paired_filter

Afrooz Razi

2024-04-26

Load the data

```
paired<-readRDS("data/paired.rds")
battle<-read_tsv("data/battle.tsv")
```

```
## Rows: 65405 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (12): study, sample, tissue, organ, biopsy, cell, disease, source, des, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

paired$frag_mode<-paired$bc_frag.mode_length
paired$overlap<- paired$bc_frag.mode_length-(paired$avg_len*2)

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

Starting numbers

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

```
## [1] 117703
```

Filter #1:

Filter based on single cell

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

```
## [1] 3175
```

```
paired<-paired[-scRNA,]
```

Filter #2:

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

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

```
##     0%    25%    50%    75%   100%
##      8     27     41     58 115324
```

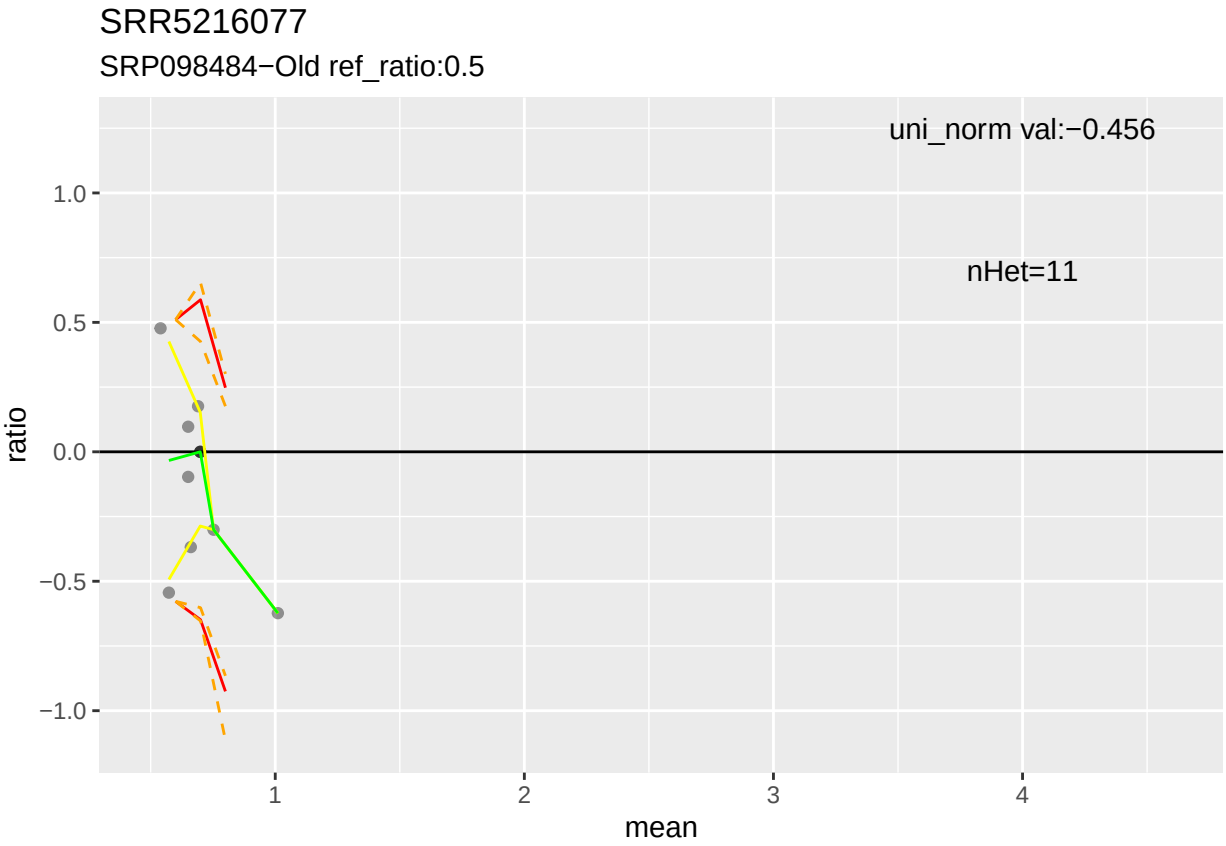
```
low_count<-paired %>% filter(read75>10)

