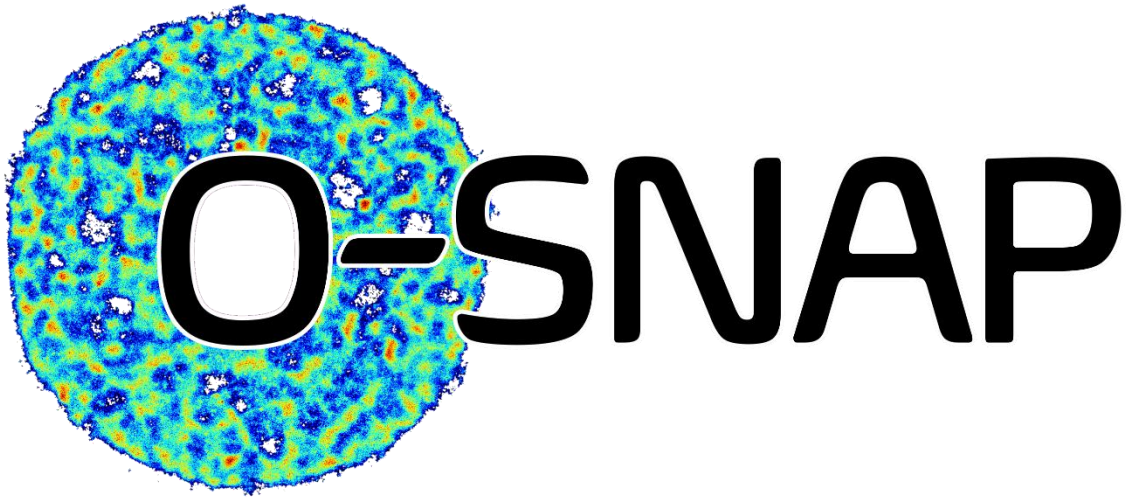


Objective **S**ingle molecule **N**ucleus **A**rchitecture **P**rofiler Tutorial

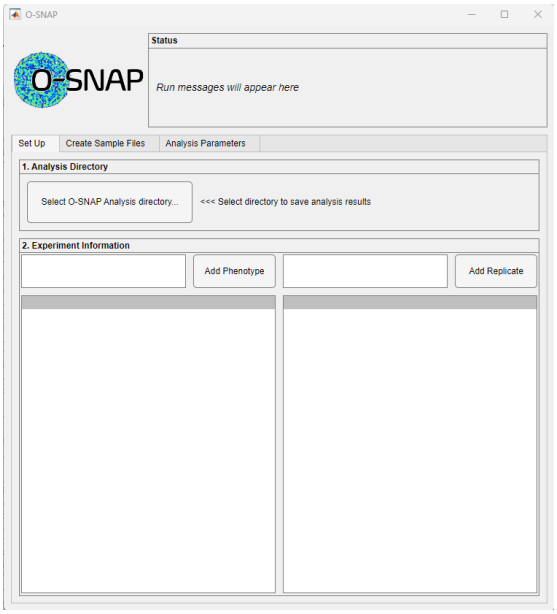
v.1.1

Author:

Hannah H. Kim
hannah.kim3@pennmedicine.upenn.edu

Lakadamyali Lab
Department of Physiology
University of Pennsylvania

Installation



O-SNAP has been tested for compatibility on Windows and Mac. The code was developed on an Intel® Xeon® Silver 4214R CPU, 2.40GHz (48 GB RAM) desktop tower with Windows 64-bit.

MATLAB

O-SNAP was originally implemented in MATLAB R2024b but should be compatible with all versions R2023 and above.

The application requires the Bioinformatics, Deep Learning, Parallel Computing, Signal Processing, and Statistics and Machine Learning Toolboxes. For some speedups, a Mex setup is required.

O-SNAP also relies on parallel processing to improve calculation times. Please check that your system can run on multiple workers.

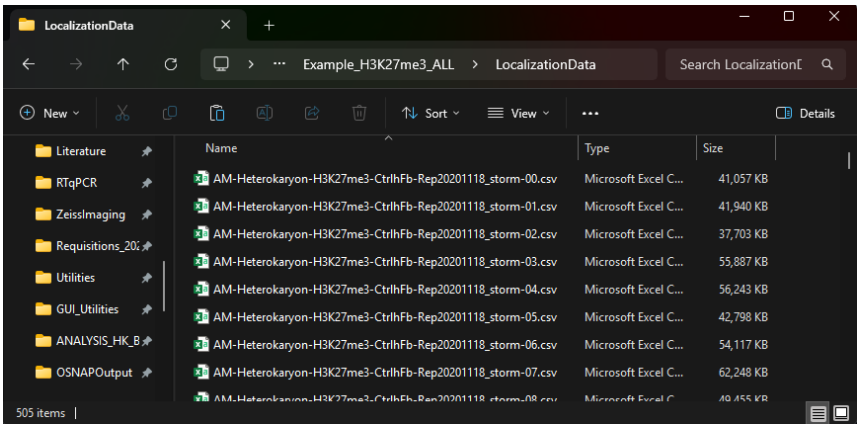
Once O-SNAP is installed, ensure that you add the O-SNAP folder and all subfolders to the MATLAB path. Under Home > Set Path, navigate to the O-SNAP directory and select **Add directory and subdirectories**. Save the changes.

R & dynverse

The pseudotimeline is performed using R (https://www.r-project.org/) and the *dynverse* package.

