BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2844–2851
doi:10.1093/bioinformatics/btt508

Genome analysis

Advance Access publication September 17, 2013

Assessing the validity and reproducibility of genome-scale predictions

Lauren A. Sugden[1], Michael R. Tackett[2], Yiannis A. Savva[3], William A. Thompson[1] and Charles E. Lawrence[1,*]

[1]Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, RI 02912, USA, [2]St. Laurent Institute, 317 New Boston St, Woburn, MA 01801, USA and [3]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI 02912, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Validation and reproducibility of results is a central and pressing issue in genomics. Several recent embarrassing incidents involving the irreproducibility of high-profile studies have illustrated the importance of this issue and the need for rigorous methods for the assessment of reproducibility.

Results: Here, we describe an existing statistical model that is very well suited to this problem. We explain its utility for assessing the reproducibility of validation experiments, and apply it to a genome-scale study of adenosine deaminase acting on RNA (ADAR)-mediated RNA editing in *Drosophila*. We also introduce a statistical method for planning validation experiments that will obtain the tightest reproducibility confidence limits, which, for a fixed total number of experiments, returns the optimal number of replicates for the study.

Availability: Downloadable software and a web service for both the analysis of data from a reproducibility study and for the optimal design of these studies is provided at http://ccmbweb.ccv.brown.edu/reproducibility.html

Contact: Charles_Lawrence@Brown.edu

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

Received on June 5, 2013; revised on August 26, 2013; accepted on August 28, 2013

1 INTRODUCTION

The issue of validation and reproducibility of scientific results has recently been the subject of intense discussion in the scientific community. Several eye-opening reports have either claimed insufficient validation of bold research findings or shown an inability to replicate such results in genomics (DeVeale *et al*., 2012; Gregg *et al*., 2010; Kleinman and Majewski, 2012; Lin *et al*., 2012; Li *et al*., 2011b; Pickrell *et al*., 2012), genetics (Hunt *et al*., 2012; Surolia *et al*., 2010), oncology (Begley and Ellis, 2012), neuroscience (Button *et al*., 2013), pharmacology (Prinz *et al*., 2011), proteomics (Bell *et al*., 2009) and psychology (Shanks *et al*., 2013; Yong, 2012). Problems with reproducibility have been demonstrated with widely-used technologies such as microarrays (Ioannidis *et al*., 2009), siRNA-based screens (Barrows *et al*., 2010) and mass spectrometry (Bell *et al*., 2009). This has led to appeals for increased statistical rigor (Macleod, 2011; Vaux, 2012), platforms for the publication of neutral studies (Macleod, 2011) and attempted replicates, whether successful or not (Editorial, 2012a), and a system-wide committed effort toward generating work that is reproducible (MacArthur, 2012), placing at least as much emphasis on reproducibility as is currently placed on novelty (Editorial, 2012b; Russell, 2013).

Recent attempts at addressing the issue of reproducibility include the Reproducibility Initiative, which, for a fee, will carry out independent validation of research findings and issue a 'certificate of reproducibility' for those studies that validate (https://www.scienceexchange.com/reproducibility), and ScienceCheck, which provides a platform for researchers to report on the 'reproducibility and utility of the literature method(s) that they have worked with' (http://www.sciencecheck.org). Although these are extremely important contributions, neither organization provides a quantitative measure of reproducibility.

In light of this, there is an urgent need for statistical tools for quantitatively evaluating reproducibility. To help address this need, we introduce the application of a well-suited Bayesian hierarchical model for assessing the reproducibility of validation experiments in the context of evaluating top-tier predictions of high-throughput genomic studies. We focus on studies in which a large number of predictions are made concerning a biological phenomenon of interest. There are many studies of this type in the recent literature, in *Drosophila* alone; Hoskins *et al*. (2011) predict 2000 new gene promoters, Li *et al*. (2008) identify thousands of targets of six transcription factors involved in regulation of the anterior–posterior axis in the embryo, Nègre *et al*. (2010) find >14 000 binding sites of six proteins associated with insulators, DNA sequences that block the spread of regions of modified chromatin and interaction between other regulatory elements, and Zeitlinger *et al*. (2007) find evidence for 1600 genes whose transcription start sites are sites of polymerase II stalling. Because validation of all predictions is typically infeasible, often a few compelling and biologically interesting cases are selected for further study (Hughes, 2009), leaving a long list of unvalidated predictions. The reader is left unsure about both the fraction of the list that is valid and the effect of biological and sample preparation variation.

Our model takes advantage of multiple biological and technical replicates, in each of which validation of a random sample

*To whom correspondence should be addressed.

of the top-tier list is carried out. From these data, we can assess the reproducibility of the validation studies and predict what another investigator could reasonably expect to see in a follow-up study.

The use of replicates, whether technical, biological or simulated, has been shown to be useful in many contexts. McShane *et al.* (2002) and Kerr and Churchill (2001) simulate microarray replicates to determine the stability of clusters of genes that exhibit similar expression patterns. In the search for differentially expressed genes, technical replicates provide additional power for microarrays (Pan *et al.*, 2002), and biological replicates reduce false positives in conclusions drawn from serial analysis of gene expression data (Baggerly *et al.*, 2003; Vêncio *et al.*, 2004) and improve accuracy in calls made from RNAseq data (Glaus *et al.*, 2012). Xia *et al.* (2011) use replicate time series datasets to capture time-delayed associations between microbes. In a larger-scale take on the replicates, meta-analyses of genome-wide association studies like those described in Zeggini and Ioannidis (2009) combine datasets from multiple laboratories to gain enough power to detect associations between particular genes and diseases such as type II diabetes (Zeggini *et al.*, 2008) and Crohn's disease (Barrett *et al.*, 2008).

