HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2021 June 29.

Published in final edited form as:

Nat Methods. 2020 December ; 17(12): 1207–1213. doi:10.1038/s41592-020-00978-4.

Quantify and Control Reproducibility in High-throughput Experiments

Yi Zhao[1], **Matthew Sampson**[2], **Xiaoquan Wen**[*,1]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[2]Division of Nephrology, Boston Childrens Hospital, Boston, MA, USA

Abstract

We propose a set of computational methods, named INTRIGUE, to evaluate and control reproducibility in high-throughput experiments. Our approaches are built upon a novel definition of reproducibility, which emphasizes directional consistency (DC) when experimental units are assessed with signed effect size estimates. The proposed methods are designed to i) assess the overall reproducible quality of multiple studies; and ii) evaluate reproducibility at the level of individual experimental units. We demonstrate the proposed methods in detecting unobserved batch effects in high-throughput experiments via simulations. In an application of assessing reproducibility in transcriptome-wide association studies (TWAS), we illustrate the versatility of the proposed methods: in addition to reproducible quality control, they are also suited for investigating genuine biological heterogeneity. Finally, we discuss the extensions of the proposed reproducibility measures and potential applications in other vital areas of reproducible research (e.g., publication bias and conceptual replications).

Introduction

Reproducibility, or results reproducibility, concerns producing corroborating results across different experiments that aim to tackle the same scientific question. In literature, it is also referred to as replicability [1]. The reproducibility of scientific research has attracted rising attention in both the scientific community and the general public [1, 2, 3]. Biomedical research faces some unique challenges in reproducibility, especially with its wide adoption of high-throughput experimental technologies. High-throughput experiments enable simultaneous measurements of a large number of biological variables. However, the accuracy and the reproducibility of the results are often susceptible to unobserved confounding factors, commonly known as batch effects [4, 5, 6]. In recent years, many incidents of irreproducible research [7, 8] can be traced to failures of quality control in high-throughput experiments. While irreproducibility is often associated with unwanted batch effects, the very concept can be applied to study genuine biological heterogeneity. For

[*]**Corresponding author** Correspondence to Xiaoquan Wen (xwen@umich.edu).

Contributions

Y.Z., M.S., and X.W. conceived the ideas. Y.Z. and X.W. designed the experiments. Y.Z. and X.W. developed methods, implemented software, and performed analyses. Y.Z, M.S., and X.W. wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

example, mapping tissue-specific expression quantitative trait loci (eQTLs) using high-throughput assays can be framed as a problem of identifying irreproducible genetic effects on gene expressions in different cellular environments [9, 10]. In this context, strong irreproducible eQTL effects across tissues may imply altered molecular mechanisms for gene regulation [11]. In both scenarios above, assessing irreproducibility is to evaluate the extent of discordance in outcomes among multiple experiments.

Although there is a rich literature on quantifying reproducibility/replicability, the available results primarily focus on the consistency of (statistically) significant findings [12, 13]. As a result, most existing methods do not address the unique settings of high-throughput experiments, nor do they provide an overall assessment for the whole replication experiment. One of the pioneering works in quantifying irreproducibility in high-throughput experiments is [14] motivated by the analysis of CHIP-seq data. The authors characterized reproducible signals as highly and consistently ranked experimental units across experiments and proposed a computational framework, *irreproducible discovery rate* (IDR), to enable rigorous false positive control of such reproducible findings. Although rank information is universally available in all types of high-throughput experiments, we note that a vast majority of the experiments offer much richer information. Notably, many experiments provide a signed effect estimate (along with its standard error) for each experimental unit. For example, the estimated gene-level effect in differential gene expression experiments quantifies the difference in expression abundance. It is natural to consider the consistency of estimated effects among multiple experiments as a measure of reproducibility.

In this paper, we re-examine the characterization of reproducible signals in a setting where each experimental unit is assessed with a signed effect size estimate. In this context, reproducibility is intuitively defined by the quantitative concordance of estimates from multiple experiments. Despite this simple intuition, several conceptual considerations are critical in implementing this idea. First, we argue that the concordance criteria should be applied to the underlying unobserved true effects (rather than the observed/estimated effects). Specifically, the estimated effect should be viewed as a noise-contaminated observation from the real effect: consider a replication experiment with a much larger noise level than the original experiment. Despite that the observed effects from the two experiments are inconsistent, it is illogical to conclude that the initial experiment is irreproducible because the noisy observations in the second experiment are simply less informative (for recovering the true underlying effects). Second, it is unrealistic to assume that the underlying effects of reproducible signals in all experiments are *exactly the same*. However, they should be reasonably similar. The key is to define an acceptable level of heterogeneity in effect sizes for reproducible signals. To the best of our knowledge, there are no known results for such a definition available in the literature. Finally, we note that by imposing some consistency criteria on effect sizes, all experimental units can be classified into three mutually exclusive categories. They represent non-signals (which have consistent zero true effects in all experiments), reproducible signals (which have consistent non-zero true effects in all experiments), and irreproducible signals (which have inconsistent effects across experiments), respectively. This classification scheme is fundamentally different from the rank-based reproducibility quantification methods, which allow only two latent categories: consistent vs. inconsistent.

In this paper, we formalize the concept of reproducibility and propose a novel statistical framework to quantify and control reproducibility in high-throughput experiments. We examine our proposed computational approaches in both simulation studies and real data applications. The proposed statistical methods are implemented in the software package INTRIGUE (quantIfy and coNTROl reproducIbility in hiGh-throUghput Experiments), which is freely available at https://github.com/artemiszhao/intrigue. A docker image that enables complete replications of the numerical results from our simulations and real data analysis is also included.

Results

Overview of computational methods

Under our specified setting, the fundamental distinction between reproducible and irreproducible signals is solely reflected by the heterogeneity among the underlying effects in repeated measurements. To provide a reasonable standard to distinguish exceedingly high and acceptable low levels of heterogeneity, we propose to characterize reproducibility by specifying ***a minimum requirement of the maximum tolerable heterogeneity*** for a common-sense reproducible signal. Based on this consideration, we argue that, with a high probability, the underlying effects of reproducible signals are expected to have the same (positive or negative) sign. Henceforth, we refer to this criterion as the requirement of ***directional consistency*** (DC). Given the true effect, the DC criterion establishes a range of reasonable variability for a reproducible signal.

The directional consistency criterion is a natural extension of Tukey's argument for detectable effects [15, 16], i.e., an effect is reliably detected if its direction (positive or negative) can be confidently determined. In our context, DC characterizes the reliability of a reproducible signal in repeated measurements. DC also enjoys the scale-free property, i.e., the criterion is invariant even if the measurement scales are different in different experiments. For example, it is possible to consider the reproducibility of multiple differential gene expression experiments conducted using different technologies (e.g., microarray vs. RNA-seq) under our proposed theoretical framework.

We propose two Bayesian hierarchical models, the CEFN and the META models, to implement the DC criteria for high-throughput data analysis. Both models explicitly parameterize and quantify the heterogeneity between the true underlying effects for a given experimental unit (e.g., a gene) across multiple experiments. Importantly, the measures of heterogeneity can be translated into precise probabilistic statements for interpreting the DC criteria. In the CEFN model, the tolerable level of heterogeneity is adaptive to the underlying effect, i.e., a higher level of heterogeneity is expected if the true underlying effect increases. We refer to such a property as the *adaptive expected heterogeneity*. In comparison, the expected (tolerable) heterogeneity is invariant for the magnitude of true effects in the META model. Both properties can be useful depending on the application context.

In both statistical models, each experimental unit can be classified into three mutually exclusive latent classes according to its underlying (grand) effect and the heterogeneity exhibited in multiple experiments: i) the null signals (which have consistent zero effects); ii)

reproducible signals (which exhibit consistent non-zero effects); and iii) irreproducible signals (whose effect size heterogeneity exceeds the expectation by the DC criteria).

