Systems Virology
of Emerging Respiratory Viruses

M001.0P - Statistical Protocol: Normalization, QA/QC and Differential Expression Analysis for Agilent Arrays

Author(s): Dr. Armand Bankhead, PhD and Dr. Shannon K McWeeney, PhD
Version: 1.0
4-28-10

Normalization and QA/QC: Raw data consists of the output from the Agilent Feature Extraction (AFE) software. Data is read into the R/Bioconductor Framework. Data visualization of raw intensity levels is performed using boxplots, histograms and correlation plots. All intensities are background corrected. Quantile normalization is performed and data visualization is again performed to assess effectiveness of normalization. Distribution of control probes is examined and then those probes are removed. Descriptive statistics are reported for non-control replicate probes and the intensity distributions are examined via histograms. All non-control replicate probes are then summarized by taking the mean intensity measurement. To assess the quality of each probe's reading, five Agilent QC flags are considered:

1. gIsFound – tests for minimal spot diameter and signal greater than 1.5 times background noise
2. gIsWellAboveBG – requires positive intensity and background-subtracted intensity greater than 2.6 times background noise
3. gIsSaturated – true if more than 50% of spot pixels are greater than saturation threshold
4. gIsFeatNonUnifOL—examines variance of spot pixels
5. gIsFeatPopnOL—checks if replicate probe is greater than 1.42 times the inter-quartile range

A given probe must pass all 5 QC flags for all treatment (i.e., infected) replicates of at least one experimental time point. This allows a conservative criteria regarding probe quality, but yet is moderate enough to detect transcripts that may be expressed only briefly (i.e., as little as one time point) during an infection time course. It is imperative that this filtering occurs before differential expression analysis so that low intensity (below background) and low quality probes are not considered in the analysis. Final QA/QC involves CV plots, heatmaps, cluster plots, scatter plots and PCA as needed in order to assess if there are sample outliers or additional batch effects.

Differential Expression Analysis: The statistical model utilized will be dependent up on the experimental design for the experiment being analyzed. A linear model is fit for each transcript/probe. All p-values will be False Discovery Rate (FDR) adjusted. Significant differential expression of a probe is defined using the following

criteria: $\log_2$ fold change > |1.5| and an FDR adjusted p-value of less than .05. Fold changes are calculated by taking the average $\log_2$ probe intensity of infected replicates and subtracting the average $\log_2$ probe intensity of mock-infected replicates from the same time point. For data sets in which the effect sizes are small, non-$\log_2$ fold-change threshold may be considered (FDR adjusted p-value remains at 0.05).

Software Utilized: Custom R Bioconductor workflow that includes the Agi4x44PreProcess Bioconductor package plus in-house scripts for QA/QC visualization and differential expression analysis.