Cryptic expression in LCM Neurons

2023-02-09

R Markdown

We wanted to know if we could detect TDP43 cryptic events in LCM neurons from ALS or control patients. AL sent me the counts of spliced reads supporting the inclusion of various TDP-43 related cryptic events from a 2018 study.

```
spliced_reads_orig <- read.csv("spliced_read_from_lcm.csv")
spliced_reads <- spliced_reads_orig
```

Looking at the spliced reads data

```
n_samp <- spliced_reads |>
  distinct(sample_name) |>
  nrow()

n_splice <- spliced_reads |>
  distinct(junction_name) |> nrow()
```

There are 21 unique samples and 770 splice junctions in the data.

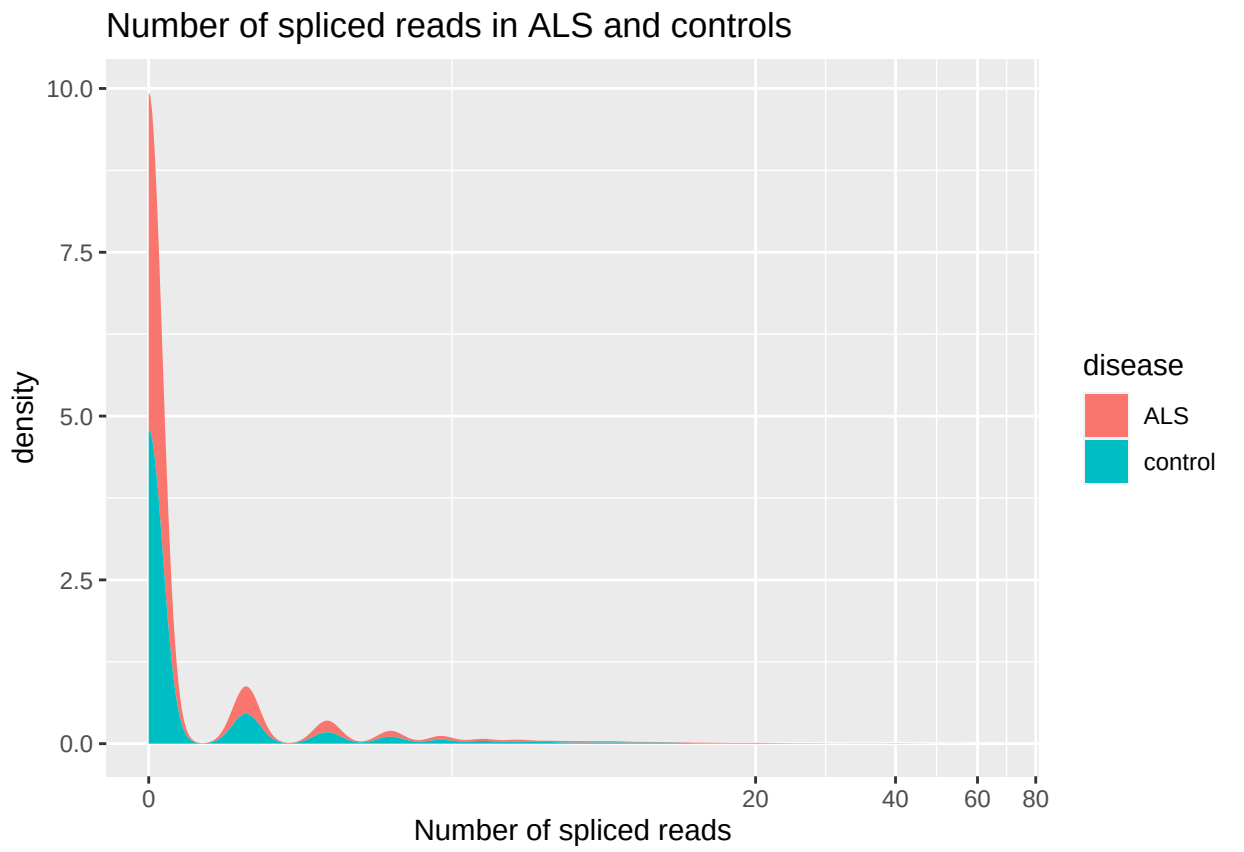
The dataframe was extended to include counts for all samples and all junctions, inserting spliced reads counts as '0'. A new column was added to include information on the disease group for each sample (i.e., ALS patients or control samples).

```
    group_column <- spliced_reads |>
      select(junction_name, sample_name, n_spliced_reads) |>
      complete(junction_name, sample_name,fill=list(n_spliced_reads = 0)) |>
      mutate(disease = ifelse(grepl("ALS", sample_name),
                              "ALS",
                              "control"))
```

To visualise the spliced reads data, a density plot was made with a pseudo log scale on the x-axis to compare the number of spliced reads in the ALS and control samples.

```
group_column |>
  ggplot(aes(x = n_spliced_reads,
             fill = disease)) +
  stat_density() +
  scale_x_continuous(trans = scales::pseudo_log_trans()) +
  labs(
```

```
    title = "Number of spliced reads in ALS and controls",
    x = "Number of spliced reads",
    color = "Group",
  )
```



The density plot illustrated that there are more spliced reads in the ALS cohort, with some more extreme values.

