main_r

March 8, 2022

1 eQTL boxplot

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

[1]:
```
suppressPackageStartupMessages({
    library(tidyverse)
    library(ggpubr)
})
```

1.1 Functions

[2]:
```
feature = "genes"
```

1.1.1 Cached functions

[3]:
```
get_residualized_df <- function(){
    expr_file = "../../_m/genes_residualized_expression.csv"
    return(data.table::fread(expr_file) %>% column_to_rownames("gene_id"))
}
memRES <- memoise::memoise(get_residualized_df)

get_biomart_df <- function(){
    biomart = data.table::fread("../_h/biomart.csv")
}
memMART <- memoise::memoise(get_biomart_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_caudate_eqtls <- function(){
    mashr_file = "../../../mashr/summary_table/_m/BrainSeq_caudateSpecific_eQTL.
 ↪txt.gz"
    return(data.table::fread(mashr_file) %>%
           filter(Type == feature_map(feature)))
}
memCAUDATE <- memoise::memoise(get_caudate_eqtls)

get_eqtl_df <- function(){
    fastqtl_file = paste0("../../../mashr/_m/", feature, "/
 ↪lfsr_allpairs_3tissues.txt.gz")
    eqtl_df = data.table::fread(fastqtl_file) %>%
        filter(gene_id %in% memCAUDATE()$gene_id)
    return(eqtl_df)
}
memEQTL <- memoise::memoise(get_eqtl_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)
```

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))
}
```

```r
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)

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_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="Region", color="Region",␣
 ↪add="jitter",
                  xlab=variant_id, ylab="Residualized Expression", outlier.
 ↪shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4, legend="bottom",
                  palette="npg", 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_gwas_snps <- function(){
    gwas_snp_file = paste0('/ceph/projects/v4_phase3_paper/inputs/sz_gwas/pgc3/
 ↪',
                          'map_phase3/_m/libd_hg38_pgc2sz_snps_p5e_minus8.tsv')
    gwas_df = data.table::fread(gwas_snp_file) %>% arrange(P)
    return(gwas_df)
}
memGWAS <- memoise::memoise(get_gwas_snps)

get_gwas_snp <- function(variant){
    return(memGWAS() %>% filter(our_snp_id == variant))
}

get_risk_allele <- function(variant){
    gwas_snp = get_gwas_snp(variant)
    if(gwas_snp$OR > 1){
        ra = gwas_snp$A1
    }else{
        ra = gwas_snp$A2
    }
    return(ra)
}

get_eqtl_gwas_df <- function(){
    return(memCAUDATE() %>% inner_join(memGWAS(),␣
 ↪by=c("variant_id"="our_snp_id")))
}

