DMR Comparison

Coleman Breen

2022-08-04

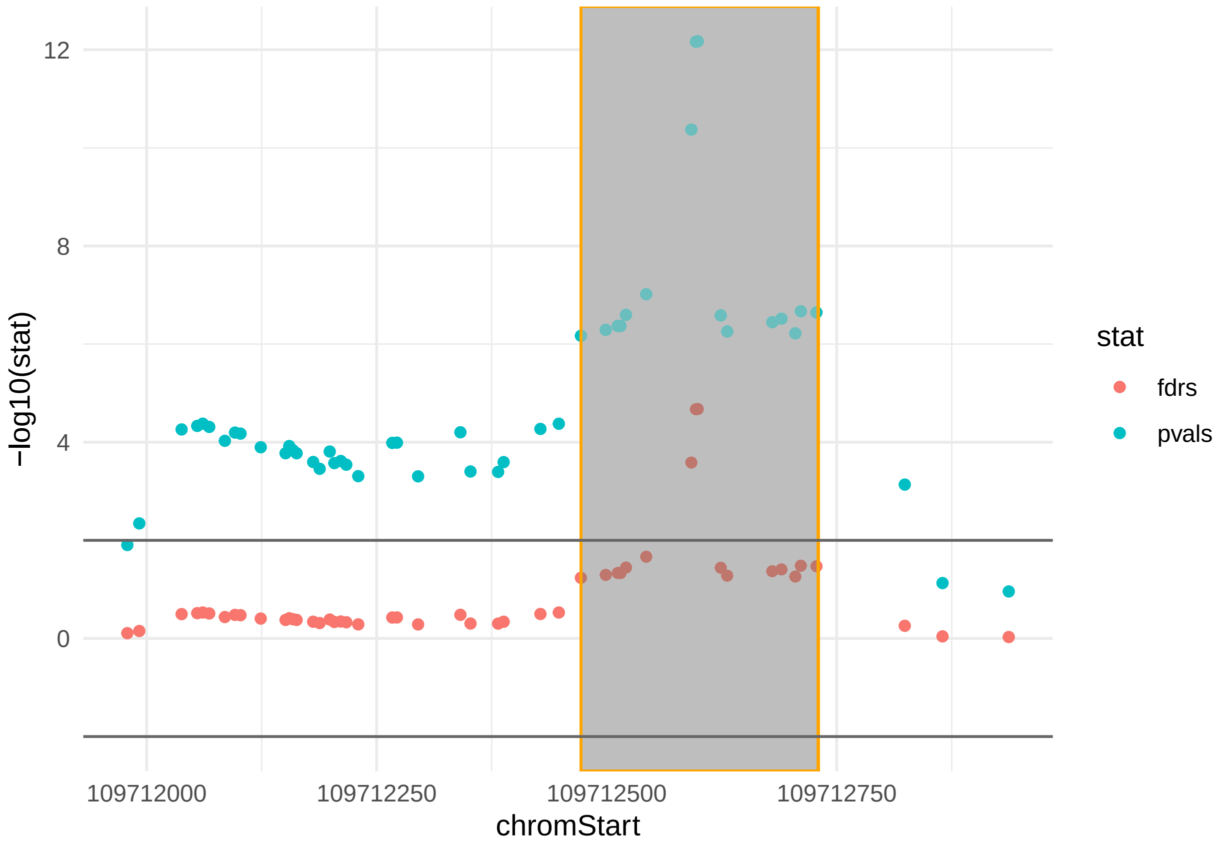
```r
cols <- c("chrom", "chromStart", "pvals",   "fdrs", "LOAD.minus.control")

dmrs.df <- fread("../data/07-counts-by-chrom-imputed-subset1/DMRs.adj.bed") %>%
  arrange(-nCG)
dmps.df <- fread("../data/07-counts-by-chrom-imputed-subset1/DMPs.bed",
                 select = cols)
```

