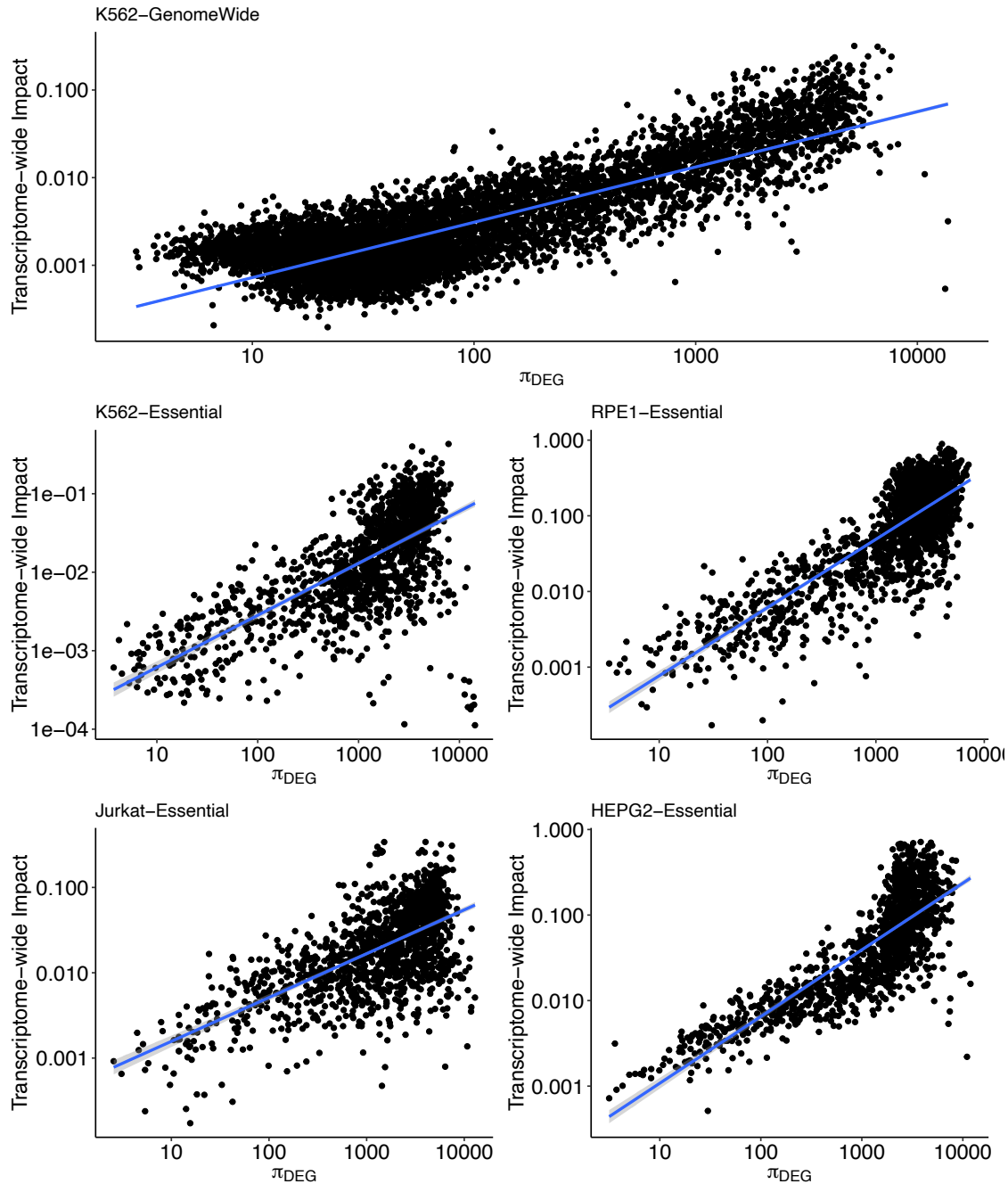


Supplementary Appendix 1: Transcriptome-wide impact and π_{DEG}

Transcriptome-wide impact and π_{DEG} capture different aspects of the distribution of differential effects. Transcriptome-wide impact is the variance of the effect-size distribution, and π_{DEG} is a function of the kurtosis of the effect size distribution, that captures the “effective number of differentially expressed genes”. The relationship between transcriptome-wide impact and π_{DEG} gives insight into how effects are spread across genes.

