



Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing

Paul Datlinger^{1,3}, André F. Rendeiro^{1,3}, Thorina Boenke¹, Martin Senekowitsch¹, Thomas Krausgruber¹, Daniele Barreca¹ and Christoph Bock^{1,2}✉

Cell atlas projects and high-throughput perturbation screens require single-cell sequencing at a scale that is challenging with current technology. To enable cost-effective single-cell sequencing for millions of individual cells, we developed ‘single-cell combinatorial fluidic indexing’ (scifi). The scifi-RNA-seq assay combines one-step combinatorial preindexing of entire transcriptomes inside permeabilized cells with subsequent single-cell RNA-seq using microfluidics. Preindexing allows us to load several cells per droplet and computationally demultiplex their individual expression profiles. Thereby, scifi-RNA-seq massively increases the throughput of droplet-based single-cell RNA-seq, and provides a straightforward way of multiplexing thousands of samples in a single experiment. Compared with multi-round combinatorial indexing, scifi-RNA-seq provides an easy and efficient workflow. Compared to cell hashing methods, which flag and discard droplets containing more than one cell, scifi-RNA-seq resolves and retains individual transcriptomes from overloaded droplets. We benchmarked scifi-RNA-seq on various human and mouse cell lines, validated it for primary human T cells and applied it in a highly multiplexed CRISPR screen with single-cell transcriptome readout of T cell receptor activation.

Microfluidic droplet generators are currently the most popular technology platform for single-cell sequencing^{1–3}. With their high throughput, straightforward handling and consistent data quality, they have contributed to the wide adoption of single-cell RNA-seq (scRNA-seq) in many areas of basic biology and biomedical research. In a typical droplet-based scRNA-seq experiment, the single-cell suspension is processed on a microfluidic chip, together with uniquely barcoded microbeads, reverse transcription reagents and carrier oil (Fig. 1a). When aqueous and oil phases are combined at controlled flow rates, emulsion droplets coencapsulate individual cells with individual microbeads. Cells are then lysed, RNA molecules anneal to bead-tethered primers, and reverse transcription is performed inside the droplets. All primers on a given bead carry the same barcode, which uniquely identifies the coencapsulated cell. Once the transcripts have been indexed inside the microfluidic droplets, the emulsion is broken and the remaining steps of the library preparation are performed in bulk.

In droplet-based scRNA-seq, two cells that are loaded into the same droplet will have their transcriptomes labeled with the same cell barcode. As a result, transcripts from these two cells are indistinguishable during data analysis, creating a cell doublet that may bias the results (for example, by forming spurious intermediate cell populations). To minimize the number of cell doublets, the single-cell suspension is loaded into the microfluidic device at very low concentrations, making it unlikely that two cells enter the same droplet. This workaround is sufficient for many applications, but it leads to a major conceptual limitation: most emulsion droplets are fully functional (that is, they contain a barcoded microbead and reverse transcription reagents) but never receive a cell and

therefore cannot yield any single-cell transcriptomes (Fig. 1a). For this reason, droplet-based scRNA-seq uses its reagents very inefficiently, which contributes to the method’s high costs and renders it cost-prohibitive for large studies.

Cell hashing, which is used to multiplex several samples in one microfluidic droplet experiment, can help detect cell doublets, which are transcription profiles derived from droplets containing more than one cell. Through DNA-labeled antibodies⁴, lipid-tagged indices⁵ or expressed genetic barcodes⁶, each cell’s sample origin is encoded by a dedicated index sequence, and profiles from droplets with more than one index sequence are discarded as cell doublets. Alternatively, genotype information derived from the transcriptome sequences^{7–9} can be used to identify and remove cell doublets when samples from genetically different individuals are pooled. With improved detection of cell doublets, it becomes feasible to load the droplet generator at slightly increased cell concentrations without large biases due to spurious profiles. However, it is not possible to resolve the transcriptional profiles of cell doublets because most transcripts do not carry any cell-specific index. Therefore, the utility of cell hashing or cell genotypes for enhancing the throughput of scRNA-seq is modest, and more powerful approaches are needed.

To overcome the inefficiency of droplet-based scRNA-seq, we pursue cell-specific barcoding of all transcripts followed by massive overloading of microfluidic droplets. Transcript-level barcoding allows us to resolve each individual cell’s transcriptome when droplets contain several cells. We drew inspiration from recently published combinatorial indexing protocols^{10–12}, and we considered that one round of whole-transcriptome preindexing inside permeabilized cells, followed by droplet-based scRNA-seq with massive droplet overloading, could markedly increase the yield of existing

¹CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria. ²Institute of Artificial Intelligence, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria. ³These authors contributed equally: Paul Datlinger, André F. Rendeiro. ✉e-mail: cbrock@cemm.oew.ac.at

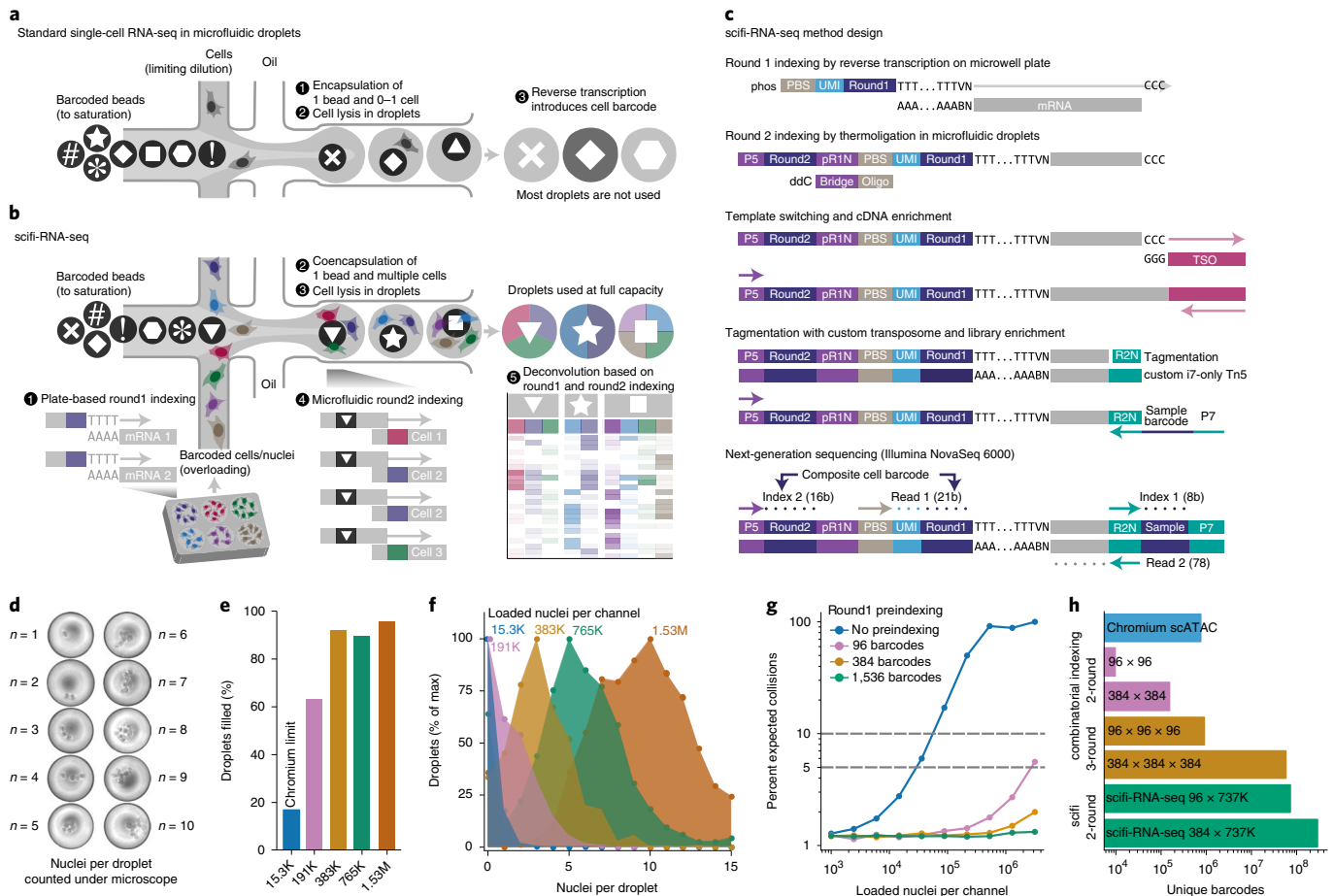


Fig. 1 | scifi-RNA-seq combines preindexing of whole transcriptomes with droplet-based scRNA-seq. **a**, Standard droplet-based scRNA-seq, where cells are loaded at a low concentration (limiting dilution) to avoid cell doublets, and most droplets do not receive a cell. **b**, scifi-RNA-seq, which uses preindexing and droplet overloading to boost the throughput of droplet-based scRNA-seq. **c**, Detailed method design of scifi-RNA-seq. **d**, Representative images of droplets containing between one and ten nuclei, showing the overloading of a standard microfluidic droplet generator (10x Genomics Chromium). **e**, Droplet overloading boosts the percentage of droplets filled with nuclei from 16.4% (obtained for the maximum loading concentration of the standard Chromium protocol) to 95.5% (obtained for 100-fold overloading using 1.53 million nuclei per channel). **f**, Droplet overloading causes the average number of nuclei per droplet to increase in a controlled fashion while maintaining the desired Poisson-like loading distribution. **g**, Expected collision rate as a function of the cell/nuclei loading concentration per channel for standard droplet-based scRNA-seq and for scifi-RNA-seq with different numbers of round1 barcodes. **h**, Due to the high number of microfluidic (round2) barcodes, scifi-RNA-seq exceeds the barcoding capability of three-round combinatorial indexing protocols.

microfluidic droplet generators. To distinguish our technology from single-cell combinatorial indexing (‘sci’) assays using microwell plates, while paying tribute to the elegant design of these methods, we refer to our barcoding strategy as ‘single-cell combinatorial fluidic indexing’ (scifi).

In scifi-RNA-seq (Fig. 1b,c), cells or nuclei are permeabilized and their transcriptomes are labeled with preindexing (round1) barcodes by reverse transcription in ‘split pools’, which are random bulk aliquots of cells or nuclei on 96-well or 384-well plates. Next, the cells or nuclei containing preindexed cDNA are pooled, randomly mixed and encapsulated using a standard microfluidic droplet generator with a high degree of overloading, such that most droplets receive several cells or nuclei. Inside these overloaded droplets, transcripts are labeled with a droplet-specific microfluidic (round2) barcode. The round1 barcodes are shared between all cells in the same split pool, the round2 barcodes are shared between all cells in the same droplet, but the combination of the two barcodes uniquely identifies transcripts derived from the same single cell.

Here we demonstrate a highly efficient implementation of scifi-RNA-seq, which makes it possible to obtain up to 150,000

single-cell transcriptomes per channel on the popular Chromium system (10x Genomics)³ and more than 1 million single-cell transcriptomes per microfluidics chip with its eight channels. We achieved high transcriptome complexity per cell, and successfully validated our method on various cell lines and on human primary material, studying T cell activation *ex vivo*. Finally, we demonstrate our method’s built-in support for massive sample multiplexing in an arrayed CRISPR screen for T cell receptor activation with scRNA-seq readout.

Results

Microfluidic droplet generators support massive overloading. Our method is built on the hypothesis that microfluidic droplet generators can tolerate substantial overloading (that is, encapsulation of several single cells or nuclei in the same droplet) while maintaining a stable, monodispersed droplet emulsion that does not clog the device. We confirmed this hypothesis for the Chromium system commercialized by 10x Genomics, on which we focused our technology development due to its large user base and the ensuing practical impact. Beyond the Chromium system, our method may also

be implemented on other microfluidic droplet generators including those used in Drop-seq¹ and inDrop², and it can be adapted to subnanoliter well plate assays such as CytoSeq¹³, Seq-Well¹⁴, Microwell-seq^{15,16} and the BD Rhapsody system¹⁷.

To test the feasibility of droplet overloading on the Chromium system, we prepared a single-nuclei suspension from human Jurkat cells and loaded the microfluidics chip at different concentrations. We did not include lysis reagents in these experiments, hence the nuclei inside the droplets stayed intact and could be counted under a light microscope (Fig. 1d and Extended Data Fig. 1a,b). We first assessed the maximum recommended loading concentration (15,300 nuclei per microfluidic channel). Counting 609 droplet images, we found that only 16.4% of droplets contained one or more nuclei (mean number of nuclei per droplet: 0.2). Remarkably, even 100-fold overloading (1.53 million nuclei per channel) resulted in a stable droplet emulsion and did not clog the microfluidic system (Extended Data Fig. 1c). As we increased the loading concentration, droplet fill rate and number of nuclei per droplet increased in a controlled manner (Fig. 1e–f), up to a droplet fill rate of 95.5% and an average of 9.6 nuclei per droplet (1.53 million nuclei per channel), with highly consistent droplet diameter (Extended Data Fig. 1d,e).

Of note, scifi-RNA-seq uses Chromium ATAC-seq reagents (rather than Chromium RNA-seq reagents) because the reverse transcription in scifi-RNA-seq is already completed when loading the droplet generator and because the gel bead oligonucleotides of the Chromium ATAC-seq kit contain a primer binding site suitable for ligation of the microfluidic (round2) barcodes to the preindexed cDNA. Given that the manufacturer recently changed the design of these microfluidic chips, we confirmed our results both with the previous Chromium Single Cell ATAC v.1.0 E chip (Fig. 1d–f and Extended Data Fig. 1) and with the new Chromium Single Cell ATAC v.1.1 (Next GEM) chip (Extended Data Fig. 1 and Extended Data Fig. 2a,b) and observed consistent results for both platforms.

Having established that droplet overloading is technically feasible, we sought to quantify the scale of preindexing required to uniquely resolve single-cell transcriptomes for 100,000s to millions of nuclei in a single scifi-RNA-seq experiment. To that end, we modeled the nuclei packaging process using four probability distributions (Extended Data Fig. 2d–f) and found the zero-inflated Poisson distribution most suitable (Extended Data Fig. 2g–i). We then estimated the expected percentage of unresolvable doublets due to barcode collisions for defined numbers of preindexing (round1) barcodes as a function of the number of loaded nuclei (Fig. 1g and Extended Data Fig. 2c). We also modeled the process with Monte Carlo simulations (Extended Data Fig. 2j). Using the 737,280 distinct microfluidic (round2) barcodes provided by the Chromium ATAC reagents, our analyses indicated that scifi-RNA-seq can resolve 1 million single-cell transcriptomes already with 96 round1 indices, which has practical advantages such as convenient handling and low setup cost. Moreover, scifi-RNA-seq with 384-well preindexing vastly exceeds the barcoding capacity of three-round (384 × 384 × 384) combinatorial indexing (Fig. 1h), with the prospect of higher transcriptome complexity and an easier workflow due to only one round of plate-based indexing.

Preindexing and droplet overloading boost scRNA-seq throughput. The experimental design of scifi-RNA-seq is outlined in Fig. 1c. Further technical details, such as oligonucleotide sequences, are provided in Extended Data Fig. 3a and Supplementary Table 1. We also prepared a user-friendly, step-by-step experimental protocol for scifi-RNA-seq, included here as a Supplementary Protocol. Future updates of this protocol will be shared via <http://scifi-rna-seq.bocklab.org>.

