Different Alternative Types

David Gerard

2016-05-28

Abstract

This the same setup as here except I use bicrossvalidation (using the package `cate`) to estimate the number of hidden confounders rather than `sva::num.sv`. The results are basically the same, with SUCCOTASH and RUVASH doing very well compared to other methods.

Simulation Setup

I ran through 200 repetitions of generating data from GTEX muscle data under the following parameter conditions:

- $n \in \{10, 20, 40\}$,
- $p = 1000$.
- $\pi_0 \in \{0.5, 0.9\}$,
- The alternative distribution being either spiky, near-normal, flattop, skew, big-normal, or bimodal, where these are the same alternatives defined in Stephens (2016) and the following table. New alternatives are generated every iteration.

Scenario	Alternative Distribution
Spiky	$0.4N(0, 0.25^2) + 0.2N(0, 0.5^2) + 0.2N(0, 1^2), 0.2N(0, 2^2)$
Near Normal	$2/3N(0, 1^2) + 1/3N(0, 2^2)$
Flattop	$(1/7)N(-1.5, .5^2) + N(-1, .5^2) + N(-.5, .5^2) + N(0, .5^2) + N(0.5, .5^2) + N(1.0, .5^2) + N(1.5, .5^2)$
Skew	$(1/4)N(-2, 2^2) + (1/4)N(-1, 1.5^2) + (1/3)N(0, 1^2) + (1/6)N(1, 1^2)$
Big-normal	$N(0, 4^2)$
Bimodal	$0.5N(-2, 1^2) + 0.5N(2, 1^2)$

I extracted the most expressed p genes from the GTEX muscle data and n samples are chosen at random. Half of these samples are randomly given the "treatment" label 1, the other half given the "control" label 0. Of the p genes, $\pi_0 p$ were chosen to be non-null. Signal was added by a Poisson-thinning approach, where the log-2 fold change was sampled from one of five the alternative models above. That is

$$A_1, \ldots, A_{p/2} \sim f \tag{1}$$

$$B_i = 2^{A_i} \text{ for } i = 1, \ldots, p/2, \tag{2}$$

where f is from the table above. If $A_i > 0$ then we replace $Y_{[1:(n/2),i]}$ with $Binom(Y_{[j,i]}, 1/B_i)$ for $j = 1, \ldots, n/2$. If $A_i < 0$ then we replace $Y_{[(n/2+1):n,i]}$ with $Binom(Y_{[j,i]}, B_i)$ for $j = n/2 + 1, \ldots, n$.

I now describe the justification for this. Suppose that

$$Y_{ij} \sim Poisson(\lambda_j). \tag{3}$$

Let x_i be the indicator of treatment vs control for individual i. Let Ω be the set of non-null genes. Let Z be the new dataset derived via the steps above. That is

$$Z_{ij}|Y_{ij} = \begin{cases} Binom(Y_{ij}, 2^{A_j x_i}) & \text{if } A_j < 0 \text{ and } j \in \Omega \\ Binom(Y_{ij}, 2^{-A_j(1-x_i)}) & \text{if } A_j > 0 \text{ and } j \in \Omega \\ Y_{ij} & \text{if } j \notin \Omega. \end{cases} \tag{4}$$

Then

$$Z_{ij}|A_j, A_j < 0, j \in \Omega \sim Poisson(2^{A_j x_i}\lambda_j) \tag{5}$$
$$Z_{ij}|A_j, A_j > 0, j \in \Omega \sim Poisson(2^{-A_j(1-x_i)}\lambda_j), \tag{6}$$

and

$$E[\log_2(Z_{ij}) - \log_2(Z_{kj})|A_j, A_j < 0, j \in \Omega] \approx A_j x_i - A_j x_k, \text{ and} \tag{7}$$
$$E[\log_2(Z_{ij}) - \log_2(Z_{kj})|A_j, A_j > 0, j \in \Omega] \approx -A_j(1-x_i) + A_j(1-x_k). \tag{8}$$

if individual i is in the treatment group and individual k is in the control group, then this just equals A_j. I treat the A_j's as the true coefficient values when calculating the MSE below.

Methods

I first normalized the counts by $\log_2(COUNTS + 1)$. The number of hidden confounders was estimated using `cate::est.confounder.num` in `R`.

The confounder adjustment methods I look at in this write-up are:

- OLS + qvalue.
- OLS + ASH
- SUCCOTASH using normal mixtures and heteroscedastic PCA as the factor-analysis method.
- The robust regression version of CATE using PCA as the factor analysis method + qvalue.
- Ridge version of LEAPP + qvalue.
- SVA + qvalue.
- RUVASH
- Negative control version of CATE using PCA as the factor analysis method + qvalue.
- RUV2 + qvalue.
- RUV4 + qvalue.
- RUV4 + ASH (no variance inflation).

Results

Note that in the plots below, n refers to the size of each group, not the total size.

Estimates of π_0

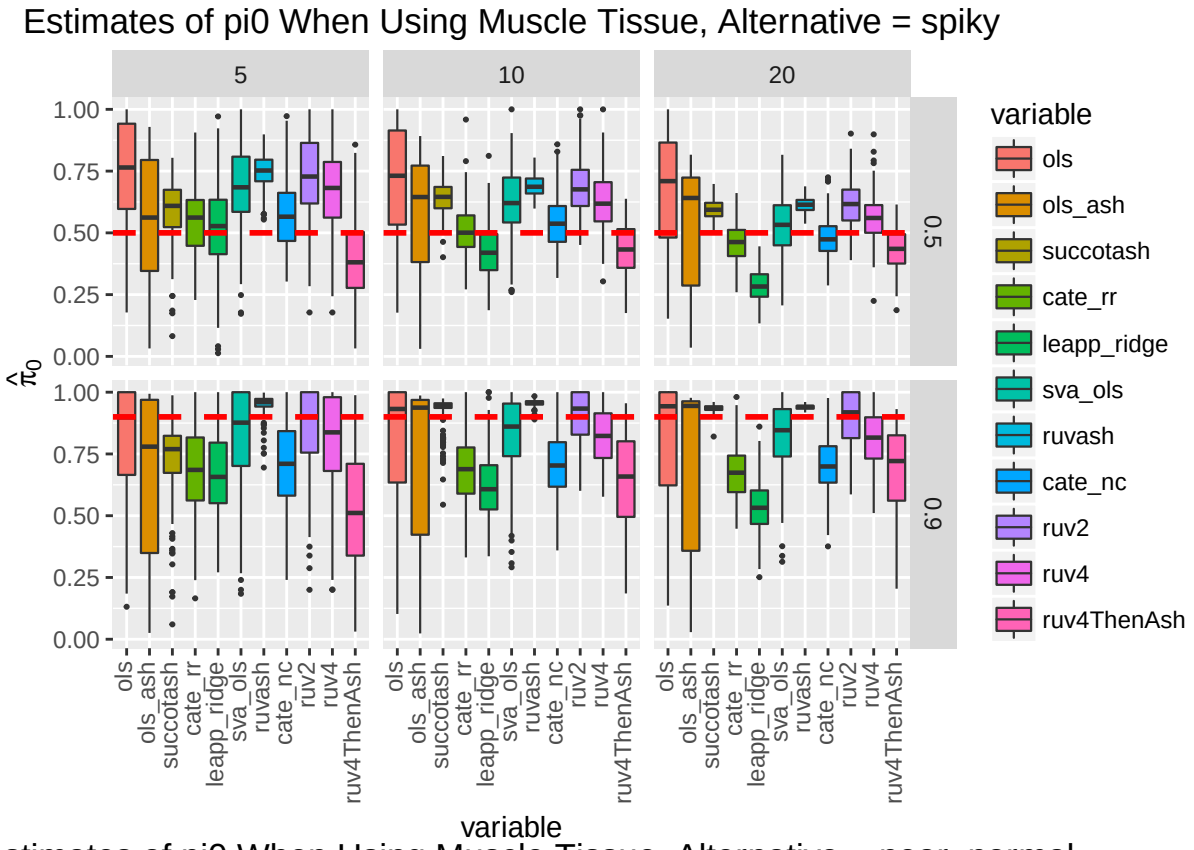
- SUCCOTASH has slightly anti-conservative estimates of π_0 in the Flattop and bimodal Scenarios. It does well for every other scenario for larger n. RUVASH also does very well.
- Without the variance inflation, RUV4ThenASH does very poorly.
- No method using qvalue ever performed as well as succotash. Indeed, none exhibited this "conservative with high-probability" behavior that is desirable.

