STARS is a gene-ranking algorithm for genetic perturbation screens. This algorithm takes a list of perturbations and associated numerical scores as input, and computes a score using the probability mass function of a binomial distribution:

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where, *n* is the total number of perturbations targeting a gene, *k* is the within-gene-rank of the perturbation, and *p* is the ratio of the rank of the *k*th perturbation over the total number of perturbations in the experiment. This 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 is written in Python. To run these two scripts successfully, you will need the following: Python 2.7; pandas; Numpy; Scipy; Statsmodels

Step1: Null distribution -- *stars_null_v1.2.py*

This Python code generates a null distribution for the list of perturbations used in an experiment. **The null distribution needs to be generated only once for a given set of perturbations at a particular threshold.**

v1.2:

- ***Same input file can be used in Step1 and Step2***
- ***Header row is now required for Step1 input file***

v1.1: Additional input argument included

Inputs:

--input-file: Tab-delimited file with the list of unique perturbations in the first column. Usually, this is the sequence of the perturbation. **A header row is required.**

--chip-file: Tab-delimited file to map the perturbations to genes. The first column should be the same perturbations as in the input-file, but this file does not require that every perturbation be unique. For example, if one perturbation maps to two genes, then that perturbation should be listed twice, once for each gene. The second column should be the gene identifier, such as Gene Symbol or Gene ID.

--thr: Threshold percentage, a number ranging from 0 - 100. This indicates the x% of perturbations for which a STARS score will be calculated. A value of 10 is standard, but can be specified based on the signal of the particular biological assay, e.g. what fraction of perturbations are receiving a meaningful rank. This value **must be matched** to the value used in Step 2.

--num-ite: Number of permutations to run. We recommend 100 - 1000. The more permutations, the lower the bound of the p-value and FDR calculations in Step 2.
--use-first-pert: Specify whether the first ranking perturbation for each gene should be used in the calculation of STARS score. STARS v1 required that at least two perturbations rank above the user-defined threshold for a gene to receive a STARS Score. *This is an **optional** argument; if it is not specified, by default the first perturbation will NOT be used to calculate the STARS Score for each gene.* This value **must be matched** to the value in Step 2.

Output:
Tab-delimited file of null distribution to be used as input in STARS in Step 2.

To run the code, type the following in your terminal:
python stars_null_v1.1.py --input-file <Path to inputfile> --chip-file <path to perturbations-to-gene mapping file> --thr <threshold> --num-ite <Number of iterations> ***--use-first-pert <Y/N>***

Step 2: STARS -- *stars_v1.2.py*

This Python code performs STARS analysis on a list of perturbations that have an associated numerical value. STARS will first rank the perturbations by the value. It will then use that ranked list of perturbations and the redundancy of perturbations mapping to a gene to assign a STARS score.

Additional input argument included in v1.1

Inputs:
--null: Null distribution generated in Step 1. Please note that this file is only mathematically valid when used with the same chip-file and same threshold value.
--chip-file: Tab-delimited file to map the perturbations to genes. This is the same file as used in Step 1 in the generation of the null distribution file.
--thr: Threshold percentage, 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 of the particular biological assay, e.g. what fraction of perturbations are receiving a meaningful rank. This value **must be matched** to the value used in Step 1.
--input-file: Tab-delimited file with the **list of perturbations** in the first column and one or more columns of **numerical values** assigned to the perturbations. **A header row is required.** The name of the header row for each column of numerical values is used to generate the name of the output file.
--dir: Directionality of numerical values in the input file. Use “P” if the best perturbation has the highest/most positive value and "N" if the best perturbation has the lowest/most negative value. All the columns will be sorted in the specified direction. If the input has multiple columns, a different direction **cannot** be specified for each column.

***--use-first-pert*:** Specify whether the first ranking perturbation for each gene should be used in the calculation of STARS score. This value **must be matched** to the value used in Step 1. STARS v1 required at least two perturbations rank above the user-defined threshold for a gene to receive a STARS Score. *This is an **optional** argument; if it is not specified, by default the first perturbation will NOT be used to calculate the STARS Score for each gene.*

To run the code, type the following in your terminal:
python stars_v1.1.py --input-file <Path to inputfile> --chip-file <path to guides-to-gene mapping file> --thr <threshold> --dir <direction of scores (P/N)> --null <Path to null distribution file> ***--use-first-pert <Y/N>***

Output:
This code generates a separate output file for every column in your input file. By default, 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 the output file. All the output files will be stored in your working directory.

The output file contains 9 columns:

1. Gene identifier, from column 2 of the chip-file
2. Number of perturbations targeting the gene
3. Ranks of perturbations targeting the gene
4. **Identity of perturbations**
5. Within-gene-rank of the least probable perturbation
6. STARS Score: -log10(value of the least probable perturbation). By default, the value of the first-ranked perturbation for every gene is **not** considered for the STARS Score, thereby ensuring that all genes in the resulting hit list have evidence from at least two perturbations. But, if the "--use-first-pert" argument has been set to "Y", the first-ranked perturbation for every gene **is** considered to calculate the STARS Score.
7. Average Score: Average of the negative log of the values of all perturbations ranking above the threshold. By default, Average Score **does** take into consideration the value of the perturbation ranking first **if** there is a second perturbation that scores above the threshold.
8. p-values calculated using the null distribution specified
9. False Discovery Rate (FDR)
10. q-value corrected FDR