

Quality Assessment of Microarray Data and Optimal Filtering Criteria

Takashi Kido^{1*}, Janos Demeter², Gavin Sherlock¹

Departments of Genetics¹ and Biochemistry², Stanford University, Stanford, CA 94305-5120,
USA

*email: msc-kido@mbf.ocn.ne.jp

Summary

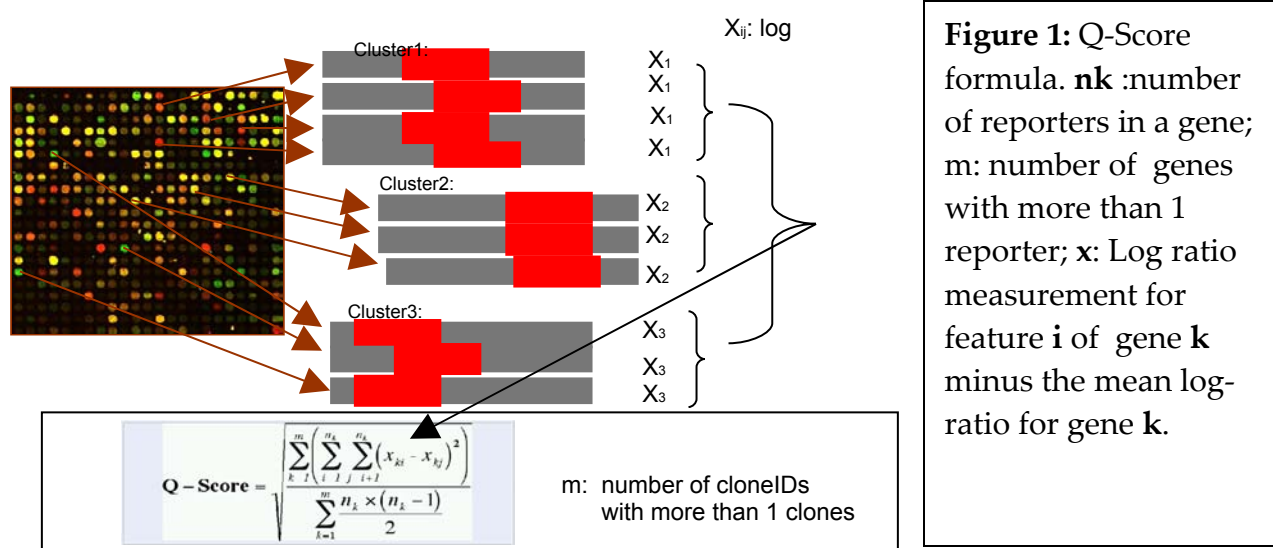
Quality assessment of microarray data is of vital importance, but standard approaches have not yet been established. Thus, researchers use *ad hoc* filtering criteria to select data for analysis, because there is little or no guidance on optimal filtering parameters. The goals of this project are (1) to generate quality metrics for individual microarrays and microarray datasets, and (2) to learn how best to apply quality filters to microarray data using these metrics. Here, we describe the “Q-Score” as a quality assessment measure, and present a tool that visualizes the dynamics of the Q-Score with respect to several filters, to allow researchers to explore the quality of their data. Our goal is to further develop this tool, such that it can suggest optimal filtering parameters.

Definition of the Q-Score

It is very common that the expression of a gene is measured multiple times, using different sequence probes (reporters), on the same array. The average spread of the log-ratios of reporters mapping to the same gene may be used to assess the quality of an array. The Q-Score is a measure of the ‘within gene’ spread of the data, and is calculated using the formula shown in **Figure 1**. A lower Q- score means a narrower spread of the data and a better quality array, while a higher value results from a wider spread of the data and a lower quality array.

Quality Assessment and Filtering with the Q-Score

By calculating the Q-Score with different fractions of the data removed, based on a spot metric’s value, the Q-Score can be used to guide for the selection of a cut-off value for that filter to remove low quality spots. The expectation is that the Q-Score will decrease (improve) with increasingly stringent filter values. Using a filter that is removing lower quality spots first, we might expect the score to drop faster initially, then reach a region of plateau or slower rate of decrease. An example is shown on the **Q vs Fraction** graph in **Figure 2 (a)**. The arrows show clear inflection points for the arrays indicated.



