BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 16 2011, pages 2173–2180
doi:10.1093/bioinformatics/btr359

Genome analysis Advance Access publication June 17, 2011

Noise reduction in genome-wide perturbation screens using linear mixed-effect models

Danni Yu[1], John Danku[2], Ivan Baxter[3], Sungjin Kim[4], Olena K. Vatamaniuk[4], David E. Salt[2,5] and Olga Vitek[1,6,*]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA, [2]Department of Plant and Soil Science, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UK, [3]USDA-ARS Plant Genetics Research Unit, Donald Danforth Plant Science Center, St. Louis, MO 63132, [4]Department of Crop and Soil Sciences, Cornell University, Ithaca, NY 14853, USA, [5]Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UK and [6]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: High-throughput perturbation screens measure the phenotypes of thousands of biological samples under various conditions. The phenotypes measured in the screens are subject to substantial biological and technical variation. At the same time, in order to enable high throughput, it is often impossible to include a large number of replicates, and to randomize their order throughout the screens. Distinguishing true changes in the phenotype from stochastic variation in such experimental designs is extremely challenging, and requires adequate statistical methodology.

Results: We propose a statistical modeling framework that is based on experimental designs with at least two controls profiled throughout the experiment, and a normalization and variance estimation procedure with linear mixed-effects models. We evaluate the framework using three comprehensive screens of *Saccharomyces cerevisiae*, which involve 4940 single-gene knock-out haploid mutants, 1127 single-gene knock-out diploid mutants and 5798 single-gene overexpression haploid strains. We show that the proposed approach (i) can be used in conjunction with practical experimental designs; (ii) allows extensions to alternative experimental workflows; (iii) enables a sensitive discovery of biologically meaningful changes; and (iv) strongly outperforms the existing noise reduction procedures.

Availability: All experimental datasets are publicly available at www.ionomicshub.org. The R package `HTSmix` is available at http://www.stat.purdue.edu/~ovitek/HTSmix.html.

Contact: ovitek@stat.purdue.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on January 13, 2011; revised on May 31, 2011; accepted on June 8, 2011

1 INTRODUCTION

Perturbation screens (Boutros and Ahringer, 2008; Forsburg, 2001) subject model organisms to stresses that are external (e.g. heat shock or chemical treatments) or genetic (e.g. disruption or deletion of genes). A variety of phenotypes can be measured in association with the stresses. These can be univariate phenotypes such as cell growth rate or activity of a reporter gene, low-dimensional phenotypes such as cellular morphology or high-dimensional phenotypes such as gene expression or protein abundance. When conducted on a genome-wide scale, perturbation screens provide invaluable insight into the function of living organisms (Markowetz, 2010; Markowetz and Spang, 2007). They are increasingly used in functional biology (Boone *et al.*, 2007; Gstaiger and Aebersold, 2009), and in biomedical (Ideker and Sharan, 2008) and biopharmaceutical research (Bharucha and Kumar, 2007).

The throughput of genome-wide screens is a primary concern in these investigations. Since it can take weeks and sometimes months to measure the phenotypes, it is often impossible to fully implement the fundamental principles of statistical experimental design. In particular, the screens can incorporate little replication, and a full randomization of the order of the replicates is often impractical. At the same time, the measured phenotypes are subject to large variation, which is due to both natural between-sample variation, and technical variation in the sample handling and measurement procedures. The problem is compounded by changes in experimental characteristics (e.g. instruments, labor, reagents) that are unavoidable in large-scale screens. Interpretation of the screens is, therefore, a key and non-trivial step, which must take the specifics of the experiments into account.

In this article, we propose a statistical modeling framework for accurate interpretation of high-throughput screens, most specifically in cases of low-dimensional phenotypes. We focus on screens which have a limited number of replicate samples and a sensitive phenotype (i.e. the phenotype that is affected in a non-negligible proportion of the samples). Distinguishing the systematic signal from noise is particularly challenging in such situations.

2 BACKGROUND

Statistical design and analysis of perturbation screens involve (i) experimental design, (ii) normalization, (iii) summarization of the phenotype of each sample from multiple replicates and estimation of the associated variation, (iv) determination of 'hits', i.e. samples

*To whom correspondence should be addressed.

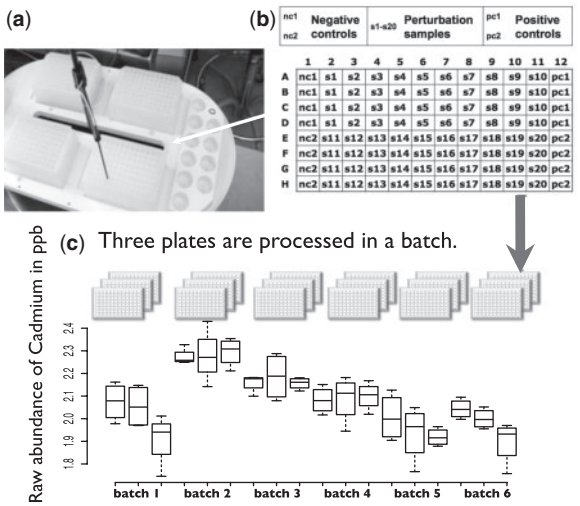

Fig. 1. Experimental design of the knock-out screen in Section 4. (**a**) Samples are processed in 96-well plates. (**b**) Two negative controls (*nc*1 and *nc*2) are quadruplicated in the first column, and two positive controls (*pc*1 and *pc*2) are quadruplicated in the last column of the plate. Samples with 20 different genetic perturbations ($s1, \ldots, s20$) are quadruplicated in the remaining columns. (**c**) Distribution of the scored abundance of Cadmium for the positive control YPR065W, in the first 18 plates of the screen, separately for each plate and batch. The distributions show systematic effects of plates and batches on the phenotype.

with systematic changes in phenotype and (v) evaluation and quality control.

Experimental design: a typical design of a perturbation screen is overviewed in Figure 1. Samples in the screen are processed in 96-well or similar plates. To enable high throughput, the samples are profiled with a small number of replicates [e.g. 4, as recommended by (Zhang and Heyse, 2009)], and all replicates of a sample are systematically allocated to the same plate. The screens typically require tens, hundreds or even thousands of plates. Therefore, the plates are handled in batches defined by the availability of biological material and capacity of equipment.

