Different Alternative Types

David Gerard

2016-05-24

Abstract

I compare SUCCOTASH to various competitors under the same alternative scenarios in Stephens (2016). SUCCOTASH generally has superior performance in terms of (1) estimating π_0, (2) having higher AUC, and (3) having lower MSE. LEAPP does better under some scenarios in terms of estimating π_0 but has far worse AUC and MSE. SUCCOTASH accurately estimates the average sign error rate, though not in as conservative manner as one might hope in some scenarios.

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 the methods of Buja and Eyuboglu (1992) implemented in the `num.sv()` function in the `sva` package in `R`.

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

- OLS + qvalue.
- RUVols + estimate variance inflation using controls + ASH
- RUVgls + estimate variance inflation using controls + ASH
- RUVgls + estimate variance inflation using controls + MLE to UMVUE motivated scaling + 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.
- SVA + qvalue.
- Negative control version of CATE using PCA as the factor analysis method + qvalue.
- RUV2 + qvalue.
- RUV4 + qvalue.
- Sparse version of LEAPP. Since this is a sparsity-inducing procedure, I used the proportion of zeros as the estimate of π_0.
- Ridge version of LEAPP + qvalue.

RUVASH

The model for the second step in RUV4, LEAPP, CATE, and SUCCOTASH is

$$Y = \beta_{p\times 1} + \hat{\alpha}_{p\times q} Z_{q\times 1} + E_{p\times 1} \tag{9}$$
$$E \sim N_p(0, \lambda\hat{\Sigma}) \tag{10}$$
$$\hat{\Sigma} = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2), \tag{11}$$

with $\hat{\Sigma}$ and $\hat{\alpha}$ known. λ is a variance inflation parameter that is novel to SUCCOTASH.

Let $\mathcal{C} \subseteq \{1, \ldots, p\}$ denote the indices for the negative controls. I.e. $\beta_{\mathcal{C}} = 0$. Then the model for the negative controls is

$$Y_{\mathcal{C}} = \alpha_{\mathcal{C}} Z + E_{\mathcal{C}}. \tag{12}$$

The maximum likelihood estimates of Z and λ are easy:

$$\hat{Z} = (\alpha_{\mathcal{C}}^T \hat{\Sigma}^{-1} \alpha_{\mathcal{C}})^{-1} \alpha_{\mathcal{C}}^T \hat{\Sigma}^{-1} Y_{\mathcal{C}}. \tag{13}$$

Let $R_{\mathcal{C}} = Y_{\mathcal{C}} - \hat{\alpha}_{\mathcal{C}} \hat{Z}$. Then the MLE of λ is

$$\hat{\lambda} = \frac{1}{|\mathcal{C}|} \sum_{i\in\mathcal{C}} R_i^2/\hat{\sigma}_i^2. \tag{14}$$

RUVASH runs ASH with means $\hat{\beta} = Y - \hat{\alpha}\hat{Z}$ and variances $\hat{\lambda}\hat{\sigma}_i^2$.

In the same way that I made an ad-hoc correction to the SUCCOTASH variance inflation parameter, I also explore using this for RUVASH. Specifically, I set

$$\tilde{\lambda} = \frac{n}{n-k-q}\hat{\lambda} \tag{15}$$

and run ASH with variances $\tilde{\lambda}\hat{\sigma}_i^2$. These are called SRUVASH in the plots below.

Results

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

Estimates of π_0

- RUV + ASH Works much better here than vanilla ASH at estimating π_0, but it is slightly anti-conservative.
- RUV + ad-hoc scaling + ASH has conservative FDR with high probability in almost every scenario. The only one where it isn't conservative is the Bimodal alternative scenario.

- 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.
- LEAPP does amazingly well in the Big-normal and bimodal scenarios, even for $n = 10$. However, it is far too conservative in every other non-null scenario. This seems to indicate that LEAPP functions best if there is a separation of the alternative signal from zero.
- No method using qvalue ever performed as well as succotash. Indeed, none exhibited this "conservative with high-probability" behavior that is desirable.

AUC performance.

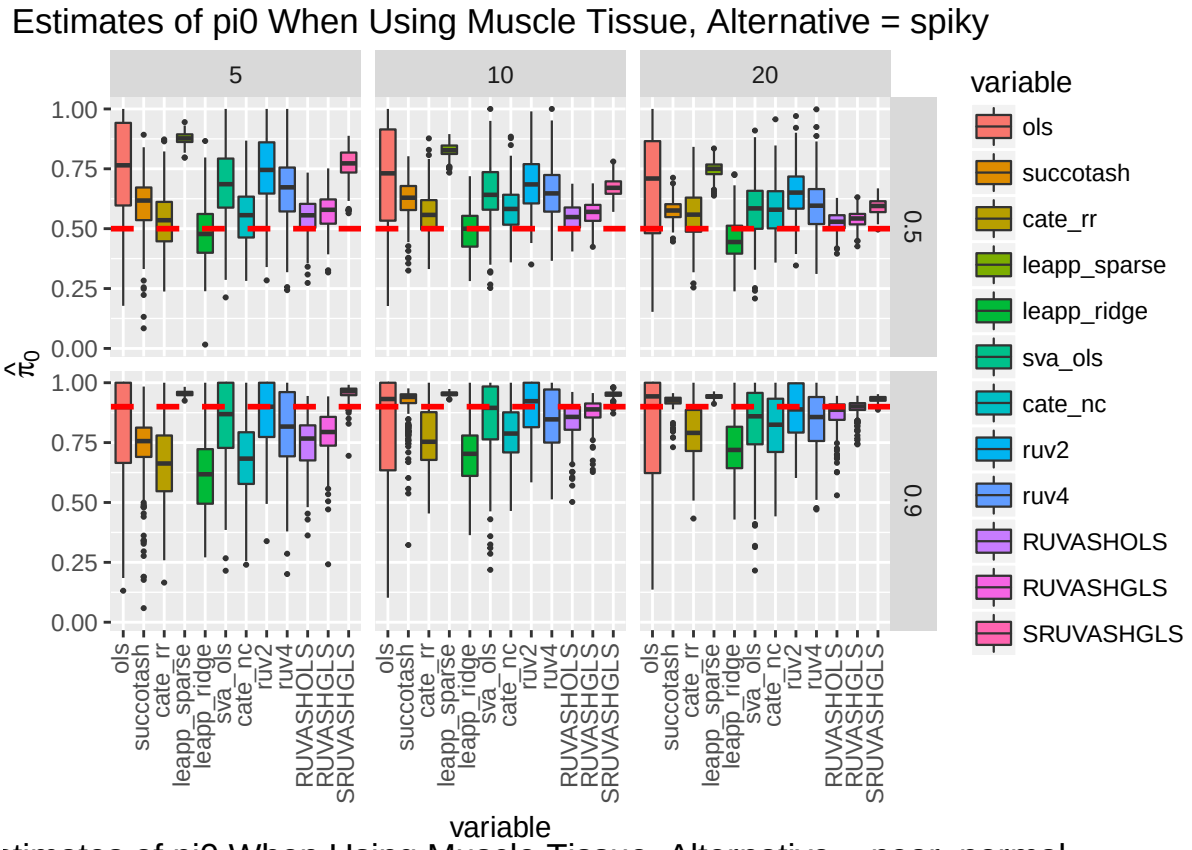
- SUCCOTASH always has higher AUC, even in the bimodal and big-normal scenarios where LEAPP estimated π_0 more accurately.
- All of the RUV + ASH methods have about the same AUC as SUCCOTASH — including the one with the ad-hoc scaling value that works so well with estimating π_0.

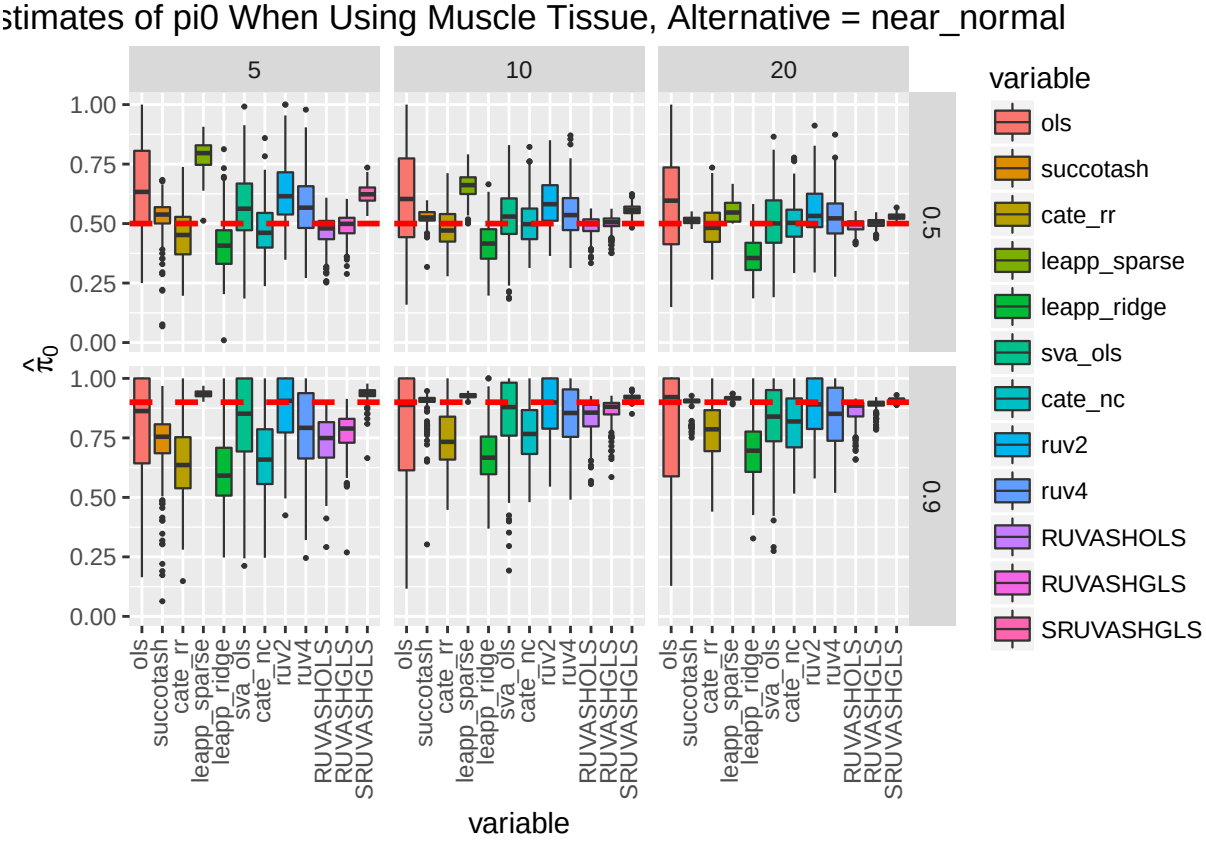
ROC

I calculated the point-wise mean true positive rate (TPR) and false positive rate (FPR) at each ordered position under each scenario. These are plotted after the AUC boxplots. SUCCOTASH and the RUV + ASH methods always have higher and further left point-wise average ROC curves, even under the bimodal scenario.

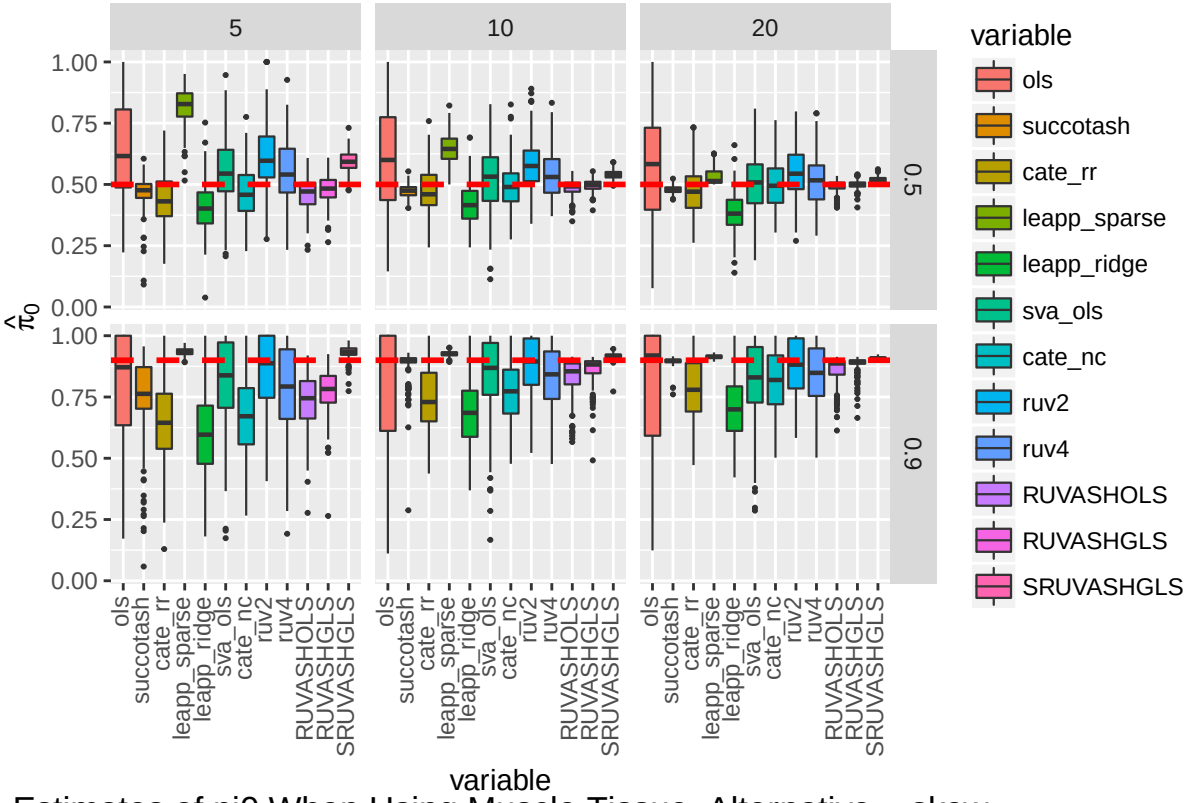
MSE

- SUCCOTASH has superior performance in term of MSE compared to other methods.
- LEAPP has terrible MSE performance in all cases except the all-null setting.
- Even then, SUCCOTASH has better performance.
- RUV + ASH has pretty good MSE performance.
- The ad-hoc scaling version of RUV + ASH seems to suffer in MSE a little bit, especially at low sample sizes.

