Prosopis cineraria project

Project report

1. Workflow analysis

1.1

2. Transcriptome analysis

2.1. FASTQC quality analysis

```
## folders with raw data are saved in ../data/raw_data

for i in tree_3 tree_4 tree_5;
do

#make out folders

  mkdir -p ../data/raw_data/${i}_quality

  for k in ../data/raw_data/${i}/*.gz; do
  echo ${k}

  fastqc ${k} -o ../data/raw_data/${i}_quality/ #run fastqc in every sample saved in prosipis samples and sav

  done
done

#made multiqc analysis in tree_3_quality tree_4_quality tree_5_quality directories

for i in tree_3 tree_4 tree_5; do

cd ../data/raw_data/${i}_quality/

multiqc .  #run multiqc analysys inside folfers

done
```

2.2. Plots
2.3. Trimming
Trimmomatic 0.39 command line

```
trimmomatic PE -threads 16 -phred33 -trimlog trimlog.txt infiles.fastq outfiles.fastq ILLUMINACLIP:TruSeq3-PE
LEADING:15 TRAILING:15 MINLEN:75 SLIDINGWINDOW:4:25
```

2.4. Plots

3. Count matrix

A local aligment was performed using Bowtie2 v.2.4.2

```
bowtie2 --local --no-unal -p 16 -x prosopis_index
        -1 input_1P.fastq  -2 input_2P.fastq
        -S out.sam
```

The *Prosopis* reference genome was used for sequence mapping
The matrix of row counts was extracted using *featureCounts* from Subread 2.0.1 program. But first we convert the gff file in gft format file with GffRead 0.12.7 package

```
gffread annotation.gff annotation.gft
```

```
featureCounts -T 4 -t "exon" -g "transcript_id" -O -a PC_final_gene_all_function_stringtie.gtf -o prosopis_co
```

```
-g <string> Specify attribute type in GTF annotation. 'gene_id' by
                      default. Meta-features used for read counting will be
                      extracted from annotation using the provided value

-t <string>     Specify feature type in GTF annotation. 'exon' by
                      default. Features used for read counting will be
                      extracted from annotation using the provided value.
-O  Assign reads to all their overlapping meta-features (or
                      features if -f is specified).
```

