

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

Marie Cariou

October 2020

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	Write new table	5
2	Comparisons Primates	5
2.1	DGINN results on Janet Young's alignments (DGINN-Young-primate) VS Janet Young's results	5
2.2	DGINN results on Janet Young's alignments (DGINN-Young-primate) VS DGINN-full's results	6
2.3	Janet Young's results (Young-primate) VS DGINN-full's results	8
3	Overlap	9
3.1	Mondrian	9
3.2	subsetR	13
4	Gene List	15
5	Shiny like	16

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

#names(tab)
```

1.2 Read DGINN Young table

DGINN-Young-primate table correspond to DGINN results, on the SAME alignment as Young-primate.

I will merge the 2 tables.

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

dim(dginnY)

## [1] 1992 7

names(dginnY)

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

1.3 Joining Young and DGINN Young table

I hide some code corresponding to verifications of gene names coherence between tables

```

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"
## [7] "omegaM0codeml" "BUSTED" "BUSTED.p.value" "MEME.NbSites"
## [13] "BppM1M2.p.value" "BppM1M2.NbSites" "BppM1M2.PSS" "BppM7M8"
## [19] "BppM7M8.PSS" "codemlM1M2" "codemlM1M2.p.value" "codemlM1M2.NbSites"
## [25] "codemlM7M8.p.value" "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
## [8] CEP43 COQ8B COQ8A CSNK2A1
## [15] DDX21[0-717] DDX21[716-2538] DDX50 DNAJC15
## [22] ELOC ERO1B EXOSC3[0-1446] EXOSC3[1445-1980]
## [29] GNB3 GOLGA7[0-312] GOLGA7[311-549] GPX1[0-1218]
## [36] IMPDH1 ITGB1[0-2328] ITGB1[2327-2844] LMAN2L
## [43] MGRN1 NDFIP2[0-768] NDFIP2[767-1314] NDUFAF2[0-258]
## [50] NUP58[0-1824] NUP58[1823-2367] PABPC3 POTPABPC1
## [57] PRIM2[0-1071] PRIM2[1070-1902] PRKACB PRKACG
## [64] RAB13 RAB18[0-855] RAB18[854-1815] RAB2B
## [71] RALB EZR EZR[0-1458] EZR[1457-3771]
## [78] RHOC SLC44A2[0-2577] SLC44A2[2576-3657] SPART
## [85] STOM[1046-1800] STOML3 TIMM29 TLE4
## [92] TMRSS2 TOMM70 TOR1B WASHC4
## 411 Levels: AAR2 AASS AATF ABCC1 ACAD9 ACADM ACE2 ACSL3 ADAM9 ADAM9[0-3120] ADAM9[3119-3927]

# 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"
## [13] "SPG20" "TCEB1" "TCEB2" "TOMM70A" "USP13" "VIMP" "WHSC1"

names(dginnT)<-c("File", "Name", "Gene.name", "GeneSize", "dginn-primate_NbSpecies",
"dginn-primate_omegaM0codeml", "dginn-primate_BUSTED", "dginn-primate_BUSTED.p.value")
```

```
"dginn-primate_MEME.NbSites", "dginn-primate_MEME.PSS", "dginn-primate_BppM1M2.p.value", "dginn-primate_BppM1M2.NbSites", "dginn-primate_BppM7M8", "dginn-primate_BppM7M8.p.value", "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"
```

1.5 Join Table and DGINN table

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

1.6 Write new table

