Amplicon result supplementary instruction (06)

2023 / 8 / 28

06.StatisticalTest

1 Adonis 【Adonis】

1.1 group1 【adonis analysis result of group1】

1.1.1 bray_adonis.txt 【Statistical result of Adonis】

Column number	Column name	Instruction
1	Group	Group name
2	Df	Degree of freedom
3	SumsOfSqs	Sum of square
4	MeanSqs	Mean square error (SumsOfSqs/Df)
5	F.Model	F value
6	R2	The higher the R square means the higher contribution of the grouping scheme to the difference between the sample
7	Pr(>F)	P value

Reminder: the data in the bracket comes from the residual。

2 Anosim 【Anosim】

2.1 stat_anosim.txt【Statistical result of Anosim 】

Column number	Column name	Instruction
1	Group	Group name
2	R-value	-1 < R value <1 when R-value is positive, it means the inter-group variation is significant, while the negative R-value means the inner-group variation is larger than inter-group variation
3	P-value	P value, $p < 0.05$ (statistical significantly)

3 MRPP 【Mrpp】

3.1 group1 【Mrpp analysis result】

3.1.1 stat_mrpp.txt【Statistical result of Mrpp analysis】

Column number	Column name	Instruction
1	Group	Group name
2	A	A>0， inner-group variation < inter-group variation A<0, inner-group variation > inter-group variation
3	observed-delta	Smaller observed-delta means smaller inner-group variation
4	expected-delta	Larger expected-delta means larger inter-group variation
5	Significance	Significance < 0.05 (statistical significantly)

4 metagenomeseq【Metagenomeseq】

4.1 group1 【Metagenomeseq analysis result of group1 】

4.1.1 phylum/...【Metagenomeseq result at each taxonomy level】

4.1.1.1 diffHeatmap 【heatmapofabundancetop10taxa】

4.1.1.1.1 annotation.heatmap.{pdf,png} 【heatmap of abundance top10 taxa that are significant between group】

4.1.1.2 *-* 【Metagenomeseq analysis result】

4.1.1.2.1 *.psig.xls 【source from Metagenomeseq analysis result, use p | logFC | >1 value <0.05 as threshold】

4.1.1.2.2 *.qsig.xls 【source from Metagenomeseq analysis result, use q value <0.05 as threshold】

4.1.1.2.3 *.Volcano.{pdf,png} 【volcano plot of significant taxa】

4.1.1.2.4 *.Volcano.label.{pdf,png} 【volcano plot of significant taxa】

4.1.1.2.5 boxplot 【boxplot of significant taxa 】

Metagenomeseq software (http://cbcb.umd.edu/software/metagenomeSeq) is performed and the result are obtained by hypothesis testing of abundance between group, MetagnomeSeq analysis is often used to compare objects at different classification levels (such as phylum, genus, etc.).The difference in abundance between the two groups of samples is screened using | logFC |>1 and P value<0.05 as the significance threshold for the difference Identify microbial species with significant differences between the two sets of samples. If there is no relevant information in the table, it indicates that there is no significant difference in species.The multiple hypothesis test corrects the P-value of FDR to obtain the q value. (group number > =2, sample number in group >=3)

.psig.xls、.qsig.xls 表格说明如下：

Column number	Column name	Instruction
1		Taxa name
2	logFC	log-fold-changes between each pair of groups
3	se	Standard Error
4	pvalues	P value
5	adjPvalues	q value (adjust p-value)

5 simper 【simperanalysis】

5.1 group1 【simper analysis of group1】

5.1.1 phylum/...【simper analysis result at each taxonomy level】

5.1.1.1 *-*【simper analysis of top10 taxa】

5.1.1.1.1 *-*.{xls,pdf,png}【statistical result of simper analysis】

-.xls table：

Column number	Column name	Instruction
1	species	Taxa name
2	average	Average contribution to overall difference
3	sd	Standard deviation
4	ratio	The standard deviation of groupA
5	ava	Average abundance in groupA
6	avb	Average abundance in groupB
7	cumsum	Accumulation contribution after sort
8	Contribution	Percentage of contribution to difference between group
9 to last column	Sample name	Sample name

6 ttest【ttest】

6.1 group1【ttest of group1】

6.1.1 phylum/...【ttest analysis at each taxonomy level】

6.1.1.1 *-*【statistical result of ttest of top10 taxa】

6.1.1.1.1 *.psig.xls【source from ttest analysis result，use p value <= 0.05 as threshold】

6.1.1.1.2 *.qsig.xls【source from ttest analysis result, use q value <= 0.05 as threshold】

6.1.1.1.3 *.Volcano.{pdf,png}【volcano plot of significant taxa without sample name】

6.1.1.1.4 *.Volcano..label.{pdf,png}【volcano plot of significant taxa with sample name】

