Enrichment and propagation of metagenomic experimental metadata

BioHackathon 2018 Paris

Ola Tarkowska & Maxim Scheremetjew

Senior Software Engineers

EMBL-EBI

EMBL-EBI

MGnify BioHackathon Dev Team

Maxim Scheremetjew

Ola Tarkowska

Miguel Boland

Motivation

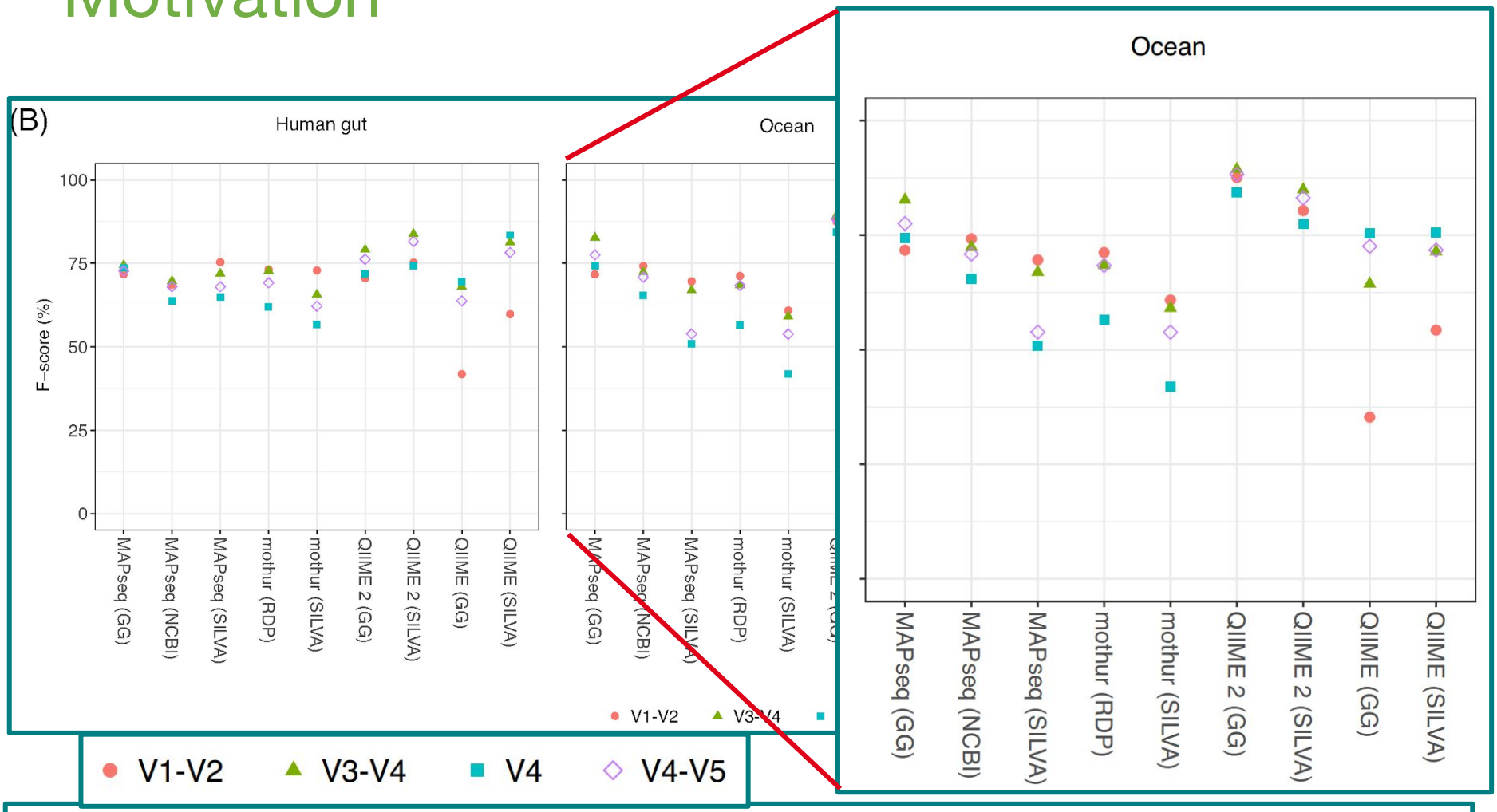
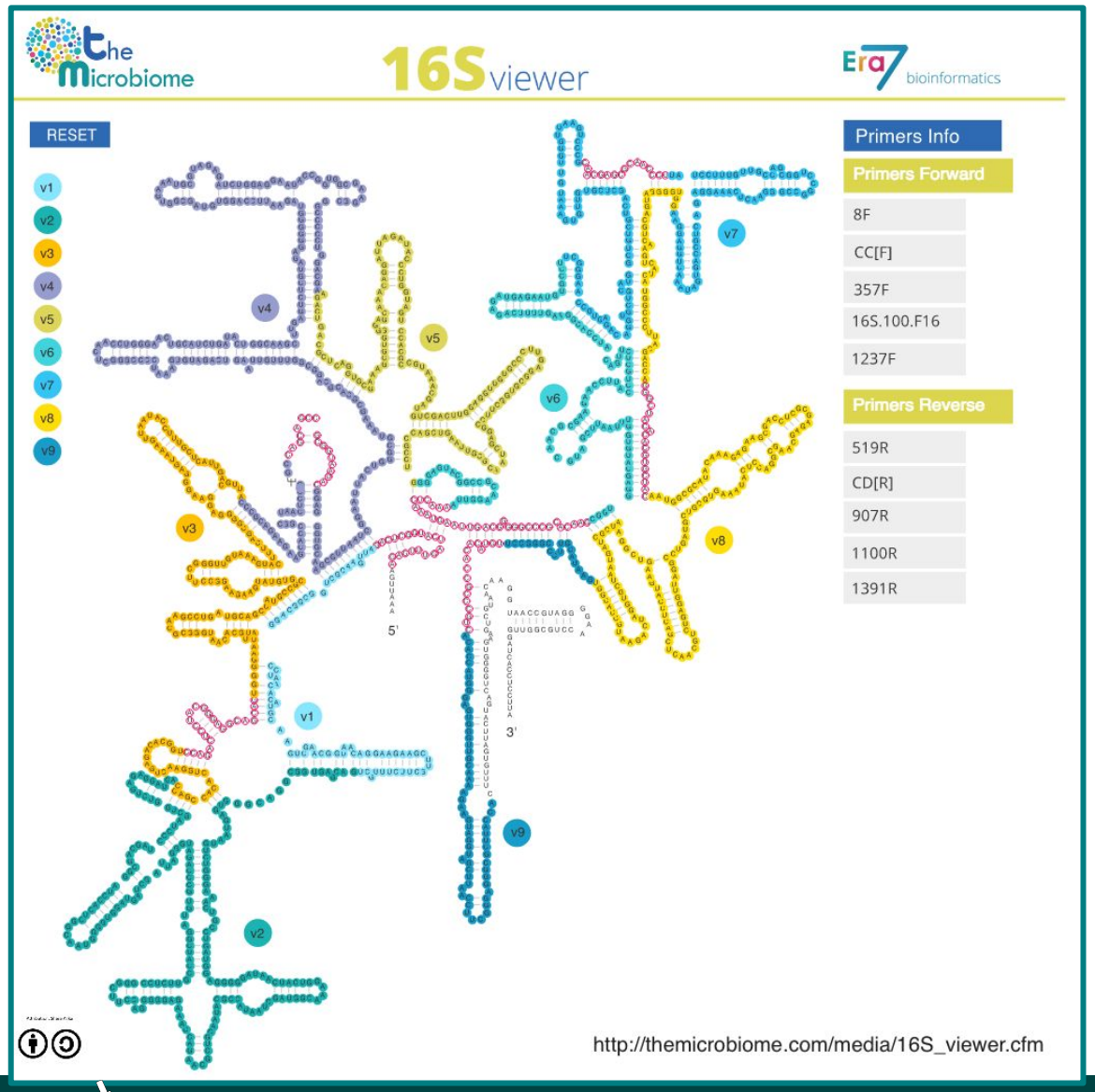


Fig. B: *F-scores calculated for some of the most commonly tested sub-regions of the 16S rRNA gene: V1-V2, V3-V4, V4, and V4-V5.*

Variable regions of the 16S ribosomal RNA of *E.coli*



- bacterial 16S ribosomal RNA genes contain nine “hypervariable regions” (V1 – V9)
- We’ve chosen E.coli as a representative for Bacteria -> well described

Yang B, Wang Y, Qian PY
BMC Bioinformatics 2016
doi: 10.1186/s12859-016-0992-y

Goal

1. New **tool** to infer amplified SSU rRNA variable region, based on sequence analysis by MGnify
   - **statistical analyses** and **machine learning algorithms** to process analysis results from MGnify resource

2. **Pipeline** to push this data to the various archiving resources, such as ENA and BioSamples.
   - representatives of the relevant resources

Who can help us?

- Statistical Data Analysts
- Machine Learning Experts