scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data

Stefan Peidli[1,2,x], Tessa D. Green[3,x], Ciyue Shen[4,5,6], Torsten Gross[7], Joseph Min[3], Samuele Garda[2], Jake P. Taylor-King[7], Debora S. Marks[3,6], Augustin Luna[4,5,6], Nils Blüthgen[1,2,+], Chris Sander[4,5,6,+]

1: Institute of Pathology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
2: IRI Life Sciences, Humboldt-Universität zu Berlin, Berlin, Germany
3: Department of Systems Biology, Harvard Medical School, Boston, MA, USA
4: Departments of Cell Biology and Systems Biology, Harvard Medical School, Boston, MA, USA
5: Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
6: Broad Institute, Cambridge, MA, USA
7: Relation Therapeutics, London, UK
x: joint first authors
+: joint senior authors

Abstract

Recent biotechnological advances led to growing numbers of single-cell studies, which reveal molecular and phenotypic responses to large numbers of perturbations. However, analysis across diverse datasets is typically hampered by differences in format, naming conventions, data filtering and normalization. In order to facilitate development and benchmarking of computational methods in systems biology, we collect a set of 44 publicly available single-cell perturbation-response datasets with molecular readouts, including transcriptomics, proteomics and epigenomics. We apply uniform pre-processing and quality control pipelines and harmonize feature annotations. The resulting information resource enables efficient development and testing of computational analysis methods, and facilitates direct comparison and integration across datasets. Using these datasets, we demonstrate the application of E-distance for quantifying perturbation similarity and strength. This work provides an information resource and guide for researchers working with single-cell perturbation data and highlights conceptual considerations for new experiments. The data is publicly available at scperturb.org.

Introduction

Definition of single-cell perturbation data

Perturbation experiments probe the response of cells or cellular systems upon changes in conditions. These changes were traditionally acting equally on all cells in the model system, such as by modifying temperature or adding drugs. Nowadays, with the latest functional genomics techniques, single-cell genetic perturbations which act on individual cellular components have become available. Perturbations target different layers of the hierarchy of protein production (Figure 1) depending on the technology used. CRISPR-cas9 at the lowest layer acts directly on the genome, using indels to induce frameshift mutations which effectively knock out one or multiple specified genes (Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016). The more recent CRISPRi and CRISPRa technologies inhibit or activate transcription, respectively (Gilbert et al., 2014). CRISPR-cas13 acts on the next layer in the hierarchy of protein production to promote RNA degradation (Wessels et al., 2022). Most small molecule drugs, in contrast, act directly on protein products like enzymes and receptors and can have inhibitory or activating effects. When these techniques are applied to large-scale screens, they create a map between genotype, transcriptome, protein, chromatin accessibility, and in some cases phenotype (Frangieh et al., 2021). Single cells are perturbed using unique CRISPR guides, and their corresponding individual barcodes are read out alongside scRNA-seq, CITE-seq or scATAC-seq reads to identify each cell's perturbation condition (Adamson et al., 2016; Dixit et al., 2016; Frangieh et al., 2021; Rubin et al., 2019). Sequencing with multi-omic readout has been applied to perturbation experiments only recently. CITE-seq, which assesses surface protein counts using oligonucleotide-tagged antibodies measured alongside the transcriptome, has been applied successfully as Perturb-CITE-seq (Frangieh et al., 2021). Efforts have also been made to link Perturb-seq to an ATAC-seq readout (Rubin et al., 2019).

Uses of single-cell perturbation data

Large-scale single-cell perturbation-response screens enable exploration of complex cellular behavior not accessible from bulk observation. Directionality in regulatory network models cannot be inferred without interventional data about the system (Gross et al., 2019). Experiments with targeted perturbations can be modeled as affecting individual nodes of a regulatory network model, while the molecular readouts provide information on the state changes. This creates the opportunity to investigate mechanistic processes and can be used to infer regulatory interactions and their directionality (Pratapa et al., 2020). Typically, however, perturbation datasets have still been too small to elucidate the complexity of a cellular system, and thus accurately predictive models of regulatory interactions remain difficult to infer (Gross and Blüthgen, 2020). This limitation will be reduced as dataset size continues to increase. More directly, drug screens have been used to suggest therapeutic interventions by analyzing detailed molecular effects of targeted drugs,

and designing new single or combinations of perturbations (Bertin et al., 2022; Franz et al., 2021; Preuer et al., 2018).

Motivation for a distance measure for high-dimensional expression profiles

Reliable analysis of increasingly large perturbation datasets requires statistical tools powerful and efficient enough to harness both the massive number of cells and perturbations, as well as the inherently high dimensionality of the data. This high dimensionality complicates calculation of distances between perturbations, as does cell-cell variation and sparsity of data (Kharchenko, 2021). A multivariate distance measure between sets of cells resulting from different perturbations is needed to perform effective statistics comparing perturbations. With such a distance measure, one can describe the difference or similarity between sets of cells treated with distinct perturbations to infer difference or similarity in terms of mechanism or perturbation target; shared mechanisms tend to produce similar shifts in molecular profiles (Replogle et al., 2022; Tian et al., 2021). A number of distance measures for scRNA-seq have been explored by the single-cell community in recent years, including Wasserstein distances (Chen et al., 2020), maximum mean discrepancy (Lotfollahi et al., 2020), neighborhood-based measures (Burkhardt et al., 2021; Dann et al., 2022), E-distance (Replogle et al., 2022), and distances in learnt autoencoder latent spaces (Lotfollahi et al., 2019). Here we exclusively use the E-distance, a fundamental statistical measure of distances between point clouds that can be used in a statistical test to identify strong or weak perturbations as well as to distinguish between perturbations affecting distinct cellular sub-processes (see Methods). This test is a statistically reliable tool for computational diagnostics of information content for a specific perturbation and can inform design of experiments and data selection for training models.

Motivation for unifying datasets

Each large perturbation screen is specifically designed to study a particular system under a set of perturbations of interest. This results in a heterogeneous assortment of single-cell perturbation-response data with a wide range of different cell types, such as immortalized cell lines and iPSC-derived models, and different perturbation technologies, including knockouts, activation, interference, base editing, and prime editing (Przybyla and Gilbert, 2022). Novel computational methods to efficiently harmonize these different perturbation datasets on a large scale are needed. Such integrative analysis is complicated by batch effects and biological differences between primary tissue and cell culture (Forcato et al., 2021; Luecken et al., 2022). Published computational methods for perturbation data are primarily focused on individual datasets (Duan et al., 2019; Jin et al., 2022; Lotfollahi et al., 2019). Moving from single-dataset to multi-dataset analysis will require development of principled quantitative approaches to perturbation biology; the dataset resource based on this work can serve as a foundation for building these models going forward.

Prior work

While large databases of perturbations with bulk readouts exist, single-cell perturbation technologies are newer and data is still not unified (Stathias et al., 2020; Tsherniak et al., 2017). Many collections of datasets are primarily a means for filtering datasets but do not supply a unified format for perturbations. Some of these collections of single-cell datasets were produced to benchmark computational methods for data integration (Lance et al., 2022). Another collection specifically aggregated all available data in the single-cell literature, with well over a thousand datasets described in a sortable table, but with no attempt at data unification and labels focused on observational studies (Svensson et al., 2020). The Broad Institute's Single Cell Portal provides .h5 files for 478 studies with cell type names from a common list but does not harmonize the datasets or allow for filtering by perturbation (Broad Institute, 2022). Yet, unified datasets are key for developing generalizable machine learning methods and establishing multimodal data integration. A recent review and repository of single-cell perturbation data for machine learning lists 22 datasets but supplied cleaned and format-unified data for only 6 (Ji et al., 2021). An existing unified framework for single cell data, called 'sfaira', is ideal for model building and memory efficient data loading, but the public 'data zoo' does not currently supply perturbation datasets or standardized perturbation annotations (Fischer et al., 2021).

Our contribution: scPerturb.org

We aim to provide a resource of standardized datasets reporting targeted perturbations with single-cell readouts and to facilitate the development and benchmarking of computational approaches in system biology. We collected a set of 44 publicly available perturbation-response datasets from 25 papers (Table 1, Supp Fig 3B). Our perturbation strength quantification and comparison of perturbation-specific variables, such as the number of perturbations and the number of cells per perturbation, across experiments may serve as a reference for optimal experimental design of future single-cell perturbation experiments. A web interface for data access, analysis and visualization is available at scperturb.org.

