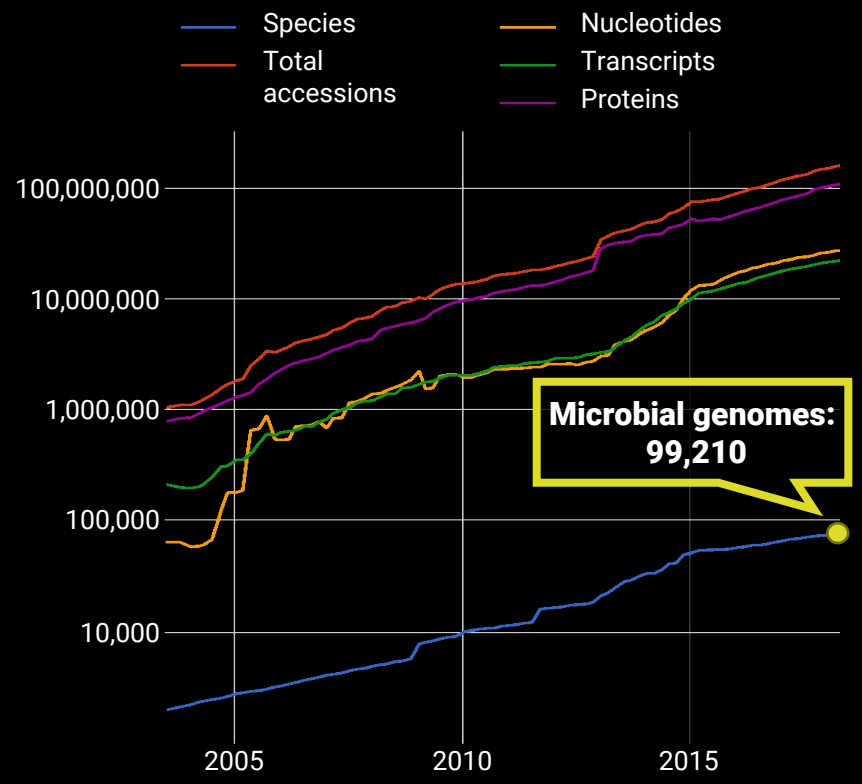
Multi-resolution phylogenetic analysis of known and unexplored members of the human microbiome with PhyloPhlAn 2

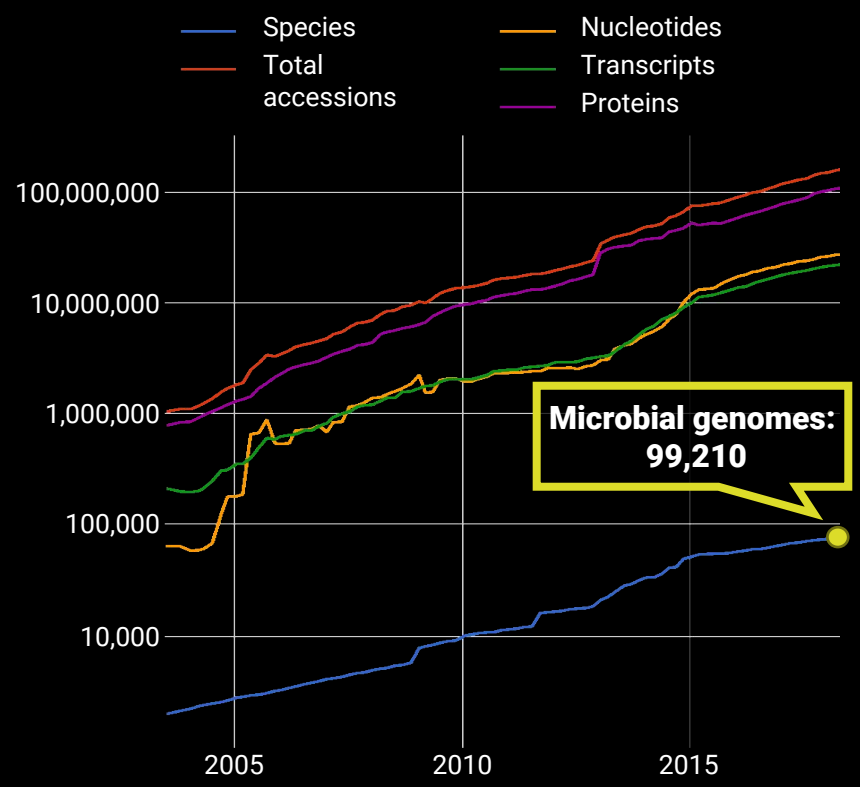
3rd Workshop on Statistical and Algorithmic Challenges in Microbiome Data Analysis

Francesco Asnicar - University of Trento
Computational Metagenomics lab - Prof. Nicola Segata

Increasing availability of genomes from isolates and MAGs

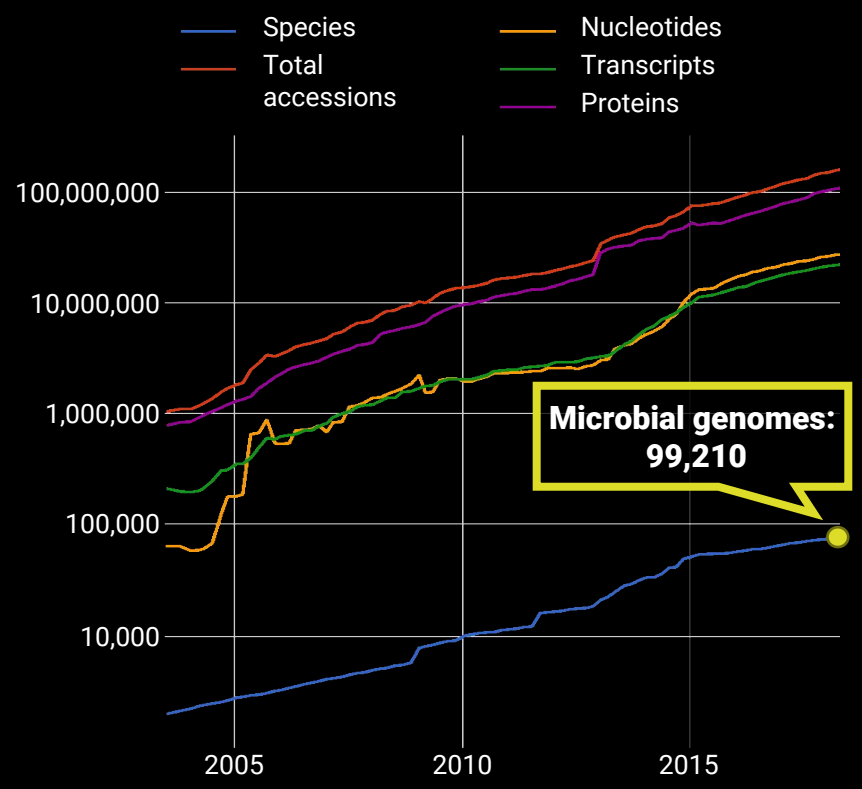


Increasing availability of genomes from isolates and MAGs



10 most sequenced species	# gen.
Streptococcus pneumoniae	>7500
Staphylococcus aureus	>6500
Escherichia coli	>5000
Mycobacterium tuberculosis	>3800
Pseudomonas aeruginosa	>2000
Salmonella enterica	>1700
Acinetobacter baumannii	>1400
Klebsiella pneumonia	>1100
Campylobacter coli	>1000
Neisseria meningitidis	>1000

Increasing availability of genomes from isolates and MAGs



10 most sequenced

Streptococcus pneum

Staphylococcus aure

Escherichia coli

Mycobacterium tube

Pseudomo

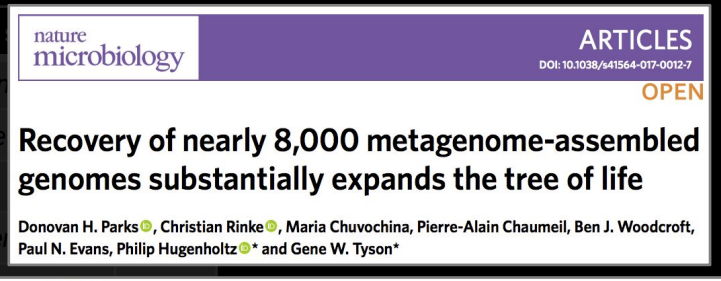
Salmonella

Acinetobac

Klebsiella p

Campyloba

Neisseria meni



nature microbiology

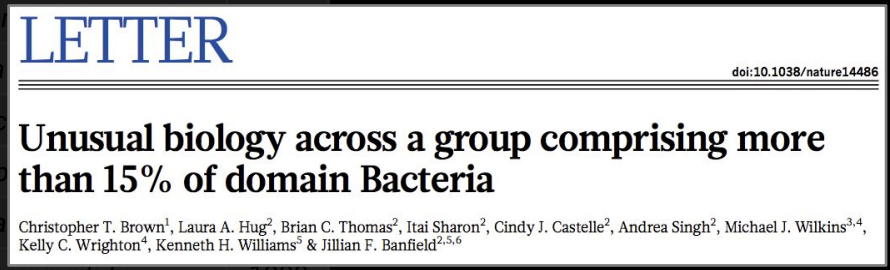
ARTICLES

DOI: 10.1038/s41564-017-0012-7

OPEN

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

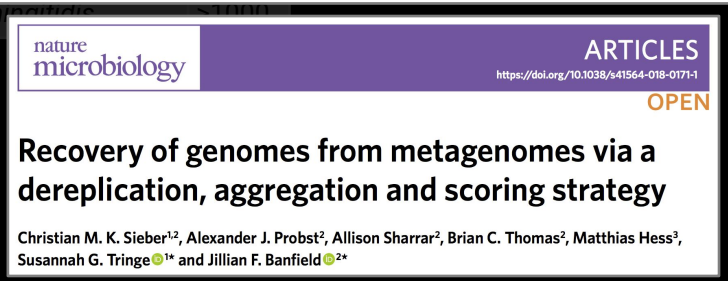


LETTER

doi:10.1038/nature14486

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]



nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-018-0171-1

OPEN

Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy

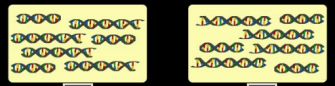
Christian M. K. Sieber[1,2], Alexander J. Probst[2], Allison Sharrar[2], Brian C. Thomas[2], Matthias Hess[3], Susannah G. Tringe[1*] and Jillian F. Banfield[2*]

