**Purpose:**

This script will classify if patients/samples from a “Test” dataset are similar to an input “Training” cohort.

This script uses Support Vector Machine (SVM) classification to identify patients that are similar to one or multiple Oncosplice clusters. For more information on SVM, see the below resources:

* <https://scikit-learn.org/stable/modules/svm.html>
* <https://www.youtube.com/watch?v=efR1C6CvhmE&ab_channel=StatQuestwithJoshStarmer>

This script is often used to validate if Oncosplice-identified clusters can be identified in an independent cancer cohort. In this case, the TCGA Oncosplice clusters and splicing profiles are used as the “Training” data, whereas the patient splicing profile from the validation cancer cohort is the “Test” dataset. Whereas splicing information and Oncosplice-derived patient groups are the most common input for this script, gene expression information and non-Oncosplice clusters/groups can also be used with this script.

**Input Files and Processing Steps:**

1. Get the splicing or gene expression file used to generate your group(s) of interest. If you are using a splicing matrix file (i.e. Event Annotation file), you will need to do some pre-processing before running the supervised analysis script:
   * The Event Annotation file should be filtered for splicing events that are not-missing for 75% of samples (EventAnnotation-75p). This file is generated by default when running Oncosplice.
   * The Event Annotation script needs to be reformatted to only keep the UID column and the PSI values. You can also do this manually in Excel is you hate efficiency, or you can run the reformat\_psi.py script.
     + Location: /Volumes/salomonis2/LabFiles/Tutorials/RunSupervisedAnalysis/reformat\_psi.py
     + Manually change file path in the script and the output file path/name and run using python 2.7.
   * Note: if you’re input is a gene expression file generated by AltAnalyze, none of these pre-processing steps apply.
2. Get the splicing or gene expression file from your independent (i.e. “Test”) cohort (e.g. the Event Annotation file from a non-TCGA breast cancer cohort). If you are using an Event Annotation file, the same pre-processing steps from Step 1 still apply.
3. Your “groups” file. This will be *the SVMOutputs/MergedResult.txt* file if running this analysis to identify Oncosplice clusters across cancer cohorts. You must transpose your MergedResult.txt file before running this script. This can easily be done in Excel or by using the transpose.py script.
   * Location for transpose.py: Volumes/salomonis2/LabFiles/Tutorials/RunSupervisedAnalysis/transpose.py
   * Command: transpose.py --groups [filepath/filename.txt]
   * Note: Your groups file MUST have your sample names as the column headers and the groups as the row names. This means that a three-column groups file will NOT work with this script and must be reformatted before running this script.
4. You will need a folder that contains the differential splicing/gene expression results for all of your groups in your groups file. If you do not already have this, generate these results with the metaDataAnalysis.py script:
   * Splicing:
     + python /data/salomonis2/software/AltAnalyze/stats\_scripts/metaDataAnalysis.py --dPSI 0.1 --p PSI --s Hs --adjp yes --pval 0.05 --i $DIR/Hs\_RNASeq\_top\_alt\_junctions-PSI\_EventAnnotation-75p.txt --m $DIR/MergedResult.txt --mf $DIR/mf\_File.txt
   * Gene expression:
     + python /data/salomonis2/software/AltAnalyze/stats\_scripts/metaDataAnalysis.py --p RNASeq --s Hs --adjp yes --pval 0.05 --f 1.5 --i $DIR/exp.TCGA-COAD-steady-state.txt --m $DIR/SVMClasses-COAD.txt --mf $DIR/comparisons.txt

**Script Command:**

python /data/salomonis2/LabFiles/Audrey/TestCode/Classification\_SVM.py \

--Training $DIR/Hs\_RNASeq\_top\_alt\_junctions-PSI\_EventAnnotation-75p.txt \

--group $DIR/MergedResult.txt \

--diffevents /data/salomonis2/path/to/PSI/Results/Events-dPSI\_0.1\_adjp \

--Test $DIR /Hs\_RNASeq\_top\_alt\_junctions-PSI\_EventAnnotation-75p.txt \

--normalize True \

--Multigroup True \

--adjp True \

--o-v-o yes

Bash script example: /Volumes/salomonis2/LabFiles/Tutorials/RunSupervisedAnalysis/SupervisedAnalysis.sh

From experience, changing the –normalize settings doesn’t change the classification results. Neither does the –Multigroup setting. –o-v-o is a way to differentiate between using Support Vector Classification (SVC), which uses a “one-versus-one” approach for multi-class classification. Alternatively, LinearSVC uses a “one-versus-rest” approach. Using SVC vs. LinearSVC has not been benchmarked in the lab yet, but SVC (--o-v-o) has anecdotally gotten the best results. Both settings can be run on your dataset, but make sure to keep this setting consistent if doing a supervised analysis on multiple “Test” datasets from the same “Training” dataset.

**Script Commands, Further Explained:**

|  |  |
| --- | --- |
| --Training | Input files used to generate previously identified clusters (Event Annotation or gene expression file).  If Event Annotation file, only keep UID and PSI value columns only (like a gene expression file).  Event annotation file should be filtered for splicing events that are not-missing for 75% of samples (EventAnnotation-75p). |
| --Test | Independent dataset to confirm presence of clusters from Training set.  If Event Annotation file, only keep UID and PSI value columns only (like a gene expression file).  Event annotation file should be filtered for splicing events that are not-missing for 75% of samples (EventAnnotation-75p). |
| --group | Binary groups file specifying previously identified clusters from samples in Training file (e.g. MergedResult file within SVMOutputs folder from Oncosplice).  Samples must be columns and groups must be rows (i.e. transpose MergedResult file). |
| --diffevents | Complete path to the folder containing the results from MetaDataAnalyses (prefix: Events-dPSI). |
| --normalize | Specifies whether to perform median normalization on training and test dataset. Default is False. |
| --Multigroup | Specifies if comparison is multi group or binary. Default is False |
| --adjp | Specifies using adjusted p-value for selecting the most differential splicing events. Default=False. |
| --o-v-o | “yes” or “no”  Support Vector Classification (SVC), which uses a “one-versus-one” approach for multi-class classification.  Alternatively, LinearSVC uses a “one-versus-rest” approach. |

**Troubleshooting Tips:**

If an error occurs, check the following first:

* Sample names between training and group file are in the same format
* Make sure the same exact patients are in the groups and training data files
* Make sure all of the groups mentioned in your groups file are also in your –-diffevents directory
* Groups file is formatted so that the clusters are along the rows and patient IDs are the columns

**Outputs:**

The SVMClasses file will contain your supervised analysis results. This file will be in a similar format to the MergedResult.txt groups file, with the group names as the columns and the patient/sample IDs as the rows. Note that the output of this script used “2” to denote a patient/sample is not associated with a given cluster instead of “0” like the MergedResult.txt file. This is a simple fix with the find – replace command in Excel.

* Note: Make sure you DO NOT do this replacement on the group names, which often contain numbers (R1-V2) when using Oncosplice clusters as the input groups.

There will also be a lot of filtered versions of your Test and Training files (-filtered.txt). You don’t need these files once the script has finished running and can delete these files.