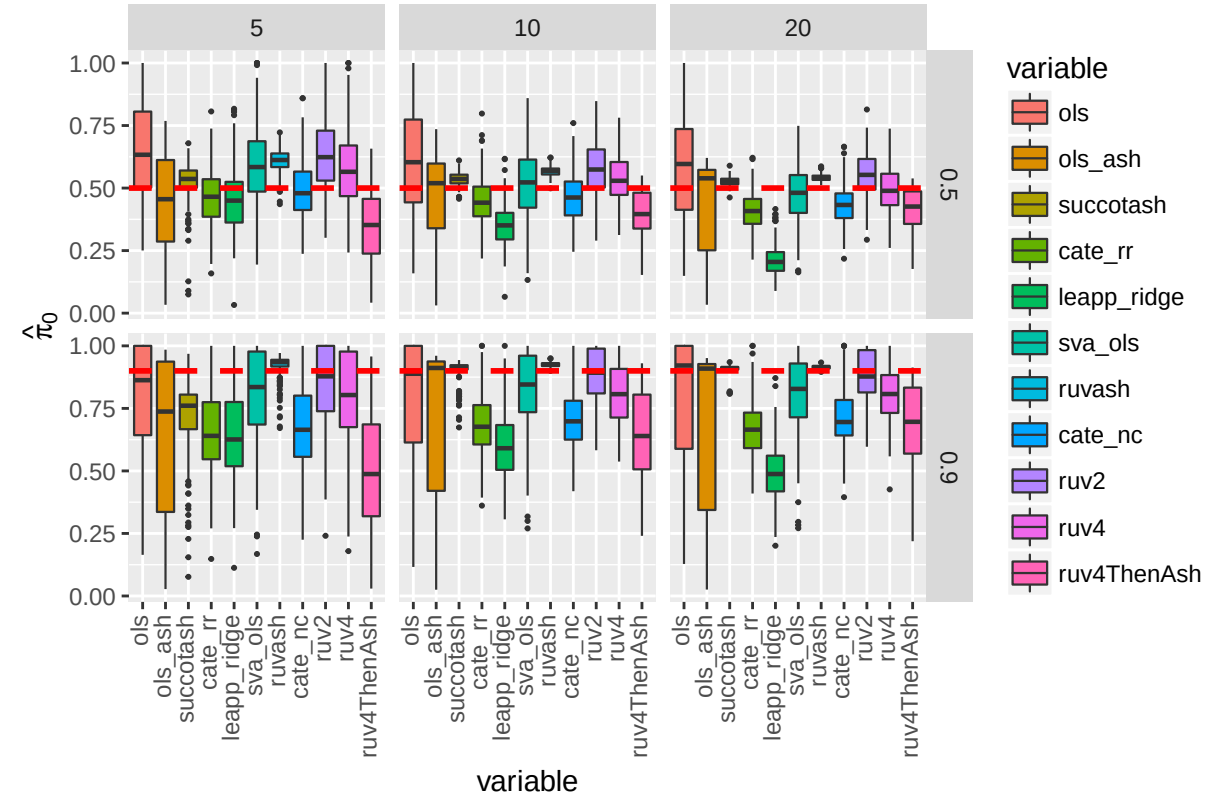
AUC performance.

The ASH-like methods always have higher AUC, even in the bimodal and big-normal scenarios where LEAPP estimated π_0 more accurately.

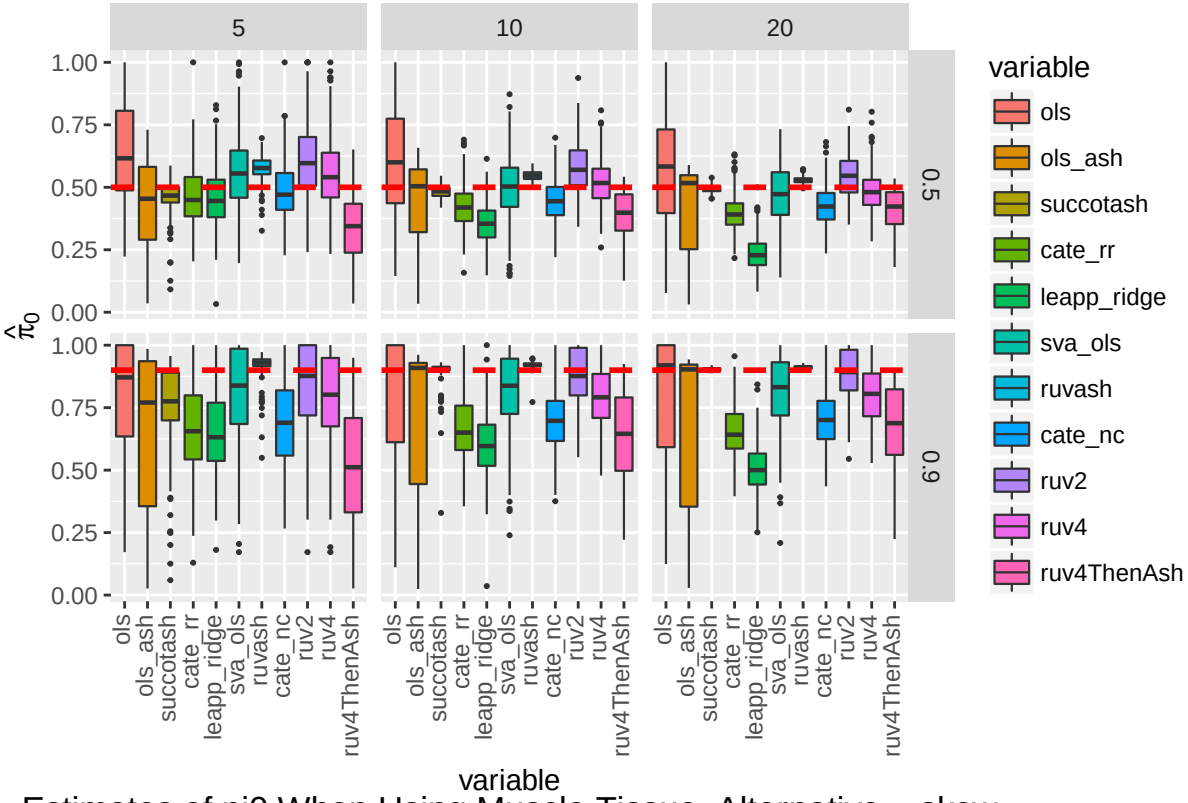
MSE

The ASH-like methods have superior performance in term of AUC compared to other methods. LEAPP has terrible MSE performance in all cases except the all-null setting. Even then, SUCCOTASH has better performance.

