Quantifying the Reproducibility of Cell-Perturbation Experiments

Jackson Loper, Robert Barton, Meena Subramaniam, Maxime Dhainaut, Jeffrey Regier

2022

Abstract

Experiments adhering to the same protocol can nonetheless lead to different conclusions, for instance, due to batch effects or lab effects. A statistical test applied to measurements from one experiment may yield a vanishingly small p-value, yet applying the same test to measurements from a replicate experiment may yield a large p-value. Recent work has highlighted this lack of reproducibility in cell-perturbation experiments. We introduce the Reproducible Sign Rate (RSR), a new reproducibility metric for settings in which each hypothesis test has two alternatives (e.g., upregulation and downregulation of gene expression). The RSR identifies the proportion of discoveries that are expected to reproduce in a future replicate. We provide conditions under which the RSR can be estimated accurately—even when as few as two experimental replicates are available. We also provide conditions under which high RSR implies a low Type S error rate. We demonstrate the uses of RSR with experiments based on several high-throughput technologies, including L1000, Sci-Plex, and CRISPR.

1 Introduction

Performing the same experiment can lead to different conclusions each time it is replicated [Boos and Stefanski, 2011, Colquhoun, 2017, Lim and Pavlidis, 2021, McShane et al., 2022]. This lack of reproducibility can arise if measurements from a single experimental batch exhibit dependencies due to factors that cannot be controlled; these are known as "batch effects" [Tung et al., 2017]. Individual laboratories that perform experiments may also introduce biases; these are known as "lab effects" [Rosenbloom et al., 2019]. Quantifying reproducibility is difficult. The seminal work of Shao and Chow [2002] described reproducibility as "a person's subjective probability of observing a significant clinical result from a future trial," but the authors were unable to find a mathematical definition that matched this intuitive description.

This article is motivated by the pressing need to quantify reproducibility in the study of gene expression. Each cell has a collection of genes encoded in its DNA, but these genes are not expressed equally; even cells with identical DNA may express their genes differently. This diversity allows organisms to grow and develop [Wang et al., 2022, Shi et al., 2022] and underlies many disease mechanisms [Yoon et al., 2022, Xie et al., 2022]. Understanding this diversity promotes the discovery of new treatments [Subramanian et al., 2017, Srivatsan et al., 2020]. High-throughput cell-perturbation experiments study this diversity by subjecting a population of cells to a variety of stimuli and measuring how each stimuli affects the expression of each gene. A statistical test is then performed for each stimulus and gene, investigating three possible hypotheses: upregulation, downregulation, or no effect (the null hypothesis). A given experiment may investigate billions of these hypothesis tests. Two or three experimental replicates are typically performed [Schmidt et al., 2022, Srivatsan et al., 2020, Subramanian et al., 2017]. To assess reproducibility in this context, it is necessary to distinguish between the upregulation alternative hypothesis and the downregulation alternative hypothesis. Analyzing one experimental replicate may lead to rejection of a null hypothesis in favor of the downregulation alternative, but another replicate might suggest rejecting the same null hypothesis in favor of the upregulation alternative. The null hypothesis is rejected by both tests, so a naïve user of some reproducibility metrics (e.g., [Li et al., 2011]) may consider this to be a successfully reproduced result—but in fact, the *scientific implications* of the two replicates are diametrically opposed, and at least one of the selections must be erroneous. Incorrect selection of an alternative is known as a "sign error" or "Type S error" in two-tailed hypothesis testing [Gelman and Tuerlinckx, 2000, Gelman and Carlin, 2014].

To quantify the reproducibility of experiments with many two-alternative hypothesis tests, we propose the Reproducible Sign Rate (RSR). It is defined in terms of two independent replicates of an experiment: a training replicate and a validation replicate. First, data from the training replicate is used to reject some of the null hypotheses. In tests where the null hypothesis is rejected, data from the training replicate is then used to select an alternative hypothesis (e.g., upregulation or downregulation); we refer to these as the "proposal selections." Data from the validation replicate is also used to select alternatives; we refer to these as the "testing selections." The Reproducible Sign Proportion (RSP) indicates the proportion of rejected hypotheses for which both replicates yield the same selections. The RSR is defined as the expected value of the RSP, conditioning on the training replicate. We formalize a notion of "experimental validity" such that a high RSR implies a low proportion of Type S errors (Section 2.1). This guarantee requires no parametric assumptions about the forms of the distributions of the measurements.

If each experimental replicate includes many weakly dependent subexperiments, the RSR can be inferred

with confidence even when an experiment has only been performed twice (Section 2.2). For example, consider a cell-perturbation experiment in which each subexperiment tests a different drug. Each subexperiment will be influenced by batch effects, but it may be reasonable to assume that the batch effects are only weakly dependent among subexperiments. If the dependency is sufficiently weak, we can obtain nontrivial confidence intervals for the RSR. These theoretical confidence intervals are found to be conservative in semi-synthetic experiments; accurate estimation of the RSR may be easier than our theoretical results suggest (Section 3).

We present a series of case studies of the RSR based on real experimental data (Section 4.1). The RSR demonstrates the superior reproducibility of the Sci-Plex technique for conducting cell-perturbation experiments (Section 4.2). We also find that a preprocessing strategy developed by Qiu et al. [2020] improves the reproducibility of the L1000 technique for conducting cell-perturbation experiments (Section 4.3). A final case study highlights how reproducibility is sensitive to the precise manner in which hypotheses are formulated (Section 4.4). Simple quantifications such as the RSR enable rapid detection of reproducibility failures (Section 5).

2 The Reproducible Sign Rate

Consider n real-valued parameters of interest, $\theta_1, \ldots \theta_n$. For example, in cell-perturbation experiments, a parameter of interest may indicate how a given gene g is regulated when its host cell is exposed to a given perturbation d, with positive values corresponding to upregulation and negative values corresponding to downregulation.

For each parameter θ_i, suppose we have an estimate of its sign, denoted $\hat{Y}_i \in \{-1, 0, 1\}$. We will refer to the vector $\hat{Y}$ as the "proposal selections." Suppose further that we have a quantity ρ_i indicating confidence in the null hypothesis that $\theta_i = 0$; typically each ρ_i will be a p-value (designed by the experimenter to be super-uniform under the null hypothesis), but this is not necessary in the analysis that follows. For any given threshold α, we may reject each null hypotheses where $\rho_i \geq \alpha$.

Now let us perform an independent replicate experiment attempting to measure θ. For each hypothesis $i \in \{1, \ldots, n\}$, suppose we have a function that can be applied to the experimental measurements to produce an estimate (denoted Y_i) of the sign of θ_i. We require that these estimators cannot return the value zero. We will refer to the vector $Y \in \{-1, 1\}^n$ as the "testing selections." Each testing selection is a random variable that is influenced by independent per-hypothesis noise as well as per-replicate noise arising from batch effects. Among the hypotheses rejected at a given threshold α, what proportion of the testing selections agree with

the proposal selections? We define this quantity formally as follows.

Definition 1. The *Reproducible Sign Proportion* of the proposal selections $\hat{Y} \in \{-1, 0, 1\}^n$ with inverse-confidences $\rho \in \mathbb{R}^n$, threshold α, and testing selections $Y \in \{-1, 1\}^n$ is

$$\mathrm{RSP}(Y; \rho, \hat{Y}, \alpha) \triangleq \frac{|\{i : \ \rho_i \leq \alpha \text{ and } Y_i = \hat{Y}_i\}|}{|\{i : \ \rho_i \leq \alpha\}|}.$$

If $\{i : \ \rho_i \leq \alpha\} = \emptyset$, then $\mathrm{RSP}(Y; \rho, \hat{Y}, \alpha) = 1$.

The RSP is a random quantity, observable by examining a replicate experiment. To make statistical statements about reproducibility, we will investigate its expected value.

Definition 2. The *Reproducible Sign Rate* (RSR) is the expected value $\mathrm{RSR}(\rho, \hat{Y}, \alpha) \triangleq \mathbb{E}_Y[\mathrm{RSP}(Y; \rho, \hat{Y}, \alpha)]$.

