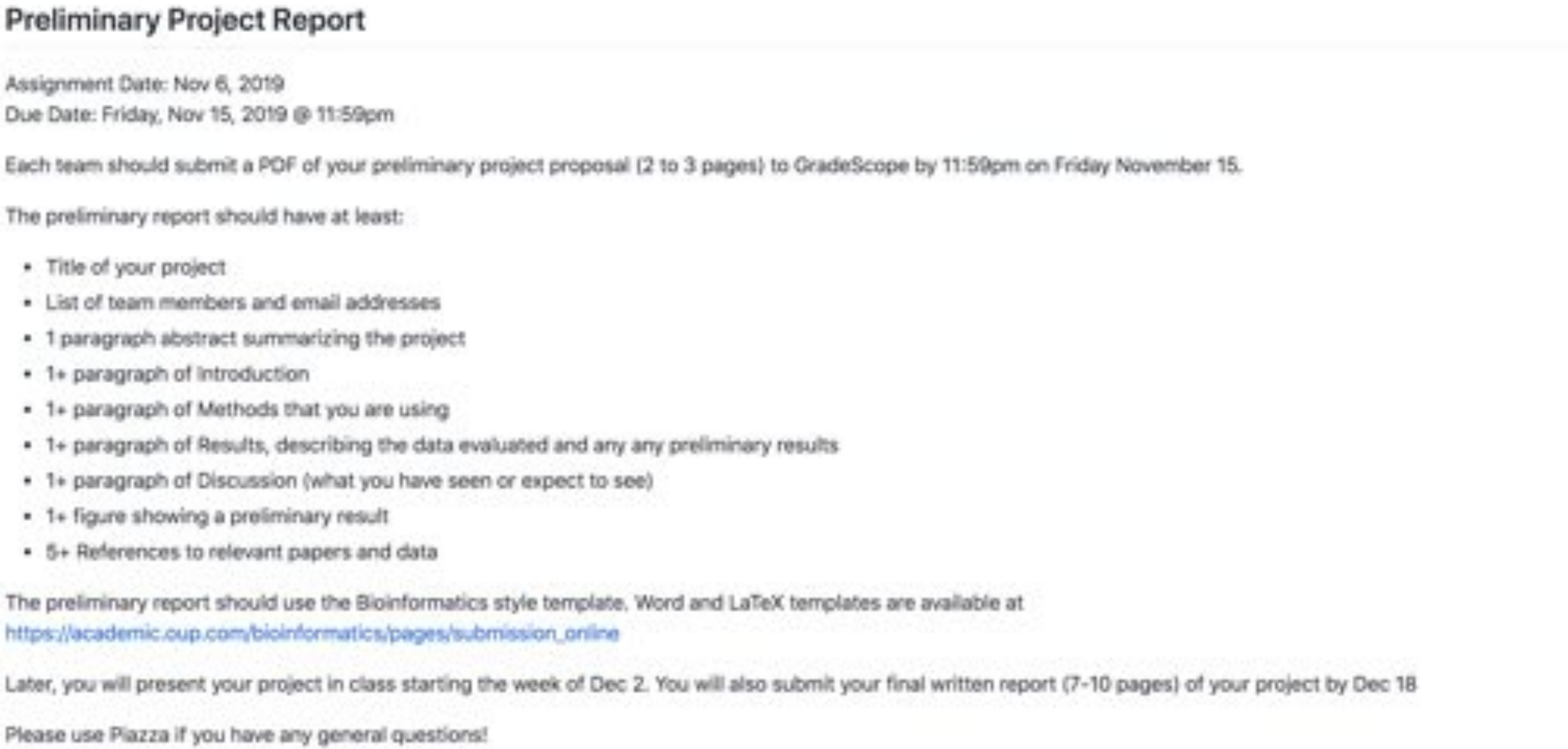
Microbiology & Metagenomics

Michael Schatz

Nov 11, 2019

Lecture 21: Computational Biomedical Research



Preliminary Project Report

Assignment Date: Nov 6, 2019
Due Date: Friday, Nov 15, 2019 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Friday November 15.

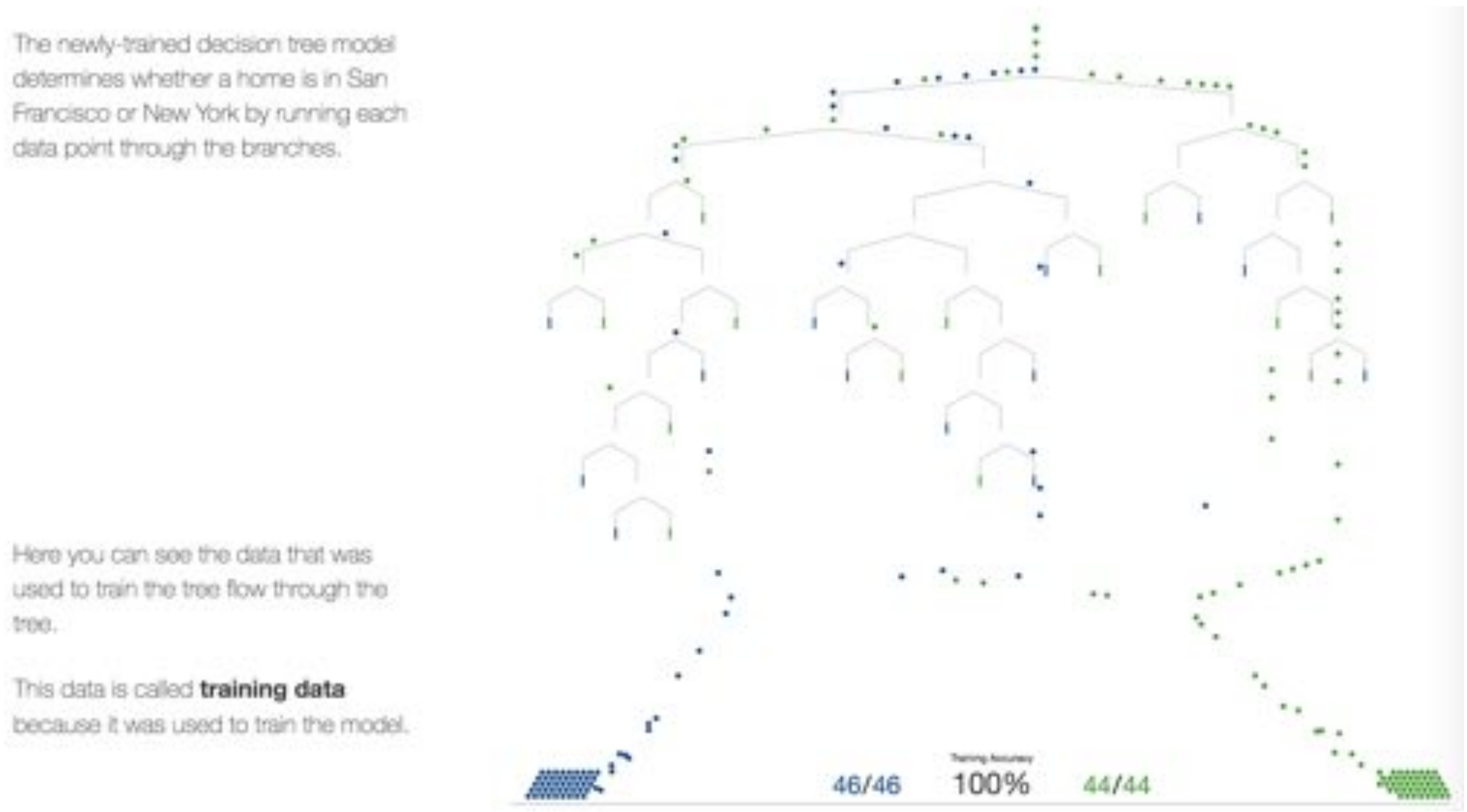
The preliminary report should have at least:

- 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 any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Later, you will present your project in class starting the week of Dec 2. You will also submit your final written report (7-10 pages) of your project by Dec 18

Please use Piazza if you have any general questions!

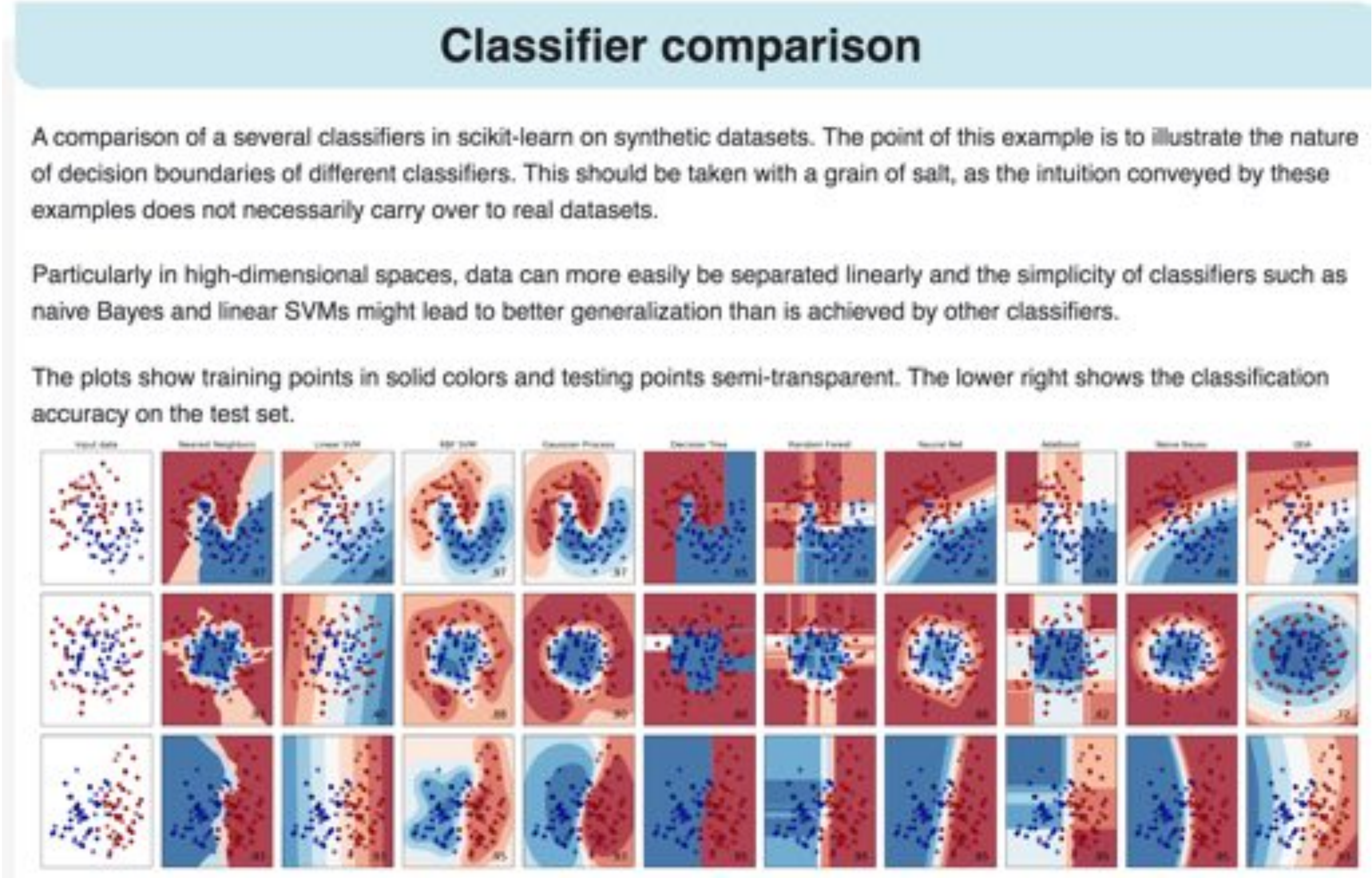


The newly-trained decision tree model determines whether a home is in San Francisco or New York by running each data point through the branches.

Here you can see the data that was used to train the tree flow through the tree.

This data is called **training data** because it was used to train the model.

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/



Classifier comparison

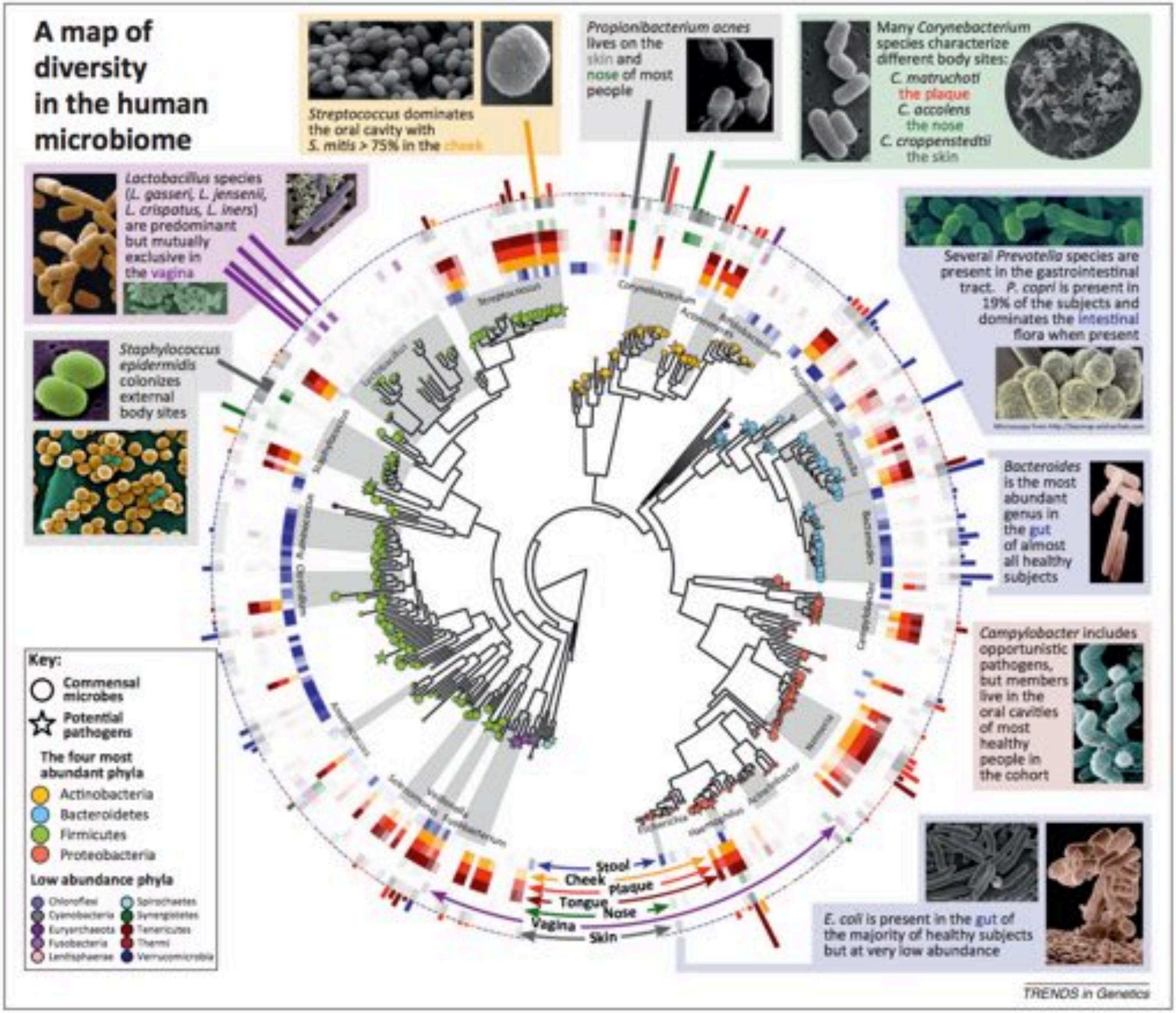
A comparison of a several classifiers in scikit-learn on synthetic datasets. The point of this example is to illustrate the nature of decision boundaries of different classifiers. This should be taken with a grain of salt, as the intuition conveyed by these examples does not necessarily carry over to real datasets.

