

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

Marie Cariou

October 2020

## Contents

<b>1</b>	<b>Data</b>	<b>2</b>
<b>2</b>	<b>Comparisons Primates</b>	<b>2</b>
2.1	Janet Young's results (Young-primate) VS DGINN-full's results	2
2.2	Janet Young's results (Young-primate) VS Cooper's result . .	4
2.3	Cooper's results (Cooper-primate) VS DGINN-full's results .	6
<b>3</b>	<b>Overlap</b>	<b>8</b>
3.1	Mondrian . . . . .	8
3.2	subsetR . . . . .	11
<b>4</b>	<b>Gene List</b>	<b>13</b>
<b>5</b>	<b>Shiny like</b>	<b>15</b>

# 1 Data

Analysis were formatted by the script covid\_comp\_script0\_table.Rnw.

```
home<-" /home/adminmarie/Documents/"
workdir<-paste0(home, "CIRI_BIBS_projects/2020_05_Etienne_covid/")

tab<-read.delim(paste0(workdir,
  "covid_comp/covid_comp_complete.txt"), h=T, sep="\t")
dim(tab)

## [1] 332 141

tab$Gene.name<-as.character(tab$Gene.name.x)
tab$Gene.name[tab$PreyGene=="MTARC1"]<-"MTARC1"
```

## 2 Comparisons Primates

### 2.1 Janet Young's results (Young-primate) VS DGINN-full's results

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

```
tab$dginn.primate_omegaMOBpp[tab$dginn.primate_omegaMOBpp=="na"]<-NA
tab$dginn.primate_omegaMOBpp<-as.numeric(as.character(
  tab$dginn.primate_omegaMOBpp))

plot(tab$whole.gene.dN.dS.model.0,
  tab$dginn.primate_omegaMOBpp,
  xlab="Omega Young-primate",
  ylab="DGINN-full's",
  cex=0.3)
abline(0,1)
abline(lm(tab$dginn.primate_omegaMOBpp~tab$whole.gene.dN.dS.model.0),
  col="red")

outlier<-tab[tab$whole.gene.dN.dS.model.0<0.4 &
  tab$dginn.primate_omegaMOBpp>0.5,]
```

```

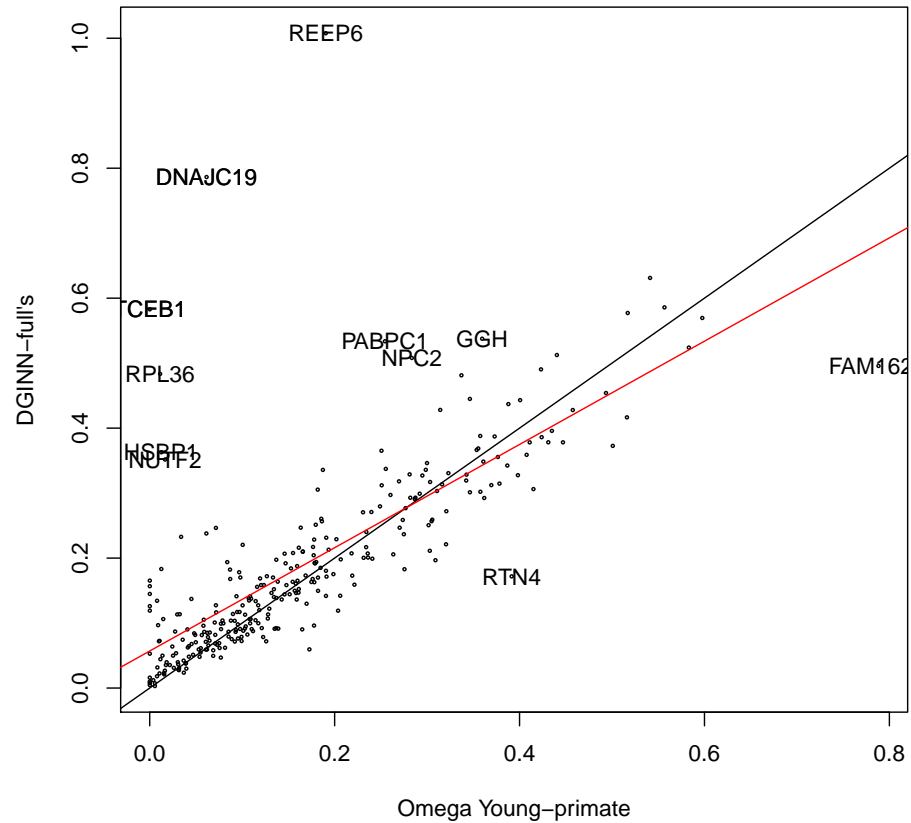
text(x=outlier$whole.gene.dN.dS.model.0,
y=outlier$dginn.primate_omegaMOBpp,
outlier$Gene.name)

outlier<-tab[tab$whole.gene.dN.dS.model.0<0.1 &
             tab$dginn.primate_omegaMOBpp>0.3,]
text(x=outlier$whole.gene.dN.dS.model.0,
y=outlier$dginn.primate_omegaMOBpp,
outlier$Gene.name)

outlier<-tab[tab$whole.gene.dN.dS.model.0>0.33 &
             tab$dginn.primate_omegaMOBpp<0.2,]
text(x=outlier$whole.gene.dN.dS.model.0,
y=outlier$dginn.primate_omegaMOBpp,
outlier$Gene.name)

outlier<-tab[tab$whole.gene.dN.dS.model.0>0.6 &
             tab$dginn.primate_omegaMOBpp<0.6,]
text(x=outlier$whole.gene.dN.dS.model.0,
y=outlier$dginn.primate_omegaMOBpp,
outlier$Gene.name)

