Lecture 21. Microbiology & Metagenomics

Michael Schatz

April 16, 2018

JHU 600.749: Applied Comparative Genomics

Part I: Introduction

Microbial Taxonomy

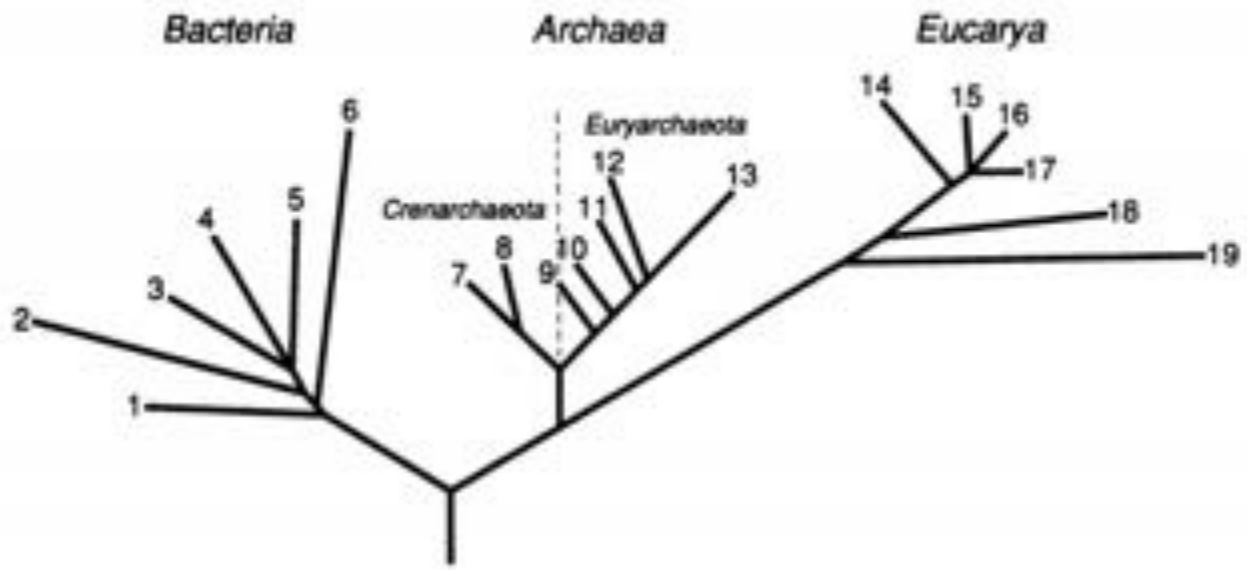
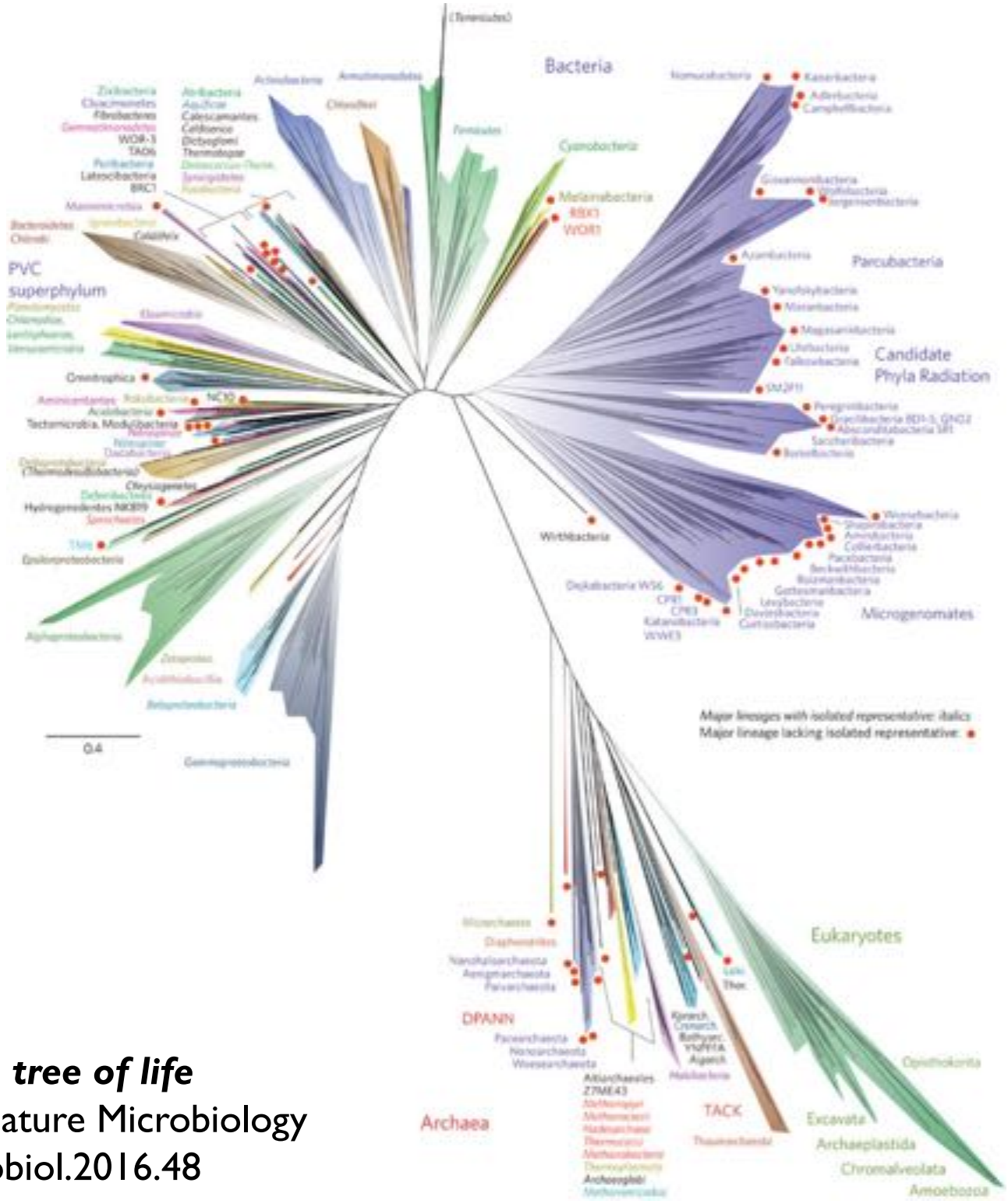


FIG. 1. Universal phylogenetic tree in rooted form, showing the three domains. Branching order and branch lengths are based upon rRNA sequence comparisons (and have been taken from figure 4 of ref. 2). The position of the root was determined by comparing (the few known) sequences of pairs of paralogous genes that diverged from each other before the three primary lineages emerged from their common ancestral condition (27). [This rooting strategy (28) in effect uses the one set of (aboriginally duplicated) genes as an outgroup for the other.] The numbers on the branch tips correspond to the following groups of organisms (2). Bacteria: 1, the Thermotogales; 2, the flavobacteria and relatives; 3, the cyanobacteria; 4, the purple bacteria; 5, the Gram-positive bacteria; and 6, the green nonsulfur bacteria. Archae: the kingdom Crenarchaeota: 7, the genus *Pyrodictium*; and 8, the genus *Thermoproteus*; and the kingdom Euryarchaeota: 9, the Thermococcales; 10, the Methanococcales; 11, the Methanobacteriales; 12, the Methanomicrobiales; and 13, the extreme halophiles. Eucarya: 14, the animals; 15, the ciliates; 16, the green plants; 17, the fungi; 18, the flagellates; and 19, the microsporidia.

Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.
Woese et al (1990) PNAS. doi: 10.1073/pnas.87.12.4576



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

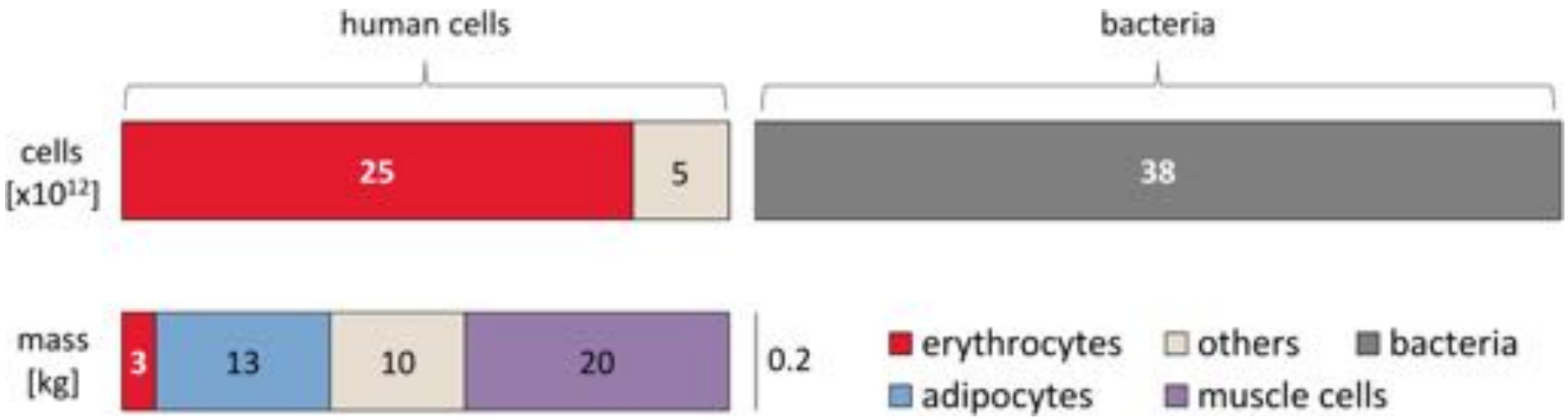
Human brain:
~3.3 lbs

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! ☺

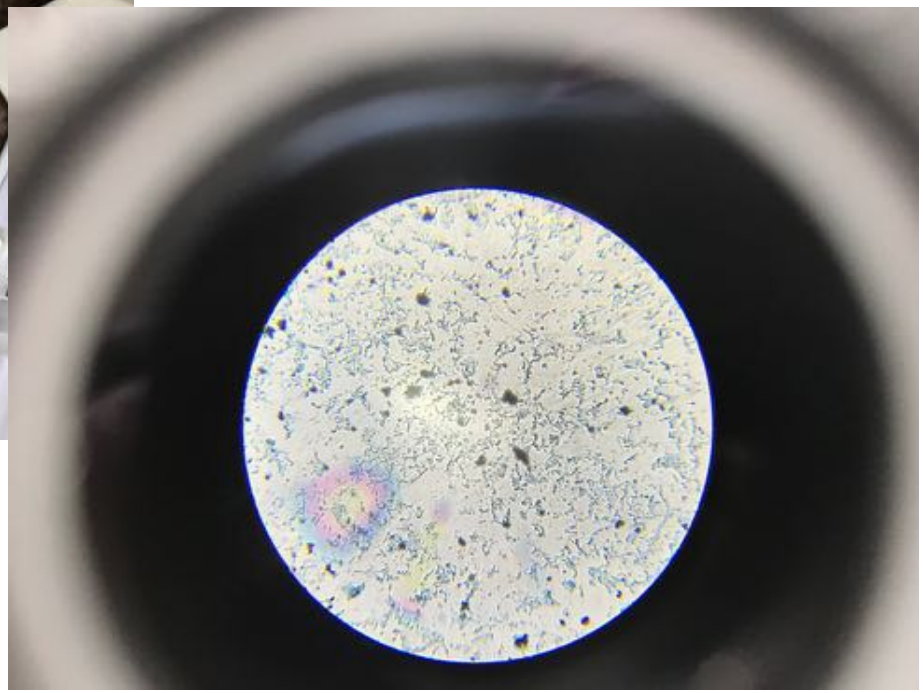


population segment	body weight [kg]	age [y]	blood volume [L]	RBC count [10^{12}/L]	colon content [g]	bac. conc. [10^{11}/ g wet] (1)	total human cells [10^{12}] (2)	total bacteria [10^{12}]	B:H
ref. man	70	20–30	4.9	5.0	420	0.92	30	38	**1.3**
ref. woman	63		3.9	4.5	480	0.92	21	44	**2.2**
young infant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

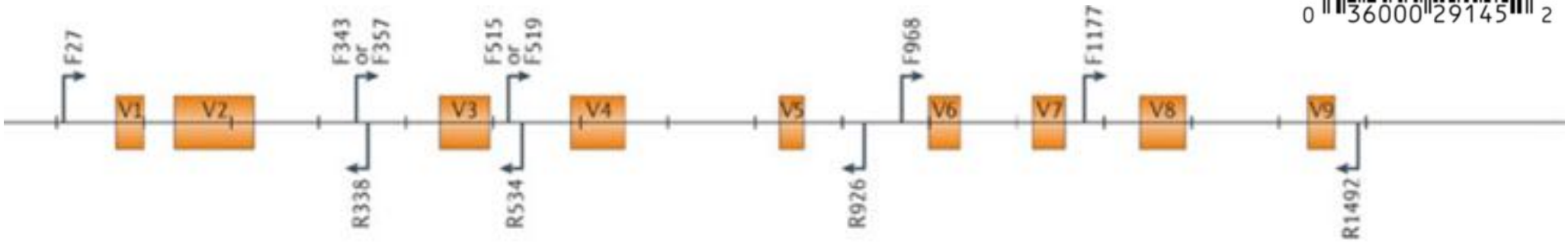
Revised Estimates for the Number of Human and Bacteria Cells in the Body
Sender et al (2016) PLOS Biology. https://doi.org/10.1371/journal.pbio.1002533

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

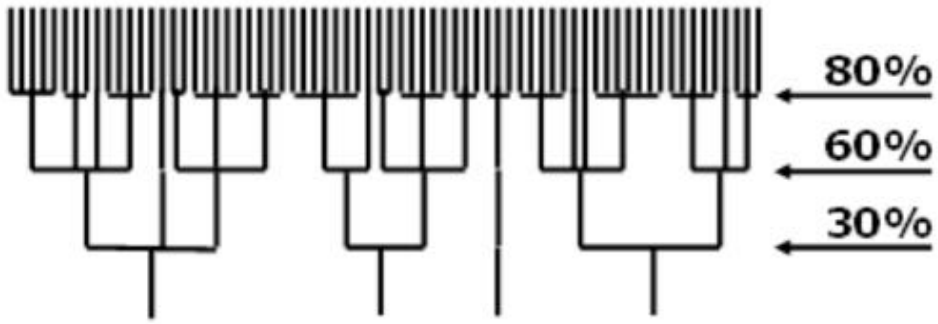
Various concepts have been suggested for microbial species, but none have been generally accepted[9]. The following quotes represent several published concepts that were chosen to illustrate the lack of consensus:

- "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)

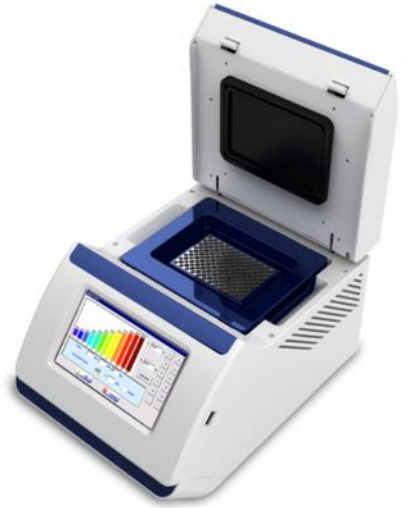
OTUs take the place of "species" in many microbiome diversity analyses because named species genomes are often unavailable for particular marker sequences.



