2020年　AJACSオンライン3

11月11日

メタゲノム解析
（webアプリで解析）

森　宙史, Ph.D.
（Hiroshi Mori）
国立遺伝学研究所
情報研究系
hmori@nig.ac.jp

- メタゲノム解析概論
- マイクロバイオーム解析の情報解析について

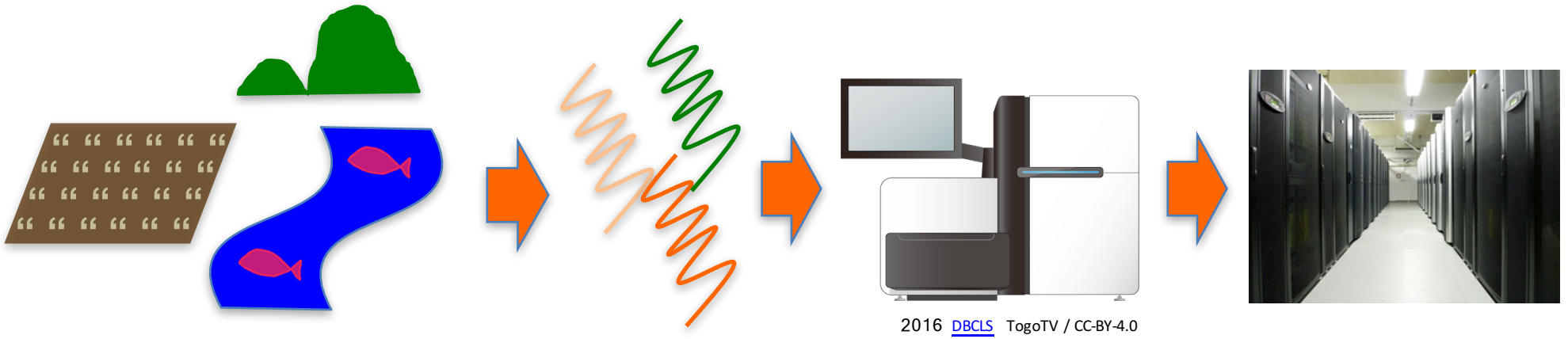
微生物＆微生物群集（細菌群集）

- 小さい
- 多様
- 様々な機能を担う（代謝、宿主の免疫の獲得等）

数%ぐらいの菌しか
培養できない

Metagenomics (since 1998)

Genome analysis against “Microbial community” to know member compositions and functions



Microbiota sampling

DNA extraction from microbes

DNA sequencing

Bioinformatics analysis

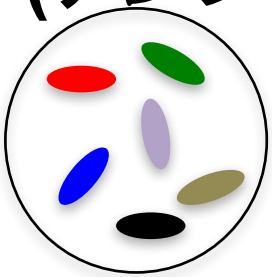
What's metagenomics?

- **Microflora, microbiota, microbial community**: 微生物群集
  Total collection of microorganisms within a community
  Microflora: [Greaves. 1926, 少なくとも]

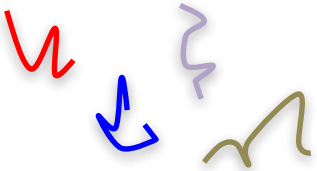
- **Metagenome**: ある群集の遺伝情報の総体
  Total genomic potential of a community
  [Handelsman et al. 1998, Chem. & Biol.]

- **Microbiome**: マイクロバイオーム
  Micro+biome or Microbio + ome?
  Microbiota and metagenome in a microbial community

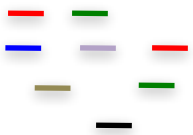
amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)



DNA extraction



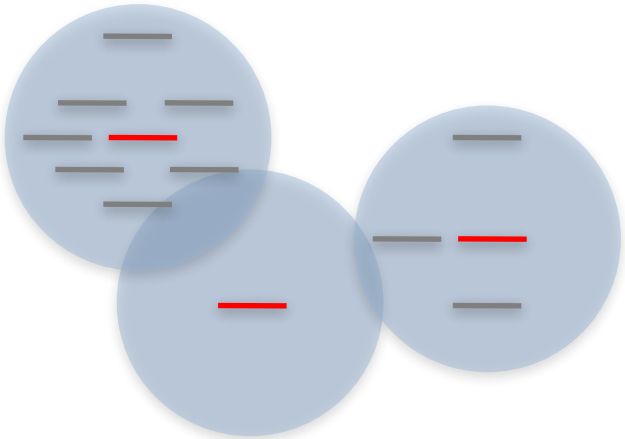
PCR amplification



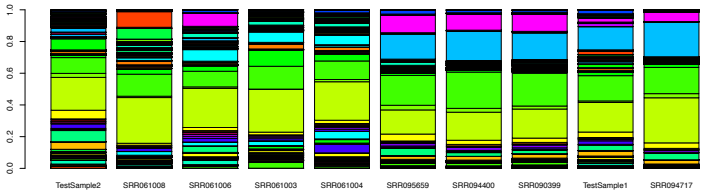
DNA Sequencing

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples

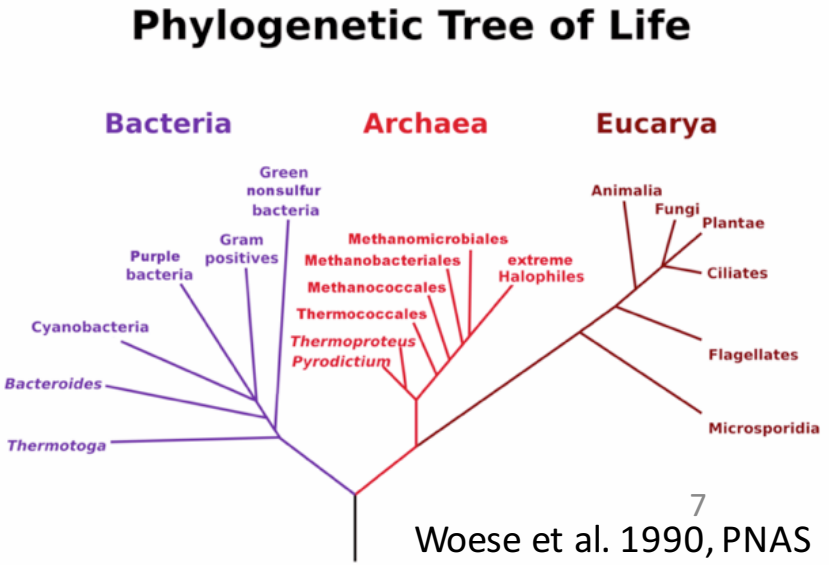


Who's there?

16S ribosomal RNA (16S rRNA)

- リボソームの核となるRNAの一つ
- 全ての細菌が所持
- 配列間の結合によって高次構造を形成 (保存されているサイトと多様なサイトがモザイク状に存在)
- 系統マーカー遺伝子の代表例
- 100万本以上の配列がデータベースに登録済み
- 多くの細菌がゲノム内に複数の遺伝子コピーを所持
- 全長約1500 base

16S rRNA遺伝子は広範囲の細菌における系統推定を行う上で適した遺伝子



Woese et al. 1990, PNAS

Metagenomic sequencing analysis
(メタゲノム解析，ショットガンメタゲノム解析)
(メタトランスクリプトーム解析)

DNA extraction / RNA extraction

Reverse transcription

DNA Sequencing

2016 DBCLS TogoTV / CC-BY-4.0

Metagenomic reads

Assemble — MEGAHIT, metaSPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

Sequence similarity search — BLASTP, DIAMOND, MMseqs2, etc.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

UniProt

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Sample1 Metadata

Sample2 Metadata

Comparative metagenomics

Who's there?

What are they doing?

アンプリコン解析もメタゲノム解析も、DNA抽出法が極めて重要

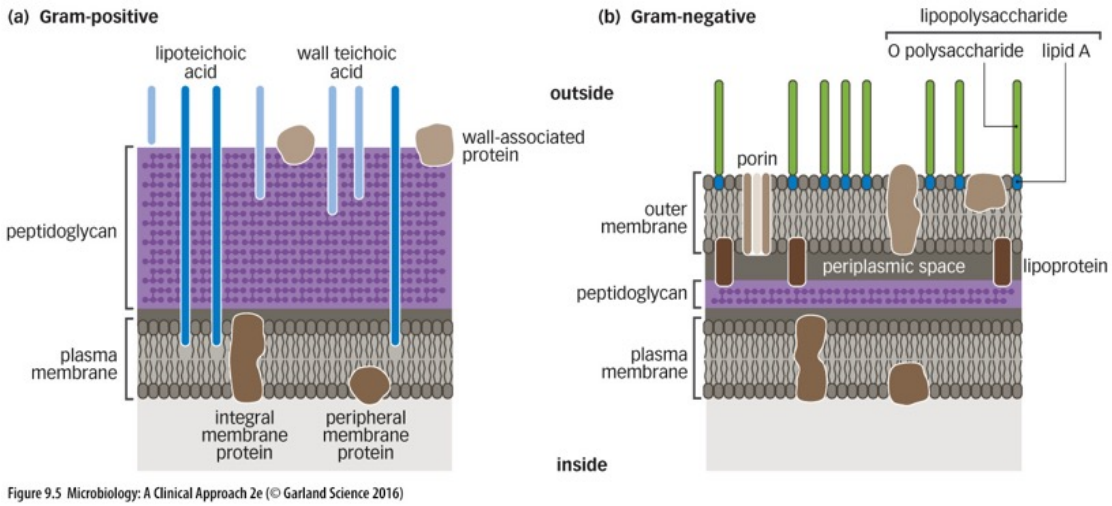


Figure 9.5 Microbiology: A Clinical Approach 2e (© Garland Science 2016)

微生物は細胞膜外の構造が多様。胞子形成の有無もDNA抽出効率に影響。

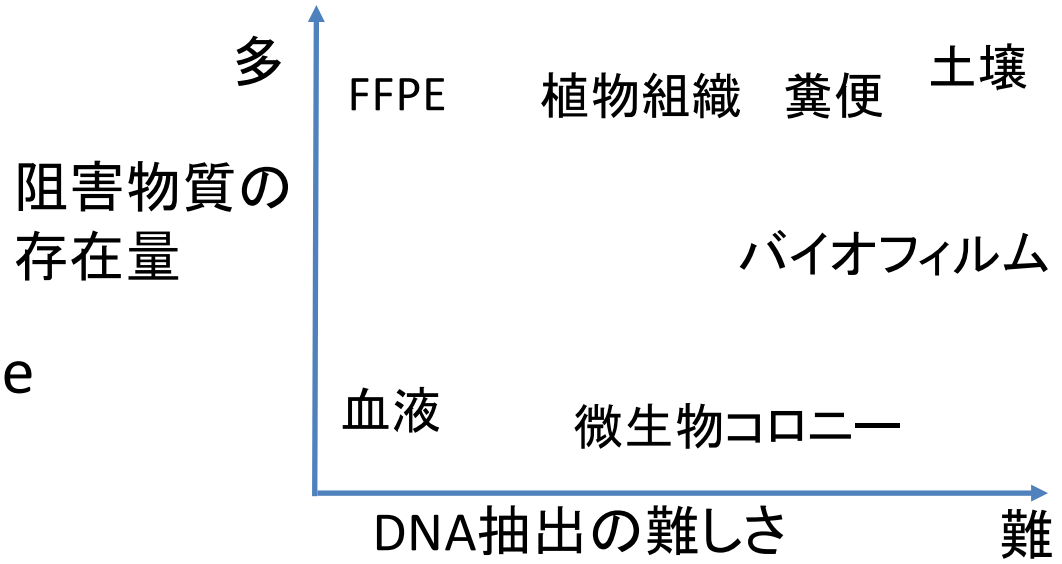
- **微生物群集解析用の代表的なDNA抽出法**
  - 酵素法（Lysozyme, Proteinase, RNase等の組み合わせ)
  - ビーズ破砕法
  - フェノールクロロホルム法
  - アルカリ法
  - 煮沸法

- **短鎖型か長鎖型のシーケンサーを使うかによっても変わる**
  - せっかく長く読めるのに、元のDNAがズタズタだと意味が無い

環境サンプルからのDNA抽出の難しさ

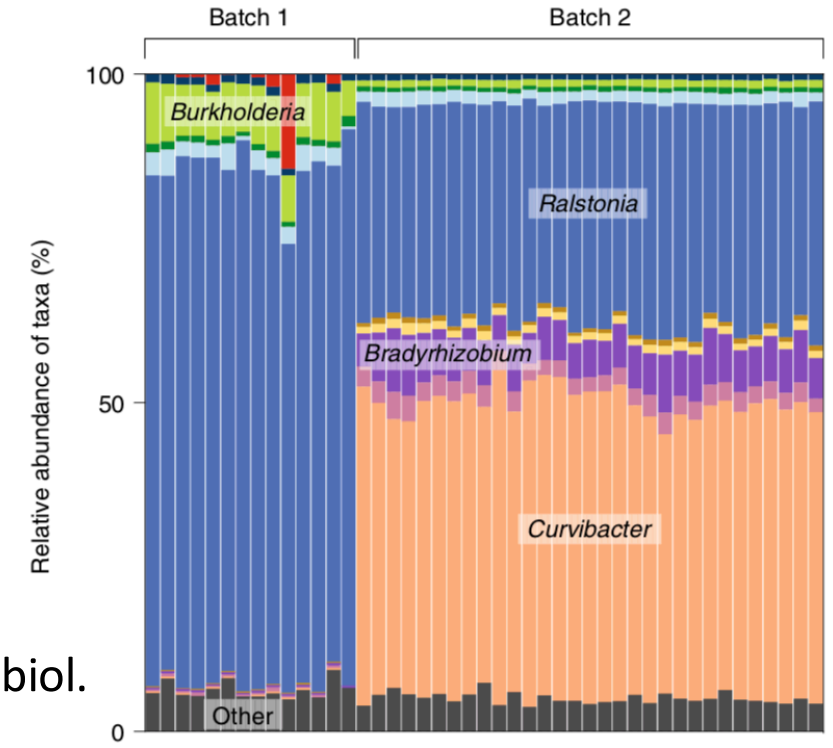
様々な夾雑物の存在

- 酵素反応阻害物質
  - フミン酸
  - DNase, RNase, proteinase
  - DNAの架橋促進物質



ターゲットのDNAが極少量の場合

- Host DNA/RNA
- 大量のRibosomal RNA
- キットや水、試薬中のコンタミDNA



Goffau et al. 2018, Nature Microbiol.