Particularly in high-dimensional spaces, data can more easily be separated linearly and the simplicity of classifiers such as naive Bayes and linear SVMs might lead to better generalization than is achieved by other classifiers.

The plots show training points in solid colors and testing points semi-transparent. The lower right shows the classification accuracy on the test set.

https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Part I: Introduction



Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites: *C. matruchoti* the plaque *C. accolens* the nose *C. croppenstedtii* the skin

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina

Staphylococcus epidermidis colonizes external body sites

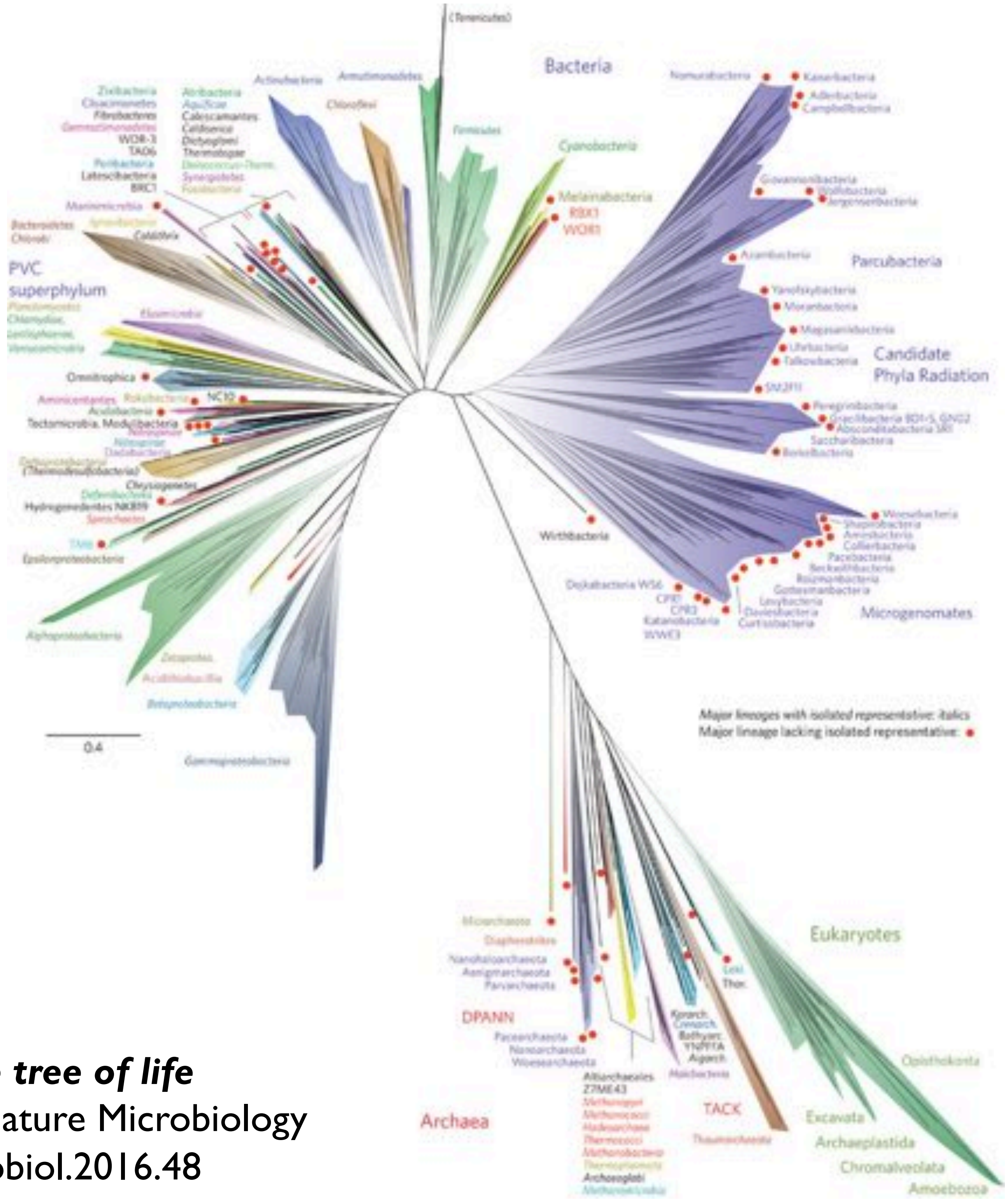
Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort

E. coli is present in the gut of the majority of healthy subjects but at very low abundance

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

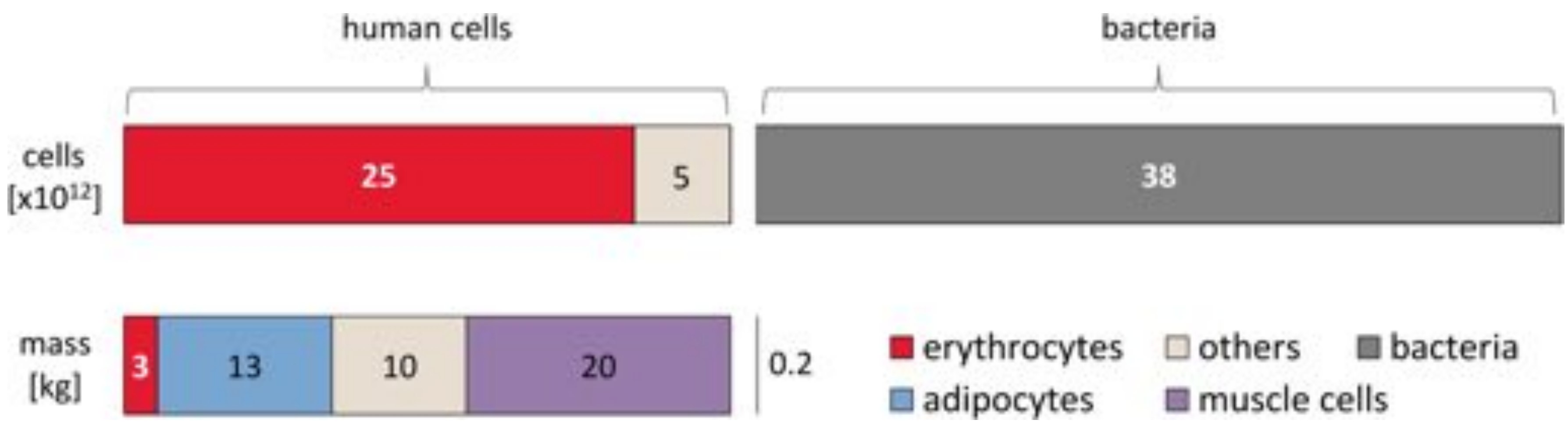
Human brain:
~3.3 lbs

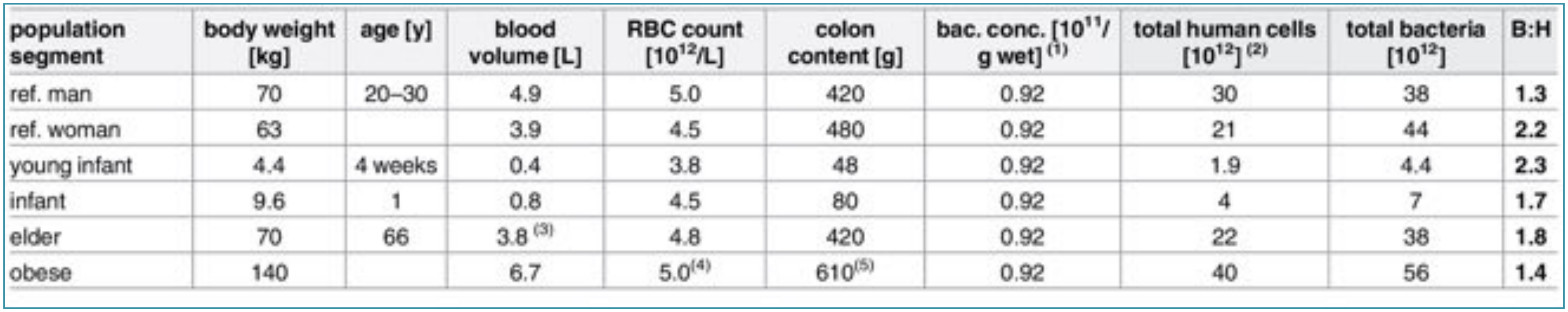
Microbiome
~100 trillion cells

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! ☺

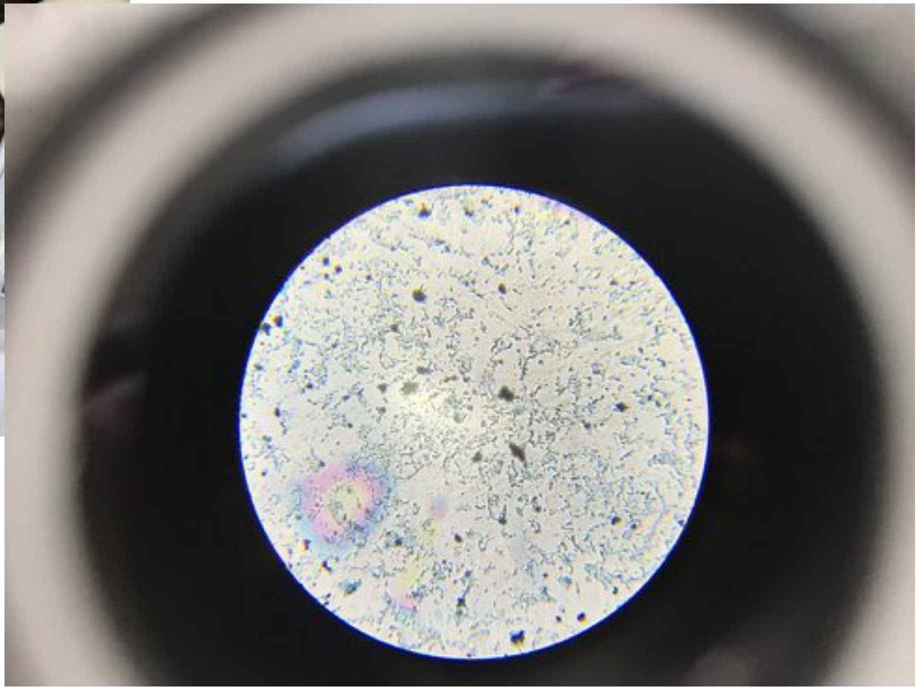




population segment	body weight [kg]	age [y]	blood volume [L]	RBC count [10^{12}/L]	colon content [g]	bac. conc. [10^{11}/ g wet] (1)	total human cells [10^{12}] (2)	total bacteria [10^{12}]	B:H
ref. man	70	20–30	4.9	5.0	420	0.92	30	38	**1.3**
ref. woman	63		3.9	4.5	480	0.92	21	44	**2.2**
young infant	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**

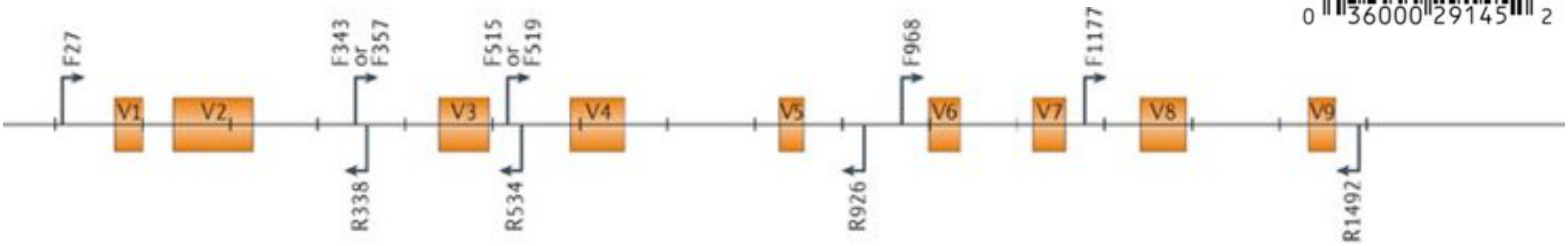
Revised Estimates for the Number of Human and Bacteria Cells in the Body
Sender et al (2016) PLOS Biology. https://doi.org/10.1371/journal.pbio.1002533

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

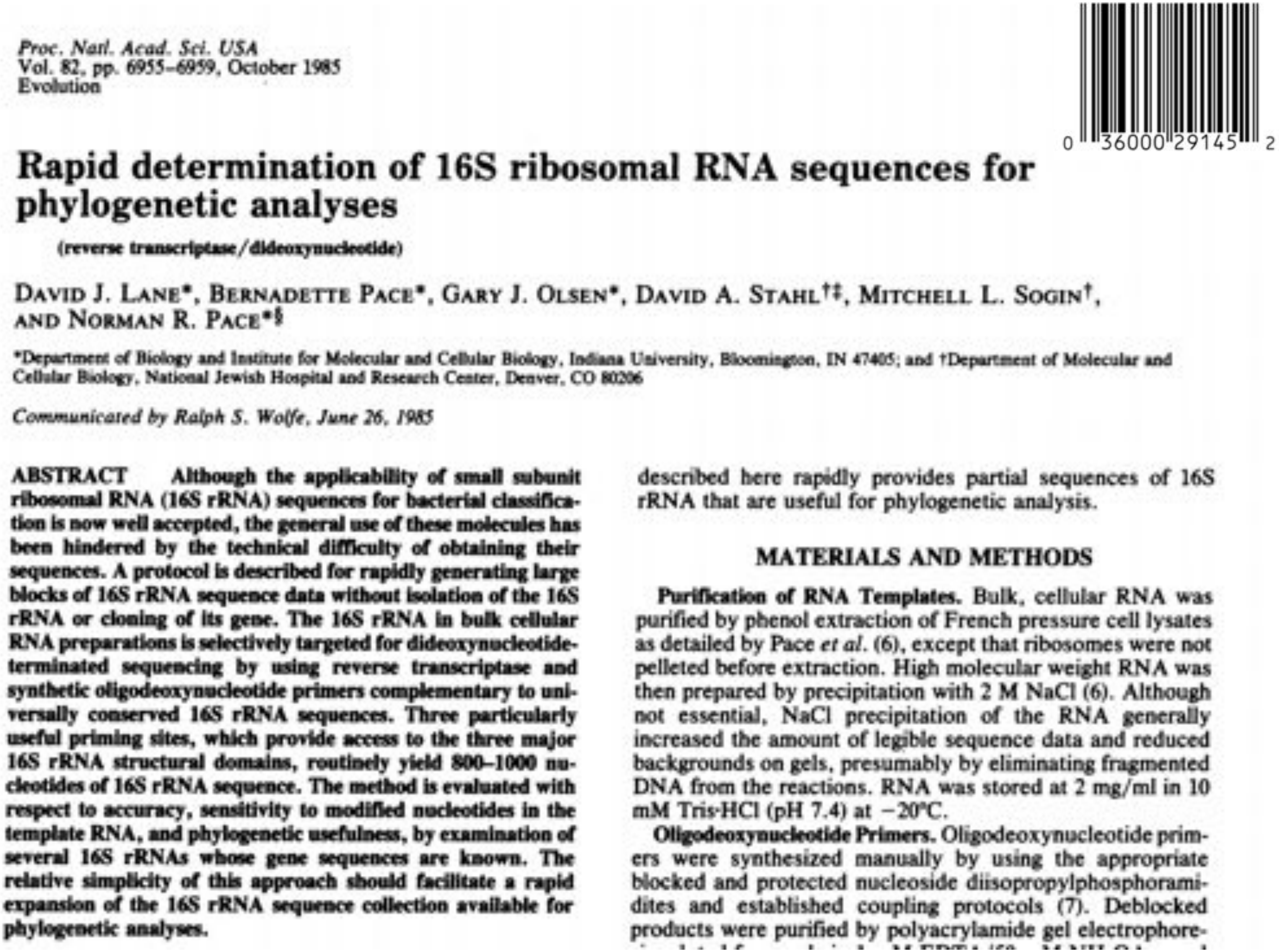
16S rRNA



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



Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 6955–6959, October 1985
Evolution

Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses

(reverse transcriptase/dideoxynucleotide)

