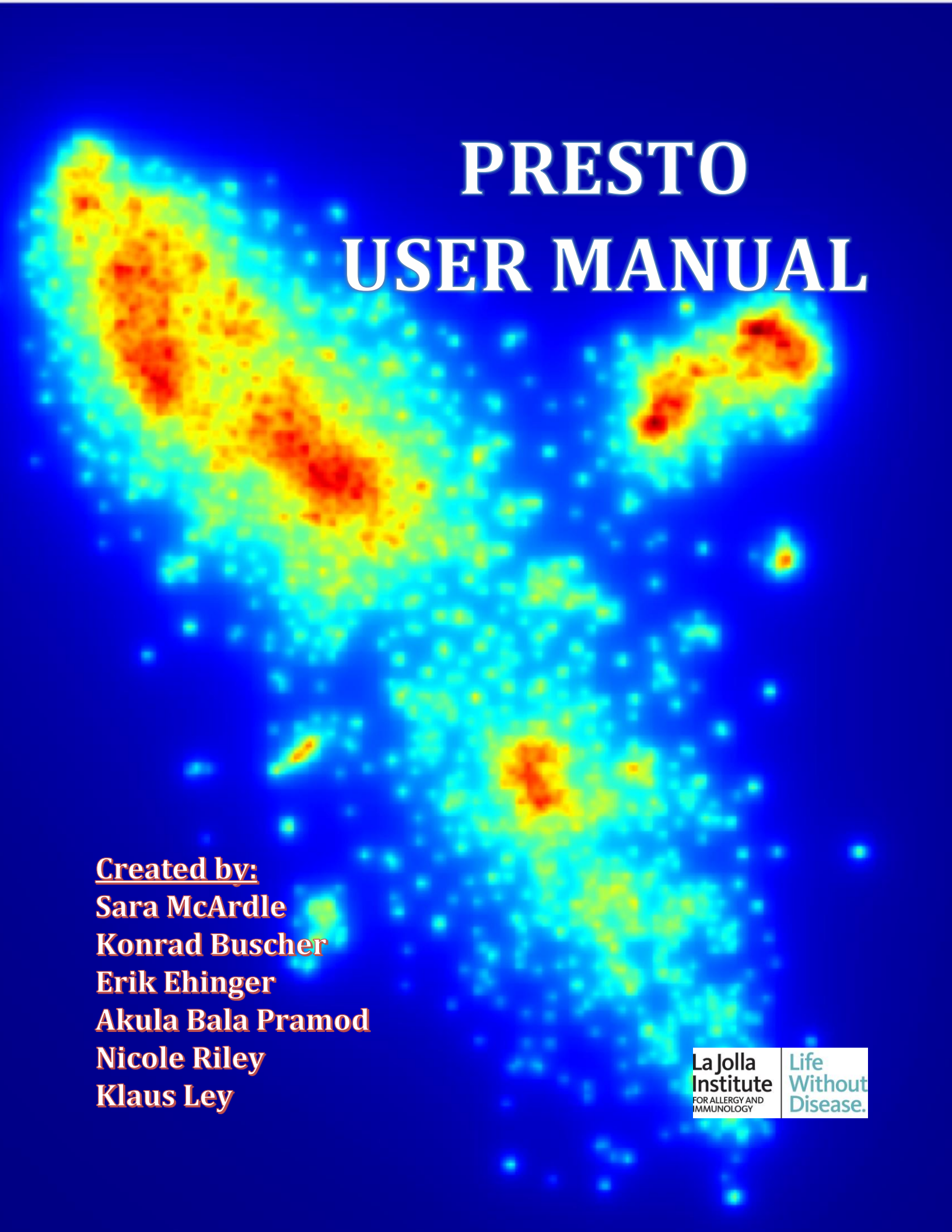


PRESTO USER MANUAL

Created by:

Sara McArdle
Konrad Buscher
Erik Ehinger
Akula Bala Pramod
Nicole Riley
Klaus Ley

La Jolla Institute FOR ALLERGY AND IMMUNOLOGY | Life Without Disease.

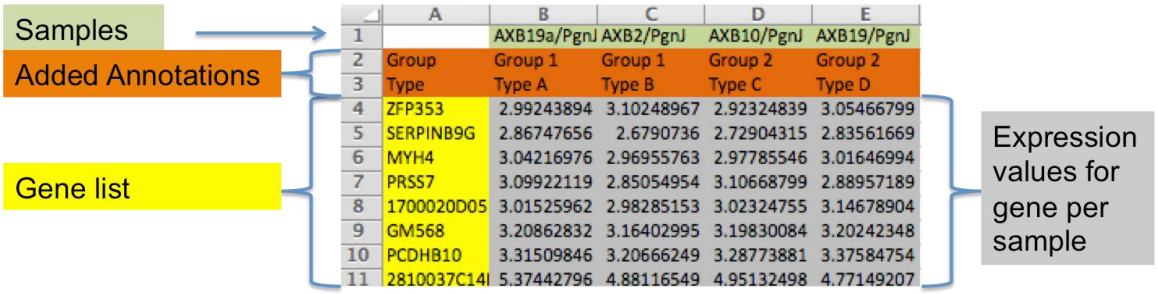
I. Introduction

Cohesive visualization and interpretation of hyperdimensional, large-scale omics data is an ongoing challenge. Here, we present PRESTO, a 'PREdictive Stochastic neighbor embedding Tool for Omics', which allows unsupervised dimensionality reduction of multivariate data matrices with thousands of subjects or conditions. Using PRESTO on genome-wide transcriptomic data of mouse macrophages across 75 mouse strains, highly variant genes are grouped to visualize their multi-dimensional pattern. Core co-expressed gene networks are identified and assigned to physiological and pathological functions that cannot be gleaned from conventional bioinformatics analyses. We applied PRESTO to clinical biopsy transcriptomic and proteomic data from large patient cohorts and detected disease-defining signatures in antibody-mediated kidney transplant rejection, renal cell carcinoma, and relapsing leukemia. Diagnostic signatures correlated patient with survival, and were validated in independent test data sets, demonstrating the versatility and clinical applicability of PRESTO.

II. Getting Started with PRESTO

Requirements:

1. Matlab and folder of PRESTO m-files OR PRESTO Stand-alone Application (PC only)
2. At least one excel sheet with omics data in this format:



	A	B	C	D	E
1		AXB19a/PgnJ	AXB2/PgnJ	AXB10/PgnJ	AXB19/PgnJ
2	Group	Group 1	Group 1	Group 2	Group 2
3	Type	Type A	Type B	Type C	Type D
4	ZFP353	2.99243894	3.10248967	2.92324839	3.05466799
5	SERPINB9G	2.86747656	2.6790736	2.72904315	2.83561669
6	MYH4	3.04216976	2.96955763	2.97785546	3.01646994
7	PRSS7	3.09922119	2.85054954	3.10668799	2.88957189
8	1700020D05	3.01525962	2.98285153	3.02324755	3.14678904
9	GM568	3.20862832	3.16402995	3.19830084	3.20242348
10	PCDHB10	3.31509846	3.20666249	3.28773881	3.37584754
11	2810037C14I	5.37442796	4.88116549	4.95132498	4.77149207

Colors are used here to highlight certain areas. Colors can be optionally used, but only highlight the areas that have data listed. If you have any formatting or data outside the grid PRESTO will not import correctly.

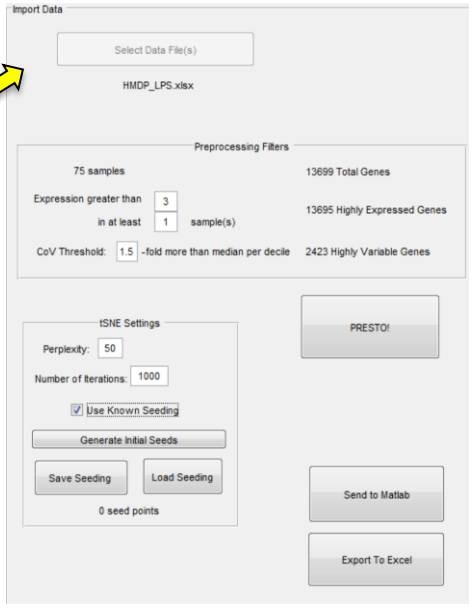
The data can be of any –omics type: RMA-normalized microarray sequencing, RPKMs, gene counts, proteomics spectral counts, etc.

The annotations are optional and can include as many rows of text as desired. Here the example shows 2 fabricated annotations: Group and Type. Clinical parameters, gender, experimental group, or any other note can be included as long as it is formatted as text.

