ReRCoP

Recombination Removal for Core-genome Phylogeny

Prerequisites

1. Linux environment
2. Python 2.7
3. Python package: ‘scipy’
4. R (if removal plot is needed)
5. BLAST package (makeblastdb, blastn)

Usage

```
python ReRCoP.py [options] Genomes.fasta

Options:
    --version           show program version number and exit
    -h, --help          show this help message and exit

  Input Options:
    -a, --aligned       Set this if genome sequences in the input file are
                        already aligned.
                        # For aligned genomes

    --gbk=GBK           Input GenBank file of the reference genome.
                        # For aligned genomes using core genome approach

    -w, --window        Set this if sliding windows instead of genes are to be
                        considered.
                        # For aligned genomes using complete genome approach

    --fSize=FSIZE       Fragment size if using sliding window. [Default: 1000]
                        # For aligned genomes using complete genome approach

    --sSize=SSIZE       Step size if using sliding window. [Default: 500]
                        # For aligned genomes using complete genome approach

    --cds=CDS           Input coding sequences in fasta format. Used to determine
                        the core genome when input genomes are not aligned.
                        # For unaligned genomes using core genome approach

  Core Gene Identification Options:
    --cov=COV           Minimum sequence coverage to regard genes as present.
                        [Default: 0.7]
                        # For aligned or unaligned genomes using core genome
                        # approach

    --sim=SIM           Minimum sequence similarity to regard genes as present.
                        [Default: 70]
```

```
                        # For unaligned genomes using core genome approach

    Outlier Removal Options:
      -m METHOD, --method=METHOD
                        Outlier removal method. Can be 'Grubbs', 'kNN', or
                        'DBSCAN', or can be multiple methods separated by ','.

      --alpha=ALPHA     For 'Grubbs' method: Significance level in Grubbs test.
                        [Default: 0.05]

      --radius=RADIUS   For 'kNN' method: Maximum number of differences for
                        a point to be considered as a neighbor (in the unit of
                        standard deviation of all pair-wise nubmer of differences
                        ). [Default: 1.5)]

      --k=K             For 'kNN' method: Minimum number of neighbors for a
                        non-outlier point (in the unit of total number of points
                        ). [Default: 0.2]

      --eps=EPS         For 'DBSCAN' method: Maximum number of differences
                        between two points for them to be considered as in the
                        same neighborhood (in the unit of standard deviation of
                        all pair-wise nubmer of differences). [Default: 1]

      --minP=MINP       For 'DBSCAN' method: Minimum number of
                        points required to form a dense region (in the unit of
                        total number of points). [Default: 0.2]

    Output Options:
      -o OUTDIR, --outdir=OUTDIR    Output directory. [Default: running directory]
      -p PREFIX, --prefix=PREFIX    Output prefix. [Default: ReRCoP]
```

Input files

Input files for aligned genomes are straightforward and thus not stated agtain here. This part wil be focused on unaligned genomes, where core genomes are to be identified and extracted by ReRCoP.
The following input files are required:

1. A file in multiple nucleotide fasta format with gene coding sequences (All the coding sequences from any one of the samples will do).
    - If one of the input is complete genome with annotation from the public database:
      1> Download coding sequence in multiple nucleotide fasta format.
      2> Remove duplicated genes.
      3> Remove phage genes.
      4> Remove genes with CRISPR sequence.
    - Else if one of the input is complete genome without annotations from the public database:
      1> Predict coding sequences with software like prodigal.
      2> Remove duplicated genes.

3> Remove phage genes.
4> Remove genes with CRISPR sequence.

- Else if none of the input files are complete genomes but are raw sequencing reads, do the following:
  1> Use assmbly tools for *de novo* assembly to get files of contigs for each sample.
  2> Use one of the samples with good assembly quality, predict coding sequences with software like prodigal.
  3> Remove duplicated genes.
  4> Remove phage genes.
  5> Remove genes with CRISPR sequence.

2. A file in multiple nucleotide fasta format with genome sequences.
   - Complete sequences: Should be in fasta format.
   - Assembled contigs: Concatenate the contigs to form one fake genome sequence, which can be done with the script FormatContig.pl in ./scripts. Then concatenate all genome sequences or fake genome sequences to form a multiple-fasta file.

```
perl FormatContig.pl <contig fasta file> <header of the output fasta file > <output fasta file>
```

Cautions

Take note about giving different header names for the fasta file. If the header file contains blanks, the first column should all be different.

Output files

- **.core.fasta** The concatenated core genomes before recombination removal.
- **.concatenation.log** The concatenation log of the core.fasta file that is composed of each gene sequence name and the respective start and end position in the concatenated core genome.
- **.snpmat** A matrix of scaled number of SNPs in each gene in each genomic sequence.
- **.DBSCAN.outliermat** A matrix of recombinant genes identified by DBSCAN with '1' denoting recombinant while '0' denoting non-recombinant.
- **.DBSCAN.removal.fasta** The concatenated core genomes after DBSCAN recombination removal.
- **.Grubbs.outliermat** A matrix of recombinant genes identified by Grubbs with '1' denoting recombinant while '0' denoting non-recombinant.
- **.Grubbs.removal.fasta** The concatenated core genomes after Grubbs recombination removal.
- **.kNN.outliermat** A matrix of recombinant genes identified by kNN with '1' denoting

recombinant while ‘0’ denoting non-recombinant.

- **.kNN.removal.fasta** The concatenated core genomes after kNN recombination removal.