nature research

https://doi.org/10.1038/s41592-020-00978-4

Supplementary information

Quantify and control reproducibility in high-throughput experiments

In the format provided by the authors and unedited

Supplementary Tables

		π_{Null}	π_{R}	π_{IR}
S1	$N_1 = 400, N_2 = 500$	0.800	0.160	0.040
	CEFN	0.799 (0.787, 0.811)	0.168 (0.154, 0.182)	0.033 (0.024, 0.042)
	META	0.806 (0.794, 0.818)	0.159 (0.146, 0.172)	0.035 (0.027, 0.043)
S2	$N_1 = 80, N_2 = 500$	0.800	0.160	0.040
	CEFN	0.798 (0.785, 0.810)	0.168 (0.153, 0.182)	0.034 (0.024, 0.044)
	META	0.805 (0.795, 0.819)	0.157 (0.142, 0.171)	0.036 (0.026, 0.046)
S3	$N_1 = 300, N_2 = 600$	0.500	0.300	0.200
	CEFN	0.521 (0.504, 0.536)	0.306 (0.281, 0.330)	0.174 (0.151, 0.196)
	META	0.543 (0.528, 0.558)	0.289 (0.268, 0.310)	0.168 (0.150, 0.186)

Supplementary Table 1: **Estimation results of simulated datasets** Each dataset consists of 1,000 pairs of observations and are simulated by three schemes with different combinations of sample sizes and proportions of signal categories using linear regression models. The residual errors of the regression models and the latent grand effects for non-null signals are drawn from the distribution of $\text{N}(0, 1)$. The table shows the averaged point estimate and the corresponding 95% confidence intervals (across 100 simulated datasets) for proportion parameters by both the CEFN and the META models using the EM algorithm. The reproducible proportions (π_{R}) in all three schemes are all well estimated by both prior models. The estimates by the META model are consistently more conservative than the estimates by the CEFN model.

Supplementary Notes for "Quantify and Control Reproducibility in High-throughput Experiment"

Contents

1 Method Details

1.1 DC implication under the META model

Here we derive the probabilistic directional consistency condition from the META model.

Note that

$$\Pr(\beta_{i,j} \text{ and } \bar{\beta}_i \text{ have the same sign}) = \Pr(\beta_{i,j} > 0, \bar{\beta}_i > 0) + \Pr(\beta_{i,j} < 0, \bar{\beta}_i < 0).$$

Under the prior specifications of the META model,

$$\begin{aligned}
\Pr(\beta_{i,j} > 0, \bar{\beta}_i > 0) &= \int_0^{+\infty} \Pr(\beta_{i,j} > 0 | \bar{\beta}_i) p(\bar{\beta}_i) d\bar{\beta}_i \\
&= \int_0^{+\infty} \int_0^{+\infty} f(\beta_{i,j} | \bar{\beta}_i) p(\bar{\beta}_i) d\beta_{i,j} d\bar{\beta}_i \\
&= \int_0^{+\infty} \int_0^{+\infty} \frac{1}{\sqrt{2\pi\phi^2}} e^{-\frac{(\beta_{i,j} - \bar{\beta}_i)^2}{2\phi^2}} \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{\bar{\beta}_i^2}{2\omega^2}} d\beta_{i,j} d\bar{\beta}_i \\
&= \int_0^{+\infty} \Phi\left(\frac{\bar{\beta}_i}{\phi}\right) \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{\bar{\beta}_i^2}{2\omega^2}} d\bar{\beta}_i,
\end{aligned}$$

where $\Phi(\cdot)$ denotes the CDF of the standard normal distribution. Let us define $\bar{z}_i = \bar{\beta}_i \, \omega$. It follows that

$$\Pr(\beta_{i,j} > 0, \bar{z}_i > 0) = \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \Phi(\bar{z}_i \cdot \frac{\omega}{\phi}) e^{-\frac{\bar{z}_i^2}{2}} d\bar{z}_i.$$

Similarly, it can be shown that the same result holds for $\Pr(\beta_{i,j} < 0, \bar{\beta}_i < 0)$. Therefore,

$$\Pr(\beta_{i,j} \text{ and } \bar{\beta}_i \text{ have the same sign}) = 2 \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \Phi(\bar{z}_i \cdot \frac{\omega}{\phi}) e^{-\frac{\bar{z}_i^2}{2}} d\bar{z}_i.$$

Recall that $r = \frac{\phi^2}{\phi^2+\omega^2}$, and $\frac{\omega}{\phi} = \sqrt{\frac{1-r}{r}}$. It follows that

$$\Pr(\beta_{i,j} \text{ and } \bar{\beta}_i \text{ have the same sign}) = \sqrt{\frac{2}{\pi}} \int_0^{+\infty} \Phi(\bar{z}_i \cdot \sqrt{\frac{1-r}{r}}) e^{-\frac{\bar{z}_i^2}{2}} \, d\bar{z}_i,$$

which is a sole function of heterogeneity parameter r.

1.2 Grid specification for partitions of model space

In this section, we detail the scheme to use a dense grid of (k, ω) (or (r, ω)) values to represent the model partitions of non-signals, reproducible signals, and irreproducible signals, respectively.

We always include the singular grid point $\omega = 0$ to represent the null signals.

The irreproducible and reproducible signals are classified based on the levels of heterogeneity conveyed in the parameter values of k and r for the CEFN model and the META model. To this end, we obtain the corresponding k and r values by mapping the probabilities implied from the DC requirement in the respective models (i.e., by applying Eqn (5) for the CEFN model and Eqn (7) for the META model in the Methods section). By default, we set DC probabilities 0.99, 0.975, and 0.95 to represent reproducible signals and DC probabilities 0.75, 0.70, and 0.65 to characterize the irreproducible signals. In our software implementation, we also allow the users to specify different sets of probability values. In practice, we find that there is little practical difference by imposing more dense values of DC probabilities for each category.

We follow the scheme in [1] to set up the grid values for ω, which set up a scaled mixture of the normal distribution to model non-zero effects. Although a single ω value is capable of modeling all possible effects (i.e., the support of a normal distribution covers $(-\infty, \infty)$), a scaled mixture normal distribution has the unique flexibility to characterize the long-tailed effect distributions that are commonly observed in high-throughput experiments.

We specify a range of values for $\zeta^2 := (k^2+1)\omega^2$. For the ith experimental unit, we compute a summary statistic

$$T_i = \frac{1}{M}\sum_j \left(\hat{\beta}_{i,j}^2 + \mathrm{se}(\hat{\beta}_{i,j})^2\right).$$

Recall that M represents the number of replication experiments and note that T_i is a natural estimate of the overall effect $(k^2+1)\omega^2$ when M is large. We set $\zeta^2_{\min}$ and $\zeta^2_{\max}$ as the mean and the 99% quantile of the T values, respectively. We then create a grid of ζ^2 values ranges from $\zeta^2_{\min}$ to $\zeta^2_{\max}$, and $\zeta^2_{i+1}/\zeta^2_i = 2$ For each ζ_l value in the grid, we complete the combination of (k_p, ω_o) for each pre-defined k_p value by setting

$$\omega_p = \sqrt{\frac{\zeta_l^2}{k_p^2+1}}.$$

The resulting grid covers a wide range of overall effects (ζ^2) for non-null signals informed by the observed data. Within each overall effect, both reproducible and irreproducible scenarios are well represented.