III. Adding 1 Omics Excel File to Presto

a. Import of the Data

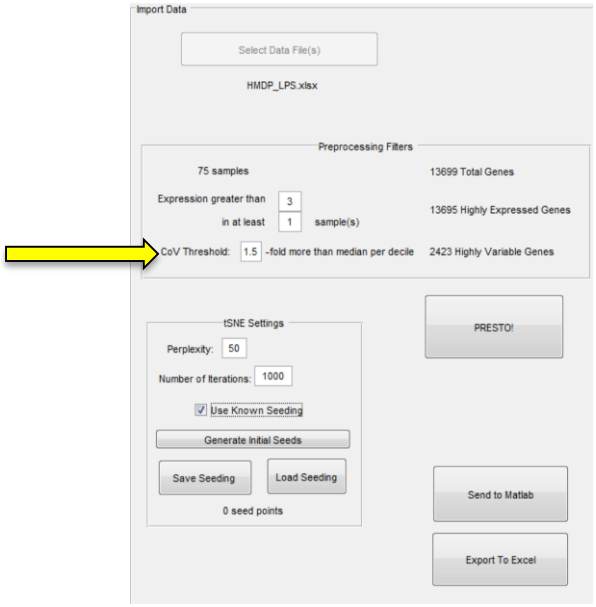
1. Start the PRESTO interface by either:
   a. Open Matlab and type PRESTO4 (using uppercase letters) and hit enter
   b. Double click on the application icon
2. Import your data excel file by clicking the "Select Data File(s)" button and select the excel sheet with the data
3. After a few seconds, the preprocessing filters area will tabulate with the detected number of genes and samples. Double check these numbers to ensure the data imported correctly.



b. Preprocessing Filters

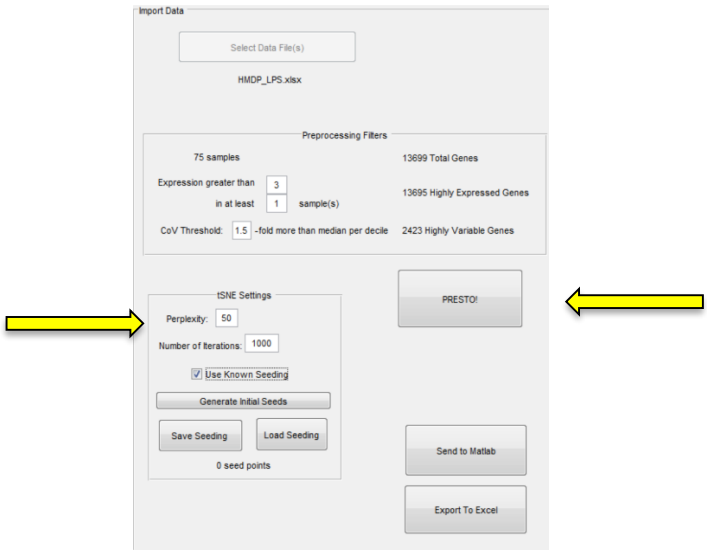
1. Edit the expression threshold, which is the lowest value that will be considered while analyzing the data by PRESTO. Therefore, it will filter out any genes that have an expression value that is below the entered expression number for the minimum amount of samples analyzed (e.g., a gene must be have achieved an RMA value of 3 in at least 1 sample, or in at least half of samples, etc.) This is important to remove noise from genes at or near the detection limit. If you are working with counts or RPKM values, it is essential to remove those genes that were undetected in all samples. The number of "highly expressed" genes will update automatically.

2. Adjust the coefficient of variation (CoV) threshold to filter out genes that do not vary between samples. This is based on the median CoV in each decile of the highly expressed



list. We recommend adjusting this number until 2000 -4000 genes remain. This can be set to 0 to select all highly expressed genes.

3. You can click PRESTO! now or adjust the tSNE Settings. Clicking PRESTO will start the optimization iterations. Once clicked, you can no longer import new data, though you can adjust the settings. To change files, restart the application.



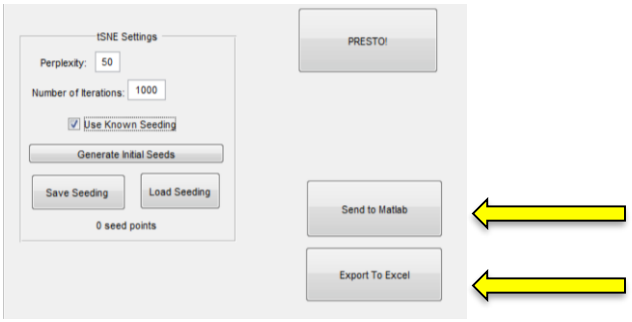
c. tSNE Settings

1. The suggested perplexity is 50; there is an ideal range of numbers that give good boundary definitions, but either extremes will not give a clear analyzable tSNE image (either 2 or 1000 will give you a scattered plot with no clear regions).
2. The default number of iterations is 1000. That number also be adjusted as needed, but it is not recommended to go below 500 (recommended is 2000 – 5000 for final images)
3. If it is required to keep the orientation of the analyzed image, check the box "Use Known Seeding", then:
   a. Load a saved seeding OR
   c. Click "Generate an initial seed" which then can be saved as well
   d. When finished editing the tSNE Settings click PRESTO!
4. If any changes were made to the Pre-Processing Filter settings in step III. d., be sure to click "Generate an initial seed" again

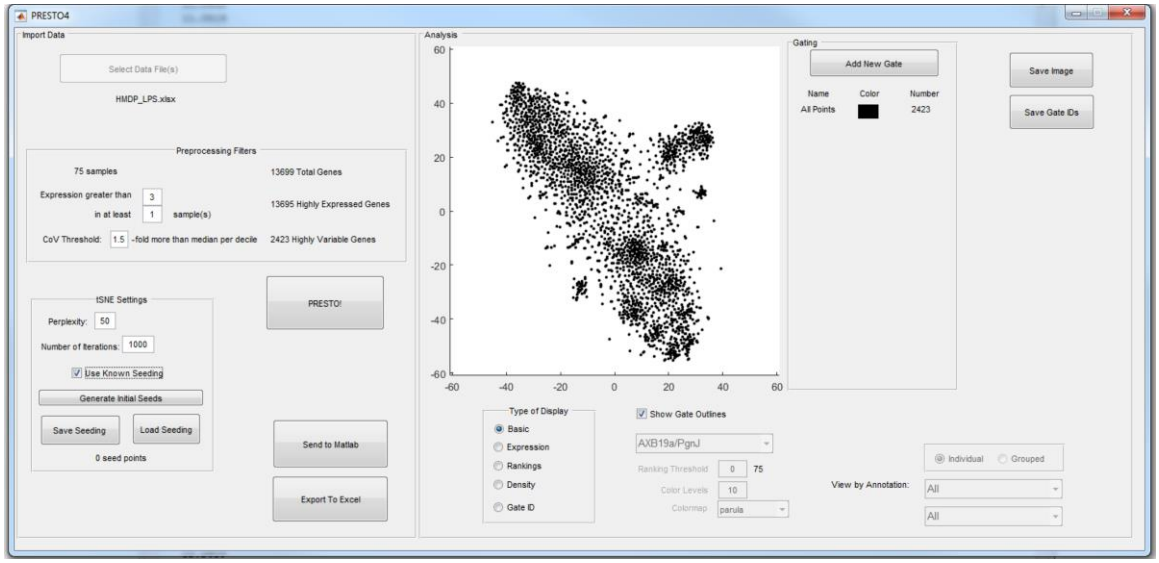
The next steps can be followed in any order.

d. Send to Matlab and Export to Excel

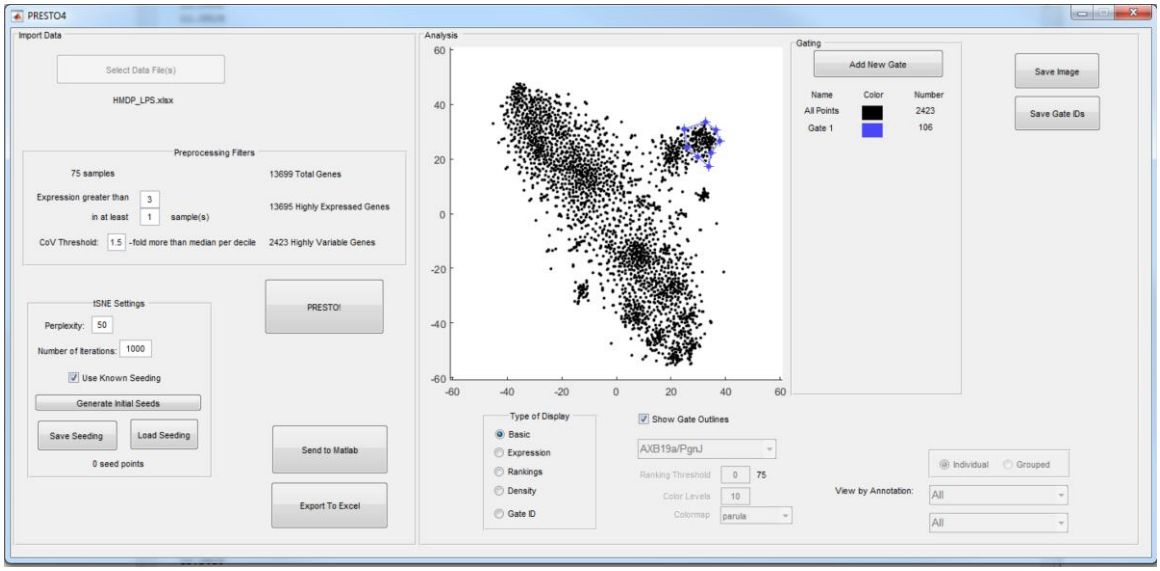
The filtered data can be sent to Matlab and saved as a .mat file or exported to an excel sheet for further analysis at any point during your analysis



