visAPPprot Data Specifications

visAPPprot provides alternative visualizations for common omics (e.g., RNAseq; Genomic or Proteomic Microarrays) outputs. In the following pages we detail data specifications for our visualization application *visAPPprot*. There are precise formatting instructions for the data input for visAPPprot. The toy datasets provided with the system (see Part 3 of this document) demonstrate how input data should be formatted.

visAPPprot is designed to work with the input data format listed in this document. Using other formats may result in undefined behavior.

1. Microarray Data Input Requirements

All microarray data should be placed in the *microarray_data* directory.

Please only include alphanumeric characters in your data file names.

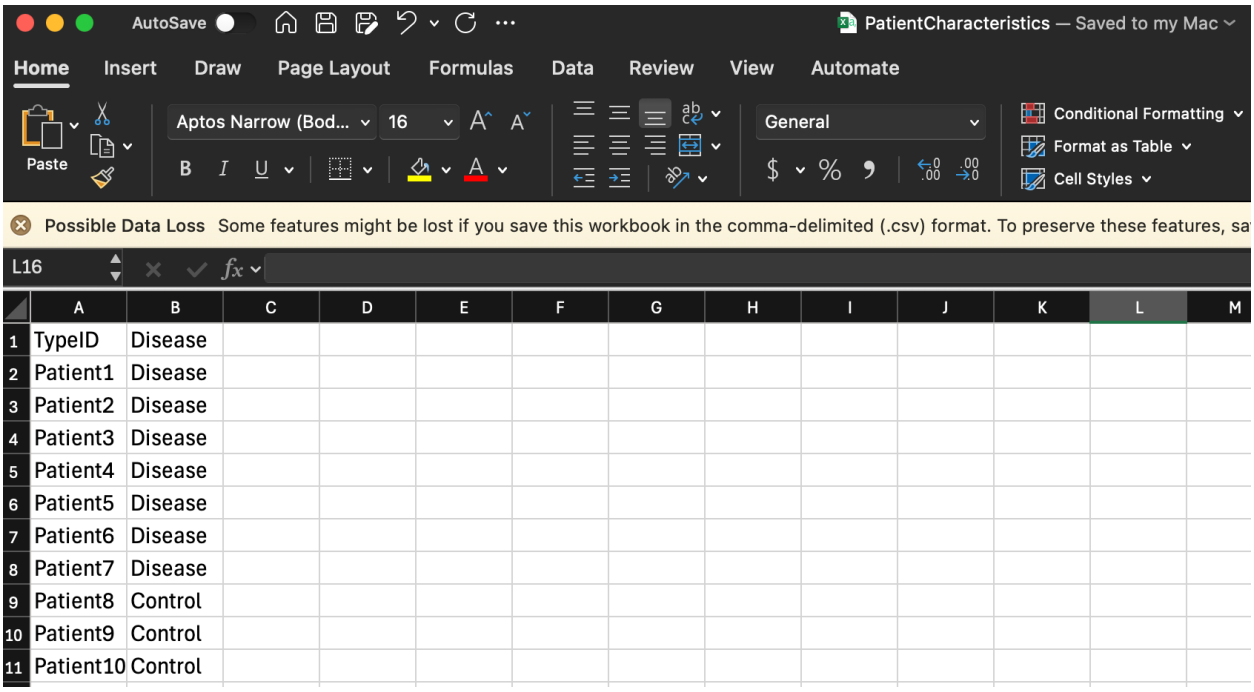
This system requires that the following columns be present in the data files:
"Block" "Row" "Column" "ID" "Name"

You also need a *PatientCharacter.csv* file that corresponds to the data in the *microarray_data* folder. The *PatientCharacter.csv* goes in the *processed*_data directory. Here are the details for constructing your *PatientCharacter.csv* file: In your *PatientCharacter.csv* file please make sure the entries for each data file are listed in the alphanumeric order of the TypeID. Please make sure you include columns "Disease" and "TypeID", where "Disease" specifies the phenotype (Disease, Control) and TypeID specifies the name of the data file (include all text before the file extension, which means do NOT include the file extension in the TypeID). The only phenotypes the system recognizes are "Disease" and "Control".

