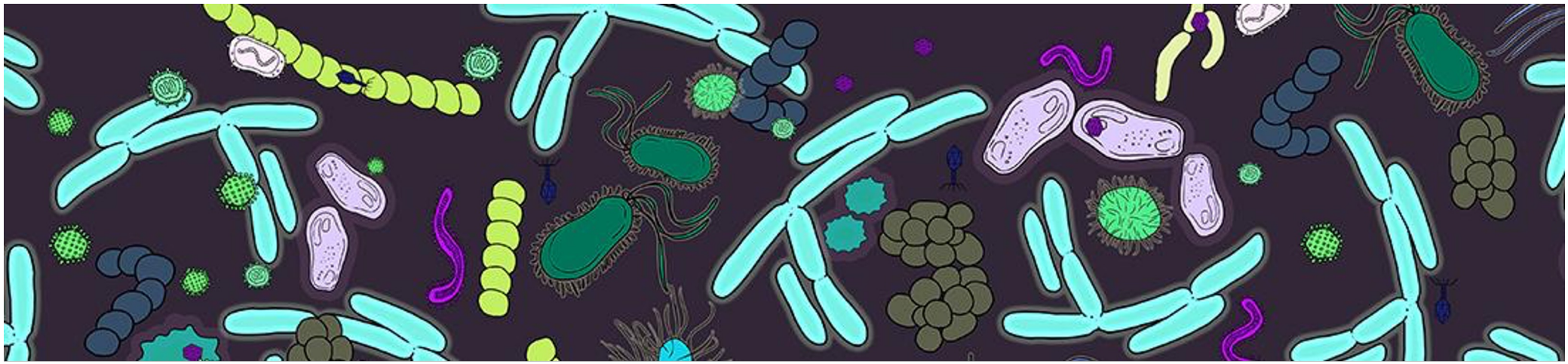


Fear and Trembling in Bacterial Genome Assembly

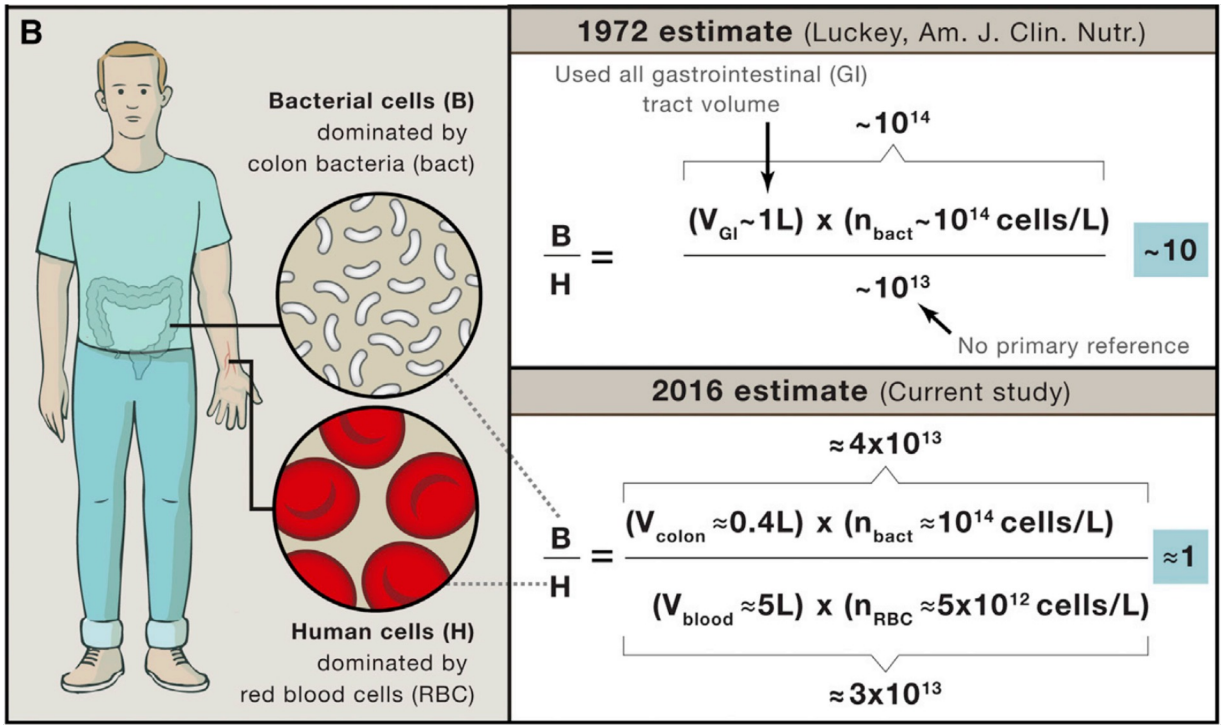
2023-03-15

PennCHOP MICROBIOME PROGRAM

Kyle Bittinger
Division of Gastroenterology, Hepatology, and Nutrition
CHOP Microbiome Center

Illustration: Arwa Abbas

We co-exist with about as many bacterial cells as human cells

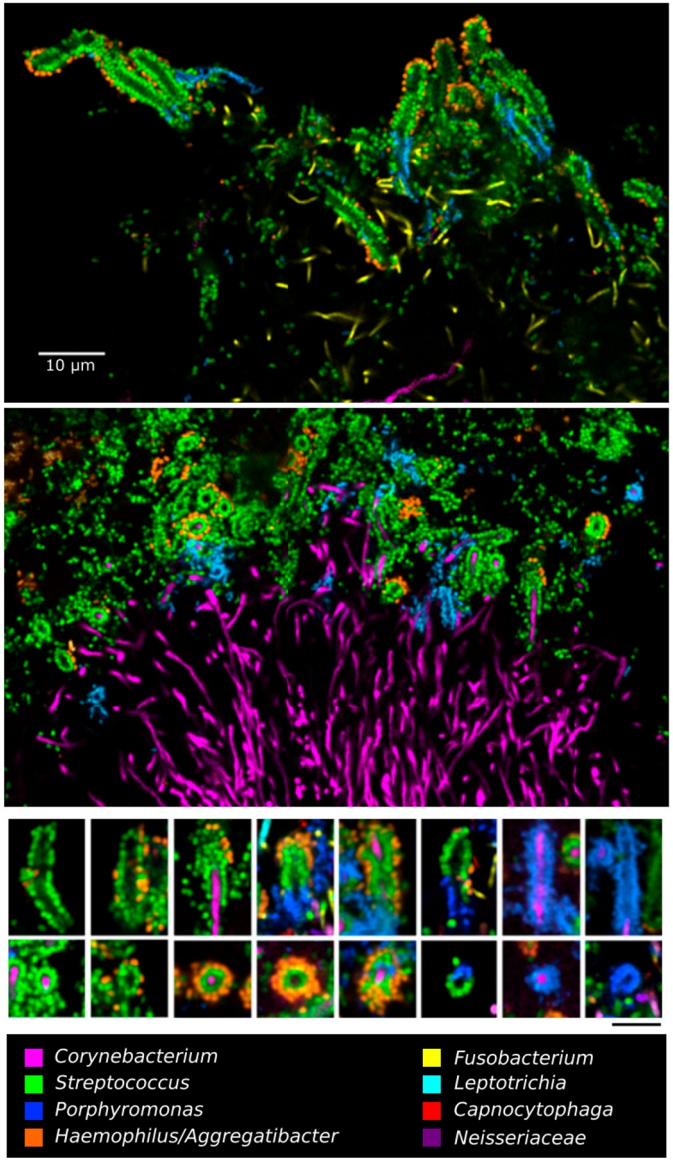


Sender R. *Cell* **164**, 337 (2016).

