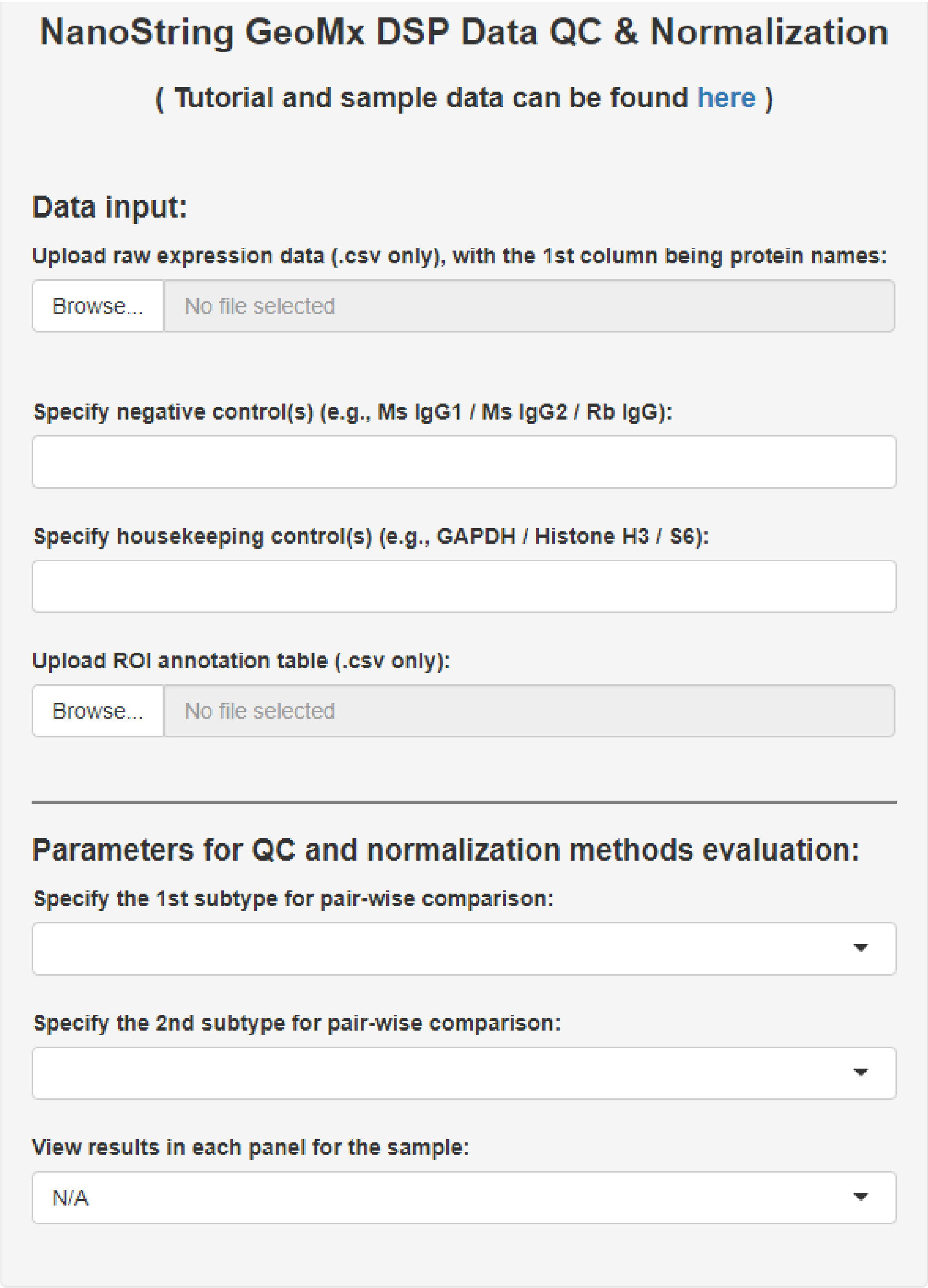
Evaluation of DSP Normalization Methods – Tutorial

Chi Wang and Daheng He
Biostatistics & Bioinformatics Shared Resource Facility (BB SRF)
Markey Cancer Center, University of Kentucky