3.1. Summary statistics of the count

```
|| Load annotation file PC_final_gene_all_function_stringtie.gtf ...          ||
||    Features : 345371                                                       ||
||    Meta-features : 77218                                                   ||
||    Chromosomes/contigs : 2226                                              ||
||                                                                            ||
|| Process BAM file Ghaf12DT_002_CGATGT_L007_P.sorted.bam...                   ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 39977207                                             ||
||    Successfully assigned alignments : 26360536 (65.9%)                     ||
||    Running time : 0.66 minutes                                             ||
||                                                                            ||
|| Process BAM file Ghaf2DT_005_ACAGTG_L007_P.sorted.bam...                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 39456863                                             ||
||    Successfully assigned alignments : 30636388 (77.6%)                     ||
||    Running time : 0.68 minutes                                             ||
||                                                                            ||
|| Process BAM file Ghaf4DT_006_GCCAAT_L007_P.sorted.bam...                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 32763492                                             ||
||    Successfully assigned alignments : 25829185 (78.8%)                     ||
||    Running time : 0.57 minutes                                             ||
||                                                                            ||
|| Process BAM file Ghaf6DT_007_CAGATC_L007_P.sorted.bam...                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 39905728                                             ||
||    Successfully assigned alignments : 31348056 (78.6%)                     ||
||    Running time : 0.68 minutes                                             ||
||                                                                            ||
|| Process BAM file Ghaf8DT_009_GATCAG_L007_P.sorted.bam...                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 35868644                                             ||
||    Successfully assigned alignments : 27169439 (75.7%)                     ||
||    Running time : 0.62 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-10b_P.sorted.bam...                                 ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 77068406                                             ||
||    Successfully assigned alignments : 34043083 (44.2%)                     ||
||    Running time : 1.02 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-12_P.sorted.bam...                                  ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 72859471                                             ||
||    Successfully assigned alignments : 35016725 (48.1%)                     ||
||    Running time : 1.00 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-2b_P.sorted.bam...                                  ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 87574928                                             ||
||    Successfully assigned alignments : 41407346 (47.3%)                     ||
||    Running time : 1.15 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-4b_P.sorted.bam...                                  ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 81541196                                             ||
||    Successfully assigned alignments : 43583422 (53.4%)                     ||
||    Running time : 1.11 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-6_P.sorted.bam...                                   ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 79045673                                             ||
||    Successfully assigned alignments : 41045983 (51.9%)                     ||
||    Running time : 1.11 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT3-8_P.sorted.bam...                                   ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 78998369                                             ||
||    Successfully assigned alignments : 33000361 (41.8%)                     ||
||    Running time : 1.06 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT4-10b_P.sorted.bam...                                 ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 76756525                                             ||
||    Successfully assigned alignments : 38778464 (50.5%)                     ||
||    Running time : 1.01 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT5-4b_P.sorted.bam...                                  ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 83151642                                             ||
||    Successfully assigned alignments : 44955784 (54.1%)                     ||
||    Running time : 1.07 minutes                                             ||
||                                                                            ||
|| Process BAM file PCDT5-8b_P.sorted.bam...                                  ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 71987791                                             ||
||    Successfully assigned alignments : 38389591 (53.3%)                     ||
||    Running time : 0.97 minutes                                             ||
||                                                                            ||
|| Process BAM file PDT5_10_P.sorted.bam...                                   ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 66364319                                             ||
||    Successfully assigned alignments : 35564759 (53.6%)                     ||
||    Running time : 0.90 minutes                                             ||
||                                                                            ||
|| Process BAM file PDT5_2_P.sorted.bam...                                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 69764299                                             ||
||    Successfully assigned alignments : 33399128 (47.9%)                     ||
||    Running time : 0.92 minutes                                             ||
||                                                                            ||
|| Process BAM file PDT5_6_P.sorted.bam...                                    ||
||    Paired-end reads are included.                                          ||
||    Assign alignments to features...                                        ||
||    Total alignments : 78887271                                             ||
||    Successfully assigned alignments : 40786816 (51.7%)                     ||
||    Running time : 1.06 minutes
```

In order to get a better overview of the mapping results we run Qualimap

```
qualimap multi-bamqc -d data.txt -gff PC_final_gene_all_function_stringtie.gtf  -outdir stats/  -outformat PD
```

```
-d,--data <arg>               File describing the input data. Format of the
                              file is a 2-column tab-delimited table.
                              Column 1: sample name
                              Column 2: either path to the BAM QC result or
                              path to BAM file (-r mode)
-gff,--feature-file <arg>     Only for -r mode. Feature file with regions of
                              interest in GFF/GTF or BED format

-outdir <arg>                 Output folder for HTML report and raw data.


-outformat <arg>              Format of the ouput report (PDF or HTML, default
                              is HTML).
-r,--run-bamqc                Raw BAM files are provided as input. If this
                              option is activated BAM QC process first will be
                              run for each sample, then multi-sample analysis
                              will be performed.
```

4. PCA

In order to evaluate the data we performed a PCA analysis in R. The results shows that in global sense there are not changes in the transcriptome respect the moth of the year, due to the samples are very similar between them.

