R-project

Neha Gupta

12/17/2018

Background

- Eukaryotic translation initiation involves a dozen of known translation initiation factors
- Ded1- ATP-dependent RNA helicase- plays an important role in translation initiation
- Dbp1- a paralog of Ded1, role in translation initiation is unclear
- Ribosome footprint profiling and mRNA-seq analyses- determined the changes in translation efficiency (TE) in Δdbp1, ded1^{ts}, and Δdbp1ded1^{ts}
- Goal: Select few candidate mRNAs to test in purified in vitro reconstituted translation initiation assay

Setting up directory and attaching packages

```
> setwd ("/Users/guptan8/Desktop/R_project")
> library(tidyverse)
```

```
── Attaching packages ──────────────────────────────────────────────────────────
── tidyverse 1.2.1 ──
✓ ggplot2 3.0.0     ✓ purrr   0.2.5
✓ tibble  1.4.2     ✓ dplyr   0.7.6
✓ tidyr   0.8.1     ✓ stringr 1.3.1
✓ readr   1.1.1     ✓ forcats 0.3.0
```

Reading a .csv file and understanding the data

```
> ded <- read_csv("ded.csv")
> dim(ded)
[1] 523  35
```

```
> str(ded)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      523 obs. of  35 variables:
 $ X1                    : chr  "Q0140" "YAL002W" "YAL009W" "YAL016W" ...
 $ X2                    : chr  "Q0140" "YAL002W" "YAL009W" "YAL016W" ...
 $ Pelechano 2013        : chr  "0" "95" "25" "81" ...
 $ Nagalakshmi           : chr  "0" "0" "26" "0" ...
 $ Xu 2009               : chr  "#N/A" "110" "33" "81" ...
 $ Lawless               : chr  "#N/A" "415" "#N/A" "97" ...
 $ X7                    : chr  "#N/A" "YAL002W" "YAL009W" "YAL016W" ...
 $ Techange_Dbp1         : chr  "#N/A" "-0.3312571" "-0.467667" "0.0063495" ...
 $ Techange_Ded1ts       : chr  "#N/A" "-0.75009" "-1.17144" "-0.77361" ...
 $ Techange_dd           : chr  "#N/A" "-1.67665" "-2.31827" "-1.03332" ...
 ….more
```

Several columns are character instead of numeric type and have “NA”

Some column headings are not very useful

```
> names(ded)
 [1] "X1"                    "X2"                    "Pelechano 2013"
 [4] "Nagalakshmi"           "Xu 2009"               "Lawless"
 [7] "X7"                    "Techange_Dbp1"         "Techange_Ded1ts"
[10] "Techange_dd"           "Techange_conditionalDed1" "teChange_conditionalDbp1"
[13] "dbp1"                  "ded1"                  "X15"
[16] "biolmrna"              "biolFT"                "treatded1dbp1"
[19] "biolFT.treatded1dbp1"  "logFDREffect"          "logFDRTrl"
[22] "mrnawtts"              "FTwtts"                "mrnaded1dbp1"
[25] "FTded1dbp1"            "mrnaChange"            "FTChange"
[28] "tewtts"                "teded1dbp1"            "teChange"
[31] "Gene"                  "Description"           "X33"
[34] "X34"                   "X35"
```

Changing column names

```
> names(ded)[1] <- "Genes"
> names(ded)[8] <-"TE_Dbp1"
> names(ded)[9] <-"TE_Ded1"
> names(ded)[10] <-"TE_Dbp1Ded1"
> names(ded)
 [1] "Genes"                  "X2"                     "Pelechano 2013"
 [4] "Nagalakshmi"            "Xu 2009"                "Lawless"
 [7] "X7"                     "TE_Dbp1"                "TE_Ded1"
[10] "TE_Dbp1Ded1"            "Techange_conditionalDed1" "teChange_conditionalDbp1"
[13] "dbp1"                   "ded1"                   "X15"
[16] "biolmrna"               "biolFT"                 "treatded1dbp1"
[19] "biolFT.treatded1dbp1"   "logFDREffect"           "logFDRTrl"
[22] "mrnawtts"               "FTwtts"                 "mrnaded1dbp1"
[25] "FTded1dbp1"             "mrnaChange"             "FTChange"
[28] "tewtts"                 "teded1dbp1"             "teChange"
[31] "Gene"                   "Description"            "X33"
[34] "X34"                    "X35"
```

