scDecouple: Decoupling cellular response and infection bias in scCRISPR-seq

Qiuchen Meng[1], Ming Shi[2], Lei Wei[1], Yinqing Li[3,*], Xuegong Zhang[1,4,*]

1 MOE Key Lab of Bioinformatics & Bioinformatics Division BRNIST, Department of Automation, Tsinghua University, Beijing 100084, China
2 School of Computer Science, Hubei University of Technology, Wuhan, China
3 School of Pharmaceutical Science, Tsinghua University, Beijing 100084, China
4 Center for Synthetic and Systems Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China
*Correspondence: XZ: zhangxg@tsinghua.edu.cn; YL: yinqingl@tsinghua.edu.cn;

Abstract

scCRISPR-seq is an emerging high-throughput CRISPR screening technology that combines CIRPSR screening with single-cell sequencing technologies. It provides rich information on gene regulation. When performing scCRISPR-seq in a population of heterogeneous cells, the observed cellular response in perturbed cells may be caused not only by the perturbation, but also by the infection bias of guide RNAs (gRNAs) mainly contributed by intrinsic differences of cell clusters. The mixing of these effects poisons gene regulation studies. We developed scDecouple to decouple the true cellular response of the perturbation from the influence of infection bias. It models the distribution of perturbed cells and iteratively finds the maximum likelihood of cell cluster proportions as well as the real cellular response for each gRNA. We demonstrated its performance on a series of simulation experiments. By applying scDecouple to real CROP-seq data, we found that scDecouple could enhance biological discovery by detecting perturbation-related genes more critically. It helps to better study gene function and identify disease targets via scCRISPR-seq, especially with heterogeneous samples or complex gRNA libraries.

Introduction

With the development of single-cell technology, there emerge a group of CRISPR screening methods named scCRISPR-seq[1] that adopt CRISPR to perturb a set of genes and then assess the resulting profiles of each perturbation by single-cell sequencings, such as Perturb-seq[2–4], CROP-seq[5], CRISP-seq[6], Mosaic-seq[7], Spear-ATAC[8], and CRISPR-sciATAC[9]. Usually, a group of guide RNAs (gRNAs) targeting different genes are packed into a pool of lentivirus and then delivered into each cell. These gRNAs each introduces perturbation to pools of cells. Then, single-cell sequencing[10-14] is used to measure one or more types of profiles for each cell, like single-cell RNA sequencing (scRNA-seq)[11,12], single-cell ATAC sequencing (scATAC-seq)[10,13], and Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq)[14]. These scCRISPR-seq methods can perform high-throughput perturbations as well as data-rich read-outs for each perturbation, providing informative data for gene regulation study[2,15], disease target identification[4], and drug development[16].

One of the basic tasks for analyzing scCRISPR-seq is to obtain the exact effects for each perturbation. Noise and uncertainties exist in several steps of scCRISPR-seq protocols[2,17,18]

like editing efficiency, off-target effects, and single-cell sequencing noise, bringing huge challenges to data analysis. One key issue among them is that the exact original expression profile of each perturbed cell cannot be measured directly but must be estimated. In typical experimental settings, a control group is introduced to estimate the profiles of cells before perturbations[2,15,19]. It can be either a group of unperturbed cells or groups of cells infected by non-targeting (NT) gRNAs. However, such methods are not valid enough to deal with scCRISPR-seq experiments on samples composed of different cell clusters. We observed that the cell cluster proportion may vary a lot between different gRNAs in actual experiments[2,4]. This could be due to the different infection efficiencies of gRNAs or growth rates across different cell clusters. Besides, the random effects in gRNA infection and sampling during sequencing introduce additional noise to the estimation of the expression profiles before perturbation, especially when the gRNA library size is large. These will all result in a different proportion of the real infected cell clusters compared to the control group. Here, we refer to this type of noise as infection bias. With the development of scCRISPR-seq, the heterogeneity of samples[4] and the complexity of gRNA libraries[15] rapidly increases, making the above problems even more nonnegligible. A bioinformatics tool is urgently needed to decouple the true cellular response and infection bias in scCRISPR-seq.

Here, we developed scDecouple to decouple the cellular response and bias by solving the maximum likelihood estimation of the original ratios of cell clusters for each gRNA. We modeled the distribution of cells in the principal component (PC) space as a Gaussian mixture model and used the expectation-maximization algorithm to iteratively estimate each cell's original cluster and real cellular response to perturbation. We conducted a series of simulations on generated data and real datasets to investigate the performance with different settings on parameters, including the number of cell clusters, the strength of the perturbation effect, and the strength of bias. scDecouple performed well on all simulation data. Then we applied our method to CROP-seq[5]. Results show that scDecouple can receive more precise cellular responses, more significant pathways, and more reasonable gene ranks. scDecouple helps to better understand perturbation effects and provides support for more complicated scCRISPR-seq protocols in the future.

