



# AtoMx® SIP Best Practices

Volume I

# Tips when performing CosMx<sup>®</sup> data analysis with AtoMx SIP

Single-cell spatial biology opens doors to unprecedented scientific questions requiring the development of new computational workflows and approaches to analyze and interpret data. Whether you are new to single-cell spatial transcriptomics data analysis or are an experienced computational biologist, the AtoMx platform provides avenues to explore, compute, iterate, and export CosMx SMI data.

In this post, we walk through a standard single-cell spatial transcriptomic cell typing pipeline in AtoMx ([Figure 1](#)), which is a typical pre-requisite to performing downstream spatial analyses.

**NOTE:** Recommendations in this guide are based on AtoMx v1.3.2.

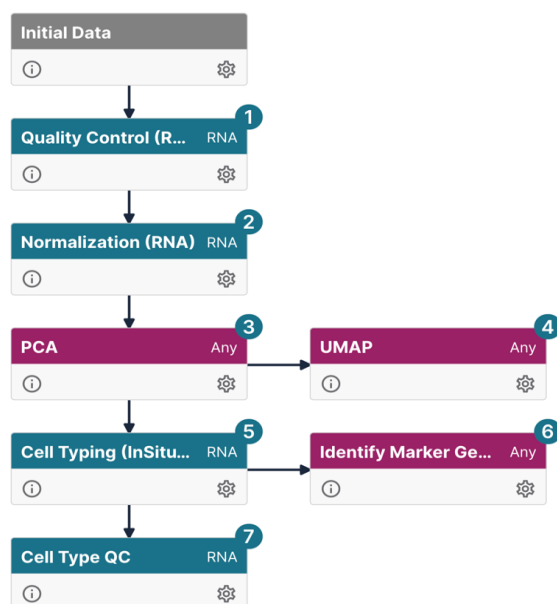


Figure 1: the standard AtoMx cell typing pipeline for CosMx RNA assays.

## Tip 1: Run Quality Control (QC)

The QC module in AtoMx is currently intended for flagging potentially lower quality cells for downstream removal after data export. QC and FOV QC are both recommended for every dataset, while additional methods can be explored.

**Cell QC** – flags cells with specific characteristics.

- **Minimum counts per cell:** 20 (for 1K RNA) and 50 (for 6K RNA); minimum counts per cell will be higher for higher plex assays and ultimately varies by dataset.

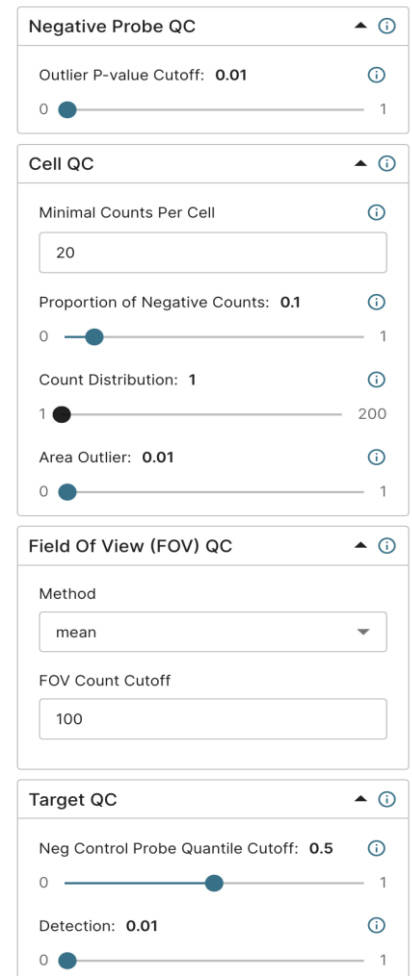
**FOV QC** – flags outlier FOVs that have overall low average expression or inconsistent clustering compared to neighboring FOVs.

- **Method:** mean
- **FOV Count Cutoff:** 100 (varies based on assay and dataset); minimum counts per cell will be higher for higher plex assays and ultimately varies by dataset. Outlier FOVs may also require detection from downstream analyses like UMAP.

