

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

Marie Cariou

Mai 2020

1 Files manipulations

I will compare Janet results to DGINN results, on the SAME alignment.

1.1 Read Janet table

```
tab<-read.delim("/home/adminmarie/Documents/CIRI_BIBS_projects/2020_05_Etienne_covid/
               fill=T, h=T, dec=",")
dim(tab)

## [1] 332 84

names(tab)

## [1] "PreyGene"
## [2] "PreyGene_JYname"
## [3] "BaitShort"
## [4] "Gene.name"
## [5] "list"
## [6] "description"
## [7] "other.names"
## [8] "top40_posSeln"
## [9] "Num.primite.seqs"
## [10] "Alignment.length..nucleotides."
## [11] "Alignment.length..codons."
## [12] "whole.gene.dN.dS.model.0"
## [13] "total.tree.length"
```

```

## [14] "total.dN.tree.length"
## [15] "total.dS.tree.length"
## [16] "p.value.M8vsM8a..raw."
## [17] "p.value.M8vsM8a..BH.corrected."
## [18] "pVal.M8vsM7"
## [19] "pVal.M8vsM7.adj"
## [20] "pVal.M2vsM1"
## [21] "pVal.M2vsM1.adj"
## [22] "X..codons.under.positive.selection"
## [23] "dN.dS.of.positively.selected.codons"
## [24] "Number.of.codons.with.BEB...0.9"
## [25] "Codons.under.positive.selection..BEB..0.9...alignment.position."
## [26] "cooper.batsGene"
## [27] "cooper.batsGene_Ensembl_ID"
## [28] "cooper.batsIsoform_Ensembl_ID"
## [29] "cooper.batsSpecies"
## [30] "cooper.batsReference_length.aa."
## [31] "cooper.batsPercent_analyzed"
## [32] "cooper.batsAverage_dNdS"
## [33] "cooper.batsMaximum_dS"
## [34] "cooper.batsAverage_M7_tree"
## [35] "cooper.batsAverage_M8_tree"
## [36] "cooper.batsM7_log_likelihood"
## [37] "cooper.batsM8_log_likelihood"
## [38] "cooper.batsM7.M8_p_value"
## [39] "cooper.batsM8a_log_likelihood"
## [40] "cooper.batsM8.M8a_pvalue"
## [41] "cooper.batsBEB_hits.pp.0.95."
## [42] "cooper.batsBEB_sites"
## [43] "cooper.primates.Gene"
## [44] "cooper.primates.Gene_Ensembl_ID"
## [45] "cooper.primates.Isoform_Ensembl_ID"
## [46] "cooper.primates.Species"
## [47] "cooper.primates.Reference_length.aa."
## [48] "cooper.primates.Percent_analyzed"
## [49] "cooper.primates.Average_dNdS"
## [50] "cooper.primates.Maximum_dS"
## [51] "cooper.primates.Average_M7_tree"
## [52] "cooper.primates.Average_M8_tree"

```

```
## [53] "cooper.primates.M7_log_likelihood"
## [54] "cooper.primates.M8_log_likelihood"
## [55] "cooper.primates.M7.M8_p_value"
## [56] "cooper.primates.M8a_log_likelihood"
## [57] "cooper.primates.M8.M8a_pvalue"
## [58] "cooper.primates.BEB_hits.pp.0.95."
## [59] "cooper.primates.BEB_sites"
## [60] "hawkins_Gene"
## [61] "hawkins_Positive.Selection..M8vM8a.p.value"
## [62] "hawkins_Positive.Selection..M8vM8a.FDR.corrected.p.value"
## [63] "hawkins_Gene.Name.Alias"
## [64] "hawkins_Connection.to.immunity.or.pathogens"
## [65] "hawkins_Connection.to.reproduction"
## [66] "hawkins_Connection.to.collagen"
## [67] "hawkins_Connection.to.peroxisome"
## [68] "hawkins_Gene.Description.for.Human.Ortholog..from.Genbank.GENE.database."
## [69] "CpGmask.numNT"
## [70] "CpGmask.numAA"
## [71] "CpGmask.overall.dN.dS"
## [72] "CpGmask.total.tree.length"
## [73] "CpGmask.total.dN.tree.length"
## [74] "CpGmask.total.dS.tree.length"
## [75] "CpGmask.pVal.M8vsM8a"
## [76] "CpGmask.pVal.M8vsM8a.adj"
## [77] "CpGmask.pVal.M8vsM7"
## [78] "CpGmask.pVal.M8vsM7.adj"
## [79] "CpGmask.pVal.M2vsM1"
## [80] "CpGmask.pVal.M2vsM1.adj"
## [81] "CpGmask.percent.sites.under.positive.selection"
## [82] "CpGmask.dN.dS.of.selected.sites"
## [83] "CpGmask.num.sites.with.BEB...0.9"
## [84] "CpGmask.which.sites.have.BEB...0.9"
```