The scored phenotypes can be systematically altered by batches and plates (Malo *et al.*, 2006), within-plate effects due e.g. to rows and columns on the plate (Malo *et al.*, 2006) or excessive evaporation of media around the edges (Wiles *et al.*, 2008). To account for these artifacts, one or more control samples are included in all plates. These can be negative controls (e.g. unperturbed samples) and positive controls (e.g. samples with known changes in the phenotype). Malo *et al.* (2006) recommend to allocate the controls around the edges of the plate, in order to limit the negative effects of evaporation on the perturbed samples.

The limited capacity of plates, the limited within-plate replication of perturbed samples, the absence of between-plate replication and a small number of controls makes elimination of experimental artifacts in perturbation screens extremely challenging in practice. We argue in Section 3 that the existing statistical methods under-use the information provided by the controls in these situations, and that it is possible to obtain a more specific detection of hits by a separate use of two or more distinct positive or negative controls.

Normalization: scored phenotypes undergo quality control to eliminate the outlying or failed samples or plates. After that, a normalization procedure accounts for confounding and for experimental artifacts, and makes the scored phenotypes comparable across samples, batches and plates.

Two most frequently used families of normalization are sample based and control based. Sample-based normalization methods (detailed in Supplementary Section 1) assume that the majority of perturbations do not affect the phenotype. Examples are *B*-score (Tukey, 1960), *Z*-score and plate-wise median (Collins *et al.*, 2006). Malo *et al.* (2006) reviewed sample-based normalizations for perturbation screens and recommended using *B*-score. Another popular method is quantile normalization, which was introduced for the analysis of gene expression microarrays (Bolstad *et al.*, 2003; Yang *et al.*, 2002), and is applied to perturbation screens (Bankhead *et al.*, 2009). Principal component analysis can be used to account for the batch effect (Leek *et al.*, 2010), and surrogate variable analysis can help remove the heterogeneous effect of plates between batches (Leek and Storey, 2007). Within-plate artifacts can be normalized using lowess smoothing (Baryshnikova *et al.*, 2010).

Sample-based normalization is attractive because it is based on the entire collection of measurements in the experiment, uses the maximal number of observations and therefore produces an accurate estimate of the normalized phenotype. However, it is not appropriate for screens where many perturbations affect the phenotype, and also in secondary and confirmatory screens. Alternative normalization procedures, based on controls, are more appropriate in these situations (Birmingham *et al.*, 2009). Examples of control-based normalization are detailed in Supplementary Section 2. Given a relatively small number of controls in a plate, control-based normalization can only account for limited types of experimental artifacts, and can yield highly variable estimates of bias. Wiles *et al.* (2008) compared the performance of seven sample-based and control-based normalization methods, and found, in the words of Birmingham *et al.* (2009), that 'no single method excelled' in all situations. Software implementations, such as the ones in the open-source Bioconductor packages `RNAither` (Rieber *et al.*, 2009) and `cellHTS2` (Boutros *et al.*, 2006) offer multiple above-mentioned alternatives.

In this work, we demonstrate that control-based normalization can improve the accuracy of results, as compared to the currently available methods, in screens where a large proportion of samples show changes in the phenotypes. We argue that such procedure should involve more than one control sample, and should be used not only for normalization, but also for estimation of residual between-plate variation.

Summarization of phenotypes and estimation of variation: this step summarizes the normalized phenotypes of a biological sample across replicates in a single value, typically by averaging, and estimates the associated variation. Estimation of variation is important, as it allows us to distinguish random variation from stress-related changes in the phenotype. Most existing methods estimate the variation by sample variance (Collins *et al.*, 2006), or by its robust alternatives. Malo *et al.* (2006) recommended using Empirical Bayes approach to variance stabilization, which was originally introduced in the context of gene expression microarrays (Smyth, 2004, 2005), but is applicable directly to the context of perturbation

screens. The approach is summarized in mathematical formulation in Supplementary Section 3.

The goal of this article is to demonstrate that such estimation of variation has serious deficiencies, in particular in screens with sensitive phenotypes and no between-plate replication. If the experimental design allocates all replicates of a biological sample in the same plate, these methods only estimate the variation within the plate. In other words, the methods assume that within-plate variation represents the full extent of variation of the normalized phenotypes.

We argue in Section 3.3 that this assumption oversimplifies the structure of variation in the screens, and is rarely verified. We note that in control-based normalization, where normalizing quantities are estimated from a small number of observations, estimates of plate- and batch-specific bias are subject to uncertainty. Moreover, the effect of batches and plates on the phenotype can differ somewhat across biological samples, and further contribute to the variation. We show that appropriately accounting for this residual variation can play an important role in the determination of hits.

Determination of hits: determination of hits is formalized as testing the null hypothesis 'the perturbed phenotype is consistent with the phenotype of a control' or 'the perturbed phenotype is consistent with the average phenotype of all perturbations' against the corresponding alternative. The test is conducted using a test statistic, such as the Student's T or the moderated T above, which compares the summary quantification of the phenotype to its estimate of variation. Depending on the experiment, the reference distribution of the statistic is assumed Student or Normal, or is estimated empirically based on controls. Non-parametric alternatives, e.g. the Mann–Whitney test and the Rank Product test (Rieber *et al.*, 2009) can also be used, but have lower power.

The second aspect of determination of hits is the selection of the test statistic cutoff, which controls the rate of false positive hits at the desired level. Multiple testing procedures controlling for the false discovery rate (FDR), such as Benjamini and Hochberg (1995) or Efron (2008), can be used directly. Alternatively, Zhang *et al.* (2008) developed a specialized Bayesian procedure, which directly models the probabilities of phenotypes and controls FDR. Using ordered *Z*-scores, Kaplow *et al.* (2009) designed a tool called RNAiCut for automated identification of pathway-relevant hits. Although all these approaches are appropriate, their sensitivity and specificity depend on the choice of the test statistic, and in particular on its estimate of variation.

