Lecture 22. Metagenomics

Matthew Nguyen & Michael Schatz

Nov. 11, 2024

JHU 600.649: Applied Comparative Genomics

Preliminary Report Due Tonight

Preliminary Project Report

Assignment Date: October 28, 2024
Due Date: Monday, November 11, 2024 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday November 11

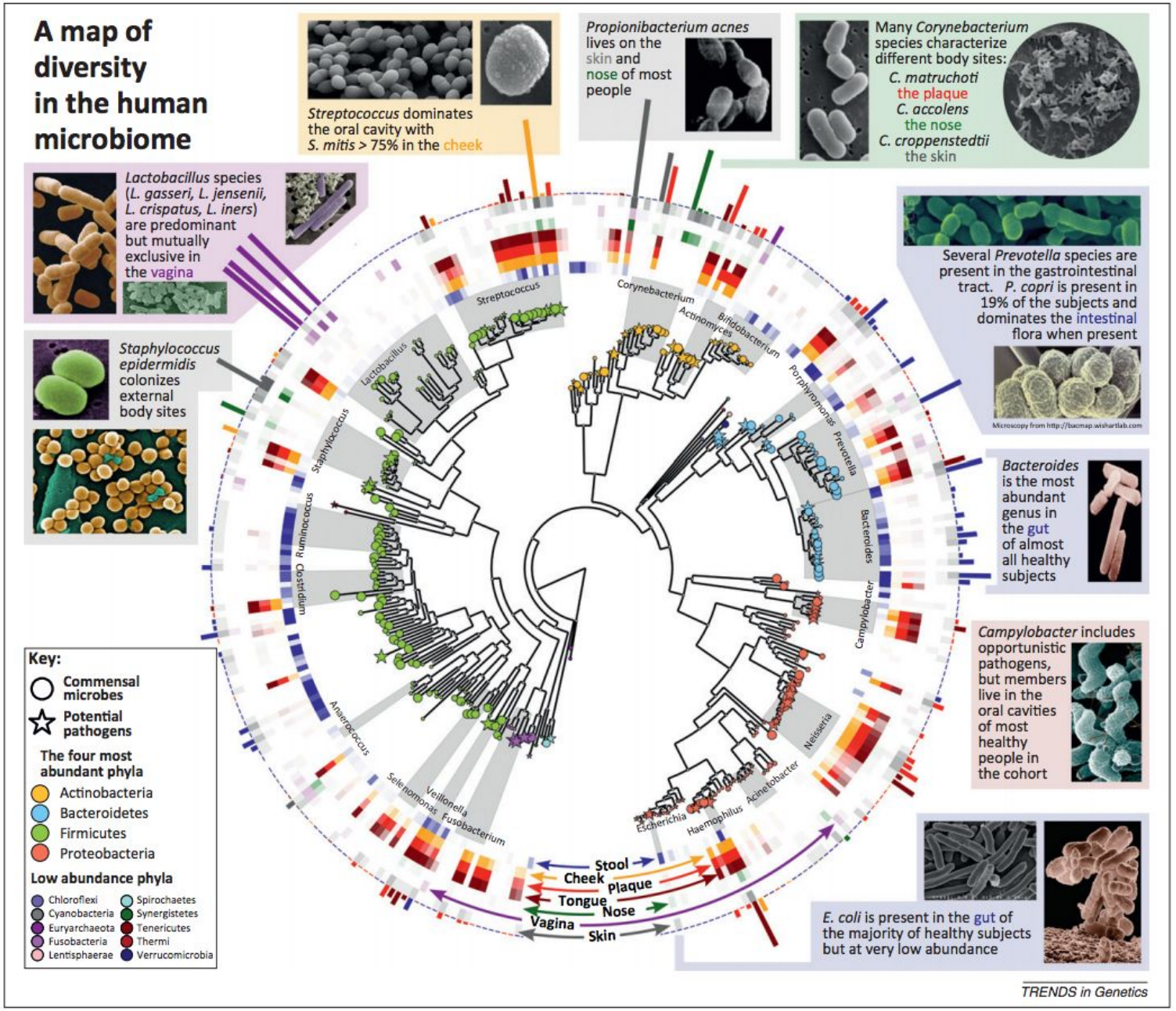
The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result (typically a summary of the data you have identified for your project)
- 5+ References to relevant papers and data

The preliminary report must use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

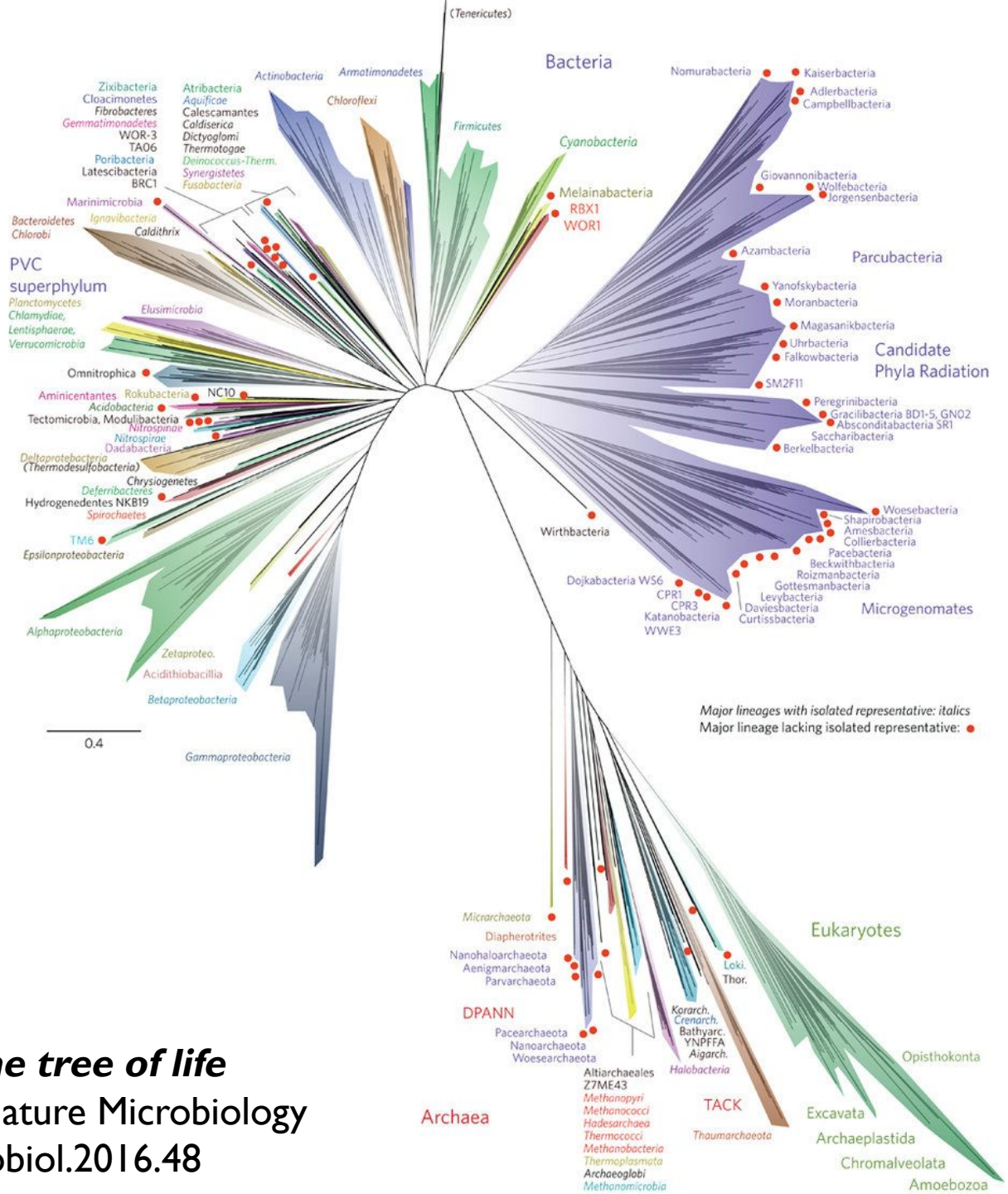
Later, you will present your project in class starting the week of November 25. You will also submit your final written report (6-8 pages) of your project by Dec 16

Please use Piazza if you have any questions!



Biodiversity and functional genomics in the human microbiome

Morgan et al (2013) Trends in Genetics. http://doi.org/10.1016/j.tig.2012.09.005



A new view of the tree of life

Hug et al. (2016) Nature Microbiology

doi:10.1038/nmicrobiol.2016.48

Your second genome?

Human body:
~10 trillion cells

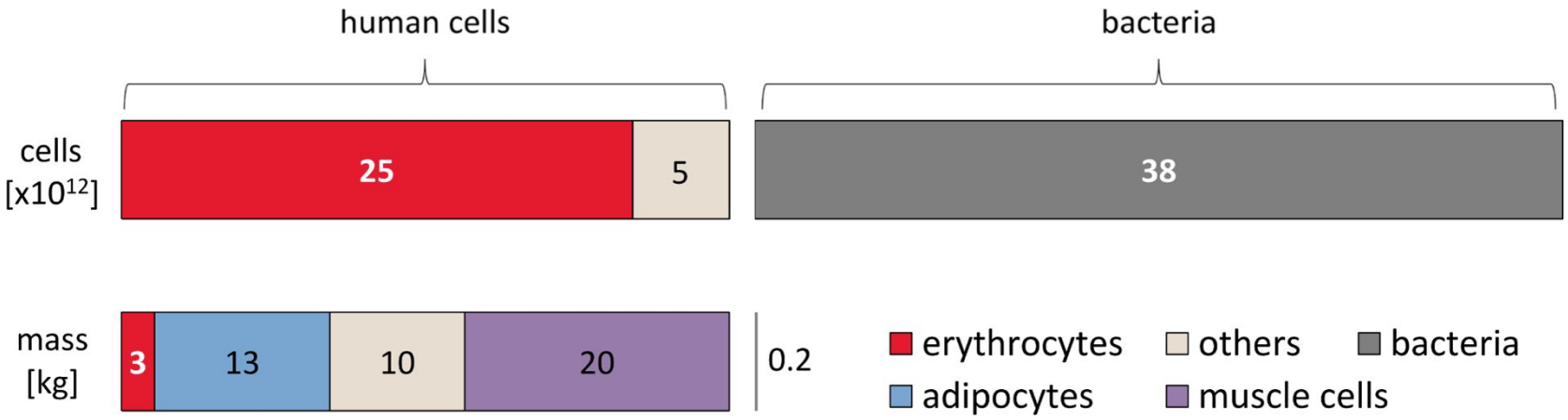
Human brain:
~3.3 lbs

Microbiome
~100 trillion cells

Total mass:
~3.3 lbs

Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans
Sender et al (2016) Cell. http://doi.org/10.1016/j.cell.2016.01.013

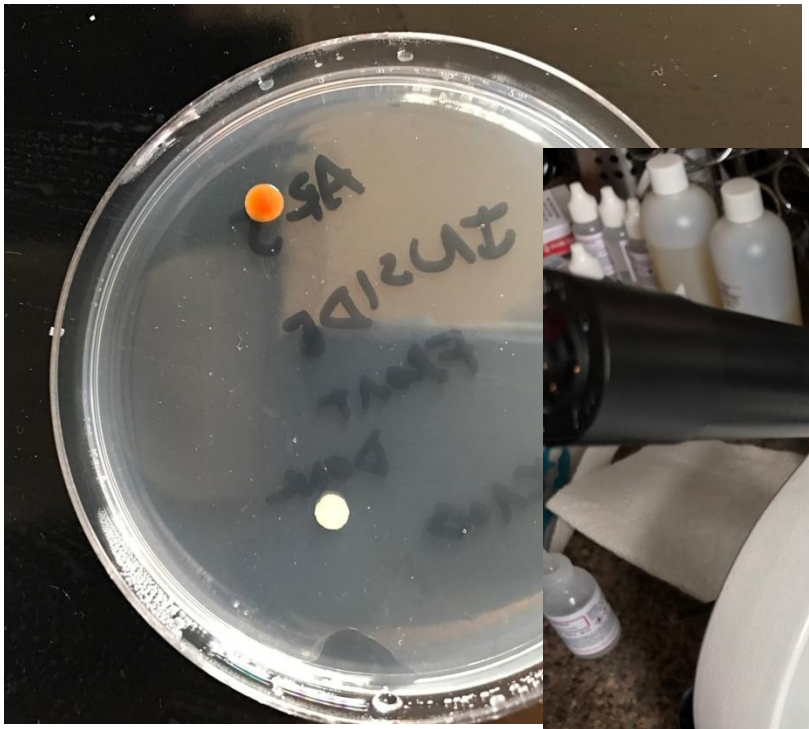
Okay, maybe not 10x more cells but still a lot! ☺

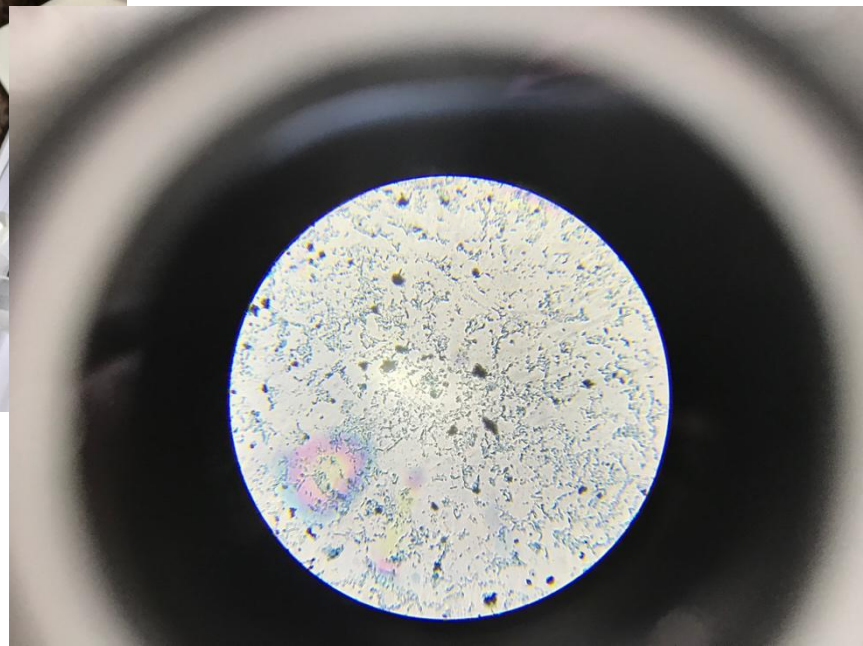


population segment	body weight [kg]	age [y]	blood volume [L]	RBC count [10^{12}/L]	colon content [g]	bac. conc. [10^{11}/ g wet] (1)	total human cells [10^{12}] (2)	total bacteria [10^{12}]	B:H
ref. man	70	20–30	4.9	5.0	420	0.92	30	38	**1.3**
ref. woman	63		3.9	4.5	480	0.92	21	44	**2.2**
young infant	4.4	4 weeks	0.4	3.8	48	0.92	1.9	4.4	**2.3**
infant	9.6	1	0.8	4.5	80	0.92	4	7	**1.7**
elder	70	66	3.8 (3)	4.8	420	0.92	22	38	**1.8**
obese	140		6.7	5.0(4)	610(5)	0.92	40	56	**1.4**

Revised Estimates for the Number of Human and Bacteria Cells in the Body
Sender et al (2016) PLOS Biology. https://doi.org/10.1371/journal.pbio.1002533

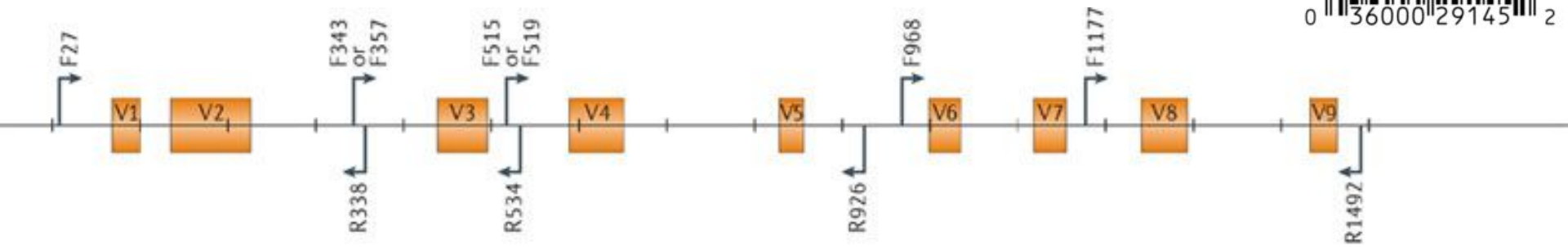
Pre-PCR: Gram-Staining





Gram staining differentiates bacteria by the chemical and physical properties of their cell walls by detecting peptidoglycan, which is present in the cell wall of Gram-positive bacteria

16S rRNA



The 16S rRNA gene is a section of prokaryotic DNA found in all bacteria and archaea. This gene codes for an rRNA, and this rRNA in turn makes up part of the ribosome.