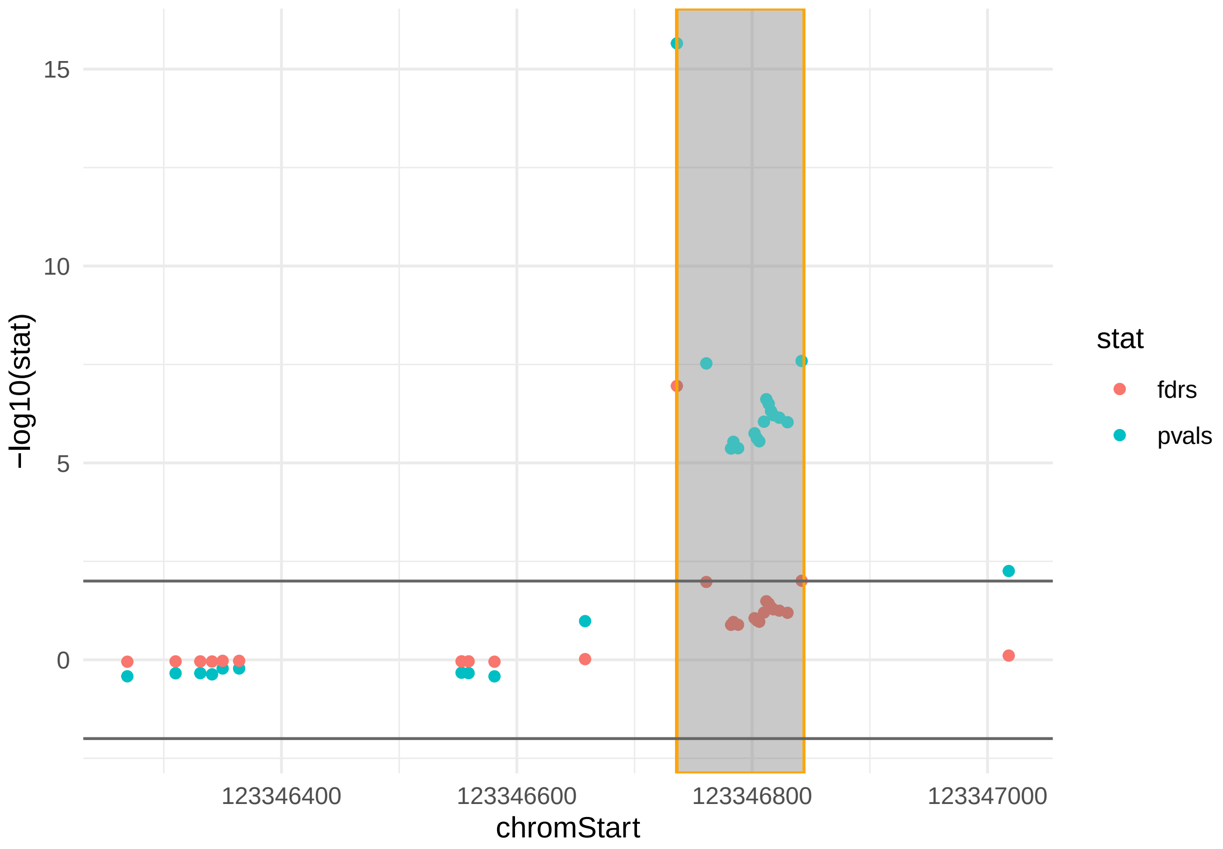
DMR calling in DSS is determined by p-values. I adjust the p-values using the same `fdrtool` that I've been using in the past (previously I would just plot $-\log_{10}(LFDR)$) but now I plot that in red, and also the corrected $-\log_{10}(p)$ in blue. What DSS considers a DMR is shaded in grey. Significance lines are in black at the $p < 0.01$ level.

I plot 500 bases additional on either side of the DMR to get a sense of what's happening nearby. Y-axes are not fixed, so be mindful of how small some of the p-values can get.

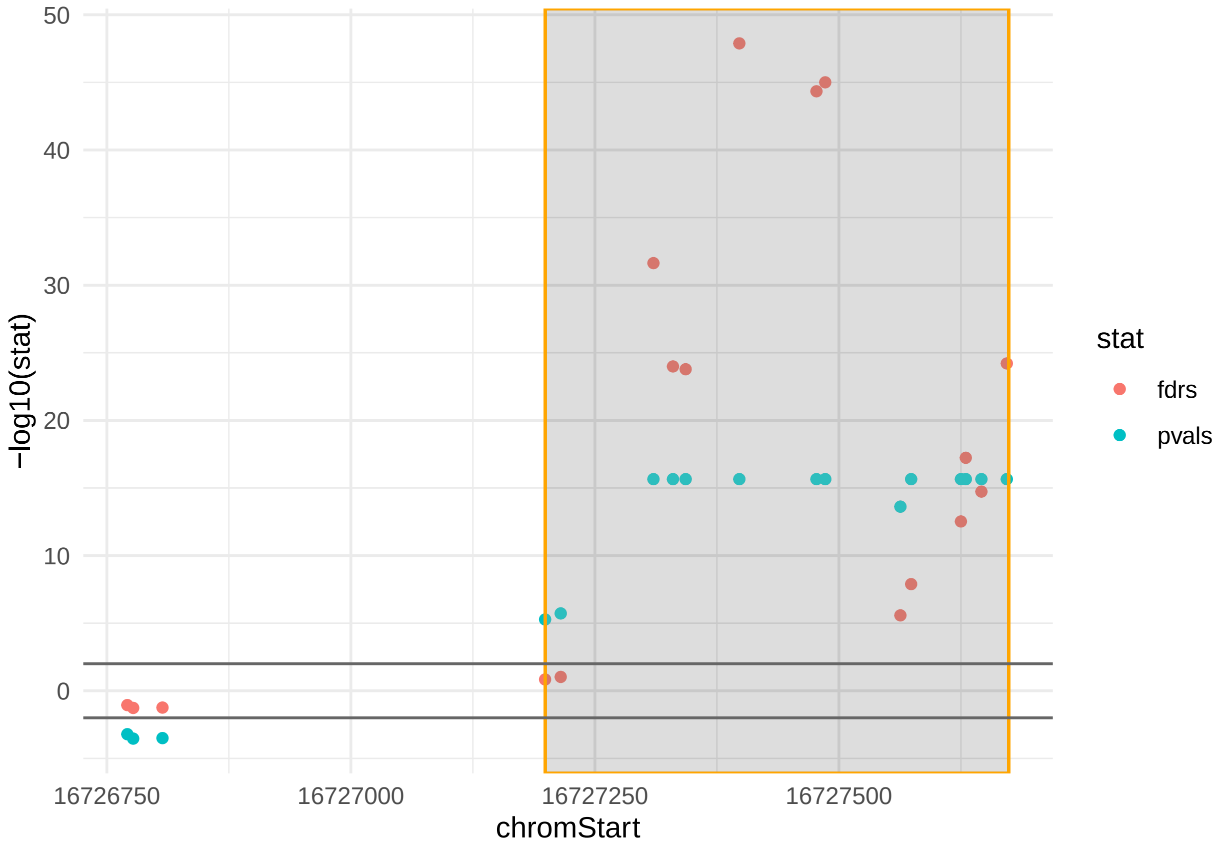
Here are 10 DMRs, in order of number of CpGs residing in the DMR.