Evaluation: development of statistical methods for high-throughput screens is challenging in part because of difficulties in their evaluation on experimental datasets. The evaluation is facilitated in the case of multivariate phenotypes, where we can examine the consistency of normalized phenotypes of the controls in a multivariate space. We use such multivariate phenotypes, and both control-based and sample-based evaluation in Section 5.

3 METHODS

In the following, we consider high-throughput perturbation screens with 1D or low-dimensional quantitative phenotypes. To be specific, we focus on genetic perturbations, and refer to the screened samples as mutants. However, the discussion is applicable to all perturbation types. The proposed method is particularly relevant for screens with highly disruptive perturbations, or with sensitive phenotypes, where we cannot expect a relatively small number of hits.

We propose a stepwise interpretation procedure based on linear mixed-effects models. In large-scale experiments, stepwise linear modeling is a computationally efficient alternative to a global mixed-effects model that is used to fit the entire dataset. In the past, stepwise procedures were successfully applied in the context of gene expression microarrays (Dobbin and Simon, 2002; Wolfinger *et al.*, 2001), and the proposed approach is similarly effective for perturbation screens.

3.1 Experimental design

We consider experiments which utilize 96-well or similar plates, and profile all replicates of a sample in the same plate. One can use all within-plate allocations of samples, e.g. suggested by Malo *et al.* (2006), and any number of biological replicates, e.g. 4 recomended by Zhang and Heyse (2009).

A key requirement of the proposed approach is the presence of at least two distinct control samples, profiled in all batches and all plates. The first control is used for normalization of the phenotype across batches and plates. The second control is used to estimate the associated variation, and to derive the summary statistic for each mutant. Incorporating one or two additional control samples, complementing the previous two, is beneficial to evaluate the quality of the results.

3.2 Normalization

Basic model-based normalization: we denote X_{gkbp} a scored univariate phenotype, where g is the mutant *gene*, k is the replicate sample of that mutant, b is the *batch* index and p is the *plate* index. For multivariate phenotypes we consider each dimension separately, and use the convention that X_{gkbp} represents one particular dimension.

The major sources of variation in a screen are batches, plates, and biological and technical variation. The basic normalization model assumes that these effects are non-systematic Normal random variables, and represents these assumptions with the linear model

$$X_{gkbp} = \mu_g + B_{gb} + P(B)_{gp} + \varepsilon_{gkbp} \tag{1}$$

$$B_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B_g}), P(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P_g}),\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g})$$

where B_{gb} is the batch effect, $P(B)_{gp}$ is the plate effect nested within the batch and ε_{gkbp} is the combination of the biological and technical variation. B_{gb}, $P(B)_{gp}$ and ε_{gkbp} are independent.

Parameters μ_1, B_{gb} and $P(B)_{gp}$ can be estimated with a sample-based approach, i.e. using all the samples in the batch or plate. However, such estimation is undesirable in screens with disruptive perturbations or sensitive phenotypes, as it will produce biased estimates. Therefore, we focus on control-based normalization, and estimate $\hat{\mu}_1$, $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ by fitting the model in Equation (1) to the first control (i.e. to biological samples with $g=1$ in the notation above). In linear mixed models, such estimates are typically obtained by maximizing the restricted/residual maximum likelihood (REML) using Expectation–Maximum (EM) or Newton–Raphson algorithms. The ridge-stabilized Newton–Raphson algorithm allows a faster convergence (Lindstrom and Bates, 1988), and we use this algorithm as implemented in the R package `lme4`. The resulting model-based estimates differ from sample averages, and are derived to ensure an unbiased estimation of variances $\sigma^2_{P_1}$, $\sigma^2_{B_1}$ and $\sigma^2_{\varepsilon_1}$. The normalization accounts for the batch- and plate-specific deviations in quantitative phenotypes (also known as batch- and plate-specific additive effects in statistical literature) by subtracting their control-based estimates $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ from all the scored phenotypes

$$r_{gkbp} = X_{gkbp} - [\hat{B}_{1b} + \hat{P}(B)_{1p}] \tag{2}$$

Here, r_{gkbp} denotes the normalized phenotype X_{gkbp} of the k-th replicate of the mutant sample g, located in the b-th batch and on the p-th plate.

Extensions: the linear model above is flexible, and can be extended in a variety of ways to account for within-plate effects, confounding effects or time-dependent correlation effects. For example, the systematic changes in

phenotype due to the position of a sample within a plate can be accounted for similarly as with B-score [Equation (3) in Supplementary Section 1]:

$$X_{gkbp} = \mu_g + R_{ip} + C_{jp} + B_{gb} + P(B)_{gp} + \varepsilon_{gkbp}, \quad (3)$$

$$\sum_i R_{ip} = 0,\ \sum_j C_{jp} = 0,$$

$$P(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P_g}),\ B_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B_g}),\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g})$$

where R_{ip} and C_{jp} are the deviations on row i and column j on the p-th plate, and the remaining notation is as in Equation (1). A lowess-based smoothing of these effects can be used when rows or columns only contain a small number of distinct biological samples (Baryshnikova *et al.*, 2010).

The model can also be extended to account for confounding effects on the scored phenotypes. For example, to account for the confounding effect of growth rate of the mutants, one can normalize both the phenotype and the growth rate as in Equations (1) and (2). If we denote gr as the normalized growth rate, then a linear model can be fit to estimate a single linear relationship between the confounding factor and the phenotype across all the biological samples.

$$r_{gkbp} = \beta_0 + \beta_1 gr_{gkbp} + \epsilon_{gkbp},\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g}) \quad (4)$$

and the adjusted normalized values are obtained as

$$r'_{gkbp} = r_{gkbp} - \hat{\beta}_1 gr_{gkbp} \quad (5)$$

The normalization steps above yield scored phenotypes that are comparable across biological samples. For the experimental datasets in Section 4, the within-plate quality control procedures (Supplementary Section 4) indicate that the row and column effects are negligible. Therefore, the analysis performed the basic normalization in Equations (1) and (2) and the adjustment for growth rate in Equations (4) and (5) without within-plate spatial normalization.

