```
In [53]: suppressMessages(library(data.table))
          suppressMessages(library(matrixStats))
          suppressMessages(library(ggplot2))
          suppressMessages(library(RColorBrewer))
          suppressMessages(library(ggsci))
          suppressMessages(library(magrittr))
          suppressMessages(library(ggrepel))
          suppressMessages(library(dplyr))
          suppressMessages(library(ggalluvial))
          suppressMessages(library(ggpubr))
          suppressMessages(library(valr))
          suppressMessages(library(ggh4x))
          suppressMessages(library(seqminer))
          options(warn=-1) # use options(warn=0) To turn warnings back on
          options(dplyr.summarise.inform = FALSE)
          rescale <- function(x, CDS start, CDS end) ( (x-CDS start)/(CDS end - CDS start) * 100)
          confidence interval upper <- function(vector, interval=0.95) {</pre>
           # Standard deviation of sample
           vec sd <- sd(vector, na.rm = TRUE)</pre>
           # Sample size
           n <- length (vector)
           # Median of sample
           vec mean <- mean(vector, na.rm = TRUE)</pre>
           # Error according to t distribution
           error \leftarrow qt((interval + 1)/2, df = n - 1) * vec sd / sqrt(n)
           # Confidence interval as a vector
           result <- vec mean + error
           return (result)
         confidence interval lower <- function(vector, interval=0.95) {</pre>
           # Standard deviation of sample
           vec sd <- sd(vector, na.rm = TRUE)</pre>
           # Sample size
           n <- length (vector)
           # Median of sample
           vec_mean <- mean(vector, na.rm = TRUE)</pre>
           # Error according to t distribution
           error \leftarrow qt((interval + 1)/2, df = n - 1) * vec sd / sqrt(n)
           # Confidence interval as a vector
           result <- vec mean - error
           return(result)
          source("../src/plot iso structure.r")
In [13]: setwd("/Users/inamojun/TMDU-LR isoform atlas/Figures")
In [14]: | #load data
          load("../data/data Fig02.RData")
          load("../data/data Fig02extra.RData")
In [15]:
          # head(phastcons novel)
          # head(phastcons lnc)
In [16]:
In [62]: # extract predicted-coding genes only in Isoform Atlas (3448 genes, Figure 1C) and count
           dplyr::filter(associated gene %in% setdiff(out[out$coding,]$associated gene,
                                                         gencode[gencode$gene type=="protein coding"
```

dplyr::select(associated gene, gene biotype) %>%

dplyr::distinct(associated gene, .keep all = TRUE) %>%

```
dplyr::mutate(type = dplyr::case when(
    grepl("^novelGene ",associated gene) ~ "novel gene",
    grepl(" ",associated gene) & !grepl("^novelGene ",associated gene) ~ "read-through",
    grepl("pseudogene", gene biotype) ~ "pseudogene",
    grepl("lncRNA", gene biotype) | grepl("lincRNA", gene biotype) ~ "lncRNA",
   TRUE ~ "others"
 )) %>%
 dplyr::group by(type) %>%
 dplyr::summarise(count = n()) %>%
 dplyr::mutate(proportion = 100*count/sum(count),
                type = factor(type, levels = c("read-through", "novel gene", "pseudogene",
options(repr.plot.width=5, repr.plot.height=2)
q a = out %>%
 dplyr::filter(associated gene %in% setdiff(out[out$coding,]$associated gene,
                                             gencode[gencode$gene type=="protein coding"
  dplyr::select(associated gene, gene biotype) %>%
 dplyr::distinct(associated gene, .keep all = TRUE) %>%
 dplyr::mutate(type = dplyr::case when(
   grepl("^novelGene ",associated gene) ~ "novel gene",
   grepl(" ",associated gene) & !grepl("^novelGene ",associated gene) ~ "read-through",
    grepl("pseudogene", gene biotype) ~ "pseudogene",
    grepl("lncRNA",gene biotype) | grepl("lincRNA",gene biotype) ~ "lncRNA",
   TRUE ~ "others"
 )) 응>용
 dplyr::group by(type) %>%
 dplyr::summarise(count = n()) %>%
 dplyr::mutate(proportion = 100*count/sum(count),
                type = factor(type, levels = c("read-through", "novel gene", "pseudogene",
 ggplot(data=., aes(x=reorder(type,count), y=count), fill="black") +
 geom bar(stat="identity", color="black", position=position dodge())+
 theme classic() +
  # ggtitle ("the coding genes identified only in the Isoform Atlas ") +
 ylab("number of genes") +
 xlab("") +
 coord flip() +
 theme(strip.text.x=element text(size=20, color="black", face="bold"),
        strip.text.y=element text(size=20, color="black", face="bold"),
        legend.position = "bottom",
        plot.title = element text(size=20),
        axis.title.x = element text(size=20),
        axis.title.y = element text(size =20),
        axis.text.y = element text(size = 20),
        axis.text.x = element text(size = 20),
        legend.text = element text(size = 15),
        legend.key.size = grid::unit(1, "lines"),
        legend.title = element text(size = 0, hjust = 0))
g_a
```