A new column was added to include information on the average count of each junction in ALS patients and control samples. This column was then split into two columns: ALS and control.

```
mean_per_junction <- group_column |>
  group_by(junction_name,disease) |>
  mutate(mean_n_spliced_reads = mean(n_spliced_reads)) |>
  ungroup() |>
  select(junction_name,disease,mean_n_spliced_reads) |>
  unique() |> #mean_n_spliced_reads for each junction_name separately for control and ALS
  pivot_wider(names_from = 'disease',
              values_from = 'mean_n_spliced_reads')
```

Wilcoxon test

The data were nested by junction name and a wilcoxon test was conducted to see if any junctions were differentially expressed in the ALS and control samples. The junctions were filtered for only those that were significant (i.e., wilcoxon test p-value < 0.05). Most of the 770 junctions had no significant difference in expression between the ALS and control samples, but a small fraction of them did - specifically 33.

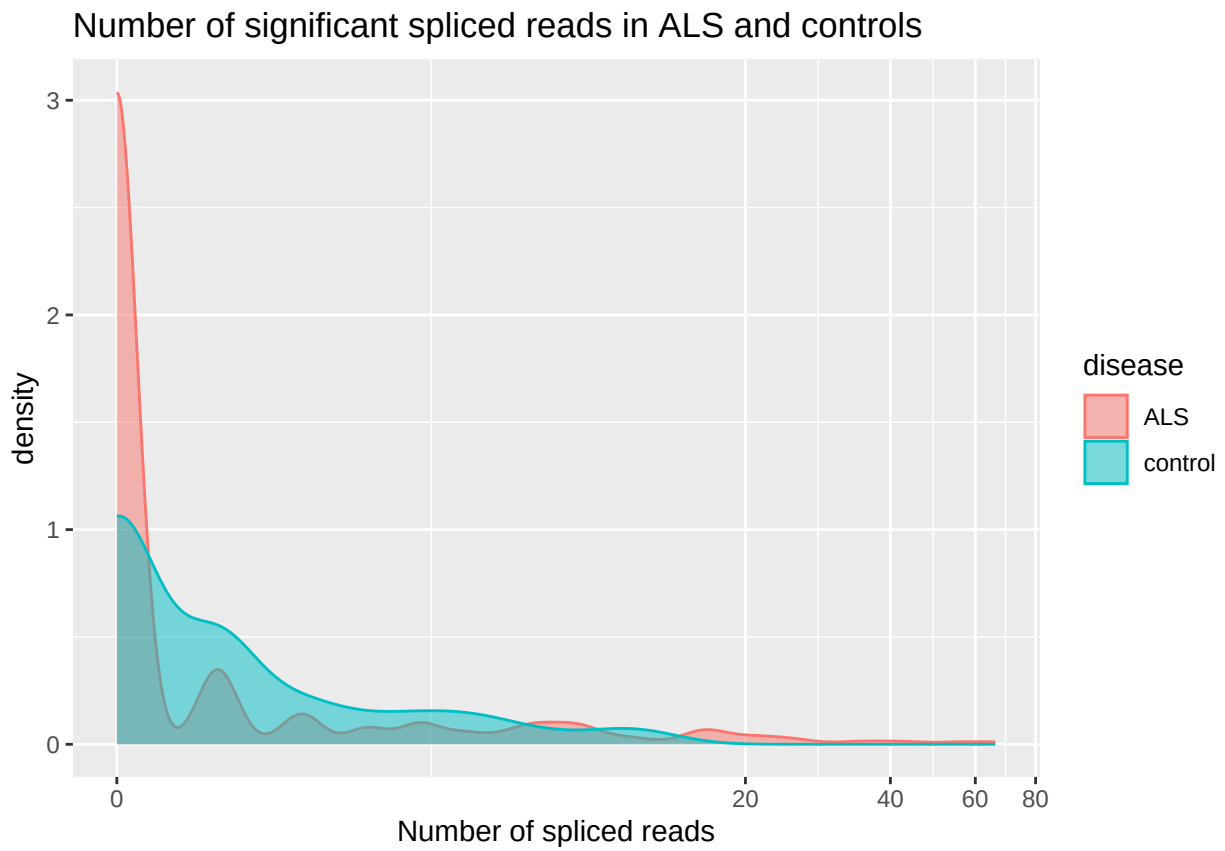
```
group_column_nested <- group_column |>
  group_by(junction_name) |>
  nest()

wilcox_tested_nested = group_column_nested |>
  mutate(wc_res = map(data,~{broom::tidy(wilcox.test(.x$n_spliced_reads ~ .x$disease, exact=FALSE))}))
  unnest(wc_res) |>
  arrange(p.value)

significant_junctions <- wilcox_tested_nested |>
  filter(p.value < 0.05)
```

These significant junctions were grouped according to their disease group ("ALS" or "control") and the percentage of junctions more highly expressed in the ALS and control samples was calculated.

```
mean_per_junction_sig <- mean_per_junction |>
  semi_join(significant_junctions, by=("junction_name")) |>
  mutate(higher_value = ifelse(ALS>control, "ALS", "control")) |>
  janitor::tabyl(higher_value)
```

Most of the significant events were expressed more highly in the controls than the ALS motor neurons. 76% of the significant junctions were more highly expressed in the controls, whereas only 24% were more highly expressed in ALS patients.

```
sig_reads_groups <- group_column |>
  semi_join(significant_junctions, by=("junction_name"))

ggplot(sig_reads_groups, aes(x = n_spliced_reads, color = disease, fill = disease)) +
  geom_density(alpha = 0.5) +
  scale_x_continuous(trans = scales::pseudo_log_trans()) +
  labs(
    title = "Number of significant spliced reads in ALS and controls",
    x = "Number of spliced reads",
  )
```