**Note:** other methods/parameters can be explored as shown [Figure 2](#) but are optional and dependent on the dataset.

Signal (counts per cell) is the result of multiple factors, including tissue, image quality, and cell segmentation (see [CosMx 1000-plex RNA Assays: Considerations When Generating Single-Cell Spatial Data](#)). For FOV QC, we flag outliers that stand out with low average expression or disparate cell typing/spatial clustering compared to neighboring FOVs of similar biology.

**Note:** the quality control (QC) module in AtoMx flags but does not remove data; the flags are preserved in the data and can be removed after data export.



The screenshot displays the 'Quality Control (RNA)' module interface with the following settings:

- Negative Probe QC:** Outlier P-value Cutoff: 0.01
- Cell QC:**
  - Minimal Counts Per Cell: 20
  - Proportion of Negative Counts: 0.1
  - Count Distribution: 1
  - Area Outlier: 0.01
- Field Of View (FOV) QC:**
  - Method: mean
  - FOV Count Cutoff: 100
- Target QC:**
  - Neg Control Probe Quantile Cutoff: 0.5
  - Detection: 0.01

*Figure 2: The Quality Control (RNA) module in AtoMx with recommended parameters for 1K RNA assays.*

### What is the purpose of negative probes?

Every CosMx RNA panel contains a series of negative probes that target no known RNA transcripts. The average of the negative probe count helps assess background noise. Compare total transcripts (counts) per cell against the average of negative probes to evaluate relative sample performance across tissues in a study.

## Tip 2: Apply the appropriate normalization

CosMx RNA normalization adjusts for cell-specific total transcript abundance and distribution of counts (may vary between some FOVs and between samples) to minimize influence on downstream visualization and data analyses.

AtoMx includes three normalization methods for RNA assays:

- **Total Counts normalization (recommended)** – a global-scaling normalization method that normalizes gene counts for each cell by the total expression of each cell.
- **Seurat normalization** – total counts normalization that is multiplied by a scale factor (10,000) and natural-log transformed after adding 1.
- **Pearson Residuals normalization** – based on the estimated mean and variance:  $(\text{raw gene count in a cell} - \text{mean gene count in the cell}) / \text{SD of gene counts in the cell}$ .

Input Parameters

Normalization method

Total Counts

Seurat

Pearson Residuals

Total Counts

Figure 3: Normalization method options from the Normalization (RNA) module in AtoMx.

A reasonable assumption to make is that a cell's detection efficiency is estimated by its total counts; thus, one normalization approach is to scale each cell's profile by its total counts for normalization.

### When is Seurat or Pearson Residuals normalization preferred over Total Counts normalization?

Seurat normalization can be used to optimize the UMAP or run Leiden clustering. Seurat normalization may offer more flexible visualization capabilities, especially for studies with high-count genes (e.g., high-expressing housekeeping genes).

When working with smaller datasets ( $\leq 500,000$  total cells per study), Pearson Residuals normalization can outperform other normalization methods, especially when [identifying biologically variable genes](#). In larger datasets ( $> 500,000$  total cells per study), the Pearson residuals transformation is not as efficient computationally; in these cases, we typically recommend a total counts normalization.

**Note:** For differential expression and many advanced analysis modules, unnormalized data is utilized; in these cases, however, total counts is still an input variable in the algorithms.

## Tip 3: Adjust UMAP parameters

UMAP (Uniform Manifold Approximation and Projection) is an established technique for visualizing and distinguishing clusters of data through dimensionality reduction via non-linear estimation of related groups of cells or features. The input parameters used have a direct impact on the shape and structure of your UMAP, which can influence data interpretation. We recommend starting with the following workflow to plot your initial UMAP:

1. After normalization, run the PCA (principal component analysis) module set with 50 PCs.
2. Run the UMAP module with the following parameters:
  - Number of Neighbors: 30
  - Minimum Distance: 0.01
  - Spread: 5
  - Distance Metric: Cosine
  - Data Fraction: 1
3. Visualize the UMAP by cell types and overlay on the tissue image for further analysis.
4. If necessary, repeat Steps 2 and 3 by altering the parameters until satisfactory cluster separation is achieved.