Cleaning data

```
> ded2 <- ded [,1:10] %>%  #selecting first 10 columns
+   select(-X2,-X7)  %>%  # deleting columns 2 and 7
+   mutate(Genes = as.factor(Genes)) %>%
+   mutate_if(is.character, as.numeric) %>% # changing columns 2 to 8 to numeric
+   mutate(Genes = as.character(Genes)) %>%
+   filter(!is.na(TE_Dbp1)) # removing NAs
> head(ded2)
# A tibble: 6 x 8
  Genes   `Pelechano 2013` Nagalakshmi `Xu 2009` Lawless  TE_Dbp1 TE_Ded1 TE_Dbp1Ded1
  <chr>            <dbl>      <dbl>    <dbl>   <dbl>    <dbl>   <dbl>       <dbl>
1 YAL002W             95          0      110     415 -0.331    -0.750      -1.68
2 YAL009W             25         26       33      NA -0.468    -1.17       -2.32
3 YAL016W             81          0       81      97  0.00635  -0.774      -1.03
4 YAL017W              0        368      363     367 -0.0783   -0.832       -1.71
5 YAL022C             16         34       36      NA  0.228    -1.01       -1.67
6 YAL029C            252         48      244      NA  0.135    -0.594      -1.10
```

Cleaner data with desirable data type

```
> str(ded2)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	522 obs. of  8 variables:
 $ Genes          : chr  "YAL002W" "YAL009W" "YAL016W" "YAL017W" ...
 $ Pelechano 2013: num  95 25 81 0 16 252 321 97 145 0 ...
 $ Nagalakshmi    : num  0 26 0 368 34 48 319 227 151 0 ...
 $ Xu 2009        : num  110 33 81 363 36 244 330 88 161 NA ...
 $ Lawless        : num  415 NA 97 367 NA NA NA 89 NA NA ...
 $ TE_Dbp1        : num  -0.33126 -0.46767 0.00635 -0.07831 0.22817 ...
 $ TE_Ded1        : num  -0.75 -1.171 -0.774 -0.832 -1.006 ...
 $ TE_Dbp1Ded1    : num  -1.68 -2.32 -1.03 -1.71 -1.67 ...
```

Checking for 'NA'

```
> apply(ded2, 2, function(x) sum(is.na(x)))
```

Genes	Pelechano 2013	Nagalakshmi	Xu 2009	Lawless	TE_Dbp1
0	1	8	31	198	0

TE_Ded1	TE_Dbp1Ded1
0	0

Columns starting with 'TE' and Genes have 0 NAs. It's good enough for the subsequent data analysis.

TE_Dbp1Ded1 and TE_Dbp1 have very weak (if any) correlation

> pdf(file = "A.pdf", width=5, height=5)

> ggplot(ded2, aes(x = TE_Dbp1Ded1, y = TE_Dbp1)) + geom_point() + scale_x_reverse() + geom_smooth(method='lm')

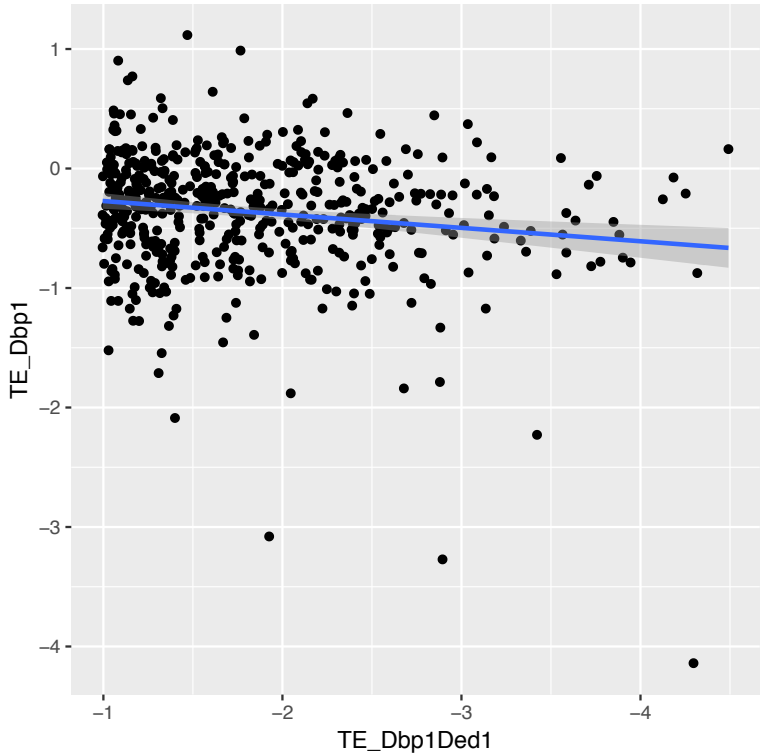
> dev.off()