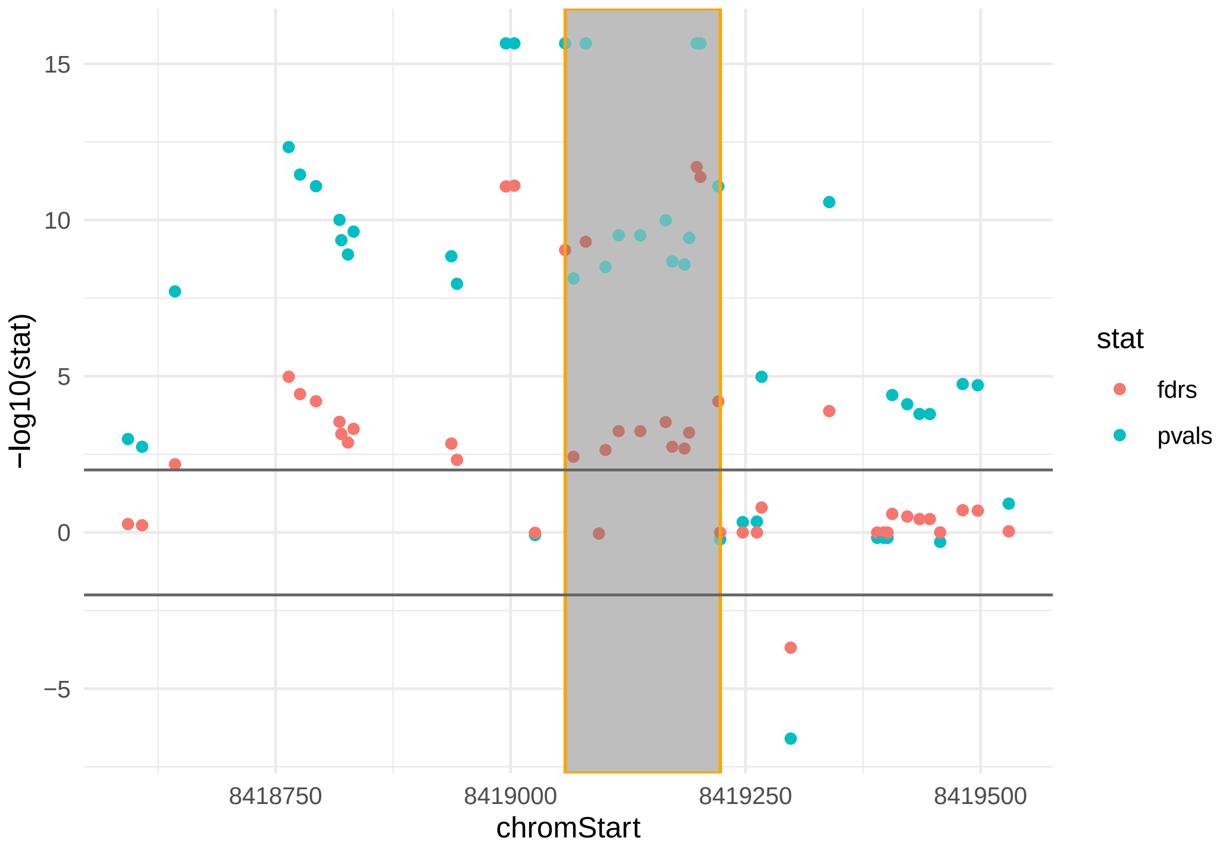
```r
pad <- 500


# nrow(dmrs.df
for (ii in 1:10){
  my_chrom <- dmrs.df$chrom[ii]
  my_start <- dmrs.df$chromStart[ii]
  left_lim <- my_start - pad

  my_end <- dmrs.df$chromEnd[ii]
  right_lim <- my_start + pad

  dmps.df %>%
    dplyr::filter(chrom == my_chrom) %>%
    dplyr::filter(chromStart > left_lim, chromStart + 1 < right_lim) %>%
    pivot_longer(cols = c(pvals, fdrs), values_to = "y", names_to = "stat") %>%
    dplyr::mutate(y = -1 * sign(LOAD.minus.control) * log10(y)) %>%
  ggplot(aes(x = chromStart, y = y, color = stat)) +
    geom_point() +
    theme_minimal() +
    geom_rect(aes(xmin = my_start, xmax = my_end, ymin = -Inf, ymax = Inf),
              alpha = 0.01, fill = "grey", color="orange") +
    geom_hline(yintercept = -log10(.01), color = "grey40") +
    geom_hline(yintercept = log10(.01), color = "grey40") +
    ylab("-log10(stat)") -> p
  print(p)
}
```

