



Title: Project X: Data Analysis
Project: Project
CICPT: 007
Date: 21 May, 2020
Investigators: ,
Organization: Analysis Group

Contents

1 Overview	2
2 Load Data	2
3 Run correlations	2
4 Analysis of correlations	2
5 Gene Correlations	5
5.1 TCC	5
5.2 CCLE	11
5.3 EXPO	17
6 ABL-IRF1	23
7 TCC	23
8 CCLE	23
9 expo	24

1 Overview

The goal of this analysis is to examine the correlation of individual probesets to the RSI score, across several large datasets. We will use three different datasets: TCC (HuRSTA array), EXPO (U133 arrays) and CCLE (U133 arrays).

```
suppressPackageStartupMessages({  
  library(readxl)  
  library(knitr)  
  library(kableExtra)  
  library(ggplot2)  
  library(dplyr)  
  library(BioBase)  
})
```

2 Load Data

Load the three different datasets.

```
ccle<-readRDS(file="../data.raw/ccle_cel.rds")  
expo<-readRDS(file="../data.raw/expo_rma_exprs.rds")  
tcc<-readRDS(file="../data.raw/tcc.primary.rds")  
  
translation<-read_excel("../data.raw/Translation.xlsx")  
  
## New names:  
## * `` -> ...11
```

3 Run correlations

```
ccle_correlations<-apply(exprs(ccle), 1, function(y){cor(y, ccle$RSI)})  
expo_correlations<-apply(exprs(expo), 1, function(y){cor(y, expo$rsi)})  
tcc_correlations<-apply(exprs(tcc), 1, function(y){cor(y, tcc$RSI)})
```

4 Analysis of correlations

This shows that most genes are not correlated to RSI.

```
hist(ccle_correlations, main="CCLE")
```

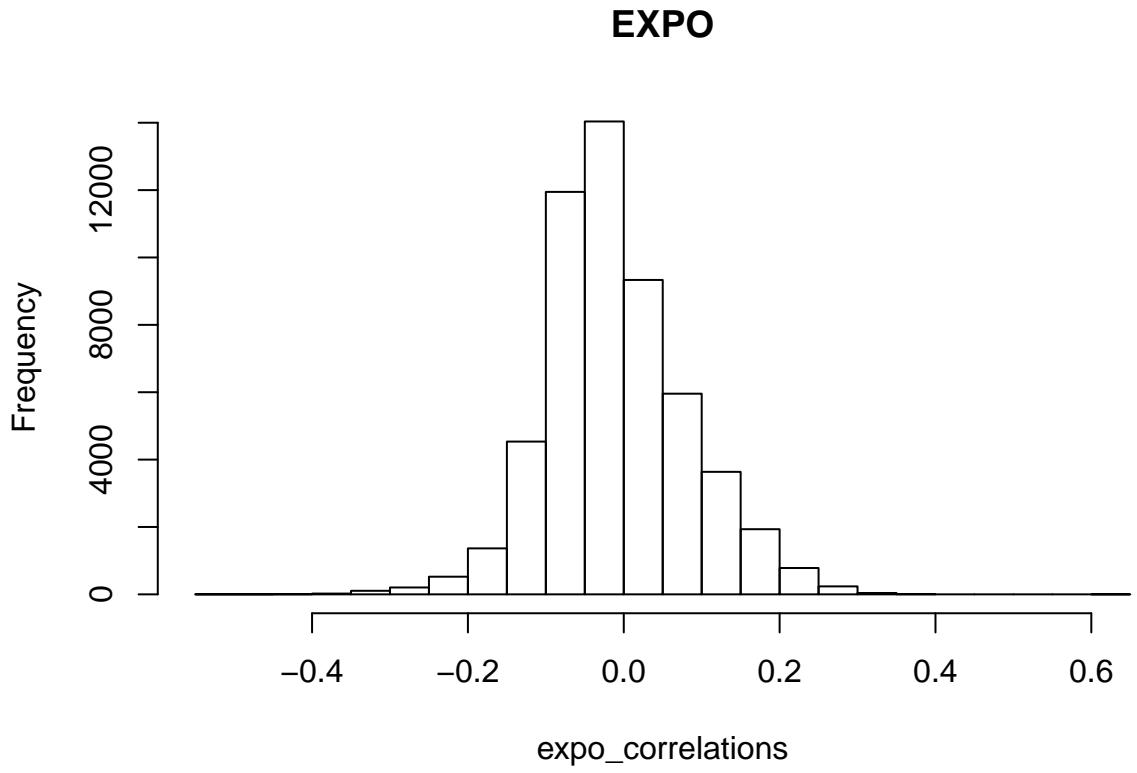
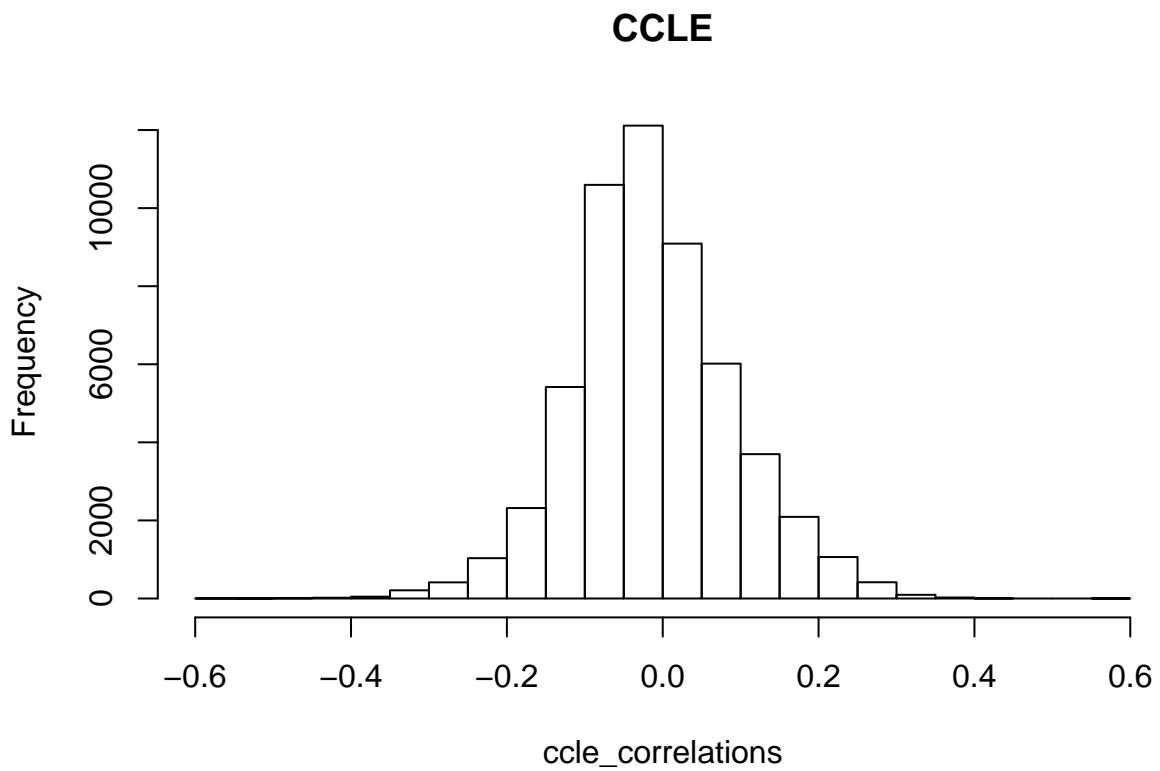
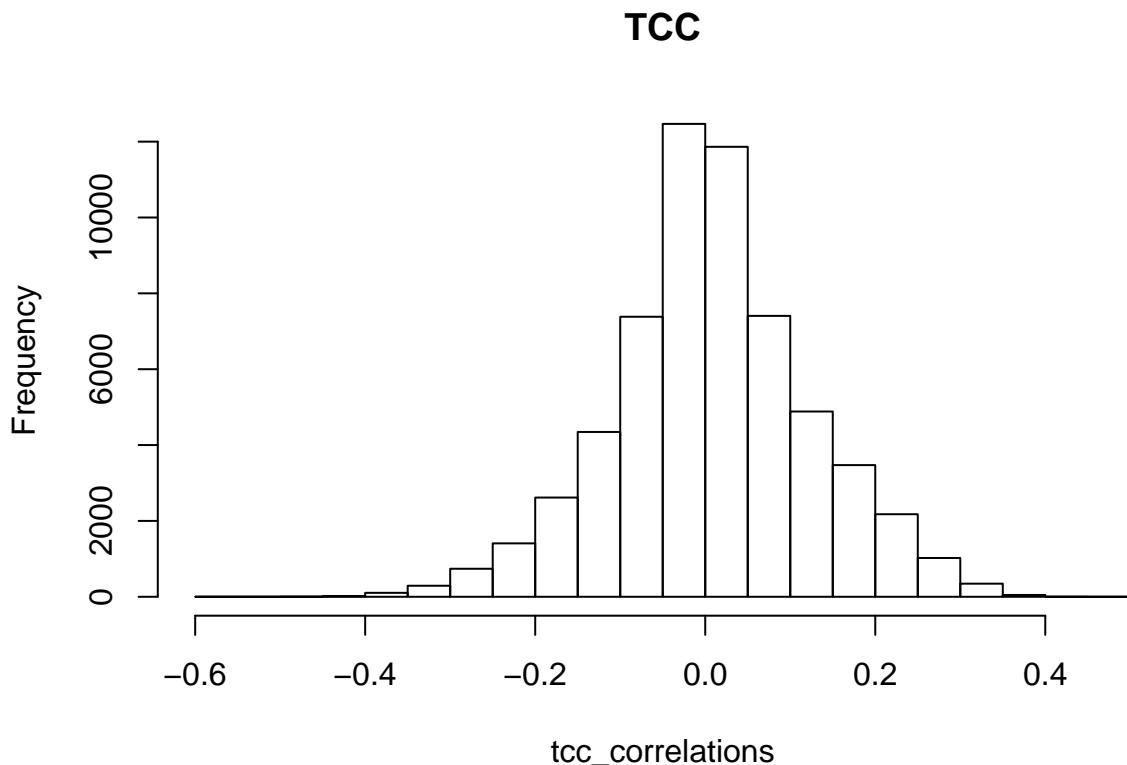


Table 1: TCC Correlated probesets

	R
merck-NM_176783_at	-0.5017434
merck-NM_002198_at	-0.5834575
merck2-AW137708_at	-0.5581379
merck2-DQ891838_at	-0.5115138

Table 2: EXPO correlated probesets

	R
202123_s_at	0.6209414
202531_at	-0.5324200



What probesets are correlated at a modest R=0.5?

```
kable(tcc_correlations[which(abs(tcc_correlations)>0.5)], col.names=c("R"), caption="TCC Correlations")

kable(expo_correlations[which(abs(expo_correlations)>0.5)], col.names=c("R"), caption="EXPO Correlations")

kable(ccle_correlations[which(abs(ccle_correlations)>0.5)], col.names=c("R"), caption="CCLE Correlations")

u133_coeffs<-c("AFFX-HUMISGF3A/M97935_MA_at"=0.0254552046046931,
               "201209_at" = -0.020446868981449,
```

Table 3: CCLE correlated probesets

	R
202123_s_at	0.5855299
202531_at	-0.5848877
230405_at	-0.5149779
238725_at	-0.5830651

```

"201466_s_at" = 0.0128282697929755,
"201783_s_at" = -0.00381713588908642,
"202123_s_at" = 0.107021274358643,
"202531_at" = -0.0441682719957224,
"205962_at" = -0.00924314364533049,
"207957_s_at" = -0.00175888592095915,
"208762_at" = -0.00025093287356907,
"211110_s_at" = -0.00980092741843779)

hursta_coeffs<-c("merck-NM_139266_at"=0.0254552046046931,
"merck-NM_004964_at" = -0.020446868981449,
"merck-NM_002228_at" = 0.0128282697929755,
"merck2-BC069248_at" = -0.00381713588908642,
"merck-NM_007313_s_at" = 0.107021274358643,
"merck-NM_002198_at" = -0.0441682719957224,
"merck-BQ646444_a_at" = -0.00924314364533049,
"merck2-X06318_at" = -0.00175888592095915,
"merck-NM_001005782_s_at" = -0.00025093287356907,
"merck-NM_000044_a_at" = -0.00980092741843779)

```

5 Gene Correlations

Show the correlations of the ten genes.

