

Positive selection on genes interacting with SARS-Cov2, comparing GWAS and PS on GWAS

Marie Cariou

March 2021

Contents

1	Data	2
1.1	GWAS 1	2
1.2	GWAS 2	2
1.3	DGINN data	2
1.4	FYCO1	3
1.5	Objectif	3
2	GWAS data	3

1 Data

```
library(shape)
```

1.1 GWAS 1

-Article:

<https://www.nejm.org/doi/full/10.1056/NEJMoa2020283>

-Data:

https://ikmb.shinyapps.io/COVID-19_GWAS_Browser/

-Locally:

```
home<-"/home/adminmarie/Documents/CIRI_BIBS_projects/"
datapath<-paste0(home, "2020_05_Etienne_covid/data/GWAS/")
gwas1<-paste0(datapath, "meta_analysis_II.hg38.gwascatalogformat.tsv")
```

1.2 GWAS 2

-Article:

Genetic mechanisms of critical illness in COVID-19

<https://www.nature.com/articles/s41586-020-03065-y>

-Data:

<https://genomicc.org/data/>

-Locally:

```
gwas2<-list.files(datapath, pattern="genomicc")
gwas2<-paste0(datapath, gwas2)
```

1.3 DGINN data

```
home<-"/home/adminmarie/Documents/CIRI_BIBS_projects/"
tabpath<-paste0(home, "2020_05_Etienne_covid/2020_dginn_covid19/")
```

```
#table

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

## [1] 442 56

# fasta
fasta<-list.files(datapath, pattern="FYCO")
```

1.4 FYCO1

https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000163820;r=3:45917899-45995824

coordinate: Chromosome 3: 45,917,899-45,995,824

1.5 Objectif

Table :

- pos dans genome de ref
- pos dans l'aln
- under PS oui/non
- GWAS1 oui/non

2 GWAS data

```
deb<-45917899
end<-45995824
len<-45995824-45917899
thres<-0.01

posref<-deb:end

fyco1tab<-as.data.frame(posref)
```

```

# GWAS 1
cmd<-paste0("cat ", gwas1, " | grep '^3' > tmp")
system(cmd)
gwastab<-read.table("tmp", h=T)
system("rm tmp")

gwastab<-gwastab[(gwastab$chromosome==3 & gwastab$base_pair_location>deb & gwastab$base_pair_location<end),]

## filtrer sur pvalue et créer colonne F/T GWAS
gwas1pos<-gwastab$base_pair_location[gwastab$p_value<thres]
fyco1tab$gwas1<-ifelse(posref %in% gwas1pos, TRUE, FALSE)

## Same for GWAS 2
gwastab<-read.table(gwas2[1],h=T, sep="\t")
gwastab<-gwastab[(gwastab$CHR==3 & gwastab$POS>deb & gwastab$POS<end),]
gwas1pos<-gwastab$POS[gwastab$Pval<thres]
file1<-ifelse(posref %in% gwas1pos, TRUE, FALSE)

gwastab<-read.table(gwas2[2],h=T, sep="\t")
gwastab<-gwastab[(gwastab$CHR==3 & gwastab$POS>deb & gwastab$POS<end),]
gwas1pos<-gwastab$POS[gwastab$Pval<thres]
file2<-ifelse(posref %in% gwas1pos, TRUE, FALSE)

gwastab<-read.table(gwas2[3],h=T, sep="\t")
gwastab<-gwastab[(gwastab$CHR==3 & gwastab$POS>deb & gwastab$POS<end),]
gwas1pos<-gwastab$POS[gwastab$Pval<thres]
file3<-ifelse(posref %in% gwas1pos, TRUE, FALSE)

gwastab<-read.table(gwas2[4],h=T, sep="\t")
gwastab<-gwastab[(gwastab$CHR==3 & gwastab$POS>deb & gwastab$POS<end),]
gwas1pos<-gwastab$POS[gwastab$Pval<thres]
file4<-ifelse(posref %in% gwas1pos, TRUE, FALSE)

fyco1tab$gwas2<-((file1 | file2) | file3 ) | file4

## Voir cohérence entre les 2

table(fyco1tab[,2:3])

```

```
##          gwas2
## gwas1    FALSE  TRUE
##    FALSE 77885    41
```

How to convert coordinates from alignments to genomic coordinate to the multi species alignment coordinates?

1. I downloaded the cds sequence to know for which transcript I should refer to FYCO1-205 is the right length. ENST00000535325.5, le cds de référence fait 4497 bp, l'alignement fait 4503 pb FYCO1-201 has one exon missing, as most of the sequences in the alignment

I will get the coordinates of exons within this transcript ENST00000535325.5

```
"cat /Xnfs/ciridb/shared_data/Genomes/Ensembl/Hs/GRCh38/Annot/v77/Homo_sapiens.GRCh38.7
```

```
— grep "ENST00000535325" — awk 'print 1"3 " " 4"5 " " 10"14 " "
17"18' ; human_fyco1.gtf"
```

There might be a max 6bp discrepancy between positions, we will see...

gtf

Le gene est en antisens co

```
gtf<-read.table("../data/GWAS/human_fyco1.gtf", h=F)

#gtf$V3=gtf$V3-45900000
#gtf$V4=gtf$V4-45900000

cds<-gtf[gtf$V2 %in% c("CDS"),]
cds$len=cds$V4-cds$V3
sum(cds$len)

## [1] 4476

# I think there is an exon missing outdated gtf?

#make a vector with positions

cat<-numeric()
for (i in 1:nrow(cds)){
  tmp<-cds$V3[i]:cds$V4[i]
  cat<-c(cat,tmp)
}

cat<-sort(cat)
```

```

cds_tab<-cbind(cat, length(cat):1)
colnames(cds_tab)<-c("posref", "pos_aln")

fyco1tab<-merge(fyco1tab, cds_tab, by="posref", all.x=T)

```

add PS sites

```

PS<-tab[tab$Gene.name=="FYC01", "dginn.primate_MEME.PSS"]
PS<-as.numeric(unlist(strsplit(as.character(PS), split = ", ")))

```

```

head(fyco1tab)

##      posref gwas1 gwas2 pos_aln
## 1 45917899 FALSE FALSE      NA
## 2 45917900 FALSE FALSE      NA
## 3 45917901 FALSE FALSE      NA
## 4 45917902 FALSE FALSE      NA
## 5 45917903 FALSE FALSE      NA
## 6 45917904 FALSE FALSE      NA

fyco1tab$PS<-FALSE
fyco1tab$PS[fyco1tab$pos_aln %in% PS]<-TRUE

plot(fyco1tab$posref, rep(1, nrow(fyco1tab)), cex=0.2,
      xlab="chromosome 3", ylab="")

points(fyco1tab$posref[is.na(fyco1tab$pos_aln)==FALSE],
        rep(1, length(fyco1tab$posref[is.na(fyco1tab$pos_aln)==FALSE])),
        cex=3, pch=15, col="lightblue")

posgwas2<-fyco1tab$posref[fyco1tab$gwas2==TRUE]

Arrows(x0=posgwas2, y0=rep(0.94, length(posgwas2)),
        x1=posgwas2, y1=rep(0.96, length(posgwas2)),
        arr.type="triangle", arr.length=0.6)

text(45930000, 1.2, "PS")

posps<-fyco1tab$posref[fyco1tab$PS==TRUE]

```

```

Arrows(x0=posps, y0=rep(1.06, length(posps)),
       x1=posps, y1=rep(1.04, length(posps)),
       arr.type="triangle", arr.length=0.6)

text(45930000, 0.8, "GWAS, pval<0.01, Pairo-Castineira et al.")

```