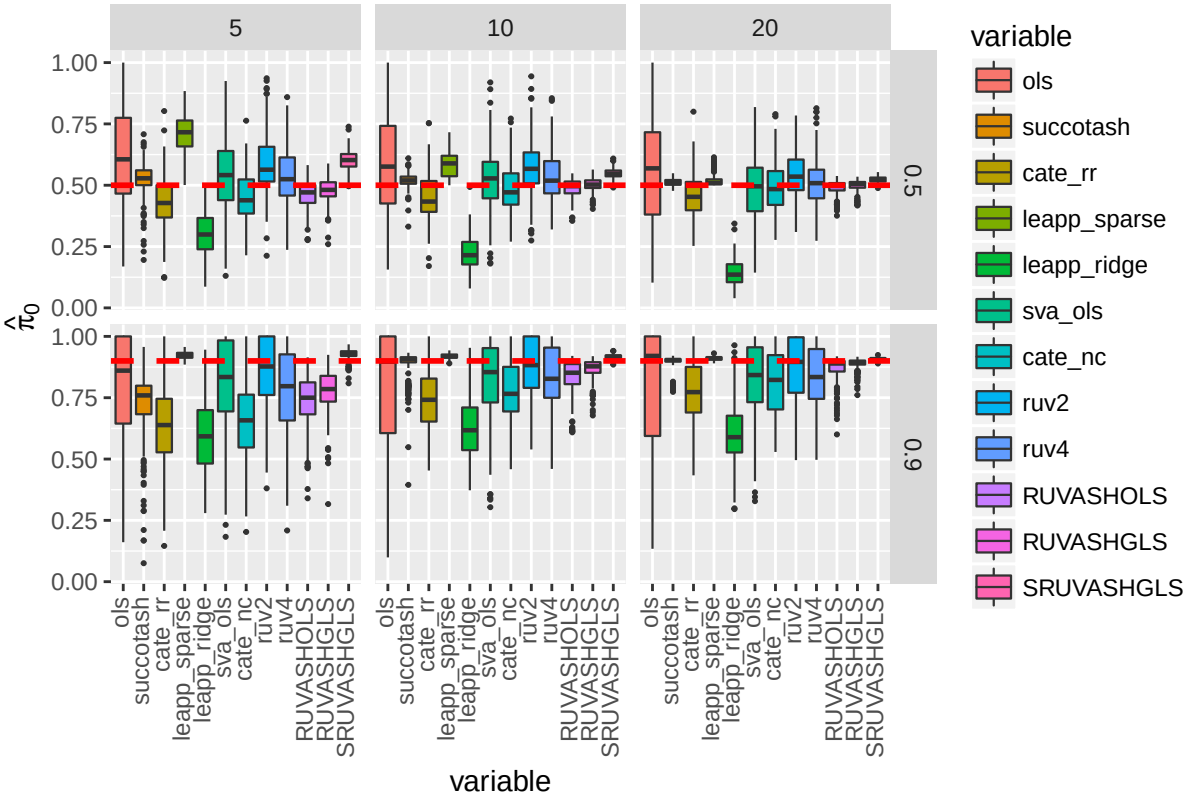

Estimates of pi0 When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc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


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
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

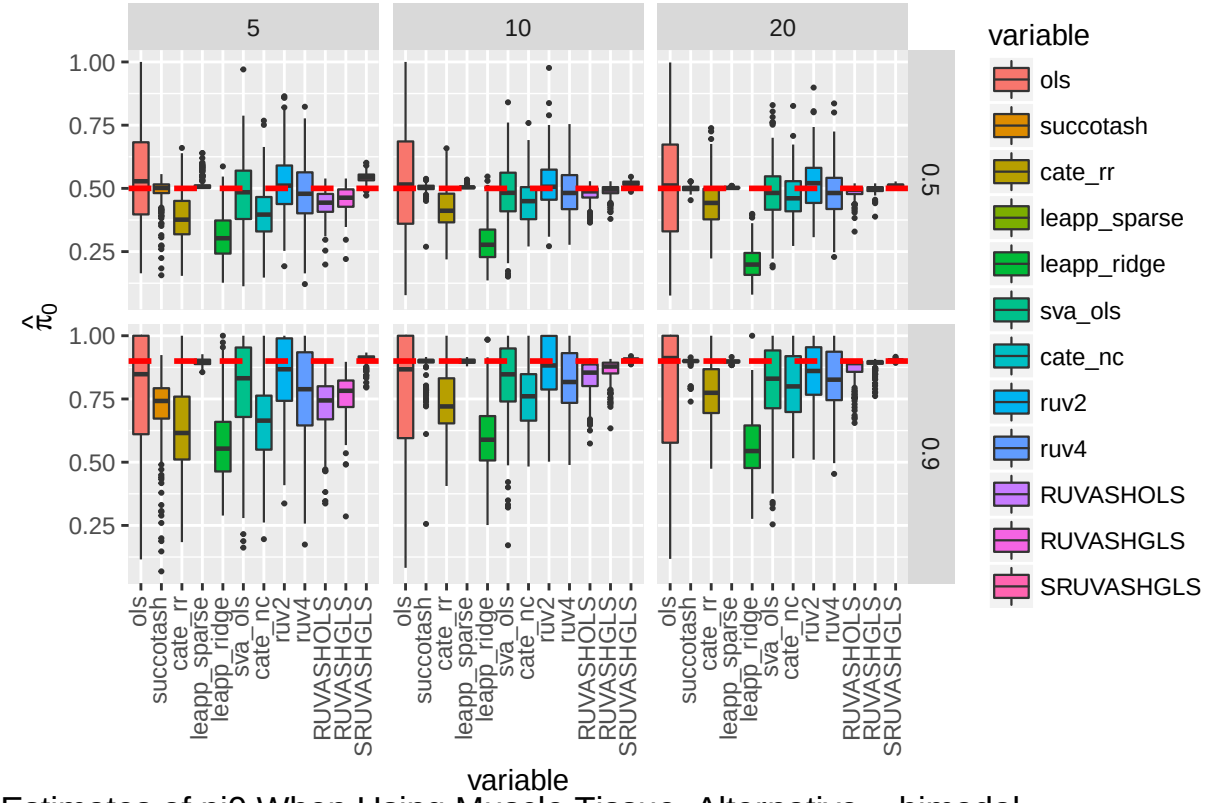
Estimates of pi0 When Using Muscle Tissue, Alternative = flattop



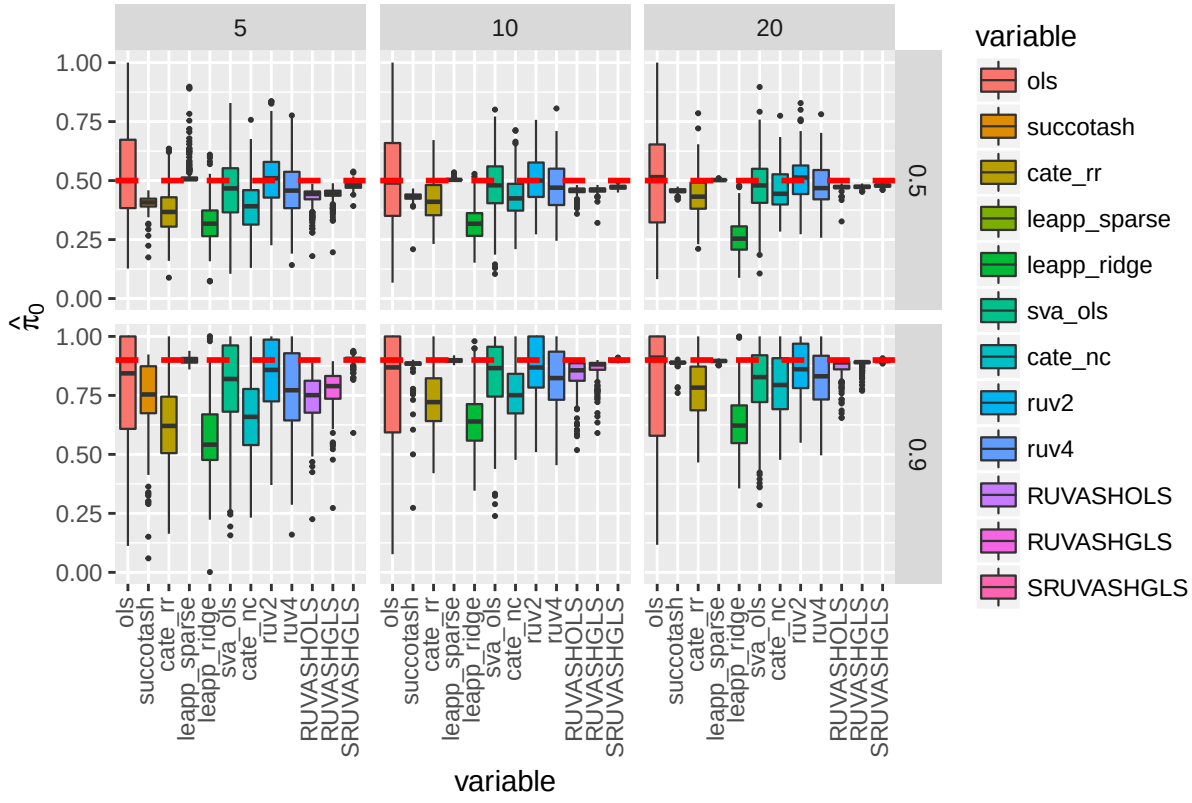
Estimates of pi0 When Using Muscle Tissue, Alternative = skew

