Large-scale strain-level population genomics from metagenomics

Nicola Segata

Laboratory of Computational Metagenomics

Centre for Integrative Biology
University of Trento
Italy

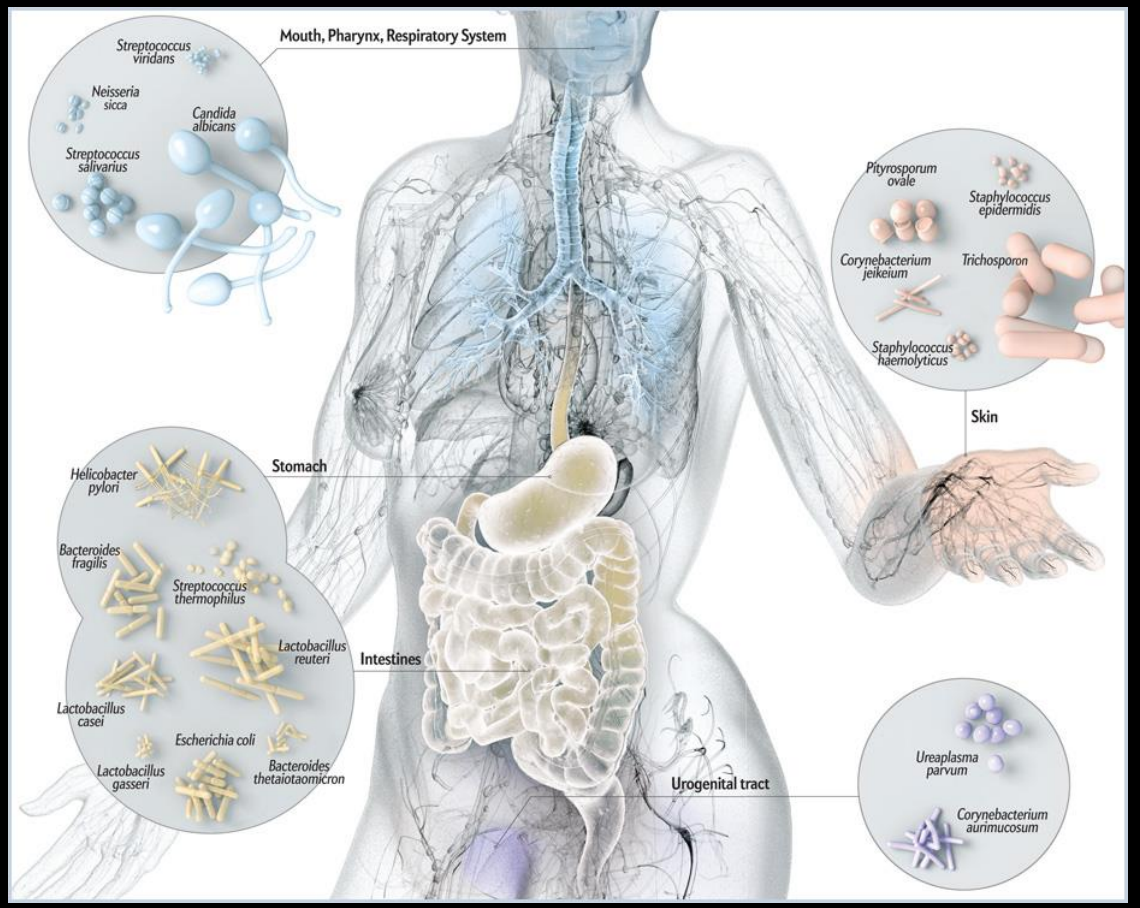
BARI

June 20th , 2017

UNIVERSITÀ DEGLI STUDI
DI TRENTO

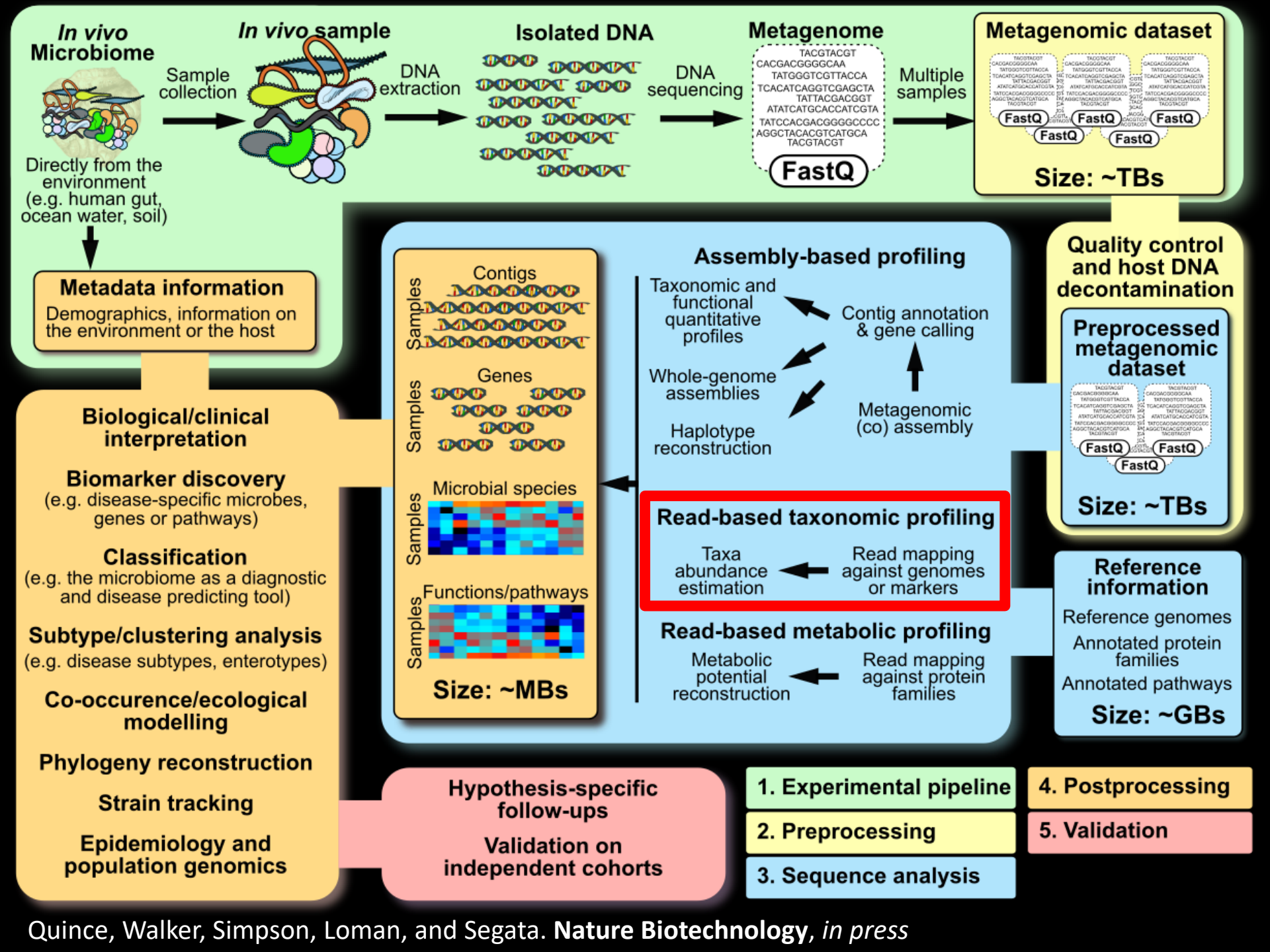
CIBIO

The human microbiome



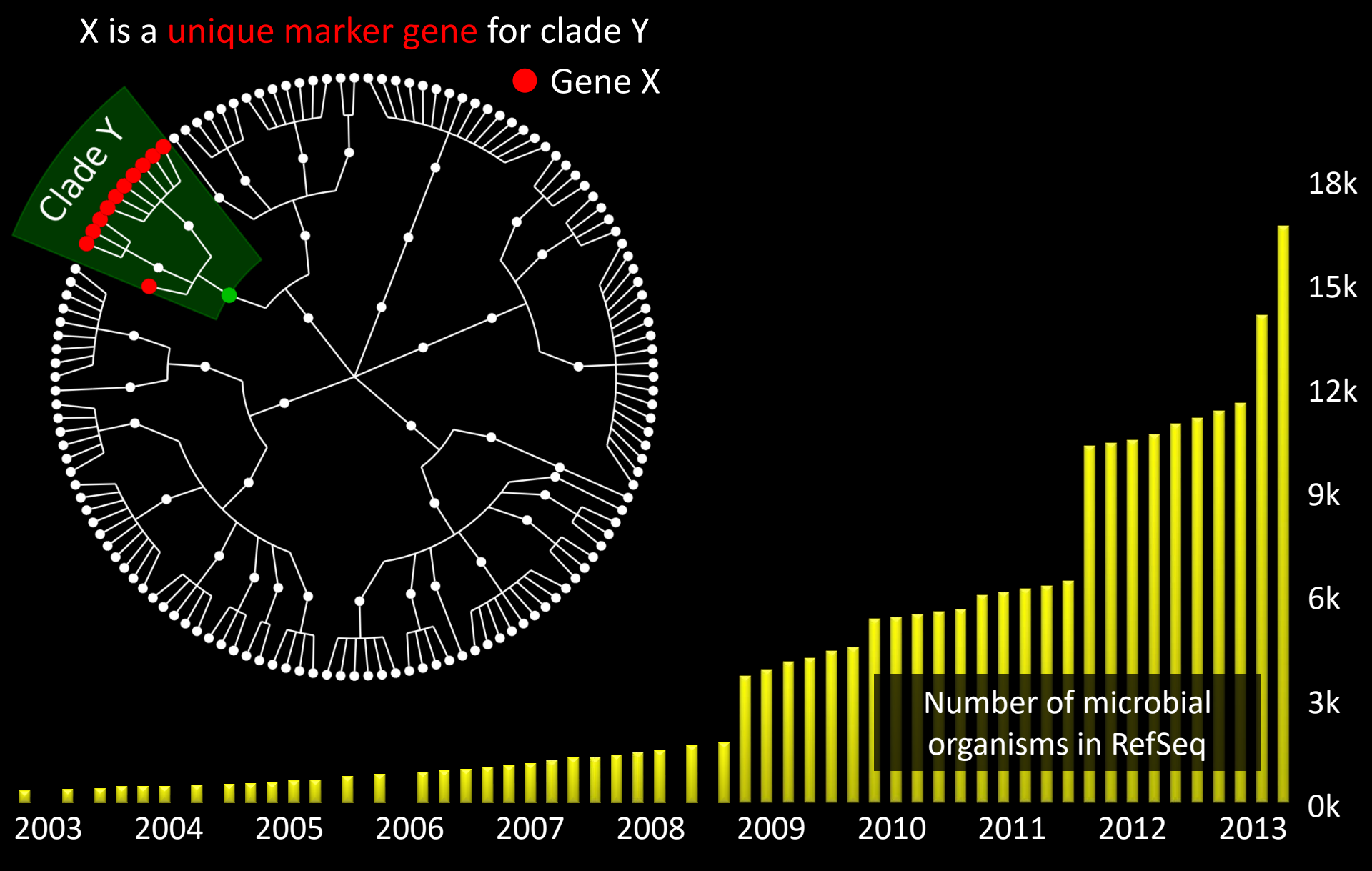
- 1-10x more microbial than human cells
- 1M times as many microbes inside each of us than humans on earth
- 100x more microbial than human genes

Not only bacteria!
Not only the gut microbiome!



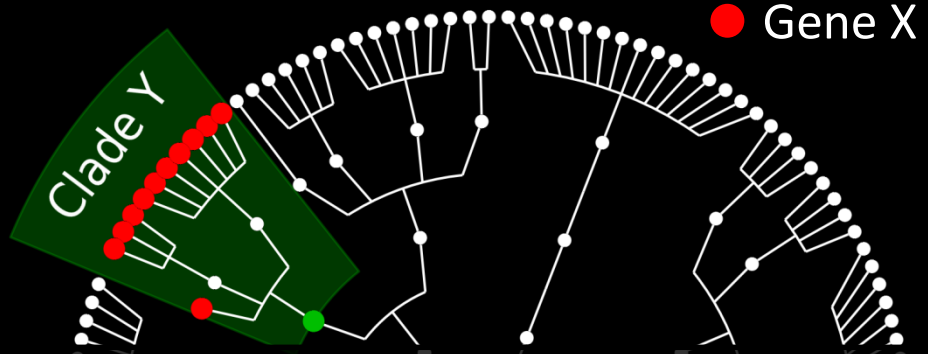
Quince, Walker, Simpson, Loman, and Segata. **Nature Biotechnology**, *in press*

MetaPhlAn2: trans-kingdom profiling using marker genes



MetaPhlAn2: trans-kingdom profiling using marker genes

X is a unique marker gene for clade Y



THE INPUT

- 2,887 genomes (107 Archaea)
- ~half of genomes are final, ~half draft
- 1,222 species, 652 genera, 278 families, 130 orders, 66 classed, 33 phyla, 2,383 total clades

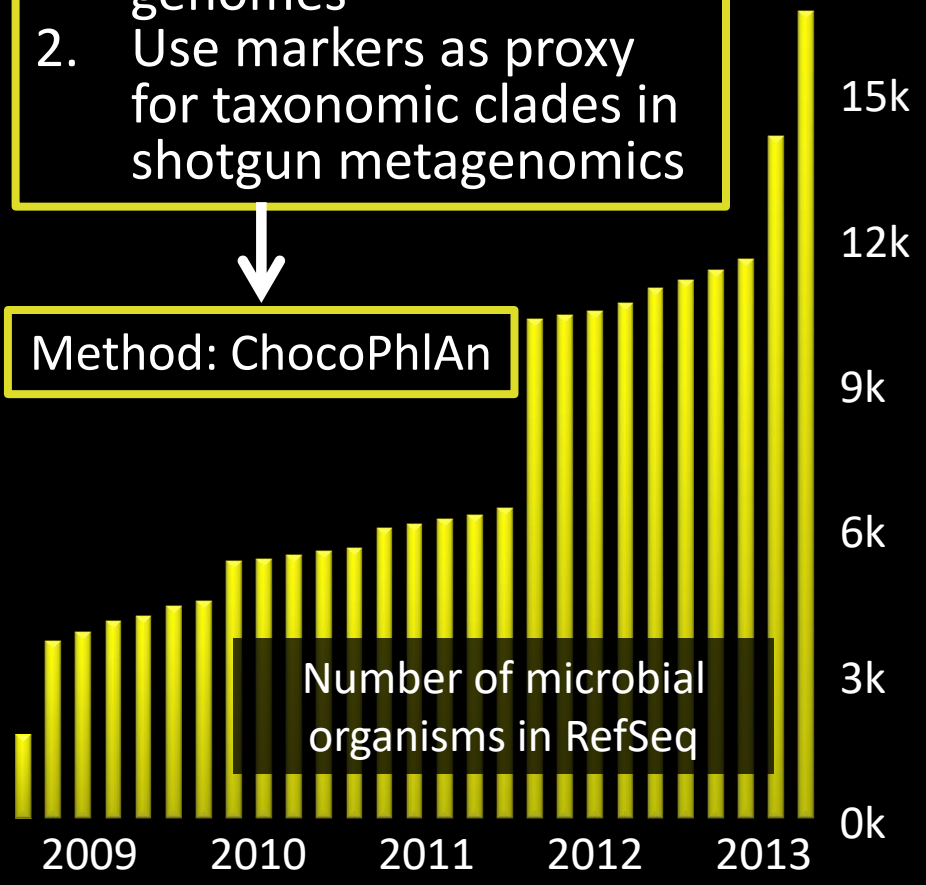
THE RESULTING DATABASE

- ~2M total unique marker genes
- ~400k "most representative" unique marker genes
- 231±107 markers per species (350 fixed max)
- Only 12 species with <15 markers (9 *Brucella* spp.)
- The 400k database represent ~4% of the total microbial sequence data available

