main_junctions

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 = "junctions"
```

1.1.1 Basic loading functions

[3]:
```
get_residualized_df <- function(){
    expr_file = paste0("/ceph/projects/v4_phase3_paper/analysis/eqtl_analysis/
 ↪all/",
                       feature,"/expression_gct/covariates/
 ↪residualized_expression/_m/",
                       feature, "_residualized_expression.csv")
    return(data.table::fread(expr_file) %>% column_to_rownames("gene_id"))
}
memRES <- memoise::memoise(get_residualized_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_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 eQTL and helpful functions

[4]:

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

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_caudate_eqtls <- function(){
    mashr_file = "../../summary_table/_m/BrainSeq_caudate_eQTL.txt.gz"
    return(data.table::fread(mashr_file) %>%
           filter(Type == feature_map(feature)) %>%
           select(gene_id, variant_id, AA, EA))
}
memCAUDATE <- memoise::memoise(get_caudate_eqtls)

get_eqtl_df <- function(){
    eGenes_file = paste0("../../_m/", feature, "/lfsr_allpairs_ancestry.txt.gz")
    eGenes = data.table::fread(eGenes_file)
    return(eGenes)
}
memEQTL <- memoise::memoise(get_eqtl_df)
```

1.1.3 Basic eQTL plotting functions

[5]:

```
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))
}

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("BrNum") %>% inner_join(memPHENO(), by="BrNum")
    ## 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)

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)
    }
}
```

[6]:

```
plot_simple_eqtl <- function(fn, gene_id, variant_id, eqtl_annot, prefix,␣
 ↪y0=NULL, y1=NULL){
    if(is.null(y0)){ y0 = quantile(memDF(variant_id, gene_id)[[gene_id]],␣
 ↪probs=c(0.01))[[1]] - 0.2 }
    if(is.null(y1)){ y1 = quantile(memDF(variant_id, gene_id)[[gene_id]],␣
 ↪probs=c(0.99))[[1]] + 0.2 }
    bxp = memDF(variant_id, gene_id) %>%
        ggboxplot(x="ID", y=gene_id, fill="Race", color="Race", add="jitter",
                  xlab=variant_id, ylab="Residualized Expression", outlier.
 ↪shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4, legend="bottom",
                  palette="npg", ylim=c(y0,y1),␣
↪ggtheme=theme_pubr(base_size=20, border=TRUE)) +
        font("xy.title", face="bold") +
        ggtitle(paste(prefix, gene_id, eqtl_annot, sep='\n')) +
        theme(plot.title = element_text(hjust = 0.5, face="bold"))
    print(bxp)
    save_ggplots(fn, bxp, 7, 7)
}
```

1.1.4 GWAS plots

[7]:

```
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,␣
 ↪pgc3_a1_same_as_our_counted,
                           OR, title){
    dt = get_gwas_ordered_snp_df(variant_id, gene_id,␣
 ↪pgc3_a1_same_as_our_counted, OR)
    bxp = dt %>% mutate_if(is.character, as.factor) %>%
        ggboxplot(x="ID", y=gene_id, fill="Race", color="Race", add="jitter",
                  xlab=variant_id, ylab="Residualized Expression", outlier.
 ↪shape=NA,
                  add.params=list(alpha=0.5), alpha=0.4, legend="bottom",␣
 ↪#ylim=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

[8]:

```
get_drd2_junction_annotation <- function(junction_id){
    return(list(
        'chr11:113424683-113474229(-)'= "DRD2 junction 1L-2",
        "chr11:113424683-113475075(-)"= "DRD2 junction 1-2",
        "chr11:113418137-113424366(-)"= "DRD2 junction 2-3",
        "chr11:113417000-113418026(-)"= "DRD2 junction 3-4",
        "chr11:113415612-113416862(-)"= "DRD2 junction 4-5",
        "chr11:113414462-113415420(-)"= "DRD2 junction 5-6",
        "chr11:113412884-113415420(-)"= "DRD2 junction 5-7",
        "chr11:113412884-113414374(-)"= "DRD2 junction 6-7",
        "chr11:113410921-113412555(-)"= "DRD2 junction 7-8")[[junction_id]])
}

get_drd2_junctions <- function(){
    cmd = paste0("cat <(head -1 /ceph/projects/v4_phase3_paper/analysis/
 ↪differential_expression/_m/junctions/diffExpr_szVctl_full.txt)",
                 " <(grep -i drd2 /ceph/projects/v4_phase3_paper/analysis/
 ↪differential_expression/_m/junctions/diffExpr_szVctl_full.txt)")
    return(data.table::fread(cmd=cmd) %>% rename("Feature"="V1"))
}
```

1.2.1 DRD2 plot

[9]:
```
drdj = get_drd2_junctions() %>% filter(str_detect(gencodeTx, "ENST00000362072.
↪7|ENST00000346454.7"))
drdj
```

A data.table: 8 × 22

Feature <chr>	inGencode <lgl>	inGencodeStart <lgl>	inGencodeEnd <lgl>	gencodeGe <chr>
chr11:113410921-113412555(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113415612-113416862(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113412884-113415420(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113417000-113418026(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113424683-113475075(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113418137-113424366(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113412884-113414374(-)	TRUE	TRUE	TRUE	ENSG0000
chr11:113414462-113415420(-)	TRUE	TRUE	TRUE	ENSG0000

[10]:
```
drd2_df0 = memCAUDATE() %>% filter(gene_id %in% drdj$Feature) %>%
    arrange(AA, EA) %>% group_by(gene_id) %>% slice(1) %>% arrange(AA, EA)
drd2_df0
```

