scPerturb: 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[2,8], Bo Yuan[4,5,6], Linus J. Schumacher[9], 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
8: Institute for Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany
9: Centre for Regenerative Medicine, University of Edinburgh, Edinburgh, UK
x: joint first authors
+: joint senior authors

Abstract

Recent biotechnological advances led to growing numbers of single-cell perturbation 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, and data filtering. 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. In addition, we introduce E-statistics for perturbation effect quantification and significance testing, and demonstrate E-distance as a general distance measure for single cell data. Using these datasets, we illustrate the application of E-statistics for quantifying perturbation similarity and efficacy. The data and a package for computing E-statistics is publicly available at scperturb.org. This work provides an information resource and guide for researchers working with single-cell perturbation data, highlights conceptual considerations for new experiments, and makes concrete recommendations for optimal cell counts and read depth.

Introduction

[Definition of single-cell perturbation data]

Perturbation experiments probe the response of cells or cellular systems to changes in conditions. These changes traditionally acted 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 using different technologies target different layers of the hierarchy of protein production (Fig 1). At the lowest layer, CRISPR-cas9 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). Newer 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 or time-series 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 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 and the inherently high dimensionality of the data. This high dimensionality complicates calculation of distances between perturbations, as does cell-cell variation and data sparsity (Kharchenko, 2021). There is not presently a convention for the statistical comparison measure used in perturbation studies. Some studies calculate pseudo-bulk by combining all cells in a given perturbation (Adamson et al., 2016; Datlinger et al., 2017). This means losing any information about the variation within each cell type. Studies with mixtures of cell types do the opposite, developing complex methods for quantifying similarity between heterogeneous cell populations (Burkhardt et al., 2021; Dann et al., 2022; Gehring et al., 2020). Ideally, one would perform statistical comparisons to identify similar perturbations and classify perturbation strength using a multivariate distance measure between sets of cells. Such a distance measure describes the difference or similarity between sets of cells treated with distinct perturbations, thus inferring 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), 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). Existing 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]

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. We also describe the E-distance and E-test as tools for statistical comparisons of sets of cells and benchmark their robustness and applicability for distinguishing both perturbations and cell types across different datasets and modalities. A web interface for data access, analysis and visualization is available at scperturb.org, and a Python implementation of e-distance based statistics for single cell data is publicly available as `scperturb` on PyPI.

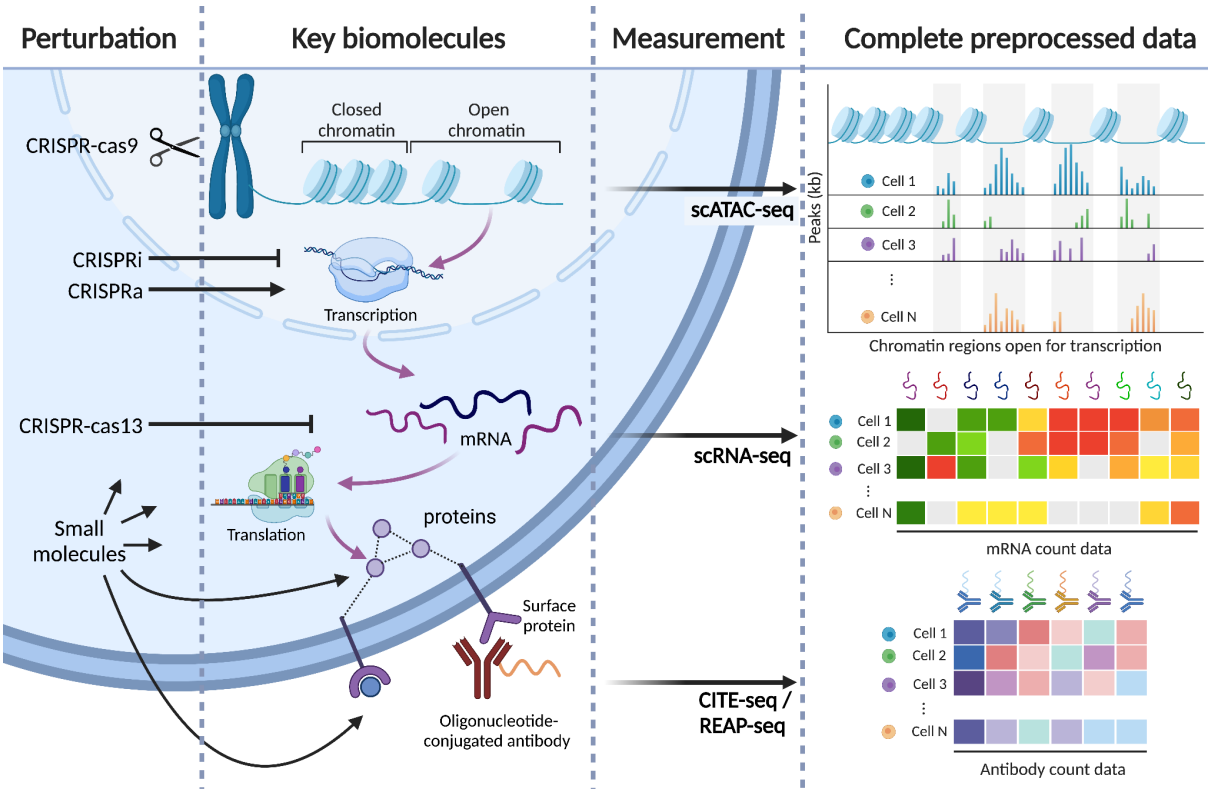


Fig 1: Perturbation-response profiling for single cells. Different perturbations act at different layers in the hierarchy of gene expression and protein production (purple arrows). Perturbations included in scPerturb include CRISPR-cas9, which directly perturbs the genome; CRISPRa, which activates transcription 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 transcriptomes, proteins and epigenomes (Table 1, Fig 2A). Metadata was harmonized across datasets (Supp Table 2). 32 datasets in this resource were perturbed using CRISPR and 9 datasets perturbed with drugs. This paucity of drug datasets is likely due to the experimental hurdle of applying large numbers of perturbations to cells; 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. In contrast, the mixed set of single guides for CRISPR perturbations can be applied in parallel, allowing

these experiments to be scaled up massively. While 32 datasets measure scRNA-seq exclusively, we also include scATAC from three papers, including one with simultaneous protein measurements (Mimitou et al., 2019). 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 (Fig 2B). Total unique molecular identifier (UMI) counts per cell and number of genes per cell are calculated as described in (Luecken 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 UMI counts measured even for lowly expressed 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 number of cells per perturbation in a dataset (Fig 2C). 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.