Roughly 10 trillion bacteria per human

Number of bacteria in colon about 10-50x that in rest of body

Red blood cells account for ~84% of human cells



Complex bacterial communities have spatial structure

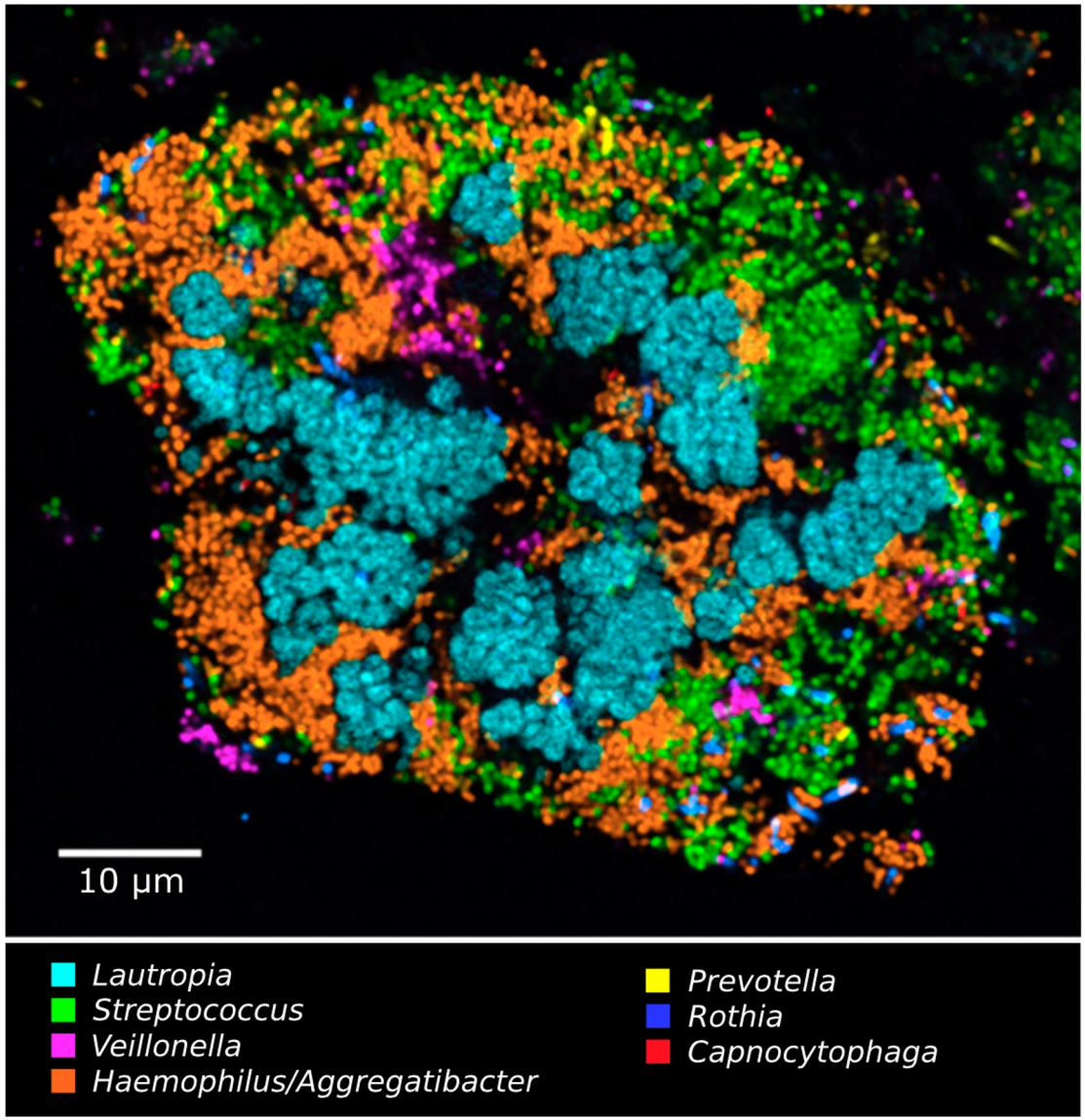
Example: oral bacteria in plaque

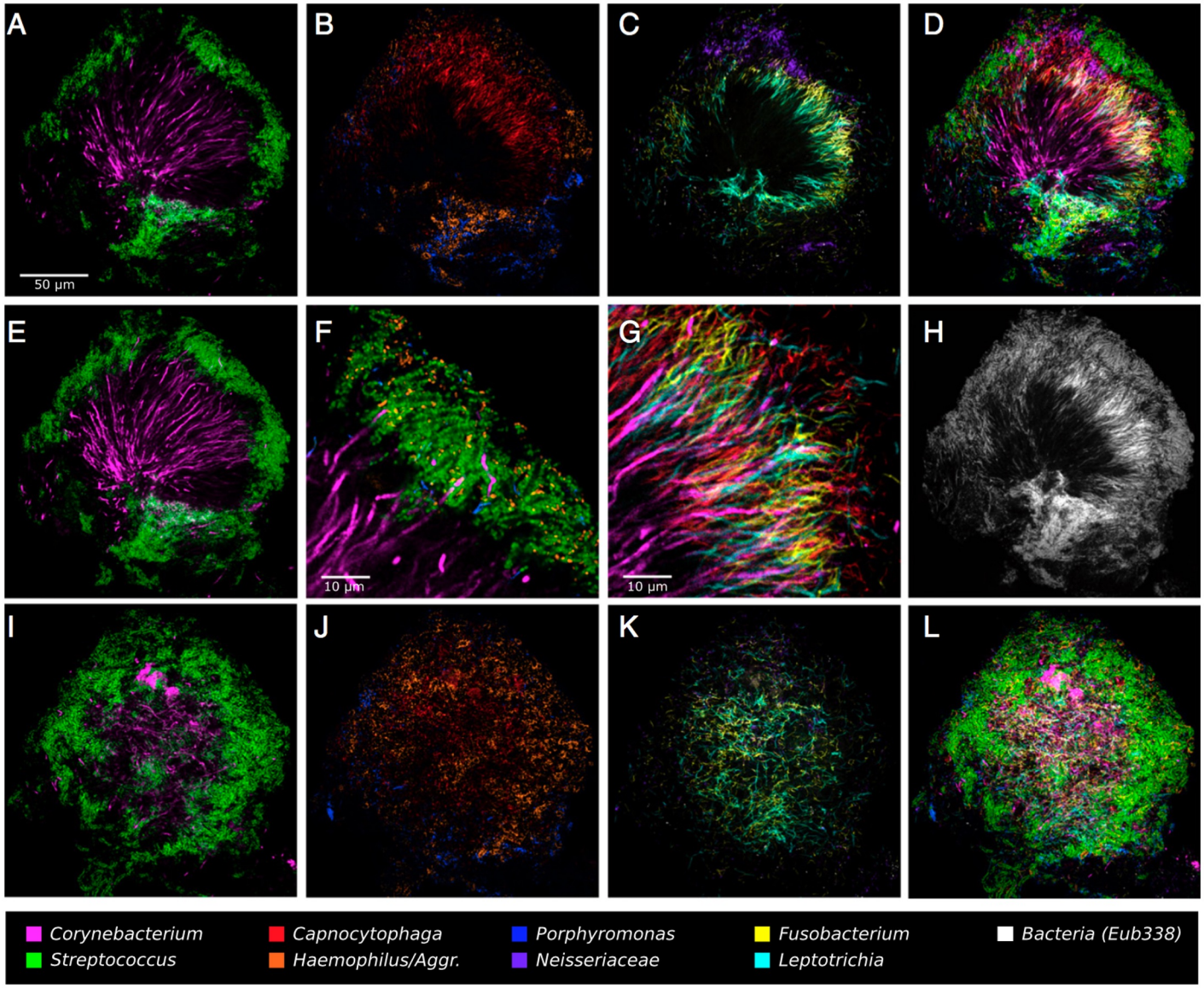
Base of "hedgehog structure" formed by Corynebacterium.

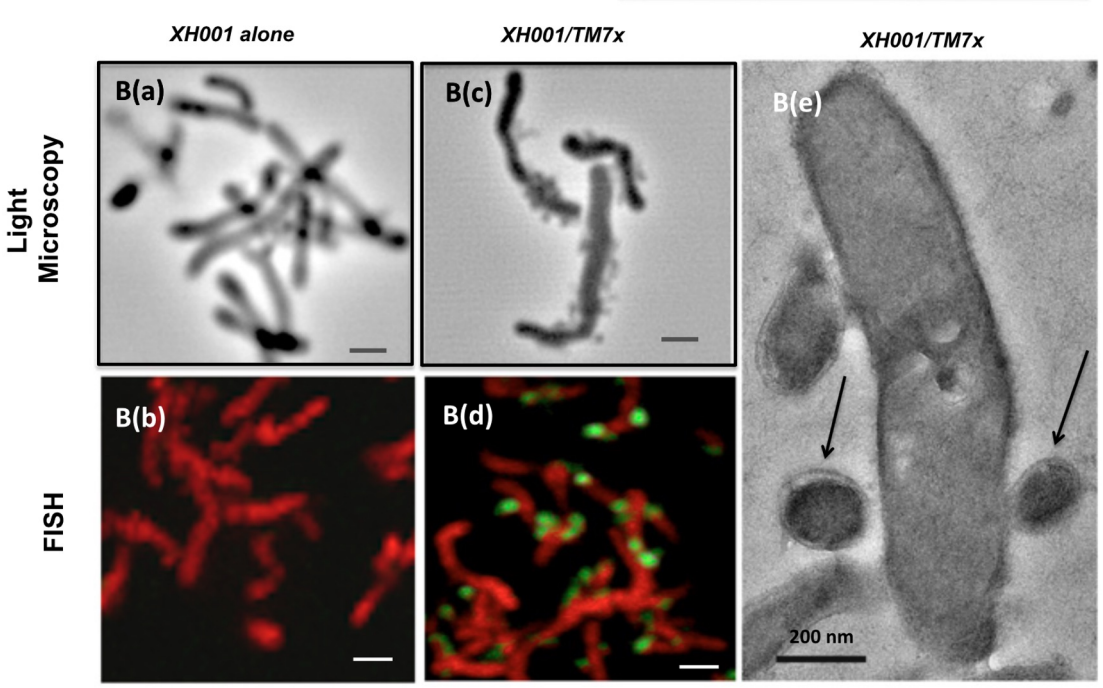
Streptococcus forms "corncob" near tip.

Haemophilus located in a third layer, growing on Streptococcus.

Mark Welch JL. *PNAS* E791 (2016).


10 µm
Lautropia
Streptococcus
Veillonella
Haemophilus/Aggregatibacter
Prevotella
Rothia
Capnocytophaga


A
B
C
D
50 µm
E
F
G
H
10 µm
10 µm
I
J
K
L
Corynebacterium
Streptococcus
Capnocytophaga
Haemophilus/Aggr.
Porphyromonas
Neisseriaceae
Fusobacterium
Leptotrichia
Bacteria (Eub338)



He *PNAS* **112**, 244 (2015).

TM7 first observed in DNA-based studies, 18 years later in culture

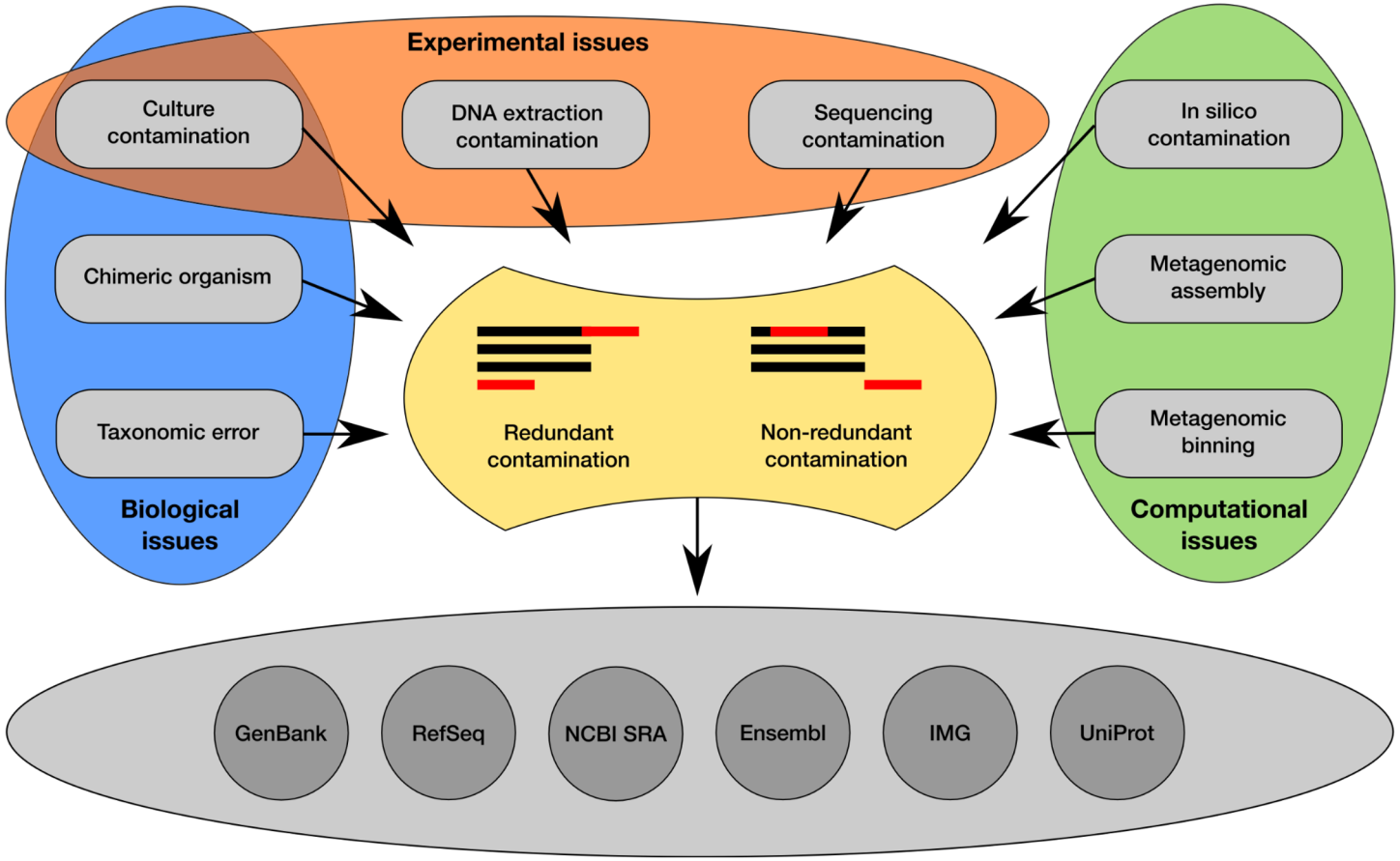
Candidate division TM7 was established using 16S sequences recovered from a peat sample in 1996. (Rheims H. *J Ind Microbiol.* **17**, 159)

TM7 bacteria not cultured until 2014. (Soro V. *Appl Environ Microbiol.* 80, 6480)