DAVID J. LANE[*], BERNADETTE PACE[*], GARY J. OLSEN[*], DAVID A. STAHL[†‡], MITCHELL L. SOGIN[†], AND NORMAN R. PACE[*§]

[*]Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, IN 47405; and [†]Department of Molecular and Cellular Biology, National Jewish Hospital and Research Center, Denver, CO 80206

Communicated by Ralph S. Wolfe, June 26, 1985

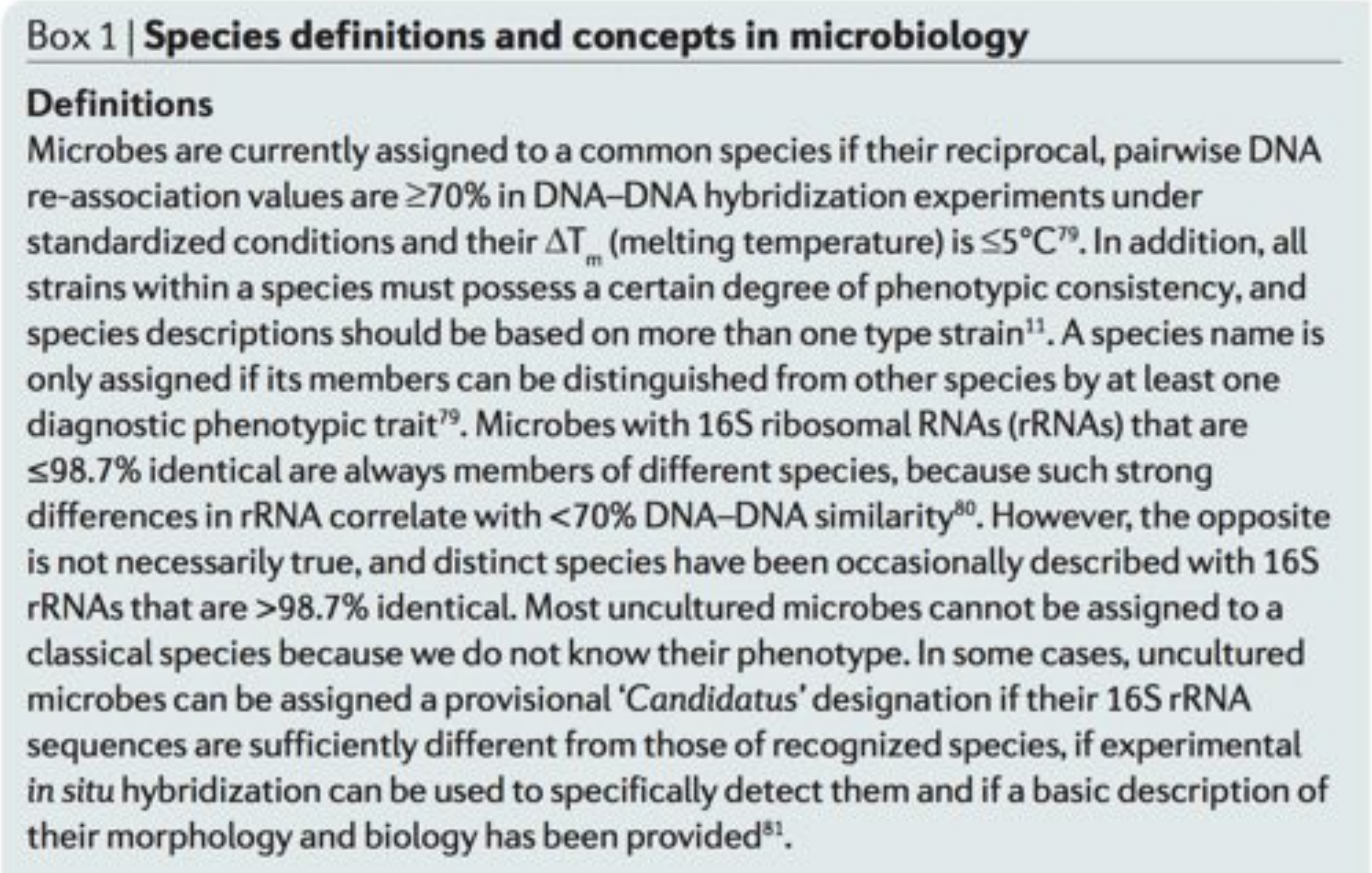
ABSTRACT Although the applicability of small subunit ribosomal RNA (16S rRNA) sequences for bacterial classification is now well accepted, the general use of these molecules has been hindered by the technical difficulty of obtaining their sequences. A protocol is described for rapidly generating large blocks of 16S rRNA sequence data without isolation of the 16S rRNA or cloning of its gene. The 16S rRNA in bulk cellular RNA preparations is selectively targeted for dideoxynucleotide-terminated sequencing by using reverse transcriptase and synthetic oligodeoxynucleotide primers complementary to universally conserved 16S rRNA sequences. Three particularly useful priming sites, which provide access to the three major 16S rRNA structural domains, routinely yield 800–1000 nucleotides of 16S rRNA sequence. The method is evaluated with respect to accuracy, sensitivity to modified nucleotides in the template RNA, and phylogenetic usefulness, by examination of several 16S rRNAs whose gene sequences are known. The relative simplicity of this approach should facilitate a rapid expansion of the 16S rRNA sequence collection available for phylogenetic analyses.

described here rapidly provides partial sequences of 16S rRNA that are useful for phylogenetic analysis.

MATERIALS AND METHODS

Purification of RNA Templates. Bulk, cellular RNA was purified by phenol extraction of French pressure cell lysates as detailed by Pace *et al.* (6), except that ribosomes were not pelleted before extraction. High molecular weight RNA was then prepared by precipitation with 2 M NaCl (6). Although not essential, NaCl precipitation of the RNA generally increased the amount of legible sequence data and reduced backgrounds on gels, presumably by eliminating fragmented DNA from the reactions. RNA was stored at 2 mg/ml in 10 mM Tris·HCl (pH 7.4) at −20°C.

Oligodeoxynucleotide Primers. Oligodeoxynucleotide primers were synthesized manually by using the appropriate blocked and protected nucleoside diisopropylphosphoramidites and established coupling protocols (7). Deblocked products were purified by polyacrylamide gel electrophore-



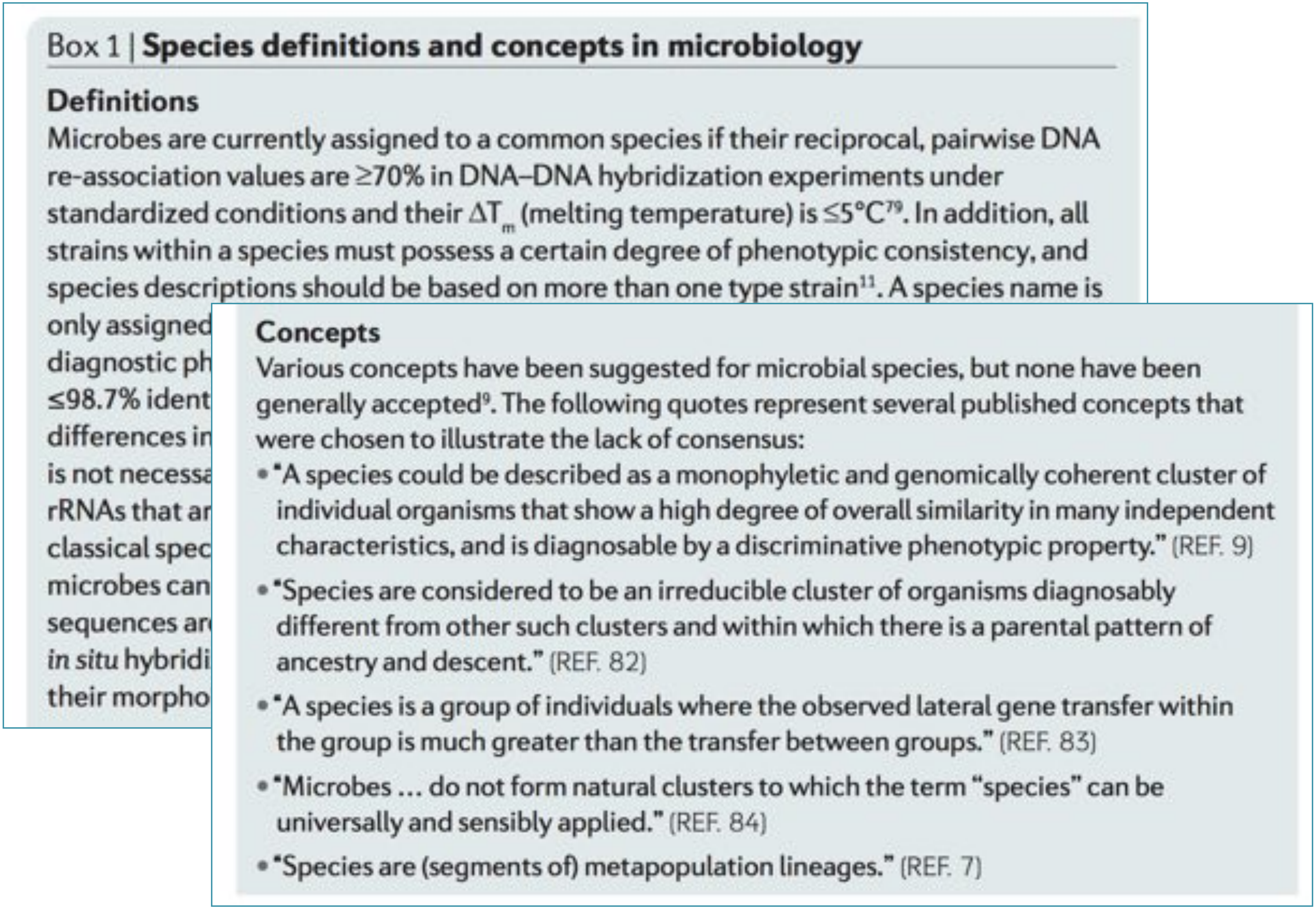
Box 1 | Species definitions and concepts in microbiology

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:

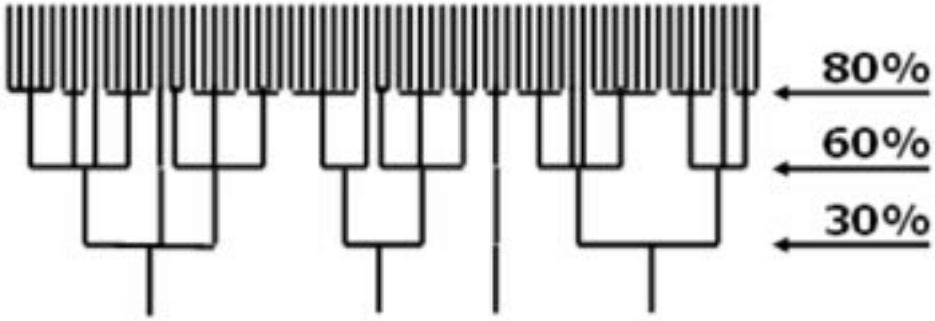
- "A species could be described as a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics, and is diagnosable by a discriminative phenotypic property." (REF. 9)
- "Species are considered to be an irreducible 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.

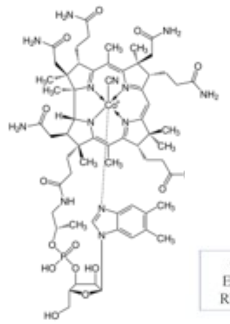
The Importance of Sub-species/strain Level Identification

E. coli

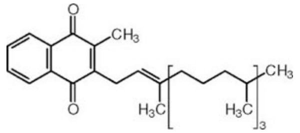
?

Commensal?

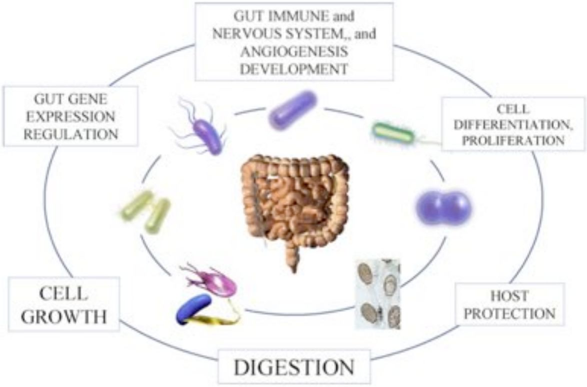
Pathogenic?



Vitamin B12



Vitamin K


GUT IMMUNE and NERVOUS SYSTEM, and ANGIOGENESIS DEVELOPMENT
CELL DIFFERENTIATION, PROLIFERATION
HOST PROTECTION
DIGESTION
CELL GROWTH
GUT GENE EXPRESSION REGULATION

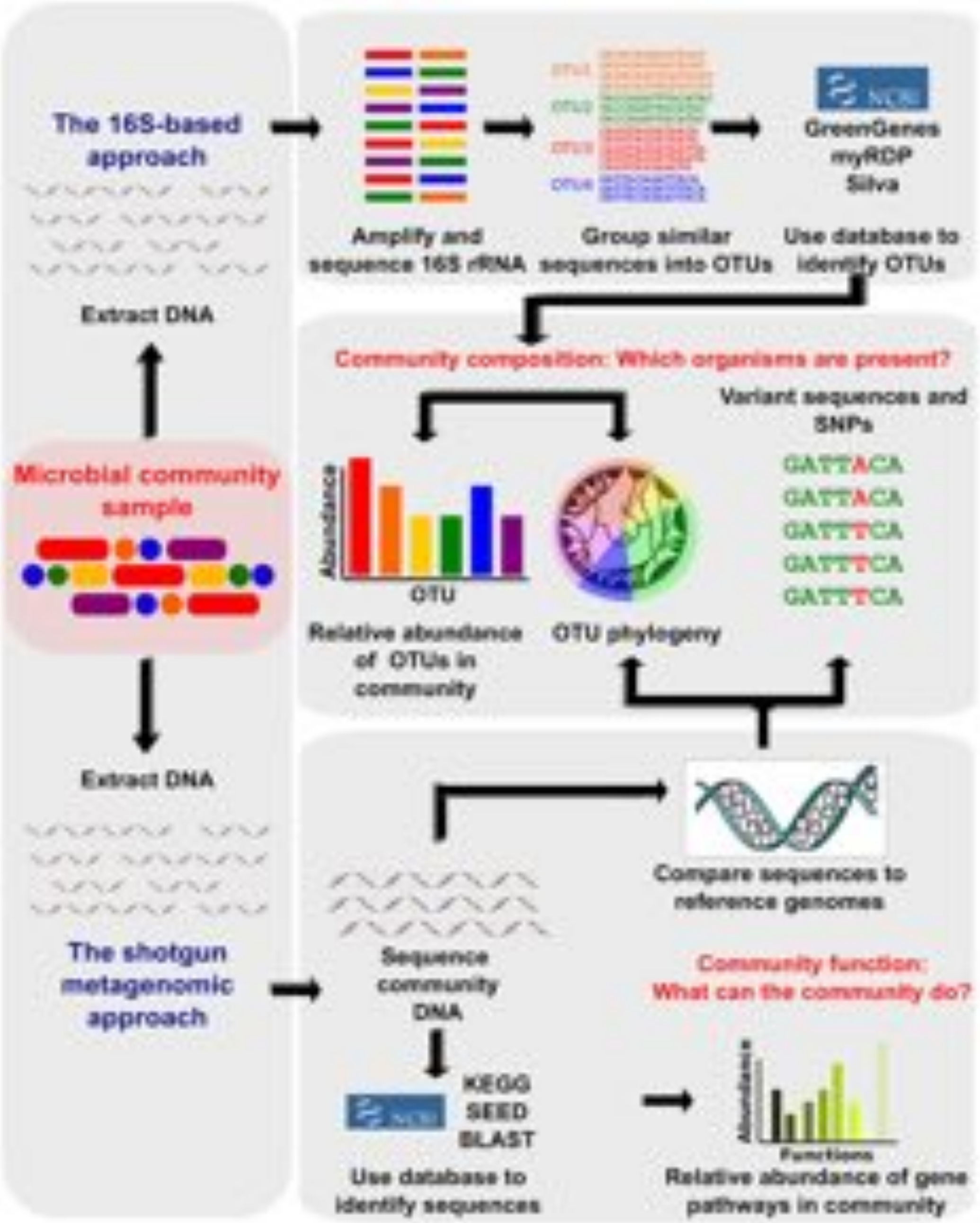
Colonization Barrier Effect Against Pathogens

