Single assay-wide variance experimental (SAVE) design for high-throughput screening

Carl Murie[1,2], Caroline Barette[3], Laurence Lafanechère[3,4,5,6] and Robert Nadon[1,2,*]

[1]McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 0G1, Canada, [2]McGill Department of Human Genetics, McGill University, Montreal, Quebec H3A 1B1, Canada, [3]Equipe Criblage pour des Molécules Bio-Actives (CMBA), CEA Grenoble, Grenoble Cedex 09, France, [4]INSERM, U823, [5]Université Joseph Fourier-Grenoble 1 and [6]Institut Albert Bonniot, Grenoble F-38706, France

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Advantages of statistical testing of high-throughput screens include P-values, which provide objective benchmarks of compound activity, and false discovery rate estimation. The cost of replication required for statistical testing, however, may often be prohibitive. We introduce the single assay-wide variance experimental (SAVE) design whereby a small replicated subset of an entire screen is used to derive empirical Bayes random error estimates, which are applied to the remaining majority of unreplicated measurements.

Results: The SAVE design is able to generate P-values comparable with those generated with full replication data. It performs almost as well as the random variance model t-test with duplicate data and outperforms the commonly used Z-scores with unreplicated data and the standard t-test. We illustrate the approach with simulated data and with experimental small molecule and small interfering RNA screens. The SAVE design provides substantial performance improvements over unreplicated screens with only slight increases in cost.

Contact: robert.nadon@mcgill.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

High-throughput screening (HTS) provides concurrent testing of large libraries for detecting biological activity among large numbers of compounds. For ease of exposition, we refer throughout to compounds although the points made apply to other screened objects such as small interfering RNAs (siRNAs). Advances in combinatorial chemistry, robotic processing and plate miniaturization have dramatically increased throughput. Large screens of one million compounds or more, termed ultra-HTS, are increasingly common in industry (Bobkova *et al.*, 2010; Chung *et al.*, 2010; Madoux *et al.*, 2010). This has led to pressure to reduce costs associated with the increased volume of screens.

Techniques that can maintain performance in identifying rare biological events while reducing costs are invaluable in the drug discovery process.

Replicated screens, which allow for statistical testing of biological activity, are becoming more common (Boutros and Ahringer, 2008; http://nsrb.med.harvard.edu/_downloads/ NSRB_newscreener201106.pdf; Malo *et al.*, 2006; Seiler *et al.*, 2008). Statistical testing offers numerous advantages for detecting active compounds relative to ranking or using biologically motivated thresholds, neither of which takes compound variability into account. With statistical testing, estimates of uncertainty can be associated with activity measurements; power analyses can be performed to determine cost-effective number of replicates for future screens; false discovery rate procedures can guide decisions for selecting compounds for follow-up screen validation. Obtaining replicates can be prohibitively expensive, however, particularly for ultra-HTS (Wahome and Mantis, 2001).

One way to reduce costs is to use shrinkage statistical tests that combine individual compound variances obtained from replicates with a screen-wide estimate of the variance. Developed specifically for studies with few replicates, these methods produce more accurate and more precise variance estimates for statistical testing than do classical methods such as the t-test (Allison *et al.*, 2006; Efron, 2005; Tong *et al.*, 2007). The random variance model (RVM), for example, has been shown to be substantially superior to the standard t-test (Murie *et al.*, 2009; Wright and Simon, 2003) and has been successfully applied to HTS data (Malo *et al.*, 2006, 2010). Although the RVM test performs well with as few as two replicates, cost considerations may argue against obtaining even this minimal level of replication for an entire screen.

We introduce the single assay-wide variance experimental (SAVE) design as a means of generating an RVM error estimate that is based on a small replicated subset of a screen, which can be applied to the remaining majority of unreplicated measurements. We show that SAVE produces P-value distributions and rank-order statistics comparable with fully replicated screens for both simulated and experimental data. SAVE also outperforms the popular Z-score and standard one-sample t-test. Finally, we provide guidelines for determining the number of compounds and replicates required to produce error estimates for use within the SAVE design.

