NASTIC WRANGLER USER MANUAL

Tristan Wallis t.wallis@uq.edu.au

Alex McCann a.mccann2@uq.edu.au

Single Molecule Neuroscience Laboratory (Fred Meunier)

Queensland Brain Institute

The University of Queensland, Australia

Introduction

The NASTIC family of programs output a large amount of metrics related to spatiotemporal clustering in super-resolution datasets. While these are useful and edifying by themselves, the real point of the exercise is to be able to establish whether these metrics change between experimental conditions (for example stimulated vs unstimulated cells). This can't be done within NASTIC itself, which only works on a single dataset at a time. Instead, the comparative stuff is handed off to NASTIC Wrangler, rocking a simple GUI which allows you to load the multiple NASTIC-generated metrics files from two experimental conditions and perform comparative analyses on the metrics.

Primitive versions of NASTIC Wrangler were written by Tristan Wallis as the original NASTIC suite matured. To coincide with the publication of NASTIC in Nature Communications, the comparative analysis routines were overhauled, and the NASTIC Wrangler GUI was largely rewritten by Alex McCann. As for the rest of NASTIC, this software is released under a Creative Commons licence.

Computer requirements

NASTIC is a Python script and requires Python 3.8 or later, and a number of python modules to run. If you've got as far as using NASTIC Wrangler then I can safely assume you downloaded the NASTIC suite from GitHub (which includes NASTIC Wrangler), and have Python and required modules installed for NASTIC. Wrangler requires a few more:

```
python -m pip install colorama matplotlib numpy pandas scipy Pillow pysimplegui seaborn scikit-learn
```

Data requirements

NASTIC Wrangler requires you to have your data in two separate directories, one for each experimental condition. For example, let's say you'd NASTIC analysed 3 files from unstimulated control and 3 from stimulated samples. Most likely this data was already in separate directories, but if it wasn't you could move the output directories created by NASTIC so that you had the following directory structure:

/control/sample1_NASTIC_20230607-123456/metrics.tsv
/control/sample2_NASTIC_20230607-134562/metrics.tsv
/control/sample3_NASTIC_20230607-143456/metrics.tsv

/stimulation/sample1_NASTIC_20230608-091234/metrics.tsv
/stimulation/sample2_NASTIC_20230608-114523/metrics.tsv
/stimulation/sample3_NASTIC_20230608-123412/metrics.tsv

NASTIC Wrangler will actually search recursively for metrics.tsv files, so they could be buried in additional directories:

/control/another_directory/sample1_NASTIC_20230607-123456/metrics.tsv
/control/sample2_NASTIC_20230607-134562/metrics.tsv
/control/another_directory/and_another_one/sample3_NASTIC_20230607-143456/metrics.tsv

