KNIME Pipeline User Manual

Eva Freckmann

Contents

1 Before starting

The computer you will be running the KNIME analysis from should have either access to, or a local, copy of the directory containing your phase images, and output from CellProfiler. For reference, this directory should have the following structure:

1. DatasetName
    - Experiment___1
        - Data
        - Experiment Key
        - Phase
        - PhaseGrayCystOutlines
    - Experiment___n
    - Output

The "Experiment Key" and "Phase" folders should be populated by the user prior to analysis with CellProfiler, which will in turn populate "Data" and "PhaseGrayCystOutlines" upon analysis completion.

The default output directory for KNIME analysis results will be the "Output" subdirectory of this folder.

Experiment Key and directory structure templates can be found in the **davebryantlab/MethodsPaper2020** Github repository.

Tip for those new to KNIME: Double-click on a meta-node/node to configure it.

1.1 KNIME installation

KNIME can be downloaded from here.

1.1.1 Notes for KNIME workflow

Required KNIME version and extensions:

Name	Version
KNIME Analytics Platform	4.0.2.v201909300912
KNIME Interactive R Statistics Integration	4.0.1.v201908131226
KNIME Core	4.0.2.v201909300912
KNIME Quick Forms	4.0.2.v201909242005
KNIME Math Expression (JEP)	4.0.2.v201909242005
KNIME Python Integration	4.0.0.v201906241606
KNIME Image Processing	1.8.0.201911140609
KNIME SVG Support	4.0.1.v201908131226
KNIME Virtual Nodes	4.0.0.v201905311239
KNIME Distance Matrix	4.0.2.v201909260824
KNIME Data Generation	4.0.0.v201905311239
KNIME File Handling Nodes	4.0.1.v201908131226
Vernalis KNIME Nodes	1.24.2.v201911141223
KNIME Excel Support	4.0.1.v201908131226
KNIME HCS Tools	4.0.0.v201906200802

1.2 R Integration

The KNIME workflow uses an R integration for some steps of the analysis. The workflow uses R version **version**, which can be downloaded here. Instructions for setting up the R integration can be found here. **Windows users:** Do not install Rserve version 1.7-3.1 as is suggested in point 1 of the "R packages installation" section of these instructions. Instead, go straight to point 2 of the section, to install Rserve v1.8-6. More information on installing Rserve can be found here.

All users: In KNIME Analytics Platform go to File → Preferences. From the list on the left, select R under KNIME. Set the "Rserve receiving buffer size limit" to 0.

The KNIME workflow should automatically download and install any missing R packages that are required for the analysis. **Check that cytofkit does this** The following packages are utilised:

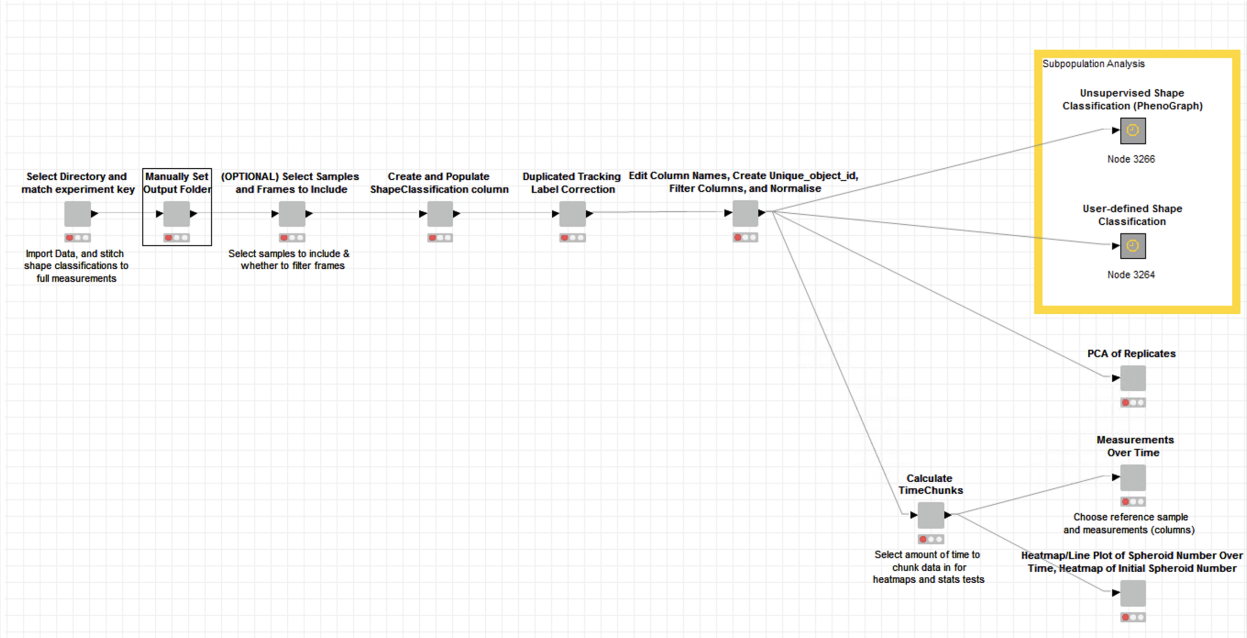
Package	Version
ggplot2	
reshape2	
ggnewscale	