The 16S rRNA gene is a commonly used tool for identifying bacteria for several reasons. First, traditional characterization depended upon phenotypic traits like gram positive or gram negative, bacillus or coccus, etc. Taxonomists today consider analysis of an organism's DNA more reliable than classification based solely on phenotypes. Secondly, researchers may, for a number of reasons, want to identify or classify only the bacteria within a given environmental or medical sample. Thirdly, the 16S rRNA gene is relatively short at 1.5 kb, making it faster and cheaper to sequence than many other unique bacterial genes.

http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-16S.cgi

Box 1 | **Species definitions and concepts in microbiology**

Definitions

Microbes are currently assigned to a common species if their reciprocal, pairwise DNA re-association values are ≥70% in DNA–DNA hybridization experiments under standardized conditions and their ΔT_m (melting temperature) is ≤5°C[79]. In addition, all strains within a species must possess a certain degree of phenotypic consistency, and species descriptions should be based on more than one type strain[11]. A species name is only assigned if its members can be distinguished from other species by at least one diagnostic phenotypic trait[79]. Microbes with 16S ribosomal RNAs (rRNAs) that are ≤98.7% identical are always members of different species, because such strong differences in rRNA correlate with <70% DNA–DNA similarity[80]. However, the opposite is not necessarily true, and distinct species have been occasionally described with 16S rRNAs that are >98.7% identical. Most uncultured microbes cannot be assigned to a classical species because we do not know their phenotype. In some cases, uncultured microbes can be assigned a provisional '*Candidatus*' designation if their 16S rRNA sequences are sufficiently different from those of recognized species, if experimental *in situ* hybridization can be used to specifically detect them and if a basic description of their morphology and biology has been provided[81].

Microbial diversity and the genetic nature of microbial species
Achtman & Wagner (2008) Nature Reviews Microbiology. doi:10.1038/nrmicro1872

Box 1 | **Species definitions and concepts in microbiology**

Definitions

Microbes are currently assigned to a common species if their reciprocal, pairwise DNA re-association values are ≥70% in DNA–DNA hybridization experiments under standardized conditions and their ΔT_m (melting temperature) is ≤5°C[79]. In addition, all strains within a species must possess a certain degree of phenotypic consistency, and species descriptions should be based on more than one type strain[11]. A species name is only assigned
diagnostic ph
≤98.7% ident
differences in
is not necessa
rRNAs that ar
classical spec
microbes can
sequences ar
in situ hybridi
their morpho

Concepts

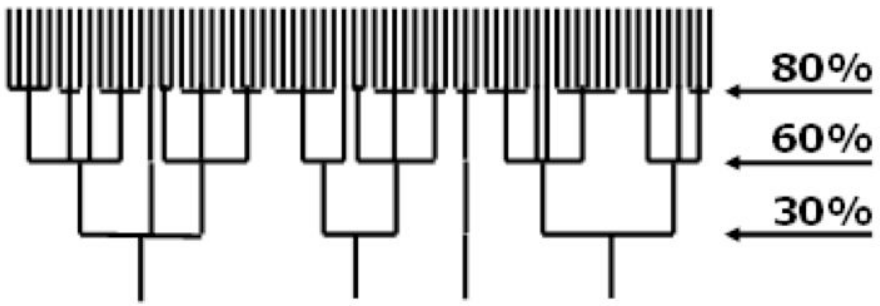
Various concepts have been suggested for microbial species, but none have been generally accepted[9]. The following quotes represent several published concepts that were chosen to illustrate the lack of consensus:

- "A species could be described as a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics, and is diagnosable by a discriminative phenotypic property." (REF. 9)
- "Species are considered to be an irreducible cluster of organisms diagnosably different from other such clusters and within which there is a parental pattern of ancestry and descent." (REF. 82)
- "A species is a group of individuals where the observed lateral gene transfer within the group is much greater than the transfer between groups." (REF. 83)
- "Microbes … do not form natural clusters to which the term "species" can be universally and sensibly applied." (REF. 84)
- "Species are (segments of) metapopulation lineages." (REF. 7)

Microbial diversity and the genetic nature of microbial species
Achtman & Wagner (2008) Nature Reviews Microbiology. doi:10.1038/nrmicro1872

Operational Taxonomic Units (OTUs)

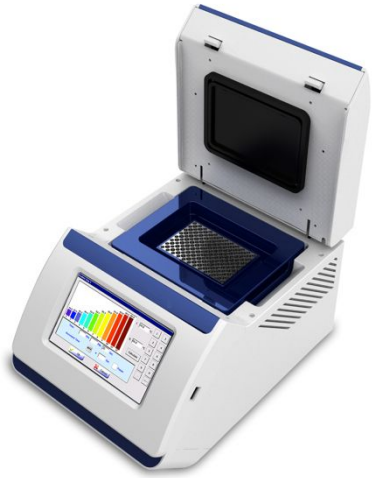
OTUs take the place of "species" in many microbiome diversity analyses because named species genomes are often unavailable for particular marker sequences.



- Although much of the 16S rRNA gene is highly conserved, several of the sequenced regions are variable or hypervariable, so small numbers of base pairs can change in a very short period of evolutionary time.
- Because 16S regions are typically sequenced using only a single pass, there is a fair chance that they will thus contain at least one sequencing error. This means that requiring tags to be 100% identical will be extremely conservative and treat essentially clonal genomes as different organisms.
- Some degree of sequence divergence is typically allowed - 95%, 97%, or 99% are sequence similarity cutoffs often used in practice [18] - and the resulting cluster of nearly-identical tags (and thus assumedly identical genomes) is referred to as an Operational Taxonomic Unit (OTU) or sometimes phylotype.

Chapter 12: Human Microbiome Analysis
Morgan & Huttenhower (2012) PLOS Comp Bio.https://doi.org/10.1371/journal.pcbi.1002808

16S versus shotgun NGS



16S

Fast (minutes – hours)
Directed analysis
Cheap per sample
Family/Genus Identification

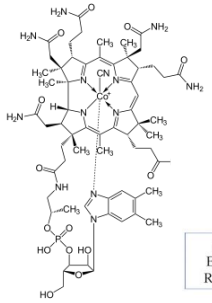
NGS

Slower (hours to days)
Whole Metagenome
More expensive per sample
Species/Strain Identification
Genes presence/absence
Variant analysis
Eukaryotic hosts
Can ID fungi, viruses, etc.

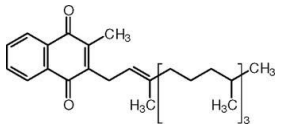
E. coli

?

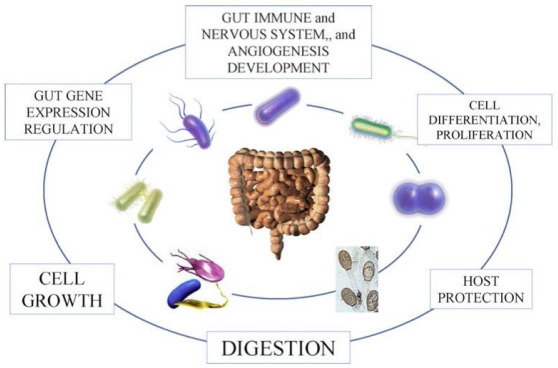
Commensal?



Vitamin B12



Vitamin K



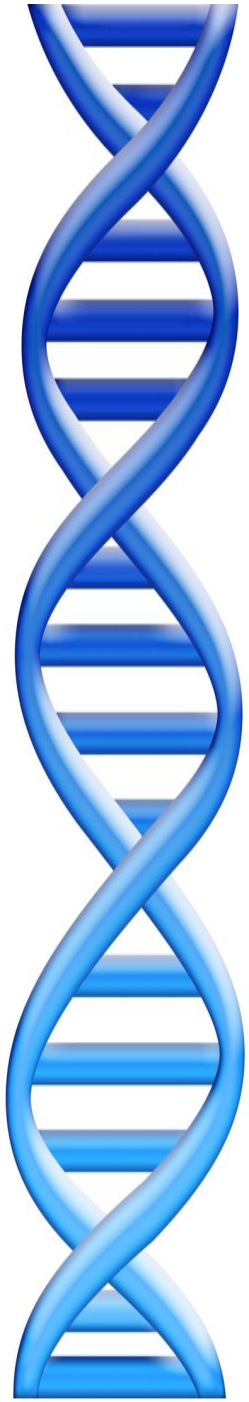
Colonization Barrier Effect Against Pathogens

Pathogenic?

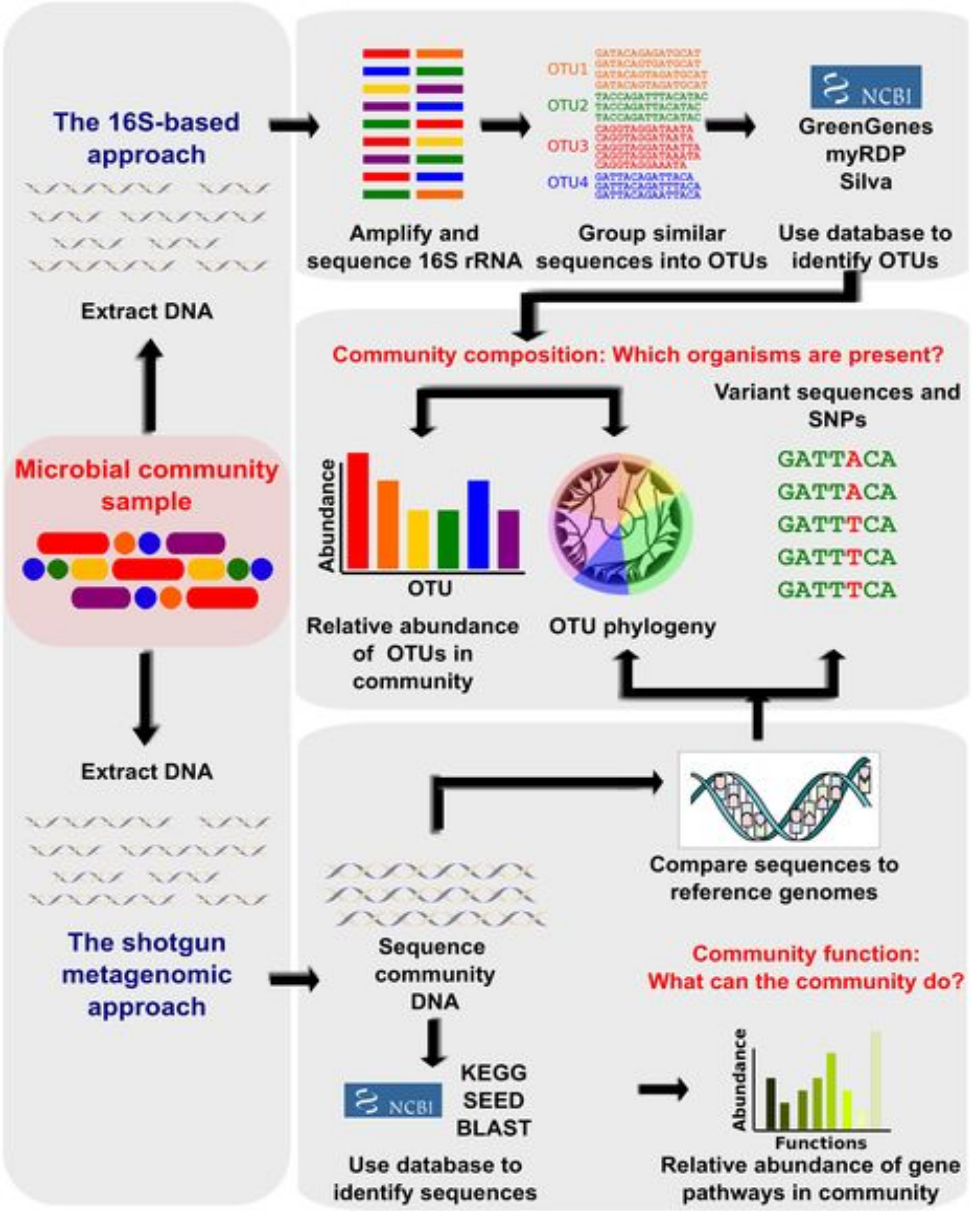
Strain O157:H7

Hemorrhagic enteritis


PRIUS



Part II:
Metagenomics Methods



Chapter 12: Human Microbiome Analysis

Morgan & Huttenhower (2012) PLOS Comp Bio.https://doi.org/10.1371/journal.pcbi.1002808

ML for Metagenomics

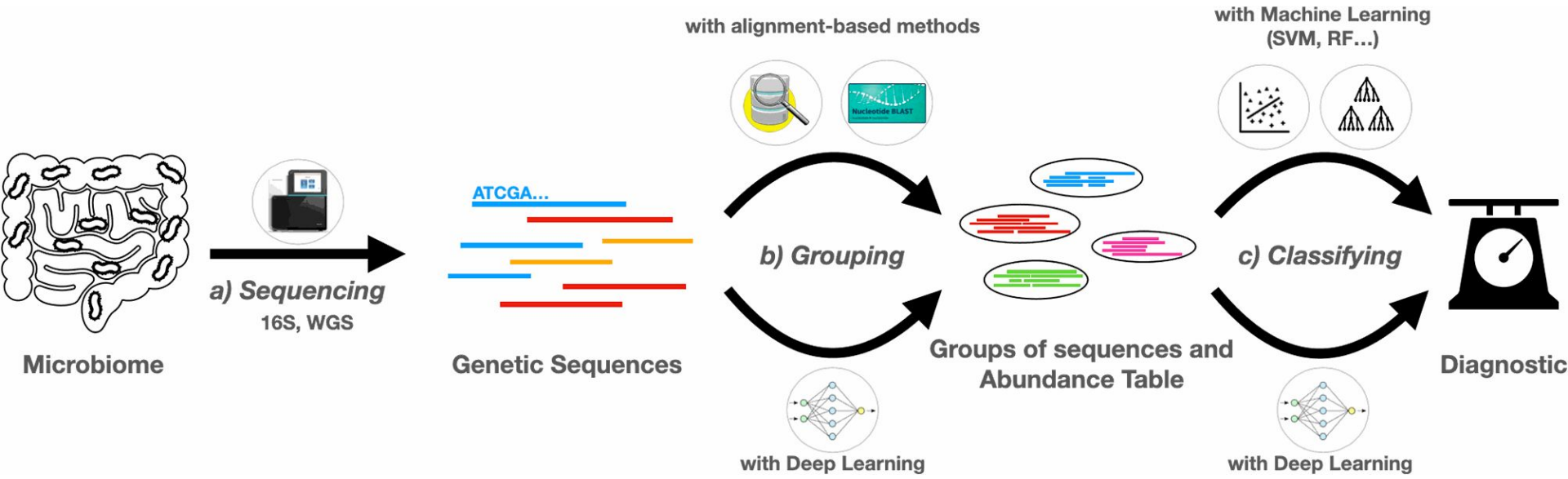
MICROBIAL GENOMICS
Volume 10, Issue 4

Review Article | Open Access

Deep learning methods in metagenomics: a review

Gaspar Roy[1], Edi Prifti[1,2], Eugeni Belda[1,2] and Jean-Daniel Zucker[1,2]
View Affiliations