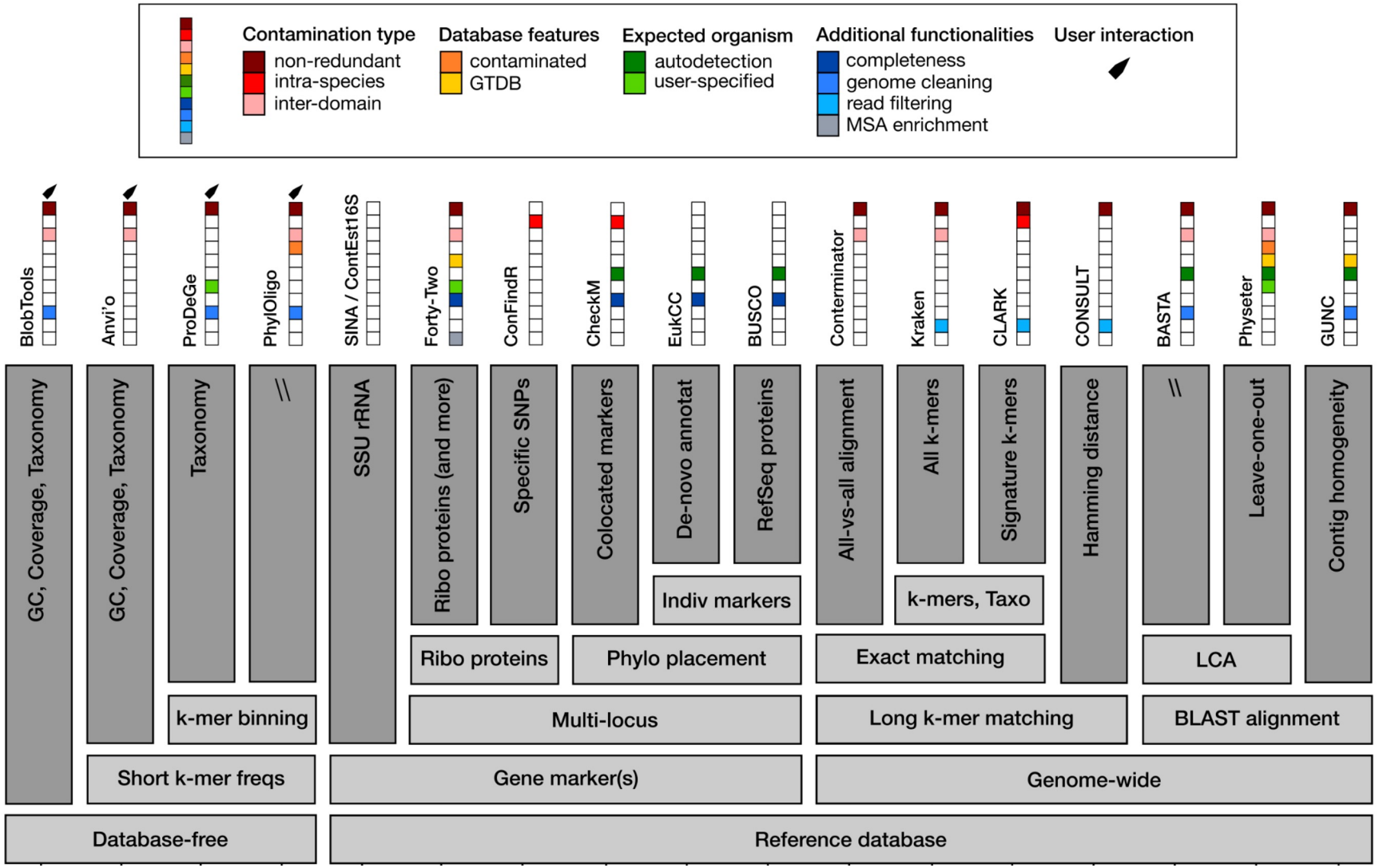
TM7x, pictured, is an obligate epibiont of Actinomyces.

Complete genome: 705kb

1 Contamination detection in bacterial genomes



Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. *Genome Biol.* 2022 Feb 21;23(1):60.


Contamination type
non-redundant
intra-species
inter-domain
Database features
contaminated
GTDB
Expected organism
autodetection
user-specified
Additional functionalities
completeness
genome cleaning
read filtering
MSA enrichment
User interaction
BlobTools
Anvi'o
ProDeGe
PhylOligo
SINA / ContEst16S
Forty-Two
ConFindR
CheckM
EukCC
BUSCO
Conterminator
Kraken
CLARK
CONSULT
BASTA
Physeter
GUNC
GC, Coverage, Taxonomy
GC, Coverage, Taxonomy
Taxonomy
\\
SSU rRNA
Ribo proteins (and more)
Specific SNPs
Colocated markers
De-novo annotat
RefSeq proteins
All-vs-all alignment
All k-mers
Signature k-mers
Hamming distance
\\
Leave-one-out
Contig homogeneity
Indiv markers
k-mers, Taxo
Ribo proteins
Phylo placement
Exact matching
LCA
k-mer binning
Multi-locus
Long k-mer matching
BLAST alignment
Short k-mer freqs
Gene marker(s)
Genome-wide
Database-free
Reference database

2

Genomes as metagenomes

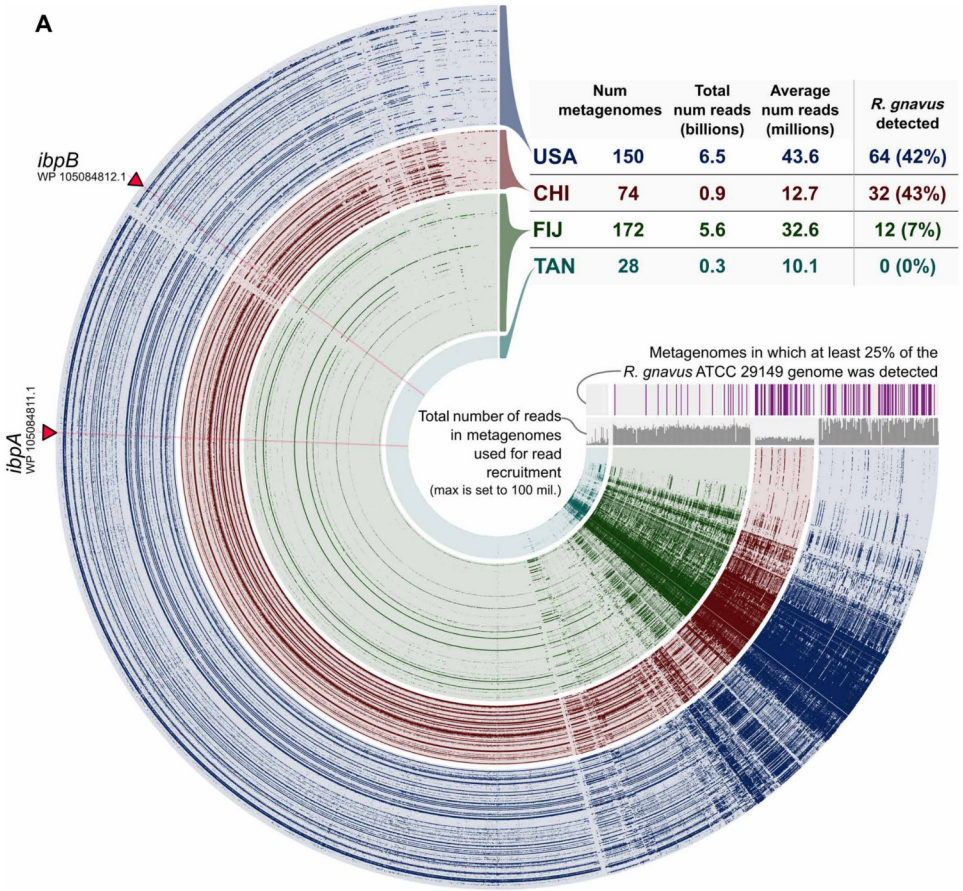
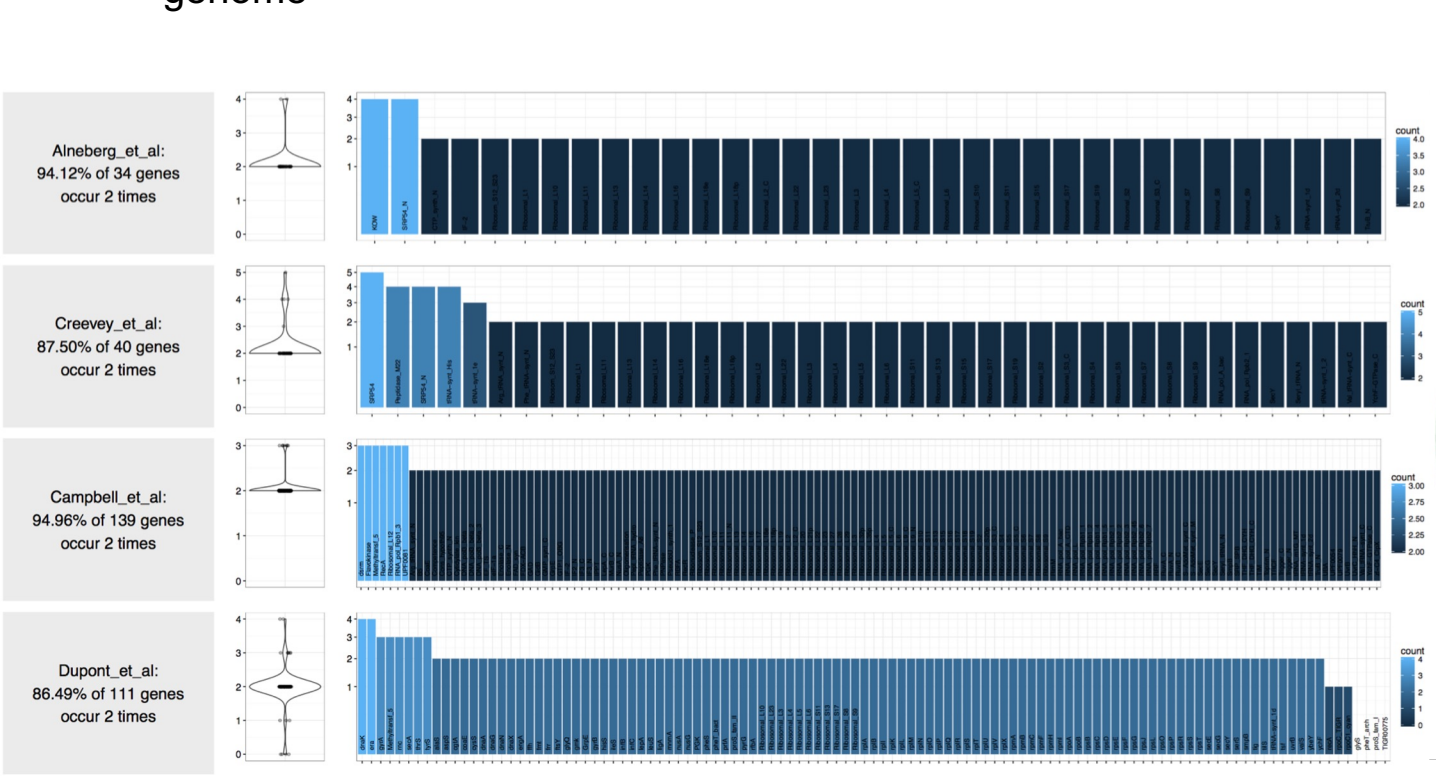
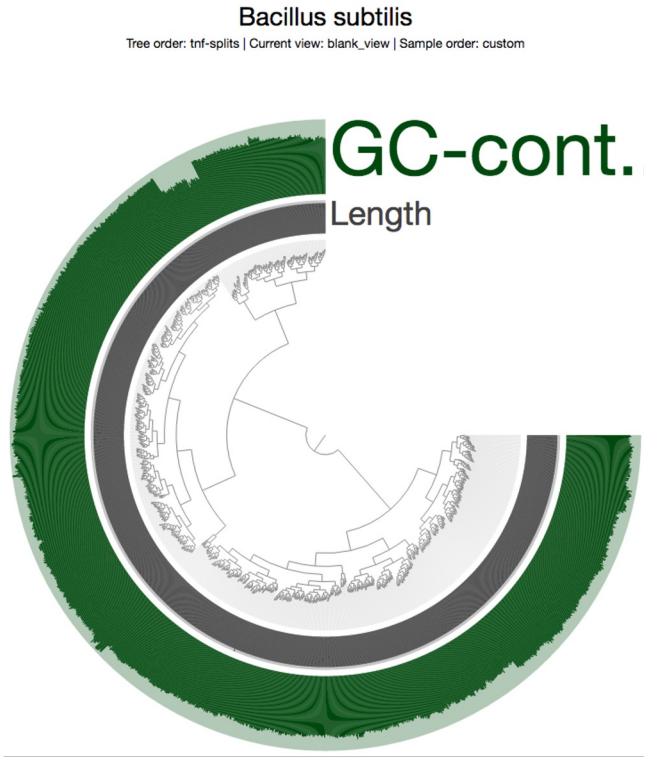