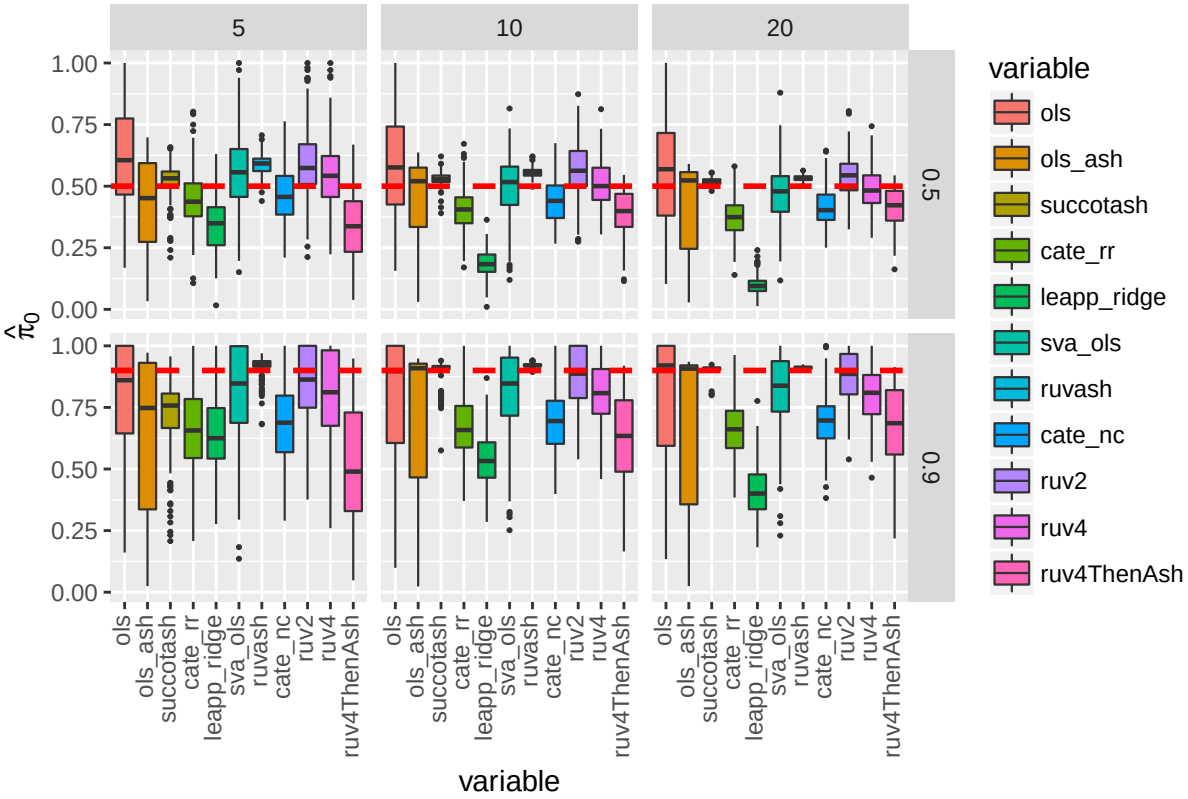

Estimates of pi0 When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh


stimates of pi0 When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

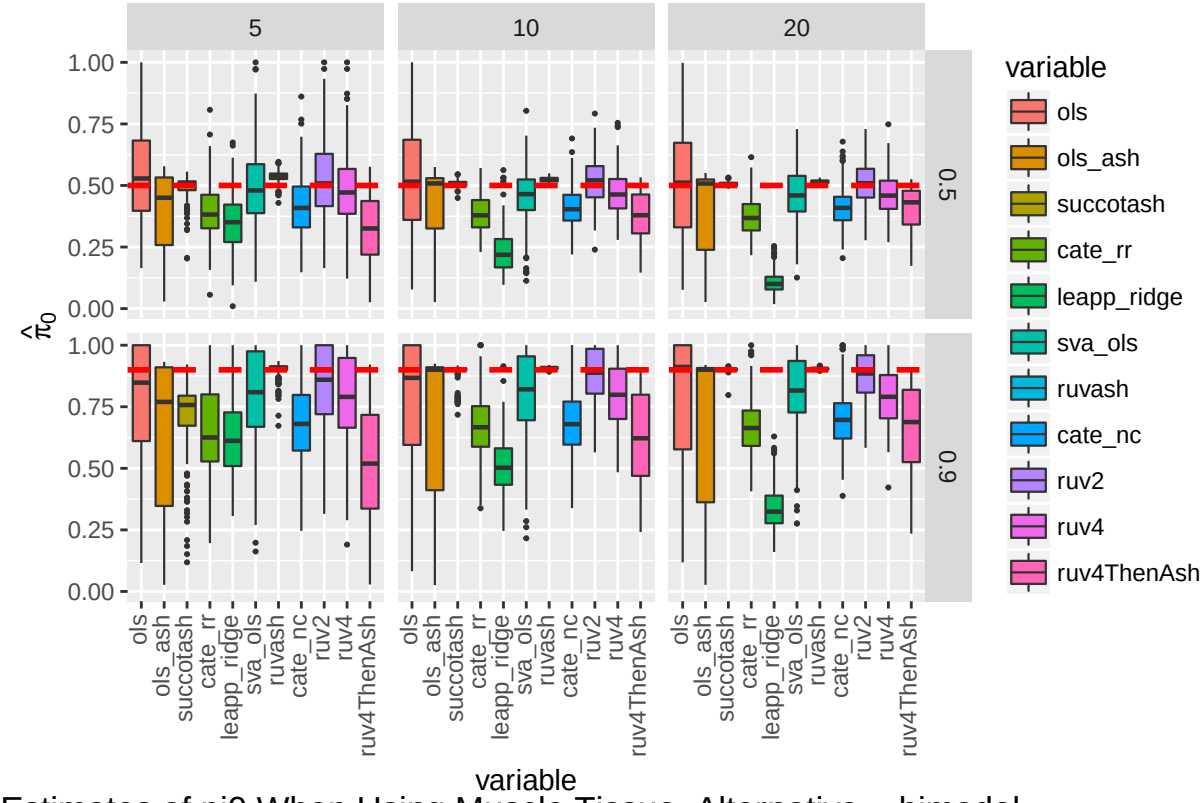
Estimates of pi0 When Using Muscle Tissue, Alternative = flattop



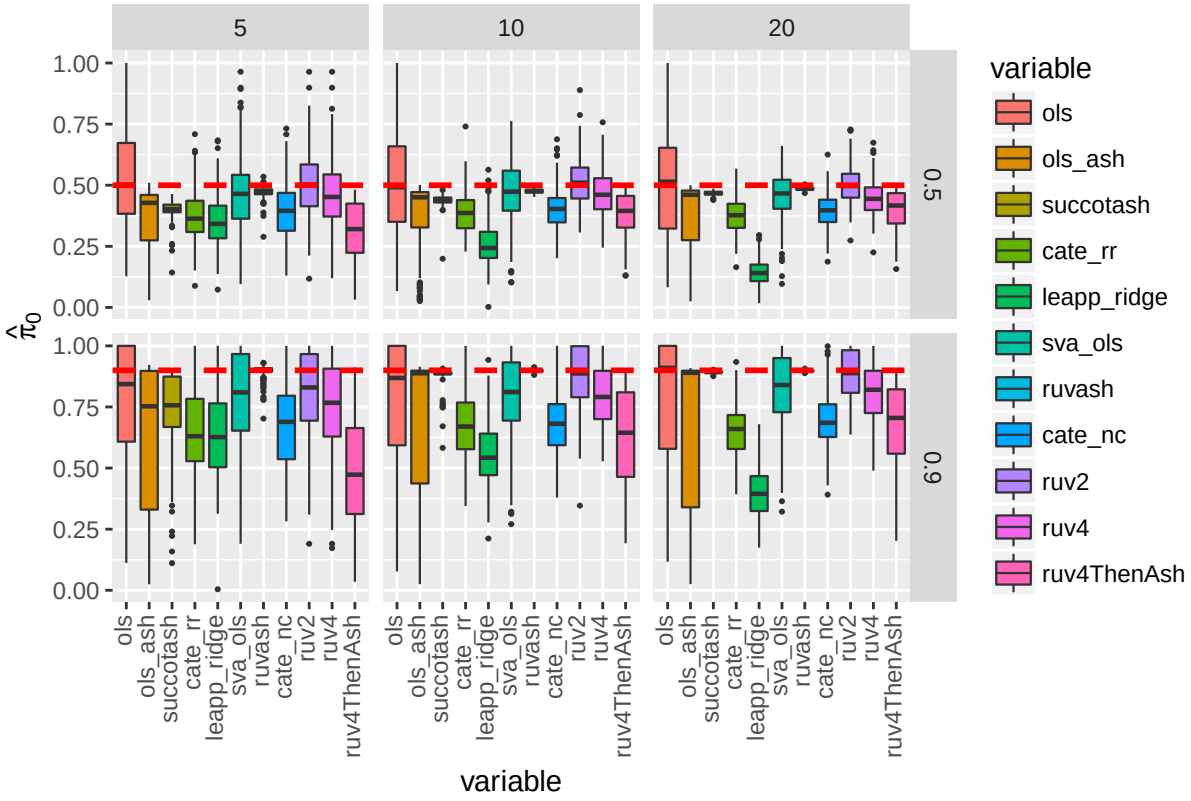
Estimates of pi0 When Using Muscle Tissue, Alternative = skew