1.3 Bayes factor calculation for the CEFN model

Under the CEFN model, let BF_i denote the Bayes factor of the i-th experimental unit with pre-defined hyper-parameter value (ω, k), i.e.,

$$\begin{aligned}\mathrm{BF}_i &= \frac{P\left(\hat{\beta}_{i,1},\ldots,\hat{\beta}_{i,M} \mid \omega, k\right)}{P\left(\hat{\beta}_{i,1},\ldots,\hat{\beta}_{i,M} \mid \omega \equiv 0\right)} \\ &= P\left(\hat{\beta}_{i,1},\ldots,\hat{\beta}_{i,M} \mid \omega, k\right) \Big/ \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{\hat{\beta}_{i,j}^2}{2\sigma_{i,j}^2}}\end{aligned}$$

Hence, the key is to compute the numerator of the above expression. Note that

$$
\begin{aligned}
p(\bar{\beta}_i) &= \frac{1}{\sqrt{2\pi\omega^2}} e^{-\frac{\bar{\beta}_i^2}{2\omega^2}} \\
p(\beta_{i,j}|\bar{\beta}_i) &= \frac{1}{\sqrt{2\pi k^2 \bar{\beta}_i^2}} e^{-\frac{(\beta_{i,j}-\bar{\beta}_i)^2}{2k^2\bar{\beta}_i^2}}, \ j = 1, 2, \ldots, M, \\
p(\hat{\beta}_{i,j}|\beta_{i,j}) &= \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{(\hat{\beta}_{i,j}-\beta_{i,j})^2}{2\sigma_{i,j}^2}}, \ j = 1, 2, \ldots, M.
\end{aligned}
$$

Given the conditional independence relationships implied from the proposed hierarchical model, it follows that

$$
p\left(\hat{\beta}_{i,1}, \ldots, \hat{\beta}_{i,M}, \beta_{i,1}, ..., \beta_{i,M}, \bar{\beta}_i \mid k, \omega\right) = \prod_j p(\hat{\beta}_{i,j}|\beta_{i,j}) \cdot \prod_j p(\beta_{i,j}|\bar{\beta}_i) \cdot p(\bar{\beta}_i).
$$

Subsequently, the marginal probability $P\left(\hat{\beta}_{i,1}, \ldots, \hat{\beta}_{i,M} \mid \omega, k\right)$ can be evaluated by integrating out $\beta_{i,1}, \ldots, \beta_{i,M}$ and $\bar{\beta}_i$ sequentially, i.e.,

$$
\begin{aligned}
& P\left(\hat{\beta}_{i,1}, \ldots, \hat{\beta}_{i,M} \mid \omega, k\right) \\
&= \int \cdots \int \prod_{j=1}^{M} p(\hat{\beta}_{i,j}|\beta_{i,j}) \cdot \prod_{j=1}^{M} p(\beta_{i,j}|\bar{\beta}_i) \cdot p(\bar{\beta}_i) \, d\beta_{i,1} \cdots d\beta_{i,M} \, d\bar{\beta}_i \\
&= \int_{-\infty}^{+\infty} \left(\prod_{j=1}^{M} \frac{1}{\sqrt{2\pi(\bar{\beta}_i^2 k^2 + \sigma_{i,j}^2)}} e^{-\frac{(\hat{\beta}_{i,j}-\bar{\beta}_i)^2}{2(\bar{\beta}_i^2 k^2 + \sigma_{i,j}^2)}} \right) \cdot \frac{1}{\sqrt{2\pi\omega^2}} \, e^{-\frac{\bar{\beta}_i^2}{2\omega^2}} \, d\bar{\beta}_i.
\end{aligned}
$$

Thus, the numerator of the Bayes factor can be evaluated by a one-dimensional numerical integration. In our software implementation, we apply the QAGI algorithm to carry out the numerical evaluation.

1.3.1 Small sample corrections for Bayes factor calculation

Consider a linear regression with a relatively small sample size, such that the effect size estimate and its standard error does follow the asymptotic normal distribution under the null. Instead, $\hat{\beta}_{i,j}/\sigma_{i,j}$ follows a t-distribution with $n-2$ degree of freedom. In this case, directly computing the Bayes factors by either the CEFN or the META model becomes inappropriate.

We consider a simple transformation based on [2]. Let $p_{i,j}$ denote the p-value under the correct t-distribution. We define

$$z_{i,j} = \text{sgn}(\hat{\beta}_{i,j}) \left| \Phi^{-1}\left(p_{i,j}/2\right) \right|. \tag{1}$$

Note that the transformed $z_{i,j}$ follows the standard normal distribution under the null. We can further define a corrected standard error due to small sample size, i.e.,

$$\tilde{\sigma}_{i,j} = \frac{\hat{\beta}_{i,j}}{z_{i,j}}. \tag{2}$$

Note that if the original data satisfy the normality assumption under the null, $\tilde{\sigma}_{i,j} = \sigma_{i,j}$. Instead of using original data $(\beta_{i,j}, \sigma_{i,j})$, we use the standard error-corrected version, $(\beta_{i,j}, \tilde{\sigma}_{i,j})$, for the Bayes factor computation.

Finally, we note that the transformation (1) is generally applied to any valid p-values derived from arbitrary null distributions.

1.4 Estimating $\boldsymbol{\pi}$ via EM algorithm

This section details the EM algorithm for estimating the hyper-parameter $\boldsymbol{\pi} = (\pi_0, ..., \pi_{L-1})$.

Regarding $\boldsymbol{\gamma}_i$'s as missing data, the complete data likelihood can be written as,

$$\begin{aligned} L(\boldsymbol{\pi}) &= \prod_{i=1}^{p} P\left(\hat{\beta}_{i,1}, ..., \hat{\beta}_{i,M}, \boldsymbol{\gamma}_i \mid \boldsymbol{\pi}\right) \\ &= \prod_{i=1}^{p} P\left(\hat{\beta}_{i,1}, ..., \hat{\beta}_{i,M} \mid \boldsymbol{\gamma}_i\right) \Pr(\boldsymbol{\gamma}_i|\boldsymbol{\pi}) \\ &= \prod_{i=1}^{p} \prod_{l=0}^{L-1} \left(\pi_l \, P(\hat{\beta}_{i,j} \mid k_l, \omega_l)\right)^{\mathbb{1}_{\{\gamma_{i,l}=1\}}}. \end{aligned}$$

Thus, the corresponding complete data log-likelihood function is given by

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{p} \sum_{l=0}^{L-1} (\log(\mathrm{BF}_{i,l}) + \log \pi_l) \cdot \mathbb{1}_{\{\gamma_{i,l}=1\}},$$

where $\mathrm{BF}_{i,l}$ denote the Bayes factor computed for the ith experimental unit using the lth grid value, (k_l, ω_l).

E-step In the $(t+1)$th iteration of the EM algorithm, let $\boldsymbol{\pi}^{(t)}$ denote the current estimate of $\boldsymbol{\pi}$. In the E-step, we compute

$$\begin{aligned} \mathrm{E}\left(\mathbb{1}_{\{\gamma_{i,l}=1\}} \mid \hat{\beta}_{i,1}..., \hat{\beta}_{i,M}, \boldsymbol{\pi}^{(t)}\right) &= \Pr\left(\gamma_{i,l} = 1 \mid \hat{\beta}_{i,1}..., \hat{\beta}_{i,M}, \boldsymbol{\pi}^{(t)}\right), \\ &= \frac{\pi_l^{(t)} \cdot \mathrm{BF}_{i,l}}{\sum_{l'=0}^{L-1} \pi_{l'}^{(t)} \cdot \mathrm{BF}_{i,l'}}. \end{aligned}$$

M-step Subsequently in the M-step in the $(t+1)$th iteration of the EM algorithm, we update the current estimate of $\boldsymbol{\pi}$ by finding

$$\boldsymbol{\pi}^{(t+1)} = \arg\max_{\boldsymbol{\pi}} \left(\sum_{i=1}^{p} \sum_{l=0}^{L-1} \log \pi_l \cdot \frac{\pi_l^{(t)} \cdot \mathrm{BF}_{i,l}}{\sum_{l'=0}^{L-1} \pi_{l'}^{(t)} \cdot \mathrm{BF}_{i,l'}} \right), \text{ subject to } \sum_l \pi_l = 1.$$

