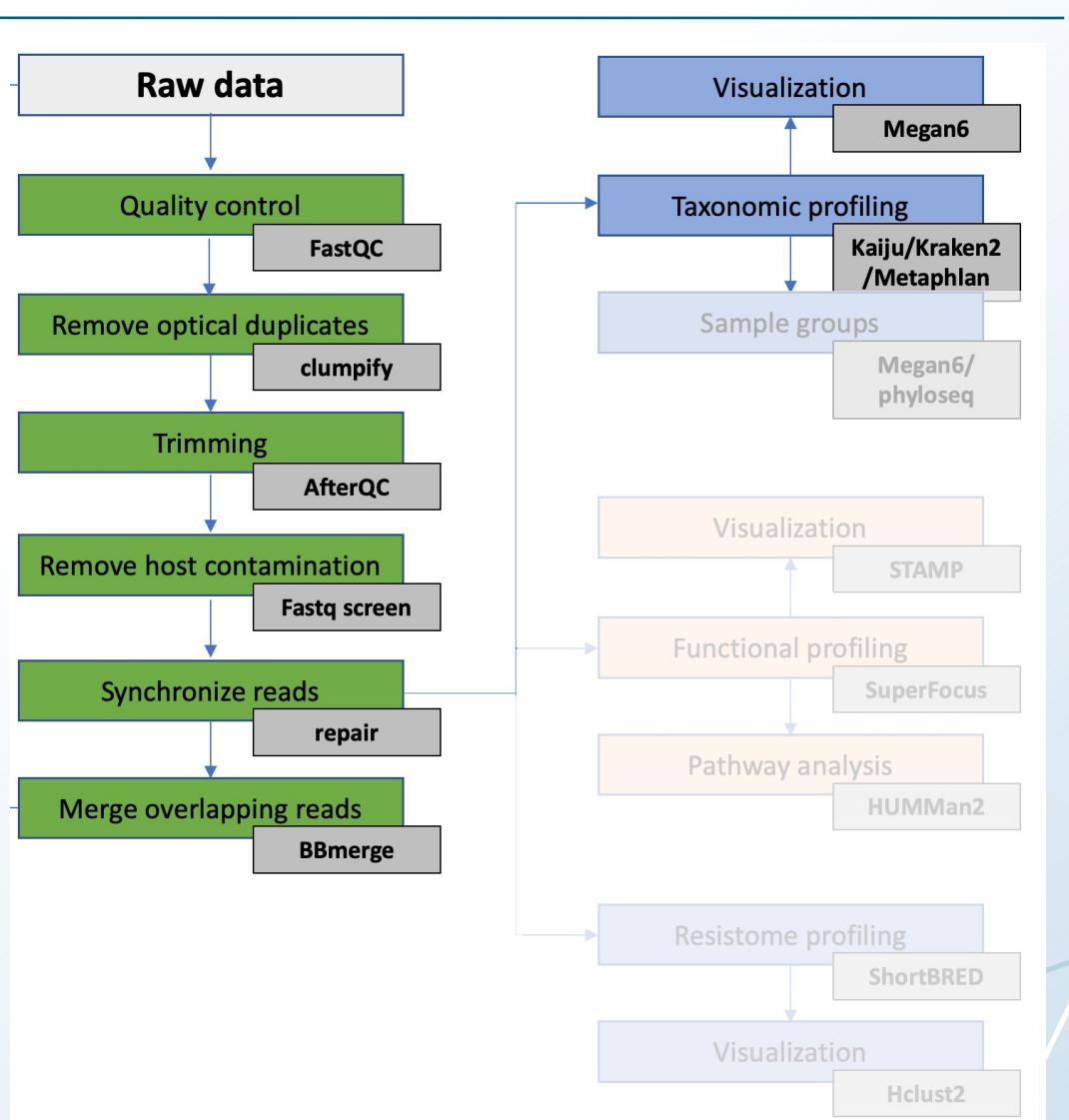
# Taxonomic assignment

# Overview of this talk

Taxonomic analysis of metagenomes

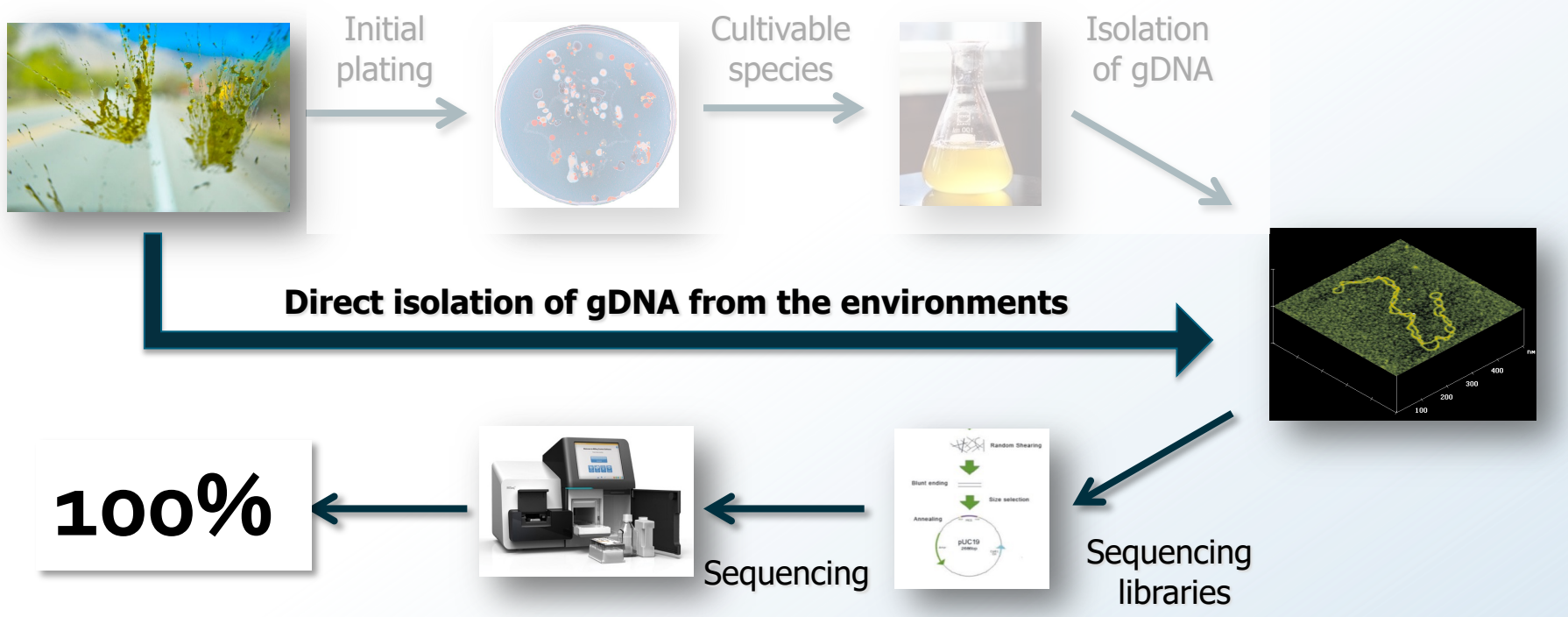
Visualization of taxonomic profiles

What's in the databases



# Recap - How do we study microbiomes?

Cultivation: Only 1% in most environmental samples



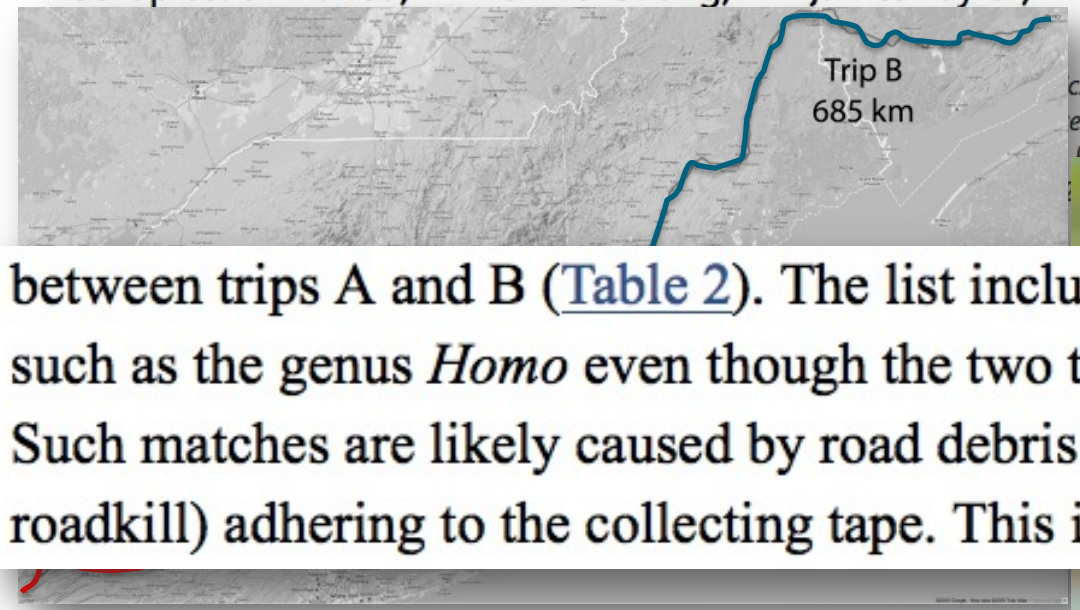
# A ”typical”☺ metagenomic study

Resource

## Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond,[1,2,6,9] Samir Wadhawan,[3,6,7] Francesca Chiaromonte,[4] Guruprasad Ananda,[1,3] Wen-Yu Chung,[1,3,8] James Taylor,[1,5,9] Anton Nekrutenko,[1,3,9]

*cs, School of Medicine University of*
*es, Penn State University, University Park,*




between trips A and B (Table 2). The list included unexpected entries such as the genus *Homo* even though the two trips were uneventful. Such matches are likely caused by road debris (which often includes roadkill) adhering to the collecting tape. This illustrates, at least at genus

