Nucleic Acids Research, 2025, **53**, D1099–D1111
https://doi.org/10.1093/nar/gkae858
Advance access publication date: 8 October 2024
Database issue

PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization

Zhiting Wei[1,2,†], Duanmiao Si[1,2,†], Bin Duan[1,2,†], Yicheng Gao[1,2,†], Qian Yu[3], Zhenbo Zhang[2,*], Ling Guo[3,*] and Qi Liu[1,2,3,4,*]

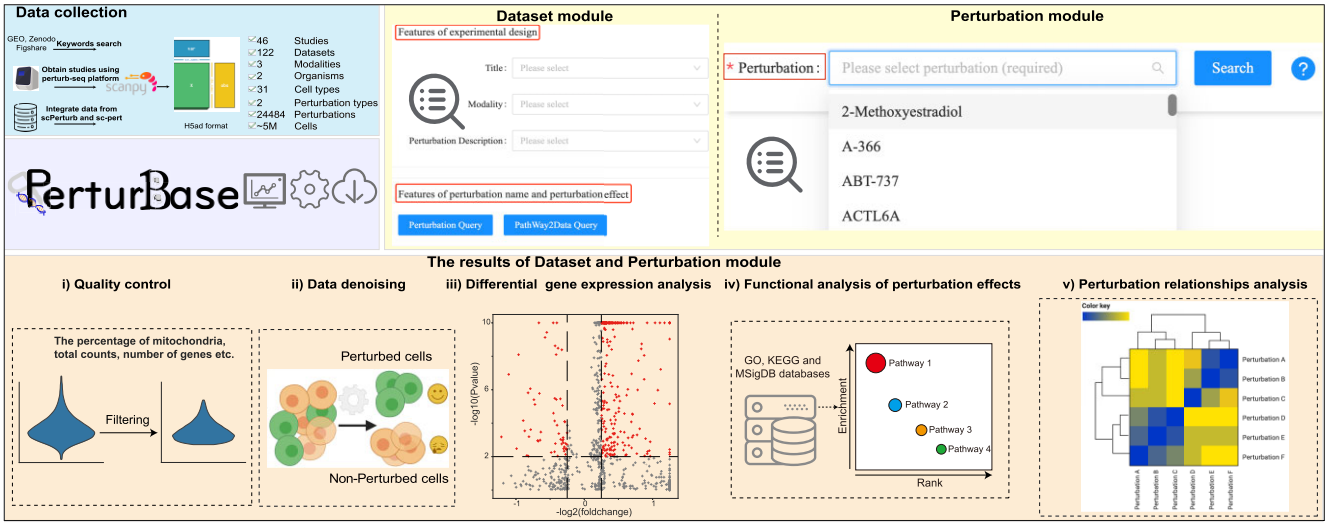
[1]State Key Laboratory of Cardiology and Medical Innovation Center, Shanghai East Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai 200092, China
[2]Reproductive Medicine Center, Department of Obstetrics and Gynecology, Tongji Hospital, School of Medicine, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai 200092, China
[3]Zhejiang Lab, Kechuang Avenue, Zhongtai Subdistrict, Yuhang District, Hangzhou 311121, China
[4]Shanghai Research Institute for Intelligent Autonomous Systems, 55 Chuanhe Road, Shanghai 200092, China

[*]To whom correspondence should be addressed. Tel: +86 021 65980296; Email: qiliu@tongji.edu.cn
Correspondence may also be addressed to Ling Guo. Email: guoling@zhejianglab.com
Correspondence may also be addressed to Zhenbo Zhang. Email: zhangzhenbozzb@tongji.edu.cn
[†]The first four authors should be regarded as Joint First Authors.

Abstract

Single-cell perturbation (scPerturbation) sequencing techniques, represented by single-cell genetic perturbation (e.g. Perturb-seq) and single-cell chemical perturbation (e.g. sci-Plex), result from the integration of single-cell toolkits with conventional bulk screening methods. These innovative sequencing techniques empower researchers to dissect perturbation effects in biological systems at an unprecedented resolution. Despite these advancements, a notable gap exists in the availability of a dedicated database for exploring scPerturbation data. To address this gap, we present PerturBase, the most comprehensive database designed for the analysis and visualization of scPerturbation data (http://www.perturbase.cn/). PerturBase curates 122 datasets from 46 publicly available studies, covering 115 single-modal and 7 multi-modal datasets that include 24 254 genetic and 230 chemical perturbations from approximately 5 million cells. The database, comprising the 'Dataset' and 'Perturbation' modules, provides insights into various results, encompassing quality control, denoising, differential gene expression analysis, functional analysis of perturbation effects and characterization of relationships between perturbations. All the datasets and results are presented on user-friendly, easy-to-browse web pages and can be visualized through intuitive and interactive plot and table formats. In summary, PerturBase stands as a pioneering, high-content database intended for searching, visualizing and analyzing scPerturbation datasets, contributing to a deeper understanding of perturbation effects.

Graphical abstract



Received: June 1, 2024. **Revised:** September 10, 2024. **Editorial Decision:** September 11, 2024. **Accepted:** September 19, 2024

Introduction

Perturbation experiments attempt to establish causal links between perturbations and responses, which can be broadly categorized into two classes: genetic perturbations and chemical perturbations. Perturbation-based omics has become a powerful tool for studying gene functions and a cornerstone of the pharmaceutical drug discovery pipeline (1–3). For example, the Library of Integrated Network-based Cellular Signatures database piloted by the Broad Institute has become a rich data source for genetic and chemical perturbation, and clustered regularly interspaced short palindromic repeat (CRISPR) screening techniques have been broadly used for drug target identification and drug resistance research (4–6). Nonetheless, conventional perturbation technology has at least two major limitations. First, the readout is generally restricted to gross cellular phenotypes, e.g. proliferation, morphology or a highly specific molecular readout. Second, even in conjunction with more comprehensive molecular phenotyping methods, such as next-generation sequencing, a limitation of bulk assays is that cells ostensibly of the same 'type' can exhibit heterogeneous responses (7–9).

Single-cell transcriptome sequencing (scRNA-seq) represents a form of high-content molecular phenotyping that, when combined with conventional perturbation technology, can overcome both limitations. In 2016, single-cell CRISPR (scCRISPR), which couples CRISPR screening and scRNA-seq to enable pooled genetic screens at large-scale single-cell resolution, was developed (10–12). The key technical innovation of scCRISPR is the creative design of the lentiviral vector to allow the identification of the sgRNA in each cell by sequencing. Moreover, in 2019, sci-Plex, a method that couples chemical screening with scRNA-seq to cost-effectively quantify transcriptional responses to hundreds of chemicals in parallel, was proposed by Srivatsan *et al*. (9). In contrast to traditional perturbation screening, single-cell perturbation (scPerturbation) allows high-content phenotypes to be obtained, thus facilitating the dissection of complex effects of genes and chemicals in heterogeneous cell populations.

Currently, numerous alternative scPerturbation platforms have emerged. Based on readout omics, these platforms can be classified into four primary categories: transcriptome-, epigenome-, proteome- and imaging-based platforms. The mainstream scPerturbation platforms are transcriptome-based platforms that combine screens with scRNA-seq, such as Perturb-seq and sci-Plex (9–13). Transcriptome-based platforms have evolved rapidly, with innovations such as CROP-seq optimizing Perturb-seq vector design and reducing complexity and cost (14). Genome-scale Perturb-seq, introduced by Replogle *et al*., enables unbiased and comprehensive profiling of genome-scale genetic perturbations affecting 9867 genes (15). Moreover, by applying the technique to multi-omics data simultaneously, multi-modal scPerturbation was developed. In 2019, Rubin *et al*. developed an epigenome-based scPerturbation named Perturb-ATAC, which combines CRISPR interference or knockout with chromatin accessibility profiling in single cells based on the simultaneous detection of CRISPR guide RNAs and open chromatin sites by assay of transposase-accessible chromatin with sequencing (ATAC-seq) (16). Mimitou *et al*. developed ECCITE-seq, which allows simultaneous detection of transcriptomes, proteins, clonotypes and CRISPR perturbations from single-cell preparations (17). Recently, new multi-modal screening platforms called Perturb-map and Perturb-FISH, which combine CRISPR with imaging and spatial transcriptomics, have been developed to identify genetic determinants of tumor composition, organization and immunity (13,18).

scPerturbation is widely applied in various fields because of its powerful capabilities, including linking genotype to phenotype (11,12,15), dissecting genetic regulations and deciphering drug mechanisms. For example, Jaitin *et al*. revealed the effects of 22 transcription factors on the regulation of antiviral, inflammatory or developmental processes in lipopolysaccharide-stimulated bone marrow cells by CRISP-seq (12). Perturb-ATAC, Spear-ATAC and CRISPR-sciATAC could reveal epigenetic landscape remodelers in human B lymphocytes and leukemia cells (19,20). Using a perturbation map, Dhainaut *et al*. discovered that knockout *TGFB2* in lung cancer cells promoted tumor microenvironment remodeling and immune exclusion (18). In combination with sci-Plex, Srivatsan revealed substantial intercellular heterogeneity in response to specific chemicals and found that the main transcriptional responses to HDAC inhibitors involved cell cycle arrest (9). However, analyzing scPerturbation data presents significant challenges due to its inherent noise, which primarily stems from two main sources: (i) Cellular heterogeneity: Even within a supposedly homogeneous population, cells can exhibit significant variability in gene expression due to differences in cell cycle stage, metabolic state and microenvironmental influences. (ii) Variable perturbation efficiency: The efficiency of the perturbation can vary from cell to cell, leading to inconsistent biological responses. This inherent noise can lead to the identification of false positives or false negatives, reducing the statistical power. To address these challenges, a series of specialized bioinformatic methods for denoising have been developed, such as Mixscape for non-perturbed cell filtering and GSFA for latent component factor decomposition (21–23).