In some cases, replicates have been incorporated into optimal study design. Many articles have been written on the number of replicates required to detect a certain fold-change in gene expression via microarray studies (Black and Doerge, 2002; Pan *et al.*, 2002; Tibshirani, 2005; Wei *et al.*, 2004). For genome-wide association studies, Moonesinghe *et al.* (2008) find the required number of samples to replicate an association across studies with a certain level of between-study heterogeneity, and Pahl *et al.* (2009) propose multistage designs which, for a given budget, maximize the power to find associations. Auer and Doerge (2010) advocate careful design of RNA-seq experiments, including sampling, randomization, replication and blocking. To our knowledge, no one in the biology community has used biological or technical replicates to assess the reproducibility of validation studies like those discussed here, or has proposed a method for optimal design of such experiments with respect to reproducibility.

There has been considerable work on assessing the reproducibility of high-throughput experiments, especially in the context of ranked lists of putative sites (Boulesteix and Slawski, 2009). In this context, reproducibility is most closely related to precision (or stability), in that the relevant issue is the similarity of two ranked lists generated from biological replicates (or different high-throughput platforms, different ranking algorithms, etc…). There are many different measures used to assess the similarity of two or more ranked lists, from Spearman's rank correlation (Kuo *et al.*, 2006; MAQC Consortium, 2006) to overlap counts for the top k sites (Zhang *et al.*, 2009) to weighted overlap counts that emphasize correlation between high ranking sites over that of low ranking sites (Yang *et al.*, 2006). Li *et al.* (2011a) improve on these measures with a mixture model consisting of reproducible and irreproducible sites, which assigns each signal a reproducibility index based on its consistency across replicates, which approximates its probability of being reproducible. They define the 'irreproducible discovery rate' (IDR), an analog of the false discovery rate for multiple hypothesis testing (Storey, 2002), which determines the 'expected rate of irreproducible discoveries' for sites whose probability of being irreproducible is below some threshold γ. Their methods provide a principled method for selecting sites for further study and for evaluating ranking algorithms. Although here we also address the issue of reproducibility, our focus is different. We are not concerned with the precision of high-throughput technologies or ranking algorithms, but rather with the reproducibility of independent validation experiments that seek to verify findings of such high-throughput experiments. The validation experiments taken individually give us information about the accuracy of the findings, whereas our model of biological replicates assesses the reproducibility of the given validation scheme in the face of biological and sample preparation variation.

Because the model we describe depends on validation of random samples, here we first review how a single simple random sample drawn from the top-tier list can be used to estimate the valid fraction of top-tier predictions. Because this method does not account for biological and sample preparation variability, it is not sufficient to assess reproducibility, as factors as seemingly benign as laboratory conditions, reagent lots, cell generations and individual experimenter techniques have been shown to affect results of biological experiments (Barrows *et al.*, 2010; Leek *et al.*, 2010; Van Hijum *et al.*, 2005). So motivated, we describe how our hierarchical model uses data from multiple replicates to compute a probability distribution of validation results for an as-yet-unseen replicate. Hierarchical models, described in many statistical textbooks including Gelman *et al.* (2003), have many uses in computational genomics (Ji and Liu, 2010), and are well suited to the task of assessing reproducibility, as they provide a way to simultaneously model similarities and differences between groups.

2 METHODS

2.1 Estimating validity

In a genome study with thousands of predictions, validation of select predictions is an important step toward lending credibility to those particular findings, but provides little, if any, support for the validity of the other predictions of the study. Thus, follow-up studies must be carried out without any confidence in the validity of the findings they are pursuing. If we ignore biological and sample preparation variability, a simple yet rigorous way to address this is for the original investigators to draw a random sample of their predictions to validate. The number valid can be modeled by a binomial distribution, so the investigator can estimate the fraction valid in the full top-tier list, complete with confidence limits to assess uncertainty (equations in Supplementary Table S1). These confidence limits can then be translated into lower confidence bounds, to give an idea of the worst-case scenario. For example, in a study with a top-tier list of 1000 predictions, if the lower bound of the 90% confidence interval is 0.8, we are assured that at least 800 of the predictions are valid with 95% confidence. Our model builds on this idea by addressing the effects of biological and sample preparation variation on reproducibility.

2.2 Hierarchical model and predictive distribution

We consider a result reproducible if it can be obtained in an independent analysis, following the exact protocol provided by the original investigators, under the same experimental conditions. Because the confidence intervals described above are based on a single replicate and do not take into account the effects of biological and sample preparation variability, they do not provide a basis for assessing the reproducibility of the validation study. In particular, such intervals will be deceptively narrow for predicting what will happen in a new replicate. With this in mind, we must allow the proportions found to be valid in each replicate to vary from that in the original validation assay.

