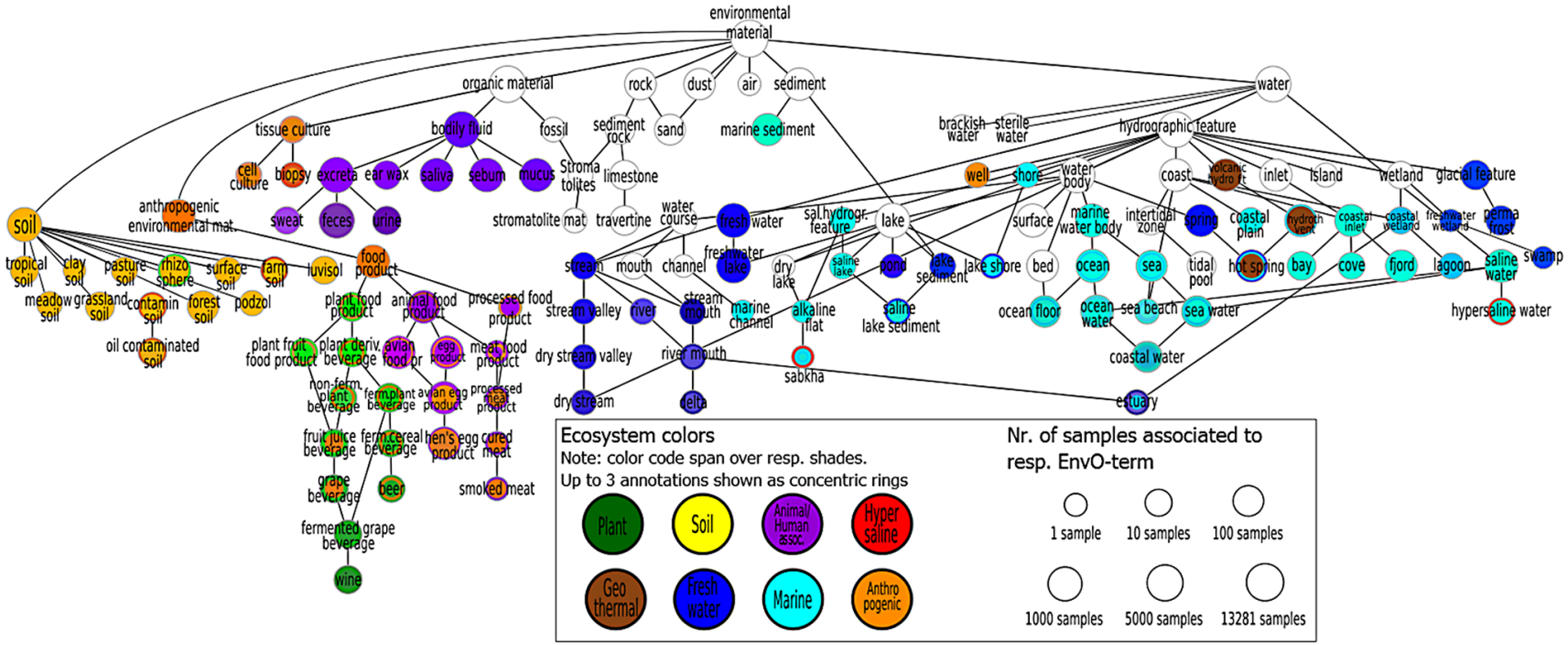
Cube
Computational Systems Biology

universität wien

The future of microbial genomics: next-generation bioinformatics for millions of genomes

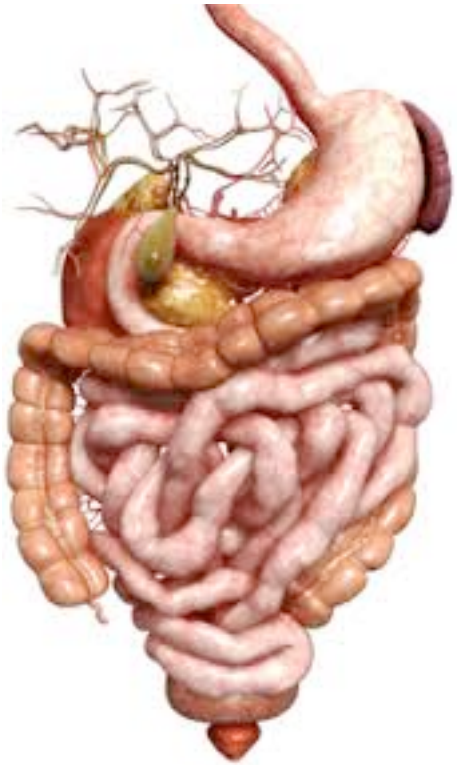
Thomas Rattei
Department of Microbiology and Ecosystem Science
University of Vienna