Given the estimated effects and their corresponding standard errors for all experimental units in multiple experiments, we develop an empirical Bayes procedure to conduct inference. (Note that our methods also work with *z*-statistics, or signed *p*-values, as the estimated effects at the signal-to-noise ratio scale.) The procedure is implemented in an expectation-maximization (EM) algorithm by treating each experimental unit's latent class status as missing data. Particularly, the procedure estimates the proportions of null (π_{Null}), reproducible (π_{R}), and irreproducible (π_{IR}) signals for all experimental units. Additionally, we report a quantity, ρ_{IR}, to measure the relative proportion of irreproducible findings in non-null signals, i.e.,

$$\rho_{\text{IR}} := \frac{\pi_{\text{IR}}}{\pi_{\text{IR}} + \pi_{\text{R}}} \tag{1}$$

The combination, (π_{IR}, ρ_{IR}), serves an informative indicator to quantify the severity of lack of reproducibility from observed data. Each experimental unit is subsequently assessed with three posterior classification probabilities corresponding to the three latent classes. The resulting posterior probabilities are used in the false discovery rate (FDR) control procedures [17] to identify reproducible and irreproducible signals.

Simulation studies

Evaluation of computational methods—We design a series of simulations to evaluate the performance of the proposed methods and investigate the practical factors that impact the assessment of reproducibility in high-throughput experiments. This set of simulations mimics the genome-wide quantitative trait locus (QTL) mapping experiments, and we implement different schemes by changing the proportions of three types of signals, adjusting sample sizes in replication studies, and varying the numbers of involved studies.

First, we examine the accuracy of the estimated proportions for π_{Null}, π_{R}, and π_{IR}. The results, summarized in Supplementary Table 1, indicate that the proposed EM algorithm provides accurate proportion estimates for both the CEFN and the META models and is robust to uneven sample sizes in multiple experiments. Next, we examine the accuracy of reproducible probabilities assessed at the level of individual experimental units using the simulated data sets. Specifically, we investigate the calibration of the inferred posterior probability for reproducible classifications by contrasting the assessed posterior probabilities with the corresponding frequencies of true reproducible signals. If the posterior probabilities are calibrated, a group of experimental units assessed with posterior reproducible probability *p* should have approximately (100 × *p*)% of truly reproducible observations. Overall, we find that both the CEFN and the META models yield reasonably calibrated probabilistic quantification of individual reproducibility (Figure 1). For both models, the calibration is particularly accurate for modest to high values of reproducible probabilities. Both models seem to be conservative in the range of lower values, but it should not lead to inflation of type I errors. Finally, we investigate the quantitative relationship between the number of

replications and the power to detect reproducible and irreproducible signals. The receiver operating characteristic (ROC) curves shown in Figure 1 indicates that the area under the curve (AUC) monotonically increases with increased replication numbers for classifying reproducible and irreproducible signals.

Simulation for batch effect detection—One of the essential practical applications for the proposed methods is to detect batch effects in high-throughput experiments. Provided that a high-quality reference dataset generated from a similar experimental setting is available, we illustrate our methods in assessing the reproducibility of a new dataset in the presence of batch effects. Following the scheme in [18], we consider a differential gene expression (DE) experiment of 1,000 genes with 20 cases and 20 controls and use a location-and-scale model [19] to simulate various magnitude of batch effects that are also partially correlated with the case-control status. Consequently, the DE analysis likely produces false-positive findings without proper controls of latent batch effects. In all our simulations, half of the genes are contaminated by the batch effects.

Both the CEFN and the META models are shown excellent sensitivity in detecting modest batch effects: their performance is comparable to the gold standard two-sample Kolmogorov-Smirnov (K-S) test (Table 1 and Extended Data Figure 1). More importantly, the proposed approaches show the unique ability to identify reliable findings and irreproducible results through rigorous FDR control of reproducibility at gene-level (Table 2). Additionally, we examine the performance of batch removal procedures and their impacts on reproducibility. We find that although such a procedure is generally effective, it does not completely eliminate the false positive findings. Nevertheless, when combined with the proposed methods, our ability to identify genuine biological signals is greatly enhanced even in the presence of batch effects (Table 2).

Real data applications

We demonstrate the proposed methods in an application of transcriptome-wide association study (TWAS) utilizing multi-tissue eQTL data from the GTEx project. TWAS analysis is a type of integrative analysis approach that tests the association between genetically predicted gene expression levels (based on tissue-relevant eQTL information) and the complex trait of interest [20, 21]. Significant associations identified from this procedure imply potential causal relationships from gene expressions to the trait under certain causal assumptions [22].

Reproducibility of height GWAS between UK Biobank and GIANT consortium —In the first analysis, we examine two independent GWAS of standing heights from the UK Biobank (UKB) and the GIANT consortium (GIANT), respectively. For both datasets, we use the GTEx skeletal muscle eQTL data to compute the genetically predicted gene expressions. If the two large-scale GWAS results are consistent, the two sets of the TWAS results are expected to be consistent. We use the z-scores from the two TWAS datasets as input and assess the reproducibility across 32,362 genes. We note that the UKB data has a larger sample size (337K individuals) than the GIANT data (253K individuals), and, as expected, the absolute values of z-scores for significant genes are often stronger in the UKB

data. Nevertheless, as shown in Figure 2, the overall trend of directional effects remains highly consistent with only a few outlying data points showing directional mismatches.

The proportion estimates by both the CEFN and the META models confirm that the two GWAS datasets are highly concordant (Table 3). Both models estimate that the proportion of irreproducible signals is close to 0. As expected, the META model yields slightly more conservative reproducibility assessment due to its assumption of non-adaptive heterogeneity, which emphasizes consistency in both signal directions and magnitude.

At 5% FDR level, the CEFN model identifies 3166 reproducible genes. The *q*-value procedure [23] identifies 1437 and 2653 significant TWAS genes at 5% FDR level in separate analyses of the GIANT and the UKB datasets, respectively. 97% of the GIANT TWAS genes and 91% of the UK Biobank TWAS genes are in the identified reproducible TWAS gene sets. The much-improved power over the analysis of individual datasets illustrates that the proposed approach achieves a similar benefit as a traditional meta-analysis while incorporating tolerable heterogeneity.

This example represents a large class of applications for which reproducibility is evaluated by the measurements of signal-to-noise ratios (e.g., signed *p*-values, *t*-statistics). Given the context of this experiment, the systematic difference in *z*-scores for associated genes between the two studies is well expected due to the discrepancy of the sample sizes. We perform the Cochran's Q test for each gene using the observed *z*-scores and observe that the empirical distribution of the derived *p*-values are distinctively right-skewed (Section 4.2.1 of the Supplementary Notes). The CEFN model is uniquely suited in this example: it not only provides the sensible evaluation of reproducibility but also automatically aggregates the evidence for deemed reproducible genes across studies. The latter feature is particularly important, especially when the traditional fixed-effect meta-analysis approach is not readily applicable for a good proportion of reproducible genes (Section 4.2.1 of the Supplementary Notes).

Reproducibility of TWAS using blood and muscle transcriptome—In the second experiment, we focus on the height GWAS data from the UK Biobank but predict the gene expressions based on the GTEx eQTL data from two distinct tissues, namely, skeletal muscle and whole blood. Because the genetically predicted gene expressions reflect the tissue-specific gene regulation, the comparison of the corresponding TWAS analyses should shed light on the tissue-specificity of identified gene-level associations in TWAS.

Our main results focus on a subset of 7,734 "eQTL genes", or eGenes [11, 24], which are pre-selected and satisfy two criteria: i) they are considered well-expressed in both tissues (Section 4.2 of the Supplementary Notes); and ii) they harbor at least one *cis*-eQTL (not necessarily identical) in both tissues. For this subset of genes, the inconsistent TWAS results between tissues are likely due to differential regulation patterns. We find that a substantial proportion of eGenes are deemed irreproducible in tissue-specific TWAS, indicating that potential causal molecular mechanisms from genes to traits are qualitatively different in different tissues.

There is a caveat to the application of the proposed methods in this specific setting. Both the META and the CEFN models assume that the observed effects from the null signals are independent across experiments. This assumption is likely violated because the analyses in different tissues utilize the same GWAS data set, which introduces correlations in observed *z*-scores even under the null. This violation potentially leads to anti-conservative estimates of null proportions. Nevertheless, we find the estimates of null proportions of all genes by both models are very close to the estimates using multiple GWAS data (Table 3). We also repeat the tissue analysis using the TWAS results from i) UKB Height GWAS and GTEx whole blood eQTL data; and ii) GIANT Height GWAS and GTEx skeletal muscle eQTL data. In this setting, the independent assumption under the null is met. The results, summarized in Section 4.3.1 of the Supplementary Notes, are overly similar to Table 3 with increased π_{IR} and π_{Null} estimates. Although the interpretation in this new analysis may be more complicated given the difference in GWAS data, our main conclusions remain the same: a significant portion of TWAS signals for heights are tissue-specific.