1.3 Python Integration

The KNIME workflow requires Python in order to run the GeoSketch algorithm for subsampling data. Instructions for setting up the Python integration can be found here. First, follow the "Quickstart" and "Anaconda Setup" instructions, and download the Python environments provided on the **davebryantlab/MethodsPaper2020** Github repository. Load these environments into Anaconda. Then follow the "Setting up the KNIME Python Integration" instructions, and select the provided environments within your KNIME Python preferences.

Package	Version
GeoSketch	

2 Workflow structure



The KNIME workflow is comprised of four parts, each addressed in one section of this manual:

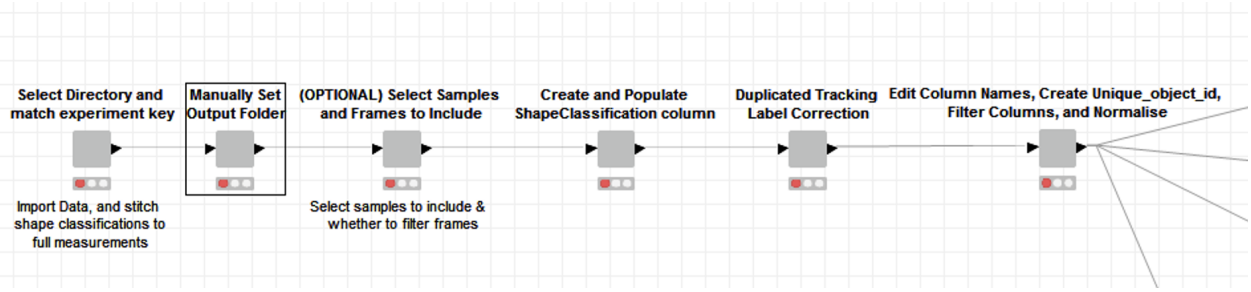
Data Import and Pre-processing

Data Summary Files

Unsupervised Shape Classification

User-Defined Shape Classification

3 Data import and pre-processing



3.1 Select Directory and match experiment key

This metanode reads in the CSV files output by CellProfiler and combines them with their respective experiment keys.

Configure the node to set the dataset root directory containing your images, Experiment Keys and CellProfiler outputs. **Get description from Dave for the rest of the user inputs**

3.2 Manually Set Output Folder

This metanode can override the default output directory.

If you wish to override the default output directory ("Output" subdirectory of your root directory), select "Yes", and choose an alternative directory for output. Otherwise, select "No".

3.3 Select Samples and Frames to Include

This metanode applies sample, frame, and lifetime filtering of objects in the dataset.

To filter out samples: Samples to be retained for analysis should be in the green "Include" box, and others in the red "Exclude" box.

To filter by frame number (timepoint): In order to include all frames, set "Filter by timepoints" to "No". Otherwise, set it to "Yes", and set the "Start Frame Number" and "Stop Frame Number" values to the lower- and upper-bound frames of interest, respectively.

To filter by lifetime: Set "Filter objects by lifetime" to "Yes", and set the maximum and minumum lifetime values.

Lifetime: the number of frames an object was tracked for by CellProfiler.

3.4 Create and Populate ShapeClassification column

If user-defined shape classifications are present, this metanode parses the classification columns from CellProfiler into a new "ShapeClassification" column. The new column is populated with the classification names as defined in CellProfiler.

Tick the box to indicate that user-defined shape classifications are present.

3.5 Duplicated Tracking Label Correction

In CellProfiler, when a track object splits into two, or more, new objects, the software assigns the tracking label of the parent to the daughter objects. This creates conflicts downstream in the KNIME workflow, whereby multiple objects have the same ID. This metanode ensures that the tracked objects in each Experiment, Plate, Well, and Site combination, each have a unique tracking label. Where duplicate tracking labels are present, the object that is largest at the initiation of the duplication retains this label, and the smaller object(s) is assigned a new tracking label.