Metrics files

Regardless of whether you've used NASTIC, segNASTIC or BOOSH, the metrics.tsv files contain largely the same information, and differ only in reporting the different parameters that each program used.

```
NASTIC: NANOSCALE SPATIO TEMPORAL INDEXING CLUSTERING - Tristan Wallis t.wallis@uq.edu.au
TRAJECTORY FILE:    W:/python/NASTIC/sx1a_pranesh.trxyt
ANALYSED:   20230424-131714
TRAJECTORY LENGTH CUTOFFS (steps): 8 - 100
TIME THRESHOLD (s): 20
CLUSTER THRESHOLD: 3
RADIUS FACTOR: 1.2
MSD FILTER THRESHOLD (um^2):   0.006054034090909118
CLUSTER MAX RADIUS (um):   5.0
SELECTION AREA (um^2): 100.0
SELECTED TRAJECTORIES: 1687
CLUSTERED TRAJECTORIES: 798
UNCLUSTERED TRAJECTORIES:   889
TOTAL CLUSTERS: 105
CLUSTERED TRAJECTORIES AVERAGE INSTANTANEOUS DIFFUSION COEFFICIENT (um^2/s):   0.019992491368949327
UNCLUSTERED TRAJECTORIES AVERAGE INSTANTANEOUS DIFFUSION COEFFICIENT (um^2/s): 0.43539409459456313
HOTSPOTS (CLUSTER SPATIAL OVERLAP AT 1/2 AVERAGE RADIUS):   8
TOTAL CLUSTERS IN HOTSPOTS: 18
AVERAGE CLUSTERS PER HOTSPOT:   2.25
PERCENTAGE OF CLUSTERS IN HOTSPOTS: 17.143

MSD CURVE DATA:
TIME (S):   0.02    0.04    0.06    0.08    0.1 0.12    0.14
UNCLUST MSD (um^2): 0.015310347720889616    0.02425511726873134 0.0330228494116826  0.041433993396563396    0.050173223621925625    0.05870554315565821 0.0673305267045077
UNCLUST SEM:    0.0002580805180023827   0.0006025807777302412   0.0009982640295434252   0.0014297471874805288   0.0019301385382886028   0.00246577308202073     0.0030006340745481674
CLUST MSD (um^2):   0.0061988299735618235   0.0067656312652331  0.00710147441988209     0.007398379455698783    0.00791090915408602 0.008165107283353498    0.008585764861626256
CLUST SEM:  7.66881421807763e-05    0.00011300954941195511  0.0001521750423453487   0.000201695918411175    0.0002668401861703863   0.0003407405227802364   0.0004255360308629564
ALL MSD (um^2): 0.011000335176510495    0.01598208239570728 0.02076128613755793     0.025334159416237397    0.03018192134253262 0.03479844900859296     0.039542548073435144
ALL SEM:    0.00017910681989877259  0.0003858604880084264   0.0006174137896523729   0.0008648334703252334   0.0011464709583294947   0.001446314734184144    0.0017466467888756936

INDIVIDUAL CLUSTER METRICS:
CLUSTER MEMBERSHIP  LIFETIME (s)    AVG MSD (um^2)  AREA (um^2) RADIUS (um) DENSITY (traj/um^2) RATE (traj/sec) AVG TIME (s)
1   3   7.549999999999983   0.00759363919191935     0.011454800000000015    0.06038357462305532 261.89894192827427  0.3973509933774436  140.10999999999999
2   12  27.939999999999998  0.0061956353503046549   0.0294019999999978  0.096741652216471   408.13550098633056  0.4294917680744453  279.56333333333
3   13  23.389999999999986  0.0046054570225892025   0.018315500000000078    0.07635446758637797 709.7813327509456   0.5557930739632325  259.8930769230769
etc
etc
105 3   0.5699999999999932  0.0060540340909099118   0.009092999999999801    0.05379955199691857 329.924117452993    5.263157894736905   268.32666666666665
AVG 7.6 16.317238095238093 0.006270059819242611    0.028683821428571423    0.0883959460627295  335.4849207818186   0.8051764827323206  178.94204989328392
SEM 0.5630496810323734  1.4760143888540158  0.00011400708600935275  0.0027689820084674985   0.0035409181670722698   16.42157468980873   0.09868004109745801 8.98171863053642
```
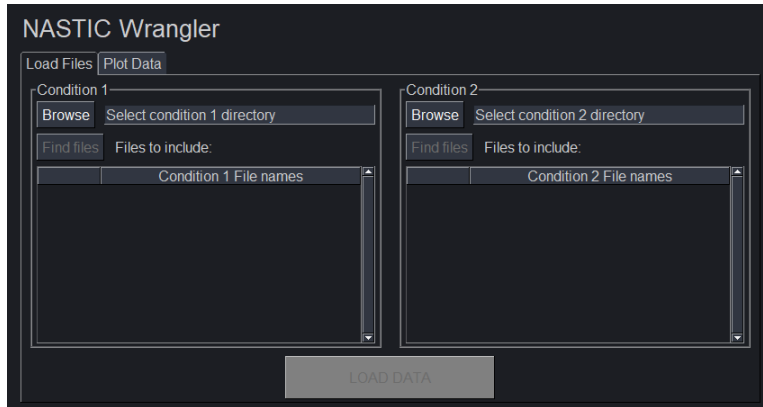
Please do not change the formatting of the files in any way.