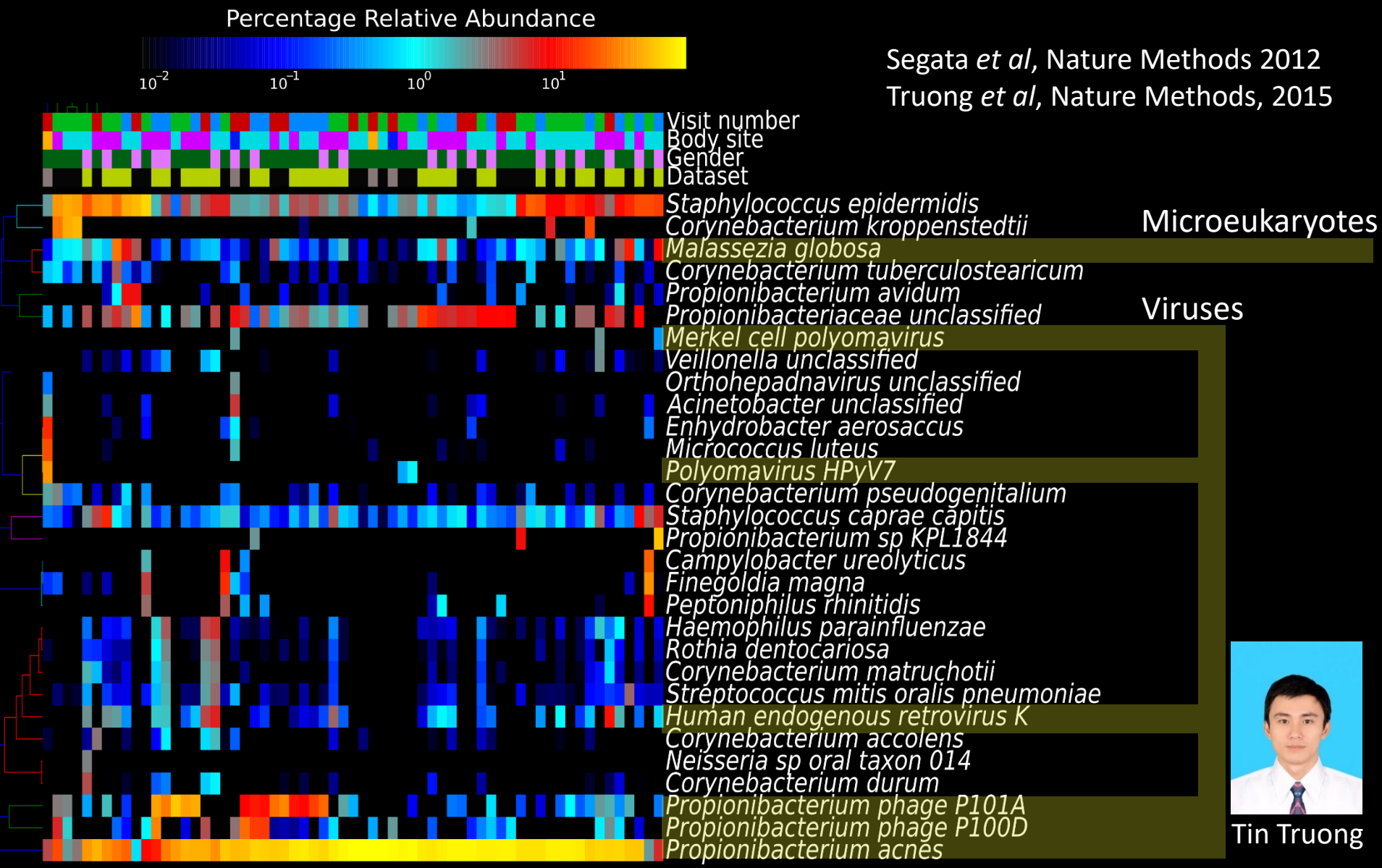
IDEA

1. Pre-identify markers from reference genomes
2. Use markers as proxy for taxonomic clades in shotgun metagenomics

Method: ChocoPhlAn



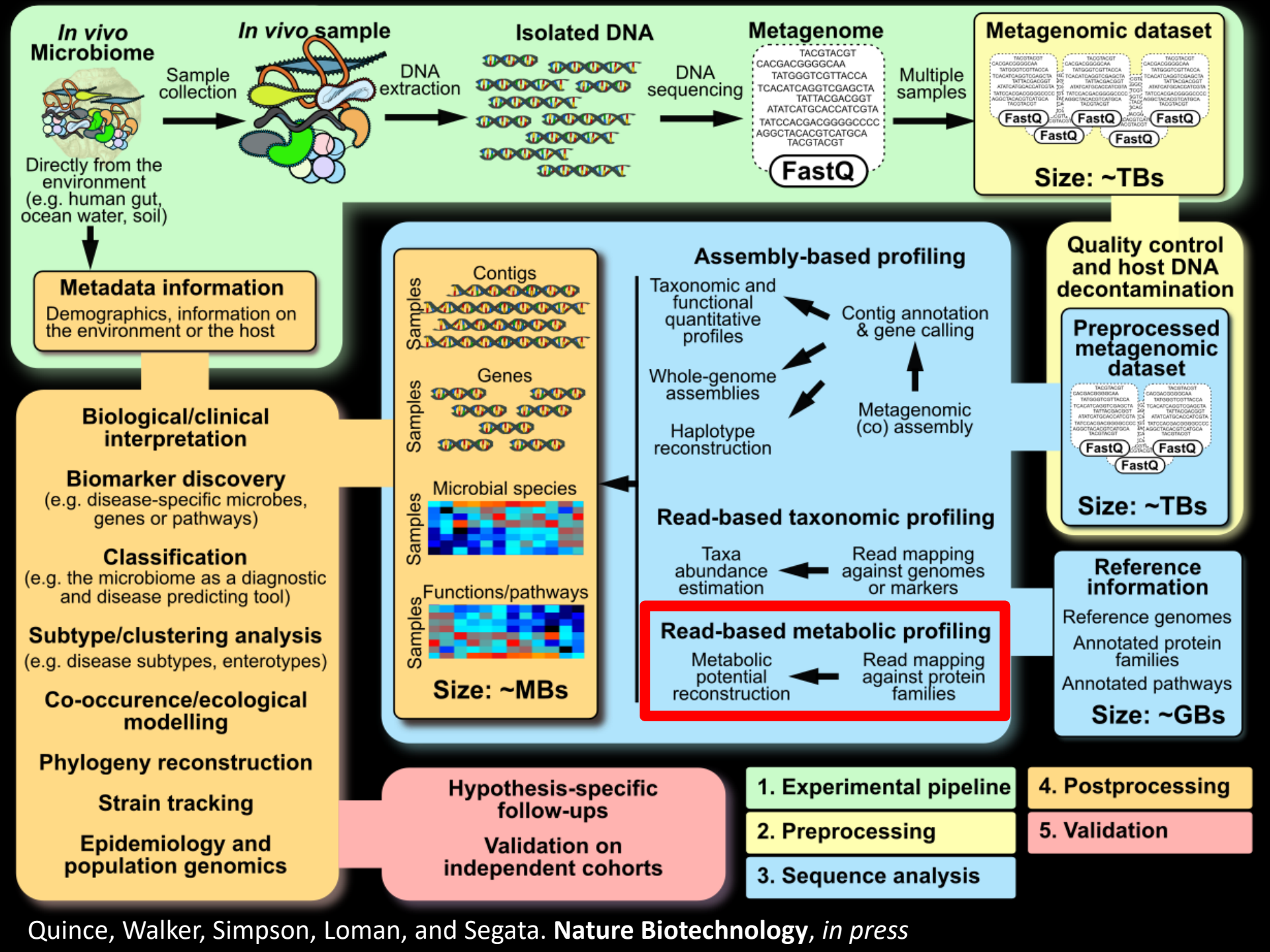
MetaPhlAn2: trans-kingdom profiling using marker genes



Segata *et al*, Nature Methods 2012

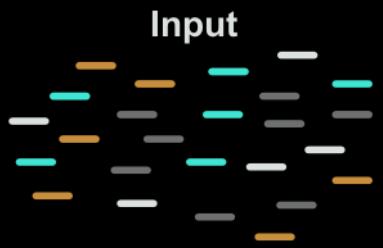
Truong *et al*, Nature Methods, 2015

Tin Truong



Quince, Walker, Simpson, Loman, and Segata. **Nature Biotechnology**, *in press*

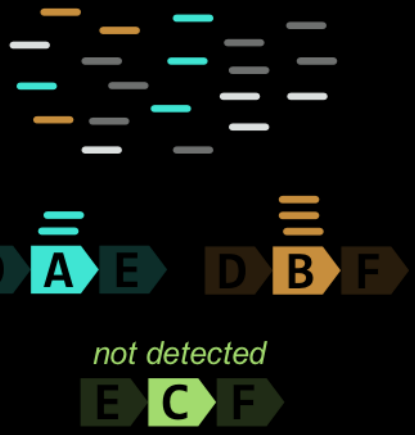
HUMAnN2: species-specific functional profiling of meta'omes



A quality-controlled metagenome or metatranscriptome (DNA/RNA reads)

- Species1
- Species2
- Ambiguous
- Novel

Taxonomic prescreen

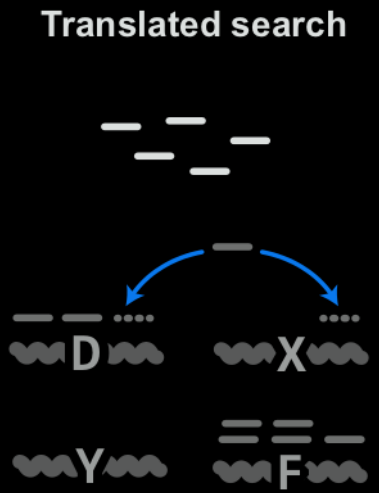


Reads mapped to clade-specific marker genes to rapidly identify community species

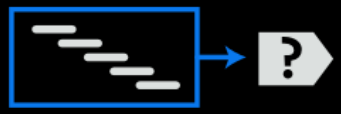
Pangenome mapping

Reads mapped to pangenomes of identified species (*Note Species2 isolate missing GeneF*)

Translated search



Unclassified reads multi-hit mapped to comprehensive protein database



Unmapped reads are isolated for assembly of novel genes (*external to HUMAnN2*)


Curtis Huttenhower


Eric Franzosa


Lauren McIver

Franzosa *et al*, in revision

- **INPUT**: Quality controlled metagenomes and/or metatranscriptomes
- Better if paired!

HUMAnN2: species-specific functional profiling of meta'omes

UniRef gene cluster	Gene name	Total gene abundance (RPK)
UniRef90_R6K3Z5:	IMP dehydrogenase	600.95
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_caccae	234.76
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_dorei	107.38
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_ovatus	92.18
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_stercoris	83.95
UniRef90_R6K3Z5:	IMP dehydrogenase|Bacteroides_vulgatus	57.27
UniRef90_R6K3Z5:	IMP dehydrogenase|unclassified	25.41

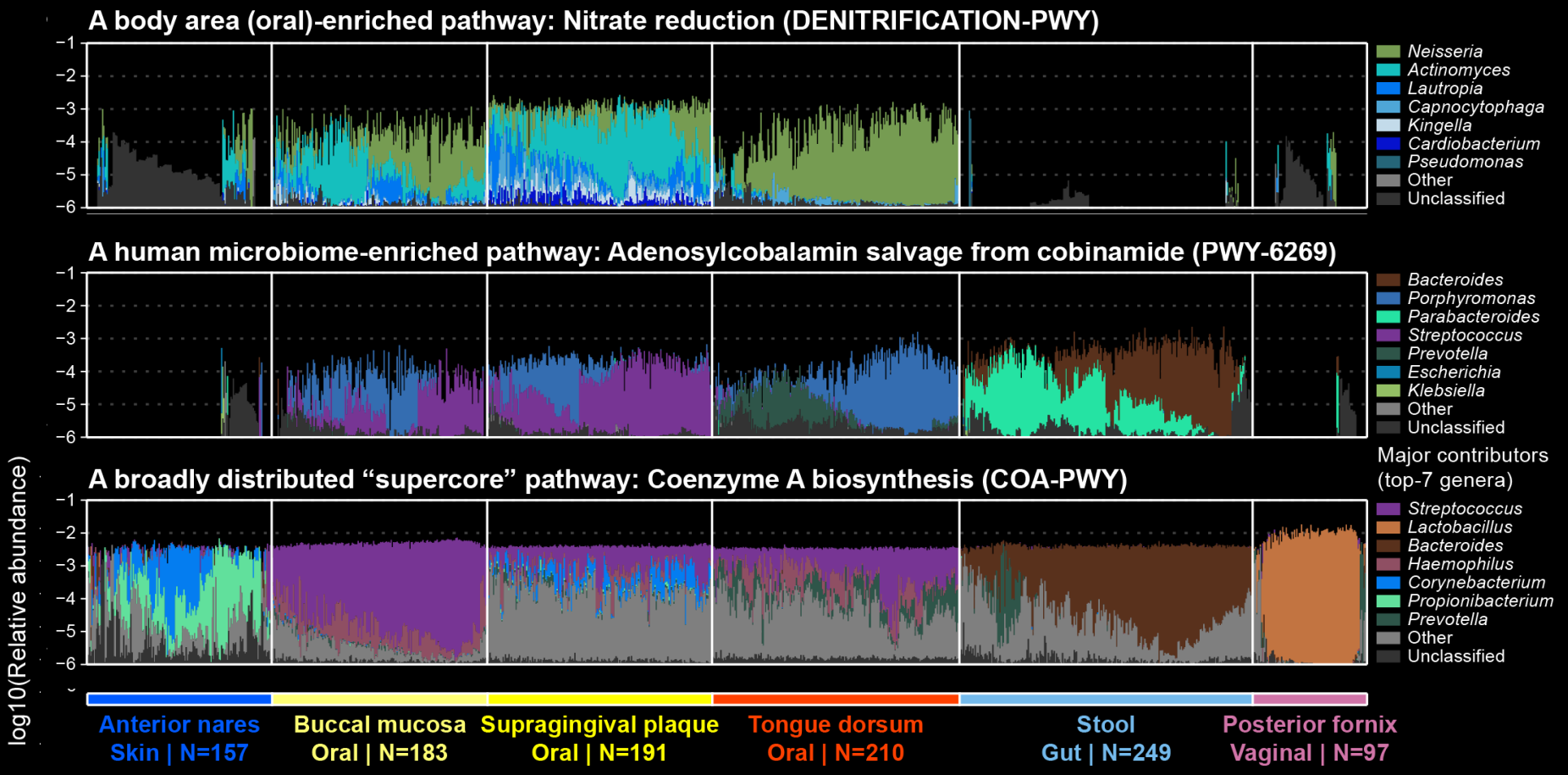
Σ

~HUMAnN1

Per-species & **unclassified** stratifications

MetaCyc pathway	Pathway **abundance** & **coverage**	
PWY-7221: GTP biosynthesis	200.35	1
PWY-7221: GTP biosynthesis|Bacteroides_caccae	120.23	1
PWY-7221: GTP biosynthesis|Bacteroides_dorei	11.12	0

HUMAnN2: species-specific functional profiling of meta'omes



- Core pathways: found in >75% of subjects w/ confident taxonomic attribution
- Housekeeping functions are core to all body sites
- Some core functions are enriched in human-associated microbes
- Other core functions are site-specific, and potentially niche-adaptive