Despite the widespread use of scPerturbation, a significant gap remains in the availability of a dedicated database for exploring and querying scPerturbation data. Recently, scPerturb has been developed for scPerturbation data exploration; however, its utility is constrained by the absence of dedicated features for querying, visualizing and further interpreting the data (24). In addition, we acknowledge the limitation of the dataset selection window in scPerturb, which stops at 2021. This is a significant advantage of PerturBase, as it includes more recently generated and published datasets. By including the latest data, PerturBase ensures that users have access to the most up-to-date information, enhancing its utility and relevance in the field. To this end, we introduce PerturBase, the most comprehensive database that integrates 122 scPerturbation datasets from 46 publicly studies. The molecular readouts of these datasets include 115 single-modal datasets and 7 multi-modal datasets. Among these datasets, 101 datasets were subjected to genetic perturbations, whereas the remaining 21 datasets were subjected to chemical perturbations. A total of 113 of the 122 datasets were derived from *Homo sapiens* studies. Among the 122 datasets collected in PerturBase, 61 contained combinatorial perturbations (Supplementary Table S1). PerturBase features two modules: the 'Dataset' module and the 'Perturbation' module. The 'Dataset' module facilitates streamlined exploration of all 122 datasets, offering filters by organism, modality, perturbation type, perturbation name and perturbation effect. After selecting a dataset of interest, users can gain insights into

a range of analysis results, including (i) quality control, (ii) denoising, (iii) differential gene expression analysis, (iv) functional analysis of perturbation effects and (v) characterization of relationships between perturbations. Moreover, the 'Perturbation' module integrates a range of analysis results across datasets of a chosen perturbation, including (i) quality control, (ii) denoising, (iii) differential gene expression analysis and (iv) functional analysis of perturbation effects. These results provide a comparison of perturbations across various cellular contexts, offering valuable insights into their effects. In summary, PerturBase stands as the pioneering high-content database designed for the searching, visualization and analysis of scPerturbation data. Its extensive data repository and diverse functionalities make it an indispensable resource in the scPerturbation research community.

Materials and methods

Data collection

In our current study, the scPerturbation datasets were obtained primarily through three methods (Figure 1A). (i) We collected the scPerturbation data through a large-scale search in the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) (25), Zenodo (https://zenodo.org) and Figshare (https://figshare.com) using keywords such as 'perturb seq', 'high content crispr screening', 'single-cell crispr screening' and 'single-cell perturbation'. (ii) We identified 10 representative scPerturbation platforms in the scPerturbation field through comprehensive review articles, namely Perturb-seq (10,11), CRISP-seq (12), CROP-seq (14,26), Mosaic-seq (27), Perturb-ATAC (28), ECCITE-seq (29), direct-capture Perturb-seq (30), Direct-seq (31), TAP-seq (32) and SHARE-seq (33). Specifically, we referred to the articles 'Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution' by Cheng *et al*. (7) and 'High-content CRISPR screening' by Bock *et al*. (8). We subsequently obtained studies using single-cell CRISPR screening platforms by determining the articles citing the representative platforms through Google Scholar. The BioProject (34) accession numbers of the studies were retrieved using NCBI's eSearch application programming interface. We subsequently manually confirmed the presence of a scPerturbation dataset from the identified BioProject accession number. (iii) The scPerturbation data mentioned in scPerturb and sc-pert were also contained in PerturBase (24,35). We downloaded the raw data, and a uniform preprocessing approach defined by PerturBase was applied. In summary, PerturBase contains 122 scPerturbation datasets (Figure 1A and Supplementary Table S1). In terms of perturbation type, the collection encompasses 24 254 genetic and 230 chemical perturbations (Figure 1B, bottom right). The same gene perturbations across different datasets are counted once, whereas different doses of the same drug are considered separate chemical perturbations. The perturbation modality covers 115 single-modal data points and 7 multi-modal data points. In terms of species, it contains *H. sapiens* and *Mus musculus*. Notably, most perturbations are predominantly applied in a single dataset, particularly in the case of genetic perturbation (Figure 1B). The total number of cells per dataset is usually restricted by experimental limitations, although it has increased over time (Supplementary Figure S1). Therefore, there is a trade-off between the number of perturbations and the mean number of cells per perturbation in a dataset (Supplementary Figure S1). Other statistics regarding the PerturBase datasets are shown in Supplementary Figure S2.

Processing of the scPerturbation RNA-seq data

Assignment of perturbation to a cell

For the scCRISPR data, we allocated guide RNAs (gRNAs or sgRNAs, indicating the targeted gene in a cell) to cells based on two criteria. (i) For cells with sgRNA information already provided in the original study, we used these results directly. (ii) For cells where the original study included sgRNA count matrix data, we allocated sgRNAs using a threshold strategy, specifically considering an sgRNA valid if it had a minimum of five counts. Notably, our current method of setting a threshold at 5 is somewhat arbitrary, and a mixed distribution approach (e.g. as provided by 10X Genomics) or similar approach is warranted (Supplementary Note S1). The cells without sgRNA induction were used as blank controls, and the cells with nontargeting sgRNA (e.g. green fluorescent protein) were labeled 'CTRL' (negative control). For single-cell chemical perturbation data, we used the label information provided in the original study. The blank controls were filtered out for downstream data analysis.

Data quality control

For each scPerturbation dataset, we utilized the anndata and Scanpy Python packages to process it uniformly into the h5ad format (36). (i) Cells with <200 expressed genes, classified as blank controls or containing a large fraction of mitochondrial genes (over 10%) were filtered. (ii) Genes expressed in less than three cells were filtered. (iii) Datlinger *et al*. reported that at least 30 cells are required to capture each perturbation phenotype (14). Therefore, the perturbations, with the exception of the negative control with <30 perturbed cells (default), were not considered in PerturBase. Notably, if no perturbation meets these criteria, the dataset is not subjected to further analysis. (iv) PerturBase adopts the global scaling normalization method in Scanpy to scale the expression in each cell to 10 000, followed by natural logarithmic transformation. (v) After normalization, PerturBase adopts 'highly variable genes' (HVGs) with the default parameters in Scanpy to identify highly variable features for the scPerturbation data. To balance computational efficiency with data information, PerturBase maintains 4000 HVGs. If the raw data include <4000 genes, all the genes are retained; however, focusing solely on HVGs can filter out perturbation-specific differentially expressed genes (DEGs), potentially biasing downstream analyses, such as functional analyses (Supplementary Note S1). (vi) After that, PerturBase performs principal component analysis (PCA) to reduce data dimensionality (n_components = 50). (vii) After dimensionality reduction, PerturBase performs clustering based on the Leiden algorithm with default parameters (resolution = 1). (viii) For user convenience, we employed clusterProfiler to obtain gene symbols and Entrez and Ensembl IDs for genes in each dataset, making them readily accessible to users within their dataset of interest (37). Among the 98 RNA-seq datasets, 89 met the quality control criteria and were suitable for downstream data analysis.

Processing of the scPerturbation ATAC-seq data

Assignment of perturbation to a cell

The procedures parallel those of the scPerturbation RNA-seq in guide RNA assignment.