*To whom correspondence should be addressed.

2 METHODS

2.1 Z-score

The *Z*-score normalizes for additive and multiplicative effects across plates:

$$Z = \frac{x_i - \mu_p}{\sigma_p} \tag{1}$$

where x_i is the signal intensity of the ith compound and μ_p and σ_p are the mean and standard deviation, respectively, of the raw well intensities per plate excluding controls. *P*-values are derived from the standard normal distribution.

We note that *Z*-scores are linear transformations of other normalization methods that likewise apply within-plate arithmetic operations (e.g. percent of control, normalized percent inhibition/activation and plate median normalization). We note also that *Z*-scores are not to be confused with the quality control *Z* and *Z'* factor scores (Zhang *et al.*, 1999) commonly used by HTS screeners. Although selection of hit detection thresholds differs (e.g. compounds with *Z*-scores exceeding 2 or 3 are often considered active, whereas thresholds for normalized percent inhibition/activation are tailored to each assay), scores of the various methods correlate one on a plate-by-plate basis.

2.2 Spatial polish and well normalization (SPAWN)

SPAWN normalization is a two-step method, which first uses a trimmed mean polish (here set to a trim value of 0.2) on individual plates to remove row and column effects. The first step fits the assay plates to the following model:

$$y_{ijp} = \mu_p + R_{ip} + C_{jp} + e_{ijp} \tag{2}$$

where y_{ijp} is the well intensity for the ith row and j^{th} column of the pth plate. μ_p is the grand mean of the pth plate, R_{ip} is the ith row effect and C_{jp} is the jth column effect of the pth plate. e_{ijp} are the residuals after removing the grand mean, row and column effects. The residuals are rescaled by dividing by the median absolute deviation of the plate and are used as the post-normalized well scores. A second (optional) individual well normalization step used here calculates a spatial bias template (SBT_{ij}), which is the median of the scores at each well location, the ith row and j^{th} column, across all plates. The spatial bias template scores are then subtracted from the scores from step 1 at each well location ($\hat{y}_{ijp} = y_{ijp} - SBT_{ij}$). The resulting scores are again rescaled by dividing by the median absolute deviation of each plate.

2.3 Standard one-sample t-test

The one-sample *t*-test is calculated as follows:

$$t = \sqrt{n}\frac{\hat{\mu} - c}{\hat{\sigma}} \tag{3}$$

where $\hat{u}$, $\hat{\sigma}$ and n are the compound average, compound standard deviation and number of replicates, respectively. c is the measure of null activity, typically zero. *P*-values are derived from the *t* distribution with *n-1* degrees of freedom.

2.4 RVM one-sample t-test

In the RVM test, the empirical Bayes posterior variance for each compound is obtained based on a weighted combination of the compound-specific variance and a prior variance estimate that is derived from the variances of all the replicated compounds in the screen. The shape (*a*) and scale (*b*) parameters for the prior variance distribution are estimated by fitting the observed variances of all compounds to an inverse gamma distribution. The RVM one-sample *t*-statistic takes the form:

$$\tilde{t} = \sqrt{n}\frac{\hat{\mu} - c}{\tilde{\sigma}_{pool}} \tag{4}$$

where $\hat{u}$, n and c are as for the standard one-sample *t*-test. The posterior, or pooled, variance is

$$\tilde{\sigma}_{pool}^2 = \frac{(n-1)\hat{\sigma}^2 + 2a(ab)^{-1}}{n - 1 + 2a} \tag{5}$$

where $\hat{\sigma}^2$ is the compound-specific sample variance and $(ab)^{-1}$, the prior variance estimate, is the mean of the fitted inverse gamma distribution. A large number of replicates give increased weight to the sample variance. A large shape value (*a*), which indicates that the inverse gamma is highly peaked (i.e. most compound-specific variances are close to the mean), gives increased weight to the prior variance. *P*-values are derived from a *t* distribution with $(n-1) + 2a$ degrees of freedom.