Preeliminary this can be because we are comparing same species same tissue. In other hand, the tree_3 is very different to the other two.

```
library(tidyverse)
library(ggrepel)
library(dplyr)

#set workdir
setwd("~/Documentos/Prosopis_project/bin/")
#load data
count_matrix<-read.table("../out/count_matrix/prosopis_count_matrix.txt", header = TRUE,
#load metadata
meta <- read.table("../metadata/meta.txt", header = T)
#change colnames
colnames(count_matrix) <- c("Gene_id","Chr","Start","End","Strand","Length",
                            "Ghaf12DT_002_CGATGT_L007","Ghaf2DT_005_ACAGTG_L007",
                            "Ghaf4DT_006_GCCAAT_L007","Ghaf6DT_007_CAGATC_L007",
                            "Ghaf8DT_009_GATCAG_L007","PCDT3.10b","PCDT3.12","PCDT3.2b",
                            "PCDT3.4b","PCDT3.6","PCDT3.8","PCDT4.10b","PCDT5.4b",
                            "PCDT5.8b","PDT5_10","PDT5_2","PDT5_6")

count_matrix <- count_matrix[,c(1, 7:23)]

pp <- as.data.frame(as.matrix(t(count_matrix)))

colnames(pp) <- pp[1,]

pp <- pp[-1,]
```

```r
pp <- as.data.frame(lapply(pp, as.numeric))
#eliminate all columns with colsum = 0
new_pp <- pp[, which(colSums(pp) != 0)]
#write the table without columns with 0
write.csv(new_pp, "../out/prub_new.csv")
#PCA using scale
PCA <- prcomp(new_pp, scale. = T)

PCA$rotation

PC <- as.data.frame(PCA$x)

pc_eigenvalues <- PCA$sdev^2

pc_eigenvalues <- tibble(PC = factor(1:length(pc_eigenvalues)),
                         variance = pc_eigenvalues) %>%
  # add a new column with the percent variance
  mutate(pct = variance/sum(variance)*100) %>%
  # add another column with the cumulative variance explained
  mutate(pct_cum = cumsum(pct))

# print the result
pc_eigenvalues
#plot
exp_pp <- pc_eigenvalues %>%
  ggplot(aes(x = PC)) +
  geom_col(aes(y = pct)) +
  geom_line(aes(y = pct_cum, group = 1)) +
  geom_point(aes(y = pct_cum)) +
  labs(x = "Principal component", y = "Fraction variance explained") +
  theme(legend.title = element_text(size=16),
        legend.text = element_text(size=12),
        legend.text.align = 0,
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 12),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))

#save the plot
ggsave(exp_pp, file="../figures/pca_expl_prosopis.png", device="png", dpi = 300, width = 10, height = 8)

#bind meta data with PC table
PC <- cbind(meta, PC)
#explre data
str(PC)

PC %>%
  mutate(tree = as.factor(tree), month.Treatment = as.factor(month.Treatment)) %>%
  ggplot(aes(x = PC1, y = PC2, color = tree, shape = month.Treatment)) +
  geom_point(size = 4)

PCA$rotation
#write table
write.csv(as.data.frame(PCA$rotation), "../out/pca_ss.csv")

#explore PC_1
cp_1 <- data.frame(PCA$rotation) %>%
  rownames_to_column() %>%
  dplyr::select(rowname, PC1) %>%
  rename(Variable_C1 = rowname, Component_1 = PC1) %>%
  arrange(desc(Component_1)) %>%
  mutate(Component_1 = round(Component_1, 3))

#explore PC2
cp_2 <- data.frame(PCA$rotation) %>%
  rownames_to_column() %>%
  dplyr::select(rowname, PC2) %>%
  rename(Variable_C2 = rowname, Component_2 = PC2) %>%
  arrange(desc(Component_2)) %>%
  mutate(Component_2 = round(Component_2, 3))

PC %>%
  mutate(tree = as.factor(tree), month.Treatment = as.factor(month.Treatment)) %>%
  ggplot(aes(x = PC1, y = PC2, color = tree, shape = month.Treatment)) +
  geom_point(size = 4)

# Extract loadings of the variables
PCAloadings <- data.frame(Variables = rownames(PCA$rotation), PCA$rotation) %>%
  filter(Variables %in% cp_1[1:10,1] | Variables %in% cp_2[1:10,1])

pl1 <- PC %>%
  mutate(tree = as.factor(tree), month.Treatment = as.factor(month.Treatment)) %>%
  ggplot() +
  geom_segment(data = PCAloadings, aes(x = 0, y = 0, xend = (PC1*13000),
                                       yend = (PC2*13000)), arrow = arrow(length = unit(1/2, "picas")),
               color = "black") +
  geom_point(aes(x = PC1, y = PC2, color = tree, shape = month.Treatment), size = 5)

pl2 <- pl1 +  geom_point(data = PCAloadings, aes(x = (PC1*13000), y = (PC2*13000)), size = 0.5) +
  geom_label_repel(data = PCAloadings, aes(x = (PC1*13000), y = (PC2*13000), label = Variables),
                   size = 4,
                   nudge_y       = 36,
                   segment.size  = 0.2,
                   segment.color = "grey50",
                   direction     = "y")

#save PCA
ggsave(pl2, file="../figures/pca_prosopis.png", device="png", dpi = 300, width = 16, height = 10)
```