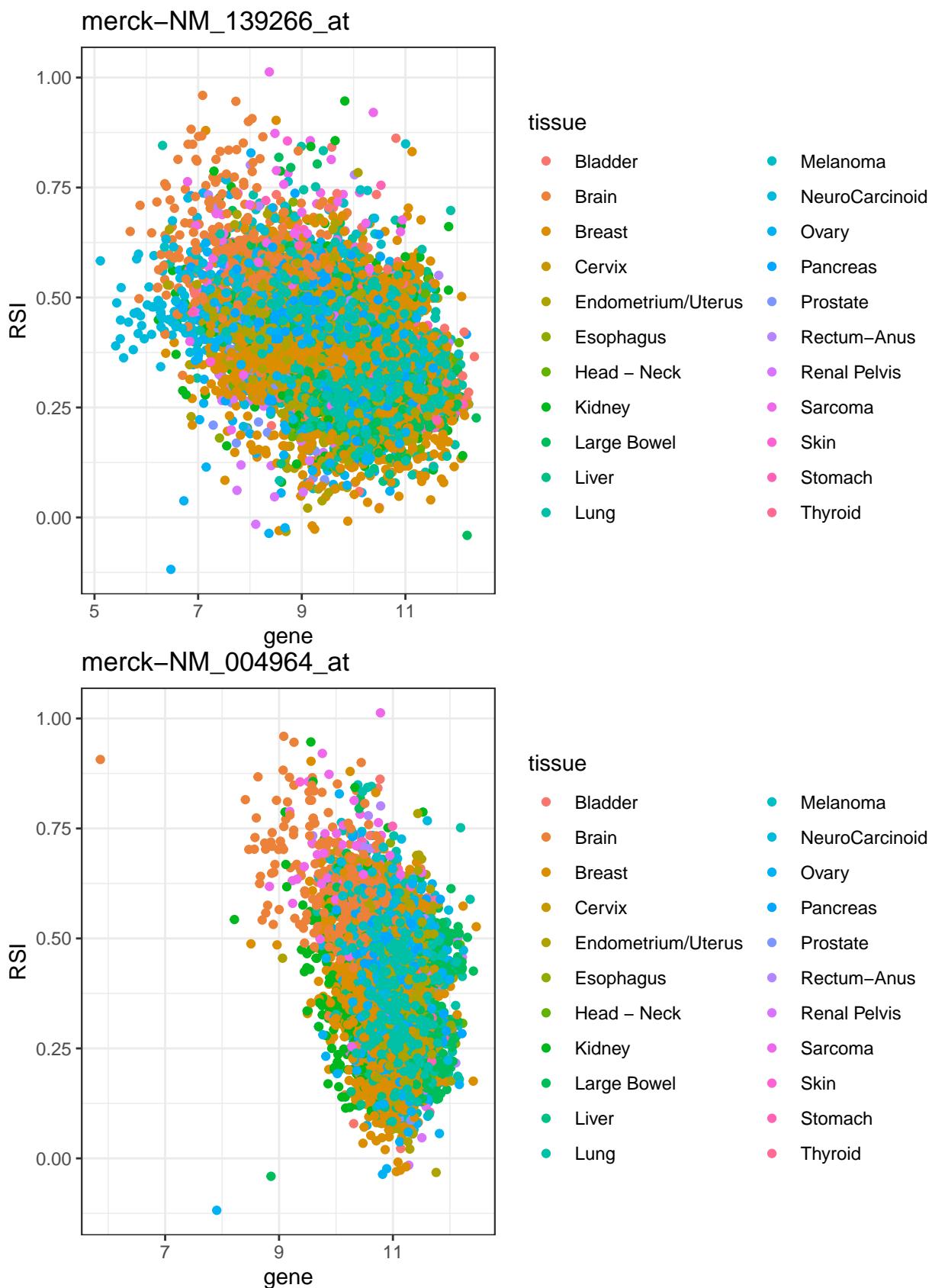
5.1 TCC

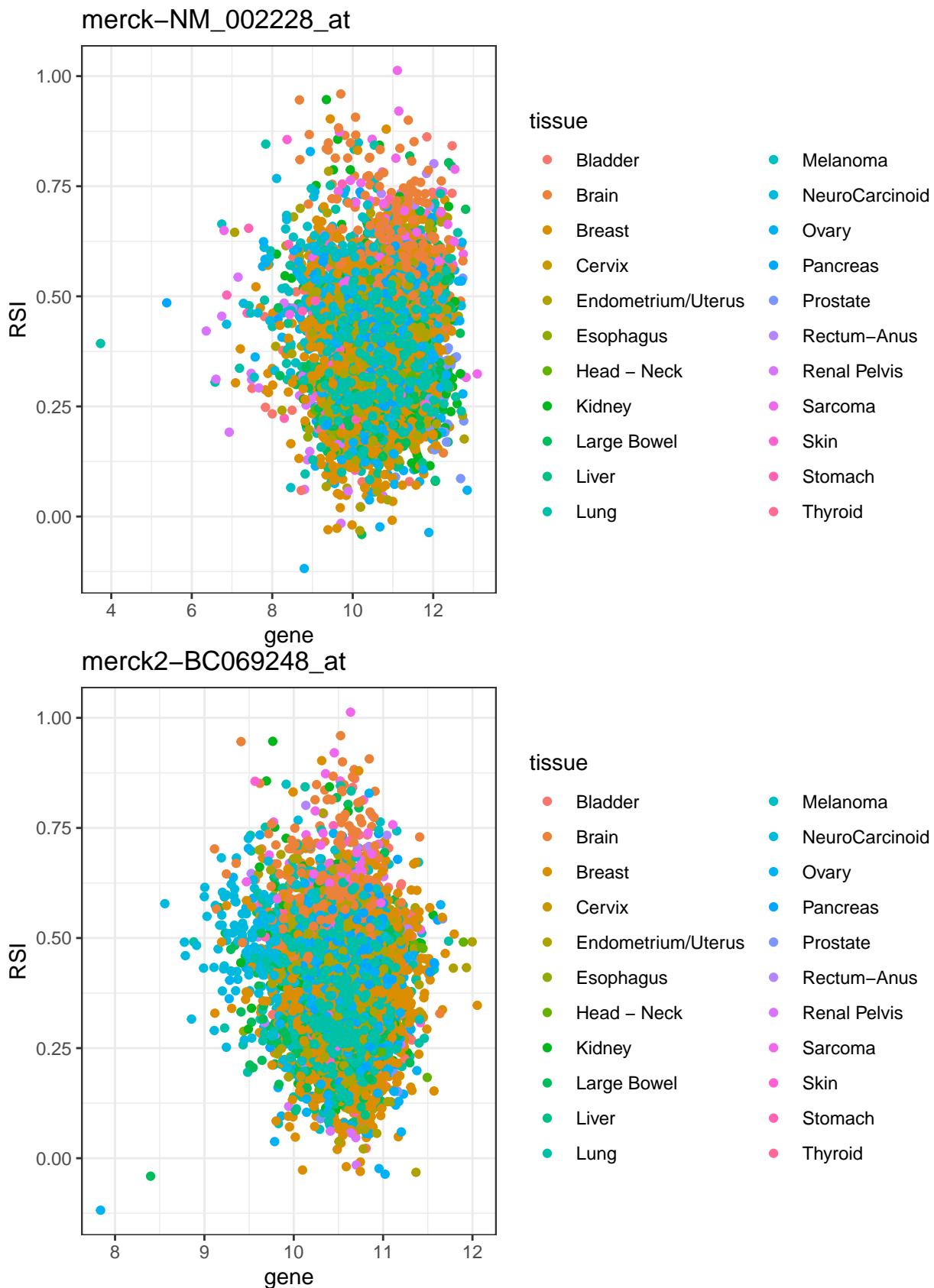
```

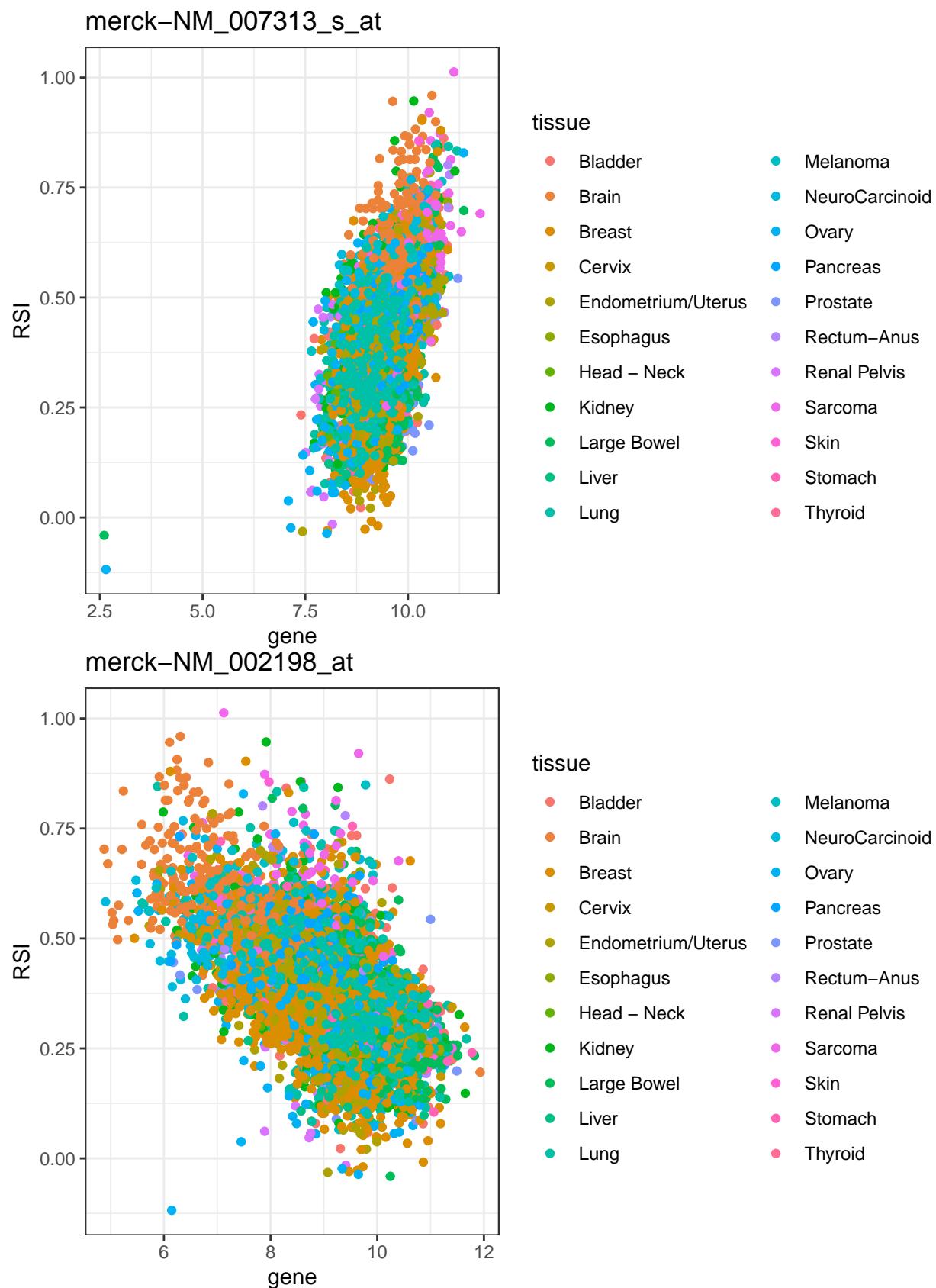
for (g in names(hursta_coeffs)) {
  myplot<-ggplot(data.frame(gene=exprs(tcc)[g,], RSI=tcc$RSI, tissue=tcc$SO0_Conformed), aes(x=
    geom_point() +
    theme_bw() +
    ggtitle(g)

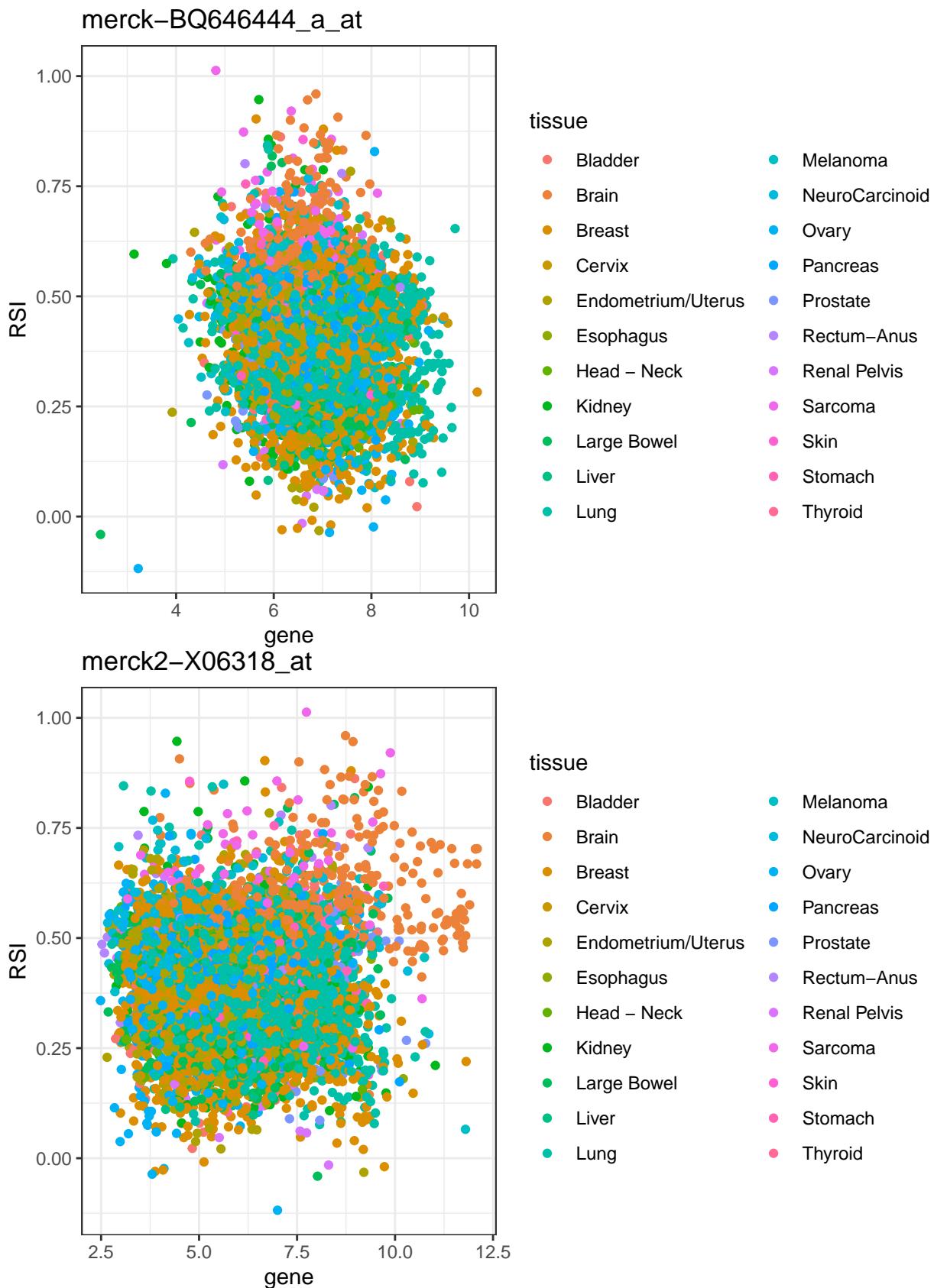
  print(myplot)
}

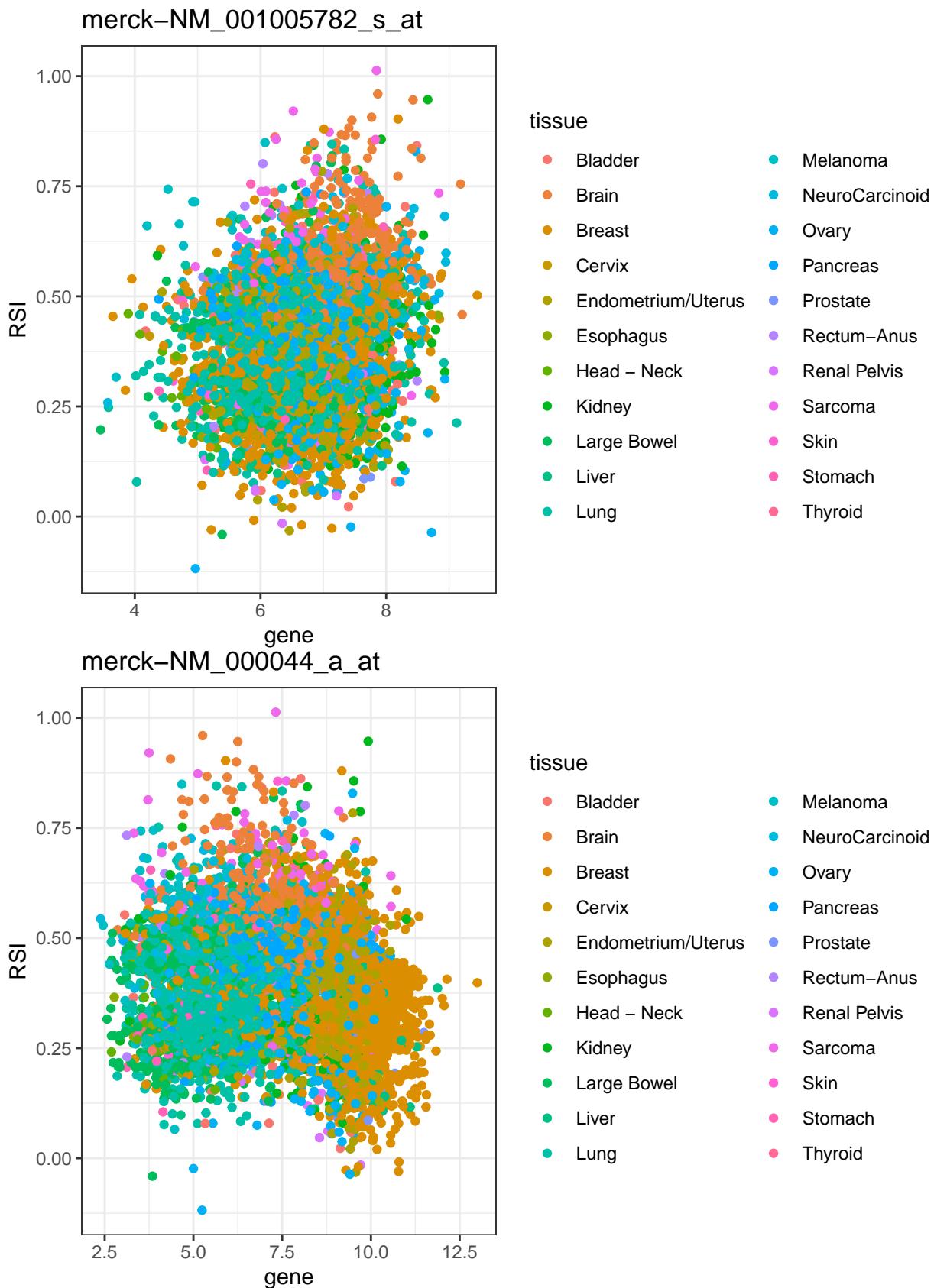
```