get_gwas_ordered_snp_df <- function(variant_id, gene_id,␣
 ↪pgc3_a1_same_as_our_counted, OR){
    df = memDF(variant_id, gene_id)
    if(!pgc3_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 <- function(fn, gene_id, variant_id, eqtl_annot,
                           pgc2_a1_same_as_our_counted, OR, title){
    dt = get_gwas_ordered_snp_df(variant_id, gene_id,␣
 ↪pgc2_a1_same_as_our_counted, OR)
    y0 = quantile(dt[[gene_id]], probs=c(0.05))[[1]] - 0.26
    y1 = quantile(dt[[gene_id]], probs=c(0.95))[[1]] + 0.26
    bxp = dt %>% mutate_if(is.character, as.factor) %>%
        ggboxplot(x="ID", y=gene_id, fill="Region", color="Region",␣
 ↪add="jitter",
                  xlab=variant_id, ylab="Residualized Expression", outlier.
 ↪shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4, legend="bottom",␣
 ↪lims=c(y0,y1),
                  palette="npg", 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, 8)
}
```

1.2 Plot eQTL

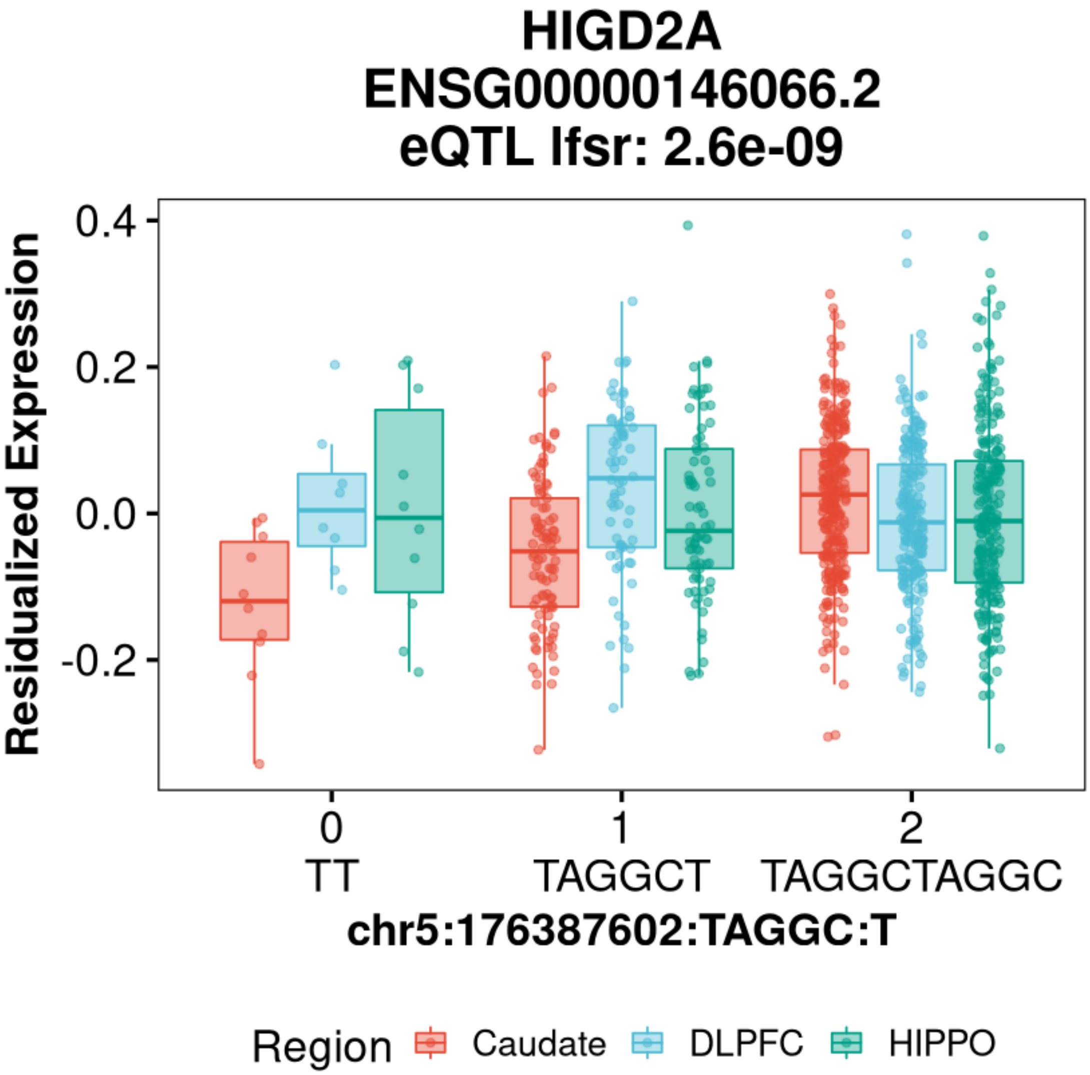
[6]:
```
eGenes <- memCAUDATE() %>% arrange(Caudate) %>% group_by(gene_id) %>% slice(1)␣
 ↪%>% arrange(Caudate)
eGenes %>% head(5)
```

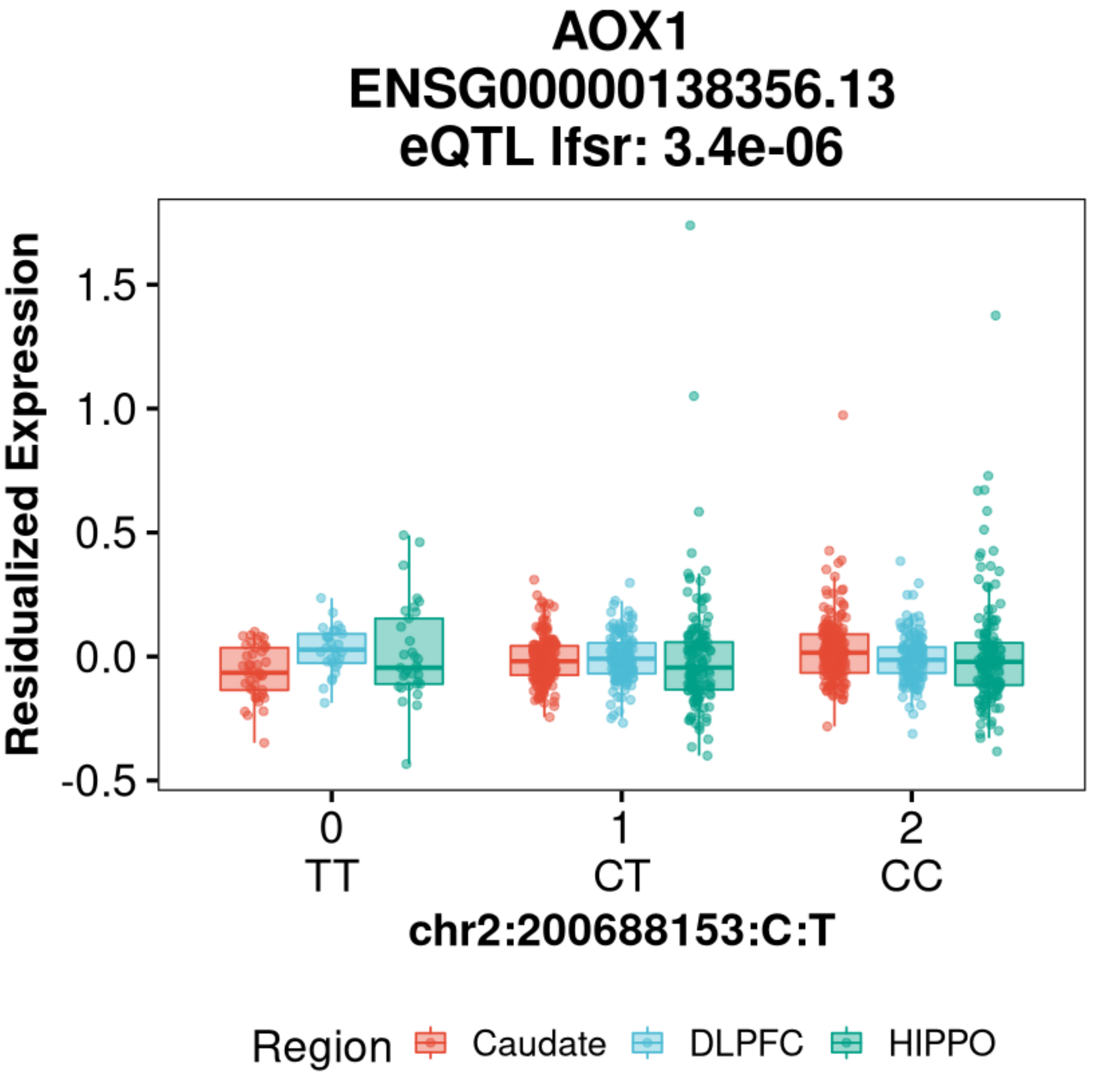
A grouped_df: 5 × 9

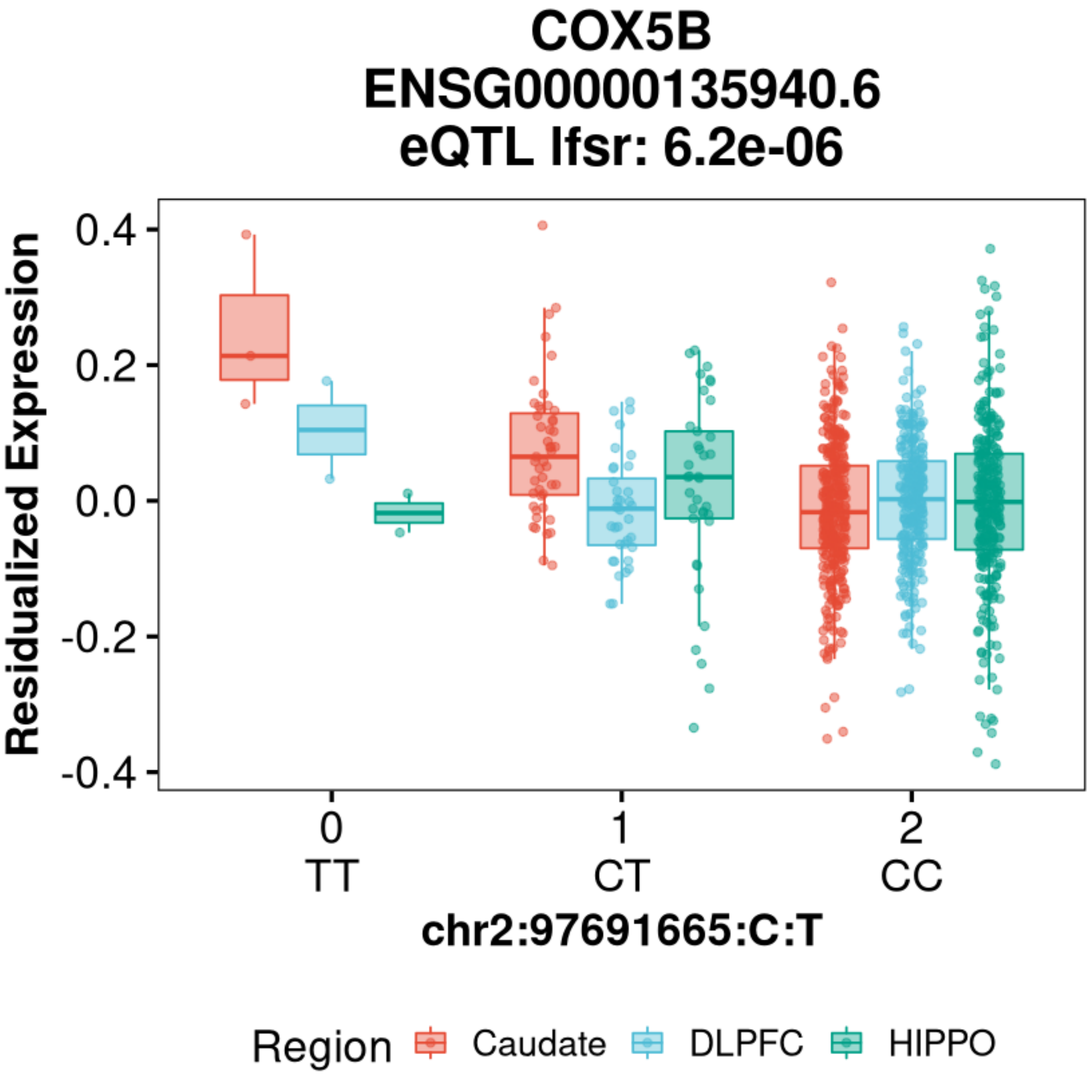
effect	gene_id	variant_id
<chr>	<chr>	<chr>
ENSG00000146066.2_chr5:176387602:TAGGC:T	ENSG00000146066.2	chr5:176387602
ENSG00000138356.13_chr2:200688153:C:T	ENSG00000138356.13	chr2:200688153
ENSG00000135940.6_chr2:97691665:C:T	ENSG00000135940.6	chr2:97691665:
ENSG00000171189.17_chr21:30089992:A:G	ENSG00000171189.17	chr21:30089992
ENSG00000154640.14_chr21:17674435:T:C	ENSG00000154640.14	chr21:17674435

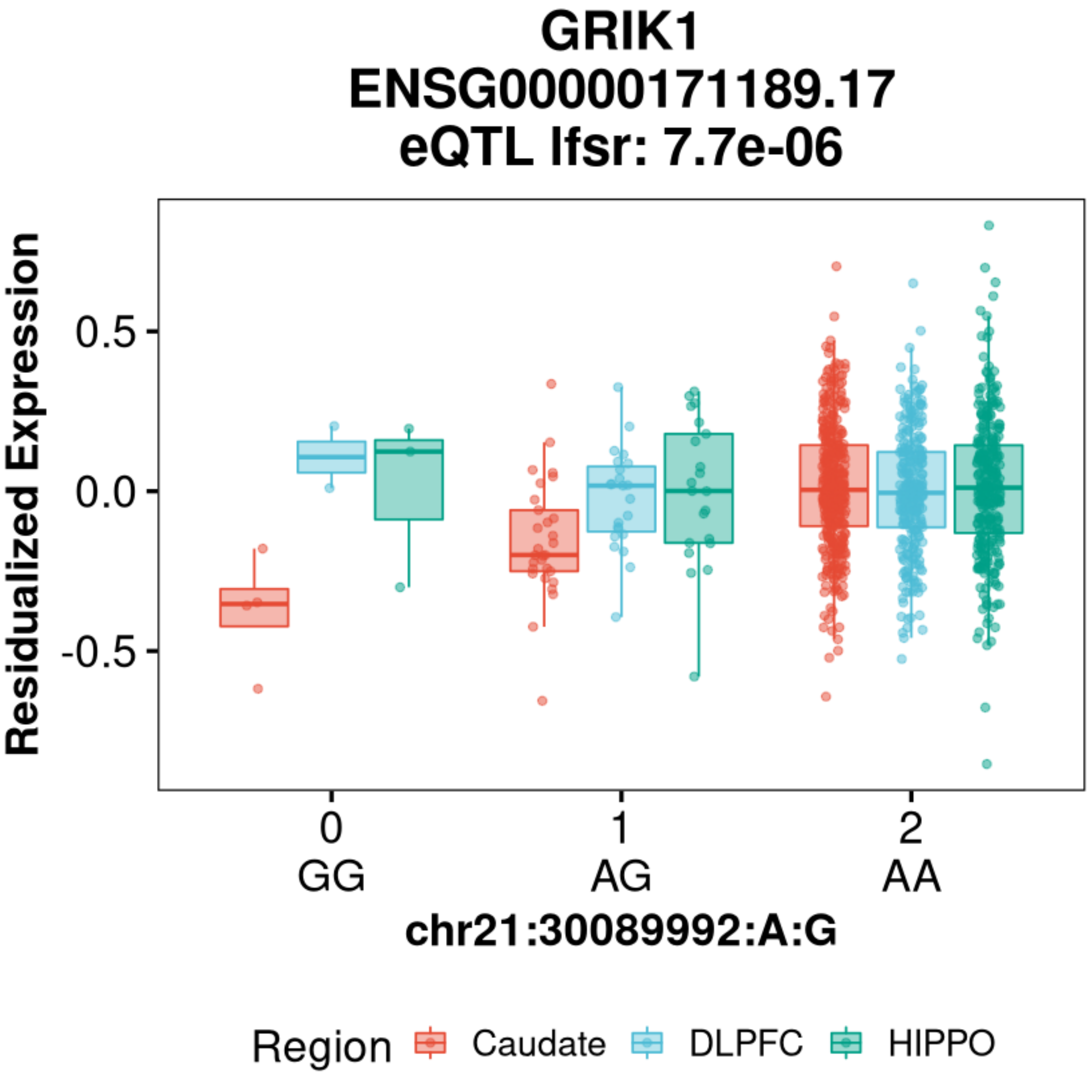
1.2.1 Top 5 eQTLs

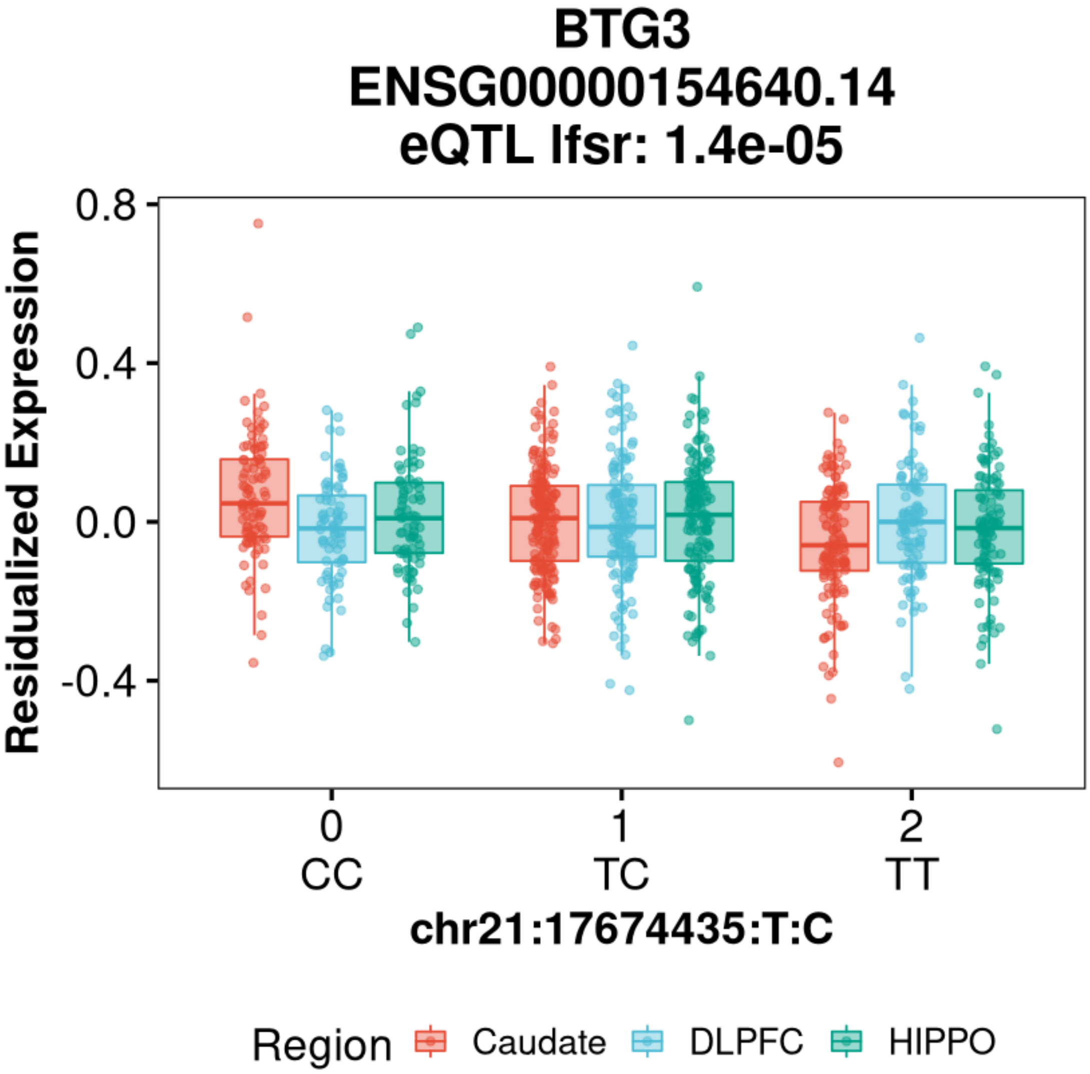