Strain O157:H7

Hemorrhagi c enteritis

PRIUS

Part II: Methods


The 16S-based approach
Amplify and sequence 16S rRNA
OTU1
OTU2
OTU3
OTU4
Group similar sequences into OTUs
NCBI
GreenGenes
myRDP
Silva
Use database to identify OTUs
Extract DNA
Community composition: Which organisms are present?
Variant sequences and SNPs
GATTACA
GATTACA
GATTTCA
GATTTCA
GATTTCA
Microbial community sample
Abundance
OTU
Relative abundance of OTUs in community
OTU phylogeny
Extract DNA
Compare sequences to reference genomes
The shotgun metagenomic approach
Sequence community DNA
Community function: What can the community do?
NCBI
KEGG
SEED
BLAST
Use database to identify sequences
Abundance
Functions
Relative abundance of gene pathways in community

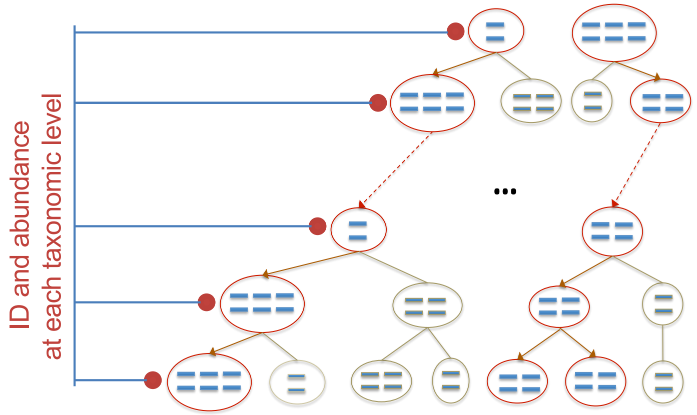
Chapter 12: Human Microbiome Analysis

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

CosmosID Curated Database

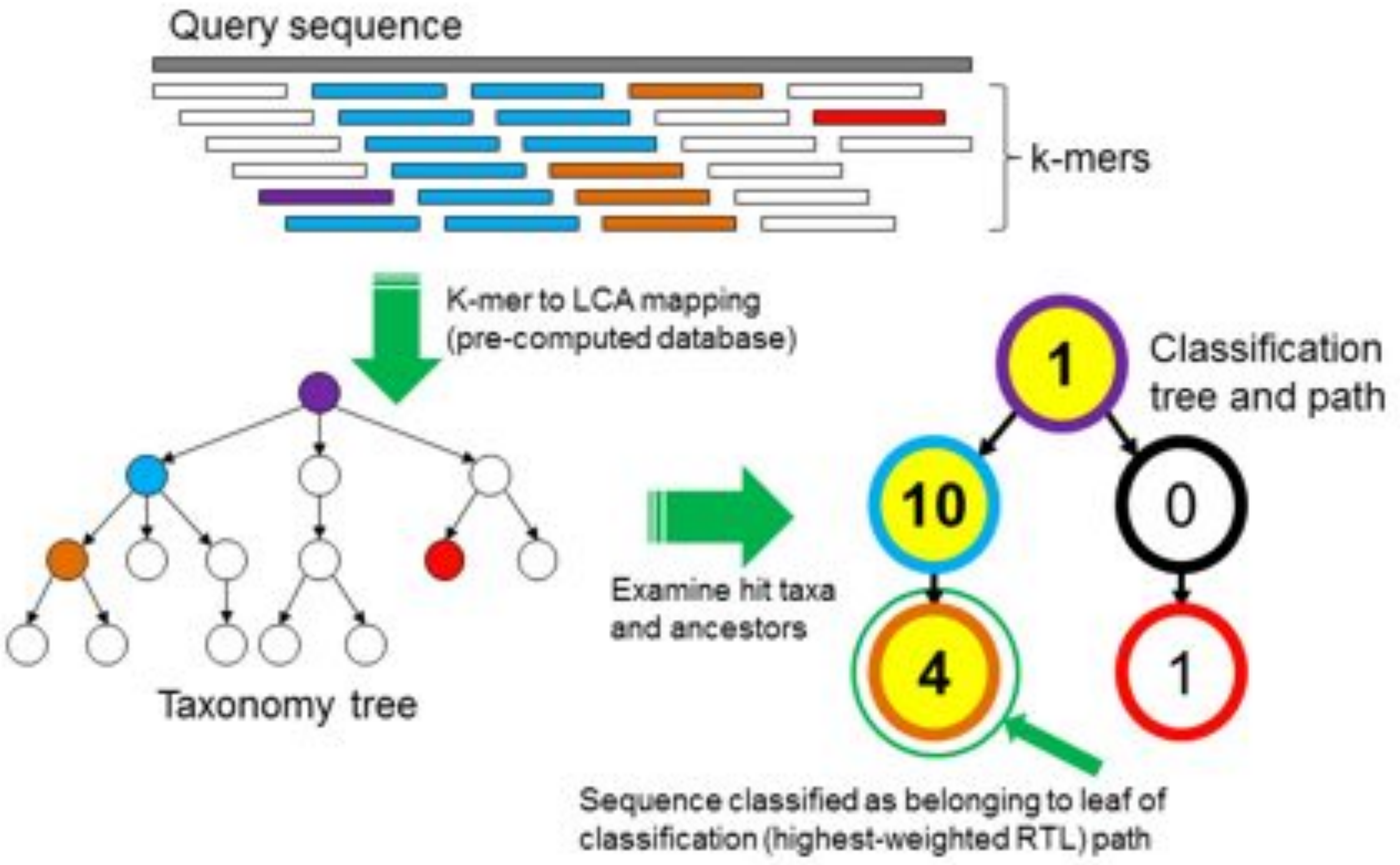
Identify and characterize what matters most to you

- World’s larges and most comprehensive database
  - 10 Years of curation of public and proprietary data
  - 150,000+ genomes and gene sequences
  - commensals, pathogens, and environmental microbes
- Database ontology follows the phylogenetic hierarchy
- Genomic biomarkers uniquely identify microbes at each taxonomic level of a phylogenetic lineage
- Customizable content



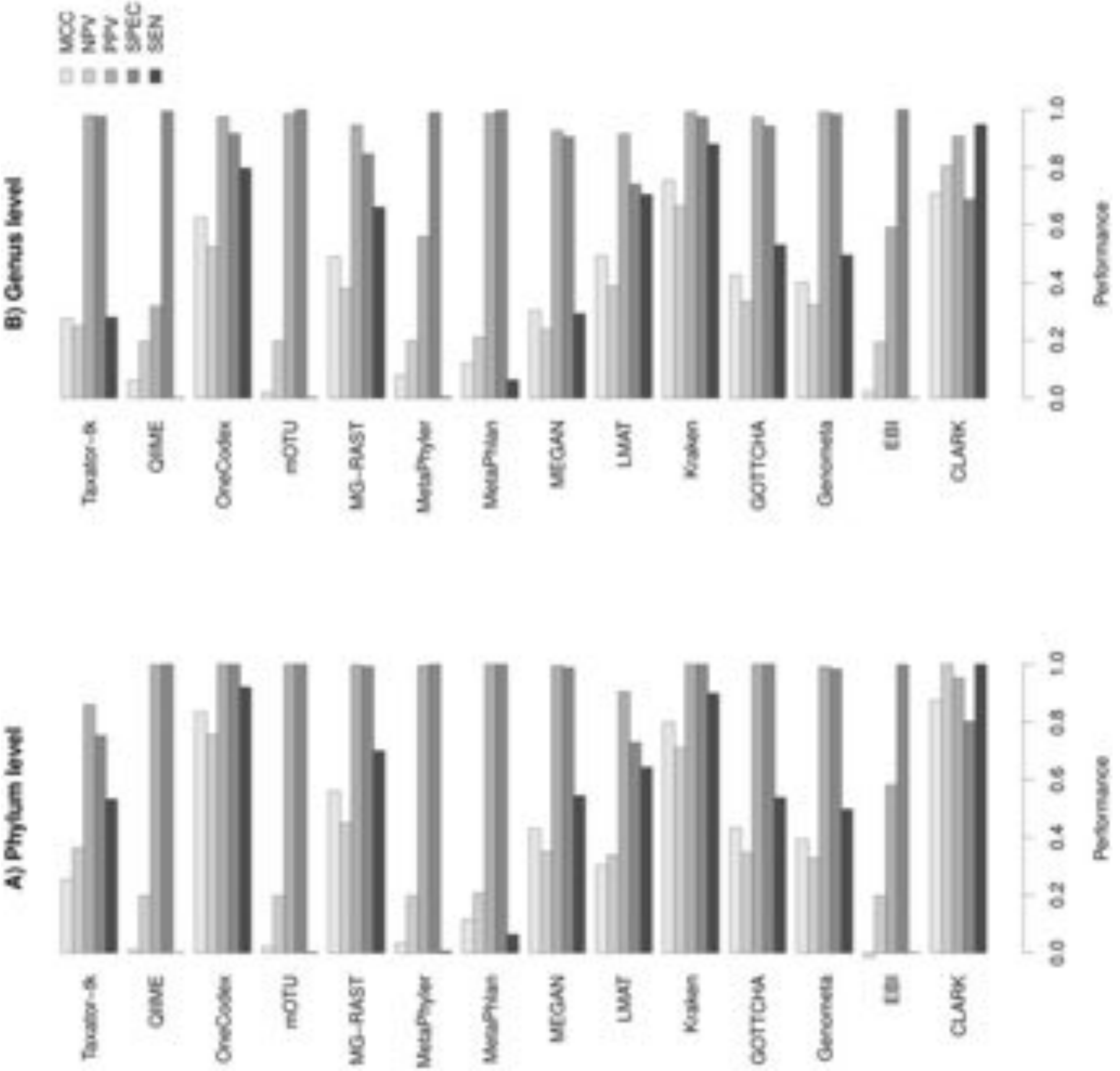
Identify
Bacteria
Viruses
Fungi
Protists
Speciality Databases
(MDR Pathogens & HAI, Biothreat Agents, Food Safety)
Characterize
Antibiotic Resistance
Virulence Factors
Plasmids
Bacteriophage

Kraken


Query sequence
k-mers
K-mer to LCA mapping (pre-computed database)
Taxonomy tree
Examine hit taxa and ancestors
1
10
0
4
1
Classification tree and path
Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

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

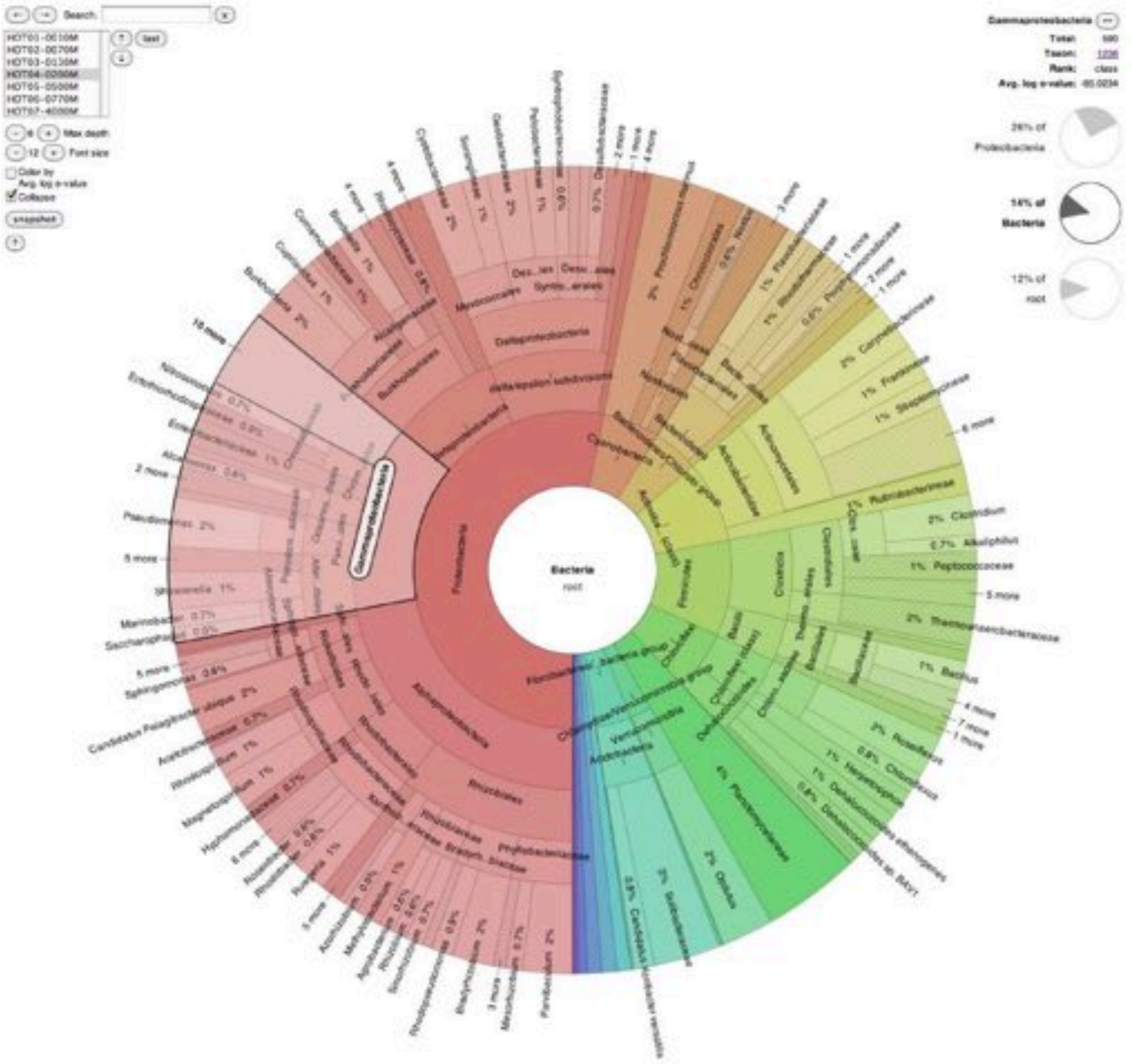
Metagenomics Benchmarking



An evaluation of the accuracy and speed of metagenome analysis tools

Lindgreen et al (2016) Scientific Reports. doi:10.1038/srep19233

Krona Plots



Search:

HOT01-0010M
HOT02-0070M
HOT03-0130M
HOT04-0200M
HOT05-0500M
HOT06-0770M
HOT07-4000M

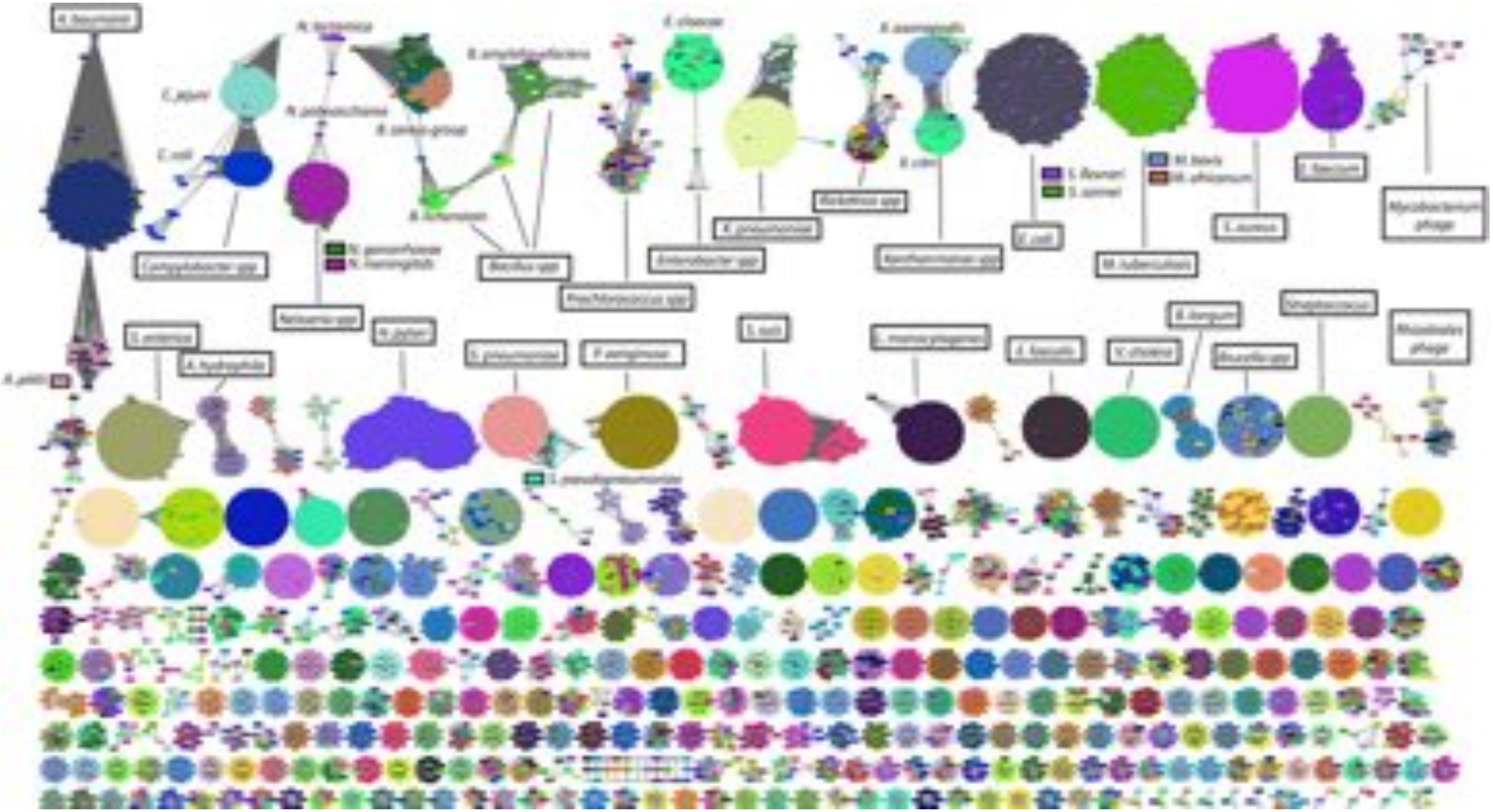
6 Max depth
12 Font size
□ Color by Avg. log e-value
☑ Collapse

Gammaproteobacteria
Total: 590
Taxon: 1236
Rank: class
Avg. log e-value: -65.0234

26% of Proteobacteria
14% of Bacteria
12% of root

Interactive metagenomic visualization in a Web browser
Ondov et al (2011) BMC Bioinformatics. DOI: 10.1186/1471-2105-12-385

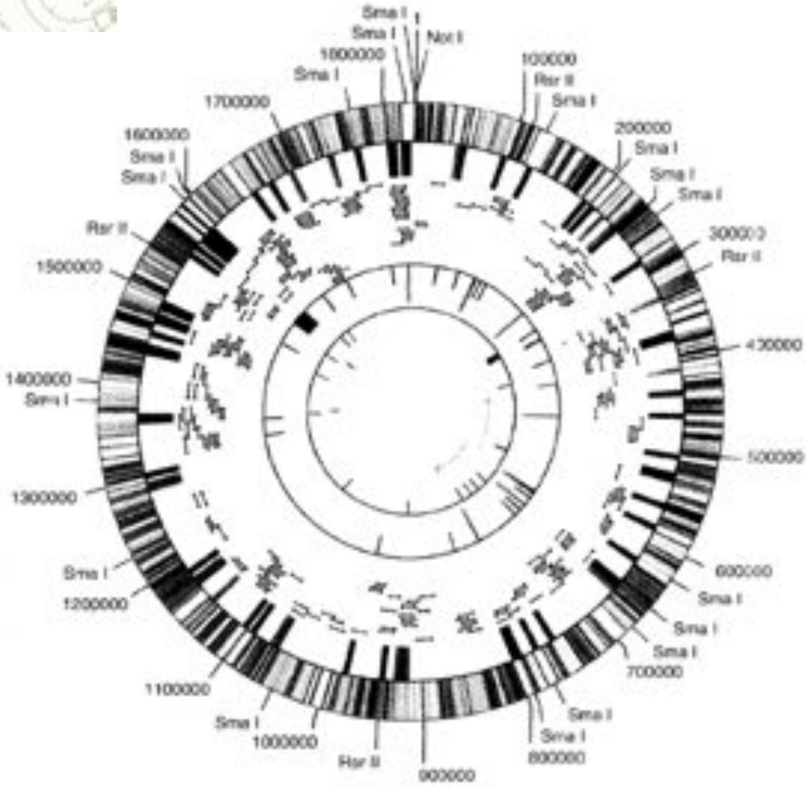
Min-Hash: Comparing all 54,118 RefSeq genomes in 1 day on a laptop



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

Part III: Results

The first microbial genomes



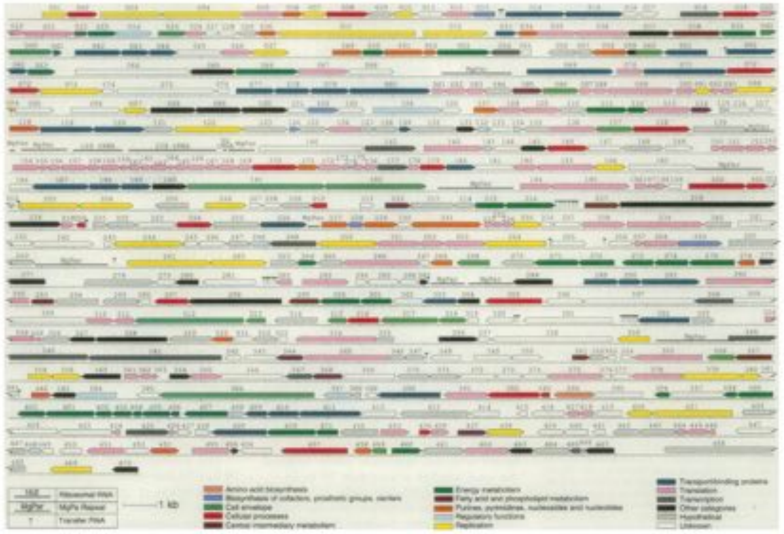


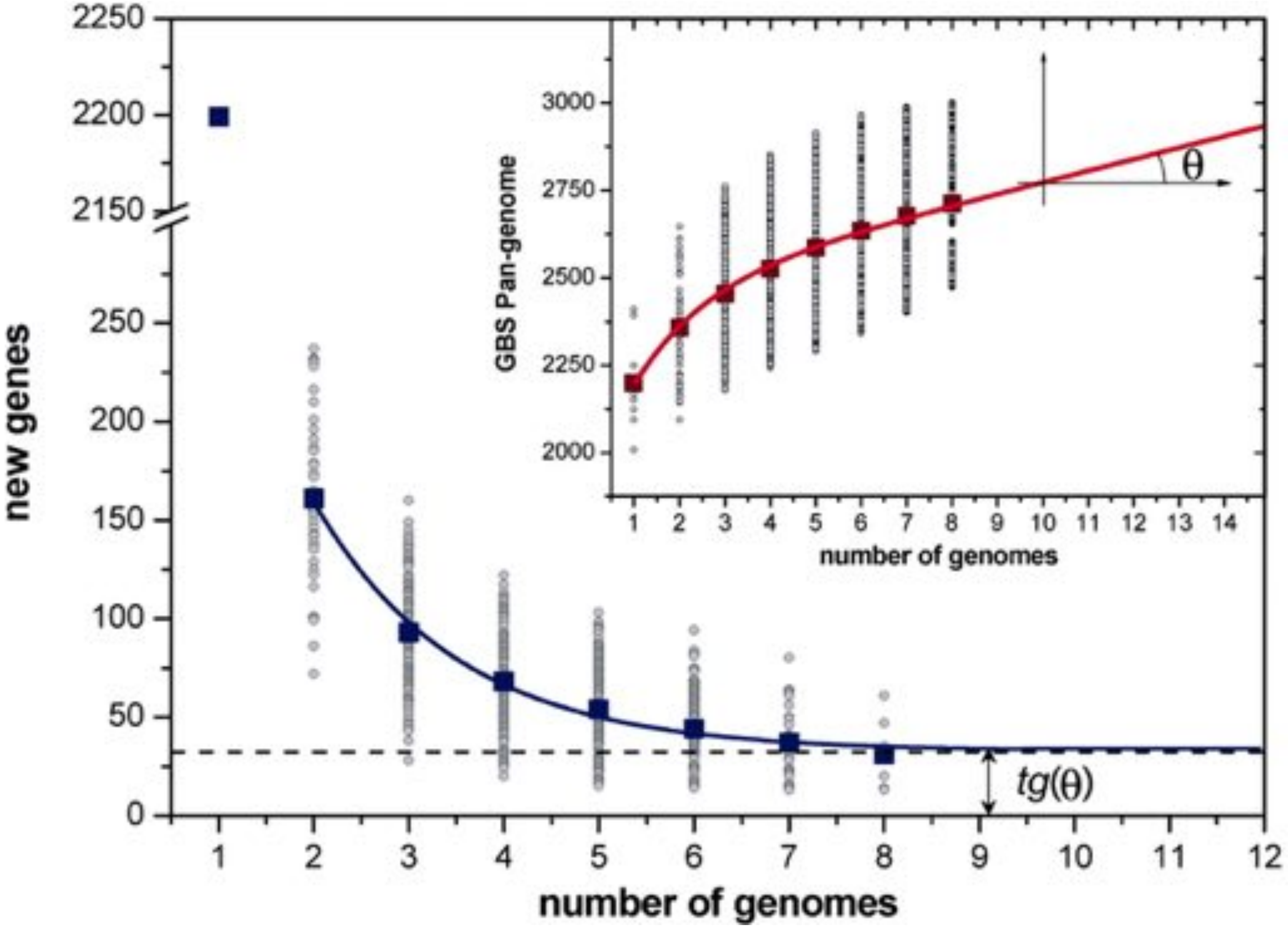
Fig. 1. Gene map of the *M. genitalium* genome. Predicted coding regions are shown, and the direction of transcription is indicated by arrows. Each line in the figure represents 24,000 bp of sequence in the *M. genitalium* genome. Genes are color-coded by role category as described in the key. Gene identificaton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

398 SCIENCE • VOL. 270 • 20 OCTOBER 1995

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

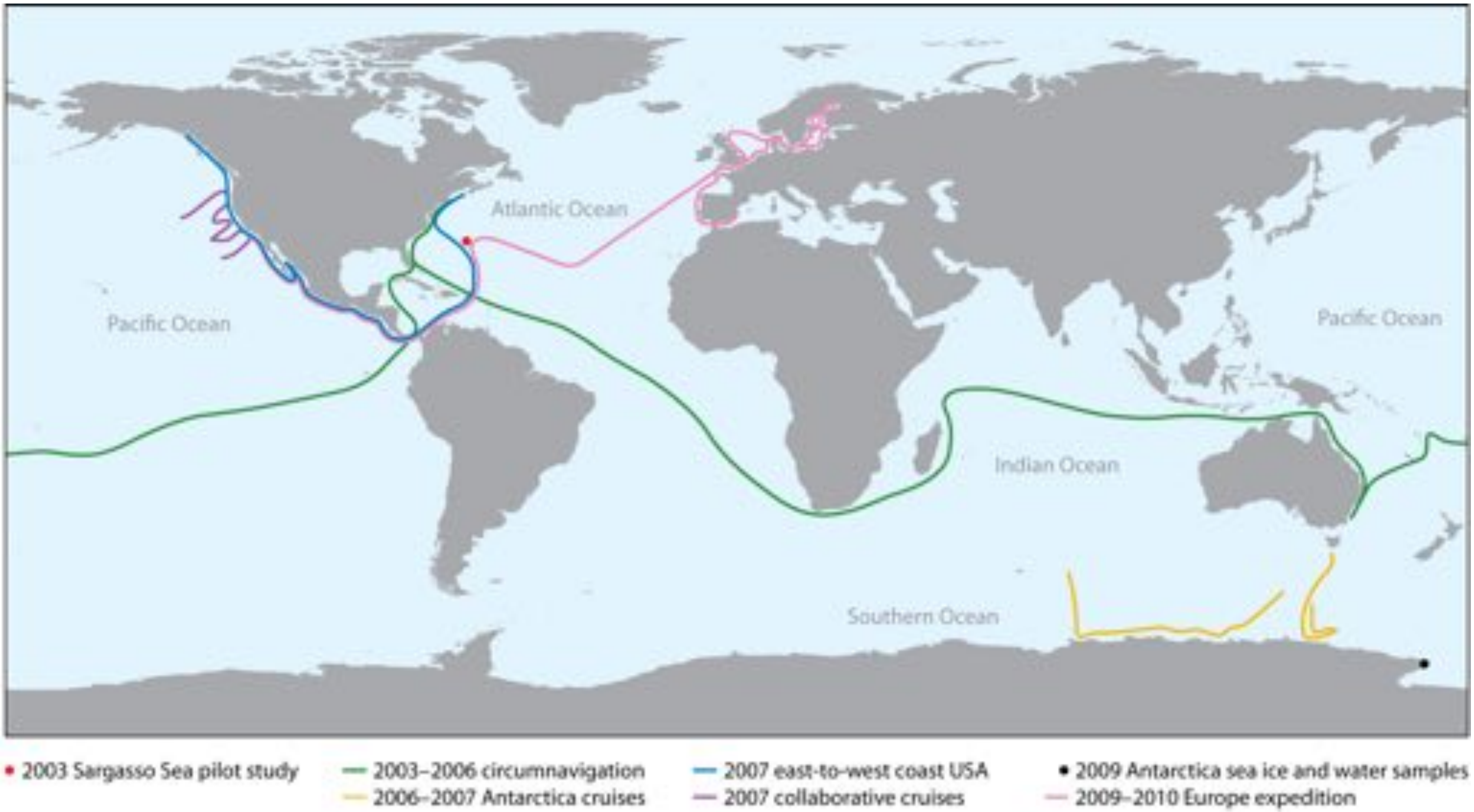
The Minimal Gene Complement of *Mycoplasma genitalium*
Fraiser et al (1995) Science. doi: 10.1126/science.270.5235.397

