HIT'nDRIVE: Data Preparation Guide

Raunak Shrestha

July 22, 2019

Disclaimer: Below are recommended suggestions to be considered while preparing input data for HIT'nDRIVE. The suggestions may or may not be applicable depending on the biological or computational question the user is trying to address or the nature of omic-data available. It is up to the user's discretion to follow, partially follow, not follow, or modify this guide. It is up to the user, how the input data is prepared. As long as the basic data structure (or format) is strictly followed, HIT'nDRIVE should run smoothly.

1 How to prepare the Alteration data

Alteration data is one of the input essential to run HIT'nDRIVE algorithm. The file contains the name of genes that are identified to have evidence of sequence level or structural alterations such as mutations, copy-number aberrations, in-dels, structural variations, gene-fusions, etc per sample (or patient) analyzed. The following guide is intended for mutation and copy number aberration data preparations.

In essence, we want to feed the most confident set of pathogenic alterations to HIT'nDRIVE.

1.1 Mutation data (for protein coding mutation changes)

- First, get the final confident list of mutation calls.
- Annotate the mutations using mutation annotation software such as Annovar.
- Convert the file to the Mutation Annotation Format (MAF) file.
- Select only **non-silent mutations** i.e. select those mutations that effects the protein.
- Remove those genes that encodes proteins of more than 1500 amino acids in length (i.e. filter out longer genes) but rescue (i.e. retain) any genes that are present in the Cancer Genes Census (CGC) database.

1.2 Copy Number Alterationa data

- Since copy number alteration spanning different chromosomal regions effects a large number of genes across the region, it is recommended that only those genes with "High Copy Gain" and "Homozygous Copy Loss" status be selected.
- If using GISTIC processed data, select genes with copy number states "+2" or "-2".
- The Mutation and Copy number alteration data may be analyzed separately using HIT'nDRIVE and then combined the results later.

1.3 Data Format

Tab separated file with the sample in the first column and the altered gene in the second column.

- The first column must have a header name as 'SampleID'. This column contains id of the sample or patient analyzed. Please avoid any symbol such as '-' that resembles mathematical operative symbol. If present the CPLEX may treat these entities as mathematical operations symbols.

SampleID	GeneName
PATIENT_01	GENE_A
PATIENT_01	GENE_B
PATIENT_01	GENE_C
PAITENT_01	GENE_D
PATIENT_02	GENE_A
PATIENT_02	GENE_B
PATIENT_03	GENE_C
PATIENT_04	GENE_D
...	...

- The second column must have a header name as 'GeneName'. These genes are those that have any evidence of sequence level or structural alterations such as mutations, copy-number aberrations, in-dels, structural variations, gene-fusions, etc. Note: Gene names containing '-' symbols will be automatically converted to that with non-mathematical symbol that are permitted by CPLEX.
- Note: All 'GeneName' present in this file MUST be a subset of genes present in the interaction network. Please remove any genes that are not represented in the interaction network.
- Note: All unique 'SampleID' present in this file MUST have one-to-one relation in the expression outlier data file. Please remove any samples that does not have matching expression outlier data.

2 How to prepare the Expression Outlier data

- Expression Outlier genes are those genes that are aberrantly expressed in a patient sample.
- The procedure to identify outlier genes largely depends on the biological or computational question the user is trying to address or the nature of omic-data available.
- Either RNA expression data or Protein expression data can be used identify expression outlier genes.

2.1 Identifying expression outlier genes

- Here, we define the outlier genes as those that aberrantly express in tumor sample as compared to the normal samples.
- For this, we will use **Generalized Extreme Studentized Deviate (GESD)** test. Follow the instructions to run GESD as described in the follwing link: `https://github.com/raunakms/GESD`
- Use GESD test to compare the transcriptome profile of each tumor sample (one at a time) with that from a number of available normal samples.
- For each gene, if the tumor sample is identified as the most extremely deviated sample (using critical value $\alpha = 0.1$, the corresponding gene is labeled as an expression outlier gene for that tumor sample. Note: the critical value alpha in the GESD function many be modified as per user's requirement.
- This procedure should be repeated for every tumor sample.
- In this way, we will obtain a set of expression outlier genes per tumor sample.