At individual gene-level, the CEFN model almost exclusively identifies irreproducible eGenes corresponding to the outlying data points in the second and the fourth quadrants in Figure 3. In comparison, the META model also flags the data points with a noticeable magnitude difference in *z*-scores despite the directional consistency. In this context, both models are uniquely valuable: the CEFN model effectively identifies qualitative interactions between genes and tissue/cellular environments; the additional quantitative interactions identified by the META model are also of biological interest.

Discussion

In this paper, we propose a statistical framework to aid quantification and control of reproducibility in high-throughput experiments. Assuming signed effect estimates are available from multiple replicated experiments, we formally introduce a probabilistic definition of reproducibility based on the directional consistency criterion. Built on this definition, we propose the probabilistic models to i) assess the overall reproducibility of multiple studies through parameter estimation; and ii) evaluate reproducibility for individual experimental units. The proposed methods are applicable to a wide range of analyses for high-throughput experiments, including differential expression analysis, molecular QTL mapping, genome-wide genetic association studies (GWAS), and more.

The DC criterion is one of our key contributions to reproducibility research, and its application goes beyond the scope of high-throughput experiments. The DC criterion provides a principled way to define a maximum level of acceptable heterogeneity. If the observed variation between replications is deemed to exceed this maximum tolerable level, the practical factors that lead to irreproducible results should be carefully investigated. To illustrate, we provide an example of detecting publication bias based on the proposed META model and a Bayesian model comparison approach. Publication bias refers to the phenomenon that reported effect estimates in scientific publications tend to be exaggerated. It is known that naturally occurred between-study heterogeneity and publication bias are two convoluted factors, and objective assessment of publication bias requires explicit consideration of between-study heterogeneity [25, 26]. Ignoring between-study

heterogeneity when assessing publication bias often yield misleading conclusions [27, 25, 28]. In our illustrative example, we simulate two studies estimating a treatment effect (Section 5 of the Supplementary Notes). The effect size estimate for the smaller study is generated from a selection process that mimics aggressive *p*-hiking behaviors. For each observation, we compute a Bayes factor to compare the evidence for irreducible and reproducible models based on the reported estimates. We find that this approach exhibits excellent sensitivity and specificity in detecting publication bias with merely two studies (see results in Section 5 of the Supplementary Notes). In comparison, other existing approaches that rely on detecting funnel plot asymmetry (e.g., Egger regression) typically requires more than ten replication studies [29].

Our proposed approaches also have close connections to assessing heterogeneity in a meta-analysis: both the CEFN and the META model resemble a random-effects meta-analysis model, and both models converge to the fixed-effect meta-analysis model in the limiting case of no heterogeneity (i.e., $k \to 0$; $r \to 0$). Similar to meta-analysis models, the quantification of heterogeneity in INTRIGUE is achieved by evaluating the between-study variance of the underlying effect sizes. However, the traditional meta-analysis often assesses the *extent of heterogeneity* by contrasting to a fixed-effect model (e.g., using Cochran's Q test) [30]. We argue that for assessing reproducibility, the fixed-effect meta-analysis model may not be a natural benchmark because the diversity between studies is inevitable. In comparison, the DC criterion formulates a much relaxed and reasonable standard and also enables a refined categorization of null, irreproducible, and reproducible signals. Therefore, we consider applying the DC criterion to define an expected level of heterogeneity *a priori* is the main distinction and novelty compared to the traditional meta-analysis approaches. Finally, we note that traditional meta-analysis typically considers only a few experimental units/effects, whereas the proposed methods are designed for high-dimensional measurements generated from high-throughput experiments. We are able to take advantage of this setting and apply Bayesian hierarchical modeling to achieve effective partial pooling of information on reproducibility from strong, modest, and even weak signals of all types.

Our work also complements the seminal work by Li et al. [14], which quantifies reproducibility via rank consistency. The proposed approaches utilize different quantitative information than the relative rankings of experimental units in a high-throughput setting. When such information is available, we find that our proposed approaches refine the classification of signals and overcome some computational difficulties in the IDR method. Nevertheless, the rank-based methods are more general for requiring less information content, and their applications are also broader.

The concepts and general ideas proposed in this paper are also applicable to *conceptual replications*, which refer to applying different methods to reproduce the result from a previous research work without replicating the complete experimental procedure [31]. One critical limitation of our current methods is its independent assumption of observed null signals from replications, which we have discussed in the example of multi-tissue TWAS analysis. As the applications of conceptual replications often share the same underlying data, explicit modeling of the correlations between null observations becomes necessary in such application scenarios. In some specific statistical settings, e.g., a multivariate linear

regression model (MVLR), the proposed META model can be conveniently incorporated into the framework proposed in [32] to resolve this issue. Nevertheless, more general solutions, especially for the CEFN model, require future work.

Methods

Model details

We consider a general setting where p experimental units (e.g., genes) are measured in M experiments. In the j-th experiment, we assume that the effect of unit i is represented by a point estimate, $\hat{\beta}_{i,j}$, and its corresponding standard error, $\sigma_{i,j}$. Let $\beta_{i,j}$ denote the underlying true effect of unit i, we assume that

$$\hat{\beta}_{i,j} \mid \beta_{i,j} \sim \mathrm{N}(\beta_{i,j}, \sigma^2_{i,j}), \tag{2}$$

where $\sigma_{i,j}$ models the noise contamination to the true effect.

The reproducibility is directly reflected by the variability of $\beta_{i,j}$'s among different experiments. To this end, we introduce a grand effect for unit i, $\bar{\beta}_i$, for a hypothetical population of possible replicating experiments. We assume

$$\bar{\beta}_i \sim \mathrm{N}(0, \omega^2), \tag{3}$$

where ω characterizes the magnitude of the overall effect. For a reproducible signal, $\beta_{i,j}$s are expected closely clustered around $\bar{\beta}_i$. To implement the DC criterion, we propose two different prior models to describe the directional relationships between $\beta_{i,j}$'s and $\bar{\beta}_i$.

Curved exponential family normal prior model (CEFN)—In this prior model, we assume

$$\beta_{i,j} \mid \bar{\beta}_i,\ k \sim \mathrm{N}(\bar{\beta}_i, k^2\bar{\beta}_i^2),\ k \geq 0,\ j = 1, 2, \ldots, M\,. \tag{4}$$

Notably, the variance of the assumed distribution (4) depends on its mean parameter and is scaled by a heterogeneity parameter, k. In the extreme case when $k \to 0$, (4) degenerates to the fixed-effect assumption in meta-analysis. Importantly, (4) has the critical implication that directly connects to the DC criterion, i.e.,

$$\Pr(\beta_{i,j} \text{ has the correct sign as } \bar{\beta}_i) = \Phi\left(\frac{1}{k}\right), \tag{5}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution (a detailed derivation is provided in Section 6 of the Supplementary Notes). Eqn (5) shows that when $k \to 0$, the probability $\to 1$, whereas $k \to \infty$, the probability $\to 0.50$. The adaptive expected heterogeneity of the CEFN model is due to the explicit mean-variance relationship in (4). The variability of $\beta_{i,j}$'s increases with $\bar{\beta}_i^2$ for a fixed k (representing a fixed DC probability) value.

Meta-analysis prior model (META)—In this model, we adopt a Bayesian meta-analysis model [33] to describe the heterogeneity of $\beta_{i,j}$'s. Particularly, we assume that

$$\beta_{i,j} \mid \bar{\beta}_i, \phi^2 \sim \mathrm{N}(\bar{\beta}_i, \phi^2), \; j = 1, 2, \ldots, M. \tag{6}$$

Let $r := \frac{\phi^2}{\phi^2 + \omega^2}$, the DC criteria has a probabilistic interpretation through the following integral equation,

$$\Pr(\beta_{i,j} \text{ has the correct sign as } \bar{\beta}_i) = \sqrt{\frac{2}{\pi}} \int_0^{+\infty} \Phi\left(\sqrt{\frac{1-r}{r}} t\right) e^{-\frac{t^2}{2}} dt. \tag{7}$$

Note that the limiting behaviors of the META model for extreme heterogeneity, i.e., as $r \to 0$ and $r \to 1$, coincide with the CEFN model.