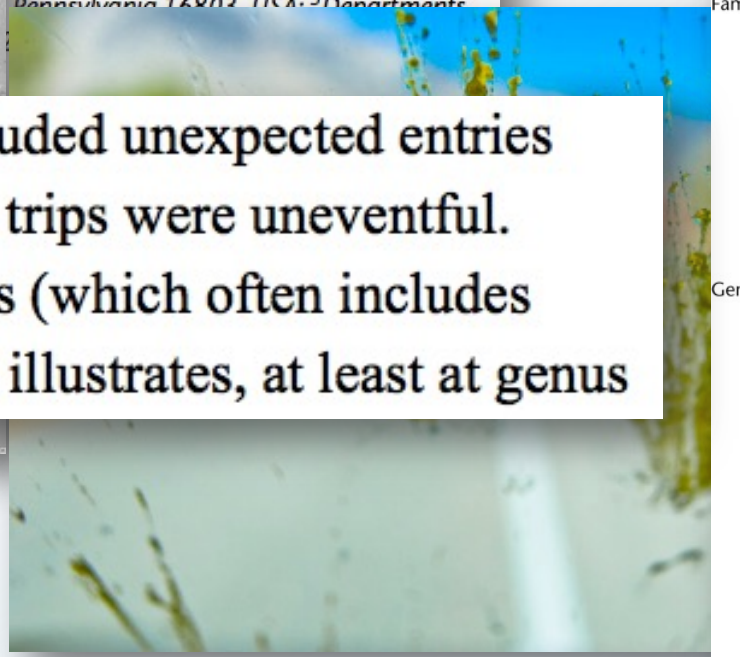


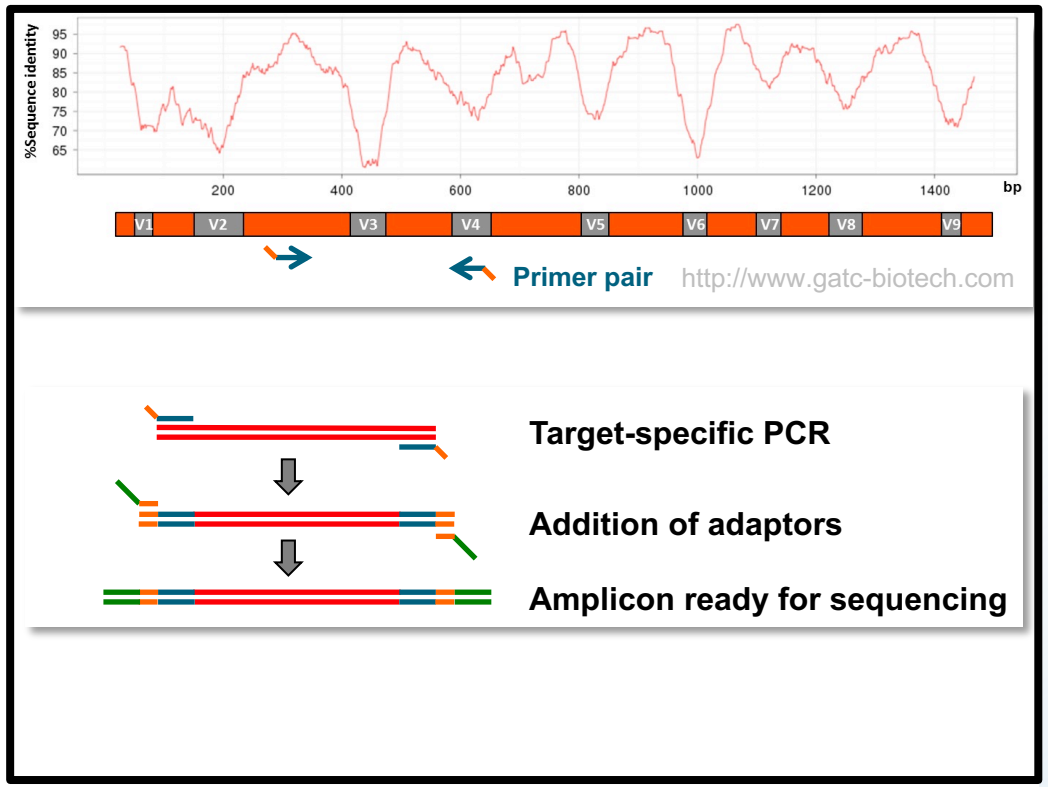
**Table 2.** Taxa with significant (at 1% level) differences in read abundance between trip A and trip B

| Rank | Name | Trip A | Trip B |
|---|---|---|---|
| Phylum | Arthropoda | 711 | 1531 |
| | Chordata | 300 | 272 |
| | Cnidaria | 10 | 87 |
| | Firmicutes | 12,927 | 5623 |
| | Proteobacteria | 45,946 | 24,663 |
| Class | Bacilli | 10,748 | 4004 |
| | Betaproteobacteria | 228 | 45 |
| | Clostridia | 2178 | 1616 |
| | Gammaproteobacteria | 44,934 | 24,413 |
| | Hydrozoa | 10 | 87 |
| | Insecta | 711 | 1516 |
| | Mammalia | 294 | 256 |
| Order | Aeromonadales | 540 | 21 |
| | Bacillales | 83 | 58 |
| | Clostridiales | 2178 | 1615 |
| | Diptera | 296 | 350 |
| | Enterobacteriales | 41,174 | 23,729 |
| | Hemiptera | 383 | 1027 |
| | Hydroida | 10 | 87 |
| | Lactobacillales | 10,643 | 3943 |
| | Primates | 112 | 10 |
| | Pseudomonadales | 1792 | 408 |
| | Rhodospirillales | 56 | 1 |
| Family | Aeromonadaceae | 540 | 21 |
| | Aphididae | 382 | 1016 |
| | Clostridiaceae | 2170 | 1608 |
| | Culicidae | 86 | 64 |
| | Drosophilidae | 32 | 95 |
| | Enterobacteriaceae | 41,172 | 23,729 |
| | Enterococcaceae | 706 | 1512 |
| | Hominidae | 97 | 6 |
| | Hydridae | 10 | 87 |
| | Lactobacillaceae | 5837 | 209 |
| | Leuconostocaceae | 2978 | 1498 |
| | Pseudomonadaceae | 1703 | 391 |
| | Streptococcaceae | 928 | 545 |
| Genus | *Acyrthosiphon* | 381 | 995 |
| | *Aeromonas* | 540 | 21 |
| | *Anopheles* | 80 | 45 |
| | *Anopheles* | 80 | 1 |
| | *Buchnera* | 9 | 59 |
| | *Clostridium* | 2170 | 1607 |
| | *Drosophila* | 31 | 94 |
| | *Enterobacter* | 4142 | 5507 |
| | *Enterococcus* | 706 | 1511 |
| | *Erwinia* | 2 | 240 |
| | *Homo* | 96 | 4 |
| | *Hydra* | 10 | 87 |
| | *Klebsiella* | 15,169 | 1695 |
| | *Lactobacillus* | 5740 | 167 |
| | *Lactococcus* | 809 | 509 |
| | *Leuconostoc* | 2971 | 1496 |
| | *Photorhabdus* | 57 | 1 |
| | *Providencia* | 123 | 3 |
| | *Pseudomonas* | 1648 | 390 |

# Two methods for performing taxonomic profiling of microbiomes

## Amplicon sequencing (16S rRNA) and random sequencing

**16S rRNA amplicon sequencing**



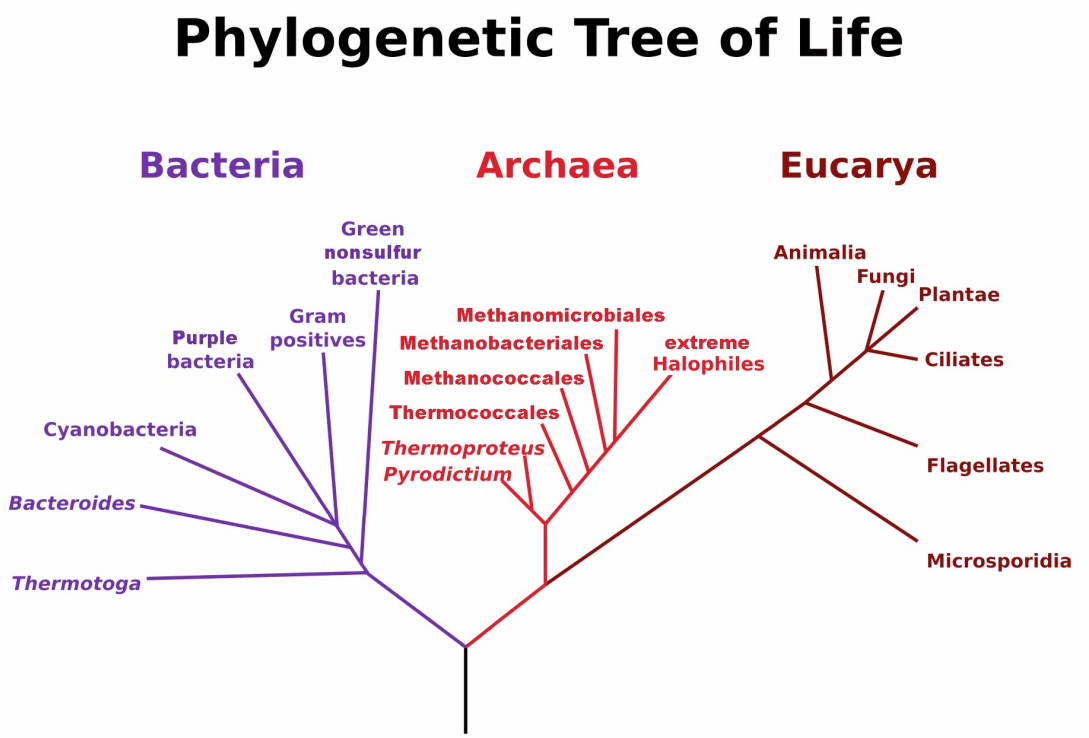

**Random sequencing**

# Amplicon vs random sequencing

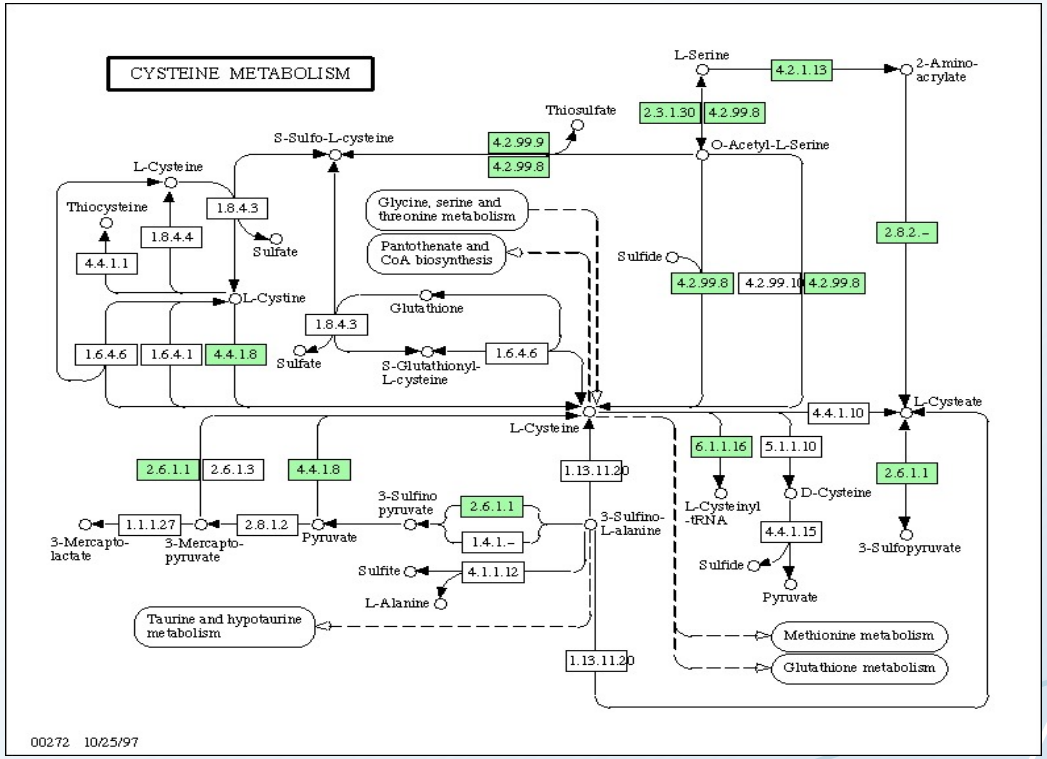
It depends on what you want to know

Main difference: taxonomic profile vs taxonomic and functional profile

**16S rRNA amplicon sequencing**




**Random sequencing**



# Amplicon vs random sequencing

16S primers are not universal - 16S rRNA amplification differences lead to biased estimates of relative abundance

This can give an over-representation or under-representation of sequences in the some genera

Eg. Clostridium and Lactobacillus contain sequences that are perfectly complementary to the primers used for amplification

Sequences in the Enterobacteriaceae family and the Clostridiales order poorly resolves using the 16S V4 or V3-V4 regions

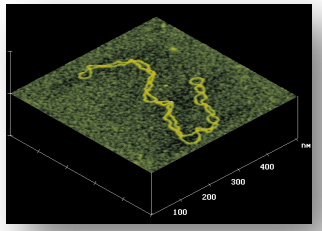
# Amplicon vs random sequencing – pros and cons

| | 16S amplicon | Random |
|---|---|---|
| Analysis of large number of samples | pro | con |
| Depth - resolution | pro | con |
| Computational resources (and skills) | pro | con |
| Expenses | pro | con |
| PCR amplification bias | con | pro |
| Discovery of new bacterial genes and genomes | con | pro |
| Simultaneous study of several domains | con | pro |

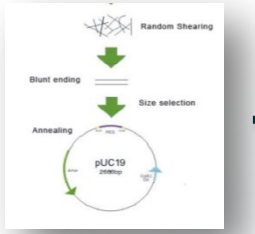
# How is taxonomic classification done?