Large-scale whole-genome comparative genomics

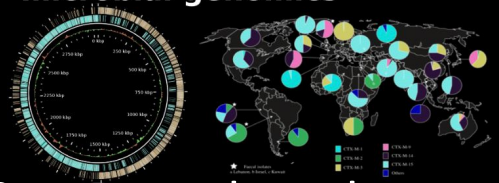
Comparative microbial genomics

Isolate and cultivate

Genome sequencing



Strain-level microbial genomics

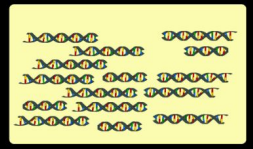


Compare to understand

- High resolution and effective
- But available for few microbes only

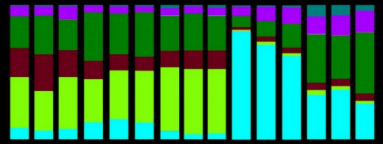
Shotgun metagenomics

Metagenome sequencing



Computational analysis

Microbiome structure comparison



Microbial ecology *in vivo*

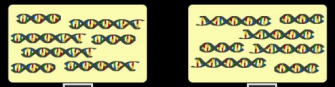
- Access uncultivable microbes (>95%)
- But strain level genomics not possible

Large-scale whole-genome comparative genomics

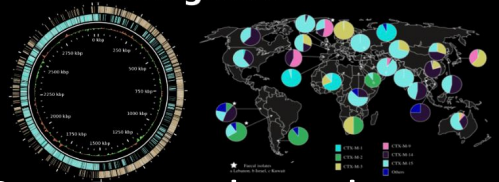
Comparative microbial genomics

Isolate and cultivate

Genome sequencing



Strain-level microbial genomics

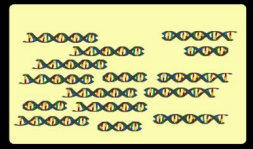


Compare to understand

- High resolution and effective
- But available for few microbes only

Shotgun metagenomics

Metagenome sequencing

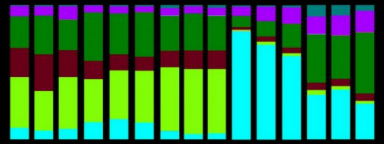


Step 1

Computational analysis

Step 2

Microbiome structure comparison



Microbial ecology *in vivo*

- Access uncultivable microbes (>95%)
- But strain level genomics not possible

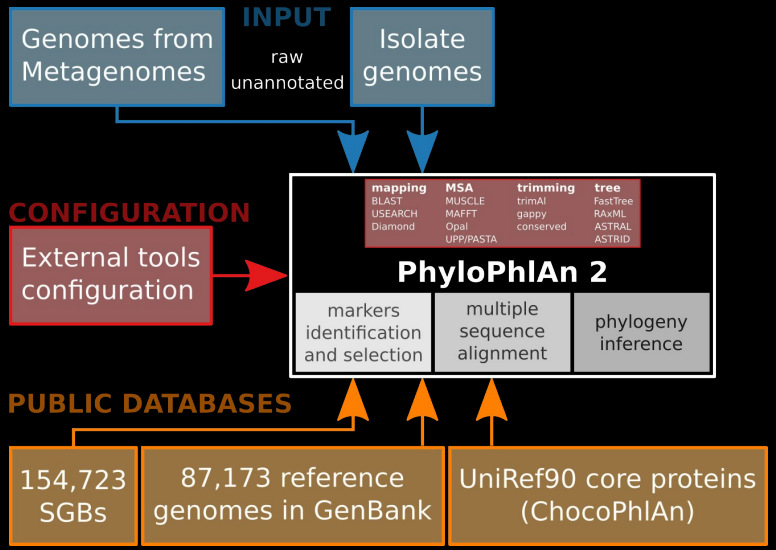
PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs

Main features:

- Scalable, flexible, automatic, modular, customizable



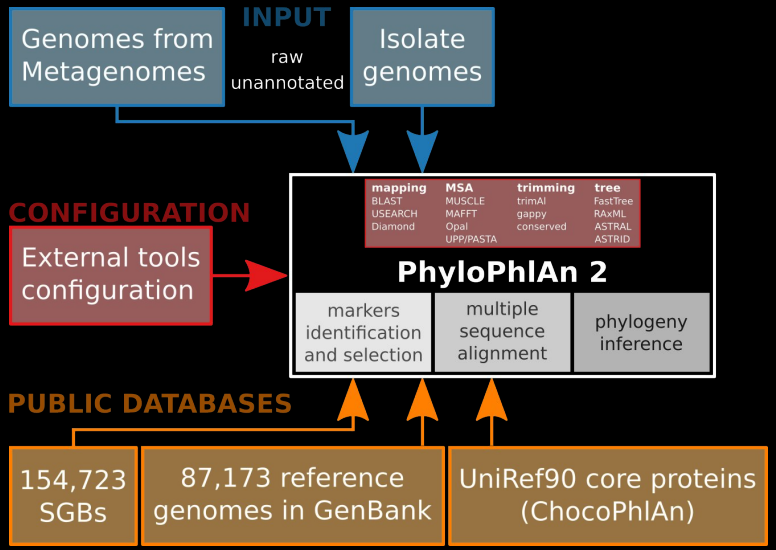
PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs

Main features:

- Scalable, flexible, automatic, modular, customizable



MSAs QC	MSAs subsample	trimming	other QCs
- several scoring position functions implemented, e.g. MUSCLE and trident	- several subsampling approaches implemented	- trimAl - gappy positions - conserved positions - all of the above	- discarding low-quality inputs - discarding low-quality markers - discarding fragmenatary sequences extracted from inputs

PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs

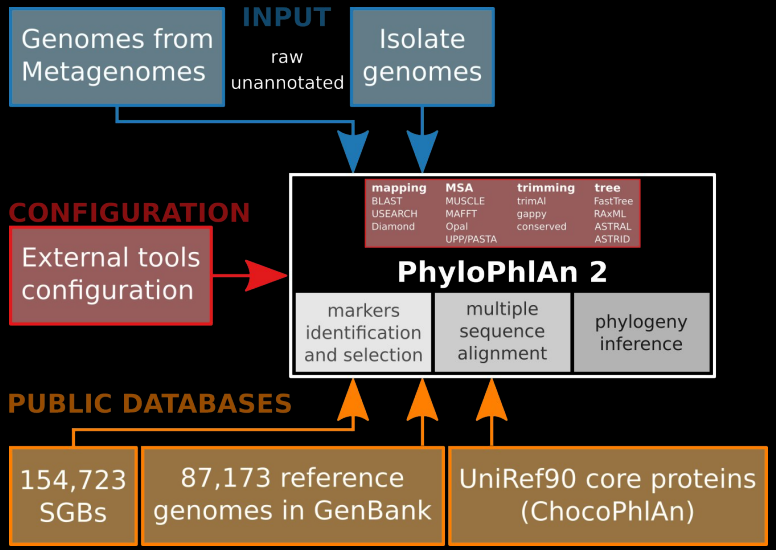
Main features:

- Scalable, flexible, automatic, modular, customizable

Examples of use-cases:

Tree-of-life size phylogenies

High number of inputs
Universal markers



PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs

Main features:

- Scalable, flexible, automatic, modular, customizable

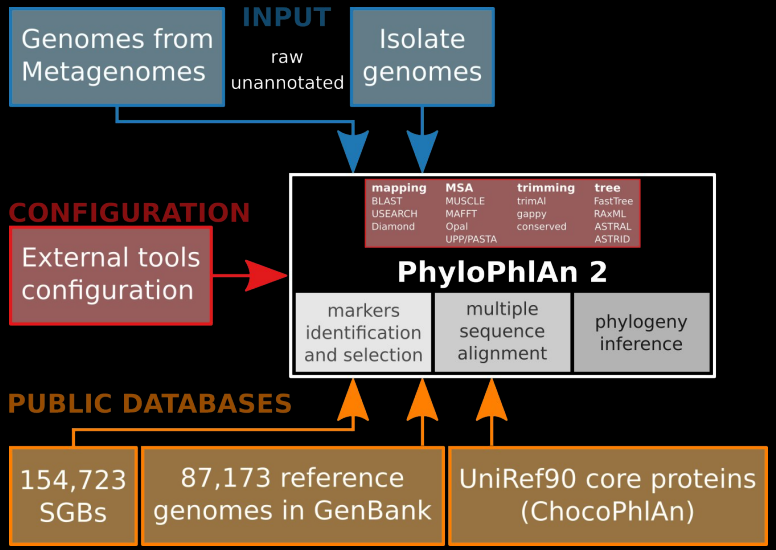
Examples of use-cases:

Tree-of-life size phylogenies

High number of inputs
Universal markers

Species-level phylogenies

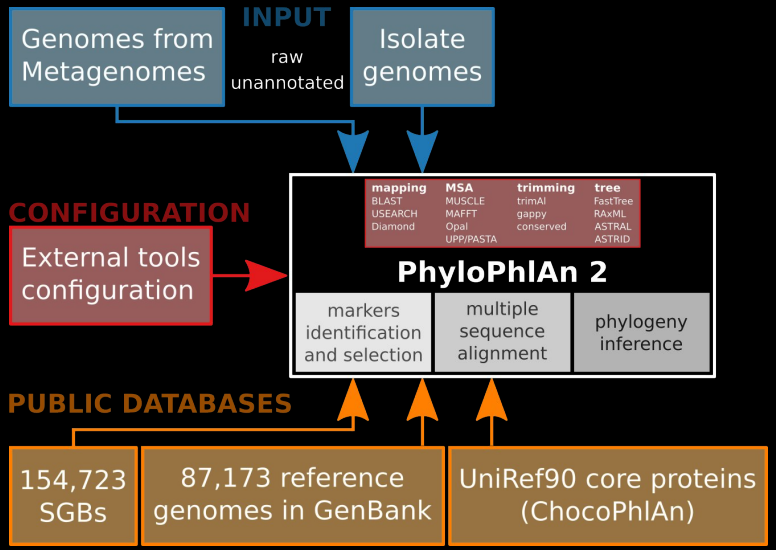
Species-specific markers



PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs



Main features:

- Scalable, flexible, automatic, modular, customizable

Examples of use-cases:

Tree-of-life size phylogenies

High number of inputs
Universal markers

Species-level phylogenies

Species-specific markers

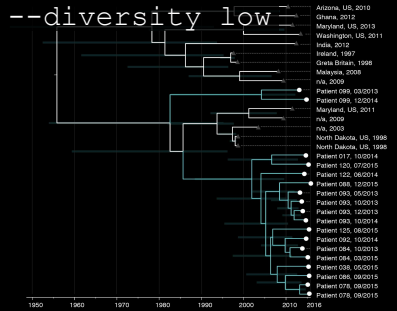
Metagenomic application

Newly assembled MAGs

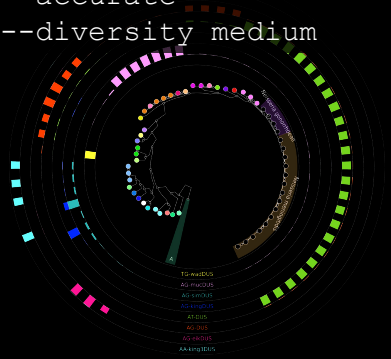
PhyloPhlAn (version 2)