6.1.1.2 boxplot【boxplot of significant taxa】

6.1.1.2 .1 *.{pdf,png}【boxplot of significant taxa】

.psig.xls、.qsig.xls table：

Column number	Column name	Instruction
1	Taxa	Taxa
2	avg(group1)	Average of groupA
3	sd(group1)	Standard deviation of groupA
4	avg(group2)	Average of groupB
5	sd(group2)	Standard deviation of groupB
6	p.value	P value
7	q.values	q value (adjust p-value)
8	interval lower	Lower confidence limit
9	interval upper	Upper confidence limit

7 lefse【LEfSe】

7.1 group1【lefse analysis result of group1】

7.1.1 *-*【lefse analysis result】

7.1.1.1 LDA.*.all.res【statistical result of lefse analysis】

7.1.1.2 LDA.*.draw.res【data for lefse visual display】

7.1.1.3 LDA.*.{svg,png}【barplot of LDA score】

7.1.1.4 LDA.*.tree.{svg,png}【tree plot of Lefse】

8 randomForest【random forest result】

At least two group to perform random forest analysis, and at least 15 sample in each group.

8.1 group1 【random forest analysis result】

8.1.1 class/...【random forest analysis at (c,o,f,g,s) taxonomy level】

8.1.1.1 trainset_group_max_roc.{pdf,png} 【maximum AUC value of training set of variable, ROC curve plot】

8.1.1.2 *-* 【random forest result among group】

8.1.1.2.1 imptop10MD.{pdf,png} 【top10 variable averageaccuracydeclineplot】

8.1.1.2.2 imptop10MG.{pdf,png} 【top10 variable average Giniindexdeclineplot】

8.1.1.2.3 testset_auc.{pdf,png} 【ROC curve of testing set of important variable】

8.1.2.4 testset.point_auc.{pdf,png} 【AUC plot of testing set of important variable】

8.1.1.2.5 trainset_auc.{pdf,png} 【ROC curve of training set of important variable】

8.1.1.2.6 trainset.point_auc.{pdf,png} 【AUC plot of training set of important variable】

8.1.1.2.7 3/5/... 【random forest with (3,5,10,50,70,100,150,200) number of variable】

8.1.1.2.7.1 cross_validation.*testset.marker.predict.xls 【prediction probability of testing set of cross-validation, prediction index is AUC】

8.1.1.2.7.2 cross_validation.*trainset.marker.predict.xls 【prediction probability of training set of cross-validation, prediction index is AUC】

8.1.1.2.7.3 cross_validation_pick_*.im.var.txt 【feature of important factor through cross-validation】

8.1.1.2.7.4 cverrof.{pdf,png} 【screening number of variable and error ratio】

8.1.1.2.7.5 impplot_MeanDecreaseAccuracy*.{pdf,png}【average accuray decline plot】

8.1.1.2.7.6 impplot_MeanDecreaseGin*.{pdf,png} 【average Gini index decline plot】

8.1.1.2.7.7 testset.probabilty.xls 【prediction probability of testing set of sample】

8.1.1.2.7.8 testset.ROC.{pdf,png} 【ROC curve of testing set】

8.1.1.2.7.9 testset.roc.xls 【testing set, AUC value: 1 means average, 2 means minimum, 3 means maximum】

8.1.1.2.7.10 top_*testset.var_predict_box_plot.{pdf,png} 【predicted accuracy plot of testing set】

8.1.1.2.7.11 top_*trainset.var_predict_box_plot.{pdf,png} 【predicted accuracy plot of training set】

8.1.1.2.7.12 top_*var_imp.xls 【contribution of selected important variables to the sample and model construction】

8.1.1.2.7.13 trainset.probabilty.xls 【predicted probability of training set of sample】

8.1.1.2.7.14 trainset.ROC.{pdf,png} 【ROC curve of training set】

8.1.1.2.7.15 trainset.roc.xls 【training set, AUC value: 1 means average, 2 means minimum, 3 means maximum】

8.1.1.2.7.16 Instruction of graph result：

◆ Cross-validation

x-axis：the number of variables be used to construct the model. y-axis: corss-validation error. Greyline is the average of the ten cross-validation error rates. The graph illustrate that the local minimum of the curve correspond to the optimal number of variables, after this point, with the increase of variables number, the error rate show a steady increase trend. The importance of variables is predicted by random forest. When other variables do not change, construct a new model with the change of variables, if the new model show higher predicted error rate, it means this variables is more important.

◆ Variables screening

MeanDecreaseAccuracy MeanDecreaseAccuracy measures the extent to which the prediction accuracy of random forest is reduced when the value of a variable is changed to a random number. Higher MeanDecreaseAccuracy means the variable is more important. MeanDecrease used the Gini index to calculate the effect of each variable on the heterogeneity of the observed values at each node of the classification tree to compare the importance of variables.

◆ ROC curve

ROC curve be used to evaluate the fitness of a binary classifier. It is also an indicator based on statistics to judge the quality of grouped information. AUC (Area Under Curve) is the area under the ROC curve. In general, this value is between 0.5 and 1. When AUC>0.5, if AUC more close to 1, it means the prediction is more accurate

◆ Model prediction accuracy

This graph show the predicted accuracy when use the important variables. X-axis is the group, y-axis is the predicted accuracy