Note that the RSR does not integrate over possible values of $\hat{Y}$ and ρ. These could be considered random in some contexts, but we take them to be fixed.

The RSR can help assess the Type S error control: under a certain validity assumption, a high value of the RSR implies a high proportion of true discoveries (Section 2.1). However, this statement has no practical use unless the RSR can be estimated reliably. When an experiment contains many weakly dependent subexperiments, nontrivial confidence intervals can be devised (Section 2.2). Plotted as a function of the threshold α and accompanied by confidence intervals, the RSP depicts reproducibility in units that can be used across different experiments (Section 2.3).

2.1 High RSR in a valid replicate implies low Type S error

A high RSR implies that repeated trials of the experimental method consistently yield the same selections of the alternative hypotheses. However, even if the RSR is high, the experimental method may still lead to a high proportion of Type S errors, i.e., cases where $\hat{Y}_i \neq \mathrm{sign}(\theta_i)$.

Definition 3. The *Type S error proportion* is

$$V(\hat{Y}, \rho, \theta, \alpha) = \frac{|\{i : \ \rho_i \leq \alpha \text{ and } \hat{Y}_i \neq \mathrm{sign}(\theta_i)\}|}{|\{i : \ \rho_i \leq \alpha\}|}.$$

If the signs of the expected values of the testing selections are consistent with the signs of θ, then high RSR implies a low Type S error proportion. We formalize this intuition below.

Definition 4. Define a *valid testing selection* for θ_i as a random variable $Y_i \in \{-1, 1\}$ such that

$$\begin{aligned}
\theta_i > 0 &\implies \mathbb{P}(Y_i = -1) \leq \frac{1}{2}, \\
\theta_i < -1 &\implies \mathbb{P}(Y_i = 1) \leq \frac{1}{2}, \text{ and} \\
\theta_i = 0 &\implies \mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = \frac{1}{2}.
\end{aligned}$$

Given any proposal selections $\hat{Y}$, the RSR based on valid testing selections Y yields an upper bound on the Type S error rate.

Theorem 1. *Let $\theta \in \mathbb{R}^n$ denote n parameters of interest. Let $\rho \in \mathbb{R}^n$ denote p-values and let $\hat{Y} \in \mathbb{R}^n$ denote proposal selections. Let $Y \in \{-1, 1\}^n$ denote valid testing selections for θ. Then, the Type S error rate is bounded by $V(\hat{Y}, \rho, \theta, \alpha) \leq 2 - 2 \cdot \mathrm{RSR}(\rho, \hat{Y}, \alpha)$.*

Proof. Let $T = |\{i : \ \rho_i \leq \alpha\}|$. The RSR may be written as

$$\mathrm{RSR} = \frac{1}{T} \sum_{i: \ \rho_i \leq \alpha} \mathbb{P}(Y_i = \hat{Y}_i).$$

The validity of Y_i gives that $\mathbb{P}(Y_i \neq \mathrm{sign}(\theta_i)) \leq 1/2$. Thus $\mathbb{P}(Y_i = \hat{Y}_i) \leq 1/2$ whenever $\hat{Y}_i \neq \mathrm{sign}(\theta_i)$. Summing over i yields

$$\begin{aligned}
\mathrm{RSR} &= \frac{1}{T} \left(\sum_{i: \ \rho_i \leq \alpha, \ \hat{Y}_i = \mathrm{sign}(\theta_i)} \mathbb{P}(Y_i = \hat{Y}_i) \right) + \frac{1}{T} \left(\sum_{i: \ \rho_i \leq \alpha, \ \hat{Y}_i \neq \mathrm{sign}(\theta_i)} \mathbb{P}(Y_i = \hat{Y}_i) \right) \\
&\leq \frac{1}{T} \left(\sum_{i: \ \rho_i \leq \alpha, \ \hat{Y}_i = \mathrm{sign}(\theta_i)} 1 \right) + \frac{1}{T} \left(\sum_{i: \ \rho_i \leq \alpha, \ \hat{Y}_i \neq \mathrm{sign}(\theta_i)} \frac{1}{2} \right) \\
&= (1 - V) + \frac{1}{2} V = 1 - \frac{1}{2} V.
\end{aligned}$$

□

Among hypotheses rejected at a fixed threshold α, Theorem 1 shows that a high proportion of proposal selections correctly estimate the signs of the entries of θ whenever the RSR is high and the experiments are valid. For example, if a particular set of proposal selections reproduce in valid replicates with RSR exceeding 90%, the Type S error rate is at most 20%.

Experiments with low RSR cannot be used to build scientific consensus; to build scientific consensus it is necessary that multiple experimental replicates yield the same conclusions about the same parameters.

However, experiments with low RSR can still be useful. Consider an experiment in which each individual replicate yields accurate conclusions about different parameters. The RSR is low because information from one experiment (which is only accurate for a small handful of parameters) is uninformative about the results of another experiment (which is only accurate for a different small handful of parameters). This experimental design may be useful for hypothesis generation; we discuss this idea further in Section 5.

If the RSR is high but the experiment is invalid, we cannot bound Type S error without additional assumptions. Consider an experimental procedure in which perturbed cells are measured at one temperature, control cells are measured at a different temperature, and upregulation and downregulation are determined by calculating the differences between the two groups. The procedure may have a high RSR, yielding the same results in each replicate—but the results may have no connection to the scientific question that the experiment purports to answer. A high RSR value cannot be taken as evidence of experimental validity.

2.2 Estimation

The RSR cannot be exactly determined from a finite number of replicates. However, in many cases we can use additional independence structures present in the data to estimate the RSR with high confidence—even when an experiment has been performed only twice. Let $\{P_1, P_2, \dots P_m\}$ be a partition of the parameters into disjoint groups such that $\{1, 2, \cdots n\} = \cup_i P_i$ and $\{Y_i\}_{i \in P_1}, \{Y_i\}_{i \in P_2} \cdots \{Y_i\}_{i \in P_m}$ are weakly dependent. We will refer to each P_i as a "subexperiment." For example, in an experiment testing the effects of many drugs on many genes, we might take each distinct drug to correspond to its own subexperiment. We will consider two forms of weak dependence: one based on low correlations and one suitable for bounding uncertainty using Hoeffding's lemma.

The following two theorems show conditions where these subexperiments can be used to obtain useful confidence intervals. For each set P_i, let a_i denote the total number of rejected hypotheses and let X_i denote the number of rejected hypotheses in which the proposal selections agree with the testing selections. In these terms, the RSP may be expressed as $\sum_i X_i / \sum_i a_i$. To show that the RSR can be reliably inferred, it thus suffices to show that the RSP does not deviate too much from its expected value. We investigate this question in several different settings.

We first assume that $X_1, \dots, X_m$ are independent. In this case, we will see that

$$\mathbb{P}(|\mathrm{RSR}(\rho, \hat{Y}, \alpha) - \mathrm{RSP}(Y; \rho, \hat{Y}, \alpha) > t|) \leq 2 \exp(-2t\xi) \tag{1}$$

where $\xi = (\sum_i a_i)^2 / \sum_i a_i^2$. This bound is proven as a special case of Theorem 3, below. We will use this bound in the case studies presented in Section 4.

We next consider the possibility that $X_1, \ldots, X_m$ are weakly dependent. We will consider two different forms of weak dependency. The different forms leads to different bounds. Each bound will be expressed in terms of a quantity that measures the magnitude of the weak dependency.

Unfortunately, at least in the case studies presented in Section 4, magnitudes of weak dependencies cannot be reliably inferred. However, we can study the influence of the weak dependency on the bounds to obtain theoretical insight.

Correlation offers one natural way to characterize dependency. If the correlations between subexperiments are bounded, Chebyschev's inequality can be used to bound the difference between RSR and RSP.

