

Positive selection on genes interacting with SARS-Cov2, comparison of different analysis

Marie Cariou

Janvier 2021

Contents

1	Files manipulations	2
1.1	Read Janet Young's table	2
1.2	Read DGINN Young table	2
1.3	Joining Young and DGINN Young table	2
1.4	Read DGINN Table	3
1.5	Join Table and DGINN table	5
1.6	Add DGINN results on bat dataset	5
1.7	Write the new table	8

1 Files manipulations

1.1 Read Janet Young's table

```
workdir<-"/home/adminmarie/Documents/CIRI_BIBS_projects/2020_05_Etienne_covid/"

tab<-read.delim(paste0(workdir,
  "data/COVID_PAMLresults_332hits_plusBatScreens_2020_Apr14.csv"),
  fill=T, h=T, dec=",")
dim(tab)

## [1] 332 84
```

1.2 Read DGINN Young table

```
dginnY<-read.delim(paste0(workdir,
  "data/summary_primate_young.res"),
  fill=T, h=T)

dim(dginnY)

## [1] 1992 7
```

1.3 Joining Young and DGINN Young table

```
# correct gene names (MARC1)
val_remp=as.character(unique(dginnY$Gene)[(unique(dginnY$Gene) %in%
  tab$Gene.name)==F])

tab$Gene.name<-as.character(tab$Gene.name)
tab$Gene.name[158]<-val_remp
sum(unique(dginnY$Gene) %in% unique(tab$Gene.name))

## [1] 332
```

```

add_col<-function(method="PamLM1M2"){

tmp<-dginnY[dginnY$Method==method,
             c("Gene", "Omega", "PosSel", "PValue", "NbSites", "PSS")]

names(tmp)<-c("Gene.name", paste0("Omega_", method),
             paste0("PosSel_", method), paste0("PValue_", method),
             paste0("NbSites_", method), paste0("PSS_", method))

tab<-merge(tab, tmp, by="Gene.name")

return(tab)
}

tab<-add_col("PamLM1M2")
tab<-add_col("PamLM7M8")
tab<-add_col("BppM1M2")
tab<-add_col("BppM7M8")

# Manip pour la colonne BUSTED

tmp<-dginnY[dginnY$Method=="BUSTED",c("Gene", "Omega", "PosSel", "PValue")]
names(tmp)<-c("Gene.name", "Omega_BUSTED", "PosSel_BUSTED", "PValue_BUSTED")
tab<-merge(tab, tmp, by="Gene.name")

tmp<-dginnY[dginnY$Method=="MEME",c("Gene", "NbSites", "PSS")]
names(tmp)<-c("Gene.name", "NbSites_MEME", "PSS_MEME")
tab<-merge(tab, tmp, by="Gene.name")

```

1.4 Read DGINN Table

```

dginnT<-read.delim(paste0(workdir,
                          "data/DGINN_202005281649summary_cleaned.csv"),
                  fill=T, h=T, sep=",")

dim(dginnT)

```

```
## [1] 412 27

names(dginnT)

## [1] "File" "Name" "Gene" "GeneSize"
## [6] "omegaM0Bpp" "omegaM0codeml" "BUSTED" "BUSTED.p.valu
## [11] "MEME.PSS" "BppM1M2" "BppM1M2.p.value" "BppM1M2.NbSit
## [16] "BppM7M8" "BppM7M8.p.value" "BppM7M8.NbSites" "BppM7M8.PSS"
## [21] "codemlM1M2.p.value" "codemlM1M2.NbSites" "codemlM1M2.PSS" "codemlM7M8"
## [26] "codemlM7M8.NbSites" "codemlM7M8.PSS"

# Number of genes in dginn-primate output not present in the original table
dginnT[(dginnT$Gene %in% tab$Gene.name)==F, "Gene"]

## [1] ACE2 ADAM9[0-3120] ADAM9[3119-3927] ATP5MGL C
## [7] CEP135[3263-3678] CEP43 COQ8B COQ8A C
## [13] CSNK2B[608-2568] CYB5R1 DDX21[0-717] DDX21[716-2538] D
## [19] DPH5[0-702] DPH5[701-1326] DPY19L2 ELOC E
## [25] EXOSC3[1445-1980] FBN3 GNB4 GNB2 G
## [31] GOLGA7[311-549] GPX1[0-1218] GPX1[1217-2946] HDAC1 H
## [37] ITGB1[0-2328] ITGB1[2327-2844] LMAN2L MRPS5[0-1569] M
## [43] MGRN1 NDFIP2[0-768] NDFIP2[767-1314] NDUFAF2[0-258] N
## [49] NUP58 NUP58[0-1824] NUP58[1823-2367] PABPC3 P
## [55] PABPC5 PCSK5 PRIM2[0-1071] PRIM2[1070-1902] P
## [61] PTGES2[0-1587] PTGES2[1586-2202] RAB8B RAB13 R
## [67] RAB2B RAB5A RAB5B RAB15 R
## [73] EZR[0-1458] EZR[1457-3771] MSN RETREG3 R
## [79] SLC44A2[0-2577] SLC44A2[2576-3657] SPART SRP72[0-2604] S
## [85] STOM[1046-1800] STOML3 TIMM29 TLE4 T
## [91] TLE2[1301-3987] TMPRSS2 TOMM70 TOR1B W
## [97] WFS1[2345-3216] YIF1B

## 411 Levels: AAR2 AASS AATF ABCC1 ACAD9 ACADM ACE2 ACSL3 ADAM9 ADAM9[0-3120] ADAM9[

# This includes paralogs, recombinations found by DGINN and additionnal genes
# included on purpose

# Number of genes from the original list not present in DGINN output
tab[(tab$Gene.name %in% dginnT$Gene)==F, "Gene.name"]

## [1] "ADCK4" "ARL6IP6" "ATP5L" "C19orf52" "C1orf50" "ER01LB" "FAM134C"
## [12] "SIGMAR1" "SPG20" "TCEB1" "TCEB2" "TOMM70A" "USP13" "VIMP"
```

```
names(dginnT)<-c("File", "Name", "Gene.name", "GeneSize",
  "dginn-primate_NbSpecies", "dginn-primate_omegaMOBpp",
  "dginn-primate_omegaM0codeml", "dginn-primate_BUSTED",
  "dginn-primate_BUSTED.p.value", "dginn-primate_MEME.NbSites",
  "dginn-primate_MEME.PSS", "dginn-primate_BppM1M2",
  "dginn-primate_BppM1M2.p.value", "dginn-primate_BppM1M2.NbSites",
  "dginn-primate_BppM1M2.PSS", "dginn-primate_BppM7M8",
  "dginn-primate_BppM7M8.p.value", "dginn-primate_BppM7M8.NbSites",
  "dginn-primate_BppM7M8.PSS", "dginn-primate_codemlM1M2",
  "dginn-primate_codemlM1M2.p.value", "dginn-primate_codemlM1M2.NbSites",
  "dginn-primate_codemlM1M2.PSS", "dginn-primate_codemlM7M8",
  "dginn-primate_codemlM7M8.p.value", "dginn-primate_codemlM7M8.NbSites",
  "dginn-primate_codemlM7M8.PSS")
```

