ase_comparison_north_america.Rmd

Carlos Martinez Ruiz

23 November 2018

```
## Warning in is.na(x[[i]]): is.na() applied to non-(list or vector) of type
## 'environment'
```

```
##                _
## platform       x86_64-pc-linux-gnu
## arch           x86_64
## os             linux-gnu
## system         x86_64, linux-gnu
## status
## major          3
## minor          4.4
## year           2018
## month          03
## day            15
## svn rev        74408
## language       R
## version.string R version 3.4.4 (2018-03-15)
## nickname       Someone to Lean On
```

```
## [1] "Package ggplot2 version 3.1.0"
## [1] "Package GenomicFeatures version 1.26.4"
## [1] "Package AnnotationDbi version 1.36.2"
## [1] "Package DESeq2 version 1.14.1"
## [1] "Package SummarizedExperiment version 1.4.0"
## [1] "Package Biobase version 2.34.0"
## [1] "Package GenomicRanges version 1.26.4"
## [1] "Package GenomeInfoDb version 1.10.3"
## [1] "Package IRanges version 2.8.2"
## [1] "Package S4Vectors version 0.12.2"
## [1] "Package BiocGenerics version 0.20.0"
## [1] "Package parallel version 3.4.4"
## [1] "Package stats4 version 3.4.4"
## [1] "Package readr version 1.1.1"
## [1] "Package tximport version 1.2.0"
## [1] "Package stats version 3.4.4"
## [1] "Package graphics version 3.4.4"
## [1] "Package grDevices version 3.4.4"
## [1] "Package utils version 3.4.4"
## [1] "Package datasets version 3.4.4"
## [1] "Package methods version 3.4.4"
## [1] "Package base version 3.4.4"
```

Introduction

This script will analyse the variant specific expression for the supergene in six samples of *Solenopsis invicta* polygyne queens (SbSB genotype) from (Wurm et al., 2011). To do so, the RNAseq reads were aligned to the *S.invicta* reference genome (gnG assembly) and the variant specific read counts obtained by running GATK's ASEreadCounter with a VCF file specific for North American populations.

Import variant specific reads into R

The read counts generated by GATK are stored in a csv file, where one of the columns has the read counts for the reference allele (SB) and another one for the alternative (Sb).

```
#Loads the counts by variant generated by GATK. All samples!!
path_to_samples <- "input/renamed_north_america_ase/"
ase_files_ppq <- grep("*csv",
                      list.files(path_to_samples),
                      value = TRUE)
ase_files_ppq <- paste(path_to_samples,
                       ase_files_ppq, sep = "" )


names(ase_files_ppq) <- gsub(pattern     = "(.+/)(Queen[0-9])(.+)",
                             replacement = "\\2",
                             x           = ase_files_ppq)

ase_files_ppq
```

```
##                                                   Queen1
## "input/renamed_north_america_ase/Queen1.na_neutral.csv"
##                                                   Queen2
## "input/renamed_north_america_ase/Queen2.na_neutral.csv"
##                                                   Queen3
## "input/renamed_north_america_ase/Queen3.na_neutral.csv"
##                                                   Queen4
## "input/renamed_north_america_ase/Queen4.na_neutral.csv"
##                                                   Queen5
## "input/renamed_north_america_ase/Queen5.na_neutral.csv"
##                                                   Queen6
## "input/renamed_north_america_ase/Queen6.na_neutral.csv"
```

```
#Load all the files in R as a single table
ase <- do.call(rbind, lapply(ase_files_ppq, read.table, header = TRUE))

#The table loaded contains all ase counts for Queens from the Wurm 2011 data. Each file has different l
#have the same variants.
```

Load variant specific read counts into R

The data loaded into R gives read counts per SNP, rather than per gene. Every SNP needs to be associated to its gene, and the number of read counts per gene averaged across all its SNPs using the median. The position of the genes was taken from the *S.invicta* annotation file (gnG assembly, version GCF_000188075.1). First step, load into R using GRanges the position of all *S.invicta* genes.

```
#Transforms the table loaded into a GRanges object, readable by GenomicRanges and GenomicFeatures.
#The first two parameters define the chromosome (scaffold in this case) and the position of the feature
#(in this case, variants). The other parameters are metadata (they could be whatever) in this case,
#the counts for the reference, the alternative allele and total and the sample to which each read belon

ase_gr <- GRanges(Rle(ase$contig),
                  IRanges(start = ase$position, width = 1),
                  count_ref     = ase$refCount,
                  count_alt     = ase$altCount,
                   tot_count     = ase$totalCount,
                   sample        = gsub("([A-z]+[0-9])(\\.[0-9]+)",
                                        "\\1", rownames(ase)))


#Load the annotation for the gnG assembly of the Solenopsis invicta reference genome.
si_ann <- makeTxDbFromGFF(file = "input/GCF_000188075.1_Si_gnG_genomic.gff",
                          format="gff3")

#Generate a table with the gene names and its position in the reference
gene_ids <- keys(si_ann, "GENEID")
gene_positions <- AnnotationDbi::select(x = si_ann,
                        keys = gene_ids,
                        columns = c("GENEID","TXCHROM", "TXSTART", "TXEND"),
                        keytype = "GENEID")

colnames(gene_positions) <- c("gene","contig", "start", "end")

#Remove the last exon of gene LOC105202834, this gene is misanotated in gnG and it causes mis-assignmen
#The number in the gene positions data frame is replaced with the position of the penultimate exon
gene_positions[gene_positions$gene =="LOC105202834", ]$end <- 220646

#Generates a GRanges object. Same as before, but this time the metadata refers to the gene name directl
gene_positions_gr <- GRanges(Rle(gene_positions$contig),
                                   IRanges(start = gene_positions$start,
                                           end   = gene_positions$end),
                                   gene          = gene_positions$gene)
```

We also want to want which of these genes loaded belong to the supergene, for this the regions of the supergene in the gnG assembly need to be loaded into R. GRanges will be then used to overlap the position of the transcripts with that of the supergene.

```
#Load the regions for gnG
regions_gng <- read.table("input/gng_regions.tab", header = TRUE)

#Generate a GRanges object from regions_gng
regions_ranges <- GRanges(Rle(regions_gng$scaffold),
                          IRanges(start = regions_gng$start,
                                  end   = regions_gng$end),
                          region        = regions_gng$region)
```

Once the positions have been generated, it is just a matter of overlapping all GRanges objects, to obtain variant specific data per gene for loci in the supergene region.

```
#This function merges both GRanges objects (the one with the allele counts
# and the other with the transcript anf gene names) according to their overlapping positions.
#This returns a GRanges object with the allele counts per transcript.

ase_by_gene <- mergeByOverlaps(ase_gr, gene_positions_gr,
                                     ignore.strand = TRUE)

#Extract Transcript positions and ase counts from ase_by_gene

ase_by_gene_df <- as.data.frame(ase_by_gene)
```

```
#Select only relevant fields (keep the position of the variants only, those of the transcripts will yie

ase_by_gene_df <- subset(x= ase_by_gene_df, select = c("ase_gr.seqnames", "ase_gr.start",
                                                       "ase_gr.count_ref", "ase_gr.count_alt",
                                                       "ase_gr.tot_count", "ase_gr.sample",
                                                       "gene_positions_gr.gene"))

#Remove duplicates (arising from different transcripts belonging to the same gene):
ase_by_gene_df <- ase_by_gene_df[!duplicated(ase_by_gene_df), ]


#Generate another GRanges object with the information (includes gene position and name)

ase_by_gene_pos_tmp <- GRanges(Rle(ase_by_gene_df$ase_gr.seqnames),
                           IRanges(start = ase_by_gene_df$ase_gr.start,
                                   width = 1),
                           count_ref     = ase_by_gene_df$ase_gr.count_ref,
                           count_alt     = ase_by_gene_df$ase_gr.count_alt,
                           countTot      = ase_by_gene_df$ase_gr.tot_count,
                           gene          = ase_by_gene_df$gene_positions_gr.gene,
                           sample        = ase_by_gene_df$ase_gr.sample)

#Now, interesct this GRanges object with the recombination positions
regions_ase_tmp <- mergeByOverlaps(ase_by_gene_pos_tmp,
                                   regions_ranges, ignore.strand = TRUE)

#The GRanges object is clunky and contains a lot of information which is redundant and that I don't nee
#This generates a simplified dataframe with only the position, the allele counts and the transcript and

regions_ase_df <- cbind(as.data.frame(regions_ase_tmp$ase_by_gene_pos),
                        as.data.frame(regions_ase_tmp$regions_ranges@elementMetadata))


#The position is not needed, so this is an even more simplified version of the final dataframe.

regions_ase_df  <- regions_ase_df[, c("count_ref", "count_alt",
                                      "countTot", "gene",
                                      "sample","region")]

rm(ase_by_gene_pos_tmp)
rm(regions_ase_tmp)
```

