Biodiversité et écologie à l'ère de la génomique environnementale

Les 1000 et une façons d'estimer la biodiversité à partir des données de métagénomique (hors méta-barcodes)

Faculté des Sciences
Aix*Marseille Université

Pascal Hingamp
Pascal.hingamp@univ-amu.fr

Métagénomes ?

"Génomique microbienne"

Isoler plancton unique

Culture axenique plancton

Extraire ADN Séquencer

Génome du plancton

Métagénomes ?

"Génomique microbienne"

Moins de 1% des microorganismes marins sont cultivables !

Isoler plancton unique

Culture axenique plancton

Extraire ADN Séquencer

Génome du plancton

Métagénomes ?

"Génomique microbienne"

Isoler plancton unique

Culture axenique plancton

Extraire ADN Séquencer

Génome du plancton

Moins de 1% des microorganismes marins sont cultivables !

"Génomique Environnementale"

Extraire ADN écosystème

Metagénome

= fragments d'ADN issus de multiples génomes présents dans l'écosystème

Les génomiques environnementales

Extraire
ADN
écosystème

Métagénome

→ quelles potentiel de fonctions biologiques est présent ?

Les génomiques environnementales

→ quelles espèces sont présentes ?

Metabarcodes

Amplifier gènes marqeurs

Extraire ADN écosystème

Métagénome

→ quelles potentiel de fonctions biologiques est présent ?

Les génomiques environnementales

→ quelles espèces sont présentes ?

Metabarcodes

Amplifier gènes marqeurs

Extraire ARN écosystème

→ quelles fonctions biologiques sont actives ?

Métatranscriptome

Extraire ADN écosystème

Métagénome

→ quelles potentiel de fonctions biologiques est présent ?

Les génomiques environnementales

→ quelles espèces sont présentes ?

Metabarcodes

Amplifier gènes marqeurs

Extraire ARN écosystème

→ quelles fonctions biologiques sont actives ?

Métatranscriptome

Tri cellulaire

Single Amplified Genomes (SAGs)

→ génome complet d'espèces de plancton représentatives

Extraire ADN écosystème

Métagénome

→ quelles potentiel de fonctions biologiques est présent ?

Les génomiques environnementales

Avantages :

→ ultra haut débit !

Metabarcodes

Amplifier gènes marqueurs

Inconvéniants :

→ marqueur universel aux 3 domaines ? & les virus ?...

→ biais de PCR

→ primers universels ?...

Extraire ADN écosystème

Métagénome

→ quelles potentiel de fonctions biologiques est présent ?

Extraire ARN écosystème

→ quelles fonctions biologiques sont actives ?

Métatranscriptome

Tri cellulaire

Single Amplified Genomes (SAGs)

→ génome complet d'espèces de plancton représentatives

environmental microbiology

Environmental Microbiology (2015) doi:10.1111/1462-2920.13023

Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples

Alma E. Parada, David M. Needham and Jed A. Fuhrman*

University of Southern California, Los Angeles, CA, USA

Summary

Microbial community analysis via high-throughput sequencing of amplified 16S rRNA genes is an essential microbiology tool. We found the popular primer pair 515F (515F-C) and 806R greatly underestimated (e.g. SAR11) or overestimated (e.g. Gammaproteobacteria) common marine taxa. We evaluated marine samples and mock communities (containing 11 or 27 marine 16S clones), showing alternative primers 515F-Y (5′-GTGYCAGCMGCCGCGGTAA) and 926R (5′-CCGYCAATTYMTTTRAGTTT) yield more accurate estimates of mock community abundances, produce longer amplicons that can differentiate taxa unresolvable with 515F-C/806R, and amplify eukaryotic 18S rRNA. Mock communities amplified with 515F-Y/926R yielded closer observed community composition versus expected ($r^2 = 0.95$) compared with 515F-Y/806R ($r^2 \sim 0.5$). Unexpectedly, biases with 515F-Y/806R against SAR11 in field samples (~4–10-fold) were stronger than in mock communities (~2-fold). Correcting a mismatch to Thaumarchaea in the 515F-C increased their apparent abundance in field samples, but not as much as using 926R rather than 806R. With plankton samples rich in eukaryotic DNA (> 1 µm size fraction), 18S sequences averaged ~17% of all sequences. A single mismatch can strongly bias amplification, but even perfectly matched primers can

the choice of primers to amplify 16S genes becomes crucial to take advantage of the sequence length and coverage made possible by improved sequencing technologies. In 2010, the Earth Microbiome Project (EMP) was established to create a catalogue of microbial diversity from habitats across the world (Gilbert *et al.*, 2010) with the goal of creating a database of microbial samples analysed exactly the same way to facilitate global comparisons. The EMP proposed standard primers and protocols to permit comparisons of diversity across samples. The primers 515F/806R were chosen to maximize the global coverage of Bacteria and Archaea while also providing polymerase chain reaction (PCR) products of suitable length for sequencing with available Illumina platforms (Caporaso *et al.*, 2011; 2012). Since it is commonly assumed that one mismatch in the middle of a primer will still allow binding and amplification of target templates, these primers appeared to have comprehensive coverage *in silico*. At around the same time, reviews of various group-specific and universal primers, such as Klindworth and colleagues (2013), performed mostly *in silico* analysis of hundreds of primers. Although Klindworth and colleagues (2013) did not examine the exact reverse primer used by EMP, they reported on similar primers that also had high apparent coverage if one mismatch is allowed. Thus, this 515F/806R primer pair seemed a reasonable choice.

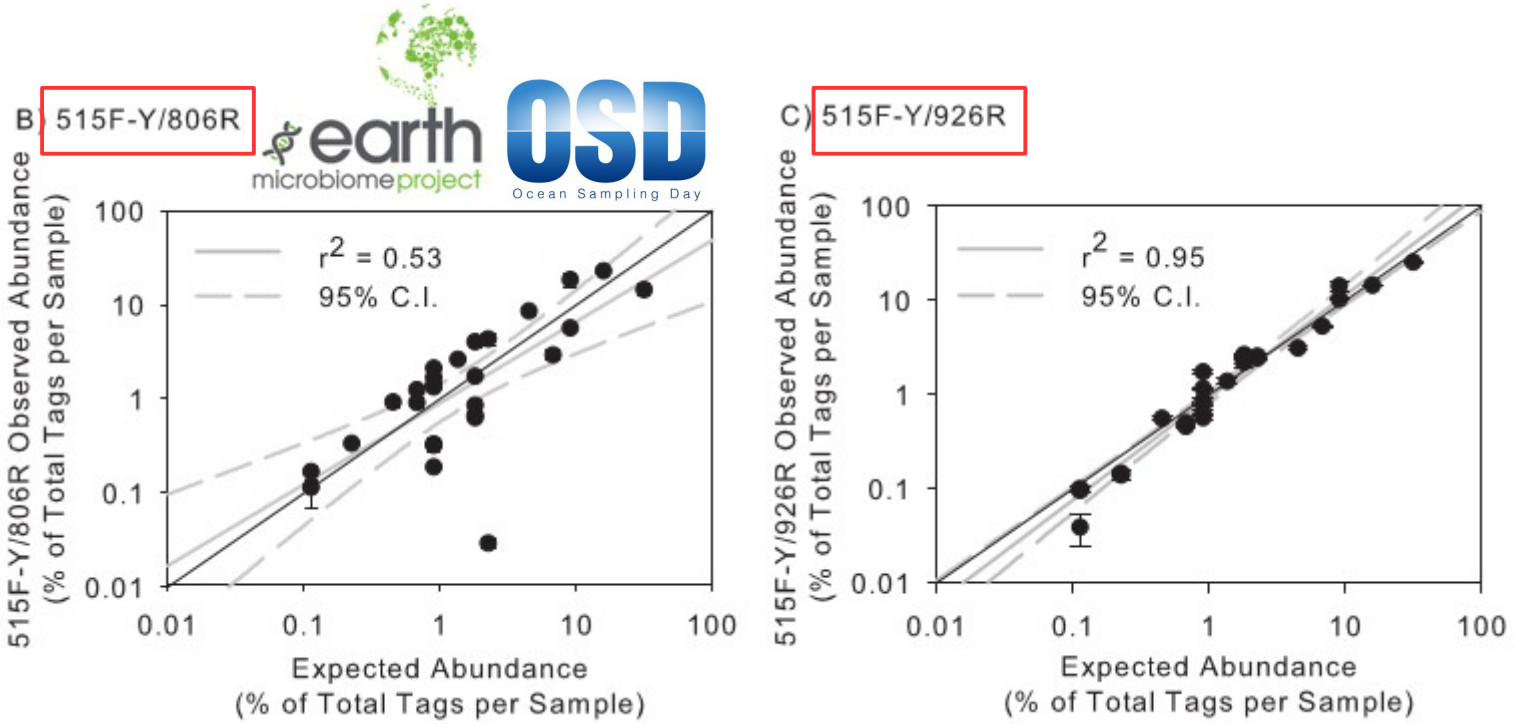
We submitted marine plankton samples from several (5–890 m) depths to the EMP and were surprised to find the SAR11 cluster, relating to the Candidate genus *Pelagibacter*, was poorly represented in the results (typically ~3%). Other studies of these samples taken from the San Pedro Ocean Time Series (SPOT) analysed via the

Estimer la biodiversité dans les métagénomes « whole shotgun »

environmental microbiology

Environmental Microbiology (2015) doi:10.1111/1462-2920.13023

Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples

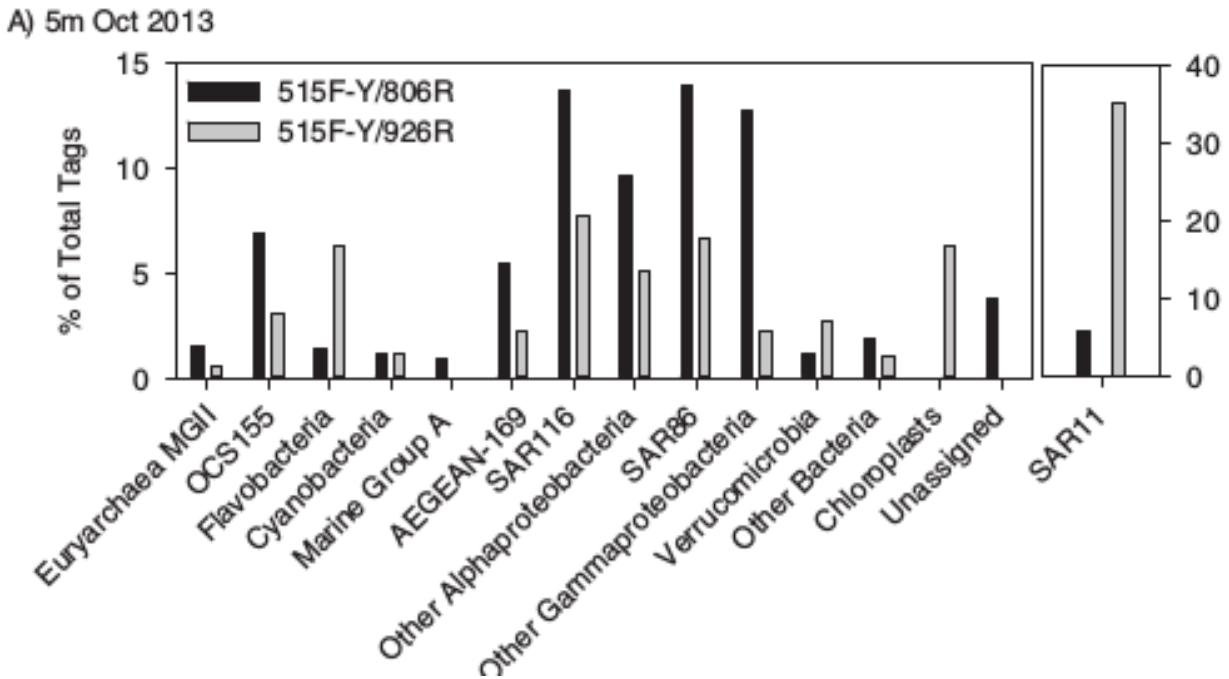


Estimer la biodiversité dans les métagénomes « whole shotgun »

environmental microbiology

Environmental Microbiology (2015) doi:10.1111/1462-2920.13023

Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples



Les génomiques environnementales

Avantages :

→ ultra haut débit !

Metabarcodes

Amplifier gènes marqueurs

Inconvéniants :

→ marqueur universel aux 3 domaines ? & les virus ?...

→ biais de PCR

→ primers universels ?

→ résolution faible, espèces cryptiques …

→ marqueurs « single copy » ?

Extraire ARN écosystème

→ quelles fonctions biologiques sont actives ?

Métatranscriptome

Tri cellulaire

Single Amplified Genomes (SAGs)

→ génome complet d'espèces de plancton représentatives

Extraire ADN écosystème

Métagénome

quelles potentiel de fonctions logiques est présent ?

Biodiversité + fonctions biologiques sans méta-barcodes...

→ quelles espèces sont présentes ?

Méta-barcodes

Amplifier gènes marqueurs

Extraire ARN écosystème

→ quelles fonctions biologiques sont actives ?

Métatranscriptome

Tri cellulaire

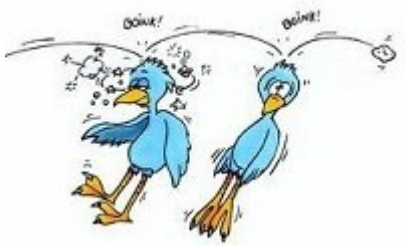
Single Amplified Genomes (SAGs)

→ génome complet d'espèces de plancton représentatives

Extraire ADN écosystème

Métagénome

→ quelles espèces sont présentes ?



→ quelles potentiel de fonctions biologiques est présent ?

Déluge de méthodes …

Estimer la biodiversité dans les métagénomes « whole shotgun »

OPEN ACCESS Freely available online

PLOS ONE

Received November 12, 2013; **Accepted** February 13, 2014; **Published** March 13, 2014

WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification

David Koslicki[1]*, Simon Foucart[2], Gail Rosen[3]

1 Mathematics Department, Oregon State University, Corvallis, Oregon, United States of America, **2** Department of Mathematics, University of Georgia, Athens, Georgia, United States of America, **3** Department of Electrical and Computer Engineering, Drexel University, Philadelphia, Pennsylvania, United States of America

Abstract

With the decrease in cost and increase in output of whole-genome shotgun technologies, many metagenomic studies are utilizing this approach in lieu of the more traditional 16S rRNA amplicon technique. Due to the large number of relatively

Published online 3 March 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 8 e73*
doi:10.1093/nar/gku169

MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences

Chengwei Luo[1,2], Luis M. Rodriguez-R[1,2] and Konstantinos T. Konstantinidis[1,2,*]

[1]Centre for Bioinformatics and Computational Genomics, and School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA and [2]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received September 30, 2013; Revised February 9, 2014; Accepted February 10, 2014

ABSTRACT

Determining the taxonomic affiliation of sequences assembled from metagenomes remains a major bottleneck that affects research across the fields of

sequence annotation and assembly (1,3–5). Perhaps most importantly, the taxonomic identity of most sequences assembled from a metagenomic dataset frequently remains elusive, making the exchange of information about an organism or a DNA sequence challenging when

Estimer la biodiversité dans les métagénomes « whole shotgun »

Wood and Salzberg *Genome Biology* 2014, **15**:R46
http://genomebiology.com/2014/15/3/R46

Genome Biology

METHOD Open Access

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]

Abstract

Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences.
utationally expensive, forcing

Wang *et al. BMC Genomics* 2014, **15**(Suppl 1):S12
http://www.biomedcentral.com/1471-2164/15/S1/S12

BMC Genomics

PROCEEDINGS Open Access

MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning

Yi Wang*, Henry Chi Ming Leung, Siu Ming Yiu, Francis Yuk Lun Chin*

From The Twelfth Asia Pacific Bioinformatics Conference (APBC 2014)
Shanghai, China. 17-19 January 2014

Estimer la biodiversité dans les métagénomes « whole shotgun »

PeerJ

Submitted 21 March 2013
Accepted 19 December 2013
Published 9 January 2014

Corresponding author
Aaron E. Darling,
aarondarling@ucdavis.edu

Academic editor
Ahmed Moustafa

PhyloSift: phylogenetic analysis of genomes and metagenomes

Aaron E. Darling[1,2], Guillaume Jospin[2], Eric Lowe[2], Frederick A. Matsen IV[5], Holly M. Bik[2] and Jonathan A. Eisen[3,4]

[1] ithree institute, University of Technology Sydney, Sydney, Australia
[2] Genome Center, University of California, Davis, CA, United States of America
[3] Department of Evolution and Ecology, University of California, Davis, CA, United States of America
[4] Department of Medical Microbiology and Immunology, University of California, Davis, CA, United States of America
[5] Fred Hutchinson Cancer Research Center, Seattle, WA, United States of America

ABSTRACT

Like all organisms on the planet, environmental microbes are subject to the forces of molecular evolution. Metagenomic sequencing provides a means to access the DNA sequence of uncultured microbes. By combining DNA sequencing of microbial
odeling and phylogenetic analysis we might obtain