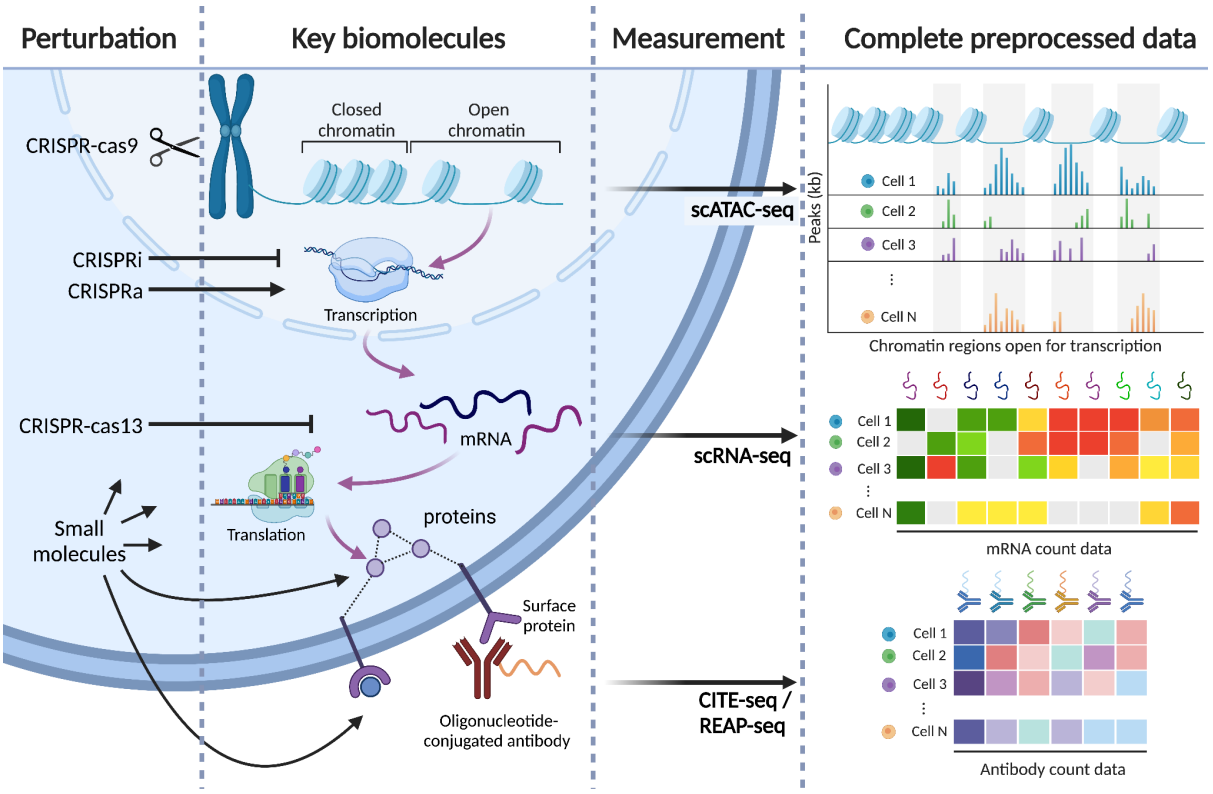


Fig 1: Perturbation-response profiling for single cells. Different perturbations act at different layers in the hierarchy of gene expression and protein production (purple arrows). Perturbations included in scPerturb include CRISPR-cas9, which directly perturbs the genome; CRISPRa, which activates transcription of a target gene; CRISPRi, which blocks transcription of targeted genes; CRISPR-cas13, which cleaves targeted mRNAs and promotes their degradation; cytokines that bind cell surface receptors; and small molecules perturb various cellular mechanisms. Single cell measurements probe the response to perturbation, also at different layers of gene expression: scATAC-seq directly probes chromatin state; scRNA-seq measures mRNA; and protein count data currently is typically obtained via antibodies bound to proteins.

Results

Overview of datasets in the information resource

Molecular readouts for our 44 publicly available single-cell perturbation response datasets include transcriptomics, proteins and epigenomics (Table 1, Figure 2A). 32 datasets in this resource were perturbed using CRISPR and 9 datasets perturbed with drugs. This difference is likely due to the experimental hurdle of applying large numbers of perturbations to cells, effectively requiring some sort of multiplexing, for example using barcodes or single nucleotide polymorphisms. Although it is possible to set up multiplexed sequencing for arrayed treatment conditions, this entails a large amount of manual labor necessary to set up hundreds of separate wells with individual drugs, limiting the total number of drug perturbations in a single experiment. For CRISPR perturbations on the other hand, the mixed set of single guides can be applied in parallel, allowing these experiments to be scaled up massively. While 32 datasets measure scRNA-seq exclusively, we also include three

scATAC datasets, as well as multimodal datasets. For each scRNA-seq dataset we supply count matrices, where each cell has a perturbation annotation, quality control metrics including gene counts and mitochondrial read percentage. Quality control plots for each dataset are also available on scperturb.org (for an example see Supp Fig 1). Three CITE-seq datasets are included with protein and RNA counts separately downloadable (Frangieh et al., 2021; Papalexi et al., 2021).

Choice of features in scATAC-seq data

In contrast to scRNA-seq data, which can be represented naturally as counts per gene, there is no single obvious feature that could be computed for scATAC-seq data, which, in its raw form, provides a noisy and very sparse description of chromatin accessibility over the entire genome. We therefore generated five different feature sets for scATAC-seq data, as motivated in prior studies (Chen et al., 2019; Granja et al., 2021). These either attempt to summarize chromatin accessibility information over different types of biologically relevant genomic intervals (e.g. gene neighborhood), or represent dense low-dimensional embeddings of the original data (see Methods for details). Different features address different biological questions in different contexts (Pierce et al., 2021), which is why we include all five feature sets in our resource.

Count statistics in scRNA-seq

Sample quality measures vary significantly across datasets (Figure 2B). Total unique molecular identifier (UMI) counts per cell and number of genes per cell are calculated as described in (Luecken, Malte D. and Theis, 2019). The average sequencing depth, i.e. the mean number of reads per cell, in each study affects the number of lowly expressed genes observed. Increasing the sequencing depth increases the number of counts measured even for low expression genes, reducing the uncertainty associated with zero counts (Haque et al., 2017; Svensson et al., 2020). These differences can affect the distinguishability of perturbations and performance of downstream analysis methods.

Cells per perturbation

The total number of cells per dataset is usually restricted by experimental limitations, though has increased over time (Supp Fig 2A). Therefore, there is a tradeoff between the number of perturbations and the mean cells per perturbation in a dataset (Fig 2C). Most datasets used one chip / eight lanes of a commercial droplet-based single-cell platform (10X Chromium), with 10,000 cells recommended per lane. The total number of cells was often in the order of multiples of 10,000 cells. The type of perturbation partially dictates the number of cells per perturbation; CRISPR datasets tend to have more perturbations than drug datasets because they are easier to scale up using multiplexing, with a corresponding smaller number of cells per perturbation. This is visible in Figure 2B, where CRISPR datasets are clustered to the left of the plot.