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

```
## [1] 1768
```

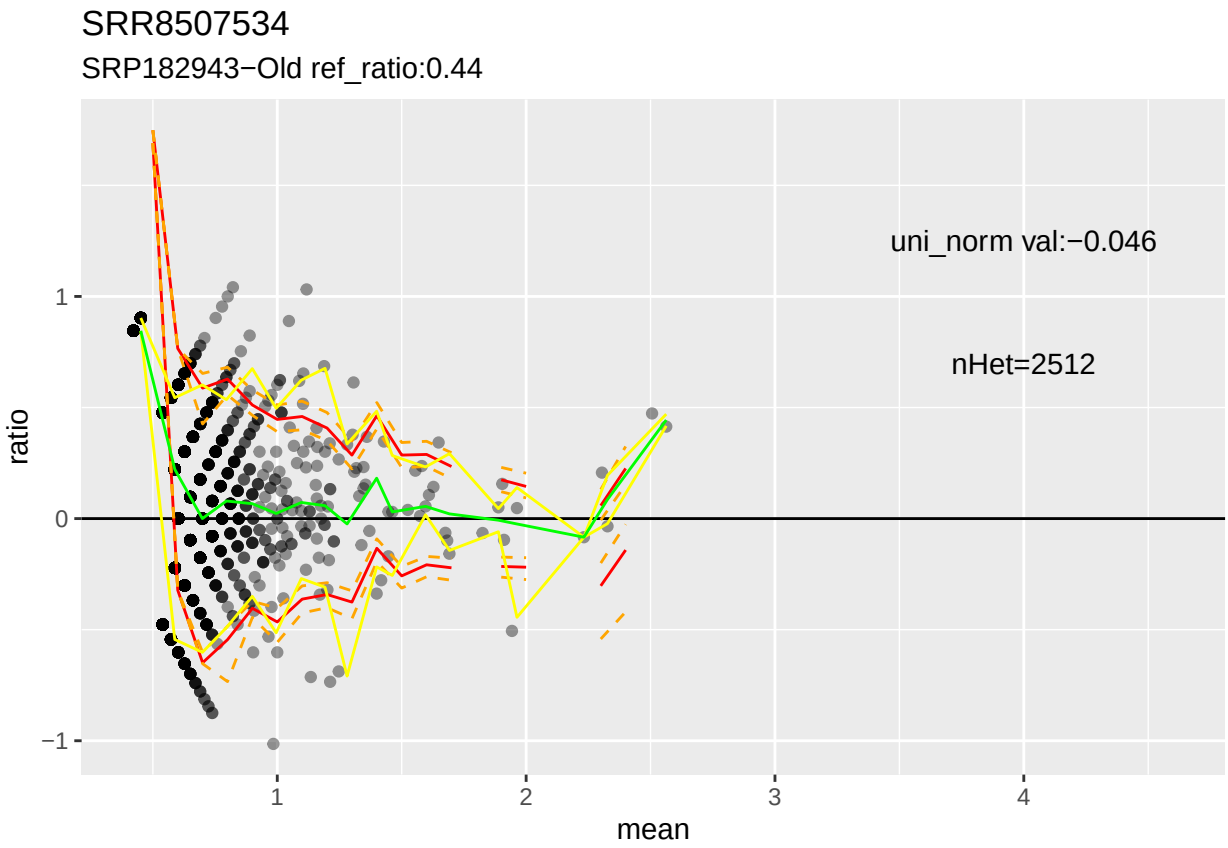
Plot filter 2

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


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

make_plot(data_samp)
```

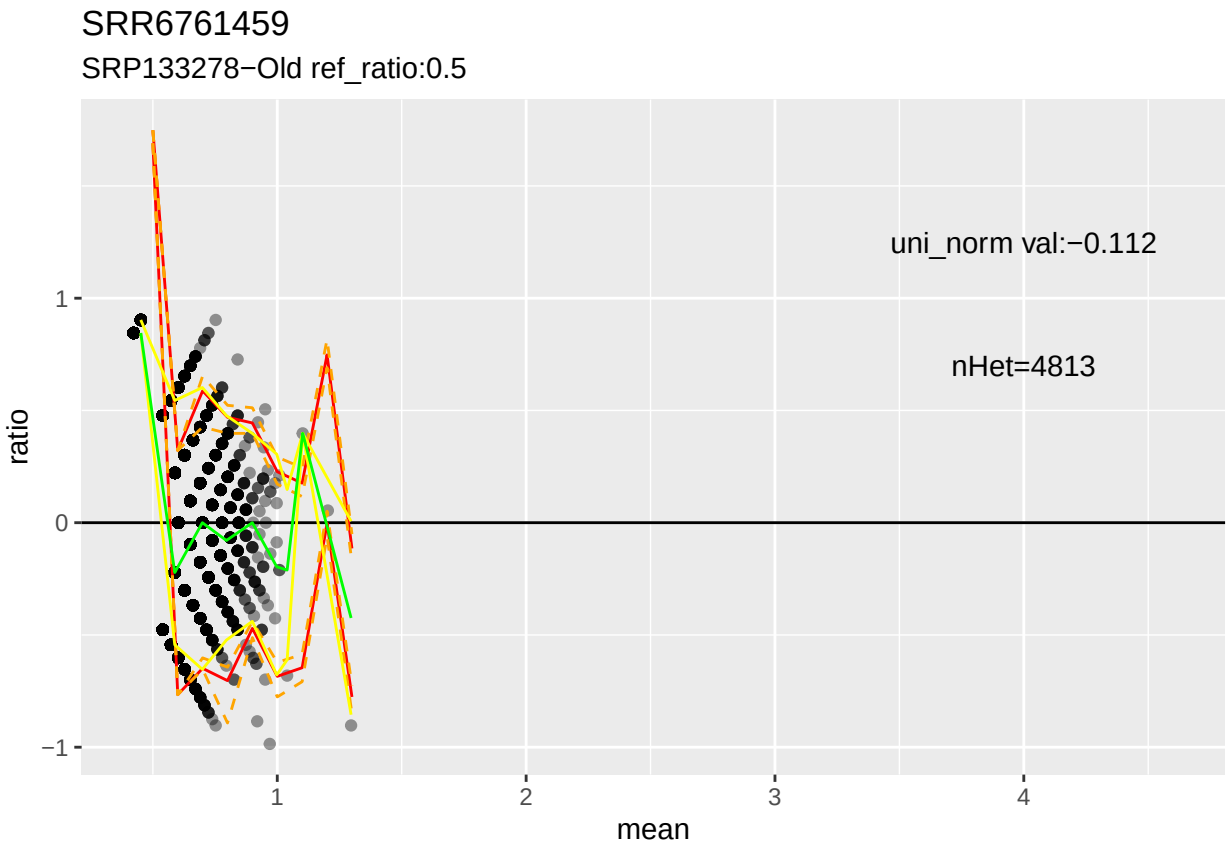
```
## [1] 1
## [1] "ase found"
```




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




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