Garcia-Etxebarria *et al. BMC Bioinformatics* 2014, **15**:90
http://www.biomedcentral.com/1471-2105/15/90

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Consistency of metagenomic assignment programs in simulated and real data

Koldo Garcia-Etxebarria, Marc Garcia-Garcerà and Francesc Calafell*

Abstract

Background: Metagenomics is the genomic study of uncultured environmental samples, which has been greatly facilitated by the advent of shotgun-sequencing technologies. One of the main focuses of metagenomics is the discovery of previously uncultured microorganisms, which makes the assignment of sequences to a particular taxon a challenge and a crucial step. Recently, several methods have been developed to perform this task, based

Estimer la biodiversité dans les métagénomes « whole shotgun »

Chapter 15

Ian T. Paulsen and Andrew J. Holmes (eds.), *Environmental Microbiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1096, DOI 10.1007/978-1-62703-712-9_15, © Springer Science+Business Media, LLC 2014

Metagenomics Using Next-Generation Sequencing

Lauren Bragg and Gene W. Tyson

Abstract

Traditionally, microbial genome sequencing has been restricted to the small number of species that can be grown in pure culture [1]. The progressive development of culture-i
15 years now allows researchers to sequence microbial communities dir
This approach is commonly referred to as "metagenomics" or "commu

Table 3
Software for "binning" reads/contigs

Tool	Sequence-binning approach	Supervised?
Carma3 [92]	Similarity	Yes
CompostBin [93]	Composition	No
MEGAN4 [94]	Similarity	Yes
PhyloPythiaS [95]	Composition	Yes
PhymmBL [87]	Composition	Yes
SGOM [96]	Composition	Semi-supervised
TACOA [97]	Composition	Yes
TaxSOM [98]	Composition	Semi-supervised
Taxy [99]	Composition	Yes
Tetra [88]	Composition	No
TreePhyla [100]	Composition	Yes

Peabody *et al. BMC Bioinformatics* (2015) 16:363
DOI 10.1186/s12859-015-0788-5

BMC Bioinformatics

RESEARCH Open Access

Evaluation of shotgun metagenomics sequence classification methods using *in silico* and *in vitro* simulated communities

Michael A. Peabody, Thea Van Rossum, Raymond Lo and Fiona S. L. Brinkman*

Abstract

Background: The field of metagenomics (study of genetic material recovered directly from an environment) has grown rapidly, with many bioinformatics analysis methods being developed. To ensure appropriate use of such methods, robust comparative evaluation of their accuracy and features is needed. For taxonomic classification of sequence reads, such evaluation should include use of clade exclusion, which better evaluates a method's accuracy when identical sequences are not present in any reference database, as is common in metagenomic analysis. To date, relatively small evaluations have been performed, with evaluation approaches like clade exclusion limited to assessment of new methods by the authors of the given method. What is needed is a rigorous, independent comparison between multiple major methods, using the same *in silico* and *in vitro* test datasets, with and without approaches like clade exclusion, to better characterize accuracy under different conditions.

Results: An overview of the features of 38 bioinformatics methods is provided, evaluating accuracy with a focus on 11 programs that have reference databases that can be modified and therefore most robustly evaluated with clade exclusion. Taxonomic classification of sequence reads was evaluated using both *in silico* and *in vitro* mock bacterial communities. Clade exclusion was used at taxonomic levels from species to class—identifying how well methods perform in progressively more difficult scenarios. A wide range of variability was found in the sensitivity, precision, overall accuracy, and computational demand for the programs evaluated. In experiments where distilled water was spiked with only 11 bacterial species, frequently dozens to hundreds of species were falsely predicted by the most popular programs. The different features of each method (forces predictions or not, etc.) are summarized, and additional analysis considerations discussed.

Conclusions: The accuracy of shotgun metagenomics classification methods varies widely. No one program clearly outperformed others in all evaluation scenarios; rather, the results illustrate the strengths of different methods for different purposes. Researchers must appreciate method differences, choosing the program best suited for their particular analysis to avoid very misleading results. Use of standardized datasets for method comparisons is encouraged, as is use of mock microbial community controls suitable for a particular metagenomic analysis.

Keywords: Metagenomics, Evaluation, Accuracy, Comparison, Taxonomic classification

Estimer la biodiversité dans les métagénomes « whole shotgun »

Peabody *et al. BMC Bioinformatics* (2015) 16:363
DOI 10.1186/s12859-015-0788-5

RESEARCH

Evaluation of shotgu sequence classificat *in silico* and *in vitro*

Michael A. Peabody, Thea Van Rossum, Raym

Abstract

Background: The field of metagenomics (s has grown rapidly, with many bioinformati of such methods, robust comparative evalu classification of sequence reads, such evalu a method's accuracy when identical sequer metagenomic analysis. To date, relatively sr like clade exclusion limited to assessment needed is a rigorous, independent compari *in vitro* test datasets, with and without app different conditions.

Results: An overview of the features of 38 focus on 11 programs that have reference evaluated with clade exclusion. Taxonomic and *in vitro* mock bacterial communities. C class—identifying how well methods perfo variability was found in the sensitivity, prec programs evaluated. In experiments where dozens to hundreds of species were falsely each method (forces predictions or not, et

Conclusions: The accuracy of shotgun me clearly outperformed others in all evaluatio methods for different purposes. Researcher suited for their particular analysis to avoid comparisons is encouraged, as is use of m metagenomic analysis.

Keywords: Metagenomics, Evaluation, Accu

Table 2 List of metagenomics sequence classification methods and their characteristics sorted by class of method

Method name	Class of method	Sequence alignment method/ Composition method	Standalone[a]/Web server	Most recent year published (first time published)[b]	Functional classification if applicable	References	Number of citations[c]
MEGAN4	Similarity	MEGABLAST, BLASTN, BLASTX, RAPSEARCH2 [32] / N/A	Yes/No	2011 (2007)	KEGG, SEED	[15, 29, 45–47]	1089
MG-RAST	Similarity	BLASTN, BLAT / N/A	No/Yes	2008	SEED, NOG, COG, KEGG	[48]	691
CAMERA	Similarity	All 6 BLAST programs / N/A	No/Yes	2007 (2011)	Pfam, TIGRFAM, COG, KOG, PRK	[49, 50]	324
CARMA3	Similarity	BLASTX, HMMER3 [51] / N/A	Yes/Yes	2011 (2008)	GO	[41, 52, 53]	201
WebMGA	Similarity	FR-HIT [54] / N/A	No/Yes	2013	Pfam, TIGRFAM, COG, KOG, PRK, GO	[55]	54
DiScRIBinATE (SOrt-ITEMS)[d]	Similarity	BLASTX, RAPSEARCH2 / N/A	Yes/No	2010 (2009)	N/A	[31, 56]	48
Ray Meta	Similarity	Exact match k-mers / N/A	Yes/No	2012	N/A	[57]	34
Kraken	Similarity	Exact match k-mers / N/A	Yes/No	2014	N/A	[28]	15
RTM	Similarity	k-mers / N/A	Yes/Yes	2012	KEGG	[58]	12
Genometa	Similarity	Bowtie [59], BWA [60] / N/A	Yes/No	2012	N/A	[61]	7
LMAT	Similarity	Exact match k-mers / N/A	Yes/No	2013	N/A	[62]	6
Sequedex	Similarity	Exact match k-mers / N/A	Yes/No	2012	N/A	[63]	5
MetaBin	Similarity	BLASTX, BLAT / N/A	Yes/Yes	2012	COG	[64]	4
TAMER	Similarity	MEGABLAST / N/A	Yes/No	2012	N/A	[65]	4
metaBEETL	Similarity	Direct comparison of compressed text indices / N/A	Yes/No	2013	N/A	[7]	2
SPANNER	Similarity	BLASTP / N/A	Yes/No	2013	N/A	[66]	2
GOTTCHA	Similarity	BWA / N/A	Yes/No	2015	N/A	[67]	0
CLARK	Similarity	k-mers / N/A	Yes/No	2015	N/A	[68]	0
MLTreeMap	Marker	BLASTX / N/A	Yes/Yes	2010 (2007)	4 Enzyme families	[69, 70]	206
AMPHORA2	Marker	HMMER3 / N/A	Yes/Yes	2012 (2008)	N/A	[13, 71, 72]	190
MetaPhlAn	Marker	MEGABLAST, Bowtie2 [73] / N/A	Yes/Yes	2012	N/A	[11]	114
MetaPhyler	Marker	BLASTN, BLASTX / N/A	Yes/No	2011	N/A	[30]	42
mOTU	Marker	HMMER3 / N/A	Yes/Yes	2013	N/A	[19]	24
Phylosift	Marker	LAST, HMMER3 / N/A	Yes/No	2014	N/A	[14]	18
phymmBL	Hybrid	MEGABLAST / IMM	Yes/No	2011 (2009)	N/A	[6, 74]	182
RITA	Hybrid	Pipeline of BLAST variations / NB	Yes/Yes	2012 (2011)	N/A	[75, 76]	38
SPHINX	Hybrid	BLASTX / k-means	No/Yes	2010	N/A	[10]	17
TaxyPro	Hybrid	CoMet web server / Mixture model	Yes/No	2013	Pfam	[77]	3
TWARIT	Hybrid	BWA short read alignment [60] / k-means	No/Yes	2012	N/A	[78]	2
PhyloPythiaS	Composition	N/A / SVM	Yes/Yes	2011 (2007)	N/A	[30, 79, 80]	269
TACOA	Composition	N/A / k-NN	Yes/No	2009	N/A	[33]	65
NBC	Composition	N/A / NB	Yes/Yes	2011 (2008)	N/A	[81, 82]	35
RAIphy	Composition	N/A / RAI	Yes/No	2011	N/A	[83]	18
ClaMS	Composition	N/A / DBC signature	Yes/No	2011	N/A	[84]	10
INDUS	Composition	N/A / k-means	No/Yes	2011	N/A	[85]	8
TAC-ELM	Composition	N/A / Neural Network	Yes/No	2012	N/A	[86]	5
MetaCV	Composition	N/A / CV	Yes/No	2013	KEGG	[87]	4
GSTaxClassifier	Composition	N/A / Bayesian	No/No	2010	N/A	[88]	2

Méthodes disponibles

1 – Biais compositionnels de l'ADN dans les fragments metaG
(« *binning* » = regroupement par classes)

2 – Extraction des gènes marqueurs &
comparaison à une base de données de référence

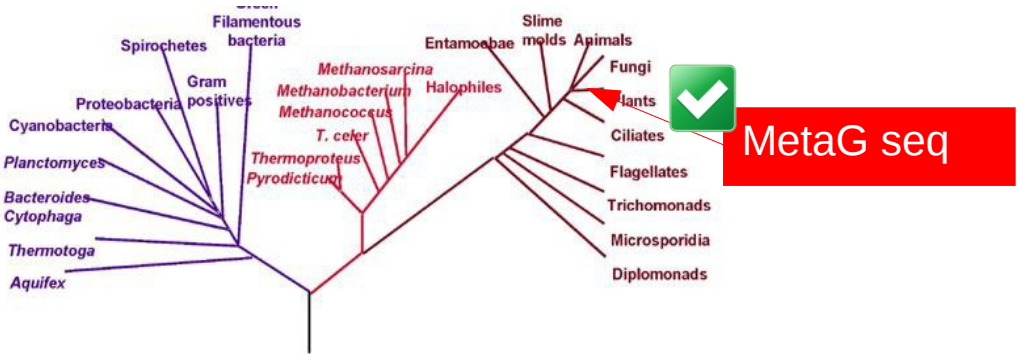
3 - Assignation taxonomique aux fragments MetaG par homologie à une base de données de référence

Dépendance à des données de références externes

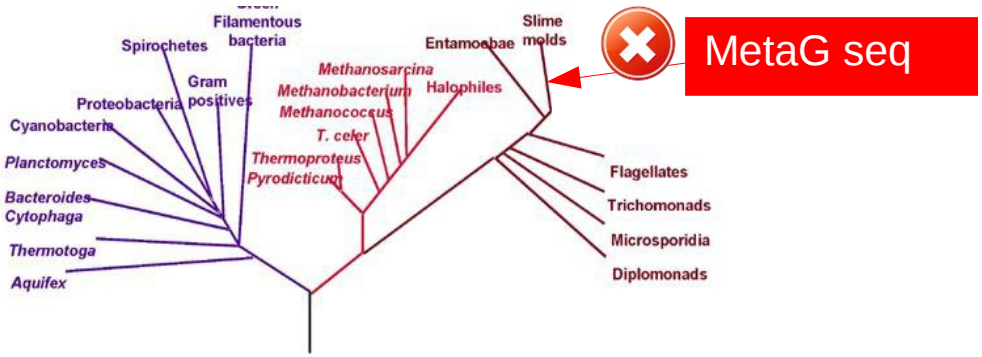
➔ Méthodes supervisées : appuyées sur BD de référence (génomes de référence, ou phylogénies de référence)

Risque : erreurs d'assignation s'il manque un clade dans la banque de référence !

Arbre de la vie « vrai »



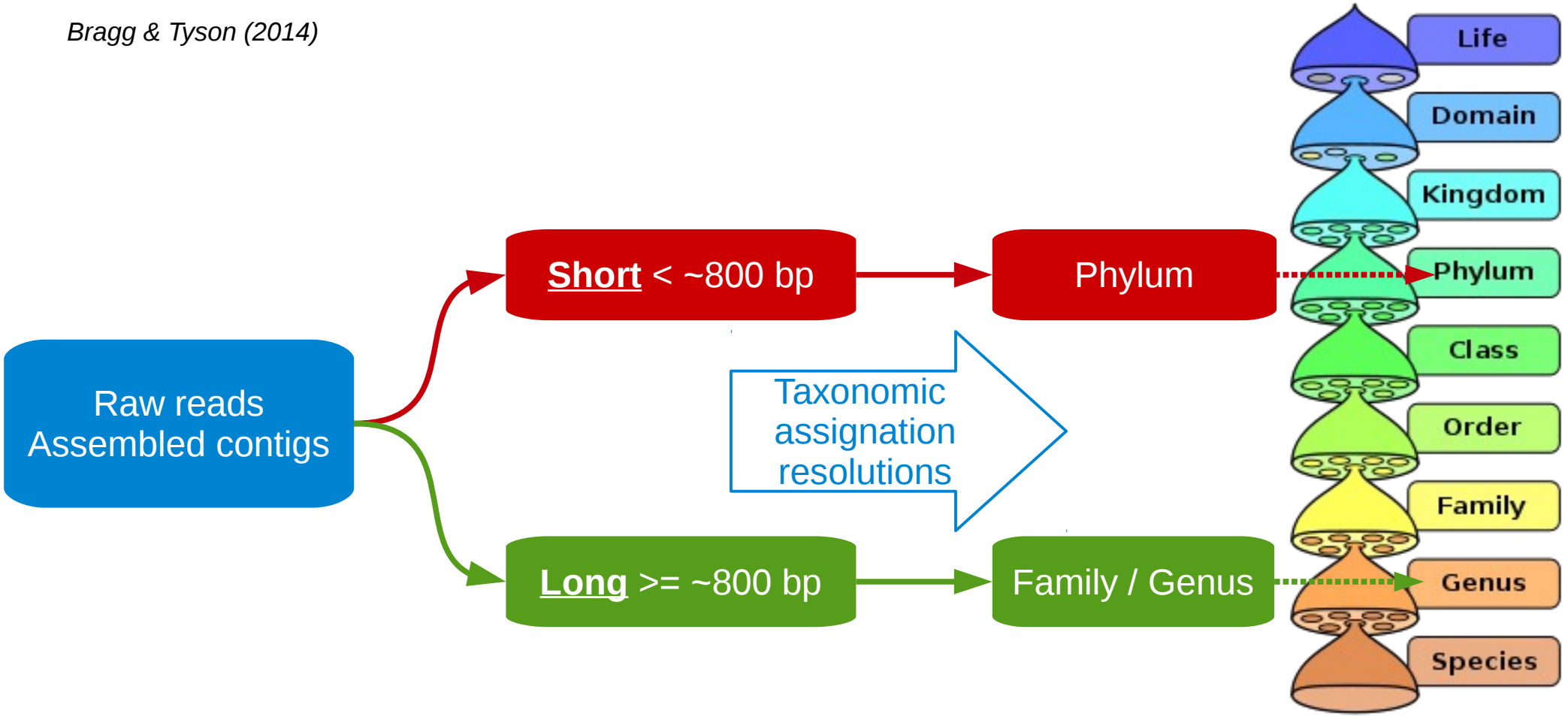
Arbre de la vie « disponible », 1 clade manquant



➔ Méthodes non supervisées : références non requises, entrainement sur les données elles même (ex biais compositionnels)

Résolution des assignations taxonomiques

Bragg & Tyson (2014)



Méthodes disponibles

1 – Biais compositionnels de l'ADN dans les fragments metaG
(« *binning* » = regroupement par classes)

2 – Extraction des gènes marqueurs &
comparaison à une base de données de référence

3 - Assignation taxonomique aux fragments MetaG par homologie à une base de données de référence

1 – biais compositionnels

Exploitation des caractèristiques intrasèques des reads (souvent combinées):

- Taux de GC
- Couverture
- Utilisation des codons (« *codon usage* »)
- Fréquences d'oligonucleotides (k-mers)

→ pour regrouper (classifier) les séquences dans des classes (« *bins* ») qui correspondent (on espère) à des espèces.