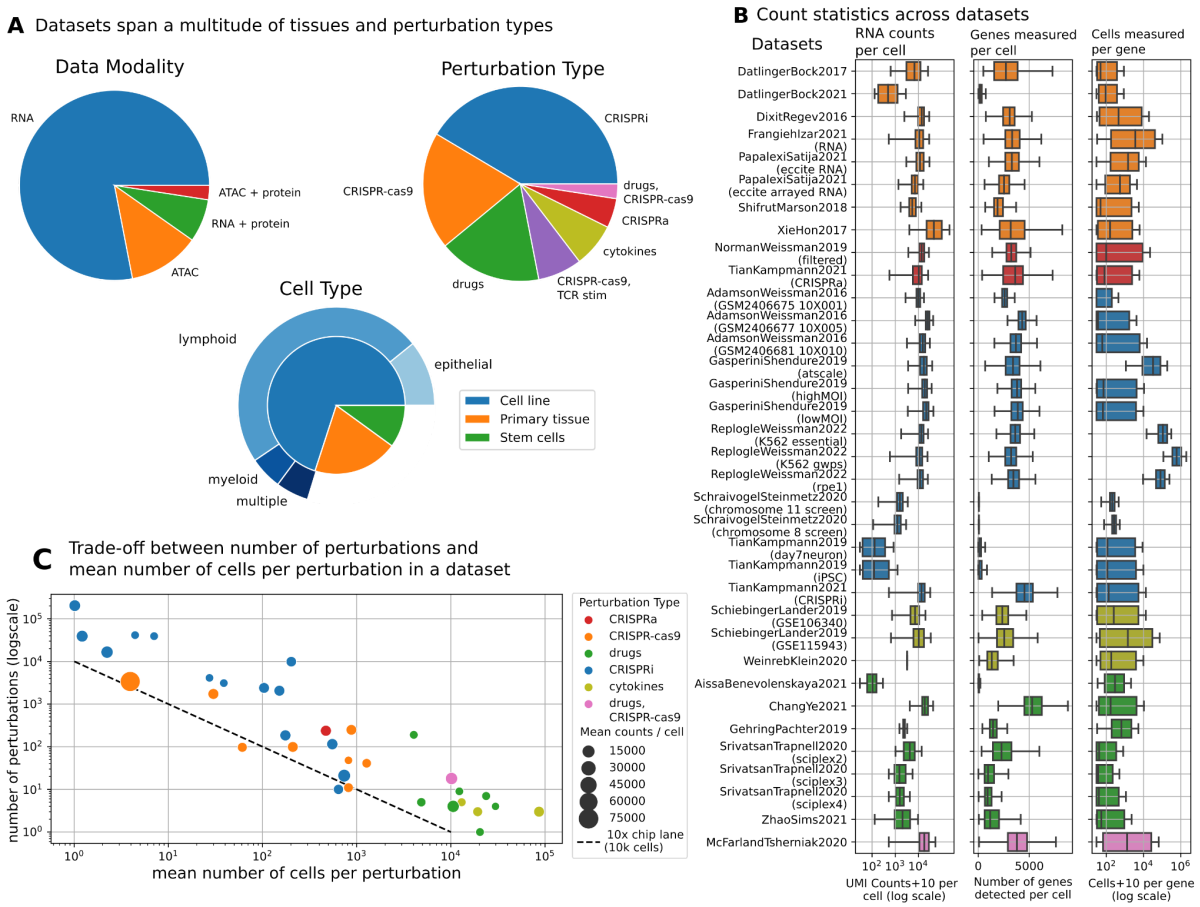


Fig 2: Single cell perturbation-response datasets are diverse in type, size, and quality. (A) The majority of included datasets result from CRISPR (DNA cut, inhibition or activation) perturbations using cell lines derived from various cancers. The studies performed on cells from primary tissues generally 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 Figure 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. 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.

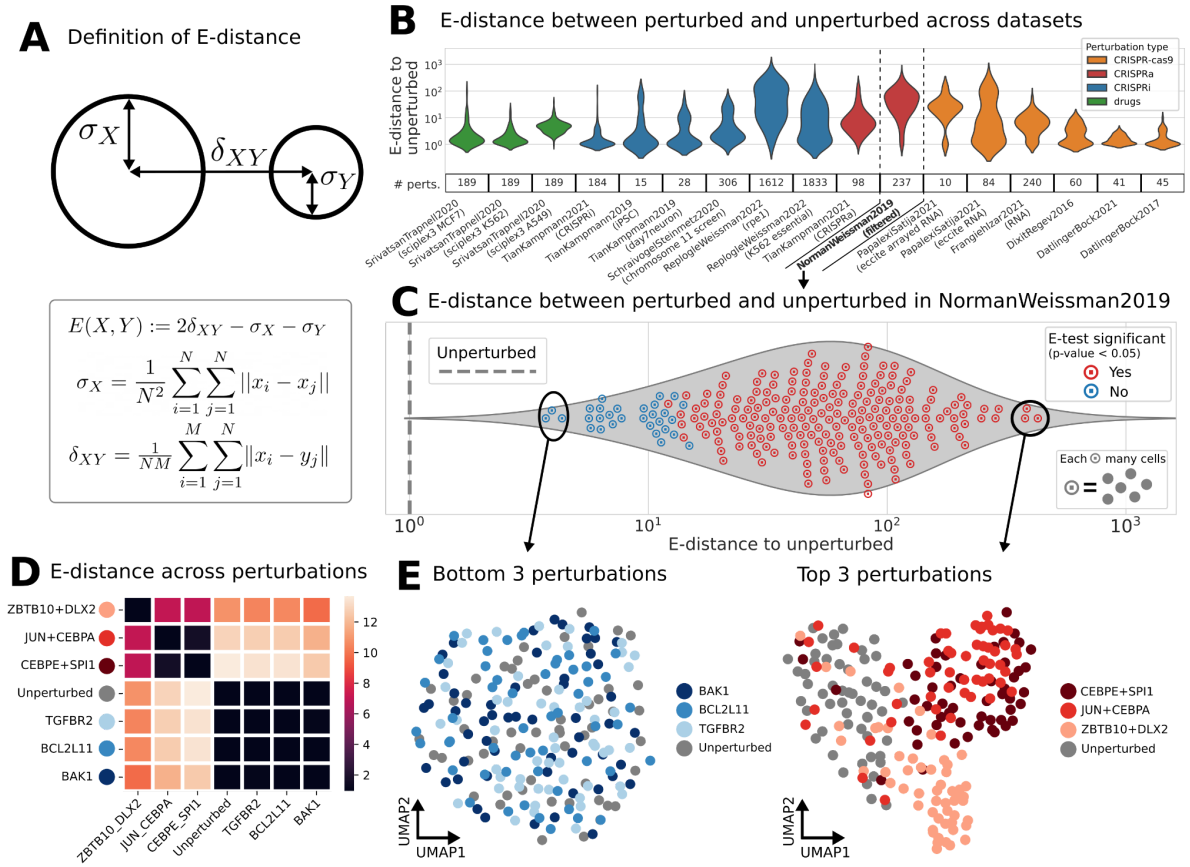


Fig 3: E-statistics describe distinctiveness of perturbations in single-cell data. (A) Definition of E-distance, relating the width of cell distributions of high-dimensional molecular profiles to their distance from each other (see Methods). A large E-distance of perturbed cells from unperturbed indicates a strong change in molecular profile induced by the perturbation. (B) Distribution of E-distances (plus 1 for log scale, same in Fig 3C) between perturbed and unperturbed cells across datasets. The number of perturbations per dataset is displayed along the bottom. Note that this plot is best used to compare the shape of the E-distance distribution rather than the magnitude; the mean E-distance will vary significantly with other dataset properties. (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 Figure 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 Figure 3C and the unperturbed cells. (E) UMAP of single cells of the weakest (left, Bottom 3) and strongest (right, Top 3) perturbations.

