

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

Marie Cariou

Mai 2020

Contents

1	Files manipulations	2
1.1	Read Janet Young's table	2
1.2	Read DGINN table	4
1.3	Joining table	4
1.3.1	Based on which column?	4
1.3.2	New columns	7
1.4	Write new table	7
2	Comparisons Primates	8
2.1	DGINN results on Janet Young's alignments (DGINN-Young-primate) VS Janet Young's results	8
2.1.1	Omega	8
2.1.2	pvalues pour M7M8	9
2.1.3	Concordance des méthodes	12
2.2	Résultats Cooper-primate VS Young-primate	13
2.2.1	How many genes in the Cooper-primate columns?	13
2.2.2	Omega	14
2.2.3	pvalues pour M7M8	15
2.3	Résultats DGINN sur alignement de Janet-Young (DGINN-Young-primate) VS Cooper-primates	17
2.3.1	Omega	17
2.3.2	pvalues pour M7M8	18
2.4	Overlap	19
2.4.1	Library and subtable	19
2.5	Mondrian	21

3	Bats Comparisons	22
3.1	Add DGINN results for Bats	22
3.2	Cooper-bats results vs DGINN-bats results	24
3.2.1	Omega	24
3.2.2	pvalues pour M7M8	25
3.3	Comparaison Cooper-Hawkins	26
3.3.1	pvalues pour M7M8	26
3.4	Comparaison dginn-Hawkins	28
3.5	Diagramme de Venn	29
3.5.1	subtab	30
3.5.2	figure	30
3.6	Mondrian	31
4	To do	32

1 Files manipulations

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

1.1 Read Janet Young's 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.primate.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_posSelIn
## 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 celle là:
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_posSelIn
## 158 mitochondrial amidoxime reducing component 1      MOSC1              no
##      Num.primite.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_posSelIn
## 158 mitochondrial amidoxime reducing component 1      MOSC1              no
##      Num.primite.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_", 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<-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")
```

1.4 Write new table

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

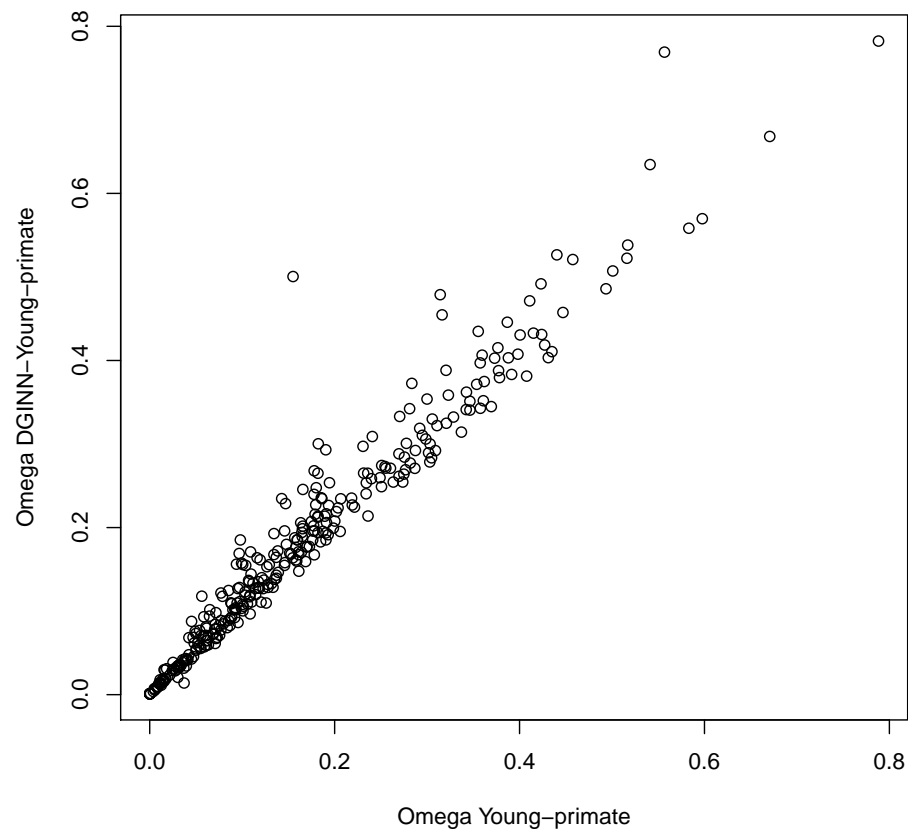
2 Comparisons Primates

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

2.1.1 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 Young-primate", ylab="Omega DGINN-Young-primate")
```



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

```

tab[tab$whole.gene.dN.dS.model.0<0.2 & tab$Omega_PamlM7M8>0.4,c("Gene.name")]

## [1] "MRPS2"

tab[tab$whole.gene.dN.dS.model.0<0.6 & tab$Omega_PamlM7M8>0.7,c("Gene.name")]

## [1] "PVR"

```

2.1.2 pvalues pour M7M8

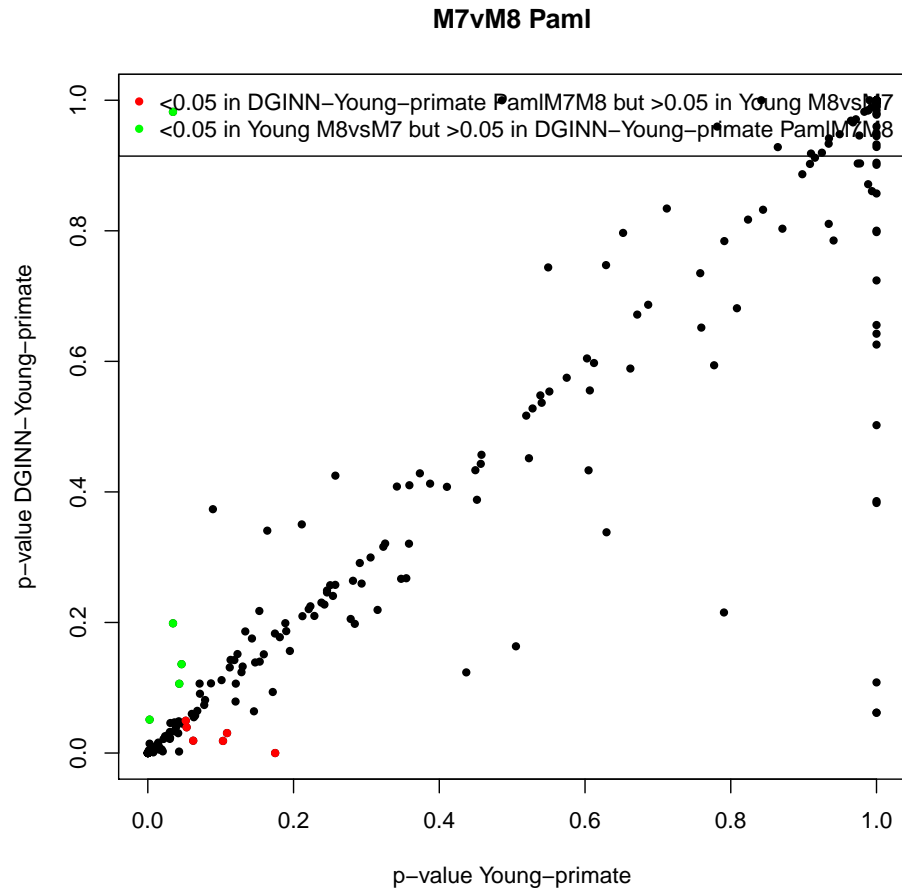
Cette fois, je compare la colonne R "pVal.M8vsM7", à la colonne "PValue" + ligne "PamlM7M8", pour la sortie de dginn.

```

plot(tab$pVal.M8vsM7, tab$PValue_PamlM7M8, pch=20,
      xlab="p-value Young-primate", ylab="p-value DGINN-Young-primate", main="M7vM8 Pa
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 DGINN-Young-primate PamlM7M8 but >0.05 in Young M8vsM7"
                    "<0.05 in Young M8vsM7 but >0.05 in DGINN-Young-primate PamlM7M8"),
      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", "Omega_P
```