The *dynverse* package (https://github.com/dynverse/dyno) used to perform pseudotime analysis is described on its website (https://dynverse.org/). This also requires the installation of either the Docker or Singularity container (Docker is recommended). Ensure that Docker is open and running prior to running training trajectory inference methods with *dynverse*.

Input data



O-SNAP accepts 2-D arrays of localization data stored in the form of CSV or TXT files. Each file should contain single-channel localization data cropped to the boundary of exactly one nucleus, also referred to here as samples. The files do not necessarily need to reside in the same directory.

Within each datafile, the rows represent each localization, and the columns are the x and y positions in pixels. The interface uses the file names to automatically recognize the phenotypes of every sample. For proper phenotype assignment, every file name should feature exactly one phenotype substring. The interface also stores information on which replicate a given sample is derived from. For a new run, this information should be provided manually by the user.

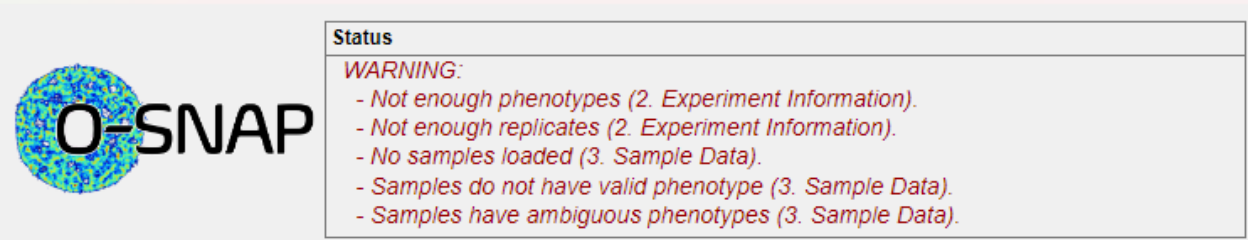
A full example input dataset is available at DOI: 10.6084/m9.figshare.29533940 under `O-SNAP_Example_Data\`. A smaller example data set is available at the O-SNAP Git repository.

Getting Started

O-SNAP provides an application to accompany the analysis code. Run `OSNAP\GUI_Utilities\OSNAP.mlapp` to open the application.

O-SNAP also includes the MATLAB scripts that perform the actual analysis. An example script of how to run these functions in the Command Window or via a script can be seen in the file: `OSNAP\Example\Example_Driver.m`.

