CDS DS 596 Assignment 3

Xiang Fu
xfu@bu.edu
Boston University Faculty of Computing & Data Sciences

Contents

In this assignment you are asked to read "Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq" and provide written responses to the questions and prompts below. Questions in section III will ask you to make comparisons with the models used in a related paper (Dixit et al). While you should read that paper in enough depth to understand the model, you are not asked to describe the results. Section I (basic information) will account for 15% of the grade, section II (factual information) 50%, and section III (comparison) 35%.

1 Basic Information

1. What are the goals of the study and what are the key claims?
2. What is the broader significance of this work if the claims are true?

This study presents the first genome-scale single-cell CRISPR screens using Perturb-seq, which combines CRISPR-based genetic perturbations with single-cell RNA sequencing readouts. The key goals and claims are:

- They performed Perturb-seq targeting all expressed genes in human cell lines to create comprehensive genotype-phenotype maps. This enabled systematic assignment of function to poorly characterized genes and in-depth dissection of complex cellular phenotypes.
- They developed a computational framework to robustly detect transcriptional phenotypes from Perturb-seq data. About 30% of genetic perturbations were found to cause significant transcriptional changes.
- Transcriptional phenotypes were used to predict gene function, uncovering new regulators of processes like ribosome biogenesis, transcription, and mitochondrial respiration. Phenotypic clustering recapitulated known biological relationships and protein complexes.
- Single-cell resolution allowed the study of phenotypic penetrance, cell-to-cell variability, and pleiotropic effects. As examples, they systematically identified genetic drivers of chromosomal instability and dissected its effects on cell cycle and stress responses.
- Mitochondrial genome expression showed stress-specific regulation in response to diverse mitochondrial perturbations. This suggests a model where a general nuclear stress response is layered with perturbation-specific changes in mitochondrial genome regulation.

2. If the claims are true, soem signifiance would be:

    - Providing an approach to systematically map gene function and genetic interactions at scale using single-cell phenotyping, which enables assigning functions to the large number of poorly characterized genes in the human genome.
    - Establishing genome-scale single-cell CRISPR screening as a generalizable platform for exploring genetic regulation of cellular processes and behaviors across different cell types and contexts.
    - It demonstrates the unique insights gained from single-cell phenotyping, like identifying variable phenotypes and studying relationships between different perturbation effects within individual cells.
    - This work woudl also establishes Perturb-seq as a powerful tool for large-scale dissection of gene function, biological pathways, and complex cellular behaviors in an unbiased manner. The information-rich genotype-phenotype maps generated provide a valuable resource.

2 Factual Information

Provide a description of what was done (rather than what was said) experimentally and computationally. It may be useful (but not necessary) to organize your response figure-by-figure (i.e., explain what was

done to generate results for Fig. 1, what was done for Fig. 2, etc.). Place more emphasis in your response on the earlier figures (especially Figs 1 and 2) but provide at least a brief description of what was done for all seven main figures.

Provide a description of the key factual findings presented. Here, you very likely want to organize your response in terms of what is presented in each figure. Again, place more emphasis on the earlier figures but provide at least a brief description of what was presented in all seven main figures.

For both responses above, the crucial information may be contained in "Supplementary Materials" (or STAR Methods).

Experimental and Computational Methods:

Figure 1

In Figure 1, the authors designed compact multiplexed CRISPRi sgRNA libraries targeting all expressed genes for Perturb-seq. They performed genome-scale Perturb-seq screens in K562 cells (a chronic myeloid leukemia cell line) at day 8 and day 6 timepoints and in RPE1 cells (a retinal pigment epithelial cell line) at day 7. The screens utilized 10x Genomics single-cell RNA-seq with sgRNA capture, collecting data for over 2.5 million cells. To analyze the data, they developed a computational framework to detect transcriptional phenotypes, which included internal normalization, a permuted energy distance test for detecting global changes, and the Anderson-Darling test for identifying gene-level differential expression.

Figure 2

For Figure 2, they analyzed 1,973 strong transcriptional phenotype perturbations from the K562 day 8 dataset. They compared expression profile correlations between perturbations to known protein-protein interactions from the CORUM and STRING databases. An unbiased clustering of perturbations was performed, and the results were visualized using a minimum distortion embedding. Cluster functions were manually annotated using databases and literature. The predicted ribosome biogenesis roles for 10 poorly annotated genes were validated by rRNA analysis.