3.3 Estimation of variation and summarization

Figure 2 shows results of the normalization procedure in the knock-out screen in Section 4, in three control samples, as applied to the sulfur accumulation phenotype. Supplementary Sections 6–8 present such plots for all the controls and all the phenotypes in Section 4. The first control (Fig. 2a) was used to derive batch- and plate-specific changes of phenotype $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$. Plate-wise medians of the normalized phenotype in the right panel of Figure 2a form a horizontal straight line, indicating that the normalization removed the systematic between-batch and between-plate variation for that control.

Figure 2b and c show normalized phenotypes of two more controls, which were not used to estimate the normalization parameters. They illustrate that, although the normalization removed large artifacts, e.g. outlying measurements in the left panel of Figure 2b and a systematic increasing trend in the left panel of Figure 2c, it did not eliminate all between-batch and between-plate deviations for these controls. This residual variation is due to the differential effect of batches and plates on mutant phenotypes (also known as non-additive effect, or batch×mutant and plate×mutant statistical interactions), as well as to the uncertainty in estimation of $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ from a small number of replicates in a plate. In screens where the interaction effects can be estimated, they can be accounted for e.g. by including them as fixed effects into the normalization model in Equation (1), or using alternative approaches (Leek *et al.*, 2010). However, in screens where all replicates of the samples are profiled in a single plate, these effects cannot be estimated directly. Omitting these effects can seriously underestimate the overall variation, and undermine the accuracy of the results.

We propose to express the residual variation in normalized phenotypes in terms of random effects the second linear model

$$r'_{gkbp} = \mu'_g + P'(B)_{gp} + B'_{gb} + \varepsilon'_{gkbp}, \quad (6)$$

$$P'(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P'_g}),\ B'_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B'_g}),$$

$$\varepsilon'_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon'_g}),\ for\ g = 2, 3, 4, 5, \ldots$$

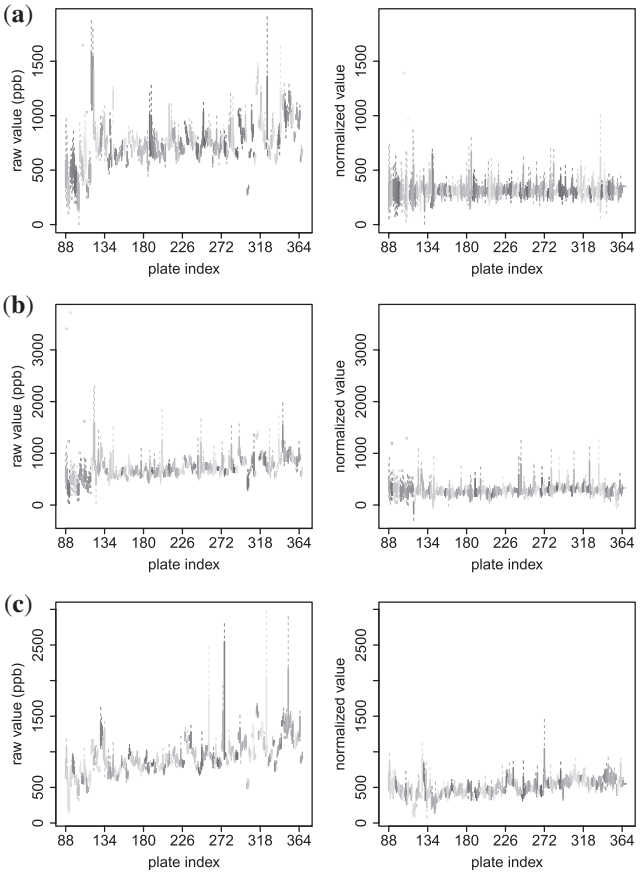

Fig. 2. Effect of normalization on elemental abundance of Sulfur in three controls of the knock-out screen in Section 4. (**a**) First control, BY4741, used for normalization. Left: before normalization. Right: after normalization. (**b**) Second control, YDL227C, not used for normalization. Left: before normalization. Right: after normalization. (**c**) Third control, YLR396C, not used for normalization. Left: before normalization. Right: after normalization. Y-axis: raw or normalized abundance. X-axis: plate id. The figures show boxplots of the phenotype in 305 plates, indicating batches by gradient colors, similarly to Figure 1c. Normalization reduces the systematic differences between batches and plates; however, residual variation is present for the second and the third controls.

where r'_{gkbp} is the normalized phenotype of sample g, and the remaining notation is as in Equation (1). For samples profiled in a single plate, the summary phenotype of mutant g is μ_g, and its estimate is equivalent to the average of the observed phenotypes over all replicates $\bar{r}'_{g..}$. The associated estimated variation is

$$\widehat{Var}(\bar{r}'_{g..}) = (\hat{\sigma}^2_{P'_g} + \hat{\sigma}^2_{B'_g} + \hat{\sigma}^2_{\varepsilon'_g}/n_g)$$

where n_g is the number of within-plate replicate samples of the mutant g. Parameter $\hat{\sigma}^2_{\varepsilon'_g}$ is estimated by the sample variance $s^2_{\varepsilon'_g}$; however $\sigma^2_{P'_g}$ and $\sigma^2_{B'_g}$ are not estimable for each mutant directly. Therefore, we propose to use one or several additional controls, which have not been previously used for normalization, to obtain plug-in estimates of $\sigma^2_{P'_g}$ and $\sigma^2_{B'_g}$. Such approach assumes that the control-based estimates accurately represent the residual variation of all the biological samples in the screen. In our experience, this assumption is frequently plausible, and yields accurate results. In screens where we cannot make this assumption, the residual variation can only be

estimated by changing the experimental design and implementing between-plate replication; however, this will reduce substantially the throughput and may be difficult to implement in practice.

In the following, we use the second control for variance estimation, i.e.

$$\hat{\sigma}^2_{P'_g} + \hat{\sigma}^2_{B'_g} \approx \sigma^2_{P'_2} + \sigma^2_{B'_2} \text{ for all } g. \quad (7)$$

Fitting the model to the second control in Figure 2 using the R package `HTSmix` yields $(\hat{\sigma}^2_{B'_2} + \hat{\sigma}^2_{P'_2})/\sigma^2_{\varepsilon'_2} = 0.23$, indicating the relative importance of the residual variation. We show in Section 5.1 that results of the proposed normalization and variance estimation procedure have little sensitivity to the specific choice of the controls.