##	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", "Omega_P

##      Gene.name pVal.M8vsM7 PValue_PamlM7M8 whole.gene.dN.dS.model.0
## 68      DCTPP1   0.0431830      0.10613016      0.29992
## 181     NDUFB9   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.primite.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
```

2.1.3 Concordance des méthodes

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

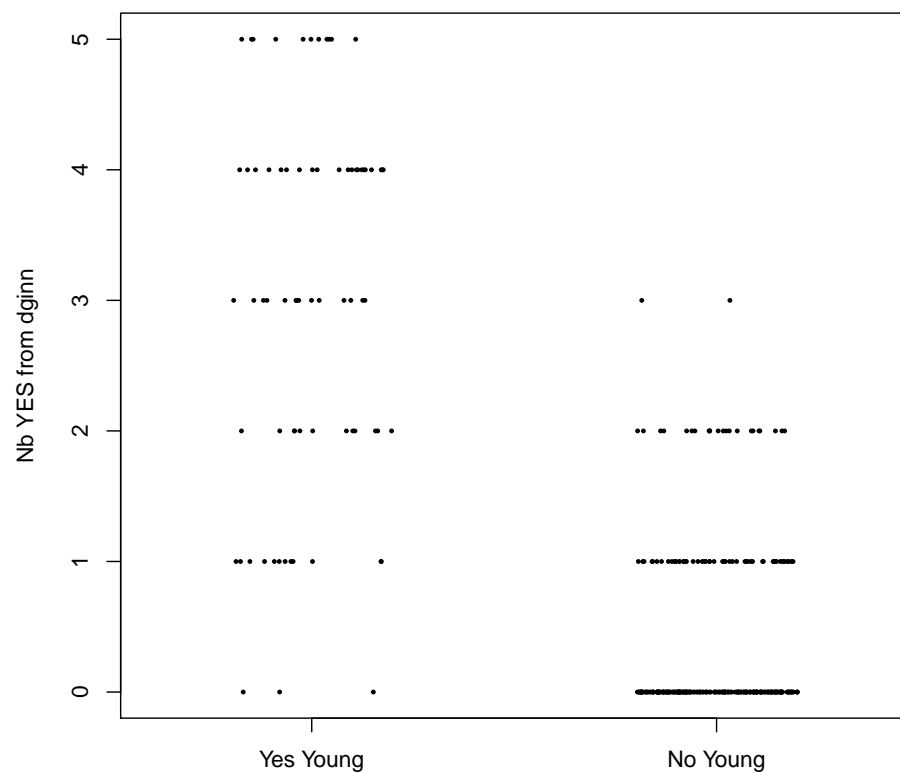
```
nontab<-tab[tab$pVal.M8vsM7>=0.05,c("Gene.name", "PosSel_PamlM1M2", "PosSel_PamlM7M8",
"PosSel_BppM7M8", "PosSel_BUSTED")]

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 Young", "No Young"),
               ylab="Nb YES from dginn")
```



2.2 Résultats Cooper-primate VS Young-primate

2.2.1 How many genes in the Cooper-primate columns?

```
# Temporary table with necessary columns

tmp<-tab[,c("Gene.name", "whole.gene.dN.dS.model.0", "pVal.M8vsM7",
"cooper.primates.Gene", "cooper.primates.Average_dNdS",
"cooper.primates.M7.M8_p_value")]
dim(tmp)

## [1] 332 6
```

```

# Lines with values in the cooper Gene names column
dim(tmp[tmp$cooper.primates.Gene!="",])

## [1] 207    6

# Line with values (no NA) in the Cooper dNdS column
sum(is.na(tmp$cooper.primates.Average_dNdS)==F)

## [1] 201

# Line with values (no NA) in the Cooper pvalue column
sum(is.na(tmp$cooper.primates.M7.M8_p_value)==F)

## [1] 207

```

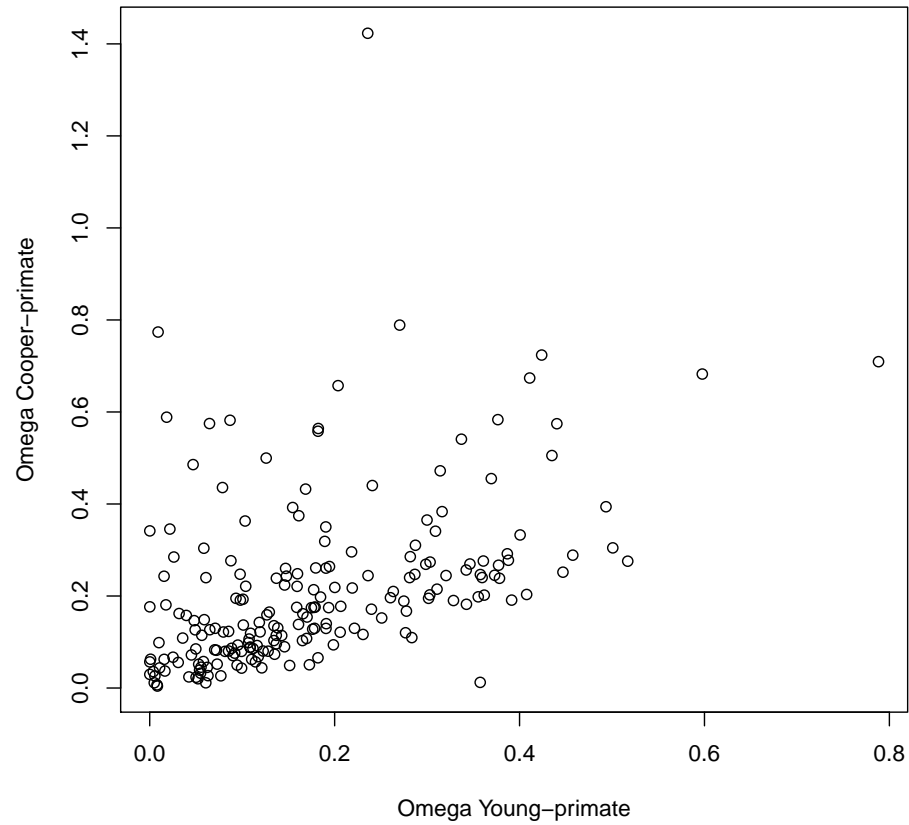
2.2.2 Omega

Comparaison des Omega: colonne L "whole.gene.dN.dS.model.0" VS colonne "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")

```



2.2.3 pvalues pour M7M8

Cette fois, je compare la colonne R "pVal.M8vsM7", à la colonne `cooper.primates.M7.M8_p_value` (p-value de l'analyse de Cooper).

