Gridsemble on Platinum Spike Dataset - Models on All Data

Jenna Landy

2023-12-21

```
source('PAPER_metrics_helpers.R')
load("PAPER_platinum_data.RData")
```

```
remove.packages('gridsemblefdr')
library(devtools)
devtools::install_github('jennalandy/gridsemblefdr')
```

```
rlang (1.1.2  -> 1.1.3 ) [CRAN]
glue  (1.6.2  -> 1.7.0 ) [CRAN]
Rcpp  (1.0.11 -> 1.0.12) [CRAN]

  There are binary versions available but the source versions are later:
      binary source needs_compilation
rlang  1.1.2  1.1.3              TRUE
glue   1.6.2  1.7.0              TRUE
Rcpp  1.0.11 1.0.12              TRUE

-- R CMD build -------------------------------------------------------------
* checking for file ‘/private/var/folders/0w/yyrmpks1285dstjz0t26td040000gn/T/RtmprhpBKS/rem
* preparing ‘gridsemblefdr’:
* checking DESCRIPTION meta-information ... OK
* checking for LF line-endings in source and make files and shell scripts
* checking for empty or unneeded directories
Omitted ‘LazyData’ from DESCRIPTION
* building ‘gridsemblefdr_0.99.0.tar.gz’
```

```r
library(gridsemblefdr)
```

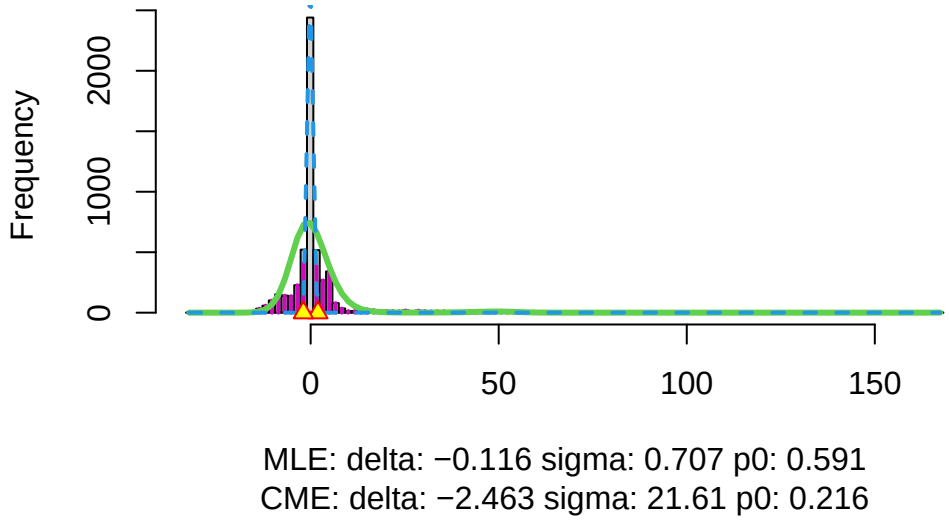
```r
library(locfdr)
library(fdrtool)
library(qvalue)
library(tidyverse)
library(ggplot2)
library(ggthemes)
library(ggdist)

color_list = list(
  "gridsemble" = "#E69F00",
  "locfdr" = "#D55E00",
  "fdrtool" = "#009E73",
  "qvalue" = "#0072B2"
)
```

Run Methods

Benchmarks

`locfdr` fails with default `pct0 = 0` because there are a few extreme outlier. We use `pct0 = 0.001` instead.

```r
locfdr_res <- locfdr(platinum_data$statistics, pct0 = 0.001)
```




```
fdrtool_res <- fdrtool(platinum_data$statistics, plot = 0)
```

```
Step 1... determine cutoff point
Step 2... estimate parameters of null distribution and eta0
Step 3... compute p-values and estimate empirical PDF/CDF
Step 4... compute q-values and local fdr
```

```
qvalue_res <- qvalue(p_from_t(platinum_data$statistics, df = 4))
```

Note that the `df` mentioned in the warning of `locfdr` refers to the degrees of freedom for fitting the marginal distribution f, NOT the degrees of freedom of our test-statistics.

