main_transcripts

September 8, 2021

1 eQTL boxplot: Enrichment and Overlap of PGC2+CLOZUK

This is script ported from python to fix unknown plotting error.

[1]:

```
suppressPackageStartupMessages({
    library(tidyverse)
    library(ggpubr)
})
```

1.1 Functions

[2]:

```
feature = "transcripts"
```

1.1.1 Cached functions

[3]:

```
get_de_df <- function(){
    de_file = paste0("../../differential_expression/_m/", feature,
                     "/diffExpr_szVctl_full.txt")
    return(data.table::fread(de_file))
}
memDE <- memoise::memoise(get_de_df)

get_eqtl_df <- function(){
    eGenes_file = paste0('../../eqtl/caudate/summary_table/_m/',
                         'Brainseq_LIBD_caudate_4features.signifpairs.txt.gz')
    eGenes = data.table::fread(eGenes_file) %>%
        filter(Type == feature_map(feature)) %>%
        arrange(pval_nominal)
    return(eGenes)
}
memEQTL <- memoise::memoise(get_eqtl_df)

get_pheno_df <- function(){
    phenotype_file = paste0('/ceph/projects/v4_phase3_paper/inputs/',
                            'phenotypes/_m/merged_phenotypes.csv')
    return(data.table::fread(phenotype_file))
}
memPHENO <- memoise::memoise(get_pheno_df)
```

```
get_residualized_df <- function(){
    expr_file = paste0("../../differential_expression/_m/", feature,
                       "/residualized_expression.tsv")
    return(data.table::fread(expr_file) %>% column_to_rownames("V1"))
}
memRES <- memoise::memoise(get_residualized_df)

get_genotypes <- function(){
    traw_file = paste0("/ceph/projects/brainseq/genotype/download/topmed/
 ↪convert2plink/",
                      "filter_maf_01/a_transpose/_m/LIBD_Brain_TopMed.traw")
    traw = data.table::fread(traw_file) %>% rename_with(~ gsub('\\_.*', '', .x))
    return(traw)
}
memSNPs <- memoise::memoise(get_genotypes)

get_gwas_snps <- function(){
    gwas_snp_file = paste0('/ceph/projects/v4_phase3_paper/inputs/sz_gwas/',
                          'pgc2_clozuk/map_phase3/_m/libd_hg38_pgc2sz_snps.tsv')
    gwas_df = data.table::fread(gwas_snp_file) %>% arrange(P)
    return(gwas_df)
}
memGWAS <- memoise::memoise(get_gwas_snps)

get_integration_df <- function(){
    return(inner_join(memGWAS(), memEQTL(),
                      by=c("our_snp_id"="variant_id"),
                      suffix=c("_PGC2", "_eQTL")) %>%
            inner_join(memDE(), by=c("gene_id"="V1")) %>%
            mutate(agree_direction=sign(OR -1) * sign(slope) * sign(t) *␣
 ↪ifelse(pgc2_a1_same_as_our_counted, 1, -1)))
}
memMERGE <- memoise::memoise(get_integration_df)

get_snp_df <- function(variant_id, gene_id){
    zz = get_geno_annot() %>% filter(SNP == variant_id)
    xx = get_snps_df() %>% filter(SNP == variant_id) %>%
        column_to_rownames("SNP") %>% t %>% as.data.frame %>%
        rownames_to_column("BrNum") %>% mutate(COUNTED=zz$COUNTED, ALT=zz$ALT)␣
 ↪%>%
        rename("SNP"=all_of(variant_id))
    yy = memRES()[gene_id, ] %>% t %>% as.data.frame %>%
        rownames_to_column("RNum") %>% inner_join(memPHENO(), by="RNum")
    ## Annotated SNPs
    letters = c()
    for(ii in seq_along(xx$COUNTED)){
        a0 = xx$COUNTED[ii]; a1 = xx$ALT[ii]; number = xx$SNP[ii]
        letters <- append(letters, letter_snp(number, a0, a1))
    }
    xx = xx %>% mutate(LETTER=letters, ID=paste(SNP, LETTER, sep="\n"))
    df = inner_join(xx, yy, by="BrNum") %>% mutate_if(is.character, as.factor)
    return(df)
}
memDF <- memoise::memoise(get_snp_df)
```

1.1.2 Simple functions

[4]:

```
feature_map <- function(feature){
    return(list("genes"="Gene", "transcripts"= "Transcript",
                "exons"= "Exon", "junctions"= "Junction")[[feature]])
}

get_geno_annot <- function(){
    return(memSNPs() %>% select(CHR, SNP, POS, COUNTED, ALT))
}

get_snps_df <- function(){
    return(memSNPs() %>% select("SNP", starts_with("Br")))
}

letter_snp <- function(number, a0, a1){
    if(is.na(number)){ return(NA) }
    if( length(a0) == 1 & length(a1) == 1){
        seps = ""; collapse=""
    } else {
        seps = " "; collapse=NULL
    }
    return(paste(paste0(rep(a0, number), collapse = collapse),
                 paste0(rep(a1, (2-number)), collapse = collapse), sep=seps))
}

save_ggplots <- function(fn, p, w, h){
    for(ext in c('.pdf', '.png', '.svg')){
        ggsave(paste0(fn, ext), plot=p, width=w, height=h)
    }
}

get_biomart_df <- function(){
    biomart = data.table::fread("../_h/biomart.csv")
}
memMART <- memoise::memoise(get_biomart_df)

get_gene_symbol <- function(gene_id){
    ensemblID = gsub("\\..*", "", gene_id)
    geneid = memMART() %>% filter(ensembl_gene_id == gsub("\\..*", "", gene_id))
    if(dim(geneid)[1] == 0){
        return("")
    } else {
        return(geneid$external_gene_name)
    }
}

plot_simple_eqtl <- function(fn, gene_id, variant_id, eqtl_annot){
    bxp = memDF(variant_id, gene_id) %>%
        ggboxplot(x="ID", y=gene_id, fill="red", add="jitter", xlab="",
                  ylab="Residualized Expression", outlier.shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4,
                  ggtheme=theme_pubr(base_size=20, border=TRUE)) +
        font("xy.title", face="bold") +
        ggtitle(paste(get_gene_symbol(gene_id), gene_id, eqtl_annot, sep='\n'))␣
 ↪+
        theme(plot.title = element_text(hjust = 0.5, face="bold"))
    print(bxp)
    save_ggplots(fn, bxp, 7, 7)
}
```

