nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-019-0372-z

Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins

Jase Gehring[1,2], Jong Hwee Park[2], Sisi Chen[2], Matthew Thomson[2] and Lior Pachter[2,3*]

We describe a universal sample multiplexing method for single-cell RNA sequencing in which fixed cells are chemically labeled by attaching identifying DNA oligonucleotides to cellular proteins. Analysis of a 96-plex perturbation experiment revealed changes in cell population structure and transcriptional states that cannot be discerned from bulk measurements, establishing an efficient method for surveying cell populations from large experiments or clinical samples with the depth and resolution of single-cell RNA sequencing.

Massively parallel single-cell RNA sequencing (scRNA-seq) is transforming our view of complex tissues and yielding new insights into functional states of heterogeneous populations. Individual scRNA-seq experiments can probe the transcriptomes of more than 10,000 cells[1,2], and the first datasets approaching and exceeding one million cells have been reported[3,4]. This raises the prospect of performing screens involving hundreds, or even thousands, of samples for high-throughput analysis of genetic, signaling and drug perturbations. Here we present an approach to scRNA-seq multiplexing that allows for fixed cells from individual samples to be rapidly chemically labeled with identifying DNA oligonucleotides in a one-pot, two-step chemical cross-linking reaction (Fig. 1a). This overcomes the limitations imposed by device operation, high reagent cost and batch effects. Our method adds to a growing family of complementary multiplexing technologies and is independent of the specific epitopes central to CITE-seq[5], chromosomal polymorphisms exploited by demuxlet[6], lipid–oligonucleotide synthesis involved in MULTI-seq[7] or genetic manipulation performed in CellTag Indexing[8].

To label cells for multiplexed scRNA-seq, methyltetrazine-modified DNA oligonucleotides, or 'ClickTags', are attached to cellular proteins using inverse electron-demand Diels–Alder (IEDDA) chemistry and the heterobifunctional, amine-reactive cross-linker NHS-*trans*-cyclooctene (NHS-TCO). After demonstrating effective labeling conditions on yeast cells (Supplementary Fig. 1), we performed a multiplexed scRNA-seq experiment in which four samples of live mouse neural stem cells (NSCs) and four samples of methanol-fixed NSCs were each labeled with two unique ClickTags. The ClickTags can be specifically amplified and sequenced with a modified 10x Genomics single-cell gene expression protocol, and we developed a computational workflow, kallisto indexing and tag extraction (kITE), to rapidly pseudoalign ClickTag reads to an index of barcodes (Supplementary Fig. 2). ClickTag reads from methanol-fixed cells accurately recapitulated the experimental design with a high correlation between unique molecular identifier (UMI) counts for pairs of tags applied to the same sample, indicating efficient single-cell labeling and facilitating sample demultiplexing (Supplementary Figs. 3 and 4). Cell doublet events were unambiguously detected as collisions of two pairs of tags corresponding to two separate samples (Supplementary Fig. 5). Live cells treated under analogous conditions did not yield readily demultiplexed scRNA-seq libraries (Supplementary Fig. 3), likely owing to competitive NHS–ester hydrolysis in aqueous buffer.

We next performed a species-mixing experiment aimed at evaluating the limits of ClickTag multiplexing and quantifying any deleterious effects on the associated cDNA libraries. Samples of human HEK293T cells and mouse NSCs were fixed and reacted individually and in combination with a series of non-overlapping sample tag pools of increasing size (Supplementary Table 1). Processing these samples as a pooled lane of scRNA-seq, we found that overall cDNA library quality was consistent with untagged methanol-fixed samples (Supplementary Fig. 6 and Supplementary Table 2), including a slight under-representation of low-expression genes, a slight over-representation of high-expression genes and reduced library complexity as compared to live cells. Samples labeled with one or two ClickTags displayed highly reproducible gene expression profiles, as did samples labeled with three, four or five ClickTags, thus validating a 'balanced' labeling scheme using the same number of ClickTags of equal concentration for all samples in multiplexing experiments (Supplementary Figs. 7 and 8). Interestingly, an examination of the observed gene expression differences between cell populations labeled with one or two versus three, four or five ClickTags showed large changes in the non-coding RNAs *MALAT1* and *Xist* (Supplementary Figs. 9 and 10), two highly expressed genes captured via internal A-rich binding sites that are frequently filtered before analysis[9,10].

Filtering, clustering and embedding ClickTag data from this species-mixing experiment revealed eight distinct clusters of cells and high concordance with the experimental design (Methods; Supplementary Figs. 11 and 12). Up to five ClickTags could be deposited on a single cell without loss of tag recovery, implying that, in principle, 15,504 experiments could be multiplexed with a panel of just 20 tags. Species and ClickTag information were used to filter out cell doublets, with identification by manual subcluster selection on the basis of *t*-distributed stochastic neighbor embedding (*t*-SNE) outperforming CellRanger (cDNA-based, interspecies doublet events) and the Scrublet[11] algorithm (ClickTag-based, intersample doublet events) (Supplementary Fig. 12c–g), and achieving successful extraction of all 28 possible intersample collisions (Supplementary Fig. 13). Barnyard plots generated from the resulting population of singlets showed near-perfect species fidelity in single-species samples, indicative of highly accurate sample assignment despite differences in read depth per species (Supplementary Figs. 14 and 15).

Finally, we performed a perturbation experiment to illustrate the utility and scope of multiplexed scRNA-seq. NSCs are known

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. [3]Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA. *e-mail: lpachter@caltech.edu