- 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



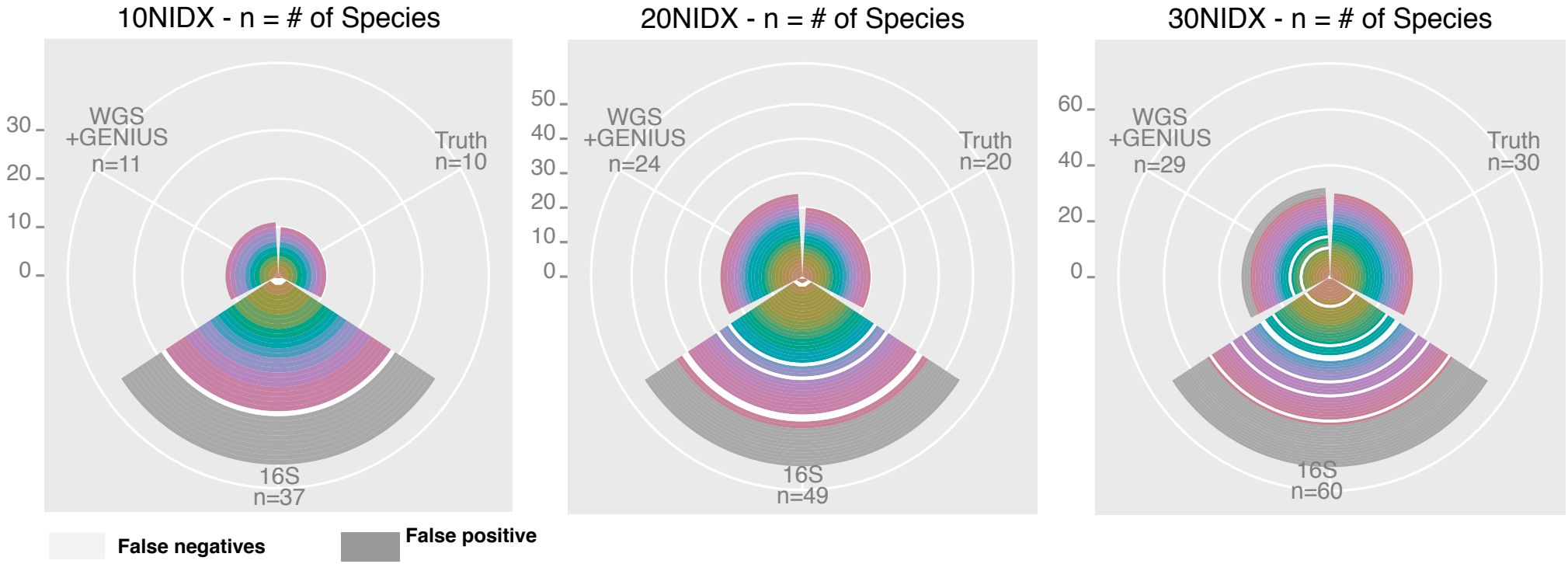
16S

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.

16S Overestimates Diversity

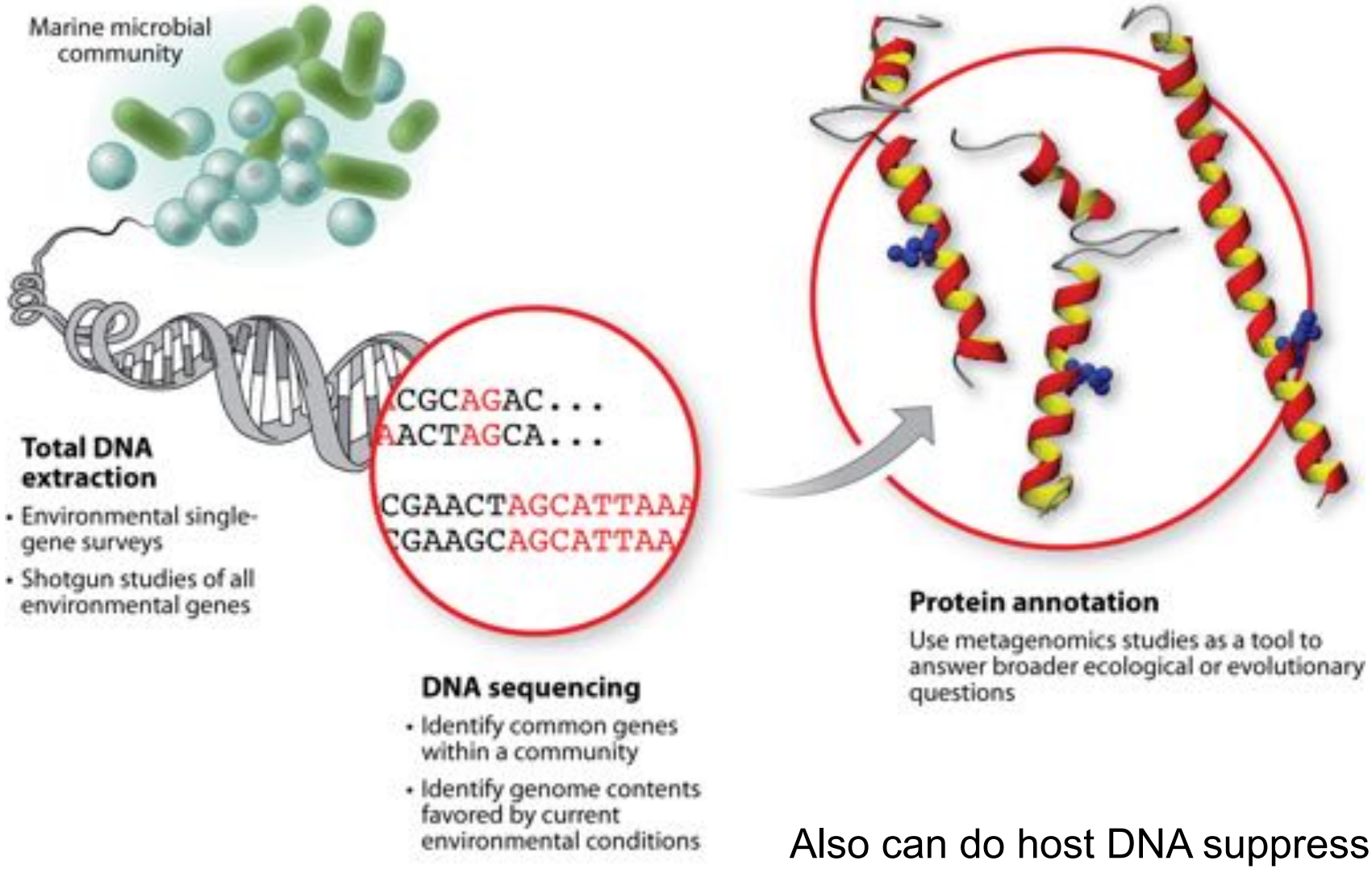


COSMOSID®


PRIUS

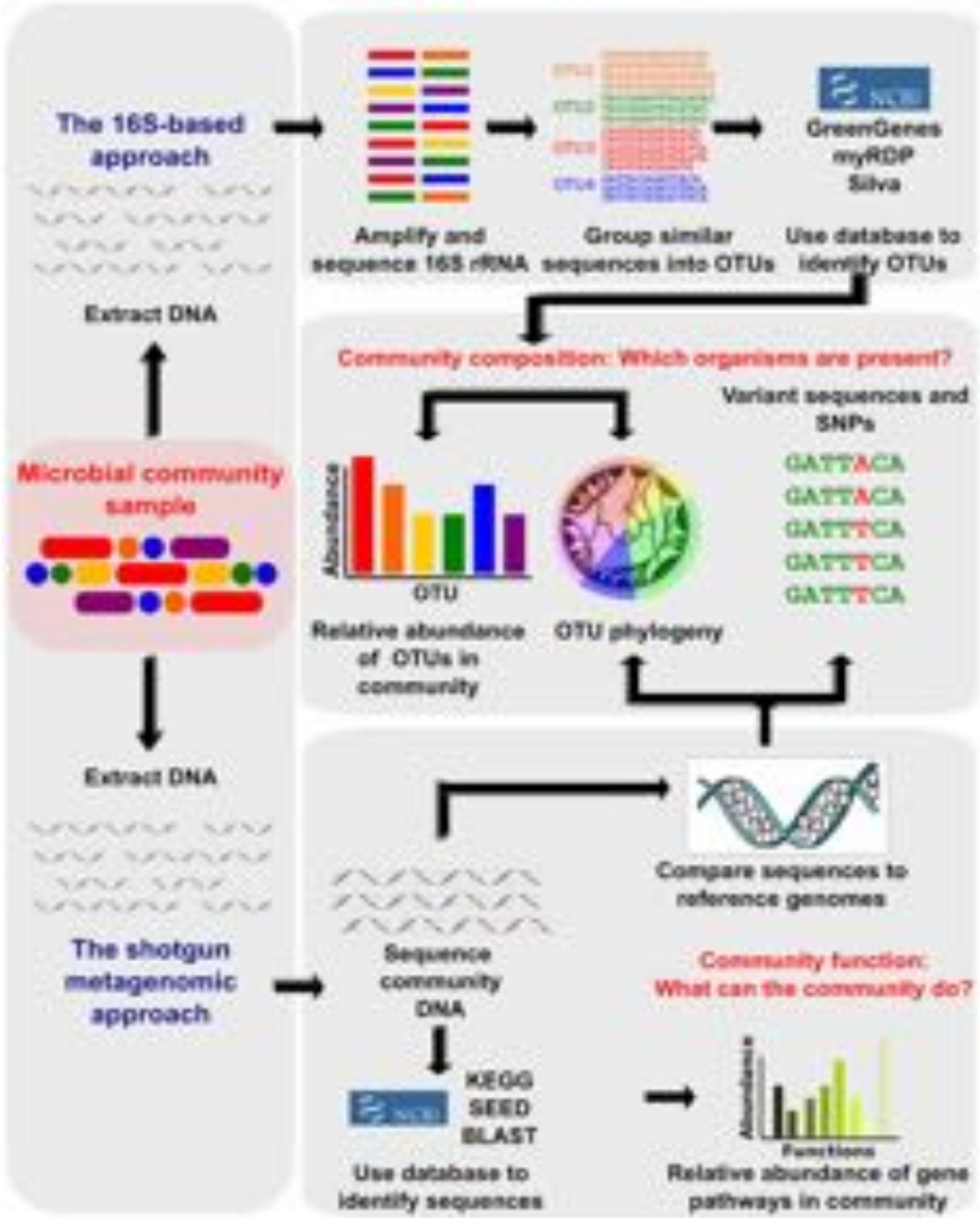
Part II: Methods

Sequencing Based Analysis



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

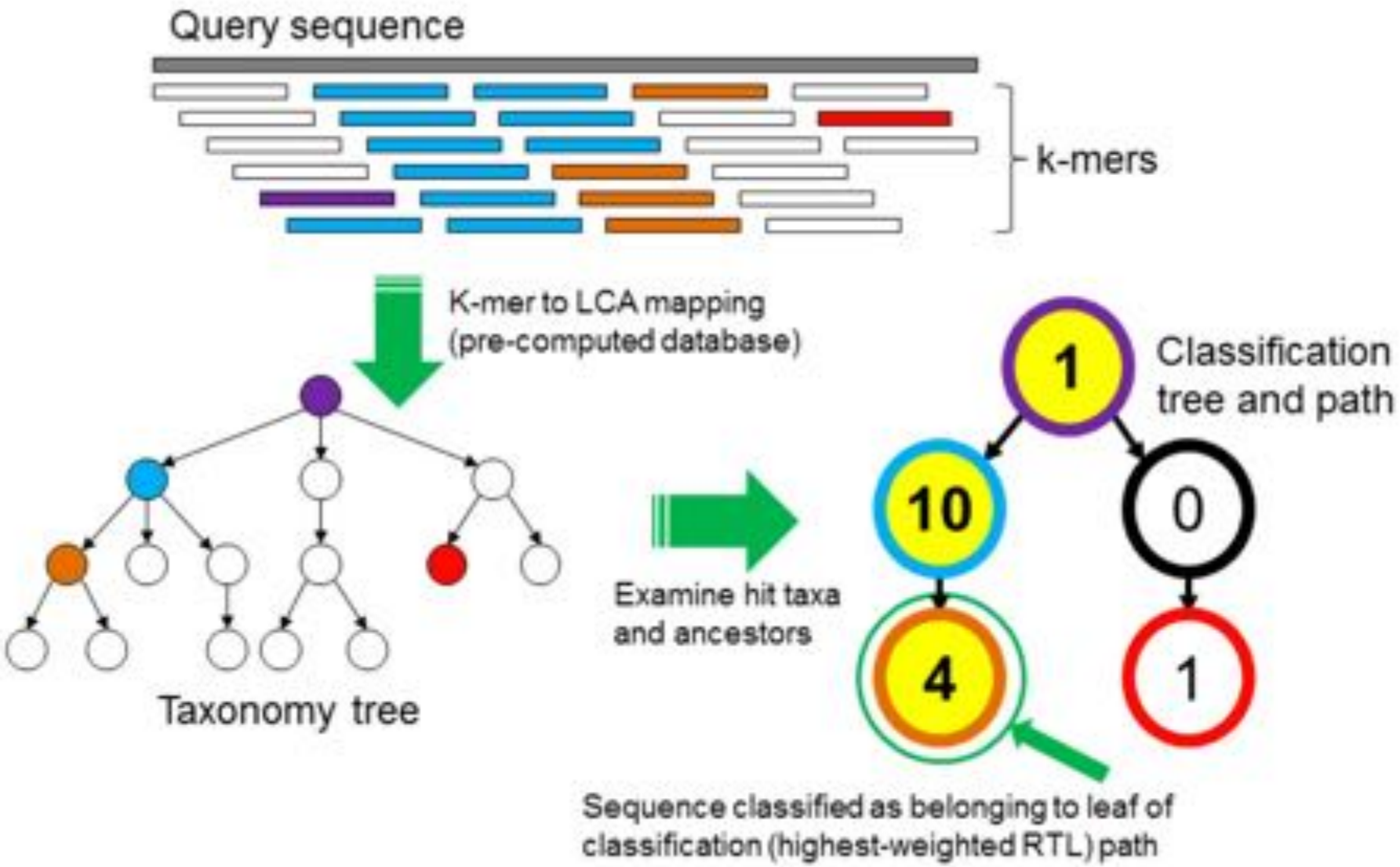
Also can do host DNA suppression/ microbial enrichment!



