A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays

Xiaohua Douglas Zhang *

Biometrics Research, Merck Research Laboratories, West Point, PA 19486, USA

Received 8 May 2006; accepted 20 December 2006
Available online 2 February 2007

Abstract

RNA interference (RNAi) high-throughput screening (HTS) enables massive parallel gene silencing and is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes. The application of genome-scale RNAi relies on the development of high-quality RNAi HTS assays. To obtain high-quality HTS assays, there is a strong need for an easily interpretable and theoretically based quality control (QC) metric. Signal-to-noise ratio (S/N), signal-to-background ratio (S/B), and Z-factor have been adopted as QC metrics in HTS assays. In this paper, I proposed a pair of new parameters, strictly standardized mean difference (SSMD) and coefficient of variability in difference (CVD), as QC metrics in RNAi HTS assays. Compared to S/B and S/N, SSMD and CVD capture the variabilities in both compared populations. Compared to Z-factor, SSMD and CVD have a clear probability interpretation and a solid statistical basis. Accordingly, the cutoff criteria of using SSMD or CVD as a QC metric in HTS assays are fully theoretically based. In addition, I discuss the relationship between the SSMD-based criterion and the popular Z-factor-based criterion and elucidate why p-value from t-test of testing mean difference fails to serve as a QC metric.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Strictly standardized mean difference; Coefficient of variability in difference; Z-factor; d’-probability; RNA interference; High-throughput screening

Method

Recently, RNA interference (RNAi), a natural mechanism for gene silencing [1,2], has made its way as a widely used method in molecular biology in both academics and industry. Academic researchers have used RNAi to elucidate gene functions through studying a loss-of-function phenotype. Pharmaceutical and biotech companies have set up libraries for large-scale screens employing thousands of short-interfering RNA- (siRNA) or short hairpin RNA- (shRNA) encoding vectors to identify new factors involved in the molecular pathways of diseases [3]. RNAi may lead to advances not only in drug target identification and validation but also in the development of a potential whole new class of RNAi-based therapeutic agents [4]. The first clinical trials based on RNAi were initiated to treat patients with age-related macular degeneration [5]. RNAi has even been seen as the third class of drug targets after small molecules and proteins [6]. Based on siRNA or shRNA libraries, RNAi high-throughput screening (HTS) enables massive parallel gene silencing to reveal the extent to which interference with the expression of specific genes alters the cell phenotype, and it is increasingly being used to reveal novel connections between genes and disease-relevant phenotypes [7–10].

Statistical methods for small-molecule HTS data have been described [11–15]. Zhang et al. [16,17] explored statistical methods for hit selection in RNAi HTS experiments. The application of genome-scale RNAi relies on the development of high-quality RNAi HTS assays. However, despite a strong need for a theoretically based and easily interpretable quality control (QC) metric in RNAi HTS assays, such a QC metric has yet to be developed. An important QC characteristic in an HTS assay is how much the positive controls, tested compounds, and negative controls differ from one another in the assay. This QC characteristic can be evaluated using the comparison of two well types in HTS assays.
Signal-to-noise ratio (S/N), signal-to-background ratio (S/B), and Z-factor have been adopted to evaluate the quality of HTS assays through the comparison of two investigated types of wells. However, S/B does not take into account any information on variability; and S/N can capture the variability only in one group and hence cannot assess the quality of assay when the two groups have different variabilities. Zhang et al. [11] proposed a screening window coefficient called “Z-factor.” The advantage of Z-factor over S/N and S/B is that it takes into account the variabilities in both compared groups. As a result, Z-factor has been broadly used as a QC metric in HTS assays (cf. [7,11,18–21]). However, its probability basis has not been explored and its statistical properties have not been investigated.

In this paper, I propose a pair of novel parameters, strictly standardized mean difference (SSMD) and its reciprocal, coefficient of variation of difference (CVD), for measuring the magnitude of difference between two populations, and I investigate their capacity as QC metrics in RNAi HTS assays. Like Z-factor, SSMD and CVD capture the variabilities in both groups; and they are simple statistical parameters. But unlike the Z-factor, this pair of parameters has a clear probability interpretation and a solid statistical basis.

**Methods**

Suppose two populations $P_1$ and $P_2$ with random value have distributions of $F_1$ and $F_2$, respectively. The first population has mean $\mu_1$ and variance $\sigma_1^2$ and the second population has mean $\mu_2$ and variance $\sigma_2^2$. The covariance between these two populations is $\sigma_{12}$. Further suppose we have one sample of size $n_1$, namely $X_{11}, X_{12}, \ldots, X_{1n_1}$, being independently identically distributed from the first population and another independent sample of size $n_2$, namely $X_{21}, X_{22}, \ldots, X_{2n_2}$, being independently identically distributed from the second population. $X_1$ and $X_2$ are respectively the sample mean and the sample standard deviation (SD) in the first sample; $\bar{X}_1$ and $\bar{X}_2$ are respectively the sample mean and the SD in the second sample.

**Signal-to-noise ratio and Z-factor**

Historically, S/N and S/B are two measures that have been used loosely in small-molecule HTS assays. Their definitions are

\[
S/N = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_2} \quad \text{and} \quad S/B = \frac{\bar{X}_1}{\bar{X}_2}
\]

The criticism for S/B (and the mean difference $\mu_1 - \mu_2$) is that it does not contain any information regarding data variability. S/N does take into account the variability, but only in a single population. It does not take into account the variability in the other population.