2.5 SAVE one-sample t-test

In the SAVE design, a subset of plates within a screen is randomly selected for replication. An RVM one-sample *t*-test prior (assay-wide) variance estimate is obtained for the replicated subset. This estimate is then used to calculate the error term for the statistical tests of the unreplicated compounds in the screen. The SAVE one-sample *t*-test is calculated as:

$$SAVEt_i = \frac{x_i - c}{\sqrt{\frac{2a(ab)^{-1}}{2a}}} = \frac{x_i - c}{\sqrt{1/(ab)}} \tag{6}$$

where x_i is the ith compound of the unreplicated data and c is the measure of null activity, typically zero. a and b are the RVM shape and scale parameters, respectively, of the inverse gamma distribution fitted to the replicated subset. The assay-wide variance is the mean of the fitted inverse gamma distribution, defined as $1/(ab)$. *P*-values are derived from a *t* distribution with $2a$ degrees of freedom.

2.6 Data

2.6.1 Simulated data methods Two sets of simulated data were generated. Simulation A examined the effect of compound number and replicate number on the prior variance distribution parameter estimates and the assay-wide variance used in the SAVE *t*-test. A set of simulated compounds with normal error, whose variances were drawn from an inverse gamma distribution with defined shape (*a*) and scale (*b*) parameters, was generated with a given number of replicates. The shape, scale and assay-wide variance [mean of the prior inverse gamma variance distribution $= 1/(ab)$] estimates for each simulation run were compared with the true parameter values. The ranges of the simulated data parameters were as follows: 1–10 plates (80 compounds/plate), 2–11 plate replicates and shape (*a*) and scale (*b*) parameters that ranged from 1 to 10. Each combination of parameters was simulated 1000 times.

For Simulation B, *P*-values were derived from one-sample *t*-tests and SAVE *t*-tests of both raw and *Z*-score data ($n = 2$), RVM *t*-tests of raw data ($n = 2$) and *Z*-scores ($n = 1$). Simulation B consisted of two separate datasets. For Simulation B1, 1000 population variances were first drawn from an inverse gamma distribution with specified shape values ($a = 1, 2,$ or 3) chosen to be comparable with values commonly found in high-throughput screens (Supplementary Information Section 1); scale value (*b*), which had no effect on output (see Results), was set to 1. Two replicate intensities were then generated for each compound by drawing from a normal distribution with the corresponding population variance. Ninety percent of the compounds were simulated to be inactive ($\mu = 0$) and the remaining 10 percent were simulated to be active (with effects ranging continuously from −2 to −3 and from 2 to 3 standard deviations away from the null value of 0).

Simulation B2 was used to generate SAVE assay-wide variance estimates for normally distributed data with the population shape and scale parameters used in B1. The number of compounds for B2 ranged from 80 to 480 (1–6 plates with 80 compounds each), and the number of replicates (*n*) ranged from 2 to 7. Sample variances for the replicates for each

compound were generated by drawing randomly from the χ^2 distribution with $n-1$ degrees of freedom and applying the following formula (Hays, 1994):

$$\hat{\sigma}_i^2 = \frac{\sigma_i^2 \chi^2(n-1)}{n-1} \tag{7}$$

where σ_i^2, $\hat{\sigma}_i^2$ and n are the ith (compound) population variance, sample variance and number of replicates, respectively. The SAVE assay-wide variance estimate [the denominator in Equation (6)] was calculated from the set of sample variances for each simulation run. Simulations B1 and B2 were repeated in parallel 1000 times for each combination of replicate number, plate number, shape (a) and scale (b) value.

2.6.2 Experimental data methods P-values were derived from one-sample t-tests and SAVE t-tests of both Z-score and SPAWN normalized data ($n=2$), RVM t-tests of SPAWN scores ($n=2$) and Z-scores ($n=1$). Two experimental datasets, an in-house (Centre de Criblage pour des Molecules Bio-Actives; CMBA) small molecule cell-based primary screen (Supplementary Information Section 2) and an siRNA screen (Wiles *et al.*, 2008), were examined. The CMBA screen, which contained 520 active and 40 inactive compounds in seven 96-well plates, was run in quadruplicate. The Wiles *et al.* data, which contained 22 915 siRNAs of unknown activity levels in 62 384-well plates, were run in duplicate.