**Input Parameters**

Number of Neighbors: 30 ⓘ

5  50

Minimum Distance ⓘ

0.01

Spread ⓘ

5

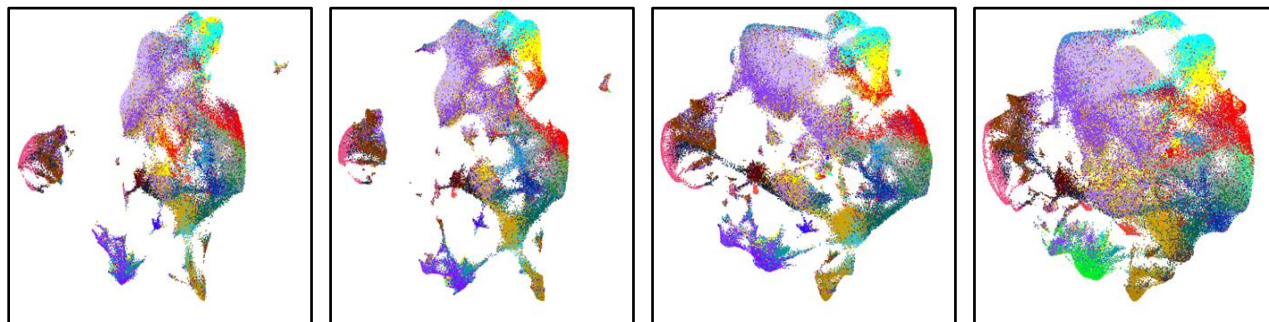
Distance Metric ⓘ

Cosine

Data Fraction ⓘ

1

*Figure 4: The UMAP module in AtoMx with recommended starting parameters.*



*Figure 5: Input parameters can greatly influence UMAP visualization and clustering, which impacts data interpretation. From left to right: spread value of 0.5, 2, 5, and 10 with the same dataset in AtoMx and all other parameters constant.*

**Note:** There are several parameters that affect the shape and structure of the UMAP embeddings. The appropriate parameterization of the UMAP is important for proper data visualization and interpretation ([McInnes et al. 2018](#); [Huang et al. 2022](#)).

## Tip 4: Enhance cell typing results through InSituType

Cell typing is a critical step in characterizing spatial single cells prior to performing downstream analyses (e.g., neighborhood, spatial correlation, and differential expression). The InSituType Cell Typing module is the recommended method to cell type CosMx data because it is specifically designed to leverage the high-plex nature of CosMx assays. Like UMAP parameterization, cell typing is a deeply iterative process and requires iteration. For most studies, providing a reference matrix (also called a cell profile library) can often enhance the quality of your cell typing results.

### Getting started with the Cell Typing (InSituType) module

This module uses the InSituType algorithm to identify and subset data based on cell types (see [Danaher et al. 2022](#)). AtoMx offers 3 clustering methods:

- **Unsupervised** – clusters cells without a reference. Use for exploration or when no appropriate references are available.
- **Supervised** – Cell type assignments are based on a reference matrix specifying the average expression profile of each cell type. Best for well-characterized samples with high-quality reference matrices.
- **Semi-supervised** – supervised typing plus discovery of novel clusters.

### What constitutes a quality reference matrix for InSituType?

- Include all the cell types present in your tissue. Cell types can be granular (e.g., separate profiles for “dendritic cell”, “M1 macrophage”, “M2 macrophage”, etc.) or broad (e.g., a single “myeloid” profile).
- Include most of the genes from your CosMx panel.
- Come from a *robust* dataset. For example, a profile based on just 20 cells from a rare cell population will be inaccurate.