In scifi-RNA-seq, permeabilized cells or nuclei are first preindexed with barcoded oligo-dT primers by reverse transcription on multiwell plates (we use one 384-well plate with well-specific

primers). This step labels the mRNA molecules inside the permeabilized cells/nuclei with a preindexing (round1) barcode specific to each well, while also introducing a unique molecular identifier (UMI) and a primer binding site (PBS) for sequencing. Upon reaching the end of the transcript, the reverse transcriptase adds untemplated C nucleotides, resulting in a defined end to the cDNA molecules. Next, the permeabilized cells/nuclei (which contain the preindexed cDNA) are pooled, washed, filtered, counted and loaded into the Chromium system at a concentration that results in several cells/nuclei per droplet. Inside the droplets, the cells/nuclei are lysed and oligonucleotides carrying the microfluidic (round2) barcode (delivered via Chromium gel beads) are ligated to the cDNA via the 5'-phosphate group of the reverse transcription primer, directed by a complementary 3'-blocked bridge oligonucleotide. Efficient ligation is achieved through repeated cycling between denaturation and ligation temperatures with a thermostable ligase. Afterward, the droplet emulsion is broken, and the 3'-ends of the cDNA molecules are extended by template switching in bulk, followed by cDNA amplification. Finally, double-stranded cDNA is tagged with a custom i7-only transposome and enriched by PCR, which results in sequencing-ready libraries. Additional library barcodes can be introduced in the final step of library preparation, allowing for pooled sequencing of several scifi-RNA-seq libraries.

Notably, preindexing inside permeabilized nuclei was only minimally damaging, with nuclei recovery rates of 53.3% (cell lines) and 41.1% (human primary material) before loading the samples on the droplet microfluidics chip (Extended Data Fig. 3b). A representative image of preindexed nuclei is shown in Extended Data Fig. 3c. Typical fragment distributions are shown in Extended Data Fig. 3d,e, and performance metrics for sequencing on the NovaSeq platform are summarized in Extended Data Fig. 3f,g and detailed in Supplementary Table 2.

In addition to the feasibility of droplet overloading, a second main feasibility concern was whether the preindexed cells or nuclei would withstand the pressures inside the microfluidic droplet generator, given their previous exposure to high-temperature incubations and reagents used in the reverse transcription step. We tested our scifi-RNA-seq protocol in a series of experiments using whole cells permeabilized by methanol, freshly isolated nuclei and formaldehyde-fixed frozen nuclei. Results were clearly positive for all three types of input material (Extended Data Fig. 4), for both the scATAC v.1.0 and v.1.1 microfluidics chips (Extended Data Fig. 5a–e), and for two alternative reverse transcriptase enzymes tested (Extended Data Fig. 5f–i). These experiments confirm that plate-based preindexing of cells/nuclei is indeed compatible with subsequent fluidic indexing, and they indicate that scifi-RNA-seq is robust across different cell permeabilization methods, microfluidic chip architectures and enzymatic reaction conditions.

To evaluate the performance of scifi-RNA-seq as a function of droplet overloading, we loaded 15,300, 383,000 and 765,000 pre-indexed nuclei into single channels of the Chromium system. We then determined the inflection point in the distribution of UMIs across single-cell barcodes that separates cells/nuclei from noise in droplet-based scRNA-seq experiments (Fig. 2a). We found that the number of recovered single-cell transcriptomes scaled with the number of loaded nuclei. Furthermore, the preindexing allowed us to determine the effective number of individual nuclei (as identified by round1 barcodes) in each droplet (as identified by round2 barcodes) from the sequencing data (Fig. 2b). The average number of nuclei inside each droplet increased in a controlled fashion, reaching an average of 4.4 nuclei per droplet when 765,000 nuclei were loaded. These results indicate that scifi-RNA-seq recovers nuclei with comparable efficiency over a wide range of loading concentrations, consistent with our statistical modeling of the nuclei count data based on visual inspection of the droplet emulsion (Extended Data Fig. 2g–i).

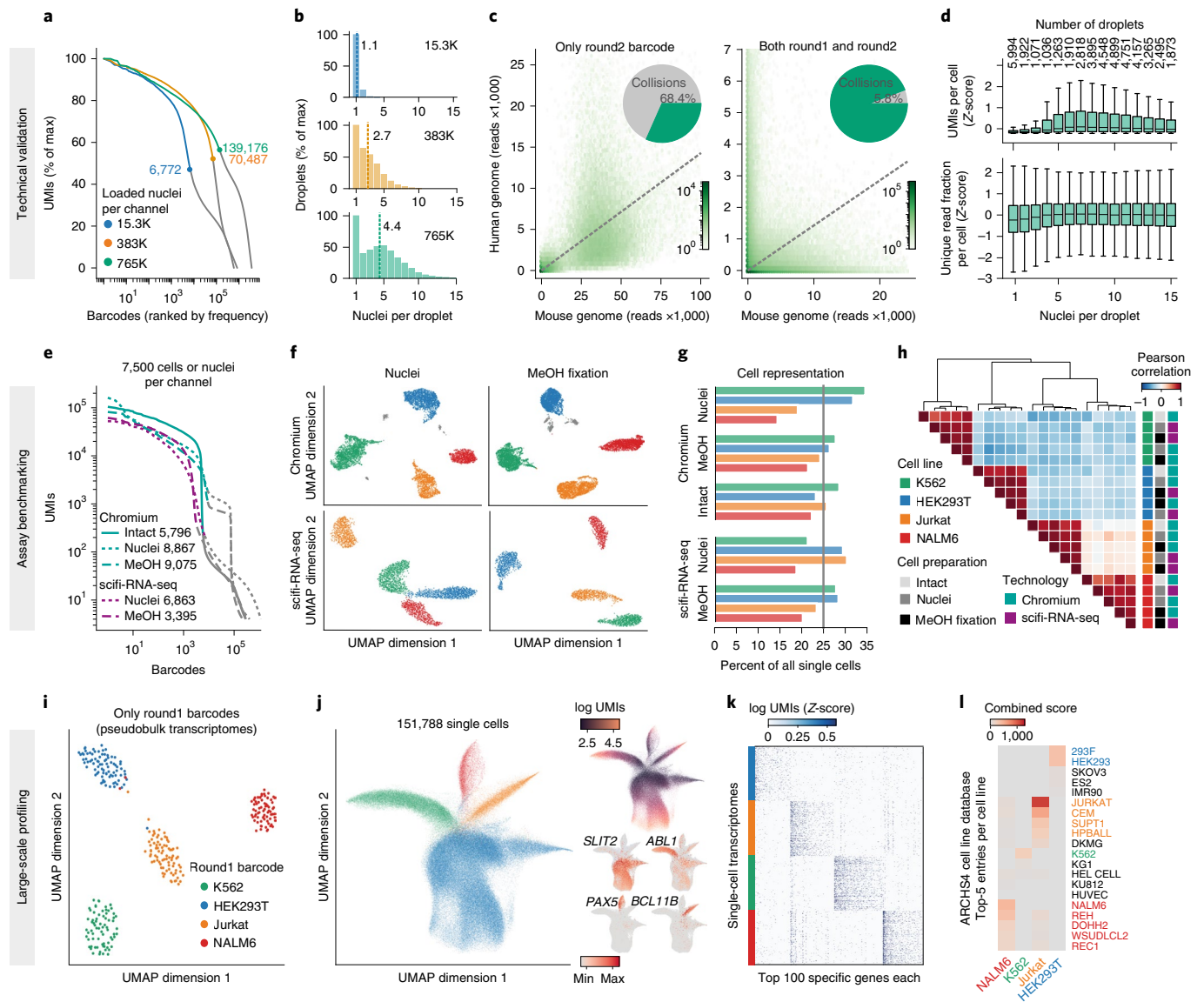


Fig. 2 | scifi-RNA-seq can produce ~150,000 single-cell transcriptomes per microfluidic channel. **a**, 'Knee plot' showing the number of UMIs (y-axis) per barcode ranked by frequency (x-axis) for scifi-RNA-seq experiments loading different numbers of nuclei. The characteristic inflection points are indicated, which separate nuclei (left, colored lines) from background noise (right, gray lines). To facilitate the comparison between samples, UMIs are normalized to percent of maximum. **b**, Distribution of the number of nuclei (round1 barcode) per droplet (round2 barcode) when loading different numbers of nuclei. The mean number of nuclei per droplet and nuclei loading concentration per channel are indicated. **c**, Species-mixing plots showing, for each droplet, the number of reads aligned to the mouse genome (x-axis) and human genome (y-axis). Transcriptomes were demultiplexed on the basis of the microfluidic round2 barcode alone (left) or on the basis of the combination of round1 and round2 barcodes (right). **d**, UMIs per cell and fraction of unique reads plotted against the number of nuclei contained in the respective droplet. Box plots depict the interquartile range with marked median and whiskers extending to 1.5 times the interquartile range. The number of droplets that each box plot summarizes is shown on top. **e**, 'Knee plot' for the comparison of scifi-RNA-seq and Chromium profiling using intact cells, nuclei or methanol-fixed cells with a standardized loading concentration of 7,500 cells or nuclei per microfluidic channel. **f**, Dimensionality reduction (UMAP) and clustering (Leiden algorithm) for the four cell lines. Spurious clusters of doublet cells (gray) are common for Chromium but absent for scifi-RNA-seq. **g**, Recovery rates for the four cell lines across technologies and cell preparation methods. **h**, Heatmap showing pairwise correlations and hierarchical clustering for the gene expression profiles across cell lines, cell preparation methods and profiling technologies. **i**, Dimensionality reduction for aggregated (pseudobulk) sample profiles in a large-scale scifi-RNA-seq experiment. **j**, Dimensionality reduction for 151,788 single-cell transcriptomes, colored by round1 barcodes corresponding to cell lines (left), UMIs per cell (top right) and marker gene expression (bottom right). **k**, Heatmap showing unfiltered, randomly sampled scifi-RNA-seq profiles for the 100 most specific genes per cell line. **l**, Gene set enrichment analysis of differentially expressed genes relative to the ARCHS4 database. Closely related cell lines are color coded.

This test dataset, which used a 1:1 mixture of human and mouse cell lines (Jurkat and 3T3, respectively), also allowed us to validate our preindexing strategy for the correct assignment of transcripts to single cells. To that end, we compared the number of human/mouse cell doublets on the basis of only the microfluidic (round2)

barcode with the number of such doublets based on the combination of preindexing (round1) and microfluidic (round2) barcodes. As expected for a loading rate of 765,000 nuclei per channel, most droplets contained both human and mouse cells (Fig. 2c, left panel), but we could resolve the vast majority of these doublets when

considering both the round1 and round2 barcode (Fig. 2c, right panel). As expected, the striking effect of preindexing for the resolution of cell doublets was seen only when the droplet generator was overloaded, while the microfluidic round2 barcode alone was sufficient for minimizing cell doublets at a standard loading rate of 15,300 nuclei per channel (Extended Data Fig. 4b).

Finally, this test dataset allowed us to conclusively resolve a third feasibility concern for scifi-RNA-seq, whether the reagents in each droplet would suffice for effective microfluidic indexing of the transcriptomes from several nuclei. When plotting UMI counts and fractions of unique reads per cell against the number of nuclei per droplet (Fig. 2d), we observed no trend toward lower transcriptome complexity in droplets containing up to 15 individual nuclei, supporting the observation that the reagents for droplet-based indexing are not a limiting factor in scifi-RNA-seq.

scifi-RNA-seq performs well in systematic benchmarking.

Having established the technical feasibility of scifi-RNA-seq, we sought to benchmark our method's empirical performance. First, we compared scifi-RNA-seq to multiround combinatorial indexing, using published data for sci-RNA-seq v.1 (ref. ¹⁰), SPLiT-seq¹¹, sci-RNA-seq v.3 (ref. ¹²) and sci-Plex¹⁸, with mouse 3T3 cells as the common cell type. We observed generally superior library quality for scifi-RNA-seq, and less read duplication (Extended Data Fig. 6a–f). Moreover, because sequencing cost is an important factor for large scRNA-seq experiments, we compared the cost effectiveness of the sequencing library designs and read structures (Extended Data Fig. 6g–h). For scifi-RNA-seq, it is possible to sequence the libraries in such a way that all sequencing cycles yield informative data, while the presence of constant ligation overhangs, primer binding sites and/or Tn5 mosaic ends in multiround combinatorial indexing renders a substantial proportion of sequencing cycles uninformative (sci-RNA-seq v.1: 42% uninformative; sci-RNA-seq v.3 and sci-Plex: 13% uninformative; SPLiT-seq: 67% uninformative). As a result, scifi-RNA-seq libraries can be sequenced with fewer sequencing cycles and/or can provide longer reads for the transcript as fewer cycles are spent on obtaining the barcode information. Moreover, scifi-RNA-seq data contained fewer cell collisions in the human/mouse species-mixing experiments, indicating a lower cell doublet rate (Extended Data Fig. 7).

Second, we compared scifi-RNA-seq to standard droplet-based scRNA-seq using the Chromium system, in a series of head-to-head experiments with the same samples and cell numbers. We prepared, as test samples, permeabilized nuclei and methanol-fixed cells for four human cell lines with variable cell size and mRNA content (HEK293T, Jurkat, K562 and NALM-6), and a species-mixing experiment using human (Jurkat) and mouse (3T3) cell lines. Samples were processed in parallel with the standard Chromium workflow and with scifi-RNA-seq, in both cases loading 7,500 cells per microfluidic channel. This experimental design allowed us to separate the effects of technology platform, permeabilization method, cell type, species and transcript content (Fig. 2e and Extended Data Fig. 8). Best results were obtained for intact cells processed with the Chromium workflow, while permeabilized nuclei and methanol-fixed cells on the Chromium platform showed high amounts of background due to cell fragments or free-floating RNA. scifi-RNA-seq did not reach the performance of the Chromium workflow on fresh cells but had less background, likely due to its stringent wash and filtration steps. Moreover, scifi-RNA-seq showed a slightly reduced cell recovery rate, which we can effectively compensate by loading a slightly higher cell concentration.

Reassuringly, all four cell lines were clearly separated by their transcriptomes (Fig. 2f and Extended Data Fig. 8a), and we recovered these cell lines with similar efficiencies despite their differences in cell size and transcriptome content (Fig. 2g). We noticed that the data obtained with the standard Chromium workflow included

spurious clusters corresponding to cell doublets (Extended Data Fig. 8a). These were absent from the scifi-RNA-seq data, indicating that scifi-RNA-seq is less susceptible to cell doublets than the Chromium workflow. For a direct comparison of transcriptome profiles across technologies and cell preparation methods, we calculated joint embeddings of single cells using five methods for dimensionality reduction (Extended Data Fig. 8b), and we quantified the cell separation using the silhouette score (Extended Data Fig. 8c). Single-cell transcriptomes consistently grouped by cell line without the need for batch effect correction and irrespective of assay, sample preparation or computational method. We further confirmed the consistency between scifi-RNA-seq and standard droplet-based scRNA-seq by comparing the transcriptome profiles (Fig. 2h), top differential genes (Extended Data Fig. 8d), log fold changes, *P* values and test statistics (Extended Data Fig. 8e).

scifi-RNA-seq scales well for cell lines and primary samples.

To demonstrate the throughput of our method, we performed a large-scale scifi-RNA-seq experiment with 383,000 nuclei loaded into a single microfluidic channel of the Chromium system. We combined four human cell lines (HEK293T, Jurkat, K562 and NALM-6), and we marked technical replicates of each cell line with different preindexing (round1) barcodes, thereby exploiting the inherent support of scifi-RNA-seq for multiplexing large numbers of samples in a single experiment. This experiment resulted in 151,788 single-cell transcriptomes passing quality control (Fig. 2i,j), constituting a 15-fold increase over the output of standard scRNA-seq on the Chromium system.

The aggregate profiles of the technical replicates for each cell line (Fig. 2i) as well as the single-cell transcriptomes (Fig. 2j, left) clustered consistently by cell line, with clear separation for all but the most shallowly sequenced cells (Fig. 2j, top right). The transcriptome data quality was high enough that even single marker genes such as *ABL1* for K562 and *PAX5* for NALM-6 were able to distinguish between the different cell lines (Fig. 2j, bottom right and Extended Data Fig. 9a), and the clustering was immediately apparent from the expression of the top-100 differential genes per cell line without filtering for high-coverage transcriptomes (Fig. 2k). Enrichment analysis against publicly available gene expression signatures accurately identified the corresponding cell lines on the basis of their gene expression profiles (Fig. 2l). Our analysis of ~150,000 single-cell transcriptomes thus confirmed that scifi-RNA-seq yields high-quality transcriptome profiles for human cell lines.

