report

Bingqing Yang

2023-12-08

Contents

1. Figure out `log(res^2)` and `log(|res|)` which is better as dependent variable;

- For 1000 genes which implement on shape constrained addictive model, calculate the residual of those model.
- Implement linear model and loess model based on Log transformed square of residual and log transformed absolute of residual as dependent variable;

```
# Reload data
library(devtools)
```

```
##      usethis
```

```
library(usethis)
library(ggplot2)
load_all("../project/array2rnaseq")
```

```
## i Loading array2rnaseq
```

```
## Warning: Objects listed as exports, but not present in namespace:
## * get_w
```

```
path = "../project/u133-array-to-tcga-rnaseq/map/"

fit_scam_r2_lm <- readRDS(paste0(path,
"models/fit_scam_r2_lm.rds"))
fit_scam_r2_loess <- readRDS(paste0(path, "models/fit_scam_r2_loess.rds"))
fit_scam_r_lm <- readRDS(paste0(path, "models/fit_scam_r_lm.rds"))
fit_scam_r_loess <- readRDS(paste0(path, "models/fit_scam_r_loess.rds"))

# PI for diff models
pred_scam_r2_lm <- readRDS(paste0(path, "pred/pred_scam_r2_lm.rds"))
pred_scam_r2_loess <- readRDS(paste0(path, "pred/pred_scam_r2_loess.rds"))
pred_scam_r_lm <- readRDS(paste0(path, "pred/pred_scam_r_lm.rds"))
pred_scam_r_loess <- readRDS(paste0(path, "pred/pred_scam_r_loess.rds"))

# R^2 summary of residual data
R2 <- read.csv(paste0(path, "pred/R2_summary.csv"), row.names = 1)
```

```r
# Microarray intensify and RNA-seq data
x <- readRDS(paste0(path, "data/x_1000.rds"))
y <- readRDS(paste0(path, "data/y_1000.rds"))
dim(x)
```

```
## [1] 1000  294
```

- Calculate mean square of R in two function, because mean of R square in $log(res^2)$ is larger than $log(|res|)$ no matter in linear model or loess model. So using $log(res^2)$ as dependent variable;

```r
# Calculate mean of R square in diff dependent vaiable and models
colMeans(R2)
```

```
##    R2_r2_lm R2_r2_loess     R2_r_lm  R2_r_loess
##  0.01145378  0.07986962  0.01095803  0.07927829
```

2. Look at genes with heteroscedasticity;

Only considering $log(res^2)$ as dependent variable, implement `lm(log(res^2) ~ x)` and `loess(log(res^2) ~ x)` for each gene. Then calculate R^2 of each gene in diff models.

- Select genes of whom $R^2 > 0.1$ of `lm(log(res^2) ~ x)`, which has 6 genes;

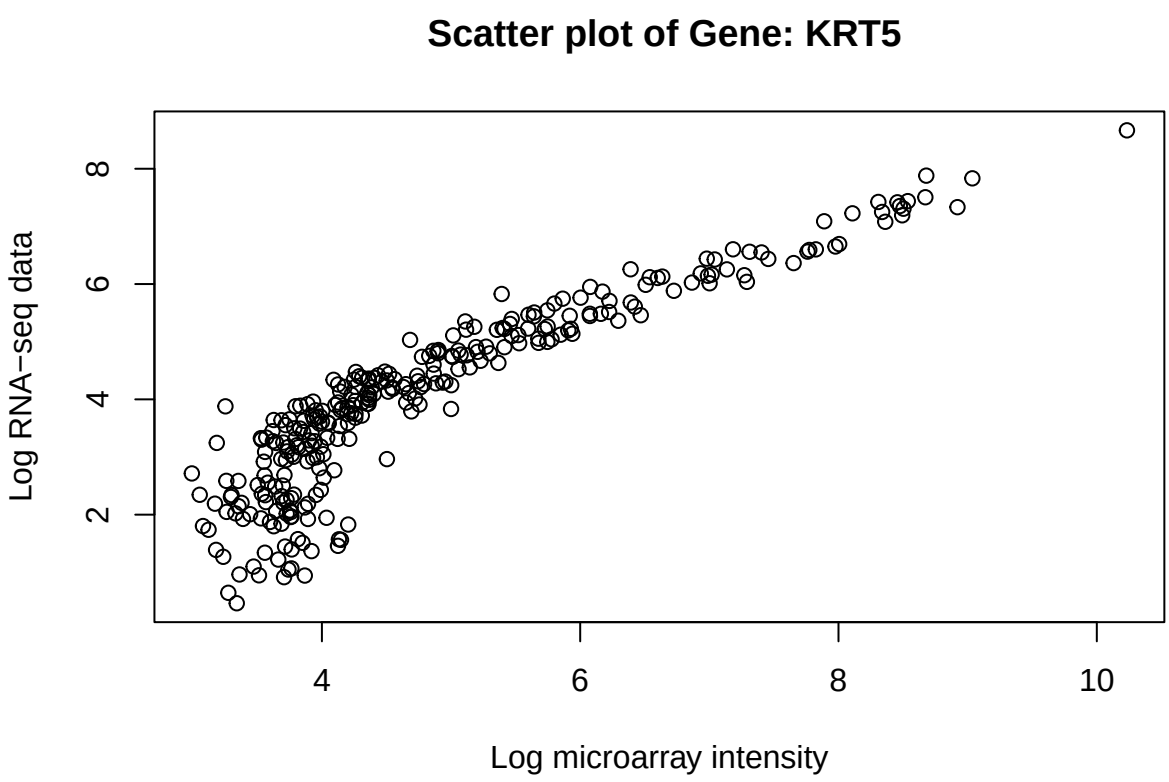
```r
# gene name
gene_lm_lst <- rownames(R2)[R2$R2_r2_lm > 0.1]
print(gene_lm_lst)
```

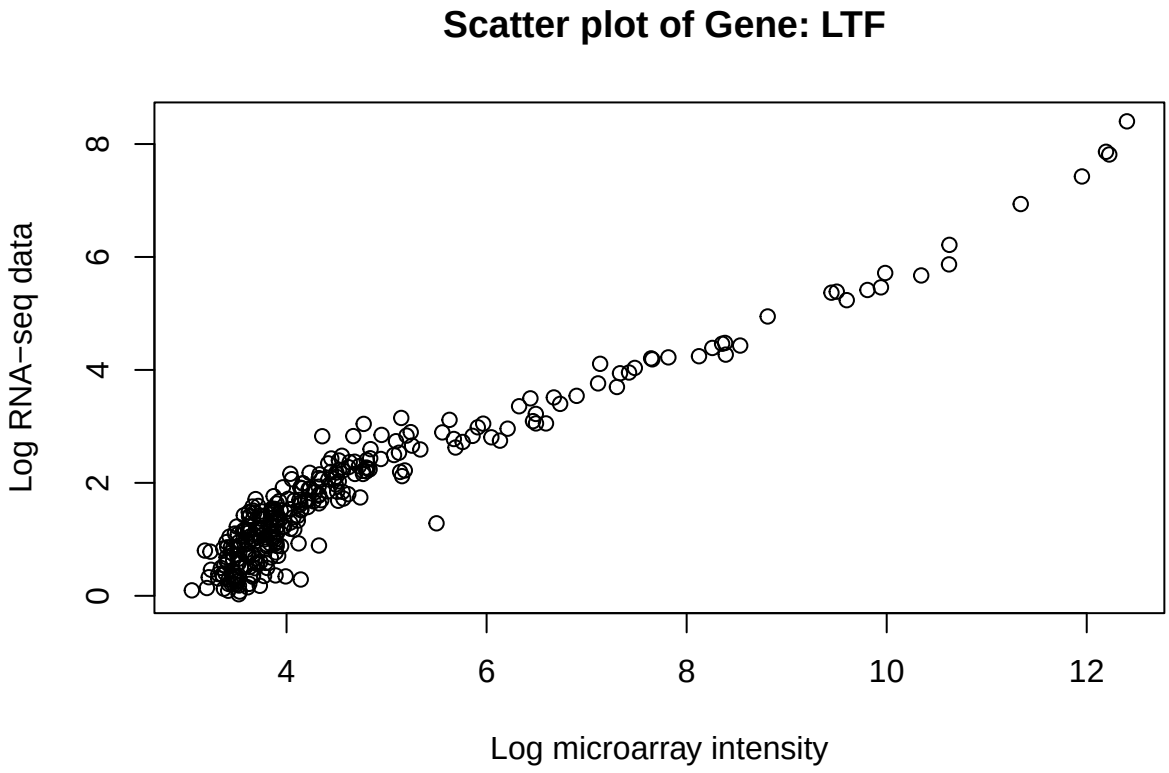
```
## [1] "KRT5"   "LTF"    "C7"     "GABBR1" "LSP1"   "AOC1"
```

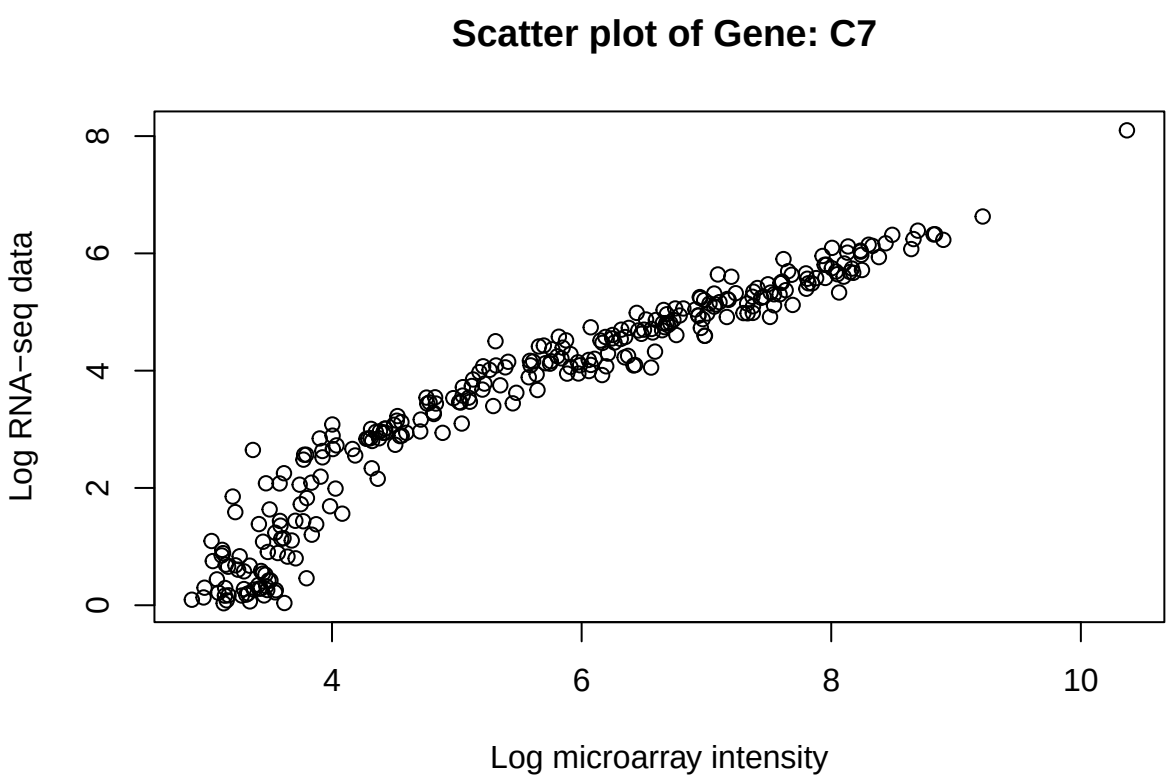
```r
# gene index
idx_R0.1 <- which(R2$R2_r2_lm > 0.1)
print(idx_R0.1)
```

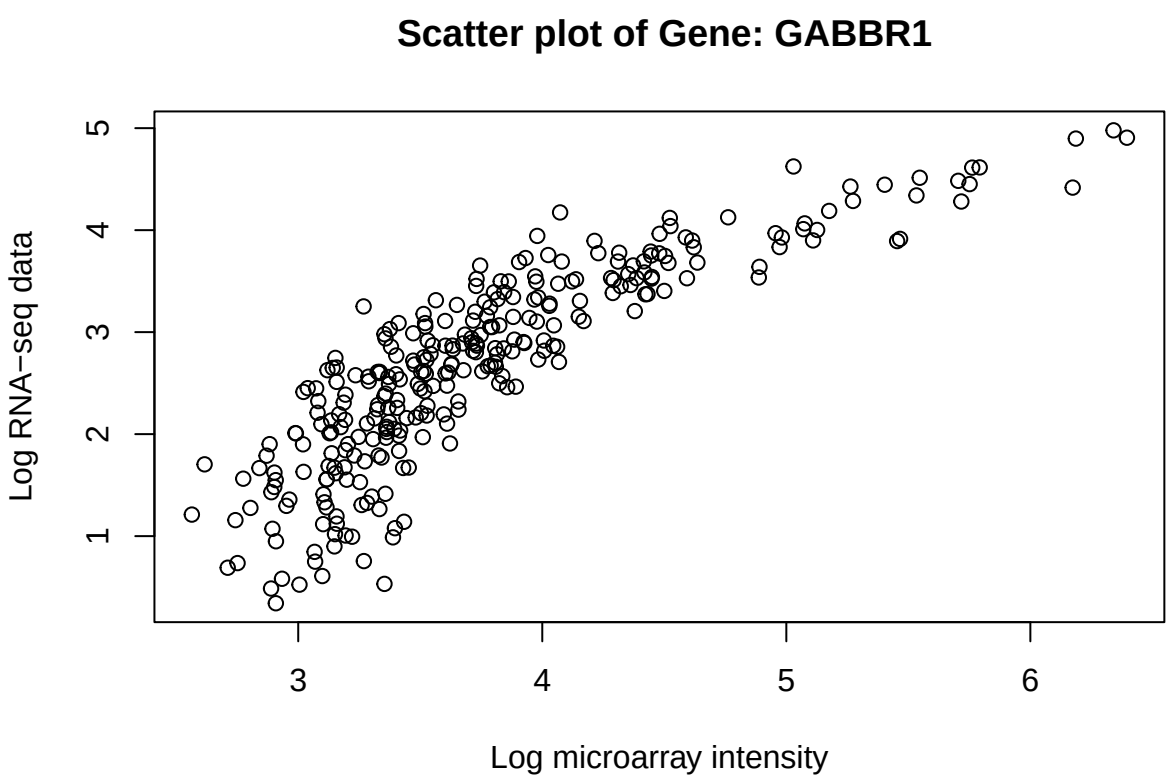
```
## [1] 356 405 706 763 894 909
```

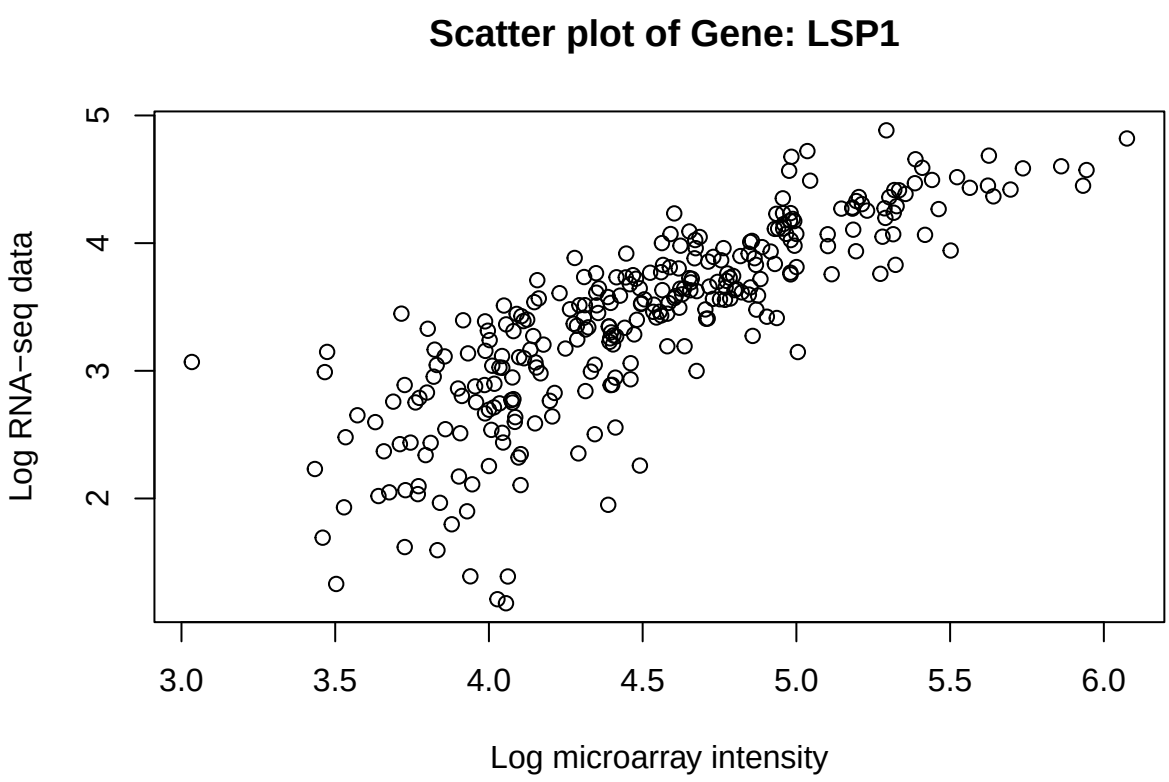
```r
for (i in idx_R0.1) {
  pic <- plot(x[i, ], y[i, ], main = paste0(" Scatter plot of Gene: ", rownames(x)[i]),
              xlab = "Log microarray intensity", ylab = "Log RNA-seq data")
}
```

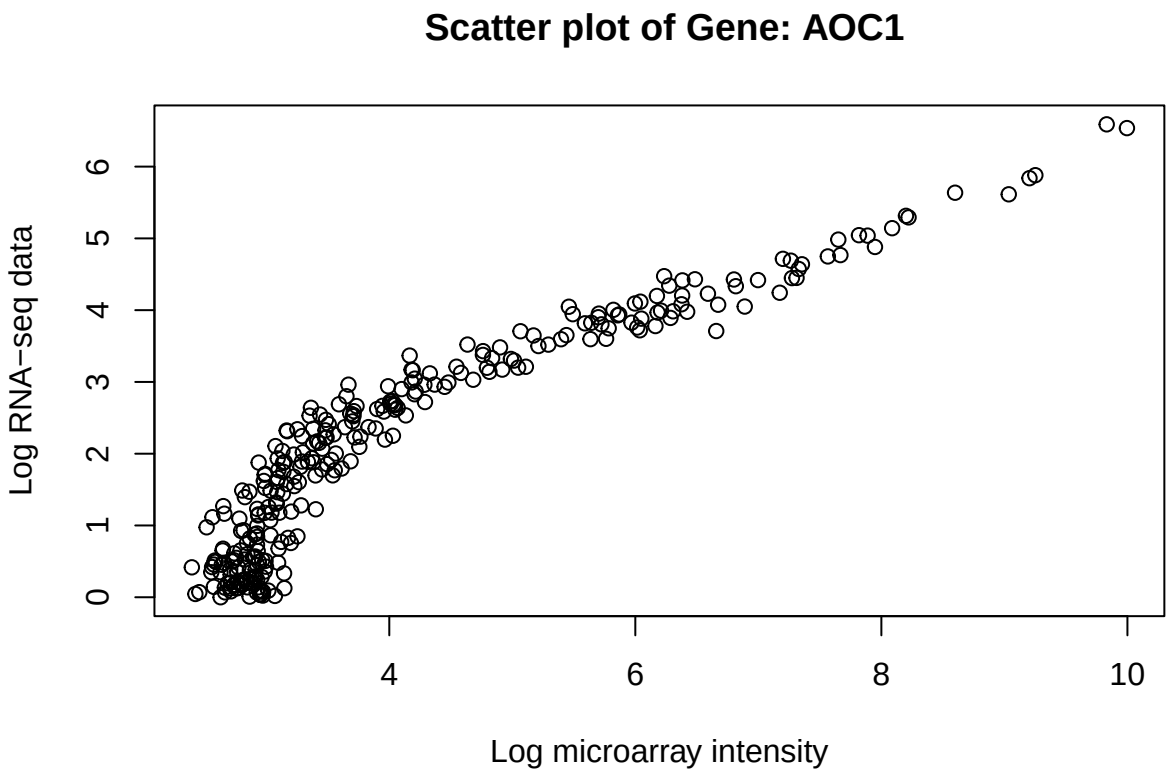

Scatter plot of Gene: KRT5
Log RNA−seq data
2
4
6
8
4
6
8
10
Log microarray intensity


Scatter plot of Gene: LTF
Log RNA−seq data
0
2
4
6
8
4
6
8
10
12
Log microarray intensity


Scatter plot of Gene: C7
Log RNA−seq data
0
2
4
6
8
4
6
8
10
Log microarray intensity


Scatter plot of Gene: GABBR1
Log RNA−seq data
1
2
3
4
5
3
4
5
6
Log microarray intensity


Scatter plot of Gene: LSP1
Log RNA−seq data
2
3
4
5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
Log microarray intensity



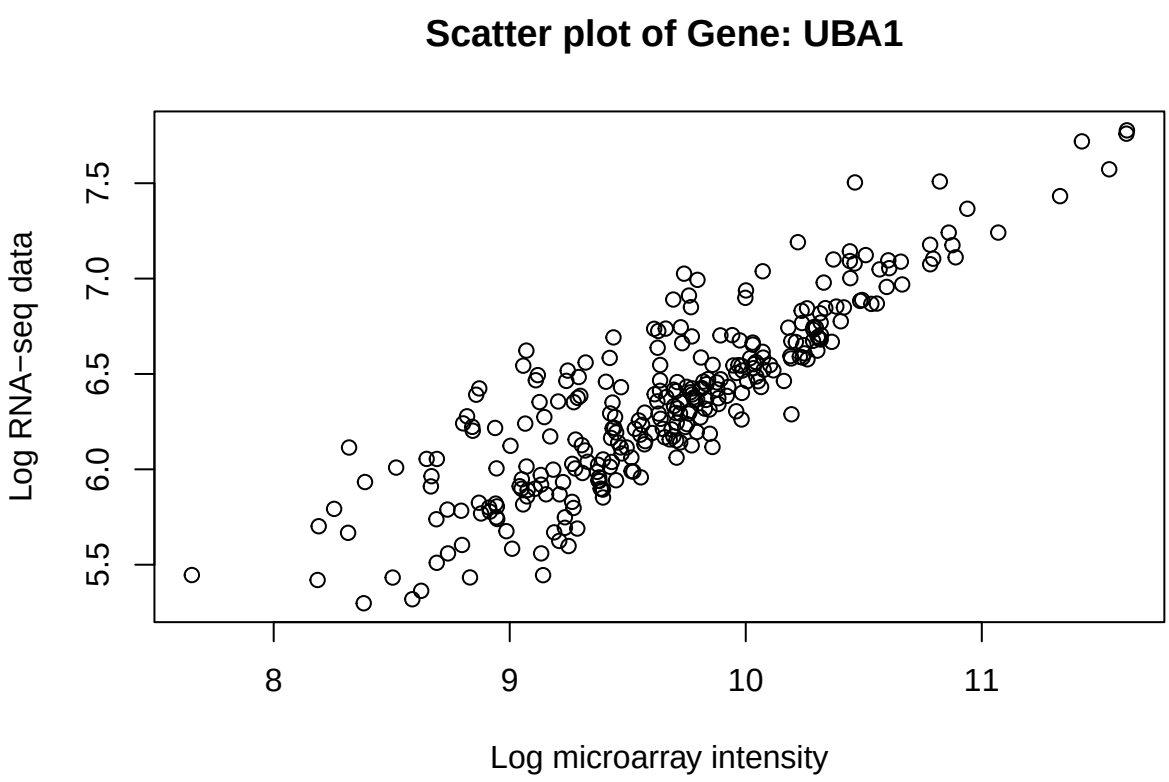
- Select genes of whom $R^2 > 0.2$ of `loess(log(res^2 ~ x)`, which has 21 genes;

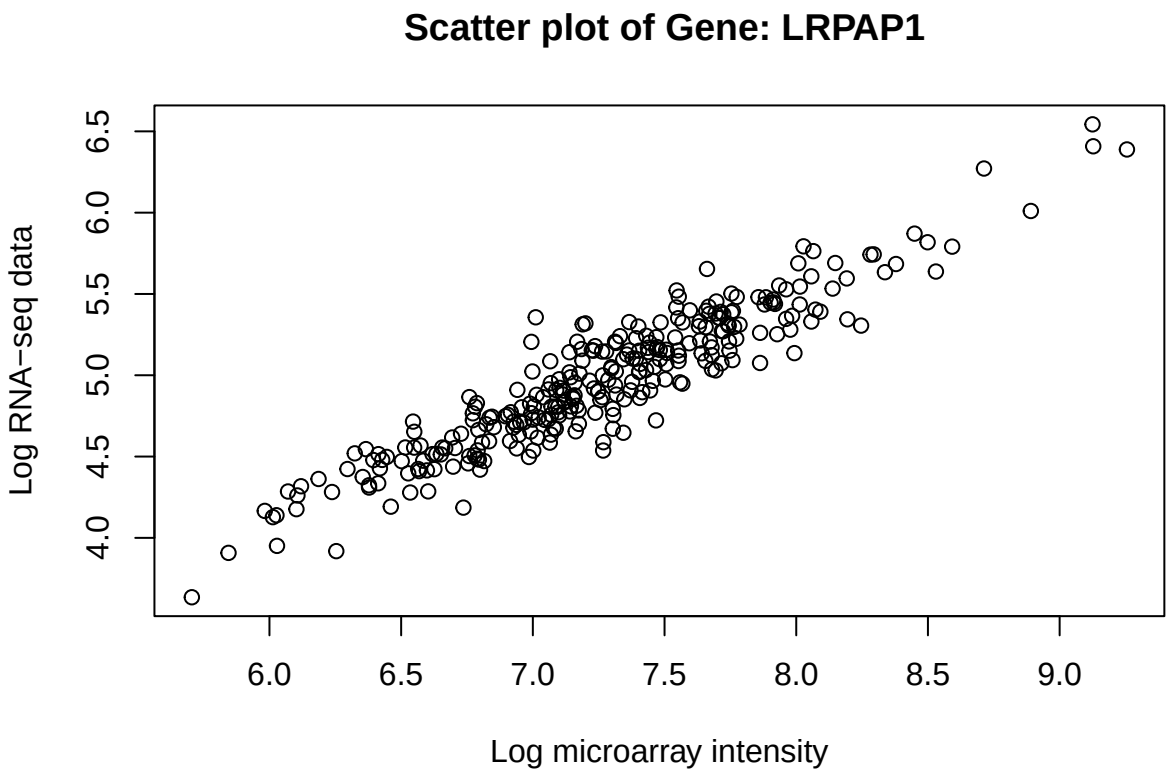
```
# gene name
gene_loess_lst <- rownames(R2)[R2$R2_r2_loess > 0.2]
print(gene_loess_lst)
```

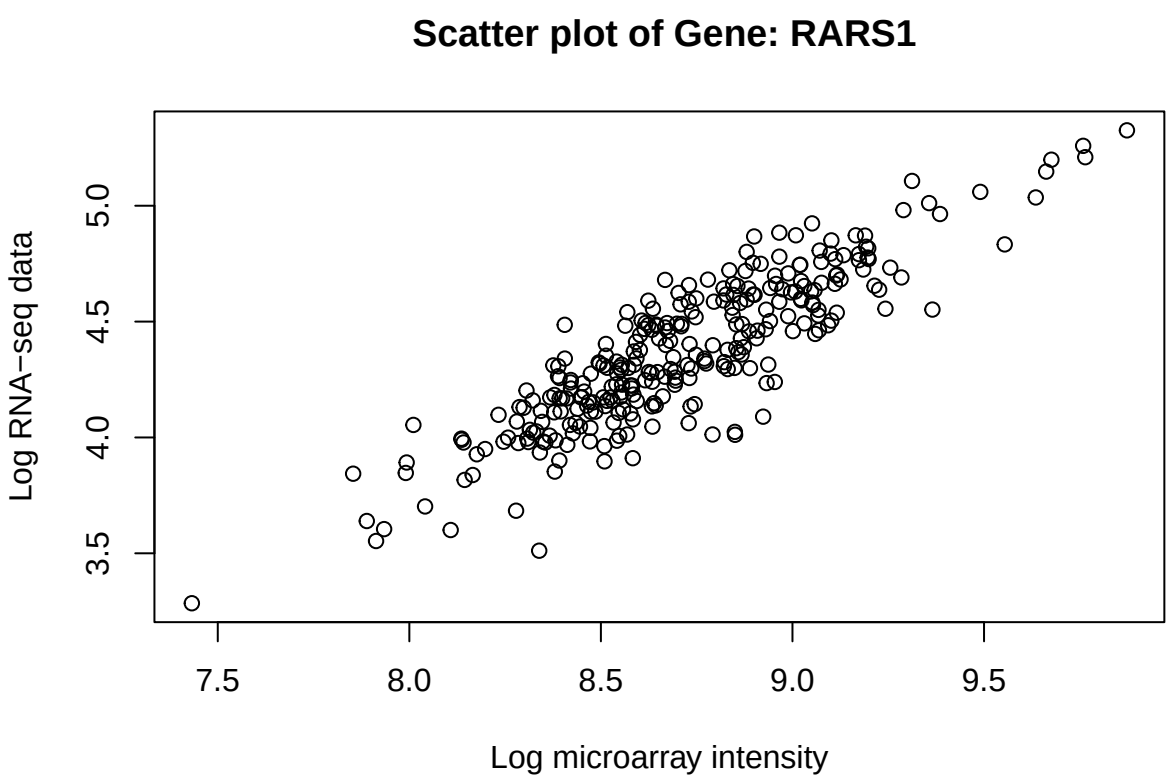
```
##  [1] "UBA1"    "LRPAP1"  "RARS1"   "NCAPD2"  "KRT5"    "CEACAM5" "PLK2"
##  [8] "LTF"     "ALDOC"   "NUTF2"   "FOSB"    "GPX2"    "AGT"     "C7"
## [15] "GABBR1"  "NECTIN2" "F13A1"   "TGIF1"   "CYP19A1" "AOC1"    "LOXL1"
```

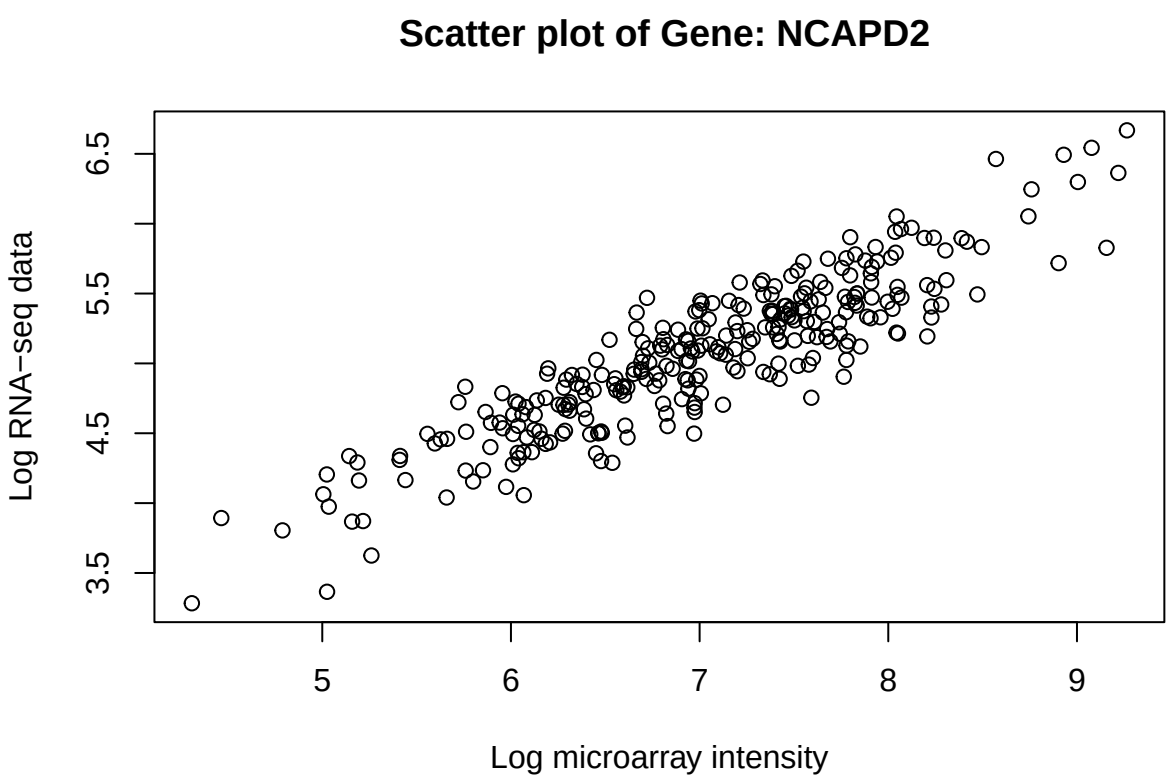
```
# gene index
idx_R0.2 <- which(R2$R2_r2_loess > 0.2)
print(idx_R0.2)
```

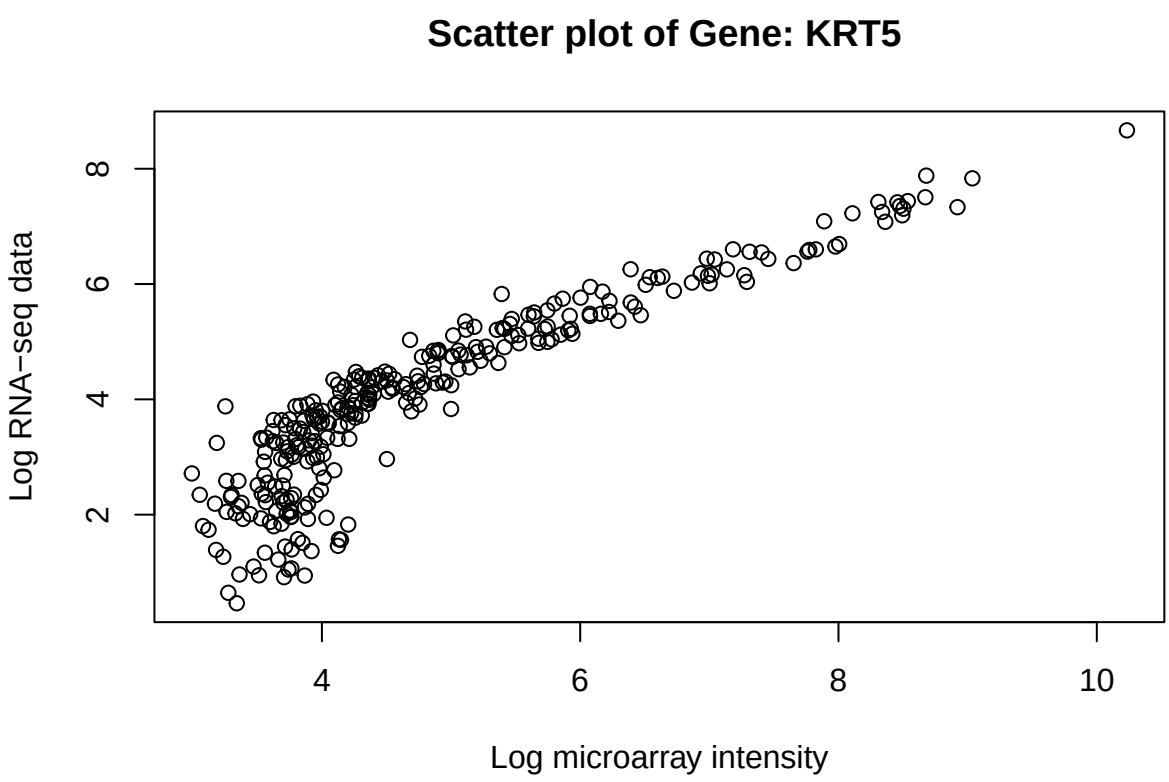
```
##  [1] 135 187 225 344 356 370 388 405 406 524 631 659 661 706 763 764 819 823 878
## [20] 909 917
```