See below for example of data file directory and corresponding *PatientCharacter.csv* file structure:

Folder structure (note that microarray data is not limited to the GPR file type):
-microarray_data/
    -Patient1.gpr
    -Patient2.gpr
    -Patient3.gpr
    -Patient4.gpr
    -Patient5.gpr
    -Patient6.gpr
    -Patient7.gpr
    -Patient8.gpr
    -Patient9.gpr
    -Patient10.gpr
-processed_data/
    -PatientCharacter.csv

PatientCharacter.csv sample file:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	TypeID	Disease											
2	Patient1	Disease											
3	Patient2	Disease											
4	Patient3	Disease											
5	Patient4	Disease											
6	Patient5	Disease											
7	Patient6	Disease											
8	Patient7	Disease											
9	Patient8	Control											
10	Patient9	Control											
11	Patient10	Control											

The *microarray_data* folder is designed to handle 1 set of data at a time. This means all the data in the *microarray_data* folder must correspond to a single *PatientCharacter.csv* file. If you want to run a different set of data, please follow this set of instructions (these steps are already detailed in the System Manual):

1. Stop the visAPPprot system by closing the terminal.
2. Delete the data files in the microarray_data folder.
3. Add your new set of data files to the microarray_data folder.
4. Restart the visAPPprot system again by following Step 8 in the Installation document.
5. Go to localhost:8888 in your browser.
6. Set the download directory to your new dataset's download directory by following Step 9 from the Installation document.
7. Go back to the tab with localhost:8888 in your browser. You are now free to analyze your new dataset.

Note that any data format outside of what was described above is not guaranteed to work with the system.

2. Pathway Input Requirements

The visAPPprot system comes with a set of 126 pathways for visualizing the pathway map. You can add more pathways of your own by placing them in the *processed_data/* folder. Each pathway is stored in a separate .txt file.

3. Output File Structure

All computed expression matrices generated using the "Compute ExpMat" feature on the interface will be saved in the *processed_data* folder. The computed expression matrix files are named after the column used to

compute the matrix. The names of the computed expression matrices do not include the name of the dataset, so an existing expression matrix computed based on a selected column of data will be overwritten by a new expression matrix computed using a different dataset if using the same column name.

Sample computed expression matrix filenames:

- expmat_B635_Median.csv – expression matrix computed using column "B635 Median" from the microarray data
- expmat_SNR mean 635.csv – expression matrix computed using column SNR mean 635.

The differential expression output will be stored under the *differential_expression_tables* folder. The *differential_expression_tables* folder has subdirectories for each of the datasets you analyze. The names of the subdirectories correspond to the distinct *PatientCharacter.csv* filenames, eg. *PatientCharacter1/* and *PatientCharacter2/* for *PatientCharacter1.csv* and *PatientCharacter2.csv*, respectively.

The differential expression tables are named after the dataset, the visualization type, and the statistic used during differential expression: res1_Dataset_visualization_statistic_#.csv. The final # is the index, starting at 0, of this specific visualization type generated; this is used to differentiate between the same visualization (eg. volcano plot) generated for the same dataset (eg. PatientCharacter1) multiple times to keep track of every visualization viewed.

Every visualization generated will result in a res1.csv file containing the full contents of the differential expression output. For the volcano plot and the pathway map, a res2Shrink.csv file will also be generated with the contents of the differential expression output after lfcShrink, if any lfcShrink statistic was selected, and with an additional column containing the UniProt function information, if available, for differentially expressed biological entities. For the pathway map, a fgsea.tsv file will be generated containing the fgsea output.

Sample differential expression output filenames:

- res1_PatientCharacter1_volcano_GLM_0.csv – the full differential expression output for dataset PatientCharacter1 when generating the volcano plot using GLM statistic. The 0 at the end indicates this is the first time the volcano plot has been generated for this dataset.
- res2Shrink_PatientCharacter1_volcano_normal_1.csv – the differential expression output for dataset PatientCharacter1 when generating the volcano plot using the lfcShrink normal statistic. The 1 at the end indicates this is the second time the volcano plot has been generated for this dataset.
- fgsea_PatientCharacter2_pathway_0 – the fgsea table for dataset PatientCharacter2 when generating the pathway map. No statistics are applied here so no names of statistics are included in the fgsea.tsv filenames. The 0 at the end indicates this is the first time the pathway map has been generated for this dataset.