If we are confident our reference profiles contain all the cell types in your tissue, supervised cell typing is recommended. However, if we intend to discover novel cell types or cell states, and/or we know the reference matrix is incomplete or inadequate, then semi-supervised cell typing is recommended.

Cell Typing (InSituType)

Cell typing using the InSituType algorithm

Input Parameters

☐ Unsupervised
 ☐ Supervised
 ☒ Semi supervised

Reference Matrix

Drag and drop file here or click Upload button

LymphNode\_A1118\_ProfileMatrix.csv

Number of Clusters

10

Semi-supervised clustering uses both a reference matrix and the number of clusters. The guidelines for each individual argument are shown in supervised (reference matrix) and unsupervised (n\_clusters). The algorithm initially fits cells using the reference matrix and cells that do not fit the given reference are split into n\_clusters.

☒ Include segmentation markers

*Figure 6: The Cell Typing (InSituType) module in AtoMx with the semi-supervised method selected and a sample reference matrix uploaded.*

*Where can I obtain a high-quality reference matrix? Can I make my own?*

We provide a [curated selection of defined reference matrixes from various human and mouse tissue](#) for cell typing CosMx data. Alternatively, it is possible to derive your own matrix from an appropriate single-cell RNA-seq (scRNA-seq) dataset. Use a defined reference matrix as a template file to produce a .csv or .RData file that will be recognized by the AtoMx platform. It is not necessary to scale the values to match the template file if all matrix data is from the same scRNA-seq dataset. If combining scRNA-seq datasets to create a single matrix, it is important to scale between datasets. For instructions to create your own matrix, refer to the section [on Cell profile matrices in the SpatialDecon vignette](#).

*What is the required number of clusters ( $n_{clus}$ ) to enter in semi-supervised and unsupervised cell typing?*

Unsupervised and semi-supervised cell typing requires you to choose the number of clusters to fit (in semi-supervised mode, this value is in addition to your reference cell types). We generally recommend a higher value for this parameter. For example, if we expect 4 new clusters, we will input 6-8 here. Redundant clusters can be merged later in the workflow.

*Do we include segmentation markers when cell typing with InSituType?*

We recommend including segmentation markers when cell typing. Cell typing can be enhanced by using protein marker data within the algorithm, which is intrinsic to every CosMx dataset. This typically results in improved cell typing results, as it employs additional information beyond the RNA transcripts measured in the assay.

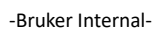
## Tip 5: Refine cell typing clusters

After running the Cell Typing module, examine these three outputs to QC your cell typing (and refine your clusters, if necessary):

- **Flightpath plot** – This shows the confidence of every cell type call and reveals the tendency of different clusters to be mixed/redundant with each other. To see this plot, click the plot icon on the InSituType box in the pipeline view.
- **Marker gene heatmap** – Run the marker gene module to create this and find it through the module’s “pipeline data” view.
- **Image viewer overlay** – For each cell type or set of related cell types, examine whether their spatial locations are consistent with tissue biology.

Refer to the [“Interpreting clustering results section” in the InSituType FAQs](#) for a detailed workflow.



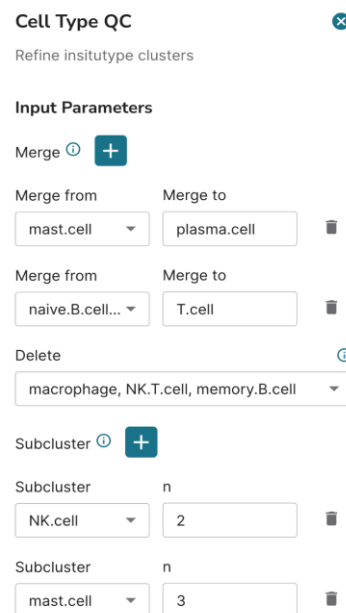




## The Cell Type QC module is a powerful tool for fine-tuning your clusters.

Initial cell typing results often require some modification based on user expertise. The Cell Type QC module enables updating of initial results in flexible and principled ways:

- **Merge clusters** – If two clusters appear to be from the same cell type, and we are not interested in fine-grained sub-clustering of that cell type, we can merge them. This will often be needed if we chose a generous value of `n_clust` as recommended. The merge clusters operation can also be used to rename clusters. This is important both for correcting mislabeled clusters and for naming unknown clusters.
- **Delete clusters** – in semi-supervised clustering, a cluster will sometimes be placed amidst reference cell types, capturing cells that could be assigned to known cell types. Deleting this unwanted cluster will send its cells to the next best-fitting cell type.
- **Subcluster** – if a cluster appears heterogeneous on the UMAP, or if we want to finely parse its biology, we can break the cluster into subclusters.



**Cell Type QC** ✕

Refine insitutype clusters

**Input Parameters**

Merge ⓘ +

Merge from	Merge to	
mast.cell	plasma.cell	<span>✕</span>
naive.B.cell...	T.cell	<span>✕</span>

Delete ⓘ

macrophage, NK.T.cell, memory.B.cell ⌵

Subcluster ⓘ +

Subcluster	n	
NK.cell	2	<span>✕</span>
mast.cell	3	<span>✕</span>

*Figure 8: The Cell Type QC module in AtoMx with example parameters.*

## Putting it all together: the essential workflow for quality cell typing

1. Select reference profiles, if appropriate.
2. Set parameters as described above and run InSituType.
3. Scrutinize cell typing data results by analyzing the flightpath plot, marker gene heatmap, and Image Viewer overlay to verify that cell types observed correspond to the cell types from the reference profile.
4. Refining Clusters: use the Cell Type QC module to refine clusters. This module enables deletion, merging, renaming (through the merging operation), and subclustering cell types.
5. Repeat Steps 3 and 4, as necessary.

For a deeper dive on the cell typing workflow, see the [InSituType FAQs](#) and [vignettes](#).

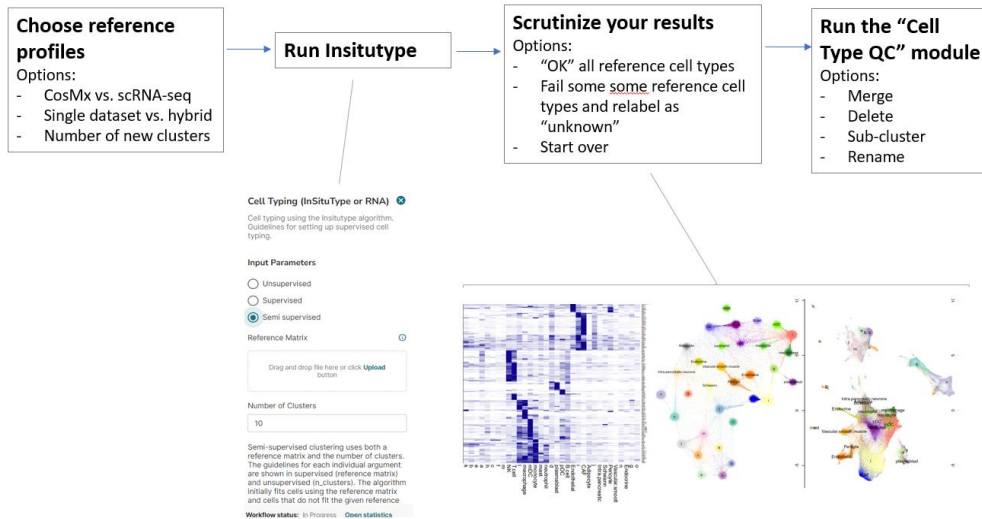


Figure 9: Recommended workflow for cell typing.

### Important note when cell typing with marker genes

We *do not* recommend cell typing using only marker gene expression. Individual marker gene expression levels can have lower signal. Methods using a cell's complete expression profile, for example Leiden clustering or InSituType, are far more robust. Instead, marker genes are quite useful for annotating unsupervised clusters.