A grouped_df: 2 × 4

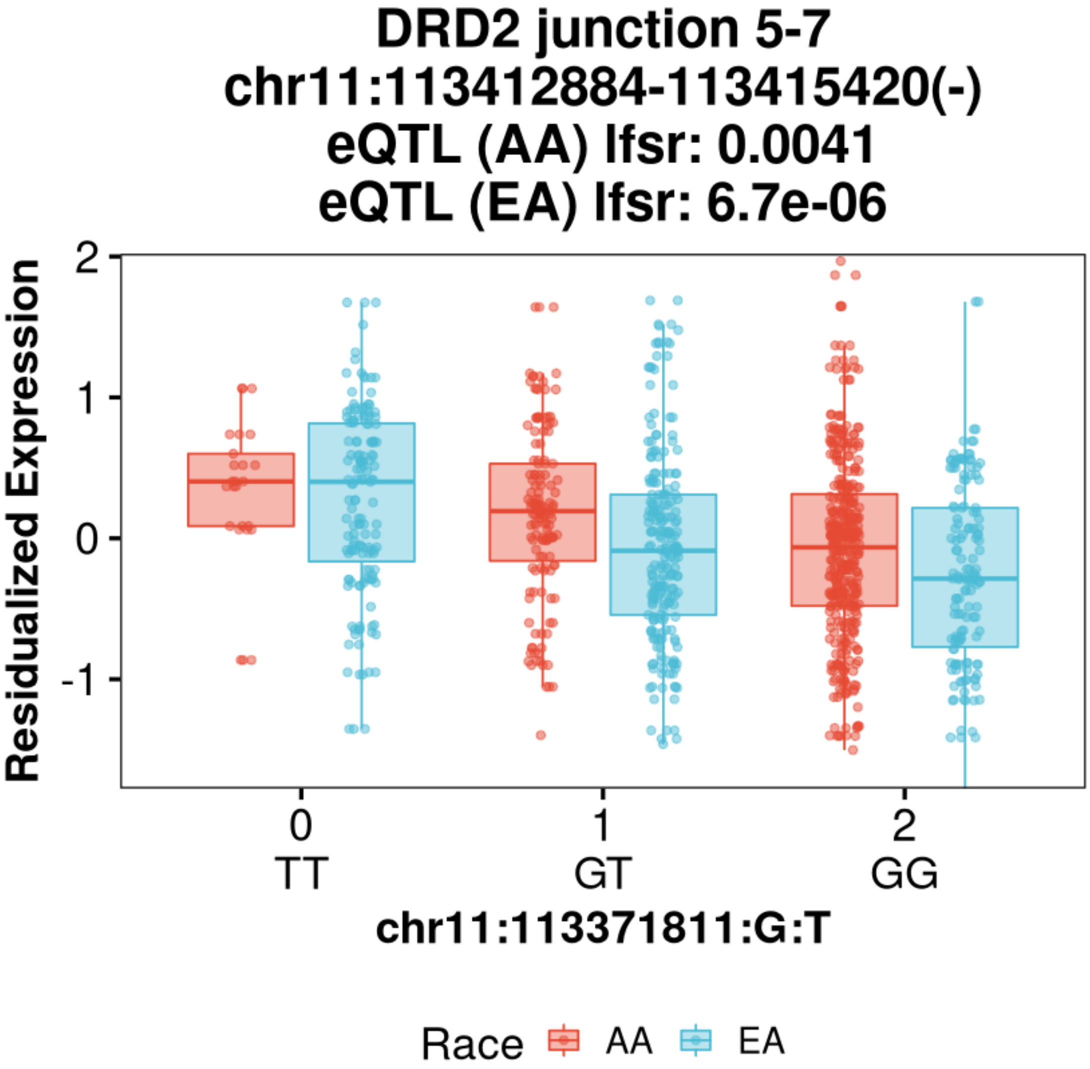
gene_id <chr>	variant_id <chr>	AA <dbl>	EA <dbl>
chr11:113412884-113415420(-)	chr11:113371811:G:T	0.004106071	6.748968e-06
chr11:113410921-113412555(-)	chr11:113434592:A:G	0.051615322	1.971606e-02

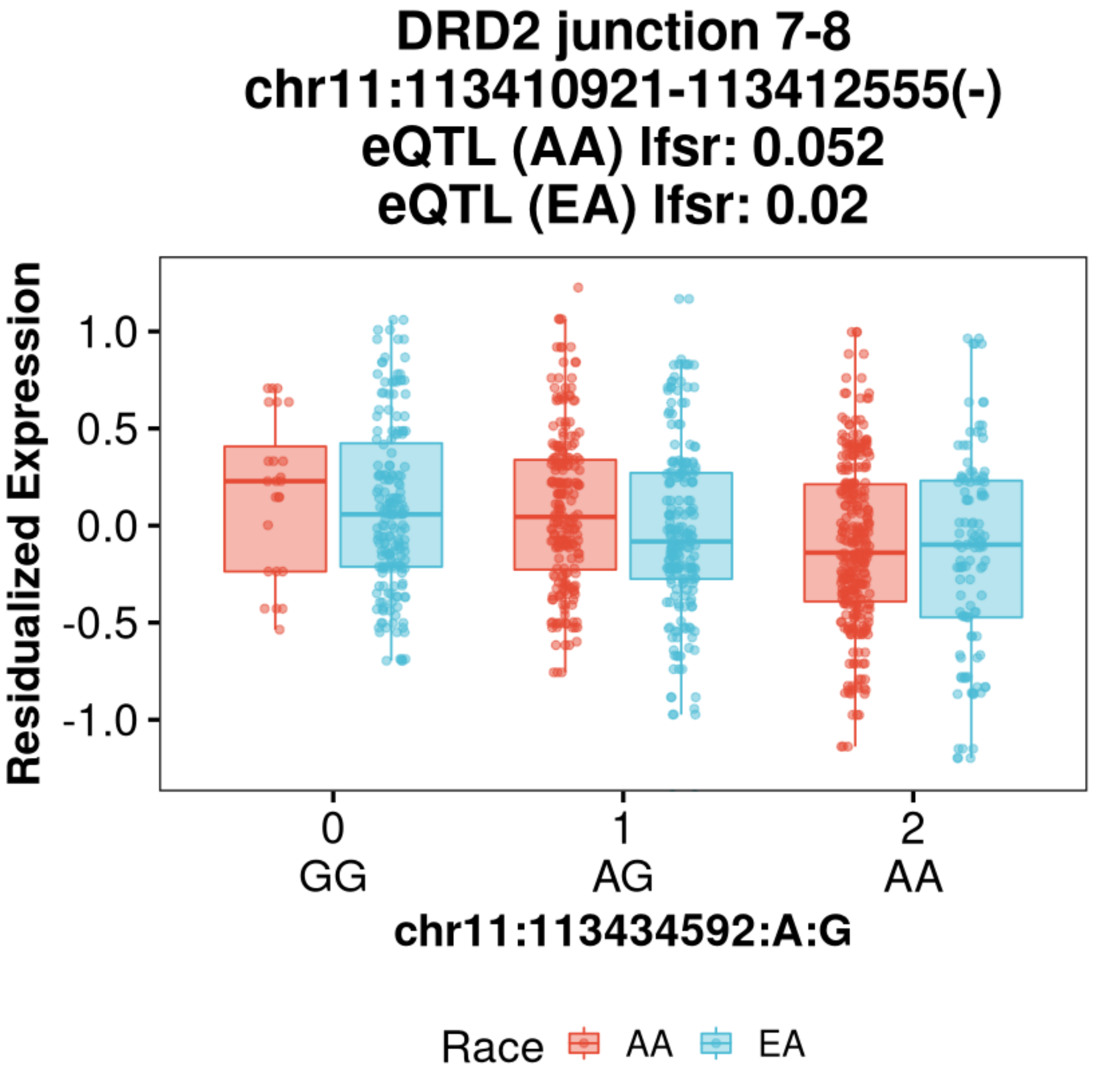
[11]:
```
drd2_df = memEQTL() %>% filter(gene_id %in% drdj$Feature) %>%
    arrange(AA, EA) %>% group_by(gene_id) %>% slice(1) %>% arrange(AA, EA)
drd2_df
```