1.1.3 GWAS plots

[5]:

```
get_risk_allele <- function(OR, A1, A2){
    ra = ifelse(OR > 1, A1, A2)
    return(ra)
}

get_df <- function(){
    return(memEQTL() %>% inner_join(memGWAS(), by="variant_id"))
}

get_gwas_ordered_snp_df <- function(variant_id, gene_id,␣
 ↪pgc2_a1_same_as_our_counted, OR){
    df = memDF(variant_id, gene_id)
    if(!pgc2_a1_same_as_our_counted){ # Fix bug with matching alleles!
        if(OR < 1){ df = df %>% mutate(SNP = 2-SNP, ID=paste(SNP, LETTER,␣
 ↪sep="\n")) }
    } else {
        if(OR > 1){ df = df %>% mutate(SNP = 2-SNP, ID=paste(SNP, LETTER,␣
 ↪sep="\n")) }
    }
    return(df)
}
```

```
plot_gwas_eqtl_pheno <- function(fn, gene_id, variant_id,␣
 ↪pgc2_a1_same_as_our_counted, OR, title){
    bxp = get_gwas_ordered_snp_df(variant_id, gene_id,␣
 ↪pgc2_a1_same_as_our_counted, OR) %>%
        mutate_if(is.character, as.factor) %>% filter(Dx %in% c("CTL", "SZ"),␣
 ↪Age > 17) %>%
        ggboxplot(x="ID", y=gene_id, fill="Dx", color="Dx", add="jitter",␣
 ↪xlab=variant_id,
                  ylab="Residualized Expression", outlier.shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4, legend="bottom",
                  ggtheme=theme_pubr(base_size=20, border=TRUE)) +
        font("xy.title", face="bold") + ggtitle(title) +
        theme(plot.title = element_text(hjust = 0.5, face="bold"))
    print(bxp)
    save_ggplots(fn, bxp, 7, 9)
}
```

1.2 Integration analysis

[6]:

```
dir.create(feature)
```

1.2.1 Enrichment

Integrate DEG with PGC2+CLOZUK SNPs

[7]:

```
dft = memMERGE() %>% mutate(agree_direction=ifelse(agree_direction == 1, "Yes",␣
 ↪ifelse(agree_direction == -1, "No", 0)))
dim(dft)
```

1. 2280801 2. 65

[8]:

```
table(dft$agree_direction)
```

```

      0      No     Yes 
   2377 1122054 1156370 
```

[9]:

```
table = matrix(c(sum((dft$P<5e-8)  & (dft$adj.P.Val < 0.05)),
                 sum((dft$P>=5e-8) & (dft$adj.P.Val < 0.05)),
                 sum((dft$P<5e-8)  & (dft$adj.P.Val >= 0.05)),
                 sum((dft$P>=5e-8) & (dft$adj.P.Val >= 0.05))),
               nrow=2)
print(table)
fisher.test(table)
```

```
      [,1]    [,2]
[1,]   989   54364
[2,] 48243 2177205
```

```
        Fisher's Exact Test for Count Data

data:  table
p-value = 4.08e-10
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.7696986 0.8749202
sample estimates:
odds ratio
 0.8210113
```

[10]:

```
dft2 = dft %>% filter(P <= 5e-8, `adj.P.Val` < 0.05) %>%
    mutate(eqtl_gwas_dir=sign(OR -1) * sign(slope) *␣
 ↪ifelse(pgc2_a1_same_as_our_counted, 1, -1),
           de_dir=sign(t), eqtl_slope=sign(OR␣
 ↪-1)*sign(slope)*ifelse(pgc2_a1_same_as_our_counted, 1, -1)) %>%
    #rowwise() %>% mutate(risk_allele=get_risk_allele(our_snp_id)) %>%
    select(gene_id, gene_name, our_snp_id, rsid, A1, A2, OR, P, pval_nominal,␣
 ↪adj.P.Val, logFC,
           t, eqtl_slope, de_dir, eqtl_gwas_dir, agree_direction,␣
 ↪pgc2_a1_same_as_our_counted) %>%
    rename("variant_id"="our_snp_id", "Symbol"="gene_name") %>%␣
 ↪mutate_all(list(~na_if(.,""))) %>%
    mutate(Symbol = coalesce(Symbol,gene_id))
dft2 %>% data.table::fwrite(paste0(feature, "/integration_by_symbol.txt"),␣
 ↪sep='\t')
dim(dft2)
```

1. 989 2. 17

[11]:

```
df = dft2 %>% group_by(gene_id) %>% slice(1) %>% arrange(P)
table(df$agree_direction)
```

```
 No Yes
  3   6
```

[12]:

```
df
```

A grouped_df: 9 × 17

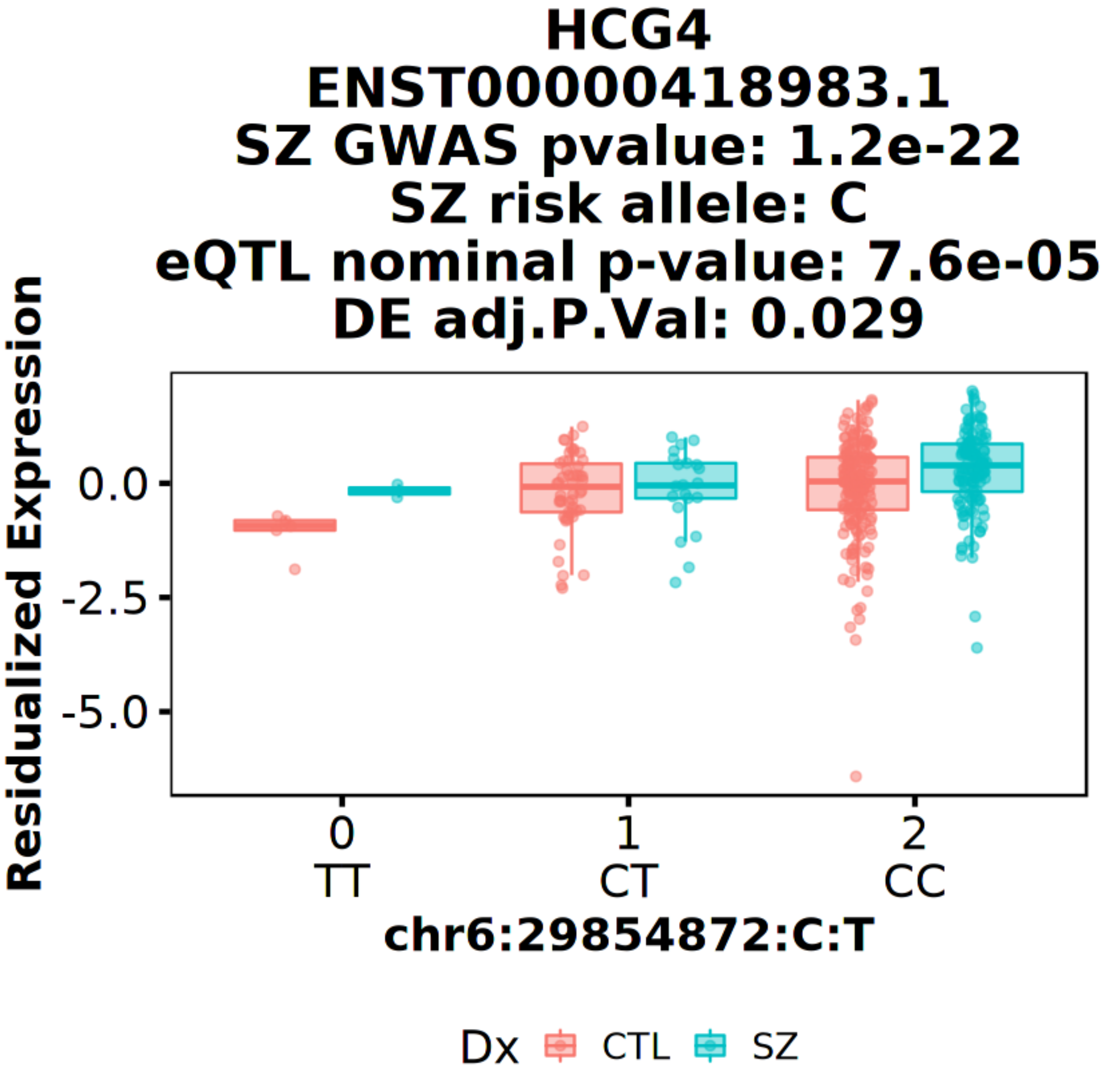
gene_id <chr>	Symbol <chr>	variant_id <chr>	rsid <chr>	A1 <chr>	A2 <chr>
ENST00000418983.1	HCG4	chr6:29854872:C:T	rs2517857	C	T
ENST00000617168.4	ZSCAN26	chr6:28600751:A:G	rs418914	A	G
ENST00000369878.8	CNNM2	chr10:102852578:T:A	rs11191419	T	A
ENST00000244576.8	ZNF391	chr6:27707511:A:T	rs1139226	A	T
ENST00000411553.2	HCG11	chr6:26466161:G:A	rs1977199	G	A
ENST00000378486.7	PLCH2	chr1:2455662:C:T	rs4648845	C	T
ENST00000293756.4	IP6K3	chr6:33773939:A:G	rs4711350	A	G
ENST00000361204.8	SREBF2	chr22:41885425:A:G	rs1052717	A	G
ENST00000344099.3	ZNF14	chr19:19633270:T:C	rs11878202	T	C

