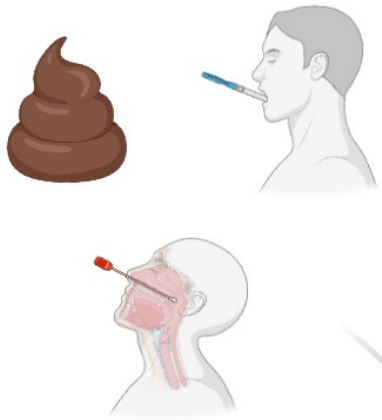
Microbiome workshop – data generation and processing

September 2024

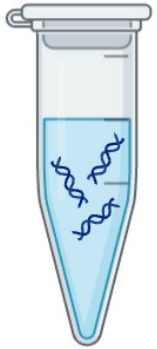
IKMB
Institute of Clinical
Molecular Biology

Microbiome science - from sample to data



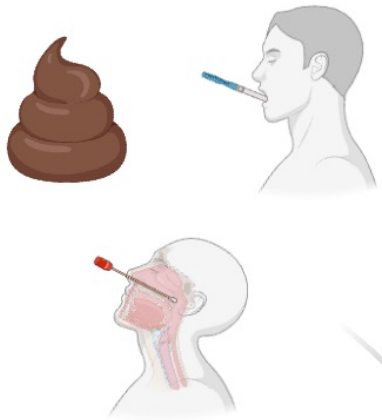
Sample collection & storage

DNA/RNA extraction



Data generation & processing

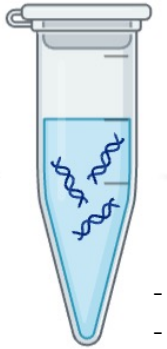
Microbiome science - from sample to data



Sample collection & storage

- Sampling method (swab, tube, …)
- Sampling buffer (RNAlater, glycerol, …)
- Storage conditions (snap freezing, -20/-80°C)

DNA/RNA extraction



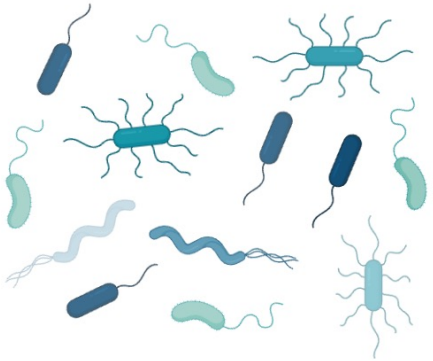
- Extraction kit used
- DNA or RNA or both?
- Cell disruption (Proteinases, Bead beating)

Data generation & processing

- 16S rRNA gene or Shotgun metagenomics
- Short reads or long reads
- Data annotation