Next, we assessed how well scifi-RNA-seq performs on human primary cells, in an application where the differences in gene expression are more subtle. We purified human CD3+ T cells from the peripheral blood of three healthy donors and maintained the isolated T cells in short-term culture using human T cell medium containing IL-2. Half of the cells were T cell receptor (TCR) stimulated with anti-CD3/CD28 activator beads, while the other half were left untreated. Nuclei were fixed with formaldehyde and cryopreserved for easier handling. All samples were processed in a single scifi-RNA-seq experiment, using the preindexing step to uniquely barcode both the donors and the stimulation states. We obtained 62,558 single-cell transcriptomes that passed quality control.

Unsupervised analysis of the single-cell transcriptomes was dominated by TCR stimulation state (Extended Data Fig. 9b,c). Our data identified characteristic markers of T cell activation (Extended Data Fig. 9d) including *IL2RA* (CD25) and *IL12RB2* (which encode cytokine receptors) as well as *IRF4* and *MIR155HG* (which are known regulators of T cell biology). Graph-based clustering identified five groups of cells characterized by similar transcriptome profiles (Extended Data Fig. 9e), which we annotated by gene set analysis (Extended Data Fig. 9f). We found significant and cluster-specific enrichment of signaling pathways relevant to T cell activation, including signaling via the TCR, IL-2, CD40L and NFκB.

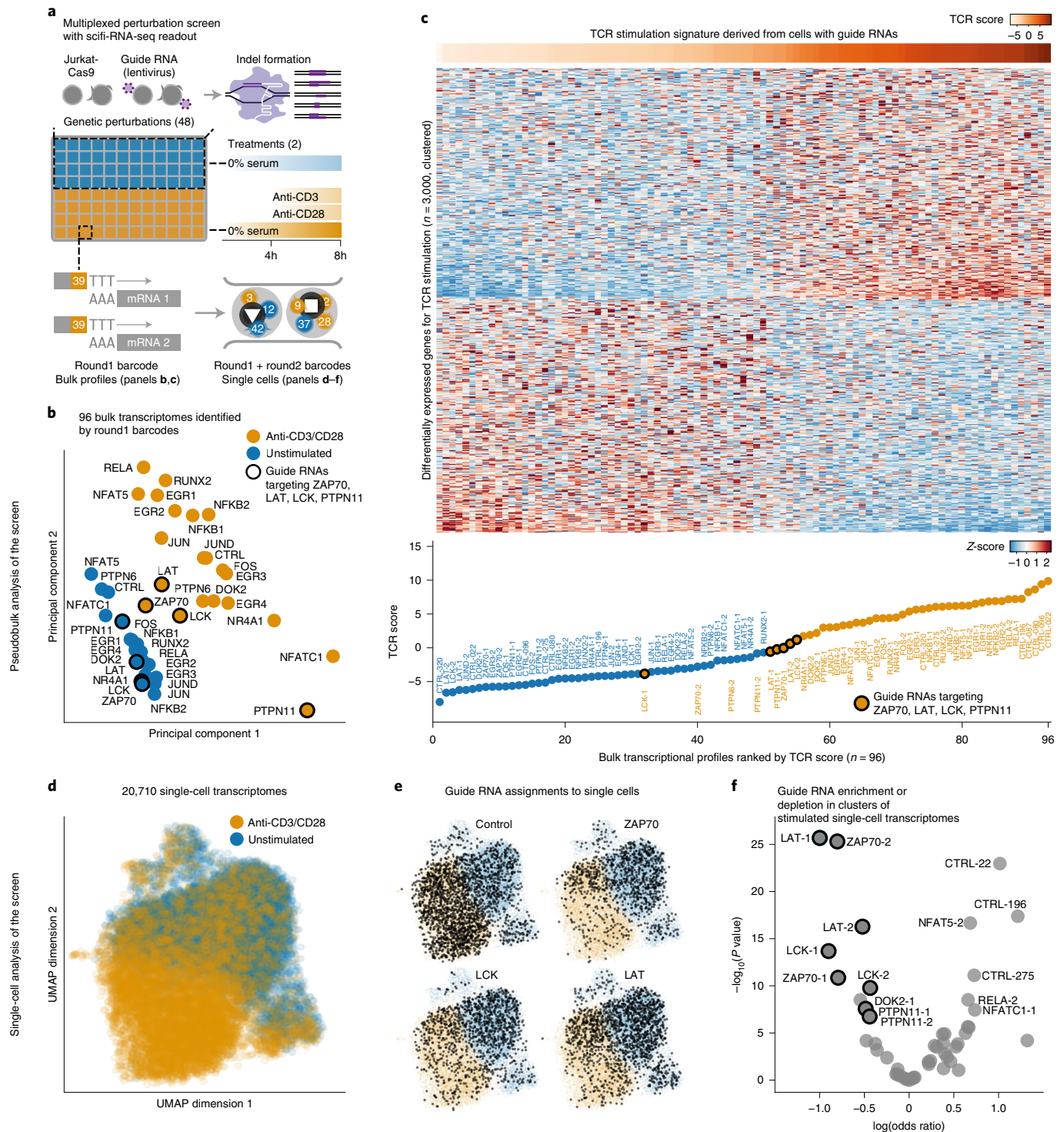


Fig. 3 | scifi-RNA-seq enables large multiplexed perturbation screens with single-cell transcriptome readout. a, Design of an arrayed CRISPR screen with single-cell transcriptome readout for TCR signaling. **b**, PCA of the resulting 96 pseudobulk transcriptomes, aggregating single cells on the basis of the experimental condition. Transcriptomes are colored by TCR stimulation status and labeled with the knockout gene. Key activators of the TCR pathway are highlighted with circles. **c**, Top 3,000 differentially expressed genes between stimulated and unstimulated cells that express control guide RNAs, which were used to assign TCR activation scores to gene knockouts. Samples in the heatmap are sorted by TCR activation score. The lower panel shows the TCR activation score for each sample. **d**, Dimensionality reduction (UMAP) for the single-cell transcriptomes of the CRISPR screen, colored by their TCR stimulation status. **e**, Cells assigned to control guide RNAs (top left) and guide RNAs targeting ZAP70, LCK and LAT are marked in black. **f**, Enrichment of guide RNAs in the stimulated versus unstimulated cluster of single cells. Guide RNAs targeting ZAP70, LAT, LCK and PTPN11 are highlighted with circles.

These results show that scifi-RNA-seq can be used with primary cells and is able to detect biologically meaningful transcription-regulatory differences between cell states.

scifi-RNA-seq enables single-cell perturbation screens. The preindexing step not only enables massive droplet overloading with subsequent resolution of doublets, but also allows us to com-

bine many samples in a single large experiment. To illustrate the opportunities that arise from the built-in multiplexing capability of scifi-RNA-seq, we performed an arrayed CRISPR screen with single-cell transcriptome readout, studying key regulators of TCR activation. Jurkat cells expressing Cas9 were transduced with lentiviral constructs encoding guide RNAs for 20 target genes downstream of the TCR (two guide RNAs per gene) and eight guide RNAs as negative controls. The genome editing resulted in 48 individual CRISPR knockout cell lines, which we split and assayed in two conditions, with and without TCR stimulation by antibodies against CD3/CD28. The resulting 96 experimental conditions were marked with different preindexing (round1) barcodes and processed in a single scifi-RNA-seq experiment (Fig. 3a).

Analyzing this dataset, we first aggregated all 20,710 single-cell transcriptome profiles on the basis of their preindexing (round1) barcodes, resulting in 96 pseudobulk transcriptomes (Fig. 3b). Across the entire dataset, we observed characteristic differences in gene expression driven by TCR stimulation. However, knockouts for known activators of TCR signaling retained transcription profiles after stimulation that fell between those of stimulated and unstimulated control samples. This observation was most pronounced for the kinases ZAP70 and LCK, the adapter protein LAT and the phosphatase PTPN11, and indicates that the ability to respond to TCR stimulation was compromised in these knockout cells. To validate this result, we inferred a gene expression signature indicative of TCR stimulation in Jurkat cells from our dataset, and we scored the TCR activation state for each pseudobulk transcriptome (Fig. 3c and Extended Data Fig. 10a). In this analysis, the stimulated knockouts for ZAP70, LAT, LCK and PTPN11 consistently fell between the nonstimulated and stimulated control samples.

To unravel potential mechanisms by which knockouts of ZAP70, LCK, LAT and PTPN11 affect TCR stimulation, we calculated the enrichment or depletion of cells expressing the respective guide RNAs, as an indicator of their effect on cell proliferation. We then correlated this cell proliferation score with the strength of the TCR activation signature in each knockout (Extended Data Fig. 10b). We found that ZAP70, LAT and LCK were compromised in their response to TCR stimulation while retaining close-to-average numbers for the proliferation score. In contrast, PTPN11 showed a much lower proliferation score, indicating that its role in TCR signaling is linked to the proliferative capabilities of Jurkat cells and is qualitatively different from that of ZAP70, LAT and LCK.

To assess the heterogeneity of TCR activation at the single-cell level, we analyzed our data on the basis of the combination of the preindexing (round1) and microfluidic (round2) barcodes. We observed global clustering of single cells by their TCR stimulation status (Fig. 3d). Cells expressing guide RNAs for ZAP70, LAT and LCK were enriched in the unstimulated single-cell cluster (Fig. 3e,f), indicating that the knockout effectively blocked TCR signaling. Our single-cell analysis also detected a small set of cells with guide RNAs targeting ZAP70, LAT or LCK that showed a strong TCR activation signature, likely because no functional knockout was achieved in the corresponding cells. In summary, these experiments illustrate how scifi-RNA-seq enables arrayed perturbation screens with single-cell transcriptome readout for the functional dissection of regulatory processes.

Discussion

We presented the scifi concept and its implementation in the scifi-RNA-seq assay, providing a method for massive-scale scRNA sequencing that is efficient, flexible and easy-to-use. scifi-RNA-seq combines the scalability of combinatorial indexing with the efficiency and ease-of-use of droplet generators. Preindexing on 96-well or 384-well plates enables us to massively overload existing microfluidic chips while still being able to resolve the single-cell transcriptomes of individual cells in the same droplet. We demonstrated

a 15-fold increase in throughput, with scope for even larger scale processing given that we were able to load 1.53 million nuclei into a single channel of the Chromium system. The microfluidic indexing with its high number of barcodes distinguishes our method from existing combinatorial indexing protocols, where at least three (and sometimes up to five) rounds of plate-based indexing are needed to achieve high throughput. With just two rounds of indexing, and only the reverse transcription done inside permeabilized cells or nuclei, scifi-RNA-seq provides an easy and efficient assay for massive-scale transcriptome profiling for a wide range of biological applications. While our method requires access to a microfluidic droplet generator (such as the Chromium system or another suitable device for microfluidic indexing), this is not a major constraint given that the protocol is compatible with fixed cells or nuclei and thus allows the shipment of samples to a sequencing core facility.

The scifi method is conceptually distinct from cell hashing using DNA-labeled antibodies⁴, lipids⁵ or expressed genetic barcodes⁶, which can flag and discard cell doublets during data analysis but cannot be used to reconstruct single-cell transcriptomes from overloaded droplets. In contrast, scifi-RNA-seq can resolve and retain single-cell transcriptomes because all transcripts of a cell receive cell-specific indices, rather than just a small number of barcoding transcripts. Therefore, scifi-RNA-seq makes it possible to overload microfluidics chips to a point where most droplets contain several cells, which provides a large gain in throughput and cost effectiveness over cell hashing. Moreover, scifi-RNA-seq has built-in support for sample multiplexing through its use of different reverse transcription primers in the preindexing step. This does not require specialized reagents such as hashing antibodies, and it scales to thousands of samples simply by adding further preindexing (round1) barcodes in the form of cheap and widely available oligonucleotides.

We benchmarked scifi-RNA-seq and found that it outperformed published multiround combinatorial indexing protocols. While neither scifi-RNA-seq nor multiround combinatorial indexing reached the high quality of droplet-based scRNA-seq using fresh intact cells, scifi-RNA-seq provided practical advantages (fewer cell doublets and reduced noise compared to droplet-based scRNA-seq on fixed cells or nuclei) and large improvements in throughput and cost effectiveness. We expect scifi-RNA-seq to be most immediately useful for four application areas: (1) Cell atlas projects pursuing the single-cell characterization of complex tissues, organs or entire organisms^{19,20}; (2) large-scale investigation of single-cell transcriptomes in cohorts of patients and/or healthy individuals^{21,22}; (3) CRISPR screens, including arrayed screens (as demonstrated in this study) and pooled screens with single-cell transcriptome readout (for example, using CROP-seq²³, Perturb-seq^{24–26} or CRISP-seq²⁷) and (4) high-throughput perturbation experiments, including drug screens with single-cell transcriptome readout^{18,28}.

While our initial demonstration of scifi-RNA-seq focuses on the widely available 10x Genomics Chromium system, we expect that scifi-RNA-seq will be straightforward to adapt to other microfluidic droplet generators and to alternative methods for scRNA-seq, including subnanoliter well plates. It is also possible to combine scifi-RNA-seq with target enrichment strategies, for example, amplifying target genes for a gene signature of interest, guide RNAs from CRISPR screens and barcoded antibodies for protein detection. More conceptually, scifi technology is not limited to scRNA-seq, but could also be used to boost the throughput of other single-cell assays such as single-cell ATAC-seq^{29,30}, whole genome sequencing³¹, DNA-methylation analysis³² and chromatin conformation capture by Hi-C³³. Our thermoligation barcoding reaction, which we optimized for use inside emulsion droplets, provides a versatile enzymatic approach to support these and other applications. For example, scifi multiomics assays may combine single-cell RNA-seq with single-cell ATAC-seq and single-cell protein profiling in the

same cell^{34–36}, with the future perspective of ultra-high-throughput assays for multiomics analysis of single cells³⁷.

In conclusion, we expect the scifi concept and its implementation in the scifi-RNA-seq assay to provide an easy-to-use, cost-effective and broadly useful method for those applications in single-cell biology that profit from profiling millions of cells or hundreds to thousands of samples in a single experiment.

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/s41592-021-01153-z>.

Received: 8 December 2019; Accepted: 8 April 2021;

Published online: 31 May 2021

References

- Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
- Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
- Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
- McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
- Guo, C. et al. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.* **20**, 90 (2019).
- Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
- Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
- Heaton, H. et al. Souporecell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
- Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
- Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
- Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
- Christina Fan, H., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
- Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
- Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
- Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
- Shum, E. Y., Walczak, E. M., Chang, C. & Christina Fan, H. Quantitation of mRNA transcripts and proteins using the BD Rhapsody single-cell analysis system. *Adv. Exp. Med. Biol.* **1129**, 63–79 (2019).
- Srivatsan, S. R. et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45–51 (2020).
- Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
- Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
- van der Wijst, M. et al. The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
- Rozenblatt-Rosen, O. et al. The Human Tumor Atlas Network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).
- Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
- Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 (2016).
- Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
- Replogle, J. M. et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).
- Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896.e15 (2016).
- McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
- Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
- Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
- Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
- Mulqueen, R. M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
- Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
- Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
- Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
- Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
- Bock, C., Farlik, M. & Sheffield, N. C. Multi-omics of single cells: strategies and applications. *Trends Biotechnol.* **34**, 605–608 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Step-by-step protocol. A step-by-step experimental protocol for scifi-RNA-seq is included as a Supplementary Protocol. Future updates will be shared via the following website: <http://scifi-rna-seq.bocklab.org>.