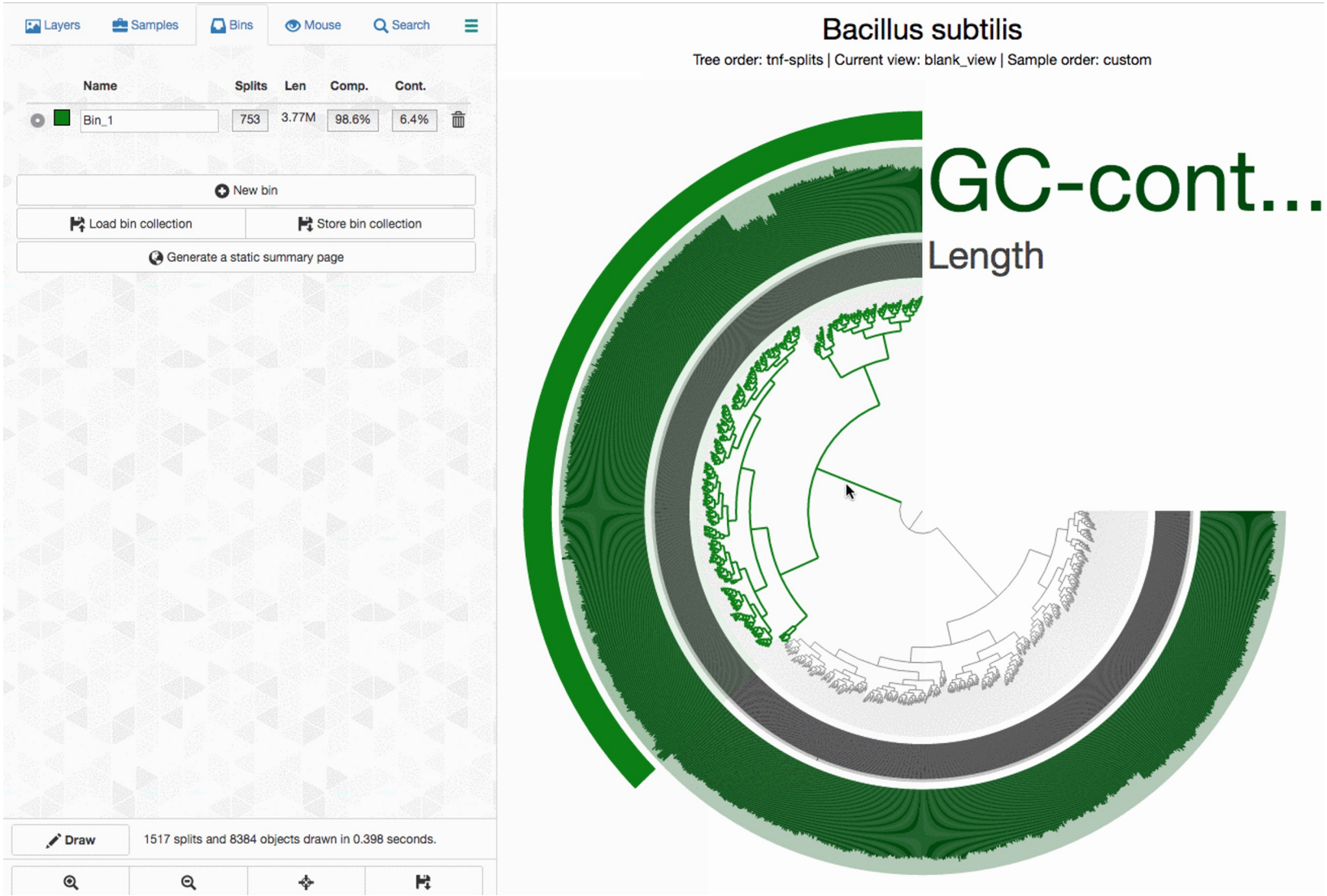
Fig. 5. Distribution of *R. gnavus* and its superantigens across human metagenomes. Dendrogram alignment of the *R. gnavus* ATCC 29149 genome to 424 human metagenomes (data files S3 and S4). Each spoke represents one gene in the *R. gnavus* genome, and each layer represents an individual human metagenome. The two superantigen genes are labeled. Intensity represents coverage of the open reading frame in the metagenome.

Eren AM, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol.* 2021 Jan;6(1):3-6.

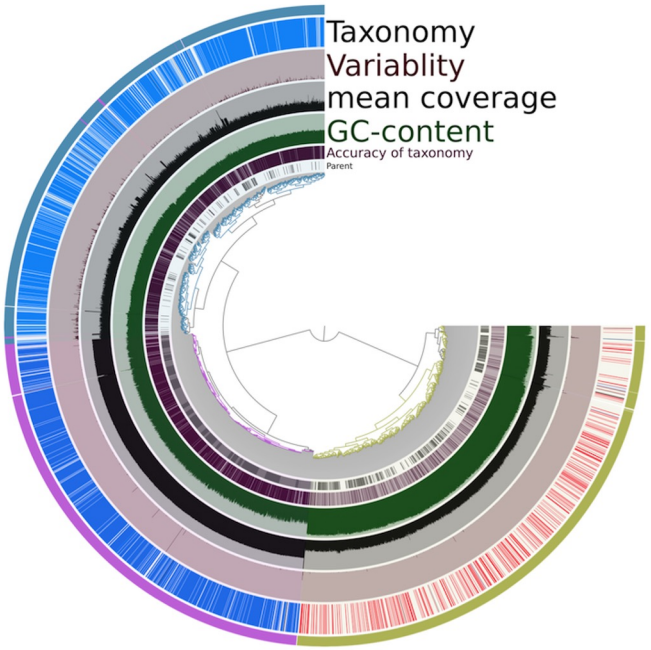
Example Combining two genomes in one cell: Stable cloning of the Synechocystis PCC6803 genome in the Bacillus subtilis 168 genome



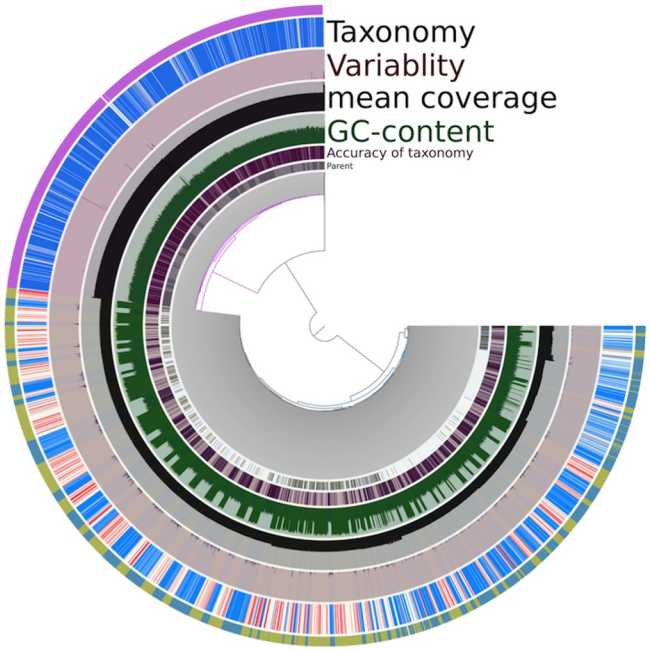



Layers
Samples
Bins
Mouse
Search
Name
Splits
Len
Comp.
Cont.
Bin_1
753
3.77M
98.6%
6.4%
New bin
Load bin collection
Store bin collection
Generate a static summary page
Draw
1517 splits and 8384 objects drawn in 0.398 seconds.
Bacillus subtilis
Tree order: tnf-splits | Current view: blank_view | Sample order: custom
GC-cont...
Length

Clustering based on sequence composition



Clustering based on sequence composition and abundance



Bin	Taxonomy	Total Size	Num Contigs	N50	GC Content	Compl.	Contam.
selection_2	*Bacillus subtilis*	3.70 Mb	54	159,263	41.56%	98.47%	8.02%
selection_3	*Unknown*	4.20 Mb	1,394	3,773	69.34%	78.37%	4.19%
selection_1	*Bacillus anthracis*	3.29 Mb	1,794	1,858	34.99%	54.45%	10.60%

NCBI Taxonomy Browser

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | BioCollections

Search for [ ] as [complete name] [x] lock [Go] [Clear]

Display [3] levels using filter: [none]

[Ruminococcus] gnavus [1)]

Taxonomy ID: 33038 (for references in articles please use NCBI:txid33038)

current name

***Ruminococcus gnavus* Moore et al. 1976 (Approved Lists 1980)** Moore et al. 1976 in [Skerman VBD et al. (1980)]

type strain of *Ruminococcus gnavus* Moore et al. 1976 (Approved Lists 1980): **ATCC:29149**, **VPI C7-9**, **JCM:6515**

homotypic synonym:

"**Mediterraneibacter gnavus**" (Moore et al. 1976) Togo et al. 2018, effective name [2)]

NCBI BLAST name: **firmicutes**

Rank: **species**

Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)

Lineage(*full*)

cellular organisms; Bacteria; Terrabacteria group; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Mediterraneibacter

Entrez records

Database name	Subtree links	Direct links	Links from type
Nucleotide	19,448	19,207	129
Protein	197,843	177,578	-
Structure	48	15	-
Genome	1	1	-
Popset	4	4	-
GEO Datasets	12	9	-
PubMed Central	10	4	-
Gene	7,503	1	-
SRA Experiments	144	122	-
Protein Clusters	2,504	2,504	-
Identical Protein Groups	78,137	76,776	-
BioProject	53	42	-
BioSample	1,408	1,382	11
Assembly	165	158	7
PubChem BioAssay	7	7	-
Taxonomy	6	1	-

[Ruminococcus] gnavus ATCC 29149

[ruminococcus] gnavus atcc 29149 is a strain of [ruminococcus] gnavus.

Browse taxonomy

Current scientific name	[Ruminococcus] gnavus ATCC 29149
Taxonomic rank	strain
NCBI Taxonomy ID	411470