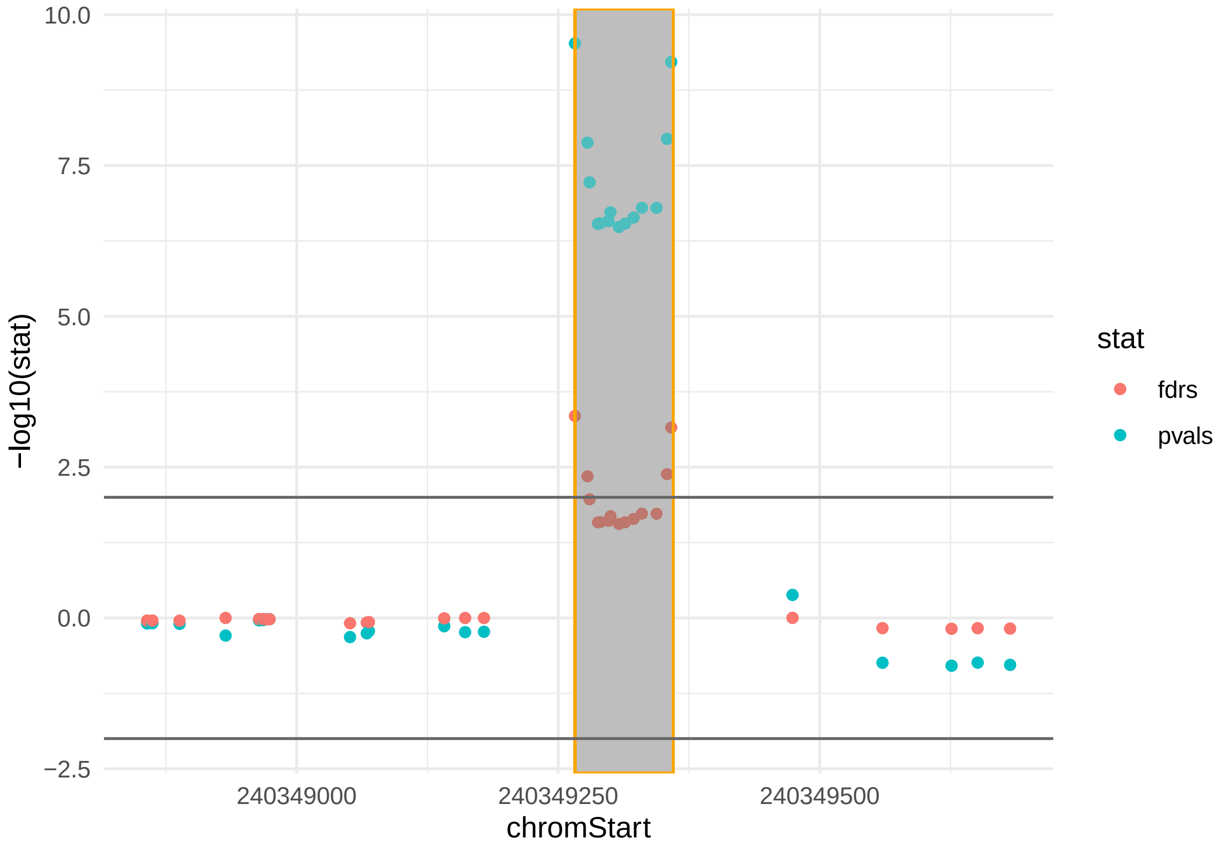

12
8
4
0
−log10(stat)
109712000
109712250
109712500
109712750
chromStart
stat
● fdrs
● pvals