Theorem 2. *Let $X_1, \ldots, X_n$ be random variables with $X_i \in [0, a_i]$ for some $a_i \geq 0$. Let $\xi = \left(\sum a_i\right)^2 / \sum a_i^2$, $r_{ij} = \mathrm{Cov}\left(X_i, X_j\right) / \sqrt{\mathrm{Var}\left(X_i\right) \mathrm{Var}\left(X_j\right)}$, $r_{\max} = \sup_{i \neq j} r_{ij}$, and $\mu_i = \mathbb{E}\left[X_i\right]$. Then,*

$$\mathbb{P}\left(\left|\frac{\sum_i \left(X_i - \mu_i\right)}{\sum_i a_i}\right| \geq t\right) \leq \frac{\left(1 - r_{\max}\right)}{4t^2\xi} + \frac{r_{\max}}{4t^2}.$$

Proof. Let $R = \sum_i a_i^2$ and $\tilde{R} = \sum_i \alpha_i$. Apply Chebyschev's inequality, yielding

$$\mathbb{P}\left(\sum_i \left(X_i - \mu_i\right) \geq t \sum_i a_i\right) \leq \frac{\mathbb{E}\left[\left(\sum_i \left(X_i - \mu_i\right)\right)^2\right]}{t^2 \tilde{R}^2} = \frac{1}{t^2 \tilde{R}^2} \sum_{ij} r_{ij} \sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}.$$

Since X_i is bounded, $\mathrm{Var}(X_i) \leq a_i^2/4$. Thus,

$$\begin{aligned}
\mathbb{P}\left(\sum_i \left(X_i - \mu_i\right) \geq t \sum_i a_i\right) &\leq \frac{\sum_{ij} a_i a_j r_{ij}}{4t^2 \tilde{R}^2} = \frac{1}{4t^2 \tilde{R}^2} \left(\sum_i a_i^2 + r^{\max} \sum_{i \neq j} a_i a_j\right) \\
&= \frac{1}{4t^2 \tilde{R}^2} \left(R + r^{\max}\left(\tilde{R}^2 - R\right)\right) = \frac{r_{\max}}{4t^2} + \frac{\left(1 - r_{\max}\right) R}{4t^2 \tilde{R}^2}.
\end{aligned}$$

□

The bound in Theorem 2 is weaker than Inequality 1 in several respects. First, the bound decays polynomially in t (whereas Inequality 1 features exponential decay). Second, the term $r_{\max}/4t^2$ does not depend on the number of rejections in each subexperiment. In particular, this theorem would not allow us to show that RSR can be estimated consistently in the limit of infinite subexperiments.

The literature concerned with dependent Hoeffding bounds [Tanoue, 2021] offers a different way to mea-

sure dependency. Given a probability measure τ on $\mathbb{R}^m$, we first construct a new measure comprising the product of the marginals of the old measure.

Definition 5. Let τ denote a probability measure on on $\mathbb{R}^m$. Let $(X_1, X_2 \dots X_m) \sim \tau$. The **product marginal** of τ is defined as the measure $\text{ProdMarg}(\tau)(X_1 \in A_1, \dots, X_m \in A_m) \triangleq \prod_i \tau(X_i \in A_i)$ for any choice of measurable sets $A_1, A_2, \cdots A_m \subset \mathbb{R}$.

We can quantify the dependence among $\{X_i\}_{i=1}^m$ as a divergence between τ and $\text{ProdMarg}(\tau)$. Existing literature uses quantifications of this form to bound the variability of sums of random variables (cf. Tanoue [2021] and the references therein). However, this literature does not contain bounds that are practical in our setting; it focuses on cases where τ carries a sparse Markov dependency structure, which is not realistic for cell-perturbation experiments. The following theorem makes no such assumptions.

Theorem 3. *Let $X_1, \dots, X_m$ be random variables with $X_i \in [0, a_i]$ for some $a_i \geq 0$. Let $\xi = \left(\sum a_i\right)^2 / \sum a_i^2$, and $\mu_i = \mathbb{E}[X_i]$. Let τ denote the law of X and $\nu = \text{ProdMarg}(\tau)$. Assume τ is absolutely continuous with respect to ν. Let $D_\chi = \int \left(\frac{d\tau}{d\nu} - 1\right)^2 d\nu$. Then,*

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\sum_i (X_i - \mu_i)}{\sum_i a_i}\right| \geq t\right) &\leq 2 \min_s \left(e^{-ts\sum a_i}\left(e^{s^2(\sum a_i^2)/8} + \sqrt{D_\chi} e^{s^2(\sum a_i^2)/4}\right)\right) \\ &\leq 2 \min\left\{e^{-2t^2\xi} + \sqrt{D_\chi}, e^{-\frac{3}{2}t^2\xi} + e^{-t^2\xi}\sqrt{D_\chi}\right\}.\end{aligned}$$

Proof. Let $R = \sum_i a_i^2$ and $\tilde{R} = \sum_i \alpha_i$. By the Chernoff bound,

$$\mathbb{P}\left(\sum_i (X_i - \mu_i) \geq t \sum_i a_i\right) \leq e^{-ts\tilde{R}} \mathbb{E}\left[e^{s\sum_i (X_i - \mu_i)}\right]. \tag{2}$$

Let $Y \sim \nu$. Applying the triangle inequality and the Cauchy-Schwarz inequality on the Hilbert space with measure ν, we obtain

$$\mathbb{E}\left[e^{s\sum_i (X_i - \mu_i)}\right] \leq \prod_i \mathbb{E}\left[e^{s(Y_i - \mu_i)}\right] + \prod_i \sqrt{D_\chi \mathbb{E}\left[e^{2s(Y_i - \mu_i)}\right]}.$$

Applying Hoeffding's lemma to each expectation and inserting the result into Equation 2 yields

$$\mathbb{P}\left(\sum_i (X_i - \mu_i) \geq t \sum_i a_i\right) \leq e^{-ts\tilde{R}}\left(e^{s^2 R/8} + \sqrt{D_\chi} e^{s^2 R/4}\right) \qquad \forall s \geq 0. \tag{3}$$

Any value of s will give a valid bound; the tightest bound may be obtained by optimizing Equation 3 over

s. This optimization problem does not appear to have an analytical solution, and so we obtain our final bound by choosing two values for s and taking whichever value yields the better bound. To choose these two values, observe that Equation 3 comprises a sum of two terms, each a convex function of s. Taking derivatives shows that the first summand is minimized when $s = 4t\tilde{R}/R$. Similarly, the second summand is minimized when $s = 2t\tilde{R}/R$. Substituting these two choices into Equation 3 gives the following bound:

$$\mathbb{P}\left(\sum_i (X_i - \mu_i) \geq t \sum_i a_i\right) \leq \min\{e^{-2t^2\xi} + \sqrt{D_\chi}, e^{-\frac{3}{2}t^2\xi} + e^{-t^2\xi}\sqrt{D_\chi}\}.$$

Finally, to obtain the desired two-sided confidence intervals, apply the same argument to $-\sum X_i$ and take a union bound. □

When D_χ is small, the bounds in Theorem 3 are superior to the bounds in Theorem 2; they offer exponential decay in both t and ξ. However, these bounds become irrelevant if the χ^2 divergence is infinite.

The discussion above is summarized in the following corollary.