Figure 3

In Figure 3, they examined the variable phenotypes of Integrator complex subunits and identified functional modules that mirrored the known architecture. They showed that C7orf26 is part of a distinct INTS10/13/14 module using techniques such as co-depletion, co-immunoprecipitation, purification, and analysis in Drosophila. The roles of different Integrator modules in snRNA processing were analyzed using splicing scores and PRO-seq.

Figure 4

Figure 4 describes the development of an approach to summarize genotype-phenotype maps by clustering co-regulated genes into expression programs and perturbations by transcriptional profiles. The authors explored data-driven gene expression programs like UPR and ISR activation. They also studied the effects of perturbations on erythroid/myeloid differentiation and validated targets by surface marker expression. Composite phenotypes such as total RNA content and TE expression were used to identify drivers of these phenotypes.

Figure 5

In Figure 5, the authors quantified single-cell phenotypic penetrance using SVD-based leverage scores. They identified chromosome segregation perturbations causing heterogeneity and inferred single-cell copy number variation to detect karyotypic changes. The effects of chromosomal instability on cell cycle and stress responses were analyzed. A CIN score was developed to find diverse drivers of chromosomal instability.

Figure 6

For Figure 6, the authors clustered mitochondrial perturbations separately by nuclear and mitochondrial transcriptional responses. They found that the nuclear response did not discriminate perturbations by function, while the mitochondrial genome expression was highly variable and stress-specific, separating perturbations to different complexes.

Figure 7

In Figure 7, the authors examined the distribution of mitochondrial genome expression to identify perturbation-specific regulation. The results were validated by bulk RNA-seq without poly-A selection. The mitochondrial transcriptome was used to predict TMEM242′s role in ATP synthase function, which was validated experimentally by a respiration assay. The authors proposed a model of a general nuclear stress response combined with specific mitochondrial genome regulation.

Key Findings

Figure 1

Figure 1 demonstrates that multiplexed CRISPRi enabled effective knockdown in genome-scale Perturb-seq. Around 30% of genetic perturbations caused a significant transcriptional phenotype, which was robustly detected by the computational framework developed by the authors.

Figure 2

In Figure 2, the authors show that transcriptional phenotype correlations recapitulated known protein complexes and interactions. Unbiased clustering identified gene functions spanning core cellular processes. The study also predicted and validated roles for poorly characterized genes in ribosome biogenesis.

Figure 3

Figure 3 reveals that the Integrator complex contained variable phenotype modules mapping to its structural architecture. C7orf26 was found to be part of a distinct INTS10/13/14 module, suggesting it is likely an Integrator subunit. The modules had distinct roles, with only the cleavage module being required for snRNA processing.

Figure 4

The genotype-phenotype maps presented in Figure 4 revealed expression programs activated by specific perturbations. Selectively essential genes like PTPN1 drove differentiation phenotypes and enhanced effects when combined with known targets. Composite phenotypes identified exosome perturbations increasing TE RNA and cell cycle perturbations increasing total RNA.

Figure 5

Figure 5 shows that heterogeneity analysis found chromosome segregation perturbations causing variable karyotypes. Chromosomal instability triggered p53-dependent G1 arrest and integrated stress response. Many diverse gene perturbations were found to cause chromosomal instability.

Figure 6

In Figure 6, the authors demonstrate that nuclear transcriptional responses to mitochondrial perturbations were relatively homogeneous. In contrast, mitochondrial genome expression was highly variable and stress-specific, differentiating perturbations to each complex. The mitochondrial transcriptome was more predictive of perturbation type than the nuclear transcriptome.

Figure 7

Figure 7 indicates that mitochondrial genome regulation occurs at multiple levels based on the stressor. TMEM242 knockdown was predicted and validated to disrupt ATP synthase function based on expression and respiration data. The authors propose a layered model of a general nuclear stress response combined with specific mitochondrial genome changes.

3 Comparison with Alternative Methods

An earlier paper (Dixit et al, Cell 2016) introduced the Perturb-Seq method and inferred perturbation effects (i.e., the effect of each perturbation on the expression of each gene) through a different approach. Briefly describe the model used in Dixit et al and compare it with the approach of Replogle et al. Discuss the relative advantages and disadvantages of each approach.

When estimating significance of perturbation effects, both approaches use a technique to correct for multiple hypotheses. What is this method? Briefly describe the motivation for the correction, and how it works. Why might this method be more or less appropriate than alternative methods for controlling error rates that we discussed in class?

Propose an alternative computational approach to inferring perturbation effects that was not used in either study and discuss what advantages your approach might have over the methods in these papers.

Comparision of Models in Dixit et al. and Replogle et al.