INCREASING RESOLUTION

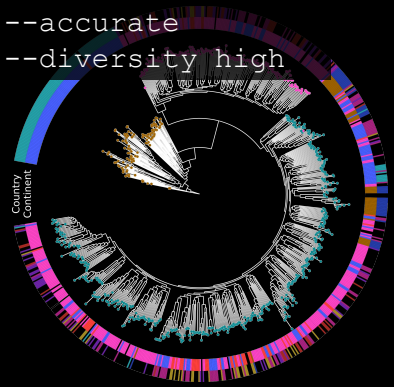
--accurate
--diversity low



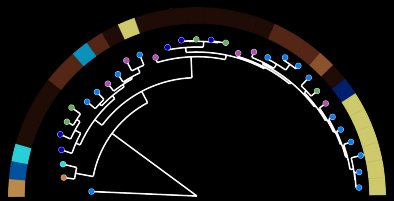
--accurate
--diversity medium



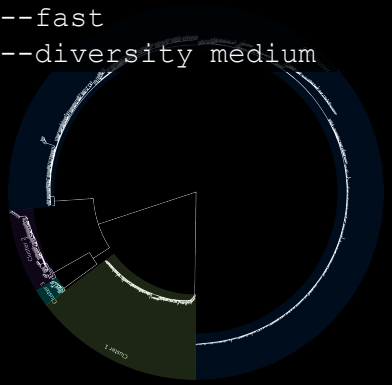
--accurate
--diversity high



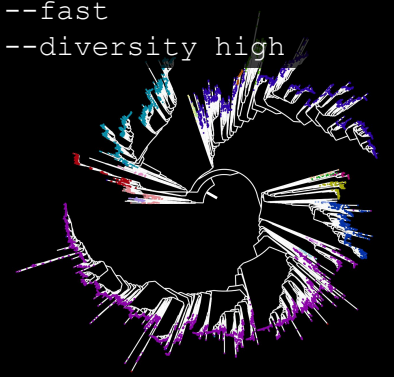
--fast
--diversity low



--fast
--diversity medium

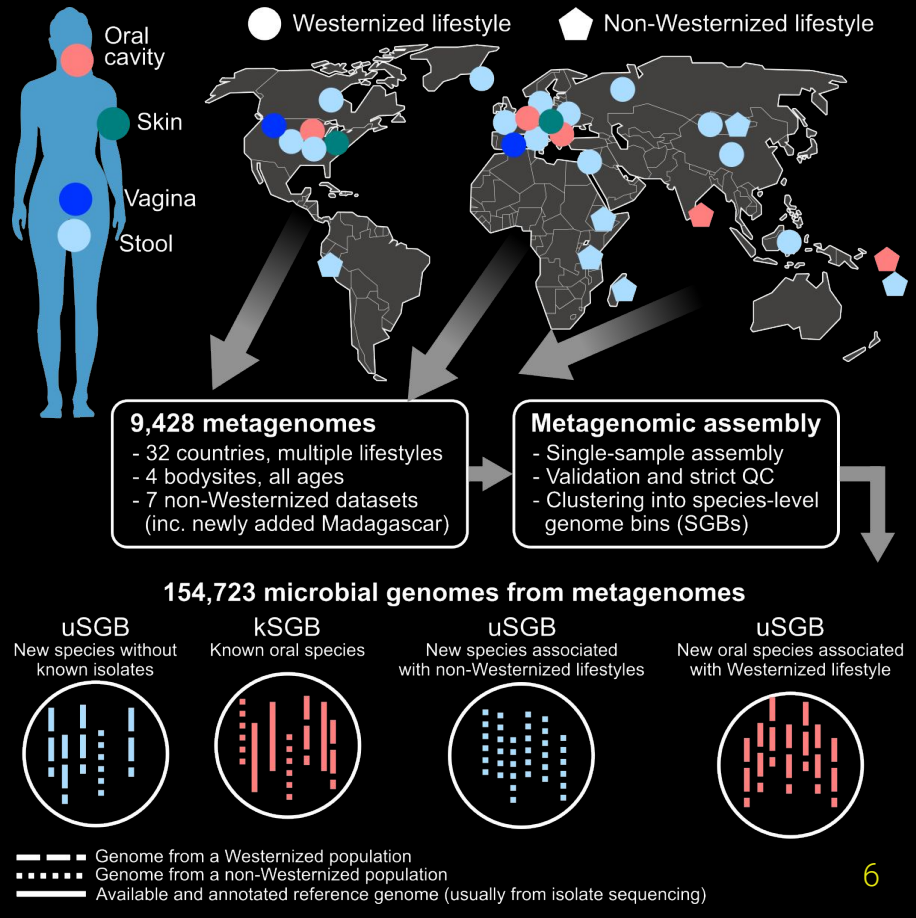


--fast
--diversity high



INCREASING DIVERSITY / SAMPLE SIZE

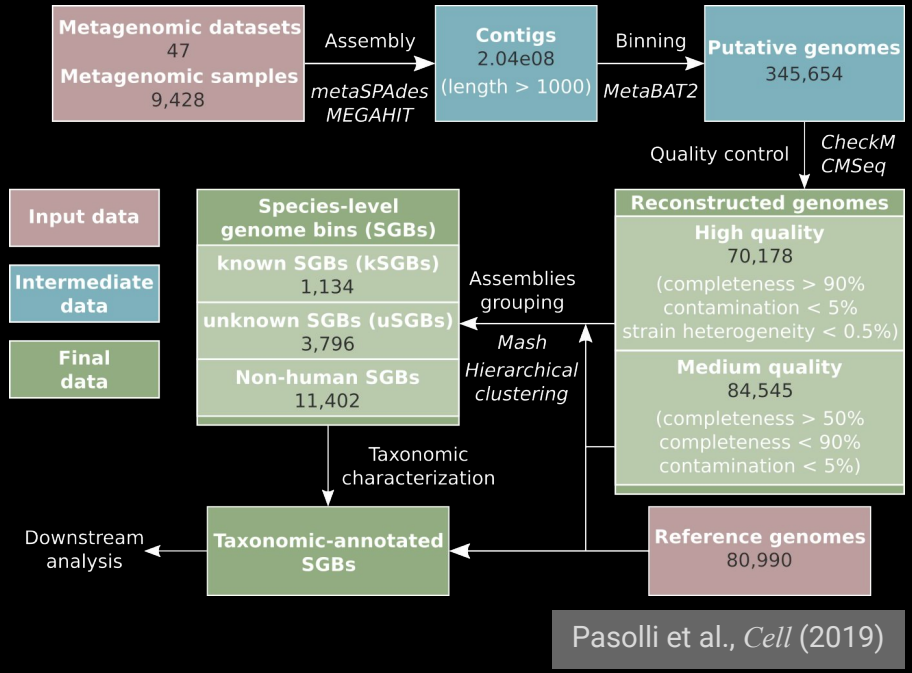
Large-scale metagenomic assembly



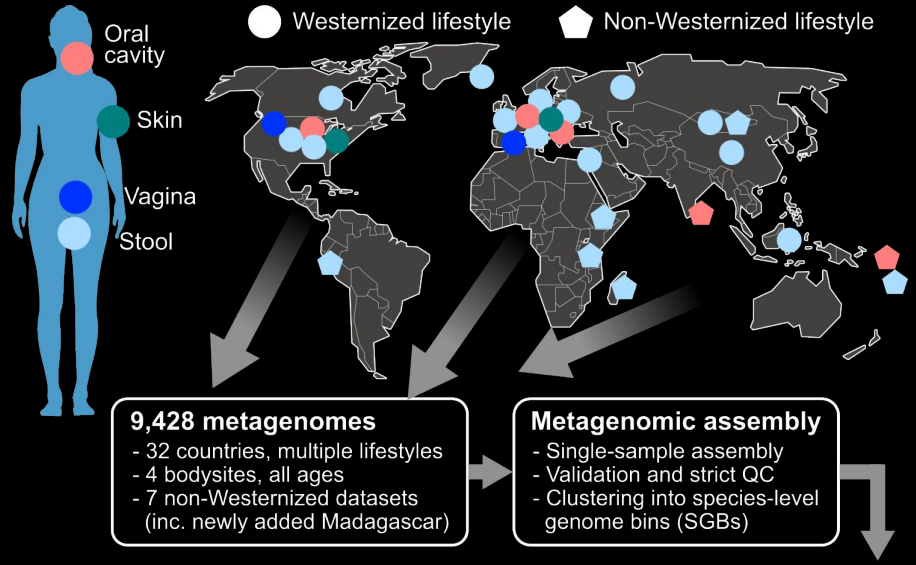
Pasolli et al., *Cell* (2019)

Large-scale metagenomic assembly

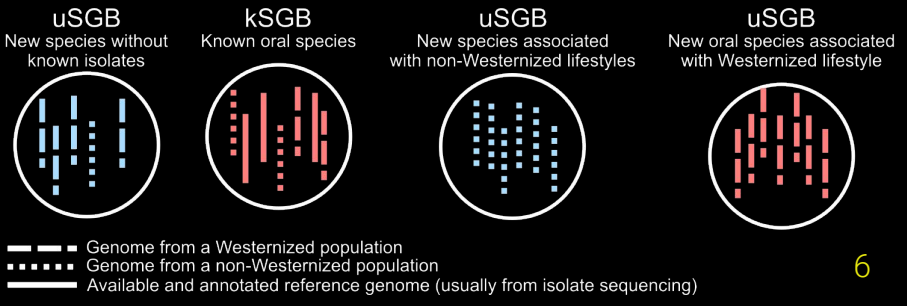
- Single-sample assembly-based
- Strict QC on reconstructed genomes
- Implemented strain-heterogeneity filtering [new]
- Species-level genome bin (SGB) clustering [new]



Pasolli et al., *Cell* (2019)

