# Integrity of TEs and expression candidates

Contents

[Integrity of TEs and expression candidates 1](#_Toc57062916)

[Integrity of TEs and expression candidates 2](#_Toc57062917)

[1. Build a file containing integrity information of all annotated TEs 2](#_Toc57062918)

[2. All annotated TEs 2](#_Toc57062919)

[3. Expression candidates of control (T=0) 2](#_Toc57062920)

[4. Expression candidates of mock treatment 3](#_Toc57062921)

[5. Expression candidates of yeast treatment 3](#_Toc57062922)

[6. Expression candidates of botrytis treatment 4](#_Toc57062923)

# Integrity of TEs and expression candidates

## Build a file containing integrity information of all annotated TEs

#!/bin/bash

#index fasta file of intact TEs to extract length of each TE family

samtools faidx TE.intact.fa

#convert gff to gtf

gffread **-**E TE.anno.gff **-**o TE.anno.gtf

#R code

#prepare a reference file containing all TEs and length and length ratio etc.

setwd**(**"/path/to/your/working/directory/"**)**

ref**=**read.table**(**textConnection**(**gsub**(**" ", "\t", readLines**(**"TE.anno.gtf"**))))**

ref**$**V10**=**gsub**(**";", "", ref**$**V10**)**

ref**$**V12**=**gsub**(**";", "", ref**$**V12**)**

ref**$**V14**=**gsub**(**";", "", ref**$**V14**)**

ref**$**V16**=**gsub**(**";", "", ref**$**V16**)**

ref**=**ref**[**,c**(**12,1,4,5,7,10,14,16**)]**

colnames**(**ref**)=**c**(**"id","chr","start","end","strand","family","superfamily","class"**)**

ref**$**length**=**ref**$**end**-**ref**$**start**+**1

reflength**=**read.table**(**"TE.intact.fai", header**=**F, sep**=**"\t"**)**

reflength**=**reflength**[**,c**(**1,2**)]**

colnames**(**reflength**)=**c**(**"family", "consensus"**)**

ref**=**merge**(**ref, reflength, by.x**=**"family", all.x**=TRUE)**

ref**=**ref**[**,c**(**2**:**6,1,7**:**10**)]**

ref**$**Lratio**=**ref**$**length**/**ref**$**consensus

write.table**(**ref, "AllTEsExpanded\_curated\_tags\_V4.txt", col.names**=**T, row.names**=**F, sep**=**"\t", quote**=**F**)**

## All annotated TEs

#R code

setwd**(**"/PhD/analysis/Reference/Curated"**)**

df**=**read.table**(**"AllTEsExpanded\_curated\_tags\_V4.txt", header**=**T, sep**=**"\t"**)**

FL**=**subset**(**df, df**$**Lratio**>**0.9**)** #full length candidates

FG**=**subset**(**df, df**$**Lratio**<=**0.9**)** #fragmented candidates

df**$**Integrity**=**df**$**Lratio**\***100

temp**=**data.frame**(**bin**=**numeric**(**0**)**, count**=**numeric**(**0**))**

**for** **(**i **in** 1**:**100**){**

x**=**subset**(**df, df**$**Integrity**<=**i **&** df**$**Integrity**>**i**-**1 **)**

y**=**nrow**(**x**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**i, count**=**y**)**

**}**

z**=**subset**(**df, df**$**Integrity**>**100**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**101, count**=**nrow**(**z**))**

temp**$**logC**=**log2**(**as.numeric**(**temp**$**count**))**

library**(**ggplot2**)**

ggplot**(**temp, aes**(**x**=**bin, y**=**logC**))** **+**

theme\_bw**()+**

geom\_col**()+**

scale\_x\_continuous**(**breaks **=** seq**(**0, 100, by **=** 10**))+**

geom\_vline**(**xintercept **=** 90, linetype **=** 2, size**=**1.5, colour **=** "firebrick1"**)**

dev.copy**(**png,"TEintegrity\_allTE.png", width **=**1200, height **=** 200**)**

dev.off**()**

## Expression candidates of control (T=0)

#R code

setwd**(**"/PhD/analysis/ECstress\_TEalignment/ExprCandidate"**)**

data**=**read.table**(**"AllExpeCandidate\_ctrl\_tag\_new.txt", header**=**T, sep**=**"\t"**)**

FL**=**subset**(**data, data**$**Lratio**>**0.9**)** #full length candidates

FG**=**subset**(**data, data**$**Lratio**<=**0.9**)** #fragmented candidates

data**$**Integrity**=**data**$**Lratio**\***100

df**=**data

temp**=**data.frame**(**bin**=**numeric**(**0**)**, count**=**numeric**(**0**))**

**for** **(**i **in** 1**:**100**){**

x**=**subset**(**df, df**$**Integrity**<=**i **&** df**$**Integrity**>**i**-**1 **)**

y**=**nrow**(**x**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**i, count**=**y**)**

**}**

z**=**subset**(**df, df**$**Integrity**>**100**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**101, count**=**nrow**(**z**))**

temp**$**logC**=**log2**(**as.numeric**(**temp**$**count**))**

library**(**ggplot2**)**

ggplot**(**temp, aes**(**x**=**bin, y**=**logC**))** **+**

theme\_bw**()+**

geom\_col**()+**

scale\_x\_continuous**(**breaks **=** seq**(**0, 100, by **=** 10**))+**

scale\_y\_continuous**(**limits**=**c **(**0, 10**))+**

geom\_vline**(**xintercept **=** 90, linetype **=** 2, size**=**1.5, colour **=** "firebrick1"**)**

dev.copy**(**png,"TEintegrity\_AllExprCandidate\_ctrl.png", width **=**1200, height **=** 200**)**

dev.off**()**

## Expression candidates of mock treatment

#R code

setwd**(**"/PhD/analysis/ECstress\_TEalignment/ExprCandidate"**)**

data**=**read.table**(**"AllExpeCandidate\_mock\_tag\_new.txt", header**=**T, sep**=**"\t"**)**