Microbial diversity in our environment

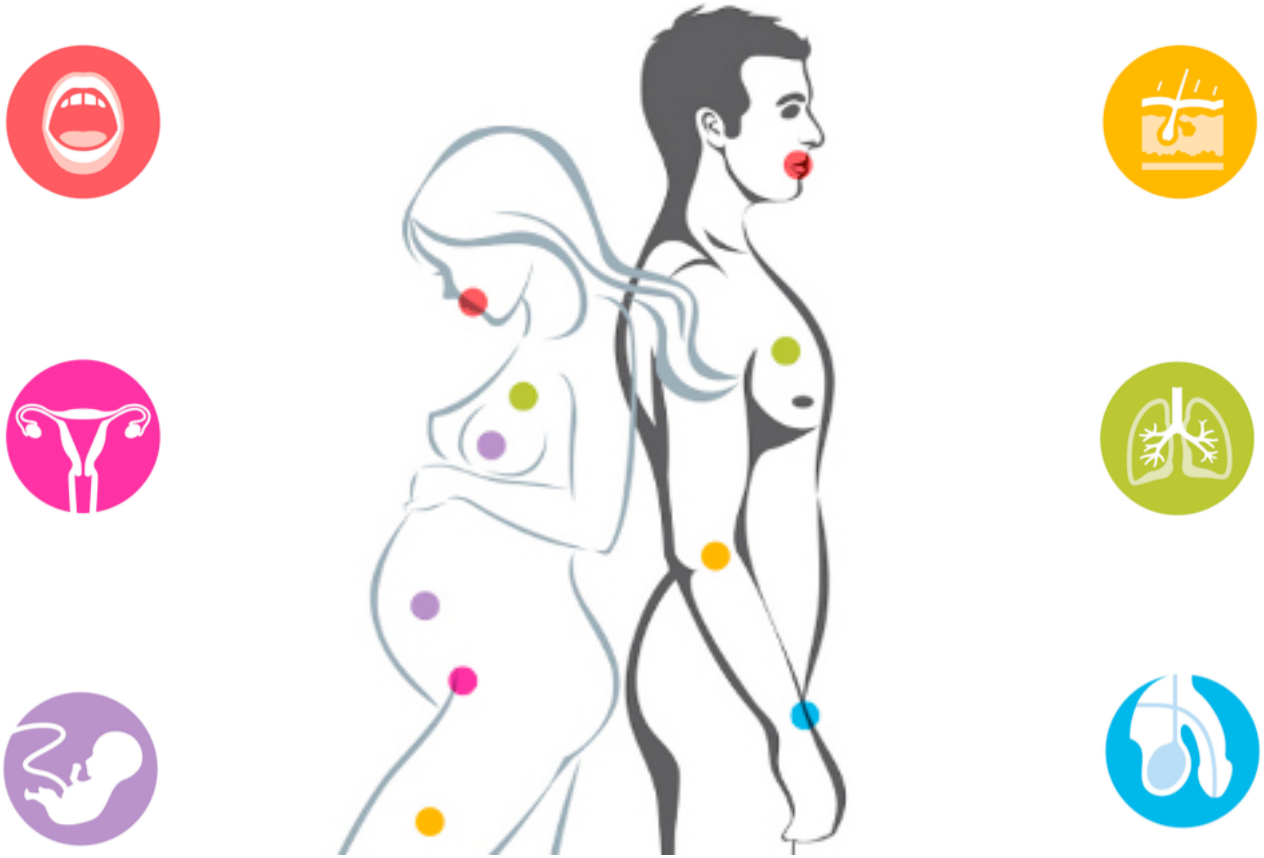


Hentschel et al., PLOS Comp Biol, Oct 2015

The body's microbiomes



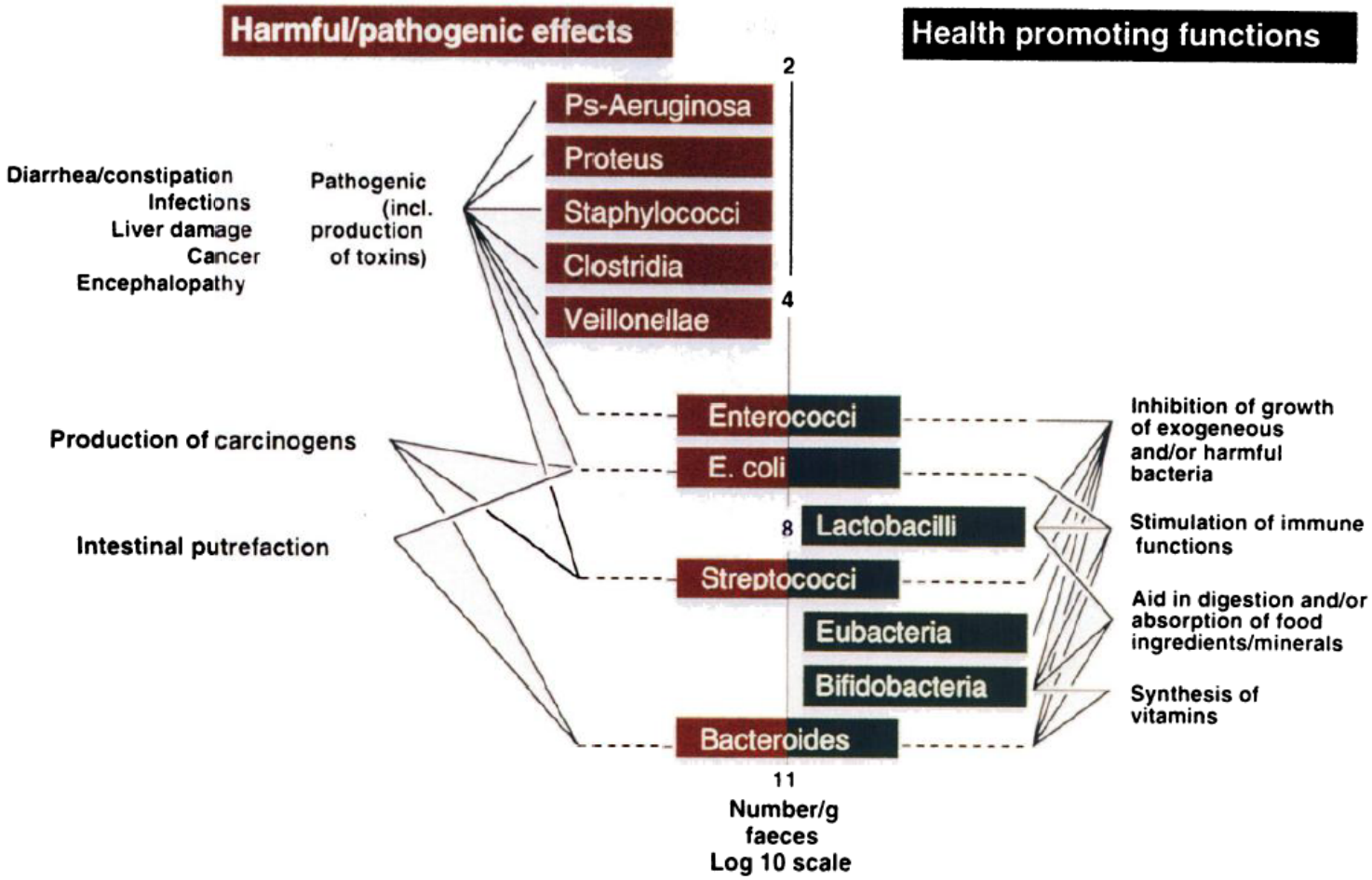
The body‘s microbiomes



Fecal Transplant At Home – DIY Instructions

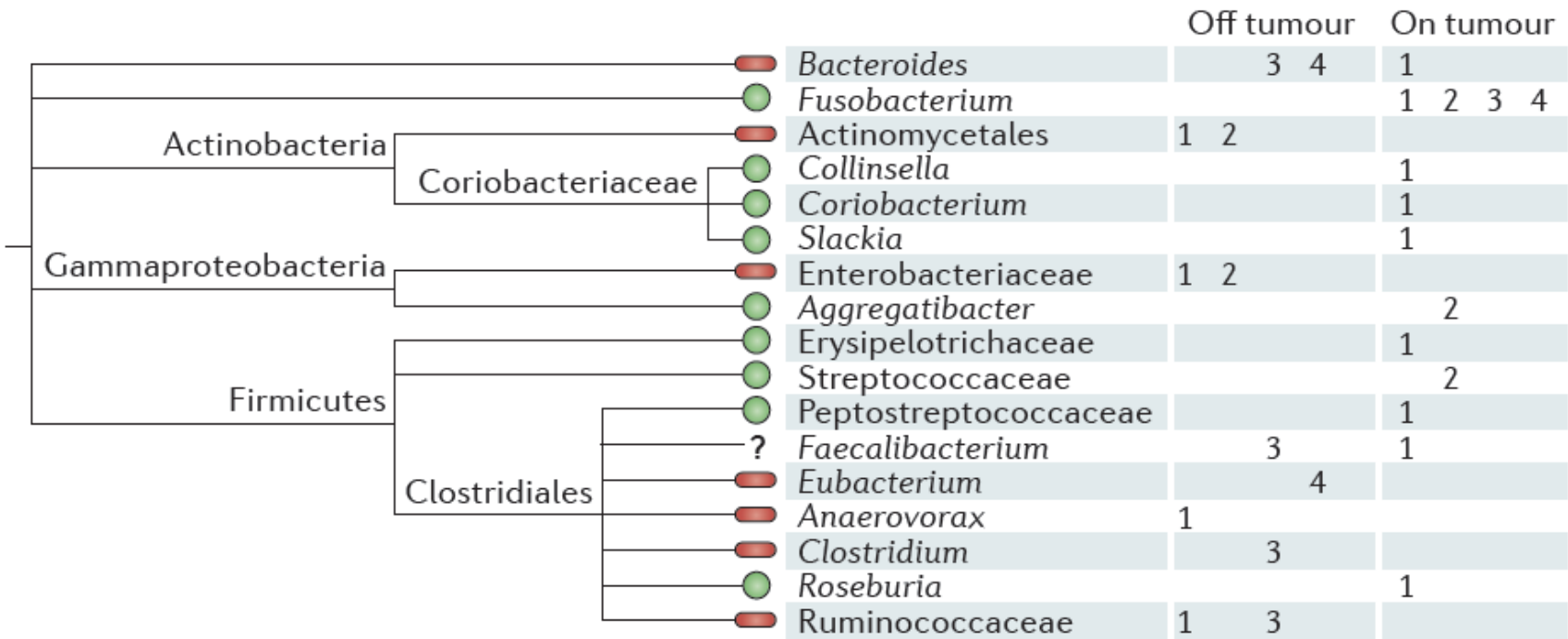
http://thepowerofpoop.com

Composition and health effects of predominant human fecal bacteria



Gibson and Roberfroid, J Nutr 1994

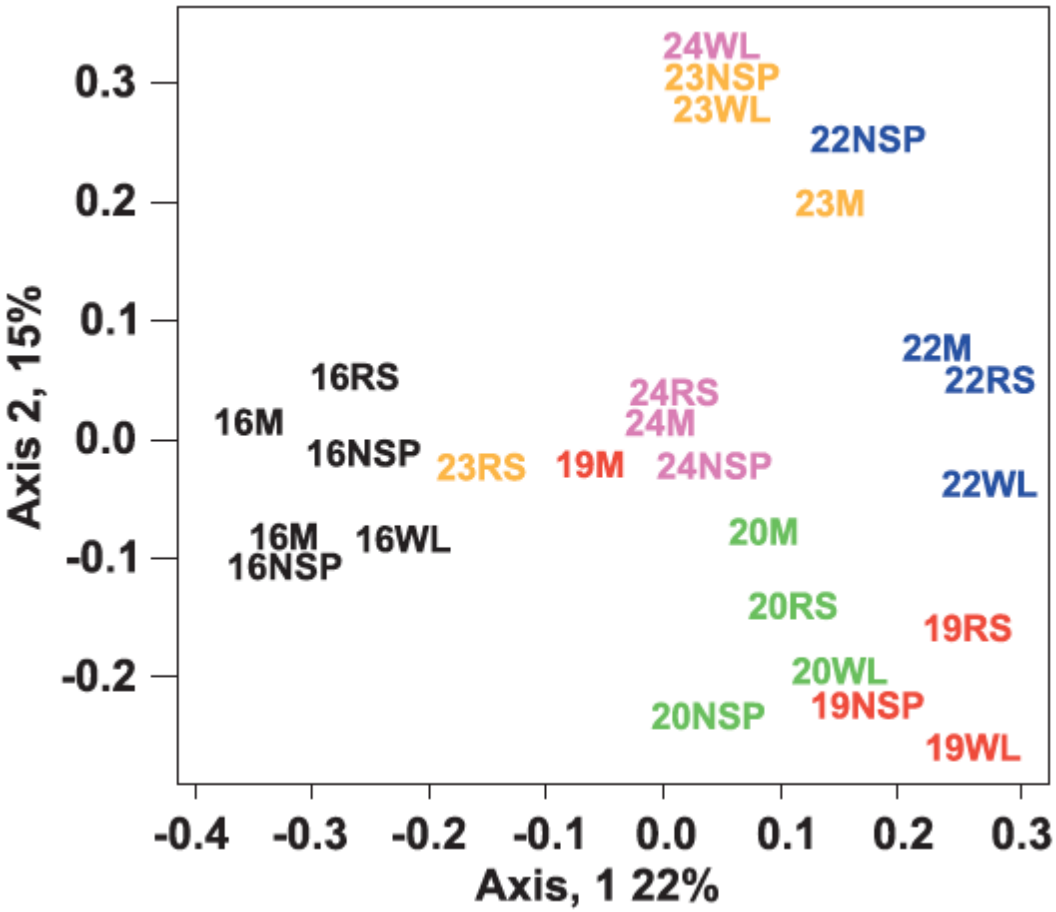
Colorectal cancer-associated microbiomes



1. 16S ribosomal RNA amplicons (V1–V3) from six Dutch individuals
2. Metagenome from nine Spanish, American and Vietnamese individuals
3. 16S rRNA amplicons (V3–V5) from 95 Spanish, American, Vietnamese individuals
4. Metatranscriptome from nine American individuals

Tjalsma et al., Nat Rev Micro 2012

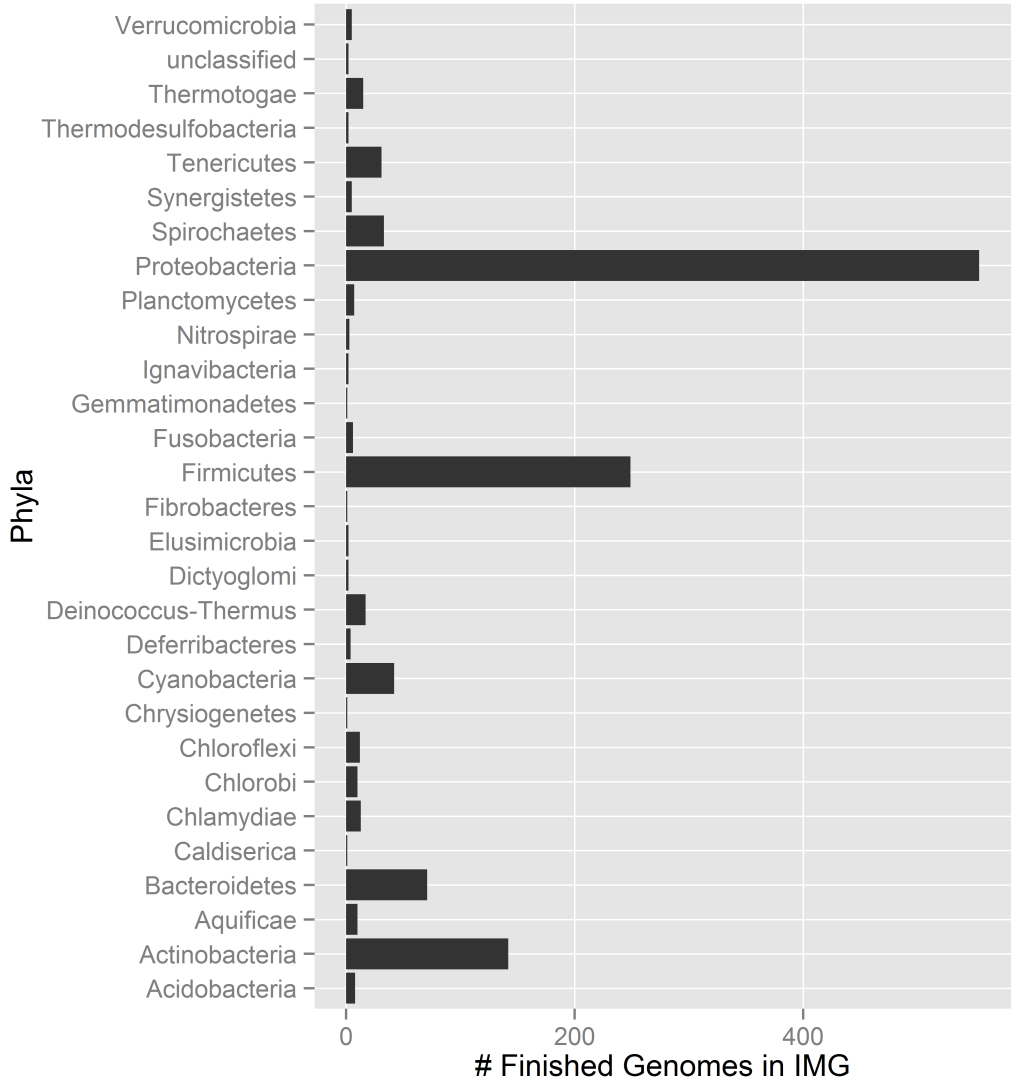
Impact of diet and individual variation upon faecal microbiota composition



M maintenance
NSP non-starch polysaccharide
RS resistant starch
WL weight loss

Walker et al., ISME J 2011

What's in the databases?



