MSstatsQC v.1.1 User Manual

Eralp DOGU, Sara TAHERI, Olga VITEK

October 10, 2016

Contents

1 MSstatsQC: Longitudinal system suitability monitoring for targeted proteomic experiments

The increasing need for defining good quality measurement and analyzing traceable quality metrics has been guiding proteomics society to focus more on system suitability analysis and discovering quantitative tools/protocols (P. A. Rudnick et al. 2009; Abbatiello et al. 2013; Abbatiello et al. 2015). Moreover, systematic longitudinal system suitability monitoring approaches are desirable to evaluate critical-to-quality measures and present most informative QC metrics over time (Ma et al. 2012; R. M. Taylor et al. 2013; Pichler et al. 2012; Bereman et al. 2014; Bereman et al. 2016). `MSstatsQC` is an open-source R-based web application for statistical analysis and monitoring of quality control (QC) and system suitability testing (SST) samples produced by spectrometry-based proteomic experiments. This document describes `MSstatsQC`, the most recent version of the application, and its use through the user interface. `MSstatsQC` uses SPC tools to track ID free system suitability metrics including total peak area, retention time, full width at half maximum (FWHM) and peak asymmetry for selection reaction monitoring (SRM) based proteomic experiments.

1.1 Applicability

`MSstatsQC` v1.1 and above is applicable to system suitability data produced from Selected Reaction Monitoring (SRM) based proteomic experiments. General framework of `MSstatsQC`is shown below.

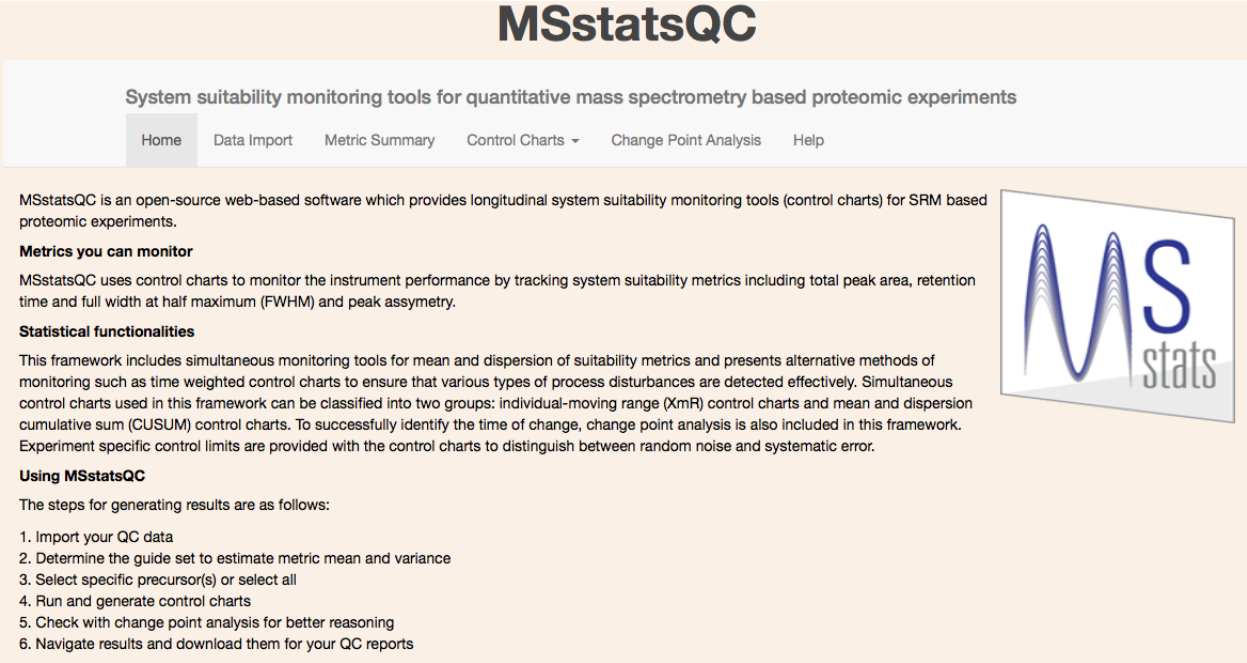

Figure 1: MSstatsQC General Framework

1.2 Statistical functionalities

Statistical process control (SPC) is a general and well-established method of quality control (QC) which can be used to monitor and improve the quality of a process such as LC MS/MS. We introduce simultaneous and time weighted monitoring tools and change point analysis to monitor mean and dispersion of system suitability metrics such as retention time. Proposed longitudinal monitoring approach significantly improves the ability of real time monitoring, early detection and prevention of chromatographic and instrumental problems of mass spectrometric assays, thereby, reducing cost of control and failure.