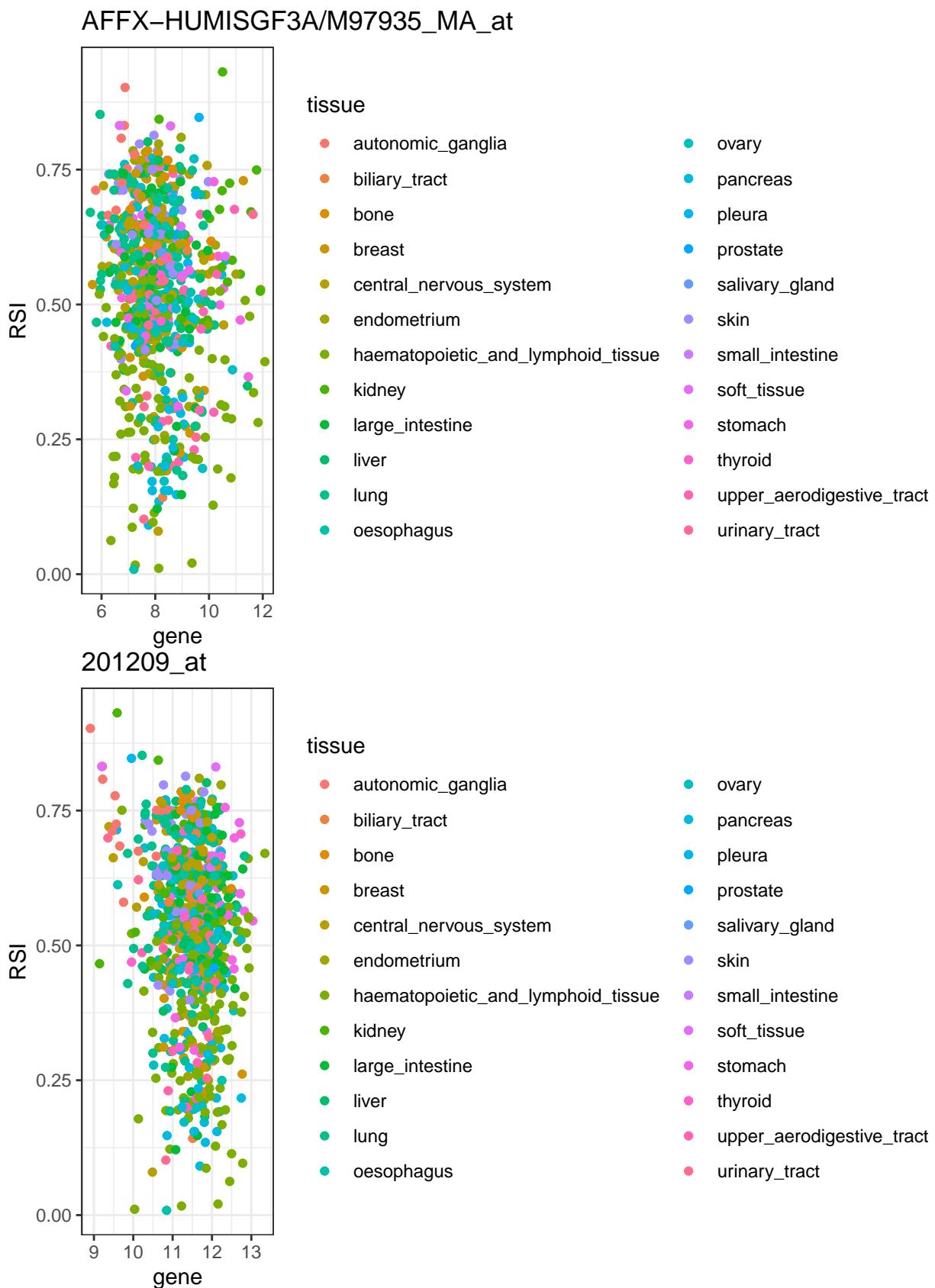


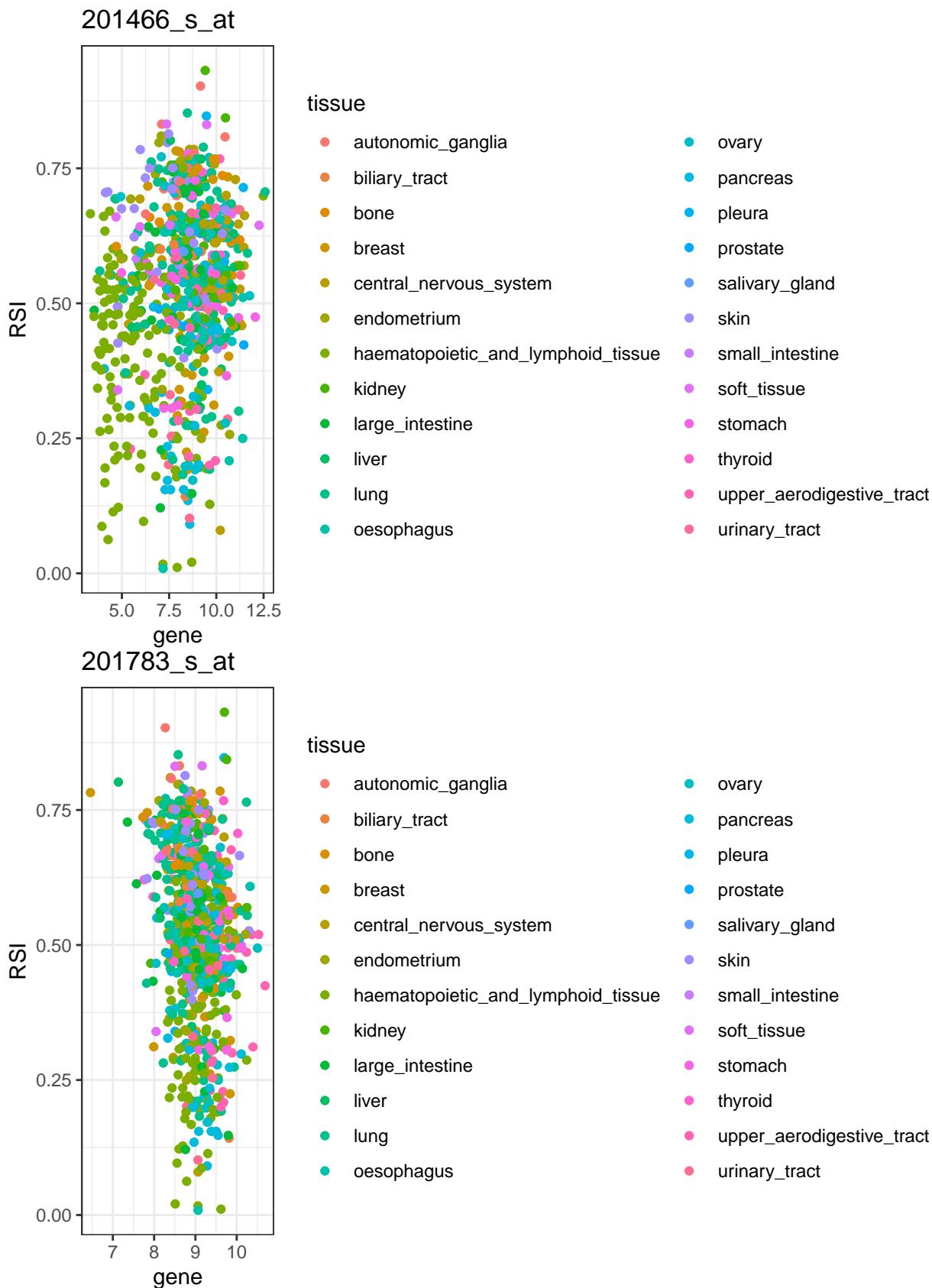


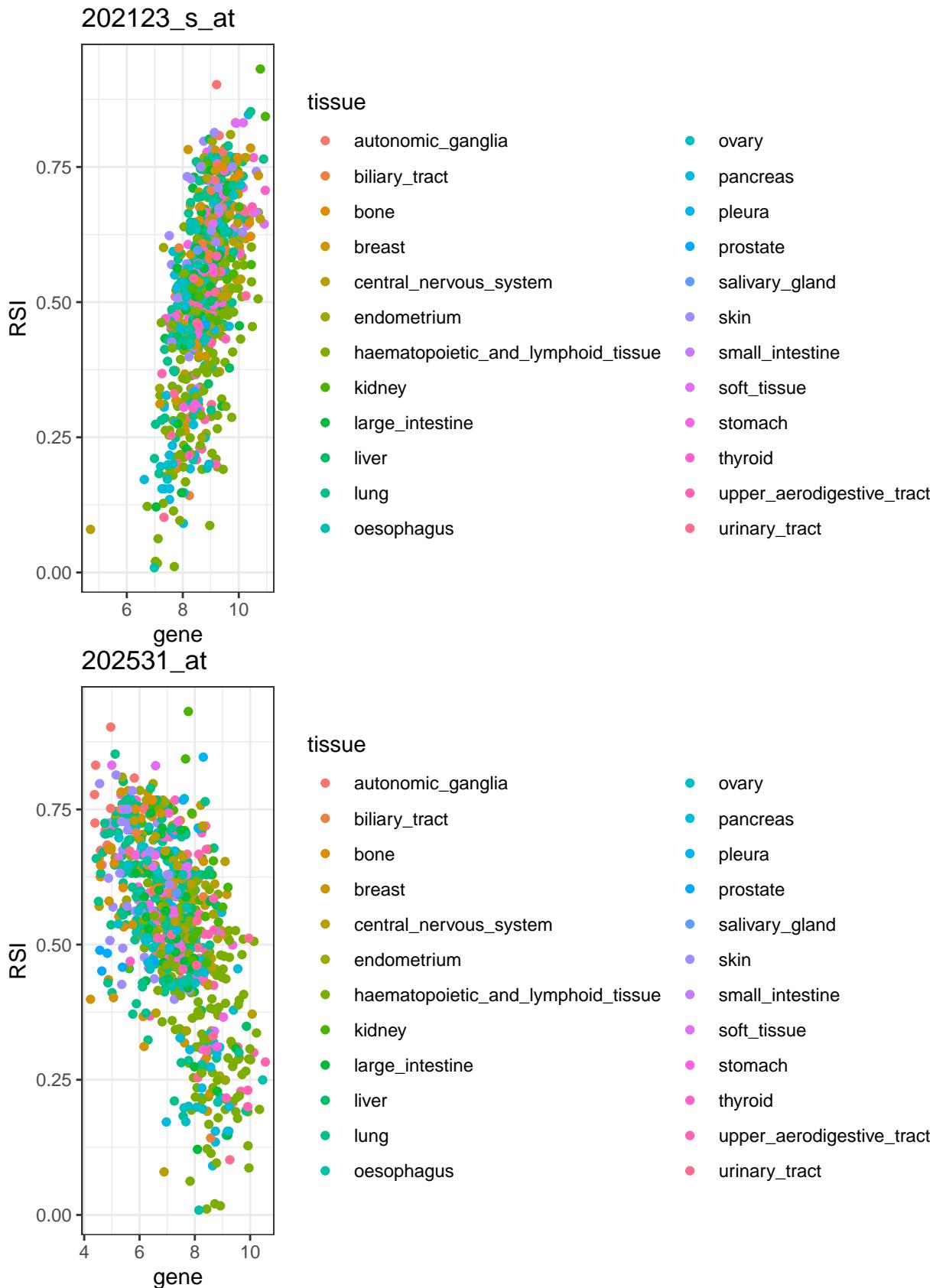


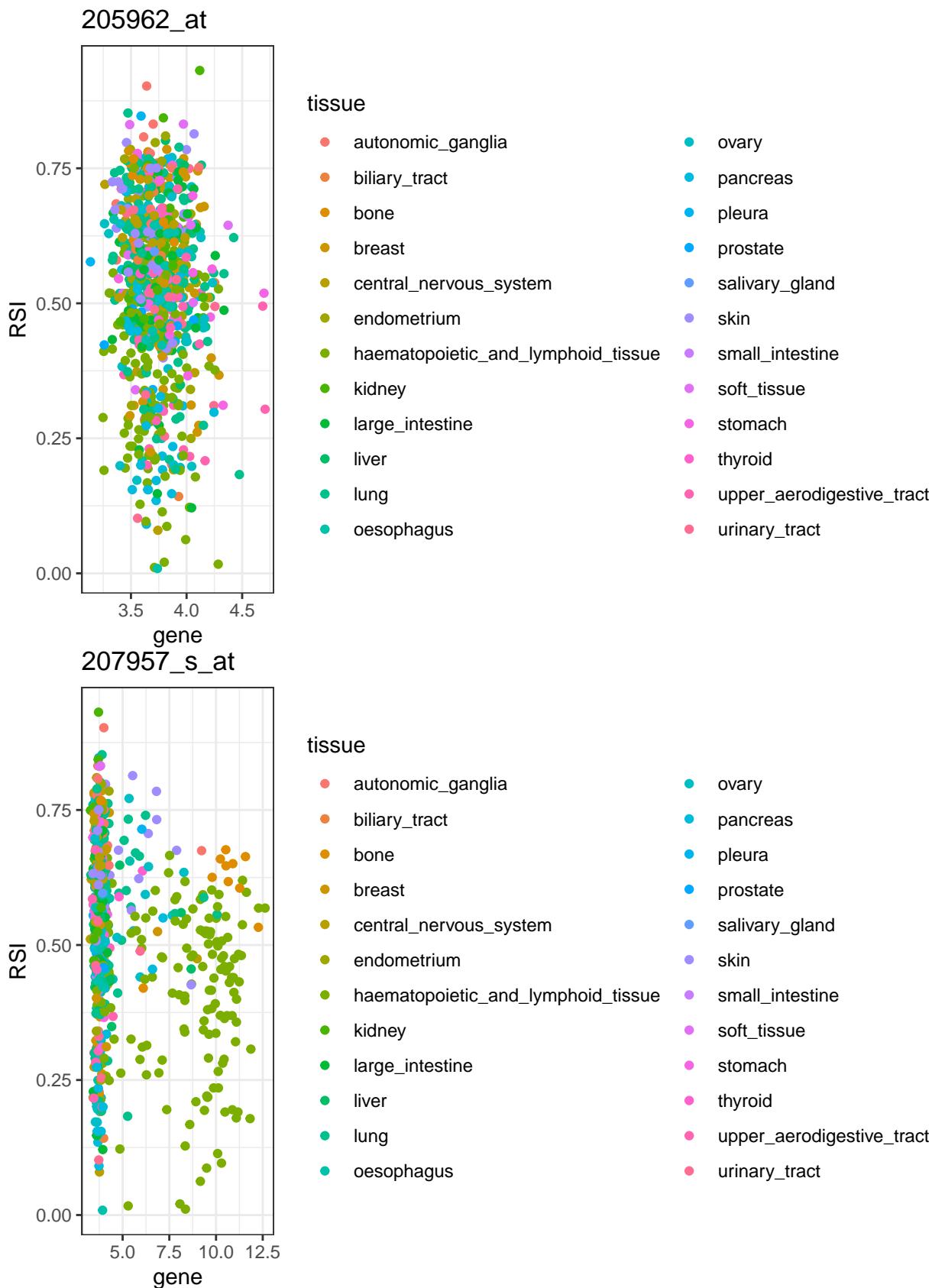
5.2 CCLE

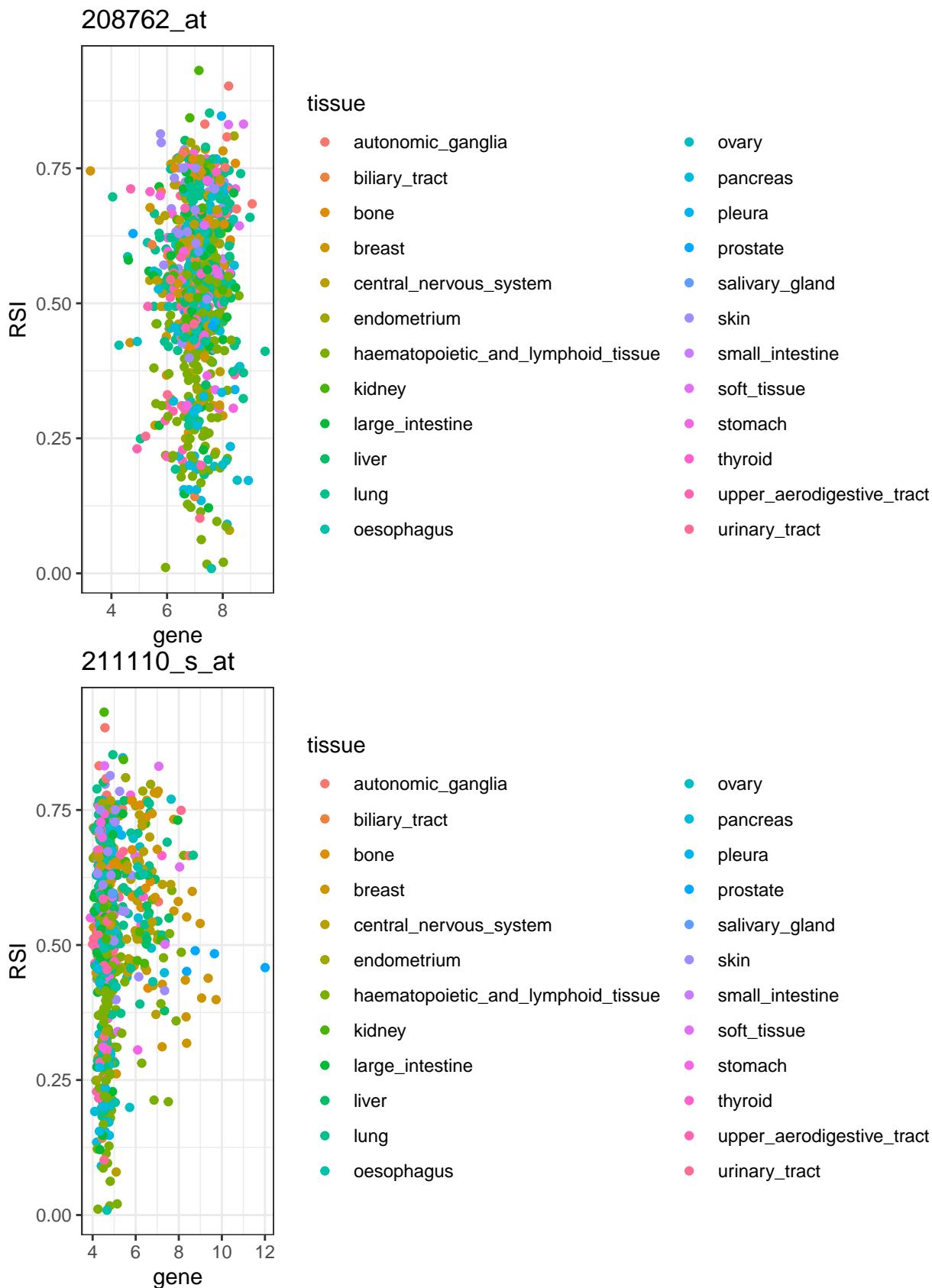
```
for (g in names(u133_coefficients)) {  
  myplot<-ggplot(data.frame(gene=exprs(ccle)[g,], RSI=ccle$RSI, tissue=ccle$`Site Primary`), a  
  geom_point() +  
  theme_bw() +  
  ggtitle(g)  
  
  print(myplot)  
}
```