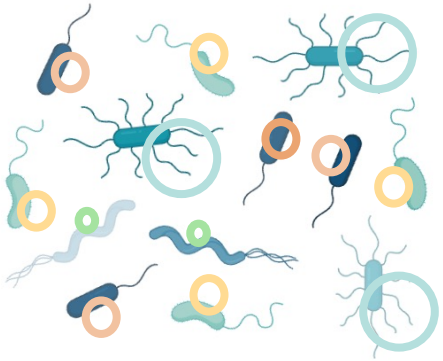
16S rRNA gene
amplicon sequencing

shotgun
metagenomics



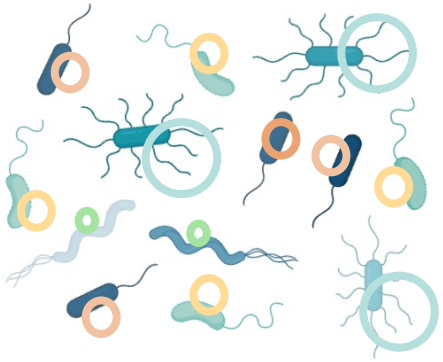
16S rRNA gene
amplicon sequencing

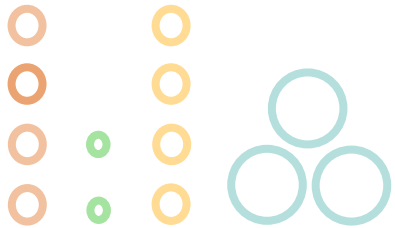
shotgun
metagenomics



16S rRNA gene
amplicon sequencing

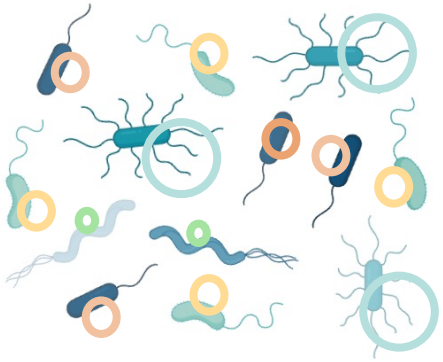
shotgun
metagenomics





16S rRNA gene amplicon sequencing

shotgun metagenomics

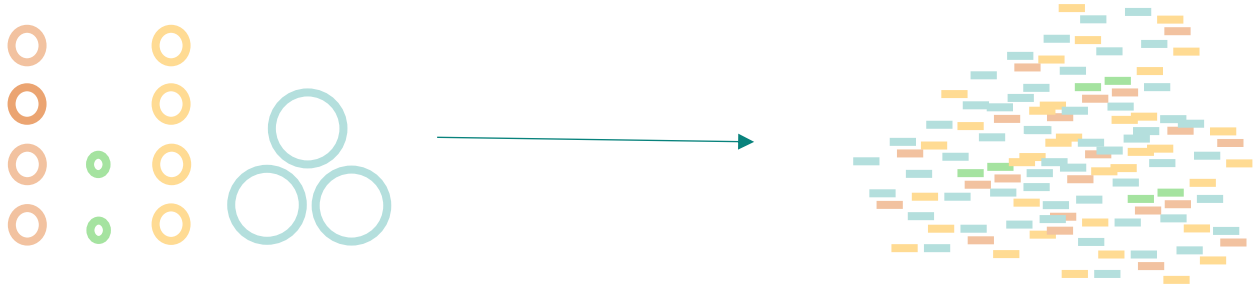


Full genome information + + +

More data + / -

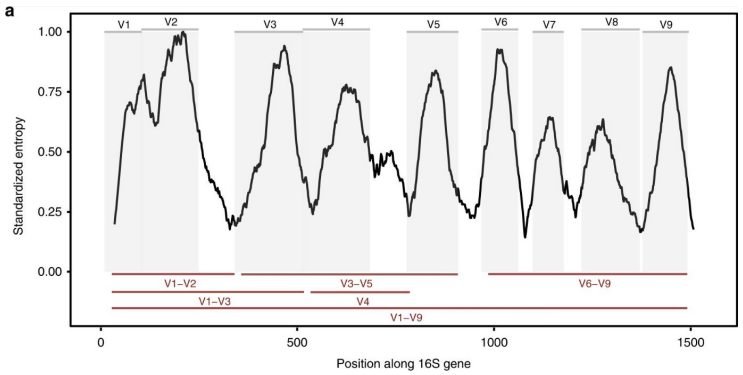
More expensive - -

Bias towards high abundant taxa -



Johnson *et al.* 2019

16S rRNA gene amplicon sequencing



16S rRNA gene: conserved and variable regions

PCR target + Molecular clock
= taxon information

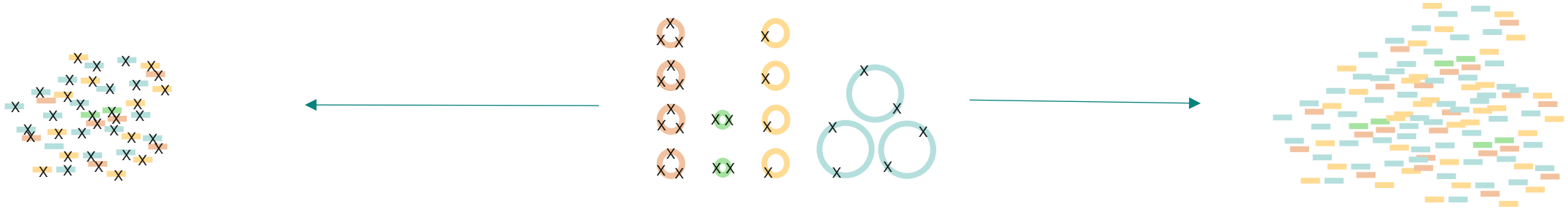
shotgun metagenomics

Full genome information + + +

More data + / -

More expensive - -

Bias towards high abundant taxa -



Johnson *et al.* 2019

What we have: The FastQ format

Identifier — @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1

Sequence — TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA

+ sign & identifier — +HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1

Quality scores — efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`

Base T

phred Quality] = 29

https://compgenomr.github.io/book/fasta-and-fastq-formats.html