1.2 Read DGINN table

```
dginn<-read.delim("/home/adminmarie/Documents/CIRI_BIBS_projects/2020_05_Etienne_covi
                fill=T, h=T)
```

```
dim(dginn)

## [1] 1992    7

names(dginn)

## [1] "Gene"      "Omega"     "Method"    "PosSel"    "PValue"    "NbSites"   "PSS"
```

1.3 Joining table

1.3.1 Based on which column?

```
head(tab)[,1:5]

##   PreyGene PreyGene_JYname BaitShort Gene.name list
## 1    PCNT          PCNT    nsp13    PCNT list26_COV_list4dataset2nonOrf
## 2    PVR           PVR     orf8     PVR list23_COV_list1orf
## 3    POLA1         POLA1    nsp1    POLA1 list24_COV_list2nonOrf
## 4 FASTKD5         FASTKD5     M FASTKD5 list26_COV_list4dataset2nonOrf
## 5    PRIM2         PRIM2    nsp1    PRIM2 list24_COV_list2nonOrf
## 6    ITGB1         ITGB1    orf8    ITGB1 list25_COV_list3dataset2orf

# gene avec un nom bizarre dans certaines colonne
tab[158,1:10]

##      PreyGene PreyGene_JYname BaitShort  Gene.name list
## 158  MTARC1      01/03/2020    nsp7 01/03/2020 list24_COV_list2nonOrf
##                                     description other.names top40_posSeln
## 158 mitochondrial amidoxime reducing component 1      MOSC1      no
##      Num.primite.seqs Alignment.length..nucleotides.
## 158                24                        1023

#
length(unique(dginn$Gene))

## [1] 332

length(unique(tab$PreyGene))

## [1] 332
```

```

length(unique(tab$Gene.name))

## [1] 332

#quelle paire de colonne contient le plus de noms identiques
sum(unique(dginn$Gene) %in% unique(tab$PreyGene))

## [1] 314

sum(unique(dginn$Gene) %in% unique(tab$Gene.name))

## [1] 331

# dginn$Gene et tab$Gene.name presque identiques sauf 1 ligne. Je soupçonne que c'est
tab[158,1:10]

##      PreyGene PreyGene_JYname BaitShort  Gene.name          list
## 158   MTARC1      01/03/2020      nsp7 01/03/2020 list24_COV_list2nonOrf
##                                     description other.names top40_posSeln
## 158 mitochondrial amidoxime reducing component 1      MOSC1          no
##      Num.primate.seqs Alignment.length..nucleotides.
## 158                24                        1023

# Verif:
tab[,1:10][!(tab$Gene.name %in% unique(dginn$Gene))==F,]

##      PreyGene PreyGene_JYname BaitShort  Gene.name          list
## 158   MTARC1      01/03/2020      nsp7 01/03/2020 list24_COV_list2nonOrf
##                                     description other.names top40_posSeln
## 158 mitochondrial amidoxime reducing component 1      MOSC1          no
##      Num.primate.seqs Alignment.length..nucleotides.
## 158                24                        1023

# yep

# Remplacement manuel par
as.character(unique(dginn$Gene)[!(unique(dginn$Gene) %in% tab$Gene.name)==F])

## [1] "MARC1"