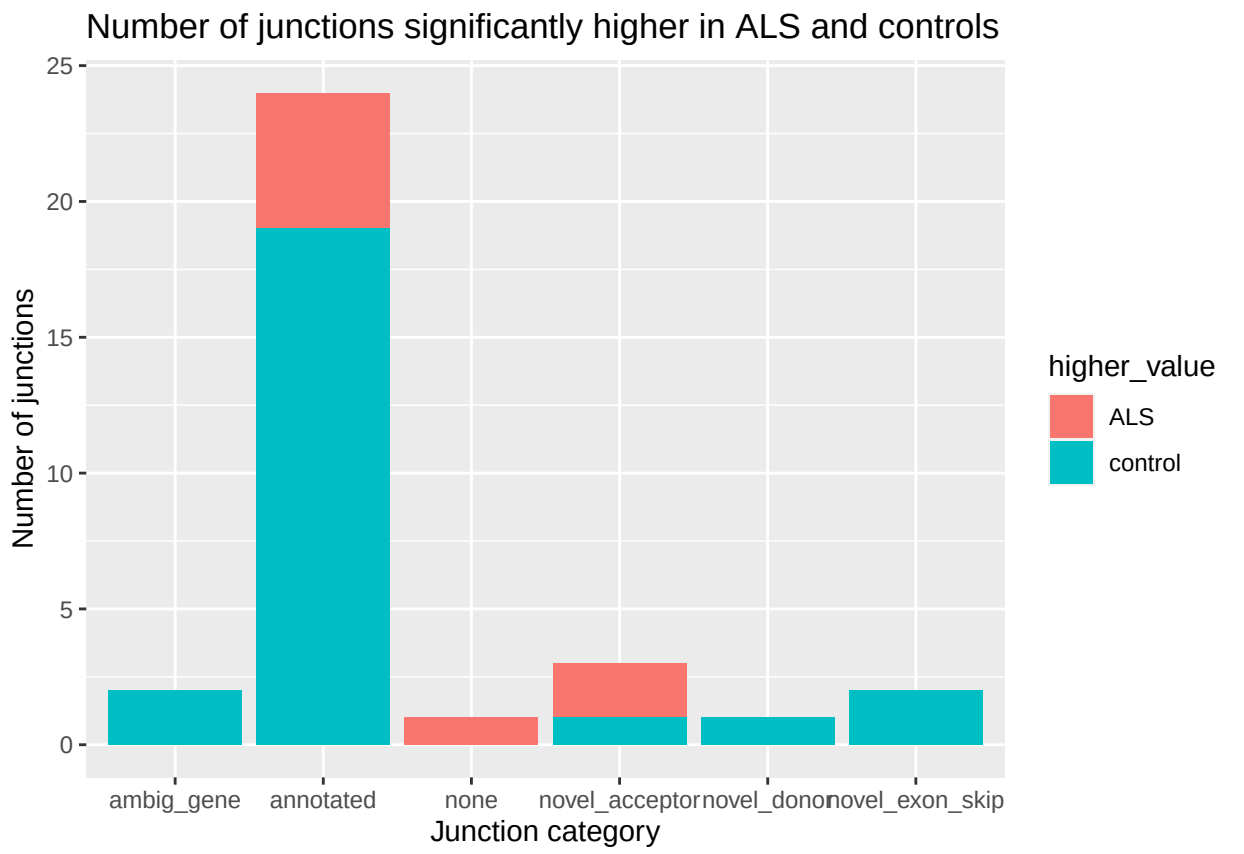

```
mean_per_junction |>
  semi_join(significant_junctions, by=("junction_name")) |>
  mutate(higher_value = ifelse(ALS>control,"ALS","control")) |>
  separate(junction_name, sep = '\\|',into = c("gene","junc_cat","n_datasets_junction_found"),convert =
  janitor::tabyl(junc_cat,higher_value)
```

```
##         junc_cat ALS control
##       ambig_gene   0       2
##        annotated   5      19
##             none   1       0
##   novel_acceptor   2       1
##      novel_donor   0       1
##  novel_exon_skip   0       2
```

Most of the junctions significantly higher in controls are annotated events, with both ends of the junctions overlapping with known introns/exons, whereas for the junctions higher in the ALS cohort either none or only 1 end overlaps a known exon.

```
mean_per_junction |>
  semi_join(significant_junctions, by=("junction_name")) |>
  mutate(higher_value = ifelse(ALS>control,"ALS","control")) |>
  separate(junction_name, sep = '\\|',into = c("gene","junc_cat","n_datasets_junction_found"),convert =
  ggplot(aes(x = junc_cat, fill = higher_value)) +
  geom_bar() +
  labs(
```

```
    title = "Number of junctions significantly higher in ALS and controls",
    x = "Junction category",
    y = "Number of junctions",
    color = "Group",
  )
```



Next, we asked: how many datasets was an event detected in if it was higher in ALS or controls?

```
# Mean 'n_datasets_junction_found' by which tissue it was higher in -------

mean_per_junction |>
  semi_join(significant_junctions, by=("junction_name")) |>
  mutate(higher_value = ifelse(ALS>control,"ALS","control")) |> group_by(higher_value) |>
  separate(junction_name, sep = '\\|',into = c("gene","junc_cat","n_datasets_junction_found"),convert =
  summarize(mean_n_datasets = mean(n_datasets_junction_found))
```