For more details see NCBI Taxonomy
View the legacy Genome page

Genome

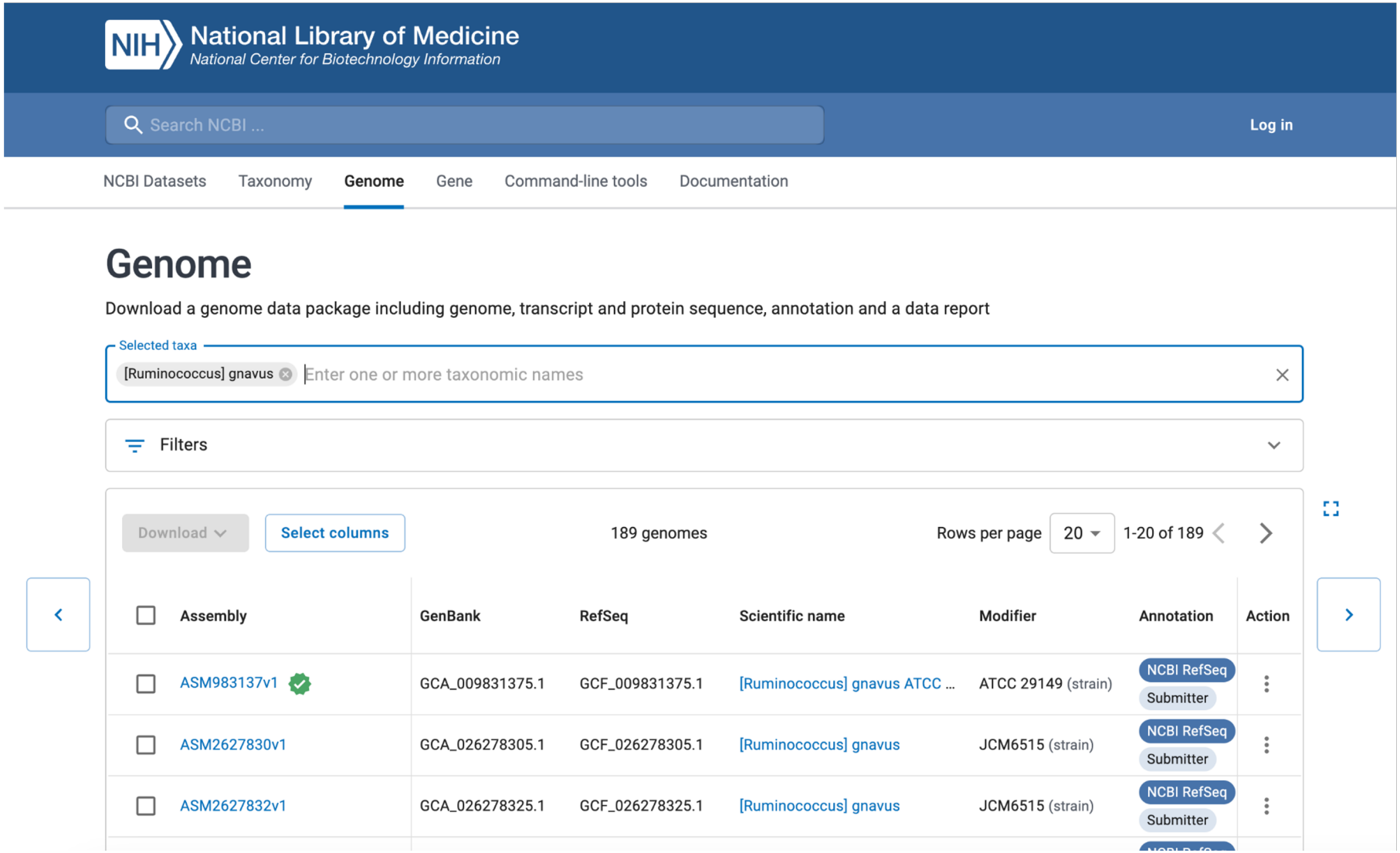
Browse all 7 genomes

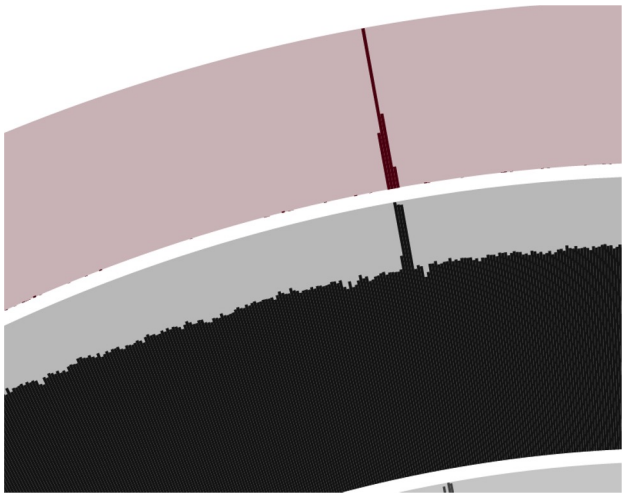
Reference genome

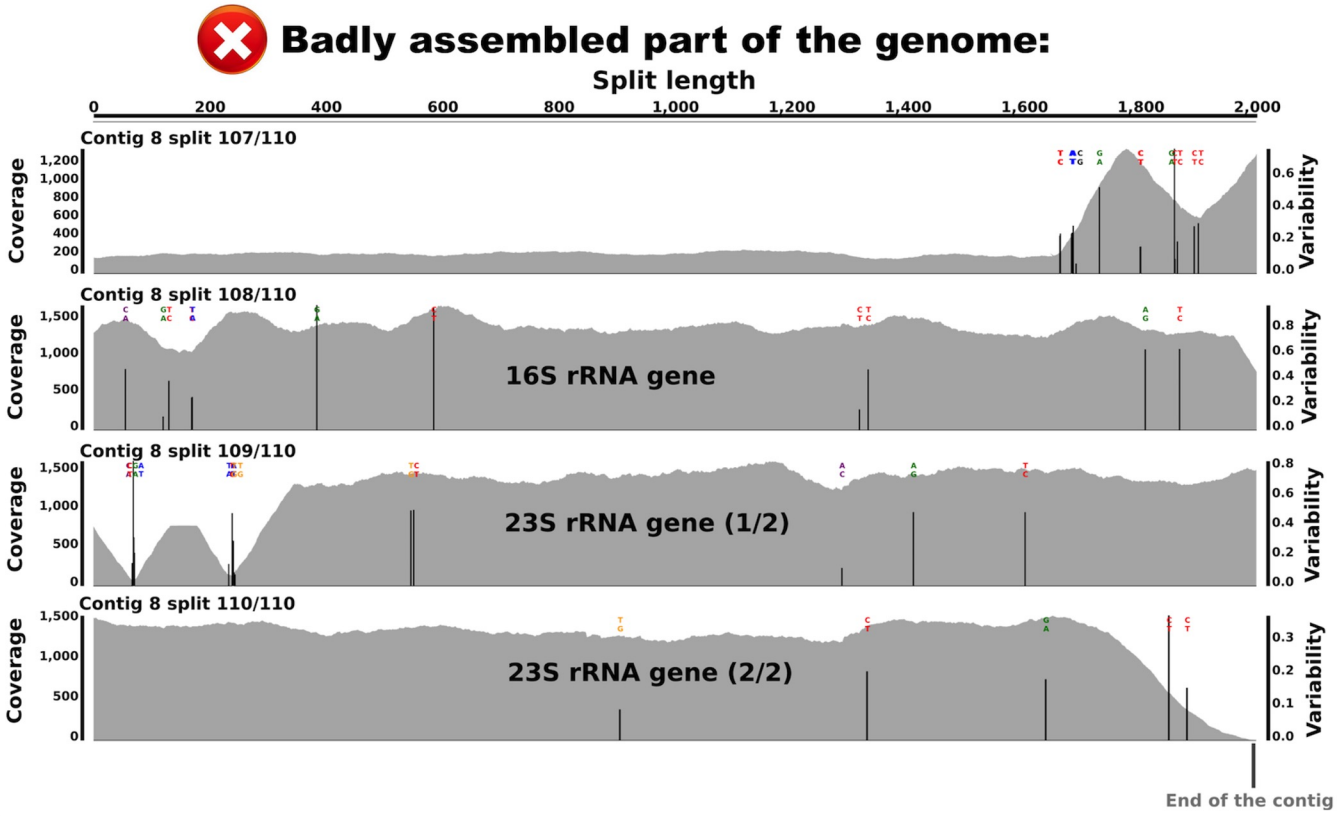
ASM983137v1

Kyungpook National University (2020). Strain: ATCC 29149.

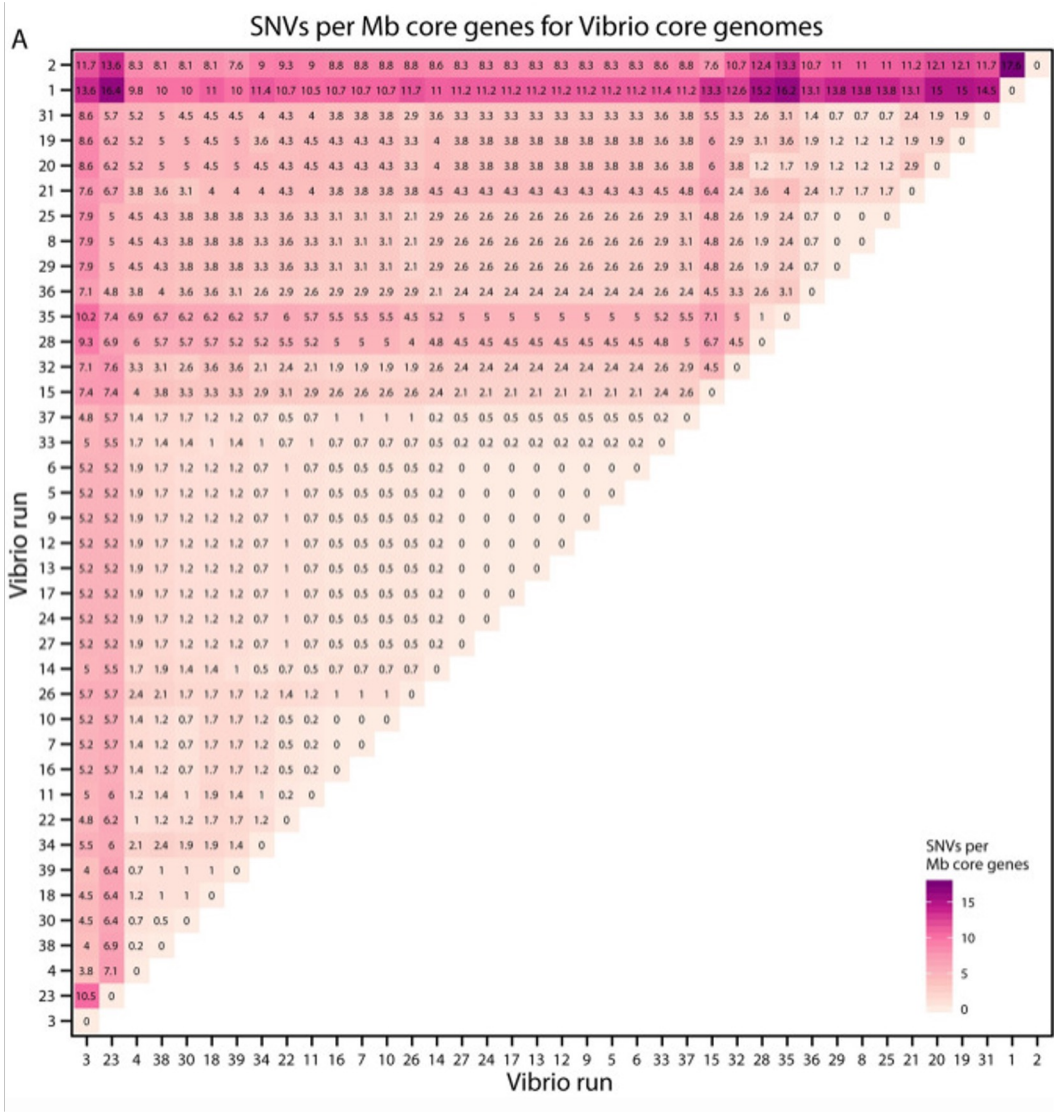
RefSeq GCF_009831375.1


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome
Download a genome data package including genome, transcript and protein sequence, annotation and a data report
Selected taxa
[Ruminococcus] gnavus
Enter one or more taxonomic names
Filters
Download
Select columns
189 genomes
Rows per page
20
1-20 of 189
Assembly
GenBank
RefSeq
Scientific name
Modifier
Annotation
Action
ASM983137v1
GCA_009831375.1
GCF_009831375.1
[Ruminococcus] gnavus ATCC ...
ATCC 29149 (strain)
NCBI RefSeq
Submitter
ASM2627830v1
GCA_026278305.1
GCF_026278305.1
[Ruminococcus] gnavus
JCM6515 (strain)
NCBI RefSeq
Submitter
ASM2627832v1
GCA_026278325.1
GCF_026278325.1
[Ruminococcus] gnavus
JCM6515 (strain)
NCBI RefSeq
Submitter