```



## 2.2 Janet Young's results (Young-primate) VS Cooper's result

Comparaison des Omega: colonne L "whole.gene.dN.dS.model.0" VS colonne "cooper.primates.Average\_dNdS".

```
tab$cooper.primates.Average_dNdS<-as.numeric(as.character(
  tab$cooper.primates.Average_dNdS))

plot(tab$whole.gene.dN.dS.model.0,
      tab$cooper.primates.Average_dNdS,
      xlab="Omega Young-primate",
```

```

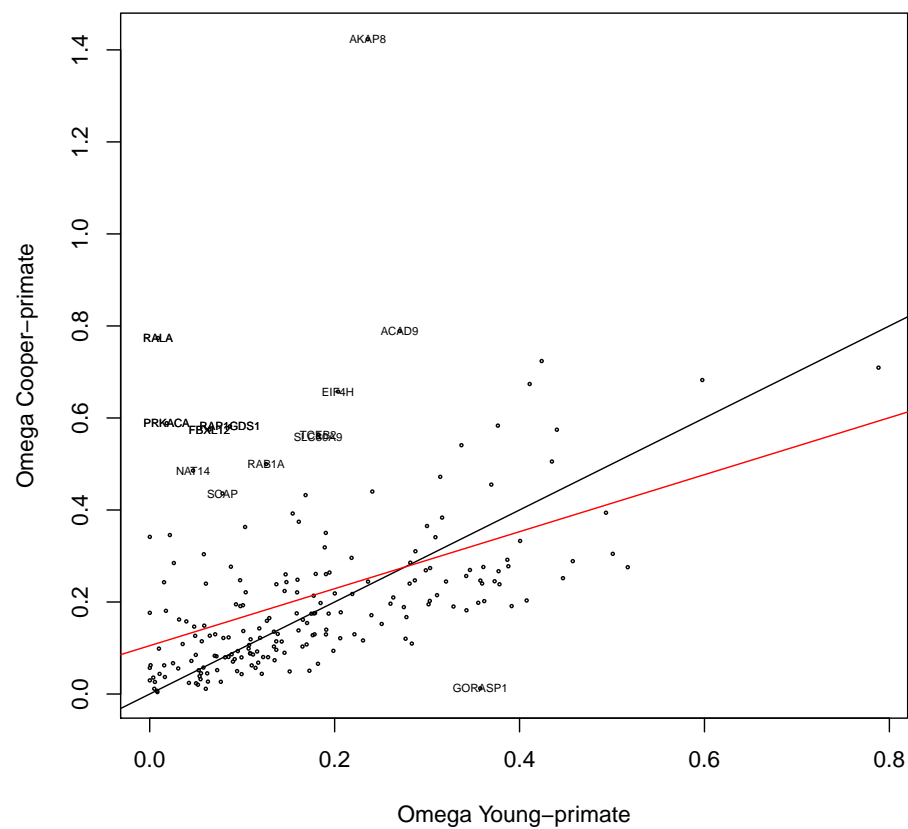
        ylab="Omega Cooper-primate",
        cex=0.3)
abline(0,1)
abline(lm(tab$cooper.primates.Average_dNdS~tab$whole.gene.dN.dS.model.0),
        col="red")

outlier<-tab[tab$whole.gene.dN.dS.model.0<0.15 &
             tab$cooper.primates.Average_dNdS>0.4,]
text(x=outlier$whole.gene.dN.dS.model.0,
     y=outlier$cooper.primates.Average_dNdS,
     outlier$Gene.name, cex=0.5)

outlier<-tab[tab$whole.gene.dN.dS.model.0<0.3 &
             tab$cooper.primates.Average_dNdS>0.5,]
text(x=outlier$whole.gene.dN.dS.model.0,
     y=outlier$cooper.primates.Average_dNdS,
     outlier$Gene.name, cex=0.5)

outlier<-tab[tab$whole.gene.dN.dS.model.0>0.3 &
             tab$cooper.primates.Average_dNdS<0.1,]
text(x=outlier$whole.gene.dN.dS.model.0,
     y=outlier$cooper.primates.Average_dNdS,
     outlier$Gene.name, cex=0.5)