```
write.table(tab,
            "COVID_PAMLresults_332hits_plusBatScreens_plusDGINN_20201014.txt",
            row.names=F, quote=F, sep="\t")
```

2 Comparisons Primates

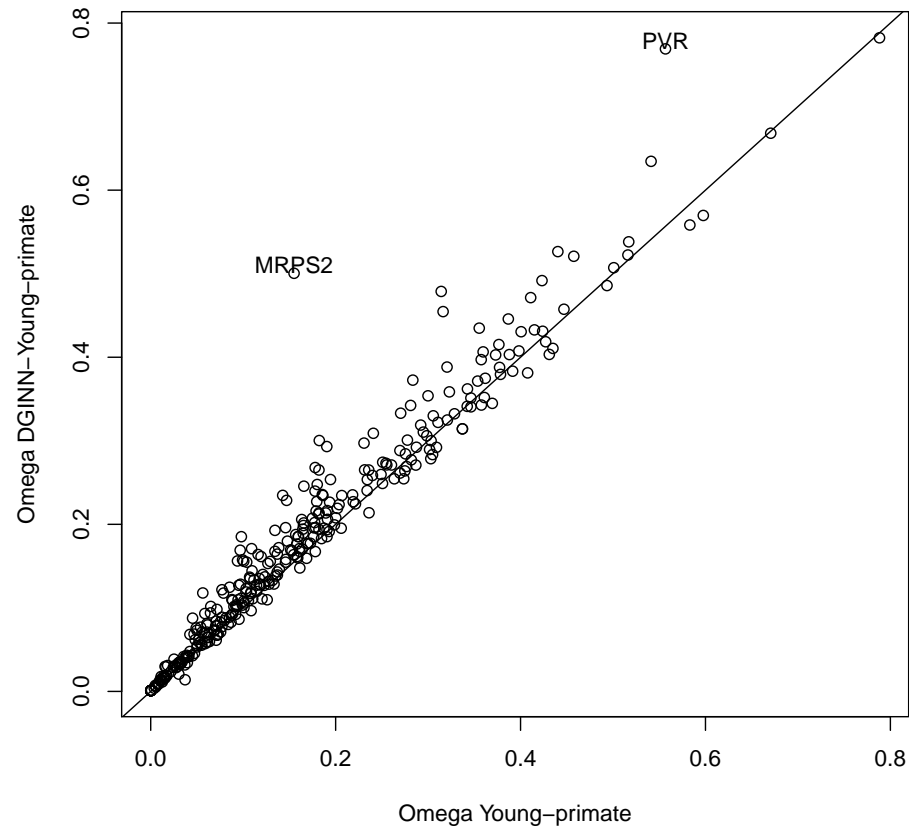
2.1 DGINN results on Janet Young's alignments (DGINN-Young-primate) VS Janet Young's results

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 Young-primate", ylab="Omega DGINN-Young-primate")
abline(0,1)
outlier<-tab[tab$whole.gene.dN.dS.model.0<0.2 & tab$Omega_PamlM7M8>0.4,]
text(x=outlier$whole.gene.dN.dS.model.0,
     y=(outlier$Omega_PamlM7M8+0.01),
     outlier$Gene.name)

outlier<-tab[tab$whole.gene.dN.dS.model.0<0.6 & tab$Omega_PamlM7M8>0.7,]
text(x=outlier$whole.gene.dN.dS.model.0,
```

```
y=(outlier$Omega_PamlM7M8+0.01),
outlier$Gene.name)
```



2.2 DGINN results on Janet Young's alignments (DGINN-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$'dginn-primate_omegaM0Bpp'<-as.numeric(as.character(tab$'dginn-primate_omegaM0Bpp'))
## Warning:  NAs introduits lors de la conversion automatique
```

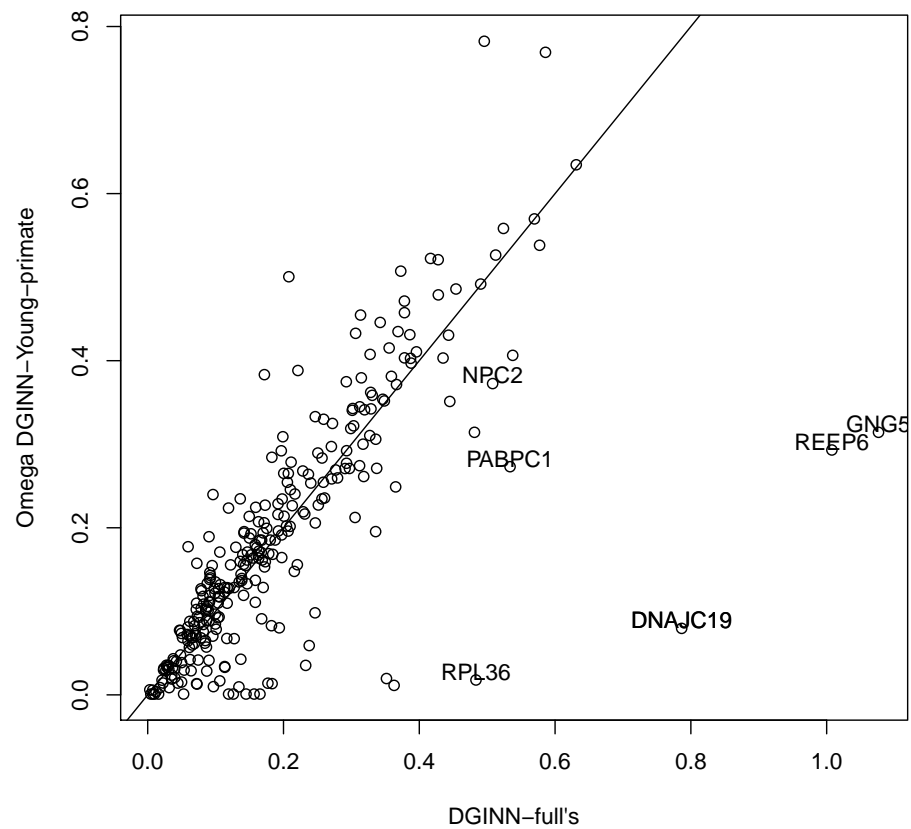
```

