compare dCNS result with plantregmap gao lab

Baoxing Song

2020-09-01

The maize conservation elements were downloaded from http://plantregmap.gao-lab.org/download.php#Cis-elements-ce
But their coordinate is in maize V3. Download the v3 to v3 chain file form http://ftp.gramene.org/CURRENT_RELEASE/assembly_chain/zea_mays/

Run `CrossMap.py gff AGPv3_to_B73_RefGen_v4.chain CE_Zma.gtf CE_Zma_v4.gtf` to get the V4 version coordinate

reformate gff into bed file `cat CE_Zma_v4.gtf | awk '{print($1"\t"$4-1"\t"$5"\t"$10"\t"$5)}' | sed '~s/"//g' |sed '~s/;//g' > CE_Zma_v4.bed`

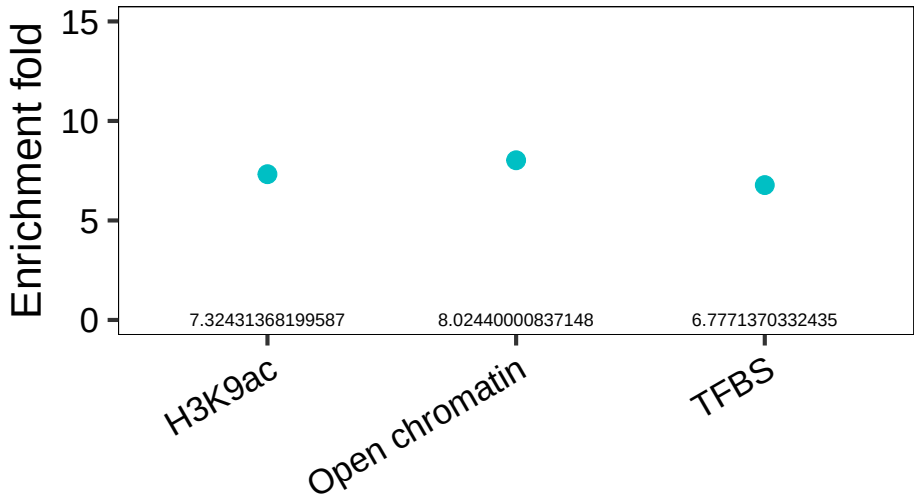
```
library(ggplot2)
# calculate the total length in V3 version
data = read.table("CE_Zma.gtf")
data$length = abs(data$V5 - data$V4)
sum(data$length)
## [1] 54959662

# calculate the total length in V4 version
data = read.table("CE_Zma_v4.gtf")
sum(data$V5 - data$V4)
## [1] 50661748


# using a custom code to classify each bp of maize genome

# non CDS region
n0 = 1976912700
#n-1 = 47856506 // CDS sequence size
n1 = 18939906  #   CE
n2 = 9648046 #   H3K9ac
n3 = 674282 # CE and H3K9ac
n4 = 22527326 # tfbs
n5 = 1543860 # ce and tfbs
n6 = 8663743  # H3K9ac and tfbs
n7 = 864214 # ce and H3K9ac and tfbs
n8 = 3130492
n9 = 290418
n10 = 632710
n11 = 82714
n12 = 9477610
```

compare dCNS result with plantregmap gao lab

```
n13 = 886575
n14 = 3734396
n15 = 472619

#
# n0 = 1952650998
# #n-1 = 39716814   // CDS sequence size
# n1 = 34928867    #   CE
# n2 = 11768654 #   H3K9ac
# n3 = 4034288  # CE and H3K9ac
# n4 = 24473236  # tfbs
# n5 = 4259411   # ce and tfbs
# n6 = 10941110  # H3K9ac and tfbs
# n7 = 4146041 # ce and H3K9ac and tfbs
# n8 = 3111981
# n9 = 369645
# n10 = 684611
# n11 = 185953
# n12 = 9421132
# n13 = 1030859
# n14 = 3864075
# n15 = 750442



n = sum(c(n0, n1, n2, n3, n4, n5, n6, n7, n8, n9, n10, n11, n12, n13, n14, n15))
n #whole backgroud size
## [1] 2058481611
a = (n/(n2+n3+n6+n10+n7+n11+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n3+n7+n11+n15))
a #H3K9ac
## [1] 7.324314

b = (n/(n4+n5+n6+n12+n7+n13+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n5+n7+n13+n15))
b
## [1] 6.777137
c = (n/(n8+n9+n10+n12+n11+n13+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n9+n11+n13+n15))
c
## [1] 8.0244
data = data.frame(x=c("TFBS", "H3K9ac", "Open chromatin"), y=c(b,a, c))

plot = ggplot(data=data, aes(x=x,y=y)) +geom_point(size=5, color="#00BFC4")+
  labs(x="", y="Enrichment fold")+ ylim(0, 15)+
  annotate(geom="text", x=1, y=0, label=a)+
  annotate(geom="text", x=2, y=0, label=c) +
  annotate(geom="text", x=3, y=0, label=b)+
  theme_bw() +theme_grey(base_size = 26) + theme(axis.line = element_blank(),
    legend.position = 'none',
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_rect(fill=NA,color="black", size=0.5, linetype="solid"),
    panel.background = element_blank(),
```