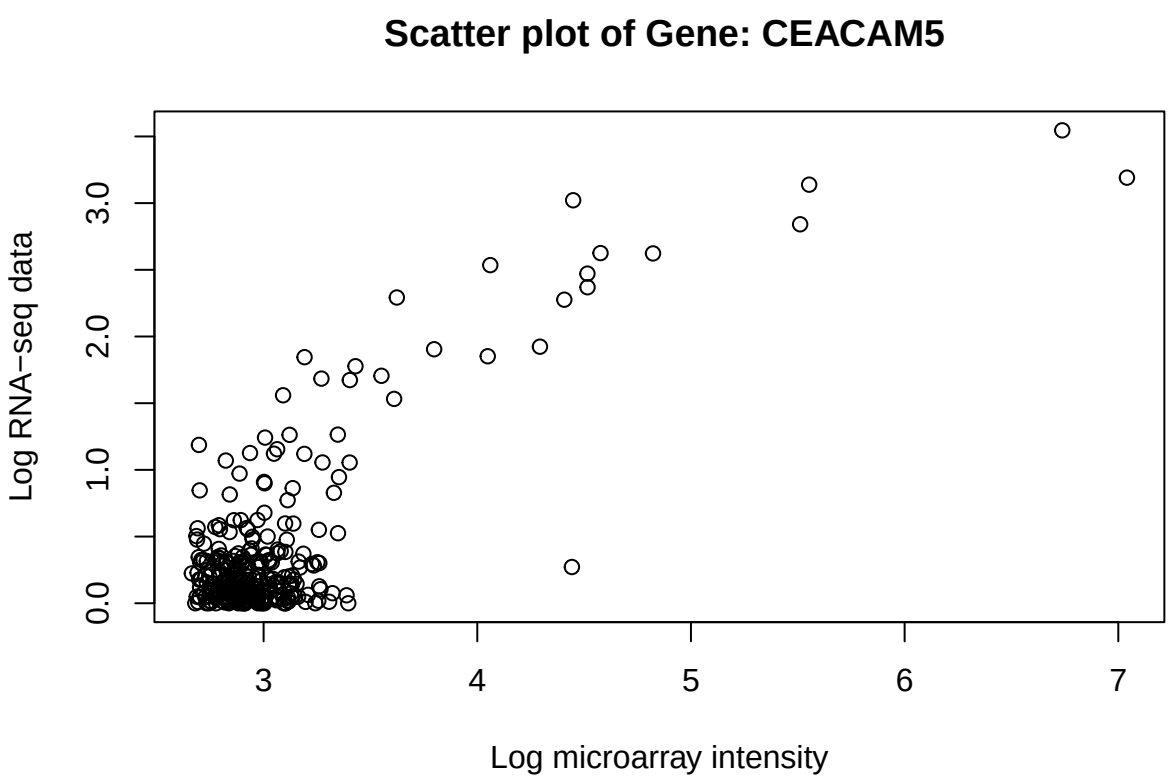

Scatter plot of Gene: UBA1
Log RNA−seq data
5.5
6.0
6.5
7.0
7.5
8
9
10
11
Log microarray intensity

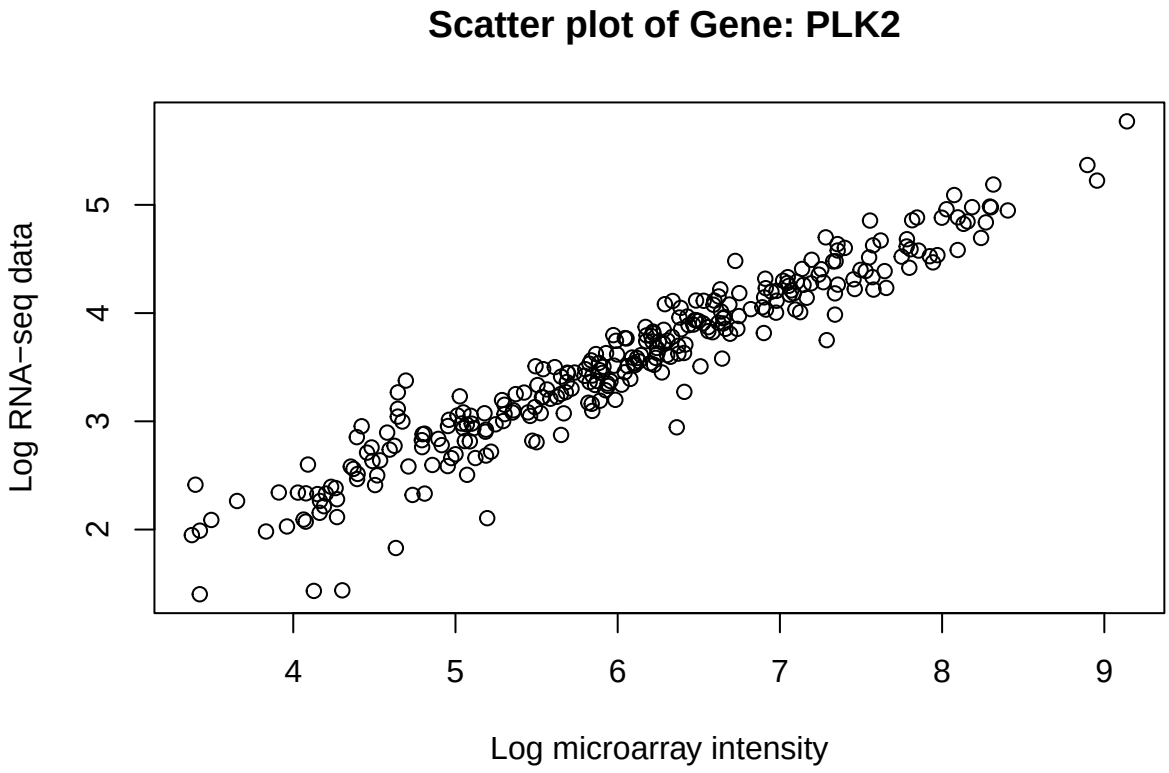

Scatter plot of Gene: LRPAP1
Log RNA−seq data
4.0
4.5
5.0
5.5
6.0
6.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Log microarray intensity

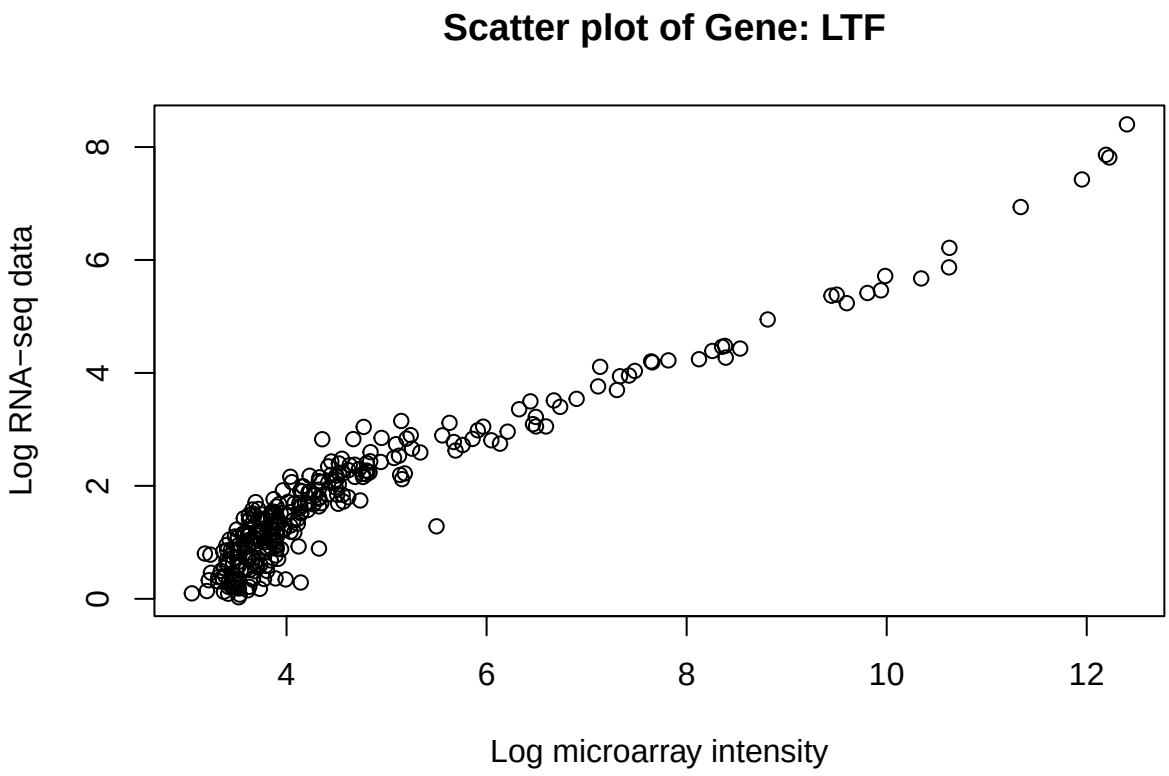

Scatter plot of Gene: RARS1
Log RNA−seq data
3.5
4.0
4.5
5.0
7.5
8.0
8.5
9.0
9.5
Log microarray intensity

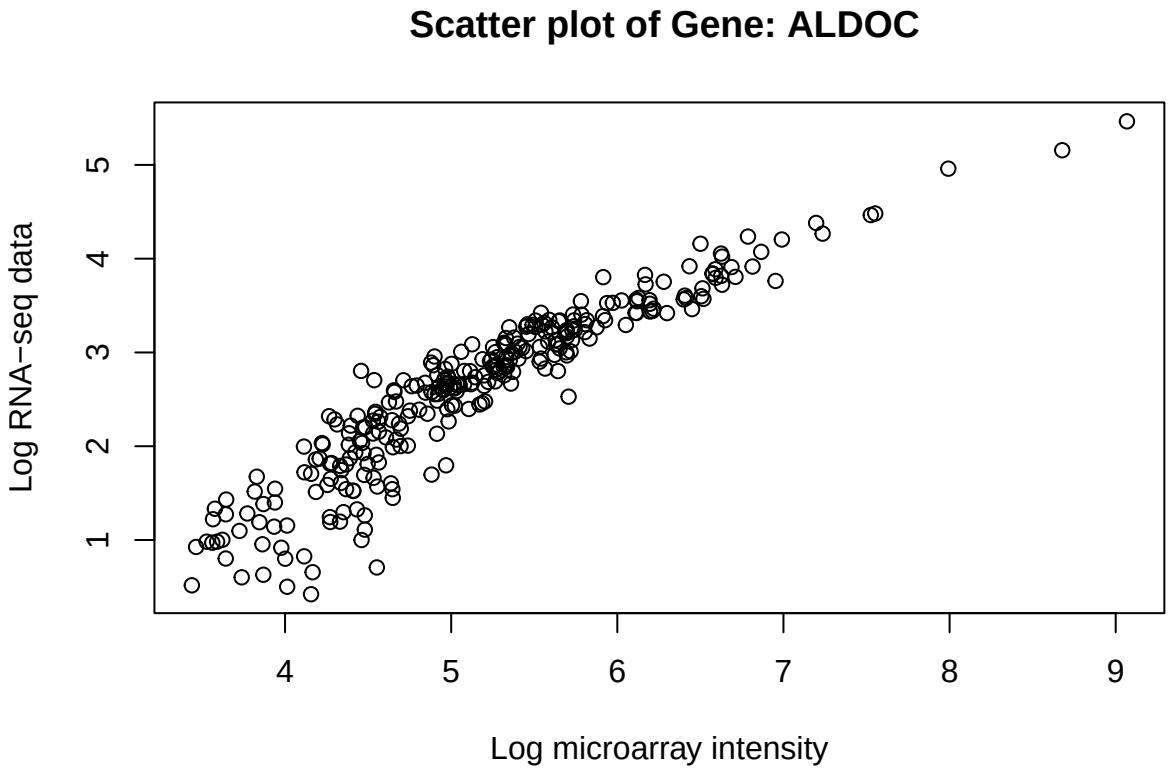

Scatter plot of Gene: NCAPD2
Log RNA−seq data
3.5
4.5
5.5
6.5
5
6
7
8
9
Log microarray intensity

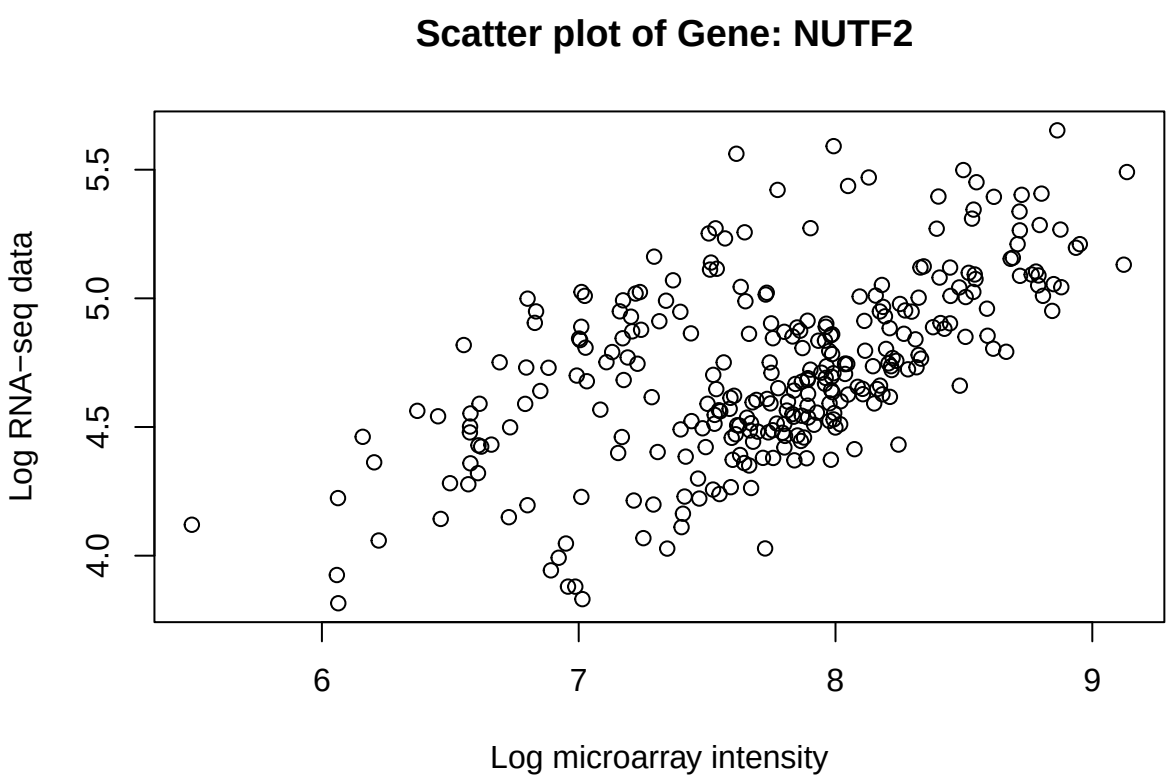

Scatter plot of Gene: KRT5
Log RNA−seq data
2
4
6
8
4
6
8
10
Log microarray intensity

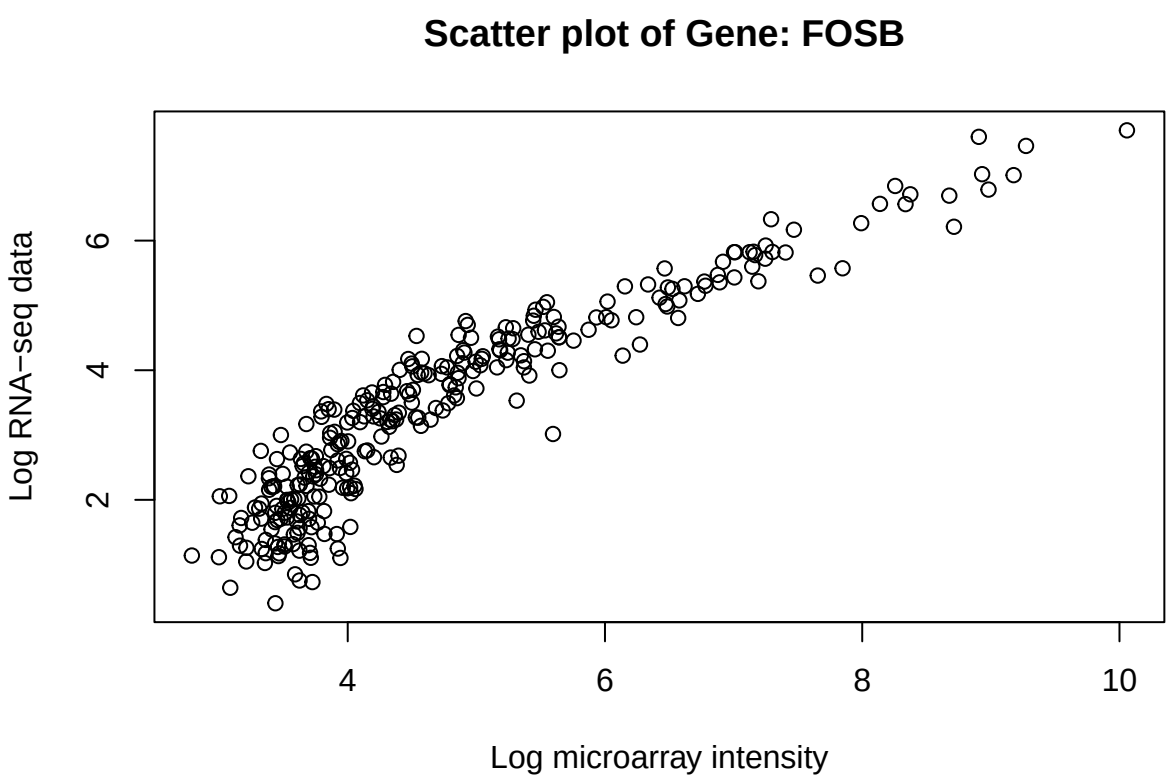

Scatter plot of Gene: CEACAM5
Log RNA−seq data
0.0
1.0
2.0
3.0
3
4
5
6
7
Log microarray intensity

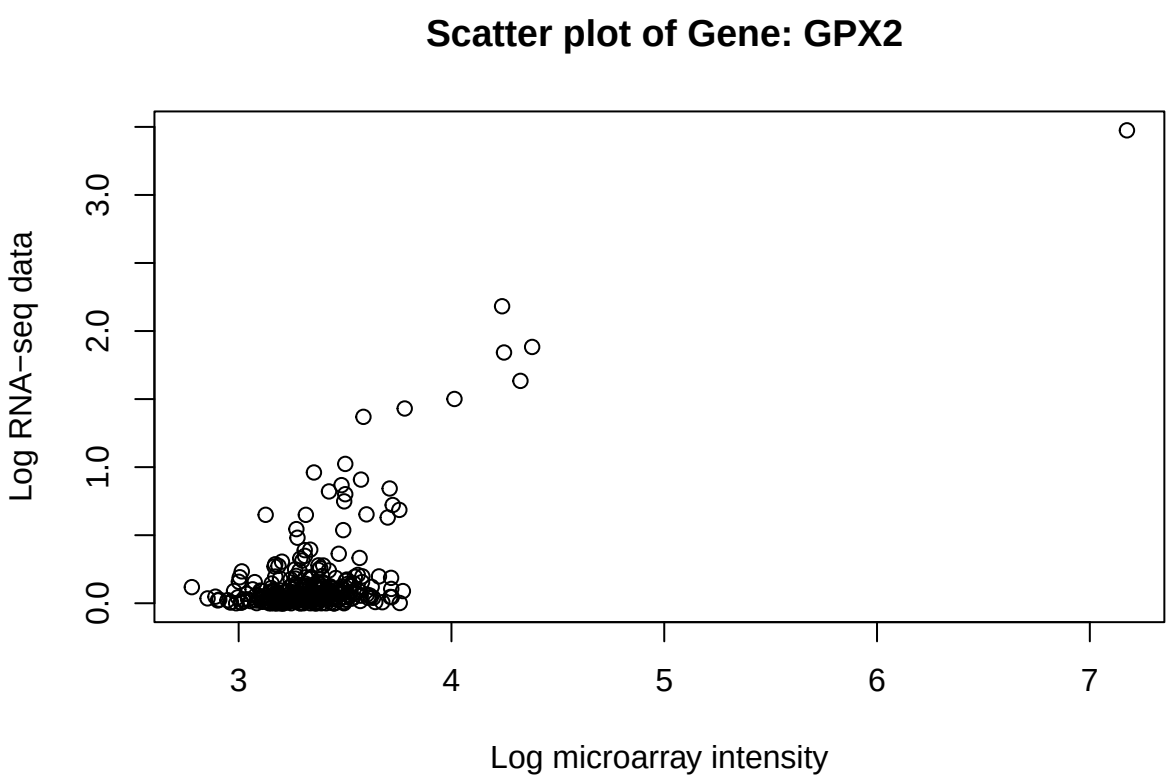

Scatter plot of Gene: PLK2
Log RNA−seq data
2
3
4
5
4
5
6
7
8
9
Log microarray intensity

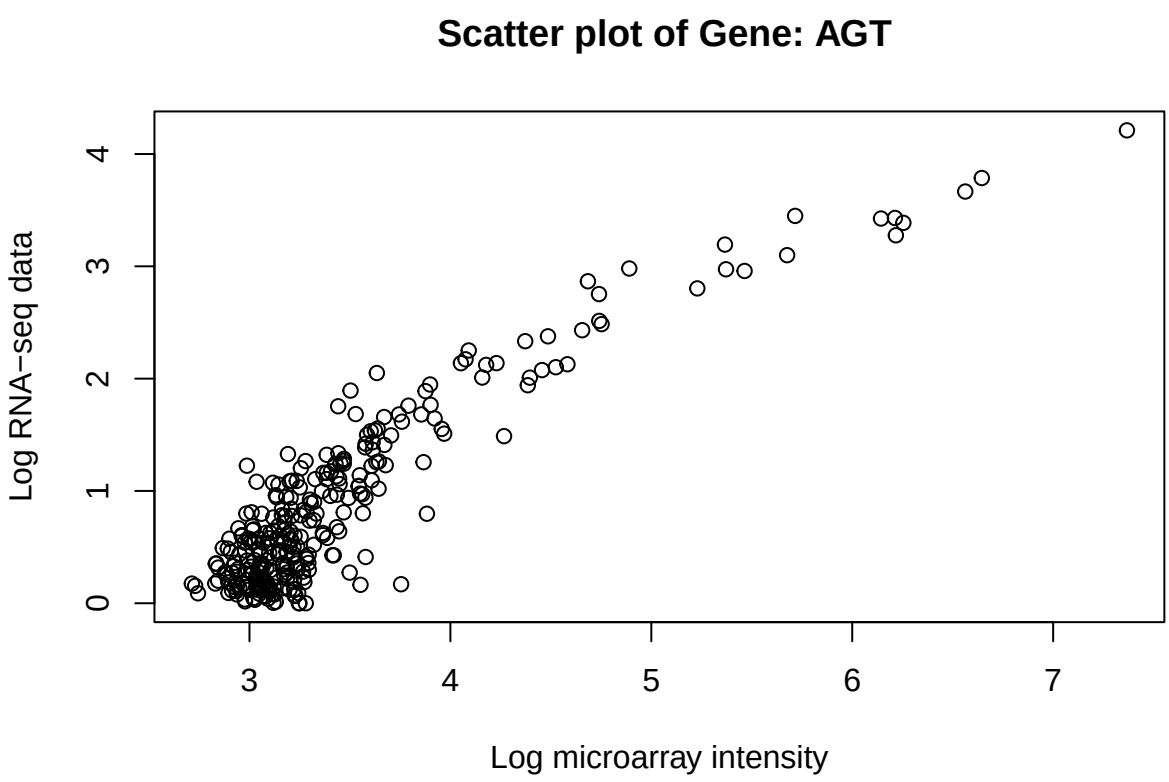

Scatter plot of Gene: LTF
Log RNA−seq data
0
2
4
6
8
4
6
8
10
12
Log microarray intensity

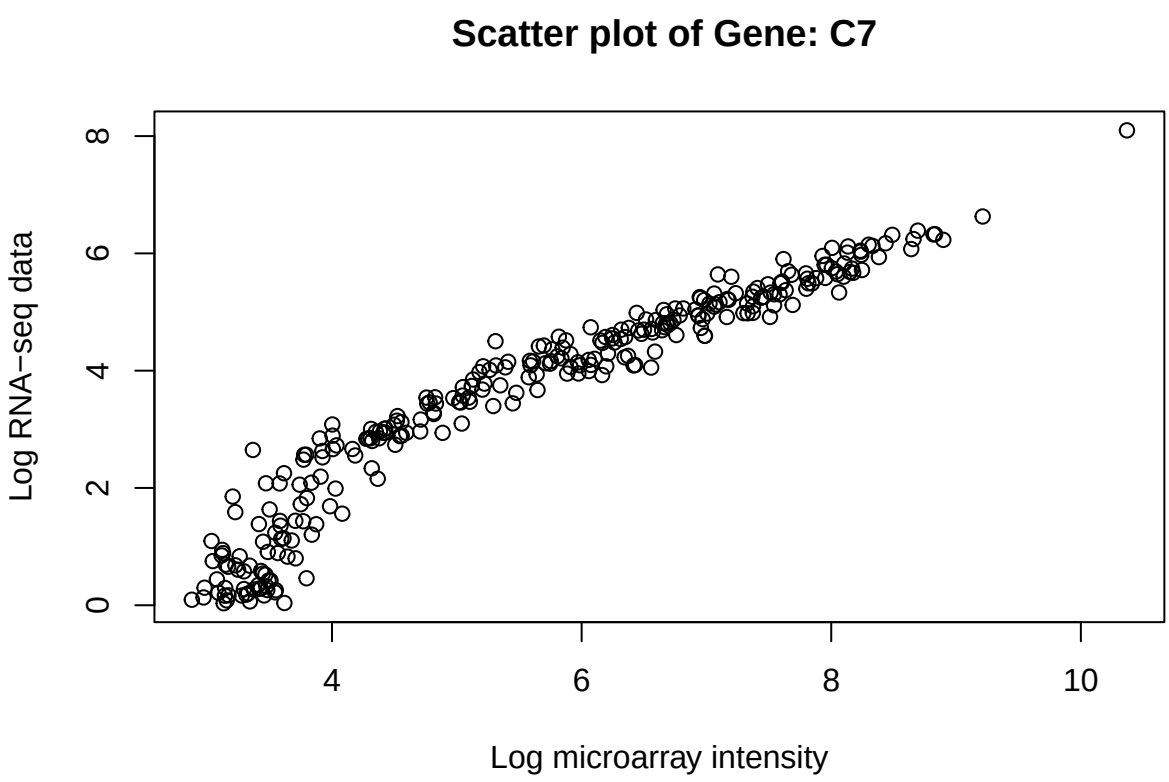

Scatter plot of Gene: ALDOC
Log RNA−seq data
1
2
3
4
5
4
5
6
7
8
9
Log microarray intensity

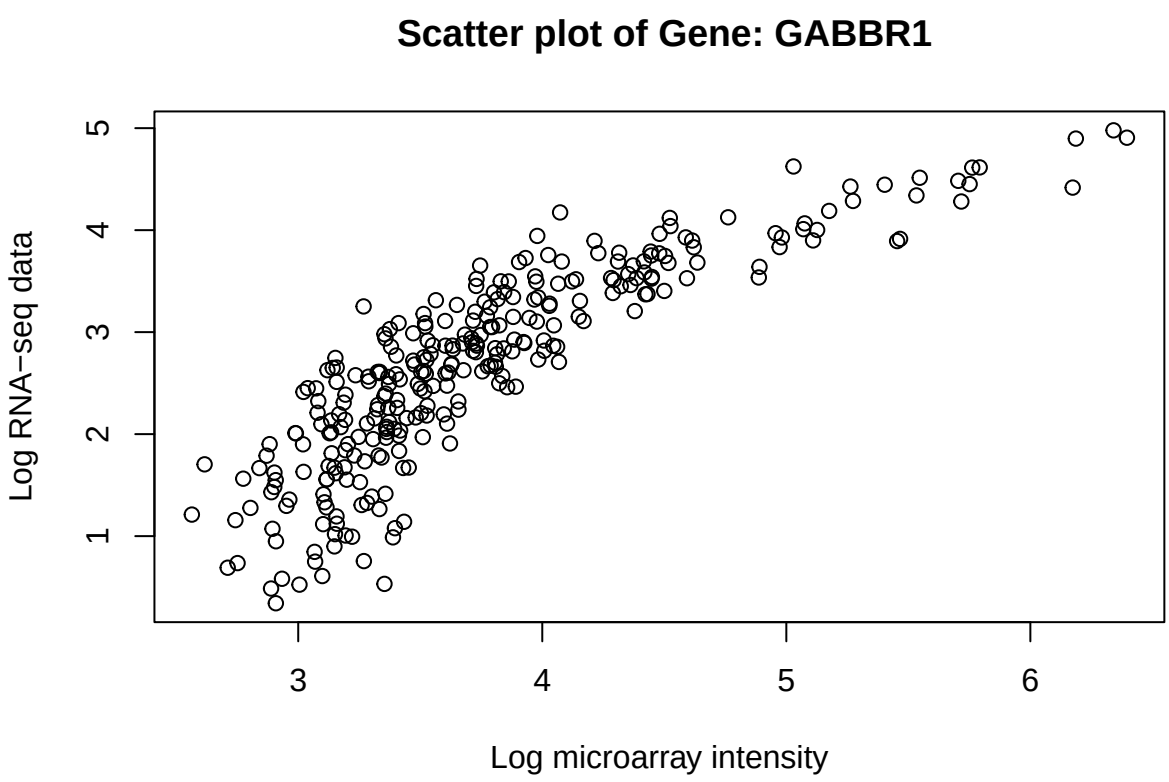

Scatter plot of Gene: NUTF2
Log RNA−seq data
4.0
4.5
5.0
5.5
6
7
8
9
Log microarray intensity

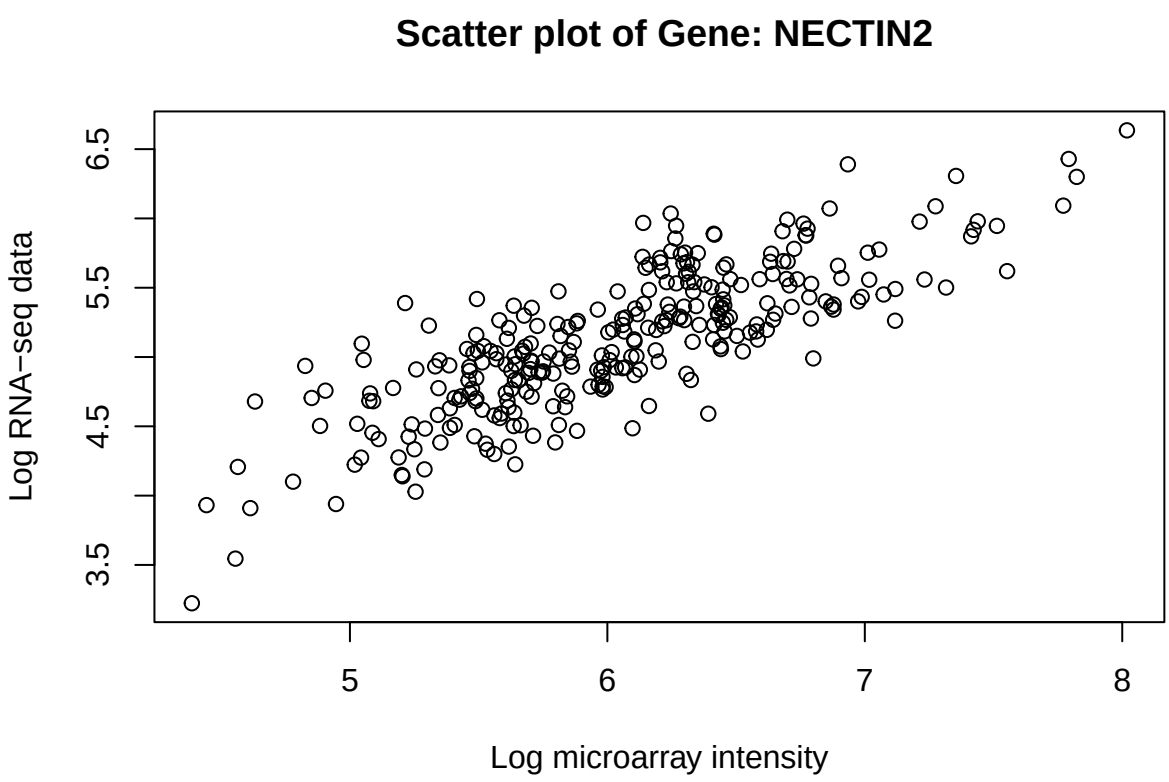

Scatter plot of Gene: FOSB
Log RNA−seq data
2
4
6
4
6
8
10
Log microarray intensity

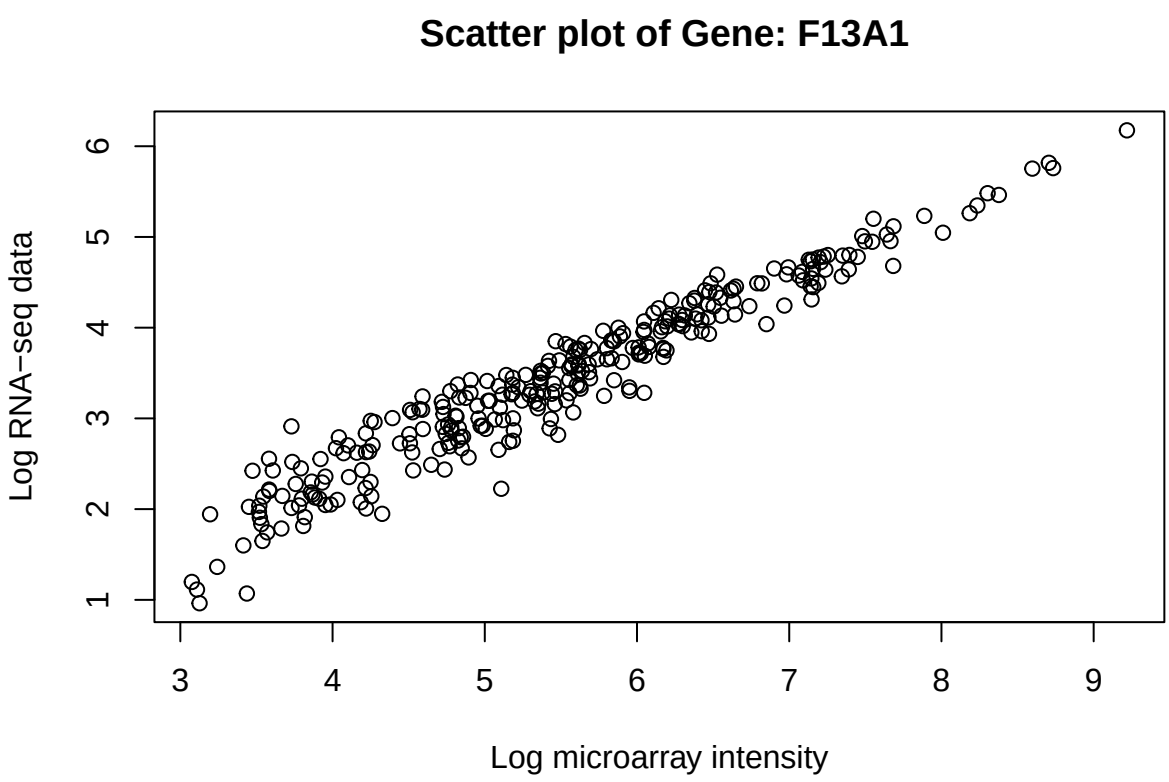

Scatter plot of Gene: GPX2
Log RNA−seq data
0.0
1.0
2.0
3.0
3
4
5
6
7
Log microarray intensity

