← Go Back (/)

ReplicatesQC, v13

Removes cell line replicate samples that fail quality control check of cell line replicate correlations and distribution scores.

Author: Barbara Weir, Sara Howell;The Broad Institute

Contact: showell@broadinstitute.org

Algorithm Version:

Introduction

ReplicatesQC is a GenePattern module that implements quality control steps during the analysis of shRNA pooled screen data. The user inputs at least one and up to three GCT files containing shRNA data, with each file containing data from a separate batch. The correlation scores for cell line replicates are calculated and a replicate sample will fail QC if the correlation score is less than the threshold, which is the 75th percentile of the non-replicate correlation scores. The distribution scores for cell line replicates are calculated and a replicate sample will fail QC if the distribution score is less than the threshold, which is the mean of the distribution scores minus the standard deviation of the distribution scores. If there are not at least 3 replicate samples passing QC for each cell line, all replicates for that cell line will be added to the failed replicate sample list.

References

Contact author

Parameters

Name	Description
shRNA data file*	Format is a GCT file where each row corresponds to a unique shRNA sequence. At least one file is required and up to three files can be uploaded.
extension *	User-defined prefix of output file names.

* - required

Input Files

1. shRNA data file
    - Format is a GCT file where each row corresponds to a unique shRNA sequence.

Output Files

1. File containing the replicates that failed quality control removelist_reps_*[extension]*.txt
    - A text file containing the replicates that failed quality control, along with the reason why each replicate failed.
2. Boxplot of non-replicate correlations NonReplicate_corr_boxplot_*[extension]*.pdf
    - A pdf of the boxplot of non-replicate correlations, displaying the distribution of non-replicate correlation values. Cell line replicates pass the correlation quality control step if their correlation values exceed the 75th percentile of the non-replicate correlation values.
3. File containing replicate correlation values Replicate_Correlations_*[extension]*.txt
    - A text file containing the pairwise replicate correlations for all cell lines. The 75th percentile of the non-replicate correlation values is included at the beginning of the file, which is the threshold a replicate needs to exceed in order to pass quality control.
4. File listing replicate samples that fail the replicate correlation requirements Reproducibility_Failing_Samples_*[extension]*.txt
    - A text file containing the replicate samples that have a correlation score lower than the threshold.
5. File listing replicate samples that fail the distribution score requirements Distribution_Failing_Samples_*[extension]*.txt
    - A text file containing the replicate samples that have a distribution score lower than the threshold, which is the 75th percentile of the distribution scores.
6. File listing all replicate samples with their distribution scores. Distribution_All_Samples_*[extension]*.txt
    - A text file containing all replicate samples, along with their distribution scores.
7. File containing replicate sample information Reps_Sample_Info_*[extension]*.txt
    - A text file containing a summary of the replicate sample, the cell line name, and the distribution score.
8. Histogram of distribution scores Histogram_75thprctile_*[extension]*.pdf
    - A pdf of the histogram of distribution scores. The red line represents the first threshold (the mean of distribution scores minus the standard deviation of distribution scores) and the magenta line represents the second threshold (the mean of distribution scores minus two times the standard deviation of distribution scores).
9. Boxplot of cell lines with worst distribution scores Worst_75th_prctile_*[extension]*.pdf
    - A pdf of the boxplot of cell lines with the worst distribution scores. The red line represents the first threshold (the mean of distribution scores minus the standard deviation of distribution scores) and the magenta line represents the second threshold (the mean of distribution scores minus two times the standard deviation of distribution scores).

Requirements

R-2.15

Platform Dependencies

Task Type:
RNAi

CPU Type:
any

Operating System:
any

Language:
R

Version Comments

Version	Release Date	Description
13	2014-10-08	updated job error status
12	2014-09-16	updated command line
11	2014-04-11	Converted documentation from pdf to html