[E-distance: definition]

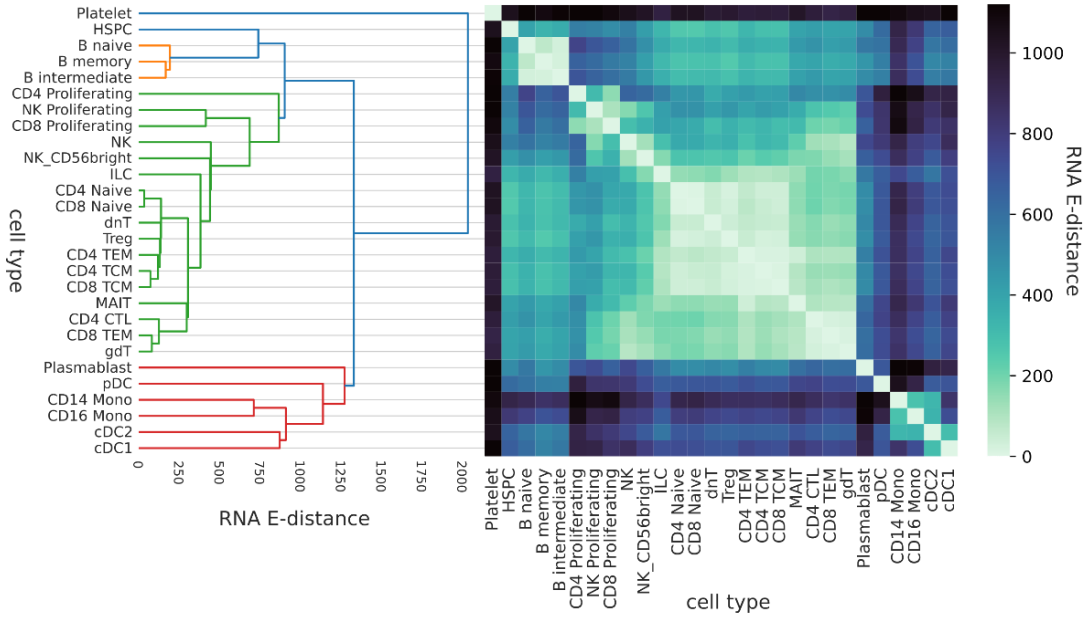
To compare and evaluate perturbations within each dataset we utilized the E-distance, a statistical distance measure between two distributions, which was used as a test statistic in (Replogle et al., 2022). Essentially, the E-distance compares the mean pairwise distance of cells across two different perturbations to the mean pairwise distance of cells within the two distributions (see Methods). If the former 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 (see Methods).

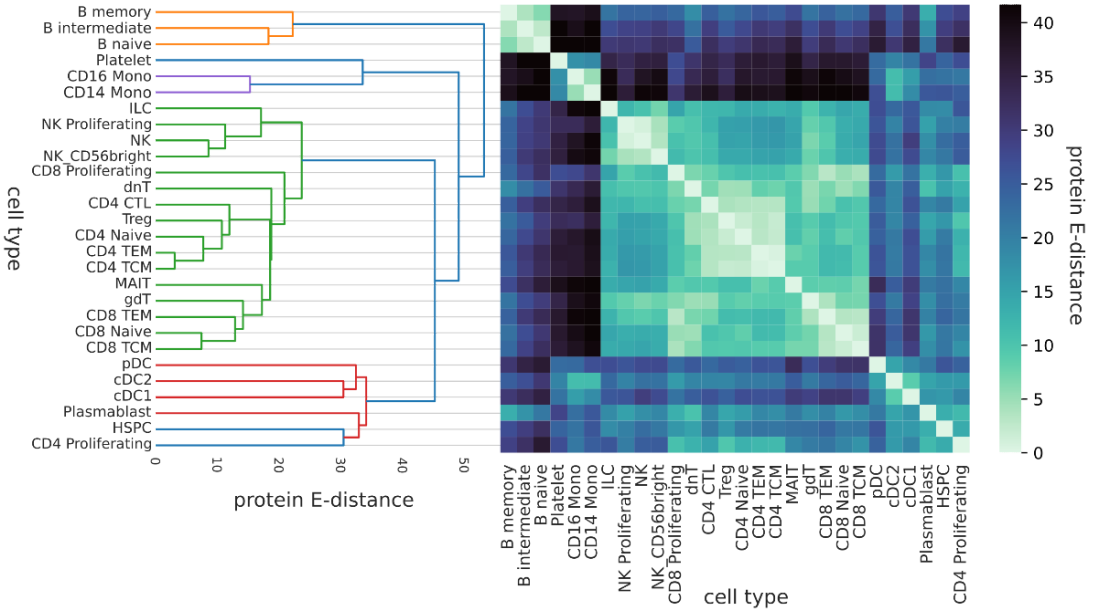
[E-distance: interpretation]

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 large shift in expression profiles, reflecting either technical problems in the experiment, ineffectiveness of the perturbation, or perturbation resistance.

A E-distance hierarchy of cell types in PBMC data based on RNA



B E-distance hierarchy of cell types in PBMC data based on protein



C UMAP with cell types from Hao et al.

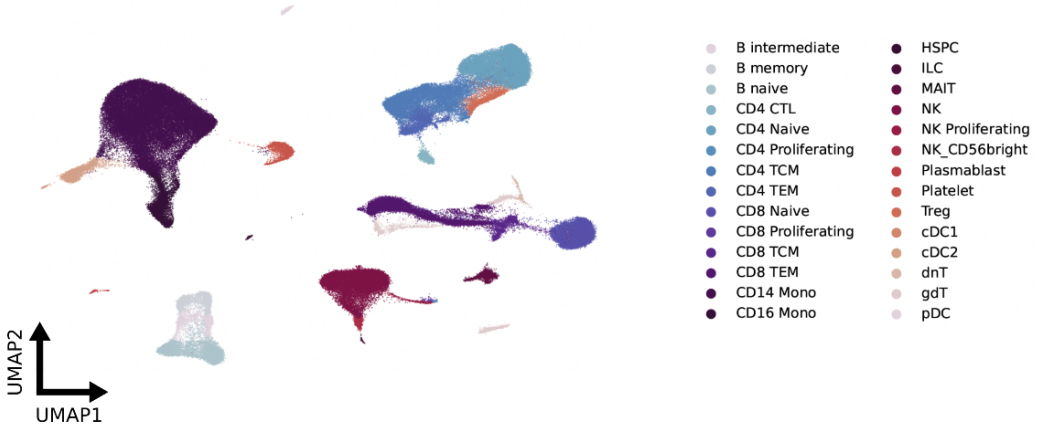


Fig 4: Cell type hierarchies computed using E-distance match known cell type relationships. (A) Hierarchical clustering of pairwise E-distances computed using RNA matches prior knowledge of transcriptome-defined cell types. Dendrogram and heatmap use the same distances. Data from (Hao et al., 2021). (B) As in (A) but using antibody-tagged surface proteins instead of RNA. (C) Visualization of cell type relationships in full multimodal dataset after batch correction. Coordinates and cell type annotations from (Hao et al., 2021).

[E-distance: distinguishing cell types]

To test whether E-distance values replicate differences between well-known cell types, we applied the E-distance to a CITE-seq human PBMC (peripheral blood mononuclear cells) dataset with existing cell type annotations (Hao et al., 2021). Separately for RNA and protein (from antibody-derived tags), we computed PCA-based E-distances between all pairs of cell types, equivalent to how perturbation E-distance is computed. The resulting pairwise distance matrices were used to compute cell type hierarchies (see Methods), which we compare to known cell type relationships (Fig 4A, 4B, Diehl et al., 2016). In both data modalities, B cell subtypes are clustered together, and platelets are the most distinct from any other cell type. Lymphoid and myeloid cells form two separate groups in the E-distance hierarchy. Notably, innate lymphoid cells (ILCs) and NK cell clusters form a distinct group as well. ILCs are innate immune cells that functionally correspond to specific types of classical lymphocytes properly expressing diversified antigen receptors; NK cells are a type of ILC also known as ILC4 and are functionally similar to cytotoxic T cells (Artis and Spits, 2015; Vivier et al., 2018). This functional similarity translates to strong similarities in transcriptional profiles, which often leads to difficulties in distinguishing NK cells and cytotoxic T cells in scRNA-seq data. These cell types are more easily disentangled by protein marker based distances, exemplifying the usefulness of CITE-seq as a method for identifying immune cell types. Likewise, when using protein, T cells are clustered primarily by CD4/CD8 type, whereas, when using RNA, they are clustered by functional phenotype (naive, proliferating, memory). For instance, clustering the cells with the E-distance in RNA-space separates proliferating cells of many types into a single cluster, likely due to shared expression of cell-cycle related genes; the cell cycle is known to have a strong effect on the transcriptome profile of cells and is not captured by surface protein measurements. We conclude that, by comparing RNA and protein representations, the protein modality more accurately represents cell type differences traditionally defined by immunologists on the basis of surface proteins, whereas the RNA representation primarily reflects functional programs of the cells such as cytotoxicity or proliferation. In both cases, the E-distance accurately captures known characteristics of each measurement modality.

