Workshop Tutorial Web Address (URL):
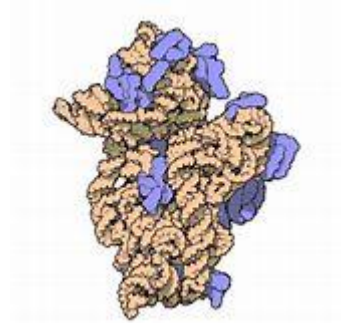
https://tsute.github.io/resispart

type the blue part into your browser

Meta-genomics

| Scope	Question	Answer
One gene – 16S rRNA	Who are there?	Genus, Species, OTUs
All genes – metagenomics (DNA)	What can they do?	Functional potentials
All genes – metatranscriptomic (RNA)	What are they doing?	Metabolic pathways Biological function

What is 16S rRNA?

- 16S ribosomal RNA - the component of the 30S small subunit of a prokaryotic ribosome that binds to the Shine-Dalgarno sequence during protein synthesis
- Approximately 1,500 nt in length
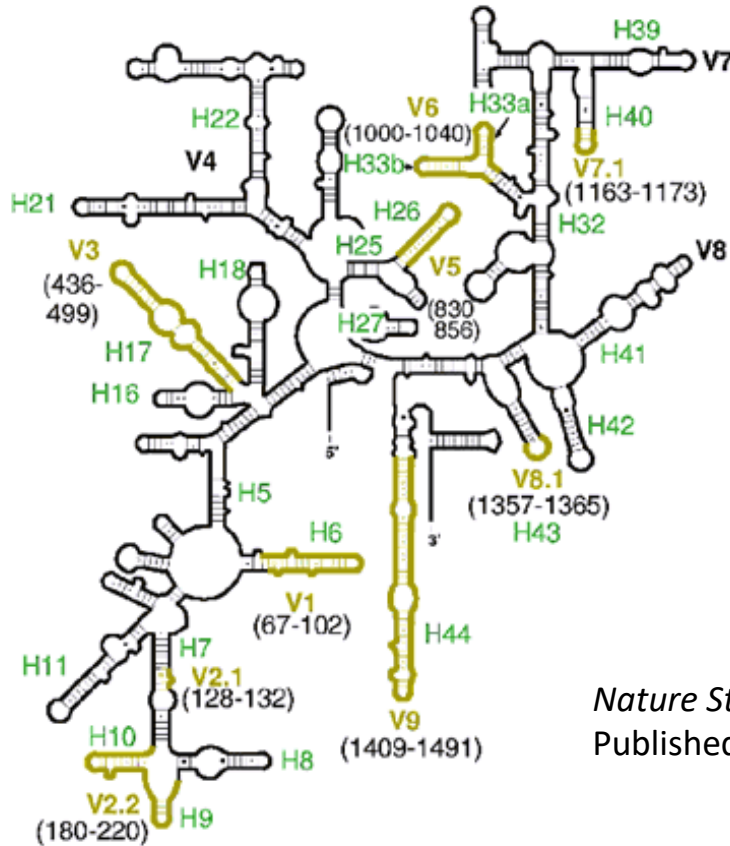
Prokaryotic 16S rRNA

E. coli

Nature Structural Biology **9**, 750 - 755 (2002)
Published online: 16 September 2002; | doi:10.1038/nsb841

CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

Alimetrics Ltd
Koskelontie 19 B
02920 Espoo
Finland

Why 16S rRNA as the marker gene?

