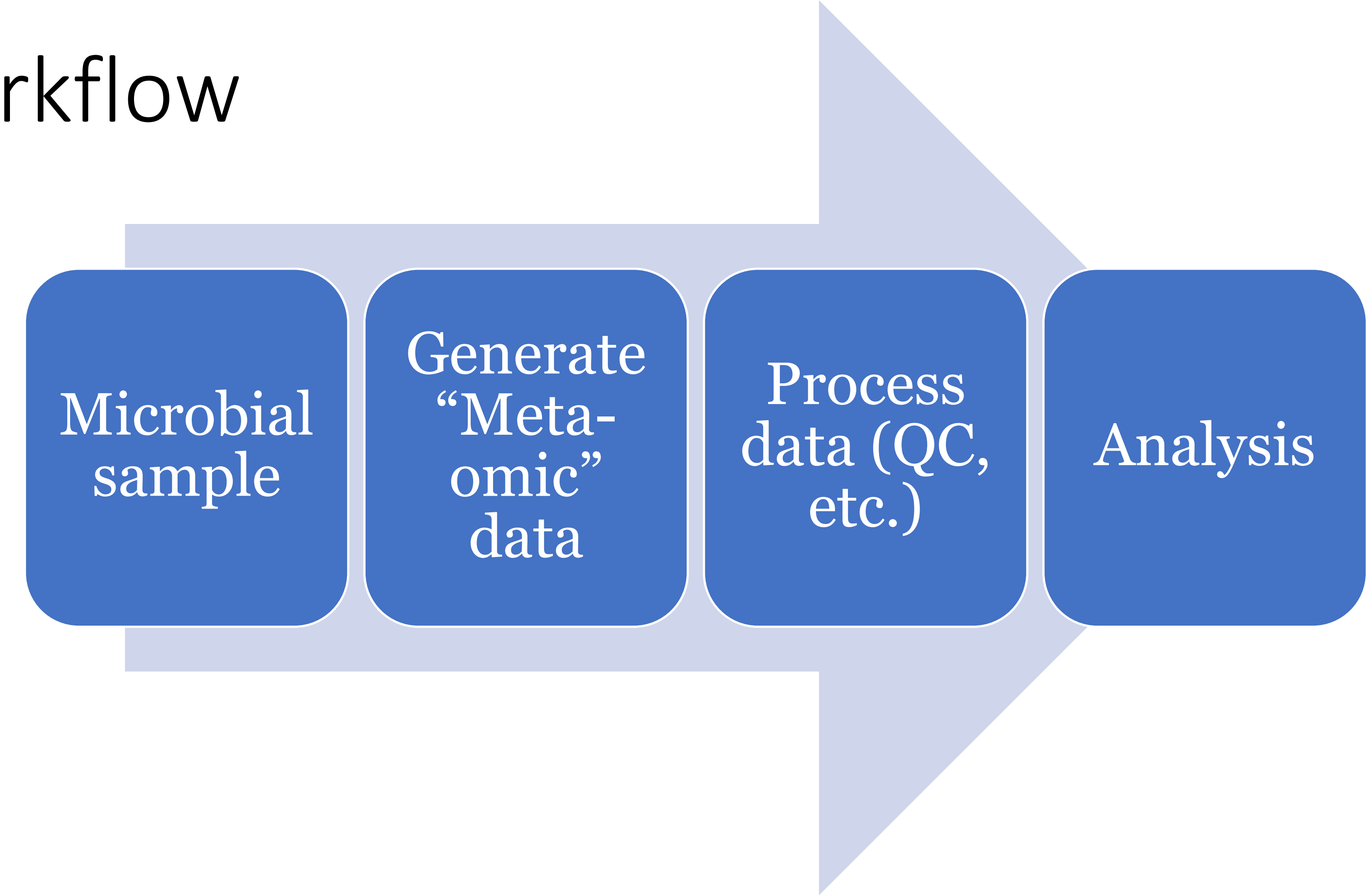
CBC Data Therapy

Metagenomics Discussion

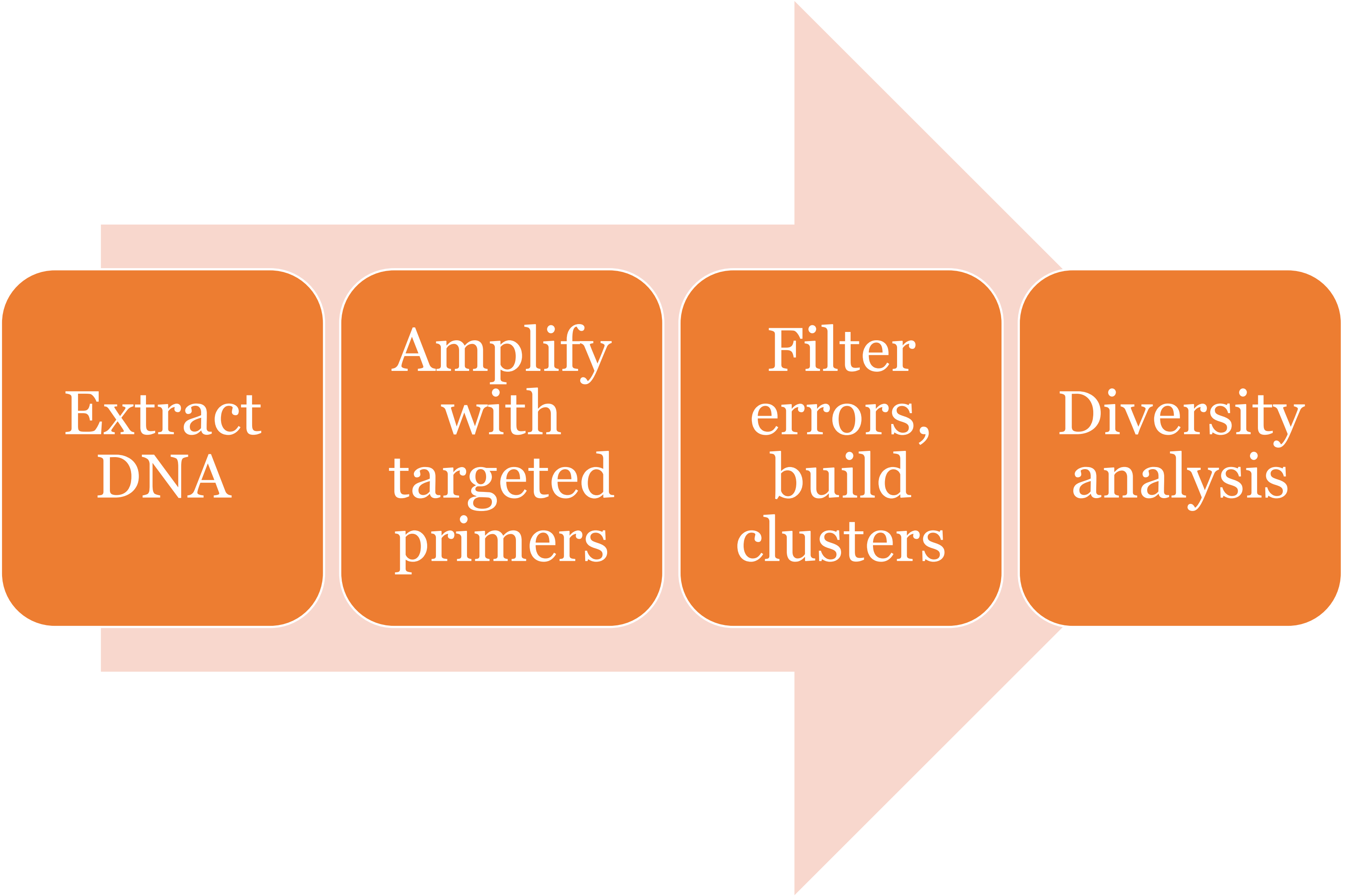
Computational Biology Core

UCONN
UNIVERSITY OF CONNECTICUT

General Workflow



Marker Genes



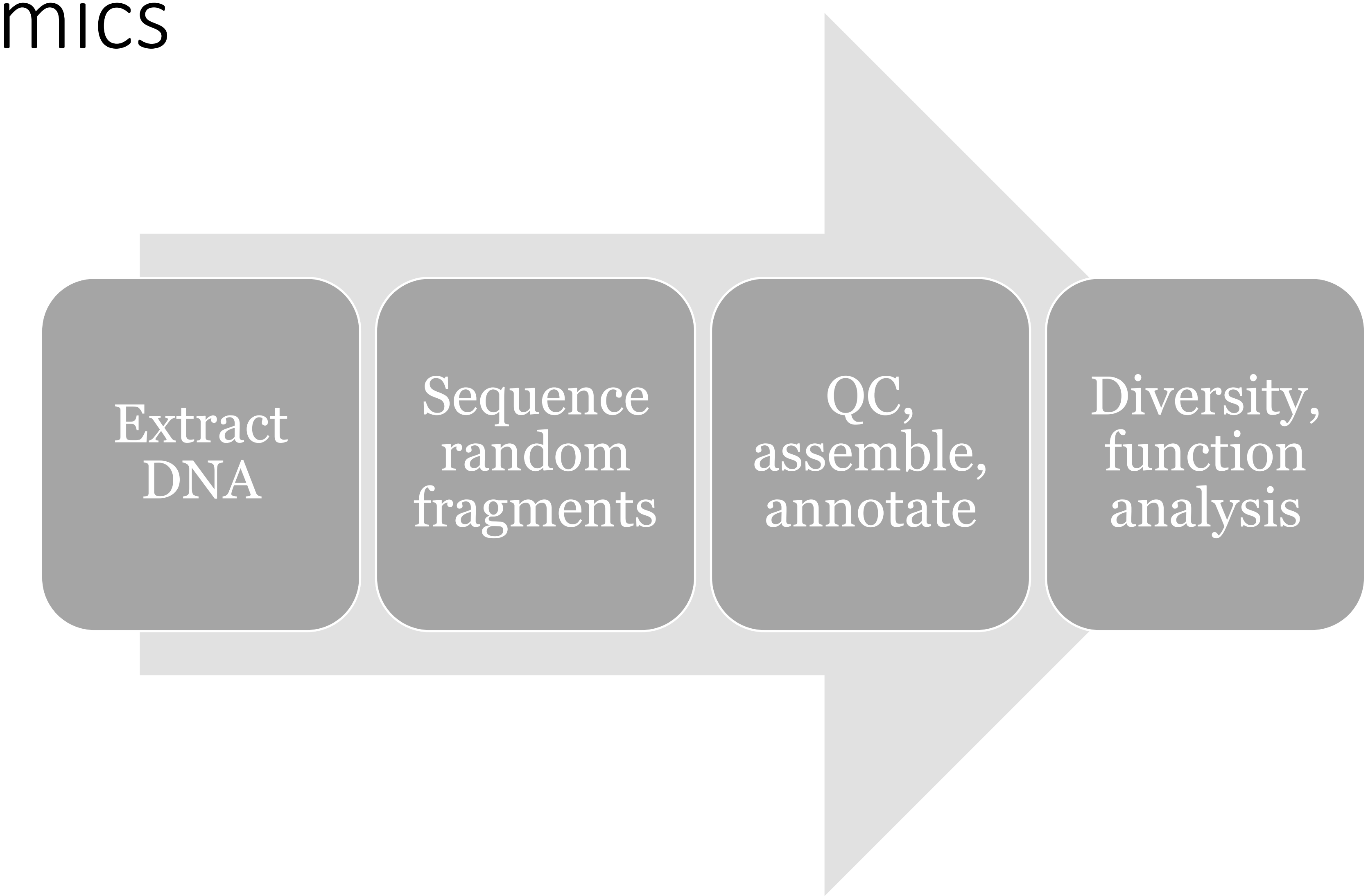
Extract DNA

Amplify with targeted primers

Filter errors, build clusters

Diversity analysis

Metagenomics



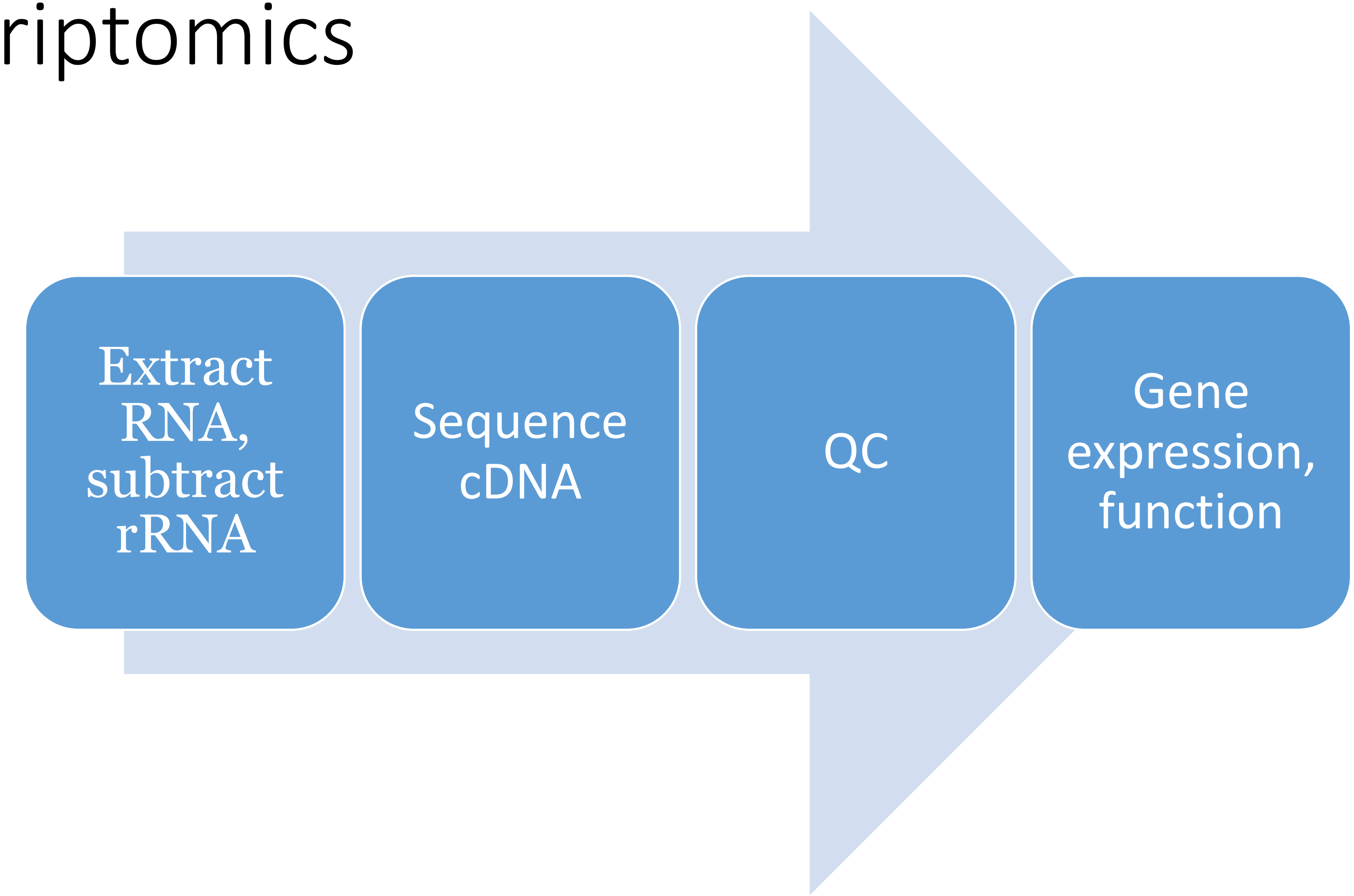
Extract DNA

Sequence random fragments

QC, assemble, annotate

Diversity, function analysis

Metatranscriptomics