The hierarchical model provides a balance between treating each replicate independently, leading to noisy estimates of the proportions valid in each pool, and combining data from all replicates, ignoring inter-pool variation. As shown in Figure 1, it achieves this balance by modeling the similarities and differences in the proportions valid in each replicate with a probabilistic function, $f(p|\alpha, \beta)$. Here, the function is a beta distribution, the natural conjugate distribution to the binomial. The proportions p_i valid in each pool are assumed to be independent samples from this distribution. Because predictions to be tested in each pool are drawn at random, p_i is the probability that a randomly drawn prediction will be valid in the i^{th} pool; thus, as in the single replicate case above, we model the number k_i found valid in the i^{th} pool with a binomial distribution, $f(k|p, N)$ in Figure 1. The distribution of parameters α and β at the top of the hierarchy is unknown, and thus must be inferred from the data coming from all samples in all pools at the bottom of the hierarchy. If the proportions in the pools are quite similar, the resulting distribution will have a tight variance, reflecting the fact that the inter-pool variation is small. On the other hand, if the proportions in the pools differ substantially, the variance of this distribution will be large. This model allows us to make predictions about the results of a new replicate, as indicated by p_{new} and k_{new} in Figure 1, and allows for tighter confidence limits in each replicate, as illustrated in Supplementary Figure S1.

2.2.1 Structure of the model Given N_T total predictions from a genome-wide study, we consider m biological and technical replicate pools in which $N_i << N_T$ randomly selected predictions are tested $(i = 1, \ldots, m)$. In a given replicate pool i, let k_i be the number of the Ni predictions that successfully validate. The number of predictions k_i that validate in pool i is modeled as a draw from a binomial distribution with parameters p_i and N_i:

$$P(k_i|p_i, N_i) = \binom{N_i}{k_i} p_i^{k_i}(1 - p_i)^{N_i - k_i} \tag{1}$$

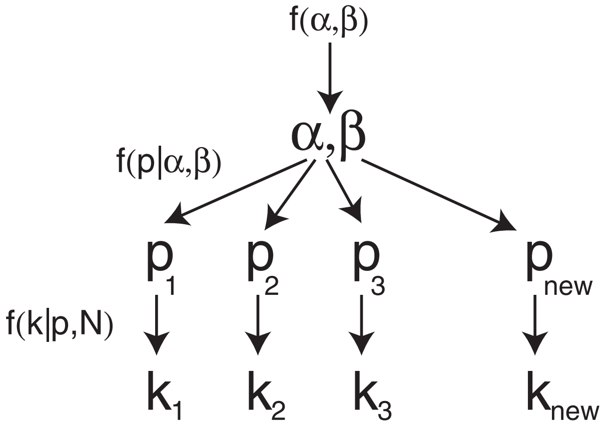

Fig. 1. Diagram of the hierarchical model. k_1, k_2 and k_3 represent the counts of successful validations in three experiments, out of N_1, N_2 and N_3 total validation experiments, respectively. Each k is a draw from a binomial distribution with corresponding parameters p and N. Each p, in turn, is a draw from a beta distribution with parameters α and β. The available data inform the common distribution of α and β, from which we can simulate the results of a new experiment

where p_i for each replicate is modeled as a draw from a common hierarchical beta distribution with parameters α and β:

$$f(p_i|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1 - p_i)^{\beta-1} \tag{2}$$

Since in general we expect that the number of replicates will be too small to result in reliable point estimates for α and β, we use Bayesian inference at the top level of the model to compute a probability distribution for these parameters, $f(\alpha, \beta)$ in Figure 1. We use a hyperprior suggested by Gelman *et al.* (2003) on α and β, a uniform distribution on transformed axes $\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{\sqrt{\alpha+\beta}}\right)$. This distribution is 'uninformative,' or 'diffuse' in the sense that it does not place any large probability mass in any one place. Because α and β govern the position and shape of the beta distribution, we use this uninformative prior to refrain from imposing prior assumptions about the amount of variation that actually exists among replicates.

2.2.2 Posterior distribution of hyperparameters There is no closed-form expression for inference of the posterior distribution of α and β, so we use the grid-based approach outlined in Gelman *et al.* (2003). The un-normalized posterior distribution of α and β given data k_i, $i = 1, \ldots, m$ is given by the following expression:

$$f(\alpha, \beta|k_1, \ldots, k_m) \propto f(\alpha, \beta)\prod_{i=1}^{m}\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha + k_i)\Gamma(\beta + N_i - k_i)}{\Gamma(\alpha + \beta + N_i)} \tag{3}$$

where $f(\alpha, \beta)$ represents the prior distribution. Details on computing and drawing from this distribution can be found in Supplementary Methods.

2.2.3 The predictive distribution To ask what the results of a validation study in a new replicate might look like, we draw 1000 samples from the posterior distribution on α and β, as described in Supplementary Methods. For each pair (α_j, β_j), $j = 1, \ldots, 1000$, we draw p_j from a beta distribution with parameters α_j and β_j. Given a sample size N, we can then draw k_j from a binomial distribution with parameters p_j and N. From these samples, we can approximate the predictive distribution and compute informative statistics such as mean, variance and percentiles.