What we have: The FastQ format

Identifier — `@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

Sequence — `TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA`

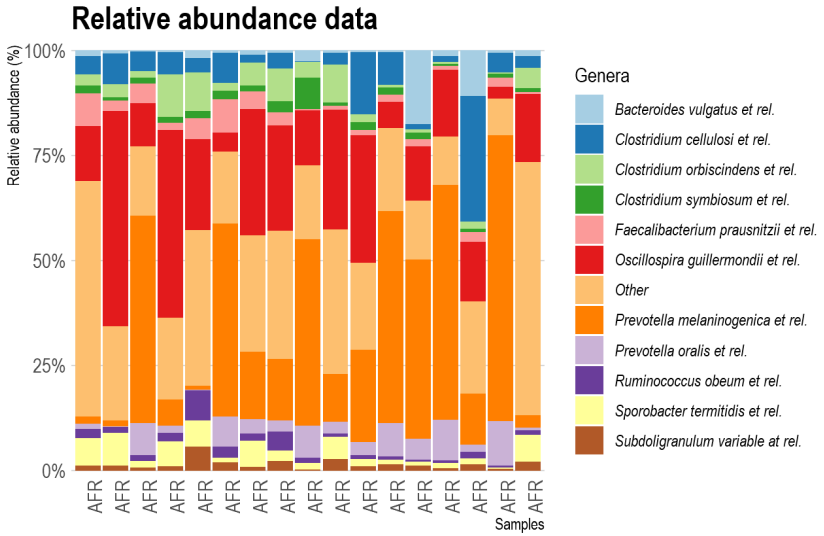
+ sign & identifier — `+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

Quality scores — `efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd``

Base T

phred Quality] = 29

What we want:



https://compgenomr.github.io/book/fasta-and-fastq-formats.html

What we have: The FastQ format

Identifier — `@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

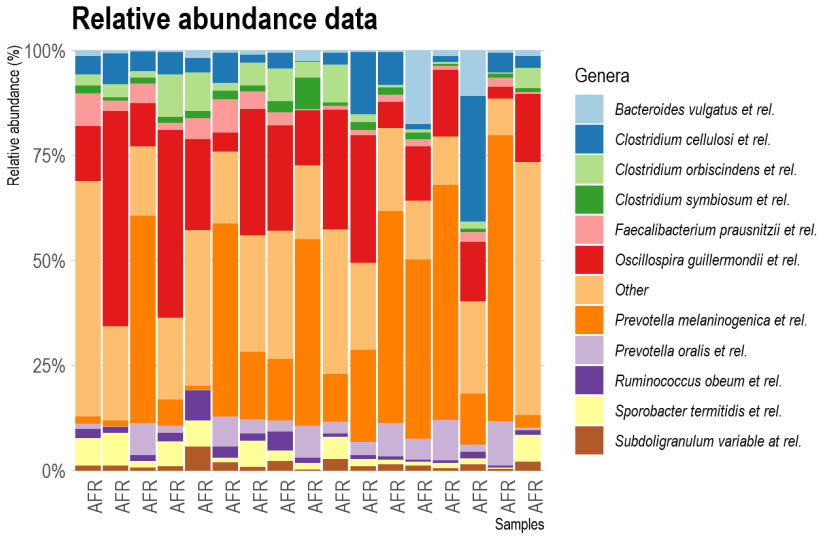
Sequence — `TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGA`

+ sign & identifier — `+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`

Quality scores — `efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd``

Base T
phred Quality] = 29

What we want:



How we get there:

qiime2

Bolyen *et al.* 2019

DADA2

Callahan *et al.* 2019

https://compgenomr.github.io/book/fasta-and-fastq-formats.html

16S data workflow with

qiime2docs

Version: 2024.5

Table of Contents

Quick search

Go

QIIME 2 user documentation

This site is the official user documentation for QIIME™ 2, including installation instructions, tutorials, and other important information. Visit http://qiime.org for information on QIIME™ 1.