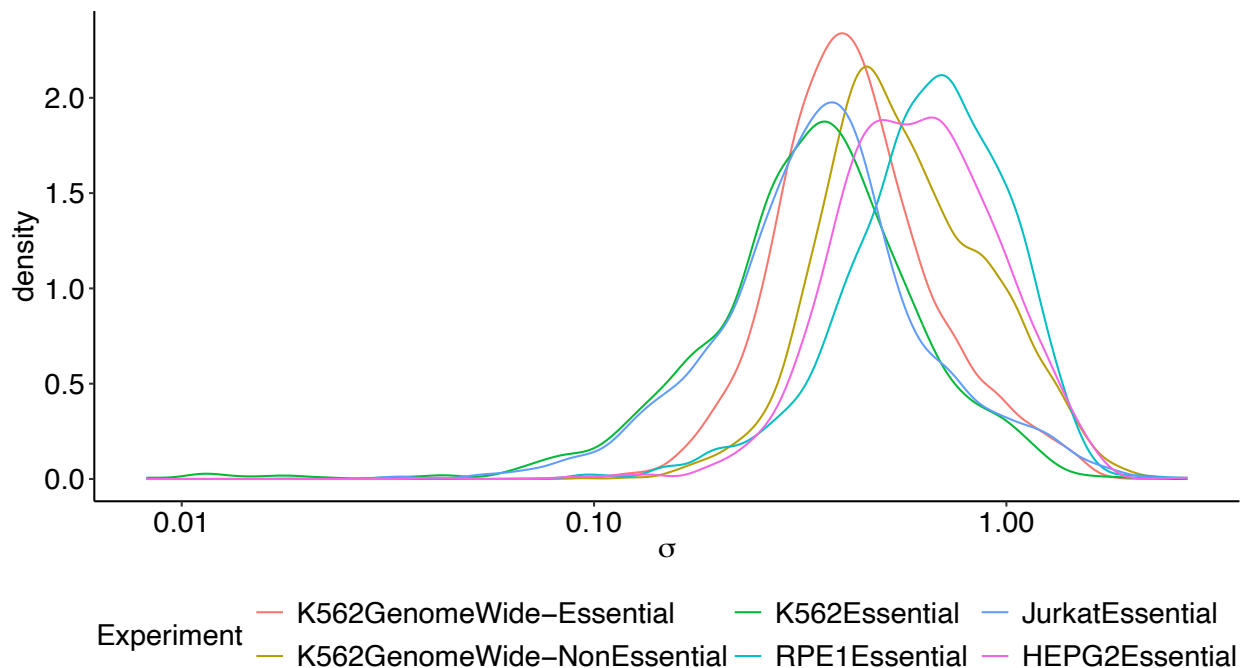


(Notably, this and the below analyses are restricted to perturbations that caused significant transcriptome-wide impact, due to the instability of π_{DEG} in the setting of low/non-significant transcriptome-wide impact)

As predicted, transcriptome-wide impact and π_{DEG} are positively associated; as the number of genes affected rises, so too does the amount of differential expression signal. Less trivially, there appear to be no perturbations that achieve a large transcriptome-wide impact while affecting a few genes (i.e. paucity of points in the top-left of the above plots). This suggests that no perturbations cause large transcriptomic change via very large effects on a small set of genes; rather, when perturbations cause large transcriptomic change, they do so by affecting many genes.

A typical $\log_2(\text{FoldChange})$ magnitude of affected genes σ can be computed with the ratio of transcriptome-wide impact to π_{DEG} :

$$\sigma = \sqrt{\frac{n_{genes} * TI}{\pi_{DEG}}}$$



Examining the distribution of σ reveals several aspects of perturbation effect architecture:

- For all experiments, the variability in σ is mostly limited to the range (0.1,1), suggesting that typical differential expression effects exhibit a limited range across perturbations
- For the K562 cell line, σ is larger in the GenomeWide experiment than in the Essential experiment, suggesting that while essential perturbations affect more genes, they do so with a smaller typical effect size
- The mean σ varies between cell types. In particular, it is largest in the RPE1 experiments, slightly smaller in the HepG2 experiments, and smaller in the K562 and Jurkat experiments

Supplementary Appendix 2: Bias of the sample correlation coefficient

This phenomenon, known as *attenuation*, was identified in Spearman (1904); the arguments from that paper are paraphrased here.

Consider two sets of effect sizes (i.e. log2FoldChanges) that are estimated with additive, independent measurement error:

$$\begin{aligned}\widehat{\beta}_1 &= \beta_1 + \epsilon_1 \\ \widehat{\beta}_2 &= \beta_2 + \epsilon_2\end{aligned}$$

The correlation between the estimated effect sizes is:

$$\begin{aligned}\text{Corr}(\widehat{\beta}_1, \widehat{\beta}_2) &= \frac{\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2)}{\sqrt{\text{Var}(\widehat{\beta}_1)\text{Var}(\widehat{\beta}_2)}} \\ &= \frac{\text{Cov}(\beta_1 + \epsilon_1, \beta_2 + \epsilon_2)}{\sqrt{\text{Var}(\beta_1 + \epsilon_1)\text{Var}(\beta_2 + \epsilon_2)}} \\ &= \frac{\text{Cov}(\beta_1, \beta_2) + \text{Cov}(\beta_1, \epsilon_2) + \text{Cov}(\epsilon_1, \beta_2) + \text{Cov}(\epsilon_1, \epsilon_2)}{\sqrt{\text{Var}(\beta_1 + \epsilon_1)\text{Var}(\beta_2 + \epsilon_2)}} \\ &= \frac{\text{Cov}(\beta_1, \beta_2)}{\sqrt{\text{Var}(\beta_1 + \epsilon_1)\text{Var}(\beta_2 + \epsilon_2)}} < \frac{\text{Cov}(\beta_1, \beta_2)}{\sqrt{\text{Var}(\beta_1)\text{Var}(\beta_2)}} = \text{Corr}(\beta_1, \beta_2)\end{aligned}$$