自分で条件検討をするのは大変なので、先行研究を参考にすべき

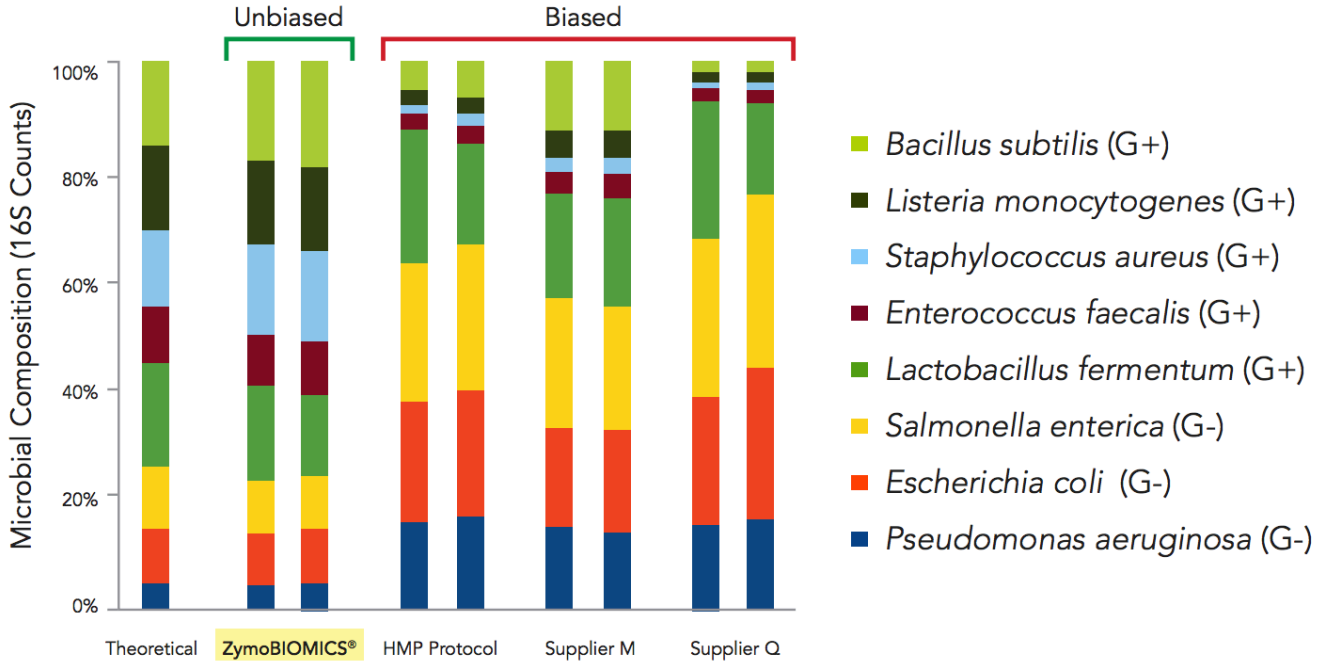
微生物の系統間でのDNA抽出バイアスの評価

Species	Avg. GC (%)	Gram Stain	gDNA Abun. (%)
Pseudomonas aeruginosa	66.2	-	12
Escherichia coli	56.8	-	12
Salmonella enterica	52.2	-	12
Lactobacillus fermentum	52.8	+	12
Enterococcus faecalis	37.5	+	12
Staphylococcus aureus	32.7	+	12
Listeria monocytogenes	38.0	+	12
Bacillus subtilis	43.8	+	12
Saccharomyces cerevisiae	38.4	Yeast	2
Cryptococcus neoformans	48.2	Yeast	2

あらかじめ細胞の混合量が
決まったmock community
をDNA抽出法間の評価に用いる

例:
HMP-ATCC mock
ZymoBIOMICS mock

https://www.zymoresearch.com/collections/zymobiomics-microbial-community-standards/products/zymobiomics-microbial-community-standard



アンプリコン解析

利点

・安価かつ少量のDNAから系統組成が得られる
・reference配列に依存しない解析も可能
・マシンパワーは少なくて済み、解析ツールも普及（QIIME2・DADA2等）

欠点

・PCRバイアスの存在
・種以下は分解能に問題あり
・個々の系統が持つ機能が不明

メタゲノム解析

利点

・系統組成と遺伝子機能組成が得られる
・実験によるバイアスが少ない
・優占系統のドラフトゲノムの構築（条件が良ければ可能）

欠点

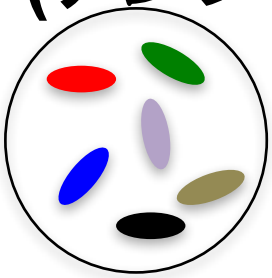
・reference配列に依存した解析
・目的依存で解析手法が変化し、マシンパワーも必要

マイクロバイオーム解析の情報解析について（アンプリコン解析）

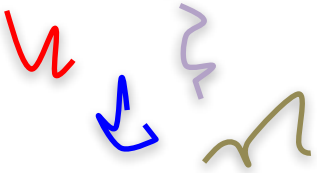
アンプリコン解析もメタゲノム解析も、Webブラウザ上やソフトウェア上でのマウスクリックのみでは解析困難

UNIXのスキルや自分のマシンでの解析スキル等が、意義ある結果を得るためにはほぼ必須

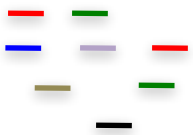
amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)



DNA extraction



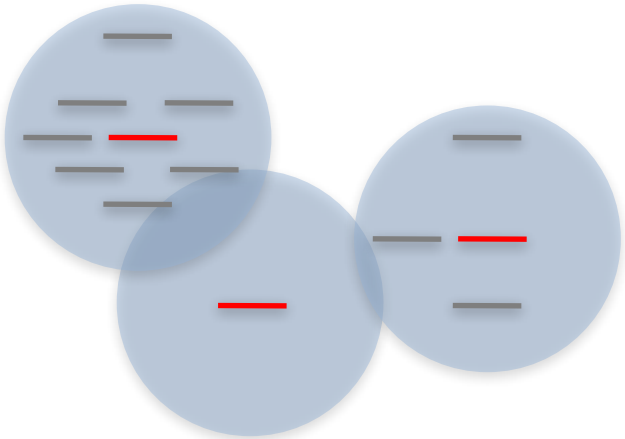
PCR amplification



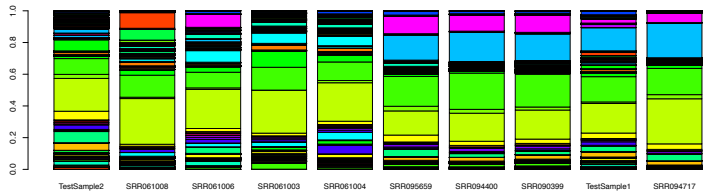
DNA Sequencing

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples



Who's there?

Bioinformatics methodology difference of amplicon sequencing before and after NGS

Before NGS: (100-1,000 reads / sample)

multiple alignment & phylogenetic tree based analysis (e.g., UniFrac, DOTUR, ARB, Bellerophon)

Accurate but slow

After NGS: (10,000-100,000 reads / sample)

sequence clustering & sequence similarity search (e.g., CD-HIT, USEARCH/UCLUST/UPARSE, Fast UniFrac, PyNAST, RDP classifier)

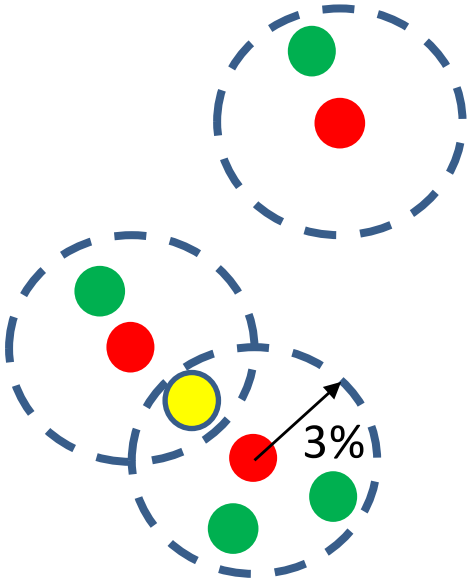
Fast but accuracy?

いわゆるOperational taxonomic unit (OTU)クラスタリング

OTU clustering problem

例: 16S rRNA遺伝子について、
配列類似性97%（種レベル）で配列をクラスタリング

- **OTUの代表配列**
- **OTUのメンバーの配列**
- **所属が曖昧な配列**



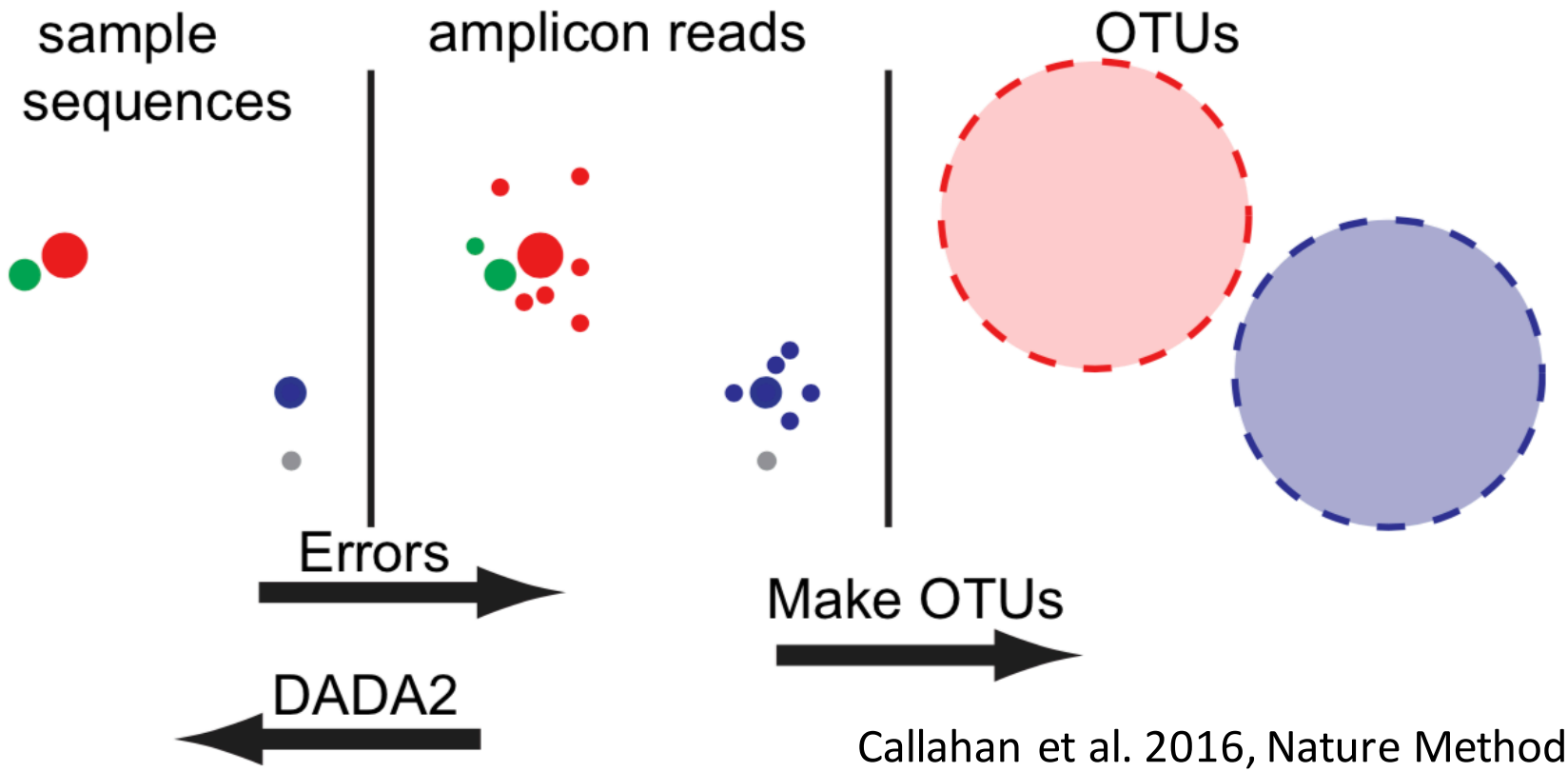
一律の配列類似性でクラスタリングする限り、所属が曖昧な配列が出ることは避けられない。

Denoising & Dereplication

e.g., DADA2 (R), Deblur

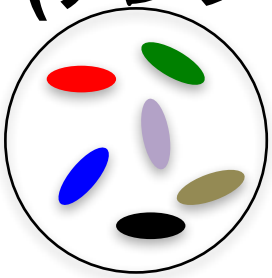
1. remove sequencing error in reads by model-based approach
2. dereplicate complete match reads to one representative read
3. conduct taxonomic assignment

Denoising & Derep. approachの結果得られた完全マッチクラスタを Amplicon sequence variant (ASV), Exact sequence variant (ESV), Zero-radius OTU (ZOTU)等と呼んでOTUと区別する

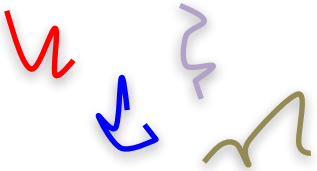


Callahan et al. 2016, Nature Methods.

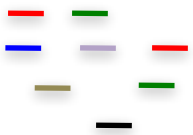
amplicon sequencing analysis
(アンプリコン解析, 16S rRNA遺伝子のアンプリコン解析)



DNA extraction



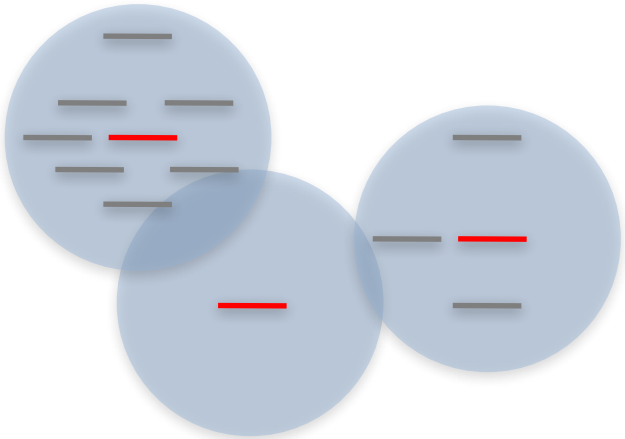
PCR amplification



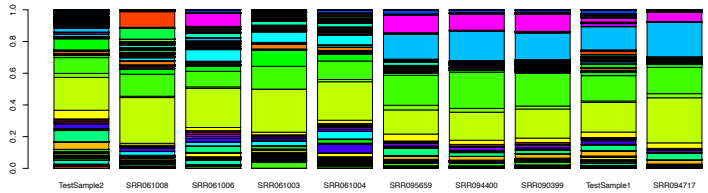
DNA Sequencing

Pre-analysis (Remove Primer, Chimera etc.)

Sequence clustering or denoising



Taxonomic assignment and
Comparison between samples

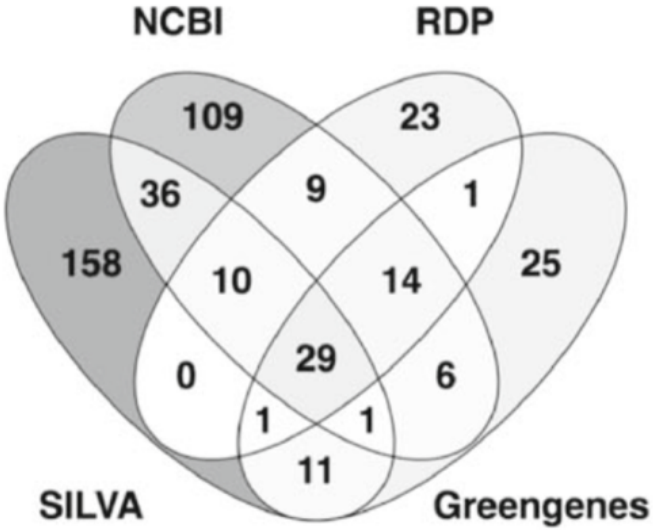


Who's there?