Measurement of bead fill rates for Chromium Single Cell ATAC chips. The Chromium Single Cell ATAC v1.0 E Chip (10x Genomics catalog no. 2000121) was loaded with 80 μ l of 1 \times Nuclei Buffer (10x Genomics catalog no. 2000153) into inlet 1, 40 μ l of Single Cell ATAC Gel Beads (10x Genomics catalog no. 2000132) into inlet 2 and 240 μ l of Partitioning Oil (10x Genomics catalog no. 220088) into inlet 3. The Chromium Single Cell ATAC v1.1 Next GEM H Chip (10x Genomics catalog no. 2000180) was loaded with 70 μ l of 1 \times Nuclei Buffer (10x Genomics catalog no. 2000207) into inlet 1, 50 μ l of Single Cell ATAC Gel Beads v1.1 (10x Genomics catalog no. 2000210) into inlet 2 and 40 μ l of Partitioning Oil (10x Genomics catalog no. 2000190) into the outlet, labeled as row 3. Because we omitted Reducing Agent B, the gel beads remained intact throughout the microfluidic run, such that they could be visualized inside the emulsion droplets using a standard light microscope. Our fill rate calculations are based on a total of 1,265 (scATAC v1.0) or 1,610 (scATAC v1.1 Next GEM) droplets that were manually counted from microscopic images.

Testing of the nuclei loading capacity for Chromium Single Cell ATAC chips. Human Jurkat cells (clone E6-1) were cultured in RPMI medium (Gibco catalog no. 21875-034) supplemented with 10% FCS (Sigma) and penicillin-streptomycin (Gibco catalog no. 15140122). Fresh nuclei were isolated as described below. We prepared samples of 15,300, 191,000, 383,000, 765,000 and 1,530,000 nuclei. For the Chromium Single Cell ATAC v1.0 E Chip, we added 1 \times Nuclei Buffer (10x Genomics catalog no. 2000153) to a total volume of 78.5 μ l, followed by 1.5 μ l of Reducing Agent B (10x Genomics catalog no. 2000087). For the Chromium Single Cell ATAC v1.1 Next GEM H chip, we added 1 \times Nuclei Buffer (10x Genomics catalog no. 2000207) to a total volume of 73.5 μ l, followed by 1.5 μ l of Reducing Agent B. This buffer did not contain detergents, hence the nuclei remained intact during the microfluidic run and could be visualized inside the emulsion droplets with a standard light microscope. At the same time, Reducing Agent B dissolves the gel beads, which might otherwise obstruct the view. The Single Cell E Chip (10x Genomics catalog no. 2000121) was loaded as follows: 75 μ l of nuclei suspension at the indicated loading concentrations into inlet 1, 40 μ l of Single Cell ATAC Gel Beads (10x Genomics catalog no. 2000132) into inlet 2 and 240 μ l of Partitioning Oil (10x Genomics catalog no. 220088) into inlet 3. The Single Cell H Chip (10x Genomics catalog no. 2000180) was loaded as follows: 70 μ l of nuclei suspension at the indicated loading concentrations into inlet 1, 50 μ l of Single Cell ATAC Gel Beads v1.1 (10x Genomics catalog no. 2000210) into inlet 2 and 40 μ l of Partitioning Oil (10x Genomics catalog no. 2000190) into the outlet, labeled as row 3. To image the resulting droplets, 15 μ l of Partitioning Oil was pipetted onto a glass slide, followed by 5 μ l of emulsion droplets, and images were taken at $\times 10$ magnification. Nuclei were counted manually from microscopic images for an average of 653 (scATAC v1.0) or 902 (scATAC v1.1) droplets per loading concentration. To compare the droplet diameter between v1.0 and v1.1 kits and across loading concentrations, microscopic images from nuclei overloading experiments on the Chromium scATAC v1.0 and v1.1 (Next GEM) platforms were processed with the ImageJ v1.53a software. The droplet diameter was measured in pixels with a straight-line segment and converted to micrometers on the basis of our microscope and camera setup. We measured droplet diameters for 100 droplets per loading condition (ten droplets on ten images), which sums to 500 analyzed droplets per chip design.

Sample preparation and permeabilization for scifi-RNA-seq. *Preparation of permeabilized cell suspension.* A total of 5 million cells were washed with 10 ml of ice-cold 1 \times PBS (Gibco catalog no. 14190-094) with centrifugation (300g, 5 min, 4°C) and fixed in 5 ml of ice-cold methanol (Fisher Scientific catalog no. M/4000/17) at -20°C for 10 min. After two further washes (centrifugation: 300g, 5 min, 4°C) with 5 ml of ice-cold PBS-BSA-SUPERase (1 \times PBS supplemented with 1% w/v BSA (Sigma catalog no. A8806-5) and 1% v/v SUPERase-In RNase Inhibitor (Thermo Fisher Scientific catalog no. AM2696)), permeabilized cells were resuspended in 200 μ l of ice-cold PBS-BSA-SUPERase, and filtered through a cell strainer (40 μ m or 70 μ m depending on the cell size). We then used 5 μ l of the sample for cell counting in duplicates on a CASY device (Schärfe System) and diluted to 5,000 cells μl^{-1} with ice-cold PBS-BSA-SUPERase. We immediately proceeded with the reverse transcription step of scifi-RNA-seq.

Preparation of fresh nuclei suspension. A total of 5 million cells were washed with 10 ml of ice-cold 1 \times PBS (Gibco catalog no. 14190-094) with centrifugation (300g, 5 min, 4°C). Nuclei were prepared by resuspending cells in 500 μ l of ice-cold Nuclei Preparation Buffer (10 mM Tris-HCl pH 7.5 (Sigma catalog no. T2944-100ML), 10 mM NaCl (Sigma catalog no. S5150-1L), 3 mM MgCl₂ (Ambion catalog no. AM9530G), 1% w/v BSA (Sigma catalog no. A8806-5), 1% v/v SUPERase-In RNase Inhibitor (Thermo Fisher Scientific catalog no. AM2696), 0.1% v/v Tween-20 (Sigma catalog no. P7949-500ML), 0.1% v/v IGEAL CA-630 (Sigma catalog no. I8896-50ML) and 0.01% v/v Digitonin (Promega catalog no. G944A)), followed by

5 min of incubation on ice. Lysis of the plasma membrane was stopped by adding 5 ml of ice-cold Nuclei Wash Buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl₂, 1% w/v BSA, 1% v/v SUPERase-In RNase Inhibitor and 0.1% v/v Tween-20). Nuclei were collected by centrifugation (500g, 5 min, 4°C), resuspended in 200 μ l of ice-cold PBS-BSA-SUPERase (1 \times PBS supplemented with 1% w/v BSA and 1% v/v SUPERase-In RNase Inhibitor) and filtered through a cell strainer (40 μ m or 70 μ m depending on the cell size). We then used 5 μ l of the sample for cell counting in duplicates on a CASY device (Schärfe System) and diluted to 5,000 nuclei μl^{-1} with ice-cold PBS-BSA-SUPERase. We immediately proceeded with the reverse transcription step of scifi-RNA-seq.

Preparation of fixed nuclei suspension. A total of 5 million cells were washed with 10 ml of ice-cold 1 \times PBS (Gibco catalog no. 14190-094) with centrifugation (300g, 5 min, 4°C). Nuclei were prepared by resuspending cells in 500 μ l of ice-cold Nuclei Preparation Buffer without Digitonin and without Tween-20 (10 mM Tris-HCl pH 7.5 (Sigma catalog no. T2944-100ML), 10 mM NaCl (Sigma catalog no. S5150-1L), 3 mM MgCl₂ (Ambion catalog no. AM9530G), 1% w/v BSA (Sigma catalog no. A8806-5), 1% v/v SUPERase-In RNase Inhibitor (Thermo Fisher Scientific catalog no. AM2696) and 0.1% v/v IGEAL CA-630 (Sigma catalog no. I8896-50ML)), followed by 5 min of incubation on ice. Lysis of the plasma membrane was stopped by addition of 5 ml of Nuclei Wash Buffer without Tween-20 (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM MgCl₂, 1% w/v BSA, 1% v/v SUPERase-In RNase Inhibitor). Nuclei were collected by centrifugation (500g, 5 min, 4°C), and fixed in 5 ml of ice-cold 1 \times PBS containing 1–4% formaldehyde (Thermo Fisher Scientific catalog no. 28908) for 15 min on ice. Fixed nuclei were collected (500g, 5 min, 4°C), the pellet was resuspended in 1.5 ml of ice-cold Nuclei Wash Buffer without Tween-20 and transferred to a 1.5 ml tube. After one more wash with 1.5 ml of ice-cold Nuclei Wash Buffer without Tween-20 (500g, 5 min, 4°C), fixed nuclei were resuspended in 200 μ l of Nuclei Wash Buffer without Tween-20, snap-frozen in liquid nitrogen and stored at -80°C .

For scifi-RNA-seq, frozen samples were thawed in a 37°C water bath for exactly 1 min, and immediately placed on ice. Following centrifugation (500g, 5 min, 4°C), fixed nuclei were resuspended in 250 μ l of ice-cold Permeabilization Buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl₂, 1% w/v BSA, 1% v/v SUPERase-In RNase Inhibitor, 0.01% v/v Digitonin (Promega catalog no. G944A), 0.1% v/v Tween-20 (Sigma catalog no. P7949-500ML)). After 5 min of incubation on ice, 250 μ l of Nuclei Wash Buffer without Tween-20 was added per sample, and nuclei were collected (500g, 5 min, 4°C). After one more wash with 250 μ l of Nuclei Wash Buffer without Tween-20, nuclei were taken up in 100 μ l of PBS-BSA-SUPERase (1 \times PBS containing 1% w/v BSA and 1% v/v SUPERase-In RNase Inhibitor). We used 5 μ l for cell counting in duplicates on a CASY device (Schärfe Systems) and diluted to 5,000 nuclei μl^{-1} with PBS-BSA-SUPERase. We immediately proceeded with the reverse transcription step of scifi-RNA-seq.

Benchmarking experiments for scifi-RNA-seq versus 10x Genomics Chromium profiling. K562, Jurkat and NALM-6 cells were cultured in RPMI medium (Gibco catalog no. 21875-034) supplemented with 10% FCS (Sigma) and penicillin-streptomycin. HEK293T and 3T3 cells were cultured in DMEM (Gibco catalog no. 10569010) with 10% FCS (Sigma) and penicillin-streptomycin. Nuclei and methanol-fixed whole cells were prepared separately for each of the cell lines, as described above. Cell suspensions were quantified on a CASY device (Schärfe System), and further diluted using PBS-BSA-SUPERase (1 \times PBS supplemented with 1% w/v BSA (Sigma catalog no. A8806-5) and 1% v/v SUPERase-In RNase Inhibitor (Thermo Fisher Scientific catalog no. AM2696)).

For scifi-RNA-seq, different cell lines were processed in an equal number of reverse transcription wells, containing unique sets of round1 indices. For 10x Genomics Chromium processing, different cell lines were mixed at equal proportions before the run. This resulted in two test samples: a mixture of four human cell lines (HEK293T, Jurkat, K562 and NALM-6) and a cross-species mixture (human Jurkat cells and mouse 3T3 cells). Both test samples were processed in parallel with scifi-RNA-seq and with the 10x Genomics Chromium Single Cell Gene Expression assay, using the newest v.3.1 (Next GEM) protocol. To allow for a direct comparison between platforms, we loaded a standardized number of 7,500 cells or nuclei per microfluidic channel.

Preparation of primary human T cell samples for scifi-RNA-seq. *Isolation of primary human T cells.* Peripheral blood from healthy donors was obtained from the Austrian Red Cross as blood packs with buffered sodium citrate as anticoagulant. The study complied with all relevant ethical regulations for working with human primary samples. Informed consent was obtained from all sample donors. The use of blood samples was approved by the ethical committees of the Austrian Red Cross and the Medical University of Vienna. For each donor, we prepared T cells from 3 \times 15 ml of peripheral blood, according to the following protocol: 15 ml of peripheral blood was mixed with 750 μ l of RosetteSep Human T Cell Enrichment Cocktail (Stemcell catalog no. 15061). After 10 min of incubation at room temperature, the sample was diluted by addition of 15 ml 1 \times PBS (Gibco catalog no. 14190-094) containing 2% v/v FCS (Sigma). SepMate tubes (Stemcell catalog no. 86450) were loaded with 15 ml of Lymphoprep density gradient medium (Stemcell catalog no. 07851) and the blood sample was poured

on top. After centrifugation (1,200g, 10 min, room temperature, brake set to 9/9), the supernatant was transferred to a fresh 50 ml tube, topped up to 50 ml with 1× PBS containing 2% FCS, and centrifuged (1,200g, 10 min, room temperature, brake set to 3/9). After one further wash with 50 ml of 1× PBS containing 2% FCS (1,200g, 10 min, room temperature, brake set to 3/9), T cells were resuspended in 10 ml of 1× PBS containing 2% FCS, filtered through a 40 µm cell strainer and counted using a CASY device (Schärfe Systems). For accurate cell counting, it was important to exclude contaminating erythrocytes, given that these cells are lysed during the subsequent nuclei preparation.

Anti-CD3/CD28 stimulation of human T cells. Freshly isolated primary human T cells were resuspended at a density of 1 million cells ml⁻¹ in Human T Cell Medium (OpTmizer medium (Thermo Fisher catalog no. A1048501) containing 1/38.5 volumes of OpTmizer supplement, 1× GlutaMax (Thermo Fisher catalog no. 35050061), 1× Penicillin/Streptomycin (Thermo Fisher catalog no. 15140122), 2% heat-inactivated human AB serum (Fisher Scientific catalog no. MT35060CI) and 10 ng ml⁻¹ of recombinant human IL-2 (PeproTech catalog no. 200-02)). The culture was split into two flasks, and one was treated with Human T-Activator CD3/CD28 Dynabeads (25 µl beads per 1 million cells, Thermo Fisher catalog no. 11131D). After 16 h, we prepared formaldehyde-fixed nuclei and snap-froze the nuclei suspension as described above.

Arrayed CRISPR screening with scifi-RNA-seq readout for T cell receptor activation. *Cloning of guide RNA constructs and lentivirus production.* We cloned 48 guide RNA cassettes into the CROP-seq-Guide-Puro plasmid (Addgene catalog no. 86708) as previously described²³, targeting 20 genes of the T cell receptor pathway with two guide RNAs each and including eight nontargeting controls. All constructs were verified by Sanger sequencing using the following primer: 5'-TTGGGCACTGACAATTCGGT-3'. For lentivirus production on 96-well plates, we seeded Lenti-X HEK293T cells (Takara catalog no. 632180) at 50,000 cells per well in 200 µl of lentivirus packaging medium (Opti-MEM I (Gibco catalog no. 51985-034), 5% FCS (Sigma) and 200 mM sodium pyruvate (Gibco catalog no. 11360-070), without antibiotics). Following overnight incubation at 37°C and 5% CO₂, cells were transfected with Lipofectamine 3000 reagents the next morning (Thermo Fisher catalog no. L3000015). We first distributed transfection mix A into a 96-well plate (per reaction: 25 µl Opti-MEM I, 0.6 µl P3000 Enhancer Reagent and 90 ng each of the three packaging plasmids pMDLg/pRRE (Addgene catalog no. 12251), pRSV-Rev (Addgene catalog no. 12253) and pMD2.G (Addgene catalog no. 12259)). We then added 170 ng of transfer plasmid encoding a single guide RNA to each well. Next, we added transfection mix B (per reaction: 25 µl Opti-MEM I and 0.7 µl Lipofectamine 3000 Reagent), mixed gently by pipetting and allowed complex formation for 20 min at room temperature. We removed 100 µl of the culture medium from the Lenti-X HEK293T cells and transfected them with 50 µl of freshly prepared lipid complexes. To avoid toxicity to the producer cells, we exchanged the medium for fresh lentivirus production medium after 6 h. The supernatant containing lentiviral particles was collected after 24 and 48 h and combined in a sterile deepwell plate (Corning catalog no. 07-200-700). Finally, producer cells were removed by filtration (Merck catalog no. MSHVS4510).

