

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

Marie Cariou

Janvier 2021

Contents

1	Data	2
2	Comparison of dataset	2
2.1	Data	2
2.2	Omega plot	2
2.3	Mondrian	4
2.4	subsetR	6
3	Which are these genes?	9
3.1	Gene under positive selection in both bats and primates . . .	9
3.2	Gene under positive selection only in primates	11
3.3	Gene under positive selection only in bats	17
3.4	Figure tableau	19

1 Data

Analysis were formatted by the script covid_comp_script0_table.Rnw.

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

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

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

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

## [1] 442 56
```

2 Comparison of dataset

2.1 Data

```
tmp<-na.omit(tab[,c("Gene.name", "bats_BUSTED", "bats_BppM1M2", "bats_BppM7M8",
  "bats_codemlM1M2", "bats_codemlM7M8", "dginn.primite_codemlM1M2",
  "dginn.primite_codemlM7M8", "dginn.primite_BppM1M2",
  "dginn.primite_BppM7M8", "dginn.primite_BUSTED")])
col<-c("Gene.name", "bats_BUSTED", "bats_BppM1M2", "bats_BppM7M8",
  "bats_codemlM1M2", "bats_codemlM7M8", "dginn.primite_codemlM1M2",
  "dginn.primite_codemlM7M8", "dginn.primite_BppM1M2",
  "dginn.primite_BppM7M8", "dginn.primite_BUSTED")
dim(tmp)

## [1] 323 11
```

2.2 Omega plot

```

x=as.numeric(as.character(tab$dginn.primate_omegaMOBpp[tab$status=="shared"]))

## Warning:  NAs introduits lors de la conversion automatique

y=as.numeric(as.character(tab$bats_omegaMOBpp[tab$status=="shared"]))

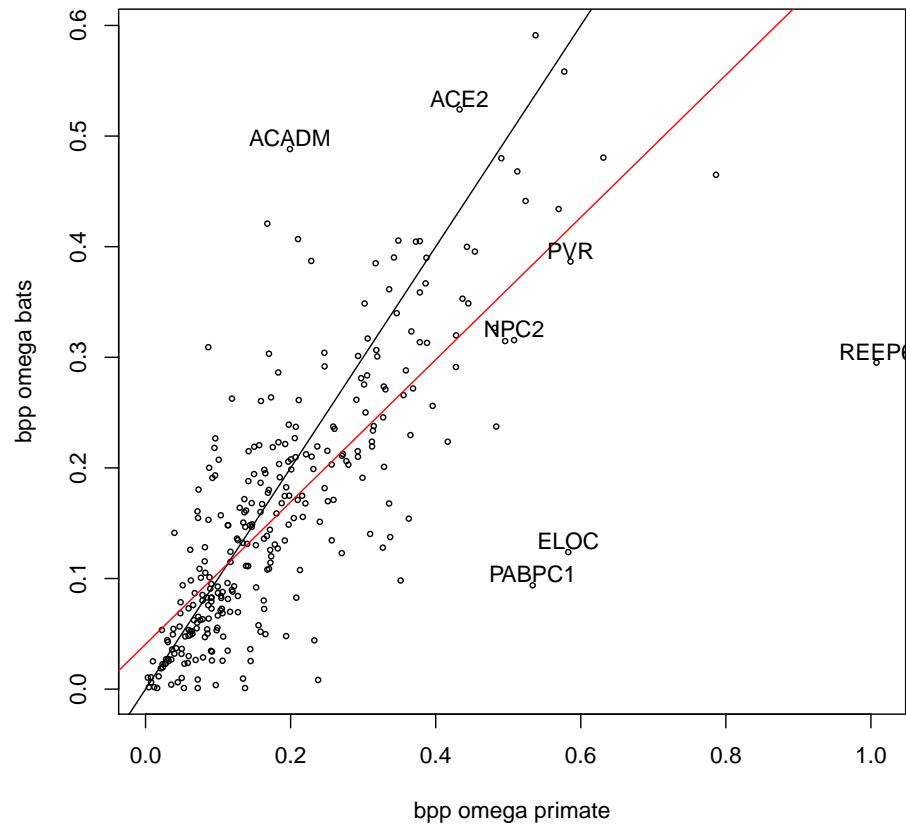
## Warning:  NAs introduits lors de la conversion automatique

names(x)<-tab$Gene.name[tab$status=="shared"]

plot(x,y, xlab="bpp omega primate", ylab="bpp omega bats", cex=0.5)
abline(0,1)
abline(lm(y~x), col="red")

text(x[x>0.5 &y<0.4], (y[x>0.5 &y<0.4]+0.01), names(x)[x>0.5 &y<0.4])
text(x[x<0.45 &y>0.45], (y[x<0.45 &y>0.45]+0.01), names(x)[x<0.45 &y>0.45])

```



2.3 Mondrian

