Bootstrap Assessment of Age-Associated Enrichment

Derek Chiu, Steven McKinney

May 19, 2017

# Introduction

To assess age-associated enrichment for intClust groups, we conducted a bootstrap analysis. We chose 50, 75, 100, 150, 200, and 300 bootstrap samples from each of the 10 intClust groups, and repeated this process 20 times. Each iteration repeat fitted a set of linear models predicting each gene target (48803 total) from age at diagnosis. An FDR of 0.01 was used for BH-adjusted p-values.

# Tests of Independence

Because of a mix of small and large expected counts in several of the tables, we obtained p-values for Fisher's exact test using Monte Carlo simulation with B = 100000 replicates. For each NiBoot size, we obtained 20 p-values, one for each bootstrap replicate. The contingency tables are cross-tabulations between intClust group (1-10) and probes exhibiting age dependent status (yes/no).

## Pearson Chi-Squared Tests

The distribution of p-values from Pearson's chi-square test for independence is shown below each bootstrap size table for the 20 replicates.

## Fisher's Exact Tests

The distribution of p-values from Fisher's exact test for independence is shown below for the 20 replicates. Because the p-values were obtained using simulation with B = 100000 replicates, the minimum p-value will be 0.0000100

## iClust group sizes

intClust number of cases

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 140 | 72 | 294 | 344 | 191 | 86 | 193 | 300 | 146 | 226 |

## Raw data iClust enrichment test

intClust number of cases

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 140 | 72 | 294 | 344 | 191 | 86 | 193 | 300 | 146 | 226 |

## 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | METABRIC Whole Series  FDR 0.01 FC>1.25 | | | |
| IntClusti | N cases | Trend | Probeset has Genomic Location | Probeset has Gene Name |
| iClust 1 | 140 | 8 | 8 | 8 |
| iClust 2 | 72 | 0 | 0 | 0 |
| iClust 3 | 294 | 878 | 876 | 765 |
| iClust 4 | 344 | 536 | 536 | 468 |
| iClust 5 | 191 | 4 | 4 | 4 |
| iClust 6 | 86 | 0 | 0 | 0 |
| iClust 7 | 193 | 12 | 12 | 10 |
| iClust 8 | 300 | 636 | 634 | 569 |
| iClust 9 | 146 | 1 | 1 | 1 |
| iClust 10 | 226 | 0 | 0 | 0 |
| N cases Total | 1992 |  |  |  |

The subgroups have varying sample size, so the difference in counts of probes showing age association reflects both underlying biology and statistical artifact. intClust groups with larger sample size will have greater power to detect differences of biological interest. Note that two of the groups showing no age-associated probes (intClust groups 2 and 6) are the subgroups with the smallest number of cases, and the intClust groups with the highest number of cases (intClust groups 3, 4 and 8) show the highest number of age-associated probes. intClust groups 5, 7 and 10 all show sample sizes near 200, yet show quite different counts of age-associated probes.

To further explore this issue we employed bootstrap methods and selected equal numbers of cases from each subtype.

## Table means

Average trend counts across bootstrap replications for each intClust group of size 50

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 10.20 | 48792.80 |
| 2 | 349.00 | 48454.00 |
| 3 | 1.80 | 48801.20 |
| 4 | 0.15 | 48802.85 |
| 5 | 4.55 | 48798.45 |
| 6 | 174.20 | 48628.80 |
| 7 | 3.20 | 48799.80 |
| 8 | 1.50 | 48801.50 |
| 9 | 10.05 | 48792.95 |
| 10 | 0.25 | 48802.75 |
| average | 55.49 | 48747.51 |

## Pearson Chi-Square Tests

## pvalue  
## 0 6.15e-140 5.99e-102 7.43e-91 4.59e-63 1.19e-49 1.48e-45   
## 9 1 1 1 1 1 1   
## 3.51e-41 8.98e-36 1.09e-15 2.99e-15 5e-04   
## 1 1 1 1 1

## Fisher's Exact Tests

## pvalue  
## 1e-05 0.00096   
## 19 1

All intClust groups had more than 50 cases.

SMQ - Why do we see observed count of zero for age associated probes for groups 2 and 6 (N=72 and 86) yet see such large counts here?