[7]:
```
for(num in 1:10){
    variant_id = eGenes$variant_id[num]
    gene_id = eGenes$gene_id[num]
    eqtl_annot = paste("eQTL lfsr:", signif(eGenes$Caudate[num], 2))
    fn = paste0("top_",num,"_eqtl")
    plot_simple_eqtl(fn, gene_id, variant_id, eqtl_annot)
}
```

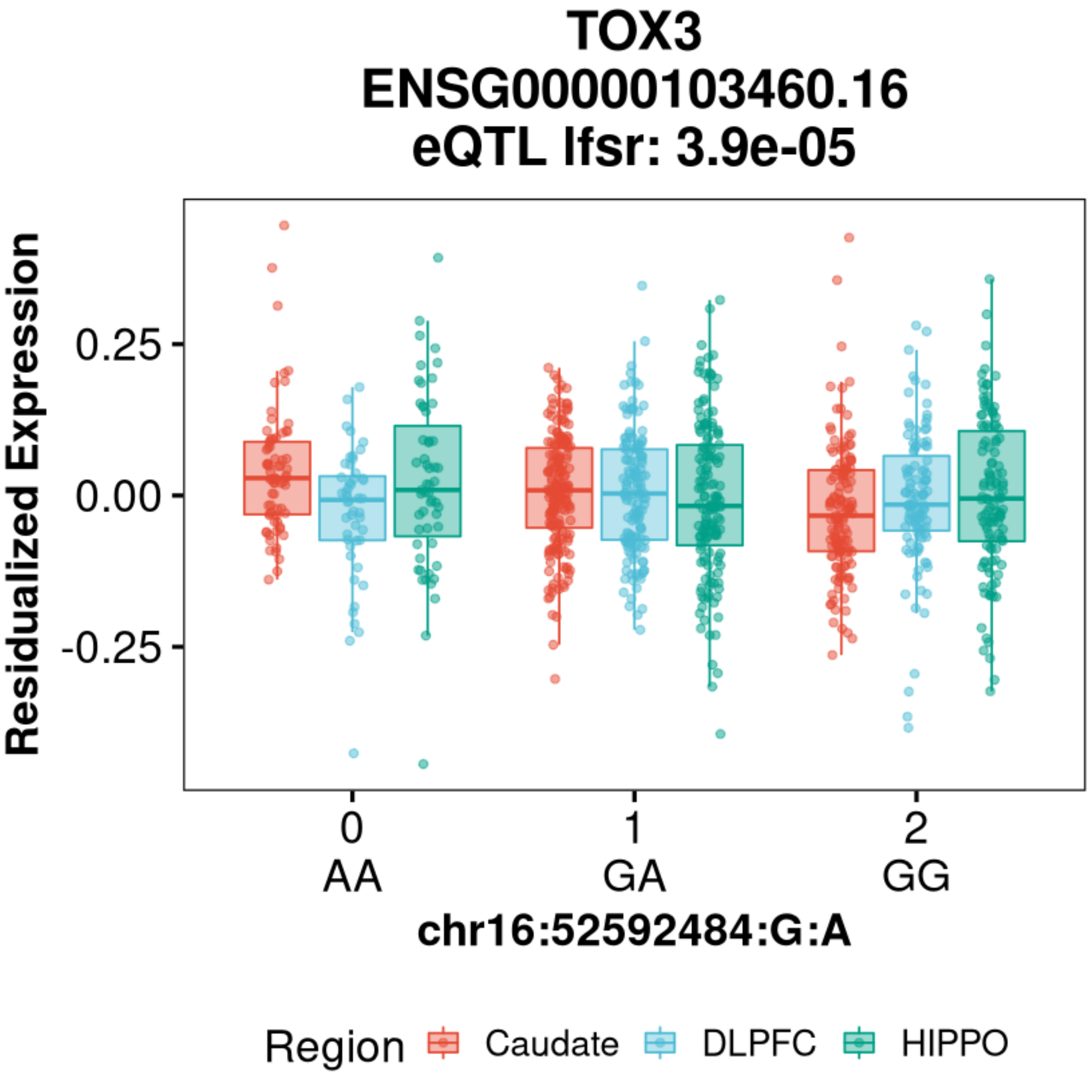

HIGD2A
ENSG00000146066.2
eQTL lfsr: 2.6e-09
Residualized Expression
0.4
0.2
0.0
-0.2
0
TT
1
TAGGCT
2
TAGGCTAGGC
chr5:176387602:TAGGC:T
Region
Caudate
DLPFC
HIPPO

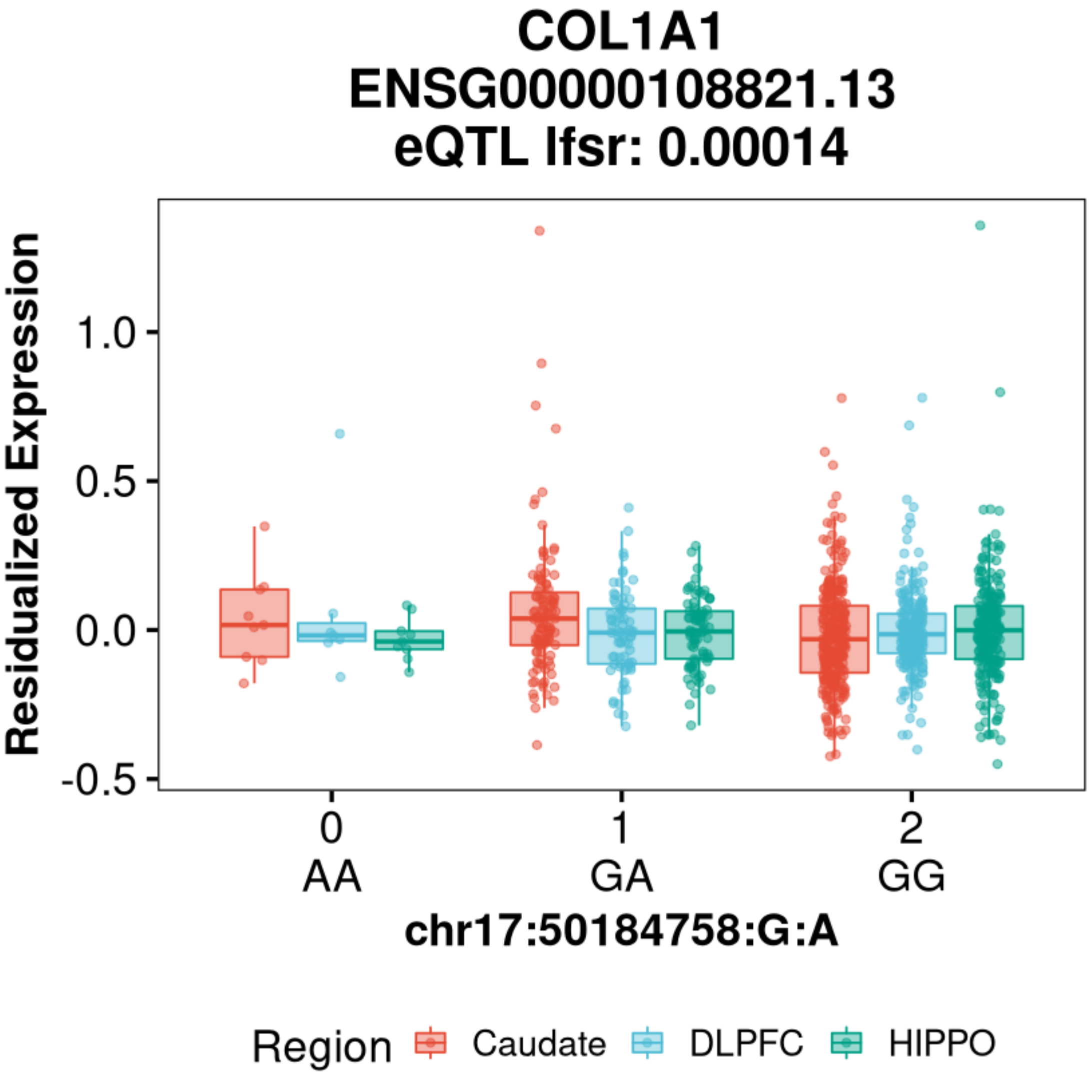

AOX1
ENSG00000138356.13
eQTL lfsr: 3.4e-06
Residualized Expression
1.5
1.0
0.5
0.0
-0.5
0
TT
1
CT
2
CC
chr2:200688153:C:T
Region
Caudate
DLPFC
HIPPO

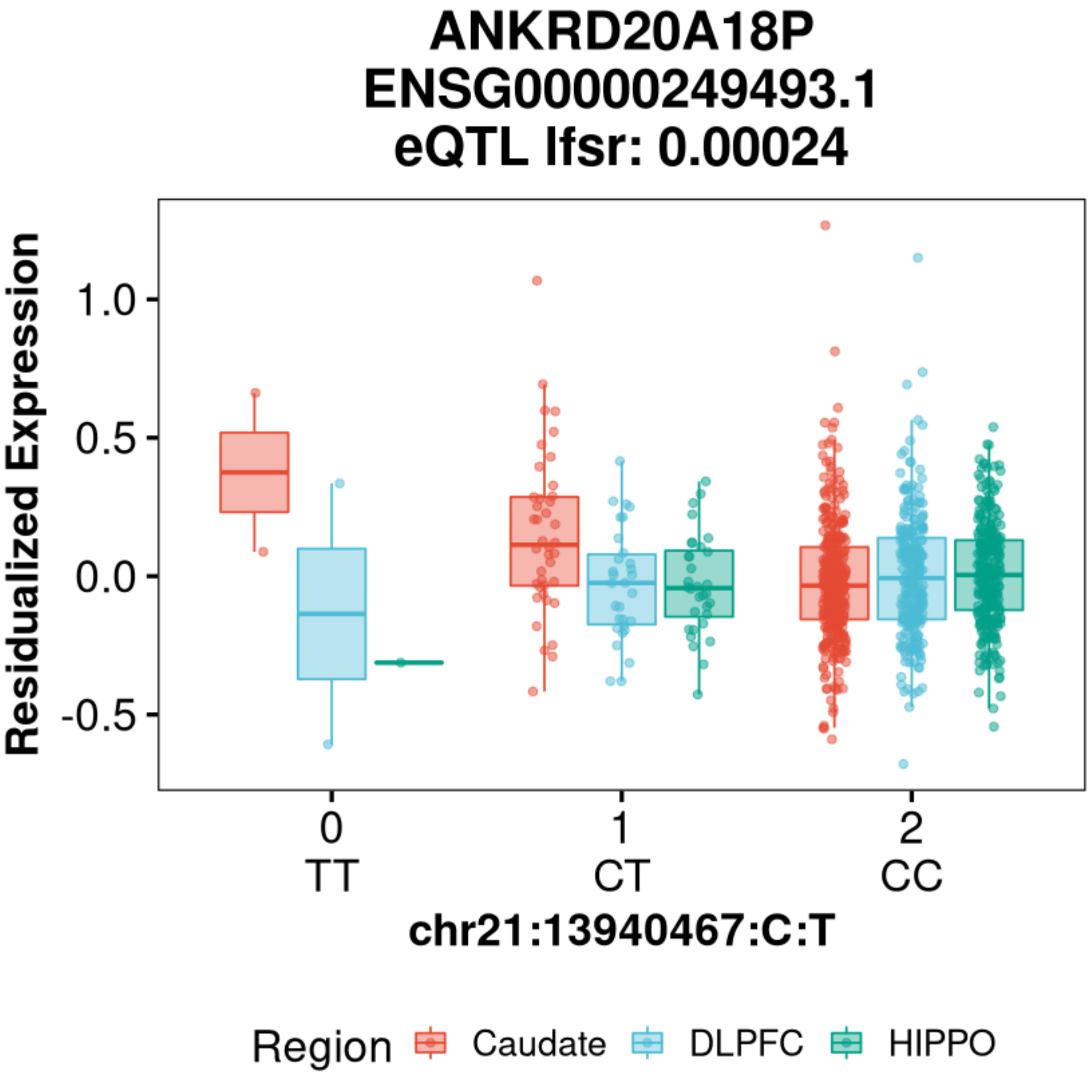

COX5B
ENSG00000135940.6
eQTL lfsr: 6.2e-06
Residualized Expression
0.4
0.2
0.0
-0.2
-0.4
0
TT
1
CT
2
CC
chr2:97691665:C:T
Region
Caudate