1 – biais compositionnels

OPEN ACCESS Freely available online

PLOS ONE

WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification

WGSQuikr
Koslicki *et al.* (2014)

David Koslicki[1]*, Simon Foucart[2], Gail Rosen[3]

1 Mathematics Department, Oregon State University, Corvallis, Oregon, United States of America, **2** Department of Mathematics, University of Georgia, Athens, Georgia, United States of America, **3** Department of Electrical and Computer Engineering, Drexel University, Philadelphia, Pennsylvania, United States of America

Abstract

With the decrease in cost and increase in output of whole-genome shotgun technologies, many metagenomic studies are utilizing this approach in lieu of the more traditional 16S rRNA amplicon technique. Due to the large number of relatively short reads output from whole-genome shotgun technologies, there is a need for fast and accurate short-read OTU classifiers. While there are relatively fast and accurate algorithms available, such as MetaPhlAn, MetaPhyler, PhyloPythiaS, and PhymmBL, these algorithms still classify samples in a read-by-read fashion and so execution times can range from hours to days on large datasets. We introduce WGSQuikr, a reconstruction method which can compute a vector of taxonomic assignments and their proportions in the sample with remarkable speed and accuracy. We demonstrate on simulated data that WGSQuikr is typically more accurate and up to an order of magnitude faster than the aforementioned classification algorithms. We also verify the utility of WGSQuikr on real biological data in the form of a mock community. WGSQuikr is a Whole-Genome Shotgun QUadratic, Iterative, K-mer based Reconstruction method which extends the previously introduced 16S rRNA-based algorithm Quikr. A MATLAB implementation of WGSQuikr is available at: http://sourceforge.net/projects/wgsquikr.

Citation: Koslicki D, Foucart S, Rosen G (2014) WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification. PLoS ONE 9(3): e91784. doi:10.1371/journal.pone.0091784

Editor: Mark R. Liles, Auburn University, United States of America

Received November 12, 2013; **Accepted** February 13, 2014; **Published** March 13, 2014

1 – biais compositionnels

WGSQuikr
Koslicki *et al.* (2014)

Problème : énormes jeux de données de « *short reads* »

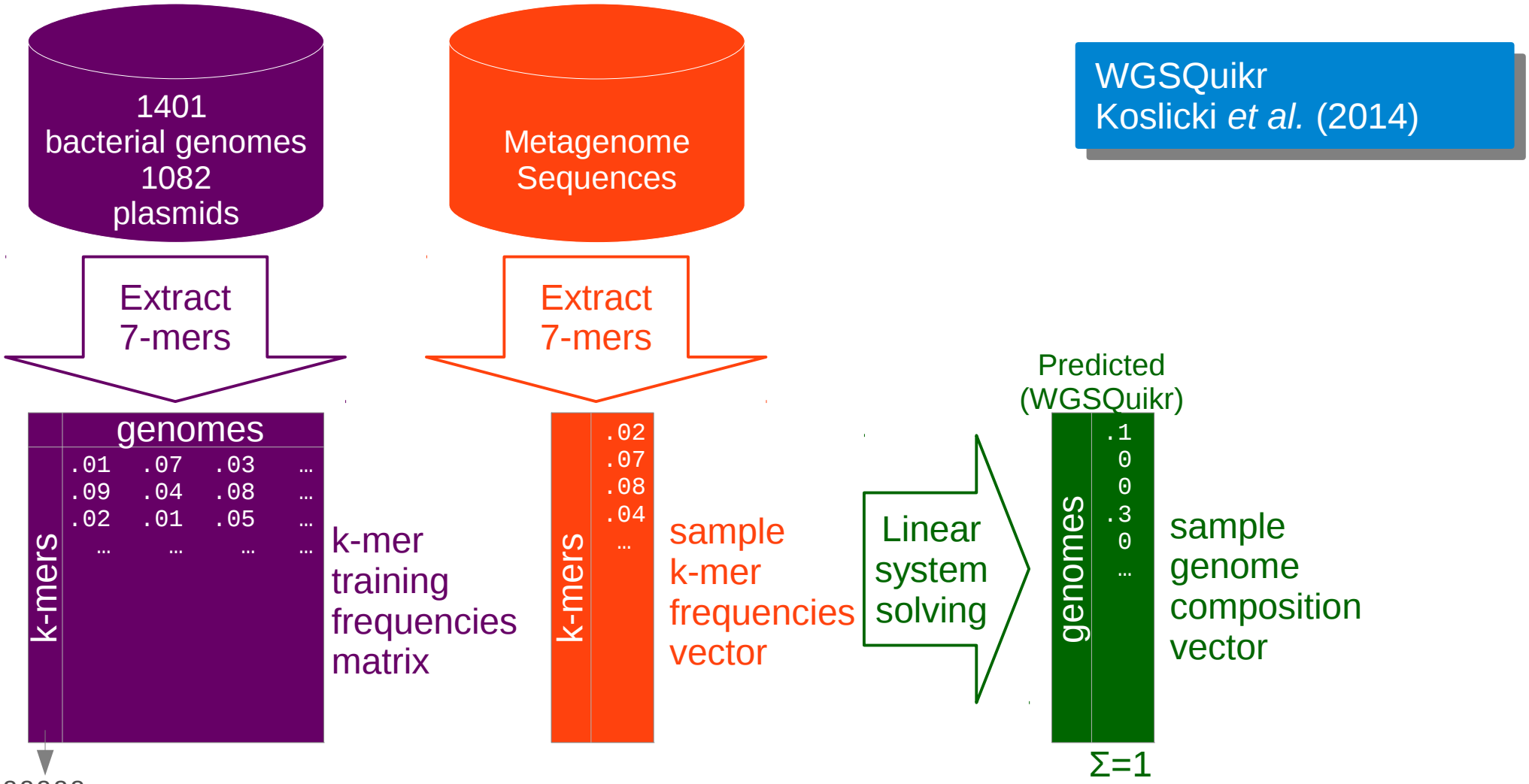
→ 15 millions x 36 bp pour Illumina **MiSeq**
→ 70 millions x 200 bp pour **Ion Torrent**
→ 3 milliards x 100 bp pour Illumina **HiSeq**

Solution : logiciels rapides, mais méthodes actuelles souvent lentes...

Exemple : mini jeu de données de 70,000 *reads* de 300 bp
→ 8 heures MetaPhyler
→ 4 jours PhymmBL

WGSQuikr : « milliards de reads sur un PC portable en moins d'une heure! »

1 – biais compositionnels



Postulat: l'échantillon metagenomique ne contient pas de taxa nouveaux absents de la BD d'entrainement (tout au moins pas trop éloignés, approx. niveau genre)

1 – biais compositionnels

WGSQuikr
Koslicki *et al.* (2014)

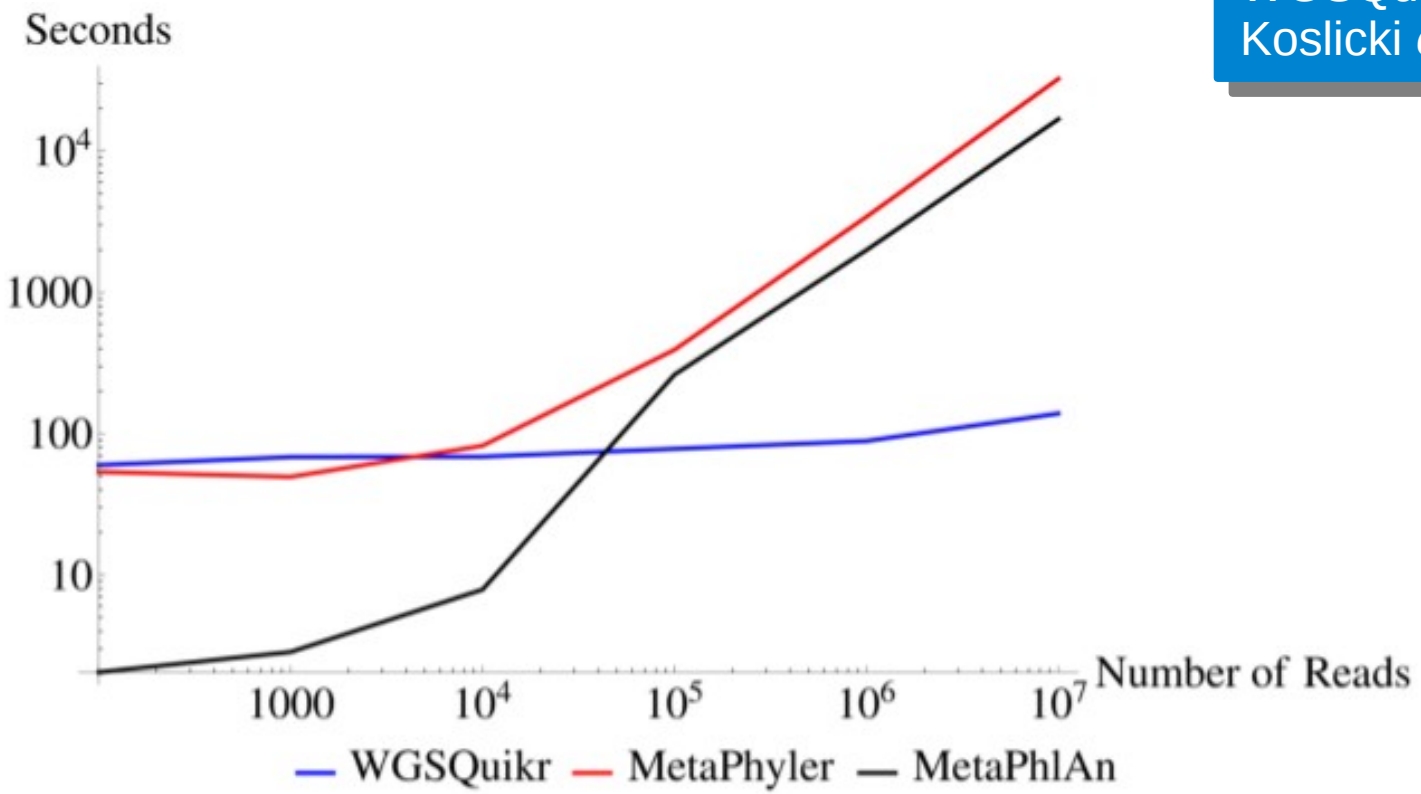
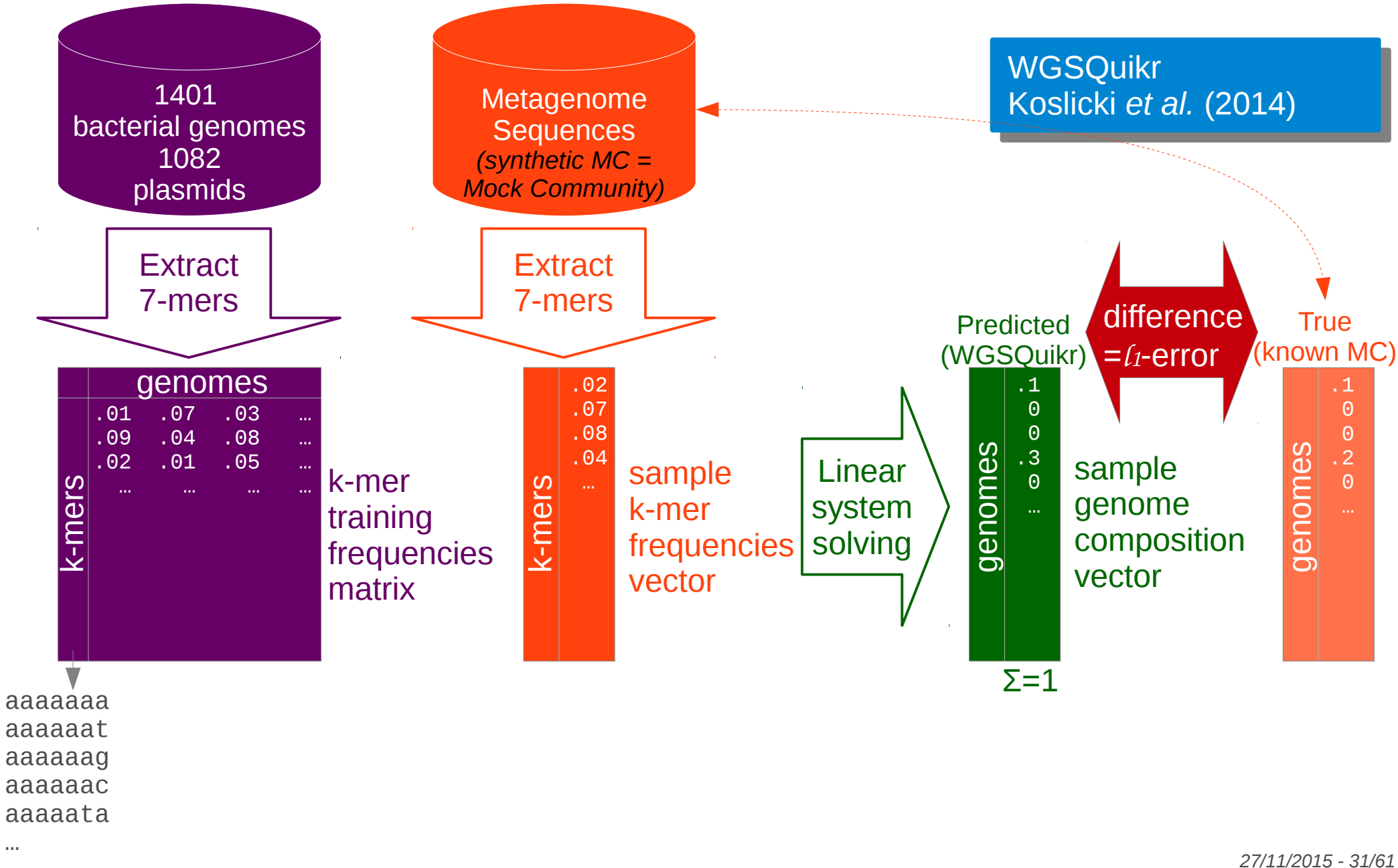


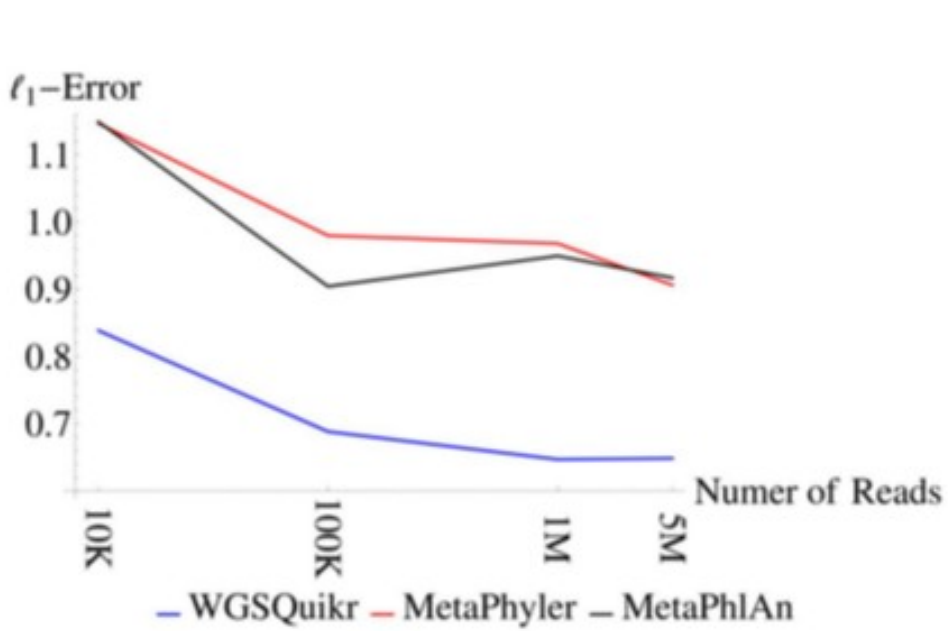
Figure 1. Log-log plot of number of reads versus execution time (seconds) for WGSQuikr, MetaPhyler and MetaPhlAn.

