ThromboSeq Source Code
https://github.com/MyronBest

Revision 19th of November 2018

Error corrections
thromboSeqTools_PreProcessing_2.R
line 624: abline in plot now drawn according to provided threshold.
Lines 648-658: re-introduced figure boxplots of age per group, only if 'Age' is available in the columns of dgeIncludedSamples$samples.

thromboSeqTools_ANOVA.R
lines 165-167: check whether figureDir exists, if not, create it.

thromboSeqTools_PSO.R
Line 233; training.samples should be evaluation.samples
Lines 311-313: check whether figureDir exists, if not, create it.
Line 315: adjusted location of PSO_optimizationPlots.pdf according to provided figureDir.
Line 609: Training Series should be Evaluation Series
Line 1997: samples.for.evaluation should be samples.for.validation

All other script no adjustments. End of revision.

Revision 28th of November 2018
thromboSeq_source_code_v1.1

New functions

- Rule-in/rule-out algorithm analysis (thromboSeqPSO.readout-function). Provide percentages at which a predictive strength cutoff may be selected in the evaluation series and apply this to the validation series. Output is stored in files with different file names.
- Provide additional sample filtering steps for the validation series (thromboSeqPSO.readout-function). Provides option to select e.g. only stage I tumors for validation readout. Output is stored using different file names based on selection criteria.
- Add additional clinical info to output svm.summary (thromboSeqPSO.readout-function). Provides option to include several clinical characteristics into the svm.summary-table of the evaluation and validation series.
- Classification of new DGE-object with unseen HTSeq files in trained algorithm (thromboSeqPSO.readout-function).
- Creation of boxplot with scatterplot overlay with number transcripts detected per group.
- Perform LOOCV analysis of a given dataset (thromboSeq.LOOCV-function). Enables for leave-one-out cross validation analysis of a provided DGE-dataset. Biomarker transcript panel is selected by thromboSeq ANOVA statistics and an SVM machine learning algorithm.

- Confusion matrix output of the evaluation and validation series are adjusted according to the maximum accuracy (predictive strength threshold) reached in the evaluation series. An additional output confusion matrix (count.matrix.updated) is added to results.classification.evaluation.RData and results.classification.validation.RData.

Error corrections
thromboSeqTools_PreProcessing_2.R
Lines 813-816: Corrected error with minor variation in validation series classifications due to selection of incorrect reference sample for TMM normalization.

thromboSeqTools_ANOVA.R
No changes

thromboSeqTools_PSO.R
Lines 940-941, 1940-1941, 2179-2180, 2221-2222, 3374-3375: Corrects error in which dge$raw.counts and dge$ruv.counts were not reduced same as dge$counts.
Lines 966, 977, 987, 999, 1014, 1030, and 1047: Corrected input argument because may cause reference error.
Lines 1014, 1030, 1050, 2135, and 2177: Replace dge by dge-object provided in input-arguments for function. Prevents that incorrect DGE previously employed in analyses is used for downstream validation analysis.
Lines 2689: Corrected error in which no numeric value was passed for replace counts function.

Revision 29th of December 2018 and 20th of March 2019
thromboSeq_source_code_v1.2, thromboSeq_source_code_v1.3

New functions
- thromboSeqPSO.readout-functions stores biomarker panel as a csv file.

Error corrections
thromboSeqTools_PreProcessing_2.R
Added headers to functions

thromboSeqTools_ANOVA.R
Added headers to functions
Line 253 Print PSO-selected FDR threshold for heatmap plotting

thromboSeqTools_PSO.R
Added headers to functions
Lines 326-329: corrected error when Inf-output was provided in logged.PSO.distribution
Lines 393, 1545, 1920, 2166, 2443, 2693, and 2923: applied RUVg normalization only to training and evaluation series. In previous versions validation series were included in this process incorrectly.
Line 1010: forwarded clinical info to be present in output to thrombo.algo.classify.training.set-function.

Lines 1351-1353: included additional input arguments for 'controls'-function, which are passed to the individual shuffled and iteration functions.
Lines 1394-1410: included if-statements to check input arguments.
Lines 1412, 1539: included dge.tool creation to circumvent overwriting of dge once snapshot RData file is loaded.
Line 1649: Forward only the dgeTraining containing training and evaluation series to evaluation-function.
Lines 1663-1669: Provide additional input variables to narrow the validation series same as for true validation function (in thromboSeqPSO.readout-function).
Line 1746: perform.RUVg.correction.validation-function has been renewed; in case of data evaluation it only selects the training and evaluation series for data RUV correction, in case of validation it first corrects training and evaluation series and RUV-corrects each validation sample one-by-one in a loop supplied by the training and evaluation series. This because it appeared that once all samples are corrected at once the validation samples do have (minor) effect on the full correction step. Adjustments in Lines 393, 1545, 1920, 2166, 2443, 2693, and 2923 are complementary to this update. Output is now only narrowed to training and evaluation series (in case of 'evaluation' or 'LOOCV') or all samples ('validation').
Lines 2179 and 2951: added refColumn-input to calcNormFactors readout function.
Lines 2249-2251 and 2263-2265: corrected rule-in/-out threshold selection.
Line 2432: skipped filtering of dge due to same effect in perform.RUVg.correction.validation-function in line 2442.
Lines 2249 and 2263: corrected best threshold correction, by adding roc.summary$xValues == / roc.summary$yValues ==; indicating true line with provided rule.in/.out threshold.

Revision 29st of August 2019
thromboSeq_source_code_v1.4

New functions

- digitalSWARM, digitalSWARMshuffled, thrombo.algo.anova.up, thrombo.algo.anova.down, and TEPscore. Functions enable for digitalSWARM analysis, readout, and shuffled labels specificity-analysis.

Error corrections
thromboSeqTools_PreProcessing_2.R
None

thromboSeqTools_ANOVA.R
Lines 121-123: Added to function thromboSeqANOVA: iterations for digitalSWARM temporary storage.

thromboSeqTools_PSO.R
Line 1781: in case no RUV-correction is applied, skip this module.
Lines 1816, 1884: added option for RUV-correction readout in digitalSWARMsetting (including evaluation series)

Revision 17st of April 2021
thromboSeq_source_code_v1.5

New functions

- filter.for.platelet.transcriptome.Train.Eval.group and thromboSeqQC.TrainEval. Functions enable to perform the previously employed filter.for.platelet.transcriptome and thromboSeqQC-functions selection procedures on only the training and evaluation series, to guarantee independence of the validation series in these steps. Also colors into the graphs can be adjusted.
- Class weights are introduced into the tune.svm functions. Class weights can be enabled via the thromboSeqPSO-function, and take during the SVM-training process unbalanced classification groups into account.
- Training of the algorithm towards 99% specificity (rule-in) can be enabled via the rule.in.optimization-function into the thromboSeqPSO-function.

Error corrections

thromboSeqTools_PreProcessing_2.R

Lines 526-528: Fix error in which more k.variables are provided than axes available in RUVg

thromboSeqTools_ANOVA.R

None

thromboSeqTools_PSO.R

The ppso-function incorrectly passed very low values of FDR (e.g. 1e-10) towards the swarm algorithm and thrombo.algo-function. To circumvent this, the thrombo.algo-function was adjusted to pass numbers of RNAs of the ANOVA list instead of FDR thresholds. In case no FDR value was passed for optimization, the number of RNAs at FDR<0.05 is included. The swarm variable is still termed 'FDR' by default. Following filtering for highly correlated RNAs, only the selected.transcripts in now filtered.

Line 1907-1910: If/else-function added when no variables in RUV need to be corrected; replace post-correction with dge$counts.