(November 2022)

The App provides a convenient interface for users to normalize the data of NanoString GeoMx protein assay in multiple ways, and evaluate their performance in order to choose an optimal normalization method for downstream analyses. The initial input panel on the left-hand side of the App looks like this:



NanoString GeoMx DSP Data QC & Normalization

(Tutorial and sample data can be found here)

Data input:

Upload raw expression data (.csv only), with the 1st column being protein names:

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG):

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6):

Upload ROI annotation table (.csv only):

Parameters for QC and normalization methods evaluation:

Specify the 1st subtype for pair-wise comparison:

Specify the 2nd subtype for pair-wise comparison:

View results in each panel for the sample:

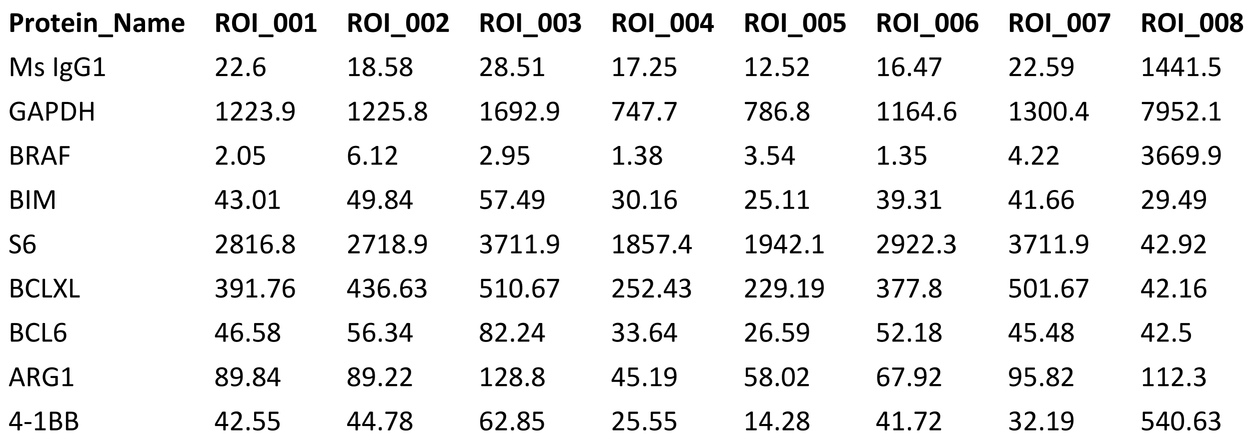
N/A

The users are expected to provide necessary info in this panel sequentially. Here are step-by-step graphical instructions on each input step:

Data input:

Upload raw expression data (.csv only), with the 1st column being protein names:

This step is to upload the raw DSP expression data, which has only been ERCC-normalized, to remote server. As indicated in the input statements, currently the App can only process input data table in csv (comma-separated values) format. In preparing the csv file of the raw expression data table, please make sure that the first column is for protein IDs, and the rest of columns for protein expression values of ROIs. The basic structure of raw expression data input should look something like this in Excel:



Protein_Name	ROI_001	ROI_002	ROI_003	ROI_004	ROI_005	ROI_006	ROI_007	ROI_008
Ms IgG1	22.6	18.58	28.51	17.25	12.52	16.47	22.59	1441.5
GAPDH	1223.9	1225.8	1692.9	747.7	786.8	1164.6	1300.4	7952.1
BRAF	2.05	6.12	2.95	1.38	3.54	1.35	4.22	3669.9
BIM	43.01	49.84	57.49	30.16	25.11	39.31	41.66	29.49
S6	2816.8	2718.9	3711.9	1857.4	1942.1	2922.3	3711.9	42.92
BCLXL	391.76	436.63	510.67	252.43	229.19	377.8	501.67	42.16
BCL6	46.58	56.34	82.24	33.64	26.59	52.18	45.48	42.5
ARG1	89.84	89.22	128.8	45.19	58.02	67.92	95.82	112.3
4-1BB	42.55	44.78	62.85	25.55	14.28	41.72	32.19	540.63