Simultaneous control charts used in this framework can be classified into two groups: *individual-moving range (XmR)* control charts and *mean and dispersion cumulative sum (CUSUM)* control charts. Experiment specific control limits are provided with the control charts to distinguish between random noise and systematic error. The QC or SST sample at which a signal is issued is considered as an evidence of nonrandom process behaviour and treated as an *out-of-control* observation. After this signal, process professionals start searching for assignable cause(s). However, the signal does not always designate that the special cause actually occurred at that certain acquire time. A remedy to this problem is to use follow-up *change point analysis* along with control charts. Change point estimation procedures have a potential to save time by narrowing the search window for special causes. In this framework, we introduce two change point models: *step shift change model for mean* and *step shift change model for variance*.

1.3 Interoperability with existing computational tools

`MSstatsQC` takes as input QC data in a tabular .csv format (Figure 1), which can be generated by any spectral processing tool. `MSstatsQC` v1.1 and above is available as an external tool and compatible with

Skyline (MacLean et al. 2010) and PanoramaWeb (Sharma et al. 2014) reports.

1.4 Availability

`MSstatsQC` is available under the Artistic-2.0 license at msstats.org/msstatsqc. We suggest to use that version if possible. The versioning of the main application is updated several times a year, to synchronise with the most recent developments. Source code can is also available through our github page (https://github.com/srtaheri/msstats-qc).

1.5 Troubleshooting

To help troubleshoot potential problems with installation or functionalities of `MSstatsQC`, a progress report is generated in a separate log file *msstatsqc.log*. The file includes information on the R session (R version, loaded software libraries), options selected by the user, checks of successful completion of intermediate analysis steps, and warning messages. If the analysis produces an error, the file contains suggestions for possible reasons for the errors. If a file with this name already exists in working directory, a suffix with a number will be appended to the file name. In this way a record of all the analyses is kept.

2 Allowable data format: 8-column format

`MSstatsQC` performs statistical analysis, to monitor system performance by tracking system suitability metrics including total peak area, retention time reproducibility, full width at half maximum (FWHM) and peak asymmetry. Therefore, input to `MSstatsQC` is the output of other software tools (such as Skyline or MultiQuant) that read raw spectral files and report system suitability metrics. The preferred structure of data for use in `MSstatsQC` is a .csv file in a "long" format with 8 columns representing the following variables: `AcquiredTime`, `PrecursorName`, `BestRetentionTime`, `TotalArea`, `MaxFWHM`, `MaxEndTime`, `MinStartTime`, and `Annotations`. The variable names are fixed, but are case-insensitive. If the user wants to use a metric which is not included in this list, he/she can parse new columns to the raw file after `Annotations` column and then `MSstatsQC` generates results for these new metrics. This required input data is generated automatically if the report format is defined or SProCop format is used in Skyline.

(a) `AcquiredTime`: This column shows the acquired time of the QC/SST sample in the format of MM/DD/YYYY HH:MM:SS AM/PM

(b) `PrecursorName`: This column shows information about Precursor id. Statistical analysis will be done separately for each unique label in this column.

(c)-(f) `BestRetentionTime`, `TotalArea`, `MaxFWHM`, `MaxEndTime`, and `MinStartTime`: The combination of these 5 columns defines a *feature* of a peak for a specific peptide. If the information for one or several of these columns is not available, please do not discard these columns but use a single fixed value across the entire dataset. For example, if the original raw data does not contain the information of `TotalArea`, assign the value NaN to the entries in the column `TotalArea` for the entire dataset. Please note that `MSstatQC`v1.1 does not currently provide plots for metrics with missing values.

(g) `Annotations`: Annotations are free-text information given by the analyst about each QC run. They can be informative explanations of any special cause or any observations related to a particular QC run. Annotations are carried in the plots provided by `MSstatsQC` interactively.

An example of an acceptable input dataset is shown below. The system suitability dataset is generated during the CPTAC Study 9.1. The dataset is stored in a .csv file in a "long" format. Each row corresponds to a single testing sample.