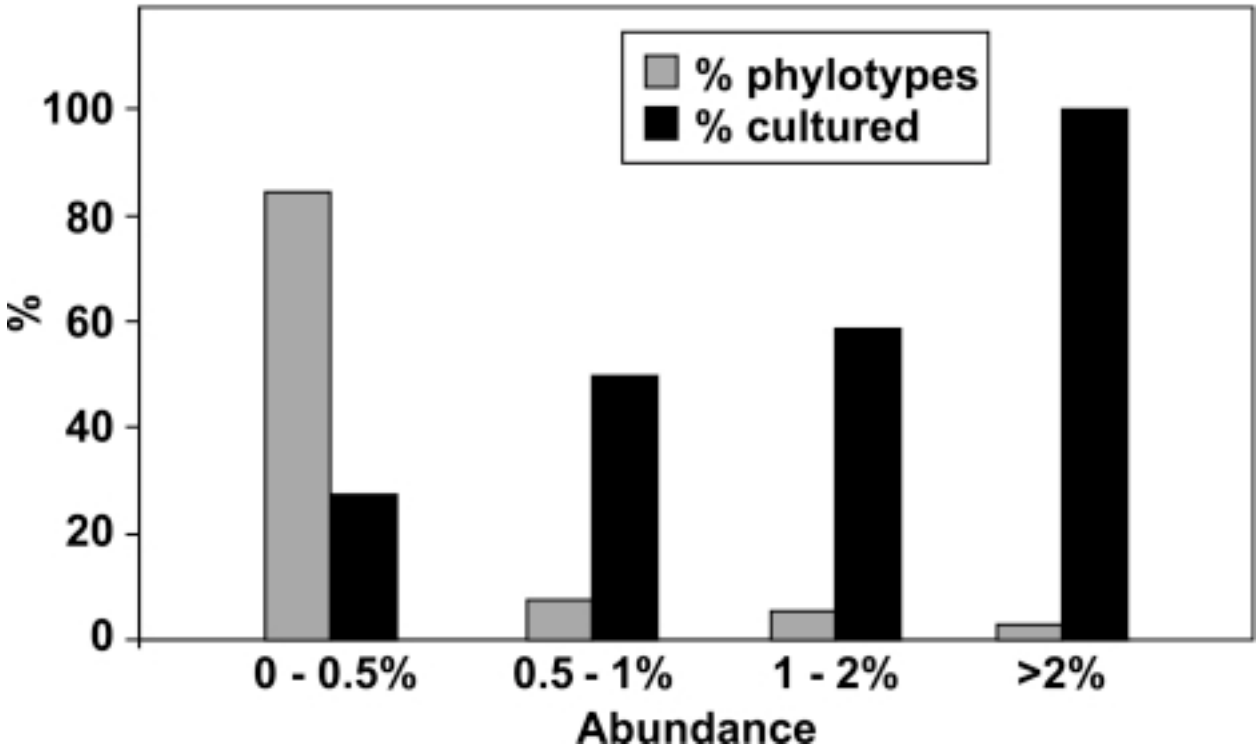
Finshed Genomes in IMG
Vs.
Greengenes 16S rRNA database

	Genomes	16S
Phyla	29	90
Class	46	249
Order	100	405
Species	1268	99322*

*97% clustering

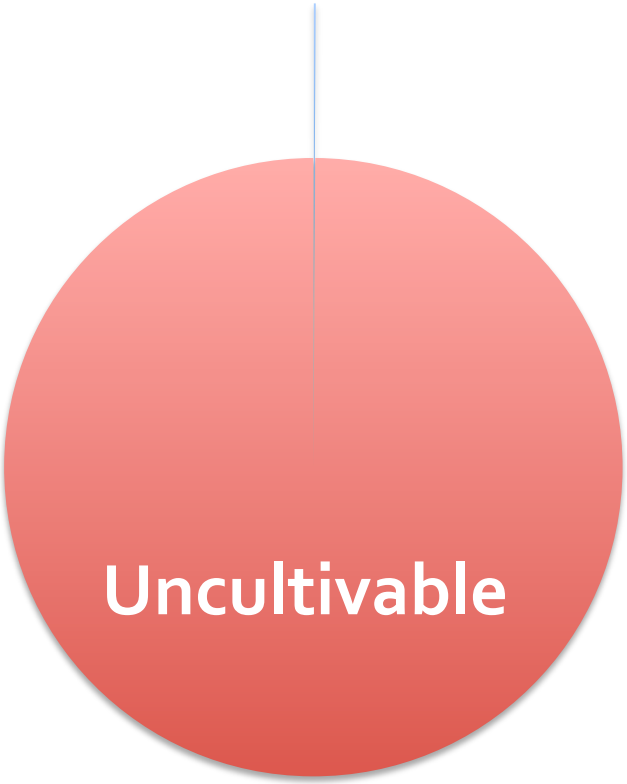
Note: only including 1 strain per species

Culturability of the microbiome



Walker et al., ISME J 2011

Microbial culturability



Aquatic and terrestrial ecosystems

Microbiomes, activated sludge

Amann et al., Microbiol Rev 1995

Genomes?

Cultivation!

Culturomics

ORIGINAL ARTICLE BACTERIOLOGY

Microbial culturomics: paradigm shift in the human gut microbiome study

J.-C. Lagier[1,*], F. Armougom[1,*], M. Million[1], P. Hugon[1], I. Pagnier[1], C. Robert[1], F. Bittar[1], G. Fournous[1], G. Gimenez[1], M. Maraninchi[2], J.-F. Trape[3], E. V. Koonin[4], B. La Scola[1] and D. Raoult[1]

1) *Aix Marseille Université, URMITE, UM63, CNRS 7278, IRD 198, INSERM 1095,* 2) *Service de Nutrition, Maladies Métaboliques et Endocrinologie, UMR-INRA U1260, CHU de la Timone, Marseille, France,* 3) *IRD, UMR CNRS 7278-IRD 198, Route des Pères Maristes, Dakar, Sénégal and* 4) *National Centre for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*

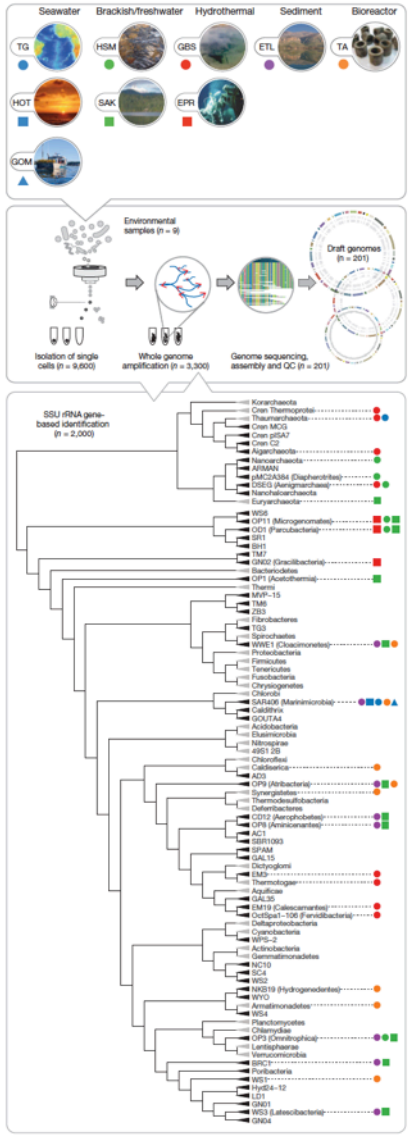
…We studied stool samples from two young lean Africans from a rural environment in Senegal and one obese French individual, using 212 different conditions, including amoebal co-culture…

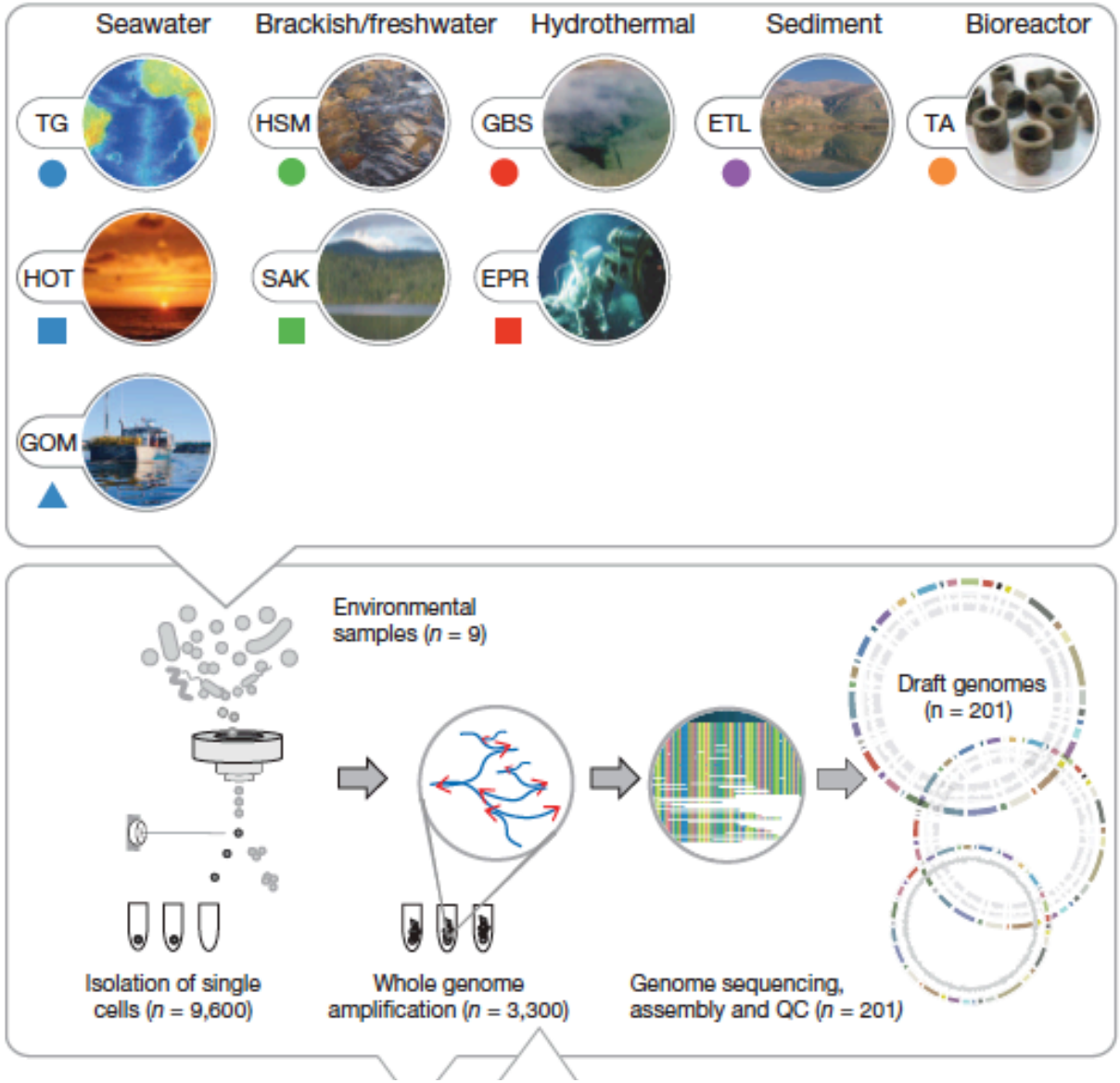
Lagier et al., Clin. Microbiol. Infect. 2012

Genomes?