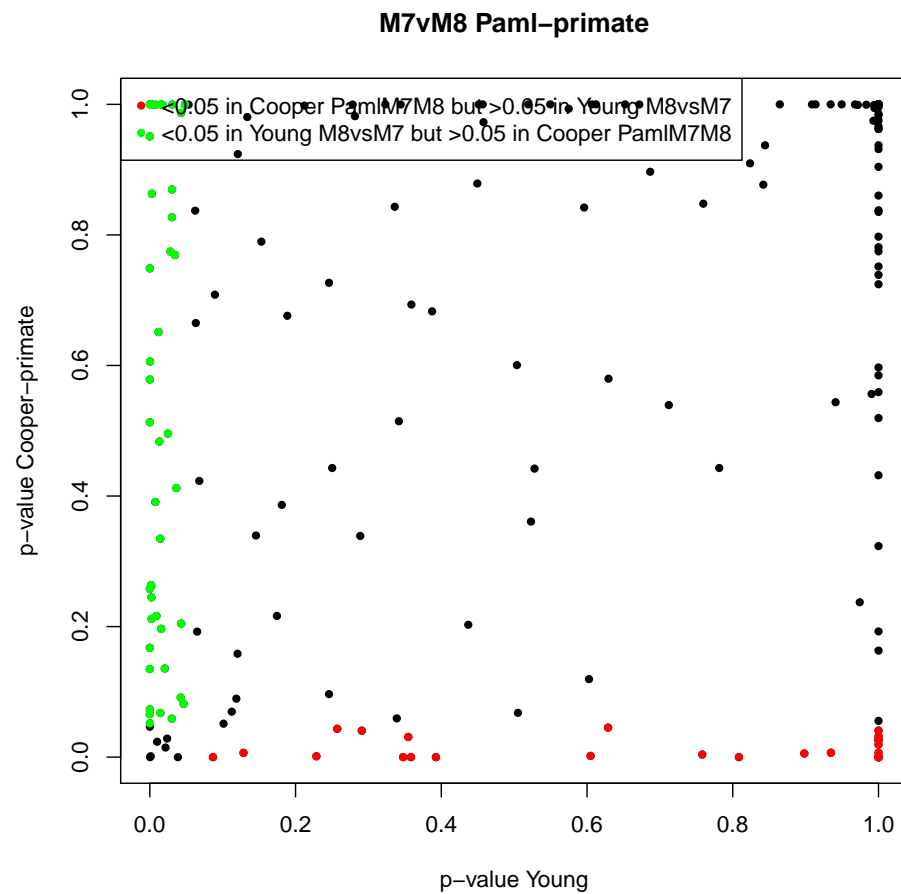
```
plot(tab$pVal.M8vsM7, tab$cooper.primates.M7.M8_p_value, pch=20,
      xlab="p-value Young", ylab="p-value Cooper-primate", main="M7vM8 Paml-primate")

points(tab$pVal.M8vsM7[tab$pVal.M8vsM7>0.05 & tab$cooper.primates.M7.M8_p_value<0.05],
       tab$cooper.primates.M7.M8_p_value[tab$pVal.M8vsM7>0.05 & tab$cooper.primates.M7.M8_p_value<0.05],
       col="red", pch=20)

points(tab$pVal.M8vsM7[tab$pVal.M8vsM7<0.05 & tab$cooper.primates.M7.M8_p_value>0.05],
       tab$cooper.primates.M7.M8_p_value[tab$pVal.M8vsM7<0.05 & tab$cooper.primates.M7.M8_p_value>0.05],
       col="green", pch=20)
```

```
tab$cooper.primates.M7.M8_p_value[tab$pVal.M8vsM7<0.05 & tab$cooper.primates.M
col="green", pch=20)
```

```
legend("topleft", c("<0.05 in Cooper PamLM7M8 but >0.05 in Young M8vsM7",
"<0.05 in Young M8vsM7 but >0.05 in Cooper PamLM7M8"),
pch=20, col=c("red", "green"))
```

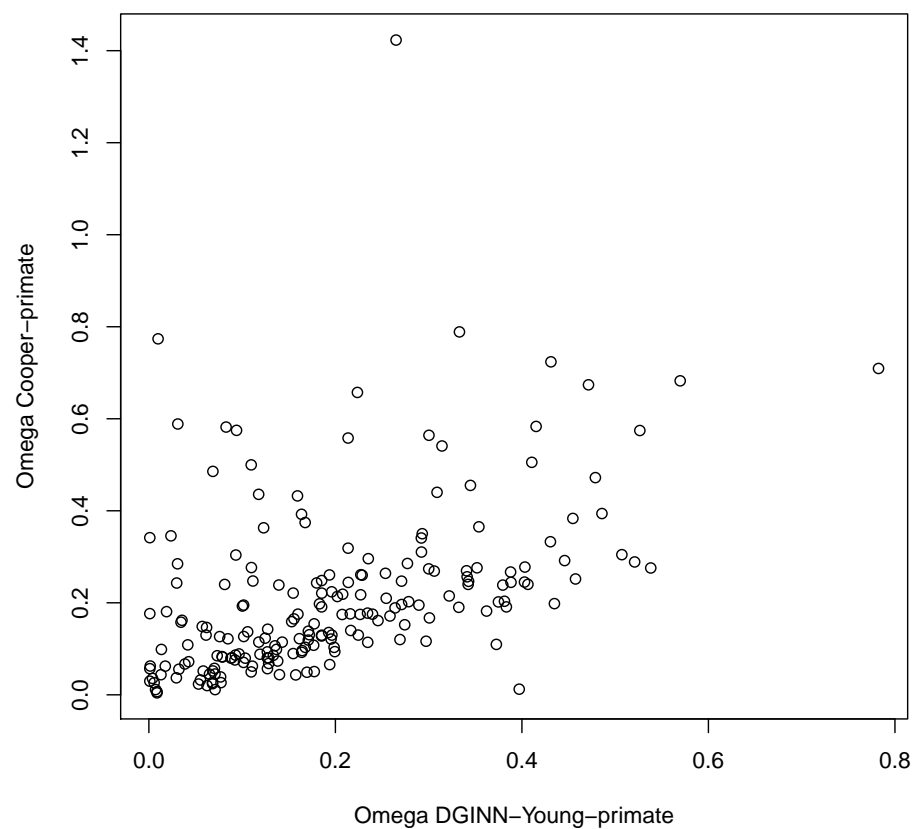


2.3 Résultats DGINN sur alignement de Janet-Young (DGINN-Young-primate) VS Cooper-primates

2.3.1 Omega

Comparaison des Omega: colonne colonne "cooper.primates.Average_dNdS"
VS omega de DGINN.

```
plot(tab$Omega_PamlM7M8, tab$cooper.primates.Average_dNdS,  
      xlab="Omega DGINN-Young-primate", ylab="Omega Cooper-primate")
```