> cor.test(ded2$TE_Dbp1, ded2$TE_Dbp1Ded1, method = "spearman")

> cor.test(ded2$TE_Dbp1, ded2$TE_Dbp1Ded1, method = "pearson")

rho = 0.1066642

cor = 0.1588643



TE_Dbp1Ded1 and TE_Ded1 are correlated

> pdf(file = "ScatterB.pdf", width=5, height=5)

> ggplot(ded2, aes(x = TE_Dbp1Ded1, y = TE_Ded1)) + geom_point() + scale_x_reverse() + geom_smooth(method='lm')

> dev.off()

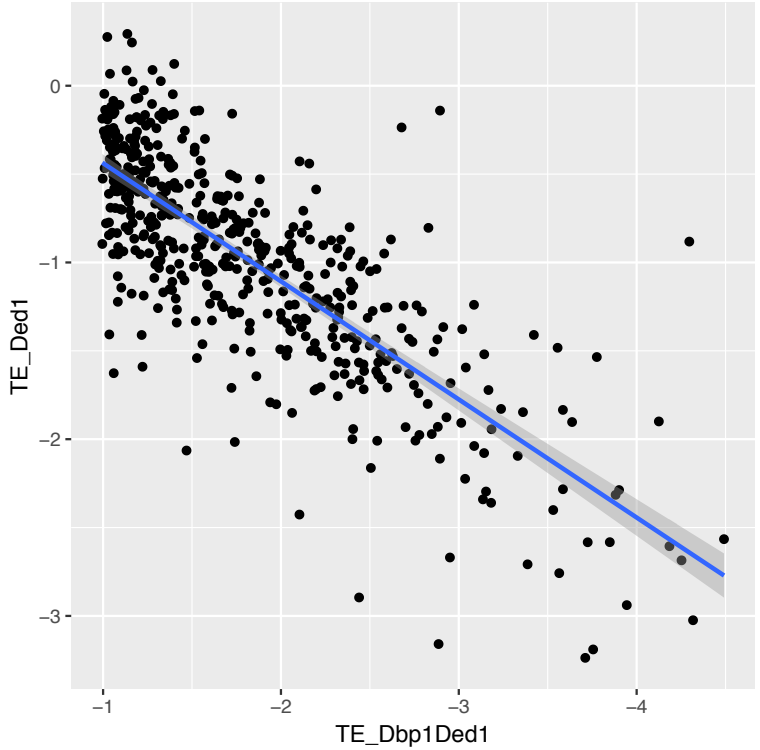
> cor.test(ded2$TE_Ded1, ded2$TE_Dbp1Ded1, method = "spearman")

> cor.test(ded2$TE_Ded1, ded2$TE_Dbp1Ded1, method = "pearson")

rho = 0.7618936 (Warning message:

Cannot compute exact p-value with ties)

cor = 0.7868674



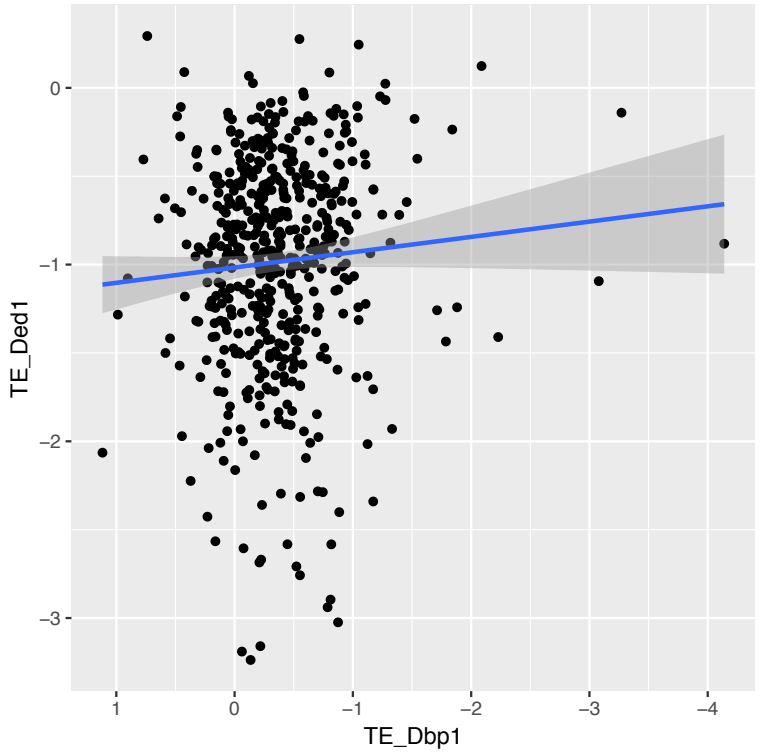
TE_Ded1 and TE_Dbp1 are not correlated

```
> pdf(file = "scatterC.pdf", width=5, height=5)

> ggplot(ded2, aes(x = TE_Dbp1, y = TE_Ded1)) + geom_point() +
scale_x_reverse() + geom_smooth(method='lm')

> dev.off()

> cor.test(ded2$TE_Ded1, ded2$TE_Dbp1, method = "spearman")

> cor.test(ded2$TE_Ded1, ded2$TE_Dbp1, method = "pearson")
```