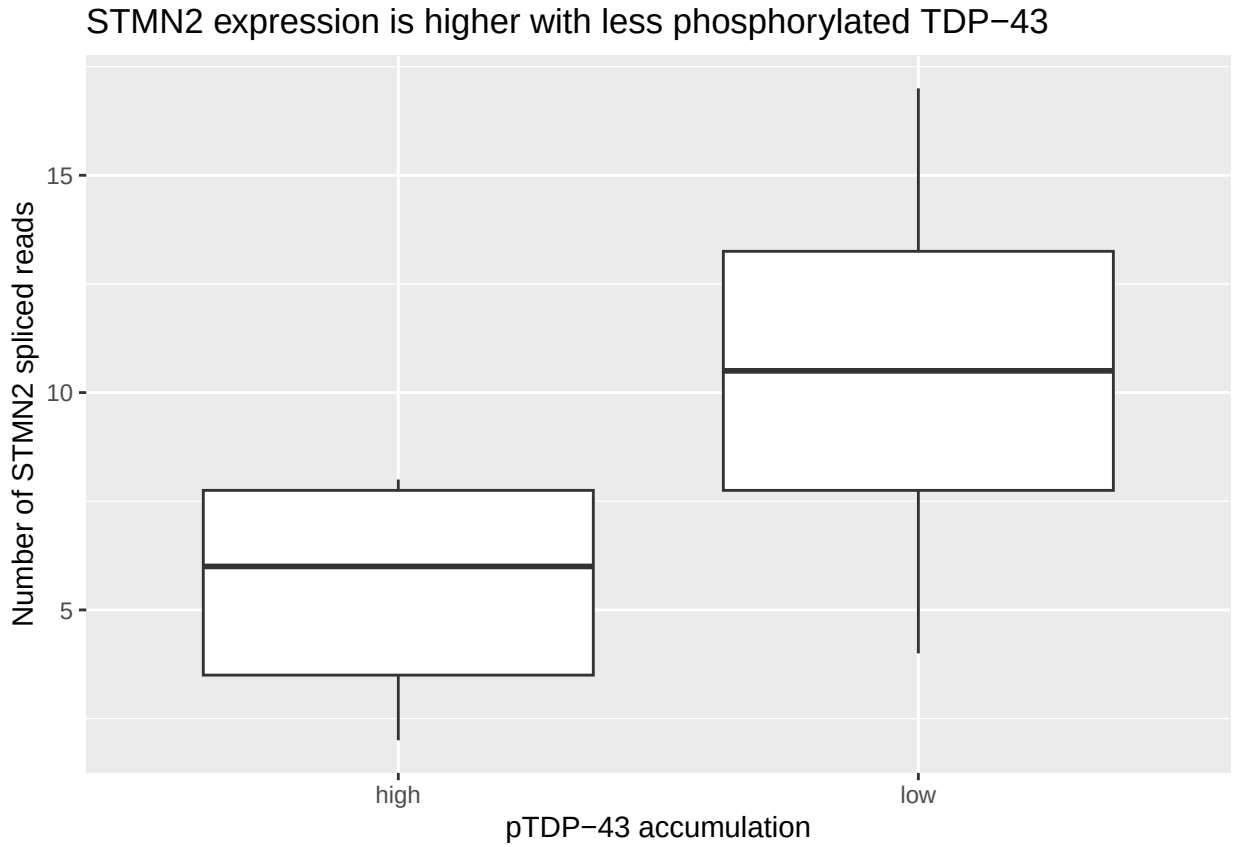
```
## # A tibble: 2 x 2
##   higher_value mean_n_datasets
##   <chr>                  <dbl>
## 1 ALS                      3.5
## 2 control                  1
```

Junctions higher in ALS motor neurons were found in, on average, 3.5 TDP-43 knockdown studies. On the other hand, junctions higher in controls were found in only 1 study. Thus, the annotated events higher in controls were unique to a single TDP-43 knockdown.

