Different Alternative Types, Including MAD Inflation

David Gerard

2016-06-10

Abstract

I add MAD inflated + ASH methods to the mix. They work pretty well, but not as well as RUVASH.

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 $\log_2(COUNTS + 1)$ (except for VLEMA below). 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.
- 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.
- SVA + qvalue.
- SVA + MAD inflation + ASH
- Voom -> limma -> eBayes -> MAD inflation -> ASH pipeline (VLEMA)
- RUVASH (inflation estimated using controls)
- RUV4 + inflation estimated using controls + qvalue
- RUV4 + MAD inflation + ASH
- RUV2 + MAD inflation + ASH
- 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

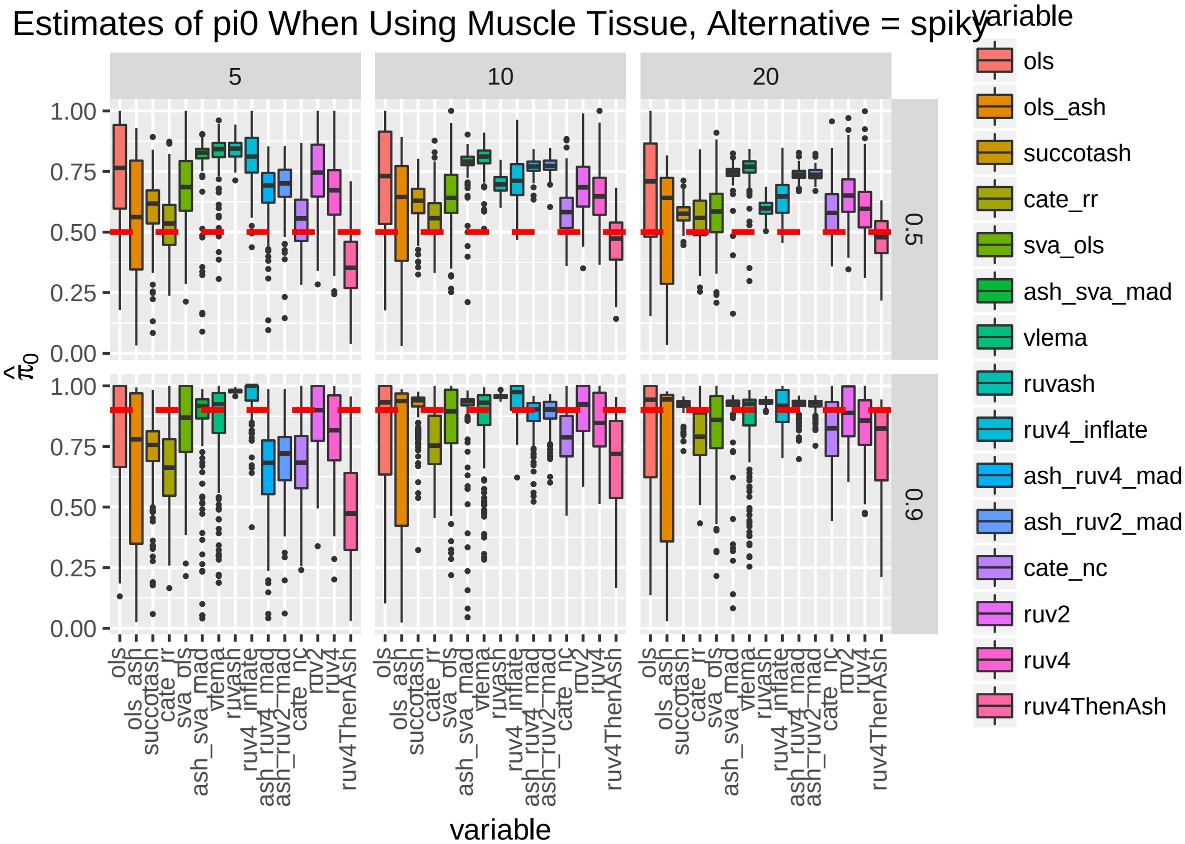
- SUCCOTASH has slightly anti-conservative estimates of π_0 in the Flattop and bimodal Scenarios. It does well for every other scenario for larger n.
- SVA + MAD + ASH has a very long left tail when $\pi_0 = 0.9$, but works pretty well when $\pi_0 = 0.5$ (though not as well as RUVASH).
- VLEMA also has very long left tails when $\pi_0 = 0.9$ but works pretty well when $\pi_0 = 0.5$ (though not as well as RUVASH).
- RUV + MAD + ASH works pretty well. It doesn't do as well as RUVASH when $n = 0.5$ and $\pi_0 = 0.9$, but it is usually conservative otherwise. It has a slightly long left tail at $\pi_0 = 0.9$ and is more conservative than RUVASH at $\pi_0 = 0.5$.
- In the all null case, only RUVASH and SUCCOTASH work really well.

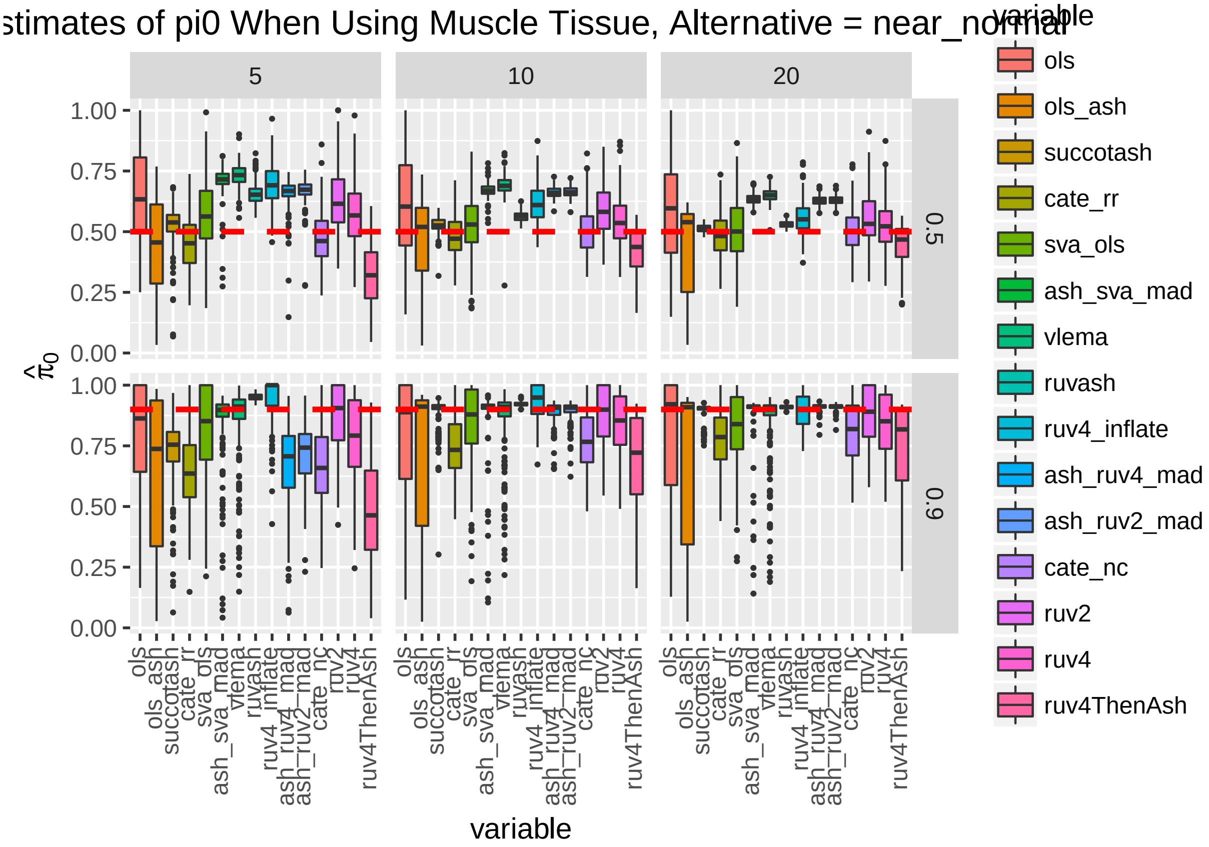
AUC performance.