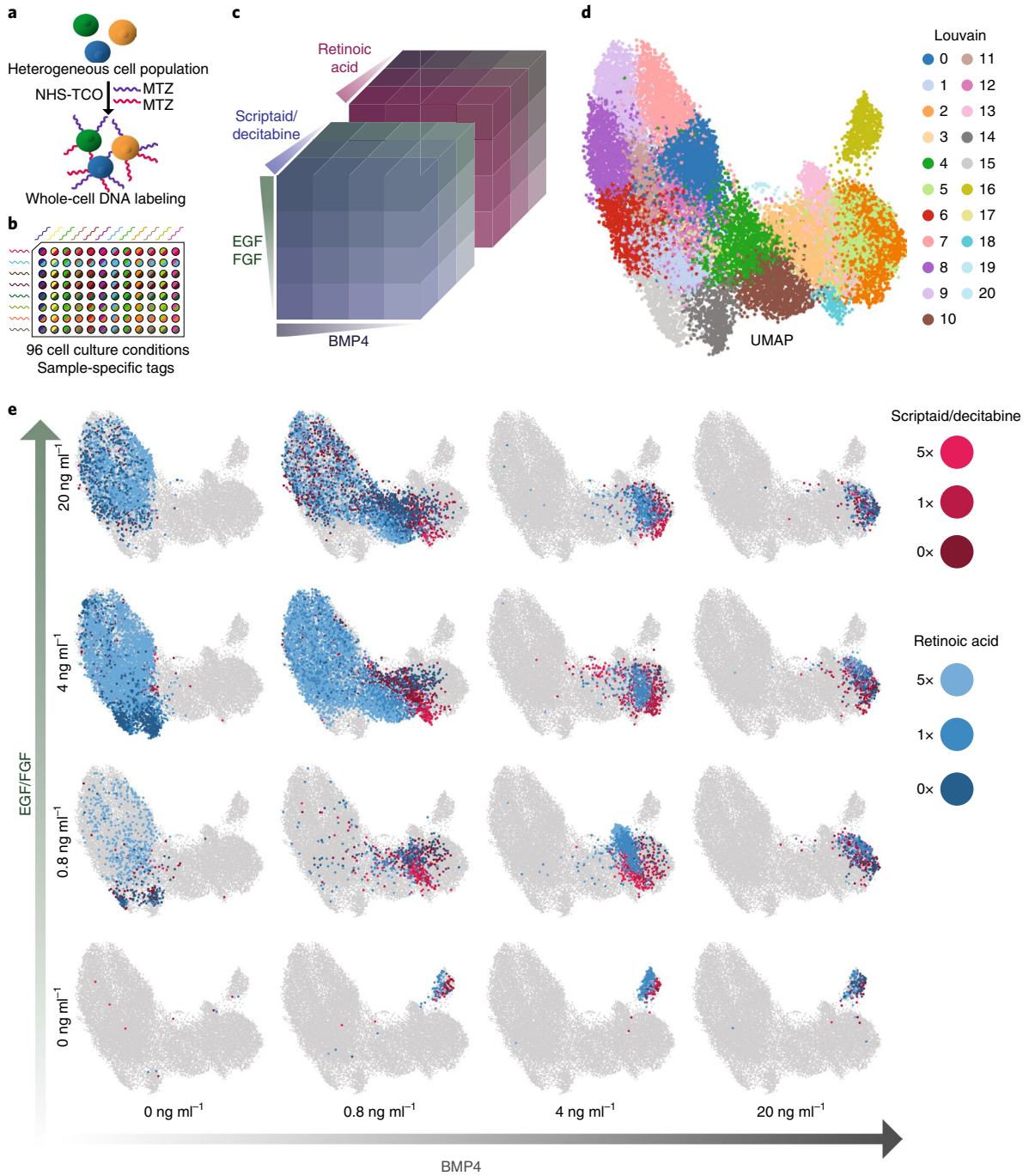

Fig. 1 | 96-plex scRNA-seq experiment. **a**, Four cellular perturbants (EGF and bFGF, BMP4, decitabine and Scriptaid, and retinoic acid) were titrated against one another to produce an array of 96 unique growth conditions. **b**, Before scRNA-seq, a one-pot, two-step reaction is performed in which cells that are already labeled with MTZ-DNA, NHS-TCO are labeled with sample-specific ClickTags. **c**, NSCs subjected to a 96-plex array of growth conditions were dual-labeled with a unique pair of ClickTags. **d**. UMAP embedding of $n = 21{,}191$ cells from the 96-plex perturbation. Cluster assignments parallel population behavior driven by experimental conditions. **e**, Visualization of the $n = 21{,}191$ cells in terms of the cell populations produced by each of the 96 experimental conditions. Each embedding corresponds to a given concentration of EGF and bFGF against a series of BMP4 concentrations and displays six samples colored by retinoic acid or decitabine and Scriptaid. Cell yields from each condition are provided in Supplementary Fig. 16c.

to differentiate into many unique cell types in vivo, primarily neurons, astrocytes and oligodendrocytes[12]. In vitro, NSCs can be forced into different differentiation trajectories by exposing the cells to a variety of synthetic chemicals, hormones and growth factors. We investigated the response of NSCs to varying concentrations of decitabine and Scriptaid, epidermal growth factor (EGF) and basic fibroblast growth factor (bFGF), retinoic acid, and bone morphogenic protein 4 (BMP4), producing a 4×4×6 perturbation array representing a large space of experimental conditions (Fig. 1c). NSCs were grown in a single 96-well plate with each culture corresponding to a unique combination of factors (Fig. 1b and Supplementary Fig. 16). After fixation and ClickTag labeling (Fig. 1a), the samples were pooled and subjected to our modified 10x protocol. A total of 23,097 cells were detected on the basis of

