Random pooling algorithm

Background

Collecting a large number of human samples is sometimes not possible, and thus often a limited number of samples are available for analysis. However, sample size is a confounding factor for computing differential gene expression. In order to minimize this effect and false positives caused by small sample sets, we developed a "random pooling algorithm" combined with "DESeq2" to identify reliable markers distinguishing any two populations.

Description and details

The aim of this algorithm is to identify stable, differentially expressed genes based on sub-sampling from the original data set followed by DESeq2. The details of the algorithm include:

1) Estimate the sample sizes of the two groups and select a small sized group as the random pooling reference group.
2) Set up *i*, which defines the range for subsampling from group1 and group2, where *i* can maximum be N-1 of the total number of samples in the reference group. In our case, *i* is from 3 to 11, as our total N number is 12.
3) Randomly select the same number *i* of the samples from group1 and group2, respectively.
4) Calculate the differentially expressed genes by using selected samples;
5) Repeat points 3-4 for *j* number of times.

Usage and arguments

The inputs include:

1) Matrix1: the read count matrix of group1, in which the rows are genes and the columns are samples.
2) Matrix2: the read count matrix of group2, in which the rows are genes and the columns are samples.
3) Sn: it represents "j", which means the repeating times in random pooling process.

4) Group1Inf: a dataframe, containing the group1's information.
5) Group2Inf: a dataframe, containing the group2's information.
6) OutDir: the folder to save the results.
7) Parallel_TF: a logical value. If it is TRUE, the DESeq will run by using BiocParallel; if it is FALSE, the DESeq will run by single core.
8) Ncores: the number of cores used when running DESeq.
9) Paired_TF: the samples are paired or not.