A tibble: 5×3

type count proportion

<fct></fct>	<int></int>	<dbl></dbl>
IncRNA	1022	29.640371
novel gene	529	15.342227
others	80	2.320186
pseudogene	452	13.109049
read-through	1365	39.588167

```
In [45]: gene="SIGLEC9|SIGLEC7|SIGLEC7 SIGLEC9"
               options (repr.plot.width=12, repr.plot.height=7)
               g b = plot iso structure(gene = gene, extra = 5e03, repeat draw = FALSE, atac draw = FAL
               g b
               [1] "plotting..."
                isoform structure [SIGLEC9|SIGLEC7|SIGLEC7_SIGLEC9]
                    481-8-11-
                                                                                 ->04da7c25-5281-4608-b13a-960b166630c3
                                                                                  ->0f6383a7-5633-4e7b-8646-659d47473c80
                    4
                                                                                 ->23c1be96-132a-40c8-b535-543b68397aa1
                                                                                  ->30067b74-931f-4634-bfcc-9089335ef29b
                                                                                 ->374d0b6a-e6b1-4473-803c-104cd4d5cd81
                                                                                  ->3b8c7667-7ad2-4e80-884e-abacb18e24ff
                                                                                  ->3ee74883-8975-40fe-a56b-355e3e80b00c
                                                                                  ->42hedf58-5d55-42hf-9e42-0e234fac3357
                                                                                  ->42c3c462-0cd0-4a0c-8fb1-3ce13ce7e3b9
                                                                                  ->6148696c-a9c6-43b5-ae18-2ed7df140306
                                                                                 ->7894e79b-8e20-4b14-b20e-66e0fdab3c01-1
                                                                                  ->817fd2a6-1535-4161-9c8b-2ccc7f69f7e5
                                                                                                                                    Isoform Atlas
                                                                                  ->992c4290-ed8a-418c-bb6f-3f23a01eb522
                    ->a36d2848-0482-42a1-9db1-04159af9abfe
                                                                                  ->a93ef0c9-7c8a-4bf0-ab77-f369c548b8ef
                                                                                  ->afda3755-dc2a-4acc-adcb-f38f5ce55e4e
                                                                                 ->cb422ab5-ed97-40e4-bc84-c7b3095eb019
                                                                                  ->dd7f9522-dde0-4005-aa57-904b2fdc8a98
                                                                                  ->ec3a505b-dcfd-4ab0-8794-9ad88a89e5a4
                                                                                         ->ENST00000250360.8
                                                                                         ->FNST00000305628.7
                                                                                         ->ENST00000317643.10
                                                                                         ->FNST00000440804.7
                                                                                         ->ENST00000536156.5
                                                                                              SIGLEC7
                                                                                                                                     GENCODE
                   SIGLEC9
                          51130000
                                             51140000
                                                                 51150000
                                       genomic position [chr19]
                                      ANCHOR ig signal peptide IDR Ig_2 V-set
```

```
In [19]: # Conservation score of novel genes in predicted-coding genes only in Isoform Atlas (529
         options(repr.plot.width=5, repr.plot.height=4)
         g c = phastcons novel %>%
           dplyr::mutate(scaled distance bin = round(scaled distance) ,
                          region = dplyr::case when (
                              region == "5UTR" ~ "5'UTR",
                              region == "CDS" ~ "ORF",
                              region == "3UTR" ~ "3'UTR",
                              region == "Intron" ~ "Intron"
                          ),
                          region = factor(region, levels = c("5'UTR", "ORF", "3'UTR", "Intron"))) %>%
           dplyr::group by(scaled distance bin, region) %>%
           dplyr::summarize(PHASTCONS mean = mean(PHASTCONS, na.rm=TRUE),
                             PHASTCONS CI up = confidence interval upper (PHASTCONS),
                             PHASTCONS CI low = confidence interval lower (PHASTCONS)
           ) 응>응
           as.data.frame() %>%
           na.omit() %>%
           ggplot(., aes(x=scaled distance bin,y=PHASTCONS mean)) +
           geom line(aes(x=scaled distance bin,y=PHASTCONS mean,color=region)) +
           geom ribbon (aes (x=scaled distance bin, y=PHASTCONS mean, ymax = PHASTCONS CI up, ymin
                        alpha=0.3) +
           ggtitle(paste0("novel predicted coding gene")) +
```