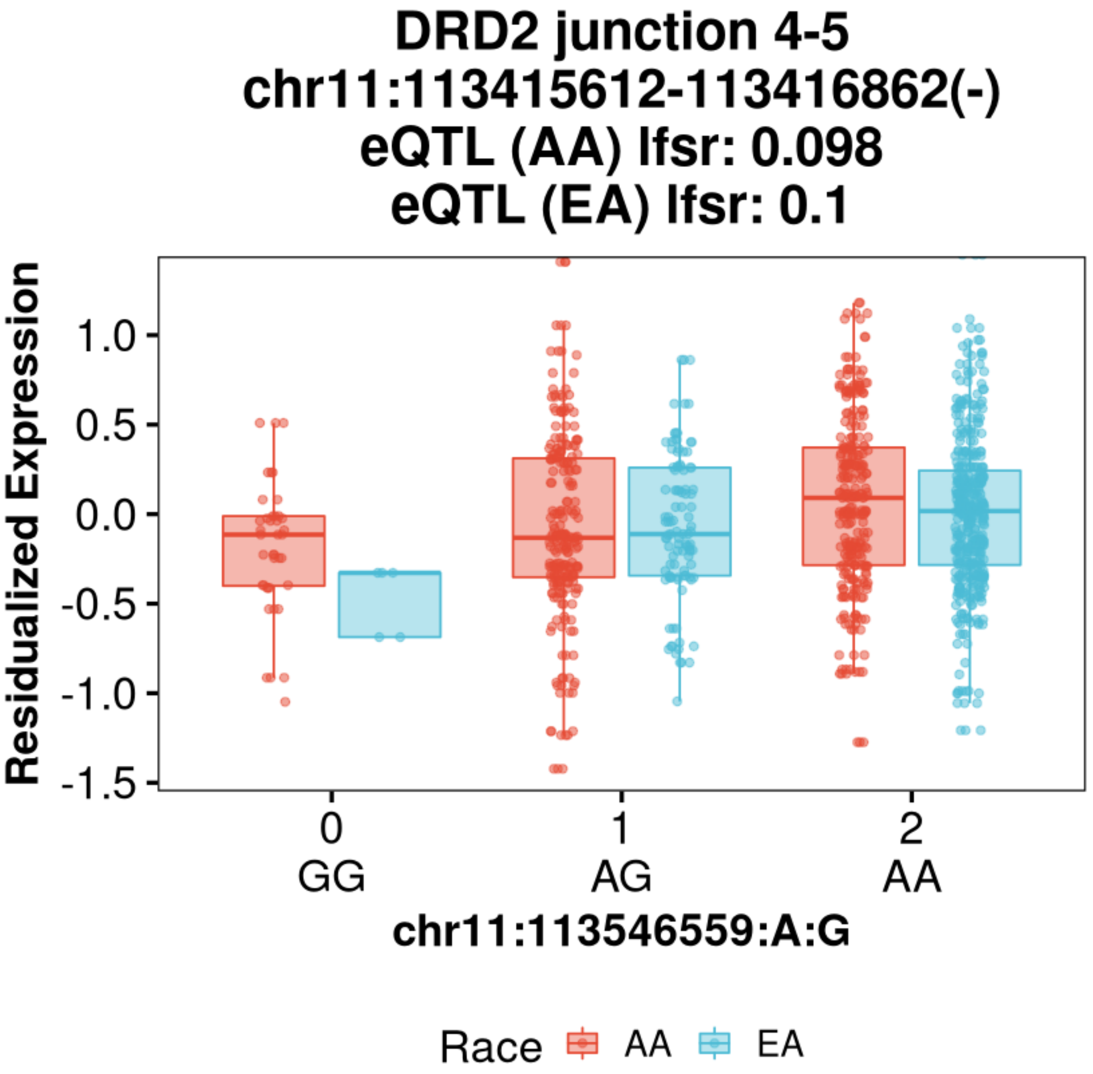
A grouped_df: 8 × 5

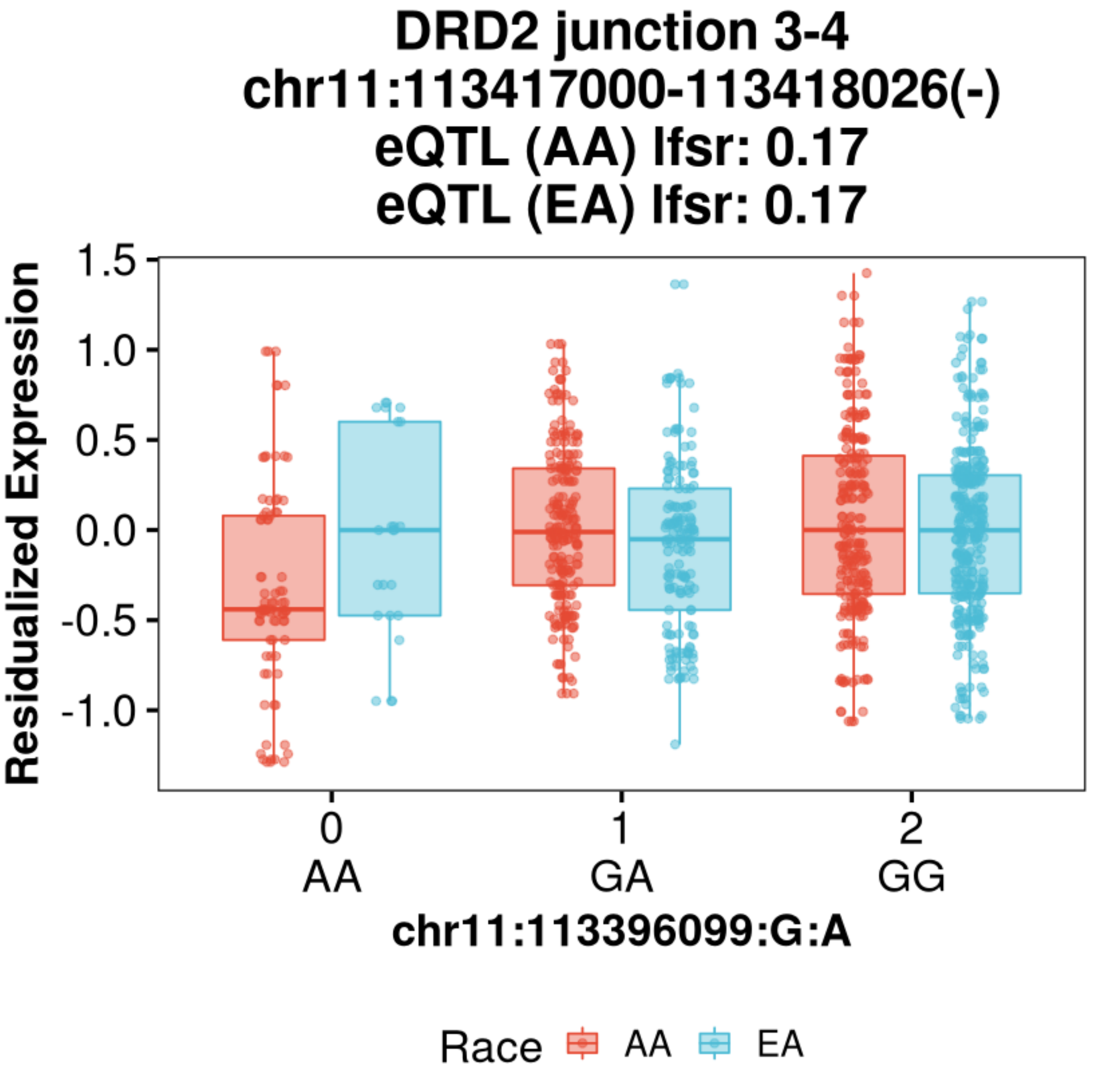
effect <chr>	gene_id <chr>	va <
chr11:113412884-113415420(-)_chr11:113371811:G:T	chr11:113412884-113415420(-)	cl
chr11:113410921-113412555(-)_chr11:113434592:A:G	chr11:113410921-113412555(-)	cl
chr11:113415612-113416862(-)_chr11:113546559:A:G	chr11:113415612-113416862(-)	cl
chr11:113417000-113418026(-)_chr11:113396099:G:A	chr11:113417000-113418026(-)	cl
chr11:113424683-113475075(-)_chr11:113192424:AG:A	chr11:113424683-113475075(-)	cl
chr11:113412884-113414374(-)_chr11:113249956:A:G	chr11:113412884-113414374(-)	cl