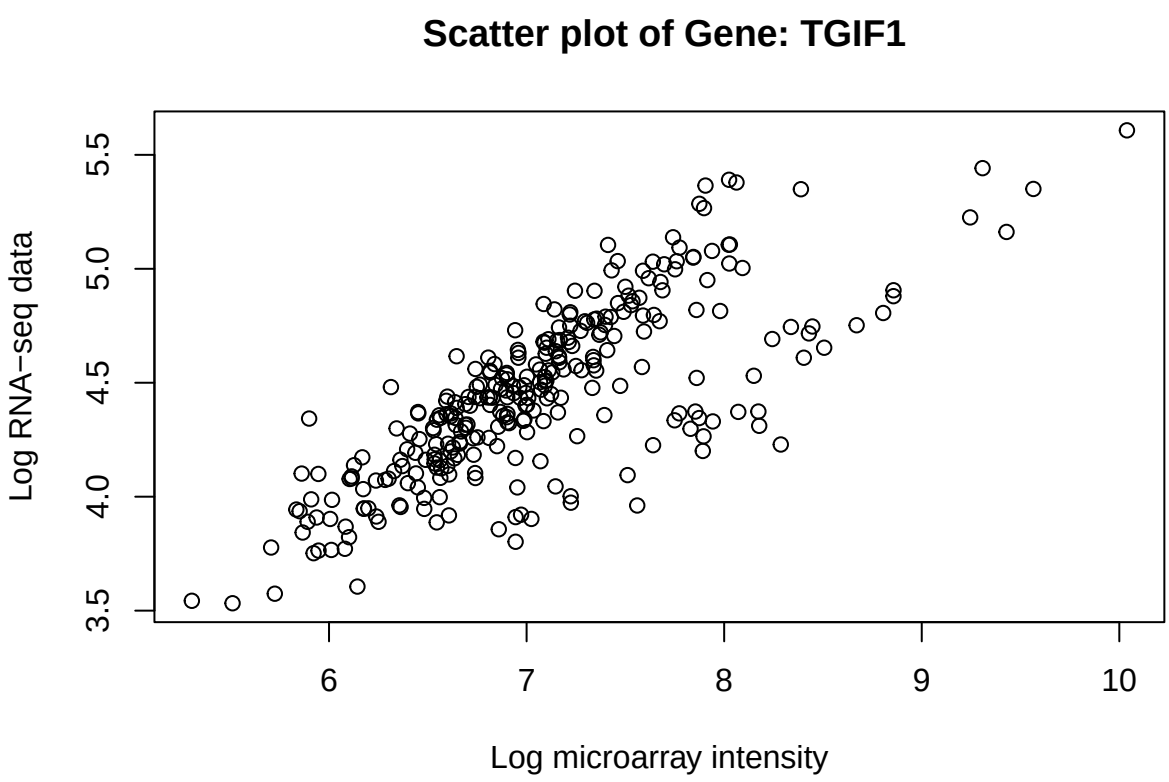

Scatter plot of Gene: AGT
Log RNA−seq data
0
1
2
3
4
3
4
5
6
7
Log microarray intensity

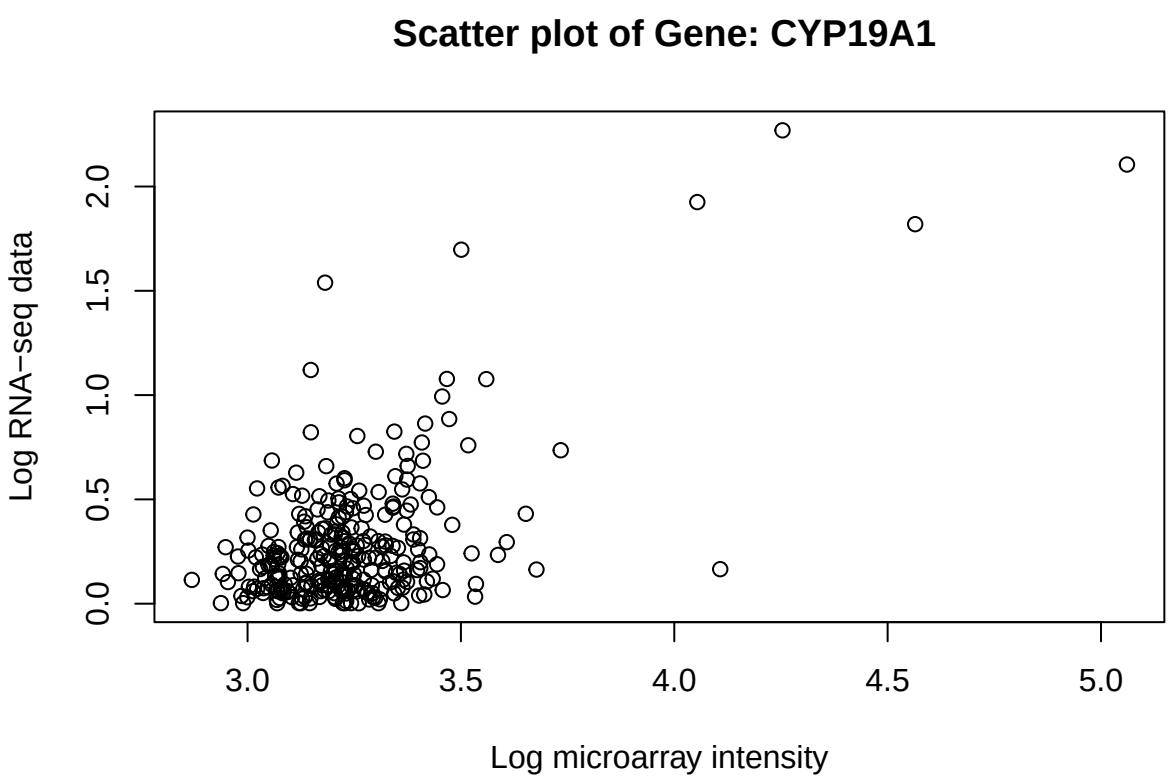

Scatter plot of Gene: C7
Log RNA−seq data
0
2
4
6
8
4
6
8
10
Log microarray intensity

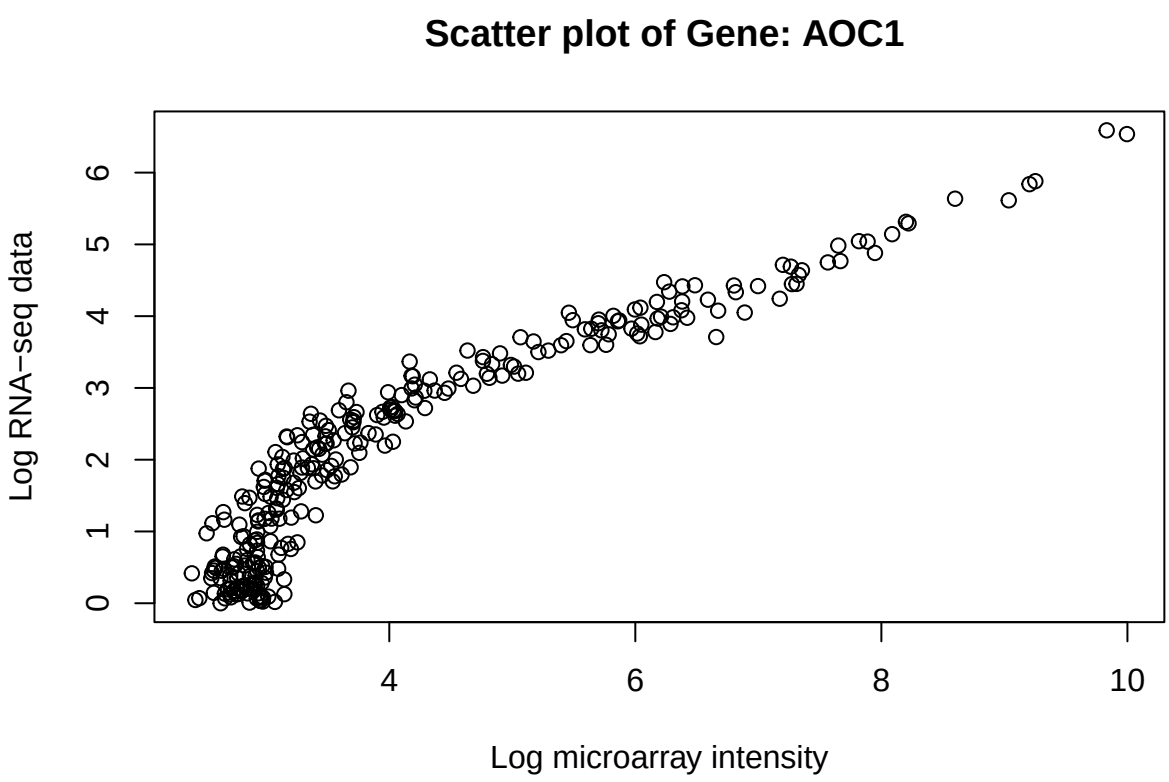

Scatter plot of Gene: GABBR1
Log RNA−seq data
1
2
3
4
5
3
4
5
6
Log microarray intensity

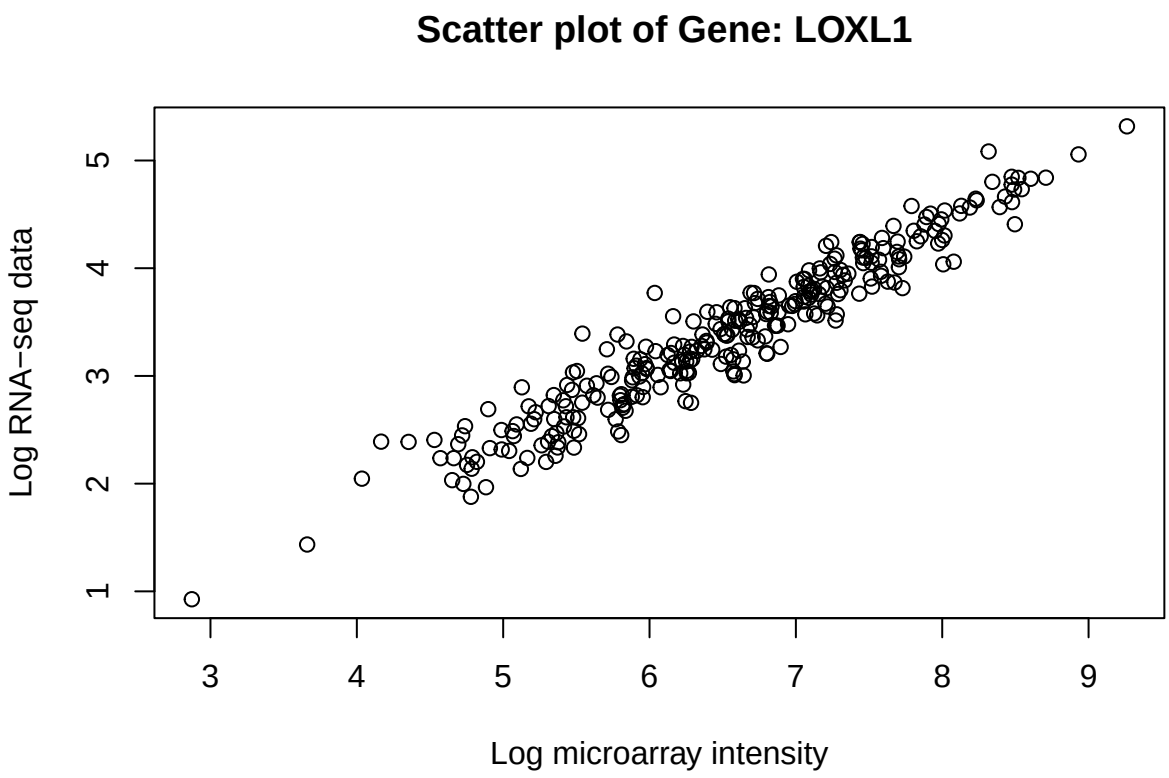

Scatter plot of Gene: NECTIN2
Log RNA−seq data
3.5
4.5
5.5
6.5
5
6
7
8
Log microarray intensity


Scatter plot of Gene: F13A1
Log RNA−seq data
1
2
3
4
5
6
3
4
5
6
7
8
9
Log microarray intensity


Scatter plot of Gene: TGIF1
Log RNA−seq data
3.5
4.0
4.5
5.0
5.5
6
7
8
9
10
Log microarray intensity


Scatter plot of Gene: CYP19A1
Log RNA−seq data
0.0
0.5
1.0
1.5
2.0
3.0
3.5
4.0
4.5
5.0
Log microarray intensity

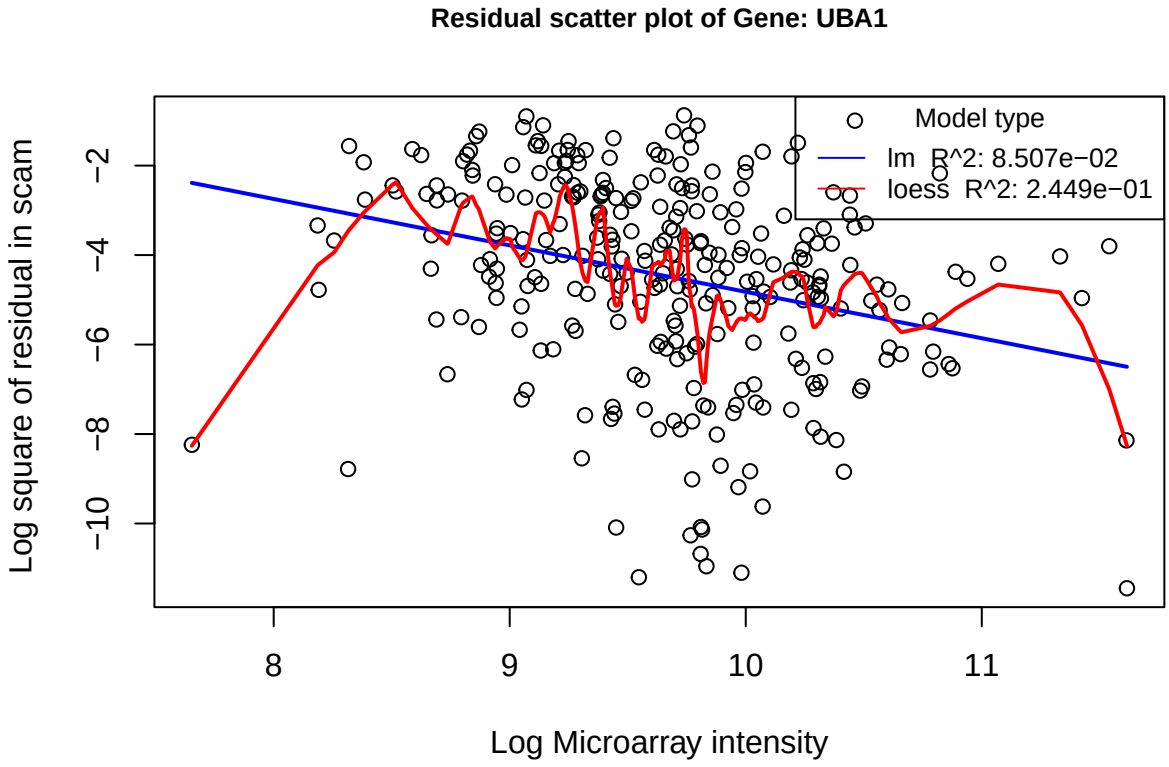

Scatter plot of Gene: AOC1
Log RNA−seq data
0
1
2
3
4
5
6
4
6
8
10
Log microarray intensity

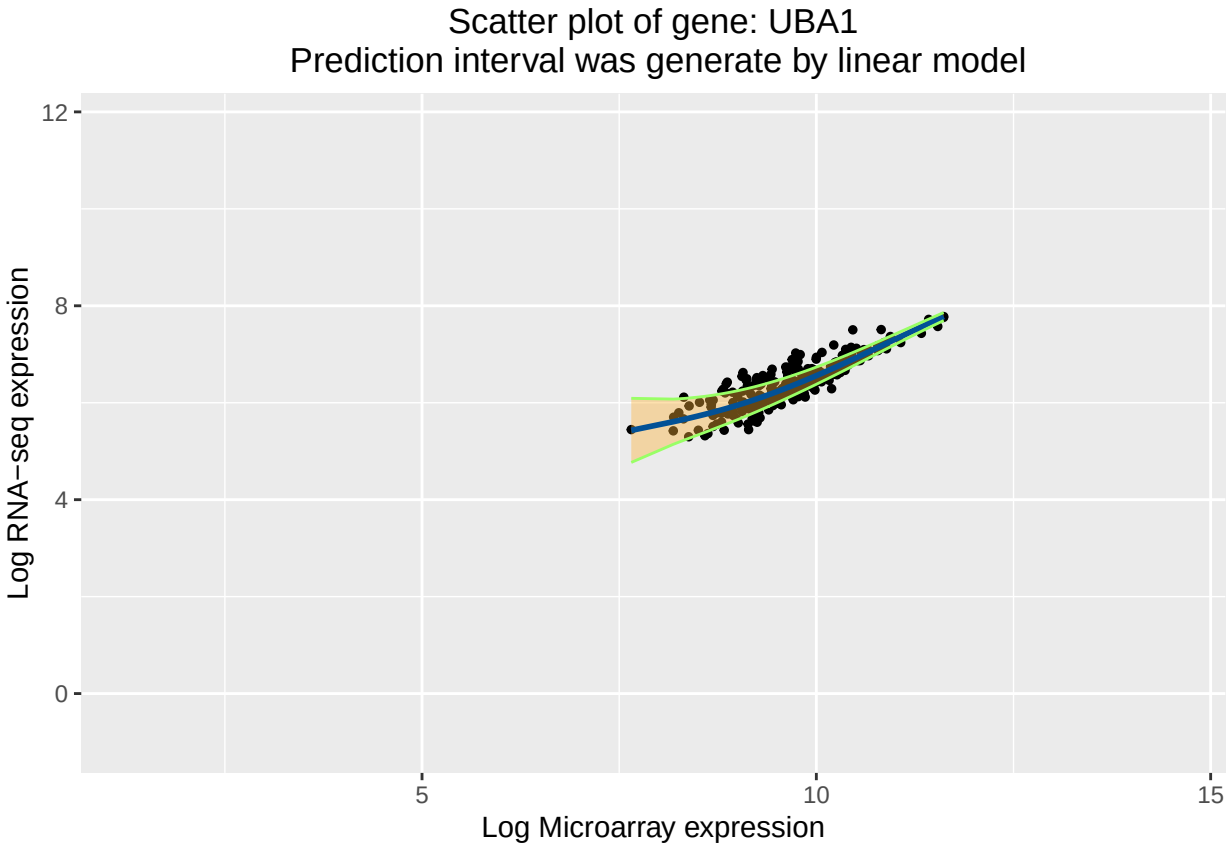


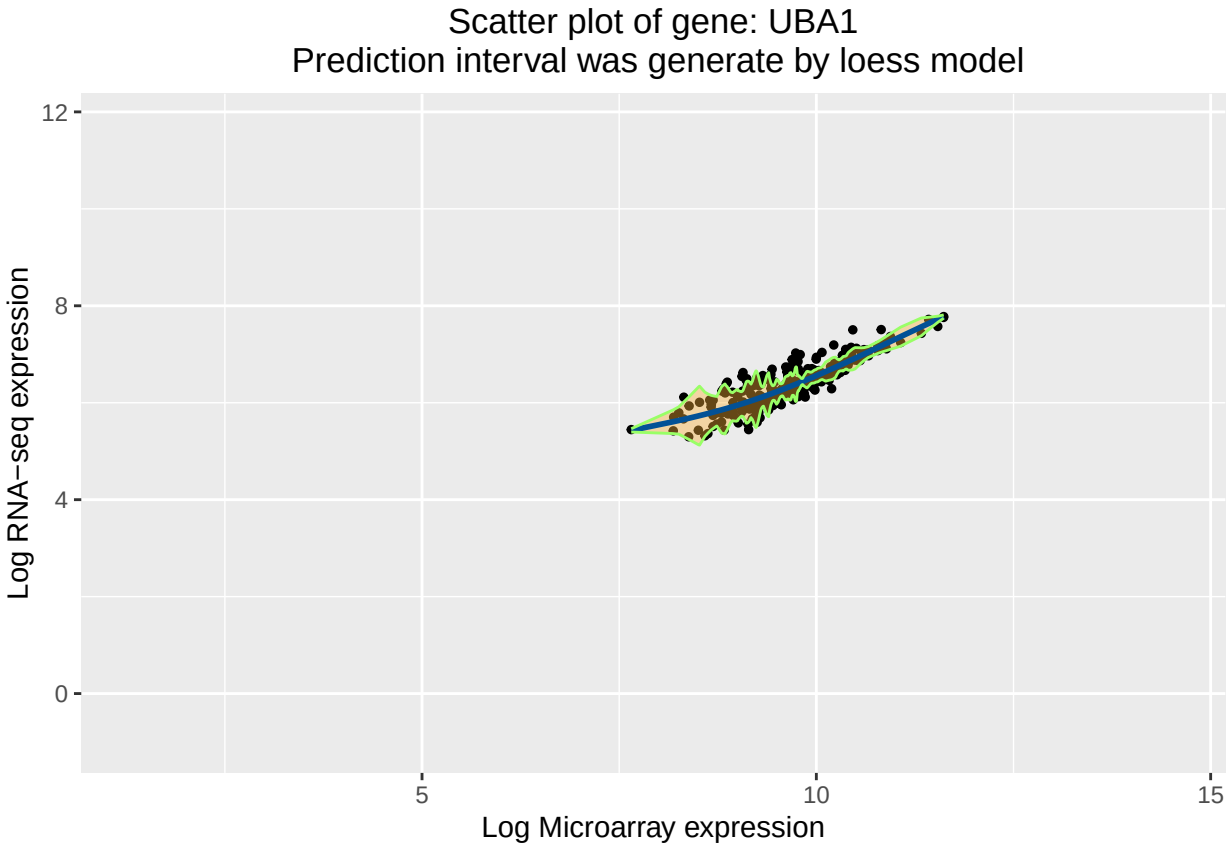
3. Using loess to fit residual model

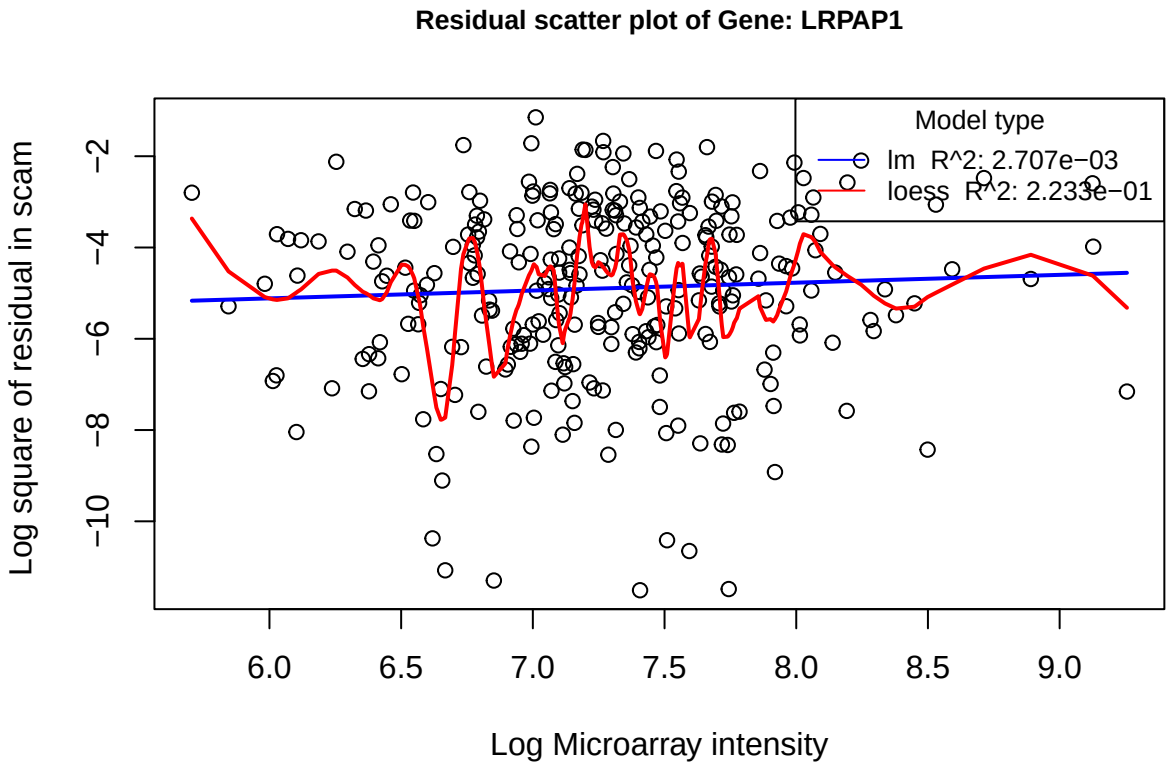
Using linear model to fit residual model isn't appropriate because it is monotonous, which showed in the prediction interval as either increase gradually or decrease in width. But the variance of our data maybe increasing in the beginning and then decreasing. So I use nonlinear model to fit residual model, like loess.

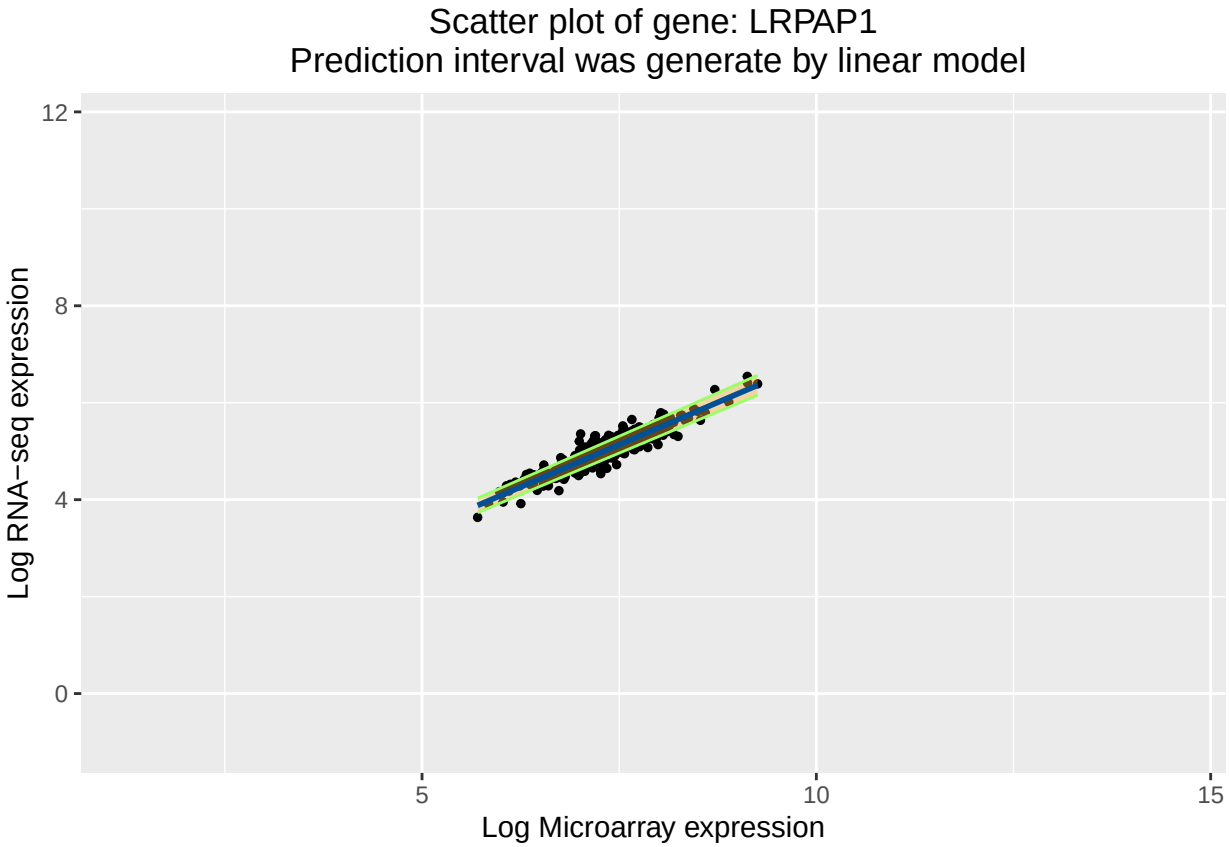
```
for (i in idx_R0.2) {
  log_res2 <- log(fit_scam_r2_lm$maps[[i]]$model$residuals^2 + 1e-5)
  plot(x[i, ], log_res2, main = paste0("Residual scatter plot of Gene: ", rownames(x)[i]),
       xlab = "Log Microarray intensity", ylab = "Log square of residual in scam ",cex.main = 0.8)
  idx <- order(x[i, ])
  lines(x[i, ][idx], fit_scam_r2_lm$maps[[i]]$res_model$fitted.values[idx], col = "blue", lwd = 2)
  lines(x[i, ][idx], fit_scam_r2_loess$maps[[i]]$res_model$fitted[idx], col = "red", lwd = 2)
  legend("topright", legend = c(paste0("lm", "  R^2: ", sprintf("%.3e", R2$R2_r2_lm[i])),
                                paste0("loess", "  R^2: ",sprintf("%.3e", R2$R2_r2_loess[i]))),
         col = c("blue", "red"), lty = 1, title = "Model type", cex = 0.8)

  scatter(i, x, y, pred = pred_scam_r2_lm, tittle = paste0("Scatter plot of gene: ", rownames(x)[i], "\n


  scatter(i, x, y, pred = pred_scam_r2_loess, tittle = paste0("Scatter plot of gene: ", rownames(x)[i],

}
```

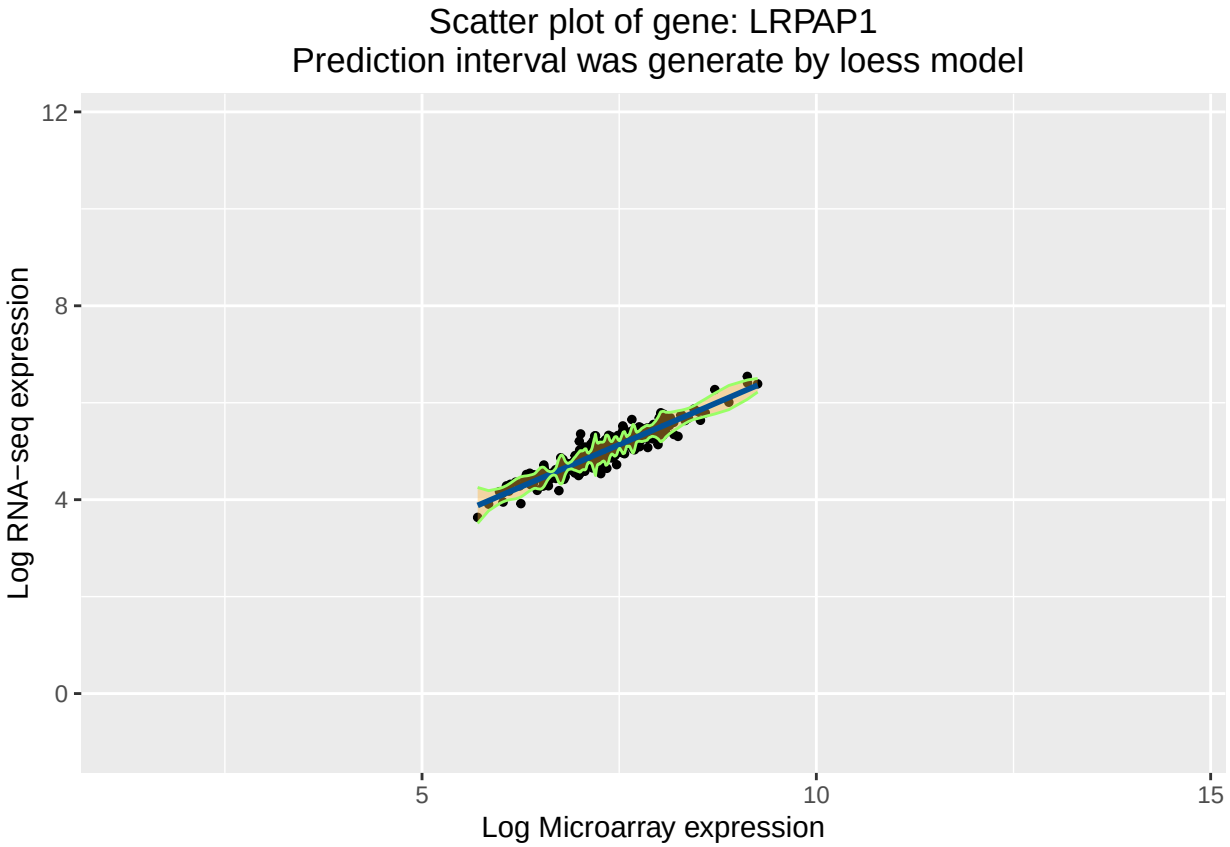

Residual scatter plot of Gene: UBA1
Model type
lm R^2: 8.507e−02
loess R^2: 2.449e−01
Log square of residual in scam
−2
−4
−6
−8
−10
8
9
10
11
Log Microarray intensity