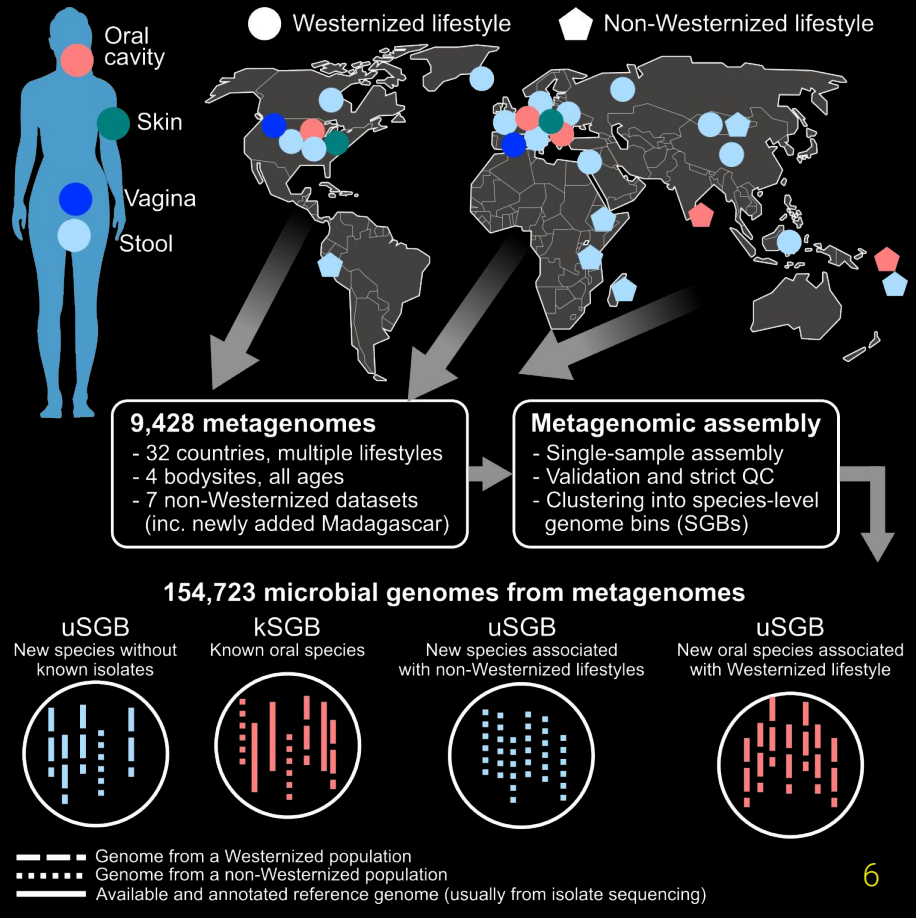


154,723 microbial genomes from metagenomes



Large-scale metagenomic assembly

9,428 metagenomes



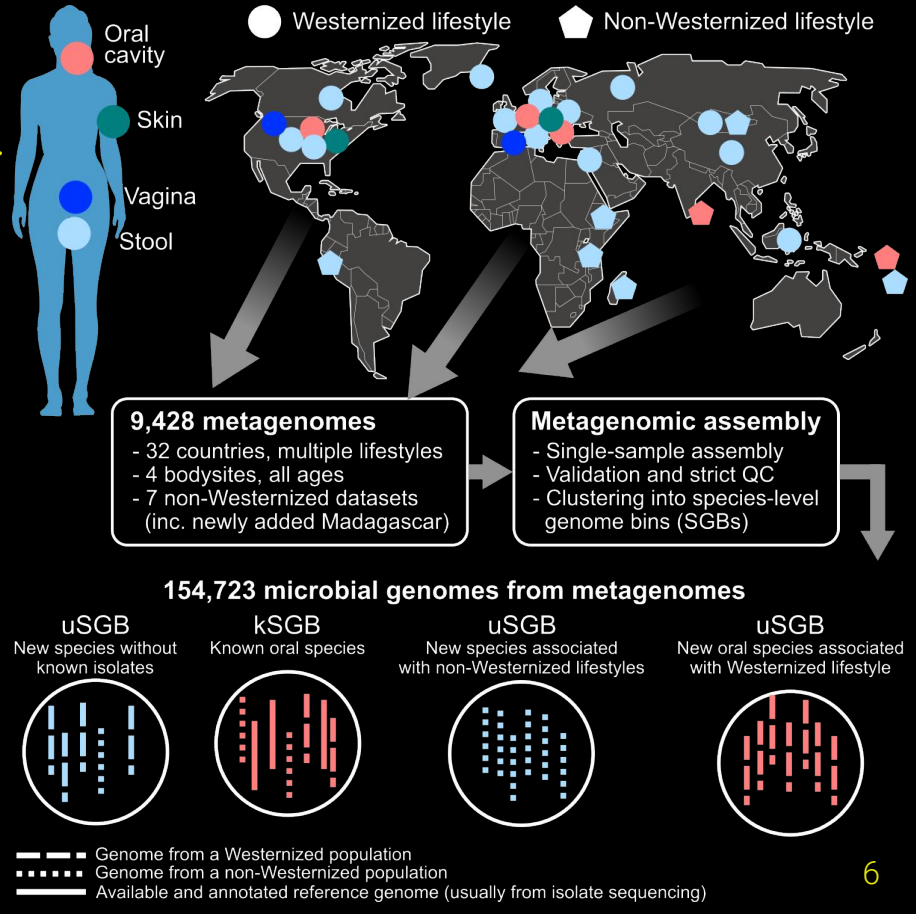
Pasolli et al., *Cell* (2019)

Large-scale metagenomic assembly

9,428 metagenomes

single-sample assembly

154,723 reconstructed genomes



Pasolli et al., *Cell* (2019)

Large-scale metagenomic assembly

9,428 metagenomes

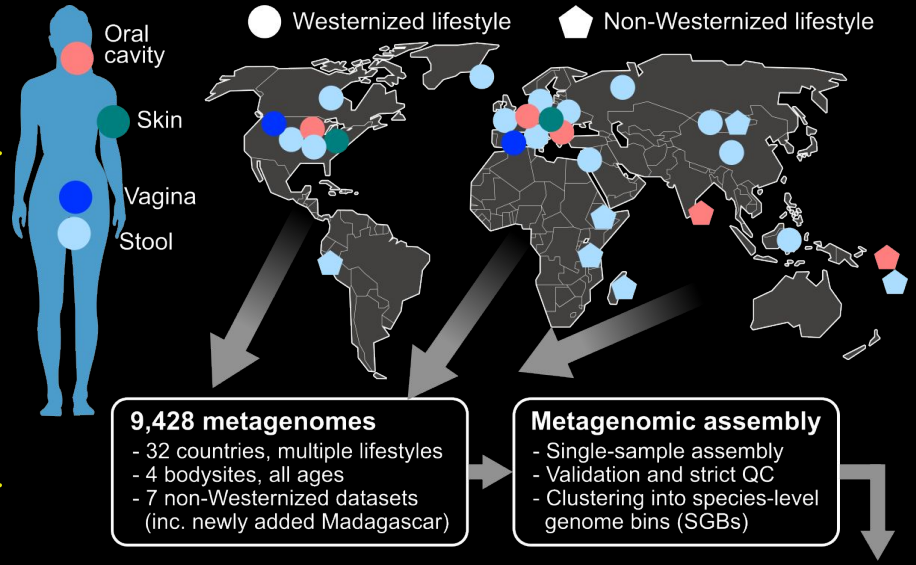
single-sample assembly

154,723 reconstructed genomes

Species-level Genome Bins clustering

4,930 SGBs
(77% without a reference genome)

Pasolli et al., *Cell* (2019)



PhyloPhlAn for microbial tree-of-life phylogenies

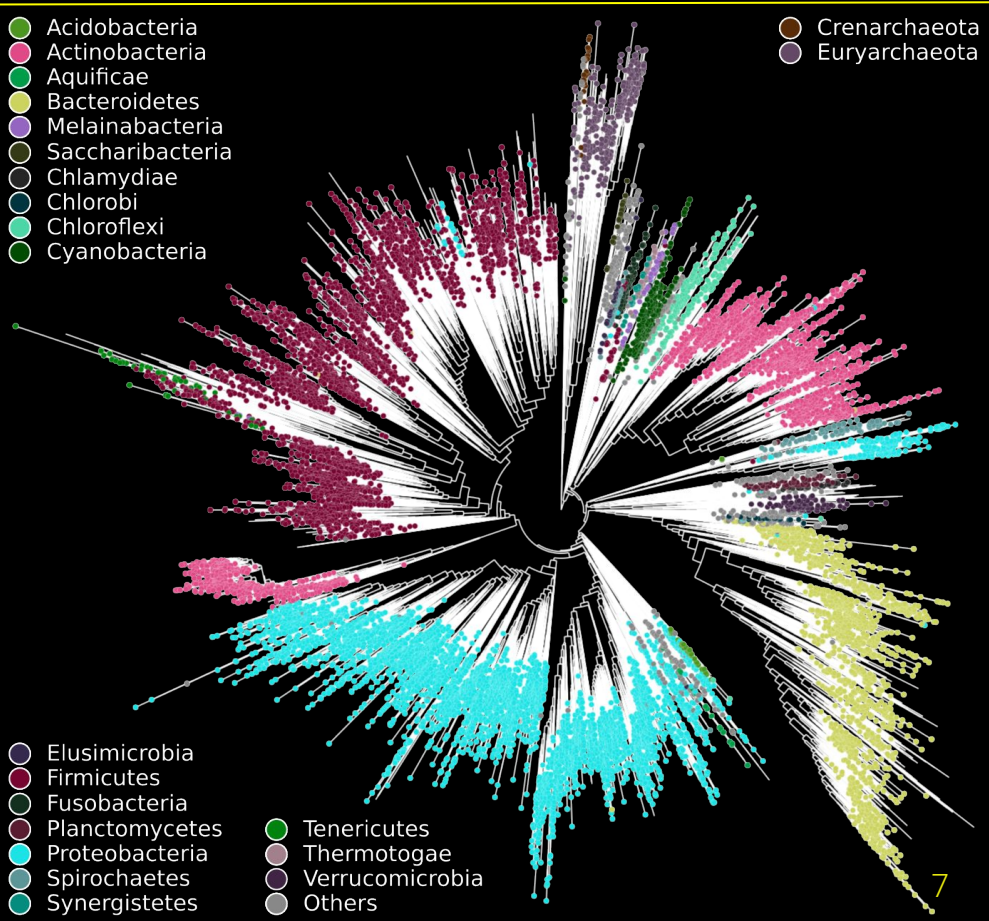
Microbial tree of life of 17,672 SGBs

Including all isolate genomes in NCBI, all MAGs from Parks (2017) and Pasolli (2019)

Phylogeny based on the 400 PhyloPhlAn marker genes validated in Qiyun (2019) with 4,522 AA positions