Average trend counts across bootstrap replications for each intClust group of size 75

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 158.90 | 48644.10 |
| 2 | 885.50 | 47917.50 |
| 3 | 19.90 | 48783.10 |
| 4 | 6.50 | 48796.50 |
| 5 | 44.90 | 48758.10 |
| 6 | 636.90 | 48166.10 |
| 7 | 4.60 | 48798.40 |
| 8 | 49.20 | 48753.80 |
| 9 | 120.65 | 48682.35 |
| 10 | 23.85 | 48779.15 |
| average | 195.09 | 48607.91 |

## Pearson Chi-Square Tests

## pvalue  
## 0 2.03e-278 5.14e-125   
## 18 1 1

## Fisher's Exact Tests

## pvalue  
## 1e-05   
## 20

Average trend counts across bootstrap replications for each intClust group of size 100

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 486.100 | 48316.90 |
| 2 | 2308.400 | 46494.60 |
| 3 | 106.800 | 48696.20 |
| 4 | 11.250 | 48791.75 |
| 5 | 100.650 | 48702.35 |
| 6 | 1212.500 | 47590.50 |
| 7 | 219.250 | 48583.75 |
| 8 | 63.600 | 48739.40 |
| 9 | 232.300 | 48570.70 |
| 10 | 7.000 | 48796.00 |
| average | 474.785 | 48328.21 |

## Pearson Chi-Square Tests

## pvalue  
## 0   
## 20

## Fisher's Exact Tests

## pvalue  
## 1e-05   
## 20

Average trend counts across bootstrap replications for each intClust group of size 150

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 1209.70 | 47593.30 |
| 2 | 5652.45 | 43150.55 |
| 3 | 502.00 | 48301.00 |
| 4 | 320.45 | 48482.55 |
| 5 | 358.50 | 48444.50 |
| 6 | 3255.30 | 45547.70 |
| 7 | 373.35 | 48429.65 |
| 8 | 266.35 | 48536.65 |
| 9 | 656.95 | 48146.05 |
| 10 | 132.85 | 48670.15 |
| average | 1272.79 | 47530.21 |

## Pearson Chi-Square Tests

## pvalue  
## 0   
## 20

## Fisher's Exact Tests

## pvalue  
## 1e-05   
## 20

Average trend counts across bootstrap replications for each intClust group of size 200

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 2211.200 | 46591.80 |
| 2 | 8462.200 | 40340.80 |
| 3 | 1050.750 | 47752.25 |
| 4 | 569.300 | 48233.70 |
| 5 | 765.300 | 48037.70 |
| 6 | 4538.200 | 44264.80 |
| 7 | 1036.550 | 47766.45 |
| 8 | 688.550 | 48114.45 |
| 9 | 2162.150 | 46640.85 |
| 10 | 352.050 | 48450.95 |
| average | 2183.625 | 46619.38 |

## Pearson Chi-Square Tests

## pvalue  
## 0   
## 20

## Fisher's Exact Tests

## pvalue  
## 1e-05   
## 20

Average trend counts across bootstrap replications for each intClust group of size 300

|  |  |  |
| --- | --- | --- |
| group | avg\_trend | avg\_no\_trend |
| 1 | 4615.30 | 44187.70 |
| 2 | 11910.05 | 36892.95 |
| 3 | 1942.95 | 46860.05 |
| 4 | 1217.65 | 47585.35 |
| 5 | 2620.50 | 46182.50 |
| 6 | 7936.55 | 40866.45 |
| 7 | 2238.05 | 46564.95 |
| 8 | 1892.40 | 46910.60 |
| 9 | 3611.10 | 45191.90 |
| 10 | 1348.55 | 47454.45 |
| average | 3933.31 | 44869.69 |