rho = -0.07614446 (Warning message:

Cannot compute exact p-value with ties)

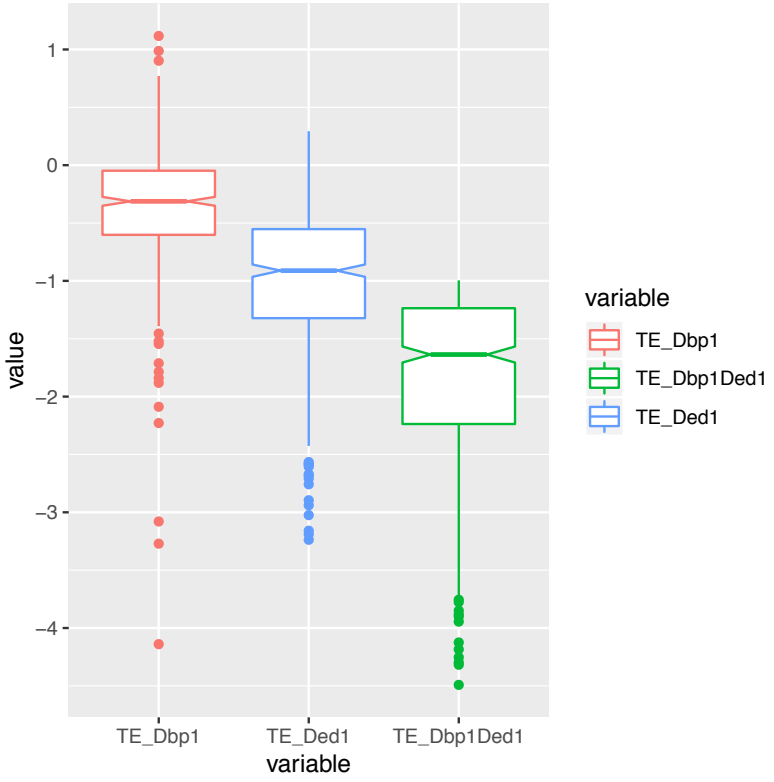
cor = -0.07204412



Boxplot may be a better way to visualize

```
> ded3 <- ded2 %>%
+   select(Genes, TE_Dbp1, TE_Ded1, TE_Dbp1Ded1) %>%
+   gather(variable, value, -Genes)
>head(ded3)
> pdf(file = "Boxplot.pdf", width=5, height=5)
> ggplot(ded3, aes(x = variable, y = value, color = variable)) +
geom_boxplot(notch = TRUE) + scale_x_discrete
(limits=c("TE_Dbp1", "TE_Ded1", "TE_Dbp1Ded1"))
> dev.off()
```

```
# A tibble: 6 x 3
  Genes    variable    value
  <chr>    <chr>       <dbl>
1 YAL002W  TE_Dbp1   -0.331
2 YAL009W  TE_Dbp1   -0.468
3 YAL016W  TE_Dbp1    0.00635
4 YAL017W  TE_Dbp1   -0.0783
5 YAL022C  TE_Dbp1    0.228
6 YAL029C  TE_Dbp1    0.135
```