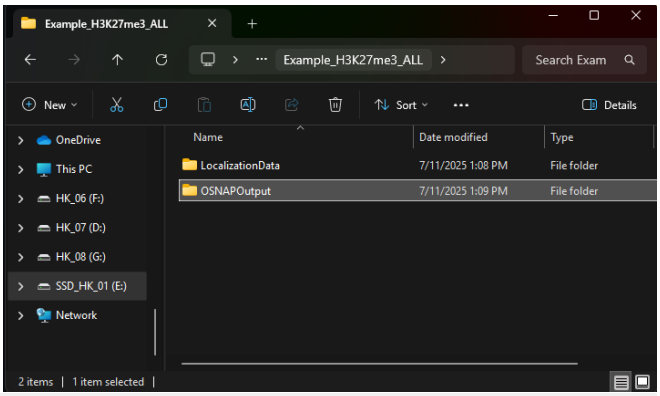
O-SNAP application overview

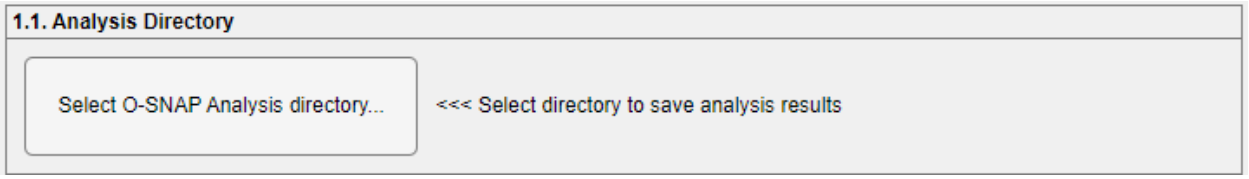


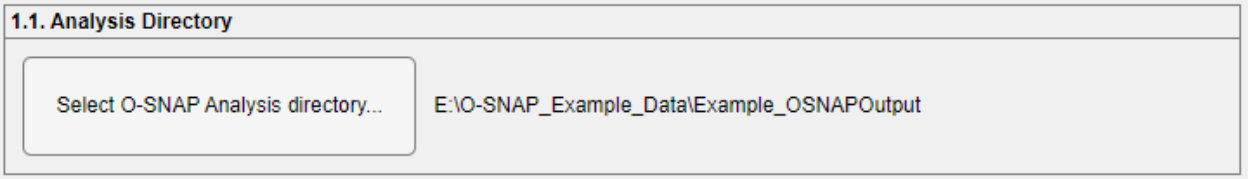
1. Set Up

1.1. Analysis directory

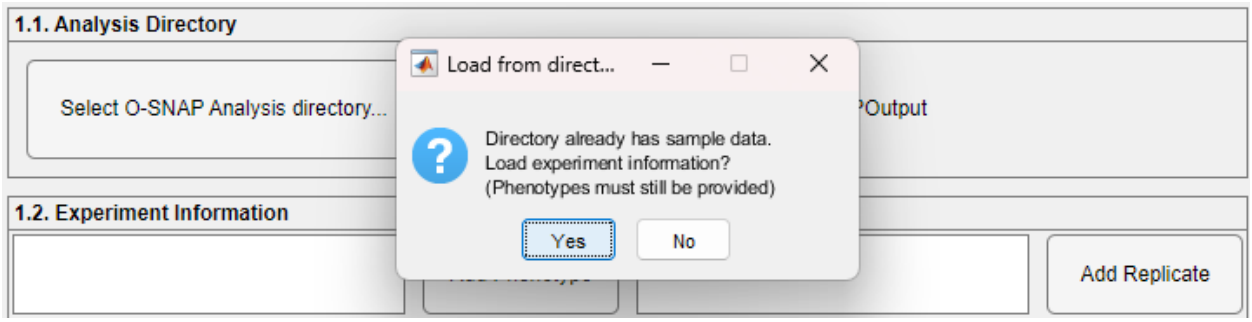
The analysis directory will contain the output to files generated by O-SNAP. It should be separate from the directory containing the localization information as new runs will remove all analysis directory contents.





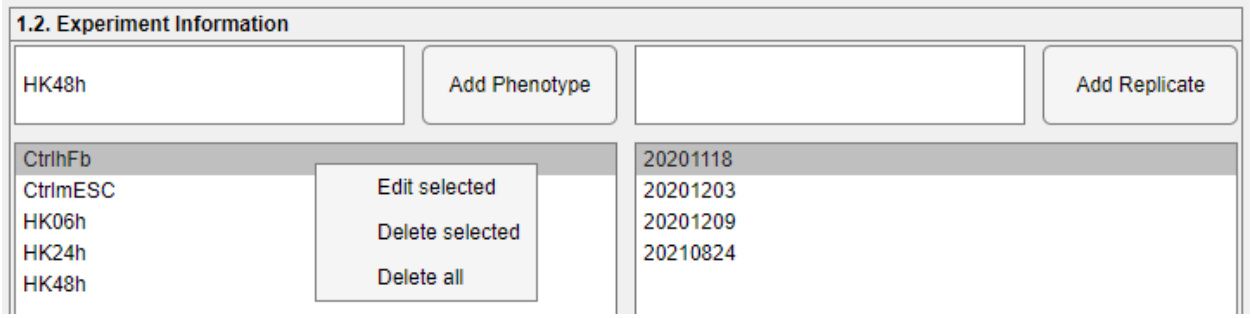


If sample files are already loaded (Sample Data Panel 2.1) and an analysis file exists in the selected analysis folder (e.g. `OSNAP\Example\Example_H3K27me3_ALL\OSNAP_Output`), the interface can retrieve the experiment information and generated sample files (see Ex.2 below).



1.2. Experiment Information

Here, the user provides input on the phenotype labels and the replicates to accompany the dataset. For the first run, the user must supply this information. The information organizes the MAT files containing sample data that are produced and can act as filters at the final run stage.

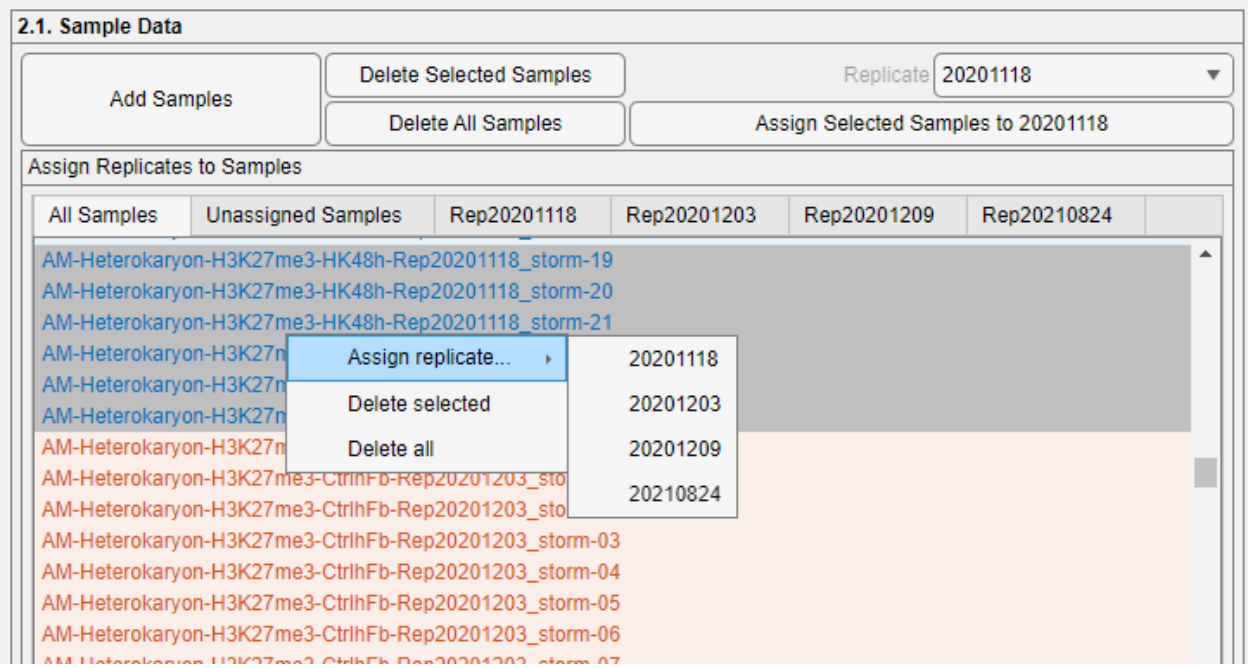


To add a phenotype or replicate, the user can type the appropriate label and then click **Add Phenotype/Replicate**. There must be at least two phenotype labels and at least one replicate label. Once the labels are added to the respective list, the user can also right-click to open a menu and select one of three options: (1) **Edit Selected**, which allows the user to change the selected label, (2), **Delete Selected**, which removes the selected label, and (3) **Delete all**, which clears all labels. When phenotypes or replicates are already assigned to samples, a prompt appears that asks the user if they'd like to delete the associated samples as well.