Integration for large scale analysis

The Human Microbiome Project Consortium* doi:10.1038/nature11234

Structure, function and diversity of the healthy human microbiome

Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes

nature biotechnology

Cell Host & Microbe

The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

A human gut microbial gene catalogue established by metagenomic sequencing

doi:10.1038/nature11450

A metagenome-wide association study of gut microbiota in type 2 diabetes

doi:10.1038/nature13568

Alterations of the human gut microbiome in liver cirrhosis

doi:10.1038/nature12198

Gut metagenome in European women with normal, impaired and diabetic glucose control

doi:10.1038/nature12506

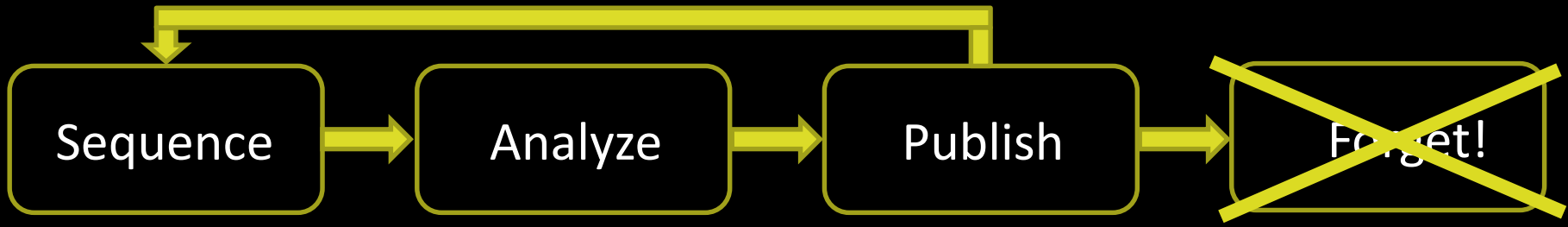
Richness of human gut microbiome correlates with metabolic markers

nature medicine

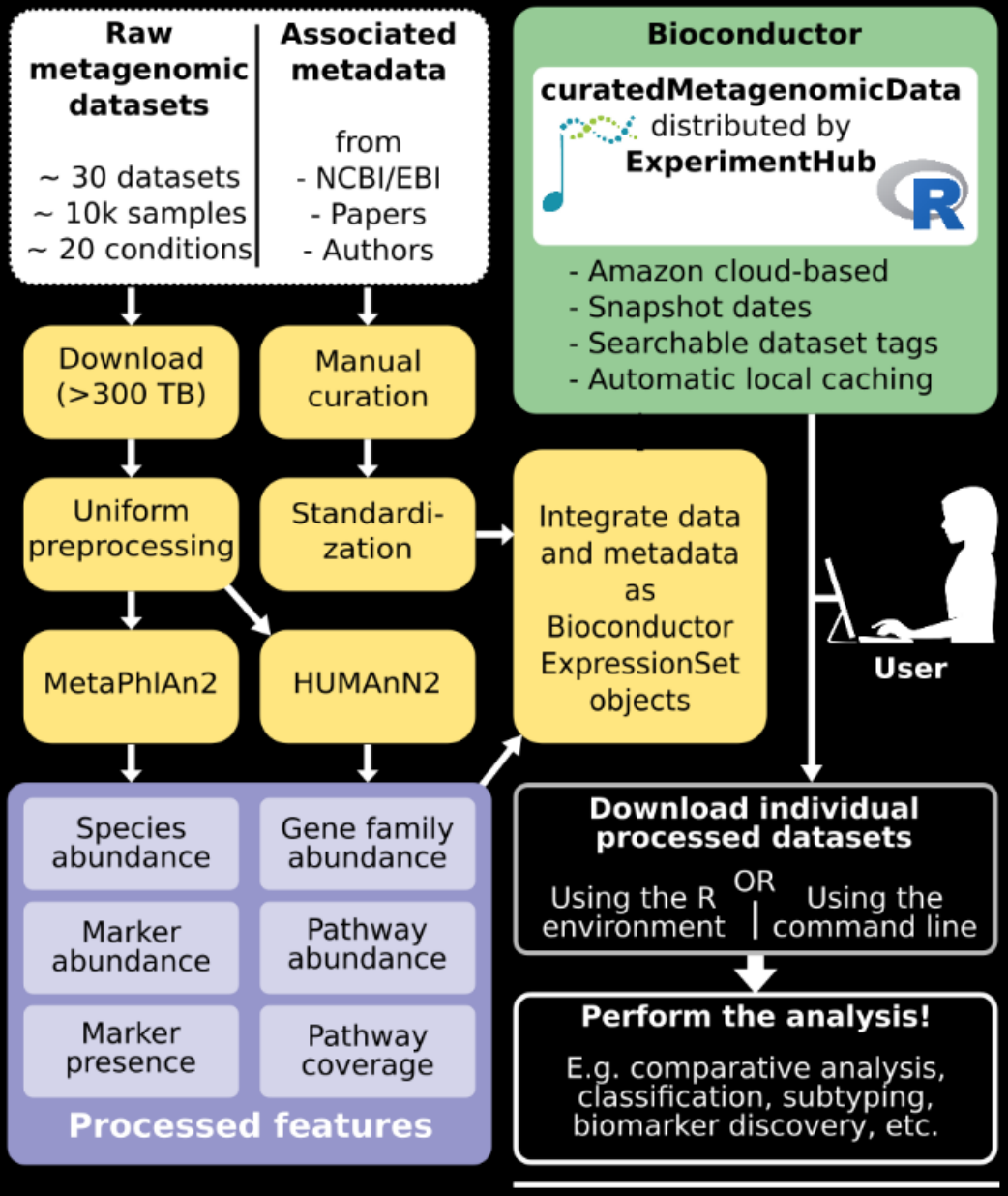
The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment

Potential of fecal microbiota for early-stage detection of colorectal cancer

molecular systems biology

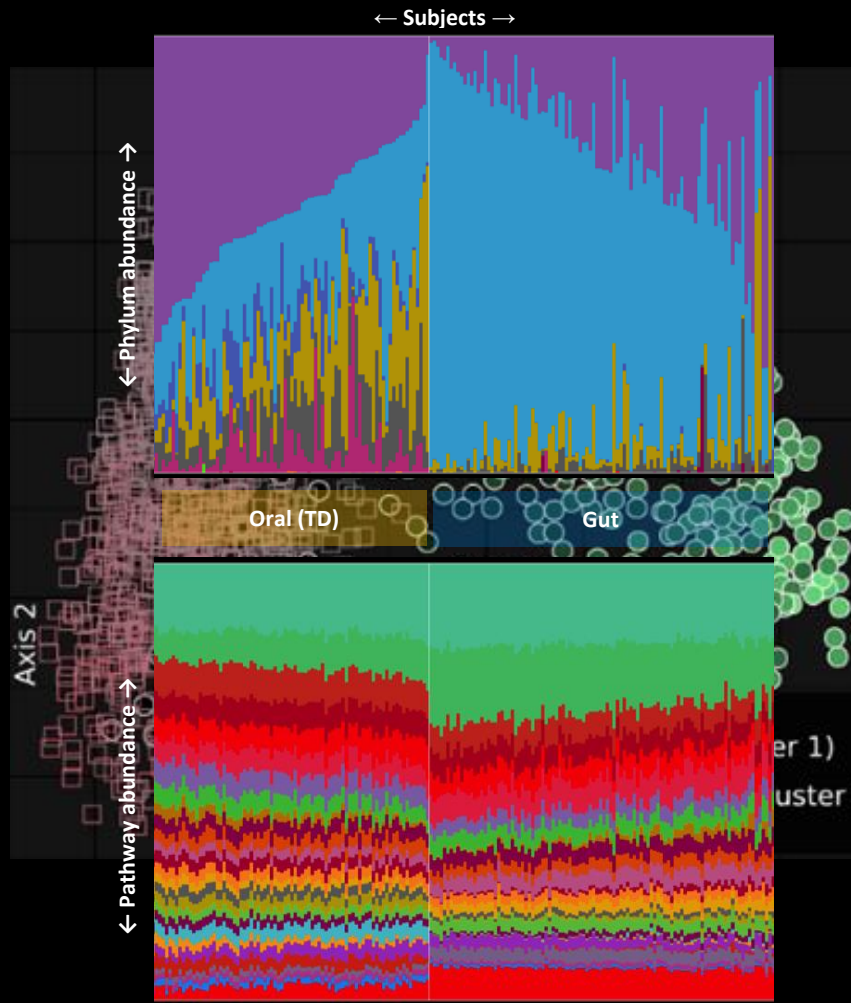


An integrative resource: curatedMetagenomicData

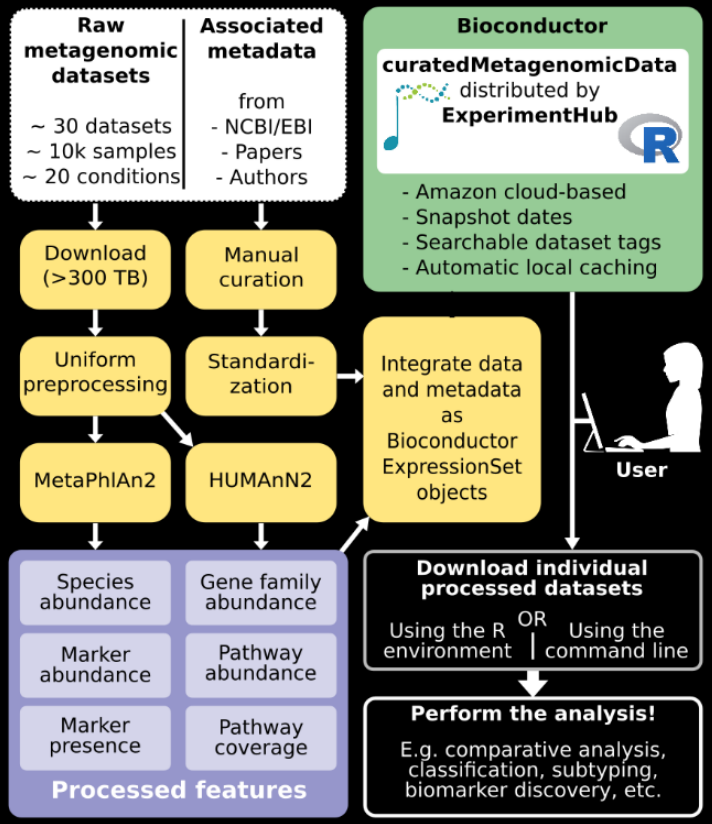


Updated numbers:

- 25 large scale studies
- 5,716 samples
- 28 different countries
- 34 distinct diseases/conditions



An integrative resource: curatedMetagenomicData



Edoardo Pasolli

Levi Waldron

- Mandatory metadata fields:

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

- Optional metadata fields: all the available ones!

Preprint: http://biorxiv.org/content/early/2017/01/27/103085
Main page: https://waldronlab.github.io/curatedMetagenomicData/
Repository: https://github.com/waldronlab/curatedMetagenomicData
incl. data, scripts, tutorials, examples

- Currently 6,000 samples
- Will continue adding datasets
  - 7,000 new samples prioritized for addition
  - 3,000 additional samples without minimal info
- Will add new analyses

Strains vs species as the building blocks of the microbiome

Commensal strains

Pathogenic strains

Environmental strains

E. coli

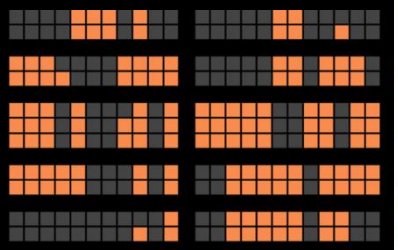
Predicting and Manipulating Cardiac Drug Inactivation by the Human Gut Bacterium *Eggerthella lenta*

Science AAAS

Henry J. Haiser,[1] David B. Gootenberg,[1] Kelly Chatman,[1] Gopal Sirasani,[2] Emily P. Balskus,[2] Peter J. Turnbaugh[1]*

Some strains of *E. lenta* inactivate digoxin (cardiac drug).

But not all strains of *E. lenta* do that!



Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis

eLIFE elife.elifesciences.org

Jose U Scher[1†], Andrew Sczesnak[2,3†], Randy S Longman[2,4†], Nicola Segata[5,6], Carles Ubeda[7,8], Craig Bielski[6], Tim Rostron[9], Vincenzo Cerundolo[9], Eric G Pamer[7], Steven B Abramson[1], Curtis Huttenhower[6], Dan R Littman[2,10*]

Some *P. copri* strains are associated with enhanced risk for arthritis

Strains matter: an additional supporting evidence

Pasolli et al., PLoS Comput Biol, 2016

Dataset name	Body site	Disease	#stages	#case samples	#control samples	Average reads per sample (std)	Reference
Cirrhosis	Gut	Liver Cirrhosis	2	118	114	51.6M (30.9M)	[33]
Colorectal	Gut	Colorectal Cancer	1	48	73	60.0M (25.5M)	[34]
HMP	Several	None	1	-	981	61.1M (51.2M)	[1]
IBD	Gut	Inflammatory Bowel Diseases	1	25	85	45.2M (18.4M)	[35]
Obesity	Gut	Obesity	1	164	89	68.2M (23.2M)	[31]
Skin	Skin	None	1	-	287	24.7M (38.1M)	[36]
T2D	Gut	Type 2 diabetes	2	170	174	40.2M (11.8M)	[37]
WT2D	Gut	Type 2 diabetes	1	53	43	31.0M (17.6M)	[32]

	Cirrhosis	Colorectal	IBD	Obesity	T2D	WT2D
#samples	232	121	110	253	344	96
Species-level features	0.922 (0.041)	0.809 (0.086)	0.862 (0.083)	0.648 (0.071)	0.663 (0.066)	0.664 (0.126)
	*				*	*
Strain-level features	**0.963 (0.027)**	0.823 (0.087)	**0.914 (0.084)**	**0.659 (0.073)**	**0.757 (0.056)**	**0.785 (0.104)**

AUC

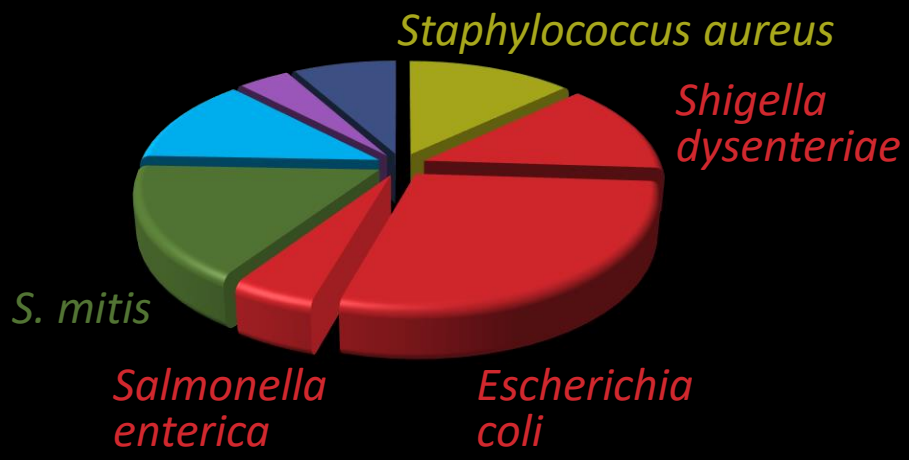
AUC

Edoardo Pasolli

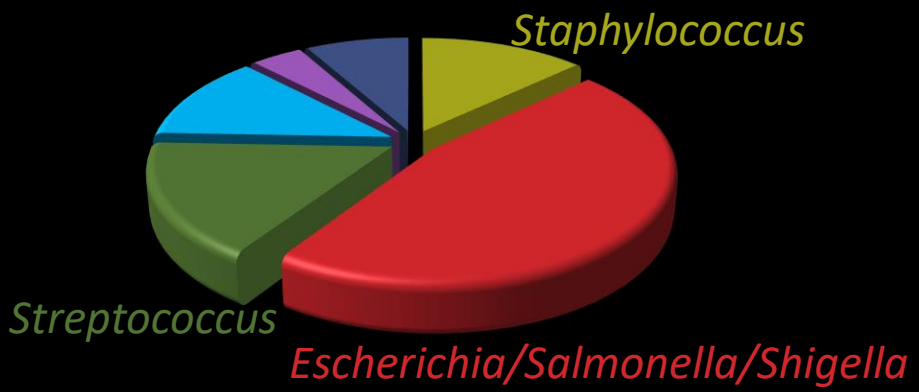
* $P < 0.05$

Goal: expose genomic strain level features

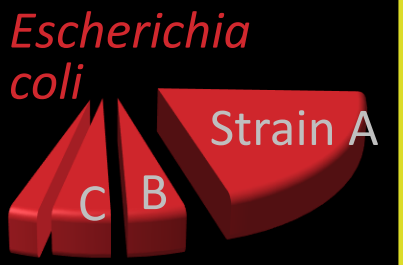
The picture using shotgun metagenomics and MetaPhlAn (and similar tools)



The picture using 16S rRNA sequencing



Next step: strain-level profiling



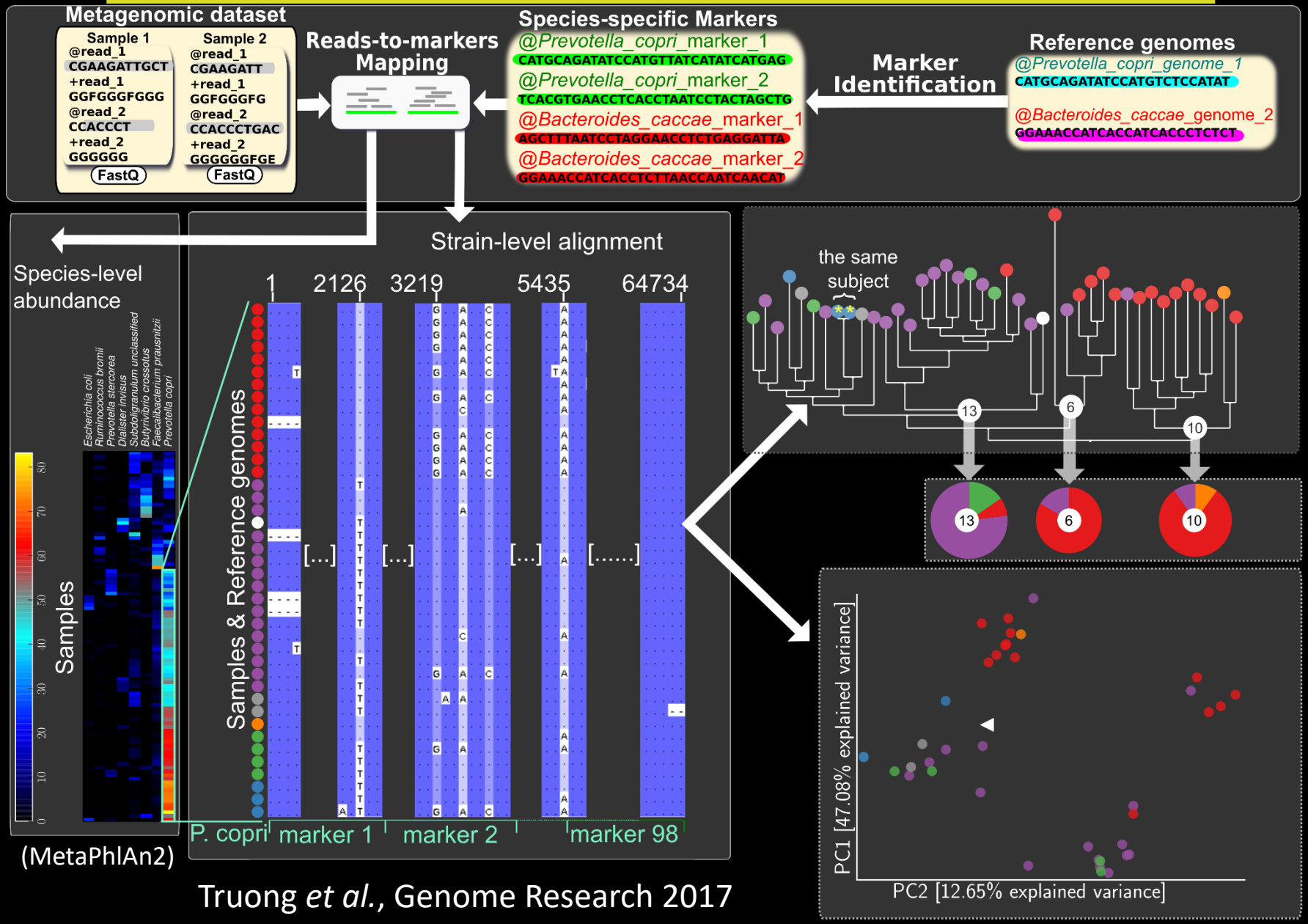
(i) Identify

(ii) Track (e.g. across samples)

(iii) Characterize (genomically)

Beghini *et al*, ISMEJ, in press
Truong *et al.*, Genome Research 2017
Asnicar *et al*, mSystems 2017
Scholz *et al.*, Nature Methods 2016.
Ward *et al.*, Cell Reports 2016
Donati *et al.*, Nature Microbiology 2016
Zolfo *et al.*, NAR 2016
Truong *et al.*, Nature Methods 2015.