Both the CEFN model and the META model can be represented by a directed acyclic graph (DAG) (Extended Data Figure 2).

Partition of model space

The space of the proposed hierarchical model based on the CEFN prior is parameterized by the combination of (k, ω), which can characterize all three latent classes in observed data.

1. Non-signal: $\omega = 0$. Under this setting, the value of k becomes irrelevant, and $\bar{\beta}_i = \beta_{i,1} = \cdots = \beta_{i,M} \equiv 0$.
2. Reproducible signal: $\omega \neq 0$ and k is small. Under this setting, the generative model will produce $\beta_{i,j}$'s with the same (correct) sign with a high probability. Hence the DC requirement is satisfied.
3. Irreproducible signal: $\omega \neq 0$ and k is large. $\beta_{i,j}$'s generated from this setting have variability exceeding the DC criteria and are considered irreproducible.

For the META model, the partition of the model space can be similarly formed by the values of (r, ω).

To ease the computational burden in inference, we choose to characterize the continuous model space by a dense grid of discrete parameter combinations. Specifically, the grid point ($k_0 = 0$, $\omega_0 = 0$) represents the non-signals and is always included. Following [16], we adopt a data-driven strategy to specify a dense grid values of (k, ω) for alternative scenarios (Section 1.2 of the Supplementary Notes). By default, we set DC probabilities {0.990, 0.975, 0.950} to represent reproducible signals and {0.750, 0.700, 0.650} to represent irreproducible signals. (Our software implementation also allows user-specified probability classifications.) Additionally, a sequence of ω values is computed to cover a wide range of possible grand effects.

For the l-th grid value in the model space, we denote its (prior) weight by π_l. Let M_0, M_R and M_{IR} denote the subsets of grids corresponding to non-signals, reproducible, and

irreproducible signals, respectively. The prior probabilities for the three latent categories can be represented by

$$\begin{aligned}\pi_{\text{Null}} &:= \Pr(M_0) = \pi_0 \\ \pi_{\text{R}} &:= \Pr(M_R) = \sum_{l \in M_R} \pi_l \\ \pi_{\text{IR}} &:= \Pr(M_{IR}) = \sum_{l' \in M_{IR}} \pi_{l'}\end{aligned} \tag{8}$$

Given the partition of the model space, the joint modeling of all experimental units can be achieved by assuming exchangeability. More specifically, for L pre-specified (k, ω) grids, we introduce a 1-of-L latent indicator vector, $\boldsymbol{\gamma}_i$ for each experimental unit i and assume that $\boldsymbol{\gamma}_i$'s are sampled with replacement from the (prior) probability distribution $\{\pi_l : l = 0, \ldots, L - 1\}$. The point of interest for inference is to compute the posterior probability, $\Pr(\boldsymbol{\gamma}_i \in M_R \mid \text{Data})$, which aids the classification of the corresponding experimental unit.

Statistical inference

We design an EM algorithm to estimate the weight parameter $\boldsymbol{\pi} := (\pi_0, \ldots, \pi_{L-1})$. Using the resulting point estimate, $\hat{\pi}$, we compute $\hat{\pi}_{\text{R}} = \sum_{i \in M_R} \hat{\pi}_i$ and $\hat{\pi}_{\text{IR}} = \sum_{j \in M_{IR}} \hat{\pi}_j$ to represent the estimated proportions of reproducible and irreproducible experimental units in the observed data, respectively. Within the empirical Bayes framework, we evaluate the posterior probability that experimental unit i is generated from the l-th model by

$$\Pr\left(\gamma_i \to l \mid (\hat{\beta}_{i,1}, \sigma_{i,1}), \ldots, (\hat{\beta}_{i,M}, \sigma_{i,M}), \hat{\pi}\right) = \frac{\hat{\pi}_l \, \text{BF}_{i,l}}{\sum_{l'} \hat{\pi}_{l'} \, \text{BF}_{i,l'}}, \tag{9}$$

where $\boldsymbol{\gamma}_i \to l$ represents that $\boldsymbol{\gamma}_i$ indicates the l-th model and $\text{BF}_{i,l}$ denote the Bayes factor/marginal likelihood for the l-th generative model. Thus, the posterior probability that experimental unit i is reproducible can be computed by

$$\Pr(\gamma_i \in M_R \mid \text{Data}) = \sum_{l \in M_R} \Pr\left(\gamma_i \to l \mid (\hat{\beta}_{i,1}, \sigma_{i,1}), \ldots, (\hat{\beta}_{i,M}, \sigma_{i,M}), \hat{\pi}\right). \tag{10}$$

The details of the EM algorithm and the numerical computation of the relevant Bayes factors are described in Section 1.4 of the Supplementary Notes.

Note that

$$lfdr_{i,\,\text{reproducible}} := 1 - \Pr\left(\gamma_i \in M_R \mid (\hat{\beta}_{i,1}, \sigma_{i,1}), \ldots, (\hat{\beta}_{i,M}, \sigma_{i,M}), \hat{\pi}\right) \tag{11}$$

is the estimated local false discovery rate (lfdr) for the reproducibility classification of experimental unit i. The posterior probabilities allow us to perform FDR control without extra cost. In comparison, deriving p-values from the complex null hypothesis stating unit i is either a non-signal or irreproducible seems considerably challenging both conceptually

and computationally. The lfdr's for the irreproducible classification can be similarly computed from the corresponding posterior probabilities for M_{IR}.

Model robustness

The proposed statistical models assume the normality of the observed summary-statistics from each experimental unit. The normal likelihood assumption is a reasonable approximation for a wide range of commonly applied parametric models involving maximum likelihood estimation, including the families of linear and generalized linear models. Many parametric models that are used in analyzing high-throughput sequencing data fall into this category. The prior distributions in the CEFN and the META models for modeling non-null signals are formed by a dense scaled-normal mixture (rather than a single normal distribution), which can be used to approximate a broad spectrum of distributions. As a result, both the marginal data probability (i.e., the unconditional distribution of observed data) and the posterior probability distribution of the shared effect size between studies are extremely flexible [16].

Like Efron's local FDR method, our proposed methods are applicable for any valid signed *p*-value derived from either parametric or non-parametric approaches. The signed *p*-values can be converted to *z*-scores via a quantile normal transformation [17]. The resulting *z*-scores follow the standard normal distribution under the null, and the distributions for the non-null signals are modeled by the flexible scaled-normal mixtures. We show an illustration of this transformation in Section 1.3.1 of the Supplementary Notes.

Additional properties of CEFN prior

The CEFN prior (4) is a key component of the proposed methods. Here we discuss some properties and implications derived from the prior, which should provide further insights.

First, we examine the directional consistency of *observed* $\hat{\beta}_{i,j}$'s. The CEFN prior and the sampling distribution (2) jointly imply that the following probabilistic statement for the directional consistency in the observed data, i.e.,

$$\begin{aligned}&\Pr\left(\hat{\beta}_{i,j}\text{'s all have the same sign} \mid \bar{\beta}_i^2,\, k^2\right)\\&= \prod_{j=1}^{M}\left[1-\Phi\left(-\frac{1}{\sqrt{k^2+\sigma_{i,j}^2/\bar{\beta}_i^2}}\right)\right]+\prod_{j=1}^{M}\Phi\left(-\frac{1}{\sqrt{k^2+\sigma_{i,j}^2/\bar{\beta}_i^2}}\right)\end{aligned} \tag{12}$$

The detailed derivation is provided in the Section 1.5 of the Supplementary Material.

Equation (12) indicates the directional consistency of observed data is determined by the interplays of the effect size heterogeneity (i.e., the property of reproducibility) and the signal-to-noise ratio (represented by $\bar{\beta}_i^2/\sigma_{i,j}^2$). Both factors can dictate the variations of the estimated effects. If the noise in the data becomes dominant, the data can be uninformative on the underlying effect size heterogeneity, which is a point that we have discussed in the introduction. In an extreme example of two experiments ($M = 2$), if one of the experiments has extremely high-level of noise on unit *i*, it follows that

$$\lim_{\frac{\sigma_{i,1}^2}{\bar{\beta}^2} \to \infty} \Pr(\hat{\beta}_{i,1} \text{ and } \hat{\beta}_{i,2} \text{ have the same sign}) = \frac{1}{2}, \tag{13}$$

regardless of the value of k. Thus, accurate inference of k is contingent on the condition that experimental noises are well-controlled.