Usage

If all is installed properly then double clicking the **nastic_Wrangler_gui.py** icon will launch the program.

Load Files tab:

Operation is very straightforward:

1. Select the top level directory for each of the conditions using "Browse"
2. Click the "Find files" button and NASTIC Wrangler will recursively search the directory for all metrics.tsv files and will populate a list
3. Select/deselect files using the tick box (use ~1s delay between clicks)
4. Click "LOAD DATA"
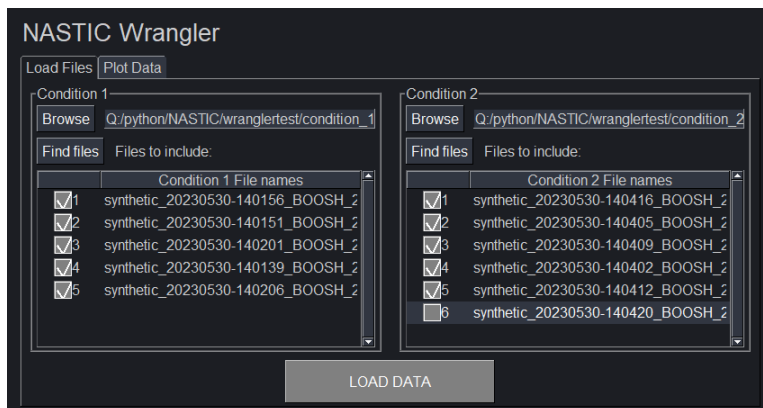
At this point NASTIC Wrangler will read all the selected metrics.tsv files, report stuff to the terminal, and will switch to the "Plot Data" tab.

Plot Data tab:

5. Give each condition a shorthand name that will show up in subsequent plots
6. Select a colour to use for plotting each of the conditions
7. Use the tick boxes to select what you want to plot
8. Click the "PLOT DATA" button. NASTIC Wrangler will save all plotted data to a datestamped directory in the same directory as NASTIC Wrangler itself: **nastic_wranger_output_YYYYMMDD-HHMMSS**.
9. Paste your plots into your manuscript and submit it to Nature.

Notes:

- If a new directory is selected using "Browse", the "LOAD DATA" and "PLOT DATA" buttons will become disabled, as the files in the newly selected directory will first have to be found using the "Find files" button before they can be read using "LOAD DATA" and subsequently analysed using "PLOT DATA".

- If after pressing the "LOAD DATA" button, the user changes the list of files to include for analysis (by unticking previously included files and/or ticking previously excluded files), the "PLOT DATA" button will become disabled until the new list of files to include in analysis is read by pressing the "LOAD DATA" button.

- Shorthand names to show on plots: must be different and cannot be blank as this causes issues with how the plots are displayed. If you really don't want a label to appear on the plots, you can set the shorthand name of one condition to a single space, and the shorthand name of the other condition to a double space, so that they appear blank on the plots but will not interfere with the generation of the plots.

Plot types

Aggregate data: These plots combine all the metrics from all of the clusters observed across all your files for a given condition. These plots are saved as PNG formatted to **nastic_wranger_output_YYYYMMDD-HHMMSS/aggregate_plots**

Metrics:

Area: average cluster area
Avmsd: average cluster MSD (at t=0)
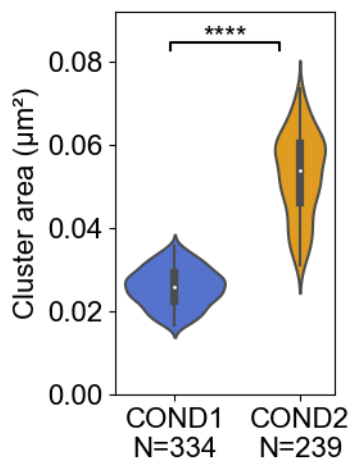Avtime: average cluster time centroid
Density: average density of trajectories within clusters
Lifetime: average cluster lifetime
Membership: average number of trajectories per cluster
Radius: average cluster radius
Rate: average rate of detections within clusters

Data is plotted as a violin plot (not individual data points) to allow assessment of data distribution. The white dot in the centre of the plot shows the median of the distribution, the thicker dark bar shows the interquartile range of the data, and the thinner dark bar shows the data that extends to 1.5 times the interquartile range. The significance, as determined by t-test, is plotted as asterisks where ns = no significance, * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001

You can optionally remove outliers from aggregate data prior to plotting, using median filtering. Median filtering works by calculating the mean for the data, establishing how far each value in the data deviates from the mean, and then dividing each deviation by the mean of all the deviations. Those values whose deviation/mean of deviation is > 2.5 are excluded.

Aggregate metrics will highlight trends in clustering behaviour between conditions, but the large N means that there is more chance of statistical significance.

Average data: These plots derive an average for each metric across the number of samples analysed. These plots are saved to **nastic_wranger_output_YYYYMMDD-HHMMSS/average_plots** and their filenames are self-explanatory. Please note that while there a large number of average metrics, not all of them are particularly informative and they are included for completeness. For example, Selection_area is down to user input and does not reflect any biology.

Metrics:

Average_cluster_area
Average_cluster_lifetime
Average_cluster_membership
Average_cluster_radius
Average_cluster_rate
Average_cluster_time_centroid
Average_clusters_per_hotspot
Average_density_of_clustered_trajectories
Average_MSD_of_clustered_trajectories
Cluster_density
Clustered_trajectories
Clustered_trajectories_Inst_diff_coeff
Hotspots
Percentage_clustered_trajectories
Percentage_clusters_in_hotspots
Selected_trajectories
Selection_area
Total_clusters
Total_clusters_in_hotspots
Unclustered_trajectories
Unclustered_trajectories_Inst_diff_coeff
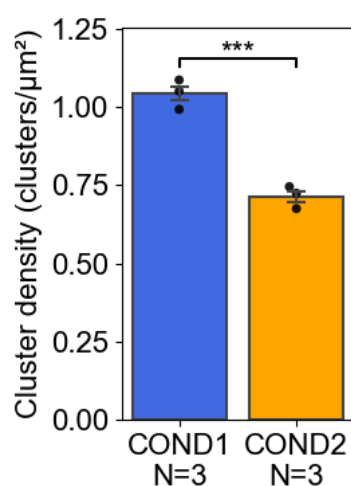
Data is presented as a bar plot of the average with datapoints plotted as a swarm plot, and significance displayed as above.

All of the aggregate and average plotted data is also saved to **nastic_wranger_output_YYYYMMDD-HHMMSS/processed_metrics.tsv** so that you may import it into other software for analysis and plotting.

PCA: These plots use Principal Component Analysis where each sample datapoint represents a number of normalised average metrics as described above.
nastic_wranger_output_YYYYMMDD-HHMMSS/pca.png
nastic_wranger_output_YYYYMMDD-HHMMSS/pca_labels.png

Metrics used:

Percentage Clustered Trajectories
Var Confined Trajectories*
Var Unconfined Trajectories*
Clustered Trajectories Average Instantaneous Diffusion Coefficient (Um^2/S)
Unclustered Trajectories Average Instantaneous Diffusion Coefficient (Um^2/S)
Average Clusters Per Hotspot
Percentage Of Clusters In Hotspots
Cluster Density (Clusters/Um^2)

