This analysis tool can be used to rank genes for genetic perturbation screens. The tool takes a list of perturbations and associated numerical scores as input, and computes a score using one of two statistical methods, negative binomial distribution (with replacement) or hypergeometric distribution (without replacement).

1. Negative Binomial Distribution (STARS)

   The STARS score is calculated using the probability mass function of a binomial distribution. The calculation is performed for all perturbations that rank above a user-defined threshold, e.g. the top *x*% of perturbations from a ranked list. The value of the least probable perturbation for each gene is then assigned to the gene as the STARS score. Unless specified, STARS requires that at least two perturbations rank above the user-defined threshold for a gene to receive a STARS score. Permutation testing is also performed on the list of perturbations used in the experiment to generate a null distribution, allowing the calculation of p-values and false discovery rates (FDR) for hit genes. STARS also provides separate outputs for sgRNAs ranked in ascending and descending direction.

2. Hypergeometric distribution

   In this method, the rank of sgRNAs is used to calculate gene p-values using the probability mass function of a hypergeometric distribution. The list of sgRNAs can be ranked in both ascending and descending directions and the resulting p-values will be different in each direction. We choose to resolve this by calculating the average -$\log_{10}$(p-value) in both directions and picking the more significant one. The top n% of sgRNAs per gene can be used to calculate the average p-value with this method. The average log-fold change per gene is also reported and this can be used to assess the magnitude of effect.

Running parameters for both methods

1. Negative Binomial Distribution (STARS)
   a. Directionality of scores: Use “Positive” if the best perturbation has the highest/most positive value (enrichment) and "Negative" if the best perturbation has the lowest/most negative value (depletion). Use “Both” if you want the screen to be analyzed in both directions.
   b. Threshold: A number ranging from 0 - 100. This indicates the x% of sgRNAs for which a STARS score will be calculated. A value of 10 is standard but can be specified based on the signal in the particular biological assay.
   c. Include first barcode in calculation: Specify whether the first ranking perturbation for each gene should be used in the calculation of STARS score.
2. Hypergeometric distribution

a. Top n% guides per gene: n% guides per gene to be used in the calculation of average p-value and average log-fold change. The default value is 100% i.e. use all guides per gene to calculate average p-value and average log-fold change.
b. Create negative control "dummy" genes with n guides per gene: Number of control guides to be used to create "dummy" control genes. Average number of guides per gene in the library is the suggested value.
c. # of genes to label: The number of genes to be labeled on the enrichment and the depletion side of the volcano plot.
d. Min. # of guides required per gene: Minimum number of guides required for a gene to be plotted on the volcano plot.
e. Max. # of guides required per gene: Maximum number of guides required for a gene to be plotted on the volcano plot.
f. Display no-site controls: Determines whether no-site controls will be displayed on the resulting volcano plot.
g. Display one-non-gene-site controls: Determines whether one-non-gene-site controls will be displayed on the resulting volcano plot.

Input formats:

Format of chip file: .txt file with first column containing the sgRNAs and second column containing the gene identifiers of the sgRNA targets.

Format of Data file: .txt file with the first column containing the list of sgRNAs as specified in the first column of the chip file and the consecutive columns containing the numerical inputs for each condition.

Output file details:

This tool generates separate output files for every column in your input file. The column name will be included in the output file name.

1. Negative Binomial Distribution (STARS)

   Only the genes with at least 2 perturbations ranking above the threshold will receive a STARS score and be reported in the output file. If the first perturbation was used to calculate the STARS score, all the genes with at least one perturbation ranking above the specified threshold will receive a STARS score and be reported in output file. The output file contains 10 columns:
   a. Gene identifier, from column 2 of the chip file
   b. Number of perturbations targeting the gene
   c. Ranks of perturbations targeting the gene
   d. Identity of perturbations
   e. Within-gene-rank of the least probable perturbation
   f. STARS score: -log10(value of least probable perturbation)
   g. Average score: Average of negative log of the values of all perturbations ranking above the threshold
   h. P-values calculated using the null distribution specified
   i. False Discovery Rate (FDR) calculated using permutation testing
   j. q-value

2. Hypergeometric distribution

   All the genes in the library will be reported in the output file along with a pdf of the volcano plot. The output file contains 10 columns:

   a. Gene identifier, from column 2 of the chip file
   b. Average log-fold change of n% guides per gene
   c. Average -log10(p-value) of n% guides per gene
   d. Number of perturbations targeting the gene
   e. Identity of perturbations
   f. Individual log-fold changes of the perturbations
   g. Ranks of the individual perturbations in the ascending direction
   h. -log10(p-values) of the individual perturbations in the ascending direction
   i. Ranks of the individual perturbations in the descending direction
   j. -log10(p-values) of the individual perturbations in the descending direction