16S rRNA gene reference DB

- **SILVA**
  (https://www.arb-silva.de/)
  Taxonomy: List of Prokaryotic Names with Standing in Nomenclature (LPSN)
  License: CC BY
  Last update: Aug. 2020

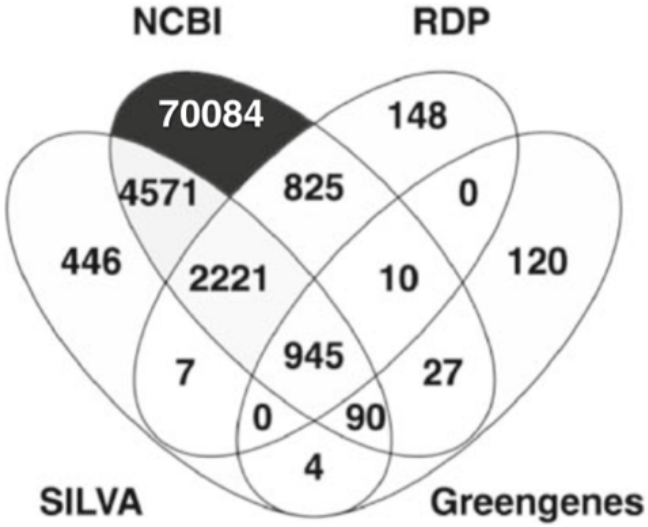
Phylum



- **Greengenes**
  (http://greengenes.secondgenome.com/)
  Taxonomy: NCBI Taxonomy (modified)
  License: CC BY SA
  Last update: Oct. 2013

- **RDP**
  (https://rdp.cme.msu.edu/)
  Taxonomy: Bergey's Manual
  License: CC BY SA
  Last update: Sep. 2016

Genus



Balvociute & Huson, 2017, BMC Genomics

サンプルごとの系統組成 or ASV組成が得られた後の解析

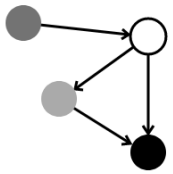
- 組成のグラフ化
- 主成分分析
- サンプル間の組成の距離計算
- 多次元尺度構成法や階層的クラスタリング
- 群間の統計的仮説検定
- 注目する系統についてのより詳細な解析および検索

得られるデータはあくまで相対量であり、
絶対量（細胞数）のデータでは無い点に注意が必要

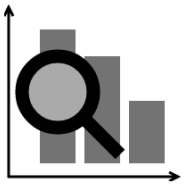
QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

https://qiime2.org/

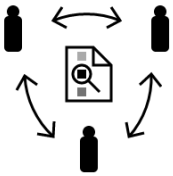
Learn more » Citing QIIME 2 »



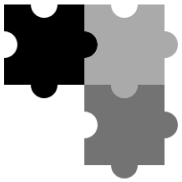
Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!



Interactively explore your data with beautiful visualizations that provide new perspectives.



Easily share results with your team, even those members without QIIME 2 installed.



Plugin-based system — your favorite microbiome methods all in one place.

Choose the interface that fits your needs

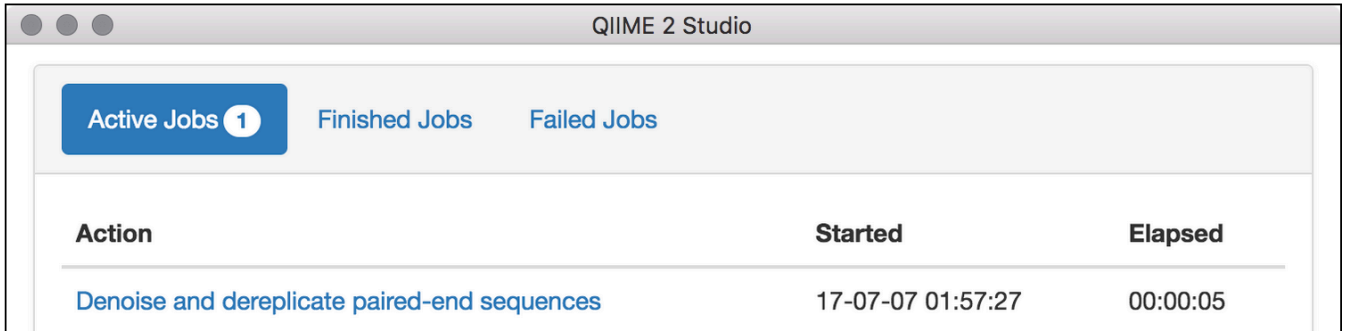
q2cli the command line interface

```
2. ~ (zsh)
$ qiime info
System versions
Python version: 3.5.3
QIIME 2 release: 2017.6
QIIME 2 version: 2017.6.0
q2cli version: 2017.6.0

Installed plugins
alignment 2017.6.0
composition 2017.6.0
dada2 2017.6.0
```

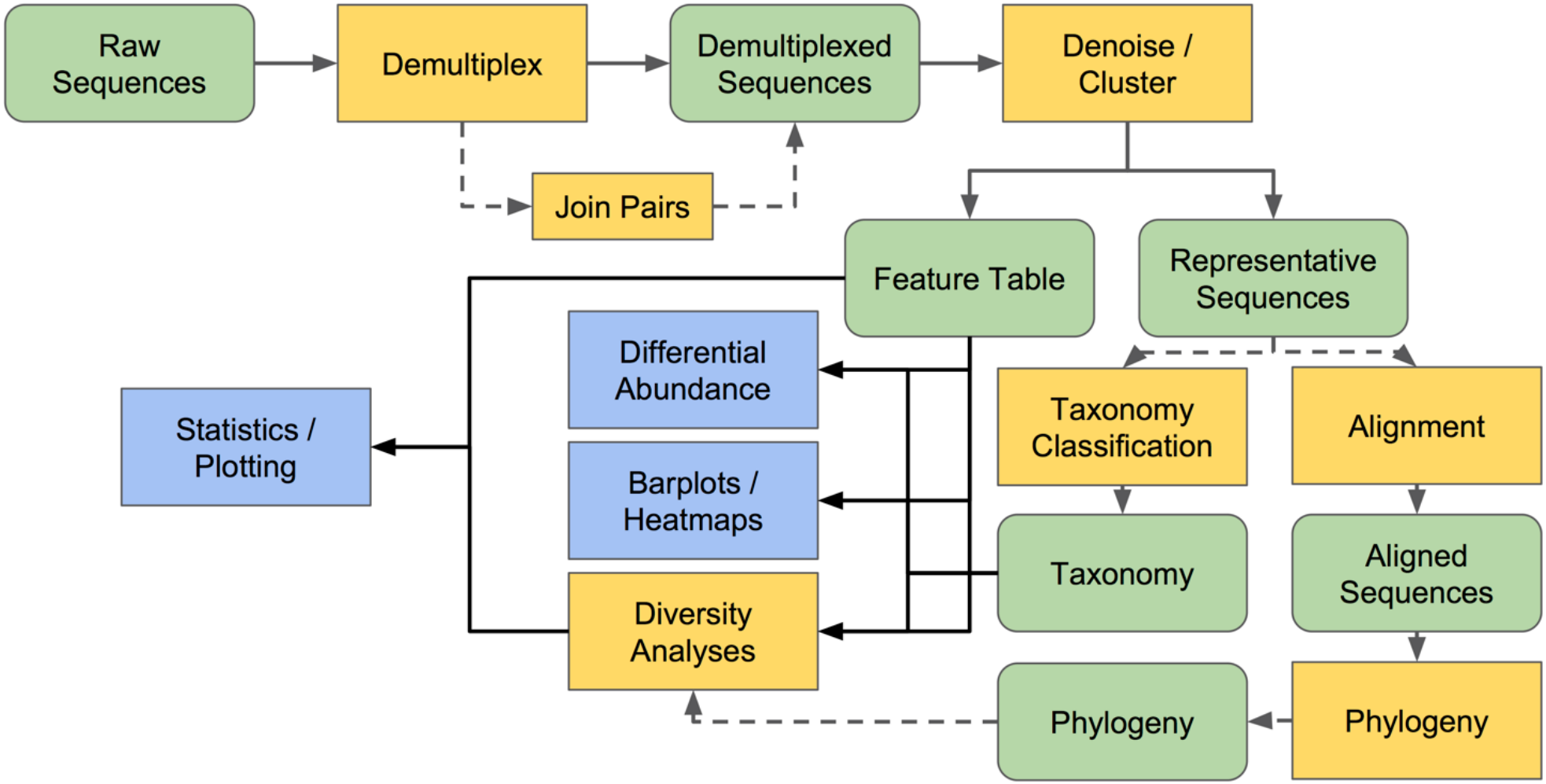
q2studio the graphical user interface (PROTOTYPE)

q2studio is a functional prototype of a graphical user interface for QIIME 2, and is not necessarily feature-complete with respect to q2cli and the Artifact API.



Overview of an example of QIIME2 workflow

https://docs.qiime2.org/2020.8/tutorials/overview/



DADA2 Pipeline Tutorial (1.16)

Here we walk through version 1.16 of the DADA2 pipeline on a small multi-sample dataset. Our starting point is a set of Illumina-sequenced paired-end fastq files that have been split (or "demultiplexed") by sample and from which the barcodes/adapters have already been removed. The end product is an **amplicon sequence variant (ASV) table**, a higher-resolution analogue of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in each sample. We also assign taxonomy to the output sequences, and demonstrate how the data can be imported into the popular phyloseq R package for the analysis of microbiome data.

https://benjjneb.github.io/dada2/tutorial.html

Starting point

This workflow assumes that your sequencing data meets certain criteria:

- Samples have been demultiplexed, i.e. split into individual per-sample fastq files.
- Non-biological nucleotides have been removed, e.g. primers, adapters, linkers, etc.
- If paired-end sequencing data, the forward and reverse fastq files contain reads in matched order.

If these criteria are not true for your data (**are you sure there aren't any primers hanging around?**) you need to remedy those issues before beginning this workflow. See the FAQ for recommendations for some common issues.

Getting ready

First we load the `dada2` package. If you don't already have it, see the dada2 installation instructions.

```
library(dada2); packageVersion("dada2")
```

```
## [1] '1.16.0'
```

Older versions of this workflow associated with previous release versions of the dada2 R package are also available: 1.6, 1.8, 1.12, .

The data we will work with are the same as those used in the mothur MiSeq SOP. To follow along, download the example data and unzip. These fastq files were generated by 2x250 Illumina Miseq amplicon sequencing of the V4 region of the 16S rRNA gene from gut samples collected longitudinally from a mouse post-weaning. For now just consider them paired-end fastq files to be processed. Define the following path variable so that it points to the extracted directory on **your** machine:

https://www.arb-silva.de/

silva
high quality ribosomal RNA databases

elixir Core Data Resource

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

Home SILVAngs Browser Search ACT Download Documentation Projects FISH & Probes Contact

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

SILVAngs

silvangs

Check out our service for Next Generation Amplicon data

SILVA Alignment, Classification and Tree (ACT) Service

The SILVA ACT service combines alignment, search and classify as well as reconstruction of trees in a single web application.

SILVA ACT is available at: → www.arb-silva.de/act

News

02.09.2020

SILVAngs updated to SILVA r138.1

We have updated our amplicon analysis pipeline SILVAngs to SILVA release 138.1.

27.08.2020

SILVA 138.1 released

SILVA 138.1 is an update of the SSU 138 full release providing corrections to the SILVA SSU taxonomy as well as updating the LSU sequence data and taxonomy to match the SSU release.

SILVAngs will be updated to 138.1 early next week.

11.04.2020

Happy Easter and Stay Healthy

A Happy Easter from the SILVA Team. Working hard from home to get the next SILVA update release on SSU and LSU 138 done. Stay healthy to fight COVID-19 together.

21.12.2019

Merry Christmans & Happy New Year 2020

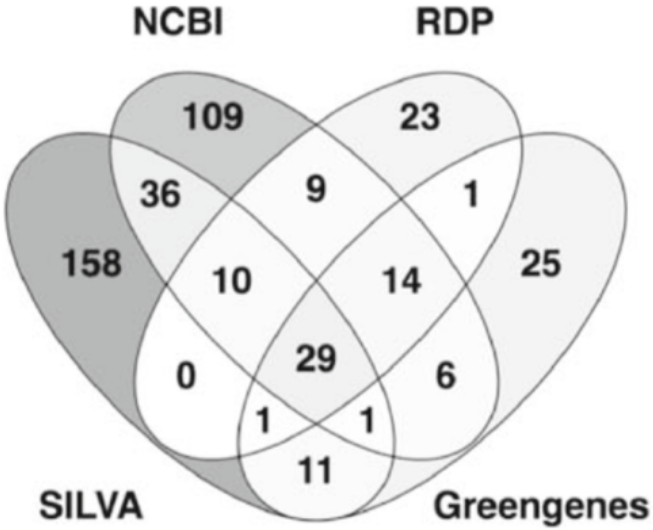
The SILVA Team wishes you a Merry Christmas & Happy New Year. Many thanks for all your feedback and support to improve SILVA and SILVAngs. Looking forward to see you again in 2020.

go to Archive ->

16S rRNA gene reference DB

- **SILVA**
  (https://www.arb-silva.de/)
  Taxonomy: List of Prokaryotic Names with Standing in Nomenclature (LPSN)
  License: CC BY
  Last update: Aug. 2020

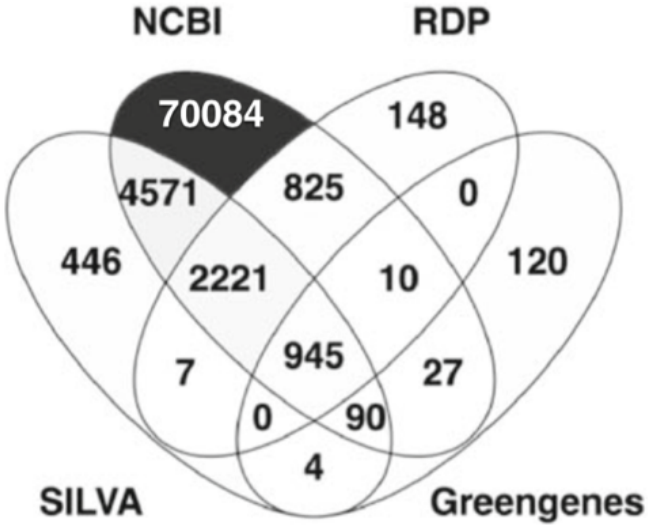
Phylum



- **Greengenes**
  (http://greengenes.secondgenome.com/)
  Taxonomy: NCBI Taxonomy (modified)
  License: CC BY SA
  Last update: Oct. 2013

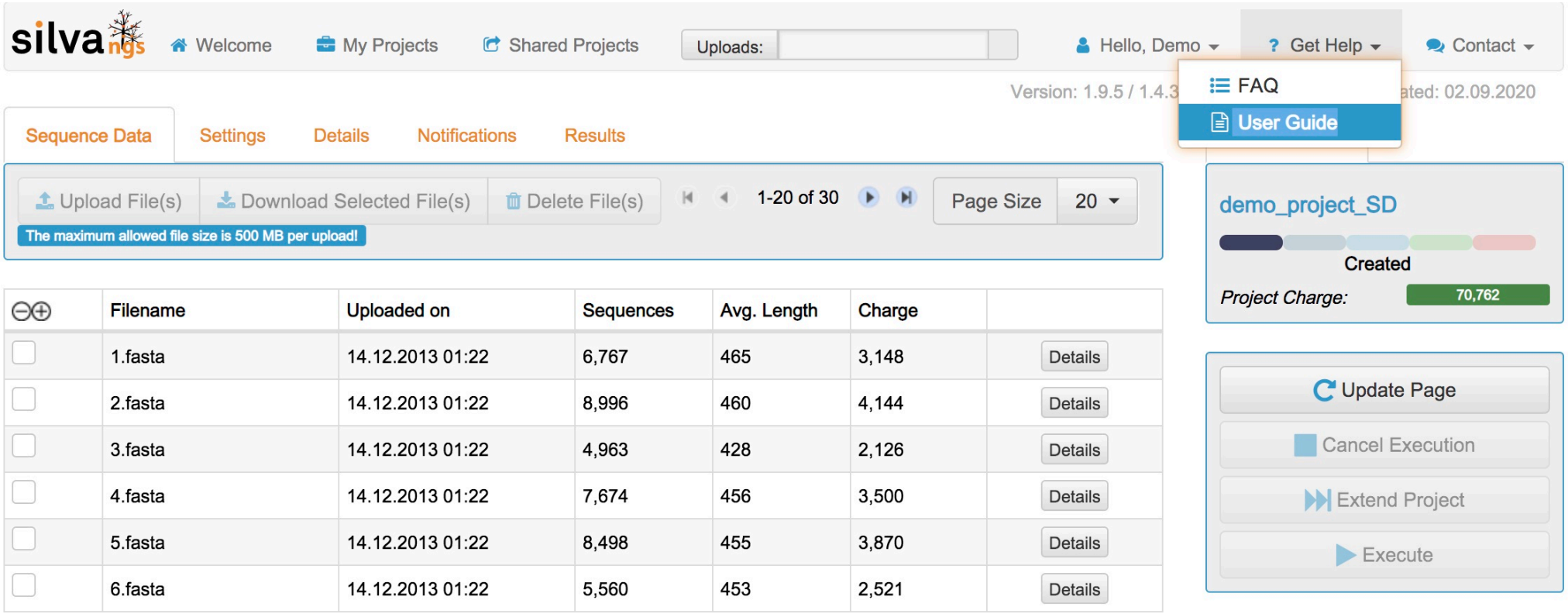
- **RDP**
  (https://rdp.cme.msu.edu/)
  Taxonomy: Bergey's Manual
  License: CC BY SA
  Last update: Sep. 2016