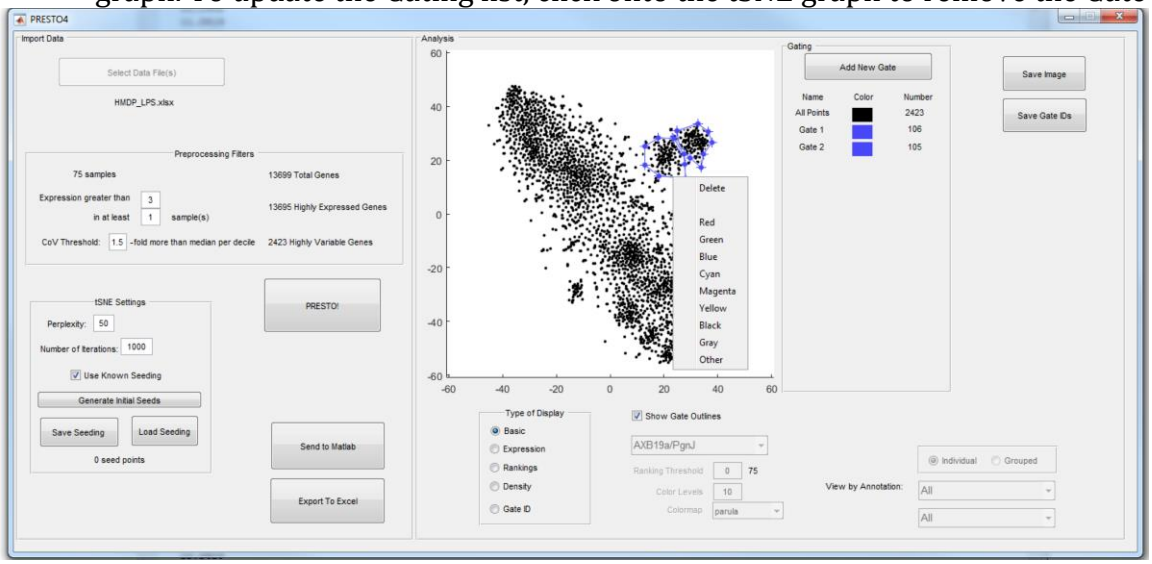
e. tSNE Analysis using Gating



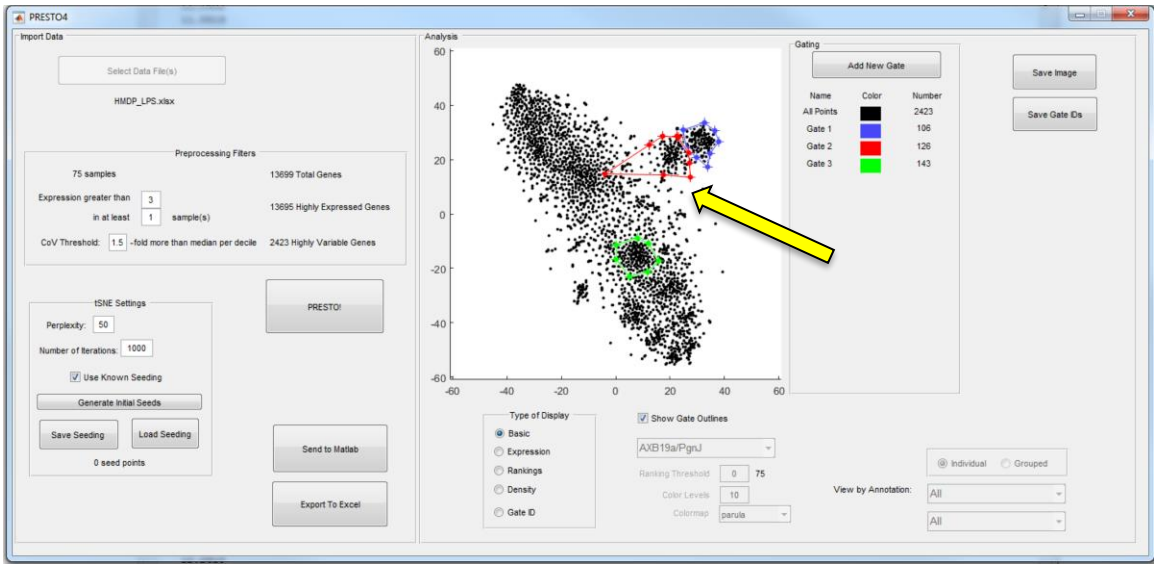
1. To add new gates for certain areas, click “Add New Gate”, then click on the tSNE image where to “frame” a specific area by clicking every time with the left mouse button to generate a corner of the area to be framed



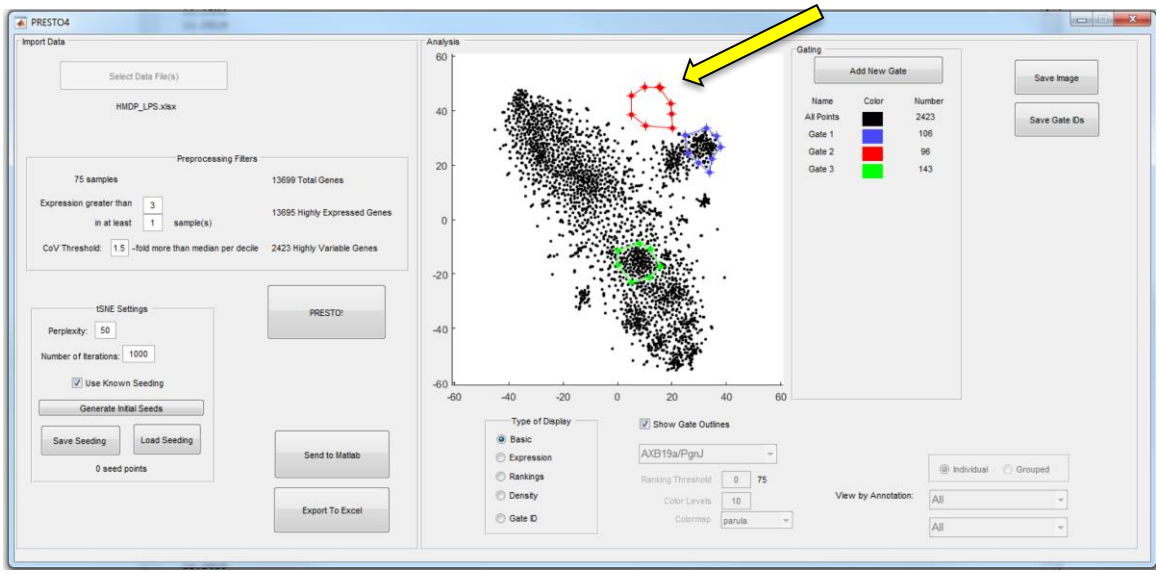
2. Once finished framing the area, do a “right mouse click” where the color can be changed or the gate can be deleted. When deleting a gate, it will first only delete the gate from the tSNE graph. To update the Gating list, click onto the tSNE graph to remove the Gate ID.



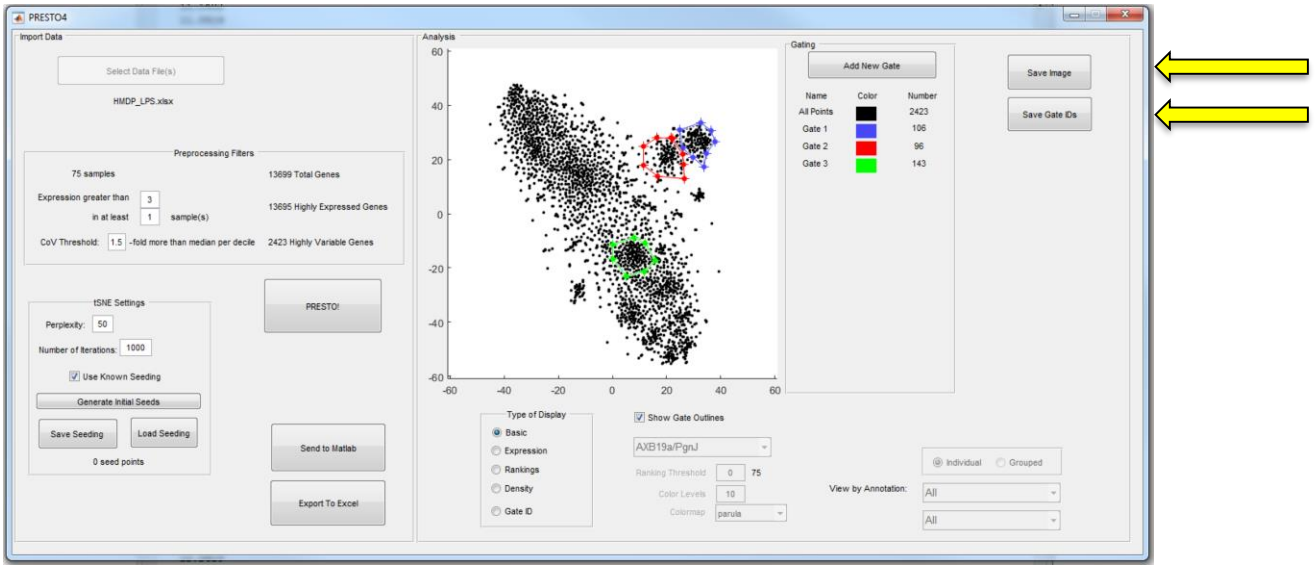
3. Now when positioning the mouse over any of the corners, the corners can be adjusted by pulling them where they need to be