Chapter 12: Human Microbiome Analysis

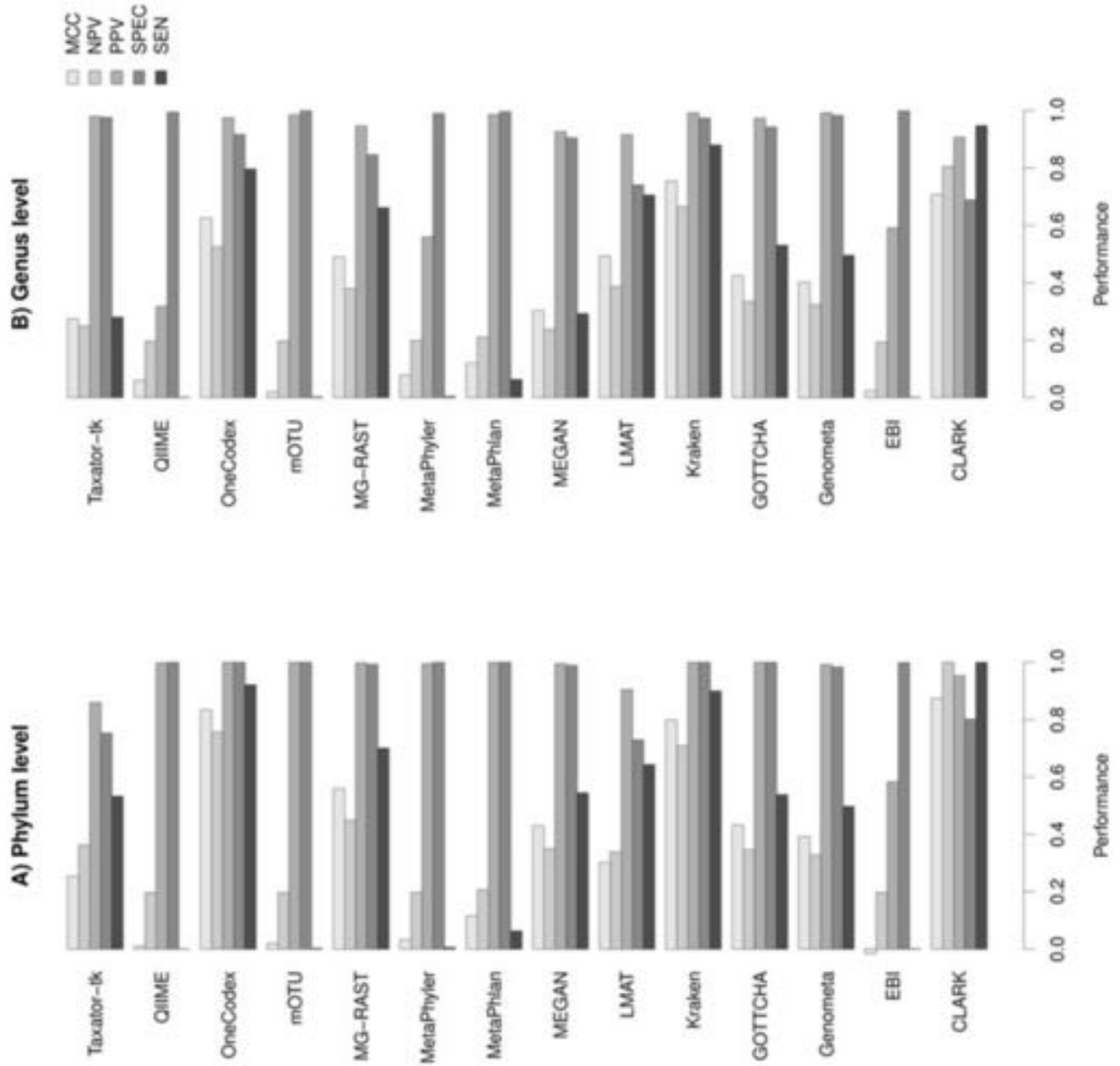
Morgan & Huttenhower (2012) PLOS Comp Bio.https://doi.org/10.1371/journal.pcbi.1002808

Kraken



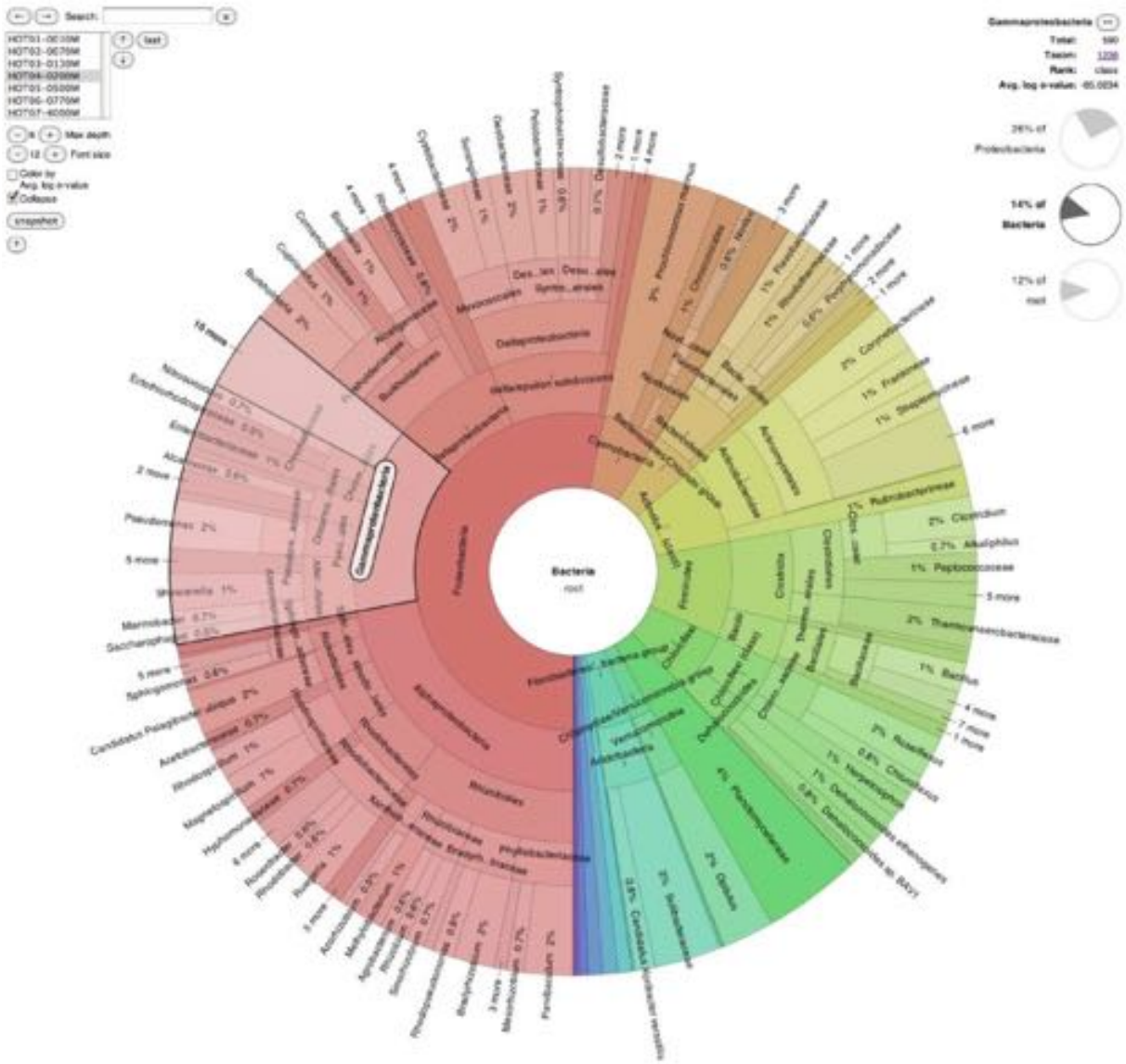
Kraken: ultrafast metagenomic sequence classification using exact alignments
Wood and Salzberg (2014) Genome Biology. DOI: 10.1186/gb-2014-15-3-r46

Metagenomics Benchmarking



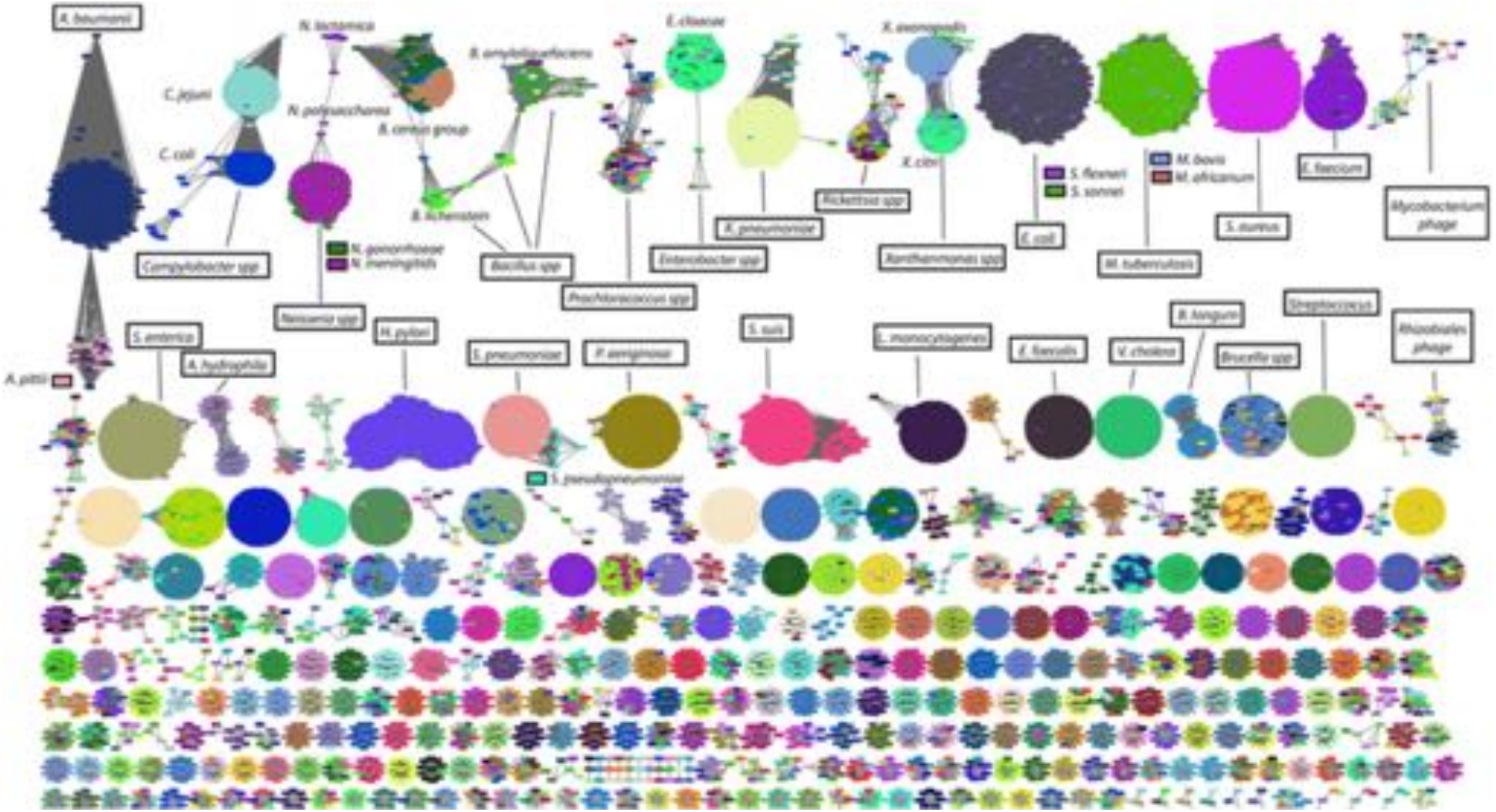
An evaluation of the accuracy and speed of metagenome analysis tools
Lindgreen et al (2016) Scientific Reports. doi:10.1038/srep19233

Krona Plots



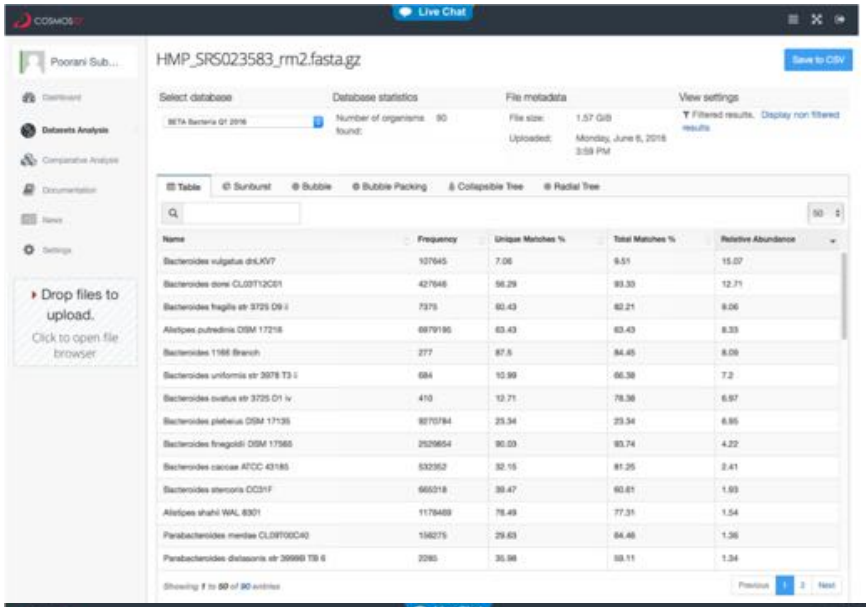
Interactive metagenomic visualization in a Web browser
Ondov et al (2011) BMC Bioinformatics. DOI: 10.1186/1471-2105-12-385