Corollary 1. *Let* $X_i = |\{j \in P_i : \ \rho_j \leq \alpha \text{ and } Y_j = \hat{Y}_j\}|$. *Let*

$$\xi = \left(\sum_i |\{j \in P_i : \ \rho_j \leq \alpha\}|^2\right) / \left(\sum_i |\{j \in P_i : \ \rho_j \leq \alpha\}|\right)^2.$$

Let $r^{\max}$ *denote the maximum correlation among* $X_1, \ldots, X_m$. *Let* D_χ *denote the* χ^2 *divergence from the law of* X *to its product marginal. The chance that* $\mathrm{RSP}(Y; \rho, \hat{Y}, \alpha)$ *deviates from* $\mathrm{RSR}(\rho, \hat{Y}, alpha)$ *by at least* t *is bounded by each of the following expressions:*

$$\text{Bound I}: \qquad \frac{(1 - r_{\max})}{4t^2\xi} + \frac{r_{\max}}{4t^2} \tag{4}$$

$$\text{Bound II}: \qquad 2e^{-2t^2\xi} + 2\sqrt{D_\chi} \tag{5}$$

$$\text{Bound III}: \qquad 2e^{-\frac{3}{2}t^2\xi} + 2e^{-t^2\xi}\sqrt{D_\chi} \tag{6}$$

Proof. Let $a_i = |\{j \in P_i : \ \rho_j \leq \alpha|$. Use Theorem 2 to obtain Bound I and Theorem 3 to obtain Bound II and Bound III. □

Each of the three bounds from Corollary 1 is tighter than the other two in some context. We present several examples in Figure 1. In each example we assume exactly one hypothesis has been rejected in each subexperiment (i.e., $a_1 = a_2 = \cdots a_m = 1$). We model dependency among variables in $[0, 1]^m$ using a

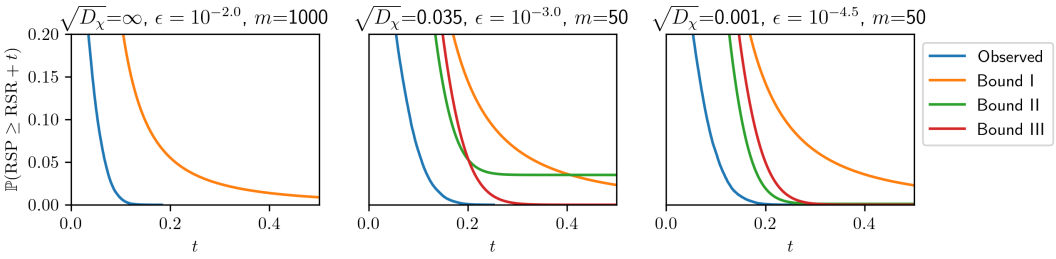

Figure 1: *Each bound of Corollary 1 is useful in some context.* **Left:** With many samples, even weak global correlations lead to infinite χ^2 divergences; the Hoeffding-based bounds do not apply in this case, but Chebyschev-based bounds still yield useful results. **Middle:** With a moderate level of dependency, Bound III is able to yield small probabilities for larger error levels. **Right:** With low dependency, Bound II approaches a Hoeffding inequality and achieves the best performance.

Gaussian copula with covariance $\Sigma_{ii} = 1\ \forall i$ and $\Sigma_{ij} = \epsilon\ \forall i \neq j$. Note that D_χ is invariant to per-coordinate bijections and can be computed exactly for any distribution with a Gaussian copula. We vary both ϵ and n, and explore the bounds as a function of the error level t. We also estimate the true probabilities of interest using $50,000$ samples. We find the Chebyschev bounds are looser than the Hoeffding bounds when the dependency is small. None of the inequalities are tight in the simulations presented, suggesting the RSR may be substantially easier to estimate than the theoretical bounds suggest.

So far we have considered estimation of a single RSR value. Similar ideas can be used to estimate the difference in reproducibility between two experimental techniques or analytical methods. In this case, one needs confidence intervals for the difference between two RSR values. Such a confidence interval can always be found by forming confidence intervals for each RSR value separately and applying a union bound. In many cases, one can obtain smaller confidence intervals using the structure of the problem. Two such cases are presented in Appendix B.

2.3 Visualization

We propose a type of visualization with units that are interpretable by a broad audience. The RSR depends on α, so a complete visualization of the RSR requires sweeping over many values of this threshold. To visualize how the different values of α yield different values of reproducibility, we plot $\text{RSP}(Y; \rho, \hat{Y}, \alpha)$ against the number of rejections made with a threshold of α. There are several ways we could render such a plot, each with different advantages. One could plot the RSR against the total number of rejected hypotheses. However, in some cases we find a single subexperiment may result in vastly more discoveries than others. To

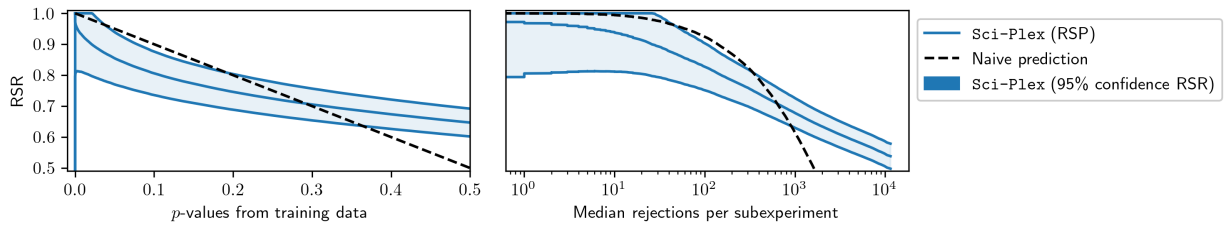

Figure 2: *Computing the maximum number of discoveries that can be made while still controlling the RSR.* We estimate the RSR using real data from replicate experiments conducted using the Sci-Plex technique. **Left:** The RSR as a function of thresholds α on the nominal p-values obtained from the data. The blue shaded area indicates a (non-simultaneous) 95% confidence band from Theorem 3. The dotted line indicates reproducibility predictions under a naive misinterpretation of p-values, namely that a discovery which is significant at the p level can be expected to reproduce with probability $1-p$. **Right:** Same as the left panel, but with the horizontal axis rescaled to match the median number of proposed discoveries for each p-value threshold.

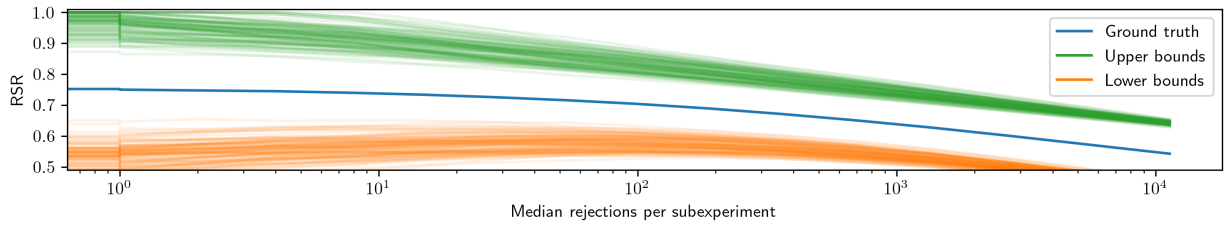

Figure 3: *The Hoeffding bounds from Theorem 3 may be loose in some cases.* We construct 100 semi-synthetic replicates of the Sci-Plex K562 data. Using each replicate we compute a 95% (non-simultaneous) confidence band for the RSR; in each case the true RSR replicate falls inside the band. This suggests the bands are conservative.

display this kind of information, one could plot the RSR against the proportion of subexperiments in which at least one hypothesis is rejected. Figure 2 shows a third approach: plotting the RSR against the median number of hypotheses rejected per subexperiment.

3 Semi-synthetic experiments

We empirically assessed the looseness of the Hoeffding bounds from Theorem 3 using semi-synthetic simulation. To generate a semi-synthetic dataset, we first obtain z-scores for the Sci-Plex data restricted to cells from the K562 cell line treated with various drugs at a dosage of one micromolar. After preprocessing, this leads to $12,207 \times 188$ z-scores for the effects of 188 different drugs on 12,207 genes. Each z-score measures

the extent to which a particular gene is affected (i.e., up- or downregulated) by a particular treatment. For example, z-score will be positive if the expression of a gene observed in treated cells is higher than that gene's expression in untreated cells. The Sci-Plex experiment and the procedure for obtaining z-scores is described in greater depth in Appendix A.