Published: 17 April 2024 | https://doi.org/10.1099/mgen.0.001231



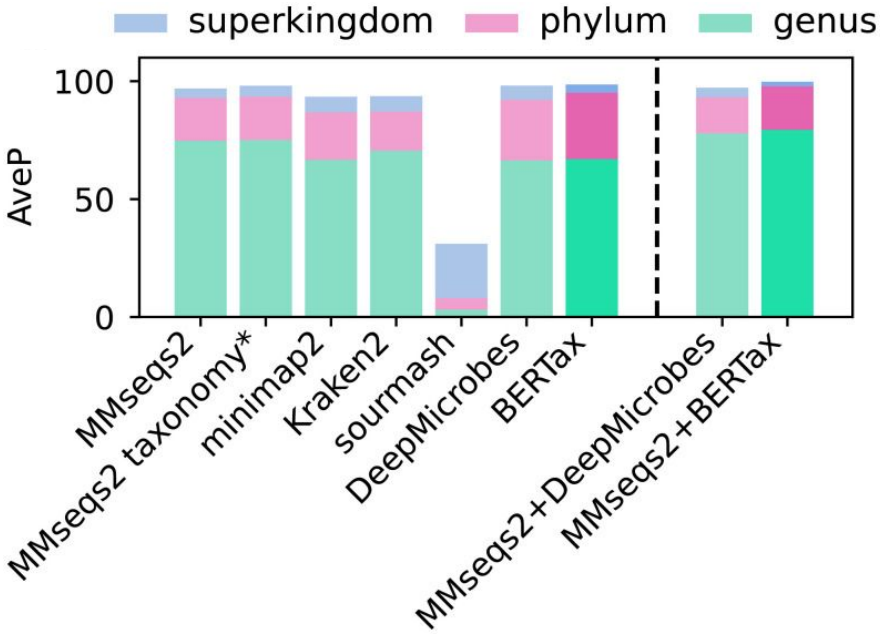
Taxonomic Classification

Input: raw reads or contigs

Traditional methods: Kraken, MetaPhlAn4

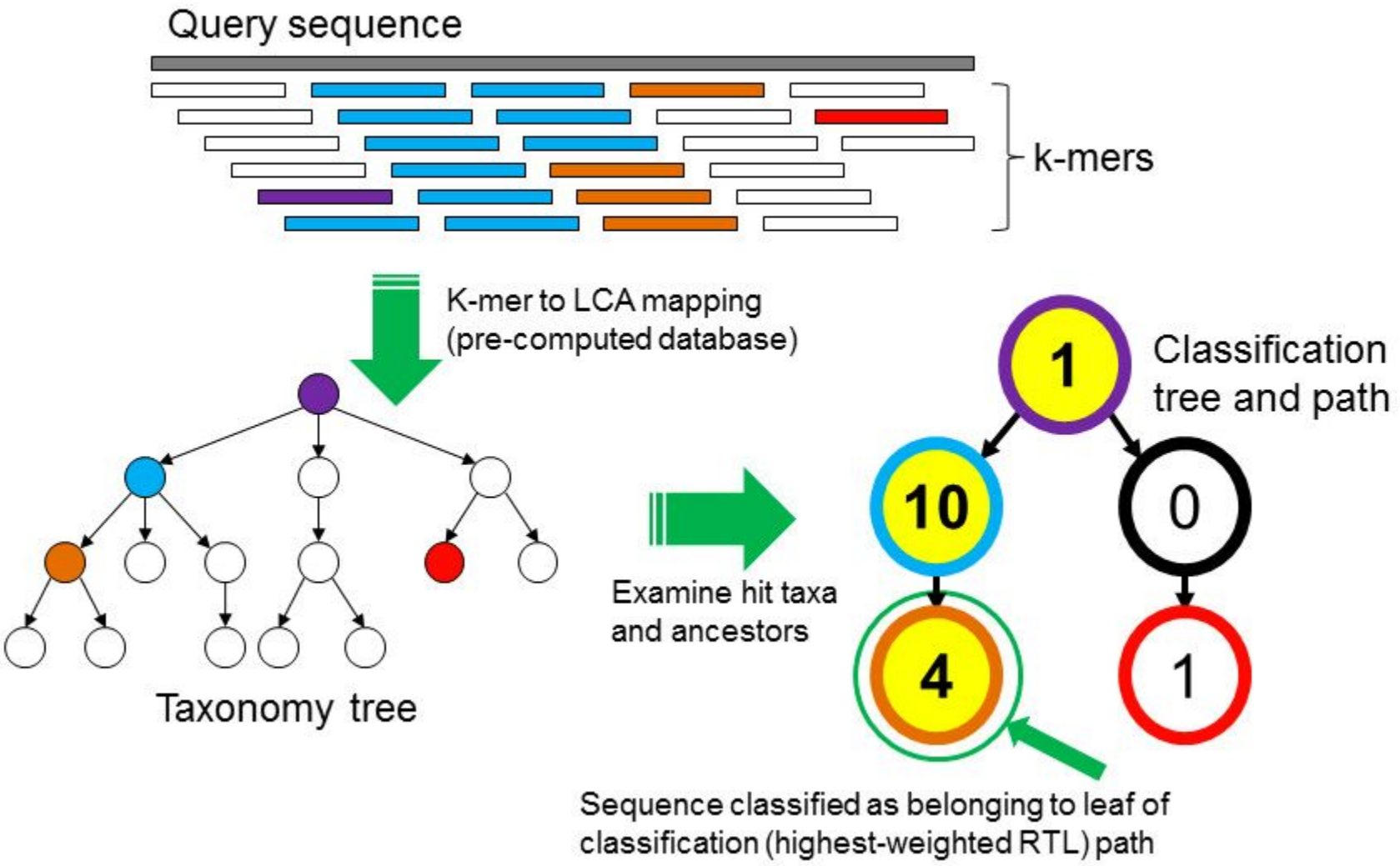
ML methods:

- DeepMicrobes
- BERTax
- VirSearcher - viral classification



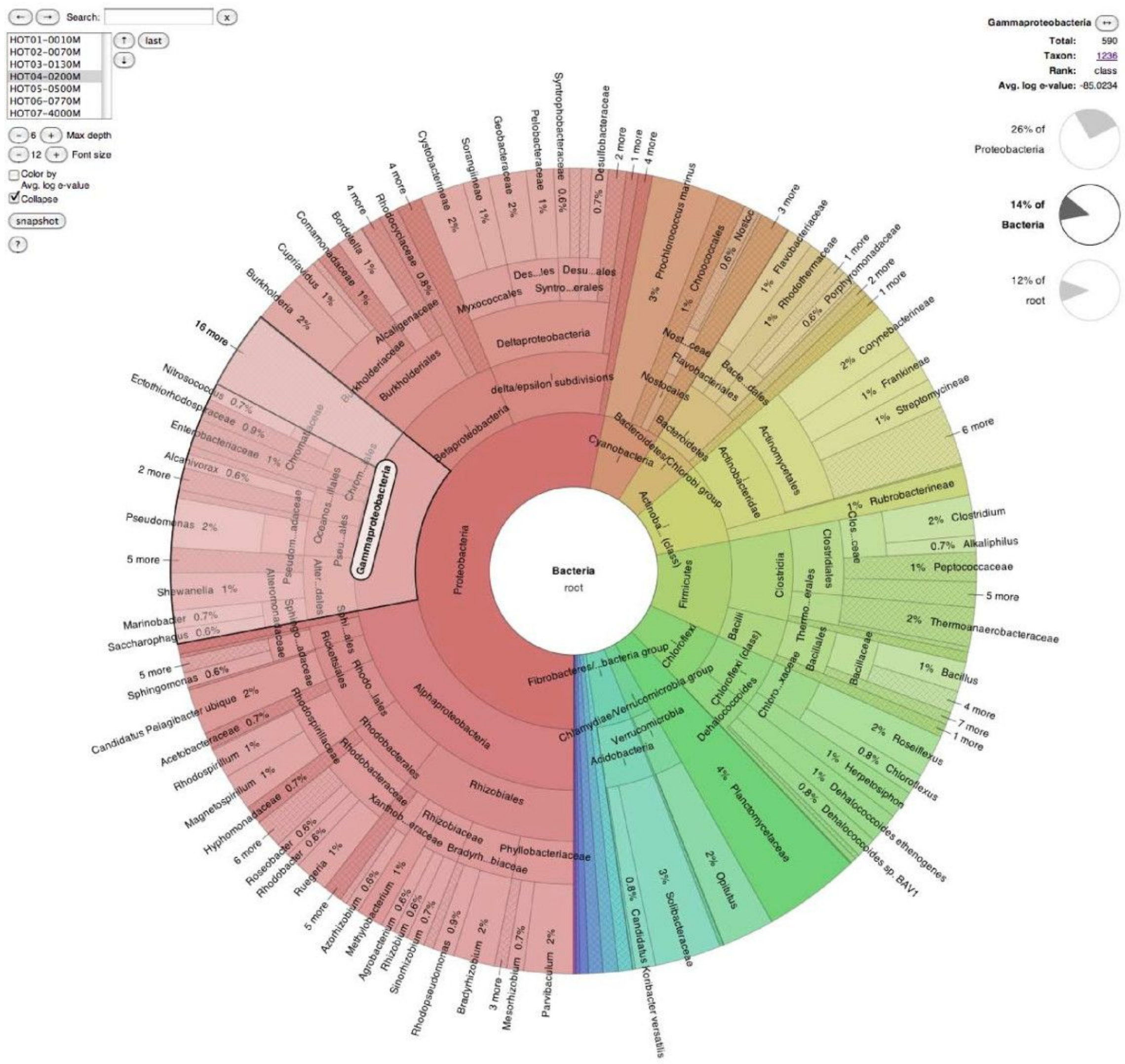
Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks
Mock et al. (2022) Proc. Natl. Acad. Sci. DOI: pnas.212263611 9

Kraken



Kraken: ultrafast metagenomic sequence classification using exact alignments
Wood and Salzberg (2014) Genome Biology. DOI: 10.1186/gb-2014-15-3-r46

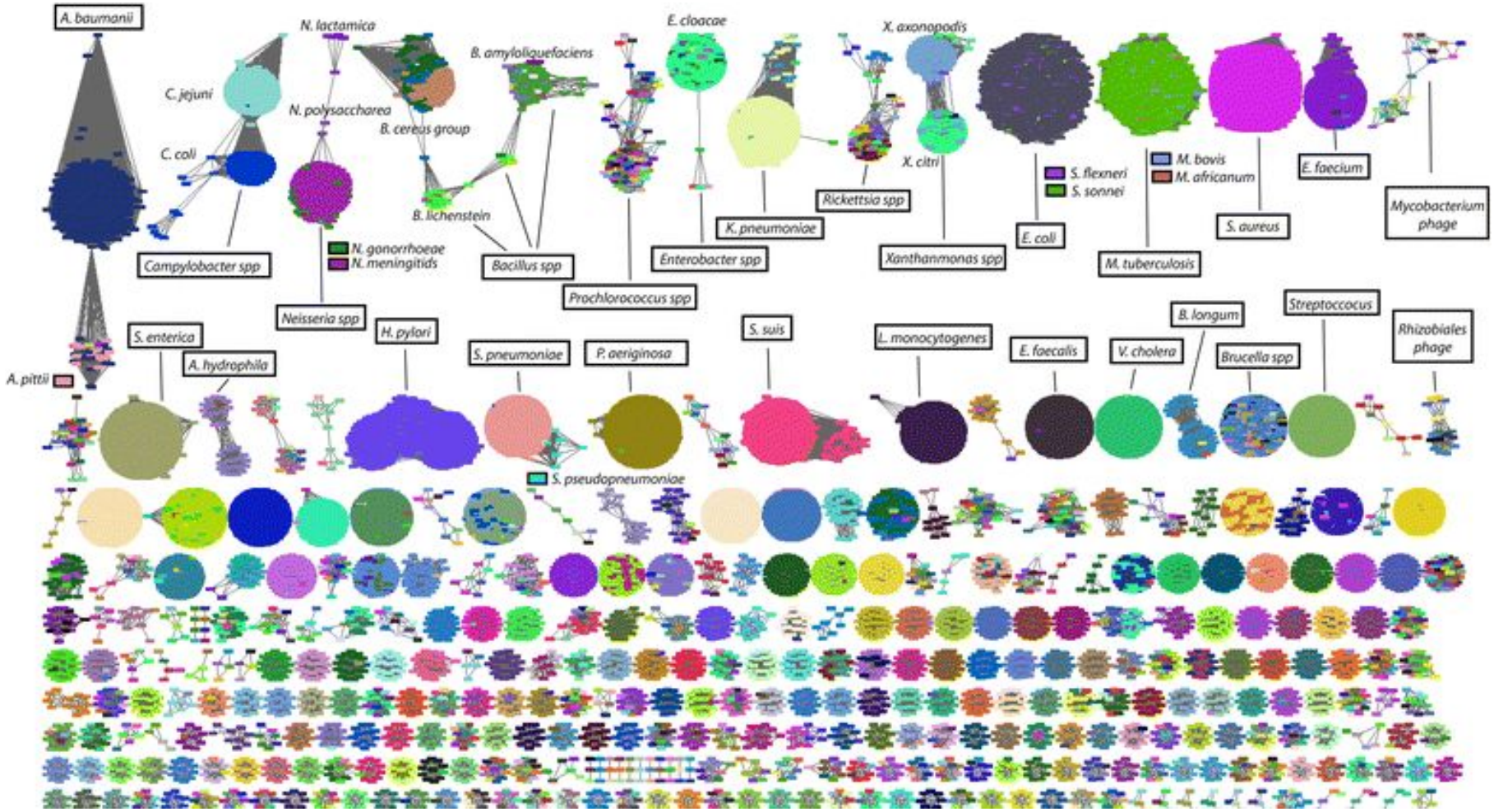
Krona Plots



Interactive metagenomic visualization in a Web browser

Ondov et al (2011) BMC Bioinformatics. DOI: 10.1186/1471-2105-12-385

Min-Hash: Comparing all 54,118 RefSeq genomes in 1 day on a laptop



Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al. (2016) Genome Biology. DOI: 10.1186/s13059-016-0997-x

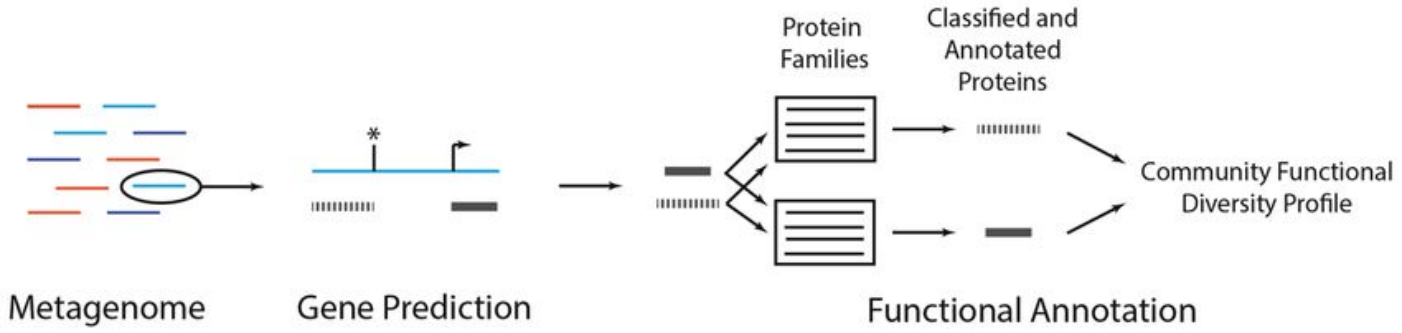
Functional Annotation

Input: raw reads or contigs

Traditional methods: sequence similarity to reference databases (e.g. BLAST)

ML methods:

- DeepARG - antimicrobial resistance genes
- Meta-MFDL - ORF
- PPR-Meta - Phage/plasmid



An introduction to the analysis of shotgun metagenomic data
Sharpton (2014) Front. Plant Sci. DOI: 10.3389/fpls.2014.00209