P-values were generated with the experimental data by first creating two mutually exclusive sets (1 and 2) for each assay. Set 1, the replicated subset of the assay as described in Section 2.5, was used to generate an RVM prior (assay-wide) variance estimate. The process was iterated across all combinations of replicate and plate numbers. Assay-wide variance estimates from Set 1 were used in the SAVE t-test of each single replicate compound in Set 2. Results of the SAVE t-tests were compared with results of other tests applied to the same compounds.

3   RESULTS

3.1   Quality of prior variance distribution parameter estimates

Simulated and experimental data were used to examine effects of compound and replicate number on the accuracy and precision of SAVE assay-wide variance estimates and to provide guidance for cost-effective experimental designs.

3.1.1 Simulated data—quality of parameter estimates   Simulated data were used to examine the effect of numbers of plates and replicates across a range of shape (a) and scale (b) parameters on the estimation of the prior variance distribution parameters (Section 2.6.1).

Shape (a) values of 1–3 are commonly found in HTS assays when estimating the assay-wide variance distribution (Supplementary Information Section 1). Figure 1, which shows simulation results for true shape values within this range, indicates that the accuracy of the SAVE shape estimates was independent of true shape value, number of replicates and number of plates. Precision of the shape estimates, by contrast, improved with increasing plate numbers, although the extent of the gains depended on both number of replicates and true shape values. At true shape = 1, increasing replicates provided little or no gains in precision; at true shapes = 2 and 3, large gains in precision were observed by increasing replicates from 2 to 3 but not from 3 to 4. The full set of graphs showing the number of replicates from 2 to 11 and true shape values from 1 to 10 are in Supplementary Information Section 3. The true scale value had no effect on
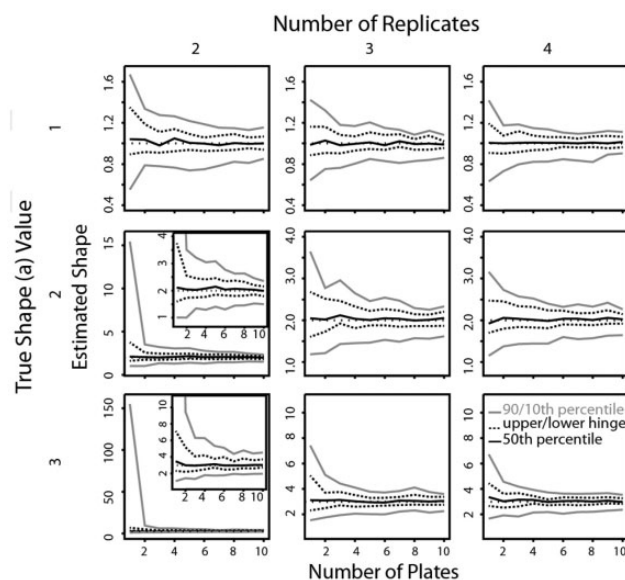
Fig. 1. SAVE estimated shape (a) values of simulated data by number of replicates, true shape and number of plates (80 wells/plate) at scale (b) = 1

either the accuracy or precision of the estimated shape parameters (Supplementary Information Section 4); the true shape value, plate number and replicate number had little to no effect on the accuracy and precision of the scale and assay-wide variance estimates (Supplementary Information Section 5).

3.1.2 Experimental data—quality of parameter estimates   The quality of the SAVE variance estimates was investigated with experimental data by comparing all possible subsets of number of plates and replicates from the small molecule CMBA and the siRNA Wiles assays to their assay-wide variances estimated from their respective full datasets.

Figure 2 shows that SAVE produced accurate estimates of shape (a) for the CMBA data with two or more plates (all replicate numbers) and with good precision at four replicates. Accurate estimates with good precision were obtained for the Wiles *et al.* data with four or more plates [see Supplementary Information Section 6 for scale (b) and assay-wide variance ($1/ab$) estimates].

