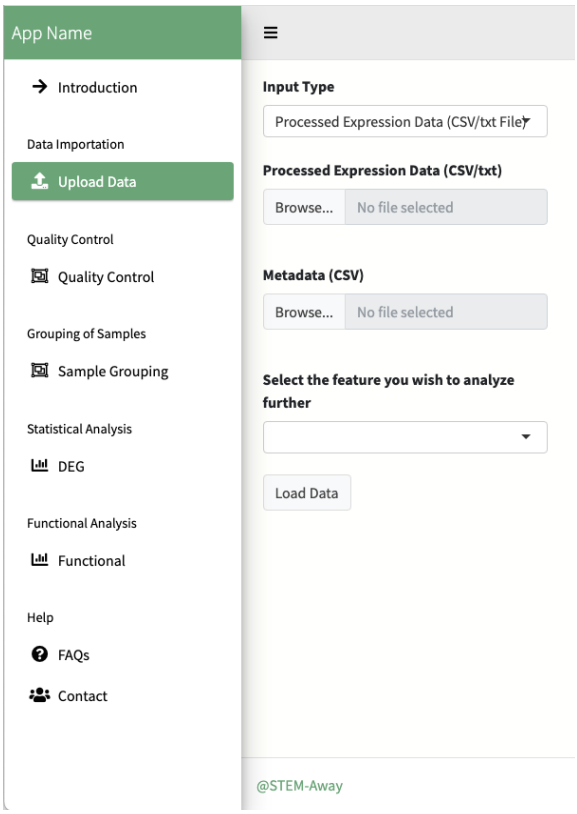
Contributors: @Arian Veyssi @Aditi Verma

Upload Data



Input Type:

- Processed Expression Data (CSV/txt file)
- Raw Affymetrix Data (.tar file containing CEL
- files)
- Raw Illumina Data (.tar file containing IDAT files)
- GEO Accession Number

Processed Expression Data (csv/txt): The selected GSE datasets from the GEO database.

Metadata: Input the simplified csv file that contains columns for sample IDs, cancer/normal identifications, and sample groups. There are a maximum of 2 sample groups.

Selecting Features: By uploading the metadata CSV file, the names of the columns will be extracted and put into this dropdown list. You will now be able to select which feature you would like to analyze further, such as comparing cancer vs. normal groups or any other condition of interest.

Instructions for Processed Expression Data and Acquiring Metadata:

1. Select and download expression data from the GEO database. Explore the GEO database here.

Example of downloading expression data from GEO:

Supplementary file	Size	Download	File type/resource
GSE8671_RAW.tar	558.0 Mb	(http)(custom)	TAR (of CEL, CHP)

Raw data provided as supplementary file
Processed data included within Sample table
Processed data provided as supplementary file

Custom GSE8671_RAW.tar archive:

Supplementary file	File size
GSM215051.CEL.gz	5.0 Mb
GSM215051.CHP.gz	6.5 Mb
GSM215052.CEL.gz	4.5 Mb
GSM215052.CHP.gz	6.3 Mb

2. Unzip the .tar file(s).

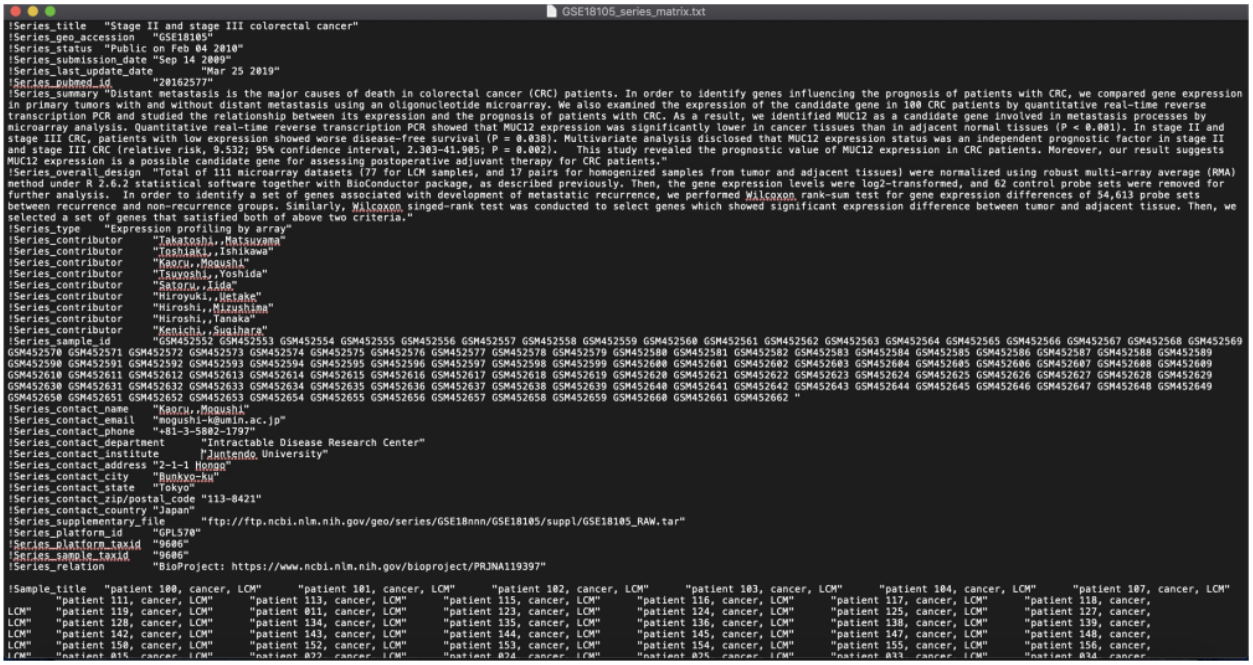
- For Windows users, you may need to download WinRAR (preferably x64).
- For Mac users, double-click the .tar folder.

