Searching further on viral matching HMM profiles

Studying further what matched the fOTUs for purposes of trying to validate the results

LJM

2019-12-22

Contents

1 Introduction

It would be nice to know further what viruses contributed to HMMs that matched to the fOTUs. Fortunately EggNOG database v 5.0 (Huerta-Cepas et al. 2019) provides that information. Let's download the relevant file from EggNOG database and taxonomic data from NCBI and create one big tsv file with all that information. Thereafter, let's check which HMMs matched significantly (with $-log_{10}$(e-value) of 2 or higher) to which fOTUs.

1.1 Load libraries

```
library(tidyverse)
library(rentrez)
library(XML)
```

2 Download from eggNOG database

2.1 Download members information on all viral HMMs

```
OUTPUT="../analyses/HMMsearch/validation_and_further_analyses/10239_members.tsv.gz"
wget --verbose --continue --output-document="$OUTPUT" \
"http://eggnog5.embl.de/download/eggnog_5.0/per_tax_level/10239/10239_members.tsv.gz"
gunzip "$OUTPUT"
```

2.2 Download annotation information on all viral HMMs

```
OUTPUT="../analyses/HMMsearch/validation_and_further_analyses/10239_annotations.tsv.gz"
wget --verbose --continue --output-document="$OUTPUT" \
"http://eggnog5.embl.de/download/eggnog_5.0/per_tax_level/10239/10239_annotations.tsv.gz"
gunzip "$OUTPUT"
```

3 Merge annotations and members files

```
INPUT1="../analyses/HMMsearch/validation_and_further_analyses/10239_annotations.tsv"
INPUT2="../analyses/HMMsearch/validation_and_further_analyses/10239_members.tsv"
SCRIPT="../scripts/merge_files.awk"
OUT="../analyses/HMMsearch/validation_and_further_analyses/10239_members_annotations.tsv"

awk -f $SCRIPT $INPUT1 $INPUT2 > $OUT
```

Now let's take a short look at the output

```
IN="../analyses/HMMsearch/validation_and_further_analyses/10239_members_annotations.tsv"
head -n 3 $IN
```

```
## taxid    HMM eggNOG_cat  mem_taxid.pfam_id   mem_taxid   description
## 10239    4QAIH   S   1229753.K7QJT8_9CAUD,948870.I7HXC4_9CAUD    1229753,948870  NA
## 10239    4QAII   S   1337877.R9VYA7_9CAUD,1589751.A0A0C5AES7_9CAUD   1337877,1589751 NA
```

Looks pretty ok.

Now we need to fetch taxonomic information from NCBI based on mem_taxid column information and append it to this previous data. R package `rentrez` can handle the accessing of NCBI and tidyverse can handle the rest.

4 Session info

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 9 (stretch)
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib/libopenblasp-r0.2.19.so
##
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] XML_3.98-1.20   rentrez_1.2.2   forcats_0.4.0   stringr_1.4.0
##  [5] dplyr_0.8.3     purrr_0.3.2     readr_1.3.1     tidyr_0.8.3
##  [9] tibble_2.1.3    ggplot2_3.2.1   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.2       cellranger_1.1.0 pillar_1.4.2     compiler_3.6.1
##  [5] tools_3.6.1      zeallot_0.1.0    digest_0.6.20    lubridate_1.7.4
##  [9] jsonlite_1.6     evaluate_0.14    nlme_3.1-140     gtable_0.3.0
## [13] lattice_0.20-38  pkgconfig_2.0.2  rlang_0.4.0      cli_1.1.0
## [17] rstudioapi_0.10  yaml_2.2.0       haven_2.1.1      xfun_0.9
## [21] withr_2.1.2      xml2_1.2.2       httr_1.4.1       knitr_1.24
## [25] vctrs_0.2.0      generics_0.0.2   hms_0.5.1        grid_3.6.1
## [29] tidyselect_0.2.5 glue_1.3.1       R6_2.4.0         readxl_1.3.1
## [33] rmarkdown_1.15   bookdown_0.13    modelr_0.1.5     magrittr_1.5
## [37] backports_1.1.4  scales_1.0.0     htmltools_0.3.6  rvest_0.3.4
## [41] assertthat_0.2.1 colorspace_1.4-1 stringi_1.4.3    lazyeval_0.2.2
## [45] munsell_0.5.0    broom_0.5.2      crayon_1.3.4
```

References

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses." *Nucleic Acids Research* 47 (D1): D309–D314. https://doi.org/10.1093/nar/gky1085.