RD-Analyzer

In silico region of difference (RD) analysis of *Mycobacterium tuberculosis* complex from sequence reads

Files

RD-Analyzer.py: the standard RD-Analyzer used for deletion prediction of previously defined RD markers and strain identification of *Mycobacterium tuberculosis* complex based on these markers.
RD-Analyzer-extended.py: the extended RD-Analyzer used for deletion prediction of user-specified RD sequences.
Reference/RDs30.fasta: sequences of previously defined RD markers used in the standard RD-Analyzer.
Reference/Lineage4.fasta: sequences of potential Lineage 4 markers identified in the manuscript.

Prerequisite

Python 2.7

Standard RD-Analyzer

Usage:

```
python2.7 RD-Analyzer.py [options] FASTQ_1 FASTQ_2(optional)
```

Options:

```
    --version           show program version number and exit
    -h, --help          show this help message and exit
    -d, --debug         enable debug mode, keeping all intermediate files

    -O OUTDIR, --outdir=OUTDIR
          output directory [Default: running directory]
    -o OUTPUT, --output=OUTPUT
          basename of output files [Default: RD-Analyzer]

    -p, --personalized
          use personalized cut-offs
    -m MIN, --min=MIN
          read depth cut-off (in the unit of average depth, 0-1), used when '-p' is
    set
     -c COVERAGE, --coverage=COVERAGE
          sequence coverage cut-off (0-1), used when '-p' is set
```

Suggestions:

Users are suggested to use the default cut-offs which are optimized by us.

Extended RD-Analyzer

Usage:

```
python2.7 RD-Analyzer-extended.py [options] REF.FASTA FASTQ_1 FASTQ_2(optional)
```

Options:

```
    --version           show program version number and exit
    -h, --help          show this help message and exit
    -d, --debug         enable debug mode, keeping all intermediate files

    -O OUTDIR, --outdir=OUTDIR
          output directory [Default: running directory]
    -o OUTPUT, --output=OUTPUT
          basename of output files [Default: RD-Analyzer]
```

Input files:

REF.FASTA - Reference sequences used should be in a fasta file with the header lines prepared as below:

- Four fields are reuiqred in the header line, which should be separated with '-'.
- Field one: reference sequence name
- Filed two: read-depth cutoff to be used (in the unit of average depth, in a 0-1 scale). Specify 'default' if want to use default parameters (0.09)
- Field three: sequence coverage cut-off (in a 0-1 scale). Specify 'default' if want to use default parameters (0.5)
- Flled four: descriptive information of the RD to be shown if the RD is detected.
- An example header line: **>Lineage4.6.1.2/1-default-default-Lineage4.6.1.2/1**
- **Notice:** 1. Don't include space in the header file. 2. Don't use '-' unless as filed deliminator