1 – biais compositionnels



1 – biais compositionnels

WGSQuikr
Koslicki *et al.* (2014)



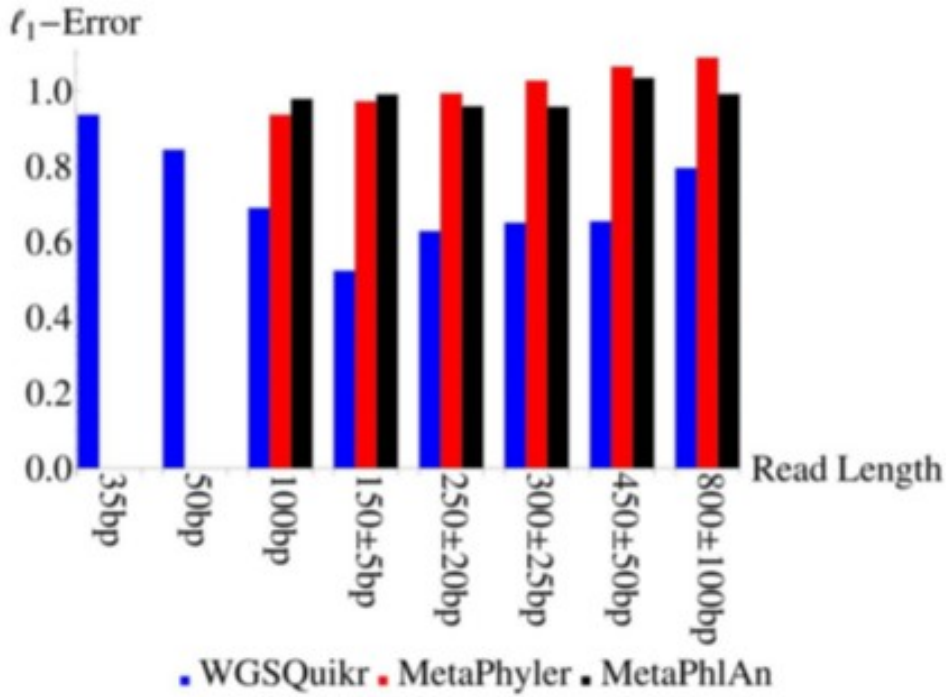


Figure 4. Mean ℓ_1-error at the genus level as a function of simulated dataset parameters for each method. MetaPhyler and MetaPhlAn failed to run on the datasets where reads were 35 bp or 50 bp in length.

1 – biais compositionnels

WGSQuikr
Koslicki *et al.* (2014)

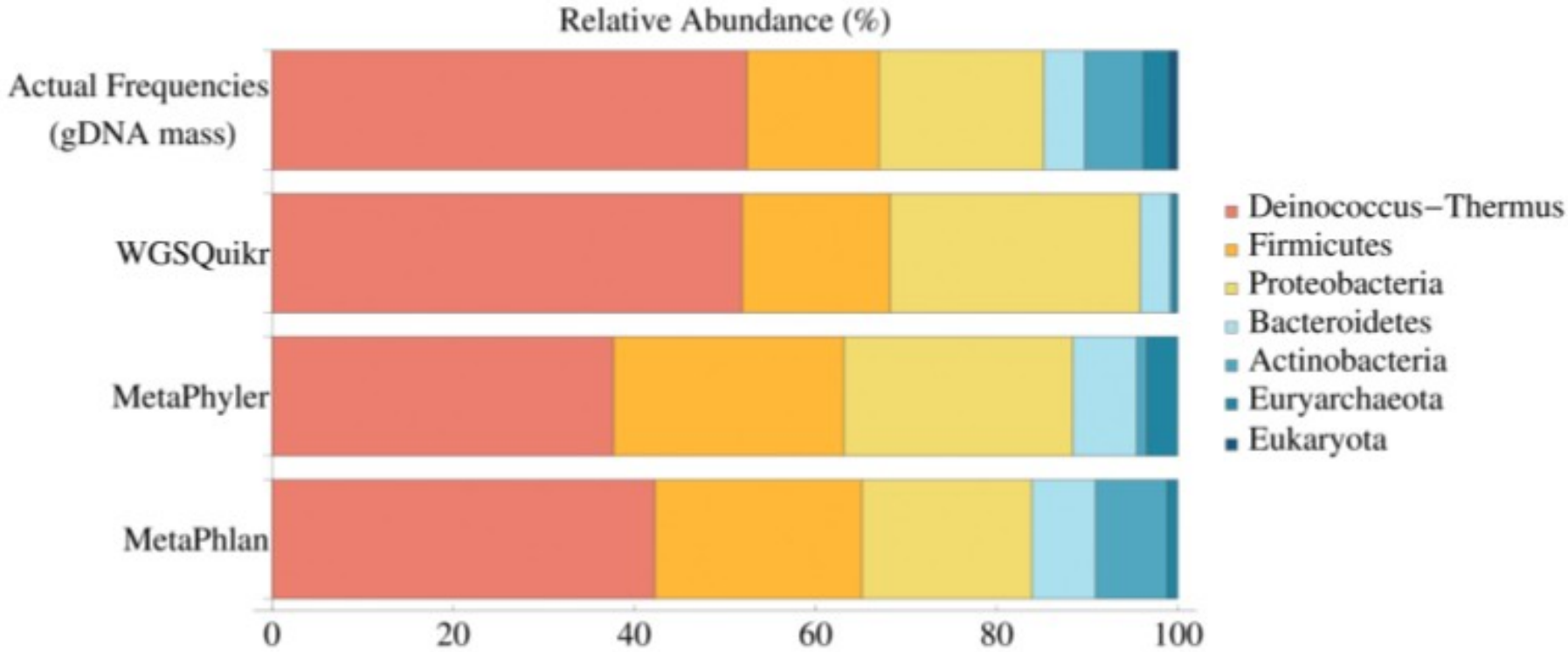


Figure 5. Relative abundances at the phylum level for reconstructions of organisms in the mock community.
doi:10.1371/journal.pone.0091784.g005

Méthodes disponibles

1 – Biais compositionnels de l'ADN dans les fragments metaG
(« binning » = regroupement par classes)

2 – Extraction des gènes marqueurs &
comparaison à une base de données de référence

3 - Assignation taxonomique aux fragments MetaG par homologie à une base de données de référence

2 – extraire des gènes marqueurs

- Requiert une base de données de gènes marqueurs de référence (supervisée)
- Approche 1 : extraire gènes 16S des metaG
- Approche 2 : extraires autres (non 16S) gènes marqueurs des metaG

2 – extraire des gènes marqueurs

- Requiert une base de données de gènes marqueurs de référence (supervisée)
- Approche 1 : extraire gènes 16S des metaG
- Approche 2 : extraires autres (non 16S) gènes marqueurs des metaG

2 – extraire des gènes marqueurs

- Approche 1 : extraire gènes 16S des metaG

→ infaisable avant les (ultra) HTS :

ex GOS (2007, Sanger) = **7.7M** reads
→ **4125** rRNA 16S genes...

ex Ghai et al. (2012, 454) = **1.5M** reads
→ **773** rRNA 16S genes...

→ possible avec les (ultra) HTS, ex Illumina :

150.000 reads ARNr 16S (0.1%) pour **150M** reads
→ run < 1000 €...
→ problème : petits reads (~50-300 bp)...

2 – extraire des gènes marqueurs

metaG **16S amplicons**

		Metagenome		SSU rDNA amplicons					
		ILM	454	V12	V13	V4	V35	V69	V48
Bacteria									
Aquificae	*Hydrogenobaculum sp. Y04AAS1*	4.9	2.4	0.0	0.3	1.8	0.2	0.1	1.0
	Persephonella marina	2.0	2.9	0.0	1.6	1.6	0.6	ND	2.0
	Sulfurihydrogenibium sp. YO3AOP1	1.3	2.5	0.0	1.4	0.7	0.6	ND	1.3
	S. yellowstonense	2.2	4.3	0.0	1.4	0.7	0.6	0.0	1.2
Thermotogae	*Thermotoga neapolitana*	1.3	1.3	1.2	0.3	1.2	0.5	ND	0.7
	Thermotoga petrophila	0.3	0.5	1.3	0.3	1.2	0.5	ND	0.7
	Thermotoga sp. RQ2	0.7	1.0	1.2	0.3	1.2	0.4	ND	0.7
Thermi/ Deinococcl	*Deinococcus radiodurans*	2.5	1.3	0.5	0.4	0.6	0.6	0.6	0.4
	Thermus thermophilus	1.3	0.1	1.1	0.5	1.9	0.7	0.5	0.3
Dictyoglomi	*Dictyoglomus turgidum*	2.7	4.3	2.2	0.4	7.9	2.4	0.1	0.8
Actinobacteria	*Salinispora arenicola*	1.0	0.3	0.5	1.2	0.1	1.0	0.4	0.3
	Salinispora tropica	1.2	0.4	0.5	1.2	0.1	1.0	0.4	0.3
Chloroflexi	*Chloroflexus aurantiacus*	1.8	1.1	2.5	2.2	1.7	2.7	0.1	0.4
	Herpetosiphon aurantiacus	1.6	1.6	2.5	1.0	1.0	0.7	0.6	1.3
Cyanobacteria	*Nostoc sp. PCC 7120*	1.5	2.1	1.3	1.6	1.3	1.8	5.9	2.2
Bacteroides	*Bacteroides thetaiotaomicron*	1.1	1.8	2.2	0.9	0.0	1.2	0.1	1.3
	Bacteroides vulgatus	1.4	1.9	2.1	0.7	ND	1.4	0.1	1.2
	Porphyromonas gingivalis	1.0	0.7	0.9	0.5	0.0	1.0	0.1	0.8
Chlorobi	*Chlorobium limicola*	0.8	1.0	1.0	1.7	0.4	0.3	0.8	0.2
	Chlorobium phaeobacteroides	1.0	1.5	0.7	1.3	0.7	0.9	1.2	0.2
	Chlorobium phaeovibrioides	0.5	0.3	0.6	1.2	0.6	0.7	0.9	0.1
	Chlorobium tepidum	1.3	1.4	1.3	1.9	0.6	1.1	1.1	0.2
	Pelodictyon phaeoclathratiforme	1.4	1.2	0.6	1.1	0.7	0.8	1.3	0.2
Firmicutes	*Caldicellulosiruptor bescii*	1.8	2.2	3.1	0.8	2.7	1.8	0.8	0.8
	Caldicellulosiruptor saccharolyticus	1.8	3.0	4.4	1.2	4.5	2.1	1.9	1.4
	Clostridium thermocellum	0.5	0.7	1.8	1.0	0.4	1.5	0.6	0.9
	Enterococcus faecalis	1.1	2.1	1.2	1.2	0.5	1.6	8.5	1.5
	Thermoanaerobacter pseudeth	1.2	2.1	0.9	0.9	1.1	2.5	1.5	0.5
Fusobacteria	*Fusobacterium nucleatum*	1.0	2.1	2.2	1.8	0.6	2.2	0.1	2.1
Verrucomicrobia	*Akkermansia muciniphila*	1.2	1.2	0.0	1.6	2.0	0.1	2.0	1.5
Gemmatimonadetes	*Gemmatimonas aurantiaca*	2.2	1.3	0.6	0.3	1.6	0.8	0.9	1.1
Planctomycetes	*Rhodopirellula baltica*	1.1	1.2	0.0	0.7	0.5	0.2	0.0	1.0
Spirochaetae	*Treponema denticola*	1.2	1.6	2.6	0.8	0.8	1.4	0.0	0.9
Acidobacteria	*Acidobacterium capsulatum*	0.8	0.5	0.7	1.3	1.0	0.9	2.4	0.8
Proteobacteria: Alpha	*Ruegeria pomeroyi*	0.5	0.3	0.8	1.0	0.6	1.0	0.7	0.9
	Sulfitobacter sp. EE-36	1.1	1.0	0.7	1.0	0.8	1.0	0.3	0.6
	Sulfitobacter sp. NAS-14.1	0.3	0.3	0.7	1.1	0.8	1.0	0.3	0.6
	Zymomonas mobilis	1.3	2.0	ND	1.2	1.6	0.5	0.3	1.5
Beta	*Bordetella bronchiseptica*	0.9	0.5	1.1	0.7	1.3	1.8	0.5	5.3
	Burkholderia xenovorans	1.2	0.9	0.7	0.7	1.0	0.9	0.1	1.5
	Leptothrix cholodnii	1.6	0.8	0.5	0.6	0.7	0.9	0.1	2.4
	Nitrosomonas europaea	0.9	1.1	0.8	0.7	2.6	1.0	0.2	1.1
Gamma	*Shewanella baltica OS185*	0.9	1.6	0.8	0.5	1.1	1.2	1.3	1.7
	Shewanella baltica OS223	1.1	1.8	0.8	0.4	1.1	1.2	1.3	1.6
Delta	*Desulfovibrio piger*	0.4	0.2	0.0	1.0	1.6	0.4	3.5	2.1
	Desulfovibrio vulgaris	1.3	0.5	ND	0.7	0.9	0.3	2.9	1.4
	Geobacter sulfurreducens	2.7	0.8	1.5	1.0	0.6	2.1	0.7	1.9
Epsilon	*Wolinella succinogenes*	1.1	0.9	1.0	1.2	0.6	0.4	2.1	1.1

Synthetic genome mix :
observed / expected abundance
Shakya *et al.* (2013)

metaG **16S amplicons**

		Metagenome		SSU rDNA amplicons			
		ILM	454	V13	V4	V4a	V48
Archaea							
Nanoarchaeaota	*Nanoarchaeum equitans*	0.6	0.1	0.1	1.0	0.4	0.3
Crenarchaeota: Thermoprotei	*Ignicoccus hospitalis*	1.8	0.9	0.2	0.5	1.2	1.8
	Pyrobaculum aerophilum	1.5	1.6	x	5.9	4.3	3.6
	Pyrobaculum arsenaticum	1.3	1.2	x	5.9	6.3	3.6
	Pyrobaculum calidifontis	1.3	0.9	0.2	5.9	2.0	3.6
	Sulfolobus tokodaii	2.4	2.9	4.6	7.9	2.7	10.3
Euryarchaeota: Thermococci	*Pyrococcus furiosus*	4.4	4.9	1.8	0.5	2.2	0.6
	Pyrococcus horikoshii	1.9	2.0	1.8	0.5	2.2	0.6
Thermoplasmata	*Aciduliprofundum boonei*	0.8	0.7	0.4	0.5	1.9	1.2
Archaeoglobi	*Archaeoglobus fulgidus*	0.8	0.8	1.0	0.0	0.4	0.2
Haloarchaea	*Haloferax volcanii*	0.4	0.0	0.2	0.1	0.2	0.1
Methanopyri	*Methanopyrus kandleri*	2.1	1.0	0.8	0.0	0.4	0.2
Methanomicrobia	*Methanosarcina acetivorans*	0.6	0.6	0.8	0.0	0.2	0.2
Methanococci	*Methanocaldococcus jannaschii*	1.1	1.8	1.2	0.4	1.7	0.8
	Methanococcus maripaludis C5	0.4	0.8	1.7	0.2	0.7	0.9
	Methanococcus maripaludis S2	0.5	1.0	2.0	0.1	0.7	0.8