Z-factor was proposed to measure the separation between “tested compound” wells and “negative control” wells or between “positive control” wells and “negative control” wells. Let $Z_F$ denote Z-factor. Z-factor is defined as [11]

\[
Z_F = 1 - \frac{3(\sigma_1 + \sigma_2)}{|\mu_1 - \mu_2|}.
\]

Zhang et al. [11] further use “Z-factor” to refer to the parameter between tested compound wells and negative control wells and “Z′-factor” to refer to the parameter between positive control wells and negative control wells. For convenience, we will use Z-factor to refer to either of them depending on the compared groups in this paper. As the authors pointed out, the Z-factor that they used is a plug-in formula as follows:

\[
Z_F = 1 - \frac{3(\sigma_1 + \sigma_2)}{|\mu_1 - \mu_2|}.
\]

They used sample mean directly in the position of population mean and sample SD in the position of population SD in the formula for Z-factor. The point estimate or interval estimation of Z-factor has not been investigated from a solid statistical basis. The meaning of Z-factor has not been explored from a probability perspective. Below, I propose a pair of new parameters with a solid probability and statistics basis that can take into account the variabilities in both groups.

**A pair of new parameters**

Let $D$ denote the difference between populations $P_1$ and $P_2$, namely $D=P_1-P_2$. The mean $\mu_2$ and variance $\sigma_2^2$ of $D$ are $\mu_2=\mu_1-\mu_2$ and $\sigma_2^2=\sigma_1^2+\sigma_2^2-2\sigma_{12}$, respectively. The ratio of mean to SD of the difference $D$ is defined as strictly standardized mean difference. Let $\beta$ denote SSMD. Then

\[
\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}.
\]

If the two populations are independent, $\sigma_{12}=0$ and $\sigma_2^2=\sigma_1^2+\sigma_2^2$. Then

\[
\beta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}.
\]

The larger the absolute value of SSMD between two populations, the greater the differentiation between the two populations. The term “standardized mean difference” [22] has been used as a type of effect size, referring to the mean difference standardized to the SD of the control group or the average SD under the assumption of equal variance. For example, S/N is such a type of standardized mean difference. Here, the mean of the random variable representing the difference is strictly standardized to SD of the random variable itself; thus we have the name “strictly standardized mean difference” for this parameter.

The coefficient of variation in difference is defined as the coefficient of variation of $D$. Let $\omega$ denote CVD. Then

\[
\omega = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{\mu_1 - \mu_2}}.
\]

If the two populations are independent,

\[
\omega = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\mu_1 - \mu_2}}.
\]

As in the original meaning of coefficient of variability for a random variable, CVD represents the relative SD of the difference with respect to mean of the difference. The larger the absolute value of CVD between two populations, the less the differentiation between the two populations. Clearly, CVD is the reciprocal of SSMD. However, CVD=∞ when $\mu_1=\mu_2$, which may make CVD less favorable than SSMD in some situations. Let $Z$ denote the standardized difference (namely $Z = \frac{D - \mu_d}{\sigma_d}$) and $F_Z(z)$ be the cumulative distribution function of $Z$. Then

\[
\Pr(D>0) = \Pr\left(\frac{D - \mu_d}{\sigma_d} > \frac{0 - \mu_d}{\sigma_d}\right) = \Pr(Z > -\beta) = 1 - F_Z(-\beta).
\]

Thus, the probability that the difference $D$ is positive is a function of $\beta$. For convenience, let us call this probability “positive difference probability” and use $d^+$-probability to denote it. That is, $d^+$-probability=$\Pr(D>0)$. When the difference $D$ is symmetrically distributed, $d^+$-probability=$\Pr(Z>-\beta)$. When $D$ is normally distributed, the relationship between SSMD and
\(d\)-probability is as simple as \(d\)-probability=\(\Phi(t)\) where \(\Phi(t)\) denotes the cumulative distribution function of the standard normal distribution \(N(0,1)\). Hence, the value of SSMD reflects the probability of the difference being greater than 0. The larger the value of SSMD, the greater the probability of the difference being greater than 0. In other words, SSMD has a clear meaning indicated by the \(d\)-probability.

### Statistical estimation and inference of SSMD and CVD

In this paper, we focus on the estimation and inference under the condition of independence between two populations, namely based on the definitions of Eqs. (4) and (6). We can obtain maximum likelihood estimates (MLEs) of SSMD and CVD and their asymptotic distributions using the Delta method and the asymptotic normality of MLE. The results are summarized in Proposition 1, which is presented and proved in the Appendix.

Based on Proposition 1, we obtain the MLE point estimate \(\hat{\beta}\) as follows:

\[
\hat{\beta} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_{ij} - \bar{X}_1)^2 + \frac{1}{n_2} \sum_{j=1}^{n_2} (X_{ij} - \bar{X}_2)^2}}.
\]

(7)

Let \(Z_{\alpha/2} = \Phi(1 - \frac{1}{\alpha})\). Using the asymptotic variance \(\sigma_0^2\) and the invariance property of MLE, the \(1-\alpha\) confidence interval of \(\beta\) is approximately

\[
\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_{ij} - \bar{X}_1)^2 + \frac{1}{n_2} \sum_{j=1}^{n_2} (X_{ij} - \bar{X}_2)^2}} \pm Z_{\alpha/2} \hat{\sigma}_0.
\]

(8)

Similarly, we can obtain the MLE point estimate and confidence interval of \(\omega\). See the Appendix for more details.