Single cells!

Microbial dark matter



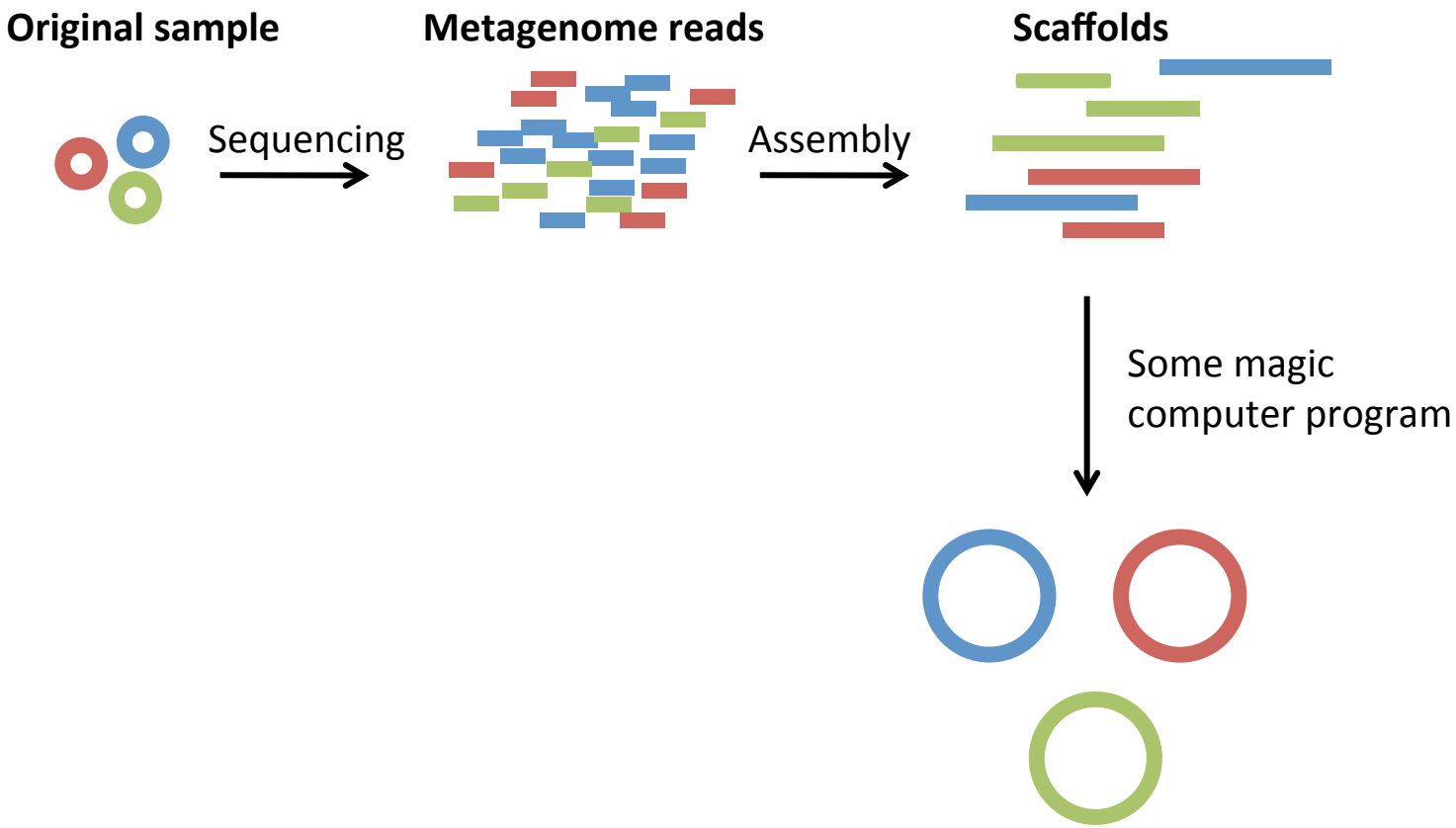


Rinke et al., Nature 2013

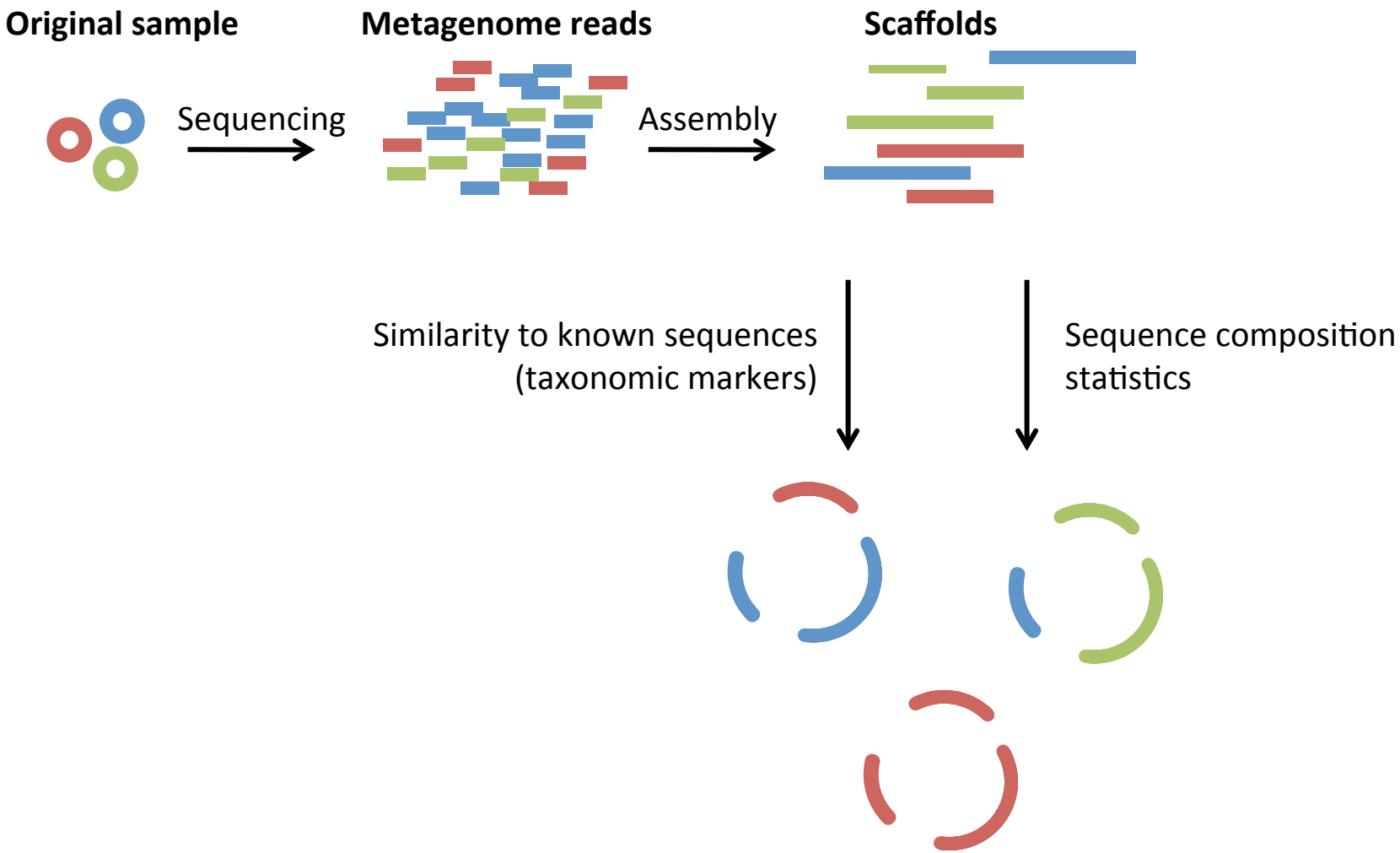
Genomes?

Metagenomics!

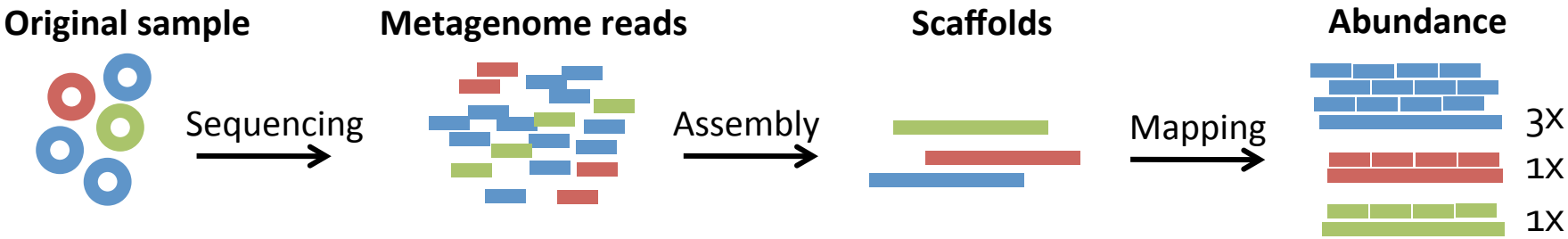
Genomes from metagenomes?