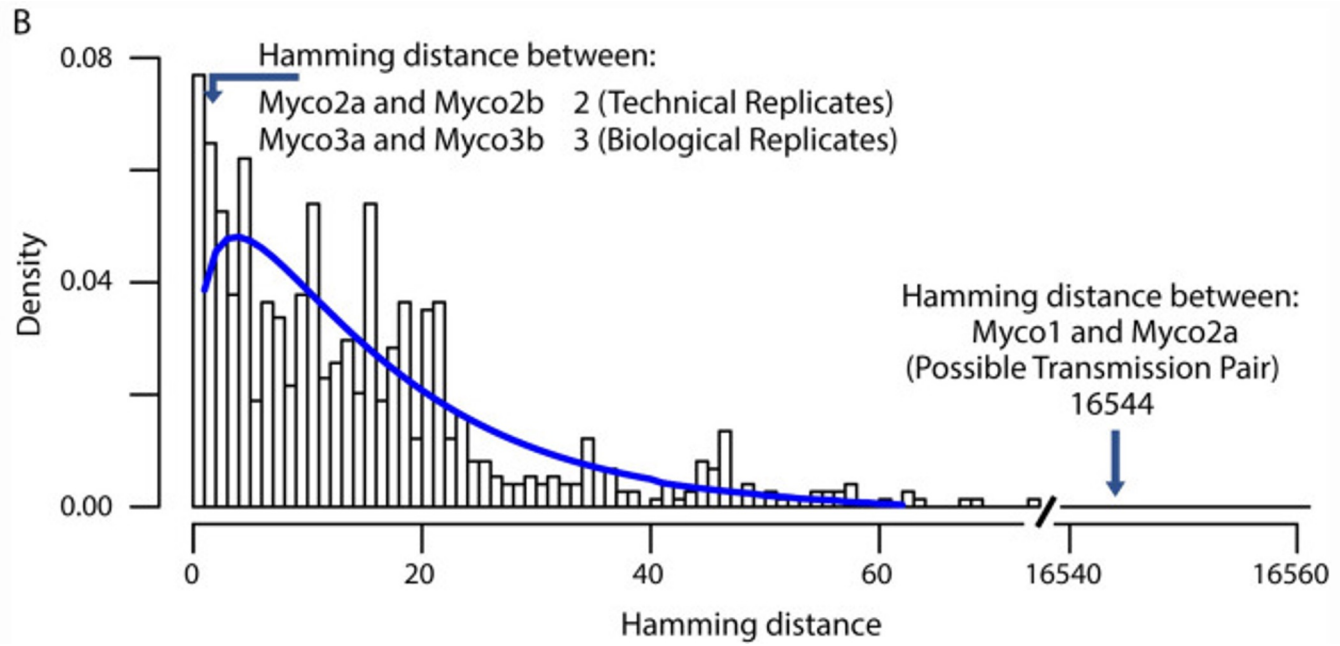

Badly assembled part of the genome:
Split length
0
200
400
600
800
1,000
1,200
1,400
1,600
1,800
2,000
Contig 8 split 107/110
Contig 8 split 108/110
16S rRNA gene
Contig 8 split 109/110
23S rRNA gene (1/2)
Contig 8 split 110/110
23S rRNA gene (2/2)
Coverage
Variability
End of the contig

3 Technical replicates and strain-level differences



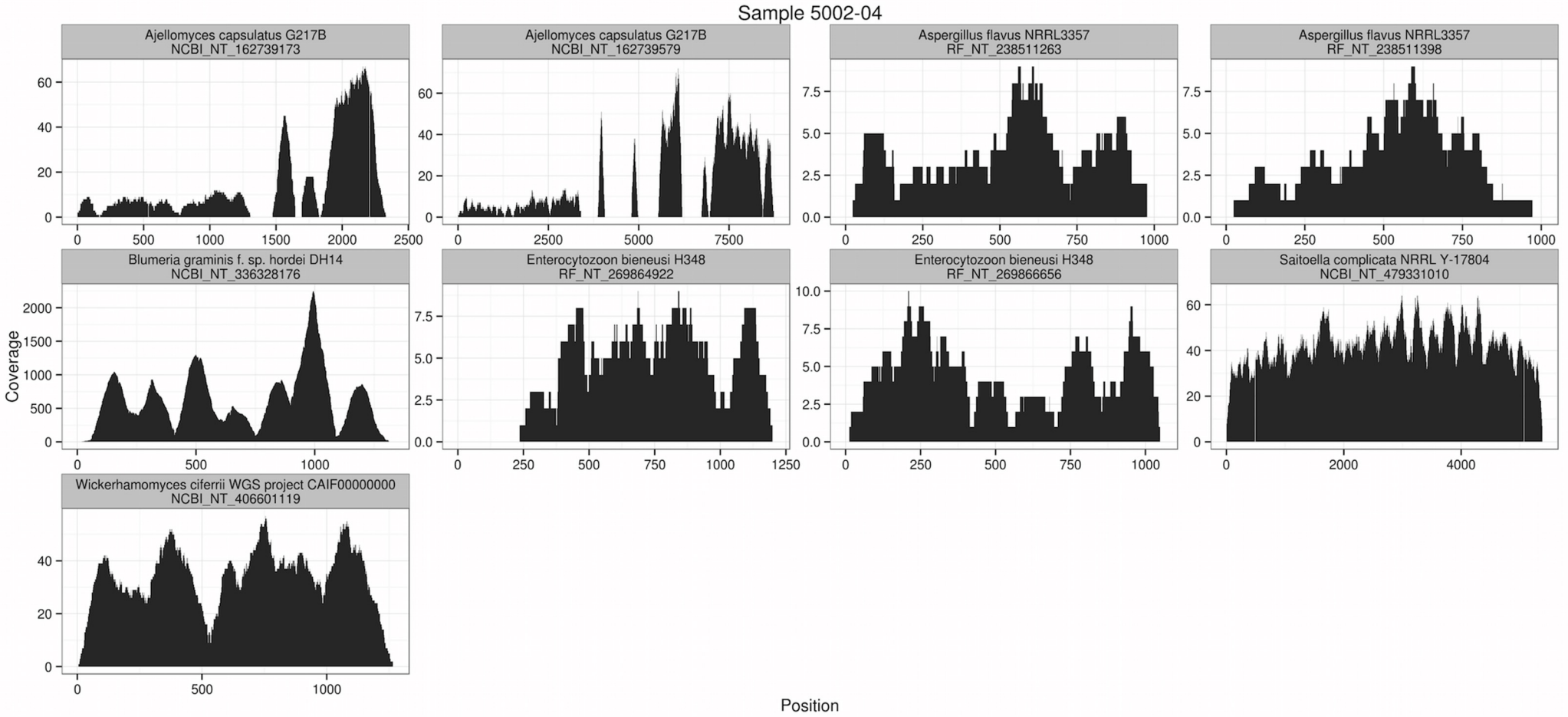
Genome assembly of a single *Vibrio campbellii* genome, 39 technical replicates

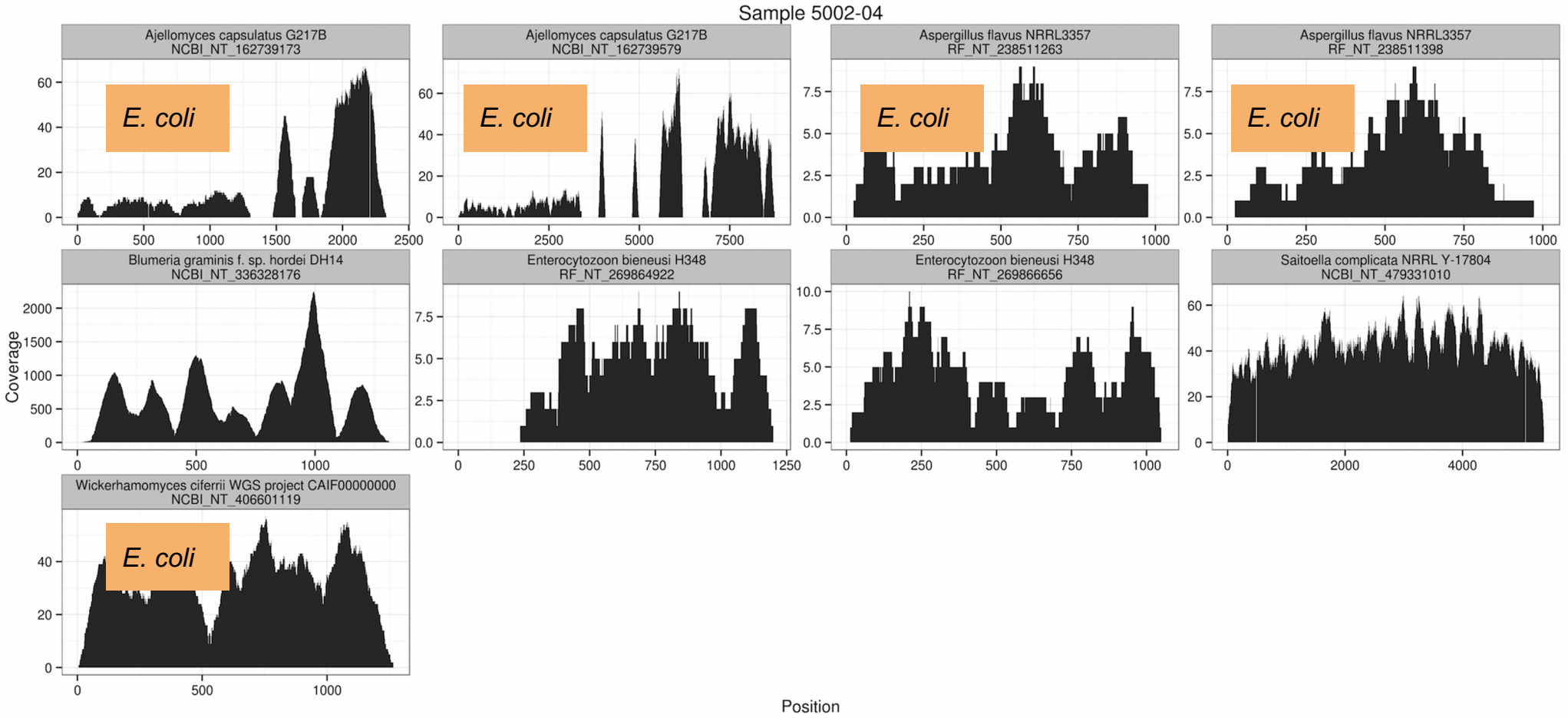
Gu CH, Zhao C, Hofstaedter C, Tebas P, Glaser L, Baldassano R, Bittinger K, Mattei LM, Bushman FD. Investigating hospital Mycobacterium chelonae infection using whole genome sequencing and hybrid assembly. *PLoS One.* 2020 Nov 9;15(11):e0236533.