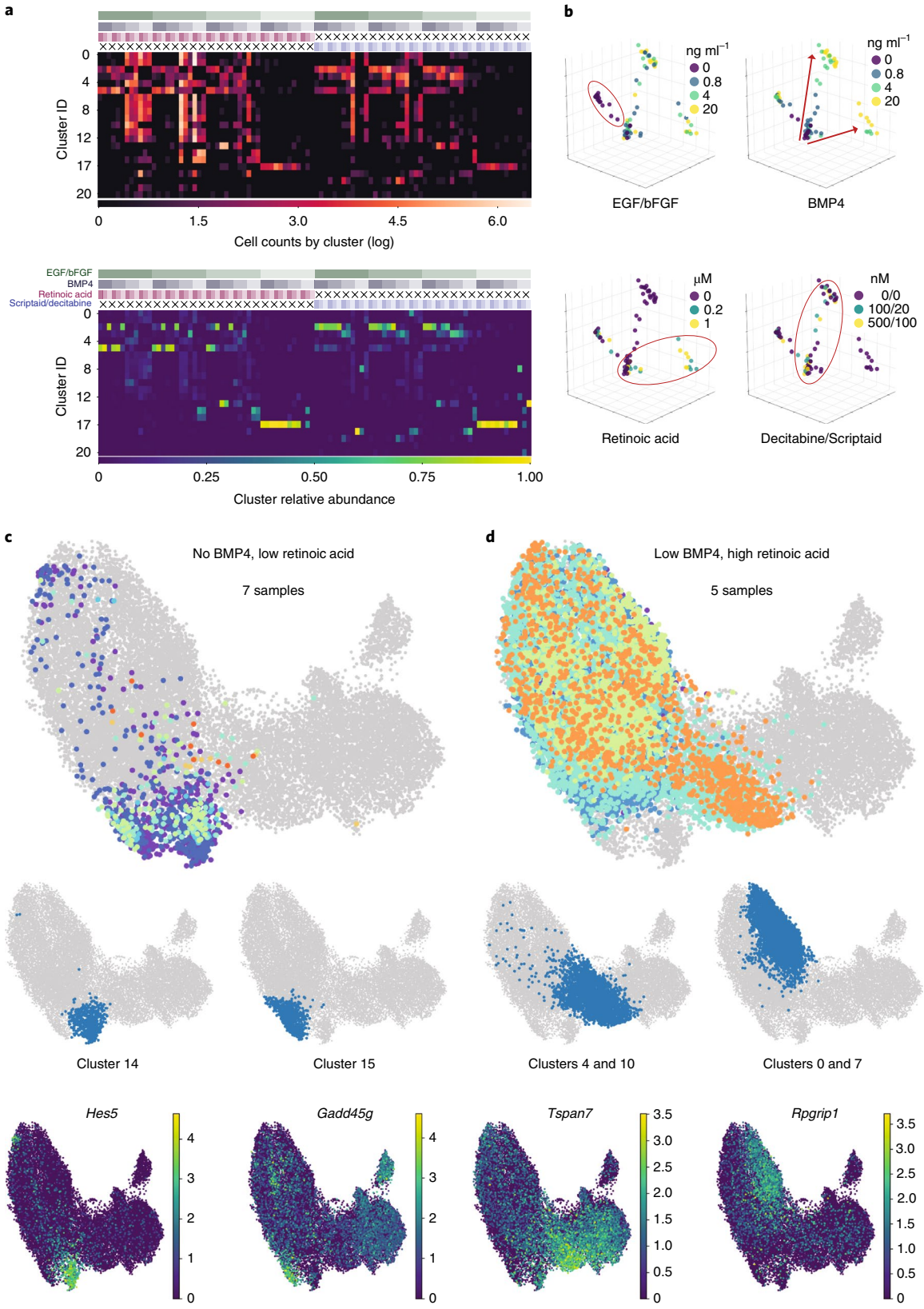

Fig. 2 | Perturbation responses at single-cell resolution. a, Cluster occupancy versus experimental condition shown as number of cells (top) or relative abundance (bottom) of $n = 21{,}191$ cells assigned to each of n = 21 clusters shown in Fig. 1. **b**, PCA of relative cluster abundance matrix from **a** (96 conditions × 21 clusters). Each point represents a cell population from one of 96 experimental conditions, and each display of the PCA reveals patterns of influence for a different experimental factor. **c**, Dissection of heterogeneous cell populations by cluster and condition for $n = 21{,}191$ cells. Seven samples from conditions with low retinoic acid and lacking BMP4 yielded cell populations predominantly mapping to clusters 14 and 15, which are distinguished by unique marker genes including the neural differentiation markers *Hes5* and *Gadd45g* (**c**). **d**, Similar segmentation to **c** was achieved for the highly proliferative cell states arising from samples treated with low BMP4 and high retinoic acid concentrations.

cDNA counts, and sample assignment was performed for these cells on the basis of a simple thresholding of ClickTag UMI counts. High concordance with the experimental design was observed, with 21,223 cells (92%) classified as positive for exactly two ClickTags, of which 99.8% corresponded to a pair in the experimental design (96 pairs were used of 190 possible combinations). Visualization of the cell populations produced by each experimental condition revealed a complex interplay between the perturbants (Fig. 1e). On a global level, cell proliferation varied widely across the experiment, revealing growth rates specific not just to each condition but also to each cell state across the experiment. Highly proliferative states (clusters 0, 1, 4, 6, 7, 8, 9, 10, 11, 12 and 17) differentially expressed various genes associated with cell growth and the cell cycle, including ribosomal, cytoskeletal and cyclin-dependent proteins (Supplementary Table 3). Conversely, samples deprived of EGF and bFGF exhibited apoptotic phenotypes including low cell counts and expression of stress-response genes such as *Cryab*, *Mt1* and *Gpx4*. We sought to define the cell states produced by the array of experimental conditions, a challenging procedure in scRNA-seq analysis and a potential roadblock to perturbation experiments, in which the presence of classical marker genes may depend on experimental conditions. We found that various distinct regions of transcriptional space were repeatedly populated by cells originating from multiple samples in localized regions of perturbation space, forming natural groupings of cells that were validated and assigned by clustering (Fig. 1d). The cluster occupancy of each sample revealed the structure of the cell populations produced across the experiment (Fig. 2a). Overall trends, such as high proliferation under low BMP4 conditions and high cluster specificity under high BMP4 conditions, were readily observed. Principal component analysis (PCA) of the relative cluster abundance × sample matrix revealed relationships between the experimental inputs (Fig. 2b). The cell populations from each scRNA-seq sample associated directly with the experimental perturbations. Absence of EGF and bFGF has a drastic effect, yielding an isolated group of samples in PCA space, while BMP4 concentration has a graded effect and a strong interaction with retinoic acid, producing low BMP4, high BMP4, and BMP4 and retinoic acid cell states. This analysis demonstrated that multiplexed scRNA-seq can be used to classify cell populations and interpret the conditions that produced them. In the context of a perturbation experiment, relevant features of the experimental space can be identified, for example, the strong effect of BMP4 concentration shown here. Of perhaps greater interest would be to extend this proof of principle to biomedical diagnostics: by applying Bayes rule to the relative cluster abundance × sample matrix, it should be possible to infer disease conditions from high-resolution cell population observations.

Next, we closely examined two regions of our experimental space occupied by cells treated with intermediate EGF and bFGF concentrations, no BMP4 and moderate to no retinoic acid. Cells from seven samples accounted for practically all of clusters 14 and 15 and little across the rest of cell state space, exhibiting strong condition dependence (Fig. 2c). Differential expression analysis showed that cells in cluster 14 are defined by *Hes5* expression and those in cluster 15 are defined by *Gadd45g* expression[13–15]. Elsewhere in experimental space, treatment with low BMP4 and high concentrations of retinoic acid generated highly proliferative cell states with complex population architectures (Fig. 2d). In this way, multiplexed scRNA-seq provides a detailed molecular dissection of heterogeneous cell populations produced from complex experimental conditions, addressing a long-standing goal in cell biology[16–20].