Dixit et al. (2016) use a linear regression framework to model the relationship between gene expression and perturbations. Their model relates the expression of each gene (Y) to the sgRNA perturbations (X), with the coefficient matrix β capturing the effect of each perturbation on each gene's expression. They account for technical covariates, such as cell quality, by including them in the design matrix X. To address the issue of ineffective perturbations in some cells, they employ an iterative refinement procedure. First, they use the initial fit of β to evaluate whether each cell's expression profile is consistent with its assigned perturbation. Then, they re-estimate the model with the corrected perturbation-to-cell assignments. This process is repeated until convergence. Additionally, they incorporate cell subtype/state classifications from unperturbed cells as covariates in the model.

On the other hand, Replogle et al. (2022) take a non-parametric approach to identify perturbation effects. They use a permutation-based energy distance test to identify perturbations that cause overall transcriptional changes. This test compares the multivariate distribution of gene expression between perturbed and control cells without assuming a specific functional form. For gene-level effects, they apply the Anderson-Darling (AD) test, which compares the distribution of each gene's normalized expression between perturbed and control cells. The AD test is sensitive to changes in the tails of the distributions and does not assume a specific distribution. To focus on high-quality perturbations, they filter out those with low on-target knockdown efficiency based on the expression of the targeted gene.

The Dixit et al. approach has the advantage of providing a unified framework to estimate perturbation effects while adjusting for confounding factors. By explicitly modeling the perturbations as covariates, it enables the assessment of interactions and the control for technical and biological variation. However, this approach relies on assumptions about the functional form of the effects and the noise distribution, which may not always hold. It also depends on the iterative refinement procedure to handle ineffective perturbations, which may be sensitive to the choice of threshold.

In contrast, the Replogle et al. approach uses flexible non-parametric tests that make fewer assumptions about the nature of the perturbation effects. The AD test, in particular, is designed to detect differences in the tails of the distributions, which may be more relevant for perturbations with heterogeneous effects. However, by not explicitly modeling the relationship between perturbations and expression while adjusting for covariates, this approach may have reduced power and interpretability compared to a well-specified parametric model.

Also, I think it might be possible to integrate the two approaches to leverage their complementary strengths. For example, the non-parametric tests could be used as a screening step to identify candidate perturbations, followed by a tailored parametric model to estimate their effects while adjusting for covariates. Alternatively, the linear model of Dixit et al. could be extended to incorporate more flexible effect sizes, such as through interaction terms, or to use generalized linear models to account for non-Gaussian distributions. Such hybrid approaches could potentially offer a balance between the robustness of non-parametric methods and the efficiency and interpretability of parametric models.

Multiple Hypothesis Correction Method

Both studies use the Benjamini-Hochberg procedure to control the false discovery rate (FDR) when testing many genes for differential expression. This method aims to limit the expected fraction of false positives among all rejected null hypotheses. It works by ordering the p-values, finding the largest p-value below a line determined by the desired FDR, and rejecting all hypotheses with smaller p-values.

In Perturb-seq experiments, a hypothesis test is conducted for each gene to determine if it is differentially expressed in response to each perturbation. With thousands of genes and perturbations, this amounts to a large number of simultaneous tests. When testing multiple hypotheses, the probability of making at least one Type I error (false positive) increases rapidly with the number of tests. For example, with a 5% significance level, the probability of at least one false positive is $1-(1-0.05)^n$, which exceeds 99% for $n = 100$ tests. Multiple testing correction is thus crucial to limit false positives.

The BH procedure for FDR control is often more appropriate than FWER methods like Bonferroni correction in high-throughput genomics settings. FWER methods control the probability of making any Type I errors, which is often too conservative when testing thousands of hypotheses. They may fail to detect many true positives. In contrast, FDR control allows a small proportion of false positives in order to achieve greater power. This is justified when the goal is to generate a list of candidate genes for further validation, rather than definitively proving the effect of each gene.

This method have the assumption that the test statistics are independent or have positive regression dependency. While gene expression levels are often correlated, independence can be approximated by testing a subset of weakly correlated genes or by using permutation-based p-values that preserve the correlation structure. Positive regression dependency is a weaker condition that holds for many common test statistics. If these assumptions are violated, the BH procedure may be liberal, but modifications exist for dependent tests (e.g., Benjamini-Yekutieli procedure).

Essentially, the BH procedure controls the FDR, defined as the expected proportion of false positives among all rejected null hypotheses.

