

# filter\_steps\_single

Afrooz Razi

2024-04-26

## Load the data

```
single<-readRDS("data/single.rds")
test_line<-readRDS("data/geuvadis_quantile.rds")
cancer<-readRDS("data/cancer_annot.rds")
potential_single_cell<-readRDS("data/potential_single_cell.rds")

seq_mean=seq(0,4.6,by=0.1)
```

## Starting numbers

```
#Total number of single-end samples:
nrow(single)
```

```
## [1] 105048
```

## Filter #1:

### Filter based on single cell

```
# number of samples removed in this step
scrNA<-which(single$external_id %in% potential_single_cell)
length(scrNA)
```

```
## [1] 1119
```

```
single<-single[-scrNA,]
```

## Filter #2:

Filter based on total read count for 75% of SNPs in each sample

```
quantile(single$read75, na.rm=T)
```

```
##      0%      25%      50%      75%     100%  
##      9.0     23.0     31.0     40.0 10468.5
```

```
low_count<-single %>% filter(read75>10)
```

```
# number of samples removed in this step  
nrow(single)- nrow(low_count)
```

```
## [1] 14517
```

## Plot filter 2

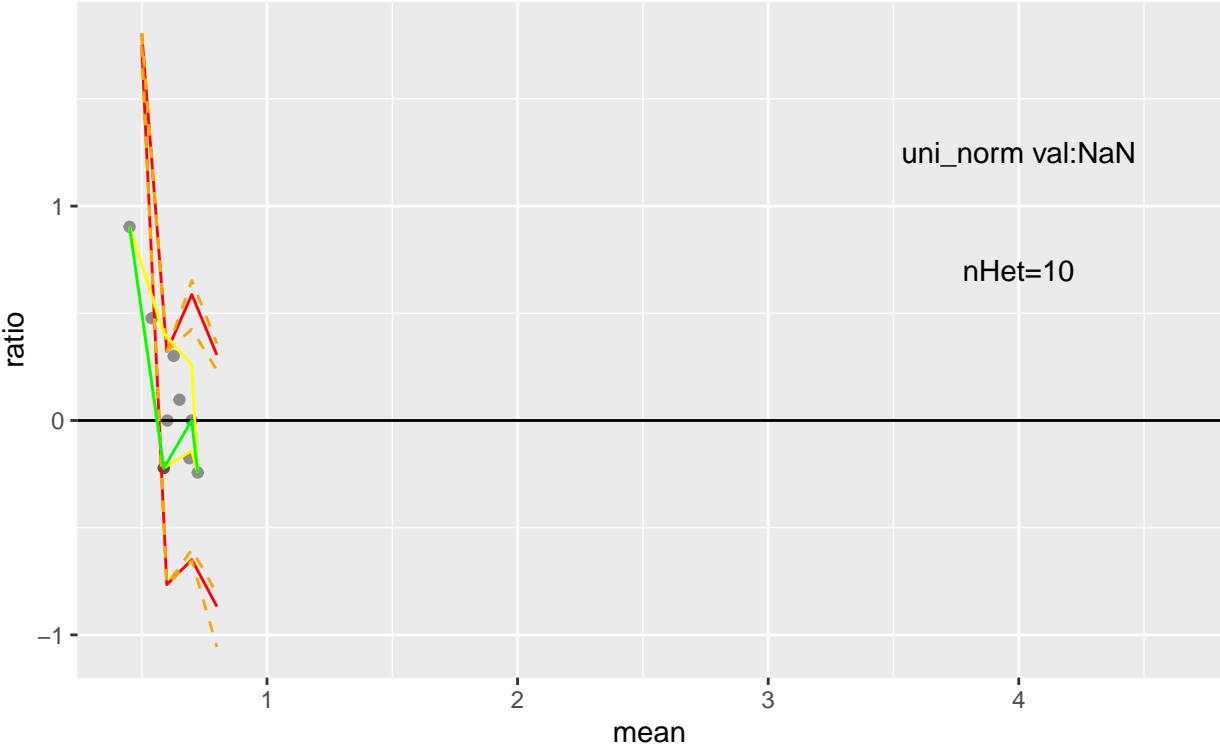
```
plot_df<-single %>% filter(read75<=10)
```

```
data_samp<-plot_df[sample(nrow(plot_df), 4),]
```

```
make_plot(data_samp)
```

```
## [1] 1  
## [1] "ase found"
```

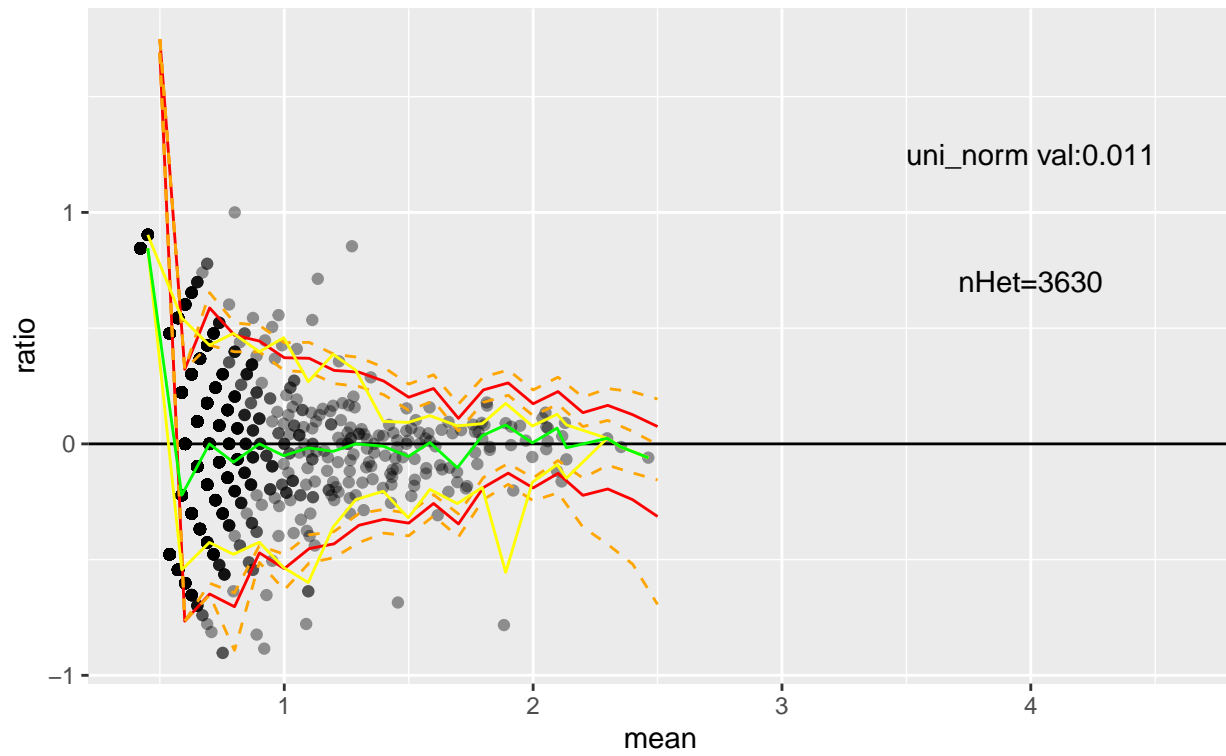
SRR9836184  
SRP216404–Old ref\_ratio:0.5



```
## [1] 2
## [1] "ase found"
```