Production of single-gene knockout cell lines. Human Jurkat cells (clone E6-1) were transduced with lentiCas9-Blast (Addgene catalog no. 52962) and selected with 25 µg ml⁻¹ of blasticidin (InvivoGen catalog no. ant-bl-5) to achieve a stable expression of Cas9. Jurkat-Cas9 cells were seeded onto a 96-well plate at 250,000 cells per well in 100 µl of complete culture medium (RPMI (Gibco catalog no. 21875-034), 10% FCS (Sigma), penicillin-streptomycin) containing 25 µg ml⁻¹ blasticidin to maintain the selection pressure for Cas9 and 8 µg ml⁻¹ polybrene (Sigma Aldrich catalog no. H9268-5G) to facilitate the interaction of viral particles with the cells. For the transduction, we added 10 µl of lentiviral particles, produced as described above, and performed spinfection at 37°C and 1,200g for 45 min. Following overnight incubation at 37°C and 5% CO₂, the medium was exchanged to select for the guide RNA construct (RPMI, 10% FCS, penicillin-streptomycin, 25 µg ml⁻¹ blasticidin, 2 µg ml⁻¹ puromycin (Thermo Fisher Scientific catalog no. A1113803)). The selective medium was renewed every 2–3 days for a duration of 10 days to allow for efficient genome editing, while maintaining the cells in a 96-well plate.

Arrayed CRISPR screen with scifi-RNA-seq readout. We produced 48 single-gene knockout cell lines by arrayed CRISPR editing as described above. The knockout cells were split into two equal parts while transferring to a 96-well V-bottom plate, pelleted by centrifugation at 300g for 5 min, and washed once with 200 µl of starvation medium (RPMI (Gibco catalog no. 21875-034), 0% FCS, penicillin-streptomycin). Cell pellets were resuspended in 200 µl of starvation medium. For T cell receptor activation, we coated one half of a flat-bottom 96-well plate with 50 µl per well of a 10 µg ml⁻¹ solution of anti-human CD3 antibody in 1× PBS (clone OKT3, Thermo Fisher Scientific catalog no. 16-0037-81). The other half of the wells was left untreated, adding only 50 µl of 1× PBS. After 2 h of surface coating at 37°C, the antibody solution was removed, and the coated plate was washed twice with 200 µl of 1× PBS. Per well, 180 µl of cells in serum-starvation medium was added, followed by either 20 µl of a 20 µg ml⁻¹ solution of anti-human CD28 antibody diluted in starvation medium for stimulated wells (clone CD28.2,

Thermo Fisher Scientific catalog no. 16-0289-81) or 20 µl of starvation medium for unstimulated cells. Cells were incubated for 4 h to allow for efficient TCR activation in antibody-treated wells.

To permeabilize cells for scifi-RNA-seq, we performed methanol fixation of whole cells in a 96-well format. Cells were transferred to a V-bottom plate and washed once with 150 µl of 1× PBS. Cell pellets were collected by centrifugation at 300g for 5 min, PBS was completely removed, and cells were resuspended in 150 µl of ice-cold methanol. The plate was covered with an aluminum seal and stored at -20°C for 10 min. Afterward, cells were washed twice with 150 µl of ice-cold PBS-BSA-SUPERase (1× PBS supplemented with 1% w/v BSA (Sigma catalog no. A8806-5) and 1% v/v SUPERase-In RNase Inhibitor (Thermo Fisher Scientific catalog no. AM2696)). Cell pellets were taken up in 12 µl of PBS-BSA-SUPERase, and 2 µl of the cell suspension was transferred to each of four wells on a 384-well plate for scifi-RNA-seq processing. Thus, each of the 96 experimental conditions (48 genetic perturbations and two treatments) was barcoded with a set of four specific round1 reverse transcription barcodes.

Detailed scifi-RNA-seq protocol description. This section provides an overview of the scifi-RNA-seq method and the enzymatic steps.

Reverse transcription. A total of 384 indexed reverse transcription primers were synthesized by Sigma Aldrich and obtained at 100 µM concentration in EB Buffer on 96-well plates (sequences are provided in Supplementary Table 1). Then 384-well plates with barcoded oligo-dT primers were prepared before the experiment and stored at -20°C (1 µl of 25 µM per well). A total of 10,000 permeabilized cells or nuclei (2 µl of a 5,000 µl⁻¹ suspension) were added to the predesigned primers in each well, and the assignment of samples to wells was recorded. The plate was incubated for 5 min at 55°C to resolve RNA secondary structures, then placed immediately on ice to prevent their reformation. Per well, a mix of 3 µl nuclease-free water, 2 µl 5× Reverse Transcription Buffer, 0.5 µl of 100 mM dithiothreitol (DTT; freshly diluted from Sigma catalog no. 646563-10x.5ML), 0.5 µl of 10 mM dNTPs (Thermo Fisher Scientific catalog no. R0193), 0.5 µl of RNaseOUT RNase inhibitor (40 U ml⁻¹, Thermo Fisher Scientific catalog no. 10777019), and 0.5 µl of Maxima H Minus Reverse Transcriptase (200 U ml⁻¹, Thermo Fisher Scientific catalog no. EP0753) were added. The reverse transcription was incubated as follows (with heated lid set to 60°C): 50°C for 10 min, 3 cycles of [8°C for 12 s, 15°C for 45 s, 20°C for 45 s, 30°C for 30 s, 42°C for 2 min, 50°C for 3 min], 50°C for 5 min, store at 4°C.

Cell/nuclei recovery and pooling. Processed cells or nuclei were recovered from the 384-well plate and pooled in one 15 ml tube per plate on ice. Wells were washed with ice-cold 1× PBS containing 1% BSA, which was transferred to the same tube for maximum recovery. The volume was topped up to 15 ml with 1× PBS containing 1% BSA, and nuclei were collected (500g, 5 min, 4°C). The resulting pellet was resuspended in 1.0 ml of 1× Ampligase Reaction Buffer (Lucigen catalog no. A0102K), filtered through a cell strainer (40 µm or 70 µm depending on the cell/nuclei size) into a 1.5 ml tube, and centrifuged (500g, 5 min, 4°C). The supernatant was removed almost completely, and the tube was centrifuged briefly (500g, 30 s, 4°C) to collect the remaining liquid at the bottom of the tube. Typically, this resulted in ~10 µl of a highly concentrated suspension. For cell counting, the concentrated sample was diluted 1:200 with 1× Ampligase Buffer and loaded into a Fuchs-Rosenthal counting chamber (Incyto catalog no. DHC-F01). The desired number of cells/nuclei was brought to a volume of 15 µl with 1× Ampligase Reaction Buffer (Lucigen catalog no. A0102K).

Microfluidic thermoligation barcoding on the Chromium scATAC v1.0 platform. Unused channels in the Chromium Chip E (10x Genomics catalog no. 2000121) were filled with 75 µl (inlet 1), 40 µl (inlet 2) and 240 µl (inlet 3) of 50% glycerol solution (Sigma catalog no. G5516-100ML). Immediately before loading the chip, a mix of 47.4 µl nuclease-free water, 11.5 µl of 10× Ampligase Reaction Buffer (Lucigen catalog no. A0102K), 2.3 µl of 100 U µl⁻¹ Ampligase (Lucigen catalog no. A0102K), 1.5 µl of Reducing Agent B (10x Genomics catalog no. 2000087) and 2.3 µl of 100 µM Bridge Oligo (sequence provided in Supplementary Table 1) was added to each sample. The microfluidic chip was loaded with 75 µl of cells or nuclei in thermoligation mix (inlet 1), 40 µl of Single Cell ATAC Gel Beads (inlet 2, 10x Genomics catalog no. 2000132) and 240 µl of Partitioning Oil (inlet 3, 10x Genomics catalog no. 220088) and run on the Chromium system.

Microfluidic thermoligation barcoding on the Chromium scATAC v1.1 (Next GEM) platform. On the Chromium Next GEM Chip H (10x Genomics catalog no. 2000180), unused channels were loaded with 70 µl (row 1), 50 µl (row 2) and 40 µl (row 3) of 50% glycerol solution (Sigma catalog no. G5516-100ML). Per sample, we added a mix of 41.9 µl of nuclease-free water, 12 µl of 10× Ampligase Reaction Buffer (Lucigen catalog no. A0102K), 2.3 µl of 100 U µl⁻¹ Ampligase (Lucigen catalog no. A0102K), 1.5 µl of Reducing Agent B (10x Genomics catalog no. 2000087) and 2.3 µl of 100 µM Bridge Oligo (sequence provided in Supplementary Table 1) immediately before the run. Seventy microliters of cell suspension in thermoligation master mix was loaded into the Chromium Next GEM Chip H (row 1), along with 50 µl of Single Cell ATAC Gel Beads v1.1 (row 2, 10x Genomics

catalog no. 2000210) and 40 µl of Partitioning Oil (row 3, 10x Genomics catalog no. 2000190), and the chip was run on the Chromium system.

GEM incubation and cleanup. For thermoligation barcoding, the droplet emulsion was incubated as follows (heated lid set to 105 °C, volume set to 100 µl): 12 cycles of (98 °C for 30 s, 59 °C for 2 min), storage at 15 °C. The emulsion was broken by addition of 125 µl Recovery Agent (10x Genomics catalog no. 220016) and 125 µl of the pink oil phase was removed by pipetting. The remaining sample was mixed with 200 µl of Dynabead Cleanup Master Mix (per reaction: 182 µl of Cleanup Buffer (10x Genomics catalog no. 2000088), 8 µl of Dynabeads MyOne Silane (Thermo Fisher Scientific catalog no. 37002D), 5 µl of Reducing Agent B (10x Genomics catalog no. 2000087) and 5 µl of nuclease-free water). After 10 min of incubation at room temperature, samples were washed twice with 200 µl of freshly prepared 80% ethanol (Merck catalog no. 603-002-00-5) and eluted in 40.5 µl of EB Buffer (Qiagen catalog no. 19086) containing 0.1% Tween (Sigma catalog no. P7949-500ML) and 1% v/v Reducing Agent B. Bead clumps were sheared with a 10 µl pipette or needle. 40 µl of the sample was transferred to a fresh tube strip and subjected to a 1.0× cleanup with SPRIselect beads (Beckman Coulter catalog no. B23318), eluting in 22 µl of EB Buffer.

Template switching. Twenty microliters of sample from the previous step was mixed with 10 µl of 5× Reverse Transcription Buffer, 10 µl of Ficoll PM-400 (20%, Sigma catalog no. F5415-50ML), 5 µl of 10 mM dNTPs (Thermo Fisher Scientific catalog no. R0193), 1.25 µl of Recombinant Ribonuclease Inhibitor (Takara catalog no. 2313A), 1.25 µl of 100 µM Template Switching Oligo (sequence provided in Supplementary Table 1) and 2.5 µl of Maxima H Minus Reverse Transcriptase (200 U ml⁻¹, Thermo Fisher Scientific catalog no. EP0753). The template switching reaction was incubated for 30 min at 25 °C, 90 min at 42 °C, storage at 4 °C and cleaned with a 1.0× SPRI cleanup, eluting in 17 µl of EB buffer.

cDNA enrichment. Fifteen microliters of sample from the step above was mixed with 33 µl of nuclease-free water, 50 µl of NEBNext High-Fidelity 2× PCR Master Mix (NEB catalog no. M0541S), 0.5 µl of 100 µM Partial-P5 primer, 0.5 µl of 100 µM TSO Enrichment Primer (sequences provided in Supplementary Table 1), and 1 µl of 100× SYBR Green in DMSO (Life Technologies catalog no. S7563). cDNA was amplified in a thermocycler as follows: 98 °C for 30 s, cycle until fluorescent signal >1,000 relative fluorescence units (RFU); 98 °C for 20 s, 65 °C for 30 s, 72 °C for 3 min, 72 °C for 5 min in another thermocycler, storage at 4 °C. cDNA was cleaned by one 0.8× SPRI cleanup followed by a 0.6× SPRI cleanup, quantified with a Qubit HS assay (Thermo Fisher Scientific catalog no. Q32854), and 1.5 ng were checked on a Bioanalyzer High-Sensitivity DNA chip (Agilent catalog no. 5067-4626 and 5067-4627).

Tagmentation. Tn5 reaction buffer was prepared: 50 mM TAPS (Sigma catalog no. T9659-100G), 25 mM MgCl₂ (Ambion catalog no. AM9530G); the pH was adjusted to 8.5 and the solution was sterile filtered. Tn5 dilution buffer was prepared: 50 mM Tris-HCl pH 7.5 (Sigma catalog no. T2944-100ML), 100 mM NaCl (Sigma catalog no. S5150-1L), 0.1 mM EDTA (Invitrogen catalog no. AM9260G), 50% glycerol (Sigma catalog no. G5516-100ML), 0.1% Triton-X100 (Sigma catalog no. X100-100ML), and supplemented with fresh 1 mM DTT (Sigma catalog no. 646563-10x.5ML) before use. To achieve maximum library complexity, the entire sample was processed in several tagmentation reactions with 1 ng input each. cDNA was diluted to 0.2 ng µl⁻¹ with nuclease-free water and 5 µl (1 ng) per reaction was distributed into a 96-well plate on ice. A mix of 11.25 µl nuclease-free water, 5 µl of 5× Tn5 reaction buffer, 2.5 µl of dimethylformamide (Sigma catalog no. D4551-250ML), and 1.25 µl of freshly diluted i7-only transposome (prepared as described below and diluted 1:4.5 in Tn5 dilution buffer) was added. Reactions were incubated for 10 min at 55 °C, then cooled for 1 min on ice. The enzyme was inactivated by addition of 2.5 µl of 1% SDS (Sigma catalog no. 71736-100ML) for 5 min at room temperature. Next, the volume was brought to 50 µl and the fragmented cDNA was purified with a 1.0× SPRI cleanup, eluting in 17 µl of EB buffer (Qiagen catalog no. 19086).

Library enrichment. Tagmented cDNA (15 µl) was mixed with 5 µl of 10 µM barcoded P7 primer (sequences provided in Supplementary Table 1). Per reaction, a mix of 28.5 µl nuclease-free water, 50 µl NEBNext High-Fidelity 2× PCR Master Mix (NEB catalog no. M0541S), 1 µl of 100× SYBR Green in DMSO (Life Technologies catalog no. S7563) and 0.5 µl of 100 µM Partial-P5 primer were added. Reactions were incubated in a qPCR cycler as follows: 72 °C for 3 min (for end fill-in after tagmentation), 98 °C for 30 s, cycle (98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s, plate read). We monitored the fluorescence signals and removed samples from the thermocycler when they reached >4,000 relative fluorescence units (RFUs). To complete unfinished PCR products, we incubated for 5 min at 72 °C in another thermocycler. Libraries were cleaned with a 0.7× SPRI cleanup with AMPure XP beads. All wells with the same P7 barcode were pooled and subjected to a 0.8× SPRI cleanup, eluting in 0.2 bead volumes of EB buffer (Qiagen catalog no. 19086). The final library concentration was measured with the Qubit High Sensitivity DNA assay (Thermo Fisher Scientific catalog no. Q32854), and 1.5 ng were run on a Bioanalyzer High Sensitivity DNA chip (Agilent catalog nos. 5067-4626 and 5067-4627).

Sequencing. Libraries were diluted to 2.0 nM with EB buffer (Qiagen catalog no. 19086) containing 0.1% Tween-20 (Sigma catalog no. P7949-500ML) and sequenced on the Illumina NovaSeq 6000 platform with standard sequencing primers and a read structure of 21 bases (Read 1), 8 bases (Index 1, i7), 16 bases (Index 2, i5) and 78 bases (Read 2). Depending on the scale of the experiment, NovaSeq 6000 SP (Illumina catalog no. 20027464), S1 (Illumina catalog no. 20012865) or S2 (Illumina catalog no. 20012862) reagent kits with a nominal number of 100 sequencing cycles (excluding indices) were used.

