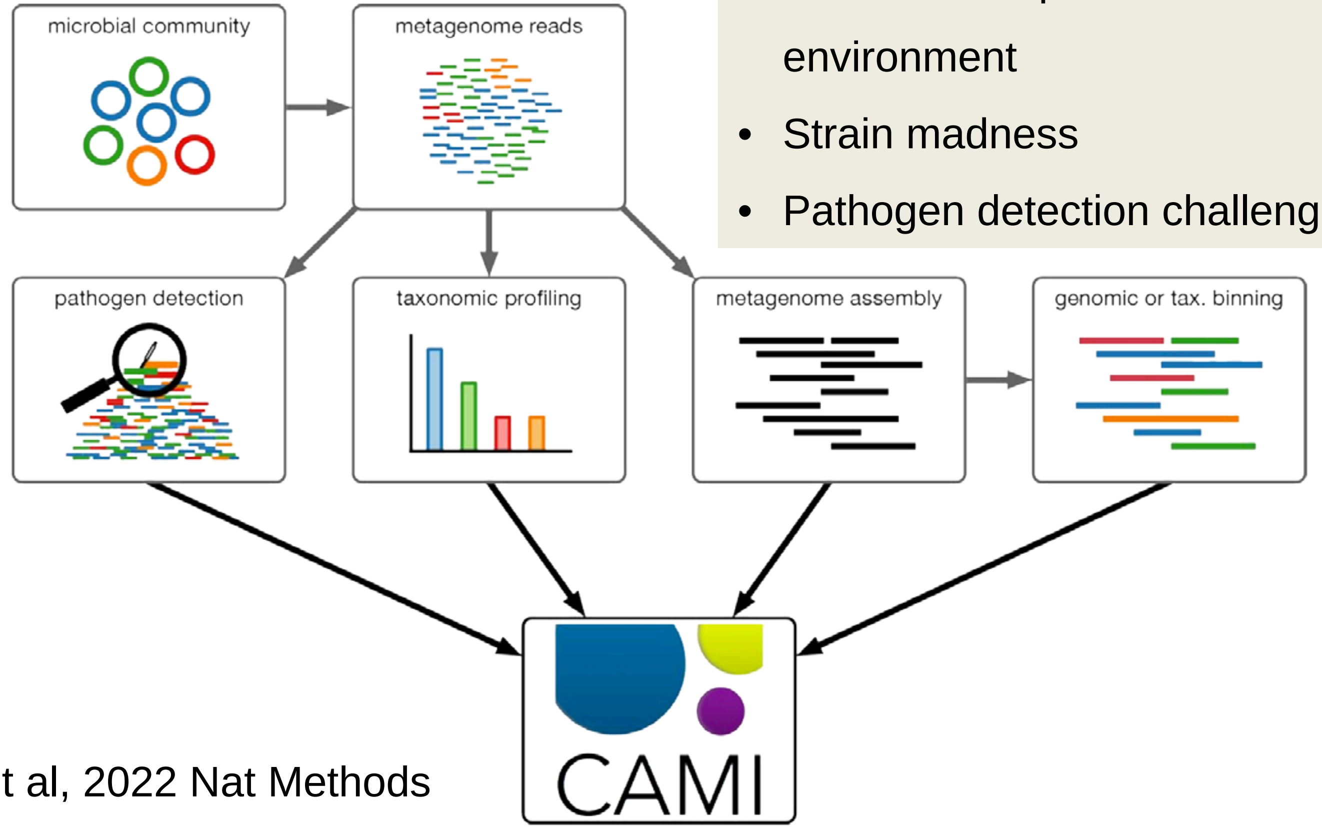
Alex Sczyrba

Thomas Rattei

Sczyrba et al, 2017 Nat Methods

CAMI 2 Challenge

- Short and long reads
- Marine/rhizosphere environment
- Strain madness
- Pathogen detection challenge



Meyer et al, 2022 Nat Methods

Summary and Outlook

- Personalized infection medicine (e.g. pathogen & AMR analyses)
- Molecular markers (e.g. AMR diagnostics, generic microbial phenotypes, microbiome-related diseases)
  - may indicate functional basis

- Shotgun metaOmics
  - increasing resolution of taxonomic analyses
  - functional markers, genomic context
  - improving sensitivity & turnaround time relative to culture-based analyses

Acknowledgements

- **E. Asgari** (BIFO / UC Berkeley)
- **A. Weimann**
- T.-H. Kuo
- **F. Meyer**
- T.R. Lesker
- P. Münch
- Z.-L. Deng, K. Hu
- **A. Fritz**
- N. Saffaei, B. Junk
- S. Häussler and lab (HZI)
- M. Mofrad (UC Berkeley)
- S. Szafranski (MHH)
- A. Sczyrba, J. Kalinowksi (U. Bielefeld)

CAMI II Contributors

F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, A. Gurevich, G. Robertson, T.-R. Lesker, M. Alser, D. Antipov, F. Beghini, D. Bertrand, J.J. Brito, C.T. Brown, J. Buchmann, A. Buluç, B. Chen, R. Chikhi, P.T.L.C. Clausen, A. Cristian, P.W. Dabrowski, A.E. Darling, R. Egan, E. Eskin, E. Georganas, E. Goltsman, M.A. Gray, L.H. Hansen, S. Hofmeyr, P. Huang, L. Irber, H. Jia, T.S. Jørgensen, S.D. Kieser, T. Klemetsen, A. Kola, M. Kolmogorov, A. Korobeynikov, J. Kwan, N. LaPierre, C. Lemaitre, C. Li, A. Limasset, F. Malcher-Miranda, S. Mangul, V.R. Marcelino, C. Marchet, P. Marijon, D. Meleshko, D.R. Mende, A. Milanese, N. Nagarajan, J. Nissen, S. Nurk, L. Oliker, L. Paoli, P. Peterlongo, V.C. Piro, J.S. Porter, S. Rasmussen, E.R. Rees, K. Reinert, B. Renard, E.M. Robertsen, G.L. Rosen, H.-J. Ruscheweyh, V. Sarwal, N. Segata, E. Seiler, L. Shi, F. Sun, S. Sunagawa, S.J. Sørensen, A. Thomas, C. Tong, M. Trajkovski, J. Tremblay, G. Uritskiy, R. Vicedomini, Zi. Wang , Zhe. Wang, Zho. Wang, A. Warren, N.P. Willassen, K. Yelick, R. You, G. Zeller, Z. Zhao, S. Zhu, J. Zhu, R. Garrido-Oter, P. Gastmeier, S. Hacquard, S. Häußler, A. Khaledi, F. Maechler, F. Mesny, S. Radutoiu, P. Schulze-Lefert, N. Smit, T. Strowig, A. Bremges, A. Sczyrba, A.C. McHardy

Universität Bielefeld

JGI DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY

University of Maryland 1856

Max-Planck-Institut für Züchtungsforschung
Max Planck Institute for Plant Breeding Research

>100 contributors from 77 institutions

HELMHOLTZ ZENTRUM FÜR INFEKTIONSFORSCHUNG

HEINRICH HEINE UNIVERSITÄT DÜSSELDORF

ETH zürich

UNIVERSITY OF TECHNOLOGY SYDNEY

PSC