Supplementary R script

Jakob Russel

October 2017

Load packages

```
library(phyloseq)
library(DAtest)
```

```
## DAtest version 2.6.6
```

```
## R version 3.3.3 (2017-03-06)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] DAtest_2.6.6    phyloseq_1.19.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.13        nloptr_1.0.4        plyr_1.8.4
##  [4] XVector_0.14.1      iterators_1.0.8     tools_3.3.3
##  [7] zlibbioc_1.20.0     lme4_1.1-14         statmod_1.4.30
## [10] digest_0.6.12       tibble_1.3.4        jsonlite_1.5
## [13] evaluate_0.10.1     nlme_3.1-131        rhdf5_2.18.0
## [16] gtable_0.2.0        lattice_0.20-35     mgcv_1.8-20
## [19] doSNOW_1.0.15       pkgconfig_2.0.1     rlang_0.1.2
## [22] igraph_1.1.2        Matrix_1.2-11       foreach_1.4.3
## [25] yaml_2.1.14         parallel_3.3.3      stringr_1.2.0
## [28] knitr_1.17          cluster_2.0.6       pROC_1.10.0
## [31] Biostrings_2.42.1   S4Vectors_0.12.2    IRanges_2.8.2
## [34] cowplot_0.8.0       multtest_2.30.0     stats4_3.3.3
## [37] rprojroot_1.2       ade4_1.7-8          grid_3.3.3
## [40] Biobase_2.34.0      data.table_1.10.4-2 snow_0.4-2
## [43] survival_2.41-3     rmarkdown_1.6       minqa_1.2.4
## [46] reshape2_1.4.2      ggplot2_2.2.1       magrittr_1.5
## [49] MASS_7.3-47         splines_3.3.3       backports_1.1.0
## [52] scales_0.5.0        codetools_0.2-15    htmltools_0.3.6
## [55] BiocGenerics_0.20.0 biomformat_1.2.0    permute_0.9-4
## [58] ape_4.1             colorspace_1.3-2    stringi_1.1.5
## [61] pscl_1.5.1          lazyeval_0.2.0      munsell_0.4.3
## [64] vegan_2.4-4
```

Import

These files are downloaded from https://www.hmpdacc.org/hmp/HMQCP/

```
otu <- import_qiime_otu_tax("v35_psn_otu.genus.fixed.txt")
map <- import_qiime_sample_data("v35_map_uniquebyPSN.txt")
```

Create phyloseq object

```
otu.tab <- otu_table(otu[[1]], taxa_are_rows = TRUE)
tax.tab <- tax_table(otu[[2]])
all_samples <- merge_phyloseq(otu.tab, tax.tab, map)
```

Subset

This is just an arbitrary subset just for testing the package

```
TS <- subset_samples(all_samples, HMPbodysubsite %in% c("Tongue_dorsum","Saliva"))
TS <- subset_samples(TS, sample_sums(TS) > 5000)
TS <- subset_samples(TS, visitno == 1)
TS <- subset_samples(TS, sex == "male")
TS <- prune_taxa(taxa_sums(TS) > 100, TS)
```