DLPFC
HIPPO

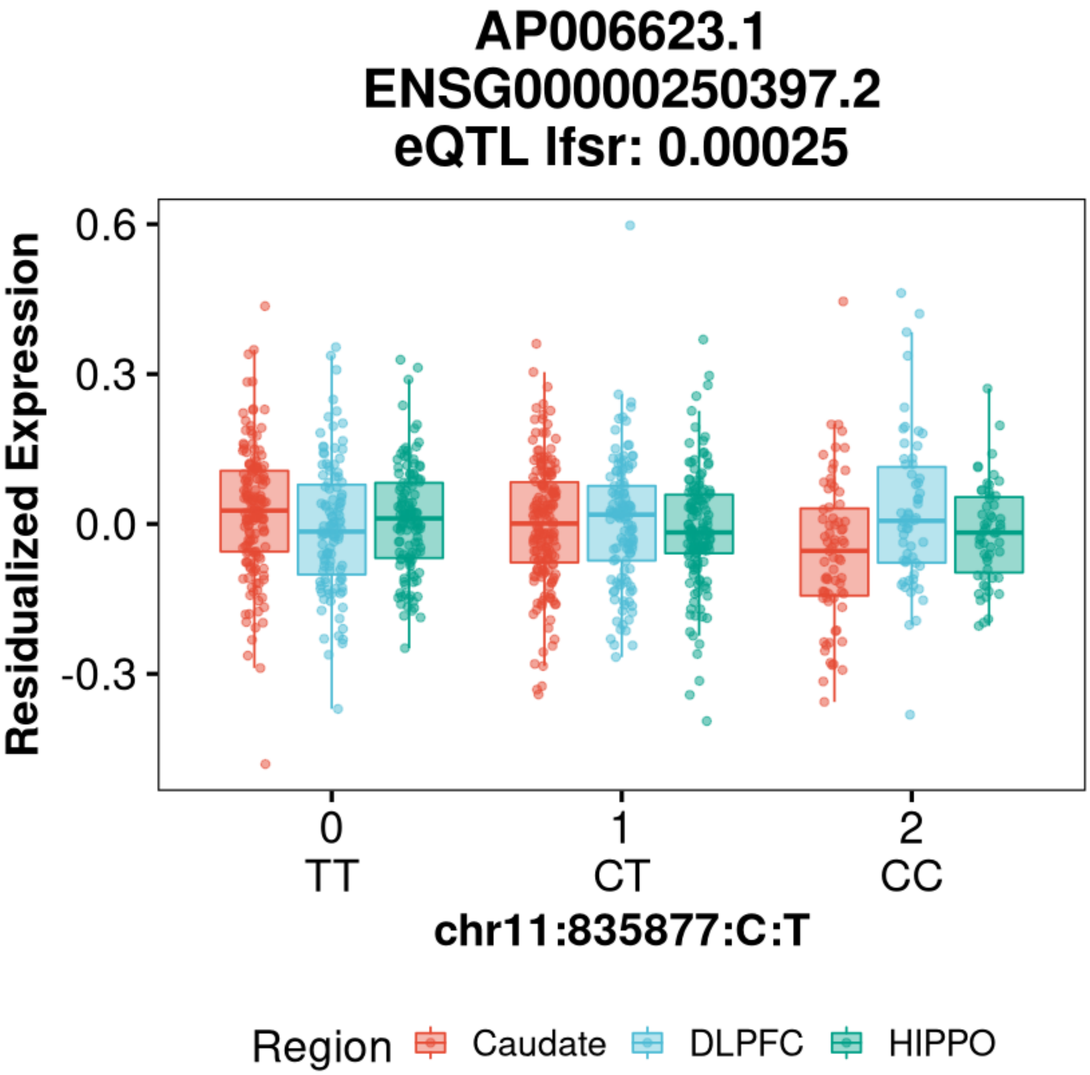

GRIK1
ENSG00000171189.17
eQTL lfsr: 7.7e-06
Residualized Expression
0.5
0.0
-0.5
0
GG
1
AG
2
AA
chr21:30089992:A:G
Region
Caudate
DLPFC
HIPPO

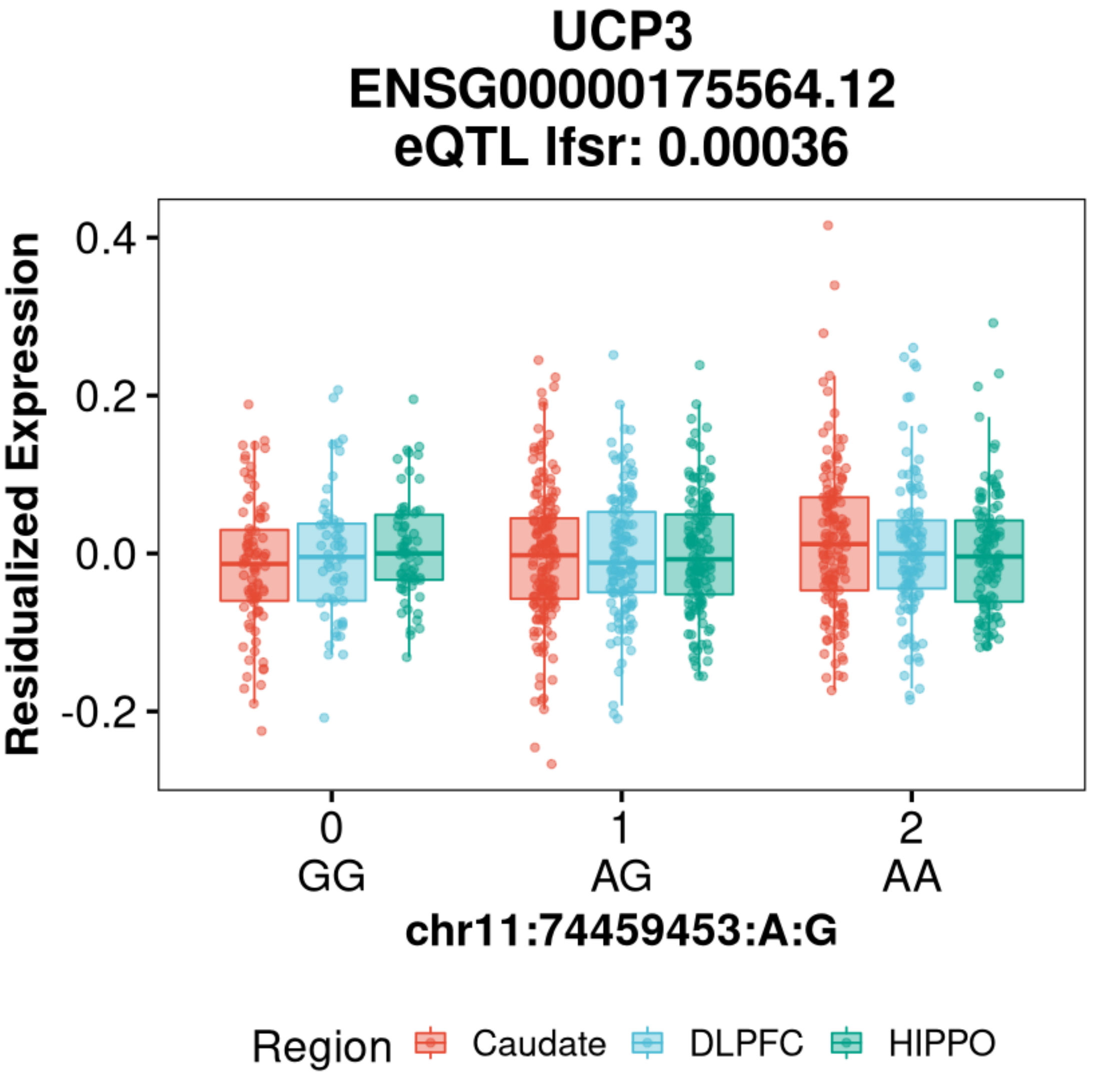

BTG3
ENSG00000154640.14
eQTL lfsr: 1.4e-05
Residualized Expression
0.8
0.4
0.0
-0.4
0
CC
1
TC
2
TT
chr21:17674435:T:C
Region
Caudate
DLPFC
HIPPO


TOX3
ENSG00000103460.16
eQTL lfsr: 3.9e-05
Residualized Expression
0.25
0.00
-0.25
0
AA
1
GA
2
GG
chr16:52592484:G:A
Region
Caudate
DLPFC
HIPPO


COL1A1
ENSG00000108821.13
eQTL lfsr: 0.00014
Residualized Expression
1.0
0.5
0.0
-0.5
0
AA
1
GA
2
GG
chr17:50184758:G:A
Region
Caudate
DLPFC
HIPPO


ANKRD20A18P
ENSG00000249493.1
eQTL lfsr: 0.00024
Residualized Expression
1.0
0.5
0.0
-0.5
0
TT
1
CT
2
CC
chr21:13940467:C:T
Region
Caudate
DLPFC
HIPPO


AP006623.1
ENSG00000250397.2