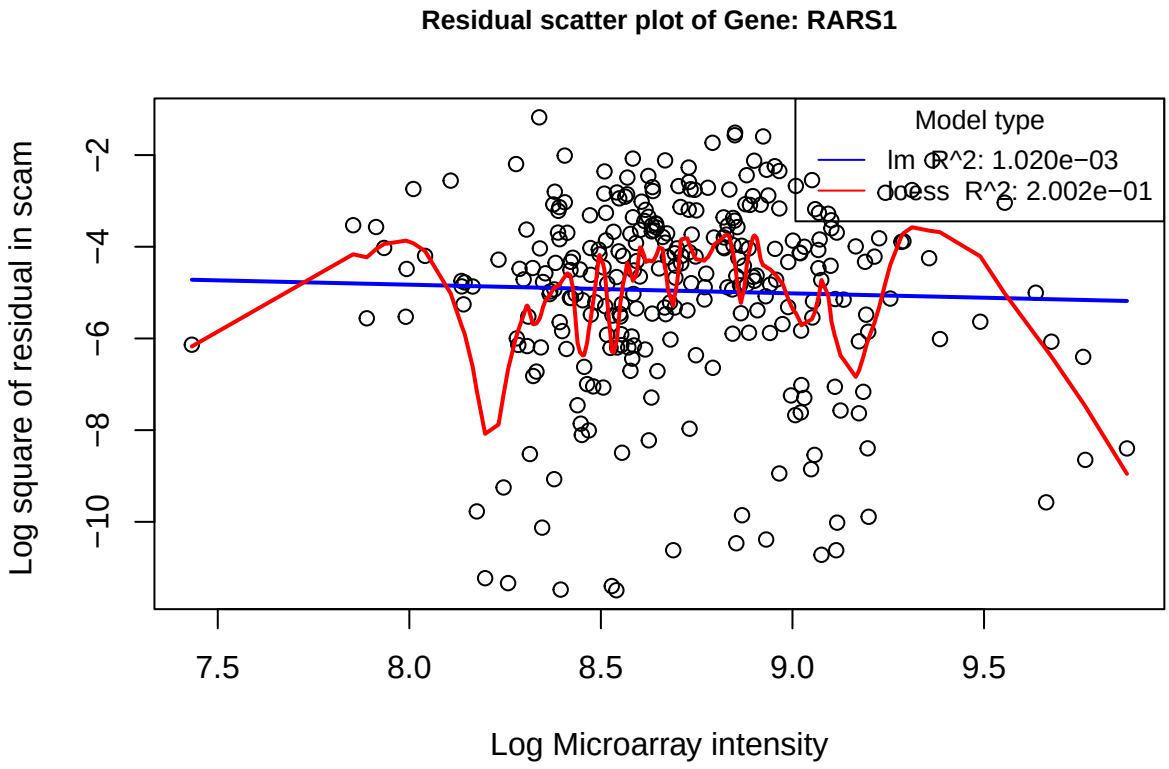

Scatter plot of gene: UBA1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

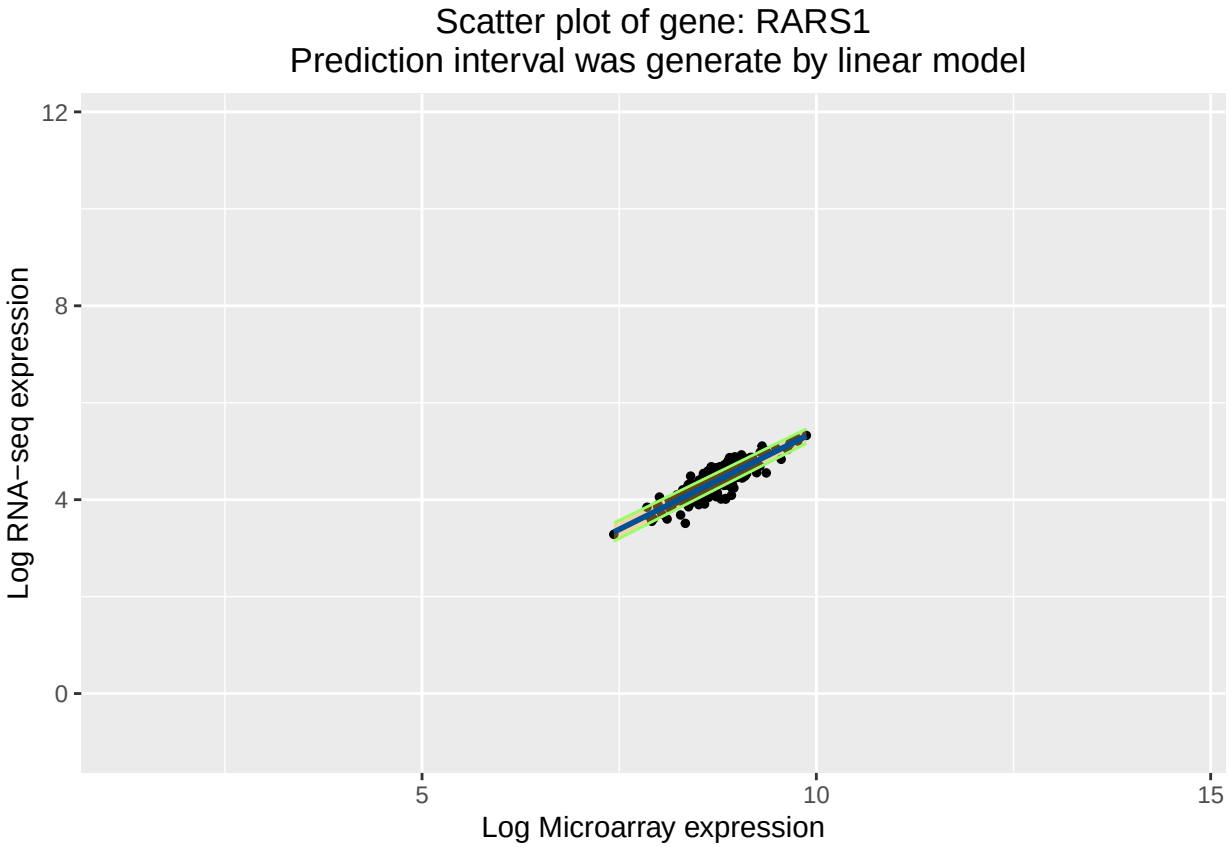

Scatter plot of gene: UBA1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

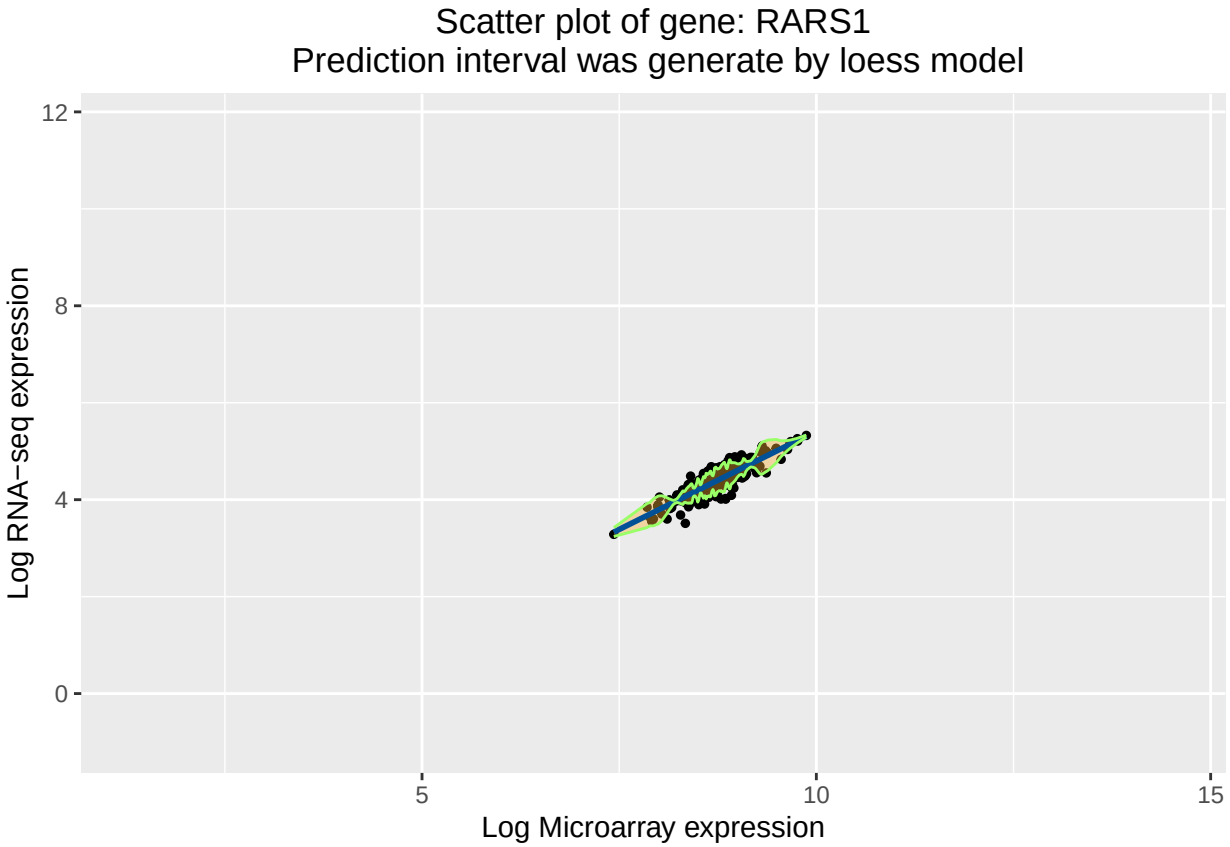

Residual scatter plot of Gene: LRPAP1
Model type
lm R^2: 2.707e−03
loess R^2: 2.233e−01
Log square of residual in scam
−2
−4
−6
−8
−10
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Log Microarray intensity

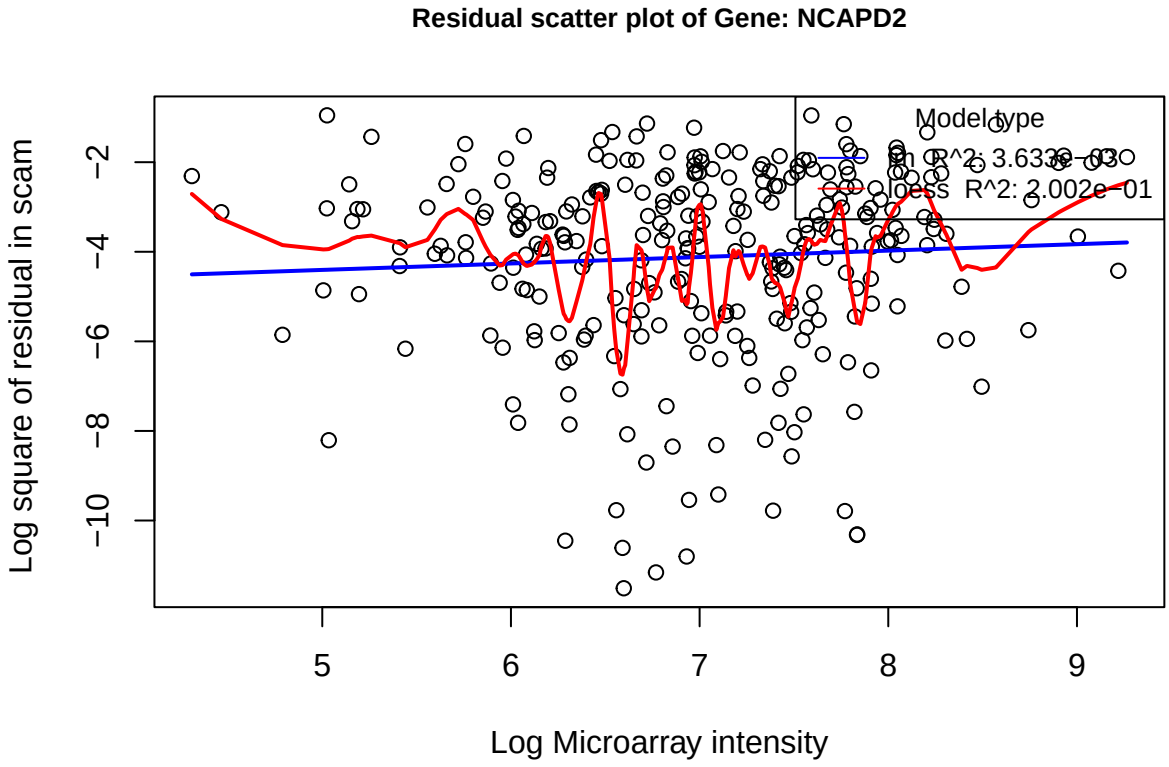

Scatter plot of gene: LRPAP1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

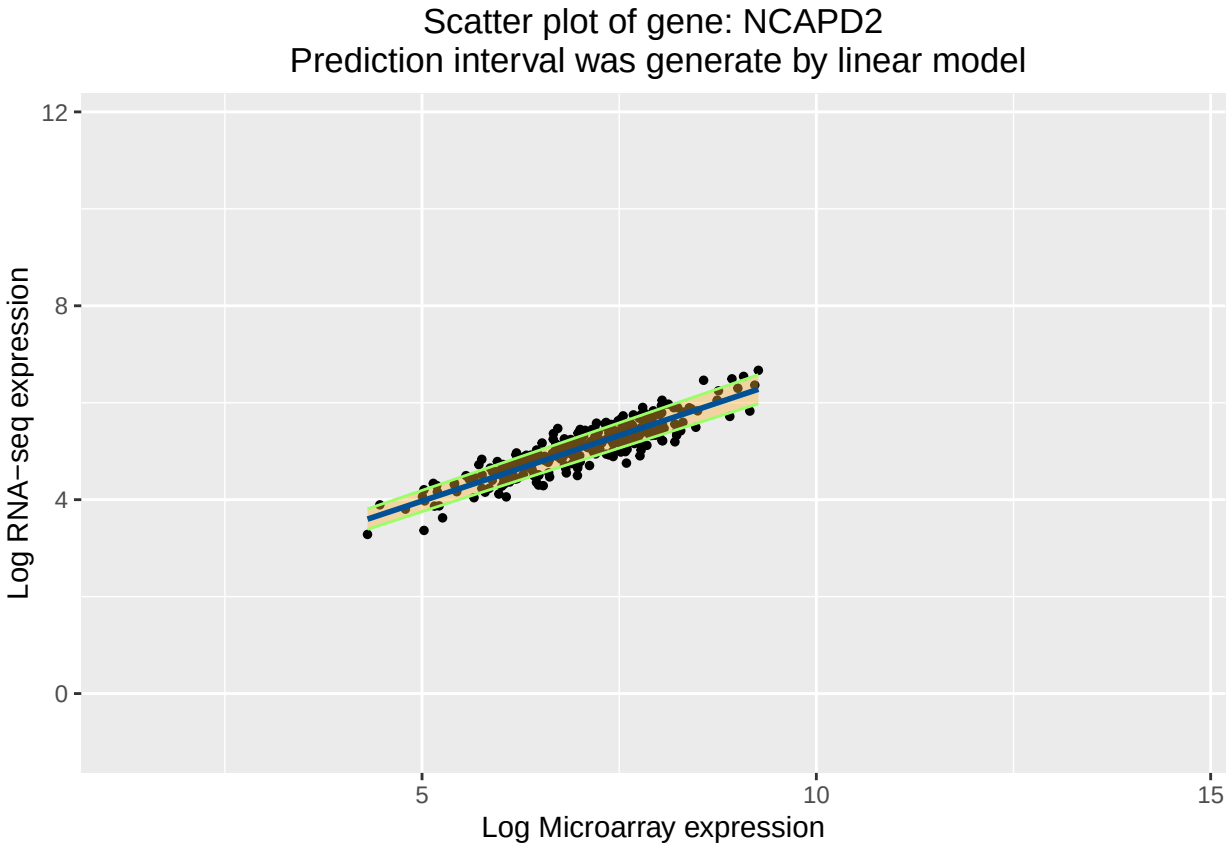

Scatter plot of gene: LRPAP1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

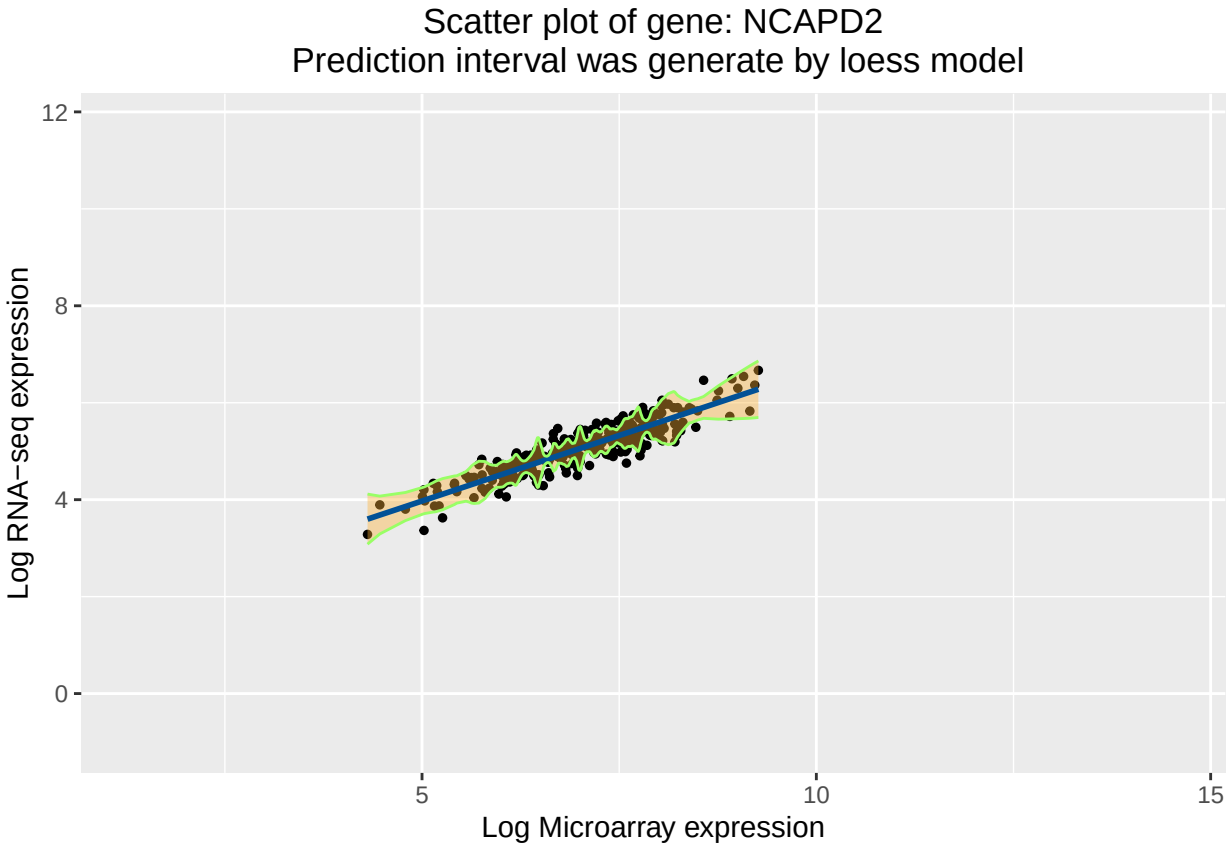

Residual scatter plot of Gene: RARS1
Model type
lm R^2: 1.020e−03
loess R^2: 2.002e−01
Log square of residual in scam
−2
−4
−6
−8
−10
7.5
8.0
8.5
9.0
9.5
Log Microarray intensity

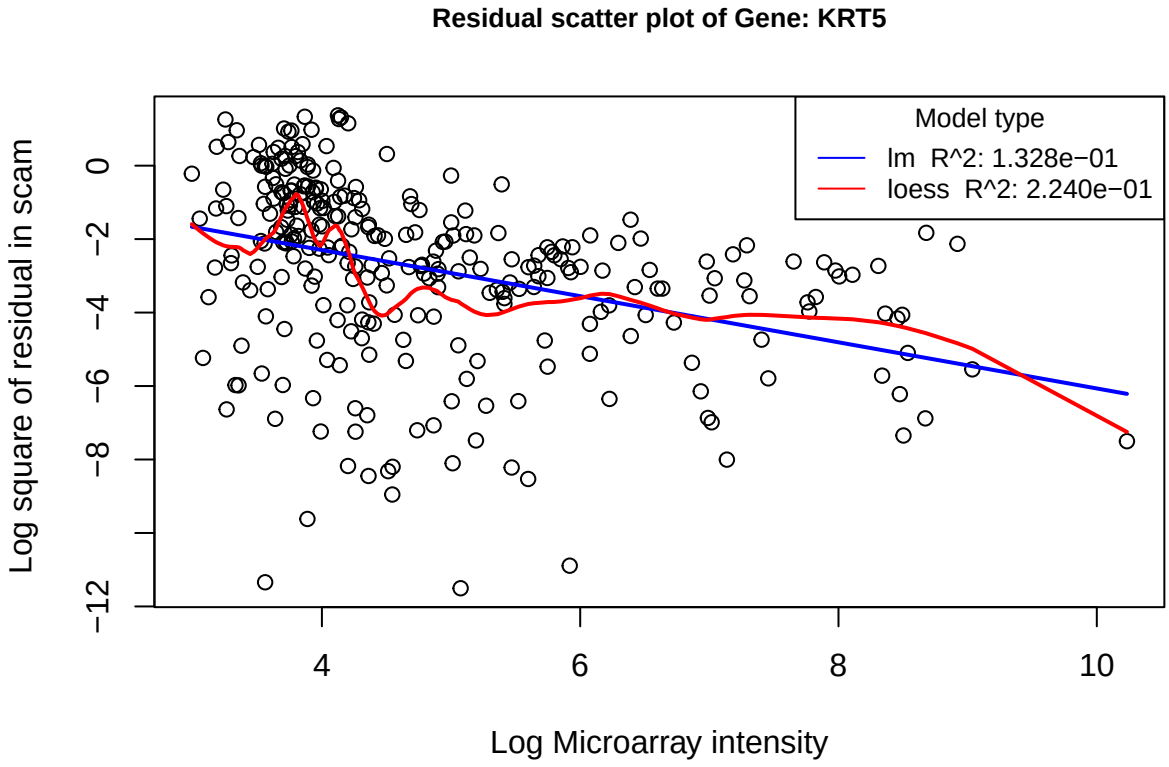

Scatter plot of gene: RARS1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

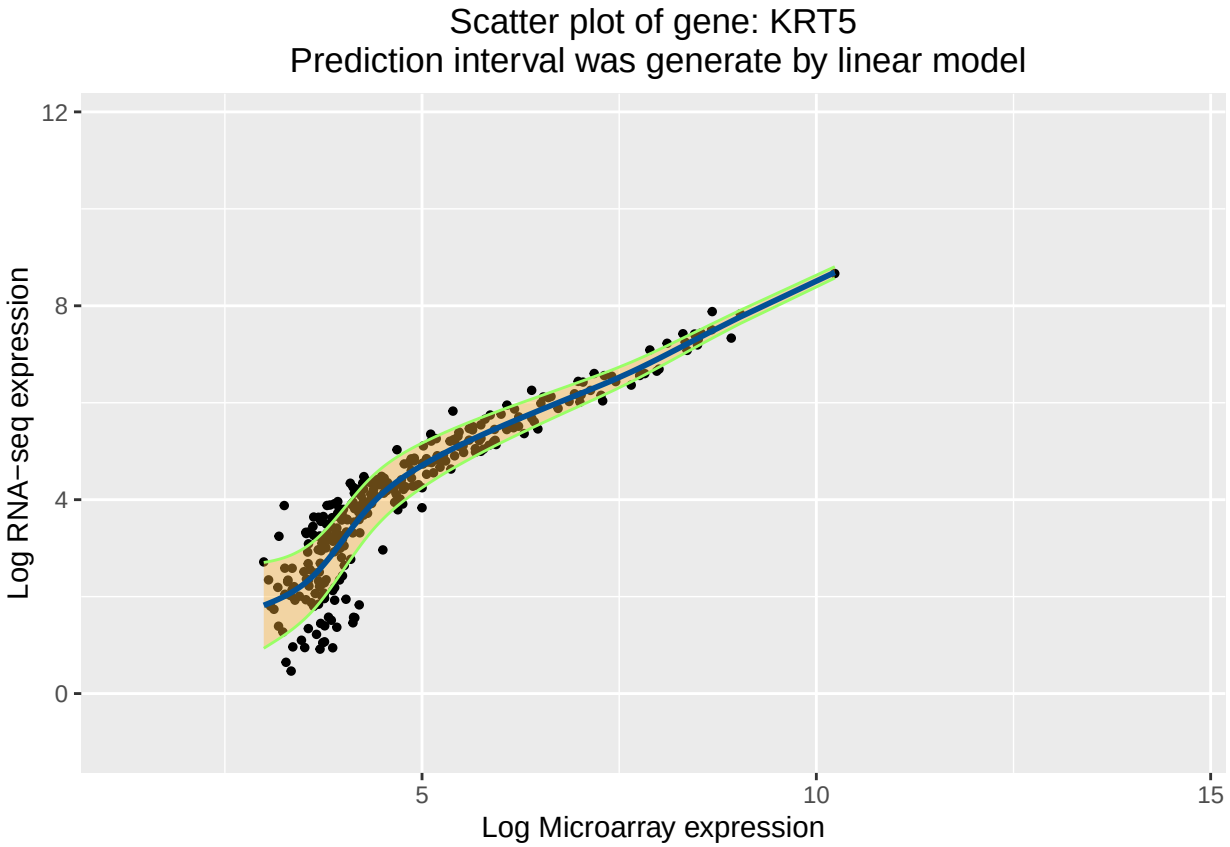

Scatter plot of gene: RARS1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

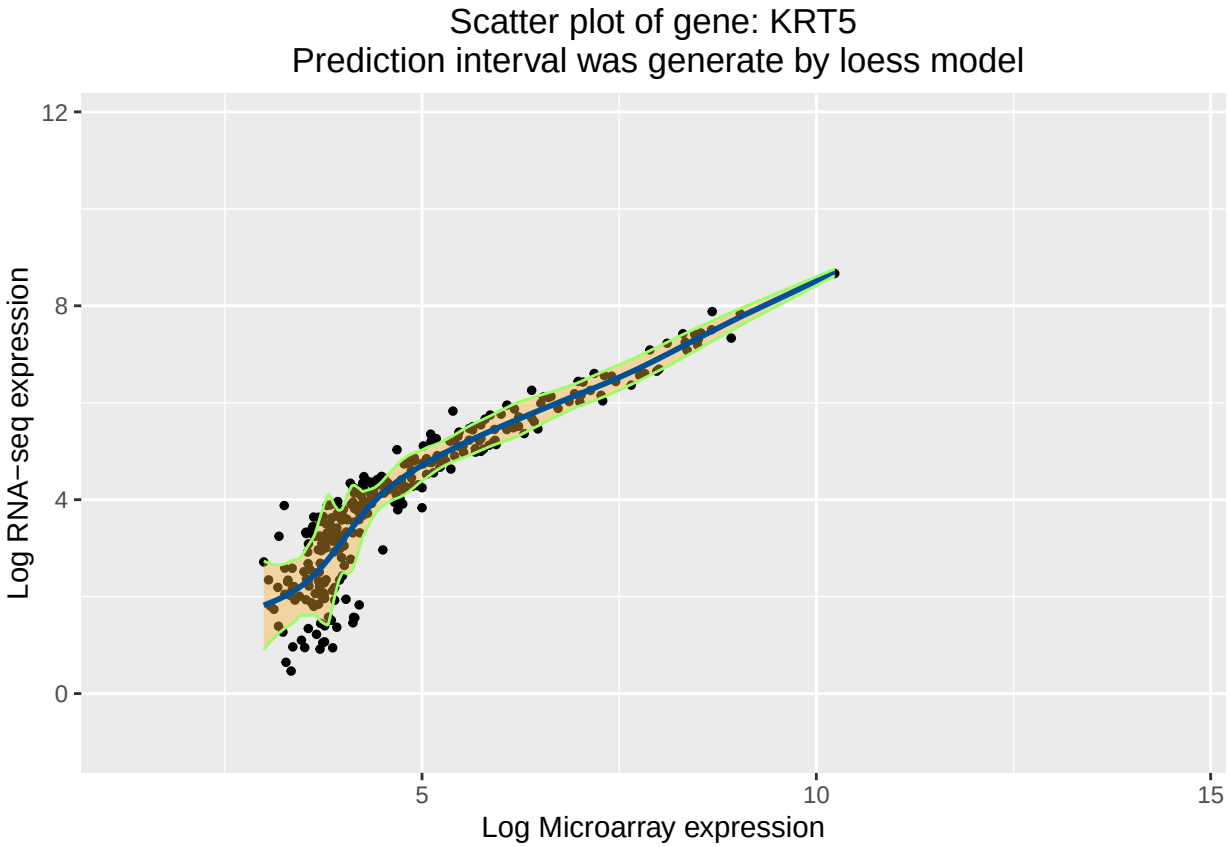

Residual scatter plot of Gene: NCAPD2
Model type
lm R^2: 3.633e−03
loess R^2: 2.002e−01
Log square of residual in scam
−2
−4
−6
−8
−10
5
6
7
8
9
Log Microarray intensity

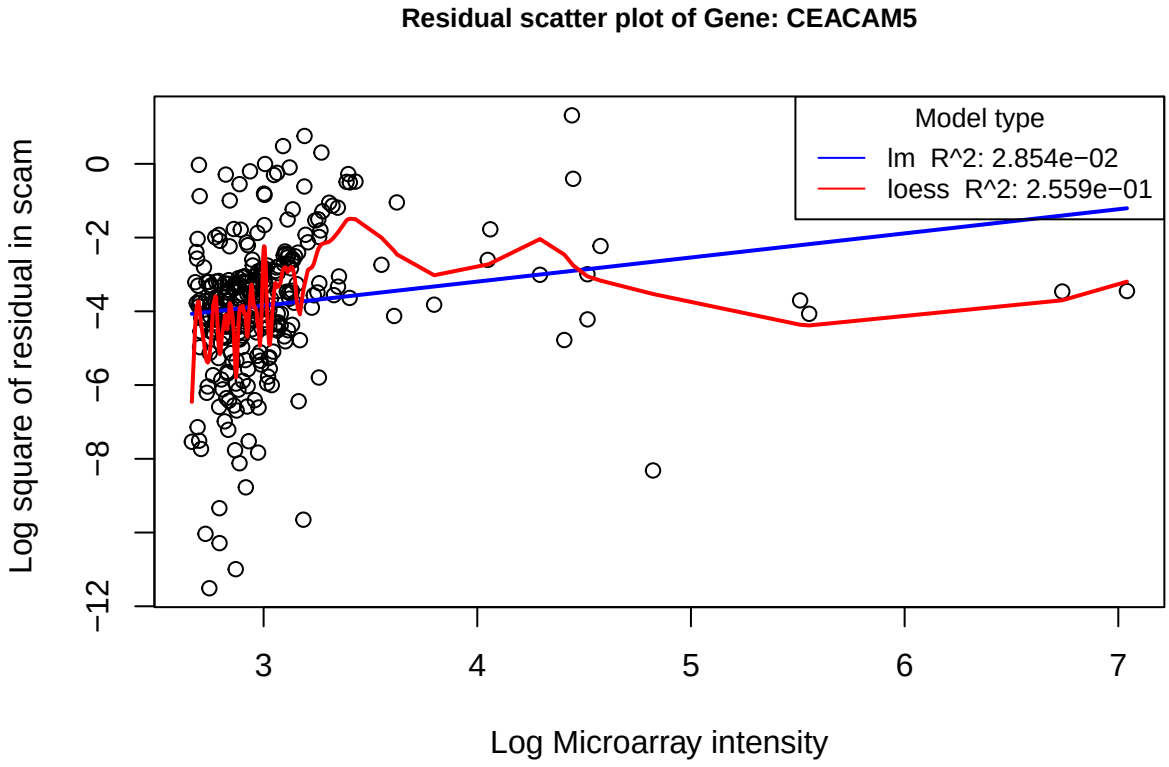

Scatter plot of gene: NCAPD2
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

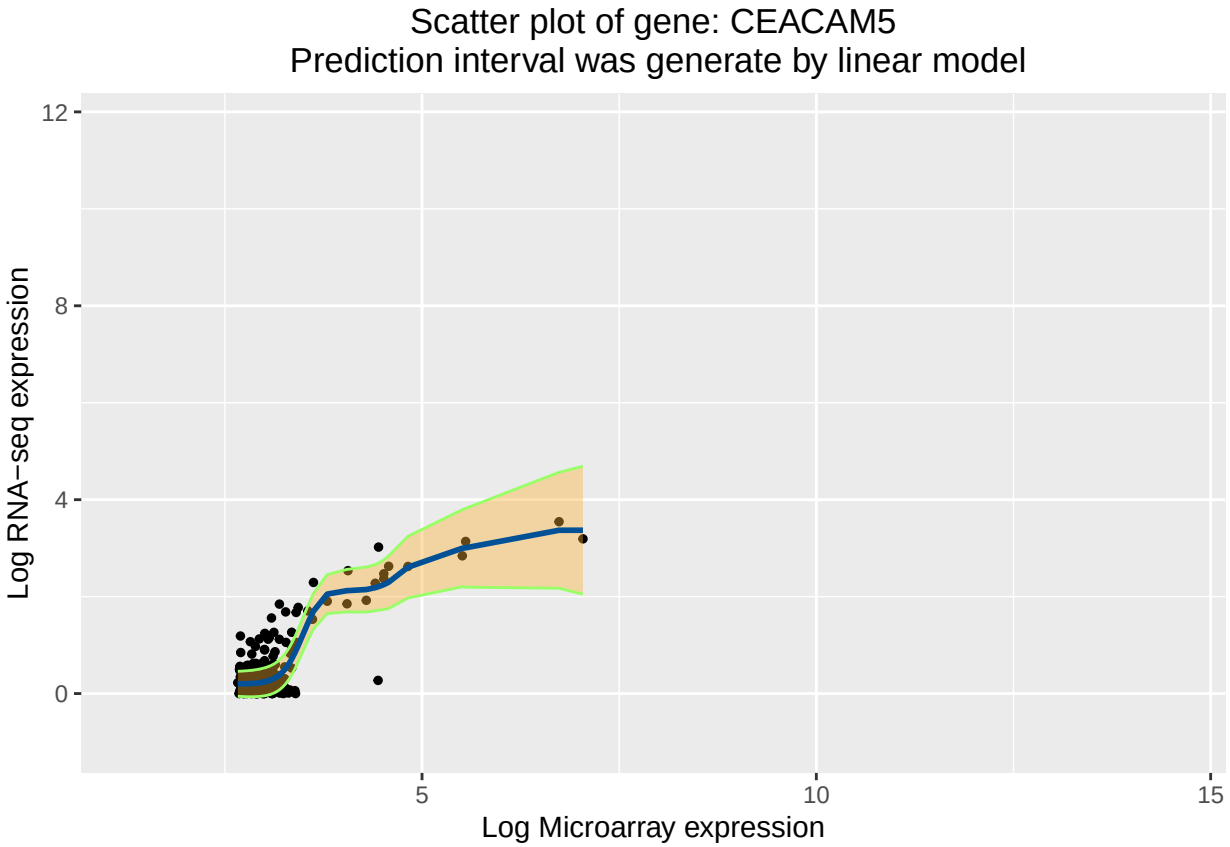

Scatter plot of gene: NCAPD2
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