```



### 2.3 Cooper's results (Cooper-primare) VS DGINN-full's results

Comparaison des Omega: colonne "cooper.primates.Average\_dNdS" VS colonne "omega" dans la sortie de dginn.

```
plot(tab$cooper.primates.Average_dNd,
     tab$dginn.primare_omegaMOBpp,
     xlab="Omega Cooper-primare",
     ylab="DGINN-full's",
     cex=0.3)
abline(0,1)
```

```

abline(lm(tab$dginn.primate_omegaMOBpp~tab$cooper.primates.Average_dNd), col="red")

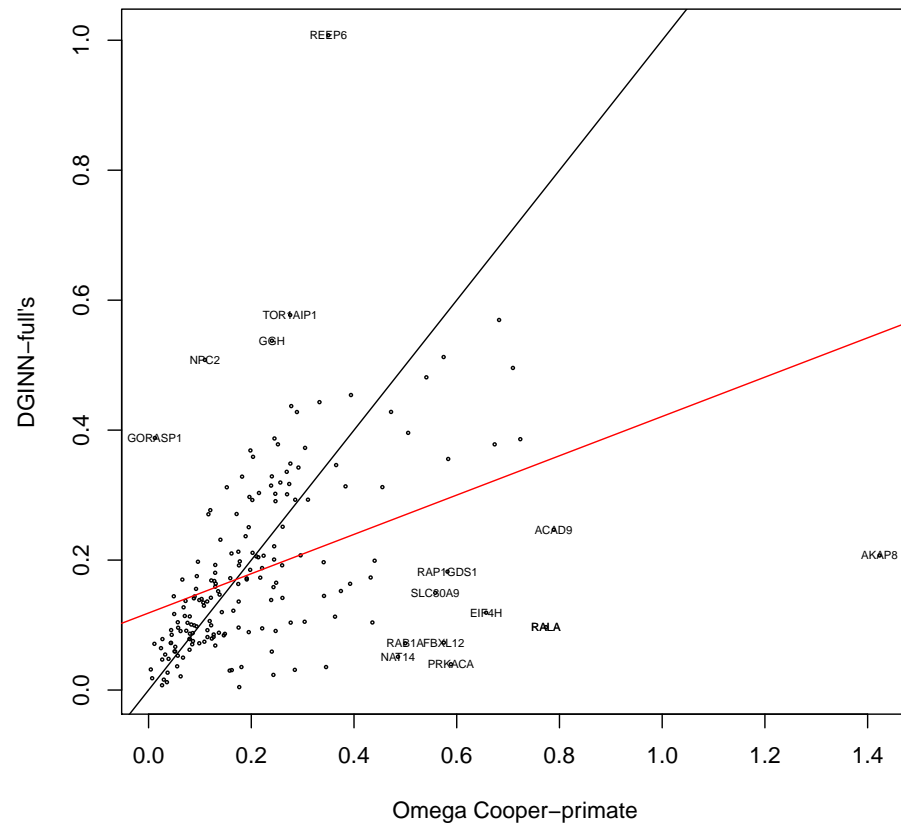
outlier<-tab[tab$cooper.primates.Average_dNd<0.4 &
             tab$dginn.primate_omegaMOBpp>0.5,]
text(x=outlier$cooper.primates.Average_dNd,
     y=outlier$dginn.primate_omegaMOBpp,
     outlier$Gene.name, cex=0.5)

outlier<-tab[tab$cooper.primates.Average_dNd<0.1 &
             tab$dginn.primate_omegaMOBpp>0.3,]
text(x=outlier$cooper.primates.Average_dNd,
     y=outlier$dginn.primate_omegaMOBpp,
     outlier$Gene.name, cex=0.5)

outlier<-tab[tab$cooper.primates.Average_dNd>0.7 &
             tab$dginn.primate_omegaMOBpp<0.3,]
text(x=outlier$cooper.primates.Average_dNd,
     y=outlier$dginn.primate_omegaMOBpp,
     outlier$Gene.name, cex=0.5)

outlier<-tab[tab$cooper.primates.Average_dNd>0.45 &
             tab$dginn.primate_omegaMOBpp<0.2,]
text(x=outlier$cooper.primates.Average_dNd,
     y=outlier$dginn.primate_omegaMOBpp,
     outlier$Gene.name, cex=0.5)