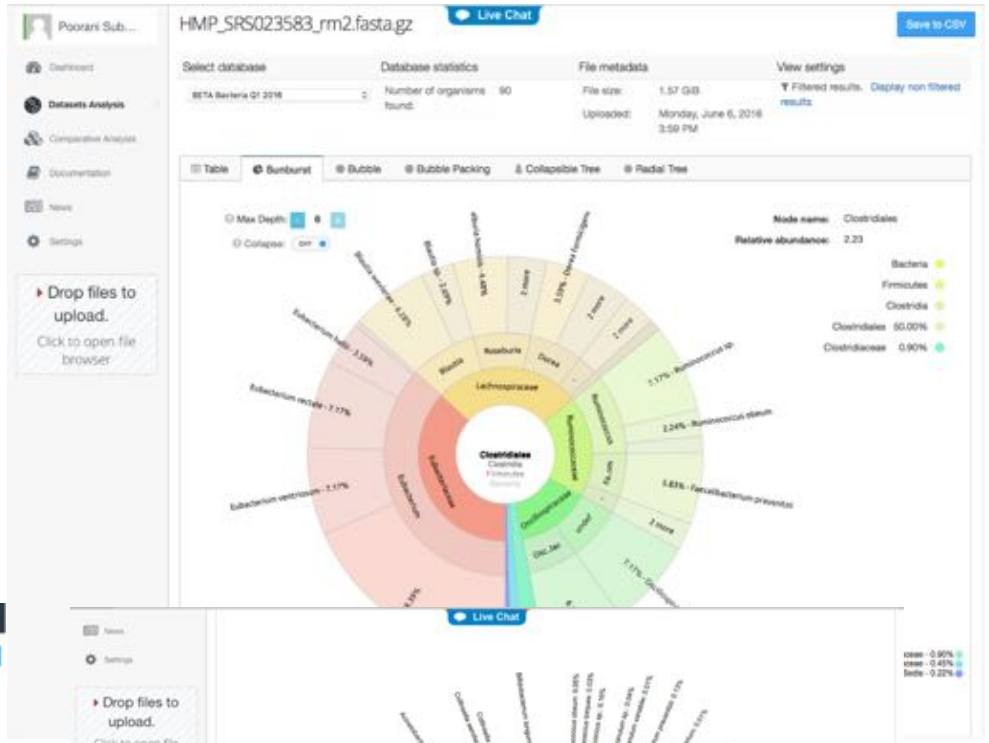
Min-Hash: Comparing all 54,118 RefSeq genomes in 1 day on a laptop

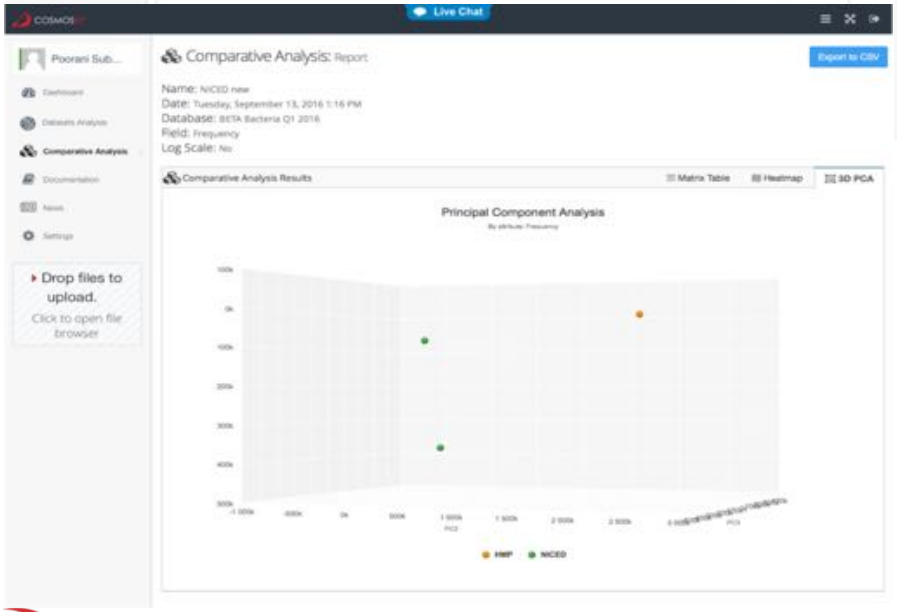


Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al. (2016) Genome Biology. DOI: 10.1186/s13059-016-0997-x

Cosmos ID: Unlocking the Microbiome



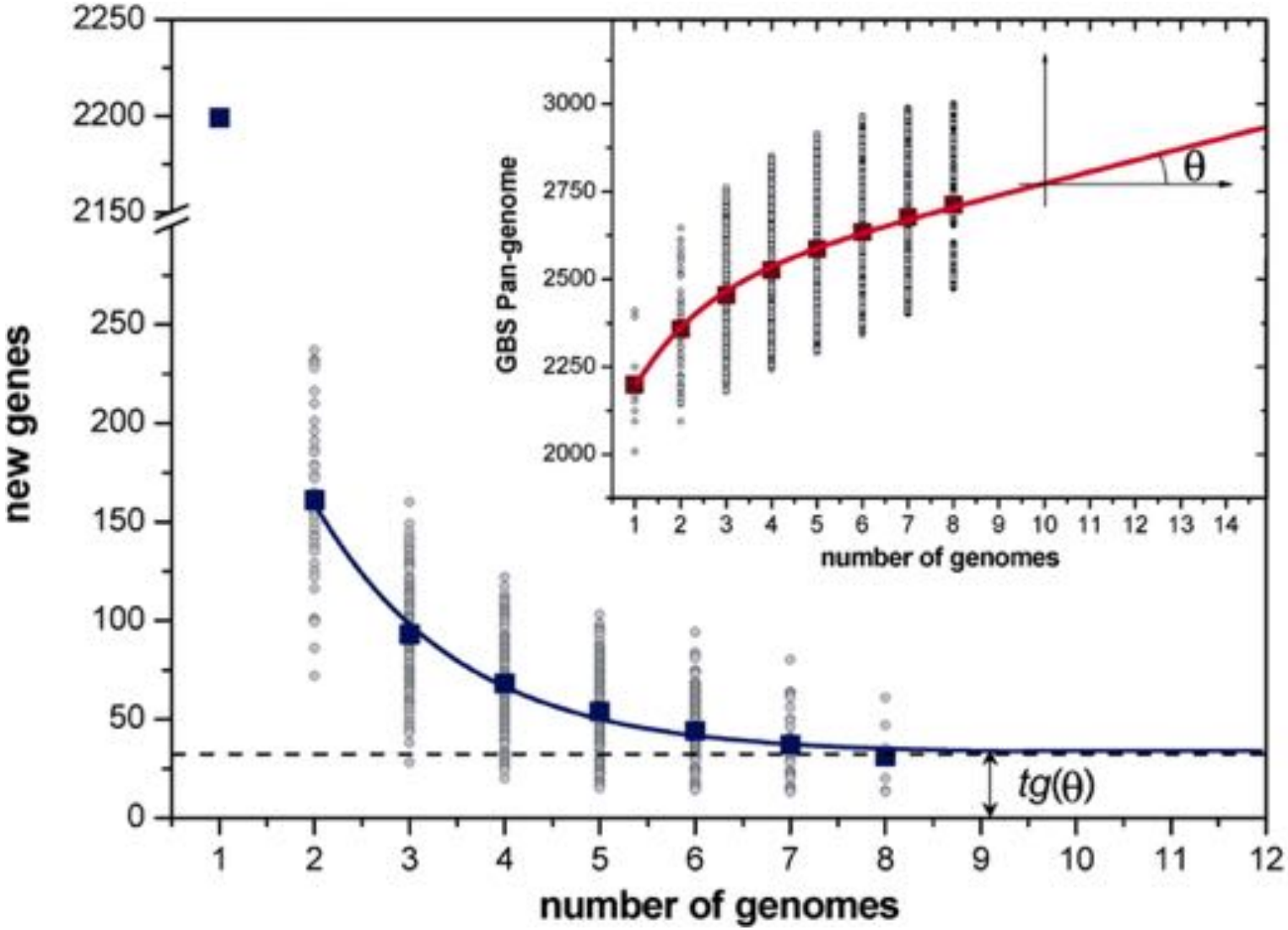




COSMOSID®

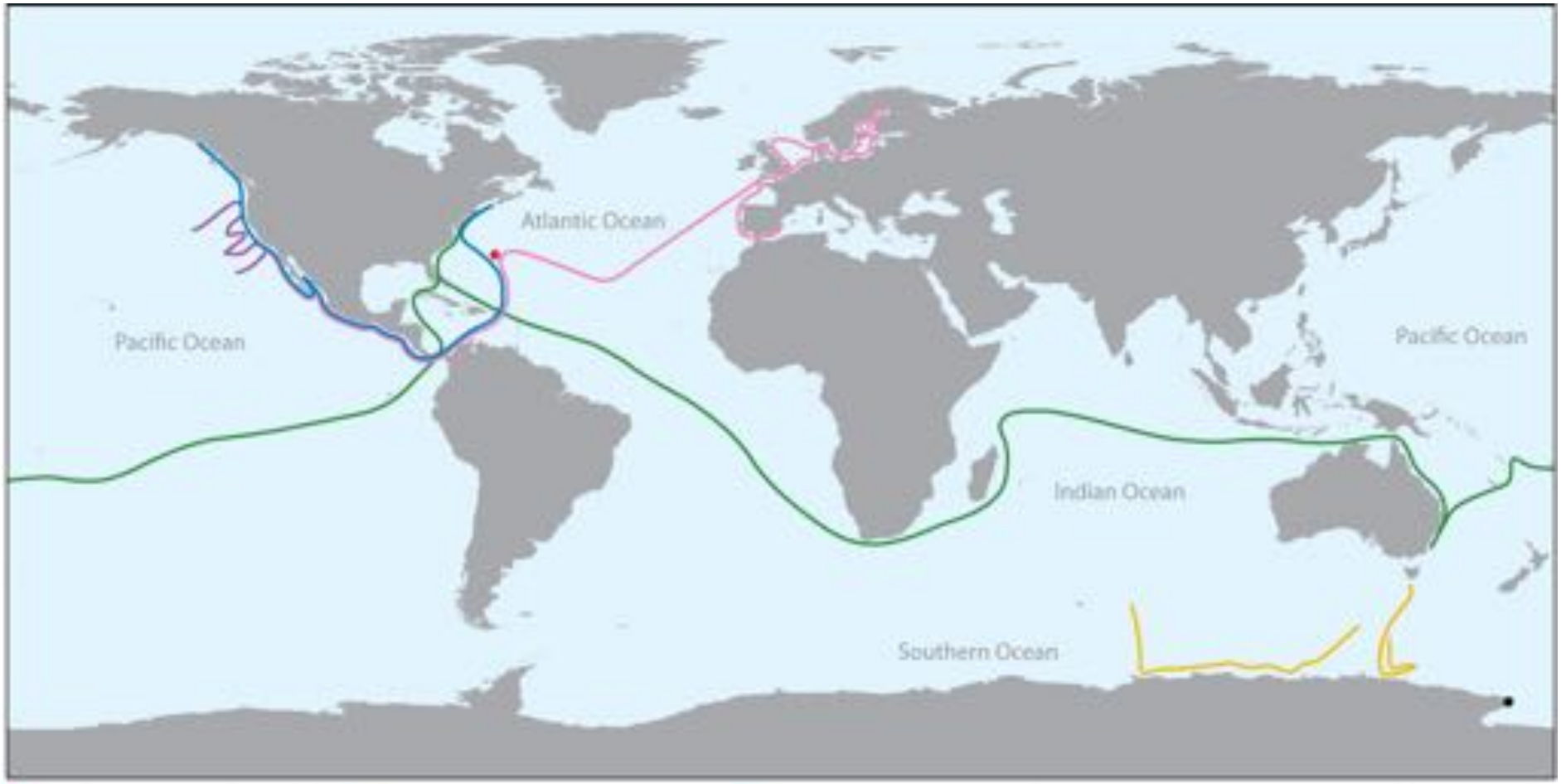
Part III: Results

Pan genome of *Streptococcus agalactiae*



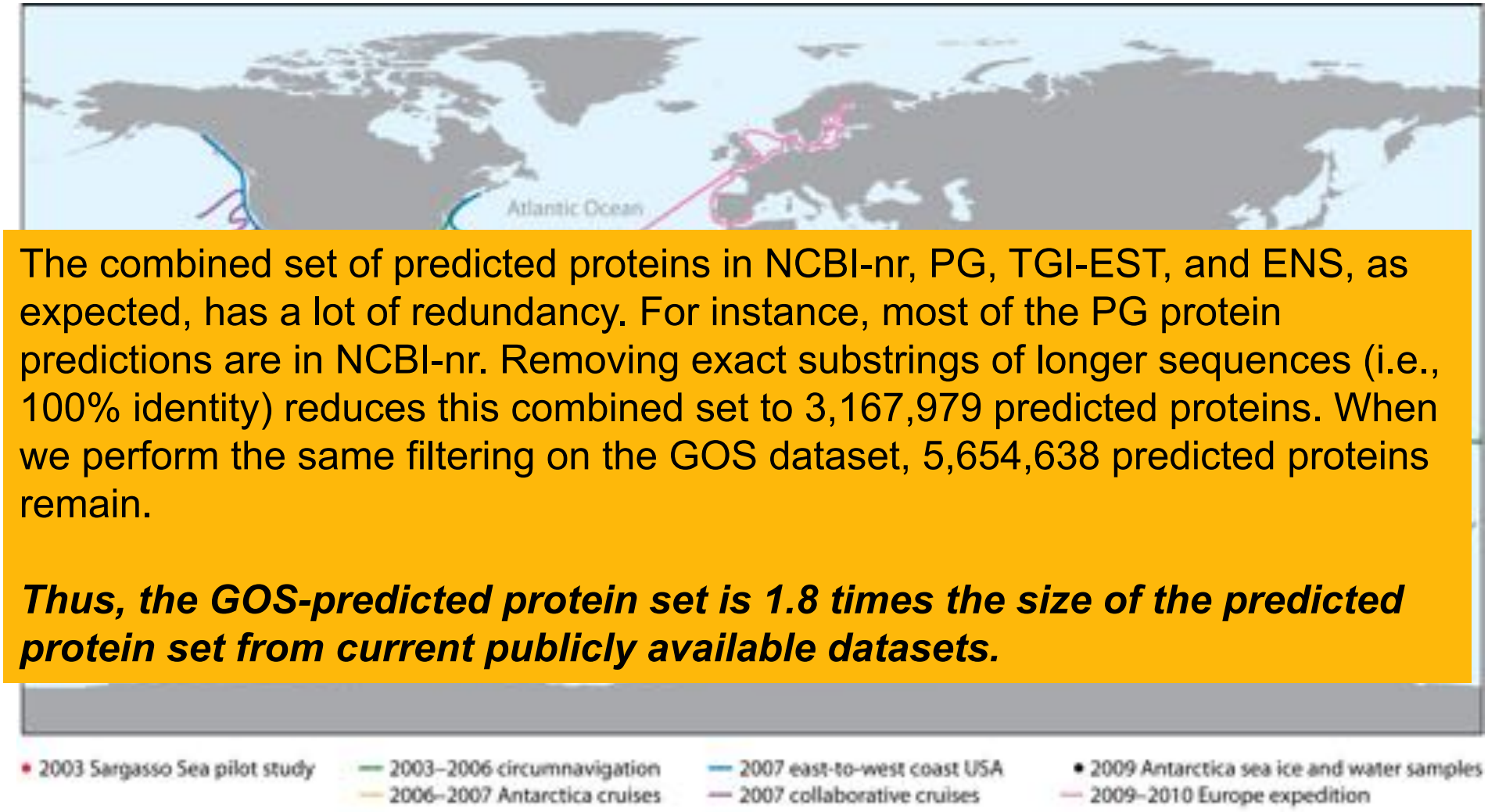
Hervé Tettelin et al. PNAS 2005;102:13950-13955