Genus



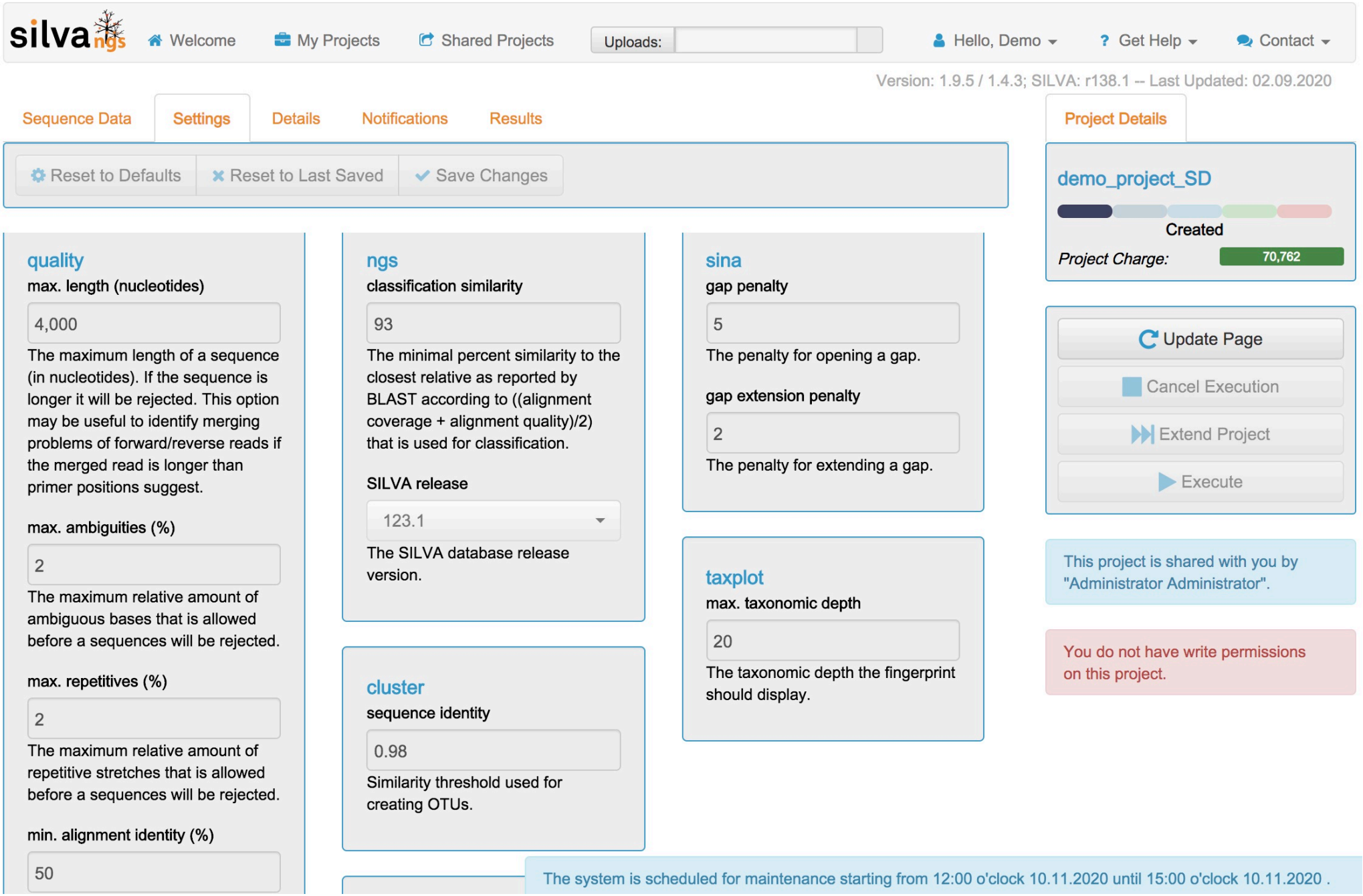
Balvociute & Huson, 2017, BMC Genomics

https://ngs.arb-silva.de/silvangs

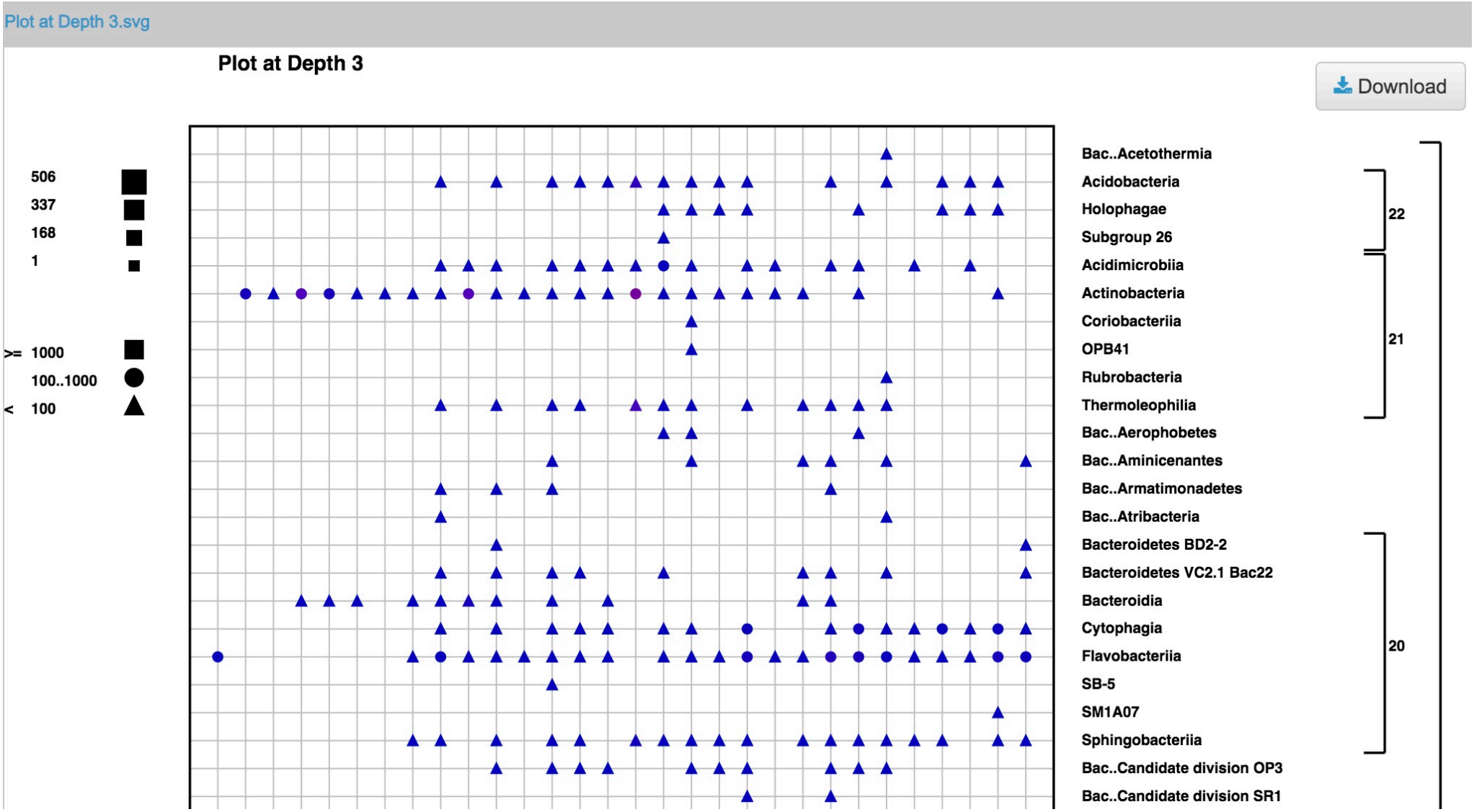


	Filename	Uploaded on	Sequences	Avg. Length	Charge	
☐	1.fasta	14.12.2013 01:22	6,767	465	3,148	Details
☐	2.fasta	14.12.2013 01:22	8,996	460	4,144	Details
☐	3.fasta	14.12.2013 01:22	4,963	428	2,126	Details
☐	4.fasta	14.12.2013 01:22	7,674	456	3,500	Details
☐	5.fasta	14.12.2013 01:22	8,498	455	3,870	Details
☐	6.fasta	14.12.2013 01:22	5,560	453	2,521	Details

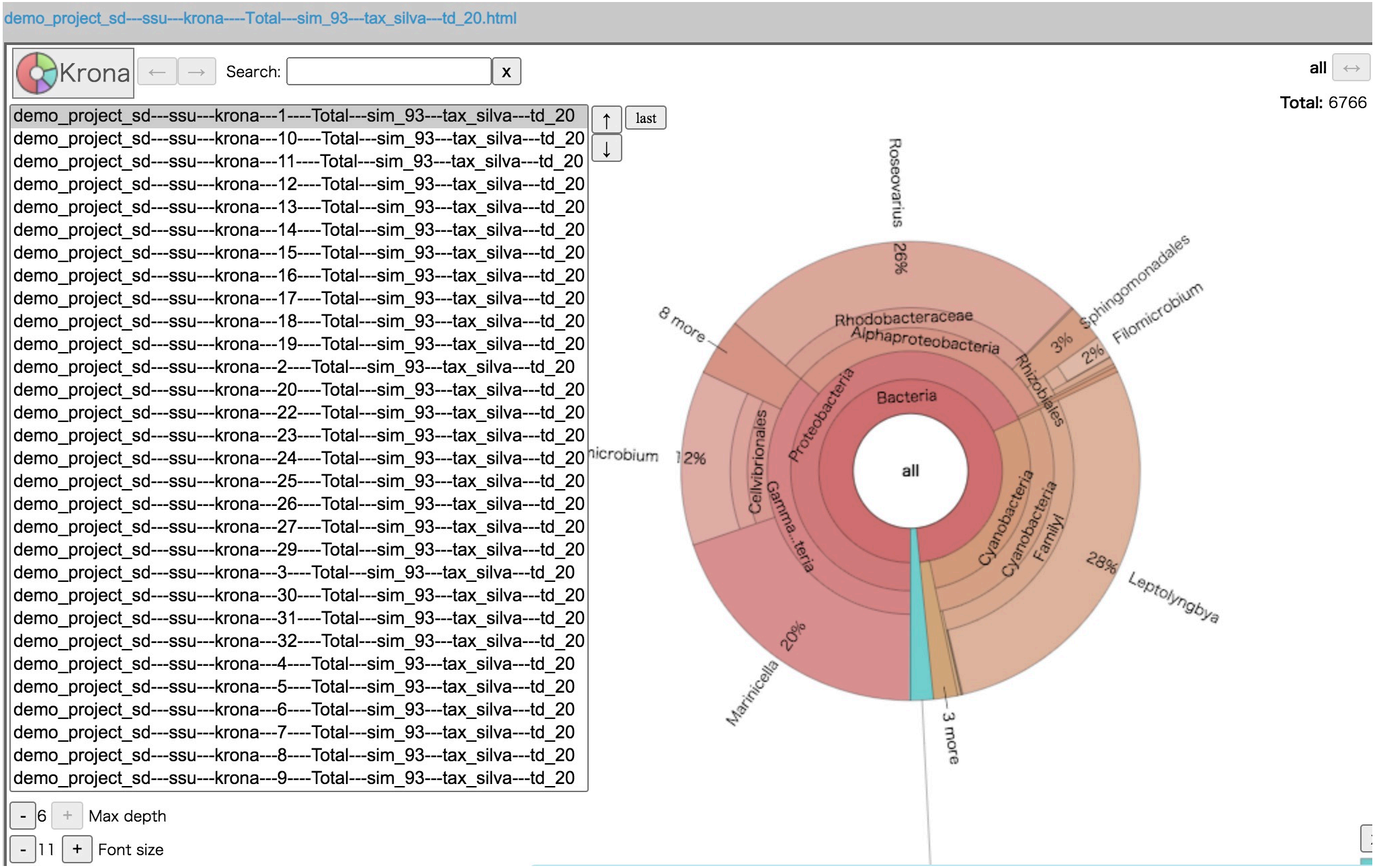
https://ngs.arb-silva.de/silvangs



https://ngs.arb-silva.de/silvangs



https://ngs.arb-silva.de/silvangs



VITCOMIC2 http://vitcomic.org

VITCOMIC2 is a visualization tool for the phylogenetic composition of microbial communities based on 16S rRNA gene amplicons and metagenomic shotgun sequencing.

Try VITCOMIC2

Metagenome/16S rRNA gene Amplicon Sequencing FASTA/FASTQ file: 選択... ファイルが選択されていません。
File format: FASTA flat FASTQ flat FASTA gzipped FASTQ gzipped
Conduct 16S rRNA gene Copy number normalization?: No Yes
Conduct 16S rRNA gene Assembly? (Shotgun metagenome only): No Yes
ID: (use [A-Za-z0-9-_])
Email:
upload&calculate clear

How to use

1. Input data

Both of a FASTA/FASTQ file and gzipped FASTA/FASTQ file are acceptable for the input data in the VITCOMIC2. Sample 16S rRNA gene Amplicon sequencing fastq data.

2. File format

File format is a file format identifier of your FASTA/FASTQ file. To reduce the size of your file, we strongly recommend that you compress your file with gzip. If you don't compress your file, please choose "flat file".

(Mori et al. 2018, BMC Systems Biol.)

