Mar 23, 2019

Working

Using CTT for comprehensive superfamily gene annotations

Zhihua Hua[1]

[1]Department of Environmental and Plant Biology, Ohio University, Athens, Ohio 45701,USA

dx.doi.org/10.17504/protocols.io.zf4f3qw

Zhihua Hua

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Hua Z, Early JM: Closing Target Trimming: a Perl Package for Discovering Hidden Superfamily Loci in Genomes. PLoS One 2019, (Under review).

PROTOCOL STATUS

Working

We use this protocol in our group and it is working

SAFETY WARNINGS

BEFORE STARTING

This protocol runs on CentOS 7 linux operating system with a Bioperl package installed. To install Bioperl, please see "BIOPERL INSTALLATION" at https://bioperl.org/INSTALL.html or some tips availabel at https://github.com/hua-lab/ctt.

Steps 1 to 3 are required only if ctt and its dependicies have not been installed.

1 Get the CTT package (<1 min)

Under the home directory, type "git clone https://github.com/hua-lab/ctt.git" to clone the ctt package.

e.g., [user@localhost ~]$ git clone https://github.com/hua-lab/ctt.git

2 Compile dependencies (~10 min).

Under the home directory, type "cd ./ctt/dependencies/" to enter the directory of "dependencies" in the ctt package, then type "make all" to compile "blast", "hmmer","CD-HIT","genewise", "pfam database", and "pfamscan" packages. An admin (sudo) user account is required in order to compile genewise program.

3 Activate WISECONFIGDIR (<1 min)

In order to allow genewise to recognize its configuration, its WISECONFIGDIR path needs to be activated. To do this, you may simply logout and login the server once.

4 Retrieve species databases (time varies)

Three files, including genome sequence, protein sequence, and gene annotation GFF3 files, are needed for ctt annotation. This protocol is based on the genome sequences organized at Ensembl genome project.

Go to Ensembl website at https://useast.ensembl.org/index.html, click "Downloads" in the list at the top of the website. At the download website, click "Download databases" located at the right corner of the webpage to enter the database page. On this page, you may find the genome sequence database of your favorite species. You may either simply browse the species whose genome sequence files are ordered alphabetically or use keyword search located at the right top corner of the spreadsheet. You can also download the files via the FTP site whose link is available on the same webpage.

Once you find the genome sequence files of your favorite organism, I recommend you to use wget to download its genome sequence, protein sequence, and gene annotation GFF3 files directly into the "./ctt/species_databases" directory. To find the link of each file, you may just right click each corresponding item in the spreadsheet and select "copy link". An example for downloading the three files of Human Ensembl genome is listed as follows.

e.g., to download human genome sequences (I recommend to use the complete genome file)

wget -r -nd -A '*dna.toplevel.fa.gz' ftp://ftp.ensembl.org/pub/release-95/fasta/homo_sapiens/dna/

e.g., to download human protein sequences (To find all possible annotations, I recommend to use the complete set of protein sequences. You may delete other protein sequence files if any. For example, for human protein sequence, we use the file "Homo_sapiens.GRCh38.pep.all.fa.gz". The ab initio annotation sequence file "Homo_sapiens.GRCh38.pep.abinitio.fa.gz" is removed).

wget -r -nd -A '*.fa.gz' ftp://ftp.ensembl.org/pub/release-95/fasta/homo_sapiens/pep/

e.g., to download human gene annotation GFF3 file (Only the complete GFF3 file will be used. You may delete the GFF3 files for single chromosome annotations)

wget -r -nd -A '*.gff3.gz' ftp://ftp.ensembl.org/pub/release-95/gff3/homo_sapiens

Since all the files downloaded are gz compressed files, type "gunzip *.gz" to decompress them.

5 Organize database files (<30 min).

Ctt is designed to work on multiple genomes, whose genome sequence, GFF3, and protein sequence file names need to be organized in order in a file, named "organismal_genome_gff3_proteome_files.tab" under the "./ctt/species_databases/" directory.

To do this, you may organize them in Excel, save it as a text file, and transfer the text file into the directory of "./ctt/species_databases/". Alternatively, you may just use vim editor to write this file under the same directory. An example for organizing 10 vertebrate genome sequence files analyzed in the paper (Hua and Early, 2019) is provided as follows.