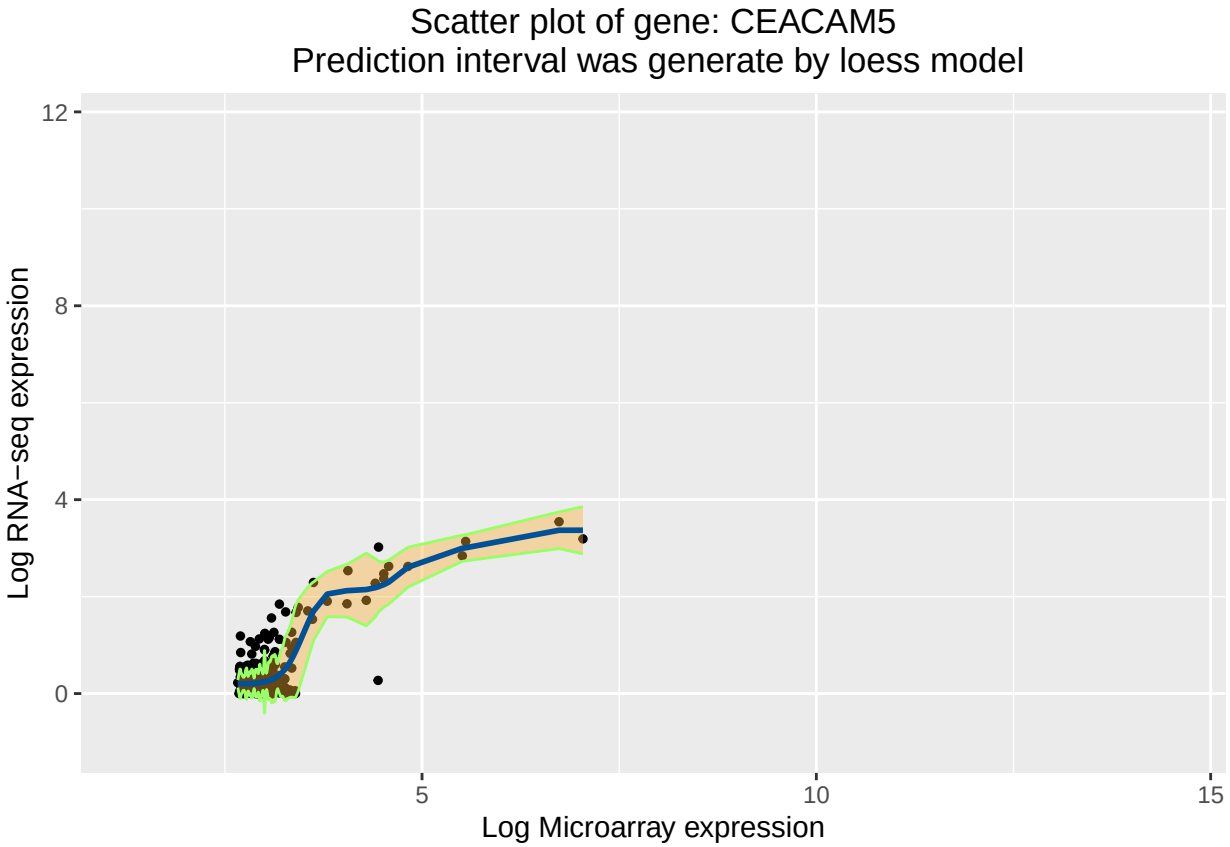

Residual scatter plot of Gene: KRT5
Model type
lm R^2: 1.328e−01
loess R^2: 2.240e−01
Log square of residual in scam
0
−2
−4
−6
−8
−12
4
6
8
10
Log Microarray intensity

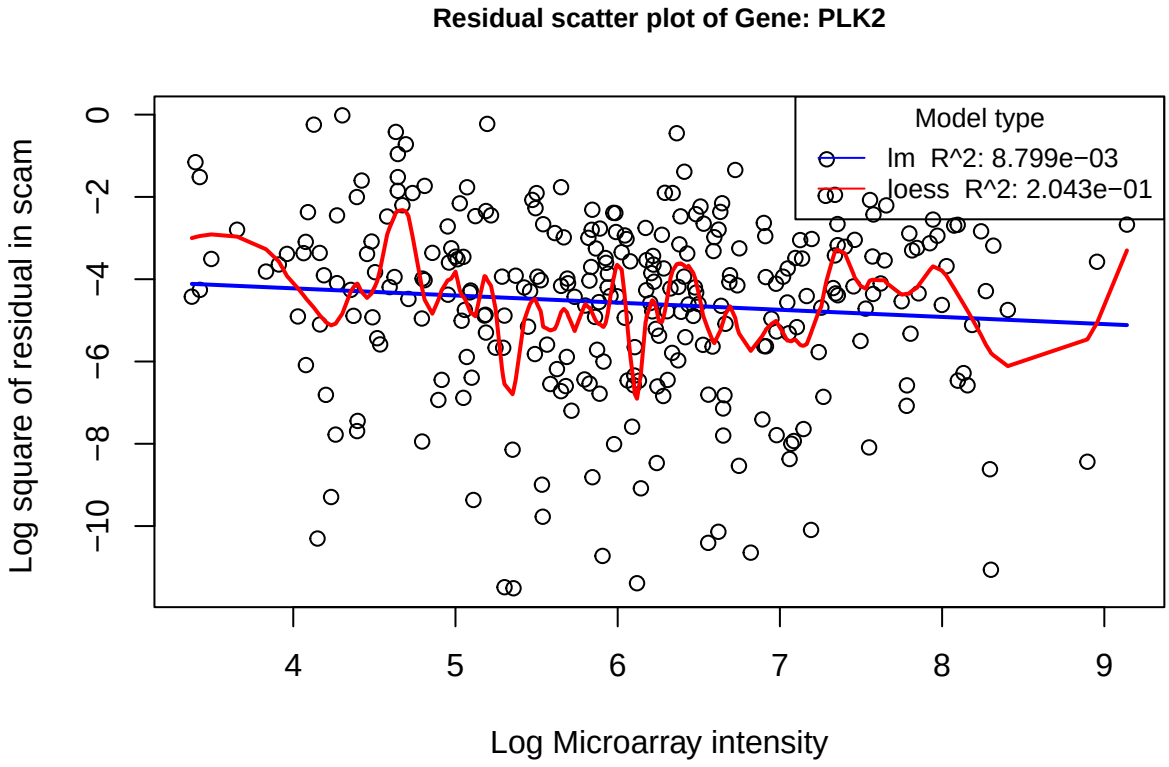

Scatter plot of gene: KRT5
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

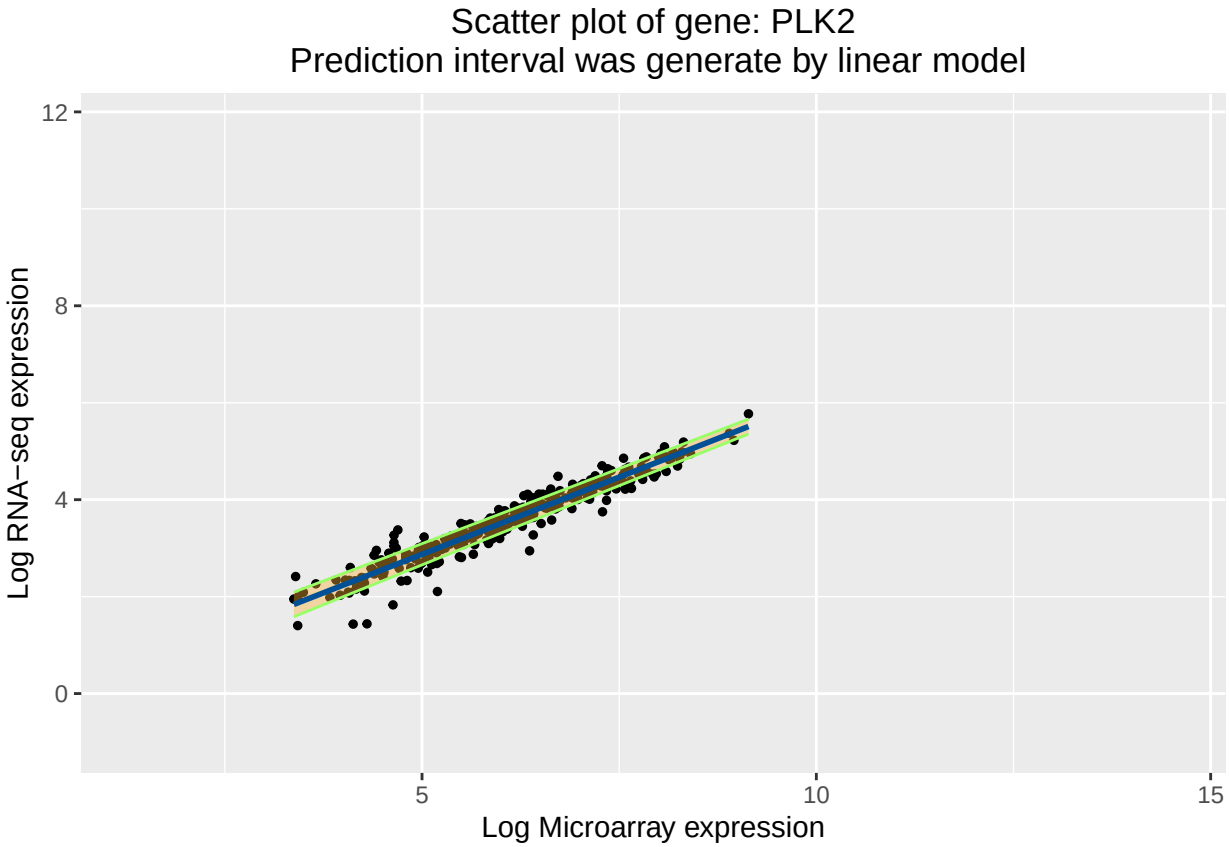

Scatter plot of gene: KRT5
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

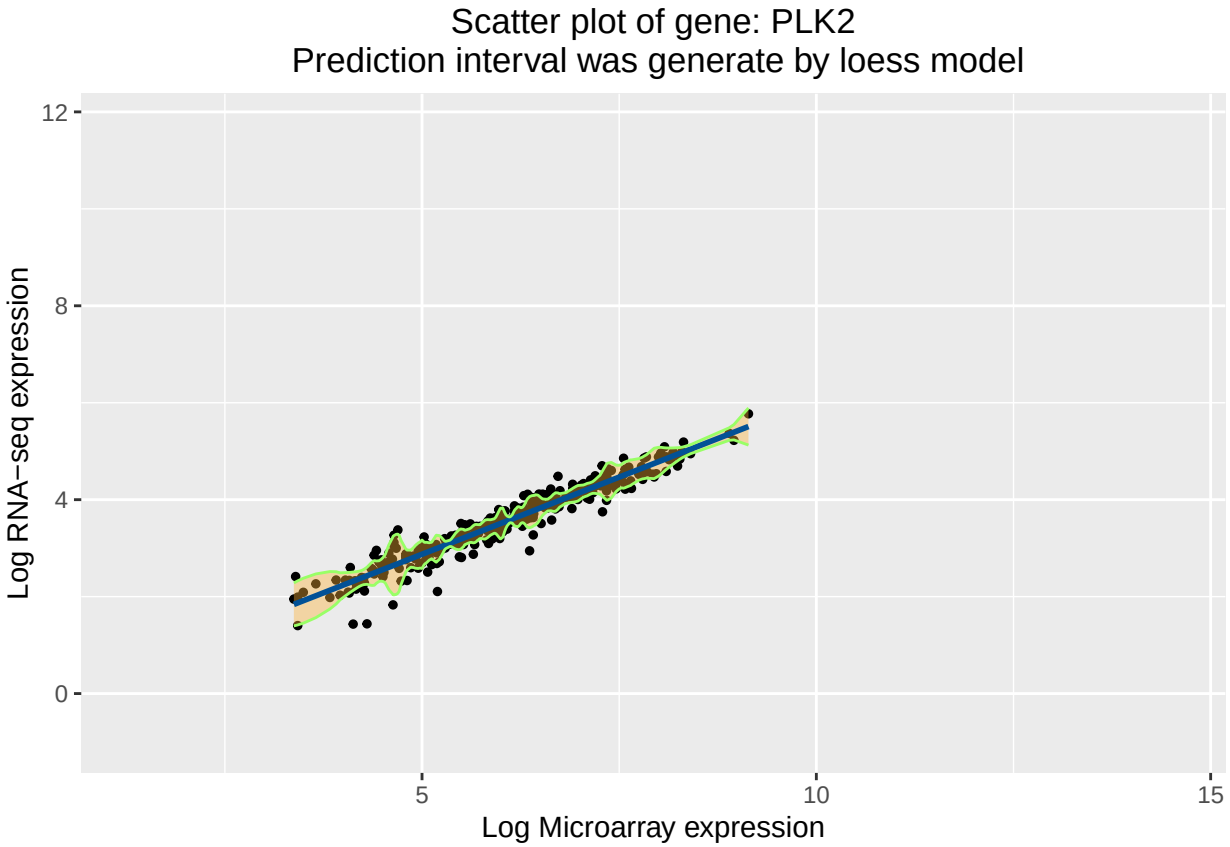

Residual scatter plot of Gene: CEACAM5
Model type
lm R^2: 2.854e−02
loess R^2: 2.559e−01
Log square of residual in scam
0
−2
−4
−6
−8
−12
3
4
5
6
7
Log Microarray intensity

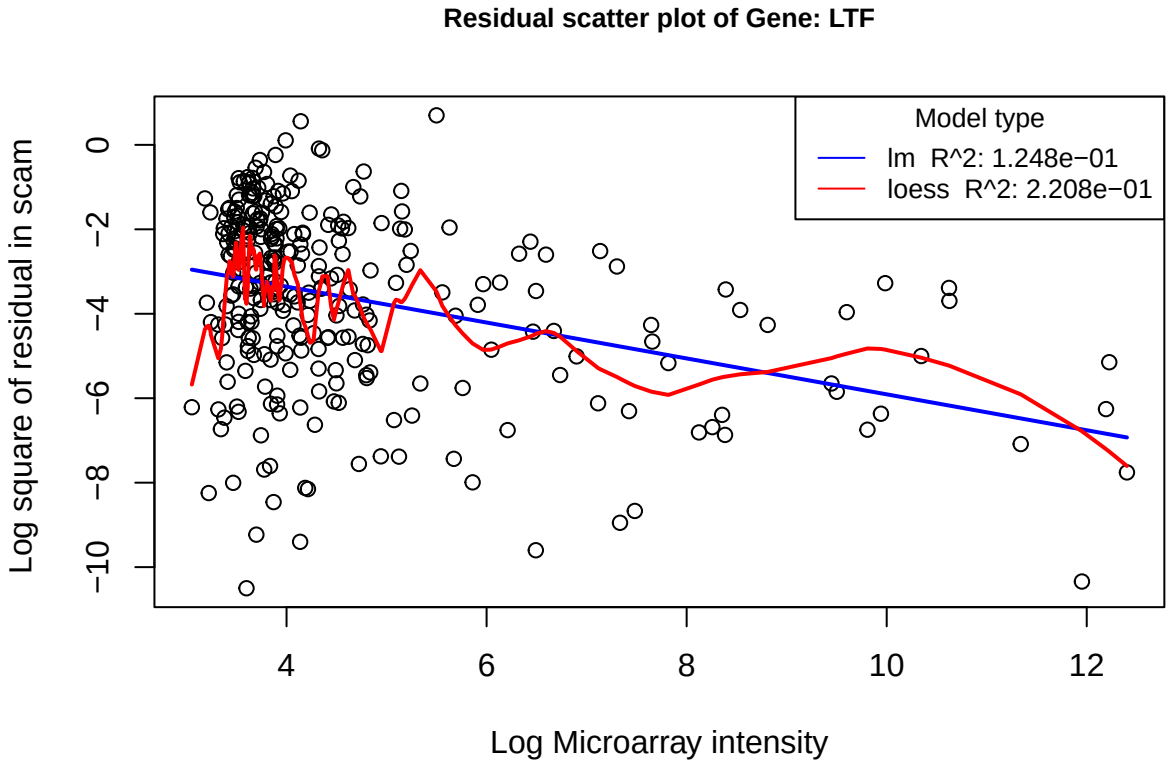

Scatter plot of gene: CEACAM5
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

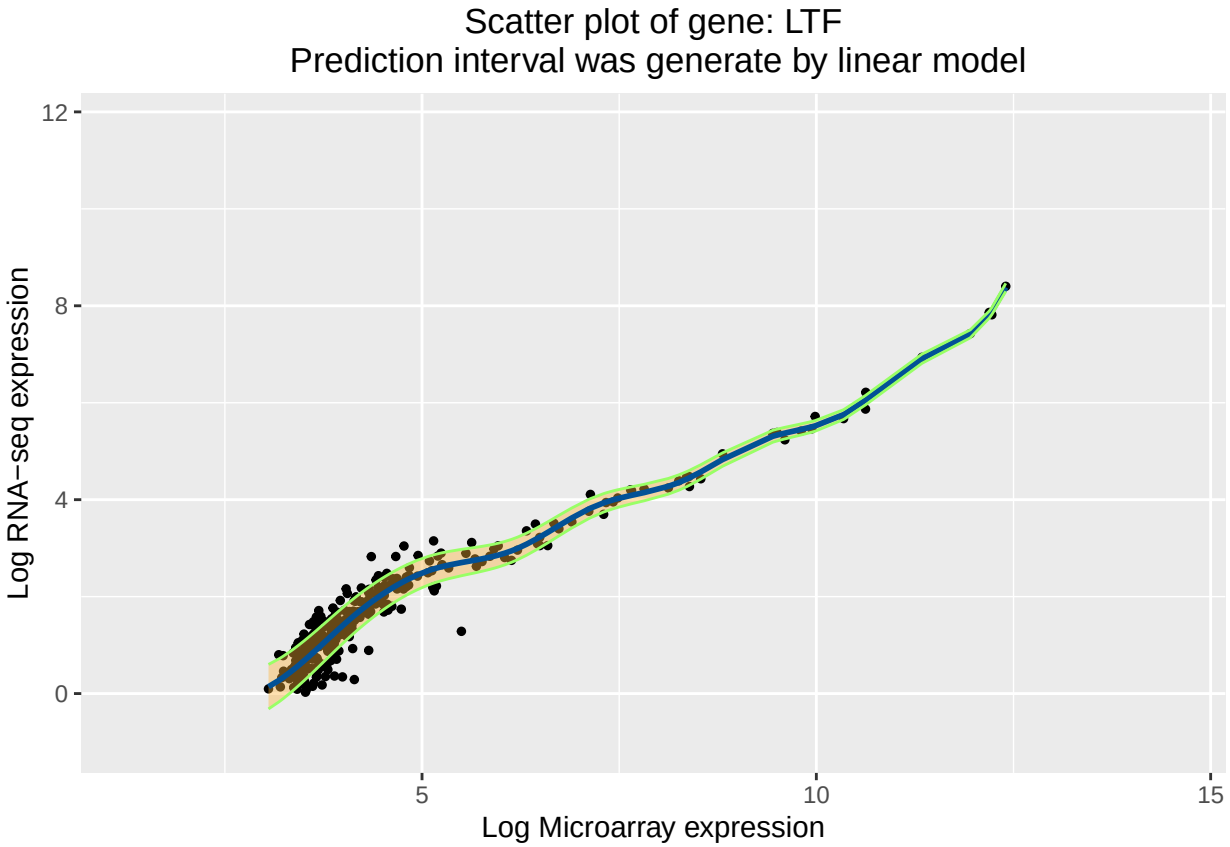

Scatter plot of gene: CEACAM5
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

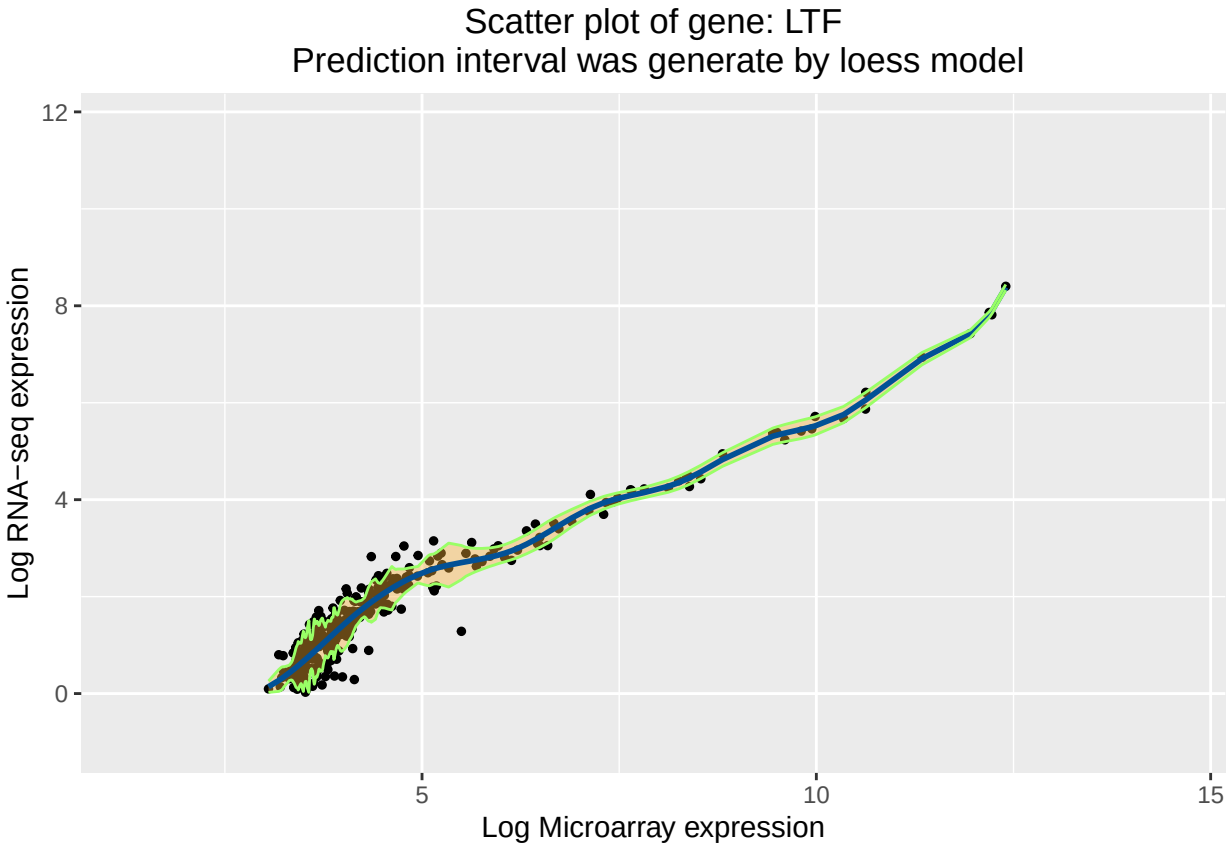

Residual scatter plot of Gene: PLK2
Model type
lm R^2: 8.799e−03
loess R^2: 2.043e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
4
5
6
7
8
9
Log Microarray intensity

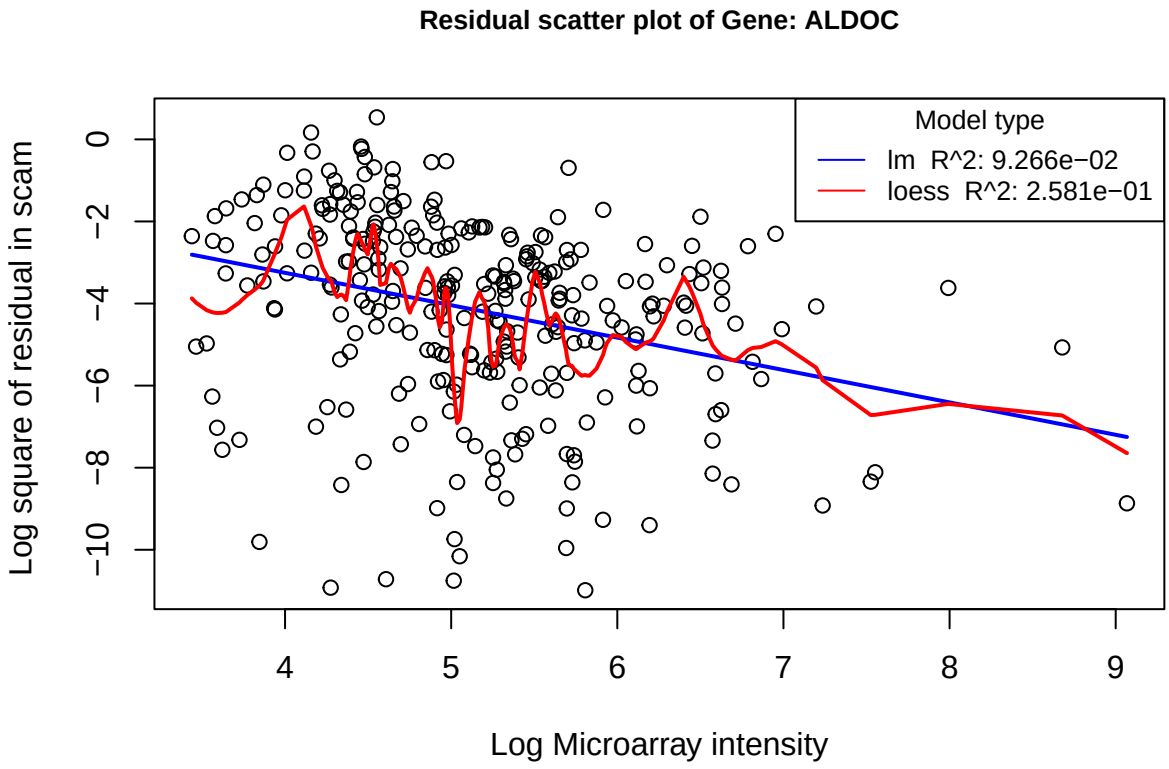

Scatter plot of gene: PLK2
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression


Scatter plot of gene: PLK2
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression


Residual scatter plot of Gene: LTF
Model type
lm R^2: 1.248e−01
loess R^2: 2.208e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
4
6
8
10
12
Log Microarray intensity

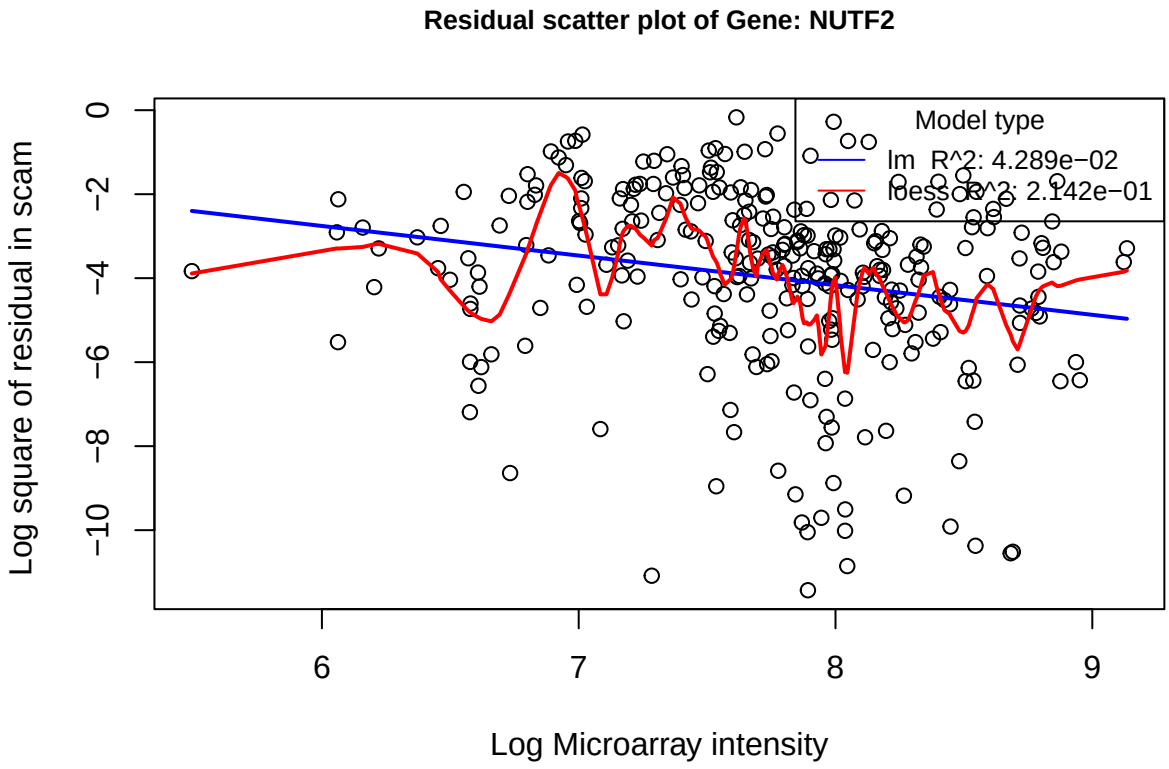

Scatter plot of gene: LTF
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

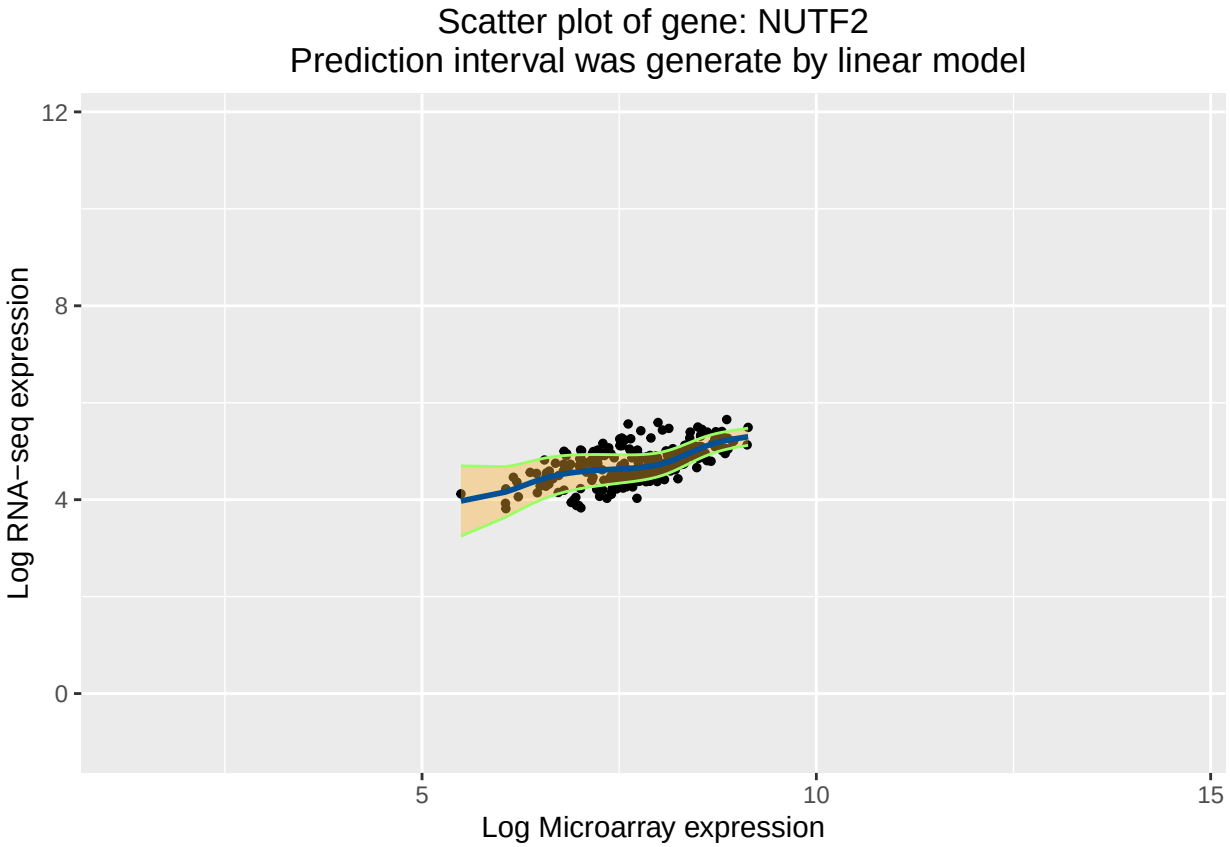

Scatter plot of gene: LTF
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

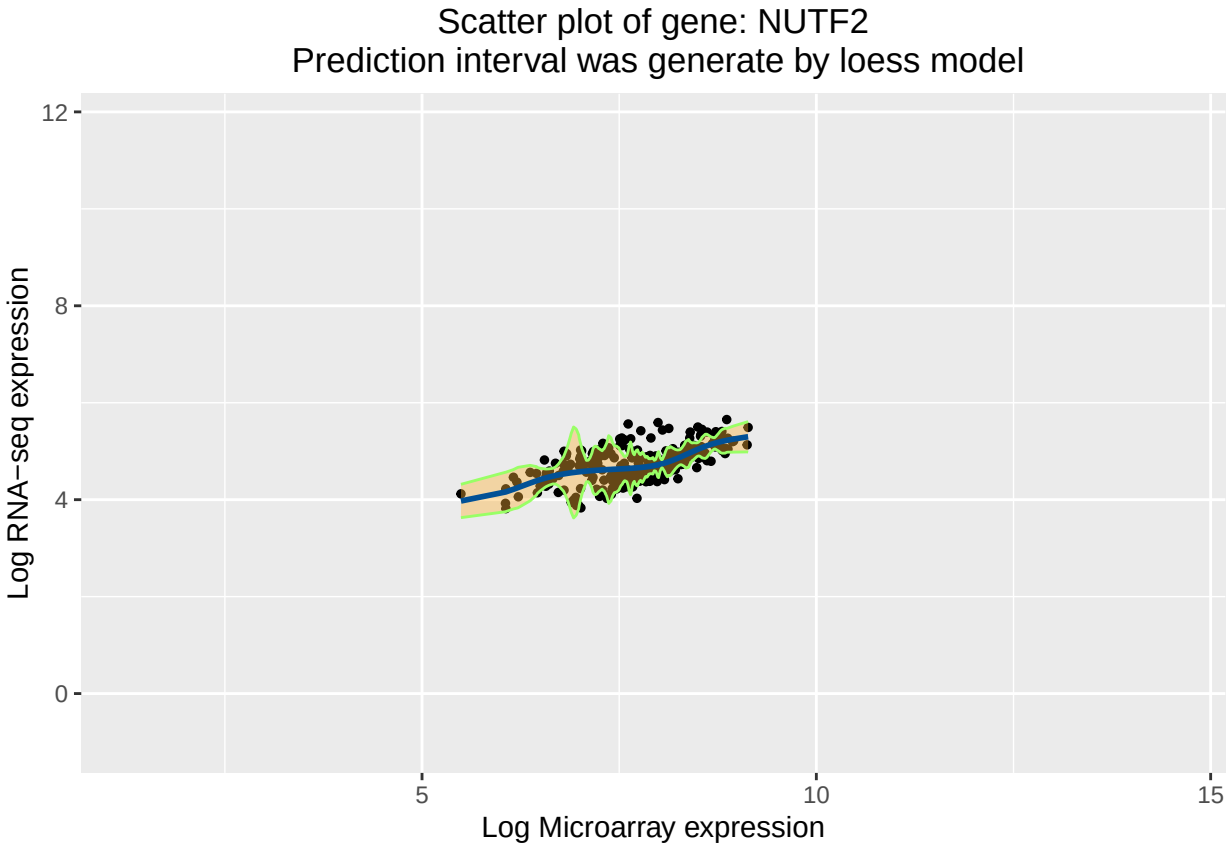

Residual scatter plot of Gene: ALDOC
Model type
lm R^2: 9.266e−02
loess R^2: 2.581e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
4
5
6
7
8
9
Log Microarray intensity