[E-distance: E-test]

The E-distance can also be used as a test statistic to assess whether cells after a perturbation are significantly different from unperturbed cells (Replogle et al., 2022), Supp Table 3). The E-test is a permutation test that uses the E-distance as a test

statistic (Székely and Rizzo, 2013, details in Methods). This permutation test requires hundreds of iterations of computing the E-distance on randomized data, but is necessary for direct comparisons of perturbations across studies, as the exact value of the E-distance depends on dataset-specific parameters. The lack of robustness is exemplified by Figure 5A, which indicates that reducing the number of cells actually increases the E-distance, while the E-test gradually loses significance. Thus, while the E-distance is a useful tool for analysis within one dataset or between experimentally similar datasets, we recommend the E-test as the appropriate statistical measure for comparisons of perturbation effects between different datasets.

[E-distance: perturbation dataset example]

Interestingly, we found that E-distances between perturbed and unperturbed cells vary significantly 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, expression profiles of most perturbations in this dataset were significantly different from those of unperturbed cells according to the E-test (Fig 3C). Plausibly, this is in part caused by two-target perturbations using CRISPRa in that dataset: targeting the same gene with two single guides increases the chances of causing a considerable change in the transcript profile. Indeed, the three perturbations with highest E-distance are double perturbations while the three closest in E-distance are not. The corresponding UMAPs for these perturbations, computed using the same PCs as the E-distance, provide a confirmatory visual intuition for high and low E-distances (Fig 3E). The top three perturbations causing the largest E-distance to unperturbed are easily distinguishable from the gray unperturbed cells, while the bottom three weakest perturbations are part of a single, uniform cloud virtually indistinguishable from the unperturbed cells. The smallest E-distance thus results from perturbations which have the least effect on the distribution of cells.

[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 low 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.

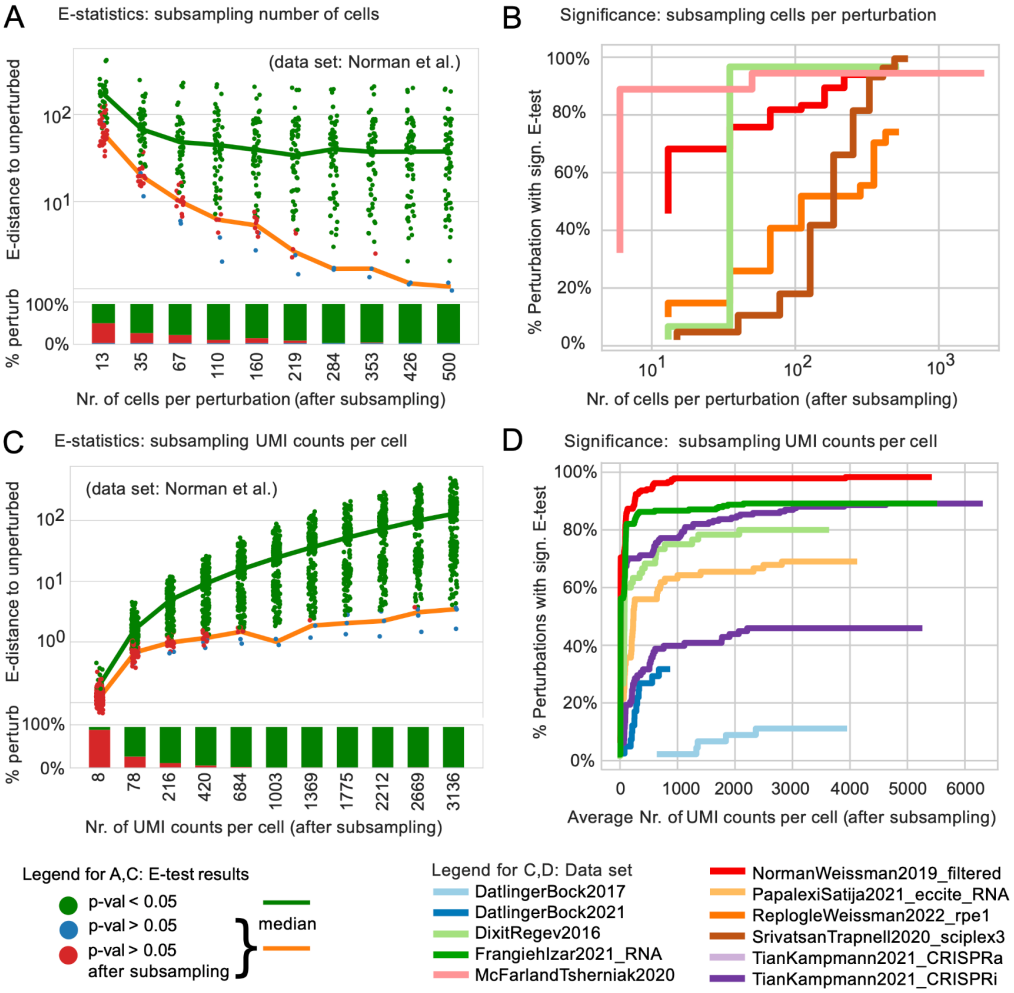


Fig 5: Effect of subsampling UMI counts per cell and number of cells per perturbation on E-statistics. (A) E-distance of each perturbation to unperturbed in Norman et al. while subsampling the number of cells per perturbation; Color indicates E-test results; "significance lost": perturbation significant when all cells are considered, but not significant after subsampling. The E-test loses significance with lower cell numbers while the E-distance actually increases. (B) Overall number of perturbations with significant E-test decreases when subsampling cells. (C) As in Figure 5A but subsampling UMI counts per cell while keeping the number of cells constant. Loss of E-test significance and dropping E-distance to unperturbed as overall signal gets deteriorated with removal of UMI counts. (D) As in Figure 5B but subsampling UMI counts per cell while keeping the number of cells constant.

[E-distance: Effect of number of cells on the E-statistics]

We investigate the robustness of E-distance and E-test scores to experimental and computational parameters using our extensive collection of harmonized single-cell perturbation datasets. We subsampled the number of cells per perturbation to create artificially smaller datasets, then examined how the E-distance and E-test results change. We find that the E-distance actually increases as the number of cells per perturbation decreases, indicating that cells per perturbation should be standardized prior to calculating E-distances (Fig 5A). This increase reflects the fact that the

E-distance, as a V-statistic, is a biased estimator (Székely and Rizzo, 2013). Despite the increase in E-distance with falling cell numbers, the number of significant perturbations correctly decreases with fewer cell counts, and only some datasets have saturated significance at full number of cells in that dataset (Fig 5B). This saturation point will depend on the strength of the perturbation and on the heterogeneity of the dataset; if all cells are similar to each other, a small set of cells will sufficiently describe every possible response to a perturbation. This suggests that, unsurprisingly, increasing sample size enables discovery of significant perturbations with smaller magnitude.

[E-distance: Effect of number of UMI counts on the E-statistics]

Similarly, we subset the number of UMI (unique molecular identifier) counts per cell, finding that E-distance increases as the number of UMI counts per cell increases (Fig 5C). The number of significant perturbations under the E-test, though, saturates around 500 counts per cell, with most perturbations that were significant at the full measured read depth maintaining that significance even with far fewer counts per cell (Fig 5D). The stability of E-test results with respect to UMI counts, in contrast to the actual E-distance value, exemplifies the necessity of the E-test as the appropriate statistical measure to evaluate perturbation effects. The optimal UMI and cell counts for a given experiment depend on downstream specific modeling tasks, as discussed in more detail elsewhere (Gross and Blüthgen, 2020). As a baseline for significant perturbations, as defined by the E-test, we suggest at least 300 cells per perturbation (Fig 5B) and 1000 average UMI counts per cell (Fig 5D) as an experimental guideline for distinguishable perturbations.