2. Sample Files

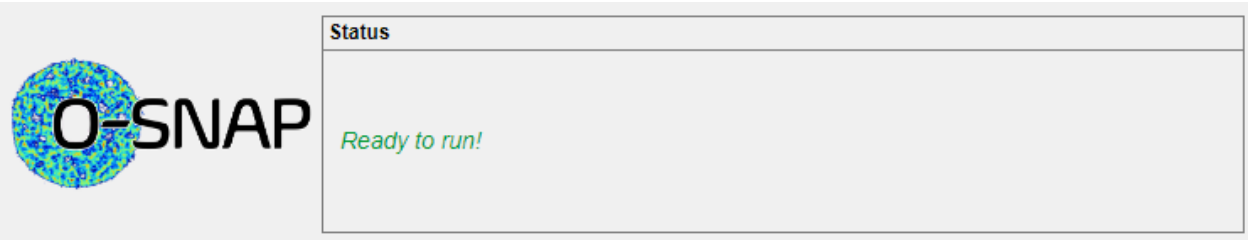
2.1. Assign Sample Data

After the experiment information is input, the Sample Data panel is where the user can add and modify replicate information for the sample files before the sample analysis MAT files are created.



Add the localization files of the samples to the interface by clicking **Add Samples** and selecting the appropriate CSV/TXT files corresponding to the samples of interest. The sample names will populate the **All Samples** and **Unassigned Samples** tab under the **Assign Replicates to Samples** panel. There are tabs for each of the user-defined replicates, which will automatically update once the respective replicate label is assigned to samples.

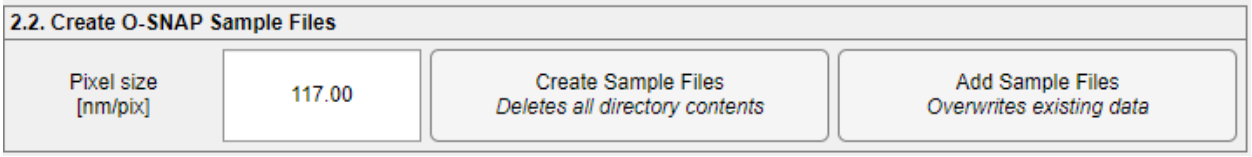
If conducting a new O-SNAP run, the user must then assign the appropriate replicate labels, which are used to create the directory structure that the sample analysis MAT files are stored in. Assign replicates by selecting samples of interest and clicking **Assign Selected Samples to (Replicate Label).** Multiple samples can be selected with shift + click or control + click. The user can also right-click and select the replicate from a dropdown menu. The samples become color coded in the **All Samples** tab based on replicate. They also move from the **Unassigned Samples** tab to the appropriate replicate tab. All samples must be assigned to exactly one replicate to create sample files and/or begin an O-SNAP analysis. Once all the samples are properly assigned, the Status panel will read, “Ready to run!”



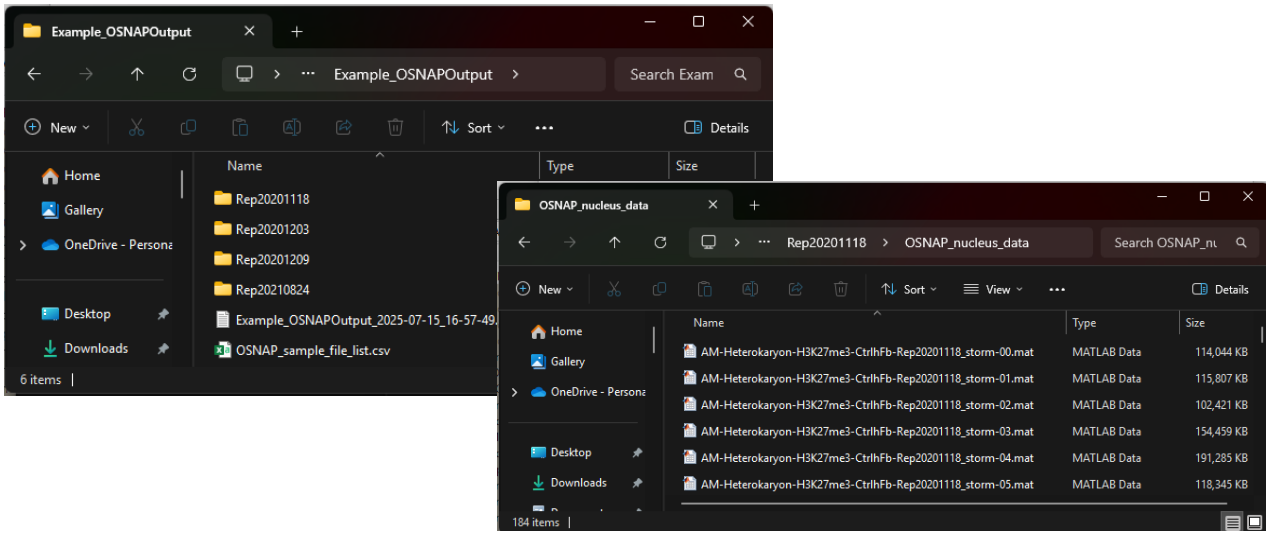
To delete samples, click the **Delete Selected Samples** button or right-click and select **Delete selected**. All the files present in the selected tab can be removed by clicking the button below **Delete Selected Samples** or right-clicking and selecting **Delete all**.