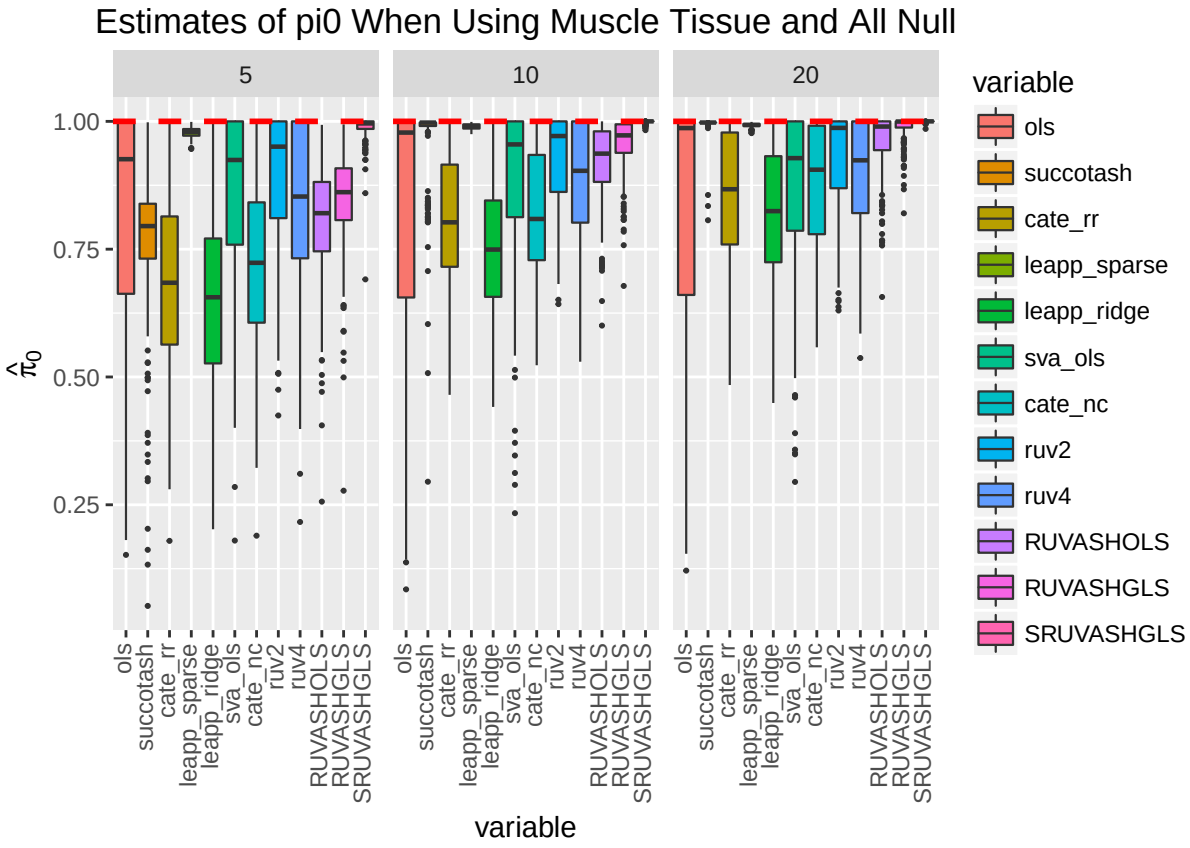


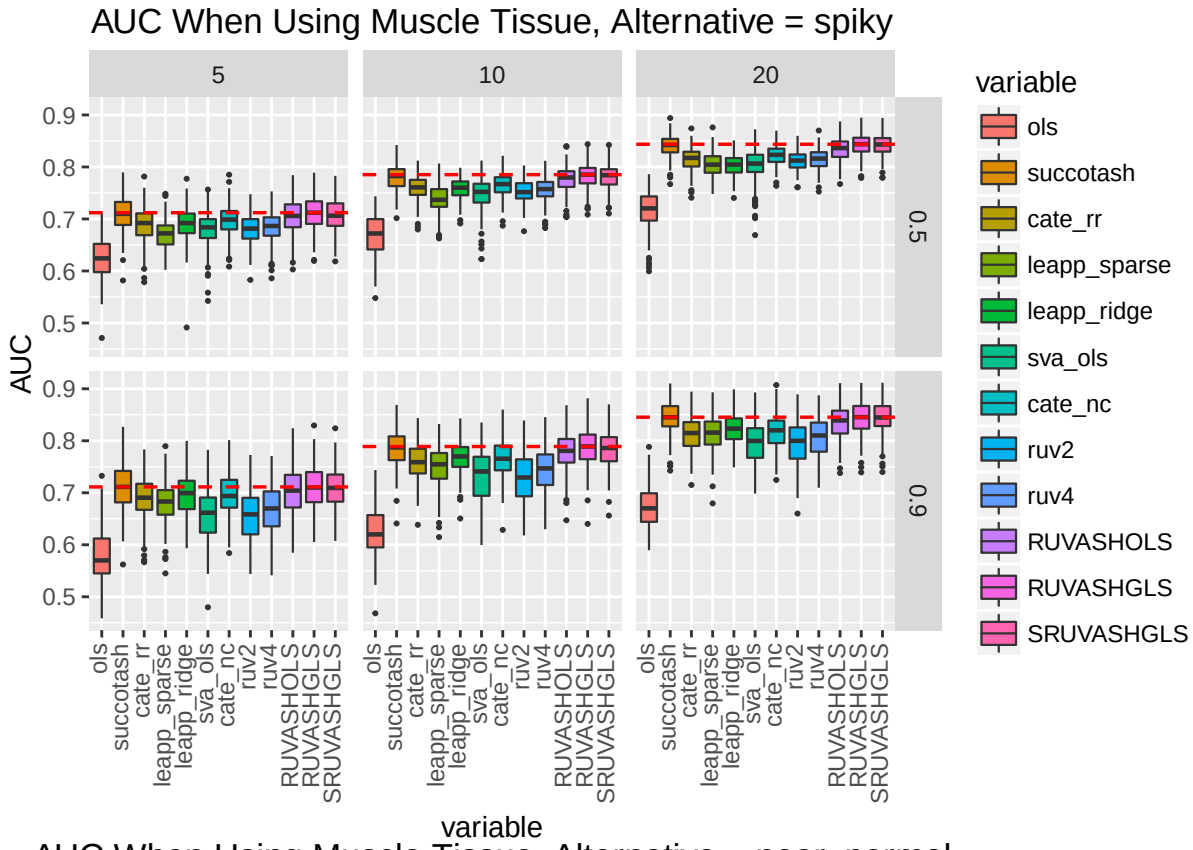
Estimates of pi0 When Using Muscle Tissue, Alternative = big_normal

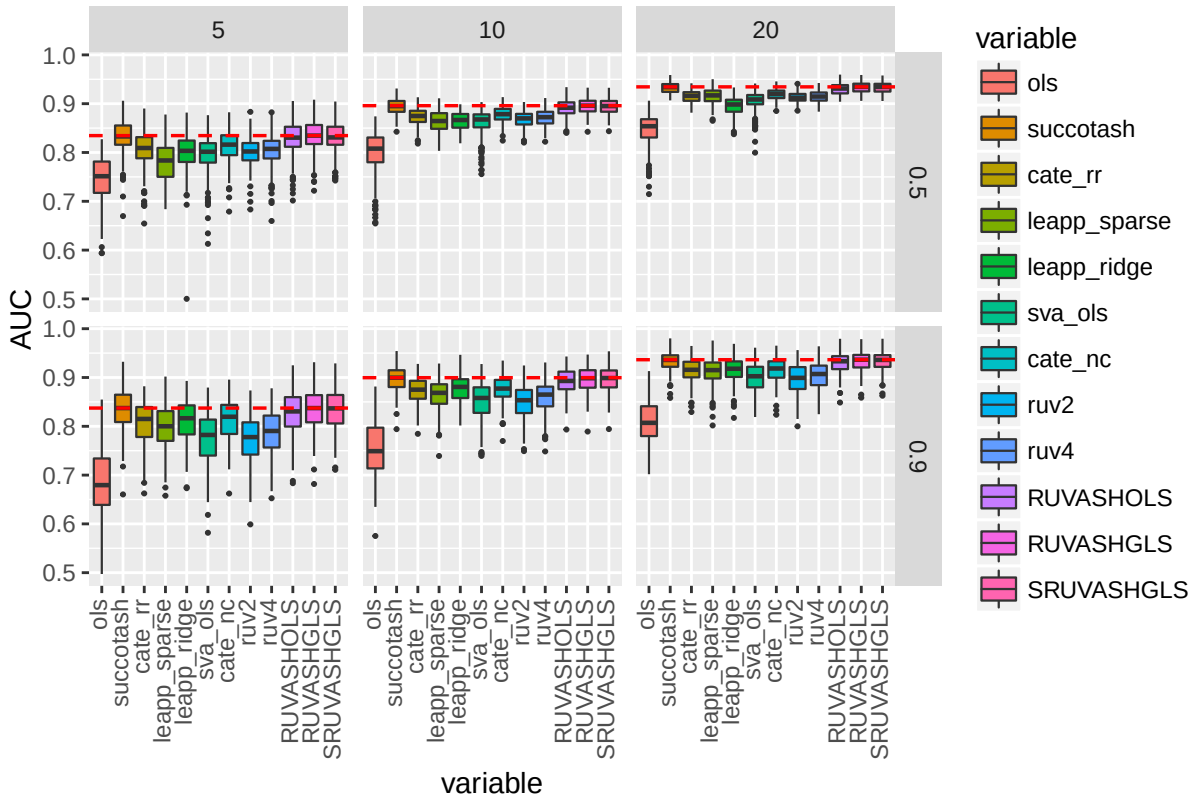


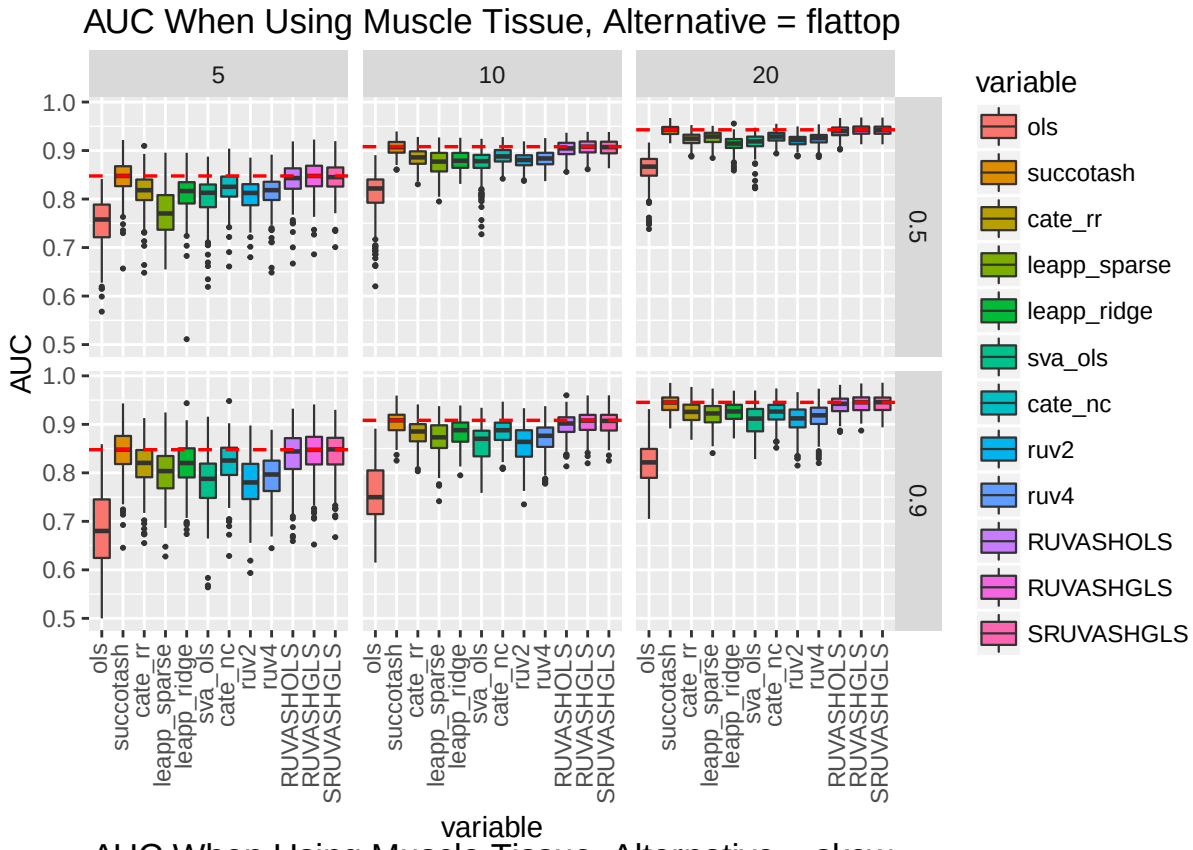
Estimates of pi0 When Using Muscle Tissue, Alternative = bimodal

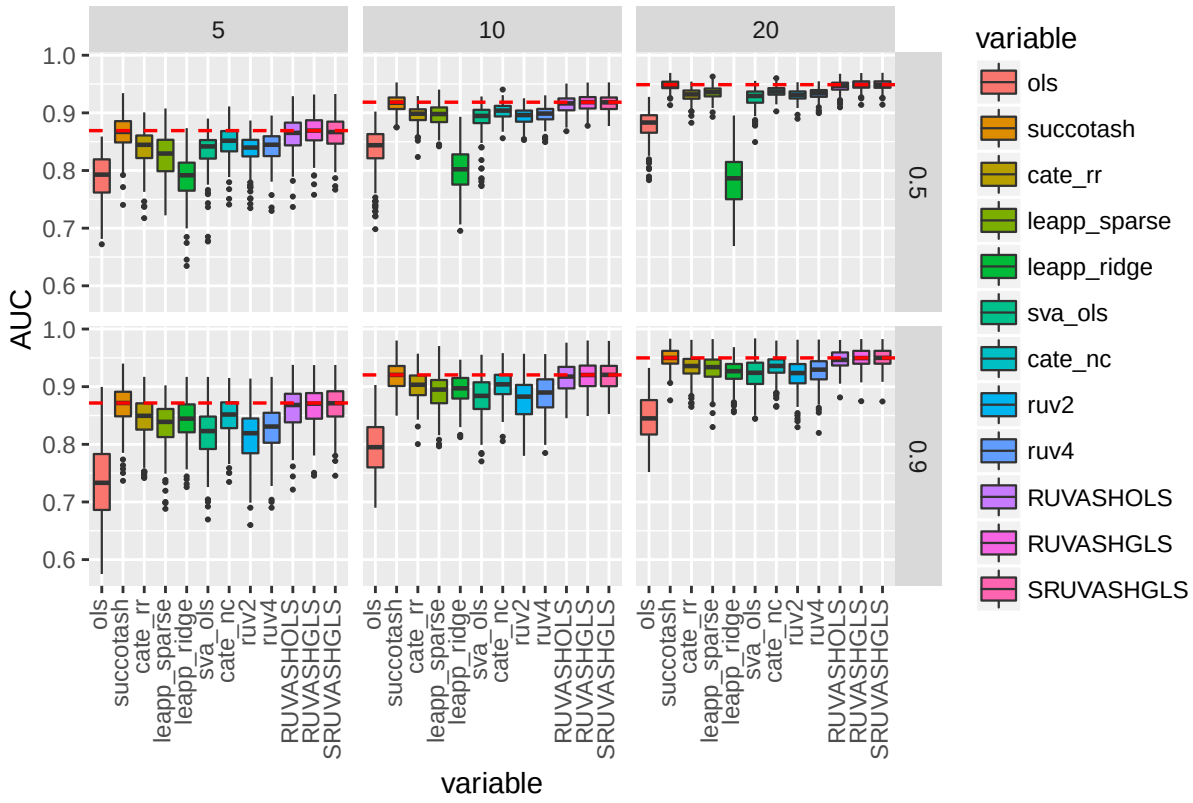


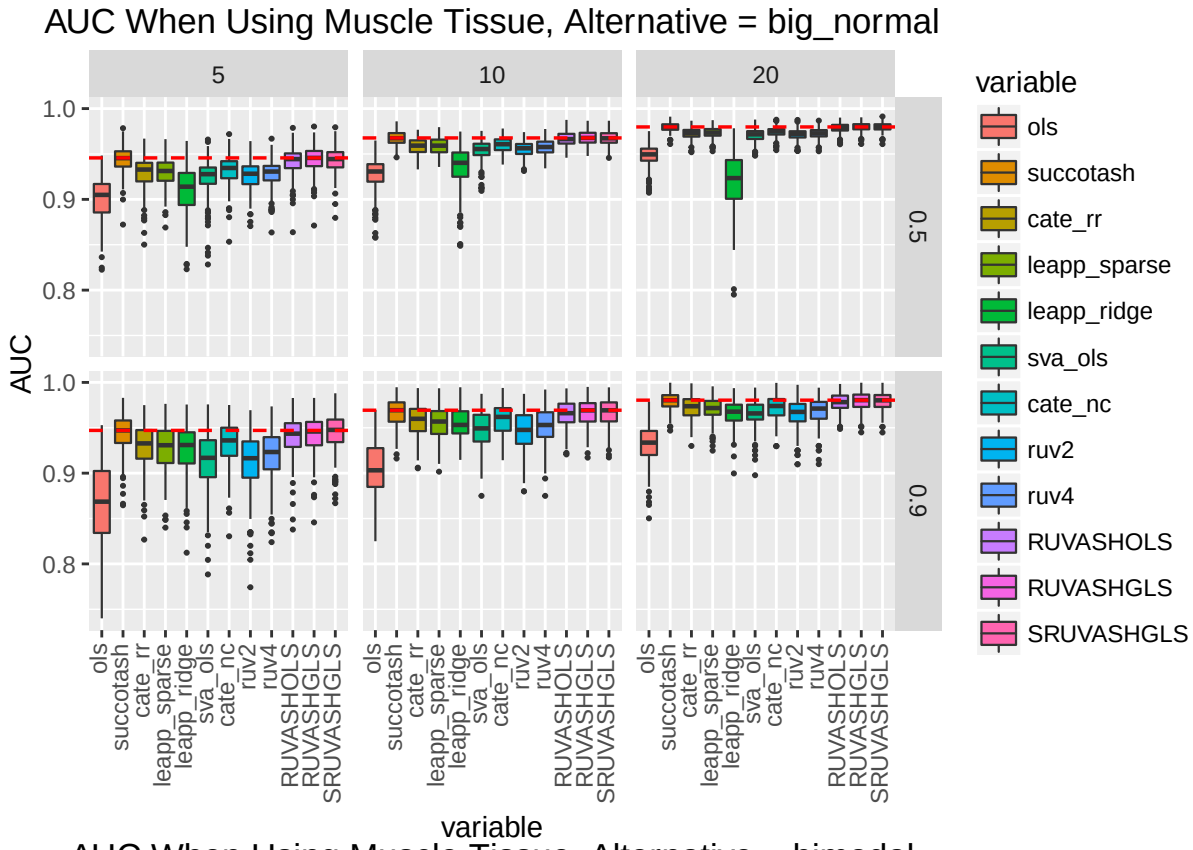

Estimates of pi0 When Using Muscle Tissue and All Null
5
10
20
1.00
0.75
0.50
0.25
$\hat{\pi}_0$
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS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