```
## [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%
##      1   5820  11827  19783 362247
```

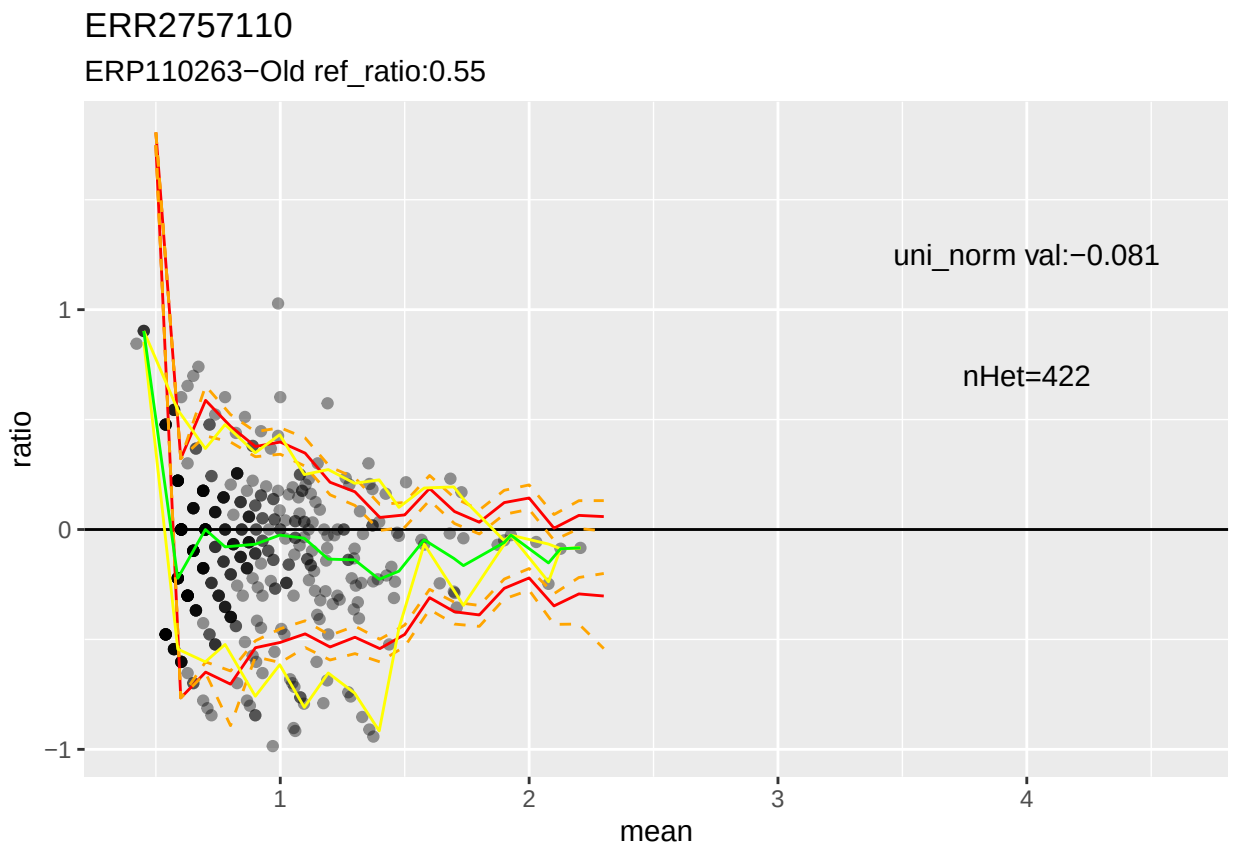
```
nhet<- low_count %>% filter(nHet>1000)