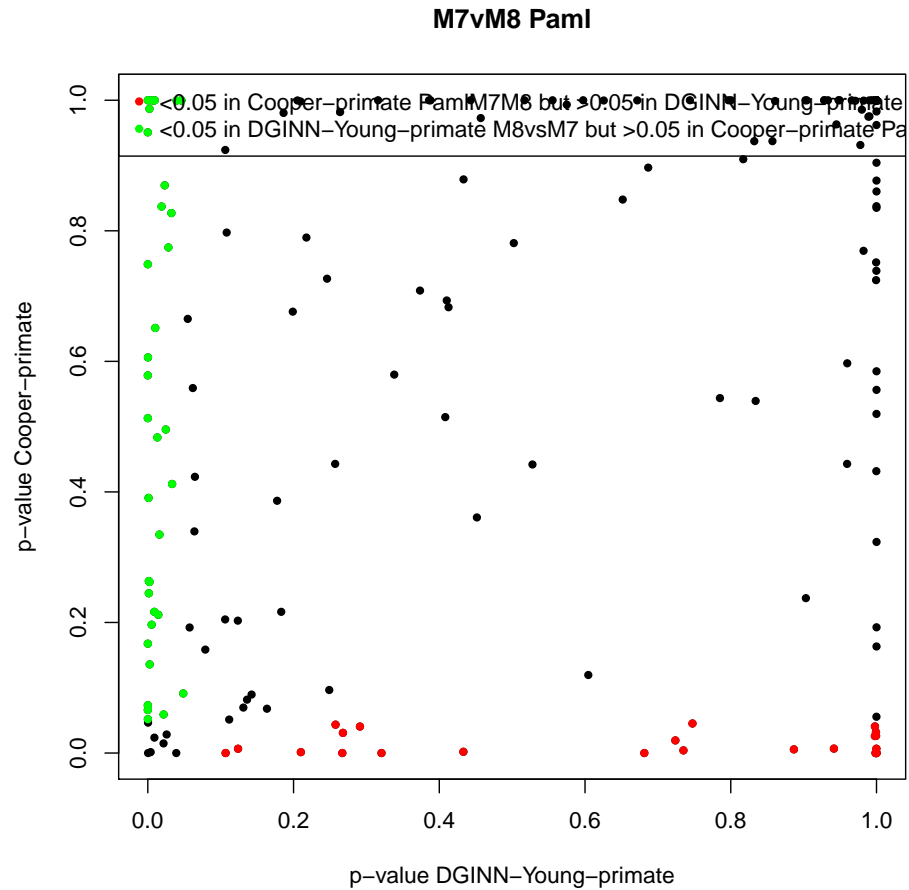


2.3.2 pvalues pour M7M8

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

```
plot(tab$PValue_PamlM7M8, tab$cooper.primates.M7.M8_p_value, pch=20,
      xlab="p-value DGINN-Young-primate", ylab="p-value Cooper-primate", main="M7vM8 P
points(tab$PValue_PamlM7M8[tab$PValue_PamlM7M8>0.05 & tab$cooper.primates.M7.M8_p_val
      tab$cooper.primates.M7.M8_p_value[tab$PValue_PamlM7M8>0.05 & tab$cooper.primat
      col="red", pch=20)
points(tab$PValue_PamlM7M8[tab$PValue_PamlM7M8<0.05 & tab$cooper.primates.M7.M8_p_val
      tab$cooper.primates.M7.M8_p_value[tab$PValue_PamlM7M8<0.05 & tab$cooper.primat
      col="green", pch=20)

legend("topleft", c("<0.05 in Cooper-primate PamlM7M8 but >0.05 in DGINN-Young-primat
      "<0.05 in DGINN-Young-primate M8vsM7 but >0.05 in Cooper-primate PamlM7M8"),
      pch=20, col=c("red", "green"))
```



2.4 Overlap

I will draw a venn diagramm for the positive genes in the 3 analyses.

2.4.1 Library and subtable

```
library(VennDiagram)

# keeps only genes analysed in all 3 experiments
tmp<-na.omit(tab[,c("Gene.name", "pVal.M8vsM7", "cooper.primates.M7.M8_p_value",
  "PosSel_PamI M7M8", "PValue_PamI M7M8")])
dim(tmp)
```

```
## [1] 186 5
```

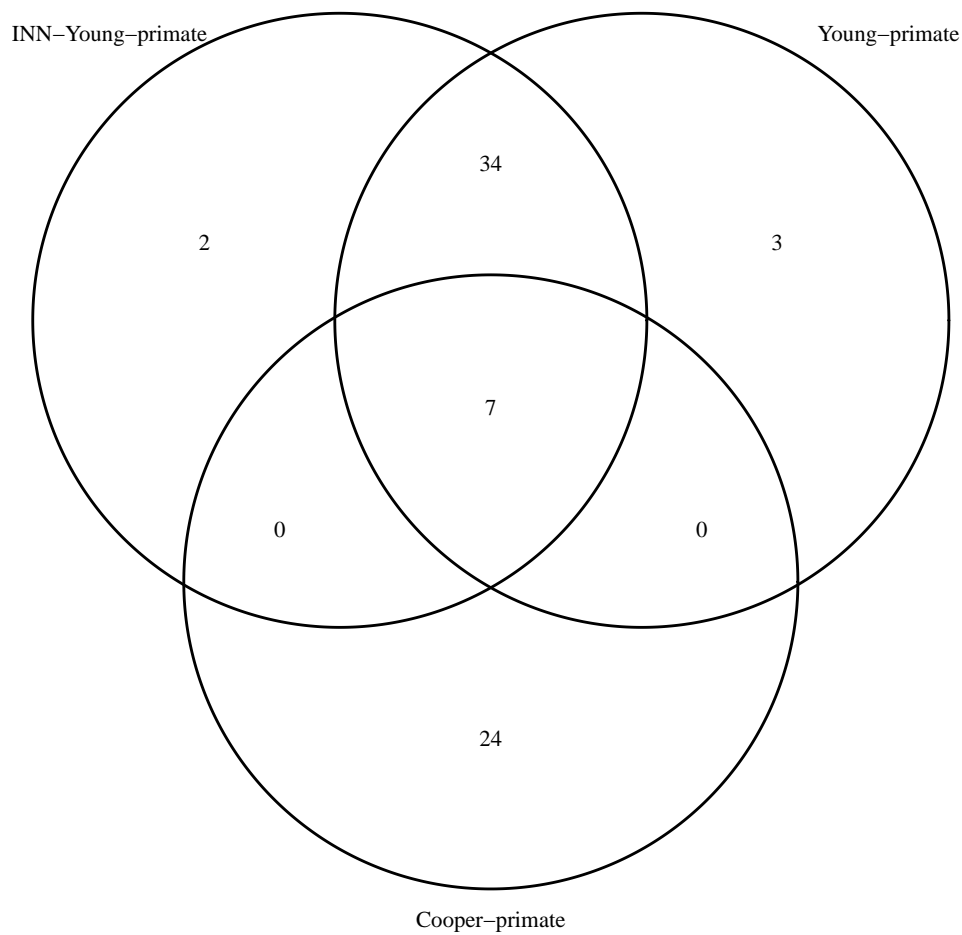
Il reste 186 gènes

```
area1dginn<-sum(tmp$PosSel_PamlM7M8=="Y")
area2jean<-sum(tmp$pVal.M8vsM7<0.05)
area3coop<-sum(tmp$cooper.primates.M7.M8_p_val<0.05, na.rm=T)

n12<-sum(tmp$PosSel_PamlM7M8=="Y" & tmp$pVal.M8vsM7<0.05)
n23<-sum(tmp$pVal.M8vsM7<0.05 & tmp$cooper.primates.M7.M8_p_val<0.05, na.rm=T)
n13<-sum(tmp$PosSel_PamlM7M8=="Y" & tmp$cooper.primates.M7.M8_p_val<0.05, na.rm=T)

n123<-sum(tmp$PosSel_PamlM7M8=="Y" & tmp$pVal.M8vsM7<0.05 &
tmp$cooper.primates.M7.M8_p_val<0.05, na.rm=T)

draw.triple.venn(area1dginn, area2jean, area3coop,
n12, n23, n13, n123,
category=c("DGINN-Young-primate", "Young-primate", "Cooper-primate"))
```