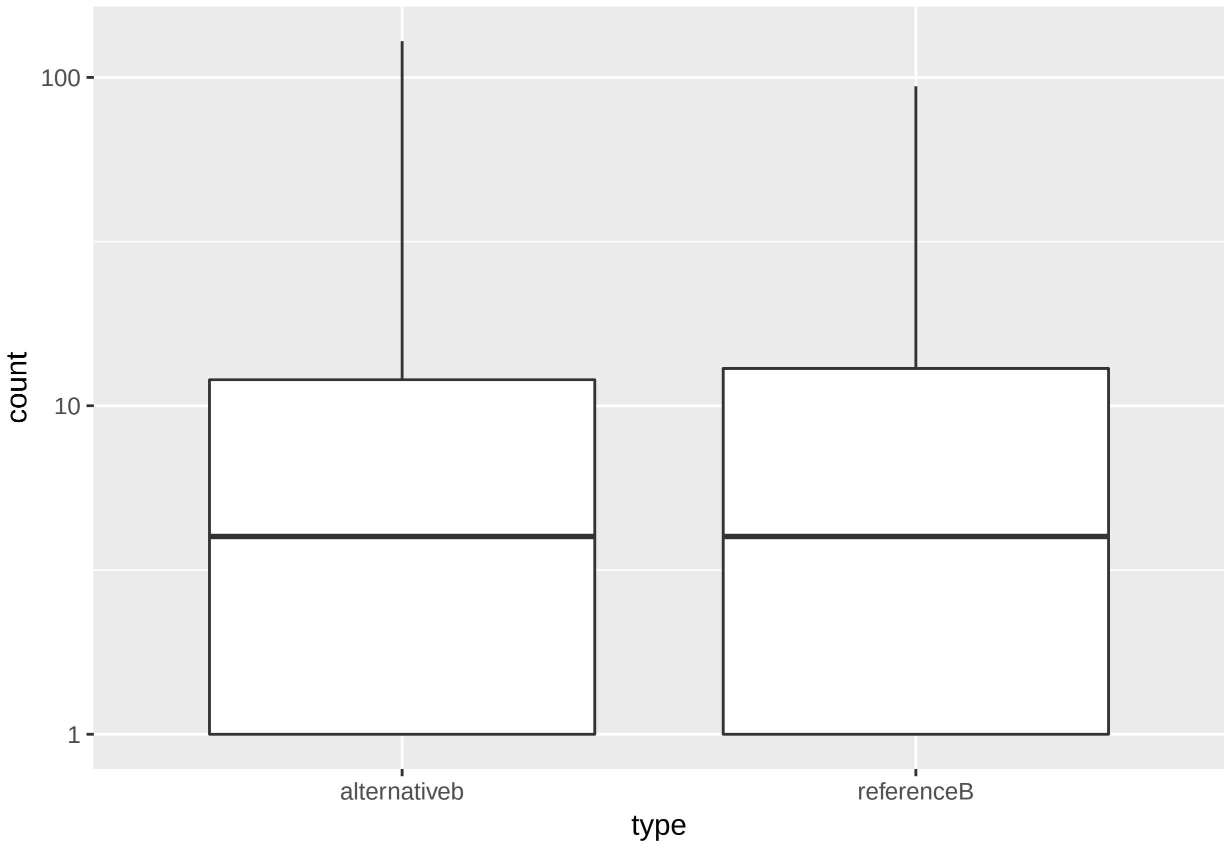
Plot the raw reads per SNP to have a crude look at the variant differences in expression.

```
#This is essentially the same dataset in this case
to_plot_ase_Bb <- subset(regions_ase_df, region == "supergene")

to_plot_ase_Bb_ref <- as.data.frame(to_plot_ase_Bb$count_ref)
to_plot_ase_Bb_ref$type <- rep("referenceB", nrow(to_plot_ase_Bb_ref))
colnames(to_plot_ase_Bb_ref) <- c("count", "type")

to_plot_ase_Bb_alt <- as.data.frame(to_plot_ase_Bb$count_alt)
to_plot_ase_Bb_alt$type <- rep("alternativeb", nrow(to_plot_ase_Bb_alt))
```

```
colnames(to_plot_ase_Bb_alt) <- c("count", "type")

to_plot_ase_Bb <- rbind(to_plot_ase_Bb_alt, to_plot_ase_Bb_ref)


ggplot(to_plot_ase_Bb, aes(x = type, y = count)) + geom_boxplot() +
  scale_y_log10()
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 7002 rows containing non-finite values (stat_boxplot).
```



Now the data is ready to be prepared for the DESeq2 analysis. In the next step, the counts per SNP will be aggregated per gene, and the read counts per read summarised using the median. That is, if a gene has 3 SNPs, the read counts per variant will be estimated as the median read count per variant of all 3 SNPs.

```
#This is essentially the same dataset in this ase, however a gene does get excluded.
#I think this is because they might be on the edges of some of the scaffolds,
#So the SNP may fall within the supergene, but most of the gene must be outside.
#In any case, the gene is present only in 4 samples, so it would be filtered out
#in the next step.
to_analyse <- subset(regions_ase_df, region == "supergene")

#Collapse the read counts per gene (instead of by variant):
to_analyse_agg     <- aggregate(cbind(count_ref, count_alt) ~ gene + sample,
                                data = to_analyse, median)

#Make sure that all samples have the same genes in the dataframe (some samples have genes that are not
```

```r
id.table <- table(to_analyse_agg$gene)
to_analyse_agg <- subset(to_analyse_agg,
                         gene %in% names(id.table[id.table == 6]))
```

The data needs to be parsed to be used in DESeq2. Only genes which have counts for all individuals are included in downstream analyses.

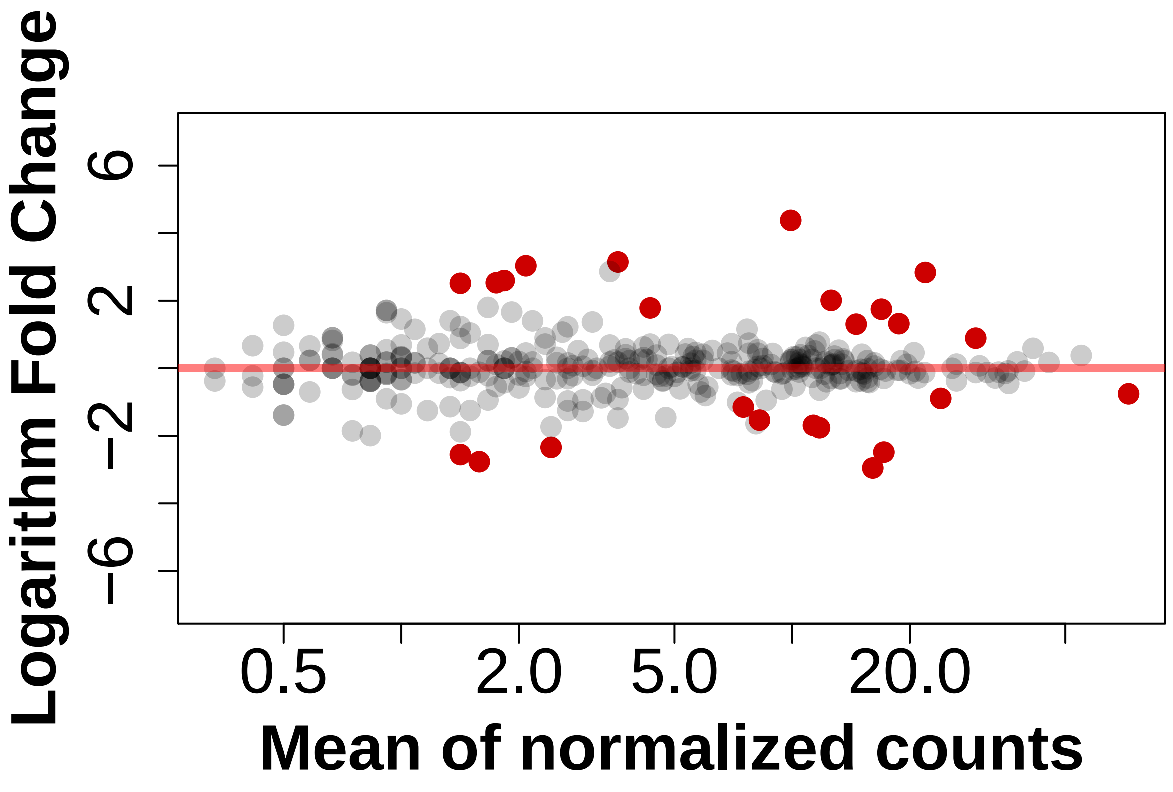
```r
#Generate matrix for B read counts:

to_analyse_B <- matrix(data = to_analyse_agg$count_ref, ncol = length(unique(to_analyse_agg$sample)))

colnames(to_analyse_B) <- paste(unique(to_analyse_agg$sample), "_B", sep = "")
rownames(to_analyse_B) <- unique(to_analyse_agg$gene)
#Select only genes which have 5 read counts or more
B_median_count <- apply(to_analyse_B, 1, median)
B_genes_to_keep <- rownames(to_analyse_B)[B_median_count >=5]

#Generate matrix for b read counts:
to_analyse_b <- matrix(data = to_analyse_agg$count_alt,
                       ncol = length(unique(to_analyse_agg$sample)))

colnames(to_analyse_b) <- paste(unique(to_analyse_agg$sample),
                                "_b", sep = "")
rownames(to_analyse_b) <- unique(to_analyse_agg$gene)

#Select only genes which have 5 read counts or more
b_median_count <- apply(to_analyse_b, 1, median)
b_genes_to_keep <- rownames(to_analyse_b)[b_median_count >=5]
#Merge the two matrixes into a single one, get data ready for DESeq2:---------------------------------

to_analyse_DESeq2_polyq <- cbind(to_analyse_b, to_analyse_B)

#Keep only the genes with 5 read counts or more in either variant
genes_to_keep_total <- unique(c(b_genes_to_keep, B_genes_to_keep))

#to_analyse_DESeq2_polyq <- to_analyse_DESeq2_polyq[genes_to_keep_total, ]
#Generate the colData data frame for DESeq2

allele <- gsub(x = colnames(to_analyse_DESeq2_polyq),
               pattern = "(Queen[0-9]_)([A-z])", replacement = "\\2")
colData_Bb <- data.frame(allele)
colData_Bb$sample <- rep(unique(to_analyse_agg$sample), 2)
rownames(colData_Bb) <- colnames(to_analyse_DESeq2_polyq)
```