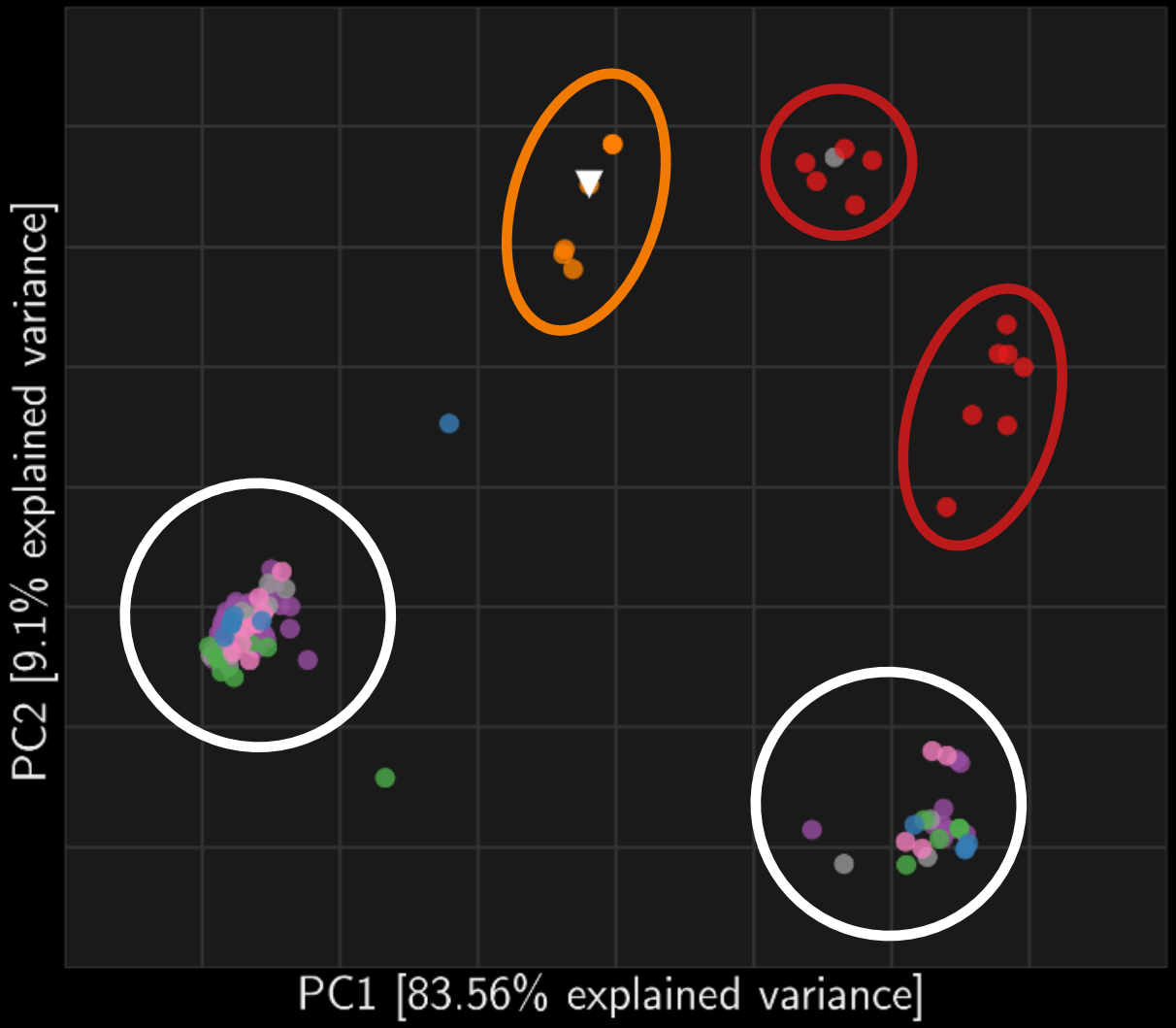
StrainPhlAn: profiling strains by SNVs



Truong *et al.*, Genome Research 2017

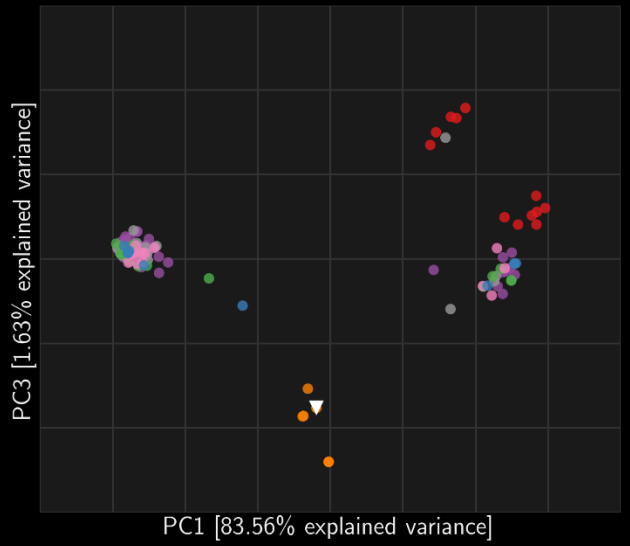
A tool for strain level population genomics

Truong *et al.*, Genome Research 2017



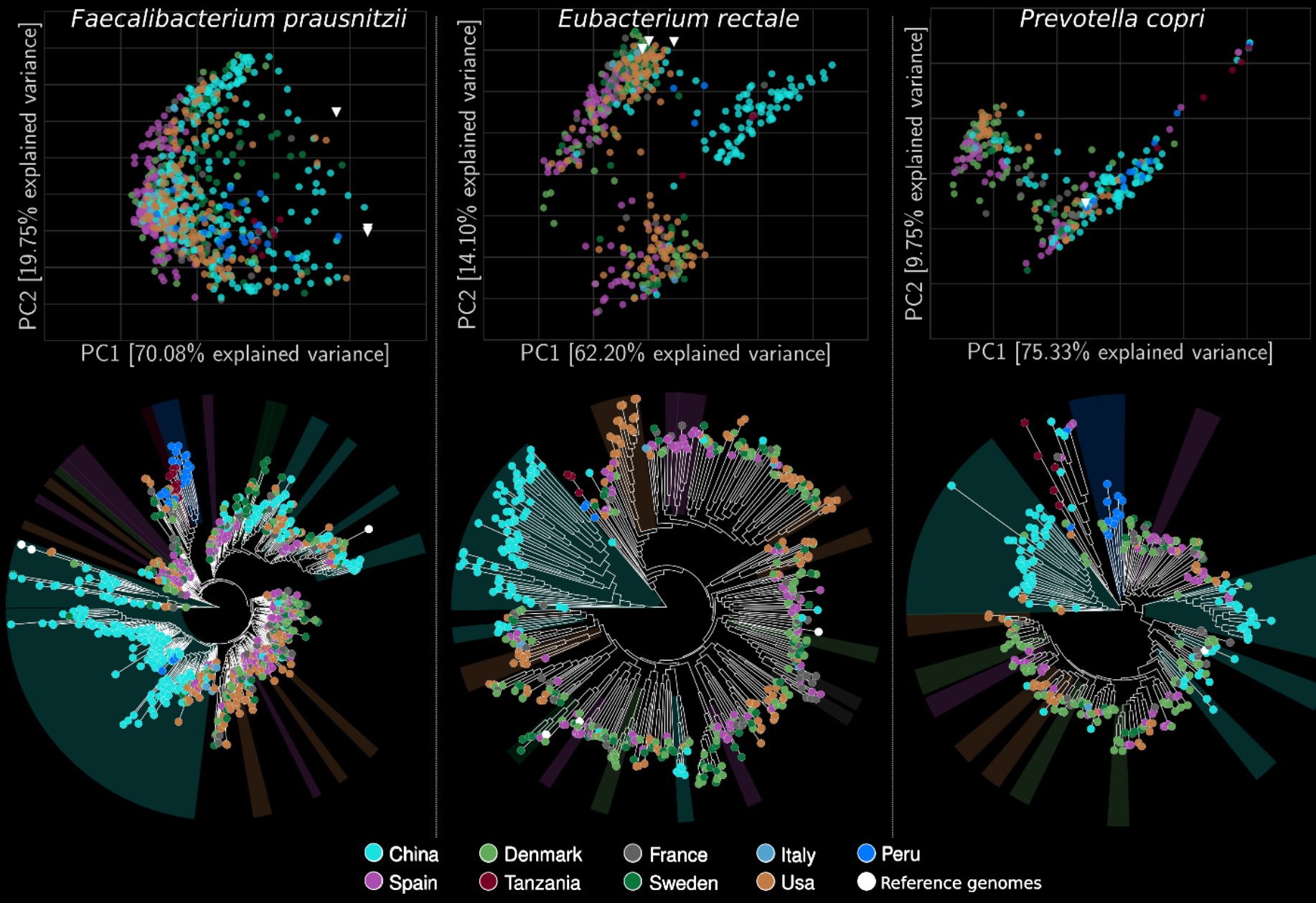
Alignment length: 62k nt
Median SNPs: 830 [1.3%]
pos. samples: 123

Butyrivibrio crossotus strains

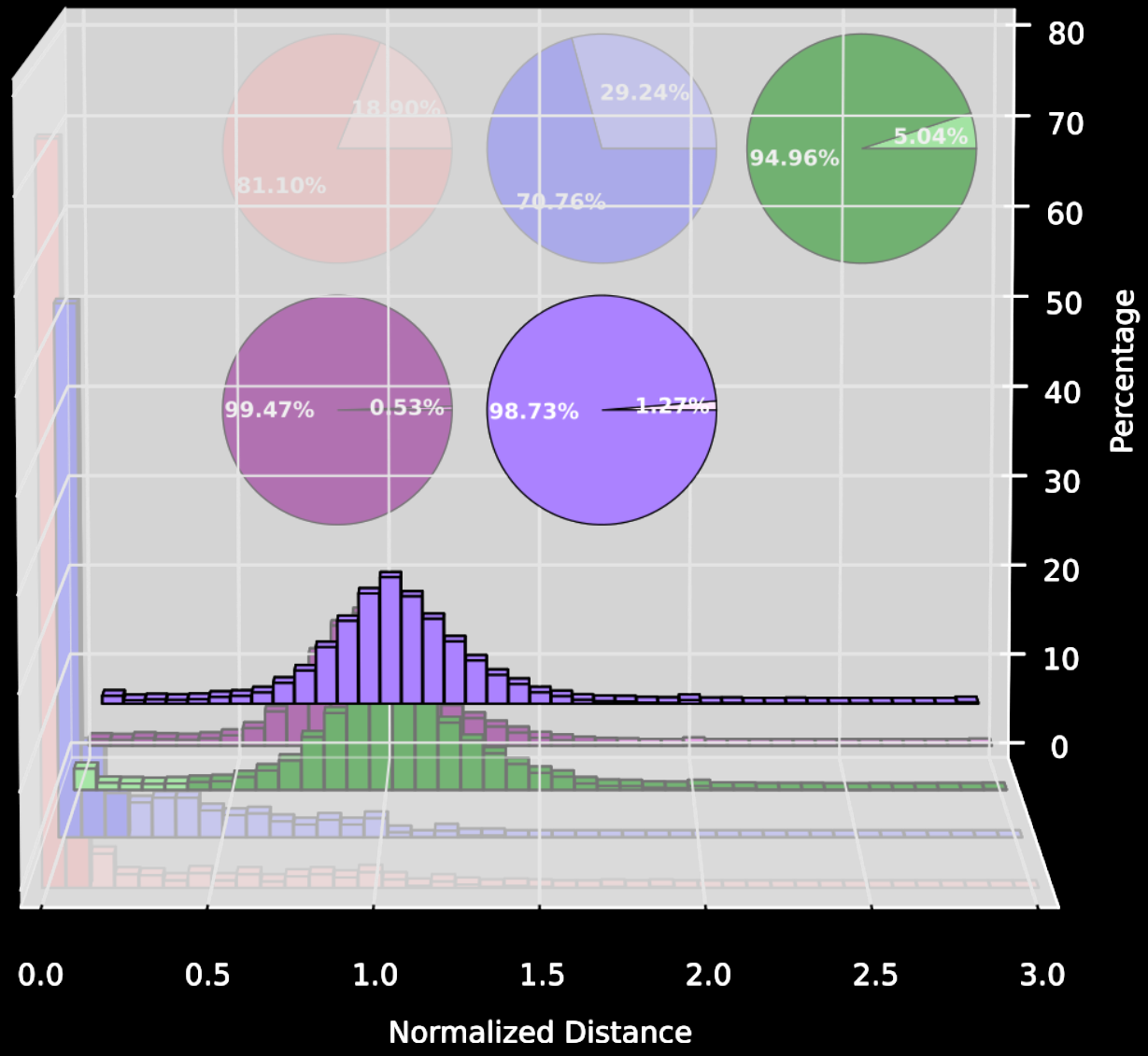


Denmark France Spain Sweden China Usa Peru Ref. Genomes

A tool for strain level population genomics



How stable are strains in the human gut?