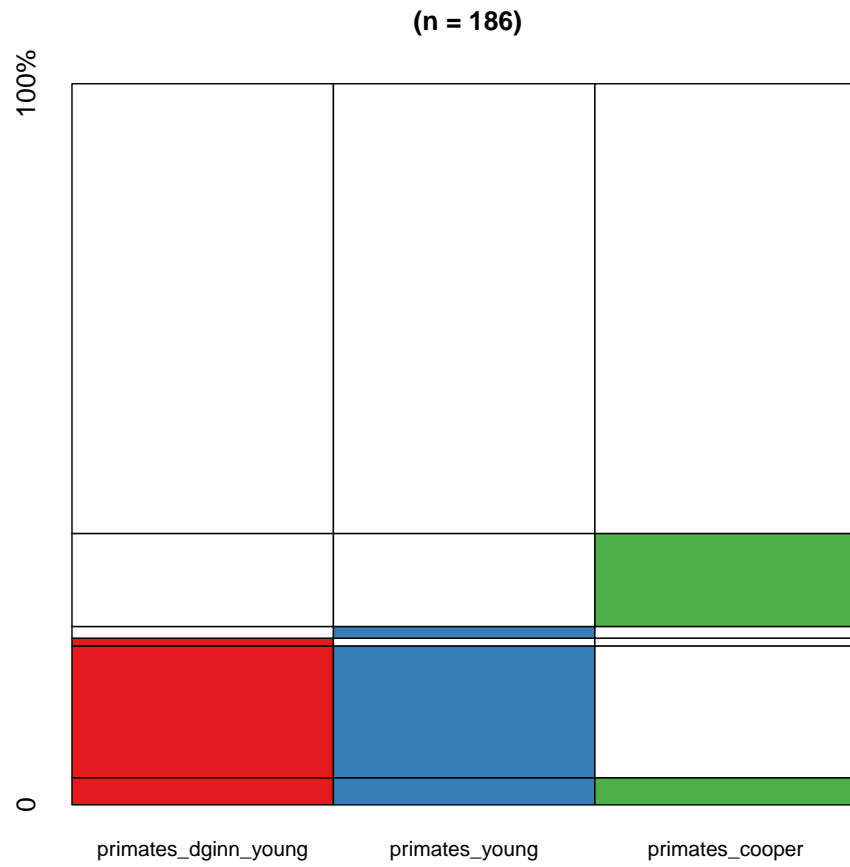
```
## (polygon[GRID.polygon.836], polygon[GRID.polygon.837], polygon[GRID.polygon.838],
```

2.5 Mondrian

```
library(Mondrian)

monddata<-as.data.frame(tmp$Gene.name)
monddata$primates_dginn_young<-ifelse(tmp$PosSel_PamlM7M8=="Y", 1,0)
monddata$primates_young<-ifelse(tmp$pVal.M8vsM7<0.05, 1, 0)
monddata$primates_cooper<-ifelse(tmp$cooper.primates.M7.M8_p_val<0.05, 1, 0)
```

```
mondrian(monddata[,2:4])
```



3 Bats Comparisons

3.1 Add DGINN results for Bats

Lecture du tableau.

```
# Tableau tel qu'il est à cet étape du script, pour travailler sur l'ajout du nouveau
tab<-read.delim("COVID_PAMLresults_332hits_plusBatScreens_plusDGINN_20200506.txt",
  header=TRUE, sep="\t")
```

```

#dginnbatsold<-read.delim("/home/adminmarie/Documents/CIRI_BIBS_projects/2020_05_Etienne_
#  fill=T, h=T)

dginnbats<-read.delim("/home/adminmarie/Documents/CIRI_BIBS_projects/2020_05_Etienne_
                      fill=T, h=T)

dim(dginnbats)

## [1] 349 27

names(dginnbats)

## [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"

length(unique(dginnbats$Gene))

## [1] 349

length(unique(tab$cooper.batsGene))

## [1] 218

table(unique(tab$cooper.batsGene) %in% unique(dginnbats$Gene))

##
## FALSE TRUE
##      3   215

```

Which genes in the Cooper table are not in the gene output?

```
unique(tab$cooper.batsGene)[unique(tab$cooper.batsGene) %in% unique(dginnbats$Gene)==
## [1]          BCS1L    C1orf50
## 218 Levels:  AAR2 AASS AATF ACADM ACSL3 ADAMTS1 AGPS AKAP8L ALG11 ALG5 ... ZYG11B
```