Description of problem

We map all cells to the PC dimension. Assuming that experimental cells X formed into two clusters with mean $\boldsymbol{\mu}_X$ on PC space(Fig. 1A). Usually, the control group is set by NT gRNAs, which means the mean of control groups Z and original experimental cells X are approximately equal. Because our observation is based on the control group, we assumed that $\boldsymbol{\mu}_Z = \boldsymbol{\mu}_X$ and used $\boldsymbol{\mu}_Z$ to represent the original state. The gRNA1 successfully infected some cells with an uneven selection of clusters (Fig. 1B). We used $\boldsymbol{\mu}_S$ to represent the mean of selected cells before perturbation. Assuming that a perturbation carried by gRNA1 can make cells tend to move in a certain direction in the PC space. After perturbation, the mean of cells turns to $\boldsymbol{\mu}_Y$ (Fig. 1C). Here, $(\boldsymbol{\mu}_Y - \boldsymbol{\mu}_S)$ represents the cellular response, which represents the real perturbation effects. $(\boldsymbol{\mu}_S - \boldsymbol{\mu}_Z)$ represents the bias caused by gRNA1 imbalanced selections. Then the observed change $(\boldsymbol{\mu}_Y - \boldsymbol{\mu}_Z)$ contains two parts: infection bias from $\boldsymbol{\mu}_z$ to $\boldsymbol{\mu}_s$ and true cellular response from $\boldsymbol{\mu}_s$ to $\boldsymbol{\mu}_Y$.

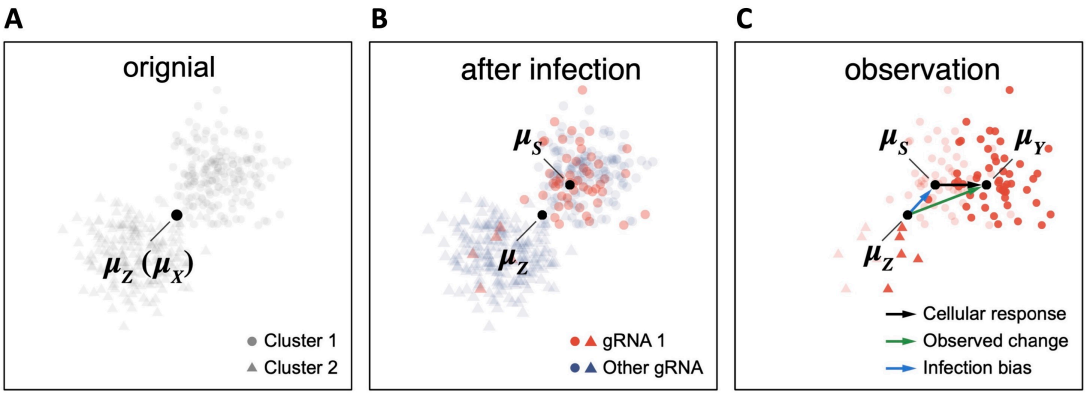

Figure 1 (A-C) The diagram of the research problem. Cells from the original group (A) are partially infected by gRNA1 and the mean of infected cells differs from the original group's (B). So the observed change contains two parts: cellular response and infection bias (C).

The general pipeline of computation

scDecouple contains four steps: data preprocessing, PC selection, decoupling, and analysis (Fig. 2A). First, cells are normalized and log-transformed. Variable genes are selected and transformed to PC space. Second, PCs with high multimodality and explained variance are selected. The multimodality is defined by dip statistic (Methods). Third, the decoupling process is performed on selected PCs to decouple observed changes after the establishment of the gaussian mixture model of control and perturbation groups (Fig. 2B). Finally, we calculated the cellular response on other PCs by observed FC and performed inverse transformation on all PCs to estimate cellular responses per gene, followed by pathway enrichment and perturbation-related gene ranking. The description of the models and details of the essential steps are shown in Methods.

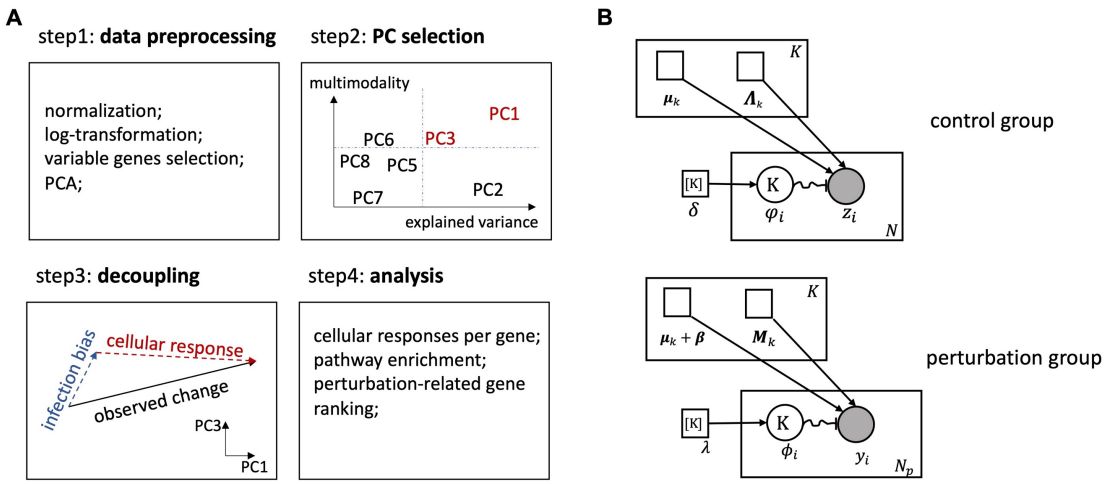

Figure 2 (A) Four steps in scDecouple: data preprocessing, PC selection considering multimodality and explained variance, decoupling observed change to two parts, following analysis. (B) the plate notation of our models. We used two gaussian mixture models to

estimate control and perturbation groups. Here, smaller squares represent fixed parameters: the cluster proportion. [K] means there are K clusters. Larger circles represent random variables and filled-in means known values. The directed edges between variables indicate dependencies between the variables and the squiggly line with a crossbar indicates the value selects from upstream variables.

