Analysis of ZNF382 with LINE-1

Zheng Zuo

```
require(dplyr)
require(ggplot2)
require(GenomicRanges)
require(TFCookbook)
```

```
load("L1 Repeat-ORF correspondence.RData")
L1P1_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1P1_orf2" &
                                 Repeat %in% c("L1HS","L1PA2", "L1PA3", "L1PA4", "L1PA5", "L1PA6"))

L1P3_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1P3_orf2" &
                                 Repeat %in% c("L1PA7", "L1PA8", "L1PA10", "L1PA11", "L1PA12"))

L1P4_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1P4_orf2" &
                                 Repeat %in% c("L1PA13", "L1PA14","L1PA15", "L1PA16", "L1PA17", "L1PA15

L1PB_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1PB_orf2" &
                                 startsWith(Repeat, "L1PB"))

L1M1_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1M1_orf2" &
                                 Repeat %in% c("L1MA1", "L1MA2", "L1MA3"))

L1M2_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1M2_orf2" &
                                 Repeat %in% c("L1MA4", "L1MA4A", "L1MA5", "L1MA6"))

L1M3_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1M3_orf2" &
                                 Repeat %in% c("L1MA7", "L1MA8", "L1MA9", "L1MA10"))

L1M4_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1M4_orf2" &
                                 startsWith(Repeat, "L1MB"))

L1MC_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1MC_orf2" &
                                 startsWith(Repeat, "L1MC"))

L1MD_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1MD_orf2" &
                                 startsWith(Repeat, "L1MD"))

L1M5_ORF.repeats = Repeat_ORF.correspondence.table %>%
                   dplyr::filter(RepeatORF=="L1M5_orf2" &
                                 startsWith(Repeat, "L1ME"))


ZNF382.L1P1.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1P1_orf2", RepeatID = L1P1_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1P3.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1P3_orf2", RepeatID = L1P3_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1P4.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1P4_orf2", RepeatID = L1P4_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1PB.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1PB_orf2", RepeatID = L1PB_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1M1.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1M1_orf2", RepeatID = L1M1_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1M2.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1M2_orf2", RepeatID = L1M2_ORF.repeats$RepeatID,
                                       start_pos = 1187, end_pos = 1213)

ZNF382.L1M3.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1M3_orf2", RepeatID = L1M3_ORF.repeats$RepeatID,
                                       start_pos = 1184, end_pos = 1210)

ZNF382.L1M4.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1M4_orf2", RepeatID = L1M4_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1M5.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1M5_orf2", RepeatID = L1M5_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1MC.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1MC_orf2", RepeatID = L1MC_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1MD.sites = TECookbook::liftOut(alignment = "hg38.fa.align",
                                       Repeat = "L1MD_orf2", RepeatID = L1MD_ORF.repeats$RepeatID,
                                       start_pos = 1190, end_pos = 1216)

ZNF382.L1P1.sites$Lineage = "L1P1"
ZNF382.L1P3.sites$Lineage = "L1P3"
```

```
ZNF382.L1P4.sites$Lineage = "L1P4"
ZNF382.L1PB.sites$Lineage = "L1PB"
ZNF382.L1M1.sites$Lineage = "L1M1"
ZNF382.L1M2.sites$Lineage = "L1M2"
ZNF382.L1M3.sites$Lineage = "L1M3"
ZNF382.L1M4.sites$Lineage = "L1M4"
ZNF382.L1MC.sites$Lineage = "L1MC"
ZNF382.L1MD.sites$Lineage = "L1MD"
ZNF382.L1M5.sites$Lineage = "L1M5"

ZNF382.sites =      c(ZNF382.L1P1.sites, ZNF382.L1P3.sites, ZNF382.L1P4.sites, ZNF382.L1PB.sites,
                      ZNF382.L1M1.sites, ZNF382.L1M2.sites, ZNF382.L1M3.sites, ZNF382.L1M4.sites, ZNF382
                      ZNF382.L1MC.sites, ZNF382.L1MD.sites)

require(BSgenome.Hsapiens.UCSC.hg38)
ZNF382.sites$Sequence = getSeq(Hsapiens, ZNF382.sites, as.character=TRUE)

save(list = "ZNF382.sites", file = "ZNF382.sites.RData")
```

```
require(dplyr)
require(GenomicRanges)
load("ZNF382.sites.RData")

Names  = c("L1P1","L1P3","L1P4","L1PB",
           "L1M1","L1M2","L1M3","L1M4","L1MC","L1MD", "L1M5")

TotalReads.Control = 58447968
TotalReads.ZNF382  = 13697977
windowSize = 10

ZNF382.forward.signals = read.table("../../ZNF382/ChIP-exo data analysis//ZNF382.Repeat.plus.bedgraph",
                                    col.names = c("Repeat", "start", "end", "Signal")) %>%
                         mutate(Strand = "Forward",
                                Signal = Signal*1e6/TotalReads.ZNF382)
ZNF382.reverse.signals = read.table("../../ZNF382/ChIP-exo data analysis/ZNF382.Repeat.minus.bedgraph",
                                    col.names = c("Repeat", "start", "end", "Signal")) %>%
                         mutate(Strand = "Reverse",
                                Signal = Signal*1e6/TotalReads.ZNF382)
rbind(ZNF382.forward.signals, ZNF382.reverse.signals) %>%
  dplyr::filter(Repeat %>% endsWith("orf2")) %>%
  mutate(start = as.integer(start/windowSize)*windowSize) %>%
  group_by(start, Strand, Repeat) %>%
  summarise(Signal = mean(Signal)) -> ZNF382.signals
```