So far, the point and interval estimations have been derived using MLE. Other point estimates for both SSMD and CVD can also be useful. For example, considering that \(X_1, X_2, \tilde{X}_1, \text{ and } \tilde{X}_2\) are unbiased estimates, respectively, for \(\mu_1, \mu_2, \sigma_1, \text{ and } \sigma_2\), a method-of-moment (MM) estimate for SSMD or CVD may be obtained as follows:

\[
\hat{\beta}_{\text{MM}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\tilde{X}_1^2 + \tilde{X}_2^2}} \quad \text{and} \quad \hat{\omega}_{\text{MM}} = \sqrt{\frac{\tilde{X}_1^2 + \tilde{X}_2^2}{\bar{X}_1 - \bar{X}_2}}.
\]

(9)

When the distributions of the two populations have non-normality, skewness, and outliers, the robust point estimates for SSMD and CVD may respectively be

\[
\hat{\beta}_{\text{robust}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\tilde{X}_1^2 + \tilde{X}_2^2}} \quad \text{and} \quad \hat{\omega}_{\text{robust}} = \sqrt{\frac{\tilde{X}_1^2 + \tilde{X}_2^2}{\bar{X}_1 - \bar{X}_2}}.
\]

(10)

where \(\bar{X}_i\) and \(\tilde{X}_i\) (\(i=1, 2\)) are respectively the median and median absolute deviation (MAD) of the sample from the \(i\)th population. To match SD with MAD in a sample from a normal distribution, we may use the rescaled MAD, namely, \(s = 1.4826\) median (\(X_i - \bar{X}_i\)) (\(i=1, 2, j=1, \ldots, n_i\)). It may be worthwhile to note that the SSMD and CVD robust estimates are approximations, which may not estimate exactly the same underlying parameter as the nonrobust measures.

### Comparison of parameters for QC in RNAi HTS assays

We compare S/R, S/N, Z-factor, CVD, and SSMD as population parameters for quality control in RNAi HTS assays on the condition of normal distributions with known means and standard deviations in a positive control and a negative control. Fig. 1 displays some typical population distributions of positive and negative controls appearing in RNAi HTS assays. The clear differentiation between these two controls indicates good quality of an HTS plate.

S/B does not contain any information regarding data variability. S/N is based on the assumption that the two compared populations compared have the same variability. If the variabilities in the two populations are different, both S/R and S/N will produce misleading results. For example, S/B=0.71 and S/N=6.67 in Cases A, C, and D of Fig. 1. However, we can clearly see that the separation between the positive and the negative controls in these three cases are very different. Furthermore, the cutoff criterion of using S/N to evaluate QC is hardly clear. For example, we see that the S/N in Cases A and D are all high and may think about a cutoff of 6 for S/N. Even if we used this big cutoff value, we would still conclude that there is good differentiation between the positive and the negative controls in Case C, which has an S/N value bigger than 6. This is clearly a misleading result as seen in Fig. 1.

The Z-factor is the ratio of separation of the [\(\bar{\mu}_1 - \bar{\mu}_2\)]/(3\(\sigma_1 + 3\sigma_2\)), to the signal dynamic range, [\(\bar{\mu}_1 - \bar{\mu}_2\)], of an assay. It is clear that Z-factor captures the mean difference and the variabilities in both compared populations. As a result, Z-factor works better than S/B and S/N as a QC metric and has thus been broadly used in HTS assays (cf. [7,11,18–21]). For the use of Z-factor as a QC metric in small-molecule HTS assays, Zhang et al. [11] proposed the following cutoff criterion: Z-factor = 1 for “ideal assay,” Z-factor > 0.5 for “excellent assay,” 0.5 > Z-factor > 0 for “doable assay,” Z-factor < 0 for “yes/no type assay,” and Z-factor < 0 for “screening essentially impossible.” Using this Z-factor-based criterion, Cases A, D, and F are classified as “doable assays” meaning “separation band is small”; Cases B, C, and E are classified as “screening essentially impossible.” However, the probability meaning of this Z-factor-based criterion is still unclear.

The major part of Z-factor, \(\frac{\sigma_1 + \sigma_2}{\bar{\mu}_1 - \bar{\mu}_2}\), looks similar to CVD, \(\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{\bar{\mu}_1 - \bar{\mu}_2}}\) (11) Like Z-factor, CVD and SSMD measure the magnitude of difference and capture the variabilities in both populations. But unlike the Z-factor, SSMD, and CVD and their cutoff criteria are easily interpretable in probability. For example, SSMD \(\geq 3\) indicates that the size of the mean difference is at least three times that of the SD of the difference between two populations. As a probability interpretation, SSMD \(\geq 3\) indicates that the probability that a value from the first population is greater that a value from the second population is greater than \(\Phi(3)=0.99865\) when the difference is normally distributed, and is greater than 0.95 when the difference has a distribution with unimodal and finite variance. Thus, we may set up the following simple criterion for quality control: in the situation in which the positive control has value greater than the negative control, a plate passes in QC if it has SSMD \(\geq 3\) and fails in QC if it has SSMD < 3; in the situation in which the positive control has value less than the negative control, a plate passes in QC if it has SSMD \(\geq -3\) and fails in QC if it has SSMD > -3. Both the |SSMD| \(\geq 3\) criterion and the mean \(\pm 3\times\)SD method originate from the well-known three sigma rule.

Using the SSMD-based criterion, the SSMD values are above 3 in Cases A, D, E, and F and are less than 3 in Cases B and C, which suggests that there is good differentiation between positive and negative controls in Cases A, D, E, and F but poor differentiation in Cases B and C. This result is sensible as displayed in Fig. 1. The \(d\)-probabilities (namely the probabilities that a value from positive control is greater than a value from the negative control) in Cases A, D, E, and F are respectively \(2.6 \times 10^{-3}\), \(3 \times 10^{-15}\), \(3.7 \times 10^{-7}\), and \(5.7 \times 10^{-3}\), all very small, while the probabilities in Cases B and C are respectively 0.123 and 0.081, both fairly high. This offers meaningful information on the magnitude of difference between positive and negative control in each plate from a probability perspective.