It has been hypothesized that cells occupy a relatively limited number of transcriptional states in response to disease or experimental perturbation, and elucidating the connections between various perturbations will help in understanding cellular behavior. Efforts such as the Connectivity Map[19] project, while impressive in scope, suffer from batch effects, averaging across cell populations and difficulty in examining conditions that yield very few cells. ClickTag multiplexing overcomes these obstacles and provides single-cell whole-transcriptome resolution at very low cost. Just as multiplexing of DNA sequencing libraries has vastly improved the utility and adoption of high-throughput DNA sequencing, ClickTags along with other solutions for multiplexed scRNA-seq[5–8] will similarly reduce costs, drive further increases in cell capacity, and extend the scope of scRNA-seq beyond bulk tissue profiling, enabling comparison of complex experimental samples with previously unattainable depth and scale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-019-0372-z.

Received: 7 August 2018; Accepted: 27 November 2019;
Published online: 23 December 2019

References

1. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
2. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
3. Han, X. et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 (2018).
4. Cao, J et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
5. Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
6. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
7. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
8. Guo, C. et al. CellTag indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.* **20**, 90 (2019).
9. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
10. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018).
11. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
12. Bond, A. M., Ming, G. & Song, H. Adult mammalian neural stem cells and neurogenesis: five decades later. *Cell Stem Cell* **17**, 385–395 (2015).
13. Imayoshi, I. et al. Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. *Science* **342**, 1203–1208 (2013).
14. Hitoshi, S. et al. Mammalian *Gcm* genes induce *Hes5* expression by active DNA demethylation and induce neural stem cells. *Nat. Neurosci.* **14**, 957–964 (2011).
15. Yuzwa, S. A. et al. Developmental emergence of adult neural stem cells as revealed by single-cell transcriptional profiling. *Cell Rep.* **21**, 3970–3986 (2017).
16. Janes, K. A. A Systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
17. Nelander, S. et al. Models from experiments: combinatorial drug perturbations of cancer cells. *Mol. Syst. Biol.* **4**, 216 (2008).
18. Sims, D. et al. High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol.* **12**, R104 (2011).
19. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
20. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).

Methods

Overview of cell tagging procedure. Barcoded DNA oligonucleotides (ClickTags) were attached to exposed NHS-reactive amines on methanol-fixed cells of interest. ClickTag labeling was achieved in a one-pot, two-step reaction by combining cell samples with methyltetrazine-activated DNA (MTZ-DNA) oligonucleotides and the amine-reactive cross-linker NHS-*trans*-cyclooctene (NHS-TCO) (Fig. 1a). NHS-functionalized oligonucleotides were formed in situ via the IEDDA reaction and nucleophilic attack by accessible cellular amines chemoprecipitated the oligonucleotides directly onto the dehydrated cells. Our one-pot reaction that is based on IEDDA chemistry improves on a previous cell-surface-modification scheme[21] that requires far higher DNA concentrations and isolation of unstable, NHS-activated DNAs immediately before use. A panel of methyltetrazine-modified ClickTags can be prepared in advance, stored at −20 °C for months and applied to many cell samples in parallel. Sequencing library preparation is derived from published methods for multimodal scRNA-seq[5,22]. In brief, the ClickTag sequences contain poly(dA) tails that are captured and copied during reverse transcription, acquiring the same cell barcodes as the associated mRNAs from the same cell. The resulting short dsDNAs are isolated during SPRI purification, specifically amplified and purified by agarose gel electrophoresis.

Oligonucleotide activation. ClickTags were prepared with either 5′- or 3′-amine-modified oligonucleotides (100- to 250-nmol scale, Integrated DNA Technologies; Supplementary Table 4). High-perfomance liquid chromatography (HPLC) purification was critical to obtain highly reactive preparations of 5′-modified oligonucleotides, while 3′-modified oligonucleotides were purchased without HPLC purification. In either case, oligonucleotides were resuspended to a concentration of 500 µM in 50 mM sodium borate buffer, pH 8.5 (Thermo Fisher Scientific). NHS esters, TCO-Cy5 and MTZ-Cy5 were resuspended in dry DMSO and stored in single-use aliquots at −80 °C at the following concentrations: NHS-TCO 20 mM, NHS-MTZ 10 mM, MTZ-Cy5 1 mM and TCO-Cy5 1 mM. For NHS-TCO and modified fluorophores, dilution in DMSO was performed immediately before use.

Oligonucleotide activation reactions were performed by combining 25 µl of oligonucleotide solution with 41.8 µl of DMSO (Sigma) and 8.2 µl of 10 mM NHS-methyltetrazine (Click Chemistry Tools). The reaction was allowed to proceed for 30 min at room temperature on a rotating platform. After 30 and 60 min, additional 8.2-µl aliquots of 10 mM NHS-methyltetrazine were added. After 90 min total reaction time, ethanol precipitation was performed by addition of 180 µl of 50 mM sodium borate buffer and 30 µl of 3 M NaCl. After mixing, 750 µl of ice-cold ethanol was added and the mixture precipitated at −80 °C overnight. The precipitate was pelleted at 20,000*g* for 30 min at 4 °C, washed twice with 1 ml of ice-cold 70% ethanol, then resuspended in 100 µl of 10 mM HEPES, pH 7.2. Yield was determined by absorbance at 260 nm. Typical final concentrations ranged between 40 and 80 µM.

Relative oligonucleotide activity was determined by electrophoretic mobility shift assay using Cy5-*trans*-cyclooctene (Click Chemistry Tools). Methyltetrazine-derivatized oligonucleotides were diluted 100-fold in 10 mM HEPES, pH 7.2, then 4 µl of this solution was added to 1 µl of a 500 nM solution of TCO-Cy5 in DMSO. All tetrazine reactions in this work were performed in the dark to protect the photoreactive *trans*-cyclooctene group. The reaction was allowed to proceed at room temperature for 20–120 min and analyzed on a 12% SDS-PAGE gel. Oligonucleotide activity varied within a twofold range across preparations. Oligonucleotides were stored at −20 °C and used without further normalization. Gel electrophoresis activity assay and confirmation of cell labeling by microscopy, as demonstrated in Supplementary Fig. 1, are strongly recommended for new users. MTZ-DNAs can be stored for months without loss of activity, but use of oligonucleotides with more than twofold difference in concentration or activity may reduce cell labeling performance across samples, and the activity of separate batches should be compared by gel electrophoresis activity assay before being used together.