3.2   *P*-value distributions

Simulated and experimental data were used to generate P-values derived from the various methods to assess their respective performances.

3.2.1 Simulated data—P-value distributions   Figure 3 shows P-values for the six methods generated with simulated data that included 10% active compounds (Section 2.6.1). Figure 3a shows that four of the methods (RVM t-test, t-test, t-test with Z-scores and SAVE) met the theoretical expectation of uniformly distributed P-values for inactive compounds, whereas two methods did not (Z-scores and SAVE with Z-scores). For P-values generated by the active compounds, SAVE showed P-values almost as small as those generated by RVM, increasingly so as shape values increased, and smaller than for any other method
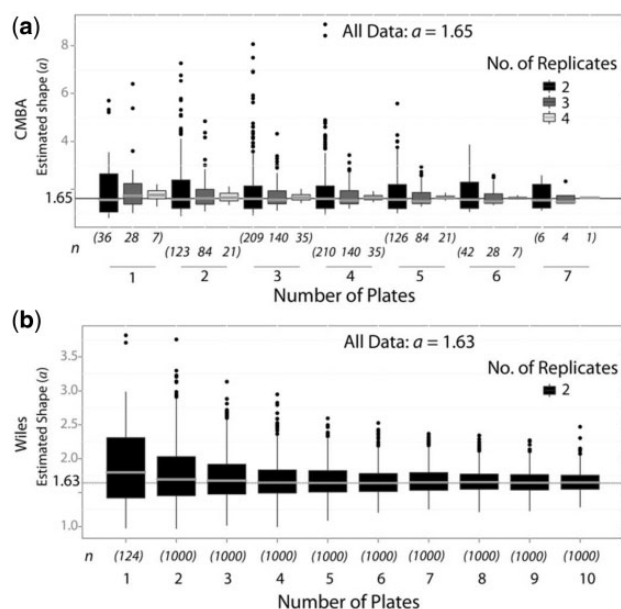
Fig. 2. Boxplots of SAVE estimated shape (*a*) values of experimental data normalized with SPAWN. (**a**) CMBA data (96-well plates with 80 non-control wells per plate). Ten extreme outliers from 1, 2 or 3 plates (two replicates) were removed for scaling purposes. (**b**) Wiles data (384-well plates with 320 non-control wells per plate) in duplicate

(Fig. 3b). These results indicate that SAVE generates correct *P*-values under the null hypothesis with almost as much statistical power as RVM and with more power than the *t*-test at considerably reduced cost in that, unlike with the *t*-test, only a small subset of the screen need be replicated.

3.2.2 Experimental data—P-value distributions Obtaining correct *P*-values with experimental HTS data is a more difficult task because of the likely presence of numerous sources of bias in the measurements. Normalization methods must remove the bias while simultaneously maintaining statistical test assumptions (e.g. normally distributed error for the *t*-test, RVM *t*-test and SAVE and the added assumption of inverse gamma distributed variances for RVM and SAVE).

Figure 4a and b show that when normalized with SPAWN, the *t*-test, RVM *t*-test and SAVE tests produced similar *P*-values with both the CMBA small molecule and the Wiles *et al.* RNA interference datasets. Importantly, all three tests showed the expected uniformly distributed *P*-values for inactive CMBA molecules (Fig. 4c). As with the simulated data, SAVE generated *P*-values almost as small as RVM for active molecules and outperformed the *t*-test (Fig. 4d). All three tests performed poorly with *Z*-score normalized data.