```



### 3 Overlap

#### 3.1 Mondrian

```
library(Mondrian)

monddata<-as.data.frame(tab$Gene.name)
dim(monddata)

## [1] 332 1

dginnfulltmp<-rowSums(cbind(tab$dginn.primite_BUSTED=="Y",
```



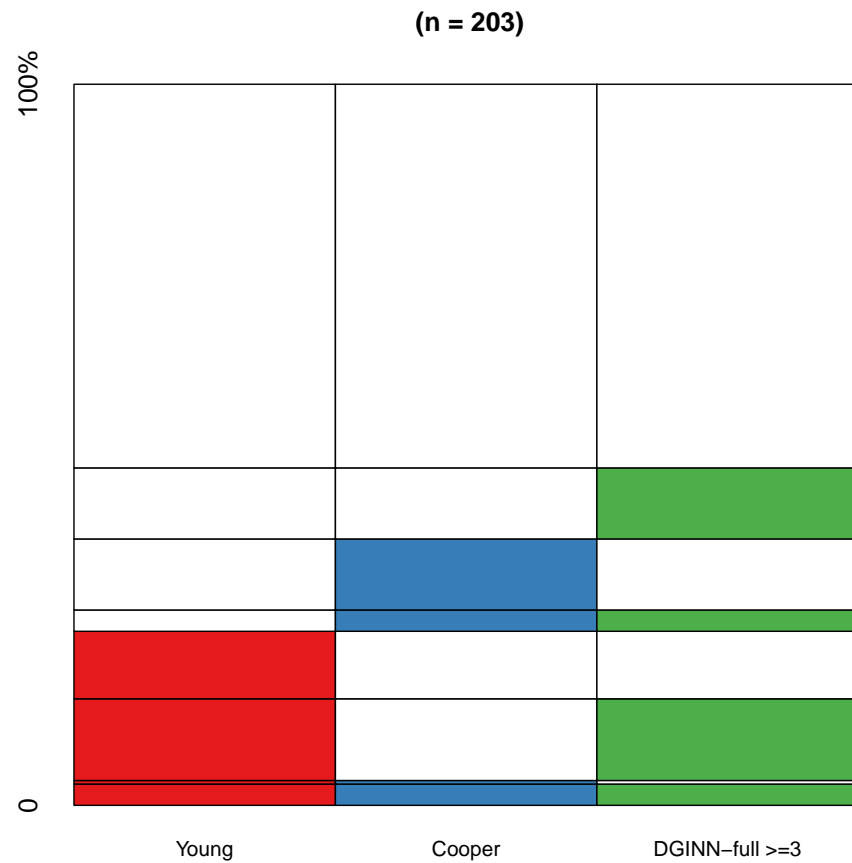
```

                                tab$dginn.primate_BppM1M2=="Y",
                                tab$dginn.primate_BppM7M8=="Y",
                                tab$dginn.primate_codemlM1M2=="Y",
                                tab$dginn.primate_codemlM7M8=="Y"))

monddata$primates_young<-ifelse(
  tab$pVal.M8vsM7<0.05, 1, 0)
monddata$primate_cooper<-ifelse(
  tab$cooper.primates.M7.M8_p_value<0.05, 1, 0)
monddata$primates_dginn_full<-ifelse(
  dginnfulltmp>=3, 1,0)

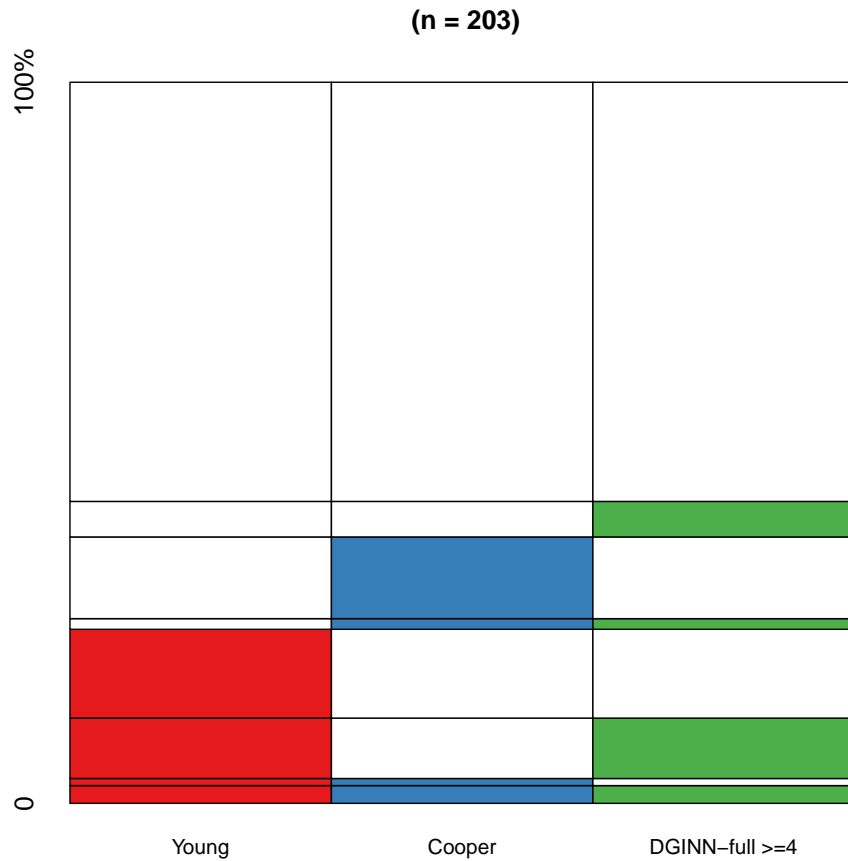
mondrian(na.omit(monddata[,2:4]),
          labels=c("Young", "Cooper", "DGINN-full >=3" ))