	A	B	C	D	E	F	G	H
1	AcquiredTime	Precursor	BestRetentio	MaxFWHM	MinStartTim	MaxEndTime	TotalArea	Annotations
234	9/19/11 23:48	FFVAPFPEVF	47.11	0.35	46.7	47.65	75423552	RT problems
235	9/20/11 1:19	FFVAPFPEVF	47.05	0.35	46.67	47.62	73847400	RT problems
236	9/20/11 16:26	FFVAPFPEVF	46.96	0.25	46.53	47.42	47423436	RT problems
237	9/20/11 17:57	FFVAPFPEVF	46.88	0.34	46.44	47.48	69410504	RT problems
238	9/20/11 19:28	FFVAPFPEVF	46.99	0.36	46.65	47.62	70255936	RT problems
239	9/21/11 3:02	FFVAPFPEVF	47.02	0.35	46.65	47.57	80757984	RT problems
240	9/21/11 15:07	FFVAPFPEVF	46.93	0.37	46.5	47.51	47260480	RT problems
241	9/21/11 16:37	FFVAPFPEVF	46.93	0.35	46.47	47.42	39373060	RT problems
242	9/21/11 18:08	FFVAPFPEVF	46.99	0.32	46.59	47.48	32859368	RT problems
243	9/22/11 5:37	FFVAPFPEVF	46.93	0.12	46.73	47.54	43599904	RT problems
244	9/22/11 15:37	FFVAPFPEVF	46.96	0.34	46.62	47.51	73333680	RT problems
245	9/23/11 1:35	FFVAPFPEVF	47.25	0.32	46.9	47.8	76762160	RT problems

Figure 2: MSstatsQC Data Format

2.1 Uploading QC Data and 'Data Import' Tab

A dataset which is in the allowable data format is uploaded via `Data Import` tab. Please follow the steps below to upload your data.

(a) Click `Choose file` and locate your file
(b) Select and upload the file you want to analyze

`MSstatsQC` uses a data validation method where slight variations in column names are compansated and converted to the standard `MSstatsQC` format. For example, our data validation function converts column names like `Best.RT`, `best retention time`, `retention time`, `rt` and `best ret` into `BestRetentionTime`. This conversion also deals with case-sensitive typing.

2.1.1 Chosing a Guide Set

Generally, a data gathering and parameter estimation step is applied to characterize in-control parameters of a given suitability metric for a specific peptide. Within that phase, control limits are obtained to test the hypothesis of statistical control. These thresholds are selected to ensure a specified type I error rate. Constructing control charts and real time evaluation are considered after achieving this phase. Along with the implementation, the analyst should follow signals given by the control charts. Each signal and non-random pattern should be examined carefully to identify the special causes of variation in mean and dispersion of a metric. Control charts in this framework are designed with the assumption of data availability to estimate the process parameters. Therefore, control limits are assumed to be available before on-line control begins.

Please select a proper and representative guide set using `Data Import` tab. The lower bound of guide set indicates the index of the first QC sample to be included in the guide set. For example, if you choose "1" as a lower bound, it means that first QC sample will be the first element of the guide set. Similarly, upper bound of guide set shows the index for the last observation. It is possible to use different guide sets for different suitability metrics and precursors.

After choosing a guide set, the user can select the precursor of interest or select all to generate a k by 2 matrix of control charts where k is the number of precursors. Mean and dispersion control chart are generated after selecting the options.