We perturb the z-scores described above to create independent synthetic replicates, $\tilde{z}$. The synthetic replicates are drawn as follows. First, for every drug, we draw a value $U_d \sim \text{Uniform}[0, 1]$. Then, for each synthetic replicate r, we draw the perturbed values for gene g and drug d according to the following process. The variable $S_{r,d} \sim \text{Uniform}[.5, 1.5]$ represents random changes in the overall variability of genes under the influence of each drug. The variable $W_{r,d} \sim \mathcal{N}(0, 1)$ represents batch effects, which may effect all genes for a given drug. The variable $\tilde{W}_{r,d,g} \sim \mathcal{N}(0, 1)$ models individual variability of each gene for each trial, reflecting the presence of independent noise in each measurement. The variable $F_{r,d} \sim \text{Bernoulli}(U_d)$ models the possibility that some drugs have been rendered impotent by exogeneous factors (in which case the observations will arise solely from the noise process). Finally, the synthetic z-scores are given by

$$\tilde{z}_{r,d,g} = F_d z_{d,g} + \tilde{W}_{r,d,g} S_{r,d} + W_{r,d}.$$

The true RSR for this model is estimated by averaging over many synthetic replicates. We also apply Theorem 3 to obtain a confidence band for the RSR from each individual replicate.

Figure 3 compares these bands with the RSR, showing that the coverage probability of the confidence intervals from Theorem 3 is much higher than the nominal level in this setting. For 100 synthetic replicates, the true RSR lay inside the 95% Hoeffding bounds in every case. This suggests that the Hoeffding bounds from Theorem 3 may be loose in typical situations. This looseness may arise because many of the sources of noise are independent across all 12,207 genes and 188 drugs in our simulations. Indeed, the RSP in these simulations is obtained by averaging over more than two million random variables that are nearly independent, whereas Theorem 3 merely assumes independence among the 188 drugs. If the measurements for different genes were known to be fully independent, the standard deviation would be $\sqrt{12,207} \approx 100$ times smaller. In such cases, accurate estimation of the RSR is easier than our current theoretical results show.

4 Case studies

We consider data from four cell-perturbation experiments (Section 4.1) and apply the RSR in three different ways. The RSR can be used to compare different experiments, even though they were performed using different experimental techniques (Section 4.2). It can also be used to compare different computational methods for analyzing the measurements of a single experiment (Section 4.3). Finally, we can investigate how the RSR depends on the scientific question being asked; in some settings researchers may be interested in absolute changes in gene expression, and in other cases researchers may be interested in how one gene's expression changes relative to another (Section 4.4).

4.1 The Data

We analyze data from four experiments. Each experiment perturbs cells in some way and then measures how the perturbations affect the cells' gene expression.

`L1000` perturbs cells by exposing them to small-molecule drugs. This experiment involves many subexperiments across seven cell types and hundreds of drug perturbations, testing between four and nine dosage levels for each perturbation [Subramanian et al., 2017]. There are three replicates of this experiment.

`Sci-Plex` also perturbs using small chemicals, measuring more genes but including only three cell types and fewer kinds of perturbations [Srivatsan et al., 2020]. There are two replicates of this experiment.

`Shifrut2018` perturbs cells by deleting certain genes (the "knockout genes") from their DNA using CRISPRn technology [Shifrut et al., 2018]. It studies T cells from two donors (each corresponding to an experimental replicate) and tests 20 different knockout genes.

`Schmidt2022` perturbs cells by inducing additional transcription of certain genes using CRISPRa technology [Schmidt et al., 2022]. It also includes two donor replicates.

In all four experiments, we can use one replicate to test null hypotheses and estimate the RSR using another replicate. Appendix A presents additional details of these datasets and our choices for hypothesis testing methods.

4.2 Comparing experimental techniques

RSR can be used to compare `L1000` and `Sci-Plex`. Both experimental techniques can be assessed in terms of the number of discoveries made and the proportion of those discoveries that are expected to reproduce. However, to properly interpret the RSR, the researcher must be aware that the two experiments measure

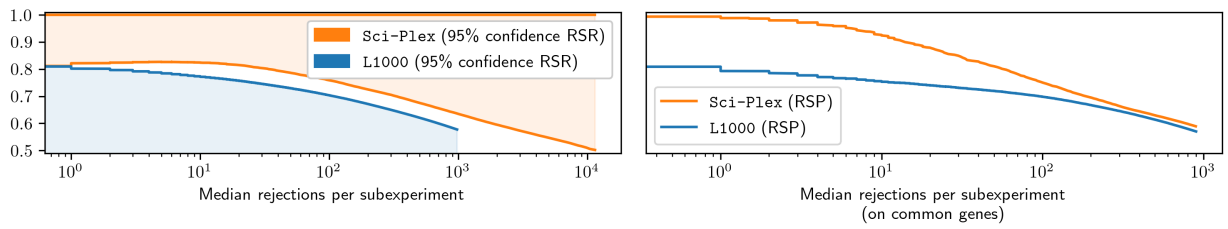

Figure 4: *RSR facilitates comparisons of experimental techniques.* We examine two different methods for measuring the effects of chemical perturbations on gene expression: `L1000` and `Sci-Plex`. For each method, we plot RSR estimates as a function of the median number of null hypotheses rejected per subexperiment. **Left:** All hypotheses with the top two dosage levels are considered, yielding enough subexperiments to obtain useful confidence intervals. Note that Sci-Plex tests more hypotheses per subexperiment. **Right:** The set of hypotheses to consider is restricted so that each subexperiment investigates the same set of genes.

a different number of hypotheses per subexperiment. `L1000` considers more drugs (1726 compared to 188 for `Sci-Plex`), more cell types (7 compared to 3), and fewer genes (978 compared to 12,207). Despite these disparities, there are several ways to use the RSR to compare these experiments.

We first consider all hypotheses concerning the the top two dosage levels in each experiment. One-sided 95% confidence intervals of the RSRs for these sets of hypotheses are shown on the left of Figure 4. Assuming that the newer technology would be better, we computed lower-bound confidence intervals for `Sci-Plex` and upper-bound confidence intervals for `L1000`. Theorem 4 from Appendix B allows us to reject the hypothesis that both methods have equal RSR when targeting a median of one rejection per subexperiment, with p-value .018. However, if we consider all hypotheses, the total number of hypotheses tested per subexperiment is quite different between the two experimental techniques: `Sci-Plex` has 12,207 and `L1000` has only 978.

Another option is to only consider hypotheses related to MCF7 cell types (which appear in both experiments), the highest dosage levels, and the 901 genes common to both experiments. With this restriction, both experiments have the same number of hypotheses per subexperiment. The right side of Figure 4 shows RSR over this restricted set. We find `Sci-Plex` reaches an RSP of 99%, whereas `L1000` reaches an RSP of 81%. This represents a twenty-fold increase in the Type S error bounds from Theorem 1; this is a major increase of importance to experimenters. However, useful confidence intervals cannot be obtained from Corollary 1 in this case; there are too few subexperiments.

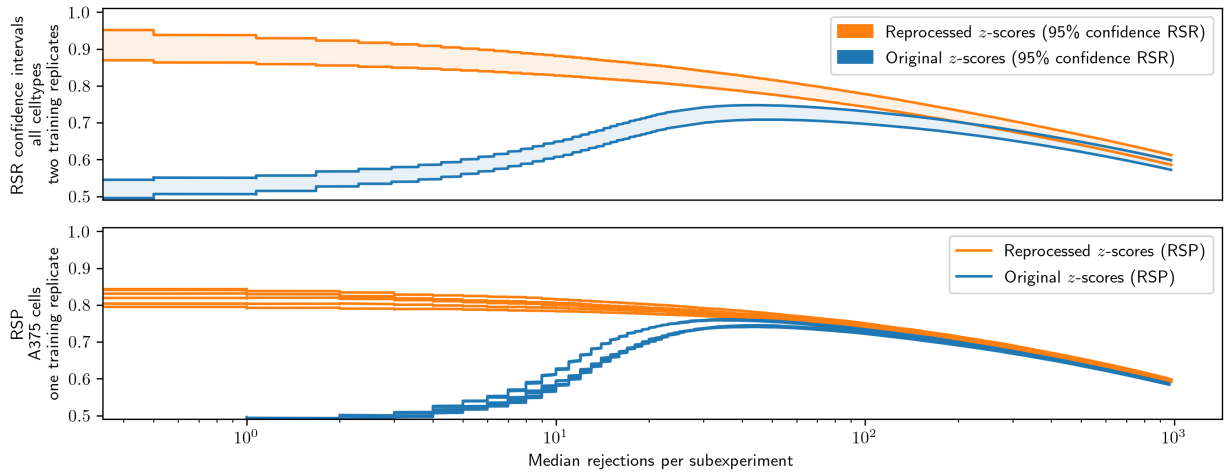

Figure 5: *RSR demonstrates that the Bayesian treatment by Qiu et al. [2020] yields greater reproducibility in the L1000 data.* The fluorescent measurements from the replicates of the L1000 dataset were reanalyzed in 2020 using a Bayesian approach. **Top:** RSR estimates across all cell types at the highest dosage levels, using two replicates for training (averaging across replicates to obtain less noisy readings) and applying Theorem 3 to obtain confidence bands. **Bottom:** RSR estimates for one cell type at the highest dosage levels, using one replicate for training and one replicate for evaluating the RSP. The six shufflings of replicates are used as an empirical measure of variability.