Cell culture and fixation. NSCs were cultured according to the following protocol. Cryopreserved mouse NSCs were thawed for 2 min at 37 °C then transferred to a 15-ml conical tube. Prewarmed NSC basal medium (SCM003, Millipore) was slowly added to a total volume of 10 ml and the resulting cell suspension was centrifuged at room temperature for 2.5 min at 200*g*. The supernatant was removed and the cell pellet was resuspended in 2 ml of prewarmed NSC basal medium and counted on the Countess II Automated Cell Counter (Thermo Fisher Scientific). Cells were seeded on 100-mm culture plates coated with poly-L-ornithine (Millipore) and laminin (Thermo Fisher Scientific) at 700,000 cells per plate in 10 ml of prewarmed NSC basal medium supplemented with EGF (Millipore) and bFGF (Millipore) at 20 ng ml^{-1} each, heparin (Sigma) at 2 µg ml^{-1} and 1% penicillin–streptomycin (Thermo Fisher Scientific). Supplemented medium was changed the next day and every other day thereafter until confluence was reached.

NSCs for the 96-sample growth factor screen were cultured according to the following protocol. After the cell culture plate described above reached ~80% confluence, stock solutions (10×) were prepared in NSC basal medium for every factor and at every concentration used (EGF and bFGF at 200 ng ml^{-1}, 40 ng ml^{-1}, 8 ng ml^{-1} and 1.6 ng ml^{-1}; BMP4 (Peprotech) at 200 ng ml^{-1}, 40 ng ml^{-1}, 8 ng ml^{-1} and 0 ng ml^{-1}; retinoic acid (Sigma) at 10 µM, 2 µM and 0 µM; Scriptaid (Selleckchem) and decitabine (Selleckchem) at 1 µM and 5 µM, 0.2 µM and 1 µM, and 0 µM and 0 µM; heparin at 20 µg ml^{-1} and penicillin–streptomycin at 10%). Twenty microliters of each stock (EGF and bFGF, BMP4, retinoic acid or Scriptaid/decitabine, and heparin/penicillin–streptomycin) were added to each well of a 96-well plate coated with poly-L-ornithine and laminin for a total of 80 µl.

NSCs previously plated on 100-mm culture plates until ~80% confluent were dissociated by incubation in 4 ml of ESGRO Complete Accutase (Millipore) for 2 min at 37 °C. After incubation, the Accutase and NSCs were transferred to a 15-ml conical tube and centrifuged at room temperature for 2.5 min at 200*g*. Supernatant was removed and the cell pellet was resuspended in 2 ml of NSC basal medium. Centrifugation and medium replacement were repeated once more and cell concentration was counted on the Countess II Automated Cell Counter. The cell suspension was then diluted with additional NSC basal medium to a concentration of 18.3 cells per microliter. From this stock, 120 µl was added to each well of the 96-well plate for a total of ~2,200 cells per well. Supplemented medium for every well in the 96-well plate was replaced every other day during the 5-d incubation.

Before NSC dissociation and fixation, 80 µl of ice-cold methanol was added to each well of twelve eight-well PCR strips on an ice block. After 5 d in culture, all medium in the 96-well plate was removed and the cells were washed three times with 150 µl of NSC basal medium. Any remaining medium was removed and replaced with 20 µl of Accutase and incubated at 37 °C for 2 min with gentle pipetting to help break up cell clumps. Next, 20 µl of dissociated NSCs in Accutase was transferred to the eight-well strip tubes containing 80 µl of 100% methanol and the entire volume was pipetted to mix. After fixation, the NSCs were stored at −20 °C until sample labeling.

For four-sample NSC labeling and species-mixing experiments, NSCs were cultured on a 100-mm culture plate coated with poly-L-ornithine and laminin according to the protocol previously described until ~80% confluence was reached. NSCs were dissociated by removing culture medium followed by incubation with 4 ml of Accutase for 2 min. NSCs in Accutase were transferred to a 15-ml conical tube and centrifuged at room temperature for 2.5 min at 200*g*. The supernatant was removed and the cell pellet was resuspended in 2 ml Hank's balanced salt solution (HBSS, Thermo Fisher Scientific) with 0.04% BSA (Sigma). Centrifugation and medium replacement were repeated once and cell concentration was determined on a Countess II Automated Cell Counter. Cells were then fixed by addition of four volumes of ice-cold methanol added slowly with constant mixing. Fixed cells were stored at −20 °C until ClickTag labeling and scRNA-seq.

Frozen stocks of HEK293T cells (ATCC) were thawed for 2 min at 37 °C with gentle agitation. Thawed cells (500 µl) were added to 5 ml of prewarmed medium (DMEM (Corning) with 10% FBS (Gemini Bio-Products) and 1% penicillin–streptomycin (Corning) and centrifuged at 1,500*g* for 5 min. The cells were resuspended in 5 ml of medium and transferred to a T-25 cell culture flask. Cells were grown at 37 °C with 5% CO_2 following standard practices. HEK293T cells were dissociated by incubation with TrypLE Select (Thermo) for 5 min at 37 °C, washed twice with HBSS and resuspended in 1 ml at a concentration of ~6×10^6 cells per milliliter. Cell number and viability were measured using a Countess II Automated Cell Counter. Four milliliters of ice-cold methanol was added slowly with constant mixing and the resulting cell suspension was incubated at −20 °C for at least 20 min. Cells were stored at −20 °C until ClickTag labeling and scRNA-seq.

Flow cytometry and fluorescence microscopy. Yeast cells (Fleischmann's Rapid Rise) were used as an abundant cellular substrate to test cell labeling reactions. Approximately 5 g of dehydrated cells were rehydrated in 4 ml of PBS with 0.1% Tween-20 (Sigma) for 10 min at room temperature with rotation. One milliliter of the resulting cell suspension was diluted with 7 ml of PBS–Tween and fixed by slow addition of 32 ml of ice-cold methanol with constant mixing. Cells were incubated at −20 °C for at least 20 min before further use.