Simulation using generated Gaussian mixture data

We first randomly generated 1,000 2-dimensional NT cells following two-cluster GMM (Fig. 3), whose bimodality concentrates on the first dimension (PC1). We inferred the bias of the FC method when there are two clusters (Methods Formula 9):

$$\sum_{k=1,2} (\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)\boldsymbol{\mu}_k = (\boldsymbol{\lambda}_1 - \boldsymbol{\delta}_1)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{1}$$

The bias is linearly correlated with two sections: the distance d between two clusters and cluster1's ratio difference r between the control and perturbation groups.

$$\begin{aligned} d &= abs(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ r &= abs(\boldsymbol{\lambda}_1 - \boldsymbol{\delta}_1) \end{aligned} \tag{2}$$

We first fixed the ratio changes $r = 0.1$ and changed the cluster distance d (Fig. 3A-E), and then fixed $d = 3$ and changed r (Fig. 3F-J). We sampled 100 times for each parameter and evaluated the estimation of cellular response for each sampling. We used Fold Change (FC) and scDecouple to get cellular responses separately. When d or r is small, the FC method and scDecouple both have low loss (Fig. 3B and Fig. 3G). With the increase of data multimodality or ratio change, the FC method gets more loss while scDecouple still performs well. In the dimension PC2 in which data is unimodal, two methods get similar results (Fig. 3D and Fig. 3I). Also, results showed that scDecouple can detect the multimodality of data (Fig. 3E, Fig. 3H and Fig. 3J) and is very sensitive to multimodality changes (Fig. 3C).

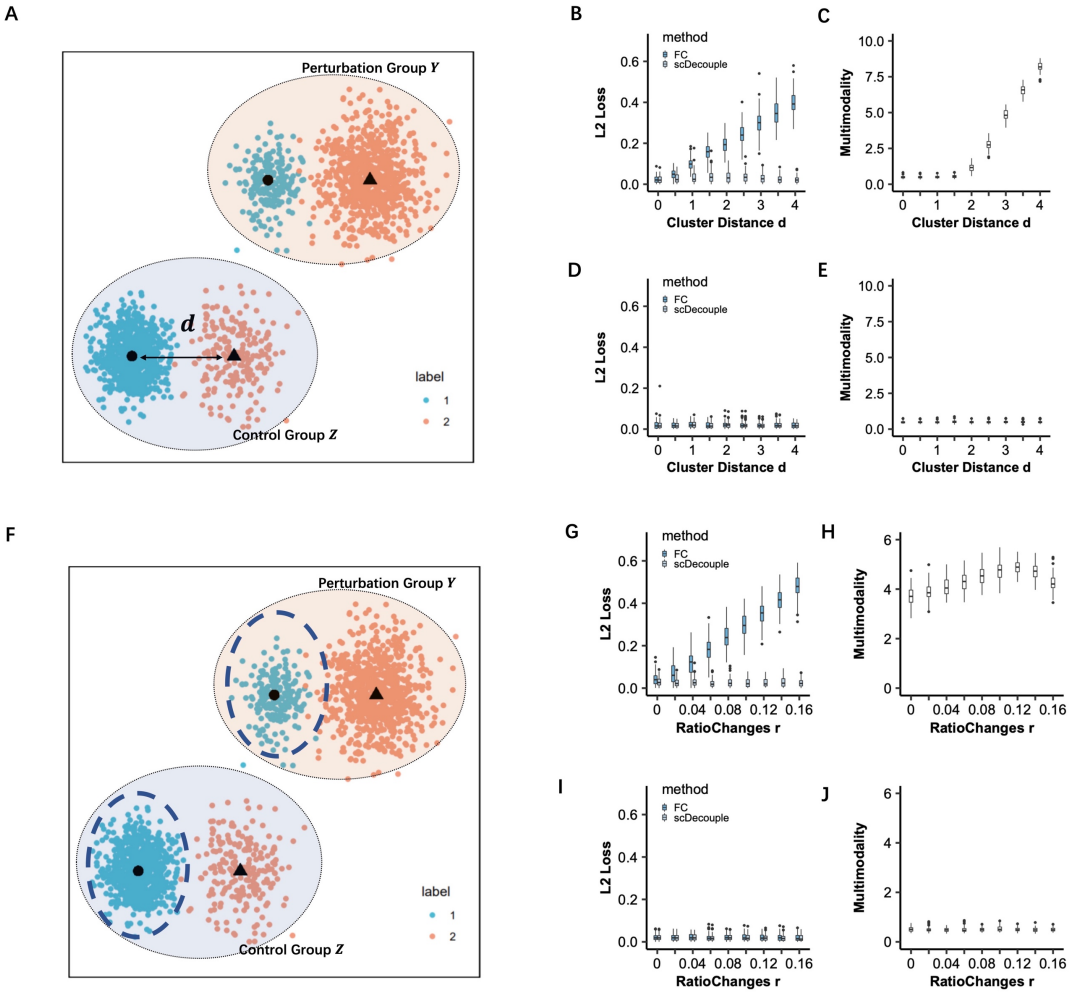

Figure 3 (A-E) The diagram **(A)** and results **(B-E)** of simulations with cluster distance changes. **(F-J)** the diagram **(F)** and results **(G-J)** of simulations with cluster ratio changes.

