Pipelines

1 DeCon

1.1 Introduction

DeCon is designed to retrieve genomic sequences of a target **phylum** from metagenomic assembly of paired next generation sequencing (NGS) reads. First, contigs below 400 base pair (bp) are removed and NGS reads are mapped to filtered assembly by minimap2 (Li 2018). Second, SprayNPray (Garber et al. 2022) is used to compute coverage, GC content and coding density of each contigs. Third, contigs are searched against non-redundant (nr) database by DIAMOND (Buchfink et al. 2015) in long read mode and assigned to phyla by MEGAN (Huson et al. 2007) in long read mode. Forth, a decision tree classifier is trained, taking coverage, GC content and coding density of contigs as training features and phylum assignments from MEGAN as target value. This classifier is used to compute phylum assignments of contigs that are not determined by MEGAN. 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.

1.2 Dependencies

Softwares

R

Python

minimap2

SAMtools

SprayNPray

DIAMOND

MEGAN tools (daa-meganizer daa2info)

seqkit

QUAST

BUSCO

Databases

DIAMOND database (nr)

MEGAN database

BUSCO database

Python modules

numpy

pandas

scikit-learn

R packages

reticulate

stringr

1.3 Usage

Modify configuration file (templated as DeCon_conf.R), and run

```
Rscript path/DeCon_pipeline.R path/DeCon_main.R path/DeCon_main.py path/DeCon_conf.R
```

2 ProtGeneCall

2.1 Introduction

ProtGeneCall is designed to call protein-coding genes from genome, combining protein-genome alignments, transcriptome-genome alignments and *ab initio* gene predictions. First, repeat elements are identified by RepeatModeler (Smit et al. 2015b) and masked by RepeatMasker (Smit et al. 2015a). Masked genome is used for downstream analysis. Second, proteins of closely related species are mapped to the masked genome by miniprot (Li 2023). Third, paired RNA-sequencing (RNA-seq) reads are mapped to masked genome by Hisat2 (Kim et al. 2019). Forth, transcriptome-genome alignments are computed by StringTie (Pertea et al. 2015). Fifth, gene structures are predicted from transcriptome-genome alignments by TransDecoder (Haas et al. 2016), combining searching against UniRef (Suzek et al. 2007) by (Buchfink et al. 2015) and PfamA (Mistry et al. 2021) by HMMER (Eddy 1992). Sixth, AUGUSTUS (Stanke et al. 2003) is trained with gene structures from TransDecoder to compute gene predictions. Seventh, BRAKER (Hoff et al. 2019) is trained with RNA-seq mapping to call genes. Eighth, GALBA (Hoff et al. 2019) is trained with proteins of closely related species. Ninth, protein-genome alignments from miniprot, transcript-genome alignments from Hisat2-StringTie, and *ab initio* gene predictions from TransDecoder, AUGUSTUS, BRAKER and GALBA are integrated into consensus gene structures by EvidenceModeler (Haas et al. 2008). Tenth, genes from EvidenceModeler are removed if they are supported by only one *ab initio* predictor and lack protein/RNA-seq evidence. Eleventh, two iterations of PASA (Haas et al. 2008) is used to update filtered gene structures from EvidenceModeler. Twelfth, genes with in-frame stop codons or incomplete coding regions (coding regions with length cannot be divided by 3) are removed and the predicted peptide set is evaluated by BUSCO (Simão et al. 2015).

2.2 Dependencies

Softwares

R

Python

RepeatModeler

RepeatMasker

miniprot

Hisat2

SAMtools

StringTie

TransDecoder

HMMER

DIAMOND

AGAT

AUGUSTUS

BLAST+

GALBA

BRAKER

EvidenceModeler

BUSCO

gffread

seqkit

MAKER

Databases

DIAMOND database (UniRef)

Pfam-A

External scrpts

cufflinks_gtf_to_alignment_gff3.pl from EvidenceModeler

augustus_GFF3_to_EVM_GFF3.pl from EvidenceModeler

gth2gtf.pl from AUGUSTUS

computeFlankingRegion.pl from AUGUSTUS

gff2gbSmallDNA.pl from AUGUSTUS

gtf2aa.pl from AUGUSTUS

simplifyFastaHeaders.pl from AUGUSTUS

aa2nonred.pl from AUGUSTUS

filterGenesIn.pl from AUGUSTUS

autoAug.pl from AUGUSTUS

evm_evidence.py in this GitHub

R packages

stringr

parallel

2.3 Usage

Modify configuration file (templated as ProtGeneCall_conf.R), and run

Rscript path/ProtGeneCall_pipeline.R path/ProtGeneCall_main.R path/ProtGeneCall_conf.R

3 Pipeline for calling repeat elements from genome (RepCall)

3.1 Introduction

Pipeline RepCall is designed to call repeat elements genes from genome. First, miniature inverted-repeat transposable elements (MITE) are called by MITE-Hunter (Han et al. 2010). Second, long terminal repeats (LTRs) are identified by incorporating LTR_FINDER_parallel (Ou et al. 2019), LTRharvest (Ellinghaus et al. 2008) and LTR_retriever (Ou et al. 2018). Third, identified MITEs and LTRs are masked by RepeatMasker (Smit et al. 2015a). Forth, RepeatModeler (Smit et al. 2015b) is used to further identify repeats in the masked genome. Fifth, the locations of MITEs, LTRs and repeats from RepeatModeler are identified by RepeatMasker and all repeats are incorporated into a consensus library.