The study of genetic material directly from a sample

Direct isolation of gDNA



Sequencing library




Sequencing

LFI5006_S7_L001_R1_001.fastq

```
@M00435:6:000000000-A23K2:1:1101:16705:1437 1:N:0:7
TTTCAGCATCGATGAATAATCTCCTAACTCTGAGTTCAGGGGGAAGATCGTAATCTGACAAATCTTTCATAAAAACAAAGCTTTCTTCCTCATCCATCACATCGATAATATTATTGACTGACAATAAATAACGCTCTTTTGCTAAGCCAT
+
=====>>>@@<@@A@@EECE>AE8EEEEEEDB.EFFFBDDE7CC+DEFDD7A>AE@E@CFFEFFEF5CFFFFFFFEEEEEEACA>>EDEEDEEEBEEDBDE,,++CD+3DEDDD++=+==+=+++4;2DDDE+0D@@**339898*1;*0
@M00435:6:000000000-A23K2:1:1101:15725:1441 1:N:0:7
CGCATTGCAATAACTAAGCTGCTTTTATACTGCTCTTGCAATGGTTTTAGACGGCGAGCGTAATCTAACGC
+
?<????BBDBDDBBDBF/CFF?CFFHFBFFFGG0AGHDHHAFHH?CFFDEGG@EHB>CACHE@FFGGGCEC
@M00435:6:000000000-A23K2:1:1101:17134:1445 1:N:0:7
CTACAGACTATTCATAATTTCGCAACGGTAAGTTGCAAAAGAGCAAGTAAACGACGATCAGGTAAACACATATTTACTTAAGTATAATAGTGATCAACTTCTCACTTCAATACCCAAACAAAAACTTAACTGTTCAAAAACTAAGATTACA
+
?????BBBDBDDDDDDFFFFFBEEHHHH6A@GHGFHHIHIFHHHHFHFFFHIDHHHH@EEHHEEDDFHHHGHHFHHIHIFHFEDFFFFHHFGGHH=CFHHCGFFHHHHH>DHCFFFFHHHFFFFEFFEEFDFFDDEFFEDBEBEE=5A,==
@M00435:6:000000000-A23K2:1:1101:16665:1448 1:N:0:7
GTGTTACAGTTATATTGATATTGCCTAAGGTCTTTTTATATTCAGTAATCCACTCAGCTTTATTATCAGCTGCTTGCTTTAATTCCAAGCGGCGTAGCTCTGAACGTGTTAGTGGTTTTGCGTCAGACTCTGATTTAGTTAAACCAATGG
+
?????BB<<<<BBBB?;0CACFEC>F;C>F0AFHFH=FFGF?FE=FDFHHFGFHF?AF//AC9CDGFG=AEGDCFE?ECGFDGGGFGH.C>7>>CACDHFHFF.7,6DD4@@7@,74?+?D6AE@E,=,=D,BD;,=BB;33;,33A,;,
@M00435:6:000000000-A23K2:1:1101:16400:1450 1:N:0:7
GTATTAATGGAGTCGTTGATGGCACCTATTATGCCTCAGTATATGACAGCTGTGGCTGAGTCGGCTGGACTGGGTTCGAGTGCATCTTCGATTTTCTTTACGTTATATTTTATTGGCTTTGCGCTGACGTTACTACCAGCAGCTAAGTTGA
+
?????BBBDDDDDDDDG?FFFGIIIIIIIHGHIIIIHHIEHHIIIIIHHHHHHFHIIIIHHFHHHIHHFCFFHIHEFHFBACBCGHHIFHHIHHHHHHHFD=DCFHHDFFHDFHB.CBFFB?D8>EGEGEGGEEEEA>A<-ACE>C>55A.
@M00435:6:000000000-A23K2:1:1101:16771:1451 1:N:0:7
GGACAACGTAACAGCCAATGTCATGGTGGCGGATGGGCAACGTAATATTATTTATATGAATGAGTCGGTTGATAAATTACTTAAAAATGCTGAAAAAGATTTCAAAAAAGAGCTACCCGACTTTAATGTAGACAATTTACTCGGCTCAAGT
+
?????B?@DDDDDDDDFFFFCFCFFHBCCFCECEEHDFFHACFD@FFFFHHIIIIIHHIHHIIFFF>7CECHFBGGHIBFGHDFHGFHHBDFFB=DFEFFDDFB.@@DE@;=,6==D<D<@BEF=ACA=ABA,5=AE,=5A*)08AAA?:*
@M00435:6:000000000-A23K2:1:1101:16128:1456 1:N:0:7
AATGTATACTTCAGATAATATTCAATTAATTGGCAATAGGGAAGAGACAAAAACTCGTACCGGTCTTCAGATCGAAGAGGGAAATGAGCAGATTAAATTACACGATAATCTTGTTTTCCATCATGAAAAGAAAAAGTGTAGCCGCAACCTT
+
55=55<77@<A@@-@@EEEEEC>C8>CCE899-8-A-9A=,-77AEEEDEDEEE@F7>@++7>5+5A--5@C-5>C++5C@E+8A====-5AAE-C=,,,6=,<)++4++4+6=++66=:+4++++4+++1*31@9*******2)))2*9*
@M00435:6:000000000-A23K2:1:1101:17464:1467 1:N:0:7
CAATTAATACTTTTAATCGTACGGGTTGATAACGATTGCGACGATACAGCAGAAACAGCGGTACATTCGGTGAGTGCCATTGTTCTTCATGCGATGCTGATATTAAAGCAGGGGCTAAGGTATTTTCCTTAATCTCAGTGTCGCATTACAG
+
?????BBB<B<<BBBB?CFACC>EHCE>F>CD/@>AFFDE>+5+,55CCFHDFDHFFFHHH+>C-55CFFEEF,5,,@,CF,@CFF=.@C?C+5CDEHFB?,??BD;B?DD*6:));BB,,,,3?,;,?,,3,3,,,;,,;;A*4)0:??*
@M00435:6:000000000-A23K2:1:1101:16930:1468 1:N:0:7
ATCTGAACCAACTGTAGATAGGTTACCCGAGATACCACTTACTTTTTGATACTCAGGTAATTTAGGGTCTAGTCCTGCTGCAAATGCGCTTTGTGAAGCGATTGCTGTCATTGATATCCCTAATGCTACAACAATTTGTTTAATATTCATT
+
```

# How is taxonomic classification done?

Each sequence read is a tiny genomic fragment from a specie in the sample

In a metagenome a sequence read is basically representing a specie

## Sample

AGTCCAGGTAACGTTACAACG

LFI5006_S7_L001_R1_001.fastq

```
@M00435:6:000000000-A23K2:1:1101:16705:1437 1:N:0:7
TTTCAGCATCGATGAATAATCTCCTAACTCTGAGTTCAGGGGGAAGATCGTAATCTGACAAATCTTTCATAAAAACAAAGCTTTCTTCCTCATCCATCACATCGATAATATTATTGACTGACAATAAATAACGCTCTTTTGCTAAGCCAT
+
=====>>>@@<@@A@@EECE>AEBEEEEEEDB.EFFFBDDE7CC+DEFDD7A>AE@E@CFFEFFEF5CFFFFFFFEEEEEEACA>>EDEEDEEEBEEDBDE,,++CD+3DEDDD++=+==+=+++4;2DDDE+0D@@**339898*1;*0
@M00435:6:000000000-A23K2:1:1101:15725:1441 1:N:0:7
CGCATTGCAATAACTAAGCTGCTTTTATACTGCTCTTGCAATGGTTTTAGACGGCGAGCGTAATCTAACGC
+
?<????BBDBDDBBDBF/CFF?CFFHFBFFFGG0AGHDHHAFHH?CFFDEGG@EHB>CACHE@FFGGGCEC
@M00435:6:000000000-A23K2:1:1101:17134:1445 1:N:0:7
CTACAGACTATTCATAATTTCGCAACGGTAAGTTGCAAAAGAGCAAGTAAACGACGATCAGGTAAACACATATTTACTTAAGTATAATAGTGATCAACTTCTCACTTCAATACCCAAACAAAAACTTAACTGTTCAAAAACTAAGATTACA
+
?????BBBDBDDDDDDFFFFFBEEHHHH6A@GHGFHHIHIFHHHHFHFFFHIDHHHH@EEHHEEDDFHHHGHHFHHIHIFHFEDFFFFHHFGGHH=CFHHCGFFHHHHH>DHCFFFFHHHFFFFEFFEEFDFFDDEFFEDBEBEE=5A,==
@M00435:6:000000000-A23K2:1:1101:16665:1448 1:N:0:7
GTGTTACAGTTATATTGATATTGCCTAAGGTCTTTTTATATTCAGTAATCCACTCAGCTTTATTATCAGCTGCTTGCTTTAATTCCAAGCGGCGTAGCTCTGAACGTGTTAGTGGTTTTGCGTCAGACTCTGATTTAGTTAAACCAATGG
+
?????BB<<<<BBBB?;0CACFEC>F;C>F0AFHFH=FFGF?FE=FDFHHFGFHF?AF//AC9CDGFG=AEGDCFE?ECGFDGGGFGH.C>7>>CACDHFHFF.7,6DD4@@7@,74?+?D6AE@E,=,=D,BD;,=BB;33;,33A,;,
@M00435:6:000000000-A23K2:1:1101:16400:1450 1:N:0:7
GTATTAATGGAGTCGTTGATGGCACCTATTATGCCTCAGTATATGACAGCTGTGGCTGAGTCGGCTGGACTGGGTTCGAGTGCATCTTCGATTTTCTTTACGTTATATTTTATTGGCTTTGCGCTGACGTTACTACCAGCAGCTAAGTTGA
+
?????BBBDDDDDDDDG?FFFGIIIIIIIHGHIIIIHHIEHHIIIIIHHHHHHFHIIIIHHFHHHIHHFCFFHIHEFHFBACBCGHHIFHHIHHHHHHHFD=DCFHHDFFHDFHB.CBFFB?D8>EGEGEGGEEEEA>A<-ACE>C>55A.
@M00435:6:000000000-A23K2:1:1101:16771:1451 1:N:0:7
GGACAACGTAACAGCCAATGTCATGGTGGCGGATGGGCAACGTAATATTATTTATATGAATGAGTCGGTTGATAAATTACTTAAAAATGCTGAAAAAGATTTCAAAAAAGAGCTACCCGACTTTAATGTAGACAATTTACTCGGCTCAAGT
+
?????B?@DDDDDDDDFFFFCFCFFHBCCFCECEEHDFFHACFD@FFFFHHIIIIIHHIHHIIFFF>7CECHFBGGHIBFGHDFHGFHHBDFFB=DFEFFDDFB.@@DE@;=,6==D<D<@BEF=ACA=ABA,5=AE,=5A*)08AAA?:*
@M00435:6:000000000-A23K2:1:1101:16128:1456 1:N:0:7
AATGTATACTTCAGATAATATTCAATTAATTGGCAATAGGGAAGAGACAAAAACTCGTACCGGTCTTCAGATCGAAGAGGGAAATGAGCAGATTAAATTACACGATAATCTTGTTTTCCATCATGAAAAGAAAAAGTGTAGCCGCAACCTT
+
55=55<77@<A@@-@@EEEEEC>C8>CCE899-8-A-9A=,-77AEEEDEDDEE@F7>@++7>5+5A--5@C-5>C++5C@E+8A====-5AAE-C=,,,6=,<)++4++4+6=++66=:+4++++4+++1*31@9*******2)))2*9*
@M00435:6:000000000-A23K2:1:1101:17464:1467 1:N:0:7
CAATTAATACTTTTAATCGTACGGGTTGATAACGATTGCGACGATACAGCAGAAACAGCGGTACATTCGGTGAGTGCCATTGTTCTTCATGCGATGCTGATATTAAAGCAGGGGCTAAGGTATTTTCCTTAATCTCAGTGTCGCATTACAG
+
?????BBB<B<<BBBB?CFACC>EHCE>F>CD/@>AFFDE>+5+,55CCFHDFDHFFFHHH+>C-55CFFEEF,5,,@,CF,@CFF=.@C?C+5CDEHFB?,??BD;B?DD*6:));BB,,,,3?,;,?,,3,3,,,;,,;;A*4)0:??*
@M00435:6:000000000-A23K2:1:1101:16930:1468 1:N:0:7
ATCTGAACCAACTGTAGATAGGTTACCCGAGATACCACTTACTTTTTGATACTCAGGTAATTTAGGGTCTAGTCCTGCTGCAAATGCGCTTTGTGAAGCGATTGCTGTCATTGATATCCCTAATGCTACAACAATTTGTTTAATATTCATT
+
```