```
library(Mondrian)

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

batstmp<-rowSums(cbind(tmp$bats_codemlM1M2=="Y", tmp$bats_codemlM7M8=="Y",
tmp$bats_BppM1M2=="Y", tmp$bats_BppM7M8=="Y", tmp$bats_BUSTED=="Y"))

primatetmp<-rowSums(cbind(tmp$"dginn.primate_codemlM1M2"=="Y",
tmp$"dginn.primate_codemlM7M8"=="Y", tmp$"dginn.primate_BppM1M2"=="Y",
```

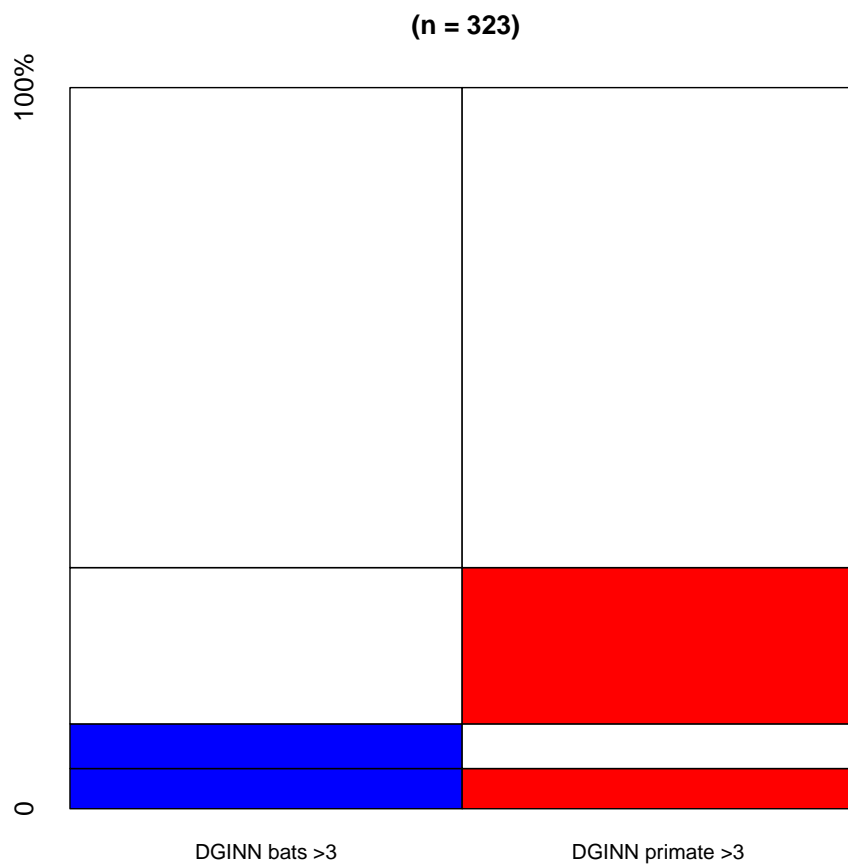
```

tmp$"dginn.primates_BppM7M8"=="Y", tmp$"dginn.primates_BUSTED"=="Y"))

monddata$bats_dginn3<-ifelse(batstmp>=3, 1,0)
monddata$primate_dginn3<-ifelse(primatetmp>=3, 1,0)
monddata$bats_dginn4<-ifelse(batstmp>=4, 1,0)
monddata$primate_dginn4<-ifelse(primatetmp>=4, 1,0)

mondrian(monddata[,2:3], labels=c("DGINN bats >3", "DGINN primate >3"))