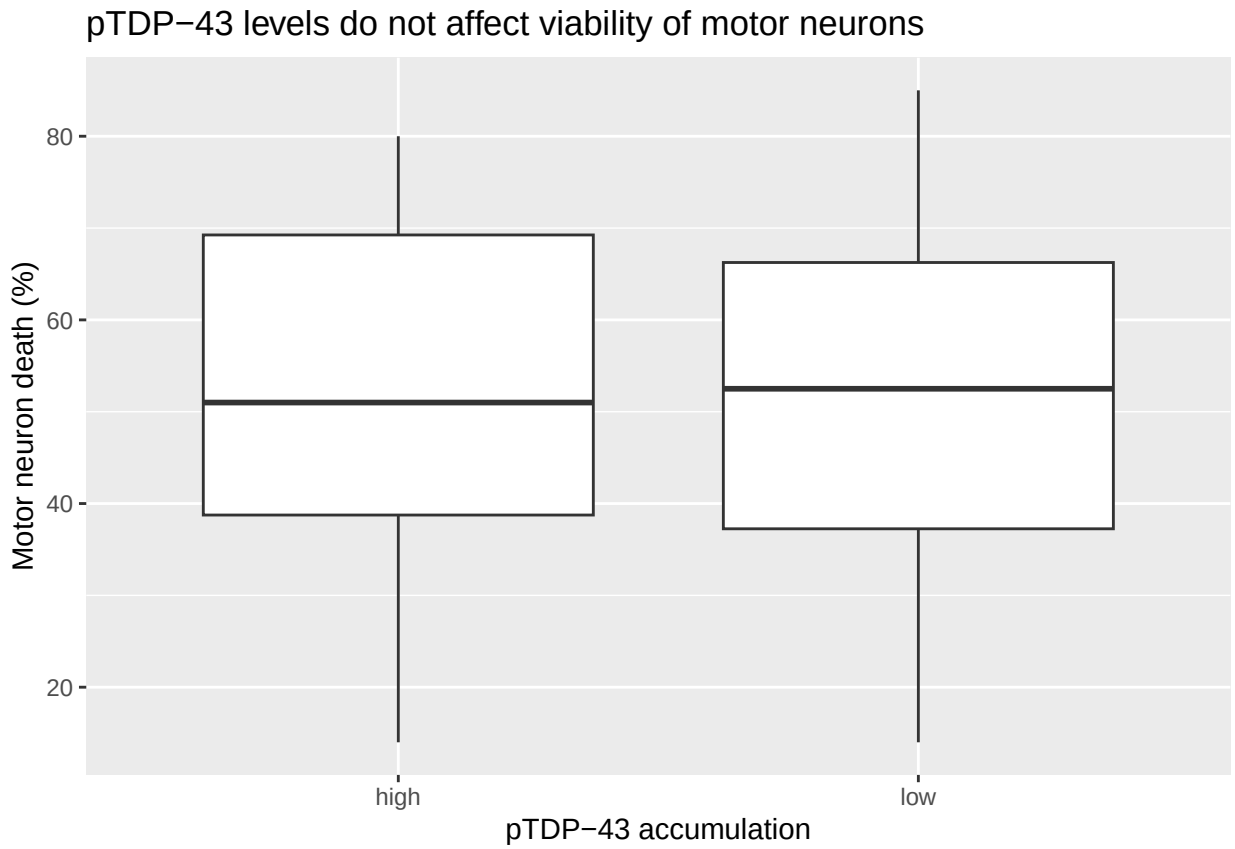
Next, AL sent me

```
# Left-join ptdp table to existing table ----------------------------------

ptdp_orig <- read.csv("ptdp_mn_death.csv")

spliced_reads$sample_name <- gsub(".SJ.out", "", as.character(spliced_reads$sample_name))

ptdp_orig <- ptdp_orig |>
  rename("sample_name" = "sample")

spliced_reads_pdtp <- spliced_reads |>
  left_join(ptdp_orig, by = "sample_name")
```

```
# Plot STMN2 expression against MN ----------------------------------------

spliced_reads_pdtp_select <- spliced_reads_pdtp |>
  select(sample_name, n_spliced_reads, junction_name, pTDP.43, MN_death, pTDP_category)

spliced_reads_pdtp_select |>
  filter(grepl("STMN2", junction_name)) |>
  drop_na() |>
  ggplot(aes(x = pTDP_category, y = n_spliced_reads)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(
    title = "STMN2 expression is higher with less phosphorylated TDP-43",
    x = "pTDP-43 accumulation",
    y = "Number of STMN2 spliced reads")
```




```
spliced_reads_pdtp_select |>
  filter(grepl("STMN2", junction_name)) |>
  drop_na() |>
  ggplot(aes(x = pTDP_category, y = MN_death)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(
    title = "pTDP-43 levels do not affect viability of motor neurons",
    x = "pTDP-43 accumulation",
    y = "Motor neuron death (%)")
```




```
# stmn2_all_junctions.bed ------------------------------------------------

stmn2_all_junctions_orig <- read_tsv("stmn2_all_junctions.bed",
                                     col_names = c("chrosome", "start", "end", "sample_name",
                                                   "n_spliced_reads", "strand"))
```

```
## Rows: 160 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (3): chrosome, sample_name, strand
## dbl (3): start, end, n_spliced_reads
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
stmn2_all_junctions <- unique(stmn2_all_junctions_orig)
stmn2_all_junctions$sample_name <- gsub(".SJ.out", "", as.character(stmn2_all_junctions$sample_name))

spliced_reads_pdtp_stmn2 <- stmn2_all_junctions |>
  left_join(ptdp_orig, by=c("sample_name"))

spliced_reads_pdtp_select |>
  filter(grepl("STMN2", junction_name))
```

```
##    sample_name n_spliced_reads          junction_name pTDP.43 MN_death
```