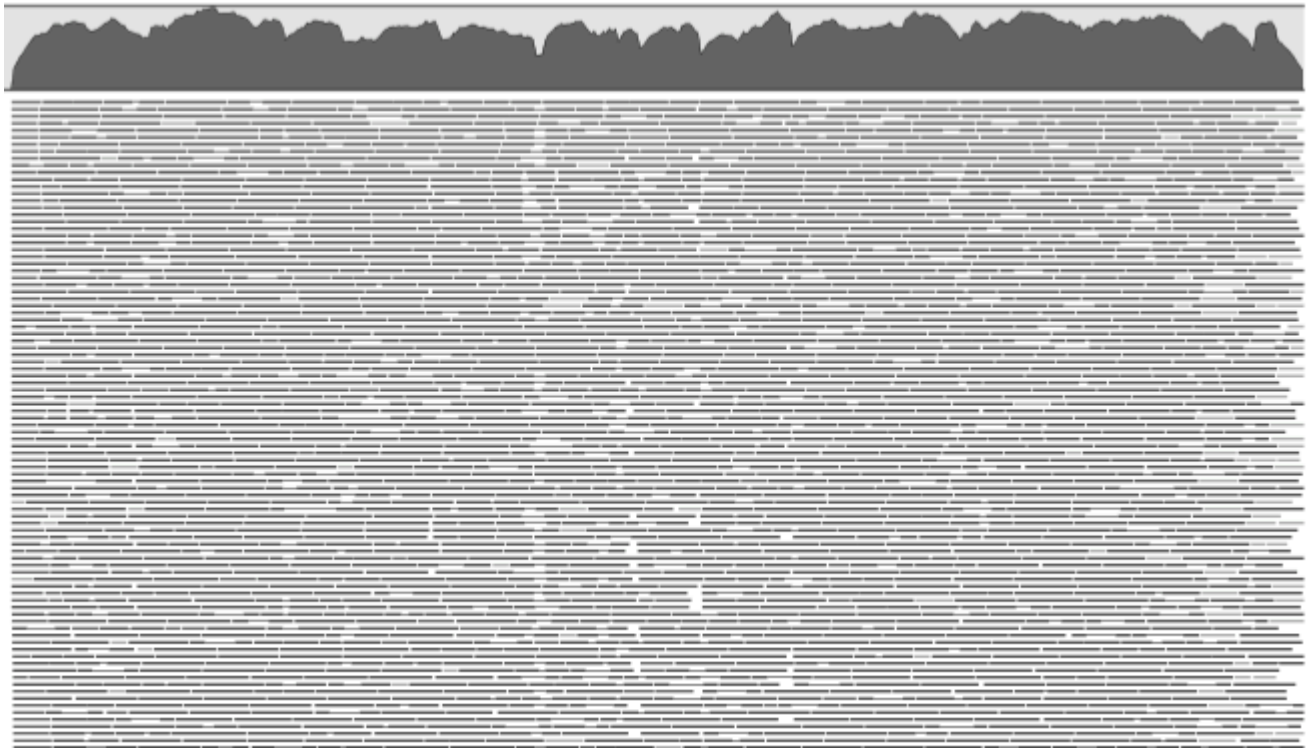


Genomes from metagenomes?

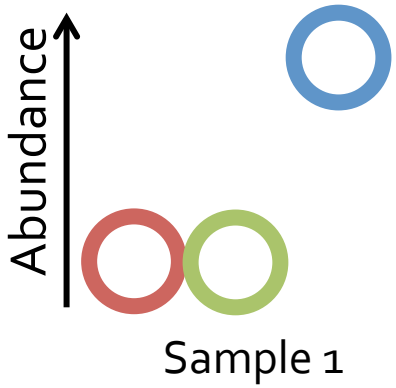


Binning by coverage



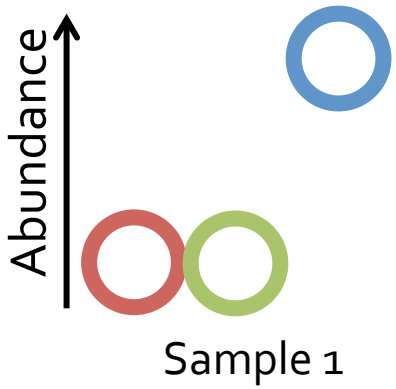


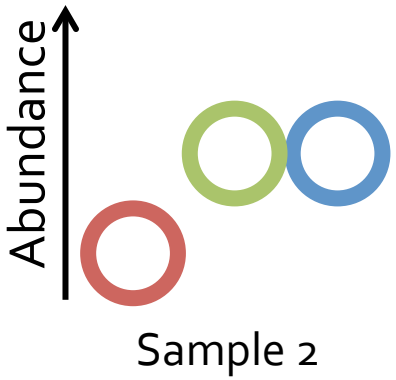
Binning by coverage



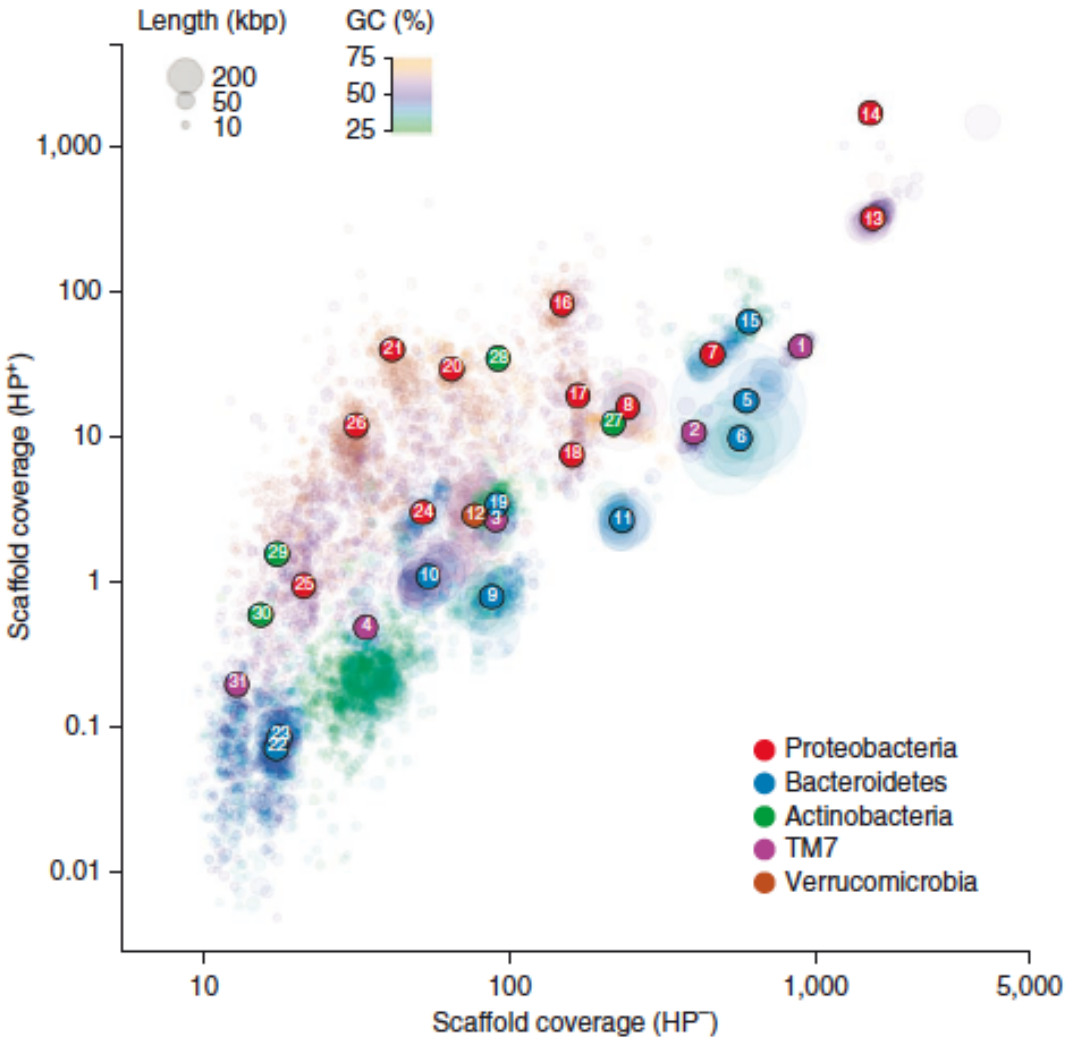
Binning by coverage

Binning by coverage





Differential coverage binning



Albertsen et al., Nature Biotechnology 2013

Upcoming software for genome-centric metagenomics (*incomplete; some unpublished*)

Differential coverage binning:

- mmgenome (Mads Albertsen/Per Nielsen)

Multi-coverage binning:

- GroopM (Michael Imelfort/Gene Tyson)
- CONCOCT (Johannes Alneberg/Christopher Quince)

Automatic evaluation, taxonomic+completeness prediction

- CheckM (Donovan Parks/Gene Tyson)

SOFTWARE EVALUATION CHALLENGE

CAMI

http://cami-challenge.org

Unpublished genome sequences

+

Assemblers

Taxonomy predictors

Binning tools

Evaluation:

- Independent
- Realistic
- Reproducible
- Continuous

Public databases?

Re-Annotation!

BLAST search at NCBI

Query: chlamydial protease like activity factor (CPAF)
[Waddlia chondrophila WSU 86-1044]

Search: BLASTP against NCBI RefSeq database

Description

chlamydial protease-like activity factor (CPAF) [Waddlia chondrophila WSU 86-1044]

putative chlamydial protease-like activity factor [Parachlamydia acanthamoebae str. Hall's coccus] >re

protease-like activity factor [Protochlamydia amoebophila UWE25]

hypothetical protein CAB712 [Chlamydophila abortus S26/3]

hypothetical protein CAB1_0732[Chlamydophila abortus LLG]

Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies

H. James Tripp[1], Ian Hewson[2], Sam Boyarsky[1], Joshua M. Stuart[1] and Jonathan P. Zehr[1,*]

[1]Department of Ocean Sciences, University of California, Santa Cruz, CA 95064, USA and [2]Department of Microbiology, Cornell University, Wing Hall 403, Ithaca, NY 14853, USA