Gridsemble

`gridsemble` takes in test statistics and, if known, the degrees of freedom for each test. In our case, there are 3 samples in each condition, so `df = (3-1) + (3-1) = 4`.

```
set.seed(321)

fdrtool_grid = build_fdrtool_grid(
  platinum_data$statistics
)
```

```r
nrow(fdrtool_grid)
```

```
[1] 22
```

```r
locfdr_grid = build_locfdr_grid(
  platinum_data$statistics
)
nrow(locfdr_grid)
```

```
[1] 128
```

```r
qvalue_grid = build_qvalue_grid(
  platinum_data$statistics
)
nrow(qvalue_grid)
```

```
[1] 120
```

```r
gridsemble_res <- gridsemble(
  platinum_data$statistics,
  df = 4,
  locfdr_grid = locfdr_grid,
  fdrtool_grid = fdrtool_grid,
  qvalue_grid = qvalue_grid
)
```

```
Warning in locfdr::locfdr(test_statistics, plot = 0): f(z) misfit = 51.9.
Rerun with increased df

Warning in locfdr::locfdr(test_statistics, plot = 0): CM estimation failed,
middle of histogram non-normal

Fitting working model

Running grid search in parallel

Ensembling
```

Evaluate Methods

pi0 estimates

```
list(
  'true' = mean(1 - platinum_data$fold_change$DE),
  'gridsemble' = gridsemble_res$pi0,
  'locfdr' = unlist(locfdr_res$fp0['mlest','p0']),
  'fdrtool' = unname(fdrtool_res$param[,'eta0']),
  'qvalue' = qvalue_res$pi0
)
```

```
$true
[1] 0.6379888

$gridsemble
[1] 0.7659173

$locfdr
[1] 0.5906014

$fdrtool
[1] 0.7181736

$qvalue
[1] 1
```

fdr metrics

```
how = "symmetric"
platinum_data$Fdr = get_true_Fdr(
  platinum_data$statistics,
  platinum_data$fold_change$DE,
  how = how
)

fdr_metrics = rbind(
  method_metrics(
    'gridsemble',
    estimated_fdr = gridsemble_res$fdr,
```

```
    test_statistics = platinum_data$statistics,
    hypothesis_labels = platinum_data$fold_change$DE,
    true_Fdr = platinum_data$Fdr,
    how_Fdr = how
  ),
  method_metrics(
    'locfdr',
    estimated_fdr = locfdr_res$fdr,
    test_statistics = platinum_data$statistics,
    hypothesis_labels = platinum_data$fold_change$DE,
    true_Fdr = platinum_data$Fdr,
    how_Fdr = how
  ),
  method_metrics(
    'fdrtool',
    estimated_fdr = fdrtool_res$lfdr,
    test_statistics = platinum_data$statistics,
    hypothesis_labels = platinum_data$fold_change$DE,
    true_Fdr = platinum_data$Fdr,
    how_Fdr = how
  ),
  method_metrics(
    'qvalue',
    estimated_fdr = qvalue_res$lfdr,
    test_statistics = platinum_data$statistics,
    hypothesis_labels = platinum_data$fold_change$DE,
    true_Fdr = platinum_data$Fdr,
    how_Fdr = how
  )
)

fdr_metrics = data.frame(fdr_metrics)
fdr_metrics
```

```
     method               roc                pr             brier
1 gridsemble 0.943782371672444  0.96216019680766 0.111726473245542
2     locfdr 0.942796882469808  0.96175016571658 0.135006128560026
3    fdrtool  0.84884290790109 0.856320589583218 0.131967876218307
4     qvalue 0.898683421859948 0.905055549355762 0.103904606251393
             Fdr.MSE
1  0.0138399008737251
```

```
2  0.0289788097388825
3 0.00685260288339469
4 0.00470635031917831
```