compare dCNS result with plantregmap gao lab

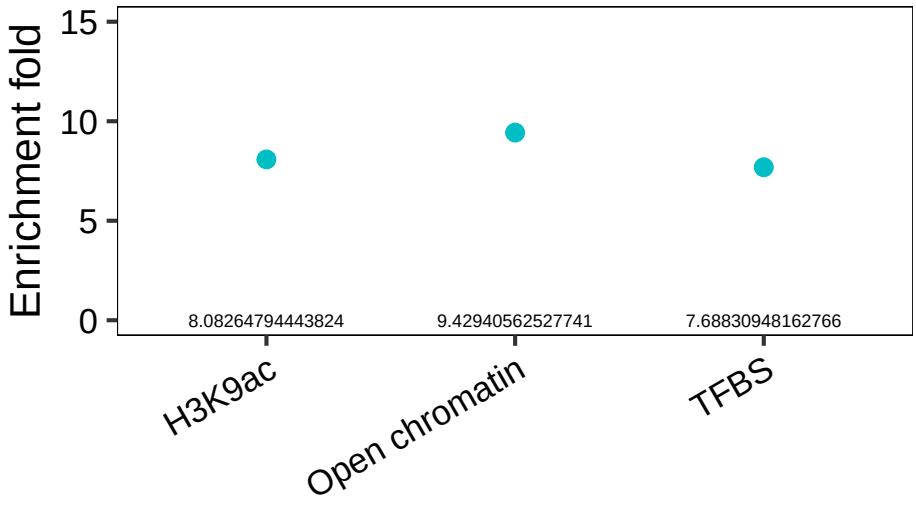
```
        axis.text.y = element_text( colour = "black"),
        axis.text.x = element_text(angle=30, hjust=1, vjust=1, colour = "black"))
plot
```




```
png("enrichment_non_cds.png" , width=500, height=460)
plot
dev.off()
## pdf
##   2
pdf("enrichment_non_cds.pdf" , width=7.5, height=6.9)
plot
dev.off()
## pdf
##   2


# the total CE out of CDS region on Chr1-10 is
(n1+n3+n5+n9+n7+n11+n13+n15)
## [1] 23754588

# this is significant smaller than core-And-CNS size

# non genetic
n0 = 1887846940
#n-1 = 161005347
n1 = 13896142
n2 = 5192145
n3 = 303985
n4 = 16713834
n5 = 964641
n6 = 4604701
n7 = 350105
n8 = 2761757
n9 = 234402
n10 = 415702
```

compare dCNS result with plantregmap gao lab

```
n11 = 41047
n12 = 8418372
n13 = 716259
n14 = 2607590
n15 = 265148


n = sum(c(n0, n1, n2, n3, n4, n5, n6, n7, n8, n9, n10, n11, n12, n13, n14, n15))
n
## [1] 1945332770
a = (n/(n2+n3+n6+n10+n7+n11+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n3+n7+n11+n15))
a
## [1] 8.082648
b = (n/(n4+n5+n6+n12+n7+n13+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n5+n7+n13+n15))
b
## [1] 7.688309
c = (n/(n8+n9+n10+n12+n11+n13+n14+n15))/((n1+n3+n5+n9+n7+n11+n13+n15)/(n9+n11+n13+n15))
c
## [1] 9.429406
data = data.frame(x=c("TFBS", "H3K9ac", "Open chromatin"), y=c(b,a, c))

plot = ggplot(data=data, aes(x=x,y=y)) +geom_point(size=5, color="#00BFC4")+
  labs(x="", y="Enrichment fold")+ ylim(0, 15)+
  annotate(geom="text", x=1, y=0, label=a)+
  annotate(geom="text", x=2, y=0, label=c) + annotate(geom="text", x=3, y=0, label=b)+
  theme_bw() +theme_grey(base_size = 26) + theme(axis.line = element_blank(),
    legend.position = 'none',
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_rect(fill=NA,color="black", size=0.5, linetype="solid"),
    panel.background = element_blank(),
    axis.text.y = element_text( colour = "black"),
    axis.text.x = element_text(angle=30, hjust=1, vjust=1, colour = "black"))
plot
```



compare dCNS result with plantregmap gao lab

```
png("enrichment_non_genetic.png" , width=500, height=460)
plot
dev.off()
## pdf
##   2
pdf("enrichment_non_genetic.pdf" , width=7.5, height=6.9)
plot
dev.off()
## pdf
##   2

# the total CE out of genetics region on Chr1-10 is
(n1+n3+n5+n9+n7+n11+n13+n15)
## [1] 16771729
```

Comparing with the dCNS result, they have low enrichment fold overlap with H3K9ac (which was 16.01333), TFBS(19.60267) and Open chromatin(23.10401)