Comparing Different subjects

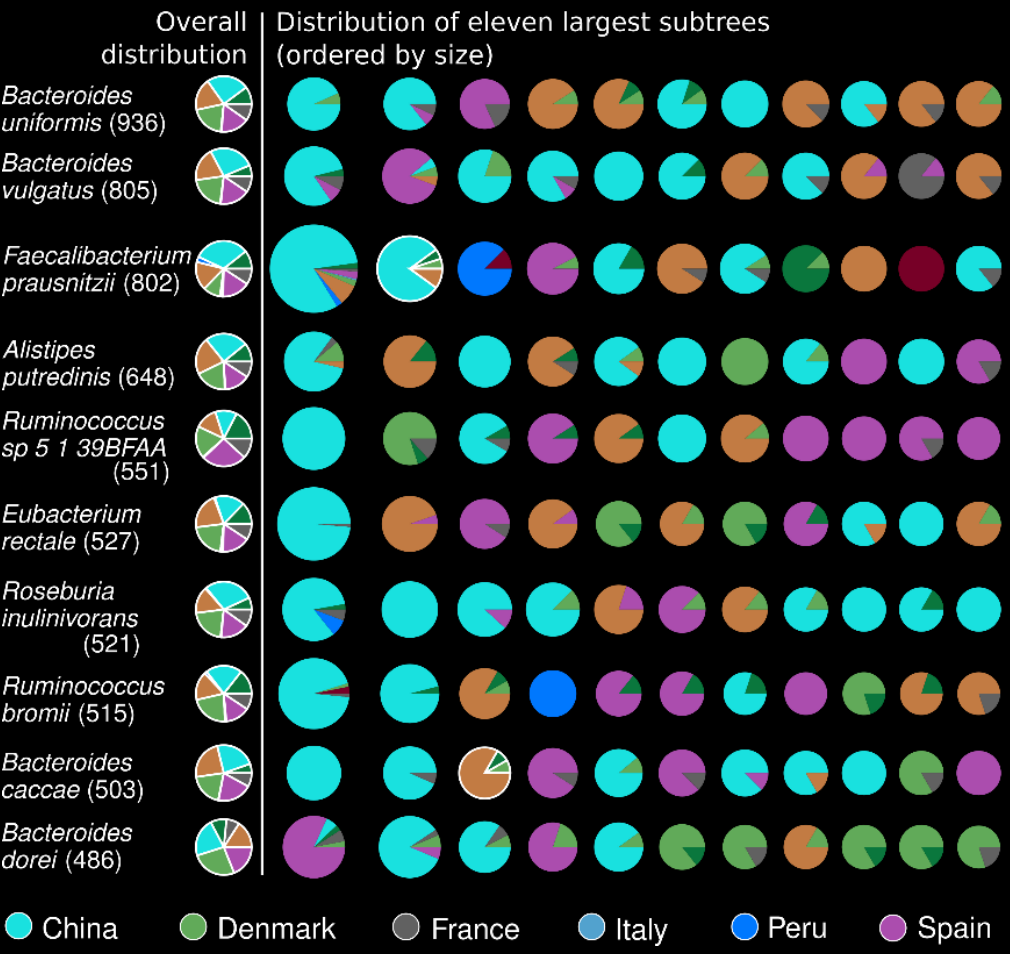
- Subjects from around the world (~3000 sbj from 4 continents)
- Subjects from EU (6 countries)
- Subjects from US (from two universities)

Comparing samples from same subjects collected at ~6 months

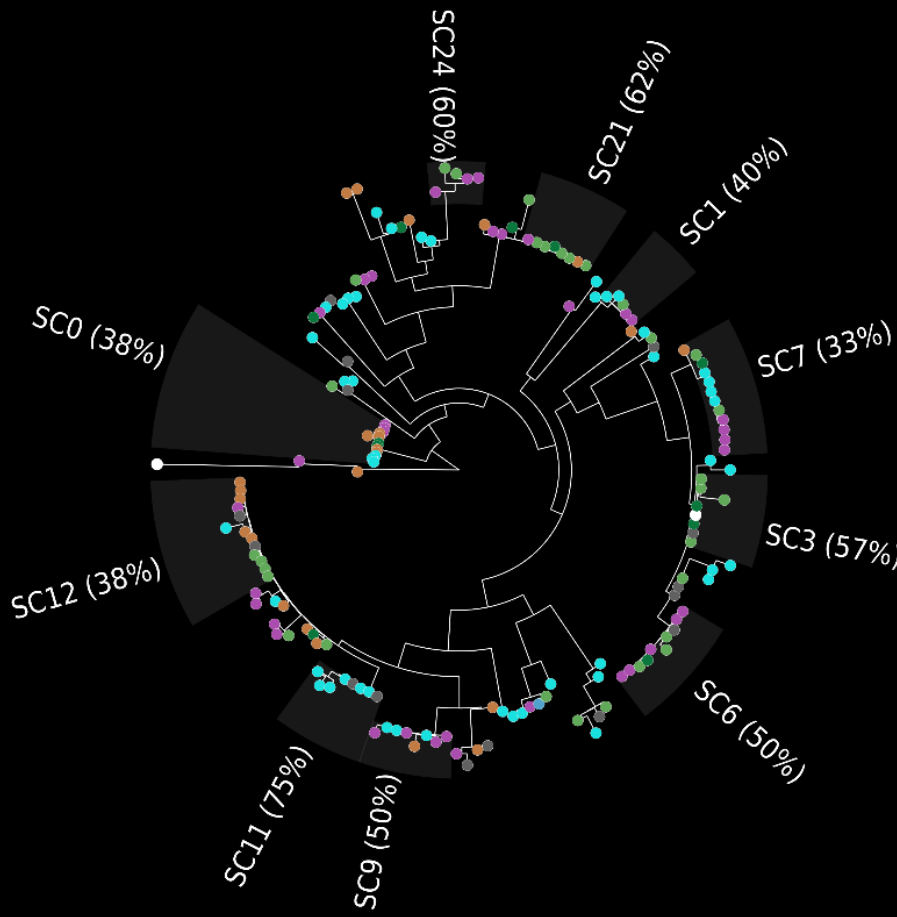
- Subjects from EU
- Subjects from US

A tool for strain level population genomics

Association of sub-species structures with geography



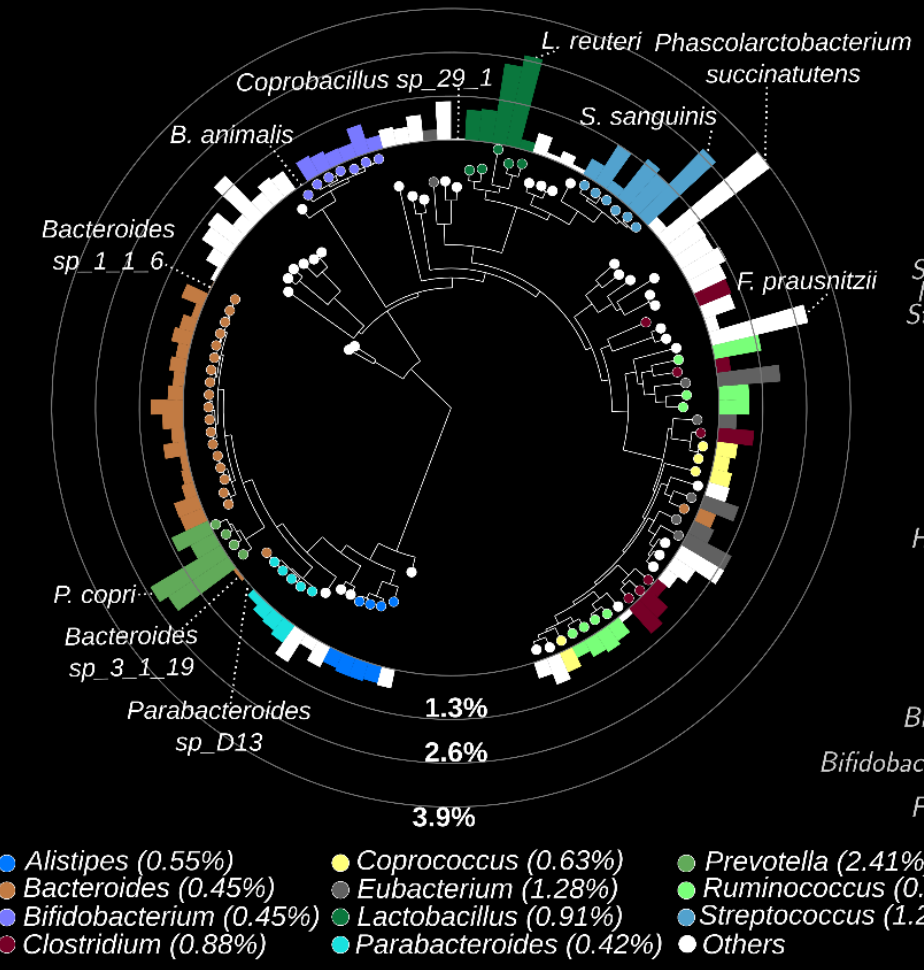
Cosmopolitan strains: the case of *Bacteroides eggerthii*

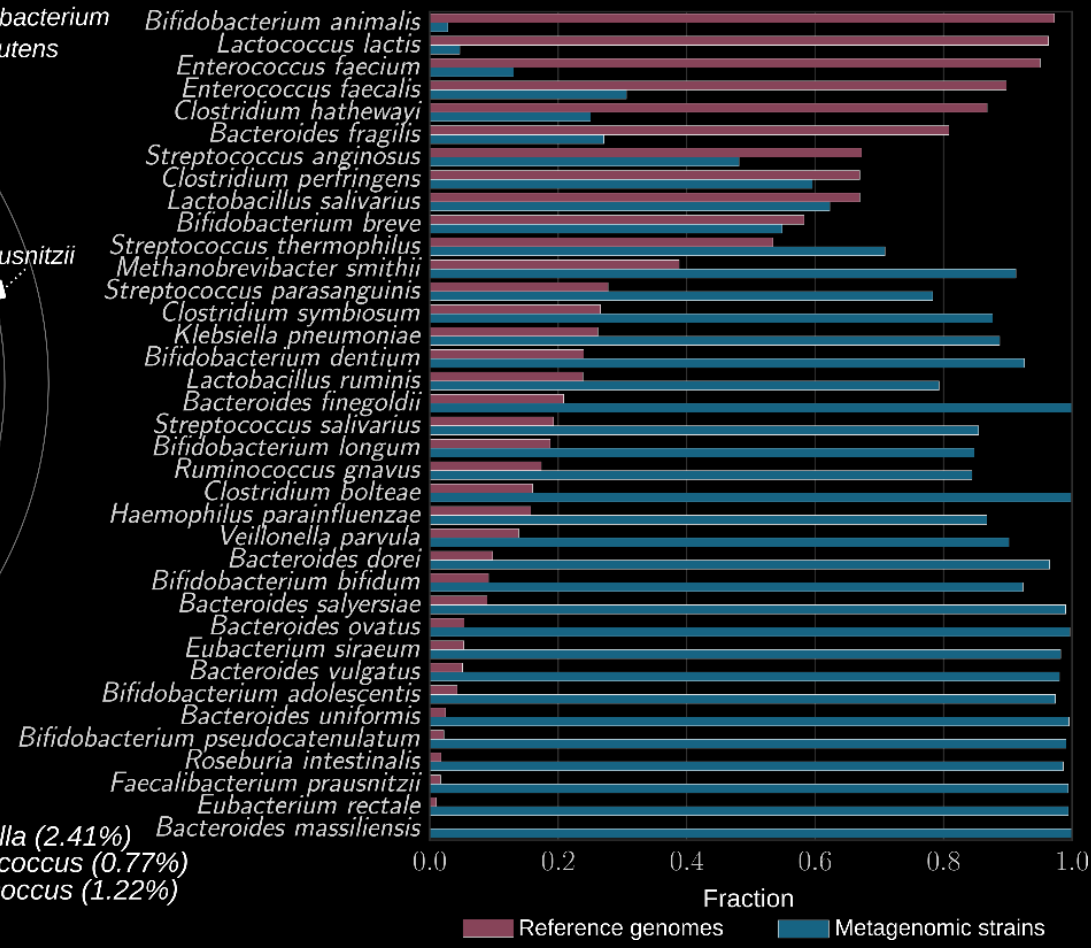


A tool for strain level population genomics

Intra-species diversity is very species and genus dependent

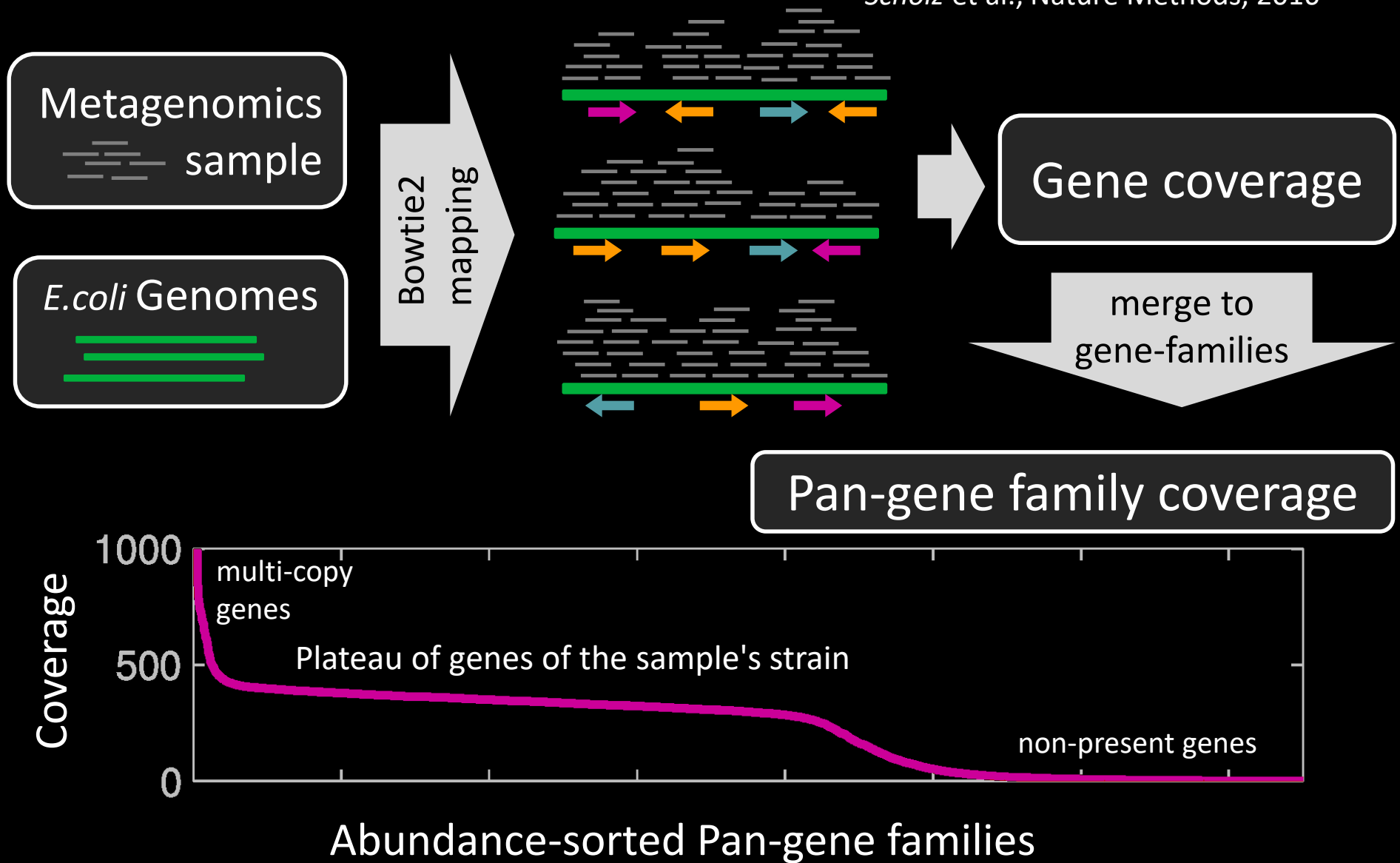
The diversity available from metagenomics





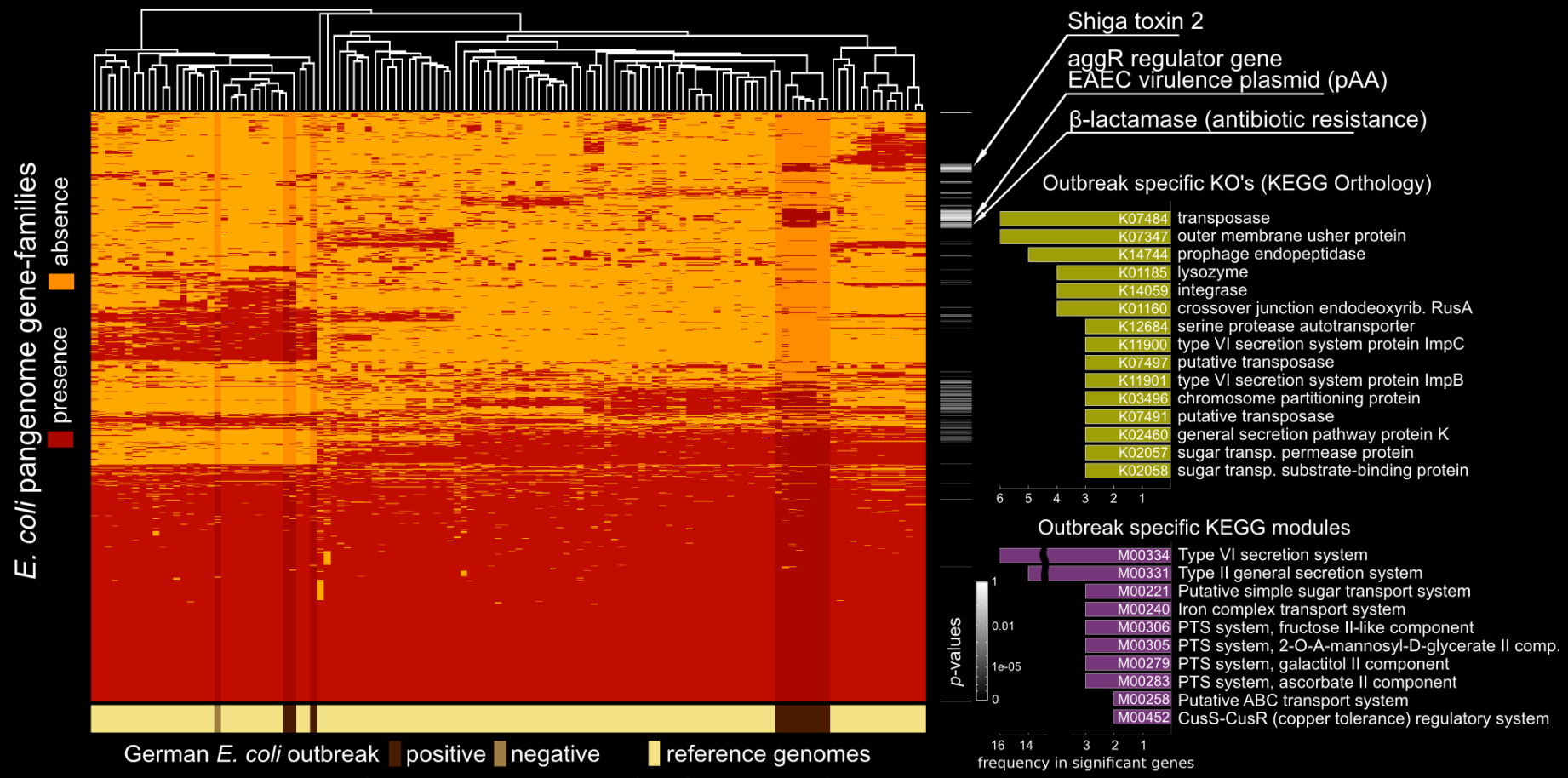
PanPhlAn: profiling strains by gene content