Classification Metrics

0.2 cutoff

```
cutoff = 0.2

classification_metrics_cutoff0.2 = rbind(
  classification_metrics(
    method = 'gridsemble',
    fdr = gridsemble_res$fdr,
    pi0 = gridsemble_res$pi0,
    test_statistics = platinum_data$statistics,
    truth = platinum_data$fold_change$DE,
    cutoff = cutoff
  ),
  classification_metrics(
    'locfdr',
    locfdr_res$fdr,
    locfdr_res$fp0['mlest','p0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = cutoff
  ),
  classification_metrics(
    'fdrtool',
    fdrtool_res$lfdr,
    fdrtool_res$param[1,'eta0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = cutoff
  ),
  classification_metrics(
    'qvalue',
    qvalue_res$lfdr,
    qvalue_res$pi0,
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = cutoff
```

```
  )
)

classification_metrics_cutoff0.2
```

```
     method       cutoff global_FDR sensitivity specificity prop_pred_T TP
[1,] "gridsemble" 0.2    0.02079395 0.2664609   0.9967893   0.09851024  518
[2,] "locfdr"     0.2    0.1755545  0.8986626   0.8914186   0.3945996   1747
[3,] "fdrtool"    0.2    0.1        0.7083333   0.9553415   0.2849162   1377
[4,] "qvalue"     0.2    0.08148148 0.6378601   0.9678926   0.2513966   1240
     FP  TN   FN   accuracy  precision f1
[1,] 11  3415 1426 0.7324022 0.979206  0.4189244
[2,] 372 3054 197  0.894041  0.8244455 0.8599557
[3,] 153 3273 567  0.8659218 0.9       0.7927461
[4,] 110 3316 704  0.8484171 0.9185185 0.752884
```

cutoff based on $\hat{\pi}_0$

```
gridsemble_cutoff <- quantile(gridsemble_res$fdr, 1-gridsemble_res$pi0)
fdrtool_cutoff <- quantile(fdrtool_res$lfdr, 1-unname(fdrtool_res$param[1,'eta0']))
locfdr_cutoff <- quantile(locfdr_res$fdr, 1-unname(locfdr_res$fp0['mlest','p0']))
qvalue_cutoff <- quantile(qvalue_res$lfdr, 1-qvalue_res$pi0)

classification_metrics_cutoff_pi0hat = rbind(
  classification_metrics(
    method = 'gridsemble',
    fdr = gridsemble_res$fdr,
    pi0 = gridsemble_res$pi0,
    test_statistics = platinum_data$statistics,
    truth = platinum_data$fold_change$DE,
    cutoff = gridsemble_cutoff
  ),
  classification_metrics(
    'locfdr',
    locfdr_res$fdr,
    locfdr_res$fp0['mlest','p0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = locfdr_cutoff
  ),
```

```
  classification_metrics(
    'fdrtool',
    fdrtool_res$lfdr,
    fdrtool_res$param[1,'eta0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = fdrtool_cutoff
  ),
  classification_metrics(
    'qvalue',
    qvalue_res$lfdr,
    qvalue_res$pi0,
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = qvalue_cutoff
  )
)

classification_metrics_cutoff_pi0hat
```

```
     method       cutoff       global_FDR sensitivity specificity prop_pred_T
[1,] "gridsemble" 0.5985333    0.07319014 0.5992798   0.9731465   0.2340782
[2,] "locfdr"     0.2306851    0.1973624  0.9079218   0.8733217   0.4094972
[3,] "fdrtool"    0.1215052    0.09630607 0.7047325   0.9573847   0.2823091
[4,] "qvalue"     2.611334e-05 0          0.001028807 1           0.0003724395
     TP   FP  TN   FN   accuracy  precision f1
[1,] 1165 92  3334 779  0.8378026 0.9268099 0.7278975
[2,] 1765 434 2992 179  0.8858473 0.8026376 0.8520396
[3,] 1370 146 3280 574  0.8659218 0.9036939 0.7919075
[4,] 2    0   3426 1942 0.6383613 1         0.002055498
```

