GENESPACE MS part 5: Grass rho paralogs

17 February, 2022

```r
wd <- "/Users/lovell/Desktop/manuscripts/genespace_2022/GENESPACE_data/results/"
library(GENESPACE)
```

```
## Loading required package: data.table
```

```r
library(ggplot2)
library(Biostrings)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:data.table':
##
##     first, second
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:data.table':
##
##     shift

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit
```

```r
load(file.path(wd, "gparRho.rda"))
gpar <- gparRho

hits <- fread(file.path(wd, "rho/results/Sviridis_Phallii_synHits.txt.gz"), na.strings = c("", "NA"))
hitsog <- subset(hits, isOg & isAnchor & isRep1 & isRep2)

# pepHal <- readAAStringSet("GENESPACE_data/results/rho/peptide/Phallii.fa")
# pepSet <- readAAStringSet("GENESPACE_data/results/rho//peptide/Sviridis.fa")
# gff <- fread("GENESPACE_data/results/rho/results/gffWithOgs.txt.gz")
# idv <- gff$id; names(idv) <- gff$ofID
# idvSet <- idv[unique(hitsog$ofID1)]
# idvHal <- idv[unique(hitsog$ofID2)]
# alignList <- mclapply(1:nrow(hitsog), mc.cores = 8, function(i){
#   s1 <- AAString(pepSet[[idvSet[hitsog$ofID1[i]]]])
#   s2 <- AAString(pepHal[[idvHal[hitsog$ofID2[i]]]])
#   alg <- pairwiseAlignment(s1, s2)
#   return(alg)
# })
# save(alignList, file = "/Users/lovell/Desktop/manuscripts/genespace_2022/analysis/Sviridis_Phallii_al
load("/Users/lovell/Desktop/manuscripts/genespace_2022/analysis/Sviridis_Phallii_align.rda")
hitsog[,`:=`(pid1 = sapply(alignList, pid, type = "PID1"),
             pid2 = sapply(alignList, pid, type = "PID2"))]
hitsog[,isRho := grepl("sec", blkID)]
hitsog[,isOverRetain := grepl("prim", blkID) & chr1 == "Chr_08" & chr2 == "Chr03"]
hitsog[,isDefOg := gsub("_", "", chr1) == chr2]
hitsog[,blkcat := ifelse(isRho, "rho", ifelse(isDefOg, "orth", ifelse(isOverRetain, "overr", "ambig")))]
with(subset(hitsog, blkcat %in% c("rho", "orth")), wilcox.test(pid2 ~ blkcat))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pid2 by blkcat
## W = 88094632, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
c8_c3 <- subset(hitsog, grepl("prim", blkID) & chr1 == "Chr_08" & chr2 == "Chr03")
print(svRng <- range(c(c8_c3$start1, c8_c3$end1)))
```

[1] 236592 3432781

```
print(phRng <- range(c(c8_c3$start2, c8_c3$end2)))
```

[1] 283046 2355430

```
abs(diff(svRng))
```

[1] 3196189

```
abs(diff(phRng))
```

[1] 2072384

```
c8 <- subset(hitsog, chr1 == "Chr_08" & start1 < 10e6)
# pdf("../figures/Fig2_window.pdf", height = 5, width = 5)
ggplot(c8, aes(x = start1/1e6, y = pid2, col = blkcat == "orth"))+
  geom_point(size = .25)+
  stat_smooth(span = .25)+
  scale_y_continuous(limits = c(50, 100), expand = c(0,0))+
  scale_x_continuous(expand = c(0,0))+
  geom_vline(xintercept = c(2180878/1e6, 283046/1e6), col = "red", lty = 2)+
  scale_color_manual(values = c("darkorange", "dodgerblue2"), guide = "none")+
  theme(panel.background = element_rect(fill = "black"),
        panel.grid.minor = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major = element_line(size = .1, linetype = 2))
```

'geom_smooth()' using method = 'loess' and formula 'y ~ x'

Warning: Removed 14 rows containing non-finite values (stat_smooth).