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

```
## [1] 6859
```

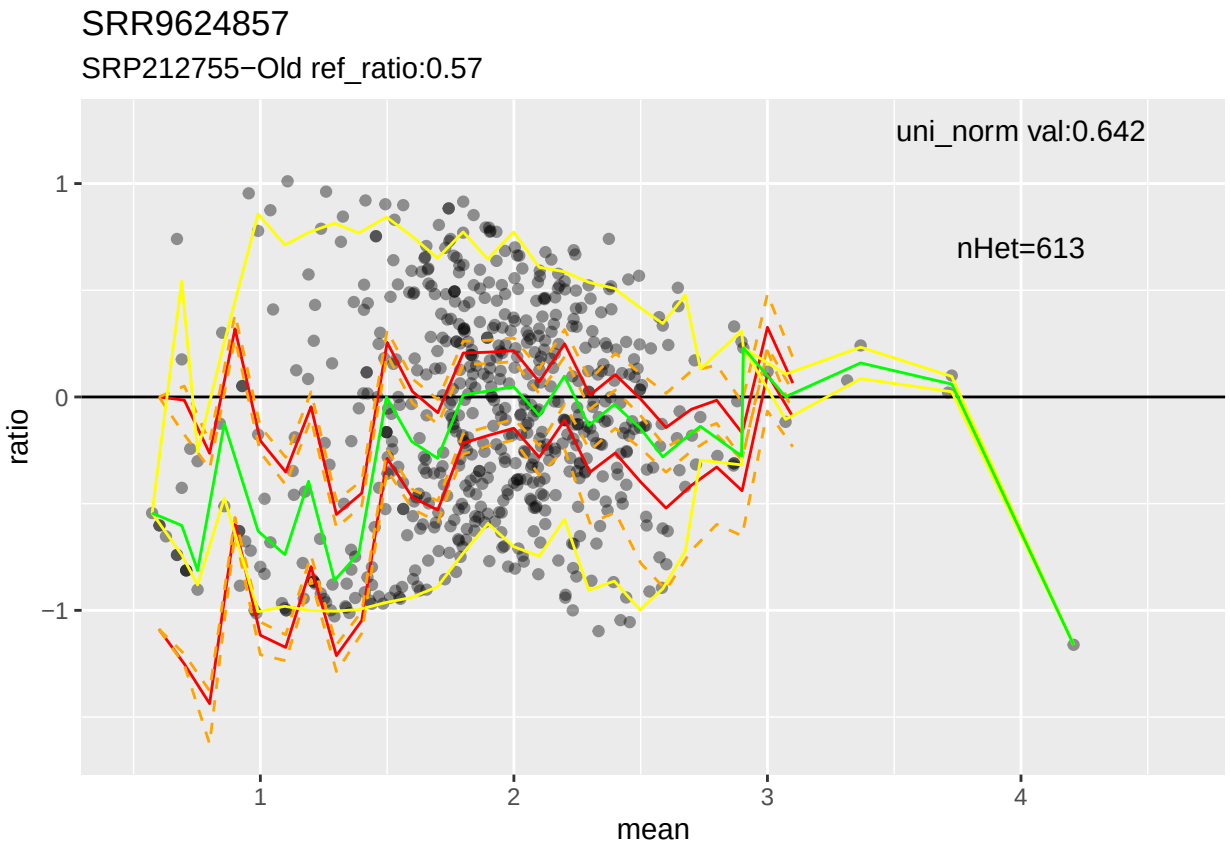
Plot filter 3

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


data_samp<-plot_df[sample(nrow(plot_df), 3),]

