Supplemental material

Algorithm for computation of **μ**

The vector **μ**, the ratio of activity levels for vehicle controls to activity levels for positive controls, is calculated as follows.

For assays with plates in format 1 or format 2 (all assays other than the Jurkat cell assay), a vector of vehicle control values and a vector of positive control values are computed for each plate.

The vehicle control vector consists of the 32 values from the column of vehicle controls (either column 3 or column 4, depending on format).

The positive control vector also has 32 values, each of which is the minimum for one of the 32 rows on the plate of the positive control data point in column 3 or 4 (depending on plate format) and the activity values in the 44 wells used for study substances. For the assay using NIH 3t3 cells, the 100 μM tamoxifen positive control actually produces much less loss of viability than do the most active study substances at their highest concentrations, so it was necessary to take the minimum row value to get an accurate description of the minimal possible measured activity. The minimum row value was used for the other assays to account for possible similar effects. Also, for the NIH 3T3 assay, the strong responses occurred mostly only for the two highest concentrations, so only the values from plates 15 and 16 (containing the two highest concentrations) were used. For the other assays, all plates were used.

For each of the 2 (for the NIH 3T3 assay) or 18 (other assays in format 1 or 2) plates used, the ratio (vehicle control activity/positive control activity) was calculated. The log of that ratio was regressed against a V-shaped length-32 vector [16 15 … 2 1 1 2 … 15 16] plus an intercept term and the regression fit was used as the value of **μ** in the algorithm. Because there are some positive control wells with very low values, there are a few points where the neutral:positive ratio is much lower than elsewhere on the plate. Using the log ratio reduces the effect of these low values. Figure S1 shows the actual and fitted values of the ratios.

For the Jurkat cell assay, full columns of vehicle control data were not available. The **μ** for the calculations using that assay's data was computed by computing the ratios for all other assays, as above, and then performing the regression, also as above, under the assumption that the Jurkat cell assay would be somewhat similar to the other assays.

Effects of varying cutoff in parameter-reduction step

This section considers how the results of the model vary when the p-cutoff parameter from step 7 of the optimization algorithm is allowed to vary. Results will be given for cutoffs of p=0.05, and 0.05/1408, as well as for a model with no cutoff.

Quality of fit

The quality of fit can be studied by looking at the value of the error standard deviation σ. Table 1 shows ratios of the value of σ for a given cutoff value to the σ from the runs with no cutoff. The ratio increases as the cutoff decreases. This is to be expected since the cutoff is a constraint on the optimization and decreasing the cutoff makes the constraint more stringent, thereby reducing the number of parameters available to fit the data.

Strength, significance, and activity classification of modeled responses

Tables 2-4 give measures of the computed significance of the modeled concentration-response and the strength of that response. Tables 2 and 3 show that the number of responses significant in the model results drops as the cutoff becomes stricter. (The third columns of tables 2 and 3 are identical. In step 7, all substances with p-values over the cutoff are forced to have v=0. Thereafter, the p-values for those substances are all 1. So when looking at the p-values at the end of the optimization, no p-values can be between the cutoff value and 1. Thus any response significant at $p<0.05$ with cutoff 0.05/1408 is also significant at $p<0.05/1408$.) Table 4 shows that over half responses for each assay except the NIH 3T3 assay are forced to have v=0 when the cutoff is 0.05. (When there is no cutoff, no responses have v exactly = 0, though some are very close.) However, as Table 5 shows, the number of substances classified as active does not change as much when the cutoff changes. Table 6 shows the classification by response strength and activity over all assays for each cutoff level. The main effect of increasing the strictness of the cutoff is to move responses from the "weak and not active" category into the "no response" category.

Comparison of replicates

Table 7 shows the correlation of the parameters v, k, and n and of the response at the highest concentration for the duplicate substances across all assays when the value of the p-cutoff is varied. (In all cases, "correlation" refers to the Pearson correlation coefficient.) The correlation is computed for duplicate pairs for which both duplicates are classified as active. The correlation of v changes hardly at all, while the other correlations increase slightly as the cutoff becomes stricter. Table 8 shows the concordance between k values as discussed in the main text. The concordance increases slightly as the cutoff becomes stricter.