Global Ocean Survey



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

Global Ocean Survey

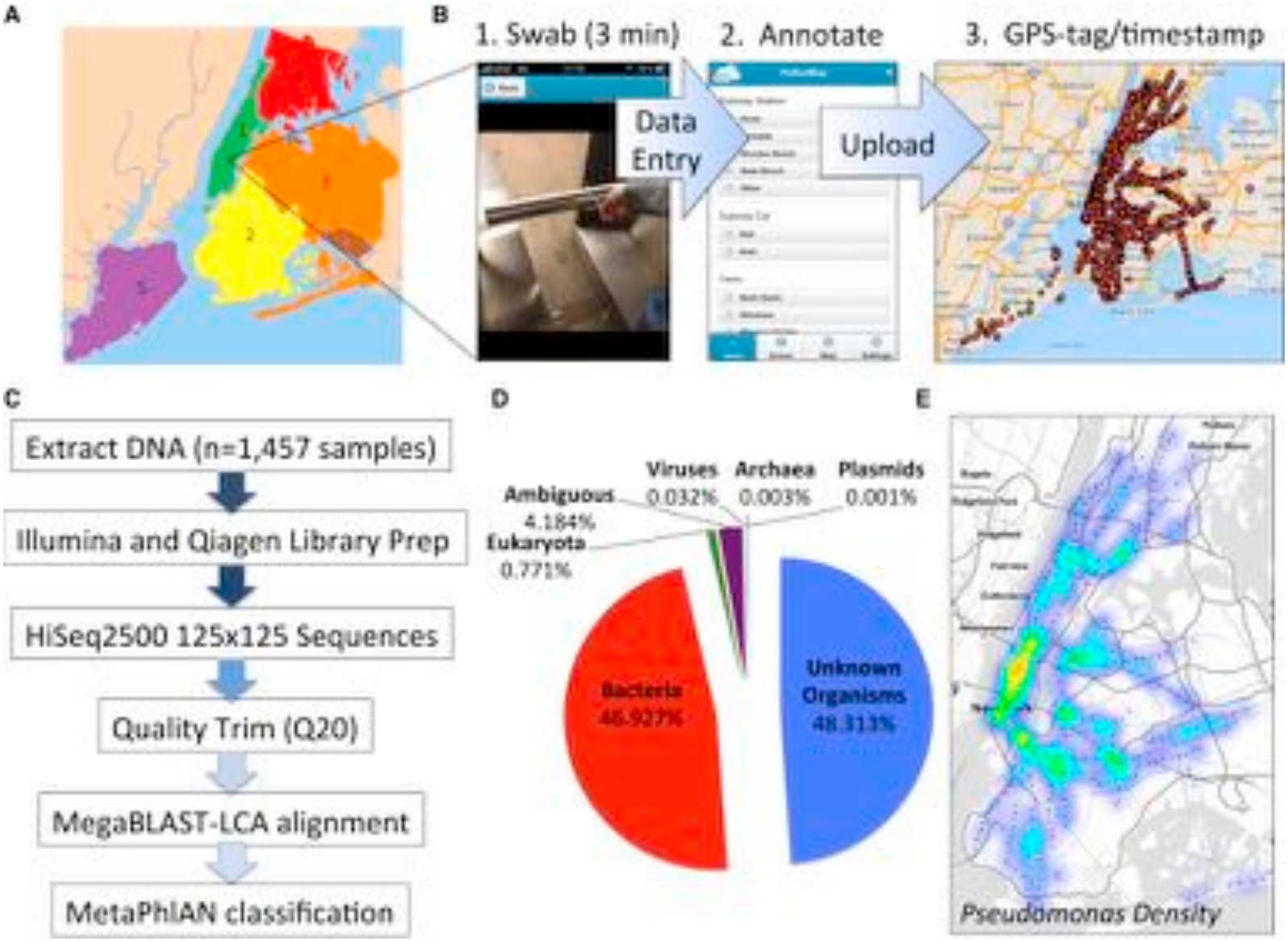


The combined set of predicted proteins in NCBI-nr, PG, TGI-EST, and ENS, as expected, has a lot of redundancy. For instance, most of the PG protein predictions are in NCBI-nr. Removing exact substrings of longer sequences (i.e., 100% identity) reduces this combined set to 3,167,979 predicted proteins.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

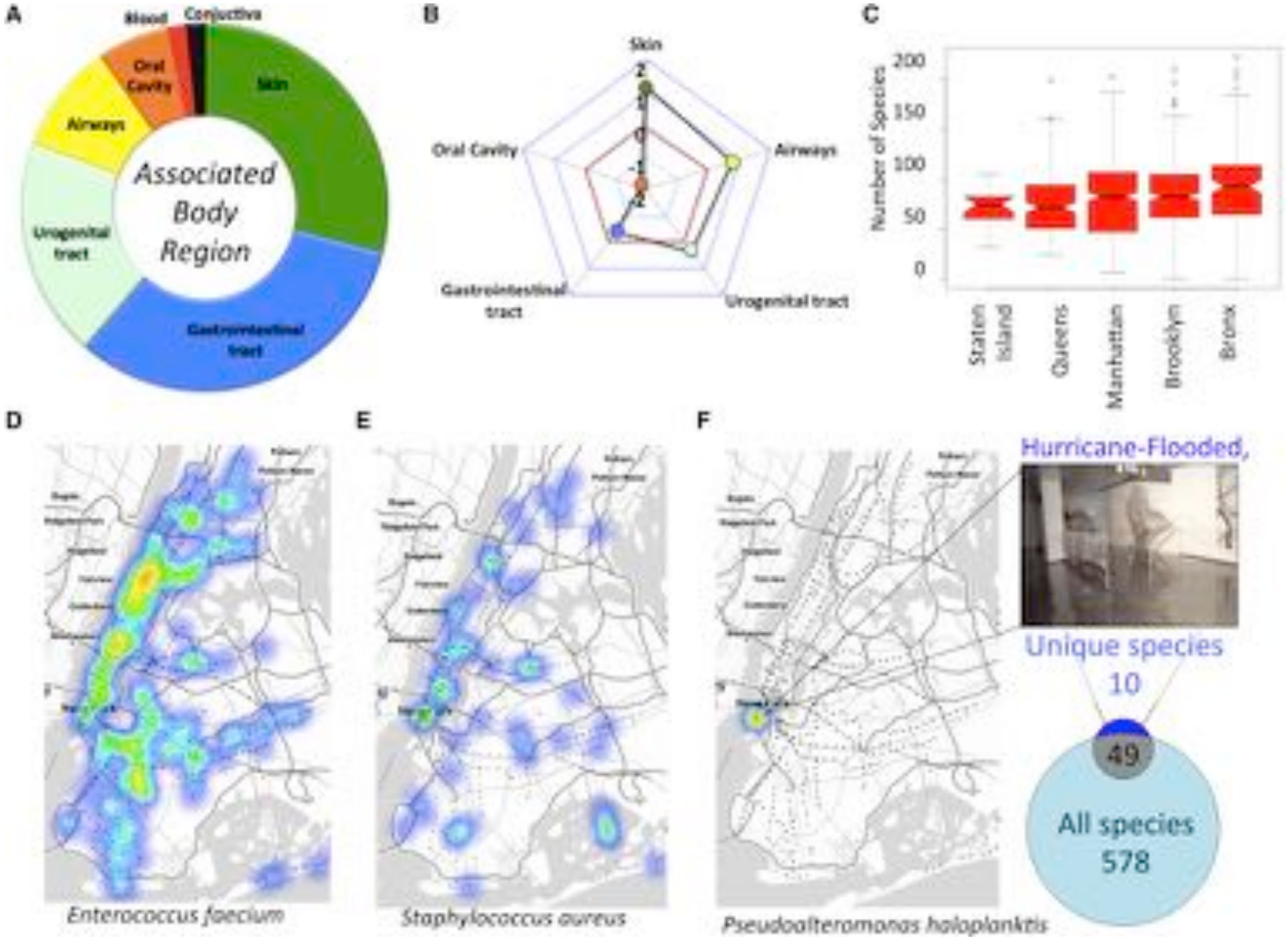
Gilbert JA, Dupont CL. 2011.
Annu. Rev. Mar. Sci. 3:347–71

Metasub



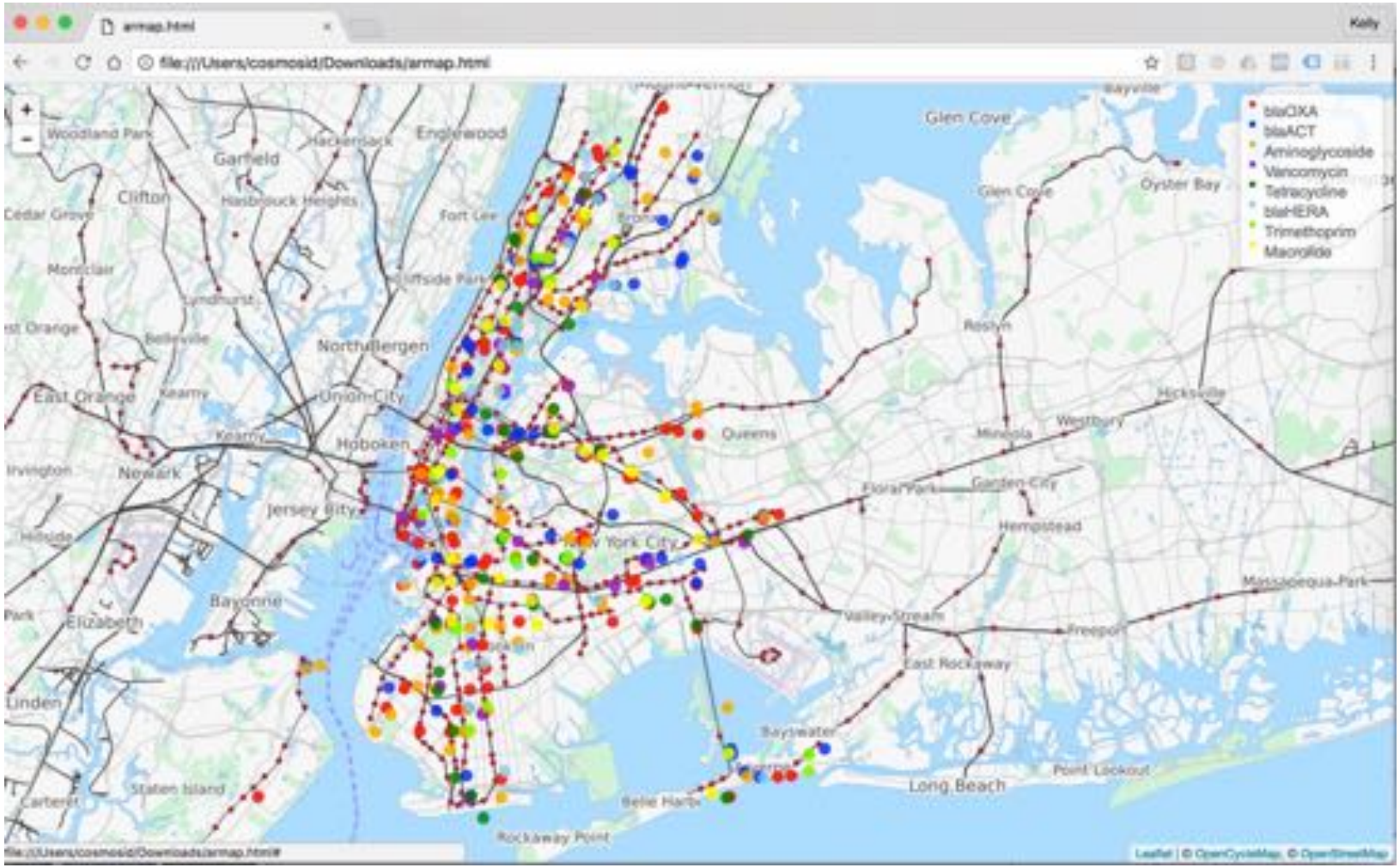
Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics
Afshinnekoo et al (2016) Cell Systems. http://dx.doi.org/10.1016/j.cels.2015.01.001

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.