4.3 Reanalysis of the L1000 dataset

The L1000 technique for measuring gene expression using the Luminex FlexMap 3D technique, yielding a collection of real-valued fluorescence intensities for each collection of cells. The publication that introduced L1000 used a heuristic algorithm to estimate gene expression from these values. Qiu et al. [2020] reanalyzed these measurements using a principled Bayesian method and argued that this reanalysis yielded higher accuracy. `L1000` includes three replicates, so we can use the RSR to probe whether the Bayesian method of Qiu et al. [2020] really does improve the reproducibility. The finite-sample guarantees of Corollary 1 allow us to make a rigorous statistical claim about this question. We combine two replicates to form proposal selections, use the final replicate for validation, and plot the RSR with 95% confidence intervals in the top panel of Figure 5. At modest thresholds, the confidence intervals do not overlap. Indeed, even confidence intervals with coverage probability ($1\text{-}10^{-50}$) do not overlap; we reject the hypothesis that both analysis methods have equal RSR. In place of confidence intervals, the bottom plot uses a heuristic approach to assess our uncertainty. The RSR depends on which replicates are used to make the proposal selections ("training replicates") and which replicates are used for validation. For example, we could compute the RSR

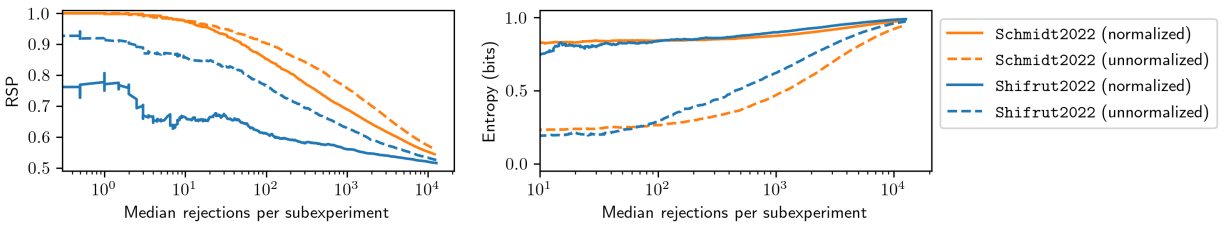

Figure 6: *Reproducibility for different scientific questions.* **Left:** We measure RSR for two different kinds of scientific questions: "normalized" (querying changes in relative gene abundance) and "unnormalized" (querying changes in absolute gene abundance). **Right:** The unnormalized questions lead to imbalances in which alternatives are selected. This imbalance can be measured using entropy.

using the third replicate to make proposal selections and the second replicate for validation. The bottom plot of Figure 5 shows the RSR for all six possible combinations. In all six cases, the same picture emerges: the reanalysis leads to higher reproducibility.

With such low RSR, it seems unlikely that the original analysis would help build scientific consensus about the effects of drugs on genes. This finding echoes those of another recent work [Lim and Pavlidis, 2021].

The RSR offers a clue about what may have gone wrong in the original analysis. The RSR for the original analysis shows a distinctive non-monotonic relationship between p-value thresholds and reproducibility. In particular, the most stringent p-value threshold (which yields the fewest rejections) leads to the worst reproducibility rates. This phenomenon was not seen in other datasets. One explanation would be that a small collection of subexperiments failed in some respect, leading to extremely small p-values but unreliable estimates of the parameters of interest; such a collection would yield the low RSR observed at the most stringent p-value thresholds. This notion is consistent with the findings of Qiu et al. [2020]: a key part of their process is to identify subexperiments with anomolous properties and to discard them.

4.4 Normalization

The Reproducible Sign Rate depends not only on the experimental technique used to obtain measurements but also on the nature of the hypotheses that are tested using those experimental measurements. We consider a simple example, investigating two kinds of hypotheses. In the first approach, which we refer to as "normalized," we compute the expression of each gene in each cell as a proportion of the total number of RNA molecules observed in that cell. For each perturbation and gene, we test the hypothesis that a

given perturbation has no effect on this relative expression. In the second approach ("unnormalized"), we instead use the raw gene measurements. For each perturbation and gene, we test the hypothesis that a given perturbation has no effect on the unnormalized gene expression. We apply each approach to two different experiments: `Shifrut2018` and `Schmidt2022`. Both experiments perturb cells by directly affecting the cells' abilities to transcribe certain genes. Figure 6 shows the RSR for each approach in both experiments.

In `Shifrut2018`, the unnormalized approach yields higher RSR. One possible explanation is that if a knockout causes a *global increase* in expression for all genes, the unnormalized approach would identify that many genes are affected. Each of those genes constitutes a discovery. Normalized measurements are invariant to total gene expression levels, so the normalized approach cannot be used to make discoveries based on global changes of this kind. Instead, the normalized approach leads to conclusions about relative expression. It seems that `Shifrut2018` contains fewer reproducible discoveries about relative expression: even at the most stringent thresholds, the normalized approach exhibits an RSP of 72%. The CRISPRa techniques used in `Schmidt2022` can make both kinds of discoveries with high reproducibility.

To gain greater understanding of these global effects, we developed a method to measure them. Recall that for each rejected hypothesis, we must choose one of two alternatives (upregulated or downregulated). For each perturbation i, we compute the number of hypotheses in which the downregulation alternative was selected and downregulation was also observed in the validation replicate; denote this by X_i^-. Similarly, let X_i^+ denote the corresponding number of cases in which upregulation was selected and observed. Let $p_i = X_i^+/(X_i^- + X_i^+)$ denote the proportion of genes for which upregulation was selected. Let $H_i = p_i \log_2 p_i + (1-p_i)\log_2(1-p_i))$ denote the corresponding entropy (in units of bits). A measurement of conditional entropy is then $\sum_i (X_i^+ + X_i^-)H_i / \sum_i (X_i^+ + X_i^-)$. When both alternatives are selected equally often, this yields an entropy of one (in units of bits). If only one of the alternatives is selected for all genes in each perturbation, this yields an entropy of zero.

The right side of Figure 6 shows these entropies as a function of the median number of rejections. For both experiments, the alternatives selected by the unnormalized approach have lower entropy. This suggests that the unnormalized approach yields discoveries about global expression (e.g., total gene expression is higher), whereas the normalized approach yields discoveries about relative expression (e.g., the relative prevalence of one gene is increased but the relative prevalence of another gene is decreased).

5 Discussion

This work proposes the Reproducible Sign Rate (RSR), a new aggregate measure of reproducibility that is applicable to experiments performing many two-alternative hypothesis tests. The RSR measures agreement between the alternatives selected in two experimental replicates after discarding hypotheses in which the first replicate fails to reject the null. The RSR has uses beyond assessing reproducibility: Theorem 1 shows that high values of RSR in valid experiments imply a high true discovery proportion. Estimating the RSR at many different thresholds yields a visualization that facilitates comparisons between different experimental techniques and computational method.

The performance of every experimental technique has many aspects (including reproducibility, financial/environmental cost, scientific importance of hypotheses tested, number of hypotheses tested, and experimental validity) and that no procedure can be fully evaluated without taking all of them into account. The RSR offers a way to measure one of those aspects: reproducibility.