chr11:113418137-113424366(-)_chr11:113630933:G:A	chr11:113418137-113424366(-)	cl
chr11:113414462-113415420(-)_chr11:112955580:G:A	chr11:113414462-113415420(-)	cl

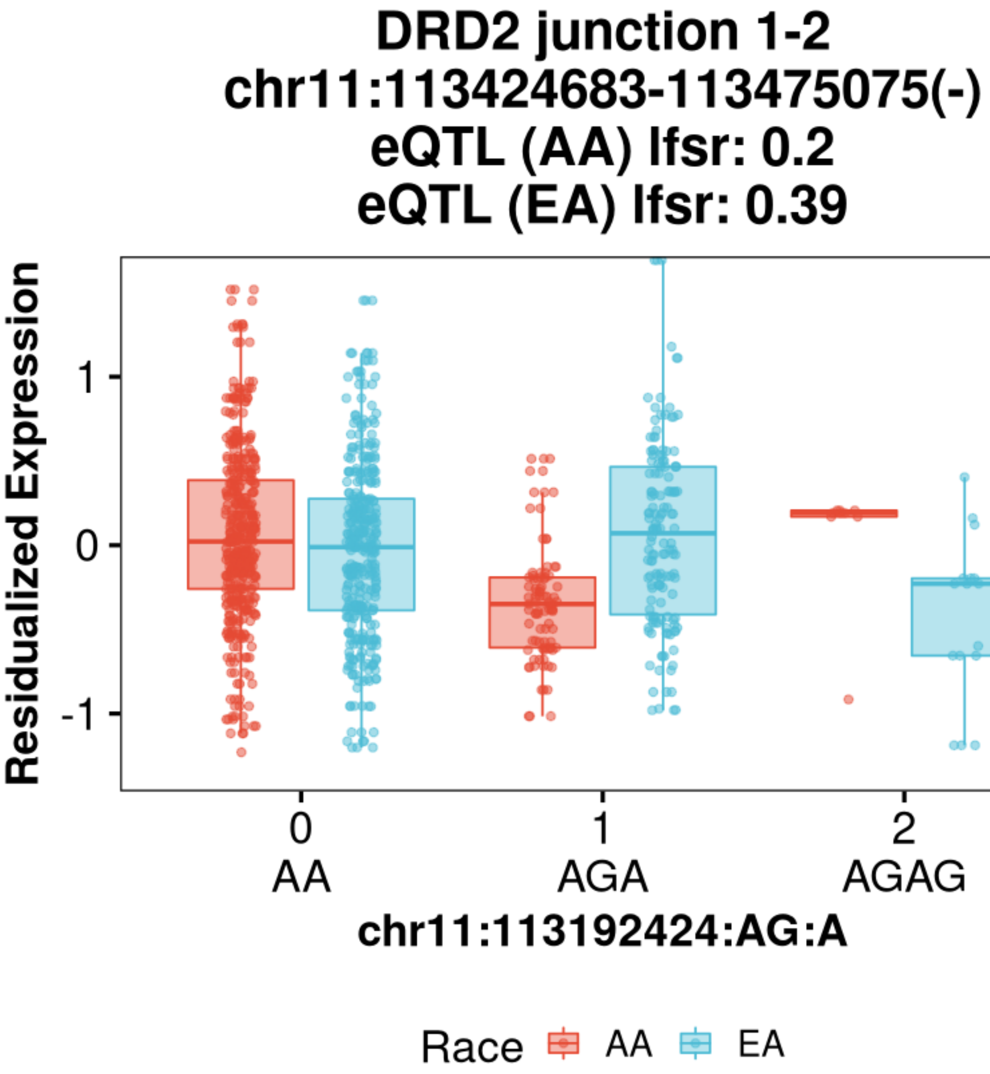
[12]:
```
for(x in seq_along(drd2_df$gene_id)){
    anno = get_drd2_junction_annotation(drd2_df$gene_id[x])
    en = gsub("-", "_", gsub(" ", "_", anno))
    fn = paste("drd2_eqtl", en, sep="_")
    eqtl_annot = paste(paste("eQTL (AA) lfsr:", signif(drd2_df$AA[x], 2)),
                        paste("eQTL (EA) lfsr:", signif(drd2_df$EA[x], 2)),␣
↪sep='\n')
    prefix = anno
    plot_simple_eqtl(fn, drd2_df$gene_id[x], drd2_df$variant_id[x], eqtl_annot,␣
↪prefix)
    #print(prefix)
}
```