SRR5019895

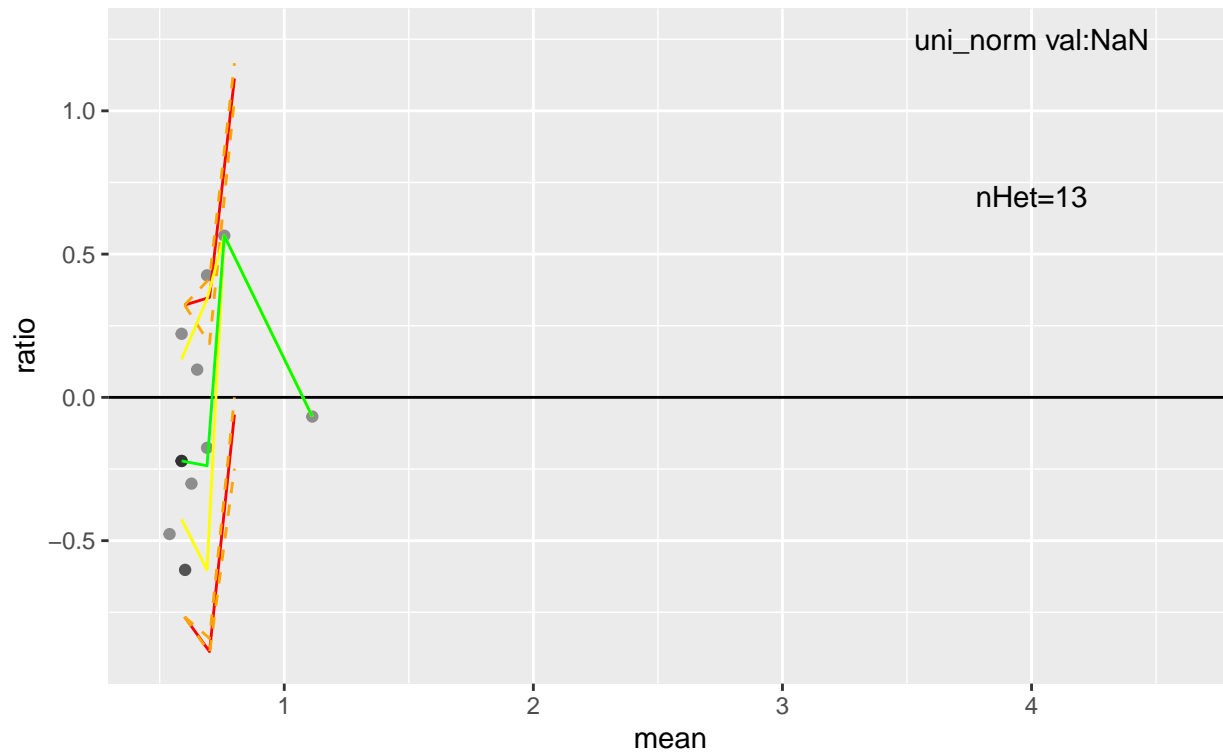
SRP093349–Old ref\_ratio:0.5



```
## [1] 3
## [1] "ase found"
```

SRR9836267

SRP216404-Old ref\_ratio:0.63



```
## [1] 4
## [1] "ase found"
```

```
## null device
##      1
```

### Filter #3:

Filter based on low number of heterozygous

```
quantile(low_count$nHet, na.rm=T)
```

```
##      0%      25%      50%      75%     100%
##       5     1377     4261     8236    247526
```

```
nhet<- low_count %>% filter(nHet>500)

# number of samples removed in this step
nrow(low_count)- nrow(nhet)
```

```
## [1] 11947
```

### Plot filter 3

```
plot_df<-low_count %>% filter(nHet<=1000)

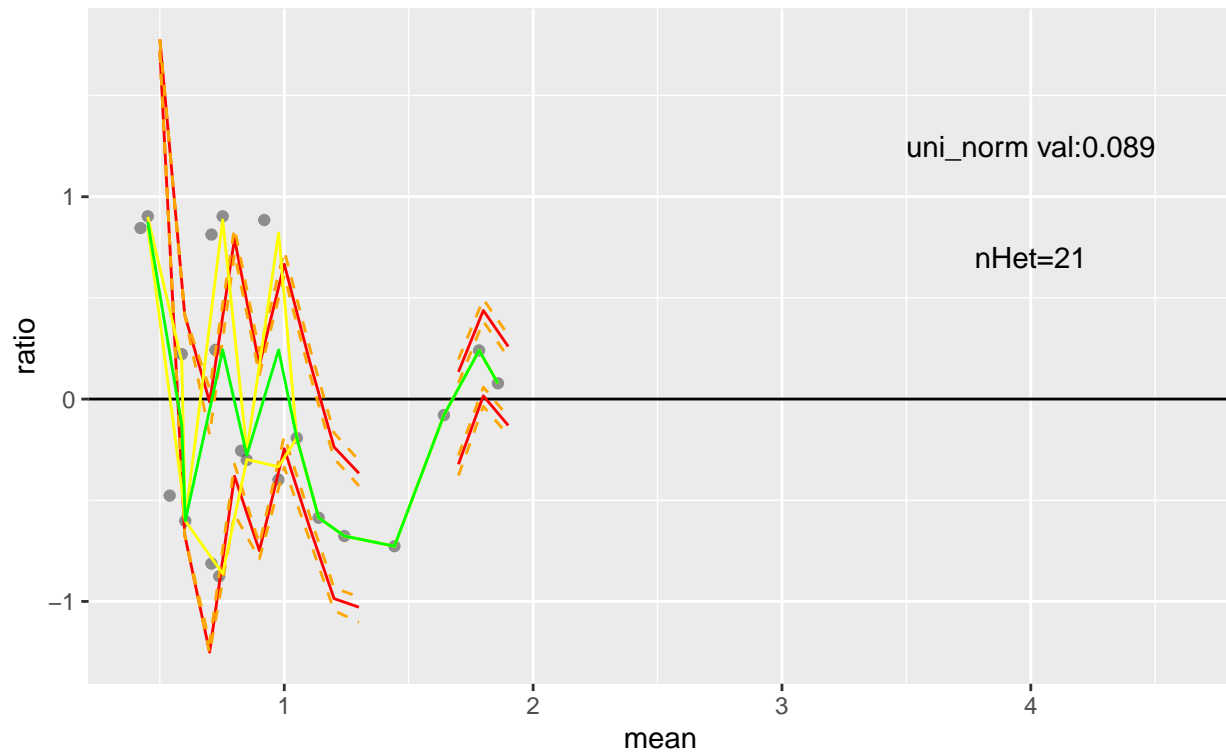
data_samp<-plot_df[sample(nrow(plot_df), 4),]

make_plot(data_samp)
```

```
## [1] 1
## [1] "ase found"
```