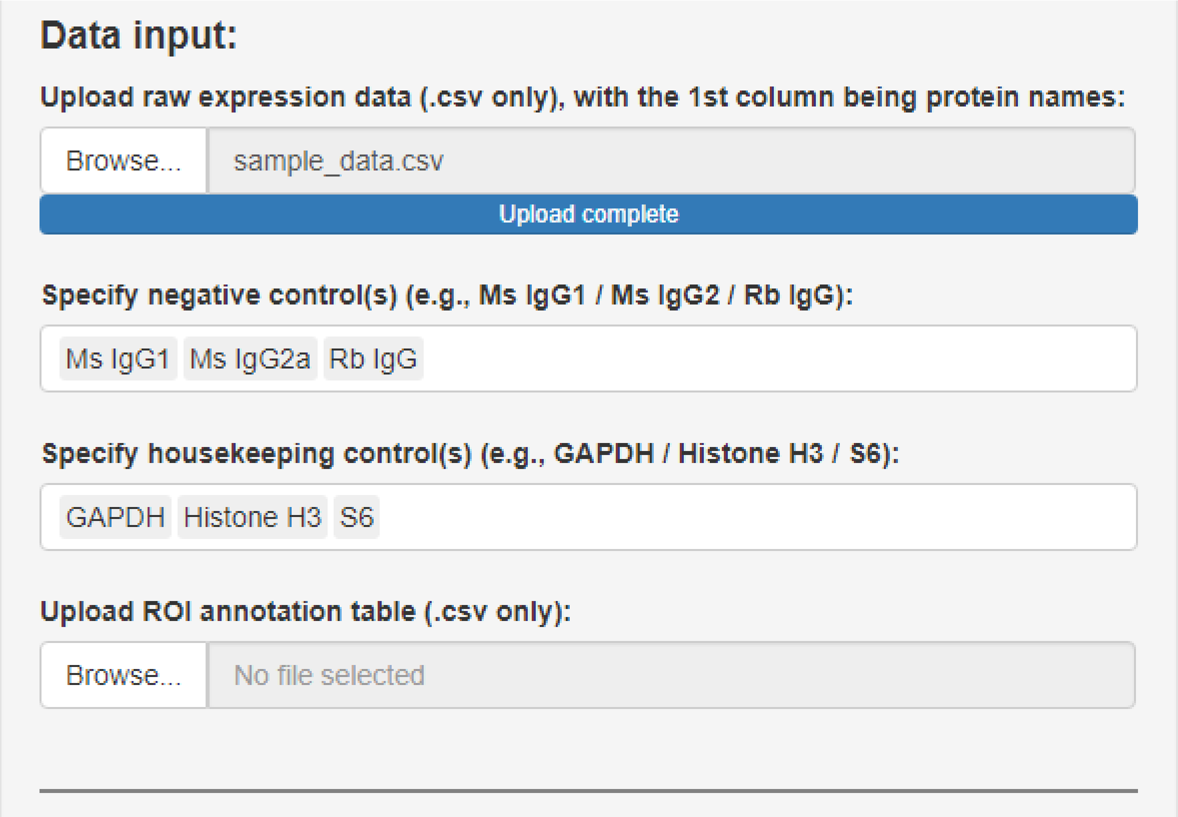
When the raw expression datafile in csv format is ready, just click on the "**Browse…**" button to initiate the local file path browsing and uploading process. You are only allowed to upload a single csv file at a time.

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG):

The App will monitor the uploading process. Once the expression data has been uploaded successfully in step 1, the App will automatically fetch the protein names from the first column of your data and summarize the protein names in the form of dropdown menu, which you may browse and click to make multiple choices informing the App which proteins are the negative controls (typically there are 3 negative controls). To speed up the protein names searching, you may simply type the first two or three letters of each wanted protein name in the text box, the App will narrow down the list of names that match your typing.