eQTL lfsr: 0.00025
Residualized Expression
0.6
0.3
0.0
-0.3
0
TT
1
CT
2
CC
chr11:835877:C:T
Region
Caudate
DLPFC
HIPPO



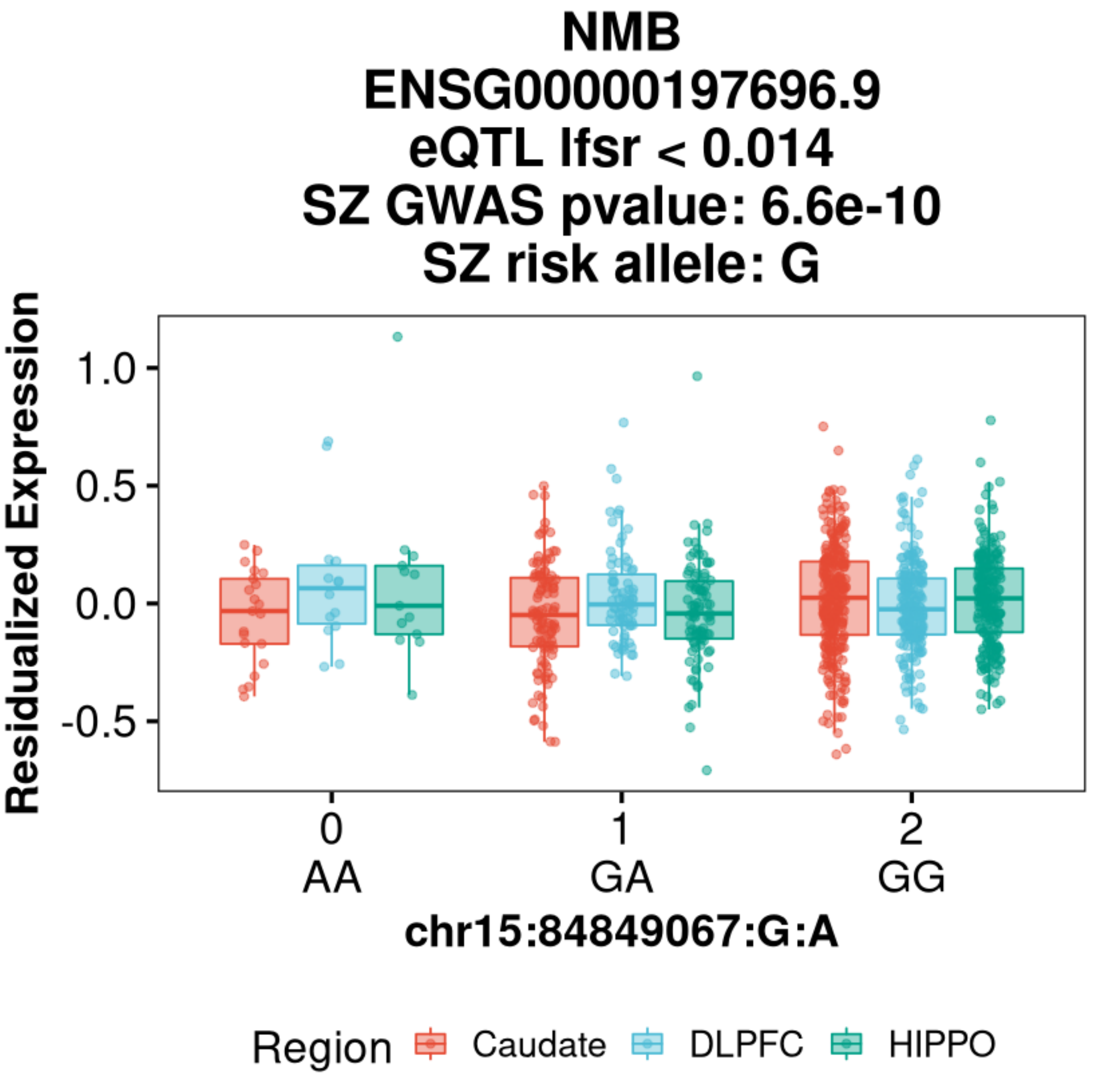
1.2.2 Top 5 GWAS associated eQTLs

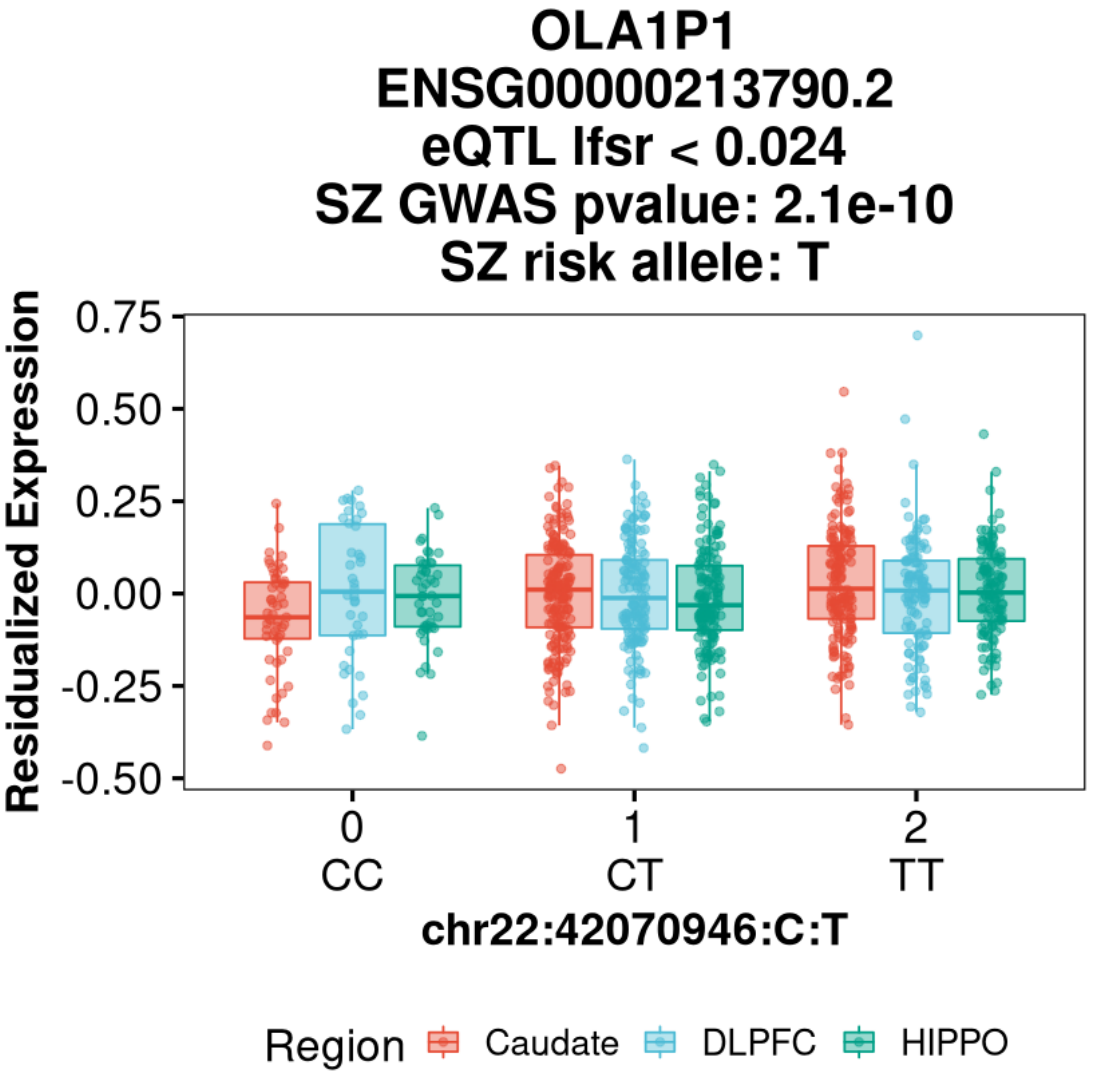
```
[8]: eGenes_gwas = get_eqtl_gwas_df() %>% arrange(Caudate, P) %>% group_by(gene_id)␣
      ↪%>% slice(1) %>% arrange(Caudate, P)
     eGenes_gwas %>% head(5)
```

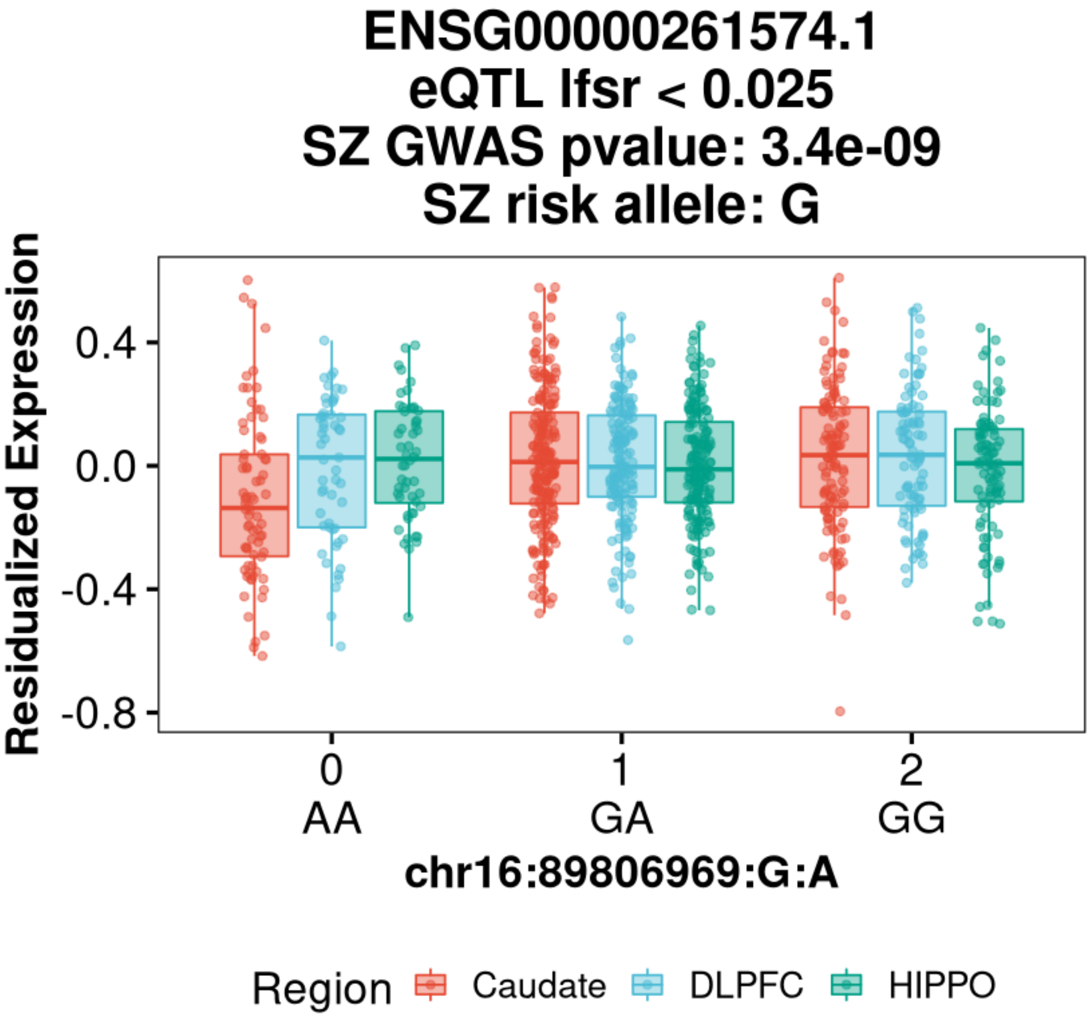
A grouped_df: 4 × 33

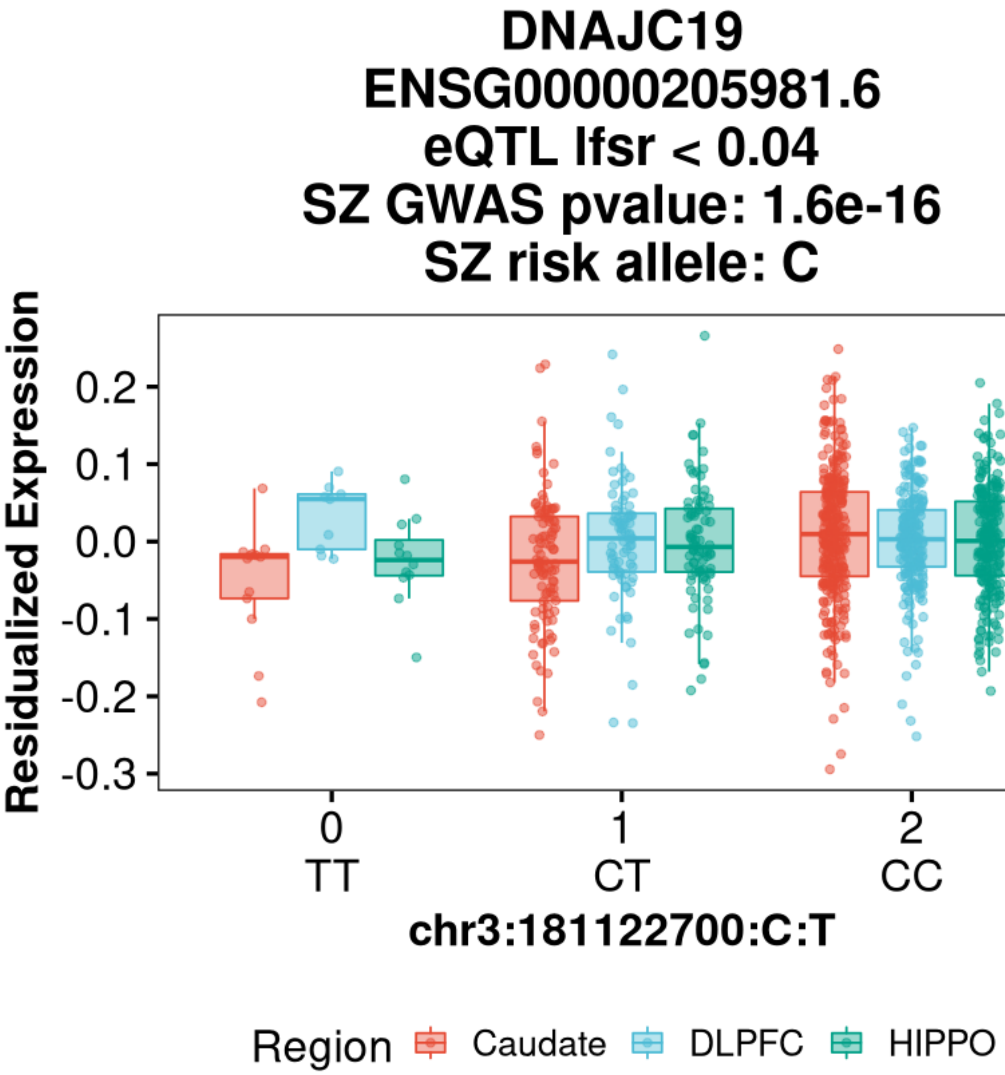
effect <chr>	gene_id <chr>	variant_id <chr>
ENSG00000197696.9_chr15:84849067:G:A	ENSG00000197696.9	chr15:84849067:G:A
ENSG00000213790.2_chr22:42070946:C:T	ENSG00000213790.2	chr22:42070946:C:T
ENSG00000261574.1_chr16:89806969:G:A	ENSG00000261574.1	chr16:89806969:G:A
ENSG00000205981.6_chr3:181122700:C:T	ENSG00000205981.6	chr3:181122700:C:T