cutoff based on π_0

```
pi0 = mean(platinum_data$fold_change$DE==0)
gridsemble_cutoff <- quantile(gridsemble_res$fdr, 1-pi0)
fdrtool_cutoff <- quantile(fdrtool_res$lfdr, 1-pi0)
locfdr_cutoff <- quantile(locfdr_res$fdr, 1-pi0)
qvalue_cutoff <- quantile(qvalue_res$lfdr, 1-pi0)

classification_metrics_cutoff_pi0 = rbind(
```

```
  classification_metrics(
    method = 'gridsemble',
    fdr = gridsemble_res$fdr,
    pi0 = gridsemble_res$pi0,
    test_statistics = platinum_data$statistics,
    truth = platinum_data$fold_change$DE,
    cutoff = gridsemble_cutoff
  ),
  classification_metrics(
    'locfdr',
    locfdr_res$fdr,
    locfdr_res$fp0['mlest','p0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = locfdr_cutoff
  ),
  classification_metrics(
    'fdrtool',
    fdrtool_res$lfdr,
    fdrtool_res$param[1,'eta0'],
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = fdrtool_cutoff
  ),
  classification_metrics(
    'qvalue',
    qvalue_res$lfdr,
    qvalue_res$pi0,
    platinum_data$statistics,
    platinum_data$fold_change$DE,
    cutoff = qvalue_cutoff
  )
)

classification_metrics_cutoff_pi0
```

```
     method       cutoff    global_FDR sensitivity specificity prop_pred_T TP
[1,] "gridsemble" 0.7394827 0.1502058  0.8497942   0.9147694   0.3620112   1652
[2,] "locfdr"     0.1292987 0.1502058  0.8497942   0.9147694   0.3620112   1652
[3,] "fdrtool"    1         0.6381077  0.9994856   0           0.9998138   1943
[4,] "qvalue"     1         0.6379888  1           0           1           1944
```

```
     FP   TN   FN  accuracy  precision f1
[1,] 292  3134 292 0.8912477 0.8497942 0.8497942
[2,] 292  3134 292 0.8912477 0.8497942 0.8497942
[3,] 3426 0    1   0.361825  0.3618923 0.5313825
[4,] 3426 0    0   0.3620112 0.3620112 0.5315833
```

What models made it into the ensemble?

```
fdrtool_rows = gridsemble_res$top_grid[gridsemble_res$top_grid$method == 'fdrtool',]$row
gridsemble_res$fdrtool_grid[fdrtool_rows,]
```

```
   cutoff.method      pct0
18          pct0 0.8736842
2         locfdr 0.7500000
16          pct0 0.8105263
17          pct0 0.8421053
19          pct0 0.9052632
15          pct0 0.7789474
```

```
locfdr_rows = gridsemble_res$top_grid[gridsemble_res$top_grid$method == 'locfdr',]$row
gridsemble_res$locfdr_grid[locfdr_rows,]
```

```
   pct  pct0 nulltype type
36   0 0.150        2    0
1    0 0.000        1    0
6    0 0.075        1    0
11   0 0.150        1    0
```