make_plot(data_samp)
```

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




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




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

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

Filter #4:

Filter cancer samples

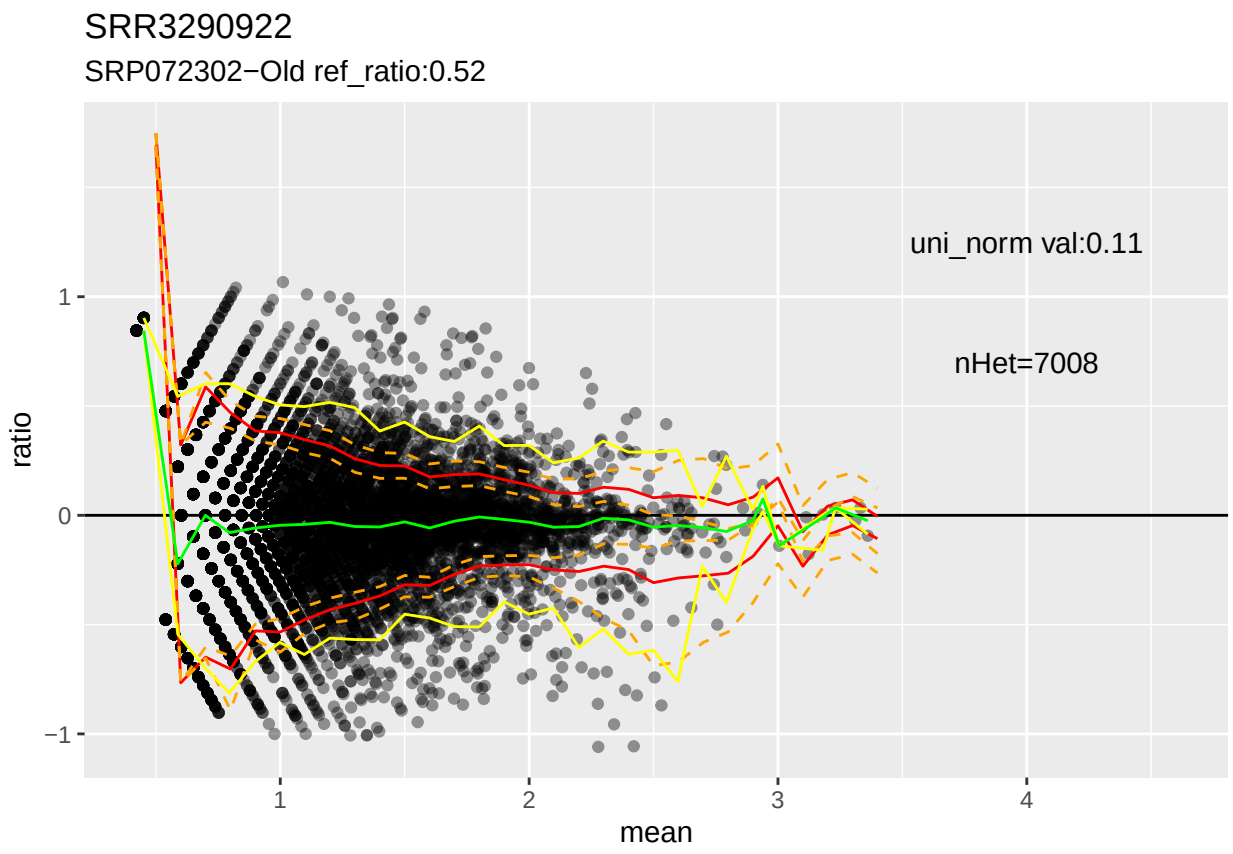
```
paired_noncancer<-nhet[-which(nhet$sample_acc %in% cancer[,1]),]

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

```
## [1] 17595
```

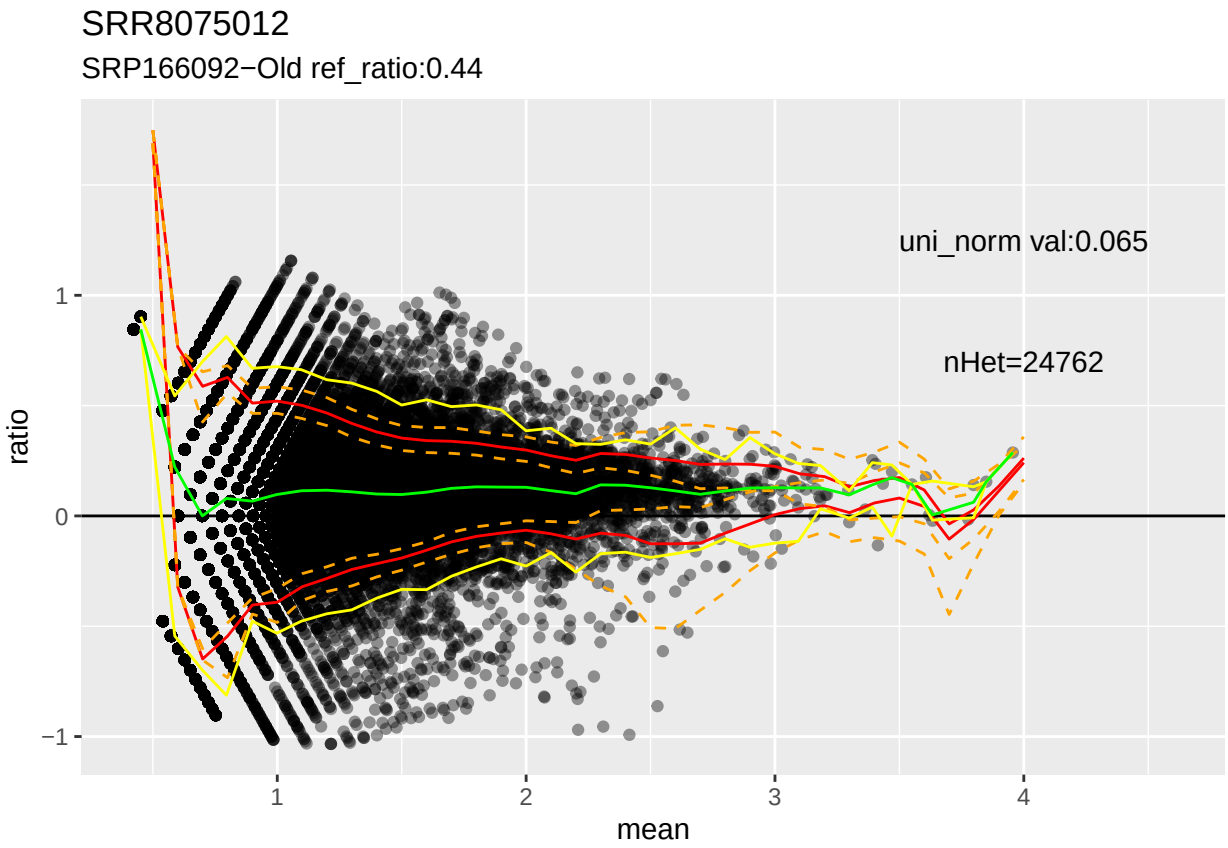
Plot filter 4

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


data_samp<-plot_df[sample(nrow(plot_df), 3),]

make_plot(data_samp)
```

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




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




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

## null device
##           1
```