VITCOMIC2の結果

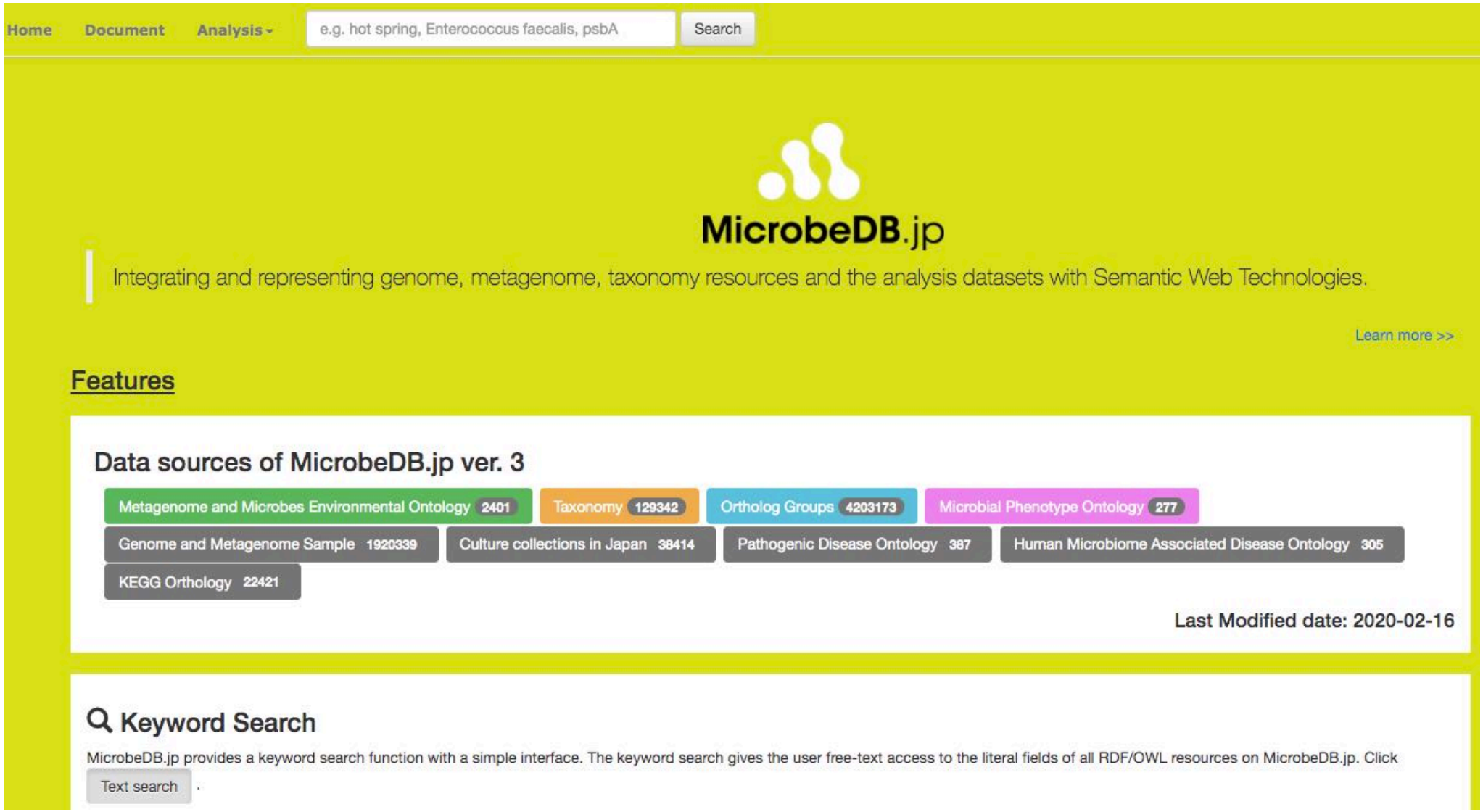
Alcaligenes	Betaproteobacteria	1
Aminobacter	Alphaproteobacteria	7
Azoarcus	Betaproteobacteria	1
Blastomonas	Alphaproteobacteria	2
Burkholderia	Betaproteobacteria	1980
Caedibacter	Gammaproteobacteria	1
Caldilinea	"Chloroflexi"	3
Comamonas	Betaproteobacteria	120
Fodinicurvata	Alphaproteobacteria	1
Gaetbulicola	Alphaproteobacteria	1
Hoeflea	Alphaproteobacteria	132
Hydrogenophaga	Betaproteobacteria	2
Hyphomicrobium	Alphaproteobacteria	2
Hyphomonas	Alphaproteobacteria	23
Insolitispirillum	Alphaproteobacteria	2
Kiloniella	Alphaproteobacteria	1
Kordiimonas	Alphaproteobacteria	5
Labrenzia	Alphaproteobacteria	1
Limnobacter	Betaproteobacteria	2
Loktanella	Alphaproteobacteria	9
Magnetospirillum	Alphaproteobacteria	3
Maribius	Alphaproteobacteria	12
Marinomonas	Gammaproteobacteria	1
Mesorhizobium	Alphaproteobacteria	1
Methylobacterium	Alphaproteobacteria	1
Methylotenera	Betaproteobacteria	2
Microvirga	Alphaproteobacteria	1
Nereida	Alphaproteobacteria	1
Nisaea	Alphaproteobacteria	167
Nitratireductor	Alphaproteobacteria	1
Nitrincola	Gammaproteobacteria	1
Novosphingobium	Alphaproteobacteria	24

サンプルの
属組成が得られる

クラスタリングしないので
メタゲノム解析にも使える

全配列のRDPデータベースへの配列類似性検索を行い、
Identity ≥ 97%のHitを集計した結果

https://microbedb.jp/

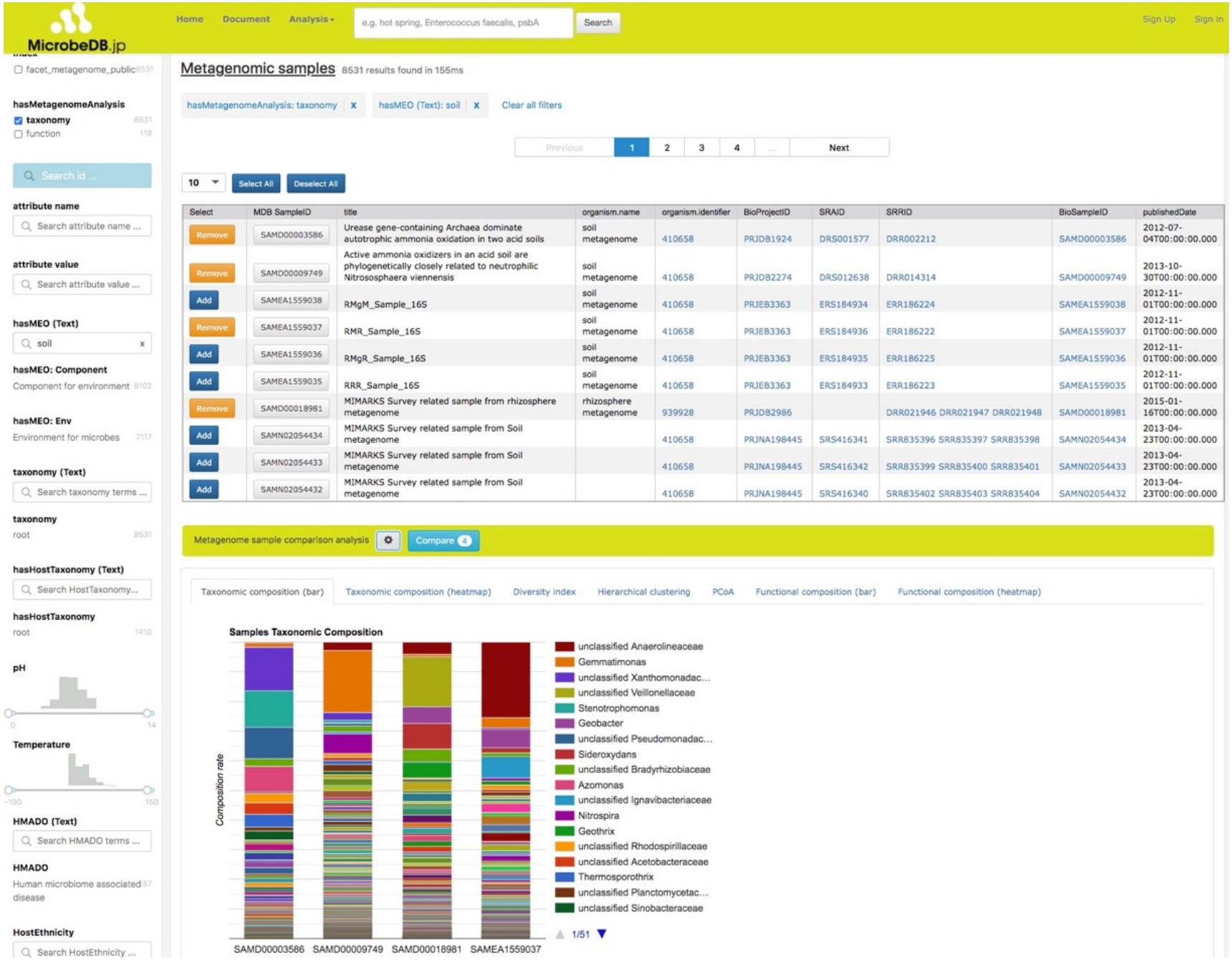


MicrobeDB.jp version 3 data

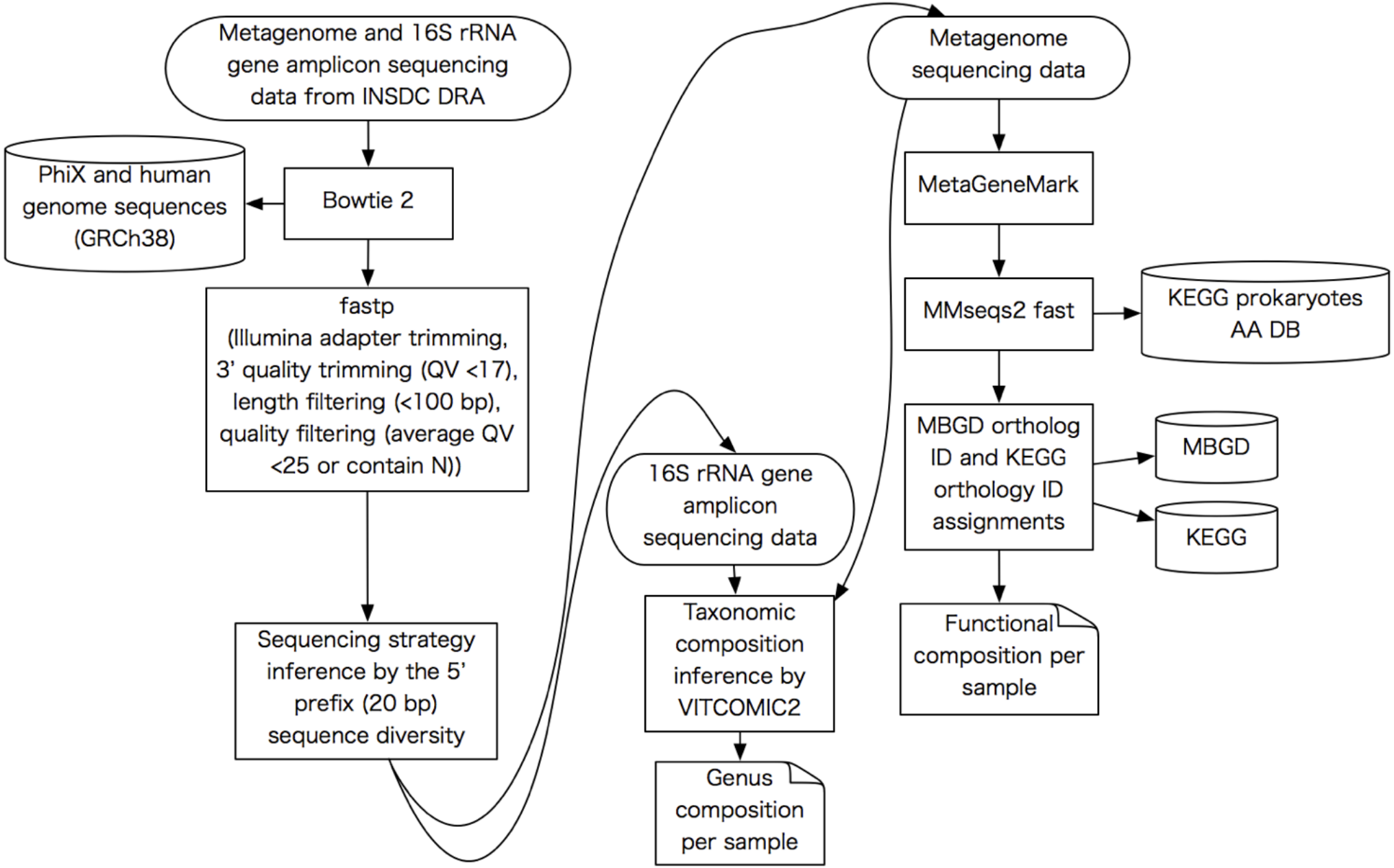
Data category	Number of entry
Prokaryote genome metadata (from RefSeq)	290,208 genomes
Culture collection strain metadata from JCM/NBRC (from RDF-Portal)	38,414 strains
Microbiome metadata (from INSDC BioSample)	1,631,611 samples
Microbiome taxonomic composition data	96,766 samples
Microbiome functional composition data	4,784 samples

Version2と比べて、ゲノムは約20倍、メタゲノムは約10倍

Search and compare microbiome samples from **soil**

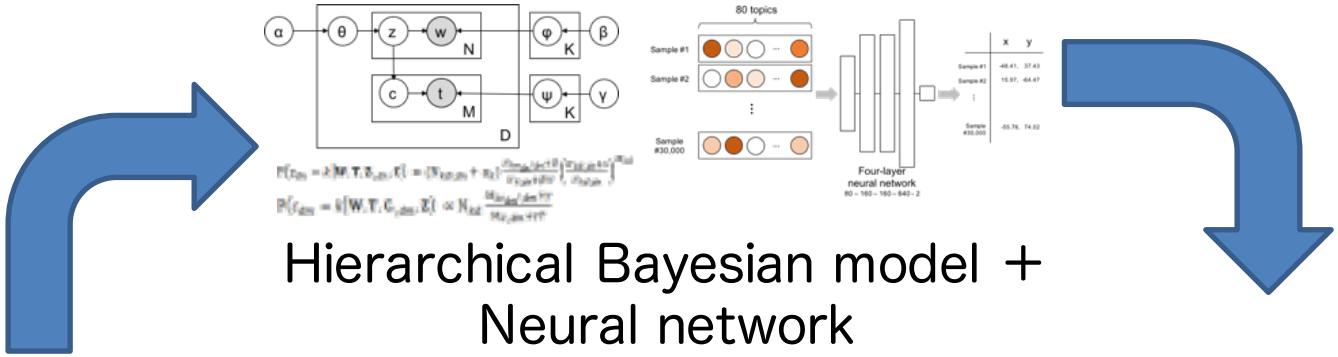


MeGAP3 (MetaGenome Analysis Pipeline for MicrobeDB.jp ver. 3）

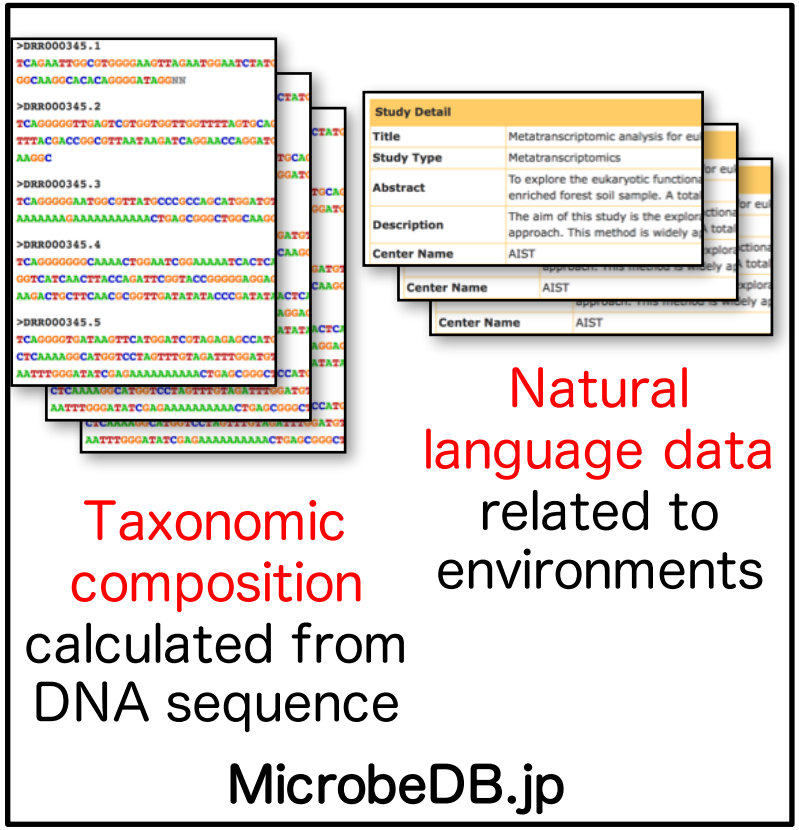


Latent Environment Allocation (LEA)