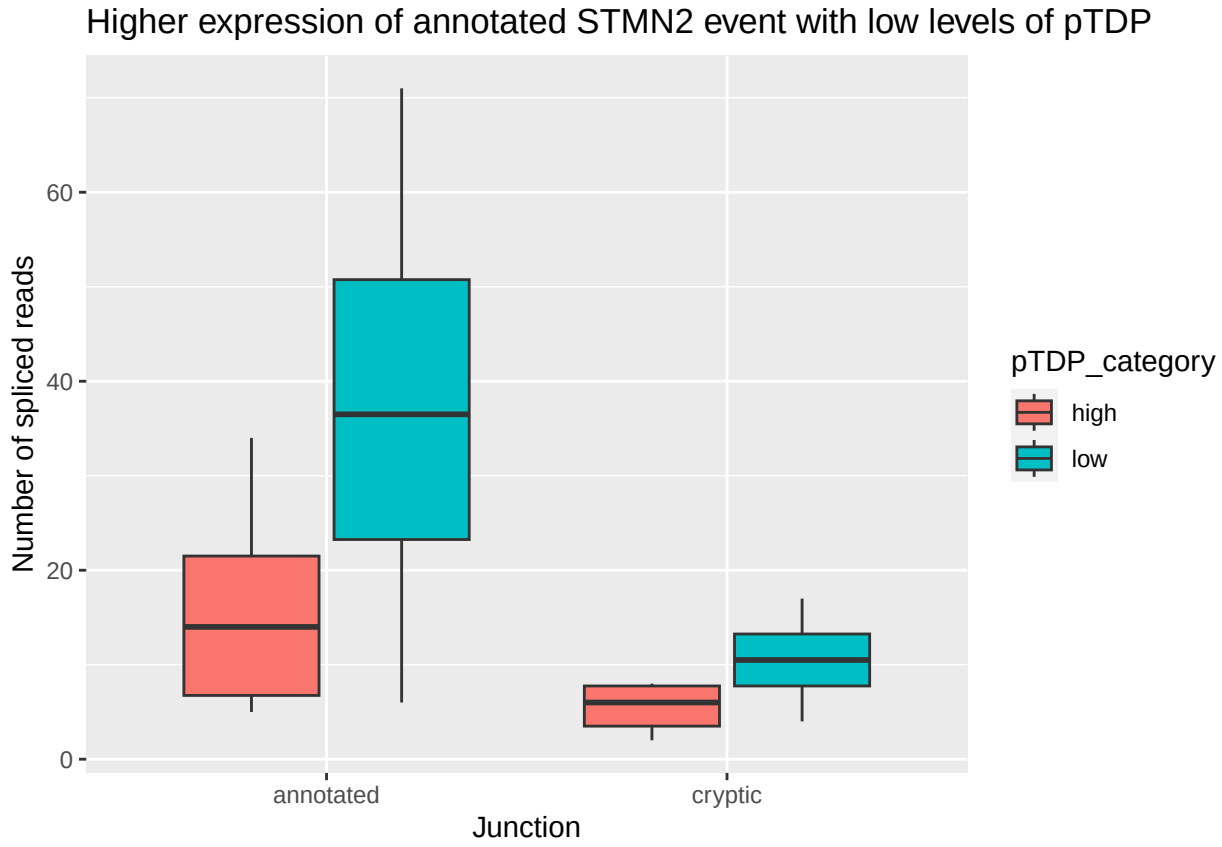
```
## 1      21_sALS              17 STMN2|novel_acceptor|9      19       45
## 2      27_sALS               3 STMN2|novel_acceptor|9      40       41
## 3      34_sALS               8 STMN2|novel_acceptor|9      43       61
## 4      48_sALS               2 STMN2|novel_acceptor|9      58       72
## 5      60_sALS               7 STMN2|novel_acceptor|9      53       38
## 6      62_sALS               5 STMN2|novel_acceptor|9      63       80
## 7      63_sALS               9 STMN2|novel_acceptor|9      37       85
## 8      79_sALS               4 STMN2|novel_acceptor|9      31       60
## 9      82_sALS               8 STMN2|novel_acceptor|9      56       14
## 10     84_sALS              12 STMN2|novel_acceptor|9      35       14
## 11     85_sALS               2 STMN2|novel_acceptor|9      NA       NA
##    pTDP_category
## 1            low
## 2           high
## 3           high
## 4           high
## 5           high
## 6           high
## 7            low
## 8            low
## 9           high
## 10           low
## 11          <NA>
```

```
stmn2_all_junctions |>
  distinct(end)
```

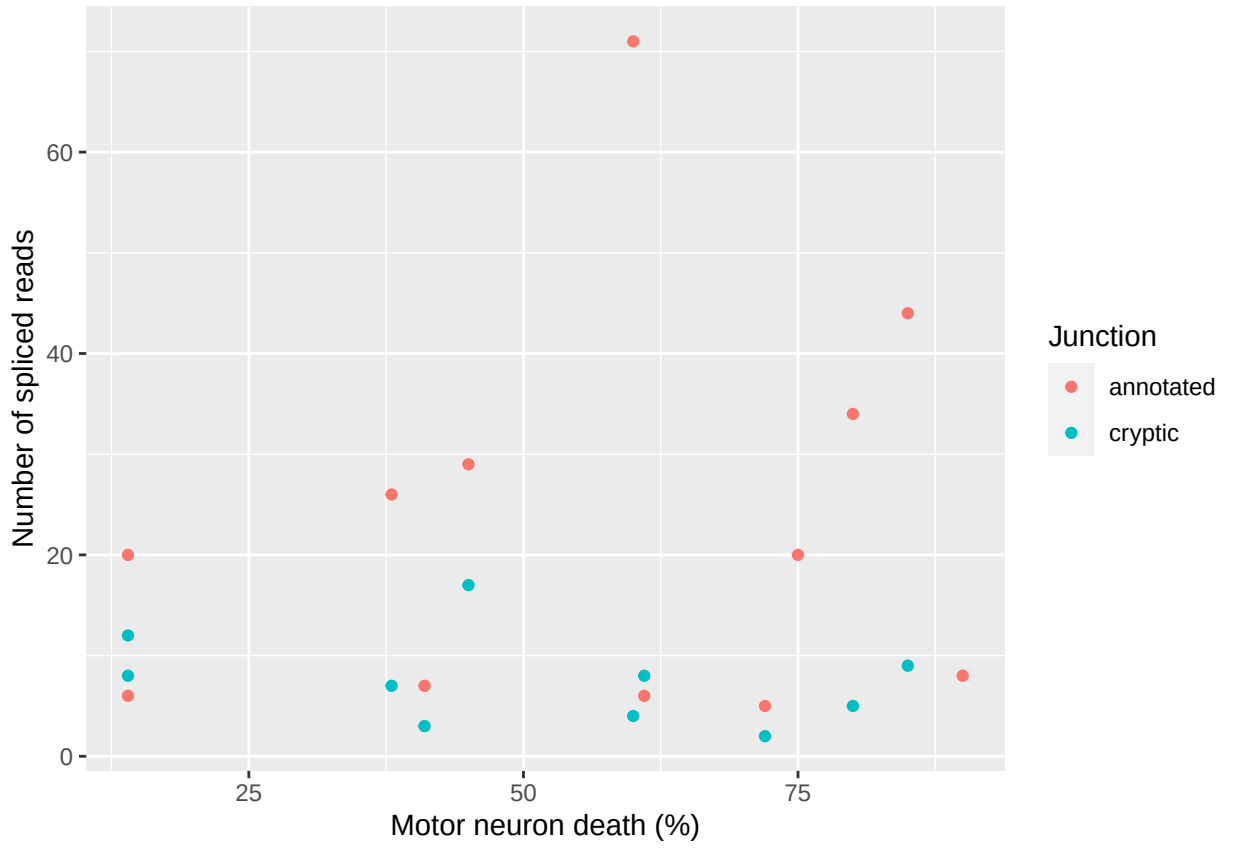
```
## # A tibble: 2 x 1
##        end
##      <dbl>
## 1 79636802
## 2 79616822
```

```
  # cryptic 79616822, annotated 79636802

spliced_reads_pdtp_stmn2 <- spliced_reads_pdtp_stmn2 |>
  mutate(junction = ifelse(grepl("79616822", end),
                           "cryptic",
                           "annotated")) |>
  mutate(disease = ifelse(grepl("ALS", sample_name),
                                "ALS",
                                "control")) |>
  relocate(disease, .after = sample_name) |>
  relocate(junction, .after = disease)

spliced_reads_pdtp_stmn2 |>
  ggplot(aes(x = junction, y = n_spliced_reads, fill = disease)) +
  geom_boxplot() +
  labs(
    title = "Annotated STMN2 event has higher expression in the control samples",
    y = "Number of spliced reads"
  )
```