ERR1096185

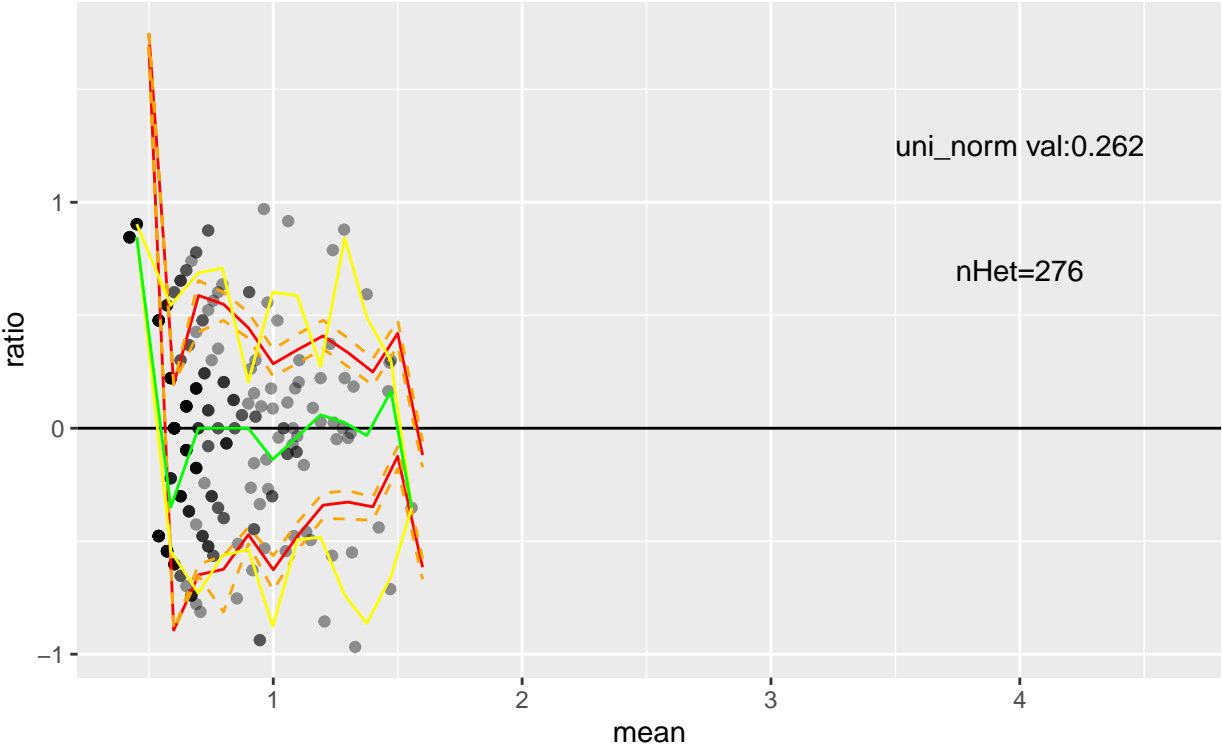
ERP012914–Old ref\_ratio:0.61



```
## [1] 2
## [1] "ase found"
```

SRR8424372

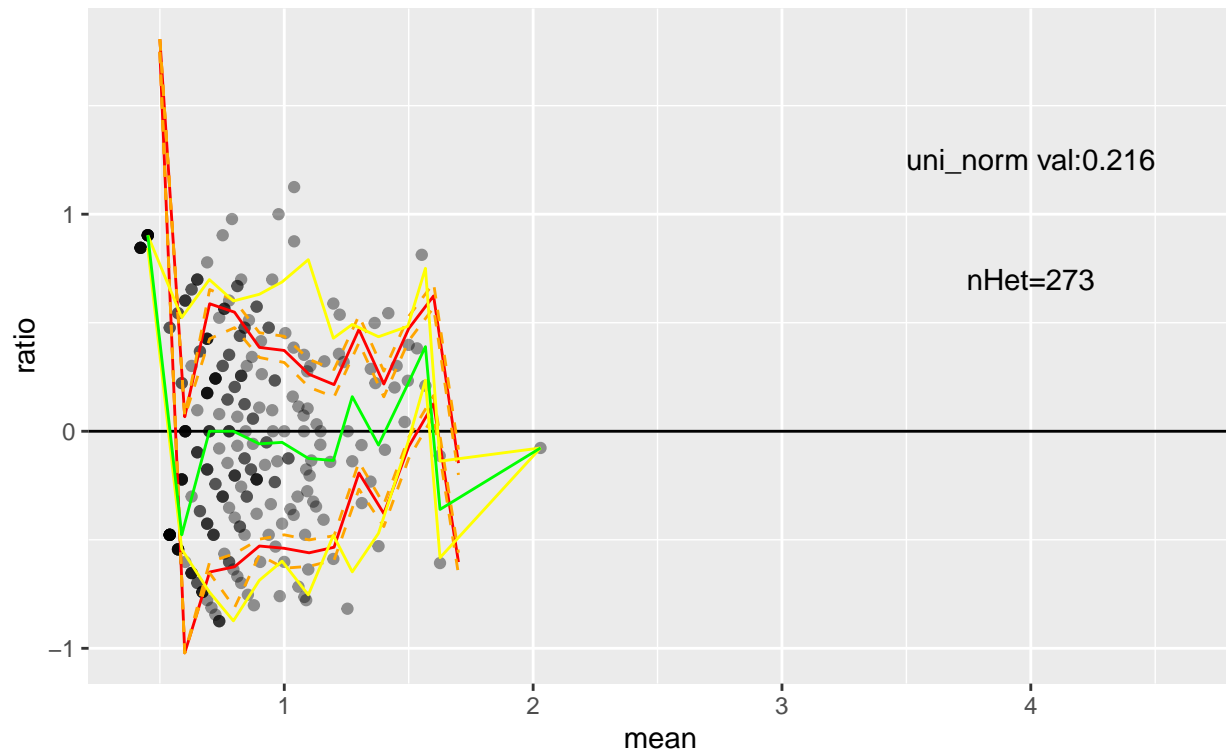
SRP178115-Old ref\_ratio:0.5



```
## [1] 3
## [1] "ase found"
```