All downloaded visualizations will be saved to the *static/* folder. Similar to *the differential_expression_tables* folder, each dataset will have its own *download_*imgs folder under the static folder. The name of the *download_imgs* folder will correspond to the distinct PatientCharacter csv filenames, eg. *download_imgs_PatientCharacter1/* and *download_imgs_PatientCharacter2/* for *PatientCharacter1.csv* and *PatientCharacter2.csv*, respectively.

The visualizations are named after the visualization type, the dataset, and any normalization if applicable: visualization_Dataset_0.svg. Again, the final # is the index, starting at 0, of this specific visualization type generated; this is used to differentiate between the same visualization (eg. volcano plot) generated for the same dataset (eg. PatientCharacter1) multiple times to keep track of every visualization viewed.

Sample downloaded visualization filenames:

- volcano_PatientCharacter1_0.svg – the volcano plot for dataset PatientCharacter1. The 0 at the end indicates this is the first time the volcano plot has been generated for this dataset.
- pathway_PatientCharacter1_0.svg – the pathway map for dataset PatientCharacter1. The 0 at the end indicates this is the first time the pathway map has been generated for this dataset.
- heatmap_PatientCharacter1_normalized_0.svg – the heatmap for dataset PatientCharacter1, generated using the VST normalization of the dataset. The 0 at the end indicates this is the first time the heatmap has been generated for the normalized data.
- heatmap_PatientCharacter1_unnormalized_0.svg – the heatmap for dataset PatientCharacter1, generated using the raw dataset. The 0 at the end indicates this is the first time the heatmap has been generated for the raw data.
- figures_PatientCharacter2_wgcna_1.svg – the WGCNA triangular heatmaps and dendrogram for dataset PatientCharacter2. The 1 at the end indicates this is the second time the WGCNA figures have been generated for this dataset.
- heatmap_PatientCharacter2_wgcna_1.svg – the corresponding color-coded heatmap for the WGCNA figures for dataset PatientCharacter2. The 1 at the end indicates this is the second time the WGCNA heatmap has been generated for this dataset.

All downloaded context map progress images will also be saved to the same *download*_imgs folder under the *static* folder. The context map images are named after the label given by the user when saving progress, followed by a semicolon, followed by "progress_#" where the final # is the index, starting at 0, of the progress saved.

Sample context map image filenames:

- part 1 liver;progress_0.png – the image corresponding to the context map for progress labeled "part 1 liver". The 0 at the end indicates this is the first time progress has been saved.
- part 2 tumor cell;progress_8.png – the image corresponding to the context map for progress labeled "part 2 tumor cell". The 0 at the end indicates this is the second time progress has been saved.

All downloaded exemplar images using the Exemplar Image Search page of the context map interface will be saved to the *static/exemplar_references/Dataset/pathway/* folder, where *Dataset* is the name of the dataset being analyzed. For example, for dataset *PatientCharacter1* exemplar images will be downloaded to *static/exemplar_references/PatientCharacter1/pathway/*. The names of the images are not determined by the visAPPprot system, but rather by the names of their source images from the Internet.

4. Toy Datasets

There are 2 toy datasets that are provided with visAPPprot: *PatientCharacter1.csv* and *PatientCharacter2.csv*. Both files can be found under the *processed_data/* folder. The same folder also contains the corresponding expression matrices associated with each dataset: *ExpMat1.csv* and *ExpMat2.csv*, for *PatientCharacter1.csv* and *PatientCharacter2.csv*, respectively. These files are sufficient to generate every visualization on the visAPPprot system.

Note that no sample microarray data files are provided with visAPPprot; in order to use the "Compute ExpMat" functionality you must provide your own data files.

You are now ready to move onto Set Up Image Search!