DESeq2 analysis

The analysis for the allele specific expression between the SB and Sb variants is performed in DESeq2. Because the variant specific counts that are being compared always come from the same sample (and therefore, the same library), performing normalisation would flatten actual differences. The analysis therefore needs to be performed ensuring that the samples are not normalised (more info herehttp://rpubs.com/mikelove/ase).

```r
#Make sure all count values are integers:
to_analyse_DESeq2_polyq <- round(to_analyse_DESeq2_polyq)
#Load the data into DESeq2
```

```r
dds_Bb <- DESeqDataSetFromMatrix(countData = to_analyse_DESeq2_polyq,
                                 colData = colData_Bb, design = ~ allele)
sizeFactors(dds_Bb) <- rep(1, ncol(to_analyse_DESeq2_polyq))
#Perform the analysis:
dds_Bb_DE <- DESeq(dds_Bb)
#Save the DESeq2 object
save(dds_Bb_DE, file = "results/dds_Sb_vs_SB_north_america.RData")

#Get results:
res_Bb <- results(dds_Bb_DE)#, contrast = c("allele", "B", "b"))

#Let's have a look:

#Create a 'transparent' black color:
trans_black <- adjustcolor("black", alpha.f = 0.2)

par(mar = c(5, 4.6, 4, 2) + 0.1, mfrow = c(1, 1))
plotMA(res_Bb, alpha = 0.05,  ylim = c(-7, 7), xlab = "Mean of normalized counts",
       colNonSig = trans_black, ylab = "Logarithm Fold Change", font.lab = 2,
       cex.axis = 2, cex.lab = 2, cex.main = 2.5, cex = 1.5)
```



Enrichment analyses

Is there a biase towards SB? To check for this, we should expect the median of the ASE LFCs to be significantly positive. THat, is to find more highly expressed genes in SB than in Sb in general.

```r
median(res_Bb$log2FoldChange)
```

```
## [1] 2.332054e-17
```

```r
wilcox.test(res_Bb$log2FoldChange)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  res_Bb$log2FoldChange
## V = 30762, p-value = 0.3782
## alternative hypothesis: true location is not equal to 0
```

Additional tests to check for enrichment: Binomial test for significantly DE genes, and chi squared tests for DE genes in SB and Sb.

```r
#How many DE genes over-expressed in Sb?
over_e_lb <- length(which(res_Bb$log2FoldChange < 0 & res_Bb$padj < 0.05))
#13

#How many DE genes over-expressed in SB?
over_e_bb <- length(which(res_Bb$log2FoldChange > 0 & res_Bb$padj < 0.05))
#16

#Total genes in the supergene region:
total_genes_Bb <- length(res_Bb$pvalue)

#Chi2 to check wether any of the variants in enriched

to_test_chi           <- matrix(c(over_e_lb, total_genes_Bb - over_e_lb, over_e_bb,
                                  total_genes_Bb - over_e_bb), nrow = 2, ncol = 2)
colnames(to_test_chi) <- c("Bb", "BB")
rownames(to_test_chi) <- c("Over-expresed", "Non-DE")

chisq.test(to_test_chi)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  to_test_chi
## X-squared = 0.043177, df = 1, p-value = 0.8354
```

Variant specific data vs social form comparisons

It has been established that the SB alleles tend to be more highly expressed than their Sb counterparts. But of the few genes whose Sb allele is more highly expressed, are those genes also more highly expressed in general in polygyne queens? To test for this, the results from the comparison between social forms needs to be loaded into R again.

```r
load("input/2019-01-08_results_deseq2_queens_mor.RData")
load("input/2019-01-08_results_deseq2_workers_mor.RData")
```

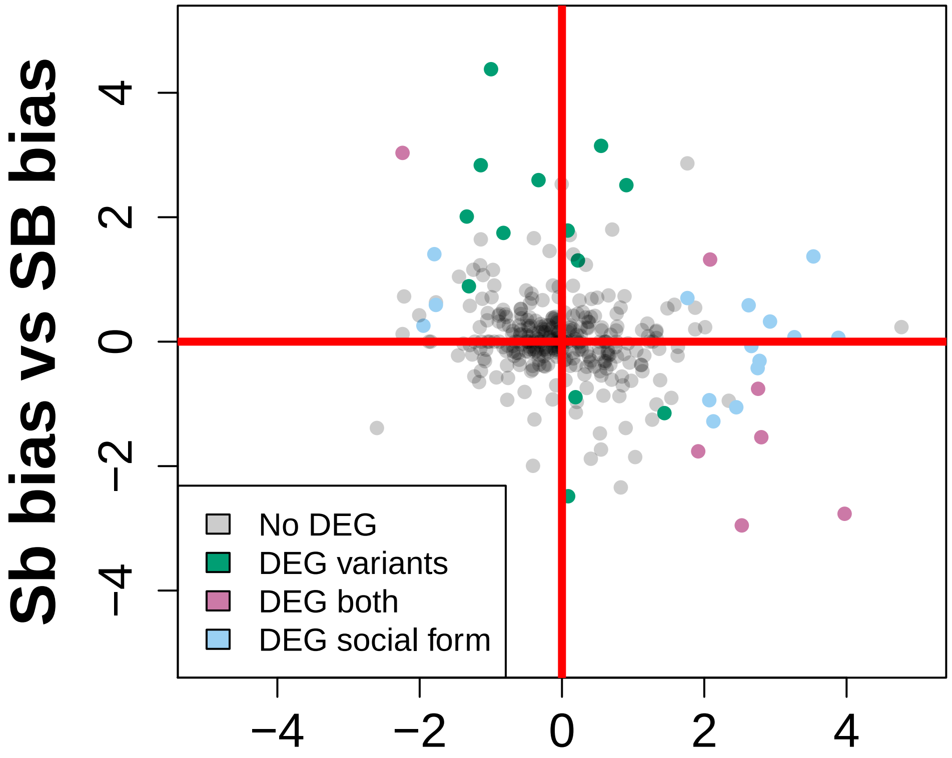
Next, a dataset with both comparisons needs to be prepared, ensuring that all the genes are the same in both datasets and are in the same order.

```r
#Genes names equivalent in the Morandin data
genes_morandin <- rownames(res_de_txi_q)

#Get LFCs for each comparison, making sure that all genes are in the same order
```

```r
lfcs_morandin           <- data.frame(res_de_txi_q[rownames(res_de_txi_q) %in% rownames(res_Bb), ])
lfcs_morandin$gene      <- rownames(res_de_txi_q[rownames(res_de_txi_q) %in% rownames(res_Bb), ])
lfcs_morandin$gene      <- as.factor(lfcs_morandin$gene)
lfcs_morandin           <- data.frame(lfcs_morandin$gene, lfcs_morandin$log2FoldChange, lfcs_morandin$pa
colnames(lfcs_morandin) <- c("gene", "lfcs_mor", "padj_mor")

lfcs_ase           <- data.frame(res_Bb[rownames(res_Bb) %in% lfcs_morandin$gene, ])
lfcs_ase$gene      <- rownames(res_Bb[rownames(res_Bb) %in% lfcs_morandin$gene, ])
lfcs_ase$gene      <- as.factor(lfcs_ase$gene)
lfcs_ase           <- data.frame(lfcs_ase$gene, lfcs_ase$log2FoldChange, lfcs_ase$padj)
colnames(lfcs_ase) <- c("gene", "lfcs_ase", "padj_ase")

#Make sure everything has the same order
lfcs_ase   <- lfcs_ase[order(lfcs_ase$gene), ]
lfcs_morandin <- lfcs_morandin[order(lfcs_morandin$gene), ]
if (!identical(lfcs_ase$gene, lfcs_morandin$gene)) {
  warning("The genes are not the same in both dataframes!")
}
```