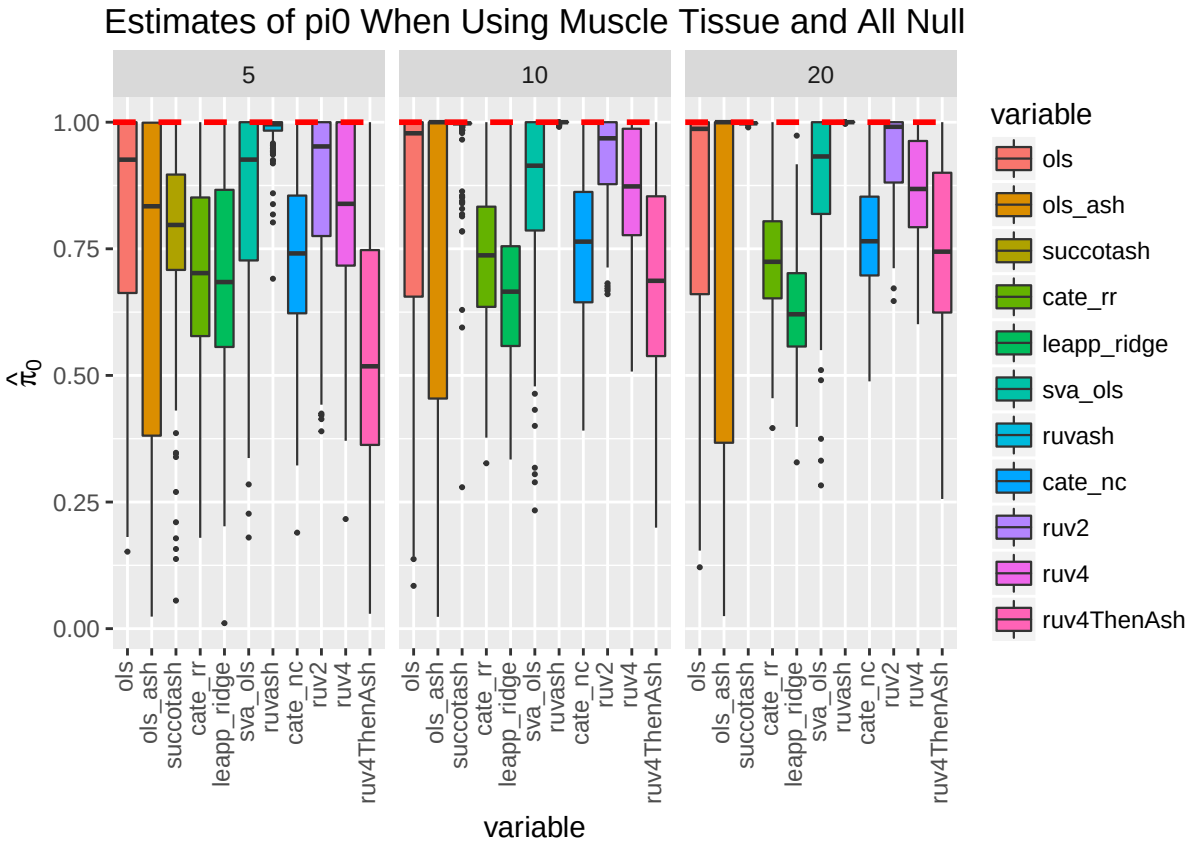


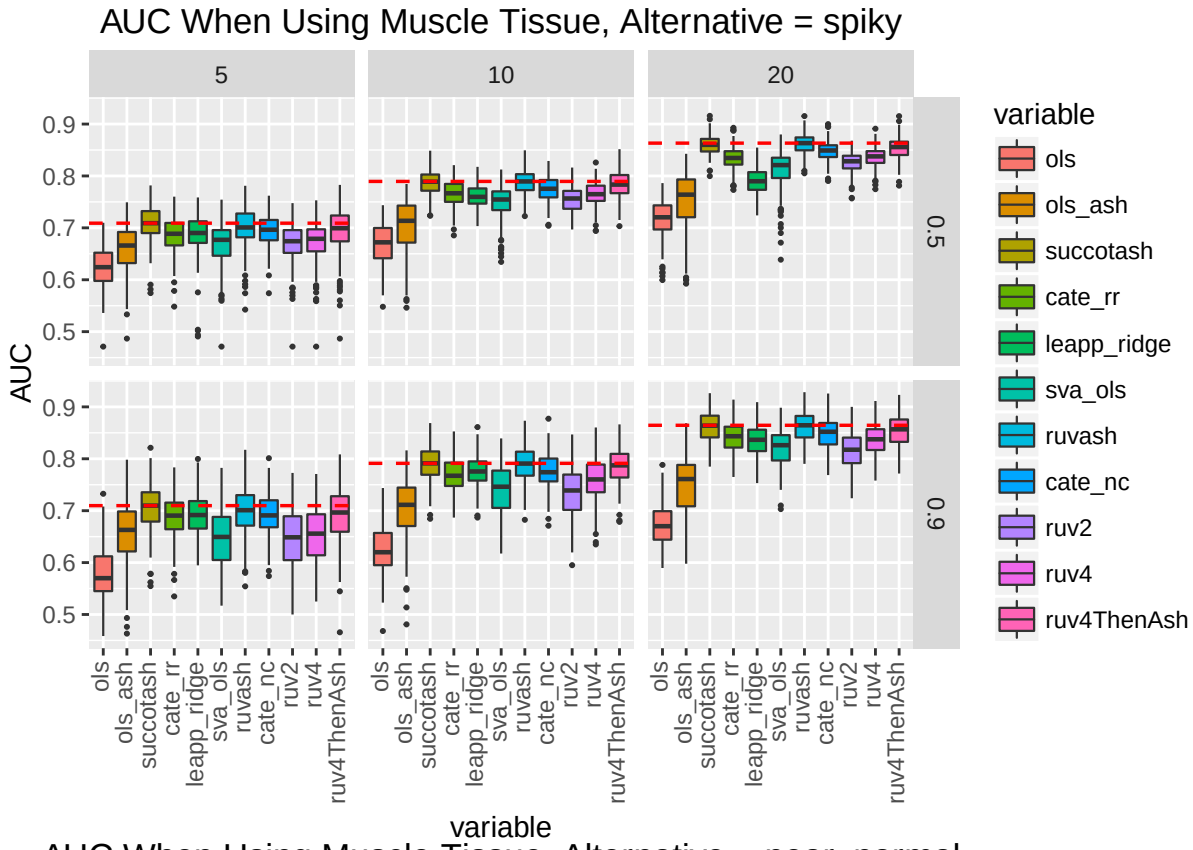
Estimates of pi0 When Using Muscle Tissue, Alternative = big_normal