ERR1625660

ERP016099–Old ref\_ratio:0.5



```
## [1] 4
## [1] "ase found"
```

```
## null device
##      1
```

## Filter #4:

Filter cancer samples

```
single_noncancer<-nhet[-which(nhet$sample_acc %in% cancer[,1]),]

# number of samples removed in this step
nrow(nhet)- nrow(single_noncancer)
```

```
## [1] 12920
```

Plot filter 4



```

plot_df<-nhet[which(nhet$sample_acc %in% cancer[,1]),]

data_samp<-plot_df[sample(nrow(plot_df), 4),]

make_plot(data_samp)

```

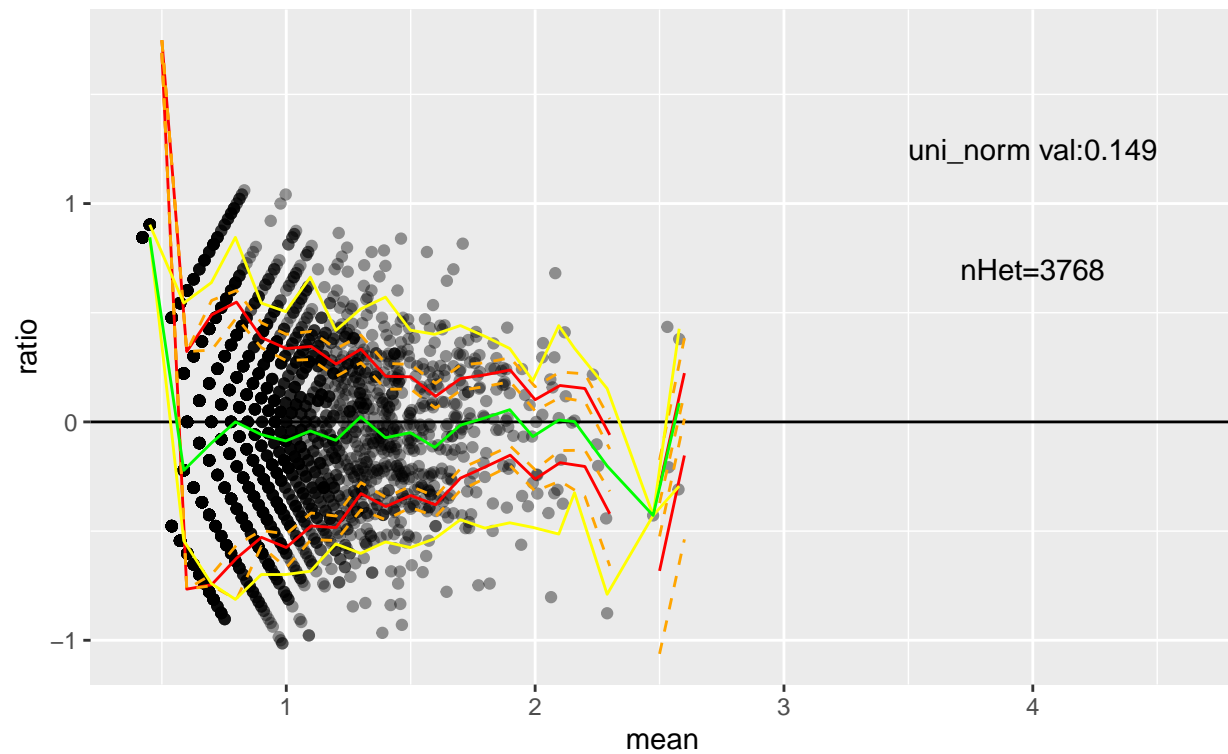
```

## [1] 1
## [1] "ase found"