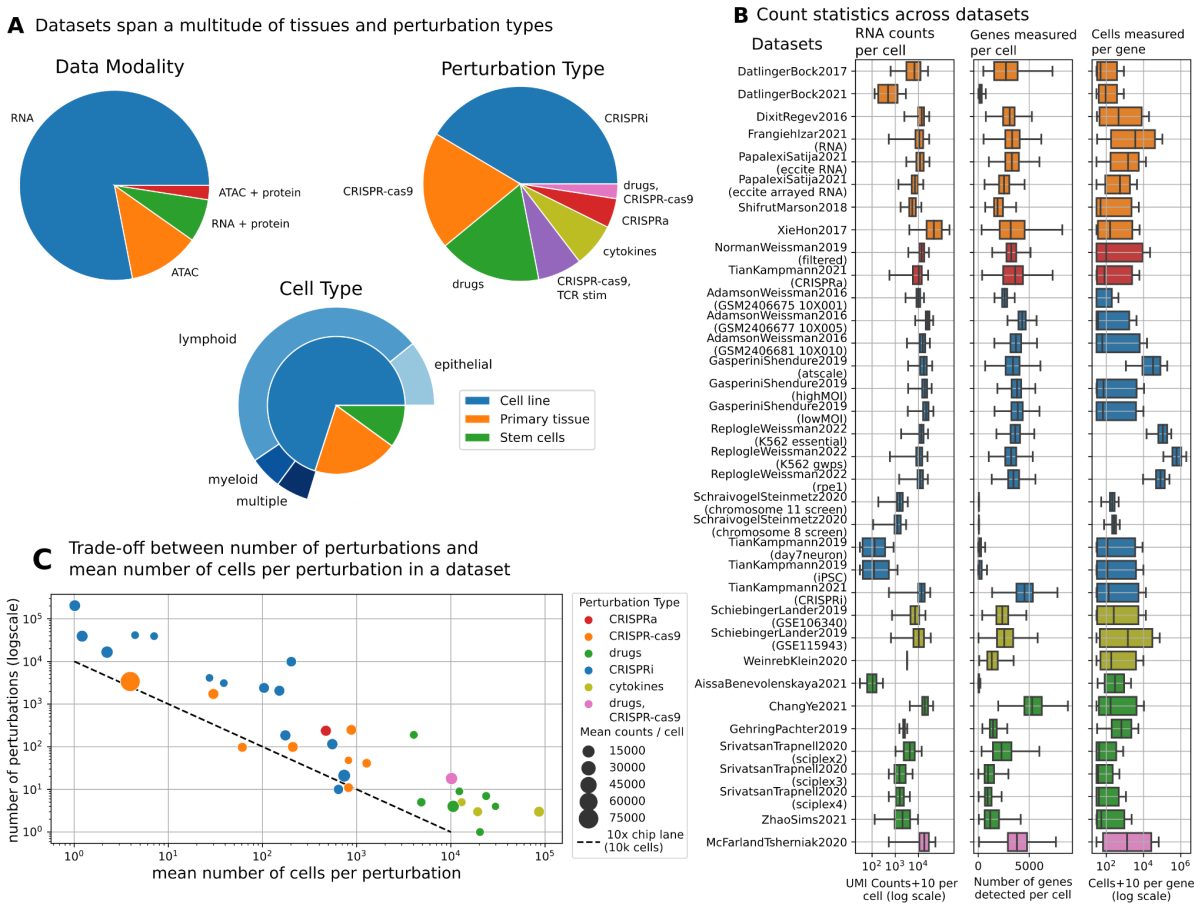


Fig 2: Single cell perturbation-response datasets are diverse in type, size, and quality. (A) The majority of included datasets result from CRISPR (DNA cut, inhibition or activation) perturbations using cell lines derived from various cancers. The studies performed on cells from primary tissues generally use drug perturbations. Primary tissue refers to samples taken directly from patients or mice, sometimes with multiple cell types. (B) Sequencing and cell count metrics across scPerturb perturbation datasets (rows), colored by perturbation type as in Fig 2C. From left to right: Distribution of total RNA counts per cell (left); distribution of the number of genes with at least one count in a cell (middle); distribution of number of cells with at least one count of a gene per gene (right). Most datasets have on average approximately 3000 genes measured per cell, though some outlier datasets have significantly sparser coverage of genes. (C) Each circle represents one dataset. Due to experimental constraints, most datasets have approximately the same number of total cells, pooled across a set of marked perturbations, resulting in a tradeoff between the number of perturbations and the number of cells in each perturbation. CRISPR-perturbation datasets, compared to drug-perturbation datasets, have fewer cells per perturbation but a larger number of perturbations.

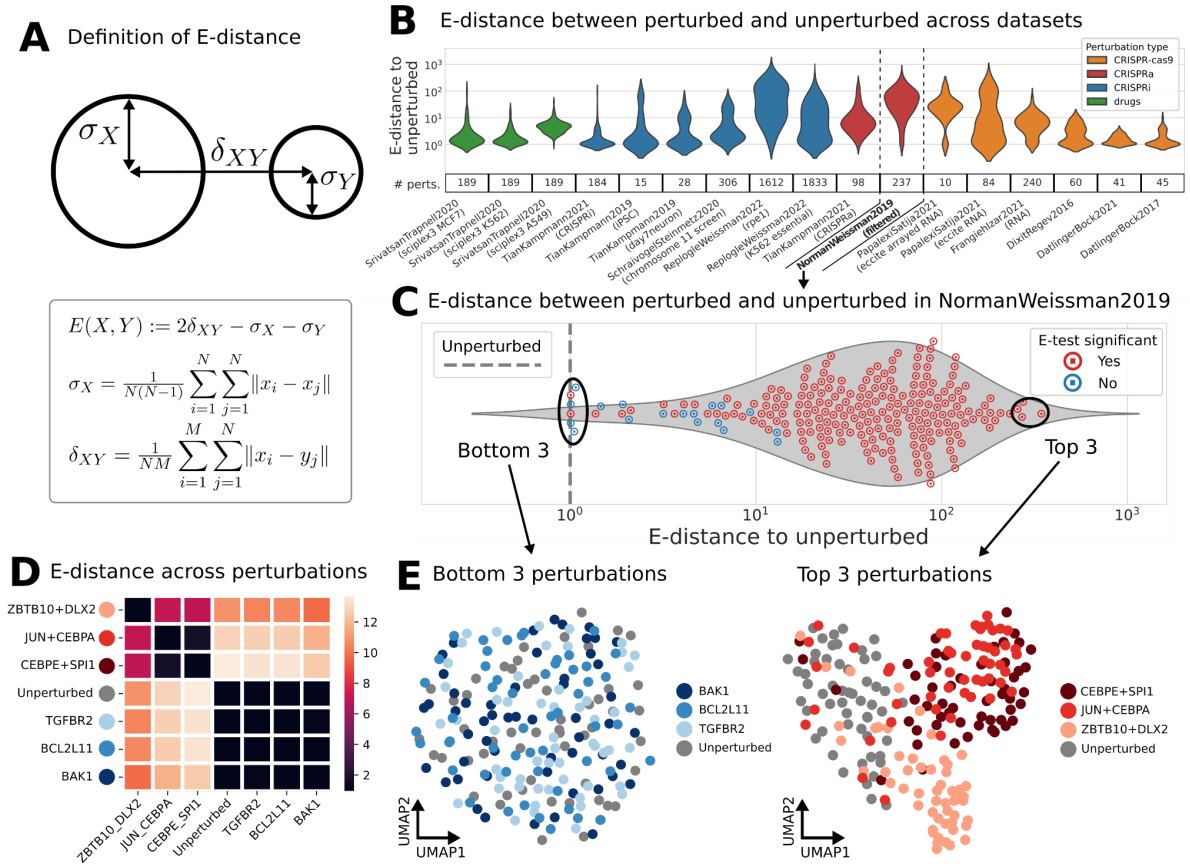


Fig 3: E-statistics describe distinctiveness of perturbations in single-cell data. (A) Definition of E-distance, relating the width of cell distributions of high-dimensional molecular profiles to their distance from each other (see Methods). A large E-distance of perturbed cells from unperturbed indicates a strong change in molecular profile induced by the perturbation. (B) Distribution of E-distances (plus 1 for log scale, same in Fig 3B) between perturbed and unperturbed cells across datasets. The number of perturbations per dataset can be found in Table 1. (C-E) Analysis based on E-statistics for one selected dataset (Norman et al., 2019): (C) Distribution of E-distances between perturbed and unperturbed cells as in Fig 3B. Each circled point is a perturbation, i.e., represents a set of cell profiles. Each perturbation was tested for significant E-distance to unperturbed (E-test). (D) Pairwise E-distance matrix across the top and bottom 3 perturbations of Fig 3C and the unperturbed cells. (E) UMAP of single cells of the weakest (left, Bottom 3) and strongest (right, Top 3) perturbations .

E-distance: introduction

To compare and evaluate perturbations within each dataset we utilized the E-distance, a statistical distance measure between two distributions, which was used as a statistical test in (Replogle et al., 2022). Essentially, the E-distance compares the mean pairwise distance of cells across two different perturbations to the mean pairwise distance of cells within the two distributions (see Methods). If the first is much larger than the latter, the two distributions can be seen as distinct. Similar to Replogle et al., we compute the E-distance after PCA (Principal Component Analysis) for dimensionality reduction. In contrast we use a slightly different, sample-corrected version for the distances within each distribution (see Methods).

E-distance: interpretation

The E-distance provides intuition about the signal-to-noise ratio in a dataset. For two groups of cells, it relates the distance between cells across the groups ("signal"), to the width of each distribution ("noise") (Fig 3A). If this distance is large, distributions are distinguishable, and the corresponding perturbation has a strong effect. A low E-distance indicates that a perturbation did not induce a significant shift in phenotype, reflecting either technical problems in the experiment, ineffectiveness of the perturbation, or perturbation resistance. The E-distance can also be used as a test statistic to test whether cells of a perturbation are significantly different from unperturbed cells ((Replogle et al., 2022), Supp Table 3).