The first pan genome: *Streptococcus agalactiae*



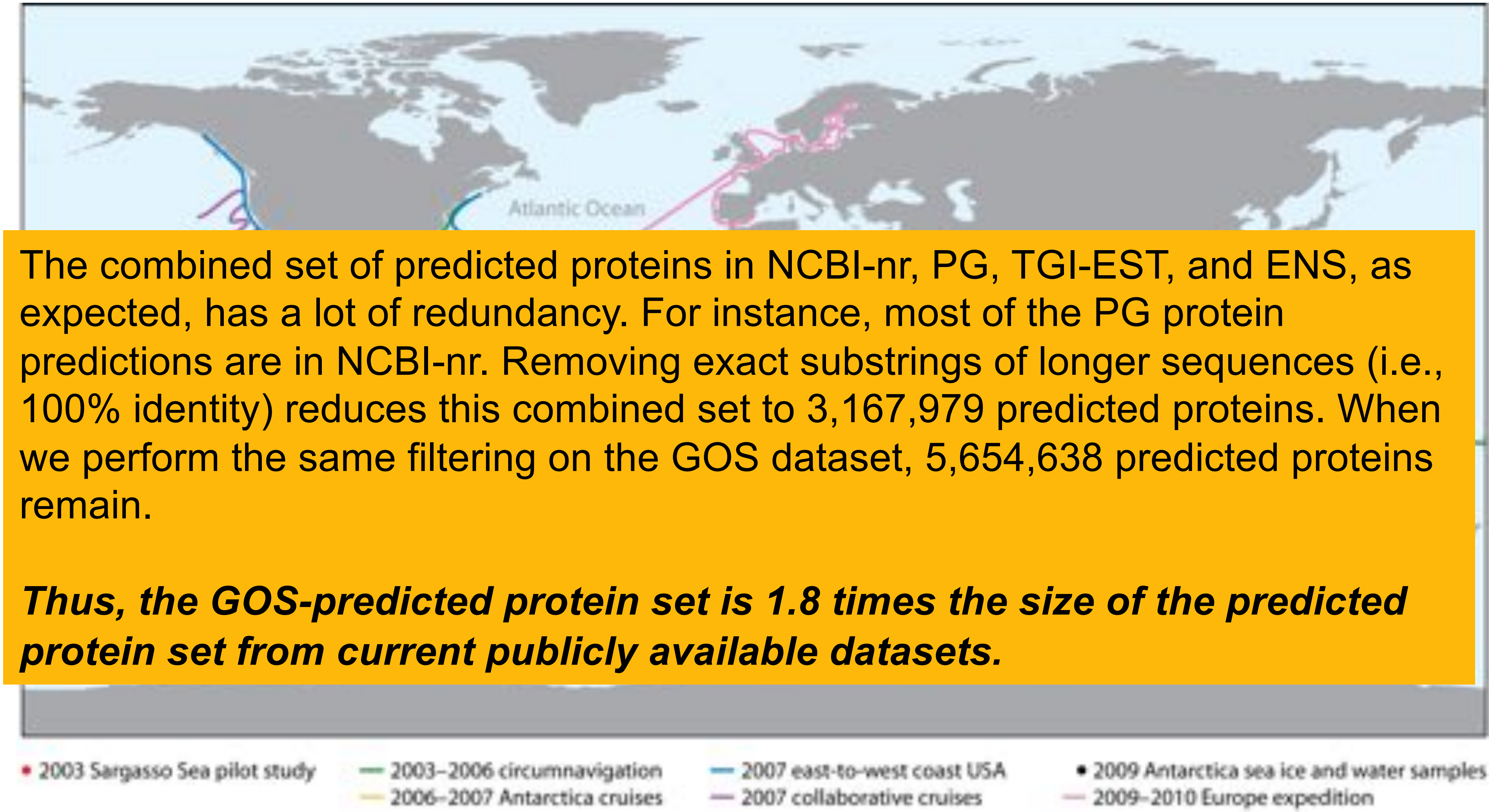
Hervé Tettelin et al. PNAS 2005;102:13950-13955

Global Ocean Survey


Atlantic Ocean
Pacific Ocean
Pacific Ocean
Indian Ocean
Southern Ocean
• 2003 Sargasso Sea pilot study
— 2003–2006 circumnavigation
— 2006–2007 Antarctica cruises
— 2007 east-to-west coast USA
— 2007 collaborative cruises
• 2009 Antarctica sea ice and water samples
— 2009–2010 Europe expedition

Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

Global Ocean Survey

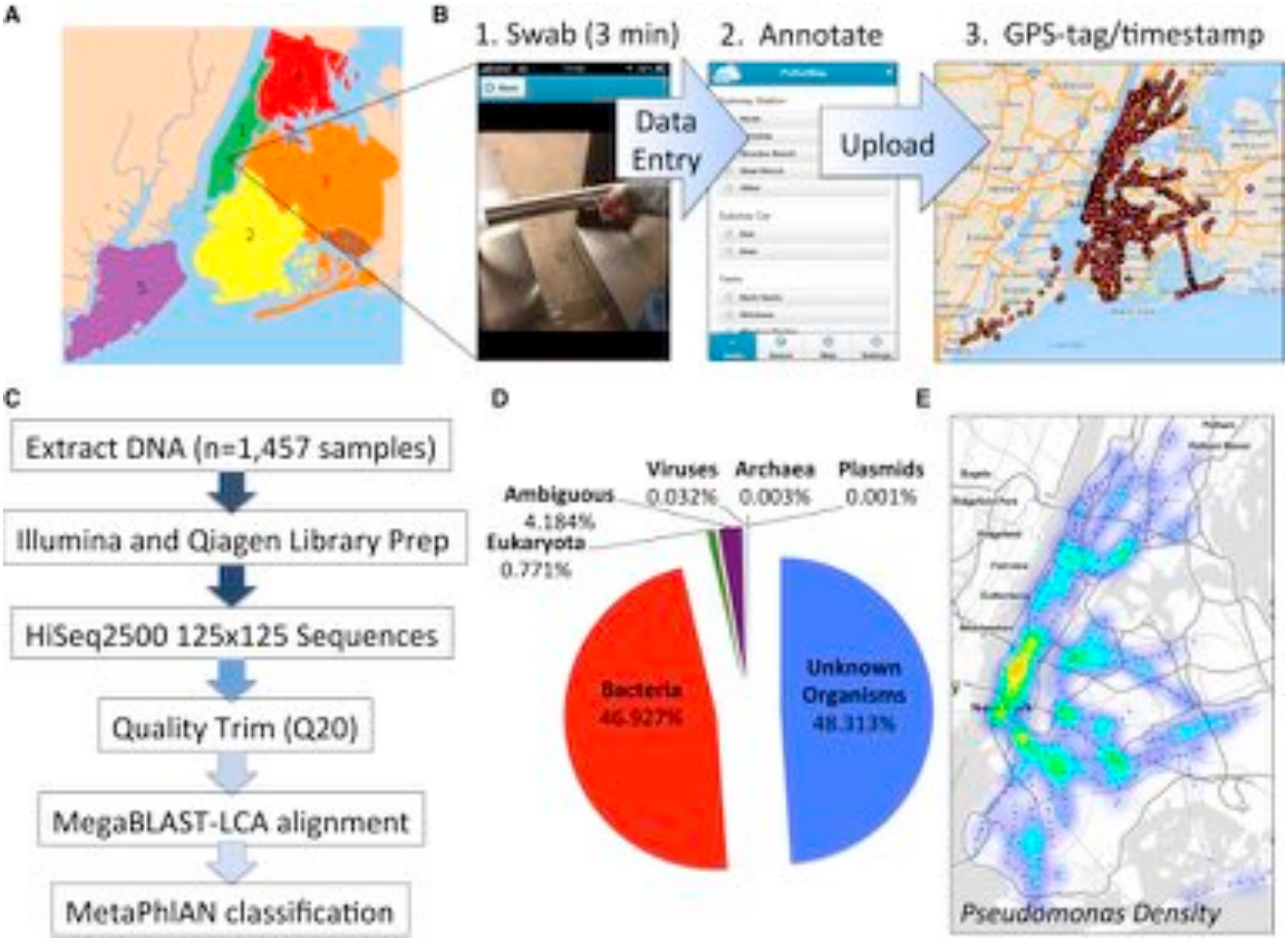

Atlantic Ocean
● 2003 Sargasso Sea pilot study
— 2003–2006 circumnavigation
— 2006–2007 Antarctica cruises
— 2007 east-to-west coast USA
— 2007 collaborative cruises
● 2009 Antarctica sea ice and water samples
— 2009–2010 Europe expedition

The combined set of predicted proteins in NCBI-nr, PG, TGI-EST, and ENS, as expected, has a lot of redundancy. For instance, most of the PG protein predictions are in NCBI-nr. Removing exact substrings of longer sequences (i.e., 100% identity) reduces this combined set to 3,167,979 predicted proteins. When we perform the same filtering on the GOS dataset, 5,654,638 predicted proteins remain.

Thus, the GOS-predicted protein set is 1.8 times the size of the predicted protein set from current publicly available datasets.

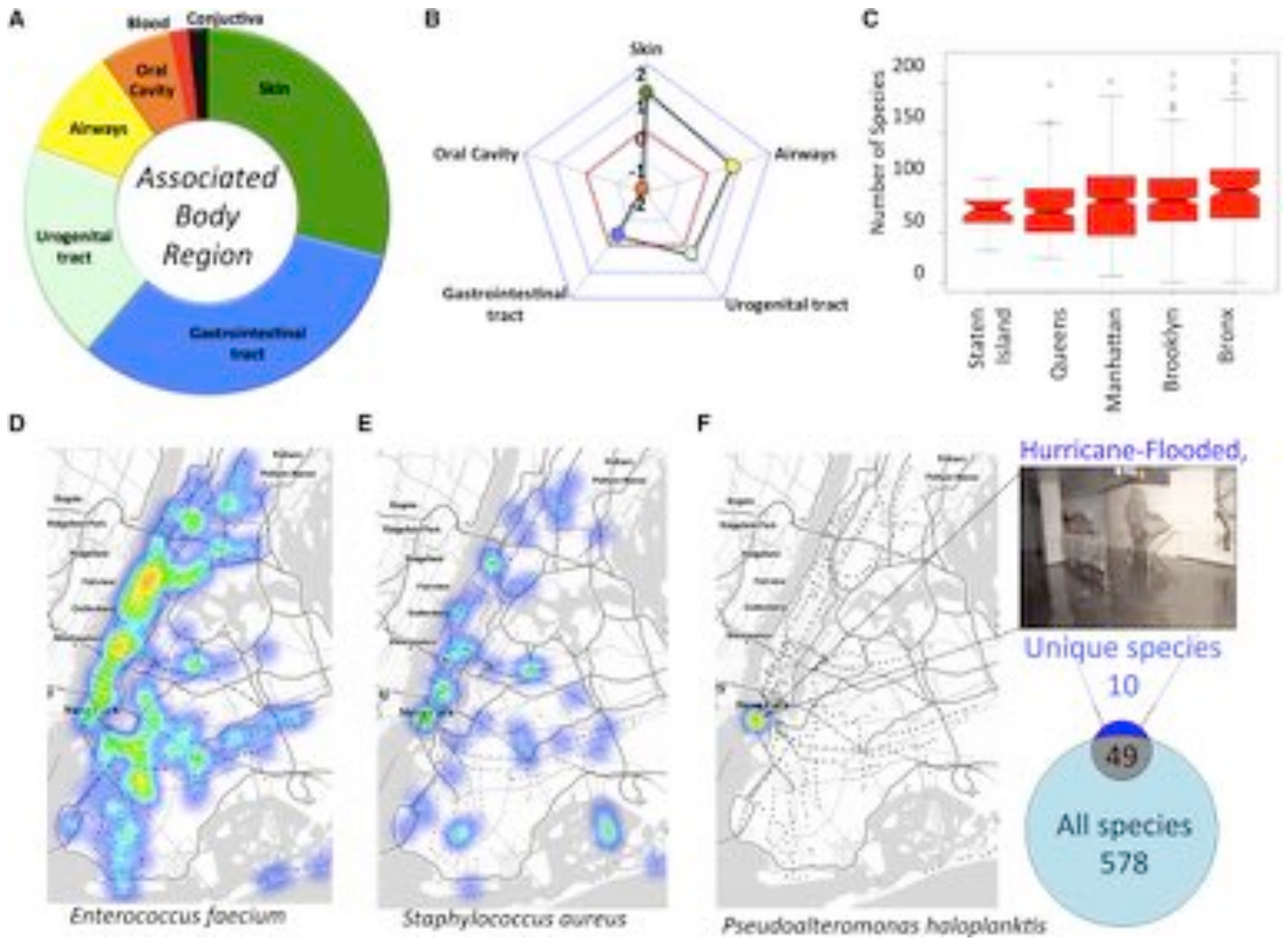
Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

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

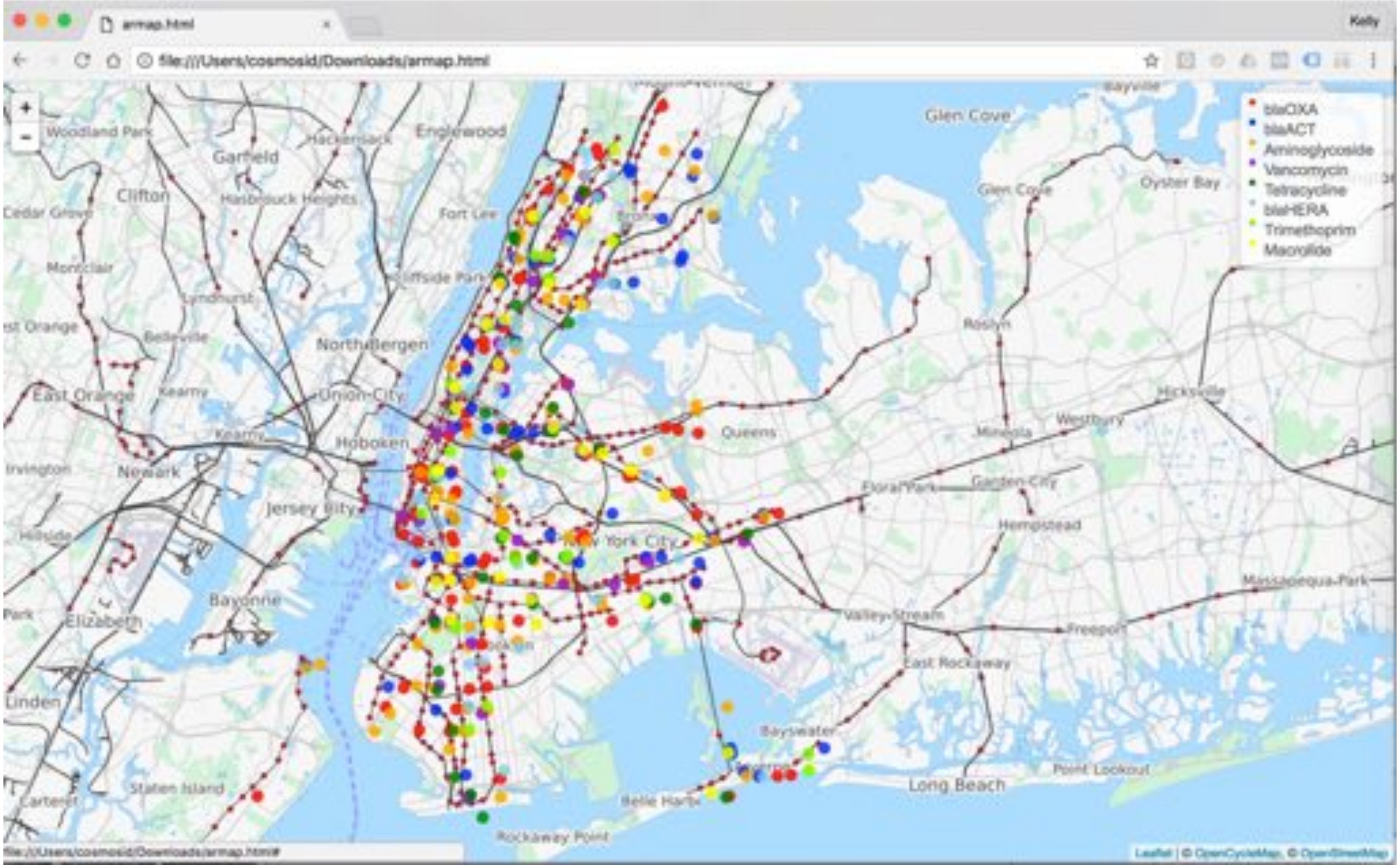
Different subway stations resembled different body sites