3. Acquire metadata by following the steps below. Metadata is important as this data will be used in batch correction and limma analysis (a method of statistical analysis on the significance of differential gene expression (DGE)).

STEP 1: Download SERIES MATRIX FILES txt from GEO and unzip.

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Once you unzip the txt file, you should see something like this:



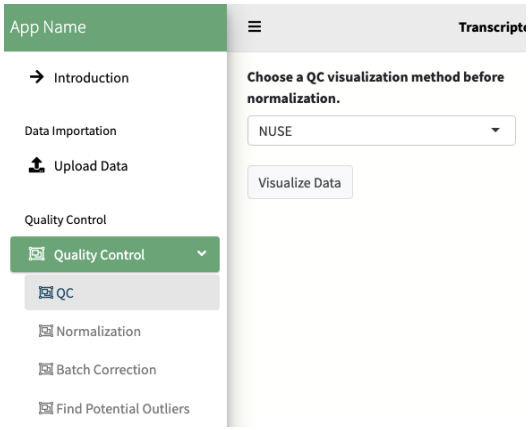
STEP 2: Find the information you need, such as the sample name and features, and paste it into an Excel Spreadsheet.

Example of Metadata

	A	B
1	Sample Name	CN
2	GSM494556	C
3	GSM494557	C
59	GSM494616	N
60	GSM494617	N

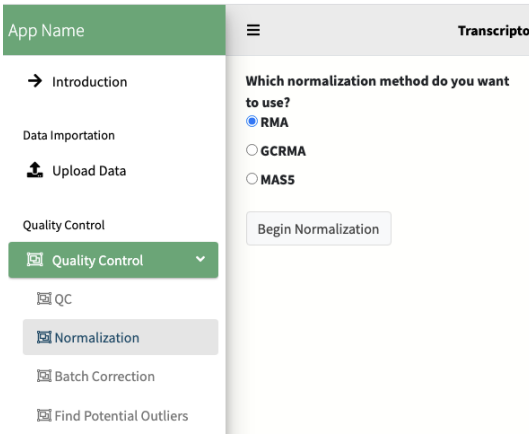
Quality Control:

QC:



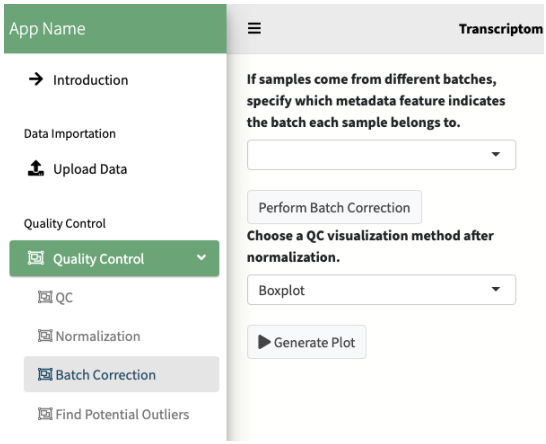
QC Visualization Methods (Un-Normalized): Select visualization method: NUSE or RLE. Click on “Visualize Data" to visualize the un-normalized gene set into a box plot.

Normalization:



Normalization Methods: Select one of the normalization methods: RMA, GCRMA, or MASS. Click “Begin Normalization” to normalize your data.

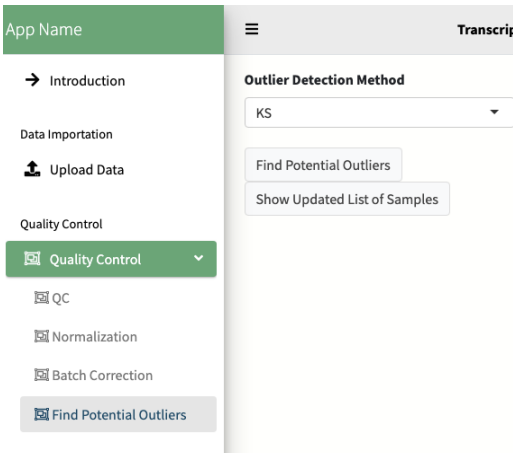
Batch Correction:



Metadata Feature (only if there are different batches): Specify which metadata feature corresponds with which batch. Click “Perform Batch Correction” to begin batch correction.

QC Visualization Method (Normalized): Select visualization method: Boxplot or PCA. Click on “Generate Plot” to visualize the normalized gene set into a box plot.

Find Potential Outliers:



Outlier Detection Method: Select the desired method of outlier identification: KS, SUM, or Upper quartile.

Find Potential Outliers: Click to find outliers.

Show Updated List of Samples: Updated list of samples will exclude outliers found using the “Find Potential Outliers” button.