(Higashi K et al. 2018, PLoS Comp Biol)



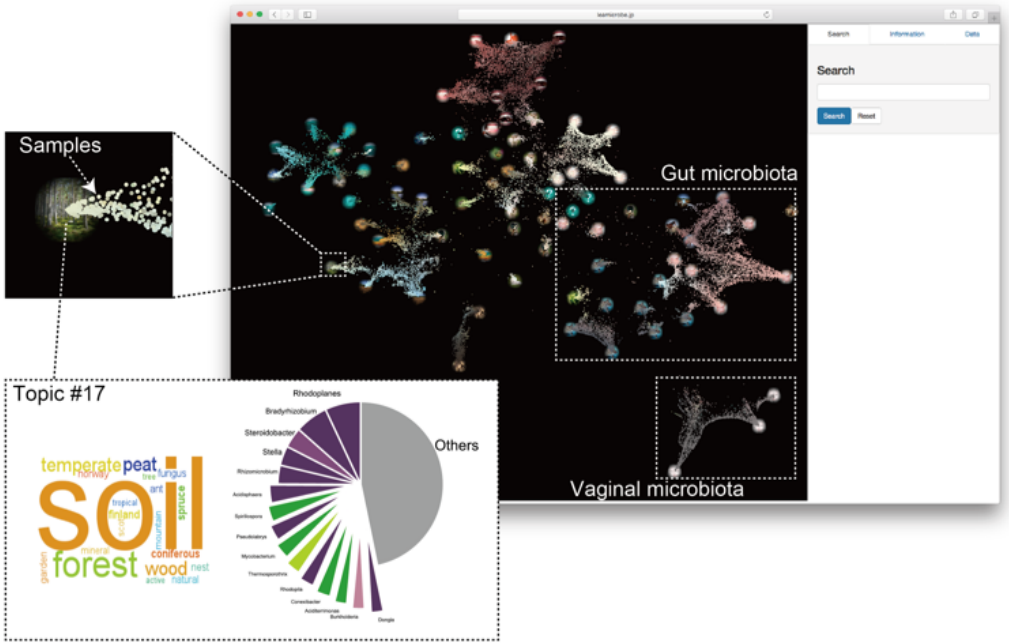
Hierarchical Bayesian model + Neural network



Natural language data related to environments

Taxonomic composition calculated from DNA sequence

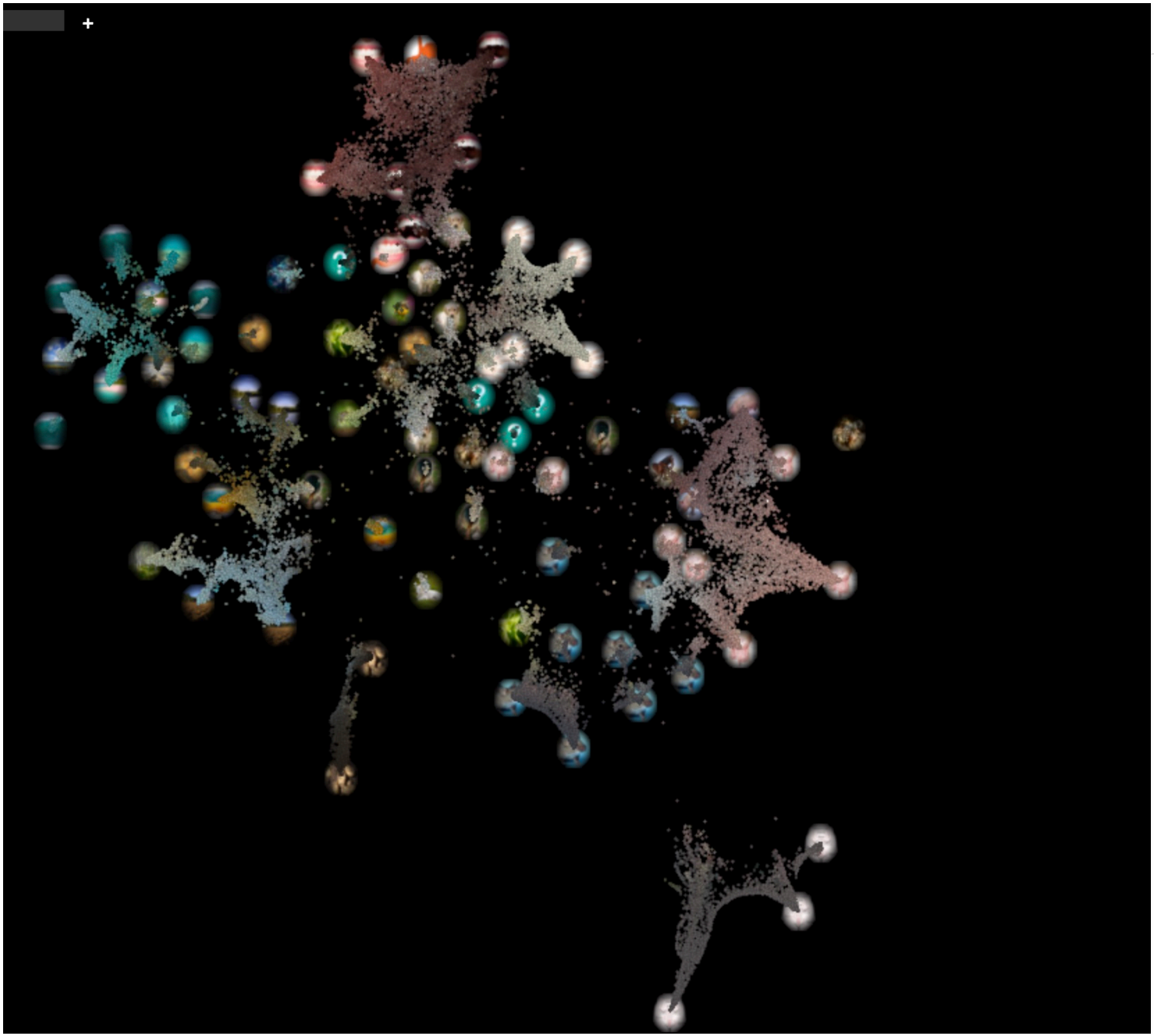
MicrobeDB.jp



Visualization of microbial community compositions with environment topics

http://leamicrobe.jp/

http://leamicrobe.jp/



Search | Information | Data

Sample metadata

Sample ID:

SRS431289

Sample Name:

human gut metagenome

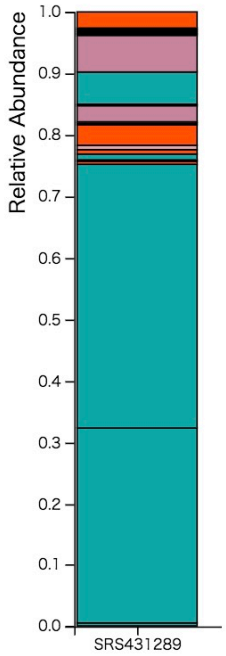
MicrobeDB.jp:

Link

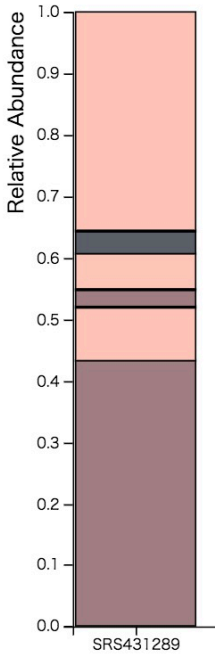
NCBI:

Link

Taxa

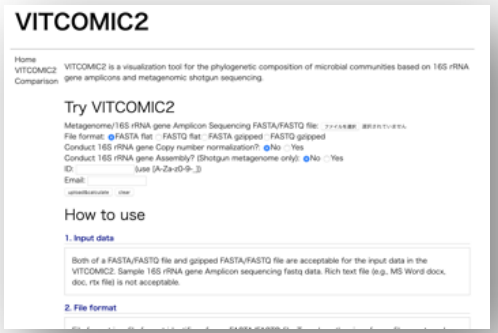


Topics

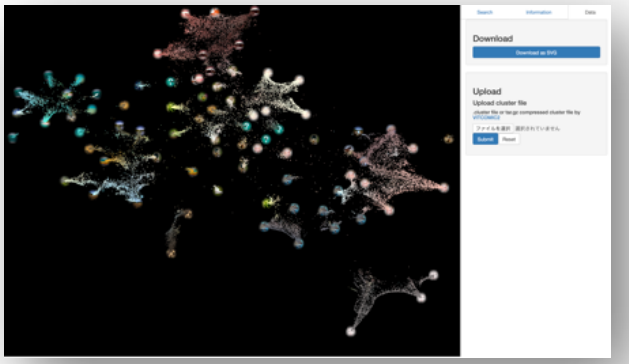


How to map your metagenome data on LEA

VITCOMIC2
(http://vitcomic.org/)



LEA
(http://leamicrobe.jp/)



fastq files

~ 6 min. / a sample (100,000 sequences)

cluster files (taxonomic compositions)

~ 0.5 sec. / a sample

Visualization of your microbial community compositions with environment topics

マイクロバイオーム解析の情報解析について（メタゲノム解析）

Metagenomic sequencing analysis
(メタゲノム解析，ショットガンメタゲノム解析)
(メタトランスクリプトーム解析)

DNA extraction / RNA extraction

Reverse transcription

DNA Sequencing

2016 DBCLS TogoTV / CC-BY-4.0

Metagenomic reads

Assemble — MEGAHIT, metaSPAdes, etc.

Contig sets

Gene finding — Prodigal, MetaGeneMark, etc.

Gene sets

Sequence similarity search — BLASTP, DIAMOND, MMseqs2, etc.

KEGG (Kyoto Encyclopedia of Genes and Genomes)

UniProt

Gene Function abundance
Pathway abundance
Pathway reconstruction
Taxonomic abundance

Sample1 Metadata

Sample2 Metadata

Comparative metagenomics

Who's there?

What are they doing?

Taxonomic assignment strategy?

	Coverage of ref. sequences	Single copy in genomes?	Can analyze eukaryotes and virus?	Robust against HGT?	Example of tools
16S rRNA genes	◯	×	×	◯	VITCOMIC2, MAPseq
Single copy genes	△	◯	×	◯	MAPLE, mOTU2
Unique marker genes	△	◯	×	◯?	MetaPhlAn2
Read mapping	△	×	◯	×	BWA-MEM, Centrifuge
k-mer	△	×	◯	×	Kraken2, Mash

パスウェイデータベース

メタゲノムでは、KEGGのKEGG Orthologyを遺伝子機能の単位として使うことが多い

https://www.genome.jp/kegg/ko.html

KO (KEGG ORTHOLOGY) Database

Linking genomes to pathways by ortholog annotation

Menu PATHWAY BRITE MODULE KO Annotation ENZYME RModule BlastKOALA

Search KO for Go

KO Database of Molecular Functions

The **KO (KEGG Orthology)** database is a database of molecular functions represented in terms of functional orthologs. A functional ortholog is manually defined in the context of KEGG molecular networks, namely, KEGG pathway maps, BRITE hierarchies and KEGG modules. For example, when a pathway map is drawn, each box is given a KO identifier (called K number) and experimentally characterized genes and proteins in specific organisms are used to find orthologs in other organisms. The granularity of "function" is context-dependent, and the resulting KO grouping may correspond to a highly similar sequence group and a limited organism group or it may be a more divergent group.

The KO system is a network-based classification of KOs shown below:

KEGG Orthology (KO)

ただし、KEGGのアミノ酸配列データとKO IDとの対応関係を手軽に取得するためのKEGG FTPサイトへのアクセスは有料

ある程度遠縁なアミノ酸配列も探せる配列類似性検索ツール

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink[1], Chao Xie[2,3] &
Daniel H Huson[1,2]

Nature Methods. 2015

MMseqs2: sensitive protein sequence searching for analysis of massive data sets

Martin Steinegger[1,2] & Johannes Söding[1]

[1]Quantitative and Computational Biology group, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany; [2]Department for Bioinformatics and Computational Biology, Technische Universität München, 85748 Garching, Germany

e-mail: johannes.soeding@mpibpc.mpg.de; martin.steinegger@mpibpc.mpg.de

Nature Biotech. 2017

Faster sequence homology searches by clustering subsequences

Shuji Suzuki[1,2], Masanori Kakuta[1], Takashi Ishida[1] and
Yutaka Akiyama[1,2,*]

Bioinformatics. 2014

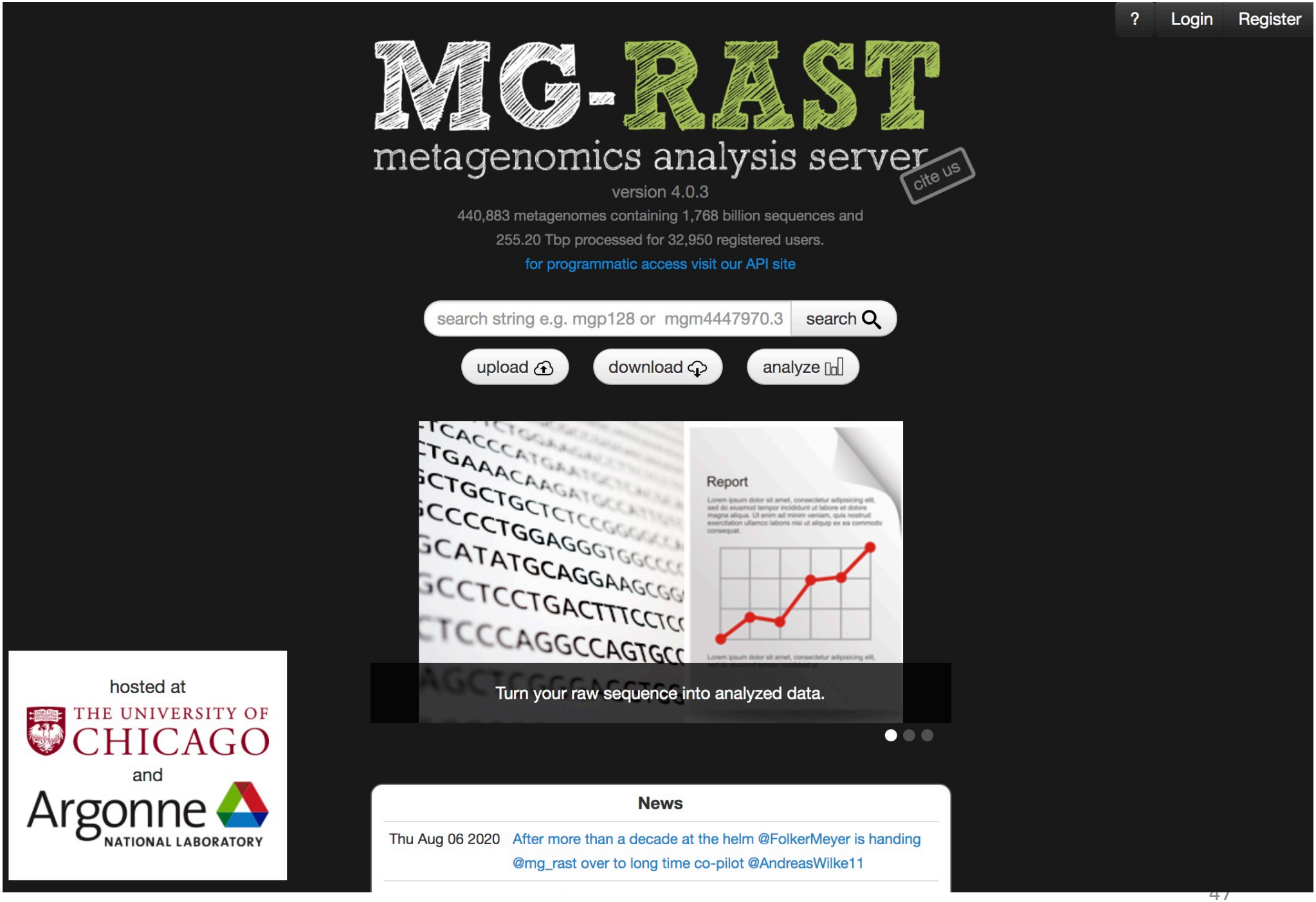
Bioinformatics analysis strategy?