1. Order the p-values from smallest to largest: $p(1) \leq p(2) \leq ... \leq p(n)$.
2. Find the largest integer k such that $p(k) \leq \left(\frac{k}{n}\right) * \alpha$, where n is the total number of hypotheses and α is the desired FDR.
3. Reject the null hypothesis for all tests with p-values $\leq p(k)$.

On the other hand, intuitively, the BH procedure compares each p-value to a threshold that depends on its rank and the desired FDR. This allows more rejections than controlling the FWER, while still limiting the expected FDR to the specified level.

Lastly, the choice of method to use depends on the desired balance between Type I and Type II errors, the assumptions about the dependence structure of the tests, and computational considerations. The BH procedure strikes a good balance for most Perturb-seq applications, but sensitivity analyses with alternative methods can help assess the robustness of the findings.

Alternative Computational Approach

An alternative approach to inferring perturbation effects is to use a Bayesian hierarchical model. In this framework, the expression of each gene in each cell is modeled using a negative binomial distribution, which accounts for the overdispersion commonly observed in single-cell RNA-seq count data. The perturbations are encoded as indicator variables in the model, and the coefficients for each perturbation are given sparse prior distributions, such as the Laplace or horseshoe prior, to induce shrinkage towards zero. This prior structure encourages sparsity in the perturbation effects, reflecting the biological expectation that most perturbations likely affect only a subset of genes. The hierarchical nature of the model allows for information sharing across genes, which can improve the estimates of perturbation effects, especially for genes with low expression levels.

In order to handle the uncertainty in perturbation assignments, the model can be extended to include a probability distribution over the possible assignments for each cell. The perturbation effect estimates can then be obtained by averaging over the posterior distribution of the assignment probabilities. This approach propagates the uncertainty in the assignments to the final effect estimates, providing a more robust inference.

Inference

For inference, it can be performed using Markov chain Monte Carlo (MCMC) methods, such as Hamiltonian Monte Carlo (HMC), which explore the posterior distribution of the model parameters. However, MCMC can be computationally intensive for large-scale datasets. An alternative is to use variational inference (VI), which approximates the posterior distribution with a simpler, tractable distribution and optimizes the approximation to minimize the divergence from the true posterior. VI can provide faster inference, albeit at the cost of some accuracy.

Advantages

1. The negative binomial likelihood accounts for the mean-variance relationship in scRNA-seq count data
2. Using a distribution appropriate for over-dispersed count data
3. Obtaining posterior probabilities for effects that incorporate uncertainty.
4. The hierarchical prior structure shares information across genes to improve effect estimates, especially for genes with low counts. The Laplace prior encourages sparsity, as most perturbations likely affect a small subset of genes.

However, the Bayesian approach also comes with challenges. Specifying appropriate prior distributions can be difficult and may require sensitivity analyses to assess the impact of different choices. The computational complexity of Bayesian inference, particularly with MCMC, can be a bottleneck for large datasets. Finally, the interpretation of posterior distributions and Bayesian credible intervals may be less familiar to researchers accustomed to frequentist p-values and confidence intervals.

I would also like to propose another approach, which is a machine learning approach. We can frame the problem as a multi-task learning problem, with each perturbation defining task. The goal would be to prerdict gene expression from perturbation assignments, while also leveraging shared structure across perturbations.

Model

We use a neural network with an architecture designed for scRNA-seq data, such as a variational autoencoder (VAE) or a deep count autoencoder (DCA). The input is the perturbation assignment vector for each cell, and the output is the predicted gene expression profile. The network is trained to minimize the reconstruction loss between the predicted and observed expression, subject to regularization penalties.

Training

The network is trained end-to-end using stochastic gradient descent. The loss function includes a term for the negative log-likelihood of the count data (e.g. negative binomial) and regularization terms to encourage sparsity and prevent overfitting. Dropout can be used for perturbations to propagate assignment uncertainty.

Inference

Perturbation effects can be estimated by inspecting the learned weights connecting the perturbation inputs to the hidden layers, or by comparing the predicted expression profiles with and without each perturbation. Perturbation similarities can be assessed by the similarity of their learned representations.

Advantages

1. Neural networks can learn complex, nonlinear relationships between perturbations and expression.
2. Sharing hidden layers allows information sharing across perturbations while still allowing perturbation-specific effects.
3. Modern architectures designed for scRNA-seq data can model count data and address challenges like overdispersion and zero-inflation.

There also come with challenges, which can include the following:

1. Some deep learning models can be computationally demanding to train and tune, requiring GPUs for large datasets.
2. These models are also highly flexible and thus prone to overfitting without careful regularization and validation.
3. The learned feature representations can be difficult to interpret compared to directly modeling coefficients.