Building took 10 days and 15 hours using 100 CPUs

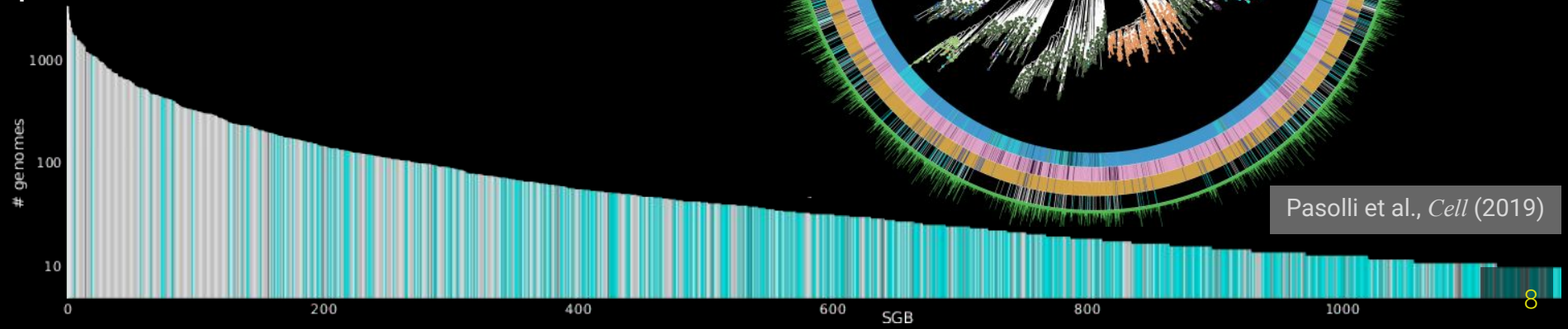
Tree-of-life size phylogenies



Diversity and prevalence of human-associated SGBs

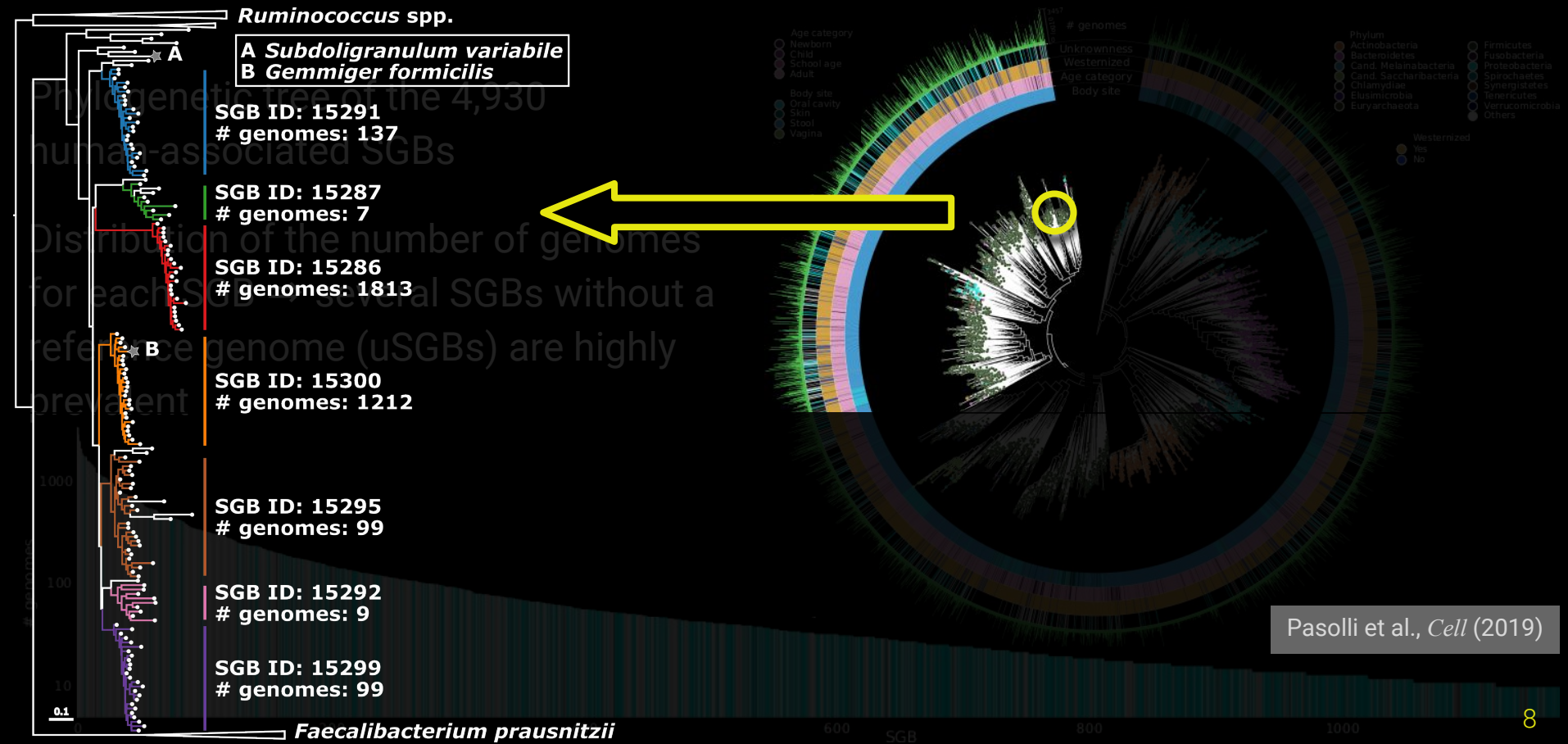
Phylogeny tree of the 4,930 human-associated SGBs

Distribution of the number of genomes for each SGB ⇒ several SGBs without a reference genome (uSGBs) are highly prevalent



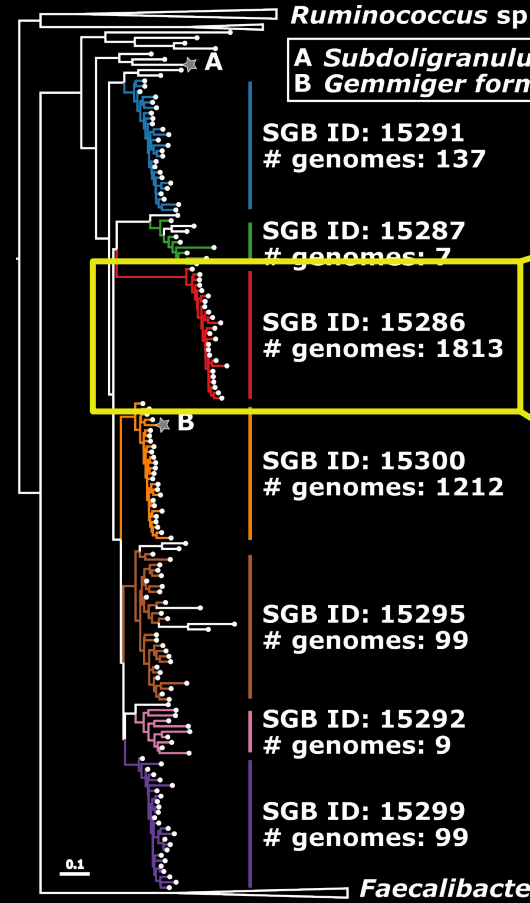
Pasolli et al., *Cell* (2019)

Characterization of lower taxonomic clades



Pasolli et al., *Cell* (2019)

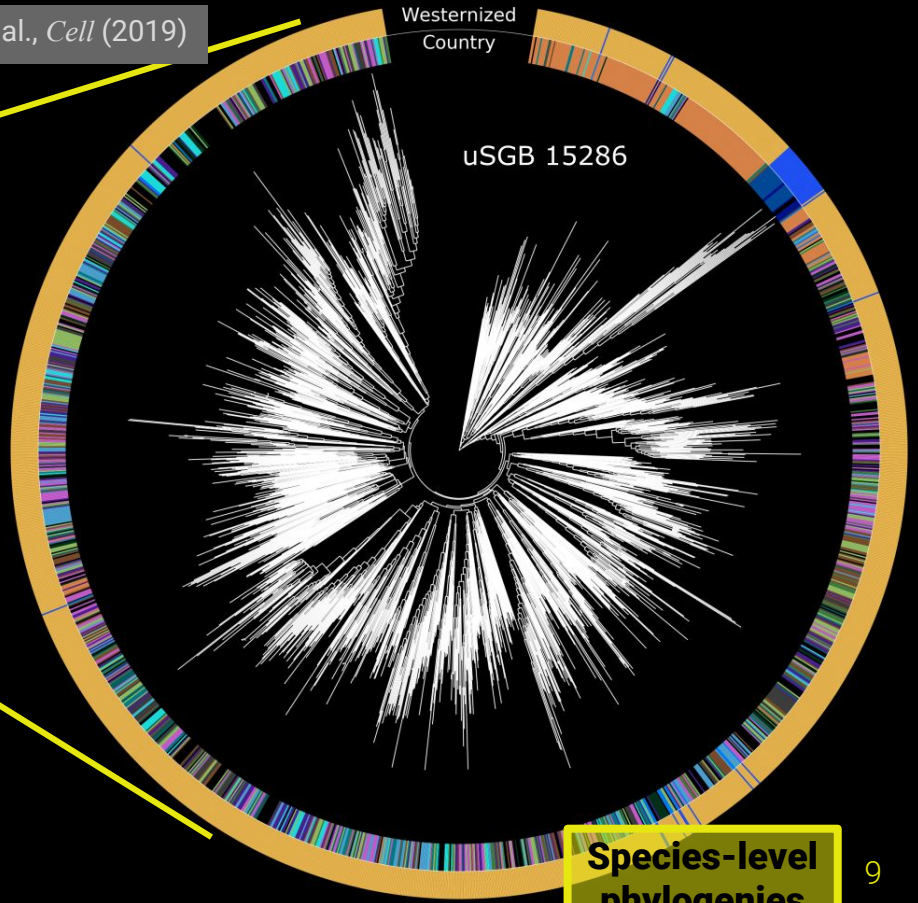
uSBG 15286: "*Candidatus* Cibiobacter qucibialis"



Most prevalent uSGB with 1,813 reconstructed genomes

Subtrees associated with geography and non-Westernized populations

Pasolli et al., *Cell* (2019)