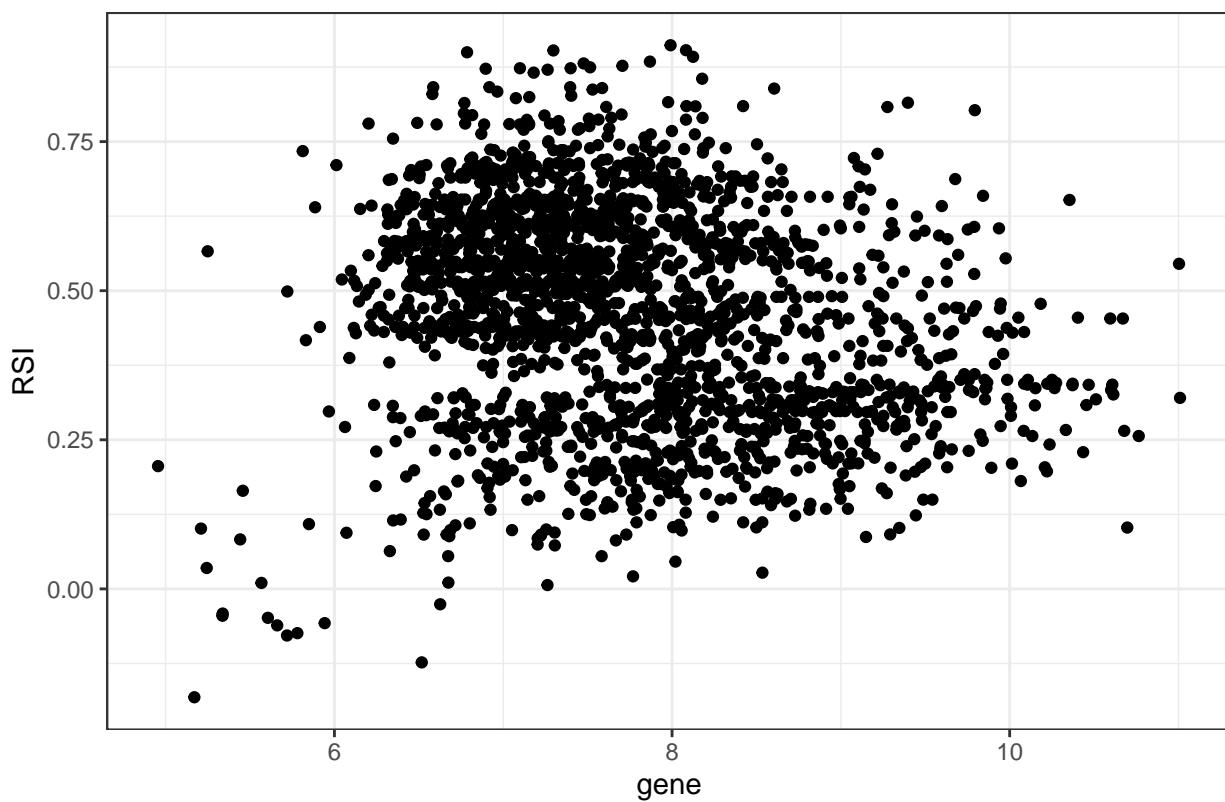




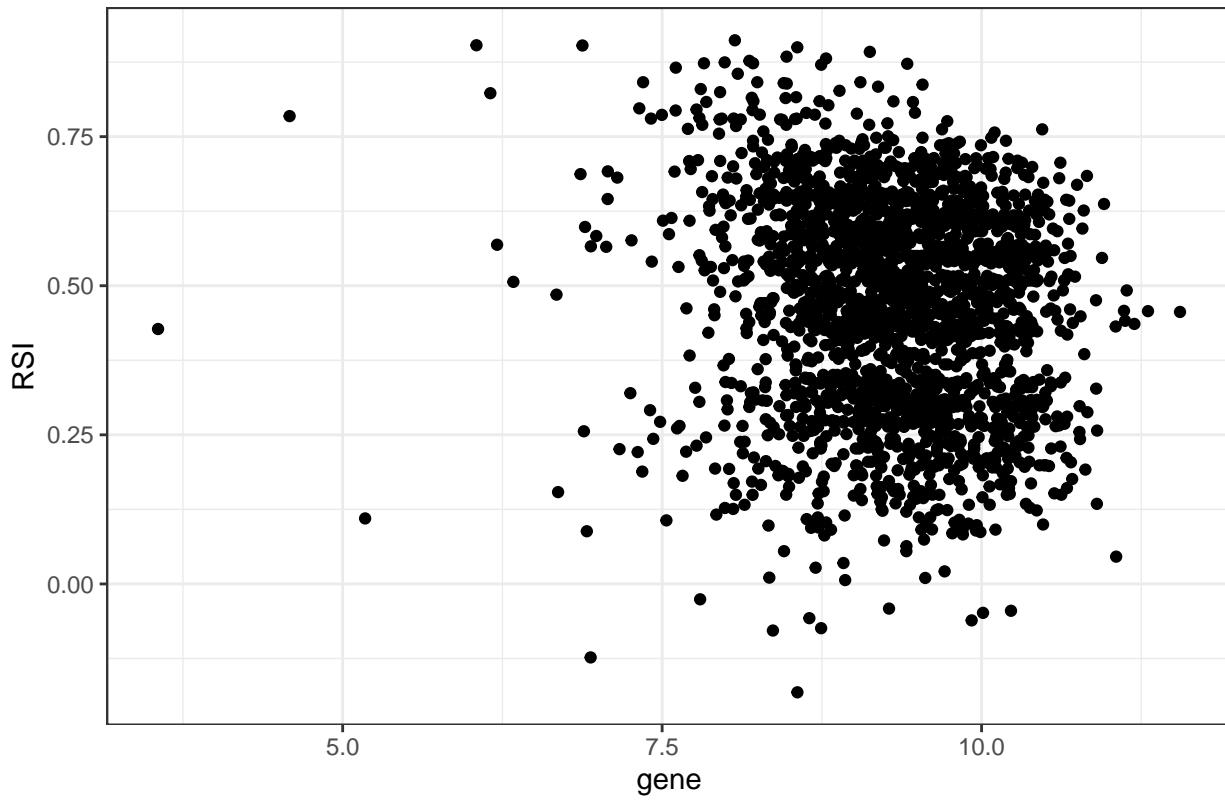
5.3 EXPO

```
for (g in names(u133_coeff)) {  
  myplot<-ggplot(data.frame(gene=exprs(expo)[g,], RSI=expo$rsi, tissue=expo$source_name_ch1),  
    geom_point() +  
    theme_bw() +  
    ggtitle(g)  
  
  print(myplot)  
}
```