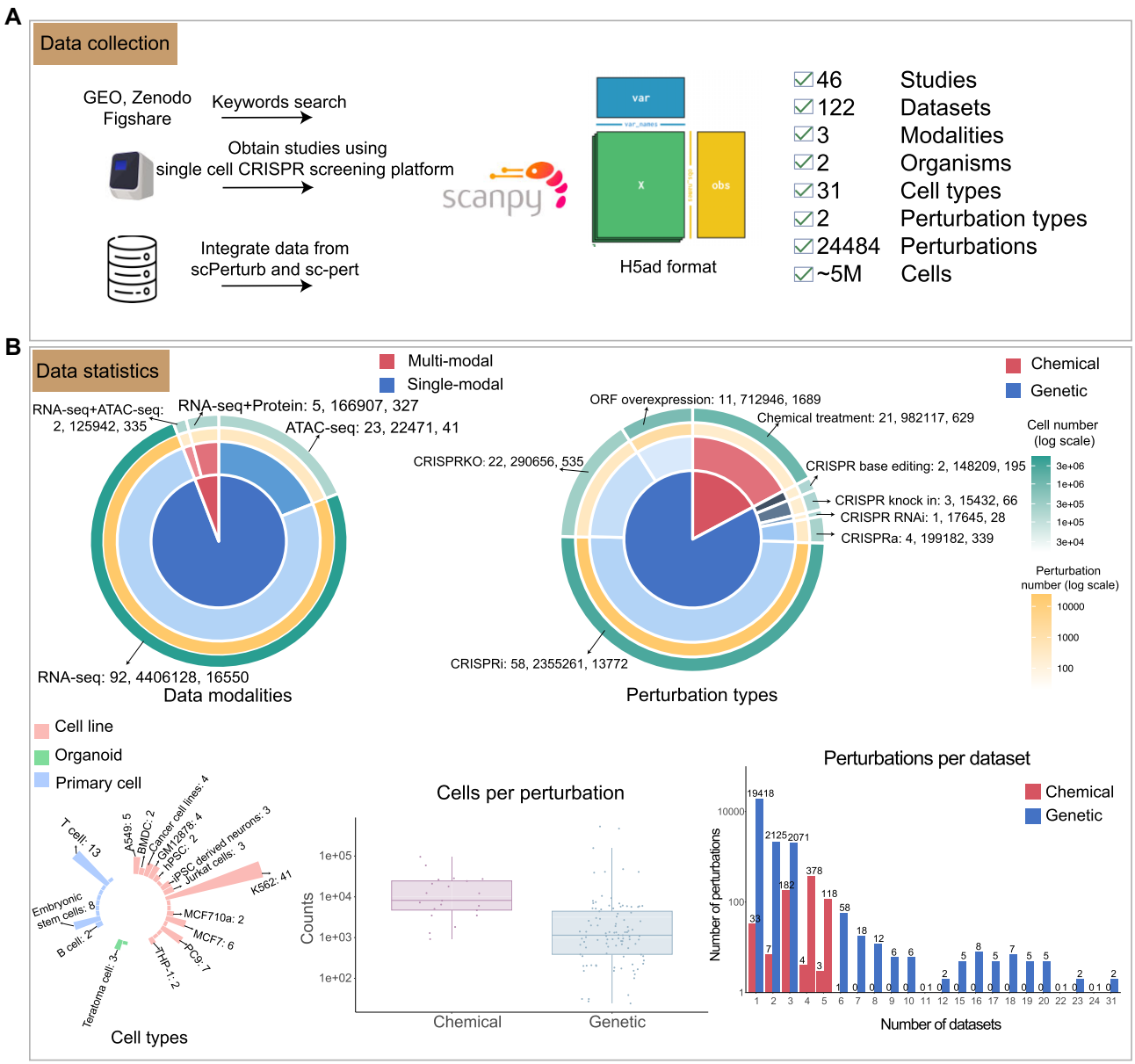


Figure 1. Overview of the data collection and statistics of PerturBase. (**A**) Data collection, construction and summary of the PerturBase resource. (**B**) Overview of the modalities, perturbation types and cell types in PerturBase. The numbers in the top left and top right pie charts represent the numbers of datasets, cells and perturbations, respectively.

Peak count matrix generation

scPerturbation ATAC-seq data exist in two primary formats: the peak count matrix and the 10x fragment TSV file. In the case of the peak count matrix file, its raw data were preserved. For the 10x fragment TSV file, we utilized macs2 within the Signac CallPeaks function to derive the peak count matrix (38,39). The peak count matrix is used for the subsequent scPerturbation ATAC-seq analysis.

Data quality control

For each scPerturbation ATAC-seq dataset, we consistently processed its peak count matrix into the h5ad format using the Seurat and Signac R packages. (i) Cells with expressed peak numbers <200 or >30 000 and a fragment ratio in the peak <0.15, a blacklist fraction >0.05, a nucleosome signal exceeding 4, a transcription start site enrichment below 4 or a blank control were filtered. (ii) Peaks expressed in <10 cells were filtered. (iii) Perturbations, with the exception of the negative control with perturbed cells <30 (default), were not considered. (iv) The peak count matrix was scaled using the term-frequency inverse-document-frequency method in Signac. Subsequent procedures aligned with those employed for scPerturbation RNA-seq in data quality control from step (v). Among the 24 ATAC-seq datasets, 8 met the quality control criteria and were suitable for further downstream data analysis.

Processing of the scPerturbation protein data

In PerturBase, protein modality data are only available for download and have not undergone data processing or subsequent analysis (Supplementary Note S1). This omission is due primarily to three factors: (i) The majority of the data

in our database consists of RNA-seq and ATAC-seq datasets, with only five datasets containing protein data. (ii) Most of the analytical methods we use are specifically designed for RNA-seq and ATAC-seq data. Consequently, their applicability to protein expression profiles is uncertain at this stage. This approach reflects our current focus on ensuring comprehensive and reliable analysis of the more prevalent RNA-seq and ATAC-seq data within our database. As the field of single-cell protein analysis continues to develop and more datasets become available, we plan to incorporate appropriate analytical methodologies to expand our capabilities in this area.

Data denoising of the scPerturbation RNA-seq data

Alternative sources of variation, including batch effects, the cell cycle stage and the activation of cellular stress responses, can confound the downstream analysis of scPerturbation data. To mitigate these issues, we employ Mixscape (21) from pertpy (36) to compute the local perturbation signature for each cell. The core concept is to isolate the effect of the genetic perturbation by subtracting the average expression of the *K* nearest cells from the negative control pool from each cell. Consequently, Mixscape extracts the component of the cell's profile that solely reflects the genetic perturbation. As recommended, we set the number of neighbors *K* to 20. This step was not performed on the scPerturbation ATAC-seq data.

Identification of non-perturbed cells after denoising

The evaluation of sgRNA knockout efficiency in genetic screening and off-target effects in chemical screening is crucial (9,40). In genetic screening, sgRNAs direct Cas9 to specific genomic loci, yet only approximately 70–80% effectively induce the desired impact on the targeted gene. This finding indicates that in 20–30% of cells harboring an sgRNA, the target gene may remain unaffected or partially impacted, resulting in a wild-type phenotype (defined as a non-perturbed cell). Like genetic screening, chemical screening also results in off-target effects (9). Such occurrences can skew the assessment of a perturbation's effect. Consequently, a filtering step to eliminate these cells is necessary. Mixscape leverages the cell's transcriptome as a phenotypic indicator of the perturbation's impact and has devised a method to systematically identify and exclude non-perturbed cells. Mixscape's fundamental premise is that each perturbation class comprises a mix of two Gaussian distributions: one representing successfully perturbed (SP) cells and the other representing non-perturbed (NP) cells. The transcriptional profile distribution of NP cells should align with that of control (CTRL) cells. Mixscape computes the posterior probability of a cell belonging to the SP class and categorizes those with a probability over 0.5 as SP cells. This approach, applied across all perturbations, enables the identification of all SP cells and assesses the targeting efficacy of genetic and chemical perturbation. Notably, in our study, further analysis was performed. We postulated that the targeting efficiency of a perturbation would not be <20%. Hence, if the SP/(NP + SP) ratio for a perturbation <20%, the perturbation will not induce a significant phenotypic change in cells. These perturbations are deemed 'weak' perturbations, and all corresponding cells are retained for analysis (21). We retain cells from 'weak' perturbations for several reasons: (i) Potential misclassification: Cells (with 'weak' perturbations) classified as non-perturbed by Mixscape might actually be perturbed but exhibit only weak effects that are not readily detectable. In other words, this potential for false negatives arises not from the low target efficiency of the induced sgRNA but rather from the subtlety of the perturbation effects. By retaining these cells, we ensure that we do not overlook weak perturbations that could still be biologically significant. (ii) Impact on the number of cells: Weak perturbations often result in a high proportion of non-perturbed cells. Filtering out these cells would significantly reduce the number of cells available for downstream analysis, thereby diminishing the statistical power and robustness of the analysis. Maintaining these cells ensures a more comprehensive dataset and improves the reliability of subsequent analyses. Conversely, if the SP/(NP + SP) ratio exceeds 20%, the perturbations are categorized as 'strong' perturbations, and the NP cells are filtered out, adhering to Mixscape's criteria. This step was not performed on the scPerturbation ATAC-seq data.

Identification of differentially expressed genes associated with a perturbation

In PerturBase, we employed five methods to detect DEGs by comparing perturbations against CTRL (control), including three methods specifically developed for scPerturbation (scMAGeCK, GSFA and SCEPTRE) and two commonly used methods in scRNA-seq provided by Scanpy (Wilcoxon test and *t* test). For the scPerturbation ATAC-seq data, three methods provided by Seurat were employed [Wilcoxon, *t* test and logistic regression (LR)]. Specifically, for scATAC data, (i) differentially accessible peaks (DAPs) between CTRL and perturbation are identified and (ii) the gene closest to each of these peaks is treated as a differentially expressed gene (identified using the 'ClosestFeature' function in Signac). If multiple peaks map to the same gene, the gene is identified as differentially expressed if at least one of the peaks is differentially accessible.

For scMAGeCK, genes with absolute regulatory scores >0.2 and *P* values <0.05 (adjusted *P* values are not available) are defined as DEGs. For GSFA, genes with *P* values (adjusted *P* values are not available) <0.05 are defined as DEGs. For SCEPTRE, genes whose absolute log (fold-change) values are >1 and whose *P* values are <0.05 (adjusted *P* value is not available) are defined as DEGs. For the Wilcoxon test and *t* test, genes (or peaks) with absolute log(fold change) values >1 and adjusted *P* values <0.05 are defined as DEGs (or DAPs). The input for the five methods is the normalized expression profile of HVGs. All the parameters in the above five methods are set to defaults. For LR, peaks with absolute log(fold change) values >1 and adjusted *P* values <0.05 are defined as DAPs.