2.3 Optimal study design

The model described above can also be used to design a validation study in a way that yields the most favorable (tightest) predictive distribution. There are four necessary input parameters: the total number of predictions to be validated (N), expected mean and standard deviation of the distribution of validation rates across replicate pools (p_E and σ_E) and the fraction of validation experiments expected to yield neither a positive nor a negative result (q_E). This last parameter acknowledges the fact that for many protocols involving techniques such as polymerase chain reaction (PCR) or other common tools, only a fraction of experiments will work (e.g. primers may fail to bind). Given these four parameters, we can determine the most effective way to distribute experiments across replicates, under the assumption that the replicate pools all have approximately equal sample sizes. This is done via simulation as follows:

(1) Split N into a representative subset of the possible numbers of replicates (N_{rep}) with $N_{\text{exp}} = \frac{N}{N_{\text{rep}}}$ experiments per replicate (see Supplementary Table S2 for details).

(2) For each N_{rep} (and corresponding N_{exp}):

 (a) Generate samples from the hierarchical model:

  (i) Generate N_{rep} values of p_i from a beta distribution with mean p_E and standard deviation σ_E

  (ii) For $i = 1, \ldots, N_{\text{rep}}$, simulate experiment failure by drawing $\hat{N}_{\text{exp},i}$ from independent binomial distributions with parameters (N_{exp}, q_E)

(iii) For $i = 1, \ldots, N_{\text{rep}}$, draw the number of positive validations k_i from a binomial distribution with parameters $\hat{N}_{\exp,i}$ and p_i

(b) From generated samples k_i and $\hat{N}_{\exp,i}$ for all i, infer the posterior distribution of α and β, and use this distribution to infer the predictive distribution of a new replicate as described above, taking note of the standard deviation and 10th percentile.

(c) Repeat from (a) 3000 times, averaging the results for standard deviation and 10th percentile

(3) Select the value of N_{rep} with the most favorable characteristics (low standard deviation, high 10th percentile). Often there is a range of options that give similar results.

3 RESULTS

3.1 Application to adenosine deaminase acting on RNA study

3.1.1 Target prediction and validation In a related manuscript (St. Laurent *et al.*, 2013), we produced a number of top-tier-predicted adenosine deaminase acting on RNA (ADAR) targets in *Drosophila*. ADAR enzymes target double-stranded RNAs (Nishikura *et al.*, 1991), catalyzing the conversion of adenosine (A) to inosine (I) through hydrolytic deamination (Bass and Weintraub, 1988). This post-transcriptional mechanism, also known as RNA editing, has the capacity to diversify genomes via amino acid recoding in functionally important protein residues (Nishikura, 2010). *Drosophila* has a single ADAR protein called dADAR, which has been shown to target protein-coding genes involved in vesicular trafficking, ion homeostasis, signal transduction, ion channels and the cytoskeleton (Hoopengardner *et al.*, 2003; Stapleton *et al.*, 2006), and is crucial for normal adult nervous system function (Palladino *et al.*, 2000). Variability in RNA editing between flies has even been suggested as a mechanism for individual differences in behavior and neuronal physiology (Jepson *et al.*, 2012).

In the related study, we used a combination of single-molecule sequencing, previously validated sites and machine learning methods to predict 1782 top-tier sites of dADAR-mediated RNA editing. As RNA editing converts an A to an I in the RNA sequence, and sequencing machinery reads an I as a G, validation consisted of Sanger sequencing to identify sites in RNA that express a G (or a mixture of A and G) where the genome contains an A. The first validation experiment was carried out in a single pool on 298 predicted sites. For ~30% of the sites, the reads were of too poor quality to assess the absence or presence of RNA editing, leaving 205 sites, of which 151 (74%) successfully validated.

To obtain data on inter-pool variation from the combined effect of biological variation and sample preparation technique, we conducted independent Sanger validations in four additional RNA pools created by a skilled investigator well versed in the protocol, and in a fifth pool created by an investigator with much less experience, which we discarded after finding it to be of much lower quality than the others (details in Supplementary Table S3). All pools in this study were isolated from wild-type *Drosophila* (Canton-S) raised at a constant 25°C on standard molasses food and under 12-h day/night cycles. RNA was extracted from adult, whole body, 1–2-day-old males (10 per sample), using TRIzol reagent (Invitrogen). For validation, cDNAs were amplified via reverse transcriptase-polymerase chain reaction using gene-specific primers. The data from the final five replicates are shown in Table 1. The first pool was larger than the others because the authors of the related manuscript wanted to investigate sequence subcategories: exons, introns and intergenic regions.

3.1.2 Predictive distribution and interpretation We used the Bayesian procedure described in Section 2 to carry out predictive inference based on the five replicates from the ADAR study. The contour plot in Supplementary Figure S2 illustrates the variation in the parameters of the beta distribution, revealing at least moderate uncertainty in our knowledge of their values based on the data, which supports our decision to avoid using point estimates. The predictive distribution, illustrated in Figure 2, has a mean of 67%, meaning that in a new validation study following the same protocol under the same experimental conditions, 67% of sites would successfully validate on average. The 80% credibility limit of the distribution has a lower bound of 55%, indicating that 90% of the time, at least 55% of sites will successfully validate. It is possible that the requirement that further studies follow the same protocol under the exact experimental conditions may be

Table 1. Data from ADAR replicates and cross-validation confidence/credibility intervals