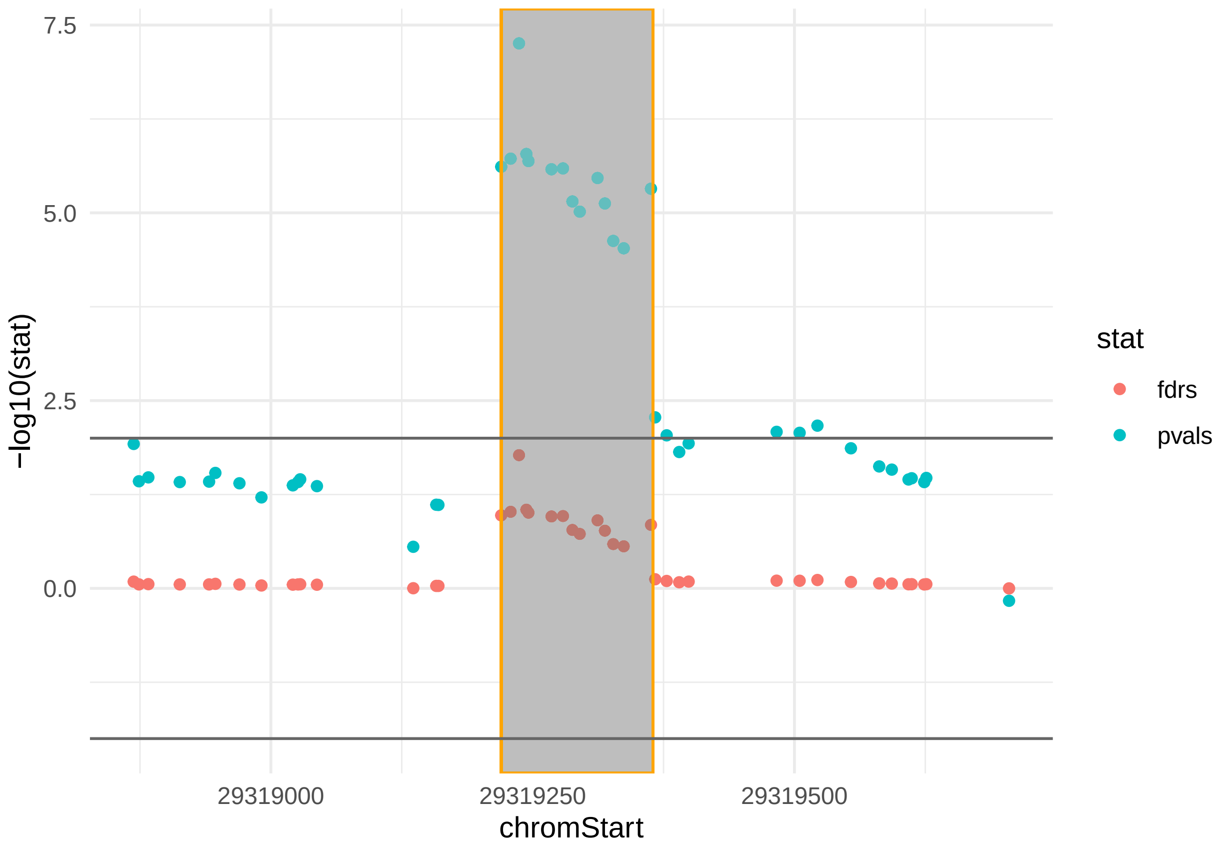


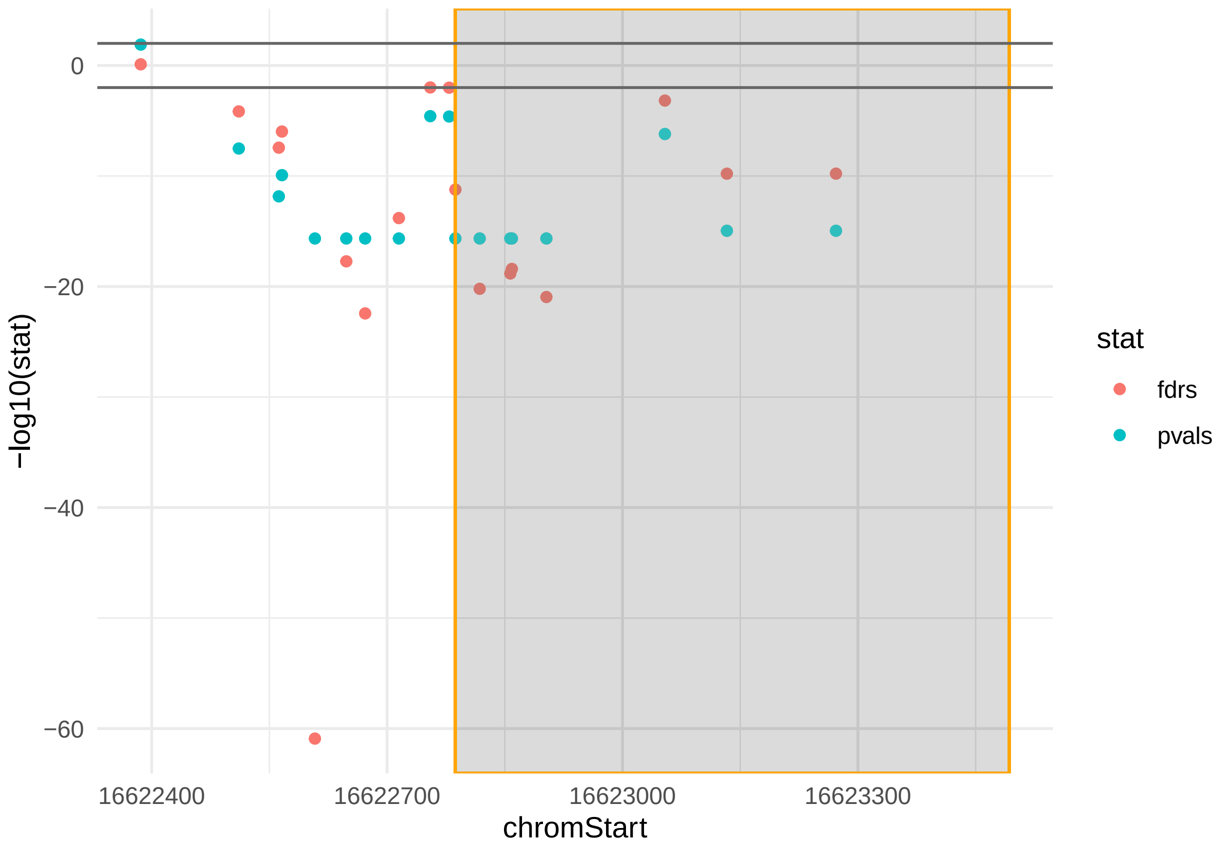
15
10
5
0
−log10(stat)
123346400
123346600
123346800
123347000
chromStart
stat
● fdrs
● pvals


50
40
30
20
10
0