Experimental Evaluations of Q-Score dynamics

We have noted the following observations in various microarray data experiments:

1. The Q-Score is reasonably consistent with the visual spot evaluations.
2. Inflection points of Q-Score may be useful for determining the optimal cutoff for filtering data.
3. Filtering with Ch2 normalized Intensity/median background and Regression Correlation improves the Q-Score the most.
4. Spot metrics can be clustered. Correlation matrix among dozens of filters is very robust, it may be useful for choosing orthogonal filters.

GUI tools

We have developed a quality assessment and filtering tool for the two-channel microarray platform based on the Q-Score dynamics. This tool provides the functions to compare the several filtering dynamics for optimizing filtering criteria. This GUI tool is implemented using Perl-Tk, and can run on the PC and Macintosh platforms.

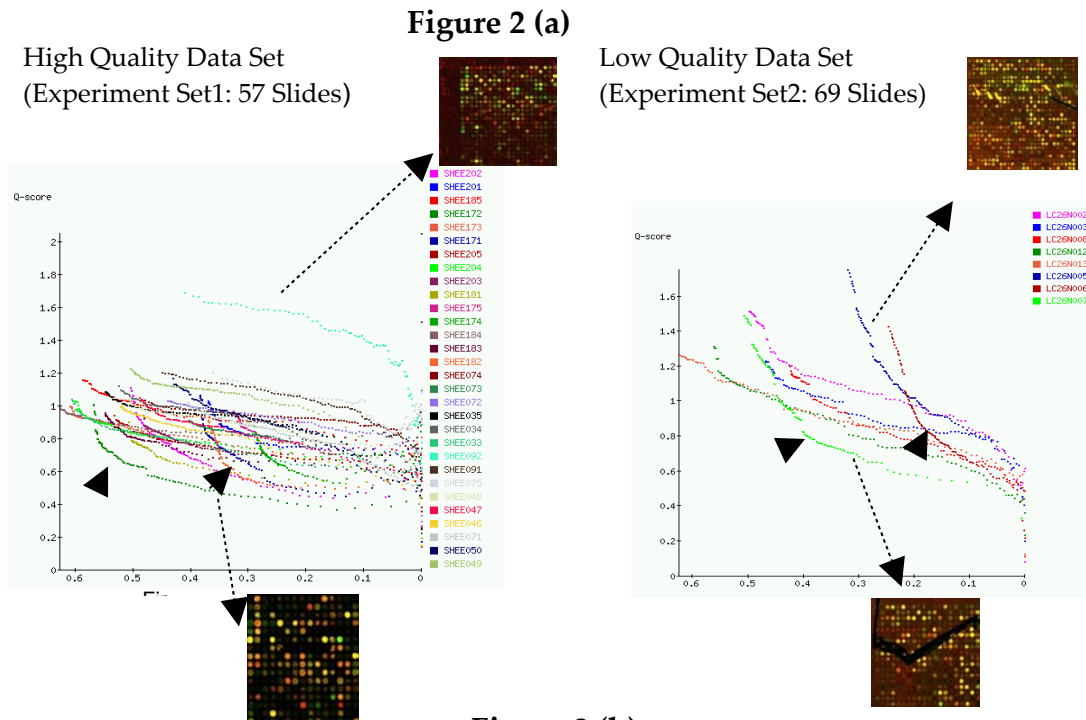


Figure 2 (b)

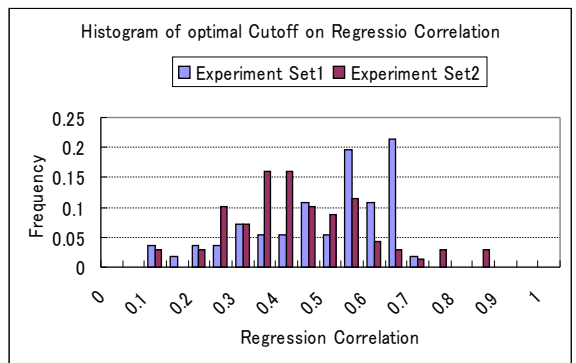


Figure 2 (a). shows a typical example of the Q-Fraction graph for two different experiment sets in Stanford Microarray Databae (SMD), which were categorized as high and low quality sets respectively by visual observation. The Q-Score works reasonably; the high quality data set has a lower Q-Score, and higher fraction of the data is retained at the inflection points than in the low quality data set.

Figure 2 (b). We detected the inflection points for Q-fraction graphs of five widely used filters in experiment set1 (57 slides) and experiment set2 (69 slides). The distribution of those inflection points differs for different filters. The usual default cutoff value, (for example, 0.6 for regression correlation) is not always the same as that of the inflection point. The experiment sets have slightly or considerably different distributions to each other.