[E-distance: robust to calculation choices]

We also examined how choices made in computing the principal components (PCs) used in distance calculations affects E-statistics. The number of PCs used from PCA to compute the E-distance had a moderate effect on E-test results, mildly decreasing the number of significant perturbations (Supp Fig 5A). Interestingly, E-test significance was lost most rapidly in a TAP-seq (targeted perturb-seq) experiment (Schraivogel et al., 2020). TAP-seq only measures approximately 3000 genes of interest, and thus has far fewer starting features than other datasets. This leads to reduced correlation between genes in the resulting expression matrix, and thus fewer PCs are needed to sufficiently describe the data. The number of highly variable genes (HVGs) used to compute the PCs had almost no effect on E-testing above 500 HVGs (Supp Fig 5B). Computing PCs separately for each perturbation rather than jointly across all perturbations in a given dataset similarly had minimal impact on the resulting E-distances (Supp Fig 5C). Taken together, this analysis indicates that E-statistics can be calculated as part of an existing computational workflow, which already includes calculating PCs.

[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. 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, either in one paper (Tian et al., 2021) or across multiple studies (Norman et al., 2019; Replogle et al., 2022). 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 all available for K562 cells (Frangieh et al., 2021; Pierce et al., 2021; Replogle et al., 2022).

Discussion

[Concise summary of results]

We present a dataset resource and 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 uniform annotations in this resource will enable data integration and benchmarking as well as exploration of shared perturbations across datasets. The use of the E-distance is motivated and applied to quantitatively compare perturbations within each dataset. We illustrate how to interpret high and low E-distances and use E-distances to identify functionally similar versus distinct perturbations. We also investigate the effect of dataset specific parameters on E-statistics, showing that E-statistics stabilize above 1000 counts per cell and 300 cells per perturbation.

[Overlap of perturbations across studies]

While this work simplifies access to datasets, 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 are 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. With more datasets anticipated in the future, we will have the unique opportunity to integrate datasets with more overlapping perturbations and nominate machine learning benchmarks for data integration.

[Towards standardization]

Lack of standardization in data sharing and processing hampered the creation of this resource. Although many processed datasets were available on the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013), there is no standard format for sharing CRISPR barcode assignments and other metadata. Starting analysis from sequencing reads may have improved interoperability of datasets in this resource, but guide assignment procedures and demultiplexing algorithms are specific to experimental setup. For scATAC data, data comparison is made more challenging by the lack of a standard method for feature assignment (see Methods). In particular, scATAC feature assignments specific to CRISPR perturbations, where known locus-of-action could be used to improve feature calls (Chen et al., 2019). In all modalities, many datasets only supplied processed data, or raw data was only available after institutional clearance. Adding more datasets to this resource, or the creation of similar resources in the future, would be much easier if there were standard formats for sharing perturbation data, and, more generally, standard formats for sharing single-cell annotations. We think a community-wide discussion on standardization of such data is urgently needed, as has been done for proteomic data (Gatto et al., 2022).

[Additional 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). 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 of the cells may be advisable to increase the signal to noise ratio without sequencing more cells. For perturbation distinguishability as defined by the E-test, regardless of experimental parameters, we find that one should have at least 300 cells per perturbation and an average of 1000 UMIs per cell.

[Conclusion]

We envision that the scPerturb collection of datasets and the suggested E-statistics analytic framework will be a valuable starting point for the analysis of single-cell perturbation data. The unified annotations and perturbation significance testing across datasets should prove especially useful to the machine learning community

for training models on this data. We expect new datasets and new experimental perturbation methods in the future will enable the community to develop novel computational approaches which 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
- RNA data at https://zenodo.org/record/7041849 and ATAC data at https://zenodo.org/record/7058382

Code Availability

Open access source code is at https://github.com/sanderlab/scPerturb/. We compiled a corresponding Python package called `scperturb` for performing E-statistics (E-distance and E-testing) in single-cell data, published on PyPI under https://pypi.org/project/scperturb/.

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.

Note that these 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).

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 relevant 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). 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 (Fig 4), cells and genes were filtered using Scanpy (versions 1.7.2–1.9.1) (Wolf et al., 2018). All .h5ad objects published on the resource were saved using Scanpy 1.9.1. 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 `scanpy.pp.find_variable_genes` with flavor 'seurat_v3'. We normalized the count matrix using `scanpy.pp.normalize_total` and log-transformed the data using `scanpy.pp.log1p`; We did not z-scale the data. Next, we computed PCA based on the highly variable 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. Large parts of our analysis were parallelized as workflows using snakemake (Mölder et al., 2021).

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

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

$$\sigma_X = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} ||x_i - x_j||$$

and σ_Y defined accordingly. We used the squared euclidean distance when calculating cell-wise distances. 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. 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 Table 3). We corrected for multiple testing using the Holm-Sidak method per dataset.

Cell type hierarchy

Cell type annotations celltype.l2 were used as provided by (Hao et al., 2021). Doublets were removed and data was subset to 91 cells per remaining cell type, which is the largest number such that all key cell types had at least that many cells. After subsetting, the data was processed as for other RNA datasets. Protein data was CLR normalized using Muon 0.1.2 and log-transformed prior to PCA (Bredikhin et al., 2022). The hierarchy was computed using `scipy.cluster.hierarchy.linkage` from scipy 1.8.0 with method “single”.

Subsampling analysis

At each subsampling point we computed detailed E-statistics (E-distances, delta, sigma, E-test results) from each perturbation to the corresponding unperturbed cells of that dataset using PCA with 50 components based on 2000 highly variable genes, except specified otherwise. We downsampled raw UMI counts using the function `scanpy.pp.downsample_counts` on raw counts, then preprocessed (normalized, log1p-transformed, etc.) the data as previously described. Cells were downsampled to the same number at each subsampling step across all perturbations to avoid comparability issues. If possible, we recalculated the PCA while keeping the highly variable genes originally obtained from the complete dataset. Figures 5C, 5D and Supplemental Figures 5A, 5B were computed as a running loss of E-test significance

(p-value<0.05) of formerly – i.e. prior to any subsampling – significant perturbations while subsampling, then normalized across datasets through division by the total number of formally significant perturbations in that datasets.

Advice for single-cell perturbation analysis

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. Additionally, the .h5ad datasets shared in this resource can be programmatically accessed using .h5py, and perturbations of interest extracted without requiring full dataset access.

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, Beyond the Exome)
- Einstein Stiftung Berlin (Einstein visiting fellow program)
- Computation was in part performed on the HPC for Research cluster of the Berlin Institute of Health.
- We appreciate informative conversations with Yuge Ji, helpful code suggestions from Garrett Wong, and computational support from Aaron Kollasch. We also appreciate preprint review comment from Arcadia Science's preprint review initiative (Gregory P. Way, Natalie Davidson, Erik Serrano, Parker Hicks, Jenna Tomkinson, Dave Bunten).