Simulation using genome-scale perturb-seq data

We further simulated on a real Perturb-seq dataset. It targeted more than 2,000 common essential genes in K562 and RPE1 cells. Their gRNAs targeted the same genes in two cell lines, but the infection numbers of these gRNAs varied a lot (Fig. 4A). Here, we combined two cell lines as two clusters of one experiment to do simulations. We use the NT gRNAs as the control group and selected gRNAs with similar effects across two cell types as target gRNAs. The infections of each cell type varied among all selected gRNAs, which caused the infection bias (Fig. 4B). According to the explained variance and modality score that scDecouple calculated, we selected PC1 to perform decoupling (Fig. 4C). We drew the distribution of control cells on selected PC and validated that it had two clusters (Fig. 4D). We used the observed FC and scDecouple separately to estimate the cellular response of each gene to each gRNA, and calculated the mean absolute error of each gene per gRNA (Fig. 4E). We also considered the infection ration between K562 and RPE1 for each gRNA. As shown in Fig. 4E, directly using FC introduced more bias to the estimations, especially when the current gRNA had great cluster ratio changes during infections. We further randomly selected

several gRNAs with different ratio changes to see the variance of estimations among genes (Fig. 4F). The cellular response estimations obtained by FC had higher variance and mean value than scDecouple estimation, especially when ratio changes are large. In general, results showed that scDecouple can help to get a more accurate cellular response and reduces the bias caused by gRNA infections.

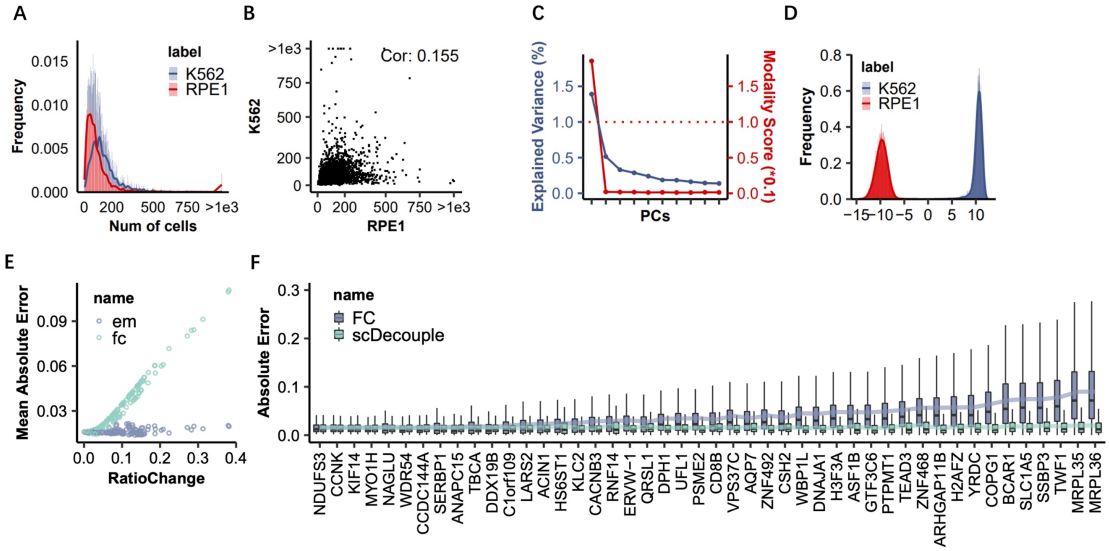

Figure 4 (A) the distribution of infection numbers among gRNAs in two cell types. **(B)** the infection numbers of each gRNA in two cell types. Each dot represents one gRNA. **(C)** The selection of PCs. Only PC1 has high variance and multimodality. **(D)** the distribution of the control group on PC1. It has two clusters. **(E)** The error of cellular response estimation using FC (green) and scDecouple (blue) on PC1. Dots are gRNAs and are sorted by cluster ratio changes. **(F)** The box plot of cellular response estimation errors across randomly selected gRNAs. The X-axis is gRNA and Y-axis is the error of genes. The gRNAs are sorted by cluster ratio changes.

Application to CROP-seq data

To further evaluate the performance of scDecouple, we applied it to a CROP-seq dataset. The CROP-seq dataset targeted 23 genes on human Jurkat cells to study the T cell receptor (TCR) signaling pathway. We used NT gRNAs to generate the control group. We first normalized the library size and log-transformed data matrix. Then, 700 highly variable genes were selected and transformed into PC space. We selected 3 PCs by the threshold of explaining variance and modality score (Fig. 5A). We plotted the cell distribution of the control group and two perturbation groups on two of the selected PCs (Fig. 5B), and found that the cell proportions of two perturbation gRNAs differ from the control group. It proved that selection bias really exists in real scCRISPR-seq datasets. We then decoupled observed changes on selected PCs and got the cellular response of each gene to each gRNA as well as the selection bias. Fig. 5C shows the inferred original cluster proportions of each gRNA. It varied among gRNAs and reflected the strength of infection bias. The decoupled cellular responses were represented by a gene-gRNA matrix, whose columns were gRNAs and rows were 165 TCR pathway

signature genes defined by CROP-seq paper[5] (Fig. 5D).