E-distance: dataset example

Interestingly, we found that E-distances between perturbed and unperturbed cells vary strongly across datasets (Fig 3B). The dataset labeled with "NormanWeissman2019" had the largest mean E-distance between all perturbations (Norman et al., 2019) compared to datasets of similar size. In fact, most perturbations in this dataset tested significantly different to unperturbed cells according to the E-test (Fig 3C). We believe this is in part caused by double perturbations with CRISPRa in that dataset; Targeting the same gene with two single guides increases the chances of provoking a considerable change in the transcriptome. This is supported by the fact that the three perturbations with highest E-distance to unperturbed are double-perturbations while the three closest ones are not. The corresponding UMAPs for these perturbations provide a visual intuition for high and low E-distances (Fig 3E). The top three furthest perturbations are easily distinguishable from the gray unperturbed cells, while the bottom three closest perturbations form a single, uniform cloud.

E-distance: perturbation-perturbation distances

The E-distance can also be used to measure similarity between different perturbations. For instance, there is a clear overlap of CEBPE+SPI1 and JUN+CEBPA perturbed cells in the UMAP (Fig 3E). This overlap is captured by the E-distance between the two perturbations; these two perturbations are closer to each other than they are to unperturbed cells or to other perturbations (Fig 3D, Supp Fig 3A). We envision that the E-distance can be used as a suitable distance for other downstream tasks such as drug embeddings and clustering of perturbations, which could allow inference of functional similarity of perturbations by similarity in their induced molecular responses measured by the E-distance.

Dataset highlights

With this assembly of datasets and quality control metrics described below, we were able to nominate notable datasets. The most extensive drug dataset is the sci-Plex 3 dataset with over 188 drugs tested across three cell lines (Srivatsan et al., 2020); 107 of those perturbations were significant according to E-test analysis (Supp Table 3). Five drugs in this dataset also appear in other drug perturbation datasets (Supp

Table 4). We hope that future large-scale drug screens will enable more detailed analysis of drug response across different cell types and conditions. Another drug dataset applies combinations of three drug perturbations at varying concentrations across samples (Gehring et al., 2020). We excluded this dataset from the E-distance analysis due to its complex study design, which was not directly comparable to any other included studies and would require a more adaptive statistical approach. By far the most detailed CRISPR dataset is from a recently published study which perturbed 9867 genes in human cells (ReplogleWeissman2022). Containing >2.5 million cells, this dataset is the largest in our database, with the number of cells each gene is detected in significantly higher than in other datasets (Fig 2B). Notably, 138 CRISPR perturbations are seen in both RNA and ATAC datasets (Supp Table 5). More than 100 genes perturbed with CRISPRa in one dataset are perturbed with CRISPRi perturbations in another dataset of the same cell line (Norman et al., 2019; Replogle et al., 2022; Tian et al., 2021). The most frequently perturbed gene, MYC, is perturbed in 9 datasets from 3 papers. Protein, RNA and ATAC readouts for CRISPRi perturbation of MYC are available from K562 cells (Frangieh et al., 2021; Pierce et al., 2021; Replogle et al., 2022).

Discussion

Concise summary of results

We present a dataset resource, as well an intuitive analytic method for quantifying and analyzing single-cell perturbation datasets. Datasets are described in detail, with additional individualized quality control metrics available on scperturb.org. The use of the E-distance is motivated and applied to quantitatively compare perturbations within each dataset. Our analysis suggests that the majority of single-cell perturbation datasets have a low E-distance to unperturbed cells for most of their perturbations (Fig 3B). For one example dataset, we illustrate how to interpret high and low E-distances and use E-distances to identify functionally similar vs. distinct perturbations. The uniform annotations in this resource will enable data integration and benchmarking as well as exploration of shared perturbations across datasets.

Overlap of perturbations across studies

Joint analysis of single-cell datasets is limited by the complexity of data integration. Across the eight drug datasets examined in this study, only 5 chemical agents occurred in more than one dataset (Supp Table 4). Shared gene targets can be found more often across the CRISPR datasets (Supp Table 5). However, multiplicity of infections and other conditions frequently differ as well, and comparisons are further complicated by distinct perturbation methods (Table 1). The considerable overlap of perturbations across studies makes this a useful resource for benchmarking model generalizability. Especially with the expected arrival of more datasets in the future, we will have the unique opportunity to integrate datasets with more overlapping perturbations and nominate machine learning benchmarks.

scRNA-seq data processing choices

Many published scRNA-seq studies only supply pre-filtered gene count matrices, with different studies using different genome versions. Although we supply ENSEMBL IDs for genes in all datasets, the variation in aligners and genome versions means that the resulting cell-by-gene matrices are not directly comparable. Ideally, one would start the analysis from sequencing reads (FASTQ files) for all datasets, allowing more uniform processing. However, starting from raw data is challenging and considerably increases computational requirements. Given that many datasets only supplied processed data or that raw data was only available after institutional clearance, we determined that the added computation would add limited value to the resource. All datasets were thus processed from the author-supplied cell-by-gene matrix when provided, or were processed from raw when not.

scATAC-seq feature definition for perturbation datasets

Defining the most useful feature matrix definition for scATAC-seq perturbation data is impeded by context dependence. The proposed features were originally developed for scATAC-seq data on non-perturbed cells, with goals such as the identification of cell types, discovery of cell type-specific regulatory elements, or reconstruction of cellular differentiation trajectories (Buenrostro et al., 2013; Satpathy et al., 2019). They were not optimized to capture the changes in chromatin accessibility induced by targeted perturbations. This reveals a potential for further research into the refinement of scATAC-seq features, particularly in the case of CRISPR perturbations where the locus of action is known and could guide feature definition (Chen et al., 2019).

Growing memory demands

Resource users should be aware that memory requirements quickly become a limiting factor, especially with the newer, larger datasets, such as ReplogleWeissman2022 with >2.5 million cells across more than 9000 perturbations (Replogle et al., 2022). For example, the E-distance presented here for calculating distances between perturbed sets of cells relies on principal component analysis (PCA), but computing PCA for all data in this dataset was not possible with 500GB of memory without modifications to accelerate computation. Going forward, computational methods will need to be modified as in (Dhapola et al., 2022) to reduce memory load, or datasets will need to be subsampled. The need for processing capacity could hamper method development in the future, particularly if cells from multiple datasets are combined. Access to high-performance clusters may be a requirement for analysis workflows of the kind described here.

Considerations for single cell perturbation experimental design

Experimental design choices such as the optimal number of cells per perturbation and the sequencing depth for each cell depend on the questions the dataset is intended to answer, and on the strength and uniqueness of the gene expression

changes caused by the perturbations (Fig 2B). If technical noise degrades the characteristic response of a perturbation, increasing the sequencing depth probably will improve distinguishability. Unfortunately, it is difficult to ascertain to what extent a low E-distance between perturbed and unperturbed cells in the data is caused by technical noise. Increasing the dose or varying the time between perturbation start and harvesting the cells may be advisable to increase the signal to noise ratio. Future investigations into downsampling counts and cells per perturbation across the scPerturb database could be informative for establishing minimum count recommendations.

Conclusion

We envision that the scPerturb collection of datasets and the suggested analytic framework using the E-distance will be a valuable starting point for the analysis of single-cell perturbation data. The unified annotations and our estimates of signal-to-noise ratios across datasets should prove especially useful to the machine learning community for training models on this type of data. We anticipate new datasets and experimental perturbation methods in the future which will enable the community to innovate novel computational approaches to exploit the increasing amount and complexity of single-cell perturbation data, aiming at the development of increasingly accurate and quantitatively predictive models of cell biological processes and the design of targeted interventions for investigational or therapeutic purposes.

Data Availability

The website scperturb.org stores harmonized datasets with the following:

- scRNA-seq and antibody-based protein datasets: .h5ad files and .mtx files are available, which can be easily read with python or R scripts.
- scATAC-seq: multiple different feature matrix definitions as separate download options.
- Access details for the original publication for each dataset
- Quality control plots for each dataset
- Filtering, e.g., by readout or type of perturbation
- Commands for direct file download using the Unix command curl

Open access source code is at https://github.com/sanderlab/scPerturb/ .
Secondary upload of RNA data at https://zenodo.org/record/7041849 and ATAC data at https://zenodo.org/record/7058382

Table 1: Key metadata for datasets on scPerturb.org. More details in Supp Table 1.