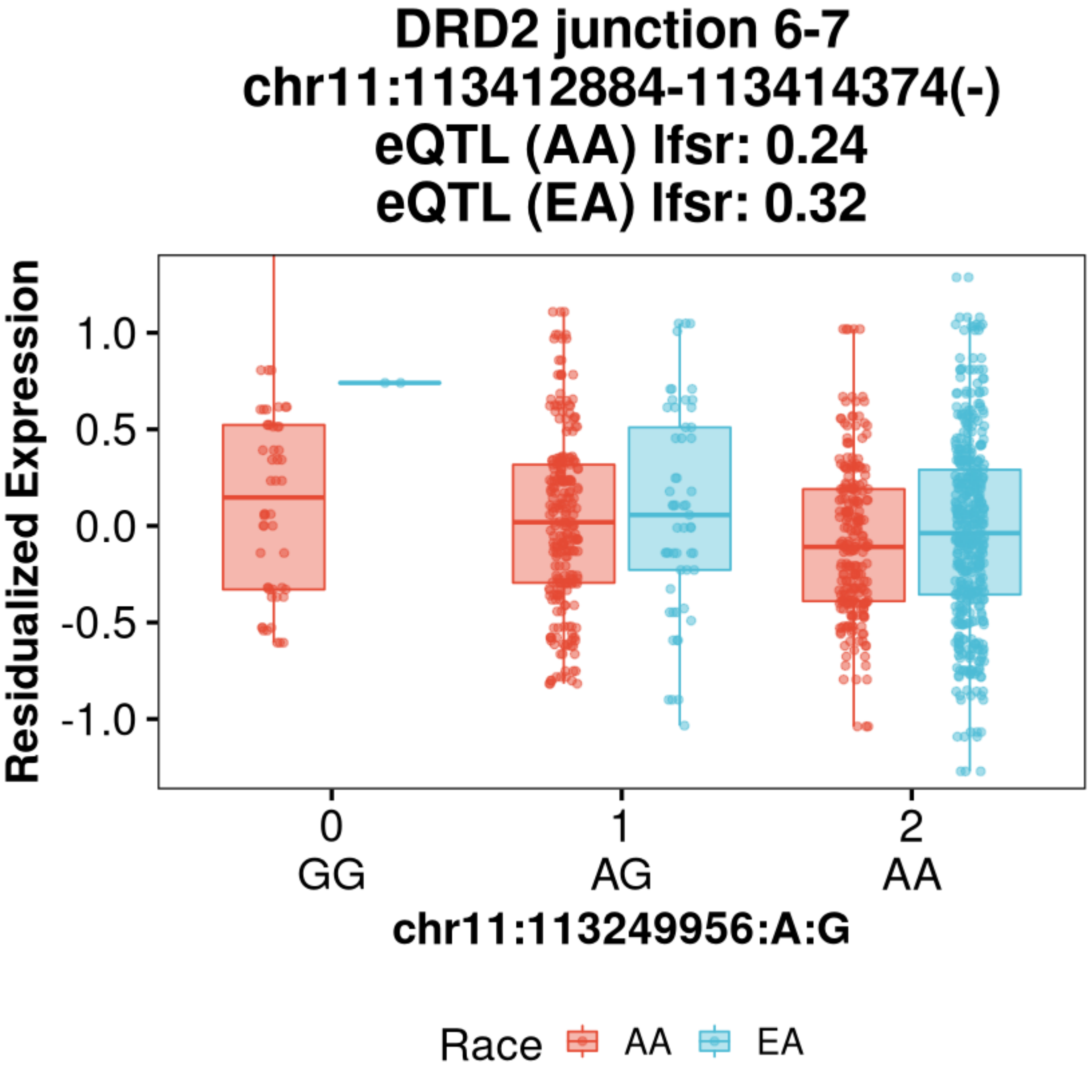


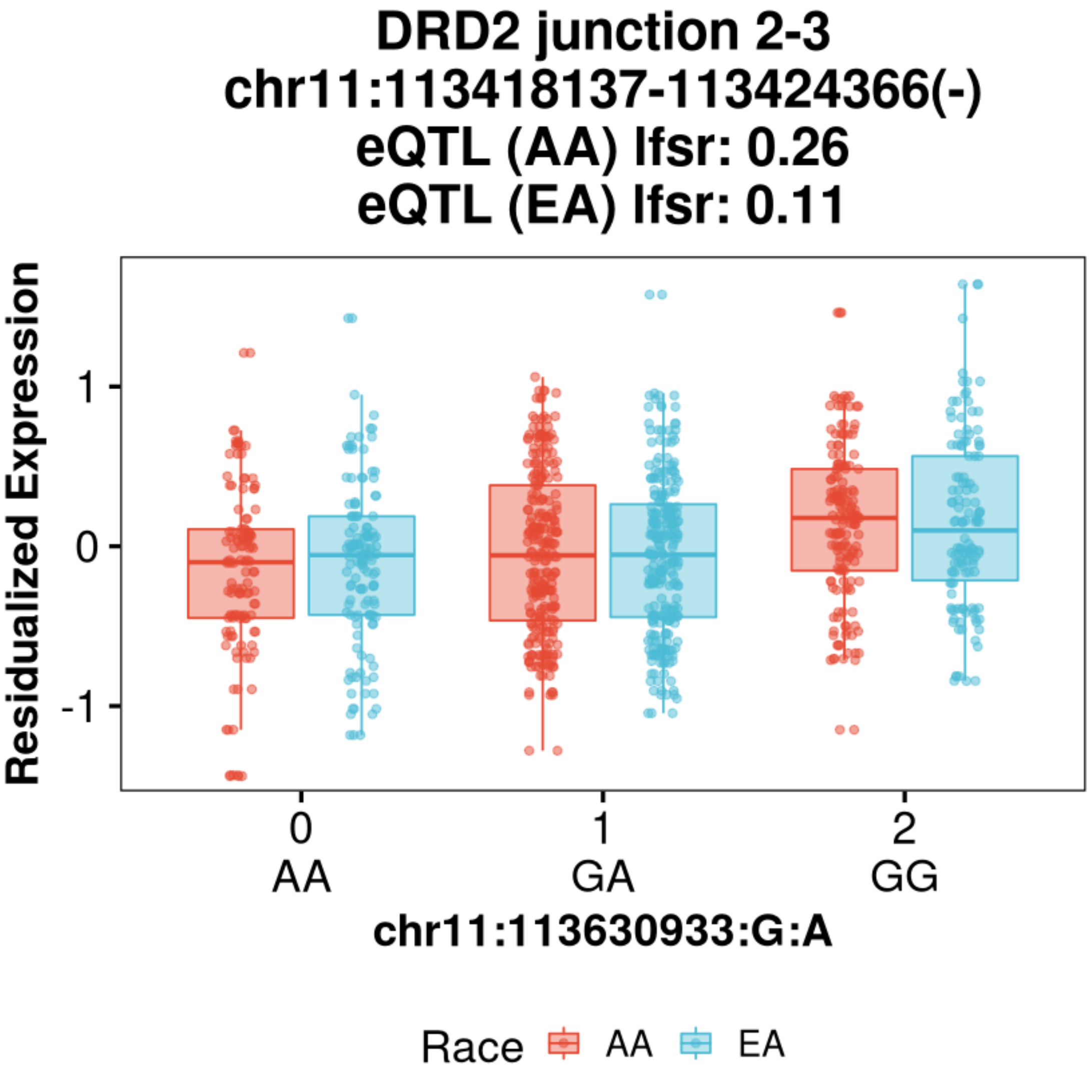

DRD2 junction 7-8
chr11:113410921-113412555(-)
eQTL (AA) lfsr: 0.052
eQTL (EA) lfsr: 0.02
Residualized Expression
1.0
0.5
0.0
-0.5
-1.0
0
GG
1
AG
2
AA
chr11:113434592:A:G
Race
AA
EA