```

```

# dans le tableau de Janet

val_remp=as.character(unique(dginn$Gene)[(unique(dginn$Gene) %in% tab$Gene.name)==F])

tab$Gene.name<-as.character(tab$Gene.name)

tab$Gene.name[158]<-val_remp

sum(unique(dginn$Gene) %in% unique(tab$Gene.name))

## [1] 332

```

1.3.2 new columns

```

add_col<-function(method="PamlM1M2"){

tmp<-dginn[dginn$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_"

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

1.3.3 Write new table

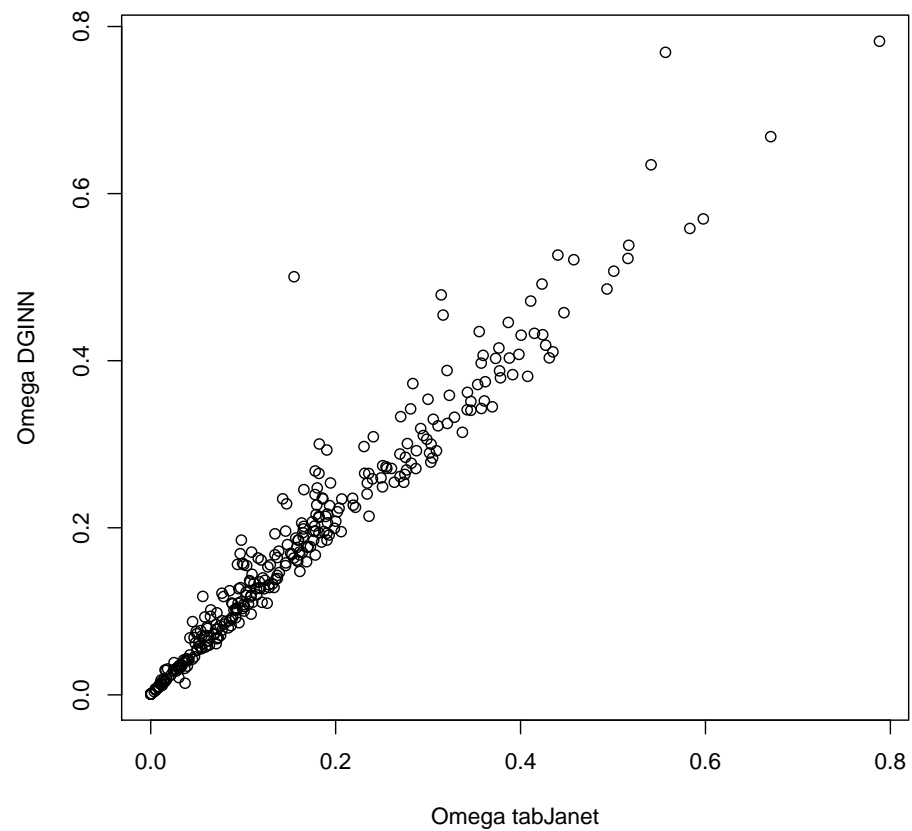
```
write.table(tab, "COVID_PAMLresults_332hits_plusBatScreens_plusDGINN_20200506.txt",
            row.names=F, quote=F, sep="\t")
```

1.4 Figure

1.4.1 Omega

Comparaison des Omega: colonne L "whole.gene.dN.dS.model.0" VS colonne "omega" dans la sortie de dginn.

```
plot(tab$whole.gene.dN.dS.model.0, tab$Omega_PamlM7M8,
      xlab="Omega tabJanet", ylab="Omega DGINN")
```



Quels sont les 2 gènes qui s'écartent de la bissectrice?

```
tab[tab$whole.gene.dN.dS.model.0<0.2 & tab$Omega_Pam1M7M8>0.4,c("Gene.name")]
## [1] "MRPS2"