1.5 Join Table and DGINN table

```
tab<-merge(tab,dginnT, by="Gene.name", all.x=T)
```

1.6 Add DGINN results on bat dataset

DGINN results from different analysis.

```
# original table
dginnbats<-read.delim(paste0(workdir,
  "data/DGINN_202005281339summary_cleaned.tab"),
  fill=T, h=T)

# rerun on corrected alignment
dginnbatsnew1<-read.delim(paste0(workdir,
  "data/DGINN_202011262248_summary.tab"),
  fill=T, h=T)
dginnbatsnew2<-read.delim(paste0(workdir,
  "data/DGINN_202012192053_summary.tab"),
  fill=T, h=T)

# colonne choice, BUSTED and Bppml form first file, codeml from the other one
```

```

dginnbatsnew<-dginnbatsnew1
dginnbatsnew$omegaM0codeml<-dginnbatsnew2$omegaM0codeml

dginnbatsnew$codemlM1M2<-dginnbatsnew2$codemlM1M2
dginnbatsnew$codemlM1M2_p.value<-dginnbatsnew2$codemlM1M2_p.value
dginnbatsnew$codemlM1M2_NbSites<-dginnbatsnew2$codemlM1M2_NbSites
dginnbatsnew$codemlM1M2_PSS<-dginnbatsnew2$codemlM1M2_PSS

dginnbatsnew$codemlM7M8<-dginnbatsnew2$codemlM7M8
dginnbatsnew$codemlM7M8_p.value<-dginnbatsnew2$codemlM7M8_p.value
dginnbatsnew$codemlM7M8_NbSites<-dginnbatsnew2$codemlM7M8_NbSites
dginnbatsnew$codemlM7M8_PSS<-dginnbatsnew2$codemlM7M8_PSS

####
## RIPK1 is actually a primate results
## 1. Take it and put it at the right place
ripk1<-as.vector(dginnbatsnew[dginnbatsnew$Gene=="RIPK1",])
tab$`dginn-primate_omegaM0Bpp`<-as.numeric(as.character(tab$`dginn-primate_omegaM0Bpp`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_BUSTED.p.value`<-as.numeric(as.character(tab$`dginn-primate_BUSTED.p.value`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_BppM1M2.p.value`<-as.numeric(as.character(tab$`dginn-primate_BppM1M2.p.value`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_BppM7M8.p.value`<-as.numeric(as.character(tab$`dginn-primate_BppM7M8.p.value`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_BppM7M8.PSS`<-as.numeric(as.character(tab$`dginn-primate_BppM7M8.PSS`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_codemlM1M2.p.value`<-as.numeric(as.character(tab$`dginn-primate_codemlM1M2.p.value`

## Warning:  NAs introduits lors de la conversion automatique
tab$`dginn-primate_codemlM1M2.PSS`<-as.numeric(as.character(tab$`dginn-primate_codemlM1M2.PSS`

## Warning:  NAs introduits lors de la conversion automatique