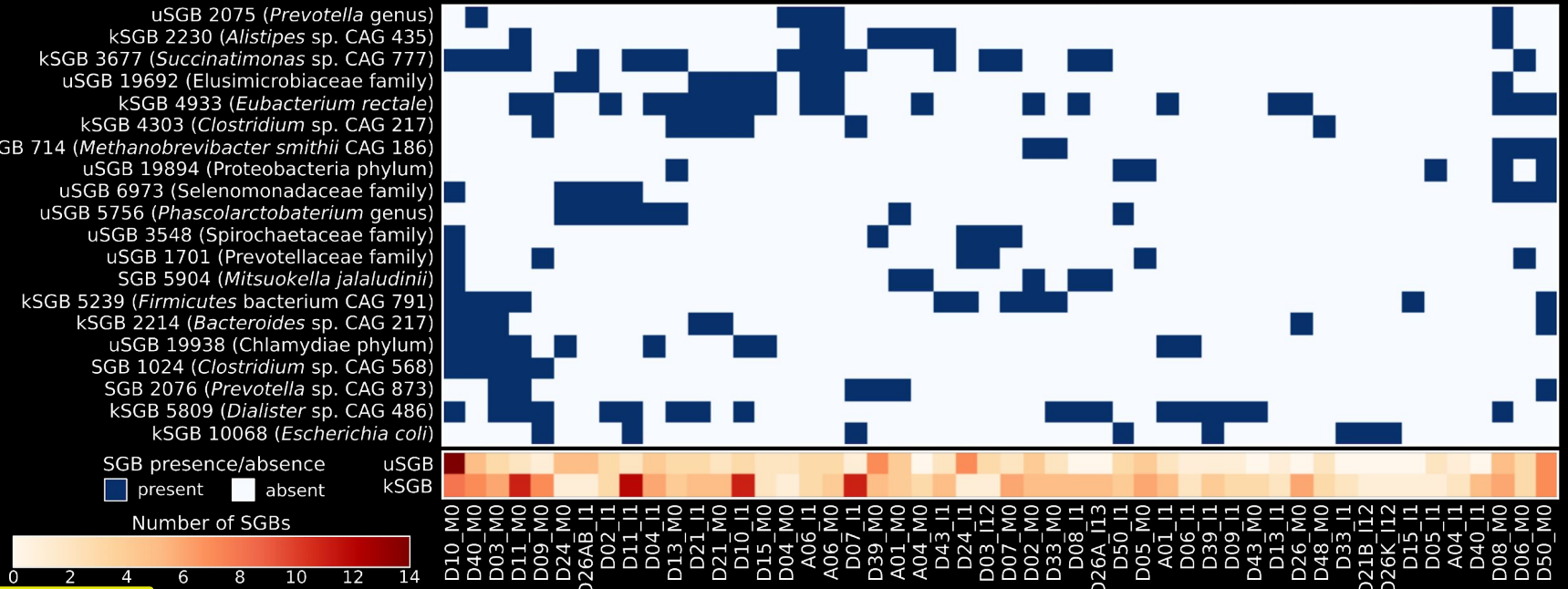
Species-level phylogenies

Automatic SGBs assignment to new MAGs

New Ethiopian non-Westernized dataset of 50 metagenomes

1. Which SGBs are present in this population?
2. Which new SGBs can be found in this population?

in collaboration with MC Collado



Metagenomic application

Automatic SGBs assignment to new MAGs

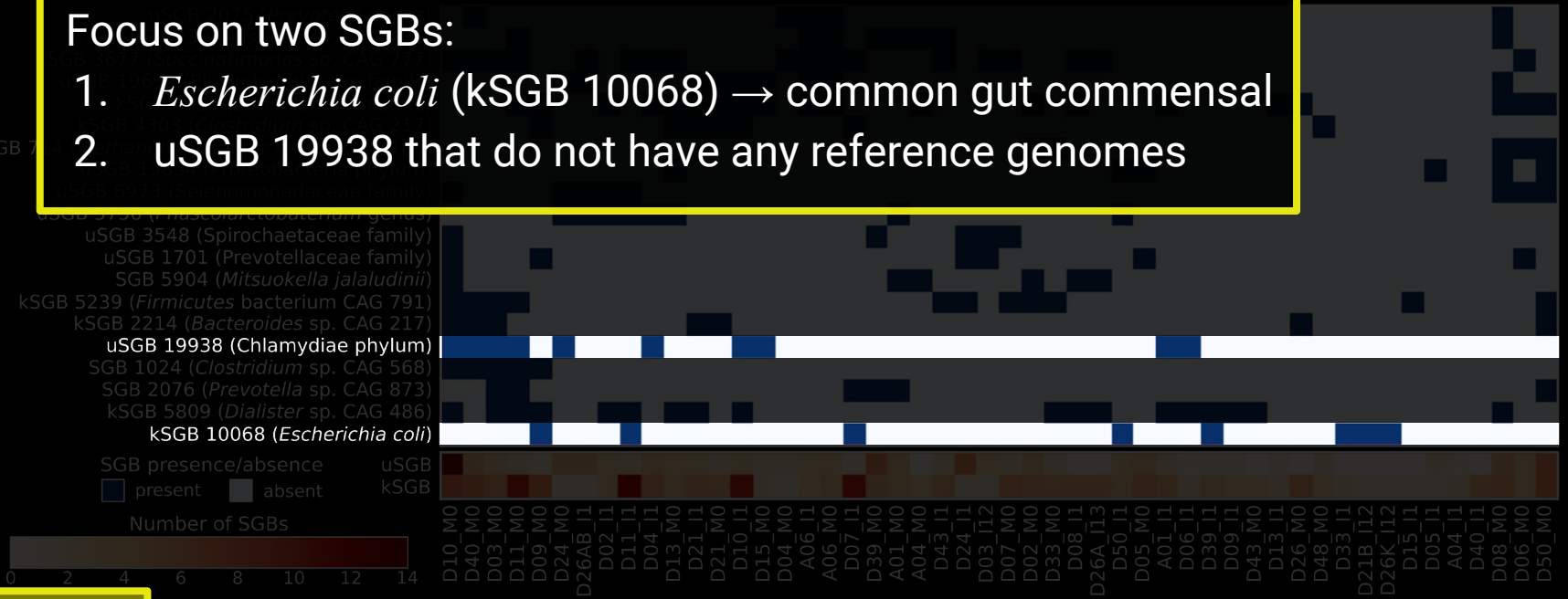
New Ethiopian non-Westernized dataset of 50 metagenomes

1. Which SGBs are present in this population?
2. Which new SGBs can be found in this population?

in collaboration with MC Collado

Focus on two SGBs:

1. *Escherichia coli* (kSGB 10068) → common gut commensal
2. uSGB 19938 that do not have any reference genomes



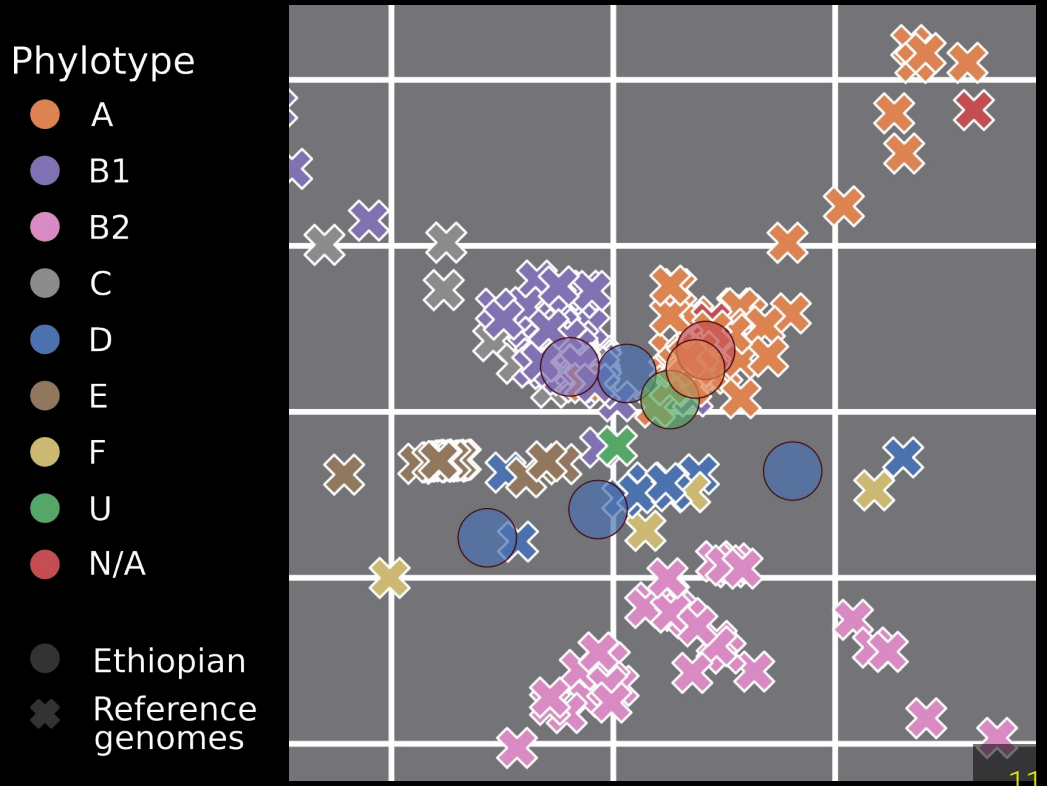
Metagenomic application

kSGB 10068: *E. coli*

Automatic download of *E. coli* core set of UniRef90 proteins

Automatic download of 1,000 *E. coli* reference genomes

Ordination of the phylogenetic distances of the *E. coli* phylogeny comprising the **8 Ethiopian MAGs**



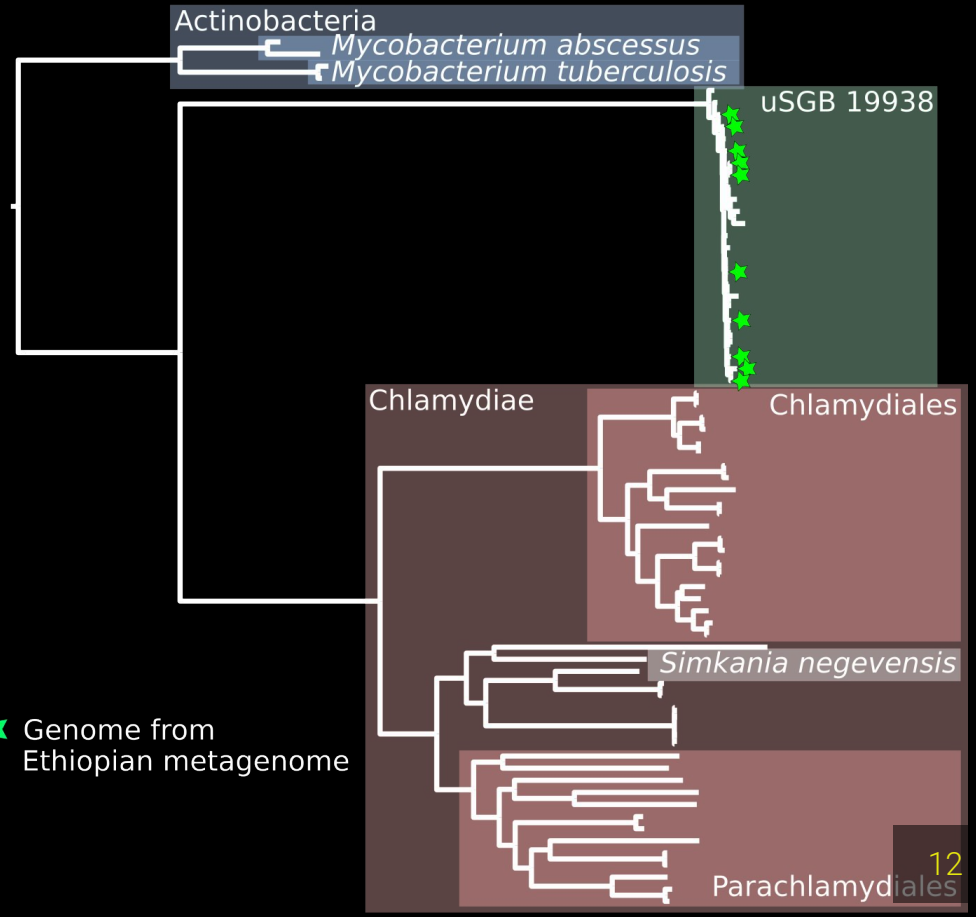
Species-level phylogenies

uSGB 19938: Chlamydiae (?)

