NanoStriDE User's Guide

Chris Brumbaugh

cbrumbau@soe.ucsc.edu

http://nanostride.soe.ucsc.edu/

October 28, 2011

Abstract

This user guide provides a brief overview of how to use the NanoStriDE web tool. Detailed information covering all of the options and output is provided to the user, including guidelines for the results. Our NanoStriDE web tool takes NanoString counts data in directly from the machine and generates a heatmap of differentially expressed features based on the user's selection of of serveral methods of statistical analysis. For details of the statistical methods used in the NanoStriDE web tool, please refer to our paper (in preparation).

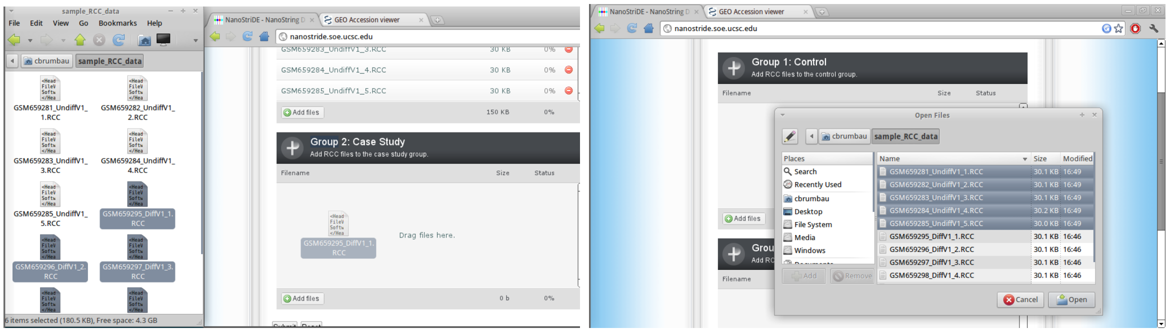
Contents

1 Getting Started

This section of the user's guide contains a brief walkthrough detailing the use of the NanoStriDE web tool, from uploading NanoString data to downloading the differential expression results from a completed job.

1.1 Uploading Files

All modern web browsers (e.g. IE 9, Firefox 7, Chrome 14, Safari 5) are compatible with the NanoStriDE web tool. Before uploading NanoString files to the NanoStriDE website, the user must first fill in the required e-mail address field, choose how many groups the data is divided into and press "Select".



GSM659281_UndiffV1_1.RCC
GSM659282_UndiffV1_2.RCC
GSM659283_UndiffV1_3.RCC
GSM659284_UndiffV1_4.RCC
GSM659285_UndiffV1_5.RCC
GSM659295_DiffV1_1.RCC
GSM659296_DiffV1_2.RCC
GSM659297_DiffV1_3.RCC

6 items selected (180.5 KB), Free space: 4.3 GB

GSM659283_UndiffV1_3.RCC	30 KB	0%
GSM659284_UndiffV1_4.RCC	30 KB	0%
GSM659285_UndiffV1_5.RCC	30 KB	0%
Add files	150 KB	0%

Group 2: Case Study
Add RCC files to the case study group.

Filename	Size	Status
Add files	0 b	0%

Drag files here.

(a) Drag & drop selection

Group 1: Control
Add RCC files to the control group.

Filename	Size	Status

Name	Size	Modified
GSM659281_UndiffV1_1.RCC	30.1 KB	16:49
GSM659282_UndiffV1_2.RCC	30.1 KB	16:49
GSM659283_UndiffV1_3.RCC	30.1 KB	16:49
GSM659284_UndiffV1_4.RCC	30.2 KB	16:49
GSM659285_UndiffV1_5.RCC	30.0 KB	16:49
GSM659295_DiffV1_1.RCC	30.1 KB	16:46
GSM659296_DiffV1_2.RCC	30.1 KB	16:46
GSM659297_DiffV1_3.RCC	30.1 KB	16:46
GSM659298_DiffV1_4.RCC	30.1 KB	16:46