	Full length genes	Gene-neighbor	Draft genome	Minor taxa	HGT	Machine power	Sensitivity against ref. seq.
Assembly	○	○	△	×	○	high	high
Assembly + Binning	○	○	○	×	△	high	high
Read mapping	×	×	×	○	×	low	low
Read-based CDS	△	×	×	○	×	middle	high

Comparison between projects?

DB名	運営組織	DB URL	公開年	受け付ける配列データ	主な系統の参照DB	主な遺伝子機能の参照DB	サンプル数（2019年末）
MG-RAST	シカゴ大学, USA	https://www.mg-rast.org/	2007	リード	SILVA, Greengenes	SEED, KEGG, eggNOG	408,442
IMG/M	JGI, USA	https://img.jgi.doe.gov/m/	2006	Contig, Scaffold	NCBI Taxonomy	Pfam, KEGG, COG	31,406
MGnify	EBI, EU	https://www.ebi.ac.uk/metagenomics/	2013	リード	SILVA	Gene Ontology	215,082
MicrobeDB.jp version 3	国立遺伝学研究所, Japan	https://microbedb.jp	2011	リード	RDP, NCBI Taxonomy	KEGG	1,631,611 (96,766)

https://www.mg-rast.org/



https://www.mg-rast.org/mgmain.html?mgpage=search

shotgun search Your search returned 21,364 results. Showing the first 20 matches. download search results

created date	study	dataset (metagenome name)	sequence type	biome	country
2010-04-23T12:31:20	soybean2_microbiota	soybean2_microbiota	shotgun metagenome	-	-
2010-03-01T17:27:27	MOB_454	MOB_454	shotgun metagenome	-	-
2008-11-15T16:56:34	S12Nov2008	S12Nov2008	shotgun metagenome	-	-
2009-03-04T17:07:05	259_GAS_phagecomp	MGPhage	shotgun metagenome	-	-
2008-10-23T16:12:54	LColon	LColon	shotgun metagenome	-	-
2008-10-20T18:42:52	S891	S89	shotgun metagenome	-	-
2010-03-10T09:57:16	2,4combine	2,4combine	shotgun metagenome	-	-
2009-09-24T08:40:53	cDNA4	cDNA	shotgun metagenome	-	-

https://www.mg-rast.org/mgmain.html?mgpage=overview&metagenome=mgm4902120.3

Analysis Statistics

Upload: bp Count	1,280,938,449 bp
Upload: Sequences Count	7,754,610
Upload: Mean Sequence Length	165 ± 33 bp
Upload: Mean GC percent	58 ± 12 %
Artificial Duplicate Reads: Sequence Count	151,074
Post QC: bp Count	1,155,767,835 bp
Post QC: Sequences Count	7,457,986
Post QC: Mean Sequence Length	155 ± 42 bp
Post QC: Mean GC percent	58 ± 13 %
Processed: Predicted Protein Features	6,357,818
Processed: Predicted rRNA Features	87,754
Alignment: Identified Protein Features	1,979,581
Alignment: Identified rRNA Features	3,665
Annotation: Identified Functional Categories	undefined

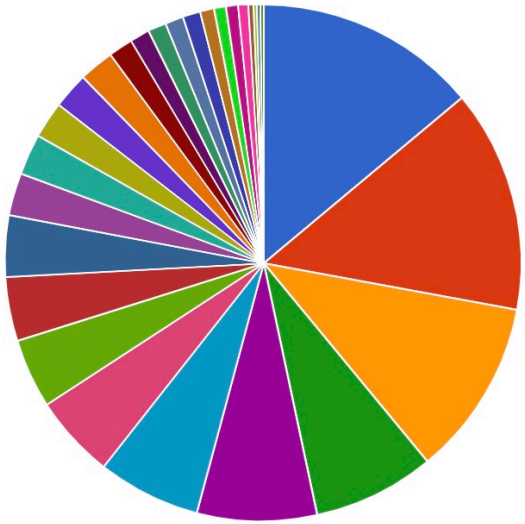
GSC MIxS Info

Investigation Type	WGS
Study Name	MCM_MG
Latitude and Longitude	-78 163
Country and/or Sea, Location	Lake Fryxell Basin
Collection Date	2019-12-27T10:51:00+1200
Environment (Biome)	cold environment
Environment (Feature)	stream
Environment (Material)	sediment
Environmental Package	sediment
Sequencing Method	illumina
More Metadata	click for full table

Table of Contents

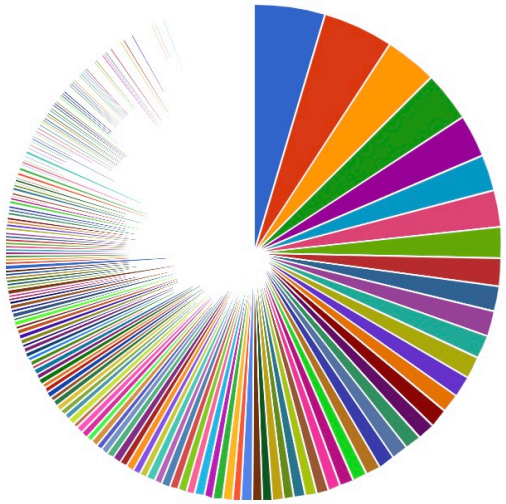
https://www.mg-rast.org/mgmain.html?mgpage=overview&metagenome=mgm4902120.3

Subsystems



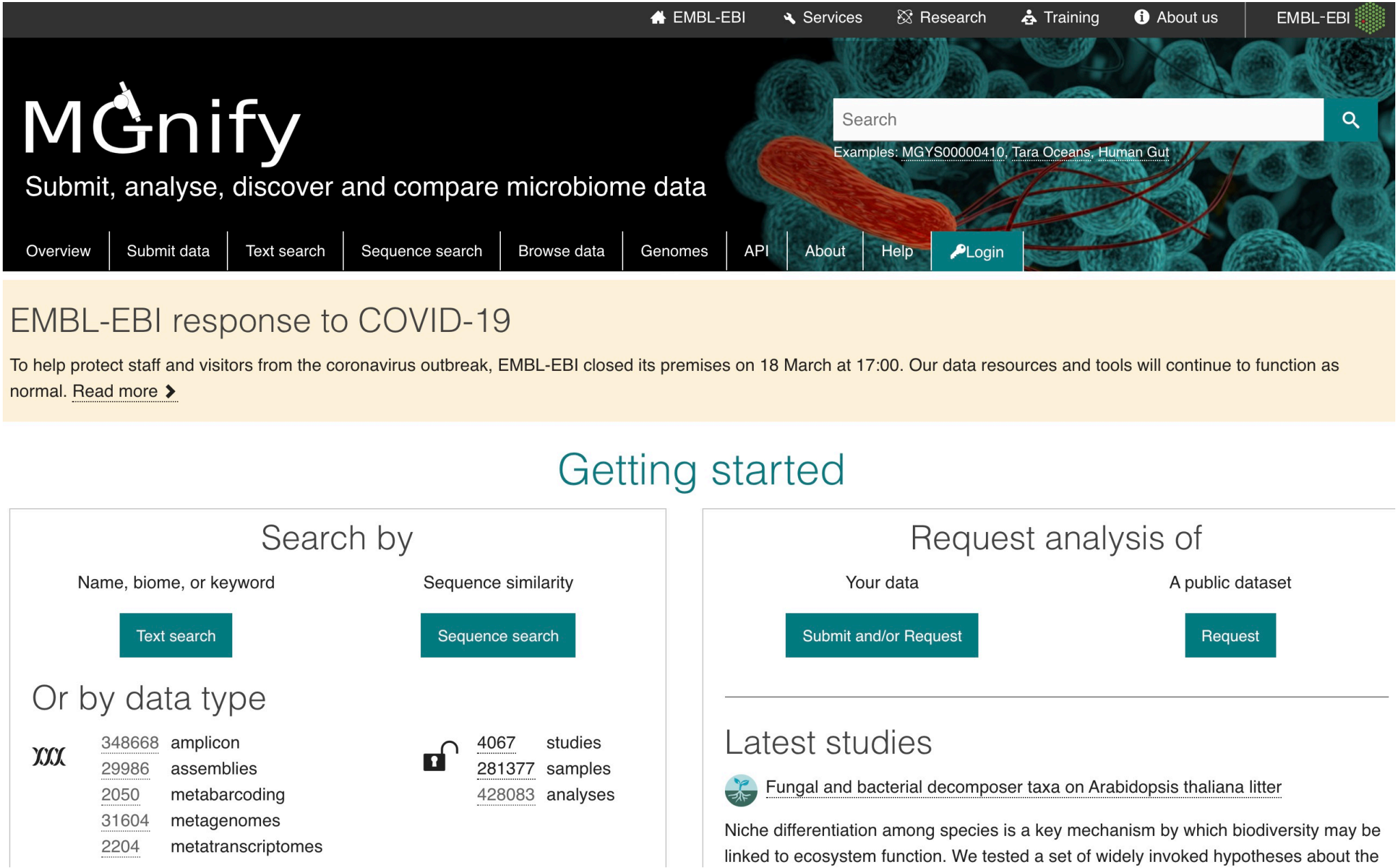
- Carbohydrates - 481,631 (14.01%)
- Clustering-based subsystems - 477,868 (13.90%)
- Amino Acids and Derivatives - 382,620 (11.13%)
- Protein Metabolism - 259,810 (7.56%)
- Miscellaneous - 258,106 (7.51%)
- Cofactors, Vitamins, Prosthetic Groups, Pigments - 222,
- RNA Metabolism - 177,542 (5.16%)
- DNA Metabolism - 150,839 (4.39%)
- Fatty Acids, Lipids, and Isoprenoids - 136,130 (3.96%)
- Cell Wall and Capsule - 133,846 (3.89%)
- Virulence, Disease and Defense - 91,854 (2.67%)
- Nucleosides and Nucleotides - 86,881 (2.53%)

Genus



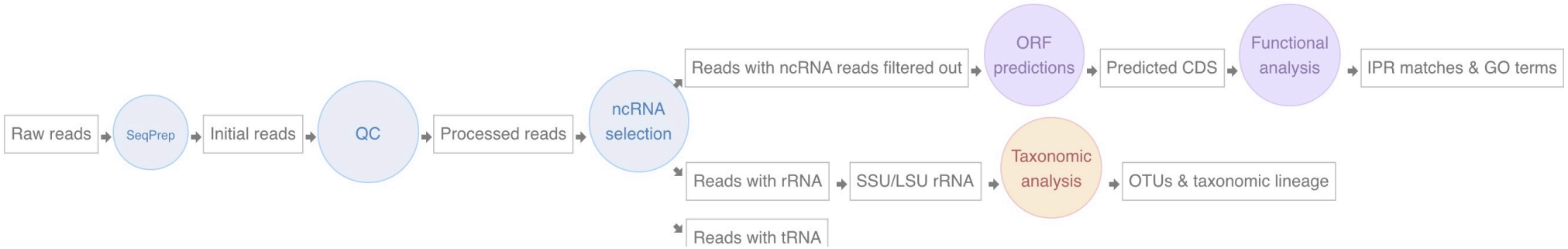
- Gemmata - 26,061 (4.69%)
- Chthoniobacter - 25,111 (4.52%)
- Sorangium - 19,099 (3.44%)
- Nostoc - 18,087 (3.26%)
- Gemmatimonas - 15,169 (2.73%)
- Nitrospira - 13,242 (2.38%)
- Candidatus Solibacter - 12,808 (2.31%)
- Plesiocystis - 11,280 (2.03%)
- Conexibacter - 9,988 (1.80%)
- unclassified (derived from Verrucomicrobia subdivision
- Chitinophaga - 8,894 (1.60%)
- Planctomyces - 8,507 (1.53%)
- Streptomyces - 8,250 (1.49%)
- Haliangium - 7,515 (1.35%)