Evaluation of the effect of a perturbation

We evaluated the effect of a perturbation through three distinct methodologies: (i) An enrichment analysis of the DEGs of a perturbation is performed. PerturBase performs Gene Ontology (GO, version 2.1) (41) enrichment analysis, including cellular component (CC), biological process and molecular function, and Kyoto Encyclopedia of Genes and Genomes (KEGG, release 107.1, 1 August 2023) (42) enrichment analysis for each perturbation using its DEGs. All DEGs were treated collectively, without distinguishing between upregulated and downregulated genes (Supplementary Note S1). Enrichment

analysis was conducted using the enrichGO and enrichKEGG functions in clusterProfiler (v4.7.1.003), which performs over-representation analysis. The enrichment terms are defined as significant if the BH (Benjamini–Hochberg) adjusted *P* value is <0.01 and the *Q* value is <0.05. (ii) The enrichment terms are defined by characterizing the associations between a perturbation with MSigDB signatures (hallmark gene sets, version 2023.2). We utilized the RRA module of scMAGeCK to link perturbations with MSigDB signatures (43). In PerturBase, 50 well-defined hallmark signatures in MSigDB were downloaded for analysis. A perturbation is considered to significantly negatively regulate a phenotype corresponding to a signature if the 'FDR.low' value is <0.01. Conversely, it is deemed to significantly positively regulate a phenotype if the 'FDR.high' value is <0.01. (iii) The clustering membership of a perturbation is evaluated (44). This analysis consists of two parts. First, we assess whether a perturbation is preferentially enriched in a specific cluster compared with the CTRL. Second, we evaluate whether its distribution across clusters significantly deviates from that of the CTRL. In this evaluation, we use the chi-square test to assess whether a perturbation significantly affects cluster membership (adjusted *P* value below 0.01) (43,44).

Characterization of the relationships between perturbations

Perturbations with shared effects or targets tend to produce similar shifts in expression profiles. Therefore, by characterizing the relationships between perturbations using expression profiles, we can describe the differences or similarities between perturbations in terms of effects or perturbation targets. In our current study, we characterize the relationships between perturbations using three methods: (i) A processed expression profile is used. First, the mean expression profiles of the perturbations in a dataset were calculated. The relationships between perturbations were subsequently calculated using cosine similarity. (ii) The E-distance function in pertpy is used (45). E-distance is a statistical metric that compares the mean pairwise distance of cells across two different perturbations to the mean pairwise distance of cells within the two distributions. A large E-distance of perturbed cells from unperturbed cells indicates a strong change in the molecular profile induced by the perturbation. We compute the E-distance after PCA (24). (iii) The latent factors output by the GSFA is used (23). The GSFA describes the effects of a perturbation through a set of latent factors that represent biological pathways or functional units. The relationships between the perturbations are calculated using cosine similarity with the latent factors of the perturbations. The GSFA results are not applied to the scPerturbation ATAC-seq data.

Database construction

PerturBase was built on a Linux server. The web services were built using Nginx (version 1.24.0). The front end of PerturBase was built with HTML5, JavaScript, CSS and React (version 18.0.0), which consists of the react UI library ant design (version 4.20.7). All the data in PerturBase are stored and managed by MySQL (version 8.0.36). PerturBase has been tested on a number of popular web browsers, including the Google Chrome, Firefox and Apple Safari web browsers. No registration or login is needed.

Results

Overview of PerturBase

PerturBase curates 122 scPerturbation datasets from 46 publicly available studies, consisting of 115 single-modal datasets and 7 multi-modal datasets, covering *H. sapiens* and *M. musculus*. Among these datasets, 101 datasets were perturbed with 24 254 genetic compounds, and 21 datasets were perturbed with 230 chemical compounds. PerturBase features two modules: the 'Dataset' and 'Perturbation' modules. The 'Dataset' module facilitates streamlined exploration of all 122 datasets, offering filters by organism, modality, perturbation type, perturbation name and perturbation effect (Figure 2A). For example, the 'Perturbation' keyword describes the perturbations a dataset contains, and users can conveniently search datasets containing the perturbation of interest. The 'PathWay2Data' keyword describes the effects of a perturbation and can be used to search datasets containing perturbations that have effects of interest ('Materials and methods' section). After selecting a dataset of interest, users can gain insights into a range of analysis results, including (i) quality control, (ii) denoising, (iii) differential gene expression analysis, (iv) functional analysis of perturbation effects and (v) characterization of relationships between perturbations (Figure 2B–F). Moreover, the 'Perturbation' module integrates a range of analysis results from datasets of a chosen perturbation, including (i) quality control, (ii) denoising, (iii) differential gene expression analysis and (iv) functional analysis of perturbation effects. These results provide a comparison of perturbations across various cellular contexts, providing valuable insights into their effects (Figure 2B–F).

Visualization of the quality control results

The quality control results provide basic information about a preprocessed dataset in the 'Dataset' module or about a perturbation in the 'Perturbation' module, such as the perturbations the dataset contains and the distribution of cell numbers of a perturbation across experiments (Figure 2B). The 'perturbation in each cell' describes the number of perturbations assigned to each cell. This information is useful if users want to access data that contain combination perturbations. We utilized UMAP to visualize the clustering results in the 'Dataset' module. In summary, the quality control results provide detailed information about a dataset or a perturbation after quality control.

Visualization of the data denoising results

Alternative sources of variation, including batch effects, the cell cycle stage and the activation of cellular stress responses, confound downstream analysis. Therefore, PerturBase adopts Mixscape (21) of pertpy to alleviate those confounding factors by calculating the local perturbation signature for each cell (Figure 2C, 'Materials and methods' section). In addition, we utilized UMAP and a heatmap to visualize and compare the clustering results before and after denoising in the 'Dataset' module. The UMAP and heatmap results are not available in the 'Perturbation' module because we do not integrate the expression profiles of the datasets. After denoising, we further employed Mixscape to identify non-perturbed cells. Non-perturbed cells were defined as cells that experienced perturbation but did not exhibit the expected phenotype because the perturbation had no effect on the target's