Source Paper	Modality	Perturbation type	Number of perturbations
(Adamson et al., 2016)	RNA	CRISPRi	9, 20, 114
(Aissa et al., 2021)	RNA	drugs	4
(Chang et al., 2022)	RNA	drugs	4
(Datlinger et al., 2017)	RNA	CRISPR-cas9+TCR[&]	97
(Datlinger et al., 2021)	RNA	CRISPR-cas9+TCR[&]	48
(Dixit et al., 2016)	RNA	CRISPR-cas9	31
(Frangieh et al., 2021)	RNA + protein	CRISPR-cas9	249
(Gasperini et al., 2019)	RNA	CRISPRi	43314*, 39087*, 16531*
(Gehring et al., 2020)	RNA	drugs	4
(Liscovitch-Brauer et al., 2021)	ATAC	CRISPR-cas9	22,84
(McFarland et al., 2020)	RNA	drugs, CRISPR-cas9	18
(Mimitou et al., 2021)	ATAC + protein	CRISPR-cas9	6
(Norman et al., 2019)	RNA	CRISPRa	237
(Papalexi et al., 2021)	RNA + protein	CRISPR-cas9	11,99
(Pierce et al., 2021)	ATAC	CRISPRi	41,41,41
(Replogle et al., 2022)	RNA	CRISPRi	2058, 2394, 9867
(Schiebinger et al., 2019)	RNA	cytokines	2,3
(Schraivogel et al., 2020)	RNA	CRISPR-cas9	3105*, 4115*
(Shifrut et al., 2018)	RNA	CRISPR-cas9+TCR[&]	49
(Srivatsan et al., 2020)	RNA	drugs	5, 8, 189
(Tian et al., 2019)	RNA	CRISPRi	27
(Tian et al., 2021)	RNA	CRISPRa, CRISPRi	101, 185
(Weinreb et al., 2020)	RNA	cytokines	5
(Xie et al., 2017)	RNA	CRISPR-cas9	229
(Zhao et al., 2021)	RNA	drugs	7

*: perturbation total treats perturbations A, B, and (A and B) as three unique perturbations

[&]: TCR receptor stimulation

Methods

scATAC-seq

Data acquisition
We included scATAC-seq data from three different sources: Spear-ATAC (Pierce et al., 2021), CRISPR-sciATAC (Liscovitch-Brauer et al., 2021), and ASAP-seq (Mimitou et al., 2019). All data that was used in our analysis can be programmatically downloaded with scripts that are provided in our code repository (https://github.com/sanderlab/scPerturb).

scATAC-seq is a biomolecular technique to assess chromatin accessibility within single cells (Buenrostro et al., 2015; Cusanovich et al., 2015). The starting point of our data processing pipeline are BED-like tabular fragment files, in which each line represents a unique ATAC-seq fragment captured by the assay. Each fragment is mapped to a genomic interval and a cell barcode. The goal of our pipeline is to extract standardized features from this information. Those are:

- Embeddings derived from Latent-Semantic-Indexing (LSI) (Cusanovich et al., 2015) with 30 dimensions for each cell (a dimensionality reduction method that is well-suited for the sparsity of the data)
- Gene scores that measure the chromatin accessibility around each gene for each cell (the weighted sum of fragment counts around the neighborhood of a gene's transcription start site where more distant counts contribute less)
- A peak-barcode matrix that quantifies the chromatin accessibility at (data-set specific) consensus peaks (genomic intervals) for each cell
- ChromVar scores (Schep et al., 2017), which quantify the activity of a set of transcription factors for each cell, using transcription factor footprints as defined in (Vierstra et al., 2020)
- Marker-peaks per perturbation target, quantifying the differential regulation of highly variable peaks for each type of perturbation

These features were computed using the ArchR framework version 1.0.1 (Granja et al., 2021) with standard parameters unless otherwise stated. We provide each feature set as a dedicated h5ad file on scperturb.org, and our analysis roughly follows the pipeline proposed in Spear-ATAC (Pierce et al., 2021), as detailed below.

Pre-processing
Filtering out cells of low quality: To ensure a consistent and homogenous quality throughout the different data sets, we filtered out cells with fewer than 1000 and more than 100,000 mapped fragments. We further required a minimum transcription start site enrichment score of 4 to ensure a sufficient signal to noise ratio. See ArchR's 'createArrowFile' function for details.

For the Spear-ATAC data set we ran ArchR's getValidBarcodes function on processed 10x cell ranger files to subset the data set to valid barcodes. For the other datasets these files were unavailable, and we relied on the original authors' pre-processing of barcodes.

Assigning sgRNAs to barcodes: For the Spear-ATAC and CRISPR_sciATAC datasets we had access to cell barcode-sgRNA count matrices (see original publications for details). We assigned the sgRNA with the highest counts to a cell barcode if the sgRNA count exceeded 20 and if that sgRNA combined at least 80% of all sgRNA counts. Cells that could not be assigned a sgRNA were left in the data set. For the ASAP-seq dataset a barcode-sgRNA matrix was not available. Instead, we relied on a sgRNA assignment downloaded from the study's GitHub repository (Lareau, 2021).

Feature computation

All features described in the overview above were computed with ArchR functions. For details inspect the "fragments2outputs.R" script in our code repository (see Data Availability).

scRNA-seq

Data acquisition

Datasets were downloaded from public databases following data availability directions in the source papers. When available from the authors, unnormalized pre-processed cell-by-gene matrices were used. Supplemental information from the papers were used in data analysis when applicable.

Data processing

Analysis started from unfiltered, unnormalized cell-by-gene matrices as provided by source papers. For one dataset, preprocessed cell-by-gene matrices were unavailable; pre-processing was performed following the procedure outlined in the original paper, directly using supplied code (Gehring et al., 2020). For datasets with cell barcodes, barcode assignments for cells were taken from the original paper when available; when not available, barcode assignment was performed as described in the methods section of the paper. If multiple guides were assigned to the same cell, the guides were listed in decreasing order of counts in the final data object. The code used for processing each individual dataset, including barcode assignment, is available in our code repository.

Datasets were imported into AnnData objects using Scanpy (versions 1.7.2–1.9.1) (Wolf et al., 2018). Metadata was taken from the original papers when available. For cell lines, information on sex, age, disease, and origin were taken from Cellosaurus

(Bairoch, 2018). Metadata columns are described in (Supp Table 2 here). Items listed in **bold** are included for all datasets.

Datasets are saved as .h5ad files and as .mtx files with obs and var as separate .csv files. Code is supplied in our code repository for the import of .mtx files into Seurat.

Data analysis

Before calculating the E-distances (Figure 4), cells and genes were filtered using Scanpy (versions 1.7.2–1.9.1) (Wolf et al., 2018). Cells were kept if they had a minimum of 1000 UMI counts, and genes with a minimum of 50 cells. 2000 highly variable genes were selected using find_variable_genes with flavor 'seurat_v3'. We computed PCA based on these genes. The E-distances were computed in that PCA space using 50 components and Euclidean distance. To avoid problems due to different numbers of cells per perturbation, we subsampled each dataset such that all perturbations had the same number of cells. We removed all perturbations with fewer than 50 cells and then subsampled to the number of cells in the smallest perturbation left after filtering.

E-distance

The E-distance is a statistical distance between high-dimensional distributions and has been used to define a multivariate two-sample test, called the E-test (Rizzo and Székely, 2016). It is more commonly known as energy distance, stemming from the original interpretation using gravitational energy in physics. Formally, it contextualizes the notion that two distributions of points in a high-dimensional space are distinguishable if they are far apart compared to the width of both distributions (Fig 3A). More specifically,

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ and $y_1, \ldots, y_M \in \mathbb{R}^d$ be samples from two distributions X,Y corresponding to two sets of *N* and *M* cells, respectively.

We define

$$\delta_{XY} = \frac{1}{NM} \sum_{i=1}^{M} \sum_{j=1}^{N} \|x_i - y_j\|$$

$$\sigma_X = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \|x_i - x_j\|$$

and σ_Y defined accordingly. Intuitively, δ_{XY} is the mean distance between cells from the two distributions, while σ_X describes the mean distance between a cell from X to another cell from X. Usually, definitions of σ_X in the literature use the factor $1/N^2$. We argue that the sum in σ_X contains zero summands when $i = j$. To correct for this, we instead use the factor $1/(N \cdot (N-1))$.