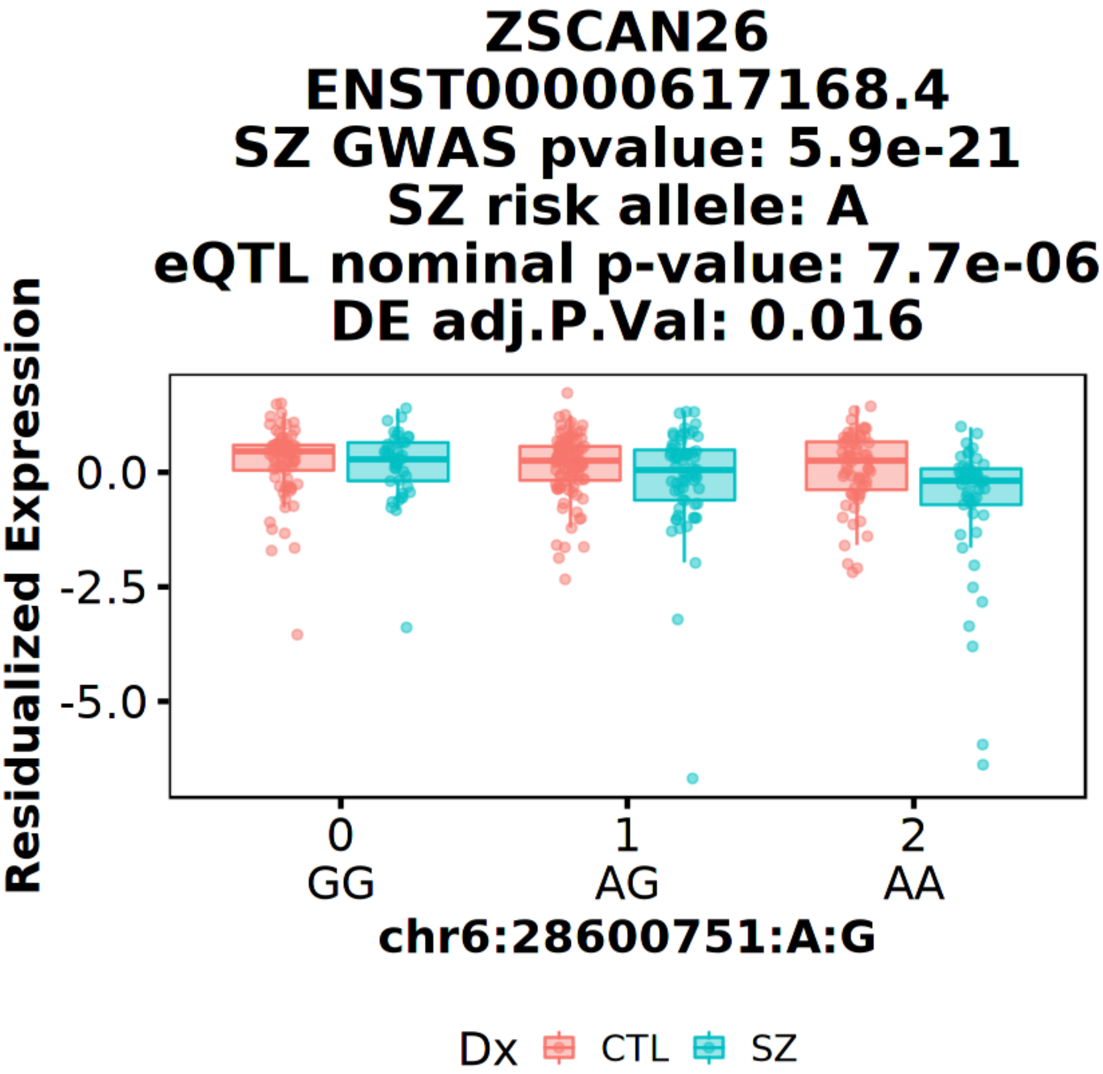
1.2.2 Plot with PGC2 risk allele

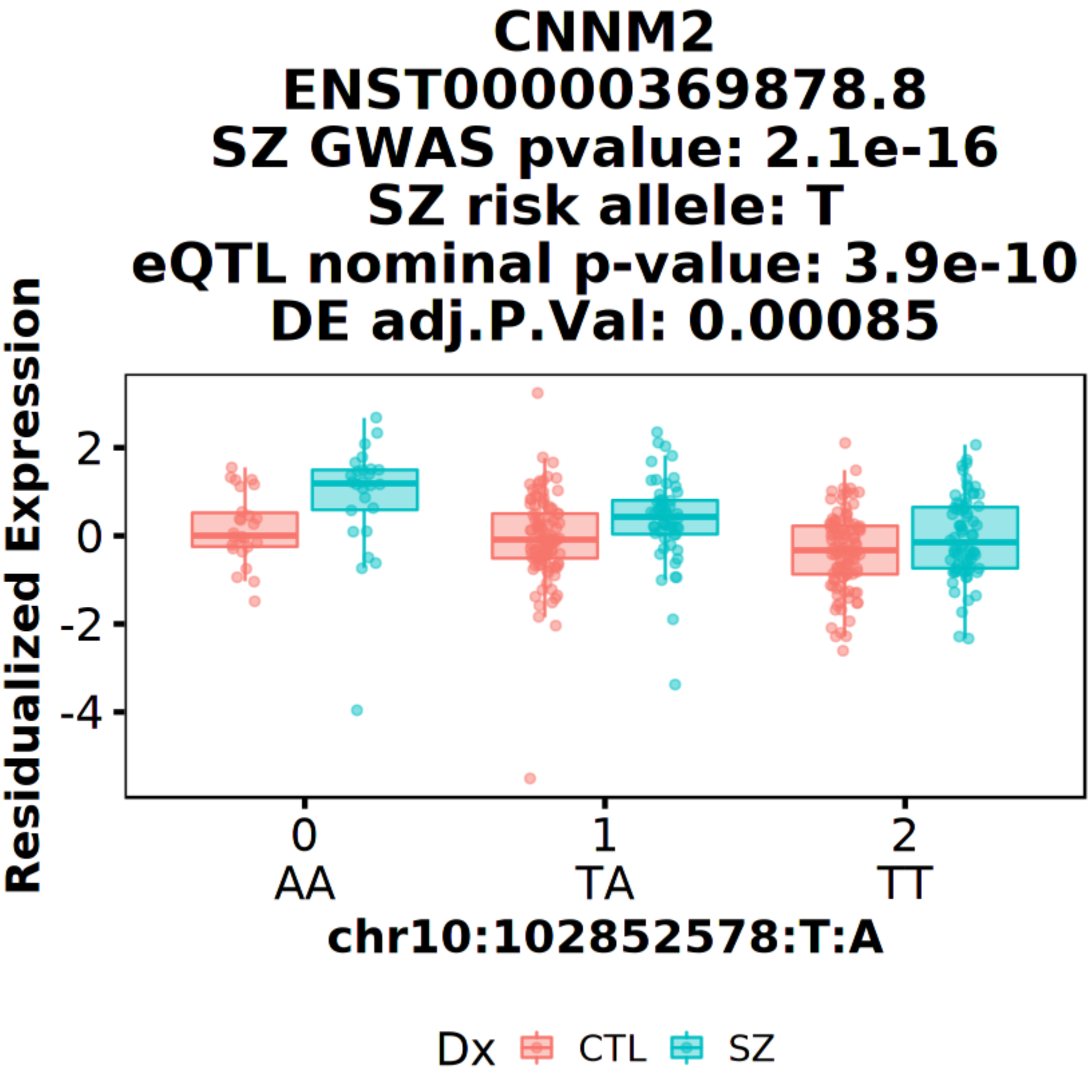
[13]:

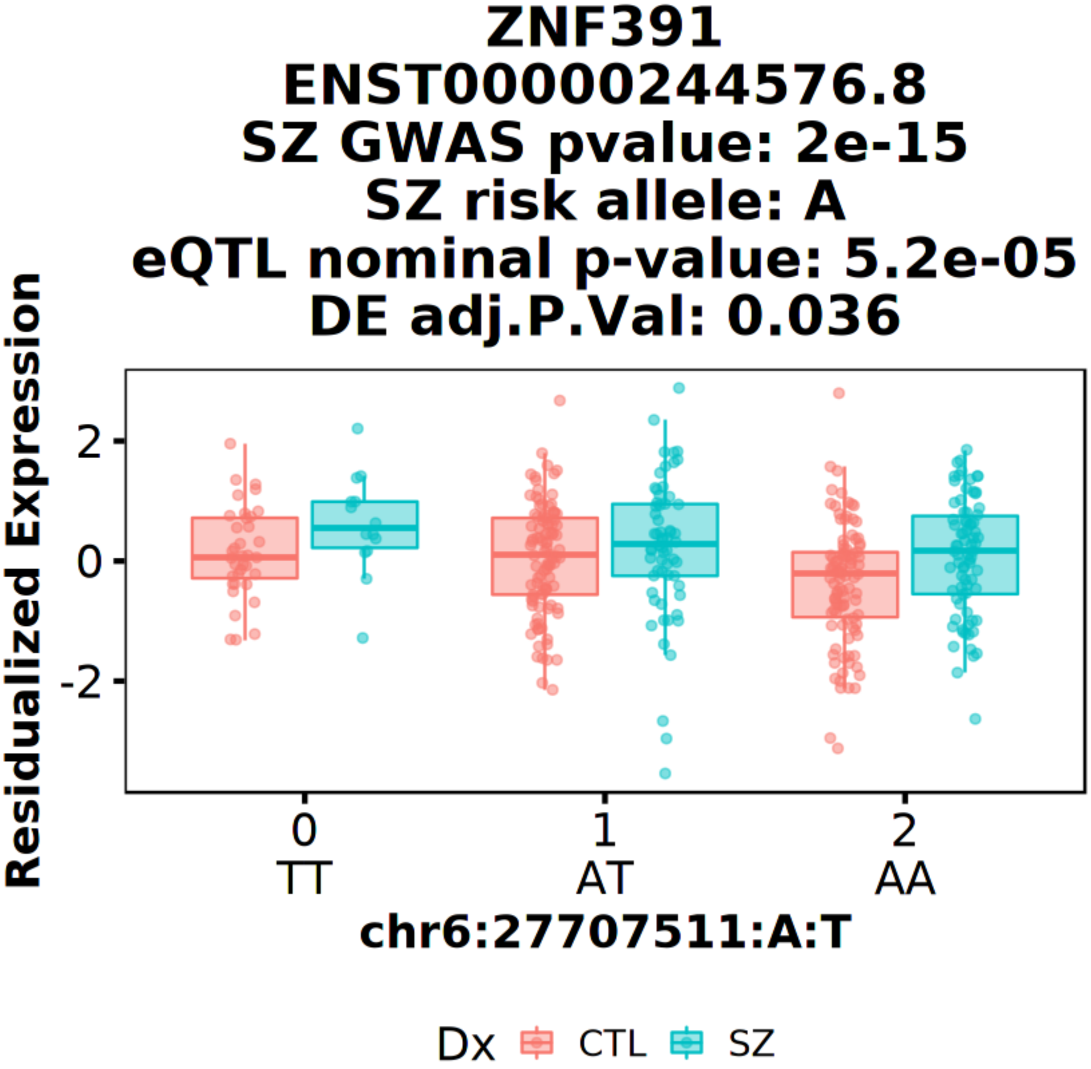
```
for(num in seq_along(df$gene_id)){
    variant_id = df$variant_id[num]
    gene_id = df$gene_id[num]
    gene_name = df$Symbol[num]
    pgc2_a1_same_as_our_counted = df$pgc2_a1_same_as_our_counted[num]
    OR = df$OR[num]; A1 = df$A1[num]; A2 = df$A2[num]
    fn = paste0(feature, "/eqtl_gwas_", gsub("\\.", "_", gene_name))
    de_annot = paste('DE adj.P.Val:', signif(df$adj.P.Val[num], 2))
    eqtl_annot = paste("eQTL nominal p-value:", signif(df$pval_nominal[num], 2))
    gwas_annot = paste("SZ GWAS pvalue:", signif(df$P[num], 2))
    risk_annot = paste("SZ risk allele:", get_risk_allele(OR, A1, A2))
    title = paste(gene_name, gene_id, gwas_annot,
                  risk_annot, eqtl_annot, de_annot, sep='\n')
    plot_gwas_eqtl_pheno(fn, gene_id, variant_id, pgc2_a1_same_as_our_counted, OR, title)
    #print(title)
}
```