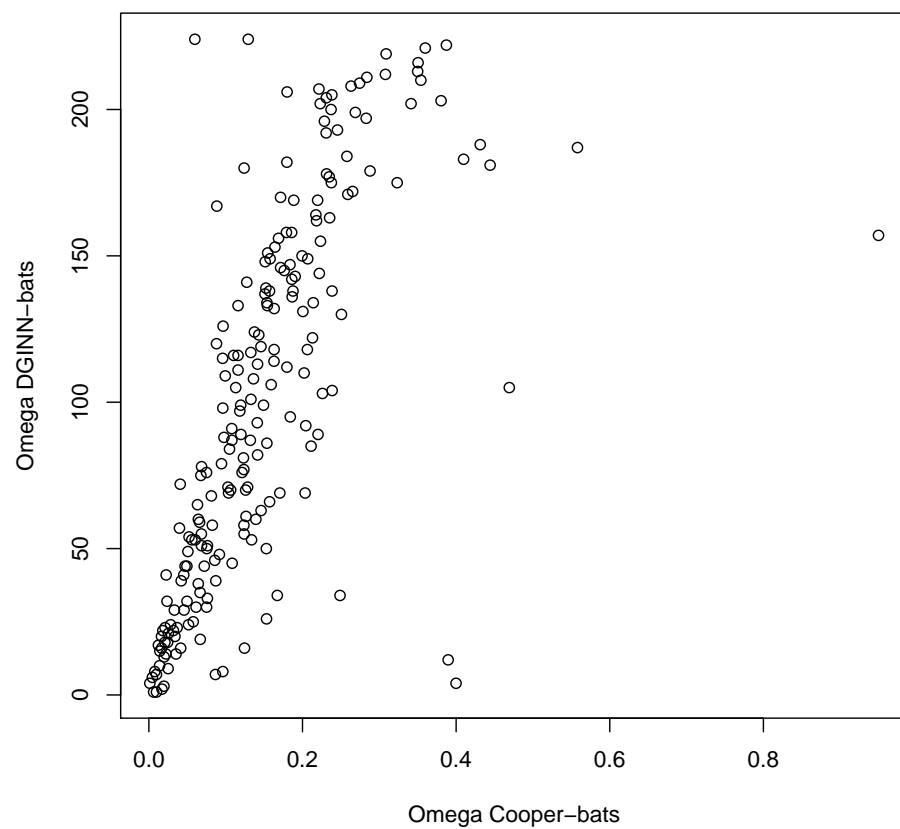
Merge tables:

```
names(dginnbats)<-c("File", "bats_Name", "cooper.batsGene", paste0("bats_", names(dgi
tab<-merge(tab,dginnbats, by="cooper.batsGene", all.x=T)
```

3.2 Cooper-bats results vs DGINN-bats results

3.2.1 Omega

```
plot(tab$cooper.batsAverage_dNdS, tab$bats_omegaM0codeml,
      xlab="Omega Cooper-bats", ylab="Omega DGINN-bats")
```



3.2.2 pvalues pour M7M8

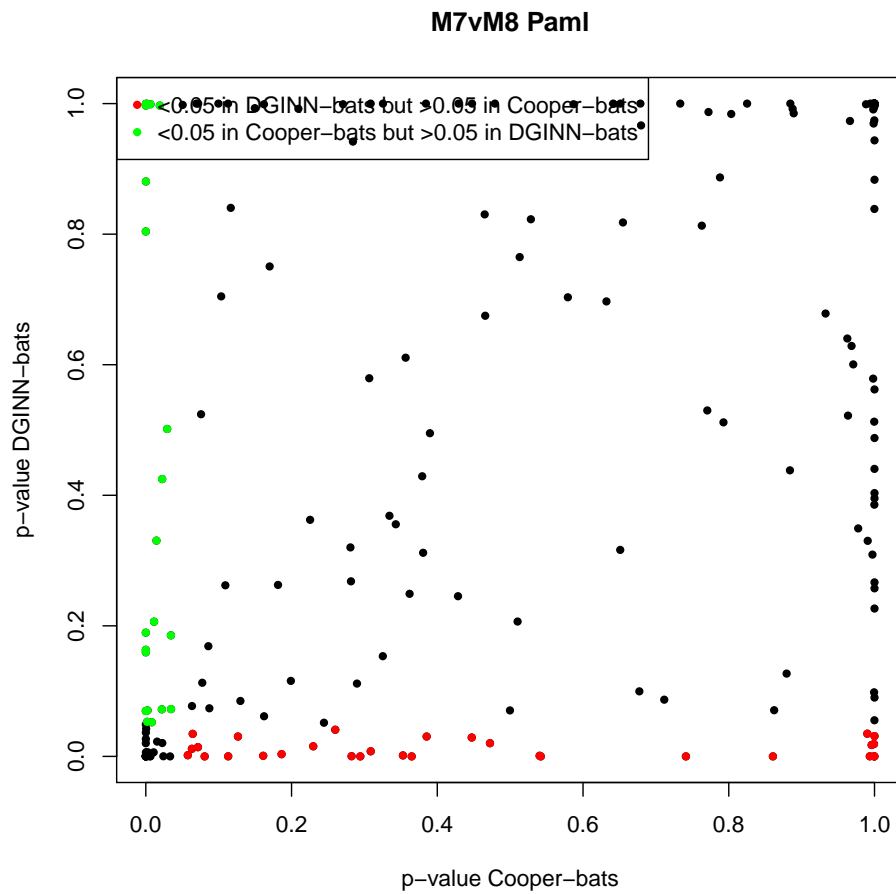
```
tab$bats_codemlM7M8.p.value<-as.numeric(as.character(tab$bats_codemlM7M8.p.value))  
## Warning: NAs introduits lors de la conversion automatique  
  
plot(tab$cooper.batsM7.M8_p_value, tab$bats_codemlM7M8.p.value, pch=20,  
      xlab="p-value Cooper-bats", ylab="p-value DGINN-bats", main="M7vM8 Pam1")  
  
points(tab$cooper.batsM7.M8_p_value[tab$cooper.batsM7.M8_p_value>0.05 & tab$bats_code  
       tab$bats_codemlM7M8.p.value[tab$cooper.batsM7.M8_p_value>0.05 & tab$bats_codem  
       col="red", pch=20)
```

```

points(tab$cooper.batsM7.M8_p_value[tab$cooper.batsM7.M8_p_value<0.05 & tab$bats_code
      tab$bats_codemlM7M8.p.value[tab$cooper.batsM7.M8_p_value<0.05 & tab$bats_codeml
      col="green", pch=20)

legend("topleft", c("<0.05 in DGINN-bats but >0.05 in Cooper-bats",
                    "<0.05 in Cooper-bats but >0.05 in DGINN-bats"),
      pch=20, col=c("red", "green"))