The probability meaning of SSMD also provides a basis for a probability interpretation to Z-factor-based criterion. Using the inequality of \(\sqrt{\sigma_1^2 + \sigma_2^2} < \sigma_1 + \sigma_2\) when \(\sigma_1 > 0\) and \(\sigma_2 > 0\), we can get |SSMD| > \(\frac{3}{1 - Z\text{-factor}}\) Therefore, given a value \(a\) (not greater than 1), if Z Factor > a then |SSMD| > \(\frac{3}{1 - a}\); however, if |SSMD| > \(\frac{3}{1 - a}\), we cannot ensure that Z-factor > a. For example, if Z-Factor > 0 (or 0.5) then |SSMD| > 3 or (6). However, if |SSMD| > 3 (or 6), we cannot ensure that Z-factor > 0 (or 0.5). An example of |SSMD| > 3 but Z-factor < 0 is displayed in Fig. 1E. Thus, Z-factor > 0 (or 0.5) is a
Comparison of SSMD and classical t-statistic

It is well known that t-test has widely been used for the comparison of two populations. The t-statistic for testing no mean difference under the situation of unequal variance is \( t = \frac{X_1 - X_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \), which looks similar to the MLE estimate of SSMD, \( \hat{\beta} \) in Eq. (7), and the MM estimate of SSMD, \( \hat{\beta}_{MM} \) in Eq. (9). The difference is in the denominators. As a result, the classical t-statistic is a function of both sample size and magnitude of difference. In general, larger sample size leads to larger absolute t-value and smaller p-value. In fact, as the total sample size \( N \rightarrow \infty \), t-value goes to \( \infty \) or \( -\infty \) and p-value goes to 0. By contrast, from Proposition 1, as sample size increases, the SSMD estimate goes to the true SSMD value, \( \beta_0 \), in probability. \( \beta_0 \) is usually a limited number. Consequently, as \( N \rightarrow \infty \), the SSMD estimate will not go to \( \infty \) or \( -\infty \). Sample size can impact only how accurately and precisely SSMD estimate can represent SSMD true value, which is reflected in the confidence interval of \( \beta \) as shown in Eq. (8). Otherwise, sample size has no impact on SSMD. Thus, we can still maintain the benefit of increasing sample size in an assay: increasing the accuracy and precision of the SSMD estimation by making the variance of \( \hat{\beta} \) smaller and the confidence interval of \( \beta \) narrower.

Let us compare the performance of SSMD and the t-statistic (for testing \( H_0: \mu_1 = \mu_2 \)) for QC in RNAi HTS assays. Suppose we observe four plates (Plates \( A_1, B_1, C_1 \), and \( A_2 \)) with Plate \( A_1 \) from Case A of Fig. 1. Plates \( B_1 \) and \( B_2 \) from Case B, and Plate \( C_1 \) from Case C. The sample means, SDs, and sample sizes of the four plates are listed in the first column of Table 1. Since Plates \( B_1 \) and \( B_2 \) are from Case B, the value of a good QC metric should be approximately the same in these two plates. However, as shown in Table 1, from Plates \( B_1 \) to \( B_2 \), t-value is doubled and p-value changes from 0.034 to 3.5 × 10^{-3}, whereas \( \beta \) changes only slightly from 1.338 to 1.197. This is one piece of evidence that classic t-statistic for testing no mean difference cannot work effectively as a QC metric in RNAi HTS assays, while SSMD can. In addition, if we used the p-values from t-tests to evaluate the quality of RNAi HTS assays, we might have concluded that Plates \( B_2 \) nd \( C_1 \) had better differentiation between positive and negative controls (and thus had
better quality) than Plate $A_1$ because the $p$-values in Plates $B_2$ and $C_1$ are smaller than in Plate $A_1$. This is clearly a misleading judgement as Plate $A_1$ is from Case A, which clearly has much better differentiation between positive and negative controls than Cases B and C.

**Results**

A good QC metric should work in a variety of experiments. Thus in this paper we concentrate on plates extracted from different RNAi HTS experiments, which may have different data ranges and different numbers of positive and negative control wells, so that we can see the impact of sample size and data range on the QC metrics. The data are extracted from several experiments on diseases such as infectious diseases and cancers, including a hepatitis C virus RNAi HTS experiment described in Zuck *et al.* [7] and Zhang *et al.* [17].

The bottom of Fig. 2 shows the data from plates 1–11, 12–15 and 16–18 of experiments A, B, and C, respectively. The data have the following notable features. The data range (between $-3.5$ and $0.5$) in Experiment A was different from those (between $-18$ and $-8$) in Experiments B and C. The numbers of control wells at each plate differed in each of the three experiments: 16 positive control wells and 16 negative control wells in Experiment A, 16 positive control wells and 8 negative control wells in Experiment B, and 4 positive control wells and 8 negative control wells in Experiment C. The measured intensities in $-\log_2$ scale were roughly symmetric with a few outliers. There were outliers in the positive controls in Plates 3, 4, 15, and 16 and in the negative controls in Plates 1, 2, 11, and 12.