Methanol-fixed cells were rehydrated by combining 700 µl of HBSS with 500 µl of fixed cells in 80% methanol. This suspension was centrifuged at 3,000*g* for 5 min, then washed twice more with HBSS. Cells were resuspended in 1 ml of HBSS and 50 µl of this cell suspension was used for cell labeling. Methyltetrazine-Cy5 (Click Chemistry Tools) was added to final concentration of 2 µM, NHS-TCO to 5 µM and DAPI to 1 µg ml^{-1}. Cell labeling reactions were incubated for 30 min at room temperature with rotation then quenched by addition of Tris-HCl to 10 mM and methyltetrazine-DBCO (Click Chemistry Tools) to 50 µM. Samples were diluted 20-fold in HBSS and analyzed on a MACSQuant VYB flow cytometer.

Fluorescence microscopy samples were prepared as above, except NHS-TCO was used at 1 µM and MTZ-Cy5 was used at 62.5 µM. Samples were imaged on a Zeiss LSM 800 laser scanning confocal microscope.

Multiplexed scRNA-seq proof of concept. Fixed NSCs were split into four aliquots with ~400,000 cells in 100 µl of 80% methanol. Live NSCs were prepared as described above, washed in HBSS and similarly aliquoted to four samples with 400,000 cells in 100 µl. Before cell labeling, eight labeling combinations were made by combining 6 µl each of two different MTZ-derivatized ClickTags. A 5-min preincubation reaction was performed in the dark at room temperature by addition of 4 µl of 1 mM NHS-TCO. After preincubation, cell suspensions were thoroughly

mixed with the entire volume of a single ClickTag labeling mix. Cell labeling proceeded for 30 min at room temperature on a rotating platform. Reactions were quenched by addition of Tris-HCl to a final concentration of 10 mM and methyltetrazine-DBCO (Click Chemistry Tools) to 50 µM. After quenching for 5 min, cells were pooled to create a single sample for fixed cells and a single sample for live cells. Two volumes of PBS of BSA were added and the cells were pelleted by centrifugation at 500*g* for 5 min. Cells were washed three times with PBS–BSA and vigorously resuspended in a final volume of 150 µl. Cells were analyzed, counted and loaded on two lanes of the Chromium Controller (10x Genomics) targeting 5,000 cells each. Library preparation was adapted from the REAP-Seq protocol[6]. The 10x Genomics v2 Single Cell 3′ sequencing reagent kit protocol was used to process samples according to the manufacturer's procedure with the following modification: after initial amplification of cDNA and ClickTags, the two libraries were separated during SPRI size-selection. The manufacturer's instructions were used to complete full-length cDNA library preparation. For ClickTags, rather than discarding the 0.6× SPRI supernatant, this fraction was combined with more SPRI beads (final SPRI ratio 1.5×) and incubated at room temperature for 5 min. The SPRI beads were washed twice with 80% ethanol and ClickTags were eluted in 20 µl of nuclease-free water. ClickTags were quantified by Qubit high-sensitivity DNA assay (Invitrogen) and amplified using primer R1-P5 and indexed reverse primers as appropriate (Supplementary Table 4). PCR was performed in a 25-µl volume including 2.5 µl of ClickTag library, 1.5 µl of 10 µM forward, 1.5 µl of 10 µM reverse primer, 7 µl of nuclease-free water and 12.5 µl of KAPA 2× HIFI PCR master mix (Kapa Biosystems). The samples were cycled as follows: 98 °C for 3 min, 16 cycles of 98 °C for 20 s, 58 °C for 30 s and 72 °C for 20 s, and a final extension step of 72 °C for 4 min. Final ClickTag libraries were obtained using a PippinPrep automated size-selection system with a 3% agarose gel set for a broad purification range from 200–250 base pairs (target library size is 225 base pairs). A Qubit assay was again used to determine library concentration for sequencing. ClickTag and cDNA libraries were analyzed on a BioAnalyzer high-sensitivity DNA kit (Agilent). Example traces are provided in Supplementary Fig. 19 for reference. ClickTag libraries were sequenced on an Illumina MiSeq using a MiSeq V3 150 cycle kit (26×98-base-pair reads) and cDNA libraries were not sequenced for this proof-of-concept experiment.

Species mixing and ClickTag Multiplexing. Methanol-fixed human HEK293T and mouse NSCs were prepared as described above. Samples were labeled with non-overlapping ClickTag sets of increasing size (Supplementary Table 1). Suspensions of both cell types were prepared at 700,000 cells per milliliter in 80% methanol. Samples of 100 µl were prepared for each condition, with species-mixing conditions comprising 50 µl of cell suspension from each species. For this experiment, 3′-modified oligonucleotides isolated by standard desalting were used as opposed to the 5′-modified, HPLC-purified oligonucleotides that were used in all other experiments. ClickTag sets were prepared by reacting 6 µl of each oligonucleotide along with 2 µl of 1 mM NHS-TCO per oligonucleotide at room temperature. After 5 min, the entire volume of each labeling mixture was added to the appropriate cell suspension. Cell labeling was performed for 30 min at room temperature on a rotating platform. Reactions were quenched as above, pooled and added to 2 ml of PBS with 1% BSA. Samples were split across two Eppendorf tubes and centrifuged at 500*g* for 5 min. Cell pellets were resuspended in 500 µl of PBS–BSA, combined and centrifuged once more. The cell pellet was washed twice more with 1 ml of PBS–BSA. Finally, the cells were resuspended in 150 µl of PBS–BSA, counted, diluted to 1×10^6 cells per milliliter and loaded on a single lane of the Chromium Controller targeting 10,000 cells. ClickTag and cDNA libraries were prepared as described. Libraries were submitted for sequencing as part of an Illumina NovaSeq library, targeting 500,000,000 reads total (2×150-base-pair reads), with ClickTags submitted at 10% of the total library concentration.