plot(tab$'dginn-primate_omegaM0Bpp', tab$Omega_PamlM7M8,
      xlab="DGINN-full's", ylab="Omega DGINN-Young-primate")
abline(0,1)

outlier<-tab[tab$'dginn-primate_omegaM0Bpp'>0.4 & tab$Omega_PamlM7M8<0.2,]
text(x=outlier$'dginn-primate_omegaM0Bpp',
     y=(outlier$Omega_PamlM7M8+0.01),
     outlier$Gene.name)

outlier<-tab[tab$'dginn-primate_omegaM0Bpp'>0.5 & tab$Omega_PamlM7M8<0.4,]
text(x=outlier$'dginn-primate_omegaM0Bpp',
     y=(outlier$Omega_PamlM7M8+0.01),
     outlier$Gene.name)

```



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

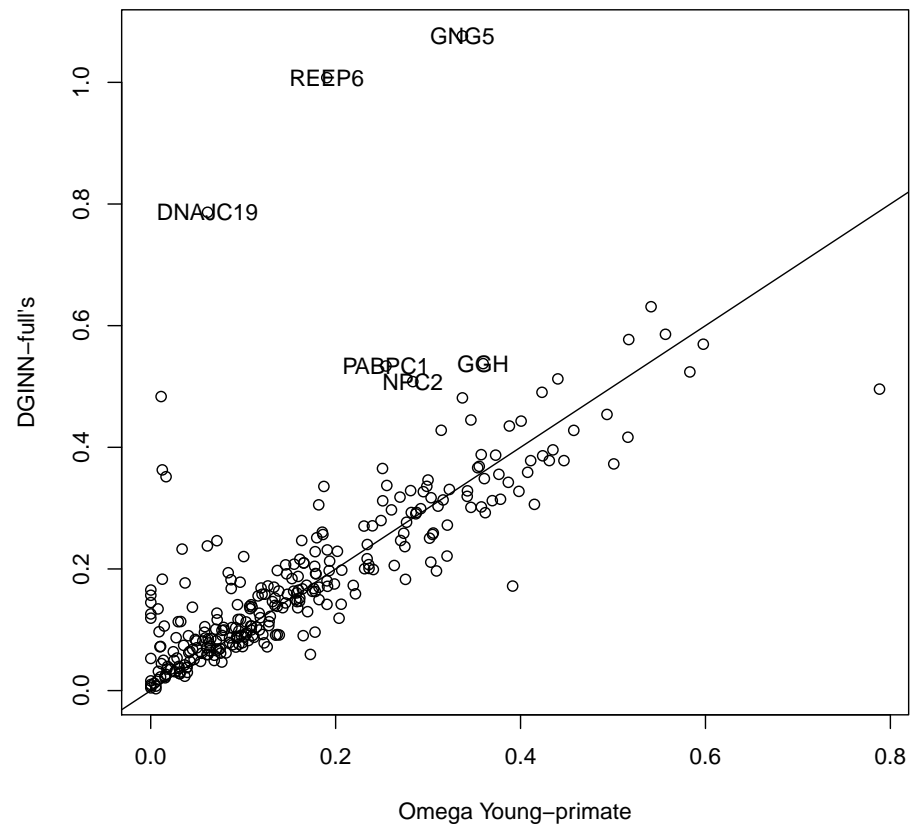
Comparaison des Omega: colonne L "whole.gene.dN.dS.model.0" VS colonne "omega" dans la sortie de dginn.