Let $W_{i,l}^{(t)} := \frac{\pi_l^{(t)} \cdot \mathrm{BF}_{i,l}}{\sum_{l'=0}^{L-1} \pi_{l'}^{(t)} \cdot \mathrm{BF}_{i,l'}}$. We maximize the following expression with respect to $\boldsymbol{\pi}$ using the method of Lagrange multipliers, i.e.,

$$\sum_{i=1}^{p} \sum_{l=0}^{L-1} \log \pi_l \cdot W_{i,l}^{(t)} - \lambda (\sum_{l=0}^{L-1} \pi_l - 1).$$

Setting the derivatives with respect to π_l for $l = 0, ...L-1$ and λ to zero, we obtain

$$\frac{\partial \left(\sum_{i=1}^{p} \sum_{l=0}^{L-1} \log \pi_l \cdot W_{i,l}^{(t)} - \lambda (\sum_{l=0}^{L-1} \pi_l - 1) \right)}{\partial \pi_l} = \frac{\sum_{i=1}^{p} W_{i,l}}{\pi_l} - \lambda = 0.$$

It follows that

$$\pi_l = \frac{\sum_{i=1}^{p} W_{i,l}^{(t)}}{\lambda}, \quad l = 0, ..., L-1.$$
$$\lambda = \sum_{l} \sum_{i} W_{i,l} = p$$

Thus, the M-step in the $(t+1)$th iteration of the EM algorithm yields

$$\pi_l^{(t+1)} = \frac{1}{p} \sum_{i=1}^{p} \frac{\pi_l^{(t)} \cdot \mathrm{BF}_{i,l}}{\sum_{l'=0}^{L-1} \pi_{l'}^{(t)} \cdot \mathrm{BF}_{i,l'}}, \quad l = 0, \ldots, L-1.$$

1.5 Directional consistency of observed effects under the CEFN model

The directional consistency of the latent effects has an impact on the observed effects, which, in our view, are contaminated by additional noise characterized by $\sigma_{i,j}$'s. In the extreme case that $\sigma_{i,j} \to 0$, the DC of the observed effects should reflect the DC of the true latent effects perfectly. In this section, we derive the DC property of the observed effects from the interplays of the noise level and the DC of the true latent effects.

Conditional on $\bar{\beta}_i$, it follows that

$$\Pr\left(\hat{\beta}_{i,j} \leq x \mid \bar{\beta}_i\right) = \Phi\left(\frac{x - \bar{\beta}_i}{\sqrt{k^2\bar{\beta}_i^2 + \sigma_{i,j}^2}}\right), \quad \forall j.$$

Thus,

$$\begin{aligned}
\Pr\left(\hat{\beta}_{i,j} \leq 0 \mid \bar{\beta}_i\right) &= \Phi\left(-\frac{\bar{\beta}_i}{\sqrt{k^2\bar{\beta}_i^2 + \sigma_{i,j}^2}}\right) \\
&= \Phi\left(-\frac{\operatorname{sgn}(\bar{\beta}_i) \cdot |\bar{\beta}_i|}{\sqrt{k^2\bar{\beta}_i^2 + \sigma_{i,j}^2}}\right) \\
&= \Phi\left(-\frac{\operatorname{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,j}^2/\bar{\beta}_i^2}}\right) \\
&= \Pr\left(\hat{\beta}_{i,j} \leq 0 \mid \bar{\beta}_i^2, \operatorname{sgn}(\bar{\beta}_i)\right)
\end{aligned}$$

Provided that $\hat{\beta}_{i,j}$'s are conditionally independent, it follows that

$$\begin{aligned}
\Pr\left(\hat{\beta}_{i,1} < 0, \cdots, \hat{\beta}_{i,M} < 0 \,|\, \bar{\beta}_i\right) &= \prod_j \Pr\left(\hat{\beta}_{i,j} < 0 \,|\, \bar{\beta}_i\right) \\
&= \prod_j \Phi\left(-\frac{\operatorname{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,j}^2/\bar{\beta}_i^2}}\right).
\end{aligned}$$

Likewise,

$$\begin{aligned}
\Pr\left(\hat{\beta}_{i,1} > 0, \cdots, \hat{\beta}_{i,M} > 0 \mid \bar{\beta}_i\right) &= \prod_j \left[1 - \Phi\left(-\frac{\operatorname{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,1}^2/\bar{\beta}_i^2}}\right)\right] \\
&= \prod_j \Phi\left(\frac{\operatorname{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,1}^2/\bar{\beta}_i^2}}\right)
\end{aligned}$$

The last equation is because $\Phi(-x) + \Phi(x) = 1, \forall x$.

Put all together, we have

$$\begin{aligned}
&\Pr\Big(\ \hat{\beta}_{i,j}\text{'s all have the same sign } \mid \bar{\beta}_i\Big) \\
=&\Pr\Big(\hat{\beta}'_{i,j}s \text{ all have the same sign } \mid \bar{\beta}_i^2, \mathrm{sgn}(\bar{\beta}_i)\Big) \\
=&\prod_j \Phi\left(-\frac{\mathrm{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,j}^2/\bar{\beta}_i^2}}\right) + \prod_j \Phi\left(\frac{\mathrm{sgn}(\bar{\beta}_i)}{\sqrt{k^2 + \sigma_{i,j}^2/\bar{\beta}_i^2}}\right).
\end{aligned}$$

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

$$\Pr\big(\mathrm{sgn}(\bar{\beta}_i) = +1\big) = \Pr\big(\mathrm{sgn}(\bar{\beta}_i) = -1\big) = \frac{1}{2}.$$

We obtain that

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

1.6 Generalization of the CEFN model

In this section, we generalize the CEFN model by assuming an arbitrary distribution for $\bar{\beta}_i$ with a density function $g(x)$. That is, we no longer require $\bar{\beta}_i$ follows a normal distribution centered at 0. However, the CEFN assumption remains intact, such that

$$\beta_{i,j} \mid \bar{\beta}_i \sim N(\bar{\beta}_i\,,\, k^2\bar{\beta}_i^2).$$

Note that,

$$\Pr(\beta_{i,j} < x \,|\, \bar{\beta}_i) = \Phi\left(\frac{x - \bar{\beta}_i}{k|\bar{\beta}_i|}\right),$$

and

$$\begin{aligned}
\Pr(\beta_{i,j} < 0 \,|\, \bar{\beta}_i) &= \Phi\left(-\frac{\mathrm{sgn}(\bar{\beta}_i)}{k}\right), \\
\Pr(\beta_{i,j} > 0 \,|\, \bar{\beta}_i) &= 1 - \Phi\left(-\frac{\mathrm{sgn}(\bar{\beta}_i)}{k}\right) = \Phi\left(\frac{\mathrm{sgn}(\bar{\beta}_i)}{k}\right),
\end{aligned}$$

It follows that

$$\begin{aligned}
&\Pr\left(\beta_{i,j} \text{ and } \bar{\beta}_i \text{ have the same sign }\right) \\
&= \Pr(\beta_{i,j} < 0, \bar{\beta}_i < 0) + \Pr(\beta_{i,j} > 0, \bar{\beta}_i > 0) \\
&= \int_{\bar{\beta}_i<0} \Pr(\beta_{i,j} < 0 \,|\, \bar{\beta}_i) g(\bar{\beta}_i)\, d\bar{\beta}_i + \int_{\bar{\beta}_i>0} \Pr(\beta_{i,j} > 0 \,|\, \bar{\beta}_i) g(\bar{\beta}_i)\, d\bar{\beta}_i \\
&= \Phi\left(\frac{1}{k}\right) \int_{\bar{\beta}_i<0} g(\bar{\beta}_i) d\bar{\beta}_i + \left[1 - \Phi\left(-\frac{1}{k}\right)\right] \int_{\bar{\beta}_i>0} g(\bar{\beta}_i) d\bar{\beta}_i \\
&= \Phi\left(\frac{1}{k}\right)
\end{aligned}$$

In summary, we have proved that the DC implication by the CEFN prior is generalizable for arbitrary distributions of $\bar{\beta}_i$.