Parameter limits

Table 9 shows how often the Hill function parameters were at their upper or lower limiting values in the optimization algorithm. Making the p-cutoff stricter decreases the number of v values at the lower limit, since it forces nonsignificant responses to zero and most of the fits with v at the small lower limit are nonsignificant. Because the criteria for classification as active includes a requirement for strength of response, no fit classified as active has v at the lower limit. Fits with v at the upper limit of 1 tend to be the stronger

concentration responses, which are less likely to be forced to zero, so the number of such fits changes less with the cutoff. Making the p-cutoff stricter also decreases the number of k and n values at either the upper or lower limits. The change is less for the active fits. Many n values for the active fits are at the upper limit. In some cases, the concentration-response curve has a number of points with more or less zero response, followed by only one high response at the highest concentration or two nearly identical responses at the two highest concentrations. In those cases, where the concentration spacing in the experiment leaves out points with intermediate response, there is really not enough information to accurately estimate the Hill shape parameter n, which can become arbitrarily large and still fit the data well.

Simulation studies

A simulation study was carried out to examine the performance of the algorithm under simulated conditions where the "true" values of the study parameters are known.

The concentration-response model used in the simulation was the same as that used in the study (equation (3)), but the simulation used 96-well plates (8 rows by 12 columns). Using a smaller plate reduces the time needed to perform the optimization of model parameters and makes it feasible to run multiple simulations. The first column in the simulated plate was of vehicle control data. The ratio μ of vehicle/positive control values was fixed to be 10 for all rows, rather than varying across rows as in the actual data analysis, and it was assumed to be a known value, rather than being calculated before the parameter optimization. This was done because it was thought that the effects of varying μ were not an important part of how well the analysis performs in terms of sensitivity and specificity. The simulated data were generated for 15 plates with the same concentration levels as were used in the actual screening.

Simulated data were generated according to equation (3) (for the concentration-response model) and then error was added (as in equation (6)). The concentration-response model requires parameters for the plate effects (row and column factors α and γ) and the concentration-response (Hill function parameters v, k, and n). The error added was normally distributed with mean 0 and σ=100.

The row and column factors α and γ were computed randomly for each plate.

For α, generate uniform random variables $f_0 \sim U(1900,2100)$, $f_1 \sim U(-2,2)$, and $f_2 \sim U(200,400)$. Then $\alpha_i = f_0 + f_2\left(\frac{i - f_1 - 4.5}{|f_1| + 3.5}\right)^2$. For γ, generate a uniform random variable $g_1 \sim U(-2,2)$. Then $\gamma_j = 0.95 + 0.1\left(\frac{j - g_1 - 6.5}{|g_1| + 5.5}\right)^2$. The values for γ are then divided by γ_1 so that γ_1=0 and the values of γ for the other columns are

relative to those in the first column. These formulas generate a control response which is higher on the edges than on the interior of the plate. The location of the lowest control response on the plate, the range between the highest and lowest responses, and the overall strength of the control response vary depending on the values of the random variables.

The parameters for the Hill function concentration-response model were generated in a multistep process. First, the response at maximum concentration was generated according to an empirical distribution resembling that found in the model results when run with no cutoff in the parameter-reduction step (step 7). That distribution is shown in Figure S2. This distribution is a discrete distribution with 88 elements. The distribution was not randomly sampled; instead, each element of the distribution was used once.

Next, the parameters n and v were randomly generated for each of the 88 responses, with distributions chosen to resemble those seen in the model results. Finally, k was chosen for each response, with k calculated to give the chosen response at maximum dose given the generated values of v and n. This gives 88 sets of parameters (v, k, n). The same parameters were used in 100 runs of the simulation, with the well locations corresponding to the parameter triplets randomized for each run. In each of the 100 runs, the optimization was performed with no parameter reduction step (step 7) or with a p-value cutoff of 0.05 or 0.05/1408 in the parameter reduction step. The model results include predicted Hill function parameters as well as p values from the likelihood ratio test.

A simulation was also performed to examine the effects of replication on the performance of the high throughput screening using the given model. In this simulation, each simulated experiment had three plates at each concentration level. The Hill function parameters were generated as above and were the same for each of the three plates in each of the 100 runs of the simulation, but the added normal error and the plate effects (row and column factors α and γ) were generated separately for the replicates. The optimization was the same as described in the main text, except that the Hill parameters for all three plates were identical and were optimized simultaneously.