Run test

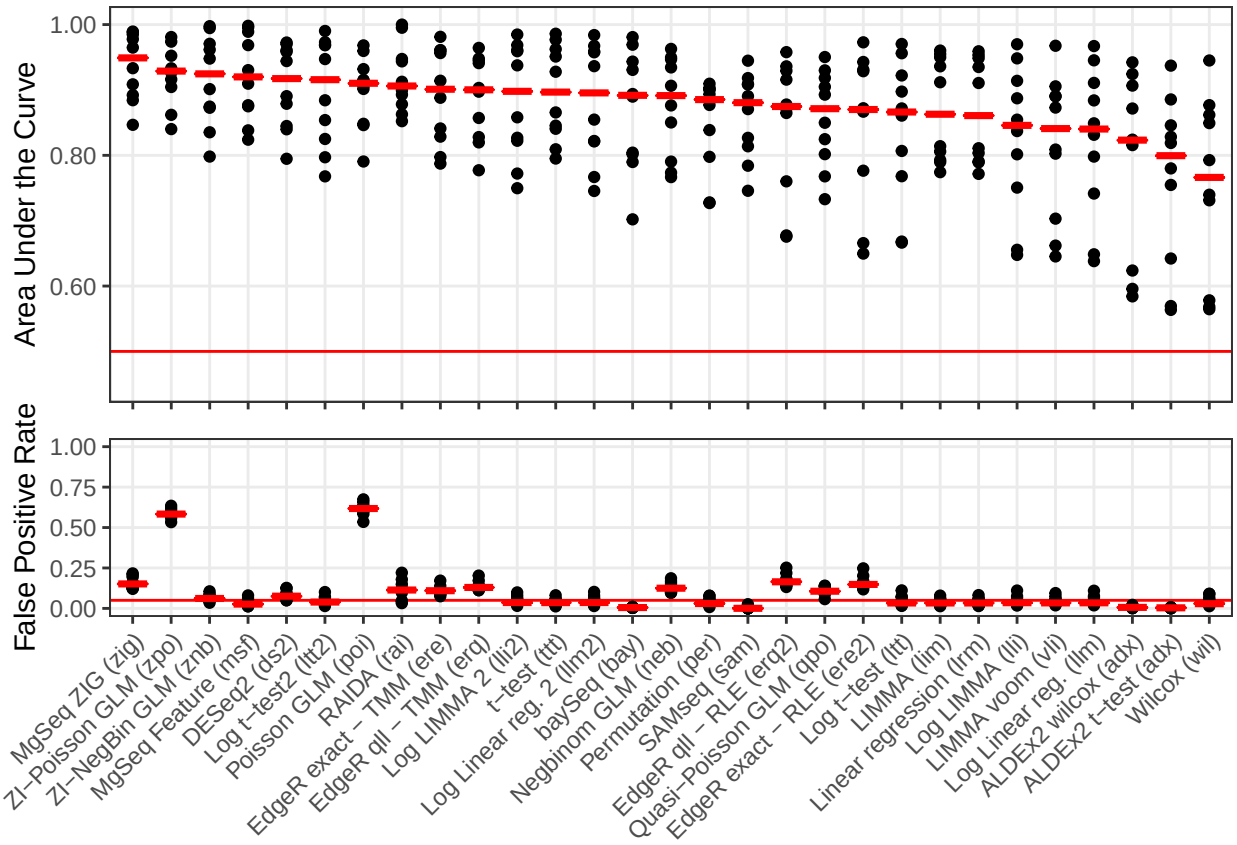
```
test <- testDA(TS, predictor = "HMPbodysubsite", effectSize = 4)
```

```
## Seed is set to 123
```

```
## predictor is assumed to be a categorical variable with 2 levels: Saliva, Tongue_dorsum
```

Check results

```
plot(test)
```




```
summary(test)
```

```
##                    Method   AUC   FPR Spike.detect.rate
##            MgSeq ZIG (zig) 0.949 0.151              0.800
##       ZI-Poisson GLM (zpo) 0.929 0.583              0.933
##        ZI-NegBin GLM (znb) 0.925 0.061              0.267
##        MgSeq Feature (msf) 0.920 0.027              0.300
##               DESeq2 (ds2) 0.917 0.075              0.233
##         Log t-test2 (ltt2) 0.916 0.039              0.200
##          Poisson GLM (poi) 0.910 0.618              1.000
##                RAIDA (rai) 0.906 0.114              0.167
##    EdgeR exact - TMM (ere) 0.901 0.109              0.267
##      EdgeR qll - TMM (erq) 0.900 0.130              0.600
##         Log LIMMA 2 (lli2) 0.898 0.036              0.133
##               t-test (ttt) 0.897 0.035              0.067
##  Log Linear reg. 2 (llm2) 0.895 0.036              0.100
##               baySeq (bay) 0.892 0.005              0.000
##         Negbinom GLM (neb) 0.891 0.125              0.500
##          Permutation (per) 0.885 0.030              0.133
##               SAMseq (sam) 0.881 0.001              0.367
##     EdgeR qll - RLE (erq2) 0.875 0.164              0.533
##   Quasi-Poisson GLM (qpo) 0.871 0.106              0.300
##  EdgeR exact - RLE (ere2) 0.870 0.148              0.467
##           Log t-test (ltt) 0.866 0.033              0.133
##                LIMMA (lim) 0.863 0.033              0.000
##   Linear regression (lrm) 0.861 0.033              0.000
```

```
##           Log LIMMA (lli) 0.846 0.035             0.067
##          LIMMA voom (vli) 0.841 0.034             0.000
##     Log Linear reg. (llm) 0.840 0.034             0.000
##       ALDEx2 wilcox (adx) 0.823 0.006             0.033
##       ALDEx2 t-test (adx) 0.799 0.004             0.000
##              Wilcox (wil) 0.766 0.029             0.033
```

MetagenomeSeq Featuremodel appears to have the highest AUC among the method with FPR below 0.05

Details from the run:

```
test$details
```

```
##
## Features             961
## Samples               75
## Predictor      Two-class
## Paired                No
## Covars
## RunTime    18.55 Minutes
## Relative            TRUE
## EffectSize             4
## RandomSeed           123
## OutAnova            TRUE
```