over-estimated

under-estimated

2 – extraire des gènes marqueurs

metaG **16S amplicons**

		Metagenome		SSU rDNA amplicons					
		ILM	454	V12	V13	V4	V35	V69	V48
Bacteria									
Aquificae	*Hydrogenobaculum sp. Y04AAS1*	4.9	2.4	0.0	0.3	1.8	0.2	0.1	1.0
	Persephonella marina	2.0	2.9	0.0	1.6	1.6	0.6	ND	2.0
	Sulfurihydrogenibium sp. YO3AOP1	1.3	2.5	0.0	1.4	0.7	0.6	ND	1.3
	S. yellowstonense	2.2	4.3	0.0	1.4	0.7	0.6	0.0	1.2
Thermotogae	*Thermotoga neapolitana*	1.3	1.3	1.2	0.3	1.2	0.5	ND	0.7
	Thermotoga petrophila	0.3	0.5	1.3	0.3	1.2	0.5	ND	0.7
	Thermotoga sp. RQ2	0.7	1.0	1.2	0.3	1.2	0.4	ND	0.7
Thermi/ Deinococcl	*Deinococcus radiodurans*	2.5	1.3	0.5	0.4	0.6	0.6	0.6	0.4
	Thermus thermophilus	1.3	0.1	1.1	0.5	1.9	0.7	0.5	0.3
Dictyoglomi	*Dictyoglomus turgidum*	2.7	4.3	2.2	0.4	7.9	2.4	0.1	0.8
Actinobacteria	*Salinispora arenicola*	1.0	0.3	0.5	1.2	0.1	1.0	0.4	0.3
	Salinispora tropica	1.2	0.4	0.5	1.2	0.1	1.0	0.4	0.3
Chloroflexi	*Chloroflexus aurantiacus*	1.8	1.1	2.5	2.2	1.7	2.7	0.1	0.4
	Herpetosiphon aurantiacus	1.6	1.6	2.5	1.0	1.0	0.7	0.6	1.3
Cyanobacteria	*Nostoc sp. PCC 7120*	1.5	2.1	1.3	1.6	1.3	1.8	5.9	2.2
Bacteroides	*Bacteroides thetaiotaomicron*	1.1	1.8	2.2	0.9	0.0	1.2	0.1	1.3
	Bacteroides vulgatus	1.4	1.9	2.1	0.7	ND	1.4	0.1	1.2
	Porphyromonas gingivalis	1.0	0.7	0.9	0.5	0.0	1.0	0.1	0.8
Chlorobi	*Chlorobium limicola*	0.8	1.0	1.0	1.7	0.4	0.3	0.8	0.2
	Chlorobium phaeobacteroides	1.0	1.5	0.7	1.3	0.7	0.9	1.2	0.2
	Chlorobium phaeovibrioides	0.5	0.3	0.6	1.2	0.6	0.7	0.9	0.1
	Chlorobium tepidum	1.3	1.4	1.3	1.9	0.6	1.1	1.1	0.2
	Pelodictyon phaeoclathratiforme	1.4	1.2	0.6	1.1	0.7	0.8	1.3	0.2
Firmicutes	*Caldicellulosiruptor bescii*	1.8	2.2	3.1	0.8	2.7	1.8	0.8	0.8
	Caldicellulosiruptor saccharolyticus	1.8	3.0	4.4	1.2	4.5	2.1	1.9	1.4
	Clostridium thermocellum	0.5	0.7	1.8	1.0	0.4	1.5	0.6	0.9
	Enterococcus faecalis	1.1	2.1	1.2	1.2	0.5	1.6	8.5	1.5
	Thermoanaerobacter pseudeth	1.2	2.1	0.9	0.9	1.1	2.5	1.5	0.5
Fusobacteria	*Fusobacterium nucleatum*	1.0	2.1	2.2	1.8	0.6	2.2	0.1	2.1
Verrucomicrobia	*Akkermansia muciniphila*	1.2	1.2	0.0	1.6	2.0	0.1	2.0	1.5
Gemmatimonadetes	*Gemmatimonas aurantiaca*	2.2	1.3	0.6	0.3	1.6	0.8	0.9	1.1
Planctomycetes	*Rhodopirellula baltica*	1.1	1.2	0.0	0.7	0.5	0.2	0.0	1.0
Spirochaetae	*Treponema denticola*	1.2	1.6	2.6	0.8	0.8	1.4	0.0	0.9
Acidobacteria	*Acidobacterium capsulatum*	0.8	0.5	0.7	1.3	1.0	0.9	2.4	0.8
Proteobacteria:									
Alpha	*Ruegeria pomeroyi*	0.5	0.3	0.8	1.0	0.6	1.0	0.7	0.9
	Sulfitobacter sp. EE-36	1.1	1.0	0.7	1.0	0.8	1.0	0.3	0.6
	Sulfitobacter sp. NAS-14.1	0.3	0.3	0.7	1.1	0.8	1.0	0.3	0.6
	Zymomonas mobilis	1.3	2.0	ND	1.2	1.6	0.5	0.3	1.5
Beta	*Bordetella bronchiseptica*	0.9	0.5	1.1	0.7	1.3	1.8	0.5	5.3
	Burkholderia xenovorans	1.2	0.9	0.7	0.7	1.0	0.9	0.1	1.5
	Leptothrix cholodnii	1.6	0.8	0.5	0.6	0.7	0.9	0.1	2.4
	Nitrosomonas europaea	0.9	1.1	0.8	0.7	2.6	1.0	0.2	1.1
Gamma	*Shewanella baltica OS185*	0.9	1.6	0.8	0.5	1.1	1.2	1.3	1.7
	Shewanella baltica OS223	1.1	1.8	0.8	0.4	1.1	1.2	1.3	1.6
Delta	*Desulfovibrio piger*	0.4	0.2	0.0	1.0	1.6	0.4	3.5	2.1
	Desulfovibrio vulgaris	1.3	0.5	ND	0.7	0.9	0.3	2.9	1.4
	Geobacter sulfurreducens	2.7	0.8	1.5	1.0	0.6	2.1	0.7	1.9
Epsilon	*Wolinella succinogenes*	1.1	0.9	1.0	1.2	0.6	0.4	2.1	1.1

Synthetic genome mix :
observed / expected abundance
Shakya *et al.* (2013)

metaG **16S amplicons**

		Metagenome		SSU rDNA amplicons			
		ILM	454	V13	V4	V4a	V48
Archaea							
Nanoarchaeaota	*Nanoarchaeum equitans*	0.6	0.1	0.1	1.0	0.4	0.3
Crenarchaeota: Thermoprotei	*Ignicoccus hospitalis*	1.8	0.9	0.2	0.5	1.2	1.8
	Pyrobaculum aerophilum	1.5	1.6	x	5.9	4.3	3.6
	Pyrobaculum arsenaticum	1.3	1.2	x	5.9	6.3	3.6
	Pyrobaculum calidifontis	1.3	0.9	0.2	5.9	2.0	3.6
	Sulfolobus tokodaii	2.4	2.9	4.6	7.9	2.7	10.3
Euryarchaeota: Thermococci	*Pyrococcus furiosus*	4.4	4.9	1.8	0.5	2.2	0.6
	Pyrococcus horikoshii	1.9	2.0	1.8	0.5	2.2	0.6
Thermoplasmata	*Aciduliprofundum boonei*	0.8	0.7	0.4	0.5	1.9	1.2
Archaeoglobi	*Archaeoglobus fulgidus*	0.8	0.8	1.0	0.0	0.4	0.2
Haloarchaea	*Haloferax volcanii*	0.4	0.0	0.2	0.1	0.2	0.1
Methanopyri	*Methanopyrus kandleri*	2.1	1.0	0.8	0.0	0.4	0.2
Methanomicrobia	*Methanosarcina acetivorans*	0.6	0.6	0.8	0.0	0.2	0.2
Methanococci	*Methanocaldococcus jannaschii*	1.1	1.8	1.2	0.4	1.7	0.8
	Methanococcus maripaludis C5	0.4	0.8	1.7	0.2	0.7	0.9
	Methanococcus maripaludis S2	0.5	1.0	2.0	0.1	0.7	0.8

Illumina

2 – extraire des gènes 16S

miTAGS : real env. data !
Logares et al. (2013)

Environmental Microbiology (2013) doi:10.

Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities

Ramiro Logares,[1] Shinichi Sunagawa,[2] Guillem Salazar,[1] Francisco M. Cornejo-Castillo,[1] Isabel Ferrera,[1] Hugo Sarmento,[1,3] Pascal Hingamp,[4] Hiroyuki Ogata,[4,5] Colomban de Vargas,[6] Gipsi Lima-Mendez,[7,8] Jeroen Raes,[7,8] Julie Poulain,[9] Olivier Jaillon,[9,10,11] Patrick Wincker,[9,10,11] Stefanie Kandels-Lewis,[2] Eric Karsenti,[2] Peer Bork[2] and Silvia G. Acinas[1*]

[1]*Department of Marine Biology and Oceanography, Institute of Marine Science (ICM), Spanish National Research Council (CSIC), Passeig Marítim de la Barceloneta, 37–49, Barcelona ES-08003, Spain.*
[2]*European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg 69117, Germany.*
[3]*Department of Oceanography and Limnology, Federal University of Rio Grande do Norte, Natal 59014-002,*
[10]*Cen*
Mixte
91057
[11]*Univ*
91025

approaches. Our results indicate that by overcoming PCR biases related to amplification and primer mismatch, *mi*tags may provide more realistic estimates of community richness and evenness than amplicon *454*tags. In addition, *mi*tags can capture expected beta diversity patterns. Using *mi*tags is now economically feasible given the dramatic reduction in high-throughput sequencing costs, having the advantage of retrieving simultaneously both taxonomic (Bacteria, Archaea and Eukarya) and functional information from the same microbial community.

Summ

Seque
(PCR)
investigating environmental prokaryotic diversity, despite the known biases introduced during PCR. Here we show that 16S rDNA fragments derived from Illumina-sequenced environmental metagenomes (*mi*tags) are a powerful alternative to 16S rDNA amplicons for investigating the taxonomic diversity

2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)

Table 1. General comparison of the different platforms and approaches

Approach[a]	Template	Coverage	16S rDNA specificity	16S rDNA recovery[b]	PCR bias[c]	16S rDNA overlap[d]	Taxonomic definition[e]	OTU clustering[f]	€/Mb[g]
$_{mi}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic	Spanning all 16S rDNA	High/moderate	Absent	Low	Variable	Mapping to reference OTUs/V region selection for *de novo*	0.1/100[h] (HiSeq)
$_{m454}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic	Spanning all 16S rDNA	Very low	Absent	Low	Variable	Mapping to reference OTUs/V region selection for *de novo*	12/12000[h] (Titanium)
$_{454}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	High/very high	Present	High	High	*De novo* and mapping to reference OTUs	12 (Titanium)
$_{¡}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	Very high	Present	High	High/moderate	*De novo* and mapping to reference OTUs	0.7 (MiSeq)

a The four basic approaches are indicated: $_{mi}$tags (metagenomic Illumina 16S tags), $_{m454}$tags (metagenomic 454 16S tags), $_{454}$tags (amplicon-based 454 16S tags) and $_{¡}$tags (amplicon-based Illumina 16S tags).

b Number of recovered 16S rDNA reads from the used template. Estimations depend on the throughput of the platform.

2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)

Table 1. General comparison of the different platforms and approaches

Approach[a]	Template	Coverage	16S rDNA specificity	16S rDNA recovery[b]	PCR bias[c]	16S rDNA overlap[d]	Taxonomic definition[e]	OTU clustering[f]	€/Mb[g]
$_{mi}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic	Spanning all 16S rDNA	High/moderate	Absent	Low	Variable	Mapping to reference OTUs/V region selection for *de novo*	0.1/100[h] (HiSeq)
$_{m454}$Tags	Metagenomic fragments	16S rDNA, functional metagenomic							2/12000[h] Titanium)
$_{454}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	High/very high	Present	High	High	*De novo* and mapping to reference OTUs	12 (Titanium)
$_{¡}$Tags	Amplicons	16S rDNA only	Specific 16S rDNA area	Very high	Present	High	High/moderate	*De novo* and mapping to reference OTUs	0.7 (MiSeq)

Tara Oceans, 3 stations :
- PCR 16S V1-V3 + 454
- illumina « whole shotgun» metaG

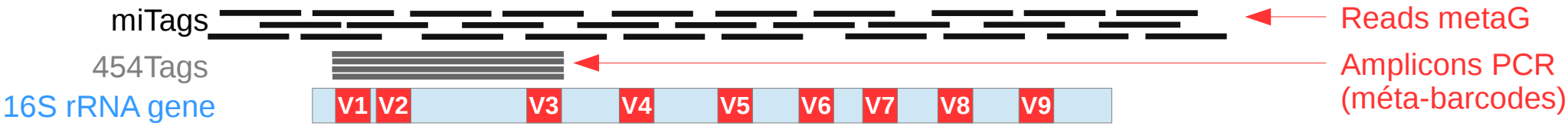
[a] The four basic approaches are indicated: $_{mi}$tags (metagenomic Illumina 16S tags), $_{m454}$tags (metagenomic 454 16S tags), $_{454}$tags (amplicon-based 454 16S tags) and $_{¡}$tags (amplicon-based Illumina 16S tags).

[b] Number of recovered 16S rDNA reads from the used template. Estimations depend on the throughput of the platform.

2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)

miTags (Illumina whole shotgun metaG)
versus
454Tags (454 V1-V3PCR amplicons)

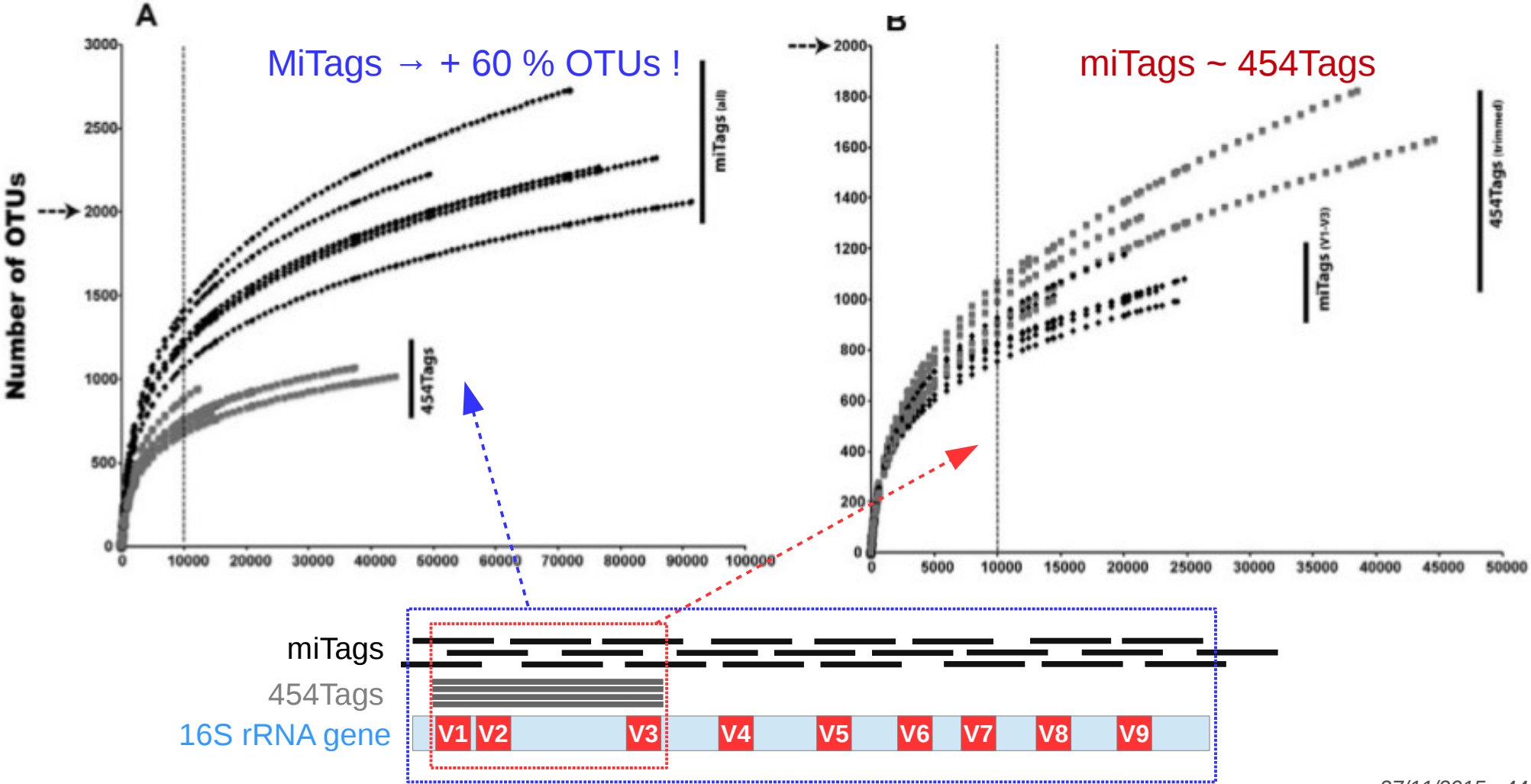


2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)

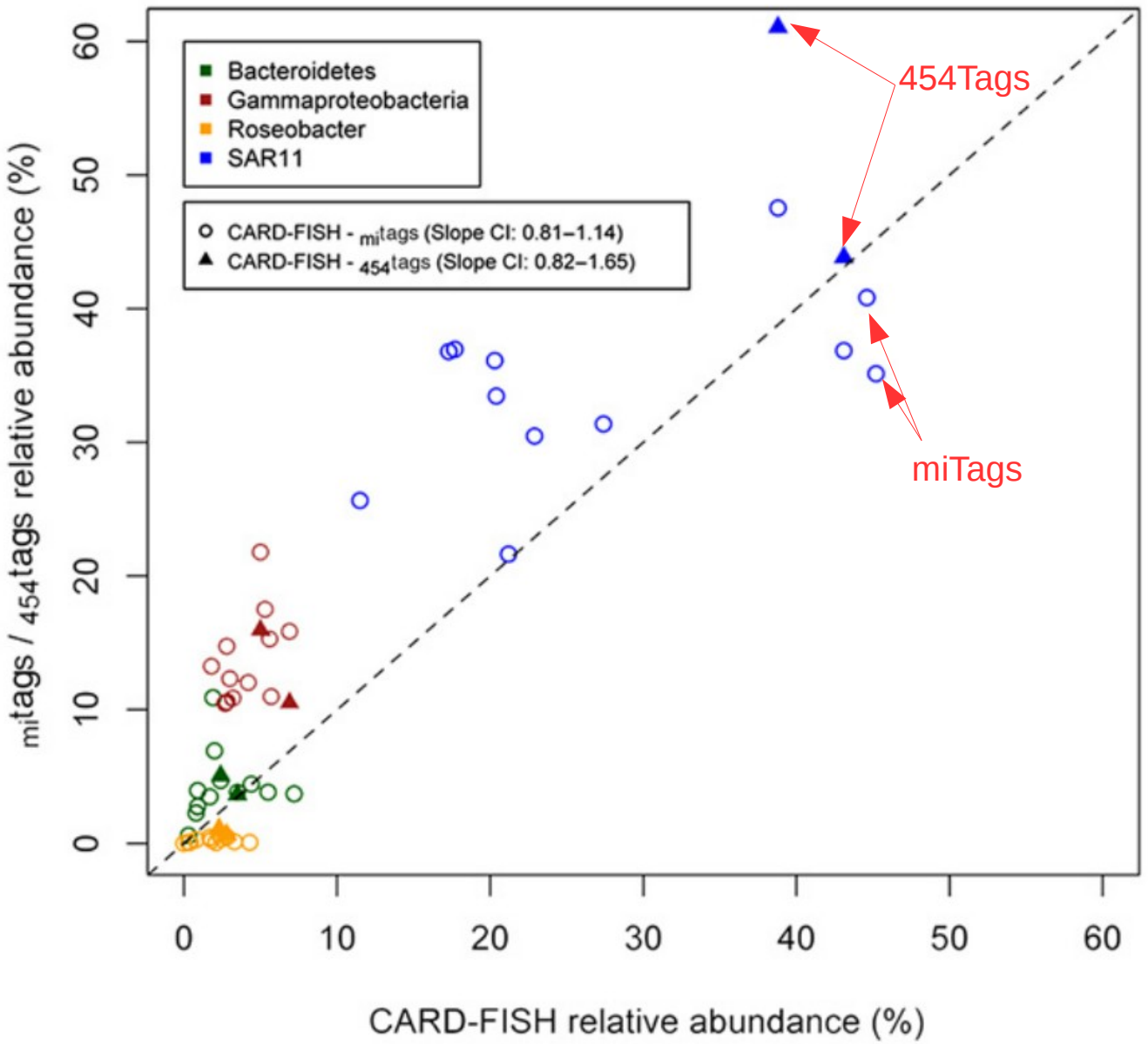
All 16S reads

Subset of reads for V1-V3 of 16S (28%)