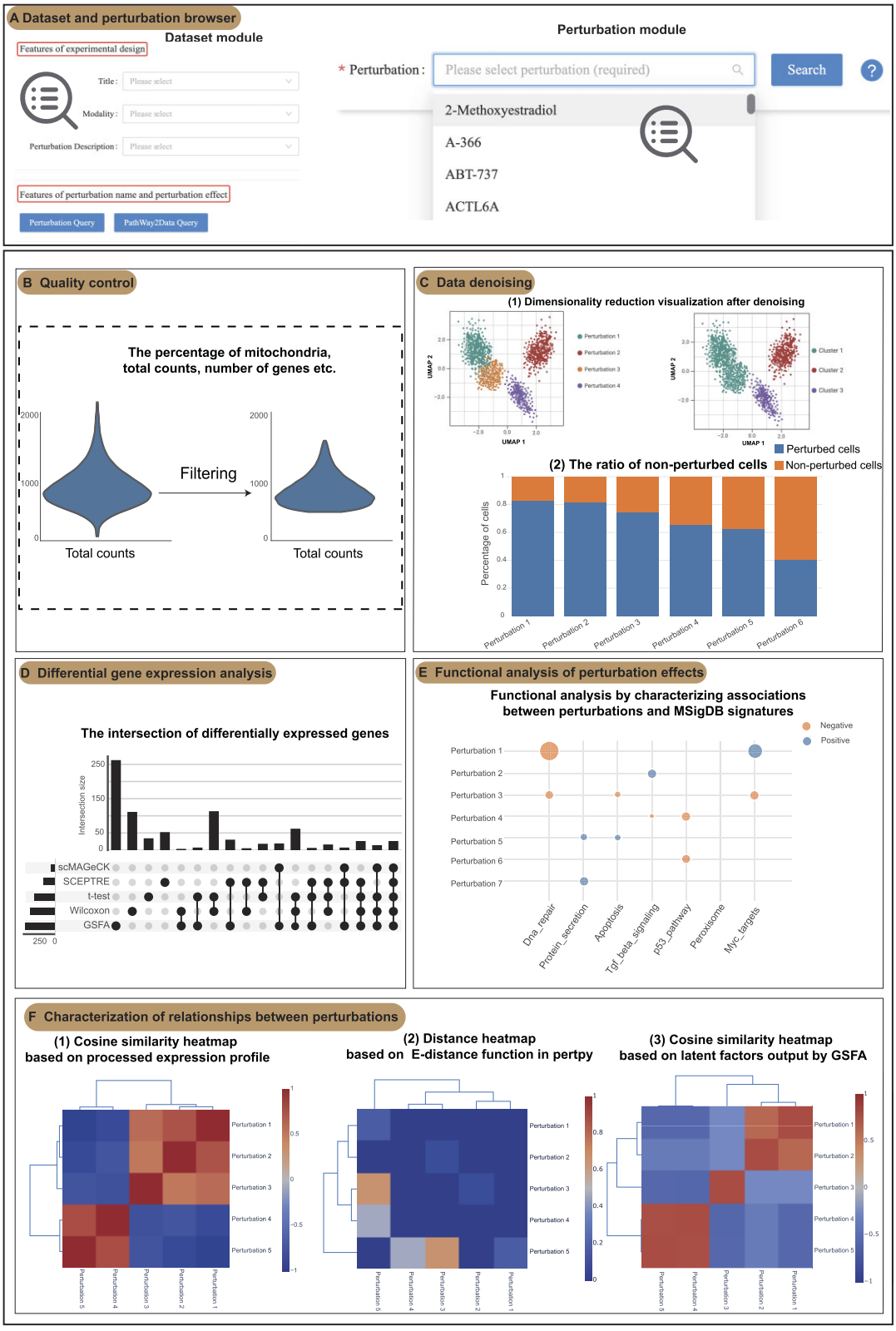


Figure 2. Overview of the 'Dataset' and 'Perturbation' modules. (**A**) The 'Dataset' module allows users to analyze specific scPerturbation data of interest based on variables such as study title, perturbation, perturbation type and cell line. The 'Perturbation' module allows users to analyze various results of a specific perturbation of interest across experiments. (**B**) The quality control section presents basic information about a preprocessed dataset or a perturbation across experiments. (**C**) PerturBase adopts Mixscape to alleviate confounding factors. The data denoising section shows the results of dimensionality reduction after data denoising and the ratio of non-perturbed cells. (**D**) PerturBase implements five distinct methods for detecting DEGs in the scPerturbation RNA-seq data and three methods for ATAC-seq data. This section presents the DEGs of a perturbation under a chosen method. (**E**) The functional analysis section employs three methodologies to evaluate the effects of perturbations. (**F**) The relationship section presents the relationships between perturbations.

transcription/translation (21). These non-perturbed cells influence the estimation of the effect of a perturbation. Therefore, Mixscape was employed to identify non-perturbed cells and evaluate the efficacy of each perturbation ('Materials and methods' section). In our current study, a further analysis was performed. We utilized the efficacy information to classify a perturbation as a 'weak' perturbation or a 'strong' perturbation (21). A strong perturbation usually has a strong effect on the cell phenotype, whereas a weak perturbation has little effect on the cell phenotype. This information provides insight into the effect of a particular perturbation, which may aid our understanding of the roles of a perturbation. All the above information, including the results output by the Mixscape and filtered datasets, can be browsed and accessed in PerturBase.

Visualization of the DEGs associated with a perturbation

In our study, we implemented five distinct methods to detect DEGs from the scPerturbation RNA-seq data (23,36,46) and three methods for the scPerturbation ATAC-seq data (38,39), as detailed in the 'Materials and methods' section (Figure 2D). In the 'Dataset' module, we used a bar plot to effectively highlight the variability in the number of DEGs identified by each of the five methods. By default, the analysis focuses on the top 25 perturbations with the most pronounced effects. However, if users are interested in a particular perturbation, they can delve deeper into the associated DEGs. This is facilitated through interactive tables and volcano plots, which become available once a specific perturbation and a method are selected. In the 'Perturbation' module, we used a bar plot to show the variability in the number of DEGs identified by each of the five methods across the experiments. Additionally, PerturBase incorporates the UpSet R package to illustrate the overlaps and distinctions in DEG identification across different methods for a given perturbation in both modules. This visualization aids in understanding the consensus and discrepancies in DEG identification among various analytical approaches and cellular contexts.

Visualization of the effect of a perturbation

As detailed in the 'Materials and methods' section, to thoroughly assess the effect of a perturbation, we employed three and two methodologies in the 'Dataset' and 'Perturbation' modules, respectively (Figure 2E). The outcomes of these analyses are effectively visualized using bar plots and heatmaps. In the 'Dataset' module, our interface prioritizes and displays only the top 25 perturbations, identified as having the most significant effects. To visualize the GO and KEGG enrichment results, we present both individual and aggregated enrichment analyses for each perturbation. Additionally, PerturBase enhances user engagement by allowing the exploration of enrichment results for DEGs identified through a specific method tailored to individual perturbations.

Visualization of the relationships between perturbations

Perturbations sharing similar effects often manifest comparable shifts in gene expression profiles. Consequently, by examining these expression profiles, we can elucidate the relationships between perturbations. In our study, we characterize their relationships through three distinct methods (Figure 2F): (i) analyzing the similarity between perturbations based on processed expression profiles; (ii) employing the E-distance function in pertpy to quantify relationships between perturbations; and (iii) leveraging latent factors derived from GSFA to further quantify their relationships. The findings from these analyses are concisely presented in a heatmap format, facilitating an intuitive understanding of the relationships between various perturbations. These results are not available in the 'Perturbation' module because we do not integrate the expression profiles of the datasets.

Case study 1

Programmed death-ligand (PD-L1) is frequently observed in human cancers and can lead to the suppression of T-cell-mediated immune responses (47–49). To demonstrate the capabilities of PerturBase, we utilized the 'Dataset' module to analyze a scPerturbation dataset from Papalexi *et al*. to investigate the regulatory mechanisms of the expression of PD-L1 (21). This comprehensive analysis highlights the efficacy of PerturBase in the analysis of the effects of perturbations. The dataset encompasses 25 perturbations, with cell counts per perturbation ranging from 33 to 1197 post-quality control (Figure 3A). Of these, 11 were categorized as 'strong' perturbations, whereas the remaining 14 were categorized as 'weak' perturbations (Supplementary Figure S3A). The sgRNA efficiencies in the strong perturbations varied from ∼50–80%, suggesting that some cells, despite receiving an sgRNA, did not exhibit the expected phenotype and were thus classified as non-perturbed cells (refer to 'Materials and methods' section for details). For example, as depicted in the left panel of Figure 3B and Supplementary Figure S3B, the identified non-perturbed cells expressing IFNGR1 presented a wild-type phenotype, and the IFNGR1 target gene PD-L1 was not affected. After filtering out non-perturbed cells, two clear groups of cells were observed, including a cluster consisting of knockouts in IFNGR1, IFNGR2, JAK2 and STAT1 and a second cluster consisting of the knockout IRF1 (Figure 3C), underscoring the necessity and effectiveness of the filtering strategy employed by PerturBase.

As shown in Supplementary Figure S3C, IFNGR1, IFNGR2, JAK2 and STAT1 were highly similar, which is concordant with prior findings (50,51). Additionally, perturbations, such as those in STAT3, CD86 and ATF2, were highly similar to those in the CTRL group, suggesting minimal phenotypic impact on the cells. This observation is consistent with their previous classification as 'weak' perturbations. Further investigation into the mechanisms driving PD-L1 downregulation revealed that these four perturbations are central to the immune response pathway, particularly in the context of interferon (IFN)-γ signaling (Figure 3D and E). When IFN-γ binds to its receptor, which consists of IFNGR1 and IFNGR2, it activates JAK2, which in turn phosphorylates STAT1. Phosphorylated STAT1 dimerizes and translocates to the nucleus, where it binds to DNA and promotes the transcription of genes involved in the immune response, including those regulating PD-L1 expression (47). However, the disruption of IFNGR1, IFNGR2, JAK2 or STAT1 can inhibit this signaling pathway, leading to decreased PD-L1 expression. This reduction in PD-L1 can enhance the ability of the immune system to target and destroy cancer cells, as PD-L1 normally acts to suppress immune responses. Moreover, gene enrichment analysis of IFNGR1, IFNGR2, JAK2 and STAT1 revealed their involvement in pathways related to MHC protein complex