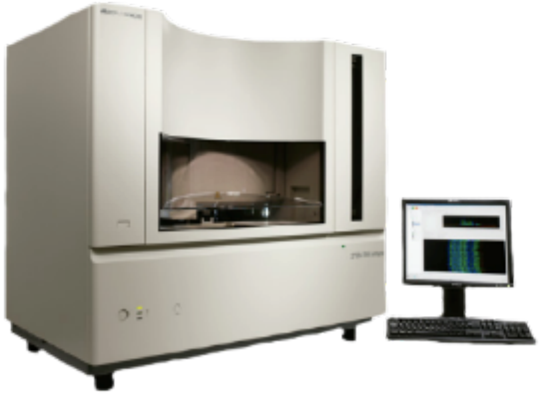
Extract RNA, subtract rRNA

Sequence cDNA

QC

Gene expression, function

Sequencing

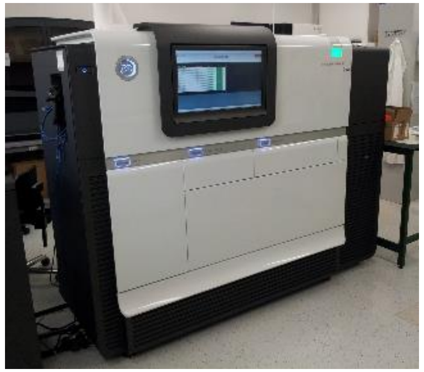


Sanger

Ion Torrent

Roche 454

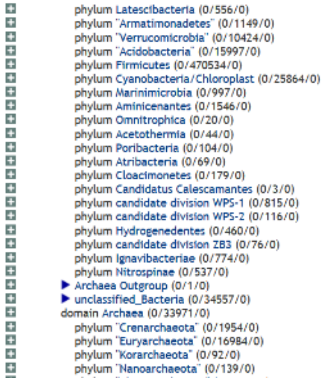
Illumina *Seq



Pacific Biosciences

Nanopore

Resources (16S)



- phylum Latescibacteria (0/556/0)
- phylum "Armatimonadetes" (0/1149/0)
- phylum "Verrucomicrobia" (0/10424/0)
- phylum "Acidobacteria" (0/15997/0)
- phylum Firmicutes (0/470534/0)
- phylum Cyanobacteria/Chloroplast (0/25864/0)
- phylum Marinimicrobia (0/997/0)
- phylum Aminicenantes (0/1546/0)
- phylum Omnitrophica (0/20/0)
- phylum Acetothermia (0/44/0)
- phylum Poribacteria (0/104/0)
- phylum Atribacteria (0/69/0)
- phylum Cloacimonetes (0/179/0)
- phylum Candidatus Calescamantes (0/3/0)
- phylum candidate division WPS-1 (0/815/0)
- phylum candidate division WPS-2 (0/116/0)
- phylum Hydrogenedentes (0/460/0)
- phylum candidate division ZB3 (0/76/0)
- phylum Ignavibacteriae (0/774/0)
- phylum Nitrospinae (0/537/0)
- ▶ Archaea Outgroup (0/1/0)
- ▶ unclassified_Bacteria (0/34557/0)