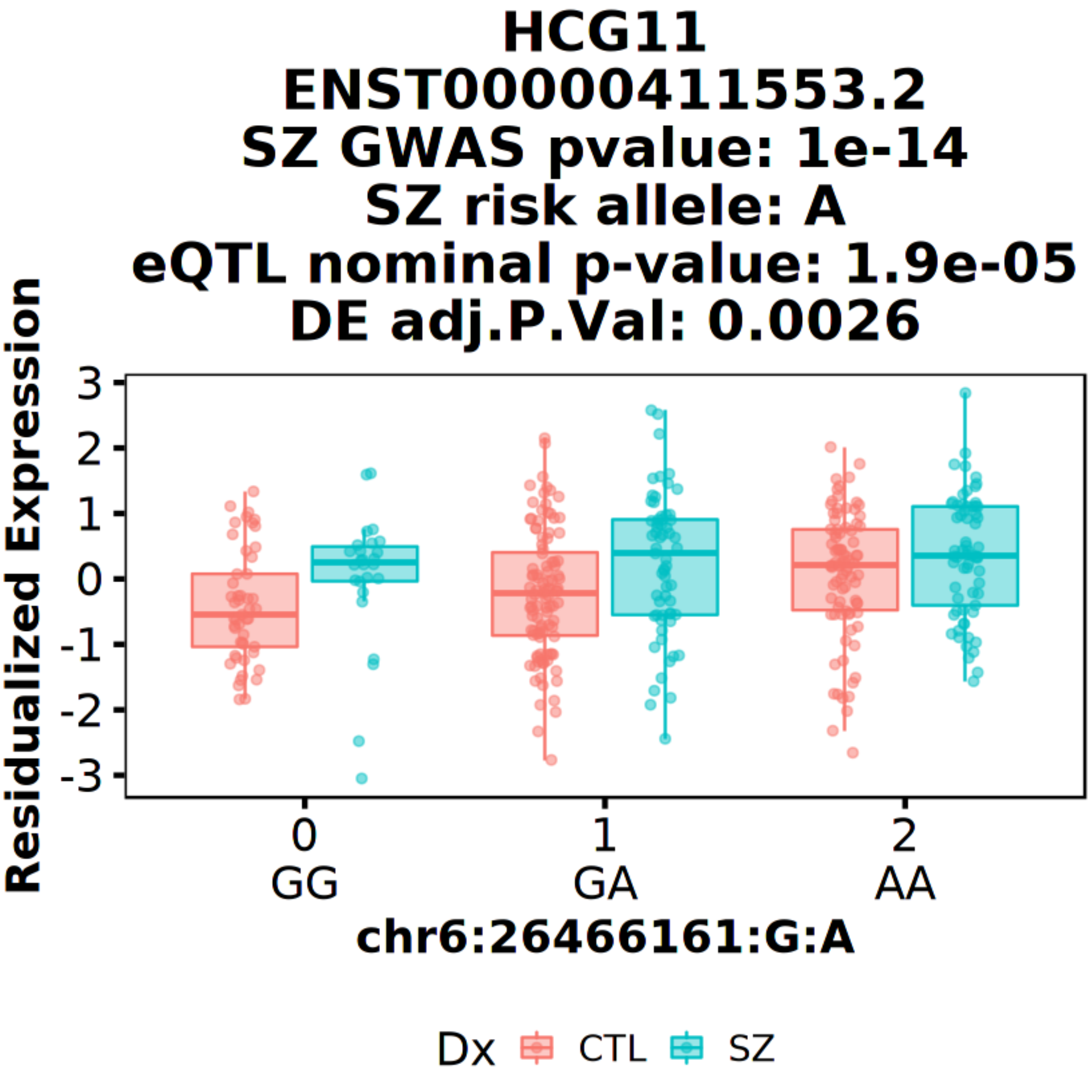
```
Warning message in data.table::fread(expr_file):
"Detected 393 column names but the data has 394 columns (i.e. invalid file).
Added 1 extra default column name for the first column which is guessed to be
row names or an index. Use setnames() afterwards if this guess is not correct,
or fix the file write command that created the file to create a valid file."
```

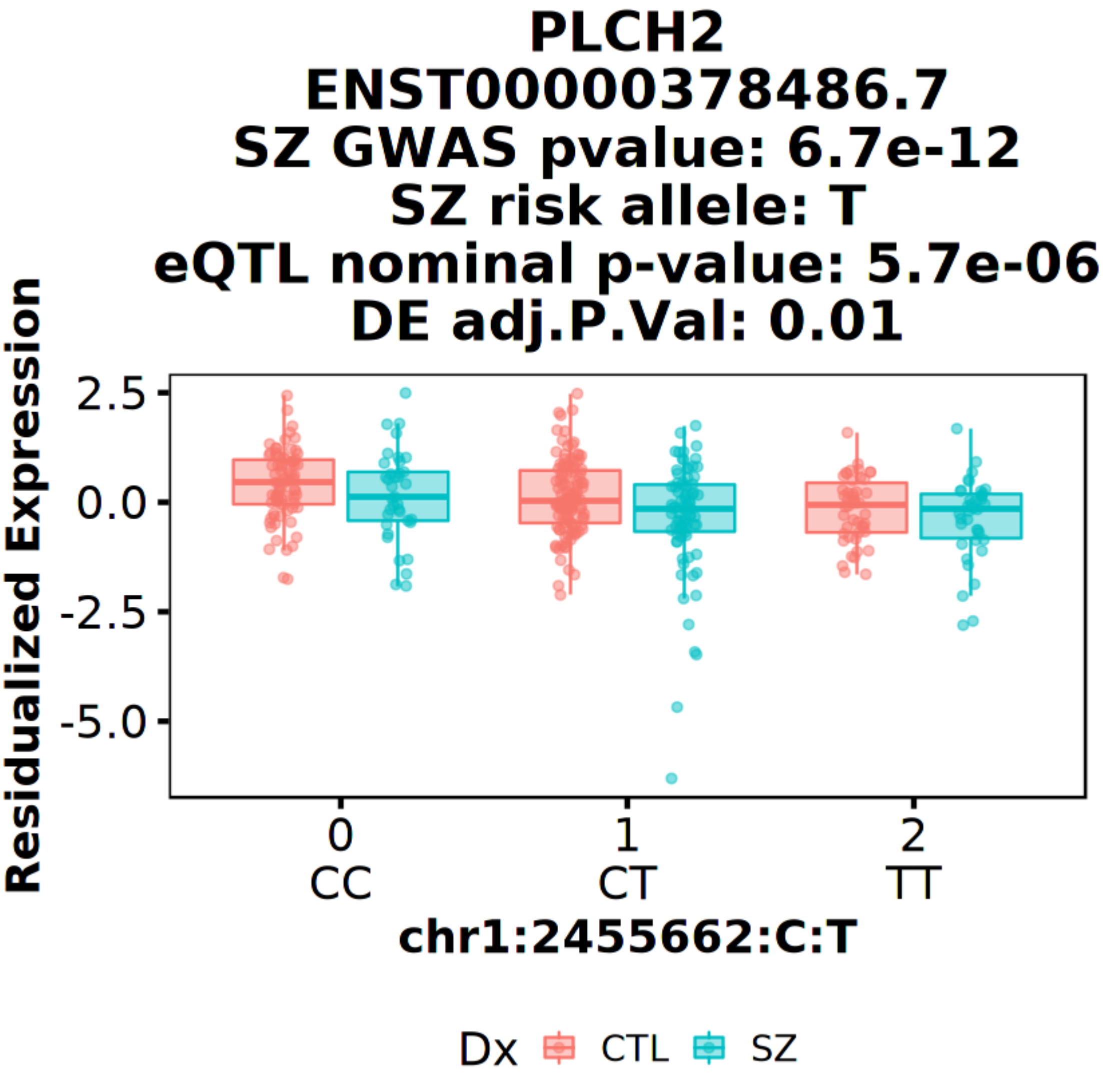

HCG4
ENST00000418983.1
SZ GWAS pvalue: 1.2e-22
SZ risk allele: C
eQTL nominal p-value: 7.6e-05
DE adj.P.Val: 0.029
Residualized Expression
0.0
-2.5
-5.0
0
TT
1
CT
2
CC
chr6:29854872:C:T
Dx
CTL
SZ