3.4 Determination of hits

Hypothesis and test statistic: in screens where we expect a relatively small number of affected phenotypes, determination of hits is equivalent to testing H_0: *Phenotype is consistent with the phenotype of control* against H_a: *Phenotype is systematically larger (or smaller) than the phenotype of the control* (Boutros and Ahringer, 2008; Malo *et al.*, 2006). However in experiments with disruptive perturbations or sensitive phenotypes, the test will result in an unpractically large number of hits. An alternative hypothesis in this case is H_0: *Phenotype is consistent with the median phenotype of all perturbed samples* against H_a: *Phenotype is systematically larger (or smaller) than the median phenotype of all perturbed samples*. We focus on the latter hypothesis in the discussion below.

The test statistic standardizes the normalized phenotype, i.e. it quantifies the phenotype in the units of its estimated standard deviation

$$D_g = \bar{r}'_{g\cdot\cdot} / \sqrt{(\hat{\sigma}^2_{P'_2} + \hat{\sigma}^2_{B'_2} + s^2_{\varepsilon'_g}/n_g)} \quad (8)$$

The denominator in Equation (8) incorporates $\hat{\sigma}^2_{P'_2}$ and $\hat{\sigma}^2_{B'_2}$, and therefore D_g will yield fewer hits as compared to the regular (or moderated) T-statistic. When the assumptions of the models in Equations (1) and (6), as well as of the estimation procedure in Equation (7) are verified, the sampling distribution of the test statistic is approximately Normal. The center and the scale of the distribution depend on the nature of the effects in the screen.

Controlling FDR in the list of hits: to produce a list of hits while controlling the FDR, we adapt the approach by Efron (2008). The approach assumes that under H_0, the sampling distribution of the test statistics D_g is the same for all g, and models the observed distribution of the test statistic as a mixture of distributions under H_0 and H_a. Similarly to Efron (2008), we apply a transformation to the test statistic to ensure that the sampling distribution under H_0 is close to the Standard Normal, i.e.

$$Z_g = \frac{D_g - median(D_g)}{median(\,|D_g - median(D_g)|\,) \cdot C}, \quad (9)$$

where $C = 1/\Phi^{-1}(3/4) \approx 1.4826$ is a normalizing constant for a robust unbiased estimation of the scale (Hoaglin *et al.*, 1983). We then use the implementation of the approach by Efron (2008) in the R package `locfdr` to fit a Normal distribution to the center of the histogram of Z_g, and determine the cutoff of Z_g that controls the FDR.

In multivariate phenotypes, the sampling distributions of Z_g are comparable across dimensions, and we suggest combining all dimensions in a single distribution to optimize the quality of fit. When the assumptions of the models in Equation (1) and Equation (6), as well as of the estimation procedure in Equation (7) are verified, the sampling distribution of Z_g under H_0 is approximately Standard Normal. Supplementary Section 5 (Figs 2 and 3) illustrate the sampling distributions of Z_g, and indicate that the data present no gross departures from the assumptions.

4 EXPERIMENTAL DATASETS

Perturbation screens: we illustrate the performance of the proposed approach using three large-scale genetic perturbation screens of *S.cerevisiae* (baker's yeast). The first perturbation screen, that we denote KO, involves the collection of 4940 viable mutants where the open reading frames in haploid cells have been disrupted one at a time. The second screen, that we denote KOd, involves the collection of 1127 viable diploid lines, with one of the two copies of the gene disrupted one at a time. The lines correspond to lethal disruptions in the haploid lines. The third screen, that we denote OE, involves the full collection of 5770 viable mutants where each of the open reading frames is expressed at a higher than normal rate. In the three experiments, the mutants were incubated in a series of 96-well plates, with 4 (and sometimes 8 or 16) replicates per strain. The majority of mutants were only grown in a single plate.

The phenotype of interest in these screens is the yeast ionome. The ionome of an organism is defined as its mineral nutrient and trace element composition (Baxter, 2009; Salt *et al.*, 2008), and includes P, Ca, K, Mg (macronutrients); Cu, Fe, Zn, Mn, Co, Ni, Se, Mo, Cl (micronutrients of significance to plant and human health); and Na, As and Cd (minerals causing agricultural, environmental or health problems). To quantify each element, a common yeast growth media was supplemented with additional elements (Danku *et al.*, 2009), and each sample was processed, in batches of three plates, using inductively coupled plasma spectroscopy combined with mass spectroscopy (ICP-MS). Peaks in the spectra were signal processed, and the absolute quantification in parts per billion (ppb) obtained through the use of calibration standards as described in Danku *et al.* (2009). A quality control procedure removed failed and outlying samples. Overall, the KO and KOd screen yields the multivariate phenotype of 14 elements, and the OE screen yields the multivariate phenotype of 17 elements for each mutant.

Each experiment included two negative and two positive control strains (BY4741, YDL227C, YLR396C and YPR065W for the KO screen, BY4743, YDL227C, YLR396C and YPR065W for the KOd screen, and YMR243C, YDL227C, YBR290W and YGL008C for the OE screen), which were grown in four replicates within each plate. The positive controls were chosen based on the results of Eide *et al.* (2005), who found observable changes in key elements such as Ni60, Cd111 and S34 for these strains. The controls help test our ability to detect such known changes in abundance.

Quality control did not identify strong spatial within-plate effects on the ionomic profiles (Supplementary Section 4). However, it was established that differences of growth rates between mutants could act as potential confounders of the ionomic phenotypes. To account for that, the growth rate of each mutant was quantified by the sample optical density (OD) using an OpsysMR plate reader (DYNEX Technologies, Chantilly, VA, USA). All measurements are publicly available at www.ionomicshub.org.

The elements constitute an integral part of most biochemical processes, and therefore a large number of mutations is expected to affect the ionomic phenotype. The goal of these experiments is, therefore, to identify the mutant strains, for which the abundance of at least one element deviates substantially from its median abundance over all mutants.