```



```
monddata$primates_dginn_full<-ifelse(
  dginnfulltmp>=4, 1,0)

mondrian(na.omit(monddata[,2:4]),
  labels=c("Young", "Cooper", "DGINN-full >=4"))
```



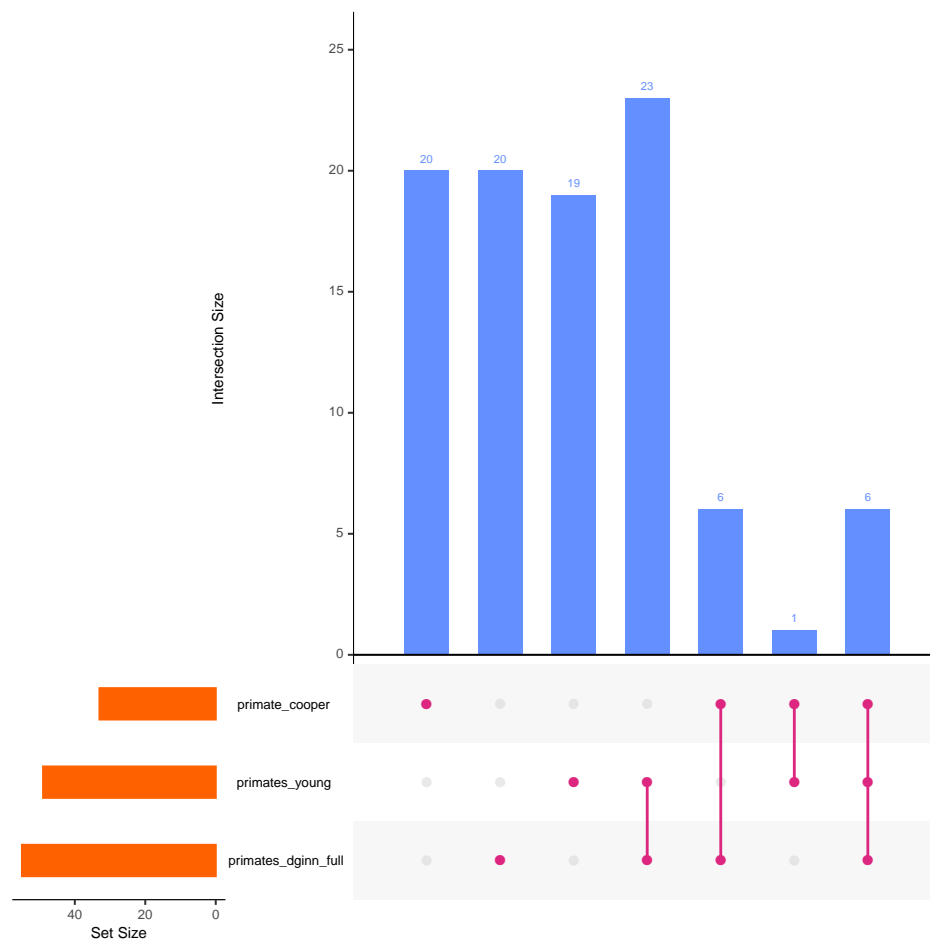
### 3.2 subsetR

Just another representation of the same result.

```
library(UpSetR)
upsetdata<-as.data.frame(tab$Gene.name)

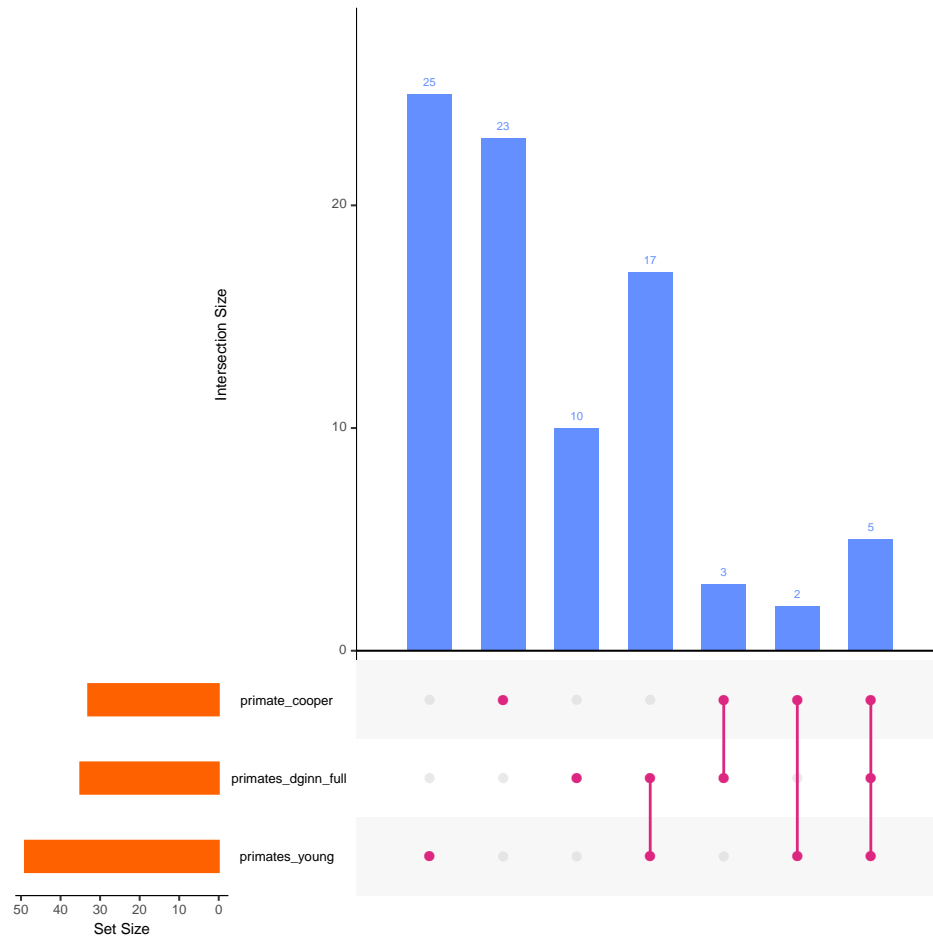
upsetdata$primates_young<-ifelse(tab$pVal.M8vsM7<0.05, 1, 0)
upsetdata$primate_cooper<-ifelse(
  tab$cooper.primates.M7.M8_p_value<0.05, 1, 0)
upsetdata$primates_dginn_full<-ifelse(dginnfulltmp>=3, 1,0)
```

```
upset(na.omit(upsetdata), nsets = 3, matrix.color = "#DC267F",
      main.bar.color = "#648FFF", sets.bar.color = "#FE6100")
```



```
###
upsetdata$primates_dginn_full<-ifelse(dginnfulltmp>=4, 1,0)

upset(na.omit(upsetdata), nsets = 3, matrix.color = "#DC267F",
      main.bar.color = "#648FFF", sets.bar.color = "#FE6100")
```