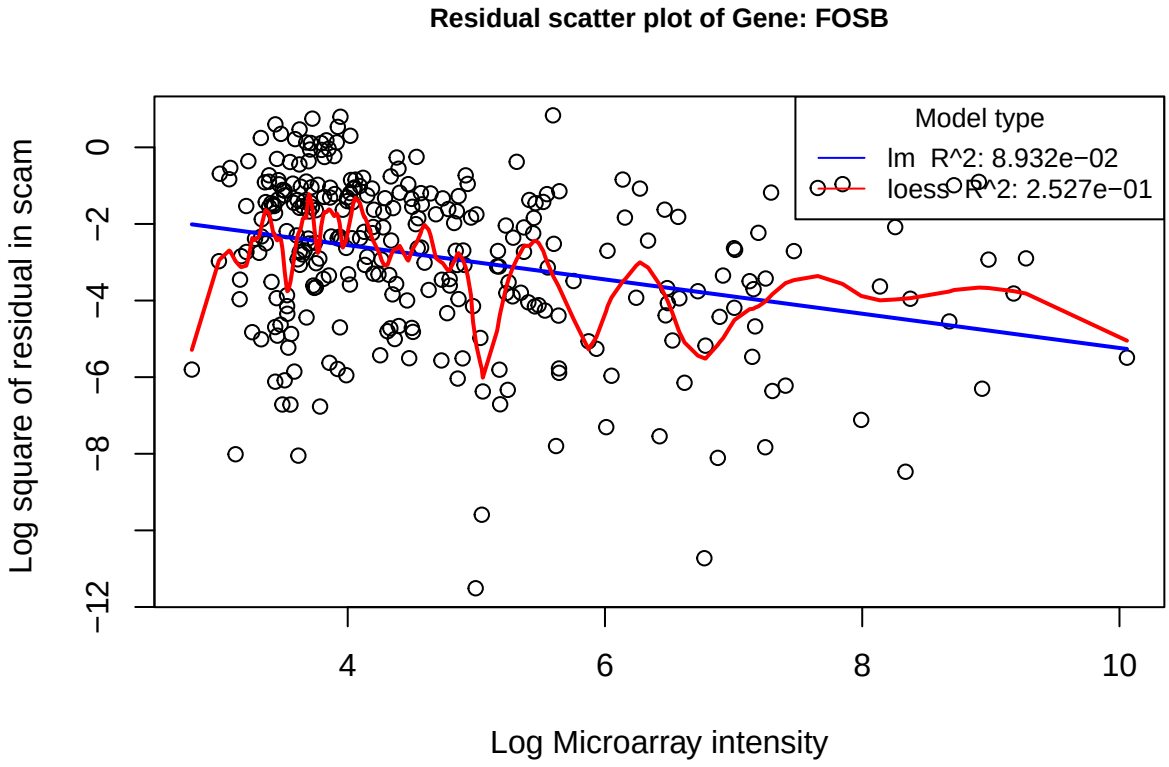

Scatter plot of gene: ALDOC
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

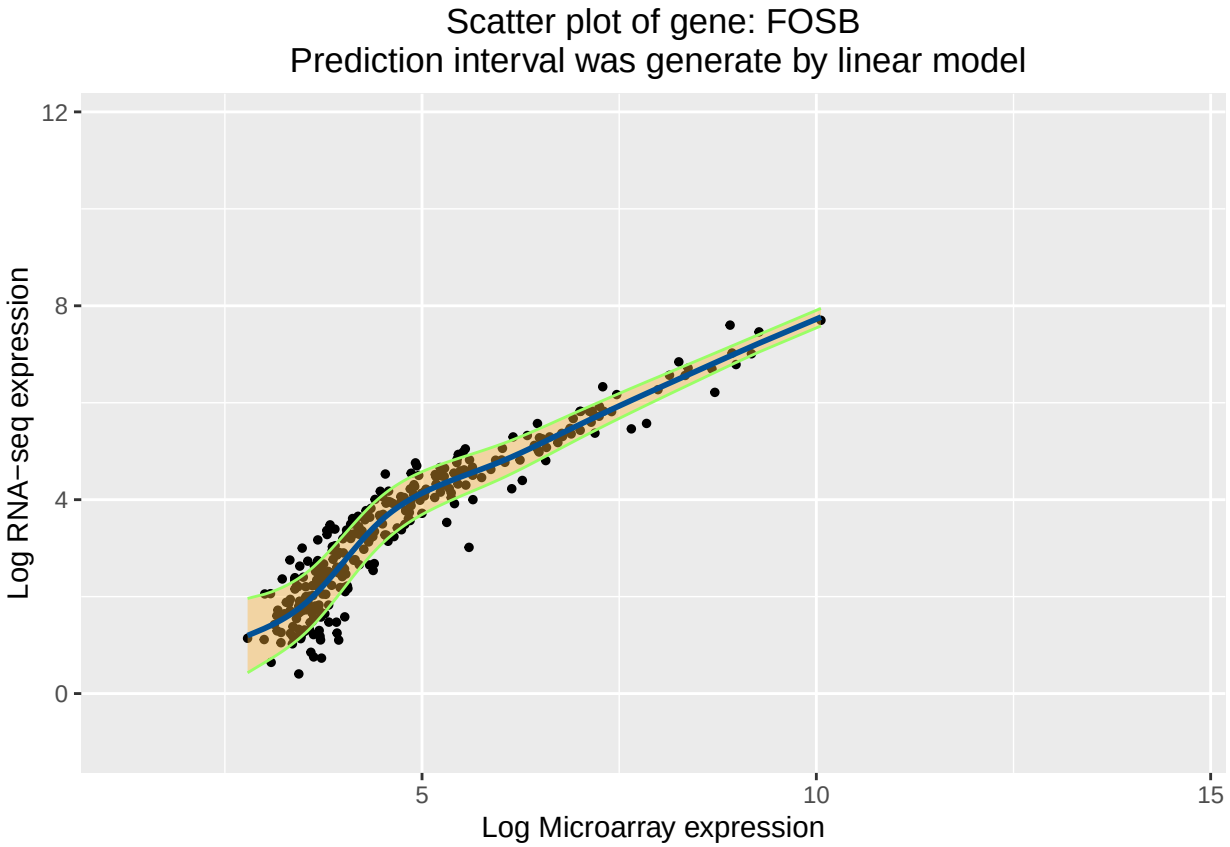

Scatter plot of gene: ALDOC
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

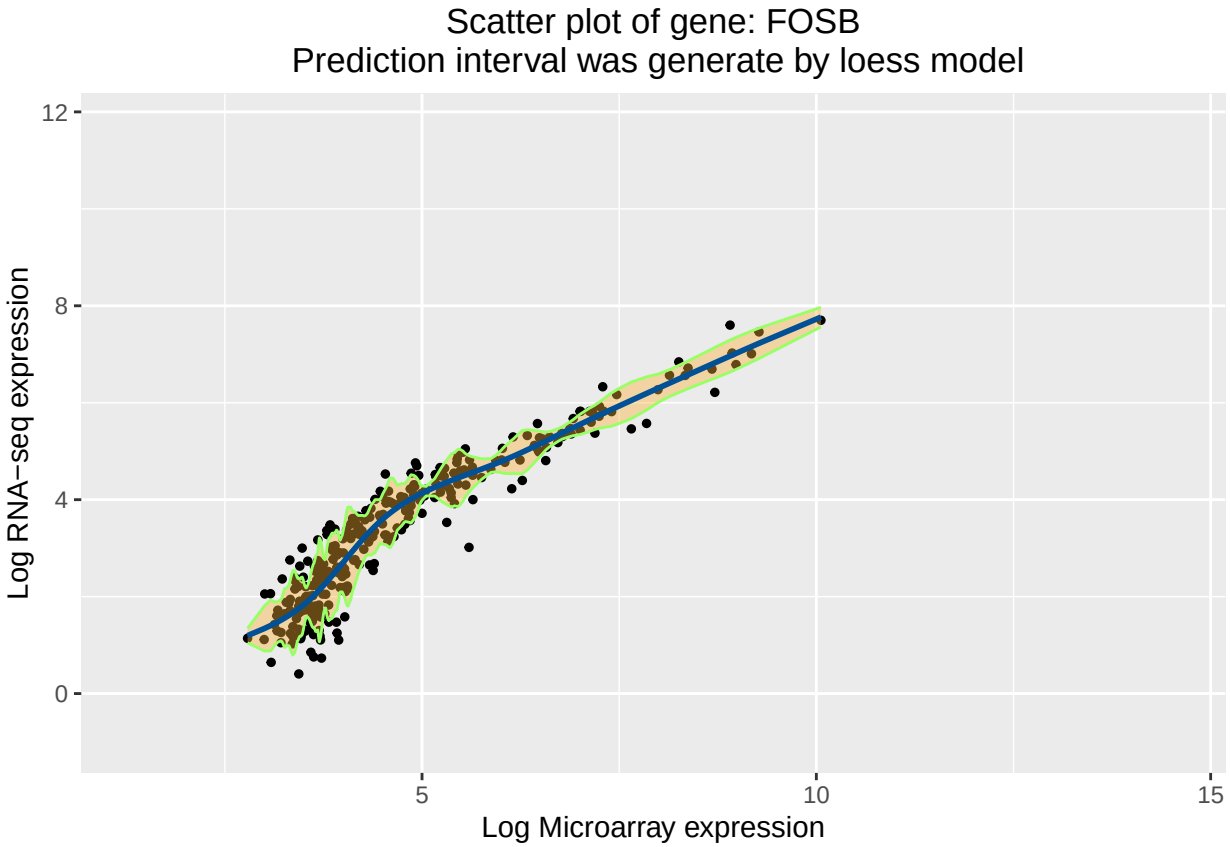

Residual scatter plot of Gene: NUTF2
Model type
lm R^2: 4.289e−02
loess R^2: 2.142e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
6
7
8
9
Log Microarray intensity

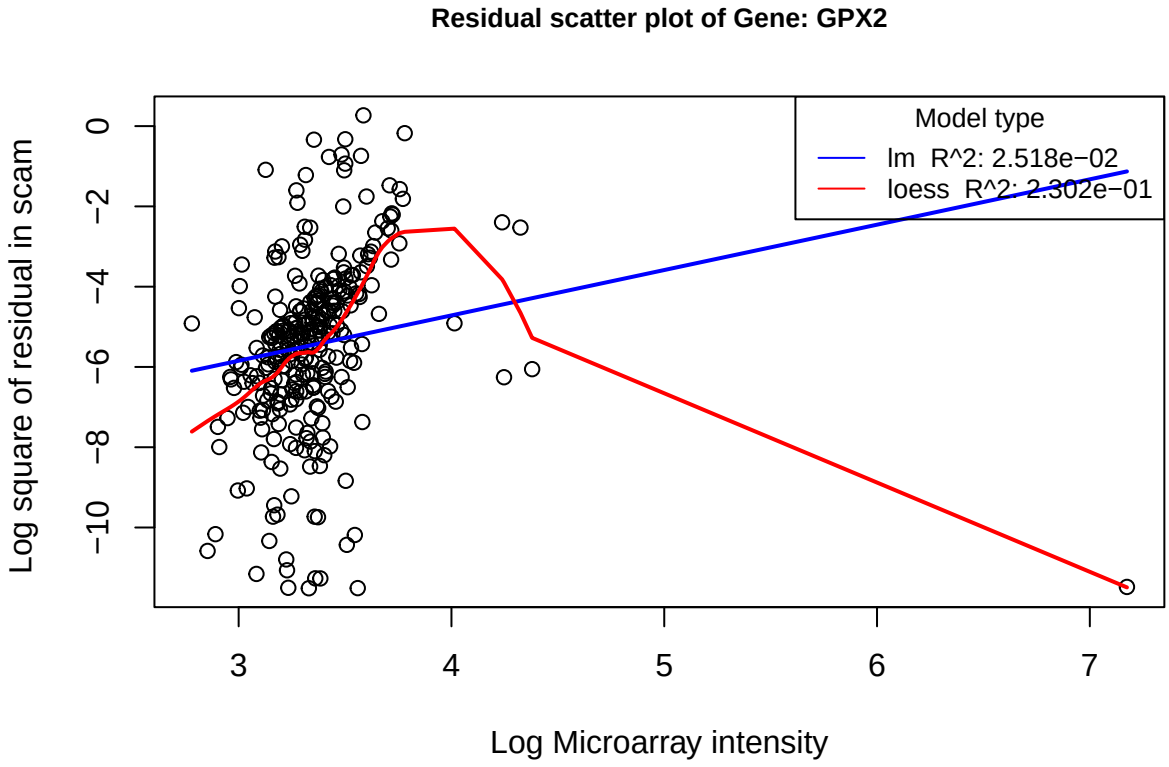

Scatter plot of gene: NUTF2
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

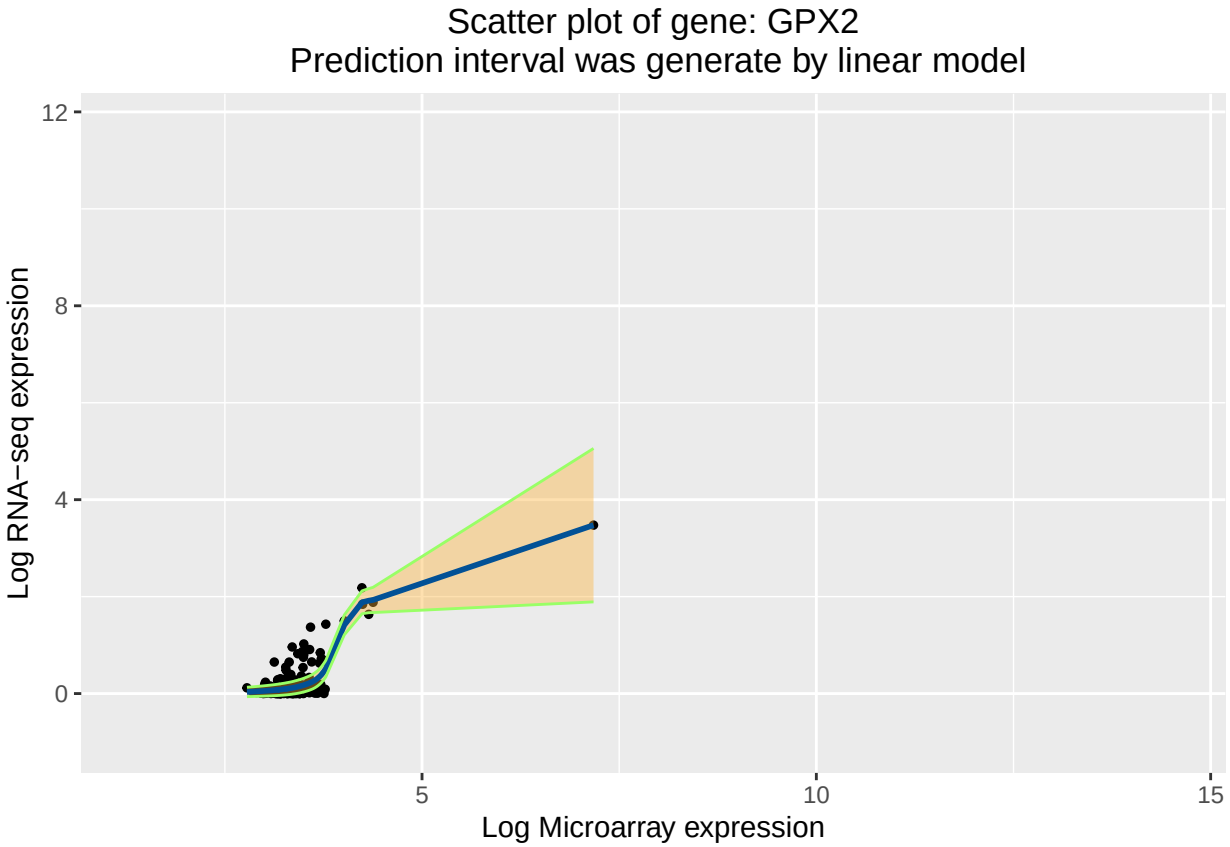

Scatter plot of gene: NUTF2
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

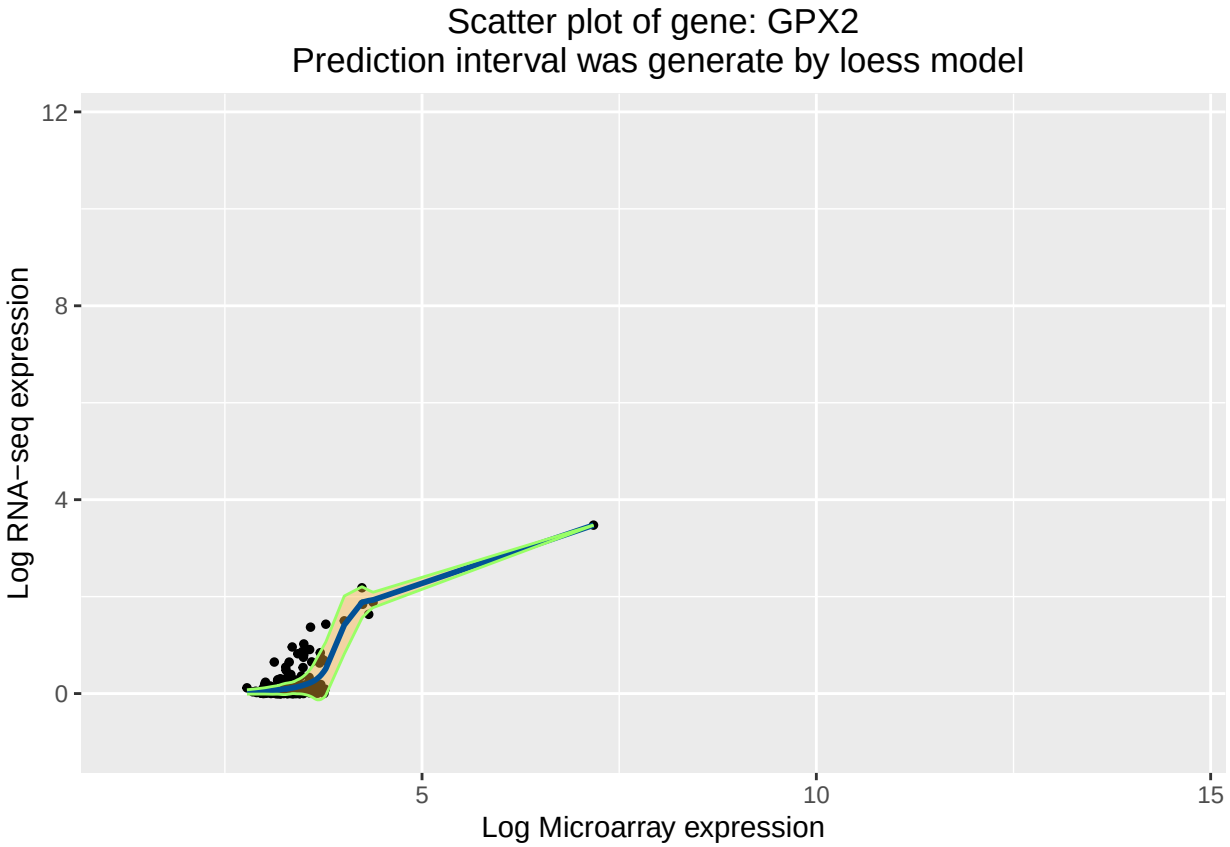

Residual scatter plot of Gene: FOSB
Model type
lm R^2: 8.932e−02
loess R^2: 2.527e−01
Log square of residual in scam
0
−2
−4
−6
−8
−12
4
6
8
10
Log Microarray intensity

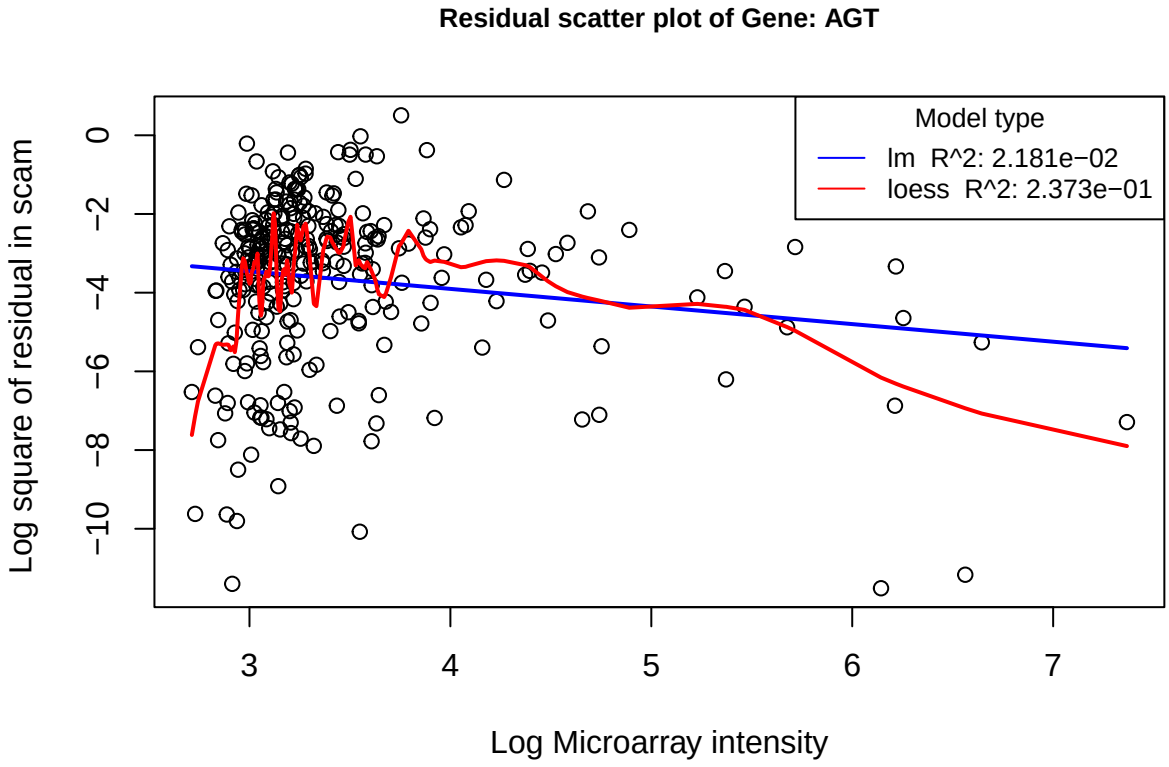

Scatter plot of gene: FOSB
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

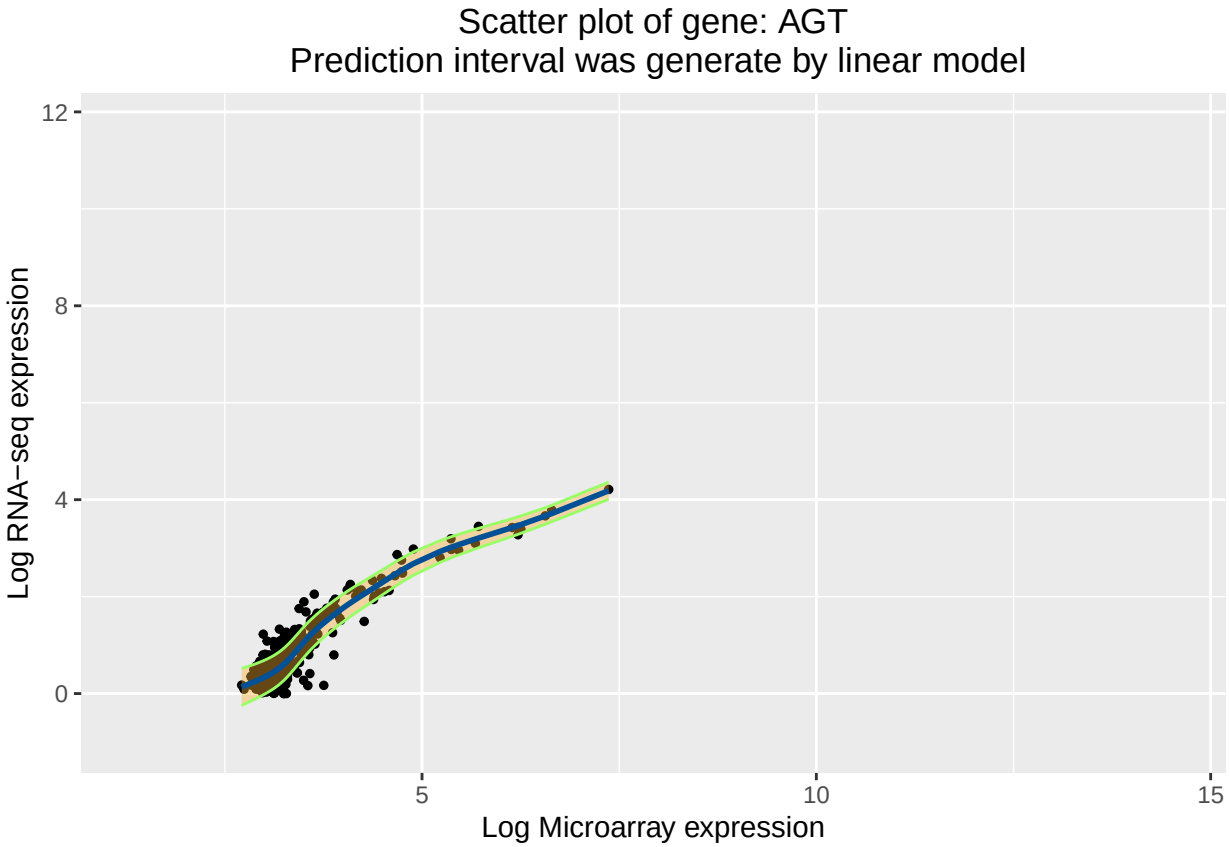

Scatter plot of gene: FOSB
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

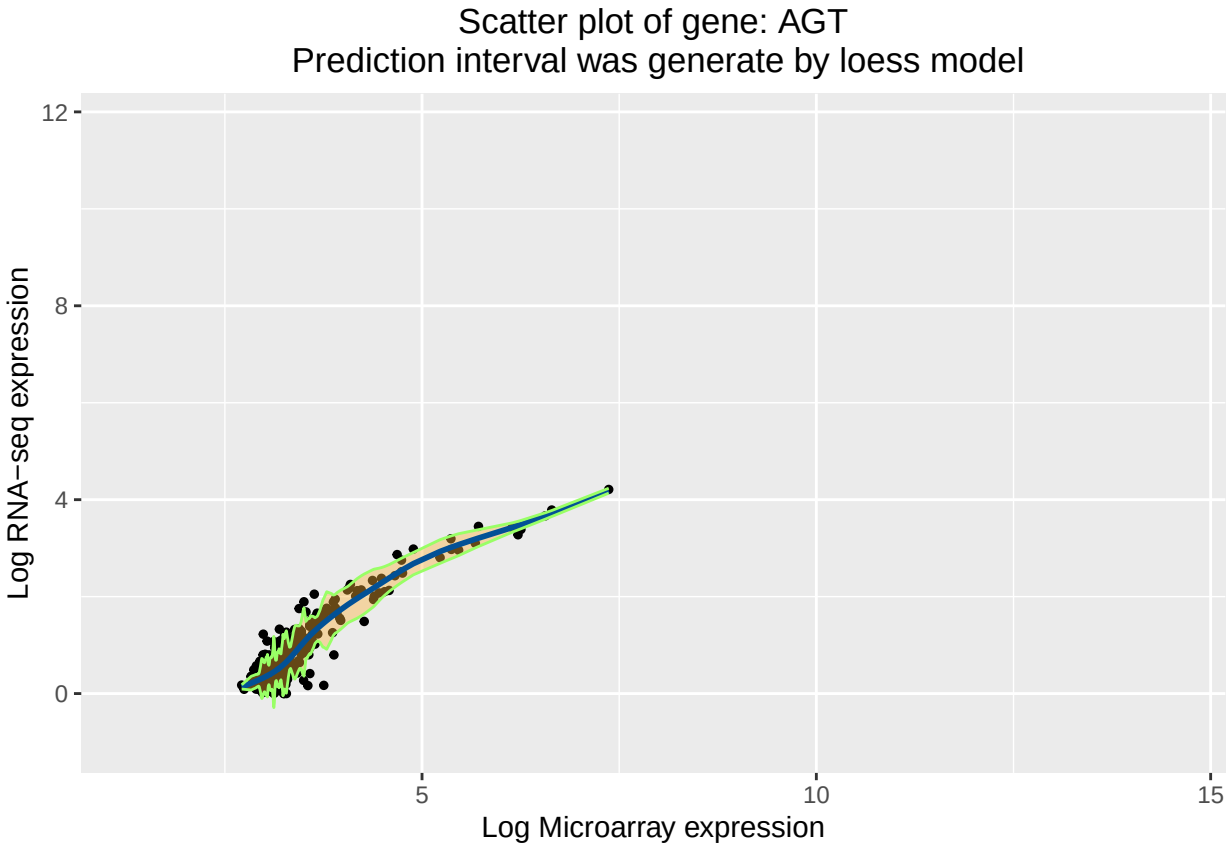

Residual scatter plot of Gene: GPX2
Model type
lm R^2: 2.518e−02
loess R^2: 2.302e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
3
4
5
6
7
Log Microarray intensity

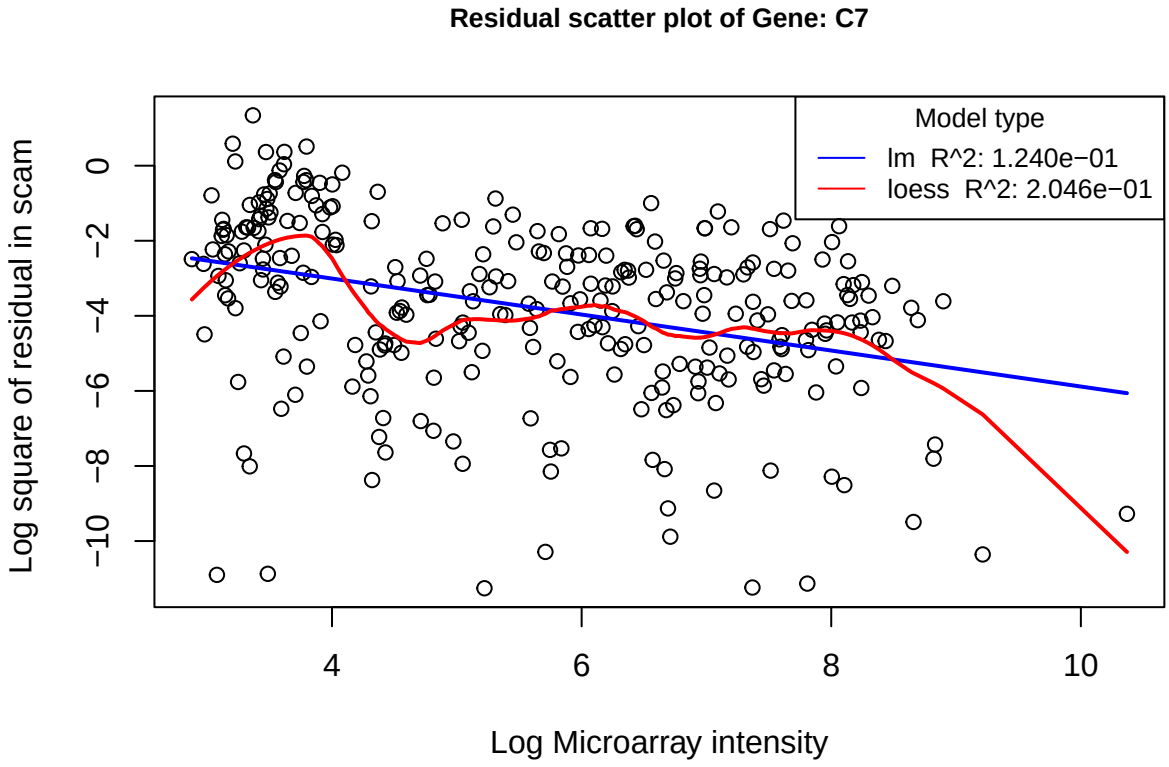

Scatter plot of gene: GPX2
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

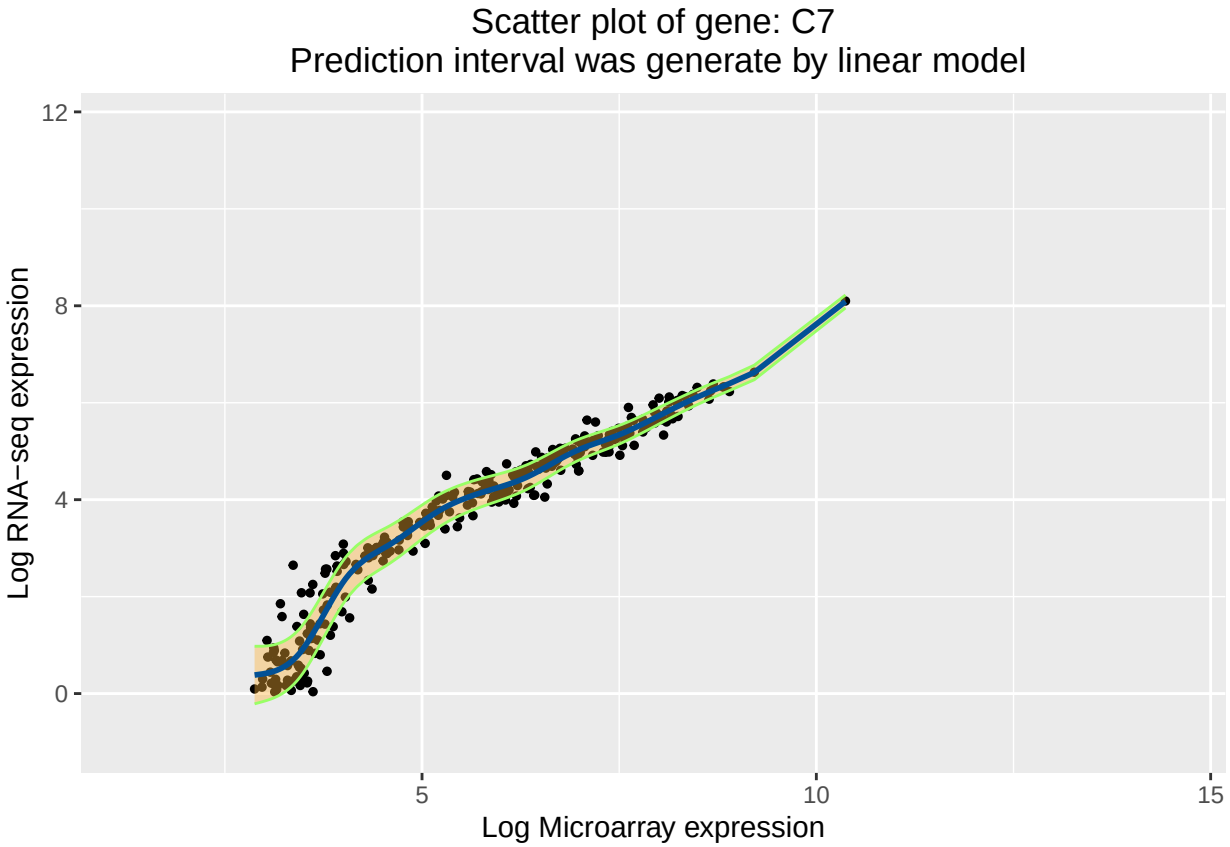

Scatter plot of gene: GPX2
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

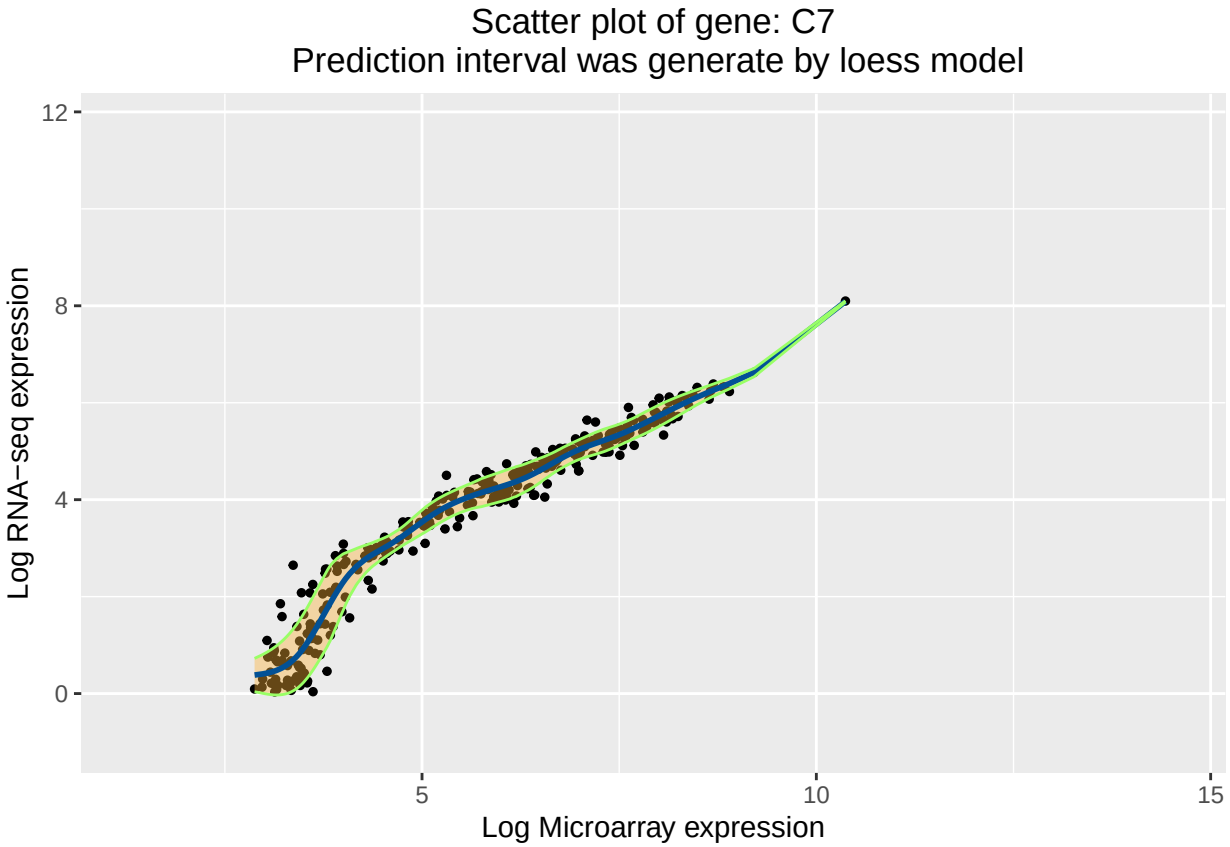

Residual scatter plot of Gene: AGT
Model type
lm R^2: 2.181e−02
loess R^2: 2.373e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
3
4
5
6
7
Log Microarray intensity