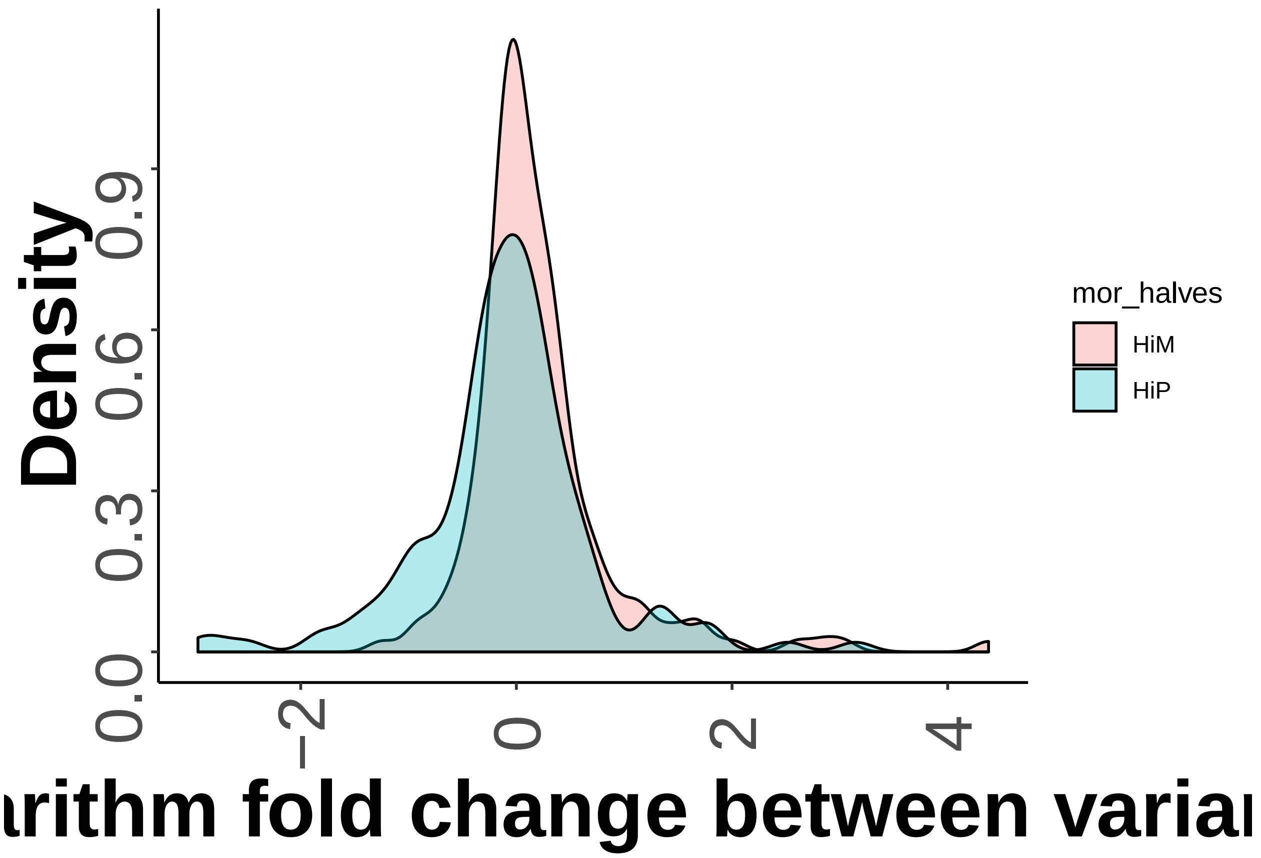
Plot the LFCs for both comparisons. This plot shows the LFCs for the comparisons between social forms in queens (from the Morandin et al., 2016 dataset) in the x axis and the LFCs for the comparisons between the SB and Sb variants in polygyne queens (from the Wurm et al., 2011 dataset) in the y axis. The plot also shows the genes which are significantly differentially expressed (according to DESeq2) between social forms only (light blue dots), between variants only (green dots) or in both comparisons (purple dots).

```r
#Get a colour vector for all values based on the significance of the padj
lfcs_joint <- data.frame(cbind(lfcs_ase, lfcs_morandin), stringsAsFactors = FALSE)


lfcs_joint$cols <- trans_black
lfcs_joint$cols[lfcs_joint$padj_ase < 0.05 & lfcs_joint$padj_mor >= 0.05] <- "#009E73"
lfcs_joint$cols[lfcs_joint$padj_ase >= 0.05 & lfcs_joint$padj_mor < 0.05] <- "#9ad0f3"
lfcs_joint$cols[lfcs_joint$padj_ase < 0.05 & lfcs_joint$padj_mor < 0.05] <- "#CC79A7"


#Plot the thing(and save in png):
#png("DC_sinvicta_2016-17 / 05-0216-CD-analyses / ase_plot.png",
#    width = 10.6, height = 6.3, units = "in", res = 800)
par(mfrow = c(1, 1), mar = c(5, 4.6, 4, 2) + 0.1, pin = c(4, 3.5))
plot(lfcs_morandin$lfcs_mor, lfcs_ase$lfcs_ase, xlab = "Single-queen bias vs Multiple-queen bias",
     ylab = "Sb bias vs SB bias", xlim = c(-5, 5), ylim = c(-5, 5), font.lab = 2, cex.axis = 1.5, cex.la
     cex.main = 2, pch = 16, col = lfcs_joint$cols)
legend("bottomleft", legend = c("No DEG", "DEG variants", "DEG both", "DEG social form"),
       fill = unique(lfcs_joint$cols))
abline(h = 0, col = "red", lwd = 4)
abline(v = 0, col = "red", lwd = 4)
```



Test for enrichment in Sb for polygyne biased genes, via a chi squared test, and by comparing the LFCs distributions across variants in highly and lowly expressed genes in polygyne queens through a Kolmogorov-Smirnov test. The plot of the distributions can be generated by un-commenting lines 417 to 424.

```
####Compare the ASE distributions across a Mono vs Poly gradient
###### Select the Morandin LFCs from 2 sections of the dataset
#Get rid of the NAs first
lfcs_joint_nona <- lfcs_joint[!is.na(lfcs_joint$lfcs_mor), ]
lfcs_joint_nona <- lfcs_joint_nona[!is.na(lfcs_joint_nona$padj_ase), ]
lfcs_joint_nona <- lfcs_joint_nona[!is.na(lfcs_joint_nona$padj_mor), ]


#Sort the new dataset by morandin LFCs
lfcs_joint_mor_sorted <- lfcs_joint_nona[order(lfcs_joint_nona$lfcs_mor), ]

#Using only two halves (M vs P)
lfcs_joint_mor_sorted$mor_halves <- NA

lfcs_joint_mor_sorted$mor_halves[1:round((1 / 2 * nrow(lfcs_joint_mor_sorted)))] <- "HiM"
lfcs_joint_mor_sorted$mor_halves[(round(1 / 2 * nrow(lfcs_joint_mor_sorted)) + 1):nrow(lfcs_joint_mor_s
lfcs_joint_mor_sorted$mor_halves[lfcs_joint_mor_sorted$lfcs_mor < 0] <- "HiM"
lfcs_joint_mor_sorted$mor_halves[lfcs_joint_mor_sorted$lfcs_mor > 0] <- "HiP"

#Now it's just a matter of comparing the B vs b distributions of LFCs between quarters
ggplot(lfcs_joint_mor_sorted, aes(x = lfcs_ase, fill = mor_halves)) + geom_density(alpha = 0.3) + theme_
  labs(x = "Logarithm fold change between variants", y = "Density") +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")) +
  theme(axis.title.x = element_text(face = "bold", size = 30),
        axis.text.x  = element_text(angle = 90, vjust = 0.5, size = 25)) +
  theme(axis.title.y = element_text(face = "bold", size = 30),
        axis.text.y  = element_text(angle = 90, vjust = 0.5, size = 25))
```



```
#Kolmogorov-Smirnov tests between quarters to see how different the distributions are
ks.test(lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves == "HiM"],
        lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves == "HiP"])
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves ==  and lfcs_joint_mor_sorted
## D = 0.24713, p-value = 0.0001488
## alternative hypothesis: two-sided
```

```
#Compare medians using Mann Whitney
wilcox.test(lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves == "HiM"],
            lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves == "HiP"])
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lfcs_joint_mor_sorted$lfcs_ase[lfcs_joint_mor_sorted$mor_halves ==  and lfcs_joint_mor_sorted
## W = 15004, p-value = 0.0004045
## alternative hypothesis: true location shift is not equal to 0
```

```
#From all Sb over-expressed genes, what is the proportion of polygyne genes? Is it different in SB?
#You need to use the clean df without NAs in the Morandin LFCs. NA in the Morandin LFCs indicate
#that the count for that gene was 0 for Morandin but not for the 2011 queen data, it is thus
```

```
#an incongruence and those genes should be removed.

#Total over-expressed genes in Sb
tot_lb <- length(which(lfcs_joint_mor_sorted$lfcs_ase < 0 & lfcs_joint_mor_sorted$padj_ase < 0.05))
#Polygyne biased genes in Sb
pol_lb <- length(which(lfcs_joint_mor_sorted$lfcs_ase < 0 & lfcs_joint_mor_sorted$padj_ase < 0.05 &
                         lfcs_joint_mor_sorted$lfcs_mor > 0 & lfcs_joint_mor_sorted$padj_mor < 0.05))

#Total over-expressed genes in SB
tot_bb <- length(which(lfcs_joint_mor_sorted$lfcs_ase > 0 & lfcs_joint_mor_sorted$padj_ase < 0.05))
#Polygyne biased genes in SB
pol_bb <- length(which(lfcs_joint_mor_sorted$lfcs_ase > 0 & lfcs_joint_mor_sorted$padj_ase < 0.05 &
                         lfcs_joint_mor_sorted$lfcs_mor > 0 & lfcs_joint_mor_sorted$padj_mor < 0.05))

#Chi2 test

to_test_chi_sb           <- matrix(c(pol_lb, tot_lb, pol_bb, tot_bb), nrow = 2, ncol = 2)
colnames(to_test_chi_sb) <- c("Sb", "SB")
rownames(to_test_chi_sb) <- c("Polygyne_only", "Total")

chisq.test(to_test_chi_sb)
```