```
plot(tab$whole.gene.dN.dS.model.0, as.numeric(as.character(tab$'dginn-primate_omegaMOBpp')),
      xlab="Omega Young-primate", ylab="DGINN-full's")
abline(0,1)
```

```
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)
```



3 Overlap

3.1 Mondrian

```
library(Mondrian)  
  
#####
```

```

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

## [1] 333 1

dginnyoungtmp<-rowSums(cbind(tab$PosSel_PamlM1M2=="Y", tab$PosSel_PamlM7M8=="Y",
tab$PosSel_BppM1M2=="Y", tab$PosSel_BppM7M8=="Y", tab$PosSel_BUSTED=="Y"))

#monddata$primates_dginn_young<-ifelse(tmp$PosSel_PamlM7M8=="Y", 1,0)

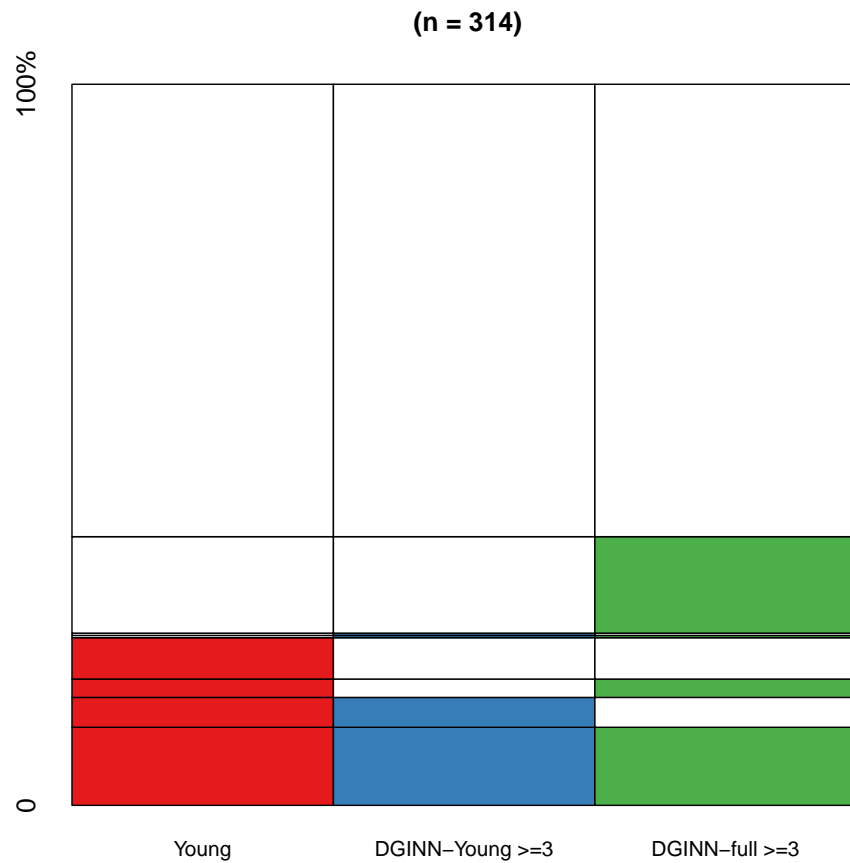
dginfulltmp<-rowSums(cbind(tab$'dginn-primate_BUSTED'=="Y", tab$'dginn-primate_BppM1
tab$'dginn-primate_BppM7M8'=="Y", tab$'dginn-primate_codemlM1M2'=="Y", tab$'dginn-pri

monddata$primates_young<-ifelse(tab$pVal.M8vsM7<0.05, 1, 0)
#monddata$primates_cooper<-ifelse(tab$cooper.primates.M7.M8_p_val<0.05, 1, 0)

monddata$primates_dginn_young<-ifelse(dginnyoungtmp>=3, 1,0)
monddata$primates_dginn_full<-ifelse(dginfulltmp>=3, 1,0)