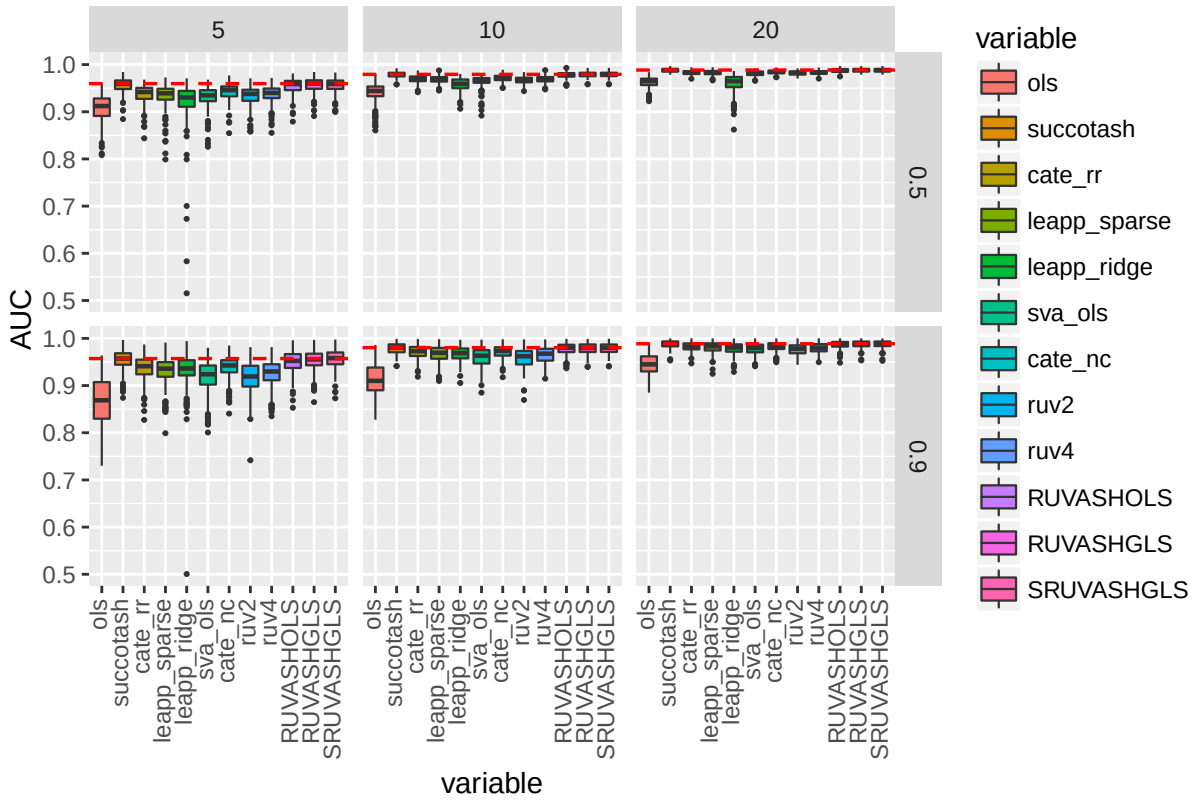

AUC When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
AUC
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

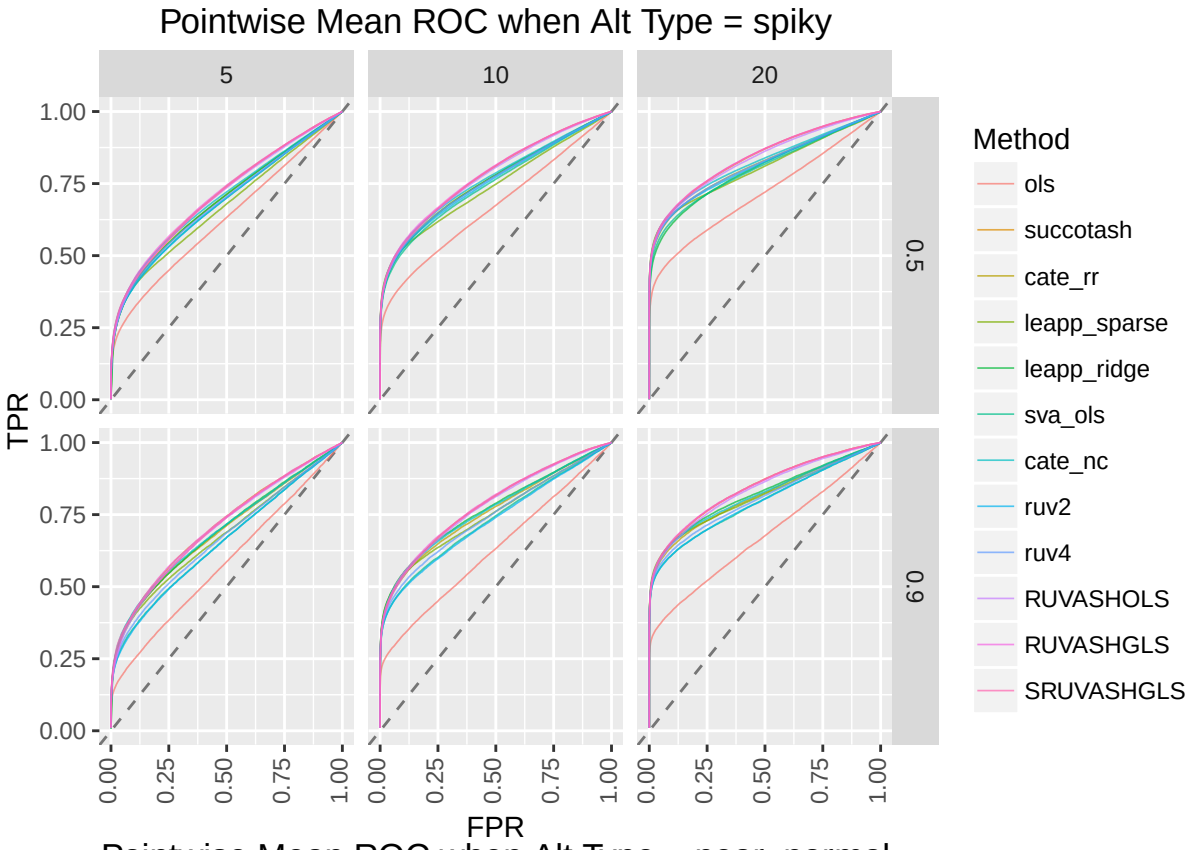

AUC When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
AUC
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