```
[9]: for(num in 1:4){
         fn = paste("top",num,"interacting_eqtl_pgc3_variants", sep="_")
         variant_id = eGenes_gwas$variant_id[num]
         gene_id = eGenes_gwas$gene_id[num]
         pgc3_a1_same_as_our_counted = eGenes_gwas$pgc3_a1_same_as_our_counted[num]
         OR = eGenes_gwas$OR[num]
         eqtl_annot = paste("eQTL lfsr <", signif(eGenes_gwas$Caudate[num], 2))
         gwas_annot = paste("SZ GWAS pvalue:", signif(eGenes_gwas$P[num], 2))
         risk_annot = paste("SZ risk allele:",␣
      ↪get_risk_allele(eGenes_gwas$variant_id[num]))
         title = paste(get_gene_symbol(gene_id), gene_id, eqtl_annot,
                       gwas_annot, risk_annot, sep='\n')
         plot_gwas_eqtl(fn, gene_id, variant_id, eqtl_annot,
                        pgc3_a1_same_as_our_counted, OR, title)
     }
```




OLA1P1
ENSG00000213790.2
eQTL lfsr < 0.024
SZ GWAS pvalue: 2.1e-10
SZ risk allele: T
Residualized Expression
0.75
0.50
0.25
0.00
-0.25
-0.50
0
CC
1
CT
2
TT
chr22:42070946:C:T
Region
Caudate
DLPFC
HIPPO


ENSG00000261574.1
eQTL lfsr < 0.025
SZ GWAS pvalue: 3.4e-09
SZ risk allele: G
Residualized Expression
0.4
0.0
-0.4
-0.8
0
AA
1
GA
2
GG
chr16:89806969:G:A
Region
Caudate
DLPFC
HIPPO



1.3 Session Info

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