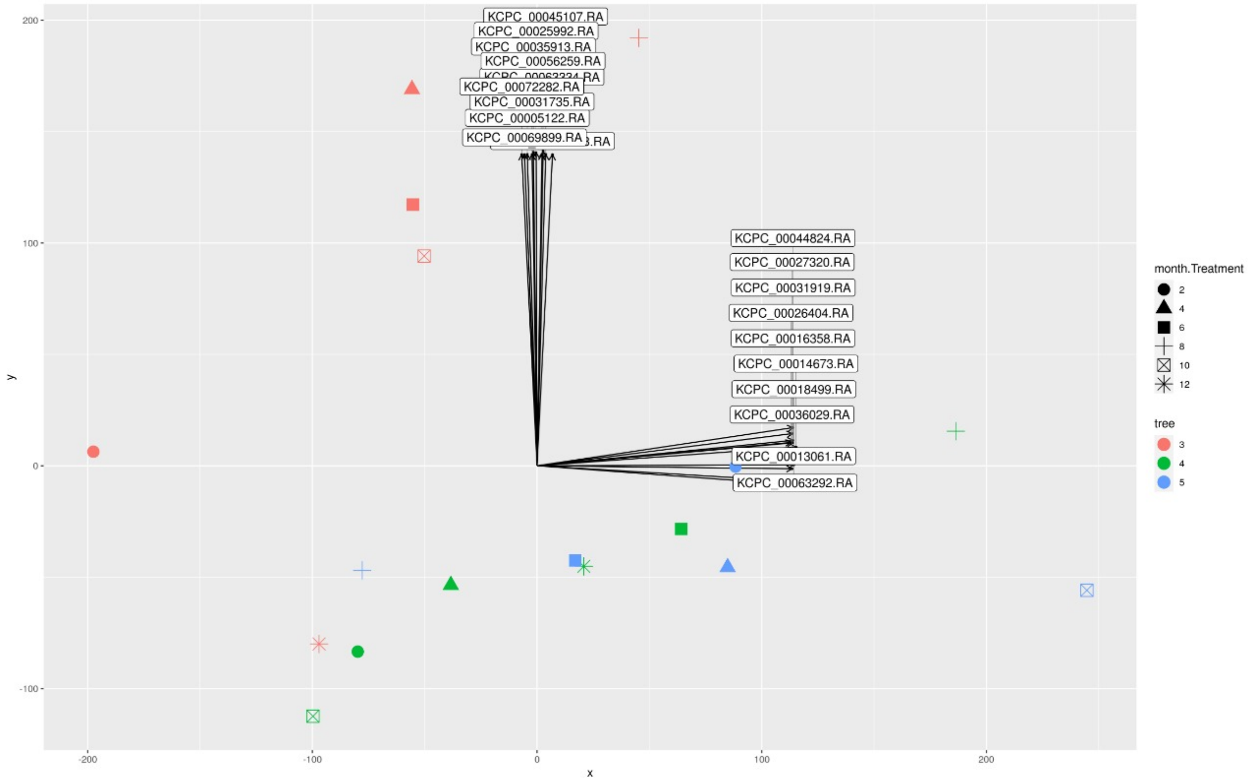


Figure 1. First two dimensions of PCA.