**Table 1**

<table>
<thead>
<tr>
<th>Sample mean, SD, and sample size</th>
<th>$\bar{\beta}$</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$\bar{X}_1 = -10, s_1 = 0.64, n_1 = 4$</td>
<td>5.265</td>
<td>9.119</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}_2 = -14, s_2 = 0.60, n_2 = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B_1$</td>
<td>$\bar{X}_1 = -11, s_1 = 1.20, n_1 = 4$</td>
<td>1.338</td>
<td>2.318</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}_2 = -13, s_2 = 1.24, n_2 = 4$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B_2$</td>
<td>$\bar{X}_1 = -11, s_1 = 1.20, n_1 = 16$</td>
<td>1.197</td>
<td>4.636</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}_2 = -13, s_2 = 1.24, n_2 = 16$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_1$</td>
<td>$\bar{X}_1 = -10, s_1 = 2.8, n_1 = 16$</td>
<td>1.443</td>
<td>5.587</td>
</tr>
<tr>
<td></td>
<td>$\bar{X}_2 = -14, s_2 = 0.6, n_2 = 16$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Data from positive and negative controls (shown at the bottom) and the estimated values of SSMD and $Z$-factor (shown at the top) for 18 plates from three RNAi HTS experiments, A, B, and C. At the bottom, a red (or green) point represents the measured intensity of a positive (or negative) control well in a plate. At the top, the two orange dashed lines represent the cutoffs of $\bar{\beta} = 3$ and $\bar{Z} = 0$, respectively. A black (or blue) dot represents the estimated (or robust) value of SSMD; a black (or blue) cross represents the estimated (or robust) value of $Z$-factor in a plate.
negative controls in Plates 5 and 12. Even in the plates without outliers in the same experiment, the data variations (indicated by $s_1$) in positive controls sometimes differ from those (indicated by $s_2$) in negative controls. For example, $s_1 = 0.29$ nearly doubled $s_2 = 0.17$ in Plate 1, whereas $s_1 = 0.11$ was nearly the half size of $s_2 = 0.20$ in Plate 11. There were also cases in which the variations in the two controls were approximately equal, such as $s_1 = 0.27$ and $s_2 = 0.29$ in Plate 7 and $s_1 = 0.20$ and $s_2 = 0.25$ in Plate 8.

Considering the unequal variabilities in the two controls and different data ranges in different experiments, $S/N$ and $S/B$ cannot work effectively as QC metrics here. Thus, we focused on investigating the use of $SSMD$ and $Z$-factor as QC metrics. For simplicity, we used the point MLE estimate of $SSMD$, $\beta$ in Eq. (7), and the commonly used plug-in estimate of $Z$-factor, $\tilde{Z}_r$ in Eq. (2), and we used the simple cutoff criteria of $\beta > 3$ for a pass in QC when using $SSMD$ and $\tilde{Z}_r > 0$ when using $Z$-factor. Considering outliers, we also used the robust estimate of $SSMD$ in Eq. (10) and a similar robust estimate of $Z$-factor as follows:

$$\tilde{Z}_{f\text{robust}} = 1 - \frac{3(\hat{s}_1 + \hat{s}_2)}{|\hat{X}_1 - \hat{X}_2|}. \quad (11)$$

The estimated values of $SSMD$ and $Z$ factor are shown respectively using the black dots and black crosses at the top of Fig. 2. The robust estimates are shown in blue. The judgment of QC for the 18 plates are summarized in Table 2. From Table 2, all the plates that passed in QC by $Z$-factor also passed in QC by $SSMD$. However, three plates (Plates 6, 7, 13) that did not have outliers passed in QC by $SSMD$ but not by $Z$-factor, which demonstrates in practice that $Z$-factor $> 0$ is a subset of $SSMD > 3$ as described under Methods.

Let us first look at Plates 1–15, in each of which there were 16 positive control wells and at least 8 negative control wells. From the data shown at the bottom of Fig. 2, we observed that the positive controls and the negative controls were not well differentiated from one another in Plates 1 and 2 and were well differentiated in Plates 6–14. We also observed outliers in the positive controls of Plates 3, 4, and 15 and in the negative controls of Plates 5 and 12. For the plates without outliers, using $SSMD$, Plates 1 and 2 failed in QC and Plates 6–11 and 13–14 all passed in QC, which matches with the observation of the differentiation in these plates. Using $Z$-factor in the plates without outliers, Plates 1, 2, 6, 7, and 13 failed in QC and Plates 8–11 and 14 passed in QC. The judgment by the $Z$-factor in Plates 6, 7, and 13 suggest that the $Z$-factor criterion is conservative.

Both $SSMD$ and $Z$-factor indicated that Plates 3 and 4 failed in QC. If we ignore the outliers, the differentiations between positive and negative controls in Plates 3 and 4 were even stronger than in Plates 8–10, which is suggested by estimated robust values of $SSMD$ and $Z$-factor. This result indicates that both $SSMD$ and $Z$-factor are affected by outliers. It is also notable that Plates 5, 12, and 15 each had one less extreme outlier; however, they all passed in QC by $SSMD$ but failed by $Z$-factor, which might suggest that $SSMD$ is more robust to outliers than $Z$-factor.

The differentiation between the positive control and the negative control in Plates 8, 9, and 10 appeared to be very similar although Plate 8 had nearly equal variabilities, Plate 9 had greater variability in the positive control, and Plate 10 had greater variability in the negative control. The estimated values $SSMD$ and $Z$-factor were nearly the same in each of the three plates (top of Fig. 2), which suggests that both $SSMD$ and $Z$-factor worked effectively in the situations of either equal or unequal variabilities in two groups. However, the robust estimates in these three plates varied, which reminds us to be cautious in using robust estimates since the distribution and variance of robust estimates are difficult to determine. Robust estimation certainly does not work as well as regular estimation in the normality situation without outliers, although it works better when there are outliers and nonsymmetric distributions.