# How is taxonomic classification done?

Compare your sample against a database of known species

**Sample**

AGTCCAGGTAACGTTACAACG

**Compare**

**Database**

# How is taxonomic classification done?

Compare your sample against a database of known species

## Sample

## Compare

## Database

# Create a taxonomic profile

Quantify occurrences

Unknown

| Sample | 4 | 2 | 5 | 5 | 6 | | 2 |
|---|---|---|---|---|---|---|---|

# Compare taxonomic profiles

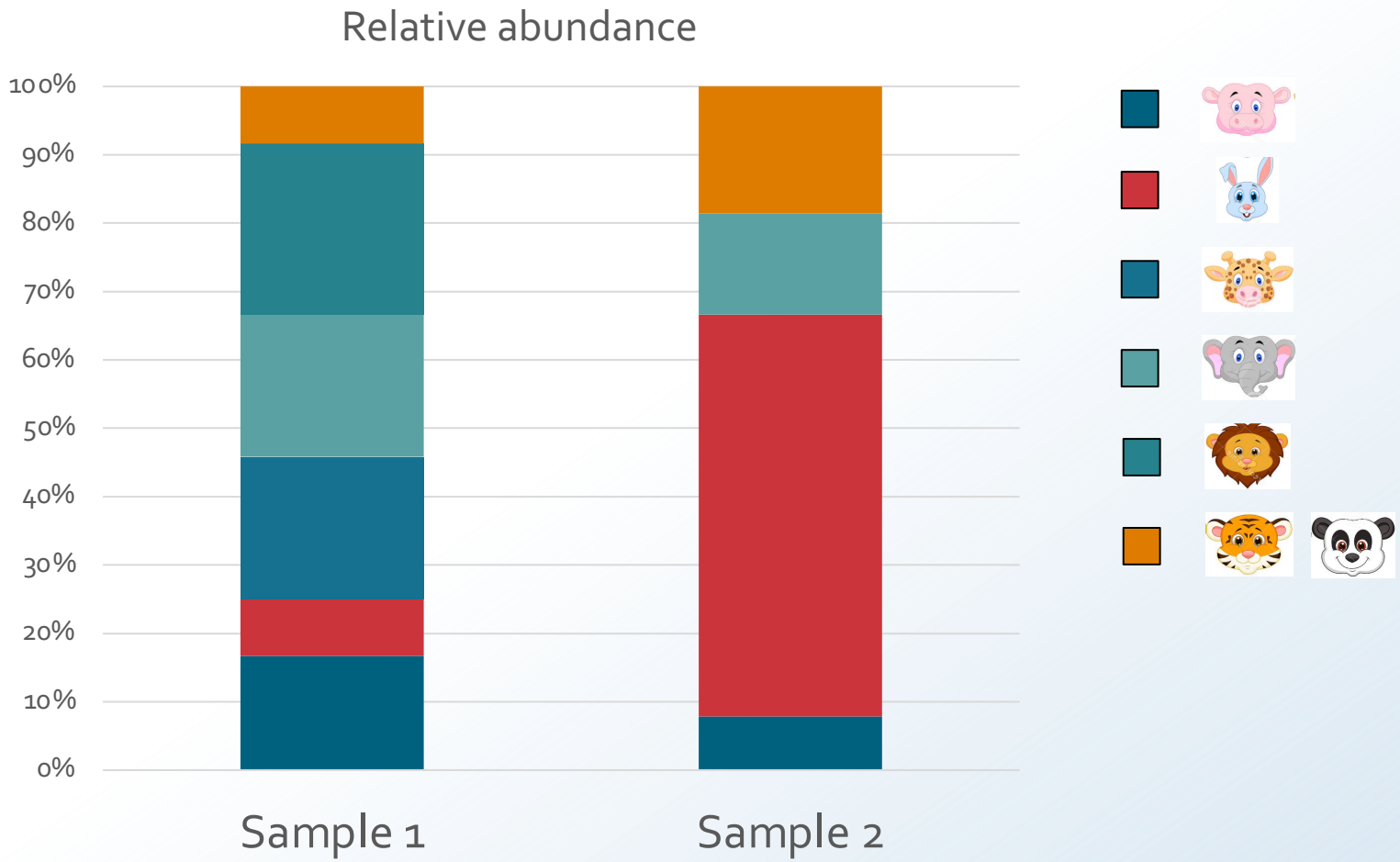
Compare two or more samples

Unknown

| Sample | 4 | 2 | 5 | 5 | 6 | | 2 |
|---|---|---|---|---|---|---|---|
| Sample 2 | 1 | 10 | 0 | 2 | 0 | | 3 |

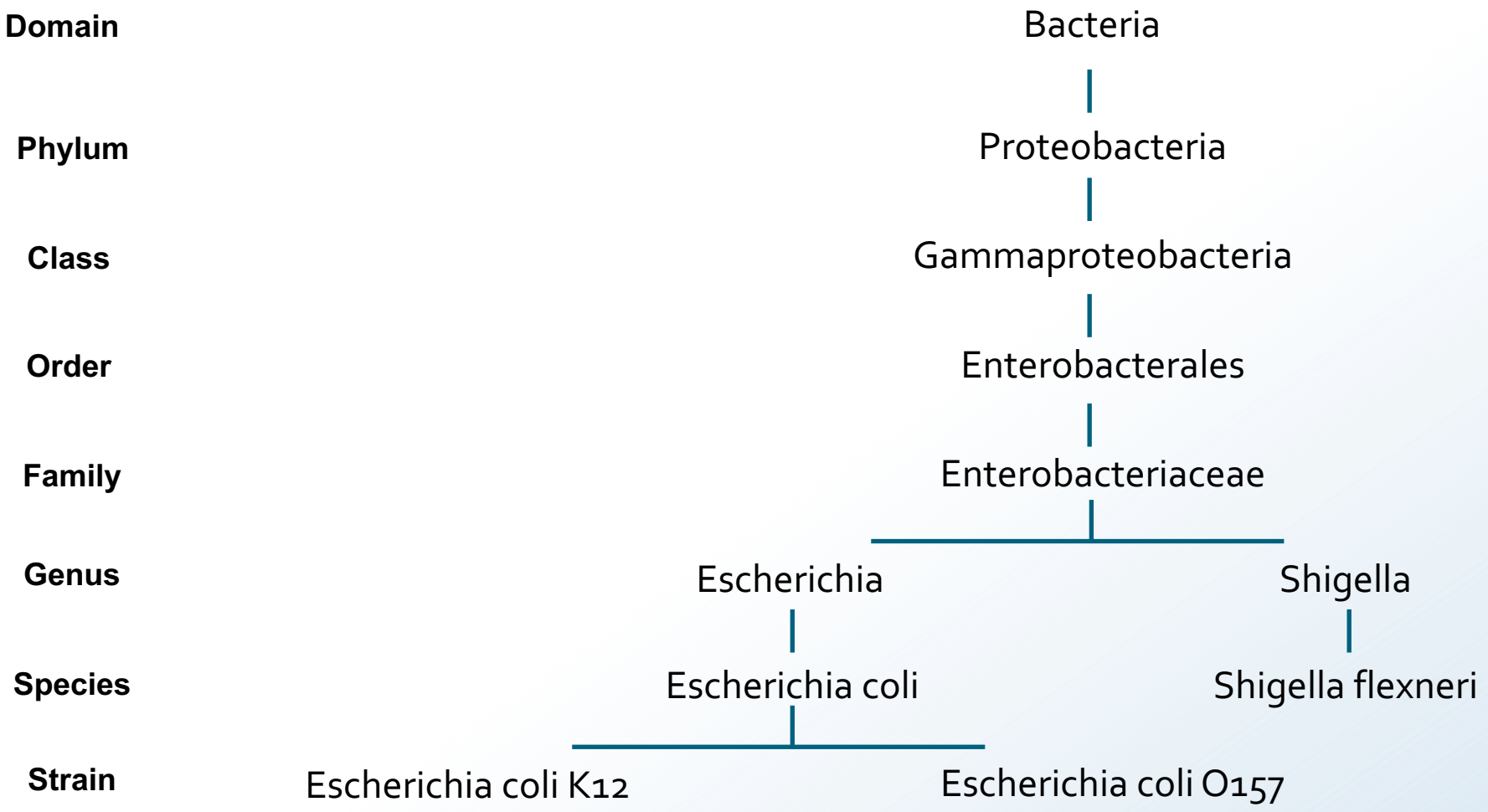
# Compare taxonomic profiles

## Compare two or more samples – relative abundance



# The taxonomy of species that contain highly similar sequences will be more difficult to resolve

When reads are too similar, they are assigned at higher levels of the taxonomy tree




Tools to hierarchically classify pre-aligned sequences on a taxonomy tree using an LCA algorithm:

- Qiime
- LCAClassifyer
- Megan
- Lots of other
- Motur

# Taxonomic profiling – K-mer based search

Kraken is a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads

- Using exact alignments of k-mers
- Kraken's default database contains just under 14 billion distinct k-mers, and requires at least 500GB of disk space (Oct 2017).
- Kraken requires enough free memory to hold the database in RAM. The default database size is 174GB (Oct 2017), and so you will need at least that much RAM if you want to build or run with the default database.
- When Kraken is run with a reduced database, it is called MiniKraken