Received March 24, 2011; Revised June 24, 2011; Accepted June 27, 2011

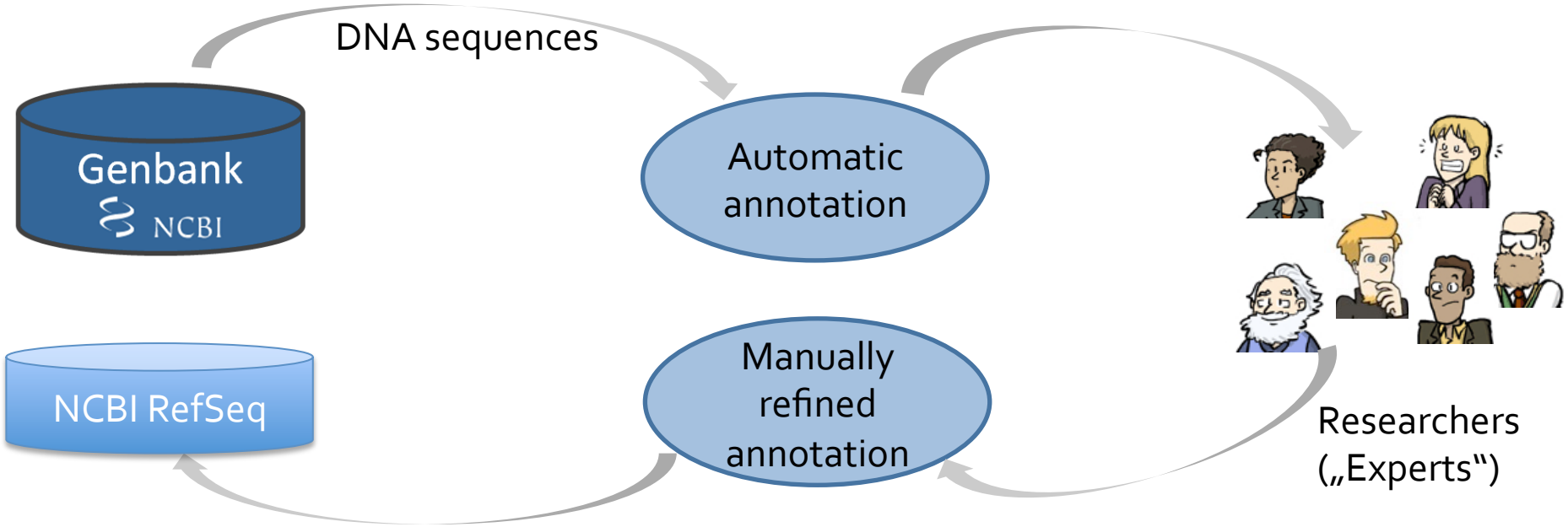
ABSTRACT

In the course of analyzing 9 522 746 pyrosequencing reads from 23 stations in the Southwestern Pacific and equatorial Atlantic oceans, it came to our attention that misannotations of rRNA as proteins is now so widespread that false positive matching of rRNA pyrosequencing reads to the National Center for Biotechnology Information (NCBI) non-redundant protein database approaches 90%. One conserved portion of 23S rRNA was consistently misannotated often enough to prompt curators at Pfam to create a spurious protein family. Detailed examination of the annotation history of each seed sequence in the spurious Pfam protein family (PF10695, 'Cw-hydrolase') uncovered issues in the standard operating procedures and quality assurance programs of major sequencing centers, and other issues relating to the curation practices of those managing public databases such as GenBank and SwissProt. We offer recommendations for all these issues, and recommend as well that workers in the field of metatranscriptomics take extra care to avoid including false positive matches in their datasets.

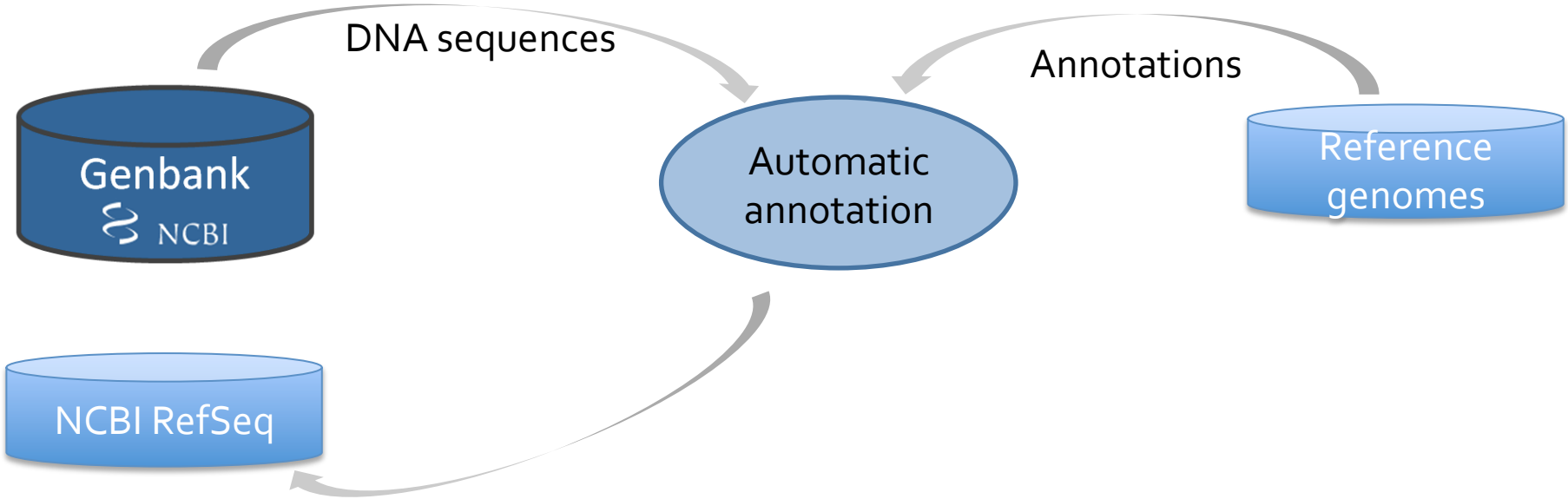
operons of *Escherichia coli* were published between 1967 and 1978 (7–10). The rRNA nucleotide sequences for *Saccharomyces cerevisiae,* which occur in ~140 tandem repeats, were published between 1972 and 1981 (11–14).

While artificial overexpression of a pentapeptide sequence adjacent to a Shine–Dalgarno motif within *E. coli* 23S rRNA was found to impart drug resistance to erythromycin (15), rRNA operons in Bacteria and Archaea are not known to contain naturally expressed protein coding regions that also code for rRNA. Also, while antisense transcription was recently reported for Bacterial and Archaeal proteins, that study did not report antisense transcription from Bacteria and Archaea rRNA (16). To be sure, insertion elements can be found in rRNA operons of Bacteria and Archaea, but not sequences that code for rRNA and protein at the same time. Therefore, annotations of Bacteria and Archaea proteins embedded in rRNA operons and overlapping with rRNA coding regions within those operons have been rightly presumed to be misannotations (17) and should continue to be, until hard evidence to the contrary emerges. While these misannotations continue to exist, they have the potential to generate false positive matches of translated environmental rRNA sequences to proteins. To our knowledge, the potential for false positives in metatranscriptomic studies due to misannotations of rRNA operons has not

NCBI RefSeq re-annotation initiatives: reference genomes



NCBI RefSeq re-annotation initiatives: other genomes



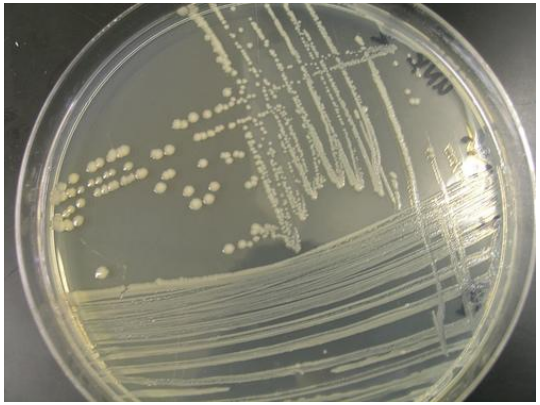
Understanding genomes?

Phenotype prediction!

Why?

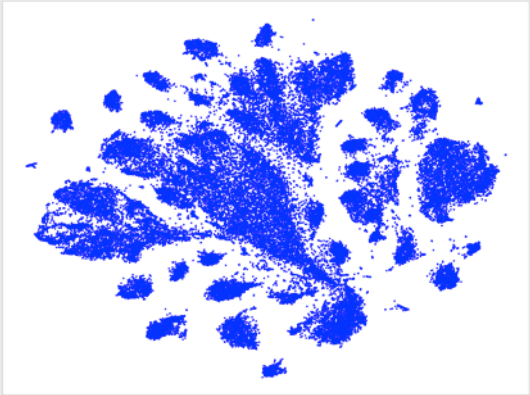
Classically

1. Interesting phenotype
2. Cultures/enrichments
3. Sequencing

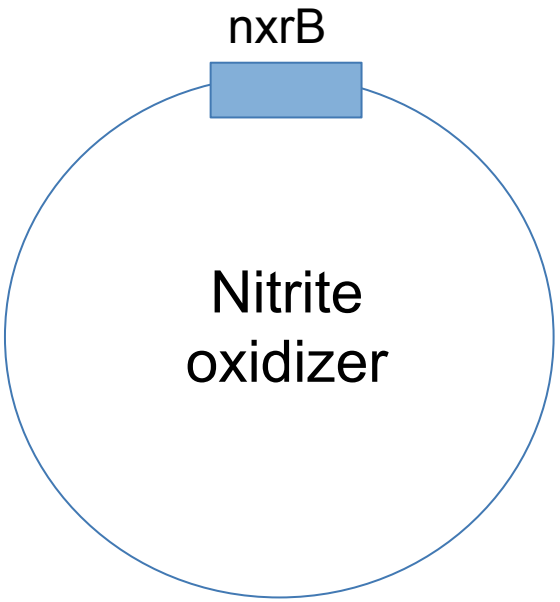


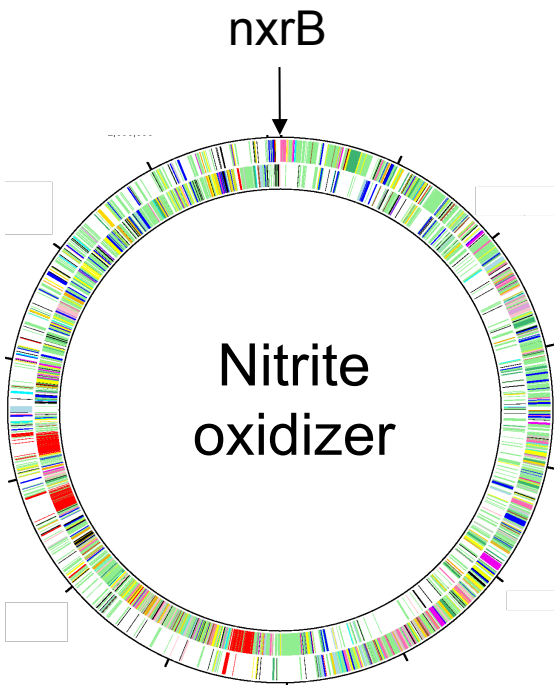
Emerging

1. Interesting habitat
2. Sequencing
3. Metagenomes



State-of-the-art (1)

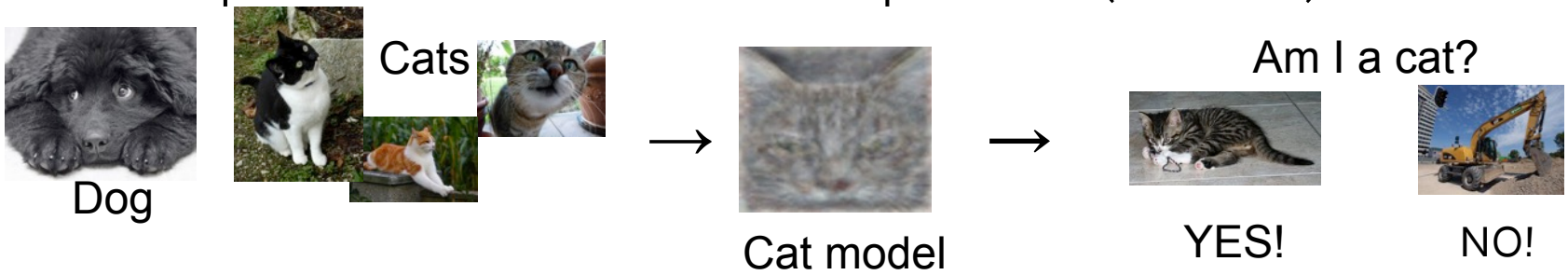




State-of-the-art (2)