FL**=**subset**(**data, data**$**Lratio**>**0.9**)** #full length candidates

FG**=**subset**(**data, data**$**Lratio**<=**0.9**)** #fgragmented candidates

data**$**Integrity**=**data**$**Lratio**\***100

df**=**data

temp**=**data.frame**(**bin**=**numeric**(**0**)**, count**=**numeric**(**0**))**

**for** **(**i **in** 1**:**100**){**

x**=**subset**(**df, df**$**Integrity**<=**i **&** df**$**Integrity**>**i**-**1 **)**

y**=**nrow**(**x**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**i, count**=**y**)**

**}**

z**=**subset**(**df, df**$**Integrity**>**100**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**101, count**=**nrow**(**z**))**

temp**$**logC**=**log2**(**as.numeric**(**temp**$**count**))**

library**(**ggplot2**)**

ggplot**(**temp, aes**(**x**=**bin, y**=**logC**))** **+**

theme\_bw**()+**

geom\_col**()+**

scale\_x\_continuous**(**breaks **=** seq**(**0, 100, by **=** 10**))+**

scale\_y\_continuous**(**limits**=**c **(**0, 10**))+**

geom\_vline**(**xintercept **=** 90, linetype **=** 2, size**=**1.5, colour **=** "firebrick1"**)**

dev.copy**(**png,"TEintegrity\_AllExprCandidate\_mock.png", width **=**1200, height **=** 200**)**

dev.off**()**

## Expression candidates of yeast treatment

#R code

setwd**(**"/PhD/analysis/ECstress\_TEalignment/ExprCandidate"**)**

data**=**read.table**(**"AllExpeCandidate\_yeast\_tag\_new.txt", header**=**T, sep**=**"\t"**)**

FL**=**subset**(**data, data**$**Lratio**>**0.9**)** #full length candidates

FG**=**subset**(**data, data**$**Lratio**<=**0.9**)** #fragmented candidates

data**$**Integrity**=**data**$**Lratio**\***100

df**=**data

temp**=**data.frame**(**bin**=**numeric**(**0**)**, count**=**numeric**(**0**))**

**for** **(**i **in** 1**:**100**){**

x**=**subset**(**df, df**$**Integrity**<=**i **&** df**$**Integrity**>**i**-**1 **)**

y**=**nrow**(**x**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**i, count**=**y**)**

**}**

z**=**subset**(**df, df**$**Integrity**>**100**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**101, count**=**nrow**(**z**))**

temp**$**logC**=**log2**(**as.numeric**(**temp**$**count**))**

library**(**ggplot2**)**

ggplot**(**temp, aes**(**x**=**bin, y**=**logC**))** **+**

theme\_bw**()+**

geom\_col**()+**

scale\_x\_continuous**(**breaks **=** seq**(**0, 100, by **=** 10**))+**

scale\_y\_continuous**(**limits**=**c **(**0, 10**))+**

geom\_vline**(**xintercept **=** 90, linetype **=** 2, size**=**1.5, colour **=** "firebrick1"**)**

dev.copy**(**png,"TEintegrity\_AllExprCandidate\_yeast.png", width **=**1200, height **=** 200**)**

dev.off**()**

## Expression candidates of botrytis treatment

#R code

setwd**(**"/PhD/analysis/ECstress\_TEalignment/ExprCandidate"**)**

data**=**read.table**(**"AllExpeCandidate\_botrytis\_tag\_new.txt", header**=**T, sep**=**"\t"**)**

FL**=**subset**(**data, data**$**Lratio**>**0.9**)** #full length candidates

FG**=**subset**(**data, data**$**Lratio**<=**0.9**)** #fragmented candidates

data**$**Integrity**=**data**$**Lratio**\***100

df**=**data

temp**=**data.frame**(**bin**=**numeric**(**0**)**, count**=**numeric**(**0**))**

**for** **(**i **in** 1**:**100**){**

x**=**subset**(**df, df**$**Integrity**<=**i **&** df**$**Integrity**>**i**-**1 **)**

y**=**nrow**(**x**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**i, count**=**y**)**

**}**

z**=**subset**(**df, df**$**Integrity**>**100**)**

temp**[**nrow**(**temp**)+**1,**]<-**c**(**bin**=**101, count**=**nrow**(**z**))**

temp**$**logC**=**log2**(**as.numeric**(**temp**$**count**))**

library**(**ggplot2**)**

ggplot**(**temp, aes**(**x**=**bin, y**=**logC**))** **+**

theme\_bw**()+**

geom\_col**()+**

scale\_x\_continuous**(**breaks **=** seq**(**0, 100, by **=** 10**))+**

scale\_y\_continuous**(**limits**=**c **(**0, 10**))+**

geom\_vline**(**xintercept **=** 90, linetype **=** 2, size**=**1.5, colour **=** "firebrick1"**)**

dev.copy**(**png,"TEintegrity\_AllExprCandidate\_botrytis.png", width **=**1200, height **=** 200**)**

dev.off**()**