2.2 Compute expression outlier gene weight

- The expression outlier gene weight are the numerical scores (weights) between [0,1] (0 is the lowest and 1 is the highest weight).
- The sum total of the expression outlier gene weight for each tumor sample should be equal to 1.

- First, assess the "group behaviour" of the expression outlier genes. For this, perform differential expression analysis comparing the tumor samples to the normal samples using Student's t-test (or wilcoxon rank sum test). This will result in *p-value* (P_{value}) of each gene analyzed. Consider only those genes that are identified as expression outlier genes (from previous steps using GESD test). Then compute *q-value* (Q_{value}) where, $Q_{value} = -log(P_{value})$

- Next, we will assess the extent of deviation of gene expression in the tumor sample as compared to the mean of the gene-expression values across all tumor samples in the cohort. For this transform the gene expression data to *z-score* (Z_{score}) values. $Z_{score} = \frac{x-\mu}{\sigma}$, where x is gene-expression value, μ mean of gene-expression values across all tumor samples, and σ is the standard deviation of gene-expression values across all tumor samples.
  Then get the absolute value of the *z-score* which we will call as Z_{abs} where, $Z_{abs} = |Z_{score}|$.

- For every expression outlier gene multiply the respective Q_{value} and Z_{abs}. $S = Q_{value} * Z_{abs}$

- Finally, for each tumor sample, get the normalized expression outlier gene weight (W) where, $W = \frac{S}{\sum S}$

2.3 Data Format

Tab separated file with the sample in the first column, the expression outlier gene in the second column, and the outlier gene weight in the third column.

SampleID	**GeneName**	**Weight**
PATIENT_01	GENE_A	0.40
PATIENT_01	GENE_B	0.05
PATIENT_01	GENE_C	0.35
PAITENT_01	GENE_D	0.20
PATIENT_02	GENE_A	0.25
PATIENT_02	GENE_B	0.10
PATIENT_02	GENE_E	0.05
PATIENT_02	GENE_W	0.30
PATIENT_02	GENE_X	0.30
PATIENT_03	GENE_A	0.06
PATIENT_03	GENE_D	0.17
PATIENT_03	GENE_W	0.46
PATIENT_03	GENE_X	0.24
PATIENT_03	GENE_Y	0.01
PATIENT_03	GENE_Z	0.02
PATIENT_04	GENE_A	0.03
PATIENT_04	GENE_B	0.05
PATIENT_04	GENE_C	0.20
PATIENT_04	GENE_F	0.15
PATIENT_04	GENE_G	0.34
PATIENT_04	GENE_H	0.03
PATIENT_04	GENE_X	0.10
PATIENT_04	GENE_Y	0.10
...	...	...

- The first column must have a header name as 'SampleID'. This column contains id of the sample or patient analyzed. Please avoid any symbol such as '-' that resembles mathematical operative symbol. If present the CPLEX may treat these entities as mathematical operations symbols.

- The second column must have a header name as 'GeneName'. These genes are those that are identified as expression outlier genes. Note: Gene names containing '-' symbols will be automatically converted to that with non-mathematical symbol that are permitted by CPLEX.

- The third column must have a header name as ‘Weight’. This is a numerical weight scaled between [0,1] given to each expression outlier gene. Note: the sum of all expression outlier gene weights for each sample must be equal to 1.
- Note: The expression outlier can be calculated using either gene microarray, RNA-seq, or Protein expression quantified from Mass spectrometry.
- Note: All ‘GeneName’ present in this file MUST be a subset of genes present in the interaction network. Please remove any genes that are not represented in the interaction network.
- Note: All unique ‘SampleID’ present in this file MUST have one-to-one relation in the expression outlier data file. Please remove any samples that does not have matching alteration data.