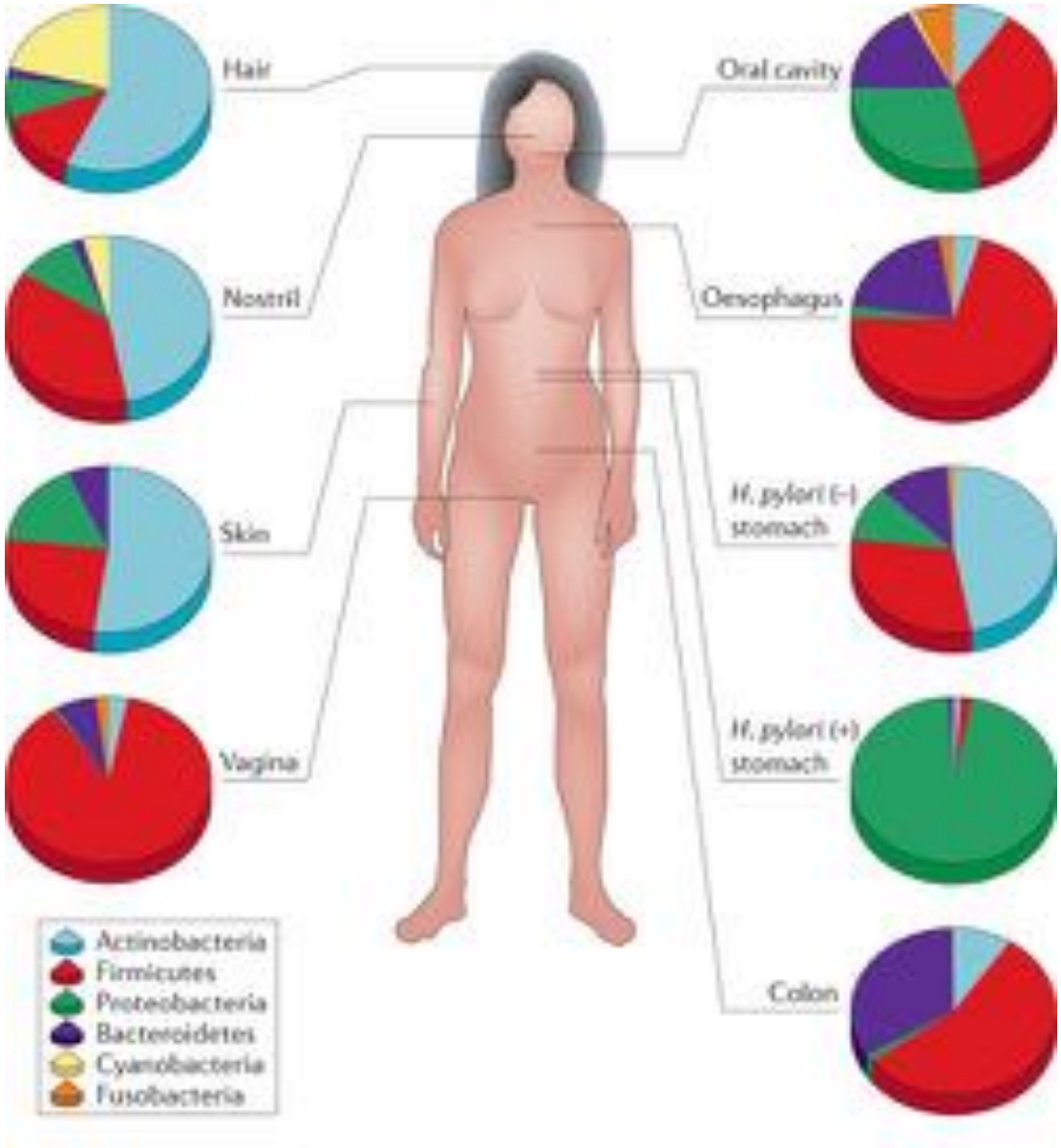
COSMOSID®

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."

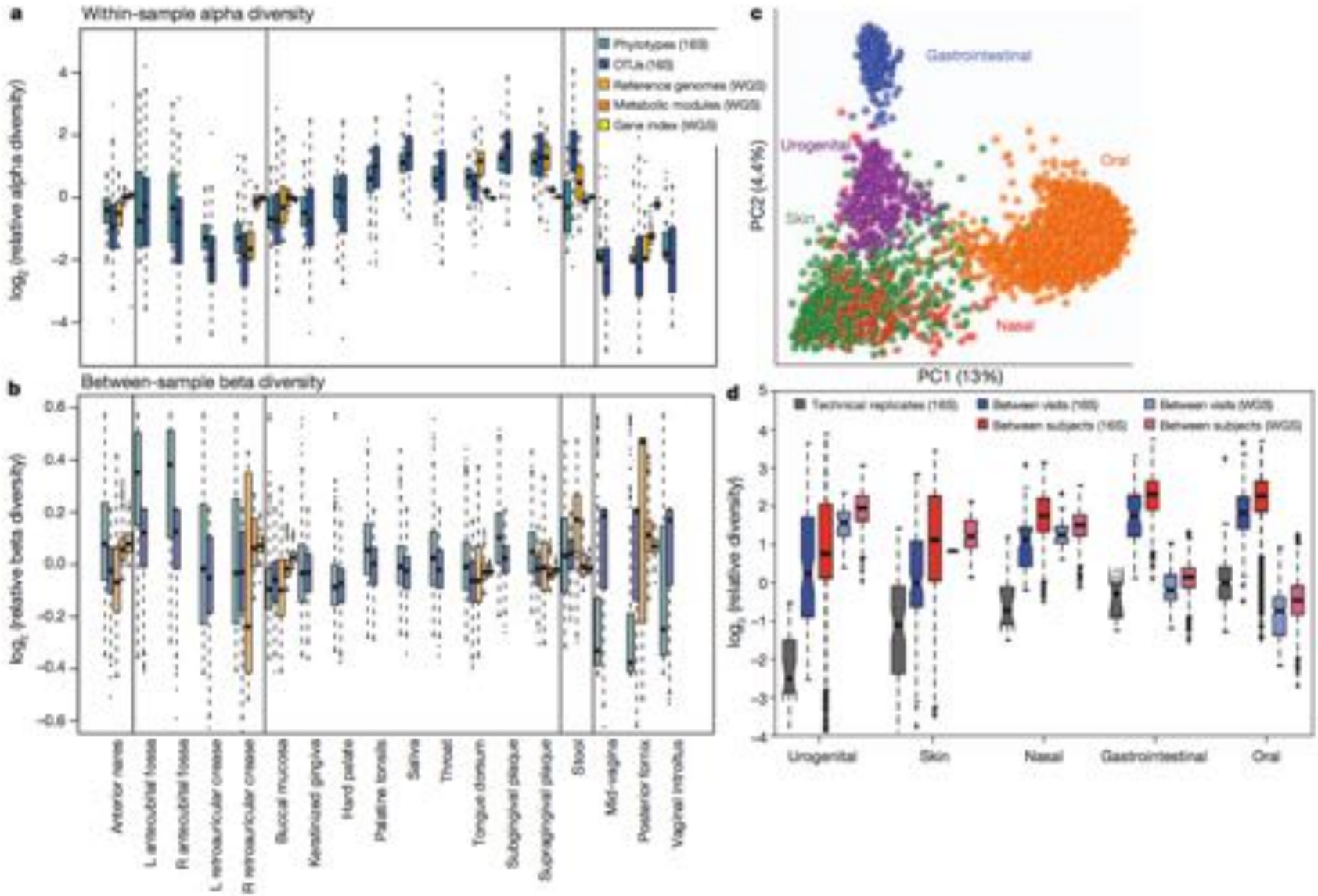
Gut Microbiota from Twins Discordant for Obesity Modulate Metabolism in Mice
Ridaura et al (2013) Science. doi: 10.1126/science.1241214

Microbes and Human Health



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

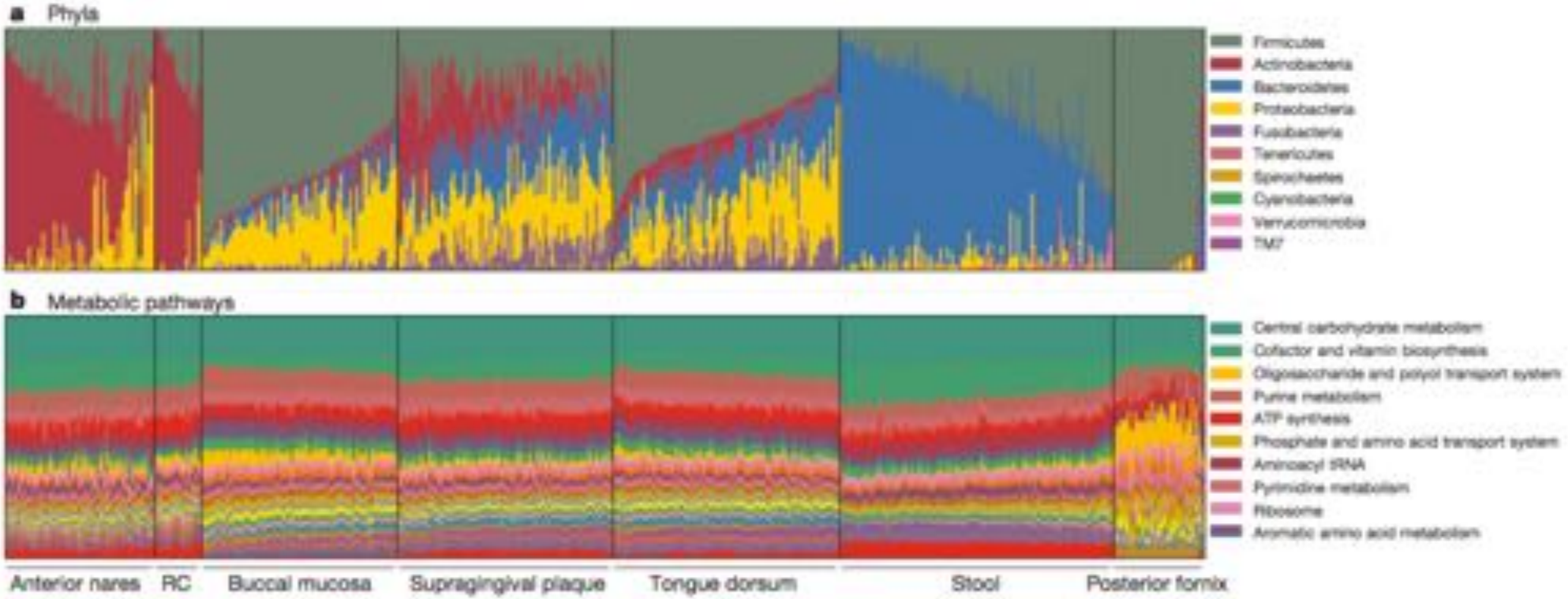
Human Microbiome Project



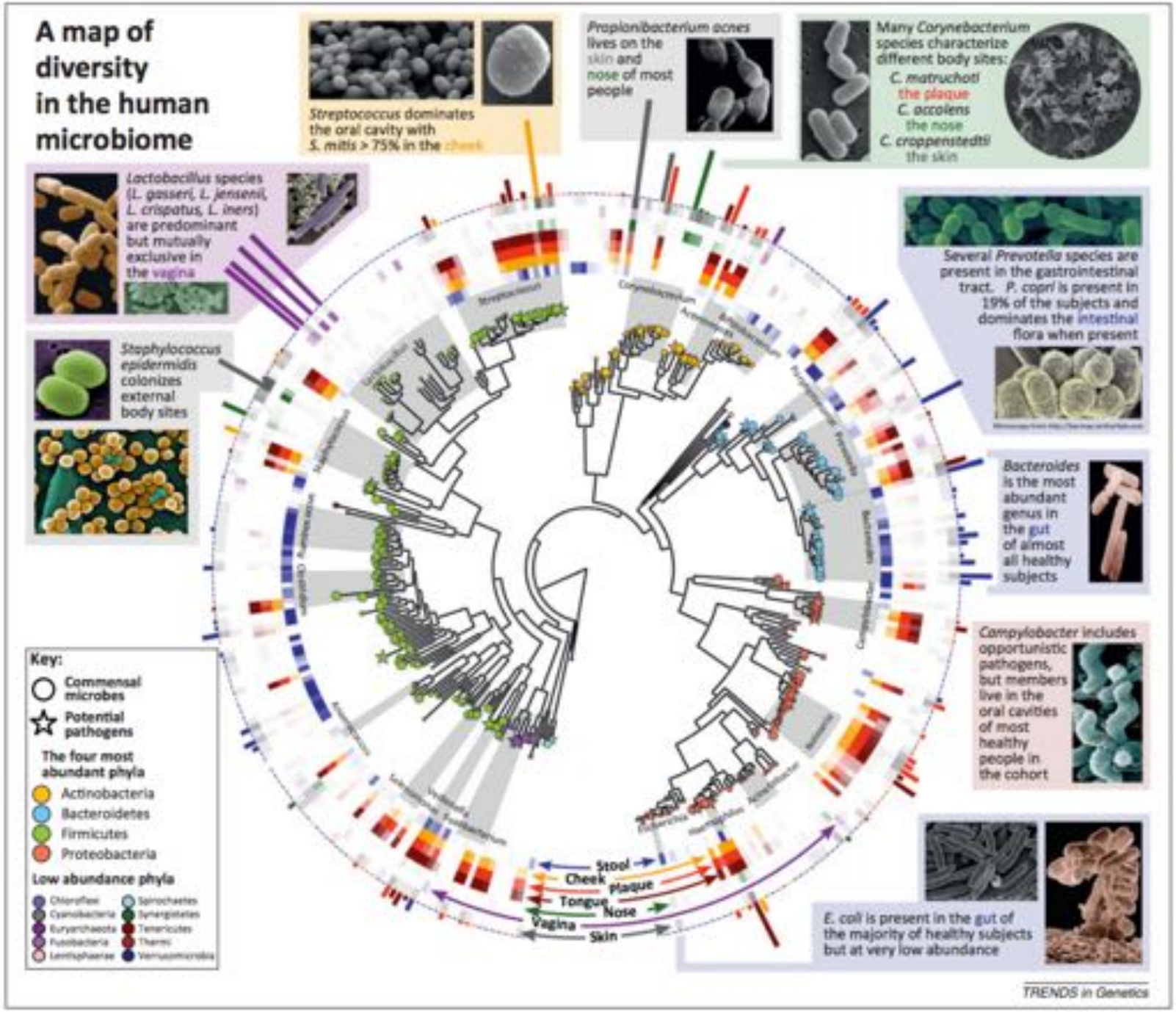
Structure, function and diversity of the healthy human microbiome

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

Functional composition tends to be more stable than genome composition



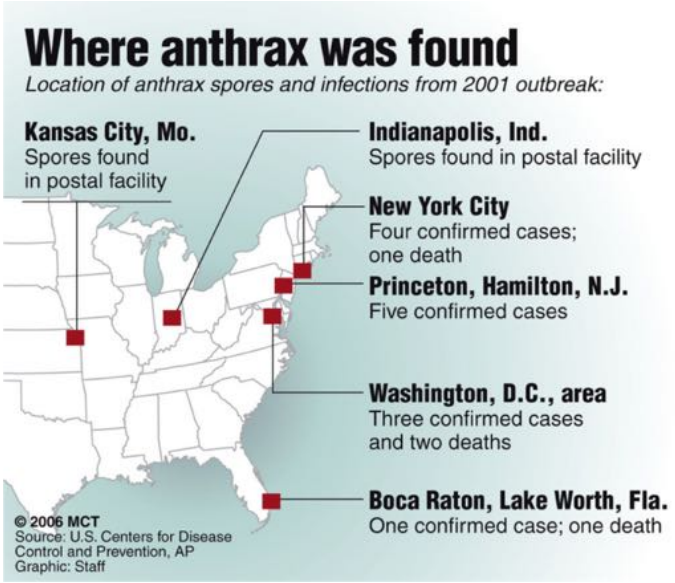
Structure, function and diversity of the healthy human microbiome
The Human Microbiome Project Consortium (2012) Nature. doi:10.1038/nature11234