Getting started

Check out the getting started guide to begin using QIIME 2.

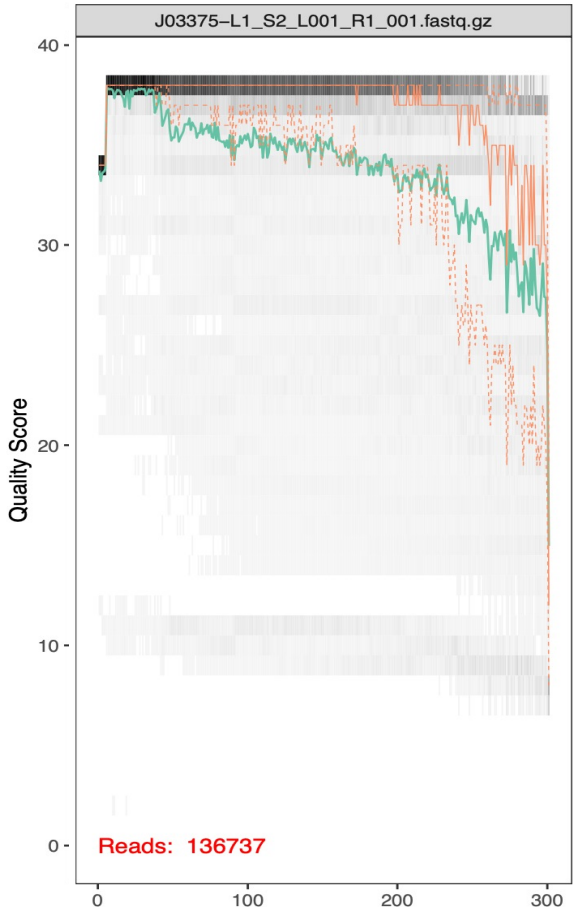
Table of contents

16S data workflow with qiime2 + DADA2

- Tutorials
  - Overview of QIIME 2 Plugin Workflows
  - QIIME 2 for Experienced Microbiome Researchers
  - “Moving Pictures” tutorial
  - “Moving Pictures” tutorial - Multiple Interface Edition
  - Fecal microbiota transplant (FMT) study: an exercise
  - “Atacama soil microbiome” tutorial
  - Parkinson’s Mouse Tutorial
  - Importing data
  - Exporting data
  - Metadata in QIIME 2
  - Filtering data
  - Training feature classifiers with q2-feature-classifier
  - Evaluating and controlling data quality with q2-quality-control
  - Predicting sample metadata values with q2-sample-classifier
  - Performing longitudinal and paired sample comparisons with q2-longitudinal
  - Identifying and filtering chimeric feature sequences with q2-vsearch
  - Alternative methods of read-joining in QIIME 2
  - Clustering sequences into OTUs using q2-vsearch
  - Utilities in QIIME 2
  - Phylogenetic inference with q2-phylogeny