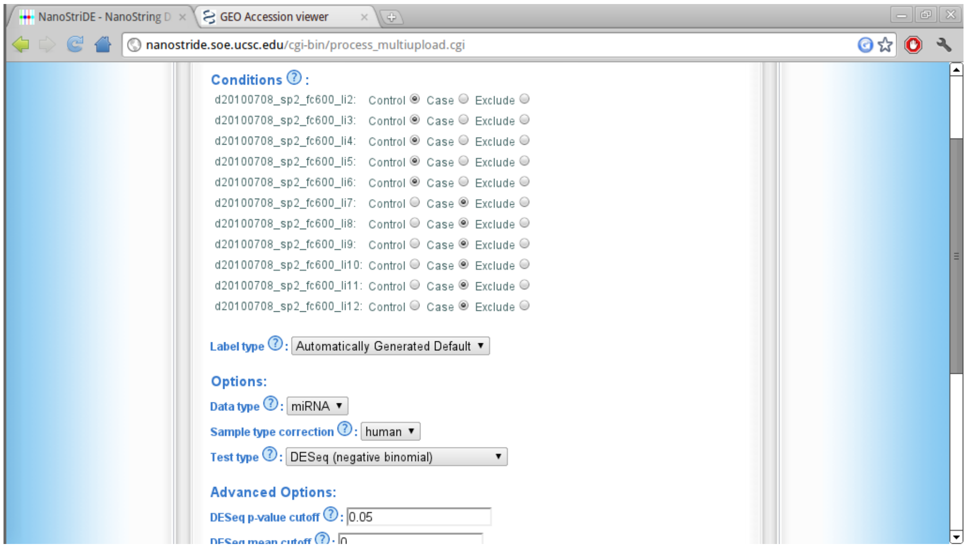
(b) Selecting multiple files in dialog

Figure 1: Illustrating uploading multiple NanoString RCC files using the NanoStriDE web tool.

The user may upload NanoString RCC files either by dragging and dropping files from the desktop and/or file manager (Figure 1(a)) or clicking "Add files" and using the Ctrl or Shift key to select multiple files to upload (Figure 1(b)). Once all of the desired files have been added to the upload queue, click 'Submit' to proceed with the upload or "Reset" to cancel.

1.2 Selecting Options

Once the upload has complete, an options page customized based on the number of files chosen is presented to the user. Warnings presented to the user are displayed at the top of the page in red text.



nanostride.soe.ucsc.edu/cgi-bin/process_multiupload.cgi

Conditions ?:

d20100708_sp2_fc600_li2: Control ● Case ○ Exclude ○
d20100708_sp2_fc600_li3: Control ● Case ○ Exclude ○
d20100708_sp2_fc600_li4: Control ● Case ○ Exclude ○
d20100708_sp2_fc600_li5: Control ● Case ○ Exclude ○
d20100708_sp2_fc600_li6: Control ● Case ○ Exclude ○
d20100708_sp2_fc600_li7: Control ○ Case ● Exclude ○
d20100708_sp2_fc600_li8: Control ○ Case ● Exclude ○
d20100708_sp2_fc600_li9: Control ○ Case ● Exclude ○
d20100708_sp2_fc600_li10: Control ○ Case ● Exclude ○
d20100708_sp2_fc600_li11: Control ○ Case ● Exclude ○
d20100708_sp2_fc600_li12: Control ○ Case ● Exclude ○

Label type ?: Automatically Generated Default ▾

Options:

Data type ?: miRNA ▾
Sample type correction ?: human ▾
Test type ?: DESeq (negative binomial) ▾

Advanced Options:

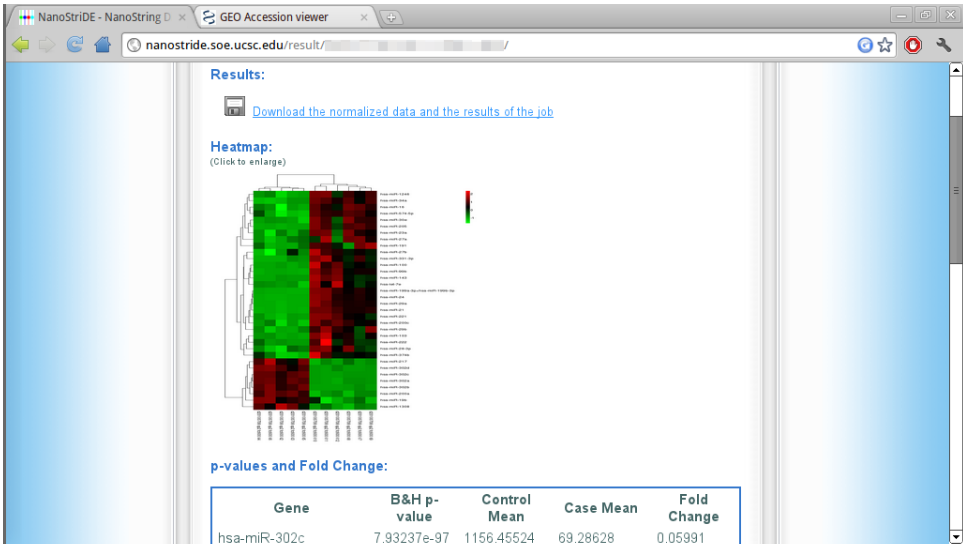
DESeq p-value cutoff ?: 0.05
DESeq mean cutoff ?: 0