3.3 Power analysis of SAVE

Figure 5 shows the power levels of the SAVE method at a *P*-value threshold of 0.01 with shape values ranging from 1 to 3 and the scale parameter set to 1. The power levels were calculated using a standard power analysis with the SAVE assay-wide variance used as the variance estimate. The power of the SAVE method improved as the shape value of the inverse gamma
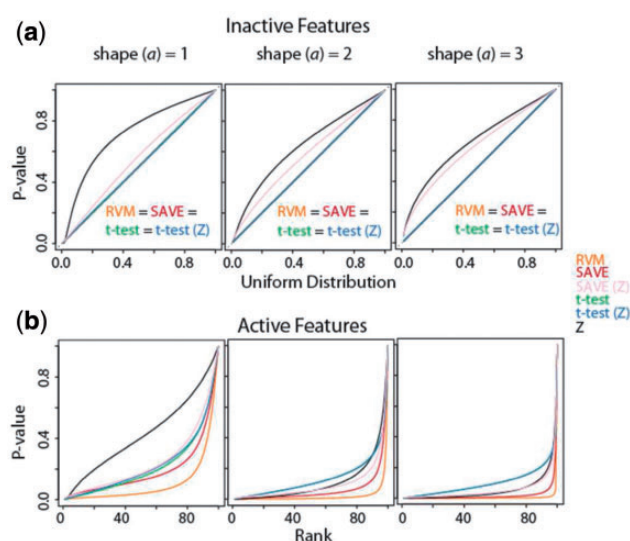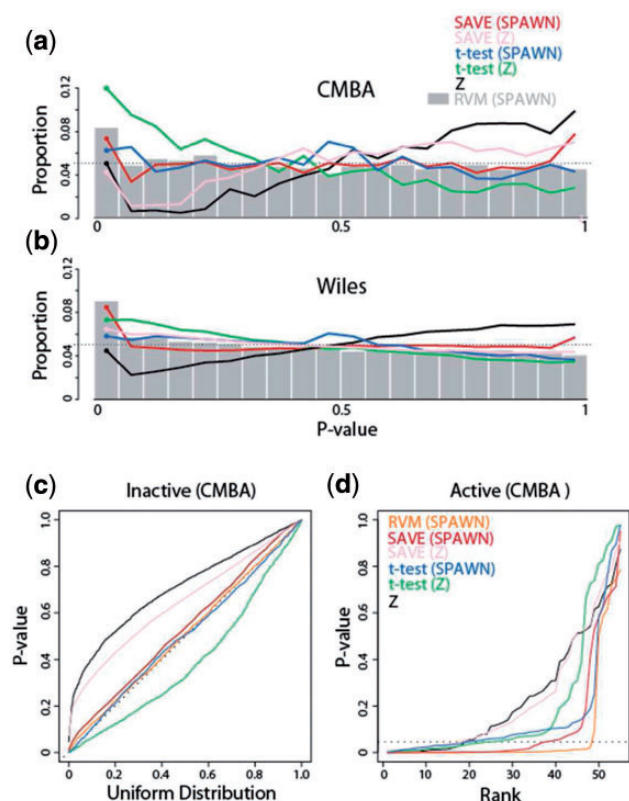

Fig. 3. Two-tailed *P*-values generated with simulated data. The SAVE assay-wide variance was generated with 240 compounds (equivalent to three 96-well plates, 80 compounds each) and 4 replicates. The RVM *t*-test, *t*-test and *t*-test with *Z*-scores were generated with 1000 compounds and 2 replicates; the SAVE, SAVE with *Z*-scores and *Z*-scores were generated with 1000 compounds and 1 replicate. The simulated data variances were generated from a prior inverse gamma variance distribution with shape values 1–3 and a scale value of 1. (**a**) QQ plot of *P*-values of the inactive compounds compared with the uniform distribution. (**b**) Rank-ordered *P*-values of the active compounds

distribution increased. This is due to the compound variances clustering more strongly around the center of the variance distribution at higher shape values. The power analysis of the shape parameter at *P*-value thresholds of 0.001 and 0.05 are shown in Supplementary Information Section 7.