5 RESULTS

5.1 Evaluation based on controls

One negative control (BY4741 for KO, BY4743 for KOd, and YMR243C for OE) was used to perform the normalization procedure in Equation (1) and one negative control (YDL227C for KO, YDL227C for KOd, and YDL227C for OE) to estimate the variation in Equation (6). Positive control samples (YLR396C and YPR065W for KO, YLR396C and YPR065W for KOd, and YBR290W and YGL008C for OE) were used to evaluate the quality of the results.

Normalization and variance estimation: univariate phenotypes: supplementary Section 1 show the results similar to Figure 2 for positive controls in all screens, and for all the phenotypes, before and after normalization with Equations (1), (2), (4) and (5). The figures show that the methods roughly succeed at removing

the systematic trend in element abundance. For illustration, Supplementary Section 5 also present results of normalization with *B*-score, *Z*-score and normalized percent inhibition (NPI). Although the methods also remove the systematic trends, they alter the scale of the phenotype, and a relative comparison in one dimension is not straightforward. We utilize multivariate phenotypes for this purpose instead.

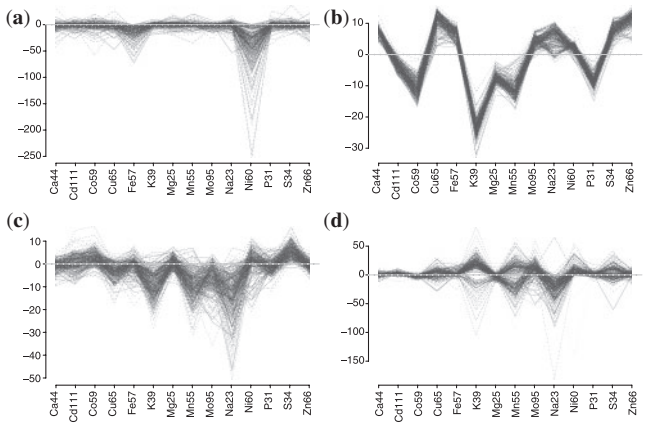

Fig. 3. Profile plots of the standardized phenotypes of the control YLR396C in the KO screen, which has not been used for normalization or standardization. (**a**) Raw phenotypes. Average pairwise correlation is 0.558. (**b**) Proposed normalization and standardization. Average pairwise correlation is 0.968. (**c**) Normalization with *B*-score, standardization with moderated T. Average pairwise correlation is 0.640. (**d**) Normalization with NPI, standardization with moderated T. Average pairwise correlation is 0.438. *X*-axis: inorganic elements. *Y*-axis: (a) raw and (b–d) normalized and standardization phenotypes. Each line represents the phenotype of the control in one plate.

Normalization and variance estimation: multivariate phenotypes: Multivariate phenotypes provide additional insight into the relative efficiency of noise reduction procedures. Since we do not expect biologically meaningful differences in phenotypes between plates for the controls, a tighter pattern of standardized phenotypes of the controls across all dimensions, as compared to the mean phenotype in each dimension, indicates a better removal of the residual batch- and plate-specific variation.

Figure 3 compares the profile plots of the standardized phenotypes for one positive control in the KO screen, obtained before normalization, after sample-based normalization with *B*-score and standardization with moderated T statistic, after normalization with control-based NPI and standardization with Moderated T, and after the proposed normalization with Equations (1), (2), (4) and (5) and standardization with Equations (6)–(9). As can be seen, *B*-score and NPI, combined with the moderated T statistic, result in noisy standardized profile, and between-plate variation exceeds the differences in standardized abundance of the elements. The average abundance of most elements is not distinguishable from zero. The proposed normalization and estimation procedure produces the tightest pattern in the profiles, which will allow us to best distinguish changes in element abundance. Supplementary Section 6–8 contain similar plots for all the screens and all the phenotypes.

We further compare the performance of the methods quantitatively by calculating the average Pearson correlations of standardized profiles (as in Fig. 3) across all pairs of plates. Table 1 shows that the proposed approach produces the highest correlation, and therefore successfully reduces the noise as compared to the other techniques.

Stability of noise reduction to choice of controls: Table 2 shows the average pairwise Pearson correlations of profiles of the controls in the KO screen, such as in Figure 3b, calculated over all pairs of plates, and using all possible combinations of controls for

Table 1. Pearson correlation of normalized and summarized profiles between pairs of plates, for two positive controls which have not been previously used for normalization or standardization

		Average pairwise Pearson correlations between plates					
		KO screen YLR396C	KO screen YPR065W	KOd screen YLR396C	KOd screen YPR065W	OE screen YBR290W	OE screen YGL008C
Current existing methods	*B*-score[a]	0.640	0.720	0.889	0.825	0.491	0.331
	Z-score[b]	0.765	0.776	0.910	0.817	0.530	0.361
	Plate-wise median[c]	0.738	0.819	0.915	0.835	0.595	0.481
	PocMean[d]	0.666	0.670	0.875	0.729	0.626	0.523
	PocMed[e]	0.765	0.806	0.896	0.834	0.554	0.424
	NPI[f]	0.438	0.508	0.689	0.686	0.759	0.595
	Quantile[g]	0.696	0.772	0.857	0.917	0.630	0.485
Proposed	Mixed model[h]	0.968	0.971	0.963	0.940	0.962	0.961

Higher values indicate better noise reduction.
[a]Normalization by *B*-score, standardization by Moderated T.
[b]Normalization by *Z*-score, standardization by Moderated T.
[c]Normalization by plate-wise median, standardization by Moderated T.
[d]Normalization by percent of mean of positive controls, standardization by Moderated T.
[e]Normalization by percent of median of negative controls, standardization by Moderated T.
[f]Normalization by normalized percent inhibition (NPI), standardization by Moderated T.
[g]Quantile normalization, standardization by Moderated T.
[h]Proposed mixed-effect modeling for normalization with Equations (1), (2), (4) and (5), and standardization with Equations (6)–(9). The methods in the above footnotes (a)–(g) are detailed in Supplementary Section 1.