−log10(stat)
16726750
16727000
16727250
16727500
chromStart
stat
fdrs
pvals

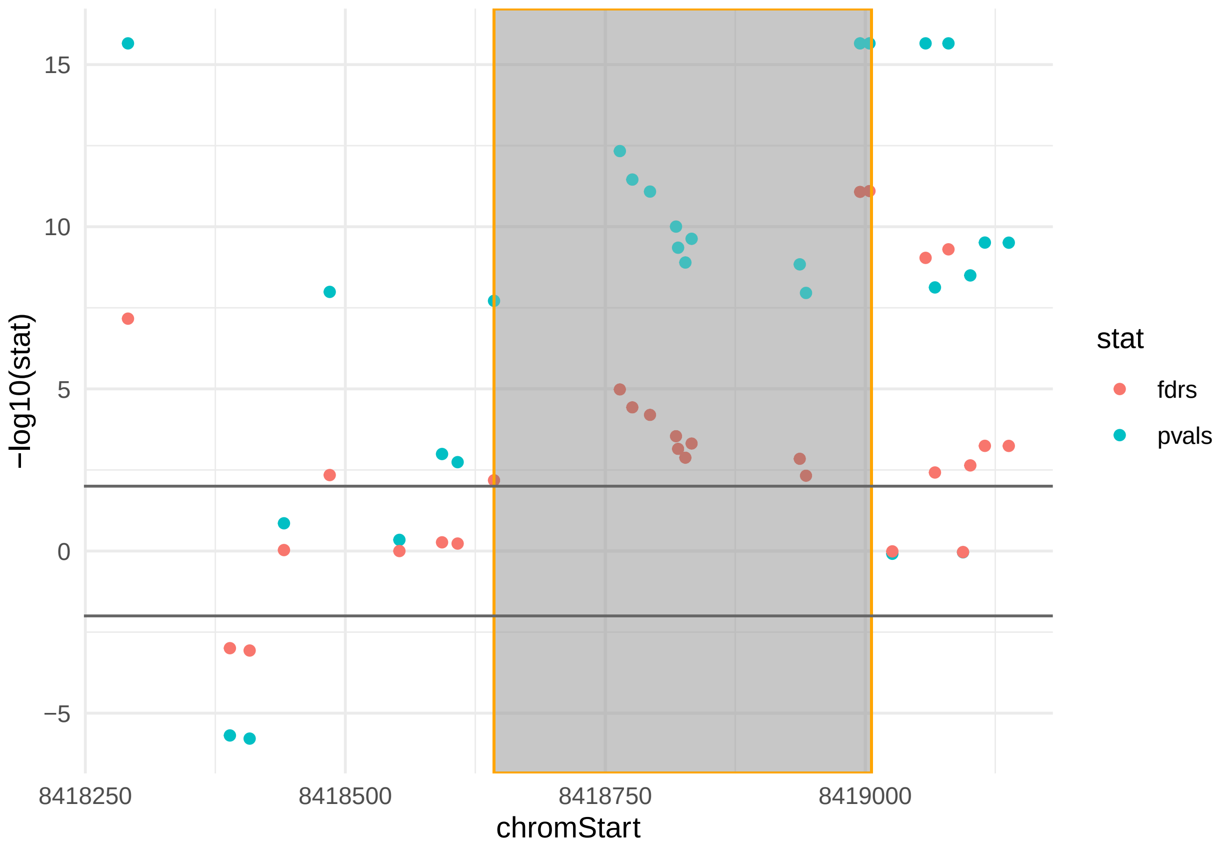

15
10
5
0
−5
−log10(stat)
8418750
8419000
8419250
8419500
chromStart
stat
● fdrs
● pvals