```

```

tab$`dginn-primate_codemlM7M8.p.value`<-as.numeric(as.character(tab$`dginn-primate_codemlM7M8.p.value`))
## Warning:  NAs introduits lors de la conversion automatique

tab$`dginn-primate_codemlM7M8.PSS`<-as.numeric(as.character(tab$`dginn-primate_codemlM7M8.PSS`))
## Warning:  NAs introduits lors de la conversion automatique

tab[tab$Gene.name=="RIPK1", "GeneSize"]<-ripk1$GeneSize
tab[tab$Gene.name=="RIPK1", "dginn-primate_NbSpecies"]<-ripk1$NbSpecies
tab[tab$Gene.name=="RIPK1", "dginn-primate_omegaM0Bpp"]<-ripk1$omegaM0Bpp
tab[tab$Gene.name=="RIPK1", "dginn-primate_omegaM0codeml"]<-ripk1$omegaM0codeml

tab[tab$Gene.name=="RIPK1", "dginn-primate_BUSTED"]<-ripk1$BUSTED
tab[tab$Gene.name=="RIPK1", "dginn-primate_BUSTED.p.value"]<-ripk1$BUSTED.p.value
tab[tab$Gene.name=="RIPK1", "dginn-primate_MEME.NbSites"]<-ripk1$MEME_NbSites
tab[tab$Gene.name=="RIPK1", "dginn-primate_MEME.PSS"]<-as.numeric(as.character(ripk1$MEME.PSS))
## Warning:  NAs introduits lors de la conversion automatique

tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM1M2"]<-ripk1$BppM1M2
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM1M2.p.value"]<-ripk1$BppM1M2.p.value
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM1M2.NbSites"]<-ripk1$BppM1M2_NbSites
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM1M2.PSS"]<-ripk1$BppM1M2_PSS

tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM7M8"]<-ripk1$BppM7M8
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM7M8.p.value"]<-ripk1$BppM7M8.p.value
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM7M8.NbSites"]<-ripk1$BppM7M8_NbSites
tab[tab$Gene.name=="RIPK1", "dginn-primate_BppM7M8.PSS"]<-ripk1$BppM7M8_PSS

tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM1M2"]<-ripk1$codemlM1M2
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM1M2.p.value"]<-ripk1$codemlM1M2.p.value
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM1M2.NbSites"]<-ripk1$codemlM1M2_NbSites
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM1M2.PSS"]<-ripk1$codemlM1M2_PSS
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM7M8"]<-ripk1$codemlM7M8
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM7M8.p.value"]<-ripk1$codemlM7M8.p.value
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM7M8.NbSites"]<-ripk1$codemlM7M8_NbSites
tab[tab$Gene.name=="RIPK1", "dginn-primate_codemlM7M8.PSS"]<-ripk1$codemlM7M8_PSS

## 2. Remove it
dginnbatsnew<-dginnbatsnew[dginnbatsnew$Gene!="RIPK1",]

```

```

## suppress redundant lines
dginnbats<-dginnbats[(dginnbats$Gene %in% dginnbatsnew$Gene)==FALSE,]
names(dginnbatsnew)<-names(dginnbats)

#####
dginnbatsnew[,4]<-as.numeric(dginnbatsnew[,4])
dginnbats[,6]<-as.numeric(as.character(dginnbats[,6]))

## Warning:  NAs introduits lors de la conversion automatique

dginnbats[,8]<-as.character(dginnbats[,8])
dginnbats[,12]<-as.character(dginnbats[,12])
dginnbats[,13]<-as.numeric(as.character(dginnbats[,13]))

## Warning:  NAs introduits lors de la conversion automatique

dginnbats[,16]<-as.character(dginnbats[,16])
dginnbats[,17]<-as.numeric(as.character(dginnbats[,17]))

## Warning:  NAs introduits lors de la conversion automatique

## replace by new data
dginnbats<-rbind(dginnbats, dginnbatsnew)

names(dginnbats)<-c("File", "bats_Name", "cooper.batsGene", paste0("bats_",
  names(dginnbats)[- (1:3)]))
names(dginnbats)

##   [1] "File"                "bats_Name"          "cooper.batsGene"
##   [5] "bats_NbSpecies"      "bats_omegaM0Bpp"    "bats_omegaM0codeml"
##   [9] "bats_BUSTED.p.value" "bats_MEME.NbSites"  "bats_MEME.PSS"
##  [13] "bats_BppM1M2.p.value" "bats_BppM1M2.NbSites" "bats_BppM1M2.PSS"
##  [17] "bats_BppM7M8.p.value" "bats_BppM7M8.NbSites" "bats_BppM7M8.PSS"
##  [21] "bats_codemlM1M2.p.value" "bats_codemlM1M2.NbSites" "bats_codemlM1M2.PSS"
##  [25] "bats_codemlM7M8.p.value" "bats_codemlM7M8.NbSites" "bats_codemlM7M8.PSS"

tab<-merge(tab,dginnbats, by="cooper.batsGene", all.x=T)

```

1.7 Write the new table

```
write.table(tab, "covid_comp_complete.txt", row.names=FALSE, quote=FALSE, sep="\t")
```