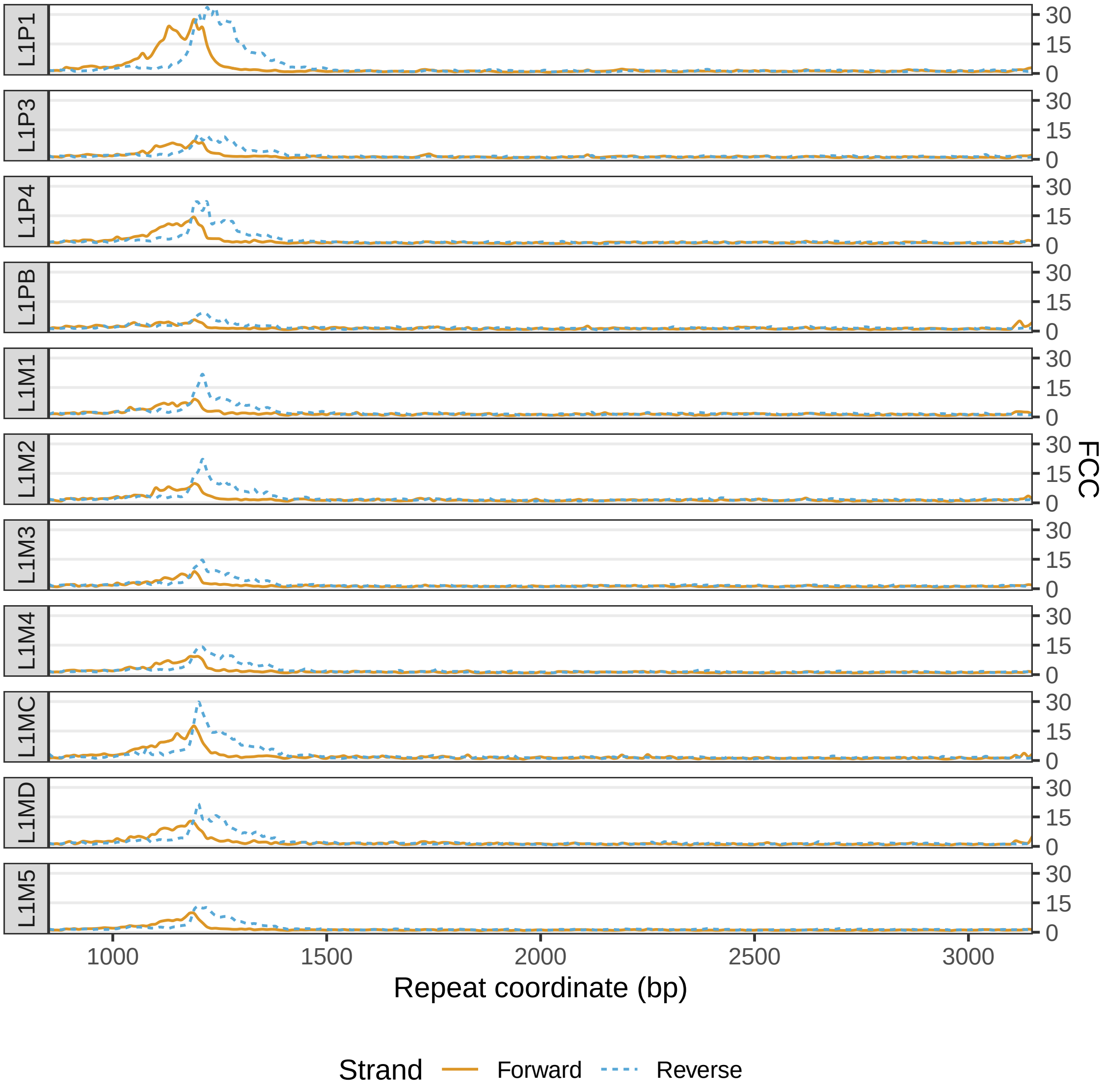
```
## `summarise()` has grouped output by 'start', 'Strand'. You can override using
## the `.groups` argument.
```

```
Control.forward.signals = read.table("../../KAP1/ChIP-exo Input/Control.Repeat.plus.bedgraph",
                                    col.names = c("Repeat", "start", "end", "Signal")) %>%
                         mutate(Strand = "Forward",
                                Signal = Signal*1e6/TotalReads.Control)

Control.reverse.signals = read.table("../../KAP1/ChIP-exo Input/Control.Repeat.minus.bedgraph",
                                    col.names = c("Repeat", "start", "end", "Signal")) %>%
                         mutate(Strand = "Reverse",
                                Signal = Signal*1e6/TotalReads.Control)

rbind(Control.forward.signals, Control.reverse.signals) %>%
  dplyr::filter(Repeat %>% endsWith("orf2")) %>%
  mutate(start = as.integer(start/windowSize)*windowSize) %>%
  group_by(start, Strand, Repeat) %>%
  summarise(Signal = mean(Signal)) -> Control.signals
```

```