2 Details of simulation I (QTL mapping)

2.1 Simulation schemes

Our first set of simulations mimic genome-wide expression quantitative trait loci (eQTL) mapping studies. Here, we provide the details for the corresponding data generative schemes.

Each simulated experiment dataset consists of genotype and expression levels of 1,000 genes. For each gene, the genotype for each individual is coded by 0 (wild type) or 1 (mutant). Let $\boldsymbol{y}_{i,j}$ and $\boldsymbol{g}_{i,j}$ denote the gene expression and genotype vectors for gene i in the j-th experiment. The normalized gene expressions are simulated by

$$\boldsymbol{y}_{i,j} = \mu \mathbf{1} + \beta_{i,j}\, \boldsymbol{g}_{i,j} + \boldsymbol{e}_{i,j}, \quad \boldsymbol{e}_{i,j} \sim \mathrm{N}(0, \sigma^2\, \boldsymbol{I}) \tag{3}$$

Without loss of generality, we set $\mu = 0$, which is equivalent to pre-centering the $\boldsymbol{y}_{i,j}$. We also set $\sigma^2 = 1$ throughout.

For genes corresponding to null signals, we set $\beta_{i,j} = 0$ for all i and j. The effect sizes for non-null signals are generated by

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

We set $k = 0.31$ and 3.00 for the reproducible and the irreproducible signals, respectively. They correspond to the DC probabilities 0.99 and 0.63. Finally, the observed effect $\hat{\beta}_{i,j}$ and $\mathrm{se}(\hat{\beta}_{i,j})$ are obtained by fitting a linear regression model based on observed $\boldsymbol{y}_{i,j}$ and $\boldsymbol{g}_{i,j}$ and used as input for INTRIGUE.

We design 5 different simulation schemes for various tasks by altering the proportions of different signal types, the sample size in each replication experiment, and the number of replications. The summary of the five schemes is as follows.

- S1: replications: $M = 2$; sample size: $N_1 = 400, N_2 = 500$; $(\pi_{\mathrm{Null}}, \pi_{\mathrm{R}}, \pi_{\mathrm{IR}}) = (0.80, 0.16, 0.04)$
- S2: replications: $M = 2$; sample size: $N_1 = 80, N_2 = 500$; $(\pi_{\mathrm{Null}}, \pi_{\mathrm{R}}, \pi_{\mathrm{IR}}) = (0.80, 0.16, 0.04)$
- S3: replications: $M = 2$; sample size: $N_1 = 300, N_2 = 600$; $(\pi_{\mathrm{Null}}, \pi_{\mathrm{R}}, \pi_{\mathrm{IR}}) = (0.50, 0.30, 0.20)$

- S4: replications: $M = 2$; sample size: $N_1 = 40, N_2 = 50$; $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.40, 0.30, 0.30)$
- S5: replications: $M = 2, 3, 5, 10$; sample size: $N = 50$ (equal sample size in all replication experiments); $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.40, 0.30, 0.30)$

2.2 IDR results

For comparison, we also apply the IDR method [3] implemented in the R package *idr* on the datasets simulated from S1, S2, and S3. Our main purpose is to compare $\hat{\pi}_{\text{R}}$ and the IDR-estimated rank-consistent proportions. To apply IDR, we use $|z_{i,j}| = |\hat{\beta}_{i,j}|/\text{se}(\hat{\beta}_{i,j})$ as its input. (The absolute z-scores are monotonic to corresponding two-sided p-values but satisfies the IDR's requirement of higher ranking for more significant findings.)

The results from IDR are summarized in Table S1. Overall, we find that the reproducible proportions are reasonably close to the truth, as well as the INTRIGUE estimates. As expected, the confidence intervals are generally larger than the INTRIGUE estimates for $\hat{\pi}_{\text{R}}$, because IDR only uses rank information from the data.

		π_{Null}	π_{R}	π_{IR}
S1	$N_1 = 400, N_2 = 500$	0.800	0.160	0.040
	IDR $\pi_{\text{consistent}}$		0.171 (0.147, 0.194)	
S2	$N_1 = 80, N_2 = 500$	0.800	0.160	0.040
	IDR $\pi_{\text{consistent}}$		0.138 (0.103, 0.173)	
S3	$N_1 = 300, N_2 = 600$	0.500	0.300	0.200
	IDR $\pi_{\text{consistent}}$		0.375 (0.336, 0.413)	

Table S1: Estimation results of simulated datasets in the IDR model. Each dataset consists of 1000 pairs of observations and is simulated by three schemes (S1, S2, and S3) with different combinations of sample sizes and proportions of signal categories presented in the main text. The estimations and the standard deviations are transformed to z-scores before throwing to the IDR model. The table shows the means and confidence intervals of the estimated proportion parameter for reproducible signal (across 100 simulated datasets) by IDR.

During the analysis, we find that IDR results are often sensitive to the pre-defined starting parameter values. The reported results in Table S1 are based on the best case scenarios by

setting initial values of the hyper-parameters close to the truth. Particularly, we set the starting value for the reproducible bivariate normal distribution mean μ_1 equals to the mean of the z-scores for reproducible signals. We utilize the same strategy for setting up the initial values for the variance σ_1^2 and correlation ρ_1 for the bivariate normal distribution and the proportion of the reproducible signal π_1.

2.3 Additional schemes and results

We also perform additional simulations to examine the performance of the proposed methods. The simulation schemes remain similar to what we described in the previous section, but we alter Eqn. (4) to generate heterogeneous effects from the META model, i.e.,

$$\begin{aligned} \beta_{i,j} \mid \bar{\beta}_i, r &\sim \mathrm{N}\left(\bar{\beta}_i\,,\, \frac{r\,\omega^2}{1-r}\right), \quad j = 1, 2, \ldots, M \\ \bar{\beta}_i &\sim \mathrm{N}(0, 1) \end{aligned} \tag{5}$$

where the r values correspond to the same DC probabilities in the original schemes. We also apply the similar simulation schemes, S1 to S5, to vary the proportions of types of signals and sample sizes in each replication. Specifically,

- SS1: replications: $M = 2$; sample size: $N_1 = 400, N_2 = 500$; $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.80, 0.16, 0.04)$
- SS2: replications: $M = 2$; sample size: $N_1 = 80, N_2 = 500$; $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.80, 0.16, 0.04)$
- SS3: replications: $M = 2$; sample size: $N_1 = 300, N_2 = 600$; $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.50, 0.30, 0.20)$
- SS4: replications: $M = 2$; sample size: $N_1 = 40, N_2 = 50$; $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.40, 0.30, 0.30)$
- SS5: replications: $M = 2, 3, 5, 10$; sample size: $N = 50$ (equal sample size in all replication experiments); $(\pi_{\text{Null}}, \pi_{\text{R}}, \pi_{\text{IR}}) = (0.40, 0.30, 0.30)$

We use these schemes to evaluate the accuracy of the estimation (Table S2), the calibration of the individual reproducibility quantification (Figure S1) and the power to identify reproducible and irreproducible signals (Figure S2).

Overall, the main characteristics of the results are virtually identical to what we report in the main text despite the difference in the data generative algorithms.