From the bottom of Fig. 2, it is fairly clear that the positive control was well differentiated from the negative control in Plate 17. However, it is not clear whether Plates 16 and 18 really passed or failed in QC although Plate 16 was classified as “fail” and Plate 18 as “pass” by both $SSMD$ and $Z$-factor. In Plate 18, the mean difference between the two controls was not large and the estimated variability in the positive control was much smaller than in the remaining plates in Experiments B and C ($s_1 = 0.18$ in Plate 18 vs $s_1 = 1.05$ and 0.39, respectively, in Plates 16 and 17). If we have one more observation in the positive control that is reasonably less than the smallest observed value in Plate 18, the estimated variability may be nearly the same in Plate 17 and it will fail in QC. On the other hand, if we remove the smallest observed value in Plate 16, then the positive control will be well differentiated from the negative control. However, by removing one value, we remove 25% of the data since there were only four observed values in the positive control. Therefore, it is not easy to judge the quality in Experiment C because the number of observed values in the positive control was too small. By contrast, it is fairly clear that Plates 12 and 15 should have good quality and they were judged as “pass in QC” by $SSMD$ even though there was one extremely low value among the 8 negative control wells in Plate 12 and one extremely low value among

<table>
<thead>
<tr>
<th>Table 2</th>
<th>The results of QC evaluation using Z factor and SSMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result category</td>
<td>Outlier in a plate</td>
</tr>
<tr>
<td>Pass in QC by both metrics</td>
<td>Without outlier</td>
</tr>
<tr>
<td>Fail in QC by both metrics</td>
<td>Without outlier</td>
</tr>
<tr>
<td>Fail in QC by both metrics</td>
<td>With outlier</td>
</tr>
<tr>
<td>Pass in QC only by SSMD</td>
<td>Without outlier</td>
</tr>
<tr>
<td>Pass in QC only by SSMD</td>
<td>With outlier</td>
</tr>
</tbody>
</table>
the 16 positive control wells in Plate 15 of Experiment B. Therefore, this observation suggests that four replicates in a control are not enough, while eight or more replicates may be enough for using the QC criterion based on estimated SSMD.

Discussion

The application of genome-scale RNAi relies on the development of high-quality RNAi HTS assays. To obtain high-quality HTS assays, there is a strong need for a generally acceptable QC metric that can be applied to various HTS assays (including RNAi HTS assays) conducted in different labs and/or at various times. This metric should have a solid theoretical basis and clear probability meanings. The classical t-test for testing no mean difference cannot work well as a QC metric in RNAi HTS assays as demonstrated under Methods. The currently available QC parameters, \( S/B, S/N, \) and \( Z \)-factor, all have disadvantages. In this paper, a pair of new parameters, SSMD and \( CVD \), is proposed for measuring the magnitude of difference between two groups and is then investigated for evaluating the quality of RNAi HTS assays.

As investigated under Methods, SSMD and \( CVD \) have clear theoretical advantages over the currently used QC parameters, \( S/B, S/N, \) and \( Z \)-factor, as QC parameters in RNAi HTS assays. Compared to \( S/B \) and \( S/N \), SSMD and \( CVD \) capture the variabilities in both compared populations. Compared to \( Z \) factor, SSMD and \( CVD \) have clear probability meanings (represented by \( d^-\)-probability) and solid statistical bases. The SSMD-based cutoff criterion has a solid probability basis, while the \( Z \)-factor-based criterion is more or less empirical. Hence, from a probability and statistics perspective, the statistical inference of SSMD and \( CVD \) is fully theoretically based. The comparison of SSMD and \( Z \)-factor in data from real RNAi HTS assays targeting different diseases further suggests that the use of SSMD as a QC metric in HTS assays leads to more sensible results than using \( Z \)-factor.

The application of SSMD in data from real RNAi HTS assays with various data ranges, data variabilities, and replicate numbers demonstrates that SSMD is robust to different data ranges and data variability. In addition, it gives suggestions on how to use SSMD or \( CVD \) as QC metrics in RNAi HTS assays. Usually, the first question to be faced in RNAi HTS assays is the determination of the number of wells for a positive control and a negative control in a plate to obtain reasonably confident results when SSMD is used as a QC metric. This question can be explored in theory, which can become very complicated. In practice, we frequently have the design of 32, 16, 8, or 4 wells in a 16-by-24 plate for either a negative control or a positive control. The application under Results suggests that 4 wells per plate for a control seems not enough while 8 wells or more may be reasonable, although the more replicates the better the performance of SSMD.

The second question to be faced is how to deal with the situation in which the measured intensity is skewed and has outliers since SSMD and \( CVD \) are based on normality assumption. This is important because it is not unusual for the measured raw intensity to have outliers and to be not normally or even not symmetrically distributed. The first strategy to deal with it is to use transformation to make the data nearly normally (or at least nearly symmetrically) distributed. The commonly used are log-transformation and square-root transformation. The second strategy is to use both regular and robust estimation of SSMD (or \( CVD \)) and double-check the plates that are disagreed by the two types of estimates for outliers.