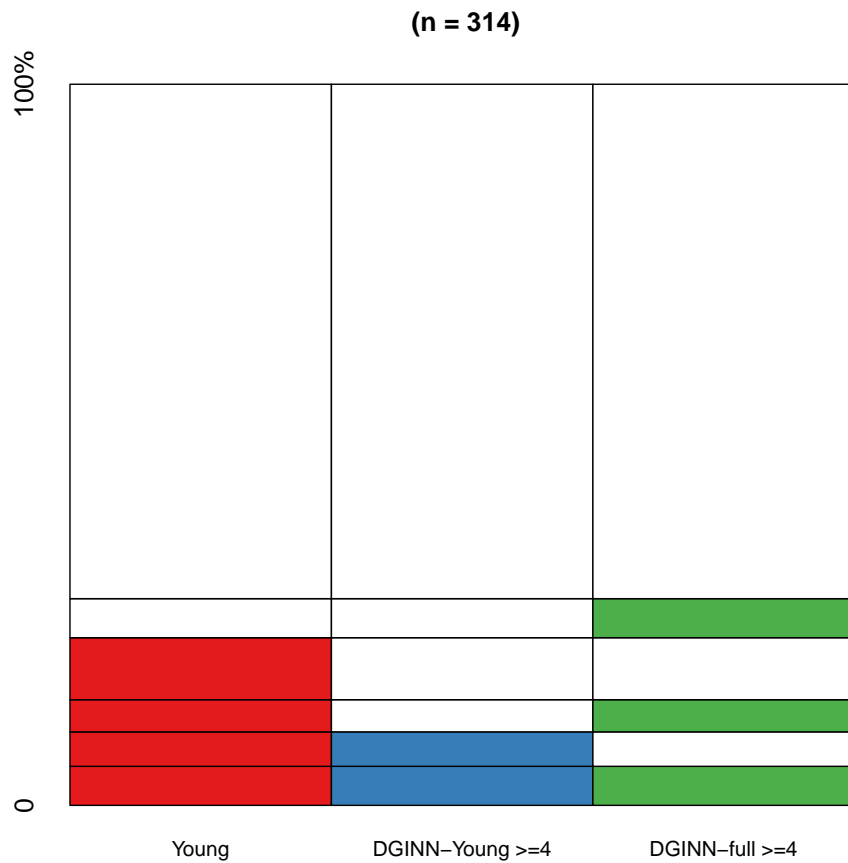
mondrian(na.omit(monddata[,2:4]), labels=c("Young", "DGINN-Young >=3", "DGINN-full >=

```



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

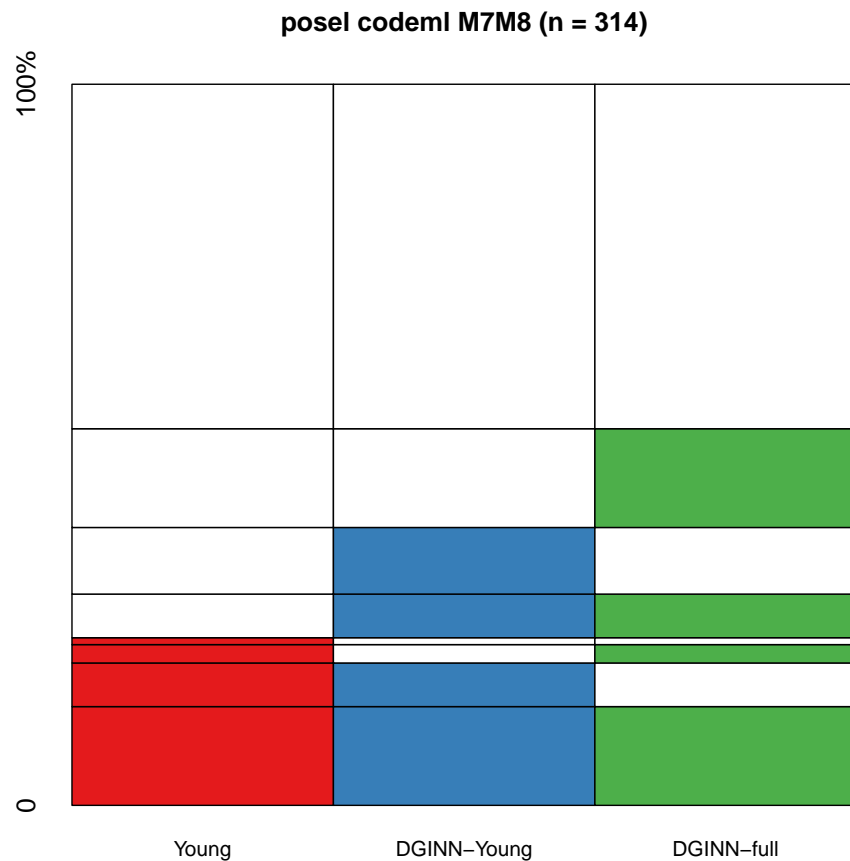
mondrian(na.omit(monddata[,2:4]), labels=c("Young", "DGINN-Young >=4", "DGINN-full >=4"))
```