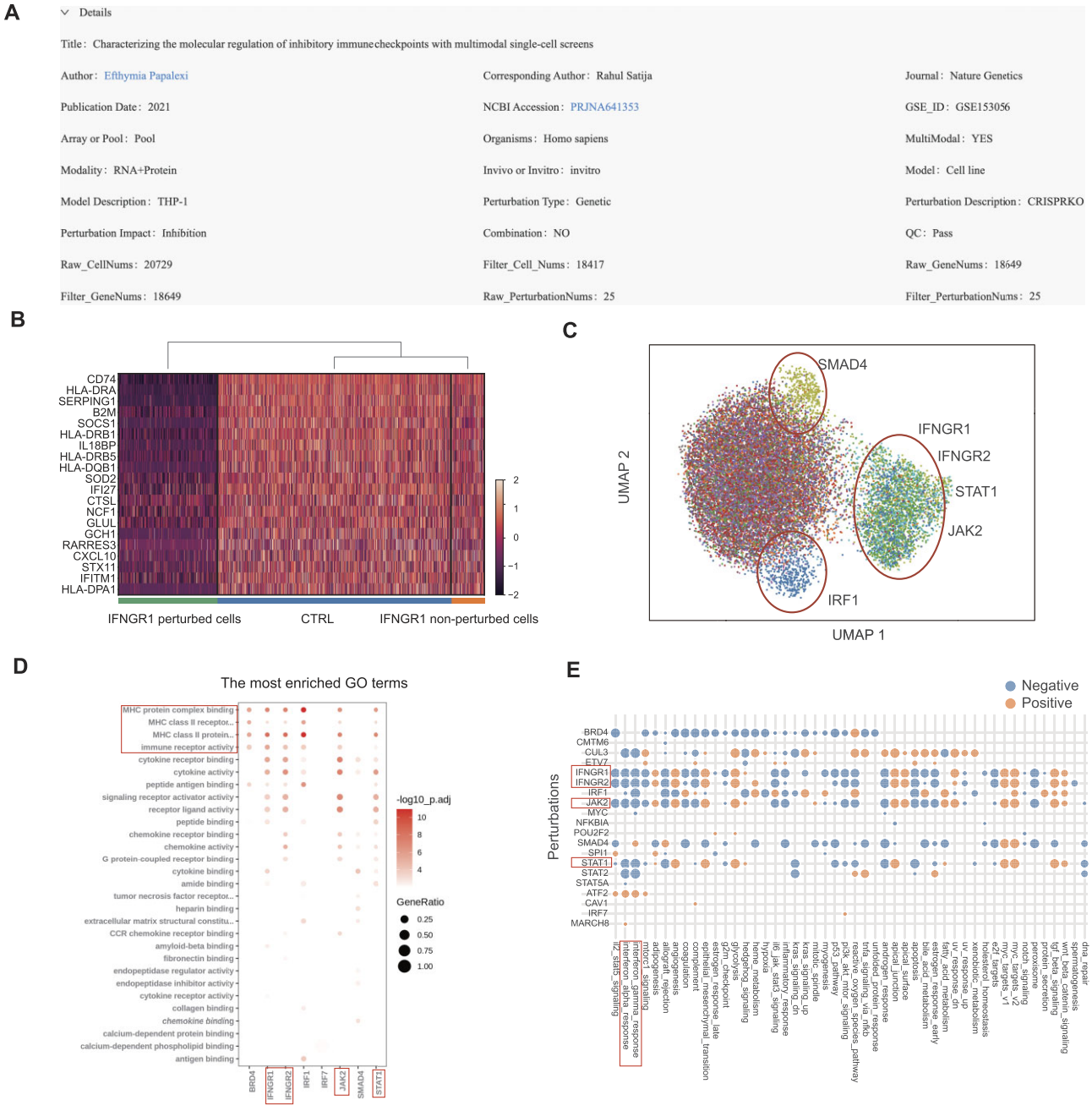


Figure 3. Case study 1 demonstrates the capabilities of the PerturBase 'Dataset' module. (**A**) Detailed information of the case study 1 dataset as provided by PerturBase. (**B**) Expression heatmap of IFNGR1-overexpressing and control cells. The heatmap was generated using Mixscape with default parameters. (**C**) UMAP plot produced by PerturBase 'Data dimensionality reduction visualization after denoising' panel. (**D**) GO enrichment results for BRD4, IFNGR1, IFNGR2, IRF1, IRF7, JAK2, SMAD4 and STAT1. (**E**) Associations between the selected perturbations and MSigDB hallmark signatures. Only associations with an adjusted *P* value <0.01 are shown. Perturbations without significant associations are excluded.

binding and immune receptor activity, further underscoring their pivotal roles in immune regulation (Figure 3D). In conclusion, these insights underscore the comprehensive nature of PerturBase as a scPerturbation database. It stands out as an indispensable resource for researchers seeking to access and analyze perturbation data, providing a robust platform for the exploration of perturbation effects.

Case study 2

To further demonstrate the functionality and value of PerturBase, we utilized the 'Perturbation' module to integrate chemical scPerturbation datasets across multiple cell lines to explore the dose–response relationship and its heterogeneity. In this context, we utilized the data generated by Srivatsan *et al.* (9) who assessed the impact of 188 compounds on A549, K562 and MCF7 cell lines at four different concentrations (0.01, 0.1, 1 and 10 μM).

Typically, there is a proportional relationship between chemical dose and concentration (known as the dose–effect relationship), meaning that the higher the dose of a compound is, the greater its impact on cell lines (52). In the A549, K562 and MCF7 cell lines, as the concentration increased from 0.01 to 10 μM, the number of DEGs also in-

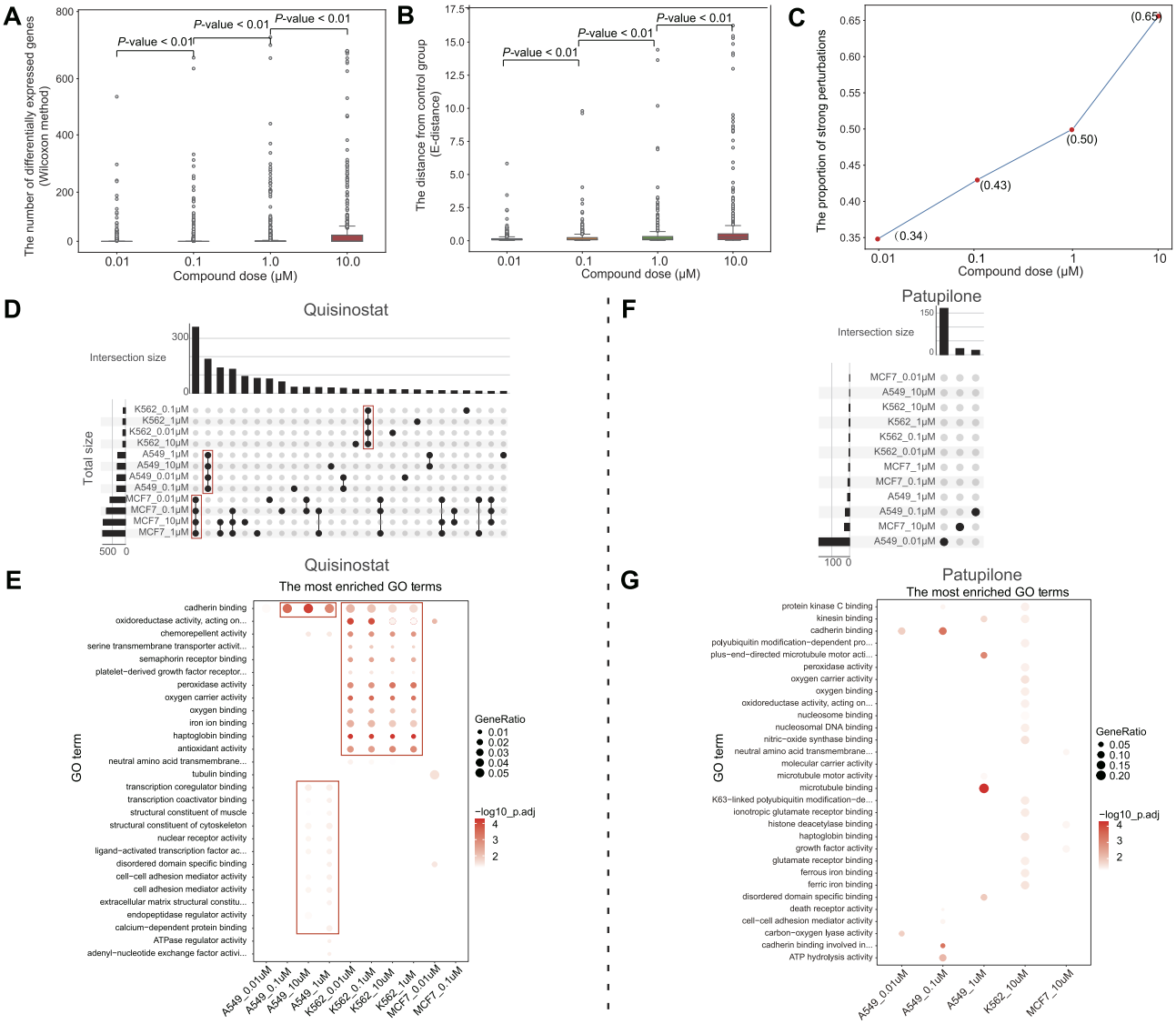


Figure 4. Case study 2 demonstrates the capabilities of the PerturBase 'Perturbation' module. (**A**) Numbers of differentially expressed genes associated with compounds at various doses. Boxes show the 25th–75th percentiles, with a line at the median. Whiskers extend to minimum and maximum values within 1.5 times the interquartile range. (**B**) Distances between the compounds and the control group at various doses. Boxes show the 25th–75th percentiles, with a line at the median. Whiskers extend to minimum and maximum values within 1.5 times the interquartile range. (**C**) Proportions of compounds categorized as strong perturbations under various doses. (**D**) Intersections of genes differentially expressed in response to quisinostat under various conditions. The *y*-axis represents various conditions. For example, A549_0.1 μM indicates A549 cells treated with 0.1 μM quisinostat. Only intersections with a size larger than 10 are shown. (**E**) The most enriched GO terms for quisinostat under various conditions. The *x*-axis represents various conditions. For example, A549_0.1 μM indicates A549 cells treated with 0.1 μM quisinostat. (**F**) Intersections of DEGs for patupilone under various conditions. The *y*-axis represents various conditions. For example, A549_0.1 μM indicates A549 cells treated with 0.1 μM of patupilone. Only intersections with a size larger than 10 are shown. (**G**) The most enriched GO terms for patupilone under various conditions. The *x*-axis represents various conditions. For example, A549_0.1 μM indicates A549 cells treated with 0.1 μM patupilone.

creased, and the distance (calculated by the E-distance function) from the control group (non-perturbed) increased (Figure 4A and B). We obtained similar conclusions with the Wilcoxon, SCEPTRE, scMAGeCK and GSFA DEG detection methods. Additionally, the percentage of compounds categorized as strong perturbations also tended to increase as the concentration increased from 0.01 to 10 μM (Figure 4C). However, the dose–effect relationships across compounds also differed. For example, compared with other compounds, quisinostat (named quisinostat-JNJ-26481585-2HCl in PerturBase), a histone deacetylase inhibitor (HDACi), reached its maximal effect at a low concentration of 0.01 μM, indicating that its effect was already saturated. As shown in Figure 4D, the number of DEGs for quisinostat remained relatively constant across low and high concentrations in the three cancer cell lines. Moreover, GO and KEGG functional enrichment analyses, along with MsigDB functional analysis, revealed that quisinostat had similar functions, such as increasing cell apoptosis at both low and high concentrations (53) (Figure 4E and Supplementary Figure S4A). Furthermore, the dose–effect relationship of a specific compound can vary drastically across cell lines. Patupilone, a microtubule function inhibitor, was classified as a weak perturbation in the K562 cell line, indicating minimal impact. In contrast, in the MCF7 cell line,

patupilone only showed significant efficacy at high concentrations, where DEGs became apparent (Figure 4F). Interestingly, in the A549 cell line, patupilone had a unique dose–response relationship, where it exerted a significant effect at low concentrations but lost its efficacy at relatively high concentrations (Figure 4F and Supplementary Figure S4B). These intricate mechanisms underscore the importance of understanding dose-dependent variations in drug efficacy, particularly in diverse cellular contexts such as those provided by PerturBase.