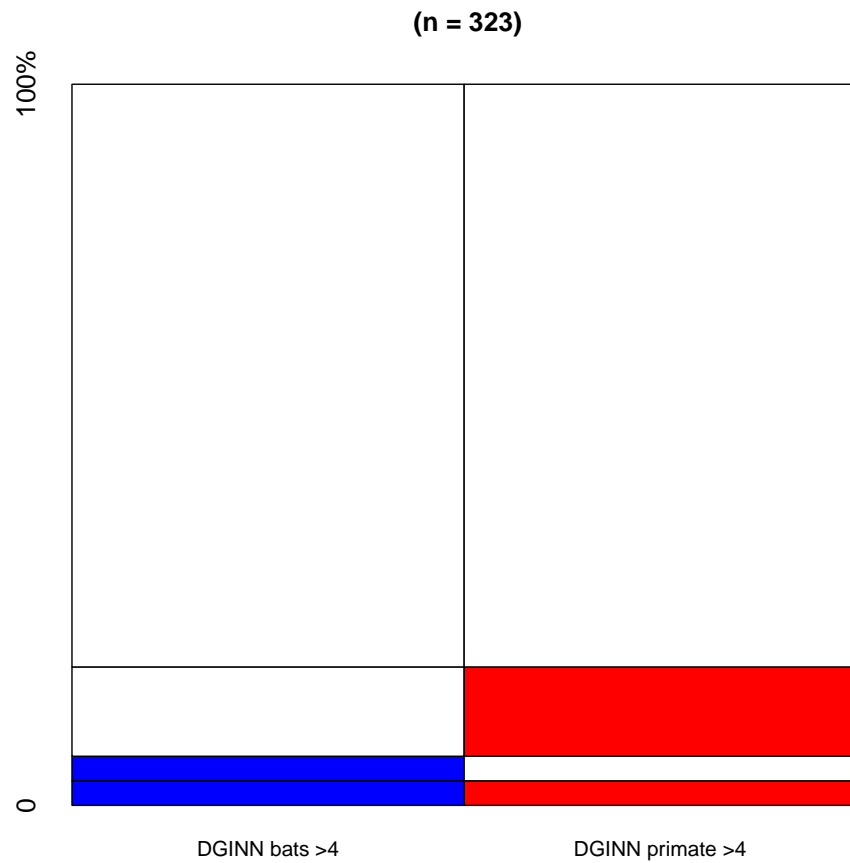
```



```

mondrian(monddata[,4:5], labels=c("DGINN bats >4", "DGINN primate >4"))

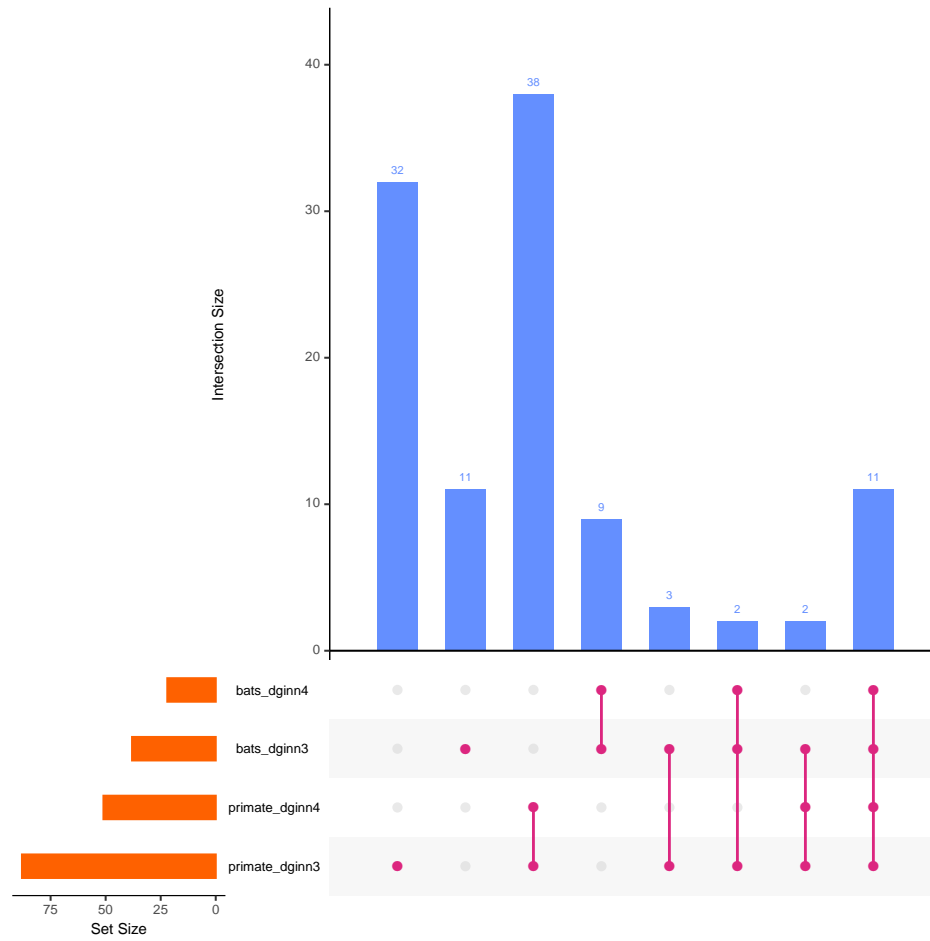
```



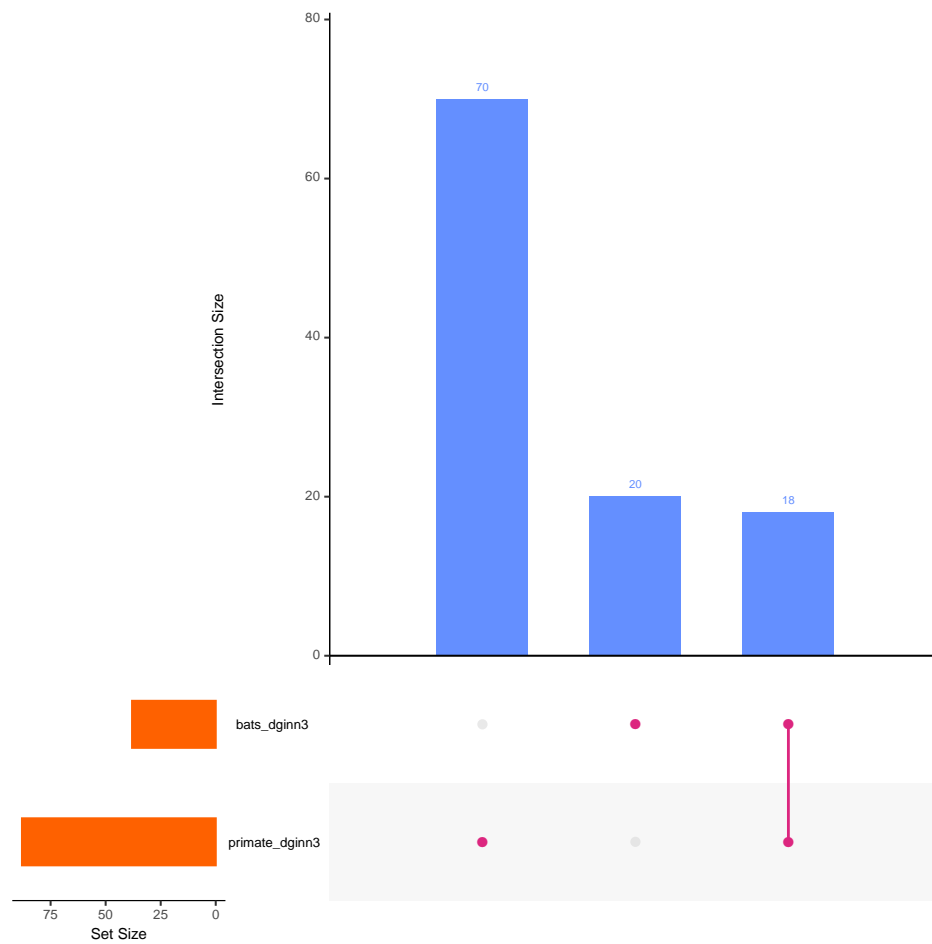
2.4 subsetR