Metagenomic Binning

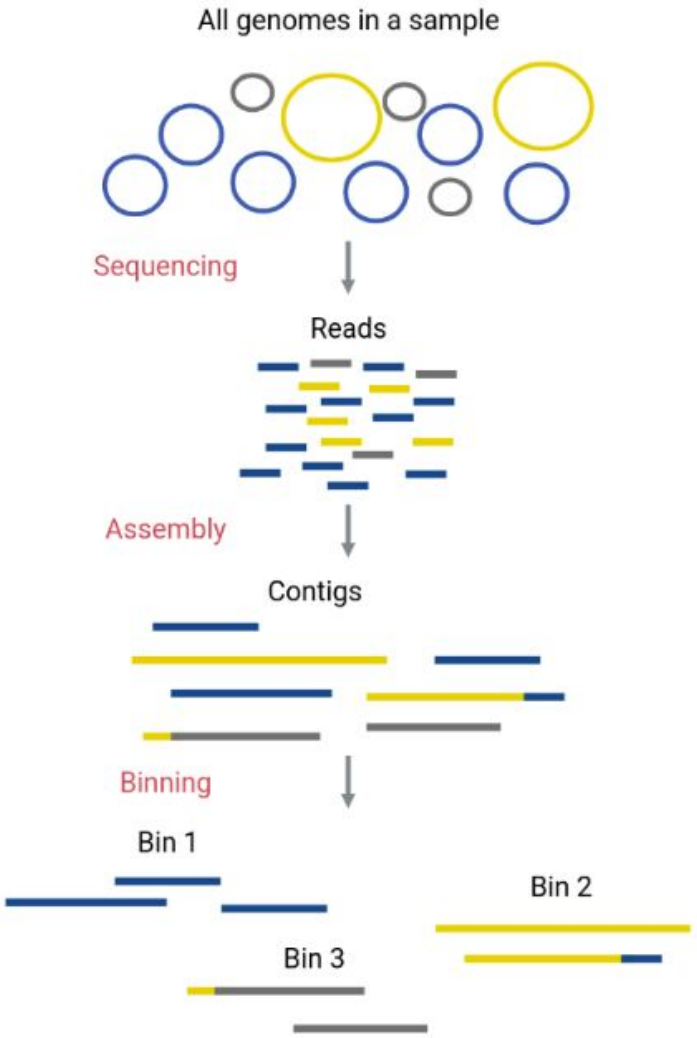
Input: contigs

Traditional methods: MetaBat2

ML methods:

- VAMB
- CCVAE
- AAMB

Produces metagenome-assembled genomes (MAGs)



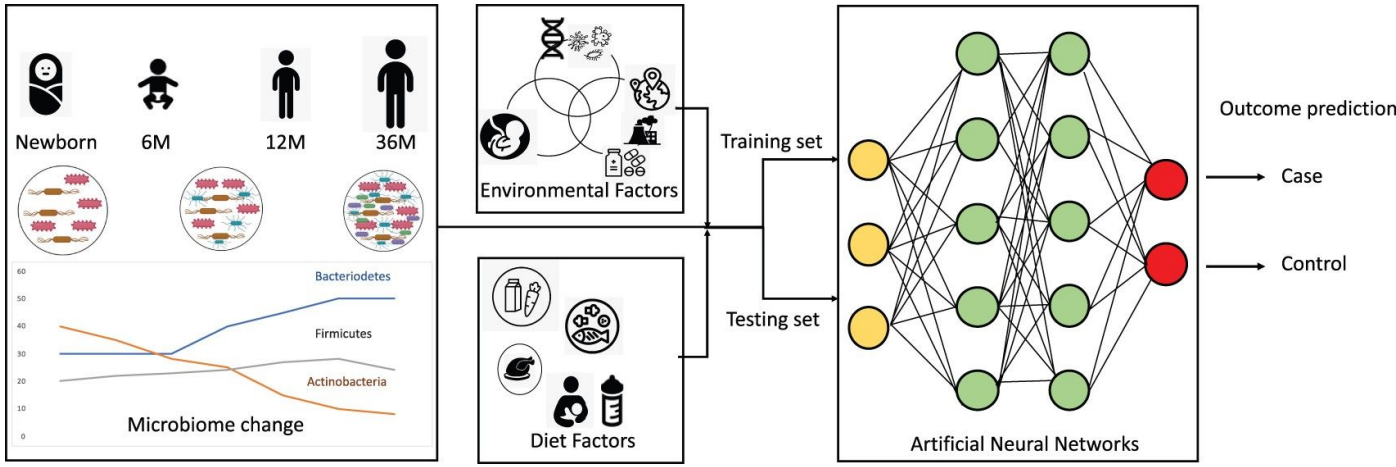
Phenotype Prediction

Input: abundance matrix

Difficult because data is sparse and the number of features is usually much larger than the number of samples

Current work is mostly proof-of-concepts:

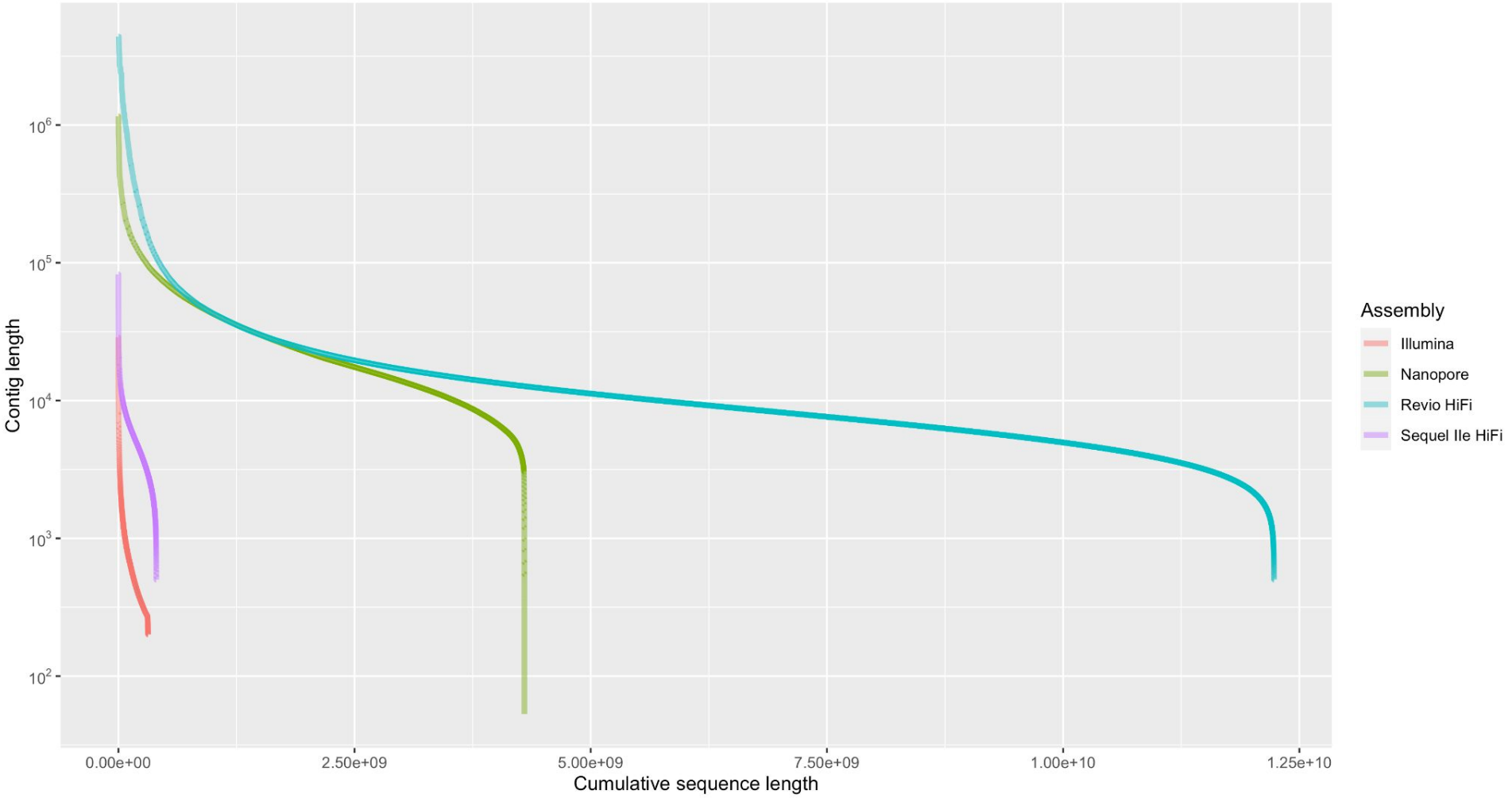
- EnsDeepDP
- MetaDR
- PhyLoSTM

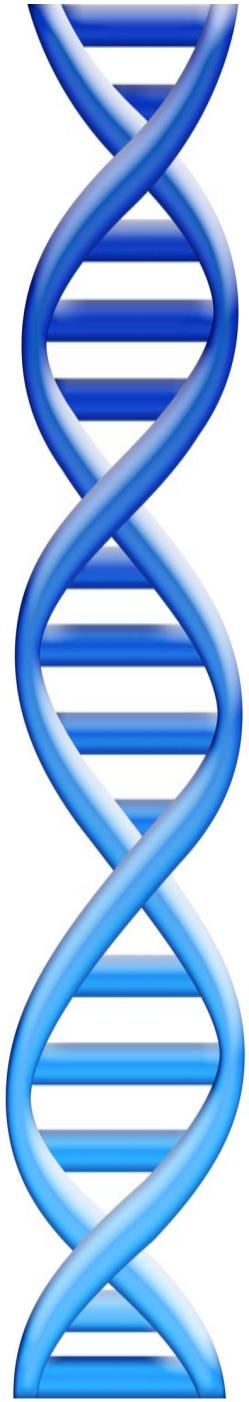


phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data
Sharma et al. (2021) Bioinformatics DOI: 10.1093/bioinformatics/btab482

What about long reads?

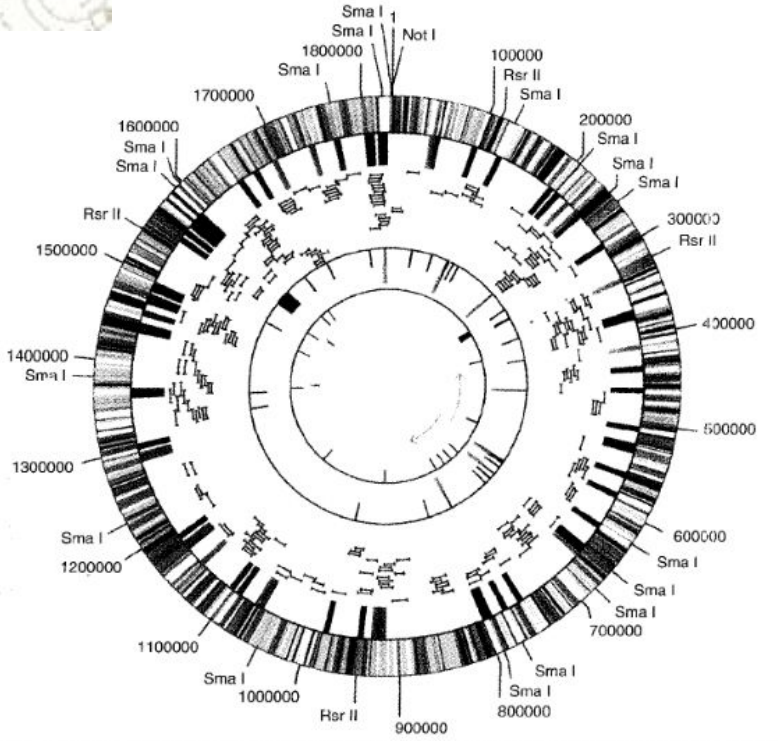
Much longer reads → Much longer contigs





Part III: Results

The first microbial genomes



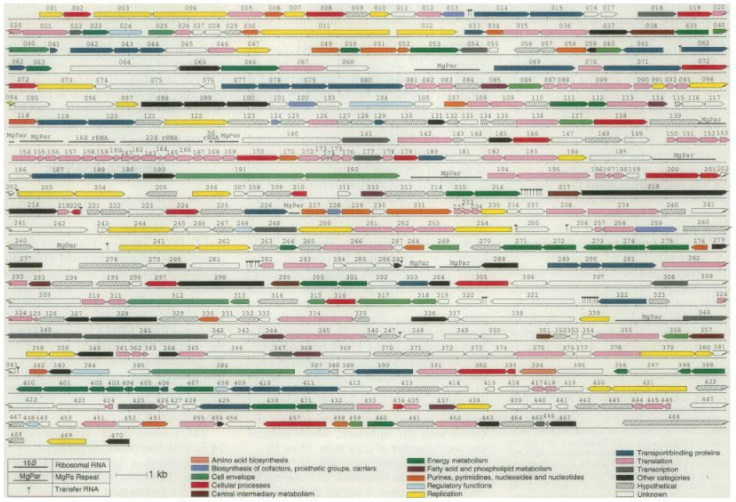
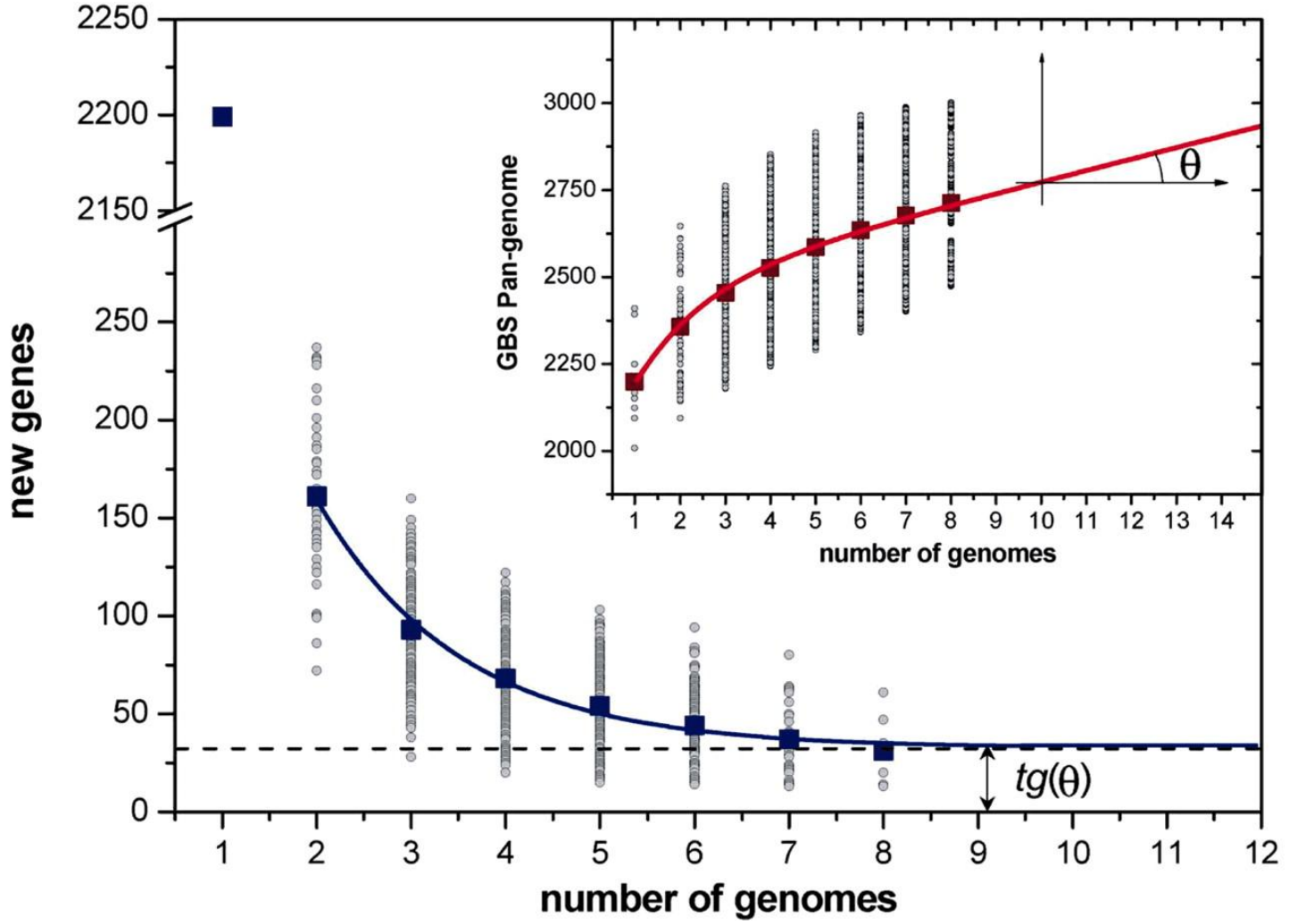


Fig. 1. Gene map of the *M. genitalium* genome. Predicted coding regions are shown, and the direction of transcription is indicated by arrows. Each line in the figure represents 24,000 bp of sequence in the *M. genitalium* genome. Genes are color-coded by role category as described in the key. Gene identificaton numbers correspond to those in Table 1. The rRNA operon, tRNA genes, and adhesin protein (MgPa) operon repeats are labeled.