Does 5'-UTR length affect TE_Dbp1Ded1?

```
> ded4 <- ded2 %>%
+   select(Genes, Nagalakshmi,TE_Dbp1, TE_Ded1, TE_Dbp1Ded1) %>%
+   filter(!Nagalakshmi == 0) %>% #NAs became 0
+   rename(Leader_length = Nagalakshmi)
> head(ded4)
```

```
# A tibble: 6 x 5
  Genes   Leader_length TE_Dbp1 TE_Ded1 TE_Dbp1Ded1
  <chr>           <dbl>   <dbl>   <dbl>       <dbl>
1 YAL009W            26 -0.468   -1.17        -2.32
2 YAL017W           368 -0.0783  -0.832       -1.71
3 YAL022C            34  0.228   -1.01        -1.67
4 YAL029C            48  0.135   -0.594       -1.10
5 YAL040C           319 -1.39    -0.719       -1.84
6 YAL061W           227 -0.280   -0.284       -1.07
```

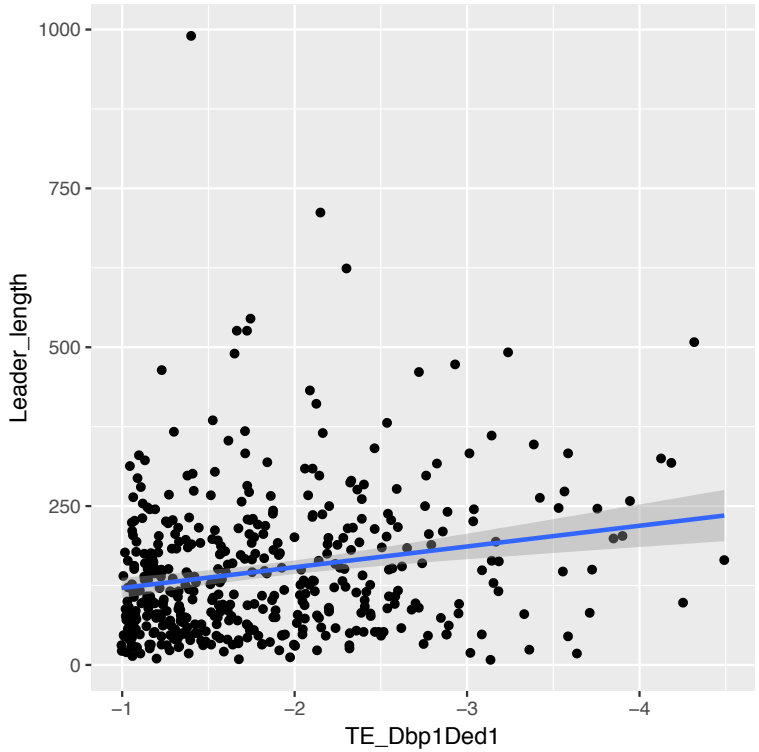
Does 5’-UTR length affect TE_Dbp1Ded1? Maybe

```
> pdf(file = "UTR.pdf", width=5, height=5)

> ggplot(ded4, aes(x = TE_Dbp1Ded1, y = Leader_length)) + geom_point() +
scale_x_reverse() + geom_smooth(method='lm')

> dev.off()

> cor.test(ded4$Leader_length, ded4$TE_Dbp1Ded1, method = "spearman")

> cor.test(ded4$Leader_length, ded4$TE_Dbp1Ded1, method = "pearson")
```

rho = -0.2098855 Warning message:

Cannot compute exact p-value with ties

cor = -0.2043999



Final list of candidate mRNAs

```
> ded5 <- ded4 %>%
+   filter(Leader_length < 300) %>%    #Want length <300nt
+   filter(TE_Dbp1 > -0.25) %>%   # Weak correlation
+   filter(TE_Ded1 > -0.25)   #Strong correlation
> ded5

# A tibble: 9 x 5
  Genes   Leader_length TE_Dbp1 TE_Ded1 TE_Dbp1Ded1
  <chr>           <dbl>   <dbl>   <dbl>       <dbl>
1 YDR466W            97 -0.144  -0.179        -1.19
2 YGR108W           116  0.425   0.0894       -1.28
3 YGR167W            52  0.0219 -0.178        -1.19
4 YKL218C           127  0.486  -0.161        -1.06
5 YKR080W           151 -0.222  -0.105        -1.07
6 YLR071C            32  0.0307 -0.249        -1.06
7 YLR455W           216 -0.155   0.0264       -1.32
8 YOR129C            79 -0.121   0.0678       -1.04
9 YOR191W            88  0.0510 -0.162        -1.02
```