```

DRR050969

DRP002866–Old ref\_ratio:0.53



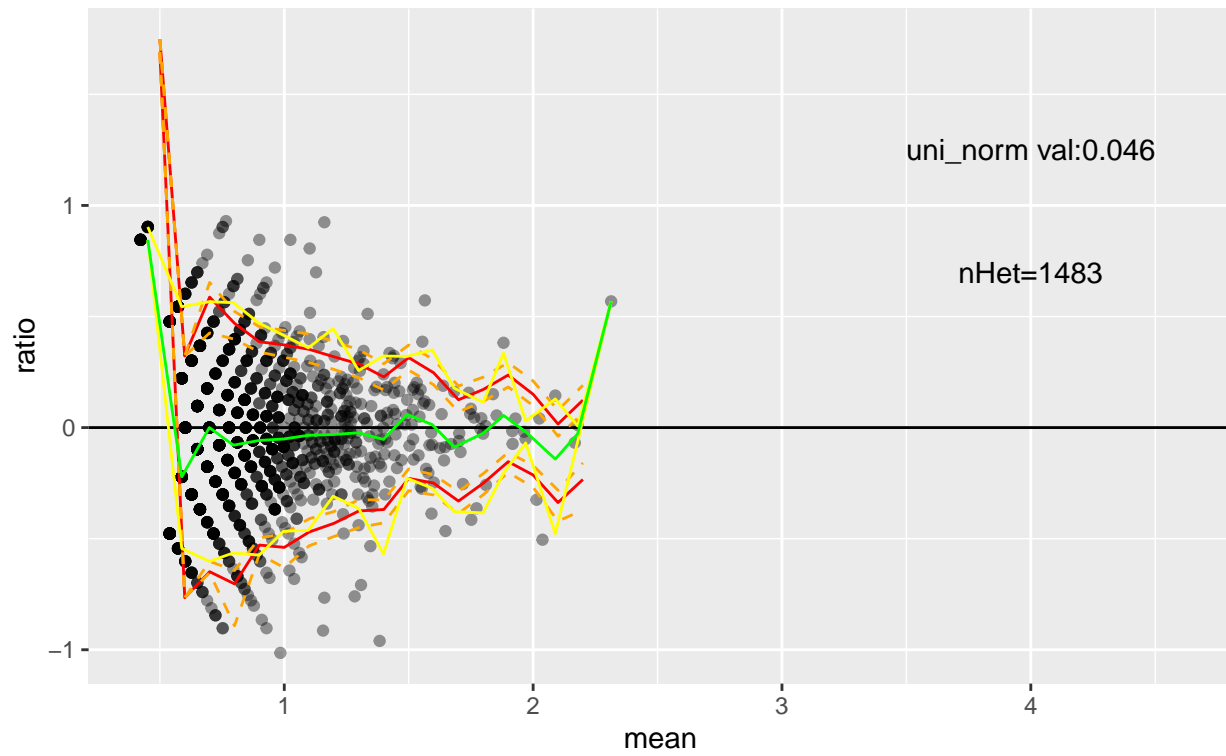
```

## [1] 2
## [1] "ase found"

```

SRR5682217

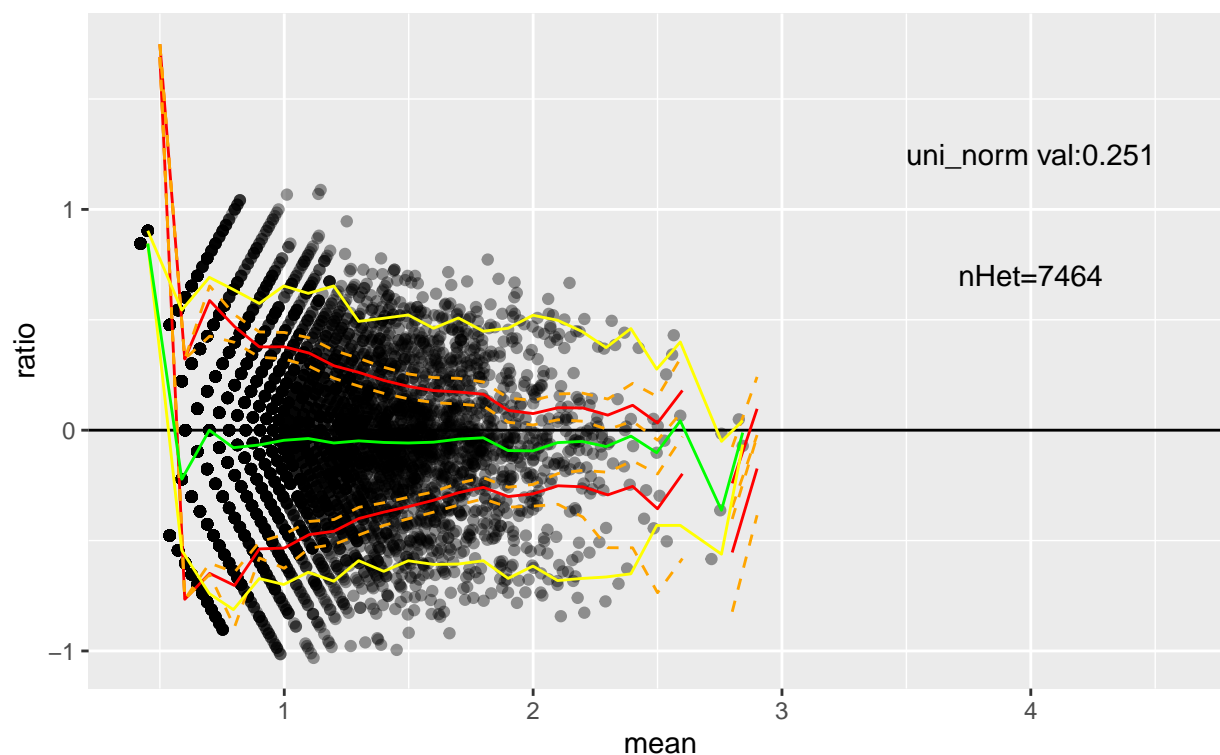
SRP103746–Old ref\_ratio:0.52



```
## [1] 3
## [1] "ase found"
```

SRR5345319

SRP101934-Old ref\_ratio:0.53



```
## [1] 4
## [1] "ase found"
```

```
## null device
##      1
```

## Filter #5:

Filter based on extreme median ref-ratio

```
quantile(single_noncancer$ref_ratio, na.rm=T)
```

```
##      0%      25%      50%      75%     100%
## 0.1764706 0.5147787 0.5263158 0.5333333 0.8034759
```

```
fold_c<- single_noncancer %>% filter(ref_ratio>0.4, ref_ratio<0.6 )
```

```
# number of samples removed in this step
nrow(single_noncancer)- nrow(fold_c)
```

```
## [1] 2359
```

## Filter #6:

Filter based on uni\_norm value

```
quantile(single$uni_norm, na.rm=T)
```

```
##           0%           25%           50%           75%           100%  
## 4.677412e-06 5.642309e-02 1.258057e-01 2.112597e-01 1.349228e+00
```

```
good<- fold_c %>% filter(uni_norm<0.185)
```

```
# number of samples removed in this step  
nrow(fold_c)- nrow(good)
```

```
## [1] 24659
```