Whole-genome random sequencing and assembly of Haemophilus influenzae Rd
Fleischmann et al (1995) Science. doi: 10.1126/science.7542800

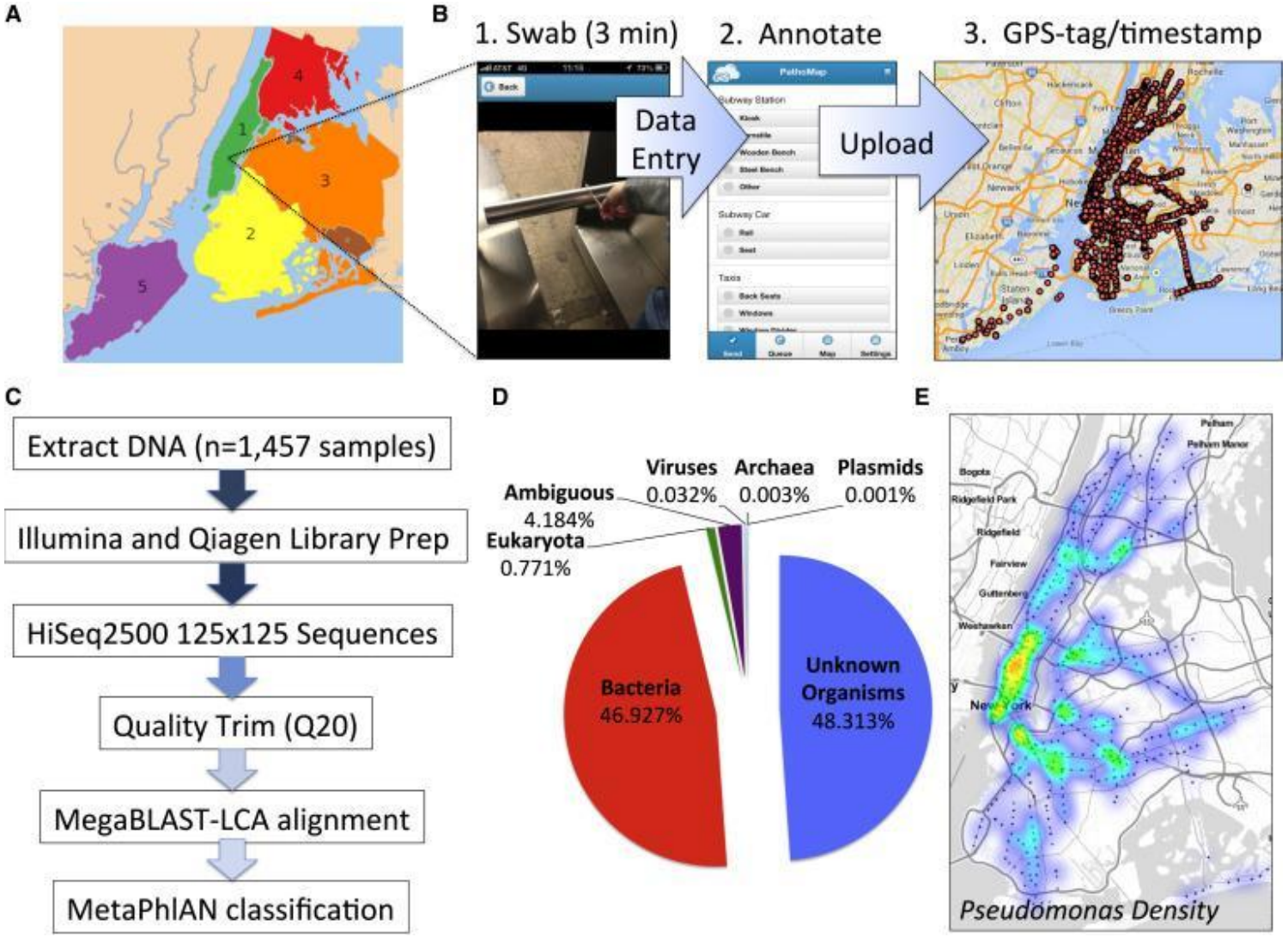
The Minimal Gene Complement of *Mycoplasma genitalium*
Fraiser et al (1995) Science. doi: 10.1126/science.270.5235.397

The first pan genome: *Streptococcus agalactiae*



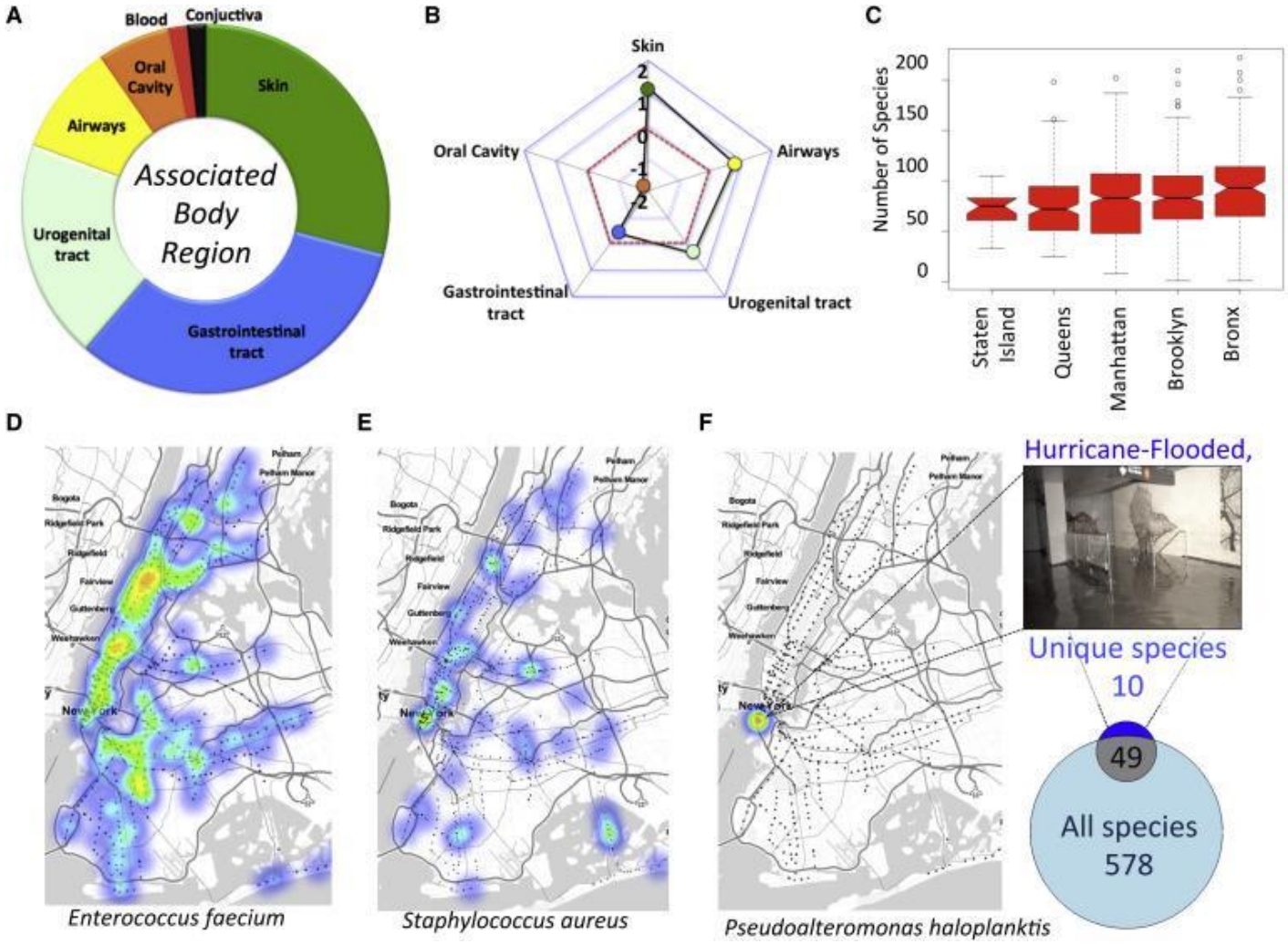
Hervé Tettelin et al. PNAS 2005;102:13950-13955

MetaSUB



Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics
Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

Different subway stations resembled different body sites



Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics
Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

Dangerous pathogens and mystery microbes ride the subway

FEBRUARY 6, 2015 / 10:42 AM / CBS NEWS

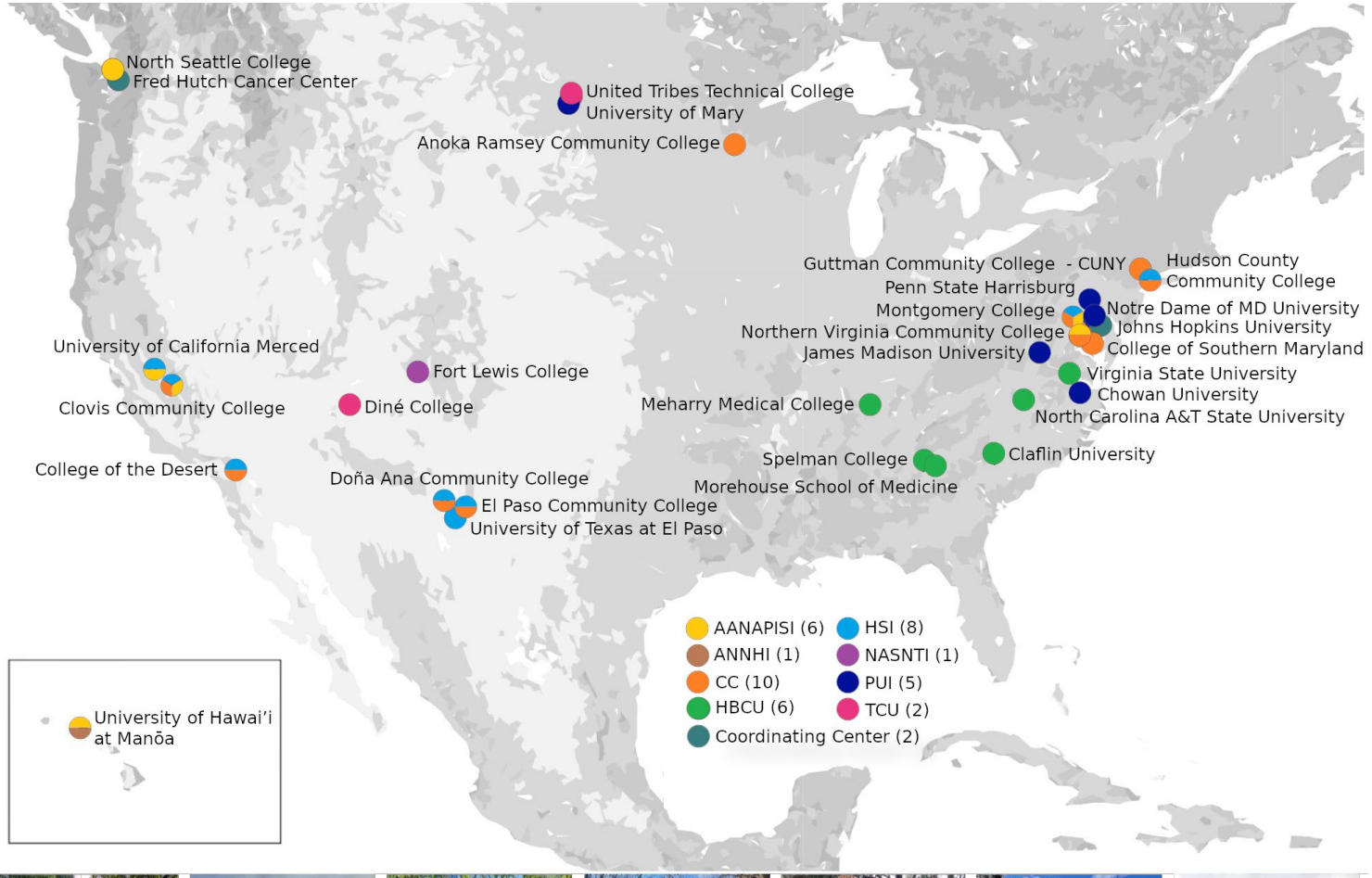
New York City's subway system has never been known for its cleanliness, but even the most jaded city dweller may be shocked and disgusted to learn just what types of microorganisms are lurking on the average subway pole.

A group of researchers led by Christopher Mason of the department of physiology and biophysics at Weill Cornell Medical College swabbed surfaces and collected specimens from the subway system to develop a map of what they called an "urban microbiome." The result, seen above, is called the PathoMap and it illustrates

Bubonic Plague in the Subway System? Don't Worry About It

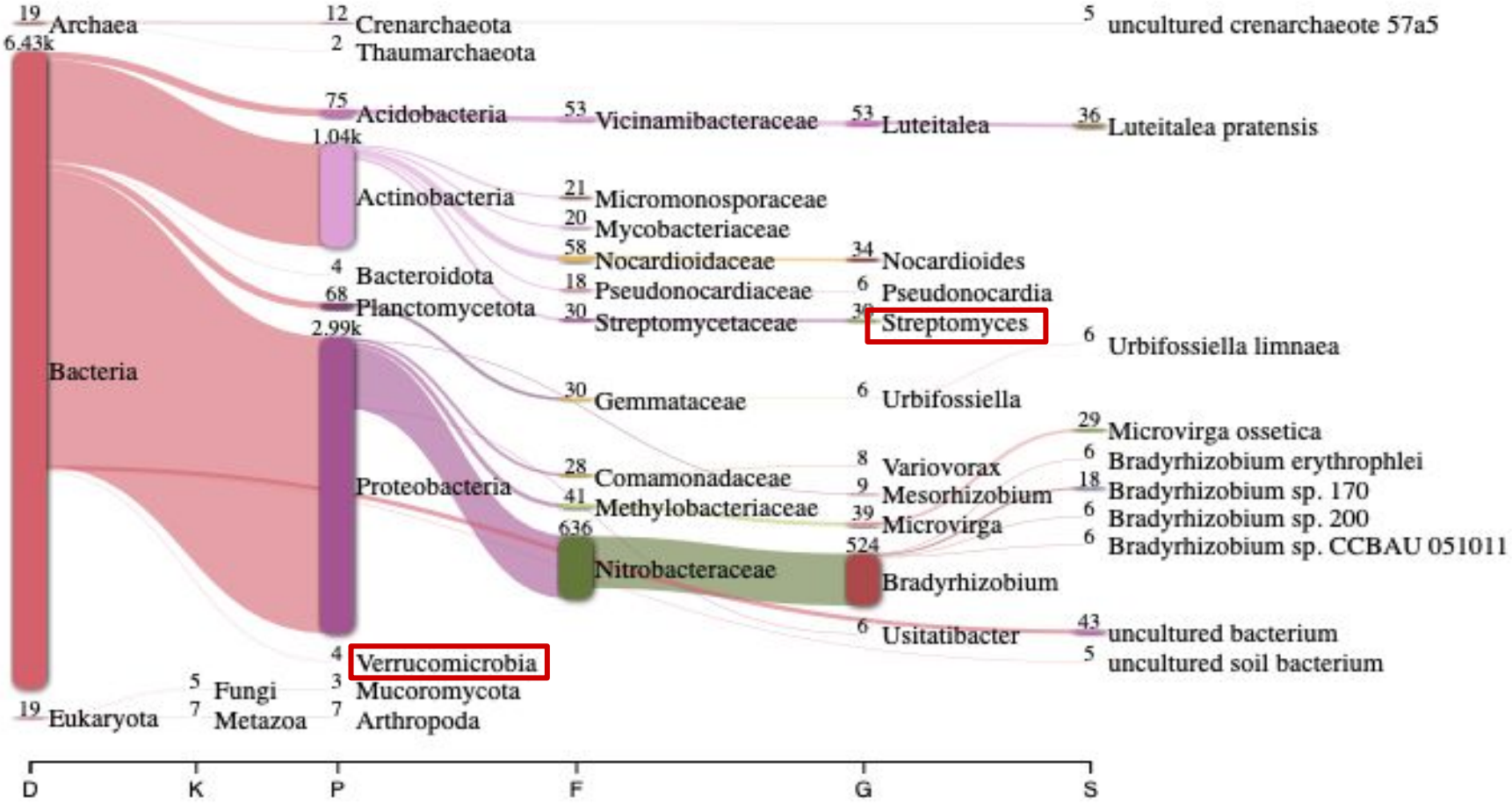
In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