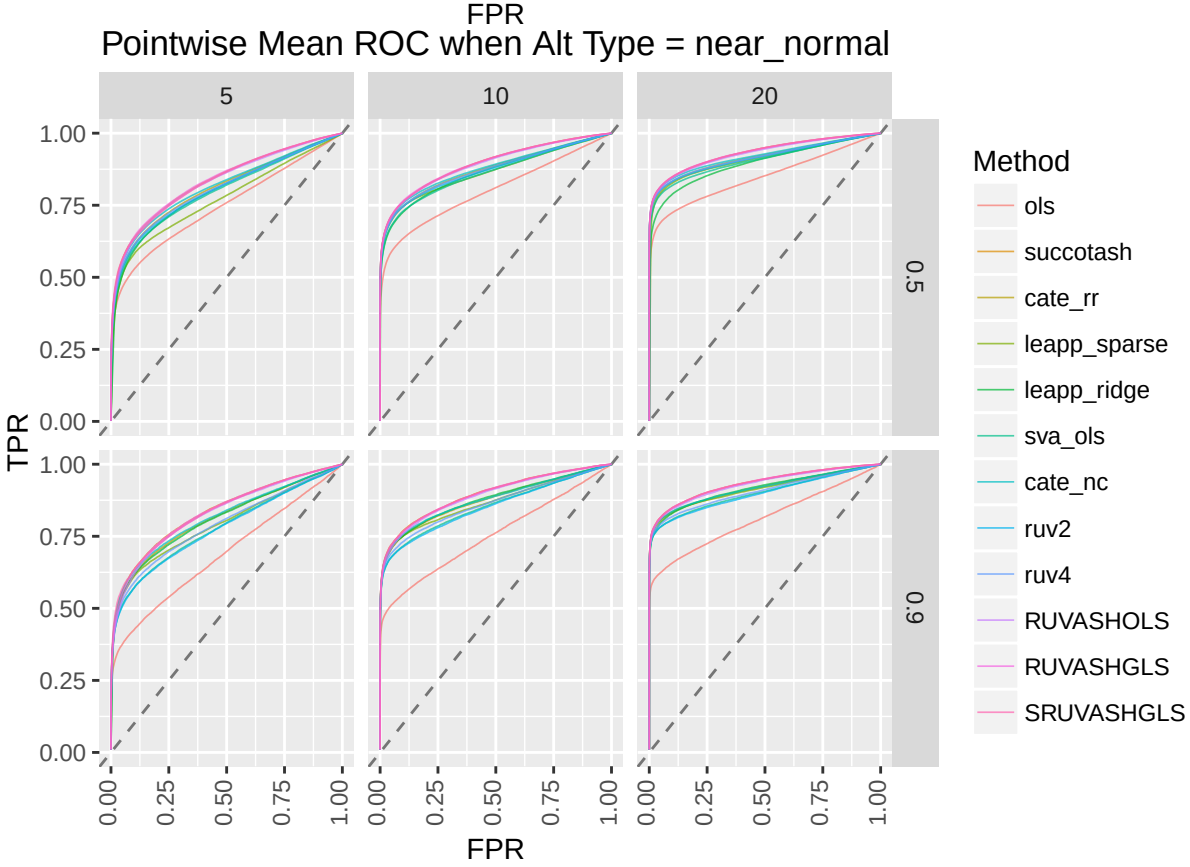

AUC When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

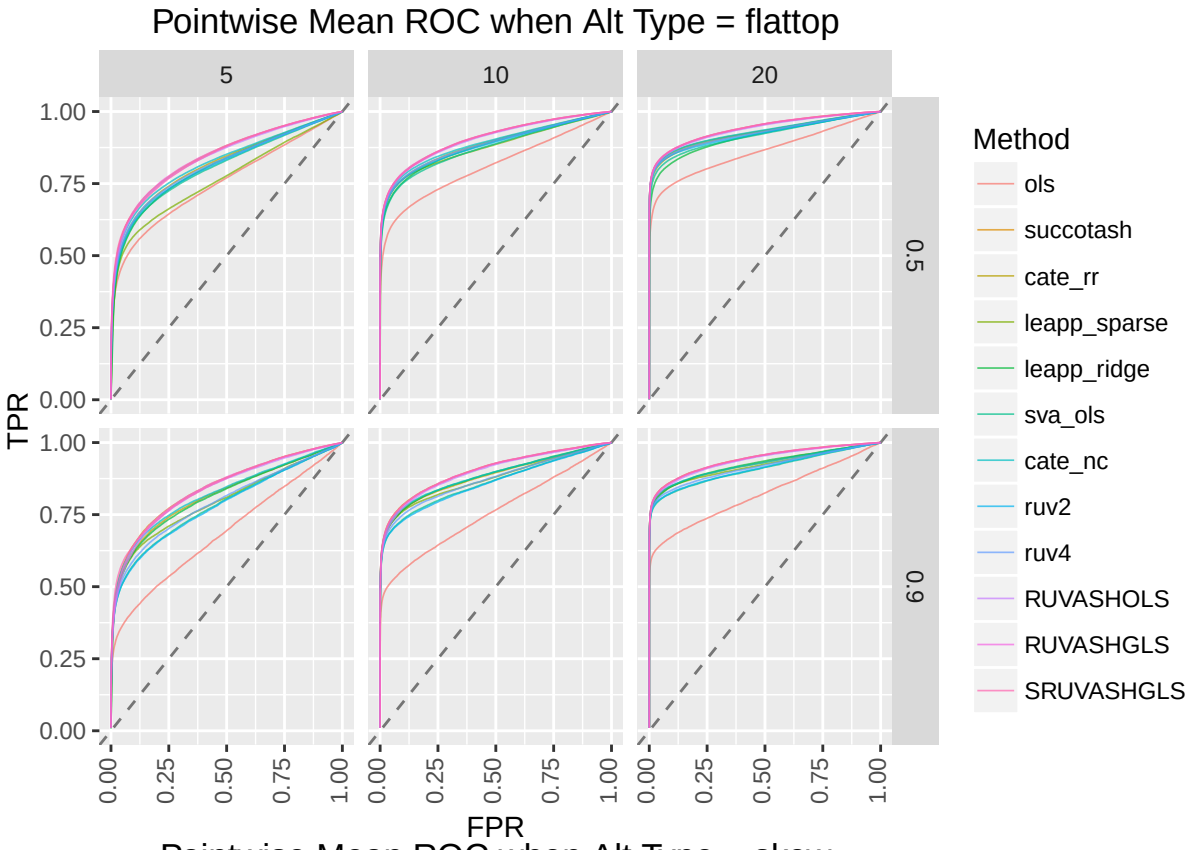

AUC When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

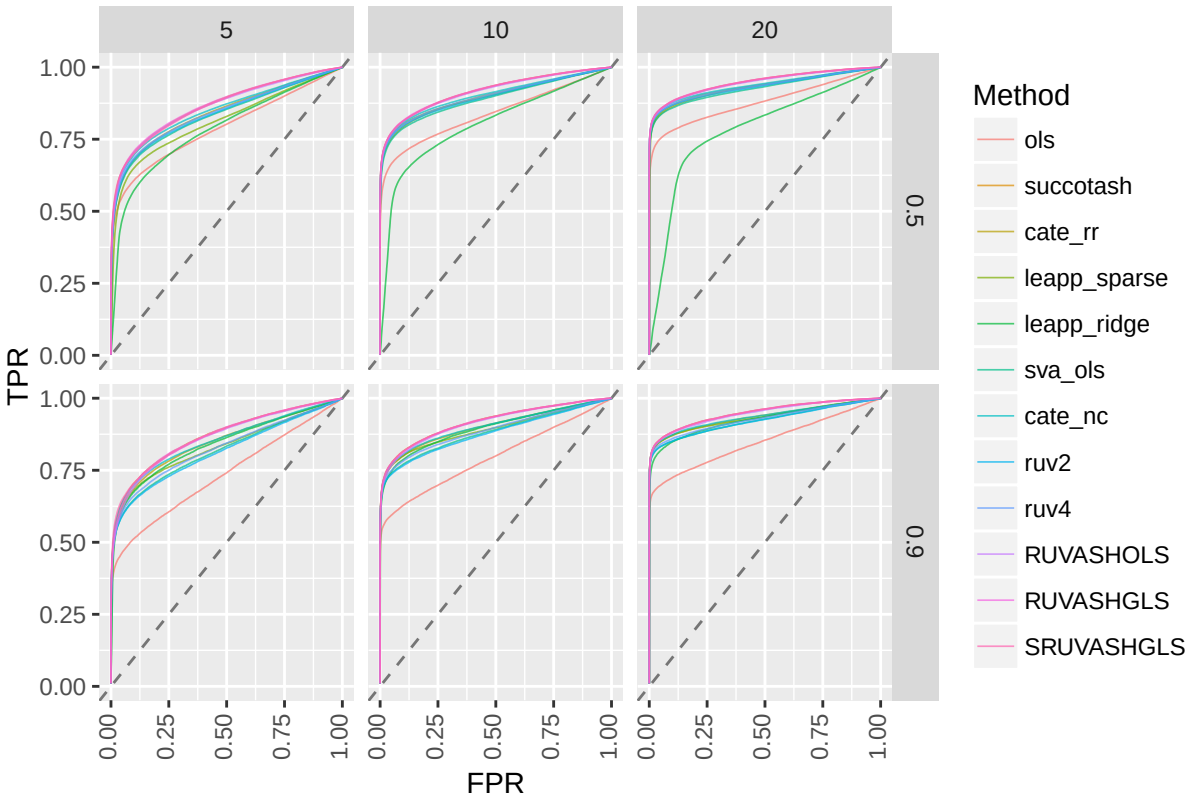

AUC When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

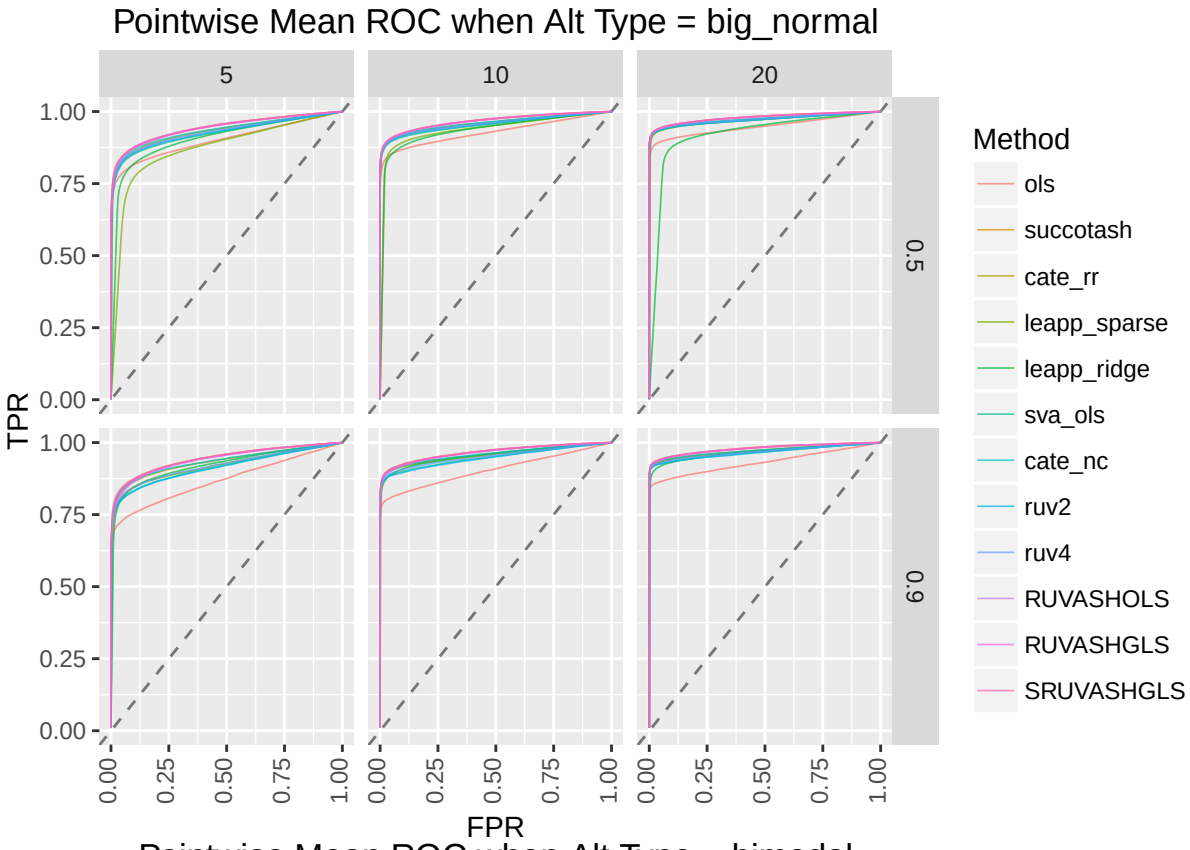

AUC When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

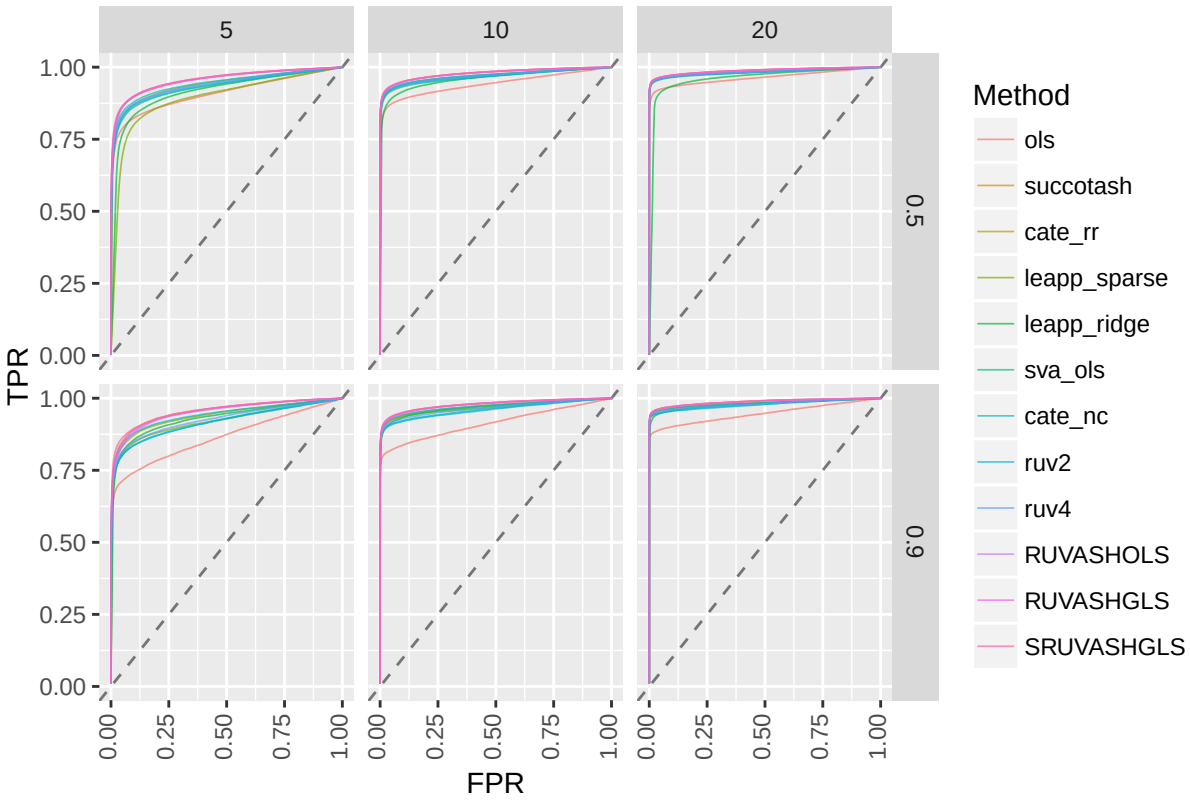

Pointwise Mean ROC when Alt Type = spiky
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

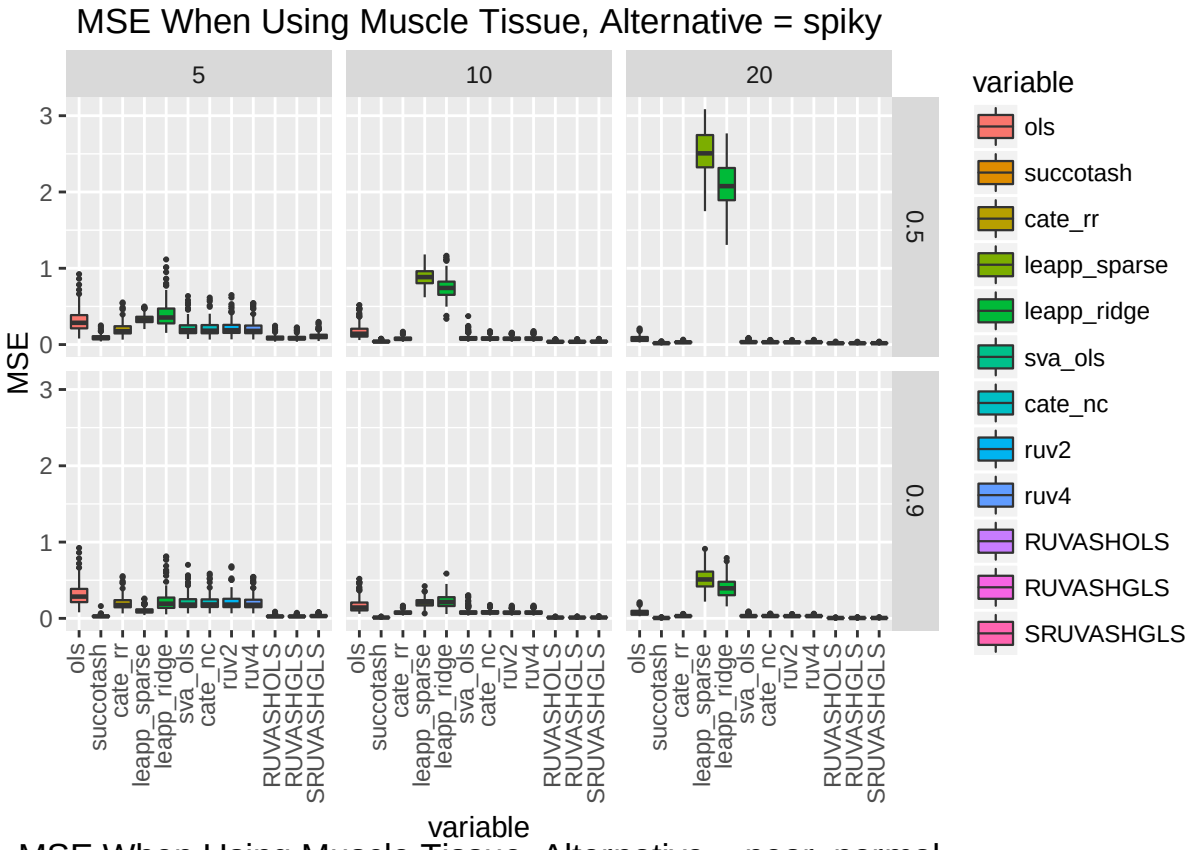

Pointwise Mean ROC when Alt Type = near_normal
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