		π_{Null}	π_{R}	π_{IR}
SS1	$N_1 = 400, N_2 = 500$	0.800	0.160	0.040
	CEFN	0.794 (0.783, 0.806)	0.170 (0.156, 0.184)	0.036 (0.027, 0.044)
	META	0.800 (0.789, 0.811)	0.161 (0.149, 0.174)	0.039 (0.033, 0.044)
SS2	$N_1 = 80, N_2 = 500$	0.800	0.160	0.040
	CEFN	0.790 (0.778, 0.802)	0.170 (0.154, 0.185)	0.040 (0.029, 0.052)
	META	0.800 (0.788, 0.811)	0.161 (0.146, 0.175)	0.040 (0.032, 0.048)
SS3	$N_1 = 300, N_2 = 600$	0.500	0.300	0.200
	CEFN	0.492 (0.479, 0.505)	0.322 (0.298, 0.347)	0.186 (0.163, 0.208)
	META	0.518 (0.505, 0.531)	0.288 (0.270, 0.305)	0.194 (0.183, 0.206)

Table S2: Estimation results using additional simulated datasets. Each dataset consists of 5000 pairs of observations and is simulated by three schemes (SS1, SS2, and SS3) with different combinations of signal-to-noise ratios and proportions of signal categories. The table shows the mean and the standard deviation of the estimated proportion parameters (across 100 simulated datasets) by both the CEFN and the META models using the EM algorithm.

3 Details of Simulation II (batch effect detection)

3.1 Simulation schemes

This set of simulations is designed to evaluate the INTRIGUE methods' ability to detect the unobserved batch effects in high-throughput experiments. Our scheme closely follows the simulation settings used in [4, 5], Particularly, we consider a differential gene expression experiment of 1,000 candidate genes for 20 cases and 20 controls. The case-control status for all samples is coded in the 40-vector $\boldsymbol{x}$. For simplicity, we assume that all experiments are conducted with 2 batch groups. In the first batch group, the samples are predominately cases, while the second

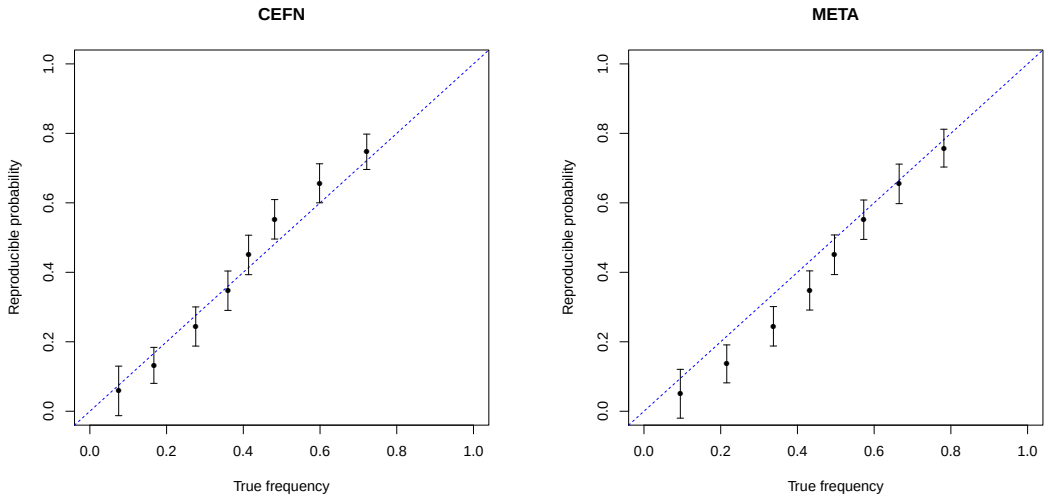

Figure S1: **Calibration of estimated reproducible probabilities** The estimated reproducible probabilities based on simulation scheme SS4 are binned into multiple frequency bins. We plot the mean of the estimated probabilities versus the frequency of the actual reproducible signals for each bin (represented by a filled circle at the center of the corresponding error bar). The error bars represent the estimated 95% confidence intervals. The statistics for each plot are computed from 200,000 independent data points pooled across 200 simulations.

batch consists of predominately control samples. The batch labels for all samples are coded in a 40-vector denoted by $\boldsymbol{\gamma}$. We use β_i to denote the genuine treatment effect on gene i and $\delta_{i,j}$ to denote the replication-specific batch effect on gene i in replication j. The normalized expression levels for gene i in the jth experiment is generated by

$$\boldsymbol{y}_{i,j} = \mu_i \mathbf{1} + \beta_i \boldsymbol{x}_j + \delta_{i,j} \boldsymbol{\gamma}_j + \boldsymbol{e}_{i,j}, \;\; \boldsymbol{e}_{i,j} \sim \mathrm{N}(0, \sigma^2 \boldsymbol{I}), \;\; j = 1, 2. \tag{6}$$

We use $j = 1$ to denote the gold-standard reference dataset which is not affected by batch effects. The replication dataset that is exposed to batch effects is labeled by $j = 2$.

Without loss generality, we set $\mu_i = 0$ and $\sigma^2 = 1$ throughout.

In the first simulation setting, β_i is set to 0 for all genes. In the second simulation setting, β_i is independently drawn from $\mathrm{N}(0, 1)$ for 200 genes, and the remaining 800 genes have $\beta_i = 0$.

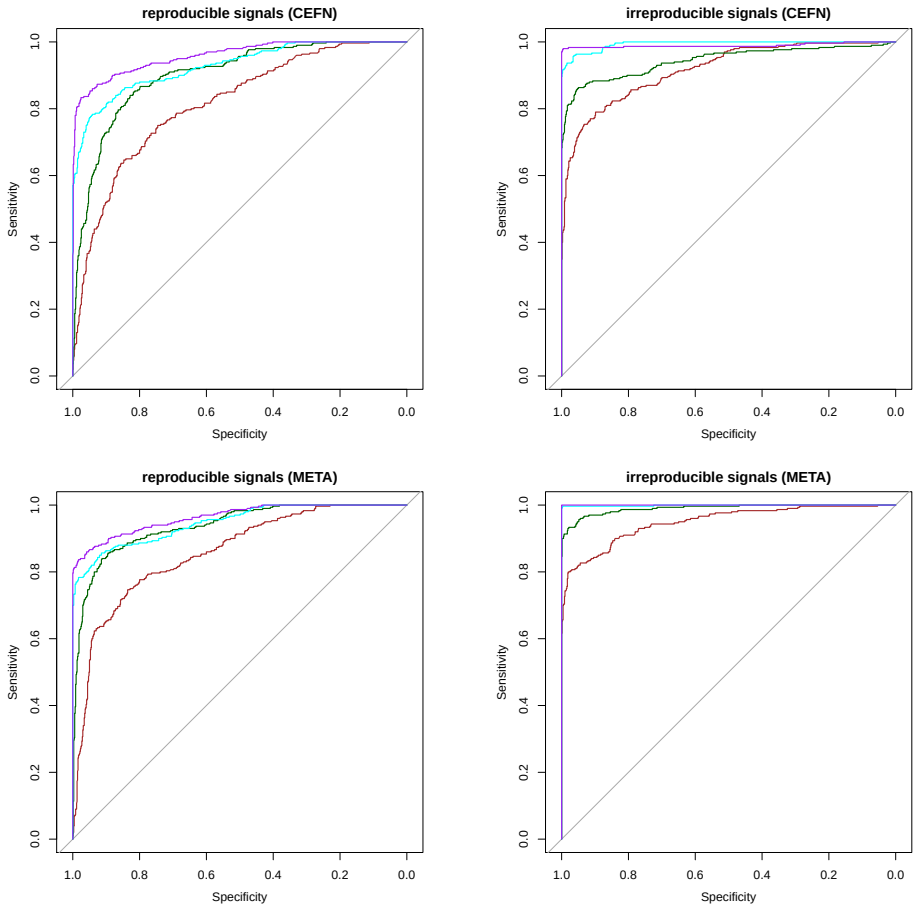

Figure S2: **Classification performance of reproducible and irreproducible signals with different numbers of replication experiments** ROC curves for classifying reproducible signals (left panel) and irreproducible signals (right panel) by both the CEFN and the META models are plotted. The simulated data are generated from scheme SS5. The performance of classifications uniformly improves as the number of replication experiments increases in all settings.