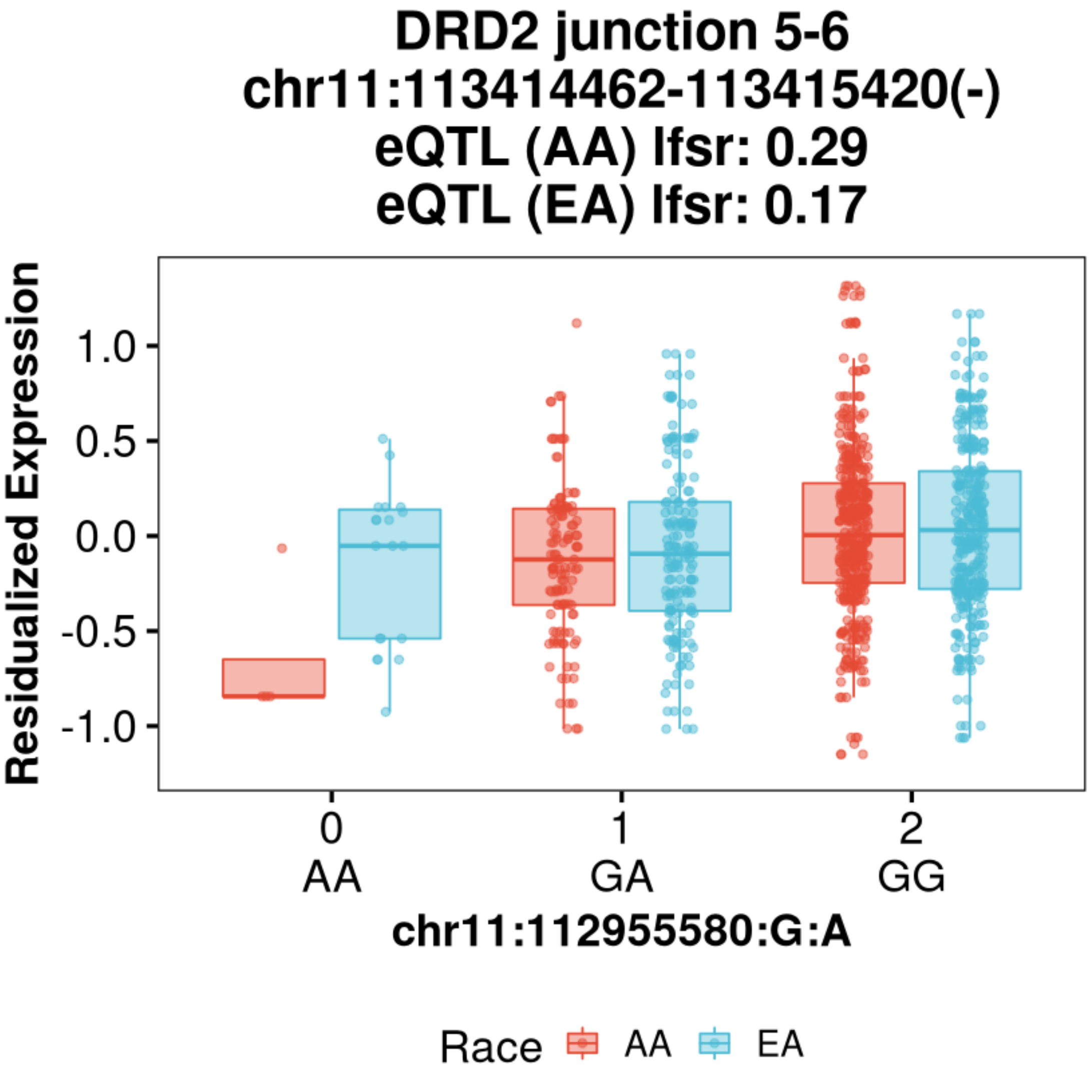

DRD2 junction 4-5
chr11:113415612-113416862(-)
eQTL (AA) lfsr: 0.098
eQTL (EA) lfsr: 0.1
Residualized Expression
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
GG
1
AG
2
AA
chr11:113546559:A:G
Race
AA
EA


DRD2 junction 3-4
chr11:113417000-113418026(-)
eQTL (AA) lfsr: 0.17
eQTL (EA) lfsr: 0.17
Residualized Expression
1.5
1.0
0.5
0.0
-0.5
-1.0
0
AA
1
GA
2
GG
chr11:113396099:G:A
Race
AA
EA


DRD2 junction 1-2
chr11:113424683-113475075(-)
eQTL (AA) lfsr: 0.2