```
library(UpSetR)

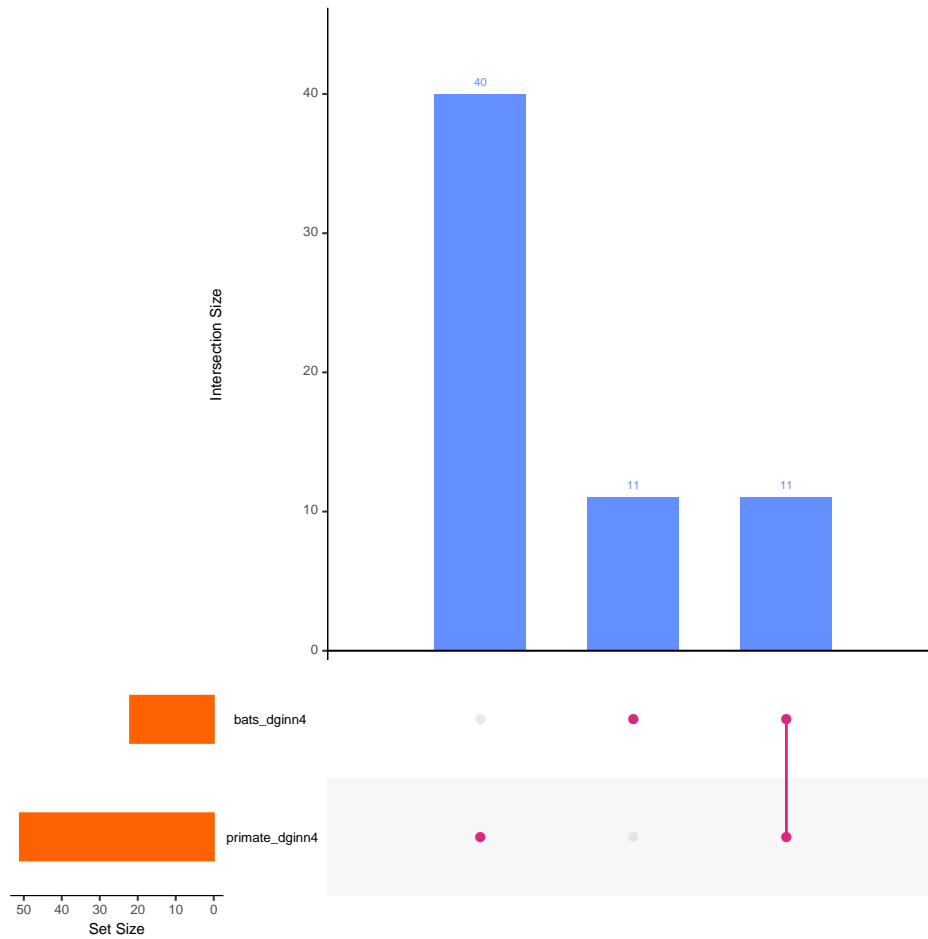
upset(monddata, nsets = 4, matrix.color = "#DC267F",
      main.bar.color = "#648FFF", sets.bar.color = "#FE6100")
```