## Plot filter 6

```
plot_df<-fold_c %>% filter(uni_norm>0.185)
```

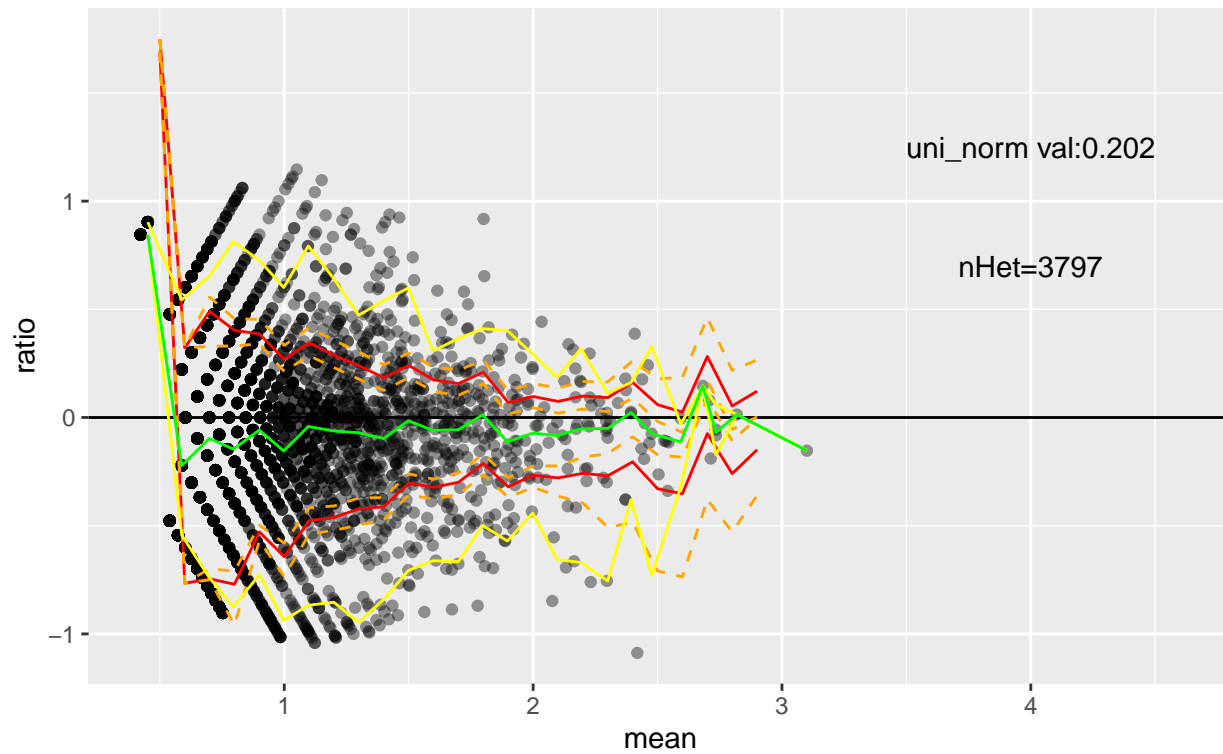
```
data_samp<-plot_df[sample(nrow(plot_df), 4),]
```

```
make_plot(data_samp)
```

```
## [1] 1  
## [1] "ase found"
```

SRR5863489

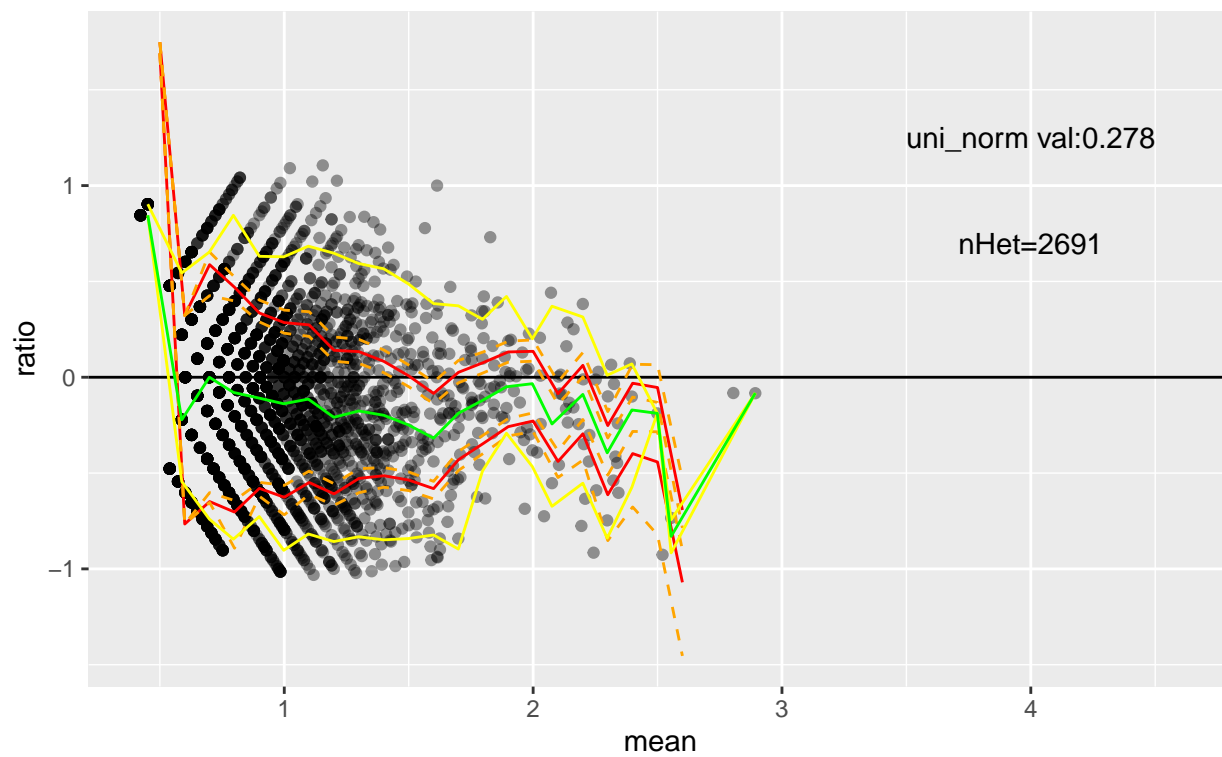
SRP113492-Old ref\_ratio:0.56



```
## [1] 2
## [1] "ase found"
```