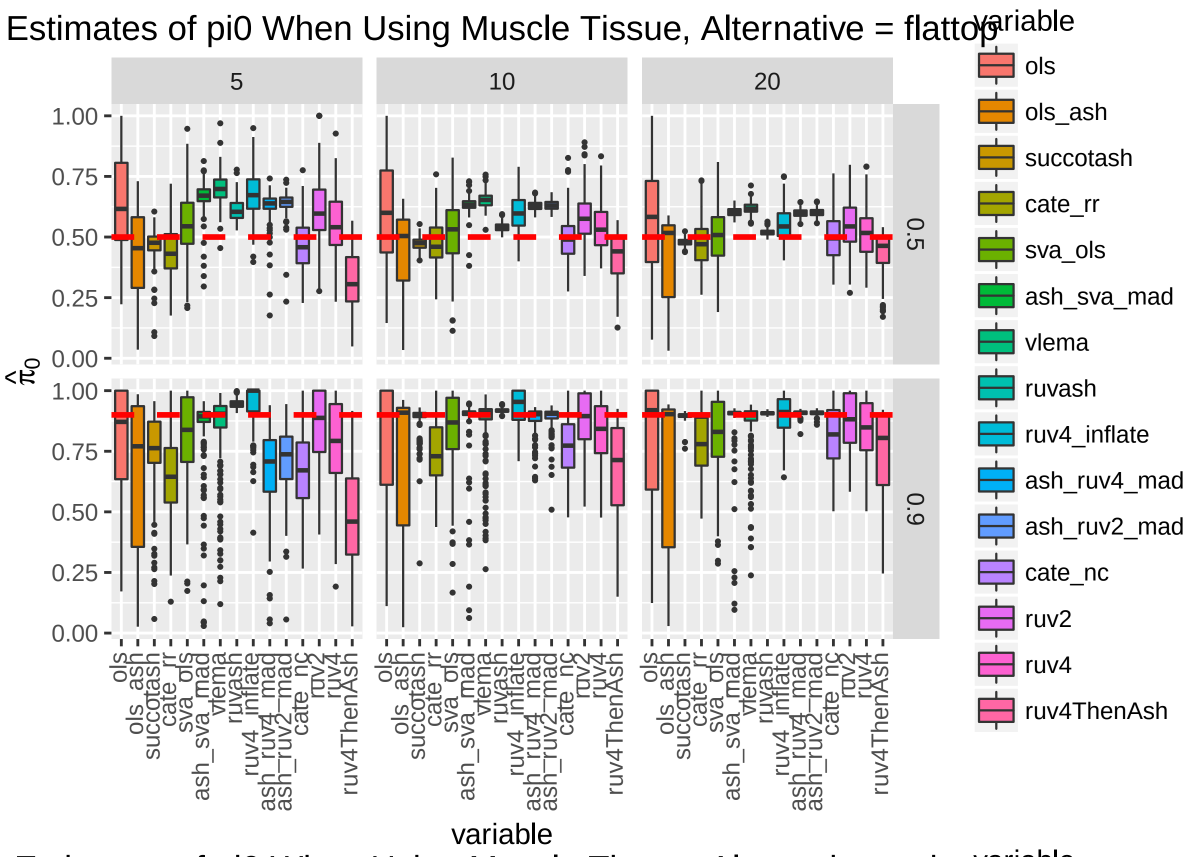
- All of the ASH-like mehtods have very similar AUC (which is higher than all of the non-ash methods). The RUV + MAD + ASH methods don't use GLS when estimating the hidden confounders, and I think that makes their AUC slightly worse in some instances.

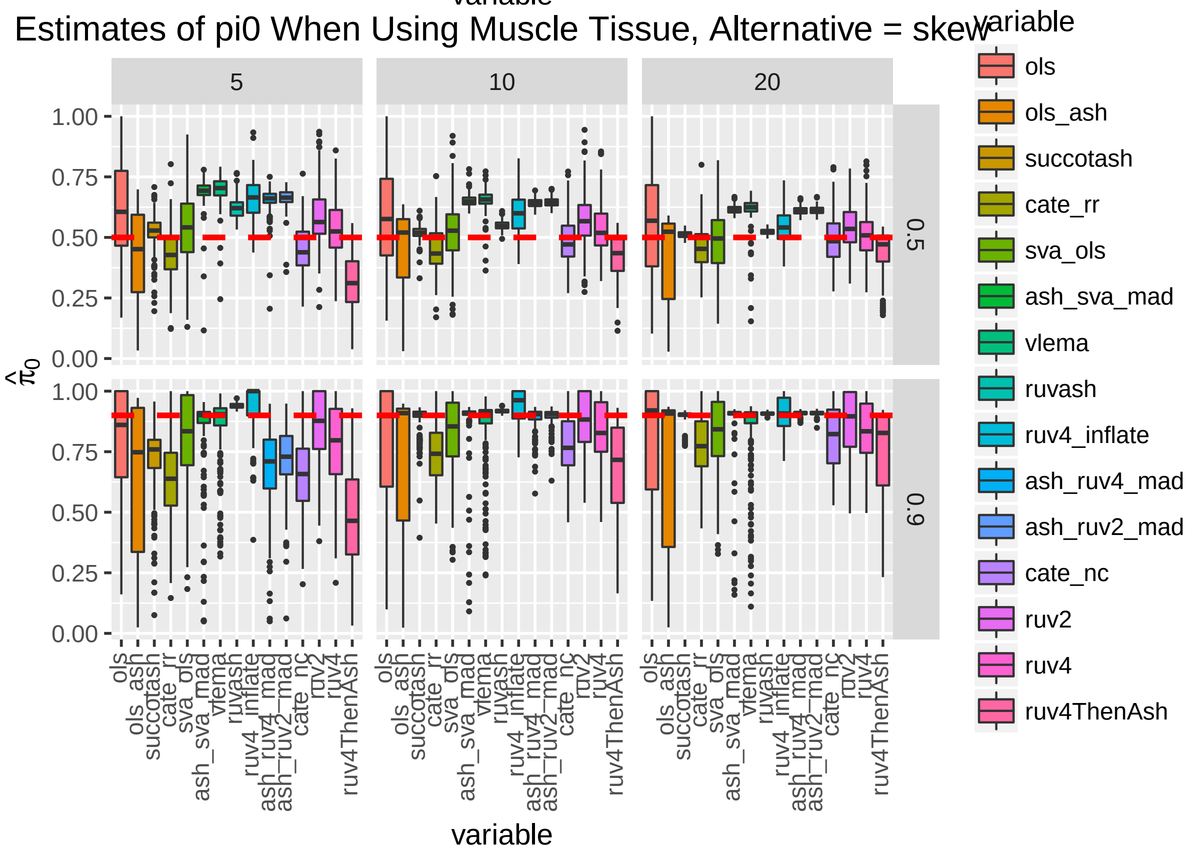
MSE

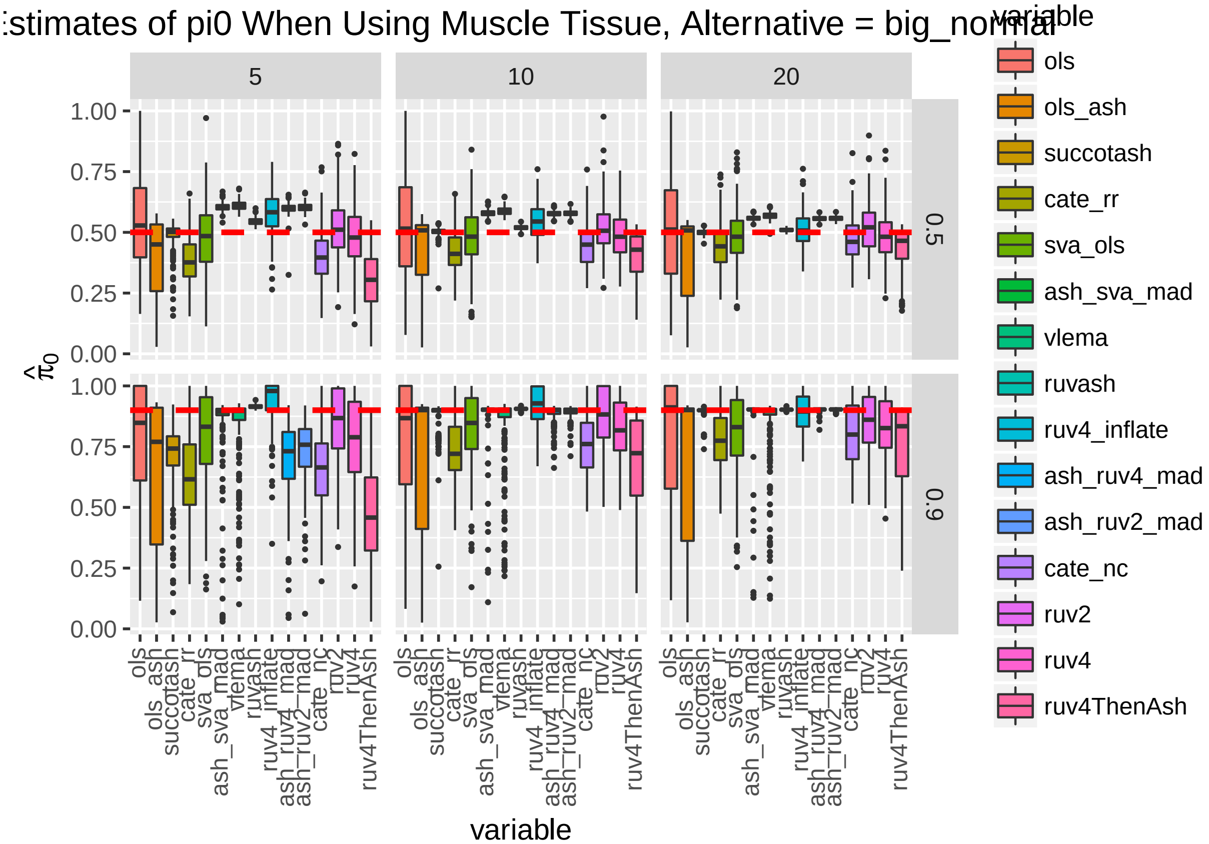
- The ASH-like methods have much better MSE than the non-ASH-like methods. RUV4 + ASH methods where we use GLS and SUCCOTASH perform the best (especially when $\pi_0 = 0.5$).