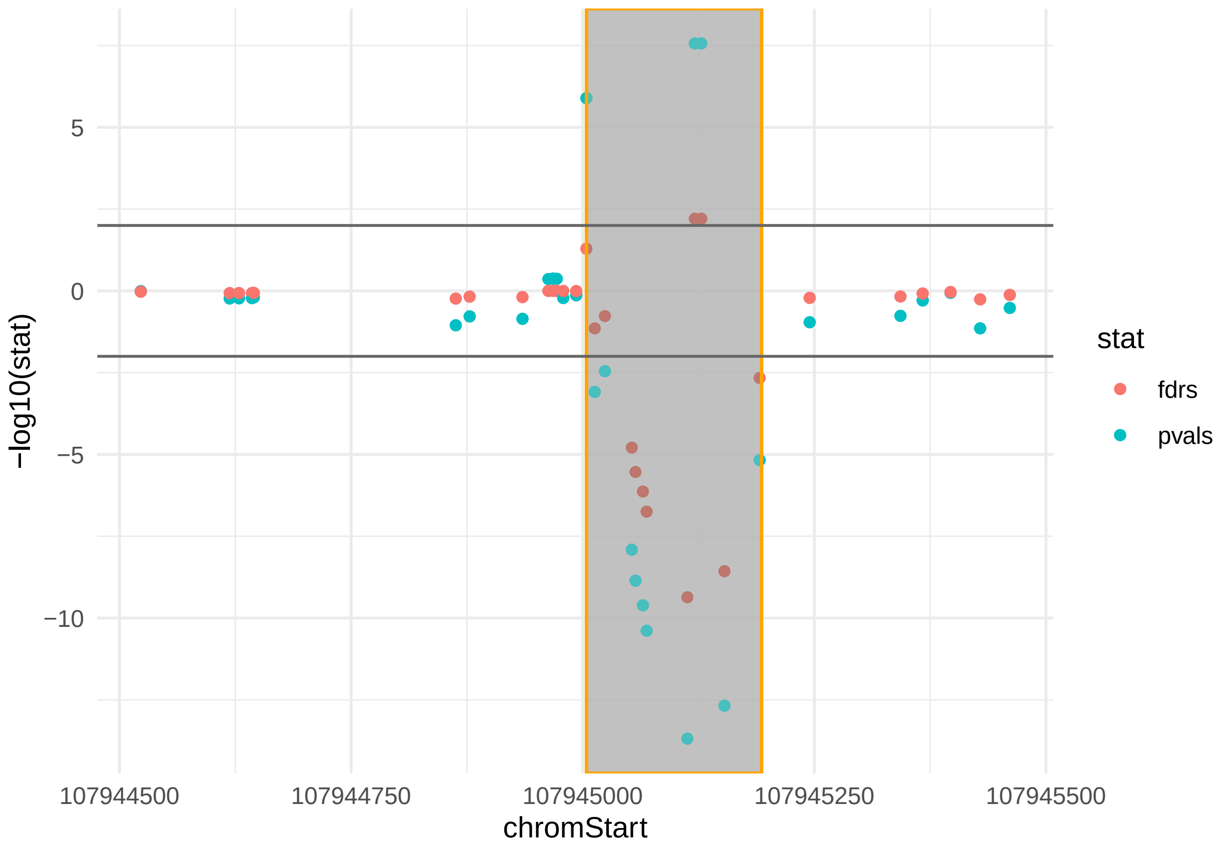

10.0
7.5
5.0
2.5
0.0
−2.5
−log10(stat)
240349000
240349250
240349500
chromStart
stat
● fdrs
● pvals