RNA pool	Valid/Total	Cross-validation 80% CI		Cross-validation 90% CI		Cross-validation 95% CI	
		Predictive distribution	Pooled	Predictive distribution	Pooled	Predictive distribution	Pooled
1	151/205 (74%)	(0.47, 0.77)	(0.60, 0.71)	(0.36, 0.83)	(0.58, 0.72)	(0.23, 0.88)	(0.57, 0.73)
2	17/32 (53%)	(0.60, 0.80)	(0.69, 0.76)	(0.52, 0.84)	(0.68, 0.77)	(0.42, 0.89)	(0.67, 0.77)
3	21/30 (70%)	(0.46, 0.82)	(0.67, 0.74)	(0.38, 0.87)	(0.66, 0.75)	(0.29, 0.92)	(0.65, 0.76)
4	24/32 (75%)	(0.47, 0.82)	(0.67, 0.73)	(0.39, 0.87)	(0.66, 0.74)	(0.26, 0.92)	(0.65, 0.75)
5	18/29 (62%)	(0.50, 0.83)	(0.68, 0.75)	(0.40, 0.87)	(0.67, 0.76)	(0.33, 0.91)	(0.66, 0.76)

Note: Column 2 shows the results from the five replicate datasets from the ADAR study. Columns 3–8 show the credibility/confidence intervals generated in the cross-validation studies described in Section 3.3.1, where the interval is generated either from the model predictive distribution, or the standard frequentist normal approximation on the pooled data, where the left-out pool is the row pool (column 1).

difficult to satisfy, in which case our limits may be too narrow. Nevertheless, this is valuable information for an investigator seeking to build on this study.

3.2 Optimal study design

As was the case in our ADAR study, the time and expense involved in validation studies often come at the level of the validation experiments themselves, and comparatively little effort is needed for the creation of multiple replicate pools. In such a situation, researchers have a good deal of flexibility in deciding how to divide a fixed number of experiments into a variable number of replicates. For example, if they are willing at the outset to do 96 experiments, they could run 48 experiments each in two replicate pools, two experiments each in 48 replicate pools or any combination in between. The shape and spread of the resulting predictive distribution will depend on these choices: the number of replicates and the associated number of experiments per replicate.

We used the experimental design procedure described in Section 2 to compute the optimal number of replicates for studies with a wide range of input parameters (total number of experiments, expected mean and standard deviation and fraction of experiments expected to fail from the start), for use by investigators planning validation studies. An example of results for one set of input variables is illustrated in Figure 3, which shows the average standard deviation and 10th percentile of the predictive distribution with increasing numbers of replicates. In most cases, as in this one, uncertainty about reproducibility achieves a minimum value after an initial decrease. This change in uncertainty is reflected in the 10th percentile, which at first increases reaching a maximum at nine replicates and then decreases. The initial increase stems from the fact that for small numbers of replicates, the increase in our knowledge about inter-pool variation outweighs the reduction of certainty about the proportion in each replicate. This holds up to nine replicates in this case, at which point intra-pool uncertainty stemming from the smaller sample sizes in each replicate begins to outweigh the gain in certainty about the inter-pool variation. The complementary trend is embodied by the standard deviation curve. Because a predictive distribution with narrow standard deviation and high 10th percentile is desirable, our algorithm would recommend around nine replicates for this validation study.

In Figure 4, we illustrate the qualitative effect of the four parameters in the system on reproducibility by displaying the 10th percentile of the optimal predictive distribution for given 4-tuples of parameters. This optimal 10th percentile rises with increasing number of experiments, expected mean and fraction of successful experiments, and falls with increasing standard deviation. A more in-depth illustration of the effect of expected standard deviation on the predictive distribution is given in Supplementary Figure S3.

3.3 Cross-validation of ADAR replicates

Because it would require a very large number of replication experiments to obtain a sufficiently large sample to have enough data to persuasively validate the confidence limits we describe, we undertook a cross-validation study based on the five available replicates. Leaving out each replicate in turn, we computed the predictive distribution given the remaining four replicates. For all five cross-validations, the proportion valid in the left-out replicate fell within the 90% credibility interval of the

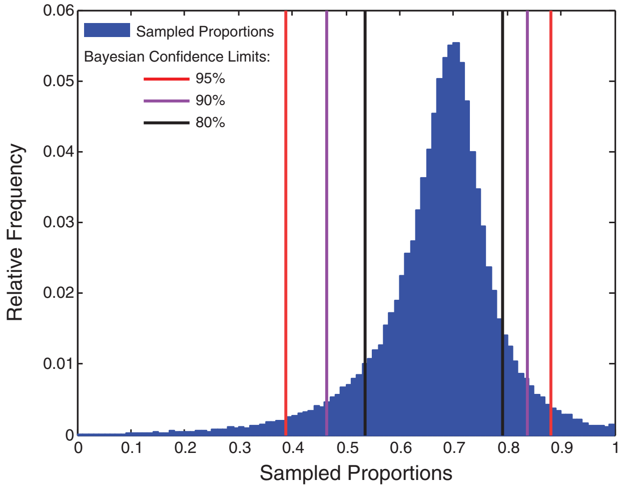

Fig. 2. The predictive distribution. The histogram represents the predictive distribution of *p* for a new replicate following the same experimental procedure that was performed in the five ADAR replicate pools. The distribution has a mean of 0.67 with 95% Bayesian confidence limits running from 0.39 to 0.88, 90% confidence limits running from 0.47 to 0.84 and 80% confidence limits running from 0.54 to 0.79. We report confidence intervals to give a sense of best- and worst-case scenarios. In particular, using the lower limits of the 80 and 90% confidence intervals, respectively, we see that only 10% of the probability mass lies <0.54 and only 5% lies <0.47