Scholz et al., Nature Methods, 2016



PanPhlAn for “meta-epidemiology”

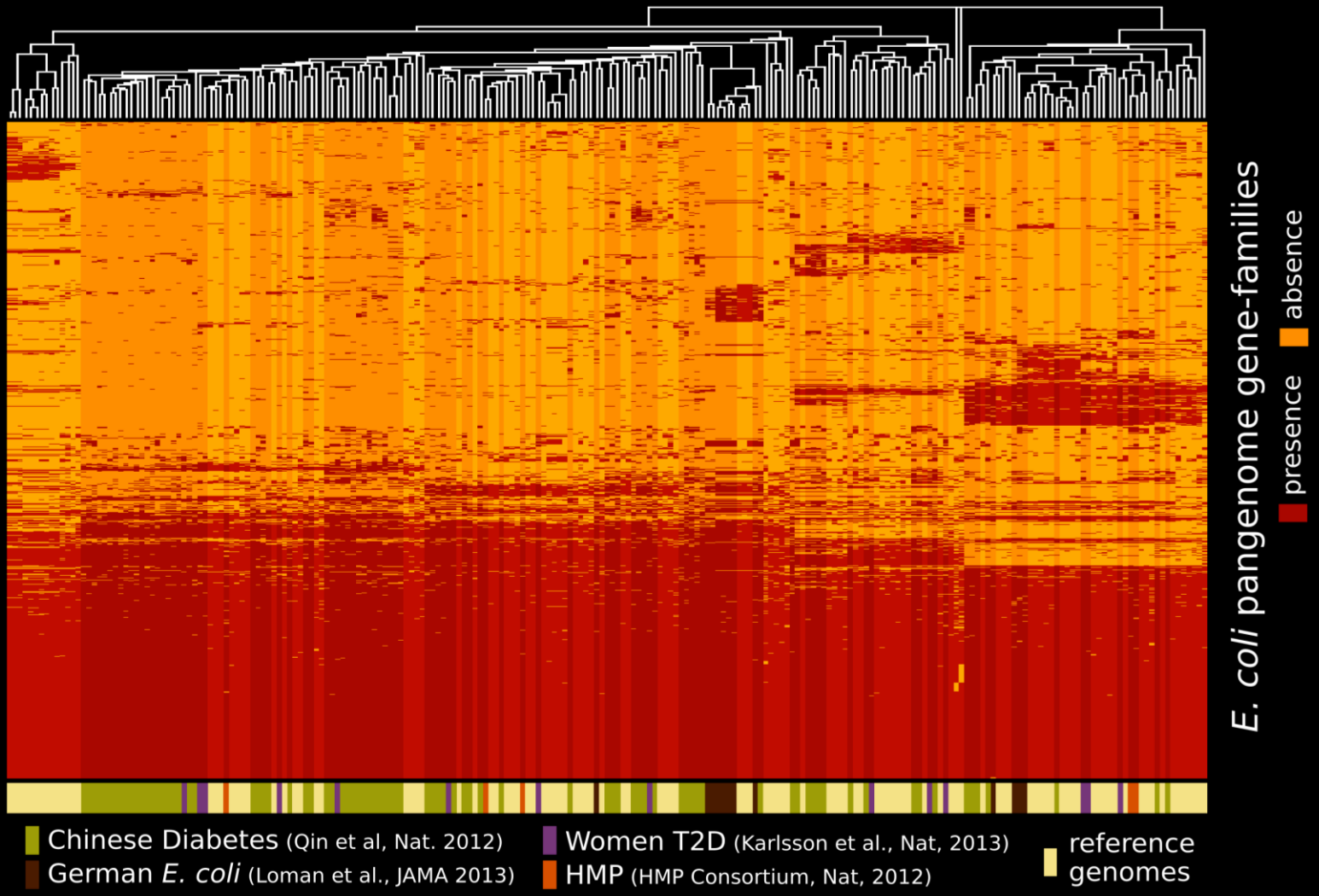
Scholz *et al.*, Nature Methods, 2016



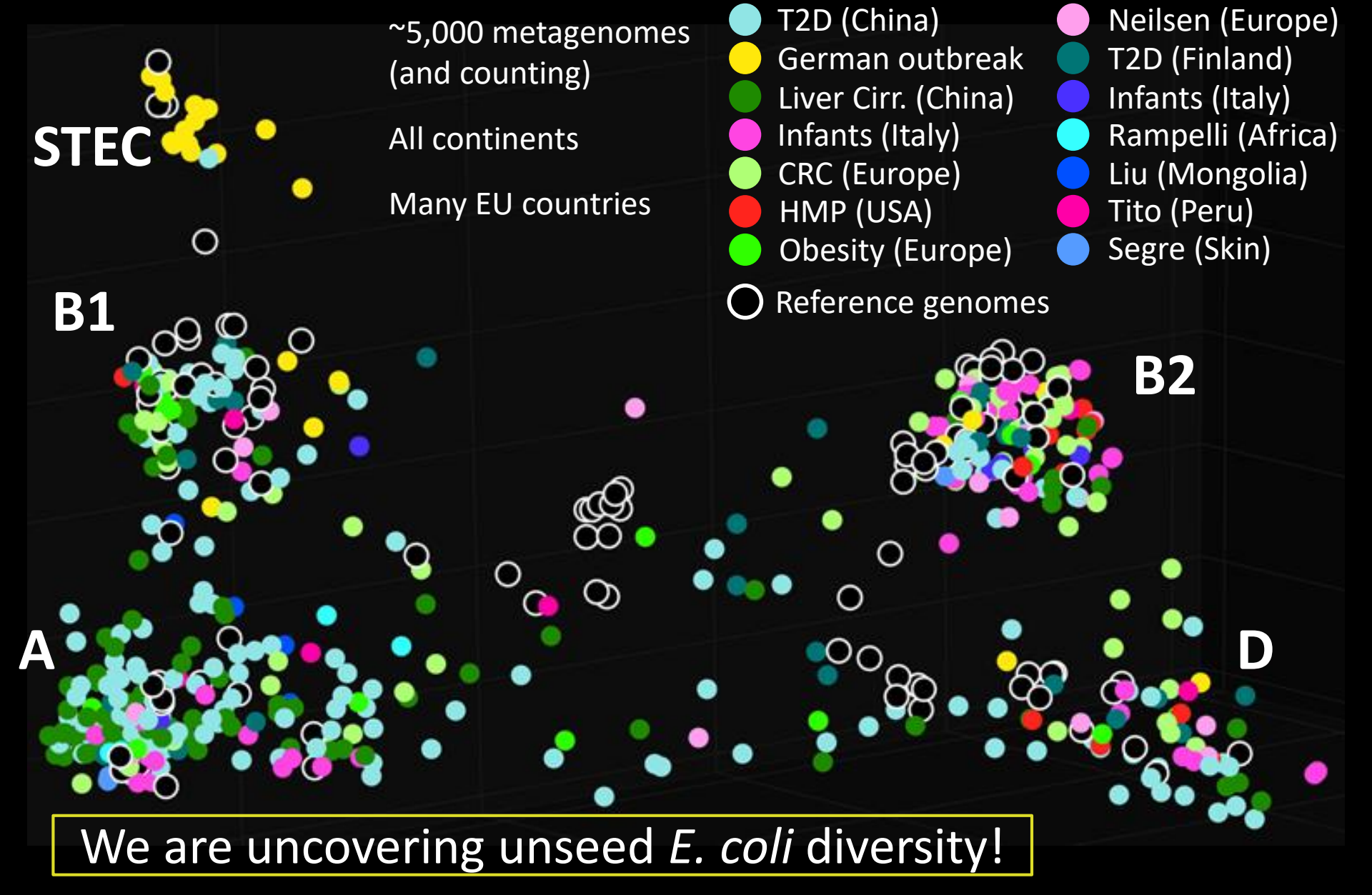
Metagenomes from [Loman et al., 2013]

E. coli population genomics with PanPhlAn

E. coli profiling from 1478 shotgun metagenomes



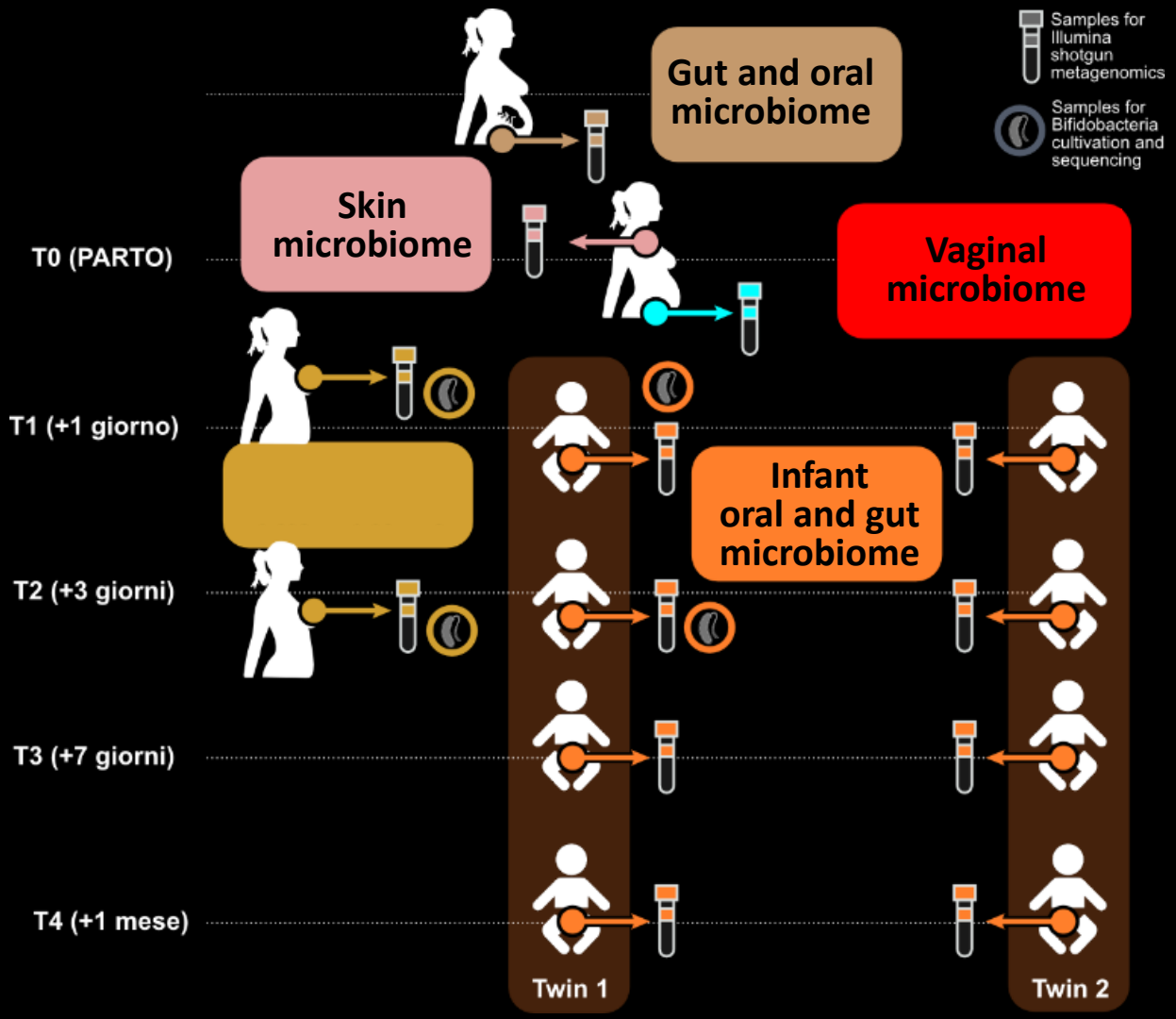
An extended human-associated *E. coli* epidemiology



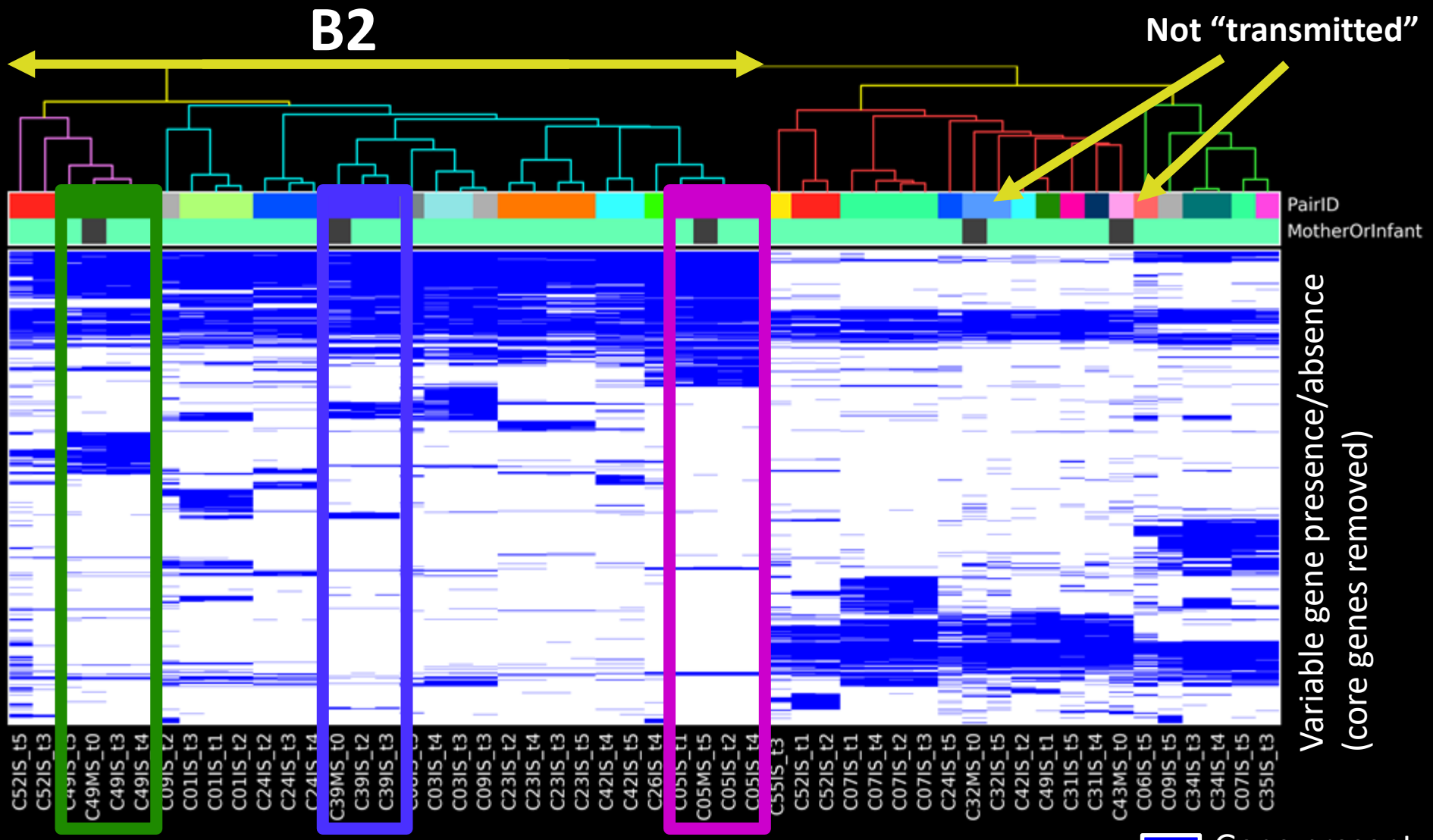
We are uncovering unseed *E. coli* diversity!

Vertical *E. coli* transmission?