https://www.ebi.ac.uk/metagenomics/



https://www.ebi.ac.uk/metagenomics/pipelines/4.1

Pipeline version 4.1 - 17-Jan-2018

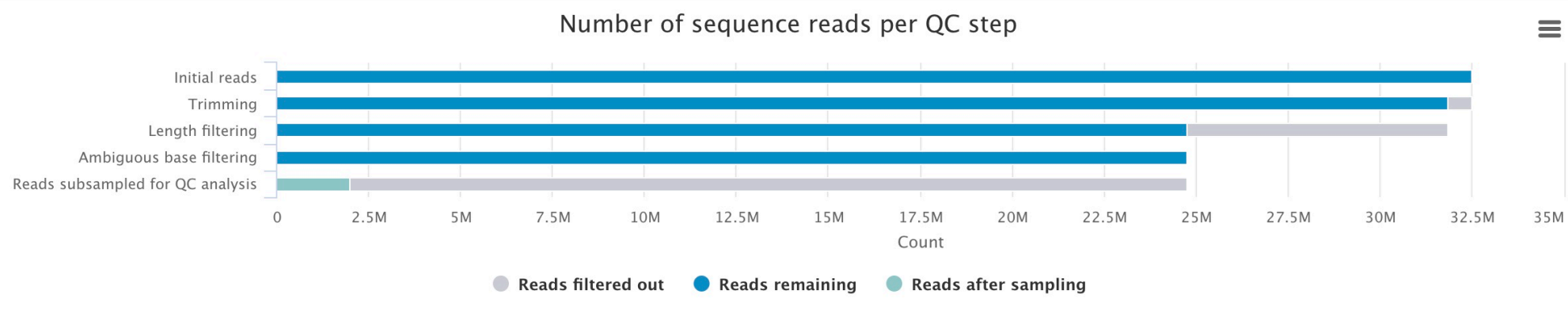


Pipeline tools & steps

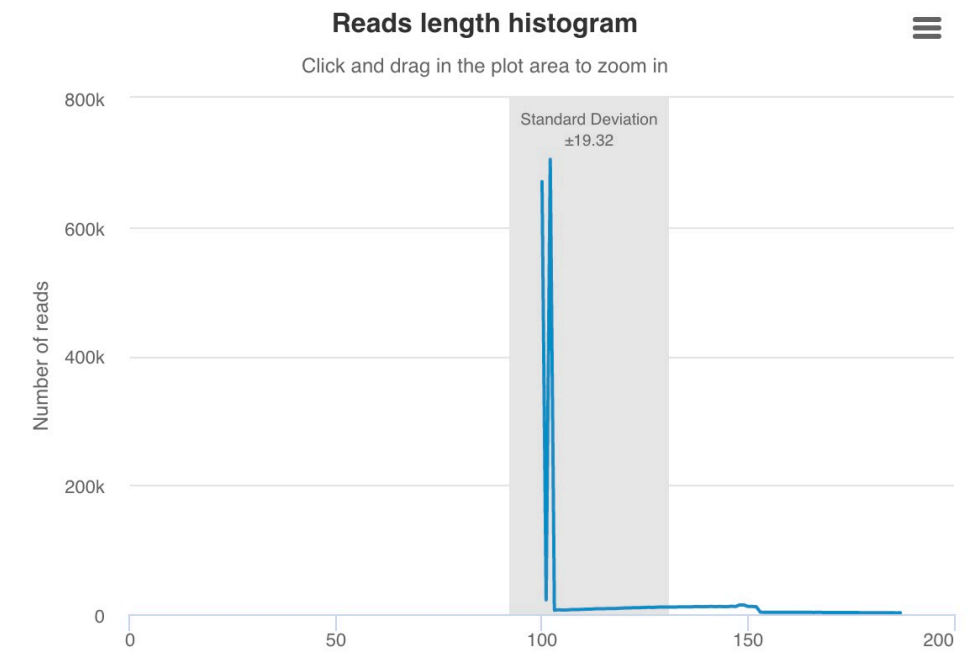
	Tools	Version	Description	How we use it
1	SeqPrep	1.2	A program to merge paired end Illumina reads that are overlapping into a single longer read.	Paired-end overlapping reads are merged - if you want your data assembled, email us.
2.1	Trimmomatic	0.35	A flexible read trimming tool.	Low quality trimming (low quality ends and sequences with > 10% undetermined nucleotides removed). Adapter sequences removed using Biopython SeqIO package.
2.2	Biopython	1.65	A set of freely available tools for biological computation written in Python.	Sequences < 100 nucleotides in length removed.
3.1	Infernal	1.1.2	Infernal ("INFERence of RNA ALignment") is for searching DNA sequence databases for RNA structure and sequence similarities. It is an implementation of a special case of profile stochastic context-free grammars called covariance models (CMs). A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus, so in many cases, it is more capable of identifying RNA homologs that conserve their secondary structure more than their primary sequence.	Identification of ncRNAs.
3.2	cmsearch deoverlap script	0.01	A tool, which removes lower scoring overlaps from cmsearch --tblout files.	Removes lower scoring overlaps from cmsearch --tblout files.
4.1	FragGeneScan	1.20	An application for finding (fragmented) genes in short reads.	Run as a combined gene caller component, giving priority to Prodigal predictions in the case of assembled sequences or FragGeneScan for short reads (all predictions from the higher priority caller are used, supplemented by any non-overlapping regions predicted by the other).

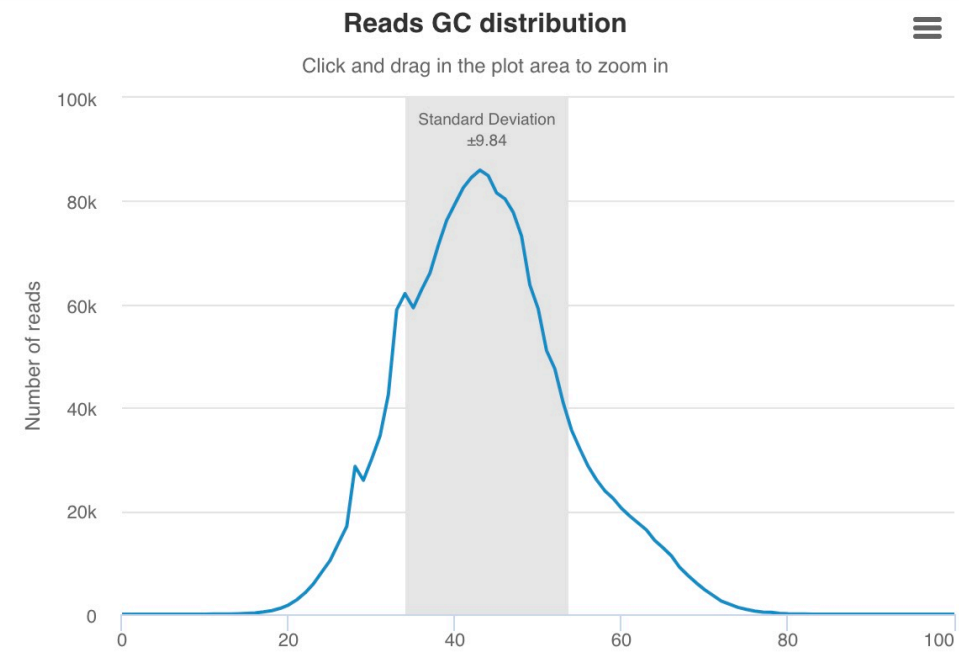
https://www.ebi.ac.uk/metagenomics/analyses/MGYA00373091#qc

The chart below shows the number of reads which pass the quality control steps in the pipeline. Paired-end sequences may have been merged, in which case the initial number of reads may differ from the number given by ENA.



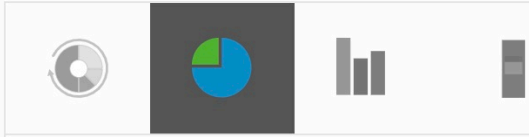
The histograms below show the distributions of sequence lengths (left) and percentage GC content (right) for the sequences having passed quality control. Note that for large files, the distributions were compiled from a random subset of 2 million reads. The standard deviations are shown on each plot. The bar chart underneath each graph indicates the minimum, mean and maximum length and mean GC and AT content, respectively.

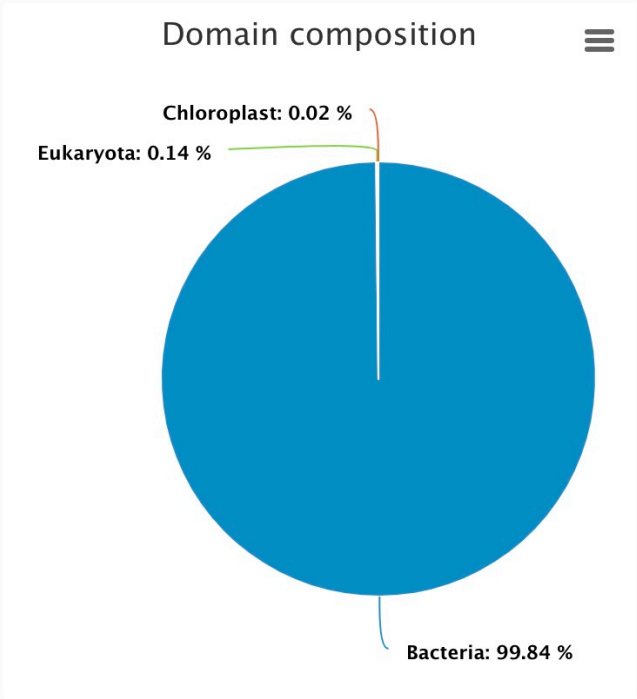


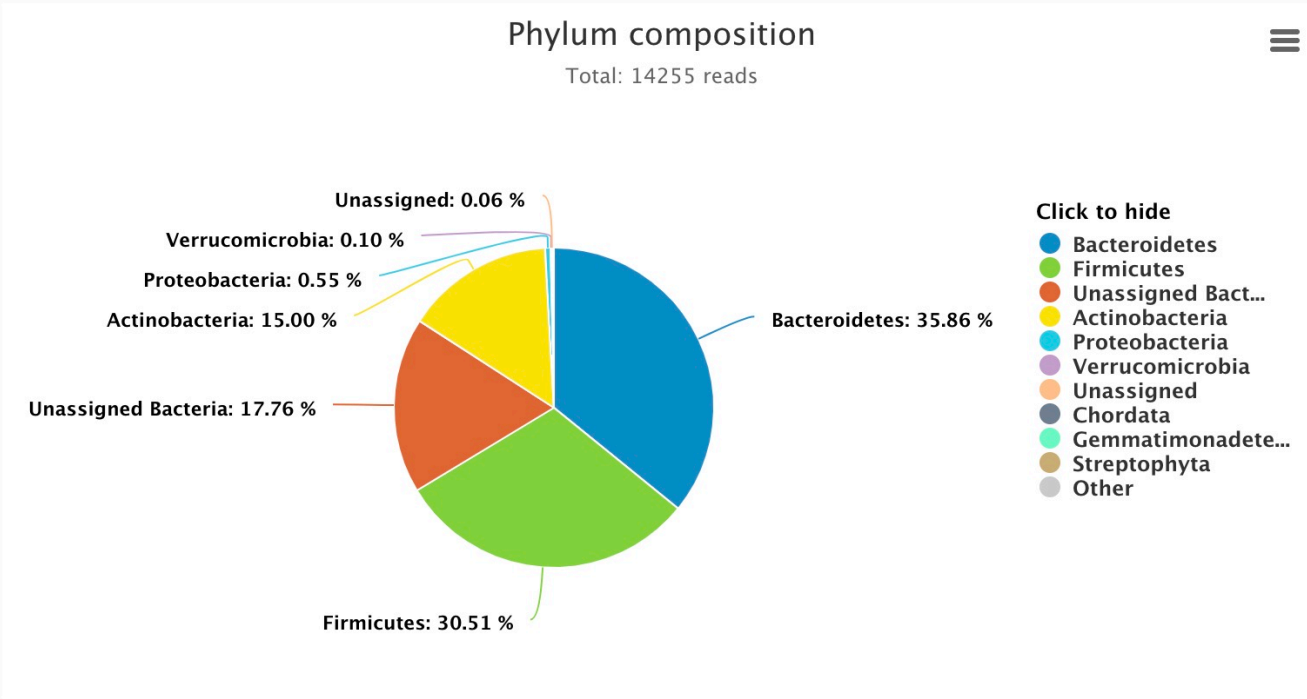


https://www.ebi.ac.uk/metagenomics/analyses/MGYA00373091#taxonomic

These are the results from the taxonomic analysis steps of our pipeline. You can switch between different views of the data using the menu of icons below (pie, bar, stacked and interactive krona charts). The data used to build these charts can be found under the "Download" tab.







Show: 25 results from a total of 15 (page 1 of 1).

Download results

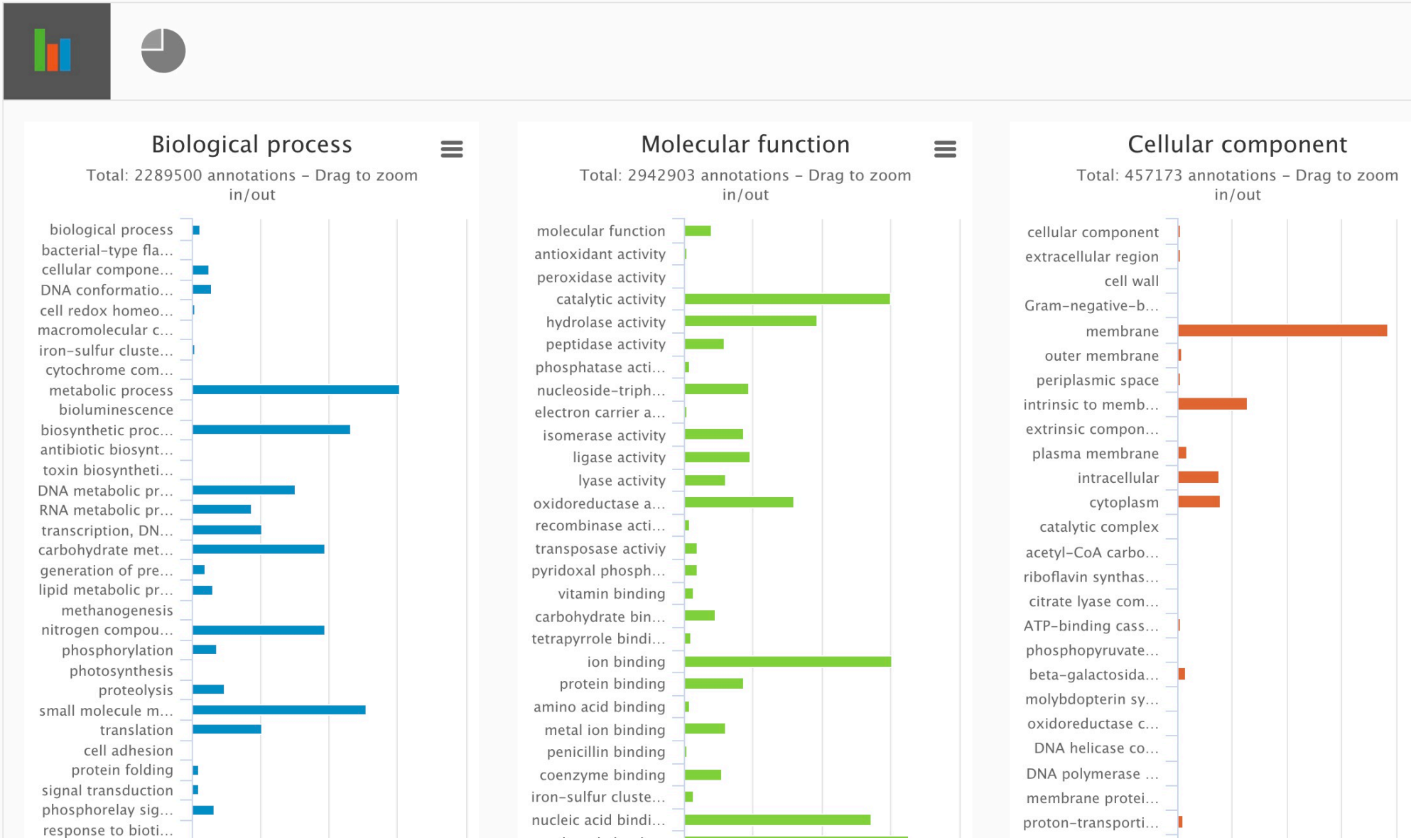
	Phylum	Domain	Unique OTUs	%
1	Bacteroidetes	Bacteria	5112	35.86

https://www.ebi.ac.uk/metagenomics/analyses/MGYA00373091#functional/go

GO terms annotation

A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.

Switch view:



Comparison between projects?

DB名	運営組織	DB URL	公開年	受け付ける配列データ	主な系統の参照DB	主な遺伝子機能の参照DB	サンプル数（2019年末）
MG-RAST	シカゴ大学, USA	https://www.mg-rast.org/	2007	リード	SILVA, Greengenes	SEED, KEGG, eggNOG	408,442
IMG/M	JGI, USA	https://img.jgi.doe.gov/m/	2006	Contig, Scaffold	NCBI Taxonomy	Pfam, KEGG, COG	31,406
MGnify	EBI, EU	https://www.ebi.ac.uk/metagenomics/	2013	リード	SILVA	Gene Ontology	215,082
MicrobeDB.jp version 3	国立遺伝学研究所, Japan	https://microbedb.jp	2011	リード	RDP, NCBI Taxonomy	KEGG	1,631,611 (96,766)