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
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

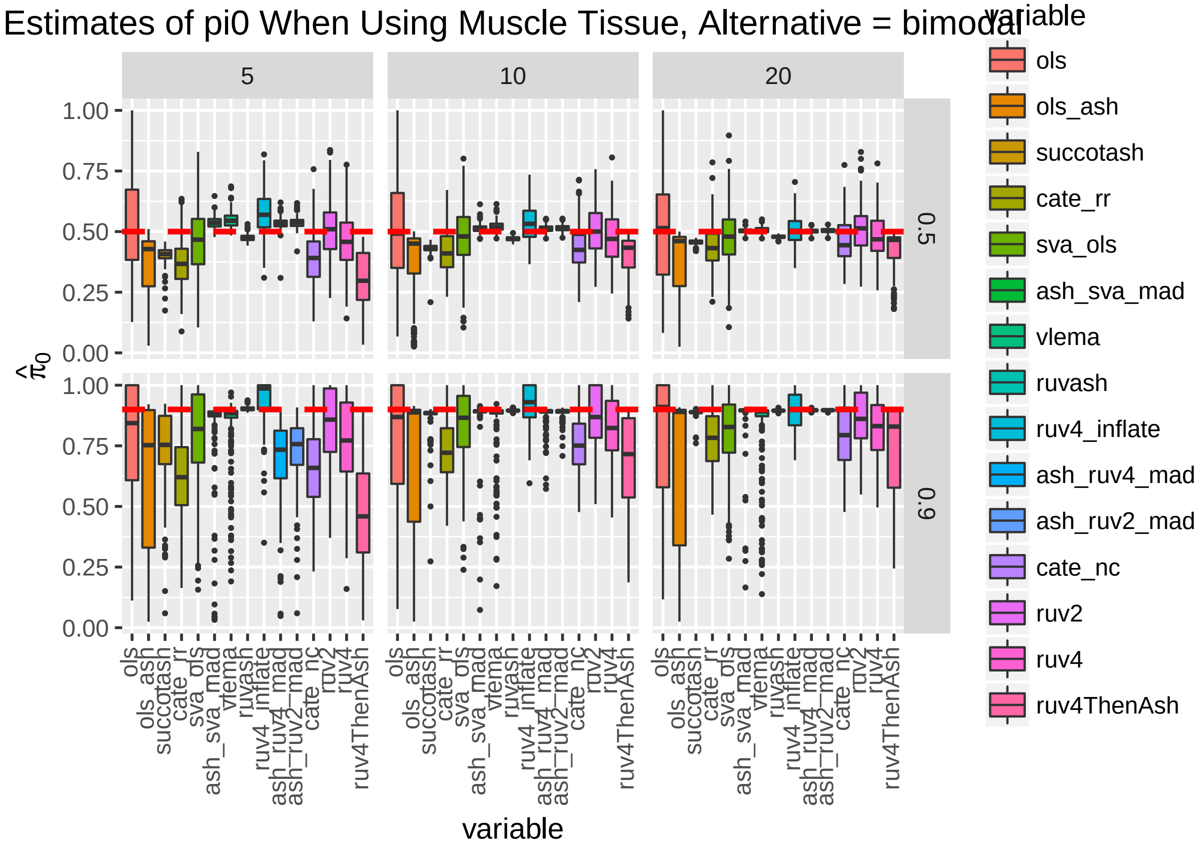

stimates of pi0 When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
1.00
0.75
0.50
0.25
0.00
$\hat{\pi}_0$
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh


Estimates of pi0 When Using Muscle Tissue, Alternative = flattop

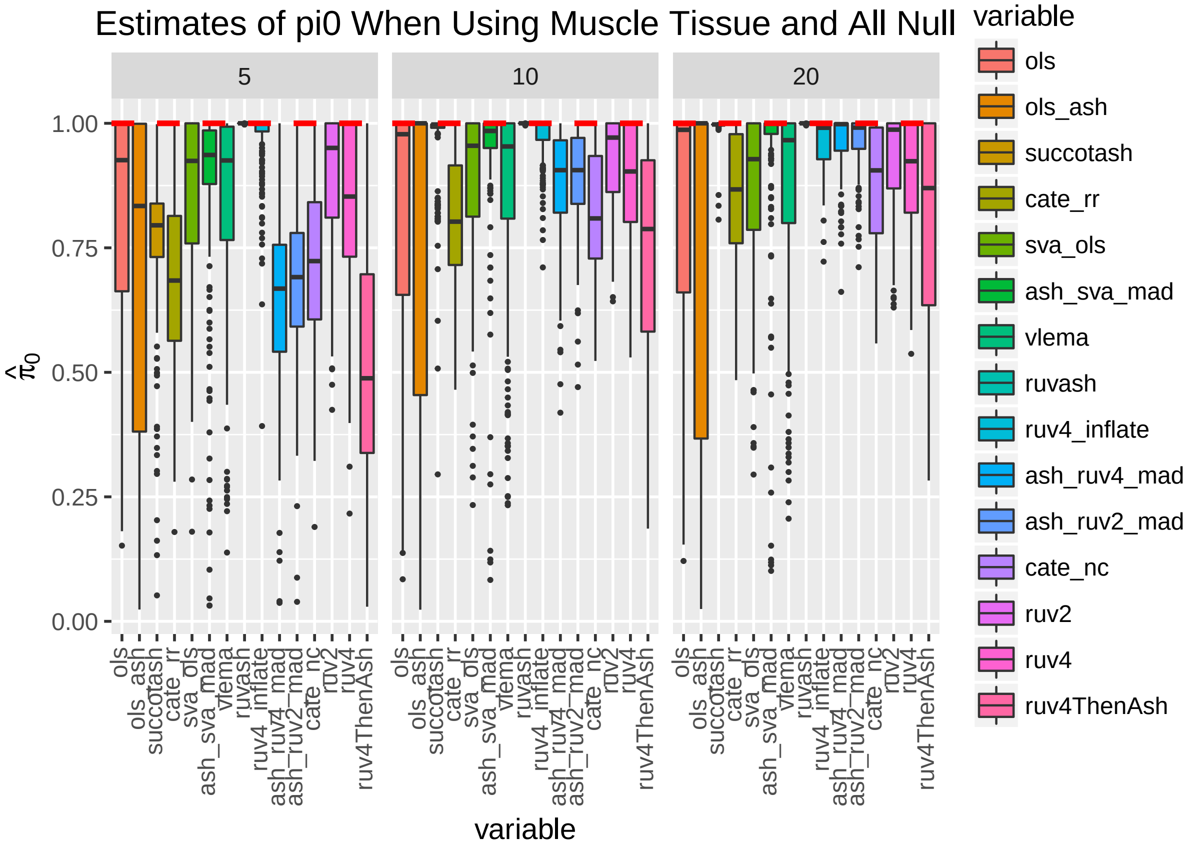

Estimates of pi0 When Using Muscle Tissue, Alternative = skew

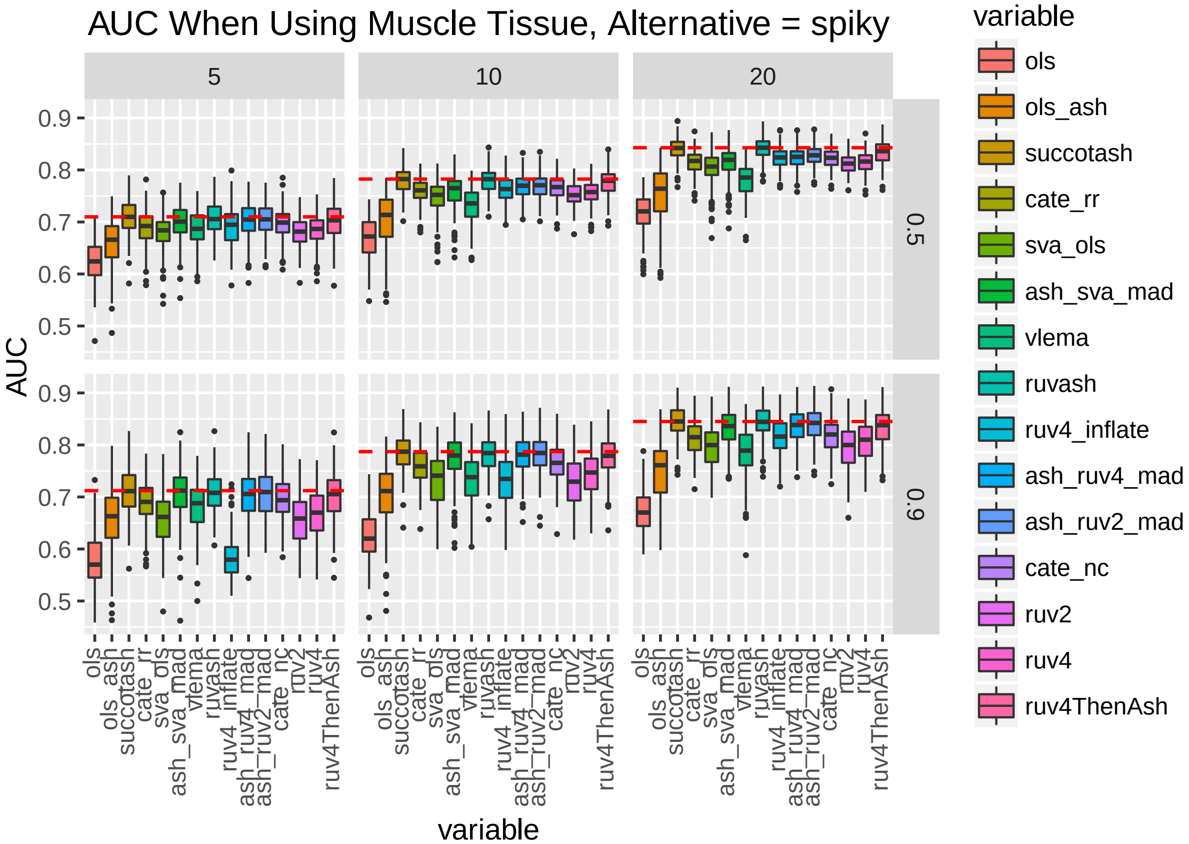


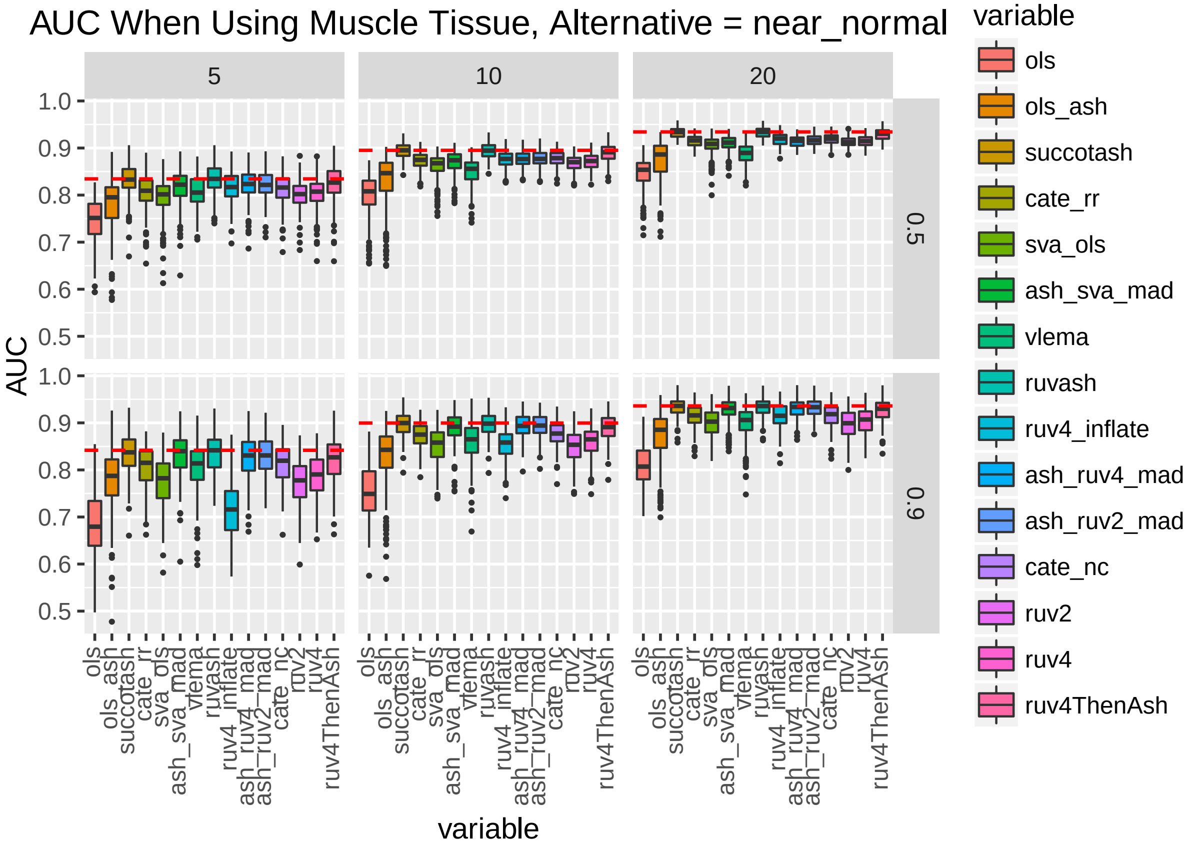
stimates of pi0 When Using Muscle Tissue, Alternative = big_normal