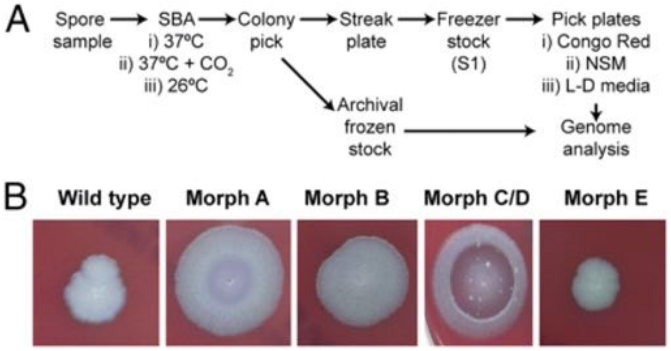


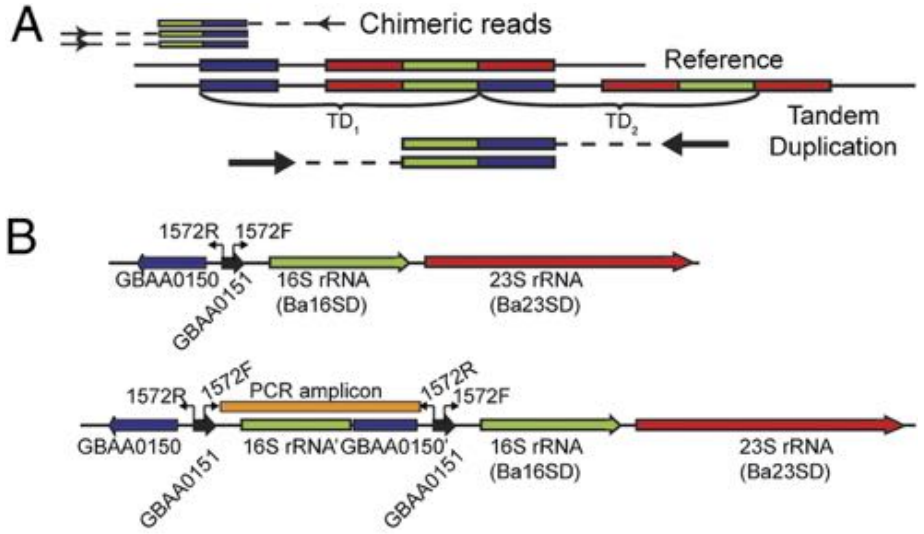
Biodiversity and functional genomics in the human microbiome

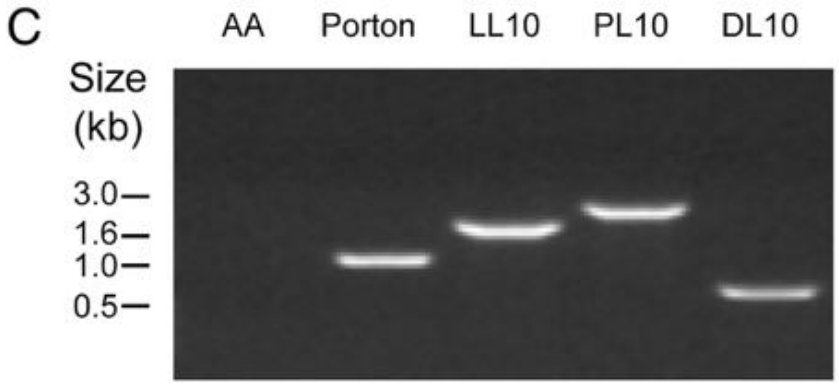
Morgan et al (2013) Trends in Genetics. http://doi.org/10.1016/j.tig.2012.09.005

Amerithrax Analysis



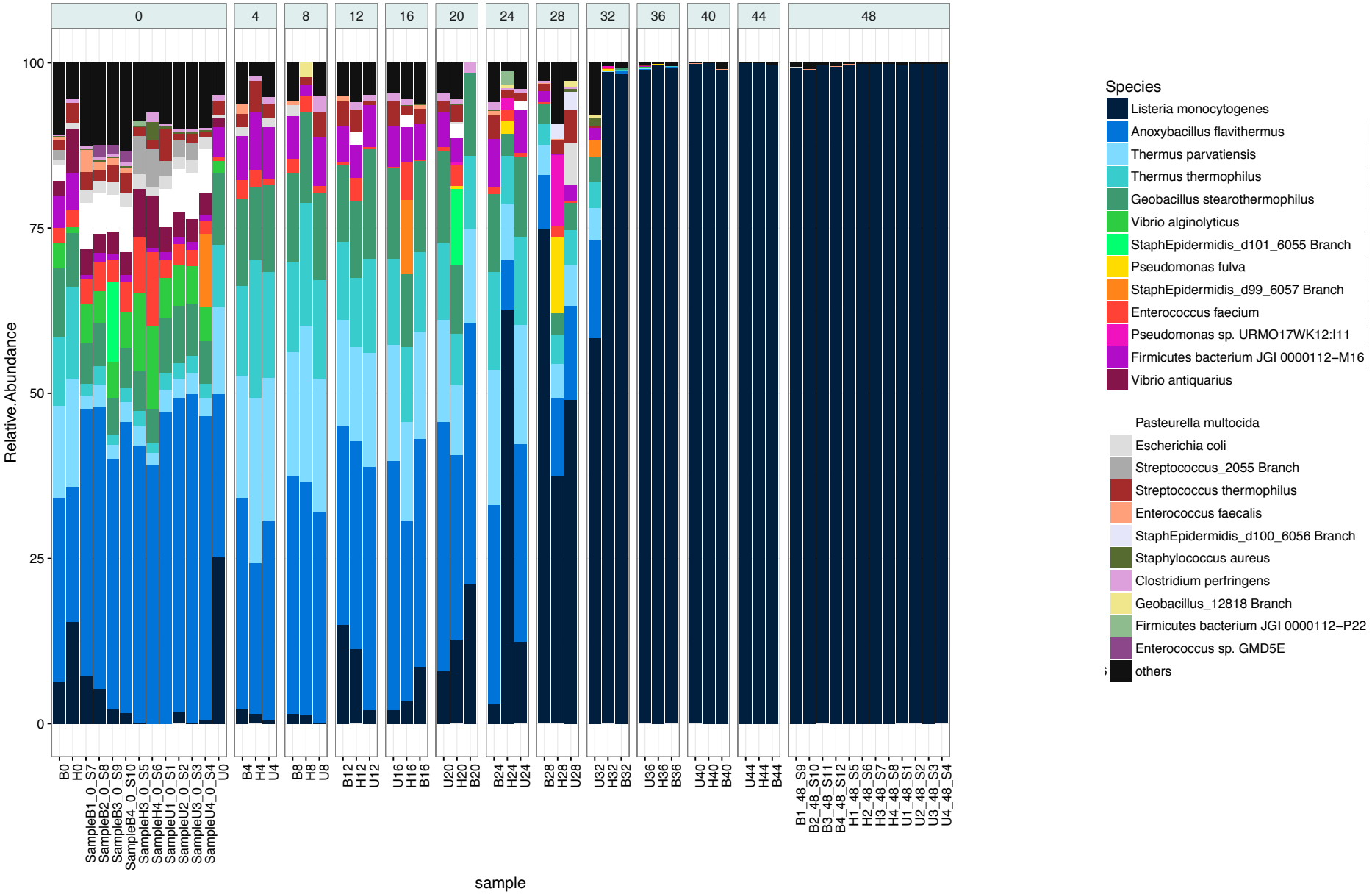






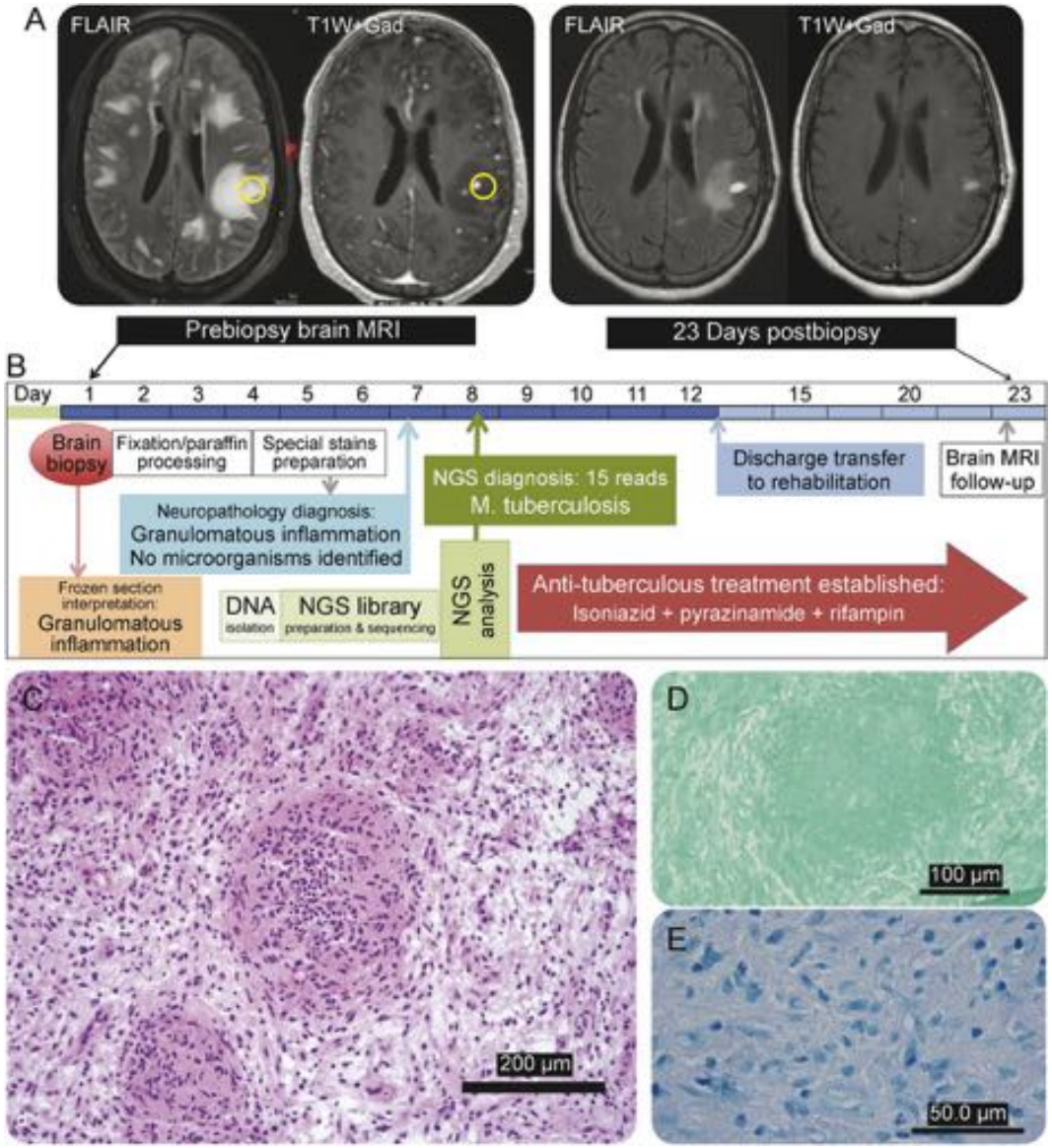
Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation
Rasko et al (2011) PNAS. doi: 10.1073/pnas.1016657108