```



3.3 Comparaison Cooper-Hawkins

3.3.1 pvalues pour M7M8

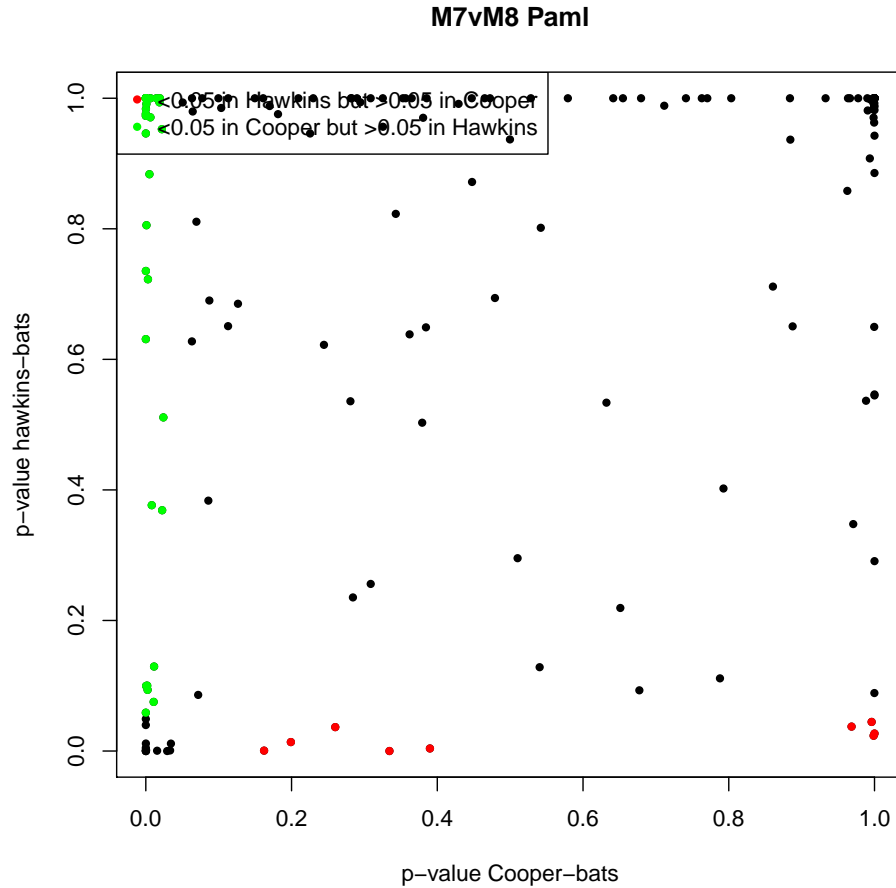
```

plot(tab$cooper.batsM7.M8_p_value, tab$hawkins_Positive.Selection..M8vM8a.p.value,
     pch=20, xlab="p-value Cooper-bats", ylab="p-value hawkins-bats", main="M7vM8

points(tab$cooper.batsM7.M8_p_value[tab$cooper.batsM7.M8_p_value>0.05 &
     tab$hawkins_Positive.Selection..M8vM8a.p.value<0.05],
     tab$hawkins_Positive.Selection..M8vM8a.p.value[tab$cooper.batsM7.M8_p_value>0.05
     tab$hawkins_Positive.Selection..M8vM8a.p.value<0.05],
     col="red", pch=20)
points(tab$cooper.batsM7.M8_p_value[tab$cooper.batsM7.M8_p_value<0.05 &
     tab$hawkins_Positive.Selection..M8vM8a.p.value>0.05],
     tab$hawkins_Positive.Selection..M8vM8a.p.value[tab$cooper.batsM7.M8_p_value<0.05
     tab$hawkins_Positive.Selection..M8vM8a.p.value>0.05],
     col="green", pch=20)

legend("topleft", c("<0.05 in Hawkins but >0.05 in Cooper",
     "<0.05 in Cooper but >0.05 in Hawkins"),
     pch=20, col=c("red", "green"))

```



3.4 Comparaison dginn-Hawkins

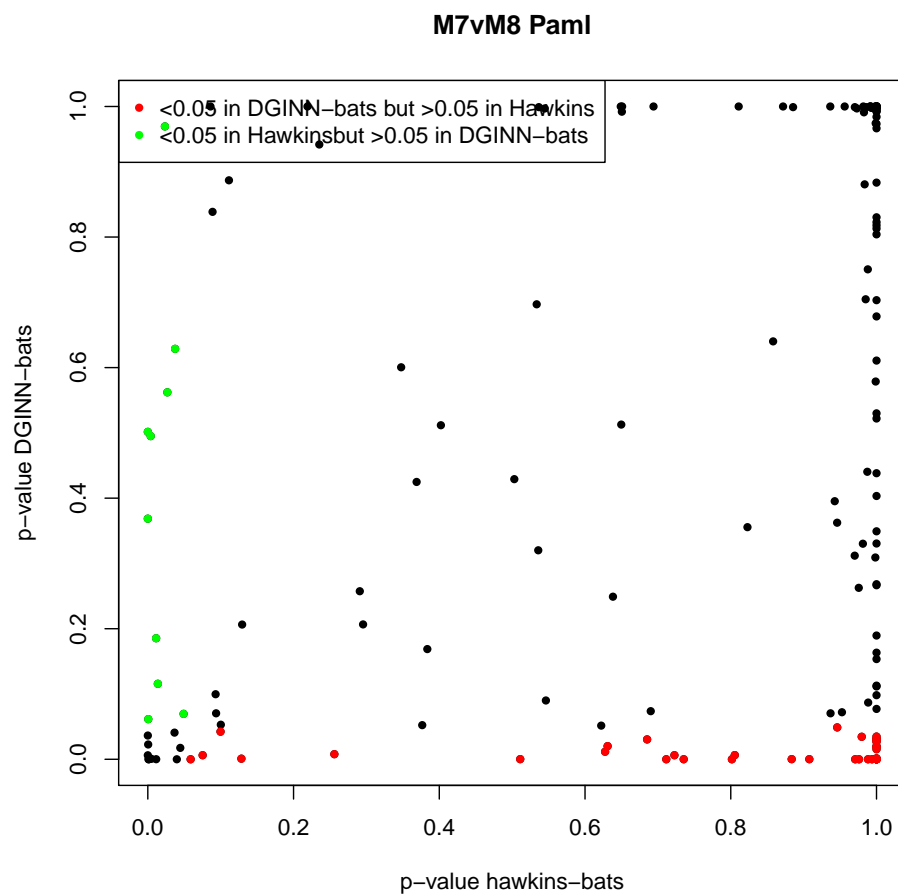
```
plot(tab$hawkins_Positive.Selection..M8vM8a.p.value, tab$bats_codemlM7M8.p.value,
     pch=20, xlab="p-value hawkins-bats", ylab="p-value DGINN-bats", main="M7vM8 P
points(tab$hawkins_Positive.Selection..M8vM8a.p.value[tab$hawkins_Positive.Selection.
      tab$bats_codemlM7M8.p.value<0.05],
      tab$bats_codemlM7M8.p.value[tab$hawkins_Positive.Selection..M8vM8a.p.value>0.05 &
      tab$bats_codemlM7M8.p.value<0.05], col="red", pch=20)
points(tab$hawkins_Positive.Selection..M8vM8a.p.value[tab$hawkins_Positive.Selection.
      tab$bats_codemlM7M8.p.value>0.05],
```

```

tab$bats_codemlM7M8.p.value[tab$hawkins_Positive.Selection..M8vM8a.p.value<0.05 &
tab$bats_codemlM7M8.p.value>0.05], col="green", pch=20)

legend("topleft", c("<0.05 in DGINN-bats but >0.05 in Hawkins",
                    "<0.05 in Hawkinsbut >0.05 in DGINN-bats"),
      pch=20, col=c("red", "green"))

```



3.5 Diagramme de Venn

I will draw a venn diagramm for the positive genes in the 3 analyses.