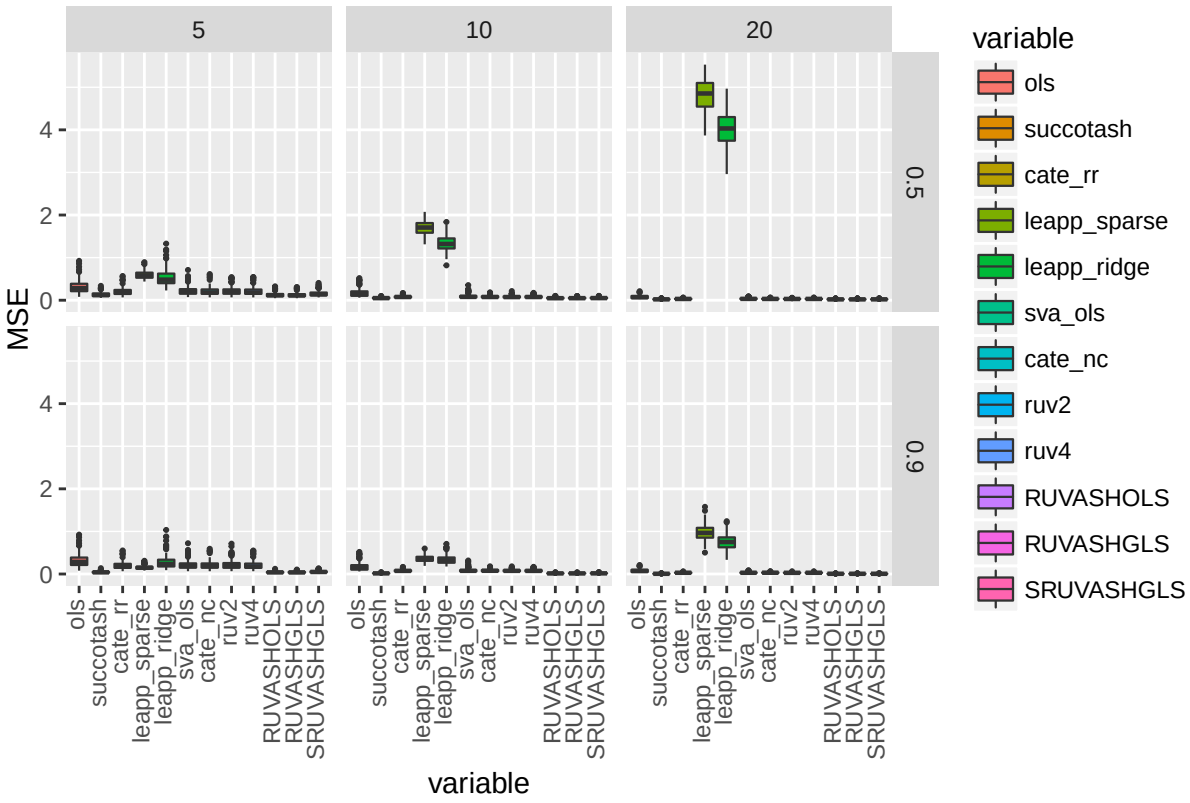

Pointwise Mean ROC when Alt Type = flattop
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

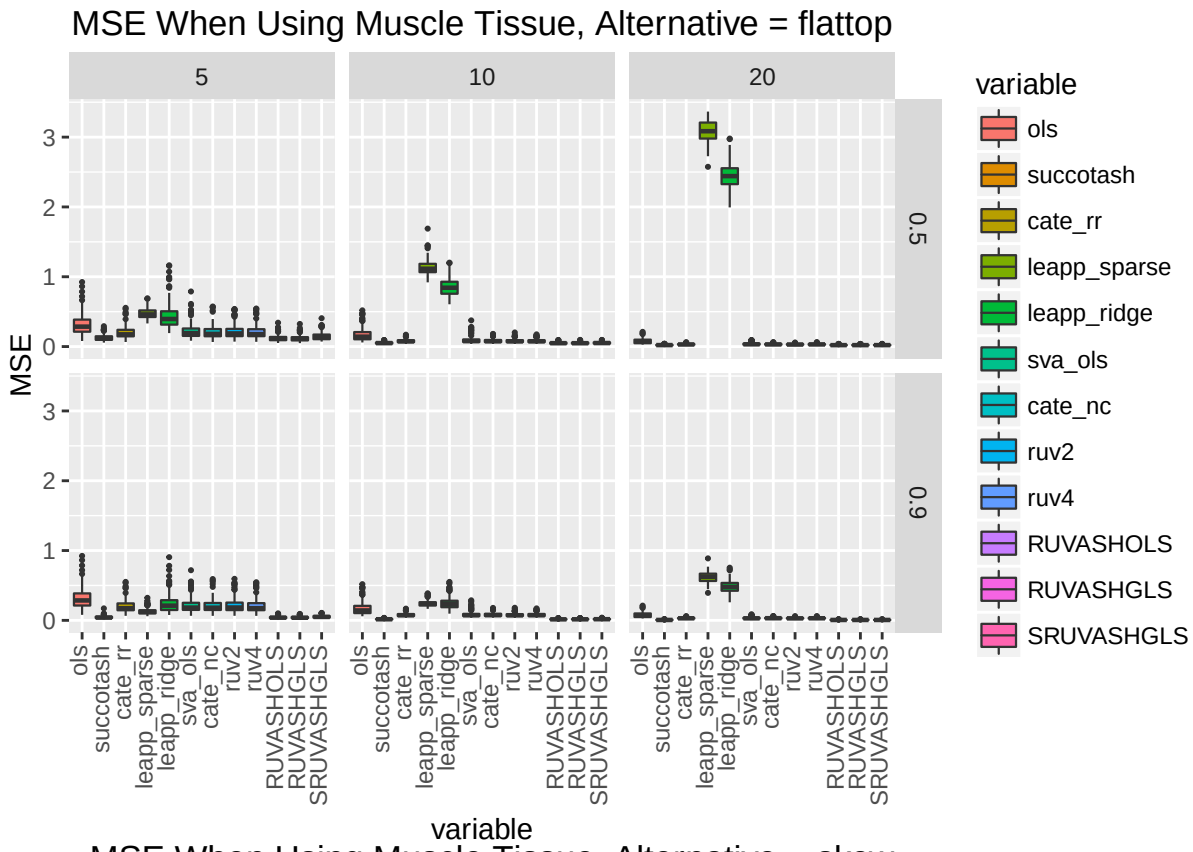

Pointwise Mean ROC when Alt Type = skew
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

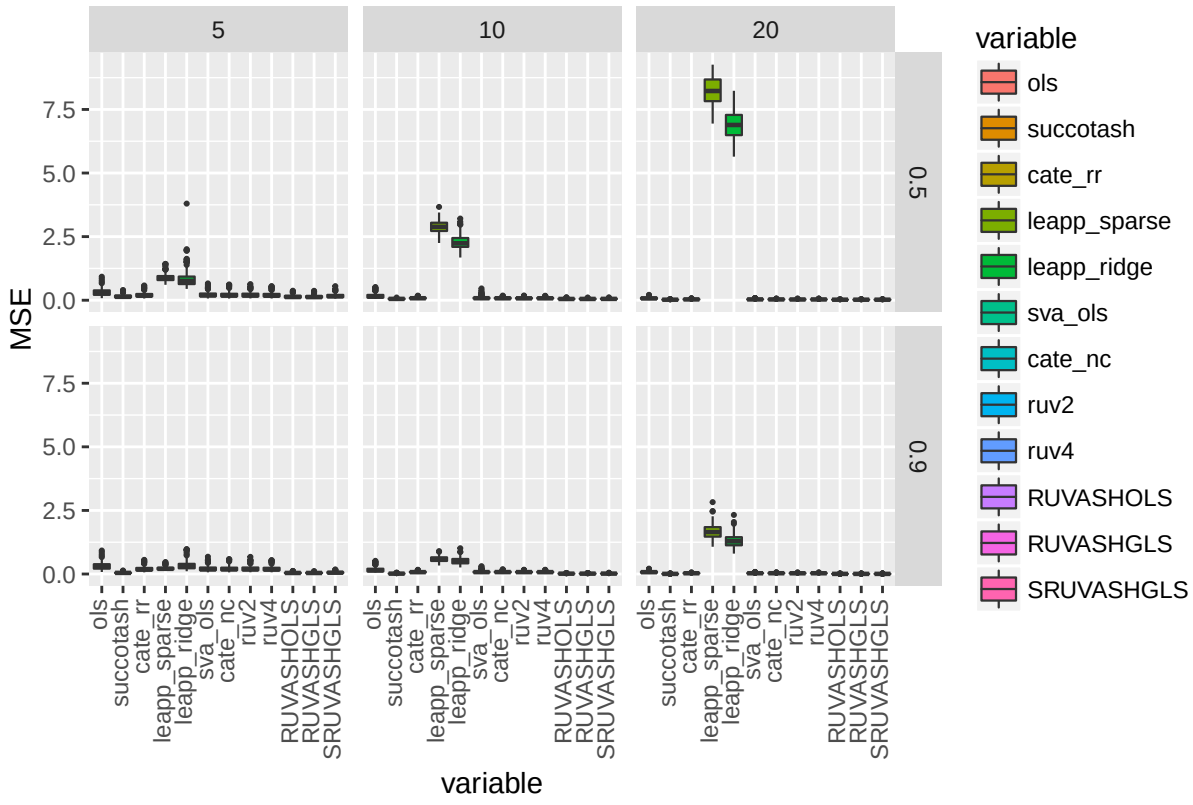

Pointwise Mean ROC when Alt Type = big_normal
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

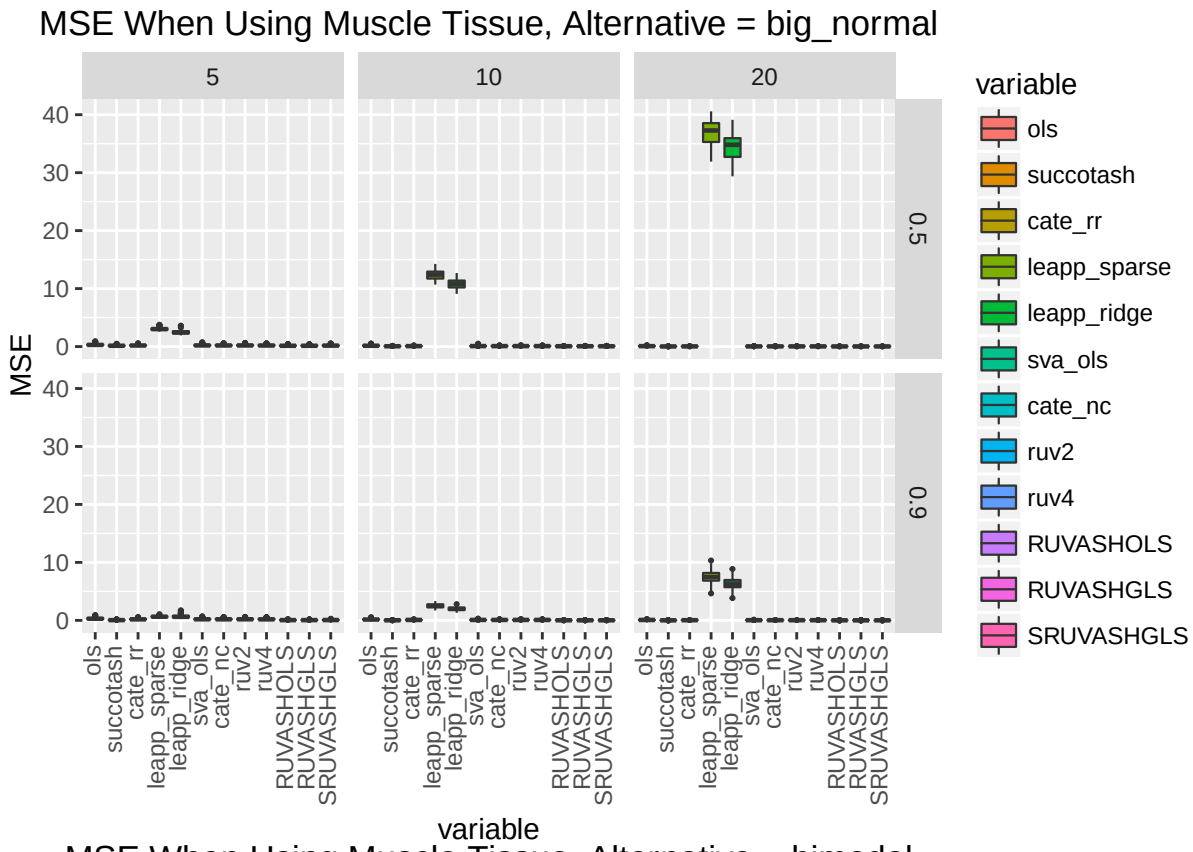

Pointwise Mean ROC when Alt Type = bimodal
5
10
20
0.5
0.9
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Method
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

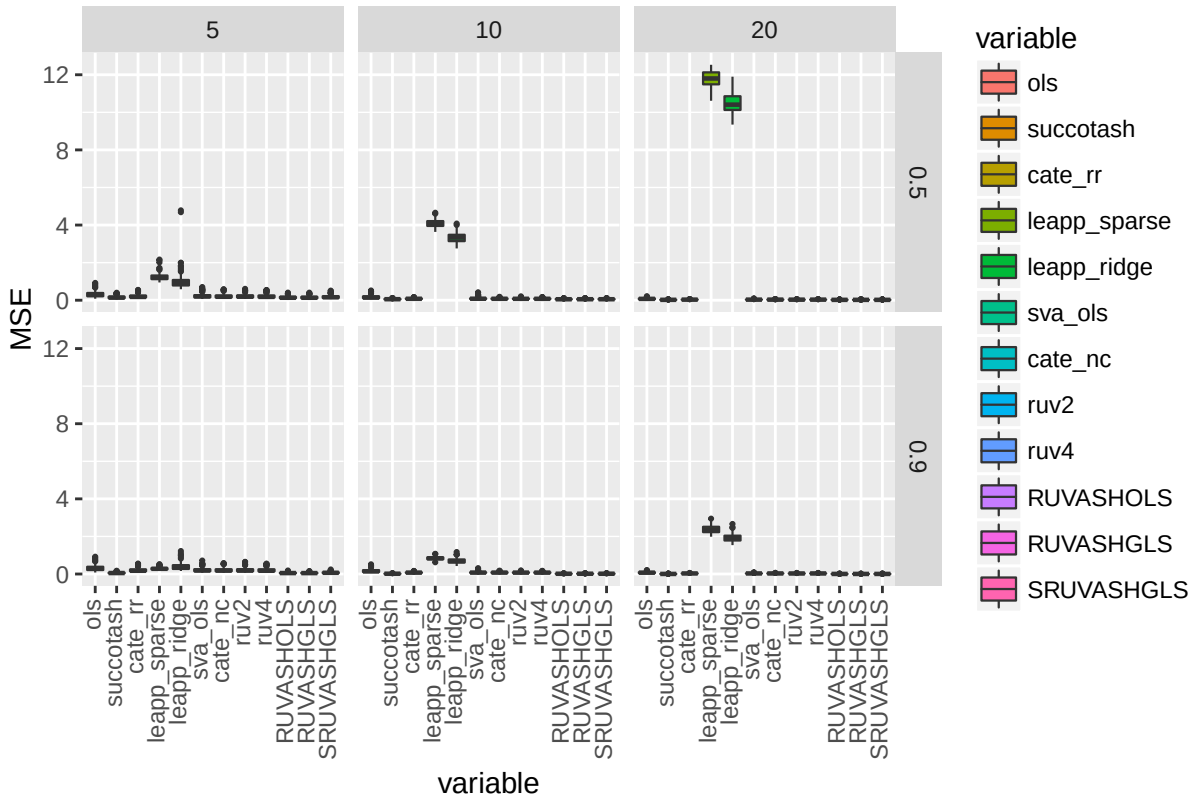

MSE When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
MSE
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