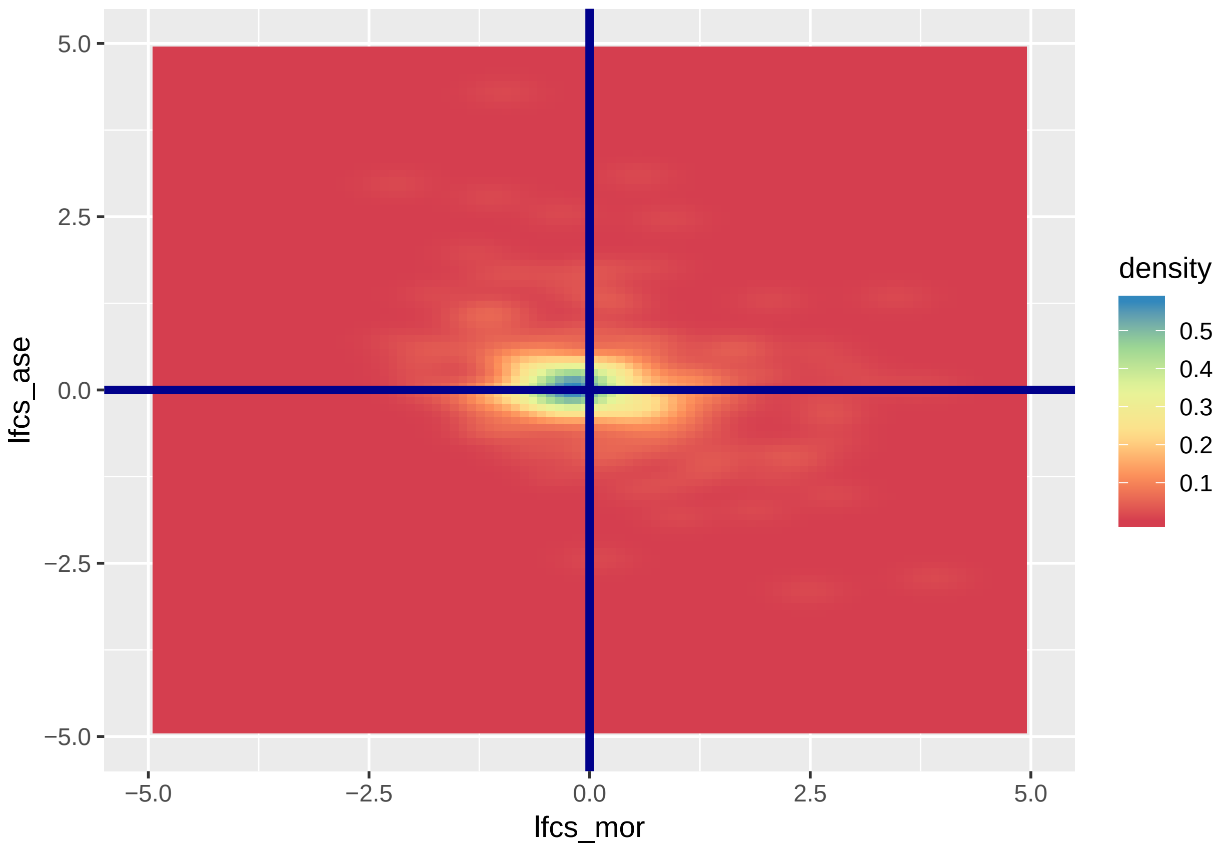
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  to_test_chi_sb
## X-squared = 1.95, df = 1, p-value = 0.1626
```