4 DISCUSSION

Statistical testing and concomitant *P*-values provide an objective way to benchmark false-positive and false-negative rates in HTS assays. The SAVE method can be used to generate *P*-values that are comparable with those generated with full replication, providing a reasonable alternative when the latter is not possible due to cost considerations. Replication that captures the true variability of the assay rather than only technical variation is the preferred approach (e.g. by replicating sample preparation and by randomly assigning replicated compounds to different well locations across plates). This approach to the minimal replication required of the SAVE design would improve generalizability of results and their validation.

As with other statistical approaches, SAVE results are best when data are adequately normalized, which is often not the case with frequently used methods such as the *Z*-score or normalized percent inhibition/activation. As a consequence of not correcting for spatial biases within plates, compound activity with data normalized by these latter methods will appear higher or lower than the true biological signal, depending on where the compound is located on a plate (Brideau *et al.*, 2003). An additional problem of using *Z*-scores for purposes of generating

(a) CMBA

SAVE (SPAWN)
SAVE (Z)
t-test (SPAWN)
t-test (Z)
Z
RVM (SPAWN)

(b) Wiles

Fig. 5. Power levels of the SAVE design across effect sizes [measured intensity difference from zero divided by the SAVE assay-wide variance $(1/ab)$] and different shape parameters of the prior variance distribution at a critical *P*-value threshold of 0.01

(c) Inactive (CMBA)   (d) Active (CMBA)

RVM (SPAWN)
SAVE (SPAWN)
SAVE (Z)
t-test (SPAWN)
t-test (Z)
Z

Fig. 4. Two-tailed *P*-values generated with experimental data. The SAVE assay-wide variance was generated with three plates and four replicates for the CMBA data and two plates and two replicates for the Wiles data. The *t*-test and RVM *t*-test *P*-values (*Z*-score and SPAWN normalized) were generated with all $n = 2$ combinations; SAVE (both *Z*-score and SPAWN normalized) and *Z*-scores were generated with $n = 1$. (**a** and **b**) Histogram of *P*-values for CMBA and Wiles assays generated with two-tailed tests. (**c**) QQ plot of *P*-values of the CMBA inactive compounds compared with the uniform distribution. (**d**) Rank-ordered *P*-values of the CMBA active compounds

P-values is the incorrect assumption that all compounds share a common population variance (Malo *et al.*, 2006). The incorrectness of the *P*-values associated with *Z*-scores can be readily observed by their deviations from a uniform distribution for inactive molecules. This graphical approach can be used to assess SAVE's appropriateness for datasets and normalization methods not addressed in our study; when used with primary screen data of mostly inactive compounds, *P*-values should be approximately uniformly distributed in all but the lower *P*-value range. We demonstrated that SPAWN normalization provides good results in line with statistical expectation; other methods that similarly correct for spatial bias, such as the R score (Wu *et al.*, 2008), may also be considered.

The standard *t*-test has been widely criticized for use with high-throughput studies because of its low power and poor random error estimation when applied to screens with few replicates (Hu and Wright, 2007; Murie *et al.*, 2009; Qin and Kerr, 2004; Tong and Wang, 2007; Xie *et al.*, 2004). We showed that SAVE error estimates provided by partial replication are
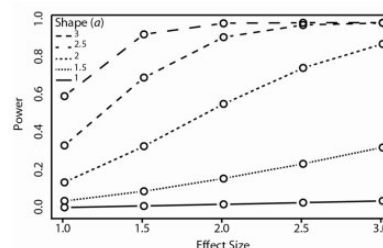
superior to full replication *t*-test estimates. This superiority is reflected in part by the typically larger degrees of freedom associated with the SAVE method. Standard *t*-tests with $n = 2$ provide only one degree of freedom (*df*), whereas the SAVE *t*-test provides $2a$ *df*s. Shape parameter values (*a*) of 1 through 3 are common. In the two empirical datasets we examined, for example, *a* was ~ 1.6, providing 3.2 *df*s for the statistical test. Thus, by using the assay-wide variance as the random error estimate rather than the individual feature compound error estimates of the standard *t*-test, SAVE provides greater statistical power that comes with more *df*s, while avoiding the unrealistically 'small denominator' problem of the *t*-test observed in high-throughput data (Tusher *et al.*, 2001).