4. The whole framed area can also be moved (data points will not move)



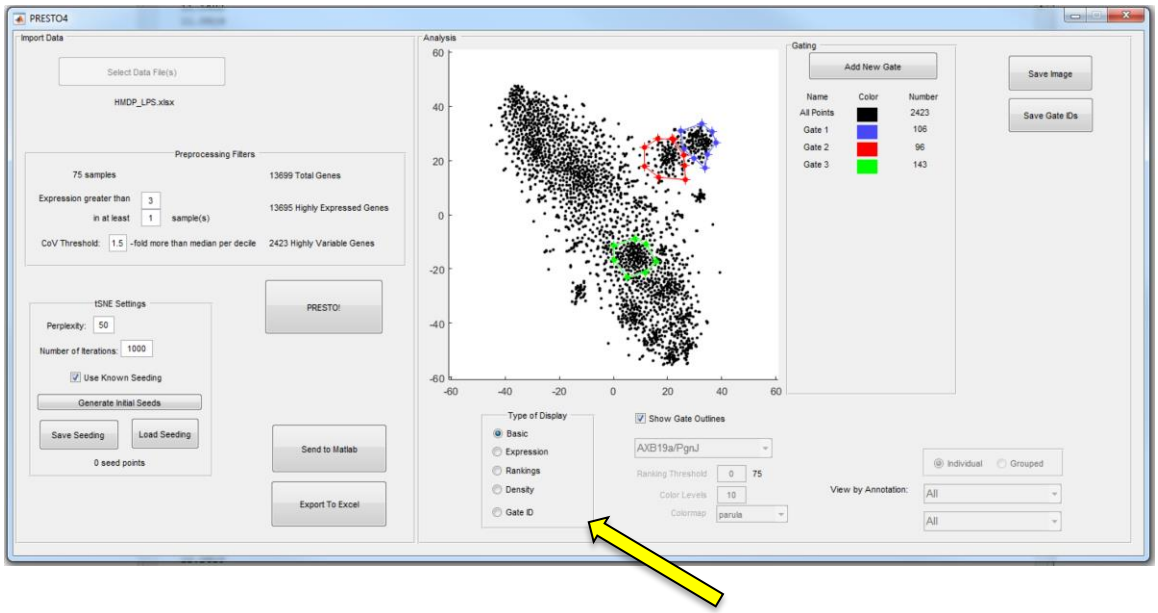
5. There can be as many areas framed as needed; if more color choices are needed click on "others" when selecting a color. Currently, only 15 gates will appear in the list, though more gates can still be added for analysis.
6. The Gate IDs can be saved as an image file at anytime



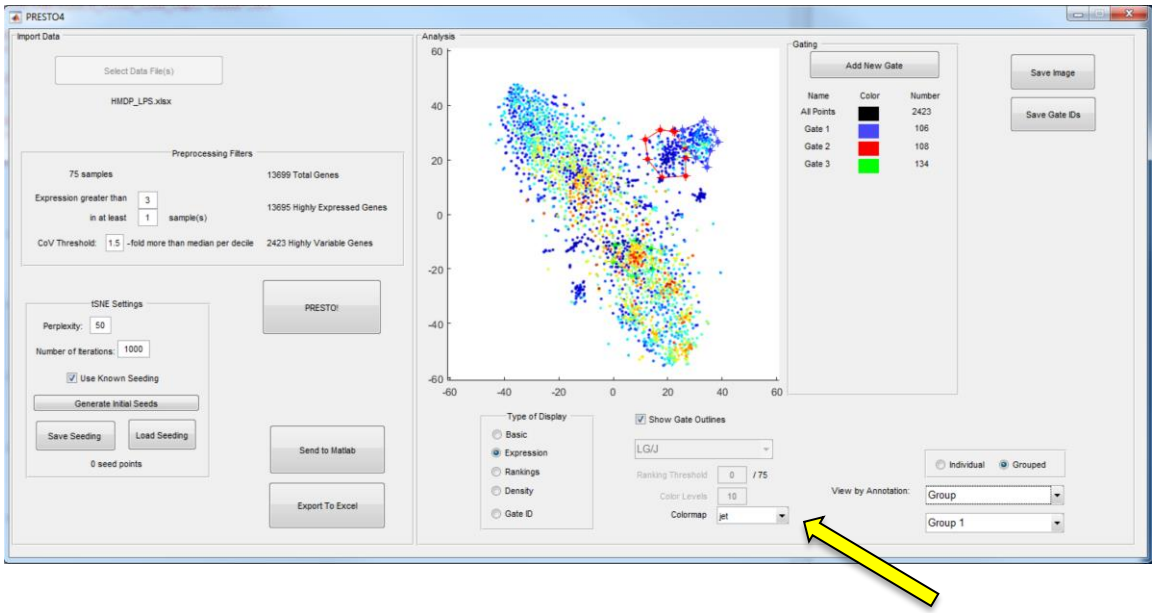
7. The image generated can also be saved at any time (see arrow above)
8. Furthermore, the Gate Outlines can be removed for visualization (that does not delete the gates permanently) by unchecking the box labelled "Show Gate Outline"

f. tSNE Analysis using different Types of Displays

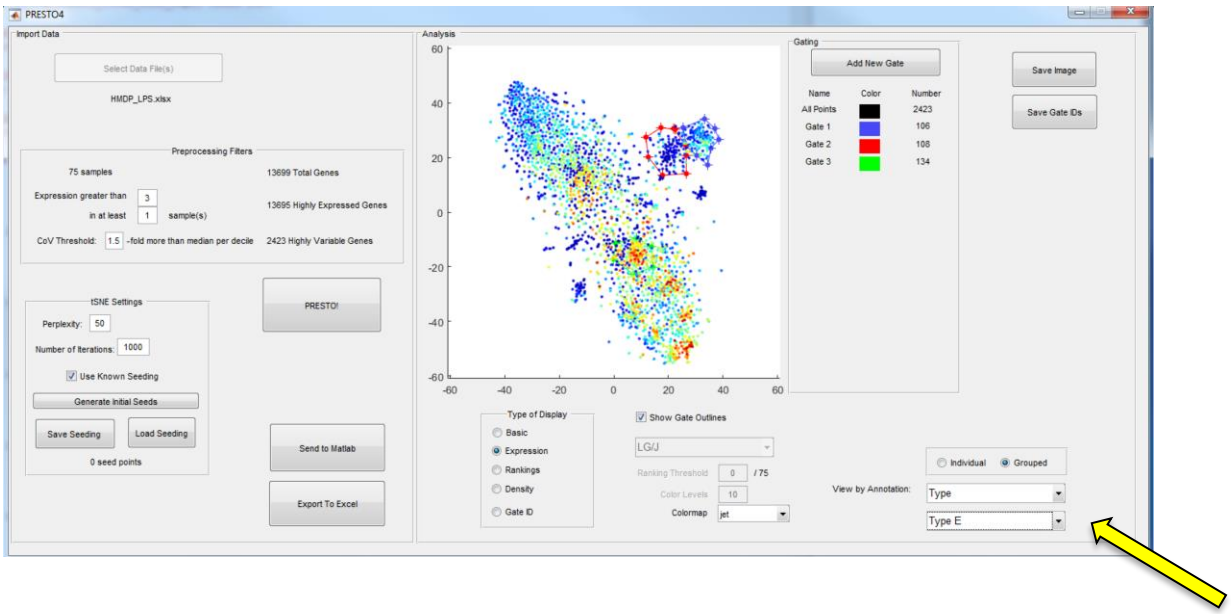
1. To analyze the image further select any of the listed "Type of Display": Basic (default), Expression, Rankings, Density, and Gate ID



2. **Expression**. This shows the expression value of each gene
    a. One can change the color of the tSNE image (jet is a good suggestion from the 4 listed choices)



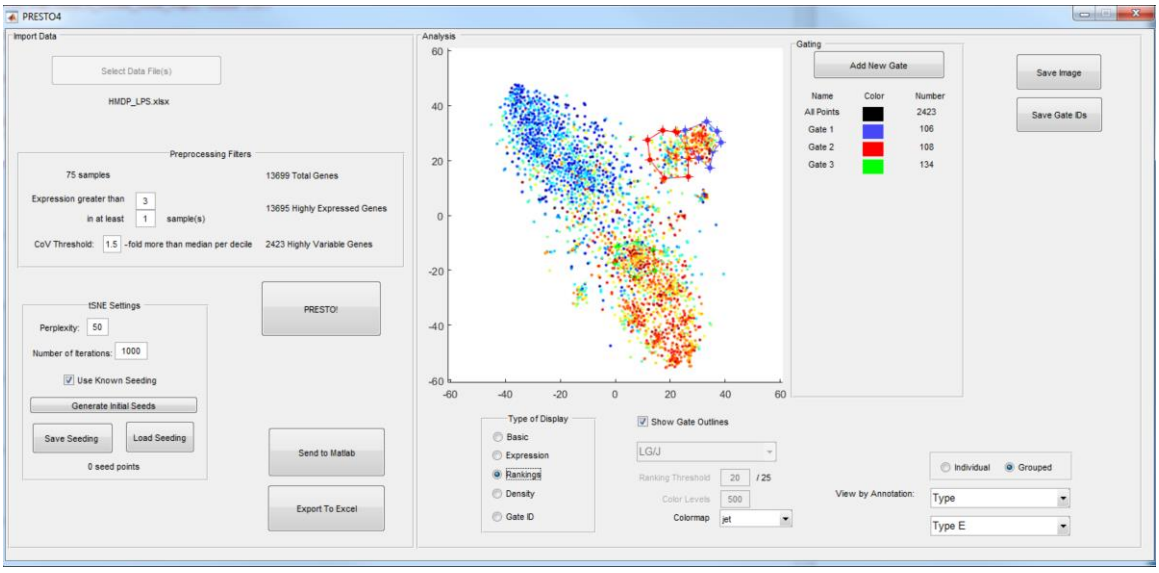
b. One can also see how the expression for each sample changes by selection which sample needs to be analyze by clicking the drop-down menu