The main technical contributions of this work lie in demonstrating conditions under which the RSR can be estimated accurately. The conditions for these confidence intervals could be used to guarantee accuracy for quantities beyond the RSR. For example, there is a line of work attempting to predict gene expression in novel conditions [Umarov et al., 2021, Lotfollahi et al., 2019, 2021]. When making such predictions, it is common practice to evaluate performance by reporting the correlation coefficient between the log-fold-change in expression predicted by a model and the log-fold-change observed on held-out data. This practice is also adopted by Lim and Pavlidis [2021]. To have confidence in this kind of method, it would be necessary to understand the variability in such correlation coefficients. Fortunately, correlation coefficients are bounded, and Theorems 2 and 3 can be applied for this purpose.

The ideas in this work can also be extended to cases for each cell typeproposal selections and the testing selections are produced using different experimental technologies. Consider an experiment in which each replicate yields a small handful of valid discoveries about a small random subset of parameters. By themselves, such experiments cannot build scientific consensus because replicates will not lead to the same conclusions about the same parameters. However, they could be useful for hypothesis generation, leading to targeted follow-up experiments. The RSR can be used to quantify reproducibility for these follow-up experiments, even though the proposal selections (from a large-scale experiment) and the testing selections (from a targeted experiment) are obtained using different experimental procedures.

The RSR and its estimators offer a simple way for experimenters to gauge the influence of batch effects on their conclusions. Simple methods for measuring reproducibility may lead to swifter recognition of potential

concerns. For example, although the original L1000 data was published in 2017, it took four years for concerns about reproducibility to rise to the level of a publication [Lim and Pavlidis, 2021]. Developing and implementing experiment-specific strategies to evaluate reproducibility is difficult. Simple best practices for assessing reproducibility that can be applied across experimental modalities would alleviate such burdens.

Acknowledgements. We thank Prayag Chatha, Roman Kouznetsov, Declan McNamara, Yash Patel, Cheng Wang, and Mallory Wang for their thoughtful comments.

Software. A python package for computing Reproducible Sign Proportions and confidence intervals can be found at `https://github.com/prob-ml/reproducible-sign-rates`.

Funding. This research was funded by Immunai Inc.

References

Dennis D Boos and Leonard A Stefanski. p-value precision and reproducibility. *The American Statistician*, 65(4):213–221, 2011.

David Colquhoun. The reproducibility of research and the misinterpretation of p-values. *Royal society open science*, 4(12):171085, 2017.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

Andrew Gelman and Francis Tuerlinckx. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational statistics*, 15(3):373–390, 2000.

Qunhua Li, James B Brown, Haiyan Huang, and Peter J Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

Nathaniel Lim and Paul Pavlidis. Evaluation of connectivity map shows limited reproducibility in drug repositioning. *Scientific Reports*, 11(1):1–14, 2021.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021.

Blakeley B. McShane, Ulf Böckenholt, and Karsten T. Hansen. Variation and covariation in large-scale replication projects: An evaluation of replicability. *Journal of the American Statistical Association*, 2022.

Yue Qiu, Tianhuan Lu, Hansaim Lim, and Lei Xie. A Bayesian approach to accurate and robust signature detection on LINCS L1000 data. *Bioinformatics*, 36(9):2787–2795, 2020.

Daniel IS Rosenbloom, Peter Bacchetti, Mars Stone, Xutao Deng, Ronald J Bosch, Douglas D Richman, Janet D Siliciano, John W Mellors, Steven G Deeks, Roger G Ptak, et al. Assessing intra-lab precision and inter-lab repeatability of outgrowth assays of hiv-1 latent reservoir size. *PLoS computational biology*, 15(4):e1006849, 2019.

Ralf Schmidt, Zachary Steinhart, Madeline Layeghi, Jacob W Freimer, Raymund Bueno, Vinh Q Nguyen, Franziska Blaeschke, Chun Jimmie Ye, and Alexander Marson. CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science*, 375(6580), 2022.

Jun Shao and Shein-Chung Chow. Reproducibility probability in clinical trials. *Statistics in Medicine*, 21 (12):1727–1742, 2002.

Bowen Shi, Yanyuan Wu, Haojie Chen, Jie Ding, and Jun Qi. Understanding of mouse and human bladder at single-cell resolution: integrated analysis of trajectory and cell-cell interactive networks based on multiple scrna-seq datasets. *Cell Proliferation*, 55(1):e13170, 2022.

Eric Shifrut, Julia Carnevale, Victoria Tobin, Theodore L Roth, Jonathan M Woo, Christina T Bui, P Jonathan Li, Morgan E Diolaiti, Alan Ashworth, and Alexander Marson. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971, 2018.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

Yuta Tanoue. Improved Hoeffding inequality for dependent bounded or sub-Gaussian random variables. *Probability, Uncertainty and Quantitative Risk*, 6(1):53, 2021.

Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(1):1–15, 2017.

Ramzan Umarov, Yu Li, and Erik Arner. DeepCellState: An autoencoder-based framework for predicting cell type specific transcriptional states induced by drug treatment. *PLoS Computational Biology*, 17(10): e1009465, 2021.

Lina Wang, Yongjun Piao, Dongyue Zhang, Wenli Feng, Chenchen Wang, Xiaoxi Cui, Qian Ren, Xiaofan Zhu, and Guoguang Zheng. Fbxw11 impairs the repopulation capacity of hematopoietic stem/progenitor cells. *Stem cell research & therapy*, 13(1):1–14, 2022.

Shishuai Xie, Wanxiang Niu, Feng Xu, Yuping Wang, Shanshan Hu, and Chaoshi Niu. Differential expression and significance of miRNAs in plasma extracellular vesicles of patients with Parkinson's disease. *International Journal of Neuroscience*, 132(7):673–688, 2022.

Sojung Yoon, Sung Eun Kim, Younhee Ko, Gwang Hun Jeong, Keum Hwa Lee, Jinhee Lee, Marco Solmi, Louis Jacob, Lee Smith, Andrew Stickley, et al. Differential expression of microRNAs in Alzheimer's disease: A systematic review and meta-analysis. *Molecular Psychiatry*, 27(5):2405–2413, 2022.

A Dataset acquisition and pre-processing

In this work we use four datasets, which we refer to as `L1000`, `Sci-Plex`, `Shifrut2018`, and `Schmidt2022`. For each dataset, we perform hypothesis tests and compute the RSR. Details about this process for each dataset are presented below.

A.1 L1000

The L1000 dataset was originally presented by Subramanian et al. [2017]. It can be viewed at several levels of pre-processing; we elected to view after it had been preprocessed to what is referred to as "level 4." The data in this level comprises a z-score for each replicate for each perturbation for each dosage for each cell type for each gene. The data can be downloaded from GEO accession GSE70138 (cf. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi`). Because this data is in the form of z-scores, our null hypotheses assume that each z-score is drawn from $\mathcal{N}(0, 1)$.

This dataset was reanalyzed by Qiu et al. [2020], leading to an alternative dataset, also containing z-scores. This data is available from `https://github.com/njpipeorgan/L1000-bayesian`.

Note that the process of Qiu et al. [2020] involves discarding some subexperiments. These subexperiments are declared "invalid." To give the original analysis method its best chance, we adopt the following policy. If a hypothesis is rejected in one replicate and declared invalid in another replicate, we compute the RSR under the convention that this discovery failed to replicate.