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6):

Specifying the housekeeping controls (typically there are 3 housekeeping controls) is very similar to the previous step of specifying negative controls. Here is what the input panel looks like after both negative and housekeeping controls have been manually specified:



Data input:

Upload raw expression data (.csv only), with the 1st column being protein names:

Browse... sample_data.csv

Upload complete

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG):

Ms IgG1 Ms IgG2a Rb IgG

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6):

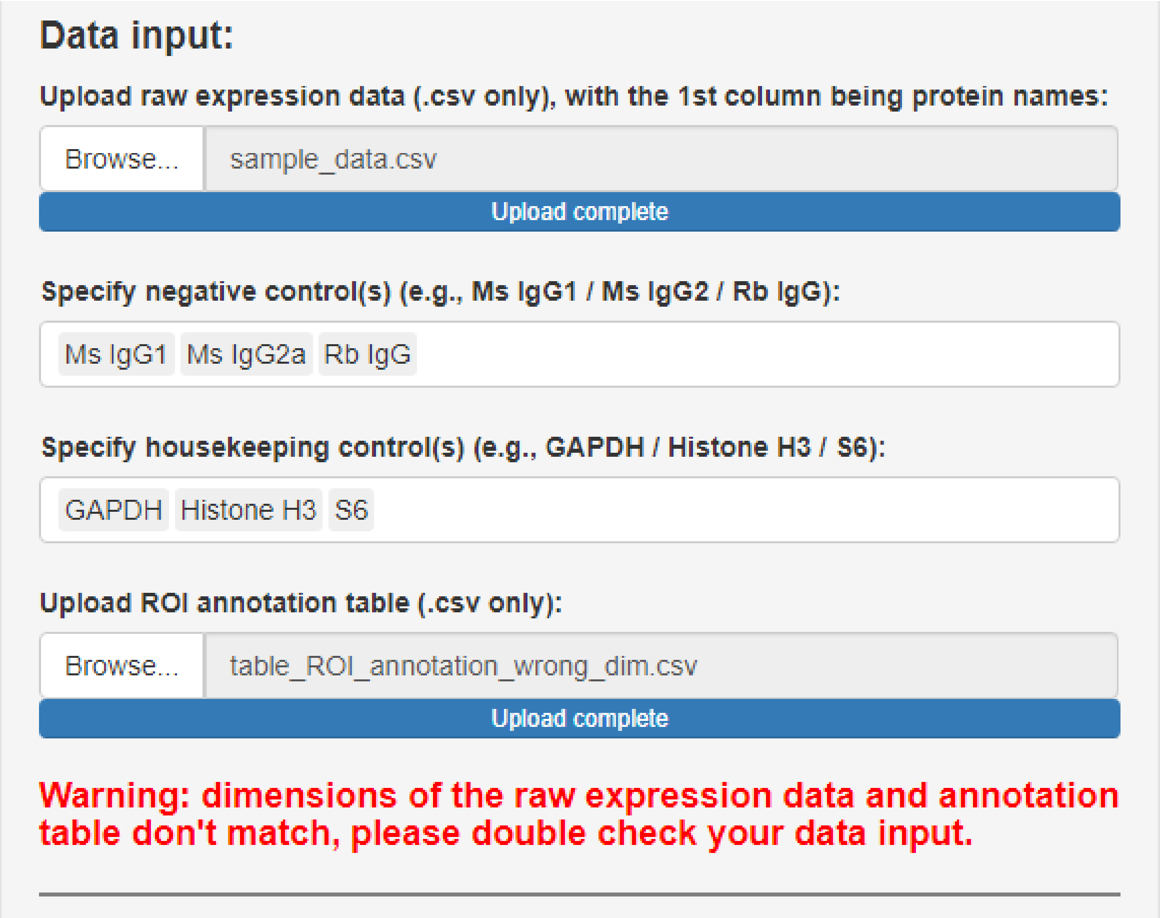
GAPDH Histone H3 S6

Upload ROI annotation table (.csv only):

Browse... No file selected

Upload ROI annotation table (.csv only):