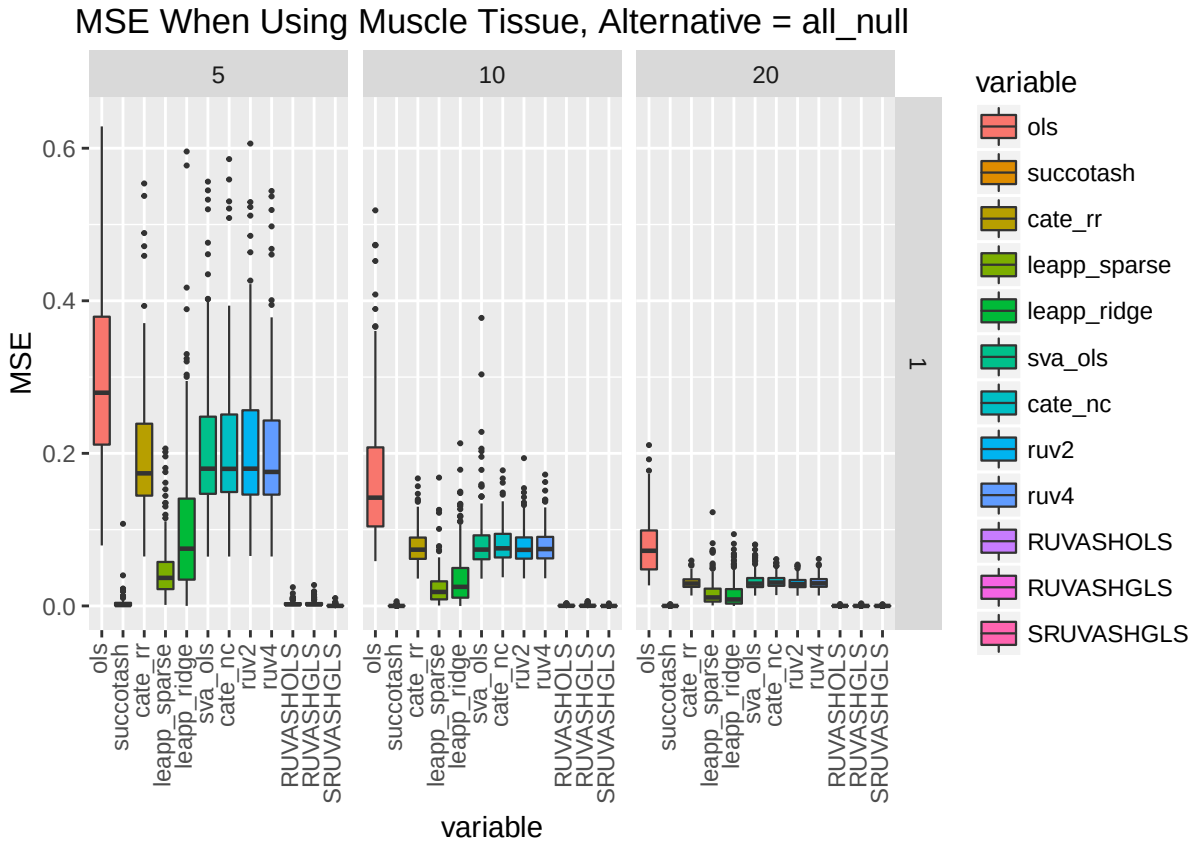

MSE When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
MSE
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS


MSE When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
MSE
0
1
2
3
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

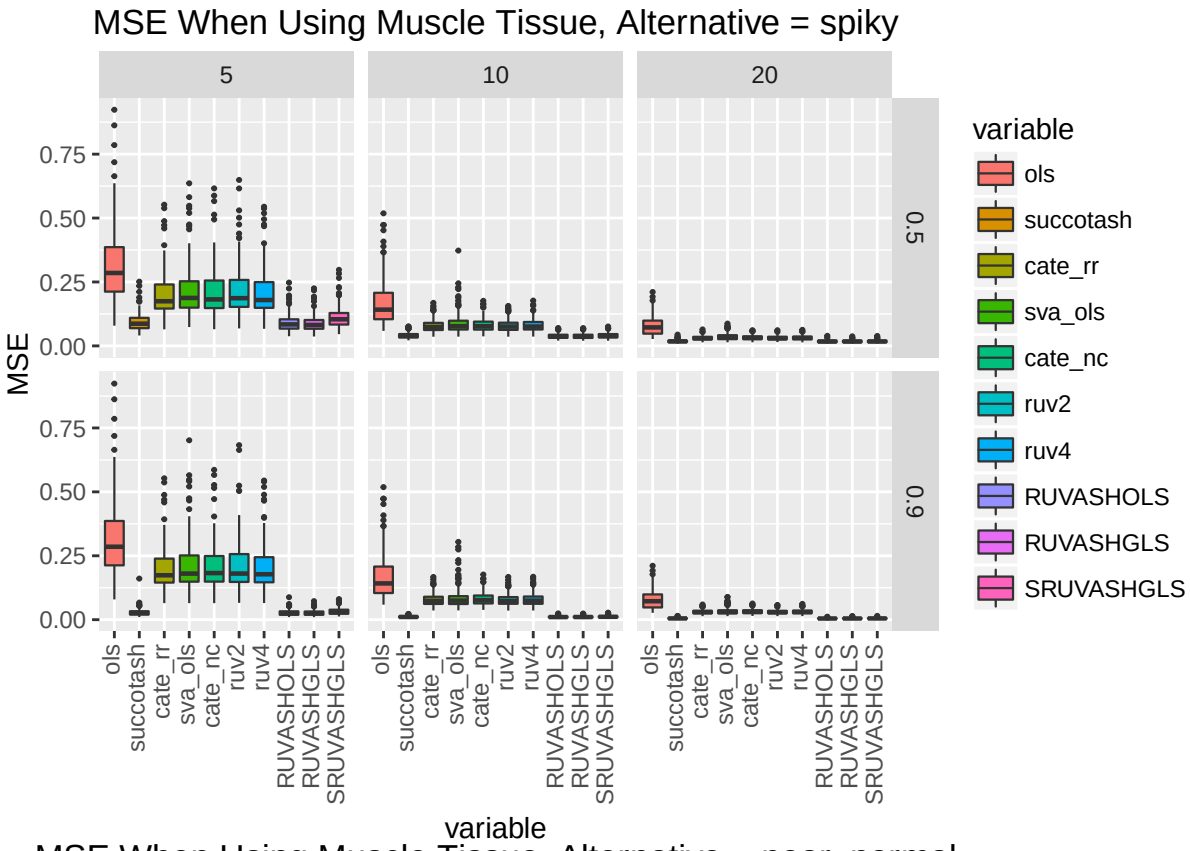

MSE When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
MSE
0.0
2.5
5.0
7.5
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

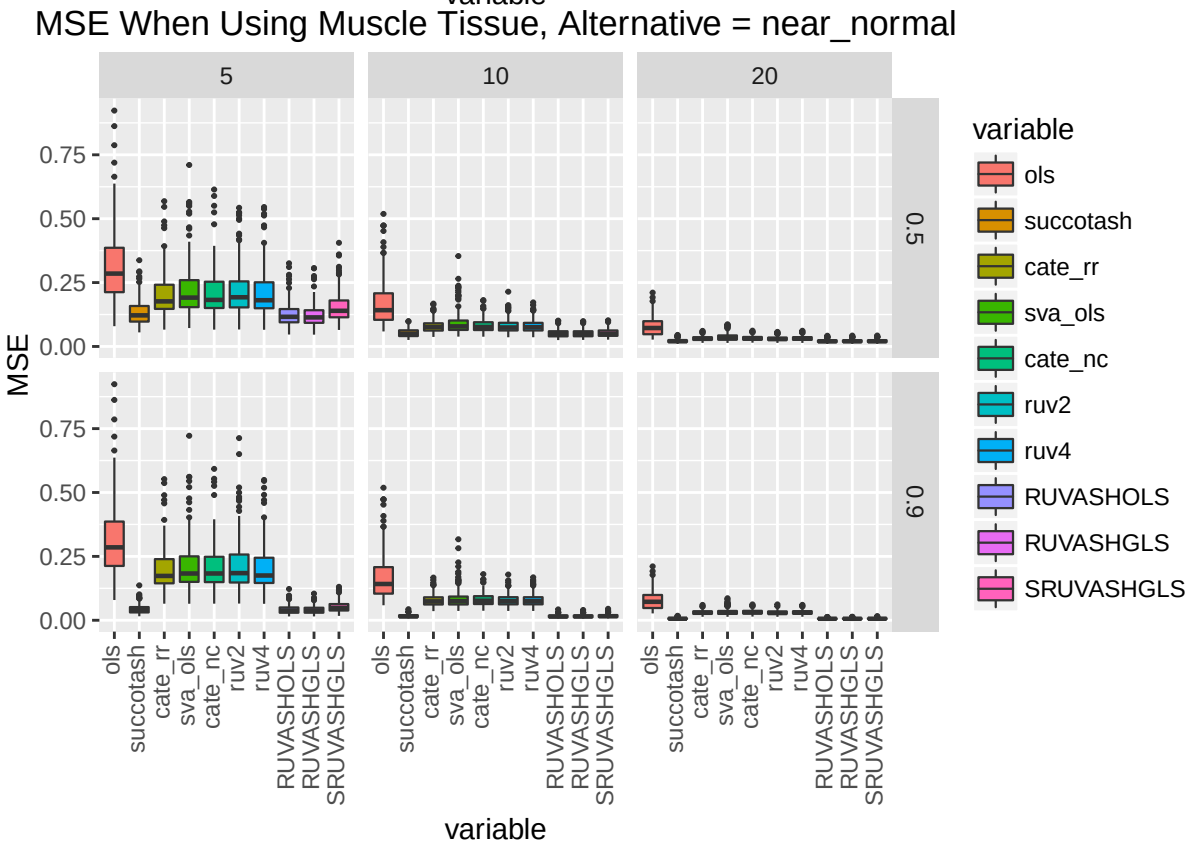

MSE When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
MSE
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

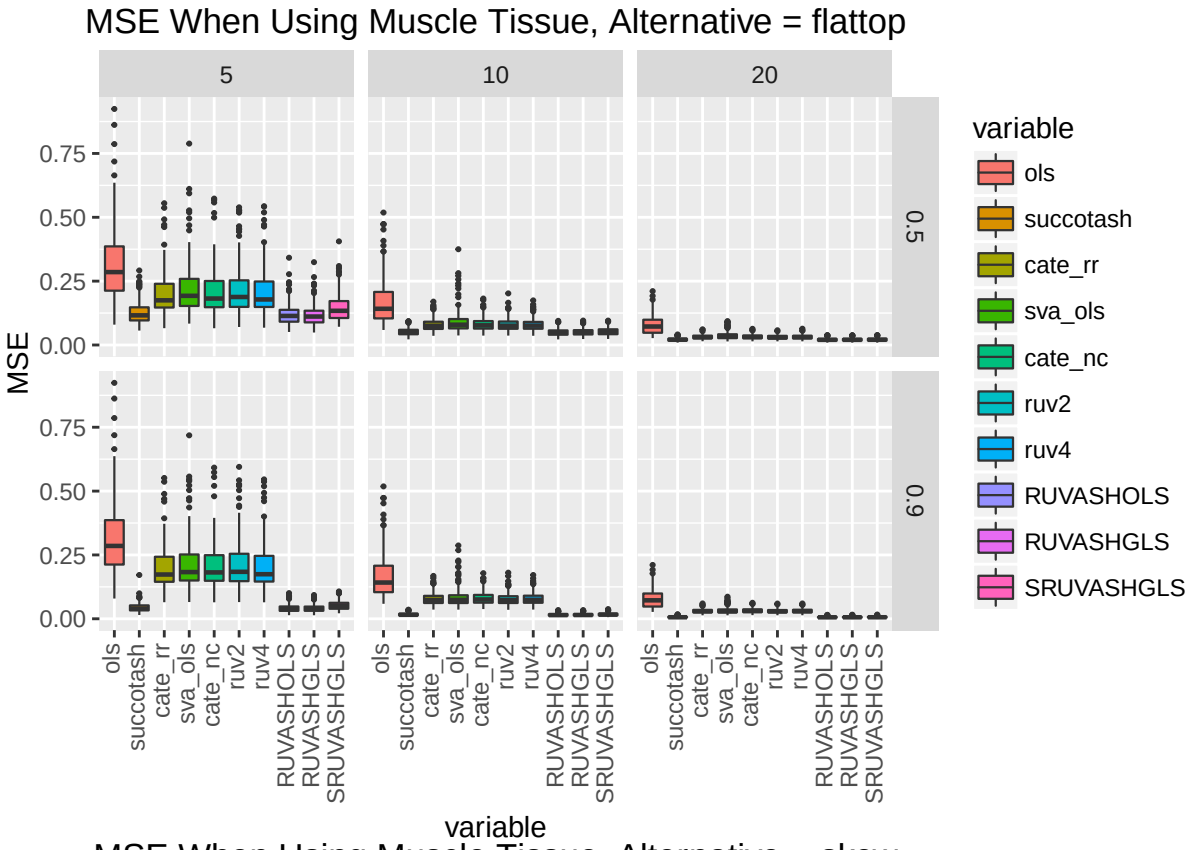

MSE When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
MSE
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