Callithrix_jacchus.ASM275486v1.dna.toplevel.fa Callithrix_jacchus.ASM275486v1.95.gff3 Callithrix_jacchus.ASM275486v1.pep.all.fa
Choloepus_hoffmanni.choHof1.dna.toplevel.fa Choloepus_hoffmanni.choHof1.95.gff3 Choloepus_hoffmanni.choHof1.pep.all.fa
Cyprinodon_variegatus.C_variegatus-1.0.dna.toplevel.fa Cyprinodon_variegatus.C_variegatus-1.0.95.gff3
Cyprinodon_variegatus.C_variegatus-1.0.pep.all.fa
Ficedula_albicollis.FicAlb_1.4.dna.toplevel.fa Ficedula_albicollis.FicAlb_1.4.95.gff3 Ficedula_albicollis.FicAlb_1.4.pep.all.fa
Haplochromis_burtoni.AstBur1.0.dna.toplevel.fa Haplochromis_burtoni.AstBur1.0.95.gff3 Haplochromis_burtoni.AstBur1.0.pep.all.fa
Ictidomys_tridecemlineatus.SpeTri2.0.dna.toplevel.fa Ictidomys_tridecemlineatus.SpeTri2.0.95.gff3
Ictidomys_tridecemlineatus.SpeTri2.0.pep.all.fa
Mesocricetus_auratus.MesAur1.0.dna.toplevel.fa Mesocricetus_auratus.MesAur1.0.95.gff3 Mesocricetus_auratus.MesAur1.0.pep.all.fa
Mus_musculus.GRCm38.dna.toplevel.fa Mus_musculus.GRCm38.95.gff3 Mus_musculus.GRCm38.pep.all.fa
Pelodiscus_sinensis.PelSin_1.0.dna.toplevel.fa Pelodiscus_sinensis.PelSin_1.0.95.gff3 Pelodiscus_sinensis.PelSin_1.0.pep.all.fa
Tupaia_belangeri.TREESHREW.dna.toplevel.fa Tupaia_belangeri.TREESHREW.95.gff3 Tupaia_belangeri.TREESHREW.pep.all.fa

6 Make blast databases for tBLASTn and BLASTp searches (<30 min).

To make tBLASTn database use the genome sequence file obtained in Step 4 under the "./ctt/species_databases" directory, type "makeblastdb -in genome_file_name -dbtype nucl -out genome_file_name.db".

e.g., to make tBLASTn database of human genome sequences

makeblastdb -in Homo_sapiens.GRCh38.dna.toplevel.fa -dbtype nucl -out Homo_sapiens.GRCh38.dna.toplevel.fa.db

To make BLASTp database use the protein sequence file obtained in Step 4 under the "./ctt/species_databases" directory, type "makeblastdb -in proteome_file_name -dbtype prot -out proteome_file_name.db".

e.g., to make BLASTp database of human protein equences

makeblastdb -in Homo_sapiens.GRCh38.pep.all.fa -dbtype prot -out Homo_sapiens.GRCh38.pep.all.fa.db

7 Download seed sequences for the superfamily of your interest (<30 min).

Go to Pfam website at http://pfam.xfam.org. Enter the ID or name of the superfamily of your interest in "JUMP TO" search engine

located in the middle of the webpage. For example, type "Pkinase" and click "Go" button for jumping to the Pkinase family Summary webpage. You need to type the exact family name that Pfam recognizes. Otherwise, Pfam may not be able to help you locate the webpage of the family.

To find the seed sequence file of the superfamily, click "Alignments" button located at the top left corner of the webpage. In the middle of the "Alignments" webpage, you will see options of "Format an alignment". Select options for "seed" alignment, "Fasta" format, "Alphabetical" order, "All upper case" sequences, "No gaps (unaligned)", and "Download". You are now able to download a fasta-formated sequence file by clicking "Generate" button. The sequence file is automatically named in a format as "(PfamID of the superfamily)_seed.txt".

Transfer the seed sequence file into the seed directory of ctt at "./ctt/seeds/".

8 Run ctt annotation (time varies).

Under the "./ctt/" directory to annotate the superfamily of your interest in selected genomes using the format as follows.

perl ctt.pl -seed family_seed_file.txt -f Pfam_family_name -superfamily simplified_family_id_you_named

For example, to annotate Pkinase in selected genomes whose genomic sequence, gff3, and protein sequence files have been organized in the "./ctt/species_databases/" directory. Got to the ctt directory and type the code as follows.

e.g. perl ctt.pl -seed PF00069_seed.txt -f Pkinase -superfamily PK