BioDIGS


North Seattle College
Fred Hutch Cancer Center
United Tribes Technical College
University of Mary
Anoka Ramsey Community College
Guttman Community College - CUNY
Hudson County Community College
Penn State Harrisburg
Montgomery College
Notre Dame of MD University
Johns Hopkins University
Northern Virginia Community College
College of Southern Maryland
James Madison University
Virginia State University
Chowan University
North Carolina A&T State University
University of California Merced
Fort Lewis College
Clovis Community College
Diné College
Meharry Medical College
College of the Desert
Spelman College
Claflin University
Morehouse School of Medicine
Doña Ana Community College
El Paso Community College
University of Texas at El Paso
University of Hawai'i at Manōa
AANAPISI (6)
HSI (8)
ANNHI (1)
NASNTI (1)
CC (10)
PUI (5)
HBCU (6)
TCU (2)
Coordinating Center (2)


Anoka, MN
Fresno, CA
Seattle, WA
Waynesboro, VA
Atlanta, GA
El Paso, TX
Durango, CO
Baltimore, MD
Augusta, GA
El Paso, TX
Bismarck, ND
Alexandria, VA
Nashville, TN
Merced, CA

Diversity of soil is still largely unknown



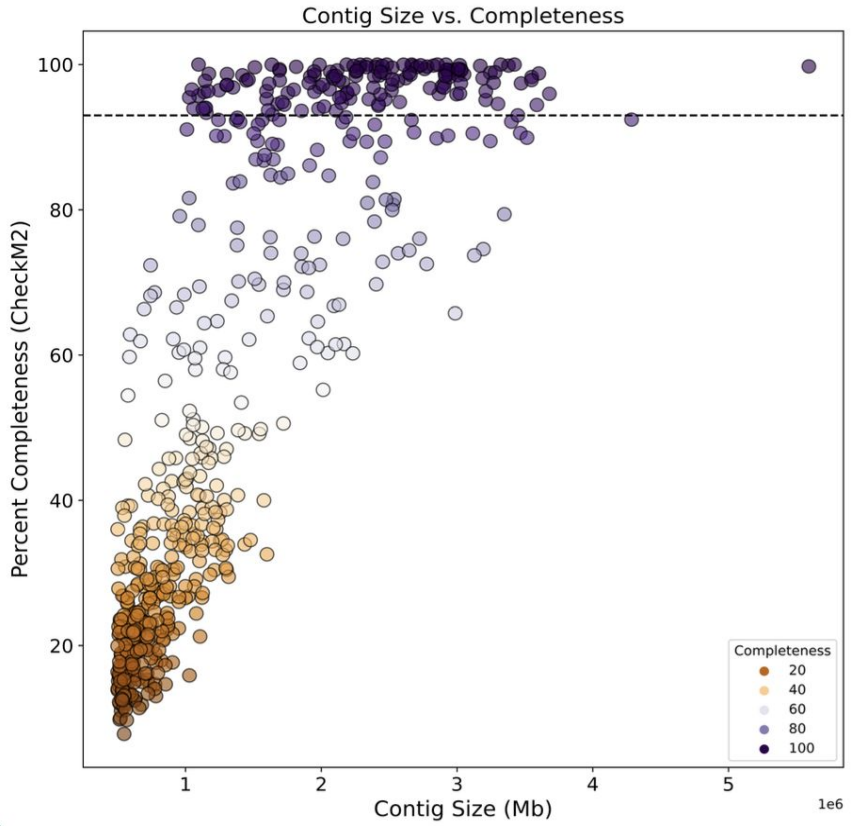
Diversity of soil is still largely unknown

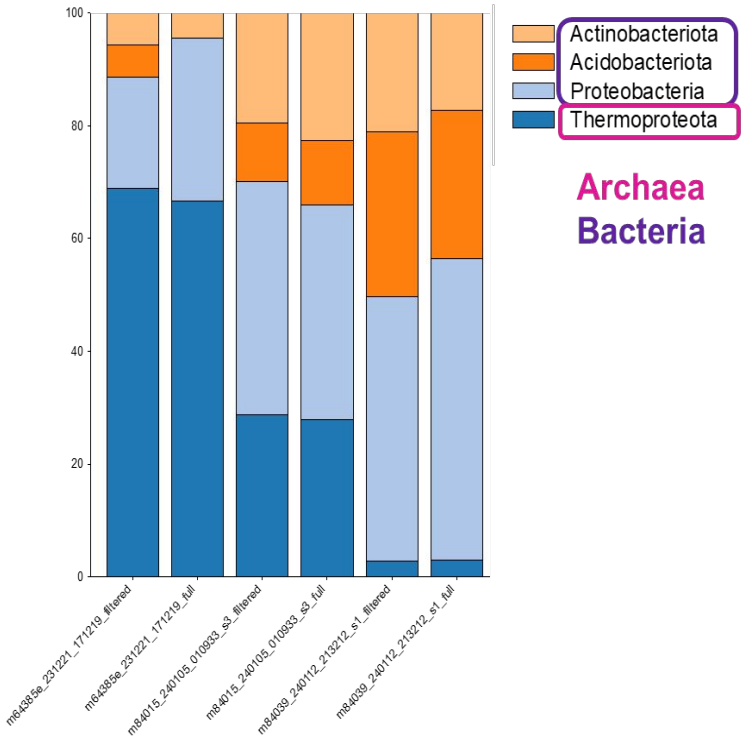
Name	Number of raw reads	Classified reads	Chordate reads	Artificial reads	Unclassified reads	Microbial reads	Bacterial reads	Viral reads	Fungal reads	Protozoan reads
HHGCVDRX3-1-ACTCGGCAAT-TTCAGTTGTC_S41_L002	5,779,462	15.5%	0.00114%	0%	84.5%	12.5%	8.79%	0.00147%	0.0535%	0.00019%
HHGCVDRX3-1-ACTCGGCAAT-TTCAGTTGTC_S41_L001	5,501,528	15.4%	0.002%	0%	84.6%	12.4%	8.7%	0.00129%	0.0515%	0.000309%
HHGCVDRX3-1-CAATCGGCTG-TTCCTACAGC_S39_L002	5,857,162	13.3%	0.00099%	0%	86.7%	11.3%	9.08%	0.00082%	0.0453%	0.000137%
HHGCVDRX3-1-CAATCGGCTG-TTCCTACAGC_S39_L001	5,617,529	13.2%	0.00141%	0.0000178%	86.8%	11.2%	8.96%	0.000819%	0.0445%	0.000196%
HHGCVDRX3-1-GAACTGAGCG-CGCTCCACGA_S1_L002	5,315,099	12.9%	0.00122%	0%	87.1%	10.9%	8.6%	0.000978%	0.0396%	0.000263%
HHGCVDRX3-1-GAACTGAGCG-CGCTCCACGA_S1_L001	5,087,303	12.7%	0.001%	0%	87.3%	10.8%	8.5%	0.00106%	0.0392%	0.000216%
HHGCVDRX3-1-GATCAAGGCA-ATTAACAAGG_S8_L001	7,260,595	11.8%	0.00169%	0%	88.2%	9.51%	7.06%	0.000689%	0.0122%	0.0000826%
HHGCVDRX3-1-GATCAAGGCA-ATTAACAAGG_S8_L002	7,676,877	11.8%	0.00168%	0%	88.2%	9.56%	7.12%	0.000391%	0.0122%	0.0000912%
HHGCVDRX3-1-ACCGGCCGTA-AATATTGCCA_S36_L002	7,379,576	11.5%	0.000474%	0.0000136%	88.5%	10.8%	9.31%	0.0011%	0.0394%	0.0000813%
HHGCVDRX3-1-CGTCTCATAT-AGCTACTATA_S3_L002	5,157,206	11.5%	0.000756%	0%	88.5%	10.7%	8.93%	0.0014%	0.0358%	0.000427%
HHGCVDRX3-1-ACCGGCCGTA-AATATTGCCA_S36_L001	6,918,535	11.4%	0.000795%	0%	88.6%	10.7%	9.19%	0.00116%	0.041%	0.0000434%
HHGCVDRX3-1-CGTCTCATAT-AGCTACTATA_S3_L001	4,917,122	11.4%	0.000936%	0%	88.6%	10.6%	8.82%	0.00122%	0.0348%	0.000285%
HHGCVDRX3-1-CTAGTGCTCT-TACTGTTCCA_S7_L002	5,701,275	11.4%	0.0016%	0%	88.6%	8.31%	6.55%	0.000368%	0.0175%	0.0000526%
HHGCVDRX3-1-GGTTGCGAGG-TTGCTCTATT_S26_L002	9,214,523	11.4%	0.00158%	0%	88.6%	11%	9.57%	0.0014%	0.0435%	0.000184%
HHGCVDRX3-1-CTAGTGCTCT-TACTGTTCCA_S7_L001	5,464,649	11.3%	0.00135%	0%	88.7%	8.22%	6.46%	0.000421%	0.0161%	0.000146%

Showing 1 to 15 of 96 entries

Previous 1 2 3 4 5 6 7 Next

Long reads facilitate assembly





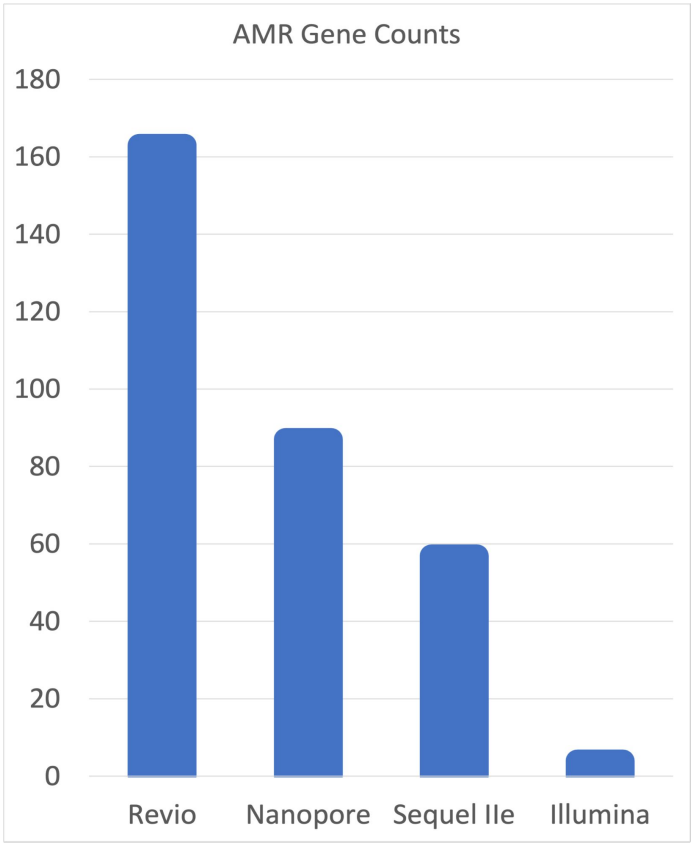
Out of 158 MAGs (55 HQ), only 1 could be assigned to a known species

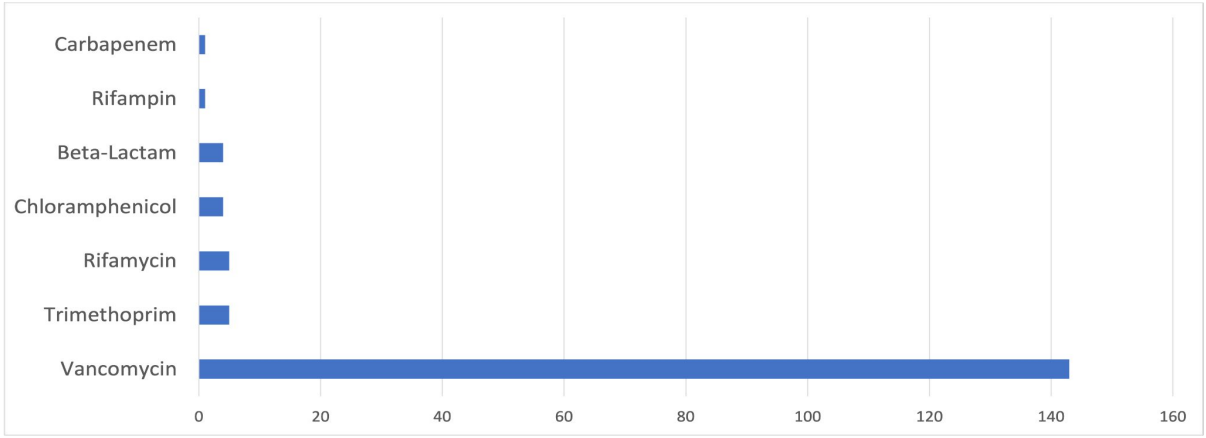
- Archaea are highly represented
- ~25x more HQ MAGs / site than short reads
- 27 single contig genomes!