```
#Try with Fisher test
fisher.test(to_test_chi_sb)
```

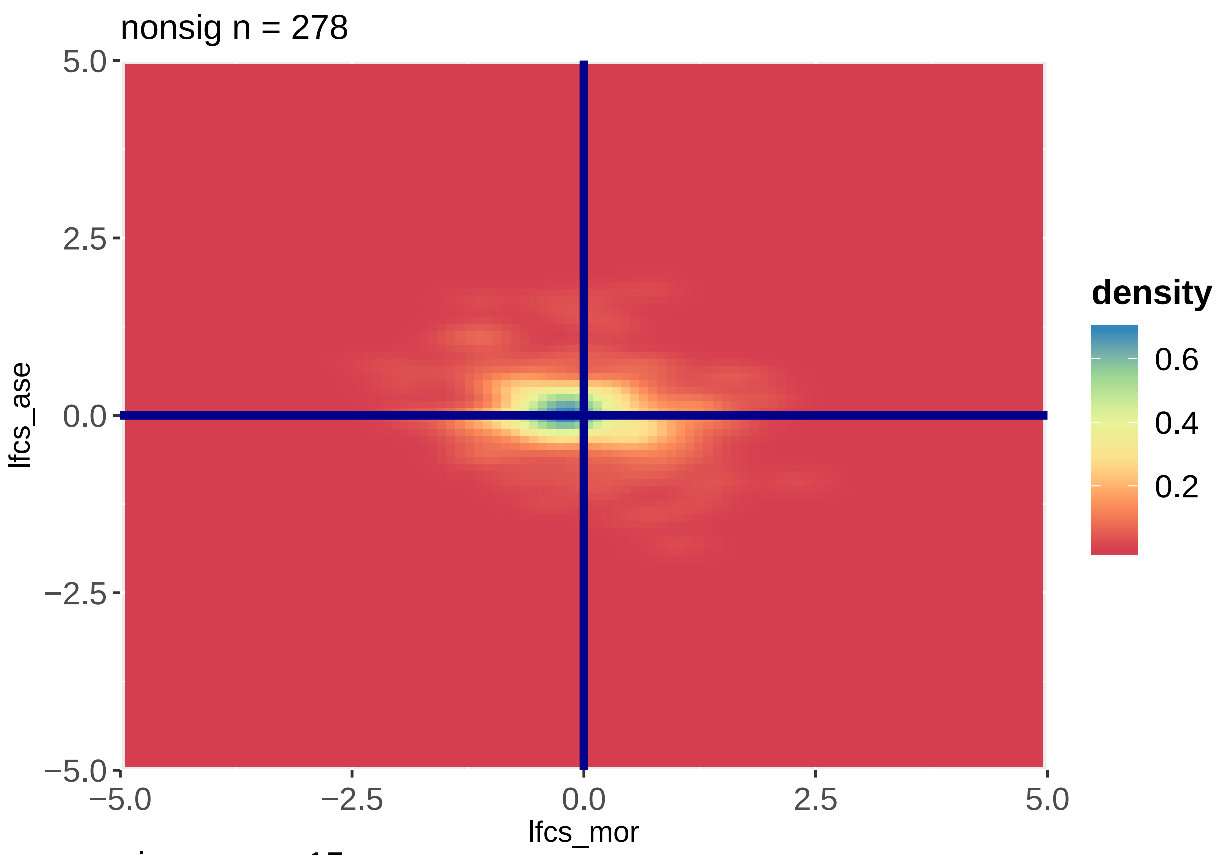
```
##
##  Fisher's Exact Test for Count Data
##
## data:  to_test_chi_sb
## p-value = 0.1602
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.6115474 382.6416315
## sample estimates:
## odds ratio
##   6.953068
```

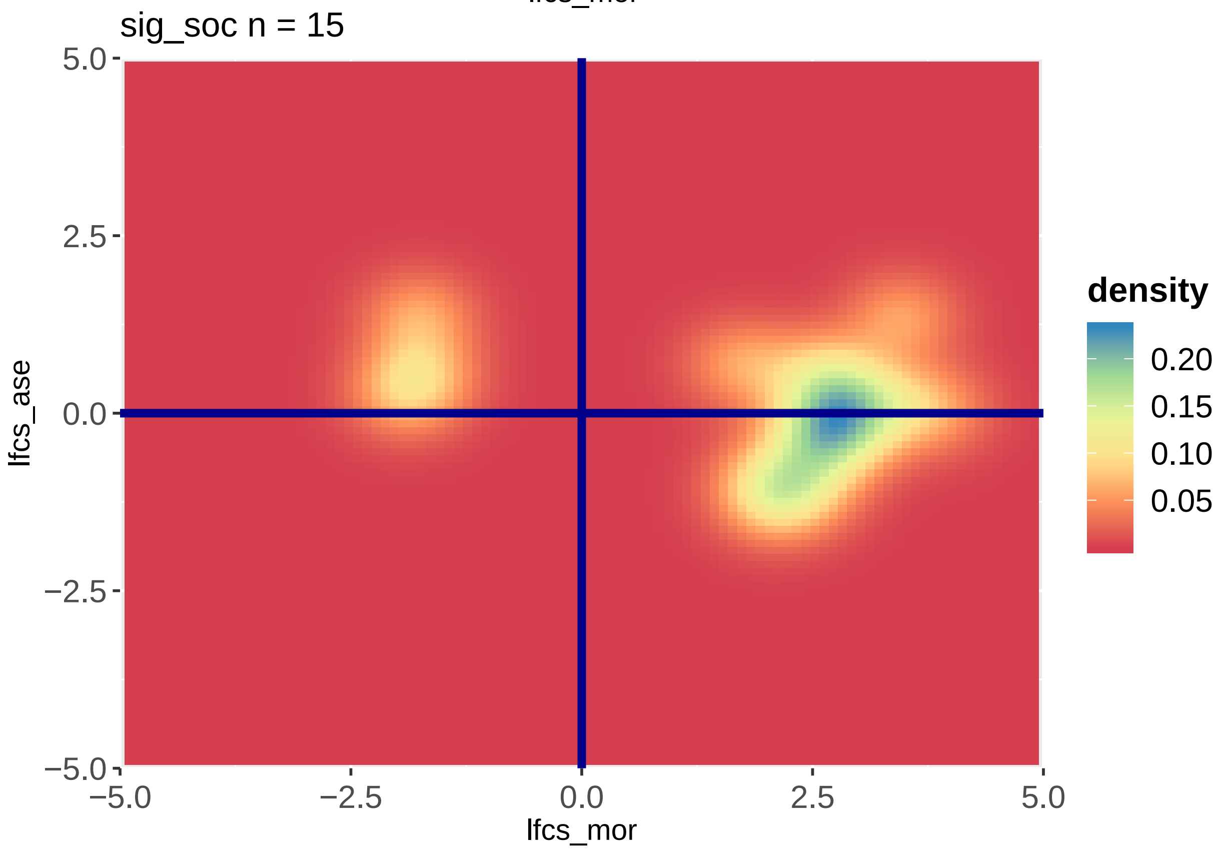
Make a density plot with social-variant information

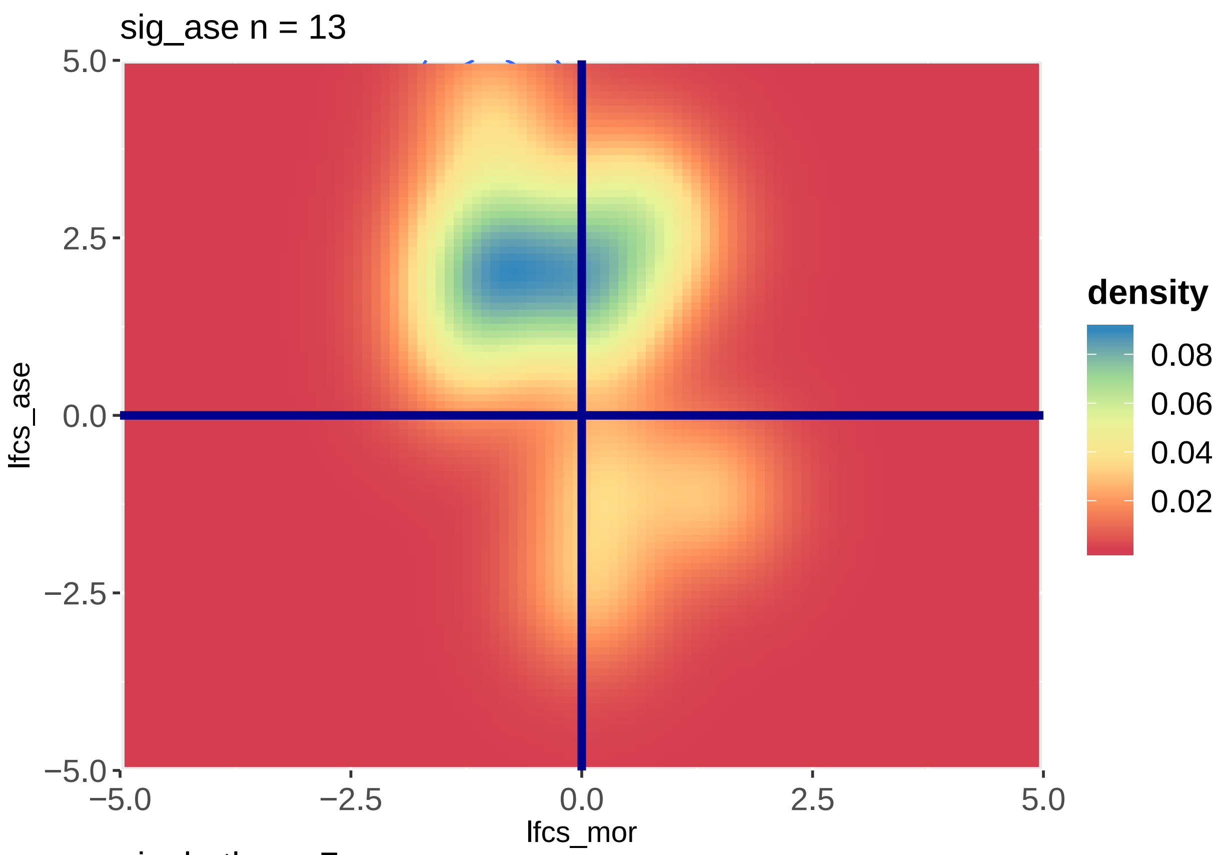
```
#Plot all the data together
ggplot(lfcs_joint_nona, aes(x = lfcs_mor, y = lfcs_ase)) + geom_density_2d() +
  stat_density_2d(aes(fill = ..density..), geom = "raster", contour = FALSE) +
   scale_fill_distiller(palette= "Spectral", direction=1) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  xlim(-5, 5) + ylim(-5, 5) +
  geom_hline(yintercept = 0, color = "darkblue", size = 1.5) +
  geom_vline(xintercept = 0, color = "darkblue", size = 1.5)
```

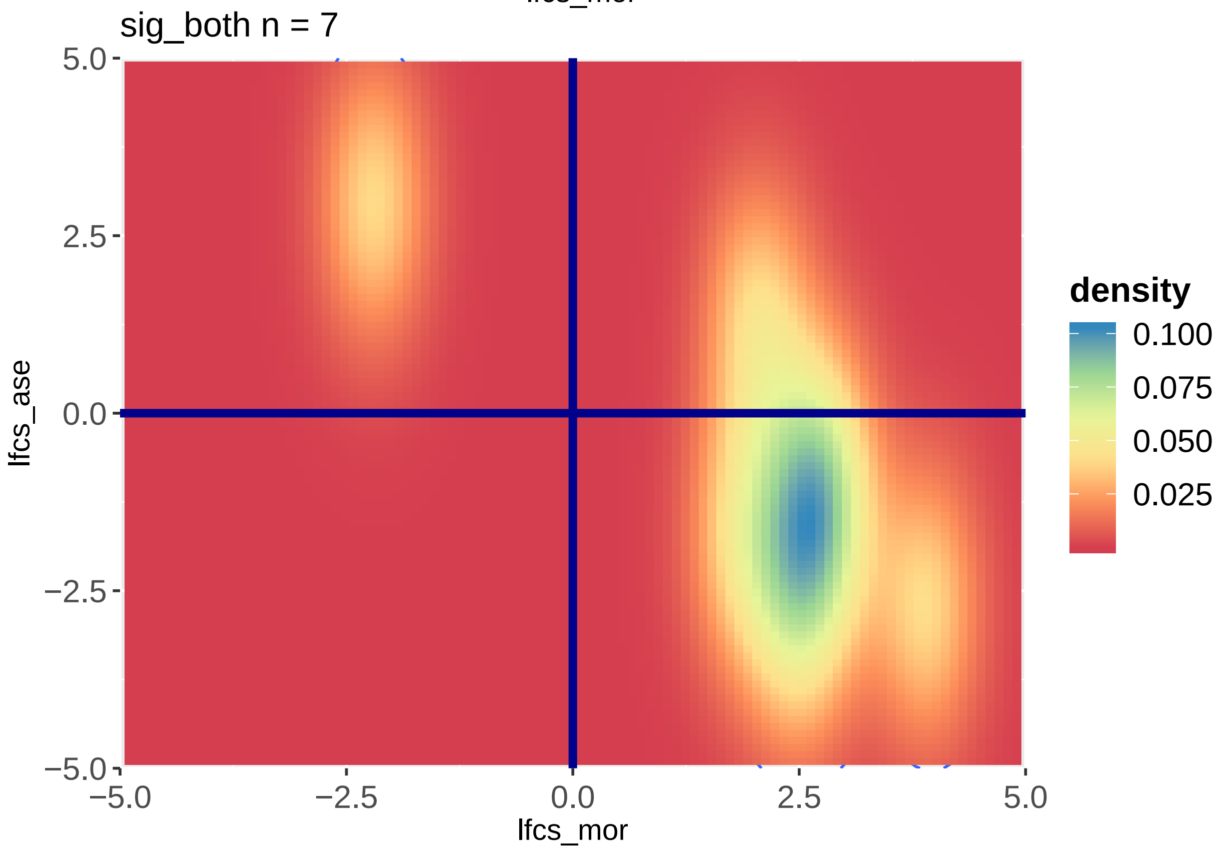


```r
#Run with different levels of significance
#Select only values which are non significant
lfcs_joint_nonsig <- subset(lfcs_joint_nona, subset = padj_ase > 0.05 & padj_mor > 0.05)
#Select only values which are significant between social forms
lfcs_joint_sig_soc <- subset(lfcs_joint_nona, subset = padj_mor < 0.05 & padj_ase > 0.05)
#Select only values which are significant between variants
lfcs_joint_sig_ase <- subset(lfcs_joint_nona, subset = padj_ase < 0.05 & padj_mor > 0.05)
#Select only values which are significant in both comparisons
lfcs_joint_sig_both <- subset(lfcs_joint_nona, subset = padj_ase < 0.05 & padj_mor < 0.05)