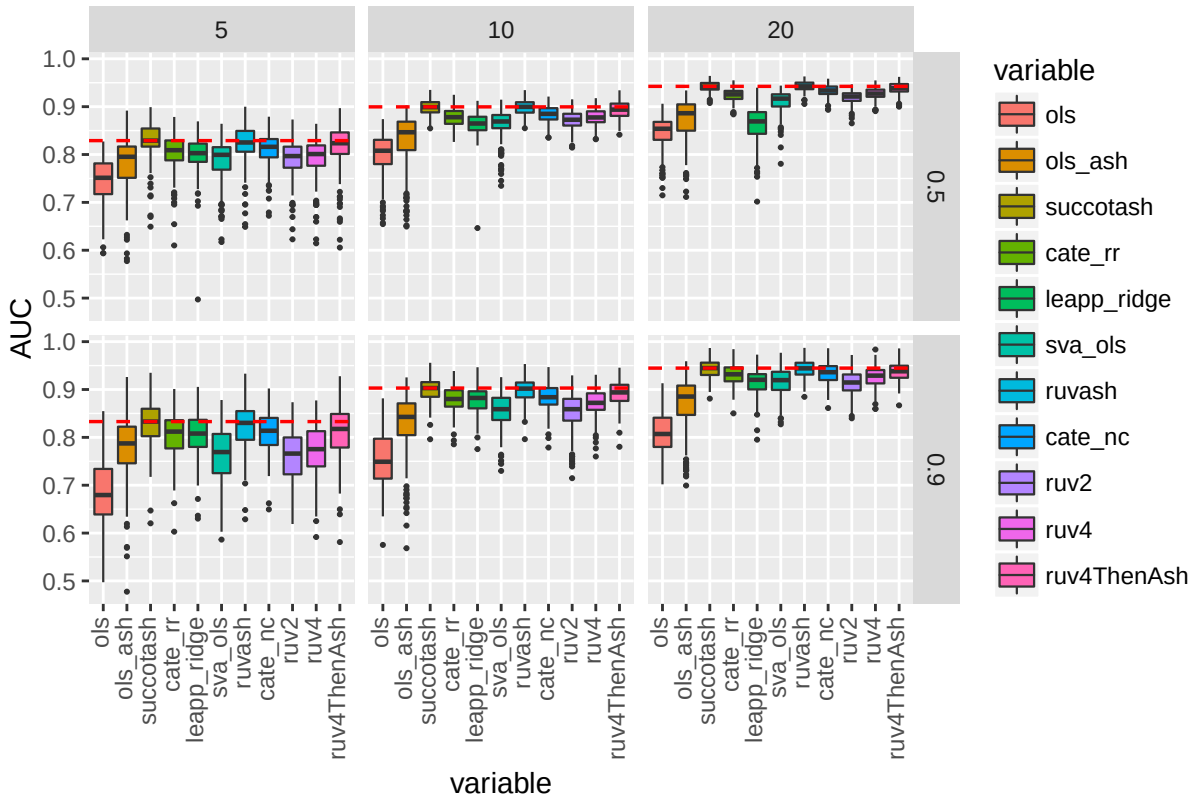


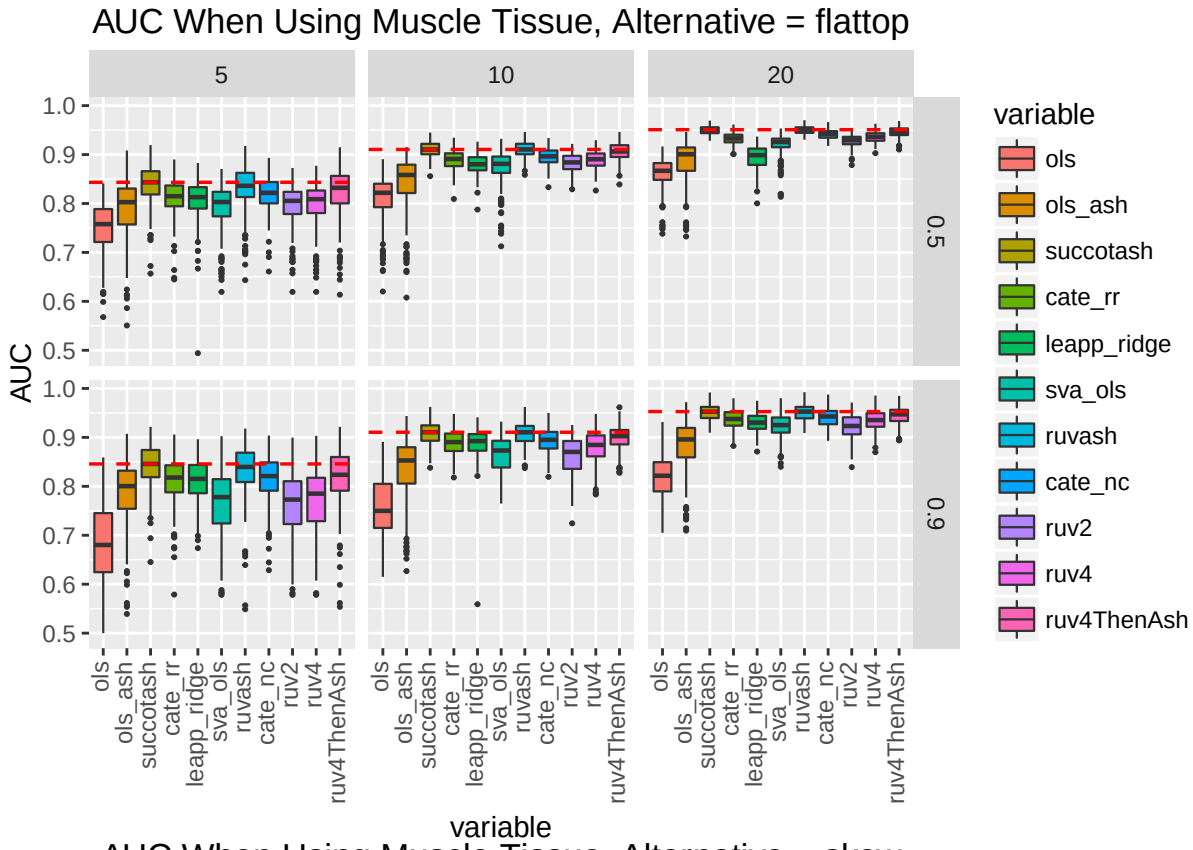
Estimates of pi0 When Using Muscle Tissue, Alternative = bimodal

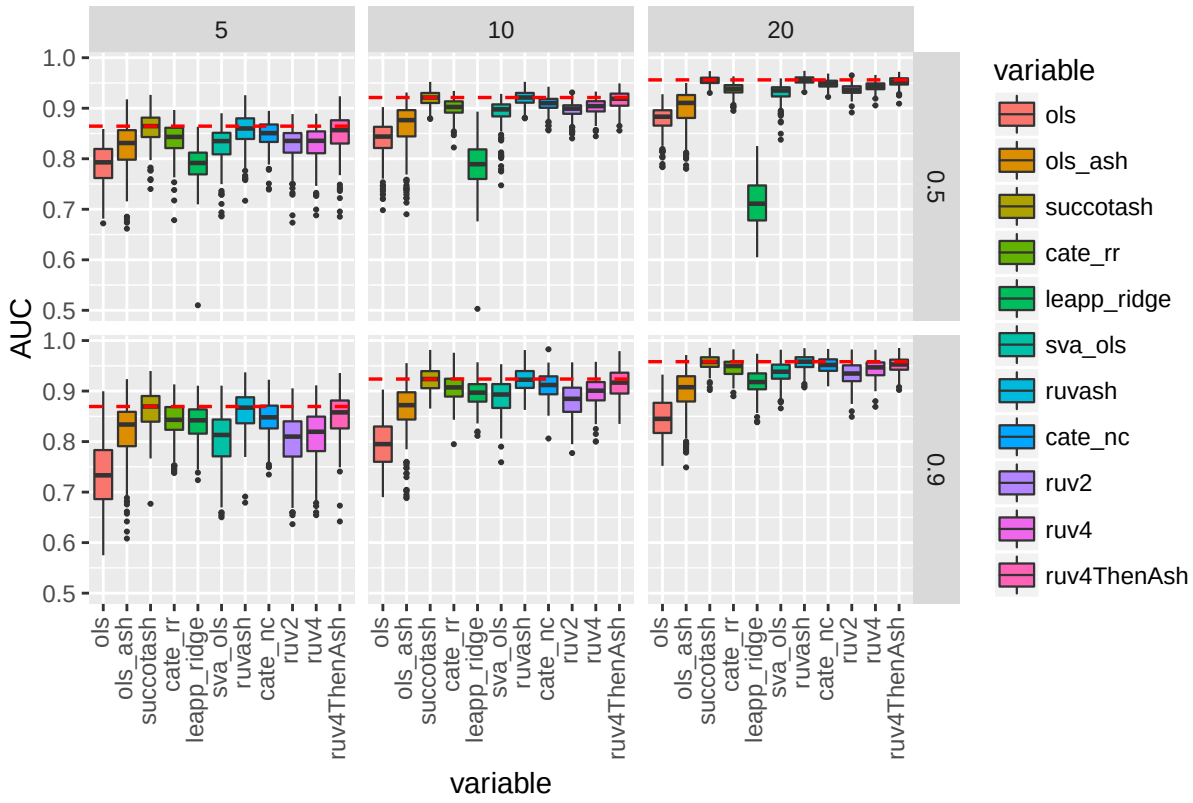


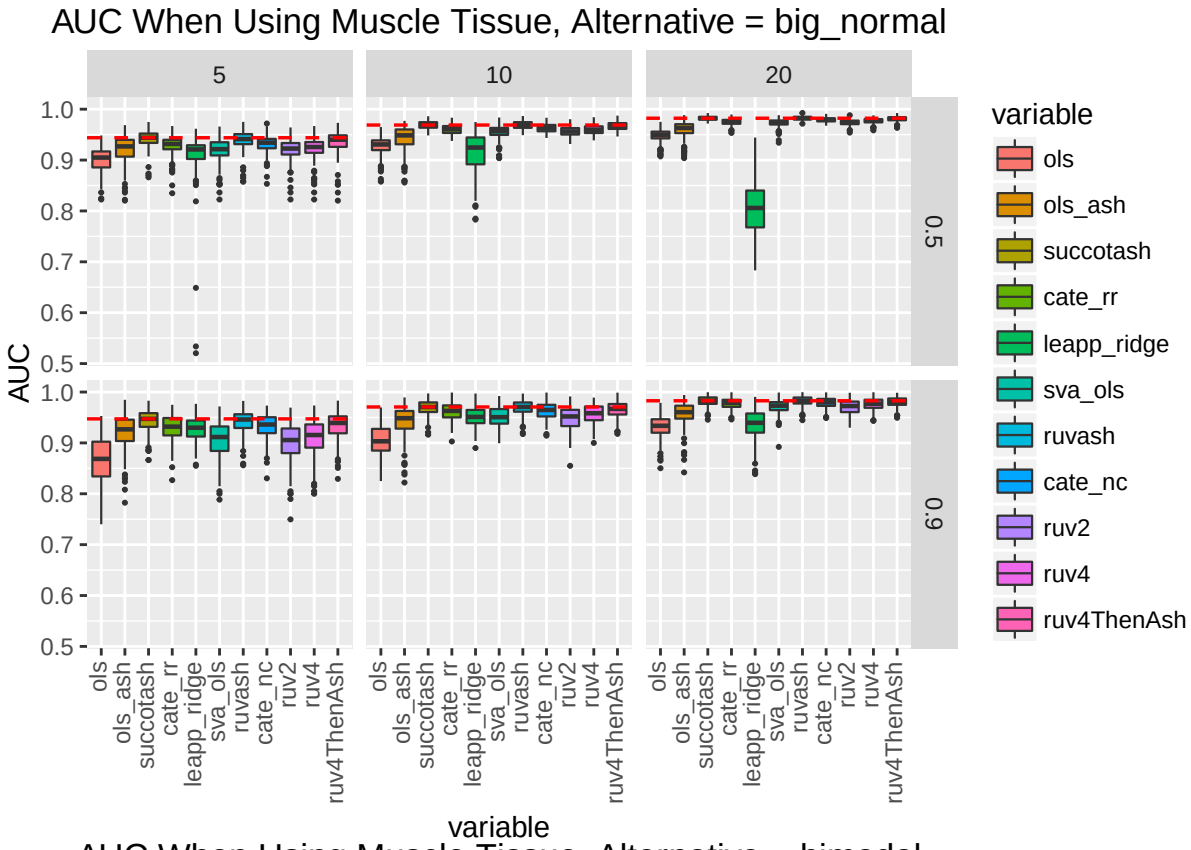

Estimates of pi0 When Using Muscle Tissue and All Null
5
10
20
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