Comparison of results with the same method.

```
#####
monddata$primates_dginn_young<-tab$PosSel_BppM7M8=="Y"
monddata$primates_dginn_full<-tab$'dginn-primate_codemlM7M8'=="Y"

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



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)

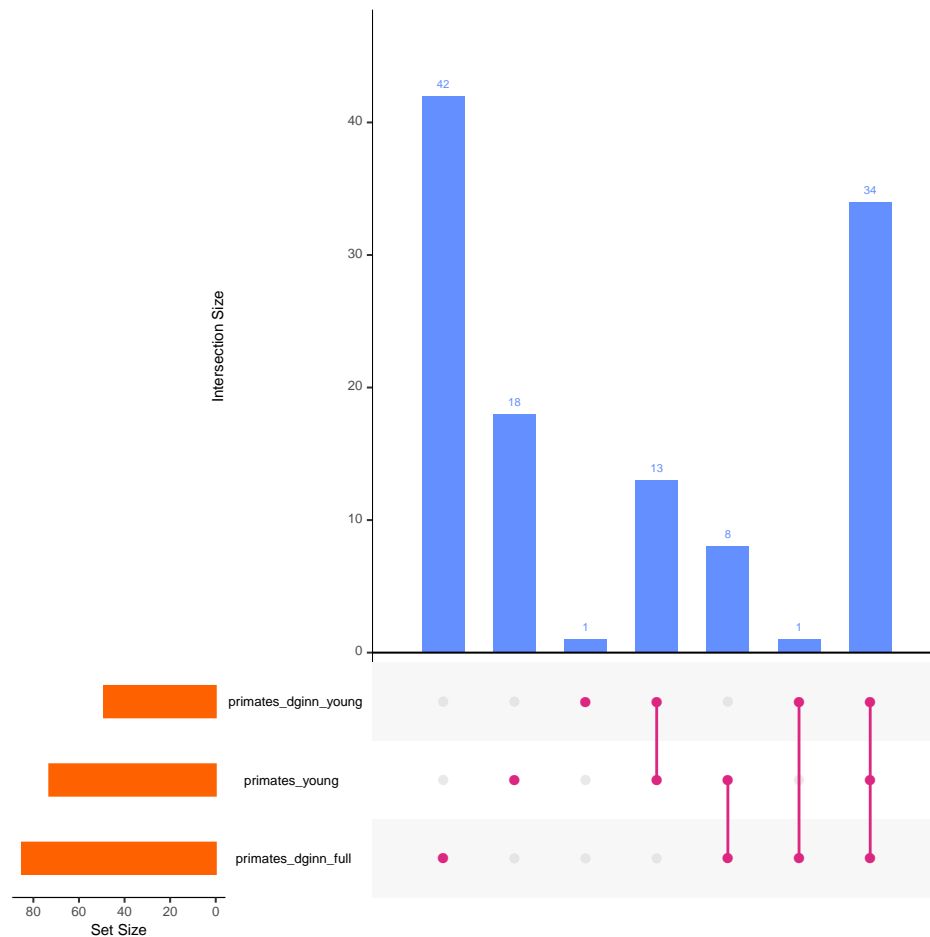
###
```

```

upsetdata$primates_dginn_young<-ifelse(dginnyoungtmp>=3, 1,0)
upsetdata$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")