Next, we consider k is the dominant factor. It can be shown that, with all else being equal, the sign consistency probability in (12) is decreasing as the number of the experiments, M, increases. However, the rate of decrease is determined by k, i.e., large k values lead to rapid decreases, whereas the rate of decrease also approaching 0 as $k \to 0$. A direct implication from this observation is that multiple experiments are more informative in identifying irreproducible signals. If the number of repeated experiments is limited, the observed data from irreproducible signals may show directional consistency by chance. However, such chance diminishes quickly as more replication experiments become available. We will further illustrate this point in our simulation studies.

Finally, we note that the proposed CEFN prior is independent of the probabilistic assumptions on $\bar{\beta}_i$. In our proposed model, the generative model for $\bar{\beta}_i$ follows the uni-modal assumption (UA) [16]. However, this is not required to utilize the CEFN prior. In Section 1.6 of the Supplementary Material, we show that the CEFN prior is valid for an arbitrary generative model of $\bar{\beta}_i$. This property is unique to the CEFN model. In comparison, the DC implication from the META model relies on the UA assumption of $\bar{\beta}_i$.

Simulation details

Evaluation of computational methods—The general setting for this set of simulations mimics a genome-wide expression quantitative trait locus (eQTL) mapping experiment investigating $p = 1000$ genes. Each gene's genetic effect in each experiment is independently simulated from one of the reproducible, the irreproducible, and the null model. The corresponding gene expressions are subsequently simulated from a linear model.

For evaluating the estimation accuracy of the EM algorithm, we create three simulation schemes, S1, S2, and S3, by varying the proportions of the null, reproducible and irreproducible signals and the combinations of the sample sizes in the two studies. Throughout this experiment, we consider $M = 2$ experiments. Notably, we select the sample sizes to mimic the study designs in the early and current eQTL studies. The simulation details are provided in Section 2.1 of the Supplementary Notes. Additional comparison to the rank-based IDR method is provided in Section 2.2 of the Supplementary Notes.

To examine the calibration of reproducible probabilities at the gene-level, we design a new scheme, S4, by further reducing sample sizes ($N_1 = 50$, $N_2 = 40$) to cover a full range of probabilistic quantification better.

Simulation scheme S5 is applied to quantify the relationship between the number of replications (M) and the power to detect reproducible and irreproducible signals. In this

scheme, the number of replications from $M = 2$ to $M = 10$ and setting the sample size in each study = 50.

Additionally, we alter the data generating model to simulate reproducible and irreproducible effects in additional sets of simulated data. The results by INTRIGUE remain mostly invariant (Section 2.3 of the Supplementary Notes).

Batch effect detection—In the first set of simulations, we assume that no true biological signals in all genes. Hence all rejections from the hypothesis testings are indeed false-positive findings. Our main interest here is to investigate the quantification and the corresponding sensitivity of various approaches in detecting unobserved batch effects. Our input to all tested methods for each gene includes the estimated DE effects and the corresponding standard errors from $M = 2$ separate experiments. Particularly, the data from the first experiment are NOT contaminated by the batch effects and are regarded as a gold-standard reference. In the replication dataset, we simulate the batch effects using the location-and-scale model for 50% of the genes, and the batch labels are partially correlated with the case-control status. The sizes of the batch effects are drawn from a random effect model, $N(0, \eta^2)$, where we vary the magnitude ratio of the batch effects to the additional white noise, η/σ, from 0 to 2 in different simulated datasets (σ^2 represents the residual error variance for the white noise). When the replication dataset is analyzed alone using the q-value procedure at the FDR 5% level, there are less than 3 rejections when $\eta/\sigma \leq 0.6$, the false rejections rise to 4.5% and 25% as η/σ reaches 1.0 and 2.0, respectively (Table 1).

We apply the META and the CEFN models on the simulated datasets and focus on their estimates on the irreproducible and reproducible proportions. For comparison, we apply the two-sample Kolmogorov-Smirnov (K-S) test to examine the concordance of the p-value distributions from the reference and the target datasets. We also apply the IDR method to analyze the replication and the reference datasets (Section 3.2 of the Supplementary Notes). The resulting estimates indicate a lack of rank-consistent signals when the estimation procedure is initialized close to the truth. However, we find that the IDR estimates can be unstable when the initial values for the model parameters are randomly altered, indicating that the latent copula model space may be multi-modal (Section 3.2 of the Supplementary Notes). Furthermore, there is no quantity similar to the irreproducible proportion estimate to alarm practitioners for potential latent batch effects in this application context.

In the second set of experiments, we additionally simulate genuine DE effects for 20% of the genes for each dataset, where the genuine effects are independently drawn from a random effect distribution, $N(0, \kappa^2)$, and we set $\kappa/\sigma = 1$ throughout. In the replication dataset, the batch effects impact 50% of the DE genes and 50% non-DE genes. We also apply the procedure, ComBAT, on each replication dataset to adjust the unobserved batch effects and create a corresponding processed replication dataset. When the replication datasets are analyzed alone (which we refer to as the “stand-alone analysis”), there are excessive false-positive errors even when the magnitude of the batch effects are well below the detectable levels by the K-S tests (Section 3.3 of the Supplementary Notes). The batch effect adjustment procedure is shown quite effective in reducing type I errors, however, the false

discovery rates are still not properly controlled in most cases, indicating that complete removal of batch effects by computational means is extremely difficult.

We apply both the CEFN and the META model to analyze the reference and the replication datasets jointly. The results are summarized in Table 2. In the joint analysis, we find that both models properly control the FDR of genuine DE signals at the pre-specified 5% level for all examined magnitudes of batch effects. More importantly, when the magnitude of the batch effects is smaller than the genuine biological effects, the proposed approaches effectively maintain the power of discovery. Nevertheless, when the batch effects overwhelm the genuine effects, i.e., $\eta/\sigma \geq \kappa/\sigma$, the power from both models diminishes to ensure the desired false positive control. The latent batch effect adjustment procedure again illustrates its usefulness. When the processed replication datasets are jointly analyzed with the gold-standard reference dataset, we observe marked power recovery even when the magnitude of the batch effects is close to the biological effects.

Assessing reproducibility of TWAS analysis

Statistically, TWAS analysis is a special case of the burden test. Given a GWAS dataset, ($\boldsymbol{y}$, $\boldsymbol{G}$), and a set of weights for relevant genetic variants derived from eQTL studies, $\{w_i\}$, it aims to examine the correlation between the phenotype $\boldsymbol{y}$ and the weighted sum, $\sum_i w_i G_i$, for a target gene. The weighted sum is often interpreted as a linear genetic prediction of the target gene. A more detailed technical overview of TWAS analysis is provided in Section 4.1 of the Supplementary Notes.

We apply INTRIGUE in two experiments to examine the reproducibility between i) two different GWAS datasets; and ii) two sets of weights derived from different tissue eQTL datasets. The details on pre-processing of the eQTL and GWAS data can be found in [24, 22].

In the first experiment, we also compare the INTRIGUE results with the traditional meta-analysis and the IDR analysis (which encounters some computational difficulty). The results are detailed in Section 4.2 of the Supplementary Notes.