Assembly and validation of custom i7-only transposome for scifi-RNA-seq.

Oligonucleotides Tn5-top_ME and Tn5-bottom_Read2N were synthesized by Sigma Aldrich (sequences are provided in Supplementary Table 1) and reconstituted in EB buffer (Qiagen catalog no. 19086) at 100 µM. We mixed 22.5 µl of each oligonucleotide with 5 µl of 10× Oligonucleotide Annealing Buffer (10 mM Tris-HCl (Sigma catalog no. T2944-100ML), 500 mM NaCl (Sigma catalog no. S5150-1L), 10 mM EDTA (Invitrogen catalog no. AM9260G)) and annealed them in a thermocycler (95 °C for 3 min, 70 °C for 3 min, ramp to 25 °C at −2 °C min⁻¹). The annealing reaction was then diluted by addition of 180 µl of water. At this point, the diluted oligonucleotide cassette can be aliquoted and frozen for future transposome assemblies. To load the Tn5 transposase, we mixed 20 µl of diluted oligonucleotide cassette from the previous step with 20 µl of 100% glycerol (Sigma catalog no. G5516-100ML) and 10 µl of EZ-Tn5 Transposase (Lucigen catalog no. TNP92110), and incubated for 30 min at 25 °C in a thermocycler. The resulting 50 µl of assembled transposome can be stored at −20 °C for at least 1 month.

Tagmented DNA flanked by two Illumina i7 adapters is suppressed in standard PCR reactions due to competition between intramolecular annealing and primer binding, as described previously³⁸. However, the enzymatic activity of the custom i7-only transposome can still be tested in a negative qPCR assay. Briefly, a defined PCR product was subjected to one tagmentation reaction and one no-enzyme control reaction, and both samples were reamplified with the same primers by qPCR. Since the tagmentation fragments the PCR product, the corresponding reaction should yield higher Ct values. The tagmentation efficiency can then be calculated from the shift of Ct values:

$$\text{Tagmentation efficiency (\%)} = 100 - (100 / (2^{(\text{average Ct tagmentation} - \text{average Ct no-enzyme control})))$$

The PCR product for testing the enzymatic activity was produced as follows. Oligonucleotides pUC19-FWD and pUC19-REV were synthesized by Sigma Aldrich (sequences are provided in Supplementary Table 1) and reconstituted in EB buffer (Qiagen catalog no. 19086) at 100 µM. Next, a 1,961-bp PCR product was generated by mixing 128.7 µl of water, 33 µl of 50 µg µl⁻¹ pUC19 plasmid (NEB catalog no. N3041S), 1.65 µl each of primers pUC19-FWD and pUC19-REV (100 µM) combined with 165 µl of 2× Q5 HotStart High-Fidelity Master Mix (NEB catalog no. M0494L). The resulting 6.6× master mix was distributed into a tube strip (six reactions of 50 µl) and amplified in a thermocycler: 98 °C for 30 s; 31× (98 °C for 10 s, 68 °C for 30 s, 72 °C for 1 min), 72 °C for 2 min and storage at 12 °C. To each 50 µl PCR reaction, we added 6.25 µl of 10× CutSmart Buffer and 6.25 µl of *DpnI* (NEB catalog no. R0176L) and incubated at 37 °C for 1 h to digest the PCR template plasmid. The six PCR reactions were pooled and cleaned with the QiaQuick PCR Purification Kit (Qiagen catalog no. 28106) using two columns and eluting with 30 µl of EB buffer per column. Eluates were pooled, and the purity of the PCR fragment was checked on a 1% agarose gel containing ethidium bromide. We then measured the concentration of dsDNA with a Qubit HS assay (Thermo Fisher Scientific catalog no. Q32854) and diluted the PCR product to 25 ng µl⁻¹ with EB buffer.

The resulting PCR product was used in tagmentation reactions, set up by mixing 2 µl of 25 ng µl⁻¹ of the PCR product, 7 µl of ATAC Buffer (10x Genomics catalog no. 2000122), and either 6 µl of custom i7-only transposome (tagmentation reaction) or 6 µl of water (no-enzyme control reaction). After 60 min of incubation at 37 °C, the Tn5 enzyme was stripped from the DNA by addition of 1.75 µl of 1% SDS solution (Sigma catalog no. 71736-100ML) followed by incubation at 70 °C for 10 min. The two reactions were diluted 1/100 with EB buffer, and qPCR reactions were set up in triplicates: 2 µl of 1/100-diluted reaction, 10 µl of 2× GoTaq qPCR Master Mix (Promega catalog no. A600A), 0.1 µl each of 100 µM pUC19-FWD and pUC19-REV primers and 7.8 µl of water. qPCR reactions were incubated as follows: 95 °C for 2 min, 40× (95 °C for 30 s, 68 °C for 30 s, 72 °C for 2 min, plate reading).

Computational modeling of the cell loading in microfluidic droplet generators.

Estimation of expected barcode collisions in scifi-RNA-seq. To obtain an initial estimate of cell doublets per compartment (droplet and/or well), we used the 'birthday problem' as described by others^{10,25}, where the probability of collision of n cells in c compartments is defined as:

$$p(n, c) \approx 1 - \left(\frac{c-1}{c} \right)^{\frac{n(n-1)}{2}}$$

Corresponding, the number of cells per compartment c at a collision rate p is given by the reverse problem:

$$n(p, c) \approx \sqrt{2c \ln \left(\frac{1}{1-p} \right)}.$$

In this formulation, plate-based preindexing combined with subsequent barcoding of the same cells using a microfluidic droplet generator is equivalent to placing cells in some compartments corresponding to the product of the number of preindexing compartments and the number of droplets with beads in the microfluidic device. These numbers provide an initial indication of the number of usable single-cell transcriptomes in overloaded experiments, although they neglect the higher-order collisions expected when loading high numbers of cells or nuclei.

Monte Carlo simulation of microfluidic device loading. As a complementary approach that makes no assumptions about specific statistical distributions in the scifi-RNA-seq protocol, we performed the following Monte Carlo simulation: We generated two vectors of size n for which we randomly sampled integers from the range of preindexing compartments $r1$ and microfluidic barcodes $r2$. We then calculated the fraction of several occurrences of the same $r1$ integer in the $r2$ compartments. This was repeated for a range of input cell numbers (n), compartments ($r1$ and $r2$), and iterations ($i = 100$).

Distribution-based modeling of microfluidic device loading. The combinatorial estimation of expected barcode collisions, as well as the Monte Carlo simulations, do not adequately consider biological doublets (such as cell clumps) and are thus at risk of underestimating the problem of cell doublets. To gain empirical insights into the loading of the Chromium system, we loaded nuclei without lysis reagents, such that they could be visualized inside the emulsion droplets under a light microscope (as described above). As we observed a fraction of droplets containing no nuclei even at high nuclei loading concentrations, we specified a Bayesian model in which the observed number of nuclei per droplet is explained by a zero-inflated Poisson distribution of latent parameters λ for mean and variance and ψ for the zero-inflated component. We imposed a prior on λ as the observed mean nuclei per droplet and estimated a posterior distribution of parameters by Markov Chain Monte Carlo (MCMC) sampling with the No U-Turn Sampler (NUTS) for 200,000 initializer iterations, 4,000 tuning draws and 4,000 draws to estimate the posterior parameters using PyMC3 (v.3.8)³⁹.

To estimate doublet rates for scifi-RNA-seq experiments, we assumed that the final doublet rate is independent of the preindexing. In this case, for the purpose of doublet estimation, the number of nuclei that can produce doublets is the number of input nuclei divided by the number of unique preindexing barcodes. We used the model parameters estimated from the optically counted numbers of nuclei per droplet to estimate and interpolate model parameters across a range of nuclei loading concentrations. Using those parameters for realistic numbers of input nuclei, the doublet rate is estimated as 1 minus the probability mass function (PMF) of the zero-inflated Poisson distribution at a given nuclei loading concentration. While these estimates become reasonably accurate when actual nuclei counts observed for the microfluidic device are taken into account, the estimates are still optimistic as they use nuclei as the countable unit, whereas in the real scifi-RNA-seq experiment the countable units are transcript molecules, and the final number of nuclei are inferred from the transcript counting procedure.

Processing of scifi-RNA-seq data. For fast parallel processing of raw sequence data, we demultiplexed the raw sequence calls on the basis of the round1 barcode (and optionally a sample barcode if several scifi-RNA-seq experiments are multiplexed for sequencing), allowing up to three mismatches for both barcode sequences together. The demultiplexed reads were written into unaligned single-read BAM files with dedicated tags for the round2 barcode ('r2' tag) and UMI ('RX' tag).

The STAR aligner (v.2.7.0e) was used to map the demultiplexed unaligned BAM files with the following parameters: '-readFilesCommand samtools view -h -readFilesType SAM SE'. We allowed trimming of poly-A stretches by setting the '-clip3pAdapterSeq AAAAAA' parameter and we retained all alignments via the following parameters: '-outFilterScoreMinOverLread 0 -outFilterMatchNminOverLread 0 -outFilterMatchNmin 0'. We set the output of STAR to aligned BAM using the '-outSAMtype BAM Unsorted' parameter.

To quantify gene expression, we used the featureCounts tool from the Subread package (v.1.6.2) to add an 'XT' tag to the aligned BAM file with the respective gene using the '-g gene_id' parameter. Only alignments with quality above 30 were tagged ('-q 30' parameter). We quantified alignments in an unstranded manner ('-s 0' parameter) for either exons only ('-t exon' parameter) or for whole gene bodies ('-t gene' parameter) since nuclei can contain high amounts of unspliced transcripts. We then used Pysam (v.0.15.2) to filter out alignments from low-quality or unmapped reads, from multimappers and from secondary or supplementary alignments. Effectively, we kept only reads which aligned uniquely to genes with an alignment quality score above 30.

To generate the gene expression matrix, we counted transcripts on the basis of reads uniquely identified by the combination of round1 and round2 barcodes, UMI, associated gene and mapping position in the chromosome. We then created a SciPy (v.1.3.0) compressed sparse row matrix of the form (UMI count, cell barcode and gene) and initialized an Anndata (v.0.6.21) object to save it as h5ad format. We used Python v.3.7.2 for all analyses.

Bioinformatic analysis of scifi-RNA-seq data. We used Scanpy (v.1.4.3) for bioinformatic analysis of the single-cell RNA-seq data. For the cell line mixture and T cell experiments, we filtered the Anndata object for cell barcodes containing at least 200 UMI counts, while keeping only those genes that were not in the lower 10th percentile of cumulative UMI counts and that were detected by at least 500 barcodes. We further excluded cells more than three standard deviations away from the mean of the fraction of the transcriptome in ribosomal or mitochondrial genes. Gene expression values were normalized to a sum of 10,000 for each cell, and log transformed. We performed principal component analysis (PCA) on scaled and centered expression values, computed a neighbor graph and performed dimensionality reduction with UMAP using the respective Scanpy functions, all with default parameters. We then used the Leiden algorithm for community detection on the neighbor graph with the 0.5 resolution parameter, and we tested for mean gene expression differences between groups of cells using the normalized and log transformed expression values and the scanpy.tl.rank_genes_groups function with a two-sided t -test with variance overestimation (t -test_overestim_var function) and corrected for multiple testing with the Benjamini–Hochberg procedure. For these differential gene sets between cell lines, we performed gene set enrichment analysis using the Enrichr API.

Comparative analysis of single-cell RNA-seq datasets. The following publicly available combinatorial indexing datasets were downloaded from GEO: (1) sci-RNA-seq: HEK293T, HeLa S3 and NIH/3T3 mixture (GSM2599699/SRX2784959); (2) sci-Plex: HEK293T and NIH/3T3 mixture (GSM4150376/SRX7101186); (3) SPLiT-seq: species mixing with 3,000 cells (GSM3017262/SRX3722699); (4) SPLiT-seq: species mixing with 300 cells (GSM3017263/SRX3722700). Files were converted to BAM format with the following command 'sam-dump {sra_run} | samtools view -bs -> {sra_run}.bam', and the cellular and molecular barcodes were written into BAM format tags to match the scifi-RNA-seq input format (unaligned BAM files of single-end transcriptome reads with additional tags). In-house-generated 10x Genomics Chromium Single-Cell Gene Expression 3' v.3.1 libraries were demultiplexed and stored in unaligned BAM format as for scifi-RNA-seq (but without any round2 barcodes). To maximize compatibility, datasets were processed uniformly with the exact same pipeline as the scifi-RNA-seq datasets, with the exception that cellular barcodes were aggregated dependent on the number of rounds of (combinatorial) indexing.

Analysis of scifi-RNA-seq data for the arrayed CRISPR screen. The scifi-RNA-seq data for the CRISPR screen in T cells were processed in the same way as the other datasets, with the exception that the resolution for Leiden clustering was set to 0.35. We aggregated mean gene expression values by guide RNAs and by genes on the basis of preindexing (round1) barcodes, and we performed dimension reduction of the resulting pseudobulk expression matrix using PCA and UMAP. To quantify the effect of each perturbation on cell proliferation, we performed Fisher's exact tests for the co-occurrence of each guide RNA or gene in the cells labeled as stimulated or unstimulated on the basis of preindexing (round1) barcodes. The rationale was that perturbations with positive or negative effects on cell proliferation or cell survival will skew the representation of the corresponding guide RNAs in the pool of sequenced cells. To quantify the effect on TCR activation, we performed Fisher's exact tests in the same way, but using the labels given by the Leiden clustering. Here, the rationale was that TCR activation, as captured by the Leiden clustering, is the dominant effect on the transcriptome and overrepresentation or underrepresentation of guide RNAs in either cluster indicates a role in TCR signaling. To obtain a numeric score for the position of each guide RNA in the range of TCR activation (as defined by the stimulated and unstimulated cells with control guide RNAs), we computed the difference between the mean square root of the errors for each guide RNA compared with cells with control guide RNAs in both states.

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

Data availability

All raw and processed datasets are available from the NCBI GEO database (GSE168620). A detailed annotation table for these datasets is provided in Supplementary Table 3. Source data are provided with this paper.

Code availability

Two versions of the source code underlying this paper are provided as separate GitHub repositories: (1) analysis code for reproducing the results in this paper (https://github.com/epigen/scifiRNA-seq_publication); (2) pipeline code for processing new scifi-RNA-seq datasets (<https://github.com/epigen/scifiRNA-seq>).

References

- Rykalina, V., Shadrin, A., Lehrach, H. & Borodina, T. qPCR-based characterization of DNA fragmentation efficiency of Tn5 transposomes. *Biol. Methods Protoc.* 2, bpx001 (2017).

39. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* 2, e55 (2016).

Acknowledgements

We thank the team of the Biomedical Sequencing Facility at CeMM for assistance with next-generation sequencing and all members of the Bock laboratory for their help and advice. P.D. would like to thank N. Winhofer and K. Winhofer for their support. This work was conducted in the context of two Austrian Science Fund (FWF) Special Research Program grants (FWF SFB F6102; FWF SFB F7001). T.K. was supported by a Lise Meitner fellowship from the Austrian Science Fund (FWF M2403). C.B. is supported by an ERC Starting Grant (European Union's Horizon 2020 research and innovation program, grant agreement no. 679146).

Author contributions

P.D. conceived and developed scifi barcoding and scifi-RNA-seq. P.D. and T.B. optimized the protocol. P.D. conducted the experiments. A.F.R. performed computational modeling and analyzed the data. P.D. and A.F.R. visualized the results. P.D. prepared the figures. M.S. prepared Chromium libraries and performed next-generation sequencing. T.K. contributed to the human primary T cell experiments. D.B. performed data

preprocessing. P.D., A.F.R. and C.B. wrote the manuscript with contributions from all authors. C.B. supervised the project.

Competing interests

P.D. and C.B. are inventors on a patent application describing scifi barcoding and the scifi-RNA-seq method. The other authors declare no competing interests.