```



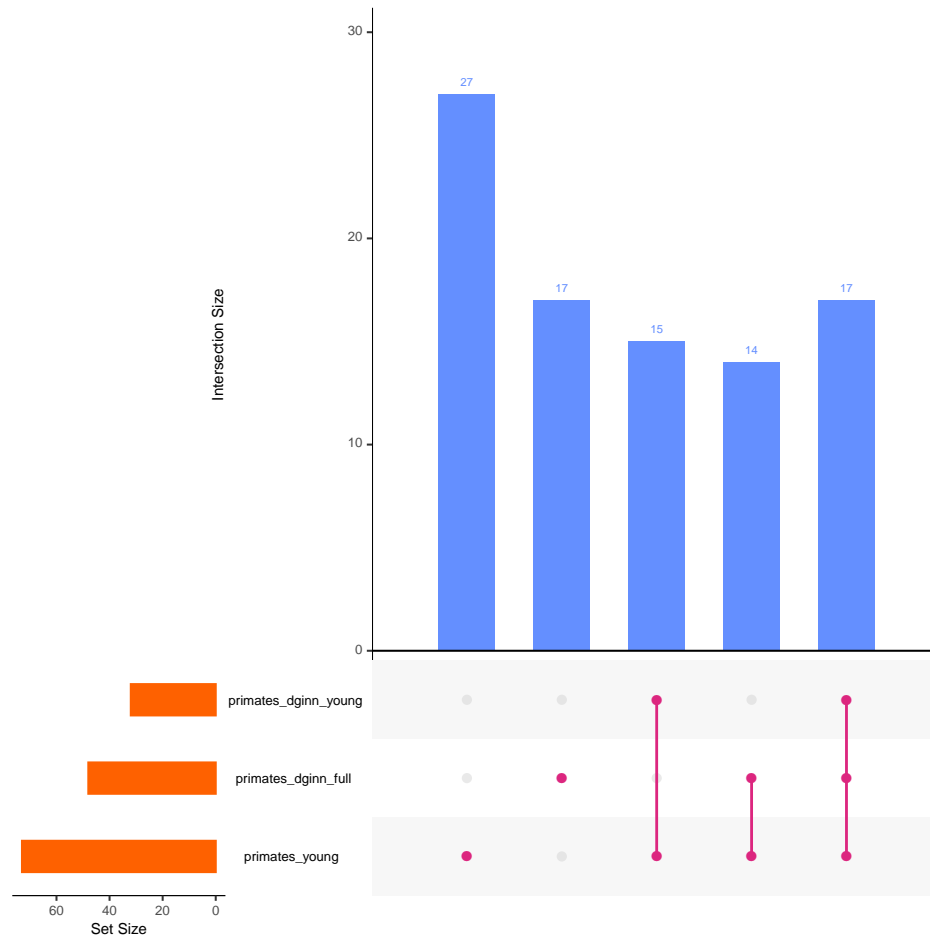
```

###
upsetdata$primates_dginn_young<-ifelse(dginnyoungtmp>=4, 1,0)
upsetdata$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"    "CEP68"     "CLIP
## [12] "GCC2"      "GGH"       "GHITM"     "GIGYF2"   "GLA"      "GOLGA7"    "HECT
## [23] "LMAN2"     "MARK1"     "MIPOL1"    "MPHOSPH10" "MYCBP2"    "NDUFAF2"   "NDUF
## [34] "PVR"       "REEP6"     "RIPK1"     "SAAL1"     "SEPSECS"   "SIRT5"     "SLC2
## [45] "UBAP2"     "UGGT2"     "VPS39"     "ZNF318"

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

## [1] "ACADM"      "ADAM9"      "AP2A2"     "ATE1"      "BCS1L"     "BRD4"      "BZW2
## [12] "CNTRL"      "DNMT1"      "DPH5"      "EDEM3"     "EIF4E2"    "EMC1"      "EXOS
## [23] "GIGYF2"     "GLA"       "GOLGA7"    "GOLGB1"    "GORASP1"   "HDAC2"     "HECT
## [34] "LARP4B"     "LARP7"     "LMAN2"     "MARK1"     "MDN1"      "MIPOL1"    "MOV1
## [45] "NDUFAF2"    "NDUFB9"    "NGLY1"     "NPC2"      "PCNT"      "PITRM1"    "PLAT
## [56] "PRIM2"      "PRKAR2A"   "PTBP2"     "PVR"       "RAB14"     "RAB1A"     "RAB2
## [67] "RPL36"      "SAAL1"     "SCCPDH"    "SEPSECS"   "SIRT5"     "SLC25A21"  "SLC2
## [78] "TRIM59"     "TRMT1"     "TUBGCP2"   "UBAP2"     "UGGT2"     "USP54"     "VPS3

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
)

```

```
}  
  
df<-read.delim(paste0(workdir,  
"/data/DGINN_202005281649summary_cleaned.csv"),  
fill=T, h=T, sep=",")  
  
makeFig1(df)
```