The batch effects are set to 0 for all genes in the reference experiments in all cases, i.e., $\delta_{i,1} \equiv 0$ for all genes. For the replication experiment, $\delta_{i,2}$ is independently drawn from the Normal distribution, $\mathrm{N}(0, \eta^2)$ for 500 (or 50% of) genes, and we set $\delta_{i,2} = 0$ for the remaining 500 genes. In the second simulation setting, 100 genes showing genuine treatment effects and 400 genes without showing treatment effects are chosen to be impacted by the batch effect. For different simulated datasets, we vary η from 0 to 2.

The input for the INTRIGUE methods is obtained by regressing $\boldsymbol{y}_{i,j}$ on $\boldsymbol{x}_j$. Note that the

batch labels are assumed to be unobserved; hence the batch effects are not explicitly controlled. The resulting least square estimates of the treatment effects and their corresponding standard errors are provided to both the CEFN and the META models. The derived p-values from the corresponding t-statistics are used for the Kolmogorov-Smirnov test.

3.2 IDR results of batch effect detection

We apply the IDR method using the absolute t-statistics computed from $\hat{\beta}_{i,j}$'s and the corresponding se($\hat{\beta}_{i,j}$)'s in the first simulation setting without genuine treatment effects.

Although the IDR method does not explicitly estimate the proportion of irreproducible signals, we hope to see the estimated proportions of reproducible signals are close to the INTRIGUE estimate. We find this indeed is the case when the initial values for the IDR EM algorithm are chosen close to the truth (Table S3). However, we again observe that the convergence of the IDR EM algorithm is highly sensitive to the starting values in this simulation setting, indicating that the likelihood space informed by rankings in this context is multi-modal. Specifically, we run the IDR EM algorithm from a set of initial values for $\pi_{\text{consistent}}$, namely, $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.4\}$. We find that the estimates obtained from different initial values typically differ. Moreover, the estimates associated with the overall highest likelihood values often overestimate the truth. Sometimes, the level of overestimation is severe (Table S3).

3.3 K-S test results for comparing reference and replication datasets with genuine biological signals

The results are summarized in Table S4.

Batch magnitude (η/σ)	Estimate of $\pi_{\text{consistent}}$	
	initializing at 0.02	global maximum likelihood estimate
0.0	0.015	0.102
0.2	0.013	0.053
0.4	0.012	0.049
0.6	0.012	0.048
0.8	0.012	0.046
1.0	0.013	0.152
1.5	0.013	0.146
2.0	0.013	0.148

Table S3: Estimation results of batch effect detection simulation by the IDR method. Each dataset is simulated by adding the corresponding magnitude of the batch effect. The results shown in the table are estimated reproducible signal proportion. The second column shows the estimation results when the initial values are set as 0.02 for the π_1 while the third column illustrates estimation results after selecting according to the global maximum likelihood from the set of different starting values.

4 Details of real data applications

4.1 Overview of TWAS analysis

Transcriptome-wide association study (TWAS) falls into the category of instrumental variable (IV) analysis, a special technique in causal inference. The general idea of TWAS can be explained by the diagram shown in Figure S3. The causal hypothesis is that relevant genetic variants ($\boldsymbol{G}$), or eQTLs in this case, regulate the gene expressions of a specific gene ($\boldsymbol{X}$), and the altered gene expression levels subsequently cause changes in complex traits ($\boldsymbol{Y}$). In the absence of a causal relationship from the target gene to the complex trait of interest, the arrow from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is missing. The relevant eQTLs are instrumental variables in this setting. The IV model is particularly attractive because of its explicit consideration of unobserved confounding factors ($\boldsymbol{U}$), which are inevitable in this genomics context. There are also three key assumptions required to establish causal conclusions in the IV analysis. In the context of TWAS, they are

1. Inclusion restriction: genetic variants are associated with expression levels, i.e., $\boldsymbol{G}$ are

Batch magnitude	K-S test p-values	
(η/σ)	No processing	ComBAT processed
0.0	0.313	0.400
0.2	0.061	0.069
0.4	0.033	0.048
0.6	0.048	0.133
0.8	7.2×10^{-3}	0.043
1.0	6.1×10^{-4}	9.6×10^{-3}
1.5	7.1×10^{-6}	1.45×10^{3}
2.0	1.18×10^{-7}	1.45×10^{-3}

Table S4: **K-S test results for comparing reference and replication datasets with genuine biological signals** The p-values are derived from two-sample two-sided K-S tests. The results are less conclusive when the genuine biological effects are presented.

eQTLs of target gene

2. Randomization restriction: no arrows between $\boldsymbol{G}$ and $\boldsymbol{U}$

3. Exclusion restriction: no direct arrow from $\boldsymbol{G}$ to $\boldsymbol{Y}$

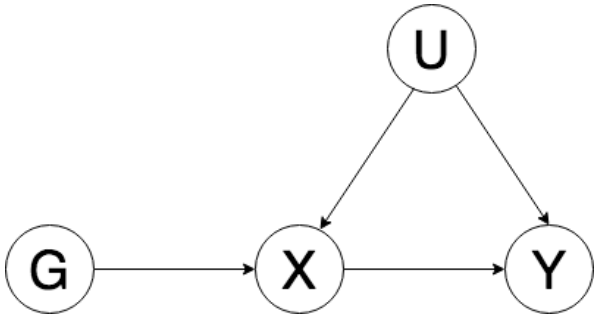

Figure S3: **Calibration of estimated reproducible probabilities** The estimated reproducible probabilities based on simulation scheme SS4 are binned into multiple frequency bins. We plot the mean of the estimated probabilities versus the frequency of the actual reproducible signals for each bin. The error bars represent the estimated 95% confidence intervals.

Under this set of causal assumptions, it is sufficient to reject the null hypothesis of no causal relationship from $\boldsymbol{X}$ to $\boldsymbol{Y}$ by showing

$\boldsymbol{G} \not\perp\!\!\!\perp \boldsymbol{Y}$,

i.e., $\boldsymbol{Y}$ and $\boldsymbol{G}$ are associated [6, 7, 8]. It is important to note that in most applications of TWAS, the main goal is to test the existence of causal links (but not to estimate the gene-to-trait effects). A popular and practical view of TWAS testing procedure is to "impute" a genetic-predicted gene expression levels for the target gene based on eQTL information, i.e.,

$$\hat{\boldsymbol{X}} = \sum_k \beta_k \boldsymbol{G}_k,$$

and subsequently test the association between $\boldsymbol{Y}$ and $\hat{\boldsymbol{X}}$. The resulting association test is in the form of a *burden test* where the weight for each genetic variant is the corresponding eQTL effect size. Various available TWAS testing approaches [9, 10, 11, 12] all follow this general form but differ in choosing weights. Note that the choice of weights in a burden test only affects its power but not the type I errors. The specific weights (β_k's) that we use in this paper are obtained from the PTWAS method [12]. They are derived from the results of the Bayesian fine-mapping analysis of eQTL data using Bayesian model averaging and represent the state-of-the-art.