In summary, these analyses underscore the value of PerturBase in providing comprehensive insights into drug responses. By facilitating the integration and analysis of scPerturbation data across various cell lines and conditions, PerturBase proves to be a valuable resource for researchers studying the dynamics of drug efficacy and cellular responses.

Conclusions and future development

PerturBase provides two major modules, namely 'Dataset' and 'Perturbation', for the exploration of high-content scPerturbation data. The 'Dataset' module provides easy exploration and accession of all 122 perturbed datasets with 12 keywords. The 'Perturbation' module integrates a range of analysis results across datasets that share the same perturbation. To enhance the interpretability of high-content perturbation resources, PerturBase offers five categories of analysis and visualization, including quality control, denoising, identification of DEGs, functional analysis of perturbation effects and characterization of relationships between perturbations. However, the data processing and analysis methods in PerturBase may have potential limitations and biases. First, preprocessing choices, such as normalization techniques and non-perturbed cell identification, can introduce variability and affect downstream analyses. Second, PerturBase's focus on HVGs can result in the exclusion of perturbation-specific DEGs, potentially introducing bias in downstream analyses, such as functional analyses (Supplementary Note S1). Additionally, the presence of confounding variables, such as the cell cycle stage or batch effect, can impact the interpretation of perturbation effects. Despite our efforts to mitigate these factors through rigorous quality control and standardized preprocessing, it is essential to consider these limitations when interpreting the results. Future work will focus on refining these methods and incorporating advanced techniques, such as Mixscale (54), to further increase the robustness and accuracy of our analyses.

Moving forward, we aim to enhance PerturBase in several key areas. First, we are committed to the continual curation of datasets, expanding our repository with the latest high-content perturbation studies, particularly those focused on protein, chemical and multi-modal experiments (Supplementary Notes S2 and S3). Second, although PerturBase offers five distinct types of analysis results, we acknowledge the need for more interactive visualizations to enhance the interpretability of each result. To address this, we are dedicated to developing a more user-friendly and interactive platform in our forthcoming version, facilitating easier access to comprehensive information. Finally, we plan to incorporate additional modules to deepen the understanding of the scPerturbation datasets (Supplementary Note S2). In essence, PerturBase stands as the pioneering high-content screening database that has been specifically designed for the efficient search, visualization and analysis of scPerturbation datasets. We believe that PerturBase will become an indispensable resource in the field, offering an extensive range of data and functionalities.

Supplementary data

Supplementary Data are available at NAR Online.

Data availability

PerturBase is an open resource for interactively visualization and analysis of the comprehensive scPerturbation data resource (http://www.perturbase.cn/). PerturBase is freely accessible, without any registration requirements. To enhance transparency and reproducibility, we have uploaded demo data to https://figshare.com/s/dddc4ddf91d0b100fd6c and code to https://github.com/bm2-lab/PerturBase and https://doi.org/10.5281/zenodo.13761517.

Acknowledgements

We gratefully acknowledge all scPerturbation dataset owners for generously sharing their data. Special thanks to Eng. Qian Yu, Prof. Ling Guo and anonymous reviewers for their valuable technical support and insightful suggestions to enhance PerturBase.

Funding

National Natural Science Foundation of China [T24250193, 32341008]; National Key Research and Development Program of China [2021YFF1201200, 2021YFF1200900]; Shanghai Pilot Program for Basic Research; Shanghai Science and Technology Innovation Action Plan—Key Specialization in Computational Biology; Shanghai Shuguang Scholars Project; Shanghai Excellent Academic Leader Project; Shanghai Municipal Science and Technology Major Project [2021SHZDZX0100]; Fundamental Research Funds for the Central Universities; Shanghai Rising-Star Program [23YF1450200]; China Postdoctoral Science Foundation [2022M722418, 2023T160485]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement

None declared.

References

1. Iorio,F., Bosotti,R., Scacheri,E., Belcastro,V., Mithbaokar,P., Ferriero,R., Murino,L., Tagliaferri,R., Brunetti-Pierri,N., Isacchi,A., *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 14621.
2. Berg,E.L. (2021) The future of phenotypic drug discovery. *Cell Chem. Biol.*, **28**, 424–430.
3. Hughes,R.E., Elliott,R.J.R., Dawson,J.C. and Carragher,N.O. (2021) High-content phenotypic and pathway profiling to advance drug discovery in diseases of unmet need. *Cell Chem. Biol.*, **28**, 338–355.
4. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K., *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.