Filter #5:

Filter based on extreme median ref-ratio

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

```
##        0%       25%       50%       75%      100%
## 0.2558140 0.4414414 0.5000000 0.5200000 0.8333333
```

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

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

```
## [1] 15800
```

Filter #6:

Filter based on uni_norm value

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

```
##          0%         25%         50%         75%        100%
## -0.20252765 -0.04123340  0.02755538  0.15332039  0.76824480
```

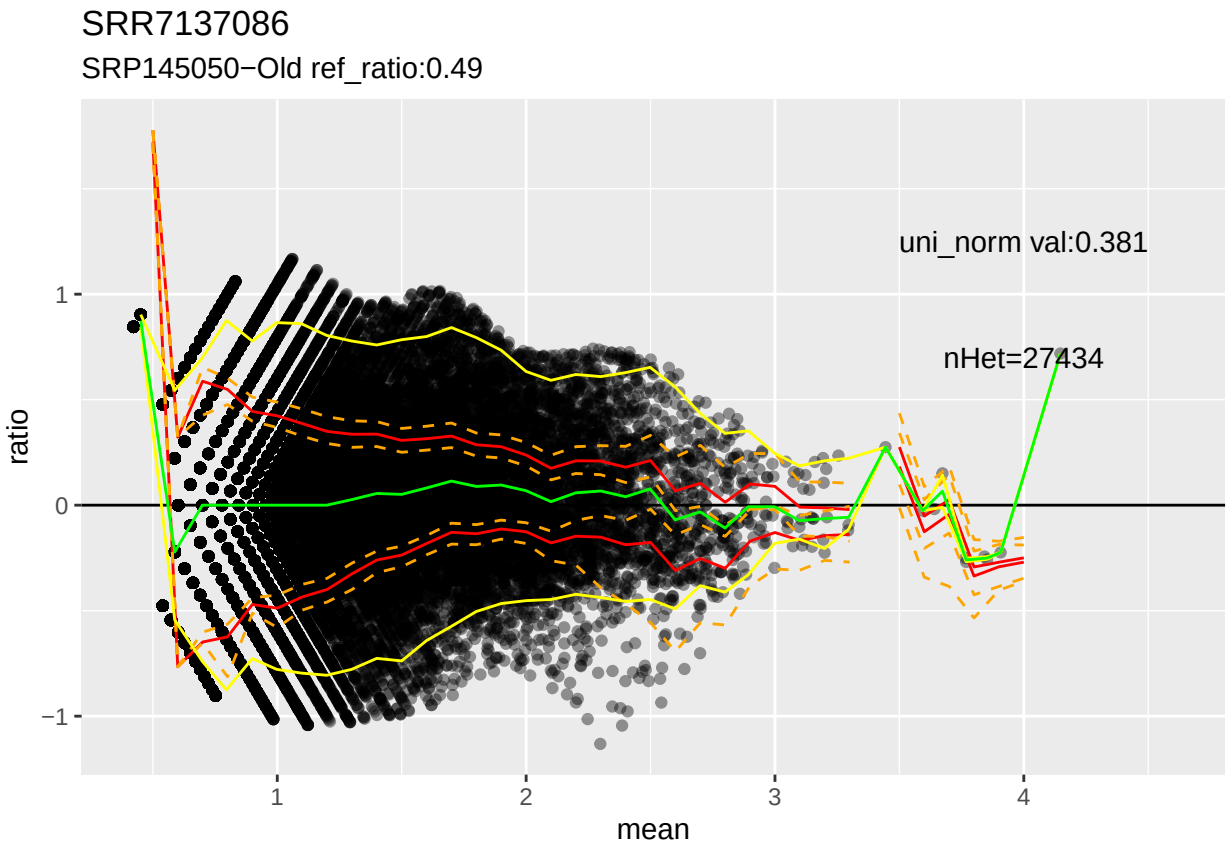
```
good<- fold_c %>%  filter(uni_norm<0.185)

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