2. Create O-SNAP Sample Analysis Files

The user must define the distance unit from pixels to nm, the unit with which O-SNAP operates.



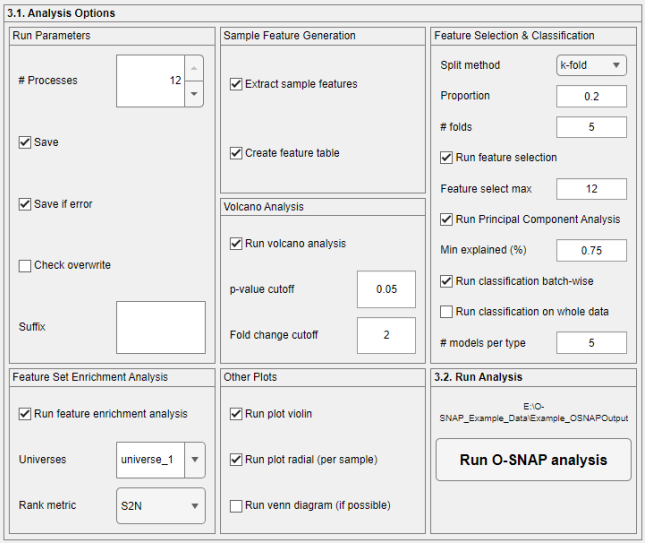
There are two approaches available to create the sample MAT files: (1) **Create Sample Files** and (2) **Add Sample Files**. The critical distinction is that **Create Sample Files** DELETES ALL DIRECTORY CONTENTS from the selected analysis file. This is to ensure a clean run with no extraneous output from a previous run that may interfere or obscure the results. The second option, **Add Sample Files**, instead will preserve all existing files, but will overwrite any samples analysis files with the same data. Once either of these two options are selected, the software will create a MAT file for every sample and save it to the analysis directory, which will then feature one directory for every replicate label. The corresponding sample MAT files are stored within the replicate directory, under the `OSNAP_nucleus_data\` subfolder.



3. Analysis

3.1. Analysis Options

The user can modify settings and run the analysis in this panel. The analysis options are grouped into six categories.



Run Parameters

These settings control how the run is performed.

- # Processes – The number of computational threads for the analysis to use when running processes in parallel.
- Save – The option on whether to save results to the directory. Recommended to keep on.
- Check if overwrite – Prompts user whenever a file may be overwritten, including the sample analysis MAT files. Recommended to keep off for a run with a large dataset to reduce intervals of wait time for user confirmation.
- Suffix – An optional suffix that will be appended to output files. Rather than saving multiple analysis outputs to the same analysis directory, it is recommended to save each distinct run in a separate analysis directory and use the suffix option to distinguish the output files.

Sample Feature Generation

These options indicate whether features must be extracted from samples. For a run on a new dataset where sample files have just been created, these two options must be performed.

- Extract sample features – For every sample analysis MAT file, use the localization data stored to calculate properties that inform the O-SNAP features. The property information is stored in the sample analysis MAT file for faster retrieval once an initial run is performed.

  WARNING: This step can be time-intensive due to the calculation of clusters for every sample. Please be patient.
- Create feature table – Once every sample MAT file has been updated with the additional properties, O-SNAP can then extract the 144 features for each sample. The feature data table is formed where the rows represent the samples (i.e. nuclei) and the columns represent the features.

  Once the feature table is generated, it can be used to perform modified analyses (e.g. filter for fewer phenotypes or run analysis with alternative parameters) without the need to run the **Extract sample features** *step, which is very time intensive. The recommended procedure is to copy the analysis MAT output file into the new analysis directory. The offshoot directories do not need to have the sample MAT files copied over, since O-SNAP will retrieve feature data directly from the analysis MAT output file rather than calculating the features from the sample files.*

Volcano Analysis

These options relate to the pair-wise comparison of feature value distributions between phenotype labels.

- Run volcano analysis – When selected, performs the volcano analysis
- p-value cutoff – The value for α, the upper threshold for the adjusted p-values
- Fold change cutoff – The lower threshold for the absolute fold-change difference between two phenotypes

Feature Selection & Classification

These options modify the feature selection, principal component analysis, and classification steps.

- Split method – Defined the method to create the folds
  - k-fold – Splits data evenly into k groups. Test and training sets are created by reserving one of the groups for testing and merging all other groups for training

data. This is repeated to create a set of k training/test sets with a different subset of samples used in the sample data for each fold.
    - replicate – If enough replicates are provided, use the replicate labels to create groups of test/training sets. For N replicate labels, N test/training sets are formed such that each test set is a different group of replicates.
    - Bootstrap – Create k folds of training/test sets, where the test set is any random selection of samples. The size of the test set is defined with the "proportion" option, so that the number of samples in the test set divided by the total number of samples is equal to this value.
    - None – Returns the full dataset as the training data with no samples in the test set.

Feature Set Enrichment Analysis (FSEA)

These options modify the FSEA.

- Universes – The name of the feature universe that defines how features are categorized into feature sets, along with information on whether each feature trends with or against the trend defined by the set.