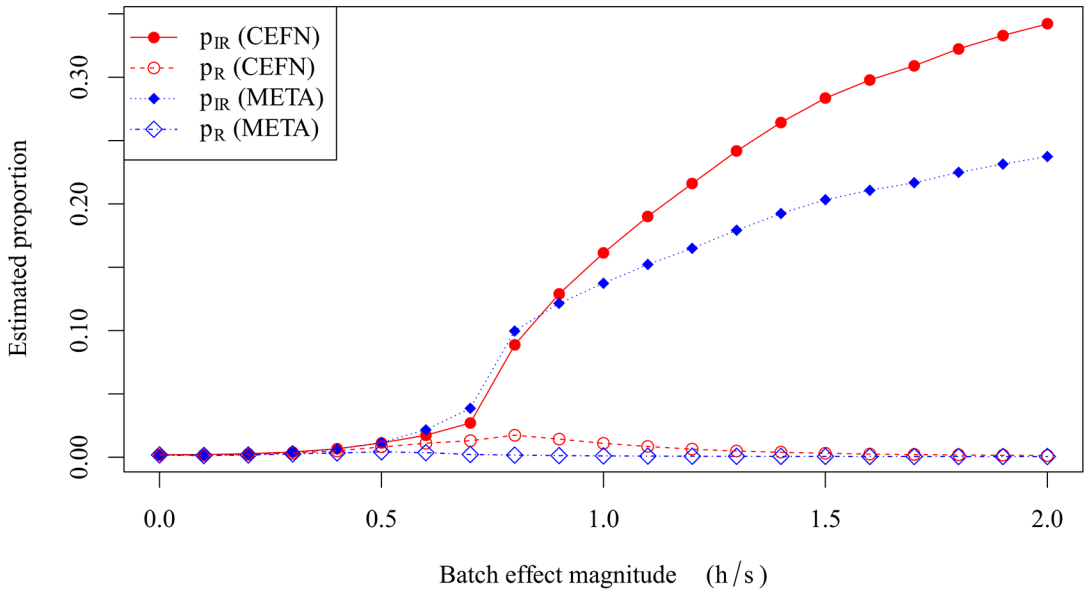
In the second experiment, we use two sets of eQTL weights derived from the whole blood and the skeletal muscle tissues. In addition to the analysis focusing on the eGenes, we also apply INTRIGUE to examine the full set of 32,363 genes. Both the CEFN and the META model yield similar estimates, and the result indicates that among the non-null genes, a majority is irreproducible between tissues, i.e., $\rho_{\mathrm{IR}} > 0.50$ (Table 3). Close inspection suggests that a large proportion of the irreproducible findings are due to tissue-specific gene expressions.

The sample sizes in GTEx skeletal muscle and whole blood eQTL data are close (706 vs. 670) but not identical. We down-sample the skeletal muscle eQTL data to match the exact sample size of whole blood and repeat the TWAS and INTRIGUE analyses. We find the proportion estimates in reproducibility analysis are virtually invariant (Section 4.3.1 of the Supplementary Notes)

Code availability

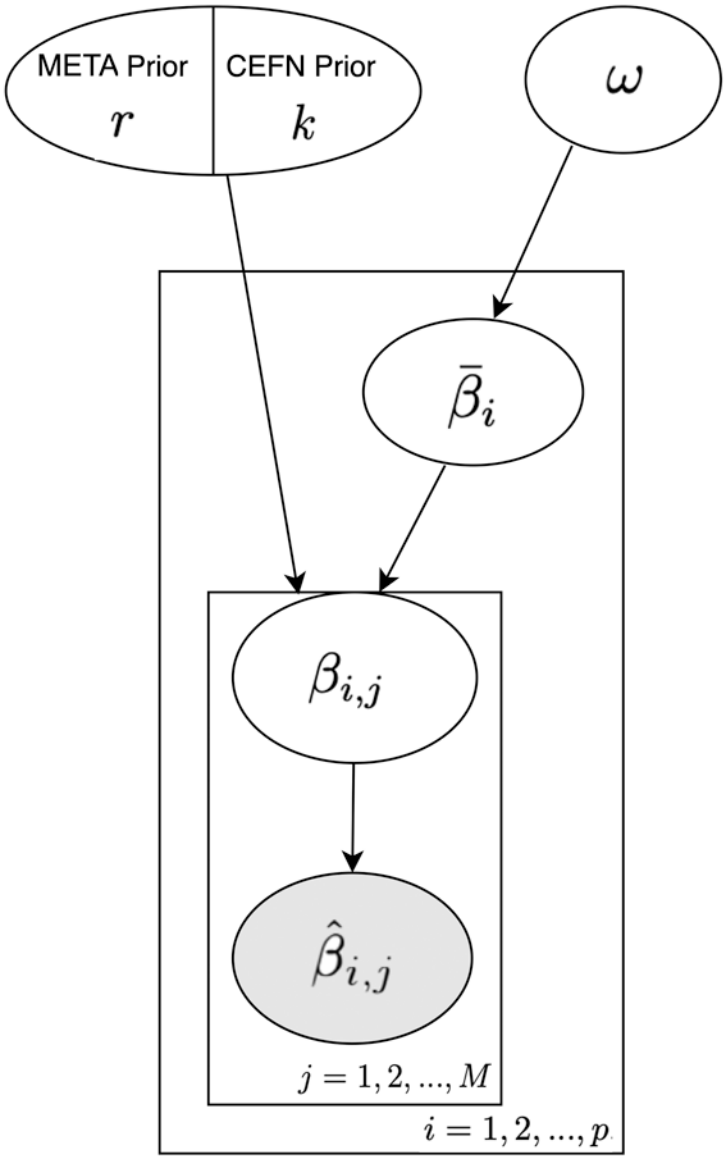
The source code for software implementation (in R and C/C++), simulation studies, and real data processing are provided in https://github.com/ArtemisZhao/INTRIGUE. A docker image that duplicates the complete computational environment for reproducing the reported results can be freely downloaded from https://hub.docker.com/r/xqwen/intrigue.

Extended Data



Extended Data Figure 1: Proportion estimates from batch effect affected high-throughput experiments with no genuine biological signals

Each simulated dataset consists of 1,000 genes. No gene is differentially expressed in the case ($N = 20$) and the control ($N = 20$) samples. In each replication dataset, 500 genes are affected by the unobserved batch effects with various magnitudes (η/σ). The figure shows the estimates of (π_{IR}, π_R) from the CEFN and the META models for all magnitudes of batch effects examined. The reproducible proportions across all datasets remain close to 0, while the estimates of the irreproducible proportions monotonically increases as the batch effects become stronger.



Extended Data Figure 2: A directed acyclic graph representation of the proposed Bayesian hierarchical model

The estimated effects, $\hat{\beta}_{i,j}$'s are observed, $\bar{\beta}_i$'s and $\beta_{i,j}$'s are latent random variables. ω, k (or r) are hyper-parameters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the anonymous referees for their insightful comments and helpful suggestions.

Data availability

- All processed data for simulations and real data analysis are available at https://github.com/ArtemisZhao/INTRIGUE/intrigue_paper.
- GWAS summary statistics for the UK Biobank and the GIANT consortium are available at doi:10.5281/zenodo.3629742.
- eQTL data for TWAS analysis are available at https://gtexportal.org/home/datasets.

References

[1]. Goodman SN, Fanelli D & Ioannidis JP What does research reproducibility mean? Science translational medicine 8, 341ps12–341ps12 (2016).

[2]. Begley CG & Ioannidis JP Reproducibility in science: improving the standard for basic and preclinical research. Circulation research 116, 116–126 (2015). [PubMed: 25552691]

[3]. Leek JT & Peng RD Opinion: Reproducible research can still be wrong: Adopting a prevention approach. Proceedings of the National Academy of Sciences 112, 1645–1646 (2015).

[4]. Leek JT et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11, 733 (2010).

[5]. AC't Hoen P et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. Nature biotechnology 31, 1015 (2013).

[6]. Goh WWB, Wang W & Wong L Why batch effects matter in omics data, and how to avoid them. Trends in biotechnology 35, 498–507 (2017). [PubMed: 28351613]

[7]. Ioannidis JP et al. Repeatability of published microarray gene expression analyses. Nature genetics 41, 149 (2009). [PubMed: 19174838]

[8]. Baggerly KA, Coombes KR et al. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. The Annals of Applied Statistics 3, 1309–1334 (2009).

[9]. Flutre T, Wen X, Pritchard J & Stephens M A statistical framework for joint eqtl analysis in multiple tissues. PLoS genetics 9, e1003486 (2013). [PubMed: 23671422]

[10]. Li G, Shabalin AA, Rusyn I, Wright FA & Nobel AB An empirical bayes approach for multiple tissue eqtl analysis. Biostatistics 19, 391–406 (2017).

[11]. Consortium G et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 (2015). [PubMed: 25954001]

[12]. Goodman SN A comment on replication, p-values and evidence. Statistics in medicine 11, 875–879 (1992). [PubMed: 1604067]

[13]. Heller R, Bogomolov M & Benjamini Y Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. Proceedings of the National Academy of Sciences 111, 16262–16267 (2014).