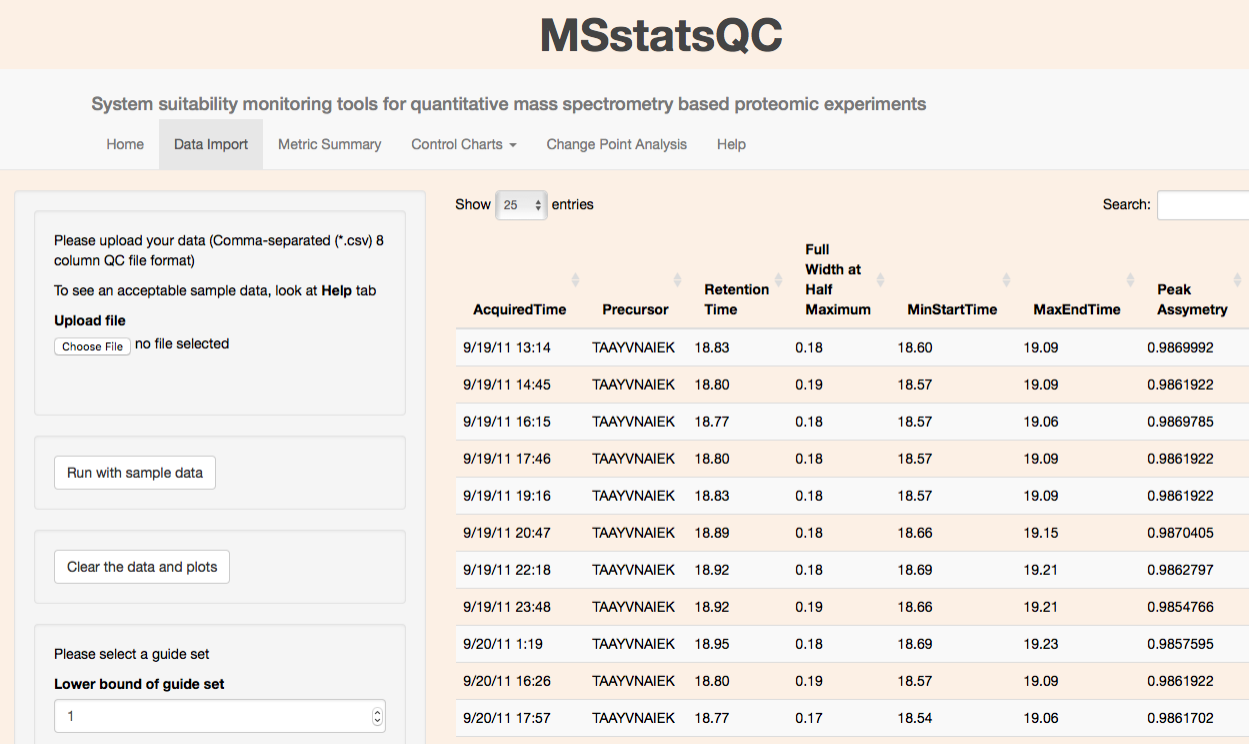

Figure 3: MSstatsQC Data Import Tab

3 'Metric Summary' Panel

The aim of the metric summary panel is to summarize results and provide a general visual summary of related results.

When analyst monitors multiple peptides, a large number of control charts need to be analyzed. For example, if 15 peptides are monitored and XmR charts are used, 30 control charts for XmR and 30 plots for change point analysis are produced. In this case, decision making becomes pretty difficult and we recommend using our summary plots for a better understanding about the problems.

Overall summary plot accumulates information using percentage of out of control peptides among all peptides monitored. Here, both increases and decreases in the mean and dispersion of a certain metric are summarized. Suppose we use an X chart, increases in the mean level of a suitability metric causes plotted points exceed the upper control limit. We count the number of observations exceeding the upper threshold and divide it to the total number of precursors for the ith QC sample. Then we plot proportions versus QC number and use a smoothing function to draw the line (orange) and confidence intervals. Similarly, another line plot is created for the peptides having observation below the lower control limit using X chart results. This line (blue) reflects decreases in the mean level of the related suitability metric. Positive and negative CUSUM statistics are similarly used to create an overall summary plot and distinguish between increases and decreases in metric mean. Overall summary plots have upper and lower part. Upper part summarizes the result for metric mean (X chart and CUSUMm charts) and lower parts summarizes the results for metric dispersion (mR and CUSUMv charts). An increasing pattern means that the problem starts to develop. Changes in metric mean and metric dispersion are plotted separately using different colors. Red lines in the plots of likelihood functions are summarized as red dots in overall summary plots. Change point estimates for mean and dispersion are plotted separately in the corresponding plotting field.

Additionally, radar charts namely precursor level summary plots are created to extract the overall contribution of each peptide. These plots help analyst distinguish the most contributing peptides for each suitability

metric separately. For example, if total peak area problems are partially observed, then we expect a higher number of out of control signals in some of the peptides marked on this plot. Panel for total peak area provides a nice example for total peak area decrease in early eluting peptides. Same color palette is used in radar plots to summarize metric mean and dispersion changes.

The boxplot tab shows boxplots for each metric. The user can investigate these charts to see if any abrubt observations are present in the dataset. If so, we recommend a preprocessing for the dataset and re-uploading it for better results. Metric summary panel also provides scatter plot matrices for each metric to show interrelations among the peptides for a specific metric.