To examine the effects of confounding between concentration-response effects and plate location effects, simulations with one or three plates per concentration level were also carried out with stronger simulated responses. Those simulations were the same as described above except for the choice of response at maximum concentration. The highest 44 fixed values from the discrete maximum-response distribution shown in Figure S2 were each used twice to give 88 non-zero responses for the non-control cells of the simulated 96-well plate. The results from these simulations are referred to under the heading “high response” below.

Sensitivity and specificity

The sensitivity (fraction of true positive responses predicted as positive) and specificity (fraction of true negative responses predicted as negative) for the simulation under various conditions are shown in Table 10. A true positive is a substance with true Hill parameter $v>0$; a true negative has true $v=0$. A predicted positive is a substance with p

value (from the likelihood ratio test) less than a given test threshold and a predicted negative is one with p-value greater than or equal to the threshold. Two run parameters can be varied: the cutoff used in the parameter reduction step (step 7 of the optimization) and the p-value threshold for the test of significance. Reducing the test threshold can only have the effect of reclassifying predicted positives as predicted negatives. Doing so can keep the sensitivity the same, or decrease it; it keeps the specificity the same, or increases it. For the "high response" results, there are no true negative results, so there can be no measure of specificity.

With no cutoff or with cutoff 0.05, changing the threshold from 0.05 to 0.05/1408 in this simulation greatly increases the specificity while producing a smaller decrease in the sensitivity. Decreasing the cutoff value from no cutoff to 0.05 to 0.05/1408 also decreases the sensitivity and increases the specificity. The sensitivity and specificity values from the simulation with 3 replicates are about the same as or noticeably better than those from the simulation with only 1 replicate. The simulations with the high response show a somewhat greater sensitivity than the others.

Correlation of parameters

Table 11 shows the correlation of parameters, over 100 simulation runs, between the optimized parameters from the simulation and the parameters used to generate the simulated data. When all of the positive responses are used or only the weakest responses are excluded in the correlation calculation, the correlation is much greater for v and the response at high concentration than for k. The correlation for n is better than that for k but not as high as for the other two variables. This pattern changes when only the strongest responses are considered. In that case, k has the highest correlation. The correlations from the 3-replicate simulation are generally nearly the same as or higher than those from the study with 1 replicate. The correlations for the "high response" simulations are mostly less than those from the other simulations. This is especially noticeable for the v parameter. It seems that the effect of using only substances with positive concentration-responses is to reduce the ability of the method to correctly model the parameters of the concentration-response curve. Examination of the same correlations using parameters from the algorithm with no p-cutoff (results not shown) also shows a reduction in the correlations, though not to the same extent. The p-cutoff in the parameter reduction step of the algorithm acts to reduce the number of false positives in the algorithm and helps to distinguish between true plate location effects and the effects of noise by forcing small and therefore nonsignificant responses to zero. This effect cannot happen if there are no true negatives in the experiment. The simulation shows that, in a sense, the non-responding substances on the plate can act like additional neutral controls.

Another measure of accuracy of fit of the parameters is the fraction of fitted k values which are within a factor of 2 of the true k parameters. This fraction was computed for all true positives classified as positive by the fitting algorithm. For the simulation with 1 replicate, including inactive substances, 0.70 of the k values were within a factor of 2 of the true values. This decreased to 0.68 for the high-response simulation. For the

simulation with 3 replicates, the fractions of similar k values were 0.67 and 0.69. These values are in the range of the same similarity measure for the duplicate substances in the fits to the real data.

Comparison of fits to two normalizations

Normalized data for the 15 assays were obtained from the NCGC and fitted using the Hill function model

$$f(d) = v_0 + \frac{(v - v_0)d^n}{k^n + d^n}$$

for each substance/assay combination. Points indicated as outliers in the analysis were not used in the fit.

Figure S3 shows the normalized data values from the NCGC and from this normalization. Data from this normalization use the same p-cutoff value (0.05/1408) as in the main paper.
The normalized data values were very similar. The largest difference is at the lower end of responses, where this algorithm's normalization using plate location effects tends to cluster values more closely around a value of -1 for full suppression of response, while the previous analysis based on plate control values includes more highly negative responses, presumably because the response for a given compound may be noticeably lower than that for the positive control on a given plate.

Figure S4 compares the fitted values from the fit to the NCGC normalized data and the values from the fit using this paper's normalized data. The fitted values are similar for the two normalizations, with the exception of some values where the reduction of parameter space step in this algorithm fixes the fitted response at 0.