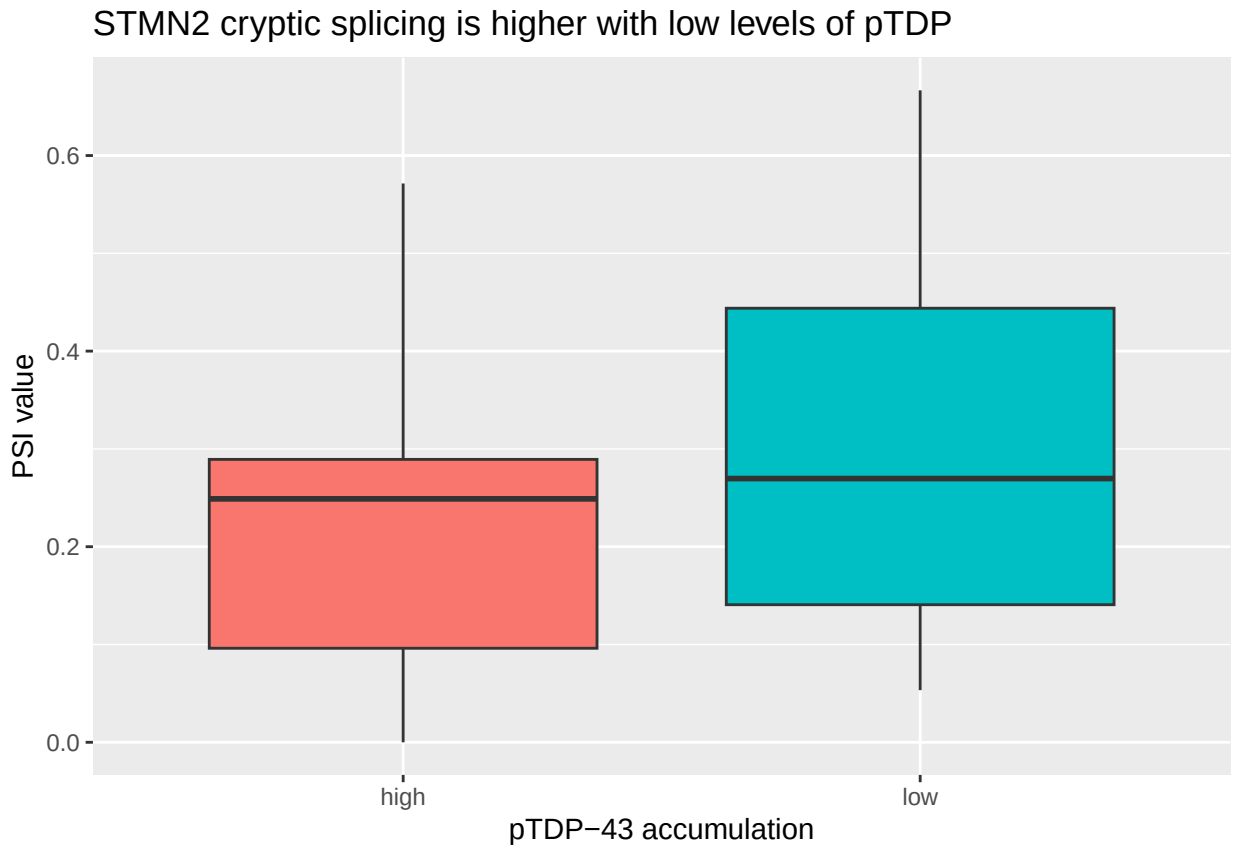

```
spliced_reads_pdtp_stmn2 |>
  drop_na() |>
  ggplot(aes(x = junction, y = n_spliced_reads, fill = pTDP_category)) +
  geom_boxplot() +
  labs(
    title = "Higher expression of annotated STMN2 event with low levels of pTDP",
    x = "Junction",
    y = "Number of spliced reads"
  )
```