```
## [1] 14863
```

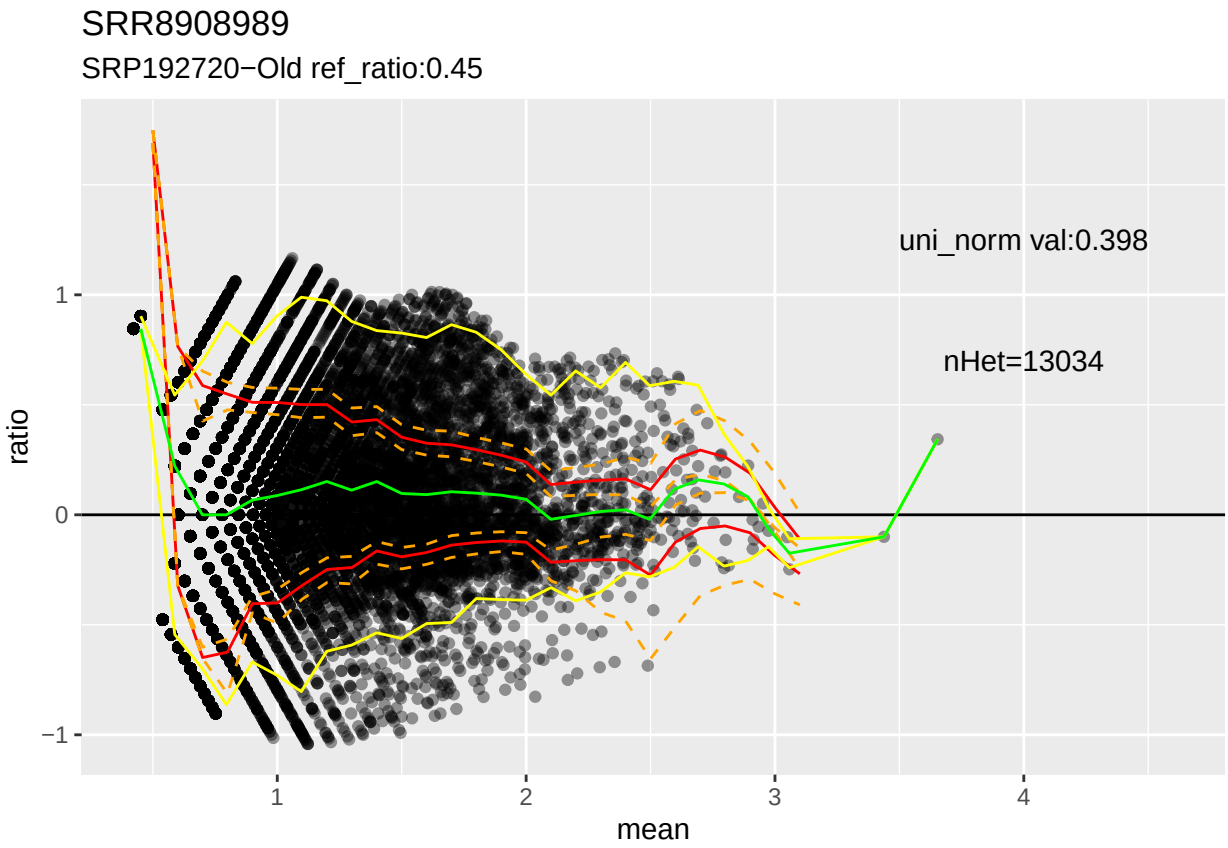
Plot filter 6

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


data_samp<-plot_df[sample(nrow(plot_df), 3),]

make_plot(data_samp)
```