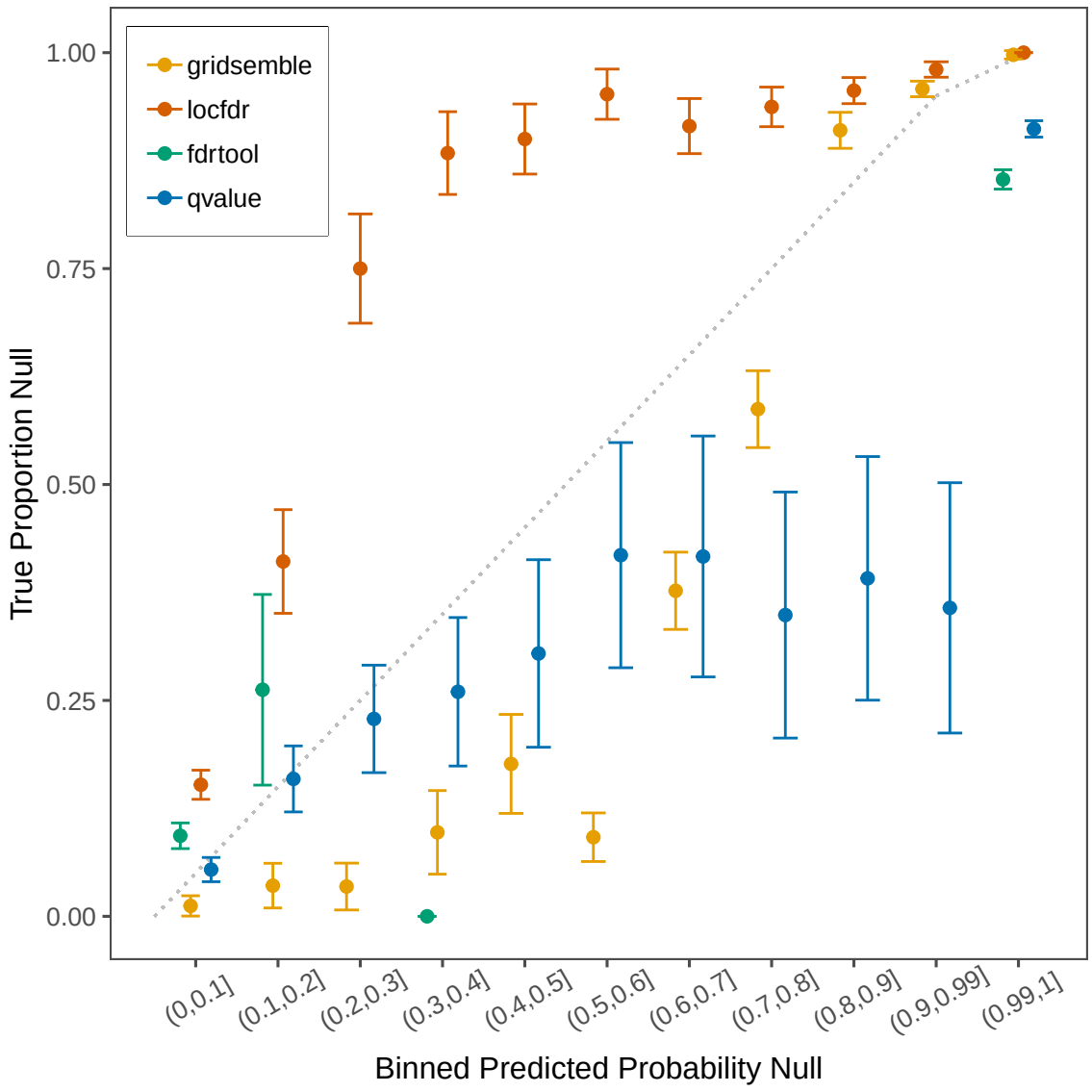
```
qvalue_rows = gridsemble_res$top_grid[gridsemble_res$top_grid$method == 'qvalue',]$row
gridsemble_res$qvalue_grid[qvalue_rows,]
```

```
[1] transf         adj            pi0.method     smooth.log.pi0
<0 rows> (or 0-length row.names)
```

Calibration

```
plot_calibration(
  fdrs = list(
    "gridsemble" = gridsemble_res$fdr,
    "locfdr" = locfdr_res$fdr,
    "fdrtool" = fdrtool_res$lfdr,
    "qvalue" = qvalue_res$lfdr
  ),
  truth = platinum_data$fold_change$DE
)
```

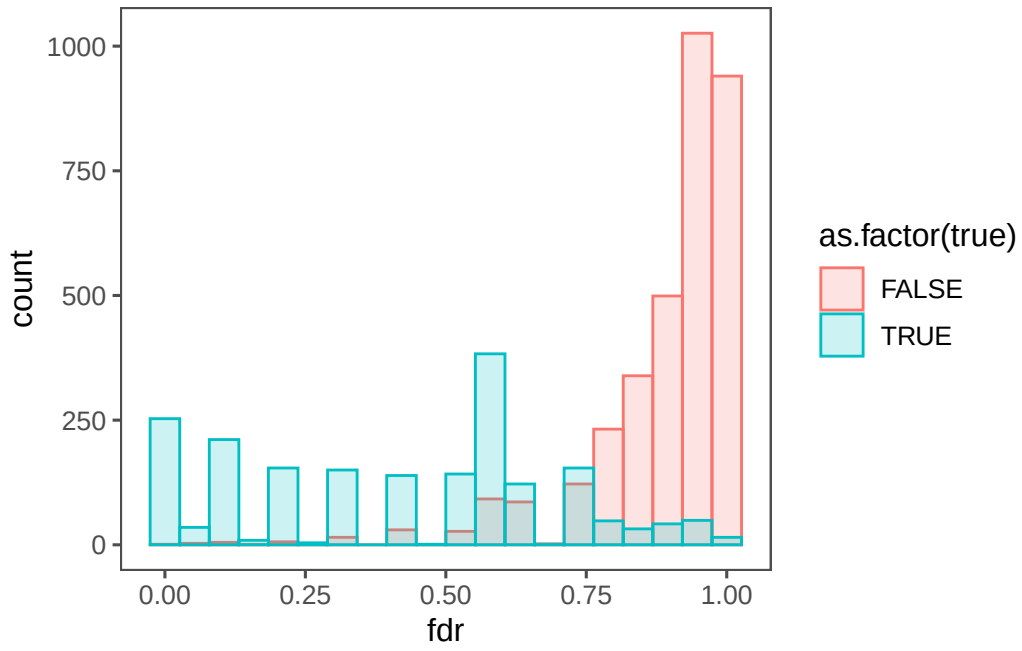
```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
ggsave("SupplementaryFigure3.png", height = 6, width = 6)
```