The energy distance between X and Y is defined as:

$$E(X,Y) := 2\delta_{XY} - \sigma_X - \sigma_Y$$

E-test calculation

The E-test was performed as a Monte Carlo permutation test using the E-distance as test statistic. For each dataset and each perturbation within that dataset, we took the cells and combined them with the unperturbed cells. Then, we shuffled the perturbation labels and computed the E-distance between the two resulting groups. We repeated this process 100 times. The number of times that this shuffled E-distance to unperturbed was larger than the unshuffled one divided by 100 yields a p-value, which we report for almost all datasets in our resource (Supp Fig 3B, Supp Table 3). We corrected for multiple testing using the Holm-Sidak method per dataset.

Advice for single-cell perturbation analysis

To our knowledge, there are not yet established best practices for analysis of single-cell perturbation data. DESeq2 is frequently used for differential expression testing, as it can be applied to pseudo-bulk profiles of each perturbation (Love et al., 2014). An optional next step would be enrichment analysis of the resulting genes. Averaging single-cell measurements over cells per perturbation simplifies analysis and reduces the effect of measurement noise significantly but comes at the cost of removing all system-intrinsic biologically relevant information in cell-to-cell variation. In many studies, these average profiles are then embedded using a dimensionality reduction method of choice and subsequently clustered to reveal groups of perturbations with potentially similar targets (Norman et al., 2019; Replogle et al., 2022).

Funding / Acknowledgements

- National Resource for Network Biology (NRNB, P41GM103504)
- Supported by the Wellcome Leap ΔTissue Program
- Deutsche Forschungsgemeinschaft (DFG, RTG2424 CompCancer)
- Einstein Stiftung Berlin (Einstein visiting fellow program)
- Computation was in part performed on the HPC for Research cluster of the Berlin Institute of Health.

References

Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., 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.e21. https://doi.org/10.1016/j.cell.2016.11.048.

Aissa, A.F., Islam, A.B.M.M.K., Ariss, M.M., Go, C.C., Rader, A.E., Conrardy, R.D., Gajda, A.M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., et al. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. Nat. Commun. *12*, 1628. https://doi.org/10.1038/s41467-021-21884-z.

Bairoch, A. (2018). The Cellosaurus, a Cell-Line Knowledge Resource. J. Biomol. Tech. JBT *29*, 25–38. https://doi.org/10.7171/jbt.18-2902-002.

Bertin, P., Rector-Brooks, J., Sharma, D., Gaudelet, T., Anighoro, A., Gross, T., Martinez-Pena, F., Tang, E.L., S, S.M., Regep, C., et al. (2022). RECOVER: sequential model optimization platform for combination drug repurposing identifies novel synergistic compounds in vitro. https://doi.org/10.48550/arXiv.2202.04202.

Broad Institute (2022). Single Cell Portal.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218. https://doi.org/10.1038/nmeth.2688.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. Nature *523*, 486–490. https://doi.org/10.1038/nature14590.

Burkhardt, D.B., Stanley, J.S., Tong, A., Perdigoto, A.L., Gigante, S.A., Herold, K.C., Wolf, G., Giraldez, A.J., van Dijk, D., and Krishnaswamy, S. (2021). Quantifying the effect of experimental perturbations at single-cell resolution. Nat. Biotechnol. *39*, 619–629. https://doi.org/10.1038/s41587-020-00803-5.

Chang, M.T., Shanahan, F., Nguyen, T.T.T., Staben, S.T., Gazzard, L., Yamazoe, S., Wertz, I.E., Piskol, R., Yang, Y.A., Modrusan, Z., et al. (2022). Identifying transcriptional programs underlying cancer drug response with TraCe-seq. Nat. Biotechnol. *40*, 86–93. https://doi.org/10.1038/s41587-021-01005-3.

Chen, H., Lareau, C., Andreani, T., Vinyard, M.E., Garcia, S.P., Clement, K., Andrade-Navarro, M.A., Buenrostro, J.D., and Pinello, L. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. *20*, 241. https://doi.org/10.1186/s13059-019-1854-5.

Chen, W.S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., and Krishnaswamy, S. (2020). Uncovering axes of variation among single-cell cancer specimens. Nat. Methods *17*, 302–310. https://doi.org/10.1038/s41592-019-0689-z.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science *348*, 910–914. https://doi.org/10.1126/science.aab1601.

Dann, E., Henderson, N.C., Teichmann, S.A., Morgan, M.D., and Marioni, J.C. (2022). Differential abundance testing on single-cell data using k-nearest

neighbor graphs. Nat. Biotechnol. *40*, 245–253. https://doi.org/10.1038/s41587-021-01033-z.
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. https://doi.org/10.1038/nmeth.4177.
Datlinger, P., Rendeiro, A.F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., and Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. Nat. Methods *18*, 635–642. https://doi.org/10.1038/s41592-021-01153-z.
Dhapola, P., Rodhe, J., Olofzon, R., Bonald, T., Erlandsson, E., Soneji, S., and Karlsson, G. (2022). Scarf enables a highly memory-efficient analysis of large-scale single-cell genomics data. Nat. Commun. *13*, 4616. https://doi.org/10.1038/s41467-022-32097-3.
Dixit, A., Parnas, O., Li, B., and Chen, J. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*. .
Duan, B., Zhou, C., Zhu, C., Yu, Y., Li, G., Zhang, S., Zhang, C., Ye, X., Ma, H., Qu, S., et al. (2019). Model-based understanding of single-cell CRISPR screening. Nat. Commun. *10*, 2233. https://doi.org/10.1038/s41467-019-10216-x.
Fischer, D.S., Dony, L., König, M., Moeed, A., Zappia, L., Heumos, L., Tritschler, S., Holmberg, O., Aliee, H., and Theis, F.J. (2021). Sfaira accelerates data and model reuse in single cell genomics. Genome Biol. *22*, 248. https://doi.org/10.1186/s13059-021-02452-6.
Forcato, M., Romano, O., and Bicciato, S. (2021). Computational methods for the integrative analysis of single-cell data. Brief. Bioinform. *22*, bbaa042. https://doi.org/10.1093/bib/bbaa042.
Frangieh, C.J., Melms, J.C., Thakore, P.I., Geiger-Schuller, K.R., Ho, P., Luoma, A.M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M.S., et al. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. Nat. Genet. *53*, 332–341. https://doi.org/10.1038/s41588-021-00779-1.
Franz, A., Coscia, F., Shen, C., Charaoui, L., Mann, M., and Sander, C. (2021). Molecular response to PARP1 inhibition in ovarian cancer cells as determined by mass spectrometry based proteomics. J. Ovarian Res. *14*, 140. https://doi.org/10.1186/s13048-021-00886-x.
Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., et al. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell *176*, 377-390.e19. https://doi.org/10.1016/j.cell.2018.11.029.
Gehring, J., Hwee Park, J., Chen, S., Thomson, M., and Pachter, L. (2020). Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. Nat. Biotechnol. *38*, 35–38. https://doi.org/10.1038/s41587-019-0372-z.
Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661. https://doi.org/10.1016/j.cell.2014.09.029.
Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for

integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411. https://doi.org/10.1038/s41588-021-00790-6.
Gross, T., and Blüthgen, N. (2020). Identifiability and experimental design in perturbation studies. Bioinformatics *36*, i482–i489. https://doi.org/10.1093/bioinformatics/btaa404.
Gross, T., Wongchenko, M.J., Yan, Y., and Blüthgen, N. (2019). Robust network inference using response logic. Bioinformatics *35*, i634–i642. https://doi.org/10.1093/bioinformatics/btz326.
Haque, A., Engel, J., Teichmann, S.A., and Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. *9*, 75. https://doi.org/10.1186/s13073-017-0467-4.
Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., Oudenaarden, A. van, and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*, 1883-1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.
Ji, Y., Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2021). Machine learning for perturbational single-cell omics. Cell Syst. *12*, 522–537. https://doi.org/10.1016/j.cels.2021.05.016.
Jin, K., Schnell, D., Li, G., Salomonis, N., Prasath, V.B.S., Szczesniak, R., and Aronow, B.J. (2022). CellDrift: Inferring Perturbation Responses in Temporally-Sampled Single Cell Data. 2022.04.13.488194. https://doi.org/10.1101/2022.04.13.488194.
Kharchenko, P.V. (2021). The triumphs and limitations of computational methods for scRNA-seq. Nat. Methods *18*, 723–732. https://doi.org/10.1038/s41592-021-01171-x.
Lance, C., Luecken, M.D., Burkhardt, D.B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., et al. (2022). Multimodal single cell data integration challenge: results and lessons learned. 2022.04.11.487796. https://doi.org/10.1101/2022.04.11.487796.
Lareau, Caleb.A. (2021). asap_reproducibility (https://raw.githubusercontent.com/caleblareau/asap_reproducibility/master/CD4_CRISPR_asapseq/output/Signac/after_filter_Signac/HTO_res_filtered.txt).
Liscovitch-Brauer, N., Montalbano, A., Deng, J., Méndez-Mancilla, A., Wessels, H.-H., Moss, N.G., Kung, C.-Y., Sookdeo, A., Guo, X., Geller, E., et al. (2021). Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. Nat. Biotechnol. *39*, 1270–1277. https://doi.org/10.1038/s41587-021-00902-x.
Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. Nat. Methods *16*, 715–721. https://doi.org/10.1038/s41592-019-0494-8.
Lotfollahi, M., Naghipourfar, M., Theis, F.J., and Wolf, F.A. (2020). Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics *36*, i610–i617. https://doi.org/10.1093/bioinformatics/btaa800.
Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.