Ferretti *et al*., in preparation



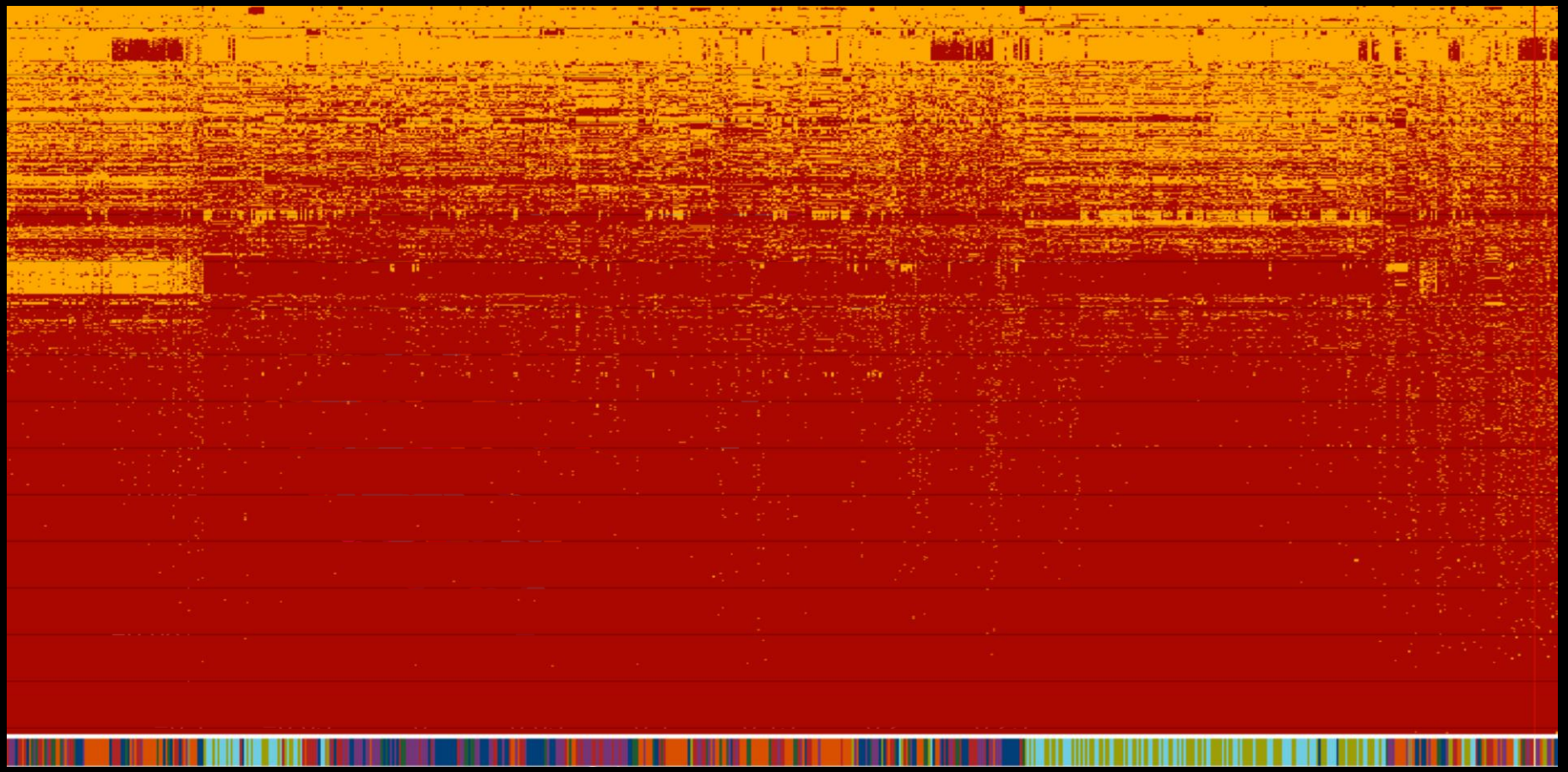
Vertical *E. coli* transmission?



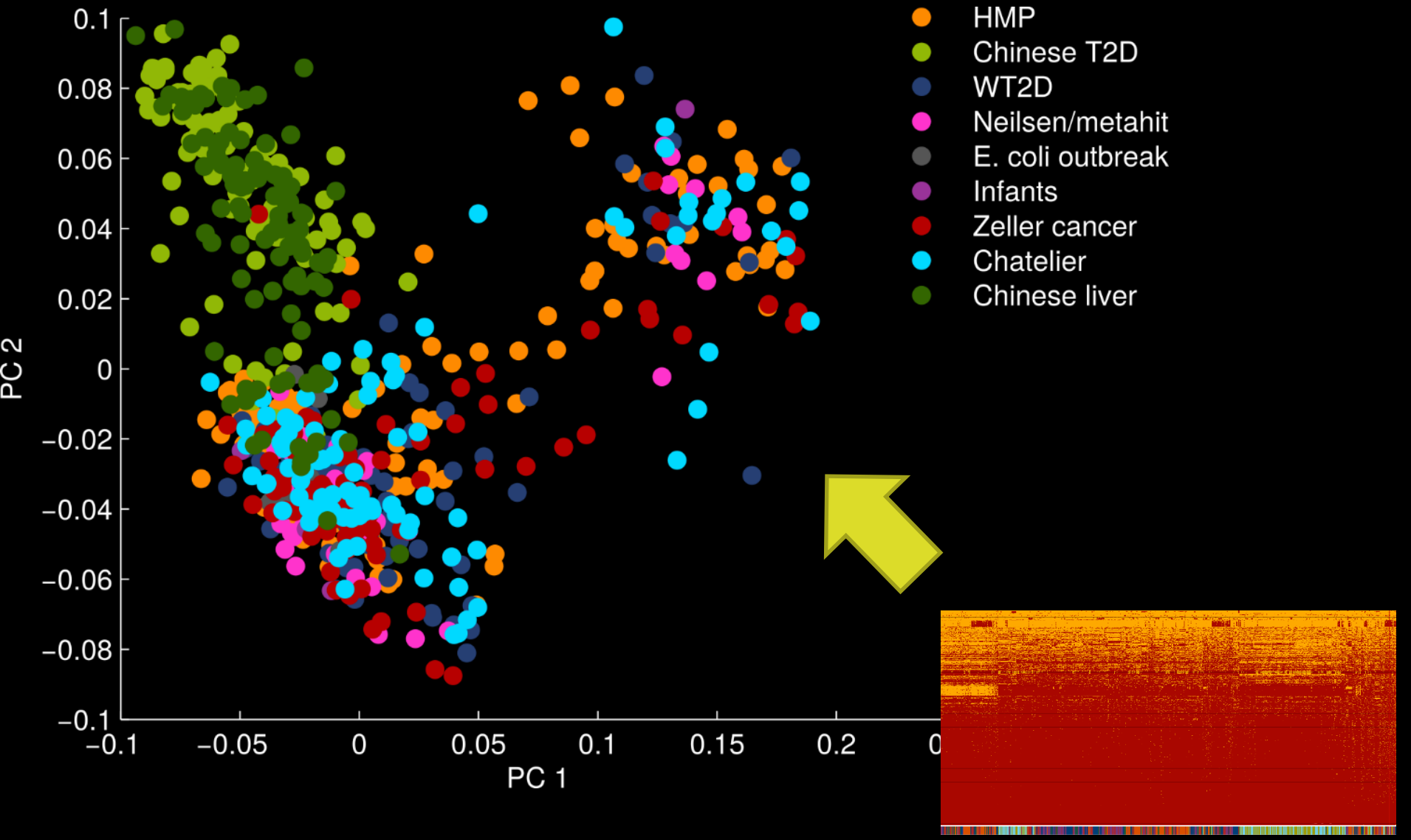
Yes! …but only for the B2 phylotype?

PanPhlAn on *Eubacterium rectale*

Only one *Eubacterium rectale* genome used here

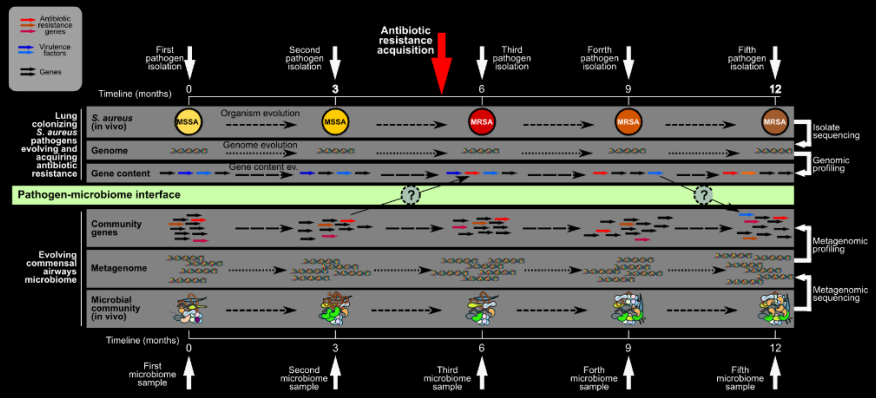


PanPhlAn on *Eubacterium rectale*

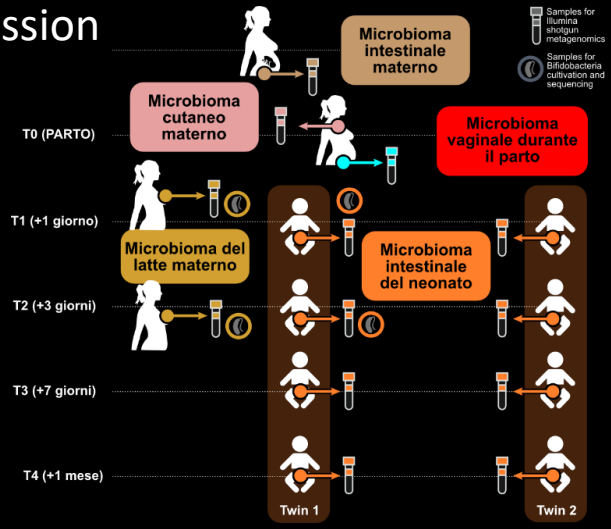


Some examples in which strains count!

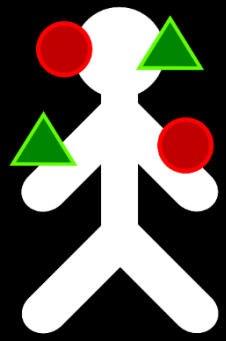
Follow the evolution of pathogenic strains in the lungs of patients with cystic fibrosis

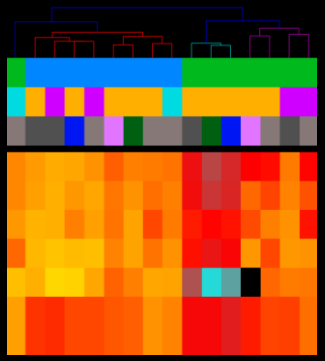


Vertical mother-to-infant microbiome transmission



The skin microbiome in disease

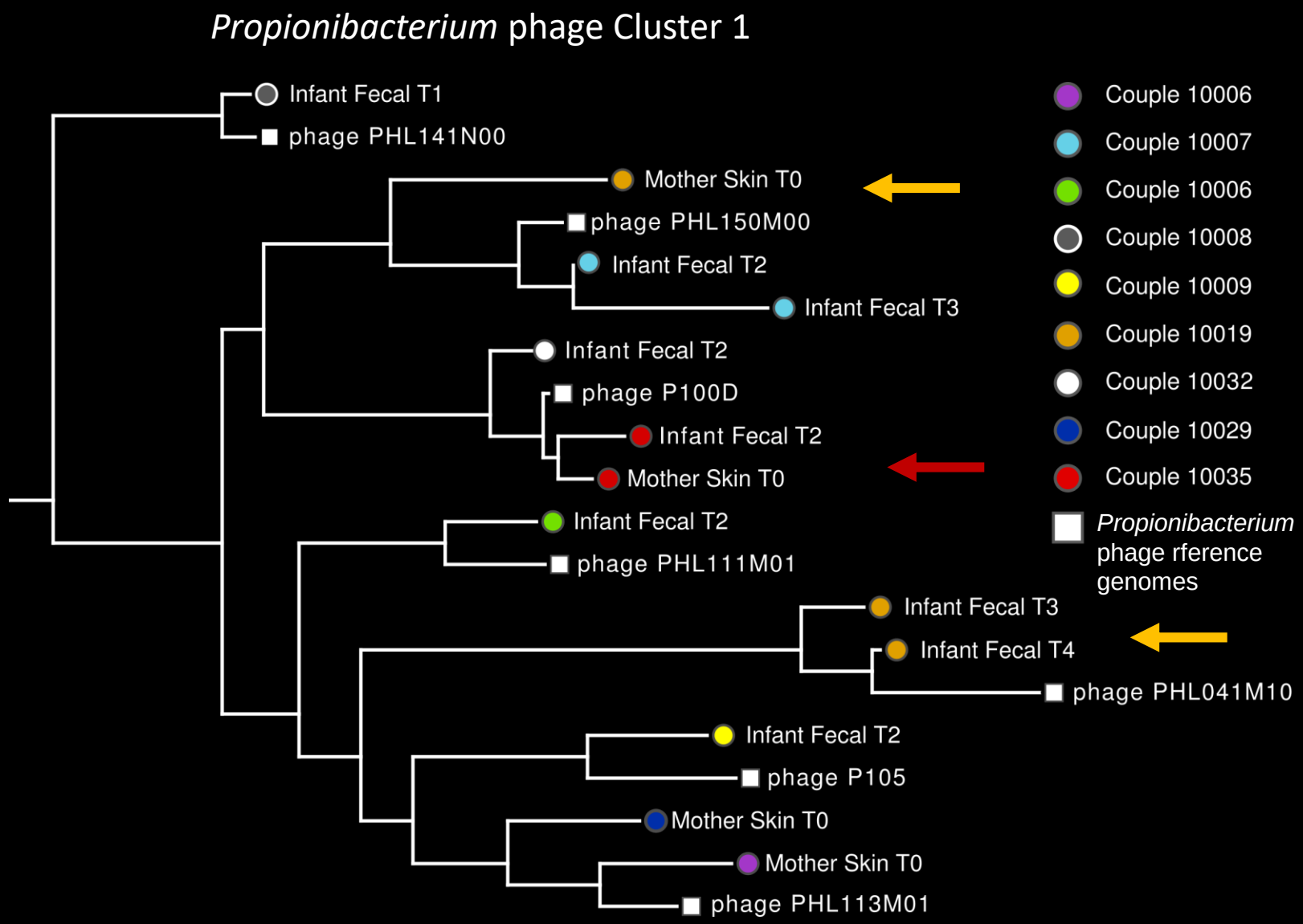




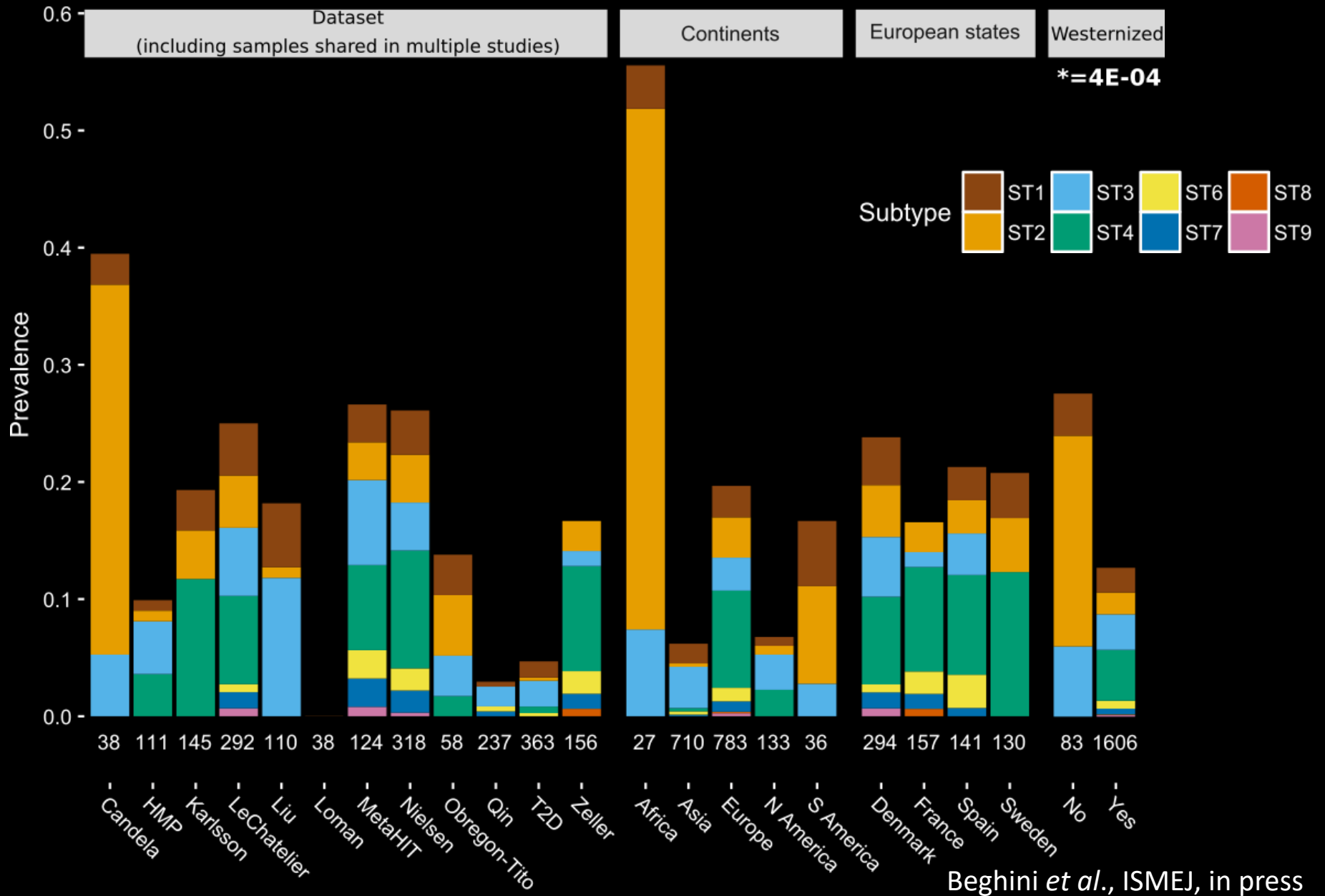
The gut microbiome and colorectal cancer