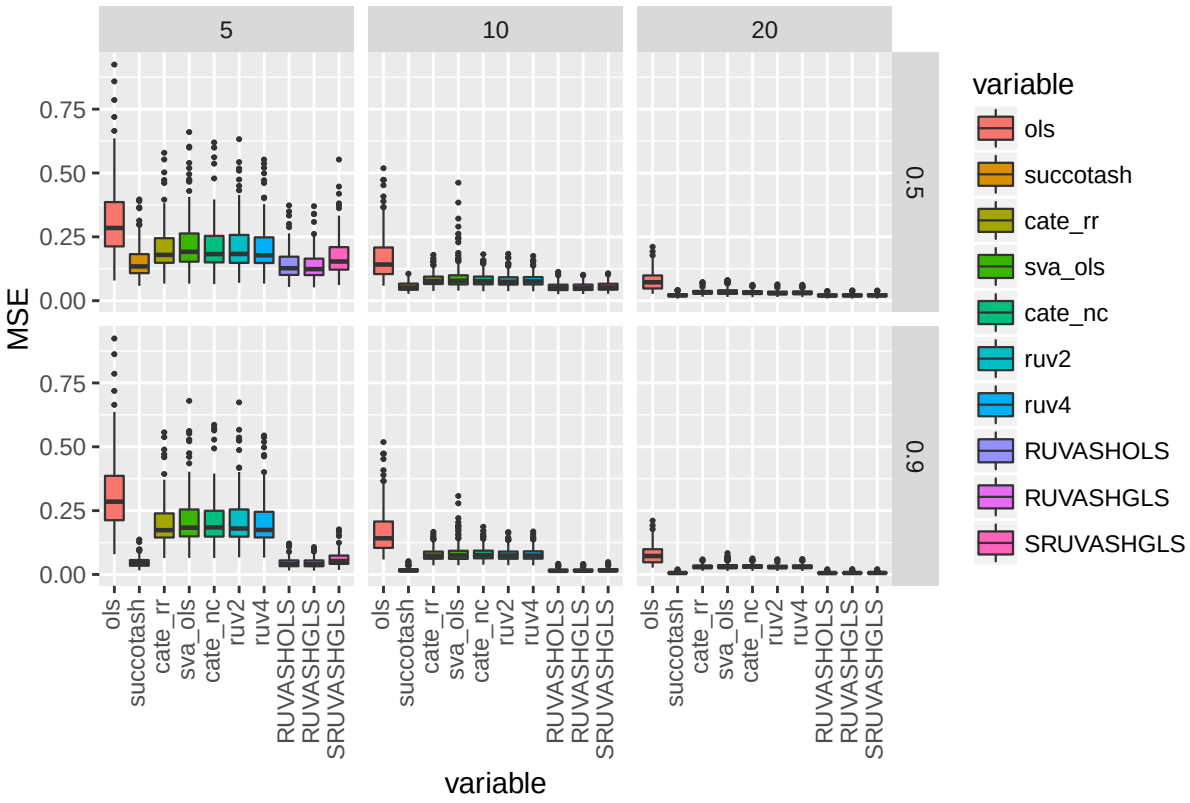

MSE When Using Muscle Tissue, Alternative = all_null
5
10
20
1
MSE
0.0
0.2
0.4
0.6
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
leapp_sparse
leapp_ridge
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS

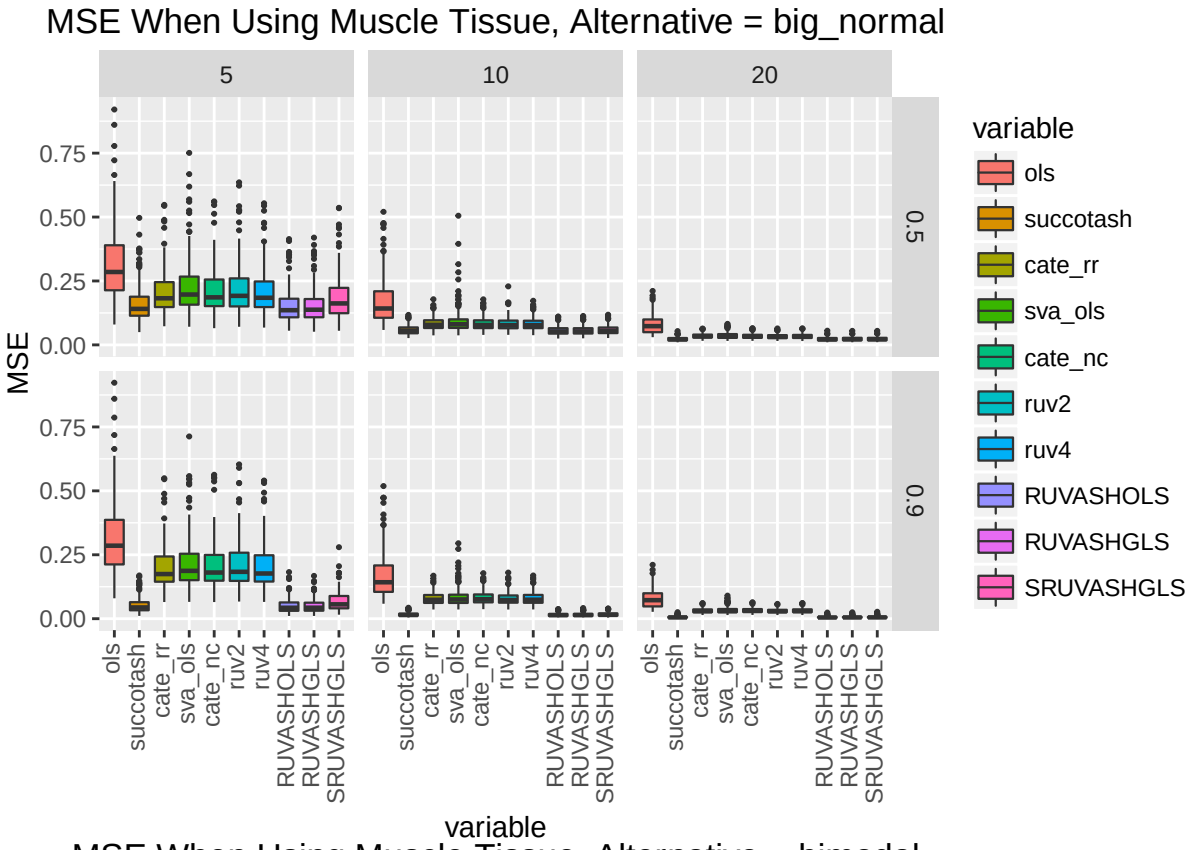
MSE without LEAPP

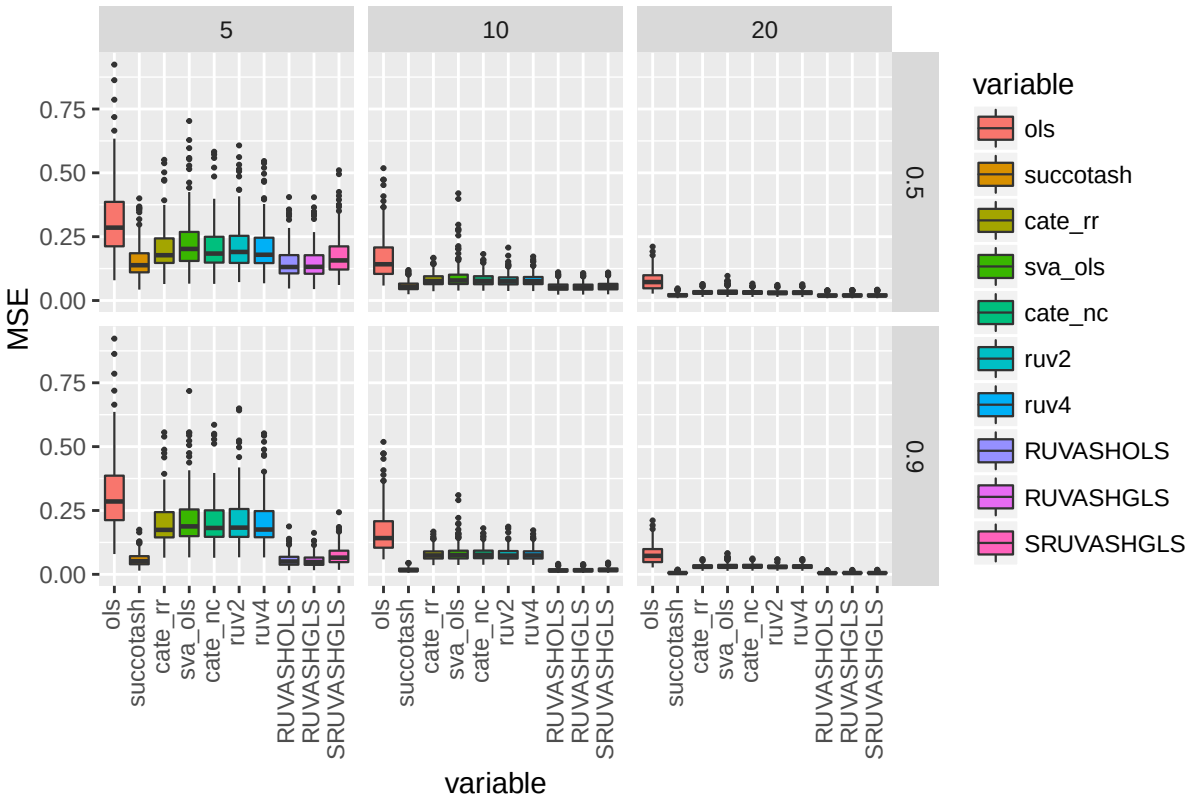


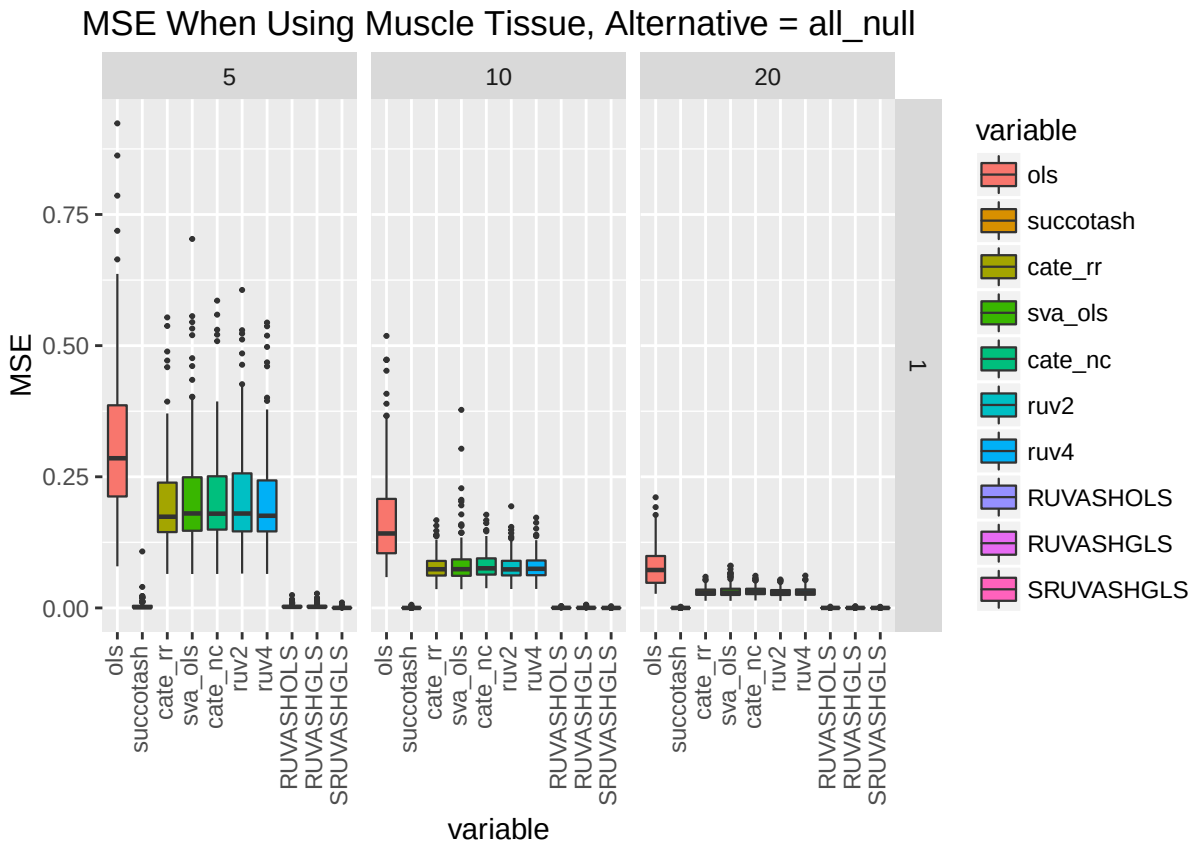



MSE When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
MSE
0.00
0.25
0.50
0.75
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS


MSE When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
MSE
0.00
0.25
0.50
0.75
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS


MSE When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
MSE
0.00
0.25
0.50
0.75
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS


MSE When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
MSE
0.00
0.25
0.50
0.75
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS
variable
variable
ols
succotash
cate_rr
sva_ols
cate_nc
ruv2
ruv4
RUVASHOLS
RUVASHGLS
SRUVASHGLS




```
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] pROC_1.8       dplyr_0.4.3    reshape2_1.4.1 ggplot2_2.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5      knitr_1.12.28    magrittr_1.5     munsell_0.4.3
##  [5] colorspace_1.2-6 R6_2.1.2         stringr_1.0.0    plyr_1.8.3
##  [9] tools_3.3.0      parallel_3.3.0   grid_3.3.0       gtable_0.2.0
## [13] DBI_0.4          htmltools_0.3.5  yaml_2.1.13      lazyeval_0.1.10
## [17] assertthat_0.1   digest_0.6.9     formatR_1.3      codetools_0.2-14
## [21] evaluate_0.9     rmarkdown_0.9.6  labeling_0.3     stringi_1.0-1
```

```
## [25] compiler_3.3.0   scales_0.4.0
```

Buja, Andreas, and Nermin Eyuboglu. 1992. "Remarks on Parallel Analysis." *Multivariate Behavioral Research* 27 (4). Taylor & Francis: 509–40.

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