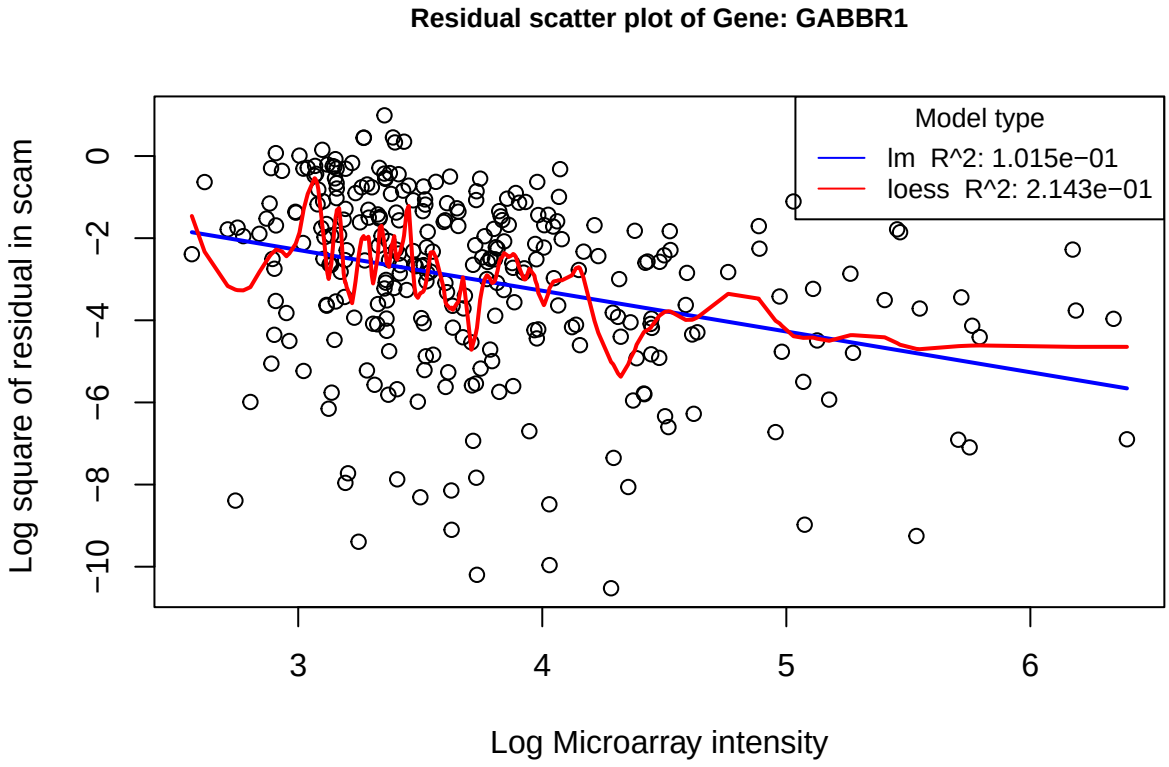

Scatter plot of gene: AGT
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

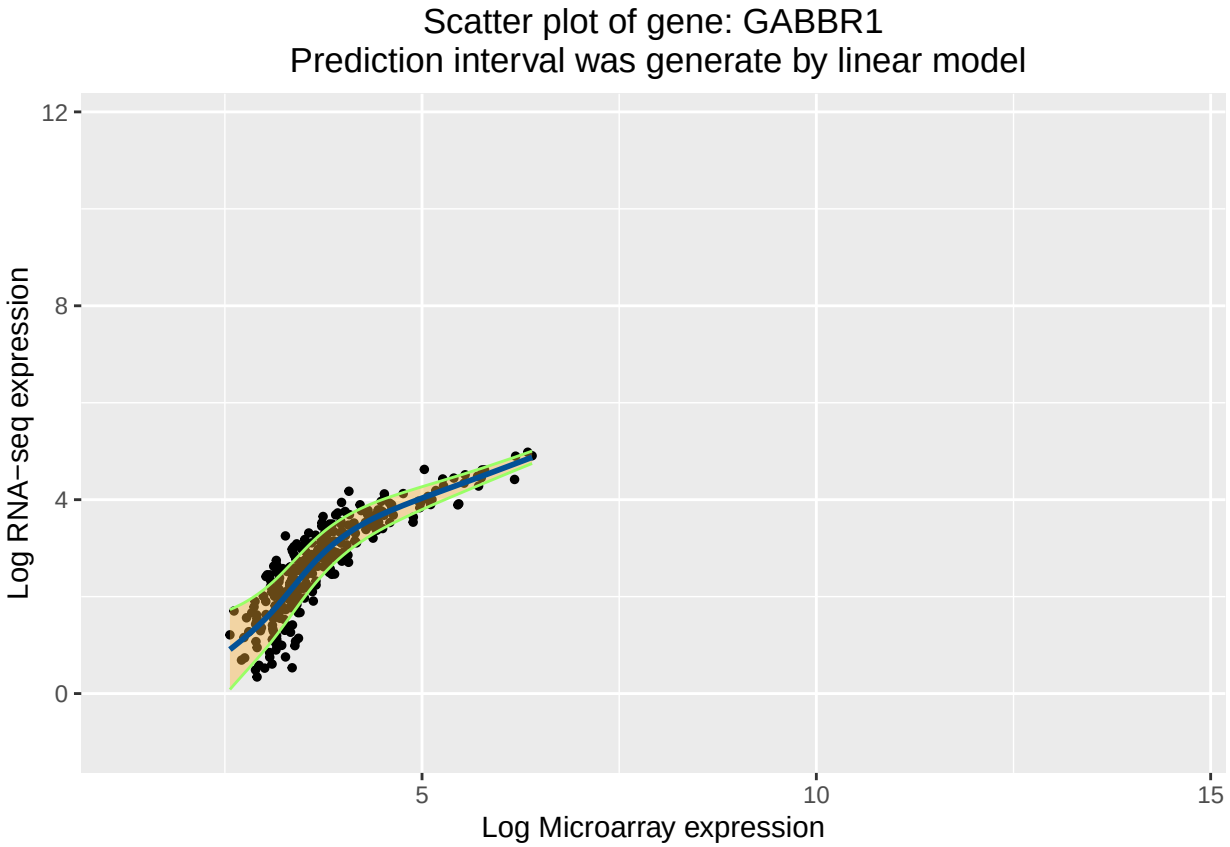

Scatter plot of gene: AGT
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

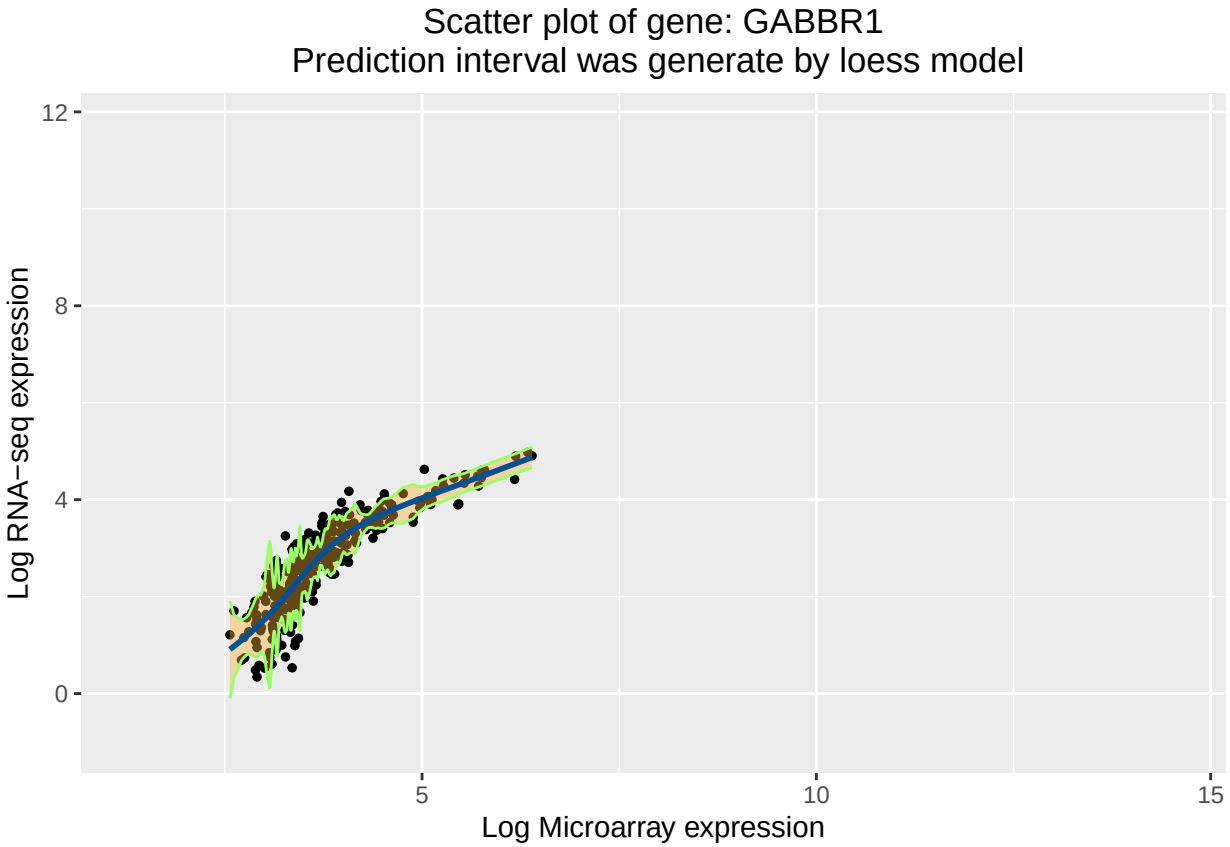

Residual scatter plot of Gene: C7
Model type
lm R^2: 1.240e−01
loess R^2: 2.046e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
4
6
8
10
Log Microarray intensity

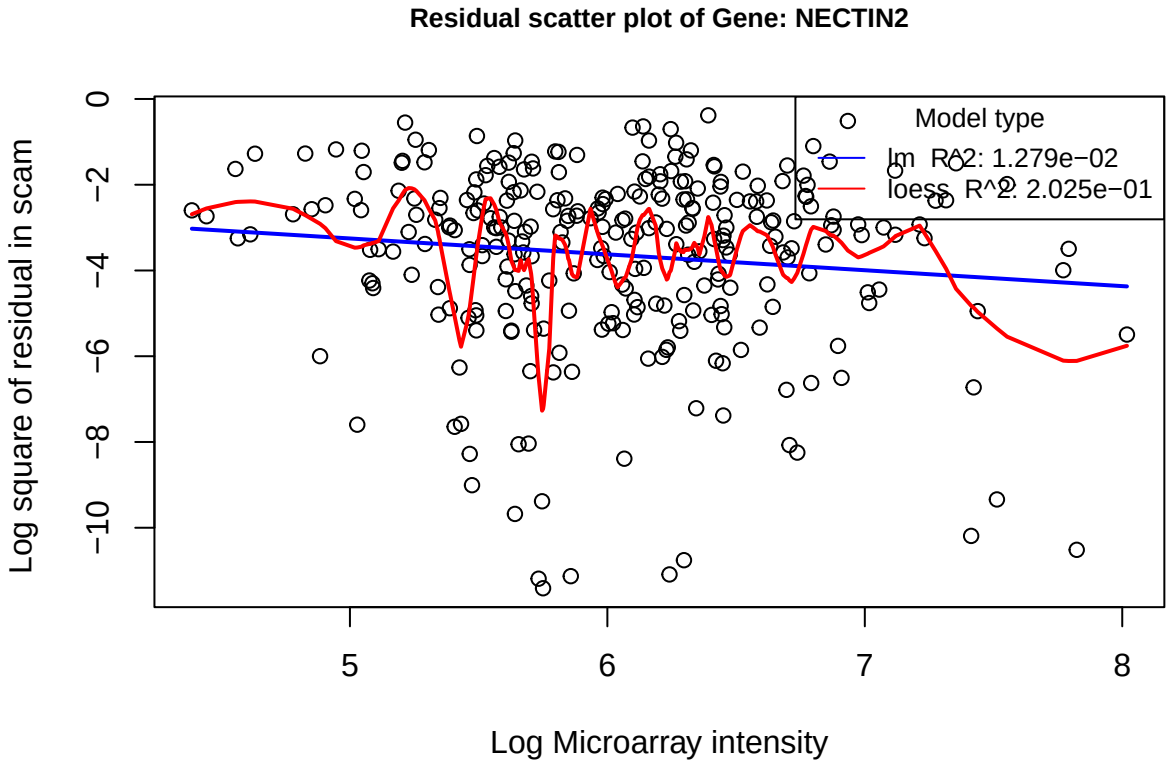

Scatter plot of gene: C7
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

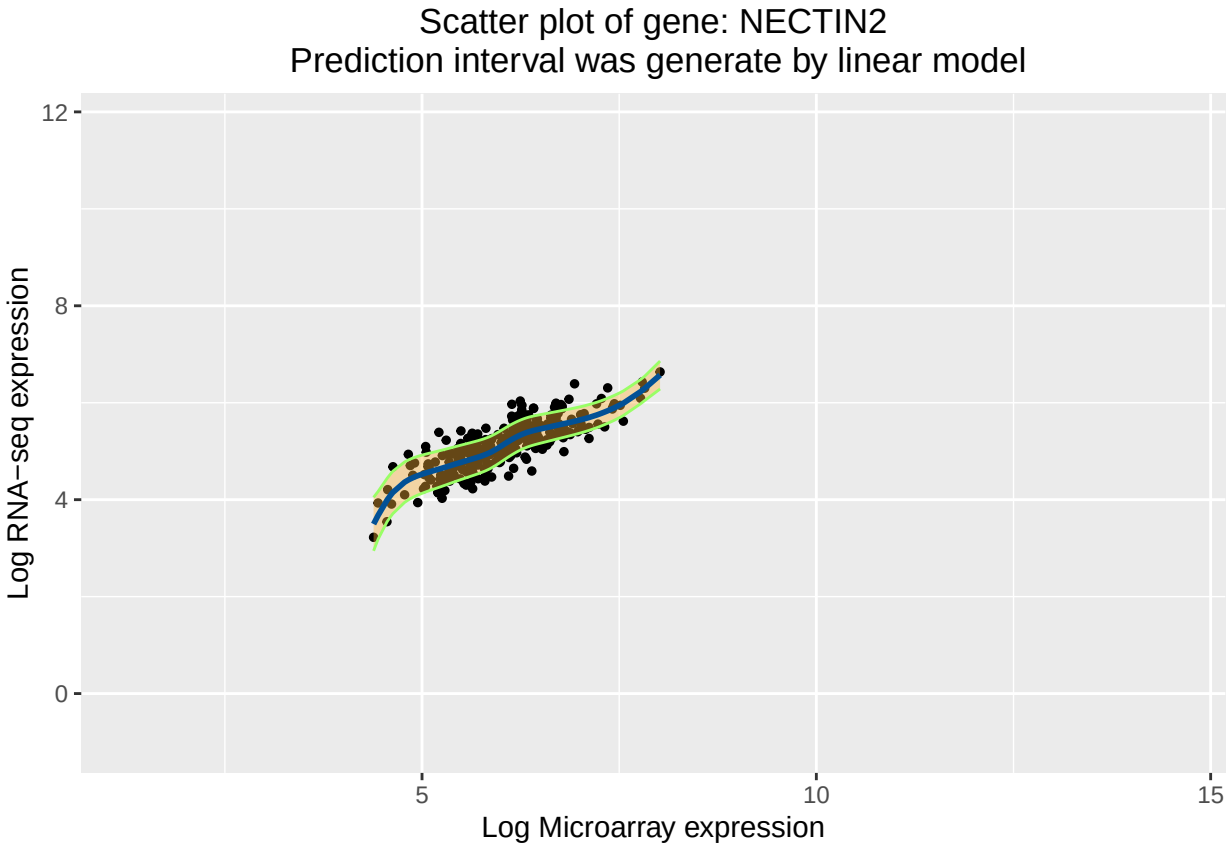

Scatter plot of gene: C7
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

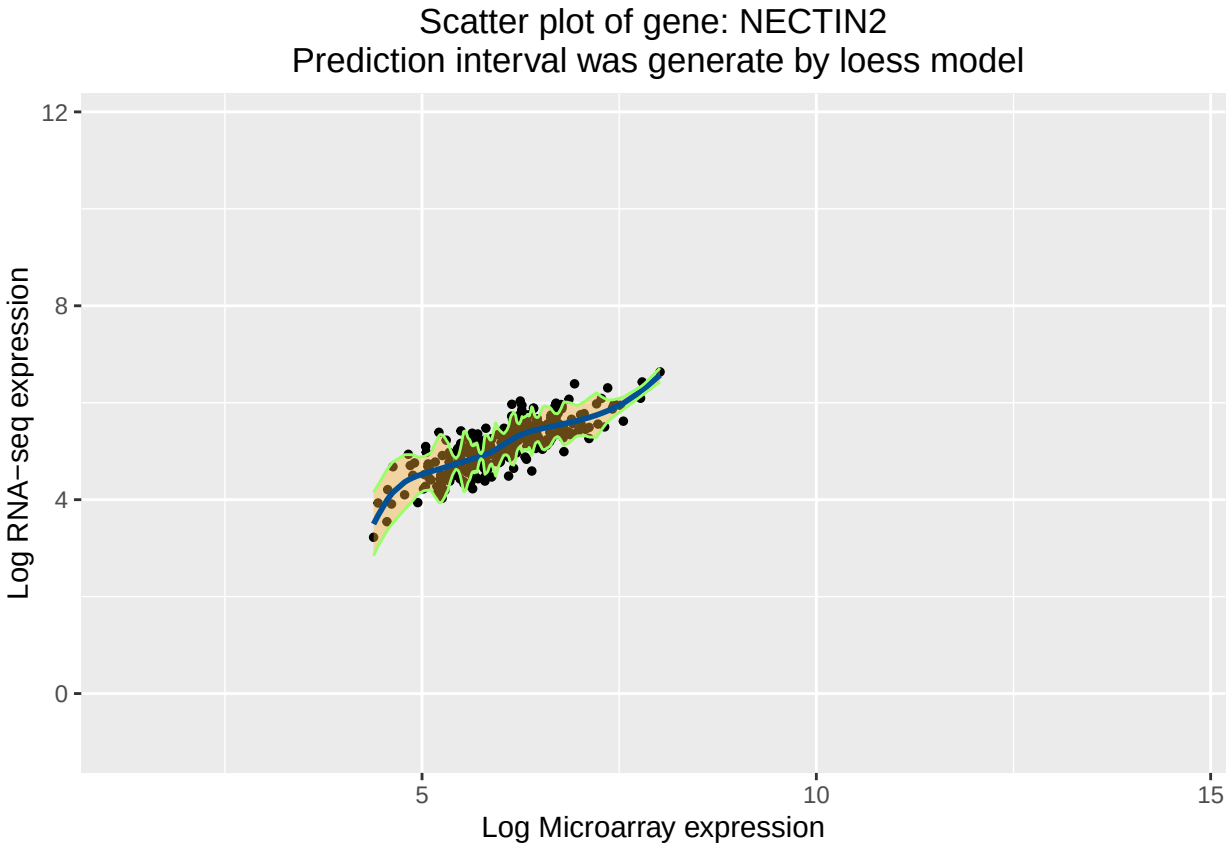

Residual scatter plot of Gene: GABBR1
Model type
lm R^2: 1.015e−01
loess R^2: 2.143e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
3
4
5
6
Log Microarray intensity

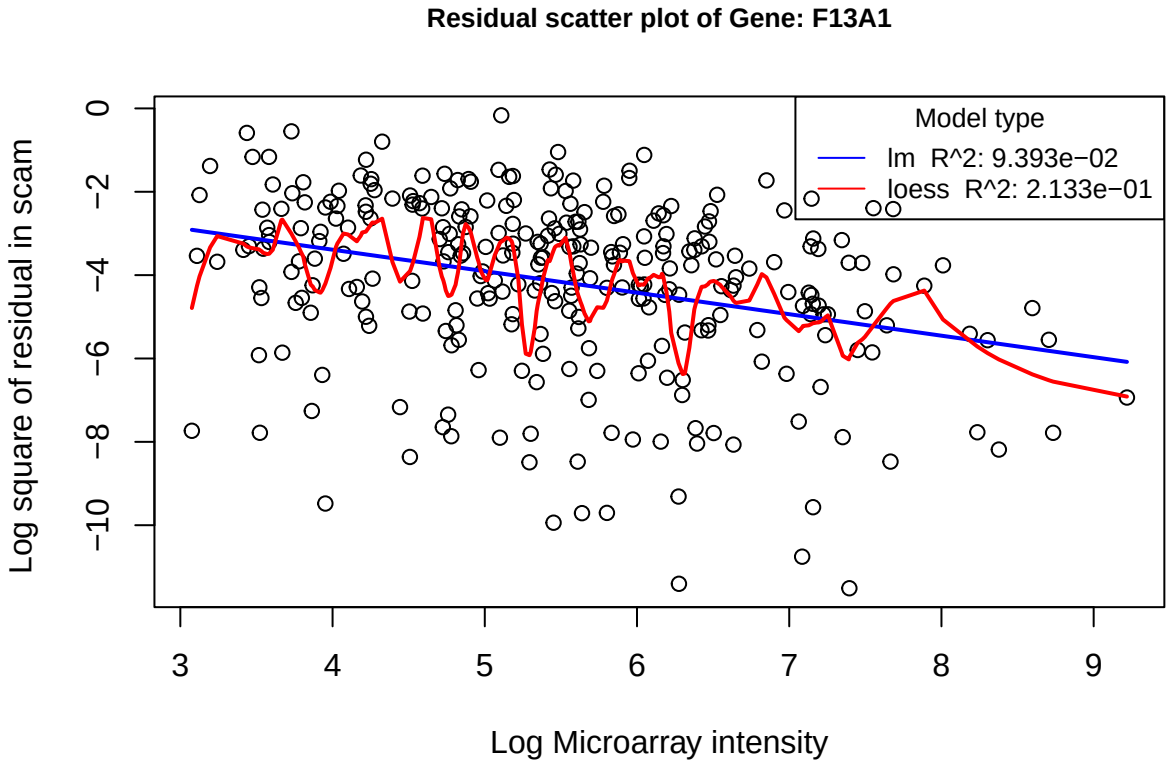

Scatter plot of gene: GABBR1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

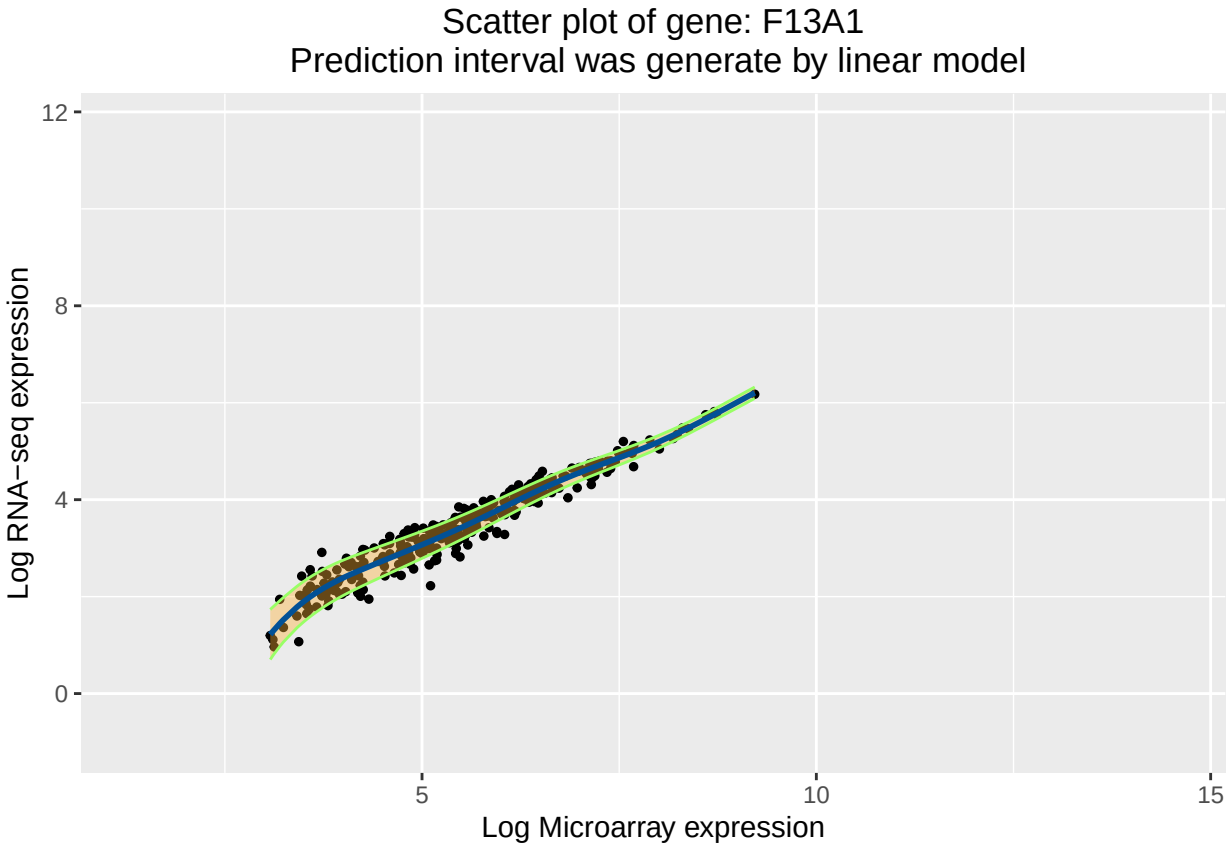

Scatter plot of gene: GABBR1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

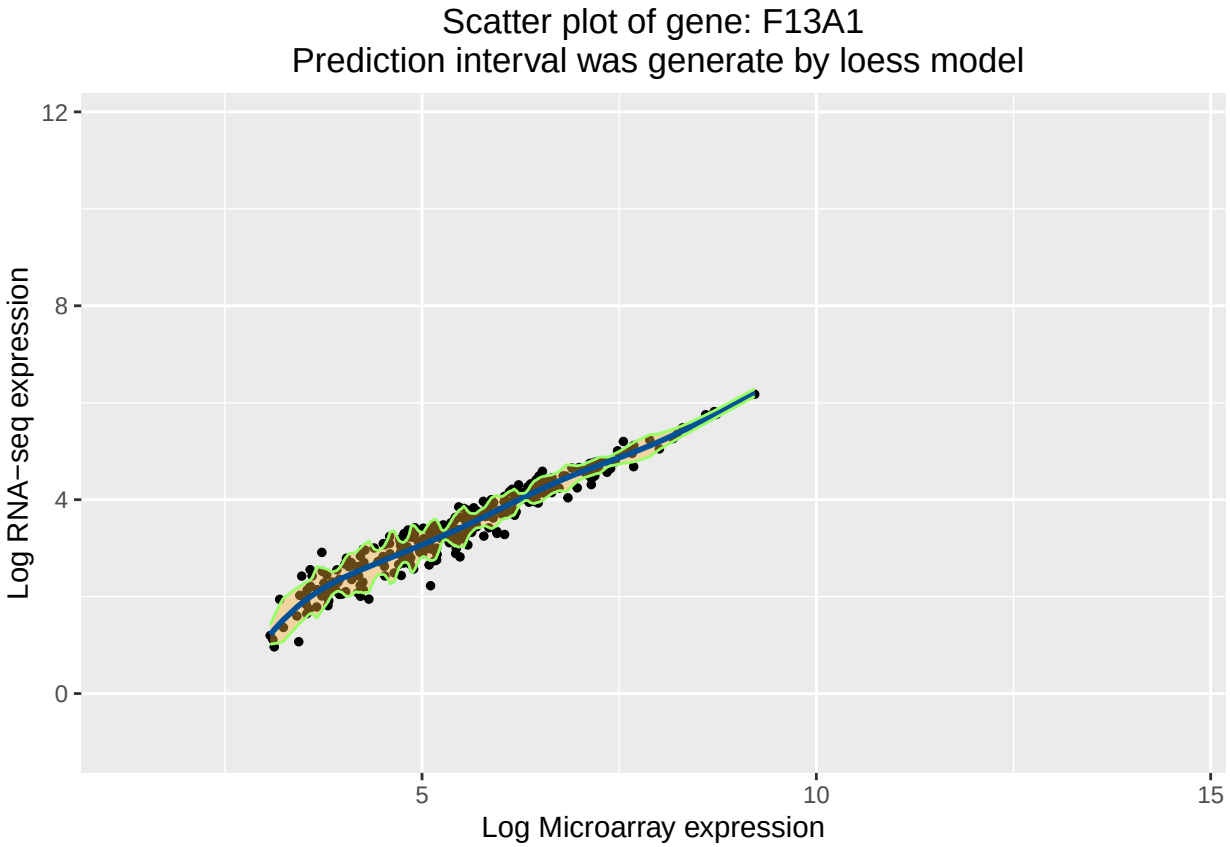

Residual scatter plot of Gene: NECTIN2
Model type
lm R^2: 1.279e−02
loess R^2: 2.025e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
5
6
7
8
Log Microarray intensity

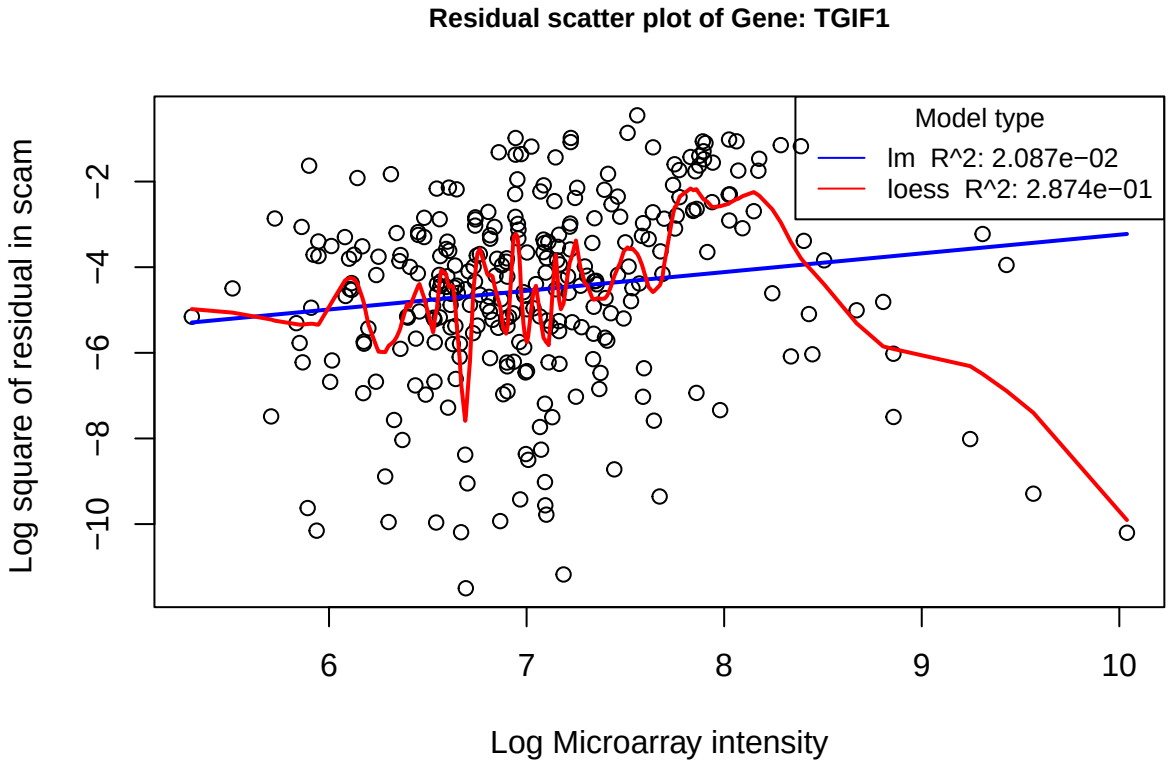

Scatter plot of gene: NECTIN2
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