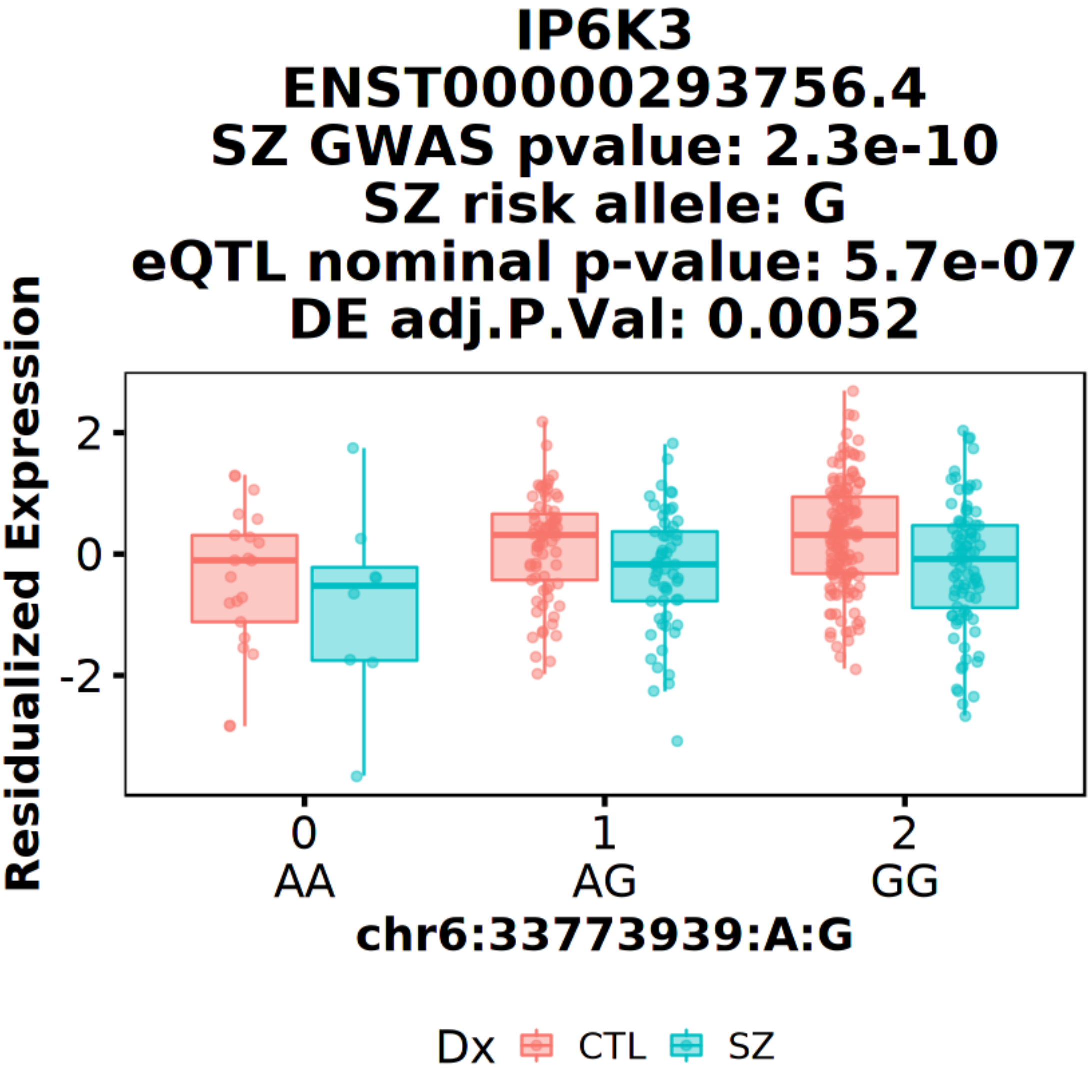

ZSCAN26
ENST00000617168.4
SZ GWAS pvalue: 5.9e-21
SZ risk allele: A
eQTL nominal p-value: 7.7e-06
DE adj.P.Val: 0.016
Residualized Expression
0.0
-2.5
-5.0
0
GG
1
AG
2
AA
chr6:28600751:A:G
Dx
CTL
SZ