- SIMILAR enough – both Bacteria and Archaea have it

- DIFFERENT enough – to differential different species

- SHORT enough – easy (easier) to sequence

- LONG enough – to contain both conserve and variable domains

The Bioinformatics Black Box
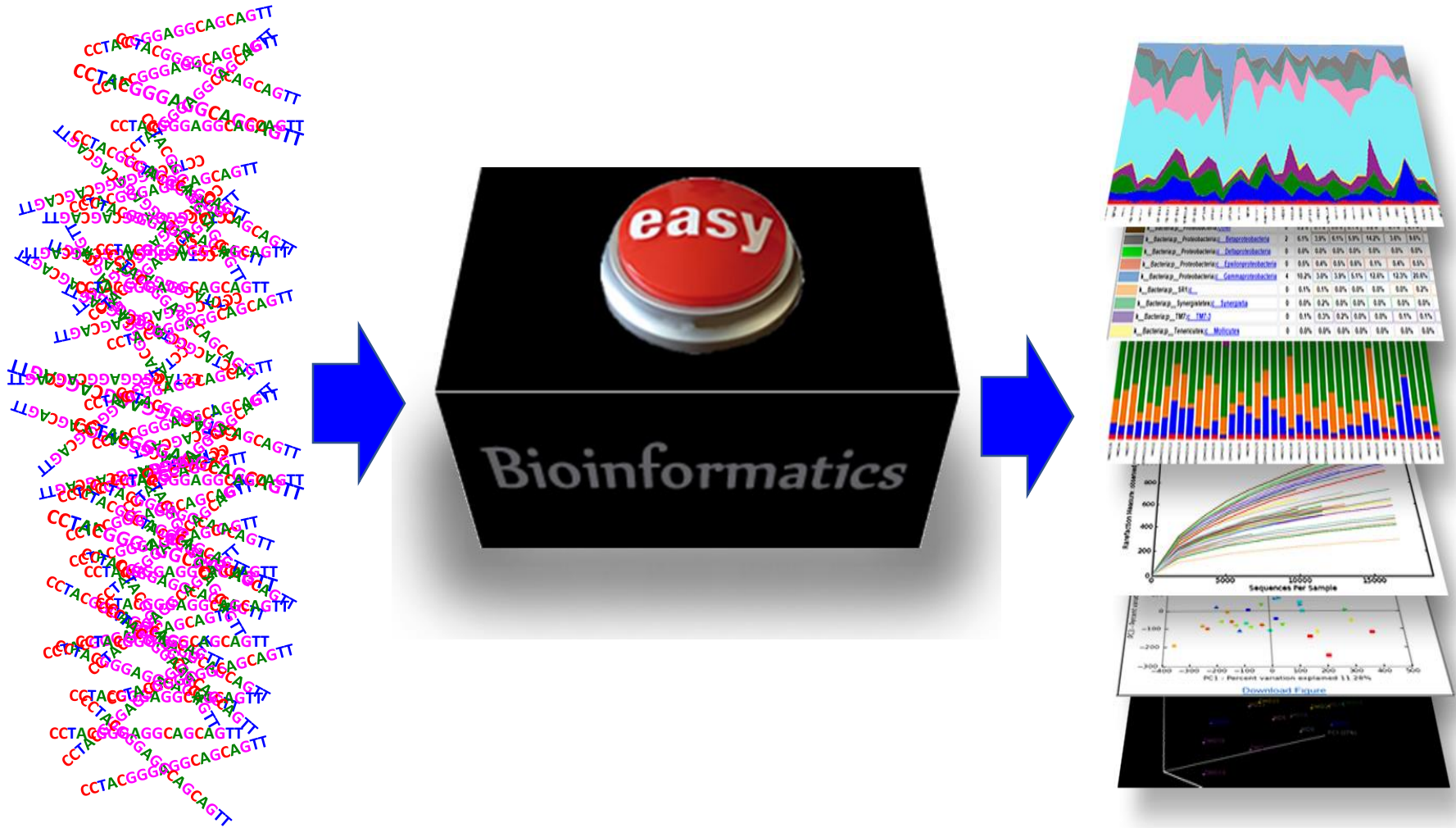
Inside the blackbox

1. **Upstream analyses** – quality control, separate samples (de-barcoding, de-multiplexing)
2. **Middle tier analyses** – organize sequences into biological groups (OTUs, ASVs, taxa)
3. **Downstream analyses** – compare microbial compositions of the samples – diversity analysis

Chaos organizing

OTU (operational taxonomic unit) picking – based on % sequence identity

TTTACTGTTTTGCATTAGTA

OTU1
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
TCTACGGGAGACAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTG
CCTACGAGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC

TTTACTGTTTTGCATTAGTA

OTU2
TTTACTGTTTTGCATTAGTA
TGTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGTATTAGTG
TTTACTGTTGTGCATTAGTA
TGTACTGTTTTGCATTAGTA

CATCCGCCACAAAGCAGTT

OTU4
CCTACGGGAGACAGCAGTG
CCTACGAGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACAGGAGACAGCAGTG

OTU3
GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCT

TTTACTGTTTTGCATTAGTA

TTTACTGTTTTGCATTAGTA

OTU5
CATAAGCCACAAAGCAGTT
CATACGCCACAACGCAGTT
CATAAGCCACAAAGCAGTT

OTU6
GGTACGGGAGCAGCAGTCA
GGTACGAGAGCAGCAGTCA
GGTACGCGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACAGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA

OTU7
CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGAACAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGACCAGGCAGCAGCC

CCTACGGGAGGCAGCAGTT

CCTACGGGAGGCAGCAGTT

CCTACGGGAGGCAGCAGTT

TTTACTGTTTTGCATTAGTA

TTTACTGTTTTGCATTAGTA

OTU10
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT

OTU8
TTTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TGTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA

OTU9
CCTACGGGAGGCAGCAGTT
CCTACGGGAGACAGCAGTT
CCTACACGAGGCAGCAGTA
CCTACGGGAGGCAGCAGTT
CCTACGGGAGGCAGCAGTT
CCTACAGGAGGCAGCAGTT

TTTACTGTTTTGCATTAGTA

OTU11
CATCCGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATCAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGCC

ASV (amplicon sequence variants) clustering – denoise sequences based on sequence quality error model

ASV1
```
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
TCTACGGGAGACAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTG
CCTACGAGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
```

ASV2
```
TTTACTGTTTTGCATTAGTA
TGTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGTATTAGTG
TTTACTGTTGTGCATTAGTA
TGTACTGTTTTGCATTAGTA
```

ASV4
```
CCTACGGGAGACAGCAGTG
CCTACGAGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACAGGAGACAGCAGTG
```

ASV3
```
GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCT
```

ASV5
```
CATAAGCCACAAAGCAGTT
CATACGCCACAACGCAGTT
CATAAGCCACAAAGCAGTT
```

ASV6
```
GGTACGGGAGCAGCAGTCA
GGTACGAGAGCAGCAGTCA
GGTACGCGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACAGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
```

ASV7
```
CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGAACAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGACCAGGCAGCAGCC
```

ASV10
```
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
```

ASV8
```
TTTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TGTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
```

ASV9
```
CCTACGGGAGGCAGCAGTT
CCTACGGGAGACAGCAGTT
CCTACGGGAGGCAGCAGTT
CCTACGGGAGGCAGCAGTT
CCTACGGGAGGCAGCAGTT
CCTACAGGAGGCAGCAGTT
```

ASV11
```
CATCCGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATCAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGCC
```

DADA2 Pipeline

Identify representative sequences

Taxonomy assignment

Fusobacterium nucleatum

CCTACGGGAGACAGCAGTG
CCTACGAGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACGGGAGACAGCAGTG
CCTACAGGAGACAGCAGTG

Streptococcus mitis

TTTACTGTTTTGCATTAGTA
TGTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGTATTAGTG
TTTACTGTTGTGCATTAGTA
TGTACTGTTTTGCATTAGTA

Lactobacillus reuteri

CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
TCTACGGGAGACAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTG
CCTACGAGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC
CCTACGGGAGGCAGCAGTC

Streptococcus mutans

GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACTGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
TGTACGGGAGCAGCAGTCT

Veillonella parvula

CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGAACAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGCCCAGGCAGCAGCC
CCTACGACCAGGCAGCAGCC

Streptococcus parasanguinis

GGTACGGGAGCAGCAGTCA
GGTACGAGAGCAGCAGTCA
GGTACGCGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA
GGTACAGGAGCAGCAGTCA
GGTACGGGAGCAGCAGTCA

Selenomonas artemidis

CATAAGCCACAAAGCAGTT
CATACGCCACAACGCAGTT
CATAAGCCACAAAGCAGTT

Aggregatibacter actinomycetemcomitans

ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT
ACTAAGAGAGGCAGCAGTT

Streptococcus parasanguinis

TTTACTGTTTTGCATTAGTA
TTTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TGTACTGTTGTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA
TTTACTGTTTTGCATTAGTA

Porphyromonas gingivalis

CCTACGGGAGGCAGCAGTT
CCTACGGGAGACAGCAGTT
CCTACACGAGGCAGCAGTA
CCTACGGGAGGCAGCAGTT
CCTACGGGAGGCAGCAGTT
CCTACAGGAGGCAGCAGTT

Lactobacillus plantarum

CATCCGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGTT
CATCAGCCACAAAGCAGTT
CATAAGCCACAAAGCAGCC

Downstream analyses

- **Alpha diversity** – analyze the complexity of the microbial community in each sample

- **Beta diversity** – compare the complexity of different groups of samples (e.g., health vs disease)

- **Differential analysis** – identify organisms with different relative abundance in different groups

- **Functional inferring** – associate functions with species identified

Workshop Objectives

- I know R
- I have used R
- I have used R to analyze NGS data
- I like R

Methodology

- Mostly copy and paste

- Provide a comprehensive online tutorial

- There are more tutorials out there (Web, Youtube)

- Google is our good friend