```
upset(monddata[,1:3], nsets = 2, matrix.color = "#DC267F",
      main.bar.color = "#648FFF", sets.bar.color = "#FE6100")
```



```
upset(monddata[,c(1,4,5)], nsets = 2, matrix.color = "#DC267F",
      main.bar.color = "#648FFF", sets.bar.color = "#FE6100")
```



3 Which are these genes?

3.1 Gene under positive selection in both bats and primates

4 methods:

```
monddata[monddata$bats_dginn4==1 & monddata$primate_dginn4==1,]

##      tmp$Gene.name bats_dginn3 primate_dginn3 bats_dginn4
## 6          ACADM          1          1          1
## 7          ACE2          1          1          1
## 109          GGH          1          1          1
```

##	117	GOLGA7	1	1	1
##	134	IDE	1	1	1
##	139	ITGB1	1	1	1
##	146	LMAN2	1	1	1
##	212	POLA1	1	1	1
##	263	SLC27A2	1	1	1
##	301	TOR1AIP1	1	1	1
##	314	VPS39	1	1	1
##		primate_dginn4			
##	6		1		
##	7		1		
##	109		1		
##	117		1		
##	134		1		
##	139		1		
##	146		1		
##	212		1		
##	263		1		
##	301		1		
##	314		1		

3 methods:

```
monddata[monddata$bats_dginn3==1 & monddata$primate_dginn3==1,]
```

##	tmp\$Gene.name	bats_dginn3	primate_dginn3	bats_dginn4
##	6	ACADM	1	1
##	7	ACE2	1	1
##	9	ADAM9	1	0
##	34	CDK5RAP2	1	0
##	71	EDEM3	1	1
##	109	GGH	1	1
##	117	GOLGA7	1	1
##	134	IDE	1	1
##	139	ITGB1	1	1
##	146	LMAN2	1	1
##	157	MIPOL1	1	0
##	159	MOV10	1	0
##	212	POLA1	1	1
##	239	RAP1GDS1	1	1

##	257	SCCPDH	1	1	0
##	263	SLC27A2	1	1	1
##	301	TOR1AIP1	1	1	1
##	314	VPS39	1	1	1
##		primate_dginn4			
##	6		1		
##	7		1		
##	9		0		
##	34		1		
##	71		0		
##	109		1		
##	117		1		
##	134		1		
##	139		1		
##	146		1		
##	157		1		
##	159		0		
##	212		1		
##	239		0		
##	257		0		
##	263		1		
##	301		1		
##	314		1		

3.2 Gene under positive selection only in primates

4 methods:

```
monddata[monddata$bats_dginn4==0 & monddata$primate_dginn4==1,]