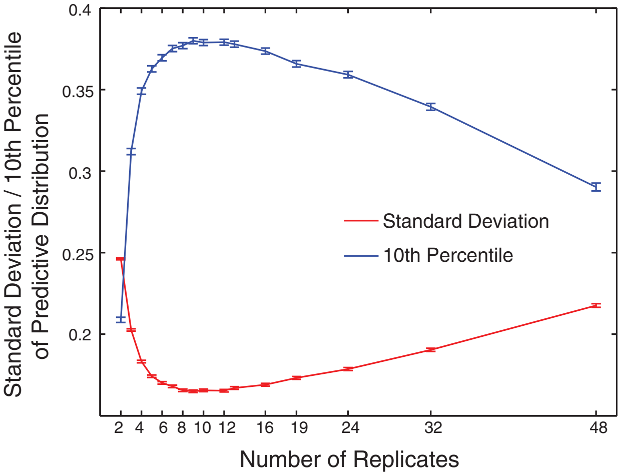

Fig. 3. Software-generated curves. Software inputs here are the total number of experiments (96), the expected overall mean (0.6) and standard deviation between replicates (0.09) and the percentage of experiments expected to work (0.6). The red curve represents the standard deviation of the predictive distribution for a new replicate, and the blue curve represents the 10th percentile of the same distribution, each computed for different numbers of replicates to identify the optimal number of replicates to use. Here, 9 replicates with 10–11 experiments each seem to give the tightest predictive distribution with the highest 10th percentile

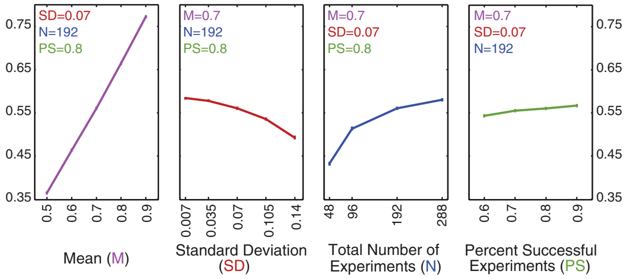

Fig. 4. Influence of parameters on optimal predictive distribution. Here we see the effect of varying the four input parameters: expected mean, expected standard deviation, expected fraction of successful experiments and total number of experiments. For each curve, we held three parameters fixed and varied the parameter of interest. For each 4-tuple, we found the maximum 10th percentile over all predictive distributions for all assignments of experiments to replicates. From the plot, we see that the desirable higher 10th percentiles result from data with high mean, low standard deviation, large numbers of experiments and high fraction of successful experiments

corresponding predictive distribution, and for four of five, the proportion valid fell within the 80% credibility interval. Thus, these results are fully consistent with the model's expectations. However, because these results are based on a small cross-validation dataset, the support for our model from this analysis is quite limited.

We carried out a similar cross-validation with our five pools, not accounting for biological or sample preparation variability. Leaving out each replicate in turn, we pooled the other four replicates and computed standard frequentist confidence limits based on a normal approximation to the binomial distribution (equations in Supplementary Table S1). In four of five trials, the left-out proportion lay outside of both the 80% and 90% confidence intervals, and in three of five, the left-out proportion lay outside of the 95% confidence interval. Cross-validation intervals for each method can be found in Table 1. These events, although not statistically significant at the standard 0.05 level, are in stark contrast with the cross-validation results above, revealing that the available data are less likely under the assumption of no biological or sample preparation variation than under our model including variation. This supports our contention that such variation plays an important role in studies of editing in *Drosophila*, and suggests that it may also be important in other genome-scale studies. We found similar results in a cross-validation using Bayesian methods to compute credibility limits for the pooled data (Supplementary Table S4), demonstrating that our model performs better not because we use Bayesian methods, but because we account for variation.

4 DISCUSSION

The need for reproducibility in scientific research has always been central, but has only recently become a major focus of the greater scientific community. Here, we present a procedure that addresses these issues in the context of high-throughput studies like that described in our companion article, where thousands of predictions are made, and only a relatively small fraction can be validated. We studied closely the example of ADAR editing sites, but the method is generalizable, and could just as easily apply to any of the high-throughput site prediction experiments described earlier. Studies of this kind have become more frequent with the advent of high-throughput technologies like Illumina and large collaborations like ENCODE, modENCODE and the 1000 Genomes Project. Most studies already have their own schemes for validation, which they carry out with varying levels of statistical rigor. The authors and any investigators hoping to carry out follow-up studies would all benefit from a carefully designed validation study using statistically random samples in biological replicates to assess reproducibility.

As the accuracy of technology inevitably grows, it may be tempting for investigators to assume a single replicate is sufficient to address reproducibility, implicitly assuming that technical variation is the only variation that matters. However, even with perfect technology, multiple biological replicates are still necessary to assess the reproducibility of a set of results. It is worth noting that each of our validation replicates was performed in a pool of 10 flies each. We expect that biological replicates will be even more crucial in studies where replicates consist of individual model organisms, such as mice or rats.