The oral microbiome and peri-implantitis

Not only microbes: profiling bacteriophages

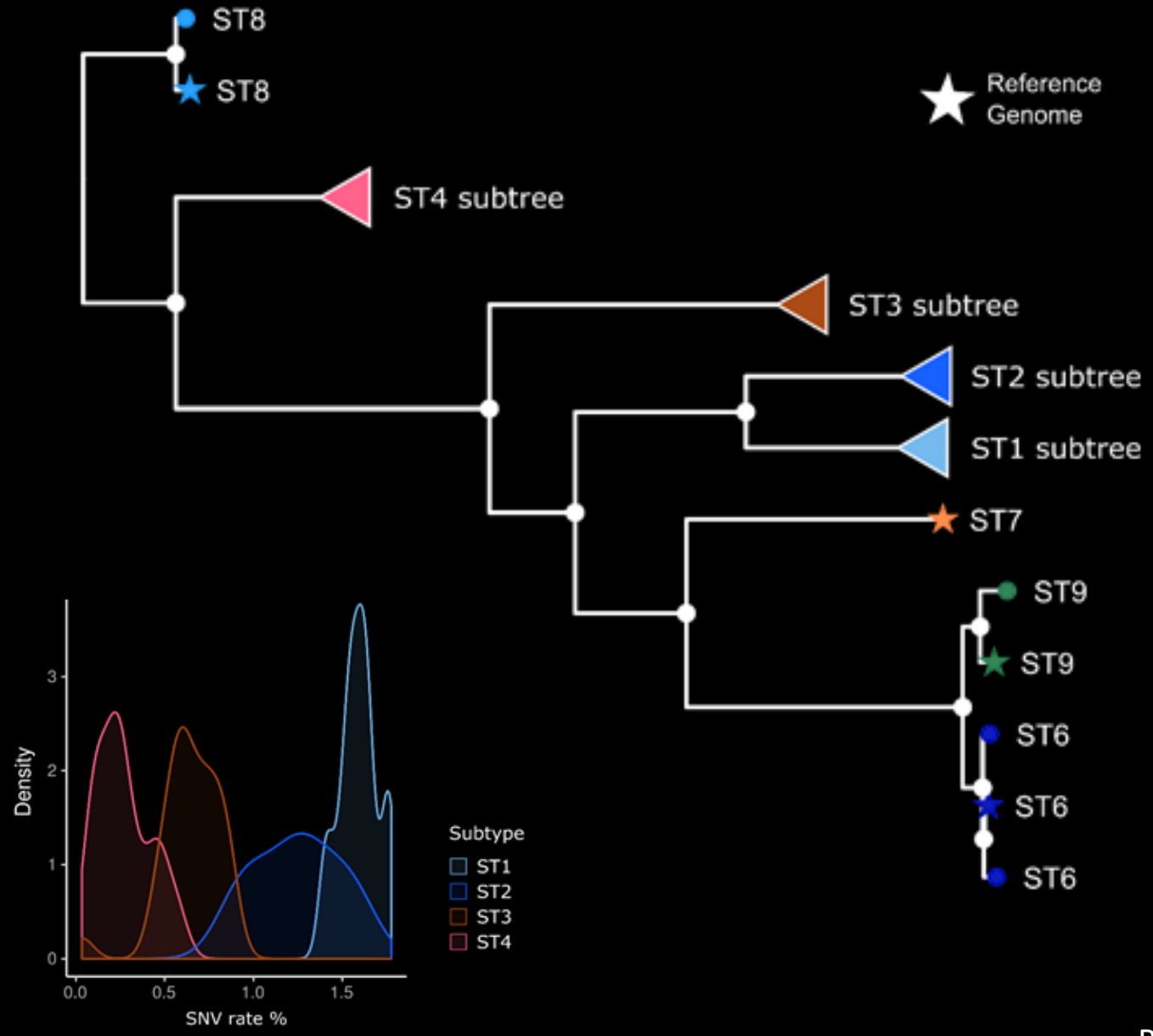


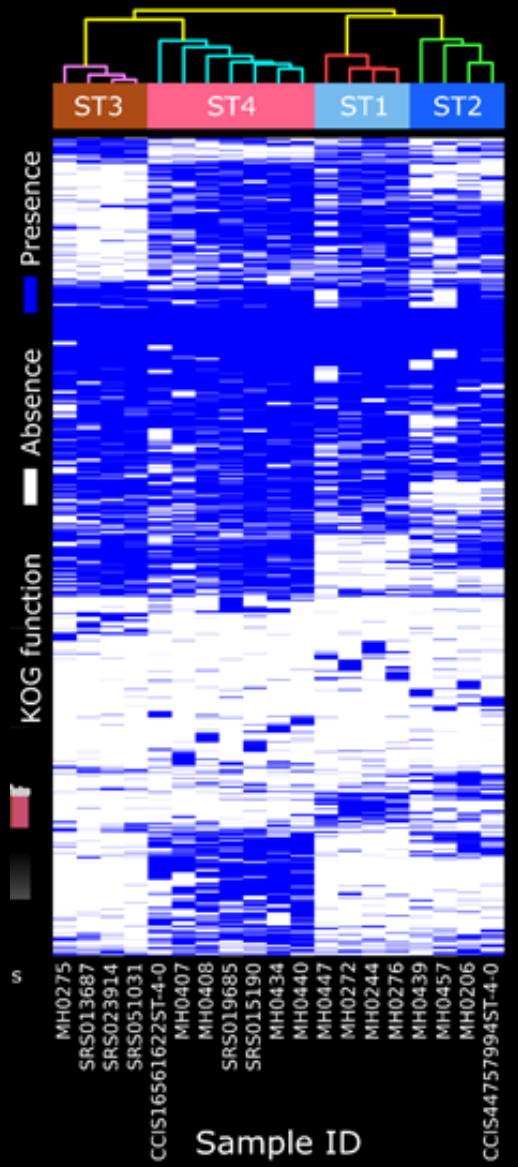
Not only bacteria: the case of *Blastocystis*



Beghini *et al.*, ISMEJ, in press

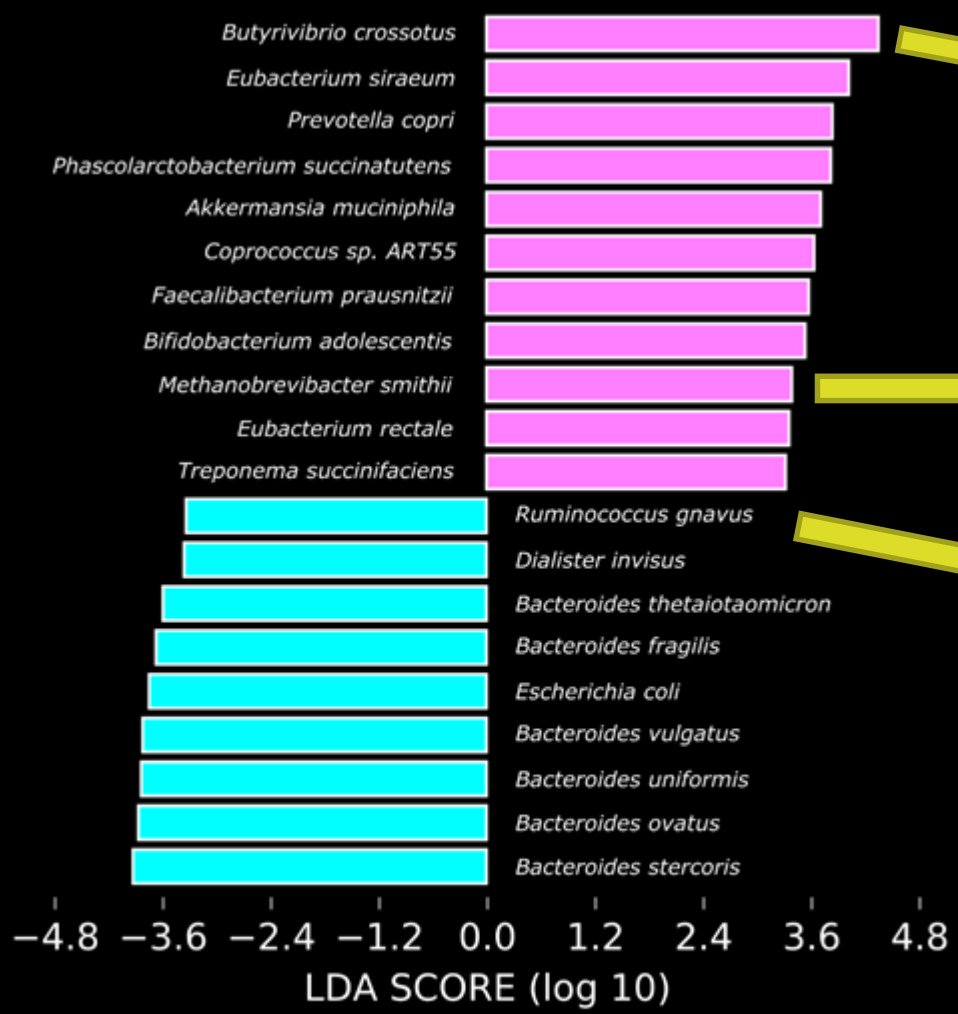
Not only bacteria: the case of *Blastocystis*

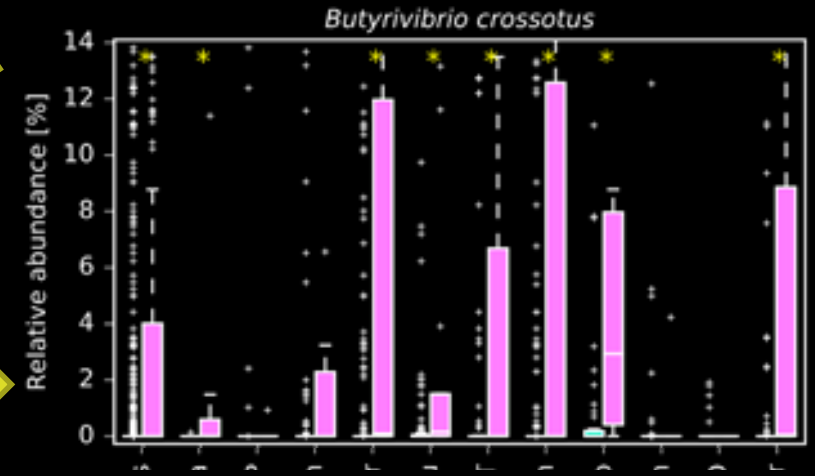


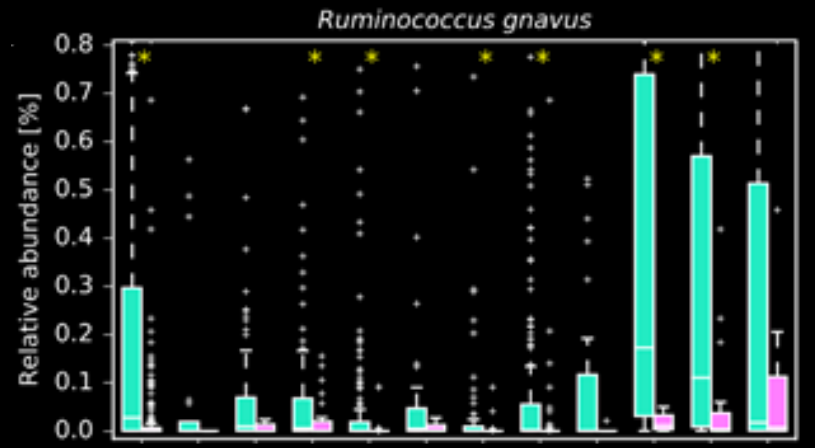


Beghini *et al.*, submitted, 2017

Blastocystis: intriguing ecological relations







	All	Candela	HMP	Karlsson	LeChatelier	Liu	MetaHIT	Nielsen	Obregon-Tito	Qin	T2D	Zeller
Cross validation	0.856	0.941	0.621	0.879	0.822	0.668	0.723	0.793	0.643	0.727	0.858	0.783
Leave-one-dataset-out	-	0.828	0.687	0.737	0.834	0.737	0.773	0.838	0.803	0.874	0.912	0.815

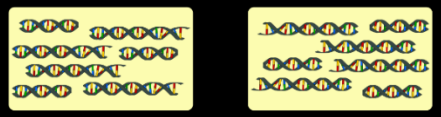
Beghini *et al.*, submitted, 2017

Toward strain-level comparative genomics from metagenomics

Comparative microbial genomics

Isolate and cultivate

Genome sequencing



Strain-level microbial genomics

Step 3

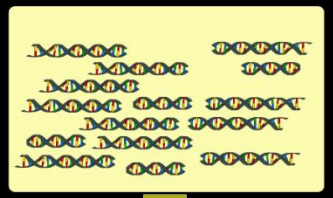
Compare to understand

- High resolution and effective

➔ But available for few microbes only

Shotgun metagenomics

Metagenome sequencing

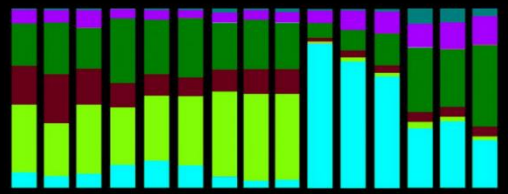


Step 1

Computational analysis

Step 2

Microbiome structure comparison



Microbial ecology *in vivo*

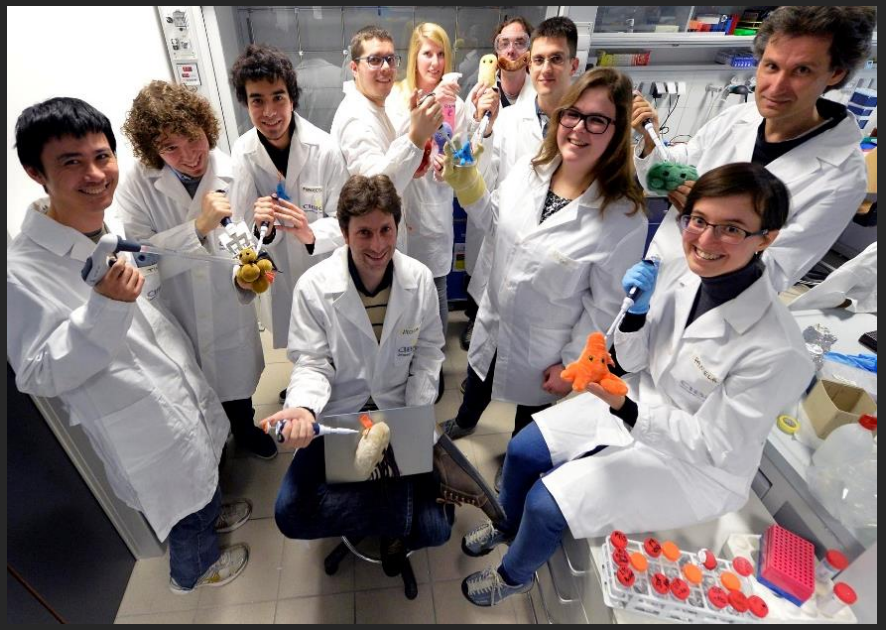
- Access uncultivable microbes (>95%)

➔ But strain level genomics not possible

Thanks!

@nsegata @cibiocm

The Laboratory of Computational Metagenomics



Adrian Tett
Tin Truong
Edoardo Pasolli
Federica Pinto
Federica Armanini
Francesco Asnicar
Serena Manara
Paolo Ghensi
Moreno Zolfo
Pamela Ferretti
Francesco Beghini
Paolo Manghi

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

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