Additional information

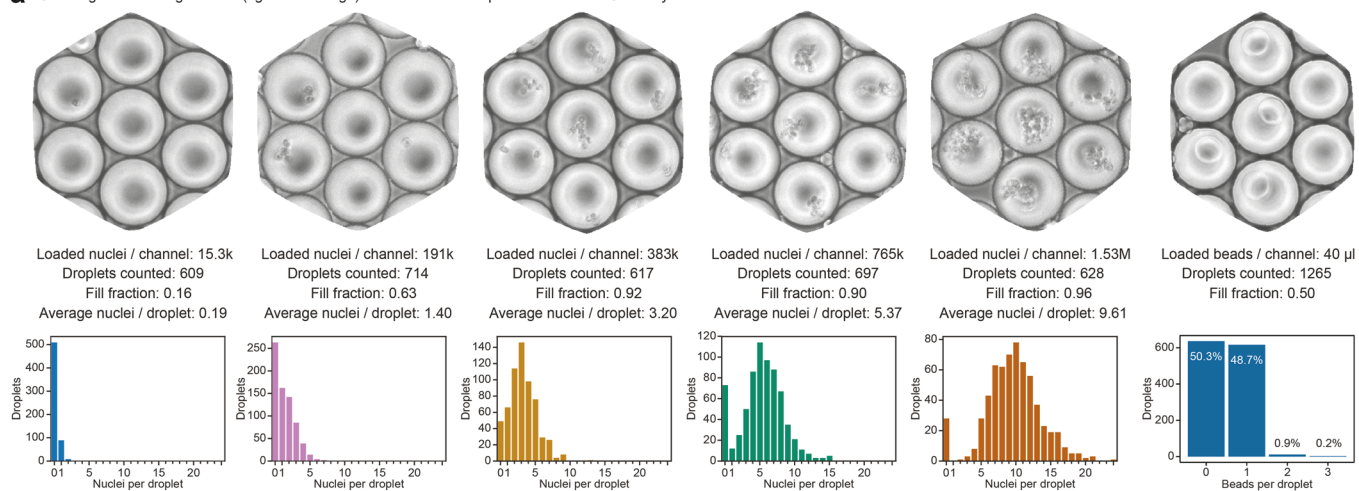
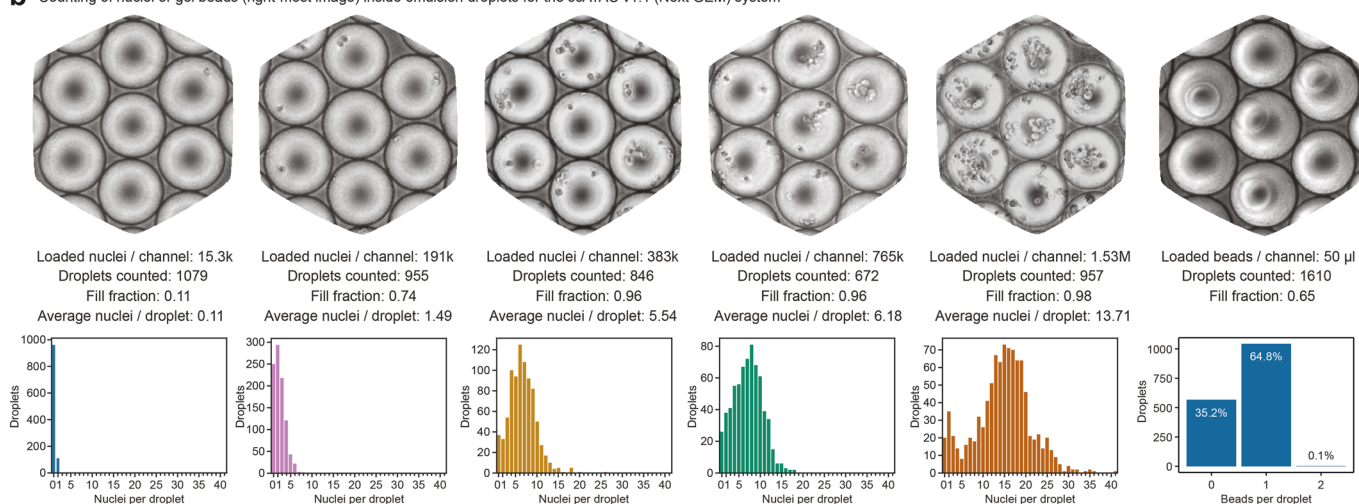
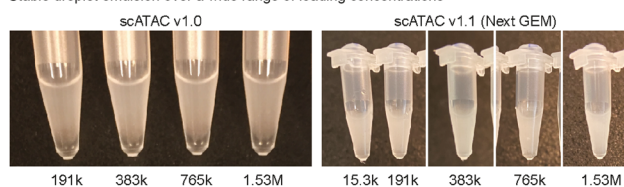
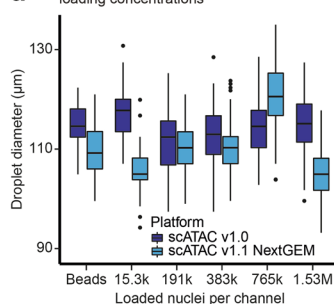
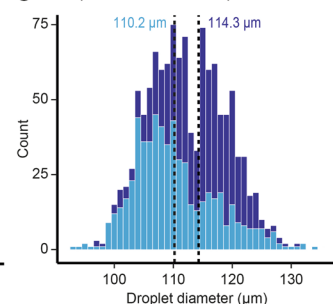
Extended data is available for this paper at <https://doi.org/10.1038/s41592-021-01153-z>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-021-01153-z>.

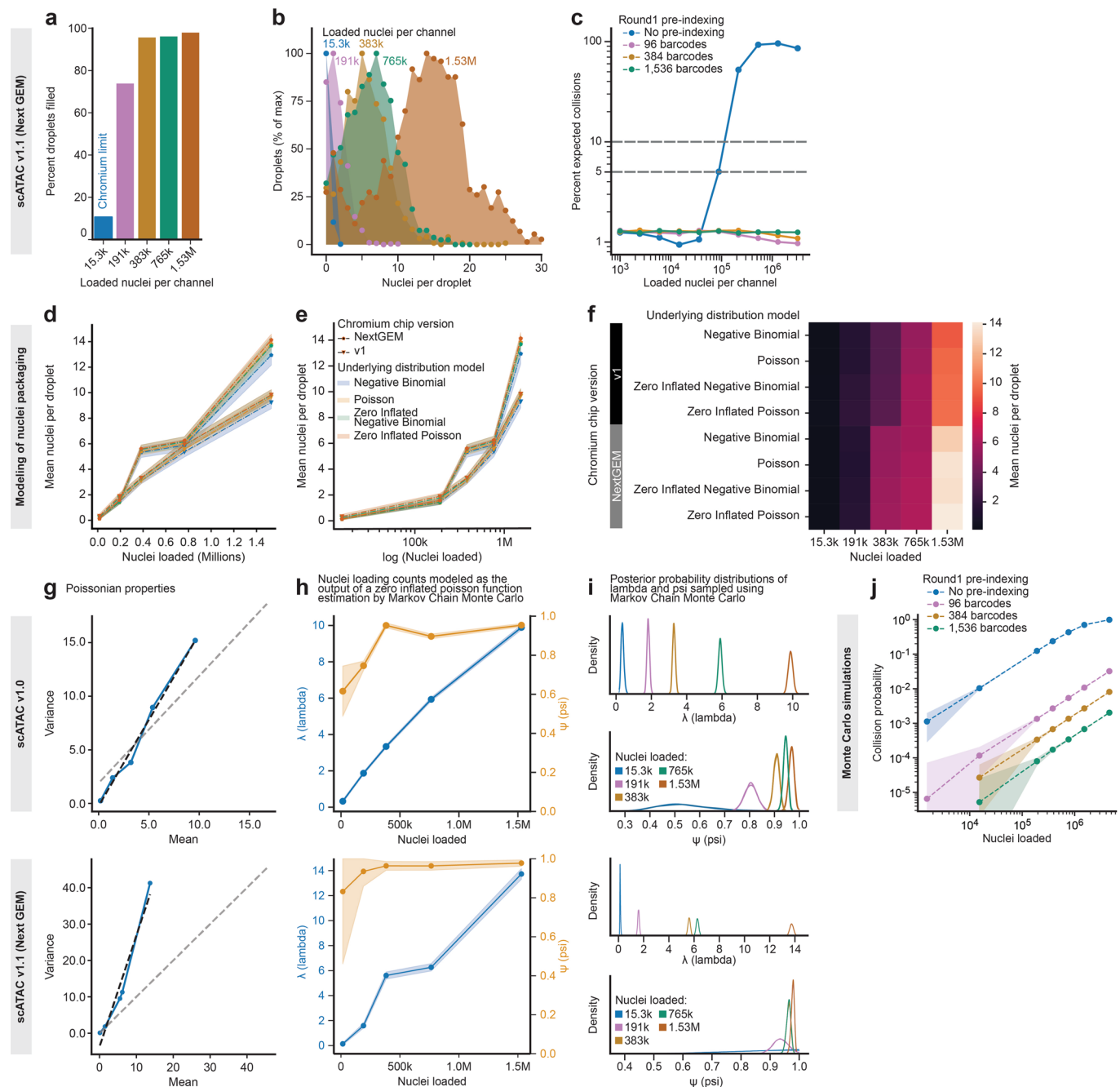
Correspondence and requests for materials should be addressed to C.B.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

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

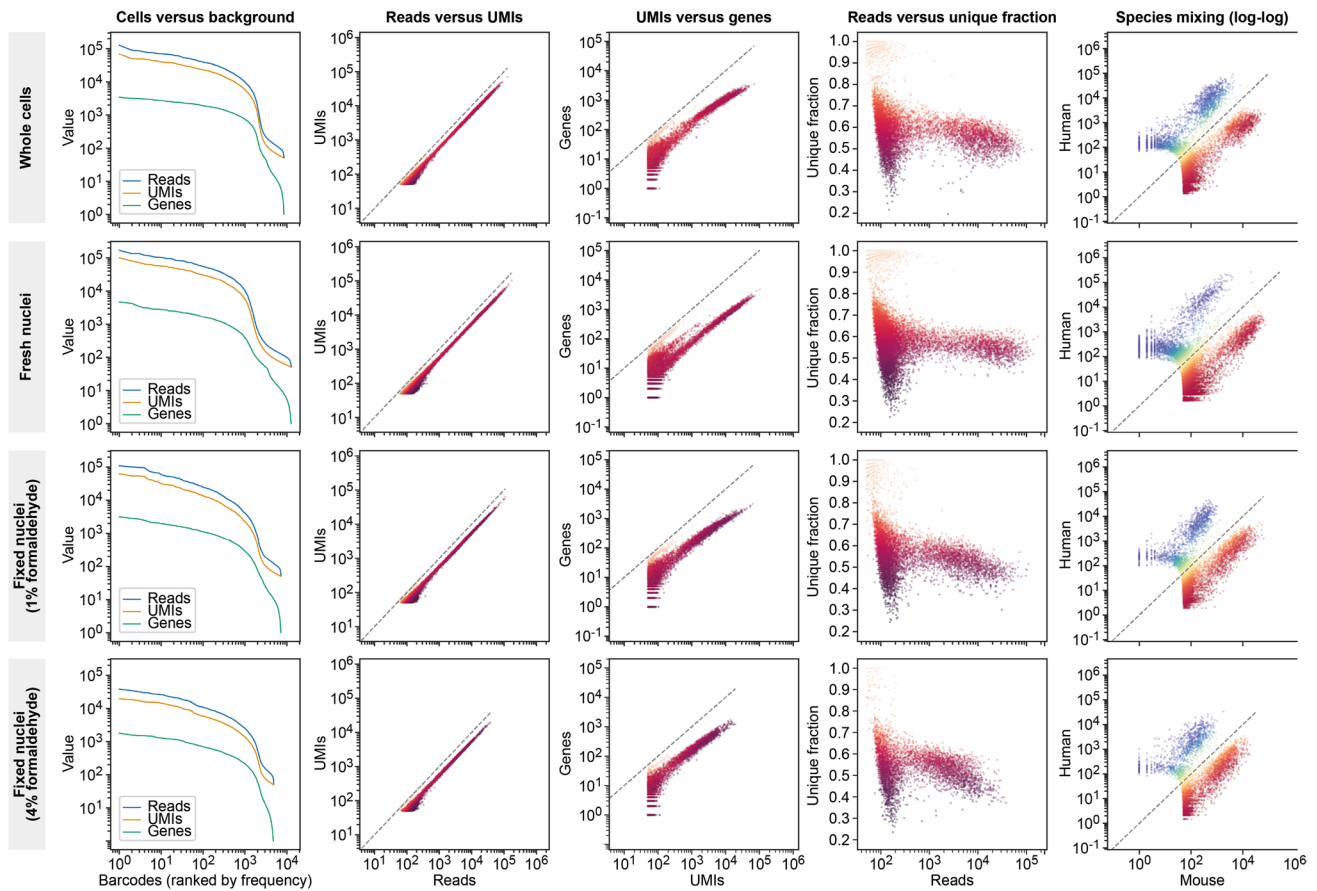
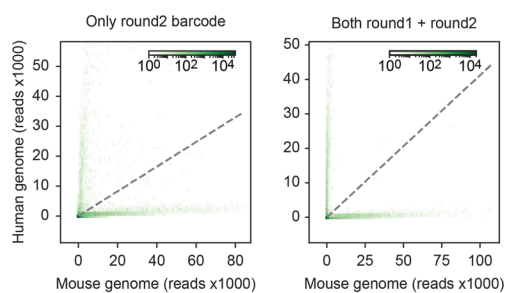
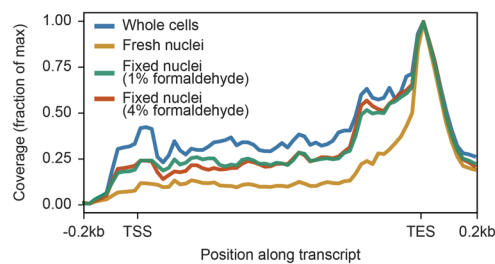
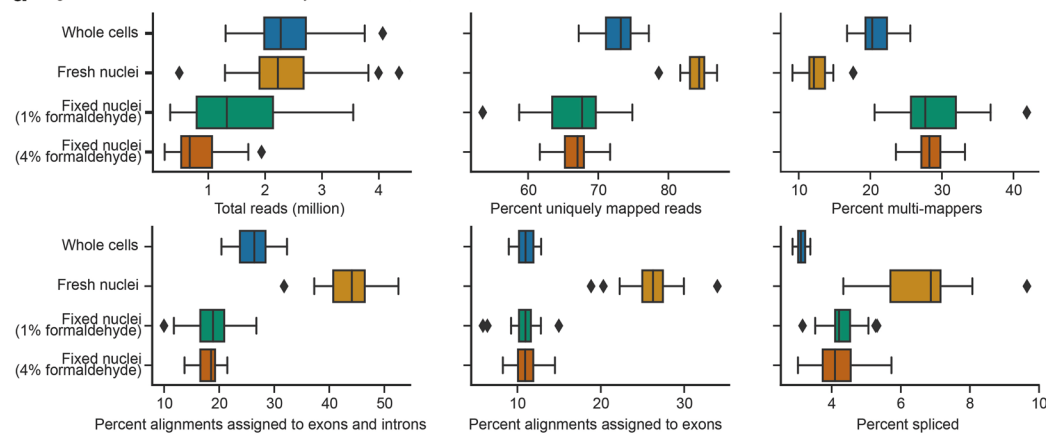
a Counting of nuclei or gel beads (right-most image) inside emulsion droplets for the scATAC v1.0 system**b** Counting of nuclei or gel beads (right-most image) inside emulsion droplets for the scATAC v1.1 (Next GEM) system**c** Stable droplet emulsion over a wide range of loading concentrations**d** Droplet diameter for increasing loading concentrations**e** Droplet diameter between platforms

Extended Data Fig. 1 | Droplet overloading for the Chromium scATAC v1.0 and v1.1 Next GEM microfluidic chips. **a–b**, Representative microscopy images of droplets (top rows) and histograms showing the number of nuclei per droplet (bottom rows) at different loading concentrations (15,300, 191,000, 383,000, 765,000, and 1,530,000 nuclei per channel) for the Chromium scATAC v1.0 chip (panel a) and for the scATAC v1.1 Next GEM chip (panel b). To obtain these images, lysis reagents were omitted from the cell loading experiment, and a total of 3,265 (scATAC v1.0) or 4,509 (scATAC v1.1 Next GEM) droplets were manually counted. Moreover, the number of beads per droplet (rightmost image and diagram) was visualized and counted based on a loading experiment in which the nuclei suspension was substituted by 1x Nuclei Buffer, while Reducing Agent B was omitted. **c**, Despite substantial droplet overloading, stable droplet emulsions were obtained for all tested conditions. **d**, Box plots showing the droplet diameters for the Chromium scATAC v1.0 and scATAC v1.1 Next GEM microfluidic chips at different loading concentrations. For each setup, 100 droplets were evaluated. Box plots depict the interquartile range with marked median and whiskers extending to 1.5 times the interquartile range. **e**, Histogram showing droplet diameters (as in panel d) pooled across different loading concentrations (500 droplets per platform).