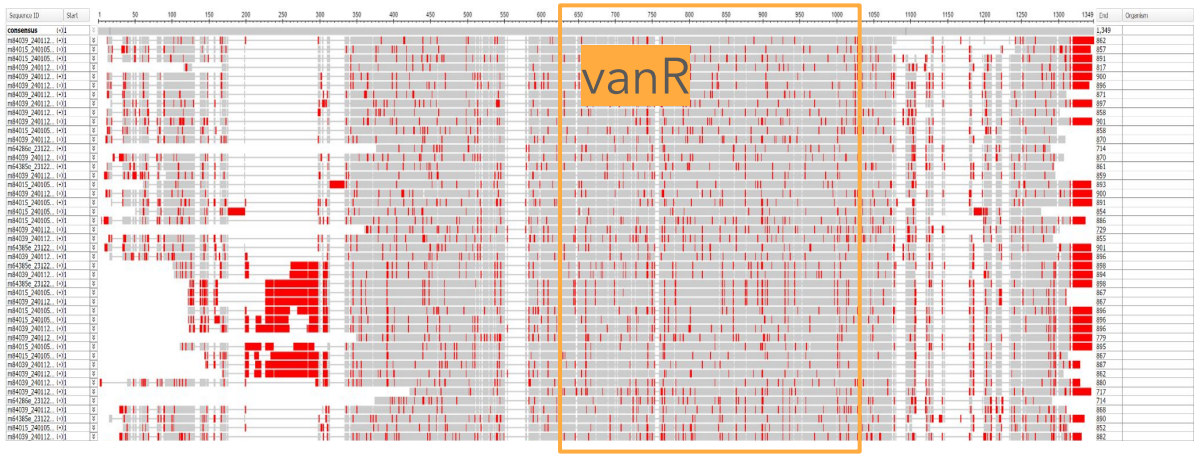
Only 1 MAG of 158 could be assigned to a species

- Lots of new diversity is present
- Archaea are highly represented

Long reads facilitate functional annotation





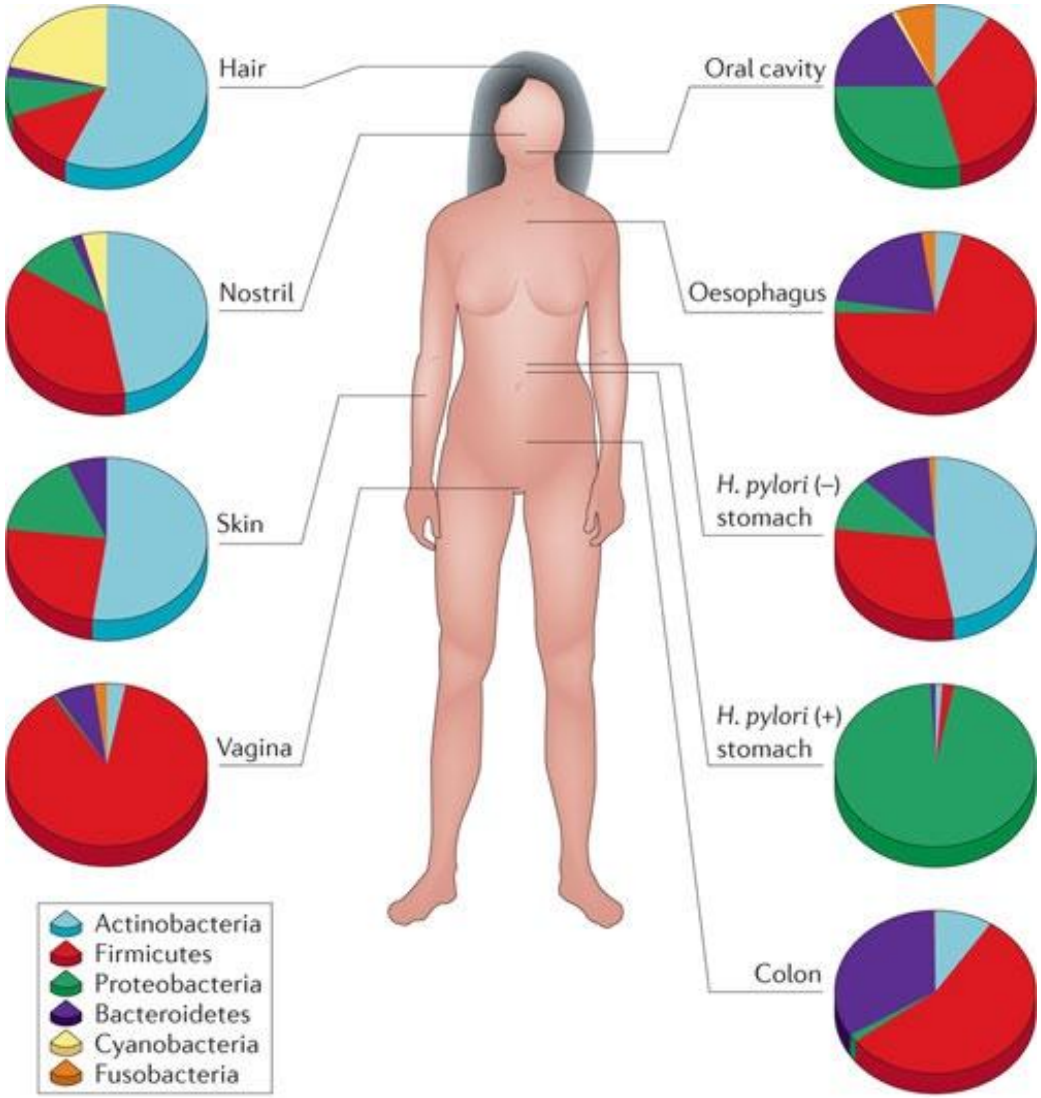


Microbes and Human Health

"MICROBE DIET Mice fed microbes from obese people tend to gain fat. Microbes from lean people protect mice from excessive weight gain, even when animals eat a high-fat, low-fiber diet."

Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice
Ridaura et al (2013) Science. doi: 10.1126/science.1241214

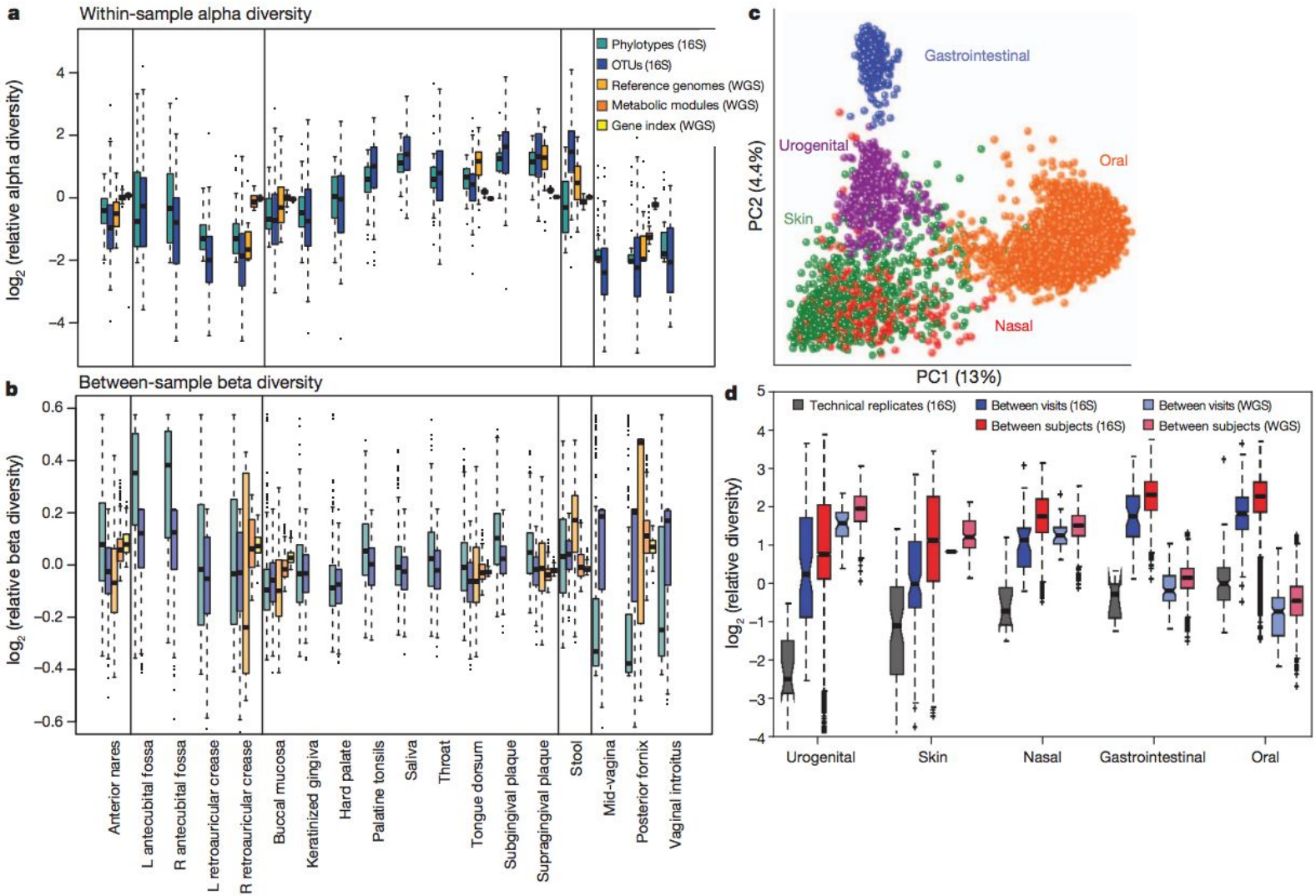
Microbes and Human Health



The human microbiome: at the interface of health and disease

Cho & Blaser (2012) Nature Reviews Genetics. doi:10.1038/nrg3182

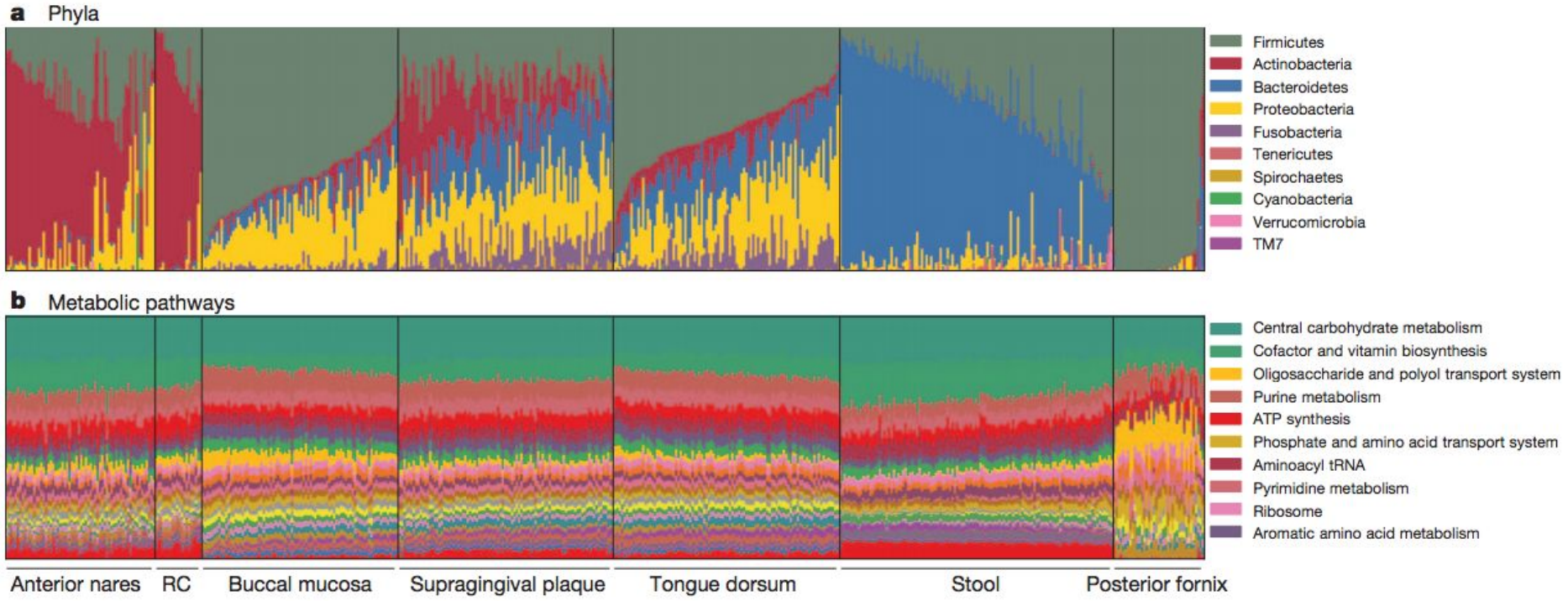
Human Microbiome Project



Structure, function and diversity of the healthy human microbiome

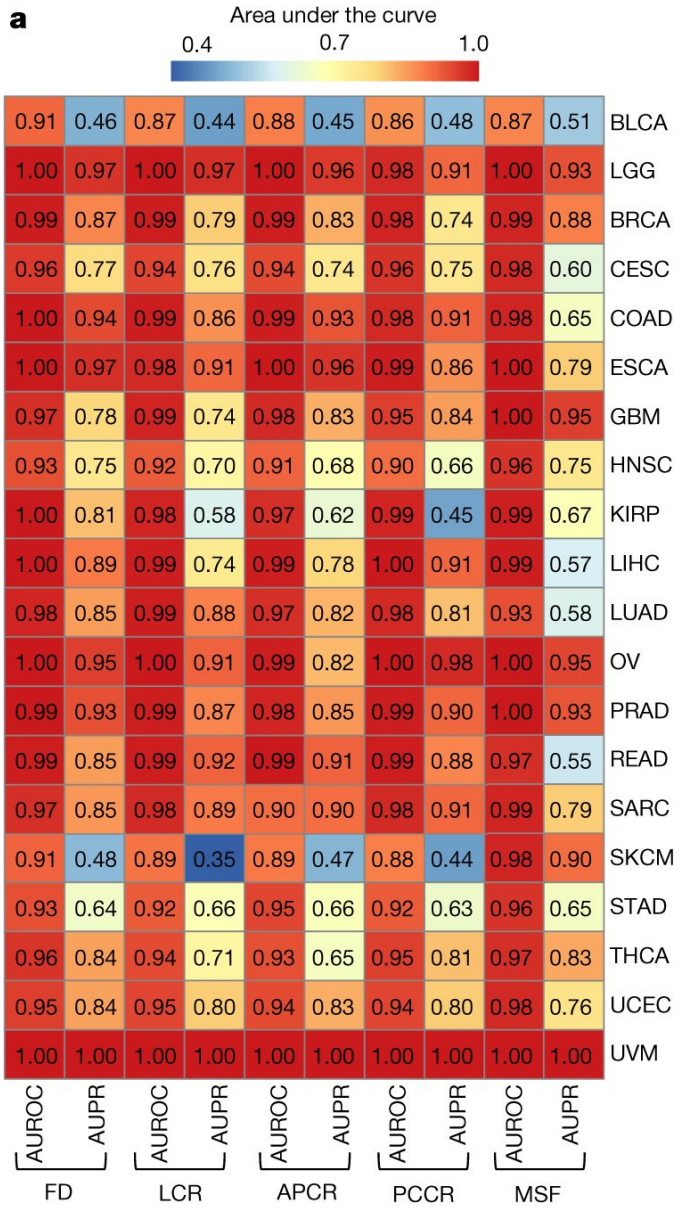
The Human Microbiome Project Consortium (2012) Nature. doi:10.1038/nature11234

Functional composition tends to be more stable than genome composition



Structure, function and diversity of the healthy human microbiome
The Human Microbiome Project Consortium (2012) Nature. doi:10.1038/nature11234

Cancer Microbiome?



nature › articles › article

Article | Published: 11 March 2020

Microbiome analyses of blood and tissues suggest cancer diagnostic approach

Gregory D. Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin Patel, Austin D. Swafford & Rob Knight

Nature **579**, 567–574 (2020) | Cite this article

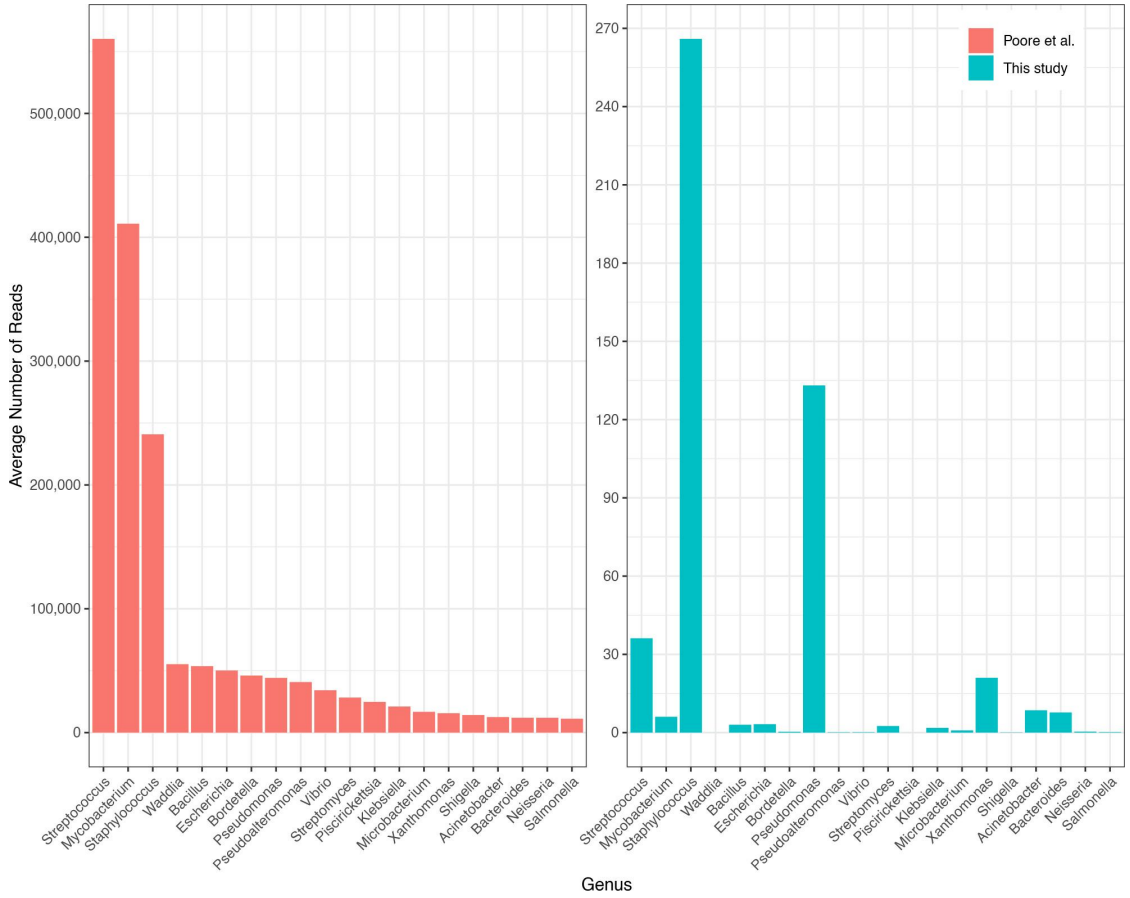
Not so fast

| Human Microbiome | Research Article | 9 October 2023

Major data analysis errors invalidate cancer microbiome findings

Authors: Abraham Gihawi, Yuchen Ge, Jennifer Lu, Daniela Puiu, Amanda Xu, Colin S. Cooper, Daniel S. Brewer, Mihaela Pertea, Steven L. Salzberg