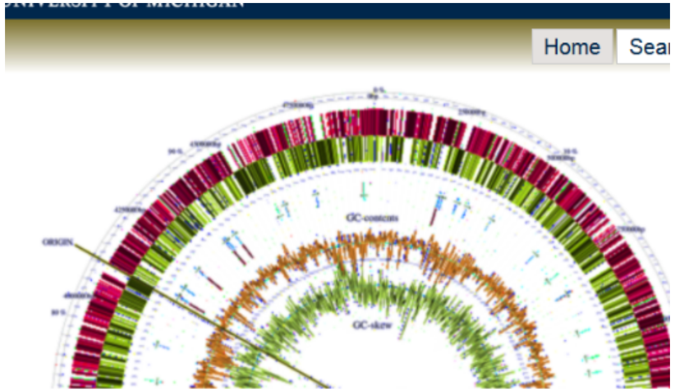
domain Archaea (0/33971/0)

- phylum "Crenarchaeota" (0/1954/0)
- phylum "Euryarchaeota" (0/16984/0)
- phylum "Korarchaeota" (0/92/0)
- phylum "Nanoarchaeota" (0/139/0)

RDP II: Cole et al.
NAR (2013)

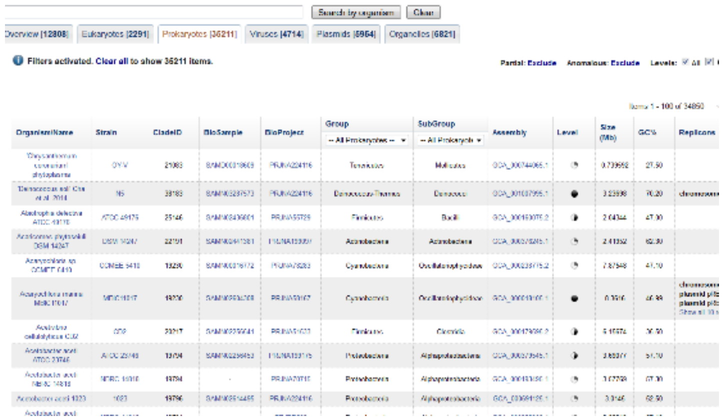
silva
high quality ribosomal RNA databases

SILVA: Quast et al.
NAR (2015)



rrnDB: Stoddard et al.
NAR (2016)

Resources (Genomes)



Search by organism | Clear

Overview [12808] | Eukaryotes [2291] | Prokaryotes [36211] | Viruses [4714] | Plasmids [5964] | Organelles [6821]

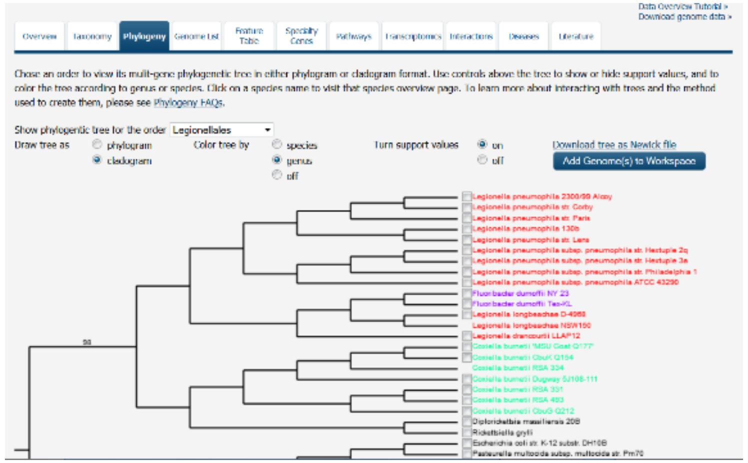
Filters activated. Clear all to show 36211 items.

Partial: Exclude Anomalous: Exclude Levels: ☑ All ☑

Items 1 - 100 of 34850

OrganismName	Strain	CladeID	BioSample	BioProject	Group -- All Prokaryotes --	SubGroup -- All Prokaryotes --	Assembly	Level	Size (Mb)	GC%	Replicons
Chrysanthemum coronarium phytoplasma	CYV	21083	SAMD00018609	PRJNA224116	Tenericutes	Mollicutes	GCA_000744065.1	◔	0.739692	27.50	
'Deinococcus soli' Cha et al. 2014	N5	39183	SAMN03237573	PRJNA224116	Deinococcus-Thermus	Deinococci	GCA_001007995.1	●	3.23998	70.20	chromosome
Abiotrophia defectiva ATCC 49176	ATCC 49176	25146	SAMN00436601	PRJNA55729	Firmicutes	Bacilli	GCA_000160075.2	◑	2.04344	47.00	
Acaricomes phytoseiuli DSM 14247	DSM 14247	22191	SAMN02441381	PRJNA193097	Actinobacteria	Actinobacteria	GCA_000376245.1	◔	2.41352	62.80	
Acaryochloris sp. CCMEE 5410	CCMEE 5410	13290	SAMN00016172	PRJNA78283	Cyanobacteria	Oscillatoriophycideae	GCA_000238775.2	◔	7.87548	47.10	
Acaryochloris marina MBIC11017	MBIC11017	19200	SAMN02604308	PRJNA58167	Cyanobacteria	Oscillatoriophycideae	GCA_000018105.1	●	8.3616	46.99	chromosome plasmid pR plasmid pR Show all 10
Acetivibrio cellulolyticus CD2	CD2	23217	SAMN02299641	PRJNA51723	Firmicutes	Clostridia	GCA_000179635.2	◑	6.19574	36.50	
Acetobacter aceti ATCC 23746	ATCC 23746	13704	SAMN02256453	PRJNA193175	Proteobacteria	Alphaproteobacteria	GCA_000379545.1	◑	3.85077	57.10	
Acetobacter aceti NBRC 14818	NBRC 14818	13794	-	PRJNA70715	Proteobacteria	Alphaproteobacteria	GCA_000193495.1	◔	3.77729	57.00	
Acetobacter aceti 1023	1023	13796	SAMN02614495	PRJNA224116	Proteobacteria	Alphaproteobacteria	GCA_000691125.1	◔	3.0146	52.50	

GenBank Genomes



Data Overview Tutorial »
Download genome data »

Overview | Taxonomy | Phylogeny | Genome List | Feature Table | Specialty Genes | Pathways | Transcriptomics | Interactions | Diseases | Literature

Chose an order to view its multi-gene phylogenetic tree in either phylogram or cladogram format. Use controls above the tree to show or hide support values, and to color the tree according to genus or species. Click on a species name to visit that species overview page. To learn more about interacting with trees and the method used to create them, please see Phylogeny FAQs.