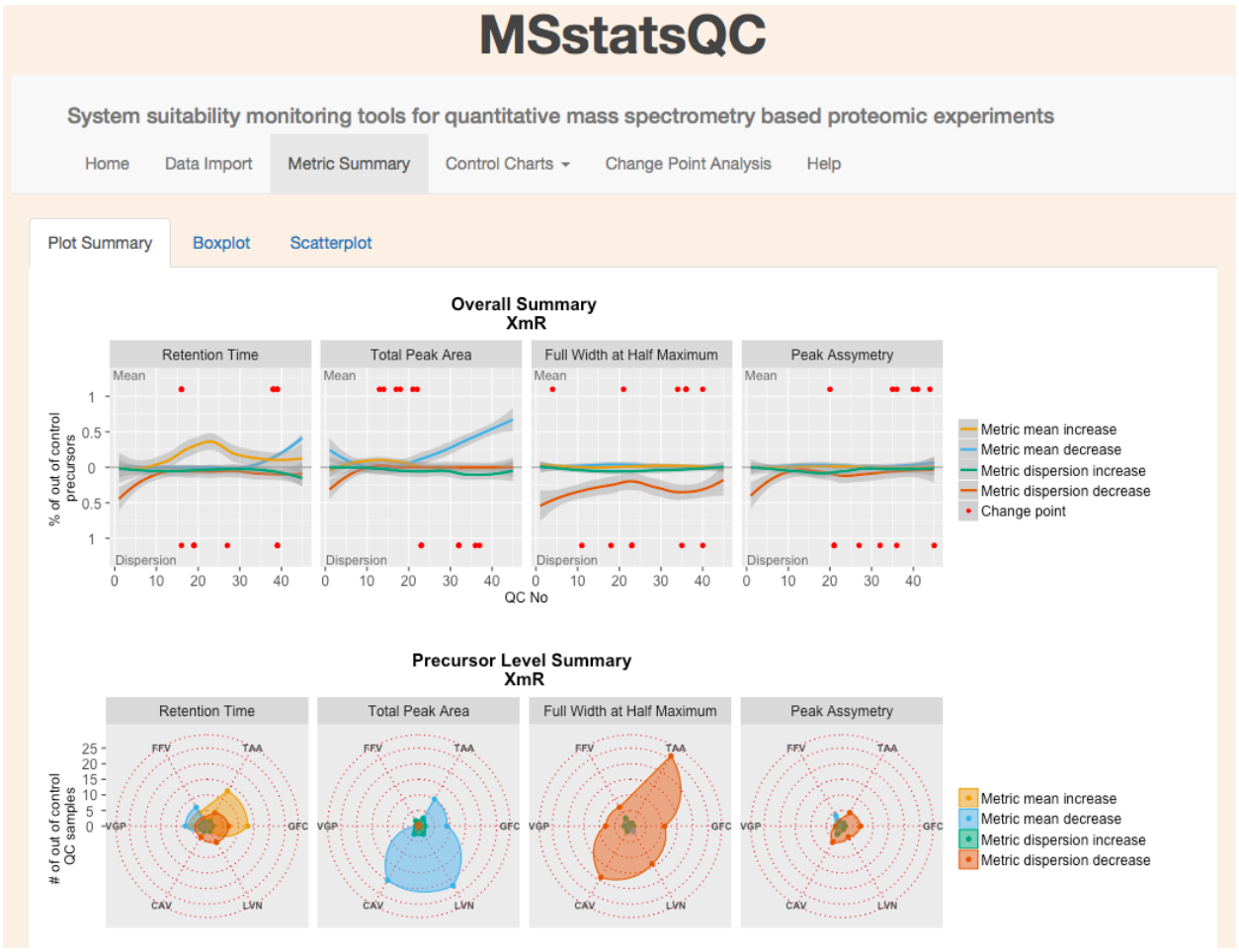

Figure 4: Metric Summary

4 'Control Charts' Panel

All control charts are generated in this tab. The drop-down menu shows the alternative control charts. `XmR` and `CUSUM` charts are available options for `MSstatsQC` v1.0. If you select `XmR` or `CUSUM`, then you will obtain a mean (right hand side) and a dispersion (left hand side) control chart for each peptide. Each control chart has limits shown in red and the relevant statistics are plotted accordingly. Any observation which exceeds the thresholds are considered as an `out-of-control` observation and shown in red. All plots are generated interactively. The user can move the cursor the the point of interest to see the original values and QC number of each observation. Additionally, the user can zoom in or out and save the plots to use in their reports.