The lower taxonomic assignment for uSGB 19938 is the Chlamydiae phylum

Automatic download of reference genomes for the Chlamydiae phylum and two *Mycobacterium* species (rooting)

Reconstruction of a phylum-level diversity phylogeny



Tree-of-life size phylogenies

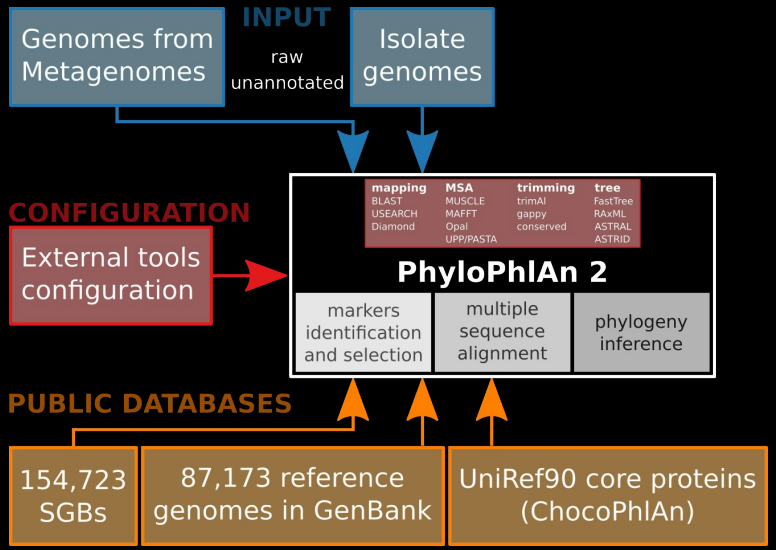
PhyloPhlAn (version 2)

An integrated framework for phylogenetic analysis

- Reference genomes from isolates
- MAGs from metagenomes
- Clade-specific phylogenetic markers
- Retrieval of additional genomes & MAGs
- Taxonomic assignment of MAGs

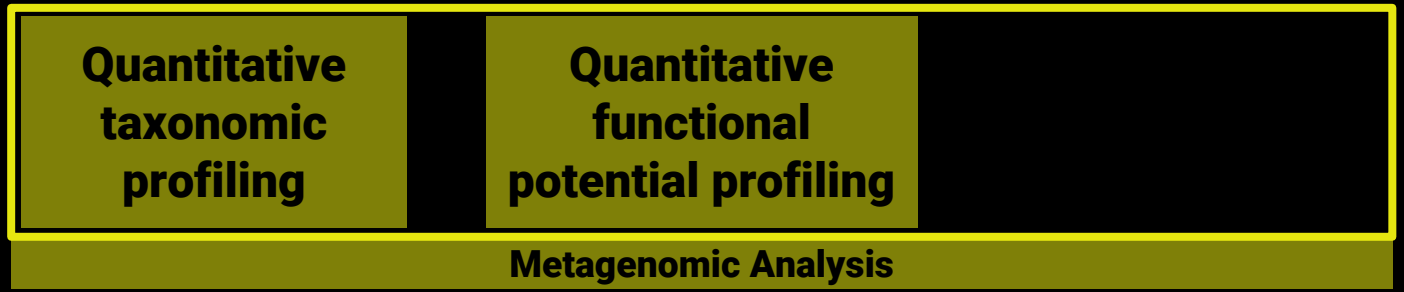
Main features:

- Scalable, flexible, automatic, modular, customizable

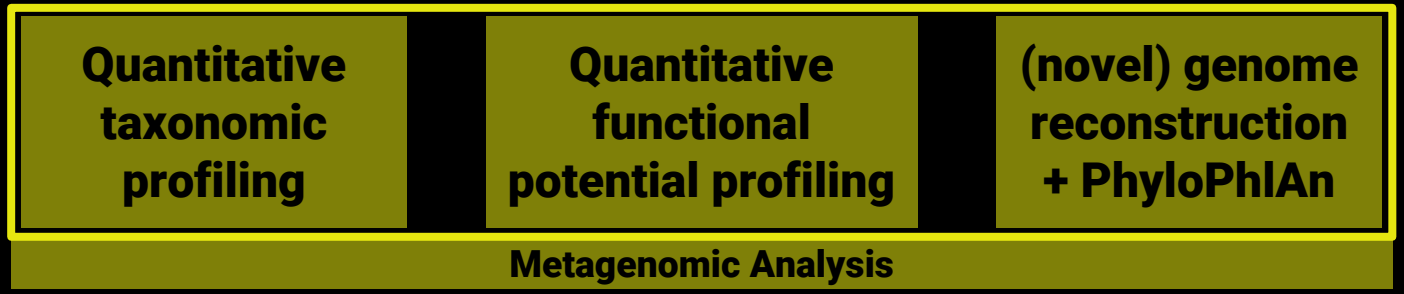


- Available open-source in Bitbucket: https://bitbucket.org/nsegata/phylophlan
- Will be soon available in BIOCONDA
- Working on tutorials to describe the new functionalities by example
- The software can be used and any feedback is appreciated!

Integrated approach for metagenomic analysis



Integrated approach for metagenomic analysis



Characterize new MAGs within the SGBs

Phylogenetic investigations including available reference genomes

New data to update the SGB resource to improve future analyses

Integrated approach for metagenomic analysis

Quantitative taxonomic profiling

Quantitative functional potential profiling

(novel) genome reconstruction + PhyloPhlAn

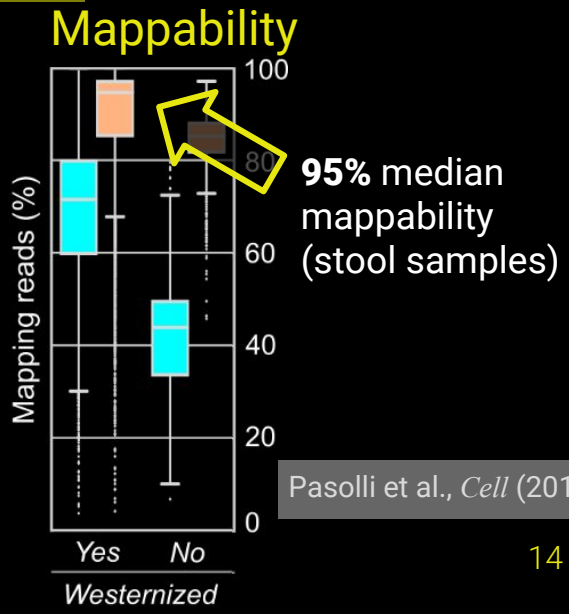
Metagenomic Analysis

Feedback to improve taxonomic and functional analyis

Characterize new MAGs within the SGBs

Phylogenetic investigations including available reference genomes

New data to update the SGB resource to improve future analyses



Pasolli et al., *Cell* (2019)

Thanks!

The Laboratory of Computational Metagenomics

Nicola Segata (PI)
Adrian Tett
Federica Pinto
Fabio Cumbo
Andrew Thomas
Giulia Masetti
Federica Armanini
Francesco Asnicar
Serena Manara
Paolo Ghensi
Moreno Zolfo
Francesco Beghini
Kun D. Huang
Nicolai Karcher
Paolo Manghi

http://segatalab.cibio.unitn.it - nicola.segata@unitn.it

Thanks to:

Curtis Huttenhower
Rob Knight
Siavash Mirarab
Qiyun Zhu
Maria C. Collado

Interested?
We are recruiting!
nicola.segata@unitn.it

ITI International Team for Implantology

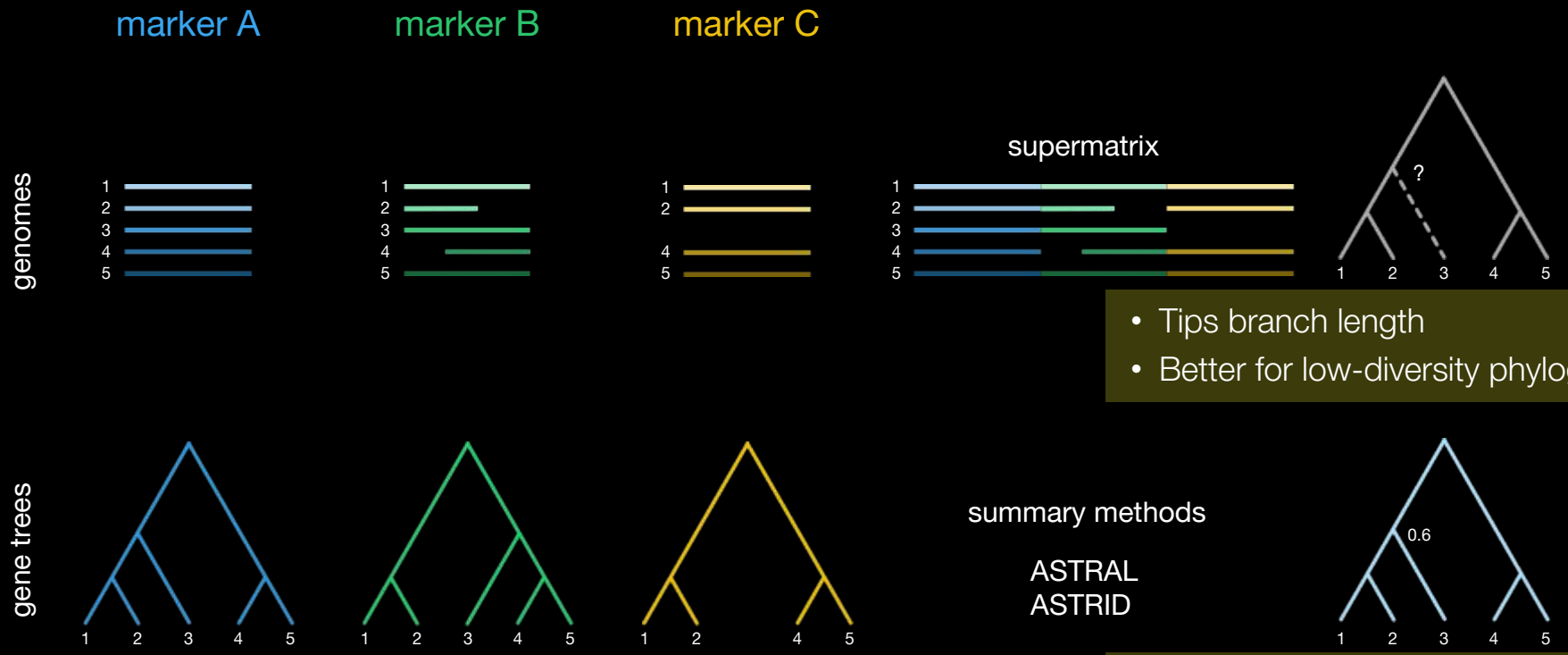
SIdP Società Italiana di Parodontologia e Implantologia