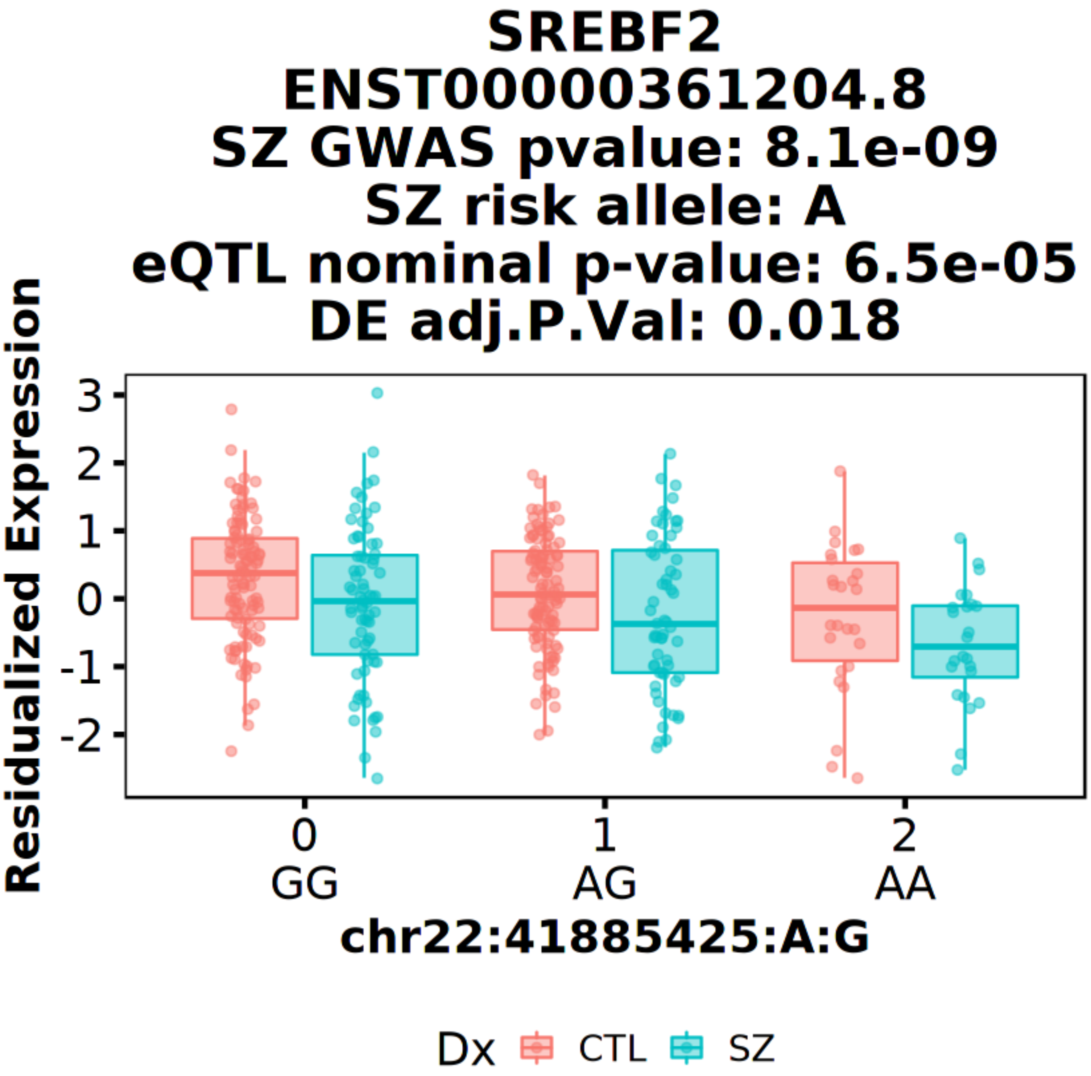

CNNM2
ENST00000369878.8
SZ GWAS pvalue: 2.1e-16
SZ risk allele: T
eQTL nominal p-value: 3.9e-10
DE adj.P.Val: 0.00085
Residualized Expression
2
0
-2
-4
0
AA
1
TA
2
TT
chr10:102852578:T:A
Dx
CTL
SZ

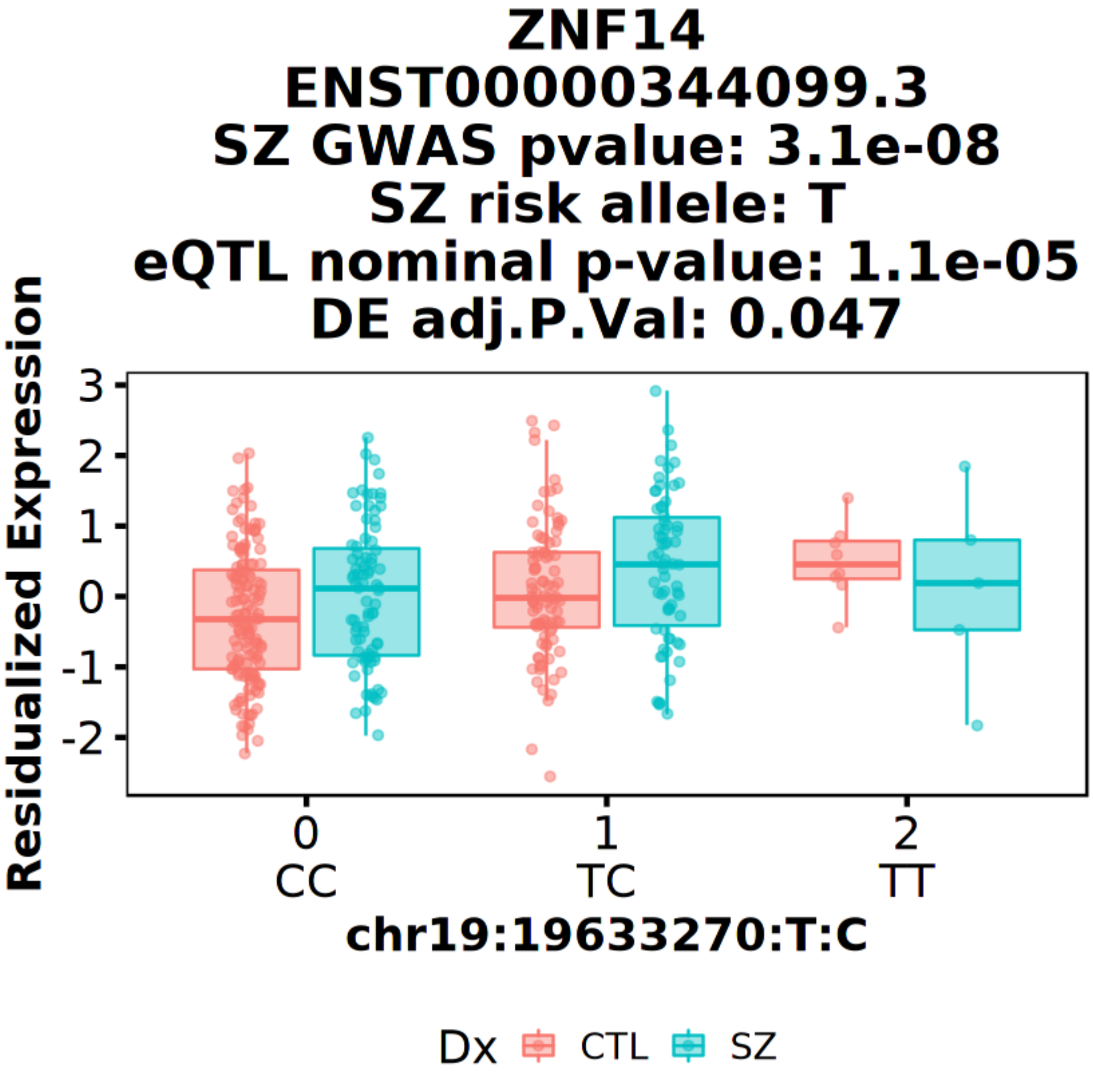

ZNF391
ENST00000244576.8
SZ GWAS pvalue: 2e-15
SZ risk allele: A
eQTL nominal p-value: 5.2e-05
DE adj.P.Val: 0.036
Residualized Expression
2
0
-2
0
TT
1
AT
2
AA
chr6:27707511:A:T
Dx
CTL
SZ


HCG11
ENST00000411553.2
SZ GWAS pvalue: 1e-14
SZ risk allele: A
eQTL nominal p-value: 1.9e-05
DE adj.P.Val: 0.0026
Residualized Expression
3
2
1
0
-1
-2
-3
0
GG
1
GA
2
AA
chr6:26466161:G:A
Dx
CTL
SZ


PLCH2
ENST00000378486.7
SZ GWAS pvalue: 6.7e-12
SZ risk allele: T
eQTL nominal p-value: 5.7e-06
DE adj.P.Val: 0.01
Residualized Expression
2.5
0.0
-2.5
-5.0
0
CC
1
CT
2
TT
chr1:2455662:C:T
Dx
CTL
SZ


IP6K3
ENST00000293756.4
SZ GWAS pvalue: 2.3e-10
SZ risk allele: G
eQTL nominal p-value: 5.7e-07
DE adj.P.Val: 0.0052
Residualized Expression
2
0
-2
0
AA
1
AG
2
GG
chr6:33773939:A:G
Dx
CTL
SZ


SREBF2
ENST00000361204.8
SZ GWAS pvalue: 8.1e-09
SZ risk allele: A
eQTL nominal p-value: 6.5e-05
DE adj.P.Val: 0.018
Residualized Expression
3
2
1
0
-1
-2
0
GG
1
AG
2
AA
chr22:41885425:A:G
Dx
CTL
SZ