    For a universe name input, there must be a function titled "get_<universe name>" that returns a struct array, where each field is the feature set that describes an overall trend and contains an [Nx2] cell array containing N features, where the first column is the name of the O-SNAP feature and the second column is the value 1 signifying a trend consistent with the feature set description and -1 for a trend opposing it.
- Rank metric – The metric used to rank features for FSEA. Please refer to **Supplementary Table 4** of the paper for more information.

Other plots

These options determine whether additional plot figures are generated with the analysis

- Run violin plot – Produces violin plots displaying the distribution of values by phenotype.

    These plots do not include any statistics.
- Run plot radial (per sample) – Creates two figures for each sample that display the radial analysis for the localization data and the chromatin packing domains.
- Run Venn diagram – Creates a Venn diagram to compare each set of pair-wise comparisons from the volcano analysis to demonstrate the number of features in common between the multiple sets of comparisons.

    Only conducted if there are 3-4 phenotypes provided.

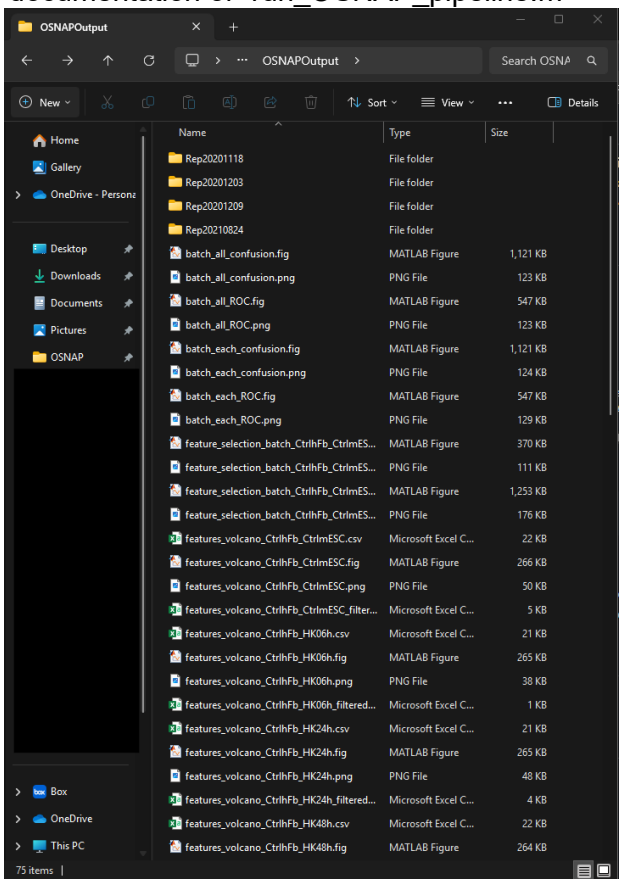
3.2. Run Analysis

When the user is ready and either (1) all samples are appropriately loaded and labeled or (2) an analysis directory with an existing analysis file is selected, the user can then run the analysis by hitting **Run Analysis button**.

Output Information

O-SNAP creates several different types of output sent to the analysis directory (*<analysis_name>*\…). Below is a description of the types of output a user can expect depending on the run they perform.

- Log file (*<analysis_name>_<YYYY-MM-DD_HH-MM-SS>.txt*)
  - A report of output from the O-SNAP run.
- Analysis MAT file (*<analysis_name>.mat*)
  - A file containing information on the last O-SNAP analysis run. Includes the following fields. They are further described in the documentation of "run_OSNAP_pipeline.m"
    - *analysis_name*
    - *classification_summary_all*
    - *classification_summary_batch_all*
    - *classification_summary_batch_each*
    - *classifiers_all*
    - *classifiers_batch_all*
    - *classifiers_batch_each*
    - *date*
    - *feature_comparisons*
    - *feature_data*
    - *groups*
    - *options*
    - *pca_result_all*
    - *pca_result_batch_all*
    - *pca_result_batch_each*
    - *replicates*
    - *starttime*
    - *test_idxs*
    - *train_idxs*
    - *vars_select_result_all*
    - *vars_select_result_batch*
    - *vars_selected_all*
    - *vars_selected_batch*
    - *venn_data*
  - *Copy and rename this table to a new directory to perform modified analysis using same feature data*



- Sample directory table (*nucleus_OSNAP_file_list.csv*)
  - A table that details information on the samples (name, replicate label, phenotype label, and full file path location)
- Sample data directories (*<Replicate Label>*)
  - Analysis MAT files (*OSNAP_nucleus_data**<sample>.mat*)
    - Files that each store data for a single sample (nucleus).
  - Voronoi density maps (*OSNAP_nucleus_images**…)*
    - Renderings of Voronoi cells colored by Voronoi cell density (*<sample>.png*).
    - Renderings of Voronoi cells colored by reduced Voronoi cell density (*<sample>_reduced.png*).