SRR8369370

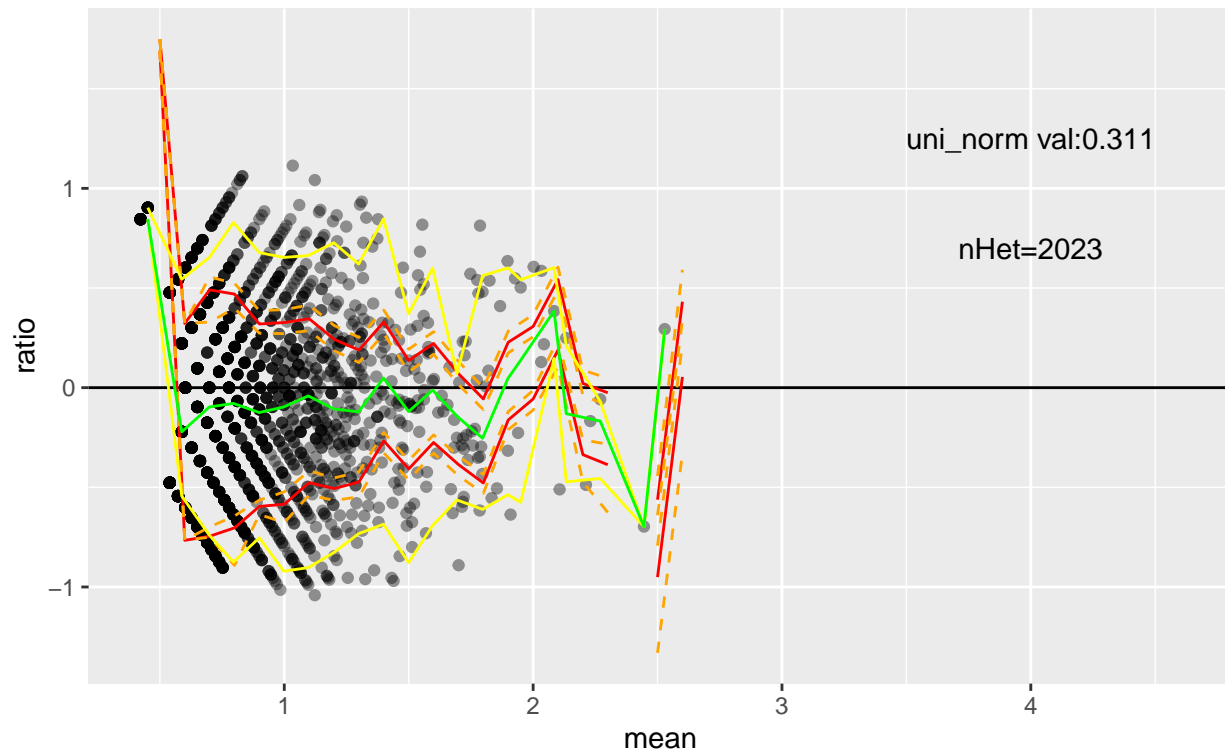
SRP174449-Old ref\_ratio:0.57



```
## [1] 3
## [1] "ase found"
```

SRR8073089

SRP166018–Old ref\_ratio:0.56



```
## [1] 4
## [1] "ase found"
```

```
## null device
##      1
```

**Final number of samples:**

```
nrow(good)
```

```
## [1] 37527
```

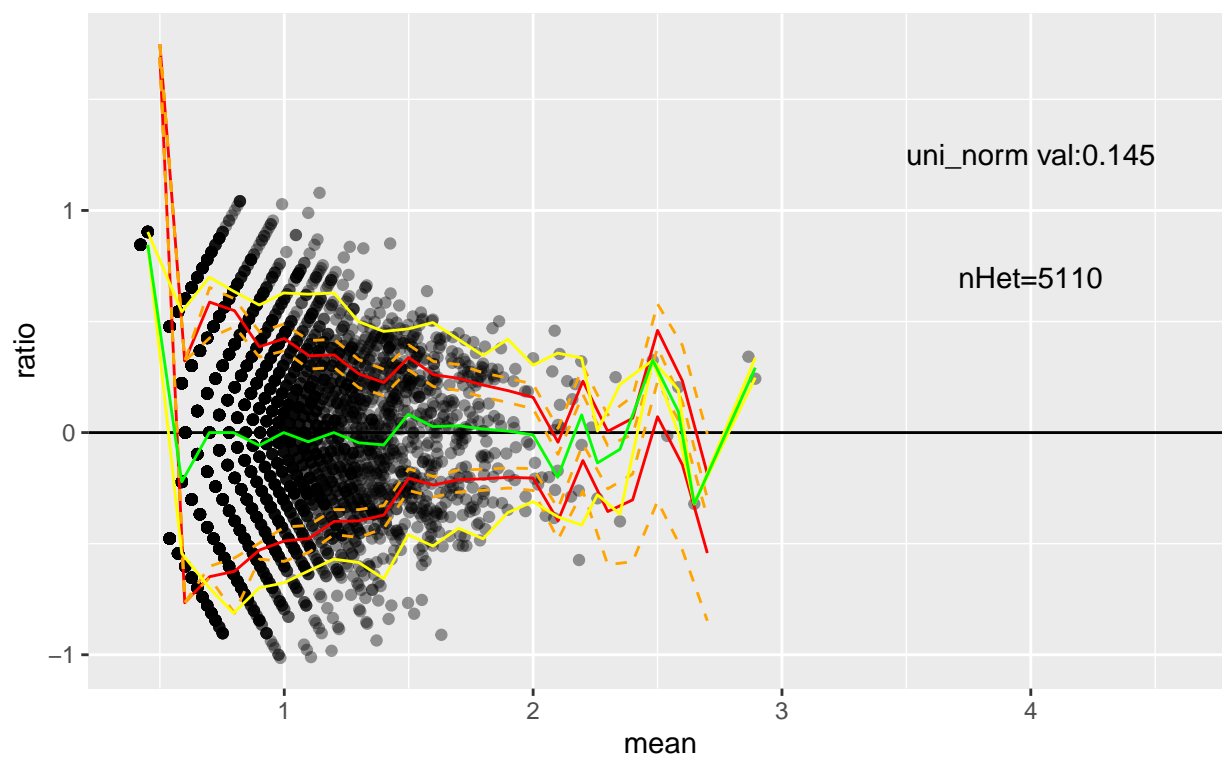
**Plot Good samples**

```
data_samp<-good[sample(nrow(good), 4),]
make_plot(data_samp)
```

```
## [1] 1
## [1] "ase found"
```