References

https://raw.githubusercontent.com/caleblareau/asap_reproducibility/master/CD4_CRISPR_asapseq/output/Signac/after_filter_Signac/HTO_res_filtered.txt

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., Pak, R.A., Gray, A.N., Gross, C.A., Dixit, A., Parnas, O., Regev, A., Weissman, J.S., 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., Feldman, L.E., Green, S.J., Lopez-Bigas, N., Frolov, M.V., Benevolenskaya, E.V., 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

Artis, D., Spits, H., 2015. The biology of innate lymphoid cells. Nature 517, 293–301. https://doi.org/10.1038/nature14189

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

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41, D991–D995. https://doi.org/10.1093/nar/gks1193

Bertin, P., Rector-Brooks, J., Sharma, D., Gaudelet, T., Anighoro, A., Gross, T., Martinez-Pena, F., Tang, E.L., S, S.M., Regep, C., Hayter, J., Korablyov, M., Valiante, N., van der Sloot, A., Tyers, M., Roberts, C., Bronstein, M.M., Lairson, L.L., Taylor-King, J.P., Bengio, Y., 2022. RECOVER: sequential model optimization platform for combination drug repurposing identifies novel synergistic compounds in vitro. https://doi.org/10.48550/arXiv.2202.04202

Bredikhin, D., Kats, I., Stegle, O., 2022. MUON: multimodal omics analysis framework. Genome Biol. 23, 42. https://doi.org/10.1186/s13059-021-02577-8

Broad Institute, 2022. Single Cell Portal [WWW Document]. URL https://singlecell.broadinstitute.org/single_cell (accessed 8.17.22).

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., 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., 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., 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., Haley, B., Evangelista, M., Malek, S., Foster, S.A., Ye, X., 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., 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., 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., 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., 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., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., 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
Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., Bock, C., 2017. Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods 14, 297–301. https://doi.org/10.1038/nmeth.4177
Dhapola, P., Rodhe, J., Olofzon, R., Bonald, T., Erlandsson, E., Soneji, S., 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
Diehl, A.D., Meehan, T.F., Bradford, Y.M., Brush, M.H., Dahdul, W.M., Dougall, D.S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C.E., Vasilevsky, N.A., Haendel, M.A., Blake, J.A., Mungall, C.J., 2016. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J. Biomed. Semant. 7, 44. https://doi.org/10.1186/s13326-016-0088-7
Dixit, A., Parnas, O., Li, B., 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., Zhang, Z., Wang, P., Sun, S., Liu, Q., 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., 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., Bicciato, S., 2021. Computational methods for the integrative analysis of single-cell data. Brief. Bioinform. 22. 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., Zhao, M., Ager, C.R., Rogava, M., Hovey, L., Rotem, A., Bernatchez, C., Wucherpfennig, K.W., Johnson, B.E., Rozenblatt-Rosen, O., Schadendorf, D., Regev, A., Izar, B., 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., 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., Trapnell, C., Ahituv, N., Shendure, J., 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

Gatto, L., Aebersold, R., Cox, J., Demichev, V., Derks, J., Emmott, E., Franks, A.M., Ivanov, A.R., Kelly, R.T., Khoury, L., Leduc, A., MacCoss, M.J., Nemes, P., Perlman, D.H., Petelski, A.A., Rose, C.M., Schoof, E.M., Van Eyk, J., Vanderaa, C., Yates III, J.R., Slavov, N., 2022. Initial recommendations for performing, benchmarking, and reporting single-cell proteomics experiments. https://doi.org/10.48550/arXiv.2207.10815

Gehring, J., Hwee Park, J., Chen, S., Thomson, M., 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., Qi, L.S., Kampmann, M., Weissman, J.S., 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., 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., 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., Blüthgen, N., 2019. Robust network inference using response logic. Bioinformatics 35, i634–i642. https://doi.org/10.1093/bioinformatics/btz326

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Haque, A., Engel, J., Teichmann, S.A., 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, 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., 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., Aronow, B.J., 2022. CellDrift: Inferring Perturbation Responses in Temporally-Sampled Single Cell Data. 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., Liu, Q., Russkikh, N., Ryazantsev, G., Ohler, U., Participants, N. 2021 M. data integration competition, Pisco, A.O., Bloom, J., Krishnaswamy, S., Theis, F.J., 2022. Multimodal single cell data integration challenge: results and lessons learned. https://doi.org/10.1101/2022.04.11.487796
Lareau, Caleb.A., 2021. asap_reproducibility.
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., Jaini, S., Smibert, P., Sanjana, N.E., 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., Naghipourfar, M., Theis, F.J., 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
Lotfollahi, M., Wolf, F.A., Theis, F.J., 2019. scGen predicts single-cell perturbation responses. Nat. Methods 16, 715–721. https://doi.org/10.1038/s41592-019-0494-8
Love, M.I., Huber, W., 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., Theis, F.J., 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, M.D., 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., Dionne, D., Bender, S., Wolpin, B.M., Ghandi, M., Tirosh, I., Rozenblatt-Rosen, O., Roth, J.A., Golub, T.R., Regev, A., Aguirre, A.J., Vazquez, F., Tsherniak, A., 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., Satija, R., Sanjana, N.E., Koralov, S.B., Smibert, P., 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., Thakore, P.I., Kibayashi, T., Wing, J.B., Hata, M., Satija, R., Nazor, K.L., Sakaguchi, S., Ludwig, L.S., Sankaran, V.G., Regev, A., Smibert, P., 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
Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. https://doi.org/10.12688/f1000research.29032.2
Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., 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., Satija, R., 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., 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., 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., 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., 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., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J.L., Jost, M., Norman, T.M., Weissman, J.S., 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., 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., Zarnegar, B.J., Greenleaf, W.J., Chang, H.Y., Khavari, P.A., 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., Shah, P., Bell, J.C., Jhutty, D., Nemec, C.M., Wang, J., Wang, L., Yin, Y., Giresi, P.G., Chang, A.L.S., Zheng, G.X.Y., Greenleaf, W.J., Chang, H.Y., 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., 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., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., Lander, E.S., 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., 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., 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., Zhang, F., Steemers, F., Shendure, J., Trapnell, C., 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., Clarke, D.J.B., Lachmann, A., Evangelista, J.E., Ma'ayan, A., Medvedovic, M., Schürer, S.C., 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., Pachter, L., 2020. A curated database reveals trends in single-cell transcriptomics. Database 2020. https://doi.org/10.1093/database/baaa073
Székely, G.J., Rizzo, M.L., 2013. Energy statistics: A class of statistics based on distances. J. Stat. Plan. Inference 143, 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018
Tian, R., Abarientos, A., Hong, J., Hashemi, S.H., Yan, R., Dräger, N., Leng, K., Nalls, M.A., Singleton, A.B., Xu, K., Faghri, F., Kampmann, M., 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
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., Ward, M.E.,

Kampmann, M., 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

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., Meyers, R.M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W.F.J., Howell, S., Merkel, E., Ghandi, M., Garraway, L.A., Root, D.E., Golub, T.R., Boehm, J.S., Hahn, W.C., 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., Rynes, E., Reynolds, A., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Kaul, R., Meuleman, W., Stamatoyannopoulos, J.A., 2020. Global reference mapping of human transcription factor footprints. Nature 583, 729–736. https://doi.org/10.1038/s41586-020-2528-x

Vivier, E., Artis, D., Colonna, M., Diefenbach, A., Santo, J.P.D., Eberl, G., Koyasu, S., Locksley, R.M., McKenzie, A.N.J., Mebius, R.E., Powrie, F., Spits, H., 2018. Innate Lymphoid Cells: 10 Years On. Cell 174, 1054–1066. https://doi.org/10.1016/j.cell.2018.07.017

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., Klein, A.M., 2020. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science 367. https://doi.org/10.1126/science.aaw3381