Table 2. Average Pearson correlations of profiles of the controls, standardized as in Figure 3b, and calculated over all pairs of plates

Normalization–standardization	Evaluation samples, KO screen			
	BY4741	YDL227C	YLR396C	YPR065W
BY4741-YDL227C			0.968	0.971
BY4741-YLR396C		0.906		0.902
BY4741-YPR065W		0.985	0.968	
YDL227C-BY4741			0.966	0.960
YDL227C-YLR396C	0.811			0.838
YDL227C-YPR065W	0.979		0.970	
YLR396C-BY4741		0.980		0.974
YLR396C-YDL227C	0.974			0.973
YLR396C-YPR065W	0.975	0.979		
YPR065W-BY4741		0.977	0.966	
YPR065W-YDL227C	0.982		0.971	
YPR065W-YLR396C	0.857	0.881		

Rows: control samples used for normalization and variance estimation. Columns: validation controls.

normalization, variance estimation and validation. All combinations yield consistently high correlations, indicating that the results have little sensitivity to the specific choice of controls for the steps of the proposed procedure.

Relative contribution of analysis steps to the overall accuracy: Table 1 in Supplementary Section 9 shows average Pearson correlations of the validation controls obtained with partial normalization or variance estimation in the three screens. Normalization with respect to the covariate and estimation of residual variance terms ($\sigma^2_{B'}$ and $\sigma^2_{P'}$) contribute more to the noise reduction than the batch- and plate-wise normalization.

5.2 Evaluation based on mutant strains

The main drawback of the existing procedures is in underestimating the between-plate variation. Therefore, the number of the resulting false positive hits can exceed the nominal FDR. To illustrate this, we considered the moderated T statistics for the KO screen in Table 1, fit the two-group model to determine the test statistic cutoff at the FDR = 0.05, and determined the number of mutants with at least one differentially abundant phenotype. Supplementary Section 5 (Fig. 2) show results of the model fit for each of the procedures.

The analysis resulted in 3497 (70%) hits using *B*-score; 3709 (75%) hits using *Z*-score; 4885 (98%) hits using NPI; 4584 (92%) hits using plate-wise median; 4044 (81%) hits using percent of positive controls; 3962 (80%) hits using percent of negative control; 3359 (68%) hits using Quantile normalization. These numbers exceed the 1303 (26%) hits obtained using the proposed procedure, and likely contain some false positive hits. Although some of the reduction in the number of hits with the proposed approach can be due to a loss of sensitivity, we show in the next section that it is specific, and helps direct the follow-up experiments towards useful targets.

Detection of known changes in abundance: Eide *et al.* (2005) assayed 4358 mutants from the knock-out library in yeast, and quantified the abundance of 13 elements, namely Ca, Co, Cu, Fe, K, Mg, Mn, Ni, P, Se, Na, S and Zn. The study quantified the ionomic

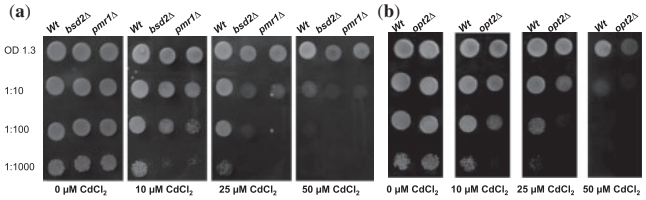

Fig. 4. Cadmium sensitivity of BY4741 wild-type (Wt) and selected mutant strains. (**a**) YBR290W (*BSD*2Δ) and YGL167C (*PMR*1Δ) to Cd supplement in growth medium. (**b**) YPR194C (*OPT*2Δ).

phenotypes with Inductively Coupled Plasma-Atomic Emission Spectroscopy (ICP-AES), is less sensitive and subject to larger variation, used different controls and no growth rate adjustments. Despite these differences in the experimental settings, the proposed approach confirmed 36 (i.e. 65%) of the KO hits reported by that study. Therefore, the proposed noise reduction procedure enables a sensitive detection of known changes in the phenotypes.

Functional annotation of differentially abundant mutant strains: to further evaluate the specificity of the proposed approach, we considered functional annotations of genes that yield mutants with at least one differentially abundant ionomic phenotype. Gene annotations were obtained from the SGD database www.yeastgenome.org, and by literature search. In particular, 37 hits in the KO screen, and 19 hits from the differentially abundant mutants in the OE screen, were involved in mineral regulation.

Three detailed examples of these mutants are YBR290W (BSD2Δ), YGL167C (PMR1Δ) and YPR194C (OPT2Δ), which were found differentially abundant in Cadmium (Cd) in the KO screen. Evidence of the involvement of these genes in Cadmium regulation has been previously established. In particular, BSD2 (bypass SOD deficiency) encodes endoplasmatic reticulum (ER)-localized membrane protein. It controls the uptake of divalent metal ions from the growth medium (Liu *et al.*, 1997). PMR1 is the major Golgi membrane-localized Ca^{2+} and Mn^{2+}-transporting P-type ATPase that has been recently shown to be essential for intracellular Cd^{2+} trafficking and detoxification (Lauer Jùnior *et al.*, 2008; Rudolph *et al.*, 1989). OPT2 is an oligopeptide transporter. The loss-of-function of OPT2 in yeast increases cells' sensitivities to anticancer drugs and divalent ion Cd (Aouida *et al.*, 2009).

Experimental validation: finally, we experimentally validated the results of a subset of 19 KO mutant strains, which were determined as differentially abundant in Cd with the proposed design and analysis methods. In the validation experiment, *S.cerevisiae cells* were grown overnight to an OD600 nm of 1.3. Aliquots of the cell suspensions were then serially diluted 10-, 100- and 1000-fold and spotted onto solid YNB medium supplemented with the indicated concentrations of $CdCl_2$. Colonies were visually assessed after incubating plates for 2 days at 30°C.