[14]. Li Q, Brown JB, Huang H, Bickel PJ et al. Measuring reproducibility of high-throughput experiments. The annals of applied statistics 5, 1752–1779 (2011).

[15]. Tukey JW The future of data analysis. The annals of mathematical statistics 33, 1–67 (1962).

[16]. Stephens M False discovery rates: a new deal. Biostatistics 18, 275–294 (2016).

[17]. Efron B et al. Size, power and false discovery rates. The Annals of Statistics 35, 1351–1377 (2007).

[18]. Leek JT & Storey JD Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS genetics 3 (2007).

[19]. Johnson WE, Li C & Rabinovic A Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics 8, 118–127 (2007). [PubMed: 16632515]

[20]. Gamazon ER et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics 47, 1091 (2015). [PubMed: 26258848]

[21]. Gusev A et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 48, 245 (2016). [PubMed: 26854917]

[22]. Zhang Y et al. Investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic twas analysis. Genome Biology (in press).

[23]. Storey JD et al. The positive false discovery rate: a bayesian interpretation and the q-value. The Annals of Statistics 31, 2013–2035 (2003).

[24]. Aguet F et al. The gtex consortium atlas of genetic regulatory effects across human tissues. BioRxiv 787903 (2019).

[25]. Peters JL et al. Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. Journal of the Royal Statistical Society: Series A (Statistics in Society) 173, 575–591 (2010).

[26]. Lin L & Chu H Quantifying publication bias in meta-analysis. Biometrics 74, 785–794 (2018). [PubMed: 29141096]

[27]. Terrin N, Schmid CH, Lau J & Olkin I Adjusting for publication bias in the presence of heterogeneity. Statistics in medicine 22, 2113–2126 (2003). [PubMed: 12820277]

[28]. Augusteijn HE, van Aert R & van Assen MA The effect of publication bias on the q test and assessment of heterogeneity. Psychological methods 24, 116 (2019). [PubMed: 30489099]

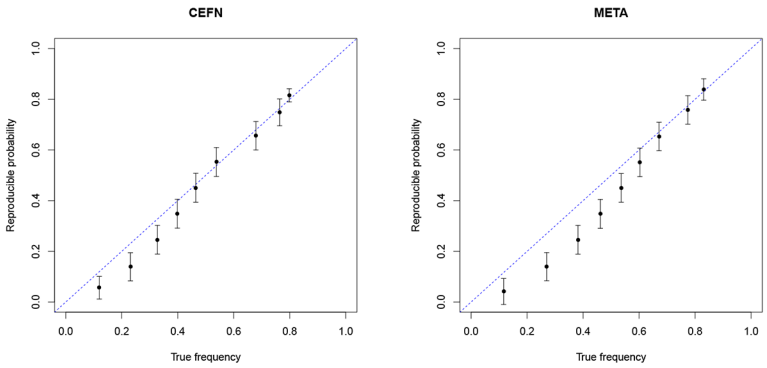
[29]. Lau J, Ioannidis JP, Terrin N, Schmid CH & Olkin I The case of the misleading funnel plot. Bmj 333, 597–600 (2006). [PubMed: 16974018]

[30]. Higgins JP & Thompson SG Quantifying heterogeneity in a meta-analysis. Statistics in medicine 21, 1539–1558 (2002). [PubMed: 12111919]

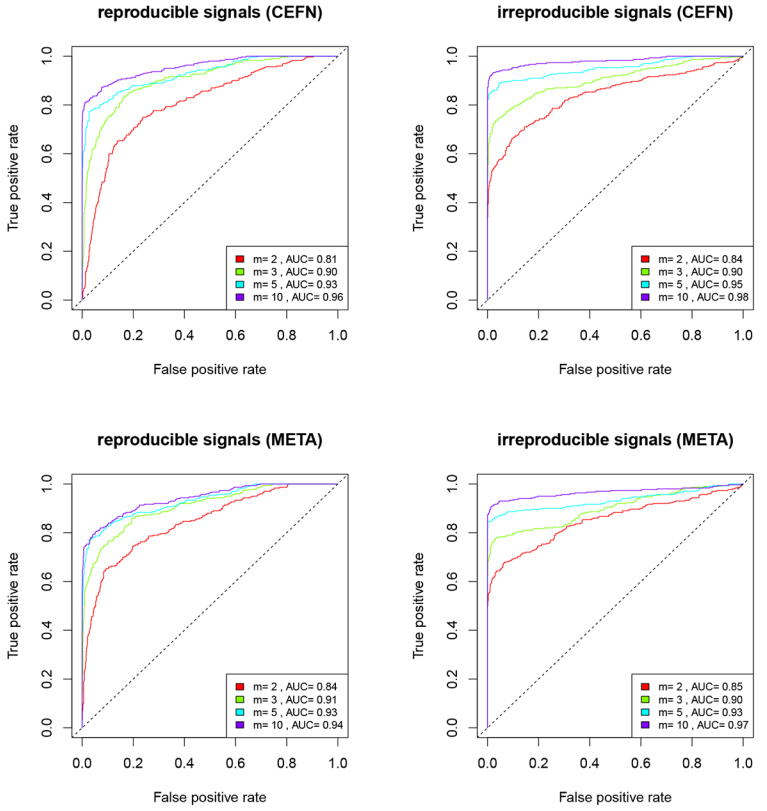
[31]. Schmidt S Shall we really do it again? the powerful concept of replication is neglected in the social sciences. Review of general psychology 13, 90–100 (2009).

[32]. Wen X Bayesian model selection in complex linear systems, as illustrated in genetic association studies. Biometrics 70, 73–83 (2014). [PubMed: 24350677]

[33]. Wen X & Stephens M Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. The annals of applied statistics 8, 176 (2014). [PubMed: 26413181]



A



B

Figure 1: Accuracy and performance of the proposed methods in simulations

Panel **A** represents the calibration accuracy of the estimated reproducible probabilities. The estimated reproducible probabilities are binned into multiple frequency bins. We plot the mean of the estimated probabilities versus the frequency of the actual reproducible signals for each bin (represented by a filled circle at the center of the corresponding error bar). The error bars represent the estimated 95% confidence intervals. The statistics in each plot are computed from 200,000 data points pooled from 200 independent simulations. Considering the small sample sizes used in the simulations, we find the reproducible probabilities are calibrated reasonably well, especially for the modest- to high-value bins. Both models are

conservative in reporting low-frequency values. Panel **B** shows the performance of classification for reproducible and irreproducible signals with different numbers of replication experiments. ROC curves for classifying reproducible signals (left) and irreproducible signals (right) by both the CEFN and the META models are plotted. In all cases, the performance of classifications uniformly improves as the number of replication experiments increases.

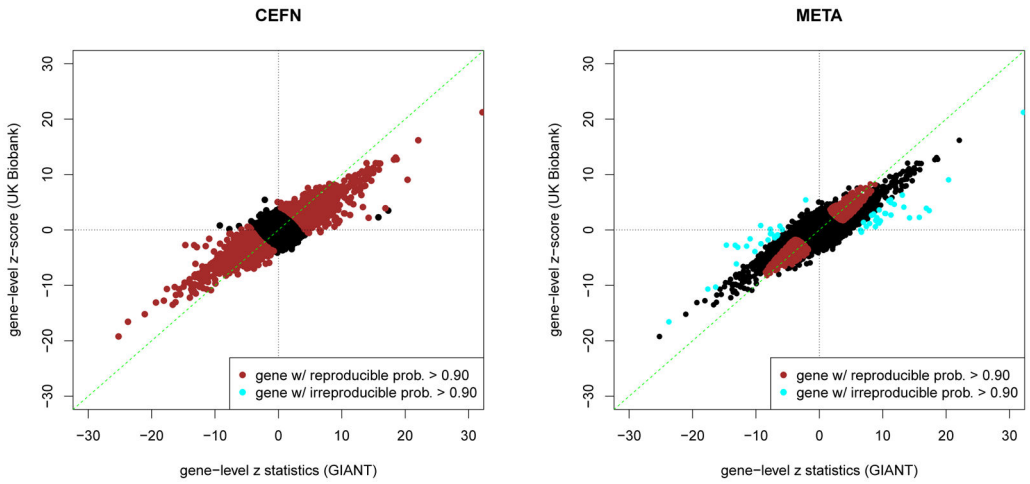


Figure 2: Highly reproducible TWAS signals identified from the height GWAS data in the UK Biobank and the GIANT consortium

The TWAS *z*-scores for all genes are plotted. The highlighted brown data points represent the gene evaluated with the reproducible probability > 0.90. It is clear that the CEFN model tolerates the magnitude difference of the *z*-scores and focuses on the directional consistency. In contrast, the META model emphasizes the consistency in both aspects and assesses only a small subset of genes with high reproducible probabilities.