## `summarise()` has grouped output by 'start', 'Strand'. You can override using
## the `.groups` argument.
```

```
inner_join(ZNF382.signals, Control.signals,
           by = c("Repeat", "start", "Strand"),
           suffix = c(".ZNF382", ".Control")) %>%
  mutate(Repeat = factor(Repeat, levels = paste0(Names, "_orf2")),
         FCC    = Signal.ZNF382/Signal.Control) %>%
  dplyr::filter(Repeat %in% levels(Repeat)) %>%
  ggplot(aes(x = start, y = FCC, color = Strand, linetype = Strand)) +
  ggalt::geom_xspline(spline_shape = 0.4)+
  theme_bw()+
  theme(legend.position = "bottom", panel.grid.minor = element_blank(), panel.grid.major.x = element_bla
  scale_color_manual(values = c("#DC9627", "#59A9D7"))+
  scale_x_continuous(limits = c(850,3150), expand = c(0,0))+
  scale_y_continuous( position = "right", breaks = c(0, 15, 30))+
  facet_wrap(~Repeat, ncol = 1, strip.position = "left", labeller = as_labeller(function(x) substr(x, s
  xlab("Repeat coordinate (bp)")
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

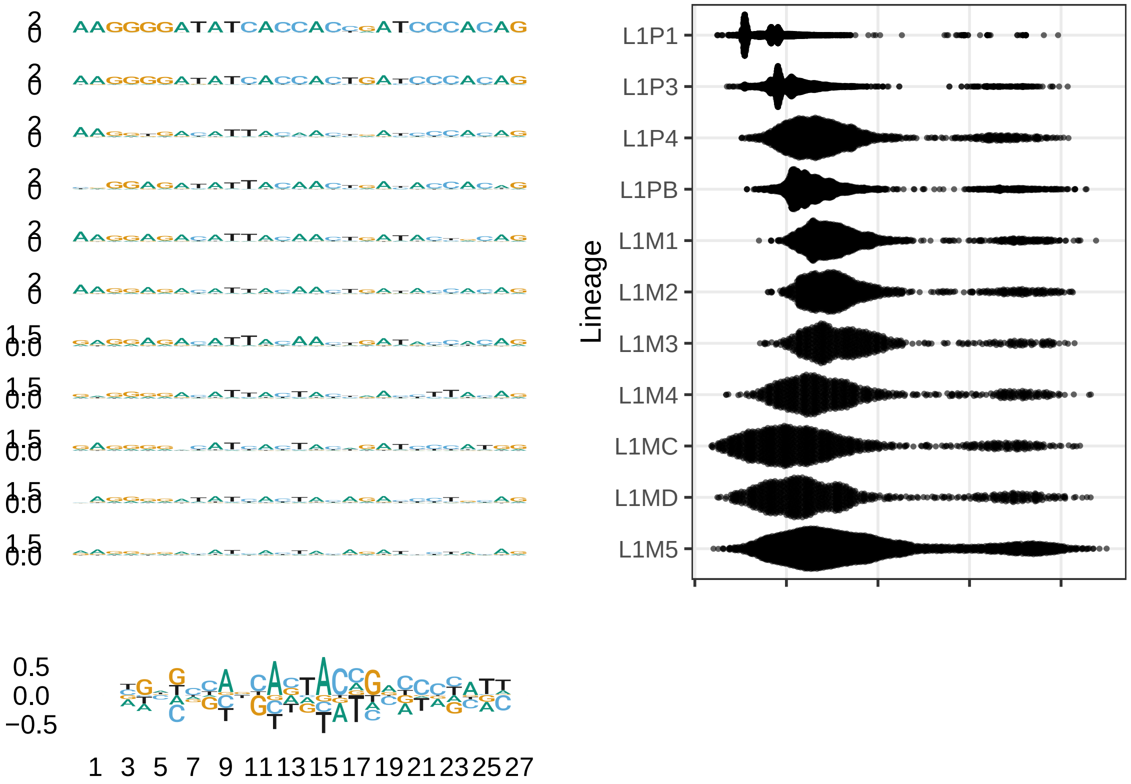