4.1 'XmR' Control Charts

This tab shows `X` and `mR` control charts for each peptide. An example for Study 9.1 is presented in Figure 5. By using the sequential differences between two successive values as a measure of dispersion, a chart for individual observations (`X` chart) and a chart for moving ranges (`mR` chart) or `XmR` chart can be created. The original observations are plotted on a `X` chart along with upper and lower control limits. Here, design parameters are particularly chosen to provide a type I error rate of 0.0027 which guarantees the well-known 3σ limits. Moving ranges are plotted on a mR chart along with their corresponding upper and lower control limits. Any points above or below the control limits are classified as out-of-control observations and need special attention as they might provide valuable information about chromatographic and instrumental problems.

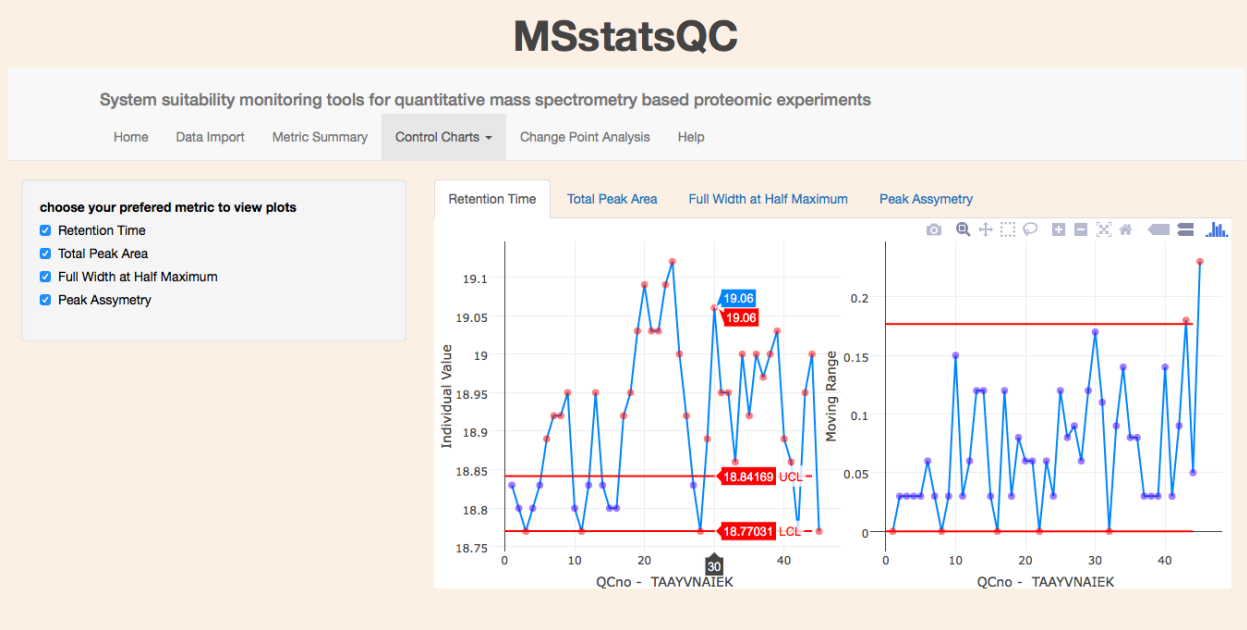

Figure 5: XmR Control Chart

4.2 'CUSUM' Control Charts

This tab shows `CUSUMm` and `CUSUMv` control charts for each precursor. An example for Study 9.1 is presented in Figure 6. Mean and dispersion CUSUM charts both have more complex design parameters when compared to `XmR` charts. However, they have proven ability to detect small shifts earlier. In order to simplify design complexity we consider standardized metrics. We use the parameters obtained from the guide set for standardization. `CUSUMm` essentially is a tabular CUSUM with the standardized QC observations and sensitive to changes in mean of a suitabilty metric. Basically, `CUSUMm` plots two types of CUSUM statistics; one for positive mean shifts and the other for negative mean shifts. Standardization enables informal benchmarking among different metrics and reduce design complexity into a considerably simple level. Similarly, it is possible to construct a variability or scale CUSUM called a `CUSUMv` chart to monitor the precision performance of the instrument.