In practice, individual-level GWAS data are often unavailable. [13, 14] show that the burden test of TWAS analysis can be carried out using summary-level statistics, namely, the z-statistics from single-SNP association testing. Specifically, a test statistic, Z, for each target gene can be constructed by

$$Z = \frac{\sum_k w_k(\beta_k) z_k}{\sqrt{\sum_k w_k(\beta_k)^2}}, \tag{7}$$

where the new weights w_k's are transformed from the β_k's. It is clear that under the null hypothesis, $Z \sim \mathrm{N}(0, 1)$. The particular form of test statistics that we apply in the paper is based on the software package GAMBIT implemented in [14].

The more detailed information on the TWAS analysis presented in the main text can be found in [12]. The histogram is clearly right-skewed, indicating a proportion of genes does not meet

the fixed effect assumption at the z-score scale.

4.2 Additional results for comparing GIANT and UKB Height GWAS

The criterion for filtering lowly expressed genes in this analysis follows the practice of the GTEx consortium [15]. A gene is considered expressed if its RNA-seq quantification satisfies

1. TPM (Transcripts Per Kilobase Million) ≥ 0.1 in $\geq 20\%$ of the tissue samples
2. unnormalized read counts ≥ 6 reads in $\geq 20\%$ of the tissue samples

4.2.1 Meta-analysis results

We apply traditional meta-analysis methods on the TWAS burden test z-scores derived from GIANT and UKB Height GWAS data using GTEx muscle-skeletal eQTLs.

We performed the Cochran's Q-tests for all genes across the two sets of z-scores. The histogram of the resulting p-values are shown in Figure S4. Because of the sample size difference, the resulting p-value distribution is right-skewed.

To compare with the INTRIGUE result, we also conduct the fixed-effect meta-analysis to combine the TWAS z-scores from the two studies. If the Q-test results are ignored, we reject 3379 genes at FDR 5% level. This set of genes overlaps with 97.1% of GIANT rejected genes and 91.1% of UKB rejected genes at the same FDR level. However, this is not a common practice because not all genes satisfy the fixed-effect assumption. By filtering out 1346 genes with Q-test p-value < 0.05, the total rejections decrease to 2638 genes, which overlap 79.3% of GIANT rejections and 67.8% of UKB rejections.

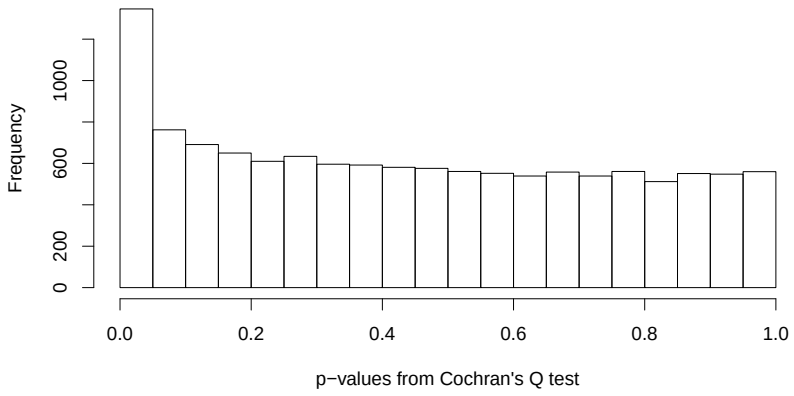

Figure S4: **Distribution of p-values from Cochran's Q tests** The p-values are derived from a one-sided χ^2 test, where the null hypothesis asserts that the underlying effects in both datasets are identical.

4.2.2 IDR results

We also applied the IDR method to compare the TWAS z-scores derived from the two GWAS datasets. Figure S5 shows the rank, the correspondence curve Ψ_n, and the changes of the correspondence curve $\Psi_n^{'}$ plots using the absolute values of TWAS z-scores from the two studies. The change point around 0.5 in the correspondence curve validates the two-component bivariate Gaussian mixtures in the GCMM space [3]. However, the IDR EM algorithm fails to converge in all initial values that we experiment with.

To run the IDR EM algorithm, we pre-define a grid of initial values for the GCMM parameters. Specifically, we set the starting values of reproducible proportion $\pi_{\text{consistent}}$ from 0.4 to 0.9, the mean μ_1 for the reproducible proportion of signal ranges from 1.0 to 4.0, the variance σ_1^2 and the correlation ρ for the reproducible bivariate normal distribution are from 0.5 to 2.5 and 0.6 to 0.9, respectively. We also extend the maximum iteration for the EM algorithm to 10000 from the default value (30). However, the EM algorithm fails to converge in all settings that we have experimented with, indicating the computational difficulty.

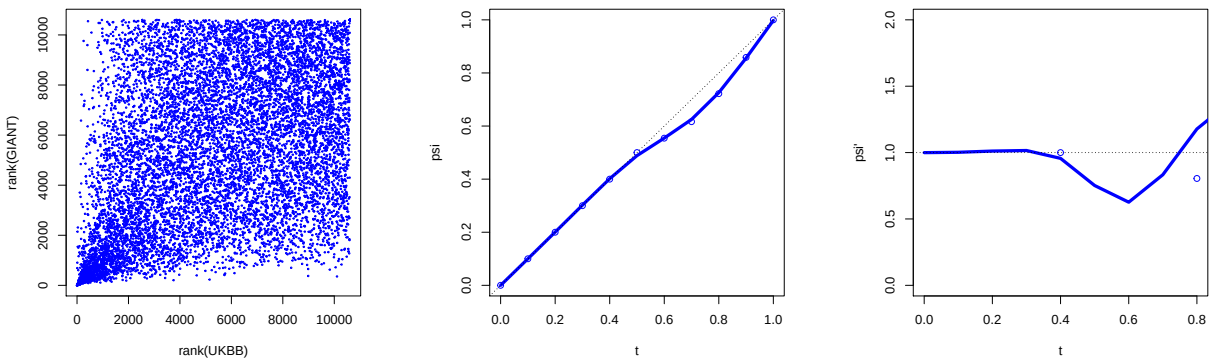

Figure S5: **IDR diagnosis plots for TWAS z-scores derived from GIANT and UKB Height GWAS data** From left to right, the three panels represent the scatter plot of the ranks of the absolute values of TWAS z-scores, the curve of Ψ values (the correspondence curve) and the curve of derivative of Ψ values. The correspondence curve indicates the proportion of the pairs that are ranked on the upper $100 \times t\%$ of $|z_{\text{UKB}}|$ and $|z_{\text{GIANT}}|$. The presence of the change point in the correspondence plot validates the GCMM assumptions.

4.3 Additional results for multi-tissue TWAS analysis

4.3.1 Results with subsampled muscle-skeletal samples

To best match the power of eQTL data in GTEx muscle-skeletal (sample size = 706) and whole blood (sample size = 670), we down-sample the muscle-skeletal eQTL data to match the whole blood data in sample size exactly. We then perform the Bayesian fine-mapping and construct the TWAS weights for the muscle-skeletal using this subset samples. Finally, we re-run the TWAS analysis using the new weights.

The INTRIGUE results based on this sample size-matched weights are virtually unchanged (Table S5)

		π_{Null}	π_{R}	π_{IR}
Full gene set	CEFN	0.795	0.069	0.136
	META	0.838	0.031	0.132
eGene set	CEFN	0.297	0.397	0.306
	META	0.407	0.366	0.227

Table S5: **Estimated null, reproducible, and irreproducible proportions between the height TWAS signals using whole blood and muscle skeletal tissues with matched sample size** The muscle TWAS weights are re-constructed using a subset of GTEx muscle-skeletal samples. The results are numerically close to the corresponding part of Table 3 in the main text.