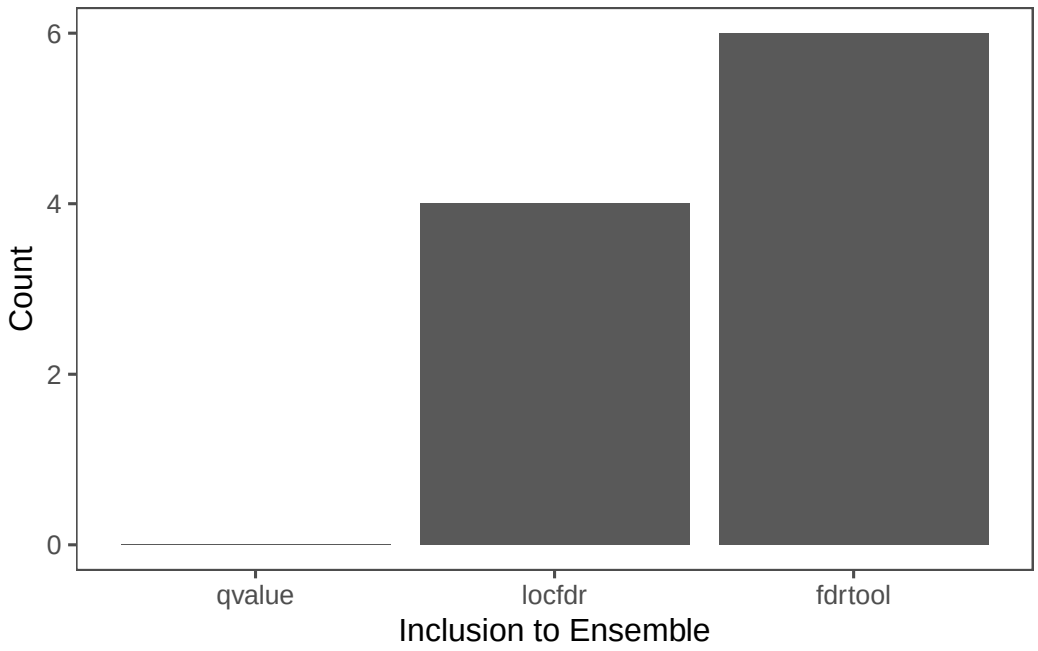
```
Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
data.frame(fdr=gridsemble_res$fdr, true =platinum_data$fold_change$DE) %>%
  ggplot(aes(x = fdr, color = as.factor(true), fill = as.factor(true))) +
  geom_histogram(position = 'identity', alpha = 0.2, bins = 20) +
  theme_few()
```



Ensemble contributions

```
gridsemble_res$top_grid %>%
    mutate(method = factor(method, levels = c('qvalue','locfdr','fdrtool'))) %>%
    pull(method) %>%
    table() %>%
    data.frame() %>%
    ggplot(aes(x = ., y = Freq)) +
    geom_bar(stat = 'identity') +
    theme_few() +
    labs(
        x = 'Inclusion to Ensemble',
        y = 'Count'
    )
```


6
4
2
0
Count
qvalue
locfdr
fdrtool
Inclusion to Ensemble