- Nucleus analysis images (*OSNAP_nucleus_images*\(*<sample>_nucleus_analysis.png*)
  - Images of the nucleus showing the 15% periphery (blue) and 85% interior regions (brown), respectively. The packing domains (black) and modeled Lamin-associated domains are also depicted (green).
- Packing domain images (*OSNAP_dbscan_cluster_images*\(*<sample>_dbscan_cluster.png*)
  - Images of the packing domains color-coded based on the value of different properties of the domains.
- Packing domain centroids (*OSNAP_dbscan_cluster_center_images*\ *<sample>_dbscan_cluster_center.png*)
  - Images of the packing domain centroids.
- Voronoi cluster images (*OSNAP_voronoi_cluster_images*\(*<sample>_voronoi_cluster.png*)
  - Images of the Voronoi clusters color-coded based on the value of different properties of the domains.
- Radial density images (*OSNAP_radial_density_images*\…)
  - Images of the radial density analysis for localizations (*<sample>_radial_loc_density.png*)
  - Images of the radial density analysis for packing domains (*<sample>_dbscan_cluster_density.png*).

➢ Feature data information (*<analysis_name>.csv*)
- The feature data, where rows represent samples and columns include the sample name, phenotype label, replicate label, and values for all 144 O-SNAP features
- *Use this table as input for pseudotime analysis*

➢ Violin plots (*violin_plots**violin_plot_<feature>.png*)
- Violin plots of the distribution between phenotypes for every O-SNAP features

➢ FSEA directories (*FSEA_plots_<universe>_<phenotype A>_<phenotype B>*\…)
- Plots of scores vs feature ranking. (…*FS_plots_<feature set>.png*)
- Summary FSEA plot. (…*FSEA_plots_<universe>.fig/png*)

➢ Volcano analysis results
- Feature comparison between phenotype A and B, for all features. (*features_volcano_<phenotypeA>_<phenotypeB>.csv*)
- Feature comparison between phenotype A and B, for features that meet the p-value and fold change cutoffs. (*features_volcano_<phenotypeA>_<phenotypeB>_filtered.csv*)
- Volcano plot highlighting significantly changing features. (*features_volcano_<phenotypeA>_<phenotypeB>.fig/png*)

➢ Feature selection results
- Top MRMR scores for feature selection
  - Performed once on whole dataset with no splits. (*feature_selection_**all**_<phenotypeA>_<phenotypeB>_...<phenotypeB>.fig/png*)

    - Performed batch-wise and then aggregated across folds using summation of individual scores. (*feature_selection_**batch**_<phenotypeA>_<phenotypeB>_...<phenotypeB>.fig/png*)
    - Performed across batches in a totally independent manner. (*feature_selection_**batch**_<phenotypeA>_<phenotypeB>_...<phenotypeB>_**batch**.fig/png*)
- PCA results
  - Plots of training data in PCA space
    - Performed with feature selection on whole data with no splits. (*PCA_<phenotypeA>_<phenotypeB>_...<phenotypeB>.fig/png*)
    - Performed using the aggregated feature selection across multiple folds. (*PCA_<phenotypeA>_<phenotypeB>_...<phenotypeB>_**all**.fig/png*)
    - Performed with feature selection on individual folds in a totally independent manner. Due to parallel framework, the indexing is not based on fold index, but a combination of the worker and sub-index values. (*PCA_<phenotypeA>_<phenotypeB>_...<phenotypeB>_**<worker #>_<sub-index #>**.fig/png*)
- Classification results
  - Confusion matrices
    - Performed with feature selection/PCA on whole data with no splits. (*all_confusion.fig/png*)
    - Performed with feature selection/PCA using the aggregated feature selection across multiple folds. (*batch_all_confusion.fig/png*)
    - Performed with feature selection/PCA on individual folds in a totally independent manner. (*batch_each_confusion.fig/png*)
  - ROC curves
    - Performed with feature selection/PCA on whole data with no splits. (*all_ROC.fig/png*)
    - Performed with feature selection/PCA using the aggregated feature selection across multiple folds. (*batch_all_ROC.fig/png*)
    - Performed with feature selection/PCA on individual folds in a totally independent manner. (*batch_each_ROC.fig/png*)

Pseudotime Analysis

The pseudotime analysis is accomplished by using the exported CSV file of the O-SNAP feature table (*<analysis_name>.csv*). The pseudotime analysis uses the the *dynverse* *R* package to perform a suite of pseudotime trajectory inference (TI) on this feature data. It is recommended that this analysis is only performed for data sets with large sample counts for each phenotype and where the topology of the cell state trajectory is simple or well-defined.

We have provided an example analysis on fibroblast reprogramming H2B and H3K27me3 data originally produced in Martinez-Sarmiento et al, Cell Reports (2024) under the `Pseudotime_Example_Heterokaryon\` directory. The O-SNAP feature data is provided for each mark under the `Input_O-SNAP_CSV\` directory, which is the same file copied from the O-SNAP

analysis directory from the respective dataset. The `Pseudotime_Exmaple_Heterokaryon.R` script performs trajectory analysis on each of these files and places the output into either the `ANALYSIS_Heterokarryon_H2B\` and `ANALYSIS_Heterokaryon_H3K27me3\` directories.

Briefly, the script imports the data stored in the CSV and filters for the phenotypes desired (in this case, 'CtrlmESC,' the control mESC phenotype, is excluded). The script then opens a Shiny app to prompt the user for information on the expected topology to offer candidate TI methods to apply. After the user provides this and closes the app, the script goes on to loop through the two markers and perform the TI analysis. The output plots, models, and pseudotime values are sent to a directory for each of the TI methods:

`Heterokaryon_Example_Data\ANALYSIS_Heterokaryon_...\<TI method name>\`

Please review the script and make modifications as needed. Additional information on the *dynverse* package is available at https://dynverse.org/.