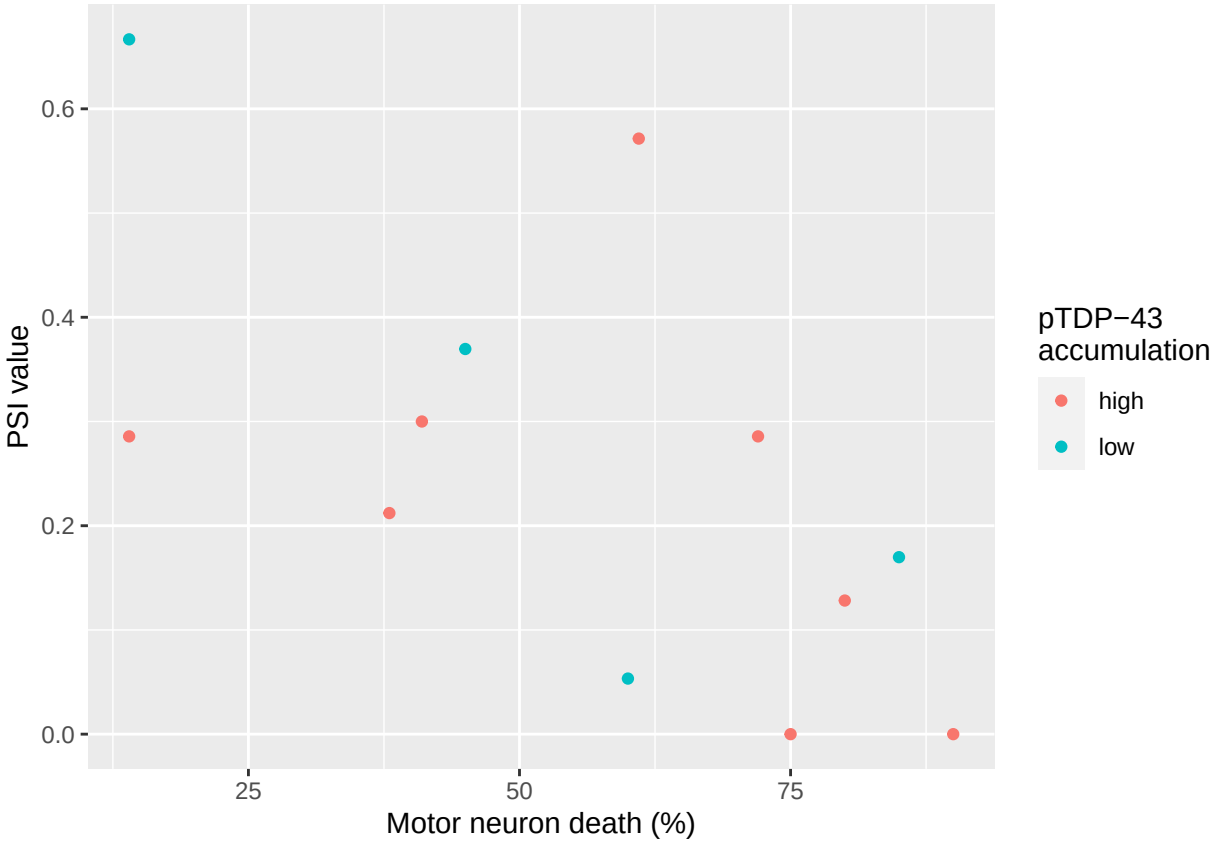

```
spliced_reads_pdtp_stmn2 |>
  drop_na() |>
  ggplot(aes(x = MN_death, y = n_spliced_reads, color = junction)) +
  geom_point() +
  labs(
    x = "Motor neuron death (%)",
    y = "Number of spliced reads",
    color = "Junction"
  )
```




```
# percent spliced in (psi) value ----------------------------------------

psi_stmn2 <- spliced_reads_pdtp_stmn2 |>
  select(-start,-end) |>
  unique() |>
  pivot_wider(names_from = 'junction',
              values_from = 'n_spliced_reads',
              values_fill = 0)

psi_stmn2 <- psi_stmn2 |>
  group_by(sample_name) |>
  mutate(total_counts = annotated + cryptic) |>
  mutate(psi = cryptic / total_counts)

psi_stmn2 |>
  drop_na() |>
  ggplot(aes(x = pTDP_category, y = psi, fill = pTDP_category)) +
  geom_boxplot() +
  theme(legend.position = "none") +
  labs(
    title = "STMN2 cryptic splicing is higher with low levels of pTDP",
    x = "pTDP-43 accumulation",
    y = "PSI value"
  )
```




```
psi_stmn2 |>
  drop_na() |>
  ggplot(aes(x = MN_death, y = psi, color = pTDP_category)) +
  geom_point() +
  labs(
    x = "Motor neuron death (%)",
    y = "PSI value",
    color = "pTDP-43 \naccumulation"
  )
```


0.6
0.4
0.2
0.0
PSI value
25
50
75
Motor neuron death (%)
pTDP−43
accumulation
high
low