For more information on cell typing and marker genes, see the [Useful References](#) section below.

## Tip 6: AtoMx data export enables additional analysis through open-source tools

AtoMx facilitates the export of data and images for analysis with open-source tools (ex: Seurat, Squidpy), allowing for integration with existing opensource analysis workflows. Flow cell images can be exported directly from the Image Viewer (Figure 7), while data can be exported using either built-in or custom export modules (Figure 10). Exporting your raw data allows for further analysis and visualization of your CosMx data using various Python- and R-based single-cell analysis tools such as Seurat, Squidpy, scanpy, anndata, napari.

**Note:** for opensource image analysis we recommend exporting images utilizing the export options shown in [Figure 10](#).

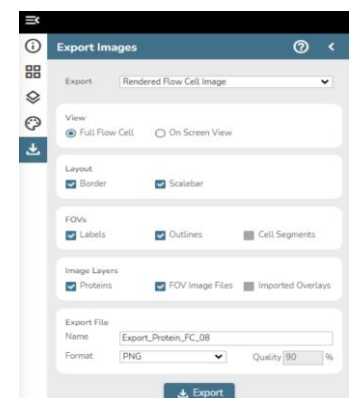


Figure 10: Image export options in Image Viewer in AtoMx.

### Export Dataset

**Description**

This module exports raw data, standalone Seurat object, corresponding TileDB array, and/or flat csv files available for download. The exported data comprises transcript counts and locations, annotation metadata, and user-initiated data transformations performed in AtoMx SIP prior to export. All results up to the point of export will be available in the Seurat object and TileDB array. While the RNA and Protein studies share the same format, the structure of the raw files folder will vary based on the analyte.

**Warning:** To prevent the duplication of files, please export raw files only once per study. Exporting raw files from multiple pipelines within the same study may result in terabytes of duplicated files.

**Input parameters**

**Flat CSV Files:**

- ☐ Export count matrix flat csv file
- ☐ Export cell metadata flat csv file
- ☐ Export transcripts flat csv file
- ☐ Export polygons flat csv file
- ☐ Export FOV positions flat csv file

**Tertiary Analysis Objects:**

- ☐ Export a Seurat Object
  - ☐ Export Seurat contains transcript coordinates (large data)
  - ☐ Export Seurat contains polygon coordinates (large data)
- ☐ Export TileDB array

**Raw Files:**

- ☐ Export Raw Files
  - ☐ Export SpotFiles folder to redo Target Decoding (large data)
  - ☐ Export Morphology2D folder (large data)
  - ☐ Export other Miscellaneous Data Files (large data, if available)

**Output Export Access**

To download your exported data, please connect using the below SFTP information through an SFTP application such as FileZilla, WinSCP, or a Command Line. Use your regular AtoMx credentials to connect.

**Hostname:** <sftp-endpoint-variable>  
**Port:** 22  
**Username:** <AtoMx user>

**Output Folder Name:** CosMx Human 6K Discovery Panel RNA study

### Custom Export Module v1.2.3

**Input Parameters**

Module version  
Version 1

- ☒ SeuratObject
- ☒ FullSeuratObject
- ☒ transcripts
- ☒ tiledbArray
- ☒ rawFiles
- ☒ exportFOVImages
- ☒ spotFiles

### Flat File Export v1.1.1

**Input Parameters**

Module version  
Version 1

- ☒ Generate count matrix file
- ☒ Generate cell metadata file
- ☒ Generate transcripts file
- ☒ Generate cell boundaries
- ☒ Generate FOV position file
- ☒ gzip files

Figure 11: Data export options in AtoMx. Left: Parameters for the built-in export function. Center: Parameters for the Custom Export Module Right: Parameters for the Custom Flat File Export Module.

## Built-in Export Function

- Raw decoded data files
- Seurat objects
- TileDB arrays
- Flat CSV files (ideal for Seurat analysis, enabling seamless integration with pre-existing workflows)

## Custom Export Module:

- Allows for direct export to an AWS S3 bucket
- Provides advanced flexibility and customization options
- Particularly useful for managing large datasets and integrating with cloud-based workflows

For instructions on exporting and analyzing CosMx RNA data in Python or R, see the [Useful References](#) section below.