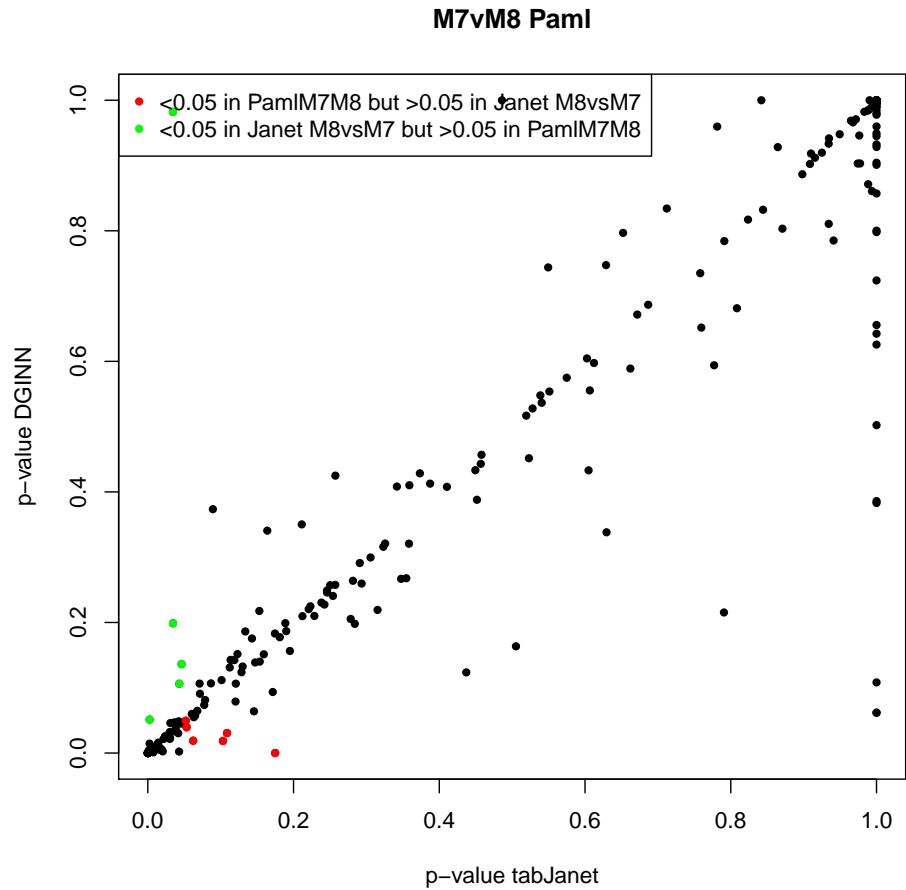
tab[tab$whole.gene.dN.dS.model.0<0.6 & tab$Omega_Pam1M7M8>0.7,c("Gene.name")]
## [1] "PVR"
```


1.4.2 pvalues pour M7M8

Cette fois, je compare la colonne R "pVal.M8vsM7", à la colonne "PValue" + ligne "PamlM7M8", pour la sortie de dginn.

```
plot(tab$pVal.M8vsM7, tab$PValue_PamlM7M8, pch=20,
      xlab="p-value tabJanet", ylab="p-value DGINN", main="M7vM8 Paml")
points(tab$pVal.M8vsM7[tab$pVal.M8vsM7>0.05 & tab$PValue_PamlM7M8<0.05],
       tab$PValue_PamlM7M8[tab$pVal.M8vsM7>0.05 & tab$PValue_PamlM7M8<0.05],
       col="red", pch=20)
points(tab$pVal.M8vsM7[tab$pVal.M8vsM7<0.05 & tab$PValue_PamlM7M8>0.05],
       tab$PValue_PamlM7M8[tab$pVal.M8vsM7<0.05 & tab$PValue_PamlM7M8>0.05],
       col="green", pch=20)

legend("topleft", c("<0.05 in PamlM7M8 but >0.05 in Janet M8vsM7", "<0.05 in Janet M8v
      pch=20, col=c("red", "green"))
```



Quels sont les gènes en couleur:

```
na.omit(tab[(tab$pVal.M8vsM7>0.05 & tab$PValue_PamIM7M8<0.05),c("Gene.name", "pVal.M8vsM7", "PValue_PamIM7M8", "whole.gene.dN.dS.model.0")])
```

##	Gene.name	pVal.M8vsM7	PValue_PamIM7M8	whole.gene.dN.dS.model.0
## 51	CIT	0.103170	1.854024e-02	0.03889
## 101	FBN2	0.174750	2.253070e-08	0.06871
## 158	MARK1	0.062265	1.890420e-02	0.08147
## 196	NUP88	0.052061	4.950260e-02	0.19123
## 316	UBXN8	0.053229	3.945009e-02	0.50084
## 322	VPS11	0.108710	3.061568e-02	0.04236
##	Omega_PamIM7M8			
## 51		0.04325399		