It is possible for multiple sets of parameters to give fits of statistically equivalent quality to a given data set, i.e., there may be several parameter sets whose likelihoods are not significantly different using the likelihood ratio test. The fit to the NCGC normalized data was performed with k and n fixed at the values from the current fit, or with v, k, and n fixed at the values from the current fit. The likelihood ratio test was used to compare those constrained fits to unconstrained fits (ones which did not fix the values of v, k, and n). Table 12 shows the fraction of fits (out of all substance/assay combinations which this paper's algorithm classified as active) for which the difference between the constrained and unconstrained fits was significant ($p<0.05/1408$ using the likelihood ratio test). The fractions do not vary much as the cutoff value in the algorithm changes. The fraction of significant differences here is much less than the fraction of triplicate HepG2 substances which could not be described by identical parameter values (see the main section); of the substances in the HepG2 study, 26% could not be fitted by identical parameter values, while for the two normalizations only 14-17% could not be fitted by identical values. With k and n fixed, 17% of the HepG2 study substances could not be fitted equally well, while for the different normalizations only 3.9-4.6% could not be fitted equally well.

For each assay, an error standard deviation σ and the significance of the concentration-response for each compound were computed by comparing the fit using the Hill model above to a fit with v-v0=0, using a likelihood ratio test. This significance was calculated both with and without the final data point. An activity measure was computed as in the main section of this paper, using the significance values, response strength, and parameter values. A concordance measure was calculated as the fraction of concentration-responses which were active in both assays or inactive in both. Table 13 shows the concordance values. They show little variation by p-cutoff. The majority of nonconcordant results (at least 71%, depending on p-cutoff) were fits that were classified as active in this paper's analysis but not in the fit to the NCGC normalized data.

Tables for supplemental material
Table 1

Assay	Cutoff level	
	$p<0.05$	$p<0.05/1408$
N2a	1.061	1.178
HUV-EC-C	1.085	1.135
NIH 3T3	1.060	1.163
H-4-IIE	1.065	1.120
mesenchymal	1.063	1.169
BJ	1.068	1.132
Jurkat	1.062	1.149
MRC-5	1.079	1.170
SK-N-SH	1.078	1.149
HEK 293	1.074	1.133
SH-SY5Y	1.071	1.145
Primary renal proximal tubule	1.080	1.140
HepG2 #1	1.071	1.144
HepG2 #2	1.063	1.170
HepG2 #3	1.069	1.140

Table 2

Assay	Cutoff level		
	no cutoff	$p<0.05$	$p<0.05/1408$
N2a	472	433	277
HUV-EC-C	330	255	158
NIH 3T3	894	635	374
H-4-IIE	460	459	380
mesenchymal	455	400	258
BJ	437	381	251
Jurkat	438	439	385
MRC-5	501	379	204
SK-N-SH	404	366	264
HEK 293	356	348	315
SH-SY5Y	534	453	298
Primary renal proximal tubule	371	325	232
HepG2 #1	424	399	273
HepG2 #2	519	489	290
HepG2 #3	416	387	263

Table 3

Assay	Cutoff level		
	no cutoff	p<0.05	p<0.05/1408
N2a	341	323	277
HUV-EC-C	183	172	158
NIH 3T3	594	431	374
H-4-IIE	368	389	380
mesenchymal	299	288	258
BJ	300	293	251
Jurkat	370	385	385
MRC-5	307	236	204
SK-N-SH	287	284	264
HEK 293	298	314	315
SH-SY5Y	338	320	298
Primary renal proximal tubule	252	245	232
HepG2 #1	289	297	273
HepG2 #2	369	355	290
HepG2 #3	290	291	263

Table 4

Assay	Cutoff level	
	p<0.05	p<0.05/1408
N2a	725	1097
HUV-EC-C	1095	1248
NIH 3T3	701	1024
H-4-IIE	841	1023
mesenchymal	783	1112
BJ	933	1146
Jurkat	717	1005
MRC-5	899	1188
SK-N-SH	907	1136
HEK 293	869	1084
SH-SY5Y	864	1108
Primary renal proximal tubule	965	1165
HepG2 #1	881	1123
HepG2 #2	758	1105
HepG2 #3	931	1138