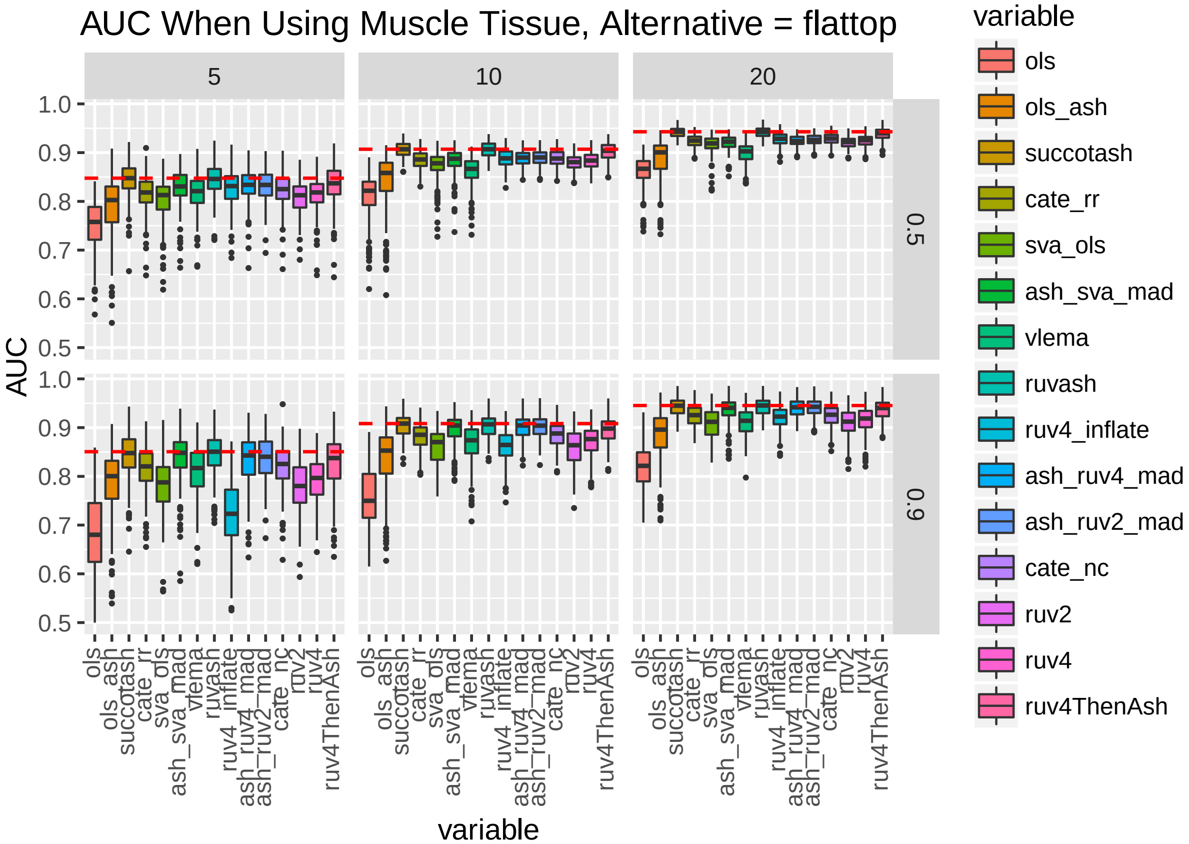


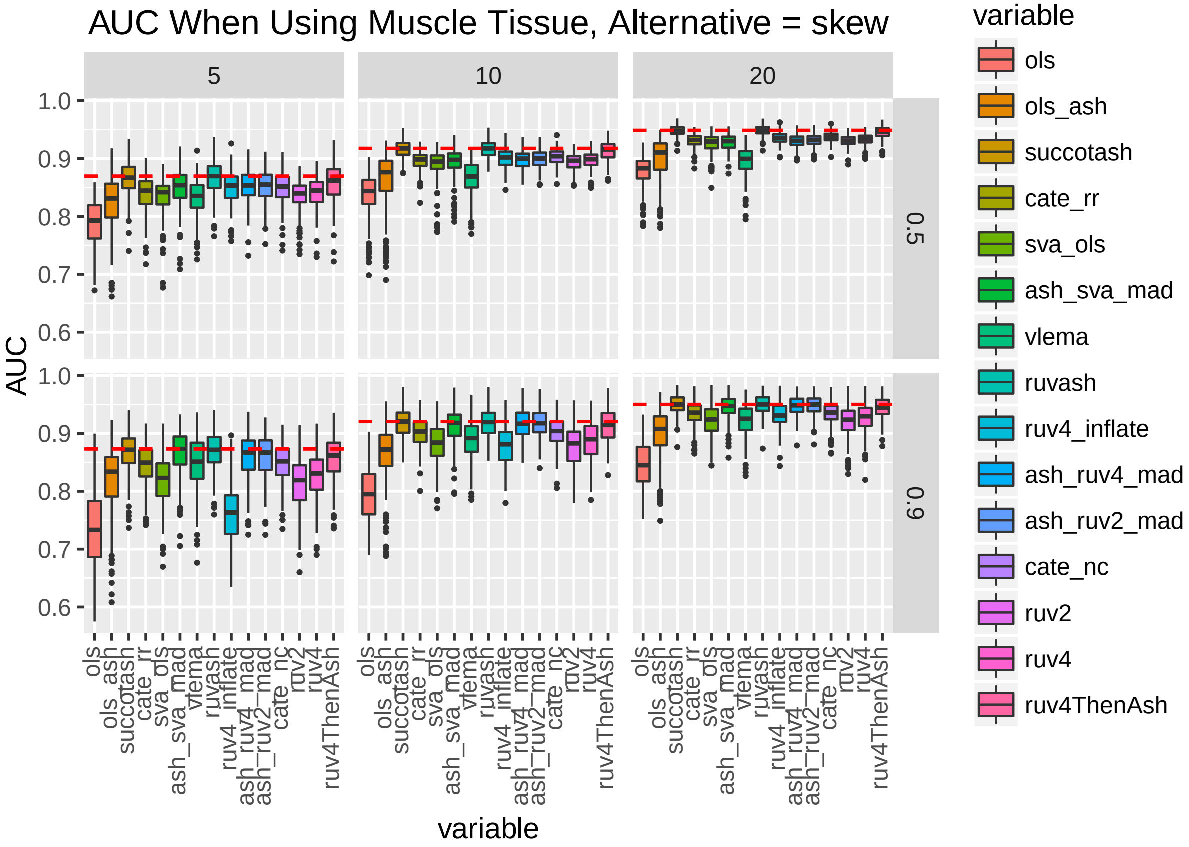
Estimates of pi0 When Using Muscle Tissue, Alternative = bimodal

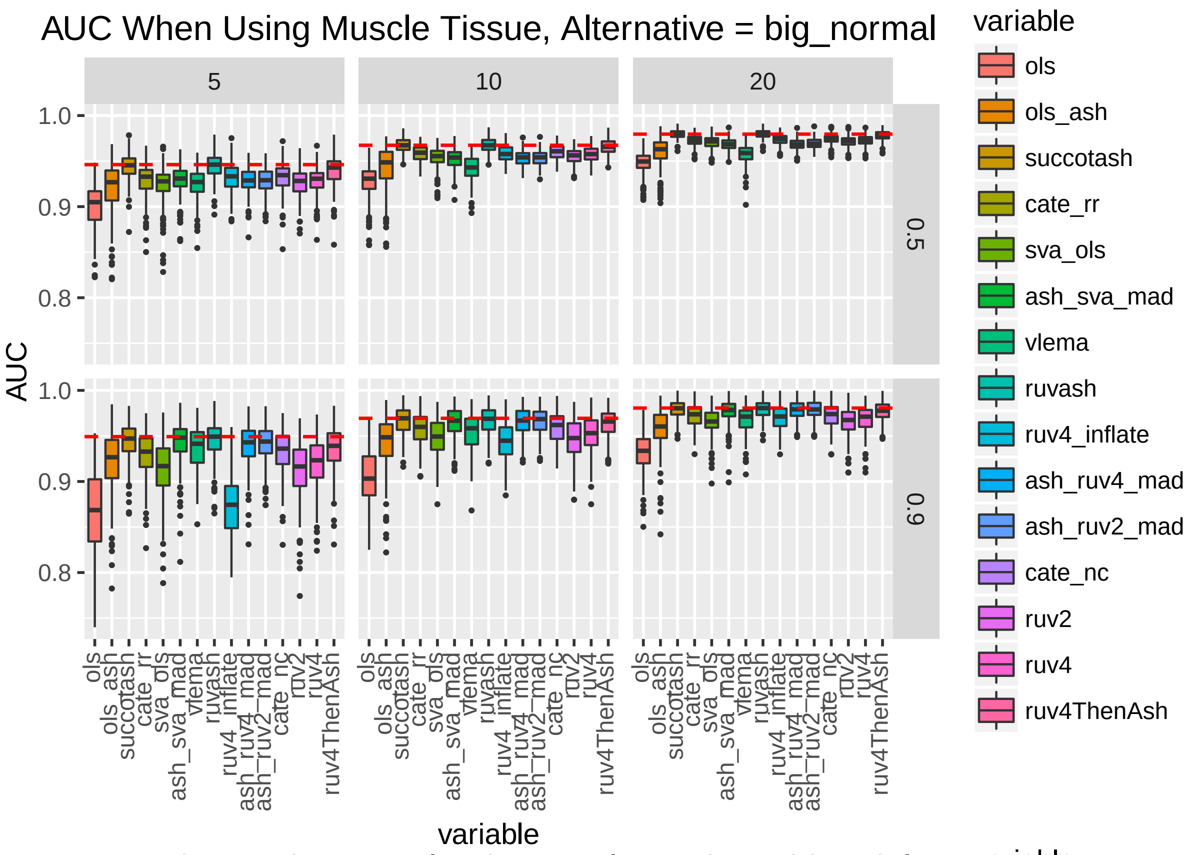

Estimates of pi0 When Using Muscle Tissue and All Null
5
10
20
$\hat{\pi}_0$
1.00
0.75
0.50
0.25
0.00
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