#Generate all plots:
sig_levels <- c("nonsig", "sig_soc", "sig_ase", "sig_both")
for(sig in sig_levels){
  dataset <- get(paste0("lfcs_joint", "_", sig))
  to_plot <- ggplot(dataset, aes(x = lfcs_mor, y = lfcs_ase)) + geom_density_2d() +
               stat_density_2d(aes(fill = ..density..), geom = "raster", contour = FALSE) +
               scale_fill_distiller(palette= "Spectral", direction=1) +
               scale_x_continuous(expand = c(0, 0), limits = c(-5,5)) +
               scale_y_continuous(expand = c(0, 0), limits = c(-5,5)) +
               geom_hline(yintercept = 0, color = "darkblue", size = 1.5) +
               geom_vline(xintercept = 0, color = "darkblue", size = 1.5) +
               theme(axis.text.x = element_text(vjust = 0.5, size = 12), axis.text.y = element_text(vju
               legend.title = element_text(size=13, face="bold"), legend.text = element_text(size=12))
               ggtitle(paste(sig, "n =", nrow(dataset)))
  print(to_plot)
}
```


nonsig n = 278
5.0
2.5
0.0
−2.5
−5.0
lfcs_ase
−5.0
−2.5
0.0
2.5
5.0
lfcs_mor
density
0.6
0.4
0.2


sig_soc n = 15
5.0
2.5
0.0
−2.5
−5.0
lfcs_ase
−5.0
−2.5
0.0
2.5
5.0
lfcs_mor
density
0.20
0.15
0.10
0.05


sig_ase n = 13
5.0
2.5
0.0
−2.5
−5.0
lfcs_ase
−5.0
−2.5
0.0
2.5
5.0
lfcs_mor
density
0.08
0.06
0.04
0.02


sig_both n = 7
5.0
2.5
0.0
−2.5
−5.0
lfcs_ase
−5.0
−2.5
0.0
2.5
5.0
lfcs_mor
density
0.100
0.075
0.050
0.025

Gene-specific dosage compensation

Some of the expression differences between variants could be due to gene-specifc dosage compensation. If this is the case, the pattern should be low expression in Sb (and hence high expression in SB) and no differential expression between social forms. Overall this should be transalted in a general bias towards SB expression once the significant genes in both comparisons have been removed. This seems to be the case in the 2D density plot, but this needs further testing. A chi2 test or a similar test to that of the Sb enrichment should do the trick.

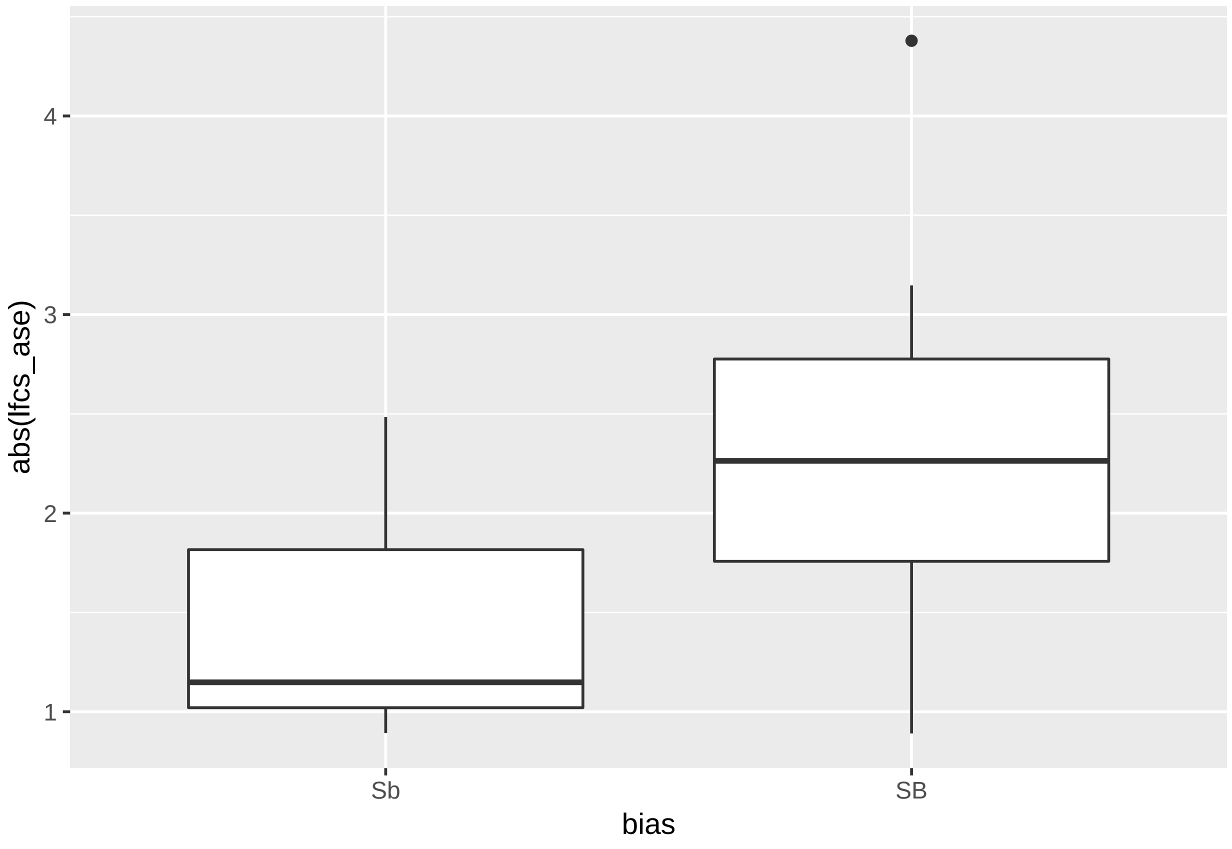
```
#Median lfcs should be biased towards SB
median(lfcs_joint_sig_ase$lfcs_ase)
```

```
## [1] 1.785649
```

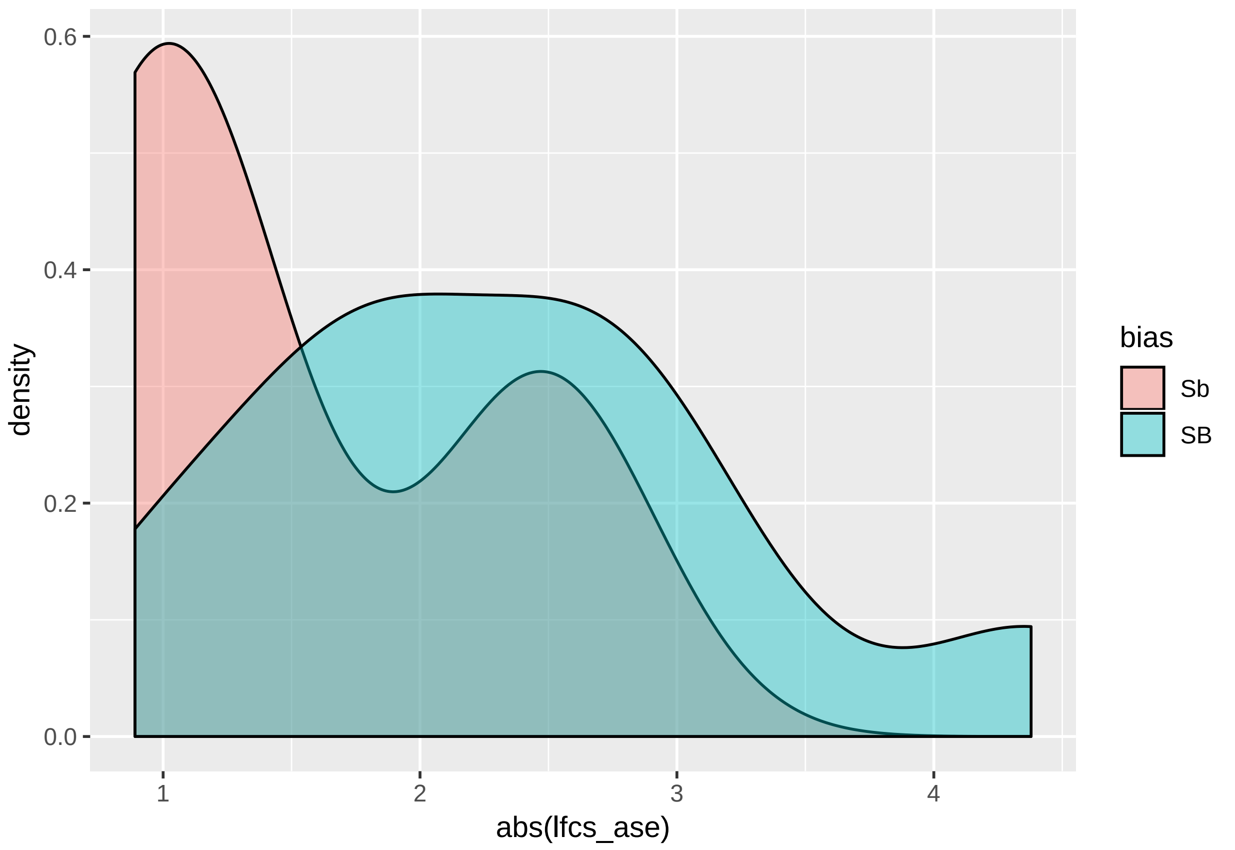
```
wilcox.test(lfcs_joint_sig_ase$lfcs_ase)
```