3.2 Dependencies

Softwares

R
seqkit
MITE-Hunter
LTR_FINDER_parallel
LTRharvest
LTR_retriever
RepeatMasker
RepeatModeler

3.3 Usage

Modify configuration file (templated as RepCall_conf.R), and run

```
Rscript path/RepCall_pipeline.R path/RepCall_main.R path/RepCall_conf.R
```

4 ncRNAcall

4.1 Introduction

ncRNAcall is designed to call non-coding RNA (ncRNA) from genome. First, transfer RNA (tRNA) is identified by tRNAscan-SE (Lowe et al. 1997). Second, microRNA is called by miRNAture (Velandia-Huerto et al. 2021). Third, target genes of microRNA are identified by searching microRNA against annotated three prime untranslated regions (3'UTR) by miRanda (Enright et al. 2003). Forth, Infernal (Nawrocki et al. 2013) searches against Rfam (Kalvari et al. 2021) database to call other non-coding RNA, *e.g.* ribosomal RNA (rRNA) and small nuclear RNA (snRNA). Fifth, all results are incorporated together.

4.2 Dependencies

Softwares

R

tRNAscan-SE

biocode

miRNAture

miRanda

bedtools

seqkit

Infernal

Databases

miRNAture database

Rfam database

R packages parallel

stringr

4.3 Usage

Modify configuration file (templated as ncRNAcall_conf.R), and run

Rscript path/ncRNAcall_pipeline.R path/ncRNAcall_main.R path/ncRNAcall_conf.R

5 buscoProt2Phylo

5.1 Introduction

buscoProt2Phylo infers phylogenetic tree using single-copy genes defined by BUSCO (Simão et al. 2015). First, from BUSCO runs complete single-copy protein sequences are collected and classified according to protein families that they belong to. Second, for protein famiies that are identified in above 4 BUSCO runs, protein sequences are aligned by MAFFT (Katoh et al. 2002). Third, multiple sequence alignments from MAFFT are trimmed by trimAl (Capella-Gutiérrez et al. 2009). Forth, gene trees are inferred from trimmed multiple sequence alignments by IQ-TREE (Minh et al. 2020) with 1,000 bootstrap replicates. Fifth, species tree is inferred from gene trees by ASTRAL (Zhang et al. 2018). Sixth, a supermatrix method was used to infer species tree from the multiple sequence alignments from MAFFT. Multiple sequence alignments contains 85%, 87.5%, 90% 92.5%, 95%, 97.5% and 100% of the total species were concatenated into supermatrixes, respectively. Missing species were represented by gaps. From each supermatrix a species tree was inferred by IQ-TREE (Minh et al. 2020) with 1,000 bootstrap replicates.

5.2 Dependencies

Softwares

R

MAFFT

trimAl

seqkit

IQ-TREE

ASTRAL

R packages

parallel

5.3 Usage

Modify configuration file (templated as buscoProt2Phylo_conf.R), and run

Rscript path/buscoProt2Phylo_pipeline.R path/nbuscoProt2Phylo_main.R path/buscoProt2Phylo_conf.R

6 metaTrans

6.1 Introduction

metaTrans is designed for taxonomic profiling of metatranscriptomic sequencing of paired NGS reads. First, metatranscriptomic reads are mapped to corresponding host genome by Hisat2 (Kim et al. 2019) and unmapped reads are extracted by SAMtools (Li et al. 2009). Second, ribosomal RNA reads are removed by SortMeRNA (Kopylova et al. 2012). Third, all reads are pooled together and assembled by rnaSPAdes (Bushmanova et al. 2019). Forth, MMseqs2 (Steinegger et al. 2017) (–cov-mode 1 -c 0.75 –min-seq-id 0.75) is used to remove redundancy in assembly from rnaSPAdes. Fifth, coding regions of assembled transcripts are identified by TransDecoder (Haas et al. 2016), combining searching against UniRef (Suzek et al. 2007) by DIAMOND (Buchfink et al. 2015) and PfamA (Mistry et al. 2021) by HMMER (Eddy

1992). Sixth, protein sequences transcribed by assembled transcripts are searched against non-redundant database by DIAMOND (Buchfink et al. 2015) and assigned to taxa by MEGAN (Huson et al. 2007). Seventh, reads are mapped to assembled transcripts by minimap2 (Li 2018) and SAMtools is used to compute coverage and depth of transcripts. Eighth, coverage, depth, coordinates of coding regions and taxonomy assignments of transcripts are taken together as comprehensive tables. Ninth, protein functions are inferred by InterproScan (Jones et al. 2014). Tenth, protein functions are inferred by eggNOG-mapper (Cantalapiedra et al. 2021).