No user configuration required.

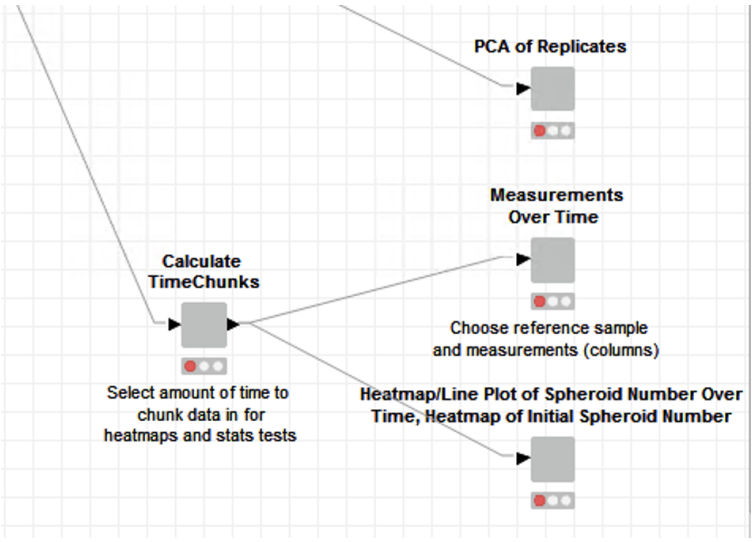
3.6 Edit Column Names, Create Unique-object_id, Filter Columns, and Normalise

This metanode removes unnecessary prefixes from column names. The AspectRatio is then calculated. Unique IDs are created for each tracked object. Finally, the filename of the originating image is extracted, and measurement columns are Z-Score normalised.

No user configuration required.

4 Data summary files

think of a better name for this section



4.1 PCA of Replicates

This metanode performs Principal Component Analysis (PCA) on the dataset replicates, and outputs results for two PCAs: one in which each point is a well, and another in which each point is an experiment. A plot of PCA Loadings is also output for each analysis. Four files are saved in the main Output directory: "PCAofReplicates_perWell.pdf", "PCAofReplicates_perWell_Loadings.pdf", "PCAofReplicates_perExperiment.pdf", "PCAofReplicates_perExperiment_Loadings.pdf".

No configuration required.

4.2 Calculate TimeChunks

Visualisation of the change in a variable over time can be complex to display when many timepoints are present. To simplify this, data can be presented in grouped segments on time, rather than at individual timepoints. This metanode calculates time interval chunks for the input data table.

Configure the metanode to indicate how many timepoints (frames) to include in each timechunk. If chunking is not required, set this value to 1. **check this works**

4.3 Measurements Over Time

This metanode generates a heatmap showing change in measurements of area, shape, and movement over time (for a set of user-specified measurements). For each sample, the average value of the measurement is calculated at each of the previously defined time chunk intervals. For the purposes of presentation, resulting values are Z-score normalised per measurement. A t-tests are performed to compare samples at each time interval, to the specified reference sample. A Bonferroni adjustment is applied to adjust for multiple testing. Two files are generated in the main Output directory:"MeasurementsOverTime_CONTROLSAMPLEControl.pdf", "MeasurementsOverTime_CONTROLSAMPLEControl.csv"