To further evaluate the results, we calculated the enrichment score of the TCR signaling pathway for each gRNA (Fig. 5E). The differential expression genes (DE genes) were selected by observed FC and scDecouple separately. The results showed that scDecouple gets similar or higher enrichment scores than the FC method. The improvement of estimation was not that large mainly because the heterogeneity of cells in this dataset was small. We also calculated the ranking of TCR-related essential genes. According to the results in the original paper[5], we further filtered TCR pathway signature genes and selected the top 60 essential genes to calculate their response ranks among all genes. Because all gRNAs were targeted on the TCR signaling pathway, these essential genes should rank high. We combined their ranks calculated by FC and scDecouple. As shown in Fig. 5F, scDecouple improved the ranking of essential genes. We then focused on CD69, a well-known early activation marker of the TCR pathway[5,20,21]. Results showed that the rankings of CD69 were greatly improved (Fig. 5G).

From the analysis of CROP-seq data, we can see that the observed changes can be well decoupled to infection bias and cellular response using scDecouple, resulting in more significant pathway enrichment results and more accurate gene ranking. This will benefit in deriving more precise gene regulatory networks and perturbation-related genes.

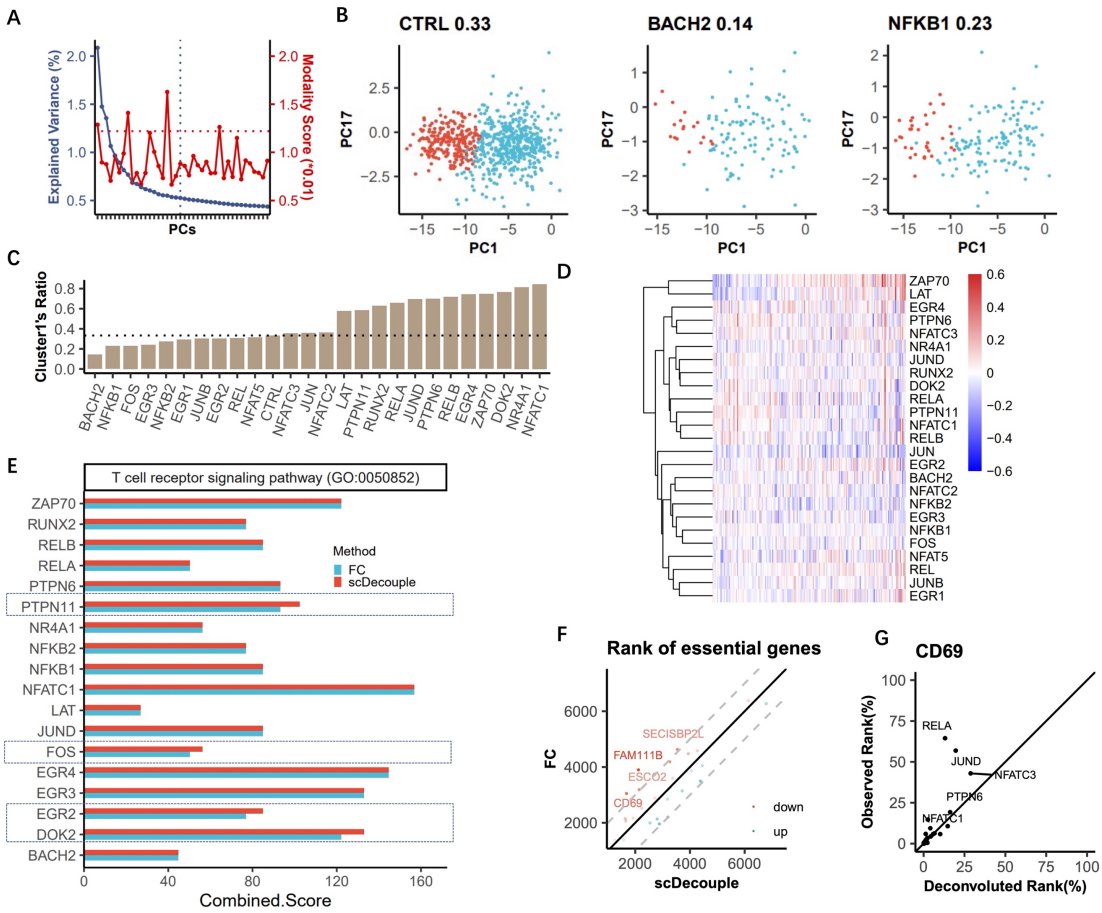

Figure 5 (A) The selection of PCs. We select 3 PCs according to explained variance and modality score. **(B)** The distribution and cluster proportion of three groups on two selected

PCs. The left is control group and the others are perturbation groups. The value is the ratio of red cluster. **(C)** The ratio of cluster 1 among all gRNAs. The dashed line is the ratio of control group. **(D)** Heatmap of gRNA-gene matrix. The rows are gRNAs and columns are TCR pathway-related signature genes. **(E)** Enrichment score of TCR pathway. Each row represents one gRNA. **(F)** The ranks of essential genes calculated by FC and scDecouple. Each dot represents one gene. The red and blue colors indicate increased or decreased ranking in scDecouple compare with FC. The alpha represents the intensity of ranking changes. The dashed line is 1K ranking changes. We labeled the genes with more than 1K ranking changes. **(G)** The rank of CD69 among all gRNAs. The axes represent the rank percentage of observed and deconvoluted responses.

Applications on drug screening

scDecouple can be applied to other scenarios for decoupling proportion changes of cell clusters and cellular response. One typical scenario is drug screening. Batch effects usually exist in drug screening due to the technical variations or other non-biological differences between the measurement of control groups and drug-treated groups. It can be considered as linear shifting on PC space, which behaves like a cellular response to the batch. Also, the influence of drugs usually causes changes in cell proportions. Thus, the observed change contains both expected cellular proportion changes and unexpected batch effects. Here, scDecouple can be used to decouple observed change to get the drug response without batch effect (Fig. 6A).