##      tmp$Gene.name bats_dginn3 primate_dginn3 bats_dginn4
## 31      BRD4           0           1           0
## 34    CDK5RAP2         1           1           0
## 37    CEP135           0           1           0
## 40    CEP68           0           1           0
## 47    CLIP4           0           1           0
## 67    DNMT1           0           1           0
## 68    DPH5           0           1           0
## 75    EMC1           0           1           0
```

## 80	ER01B	0	1	0
## 101	FYC01	0	1	0
## 105	GCC2	0	1	0
## 110	GHITM	0	1	0
## 111	GIGYF2	0	1	0
## 112	GLA	0	1	0
## 127	HECTD1	0	1	0
## 143	LARP1	0	1	0
## 144	LARP4B	0	1	0
## 150	MARK1	0	1	0
## 157	MIPOL1	1	1	0
## 160	MPHOSPH10	0	1	0
## 166	MYCBP2	0	1	0
## 171	NDUFAF2	0	1	0
## 172	NDUFB9	0	1	0
## 187	NUP58	0	1	0
## 195	PCNT	0	1	0
## 218	PRIM2	0	1	0
## 220	PRKAR2A	0	1	0
## 227	PVR	0	1	0
## 245	REEP6	0	1	0
## 248	RIPK1	0	1	0
## 253	SAAL1	0	1	0
## 259	SEPSECS	0	1	0
## 261	SIRT5	0	1	0
## 262	SLC25A21	0	1	0
## 296	TMEM39B	0	1	0
## 298	TMPRSS2	0	1	0
## 304	TUBGCP2	0	1	0
## 307	UBAP2	0	1	0
## 310	UGGT2	0	1	0
## 321	ZNF318	0	1	0
##	primate_dginn4			
## 31	1			
## 34	1			
## 37	1			
## 40	1			
## 47	1			
## 67	1			

```
## 68      1
## 75      1
## 80      1
## 101     1
## 105     1
## 110     1
## 111     1
## 112     1
## 127     1
## 143     1
## 144     1
## 150     1
## 157     1
## 160     1
## 166     1
## 171     1
## 172     1
## 187     1
## 195     1
## 218     1
## 220     1
## 227     1
## 245     1
## 248     1
## 253     1
## 259     1
## 261     1
## 262     1
## 296     1
## 298     1
## 304     1
## 307     1
## 310     1
## 321     1
```

3 methods:

```
monddata[monddata$bats_dginn3==0 & monddata$primate_dginn3==1,]

##      tmp$Gene.name bats_dginn3 primate_dginn3 bats_dginn4
```

## 19	AP2A2	0	1	0
## 23	ATE1	0	1	0
## 31	BRD4	0	1	0
## 32	BZW2	0	1	0
## 37	CEP135	0	1	0
## 40	CEP68	0	1	0
## 47	CLIP4	0	1	0
## 48	CNTRL	0	1	0
## 67	DNMT1	0	1	0
## 68	DPH5	0	1	0
## 72	EIF4E2	0	1	0
## 75	EMC1	0	1	0
## 80	ER01B	0	1	0
## 83	EXOSC2	0	1	0
## 101	FYC01	0	1	0
## 105	GCC2	0	1	0
## 110	GHITM	0	1	0
## 111	GIGYF2	0	1	0
## 112	GLA	0	1	0
## 118	GOLGB1	0	1	0
## 119	GORASP1	0	1	0
## 125	HDAC2	0	1	0
## 127	HECTD1	0	1	0
## 131	HS6ST2	0	1	0
## 143	LARP1	0	1	0
## 144	LARP4B	0	1	0
## 145	LARP7	0	1	0
## 150	MARK1	0	1	0
## 154	MDN1	0	1	0
## 160	MPHOSPH10	0	1	0
## 164	MRPS5	0	1	0
## 166	MYCBP2	0	1	0
## 168	NAT14	0	1	0
## 171	NDUFAF2	0	1	0
## 172	NDUFB9	0	1	0
## 176	NGLY1	0	1	0
## 181	NPC2	0	1	0
## 187	NUP58	0	1	0
## 195	PCNT	0	1	0

##	202	PITRM1	0	1	0
##	204	PLAT	0	1	0
##	208	PLOD2	0	1	0
##	210	PMPCB	0	1	0
##	214	POR	0	1	0
##	218	PRIM2	0	1	0
##	220	PRKAR2A	0	1	0
##	224	PTBP2	0	1	0
##	227	PVR	0	1	0
##	230	RAB14	0	1	0
##	232	RAB1A	0	1	0
##	233	RAB2A	0	1	0
##	242	RBX1	0	1	0
##	245	REEP6	0	1	0
##	248	RIPK1	0	1	0
##	250	RPL36	0	1	0
##	253	SAAL1	0	1	0
##	259	SEPSECS	0	1	0
##	261	SIRT5	0	1	0
##	262	SLC25A21	0	1	0
##	277	STOM	0	1	0
##	290	TIMM8B	0	1	0
##	296	TMEM39B	0	1	0
##	298	TMPRSS2	0	1	0
##	302	TRIM59	0	1	0
##	303	TRMT1	0	1	0
##	304	TUBGCP2	0	1	0
##	307	UBAP2	0	1	0
##	310	UGGT2	0	1	0
##	312	USP54	0	1	0
##	321	ZNF318	0	1	0
##		primate_dginn4			
##	19		0		
##	23		0		
##	31		1		
##	32		0		
##	37		1		
##	40		1		
##	47		1		

## 48	0
## 67	1
## 68	1
## 72	0
## 75	1
## 80	1
## 83	0
## 101	1
## 105	1
## 110	1
## 111	1
## 112	1
## 118	0
## 119	0
## 125	0
## 127	1
## 131	0
## 143	1
## 144	1
## 145	0
## 150	1
## 154	0
## 160	1
## 164	0
## 166	1
## 168	0
## 171	1
## 172	1
## 176	0
## 181	0
## 187	1
## 195	1
## 202	0
## 204	0
## 208	0
## 210	0
## 214	0
## 218	1
## 220	1