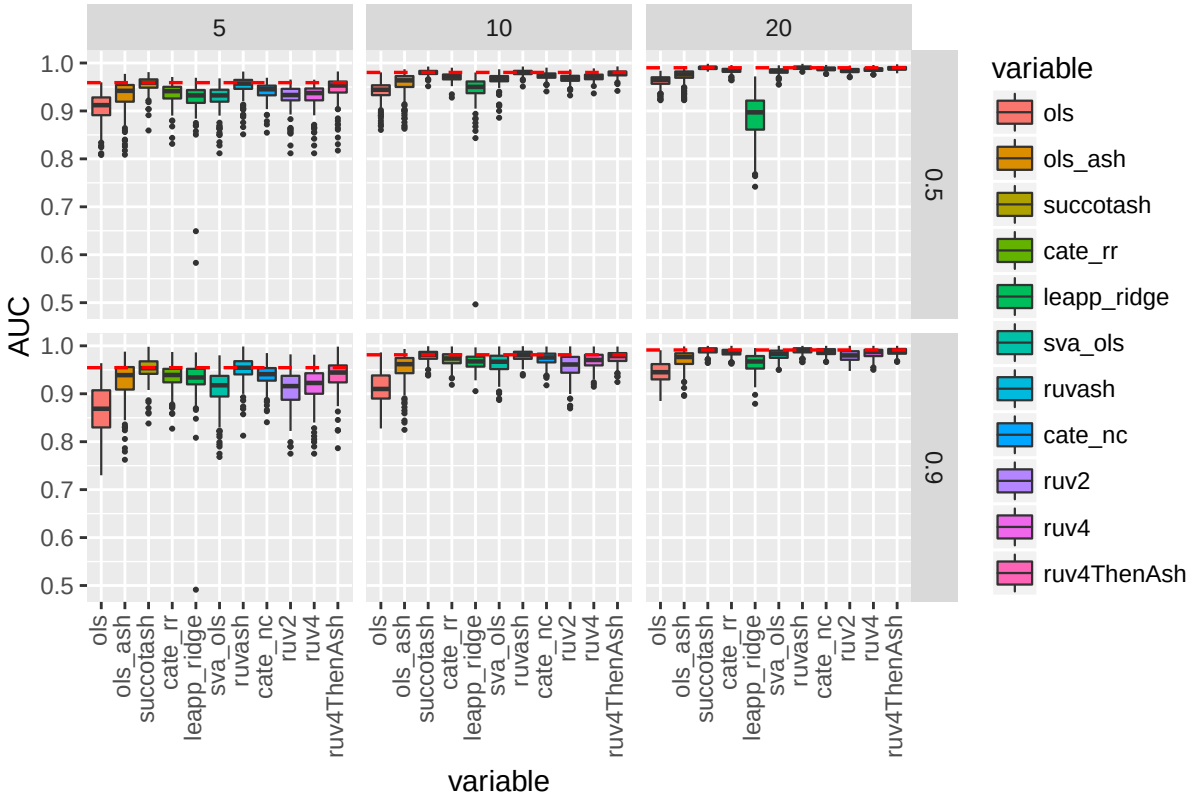

AUC When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
AUC
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

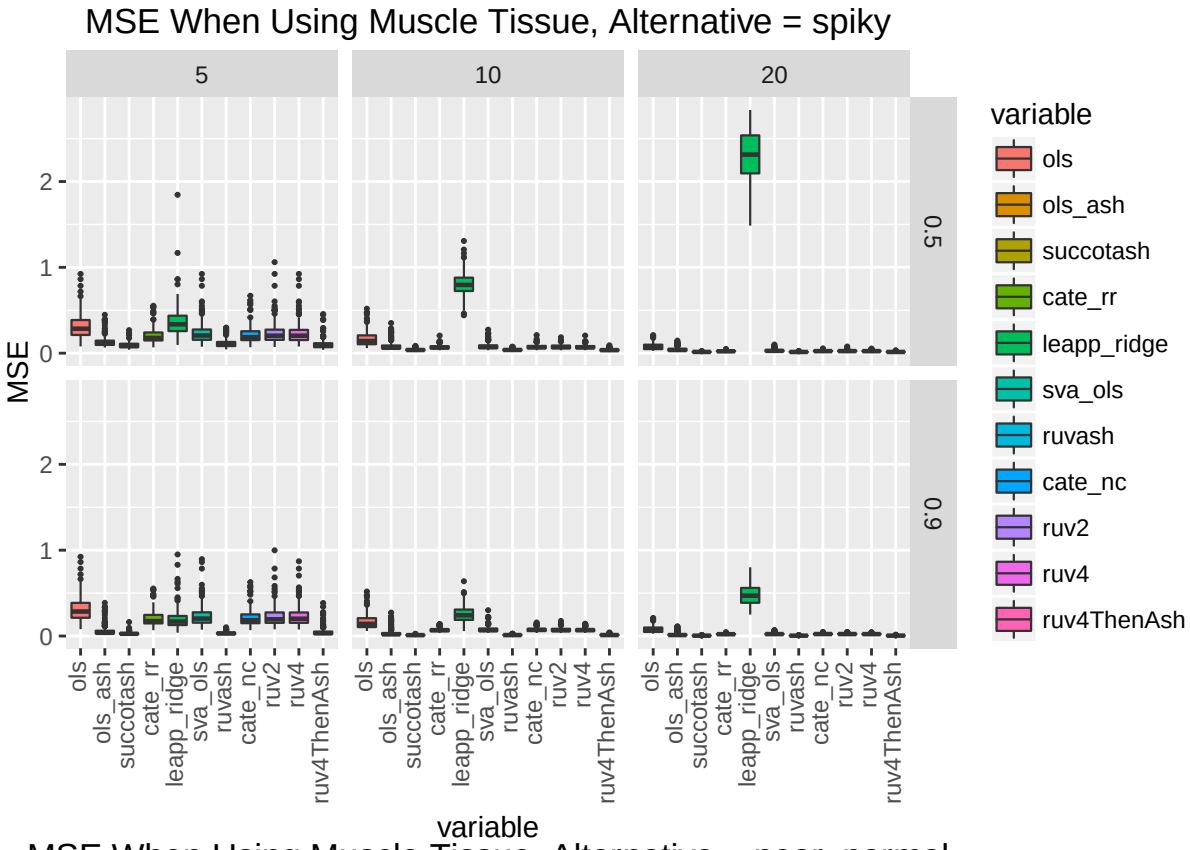

AUC When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
AUC
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