Show phylogentic tree for the order Legionellales

Draw tree as ○ phylogram ● cladogram

Color tree by ○ species ● genus ○ off

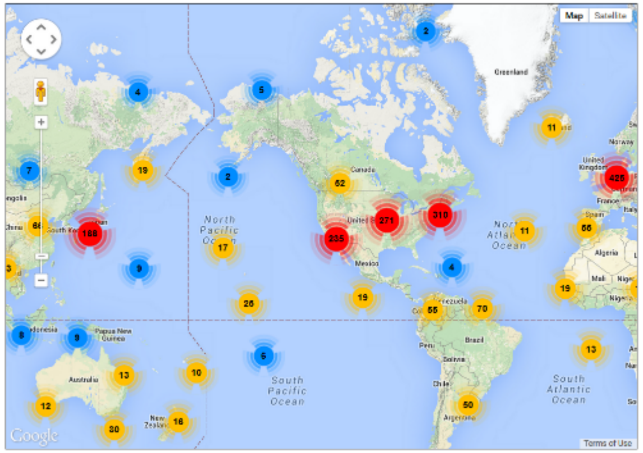
Turn support values ● on ○ off

Download tree as Newick file

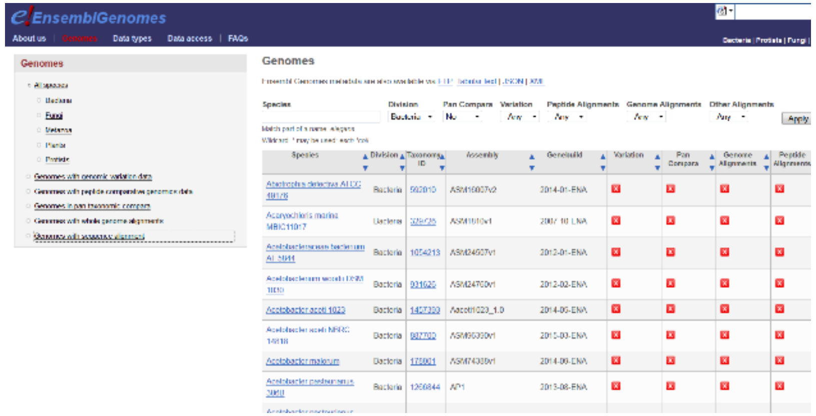
Add Genome(s) to Workspace

98
Legionella pneumophila 2300/99 Alcoy
Legionella pneumophila str. Corby
Legionella pneumophila str. Paris
Legionella pneumophila 130b
Legionella pneumophila str. Lens
Legionella pneumophila subsp. pneumophila str. Hextuple 2q
Legionella pneumophila subsp. pneumophila str. Hextuple 3a
Legionella pneumophila subsp. pneumophila str. Philadelphia 1
Legionella pneumophila subsp. pneumophila ATCC 43290
Fluoribacter dumoffii NY 23
Fluoribacter dumoffii Tex-KL
Legionella longbeachae D-4968
Legionella longbeachae NSW150
Legionella drancourtii LLAP12
Coxiella burnetii 'MSU Goat Q177'
Coxiella burnetii CbuK Q154
Coxiella burnetii RSA 334
Coxiella burnetii Dugway 5J108-111
Coxiella burnetii RSA 331
Coxiella burnetii RSA 493
Coxiella burnetii CbuG Q212
Diplorickettsia massiliensis 20B
Rickettsiella grylli
Escherichia coli str. K-12 substr. DH10B
Pasteurella multocida subsp. multocida str. Pm70

PATRIC (host and bacterial)



GOLD (JGI metagenomes)



EnsemblGenomes

About us | Genomes | Data types | Data access | FAQs

Bacteria | Protists | Fungi |

Genomes

- All species
  - Bacteria
  - Fungi
  - Metazoa
  - Plants
  - Protists
- Genomes with genomic variation data
- Genomes with peptide comparative genomics data
- Genomes in pan taxonomic compara
- Genomes with whole genome alignments
- Genomes with sequence alignment

Genomes

Ensembl Genomes metadata are also available via FTP: tabular text | JSON | XML

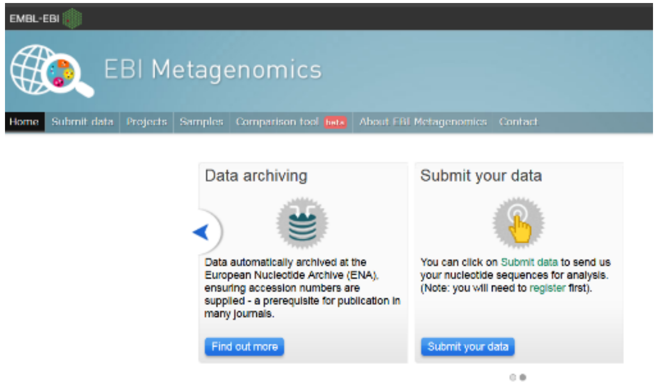
Species: Division: Bacteria | Pan Compara: No | Variation: Any | Peptide Alignments: Any | Genome Alignments: Any | Other Alignments: Any | Apply

Match part of a name: elegans
Wildcard * may be used: esch*coli

Species	Division	Taxonomy ID	Assembly	Genebuild	Variation	Pan Compara	Genome Alignments	Peptide Alignments
Abiotrophia defectiva ATCC 49176	Bacteria	592010	ASM16007v2	2014-01-ENA	☒	☒	☒	☒
Acaryochloris marina MBIC11017	Bacteria	329726	ASM1810v1	2007-10-ENA	☒	☒	☒	☒
Acetobacteraceae bacterium AT-5844	Bacteria	1054213	ASM24507v1	2012-01-ENA	☒	☒	☒	☒
Acetobacterium woodii DSM 1030	Bacteria	931626	ASM24760v1	2012-02-ENA	☒	☒	☒	☒
Acetobacter aceti 1023	Bacteria	1457393	Aaceti1023_1.0	2014-06-ENA	☒	☒	☒	☒
Acetobacter aceti NBRC 14818	Bacteria	887700	ASM96330v1	2015-03-ENA	☒	☒	☒	☒
Acetobacter malorum	Bacteria	178901	ASM74330v1	2014-09-ENA	☒	☒	☒	☒
Acetobacter pasteurianus 386B	Bacteria	1266844	AP1	2013-08-ENA	☒	☒	☒	☒