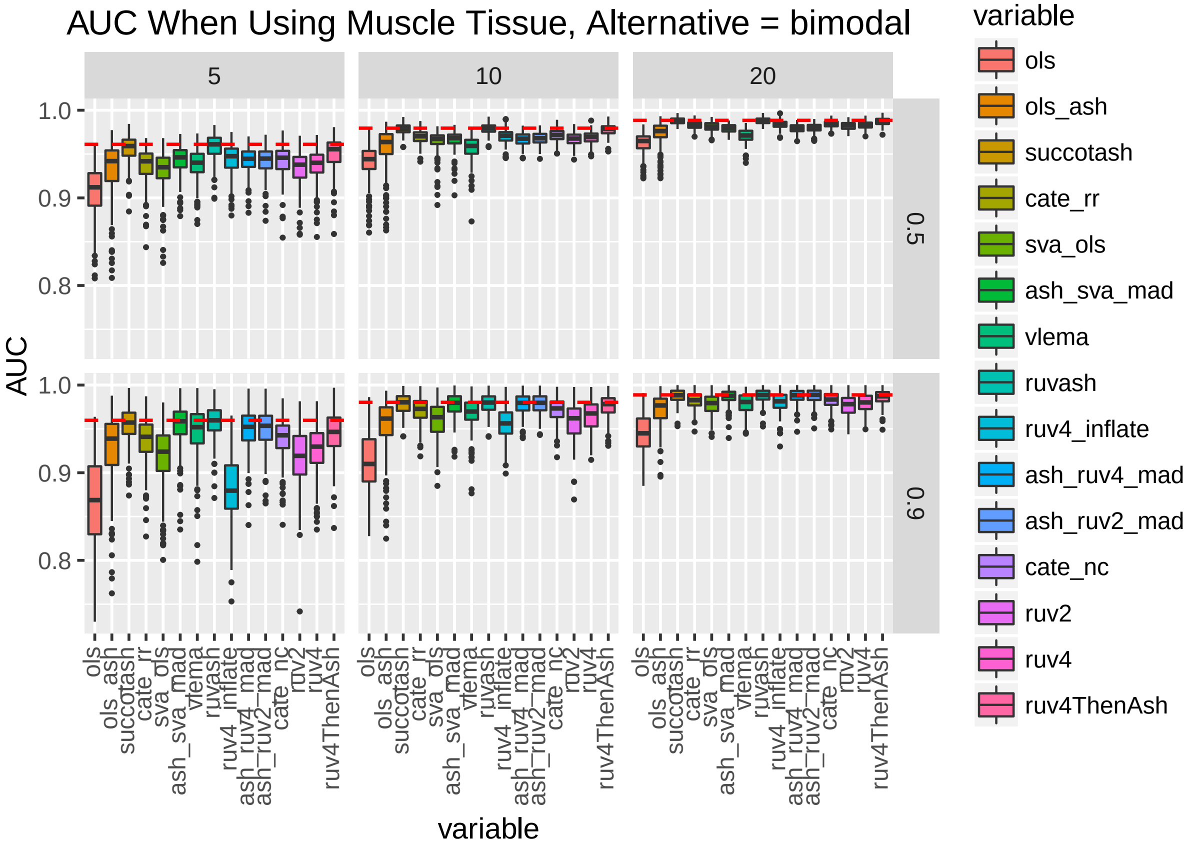

AUC When Using Muscle Tissue, Alternative = spiky

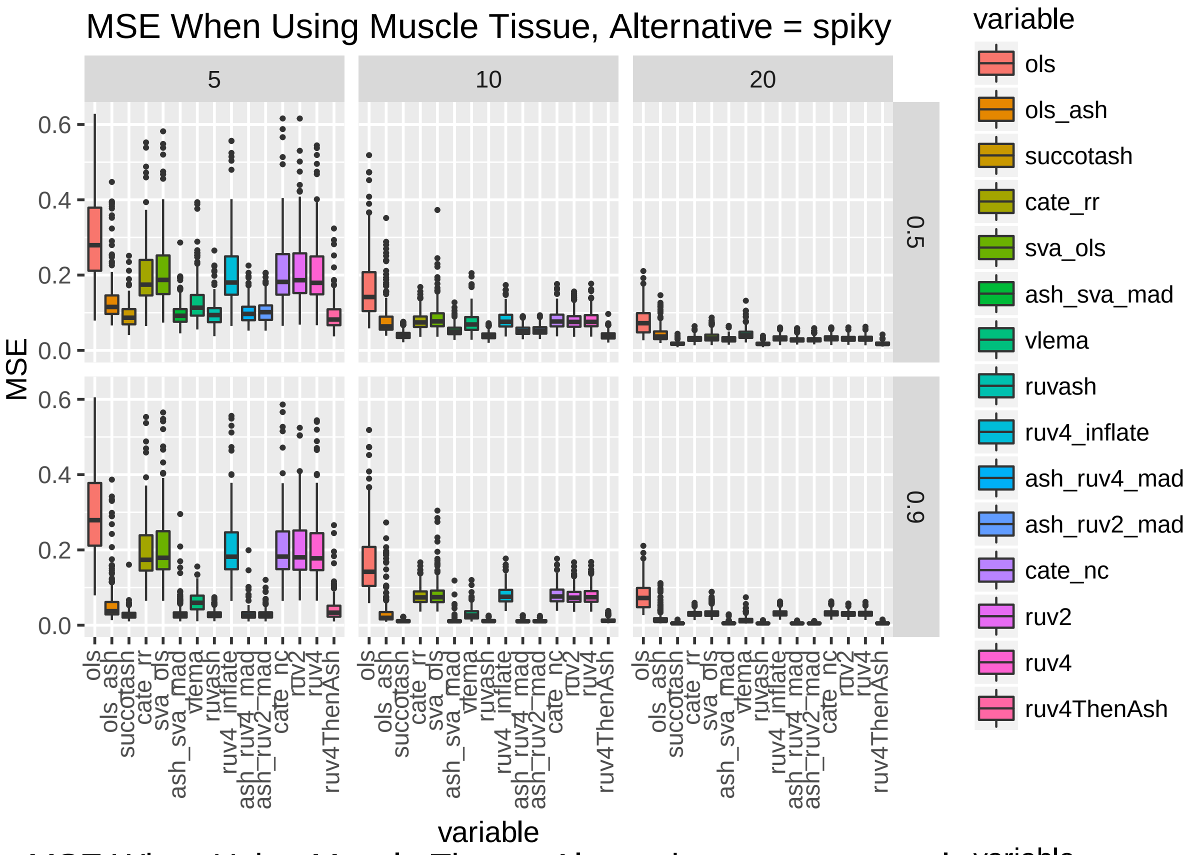

AUC When Using Muscle Tissue, Alternative = near_normal

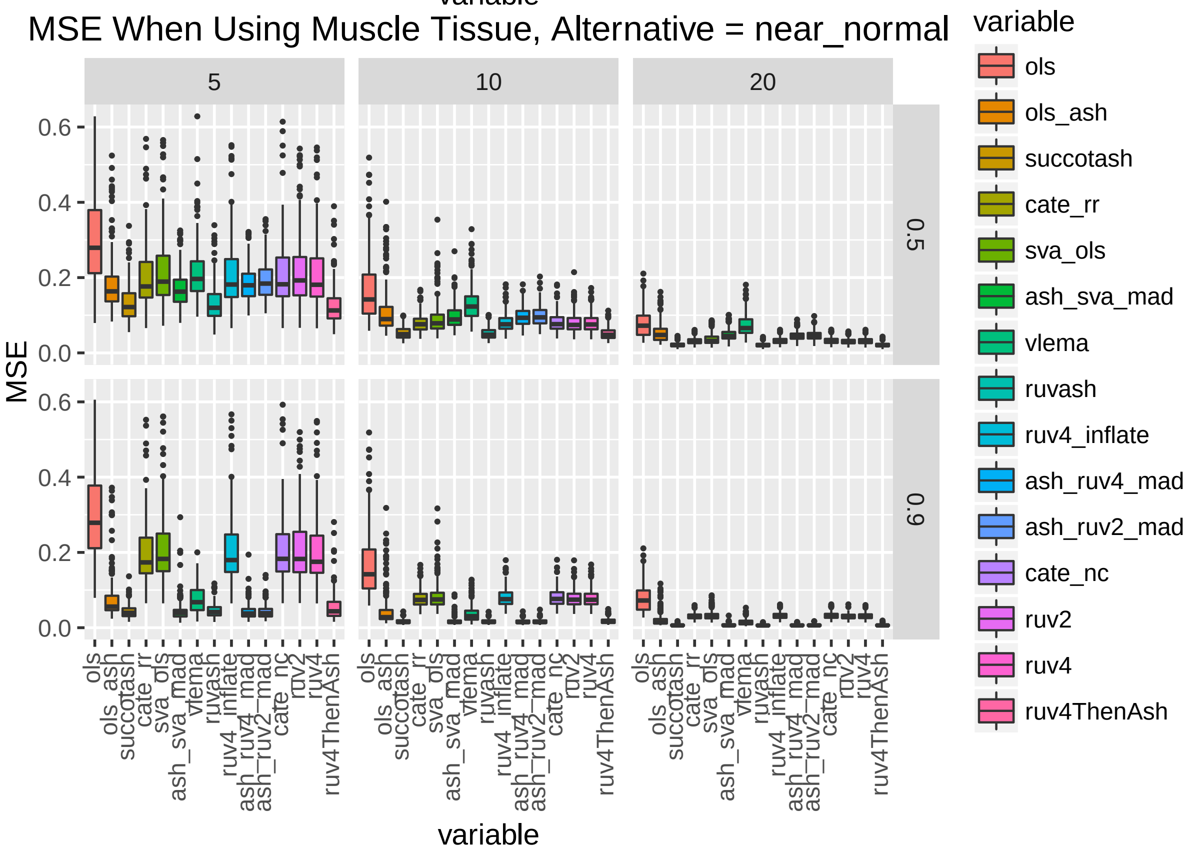

AUC When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
0.5
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