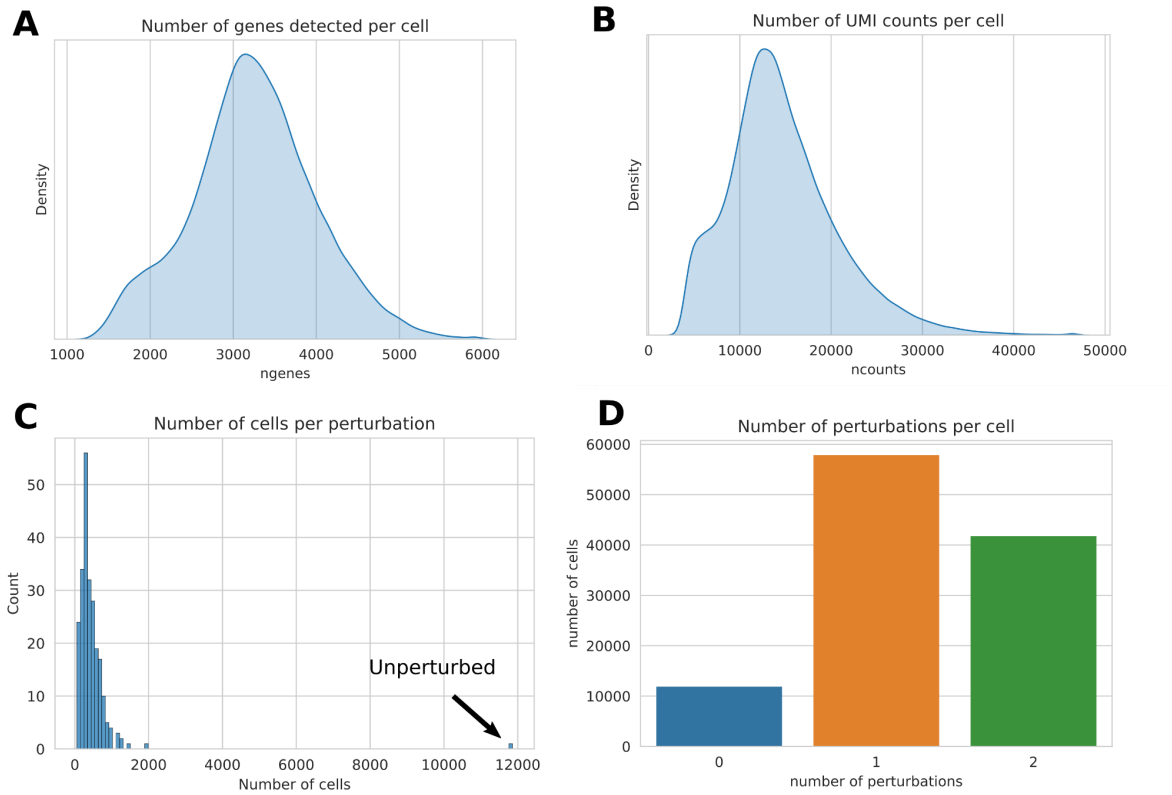
Wessels, H.-H., Méndez-Mancilla, A., Papalexi, E., Mauck, W.M., Lu, L., Morris, J.A., Mimitou, E., Smibert, P., Sanjana, N.E., Satija, R., 2022. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq (preprint). Genomics. https://doi.org/10.1101/2022.02.02.478894

Wolf, F.A., Angerer, P., 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

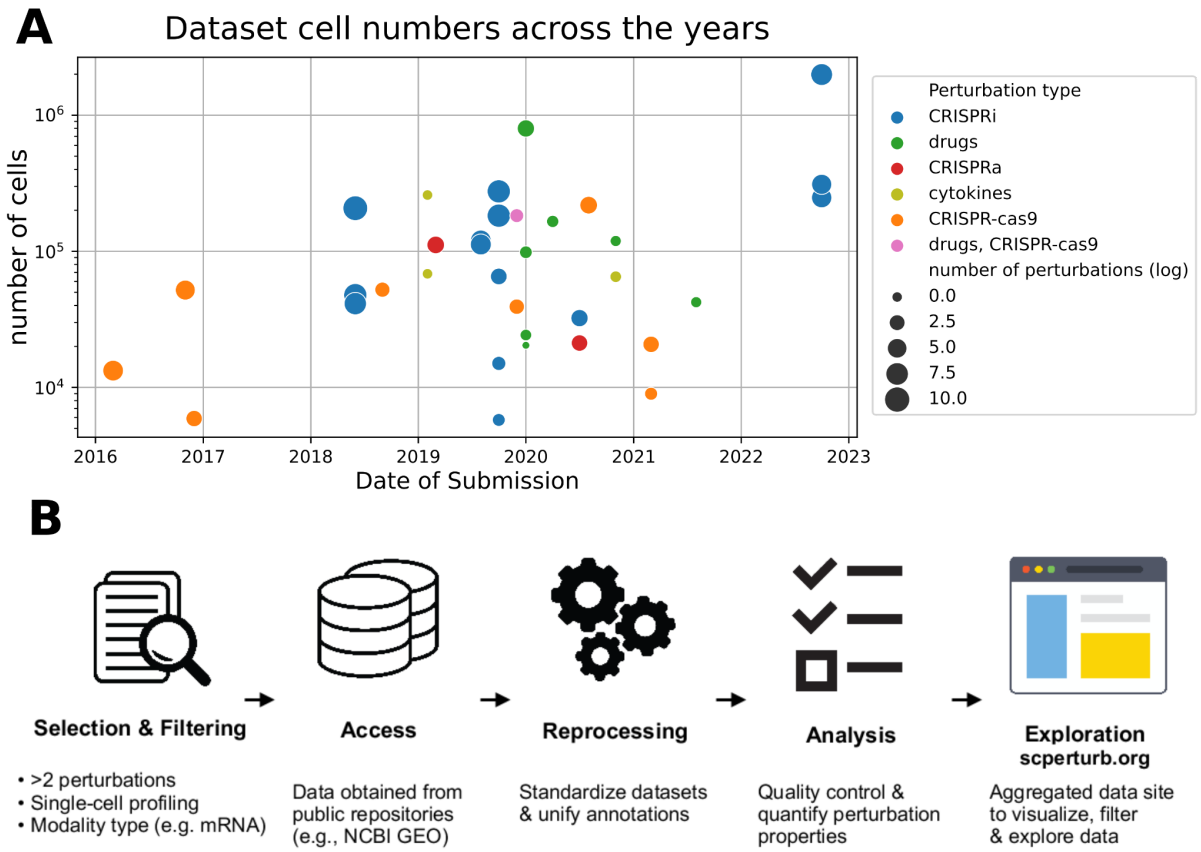
Xie, S., Duan, J., Li, B., Zhou, P., Hon, G.C., 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. Mol. Cell 66, 285-299.e5. https://doi.org/10.1016/j.molcel.2017.03.007

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., Bruce, J.N., Canoll, P., Sims, P.A., 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