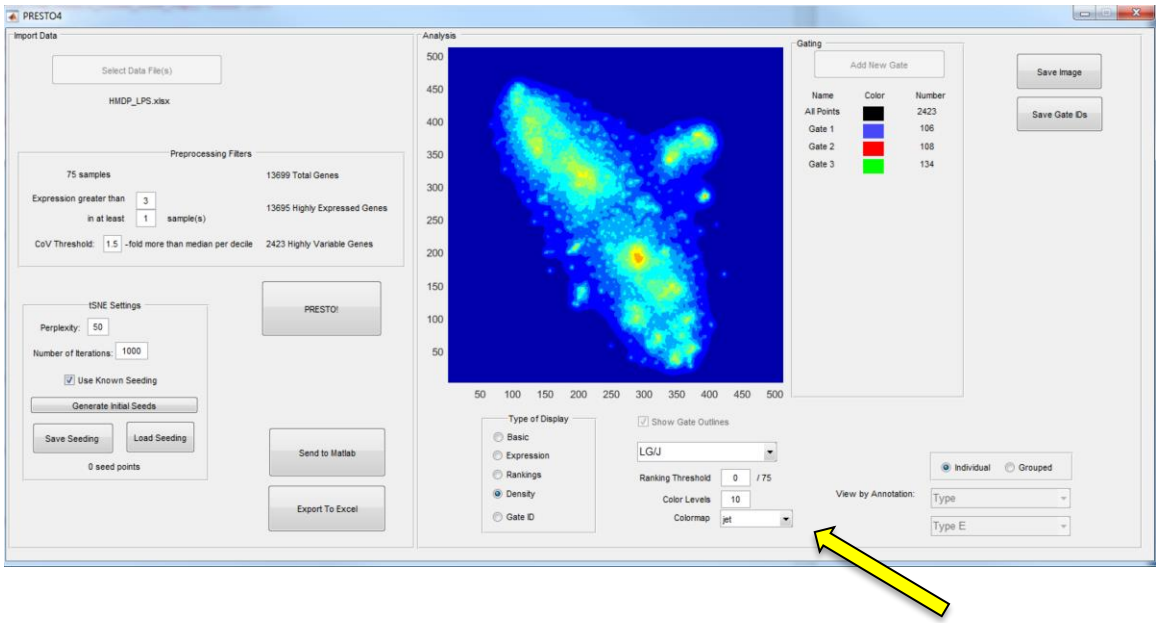
c. Furthermore, one can view the expression data grouped by annotation labels by clicking on “Grouped”. The expression values are averaged for all samples labeled with that annotation. If you click on “All”, the expression values are averaged across all genes.

d. At this point, the colors can still be changed, gates rearranged or reshaped by dragging the corners

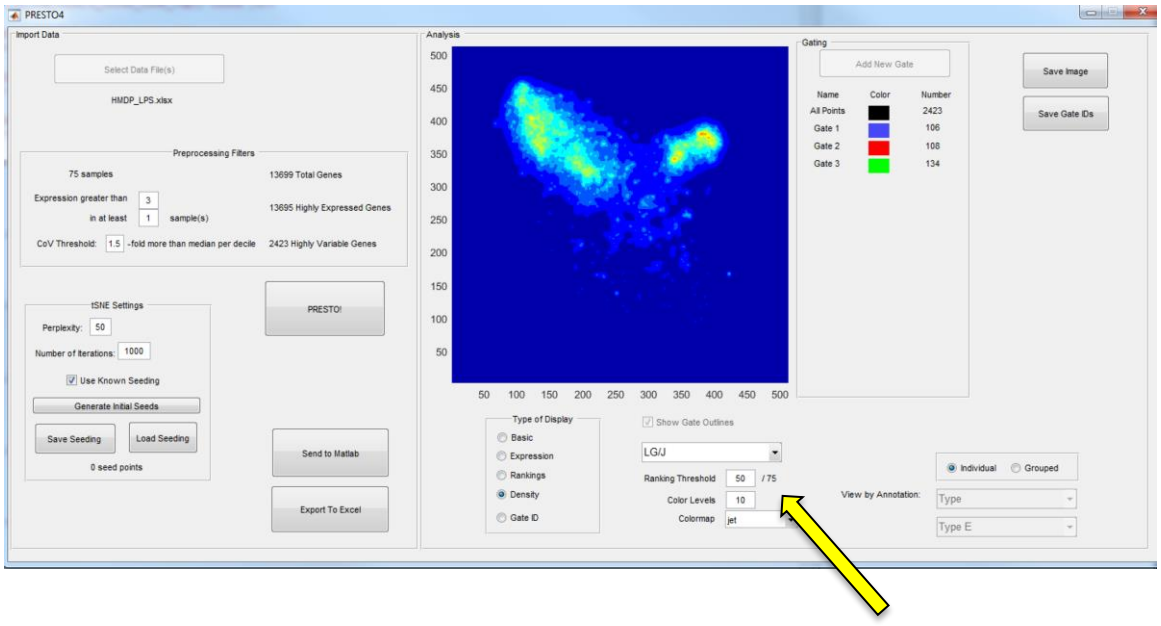
3. Same settings are available for Rankings as there are for Expression



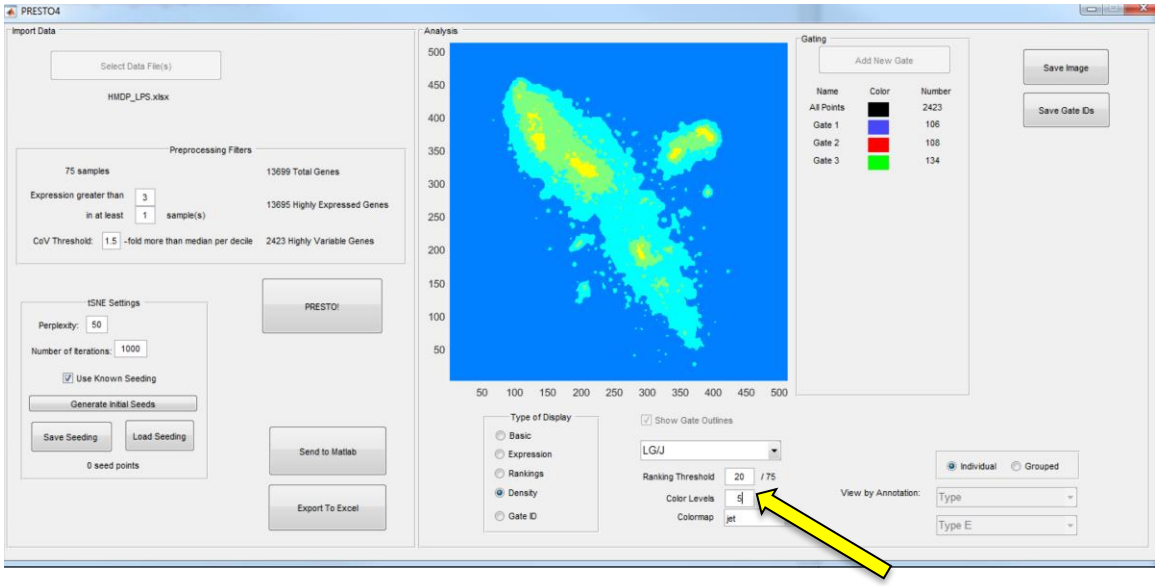
4. **Density**:
    a. When clicking on "Density" it will always take a second to load
    b. Same settings are available for Density as for Expression and Ranking
    c. Additional settings include the Ranking Threshold of the amount of samples and how many colors for the density should be shown



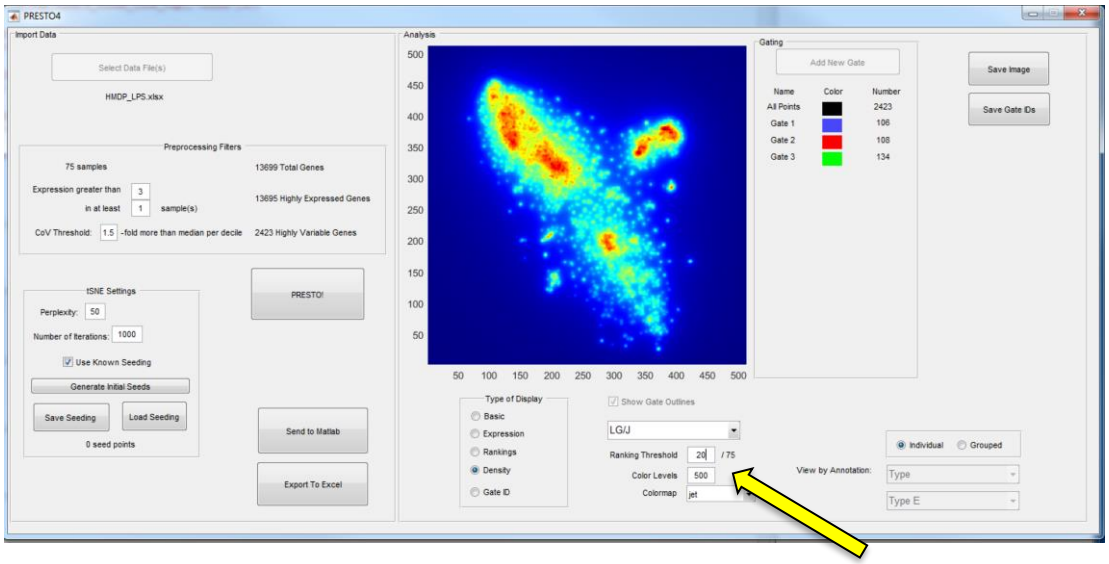
    d. Ranking Threshold changed to 50 for this sample. This removes all genes for which this sample (LG/J) is ranked in the bottom 50 (i.e., only shows those genes for which LG/J is ranked in the top 25).