In our analysis of ADAR-mediated editing data, we found that the 95% confidence intervals for individual replicates showed substantial variation. This was not unexpected, given the documented variation in ADAR activity in individual flies (Jepson *et al.*, 2012) and observations on the effects of experimental conditions described earlier, and it underscored the need for statistical tools that can address the effect of such variation on the reproducibility of results. Our software predicted that a new experiment using the same protocol under the same experimental conditions would validate on average 67% of sites, and that 90% of the time, the percentage validated would be at least 55%. We emphasize the lower bound percentile, e.g. the 10th percentile, because it represents a worst-case scenario for a future experiment. Of course, because these results are affected by our choice of the diffuse prior on α and β, the intervals we generate may be too conservative in the case where more is known about these parameters *a priori*.

There is a technical limitation to our model that arises from the statistical formulation. The predictive distribution we describe is well defined everywhere except in the precise circumstance in which every replicate pool has a validation rate of either 100 or 0% (Gelman *et al.*, 2003). We consider such extremes to be very unlikely in most validation studies; however, this limitation occasionally comes to bear in experimental design, where we simulate thousands of distributions. This affects almost all simulations with means >95% and most simulations with means ~90% and with wide variances. Otherwise, the effect is negligible. Our software will not return results in the few non-negligibly affected cases.

Our model makes two major assumptions: that the validation experiments in each replicate follow a binomial distribution, and that the proportions valid in each replicate follow a beta distribution. Because we require that each replicate test a random sample drawn from the whole population of predictions, the results of each replicate follow a hypergeometric distribution, which is very well approximated by a binomial distribution as long as the number of predictions n tested in a single replicate is much smaller than the total number of predictions N (a typical rule of thumb is $N \geq 20n$). The beta distribution was chosen for the model primarily because as a conjugate prior to the binomial distribution, it makes computation feasible. However, another real advantage is that the beta distribution is extremely flexible, in that it can approximate most smooth unimodal distributions. We illustrate this flexibility in the context of our model in Supplementary Figure S4. Together, this suggests that the assumptions of our model will be appropriate for most applications.

There are three places in our model where we rely on numerical approximations: during predictive inference, we compute the posterior distribution of the hyperparameters over a grid, as there is no closed-form expression for computing it directly, then we sample from the grid to approximate the predictive distribution; and in our optimal study design, we rely on a sampling approximation to find the average mean, standard deviation and 10th percentile of the population of predictive distributions resulting from particular input parameters. For predictive inference, we follow the procedure outlined in Gelman *et al.*, computing over a dense-enough grid that we believe captures the important features of the posterior distribution, and sampling a large number of points (1000) to approximate the predictive distribution. We found that neither varying the grid density nor increasing the number of samples noticeably changed the results (data not shown). For study design, we sample a large number of distributions (3000 is the default), and if the user downloads our software, he or she can increase the number of samples if desired, so as to obtain narrower error bars.

Finally, it should be noted that the results of any reproducibility analysis can only be generalized to the population to which the replicates belong (just as the results of any study should only be generalized to the population from which the data are drawn). We assume that follow-up validation studies follow the original protocol under the exact experimental conditions as the original experiments. To the extent that this is not possible, the credibility intervals that we report may be too narrow to accurately reflect the population of follow-up validation studies performed by other investigators in other laboratories. Therefore, care must be exercised in how claims of reproducibility are made, and authors should be sure to specify the population to which their results generalize. In some cases, large collaborations between laboratories (such as those associated with modENCODE) will be able to carry out replicates that represent a larger portion of the possible variability, and will be able to make even stronger claims about the reproducibility of their findings.

Validation and reproducibility are bedrock principles throughout science that have until recently received limited attention. We present this work as an aid in advancing these crucial principles in the field of genomics.

ACKNOWLEDGEMENTS

The authors thank their collaborators Robert Reenan, Georges St. Laurent and Philipp Kapranov for valuable discussions throughout the project, and the reviewers for their insightful comments and suggestions.

Funding: This work was supported the National Institutes of Health [NS074686] (in part) and by an Ellison Medical Foundation Senior Scholar Award to R.R.

Conflict of Interest: None declared.

REFERENCES

Auer,P.L. and Doerge,R.W. (2010) Statistical design and analysis of RNA sequencing data. *Genetics*, **185**, 405–416.

Baggerly,K.A. *et al.* (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, **19**, 1477–1483.

Barrett,J.C. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.

Barrows,N.J. *et al.* (2010) Factors affecting reproducibility between genome-scale siRNA-based screens. *J. Biomol. Screen.*, **15**, 735–747.

Bass,B.L. and Weintraub,H. (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell*, **55**, 1089–1098.

Begley,C.G. and Ellis,L.M. (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Bell,A.W. *et al.* (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Meth.*, **6**, 423–430.

Black,M.A. and Doerge,R.W. (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. *Bioinformatics*, **18**, 1609–1616.

Boulesteix,A.L. and Slawski,M. (2009) Stability and aggregation of ranked gene lists. *Brief. Bioinform.*, **10**, 556–568.

Button,K.S. *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.*, **14**, 365–376.

DeVeale,B. *et al.* (2012) Critical evaluation of imprinted gene expression by RNA-seq: a new perspective. *PLoS Genet.*, **8**, e1002600.

Editorial. (2012a) Further confirmation needed. *Nat. Biotechnol.*, **30**, 806.

Editorial. (2012b) Error prone: biologists must realize the piffalls of work on massive amounts of data. *Nature*, **487**, 406.

Gelman,A. *et al.* (2003) Hierarchical models. In: *Bayesian Data Analysis*. 2nd edn. Chapman and Hall/CRC, pp. 120–160.