Table 5

Assay	Cutoff level		
	no cutoff	p<0.05	p<0.05/1408
N2a	155	153	140
HUV-EC-C	97	93	91
NIH 3T3	282	224	196
H-4-IIE	211	218	208
mesenchymal	136	140	136
BJ	170	169	166
Jurkat	226	228	224
MRC-5	120	116	110
SK-N-SH	171	176	172
HEK 293	170	175	181
SH-SY5Y	217	212	203
Primary renal proximal tubule	121	118	114
HepG2 #1	175	175	170
HepG2 #2	186	186	168
HepG2 #3	164	161	150
Total	2601	2544	2429

Table 6

Cutoff	No loss of viability	Weak, not active	Weak, active	Medium, not active	Medium, active	Strong, not active	Strong, active
No cutoff	10586	7378	505	527	989	60	1075
p<0.05	14972	3005	459	558	983	63	1080
p<0.05/1408	16898	1161	400	576	944	76	1065

Table 7

Cutoff	Number of duplicate pairs (out of 825) with both duplicates active	v	k	n	high-concentration response
no cutoff	103	0.90	0.76	0.70	0.90
p<0.05	103	0.89	0.79	0.72	0.92
p<0.05/1408	103	0.91	0.79	0.73	0.93

Table 8

Cutoff	Duplicates with at least 1 significant		Duplicates with both significant	
	Ratio < 2	Ratio < 10	Ratio < 2	Ratio < 10
no cutoff	0.51	0.66	0.76	0.98
p<0.05	0.53	0.67	0.78	0.98
p<0.05/1408	0.54	0.69	0.78	0.99

Table 9

	v low	v high	k low	k high	n low	n high
no cutoff	0.13	0.093	0.11	0.059	0.10	0.49
0.05	0.033	0.084	0.037	0.034	0.038	0.14
0.05/1408	0.0054	0.075	0.0056	0.015	0.0059	0.066
Actives						
no cutoff	0	0.34	0.038	0.033	0.039	0.32
0.05	0	0.36	0.032	0.031	0.031	0.27
0.05/1408	0	0.37	0.015	0.017	0.014	0.22

Table 10

Cutoff	Test threshold for p	1 replicate		3 replicates	
		Sensitivity	Specificity	Sensitivity	Specificity
none	0.05	0.98	0.26	0.97	0.48
none	0.05/1408	0.87	0.69	0.91	0.84
0.05	0.05	0.90	0.58	0.94	0.74
0.05	0.05/1408	0.75	0.81	0.85	0.90
0.05/1408	0.05/1408	0.53	0.98	0.80	0.96
High response					
none	0.05	0.96	n/a	0.95	n/a
none	0.05/1408	0.86	n/a	0.88	n/a
0.05	0.05	0.95	n/a	0.94	n/a
0.05	0.05/1408	0.85	n/a	0.87	n/a
0.05/1408	0.05/1408	0.69	n/a	0.83	n/a

Table 11

X	1 replicate				3 replicates			
	v	k	n	Rmax	v	k	n	Rmax
0	0.82	0.33	0.67	0.98	0.85	0.42	0.77	0.99
0.1	0.82	0.33	0.68	0.98	0.84	0.43	0.80	0.99
0.3	0.76	0.73	0.80	0.97	0.83	0.70	0.89	0.99
0.5	0.75	0.94	0.90	0.94	0.80	0.94	0.86	0.98
0.8	0.73	0.99	0.95	0.83	0.86	1.00	0.70	0.92
High response								
0	0.81	0.29	0.61	0.97	0.77	0.34	0.67	0.99
0.1	0.81	0.31	0.63	0.97	0.78	0.34	0.71	0.99
0.3	0.62	0.58	0.87	0.96	0.84	0.91	0.84	0.98
0.5	0.55	0.95	0.92	0.95	0.77	0.93	0.88	0.97
0.8	0.59	0.99	0.94	0.83	0.78	0.99	0.95	0.90

Table 12 Comparison of normalizations: Fraction of substance/assay combinations for which a fit to the NCGC normalized data with free parameters was significantly better than one with two or three parameters fixed at the value from this paper's fit to the data

p-cutoff for this paper's fit	Fraction significant	
	k and n fixed	v, k, and n fixed
no cutoff	0.039	0.14
0.05	0.041	0.15
0.05/1408	0.046	0.17

Table 13: Comparison of normalizations: Concordance between the fit to the NCGC normalized data and fit using this paper's model.

p-cutoff for this paper's fit	Concordance
no cutoff	0.9598
0.05	0.9612
0.05/1408	0.9633

Table Captions for Supplemental Material

Table 1 Effects of varying cutoff: Ratio of σ for a given assay and cutoff to σ for the assay with no cutoff.