# Taxonomic profiling – Search against protein databases

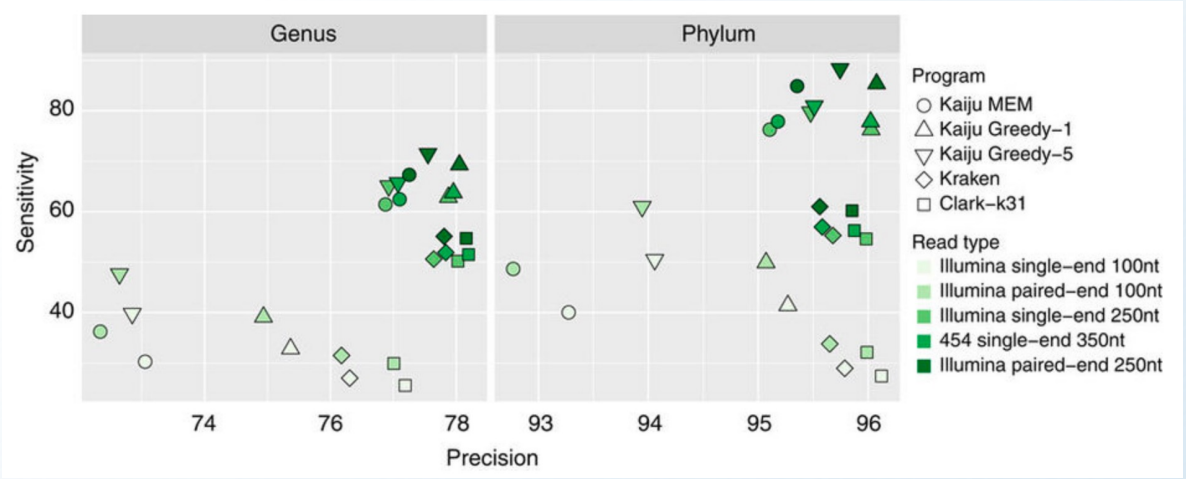
Kaiju is a taxonomic sequence classifier that use a reference database of protein sequences

- Finds maximum matches on the protein-level using the Burrows–Wheeler transform
- Reads are directly assigned to taxa using the NCBI taxonomy and a reference database of protein sequences from microbial and viral genomes
- Kaiju can be installed locally or used via a web server
- Can be run against various databases (eg. NCBI RefSeq)
- It can also be run against the Mar databases from the Marine Metagenomics Portal

# Taxonomic profiling – Search against protein databases

Kaiju is a taxonomic sequence classifier that use a reference database of protein sequences

Claim to be faster and more sensitive than K-mer based methods


Peter Menzel Nature Communications 7, Article number: 11257 (2016)

# Taxonomic profiling - Clade-specific markers

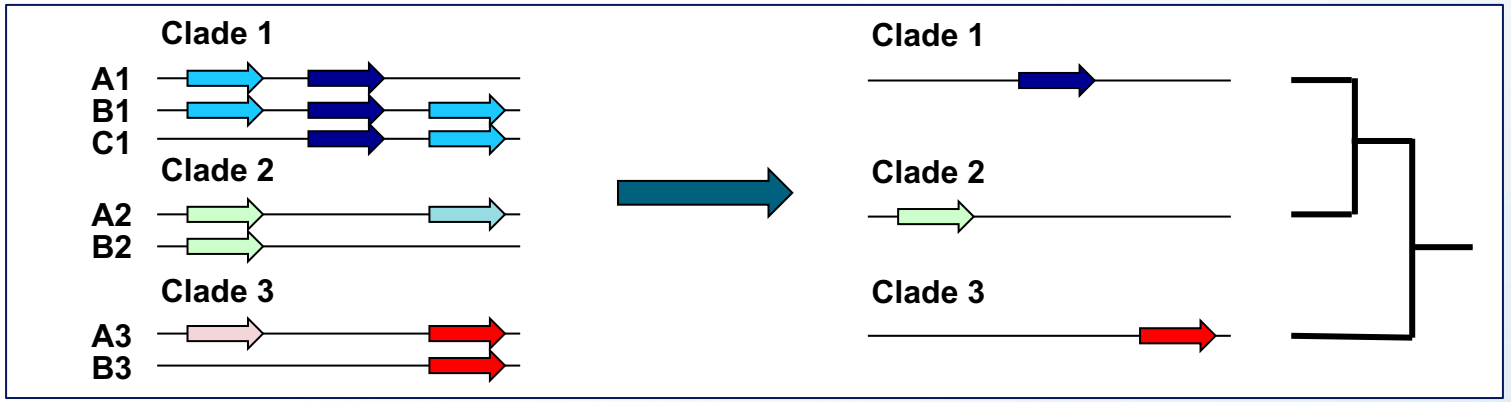
MetaPhlAn2 is a taxonomic sequence classifier that use a clade-specific marker database

- Using read coverage of clade-specific markers to detect the taxonomic clades present in a microbiome sample and estimate their relative abundance
- Map reads against clade-specific marker sequences that are pre-selected from coding sequences that identify specific microbial clades at the species or higher taxonomic levels
- The clade-specific markers cover all main functional categories
- MetaPhlAn2 includes ~1 million markers from >7,500 species

# Taxonomic profiling - Clade-specific markers

MetaPhlAn2 is a taxonomic sequence classifier that use a clade-specific marker database

- Dark blue is restricted yet universal across Clade 1
- Green genes are restricted to Clade 2, red genes to Clade 3



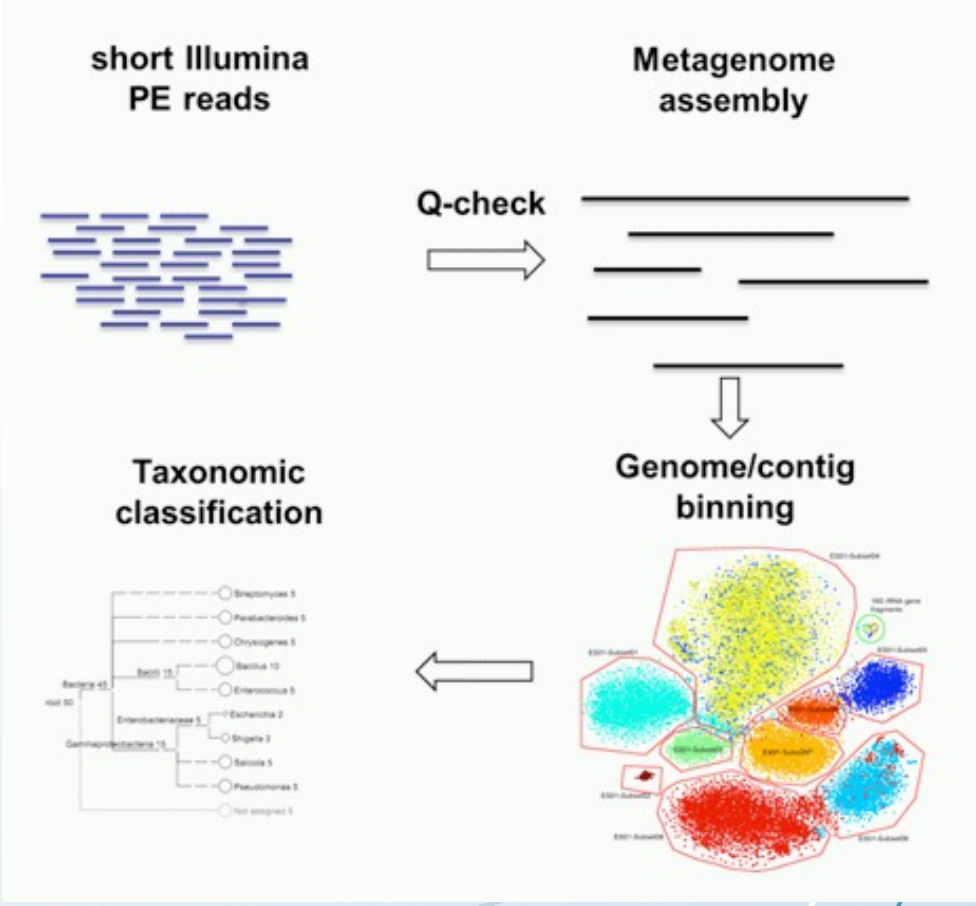
# Taxonomic binning

Clustering of assembled contigs that apparently originate from the same source population

- Assign to the closest possible taxonomy
- Enables the discovery of new microbial of new organisms

Tools for binning of contigs

- MaxBin
- MyCC
- Metawatt
- MetaBAT



# CAMI - Compared taxonomic profilers – not binning

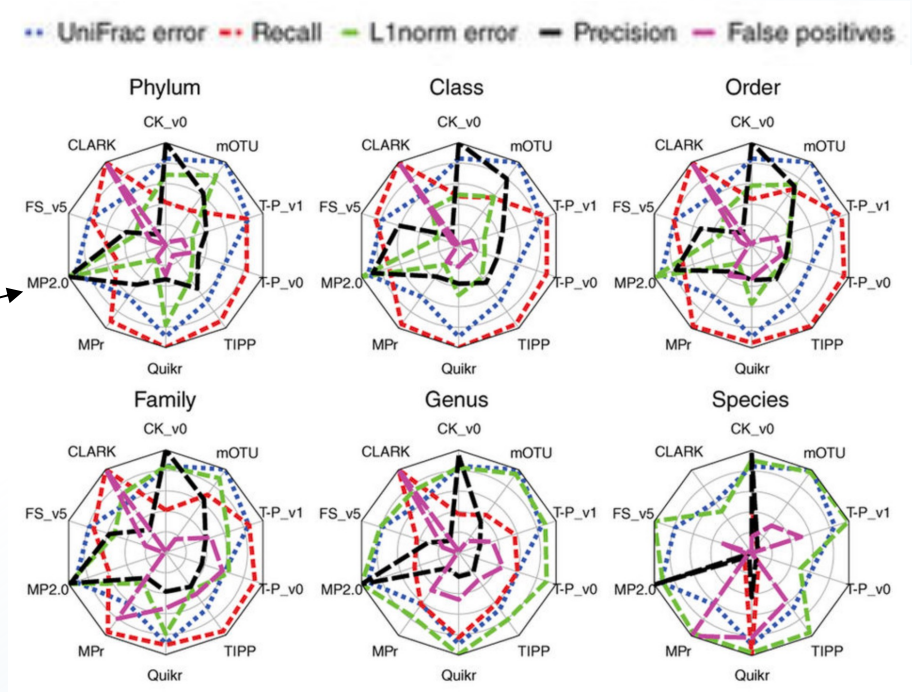
Profilers fell into three categories:

(i) profilers that correctly predicted relative abundances

(ii) precise profilers

(iii) profilers with high recall

Highest precision




*Nature Methods* **14**, 1063–1071 (2017)

# Technical variations influence results

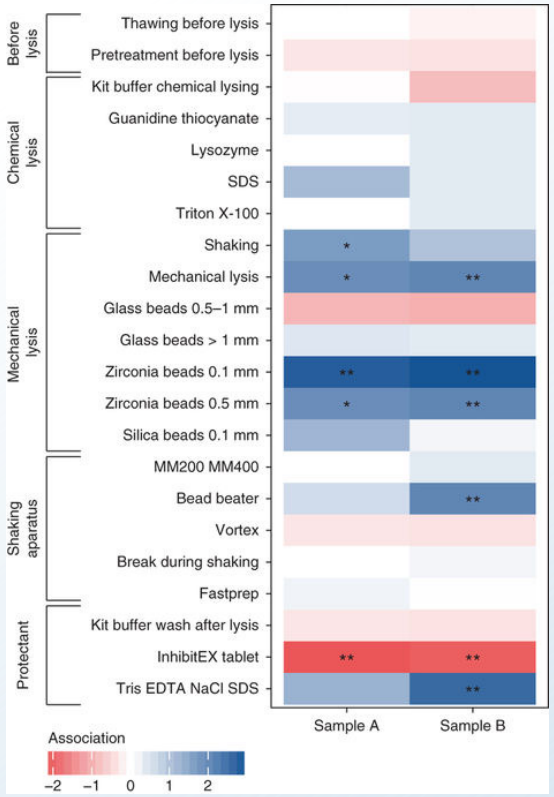
DNA extraction had the largest effect on the outcome of metagenomic analysis