As an important component of DSP data set, you need to provide an additional table in csv format, which provides annotating info for each ROIs in the raw expression table, such as the area and nuclei count of each ROI, and the comparison groups each ROI belongs to. The annotation table must be arranged in such a way that its rows from top to bottom are 1-to-1 matching the ROI columns (i.e., excluding the first column for protein names) of the expression data matrix from left to right. It goes without saying that the number of rows of the ROI annotation table must match the number of ROI columns in the expression data. The uploading of ROI annotation table is very similar to the uploading of raw expression data -- just click on the "**Browse…**" button to initiate the local file path browsing and uploading process. Please note that if the App detects that if the number of rows in the ROI annotation table does not match the number of ROI columns in the raw expression table, the App will not allow you to continue to the remaining steps, but return a warning message and stand by until you have fixed the mistake and uploaded the corrected annotation table, here is what the warning message looks like:



Data input:

Upload raw expression data (.csv only), with the 1st column being protein names:

sample_data.csv

Upload complete

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG):

Ms IgG1 Ms IgG2a Rb IgG

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6):

GAPDH Histone H3 S6

Upload ROI annotation table (.csv only):

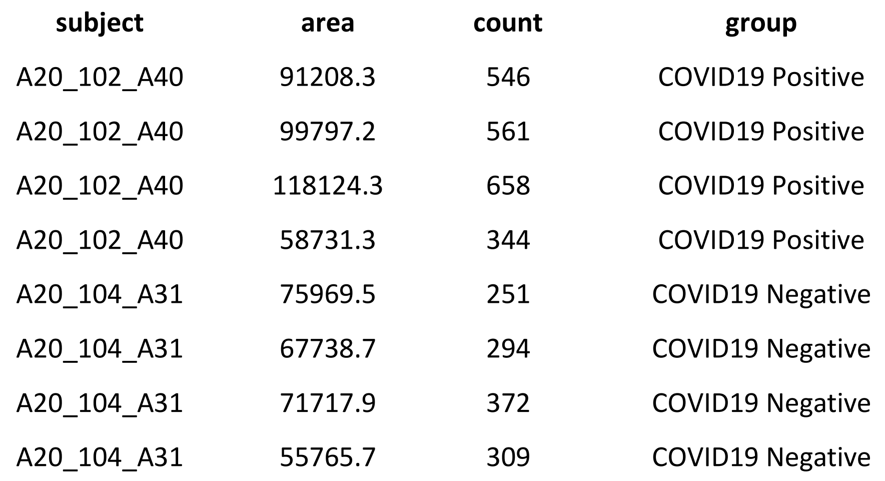
table_ROI_annotation_wrong_dim.csv

Upload complete

Warning: dimensions of the raw expression data and annotation table don't match, please double check your data input.

In our realistic practice of DSP data analysis, we have noticed that in many DSP data multiple subjects were involved in DSP data collection, and the tissue slides from each subject were often prepared and measured at different times, under different lab conditions, or even by different machines, these technical variations together with the genuine biological difference among subjects often make the subject-related heterogeneity in DSP data quite sizeable, and may significantly affect the conclusions of downstream analyses. Such a heterogeneity must be properly identified and handled. Therefore, we are making it mandatory to provide the subject info in the annotation table, even if your data really just come from a single subject.

Each column in the annotation table corresponds to a specific kind of annotating info, such as ROI area, nuclei count, subgroup and subject each ROI belongs to. You need to arrange your annotation table in such a way that its rows, from top to bottom, are in exactly the same order as those columns (i.e., ROIs) from left to right in the raw expression data. The basic structure of annotation table input should look something like this in Excel:



subject	area	count	group
A20_102_A40	91208.3	546	COVID19 Positive
A20_102_A40	99797.2	561	COVID19 Positive
A20_102_A40	118124.3	658	COVID19 Positive
A20_102_A40	58731.3	344	COVID19 Positive
A20_104_A31	75969.5	251	COVID19 Negative
A20_104_A31	67738.7	294	COVID19 Negative
A20_104_A31	71717.9	372	COVID19 Negative
A20_104_A31	55765.7	309	COVID19 Negative