Table 2 Effects of varying cutoff: Number of substances with concentration-response significant at $p<0.05$ with $v>0$

Table 3 Effects of varying cutoff: Number of substances with concentration-response significant at $p<0.05/1408$ with $v>0$

Table 4 Effects of varying cutoff: Number of responses with v forced to equal 0

Table 5 Effects of varying cutoff: Number of substances classified as active

Table 6 Effects of varying cutoff: Classification by activity and strength of response

Table 7 Effects of varying cutoff: Concordance and correlation measures. Correlations are for the given parameter, between duplicates in which both pairs are significant

Table 8 Effects of varying cutoff: Concordance of k values

Table 9 Fraction of parameter values found to be equal to the high or low constraints, for all data and for substances classified as active

Table 10 Simulation: Effects of varying cutoff: sensitivity (fraction of positives predicted as positive) and specificity (fraction of negatives predicted as negative) for various combinations of cutoff level and p-value used for test. A prediction is counted as a positive if the p-value (from the test of the null hypothesis $v>0$) is less than the listed value. “High response” indicates results from the simulation using only positive concentration-responses.

Table 11 Simulation: Correlation of modeled and true parameters across simulation results. Correlation is taken for substances with predicted response with $v>0$, true $k<0.092$, and true loss of viability at highest concentration > X. Rmax is the response at the highest concentration, which is a function of v, k, and n. Results are from simulations with cutoff of $p<0.05/1408$ for the parameter reduction step. “High response” indicates results from the simulation using only positive concentration-responses.

Table 12 Comparison of normalizations: Fraction of substance/assay combinations for which a fit to the NCGC normalized data with free parameters was significantly better than one with two or three parameters fixed at the value from this paper’s fit to the data

Table 13: Comparison of normalizations: Concordance between the fit to the NCGC normalized data and fit using this paper’s model.

Figure Captions for Supplemental Material

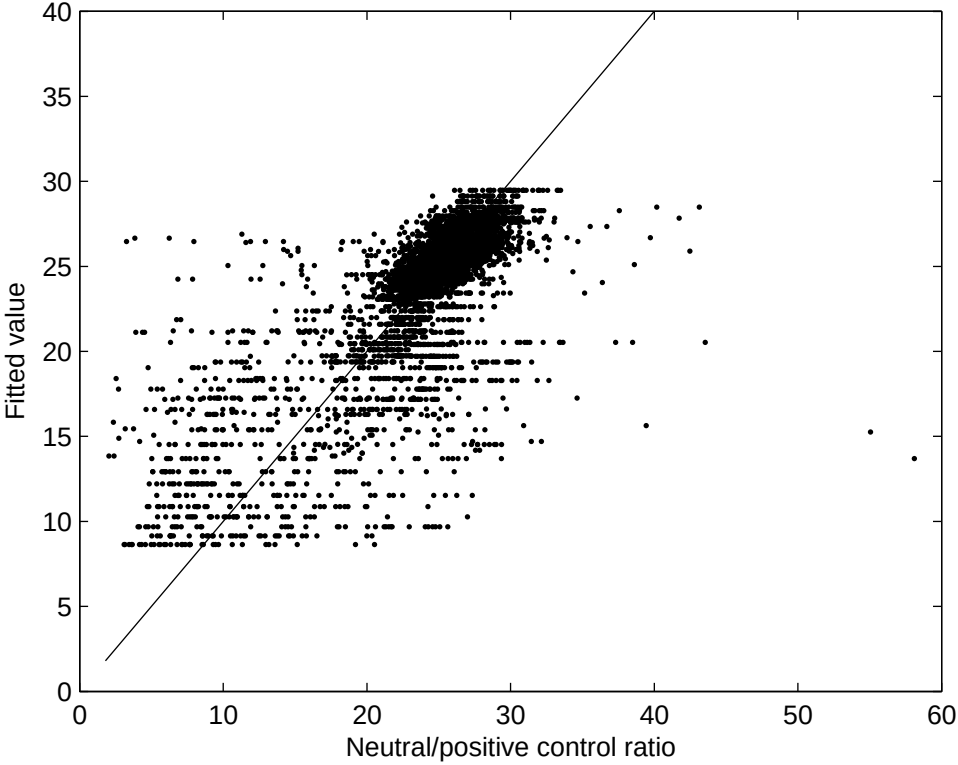
Figure S1 Data and fitted values of μ, the ratio of neutral to positive control responses