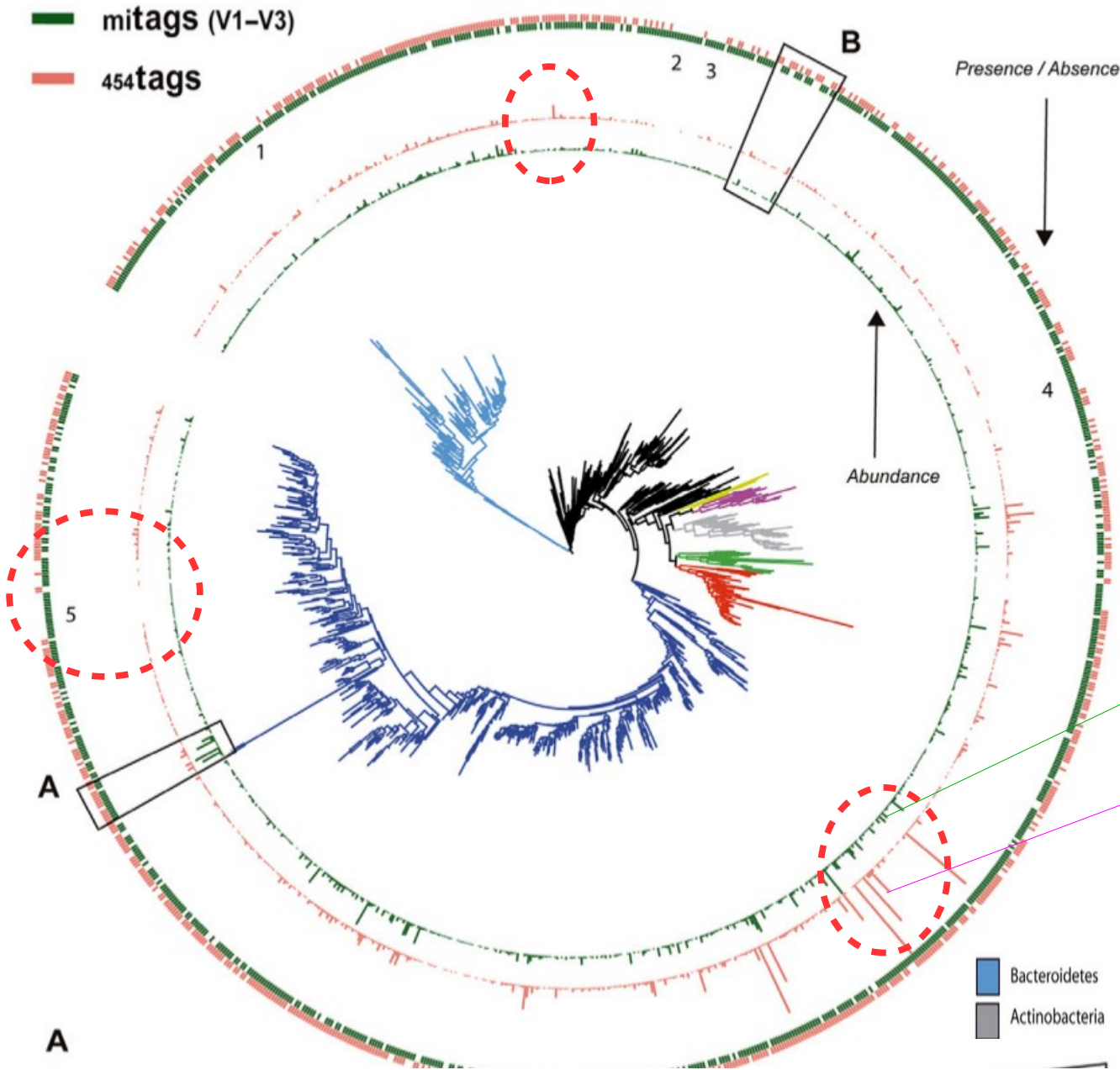
2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)



2 – extraire des gènes 16S

miTAGS
Logares et al. (2013)



miTags
(16S in metaG)

454Tags
(classic amplicons)

Bacteroidetes
Actinobacteria
Gammaproteobacteria
Firmicutes
Planctomycetes
Cyanobacteria
Verrucomicrobia
Alphaproteobacteria

2 – extraire des gènes marqueurs

- Approche 1 : extraire gènes 16S des metaG

- Approche 2 : extraires autres (non 16S) gènes marqueurs des metaG

mOTU-LG
Sunagawa et al. (2013)

2 – extraire des gènes marqueurs

40 universal prok genes, single copy, rare HGT :

10/40 marker genes

mOTU-LG
Sunagawa et al. (2013)

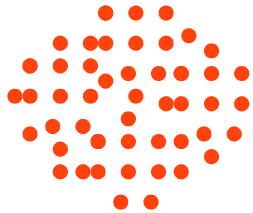
HMM profile search (fast)

3500 prokaryotic reference genomes

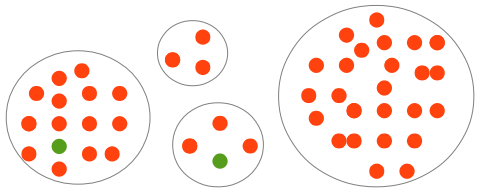
metagenomic contigs >500bp

Ref marker genes

MetaG marker genes



Marker genes clustering (93-96%)

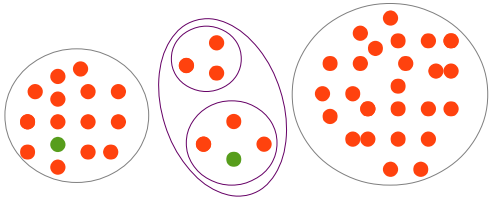


42% 58%

in 252 human gut metaG

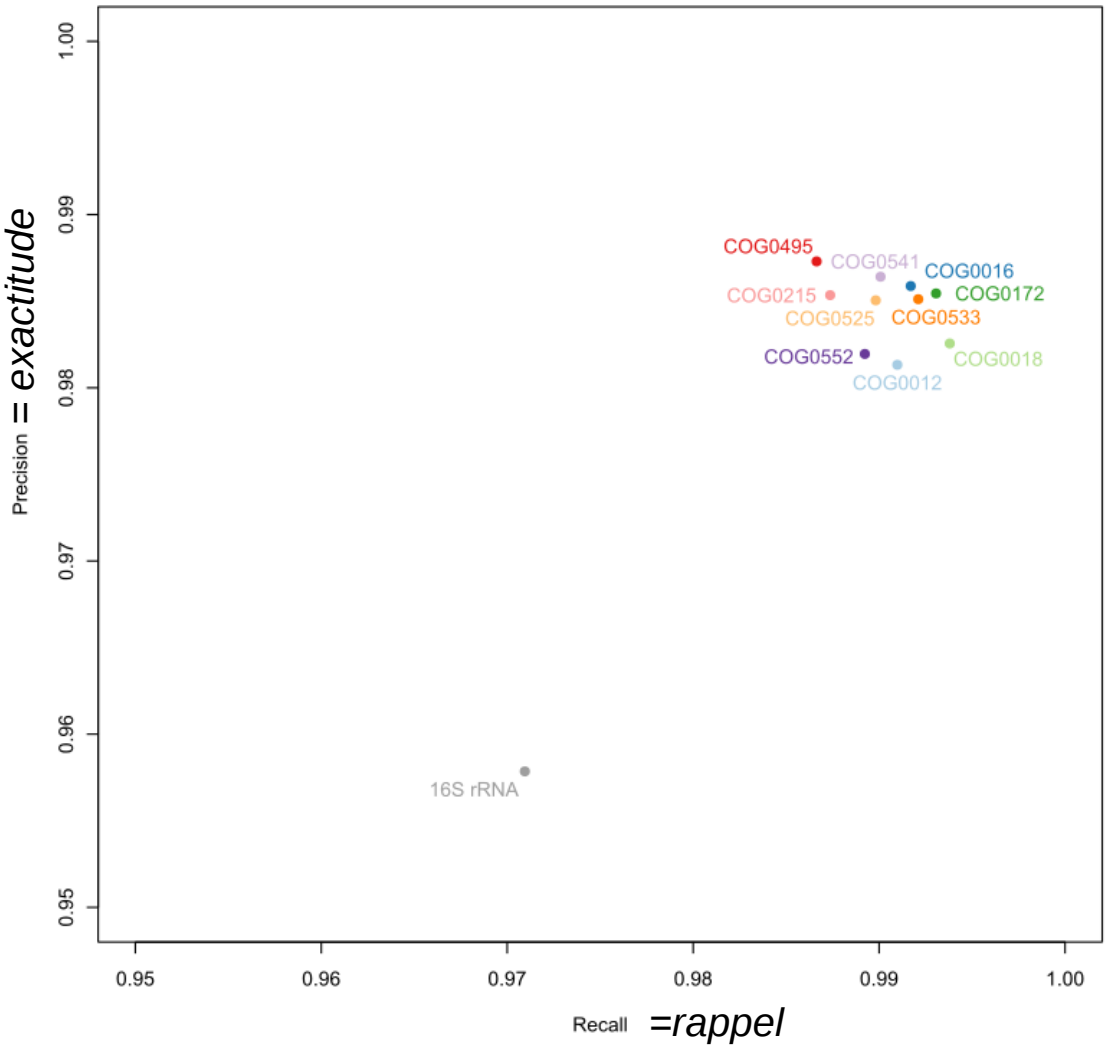
mOTUs

Covariance across samples



mOTUs – Linkage Gr.

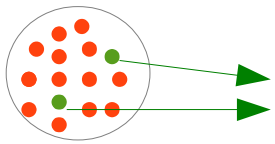
2 – extraire des gènes marqueurs



Supplementary Figure 1: Accuracy of species-level mOTU clustering. Clustering accuracy was assessed by testing whether mOTUs with at least two MGs or 16S rRNA genes that originated from a type strain reference genome were consistent regarding the taxonomic annotation of their members at the species level according to the NCBI taxonomy. According to this information, false discovery rates (FDRs) and recall values were calculated for all mOTUs. Precision (1 - FDR) and recall can take values between 0 and 1, with high values indicating a good agreement of mOTU cluster members with the NCBI taxonomy.

mOTU-LG
Sunagawa et al. (2013)

Précision des prédictions évaluées avec mOTUs ayant >1 gènes réf.



Taxonomie des marqueurs de référence concordantes ?

Méthodes disponibles

1 – Biais compositionnels de l'ADN dans les fragments metaG
(« binning » = regroupement par classes)

2 – Extraction des gènes marqueurs &
comparaison à une base de données de référence

3 - Assignation taxonomique aux fragments MetaG par homologie à une base de données de référence

3 -Homologie aux BD de réf.

MEGAN
Huson et al. (2007)

Resource

MEGAN analysis of metagenomic data

Daniel H. Huson,[1,3] Alexander F. Auch,[1] Ji Qi,[2] and Stephan C. Schuster[2,3]

[1]Center for Bioinformatics, Tübingen University, Sand 14, 72076 Tübingen, Germany; [2]Center for Comparative Genomics and Bioinformatics, Center for Infectious Disease Dynamics, Penn State University, University Park, Pennsylvania 16802, USA

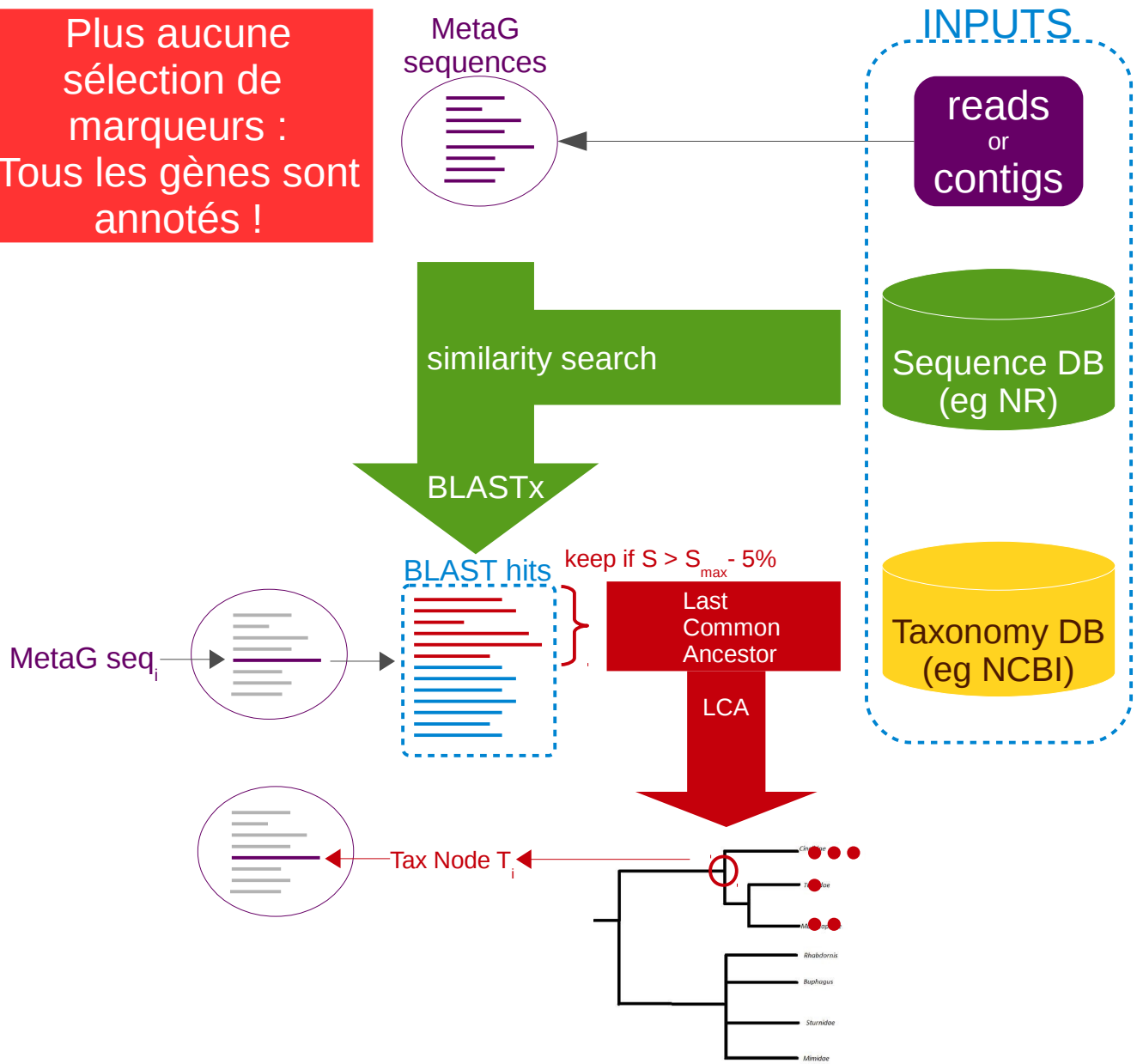
Metagenomics is the study of the genomic content of a sample of organisms obtained from a common habitat using targeted or random sequencing. Goals include understanding the extent and role of microbial diversity. The taxonomical content of such a sample is usually estimated by comparison against sequence databases of known sequences. Most published studies use the analysis of paired-end reads, complete sequences of environmental fosmid and BAC clones, or environmental assemblies. Emerging sequencing-by-synthesis technologies with very high throughput are paving the way to low-cost random "shotgun" approaches. This paper introduces MEGAN, a new computer program that allows laptop analysis of large metagenomic data sets. In a preprocessing step, the set of DNA sequences is compared against databases of known sequences using BLAST or another comparison tool. MEGAN is then used to compute and explore the taxonomical content of the data set, employing the NCBI taxonomy to summarize and order the results. A simple lowest common ancestor algorithm assigns reads to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. The software allows large data sets to be dissected without the need for assembly or the targeting of specific phylogenetic markers. It provides graphical and statistical output for comparing different data sets. The approach is applied to several data sets, including the Sargasso Sea data set, a recently published metagenomic data set sampled from a mammoth bone, and several complete microbial genomes. Also, simulations that evaluate the performance of the approach for different read lengths are presented.