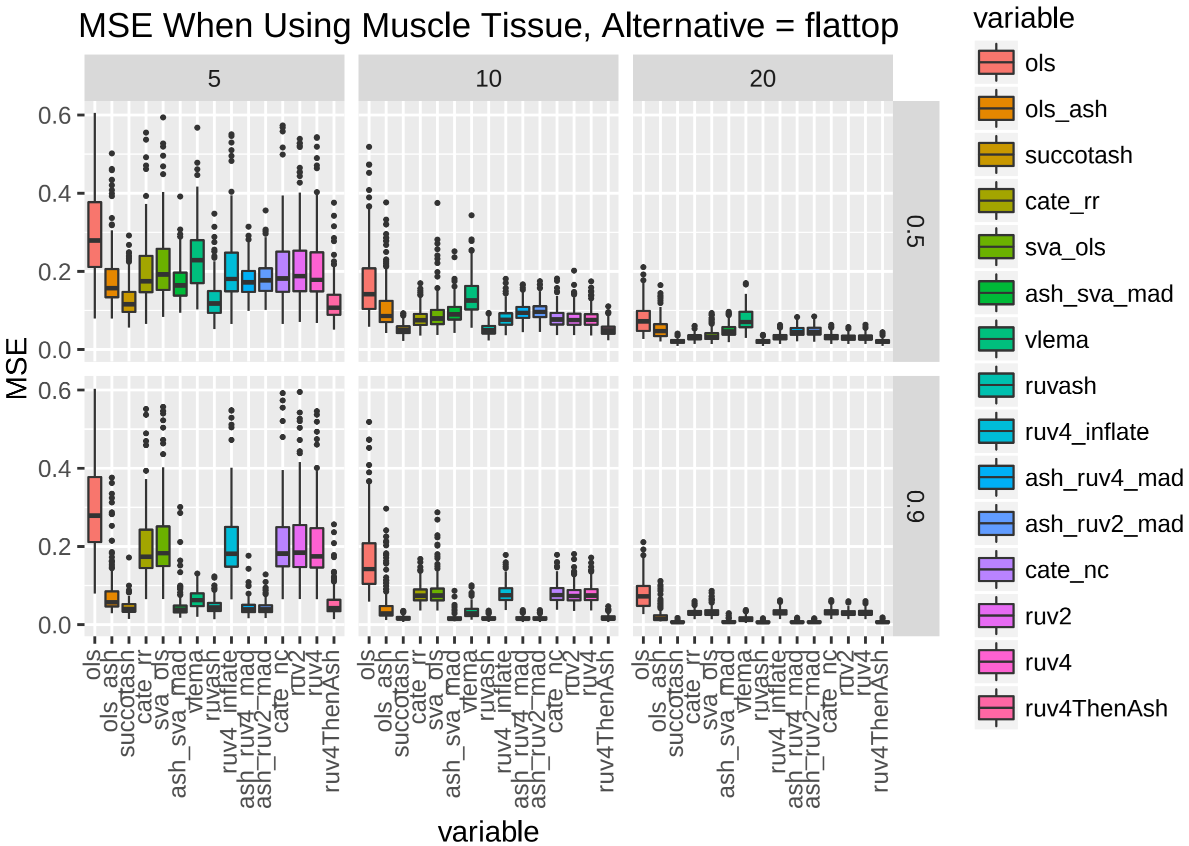

AUC When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
AUC
1.0
0.9
0.8
0.7
0.6
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

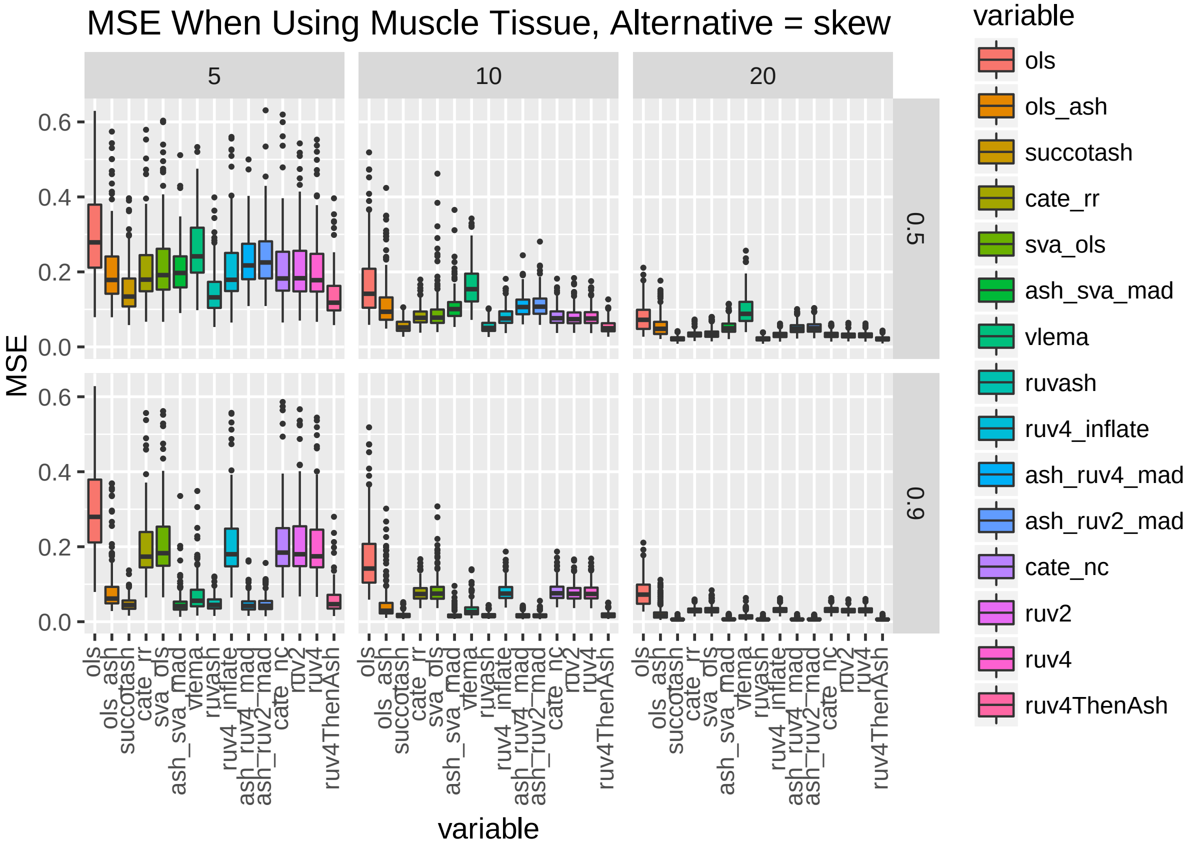

AUC When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
1.0
0.9
0.8
AUC
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