A.2 Sci-Plex

The Sci-Plex data of Srivatsan et al. [2020] can be accessed from `https://github.com/cole-trapnell-lab/sci-plex`. In this work we specifically accessed the largest experiment in that paper, referred to in the data as sciPlex3. This data takes the form of RNA counts for many cells of many cell types for many genes. Following the standard of practice in the field, we begin by dividing the counts for each cell for each gene by the total number of counts for each cell Srivatsan et al. [2020]. Each cell arises from a particular condition. Some of these are control conditions (in which the cells are not treated with drugs) and some are treatment conditions (in which the cells are treated with drugs at various dosages). Each cell also lies on a particular "plate," and each plate may be subject to batch effects Srivatsan et al. [2020]. To partially control for these batch-effects, we only compare cells within a single plate. In particular, to test the null hypothesis that a particular gene in a particular cell type is unaffected by a particular drug at a particular dosage, we computed a Mann-Whitney test between that gene's expression in the treatment group and that gene's expression in the control cells residing on the same plate. Note that there are many cases where a gene exhibits zero expression in a given cell; a proper application of Mann-Whitney would therefore require tie-corrections to handle these zero values. On the other hand, tie-corrections will yield lower p-values when there are many ties—in this case many zero values—and genes with a preponderance of zero values are noisy and difficult to reason about. Ultimately, we did not adopt tie-corrections, as we found RSR to be higher when uncorrected p-values were used. (Note that the results in this work concerning RSR do not require the p-values ρ to be superuniform under the null.) We only consider hypotheses regarding genes which appear in at least 1.25% of all cells throughout the entire experiment.

A.3 Shifrut2018

The data of Shifrut et al. [2018] can be obtained from GEO accession GSE119450 (cf. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi`). It comprises many RNA counts for many cells in many conditions. We

focused specifically on cells in the "Stimulated" condition. As described in the main text, for some analyses we perform the normalization procedure (dividing the counts for each cell for each gene by the total number of counts for each cell), and in other analyses we do not. For testing hypotheses, we adopted the same conventions as used in the Sci-Plex data, above. We only consider hypotheses regarding genes which appear in at least 100 cells throughout the entire experiment.

A.4 Schmidt2022

The data of Schmidt et al. [2022] can be obtained from Zenodo accession 5784651 (cf. `https://zenodo.org`). We focused specifically on the "Re-stimulated" condition for CD4 T cells. As described in the main text, for some analyses we perform the normalization procedure, and in other analyses we do not. For testing hypotheses, we adopted the same conventions as used in the Sci-Plex data, above. We only consider hypotheses regarding genes which appear in at least 100 cells throughout the entire experiment.

B Comparing RSR values

We here consider settings where one wishes to estimate the difference between two RSR values.

Different kinds of experiments

Consider the case that we have observed training and validation replicates of two different kinds of experiments, yielding four replicates in all. Let $\rho^{(w)}, \hat{Y}^{(w)}, \alpha^{(w)}$ denote p-values, proposal selections, and a p-value threshold for the training replicate of the wth kind of experiment. Let $Y^{(w)}$ denote testing selections from a validation replicate of the wth kind of experiment. Assume there are $m^{(w)}$ hypotheses in the wth kind of experiment, and they have been partitioned as $\left\{1, 2, \cdots m^{(w)}\right\} = \cup_i P_i^{(w)}$.

- Let $X_i^{(w)} = \left|\left\{j \in P_i^{(w)} : \ \rho_j^{(w)} \leq \alpha^{(w)}, \ Y_j^{(w)} = \hat{Y}_j^{(w)}\right\}\right|$ for each w, i.
- Let $a_i^{(w)} = \left|\left\{j \in P_i^{(w)} : \ \rho_j^{(w)} \leq \alpha^{(w)}\right\}\right|$ for each w, i.
- Let $\mathrm{RSP}^{(w)} = \mathrm{RSP}(Y^{(w)}; \rho^{(w)}, \hat{Y}^{(w)}, \alpha^{(w)})$ and $\mathrm{RSR}^{(w)} = \mathbb{E}\left[\mathrm{RSP}^{(w)}\right]$ for each w.
- Let $\Delta = \mathrm{RSP}^{(1)} - \mathrm{RSP}^{(2)}$ and $\bar{\Delta} = \mathrm{RSR}^{(1)} - \mathrm{RSR}^{(2)}$.

Further assume that $\left\{X_1^{(1)}, \ldots, X_{m^{(1)}}^{(1)}, X_1^{(2)}, \ldots, X_{m^{(2)}}^{(2)}\right\}$ are independent.

Theorem 4. $\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq \exp\left(-2t^2 / \left(\frac{1}{\xi^{(1)}} + \frac{1}{\xi^{(2)}}\right)\right)$, *where*

$$\xi^{(w)} = \left(\sum_i a_i^{(w)}\right)^2 / \sum_i \left(a_i^{(w)}\right)^2.$$

Proof. Let $\mu_i^{(w)} = \mathbb{E}\left[X_i^{(w)}\right]$. Apply the Chernoff bound and independence to obtain

$$\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq e^{-st}\mathbb{E}\left[\exp\left(s\frac{\sum_i X_i^{(1)} - \mu_i^{(1)}}{\sum_i a_i^{(1)}}\right)\right]\mathbb{E}\left[\exp\left(-s\frac{\sum_i X_i^{(2)} - \mu_i^{(2)}}{\sum_i a_i^{(2)}}\right)\right].$$

For each w, Hoeffding's lemma then yields

$$\mathbb{E}\left[\exp\left(s\frac{\sum_i X_i^{(w)} - \mu_i^{(w)}}{\sum_i a_i^{(w)}}\right)\right] \leq \prod_i \exp\left(\left(\frac{s^2}{8}\right)\frac{\left(a_i^{(w)}\right)^2}{\left(\sum_i a_i^{(w)}\right)^2}\right) = \exp\left(\frac{s^2}{8\xi^{(w)}}\right)$$

For any $s \geq 0$ we have

$$\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq e^{-st}\exp\left(\frac{s^2}{8\xi^{(1)}} + \frac{s^2}{8\xi^{(2)}}\right).$$

Letting $s = 4t/(1/\xi^{(1)} + 1/(\xi^{(2)}))$ we obtain our final result. □

Different kinds of analyses

We now consider the case that we have observed measurements from a training and validation replicate for a single kind of experiment. We would like to consider two different ways of analyzing the measurements. Let $\rho^{(w)}, \hat{Y}^{(w)}, \alpha^{(w)}$ denote p-values, proposal selections, and a p-value threshold for the training replicate analyzed with the wth method. Let $Y^{(w)}$ denote testing selections from a validation replicate analyzed with the wth method. Assume there are m hypotheses, and for both analysis methods they have been partitioned as $\{1, 2, \cdots m\} = \cup_i P_i$. We assume that $\left\{\left(X_1^{(1)}, X_1^{(2)}\right), \ldots, \left(X_m^{(1)}, X_m^{(2)}\right)\right\}$ are independent; arbitrary dependency is permitted between $X_i^{(1)}, X_i^{(2)}$ for each i, but we assume independence between $\left(X_i^{(1)}, X_i^{(2)}\right)$ and $\left(X_j^{(1)}, X_j^{(2)}\right)$ for each i, j. Adopting the notations of the previous section, we obtain the following.

Theorem 5. $\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq \exp\left(-2t^2 / \sum_i \left(\frac{a_i^{(1)}}{\sum_{i'} a_{i'}^{(1)}} + \frac{a_i^{(2)}}{\sum_i a_{i'}^{(2)}}\right)^2\right)$.

Proof. Let $\mu_i^{(w)} = \mathbb{E}\left[X_i^{(w)}\right]$. Apply the Chernoff bound and independence to obtain

$$\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq e^{-st} \prod_i \mathbb{E}\left[\exp\left(s\frac{X_i^{(1)} - \mu_i^{(1)}}{\sum_{i'} a_{i'}^{(1)}} - s\frac{X_i^{(2)} - \mu_i^{(2)}}{\sum_i a_{i'}^{(2)}}\right)\right].$$

Applying Hoeffding's lemma for each i, we obtain

$$\mathbb{P}\left(\Delta - \bar{\Delta} \geq t\right) \leq \exp\left(\frac{s^2}{8}\sum_i \left(\frac{a_i^{(1)}}{\sum_{i'} a_{i'}^{(1)}} + \frac{a_i^{(2)}}{\sum_i a_{i'}^{(2)}}\right)^2 - st\right)$$

for each $s \geq 0$. Choosing

$$s = 4t/\sum_i \left(\frac{a_i^{(1)}}{\sum_{i'} a_{i'}^{(1)}} + \frac{a_i^{(2)}}{\sum_i a_{i'}^{(2)}}\right)^2$$

we obtain the desired result. □