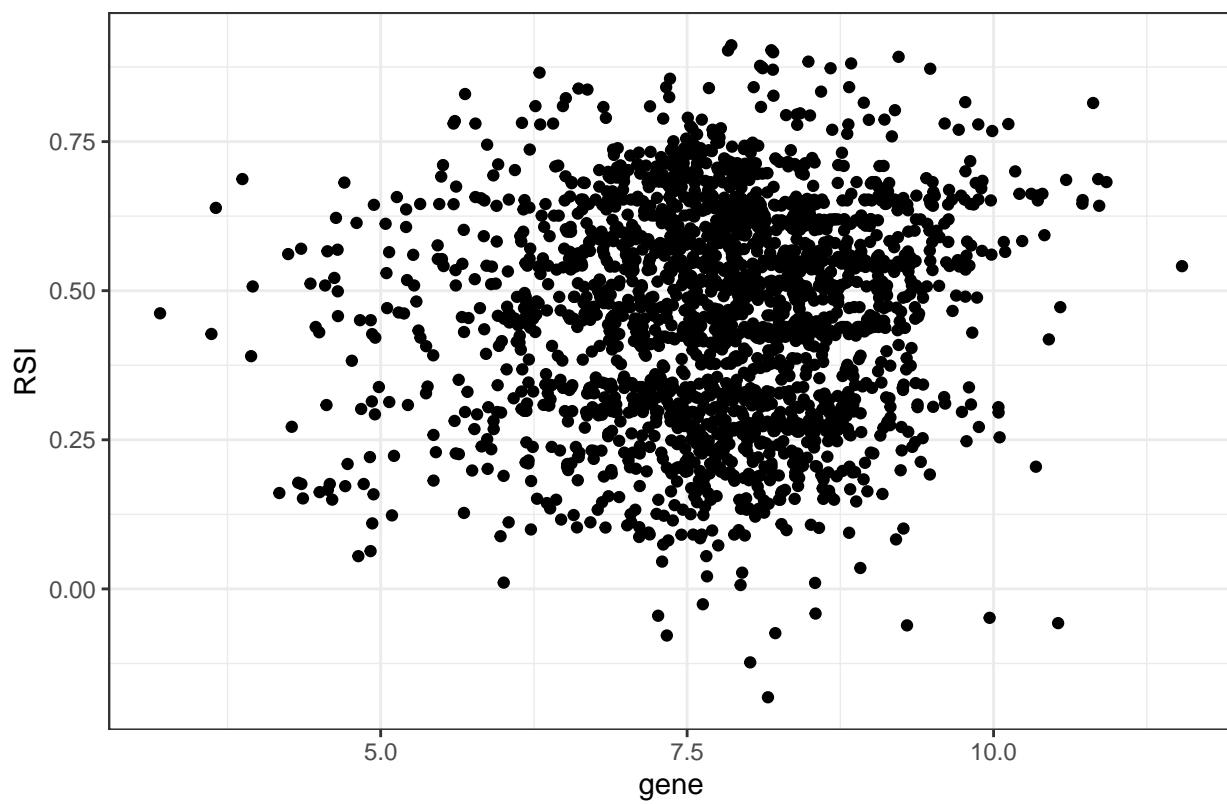
AFFX-HUMISGF3A/M97935_MA_at



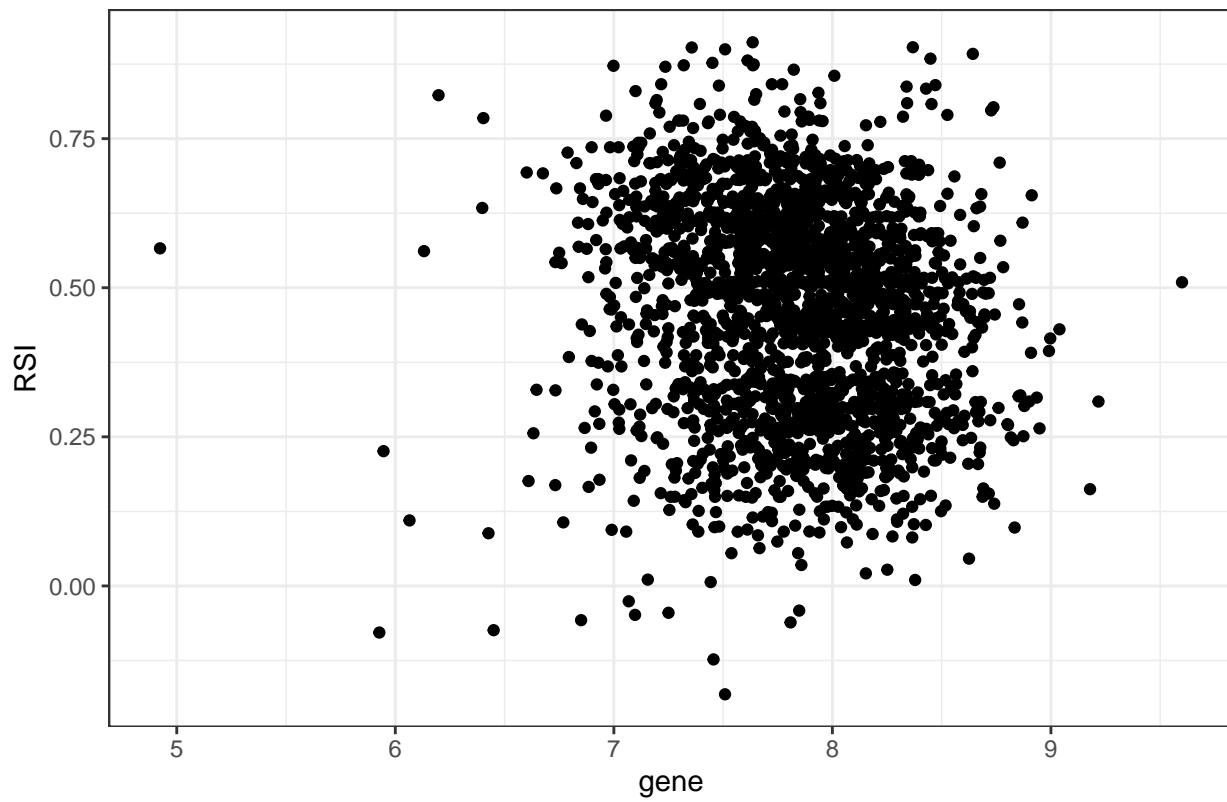
201209_at



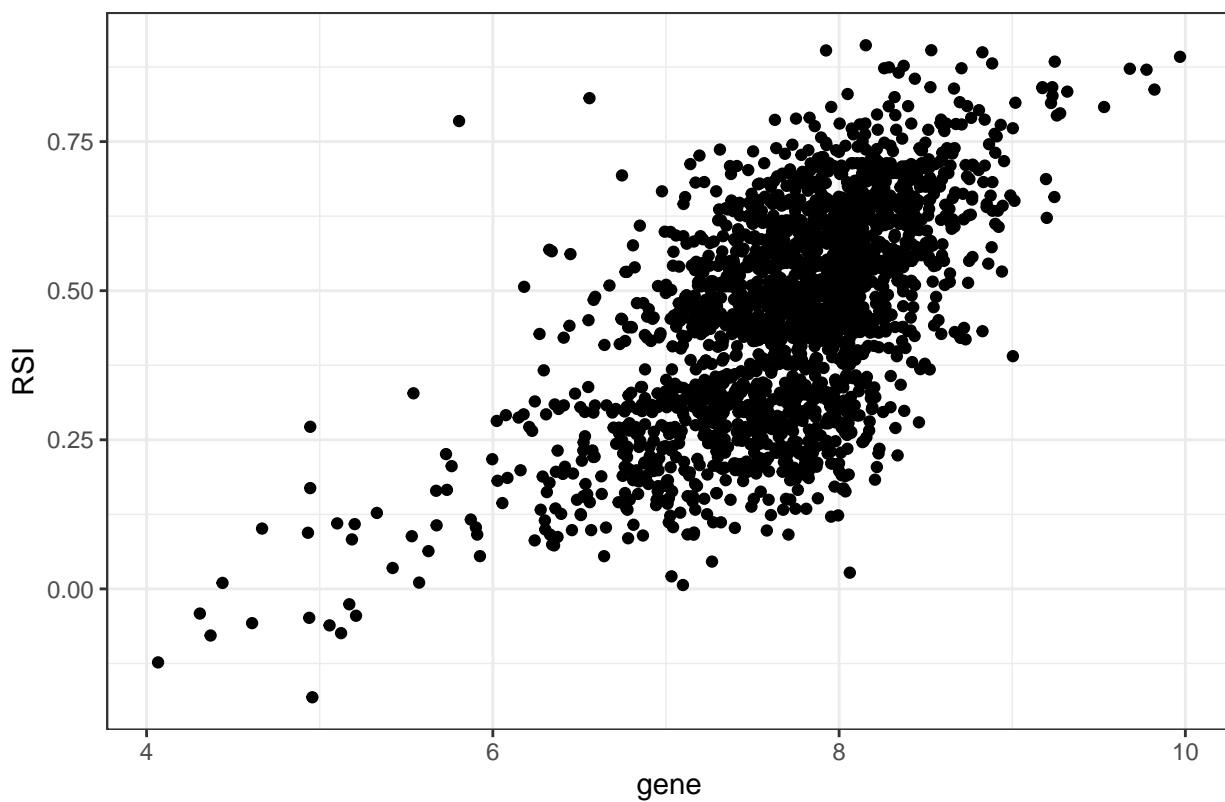
201466_s_at



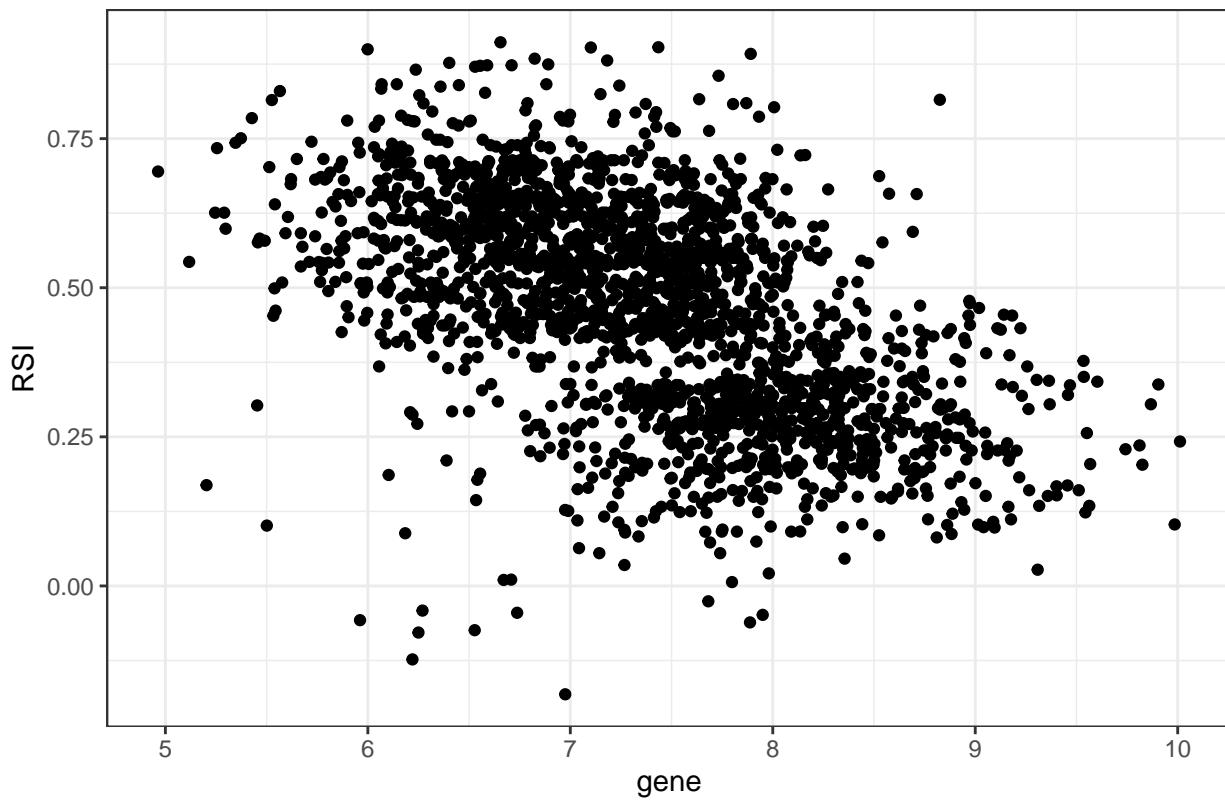
201783_s_at



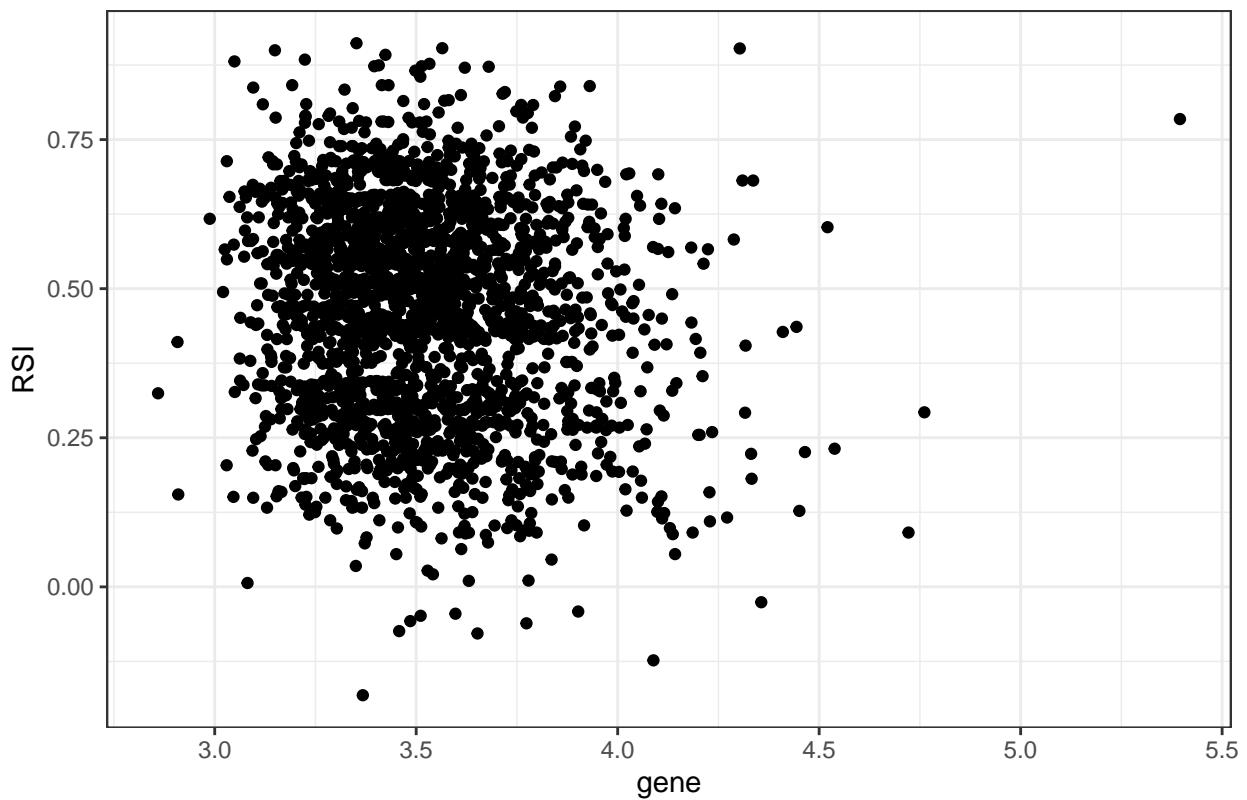
202123_s_at



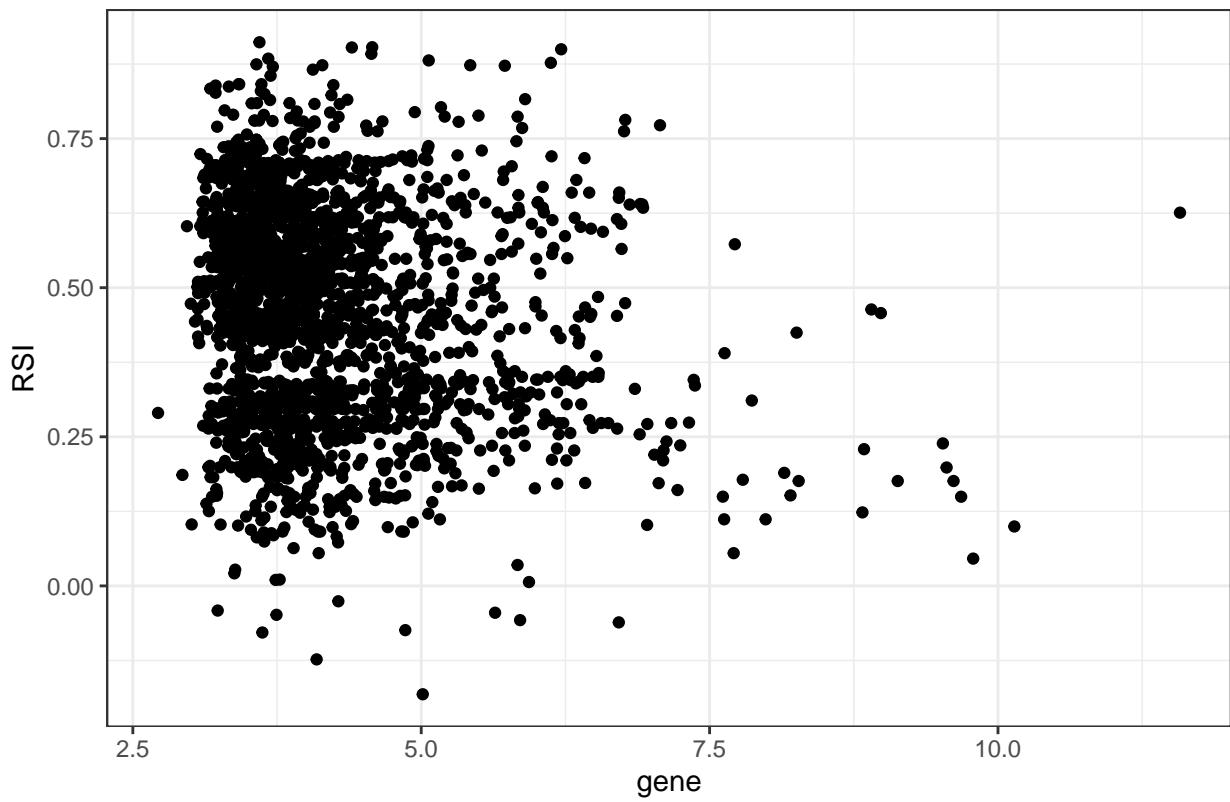
202531_at



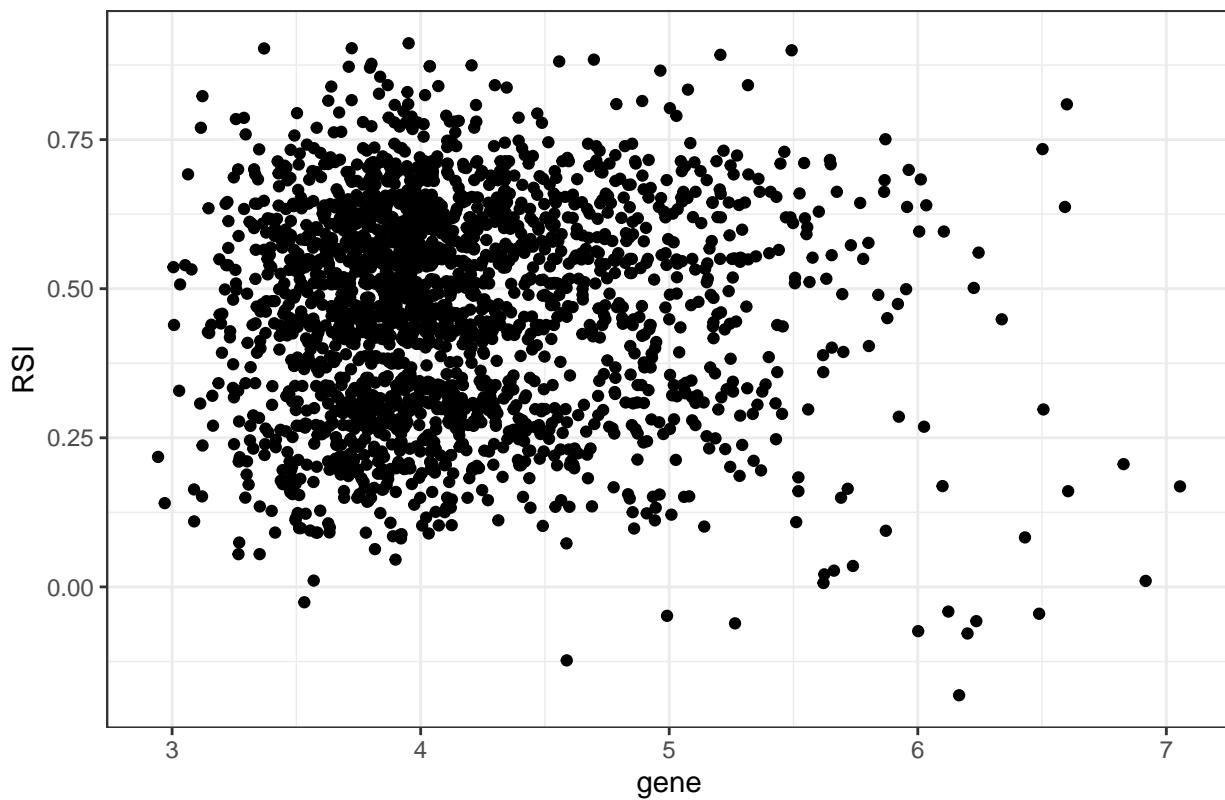
205962_at



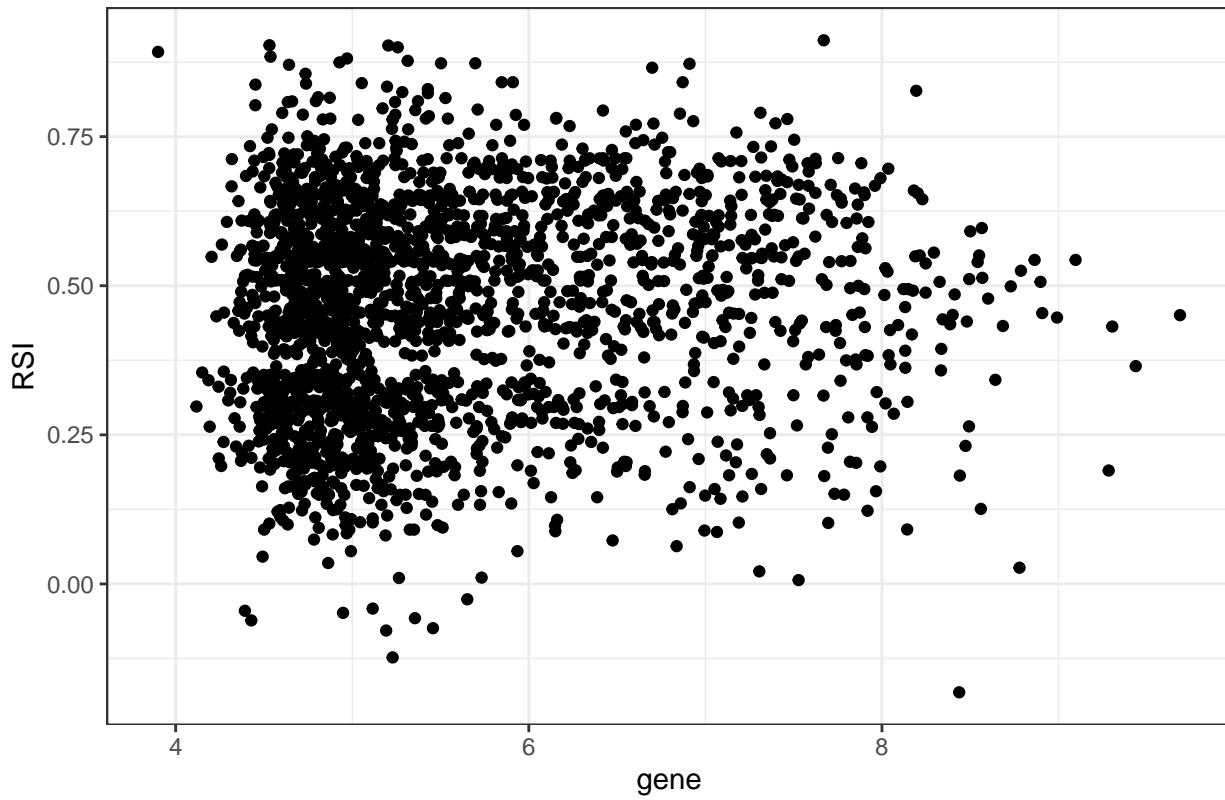
207957_s_at



208762_at



211110_s_at