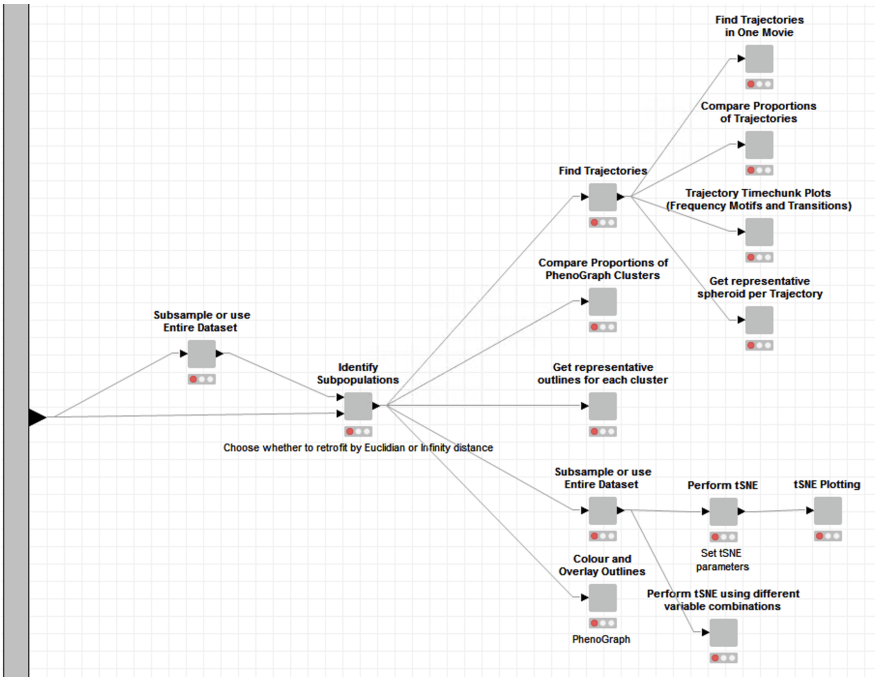
Configure the metanode to set a reference sample for statistical comparison. Measurements to be plotted in the heatmap should be included in the green "Include" box - all others should remain in the red "Exclude" box. Finally, indicate how rows (samples) in the heatmap should be ordered: alphabetically or in user-specified ordering. If a user-specified order is to be used, use the text box to define this order (from top row of the heatmap, to bottom). Sample names should be written as they appear in the Experiment Key, separated by only commas. The easiest way to ensure this is done correctly is by copy and pasting sample names directly from the Experiment Key. **Tip**: Be aware of any leading or trailing spaces in sample names in the Experiment Key, and if present, ensure not to delete them when configuring this text box.

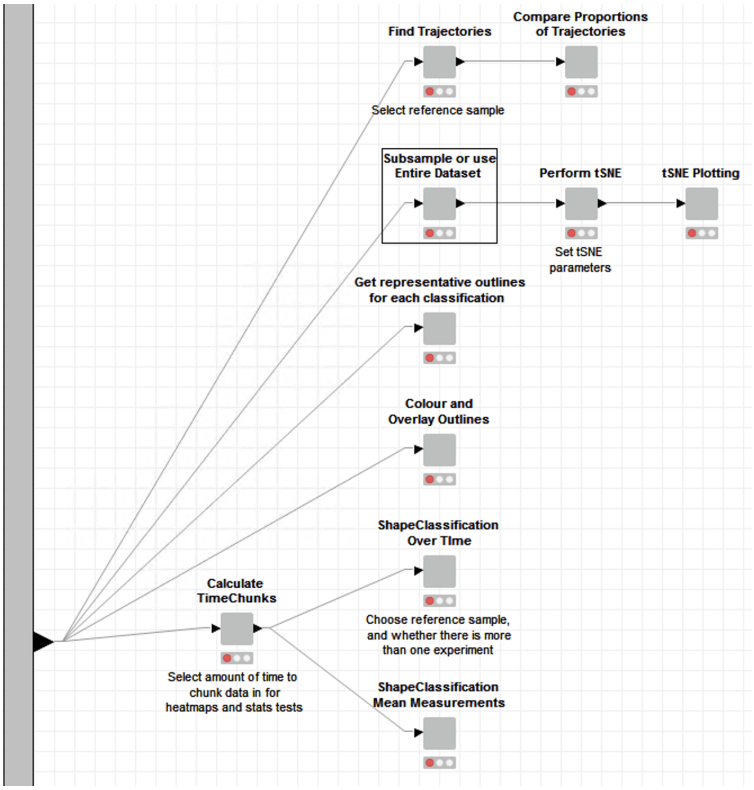
4.4 Heatmap/Line Plot of Spheroid Number Over Time, Heatmap of Initial Spheroid Number

This metanode outputs the intial counts of spheroids in each treatment group/sample per experiment (CSV file), and the relative change over time (line plot). Three files are generated in the main Output directory: "InitialSpheroidNumbersPerExperiment.csv", "SpheroidNumberOverTime.csv", "SpheroidNumberOverTime_LinePlot.pdf"

No configuration required.

5 Unsupervised Shape Classification





6.1 Find Trajectories

6.2 Compare Proportions of Trajectories

Add trajectory transition plots here?

6.3 Subsample or use Entire Dataset

6.4 Perform tSNE

6.5 tSNE Plotting

Add node for tsne using parameter combinations here?

6.6 Get representative outlines for each classification

6.7 Colour and Overlay Outlines

6.8 Calculate TimeChunks

6.9 ShapeClassification Over Time

6.10 ShapeClassification Mean Measurements