AUTHORS INFO & AFFILIATIONS

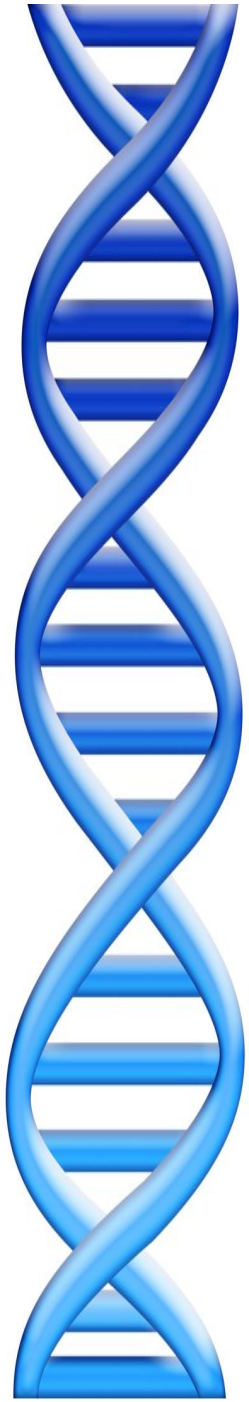


Not so fast

Issues:

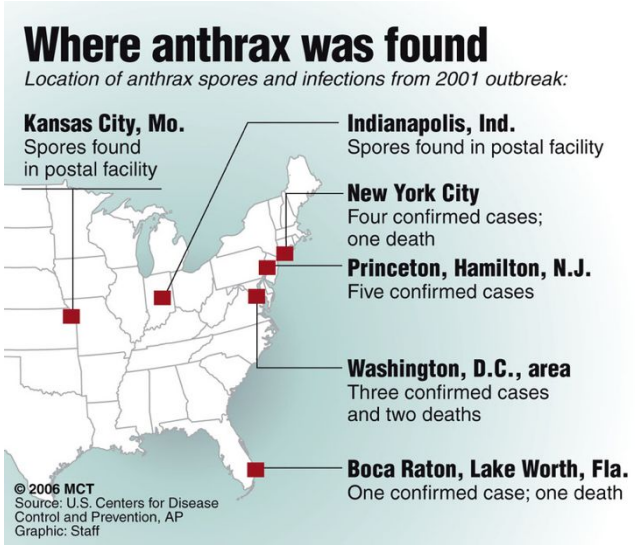
- Raw reads create the false sense of bacteria
  - Kraken database contained draft genomes
- Normalization of reads erroneously created distinct signature for each cancer
- Downstream studies have based off this original study!

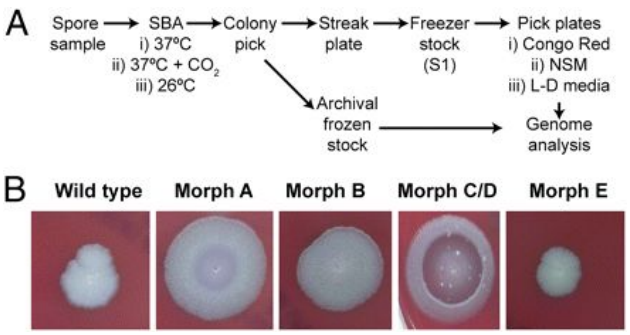
RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach

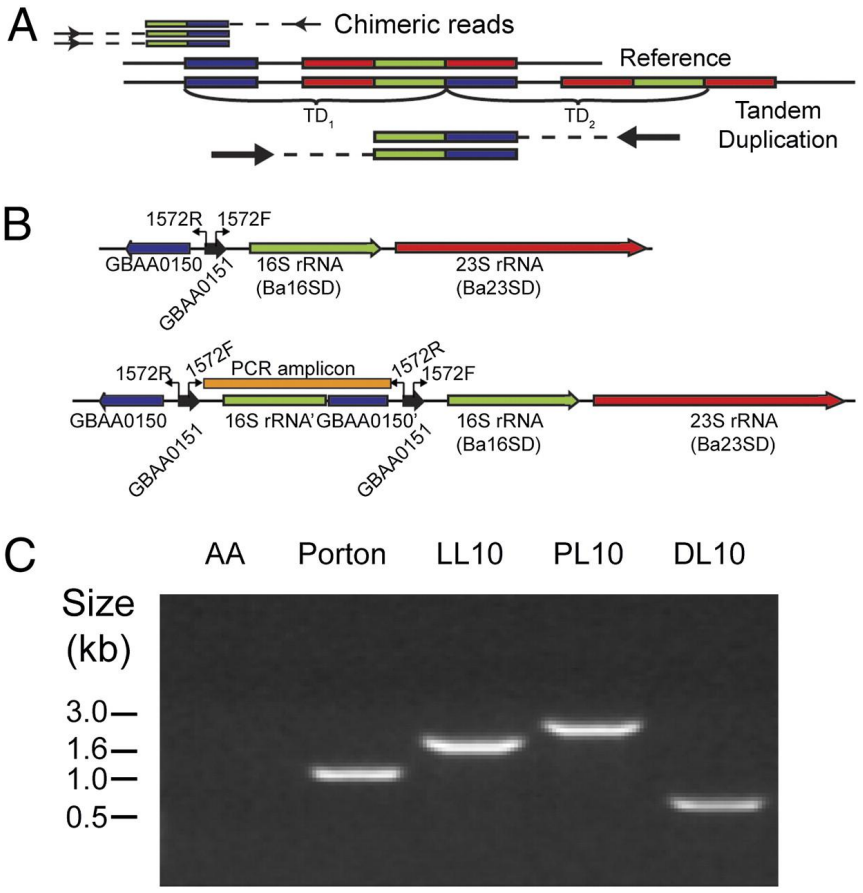


Part IV: The Future

Amerithrax Analysis

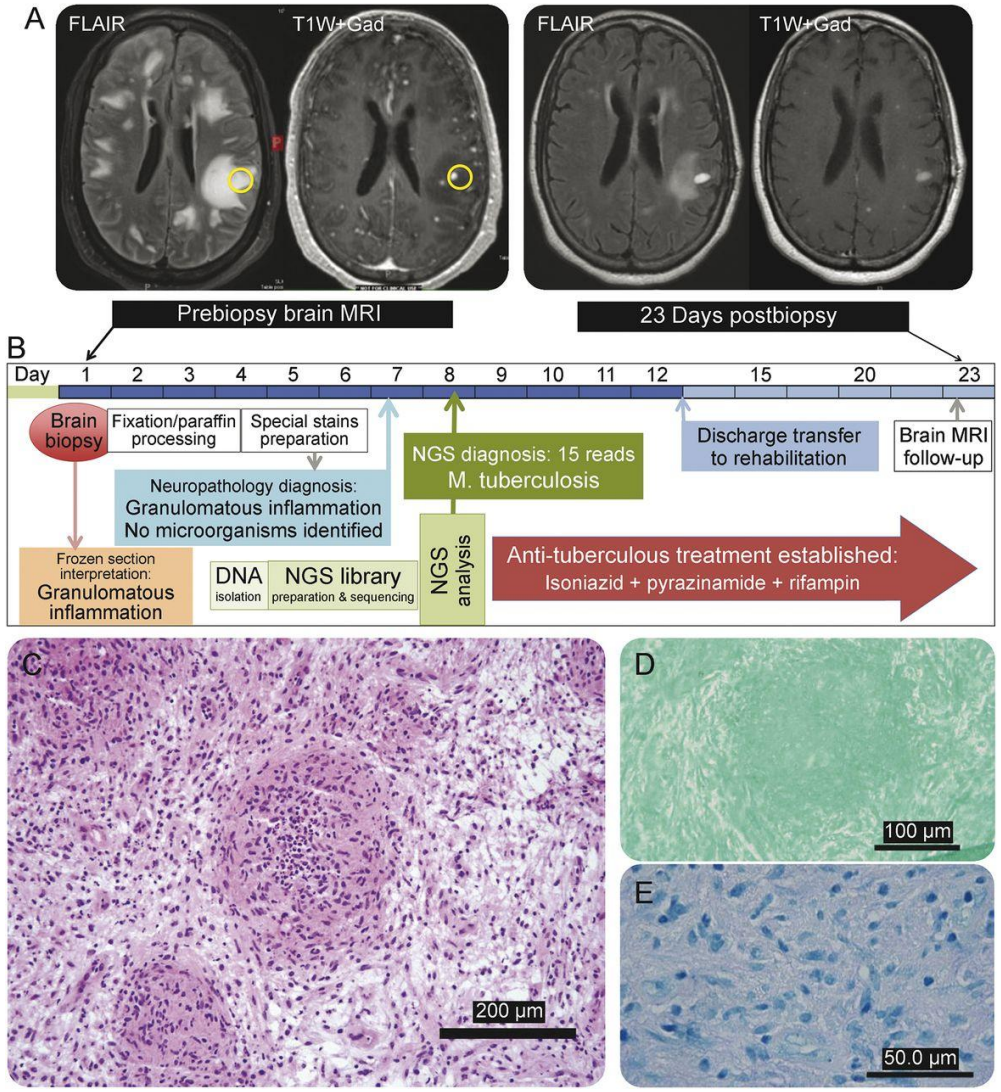






Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation
Rasko et al (2011) PNAS. doi: 10.1073/pnas.1016657108

Diagnosing Brain Infections with NGS



Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system
Salzberg et al (2016) Neurol Neuroimmunol Neuroinflamm dx.doi.org/10.1212/NXI.0000000000000251

Nextstrain / ncov / global

nextstrain.org/ncov/global

JHUMail Daily GRANTS TODO jhu Media Rm Cookies james Other Bookmarks

DOCS HELP LOGIN

Genomic epidemiology of novel coronavirus - Global subsampling

Maintained by the Nextstrain team. Enabled by data from GISAID

Showing 3771 of 3771 genomes sampled between Dec 2019 and Apr 2020.

Dataset
ncov
global

Date Range
2019-12-09 — 2020-04-10

Color By
Country

Tree Options
Layout
RECTANGULAR
RADIAL
UNROOTED
CLOCK

Branch Length
TIME DIVERGENCE

Show confidence intervals

Branch Labels
none

Search Strains

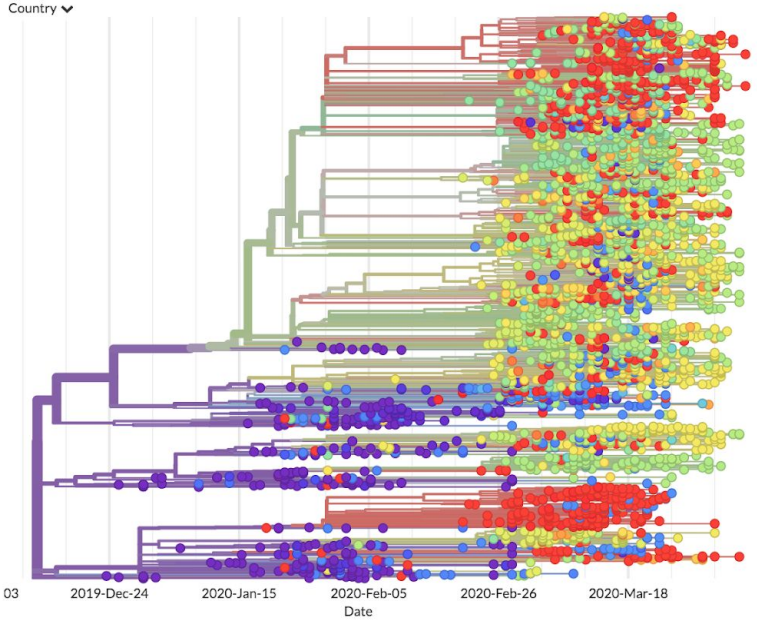
Second Tree
Select...

Map Options
Geographic resolution
country

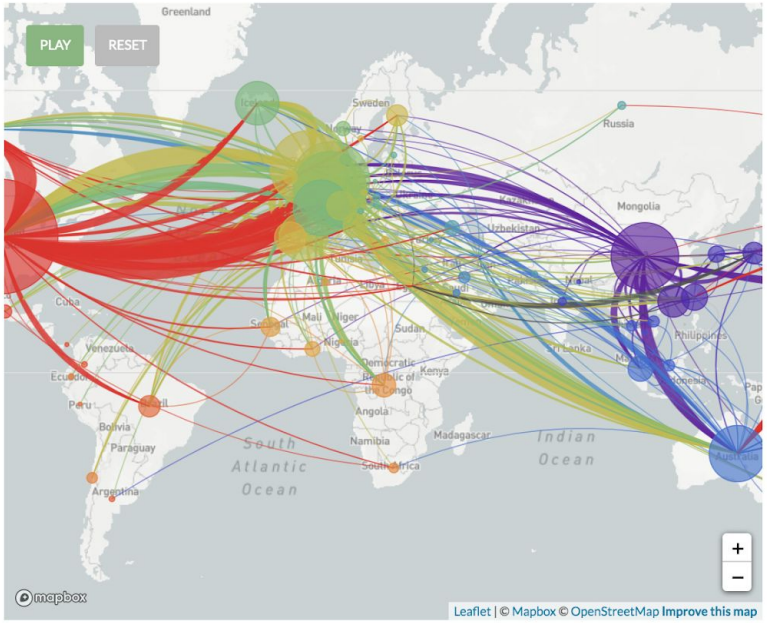
Animation Speed

Phylogeny — RESET LAYOUT

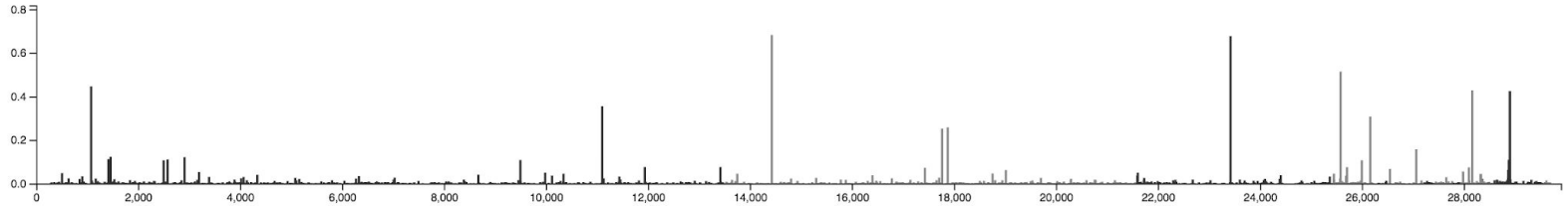
Country



Transmissions — RESET ZOOM



Diversity — ENTROPY EVENTS | AA NT



The Future of Metagenomics

- Applications:
  - WGS metagenomics in the clinic for anaerobic infections and high risk patients (NICU etc.)
  - Surveillance: bioterror agents and epidemiology

- Methods:
  - Single cell, Hi-C, and long read sequencing
  - Machine learning
  - Computational challenges
    - Species level binning of large datasets
    - Plasmid analysis (antimicrobial resistance genes)
    - Going from associations to specific mechanisms
    - Functional analysis