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




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




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

## null device
##           1
```

Final number of samples:

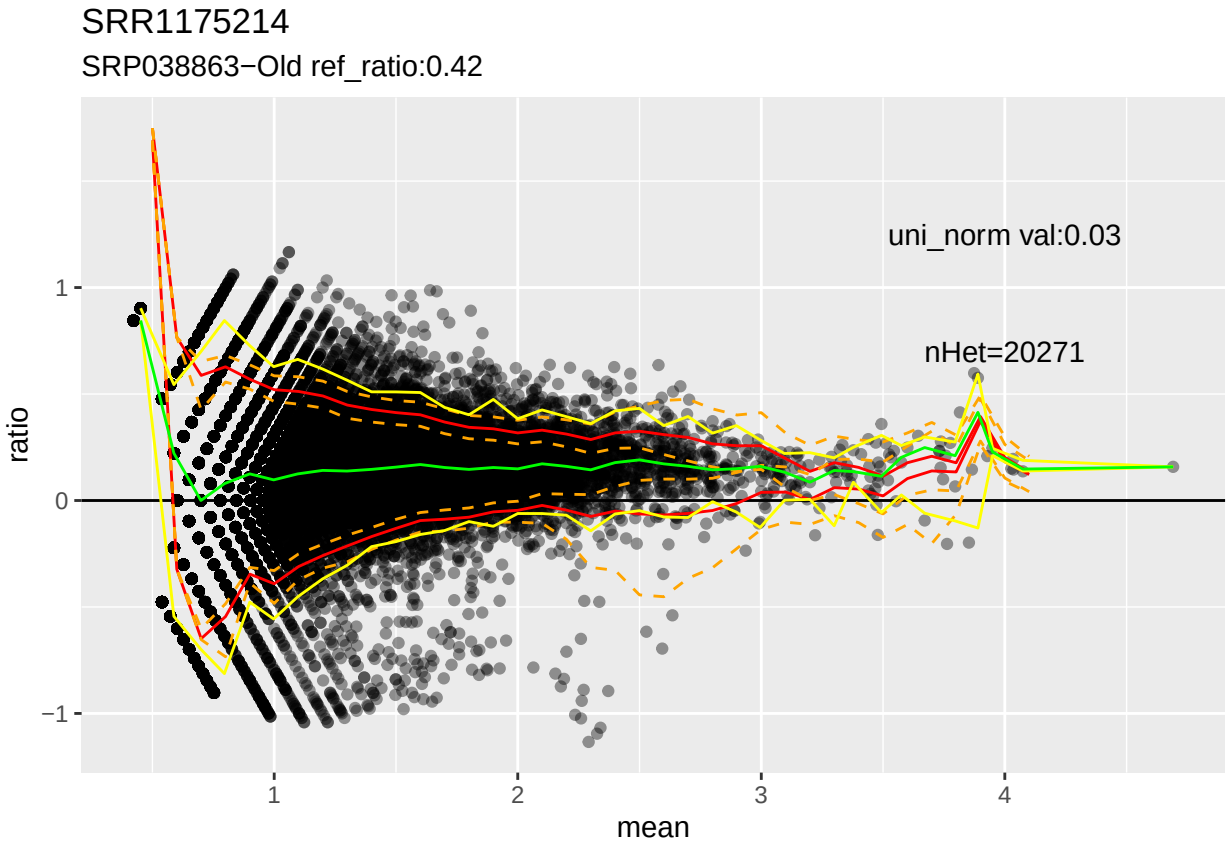
```
nrow(good)
```

```
## [1] 57643
```

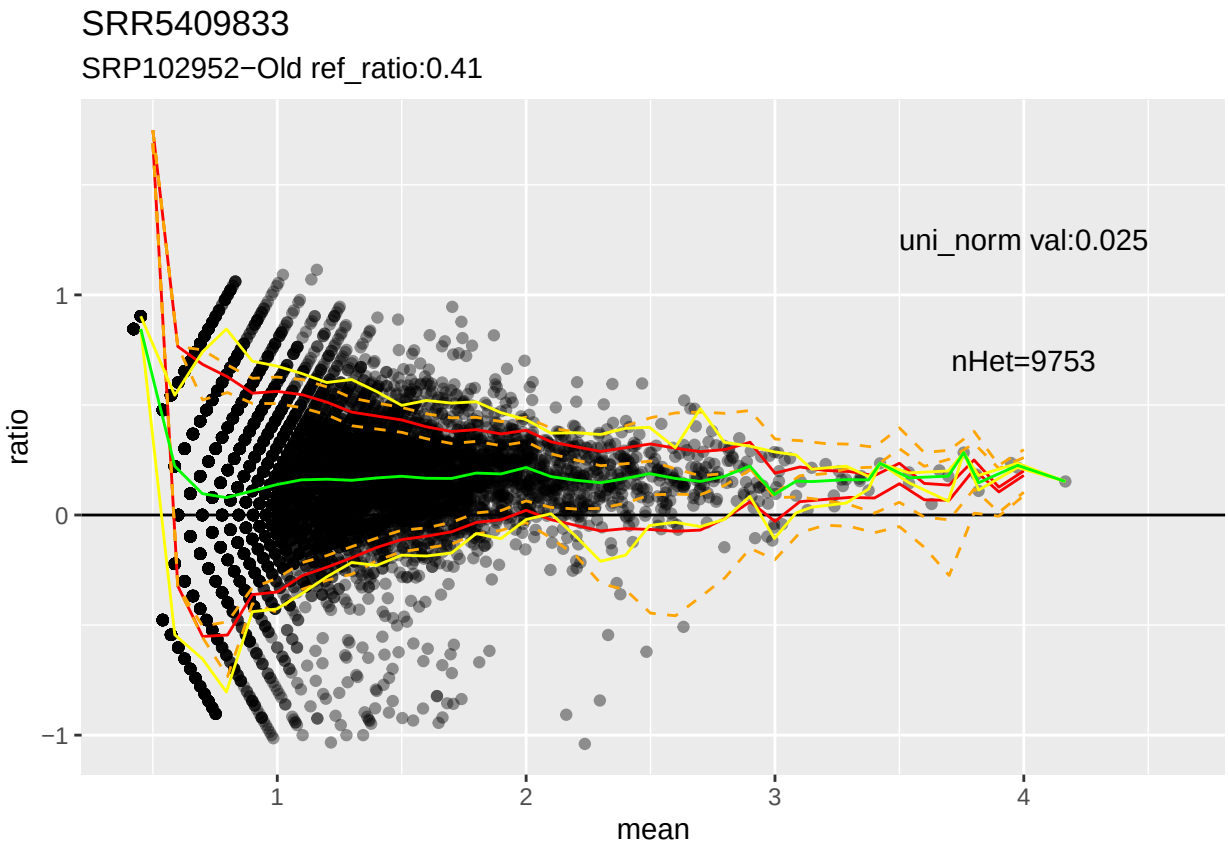
Plot Good samples

```
data_samp<-good[sample(nrow(good), 3),]

make_plot(data_samp)
```

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




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




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

## null device
##           1
```