```
#ggsave("ZNF382 ChIP-exo profies on repeat coordinates.svg", height = 6, width = 6)
```

```
require(dplyr)
require(GenomicRanges)
load("ZNF382.sites.RData")

ZNF382.sites %>% as_tibble()
```

```
## # A tibble: 33,513 x 8
##    seqnames    start      end width strand Sequence              Repea~1 Lineage
##    <fct>       <int>    <int> <int> <fct>  <chr>                 <chr>   <chr>
##  1 chr1      4863590  4863616    27 -      AAGGGGATATCACCACTGAT~ 6063    L1P1
##  2 chr1      7414585  7414611    27 -      AAGGGGATATCACCACCGAT~ 10179   L1P1
##  3 chr1     11086448 11086474    27 -      AAGGGGATATCACCACTGAT~ 17576   L1P1
##  4 chr1     11164405 11164431    27 +      AAGGGGATATCACCACCGAT~ 17727   L1P1
##  5 chr1     12815830 12815856    27 -      AAGGGCATATCATCACGGAT~ 21136   L1P1
##  6 chr1     14385143 14385169    27 -      AAGGGGTTATCACCACTGAT~ 23910   L1P1
##  7 chr1     20005538 20005563    26 +      AAGGGATATCACCACTGATC~ 36294   L1P1
##  8 chr1     25058223 25058249    27 +      AAGGGGTTATCACCACTGAT~ 48126   L1P1
##  9 chr1     25509863 25509889    27 +      AAGGGGATATCACCACTGAT~ 49175   L1P1
## 10 chr1     29402868 29402894    27 -      AAGGGGATATCACCACCAAT~ 58710   L1P1
## # ... with 33,503 more rows, and abbreviated variable name 1: RepeatID
```