```
## 224      0
## 227      1
## 230      0
## 232      0
## 233      0
## 242      0
## 245      1
## 248      1
## 250      0
## 253      1
## 259      1
## 261      1
## 262      1
## 277      0
## 290      0
## 296      1
## 298      1
## 302      0
## 303      0
## 304      1
## 307      1
## 310      1
## 312      0
## 321      1
```

3.3 Gene under positive selection only in bats

4 methods:

```
monddata[monddata$bats_dginn4==1 & monddata$primate_dginn4==0,]

##      tmp$Gene.name bats_dginn3 primate_dginn3 bats_dginn4
## 14      AKAP9          1          0          1
## 26      ATP6AP1        1          0          1
## 44      CISD3          1          0          1
## 71      EDEM3          1          1          1
## 77      ERGIC1         1          0          1
## 136     IMPDH2         1          0          1
## 137     INHBE          1          0          1
```

```
## 231      RAB18      1      0      1
## 239      RAP1GDS1    1      1      1
## 267      SLC44A2    1      0      1
## 283      TBK1      1      0      1
##      primate_dginn4
## 14      0
## 26      0
## 44      0
## 71      0
## 77      0
## 136     0
## 137     0
## 231     0
## 239     0
## 267     0
## 283     0
```

3 methods:

```
monddata[monddata$bats_dginn3==1 & monddata$primate_dginn3==0,]
```

```
##      tmp$Gene.name bats_dginn3 primate_dginn3 bats_dginn4
## 5      ACAD9      1      0      0
## 11     AGPS      1      0      0
## 14     AKAP9      1      0      1
## 26     ATP6AP1    1      0      1
## 44     CISD3      1      0      1
## 49     COL6A1     1      0      0
## 77     ERGIC1     1      0      1
## 122    GRIPAP1    1      0      0
## 123    GRPEL1     1      0      0
## 136    IMPDH2     1      0      1
## 137    INHBE      1      0      1
## 151    MARK2      1      0      0
## 185    NUP214     1      0      0
## 217    PRIM1      1      0      0
## 226    PUSL1      1      0      0
## 231    RAB18      1      0      1
## 266    SLC30A9    1      0      0
## 267    SLC44A2    1      0      1
```

```
## 268      SLC9A3R1      1      0      0
## 283      TBK1      1      0      1
##      primate_dginn4
## 5      0
## 11     0
## 14     0
## 26     0
## 44     0
## 49     0
## 77     0
## 122    0
## 123    0
## 136    0
## 137    0
## 151    0
## 185    0
## 217    0
## 226    0
## 231    0
## 266    0
## 267    0
## 268    0
## 283    0
```

3.4 Figure tableau

```
tablo<-as.data.frame(tmp$Gene.name)
tablo$nbats<-batstmp
tablo$nprimates<-primatetmp

plot(NULL, xlim=c(-0.5,5.5), ylim=c(-3,5.5), xlab="bats", ylab="primates", main="Gene

text(x=rep(-0.6, 6), y=0:5, 0:5)
text(y=rep(-0.65, 6), x=0:5, 0:5)
sapply(seq(from=-0.5, to=5.5, by=1), function(x){
  segments(x0=x, x1=x, y0=-0.5, y1=5.5)
})
```

```

## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL

sapply(seq(from=-0.5, to=5.5, by=1), function(x){
  segments(x0=-0.5, x1=5.5, y0=x, y1=x)
})

## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##

```

```

## [[6]]
## NULL
##
## [[7]]
## NULL

for (p in 0:5){
  for (b in 0:5){
    tmp<-tablo$tmp$Gene.name`[tablo$nbats==b & tablo$nprimates==p]
    if(length(tmp)>0 & length(tmp)<=8){
      text(b,seq(from=(p-0.4), to=(p+0.4), length.out = length(tmp)), tmp, cex=0.4)
    }else if (length(tmp)>8 & length(tmp)<=16){
      print(c(p, b))
      text((b-0.3),seq(from=(p-0.4), to=(p+0.4), length.out = 8), tmp[1:8], cex=0.4)
      text((b+0.3),seq(from=(p-0.4), to=(p+0.4), length.out = (length(tmp)-8)), tmp[9
    ]else if (length(tmp)>16){
      text(b,p, paste0(length(tmp), " values"))
    }
  }
}