17:377–386 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Method name	Class of method	Number of citations[c]
MEGAN4	Similarity	1089

3 -Homologie aux BD de réf.

Plus aucune sélection de marqueurs : Tous les gènes sont annotés !



MEGAN
Huson et al. (2007)

3 -Homologie aux BD de réf.

MEGAN
Huson et al. (2007)

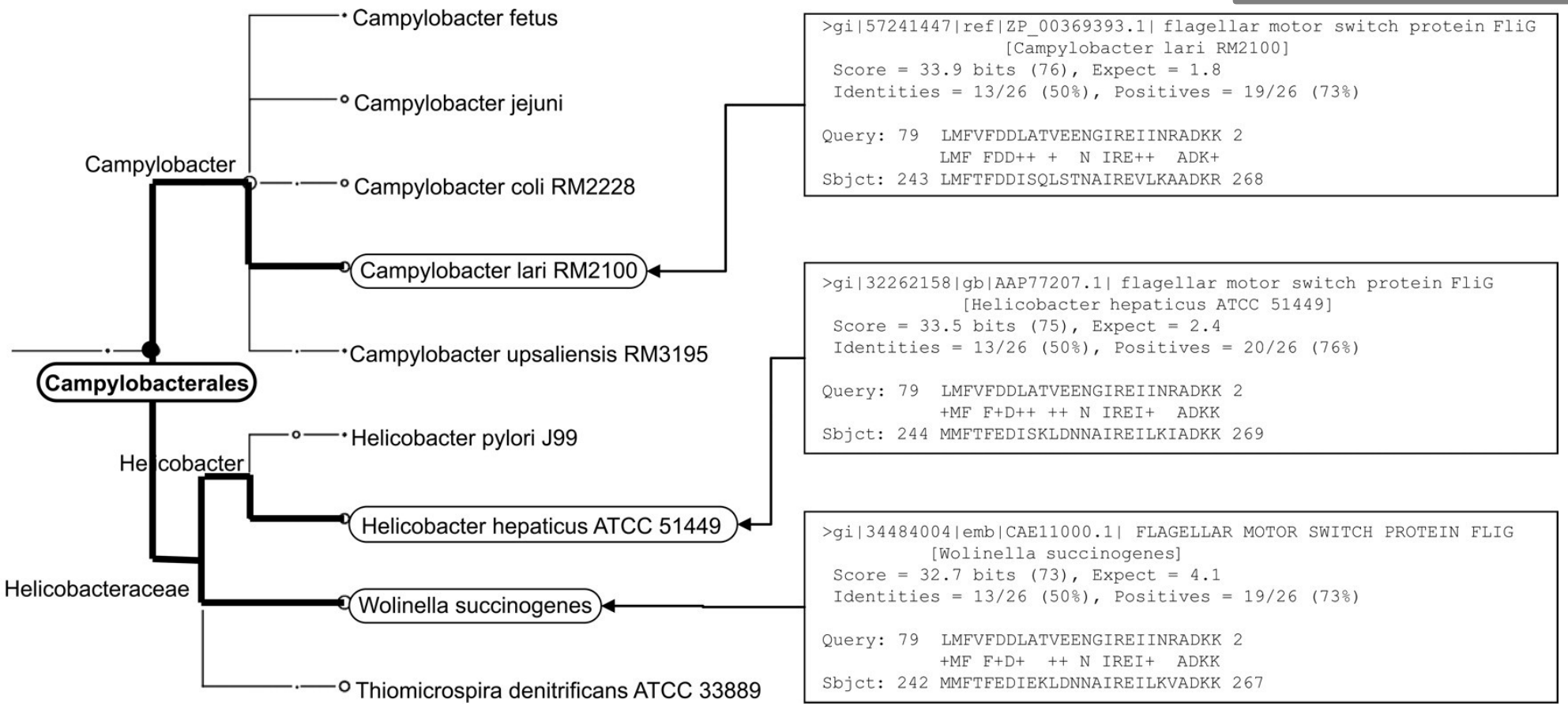


Figure 2. On the *right*, we list the three BLASTX matches obtained for a specific read *r* from the mammoth data set, to sequences representing *Campylobacter lari*, *Helicobacter hepaticus*, and *Wolinella*, respectively. The LCA-assignment algorithm assigns *r* to the taxon *Campylobacterales*, shown on the *left*, as it is the lowest-common taxonomical ancestor of the three matched species.

3 -Homologie aux BD de réf.

MEGAN
Huson et al. (2007)

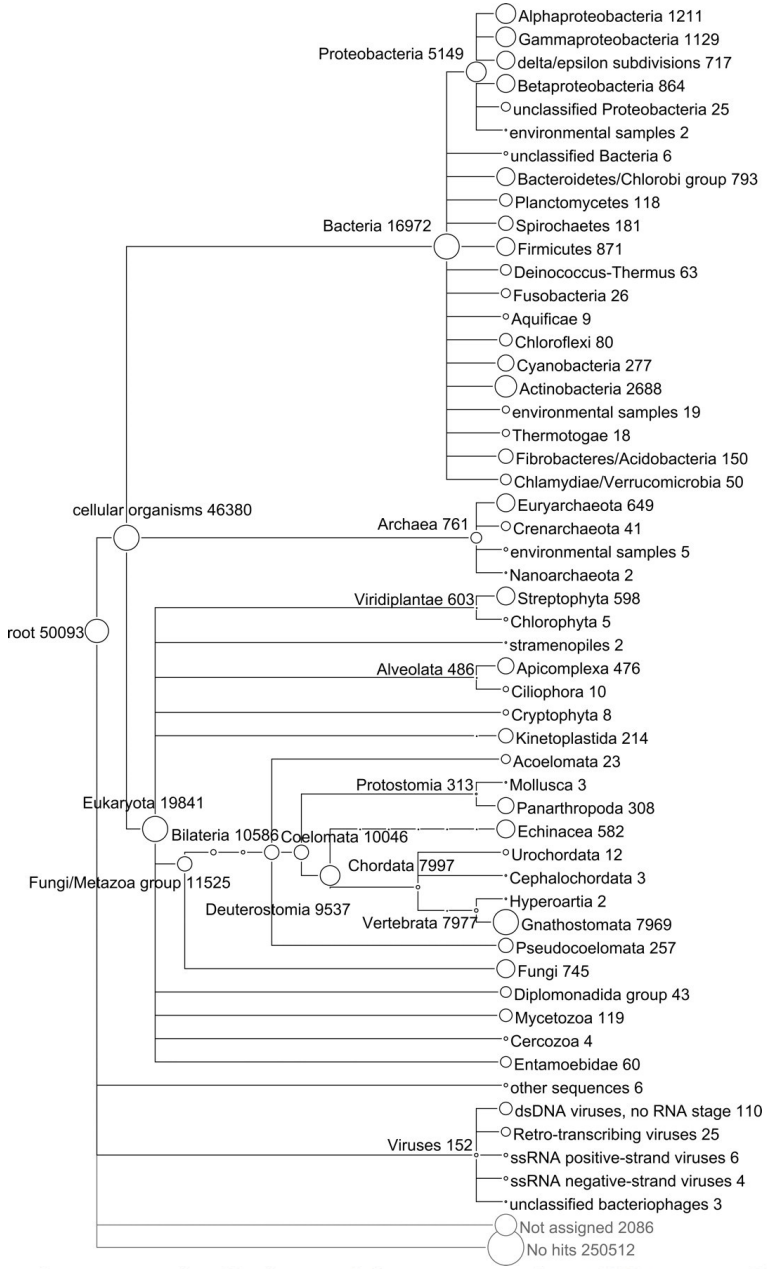


Figure 6. A low level view of the MEGAN analysis of the mammoth data set.

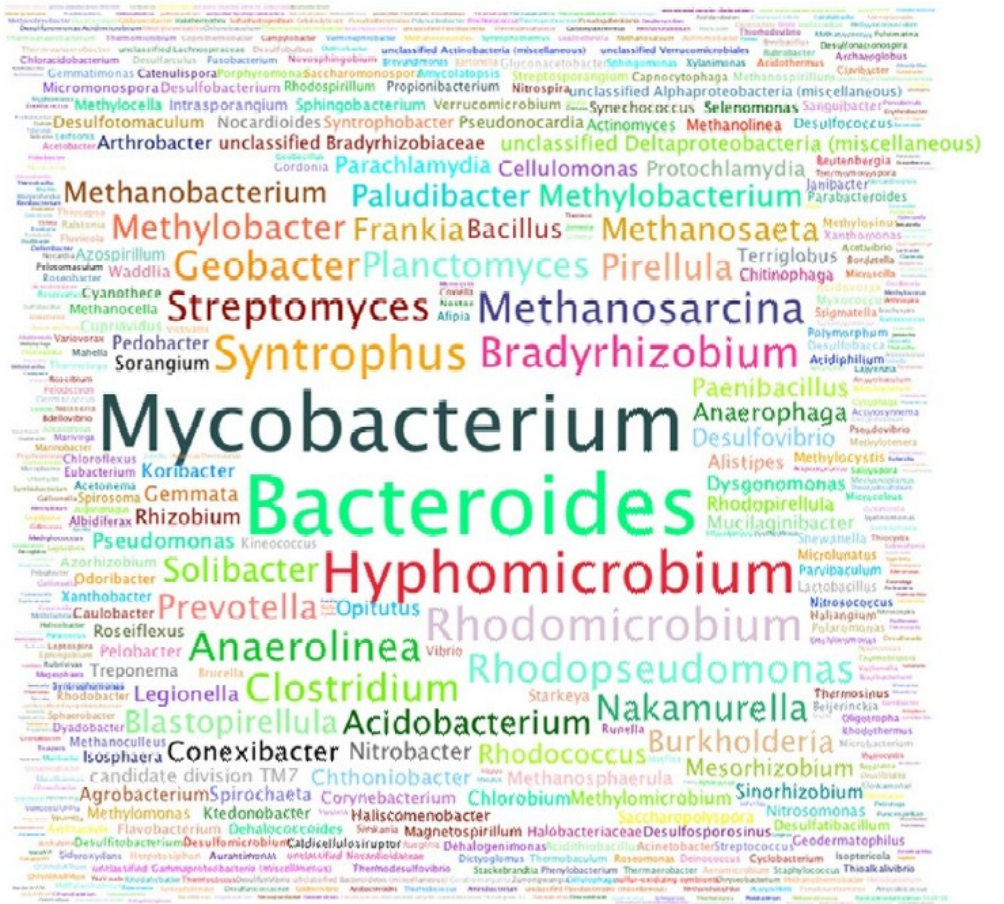
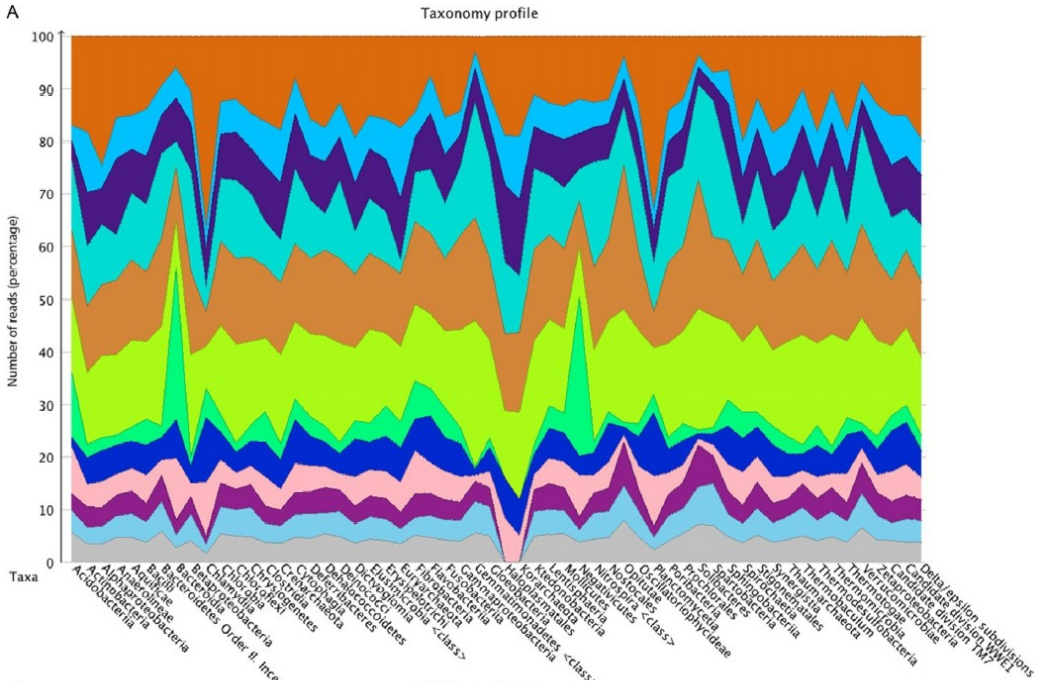


Figure 21.2 Word cloud of genus rank analysis. This type of representation makes it easy for the human eye to identify whether specific taxa have a significant presence in a sample. Fonts are scaled by number of reads assigned.

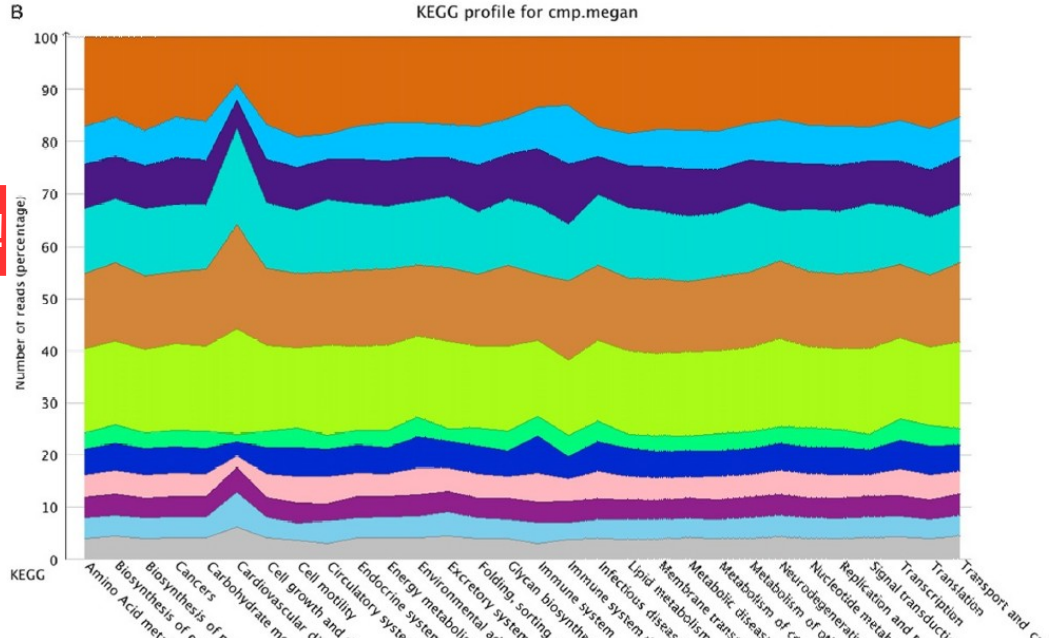
3 -Homologie aux BD de réf.

Taxonomie ...



MEGAN
Huson et al. (2007)

… & Fonctions !