Luecken, Malte D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746. https://doi.org/10.15252/msb.20188746.

McFarland, J.M., Paolella, B.R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W.N., Jones, A., Chambers, E., et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat. Commun. *11*, 4296. https://doi.org/10.1038/s41467-020-17440-w.

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. https://doi.org/10.1038/s41592-019-0392-0.

Mimitou, E.P., Lareau, C.A., Chen, K.Y., Zorzetto-Fernandes, A.L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B.Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol. *39*, 1246–1258. https://doi.org/10.1038/s41587-021-00927-2.

Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science *365*, 786–793. https://doi.org/10.1126/science.aax4438.

Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., Wessels, H.-H., Hao, Y., 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. https://doi.org/10.1038/s41588-021-00778-2.

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. https://doi.org/10.1038/s41467-021-23213-w.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154. https://doi.org/10.1038/s41592-019-0690-6.

Preuer, K., Lewis, R.P.I., Hochreiter, S., Bender, A., Bulusu, K.C., and Klambauer, G. (2018). DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics *34*, 1538–1546. https://doi.org/10.1093/bioinformatics/btx806.

Przybyla, L., and Gilbert, L.A. (2022). A new era in functional genomics screens. Nat. Rev. Genet. *23*, 89–103. https://doi.org/10.1038/s41576-021-00409-w.

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.e28. https://doi.org/10.1016/j.cell.2022.05.013.

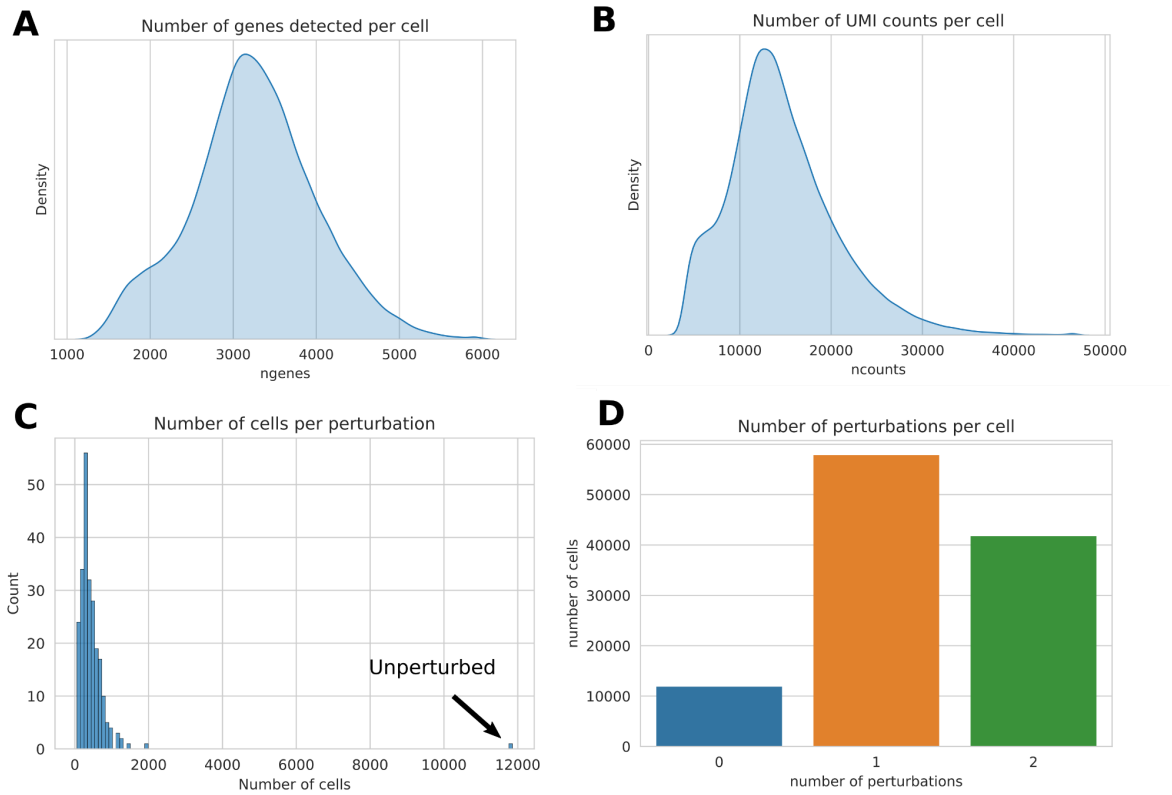
Rizzo, M.L., and Székely, G.J. (2016). Energy distance. WIREs Comput. Stat. *8*, 27–38. https://doi.org/10.1002/wics.1375.

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

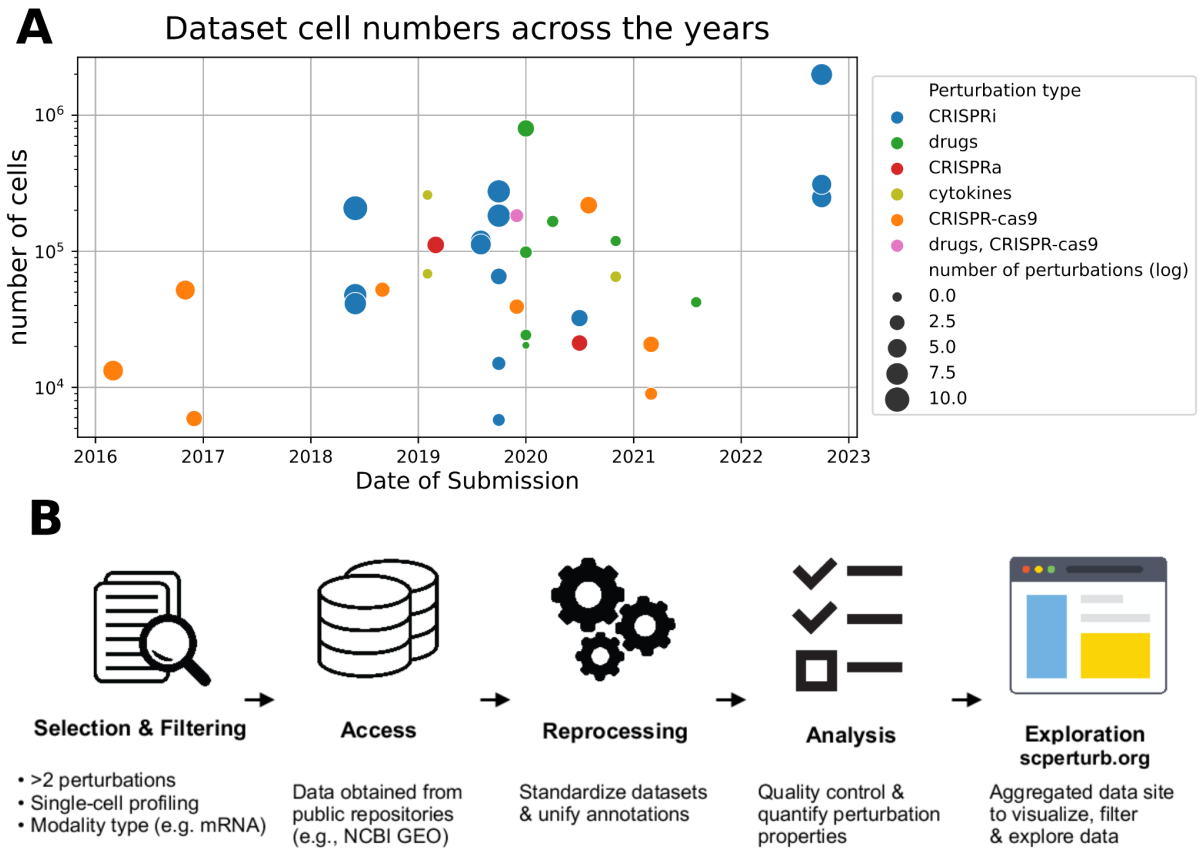
Networks. Cell *176*, 361-376.e17. https://doi.org/10.1016/j.cell.2018.11.022.
Satpathy, A.T., Granja, J.M., Yost, K.E., Qi, Y., Meschi, F., McDermott, G.P., Olsen, B.N., Mumbach, M.R., Pierce, S.E., Corces, M.R., et al. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat. Biotechnol. *37*, 925–936. https://doi.org/10.1038/s41587-019-0206-z.
Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978. https://doi.org/10.1038/nmeth.4401.
Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell *176*, 928-943.e22. https://doi.org/10.1016/j.cell.2019.01.006.
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. https://doi.org/10.1038/s41592-020-0837-5.
Shifrut, E., Carnevale, J., Tobin, V., Roth, T.L., Woo, J.M., Bui, C.T., Li, P.J., Diolaiti, M.E., Ashworth, A., and Marson, A. (2018). Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. Cell *175*, 1958-1971.e15. https://doi.org/10.1016/j.cell.2018.10.024.
Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., 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. https://doi.org/10.1126/science.aax6234.
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. https://doi.org/10.1093/nar/gkz1023.
Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. Database *2020*, baaa073. https://doi.org/10.1093/database/baaa073.
Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron *104*, 239-255.e12. https://doi.org/10.1016/j.neuron.2019.07.014.
Tian, R., Abarientos, A., Hong, J., Hashemi, S.H., Yan, R., Dräger, N., Leng, K., Nalls, M.A., Singleton, A.B., Xu, K., et al. (2021). Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nat. Neurosci. *24*, 1020–1034. https://doi.org/10.1038/s41593-021-00862-0.
Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010.
Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736.