## [1] 1 2
## [1] 2 0
## [1] 2 1
## [1] 2 2
## [1] 3 0
## [1] 3 1
## [1] 4 0
## [1] 4 1

tmp<-tablo$tmp$Gene.name`[tablo$nbats==0 & tablo$nprimates==1]
text(-0.4,-1.2, "p=1/n=0", cex=0.6)
text(seq(from=0.1, to=5.5, length.out = 18),-1.1, tmp[1:18], cex=0.4)
text(seq(from=0.1, to=5.5, length.out = length(tmp)-18),-1.3, tmp[19:length(tmp)], ce

tmp<-tablo$tmp$Gene.name`[tablo$nbats==1 & tablo$nprimates==1]
text(-0.4,-1.7, "p=1/n=1", cex=0.6)
text(seq(from=0.1, to=5.5, length.out = 18),-1.6, tmp[1:18], cex=0.4)
text(seq(from=0.1, to=4.5, length.out = length(tmp)-18),-1.8, tmp[19:length(tmp)], ce

```

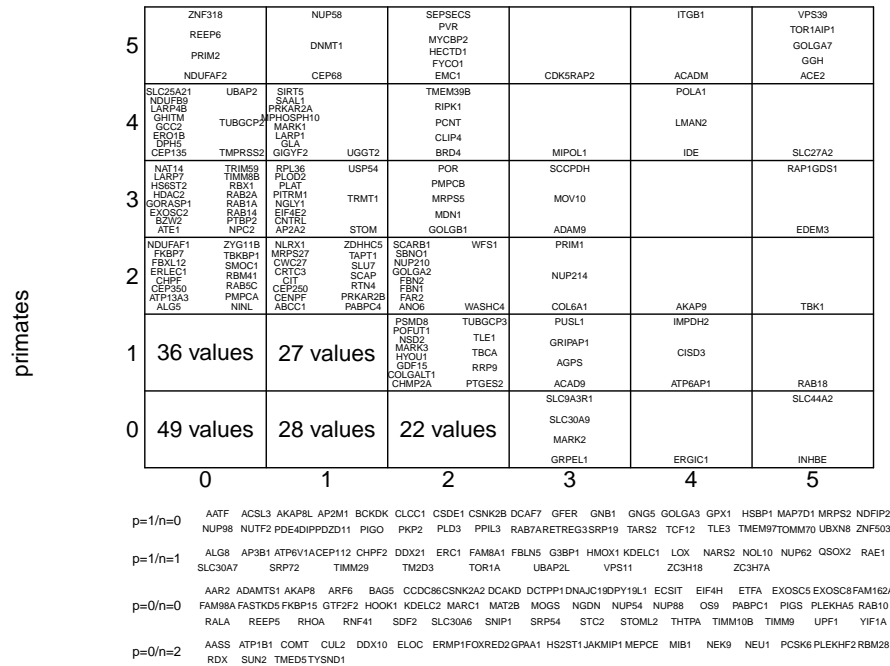
```

tmp<-tablo$`tmp$Gene.name`[tablo$nbats==0 & tablo$nprimates==0]
text(-0.4,-2.3, "p=0/n=0", cex=0.6)
text(seq(from=0.1, to=5.5, length.out = 17),-2.1, tmp[1:17], cex=0.4)
text(seq(from=0.1, to=5.5, length.out = 17),-2.3, tmp[18:34], cex=0.4)
text(seq(from=0.1, to=5.5, length.out = length(tmp)-34),-2.5, tmp[35:length(tmp)], cex=0.4)

tmp<-tablo$`tmp$Gene.name`[tablo$nbats==2 & tablo$nprimates==0]
text(-0.4,-2.9, "p=0/n=2", cex=0.6)
text(seq(from=0.1, to=5.5, length.out = 18),-2.8, tmp[1:18], cex=0.4)
text(seq(from=0.1, to=1, length.out = length(tmp)-18),-3.0, tmp[19:length(tmp)], cex=0.4)

```

Genes supported by x,y methods in bats and primates



```
write.csv(tablo[tablo$nbats>=3,"tmp$Gene.name"], "batssup3.csv", row.names=FALSE, quo
write.csv(tablo[tablo$nprimates>=3,"tmp$Gene.name"], "primatessup3.csv", row.names=FA
write.csv(tablo, "primatesVbats.csv", row.names=FALSE, quote=FALSE)
```