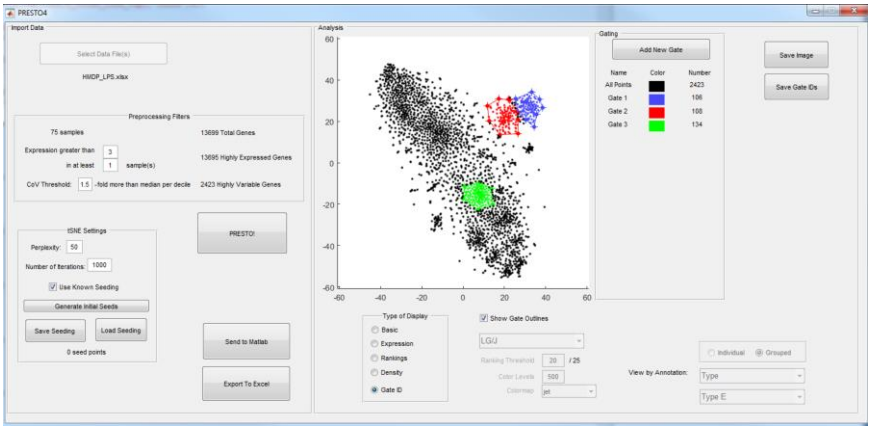
e. Ranking Threshold set for 20 and the Color Levels at 5



f. For the same Ranking Threshold as above, but the Color Levels was set to 500



5. **Gate ID**:
    a. Same settings as the "Basic" setting besides that the areas framed are filled in completely in the color assigned for it during the Basic step,



    b. You can change the color of all unassigned genes by right-clicking outside of a gate.

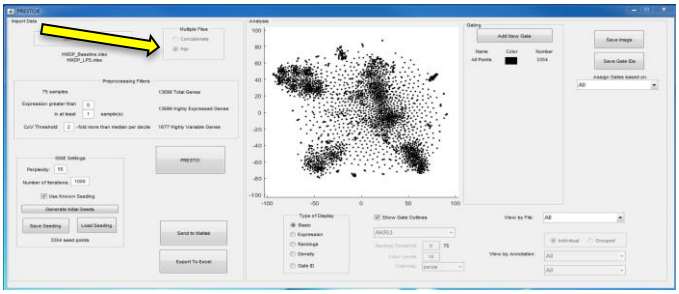
6. Starting over with the same file: All assigned gates need to be manually removed before starting over on for the Preprocessing Filters and tSNE setting steps and clicking PRESTO! again. If the gates will not be removed before hand, PRESTO will not work properly.

IV. Adding 2 or more files to PRESTO

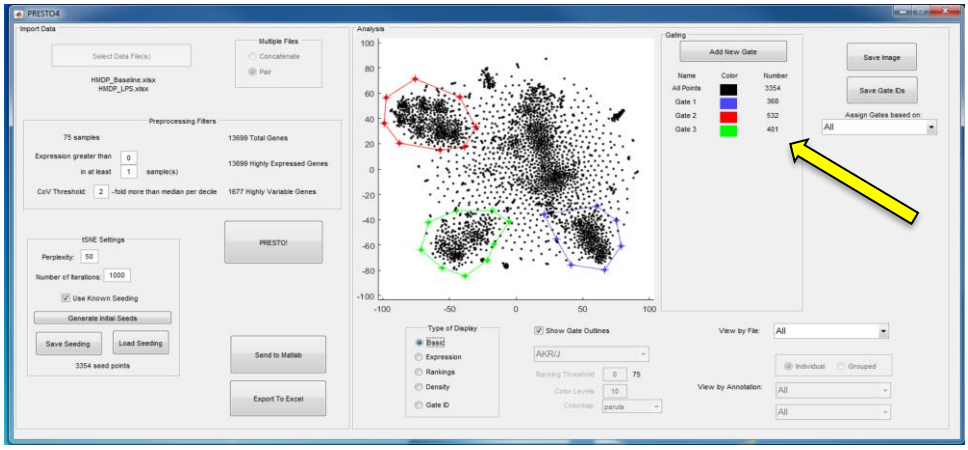
Be sure that Presto is started from the beginning. If Presto was used previously, close it out and restart it to be able to add files. Files need to be in the same folder and marked at the same time because the files can't be changed once loaded to Presto. If a mistake was made, Presto needs to be restarted and the correct files need to be loaded at the same time. Press Ctrl to select more than one file. Any number can be added (up to 3 have been tested).

a. Import of the Data

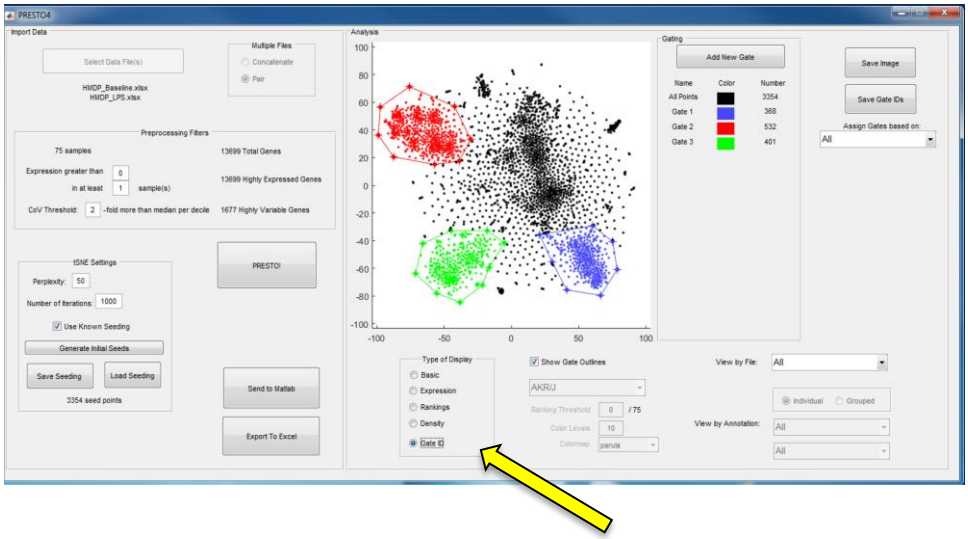
1. Once the files are loaded, either click on "Concatenate" or "Pair"
    a. "Concatenate" will combine the files via columns, so if 2 files have 75 samples each, Presto will then show 150 samples. This is the same as copying all the data into 1 file and adding an annotation row for the file name and is provided for convenience.
    b. "Pair" will combine the files via rows, so, if 2 files have each 13699 genes, it will be then
2. If "Pair" was assigned, the preprocessing filters are set and the tSNE settings are adjusted as needed, it is highly recommended to "generate initial seeds" as that will allow comparing sample sets during analysis later



3. The gates can be created by clicking “Add New Gate” just like it was mentioned during the part when only one file was added to PRESTO. The colors and the position can also be changed as mentioned previously.

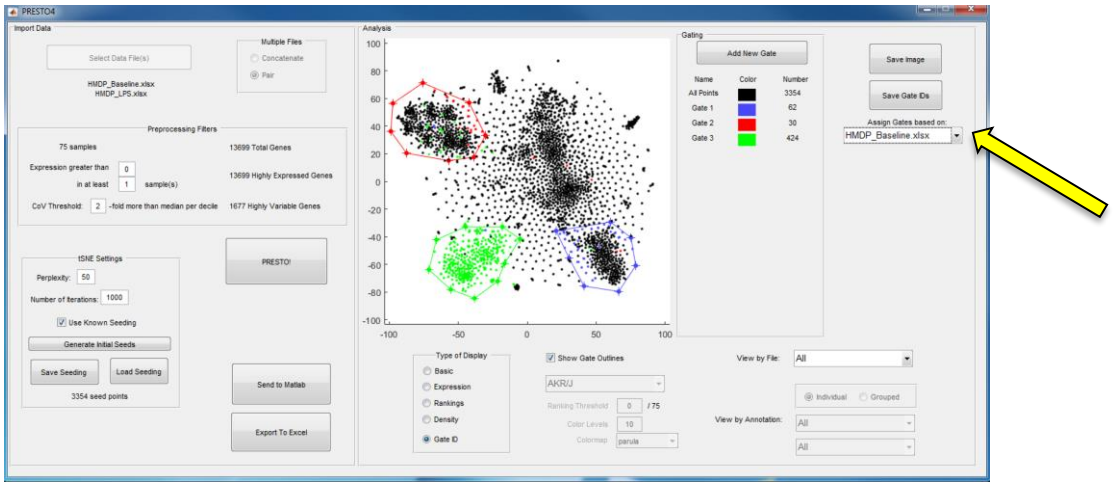


4. When changing the Type of Display to Gate ID, this example would look like seen below. The framed areas data points take on the color of the designated color for its specific gate.