Glaus,P. *et al.* (2012) Identifying differentially expressed ranscripts from RNA-seq data with biological variation. *Bioinformatics*, **28**, 1721–1728.

Gregg,C. *et al.* (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643–648.

Hoopengardner,B. *et al.* (2003) Nervous system targets of RNA editing identified by comparative genomics. *Science*, **301**, 832–836.

Hoskins,R.A. *et al.* (2011) Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.*, **21**, 182–192.

Hughes,T.R. (2009) 'Validation' in genome-scale research. *J. Biol.*, **8**, 3–5.

Hunt,K.A. *et al.* (2012) Rare and functional SIAE variants are not associated with autoimmune disease risk in up to 66,924 individuals of European ancestry. *Nat. Genet.*, **44**, 3–5.

Ioannidis,J.P.A. *et al.* (2009) Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 149–155.

Jepson,J.E.C. *et al.* (2012) Visualizing adenosine-to-inosine RNA editing in the drosophila nervous system. *Nat. Meth.*, **9**, 189–194.

Ji,H. and Liu,X.S. (2010) Analyzing omics data using hierarchical models. *Nat. Biotech.*, **28**, 337–340.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *PNAS*, **98**, 8961–8965.

Kleinman,C.L. and Majewski,J. (2012) Comment on Widespread RNA and DNA Sequence Differences in the Human Transcriptome. *Science*, **335**, 1302.

Kuo,W. *et al.* (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.*, **24**, 832–840.

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

Li,Q. *et al.* (2011a) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Li,M. *et al.* (2011b) Widespread RNA and DNA sequence differences in the Human Transcriptome. *Science*, **333**, 53–58.

Li,X. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila blastoderm*. *PLoS Biol.*, **6**, e27.

Lin,W. *et al*. (2012) Comment on widespread RNA and DNA sequence differences in the human transcriptome. *Science*, **335**, 1302.

MacArthur,D. (2012) Face up to false positives. *Nature*, **487**, 427–428.

Macleod,M. (2011) Why animal research needs to improve. *Nature*, **477**, 511–511.

MAQC Consortium. (2006) The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol*., **24**, 1151–1161.

McShane,L.M. *et al*. (2002) Methods for assessing reproducibility of clustering patterns observed in analysis of microarray data. *Bioinformatics*, **18**, 1462–1469.

Moonesinghe,R. *et al*. (2008) Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *PNAS*, **105**, 617–622.

Nègre,N. *et al*. (2010) A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet*., **6**, e1000814.

Nishikura,K. *et al*. (1991) Substrate specificity of the dsRNA unwinding/modifying activity. *EMBO J*., **10**, 3523–3532.

Nishikura,K. (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem*., **79**, 321–349.

Pahl,R. *et al*. (2009) Optimal multistage designs – a general framework for efficient genome-wide association studies. *Biostatistics*, **10**, 297–309.

Palladino,M.J. *et al*. (2000) A-to-I Pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell*, **102**, 437–449.

Pan,W. *et al*. (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol*., **3** research 0022.

Pickrell,J.K. *et al*. (2012) Comment on widespread RNA and DNA sequence differences in the human transcriptome. *Science*, **335**, 1302.

Prinz,F. *et al*. (2011) Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov*., **10**, 712–712.

Russell,J.F. (2013) If a job is worth doing, it is worth doing twice. *Nature*, **496**, 7–7.

Shanks,D.R. *et al*. (2013) Priming intelligent behavior: an elusive phenomenon. *PLoS One*, **8**, e56515.

St. Laurent,G. *et al*. (2013) Genome-wide analysis of A-to-I RNA editing via single molecule sequencing in *Drosophila*. *Nat. Struct. Mol. Biol*, in press.

Stapleton,M. *et al*. (2006) RNA editing in *Drosophila melanogaster:* new targets and functional consequences. *RNA*, **12**, 1922–1932.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol*., **64**, 479–498.

Surolia,I. *et al*. (2010) Functionally defective germline variants of sialic acid acetylesterase in autoimmunity. *Nature*, **466**, 243–247.

Tibshirani,R. (2005) A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, **7**, 106.

Van Hijum,S. *et al*. (2005) A generally applicable validation scheme for the assessment of factors involved in reproducibility and quality of DNA-microarray data. *BMC Genomics*, **6**, 77.

Vaux,D.L. (2012) Know when your numbers are significant. *Nature*, **492**, 180–181.

Vêncio,R.Z.N. *et al*. (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics*, **5**, 119.

Wei,C. *et al*. (2004) Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, **5**, 87.

Xia,L.C. *et al*. (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol*., **5**, S15.

Yang,X. *et al*. (2006) Similarities of ordered gene lists. *J. Bioinform. Comput. Biol*., **4**, 693–708.

Yong,E. (2012) Replication studies: Bad copy. *Nature*, **485**, 298–300.

Zeggini,E. and Ioannidis,J.P.A. (2009) Meta-analysis in genome-wide association studies. *Pharmacogenomics*, **10**, 191–201.

Zeggini,E. *et al*. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet*., **40**, 638–645.

Zeitlinger,J. *et al*. (2007) RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embyo. *Nat. Genet*., **39**, 1512–1516.

Zhang,M. *et al*. (2009) Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, **25**, 1662–1668.