As shown above, a complete annotation should contain 4 columns (the order of these columns in your annotation table do not matter), named exactly as "**subject**", "**area**", "**count**", "**group**", all are **case-sensitive**. Among the 4 columns, **the "subject" and "group" columns are mandatory**. In the case of a single subject without a pre-specific subject ID, you still need to have a column named as "subject", and just put some arbitrary short string, for example "S1", as the subject's ID in all of the row entries of the column. **The "area" and "count" columns, on the other hand, are optional**. If these two types of information were not recorded at all in the data collection steps, then you can simply ignore these columns, and just submit an annotation table with only the two mandatory columns "subject" and "group". In the absence of either or both of the two optional columns, the App will skip the evaluations that require the information. **There is yet another situation, in which you must ignore the "area" column** –

sometimes the DSP data may be collected from the ROIs with identical areas, in which case the area info cannot provide any meaningful value in the evaluation of normalization methods, and more importantly, providing an “area” column with all identical values will make several internal statistical evaluations divergent, triggering unexpected error messages and failure of the App.

Download normalized data (.xlsx):

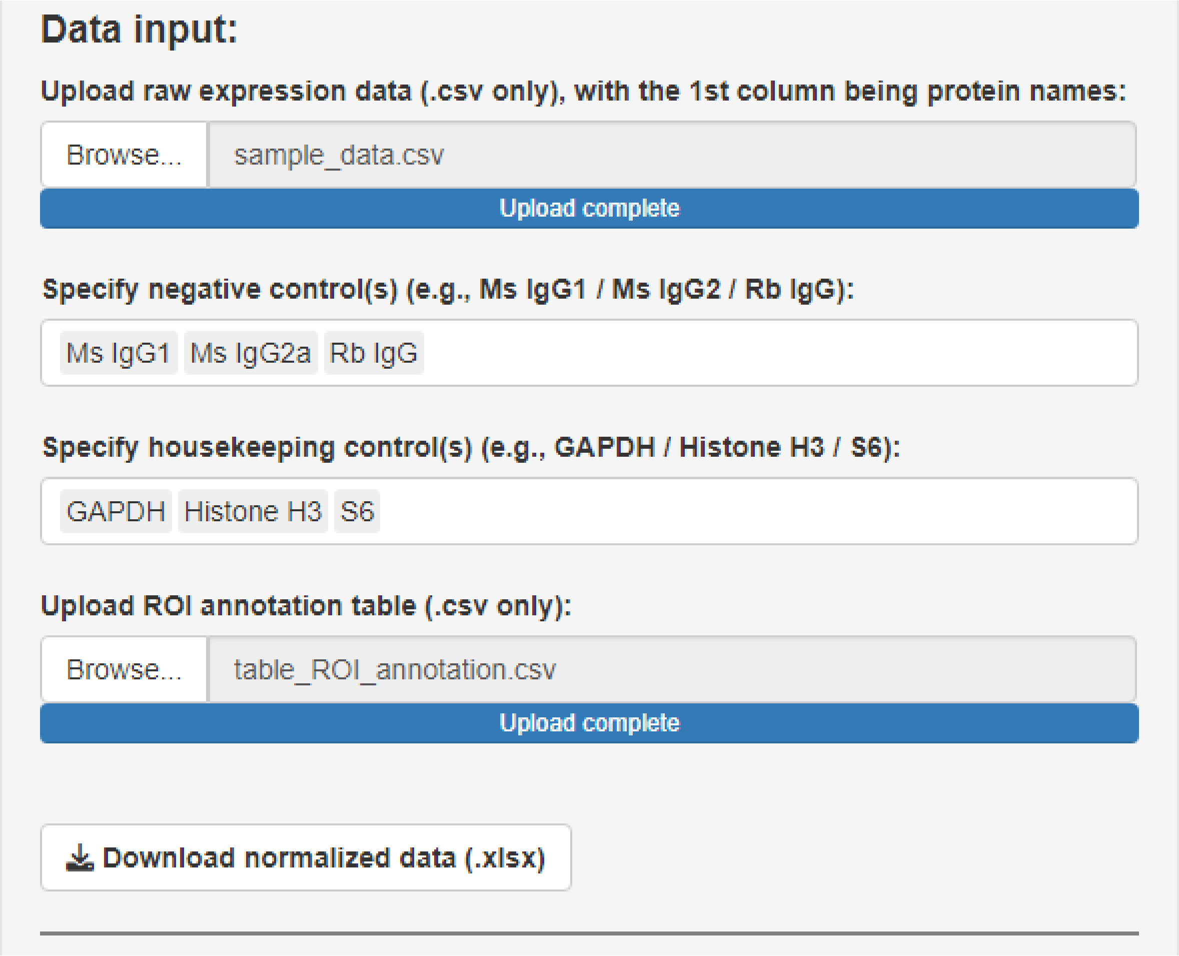
Upon providing all the necessary inputs correctly, including