Ensembl Genomes

Resources (Metagenomes)



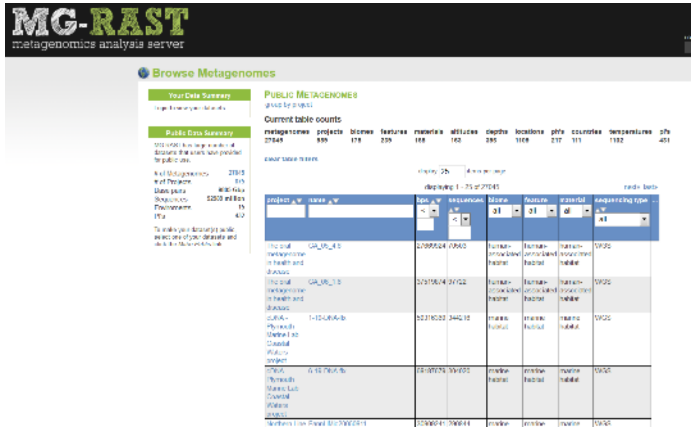
Data archiving

Data automatically archived at the European Nucleotide Archive (ENA), ensuring accession numbers are supplied - a prerequisite for publication in many journals.

Submit your data

You can click on Submit data to send us your nucleotide sequences for analysis. (Note: you will need to register first).

EBI metagenomics



Public Metagenomes

Current table counts

metagenomes	projects	biomes	features	materials	altitudes	depths	locations	ph's	countries	temperatures	pfs
27045	999	178	239	188	162	395	1108	217	111	1182	431

displaying 1 - 25 of 27045

project	name	bps	sequences	biome	feature	material	sequencing type
The oral metagenome in health and disease	CA_06_4.5	[illegible]	[illegible]	human-associated habitat	human-associated habitat	human-associated habitat	WGS
The oral metagenome in health and disease	CA_06_1.5	[illegible]	[illegible]	human-associated habitat	human-associated habitat	human-associated habitat	WGS
cDNA - Plymouth Marine Lab Coastal Waters project	1-10-DNA-fb	[illegible]	[illegible]	marine habitat	marine habitat	marine habitat	WGS
cDNA - Plymouth Marine Lab Coastal Waters project	6-10-DNA-fb	[illegible]	[illegible]	marine habitat	marine habitat	marine habitat	WGS
Northern Line [illegible]	[illegible]	[illegible]	[illegible]	marine	marine	marine	WGS

MG-RAST

Welcome to the Data Analysis and Coordination Center (DACC) for the National Institutes of Health (NIH) Common Fund supported Human Microbiome Project (HMP). This site is the central repository for all HMP data. The aim of the HMP is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. More information can be found in the menus above and on the NIH Common Fund site

- January 2015
  Metagenomic Analysis Workshop March 3-6
- September 2014
  HMC 2015 from Mar. 31 to Apr. 2
- May 2014
  Poster and Booth at ASM 2014

- Identification of a gene in Mycoplasma hominis associated with preterm...
- Diet in the pathogenesis and treatment of inflammatory bowel diseases...
- Using corticosteroids to reshape the gut microbiome: implications for ...

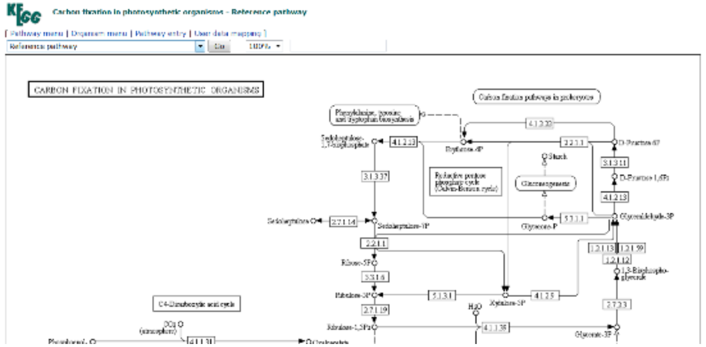
- NIH Common Fund
- NCBI HMP Data Repository

Human Microbial Sampling

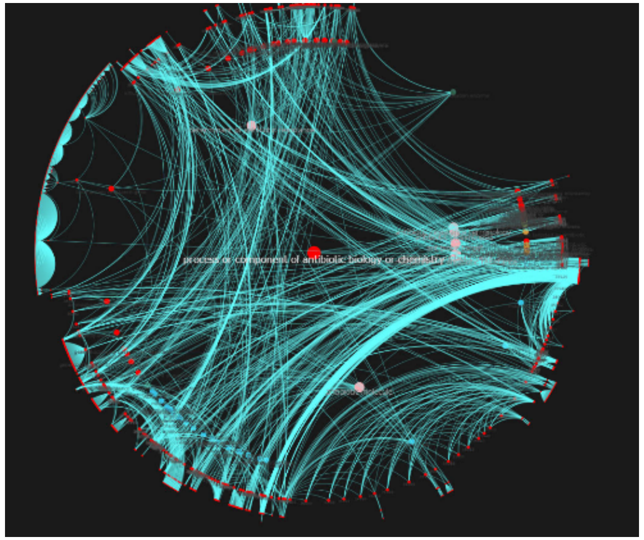
16S rRNA and whole metagenome sequencing of samples collected from 300 healthy human participants, to characterize complexity of microbial communities at individual body sites and to provide insights into functions performed by the human microbiome...