Machine learning

Example data → model → prediction (new data)



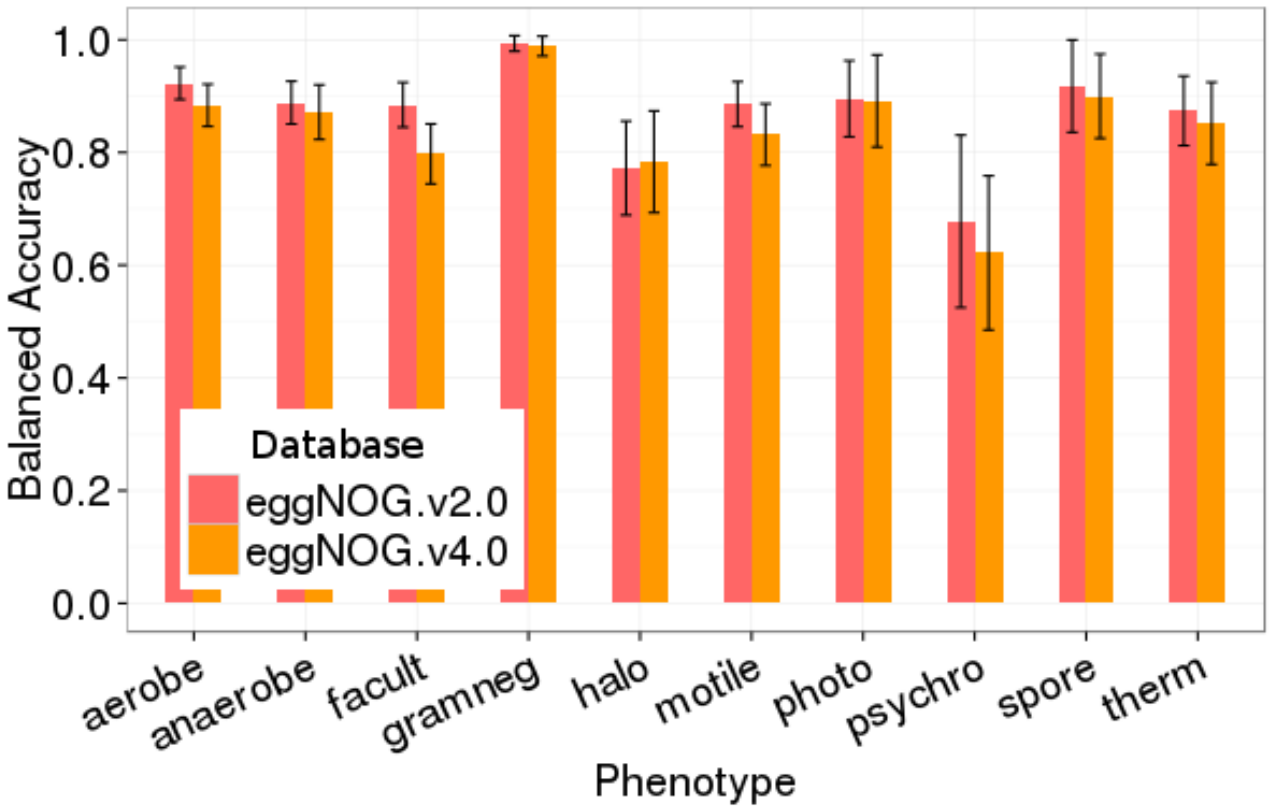
Software: PICA

- Techniques:
  Association rule mining, Support vector machines
- Phenotypes: eggNOG 2 data

Roman Feldbauer

Does it work in 2015?

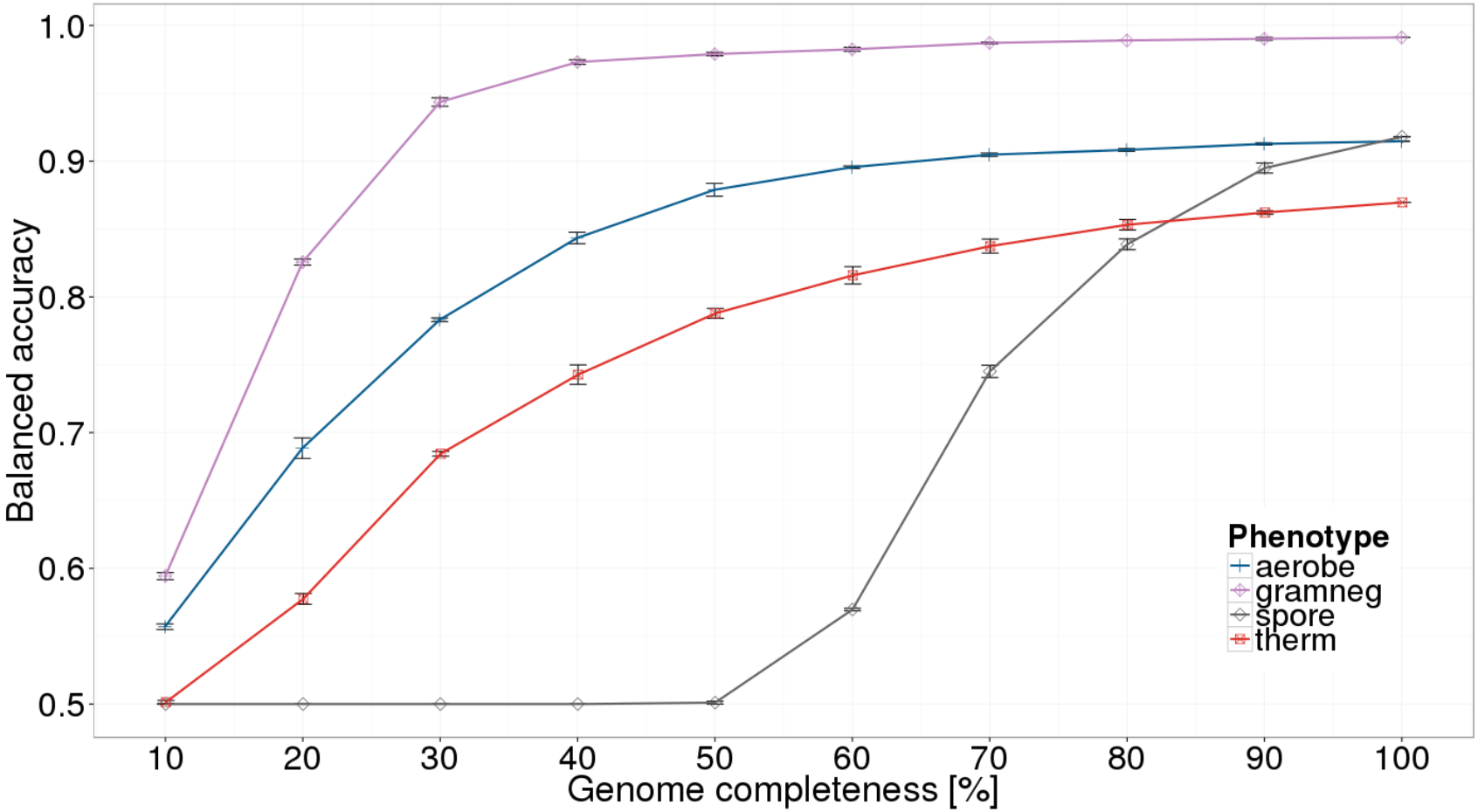
- Today: more genomes, more COGs (eggNOG 4)
- Improvement of SVM plugin



Accuracy ✔

Also:
CPU ✔
Memory ✔

Does it work on incomplete genomes?



~70% completeness often sufficient ✔

Modeling metabolic traits

Methanotroph	TRAIT	Nitrifier
pmoA mmoX mxaF	MARKERS	amoA nxrB
(+) 33 (-) 86	GENOMES in training set	(+) 35 (-) 340
97.2 ± 4.6 %	Prediction ACCURACY	97.7 ± 4.7 %
1. pmoA 4. mmoX 5. uncharacterized protein 17. mxaF	Proteins of highest PREDICTIVE POWER (novel feature ranking mechanism)	1. DUF2024 (structural similarities to nitrogen regulatory protein P-II and nitrogen fixation protein NifU)