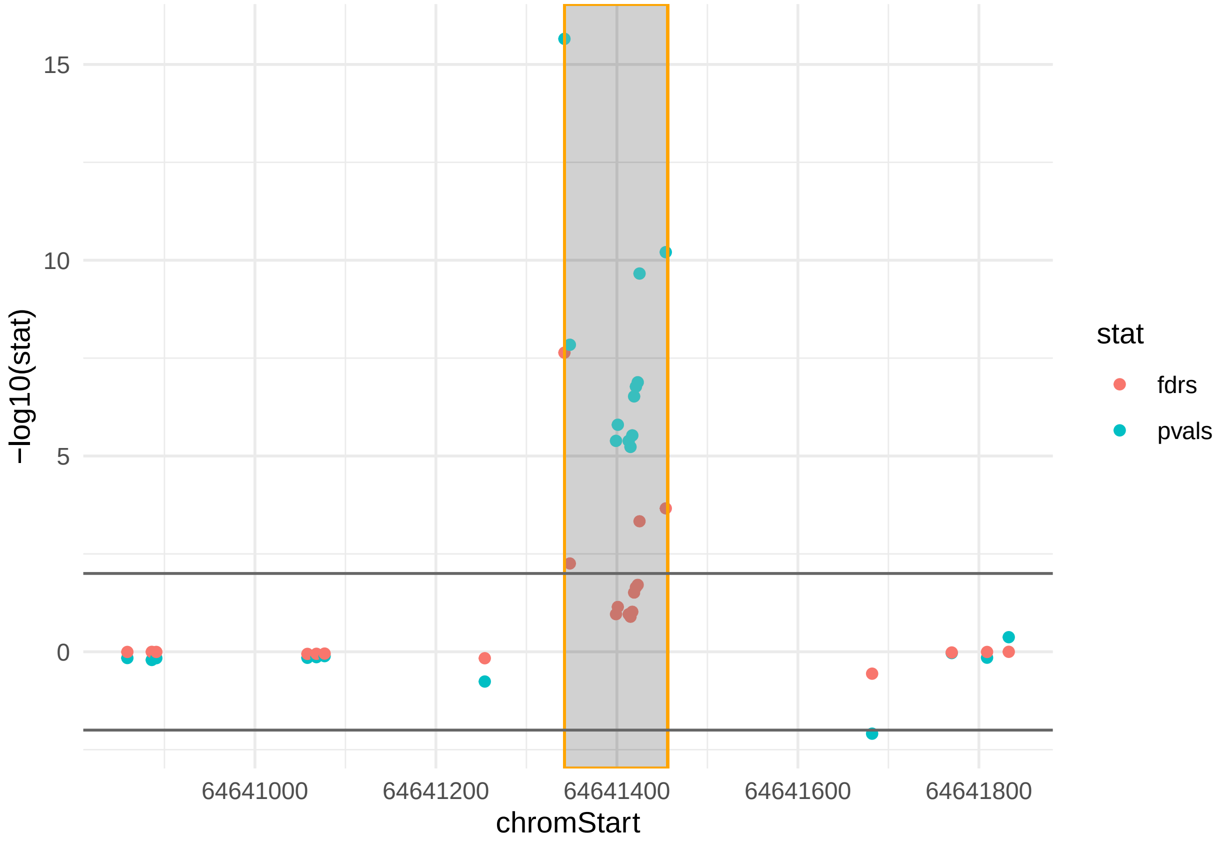

7.5
5.0
2.5
0.0
−log10(stat)
29319000
29319250
29319500
chromStart
stat
fdrs
pvals



0
−20
−40
−60
−log10(stat)
16622400
16622700
16623000
16623300
chromStart
stat
fdrs
pvals


15
10
5
0
−5
−log10(stat)
8418250
8418500
8418750
8419000
chromStart
stat
fdrs
pvals


5
0
−5
−10
−log10(stat)
107944500
107944750
107945000
107945250
107945500
chromStart
stat
fdrs
pvals



Takeaways - DMRs lok reasonable to me, if not a bit conservarvative -

Are the DMRs comparable?

- Take two subsets (recall the 42 LOADs are the same in both groups, the controls are different), and

```
dmrs.subset1 <- fread("../data/07-counts-by-chrom-imputed-subset1/DMRs.adj.bed") %>% makeGRangesFromData
dmrs.subset2 <- fread("../data/07-counts-by-chrom-imputed-subset2/DMRs.adj.bed")%>% makeGRangesFromDataF
```

```
n_dmrs_1 <- length(seqnames(dmrs.subset1))
n_dmrs_1
```

```
## [1] 565
```

```
n_dmrs_2 <- length(seqnames(dmrs.subset2))
n_dmrs_2
```

```
## [1] 670
```

```
length(subsetByOverlaps(dmrs.subset1, dmrs.subset2)) / n_dmrs_1
```

```
## [1] 0.1982301
```

```
length(subsetByOverlaps(dmrs.subset2, dmrs.subset1)) / n_dmrs_2
```

```
## [1] 0.1671642
```

Subset 1 has 565 DMRs and subset 2 has 670.

About 20% of subset 1's DMRs overlap DMRs in subset 2. About 17% of subset 2's DMRs overlap DMRs in subset 1.

What if the DMRs are close, but not overlapping?

```
p <- 0.1
gap <- 10000

while (p < 0.5){
  p <- length(subsetByOverlaps(dmrs.subset1, dmrs.subset2, maxgap = gap)) / n_dmrs_1

  gap <- gap + gap

}

print(gap)
```

```
## [1] 1280000
```

To get to the point where ~50% of DMRs in subset 1 are also in subset 2, we'd need to consider DMRs "the same" even if they are up to 1.3 megabases apart from each other.