Effects of protocol manipulations on sample composition

**SAMPLE 1 EXTRACTION A**

**SAMPLE 1 EXTRACTION B**

amazon.com



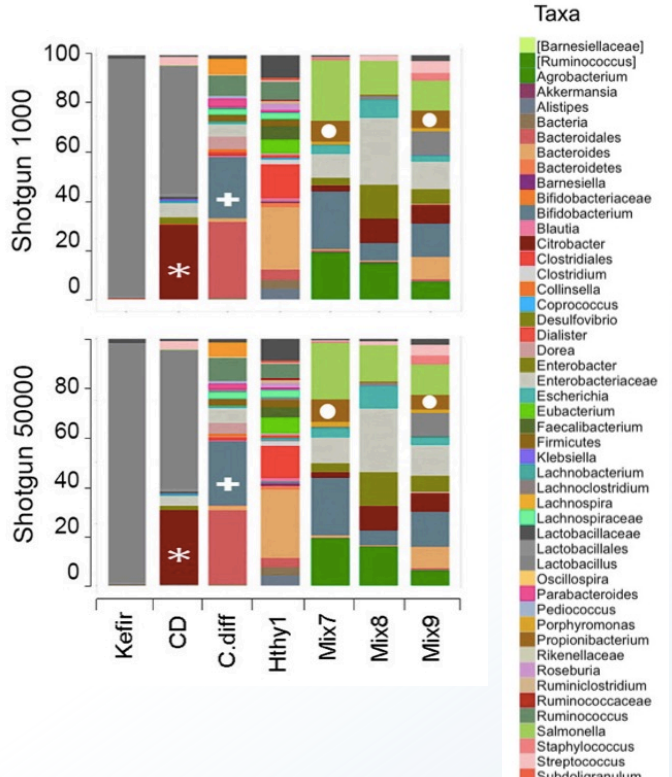

Costea et al, Nature Biotechnology 35, 1069–1076 (2017)

# Sequencing depth influence results

Increasing sampling depth = increased detection of taxa

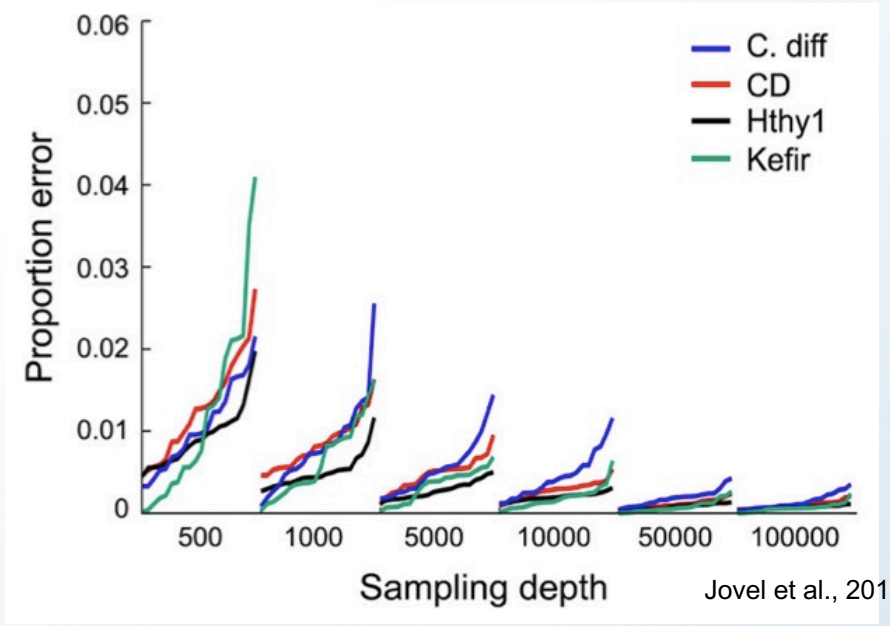
Taxonomic classification for the same library at different sequencing depths is surprisingly consistent (Jovel et al., 2016)

The proportion error and its variance decrease with increasing sampling depth



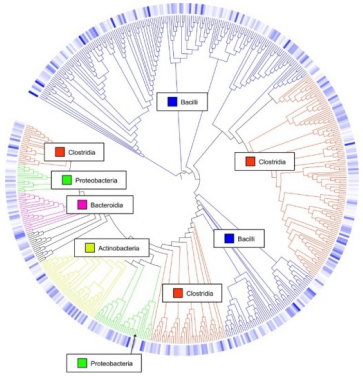


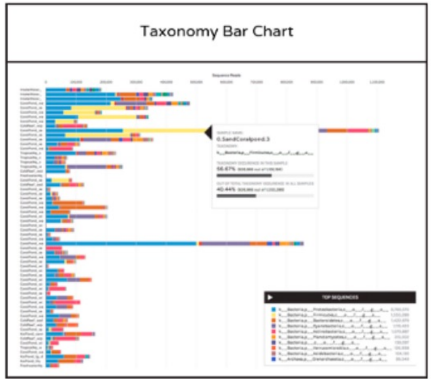


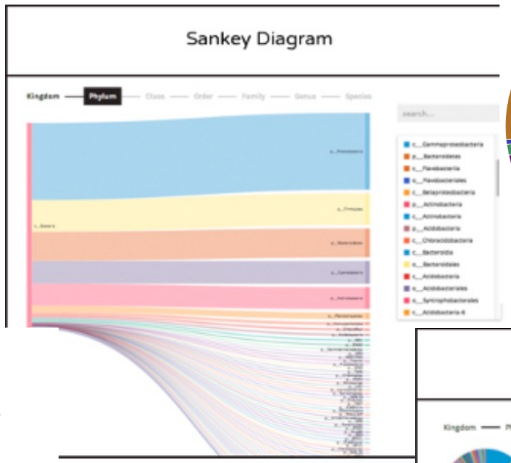


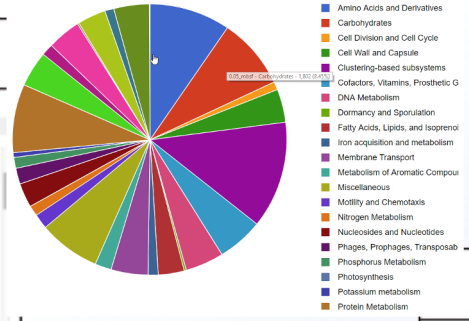

Jovel et al., 2016

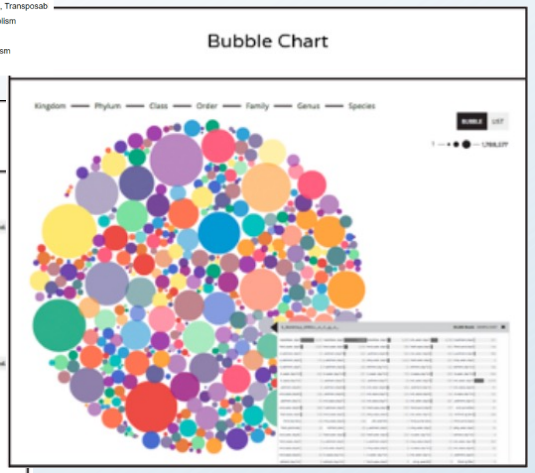
# Visualization – it is a jungle out there….

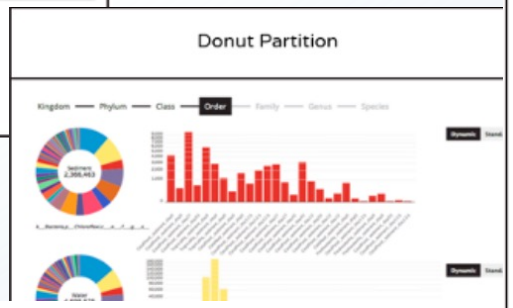


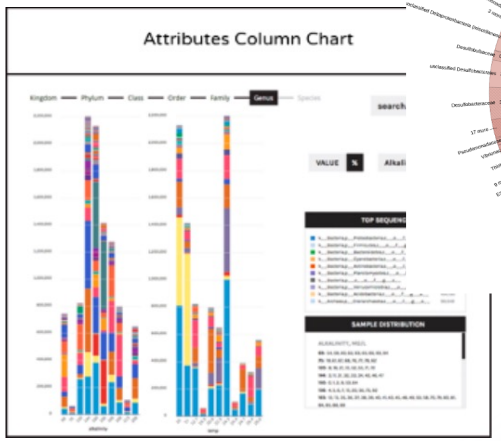



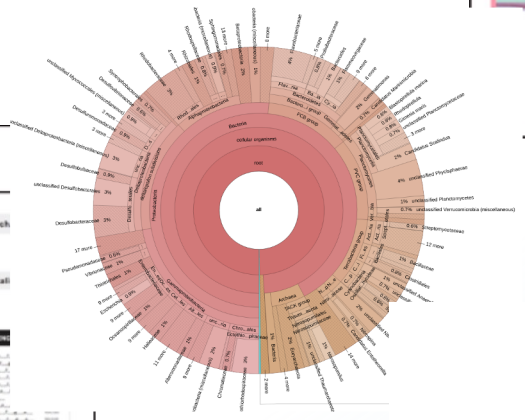



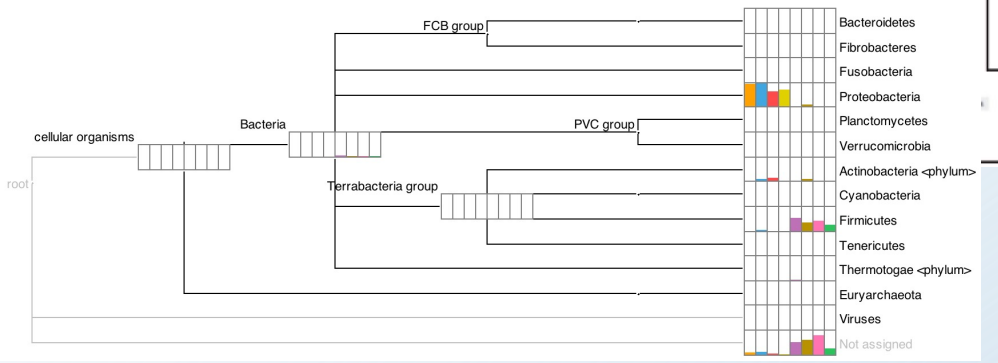



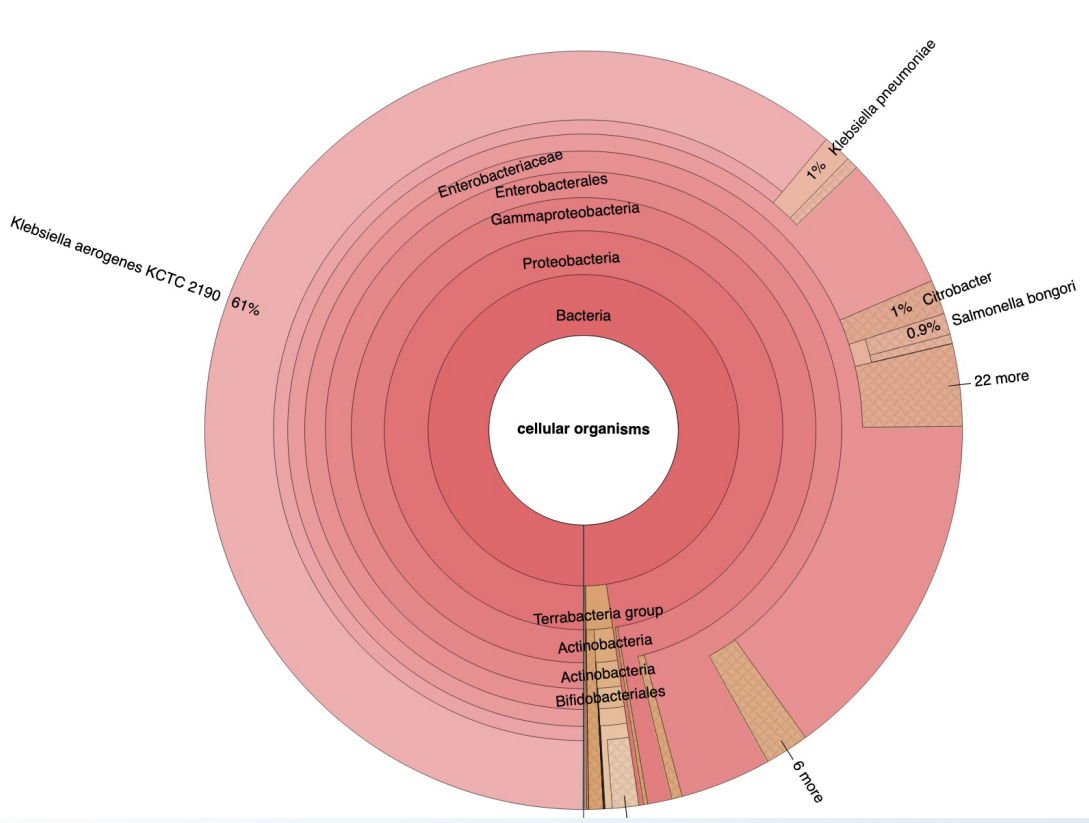





# Krona charts - Quick and easy way to visualize a taxonomic profile

Krona allows hierarchical data to be explored with zooming, multi-layered pie charts