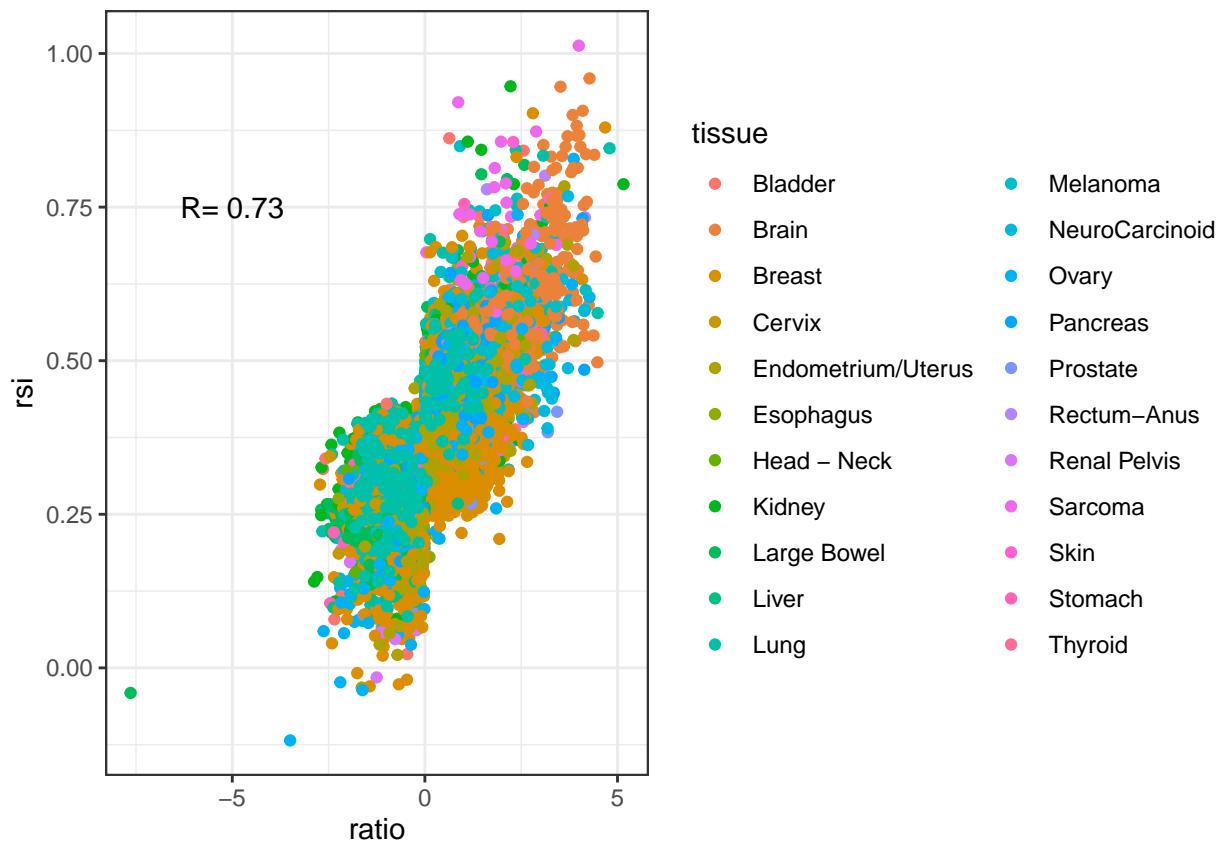
6 ABL-IRF1

7 TCC

```
ABL1<-translation %>% filter(~Gene Symbol=="ABL1") %>% pull("HuRSTA Array")
IRF1<-translation %>% filter(~Gene Symbol=="IRF1") %>% pull("HuRSTA Array")

ratio<-exprs(tcc)[ABL1,]-exprs(tcc)[IRF1,]

ggplot(data.frame(ratio=ratio, rsi=tcc$RSI, tissue=tcc$S00_Conformed), aes(x=ratio, y=rsi, col=
  geom_point() +
  theme_bw() +
  annotate(geom="text",x=-5,y=0.75, label=sprintf("R=%5.2f",cor(ratio, tcc$RSI)))
```