5 'Change Point Analysis' Panel

This tab shows change point analysis for mean and dispersion shifts for each precursor. An example for Study 9.1 is presented in Figure 7. The first change point model considers a step change in mean level of

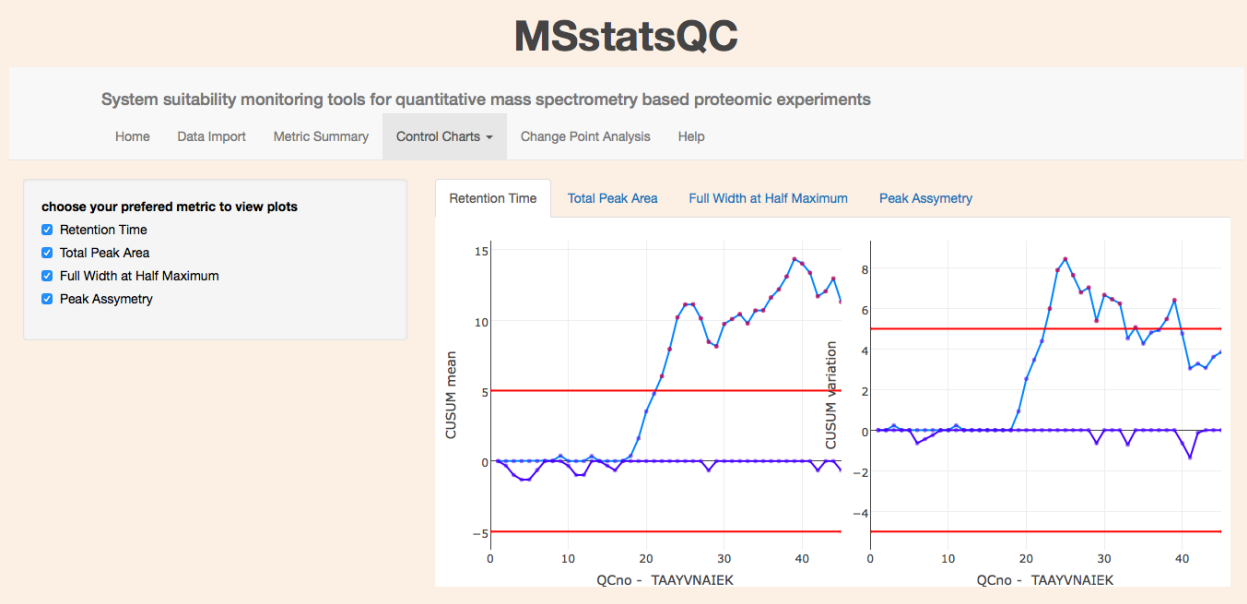

Figure 6: CUSUM Control Chart

a suitability metric. The change point estimator is the value which maximizes the change point function for process mean. Change point formulation for dispersion follows a similar approach using a change point function for process dispersion. The red vertical lines show the change point estimate which maximizes each change point function and corresponds to an estimation of change point.

6 'Help' Panel

The aim of this panel is to help user get information about the system suitability metrics and control charts used in `MSstatsQC`.

References

Abbatiello, Susan E, D R Mani, Birgit Schilling, Brendan Maclean, Lisa J Zimmerman, Xingdong Feng, Michael P Cusack, et al. 2013. "Design, implementation and multisite evaluation of a system suitability protocol for the quantitative assessment of instrument performance in liquid chromatography-multiple reaction monitoring-MS (LC-MRM-MS)." *Mol. Cell. Proteomics* 12 (9): 2623–39. doi:10.1074/mcp.M112.027078.

Abbatiello, Susan E, Birgit Schilling, D R Mani, Lisa J Zimmerman, Steven C Hall, Brendan MacLean, Matthew Albertolle, et al. 2015. "Large-scale inter-laboratory study to develop, analytically validate and apply highly multiplexed, quantitative peptide assays to measure cancer-relevant proteins in plasma." *Mol. Cell. Proteomics* 1 (409): M114.047050. doi:10.1074/mcp.M114.047050.

Bereman, Michael S., Joshua Beri, Vagisha Sharma, Cory Nathe, Josh Eckels, Brendan MacLean, and Michael J. MacCoss. 2016. "An Automated Pipeline to Monitor System Performance in Liquid Chromatography Tandem Mass Spectrometry Proteomic Experiments." *J. Proteome Res.*, September. American Chemical Society, acs.jproteome.6b00744. doi:10.1021/acs.jproteome.6b00744.