```
    col_scheme = ggseqlogo::make_col_scheme(chars=c('A', 'C', 'G', 'T'),
                                            #cols=c('darkgreen', 'blue', 'orange', 'red'))
                                            cols=c("#0E927B", "#59A9D8", "#DC9514", "#1A1A1A"))
for(x in c("L1P1", "L1P3", "L1P4", "L1PB", "L1M1", "L1M2")){
  assign(paste0("ZNF382.",x,".logo"),
         subset(ZNF382.sites, Lineage==x & width==27)$Sequence %>%
           ggseqlogo::ggseqlogo(col_scheme=col_scheme) + scale_y_continuous(limits = c(0,2),breaks = c(
           theme(axis.text.x = element_blank(), axis.title = element_blank(), axis.ticks.x = element_bla

for(x in c("L1M3", "L1M4", "L1MC", "L1MD", "L1M5")){
  assign(paste0("ZNF382.",x,".logo"),
         subset(ZNF382.sites, Lineage==x & width==27)$Sequence %>%
           ggseqlogo::ggseqlogo(col_scheme=col_scheme) + scale_y_continuous(limits = c(0,1.5),breaks =
           theme(axis.text.x = element_blank(), axis.title = element_blank(), axis.ticks.x = element_bla

cowplot::plot_grid(ZNF382.L1P1.logo, ZNF382.L1P3.logo, ZNF382.L1P4.logo, ZNF382.L1PB.logo,
                   ZNF382.L1M1.logo, ZNF382.L1M2.logo, ZNF382.L1M3.logo, ZNF382.L1M4.logo,
                   ZNF382.L1MC.logo, ZNF382.L1MD.logo, ZNF382.L1M5.logo,
                   ncol=1, align = "v") -> plot.Logos

load("../../ZNF382/ZNF382.motif.RData")

ZNF382.PEM = cbind(rep(0,4), rep(0,4), ZNF382.PEM, rep(0,4))

TFCookbook::plotEnergyLogo(ZNF382.PEM) +
  scale_x_continuous(breaks = seq(1,27,2)) + ylim(-0.8, 0.8)+
  theme(axis.title = element_blank())-> ZNF382.Spec.logo

ZNF382.sites$predicted.Energy = TFCookbook::predictEnergy(ZNF382.sites$Sequence, ZNF382.PEM)

subset(ZNF382.sites, width==27) %>%
  as_tibble() %>%
  mutate(Lineage = forcats::fct_rev(factor(Lineage, levels = Names))) %>%
  ggplot(aes(x = predicted.Energy, y = Lineage), alpha = 0.5)+
  ggbeeswarm::geom_quasirandom(groupOnX = FALSE, size = 0.4, alpha = 0.6)+
  scale_x_continuous(breaks = seq(-6, 3, 2), minor_breaks = NULL)+
  theme_bw() +
  theme(axis.title.x = element_blank(), axis.text.x = element_blank())-> plot.Energy

cowplot::plot_grid(plot.Logos, plot.Energy, nrow = 2,
                   ZNF382.Spec.logo, align = "h", rel_heights = c(1, 0.26))
```



```
#ggsave("ZNF382 logos and energy distribution.svg", width = 9, height = 6.3)
```