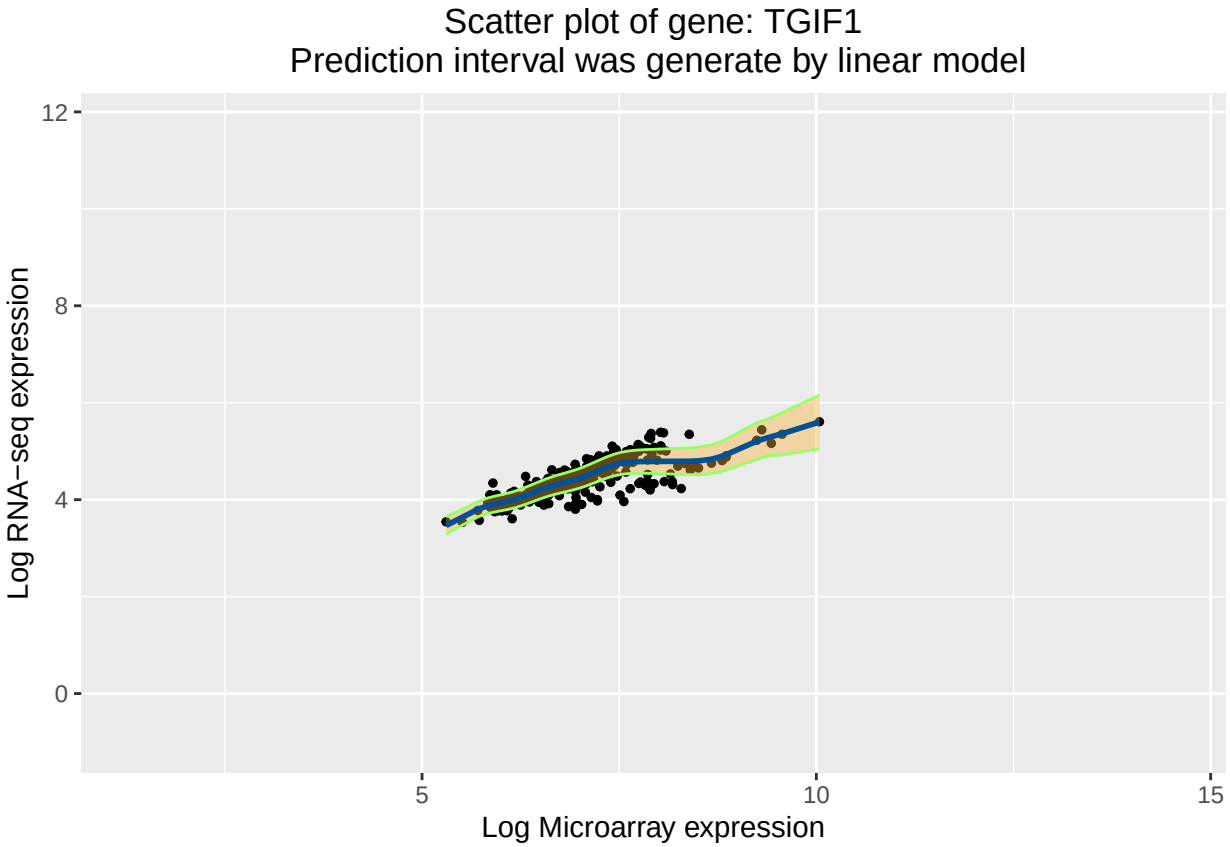

Scatter plot of gene: NECTIN2
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

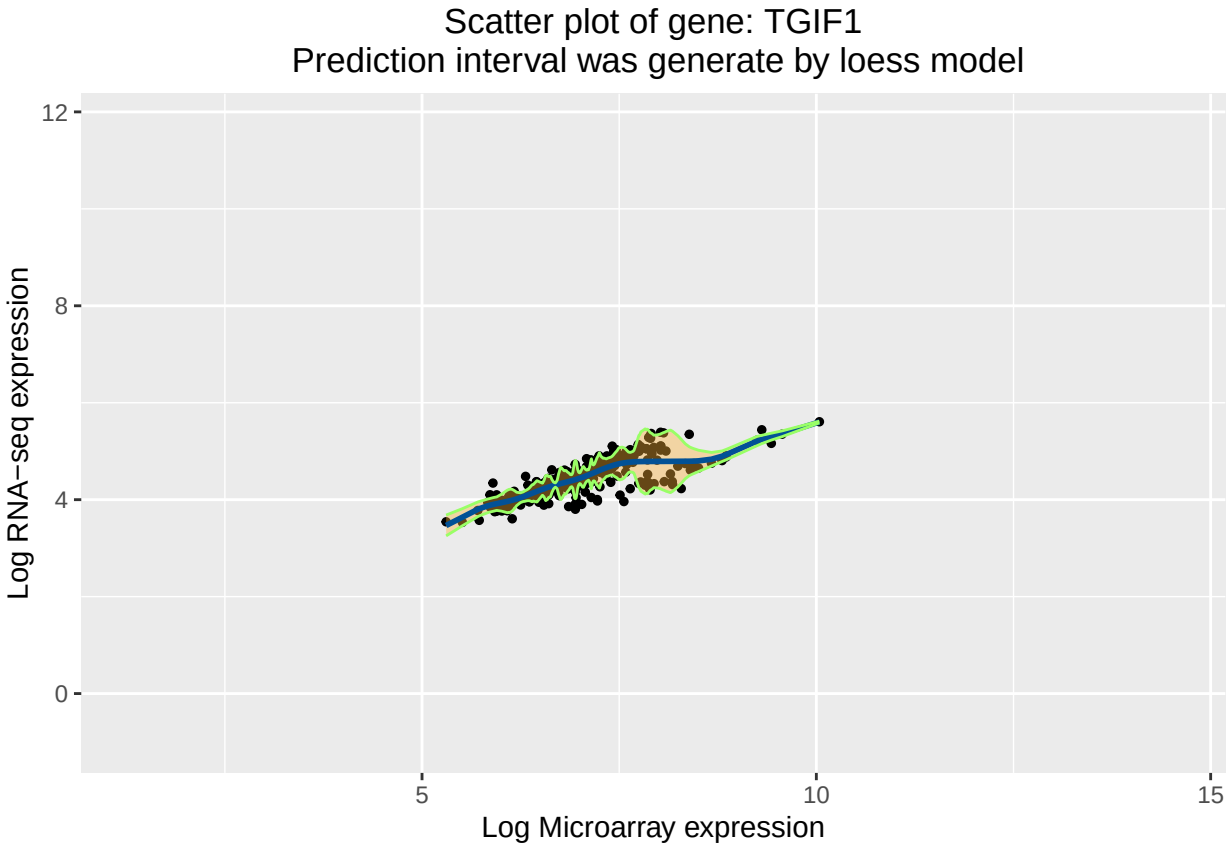

Residual scatter plot of Gene: F13A1
Model type
lm R^2: 9.393e−02
loess R^2: 2.133e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
3
4
5
6
7
8
9
Log Microarray intensity

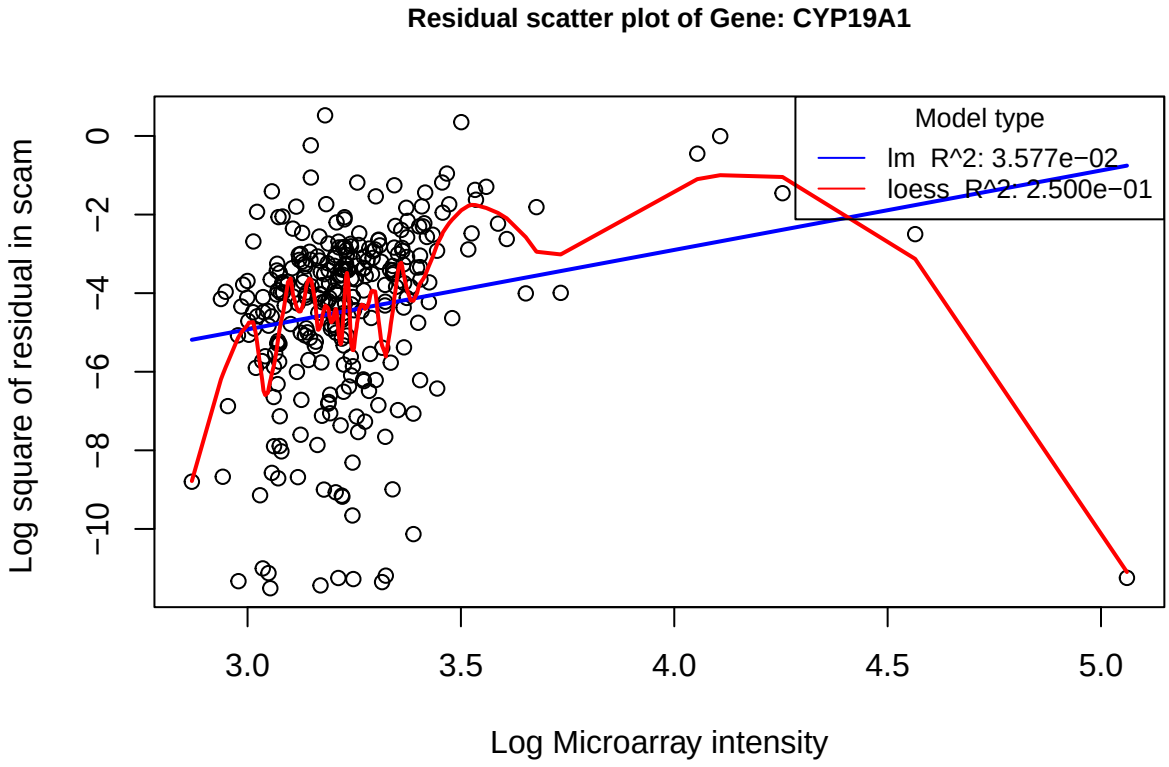

Scatter plot of gene: F13A1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

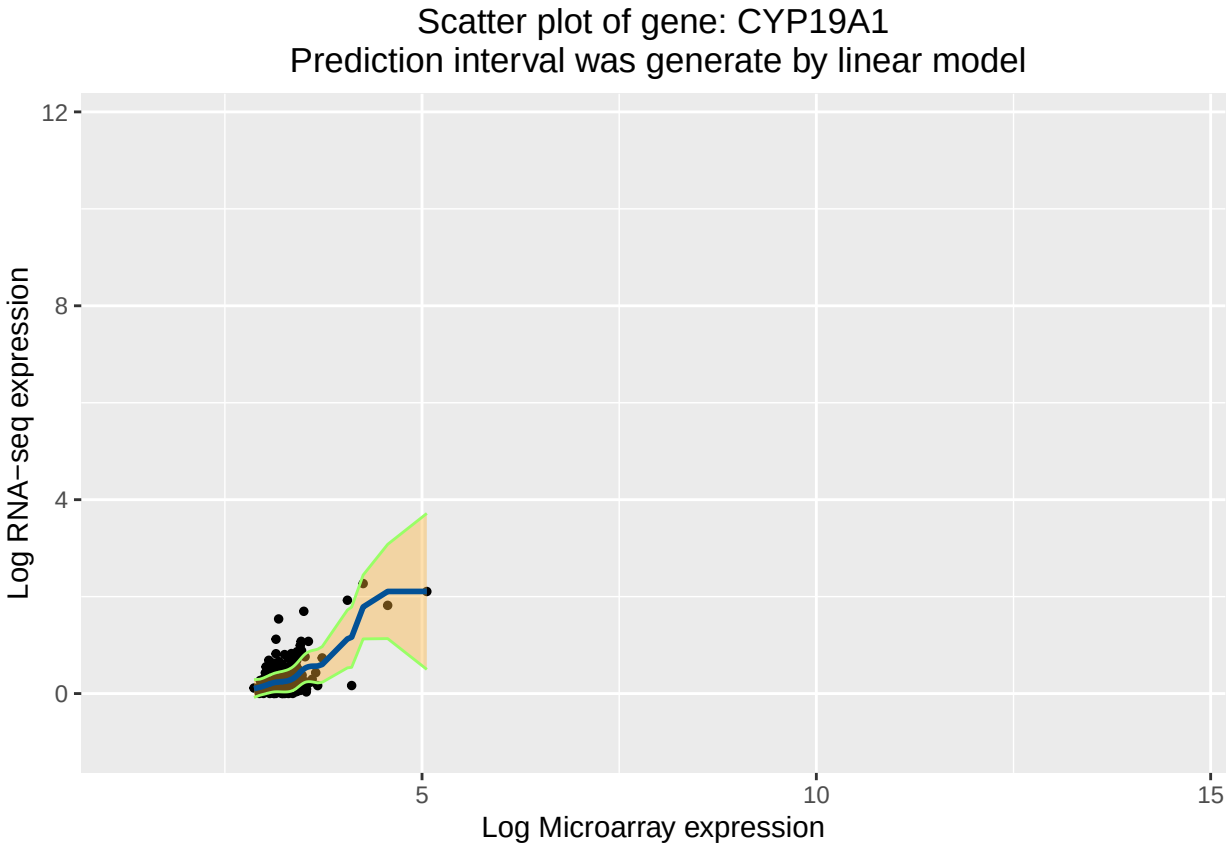

Scatter plot of gene: F13A1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

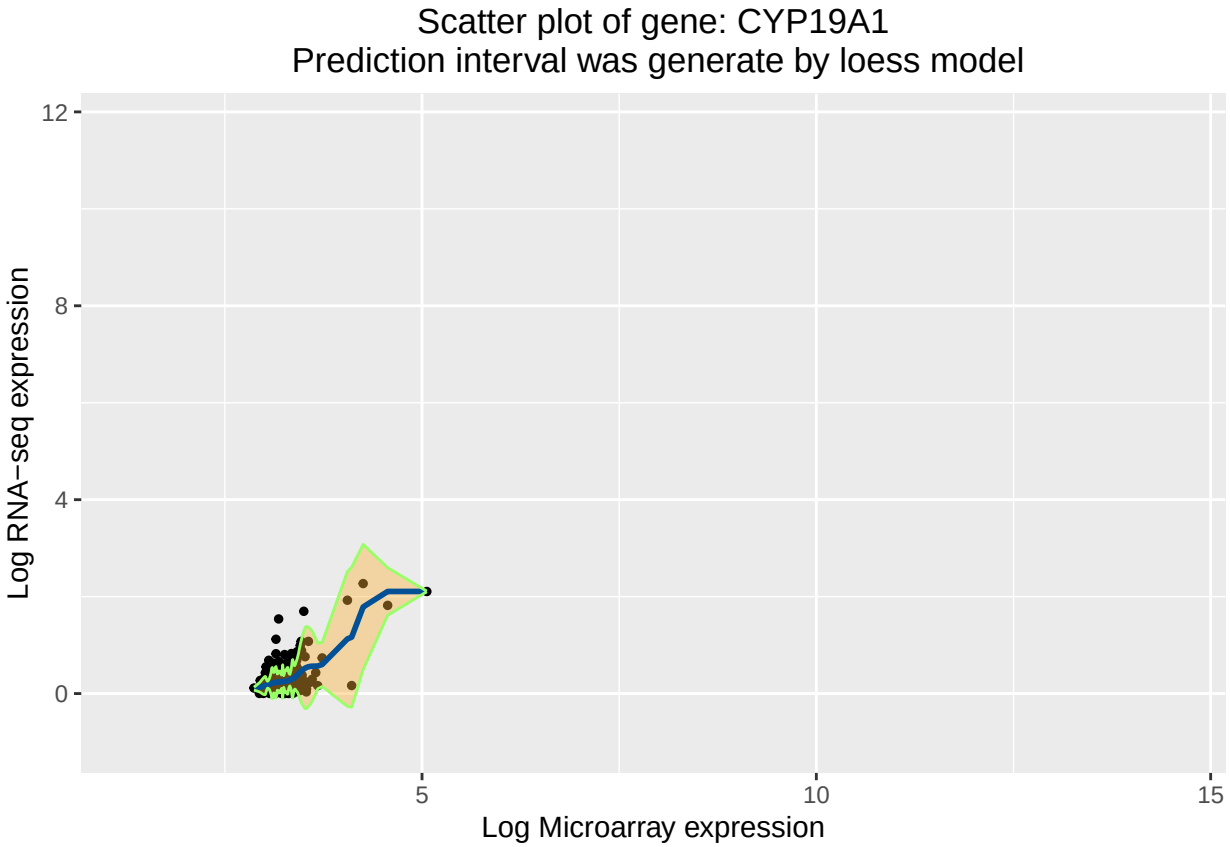

Residual scatter plot of Gene: TGIF1
Model type
lm R^2: 2.087e−02
loess R^2: 2.874e−01
Log square of residual in scam
−2
−4
−6
−8
−10
6
7
8
9
10
Log Microarray intensity

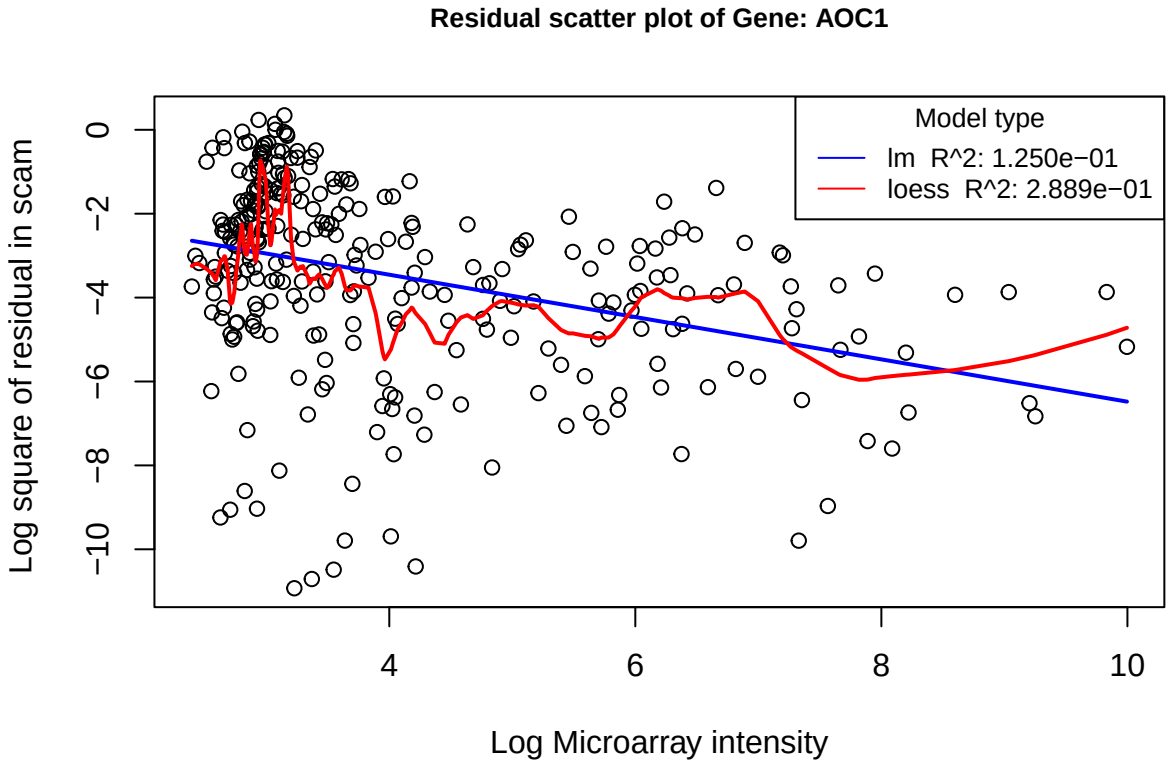

Scatter plot of gene: TGIF1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

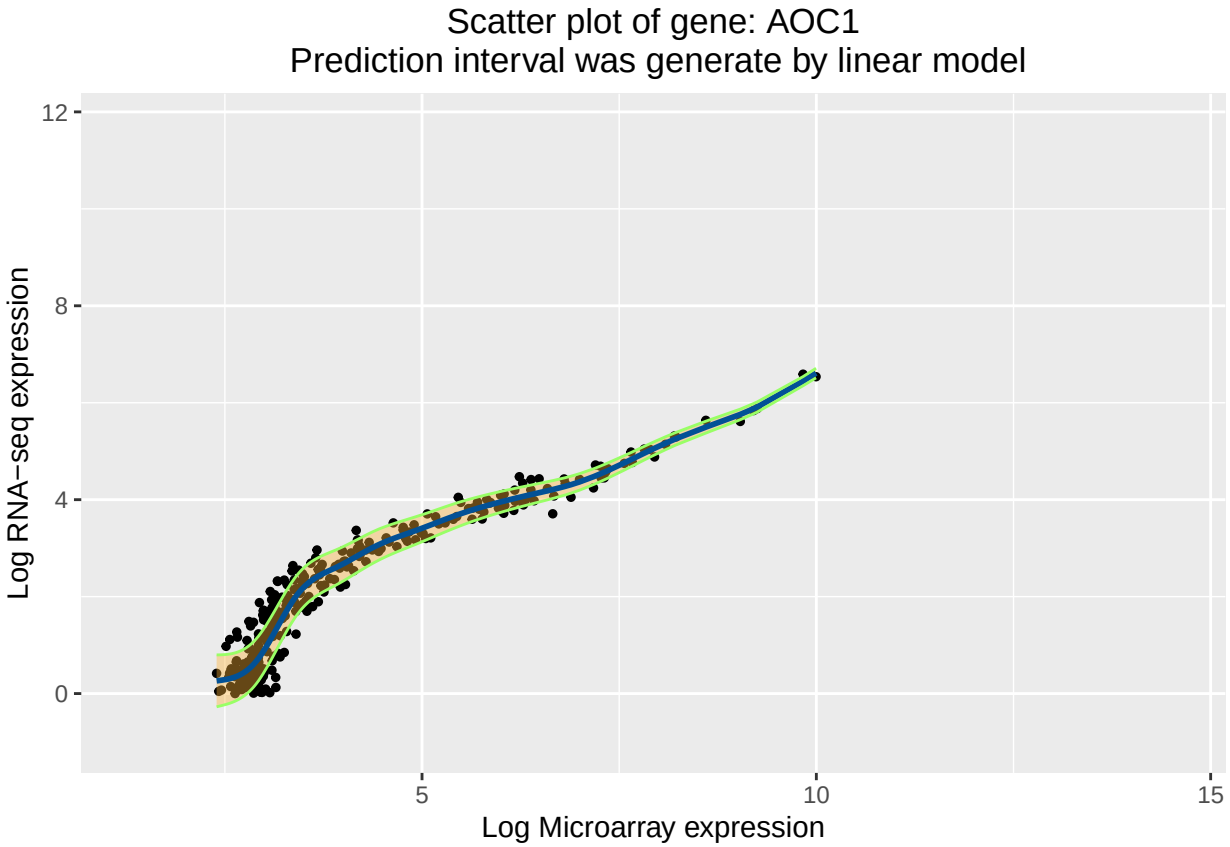

Scatter plot of gene: TGIF1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

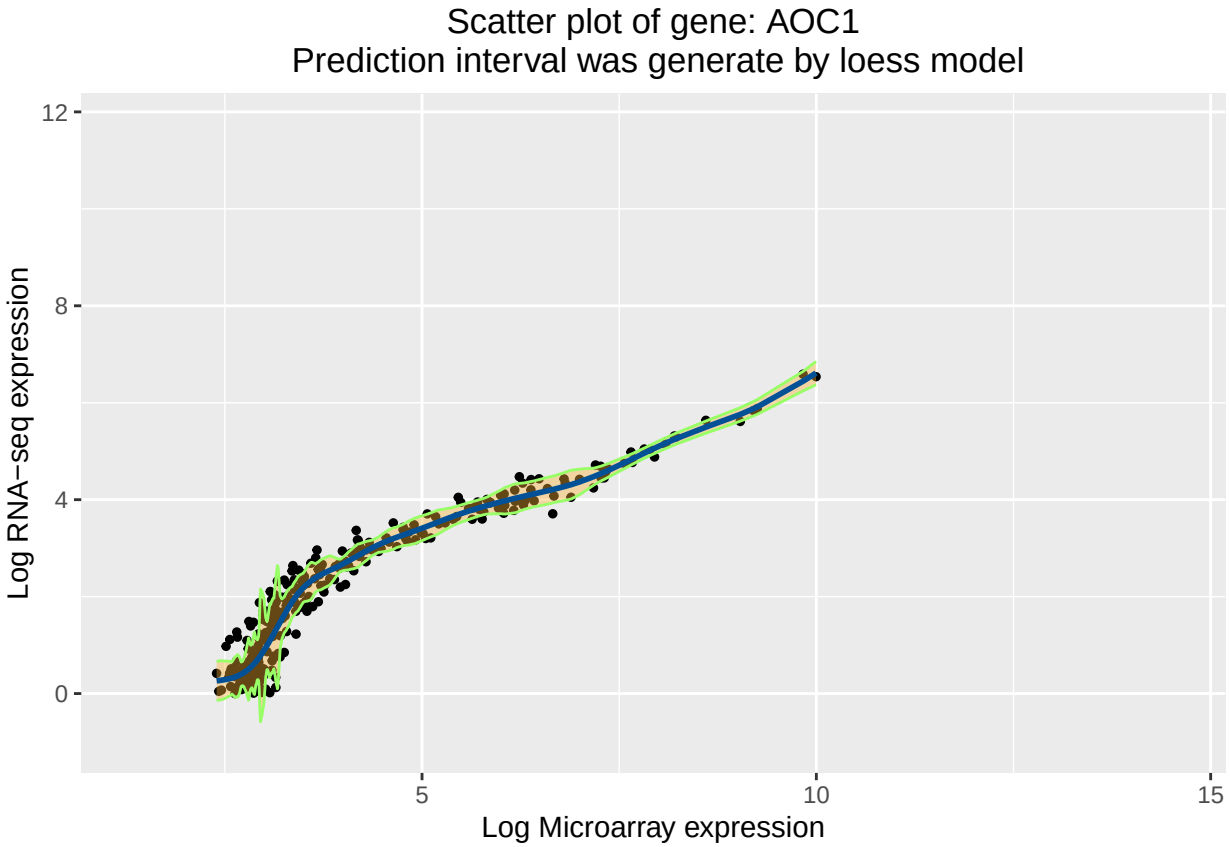

Residual scatter plot of Gene: CYP19A1
Model type
lm R^2: 3.577e−02
loess R^2: 2.500e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
3.0
3.5
4.0
4.5
5.0
Log Microarray intensity

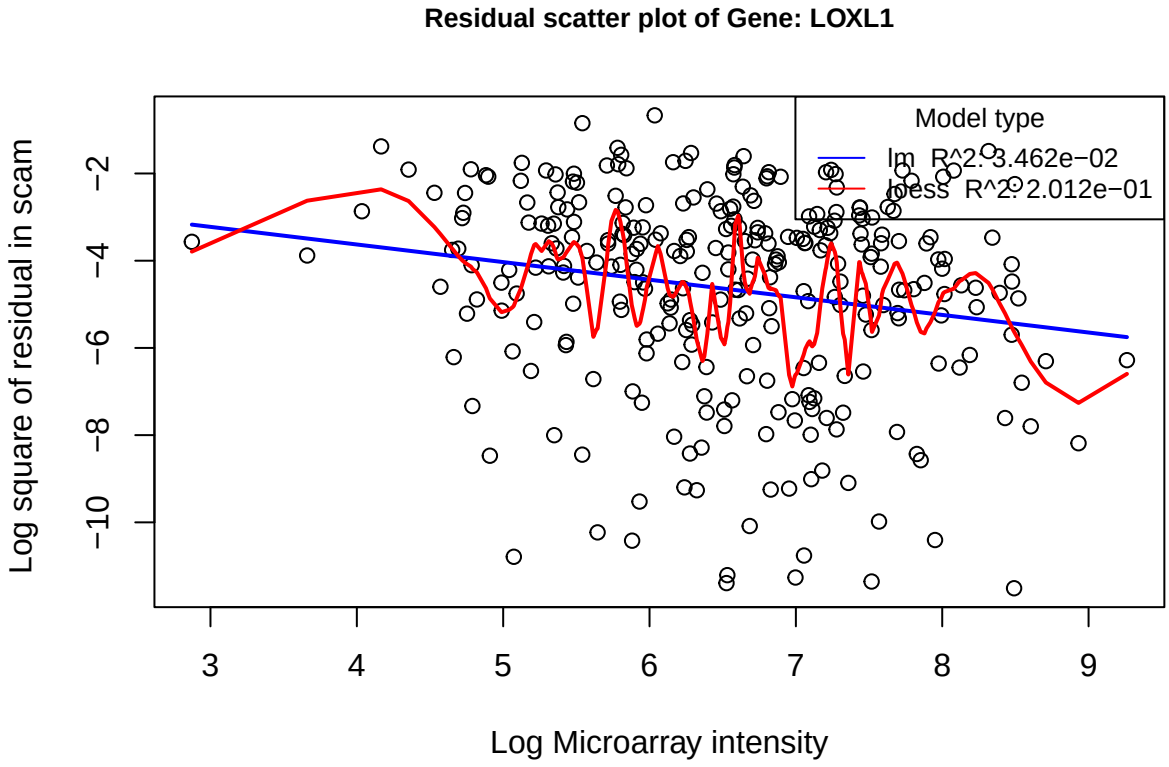

Scatter plot of gene: CYP19A1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

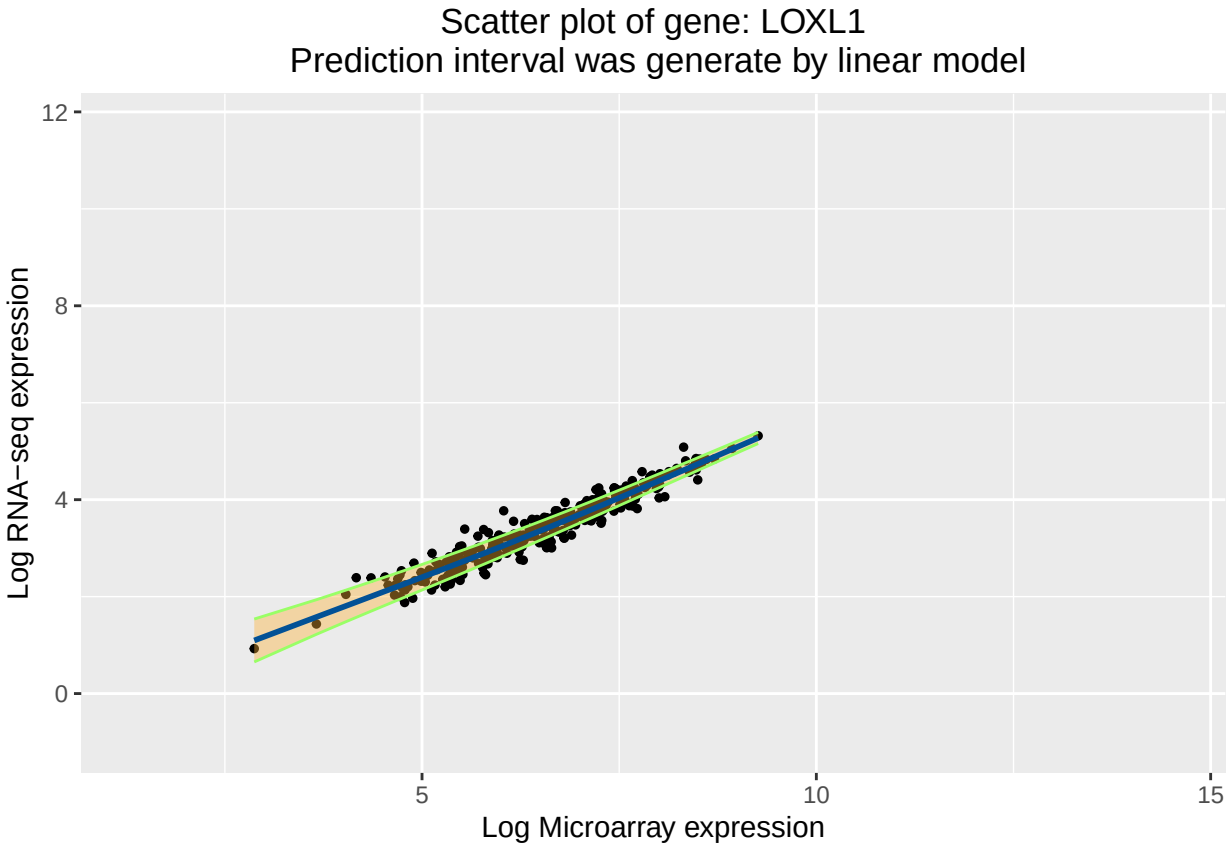

Scatter plot of gene: CYP19A1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression

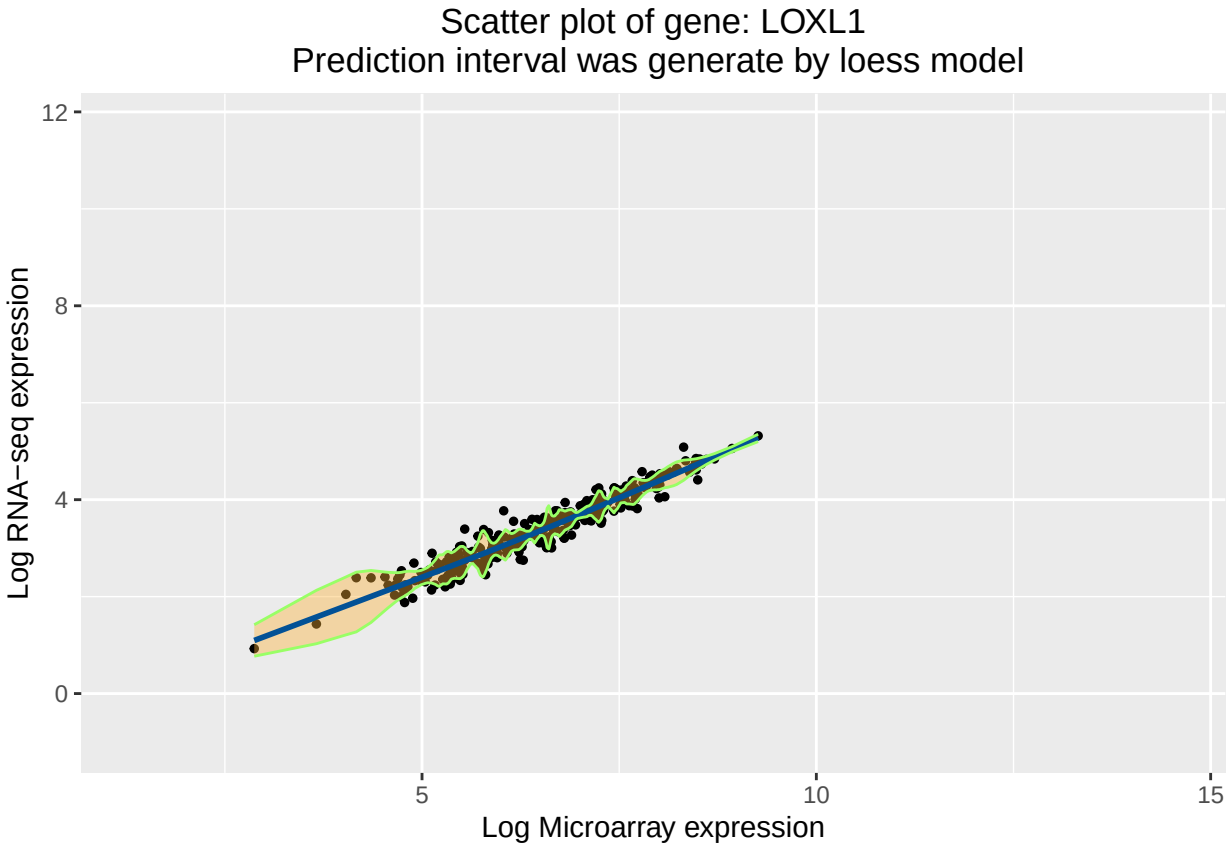

Residual scatter plot of Gene: AOC1
Model type
lm R^2: 1.250e−01
loess R^2: 2.889e−01
Log square of residual in scam
0
−2
−4
−6
−8
−10
4
6
8
10
Log Microarray intensity


Scatter plot of gene: AOC1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression


Scatter plot of gene: AOC1
Prediction interval was generate by loess model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression


Residual scatter plot of Gene: LOXL1
Model type
lm R^2: 3.462e−02
loess R^2: 2.012e−01
Log square of residual in scam
−2
−4
−6
−8
−10
3
4
5
6
7
8
9
Log Microarray intensity


Scatter plot of gene: LOXL1
Prediction interval was generate by linear model
Log RNA−seq expression
12
8
4
0
5
10
15
Log Microarray expression




```
Residual plot with lm;
Residual plot with loess;
Raw data scatter plot with prediction interval using lm;
Raw data scatter plot with prediction interval using loess;
```

Please look at "residual_plots_vs_plots_PI_loess.pdf", which have 21 genes, each genes have four plots as mentioned above