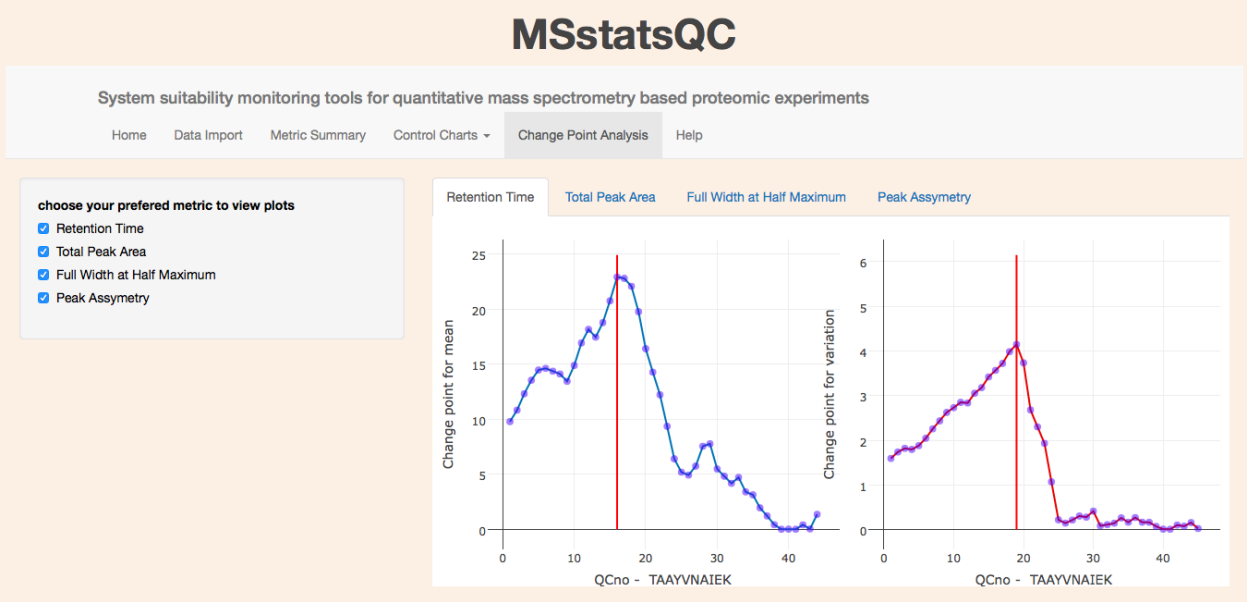

Figure 7: Change Point Analysis

Bereman, Michael S., Richard Johnson, James Bollinger, Yuval Boss, Nick Shulman, Brendan MacLean, Andrew N. Hoofnagle, and Michael J. MacCoss. 2014. "Implementation of statistical process control for proteomic experiments via LC MS/MS." *J. Am. Soc. Mass Spectrom.* 25 (4): 581–87. doi:10.1007/s13361-013-0824-5.

Ma, Ze Qiang, Kenneth O. Polzin, Surendra Dasari, Matthew C. Chambers, Birgit Schilling, Bradford W. Gibson, Bao Q. Tran, Lorenzo Vega-Montoto, Daniel C. Liebler, and David L. Tabb. 2012. "QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation." *Anal. Chem.* 84 (14): 5845–50. doi:10.1021/ac300629p.

MacLean, Brendan, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. 2010. "Skyline: an open source document editor for creating and analyzing targeted proteomics experiments." *Bioinformatics* 26 (7): 966–8. doi:10.1093/bioinformatics/btq054.

Pichler, Peter, Michael Mazanek, Frederico Dusberger, Lisa Weilnböck, Christian G. Huber, Christoph Stingl, Theo M. Luider, Werner L. Straube, Thomas Köcher, and Karl Mechtler. 2012. "SIMPATIQCO: A server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on orbitrap instruments." *J. Proteome Res.* 11 (11): 5540–47. doi:10.1021/pr300163u.

Rudnick, Paul A, Karl R Clauser, Lisa E Kilpatrick, Dmitrii V Tchekhovskoi, Pedatsur Neta, Dean D Billheimer, Ronald K Blackman, et al. 2009. "Performance Metrics for Evaluating Liquid Chromatography-Tandem Mass Spectrometry Systems in Shotgun Proteomics." *Mol. Cell. Biol.*, no. 9.2: 225–41. doi:10.1074/mcp.M900223-MCP200.

Sharma, Vagisha, Josh Eckels, Greg K Taylor, Nicholas J Shulman, Andrew B Stergachis, Shannon A Joyner, Ping Yan, et al. 2014. "Panorama: a targeted proteomics knowledge base." *J. Proteome Res.* 13 (9). American Chemical Society: 4205–10. doi:10.1021/pr5006636.

Taylor, Ryan M., Jamison Dance, Russ J. Taylor, and John T. Prince. 2013. "Metriculator: Quality assessment for mass spectrometry-based proteomics." *Bioinformatics* 29 (22): 2948–49. doi:10.1093/bioinformatics/btt510.