Figure 4 compares the growth of three mutant strains, YBR290W (BSD2Δ), YGL167C (PMR1Δ) and YPR194C (OPT2Δ) that were among 19 profiled mutants, and the wild-type BY4741 strain in the medium with or without Cd. The growth of the three KO strains on the medium without Cd is indistinguishable from the control strain. However, when the growth medium was supplemented with Cd, all KO strains showed more sensitivity to Cd than control strain, and the sensitivity increased with the increase of Cd concentration. These

results are consistent with the KO ionomic screen, which concluded that these lines accumulate more Cd than the median mutant. This is also consistent with the existing literature, which has established the role of BSD2, PMR1 and OPT2 in Cd detoxification (Aouida *et al.*, 2009; Lauer Jùnior *et al.*, 2008; Liu *et al.*, 1997). Similar experimental confirmation was obtained for 18 out of the 19 differentially abundant mutants that we profiled.

6 CONCLUSION

The requirements of high throughput impose constraints on the design and implementation of perturbation screens, and introduce challenges in their interpretation. Work in this article was motivated by the insights that (i) control-based normalization is most appropriate for the screens where a large proportion of samples show changes in the phenotypes, and (ii) residual non-additive effects of batch and plate variation are important components of the stochastic variation in the screens, and should be accounted for the optimal detection of hits. We proposed an experimental design that involves at least two control samples, and a normalization and variance estimation procedure based on linear mixed-effects models. Evaluations on three comprehensive ionomic screens showed that the proposed method:

- can be used in conjunction with a practical experimental design;
- allows extensions to alternative structures of data;
- enables a specific discovery of biologically meaningful hits; and
- strongly outperforms the existing approaches.

We therefore recommend this approach as a useful tool in high-throughput functional investigations.

ACKNOWLEDGEMENTS

We thank Dr C. Zheng for helpful discussions.

Funding: NSF BIO/DBI 1054826 award to DR O.V.; NSF (DBI-0606193) and NIH (4R33DK070290-02) awards to Dr D.E.S.; NSF (MCB-0923731) award to Dr O.K.V.

Conflict of Interest: none declared.

REFERENCES

Aouida,M. *et al.* (2009) Novel role for the Saccharomyces cerevisiae oligopeptide transporter Opt2 in drug detoxification. *Biochem. Cell Biol.*, **87**, 653–661.

Bankhead,A. *et al.* (2009) Knowledge based identification of essential signaling from genome-scale siRNA experiments. *BMC Syst. Biol.*, **3**, 80.

Baryshnikova,A. *et al.* (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods*, **7**, 1017–1024.

Baxter,I. (2009) Ionomics: studying the social network of mineral nutrients. *Curr. Opin. Plant Biol.*, **12**, 381–386.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Bharucha,N. and Kumar,A. (2007) Yeast genomics and drug target identification. *Comb. Chem. High Throughput Screen*, **10**, 618–634.

Birmingham,A. *et al.* (2009) Statistical methods for analysis of high-throughput rna interference screens. *Nat. Methods*, **6**, 569–575.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Boone,C. *et al.* (2007) Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.*, **8**, 437.

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

Boutros,M. *et al.* (2006) Analysis of cell-based RNAi screens. *Genome Biol.*, **7**, R66.

Collins,S. *et al.* (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol.*, **7**, R63.

Danku,J. *et al.* (2009) A high-throughput method for Saccharomyces cerevisiae (yeast) ionomics. *J. Anal. At. Spectrom.*, **24**, 103–107.

Dobbin,K. and Simon,R. (2002) Comparison of microarray designs for class comparison and class discovery. *Bioinformatics*, **18**, 1438.

Efron,B. (2008) Microarrays, Empirical Bayes, and the two-groups model. *Stat. Sci.*, **23**, 1–22.

Eide,D.J. *et al.* (2005) Characterization of the yeast ionome: a genome-wide analysis of nutrient mineral and trace element homeostasis in saccharomyces cerevisiae. *Genome Biol*., **6**, R77.

Forsburg,S.L. (2001) The art and design of genetic screens: yeast. *Nat. Rev. Genet.*, **2**, 659–668.

Gstaiger,M. and Aebersold,R. (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.*, **10**, 617–627.

Hoaglin,D. *et al.* (1983) *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons. pp. 404–414.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res*., **18**, 644–652.

Kaplow,I.M. *et al.* (2009) Rnaicut: automated detection of significant genes from functional genomic screens. *Nat. Methods*, **6**, 476–477.

Lauer Jùnior,C.M. *et al.* (2008) The PMR1 protein, the major yeast Ca2+-ATPase in the Golgi, regulates intracellular levels of the cadmium ion. *FEMS Microbiol. Lett.*, **285**, 79–88.

Leek,J. and Storey,J. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.

Leek,J. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

Lindstrom,M. and Bates,D. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J. Am. Stat. Assoc.*, **83**, 1014–1022.

Liu,X.F. *et al.* (1997) Negative control of heavy metal uptake by the Saccharomyces cerevisiae BSD2 gene. *J. Biol. Chem.*, **272**, 11763–11769.

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

Markowetz,F. (2010) How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol.*, **6**, e1000655.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks – a review. *BMC Bioinformatics*, **8**, 1–17.

Rieber,N. *et al.* (2009) RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens. *Bioinformatics* , **25**, 678–679.

Rudolph,H.K. *et al.* (11989) The yeast secretory pathway is perturbed by mutations in PMR1, a member of a Ca2+ ATPase family. *Cell*, **58**, 133–145.

Salt,D.E. *et al.* (2008) Ionomics and the study of the plant ionome. *Annu. Rev. Plant Biol.*, **59**, 709–733.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, Vol. 3: Iss. 1, bepress, Article 3.

Smyth,G.K. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Limma: linear models for microarray data. Springer Verlag.

Tukey,J.W. (1960) A survey of sampling from contaminated distributions. In *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford University Press, pp. 448–485.

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.

Wolfinger,R. *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

Yang,Y.H. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, **30**, e15.

Zhang,X.D. and Heyse,J.F. (2009) Determination of sample size in genome-scale rnai screens. *Bioinformatics*, **25**, 841–844.

Zhang,X.D. *et al.* (2008) Hit selection with false discovery rate control in genome-scale RNAi screens. *Nucleic Acids Res.*, **36**, 4667–4679.