Warning: Removed 14 rows containing missing values (geom_point).
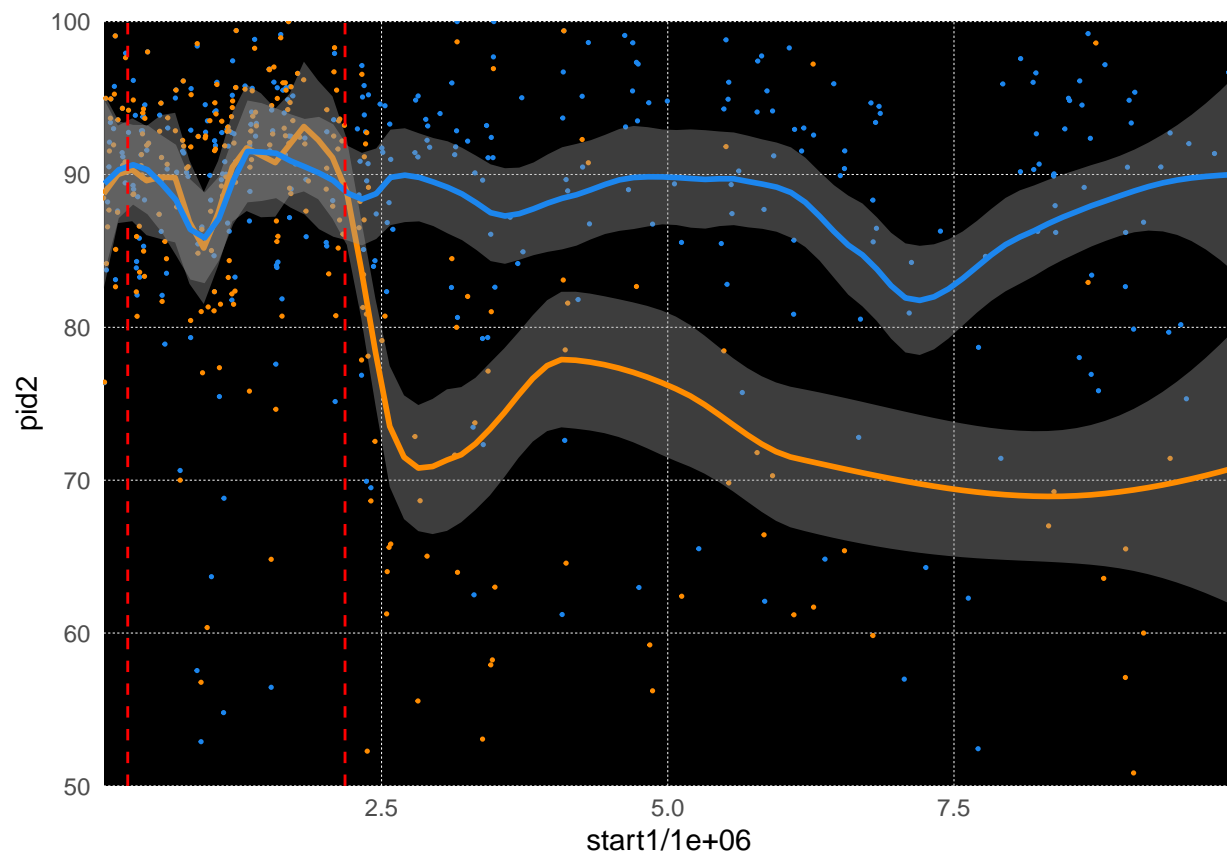
```
# dev.off()

c8 <- subset(hitsog, chr1 == "Chr_08" & blkcat %in% c("orth", "overr"))
c8 <- subset(c8, start1 >= svRng[1] & end1 <= svRng[2] &
              start2 >= phRng[1] & end2 <= phRng[2])
r8 <- subset(hitsog, chr1 == "Chr_08" & blkcat %in% c("orth", "rho"))
print(ct8 <- with(c8, tapply(pid2, blkcat, median)))
```

```
##      orth    overr
## 91.53736 90.96990
```

```
print(rt8 <- with(r8, tapply(pid2, blkcat, median)))
```

```
##      orth      rho
## 90.51627 71.42857
```

```
(ct8["orth"] - ct8["overr"])/ct8["orth"]
```

```
##        orth
## 0.006199218
```

```
(rt8["orth"] - rt8["rho"])/rt8["orth"]
```

```
##      orth
## 0.2108759
```

```
with(c8, wilcox.test(pid2 ~ blkcat))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pid2 by blkcat
## W = 16993, p-value = 0.2749
## alternative hypothesis: true location shift is not equal to 0
```

```
with(r8, wilcox.test(pid2 ~ blkcat))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  pid2 by blkcat
## W = 83140, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
orReg <- calc_blkCoords(subset(c8, blkcat == "overr"))
# -- overretained blocks on sviridis chr_08 and phallii chr03
setkey(orReg, startBp1)
knitr::kable(orReg[,c("blkID", "startBp1", "endBp1","startBp2", "endBp2", "nHits1", "orient")])
```
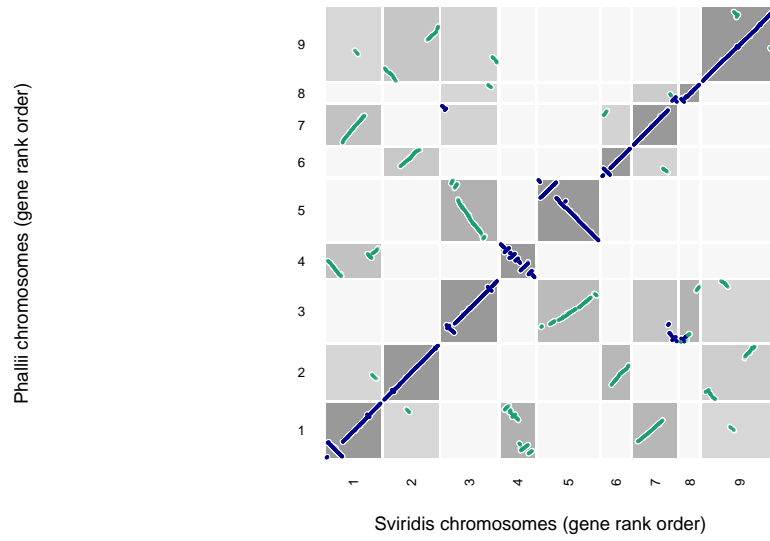
blkID	startBp1	endBp1	startBp2	endBp2	nHits1	orient
primary__133	236592	2343084	1795579	283046	92	-
primary__192	1308275	2178130	1194590	1749868	43	+
primary__12	1830798	2110510	1711034	1590208	13	-
primary__62	2119722	3432781	1712695	2355430	33	+

```
subset(c8, blkcat == "overr")[,list(SvSize = max(end1)-min(start1),
                                    PhSize = max(end2)-min(start2)),
                              by = "regID"]
```

```
##        regID  SvSize  PhSize
## 1: primary_34 3196189 2072384
```

```
hitsog[,n := uniqueN(ofID1), by = "blkID"]
h <- subset(hitsog, n > 20)
h[,regID := grepl("pri", regID)]
# pdf("../figures/Fig2_dotplots.pdf", height = 5, width = 5)
plot_hits(h,  reorderChrs = F, plotType = "regAnchor")
```

5

syntenic region anchors

Phalii chromosomes (gene rank order)

Sviridis chromosomes (gene rank order)

```
# dev.off()
```