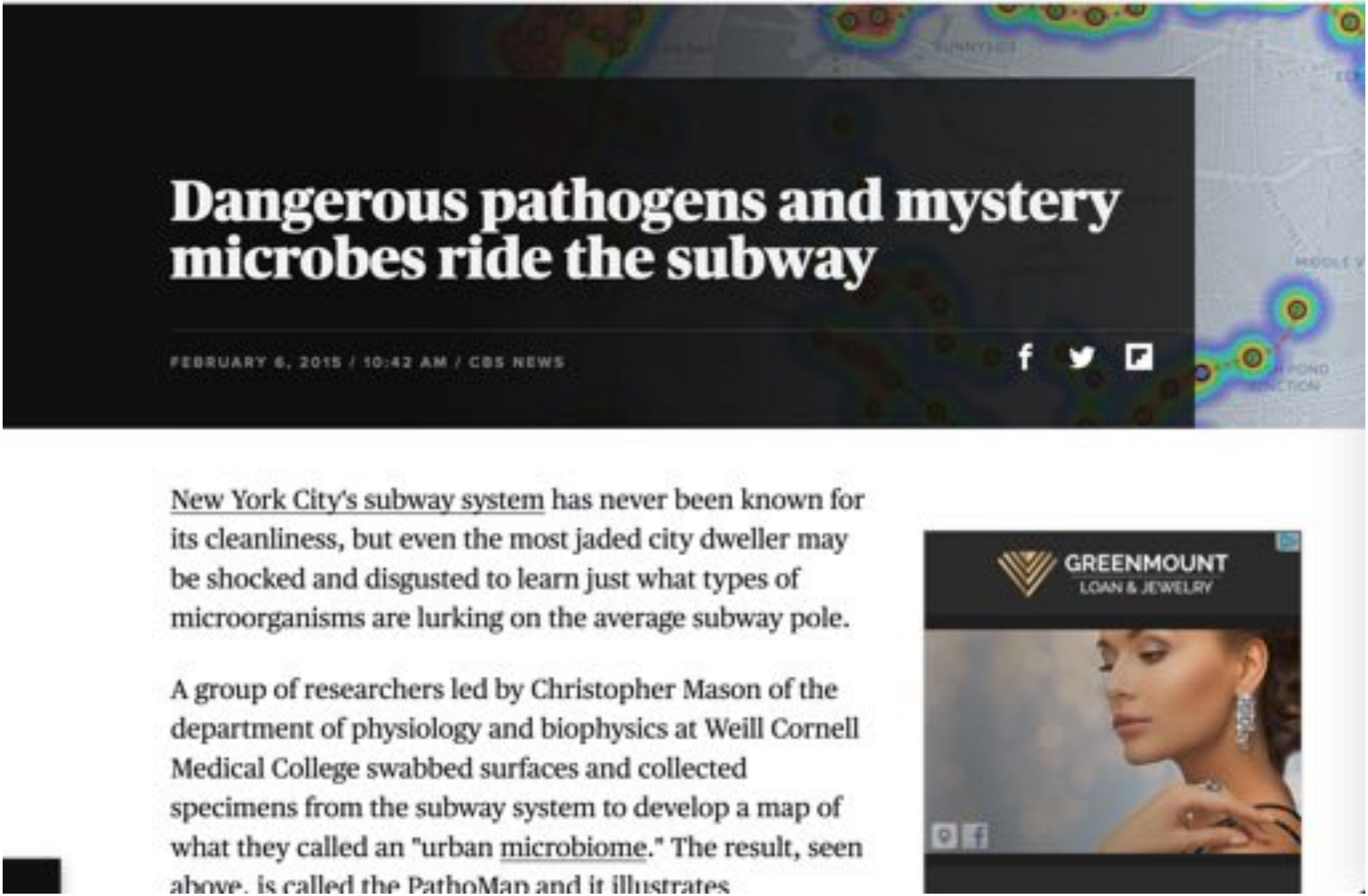
Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics

Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

Mapping Antimicrobial Resistance Factors: PathoMap



Antibiotic resistance genes that were found most frequently in samples were plotted on the map of New York City based on their origin.

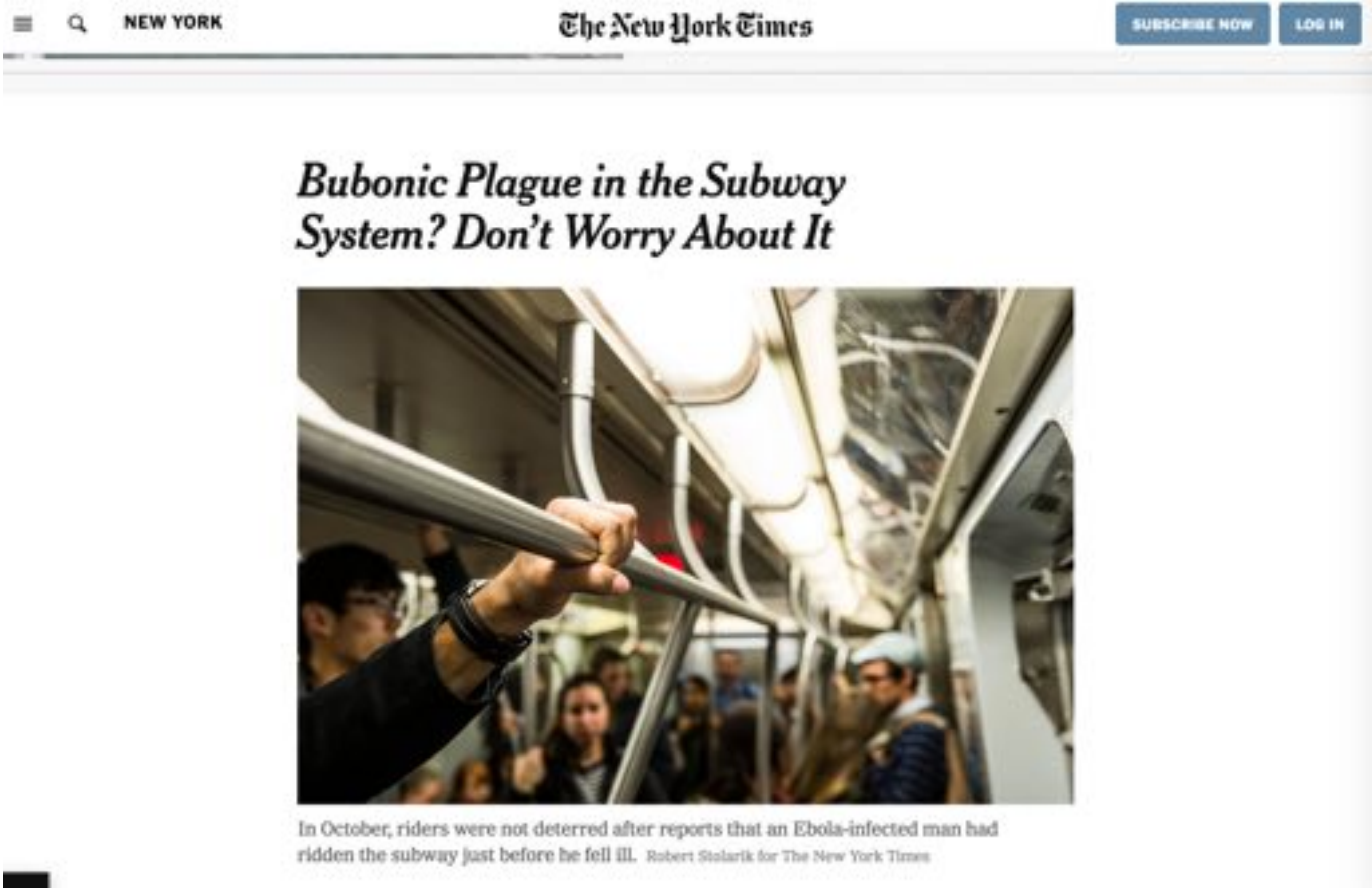


Dangerous pathogens and mystery microbes ride the subway

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

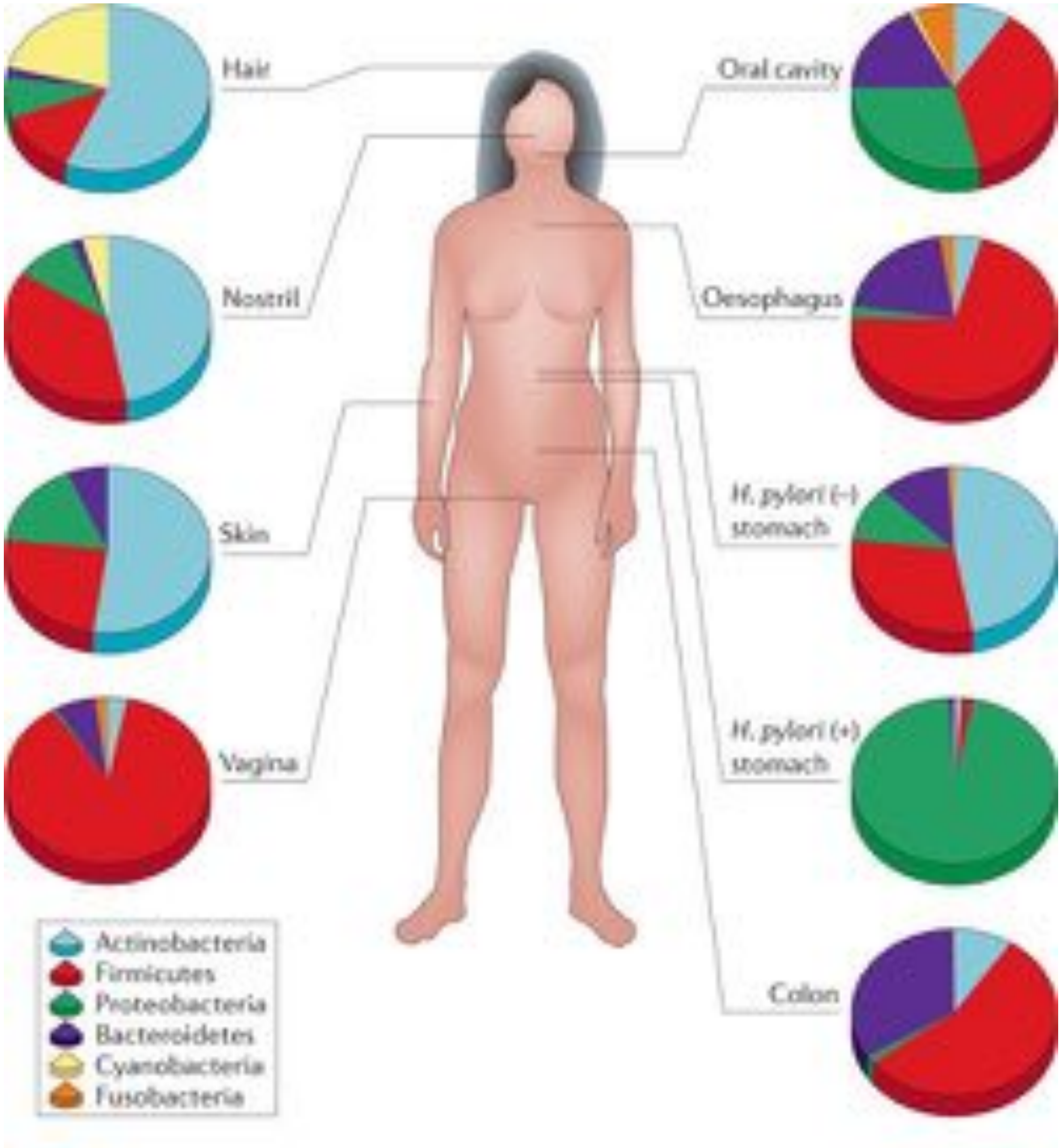
In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

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."

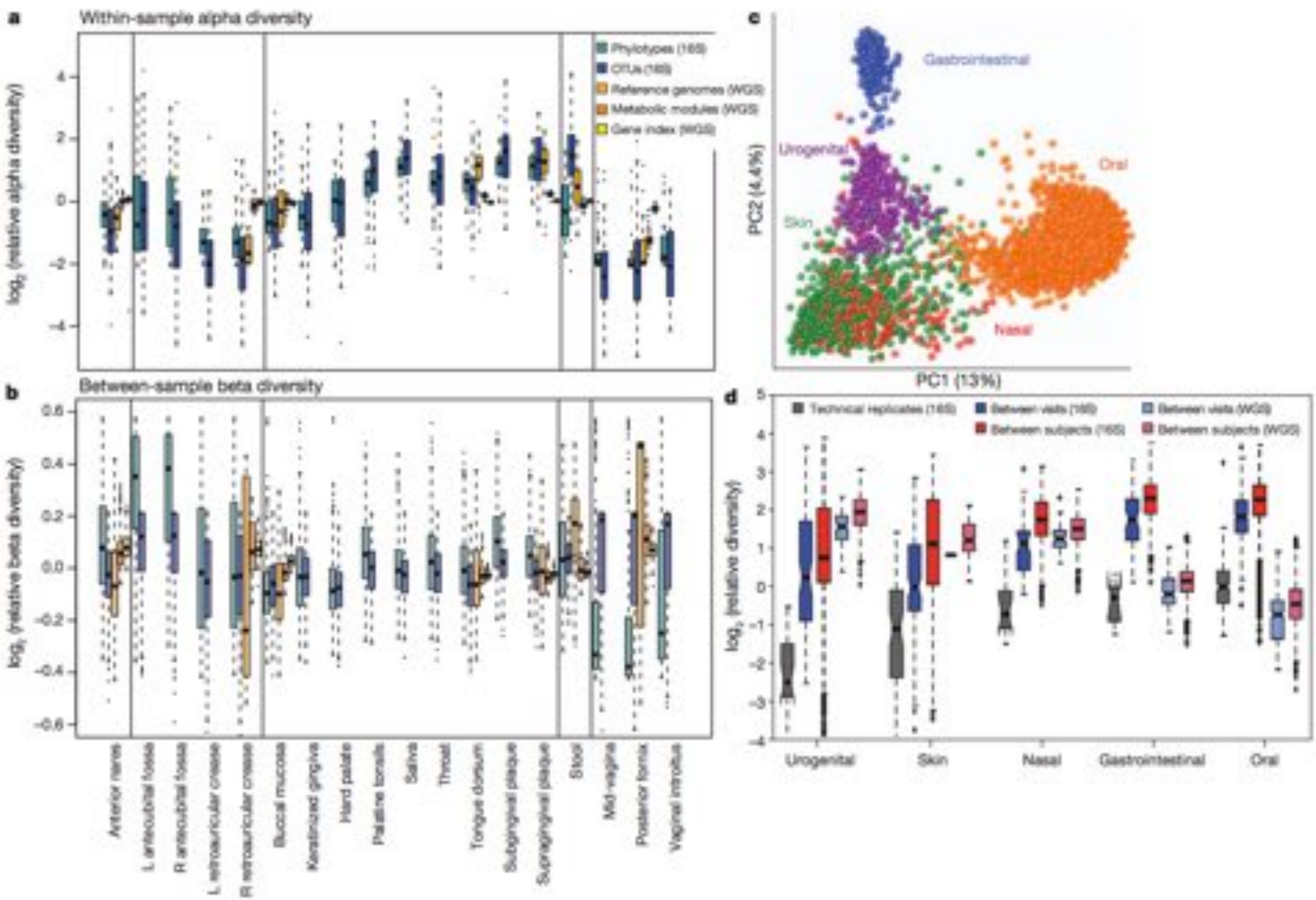
Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice
Ridaura et al (2013) Science. doi: 10.1126/science.1241214