The third question is how to choose the cutoff value of SSMD for different experiments. In practices, some HTS assays may have only strong positive (or negative) controls while other assays may have only weak positive (or negative) controls. Some assays may even have both strong and weak positive controls in a plate, such as in the HCV RNAi HTS assay [17]. The assay with good quality should have bigger separation between strong positive controls and negative controls than between weak positive controls and negative controls. Thus, to evaluate assay quality, the QC cutoff criterion using strong positive controls should be different from the criterion using weak positive controls. For example, in the situation when positive control should have value bigger than negative controls, when using the strong positive control, we may use the criterion of \( SSMD \geq 6 \) to indicate good quality while we may use the criterion of \( SSMD \geq 3 \) when using weak positive controls. In some situations, we may have only very weak biological positive controls in a whole experiment. In these situations, the criterion of \( SSMD \geq 2 \) may even be enough to indicate good quality.

We have investigated the use of SSMD or \( CVD \) as QC metrics in RNAi HTS assays. SSMD and \( CVD \) are applicable not only in various RNAi HTS assays targeting a variety of diseases including infectious diseases, cancers, and complex neurodiseases, but also in other HTS assays including small-molecule HTS assays. SSMD and \( CVD \) measure the magnitude of difference between two groups in general. Hence, in addition to being used as QC metrics, they may have other applications such as in hit selection in HTS assays. In the future, more theoretical research on SSMD and \( CVD \) can be also done. For example, the data from real experiments seem to suggest that the sample size of 8 or more is reasonably good, whereas a sample size of 4 is not big enough for the use of SSMD for QC in RNAi HTS assays. How can we determine the sample size requirement theoretically? The probability meanings of SSMD and \( CVD \) are based on the normality assumption; and the MLE estimation and properties work best in normal distributions. What will happen if the normality assumption is violated even after applying commonly used transformation? How could we adjust the SSMD or \( CVD \) estimation? In summary, SSMD and \( CVD \) have many other potential applications in HTS assays and more research can be conducted on them.
Acknowledgments

The author thanks Daniel Holder, Keith Soper, and Joseph Heyse for their supports in this research; Daniel Holder, Jason Liao at Merck, and Xiaohua Shu at Temple University for their helpful discussions; and Editor Meldrum and two anonymous referees for their valuable comments on the revision.

Appendix A

Proposition 1

If $X_{i1}, X_{i2}, \ldots, X_{in_i}$, are independently identically distributed (IID) from $N(\mu_1, \sigma_1^2)$, and $X_{j1}, X_{j2}, \ldots, X_{jn_j}$, are IID from $N(\mu_2, \sigma_2^2)$, and both samples are independent of each other, then the MLEs of SSMD and CVD are, respectively,

$$\hat{\beta} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1-1}{n_1} s_1^2 + \frac{n_2-1}{n_2} s_2^2}}$$

$$\hat{\sigma} = \frac{\sqrt{\frac{n_1-1}{n_1} s_1^2 + \frac{n_2-1}{n_2} s_2^2}}{X_1 - X_2},$$

where $\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki}$ and $s_k = \sqrt{\frac{1}{n_k-1} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)^2}$ ($k = 1, 2$) and $\hat{\beta}$ and $\hat{\sigma}$ are asymptotically unbiased estimates and are asymptotically normally distributed. That is, $\hat{\beta}$ and $\hat{\sigma}$ have the following properties:

- consistency (i.e., asymptotic unbiasedness),

$$\hat{\beta} \xrightarrow{p} \beta_0 \quad \text{and} \quad \hat{\sigma} \xrightarrow{p} \sigma_0 \quad \text{as} \quad n_1 \to \infty$$

- asymptotic normality,

$$\hat{\beta} \xrightarrow{d} N(\beta_0, \sigma_\beta^2) \quad \text{and} \quad \hat{\sigma} \xrightarrow{d} N(\sigma_0, \sigma_\sigma^2) \quad \text{as} \quad B_1 \to \infty,$$

where

$$\sigma_\beta^2 = \frac{n_1 + n_2}{n_1 + n_2} + \frac{n_1}{n_1 + n_2} (\mu_1 - \mu_2)^2, \quad \sigma_\sigma^2 = \frac{1}{2(\sigma_1^2 + \sigma_2^2)} + \frac{\sigma_1^2 + \sigma_2^2}{(\mu_1 - \mu_2)^2} \frac{n_1 + n_2}{(\mu_1 - \mu_2)^2}.$$ 

Proof

Let $X = (X_{i1}, X_{i2}, \ldots, X_{in_i}, \ldots, X_{j1}, X_{j2}, \ldots, X_{jn_j})$, $X_1 = (X_{i1}, X_{i2}, \ldots, X_{in_i}, X_{i1}), X_2 = (X_{j1}, X_{j2}, \ldots, X_{jn_j}), \theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2), \delta = (\mu_1, \beta, \mu_2, \sigma_3^2), f_1(X_i; \delta)$ be the distribution of $X_{i1}$, and $f_2(X_j; \delta)$ be the distribution of $X_{j1}$. As shown in many classical textbook, the MLEs of $\mu_1, \mu_2, \sigma_1^2$, and $\sigma_2^2$ are respectively $\bar{X}_1, \bar{X}_2, \frac{n_1}{n_1} s_1^2, \text{and} \frac{n_2}{n_2} s_2^2$. By the invariance property of MLE, the MLE of $\beta$ is $\hat{\beta} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{n_1-1}{n_1} s_1^2 + \frac{n_2-1}{n_2} s_2^2}}$. Thus, we have the MLE of $\beta$ given in Eq. (12).