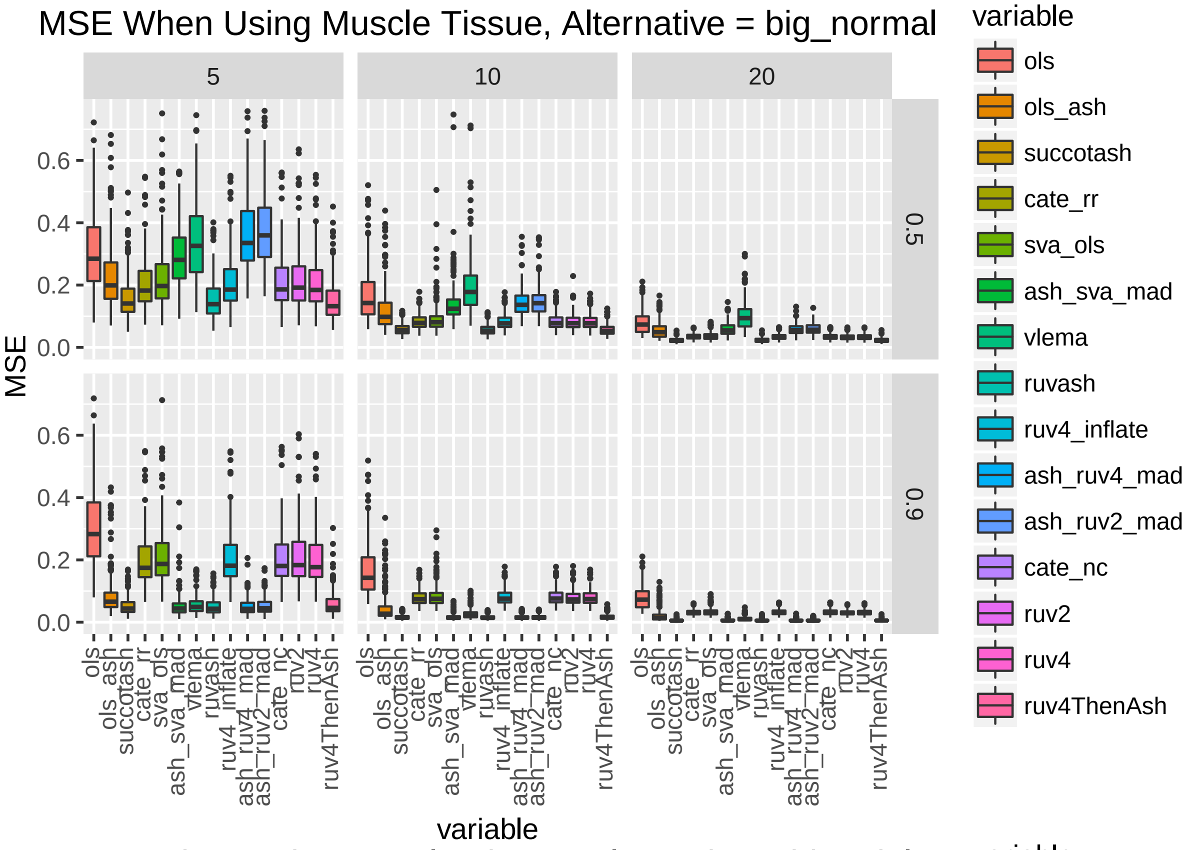

AUC When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
1.0
0.9
0.8
AUC
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh

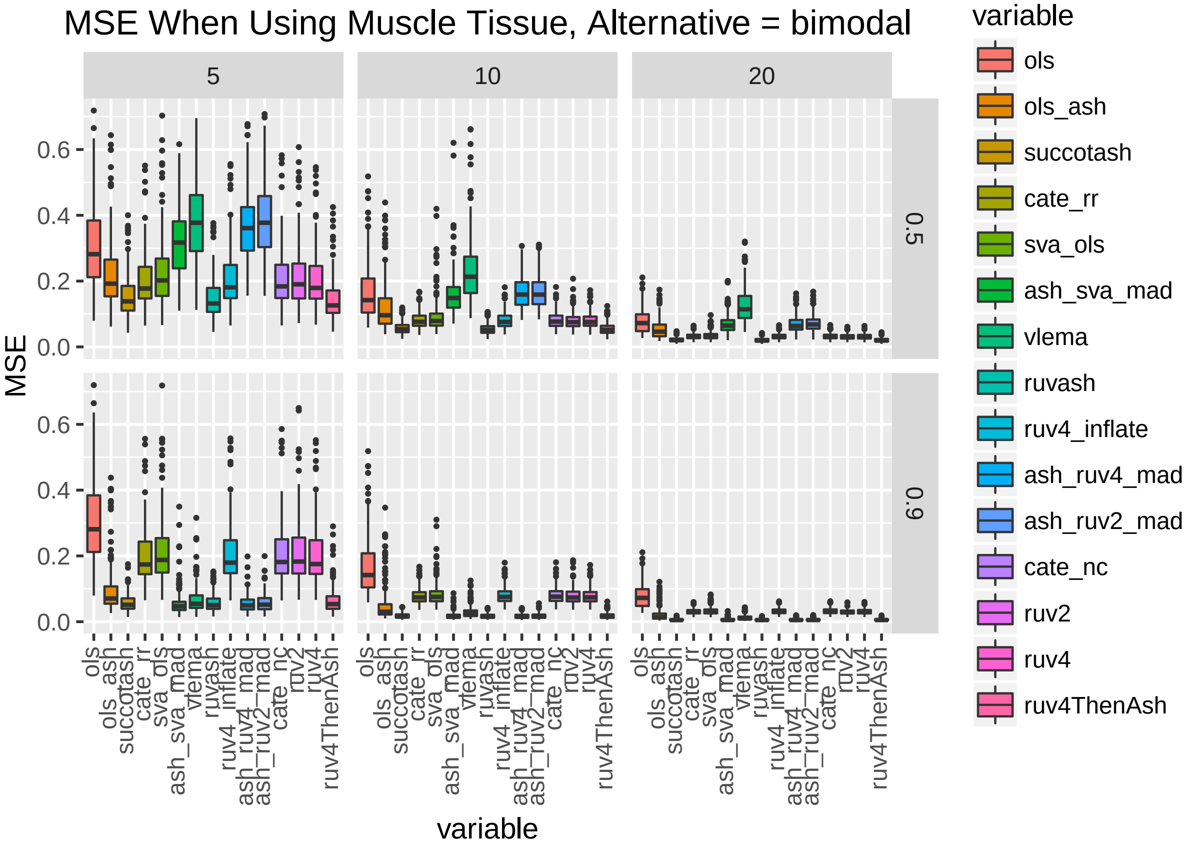

MSE When Using Muscle Tissue, Alternative = spiky
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable

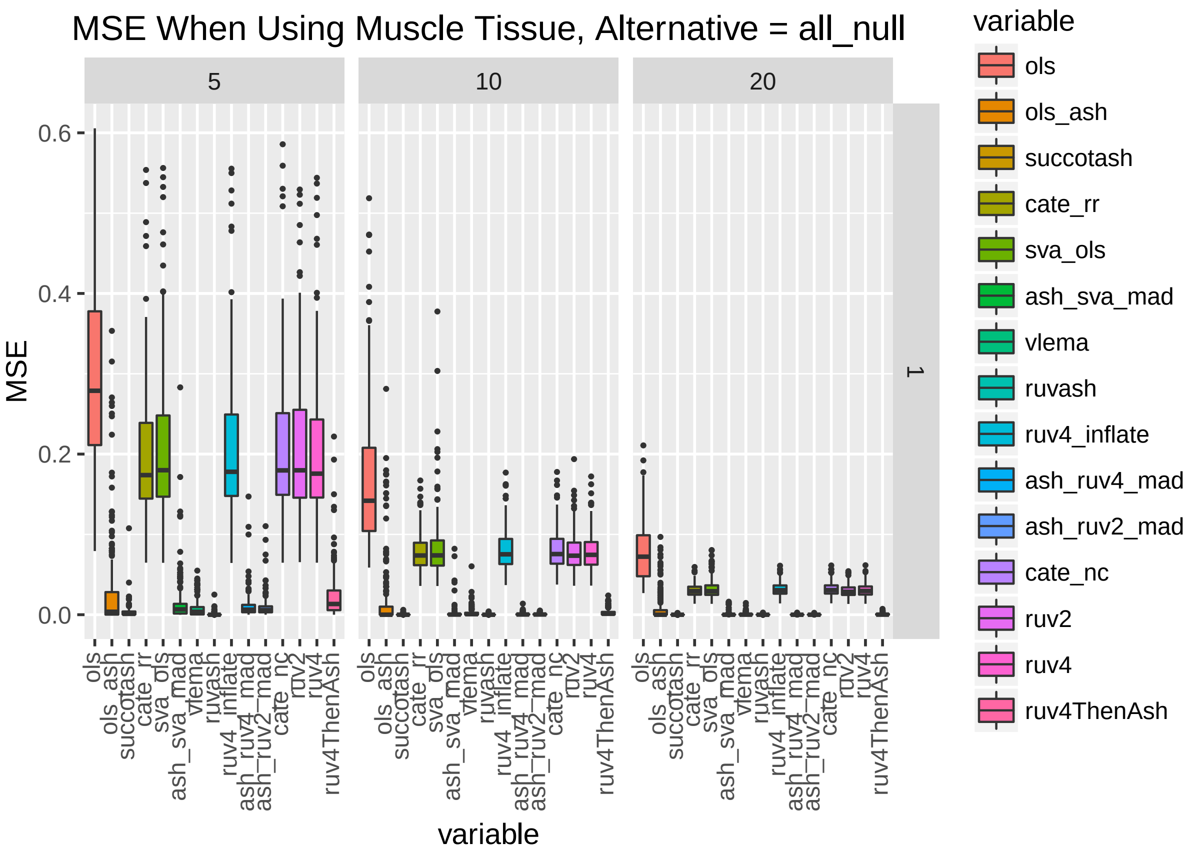

MSE When Using Muscle Tissue, Alternative = near_normal
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable


MSE When Using Muscle Tissue, Alternative = flattop
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = skew
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = big_normal
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh


MSE When Using Muscle Tissue, Alternative = bimodal
5
10
20
0.5
0.9
MSE
0.0
0.2
0.4
0.6
variable
ols
ols_ash
succotash
cate_rr
sva_ols
ash_sva_mad
vlema
ruvash
ruv4_inflate
ash_ruv4_mad
ash_ruv2_mad
cate_nc
ruv2
ruv4
ruv4ThenAsh



```r
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_1.8.4
##  [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.