## Pearson Chi-Square Tests

## pvalue  
## 0   
## 20

## Fisher's Exact Tests

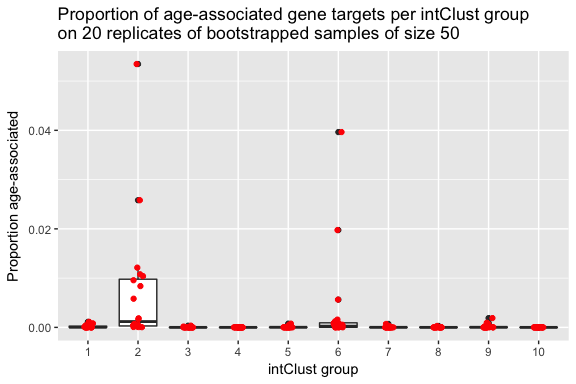
## pvalue  
## 1e-05   
## 20

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | METABRIC Whole Series | | | |
| IntClusti | N cases | Trend | Probeset has Genomic Location | Probeset has Gene Name |
| iClust 1 | 140 | 113 | 112 | 105 |
| iClust 2 | 72 | 0 | 0 | 0 |
| iClust 3 | 294 | 1584 | 1581 | 1387 |
| iClust 4 | 344 | 1235 | 1232 | 1076 |
| iClust 5 | 191 | 19 | 19 | 18 |
| iClust 6 | 86 | 0 | 0 | 0 |
| iClust 7 | 193 | 91 | 91 | 76 |
| iClust 8 | 300 | 1292 | 1289 | 1144 |
| iClust 9 | 146 | 1 | 1 | 1 |
| iClust 10 | 226 | 0 | 0 | 0 |
| N cases Total | 1992 |  |  |  |

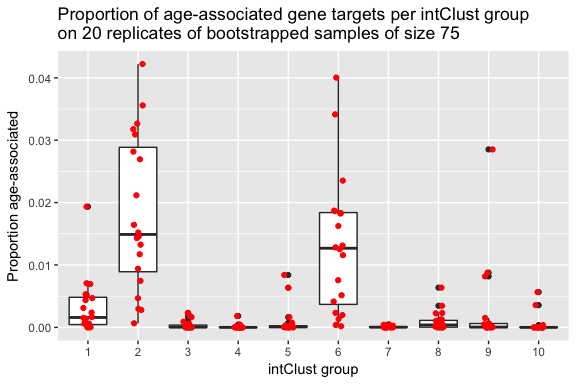
# Boxplots

We can also visualize the results using boxplots. Each of the following boxplots shows the distribution of the proportion of age-associated gene targets within each intClust group across the 20 replicates. There is a figure for each of the NiBoot sizes chosen.

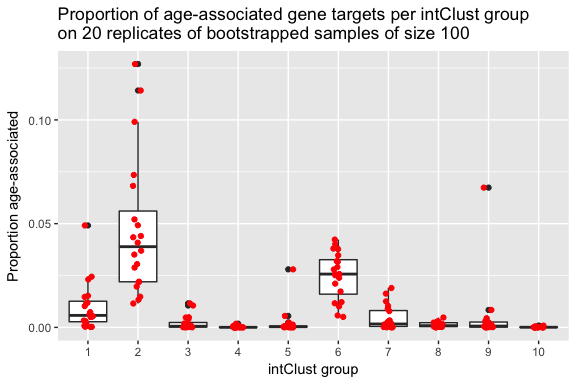
## NiBoot = 50



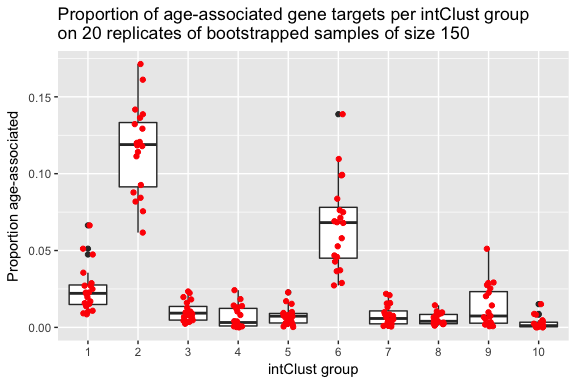
## NiBoot = 75



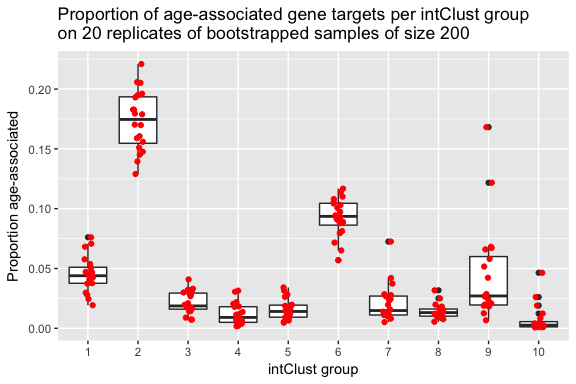
## NiBoot = 100



## NiBoot = 150



## NiBoot = 200



## NiBoot = 300