HMP DACC

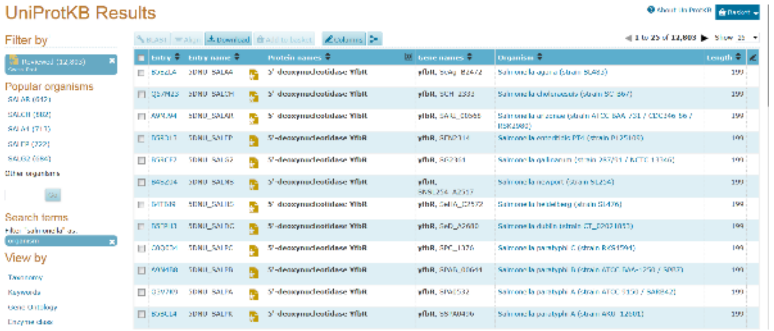
Resources (Function)


KEGG
Carbon fixation in photosynthetic organisms - Reference pathway
[Pathway menu | Organism menu | Pathway entry | User data mapping]
Reference pathway
Go
100%
CARBON FIXATION IN PHOTOSYNTHETIC ORGANISMS
Carbon fixation pathways in prokaryotes
Phenylalanine, tyrosine and tryptophan biosynthesis
Reductive pentose phosphate cycle (Calvin-Benson cycle)
Gluconeogenesis
C4-Dicarboxylic acid cycle
Starch
Glycerate-3P
Glyceraldehyde-3P
D-Fructose-6P
D-Fructose-1,6P2
Sedoheptulose-7P
Ribose-5P
Ribulose-5P
Ribulose-1,5P2
Xylulose-5P
Erythrose-4P
Phosphoenol-pyruvate
Oxaloacetate
CO2 (atmospheric)
H2O

KEGG



CARD



UniProtKB Results

About UniProtKB | Basket

Filter by

Reviewed (12,803) Swiss-Prot

Popular organisms
- SALAR (642)
- SALCH (682)
- SALA4 (713)
- SALEP (722)
- SALG2 (684)

Other organisms

Go

Search terms

Filter "salmonella" as organism

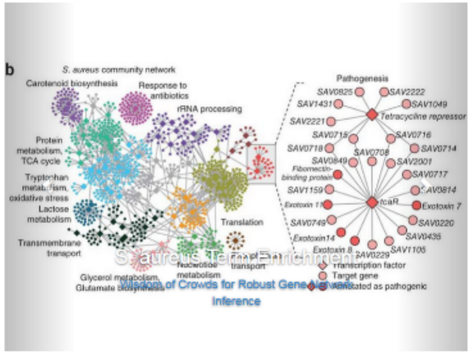
View by
- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class

BLAST | Align | Download | Add to basket | Columns

1 to 25 of 12,803 ▶ Show 25

□	Entry	Entry name		Protein names	Gene names	Organism	Length
□	[illegible]	5DNU_SALA4		5'-deoxynucleotidase YfbR	yfbR, SeAg_B2472	Salmonella agona (strain SL483)	199
□	[illegible]	5DNU_SALCH		5'-deoxynucleotidase YfbR	yfbR, SCH_2333	Salmonella choleraesuis (strain SC-B67)	199
□	[illegible]	5DNU_SALAR		5'-deoxynucleotidase YfbR	yfbR, SARI_00566	Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980)	199
□	[illegible]	5DNU_SALEP		5'-deoxynucleotidase YfbR	yfbR, SEN2314	Salmonella enteritidis PT4 (strain P125109)	199
□	[illegible]	5DNU_SALG2		5'-deoxynucleotidase YfbR	yfbR, SG2361	Salmonella gallinarum (strain 287/91 / NCTC 13346)	199
□	[illegible]	5DNU_SALNS		5'-deoxynucleotidase YfbR	yfbR, SNSL254_A2517	Salmonella newport (strain SL254)	199
□	[illegible]	5DNU_SALHS		5'-deoxynucleotidase YfbR	yfbR, SeHA_C2572	Salmonella heidelberg (strain SL476)	199
□	[illegible]	5DNU_SALDC		5'-deoxynucleotidase YfbR	yfbR, SeD_A2610	Salmonella dublin (strain CT_02021853)	199
□	[illegible]	5DNU_SALPC		5'-deoxynucleotidase YfbR	yfbR, SPC_1376	Salmonella paratyphi C (strain RKS4594)	199
□	[illegible]	5DNU_SALPB		5'-deoxynucleotidase YfbR	yfbR, SPAB_00641	Salmonella paratyphi B (strain ATCC BAA-1250 / SPB7)	199
□	[illegible]	5DNU_SALPA		5'-deoxynucleotidase YfbR	yfbR, SPA0532	Salmonella paratyphi A (strain ATCC 9150 / SARB42)	199
□	[illegible]	5DNU_SALPK		5'-deoxynucleotidase YfbR	yfbR, SSPA0496	Salmonella paratyphi A (strain AKU_12601)	199

UniProtKB



Gene Ontology

General Challenges/Considerations

- Sequencing errors
  - Error rates, error *type* (PacBio: 10% random, Illumina – 0.1% substitution)
- Chimeras
  - Amplification artifacts, cloning of restriction fragments
- 16S: different V regions give different results
- Different sequencing platforms / sampling conditions ALSO give different results
- Workflow complexity / plethora of tools

General Challenges/Considerations

- Strain-level diversity in metagenomes will often be missed by amplicon (esp. short-read) and shotgun approaches
  - This may be especially important **between** samples
- Taxonomy
  - Database predictions (RDP)
- Functional Annotation
  - Coverage versus accuracy

Marker Genes

- Eukaryotic Organisms (protists, fungi)
  - 18S (http://www.arb-silva.de)
  - ITS (http://www.mothur.org/wiki/UNITE_ITS_database)
- Bacteria
  - CPN60 (http://www.cpndb.ca/cpnDB/home.php)
  - ITS (Martiny, Env Micro 2009)
  - RecA gene
- Viruses
  - Gp23 for T4-like bacteriophage
  - RdRp for picornaviruses

Faster evolving markers used for strain-level differentiation

Marker Genes

- Focus on contamination reduction during preparation
- 16S rRNA contains 9 hypervariable regions (V1-V9)
- V4 was chosen because of its size (suitable for Illumina 150bp paired-end sequencing) and phylogenetic resolution
- Different V regions have different phylogenetic resolutions
  – giving rise to slightly different community composition results
- Sequencing:
  - MiSeq capacity allows multiple samples to be combined into a single run
  - Number of reads needed to differentiate samples depends on the nature of the studies
    - Unique DNA barcodes can be incorporated into your amplicons to differentiate samples