Expected markers ✔ + new associations ✔

Modeling complex traits

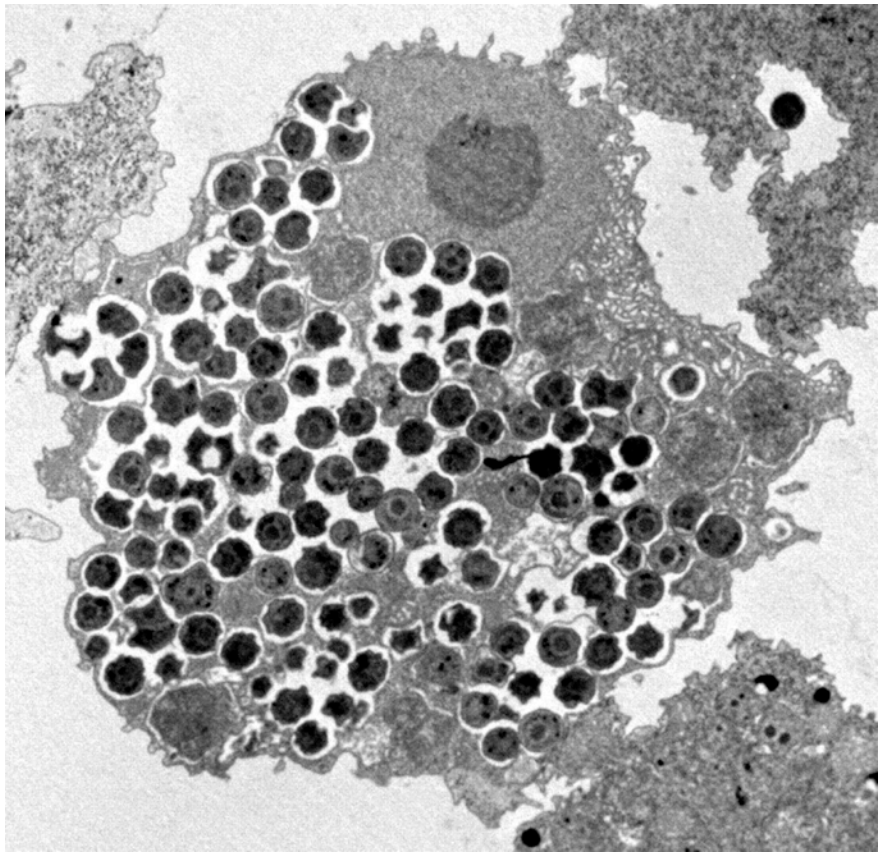
Intracellular lifestyle

- Genome reduction?
- (+) 76, (-) 56 genomes

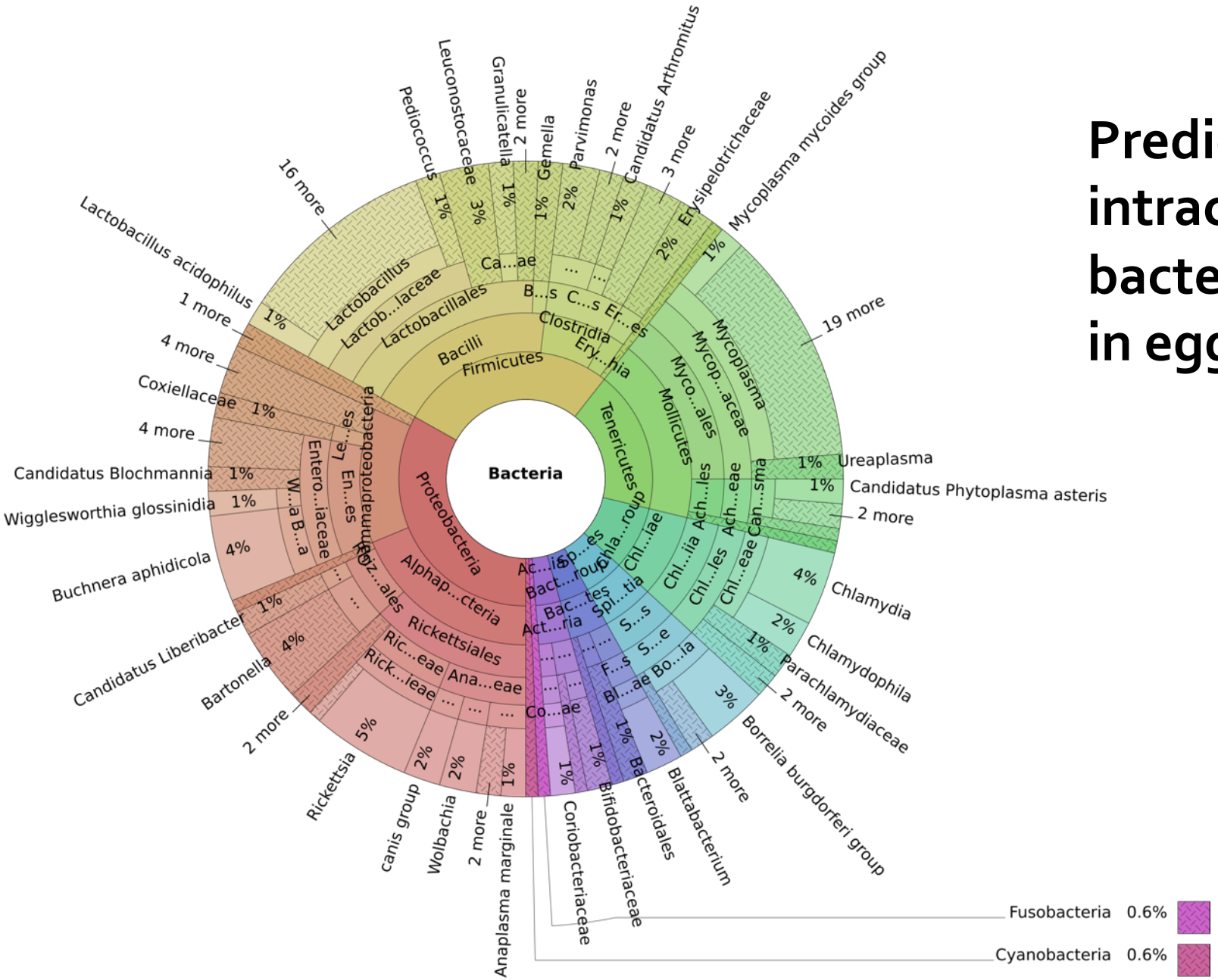
- Accuracy 96.7 ± 4.1 %

- Top 50 features:

48 negative predictors



Modeling complex traits



Predicted intracellular bacteria in eggNOG4

Outlook

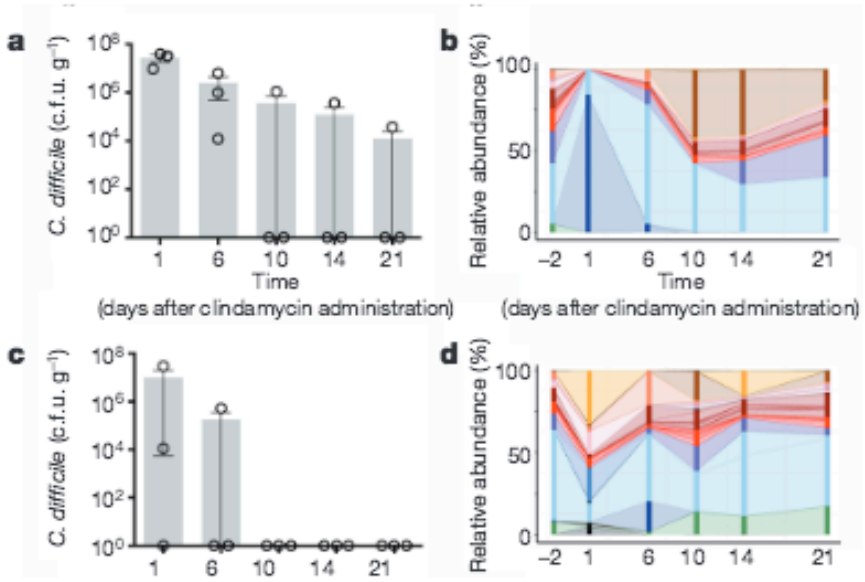
- Genome-centric metagenomics
- Independent method evaluation in metagenomics
- Genome re-annotation in public databases
- Computational prediction of simple and complex traits
  - Automatic analysis of metagenomes
  - Continuous annotation of genomes/bins from public databases

Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*

Charlie G. Buffie[1,2], Vanni Bucci[3,4], Richard R. Stein[3], Peter T. McKenney[1,2], Lilan Ling[2], Asia Gobourne[2], Daniel No[2], Hui Liu[5], Melissa Kinnebrew[1,2], Agnes Viale[6], Eric Littmann[2], Marcel R. M. van den Brink[7,8], Robert R. Jenq[7], Ying Taur[1,2], Chris Sander[3], Justin R. Cross[5], Nora C. Toussaint[2,3], Joao B. Xavier[2,3] & Eric G. Pamer[1,2,8]

The gastrointestinal tracts of mammals are colonized by hundreds of microbial species that contribute to health, including colonization resistance against intestinal pathogens[1]. Many antibiotics destroy intestinal microbial communities and increase susceptibility to intestinal pathogens[2]. Among these, *Clostridium difficile*, a major cause of antibiotic-induced diarrhoea, greatly increases morbidity and mortality in hospitalized patients[3]. Which intestinal bacteria provide resistance to *C. difficile* infection and their *in vivo* inhibitory mechanisms remain unclear. Here we correlate loss of specific bacterial taxa with development of infection, by treating mice with different antibiotics that result in distinct microbiota changes and lead to varied susceptibility to *C. difficile*. Mathematical modelling augmented by analyses of the microbiota of hospitalized patients identifies resistance-associated bacteria common to mice and humans. Using these platforms, we determine that *Clostridium scindens*, a bile acid 7α-dehydroxylating intestinal bacterium, is associated with resistance to *C. difficile* infection and, upon administration, enhances resistance to infection in a secondary bile acid dependent fashion. Using a workflow involving mouse models, clinical studies, metagenomic analyses, and mathematical modelling, we identify a probiotic candidate that corrects a clinically relevant microbiome deficiency. These findings have implications for the rational design of targeted antimicrobials as well as microbiome-based diagnostics and therapeutics for individuals at risk of *C. difficile* infection.

microbiota alpha diversity (that is, diversity within individuals) (Fig. 2a), consistent with previous studies[6]. Using weighted UniFrac[7] distances to evaluate beta diversity (that is, diversity between individuals), we found that although clindamycin and ampicillin administration induced distinct changes in microbiota structure, recovery of resistance corresponded with return to a common coordinate space shared by antibiotic-naive animals (Fig. 2b). However, these diversity metrics generally did not resolve the susceptibility status of animals harbouring microbiota with



Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown[1], Laura A. Hug[2], Brian C. Thomas[2], Itai Sharon[2], Cindy J. Castelle[2], Andrea Singh[2], Michael J. Wilkins[3,4], Kelly C. Wrighton[4], Kenneth H. Williams[5] & Jillian F. Banfield[2,5,6]

A prominent feature of the bacterial domain is a radiation of major lineages that are defined as candidate phyla because they lack isolated representatives. Bacteria from these phyla occur in diverse environments[1] and are thought to mediate carbon and hydrogen cycles[2]. Genomic analyses of a few representatives suggested that metabolic limitations have prevented their cultivation[2–6]. Here we reconstructed 8 complete and 789 draft genomes from bacteria representing >35 phyla and documented features that consistently distinguish these organisms from other bacteria. We infer that this group, which may comprise >15% of the bacterial domain, has shared evolutionary history, and describe it as the candidate phyla radiation (CPR). All CPR genomes are small and most lack numerous biosynthetic pathways. Owing to divergent 16S ribosomal RNA (rRNA) gene sequences, 50–100% of organisms sampled from specific phyla would evade detection in typical cultivation-independent surveys. CPR organisms often have self-splicing introns and proteins encoded within their rRNA genes, a feature rarely reported in bacteria. Furthermore, they have unusual ribosome compositions. All are missing a ribosomal protein often absent in symbionts, and specific lineages are missing ribosomal proteins and biogenesis factors considered universal in bacteria. This implies different ribosome structures and biogenesis mechanisms, and underlines unusual biology across a large part of the bacterial domain.

to previously unrecognized lineages (CPR1–3; Fig. 1). In total, 789 draft-quality (≥50% complete) genomes were reconstructed (Table 1). We manually curated eight genomes to completion: the first three from Microgenomates, two from Parcubacteria, one each from Kazan and Berkelbacteria, and an additional genome from Saccharibacteria. All complete and draft genomes are small and most are <1 Mb in length (Supplementary Tables 3 and 4).

In total, 1,543 bacterial 16S rRNA genes ≥800 bp were assembled and curated to eliminate assembly errors (713 sequences clustered at 97% identity; Supplementary Data 1). Relative abundance measurements show enrichment of CPR organisms in small-cell filtrates, suggesting that they have ultra-small cells (Extended Data Fig. 3). This finding is supported by a recent microscopy study[8]. Surprisingly, 31% of 16S rRNA genes encoded a large (≥10 bp) insertion sequence (maximum 2,004 bp; mean 519 bp; standard deviation (s.d.) 372 bp; Supplementary Table 5). Insertions are found in phylogenetically diverse members of CPR phyla (Fig. 1, Supplementary Fig. 1 and Supplementary Data 2). Insertion sites are clustered in several distinct locations on the 16S rRNA gene, both in variable and conserved regions (Fig. 2). Most insertions ≥500 bp encode a catalytic RNA intron (group I or II) and/or an open reading frame (ORF), suggesting that they are self-splicing. Encoded proteins frequently belong to families of homing endonucleases (LAGLIDAG 1–3 and GIY-YIG). However, 25% are not similar to known protein families or to each