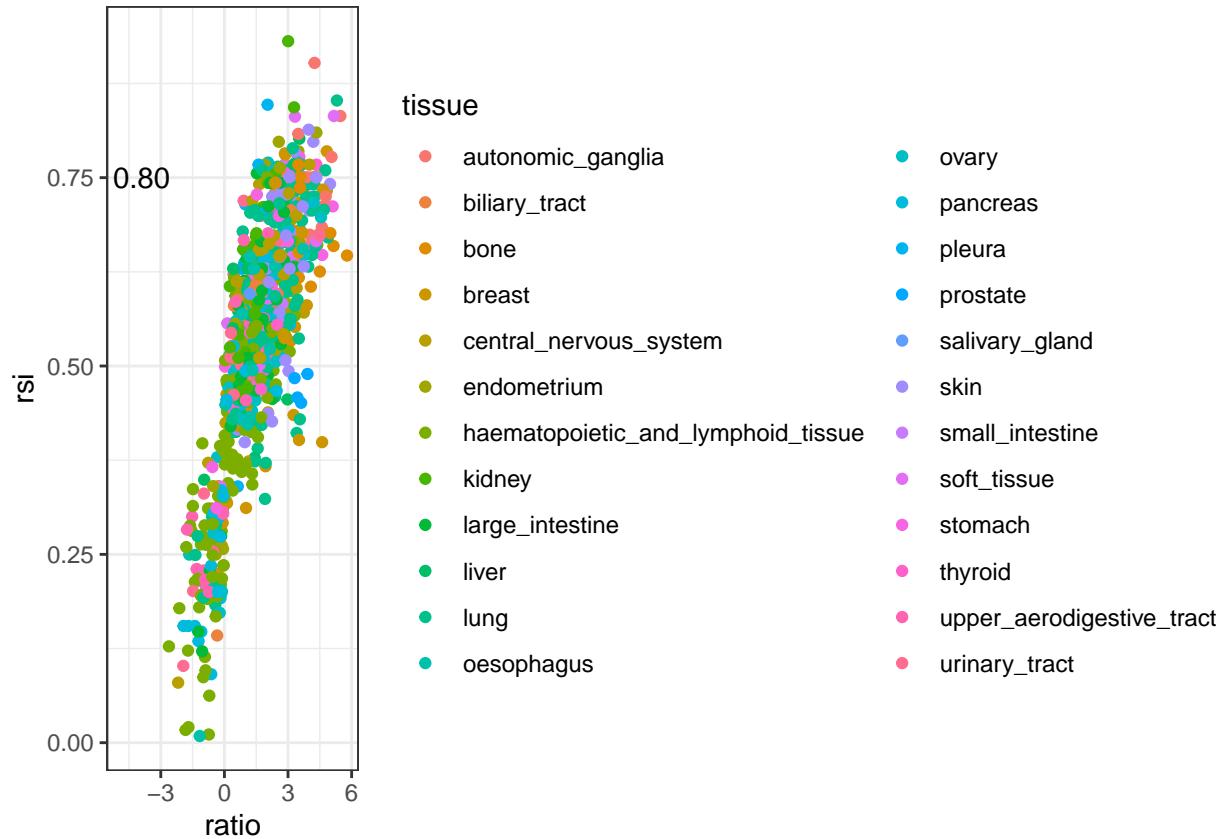


8 CCLE

```
ABL1<-translation %>% filter(~Gene Symbol=="ABL1") %>% pull("HG-U133 Plus 2.0 Probeset")
IRF1<-translation %>% filter(~Gene Symbol=="IRF1") %>% pull("HG-U133 Plus 2.0 Probeset")

ratio<-exprs(ccle)[ABL1,]-exprs(ccle)[IRF1,]
```

```
ggplot(data.frame(ratio=ratio, rsi=ccle$RSI, tissue=ccle$`Site Primary`), aes(x=ratio, y=rsi, color=tissue))
  geom_point() +
  theme_bw() +
  annotate(geom="text",x=-5,y=0.75, label=sprintf("R=%5.2f",cor(ratio, ccle$RSI)))
```

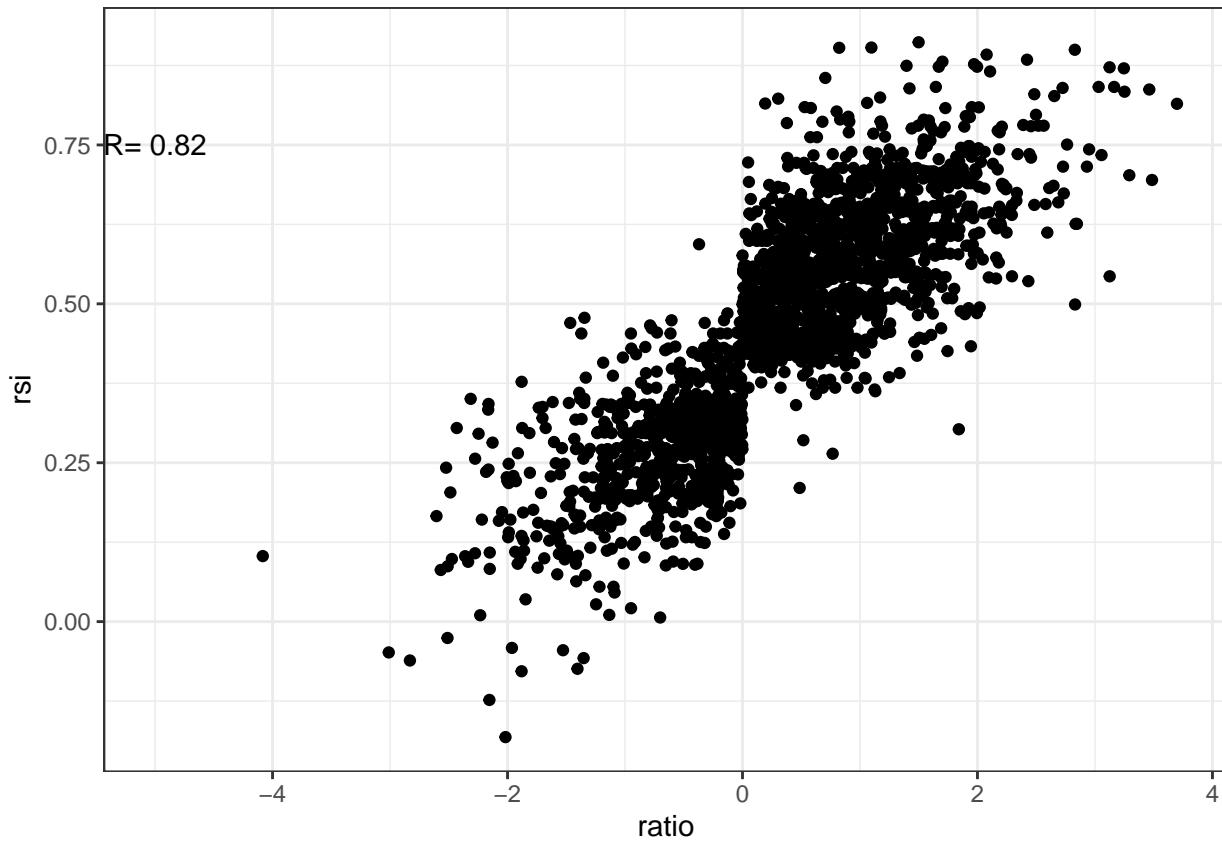


9 expo

```
ABL1<-translation %>% filter(`Gene Symbol`=="ABL1") %>% pull("HG-U133 Plus 2.0 Probeset")
IRF1<-translation %>% filter(`Gene Symbol`=="IRF1") %>% pull("HG-U133 Plus 2.0 Probeset")

ratio<-exprs(expo)[ABL1,]-exprs(expo)[IRF1,]

ggplot(data.frame(ratio=ratio, rsi=expo$rsi, tissue=expo$source_name_ch1), aes(x=ratio, y=rsi))
  geom_point() +
  theme_bw() +
  annotate(geom="text",x=-5,y=0.75, label=sprintf("R=%5.2f",cor(ratio, expo$rsi)))
```



Session Information This information is included in the report to facilitate reproducible research. The specific environment under which the code was run is detailed. Also included are the packages that are loaded during the run.

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

Table 4: Supplemental Table: Reproducibility Software Session Information

Setting	
version	R version 3.6.3 (2020-02-29)
os	Ubuntu 16.04.6 LTS
system	x86_64, linux-gnu
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Etc/UTC
date	2020-05-21

Table 5: Supplemental Table: Reproducibility Software Package Version Information

	loadedversion	date	source	library
assertthat	0.2.1	2019-03-21	CRAN (R 3.6.3)	[1]
backports	1.1.7	2020-05-13	CRAN (R 3.6.3)	[1]
Biobase	2.46.0	2019-10-29	Bioconductor	[1]
BiocGenerics	0.32.0	2019-10-29	Bioconductor	[1]
callr	3.4.3	2020-03-28	CRAN (R 3.6.3)	[1]
cellranger	1.1.0	2016-07-27	CRAN (R 3.6.3)	[1]
cli	2.0.2	2020-02-28	CRAN (R 3.6.3)	[1]
colorspace	1.4-1	2019-03-18	CRAN (R 3.6.3)	[1]
CosmosReportTemplates	0.0.4	2020-05-18	local	[1]
crayon	1.3.4	2017-09-16	CRAN (R 3.6.3)	[1]
desc	1.2.0	2018-05-01	CRAN (R 3.6.3)	[1]
devtools	2.3.0	2020-04-10	CRAN (R 3.6.3)	[1]
digest	0.6.25	2020-02-23	CRAN (R 3.6.3)	[1]
dplyr	0.8.5	2020-03-07	CRAN (R 3.6.3)	[1]
ellipsis	0.3.1	2020-05-15	CRAN (R 3.6.3)	[1]
evaluate	0.14	2019-05-28	CRAN (R 3.6.3)	[1]
fansi	0.4.1	2020-01-08	CRAN (R 3.6.3)	[1]
farver	2.0.3	2020-01-16	CRAN (R 3.6.3)	[1]
fs	1.4.1	2020-04-04	CRAN (R 3.6.3)	[1]
ggplot2	3.3.0	2020-03-05	CRAN (R 3.6.3)	[1]

Table 5: Supplemental Table: Reproducibility Software Package Version Information (*continued*)

	loadedversion	date	source	library
glue	1.4.1	2020-05-13	CRAN (R 3.6.3)	[1]
gtable	0.3.0	2019-03-25	CRAN (R 3.6.3)	[1]
hms	0.5.3	2020-01-08	CRAN (R 3.6.3)	[1]
htmltools	0.4.0	2019-10-04	CRAN (R 3.6.3)	[1]
httr	1.4.1	2019-08-05	CRAN (R 3.6.3)	[1]
kableExtra	1.1.0	2019-03-16	CRAN (R 3.6.3)	[1]
knitr	1.28	2020-02-06	CRAN (R 3.6.3)	[1]
labeling	0.3	2014-08-23	CRAN (R 3.6.3)	[1]
lifecycle	0.2.0	2020-03-06	CRAN (R 3.6.3)	[1]
magrittr	1.5	2014-11-22	CRAN (R 3.6.3)	[1]
memoise	1.1.0	2017-04-21	CRAN (R 3.6.3)	[1]
munsell	0.5.0	2018-06-12	CRAN (R 3.6.3)	[1]
pillar	1.4.4	2020-05-05	CRAN (R 3.6.3)	[1]
pkgbuild	1.0.8	2020-05-07	CRAN (R 3.6.3)	[1]
pkgconfig	2.0.3	2019-09-22	CRAN (R 3.6.3)	[1]
pkgload	1.0.2	2018-10-29	CRAN (R 3.6.3)	[1]
prettyunits	1.1.1	2020-01-24	CRAN (R 3.6.3)	[1]
processx	3.4.2	2020-02-09	CRAN (R 3.6.3)	[1]
ps	1.3.3	2020-05-08	CRAN (R 3.6.3)	[1]
purrr	0.3.4	2020-04-17	CRAN (R 3.6.3)	[1]
R6	2.4.1	2019-11-12	CRAN (R 3.6.3)	[1]
Rcpp	1.0.4.6	2020-04-09	CRAN (R 3.6.3)	[1]
readr	1.3.1	2018-12-21	CRAN (R 3.6.3)	[1]
readxl	1.3.1	2019-03-13	CRAN (R 3.6.3)	[1]
remotes	2.1.1	2020-02-15	CRAN (R 3.6.3)	[1]
rlang	0.4.6	2020-05-02	CRAN (R 3.6.3)	[1]
rmarkdown	2.1	2020-01-20	CRAN (R 3.6.3)	[1]
rprojroot	1.3-2	2018-01-03	CRAN (R 3.6.3)	[1]
rstudioapi	0.11	2020-02-07	CRAN (R 3.6.3)	[1]
rvest	0.3.5	2019-11-08	CRAN (R 3.6.3)	[1]
scales	1.1.1	2020-05-11	CRAN (R 3.6.3)	[1]
sessioninfo	1.1.1	2018-11-05	CRAN (R 3.6.3)	[1]
stringi	1.4.6	2020-02-17	CRAN (R 3.6.3)	[1]
stringr	1.4.0	2019-02-10	CRAN (R 3.6.3)	[1]
testthat	2.3.2	2020-03-02	CRAN (R 3.6.3)	[1]
tibble	3.0.1	2020-04-20	CRAN (R 3.6.3)	[1]
tidyselect	1.1.0	2020-05-11	CRAN (R 3.6.3)	[1]
usethis	1.6.1	2020-04-29	CRAN (R 3.6.3)	[1]
vctrs	0.3.0	2020-05-11	CRAN (R 3.6.3)	[1]

Table 5: Supplemental Table: Reproducibility Software Package Version Information (*continued*)

	loadedversion	date	source	library
viridisLite	0.3.0	2018-02-01	CRAN (R 3.6.3)	[1]
webshot	0.5.2	2019-11-22	CRAN (R 3.6.3)	[1]
withr	2.2.0	2020-04-20	CRAN (R 3.6.3)	[1]
xfun	0.14	2020-05-20	CRAN (R 3.6.3)	[1]
xml2	1.3.2	2020-04-23	CRAN (R 3.6.3)	[1]
yaml	2.2.1	2020-02-01	CRAN (R 3.6.3)	[1]

¹ /share/data2/eschrich/R/x86_64-pc-linux-gnu-library/3.6

² /usr/local/lib/R/site-library

³ /usr/lib/R/site-library

⁴ /usr/lib/R/library

Table 6: Supplemental Table: Reproducibility Software Environment Information

name	value
username	eschris
file name	rsi_correlations.Rmd
git	Not in git
local dir	/share/data2/RSI/Projects/RSI_RNASeq_Translation/rsi_correlations
hostname	compute-0-31.local