Figure 2: Illustrating the NanoStriDE web tool options screen.

Here, a user can change the the classifications of the samples (the default is the groups that were selected at the time of uploading the data), as well as set the options for the differential expression analysis (Figure 2). For a description of any option, hover the mouse cursor over the question mark indicators next to that option. When the options are set, the "Submit" button processes the job and "Cancel" deletes the unsubmitted job and returns to the upload page.

1.3 Viewing Results

After the job is submitted to the NanoStriDE web server, the user is presented with a status page that updates every thirty seconds, until the job is completed. When the job completes, a confirmation e-mail is sent to the e-mail address provided during the initial upload step and a results page is presented to the user.


NanoStriDE - NanoString D
GEO Accession viewer
nanostride.soe.ucsc.edu/result/[redacted]/

Results:

Download the normalized data and the results of the job

Heatmap:
(Click to enlarge)

p-values and Fold Change:

Gene	B&H p-value	Control Mean	Case Mean	Fold Change
hsa-miR-302c	7.93237e-97	1156.45524	69.28628	0.05991

Figure 3: Illustrating a sucessful NanoStriDE run results screen.

On the results page, if the differential expression of more than one probe is determined to be statistically significant, the heatmap along with a table of the statistically significant genes containing the p-values and the base means for all the conditions is displayed with a download link to a zip-compressed archive storing all of the results (Figure 3). Clicking the heatmap on the results page expands it to the full (actual) size of the heatmap image. If there were insufficient statistically significant genes or there was an error when running the chosen statistical test, only the download link to the results is presented to the user and a warning indicating that not enough statistically significant genes were found to create a heatmap.

2 Using NanoStriDE

This section of the user's guide details all of the options that are available to the user and any warning messages that may appear when running or finishing a job.

2.1 Options

2.1.1 Selecting Conditions

Under the conditions section of the otpions page, the user is allowed to change which samples in the uploaded data set belong to which condition. In addition to the conditions that are available, there is also an "Exclude" selection available for each sample. When any sample is set to this condition, it is removed from the subsequent normalization steps and differential expression analysis that is performed on the data set. This option is available to the user so that if on the results screen the user has noticed that if one sample appears to be a poor sample (e.g. due to warning messages provided, for example), the user can return to the options page by going back in the browser, exclude that one sample from the data set, and rerun the analysis again without that sample.

2.1.2 Data Type

Depending on the data type of the NanoString samples that are used (either microRNA or mRNA), different methods of sample content normalization may be appropriate. This setting sets the default for the sample content normalization to be used, which still can be changed at user discretion.

2.1.3 Statistical Test Type

A choice of four statistical tests are available for use in differential expression analysis. Selecting the correct statistical test option depends on the assumed underlying background distribution of the data and the number of conditions to be used for the analysis. If a normal background distribution is assumed, the t-test is appropriate to be used for two conditions and the one-way ANOVA is appropriate for three or more conditions. If a negative binomial distribution is assumed, which is a better assumption for NanoString count data, DESeq is appropriate to be used for two conditions and the one-way ANOVA (negative binomial) that uses DESeq's ANODEV implementation is appropriate for three or more conditions.