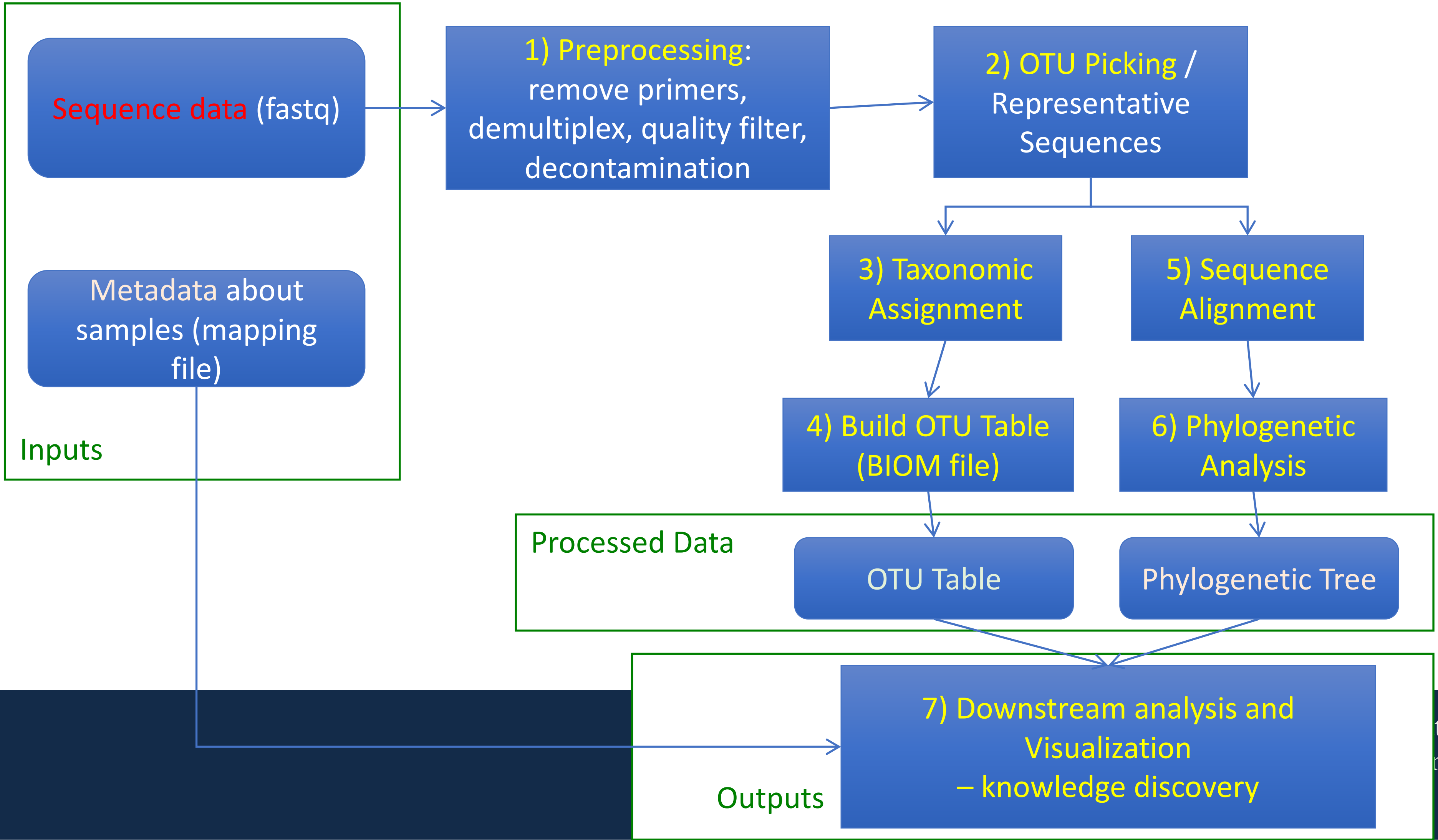
Marker Genes

- QIIME (http://qiime.org)

Quantitative Insights Into Microbial Ecology

- Mothur (http://www.mothur.org)

Bioinformatics
Overall Bioinformatics Workflow



Inputs

Sequence data (fastq)

Metadata about samples (mapping file)

1) Preprocessing: remove primers, demultiplex, quality filter, decontamination

2) OTU Picking / Representative Sequences

3) Taxonomic Assignment

4) Build OTU Table (BIOM file)

5) Sequence Alignment

6) Phylogenetic Analysis

Processed Data

OTU Table

Phylogenetic Tree

Outputs

7) Downstream analysis and Visualization – knowledge discovery

QIIME versus MOTHUR

QIIME	Mothur
A python interface to glue together many programs	Single program with minimal external dependency
Wrappers for existing programs	Reimplementation of popular algorithms
Large number of dependencies / VM available	Easy to install and setup; work best on single multi-core server with lots of memory
More scalable	Less scalable
Steeper learning curve but more flexible workflow if you can write your own scripts	Easy to learn and works the best with built-in tools
http://www.ncbi.nlm.nih.gov/pubmed/24060131	http://www.mothur.org/wiki/MiSeq_SOP

Metagenomics

- Goal: Identify the relative abundance of different microbes in a sample given using metagenomics
- Problems:
  - Reads are all mixed together
  - Reads can be short (~100bp)
  - Lateral gene transfer
- Two broad approaches
  1. Binning Based
  2. Marker Based

Metagenomics

- Attempts to “bin” reads into the genome from which they originated
- Composition-based
  - Uses GC composition or k-mers (e.g. Naïve Bayes Classifier)
  - Generally not very precise and not recommended
- Sequence-based
  - Compare reads to large reference database using BLAST (or some other similarity search method)
  - Reads are assigned based on “Best-hit” or “Lowest Common Ancestor” approach

LCA

- Use all BLAST hits above a threshold and assign taxonomy at the lowest level in the tree which covers these taxa.
- Notable Examples:
  - MEGAN: http://ab.inf.uni-tuebingen.de/software/megan5/
    - One of the first metagenomic tools
    - Does functional profiling too!
  - MG-RAST: https://metagenomics.anl.gov/
    - Web-based pipeline (might need to wait awhile for results)
  - Kraken: https://ccb.jhu.edu/software/kraken/
    - Fastest binning approach to date and very accurate.
    - Large computing requirements (e.g. >128GB RAM)

Metagenomic Assembly

- “MetaSPAdes showed the overall best assembly size statistics while also capturing a relatively large fraction of the expected diversity. The usage of this tool is relatively simple and convenient, being basically identical to that of SPAdes, and largely flexible regarding the format of the input data. A drawback may be the reduced sensitivity for micro diversity. However, for the majority of metagenome research questions, accurate and representative consensus genomes of species should be more than sufficient. ”