Here, we used the data of etoposide drug screening on human brain tumor tissue[22] as an example. They performed drug perturbation on tumor slices and then used scRNA-seq to profile transcription-level drug responses. We selected the first two PCs based on their variance and multimodality. The control and perturbation groups show batch effects on selected PCs (Fig. 6B). We applied scDecouple and annotate each cluster by malignancy score and marker genes (Fig. 6C), which are both defined by the original paper[22]. and the results showed that the decoupled proportion changes focus on the decrease of tumor cells and increase of myeloid cells (Fig. 6D). The finding is consistent with previous studies[22–24], which indicates the feasibility of scDecouple on drug screening datasets.

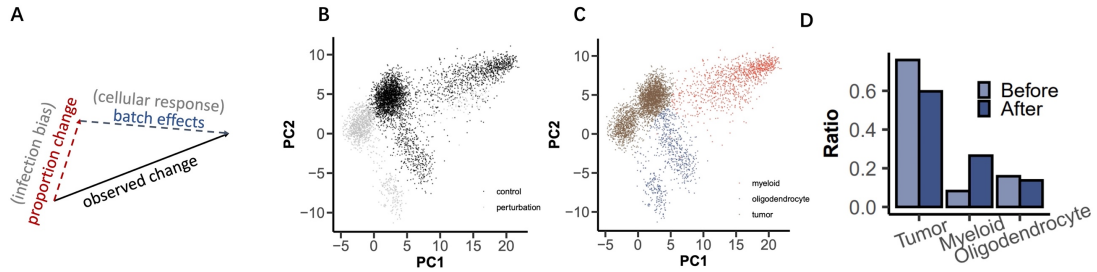

Figure 6 (A) The observed change can be decoupled into proportion change and batch effects. The red line is the signal and the blue line is the noise. **(B)** The batch effects between control group (grey) and perturbation group (black) on selected PC space. **(C)** The assigned three clusters of each cell: myeloid cells (red), oligodendrocyte cells (blue) and tumor cells

(brown). **(D)** The inferred ratios of three clusters before and after drug perturbation.

Discussion

scDecouple decoupled observed changes of scCRISPR-seq data on PC space to the cellular response and infection bias based on the maximum likelihood estimation. We verified its performance on a series of simulations and applied scDecouple to real datasets. It gets more obvious pathway enrichment and makes perturbation-related genes rank higher. scDecouple can also be extended to drug-screening datasets to reduce batch effects. The good performance of scDecouple is mainly contributed by its estimation of the real cluster proportions for cells in perturbation groups.

scDecouple is the first method that focuses on infection bias on scCRISPR-seq data. With the development of technology, scCRISPR-seq with more complex gRNA libraries and more heterogeneous cells will be developed and popularized. As a result, reducing infection bias will be more important and necessary in the following analysis of scCRISPR-seq. Also, the process of double-strand breaking causes P53 pathway activation and cell state arrest, which impacts the downstream analysis especially projects related to cell states and aging. Our method helps to discard all these impacts and focus on the cellular responses we are concerned about. In the future, we may not need to set control groups. We can use the public single-cell atlas datasets such as hECA[25] to get information of cell clusters instead of using control groups to estimate.

scDecouple has two assumptions, one is that cells obey GMM on PC space, and the other is that the number of clusters is the same between the NT and perturbation groups. Later, scDecouple can be extended to other high-dimensional space or statistical models. Also, more information such as marker genes of cell clusters can be added to deal with perturbation groups that lose one or more clusters compared with the control group.

Acknowledgments

This work is supported in part by National Key R&D Program of China grant 2021YFF1200900, NSFC grants 62250005, 61721003.

Author contributions

Conceptualization, X.Z. and Y.L.; Methodology and Investigation, Q.M.; Writing – Original Draft, Q.M.; Writing – Review & Editing, Q.M., M.S., L.W., Y.L. and X.Z.; Supervision, X.Z., Y.L.; Funding Acquisition, X.Z., Y.L.

Reference

1. Bock C, Datlinger P, Chardon F, et al. High-content CRISPR screening. *Nat Rev Methods Primer*. 2022;2(1):8. doi:10.1038/s43586-021-00093-4

2. Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*. 2016;167(7):1853-1866.e17. doi:https://doi.org/10.1016/j.cell.2016.11.038

3. Adamson B, Norman TM, Jost M, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*. 2016;167(7):1867-1882.e21. doi:https://doi.org/10.1016/j.cell.2016.11.048

4. Jin X, Simmons SK, Guo A, et al. In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*. 2020;370(6520):eaaz6063. doi:10.1126/science.aaz6063

5. Datlinger P, Rendeiro AF, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods*. 2017;14:297. doi:10.1038/nmeth.4177 https://www.nature.com/articles/nmeth.4177#supplementary-information

6. Jaitin DA, Weiner A, Yofe I, et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*. 2016;167(7):1883-1896.e15. doi:https://doi.org/10.1016/j.cell.2016.11.039

7. Xie S, Duan J, Li B, Zhou P, Hon GC. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell*. 2017;66(2):285-299.e5. doi:10.1016/j.molcel.2017.03.007

8. Pierce SE, Granja JM, Greenleaf WJ. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat Commun*. 2021;12(1):2969. doi:10.1038/s41467-021-23213-w