Obtaining accurate and precise shape estimates is crucial to the success of the SAVE approach because it is the critical variable in estimating the prior variance distribution and because it defines the degrees of freedom used in generating *P*-values (Smyth, 2004; Wright and Simon, 2003). In general, SAVE shape estimates improve with increasing numbers of plates and replicates but the overall number of plates required is small relative to the number of plates in a typical screen, in particular for ultra-HTS screens for which the SAVE design is most useful. Finally, although SAVE was created for and tested with HTS assays, the method is readily transferrable to other types of high-throughput assays whose variances are distributed according to an inverse gamma distribution (e.g. gene expression microarrays).

ACKNOWLEDGEMENTS

REFERENCES

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Bobkova,E.V. *et al.* (2010) Discovery of PDK1 kinase inhibitors with a novel mechanism of action by ultrahigh throughput screening. *J. Biol. Chem.*, **285**, 18838–18846.

Boutros,M. and Ahringer,J. (2008) The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.*, **9**, 554–566.

Brideau,C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.*, **8**, 634–647.

Chung,N. *et al.* (2010) A 1,536-well ultra-high-throughput siRNA screen to identify regulators of the Wnt/beta-catenin pathway. *Assay Drug Dev. Technol.*, **8**, 286–294.

Efron,B. (2005) Fifty Years of Empirical Bayes (lecture, Université de Montréal, Montreal, QC, September 23).

Hays,W.L. (1994) *Statistics*. Harcourt Brace & Co., Forth Worth, TX.

Hu,J. and Wright,F.A. (2007) Assessing differential gene expression with small sample sizes in oligonucleotide arrays using a mean-variance model. *Biometrics*, **63**, 41–49.

Madoux,F. *et al.* (2010) Small molecule inhibitors of Wee1 degradation and mitotic entry. In: *Probe Reports from the NIH Molecular Libraries Program*. Bethesda, MD, http://www.ncbi.nlm.nih.gov/books/NBK47352/ (25 June 2013, date last accessed).

Malo,N. *et al.* (2006) Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.*, **24**, 167–175.

Malo,N. *et al.* (2010) Experimental design and statistical methods for improved hit detection in high-throughput screening. *J. Biomol. Screen.*, **15**, 990–1000.

Murie,C. *et al.* (2009) Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, **10**, 45.

Qin,L.X. and Kerr,K.F. (2004) Empirical evaluation of data transformations and ranking statistics for microarray analysis. *Nucleic Acids Res.*, **32**, 5471–5479.

Seiler,K.P. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.

Smyth,G. (2004) Linear models and Empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 3.

Tong,T.J. and Wang,Y.D. (2007) Optimal shrinkage estimation of variances with applications to microarray data analysis. *J. Am. Stat. Assoc.*, **102**, 113–122.

Tong,D.D.M. *et al.* (2007) Application of a mixture model for determining the cutoff threshold for activity in high-throughput screening. *Comput. Stat. Data Anal.*, **51**, 4002–4012.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

Wahome,P.G. and Mantis,N.J. (2001) High-throughput, cell-based screens to identify small-molecule inhibitors of ricin toxin and related category B ribosome inactivating proteins (RIPs). In: *Current Protocols in Toxicology*. John Wiley & Sons, Inc, doi: 10.1002/0471140856.tx0223s55.

Wiles,A.M. *et al.* (2008) An analysis of normalization methods for drosophila RNAi genomic screens and development of a robust validation scheme. *J. Biomol. Screen.*, **13**, 777–784.

Wright,G.W. and Simon,R.M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**, 2448–2455.

Wu,Z.J. *et al.* (2008) Quantitative assessment of hit detection and confirmation in single and duplicate high-throughput screenings. *J. Biomol. Screen.*, **13**, 159–167.

Xie,Y. *et al.* (2004) A case study on choosing normalization methods and test statistics for two-channel microarray data. *Comp. Funct. Genomics*, **5**, 432–444.

Zhang,J.H. *et al.* (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.*, **4**, 67–73.