		π_{Null}	π_{R}	π_{IR}
Full gene set	CEFN	0.821	0.027	0.152
	META	0.868	0.003	0.129
eGene set	CEFN	0.272	0.273	0.455
	META	0.448	0.145	0.407

Table S6: **Estimated null, reproducible, and irreproducible proportions between the height TWAS signals using UKB GWAS-GTEx blood eQTL vs. GIANT GWAS-GTEx muscle eQTL data** In this setting, the GWAS data are from two different consortia. Therefore, the assumptions of the INTRIGUE models are better met. The overall results remain qualitatively similar.

4.3.2 Multi-tissue analysis with different GWAS Height datasets

5 Detection of publication bias

In this section, we illustrate the application of INTRIGUE in detecting publication bias via simulations. Here, we no longer consider a high-throughput setting and focus on evaluating an individual set of observations/estimates from a pair of studies.

We design our simulation scheme by following [4, 5]. Particularly, we consider two balanced case-control studies investigating a treatment effect with an odds ratio of $= 0.75$. The first study is well-powered with sample size $N_1 = 4000$, while the second study is under-powered with sample size $N_2 = 200$. The prevalence in the control population is drawn from the distribution Uniform$(0.1, 0.5)$.

In the first simulation setting, we impose a selection process to censor the simulated data in the small study. Specifically, we record the simulated data only if the resulting p-value for testing the treatment effect ≤ 0.01. Otherwise, the simulated data are discarded, and we repeat the simulation procedure until the criterion is met. This selection process mimics an aggressive P-hiking procedure. For comparison, we create another similar simulation setting where neither study is imposed with the selection process, and the simulated data represents the case absent of publication bias. We repeat the simulation procedures to generate 2000 sets of observations with publication bias (setting 1) and 2000 sets of observation without publication publications (setting 2)

Our analysis considers one set of observations at a time. We perform Bayesian model comparison by computing a Bayes factor for the irreproducible model (BF_{irr}) and a Bayes factor for the reproducible model (BF_{rep}) based on the two estimates from the large and the small studies. Note that the ratio of the two Bayes factors,

$$\text{BF}_{\text{irr\_vs\_rep}} = \frac{\text{BF}_{\text{irr}}}{\text{BF}_{\text{rep}}}, \tag{8}$$

is also a valid Bayes factor. $\text{BF}_{\text{irr\_vs\_rep}} > 1$ indicates that the observed data support the irreproducible model, whereas $\text{BF}_{\text{irr\_vs\_rep}}$ indicates that the data support reproducible model.

Figure S6 illustrates the ability of $\text{BF}_{\text{irr\_vs\_rep}}$ in detecting publication bias as a model comparison criterion. The Bayes factors mostly favor the irreproducible model when the small sample size data are generated with the selection bias, while they generally favor the reproducible model when there is no selection process. The META model performs better than the CEFN model, as the non-adaptive heterogeneity is better suited for detecting publication bias. Figure S7 shows the ROC for classifying publication bias using the simulated data based on $\text{BF}_{\text{irr\_vs\_rep}}$ computed from the META model.

The simulation results illustrate that the directional consistency criterion as a useful tool to

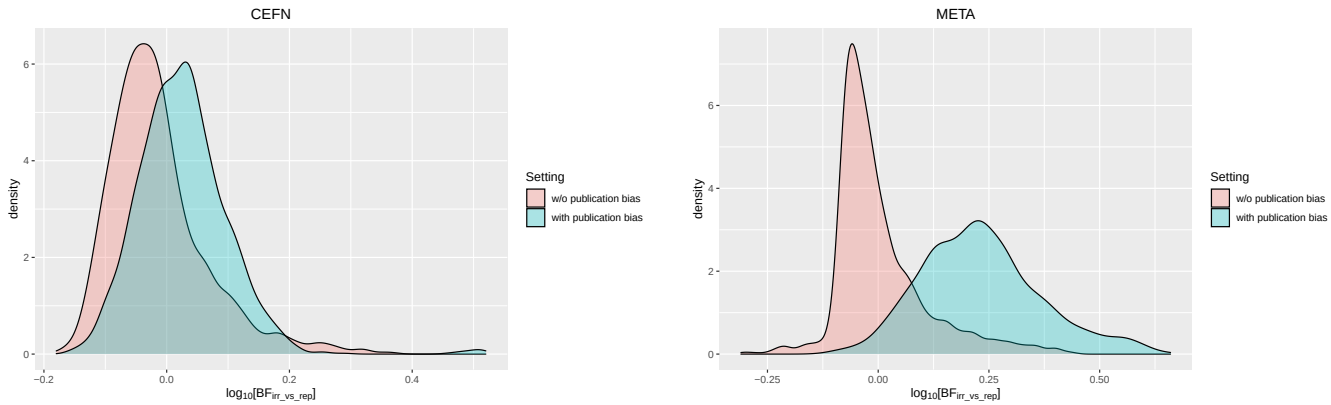

Figure S6: **Model comparison results for simulated publication bias data** The figure shows the distributions of model comparison criterion, $\log_{10}(\mathrm{BF}_{\mathrm{irr\_vs\_rep}})$, for simulated data generated with and without publication bias. The Bayes factors in the left and the right panel are computed using the CEFN and the META model, respectively. The META model shows better distinguishing power in this context.

define expected heterogeneity in repeated measurements. It is important to note that the DC criterion defines the magnitude of expected heterogeneity but does not constrain the direction of the variation. As a result, the excessive variation observed in the data, even when the estimates maintain the correct sign, leads to evidence for irreproducible results.

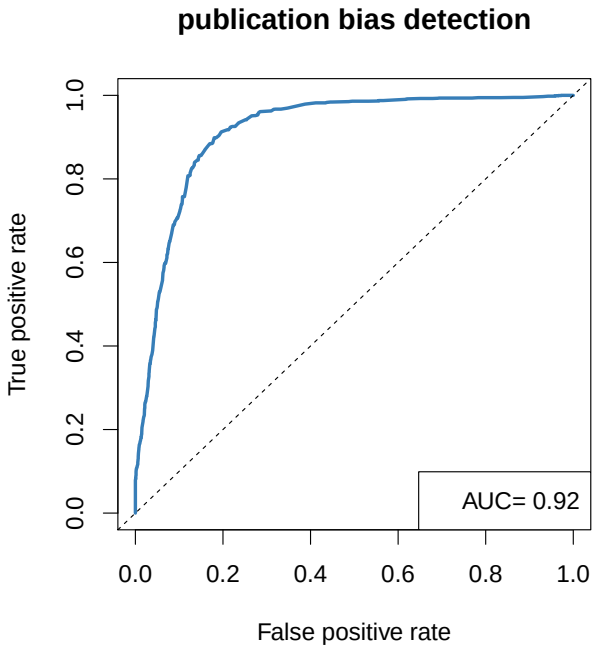

Figure S7: **ROC curve for publication bias detection using INTRIGUE with meta prior**

References

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

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

[3] Li, Q., Brown, J. B., Huang, H., Bickel, P. J. *et al.* Measuring reproducibility of high-throughput experiments. *The annals of applied statistics* **5**, 1752–1779 (2011).

[4] Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3** (2007).

[5] Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2007).

[6] Katan, M. Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet* **327**, 507–508 (1986).

[7] VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M. & Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)* **25**, 427 (2014).

[8] Sheehan, N. A. & Didelez, V. Epidemiology, genetic epidemiology and mendelian randomisation: more need than ever to attend to detail. *Human genetics* 1–16 (2019).

[9] Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091 (2015).

[10] Zhu, Z. *et al.* Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics* **48**, 481 (2016).

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

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

[13] Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications* **9**, 1825 (2018).

[14] Quick, C., Wen, X., Abecasis, G., Boehnke, M. & Kang, H. M. Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis. *BioRxiv* 732404 (2019).

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