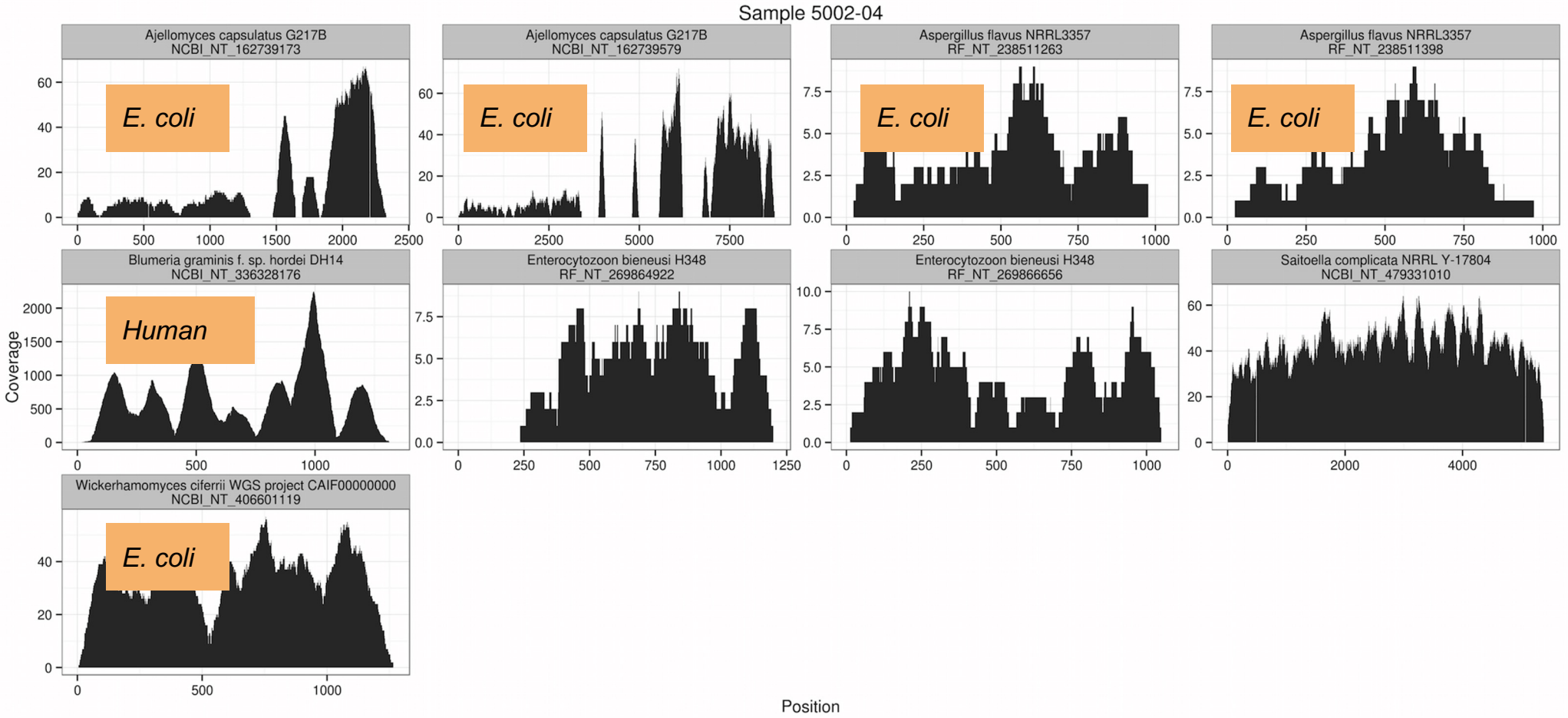

B
Hamming distance between:
Myco2a and Myco2b 2 (Technical Replicates)
Myco3a and Myco3b 3 (Biological Replicates)
Hamming distance between:
Myco1 and Myco2a
(Possible Transmission Pair)
16544
Density
0.08
0.04
0.00
0
20
40
60
16540
16560
Hamming distance

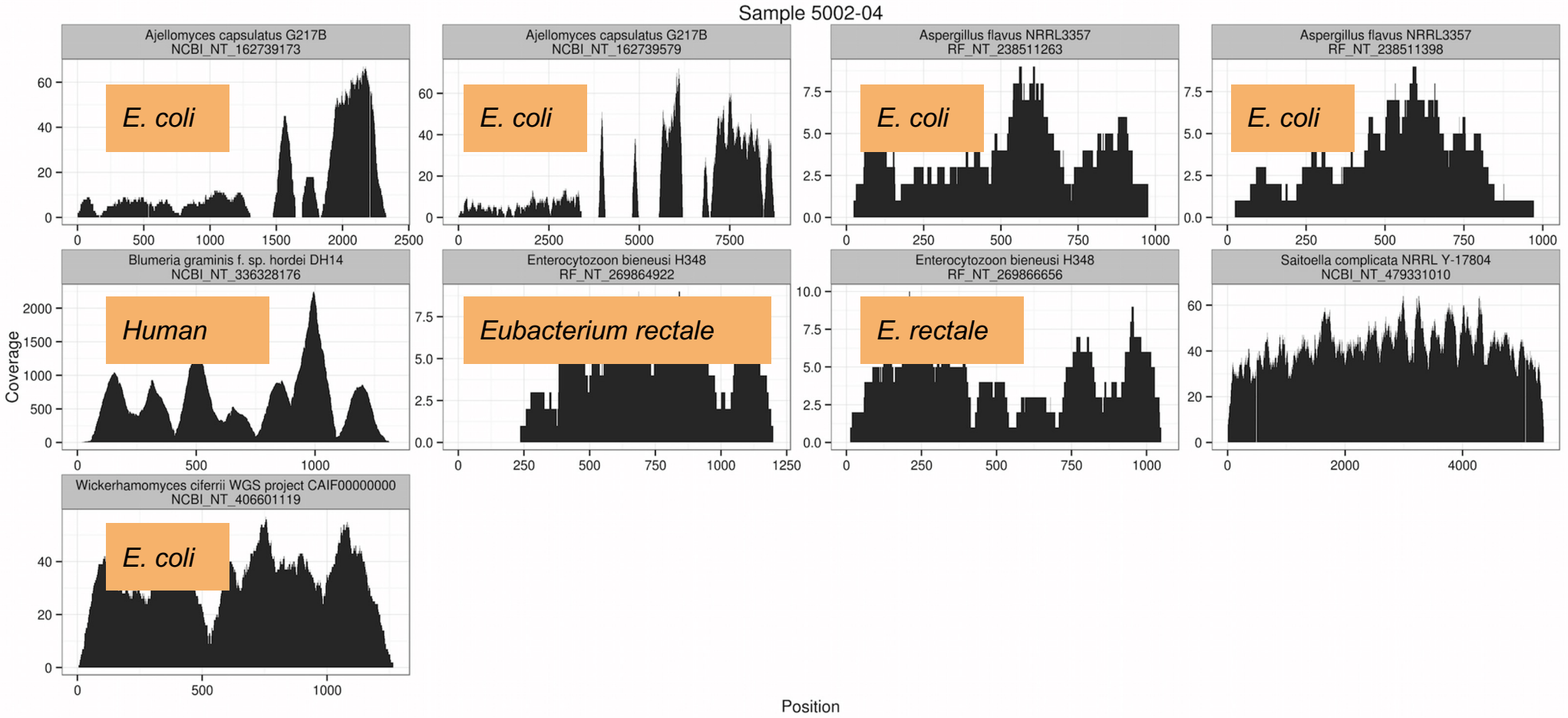
4

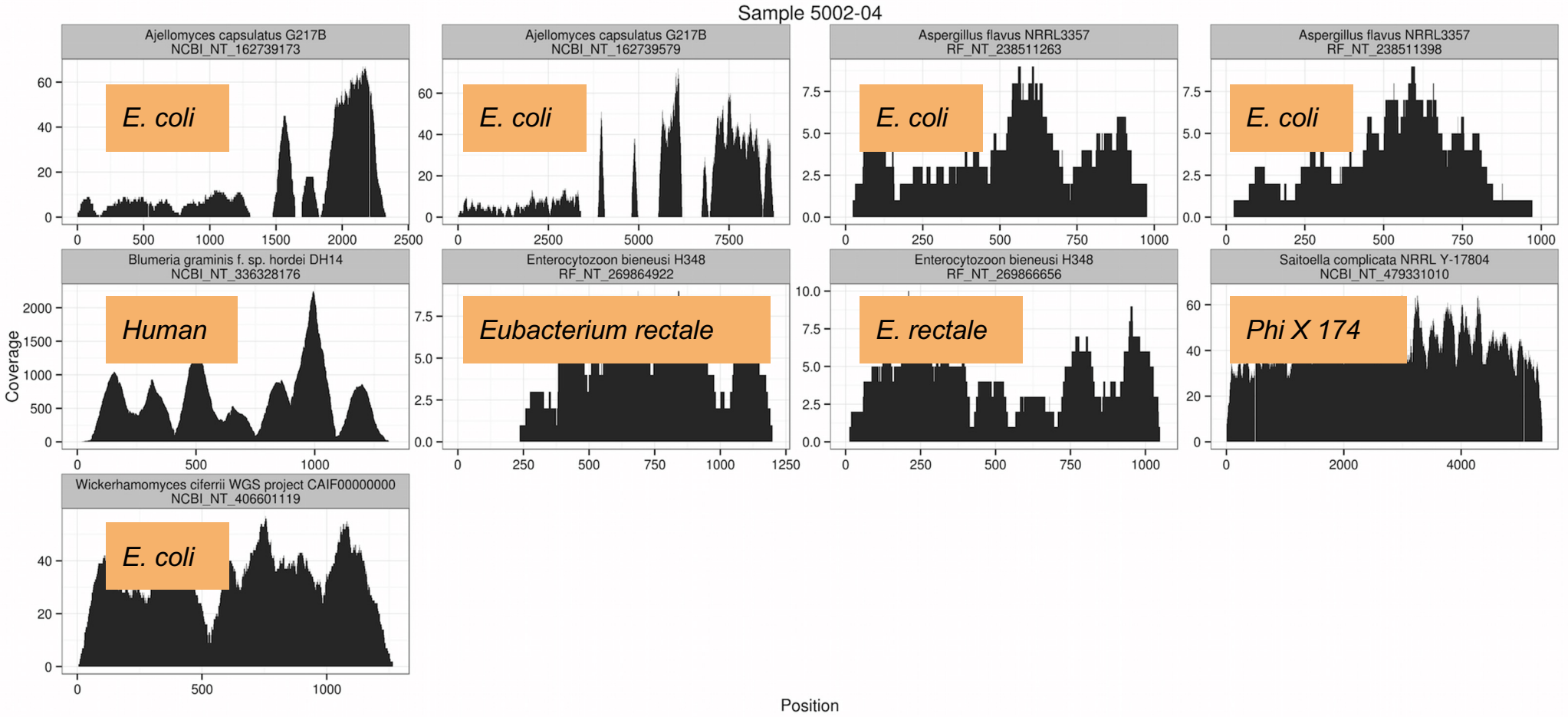
Database contamination, human DNA contamination


Sample 5002-04
Ajellomyces capsulatus G217B
NCBI_NT_162739173
Ajellomyces capsulatus G217B
NCBI_NT_162739579
Aspergillus flavus NRRL3357
RF_NT_238511263
Aspergillus flavus NRRL3357
RF_NT_238511398
Blumeria graminis f. sp. hordei DH14
NCBI_NT_336328176
Enterocytozoon bieneusi H348
RF_NT_269864922
Enterocytozoon bieneusi H348
RF_NT_269866656
Saitoella complicata NRRL Y-17804
NCBI_NT_479331010
Wickerhamomyces ciferrii WGS project CAIF00000000
NCBI_NT_406601119
Coverage
Position


Sample 5002-04
Ajellomyces capsulatus G217B
NCBI_NT_162739173
E. coli
Ajellomyces capsulatus G217B
NCBI_NT_162739579
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511263
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511398
E. coli
Blumeria graminis f. sp. hordei DH14
NCBI_NT_336328176
Enterocytozoon bieneusi H348
RF_NT_269864922
Enterocytozoon bieneusi H348
RF_NT_269866656
Saitoella complicata NRRL Y-17804
NCBI_NT_479331010
Wickerhamomyces ciferrii WGS project CAIF00000000
NCBI_NT_406601119
E. coli
Coverage
Position


Sample 5002-04
Ajellomyces capsulatus G217B
NCBI_NT_162739173
E. coli
Ajellomyces capsulatus G217B
NCBI_NT_162739579
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511263
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511398
E. coli
Blumeria graminis f. sp. hordei DH14
NCBI_NT_336328176
Human
Enterocytozoon bieneusi H348
RF_NT_269864922
Enterocytozoon bieneusi H348
RF_NT_269866656
Saitoella complicata NRRL Y-17804
NCBI_NT_479331010
Wickerhamomyces ciferrii WGS project CAIF00000000
NCBI_NT_406601119
E. coli
Coverage
Position