16S data workflow with qiime2 + DADA2

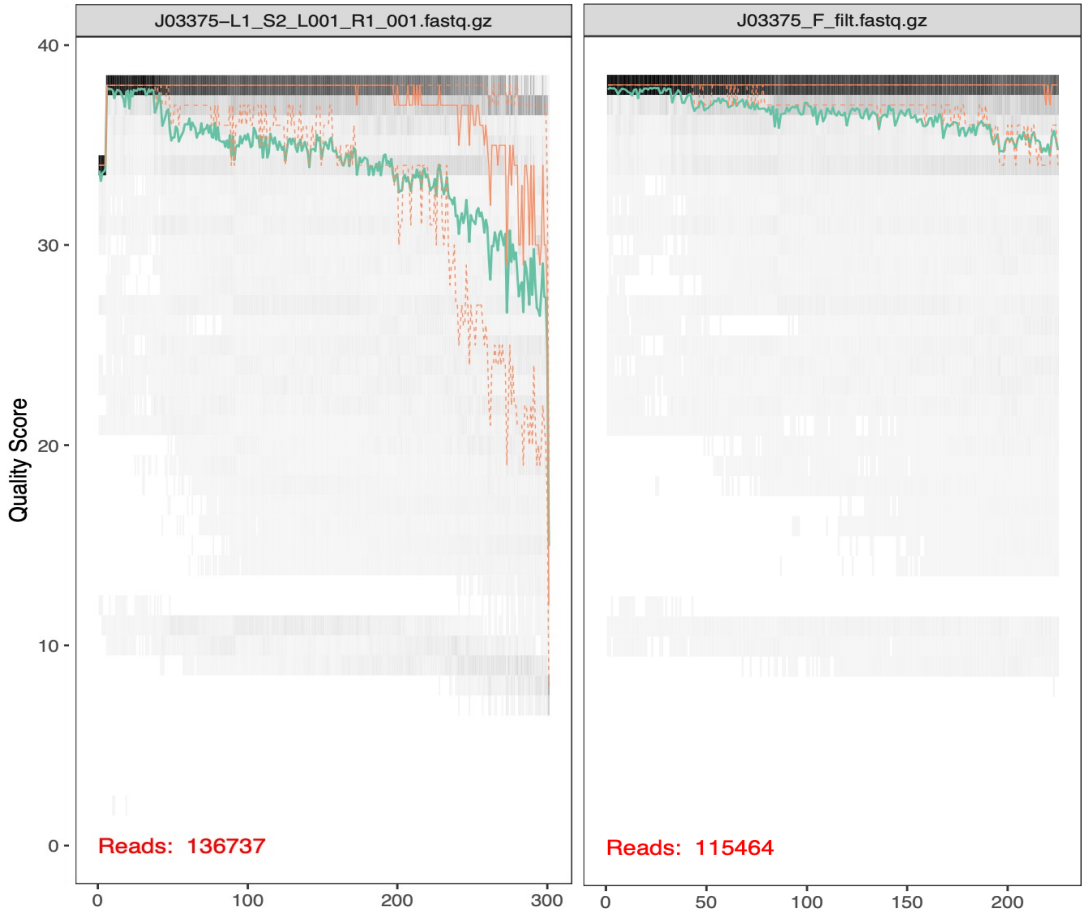
Step 1: QC and trimming



Raw

16S data workflow with qiime2 + DADA2 (simplified)

Step 1: QC and trimming



Raw

After QC

16S data workflow with qiime2 + DADA2 (simplified)

Step 1: QC and trimming

Raw

After QC

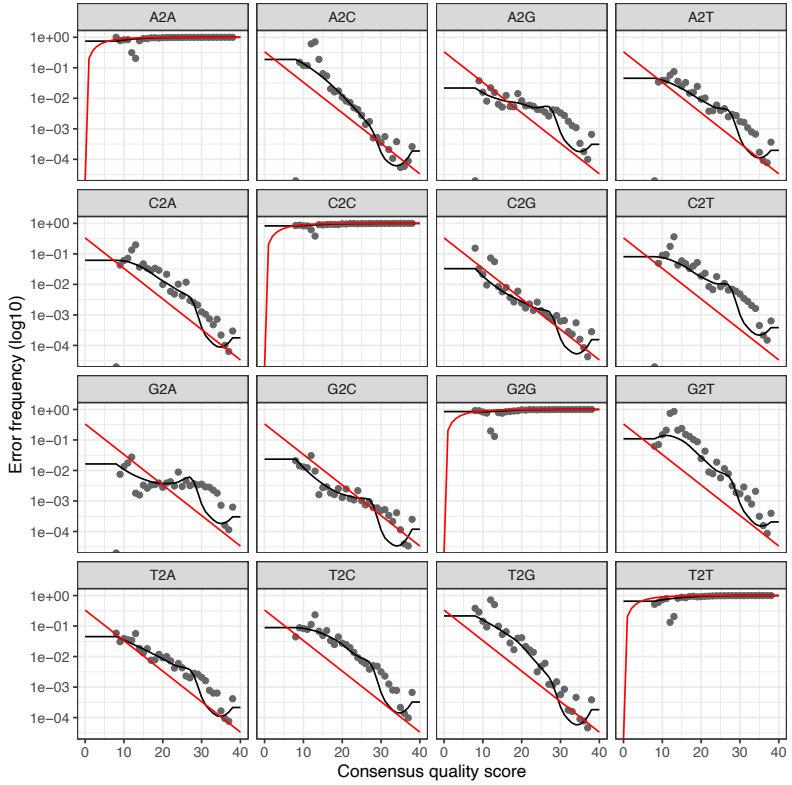
16S data workflow with qiime2 + DADA2 (simplified)