1.3 Session Info

```
[14]: Sys.time()
      proc.time()
      options(width = 120)
      sessioninfo::session_info()
```

```
[1] "2021-09-08 11:10:47 EDT"

    user   system  elapsed
3793.807 2536.544 1051.719

 Session info
setting  value
```

```
version  R version 4.0.3 (2020-10-10)
os       Arch Linux
system   x86_64, linux-gnu
ui       X11
language (EN)
collate  en_US.UTF-8
ctype    en_US.UTF-8
tz       America/New_York
date     2021-09-08

 Packages
package     * version  date       lib source
abind         1.4-5    2016-07-21 [1] CRAN (R 4.0.2)
assertthat    0.2.1    2019-03-21 [1] CRAN (R 4.0.2)
backports     1.2.1    2020-12-09 [1] CRAN (R 4.0.2)
base64enc     0.1-3    2015-07-28 [1] CRAN (R 4.0.2)
broom         0.7.9    2021-07-27 [1] CRAN (R 4.0.3)
cachem        1.0.6    2021-08-19 [1] CRAN (R 4.0.3)
Cairo         1.5-12.2 2020-07-07 [1] CRAN (R 4.0.2)
car           3.0-11   2021-06-27 [1] CRAN (R 4.0.3)
carData       3.0-4    2020-05-22 [1] CRAN (R 4.0.2)
cellranger    1.1.0    2016-07-27 [1] CRAN (R 4.0.2)
cli           3.0.1    2021-07-17 [1] CRAN (R 4.0.3)
colorspace    2.0-2    2021-06-24 [1] CRAN (R 4.0.3)
crayon        1.4.1    2021-02-08 [1] CRAN (R 4.0.3)
curl          4.3.2    2021-06-23 [1] CRAN (R 4.0.3)
data.table    1.14.0   2021-02-21 [1] CRAN (R 4.0.3)
DBI           1.1.1    2021-01-15 [1] CRAN (R 4.0.2)
dbplyr        2.1.1    2021-04-06 [1] CRAN (R 4.0.3)
digest        0.6.27   2020-10-24 [1] CRAN (R 4.0.2)
dplyr       * 1.0.7    2021-06-18 [1] CRAN (R 4.0.3)
ellipsis      0.3.2    2021-04-29 [1] CRAN (R 4.0.3)
evaluate      0.14     2019-05-28 [1] CRAN (R 4.0.2)
fansi         0.5.0    2021-05-25 [1] CRAN (R 4.0.3)
farver        2.1.0    2021-02-28 [1] CRAN (R 4.0.3)
fastmap       1.1.0    2021-01-25 [1] CRAN (R 4.0.2)
forcats     * 0.5.1    2021-01-27 [1] CRAN (R 4.0.2)
foreign       0.8-80   2020-05-24 [2] CRAN (R 4.0.3)
fs            1.5.0    2020-07-31 [1] CRAN (R 4.0.2)
generics      0.1.0    2020-10-31 [1] CRAN (R 4.0.2)
ggplot2     * 3.3.5    2021-06-25 [1] CRAN (R 4.0.3)
ggpubr      * 0.4.0    2020-06-27 [1] CRAN (R 4.0.2)
ggsignif      0.6.2    2021-06-14 [1] CRAN (R 4.0.3)
glue          1.4.2    2020-08-27 [1] CRAN (R 4.0.2)
gtable        0.3.0    2019-03-25 [1] CRAN (R 4.0.2)
haven         2.4.3    2021-08-04 [1] CRAN (R 4.0.3)
hms           1.1.0    2021-05-17 [1] CRAN (R 4.0.3)
htmltools     0.5.2    2021-08-25 [1] CRAN (R 4.0.3)
```

```
httr           1.4.2    2020-07-20 [1] CRAN (R 4.0.2)
IRdisplay      1.0      2021-01-20 [1] CRAN (R 4.0.2)
IRkernel       1.2      2021-05-11 [1] CRAN (R 4.0.3)
jsonlite       1.7.2    2020-12-09 [1] CRAN (R 4.0.2)
labeling       0.4.2    2020-10-20 [1] CRAN (R 4.0.2)
lifecycle      1.0.0    2021-02-15 [1] CRAN (R 4.0.3)
lubridate      1.7.10   2021-02-26 [1] CRAN (R 4.0.3)
magrittr       2.0.1    2020-11-17 [1] CRAN (R 4.0.2)
memoise        2.0.0    2021-01-26 [1] CRAN (R 4.0.2)
modelr         0.1.8    2020-05-19 [1] CRAN (R 4.0.2)
munsell        0.5.0    2018-06-12 [1] CRAN (R 4.0.2)
openxlsx       4.2.4    2021-06-16 [1] CRAN (R 4.0.3)
pbdZMQ         0.3-5    2021-02-10 [1] CRAN (R 4.0.3)
pillar         1.6.2    2021-07-29 [1] CRAN (R 4.0.3)
pkgconfig      2.0.3    2019-09-22 [1] CRAN (R 4.0.2)
purrr        * 0.3.4    2020-04-17 [1] CRAN (R 4.0.2)
R.methodsS3    1.8.1    2020-08-26 [1] CRAN (R 4.0.3)
R.oo           1.24.0   2020-08-26 [1] CRAN (R 4.0.3)
R.utils        2.10.1   2020-08-26 [1] CRAN (R 4.0.3)
R6             2.5.1    2021-08-19 [1] CRAN (R 4.0.3)
Rcpp           1.0.7    2021-07-07 [1] CRAN (R 4.0.3)
readr        * 2.0.1    2021-08-10 [1] CRAN (R 4.0.3)
readxl         1.3.1    2019-03-13 [1] CRAN (R 4.0.2)
repr           1.1.3    2021-01-21 [1] CRAN (R 4.0.2)
reprex         2.0.1    2021-08-05 [1] CRAN (R 4.0.3)
rio            0.5.27   2021-06-21 [1] CRAN (R 4.0.3)
rlang          0.4.11   2021-04-30 [1] CRAN (R 4.0.3)
rstatix        0.7.0    2021-02-13 [1] CRAN (R 4.0.3)
rstudioapi     0.13     2020-11-12 [1] CRAN (R 4.0.2)
rvest          1.0.1    2021-07-26 [1] CRAN (R 4.0.3)
scales         1.1.1    2020-05-11 [1] CRAN (R 4.0.2)
sessioninfo    1.1.1    2018-11-05 [1] CRAN (R 4.0.2)
stringi        1.7.4    2021-08-25 [1] CRAN (R 4.0.3)
stringr      * 1.4.0    2019-02-10 [1] CRAN (R 4.0.2)
svglite        2.0.0    2021-02-20 [1] CRAN (R 4.0.3)
systemfonts    1.0.2    2021-05-11 [1] CRAN (R 4.0.3)
tibble       * 3.1.4    2021-08-25 [1] CRAN (R 4.0.3)
tidyr        * 1.1.3    2021-03-03 [1] CRAN (R 4.0.3)
tidyselect     1.1.1    2021-04-30 [1] CRAN (R 4.0.3)
tidyverse    * 1.3.1    2021-04-15 [1] CRAN (R 4.0.3)
tzdb           0.1.2    2021-07-20 [1] CRAN (R 4.0.3)
utf8           1.2.2    2021-07-24 [1] CRAN (R 4.0.3)
uuid           0.1-4    2020-02-26 [1] CRAN (R 4.0.2)
vctrs          0.3.8    2021-04-29 [1] CRAN (R 4.0.3)
withr          2.4.2    2021-04-18 [1] CRAN (R 4.0.3)
xml2           1.3.2    2020-04-23 [1] CRAN (R 4.0.2)
zip            2.2.0    2021-05-31 [1] CRAN (R 4.0.3)
```

```
[1] /home/jbenja13/R/x86_64-pc-linux-gnu-library/4.0
[2] /usr/lib/R/library
```