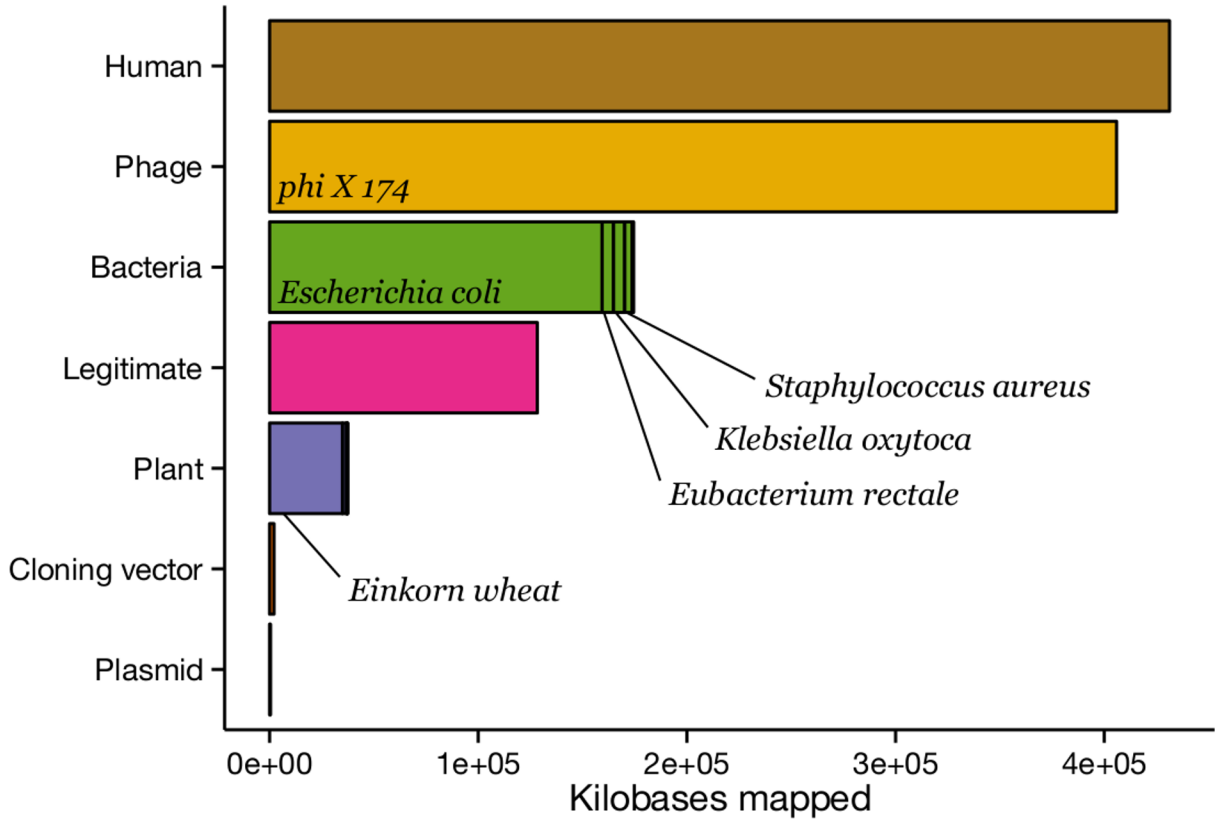

Sample 5002-04
Ajellomyces capsulatus G217B
NCBI_NT_162739173
E. coli
Ajellomyces capsulatus G217B
NCBI_NT_162739579
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511263
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511398
E. coli
Blumeria graminis f. sp. hordei DH14
NCBI_NT_336328176
Human
Enterocytozoon bieneusi H348
RF_NT_269864922
Eubacterium rectale
Enterocytozoon bieneusi H348
RF_NT_269866656
E. rectale
Saitoella complicata NRRL Y-17804
NCBI_NT_479331010
Wickerhamomyces ciferrii WGS project CAIF00000000
NCBI_NT_406601119
E. coli
Coverage
Position


Sample 5002-04
Ajellomyces capsulatus G217B
NCBI_NT_162739173
E. coli
Ajellomyces capsulatus G217B
NCBI_NT_162739579
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511263
E. coli
Aspergillus flavus NRRL3357
RF_NT_238511398
E. coli
Blumeria graminis f. sp. hordei DH14
NCBI_NT_336328176
Human
Enterocytozoon bieneusi H348
RF_NT_269864922
Eubacterium rectale
Enterocytozoon bieneusi H348
RF_NT_269866656
E. rectale
Saitoella complicata NRRL Y-17804
NCBI_NT_479331010
Phi X 174
Wickerhamomyces ciferrii WGS project CAIF00000000
NCBI_NT_406601119
E. coli
Coverage
Position

Sources of misattribution in fungal genomes detected in study of pediatric Crohn's Disease

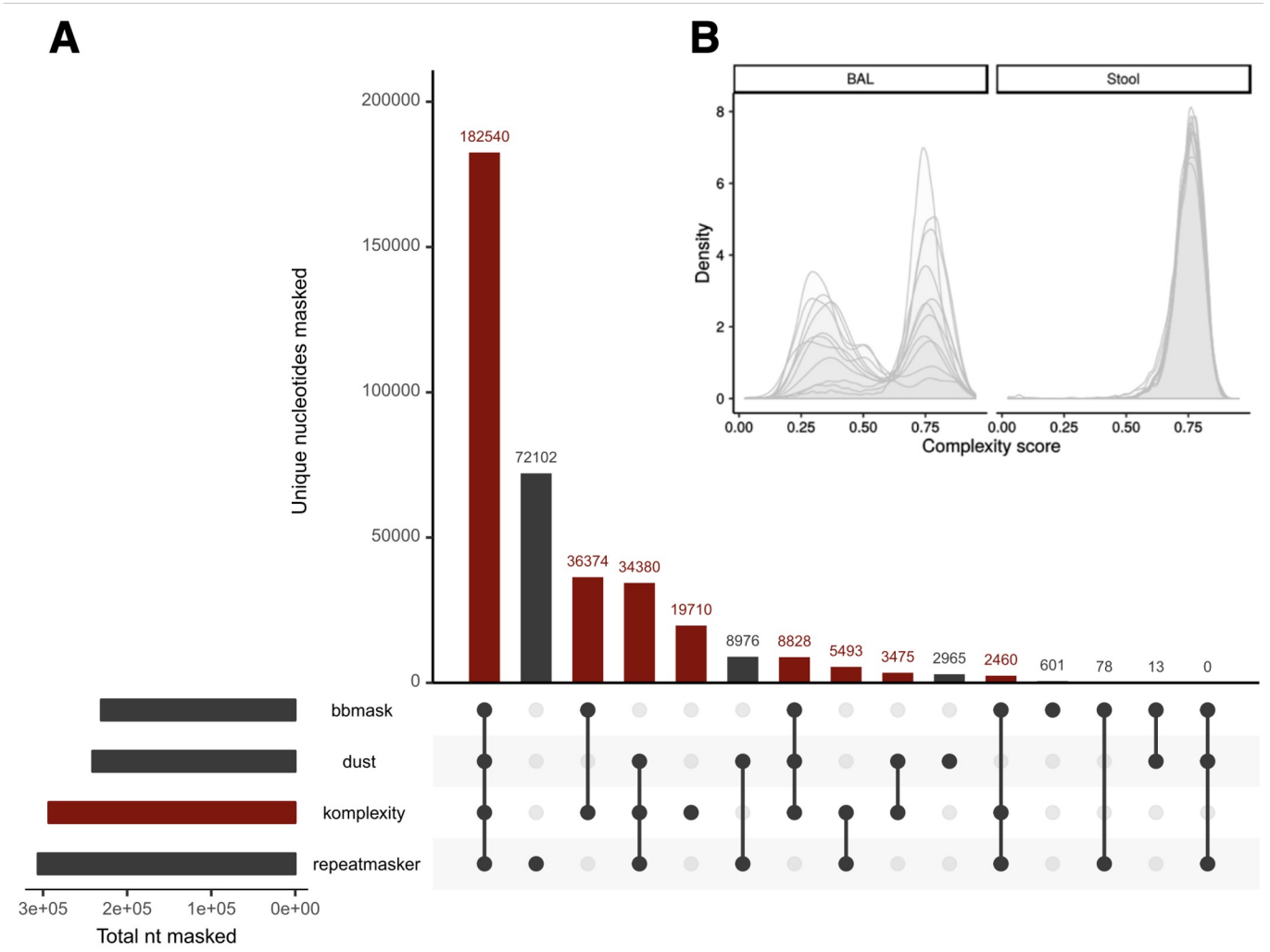
Total number of kilobases mapped to fungal genomes from NCBI.

Potential sources of misidentification evaluated by BLAST search to nt database followed by manual inspection.

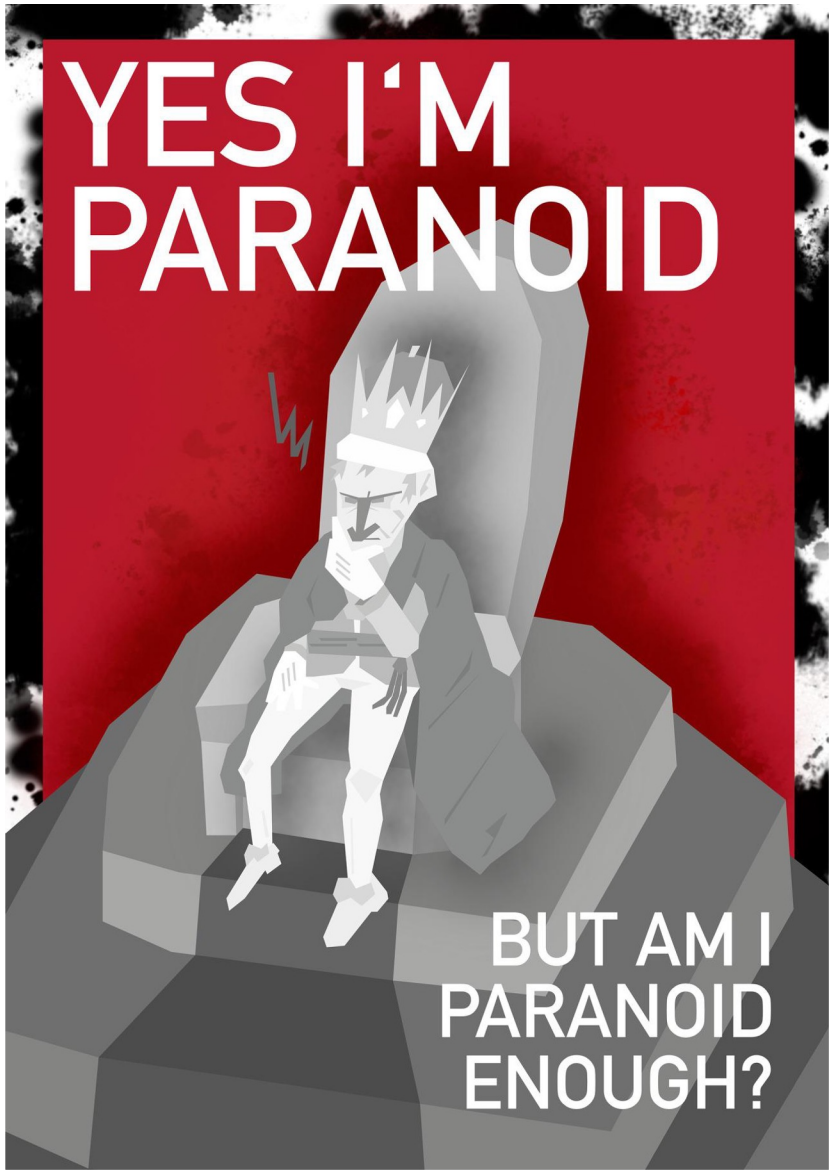


Eukaryotic genomes contain **low-complexity DNA**, which is difficult to detect by sequence alignment.

DNA complexity can be quantified in order to remove low-complexity reads.



Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, Bushman FD, Bittinger K. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome.* 2019 Mar 22;7(1):46.


YES I'M PARANOID
BUT AM I PARANOID ENOUGH?