9. Liscovitch-Brauer N, Montalbano A, Deng J, et al. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol*. 2021;39(10):1270-1277. doi:10.1038/s41587-021-00902-x

10. Lareau CA, Duarte FM, Chew JG, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol*. 2019;37(8):916-924. doi:10.1038/s41587-019-0147-6

11. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*. 2014;9(1):171-181. doi:10.1038/nprot.2014.006

12. Macosko EZ, Basu A, Satija R, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202-1214. doi:10.1016/j.cell.2015.05.002

13. Chen X, Miragaia RJ, Natarajan KN, Teichmann SA. A rapid and robust method for single cell chromatin accessibility profiling. *Nat Commun*. 2018;9(1):5345. doi:10.1038/s41467-018-07771-0

14. Simultaneous epitope and transcriptome measurement in single cells | Nature Methods. Accessed January 30, 2023. https://www.nature.com/articles/nmeth.4380

15. Replogle JM, Saunders RA, Pogson AN, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*. 2022;185(14):2559-2575.e28. doi:10.1016/j.cell.2022.05.013

16. McFarland JM, Paolella BR, Warren A, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat Commun*. 2020;11(1):4296. doi:10.1038/s41467-020-17440-w

17. Fu Y, Foden JA, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*. 2013;31(9):822-826. doi:10.1038/nbt.2623

18. Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun*. 2020;11(1):1169. doi:10.1038/s41467-020-14976-9

19. Satpathy AT, Granja JM, Yost KE, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol*. 2019;37(8):925-936. doi:10.1038/s41587-019-0206-z

20. Cibrián D, Sánchez-Madrid F. CD69: from activation marker to metabolic gatekeeper. *Eur J Immunol*. 2017;47(6):946-953. doi:10.1002/eji.201646837

21. Ziegler SF, Ramsdell F, Alderson MR. The activation antigen CD69. *Stem Cells*. 1994;12(5):456-465. doi:10.1002/stem.5530120502

22. Zhao W, Dovas A, Spinazzi EF, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Med*. 2021;13(1):82. doi:10.1186/s13073-021-00894-y

23. Montecucco A, Zanetta F, Biamonti G. Molecular mechanisms of etoposide. *EXCLI J*. 2015;14:95-108. doi:10.17179/excli2015-561

24. Etoposide, Topoisomerase II and Cancer | Bentham Science. Accessed January 30, 2023. https://www.eurekaselect.com/article/35148

25. Chen S, Luo Y, Gao H, et al. hECA: The cell-centric assembly of a cell atlas. *iScience*. 2022;25(5). doi:10.1016/j.isci.2022.104318

Supplemental information

Methods

Gaussian Mixture Model

First, we estimated control group Z using the Gaussian Mixture Model (GMM), which is a linear superposition of Gaussian components, on its PC space. The density function is

$$\begin{aligned} p(\boldsymbol{z}) &= \sum_{k=1}^{K} \delta_k \mathcal{N}\left(\boldsymbol{z} \middle| \boldsymbol{\mu}_k,\ \boldsymbol{\Lambda}_k^{-1}\right) \\ &= \sum_{k=1}^{\mathrm{K}} \delta_k \left|\frac{\boldsymbol{\Lambda}_k}{2\pi}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_k)^T \boldsymbol{\Lambda}_k (\boldsymbol{z}-\boldsymbol{\mu}_k)\right] \end{aligned} \tag{1}$$

Where K is the number of mixture components, refer to the number of cell clusters, D is the number of selected PCs. $\delta_k (k = 1,2,\cdots,K)$ is the mixture coefficient or mixture weight of cluster k. It's the prior probability of each subgroup and $\sum_{k=1}^{K} \delta_k = 1$. $\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k$represents the expectation and precision (inverse of the variance) of cluster k.

Here, we use perturbation group Y to represent cells affected by some gRNA. The mean of each cluster in Y represents by the sum of the corresponding cluster mean in control cells $\boldsymbol{\mu}_k$ and cellular responses $\boldsymbol{\beta}$. The density function of Y is:

$$\begin{aligned} p(\boldsymbol{y}) &= \sum_{k=1}^{K} \lambda_k \mathcal{N}\left(\boldsymbol{y} \middle| \boldsymbol{\mu_k} + \boldsymbol{\beta}, \boldsymbol{M}_k^{-1}\right) \\ &= \sum_{k=1}^{K} \lambda_k \left|\frac{\boldsymbol{M}_k}{2\pi}\right|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-(\boldsymbol{\mu}_k+\boldsymbol{\beta}))^T \boldsymbol{M}_k \left(\boldsymbol{y}-(\boldsymbol{\mu}_k+\boldsymbol{\beta})\right)\right] \end{aligned} \tag{2}$$

where λ_k is the mixture coefficient of Y, it reflects the cell proportion of the perturbation group. $\boldsymbol{M}_k$ represents the precision of cluster k. The plate notations of two gaussian mixture models are shown in Fig. 3B.

Inference with Expectation maximization algorithm

The likelihood function of all observed data is:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{z},\boldsymbol{y}) &= L(\boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{M}|\boldsymbol{z},\boldsymbol{y}) \\ &= \prod_{j=1}^{N_z} p\left(\boldsymbol{z}_j \middle| \boldsymbol{\delta},\boldsymbol{\mu},\boldsymbol{\Lambda}\right) \prod_{i=1}^{N_y} p(\boldsymbol{y}_i|\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{M}) \end{aligned} \tag{3}$$

It contains two parts: the likelihood of control groups L_c and the likelihood of treatment group L_t. Namely:

$$\begin{cases} L_c = \prod_{j=1}^{N_z} p(\boldsymbol{z}_j | \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ L_t = \prod_{i=1}^{N_y} p(\boldsymbol{y}_i | \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{M}) \end{cases} \tag{4}$$

Where, N_y, N_y represents the number of cells in the control group and treatment group.