The interactive charts are self-contained and can be viewed with a web browser

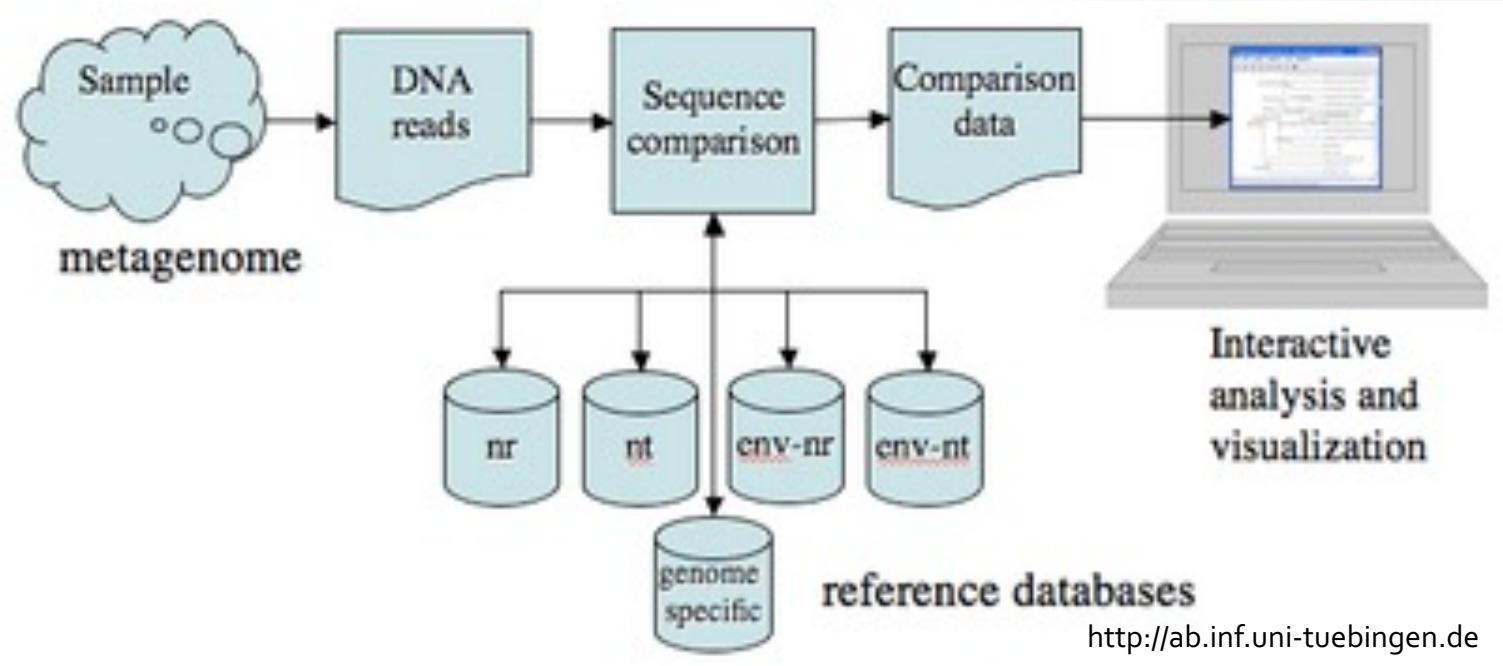


# MEGAN is a comprehensive toolbox for analysing microbiome data

MEGAN can perform both taxonomic and functional analysis

Reads are compared against a database (eg. BLAST)

The sequence comparison are imported into MEGAN where the taxonomy is automatically classified, quantified and can be visualized



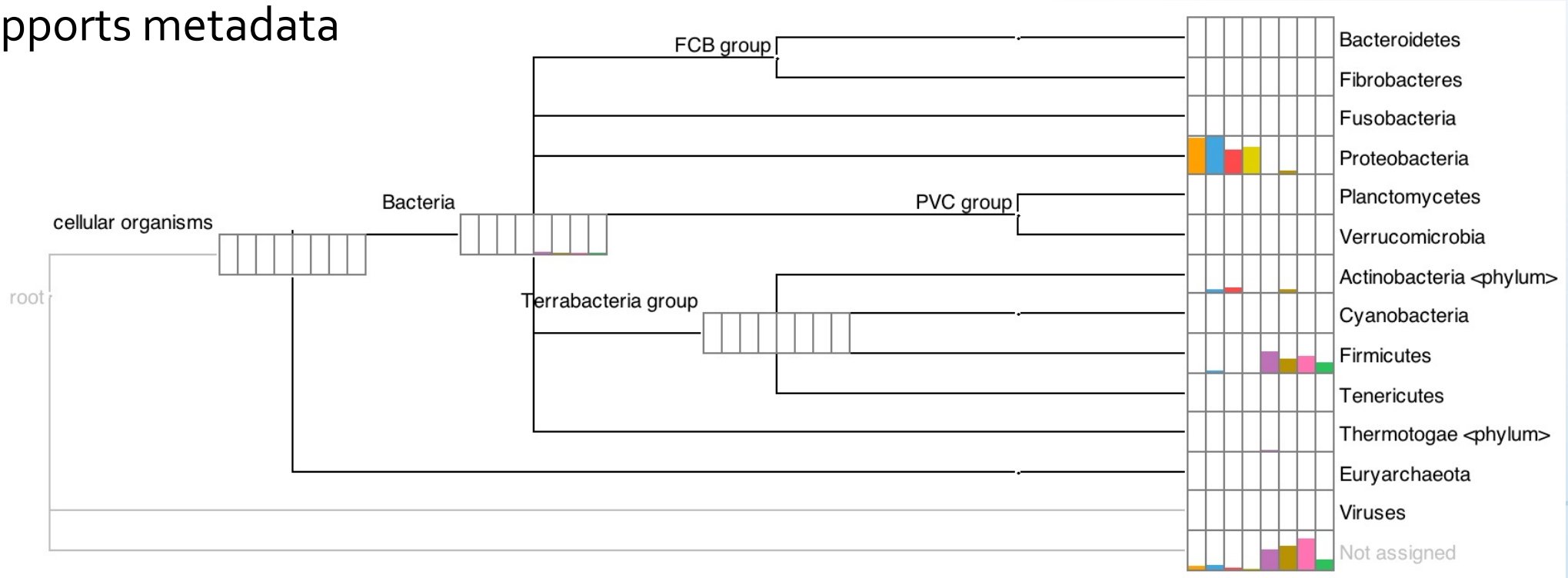
# MEGAN is a comprehensive toolbox for analysing microbiome data

Taxonomic analysis using the NCBI taxonomy or SILVA

Bar charts, word clouds, Voronoi tree maps and many other charts

PCoA, clustering and networks

Supports metadata



# Number of species on earth

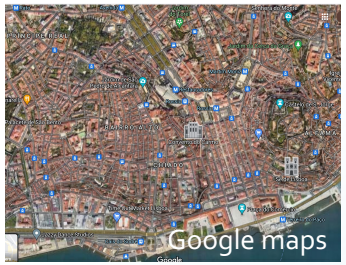
We know very few…

Earth contains 10^11 to 10^12 species of microbes (some estimate 10^19)

The total number of described bacterial species is very low 10^4

NCBI list of taxonomically approved names contain 17.989 bacterial species


**= 510 100 000 km²**

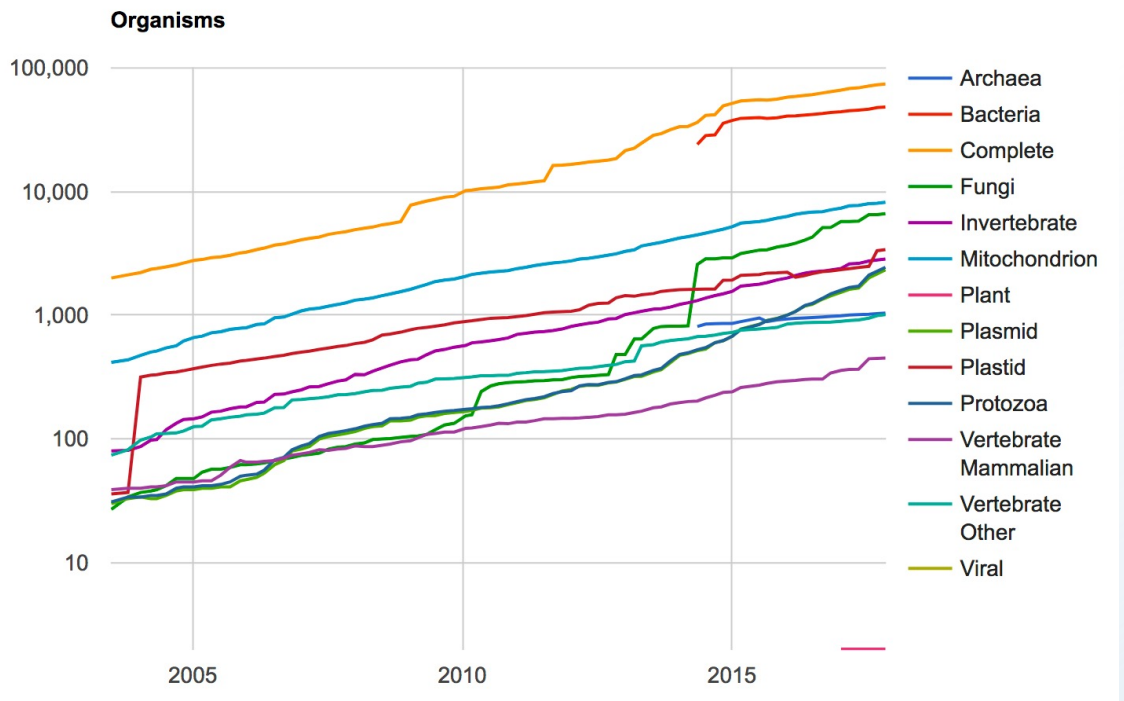


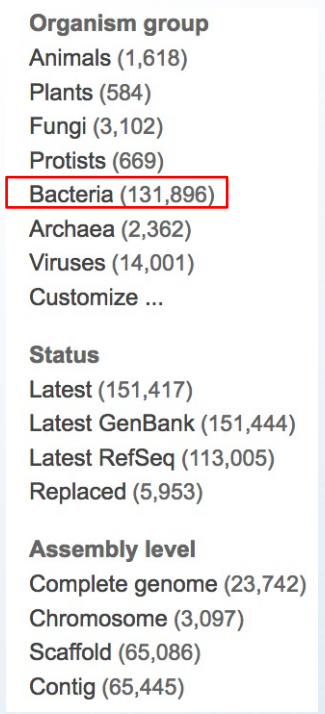

**Lisboa city centre = 5,1 km²**

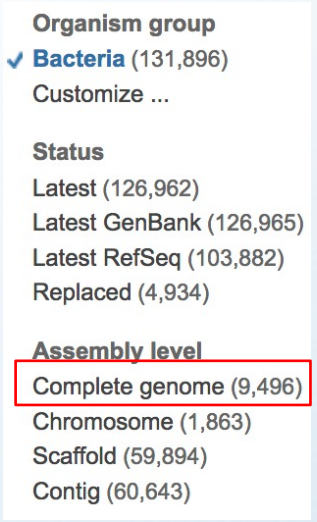
# You only find what is in the database...