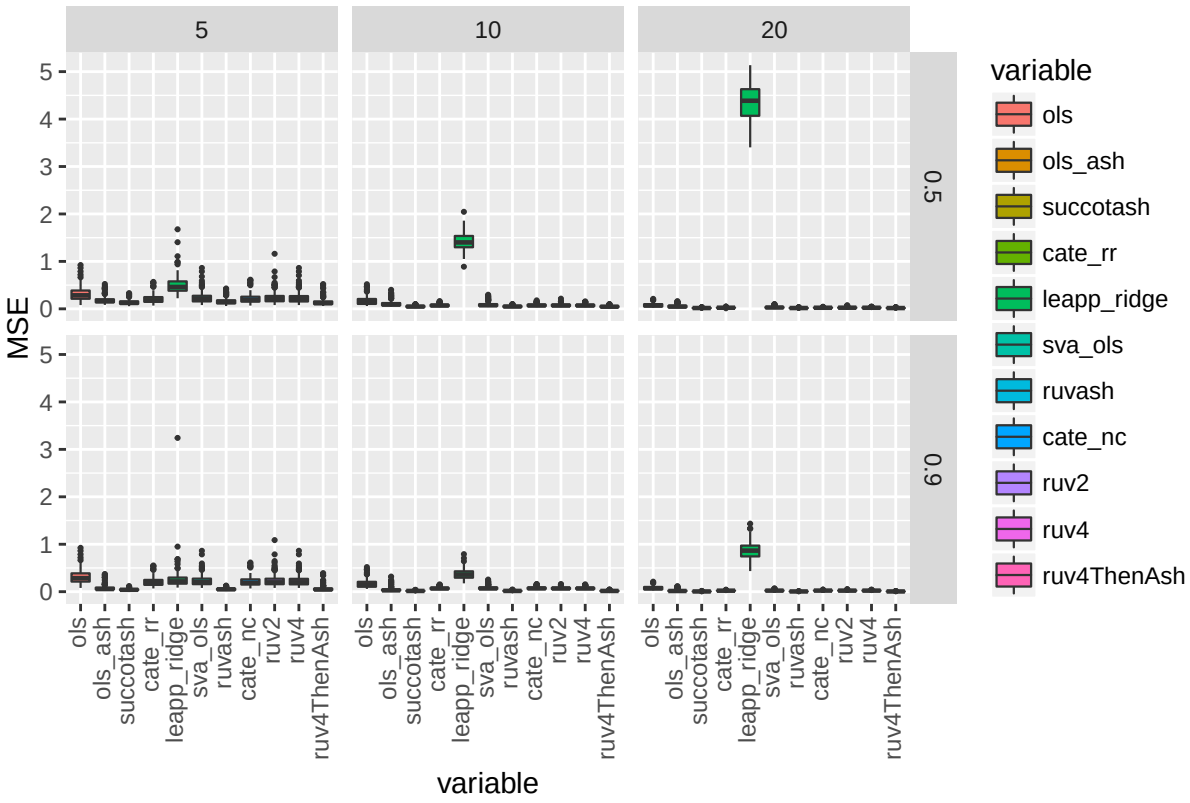

AUC When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

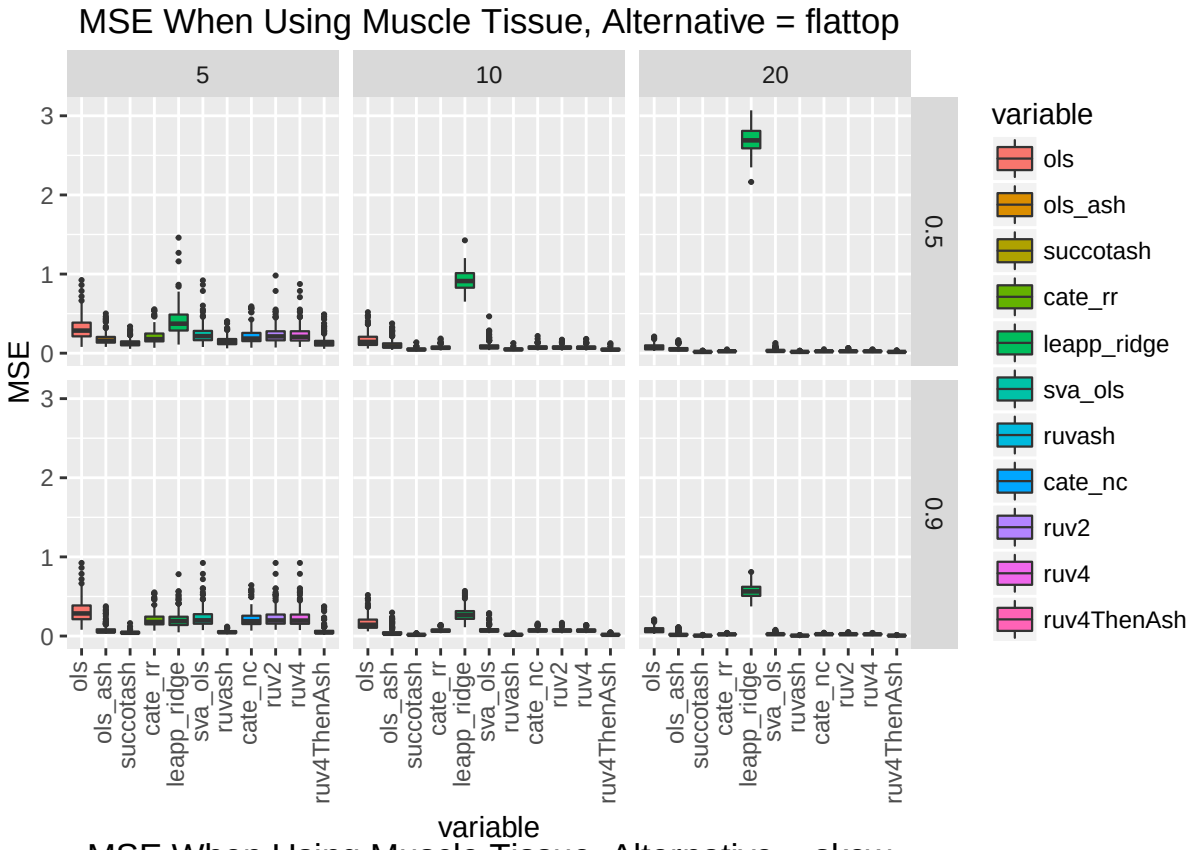

AUC When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

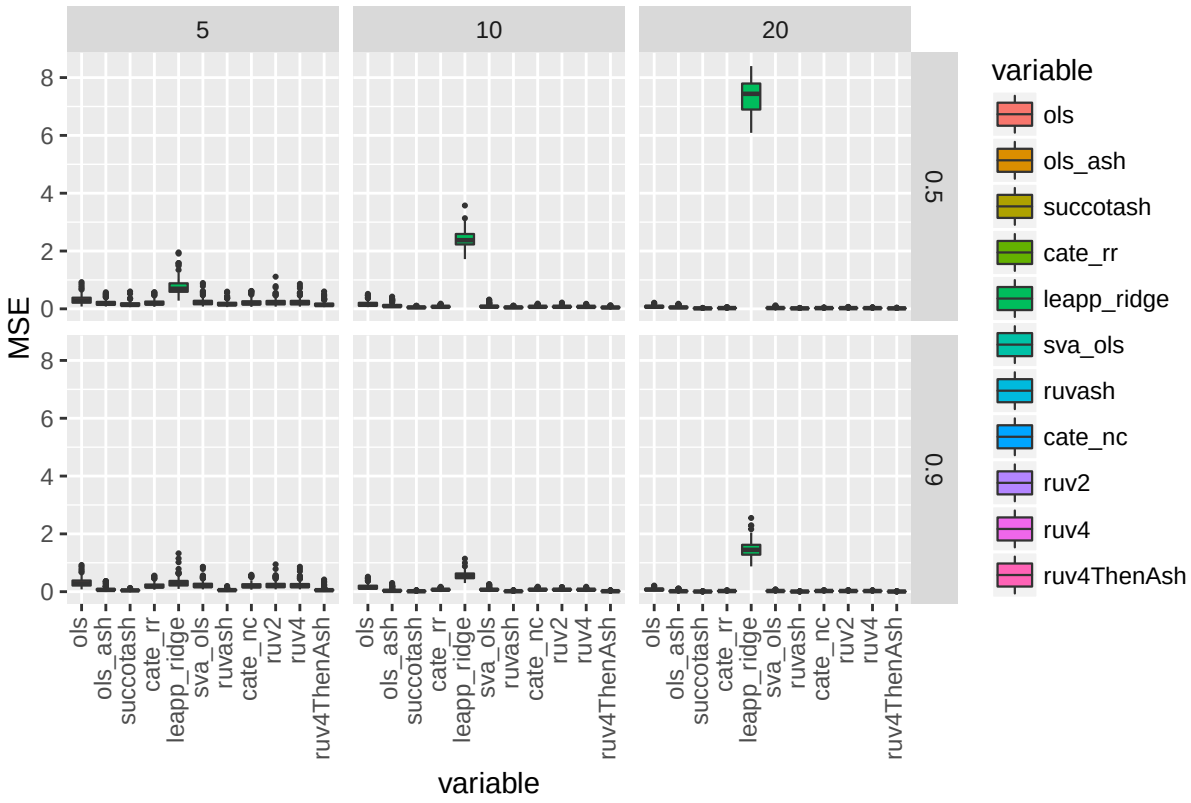

AUC When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

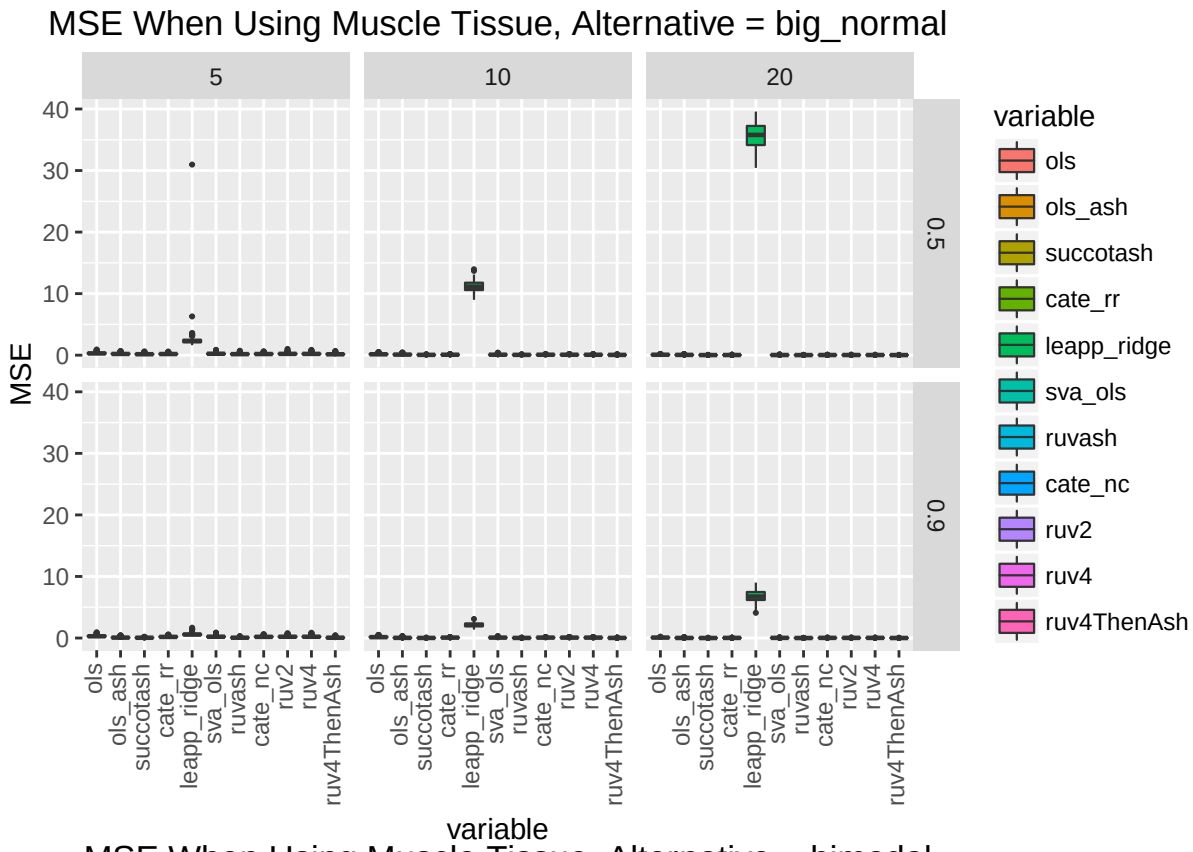

AUC When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

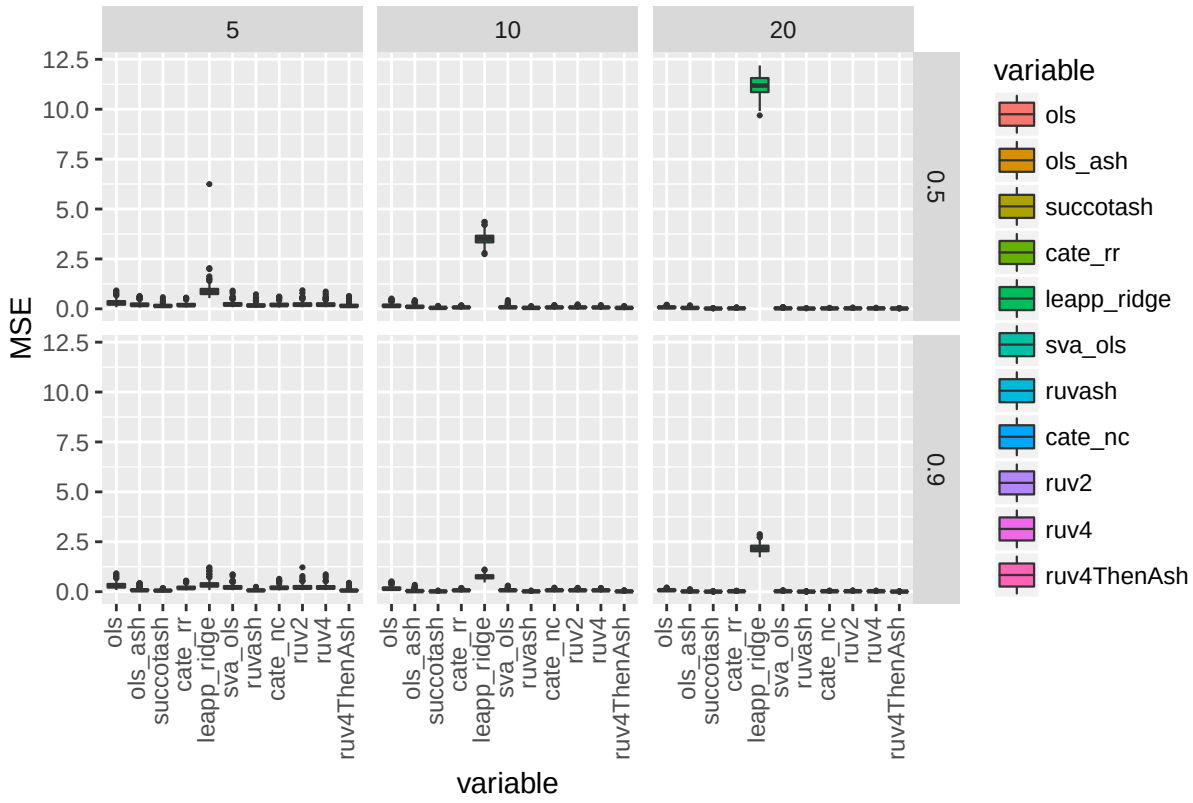

MSE When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh

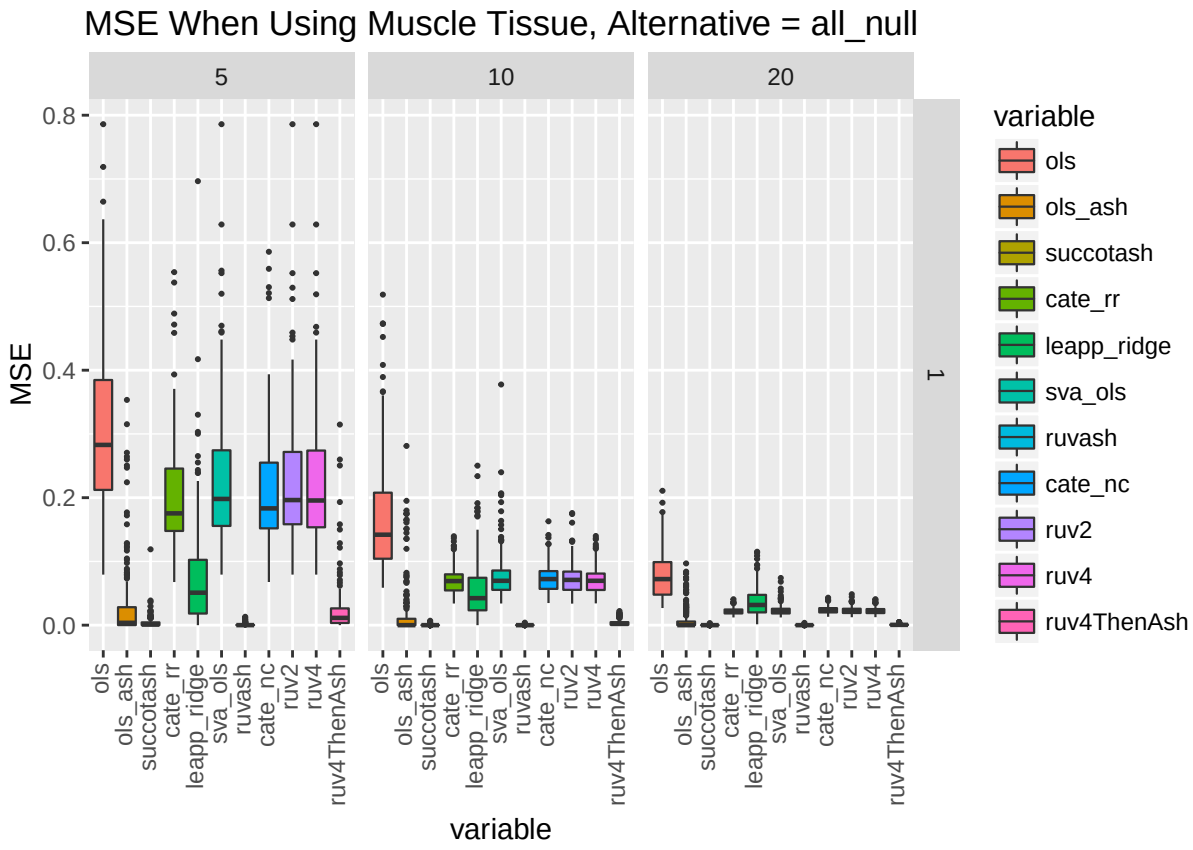

MSE When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
MSE
variable
ols
ols_ash
succotash
cate_rr
leapp_ridge
sva_ols
ruvash
cate_nc
ruv2
ruv4
ruv4ThenAsh




```
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.4 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5          compiler_3.3.0       formatR_1.3