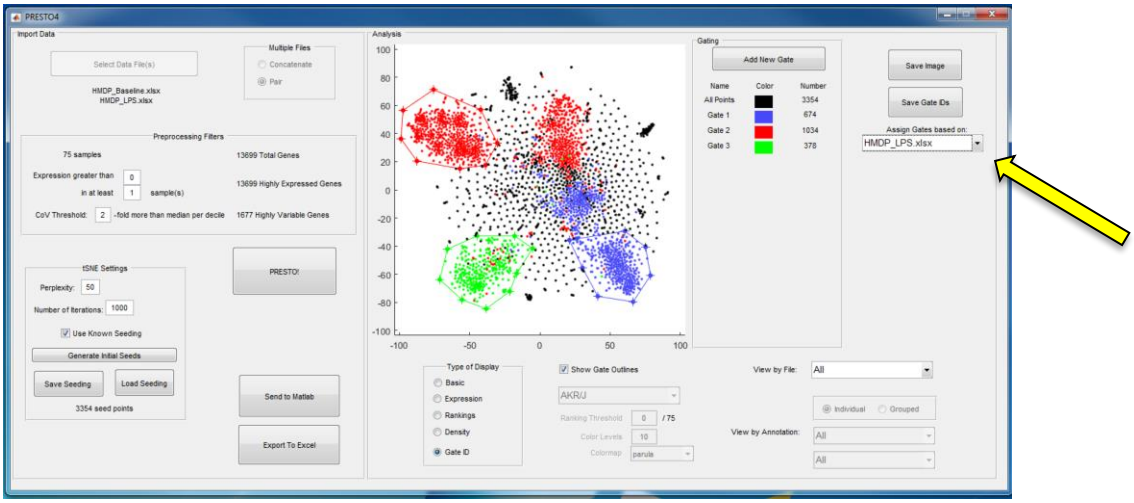


5. To see the data points from the assigned gates on the tSNE graph based on one of the added files, click the drop-down menu for “Assign Gates based on”

a. For this example, the tSNE graph is based on HMDP-Baseline

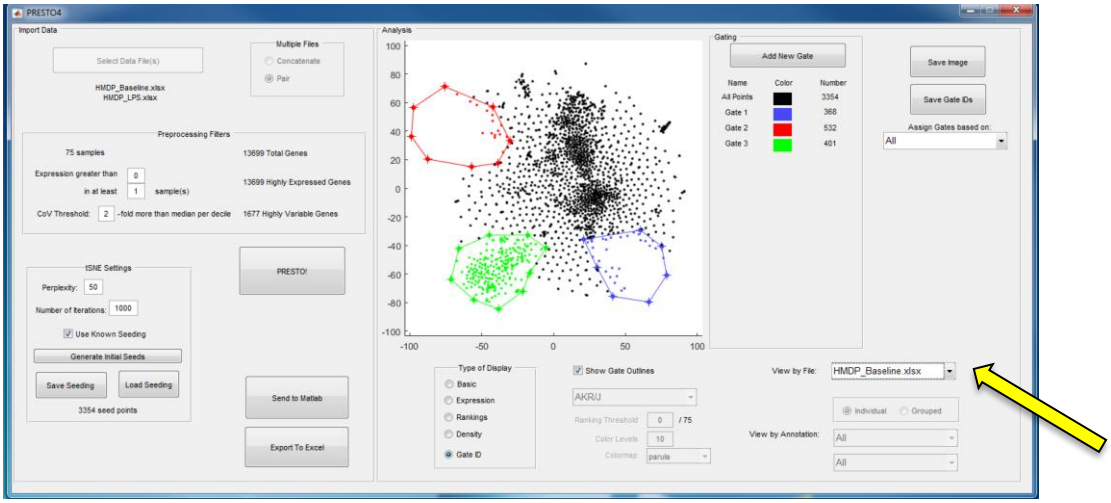


b. Whereas this picture shows the same data points when they are assigned to the HMDP_LPS file

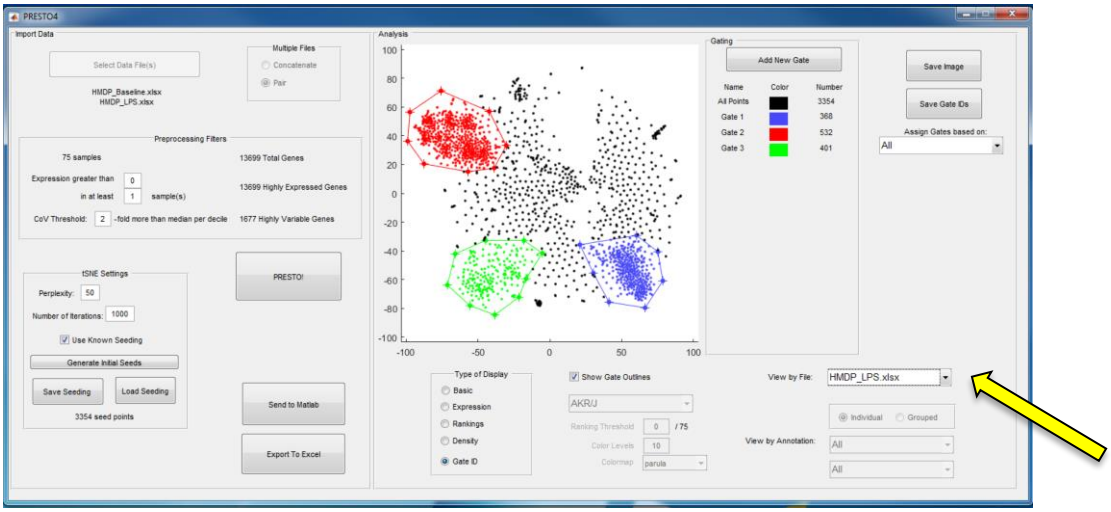


6. Furthermore, the assigned data points can be viewed by one of the added data files

a. Here one can see the data points only from the HMDP_Baseline file

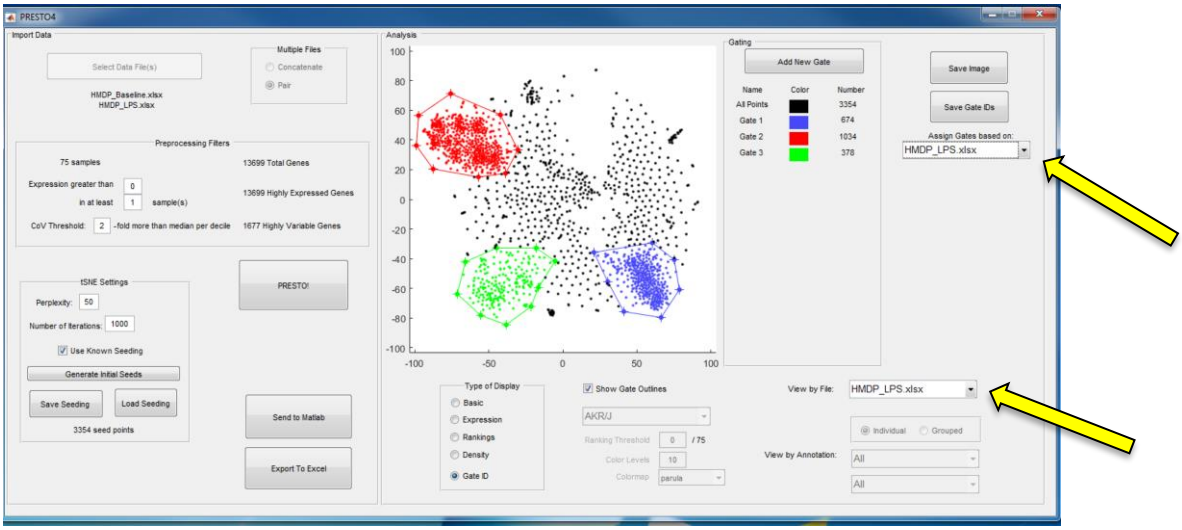


b. Below it shows the data points for the HMDP_LPS file only

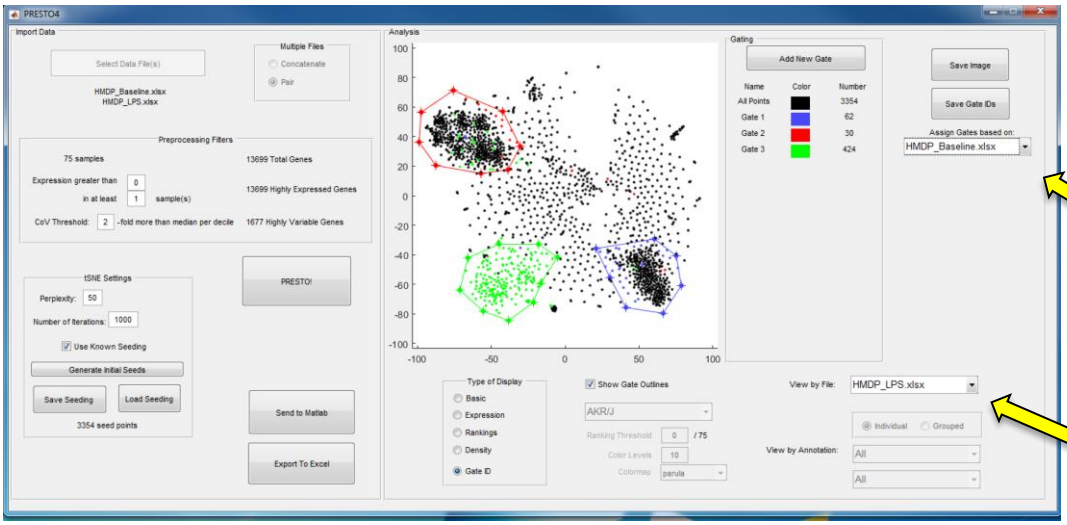


7. One can assign the gate data points to a specific file only, as well as base it on where the data points for those gates are within the file

a. Only HMDP_LPS shown as well as where the gated data points are within the HMDP_LPS file