https://doi.org/10.1038/s41586-020-2528-x.
Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381. https://doi.org/10.1126/science.aaw3381.
Wessels, H.-H., Méndez-Mancilla, A., Papalexi, E., Mauck, W.M., Lu, L., Morris, J.A., Mimitou, E., Smibert, P., Sanjana, N.E., and Satija, R. (2022). Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq (Genomics).
Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.
Xie, S., Duan, J., Li, B., Zhou, P., and Hon, G.C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell *66*, 285-299.e5. https://doi.org/10.1016/j.molcel.2017.03.007.
Zhao, W., Dovas, A., Spinazzi, E.F., Levitin, H.M., Banu, M.A., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M.L., Sisti, M.B., et al. (2021). Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. Genome Med. *13*, 82. https://doi.org/10.1186/s13073-021-00894-y.

Supplement

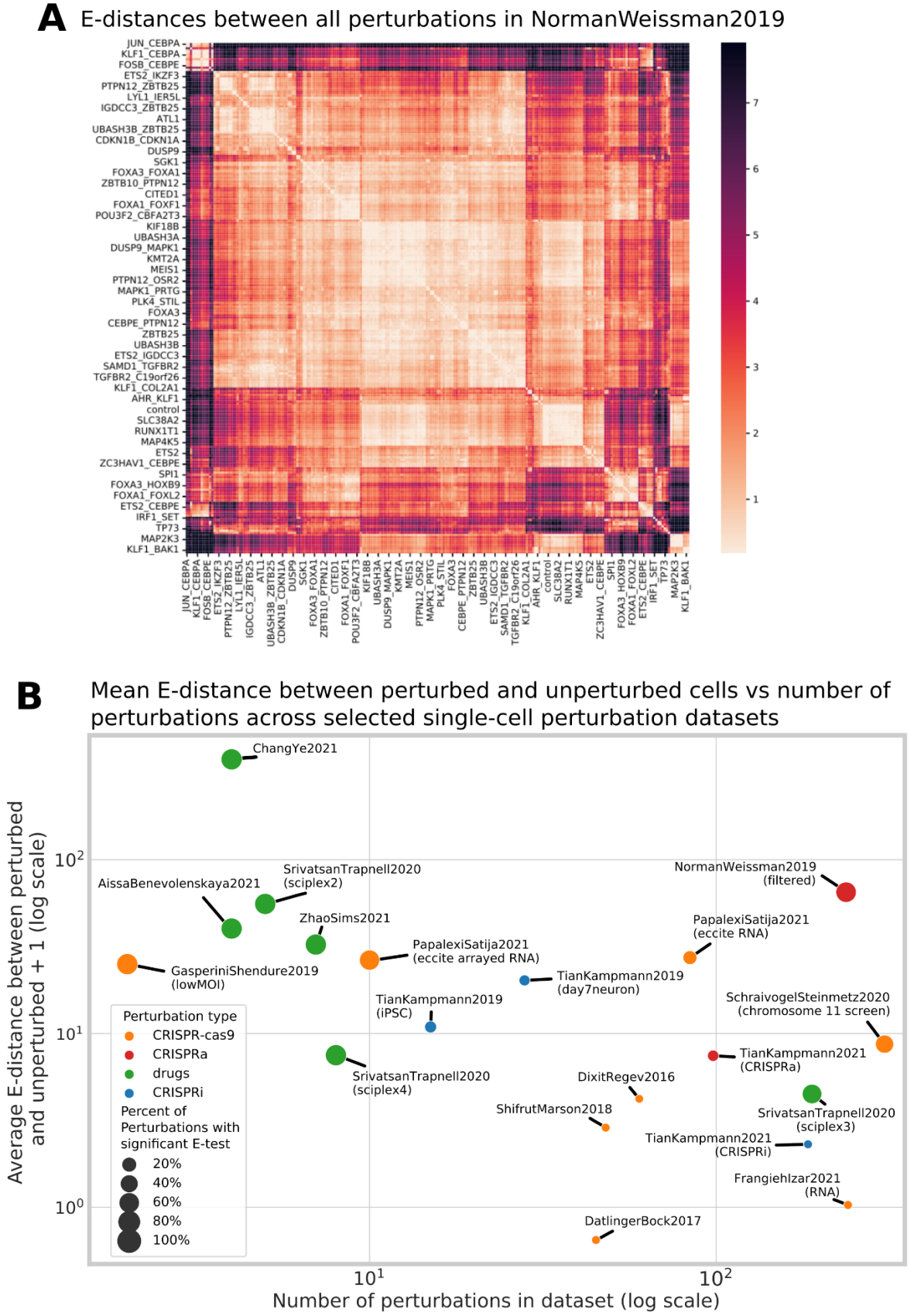


Supp Fig 1: Exemplary information provided for each scPerturb dataset (here for (Norman et al., 2019)) (A) Number of genes that are detected with at least one count in a cell across all cells. (B) Total number of UMI counts per cell across all cells. Together with Supp Fig 1A this provides an overview over both sparsity and quality of the dataset. (C) Number of cells per perturbation. Depending on the application, perturbations with few cells can be filtered out before down-stream analysis. High imbalance in cell numbers per perturbation may also lead to biases in models. (D) Number of cells which received none, a single one, or two perturbations.



Supp Fig 2A: Number of cells per dataset by submission date. There is a rapid increase in published single-cell perturbation datasets around 2019. We speculate that the slight decrease of dataset numbers after 2021 suggested by the plot is due to the ongoing impact of reduced research in the earlier phases of the COVID-19 pandemic.

Supp Fig 2B: Harmonization and analysis workflow. Perturbation datasets with single-cell molecular profiles with at least two perturbations and one control condition (e.g. unperturbed) of various modality types were identified in a literature search. Data was obtained from public repositories, and metadata (such as guide identity) from paper supplements. Datasets were reprocessed to standardize annotations and analyzed in parallel. All datasets are now available for download from scperturb.org, along with visualizations and summarizing information



Supp Fig 3: E-distance extended plots. (A) E-distances between all pairs of perturbations in the dataset NormanWeissman2019. The color scale is clipped at 5% highest and lowest percentiles.

Clusters of similar perturbations are visible, e.g. a cluster of strongly acting perturbations targeting CEBPA at the top. (B) Mean E-distance of perturbed to unperturbed cells in each dataset plotted against the total number of perturbations in the dataset (both in log scale). This dataset (Norman et al., 2019) has many perturbations with a considerable mean E-distance to unperturbed.