```
## 101      0.07233605
## 158      0.08802245
## 196      0.20601208
## 316      0.50718198
## 322      0.04780560

na.omit(tab[(tab$pVal.M8vsM7<0.05 & tab$PValue_PamlM7M8>0.05),c("Gene.name", "pVal.M8vsM7", "PValue_PamlM7M8", "whole.gene.dN.dS.model.0")])

##      Gene.name pVal.M8vsM7 PValue_PamlM7M8 whole.gene.dN.dS.model.0
## 68      DCTPP1   0.0431830      0.10613016      0.29992
## 181     NDUF9B   0.0024264      0.05119297      0.29487
## 188     NLRX1    0.0463220      0.13614538      0.17885
## 197     NUP98    0.0345210      0.98219934      0.17017
## 284     STOM     0.0345710      0.19872467      0.16126
##      Omega_PamlM7M8
## 68      0.3538646
## 181     0.3104234
## 188     0.2159544
## 197     0.1772109
## 284     0.1477986
```

Focus sur le gène CIT pour lequel la différence est vraiment assez importante:

```
dginn[dginn$Gene=="CIT",]

##      Gene      Omega  Method PosSel      PValue NbSites
## 1201  CIT 0.04325399  BUSTED      N 1.000000e+00      NA
## 1202  CIT 0.04325399  BppM1M2      N 9.999983e-01      NA
## 1203  CIT 0.04325399  BppM7M8      Y 2.254251e-05      11
## 1204  CIT 0.04325399  PamlM1M2      N 1.000000e+00      NA
## 1205  CIT 0.04325399  PamlM7M8      Y 1.854024e-02      0
## 1206  CIT 0.04325399      MEME      NA      1
##                                           PSS
## 1201
## 1202
## 1203 258, 8, 1835, 304, 369, 338, 434, 625, 151, 410, 255
## 1204
## 1205
## 1206                                           410
```

```

tab[tab$Gene.name=="CIT",1:20]

##      Gene.name PreyGene PreyGene_JYname BaitShort                                list
## 51          CIT          CIT              CIT      nsp13 list26_COV_list4dataset2nonOrf
##                                     description                                other.names
## 51 citron rho-interacting serine/threonine kinase CITK|CRIK|MCPH17|STK21
##      top40_posSeln Num.primate.seqs Alignment.length..nucleotides.
## 51              no              24              6210
##      Alignment.length..codons. whole.gene.dN.dS.model.0 total.tree.length
## 51              2070              0.03889              0.33654
##      total.dN.tree.length total.dS.tree.length p.value.M8vsM8a..raw.
## 51              0.014              0.3603              0.99887
##      p.value.M8vsM8a..BH.corrected. pVal.M8vsM7 pVal.M8vsM7.adj pVal.M2vsM1
## 51              1      0.10317      0.3747212      1

```

1.4.3 Concordance est méthodes

Est-ce que les gènes avec une faible p-value sont détecté par 1,2,3,4 ou 5 méthodes en général?

```

nontab<-tab[tab$pVal.M8vsM7>=0.05,c("Gene.name", "PosSel_PamlM1M2", "PosSel_PamlM7M8",
"PosSel_BppM7M8", "PosSel_BUSTED")]

non<-apply(nontab, 1, function(x) sum(x=="Y"))

ouitab<-tab[tab$pVal.M8vsM7<0.05,c("Gene.name", "PosSel_PamlM1M2", "PosSel_PamlM7M8",
"PosSel_BppM7M8", "PosSel_BUSTED")]

oui<-apply(ouitab, 1, function(x) sum(x=="Y"))

stripchart(x=list(oui, non), method="jitter", jitter=0.2,
               vertical=T, pch=20, cex=0.5,
               group.names=c("Yes Janet", "No Janet"),
               ylab="Nb YES from dginn")

```