## Useful References

- **CosMx Analysis Scratch Space:** an exploratory resource to accelerate open-source analysis of CosMx data. We regularly post to keep you updated with the latest how-to guides, tools, scripts, and best practices in spatial biology. Scratch Space also serves as an open forum for CosMx users to connect with our spatial data scientists.
  - Blog: <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/>
  - Repository: Scratch Space <https://github.com/Nanostring-Biostats/CosMx-Analysis-Scratch-Space>
  - Forum: <https://github.com/Nanostring-Biostats/CosMx-Analysis-Scratch-Space/issues>
- **NanoString University:** for the latest CosMx and AtoMx user manuals and tutorials (<https://university.nanostring.com>).
- **InSituType:** NanoString-developed single-cell spatial algorithm for cell typing.
  - Pre-print manuscript: <https://www.biorxiv.org/content/10.1101/2022.10.19.512902v1.abstract>
  - GitHub: <https://github.com/Nanostring-Biostats/InSituType>
- **Reference cell profiles for InSituType-based cell typing:** Reference cell profiles can be obtained from pre-existing single-cell or CosMx data. NanoString has also has a curated collection of reference matrices for use with your study. Each contain a library of cell profile matrices with accompanying statistics and metadata:
  - Cell profiles derived from CosMx data: <https://github.com/Nanostring-Biostats/CosMx-Cell-Profiles>
  - Cell profiles derived from scRNA-seq data: <https://github.com/Nanostring-Biostats/CellProfileLibrary>
- **Cell Typing and Marker Genes**
  - <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/on-cell-typing-with-marker-genes/>
  - <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/marker-gene-smoothing/index.html>
- **Custom modules for AtoMx:** get the R script for the Export, Flat File Export, and RNA QC Plots custom modules:
  - <https://github.com/Nanostring-Biostats/CosMxDACustomModules>
- **Exporting and analyzing CosMx RNA data in Python or R:**
  - <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/squidpy-essentials/squidpy-essentials.html>
  - [https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/h5ad\\_conversion/](https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/h5ad_conversion/)
  - <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/using-napari-for-cosmx-data/>
  - <https://nanosttring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/vignette-basic-analysis/>

- **Technical background for UMAPs:**
  - Primer on applying UMAPs in single-cell datasets:
    - [McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction. arXiv 2018." \*arXiv preprint arXiv:1802.03426\* 10 \(2018\).](#)
  - For a review of the mathematics underlying UMAPs: <https://arxiv.org/abs/1802.03426>
  - For a review of dimension reduction methods, including UMAPs: [Huang, H., Wang, Y., Rudin, C. \*et al.\* Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization. \*Commun Biol\* 5, 719 \(2022\).](#)
- **CosMx public datasets:** <https://nanosttring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/cosmx-smi-human-pancreas-ffpe-dataset/>
  - <https://nanosttring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/human-frontal-cortex-ffpe-dataset/>
  - <https://nanosttring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/cosmx-smi-mouse-brain-ffpe-dataset/>
  - <https://nanosttring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/human-liver-rna-ffpe-dataset/>
  - <https://nanosttring.com/products/cosmx-spatial-molecular-imager/ffpe-dataset/cosmx-smi-human-pancreas-ffpe-dataset/>

As you begin your spatial biology journey, remember:

- Spatial data analysis, especially cell typing single-cell data, is an iterative process, requiring flexibility, adaptability, and a willingness to refine your approach as needed.
- Combining multiple analytical approaches, optimizing parameters, and considering contextual information (including tissue images) is crucial for achieving accurate and meaningful results.

Should you have further questions or need additional guidance, reach out to your local Field Application Scientist or contact our Support team ([support@nanosttring.com](mailto:support@nanosttring.com)).