Figure S2 Distribution of the response at high concentration used when generating data in the simulation study

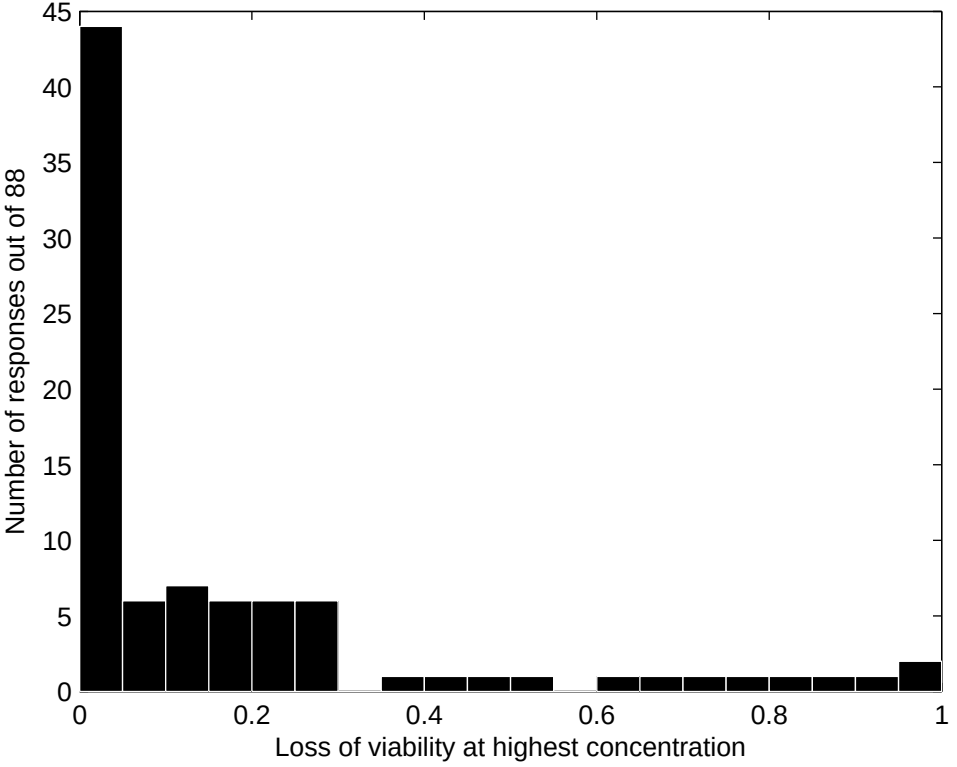
Figure S3 Normalized data values from the NCGC normalization and this paper's normalization, on the scale where 0 is no response and -1 is full loss of viability. Each dot is one data point

Figure S4 Fitted values from the Hill function model, from fits to both sets of normalized data, on the scale where 0 is no response and -1 is full loss of viability. Each dot is one data point.

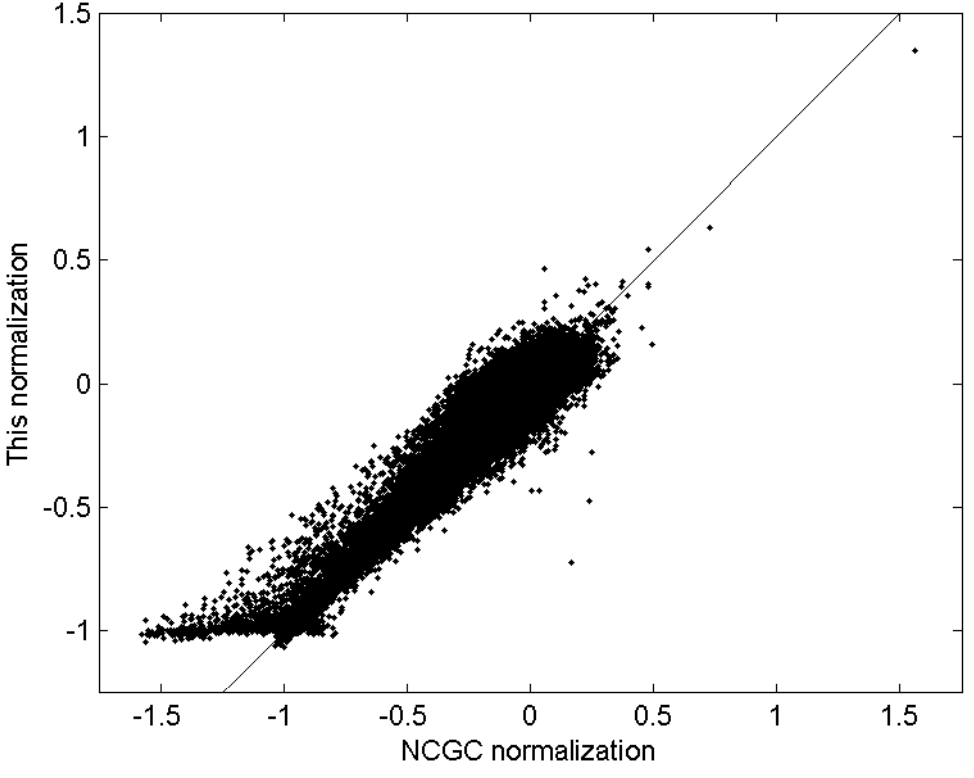
Type of file: figure
Label: FigS1
Filename: FigureS1.pdf