Listeria in ice cream



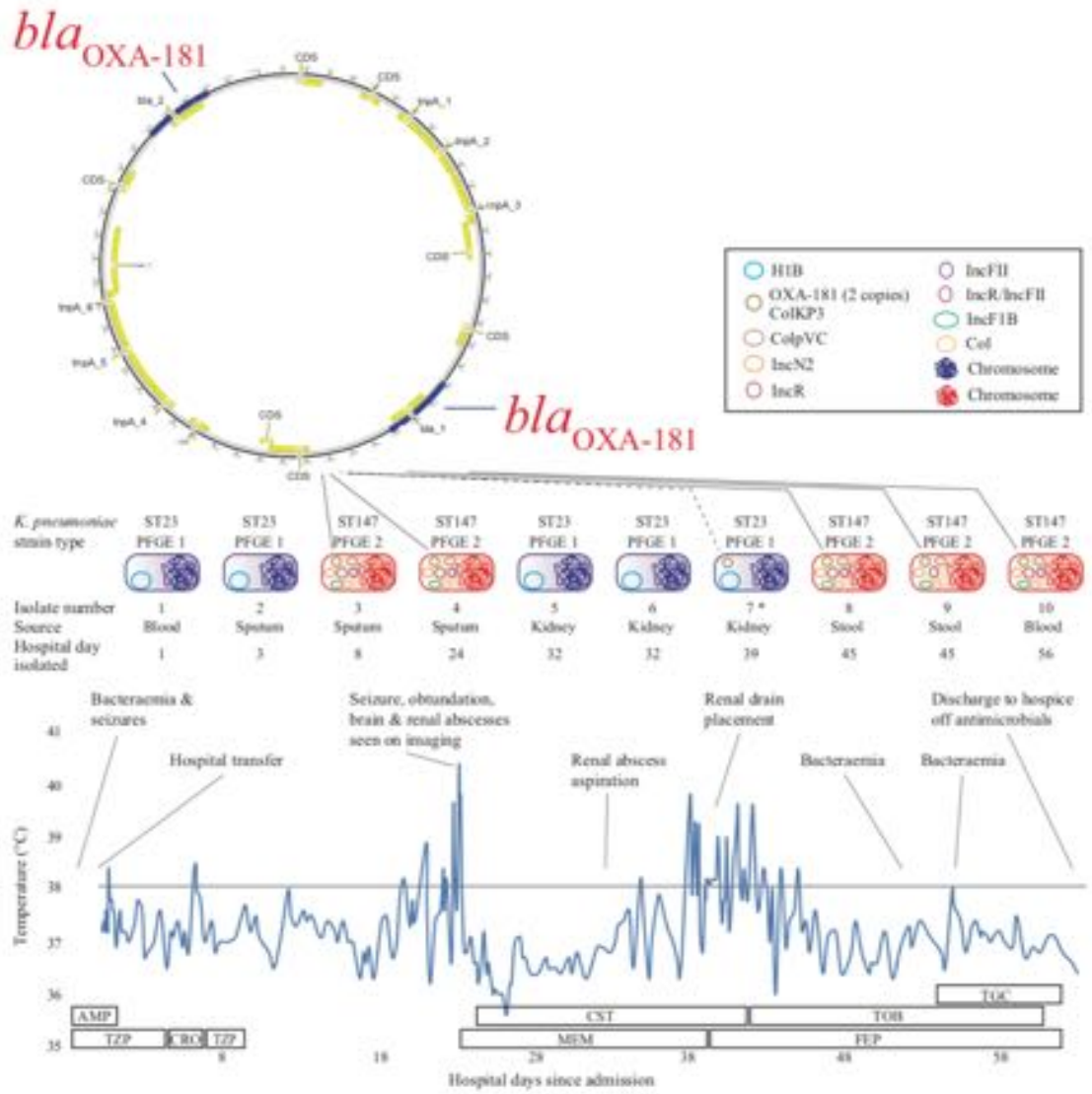
COSMOSID®

Diagnosing Brain Infections with NGS

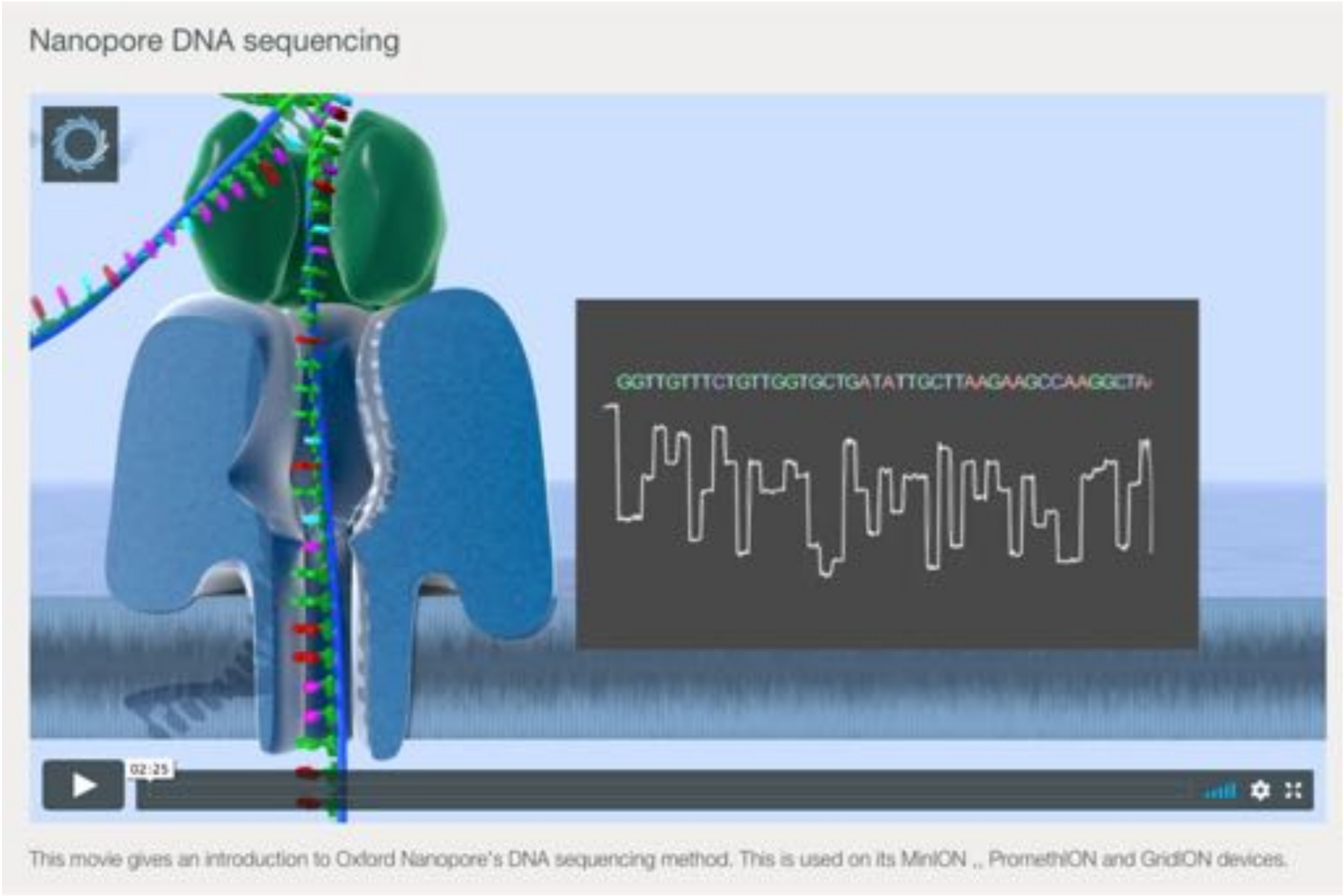


Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system
Salzberg et al (2016) Neurol Neuroimmunol Neuroinflamm dx.doi.org/10.1212/NXI.0000000000000251

Diagnosing Lung Infections with Nanopore



Antibiotic pressure on the acquisition and loss of antibiotic resistance genes in Klebsiella pneumoniae
Simner et al (2018) *Journal of Antimicrobial Chemotherapy*, https://doi.org/10.1093/jac/dky121



https://nanoporetech.com/resource-centre/videos/nanopore-dna-sequencing

Genomic Futures?

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Severi[6]*, Lauren Cowley[7]*, Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraogo[9], Babak Afrough[2,10], Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar Cabeza-Cabrerizo[2,3], Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Isabel García-Dorival[2,15], Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Michel Koropogui[2,17], Abigael Kosgey[2,18], Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc Mertens[2,20], Janine Michel[2,12], Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Portmann[2,21], Johanna Gabriella Repits[2,22], Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel L. Yemanaberhan[2,3], Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Faye[26], Oumar Faye[26], N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Emily Davis[29,30], Jon Gerlach[29,30], Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Camara[31], Hermann Somlare[31], Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Delaune[32,33], Koumpingnin Yacouba Nebie[34], Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gutierrez[35], Natacha Milhano[6,36], Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], James Taylor[38], Phillip Rachwal[38], Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthew K. O'Shea[41], Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di Caro[2,19], Roman Wölfel[2,44], Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koivogui[45], Boubacar Diallo[34], Sakoba Keita[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. Carroll[2,10,48,49]

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Se
Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[6], Amy Mikhail[7], Nobila Ouédraog
Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Ma
Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Is
Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Miche
Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc M
Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Port
Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel
Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Fa
N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Er
Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Cam
Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Dela
Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gu
Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], Jam
Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthe
Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di
Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koi
Sakoba Keita[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. C

Figure 1 | Deployment of the portable genome surveillance system in Guinea. a, We were able to pack all instruments, reagents and disposable consumables within aircraft baggage. **b**, We initially established the genomic surveillance laboratory in Donka Hospital, Conakry, Guinea. **c**, Later we moved the laboratory to a dedicated sequencing laboratory in Coyah prefecture. **d**, Within this laboratory we separated the sequencing instruments (on the left) from the PCR bench (to the right). An uninterruptable power supply can be seen in the middle that provides power to the thermocycler. (Photographs taken by J.Q. and S.D.)

Ebola Surveillance

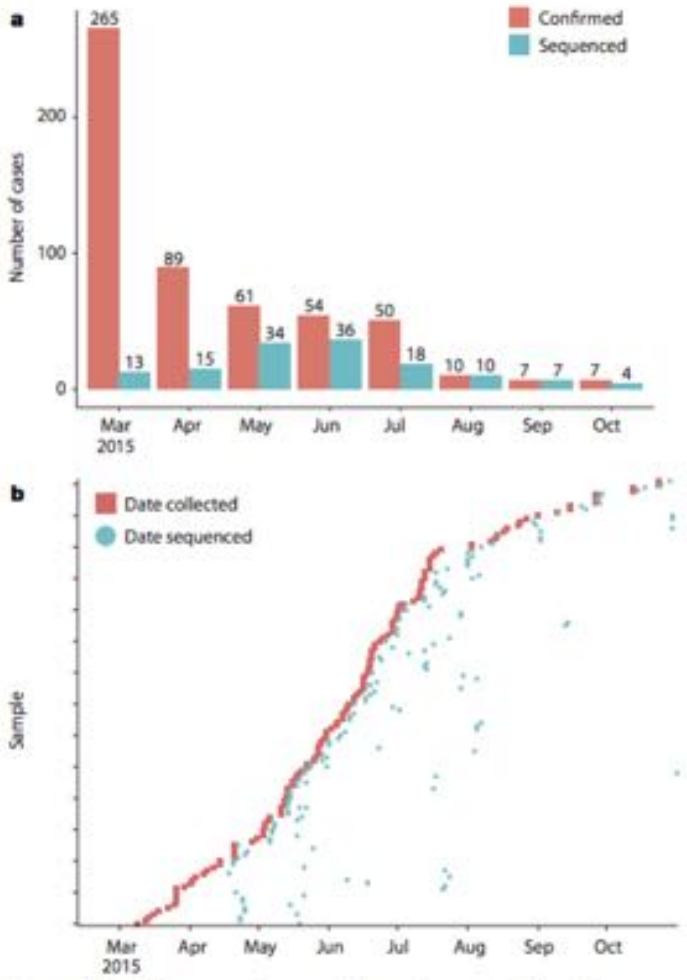


Figure 2 | Real-time genomics surveillance in context of the Guinea Ebola virus disease epidemic. a, Here we show the number of reported cases of Ebola virus disease in Guinea (red) in relation to the number of EBOV new patient samples ($n = 137$, in blue) generated during this study. **b,** For each of the 142 sequenced samples, we show the relationship between sample collection date (red) and the date of sequencing (blue). Twenty-eight samples were sequenced within three days of the sample being taken, and sixty-eight samples within a week. Larger gaps represent retrospective sequencing of cases to provide additional epidemiological context.

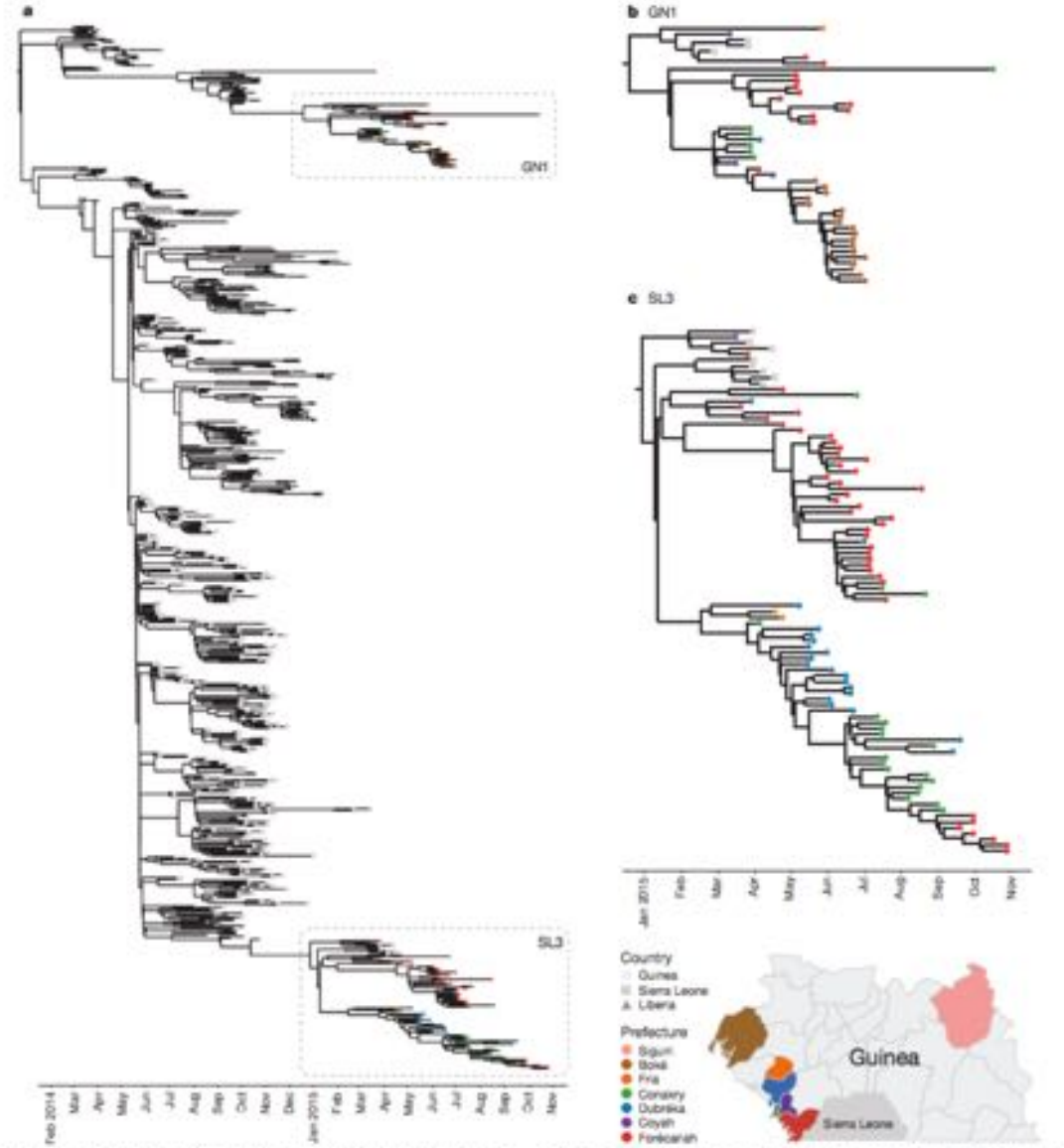


Figure 3 | Evolution of EBOV over the course of the Ebola virus disease epidemic. a, Time-scaled phylogeny of 603 published sequences with 125 high quality sequences from this study. The shape of nodes on the tree demonstrates country of origin. Our results show Guinean samples (coloured circles) belong to two previously identified lineages, GN1 and SL3. **b,** GN1 is deeply branching with early epidemic samples. **c,** SL3 is related to cases identified in Sierra Leone. Samples are frequently clustered by geography (indicated by colour of circle) and this provides information as to origins of new introductions, such as in the Boké epidemic in May 2015. Map figure adapted from SimpleMaps website (http://simplemaps.com/resources/svg-gn).

Genomic Futures?

Clive G. Brown @Clive_G_Brown · Aug 15
It's just the prototype ....

102 122

Clive G. Brown @Clive_G_Brown · Aug 15
It's does have flashing lights already ...

13 21

Genomic Futures?

The rise of a digital immune system
Schatz & Phillippy (2012) GigaScience 1:4