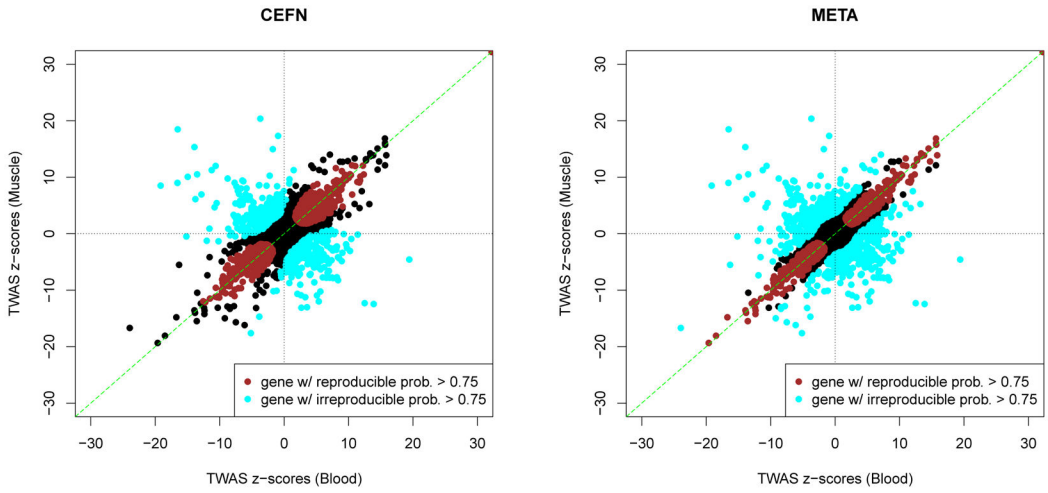


Figure 3: Tissue-consistent and Tissue-specific height TWAS signals identified from whole blood and muscle skeletal tissues

Tissue-consistent (reproducible probability > 0.75) and Tissue-specific (irreproducible probability > 0.75) height TWAS signals are highlighted in brown and cyan, respectively. We observe that the META model identifies genes with the consistent directional effect but the inconsistent magnitude of *z*-scores (i.e., cyan points in the first and third quadrants) as irreproducible signals. The irreproducible genes flagged by the CEFN model are primarily directional inconsistent.

Table 1:

Detecting batch effects in high-throughput experiments with no genuine biological signals

Each simulated dataset consists of 1,000 genes. No gene is differentially expressed in the case ($N = 20$) and the control ($N = 20$) samples. In each replication dataset, 500 genes are affected by the unobserved batch effects with various magnitudes (η/σ). The table shows that the estimates of (π_{IR}, ρ_{IR}) from both the CEFN and the META models are effective in detecting batch effects with modest to large magnitudes. The two-sample two-sided K-S tests, comparing the empirical distributions of *p*-values from the reference and the replication datasets, are slightly more sensitive. However, it is unclear how the test information can be used to guard against false-positive findings. The number of false rejections represents the number of rejections by the *q*-value procedure at 5% FDR level when analyzing the batch contaminated dataset alone. The false-positive findings become a severe problem as the batch effect increases. When both datasets are jointly analyzed, there is no false classification of reproducible signals at 5% FDR level in any setting by either the CEFN model or the META model.

Batch Effect (η/σ)	CEFN		META		K-S test	False Rejections
	$\hat{\pi}_{IR}$	$\hat{\rho}_{IR}$	$\hat{\pi}_{IR}$	$\hat{\rho}_{IR}$	*p*-value	(by *q*-value w/o ref)
0.0	2.0×10^{-3}	0.51	1.4×10^{-3}	0.44	0.219	0
0.2	2.8×10^{-3}	0.57	2.1×10^{-3}	0.54	0.134	0
0.4	6.7×10^{-3}	0.57	3.7×10^{-3}	0.65	0.200	1
0.6	0.017	0.61	0.022	0.86	0.022	2
0.8	0.089	0.84	0.100	0.98	8.67×10^{-4}	15
1.0	0.161	0.94	0.137	0.99	9.08×10^{-5}	45
1.5	0.283	0.99	0.203	0.99	1.79×10^{-6}	152
2.0	0.342	1.00	0.238	1.00	3.95×10^{-11}	252

Table 2:

Identify reproducible signals from batch effect affected high-throughput experiments with genuine biological signals

Each simulated dataset consists of 1,000 genes. 200 genes are differentially expressed in the case ($N = 20$) and the control ($N = 20$) samples. In each replication dataset, 500 genes (100 DE genes and 400 non-DE genes) are affected by the unobserved batch effects with various magnitudes. We apply the FDR control procedures based on the CEFN and the META models to identify the reproducible signals by jointly analyzing the corresponding replication and reference datasets. The table shows the realized false discovery rates and power. For reference, we report the discoveries by applying the q-value procedure on the replication dataset alone. The numbers in the parentheses represent the analysis results of the processed replication datasets by the ComBAT method, a latent batch effect adjustment procedure. All procedures are performed at the target FDR control level of 5%.

	CEFN		META		Stand-alone Analysis	
(η/σ)	FDR	Power	FDR	Power	FDR	Power
0.0	0.046 (0.037)	0.520 (0.520)	0.046 (0.046)	0.520 (0.515)	0.014 (0.015)	0.340 (0.325)
0.2	0.029 (0.029)	0.500 (0.500)	0.029 (0.029)	0.495 (0.495)	0.031 (0.031)	0.315 (0.315)
0.4	0.038 (0.029)	0.500 (0.495)	0.030 (0.030)	0.490 (0.485)	0.145 (0.103)	0.325 (0.315) †
0.6	0.049 (0.039)	0.490 (0.490)	0.024 (0.033)	0.405 (0.440)	0.239 (0.205)	0.350 (0.310) †
0.8	0.044 (0.04l)	0.430 (0.470)	0.000 (0.000)	0.085 (0.190)	0.327 (0.250)	0.380 (0.345) †
1.0	0.000 (0.026)	0.000 (0.375)	0.000 (0.000)	0.020 (0.050)	0.448 (0.345)	0.395 (0.360) †
1.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.005)	0.616 (0.496)	0.440 (0.350) †
2.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.690 (0.569)	0.475 (0.345) †

† The symbol indicates power without proper type I error control. In all simulation scenarios, the proposed methods control the false-positive findings at the target level.

Table 3:

Estimated null, reproducible, and irreproducible proportions in standing height TWAS analyses

The top panel represents the comparison between two GWAS datasets with the same tissue-eQTL data. Both the CEFN and the META models provide qualitatively similar estimates in all three categories, and the estimates indicate high level of consistency between different GWAS data. The bottom panel represents the comparison of TWAS analyses with the same GWAS data (UKB) but different tissue-eQTL data. For both the full gene set and the subset of eGenes, the META model and the CEFN model provide qualitatively similar estimates. There is a noticeable proportion of TWAS signals that are estimated irreproducible, suggesting potential tissue-specific molecular mechanisms leading to the complex trait of interest, height. Note that the genes in the eGene set are necessarily expressed in both tissues. Thus, in the analysis of full set, the irreproducibility of the genes are explained by both differential expressions (i.e., a gene is only expressed in a single tissue) and differential regulations, whereas in the analysis of eGene set, explanations can be narrowed down to the potential differential regulatory mechanisms in the two tissues.

GWAS data	eQTL data	Gene set		$\hat{\pi}_{\text{Null}}$	$\hat{\pi}_{\text{R}}$	$\hat{\pi}_{\text{IR}}$
UKB vs. GIANT	Muscle	Full set	CEFN	0.838	0.161	0.001
			META	0.841	0.141	0.018
UKB	Muscle vs. Blood	eGene set	CEFN	0.296	0.399	0.305
			META	0.406	0.368	0.225
		Full set	CEFN	0.794	0.069	0.136
			META	0.837	0.031	0.131