Diverse commands used for Eukaryotic MAGs analyses

Grégoire Michoud

02/06/2021

Metaeuk

Download the databases `MMETSP_uniclust50_MERC` and `MMETSP_zenodo_3247846_uniclust90_2018_08_seed_valid_taxids` from *http://wwwuser.gwdg.de/~compbiol/metaeuk*

Installation

`mamba create -n metaeuk -c conda-forge -c bioconda metaeuk=4.a0f584d`

Run

```
source activate metaeuk
easy-predict --threads 35 --slice-search -e 100 --metaeuk-eval 0.0001 --min-ungapped-score 35 --min-exon-aa 20 --metaeuk-tcov 0.6 --min-length 40 --disk-space-limit 200G ASSEMBLY_euk_sim.fasta
/mnt/databases/MMETSP/MMETSP_uniclust50_MERC ASSEMBLY_euk_metaeuk temp
```

And then the taxonomy

```
metaeuk taxtocontig --threads 35 ASSEMBLY_euk_sim_metaeuk_db/contigs
Metaeuk/ASSEMBLY_euk_metaeuk.fas Metaeuk/ASSEMBLY_euk_metaeuk.headersMap.tsv
../../../databases/MMETSP/MMETSP_zenodo_3247846_uniclust90_2018_08_seed_valid_taxids Metaeuk_tax temp --majority 0.5 --tax-lineage 1 --lca-mode 2
```

All MAGs proteins were extracted from the `ASSEMBLY_euk_metaeuk.fas`

Busco

Installation

`mamba create -n Busco -c conda-forge -c bioconda busco=5.0.0`

Run

All protein files (*.faa*) should be in the same folder

```
source activate Busco
for i in *faa
```

```
    do busco -m prot -c 28 -i $i -o ${i%.faa}_busco -l eukaryota
done
```

EggNog

Installation

`mamba create -n eggnog -c conda-forge -c bioconda eggnog-mapper=2.1.0`

Had some issues with the conda installation so downloaded the release (https://github.com/eggnogdb/eggnog-mapper/archive/refs/tags/2.1.2.tar.gz) and used this version inside the conda environment

Run

Concatenate all MAGs proteins to speed up the calculation

```
source activate eggnog
/work/sber/Databases/eggnog-mapper-2.1.2/emapper.py --dbmem --resume --cpu 28 -i /scratch/gmichoud/Annotation/allConcoctFinal.faa --itype proteins -m diamond --sensmode very-sensitive -o /scratch/gmichoud/Annotation/allConcoctFinal_egg
```

EUKeule

Installation

`mamba create -n EUKulele -c bioconda -c conda-forge eukulele=1.0.3`

Run

```
source activate EUKulele
for i in PhyloDB EukProt MMETSP
    do for j in Prots/*faa
        do EUKulele -m mags --reference_dir /work/sber/Databases/EUKuleleDB/$i -s ${j%.faa} -o ${j%.faa}\_$i --CPUs 7 --p_ext faa &
done
```

EukCC

Installation

`mamba create -n eukcc -c bioconda -c conda-forge eukcc=0.3`

You need to download the database prior to running the software (http://ftp.ebi.ac.uk/pub/databases/metagenomics/eukcc/eukcc_db_v1.1.tar.gz)

Run

```
for i in *faa
   eukcc --db /mnt/md1200/epfl_sber/databases/eukcc_db_20191023_1/ --proteins $i -
o ${i%.faa}_eukcc --ncorespplacer 5 --ncores 15
done
```

Homology comparisons

All Refseq and Genbank genomes and proteomes belonging to stramenopiles were downloaded using the `datasets` (v11.13.6) software of the NCBI

`./datasets download genome --exclude-gff3 --exclude-rna taxon "stramenopiles"`

Then, all protein coding genes that were absent from the genbank genomes were obtained using Metaeuk

```
source activate metaeuk
for i in GCA*fna
   do metaeuk easy-predict --threads 28 --slice-search -e 100 --metaeuk-eval
0.0001 --min-ungapped-score 35 --min-exon-aa 20 --metaeuk-tcov 0.6 --min-length 40
/work/sber/Algea/$i /work/sber/Databases/Metaeuk/MMETSP_uniclust50_MERC
/work/sber/Algea/${i%.fna}\_metaeuk /scratch/gmichoud/Algea/${i%.fna}_tmp
done
```

Add the name of the assembly to each proteins to be able to distinguish them later on

```
for i in Reference/*.faa
   do perl -pe "'s/>/>${i%.faa}_/g'" $i \> ${i%.faa}_rename.faa;
done
```

Then make a diamond database for all of them

`mamba create -n diamond -c bioconda -c conda-forge diamond=2.0.9.147`

```
source activate diamond
for j in  Reference/*faa
   do diamond makedb -in $j -o ${j%.faa}.dmnd
done

for j in Reference/*dmnd
   for i in Query/*faa
      do diamond blastp -d $j -q $i -e 1e-12 -f 6 -k 1 --very-sensitive -o
Results/${i%.faa}\_${i%.dmnd}.txt -p 7;
   done
done

 for i in *txt; do echo $i; awk '$3>60' $i | cut -f3 | wc -l; done >>
../numberHits.txt
```

Check the `numberHits.txt` file in R or excel to find the best hits