Step-by-Step Examples

Example data is available at DOI 10.6084/m9.figshare.29533940. Below are some examples of how the pipeline was applied to:

`O-SNAP_Example_Data\Example_H3K27me3_ALL\`

A smaller, quicker example to run end-to-end is included:

`O-SNAP_Example_Data\Example_H3K27me3_CtrlhFb_HK48h\`

Ex. 1: Running a new analysis from scratch, end-to-end

This example guides a user through performing an initial O-SNAP run starting with a set of CSV/TXT files containing 2-D arrays of localization data for each nucleus. It is assumed this data is already cropped to localizations within the nucleus.

1. Under **1.1. Analysis Directory**, select the directory with CSV/TXT files of localization data

   `Analysis Directory:` `Example_H3K27me3_ALL\OSNAP_Output\`

2. Under **1.2. Experiment Information**, add phenotypes labels. A label must appear exactly once as a substring in the corresponding sample CSV/TXT files.
   a. Type in the label and click **Add phenotype**
   b. Edit or delete labels with right-click menu.

   `Phenotypes:` `CtrlhFb, CtrlmESC, HK06h, HK24h, HK48h`

3. Under **1.2. Experiment Information**, add replicate labels.
   a. Type in the label and click **Add replicate**
   b. Edit or delete labels with right-click menu.

   `Replicates:` `20201118, 20201203, 20201209, 20210824`

4. Under **2.1. Assign Sample Data**, click **Add sample**. Select all desired localization data.

   `Localization files:` `Example_H3K27me3_ALL\LocalizationData\*.csv`

5. Sample names should appear in the **Assign Replicates to Samples** panel. Select multiple samples.
6. Assign a replicate to the samples either by choosing a replicate in the drop-down and clicking the **Add replicate** or right-clicking and choosing **Assign replicate** > ***…***.

    a. The **Unassigned Samples** tab can be useful to check if any samples remain unassigned.
7. Delete samples by selecting and clicking **Delete Selected Samples** or right-clicking and choosing **Delete Selected**. Delete all samples contained in the tab by clicking **Delete All Samples** or right-clicking and choosing **Delete All**.
8. If the samples are loaded and assigned properly, the **Status** panel should read: “Ready to run!”
9. Under **2.2. Create O-SNAP Sample Analysis Files**, specify the pixel size in nm/pix.

```
Pixel size [nm/pix]: 117
```

10. Click **Create Sample Files** to create sample analysis MAT files. **WARNING:** This will delete ALL existing contents in the directory.
11. Under **3.1. Analysis Options**, select all the desired options. For a first run, the following settings are strongly recommended:
    a. Under *Run Parameters*

```
# Processes: The maximum value possible
Save: Checked true
Save if error: Checked true
```

    b. Under *Sample Feature Generation*

```
Extract sample features: Checked true
Create feature table: Checked true
```

12. Under **Run Analysis**, confirm the analysis directory.
13. Click **Run O-SNAP analysis**
14. The GUI will then execute the O-SNAP analysis in the MATLAB base workspace. Output will appear in the **Command Window**.
    a. Feature extraction can take some time due to the clustering algorithms.
15. Once the run is completed, you should see the appropriate output appear in the analysis directory.

Ex. 2: Modifying a previous analysis

This example is for a user who already has run the O-SNAP pipeline at least once on all the samples and wishes to perform a modified analysis on the same feature data without the need to extract features from the samples from scratch.

1. Create a new analysis directory separate from the original, where the directory is titled with the new analysis name of choice (*<new_analysis_name>*\).

```
Old Analysis Directory: Example_H3K27me3_ALL\OSNAP_Output\
New Analysis Directory: Example_H3K27me3_ALL\OSNAP_Output_PreFusion\
```

2. Copy and paste the original analysis MAT file (*<old_analysis_name>*.mat) to the new directory.

```
Old Analysis MAT file: Example_H3K27me3_ALL\OSNAP_Output\OSNAP_Output.mat
New Analysis MAT file: Example_H3K27me3_ALL\OSNAP_Output_PreFusion\
                       OSNAP_Output.mat
```

3. Rename the file name so that it now matches the new directory name (*<new_analysis_name>**<new_analysis_name>*.mat).

Renamed Analysis MAT file: Example_H3K27me3_ALL\OSNAP_Output_PreFusion\OSNAP_Output_CtrlhFb_HK48h.mat

4. In the O-SNAP application under **1.1. Analysis Directory**, select the new directory.
   O-SNAP will load the original analysis information. There is no need to load and assign sample localization data, unless the user wants to add additional samples to the analysis that were not previously included (see Step 7).

   New Analysis Directory: Example_H3K27me3_ALL\OSNAP_Output_PreFusion\

5. Under **1.2. Experiment Information**, modify phenotype and replicate labels as desired. Phenotype labels must still be present as a substring in the sample name exactly once.
   a. Phenotypes and replicates can be removed from analysis. The associated samples should also be removed under panel **2.1. Assign Sample Data.**
   b. When prompted "Some samples have selected phenotypes. Delete these samples as well?" select "Yes."

   Phenotypes: CtrlhFb, CtrlmESC

6. Ensure the **Status** panel reads: "Ready to run!"
7. Skip **2.2. Create O-SNAP Sample Analysis Files**, unless there are additional samples to add. However, it is recommended to keep all sample analysis files in the same folder.
   a. Copy the replicate folders containing the sample MAT files into the current directory. It's a good idea to preserve the original files as a backup.
   b. Once relevant information has been updated, select **Add Sample Files** (not "Create").
8. Under **3.1. Analysis Options**, select desired options.
9. Click **Run O-SNAP analysis**