Microbes and Human Health



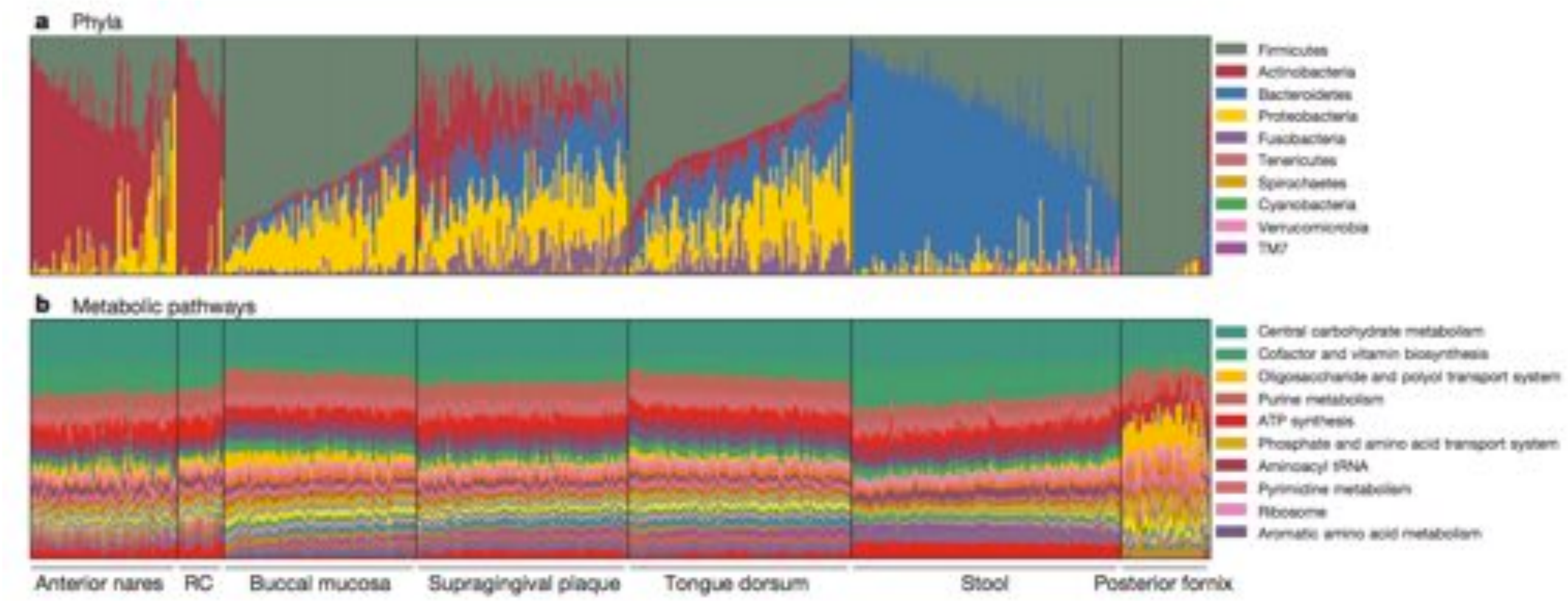
The human microbiome: at the interface of health and disease
Cho & Blaser (2012) Nature Reviews Genetics. doi:10.1038/nrg3182

Human Microbiome Project



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

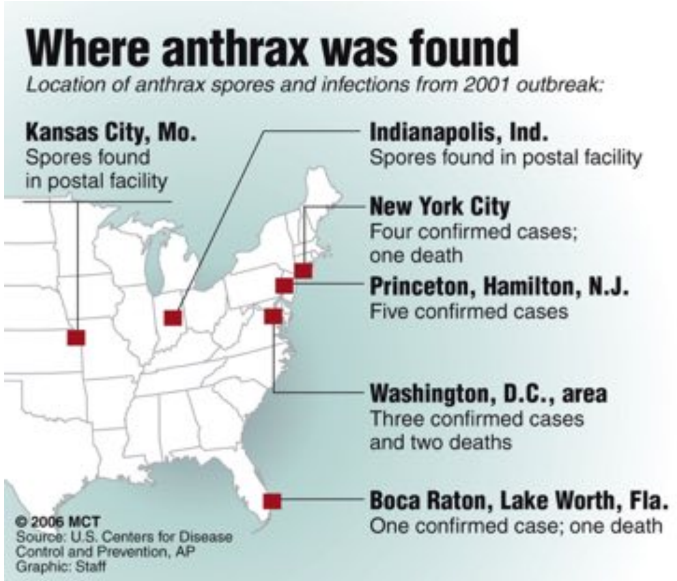
Functional composition tends to be more stable than genome composition



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

Part IV: The Future

Amerithrax Analysis



Where anthrax was found

Location of anthrax spores and infections from 2001 outbreak:

Kansas City, Mo.
Spores found in postal facility

Indianapolis, Ind.
Spores found in postal facility

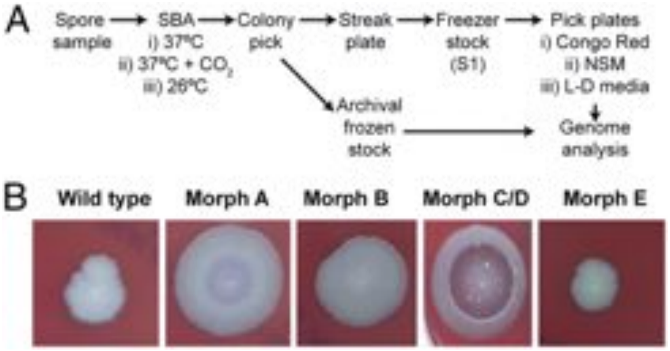
New York City
Four confirmed cases; one death

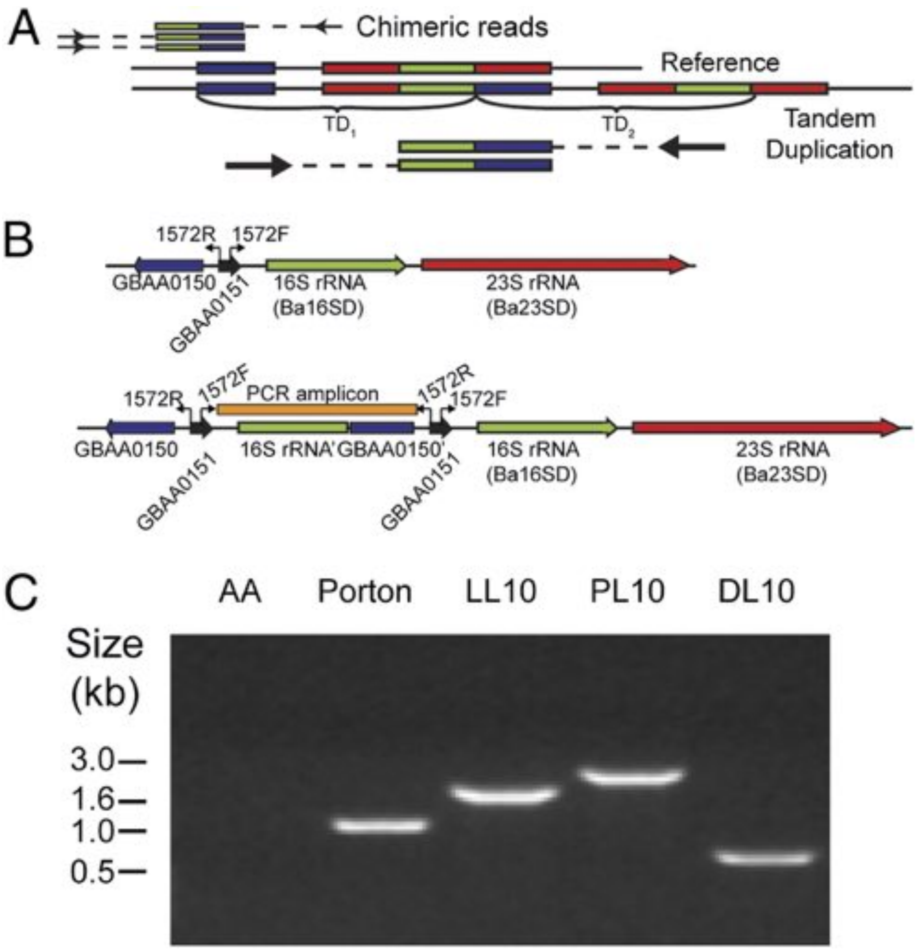
Princeton, Hamilton, N.J.
Five confirmed cases

Washington, D.C., area
Three confirmed cases and two deaths

Boca Raton, Lake Worth, Fla.
One confirmed case; one death

© 2006 MCT
Source: U.S. Centers for Disease Control and Prevention, AP
Graphic: Staff

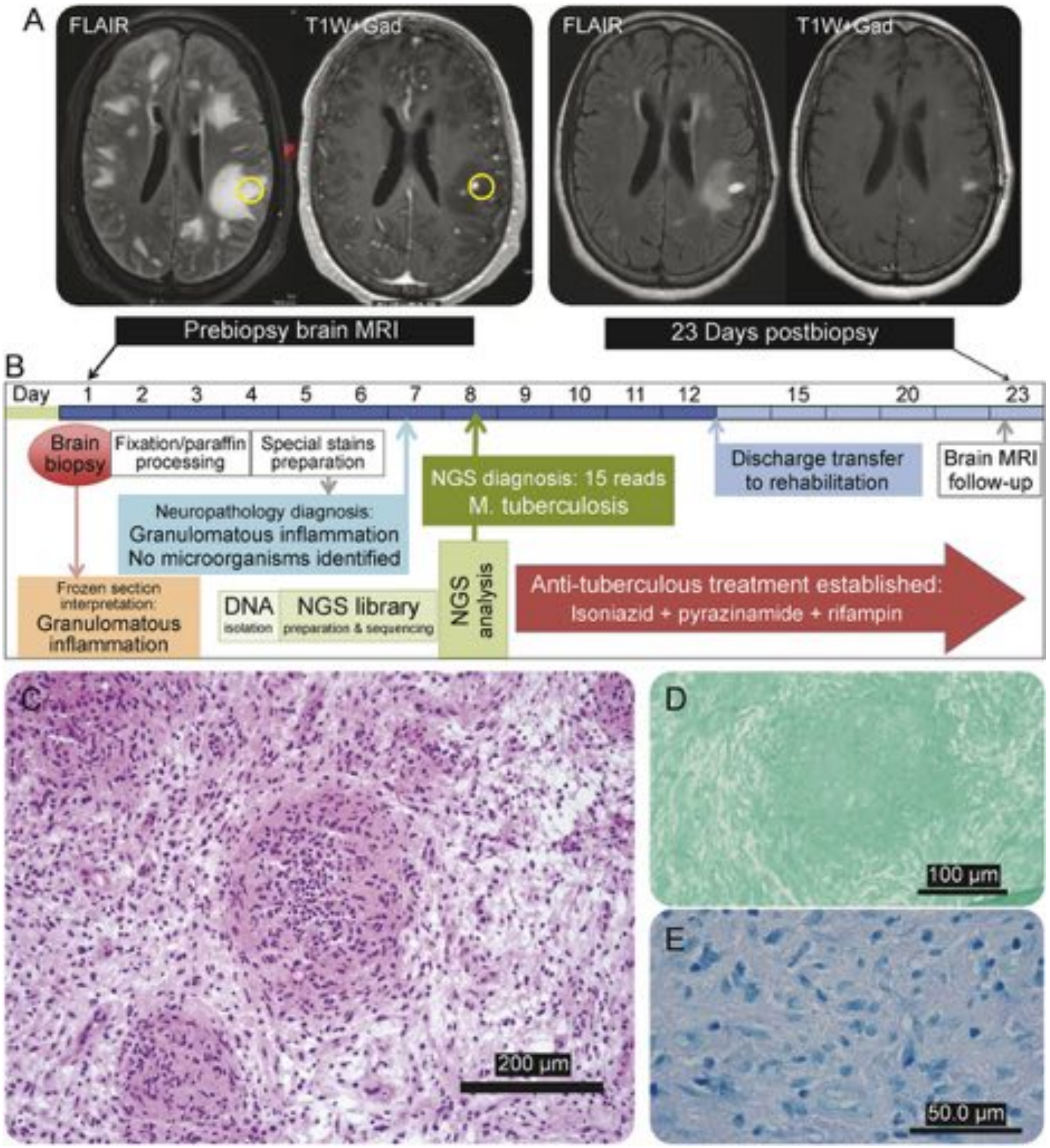




Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation

Rasko et al (2011) PNAS. doi: 10.1073/pnas.1016657108

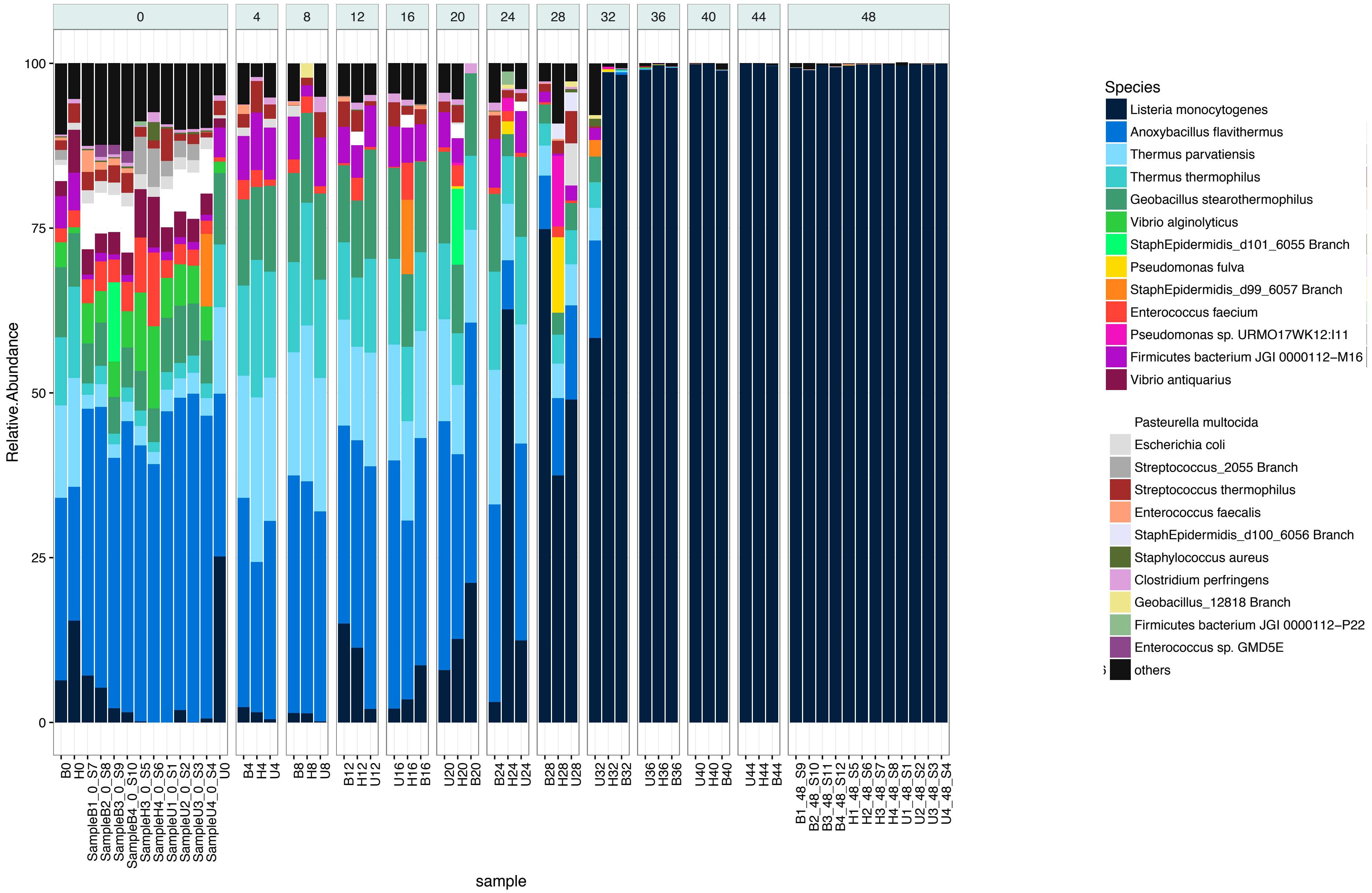
Diagnosing Brain Infections with NGS


A
FLAIR
T1W+Gad
FLAIR
T1W+Gad
Prebiopsy brain MRI
23 Days postbiopsy
B
Day
1
2
3
4
5
6
7
8
9
10
11
12
15
20
23
Brain biopsy
Fixation/paraffin processing
Special stains preparation
Neuropathology diagnosis: Granulomatous inflammation No microorganisms identified
NGS diagnosis: 15 reads M. tuberculosis
Discharge transfer to rehabilitation
Brain MRI follow-up
Frozen section interpretation: Granulomatous inflammation
DNA isolation
NGS library preparation & sequencing
NGS analysis
Anti-tuberculous treatment established: Isoniazid + pyrazinamide + rifampin
C
200 µm
D
100 µm
E
50.0 µm

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

Listeria in ice cream



COSMOSID®