The log-likelihoods of $X_1, X_2,$ and $X$ are respectively

$$l_1(\delta) = \sum_{i=1}^{n_1} \log f_1(X_{i1}; \delta), \quad l_2(\delta) = \sum_{i=1}^{n_2} \log f_2(X_{j1}; \delta), \quad \text{and} \quad l(\delta) = l_1(\delta) + l_2(\delta).$$

Because $l_2(\delta)$ depends on the parameters of $\mu_2$ and $\sigma_2^2$ only, $\frac{\partial}{\partial \beta} l_2(\delta) = 0$. $\hat{\delta}$ is the MLE of $\delta$ with respect to $l(\delta)$. Thus, $\frac{\partial}{\partial \beta} l(\hat{\delta}) = 0$. Therefore, $0 = \frac{\partial}{\partial \beta} l(\hat{\delta}) = \frac{\partial}{\partial \beta} l_1(\hat{\delta}) + \frac{\partial}{\partial \beta} l_2(\hat{\delta}) = \frac{\partial}{\partial \beta} l_1(\hat{\delta}) + \frac{\partial}{\partial \beta} l_2(\hat{\delta})$. That is, the MLE of $\hat{\delta}$ with respect to $l(\delta)$ is also the
MLE of $\delta$ with respect to $l_1(\hat{\delta})$. Considering that $(X_{11}, X_{12}, ..., X_{1n_1})$ is IID, by the consistency of MLE based on an IID sample, we have $\hat{\theta} \xrightarrow{p} \theta_0$ as $n_1 \to \infty$.

Similarly, using $\delta'=(\mu_1, \sigma_1^2, \mu_2, \beta)$, we can prove that $\hat{\theta} \xrightarrow{p} \theta_0$ as $n_2 \to \infty$.

Let $\theta_1 = \left( \mu_1, \sigma_1^2 \right)^T$ and $\hat{\theta}_1 = \left( \hat{X}_1, \frac{n_1-1}{n_1} \hat{s}_1^2 \right)^T$. By the asymptotical normality and efficiency of MLE from the IID sample $X_1$, we have

$$\sqrt{n_1}(\hat{\theta}_1 - \theta_1) \xrightarrow{d} N(0, I^{-1}(\theta_1))$$

as $n_1 \to \infty$.

where

$$I(\theta_1) = E \left( -\frac{\partial^2}{\partial \theta_1^2} \log f_1(X_{11}; \theta_1) \right).$$

It is trivial to show that

$$I(\theta_1) = E \left( -\frac{\partial^2}{\partial \theta_1^2} \log f_1(X_{11}; \theta_1) \right) = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{2\sigma_1^2} \end{pmatrix}.$$

Therefore, as $n_1 \to \infty$,

$$\sqrt{n_1} \begin{pmatrix} \bar{X}_1 - \frac{n_1-1}{n_1} \hat{s}_1^2 \\ \frac{n_1-1}{n_1} \hat{s}_1^2 \end{pmatrix} \xrightarrow{d} N \left( 0, \begin{pmatrix} 0 & \sigma_1^2 \\ \sigma_1^2 & 2\sigma_1^4 \end{pmatrix} \right).$$

Similarly, we can obtain that, as $n_2 \to \infty$,

$$\sqrt{n_2} \begin{pmatrix} \bar{X}_2 - \frac{n_2-1}{n_2} \hat{s}_2^2 \\ \frac{n_2-1}{n_2} \hat{s}_2^2 \end{pmatrix} \xrightarrow{d} N \left( 0, \begin{pmatrix} 0 & \sigma_2^2 \\ \sigma_2^2 & 2\sigma_2^4 \end{pmatrix} \right).$$

Therefore, considering the two samples are independent, as $n_1$ and $n_2 \to \infty$,

$$\begin{pmatrix} \frac{n_1-1}{n_1} \bar{s}_1^2 \\ \frac{n_2-1}{n_2} \bar{s}_2^2 \end{pmatrix} \xrightarrow{d} N \left( 0, \begin{pmatrix} \frac{\sigma_1^4}{n_1} & 0 & 0 & 0 \\ 0 & \frac{2\sigma_1^4}{n_1} & 0 & 0 \\ 0 & 0 & \frac{\sigma_2^4}{n_2} & 0 \\ 0 & 0 & 0 & \frac{2\sigma_2^4}{n_2} \end{pmatrix} \right).$$

By the Delta method (cf. [23]),

$$\hat{\beta} - \beta_0 \xrightarrow{d} N \left( 0, \frac{\partial^2}{\partial \theta^T} \right) \begin{pmatrix} \sigma_1^2 \\ \frac{2\sigma_1^4}{n_1} \\ \frac{\sigma_2^4}{n_2} \\ \frac{2\sigma_2^4}{n_2} \end{pmatrix} \frac{\partial \beta}{\partial \theta^T},$$

where

$$\frac{\partial \beta}{\partial \theta^T} = \begin{pmatrix} \frac{\partial \beta}{\partial \mu_1} & \frac{\partial \beta}{\partial \sigma_1^2} \\ \frac{\partial \beta}{\partial \mu_2} & \frac{\partial \beta}{\partial \sigma_2^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}}, & \frac{\mu_1 - \mu_2}{(\sigma_1^2 + \sigma_2^2)^{3/2}}, & \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \left( \frac{\mu_1 - \mu_2}{\sigma_1^2} - \frac{\mu_1 - \mu_2}{\sigma_2^2} \right) \end{pmatrix}.$$
By plugging the above partial derivatives into Eq. (16), we obtain Eq. (15). Therefore, the two properties of the MLE of SSMD have been proven.

Using an approach similar to the proof of the part related to SSMD in Proposition 1, we can prove the part of Proposition 1 related to CVD.

References