eQTL (EA) lfsr: 0.39
Residualized Expression
1
0
-1
0
AA
1
AGA
2
AGAG
chr11:113192424:AG:A
Race
AA
EA


DRD2 junction 6-7
chr11:113412884-113414374(-)
eQTL (AA) lfsr: 0.24
eQTL (EA) lfsr: 0.32
Residualized Expression
1.0
0.5
0.0
-0.5
-1.0
0
GG
1
AG
2
AA
chr11:113249956:A:G
Race
AA
EA


DRD2 junction 2-3
chr11:113418137-113424366(-)
eQTL (AA) lfsr: 0.26
eQTL (EA) lfsr: 0.11
Residualized Expression
1
0
-1
0
AA
1
GA
2
GG
chr11:113630933:G:A
Race
AA
EA



1.2.2 GWAS association

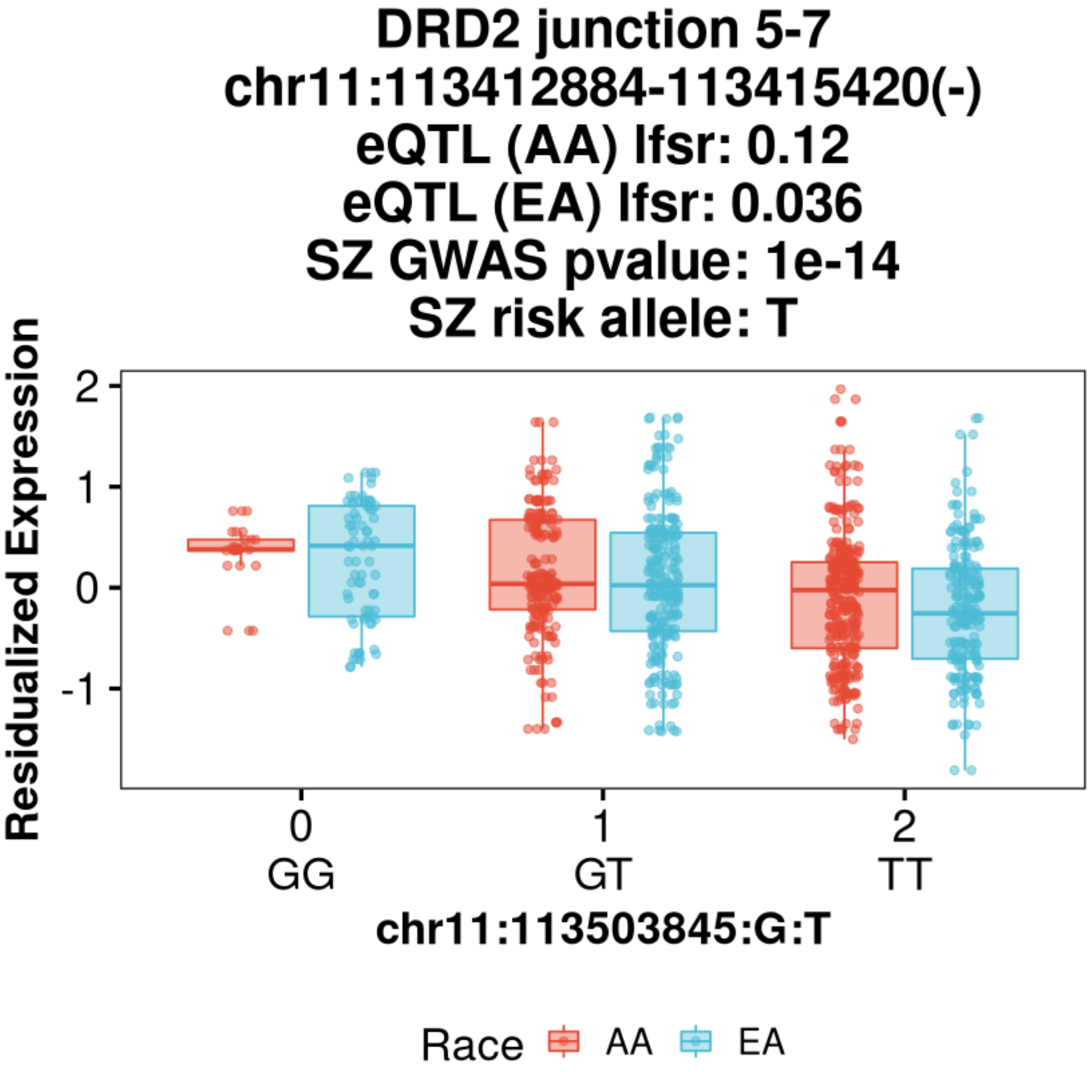
[13]:

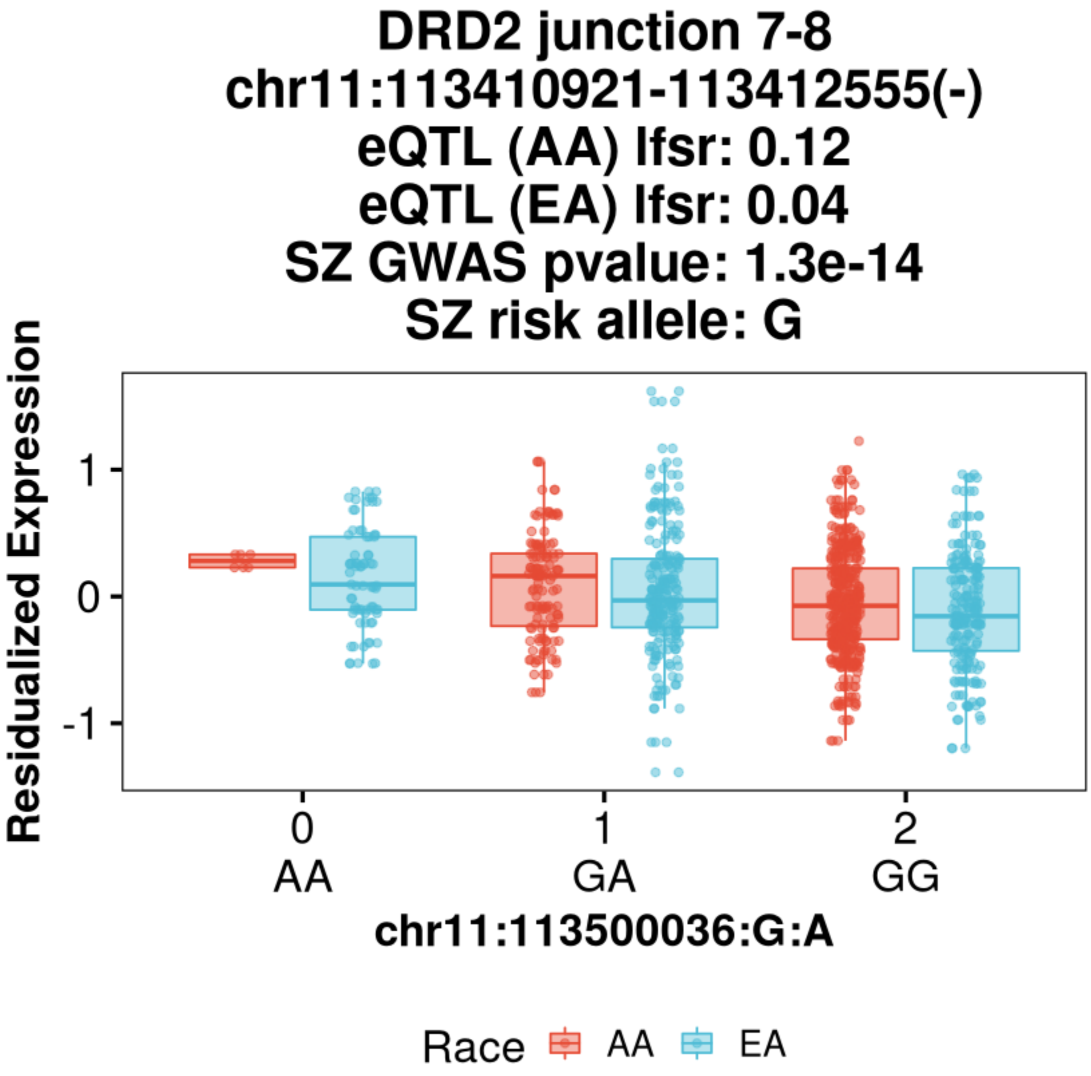
```
eGenes_gwas = get_eqtl_gwas_df() %>% filter(gene_id %in% drdj$Feature) %>%
    arrange(AA, EA, P) %>% group_by(gene_id) %>% slice(1) %>% arrange(AA, EA, P)
eGenes_gwas
```

```
                         gene_id                       variant_id             AA          EA            V1
                         <chr>                         <chr>                  <dbl>       <dbl>         <int>
A grouped_df: 2 × 28 ──────────────────────────────────────────────────────────────────────────────────────
                         chr11:113412884-113415420(-)  chr11:113503845:G:T    0.1196574   0.03557255    982
                         chr11:113410921-113412555(-)  chr11:113500036:G:A    0.1232662   0.04010518    979
```

[14]:

```
for(num in seq_along(eGenes_gwas$variant_id)){
    anno = get_drd2_junction_annotation(eGenes_gwas$gene_id[num])
    en = gsub("-", "_", gsub(" ", "_", anno))
    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(paste("eQTL (AA) lfsr:", signif(eGenes_gwas$AA[num], 2)),
                       paste("eQTL (EA) lfsr:", signif(eGenes_gwas$EA[num], 2)), sep='\n')
    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(anno, gene_id, eqtl_annot, gwas_annot, risk_annot, sep='\n')
    fn = paste("drd2_eqtl_in_gwas_significant_snp", en, sep="_")
    plot_gwas_eqtl(fn, gene_id, variant_id, eqtl_annot,
                   pgc3_a1_same_as_our_counted, OR, title)
}
```


DRD2 junction 5-7
chr11:113412884-113415420(-)
eQTL (AA) lfsr: 0.12
eQTL (EA) lfsr: 0.036
SZ GWAS pvalue: 1e-14
SZ risk allele: T
Residualized Expression
2
1
0
-1
0
GG
1
GT
2
TT
chr11:113503845:G:T
Race
AA
EA



1.3 Session Info

[15]:
```
Sys.time()
proc.time()
options(width = 120)
sessioninfo::session_info()
```

```
[1] "2022-03-08 20:51:05 EST"
```

```
     user    system   elapsed 
11848.491   639.293  4449.132 
```

$platform **$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 <chr>	ondiskversion <chr>	loadedversion <chr>	path <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	3.2.0	3.2.0	/home/jbe
colorspace	colorspace	2.0.3	2.0-3	/home/jbe
crayon	crayon	1.5.0	1.5.0	/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	1.0.8	1.0.8	/home/jbe
ellipsis	ellipsis	0.3.2	0.3.2	/home/jbe
evaluate	evaluate	0.15	0.15	/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	1.0.1	1.0.1	/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	2.1.0	2.1.0	/home/jbe
systemfonts	systemfonts	1.0.4	1.0.4	/home/jbe