Ninety-six-sample growth factor screen. NSCs for the 96-sample perturbation experiment were prepared as described above. For each sample, two ClickTags (6 µl each) were combined with 4 µl of 1 mM NHS-TCO according to an 8×12 matrix. Columns 1–12 of the 96-well plate corresponded to ClickTags BC21–BC32, while rows A–H corresponded to ClickTags BC33–BC40 (Supplementary Fig. 16). Fixed cells from each experimental condition (100 µl) were labeled with the entire volume of the corresponding ClickTag mix for 30 min at room temperature on a rotating platform. Samples were quenched as described above, pooled and combined with 15 ml of PBS–BSA. Samples were split across two 15-ml conical tubes and spun at 500*g* for 5 min. Cell pellets were resuspended in 3 ml of PBS–BSA each and centrifuged again. The pellets were washed twice with 1 ml of PBS–BSA and resuspended in a final combined volume of 200 µl. Cells were loaded on two lanes of the 10x Chromium Controller targeting 10,000 cells per lane. Sequencing libraries were prepared as two large libraries (9,000 cells each) and two small libraries (1,000 cells each). ClickTag amplicons were sequenced on two lanes of an Illumina MiSeq using MiSeq v3 150 cycle kits (26×98-base-pair reads) and cDNA libraries were pooled and sequenced on an Illumina HiSeq 4000 using two HiSeq 3000/4000 SBS 300 cycle kits (2×150-base-pair reads).

cDNA data processing. Raw sequenced reads were processed using the 10x Genomics Cell Ranger pipeline (v.3.0.0). The 'cellranger mkfastq' command was used to demultiplex libraries on the basis of sample indices and to convert the barcode and read data to FASTQ format files. The 'cellranger count' command was used to identify cell barcodes and to align reads to the mouse or human transcriptomes (mm10 and hg19) as appropriate. For the 96-sample perturbation experiment, the 'cellranger aggr' command was used to combine and normalize sequencing data from the two 10x lanes split across two HiSeq lanes. Cells were selected by Cell Ranger using the inflection point of the rank-UMI versus cell barcodes plot.

ClickTag data processing, assignment and doublet detection. Cell barcode error correction was performed using the 10x barcode whitelist. Subsequently, sequenced reads from the ClickTag libraries were processed with a new feature barcode processing workflow, kITE, which is built on the 'kallito | bustools' scRNA-seq workflow[23,24]. In kITE, the ClickTag barcodes used in a given experiment are used to generate a 'mismatch index' consisting of the whitelist feature barcodes and all of their Hamming distance 1 variations (Supplementary Fig. 2). A kallisto index is produced from the mismatch index and the 'kallisto bus' command is used to pseudoalign ClickTag reads against the mismatch index. The output is a BUS file entry for every unique feature barcode, UMI and set combination. Finally, the BUS file is converted to a cells×ClickTags digital count matrix by collapsing counts from each ClickTag feature barcode with those corresponding to its Hamming distance 1 mismatches.

Sample assignment for the four-sample NSC experiments was performed for the top 3,800 cells with the most ClickTag UMIs. For each ClickTag, a threshold was calculated using the numpy gradient function to find the maximum slope of the rank-UMI versus cell barcode plot. The cells×ClickTags matrix was further processed using the ScanPy[25] single-cell analysis package. The data were normalized to 1,000 reads per cell and log-transformed, followed by *t*-SNE embedding and clustering by Louvain community detection. The Scrublet doublet detection algorithm was used to isolate likely multiplets from the ClickTag data and this subset of cells was reclustered to generate the violin plots in Supplementary Fig. 5.

For the species-mixing experiment, 11,264 valid cell barcodes were selected by Cell Ranger using the rank-UMI versus cell barcode plot for the cDNA libraries. The same cell barcodes were extracted from a cells×ClickTags matrix created using the kITE procedure described above. After normalizing each cell to 1,000 ClickTag counts, the data were log-transformed and the number of counts regressed out using ScanPy 'regress_out'. The resulting matrix was embedded with *t*-SNE and clustered by Louvain community detection. Of the nine clusters produced, one cluster showed greatly reduced UMI counts from both ClickTag and cDNA libraries and no clear correlation with any of the experimental groups (Supplementary Fig. 7). The cells in this 'noise cluster' were discarded, resulting in 10,482 high-quality cells, which were used for downstream analysis. Doublet identification was compared across three methods: Cell Ranger (cDNA-based), Scrublet (ClickTag count-based) and manual cluster selection (ClickTag *t*-SNE-based). Manual selection was performed with the FlowJo cytometry analysis software, isolating individual sub-clusters from the *t*-SNE embedding.

For the 96-plex NSC experiment, sample assignment and doublet removal were similar to the four-sample NSC experiment, beginning with identification of 23,068 cells on the basis of cDNA UMI counts. For each ClickTag, a threshold was calculated using the numpy gradient function to find the maximum slope of the rank-UMI versus cell barcode plot. Positive sample assignments were evaluated in comparison to the experimental design and cells with sample assignments that did not exactly match a ClickTag combination used in the experiment were filtered out, yielding 21,191 high-quality cells.

cDNA data analysis. For the species-mixing experiment, the species origin of each cell was determined by Cell Ranger using cDNA counts from each genome, and droplets containing cells were selected by Cell Ranger using the rank-UMI versus cell barcode plot. After filtering out cells with low ClickTag and cDNA counts (see above), the cDNA count data were normalized and log-transformed. PCA of the resulting matrix was followed by construction of a neighborhood graph and a uniform manifold approximation and projection (UMAP) embedding. These were computed using the ScanPy 'neighbors' and 'UMAP' functions with default settings.

For the 96-sample perturbation experiment, high-quality cells were selected as described above and ScanPy was used normalize and log-transform the data. Highly variable genes were selected as those with mean normalized counts greater than 0.0125 and less than 5 and with dispersion of more than 0.5, yielding 1,860 highly variable genes. The per-cell read counts were regressed out and the data were scaled to unit variance. PCA was performed on this matrix, followed by generation of a neighborhood graph using the top 40 principal components. The neighborhood graph was used to compute a dUMAP embedding and clustering was performed using the 'louvain' command in ScanPy. For clustering on the basis of Louvain community detection, the resolution parameter was adjusted to agree with subpopulations produced by the perturbation experiment. We reasoned that these natural groupings represent reproducible, quantitatively distinct biological states under the conditions of our experiment and would thus hold the most information relevant to the changing experimental parameters. In practice, a resolution setting of 2.2 yielded clusters that were in best agreement agreed with

the sample-specific subpopulations produced by the perturbation experiment. Sample assignments were combined with cluster assignments from each cell to produce a matrix of cluster occupancy × experimental condition as well as a normalized version of the same matrix consisting of the relative abundance of clusters for each sample (Fig. 2a). PCA was performed on the matrix of relative cluster abundances to visualize relationships between the experimental conditions used in our perturbation (Fig. 2b). Differential expression analysis was performed with the rank_genes_groups function in ScanPy. The top differential genes between the cluster(s) of interest and the rest of the dataset are shown (Fig. 2c,d). Linear regression was performed using the statsmodels Python package. The concentrations of the perturbants applied to each cell were used as independent variables and the corresponding gene expression for that cell was set as the dependent variable. Regression was performed for all highly variable genes and all high-quality cells. Genes with strong condition dependence were selected on the basis of their *P* values from this model.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data from these experiments can be obtained from CaltechDATA at https://doi.org/10.22002/D1.1311.