Therefore, the correlation of the estimated effect sizes is a downwardly biased estimator of true effect size correlation, and this downward bias is related to the relative magnitude of the sampling noise. If the sampling variation is of similar or greater magnitude than the true effect size variance, this bias will be severe.

Supplementary Appendix 3: Consistency of dose-response curves

Consider the relationship between the expression of target genes of a gene perturbation. For each target gene g , the expression of the target gene at dosage 1 of the perturbed gene is equal to the expression of the target gene at full dosage of the perturbed gene times a function $f_g(dosage)$, that describes the response kinetics of g in response to changing dosage of the perturbed gene.

$$\begin{aligned}\text{Expression}_g(\text{dosage}_1) &= f_g(\text{dosage}_1) * \text{Expression}_g(\text{full dosage}) \\ \text{Expression}_g(\text{dosage}_2) &= f_g(\text{dosage}_2) * \text{Expression}_g(\text{full dosage})\end{aligned}$$

Rearranging:

$$\begin{aligned}\text{FoldChange}_g(\text{dosage}_1) &= f_g(\text{dosage}_1) \\ \text{FoldChange}_g(\text{dosage}_2) &= f_g(\text{dosage}_2)\end{aligned}$$

$$\text{FoldChange}_g(\text{dosage}_1) = \frac{f_g(\text{dosage}_1)}{f_g(\text{dosage}_2)} \text{FoldChange}_g(\text{dosage}_2)$$

When f_g is identical across target genes up to a scaling factor, then $\frac{f_g(\text{dosage}_1)}{f_g(\text{dosage}_2)}$ is constant, and $\text{FoldChange}_g(\text{dosage}_1)$ and $\text{FoldChange}_g(\text{dosage}_2)$ are perfectly correlated across genes. However, when f_g varies across genes, and $\frac{f_g(\text{dosage}_1)}{f_g(\text{dosage}_2)}$ is no longer constant, there is an imperfect correlation between $\text{FoldChange}_g(\text{dosage}_1)$ and $\text{FoldChange}_g(\text{dosage}_2)$. This reveals that consistency of perturbation effects across dosages is at least partially a function of the consistency of downstream response kinetics across target genes.

Supplementary Appendix 4: Bias of the sample Euclidean Distance

This phenomenon is described in Mahalanobis (1936), and paraphrased here. It is closely related to the bias of the sample correlation in the presence of uncorrelated noise.

Consider the Euclidean distance between expression states of two clusters of cells:

$$\text{Euclidean Distance}^2 = \sum_g (\mu_g - \nu_g)^2$$

Where μ_g is the mean expression of gene g in cluster 1, and ν_g is the mean expression of gene g in cluster 2.

In computing distance metrics for analyses such as hierarchical clustering, this quantity is typically estimated using sample means, which differ from the population mean due to sampling variation with a known distribution:

$$\begin{aligned}\widehat{\mu}_g &= \mu_g + \epsilon_g & \epsilon_g &\sim N\left(0, \frac{\sigma_g^2}{n_1}\right) \\ \widehat{\nu}_g &= \nu_g + \delta_g & \delta_g &\sim N\left(0, \frac{\tau_g^2}{n_2}\right)\end{aligned}$$

Computing the Euclidean Distance with sample means rather than population means, we have:

$$\text{Euclidean Distance}^2 = \sum_g (\mu_g + \epsilon_g - (\nu_g + \delta_g))^2$$

Taking expectation, and assuming that ϵ_g and δ_g are independent:

$$\begin{aligned}& \sum_g (\mu_g - \nu_g)^2 + \sum_g \text{Var}(\epsilon_g) + \text{Var}(\delta_g) \\ &= \sum_g (\mu_g - \nu_g)^2 + \sum_g \frac{\sigma_g^2}{n_1} + \frac{\tau_g^2}{n_2} \\ &= \text{Euclidean Distance}^2 + \frac{1}{n_1} \sum_g \sigma_g^2 + \frac{1}{n_2} \sum_g \tau_g^2\end{aligned}$$

We see that that when estimating Euclidean Distance with sample means, the resulting estimates are upwardly biased in the following manner:

- The upward bias becomes smaller as the number of cells in each cluster (n_1 and n_2) increases
- The upward bias becomes larger as the variance in expression between cells (σ_g^2 and τ_g^2) grows.
- The upward bias becomes larger as more genes are included in the calculation.