```
##
##  Wilcoxon signed rank test
##
## data:  lfcs_joint_sig_ase$lfcs_ase
## V = 78, p-value = 0.02148
## alternative hypothesis: true location is not equal to 0
```

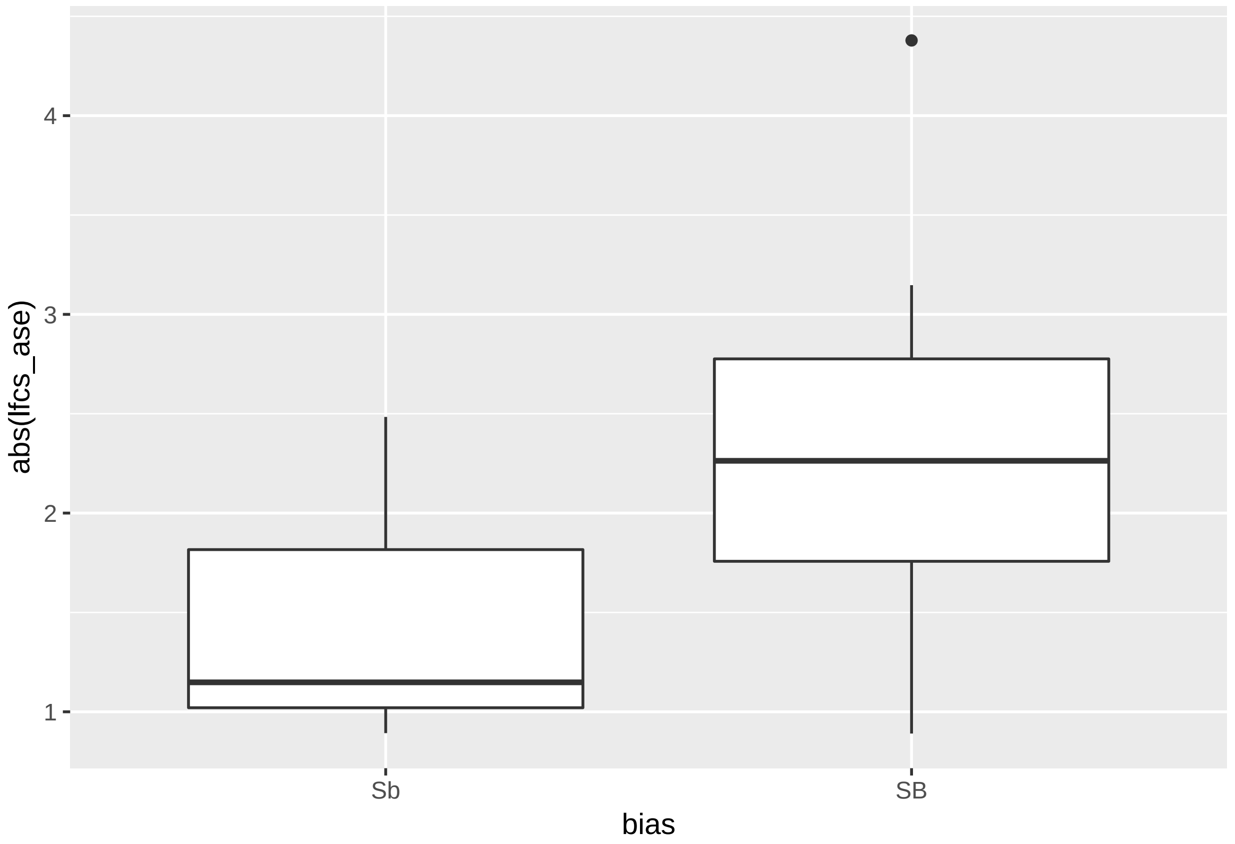
```
#Differences in lfc medians
lfcs_joint_sig_ase$bias <- ifelse(lfcs_joint_sig_ase$lfcs_ase > 0, "SB", "Sb")
ggplot(data = lfcs_joint_sig_ase, aes(x = bias, y = abs(lfcs_ase))) + geom_boxplot()
```



```
ggplot(data = lfcs_joint_sig_ase, aes(x = abs(lfcs_ase), fill = bias)) + geom_density(alpha = 0.4)
```



```
lfcs_joint_sig_ase$bias <- ifelse(lfcs_joint_sig_ase$lfcs_ase > 0, "SB", "Sb")
ggplot(data = lfcs_joint_sig_ase, aes(x = bias, y = abs(lfcs_ase))) + geom_boxplot()
```



```r
#Bimoial test
binom.test(nrow(lfcs_joint_sig_ase[lfcs_joint_sig_ase$bias == "SB", ]), nrow(lfcs_joint_sig_ase))
```

```
##
##  Exact binomial test
##
## data:  nrow(lfcs_joint_sig_ase[lfcs_joint_sig_ase$bias == "SB", ]) and nrow(lfcs_joint_sig_ase)
## number of successes = 10, number of trials = 13, p-value = 0.09229
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4618685 0.9496189
## sample estimates:
## probability of success
##              0.7692308
```

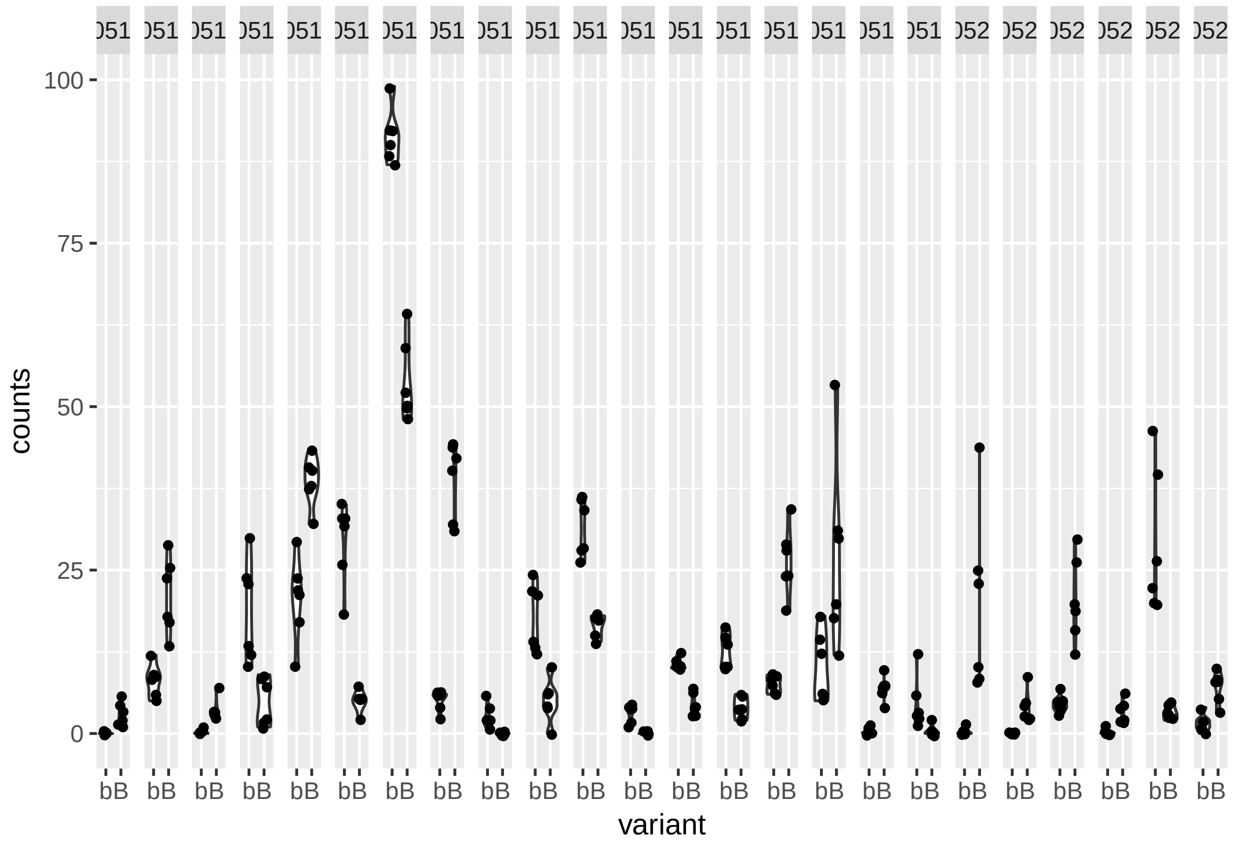
Both the wilcoxon test for the median and the binomial test are significant. Genes which are only differentially expressed between variants (but not between social forms) tend to be more highly expressed in the SB variant. This points at gene-level dosage compensation.

Plot all significant genes to check variant effect visually

DESeq2 gives a lsit of differentially expressed genes between variants. Some of those are based on very low read counts on average. The following plot can be use to assess visually those differences.

```r
#Select significant genes
sig_genes <- rownames(res_Bb[which(res_Bb$padj < 0.05), ])
counts_specific_genes <- counts(dds_Bb, norm = TRUE)[sig_genes, ]
```

```
specific_genes <- rep(rownames(counts_specific_genes), ncol(counts_specific_genes))
samples <- rep(colnames(counts_specific_genes), each = nrow(counts_specific_genes))
variant <- gsub(pattern = ".+_", replacement = "", x = samples)

specific_genes_parsed <- data.frame(c(counts_specific_genes), specific_genes, samples, variant)

colnames(specific_genes_parsed) <- c("counts", "genes", "samples", "variant")

ggplot(data = specific_genes_parsed, aes(x = variant, y = counts)) + geom_violin() + facet_grid(. ~ gene
  geom_jitter(shape=16, position=position_jitter(0.2))
```


)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)51
)52
)52
)52
)52
)52
)52
counts
bB
variant