1.) the raw expression data;

2.) the manually specified list of negative and housekeeping controls;

3.) the ROI annotation table,

the App will dynamically generate a “**Download normalized data (.xlsx)**” button that is previously invisible on the input panel:



Data input:

Upload raw expression data (.csv only), with the 1st column being protein names:

Browse... sample_data.csv

Upload complete

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG):

Ms IgG1 Ms IgG2a Rb IgG

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6):

GAPDH Histone H3 S6

Upload ROI annotation table (.csv only):

Browse... table_ROI_annotation.csv

Upload complete

Download normalized data (.xlsx)

You can click the button to download the normalized DSP expression matrices based on the build-in normalization methods. There are in total 17 build-in normalization methods, but depending on your input, if the area and/or nuclei count info is absent, certain normalization methods requiring such info will not be available. The downloaded normalized data is in xlsx format containing multiple sheets, among which the first sheet is the table giving detailed descriptions of the normalization methods, those methods that are not available due to the absence of area or nuclei count info are marked with "(not available)", the rest of sheets are for the normalized expressions by each available normalization method. You can choose the data, which is normalized by the most optimal normalization method as decided by those evaluation metrics in the output panels, for your downstream analysis for more reliable conclusions.

Parameters for QC and normalization methods evaluation:

Specify the 1st subtype for pair-wise comparison:

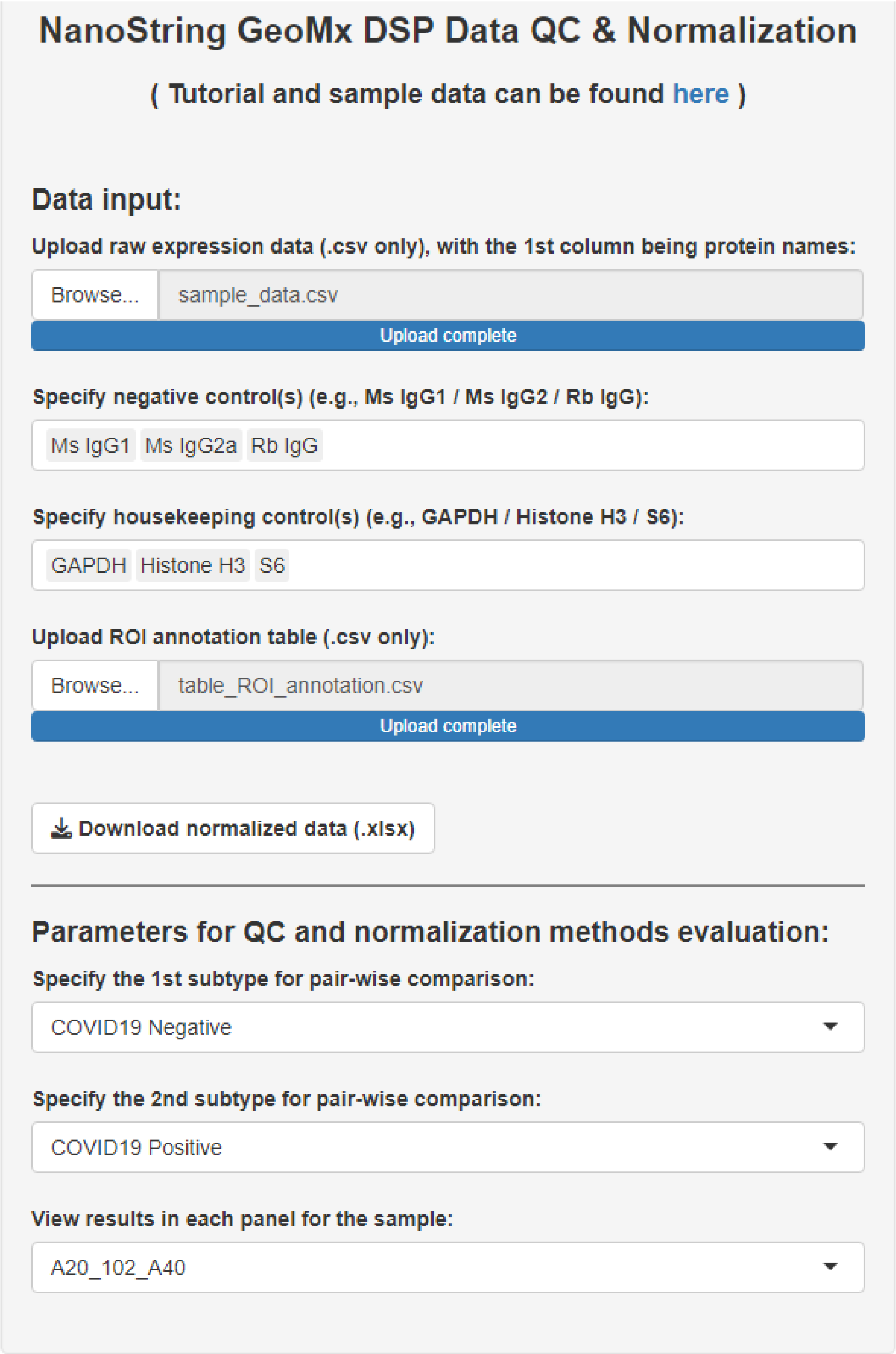
Specify the 2nd subtype for pair-wise comparison:

These two steps are to let the App know between which two biologically distinct subgroups you would like to compare. After uploading the correct ROI annotation table, the App will automatically scan the mandatory column "group" to summarize the unique values in the column, and summarize them in the form of dropdown menu, which you may browse and click sequentially to make a choice. For each subtype specification, you are only allowed to make a single choice -- currently the App only supports two-group comparison.

View results in each panel for the sample:

After uploading the correct ROI annotation table, the App will also scan the mandatory column "subject", summarize its unique values in a dropdown menu, and by default choose the first subject ID alphabetically as the input of subject for evaluation. At this stage (assuming you have chosen two distinct groups of interest in the steps right above), all of the required inputs are ready for the App, therefore the evaluation processes in the background will be initiated automatically. You can start to browse each output panel on the right-hand side of the web page to view its detailed output. You can choose a different subject from the subject menu, and all output panels will respond to the change of subject simultaneously. Please note that in some panels, computing and generating figures may take a while, so please allow a few seconds for the updating of figures.

Upon providing all the necessary inputs to the App correctly, the input panel on the left-hand side of the web page should look something like this:



NanoString GeoMx DSP Data QC & Normalization

(Tutorial and sample data can be found here)

Data input:

Upload raw expression data (.csv only), with the 1st column being protein names: sample_data.csv

Specify negative control(s) (e.g., Ms IgG1 / Ms IgG2 / Rb IgG): Ms IgG1 Ms IgG2a Rb IgG

Specify housekeeping control(s) (e.g., GAPDH / Histone H3 / S6): GAPDH Histone H3 S6

Upload ROI annotation table (.csv only): table_ROI_annotation.csv

Parameters for QC and normalization methods evaluation:

Specify the 1st subtype for pair-wise comparison: COVID19 Negative

Specify the 2nd subtype for pair-wise comparison: COVID19 Positive

View results in each panel for the sample: A20_102_A40

Acknowledgement:

In searching for relevant DSP datasets to test our models, the research groups of Desai N et al [1] and Delorey TM et al [2] kindly shared their DSP data and provided help with explaining their DSP data via private communications. We deeply appreciate their kind help.

Reference:

[1] Desai N, Neyaz A, Szabolcs A, Shih AR et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. Nat Commun 2020 Dec 9;11(1):6319. PMID: 33298930

[2] Delorey TM, Ziegler CGK, Heimberg G, Normand R et al. COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature 2021 Jul;595(7865):107-113. PMID: 33915569