## What is in the databases - for example RefSeq?

The Reference Sequence (RefSeq) collection is a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA



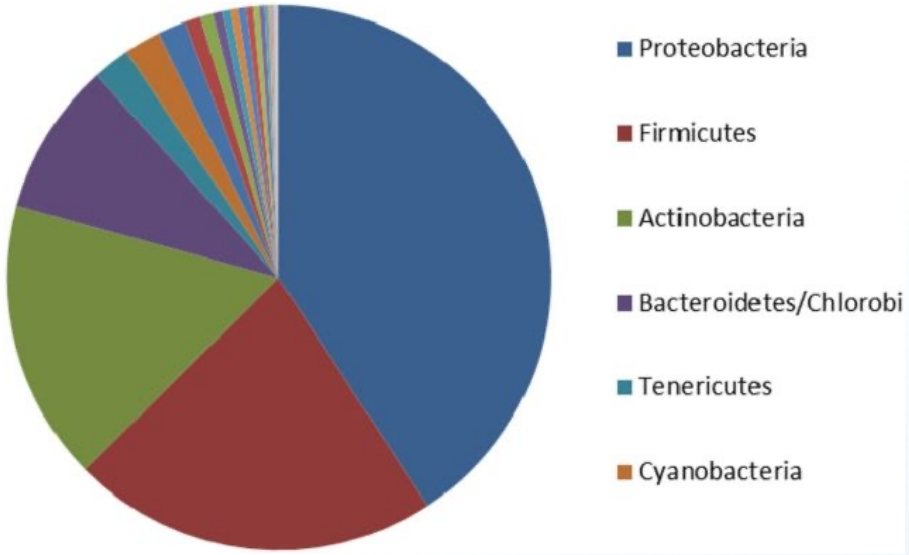



# You only find what is in the database...

What is in the databases - for example RefSeq?

Large fraction of Proteobacteria

Host-associated are overrepresented




RefSeq

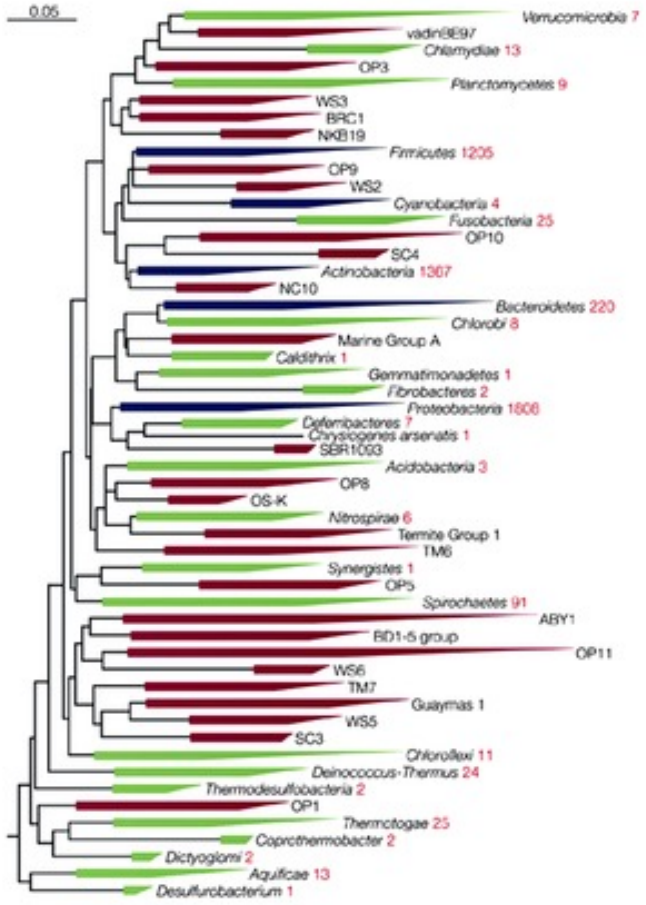
| Ecosystem | Total |
|---|---|
| **Host-associated** | **11,816** |
| Humans | 4973 |
| Animal | 1804 |
| Plants | 1410 |
| Mammals | 867 |
| Other | 2762 |
| **Environmental** | 6774 |
| Aquatic | 4559 |
| Terrestrial | 2057 |
| Other | 158 |
| **Engineered systems** | **1658** |
| Food production | 440 |
| Wastewater | 410 |
| Lab synthesis | 387 |
| Other | 418 |
| **Total** | **20,248** |

GOLD database

# You only find what is in the database...

92 named bacterial phyla – but constantly changing

The total number has been estimated to exceed 1,000 bacterial phyla




nature microbiology

## A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

*Nature Microbiology* **1**, Article number: 16048 (2016)
doi:10.1038/nmicrobiol.2016.48

Received: 25 January 2016
Accepted: 10 March 2016
Published online: 11 April 2016

Martin Keller & Karsten Zengler
Nature Reviews Microbiology volume 2, pages 141–150 (2004)

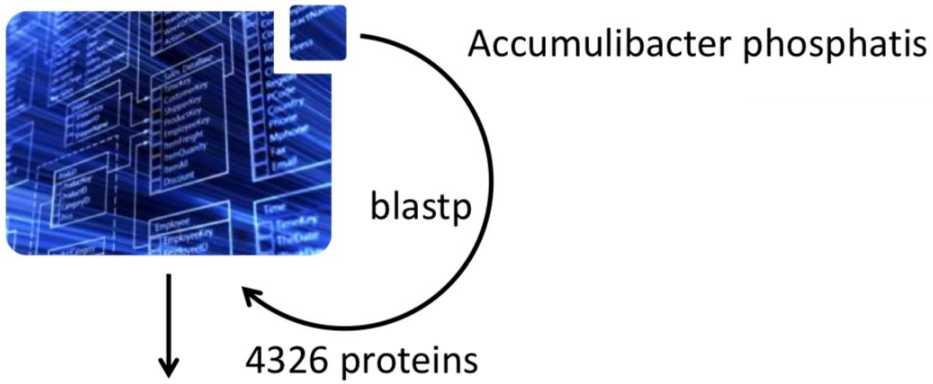
# Effect of missing genome

What is the effect of not having closely related genomes in the database?

1. Remove a genome from the database

2. Search the removed genome against the database

# Effect of missing genome



# Effect of missing genome

blastp

4326 proteins

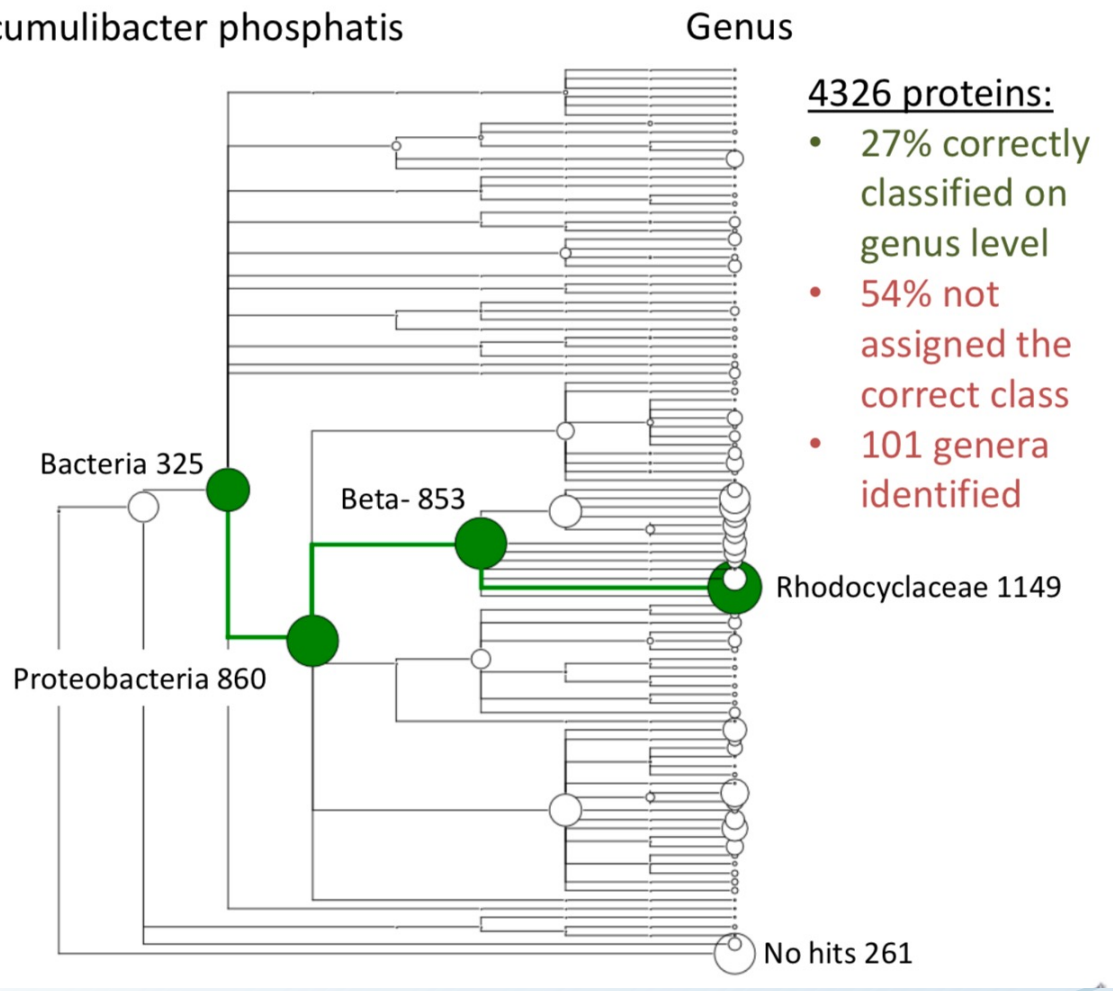
MEGAN LCA

Lowest common ancester (LCA) approach:
Hit 1: Beta-proteobacteria 80% ID
Hit 2: Gamma-proteobacteria 79% ID
Hit 3: Actinobacteria 59% ID

Assigned to Proteobacteria

## Related genomes

| | |
|---|---|
| Bacteria | 1268 |
| Proteobacteria | 564 |
| Betaproteobacteria | 84 |
| Rhodocyclales | 5 |
| Rhodocyclaceae | 5 |




4326 proteins:

- 27% correctly classified on genus level
- 54% not assigned the correct class
- 101 genera identified