## 4 Gene List

Genes under positive selection for at least 4 methods.

```
dginnfulltmp<-rowSums(cbind(tab$dginn.primate_BUSTED=="Y",
  tab$dginn.primate_BppM1M2=="Y",
  tab$dginn.primate_BppM7M8=="Y",
  tab$dginn.primate_codemlM1M2=="Y",
  tab$dginn.primate_codemlM7M8=="Y"))

tab$Gene.name[dginnfulltmp>=4 & is.na(dginnfulltmp)==F]
```

```
## [1] "ACADM"      "BCS1L"      "BRD4"      "CDK5RAP2"   "CEP135"
## [6] "CEP68"      "CLIP4"      "DNMT1"     "DPH5"       "EMC1"
## [11] "ER01LB"     "FYC01"      "GCC2"      "GGH"        "GHITM"
## [16] "GIGYF2"     "GLA"        "GOLGA7"    "HECTD1"     "IDE"
## [21] "ITGB1"      "LARP1"      "LARP4B"    "LMAN2"      "MARK1"
## [26] "MIPOL1"     "MPHOSPH10" "MYCBP2"    "NDUFAF2"    "NDUFB9"
## [31] "NUPL1"      "PCNT"       "POLA1"     "PRIM2"      "PRKAR2A"
## [36] "PVR"        "REEP6"      "RIPK1"     "SAAL1"      "SEPSECS"
## [41] "SIRT5"      "SLC25A21"   "SLC27A2"   "TMEM39B"    "TOR1AIP1"
## [46] "TUBGCP2"    "UBAP2"      "UGGT2"     "VPS39"      "ZNF318"

tab$Gene.name[dginnfulltmp>=3 & is.na(dginnfulltmp)==F]

## [1] "ACADM"      "ADAM9"      "AP2A2"      "ATE1"       "BCS1L"
## [6] "BRD4"      "BZW2"      "CDK5RAP2"   "CEP135"     "CEP68"
## [11] "CLIP4"     "CNTRL"      "DNMT1"     "DPH5"       "EDEM3"
## [16] "EIF4E2"    "EMC1"       "ER01LB"    "EXOSC2"     "FYC01"
## [21] "GCC2"      "GGH"        "GHITM"     "GIGYF2"     "GLA"
## [26] "GOLGA7"    "GOLGB1"     "GORASP1"   "HDAC2"      "HECTD1"
## [31] "HS6ST2"    "IDE"        "ITGB1"     "LARP1"      "LARP4B"
## [36] "LARP7"     "LMAN2"      "MARK1"     "MDN1"       "MIPOL1"
## [41] "MOV10"     "MPHOSPH10" "MRPS5"     "MYCBP2"     "NAT14"
## [46] "NDUFAF2"   "NDUFB9"     "NGLY1"     "NPC2"       "NUPL1"
## [51] "PCNT"      "PITRM1"     "PLAT"      "PLOD2"      "PMPCB"
## [56] "POLA1"     "POR"        "PRIM2"     "PRKAR2A"    "PTBP2"
## [61] "PVR"       "RAB14"      "RAB1A"     "RAB2A"      "RAP1GDS1"
## [66] "RBX1"      "REEP6"      "RIPK1"     "RPL36"      "SAAL1"
## [71] "SCCPDH"    "SEPSECS"    "SIRT5"     "SLC25A21"   "SLC27A2"
## [76] "STOM"      "TIMM8B"     "TMEM39B"   "TOR1AIP1"   "TRIM59"
## [81] "TRMT1"     "TUBGCP2"    "UBAP2"     "UGGT2"      "USP54"
## [86] "VPS39"     "ZNF318"

tmp<-tab[dginnfulltmp>=4 & is.na(dginnfulltmp)==F,
c("Gene.name", "dginn.primate_BUSTED", "dginn.primate_BppM1M2",
  "dginn.primate_BppM7M8", "dginn.primate_codemlM1M2", "dginn.primate_codemlM7M8")]

write.table(tmp, "geneList_DGINN_full_primate_pos4.txt", row.names=F, quote=F)
```