3.5.1 subtab

```
tmp<-na.omit(tab[,c("Gene.name", "bats_codemlM7M8.p.value", "hawkins_Positive.Selection..M8vM8a.p.value", "cooper_batsM7.M8_p.value")])
dim(tmp)

## [1] 168 4
```

154 genes (present in the 3 experiments)

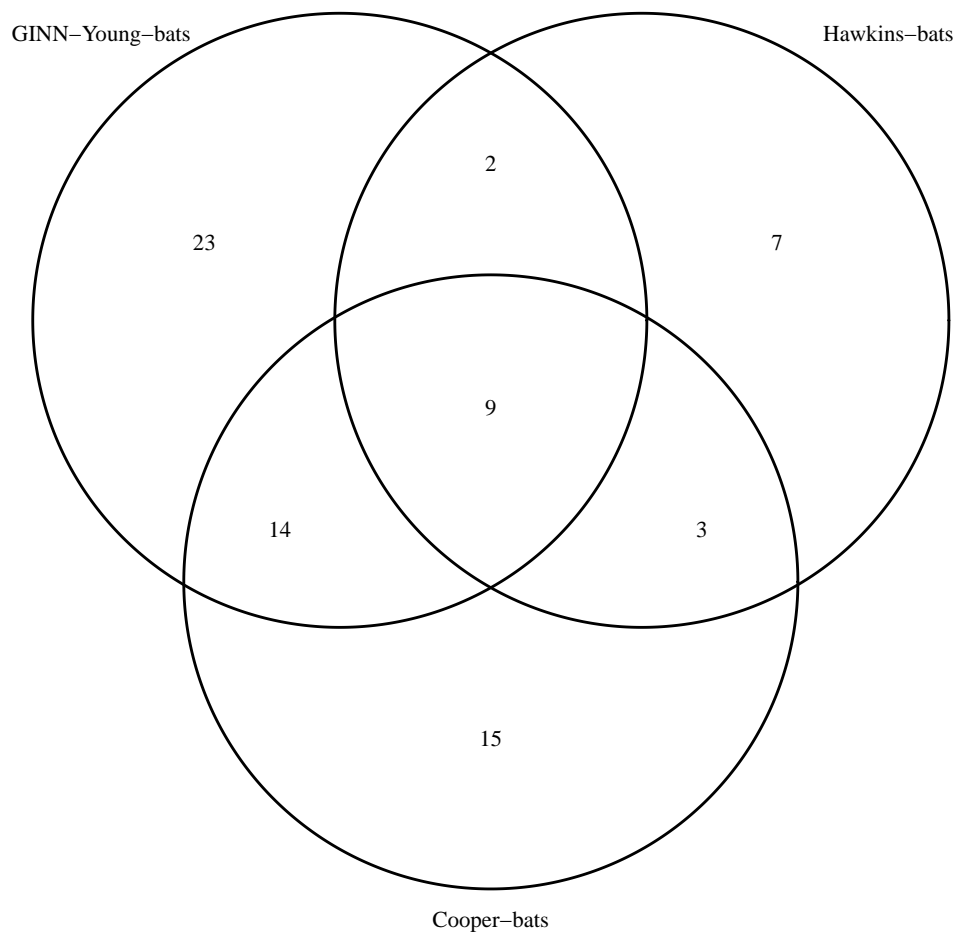
3.5.2 figure

```
area1dginn<-sum(tmp$bats_codemlM7M8.p.value<0.05, na.rm=T)
area2hawk<-sum(tmp$hawkins_Positive.Selection..M8vM8a.p.value<0.05, na.rm=T)
area3coop<-sum(tmp$cooper_batsM7.M8_p.value<0.05, na.rm=T)

n12<-sum(tmp$bats_codemlM7M8.p.value<0.05 & tmp$hawkins_Positive.Selection..M8vM8a.p.value<0.05)
n23<-sum(tmp$hawkins_Positive.Selection..M8vM8a.p.value<0.05 & tmp$cooper_batsM7.M8_p.value<0.05)
n13<-sum(tmp$bats_codemlM7M8.p.value<0.05 & tmp$cooper_batsM7.M8_p.value<0.05, na.rm=T)

n123<-sum(tmp$bats_codemlM7M8.p.value<0.05 & tmp$hawkins_Positive.Selection..M8vM8a.p.value<0.05 & tmp$cooper_batsM7.M8_p.value<0.05)

draw.triple.venn(area1dginn, area2hawk, area3coop,
n12, n23, n13, n123,
category=c("DGINN-Young-bats", "Hawkins-bats", "Cooper-bats"))
```



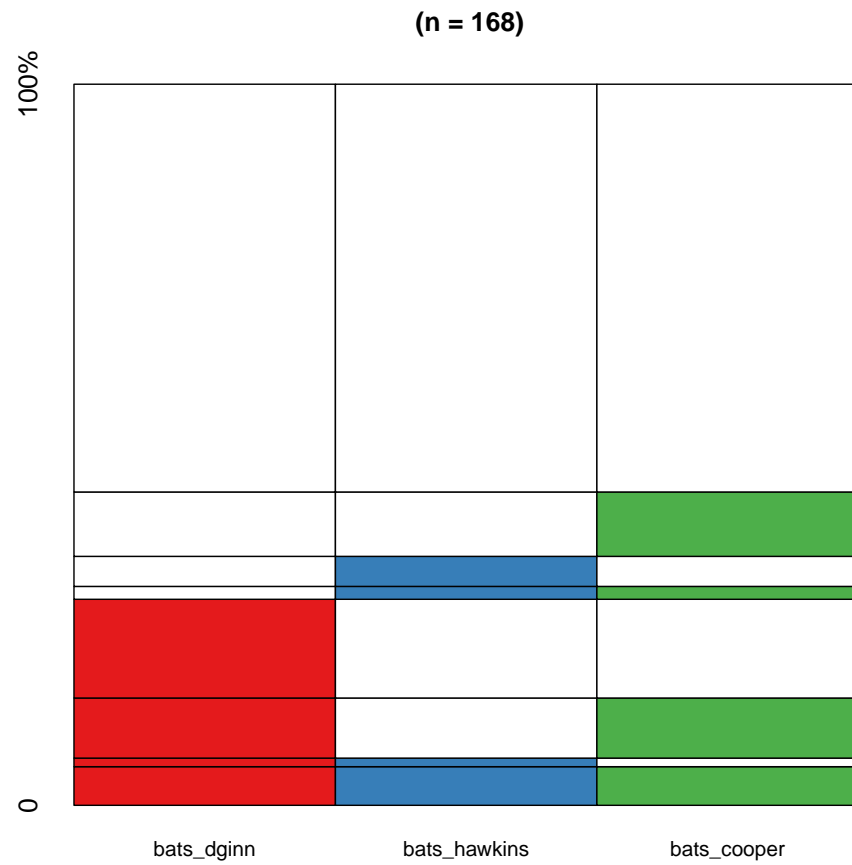
```
## (polygon[GRID.polygon.852], polygon[GRID.polygon.853], polygon[GRID.polygon.854],
```

3.6 Mondrian

```
library(Mondrian)

monddata<-as.data.frame(tmp$Gene.name)
monddata$bats_dginn<-ifelse(tmp$bats_codemlM7M8.p.value<0.05, 1,0)
monddata$bats_hawkins<-ifelse(tmp$hawkins_Positive.Selection..M8vM8a.p.value<0.05, 1,
monddata$bats_cooper<-ifelse(tmp$cooper.batsM7.M8_p_value<0.05, 1, 0)
```

```
mondrian(monddata[,2:4])
```



4 To do

Comparaison G4 pas G4