Code availability

Code and tutorials for the kITE demultiplexing workflow can be found at https://www.kallistobus.tools/kite_tutorial.html. Python notebooks used to process data and generate figures are available on GitHub at https://github.com/pachterlab/GPCTP_2019. The same GitHub repository also contains a fully reproducible reanalysis using 'kallisto | bustools' transcript alignments and a Google Colab notebook.

References

21. Hsiao, S. C. et al. Direct cell surface modification with DNA for the capture of primary cells and the investigation of myotube formation on defined patterns. *Langmuir* **25**, 6985–6991 (2009).
22. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
23. Melsted, P., Ntranos, V. & Pachter, L. The barcode, UMI, set format and BUStools. *Bioinformatics* **35**(21), 4472–4473 (2019).
24. Melsted, P. et al. Modular and efficient pre-processing of single-cell RNA-seq. Preprint at *bioRxiv* https://doi.org/10.1101/673285 (2019).
25. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Acknowledgements

We thank Z. Gartner and C. McGinnis for helpful feedback regarding the ClickTag protocol and V. Svensson for suggestions regarding analysis of multiplexed datasets. Thanks to P. Melsted and S. Booeshaghi for developing the 'kallisto | bustools' functions used in the preprocessing workflow and to P. Rivaud for assistance with 10x data processing. Additional support was provided by the the Caltech Bioinformatics Resource Center and the Single Cell Profiling and Engineering Center (SPEC) in the Beckman Institute at Caltech.

Author contributions

J.G. conceived and developed the ClickTag multiplexing strategy. J.G., J.H.P. and S.C. designed the scRNA-seq experiments and J.G. and J.H.P. performed the experiments. J.H.P. performed all tissue culture operations and J.G. developed the kITE demultiplexing workflow and analyzed the scRNA-seq data. J.G., J.H.P., S.C, M.T. and L.P. contributed to the interpretation of the results and writing of the manuscript.

Competing interests

J.G., L.P., S.C. and J.H.P. are listed as co-inventors on a patent application related to this work (US patent application 16/296,075).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0372-z.

Correspondence and requests for materials should be addressed to L.P.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Lior Pachter (lpachter@caltech.edu)
Jase Gehring (jgehring@caltech.edu)

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Illumina HiSeq sequencing data collection was performed by Fulgent Genetics. MiSeq data were collected in-house. In both cases, Illumina reads were first processed with the CellRanger version 3.0.0 mkfastq wrapper for the bcl2fastq function. MACQuantify Software 2.11 was used for flow cytometry data collection.
Data analysis	FlowJo version 10.4.1 was used for flow cytometry analysis and doublet detection in species mixing experiment. CellRagner version 3.0.0 was used to align and quantify human and mouse single-cell RNA-sequencing reads. The kallisto (v0.46), bustools (v0.39.2) and kite (v0.00) programs were used to align and quantify ClickTag sequencing data. Single-cell RNA-seq data were analyzed using the ScanPy Python package (version 1.0.4). Scrublet version 0.2.1 was also used for doublet detection.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequence data that support the findings of this study have been deposited in the CaltechDATA repository under DOI 10.22002/D1.1311

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the practical limitations of the platforms utilized. No statistical estimation of sample size was performed.
Data exclusions	No data were excluded from this study.
Replication	Cell labeling with DNA oligos has been verified in numerous independent experiments on diverse tissue types from cultured human and mouse cells to human PBMCs to zebrafish embryos and Cnidarian cells. Multiple batches tagged with orthogonal oligo sets served as replicates. In all cases, the optimized procedure described in the text yielded satisfactory sample assignment and demultiplexing with no modifications. In the case of the 96-sample multiplexing experiment, extremely limited cell numbers prevented an attempt at replication.
Randomization	In most experiments, sample arrangement was not relevant. For the 96-sample perturbation experiment, samples were arranged as described in the Supplementary Information. The complex effect of plate position mixed with sample condition in the 3-D array was not analyzed, but the interspersed arrangement of the various conditions largely precludes systematic plate effects.
Blinding	Cells and samples were grown and treated identically in all experiments performed; blinding was not used.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK 293T (ATCC) and 3T3 (ATCC)
Authentication	Cell lines were not authenticated.
Mycoplasma contamination	Cell lines not tested for Mycoplasma

Commonly misidentified lines (See ICLAC register)	None used.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Yeast cells (Fleischmann's Rapid Rise) were used as an abundant cellular substrate to test cell labeling reactions. Approximately 5 g of dehydrated cells were rehydrated in 4 ml PBS+ 0.1%Tween-20 (Sigma) for 10 minutes at room temperature with rotation. One ml of the resulting cell suspension was diluted with 7 ml PBS-Tween and fixed by slow addition of 32 ml ice-cold methanol with constant mixing. Cells were incubated at -20 •c for at least 20 minutes before further use. Methanol-fixed cells were rehydrated by combining 700 µl HBSS with 500 µl fixed cells in 80% methanol. This suspension was centrifuged at 3,000 x g for 5 minutes, then washed twice more with HBSS. Cells were resuspended in 1 ml HBSS, and 50 µl of this cell suspension was used for cell labeling. Methyltetrazine-Cy5 (Click Chemistry Tools) was added to 2 µM final concentration, NHS-TCO to 5 µM, and DAPI to 1 µg/ml. Cell labeling reactions were incubated for 30 minutes at room temperature with rotation then quenched by addition of Tris-HCl to 10 mM and methyltetrazine-DBCO (Click Chemistry Tools) to 50 µM. Samples were diluted 20-fold in HBSS and analyzed on a MACSQuant VYB flow cytometer.
Instrument	MACSQuant VYB flow cytometer
Software	MACSQuantify Software 2.11 was used to collect data, and plots were generated with FlowJo 10.4.1
Cell population abundance	Cells were analyzed based on a target number of events. The goal was to distinguish labeled and unlabeled cells, making population abundance irrelevant.
Gating strategy	Yeast cells were selected using FSC-A window 100-800 and SSC-A window 5-75. All events in this gate are shown according to their Cy5 fluorescence, demonstrating clear separation between positive and negative populations.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.