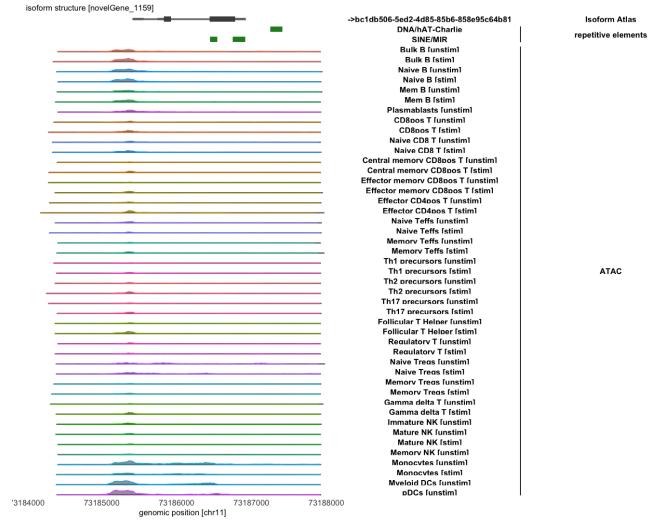
```
ylab(paste0("phastCons100way score")) +
 xlab("relative to splice junction") +
 scale x continuous (labels=c("5","3"), breaks=c(0,100), limits=c(0,100)) +
 geom hline(yintercept = 0,linetype="dashed",colour="black") +
 ylim(0,1) +
 theme minimal() +
 theme(strip.text.x=element text(size=20, color="black", face="bold"),
        strip.text.y=element text(size=20, color="black", face="bold"),
       panel.grid=element blank(),
       legend.position = "top",
       plot.title = element text(size=20),
       axis.title.x = element text(size=20),
       axis.title.y = element text(size =20),
       axis.text.y = element text(size = 20),
       axis.text.x = element text(size = 20),
       legend.text = element_text(size = 20),
       legend.key.size = grid::unit(0.8, "lines"),
       legend.title = element text(size = 0, hjust = 0))
# Conservation score of lncRNA in predicted-coding genes only in Isoform Atlas (529 genes
q d = phastcons lnc %>%
 dplyr::mutate(scaled distance bin = round(scaled distance) ,
                region = dplyr::case when(
                    region == "5UTR" ~ "5'UTR",
                    region == "CDS" ~ "ORF",
                    region == "3UTR" ~ "3'UTR",
                    region == "Intron" ~ "Intron"
                region = factor(region,levels = c("5'UTR","ORF","3'UTR","Intron"))) %>%
 dplyr::group by (scaled distance bin, region) %>%
 dplyr::summarize(PHASTCONS mean = mean(PHASTCONS, na.rm=TRUE),
                   PHASTCONS CI up = confidence interval upper (PHASTCONS),
                   PHASTCONS CI low = confidence interval lower (PHASTCONS)
 ) 응>용
 as.data.frame() %>%
 na.omit() %>%
 ggplot(., aes(x=scaled distance bin,y=PHASTCONS mean)) +
 geom line(aes(x=scaled distance bin,y=PHASTCONS mean,color=region)) +
 geom ribbon (aes (x=scaled distance bin, y=PHASTCONS mean, ymax = PHASTCONS CI up, ymin
             alpha=0.3) +
 ggtitle(paste0("lncRNA predicted coding gene")) +
 ylab(paste0("phastCons100way score")) +
 xlab("relative to splice junction") +
 scale x continuous (labels=c("5","3"), breaks=c(0,100), limits=c(0,100)) +
 geom hline(yintercept = 0,linetype="dashed",colour="black") +
 ylim(0,1) +
 theme minimal() +
 theme(strip.text.x=element text(size=20, color="black", face="bold"),
       strip.text.y=element text(size=20, color="black", face="bold"),
       panel.grid=element blank(),
       legend.position = "top",
       plot.title = element text(size=20),
       axis.title.x = element text(size=20),
       axis.title.y = element text(size =20),
       axis.text.y = element text(size = 20),
       axis.text.x = element text(size = 20),
       legend.text = element text(size = 20),
       legend.key.size = grid::unit(0.8, "lines"),
       legend.title = element text(size = 0, hjust = 0))
g c
g d
```

```
5'UTRORF 3'UTR Intron
          bhastCons100way score
0.00 0.50
0.00 0.00
                   5'
                                                    3'
                       relative to splice junction
                   IncRNA predicted coding gene
                     5'UTRORF3'UTR Intron
          phastCons100way score
             1.00
             0.75
             0.50
             0.25
             0.00
                                                    3'
                   5'
                       relative to splice junction
           gene="novelGene_1159"
In [20]:
           options(repr.plot.width=12, repr.plot.height=9)
           g e = plot iso structure(gene = gene, extra = 1e03, repeat draw = TRUE, atac draw = TRUE
```