Run MetagenomeSeq Featuremodel

which had the highest AUC and FPR below 0.05

```
results.msf <- DA.msf(TS, predictor = "HMPbodysubsite")
```

All tests

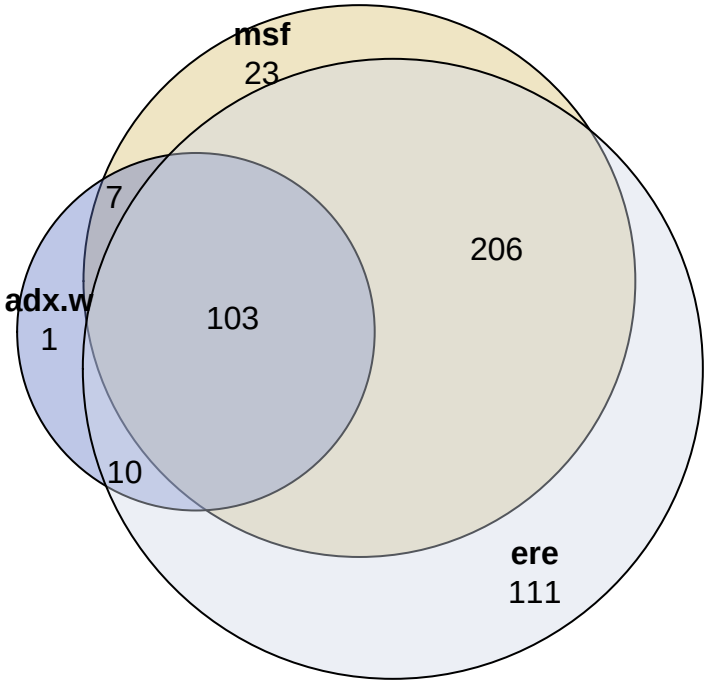
Lets try to run all tests and compare their results

```
test.all <- allDA(TS, predictor = "HMPbodysubsite")
```

```
## predictor is assumed to be a categorical variable with 2 levels: Saliva, Tongue_dorsum
## Seed is set to 123
```

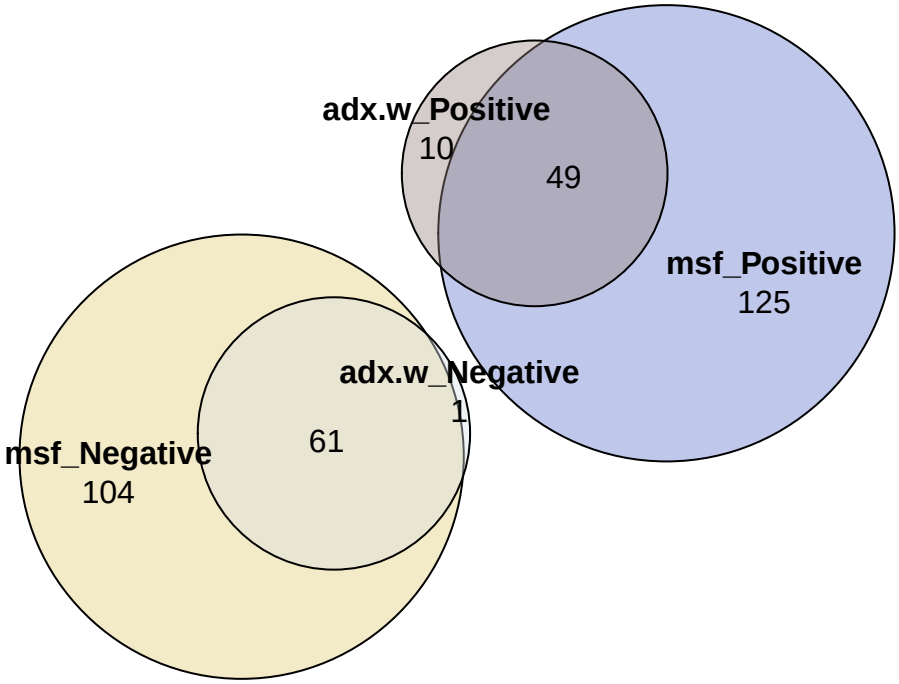
Euler diagrams of three select methods

```
vennDA(test.all, tests = c("msf","adx.w","ere"))
```



Split in negative and positive fold changes

```
vennDA(test.all, tests = c("msf","adx.w"), split = TRUE)
```


adx.w_Positive
10
49
msf_Positive
125
adx.w_Negative
1
61
msf_Negative
104