```
[1] "2022-03-08 19:29:20 EST"

    user   system  elapsed 
9901.426  734.873 1202.516 
```

$platform **$version** ’R version 4.1.2 (2021-11-01)’

$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 ’2022-03-08’

$pandoc ’2.14.1 @ /usr/bin/pandoc’

$packages A packages_info: 78 × 11

	package	ondiskversion	loadedversion	path
	<chr>	<chr>	<chr>	<chr>
abind	abind	1.4.5	1.4-5	/home/jbe
assertthat	assertthat	0.2.1	0.2.1	/home/jbe
backports	backports	1.4.1	1.4.1	/home/jbe
base64enc	base64enc	0.1.3	0.1-3	/home/jbe
broom	broom	0.7.12	0.7.12	/home/jbe
cachem	cachem	1.0.6	1.0.6	/home/jbe
car	car	3.0.12	3.0-12	/home/jbe
carData	carData	3.0.5	3.0-5	/home/jbe
cellranger	cellranger	1.1.0	1.1.0	/home/jbe
cli	cli	3.1.1	3.1.1	/home/jbe
colorspace	colorspace	2.0.2	2.0-2	/home/jbe
crayon	crayon	1.4.2	1.4.2	/home/jbe
data.table	data.table	1.14.2	1.14.2	/home/jbe
DBI	DBI	1.1.2	1.1.2	/home/jbe
dbplyr	dbplyr	2.1.1	2.1.1	/home/jbe
digest	digest	0.6.29	0.6.29	/home/jbe
dplyr	dplyr	1.0.7	1.0.7	/home/jbe
ellipsis	ellipsis	0.3.2	0.3.2	/home/jbe
evaluate	evaluate	0.14	0.14	/home/jbe
fansi	fansi	1.0.2	1.0.2	/home/jbe
farver	farver	2.1.0	2.1.0	/home/jbe
fastmap	fastmap	1.1.0	1.1.0	/home/jbe
forcats	forcats	0.5.1	0.5.1	/home/jbe
fs	fs	1.5.2	1.5.2	/home/jbe
generics	generics	0.1.2	0.1.2	/home/jbe
ggplot2	ggplot2	3.3.5	3.3.5	/home/jbe
ggpubr	ggpubr	0.4.0	0.4.0	/home/jbe
ggsci	ggsci	2.9	2.9	/home/jbe
ggsignif	ggsignif	0.6.3	0.6.3	/home/jbe
glue	glue	1.6.1	1.6.1	/home/jbe
purrr	purrr	0.3.4	0.3.4	/home/jbe
R.methodsS3	R.methodsS3	1.8.1	1.8.1	/home/jbe
R.oo	R.oo	1.24.0	1.24.0	/home/jbe
R.utils	R.utils	2.11.0	2.11.0	/home/jbe
R6	R6	2.5.1	2.5.1	/home/jbe
Rcpp	Rcpp	1.0.8	1.0.8	/home/jbe
readr	readr	2.1.2	2.1.2	/home/jbe
readxl	readxl	1.3.1	1.3.1	/home/jbe
repr	repr	1.1.4	1.1.4	/home/jbe
reprex	reprex	2.0.1	2.0.1	/home/jbe
rlang	rlang	1.0.0	1.0.0	/home/jbe
rstatix	rstatix	0.7.0	0.7.0	/home/jbe
rstudioapi	rstudioapi	0.13	0.13	/home/jbe
rvest	rvest	1.0.2	1.0.2	/home/jbe
scales	scales	1.1.1	1.1.1	/home/jbe
sessioninfo	sessioninfo	1.2.2	1.2.2	/home/jbe
stringi	stringi	1.7.6	1.7.6	/home/jbe
stringr	stringr	1.4.0	1.4.0	/home/jbe
svglite	svglite	2.0.0	2.0.0	/home/jbe
systemfonts	systemfonts	1.0.3	1.0.3	/home/jbe