novel predicted coding gene

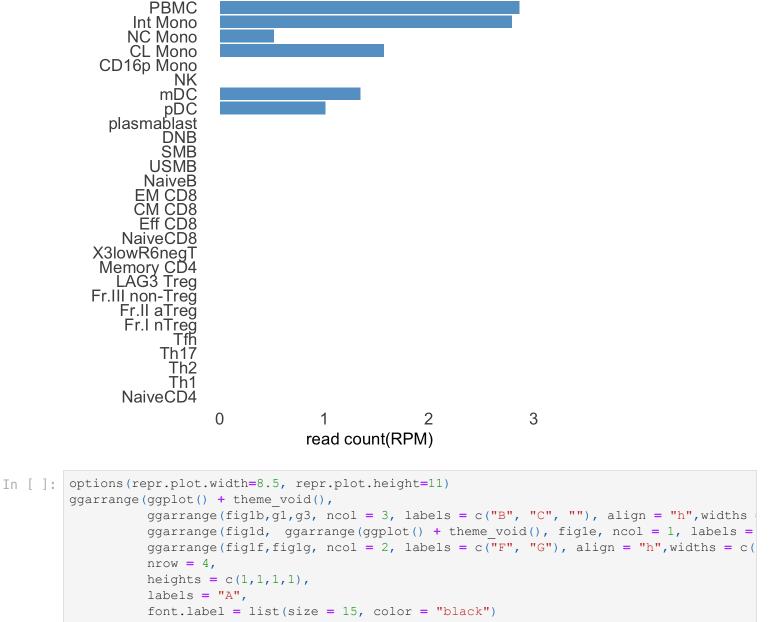
[1] "extracting ATAC-seq peak..."

[1] "plotting..."



```
In [55]: gene="novelGene_1159"
    options(repr.plot.width=8, repr.plot.height=8)

g_f = plot_iso_usage(gene = gene, method = "sum", legend = "none")
# plot_iso_usage(gene = gene, method = "ratio")
g_f
```



Neu

In []: pdf("./Figure2a.pdf", width=8.5, height=11)

g a

dev.off()