##  [4] plyr_1.8.3           tools_3.3.0          digest_0.6.9
##  [7] RSQLite_1.0.0        annotate_1.48.0      evaluate_0.9
## [10] gtable_0.2.0         cate_1.0.4           nlme_3.1-127
## [13] lattice_0.20-33      mgcv_1.8-12          Matrix_1.2-6
## [16] DBI_0.4              yaml_2.1.13          parallel_3.3.0
```

```
## [19] genefilter_1.52.0    stringr_1.0.0        knitr_1.12.28
## [22] IRanges_2.4.6        S4Vectors_0.8.6      stats4_3.3.0
## [25] grid_3.3.0           Biobase_2.30.0       ruv_0.9.6
## [28] AnnotationDbi_1.32.3 survival_2.39-2      XML_3.98-1.3
## [31] rmarkdown_0.9.6      leapp_1.2            sva_3.18.0
## [34] corpcor_1.6.8        magrittr_1.5         splines_3.3.0
## [37] scales_0.4.0         htmltools_0.3.5      MASS_7.3-45
## [40] BiocGenerics_0.16.1  svd_0.3.3-2          colorspace_1.2-6
## [43] xtable_1.8-2         esaBcv_1.2.1         stringi_1.0-1
## [46] munsell_0.4.3
```

Stephens, Matthew. 2016. "False Discovery Rates: A New Deal." *BioRxiv*. Cold Spring Harbor Labs Journals, 038216.