Fitted value
Neutral/positive control ratio
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60

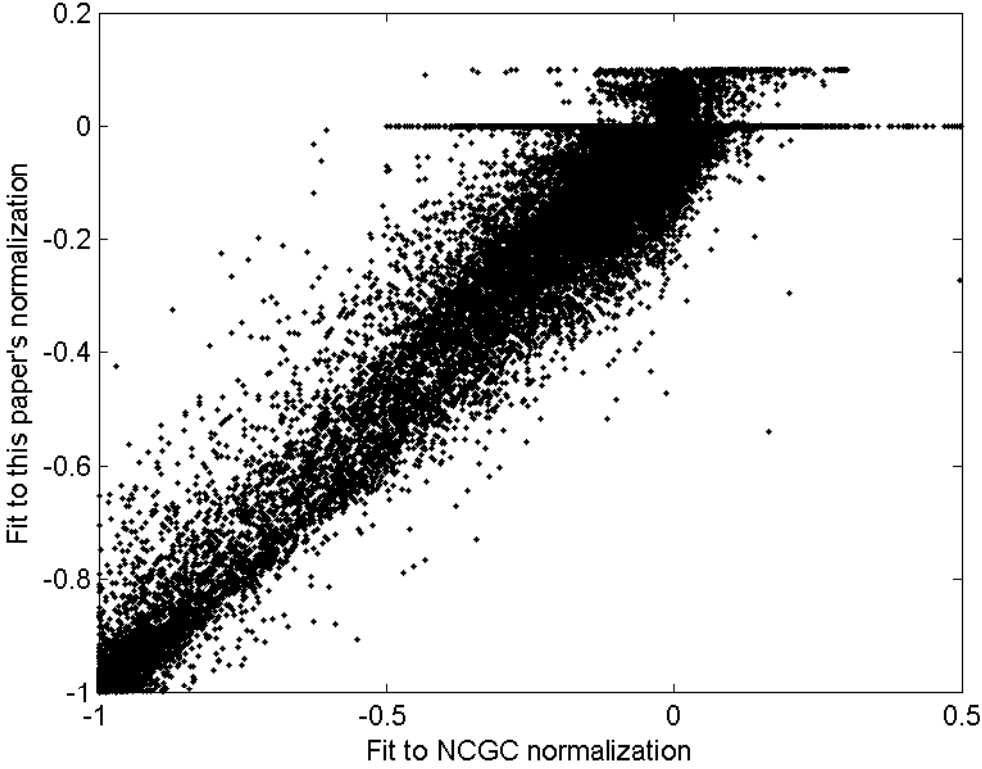
Type of file: figure
Label: FigS2
Filename: FigureS2.pdf


45
40
35
30
25
20
15
10
5
0
Number of responses out of 88
0
0.2
0.4
0.6
0.8
1
Loss of viability at highest concentration

Type of file: figure
Label: FigS3
Filename: FigureS3.pdf


1.5
1
0.5
0
-0.5
-1
This normalization
-1.5
-1
-0.5
0
0.5
1
1.5
NCGC normalization

Type of file: figure
Label: FigS4
Filename: FigureS4.pdf


0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.5
0
0.5
Fit to this paper's normalization
Fit to NCGC normalization