SRR8063570

SRP165866-Old ref\_ratio:0.5

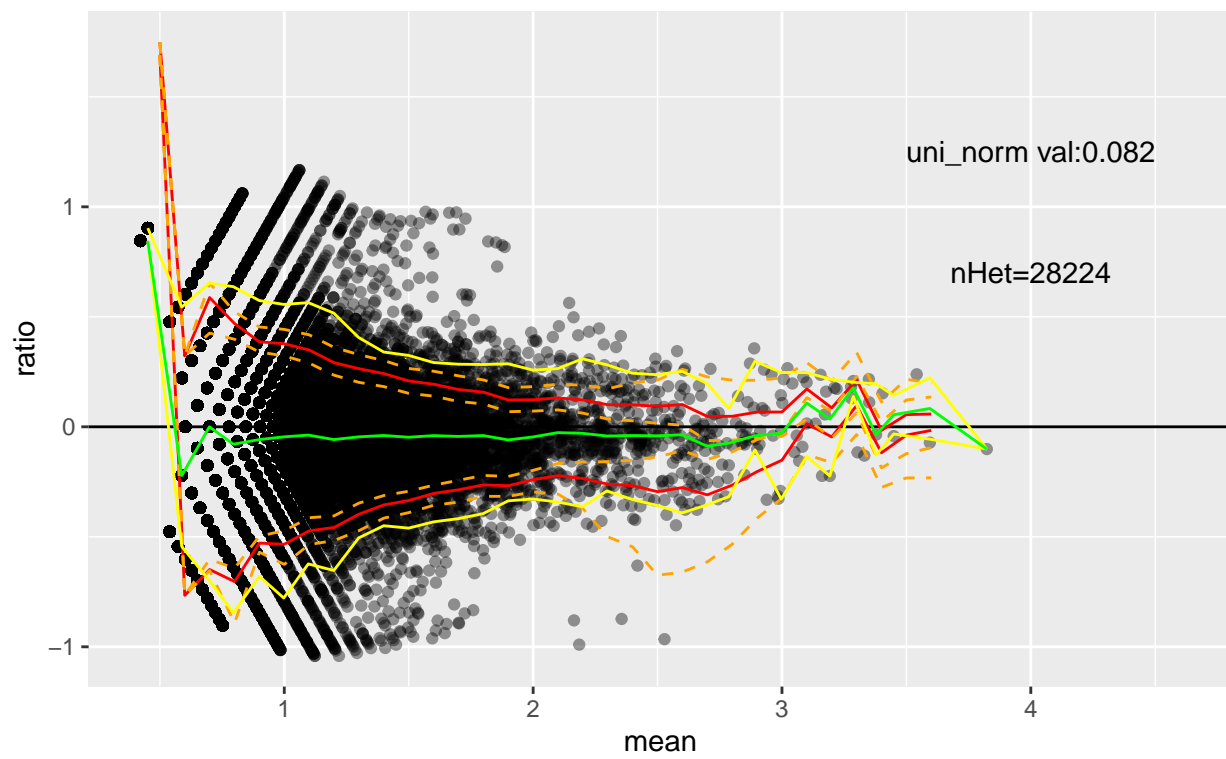


```
## [1] 2
## [1] "ase found"
```



ERR655291

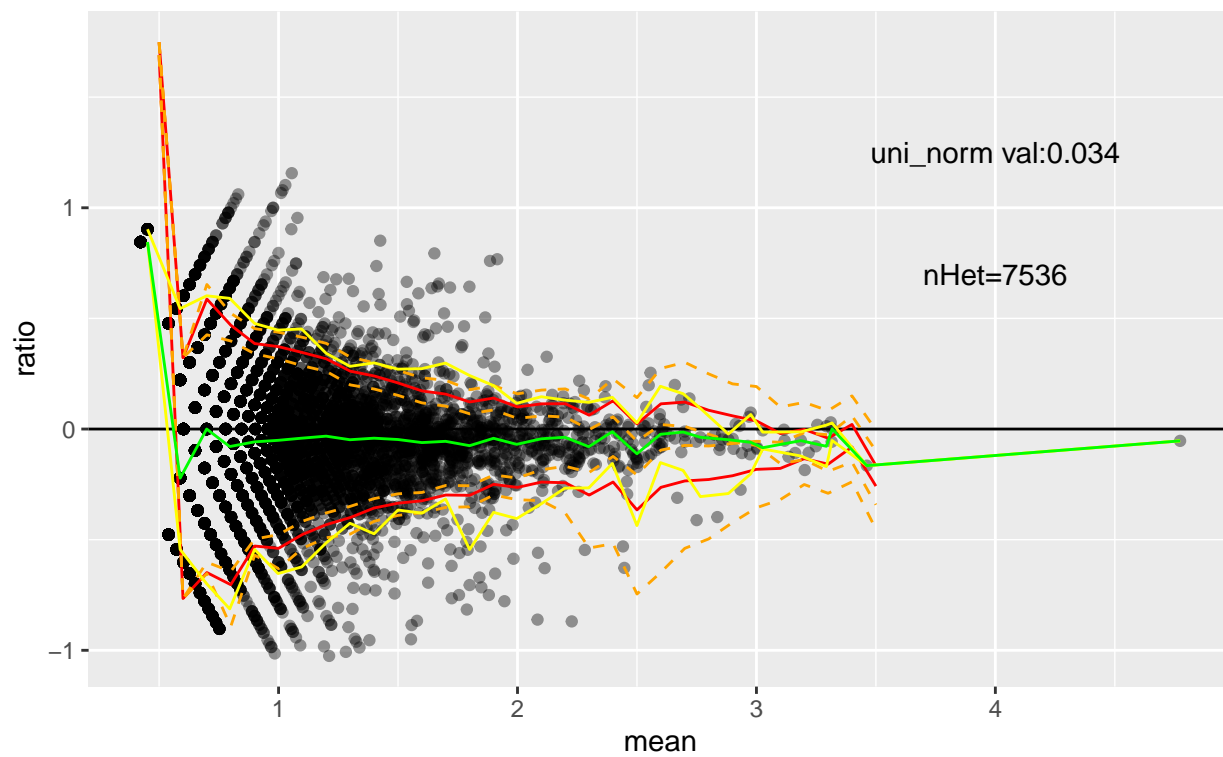
ERP008445-Old ref\_ratio:0.52



```
## [1] 3
## [1] "ase found"
```

SRR6900791

SRP136499-Old ref\_ratio:0.53



```
## [1] 4
## [1] "ase found"
```

```
## null device
##      1
```