Step 1: QC and trimming

Raw

After QC

Step 2: Error correction



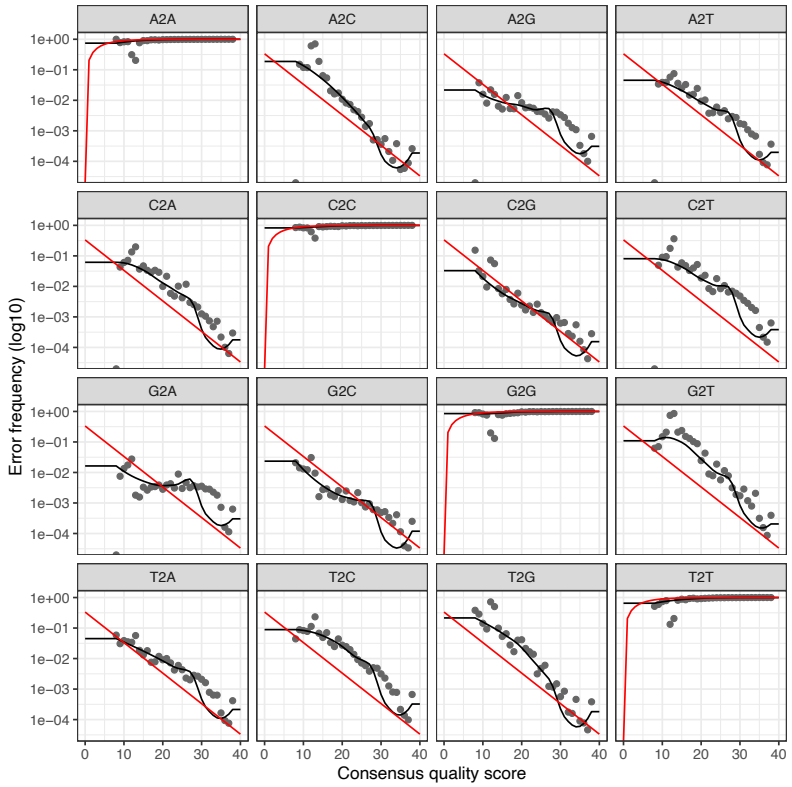
16S data workflow with qiime2 + DADA2 (simplified)

Step 1: QC and trimming

Raw

After QC

Step 2: Error correction



Step 3:
Inference of Amplicon Sequence Variants (ASVs)

Assigning taxonomic labels

Abundance Table

	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3

ASV sequences:

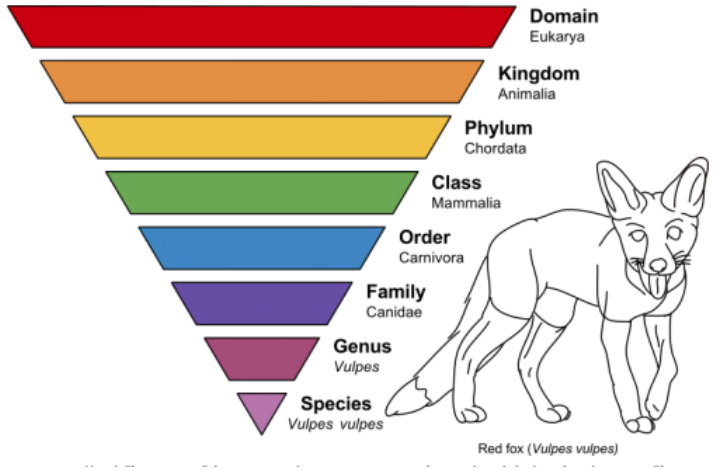
```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

Assigning taxonomic labels

What is a taxonomy?

Abundance Table

	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3



ASV sequences:

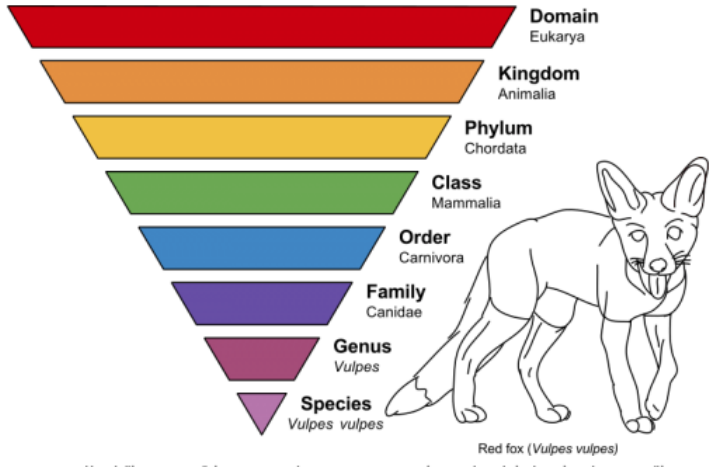
```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

Assigning taxonomic labels

What is a taxonomy?

Abundance Table

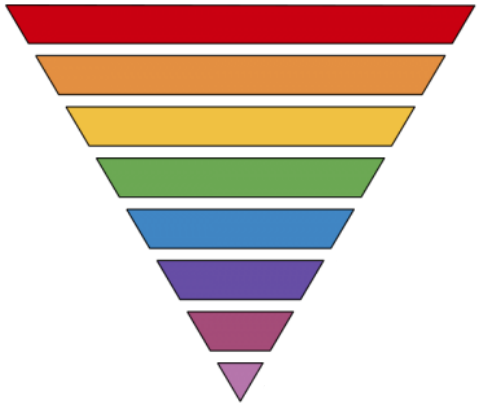
	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3



ASV sequences:

```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

Same can be done for microorganisms:



Domain	Bacteria
Kingdom	Bacteria
Phylum	Bacillota A
Class	Clostridia
Order	Oscillospirales
Family	Ruminococcaceae
Genus	Faecalibacterium
Species	Faecalibacterium prausnitzii

Assigning taxonomic labels

Abundance Table

	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3

ASV sequences:

```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

Database search
(Naïve Bayesian Classifier)

Assigning taxonomic labels

Abundance Table

	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3

ASV sequences:

```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

Greengenes2

silva high quality ribosomal RNA databases

GTDB TK

Ribosomal database project

Database search
(Naïve Bayesian Classifier)

Assigning taxonomic labels

Greengenes2

silva
high quality ribosomal RNA databases

GTDB TK

Ribosomal database project

Abundance Table

	ASV1	ASV2	ASV3	ASV4
S1	5	10	3	7
S2	5	0	3	0
S3	0	0	4	3

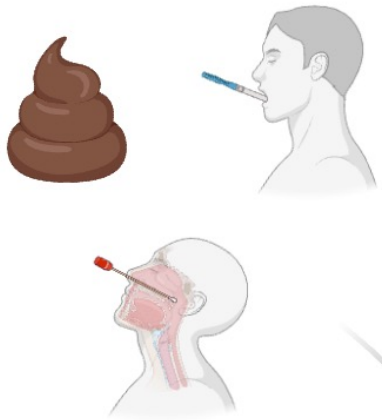
Database search
(Naïve Bayesian Classifier)

ASV sequences:

```
>ASV1
ACGCTGGCGGCATGCTTAACACATGCAAGTCGTACGAATGAAT…
>ASV2
ACGCTGGCGGTATGCCTAACACATGCAAGTCGAACGAGGTAGC…
>ASV3
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAACAG…
>ASV4
ACGCTAGCGGCAGGCTTAACACATGCAAGTCGAGGGGTAGTCA…
```

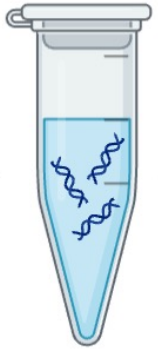
	Domain	Phylum	Class	Order	Family	Genus
ASV1	Bacteria	Bacillota_C	Negativicutes	Veillonellales	Dialisteraceae	Dialister
ASV2	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
ASV3	Bacteria	Pseudomonadota	Gammaproteo bacteria	Enterobacterales	Enterbacteriaceae	
ASV4	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Prevotella

Microbiome science - from sample to data



Sample collection & storage

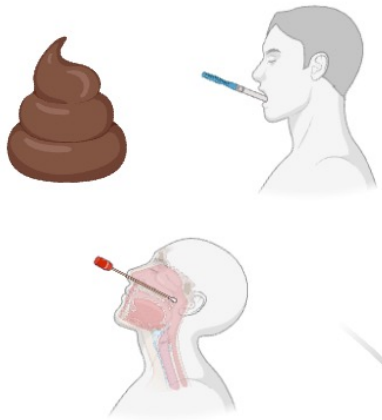
DNA/RNA extraction



Data generation & processing

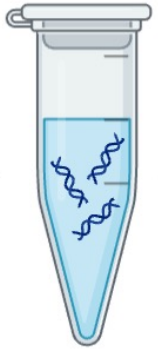
Data analysis

Microbiome science - from sample to data to analysis



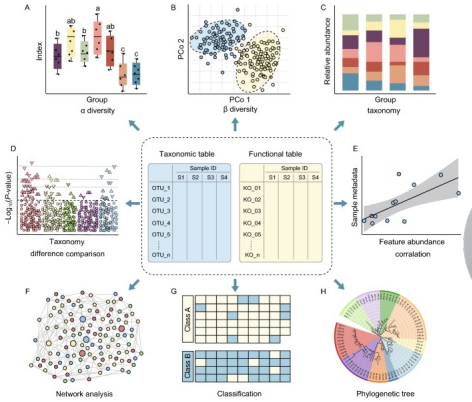
Sample collection & storage

DNA/RNA extraction



Data generation & processing

Data analysis



RStudio

AlphaDiversity.R | Untitled3* | Intro_R.R | 1_IgA_CoatingIndexCalculation_OT

```r
## My microbiome analyses

library(phyloseq)

setwd("/Users/apschaan/Documents/Kiel/UKSH-CAU/Classes/KMC_workshop")

taxTable <- read.csv("taxonomy_gmbc1_form.csv", header=TRUE, row.names=1)

head(taxTable)
```

11:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.3.1 · ~/Documents/Kiel/UKSH-CAU/Classes/KMC_workshop/

```
> library(phyloseq)
> setwd("/Users/apschaan/Documents/Kiel/UKSH-CAU/Classes/KMC_workshop")
> taxTable <- read.csv("taxonomy_gmbc1_form.csv", header=TRUE, row.names=1)
> head(taxTable)
                                  Kingdom            Phylum
675c847bccbc53942ebb7b8cbb4efc4d d__Bacteria    p__Bacteroidota
da6a7cba87e0895b9cbe6037b9bd8b3b d__Bacteria    p__Bacteroidota
```

Environment | History | Connections

Import Dataset · 994 MiB · List

R · Global Environment · Taxa

Data

TaxaForHeatm…	40 obs. of 12 variables
TaxaForHeatm…	20 obs. of 2 variables

Values

taxa	"a361d6b053ce9d38182128a316635e3e"
taxa_names	chr [1:197] "2e63e8e3f9f0f102f570e7…
taxaKeep	chr [1:2174] "9492bc05eee3b253c1557…

Values

taxa	"a361d6b053ce9d38182128a316635e3e"
taxa_names	chr [1:197] "2e63e8e3f9f0f102f570e7…
taxaKeep	chr [1:2174] "9492bc05eee3b253c1557…

Files | Plots | Packages | Help | Viewer | Presentation

Zoom · Export