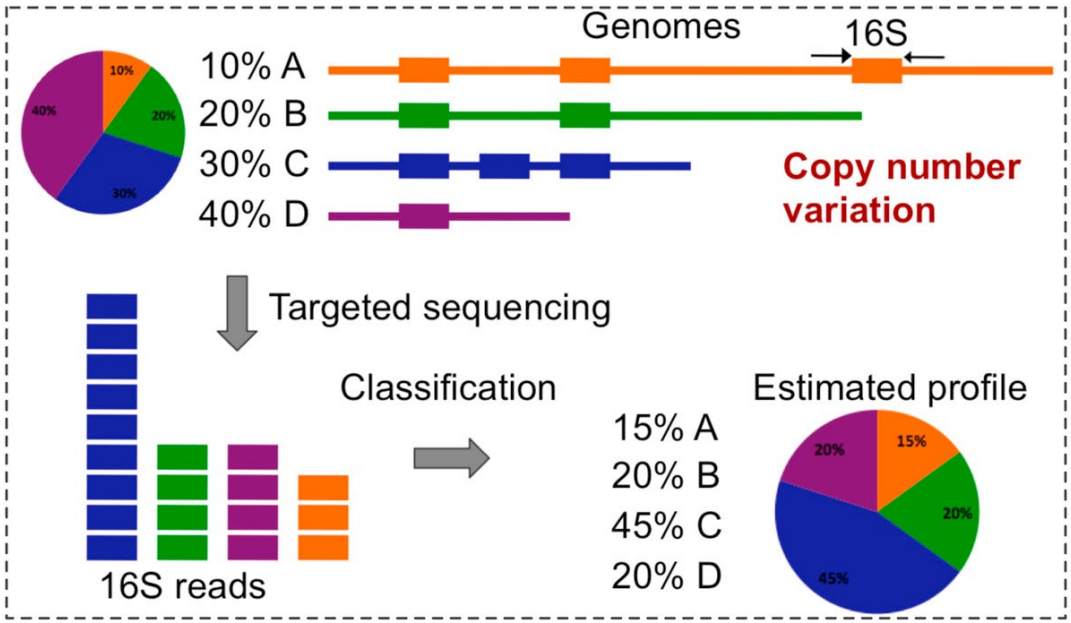
Estimation des abondances

- Gènes marqueurs : si « single copy », alors abondance gène marqueur = proxy pour le nombre d'organismes

- Méthodes de classifications (compos+similarité) nécessitent :
  - estimation des tailles de génomes

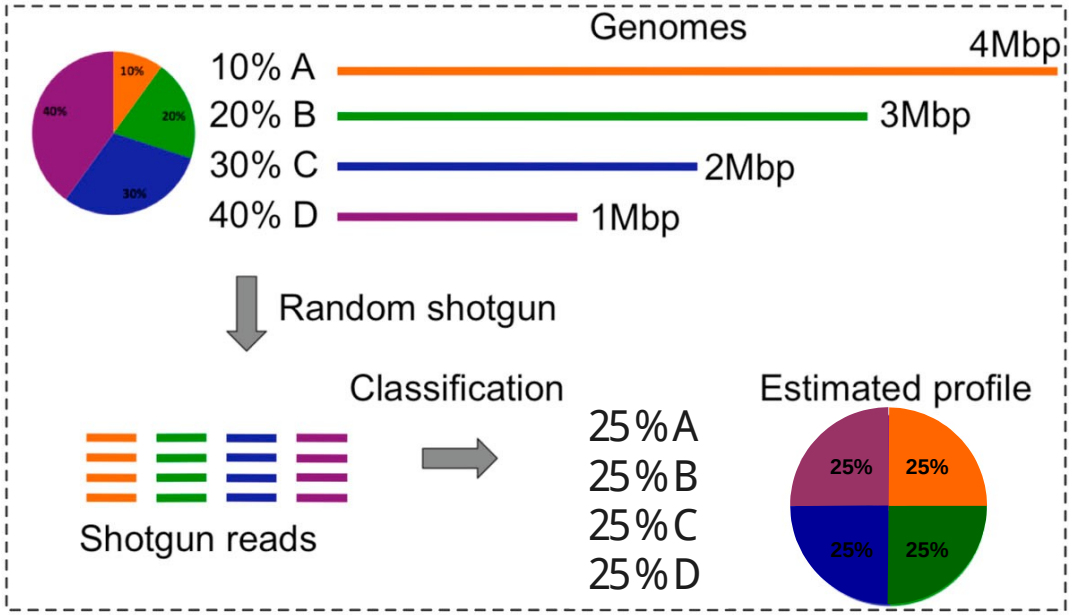
Biais d'abondances

Gènes marqueurs ARNr 16S



(a) Targeted sequencing of 16S rRNA

Classification globale métaG



(b) Metagenome shotgun sequencing

Lui *et al.* 2011

Correction des biais d'abundance

16S rRNA marker gene

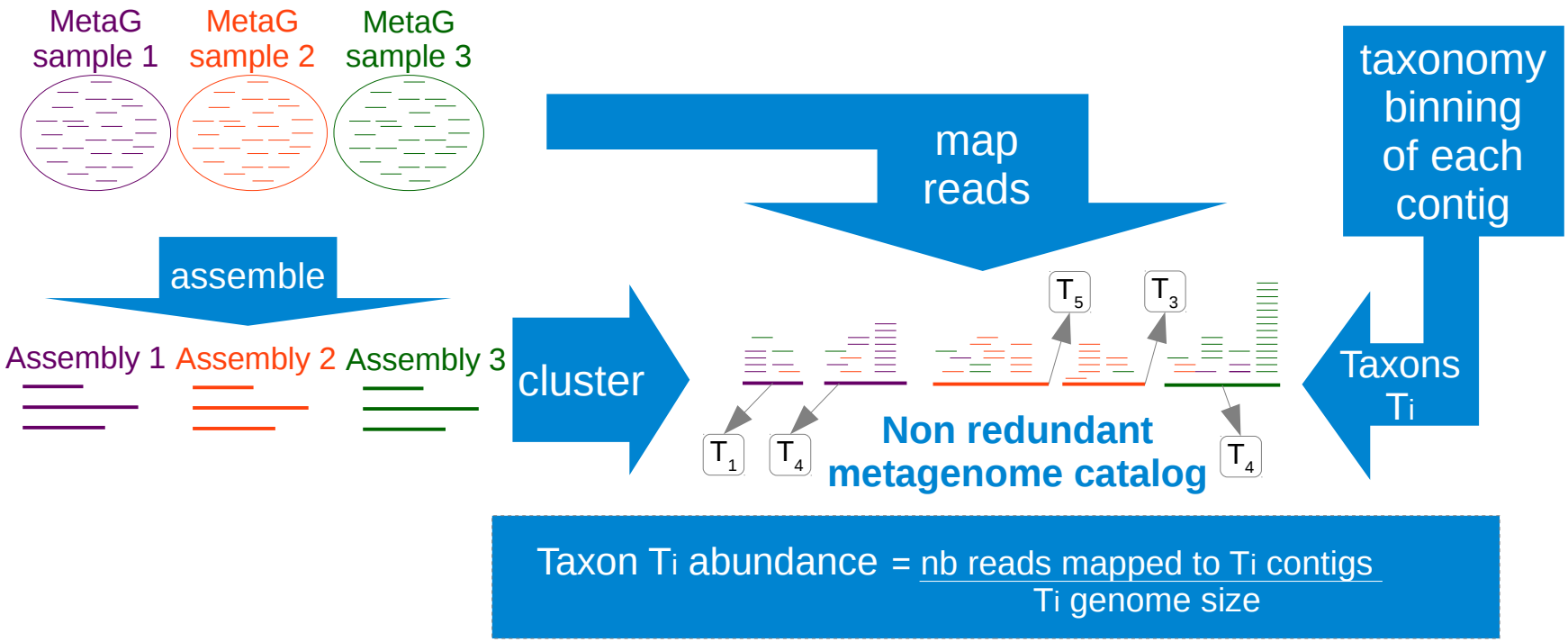
METHODOLOGY Open Access

Angly et al. (2014)

CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction

Florent E Angly[1*], Paul G Dennis[1,2], Adam Skarshewski[1], Inka Vanwonterghem[1,3], Philip Hugenholtz[1] and Gene W Tyson[1,3]

Binning whole metaG



Taxon Ti abundance = nb reads mapped to Ti contigs / Ti genome size

Méthodes disponibles

Article de revue
Segata *et al.* (2013)

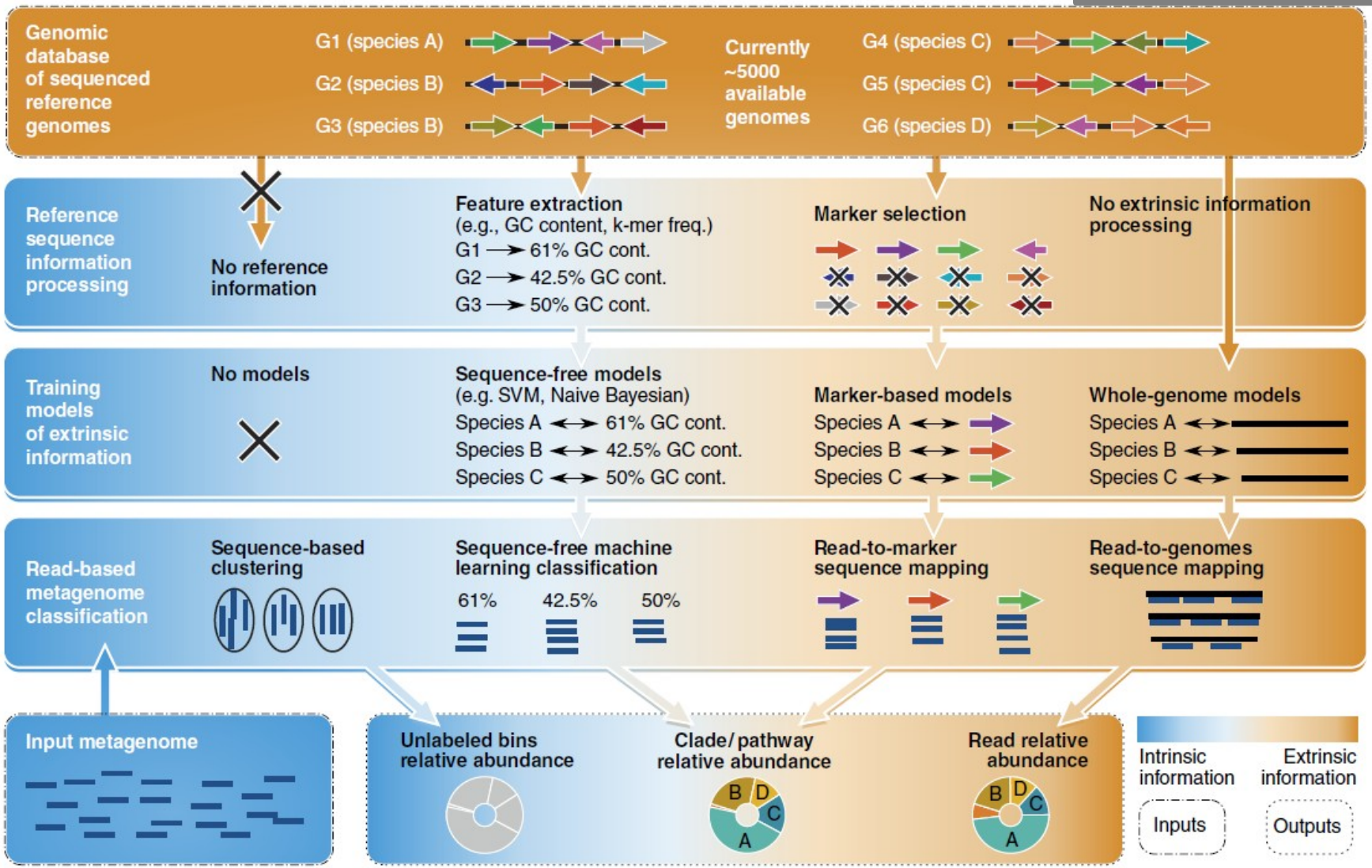


Figure 3 Intrinsic versus extrinsic metagenomic analysis can minimally, partially, or completely rely on prior knowledge from sequenced reference genomes. Methods that do not rely on any reference sequence information typically perform a sequence-based clustering of meta'omic reads, resulting in unlabeled clusters of sequences that can later be assigned to taxonomic or functional classes (analogous to Operational Taxonomic Unit clustering for 16S sequences). Available genomes can alternatively be used more extensively as references for short-read mapping, typically incurring an expense of high computational cost and possible ambiguous assignments for reads from nonunique regions. Intermediate approaches typically rely on a combination of pre-processing extrinsic reference genome information (e.g., to train a composition-based classifier) and intrinsic information (e.g., reads' nucleotide composition) to improve the discrimination power and focus the subsequent mapping operation to the most discriminative sequence-based markers.

Life beyond 16S…

Article de revue
Segata *et al.* (2013)

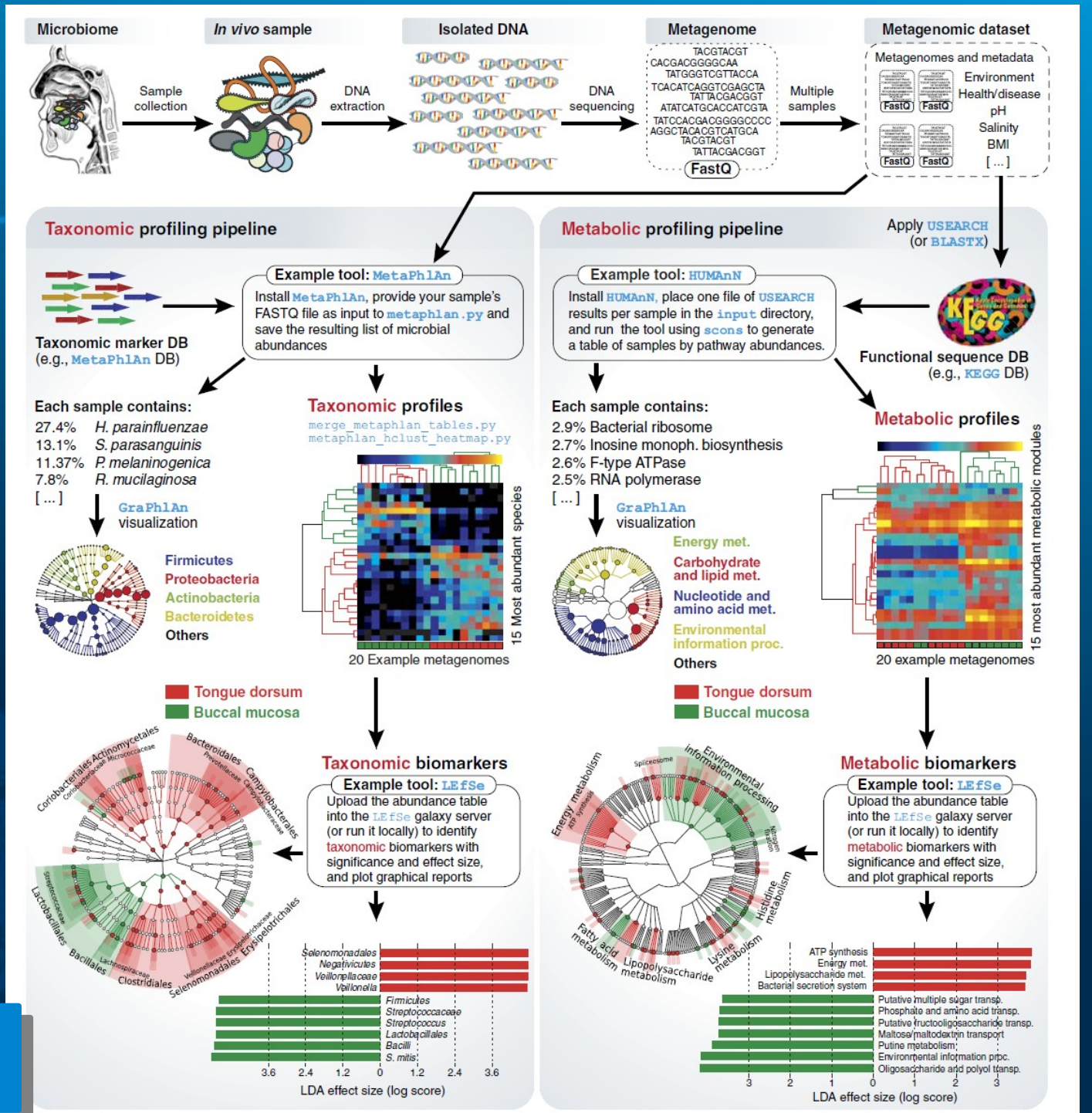


Figure 4 A typical current computational meta'omic pipeline to analyze and contrast microbial communities. After collecting microbiome samples, community DNA or RNA is extracted and sequenced, generating WMS samples (i.e., metagenomes) generally consisting of several million short reads each. This example uses 20 WMS samples from the oral cavity (10 from the buccal mucosa, and 10 from the tongue dorsum (The Human Microbiome Project Consortium, 2012b)). Complementary methods reconstruct the taxonomic characteristics (left) and metabolic potential (right) of the microbial communities. MetaPhlAn (Segata *et al.* 2012) is one of many alternatives to

Liste non exhaustive de logiciels

Compositional methods:

WGSQuikr	Koslicki et al. (2014)
Kraken	Wood et al. (2014)
PhyloPythiaS	Patil et al. (2012)
ClaMS	Pati et al. (2011)
Phymm	Brady et al. (2009)
NBC	Rosen et al. (2008)

Marker genes methods:

PhyloSift	Darling et al. (2014)
miTAGS	Logares et al. (2013)
mOTU-LG	Sunagawa et al. (2013)
MetaPhlAn	Segata et al. (2012)
MetaPhyler	Liu et al. (2011)

Reference database homology methods:

2bLCA	Hingamp et al. (2013)
CREST	Lanzen et al. (2012)
MEGAN	Huson et al. (2007, 2011)

Combination of methods :

PhymmBL (comp+homol)	Brady et al. (2011)