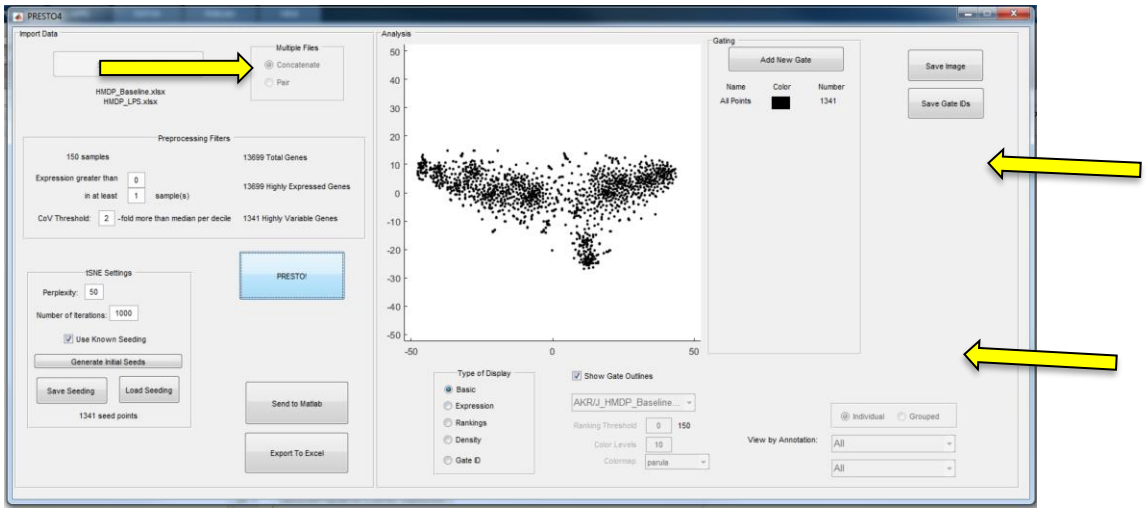


b. In comparison, below can be seen that the data points for HMDP_LPS are only shown, where also the gated data points for the HMDP_Baseline file are



8. The Expression, Rankings and Density settings can be adjusted the same way as mentioned previously when only one file was added to PRESTO

9. If “Concatenate” was assigned, the preprocessing filters are set and the tSNE settings are adjusted as needed, it is highly recommended to “generate initial seeds” as that will allow comparing sample sets during analysis later similar as mentioned for the “Pair” setting



10. All settings mentioned previously for the “Pair” setting also apply and work the same way when “Concatenate” is assigned for at least 2 files, beside that the gate points cannot be based or viewed by a specific file as those options do not exist

b. <u>What Excel files will look like</u>

One can save their analysis to excel at any time. Because everyone’s files will look slightly different, only a few examples with some basic information will be shown here.

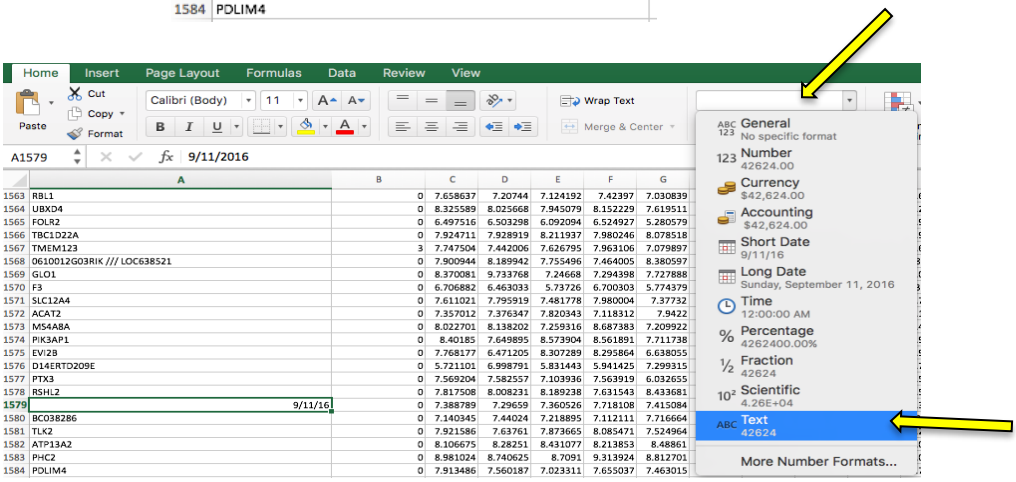
1. In all files, the names of the pre-processed genes (passing the expression and CoV filters) will be exported, along with their values for all samples. Any sample annotations will also be exported.

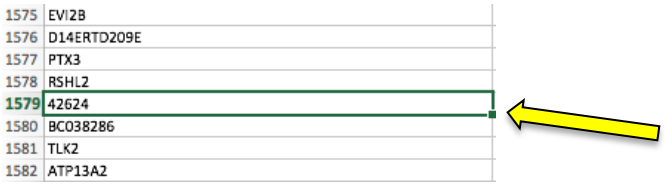
2. The column after the gene names will have the gate number assignments. Unassigned genes will be assigned a 0.

	Marker Name	Gate Number
2	Group	
3	Type	
4	NEK2	0
5	TNNT3	0
6	GLRP1	0
7	NEK6	0
8	KLK1B22 /// KLK1B9	0
9	PRM2	0
10	4922502D21RIK	0
11	SLC6A15	0
12	CYP2A4 /// CYP2A5	0
13	1700029I01RIK /// LOC433791 /// LOC666532 /// LOC671566	0
14	XLR /// 3830403N18RIK	0
15	5830443L24RIK /// LOC626578	0
16	MARVELD2	0
17	CYP39A1	0
18	CRISP2	0
19	RARB	0
20	LY6I	0
21	2610209A20RIK	0
22	LAMA1	0
23	COL4A3	0
24	1600002H07RIK	3
25	CD34	2
26	REEP6	0
27	THY1	2
28	HMBS /// LOC623818 /// LOC639157	0
29	AB041803	0
30	UBD	0
31	TMC2	0
32	BC027382	0
33	DIAPH3	2

3. By default, gene names that are only numbers will be converted by excel to dates. To get the correct gene name back, click on the gene that shows a date. Next, click on the tap where it says “Date” and change it to “Text”

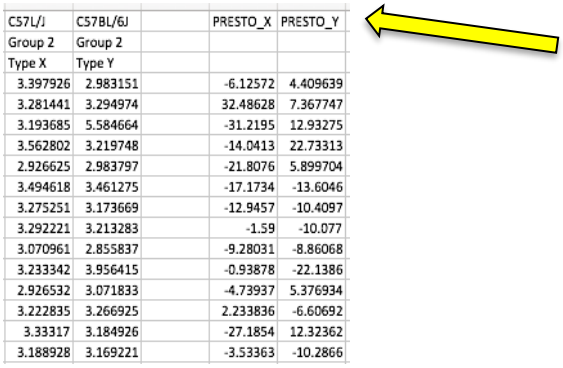
1573	MS4A8A
1574	PIK3AP1
1575	EVI2B
1576	D14ERTD209E
1577	PTX3
1578	RSHL2
1579	9/11/16
1580	BC038286
1581	TLK2
1582	ATP13A2
1583	PHC2
1584	PDLIM4





1575	EVI2B
1576	D14ERTD209E
1577	PTX3
1578	RSHL2
1579	42624
1580	BC038286
1581	TLK2
1582	ATP13A2

4. If PRESTO was performed, the excel output will also have additional columns in the back for the tSNE plots X and Y coordinates.



C57L/J	C57BL/6J		PRESTO_X	PRESTO_Y
Group 2	Group 2			
Type X	Type Y			
3.397926	2.983151		-6.12572	4.409639
3.281441	3.294974		32.48628	7.367747
3.193685	5.584664		-31.2195	12.93275
3.562802	3.219748		-14.0413	22.73313
2.926625	2.983797		-21.8076	5.899704
3.494618	3.461275		-17.1734	-13.6046
3.275251	3.173669		-12.9457	-10.4097
3.292221	3.213283		-1.59	-10.077
3.070961	2.855837		-9.28031	-8.86068
3.233342	3.956415		-0.93878	-22.1386
2.926532	3.071833		-4.73937	5.376934
3.222835	3.266925		2.233836	-6.60692
3.33317	3.184926		-27.1854	12.32362
3.188928	3.169221		-3.53363	-10.2866

5. If multiple files were concatenated, the sample names will be changed to “SampleName_FileName”.

6. If multiple files were paired, each file will be given its own section of the excel sheet, with the expression values, gate numbers, and tSNE coordinates listed independently.