2.1.4 Negative Correction

If the t-test or one-way ANOVA is used, the negative correction is applied to the positive corrected data, as per NanoString guidelines. There are several options available for the type of negative correction to use. The first three options are different thresholds with which to perform the subtractive normalization: the mean of a given gene, the mean added to 2 standard deviations, and the maximum value of the negative controls. The mean added to 2 standard deviations is the suggested default, but the mean can be used for a less stringent value and the maximum of the negative controls can be used for a more stringent value to eliminate background noise. The fourth option, the one tailed Student's t-test, differs from the previous three options. For each probe in the probe set, a one tailed Student's t-test is performed on each probe. If the resulting p-value is below the p-value cutoff (default 0.05), the subtractive normalization using the mean is applied, otherwise all counts are set to zero.

2.1.5 Sample Content Normalization

If the t-test or one-way ANOVA is used, a sample content normalization is applied to ensure that transcript quantity levels are comparable across samples by normalizing counts to the total number of transcripts seen. The first option, recommended for mRNA data, normalizes each sample to housekeeping genes present in the data. The second option, recommended for microRNA data, normalizes each sample to the entire microRNA data. The thrid option rank sorts the first sample and selects the top 75 probes present used to normalize the data to, as opposed to all probes for the previous option. The default values are recommended, but the third option is available at user discretion.

2.1.6 p-value and Mean Cutoffs

The p-value and mean count cutoffs can be adjusted for the differential expression analysis. The default p-value cutoff is 0.05 and can be adjusted to other values as desired (e.g. 0.01) to filter for the heatmap. The mean cutoff applies a minimum mean count cutoff to normalized count values and be default is inactive as it is set to 0. If the mean cutoff is set to any non-zero positive value, any probe where the mean counts is below the cutoff is excluded from the heatmap.

2.1.7 p-values for Heatmap

The p-values used for the heatmap can be unadjusted p-values or adjusted p-values. Adjusted p-values are p-values that have been adapted for multiple testing to account for the increased risk of Type I errors (false positive errors) that can arise from comparing the unadjusted p-values with each other. The p-value adjustment can be set to different methods as per the “Adjusted p-value type” option, though the default Benjamini & Hochberg is recommended.

2.1.8 Miscellaneous Options

The clustering of samples in the heatmap with hierarchical clustering using Euclidean distance can be toggled with the “Cluster samples in heatmap” option. The key/legend for the heatmap displaying the values for the colors in the heatmap can be toggled with the “Display key in heatmap” option. These values are centered around zero as the values displayed in the heatmap have been adjusted on a log scale to allow for better contrast in the resulting heatmap. The colors used to display the heatmap can be adjusted to standard presets using the “Heatmap colors” option. The normalization output as well as the results of the statistical test format can also be changed between the comma serparated value format (.csv) and tab-delimited files (.tab) with the “Output format” option.

2.2 Warnings

2.2.1 Issues with Raw Data

- FOVCounted to FOVCount ratio should be greater than 80%.

  The ratio of the number of barcodes counted in the field of view (FOV) to the expected number of barcodes in the FOV should be above 80% indicating that enough of the sample was successfully identified.

- Binding density should be between 0.05 and 2.25.

  Binding density below 0.05 indicates that not enough sample has bound to the platform and binding density above 2.25 indicates that too much sample has bound for the data to be reliably interpreted.

2.2.2 Issues with Normalization

- Positive control normalization should be between 0.3 and 3.

  If the positive control normalization factor is below 0.3, this indicates that the counts for the positive controls in that sample are far higher than for all other samples in the data set. If the positive control normalization factor is above 3.0, this indicates that the positive controls in that sample are far lower than for all other samples in the data set.

- 0.5fM control counts should be above average of negative controls in 90% of lanes.

  The 0.5fM positive control counts should be above the average of the negative control for most of the lanes, as a percentage lower than this indicates that the signal-to-noise ratio to too low for the data to be reliably interpreted.

- Linear correlation of positive controls vs. concentration should have R^2 greater than 0.95 in at least 90% of lanes.

  For most of the lanes, the count values for the positive controls should mostly be linear with respect to the known concentration values for the positive controls, as indicated by the coefficient of determination (R^2) of 0.95 or higher.

2.2.3 Issues with Differential Expression Analysis

- No/One statistically significant probes/probe identified, heatmap cannot be generated.

  Fewer than two statistically significant genes were found so a heatmap could not be generated.

3 License

The NanoStriDE web tool is free for academic use and the code is available upon request.