Pipelines

1 Pipeline for genome decontamination (DeCon)

1.1 Introduction

Pipeline DeCon is designed to retrieve genomic sequences of target **phylum** from metagenomic assembly of paired next generation sequencing (NGS) reads. First, NGS reads are mapped to assembly by minimap2 (Li 2018), generating BAM file. Second, SprayNPray (Garber et al. 2022) is used to compute coverage, GC content and coding density of each contigs. Third, all contigs are searched against non-redundant (nr) database by DIAMOND (Buchfink et al. 2015) and assigned to phyla by MEGAN (Huson et al. 2007). Forth, contigs below 400 base pair (bp) are removed. Then a decision tree classifier is trained, taking coverage, GC content and coding density as training features and phylum assignment as target value. This classifier is used to compute phylum assignment of contigs that DIAMOND and MEGAN failed to compute assignments. Fifth, contigs assigned to the target phylum are retrieved. QUAST (Gurevich et al. 2013) and BUSCO (Simão et al. 2015) are used to evaluate retrieved genome. Distributions of contig coverage and GC content of retrieved genome are plotted.

1.2 Dependencies

Softwares

R

Python

minimap2

SAMtools

SprayNPray

DIAMOND

MEGAN (blast2rma rma2info scripts)

seqkit

QUAST

BUSCO

Databases

DIAMOND database (nr)

MEGAN database

BUSCO database

Python modules

numpy

pandas

scikit-learn

R packages

reticulate

stringr

ggplot2

ggExtra

1.3 Usage

Modify configuration file (templated as DeCon.conf), and run

```
Rscript path/DeCon_pipeline.R path/DeCon_main.R path/DeCon_main.py path/DeCon.conf
```

2 Pipeline for calling protein-coding genes from genome (ProtGeneCall)

References

Buchfink, Benjamin et al. (2015). "Fast and sensitive protein alignment using DIAMOND". In: *Nature methods* 12.1, pp. 59–60.

Garber, Arkadiy I et al. (2022). "SprayNPray: user-friendly taxonomic profiling of genome and metagenome contigs". In: *BMC genomics* 23.1, p. 202.

Gurevich, Alexey et al. (2013). "QUAST: quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075.

Huson, Daniel H et al. (2007). “MEGAN analysis of metagenomic data”. In: *Genome research* 17.3, pp. 377–386.

Li, Heng (2018). “Minimap2: pairwise alignment for nucleotide sequences”. In: *Bioinformatics* 34.18, pp. 3094–3100.

Simão, Felipe A et al. (2015). “BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs”. In: *Bioinformatics* 31.19, pp. 3210–3212.