5. Joung,J., Konermann,S., Gootenberg,J.S., Abudayyeh,O.O., Platt,R.J., Brigham,M.D., Sanjana,N.E. and Zhang,F. (2017) Genome-scale CRISPR–Cas9 knockout and transcriptional activation screening. *Nat. Protoc.*, **12**, 828–863.
6. Stathias,V., Turner,J., Koleti,A., Vidovic,D., Cooper,D., Fazel-Najafabadi,M., Pilarczyk,M., Terryn,R., Chung,C., Umeano,A., *et al.* (2020) LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.*, **48**, D431–D439.
7. Cheng,J.Y., Lin,G.L., Wang,T.H., Wang,Y.Z., Guo,W.B., Liao,J., Yang,P.H., Chen,J., Shao,X., Lu,X.Y., *et al.* (2023) Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. *Adv. Sci.*, **10**, e2204484.
8. Bock,C., Datlinger,P., Chardon,F., Coelho,M.A., Dong,M.T.B., Lawson,K.A., Lu,T., Maroc,L., Norman,T.M., Song,B.A., *et al.* (2022) High-content CRISPR screening. *Nat. Rev. Methods Primers*, **2**, 9.
9. Srivatsan,S.R., McFaline-Figueroa,J.L., Ramani,V., Saunders,L., Cao,J.Y., Packer,J., Pliner,H.A., Jackson,D.L., Daza,R.M., Christiansen,L., *et al.* (2020) Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, **367**, 45–51.
10. Dixit,A., Pamas,O., Li,B.Y., Chen,J., Fulco,C.P., Jerby-Amon,L., Marjanovic,N.D., Dionne,D., Burks,T., Raychowdhury,R., *et al.* (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, **167**, 1853–1866.
11. Adamson,B., Norman,T.M., Jost,M., Cho,M.Y., Nuñez,J.K., Chen,Y.W., Villalta,J.E., Gilbert,L.A., Horlbeck,M.A., Hein,M.Y., *et al.* (2016) A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, **167**, 1867–1882.
12. Jaitin,D.A., Weiner,A., Yofe,I., Lara-Astiaso,D., Keren-Shaul,H., David,E., Salame,T.M., Tanay,A., van Oudenaarden,A. and Amit,I. (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell*, **167**, 1883–1866.e17.
13. Binan,L., Danquah,S., Valakh,V., Simonton,B., Bezney,J., Nehme,R., Cleary,B. and Farhi,S.L. (2023) Simultaneous CRISPR screening and spatial transcriptomics reveals intracellular, intercellularand functional transcriptional circuits. bioRxiv doi: https://doi.org/10.1101/2023.11.30.569494, 01 December 2023, preprint: not peer reviewed.
14. Datlinger,P., Rendeiro,A.F., Schmidl,C., Krausgruber,T., Traxler,P., Klughammer,J., Schuster,L.C., Kuchler,A., Alpar,D. and Bock,C. (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, **14**, 297–301.
15. Replogle,J.M., Saunders,R.A., Pogson,A.N., Hussmann,J.A., Lenail,A., Guna,A., Mascibroda,L., Wagner,E.J., Adelman,K., Lithwick-Yanai,G., *et al.* (2022) Mapping information-rich genotype–phenotype landscapes with genome-scale perturb-seq. *Cell*, **185**, 2559–2575.
16. Rubin,A.J., Parker,K.R., Satpathy,A.T., Qi,Y., Wu,B., Ong,A.J., Mumbach,M.R., Ji,A.L., Kim,D.S., Cho,S.W., *et al.* (2019) Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, **176**, 361–376.
17. Mimitou,E.P., Cheng,A., Montalbano,A., Hao,S., Stoeckius,M., Legut,M., Roush,T., Herrera,A., Papalexi,E., Ouyang,Z., *et al.* (2019) Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods*, **16**, 409–412.
18. Dhainaut,M., Rose,S.A., Akturk,G., Wroblewska,A., Nielsen,S.R., Park,E.S., Buckup,M., Roudko,V., Pia,L., Sweeney,R., *et al.* (2022) Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell*, **185**, 1223–1239.
19. Pierce,S.E., Granja,J.M. and Greenleaf,W.J. (2021) High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.*, **12**, 2969.
20. Liscovitch-Brauer,N., Montalbano,A., Deng,J.L., Méndez-Mancilla,A., Wessels,H.H., Moss,N.G., Kung,C.Y., Sookdeo,A., Guo,X.Y., Geller,E., *et al.* (2021) Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat. Biotechnol.*, **39**, 1270–1277.
21. Papalexi,E., Mimitou,E.P., Butler,A.W., Foster,S., Bracken,B., Mauck,W.M., Wessels,H.H., Hao,Y.H., Yeung,B.Z., Smibert,P., *et al.* (2021) Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.*, **53**, 322–331.
22. Duan,B., Zhou,C., Zhu,C.Y., Yu,Y.F., Li,G.Y., Zhang,S.H., Zhang,C., Ye,X.Y., Ma,H.H., Qu,S., *et al.* (2019) Model-based understanding of single-cell CRISPR screening. *Nat. Commun.*, **10**, 2233.
23. Zhou,Y., Luo,K., Liang,L., Chen,M. and He,X. (2023) A new Bayesian factor analysis method improves detection of genes and biological processes affected by perturbations in single-cell CRISPR screening. *Nat. Methods*, **20**, 1693–1703.
24. Peidli,S., Green,T.D., Shen,C., Gross,T., Min,J., Garda,S., Yuan,B., Schumacher,L.J., Taylor-King,J.P., Marks,D.S., *et al.* (2024) scPerturb: harmonized single-cell perturbation data. *Nat. Methods*, **21**, 531–540.
25. Sayers,E.W., Beck,J., Bolton,E.E., Brister,J.R., Chan,J., Comeau,D.C., Connor,R., DiCuccio,M., Farrell,C.M., Feldgarden,M., *et al.* (2024) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **52**, D33–D43.
26. Hill,A.J., McFaline-Figueroa,J.L., Starita,L.M., Gasperini,M.J., Matreyek,K.A., Packer,J., Jackson,D., Shendure,J. and Trapnell,C. (2018) On the design of CRISPR-based single-cell molecular screens. *Nat. Methods*, **15**, 271–274.
27. Xie,S.Q., Duan,J.L., Li,B.X., Zhou,P. and Hon,G.C. (2017) Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell*, **66**, 285–299.
28. Rubin,A.J., Parker,K.R., Satpathy,A.T., Qi,Y.Y., Wu,B.J., Ong,A.J., Mumbach,M.R., Ji,A.L., Kim,D.S., Cho,S.W., *et al.* (2019) Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell*, **176**, 361–376.
29. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, **14**, 865–868.
30. Replogle,J.M., Norman,T.M., Xu,A., Hussmann,J.A., Chen,J., Cogan,J.Z., Meer,E.J., Terry,J.M., Riordan,D.P., Srinivas,N., *et al.* (2020) Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.*, **38**, 954–961.
31. Song,Q., Ni,K., Liu,M., Li,Y., Wang,L., Wang,Y., Liu,Y., Yu,Z., Qi,Y., Lu,Z., *et al.* (2020) Direct-seq: programmed gRNA scaffold for streamlined scRNA-seq in CRISPR screen. *Genome Biol.*, **21**, 136.
32. Schraivogel,D., Gschwind,A.R., Milbank,J.H., Leonce,D.R., Jakob,P., Mathur,L., Korbel,J.O., Merten,C.A., Velten,L. and Steinmetz,L.M. (2020) Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods*, **17**, 629–635.
33. Joung,J., Ma,S., Tay,T., Geiger-Schuller,K.R., Kirchgatterer,P.C., Verdine,V.K., Guo,B.L., Arias-Garcia,M.A., Allen,W.E., Singh,A., *et al.* (2023) A transcription factor atlas of directed differentiation. *Cell*, **186**, 209–229.
34. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T., *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
35. Ji,Y.G., Lotfollahi,M., Wolf,F.A. and Theis,F.J. (2021) Machine learning for perturbational single-cell omics. *Cell Syst.*, **12**, 522–537.
36. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
37. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L., *et al.* (2021) clusterProfiler 4.0: a universal

enrichment tool for interpreting omics data. *Innovation(Camb.)*, **2**, 100141.
38. Stuart,T., Srivastava,A., Madad,S., Lareau,C.A. and Satija,R. (2021) Single-cell chromatin state analysis with Signac. *Nat. Methods*, **18**, 1333–1341
39. Hao,Y.H., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S.W., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
40. Popp,M.W. and Maquat,L.E. (2016) Leveraging rules of nonsense-mediated mRNA decay for genome engineering and personalized medicine. *Cell*, **165**, 1319–1322.
41. Gene Ontology,C., Aleksander,S.A., Balhoff,J., Carbon,S., Cherry,J.M., Drabkin,H.J., Ebert,D., Feuermann,M., Gaudet,P., Harris,N.L., *et al.* (2023) The Gene Ontology knowledgebase in 2023. *Genetics*, **224**, iyad031.
42. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
43. Yang,L., Zhu,Y.Q., Yu,H., Cheng,X.L., Chen,S.T., Chu,Y.L., Huang,H., Zhang,J. and Li,W. (2020) scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Cancer Res.*, **21**, 19.
44. Fleck,J.S., Jansen,S.M.J., Wollny,D., Zenk,F., Seimiya,M., Jain,A., Okamoto,R., Santel,M., He,Z.S., Camp,J.G., *et al.* (2022) Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, **621**, 365–372.
45. Heumos,L., Ji,Y., May,L., Green,T., Zhang,X., Wu,X., Ostner,J., Peidli,S., Schumacher,A., Hrovatin,K., *et al.* (2024) Pertpy: an end-to-end framework for perturbation analysis. bioRxiv doi: https://doi.org/10.1101/2024.08.04.606516, 07 August 2024, preprint: not peer reviewed.
46. Barry,T., Wang,X.R., Morris,J.A., Roeder,K. and Katsevich,E. (2021) SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biol.*, **22**, 344.
47. Burr,M.L., Sparbier,C.E., Chan,Y.C., Williamson,J.C., Woods,K., Beavis,P.A., Lam,E.Y.N., Henderson,M.A., Bell,C.C., Stolzenburg,S., *et al.* (2017) CMTM6 maintains the expression of PD-L1 and regulates anti-tumour immunity. *Nature*, **549**, 101–105.
48. Zhang,J., Bu,X., Wang,H., Zhu,Y., Geng,Y., Nihira,N.T., Tan,Y., Ci,Y., Wu,F., Dai,X., *et al.* (2018) Cyclin D-CDK4 kinase destabilizes PD-L1 via cullin 3-SPOP to control cancer immune surveillance. *Nature*, **553**, 91–95.
49. Yamaguchi,H., Hsu,J.M., Yang,W.H. and Hung,M.C. (2022) Mechanisms regulating PD-L1 expression in cancers and associated opportunities for novel small-molecule therapeutics. *Nat. Rev. Clin. Oncol.*, **19**, 287–305.
50. Cossetti,C., Iraci,N., Mercer,T.R., Leonardi,T., Alpi,E., Drago,D., Alfaro-Cervello,C., Saini,H.K., Davis,M.P., Schaeffer,J., *et al.* (2014) Extracellular vesicles from neural stem cells transfer IFN-γ via Ifngr1 to activate Stat1 signaling in target cells. *Mol. Cell*, **56**, 609–609.
51. Moon,J.W., Kong,S.K., Kim,B.S., Kim,H.J., Lim,H., Noh,K., Kim,Y., Choi,J.W., Lee,J.H. and Kim,Y.S. (2017) IFNγ induces PD-L1 overexpression by JAK2/STAT1/IRF-1 signaling in EBV-positive gastric carcinoma. *Sci. Rep.*, **7**, 17810.
52. Chen,B., Ma,L., Paik,H., Sirota,M., Wei,W., Chua,M.S., So,S. and Butte,A.J. (2017) Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat. Commun.*, **8**, 16022.
53. Morales Torres,C., Wu,M.Y., Hobor,S., Wainwright,E.N., Martin,M.J., Patel,H., Grey,W., Gronroos,E., Howell,S., Carvalho,J., *et al.* (2020) Selective inhibition of cancer cell self-renewal through a quisinostat-histone H1.0 axis. *Nat. Commun.*, **11**, 1792.
54. Jiang,L., Dalgarno,C., Papalexi,E., Mascio,I., Wessels,H.-H., Yun,H., Iremadze,N., Lithwick-Yanai,G., Lipson,D. and Satija,R. (2024) Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. bioRxiv doi: https://doi.org/10.1101/2024.01.29.576933, 30 January 2024, preprint: not peer reviewed.

Received: June 1, 2024. **Revised:** September 10, 2024. **Editorial Decision:** September 11, 2024. **Accepted:** September 19, 2024