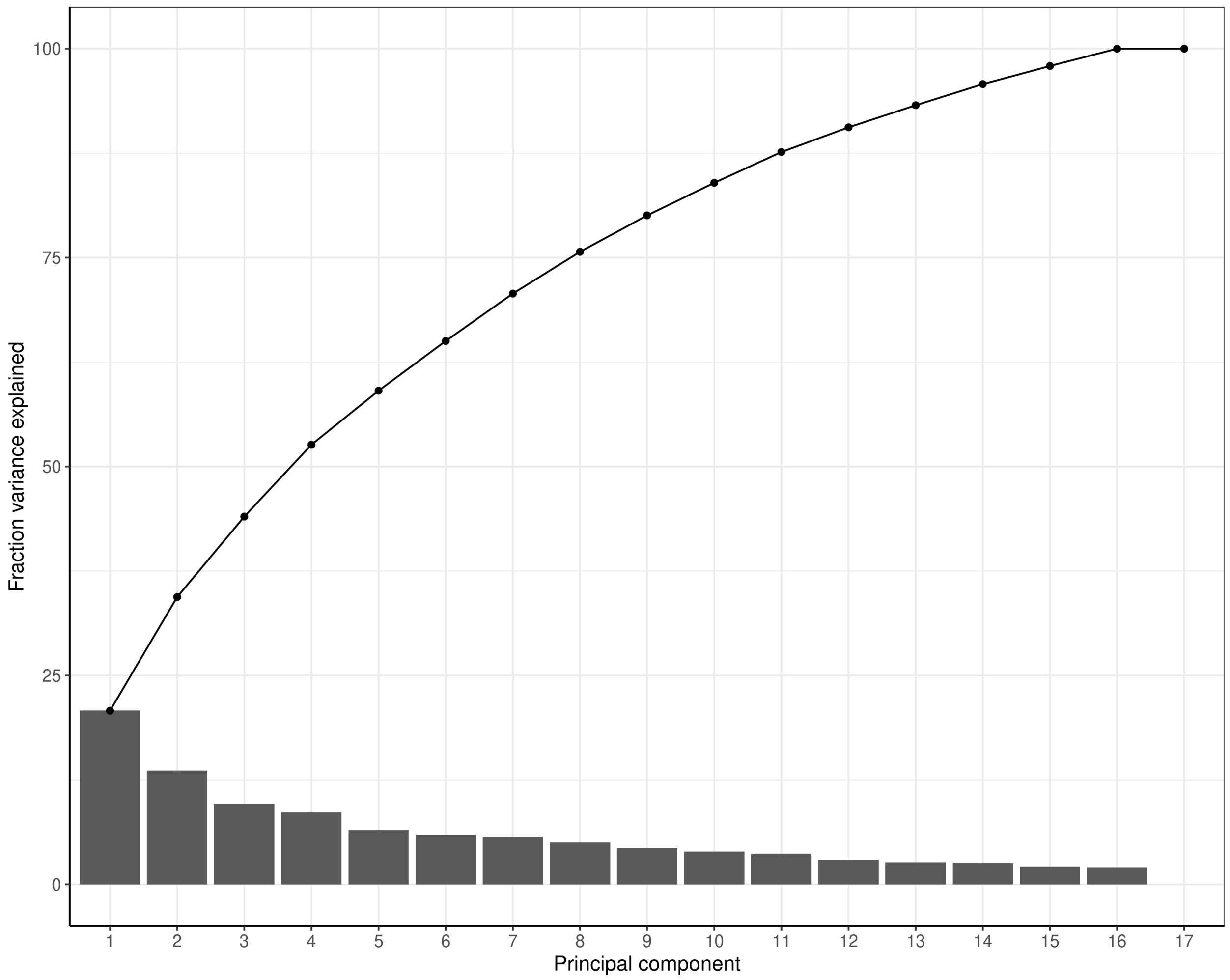


Figure 2. Proportion of explained variance by PC.