FONDAZIONE CASSA DI RISPARMIO DI TRENTO E ROVERETO

TERME DI COMANO

LILT SEZIONE PROVINCIALE DI TRENTO
LEGA ITALIANA PER LA LOTTA CONTRO I TUMORI
prevenire è vivere

Supermatrix vs. Supertree



- Tips branch length
- Better for low-diversity phylogeny

- No tips branch length
- Better for large phylogenies

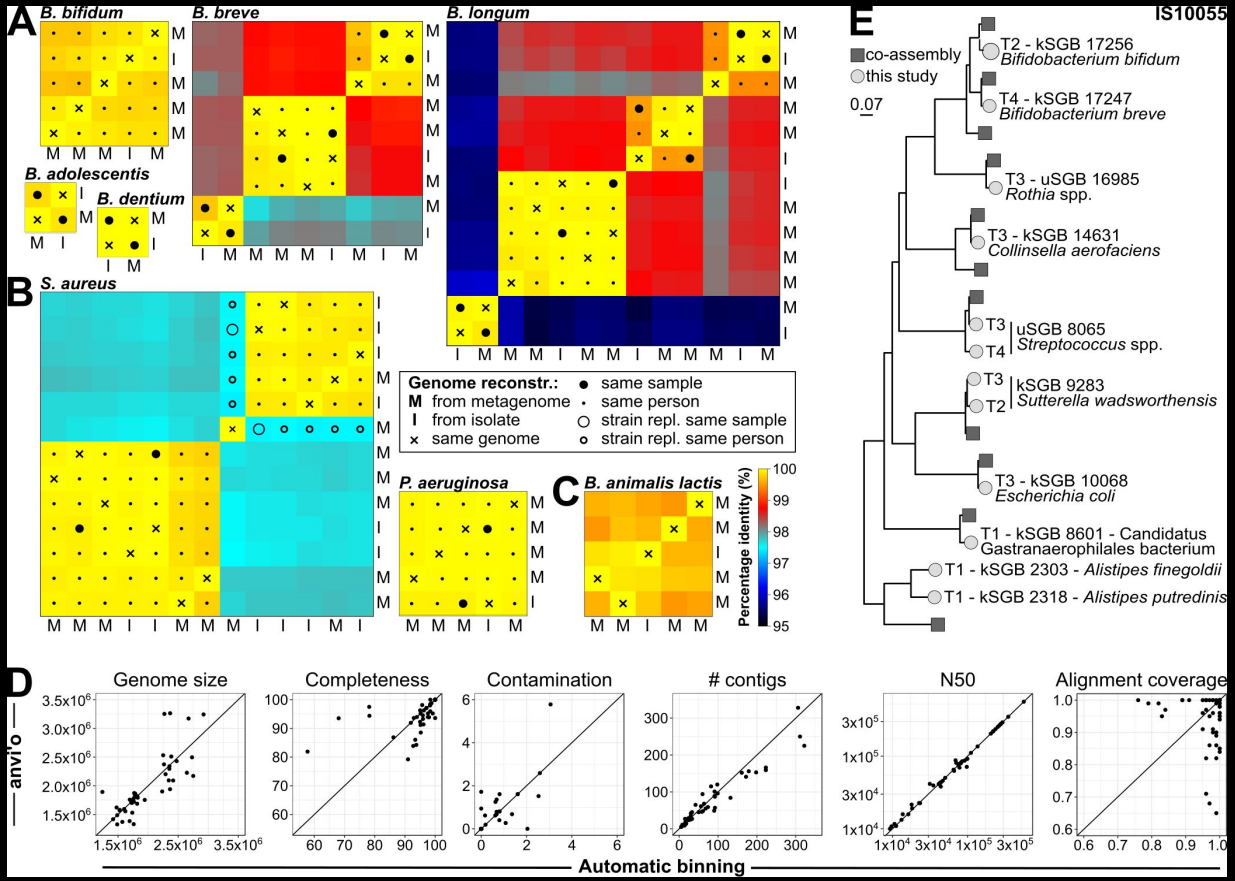
Stamatakis, Bioinformatics (2014)

Rokas et al., Nature (2003)

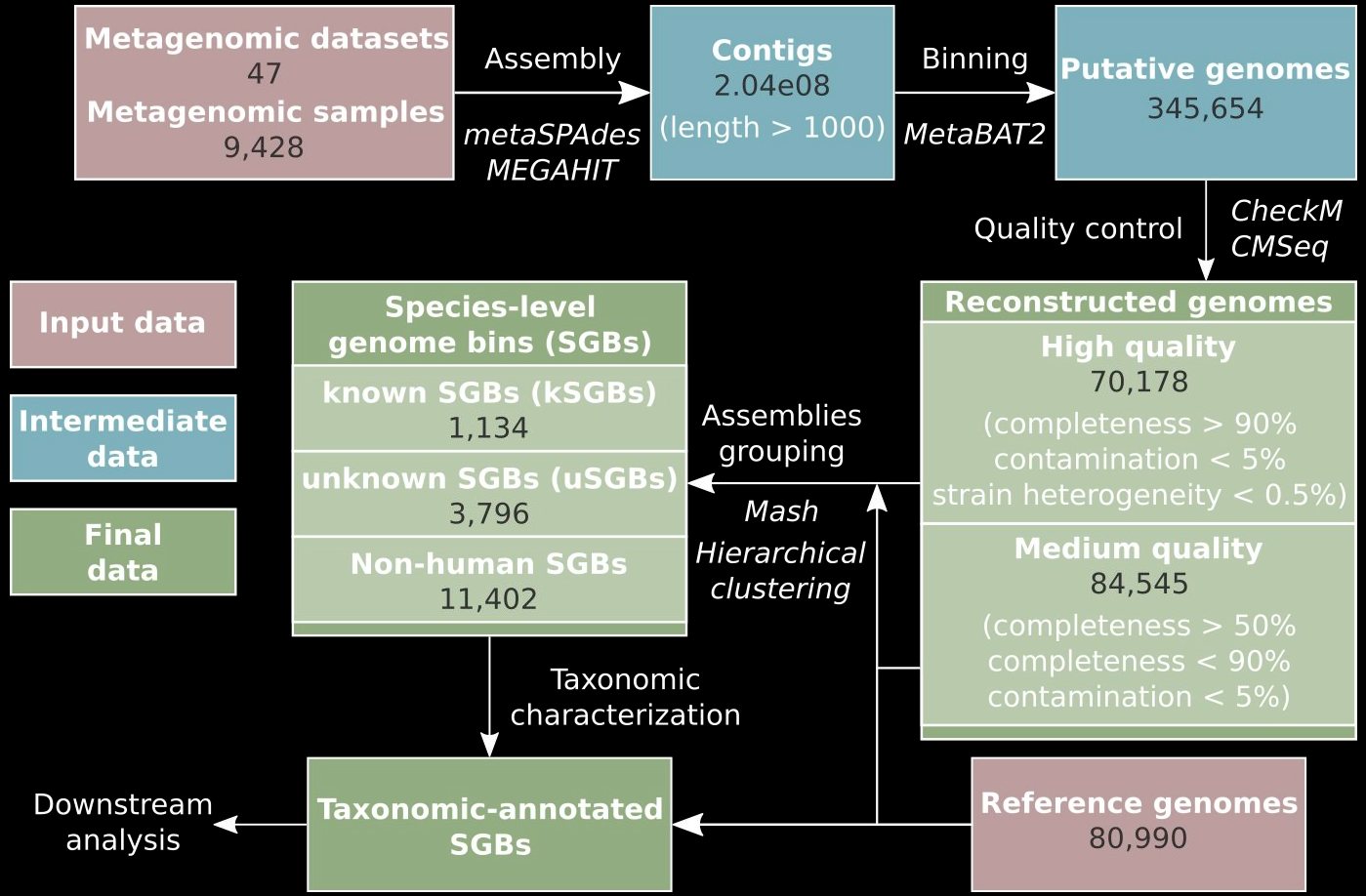
Siavash et al., Bioinformatics (2014)

Vachaspati and Warnow, BMC Genomics

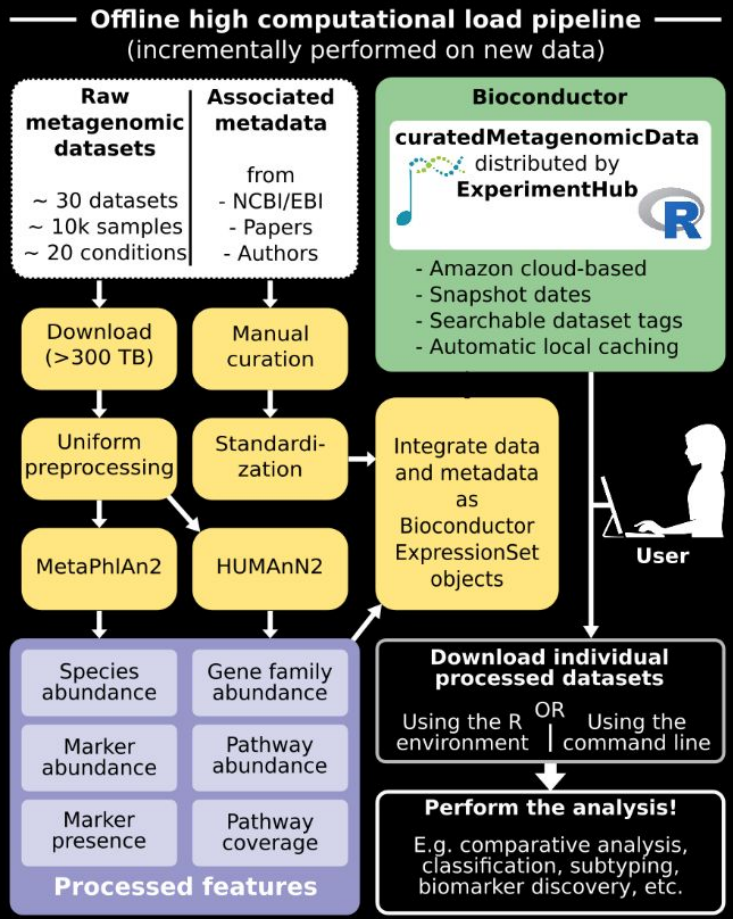
Automatic and manual single-sample assembly and co-assembly evaluation



Large-scale single-sample assembly approach



CuratedMetagenomicData



Pasolli et al., *Nat Methods* (2017)

- Mandatory metadata fields:

```
sampleID DNA_extraction_kit
subjectID
sequencing_platform
body_site                 number_reads
country                   number_bases
antibiotics_current_use
minimum_read_length
study_condition           median_read_length
disease                   NCBI_accession
age_category              PMID
gender                    non_westernized
```

- Optional metadata fields are... All the available ones!

- Currently >10,000 metagenome samples
- New available datasets are continuosly included https://waldronlab.github.io/curatedMetagenomicData