We first use the EM algorithm to maximum L_c and estimate parameters $\widehat{\boldsymbol{\delta}}$, $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\Lambda}}^{-1}$:

$$\begin{aligned} \max L_c &= \max \prod_{i=1}^{N_c} p(\boldsymbol{z}_i | \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= \max \prod_{i=1}^{N_c} \prod_{k=1}^{K} \left[p(u_i = k | \boldsymbol{\delta}) p\left(\boldsymbol{z}_i | u_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right) \right]^{I(u_i = k)} \end{aligned} \tag{5}$$

For L_t:

$$\max L_t = \max \prod_{j=1}^{N_t} p\left(\boldsymbol{y}_j | \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{M},\right) \tag{6}$$

We already estimate the parameters $\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}^{-1}$. So, we only estimate$\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{M}}$ and during EM iteration:

Expectation step (E-step):

$$\mathrm{E}\left[v_j = k \middle| \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\beta}}\right] = \frac{\lambda_k |\boldsymbol{M}_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\beta}}_k\right)\right)^T \boldsymbol{M}_k \left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\beta}}_k\right)\right)\right\}}{\sum_{l=1}^{K} \lambda_l |\boldsymbol{M}_l|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_l + \widehat{\boldsymbol{\beta}}_l\right)\right)^T \boldsymbol{M}_l \left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_l + \widehat{\boldsymbol{\beta}}_l\right)\right)\right\}} \tag{7}$$

Maximum step (M-step):

$$\begin{aligned} \hat{\lambda}_k &= \frac{1}{N_t} \sum_{j=1}^{N_t} \mathrm{E}\left[v_j = k\right] \\ \widehat{\boldsymbol{\beta}} &= \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{j=1}^{N_t} \mathrm{E}\left[v_j = k\right]\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_k\right)}{\sum_{s=1}^{N_t} \mathrm{E}\left[v_j = k\right]} \\ \widehat{\boldsymbol{M}}_k &= \frac{\sum_{j=1}^{N_t} \mathrm{E}\left[v_j = k\right]\left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\beta}}_k\right)\right)\left(\boldsymbol{y}_j - \left(\widehat{\boldsymbol{\mu}}_k + \widehat{\boldsymbol{\beta}}_k\right)\right)^T}{\sum_{i \in N_p} \mathrm{E}\left[v_j = k\right]} \end{aligned} \tag{8}$$

By iteratively applying E-step and M-step, we can maximize the likelihood function and get $\widehat{\boldsymbol{\lambda}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{M}}$. In some case, we can choose to fix the estimation of $\boldsymbol{M}_k$ to make the shape of clusters in perturbation groups be similar with control groups:

$$\widehat{\boldsymbol{M}}_k = \widehat{\boldsymbol{\Lambda}}_k, k = 1,2,\cdots,K$$

Algorithm acceleration

To accelerate the calculation and to reduce noise, we divide candidate PCs into four

categories: multimodal and high-variance, unimodal and high-variance, multimodal and low-variance, unimodal and low-variance.

In unimodal PCs, Fold-Change estimation (FC estimation) is unbiased because there are no clusters. In multimodal PCs, bias using FC estimation is:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \frac{1}{N_t}\sum_{j=1}^{N_t}\boldsymbol{y}_j - \frac{1}{N_c}\sum_{i=1}^{N_c}\boldsymbol{z}_i \\ &= \sum_{k=1}^{K}\boldsymbol{\lambda}_k(\boldsymbol{\mu}_k + \boldsymbol{\beta}_k) - \sum_{k=1}^{K}\boldsymbol{\delta}_k\boldsymbol{\mu}_k \\ &= \boldsymbol{\beta} + \sum_{k=1}^{K}(\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)\boldsymbol{\mu}_k \\ \mathbf{Bias} &= \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \sum_{k=1}^{K}(\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)\boldsymbol{\mu}_k\end{aligned} \tag{9}$$

And when PCs are multimodal and low-variance, the bias using FC estimation is very small and there are few spaces for improvement when considering multimodality using the GMM model. The bias approaches zero:

$$\sum_{k=1}^{K}(\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k)\boldsymbol{\mu}_k \approx \overline{\boldsymbol{\mu}}\sum_{k=1}^{K}(\boldsymbol{\lambda}_k - \boldsymbol{\delta}_k) = 0 \tag{10}$$

Therefore, only multimodal PCs with high variance need to be decoupled. The degree of multimodality of PCs is measured by dip statistic, which calculates the maximum difference between the empirical distribution function and the best-fitting unimodal distribution.

Define:

$$\begin{aligned}\rho(F,G) &= \sup_x |F(x) - G(x)| \\ \rho(F,\mathcal{A}) &= \inf_{G\in\mathcal{A}} \rho(F,G)\end{aligned} \tag{11}$$

Let $\mathcal{U}$ be the class of unimodal distribution functions.

The dip of a distribution function F is then defined by:

$$D(F) = \rho(F,\mathcal{U}) \tag{12}$$

We used dip statistic to assess multimodality and select PCs.