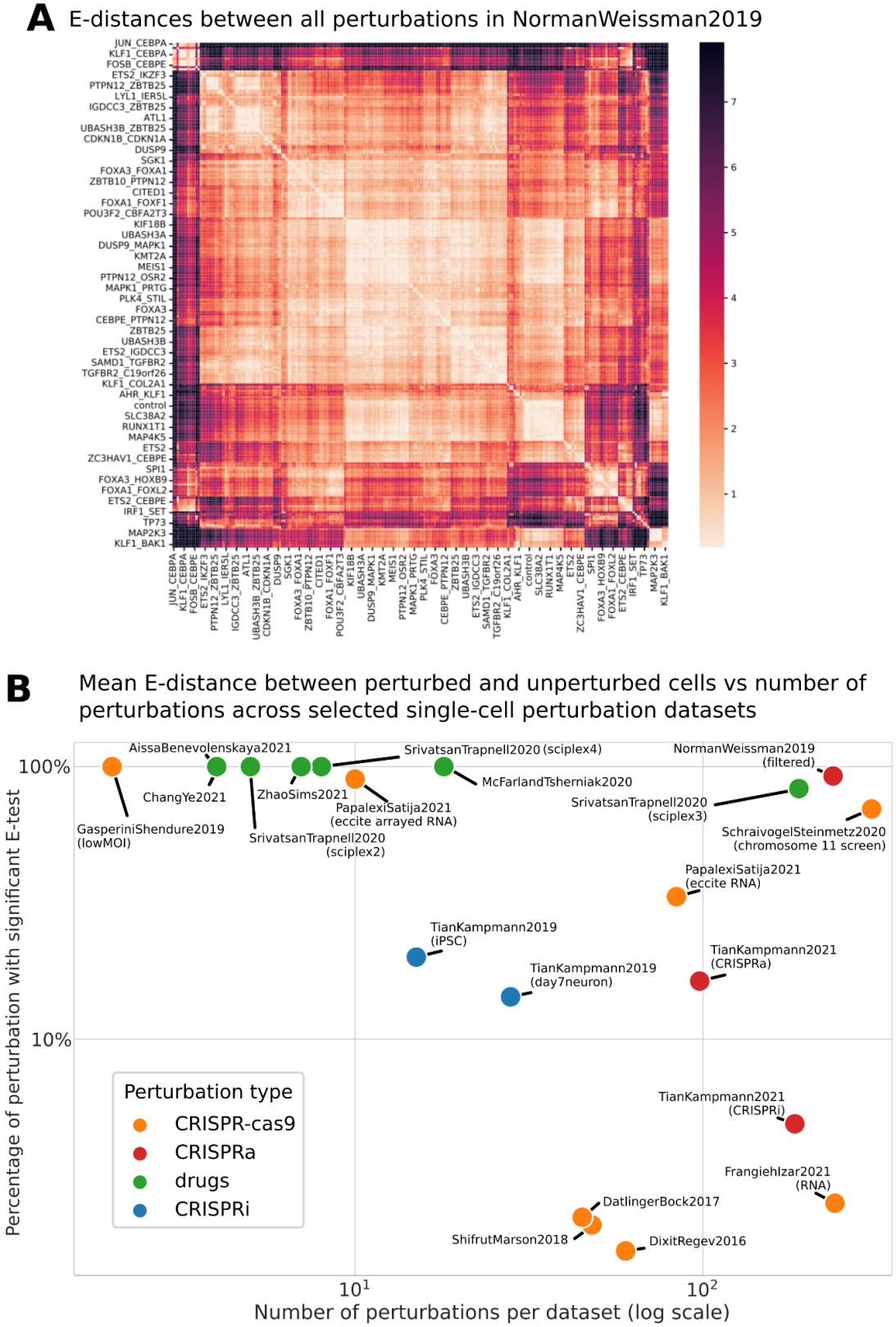


Supp Fig 1: Exemplary information provided for each scPerturb dataset (here for (Norman et al., 2019)**)** (A) Number of genes that are detected with at least one count in a cell across all cells. (B) Total number of UMI counts per cell across all cells. Together with Supp Fig 1A this provides an overview 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.

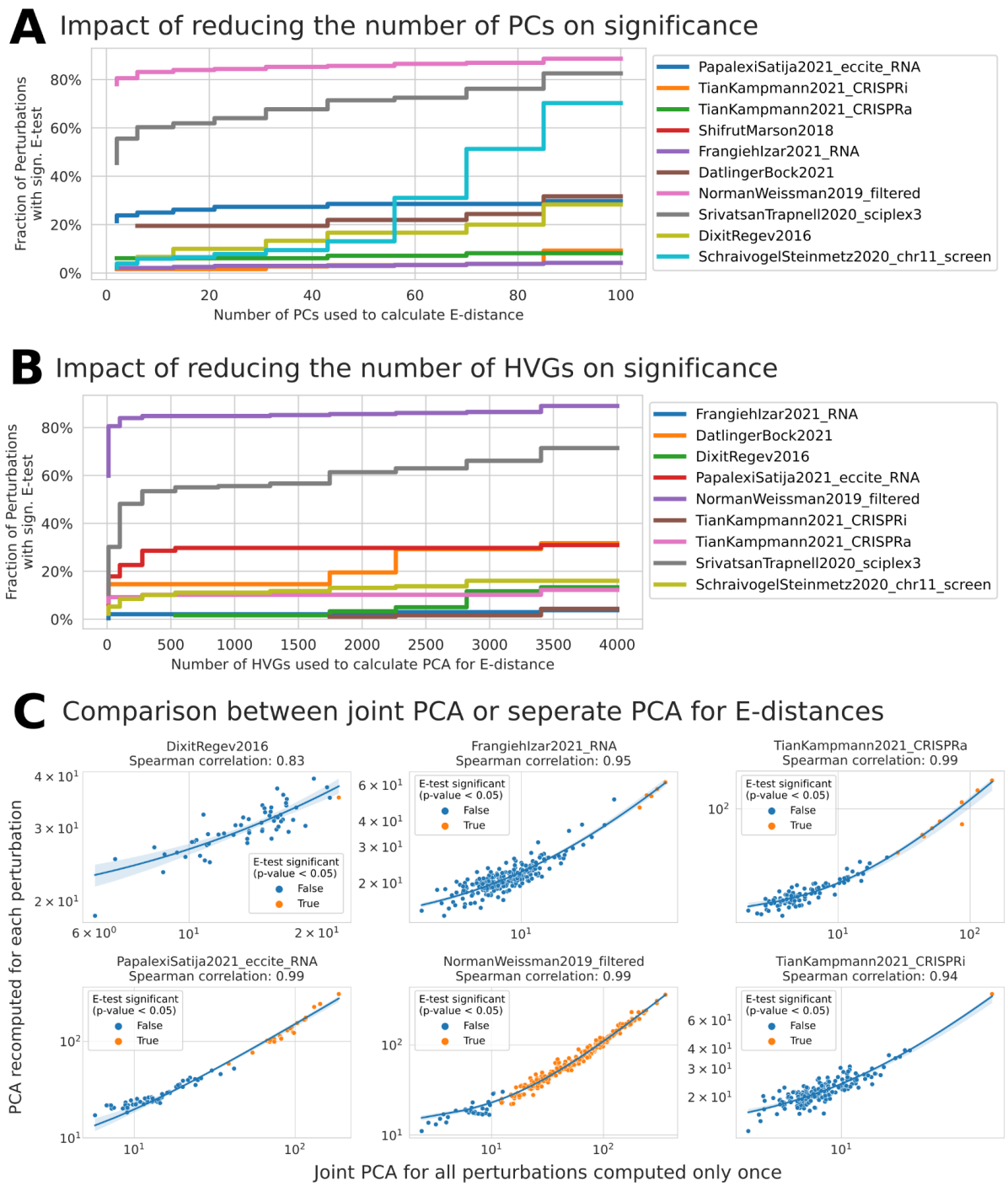


Supp Fig 2A: Number of cells per dataset by submission date. There is a rapid increase in published single-cell perturbation datasets around 2019. We speculate that the slight decrease of dataset numbers after 2021 suggested by the plot is due to the ongoing impact of reduced research in the earlier 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) Percentage of perturbations with significant E-test to unperturbed cells in each dataset plotted against the total number of perturbations in the dataset (both in log scale).



Supp Fig 5: Tests on robustness of E-statistics to dataset properties and parameters. (A) Effect of using different numbers of principal components (PCs) from PCA on the number of perturbations with significant E-test w.r.t. unperturbed cells. The SchraivogelSteinmetz2020 dataset is TAP-seq, thus has much less genes measured than all other datasets. The faster decrease in significance observed in this dataset indicates stronger sensitivity on the number of PCs with fewer features available. (B) Effect of using different numbers of highly variable genes (HVGs) for the PCA calculation prior to E-testing. For most datasets, E-test results appear to stay comparable between 500 and 4000 HVGs. (C) E-distance computed in a single, joint PCA compared to E-distance computed in a separate PCA per perturbed-unperturbed combination across three exemplary datasets. Consistently high Pearson correlations indicate strong equivalence between both approaches across datasets.