* If VAR analysis was performed
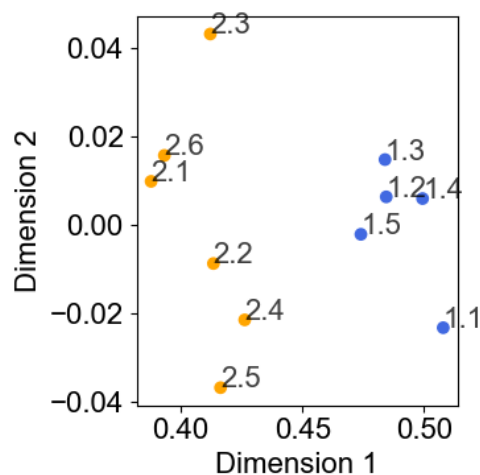
This plot allows the user to determine qualitatively if there are differences between the conditions, and whether there may be outlier samples which can be removed (using sample tick boxes as described above) for future analyses.

1.1 = condition 1, sample 1

2.3 = condition 2, sample 3

The plot is available with/without labels

Menu

Default settings:

Similarly to the other NASTIC programs, the first time you run NASTIC Wrangler it will create a file called **nastic_wrangler_gui.defaults** which contains the default values for the plots that are generated. You can load it into a text editor or spreadsheet to view it if you'd like.

Short name condition 1	COND1
Short name condition 2	COND2
Color condition 1	royalblue
Color condition 2	orange
Plot average data	TRUE
Plot aggregate data	TRUE
Outlier removal	TRUE
Plot 2D PCA data	TRUE

Values that can be changed in NASTIC Wrangler and saved to this file include: shorthand names and colors given to condition 1 and 2 (shown on plots), which plots will be generated (Average, Aggregate and 2D PCA), and whether Outlier Removal will be performed on Aggregate data. If you wish to keep these values (especially color) consistent across datasets, **"File >> Save settings"** will save them to nastic_wrangler_gui.defaults. The next time you start NASTIC Wrangler, they'll be loaded in as the defaults. If you wish to reset these values, **"File >> Default settings"** will restore the defaults shown in the above table. If you don't save them, the next time you load NASTIC you will have to reselect the values that you used previously. Similarly, if you started with your preferred values, and changed settings, **"File >> Load settings"** will restore your last saved settings. If settings for previously plotted data were not saved, the colors and shorthand names used, as well as whether outlier reduction was used, can be found in the corresponding processed_metrics.tsv output file, and then copied (without spaces) into the defaults file.

IMPORTANT NOTE: Periodically NASTIC Wrangler is updated with new functionality that requires a new **nastic_wrangler_gui.defaults** file. Newer versions of NASTIC Wrangler may crash if loading an older defaults file. You are encouraged to delete your existing defaults file and run the new version of NASTIC Wrangler to recreate the correct file with default settings.

Help:

Clicking on **"Info >> Help"** will show a popup window (left) which explains what each of the buttons do, as a useful reminder when using NASTIC Wrangler.

Help

This program allows the visualisation of the metrics.tsv files produced by NASTIC and SEGNASTIC. Comparison bar plots for each metric and statistical significance (t-test) are shown.

Load Files tab:

Browse: Select a directory for each condition.

Find files: Recursively search directories for metrics.tsv files.

Untick files to exclude from analysis (use ~1s delay between clicks).

LOAD DATA: Load selected files and extract information.

Plot Data tab:

Shorthand name: for each condition to appear on plots.

Select color: for each condition to appear on plots.

Outlier removal: uses median filtering to remove outliers from aggregate data.

Aggregate data: aggregates individual cluster metrics across all samples (N = total number of clusters).

Average data: averages cluster metrics for each sample (N = number of samples).

2D PCA analysis allows you to determine the overall relationships between samples.

PLOT DATA: plot (selected) aggregate, average and PCA. Also saves datestamped TSV of raw data used for plots.

Tristan Wallis, Alex McCann 20230608

OK