## 5 Shiny like

```
makeFig1 <- function(df){  
  
  # prepare data for colors etc  
  colMethods <- c("deepskyblue4", "darkorange", "deepskyblue3", "mediumseagreen",  
  nameMethods <- c("BUSTED", "BppM1M2", "BppM7M8", "codemlM1M2", "codemlM7M8", "MEME"  
  metColor <- data.frame(Name = nameMethods, Col = colMethods, stringsAsFactors = F)  
  
  # subset for this specific figure  
  #df <- df[df$nbY >= 1, ] # to drop genes found by 0 methods (big datasets)  
  xt <- df[, c("BUSTED", "BppM1M2", "BppM7M8", "codemlM1M2", "codemlM7M8")]  
  xt$Gene <- df$Gene  
  nbrMeth <- 5  
  # reverse order of dataframe so that genes with the most Y are at the bottom (to be  
  xt[,1:5] <- ifelse(xt[,1:5] == "Y", 1, 0)  
  # sort and Filter the 0 lines  
  xt<-xt[order(rowSums(xt[,1:5])),]  
  xt<-xt[rowSums(xt[,1:5])>2,]  
  
  row.names(xt)<-xt$Gene  
  xt<-xt[,1:5]  
  
  colFig1 <- metColor[which(metColor$Name %in% colnames(xt)), ]  
  
  ##### PART 1 : NUMBER OF METHODS  
  par(xpd = NA, mar=c(2,7,4,0), oma = c(0,0,0,0), mgp = c(3,0.3,0))  
  
  h = barplot(  
    t(xt),  
    border = NA,  
    axes = F,  
    col = adjustcolor(colFig1$Col, alpha.f = 1),  
    horiz = T,  
    las = 2,  
    main = "Methods detecting positive selection",  
    cex.main = 0.85,  
    cex.names = min(50/nrow(xt), 1.5)  
  )
```

```

axis(3, line = 0, at = c(0:nbrMeth), label = c("0", rep("", nbrMeth - 1), nbrMeth),

legend("bottomleft",
      horiz = T,
      border = colFig1$Col,
      legend = colFig1$Name,
      fill = colFig1$Col,
      cex = 0.8,
      bty = "n",
      xpd = NA
)
}

```

```
source("covid_comp_shiny.R")
```

```

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

```

```
names(df)
```

```

## [1] "File"          "Name"          "Gene"
## [4] "GeneSize"      "NbSpecies"     "omegaM0Bpp"
## [7] "omegaM0codeml" "BUSTED"        "BUSTED.p.value"
## [10] "MEME.NbSites"  "MEME.PSS"      "BppM1M2"
## [13] "BppM1M2.p.value" "BppM1M2.NbSites" "BppM1M2.PSS"
## [16] "BppM7M8"       "BppM7M8.p.value" "BppM7M8.NbSites"
## [19] "BppM7M8.PSS"   "codemlM1M2"     "codemlM1M2.p.value"
## [22] "codemlM1M2.NbSites" "codemlM1M2.PSS" "codemlM7M8"
## [25] "codemlM7M8.p.value" "codemlM7M8.NbSites" "codemlM7M8.PSS"

```

```

dftmp<-tab[,c("File", "Name", "Gene.name",
"GeneSize", "dginn.primite_NbSpecies", "dginn.primite_omegaM0Bpp",
"dginn.primite_omegaM0codeml", "dginn.primite_BUSTED", "dginn.primite_BUSTED.p",
"dginn.primite_MEME.NbSites", "dginn.primite_MEME.PSS", "dginn.primite_BppM1M2",
"dginn.primite_BppM1M2.p.value", "dginn.primite_BppM1M2.NbSites", "dginn.primite_BppM1M2.PSS",
"dginn.primite_BppM7M8", "dginn.primite_BppM7M8.p.value", "dginn.primite_BppM7M8.NbSites", "dginn.primite_BppM7M8.PSS")]

```



```
"dginn.primate_BppM7M8.PSS", "dginn.primate_codemlM1M2", "dginn.primate_codemlM1M2.NbSites", "dginn.primate_codemlM1M2.PSS", "dginn.primate_codemlM7M8.p.value", "dginn.primate_codemlM7M8.NbSites" , "dginn.primate_codemlM7M8.PSS"
```

```
names(dftmp)<-names(df)
makeFig1(dftmp)
```