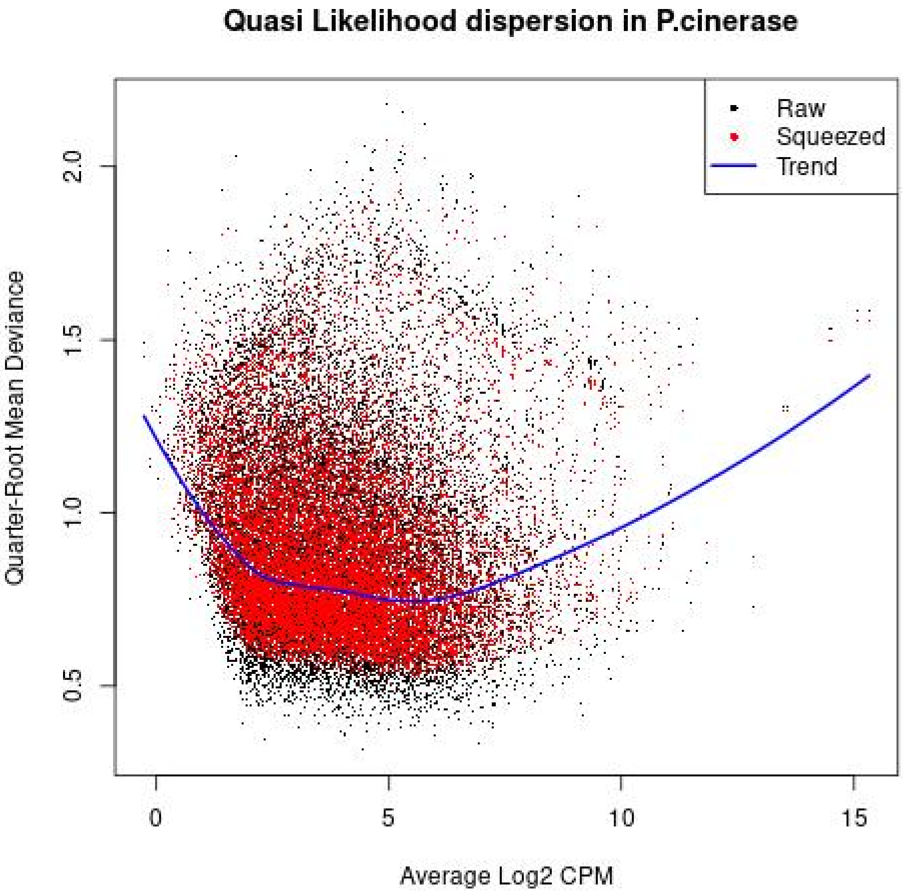
5. Differential expression analysis

The Differential expression analysis will be performed using the count matrix and *edgeR* and *DeSeq2* packages

We performed a differential expression analysis in edgeR using the prosopis data the tree_3 to tree_5 were used as a biological replicates and the months were considered as a treatment. As a result we can see that the samples have a lot o variation and replicates can not show a replicate behaviour, thats the reason why the P-value is too high.

```
comparacion           genes               logFC           unshrunk.logFC           logCPM            PValue
Length:22354       Length:22354       Min.   :-10.10308   Min.   :-144269486   Min.   :-0.2659   Min.   :0.00
Class :character   Class :character   1st Qu.: -0.36141   1st Qu.:         0   1st Qu.: 2.5909   1st Qu.:0.57
Mode  :character   Mode  :character   Median : -0.02668   Median :         0   Median : 3.7133   Median :0.80
                                      Mean   : -0.06066   Mean   :     12908   Mean   : 3.9144   Mean   :0.72
                                      3rd Qu.:  0.30854   3rd Qu.:         0   3rd Qu.: 4.9792   3rd Qu.:0.92
                                      Max.   : 11.63869   Max.   : 144269488   Max.   :15.3226   Max.   :0.99

 condition
Length:22354
Class :character
Mode  :character
```


Quasi Likelihood dispersion in P.cinerase
Raw
Squeezed
Trend
Quarter-Root Mean Deviance
0.5
1.0
1.5
2.0
0
5
10
15
Average Log2 CPM