6.2 Dependencies

Softwares

R

Hisat2

SAMtools

SortMeRNA

SPAdes

MMseqs2

TransDecoder

DIAMOND

HMMER

MEGAN tools (daa-meganizer daa2info)

minimap2

gffread

seqkit

MAKER

InterproScan

eggNOG-mapper

External scripts simplifyFastaHeaders.pl from AUGUSTUS

Databases

DIAMOND database (UniRef)

Pfam-A

R packages

stringr

6.3 Usage

Modify configuration file (templated as metaTrans_conf.R), and run

```
Rscript path/metaTrans_pipeline.R path/metaTrans_main.R path/metaTrans_conf.R
```

7 PseudoCall

7.1 Introduction

PseudoCall is designed to call pseudogenes with PseudoPipe that has been modified to (1) use stricter criteria for filtering blast hits, (2) run commands in parallel and (3) enable restarting. (**To be continued...**)

7.2 Dependencies

7.3 Usage

Modify configuration file (templated as PseudoCall_conf.R), and run

```
Rscript path/PseudoCall_pipeline.R path/PseudoCall_main.R path/PseudoCall_conf.R
```

References

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

Bushmanova, Elena et al. (2019). "rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data". In: *GigaScience* 8.9, giz100.

Cantalapiedra, Carlos P et al. (2021). "eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale". In: *Molecular biology and evolution* 38.12, pp. 5825–5829.

Capella-Gutiérrez, Salvador et al. (2009). "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses". In: *Bioinformatics* 25.15, pp. 1972–1973.

Eddy, Sean (1992). "HMMER user's guide". In: *Department of Genetics, Washington University School of Medicine* 2.1, p. 13.

Ellinghaus, David et al. (2008). “LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons”. In: *BMC bioinformatics* 9, pp. 1–14.

Enright, Anton et al. (2003). “MicroRNA targets in Drosophila”. In: *Genome biology* 4, pp. 1–27.

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.

Haas, B et al. (2016). “TransDecoder (find coding regions within transcripts)”. In: *Google Scholar*.

Haas, Brian J et al. (2008). “Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments”. In: *Genome biology* 9, pp. 1–22.

Han, Yujun et al. (2010). “MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences”. In: *Nucleic acids research* 38.22, e199–e199.

Hoff, Katharina J et al. (2019). “Whole-genome annotation with BRAKER”. In: *Gene prediction: methods and protocols*, pp. 65–95.

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

Jones, Philip et al. (2014). “InterProScan 5: genome-scale protein function classification”. In: *Bioinformatics* 30.9, pp. 1236–1240.

Kalvari, Ioanna et al. (2021). “Rfam 14: expanded coverage of metagenomic, viral and microRNA families”. In: *Nucleic Acids Research* 49.D1, pp. D192–D200.

Katoh, Kazutaka et al. (2002). “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform”. In: *Nucleic acids research* 30.14, pp. 3059–3066.

Kim, Daehwan et al. (2019). “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype”. In: *Nature biotechnology* 37.8, pp. 907–915.

Kopylova, Evguenia et al. (2012). “SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data”. In: *Bioinformatics* 28.24, pp. 3211–3217.

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

— (2023). “Protein-to-genome alignment with miniprot”. In: *Bioinformatics* 39.1, btad014.

Li, Heng et al. (2009). “The sequence alignment/map format and SAMtools”. In: *bioinformatics* 25.16, pp. 2078–2079.

Lowe, Todd M et al. (1997). “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence”. In: *Nucleic acids research* 25.5, pp. 955–964.

Minh, Bui Quang et al. (2020). "IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era". In: *Molecular biology and evolution* 37.5, pp. 1530–1534.

Mistry, Jaina et al. (2021). "Pfam: The protein families database in 2021". In: *Nucleic acids research* 49.D1, pp. D412–D419.

Nawrocki, Eric P et al. (2013). "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29.22, pp. 2933–2935.

Ou, Shujun et al. (2018). "LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons". In: *Plant physiology* 176.2, pp. 1410–1422.

— (2019). "LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons". In: *Mobile DNA* 10.1, pp. 1–3.

Pertea, Mihaela et al. (2015). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads". In: *Nature biotechnology* 33.3, pp. 290–295.

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

Smit, AFA et al. (2015a). *RepeatMasker Open-4.0. 2013–2015.*

Smit, AFA et al. (2015b). "RepeatModeler Open-1.0. 2008–2015". In: *Seattle, USA: Institute for Systems Biology. Available from: httpwww. repeatmasker. org, Last Accessed May* 1, p. 2018.

Stanke, Mario et al. (2003). "Gene prediction with a hidden Markov model and a new intron submodel". In: *Bioinformatics* 19.suppl_2, pp. ii215–ii225.

Steinegger, Martin et al. (2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature biotechnology* 35.11, pp. 1026–1028.

Suzek, Baris E et al. (2007). "UniRef: comprehensive and non-redundant UniProt reference clusters". In: *Bioinformatics* 23.10, pp. 1282–1288.

Velandia-Huerto, Cristian A et al. (2021). "miRNAture—Computational Detection of microRNA Candidates". In: *Genes* 12.3, p. 348.

Zhang, Chao et al. (2018). "ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees". In: *BMC bioinformatics* 19.6, pp. 15–30.