Extended Data Fig. 2 | Computational modeling of microfluidic chip loading and of barcode collisions. **a**, Droplet overloading boosts the percentage of droplets filled with nuclei for the scATAC v1.1 Next GEM microfluidic chip. **b**, Droplet overloading on the scATAC v1.1 Next GEM chip increases the average number of nuclei per droplet in a controlled fashion, while maintaining the desired Poisson-like loading distribution. **c**, Expected collision rates on the Next GEM chip as a function of the loaded number of cells or nuclei per channel for standard droplet-based scRNA-seq and for scifi-RNA-seq with different numbers of round1 barcodes. The cell/nuclei fill rate was modeled as a zero-inflated Poisson distribution. **d-f**, Modeling of the microfluidic device loading using alternative distributions (Negative Binomial, Poisson, Zero Inflated Negative Binomial, Zero Inflated Poisson). The number of loaded nuclei is plotted against the number of nuclei per droplet on a linear scale (panel **d**), logarithmic scale (panel **e**) and as point estimates (panel **f**). **g**, Statistical properties of the distribution of nuclei per droplet across experiments. The relationship between mean and variance that is expected for a Poisson distribution is indicated by gray lines. **h**, Computational modeling of droplet loading as a zero-inflated Poisson function. **i**, Posterior probability distributions of λ and ψ sampled using a Markov Chain Monte Carlo (MCMC) analysis. **j**, Independent estimation of the cell doublet rates using Monte Carlo simulations. Error bars in panels **d**, **e**, **h**, and **j** indicate three standard deviations around the mean.

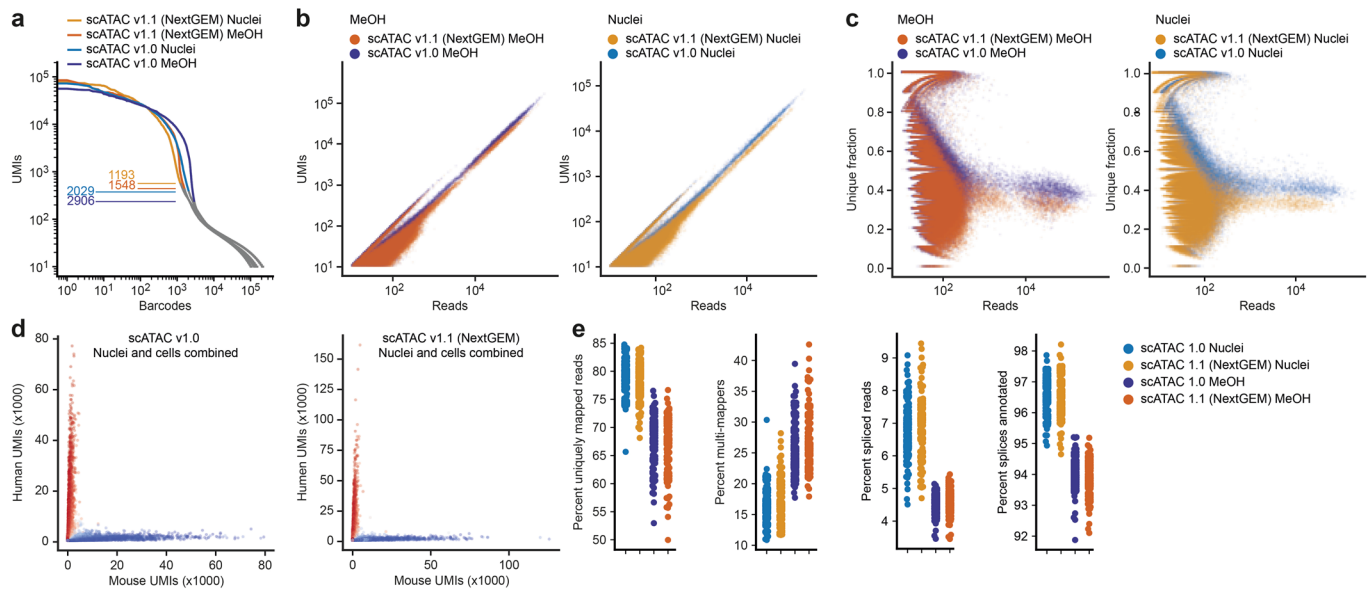
Extended Data Fig. 3 | Detailed assay design and performance metrics of scifi-RNA-seq. **a**, Schematic outline of scifi-RNA-seq including detailed oligonucleotide sequences. The reverse transcription is performed inside permeabilized cells or nuclei on a 96-well or 384-well plate, introducing well-specific round1 barcodes into the whole transcriptome. Pre-indexed cells or nuclei are pooled and encapsulated into emulsion droplets using a standard microfluidic droplet generator (10x Genomics Chromium). The round2 barcodes are introduced by thermocycling ligation with a complementary bridge oligo and thermostable ligase. The droplet emulsion is then broken, and a second defined end is introduced into the library via template switching. cDNA is enriched and tagmented with a custom i7-only transposome. Finally, the library is PCR-enriched, with the option to introduce an additional sample index. The read structure for next-generation sequencing on the Illumina NovaSeq 6000 and NextSeq 500 platforms is shown. **b**, Nuclei recovery after pre-indexing of the whole transcriptome by reverse transcription. scifi-RNA-seq achieves high recovery rates for both cell lines and primary material. **c**, Nuclei with pre-indexed transcriptome, prior to microfluidic device loading, visualized under a microscope in a counting chamber. The selected image (representative of two replicate samples) shows nuclei derived from human primary T cells. **d**, Typical size distribution of enriched cDNA obtained with scifi-RNA-seq. **e**, Typical size distribution of final scifi-RNA-seq libraries ready for next-generation sequencing. **f**, Distribution of DNA bases along scifi-RNA-seq sequencing reads, showing the characteristic sequence patterns of the UMI, round1 barcode, sample barcode, round2 barcode, and transcript. **g**, Heatmap showing sequencing quality (Qscore) for each sequencing cycle.

a Data quality metrics**b****c** Coverage along transcripts**d** Alignment metrics for whole cells, freshly isolated nuclei, and fixed nuclei

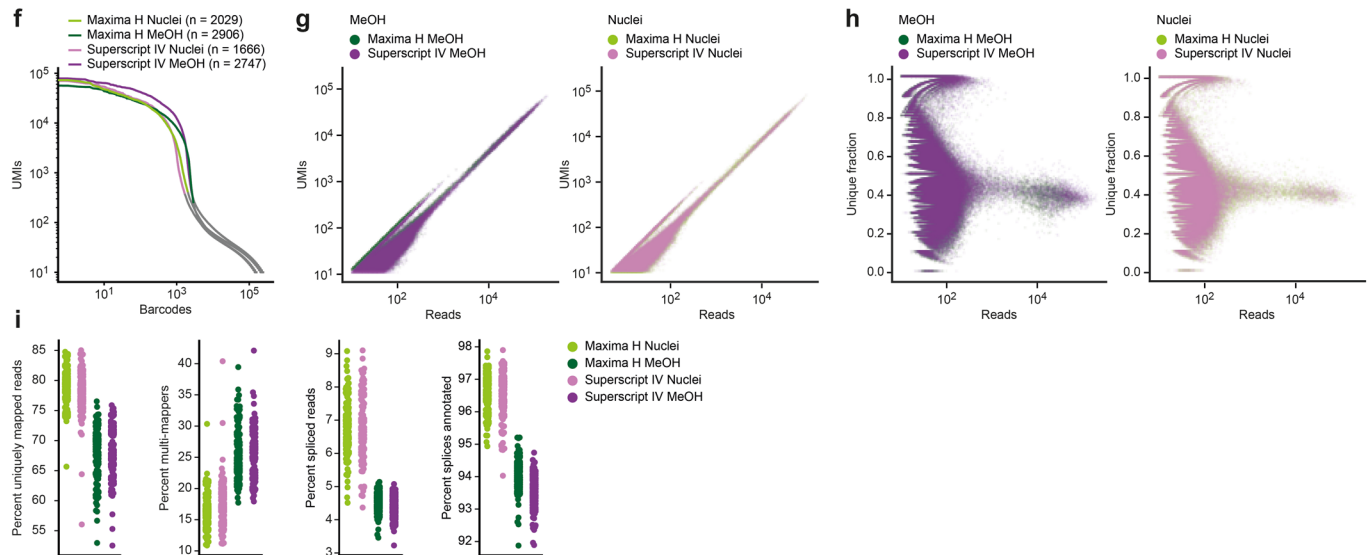
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | scifi-RNA-seq yields high-quality data for whole cells, fresh nuclei and fixed nuclei. a. Performance metrics for scifi-RNA-seq experiments using a mixture of human Jurkat cells and mouse 3T3 cells, starting from whole cells permeabilized by methanol, freshly isolated nuclei, and nuclei fixed with 1% or 4% formaldehyde (cryopreserved, re-hydrated, and permeabilized). The following plots are shown: (i) ranked barcodes plotted against reads, unique molecular identifiers (UMIs), and detected genes, distinguishing single-cell transcriptomes from background noise; (ii) reads plotted against UMIs; (iii) reads plotted against the number of detected genes; (iv) reads plotted against the fraction of unique reads; (v) species mixing plot showing the number of UMIs per cell aligning to the mouse genome (x-axis) versus the human genome (y-axis). To facilitate comparisons between the different types of input material, the axes of the performance plots use the same scale across conditions. **b.** In a species mixing experiment with pre-indexed nuclei from human (Jurkat) and mouse (3T3) cells run at the maximum loading concentration of the standard Chromium protocol (15,300 nuclei per channel), the microfluidic round2 barcode (left plot) is sufficient to resolve single cells. Nevertheless, the combination of round1 and round2 barcodes still improves the separation (right plot). **c.** Coverage along human and mouse transcripts from 200 bp upstream of the transcription start site (TSS) to 200 bp downstream of the transcription end site (TES), shown for whole cells permeabilized by methanol, freshly isolated nuclei, and nuclei fixed with 1% or 4% formaldehyde (cryopreserved, re-hydrated, and permeabilized). Freshly isolated nuclei show the strongest 3' enrichment. **d.** Box plots summarizing sequence alignment metrics across the different types of input material: Total reads sequenced, percent uniquely mapped reads, percent multi-mappers, percent alignments to exons plus introns, percent alignments to exons, and percent spliced reads. Freshly isolated nuclei showed the best performance for these alignment metrics. The box plots summarize a total of 2,299 whole cells; 2,000 fresh nuclei; 2,051 nuclei fixed with 1% formaldehyde and 1,896 nuclei fixed with 4% formaldehyde. Box plots depict the interquartile range with marked median and whiskers extending to 1.5 times the interquartile range.

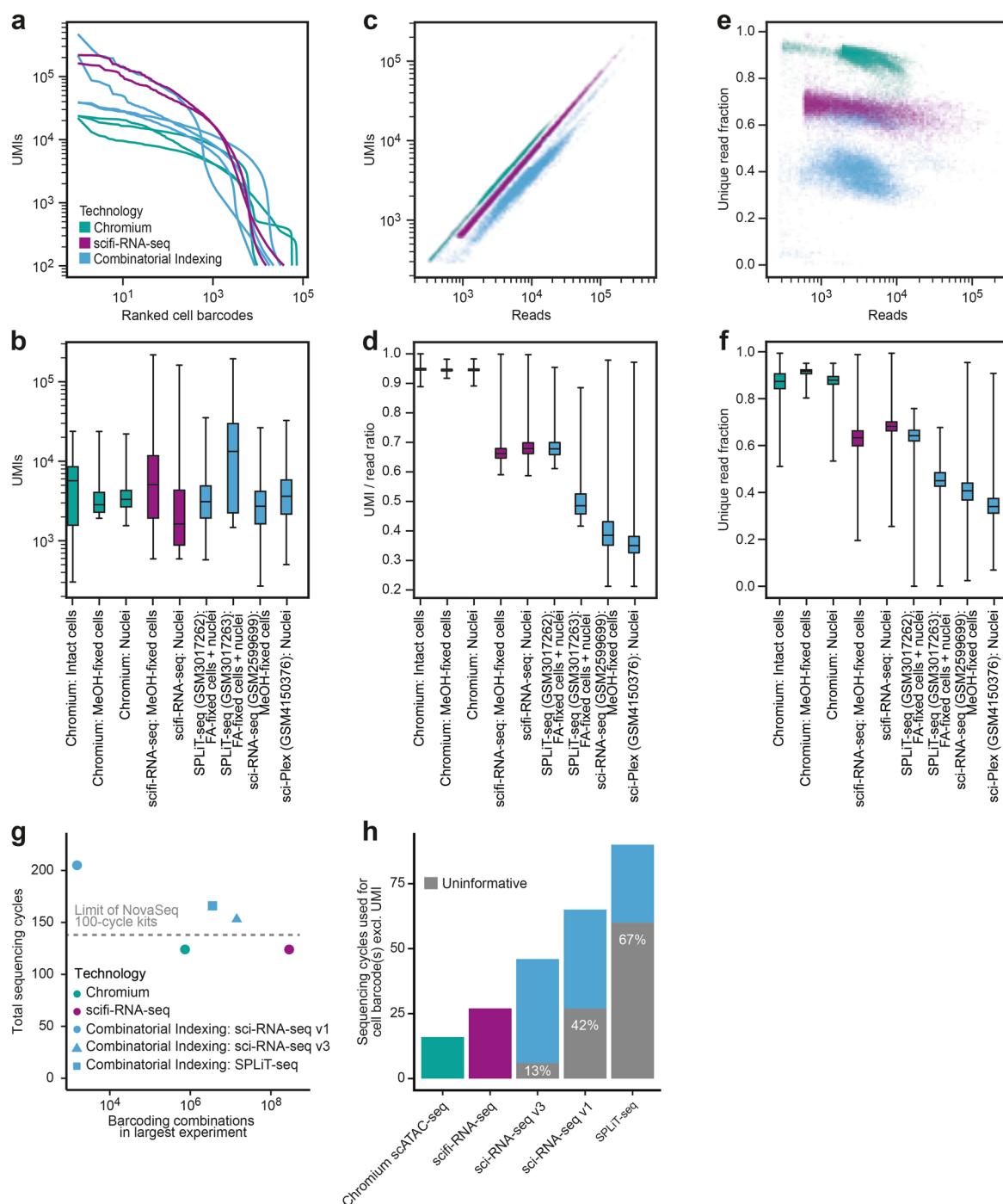
Comparison: scATAC v1.0 vs. v1.1 (NextGEM)



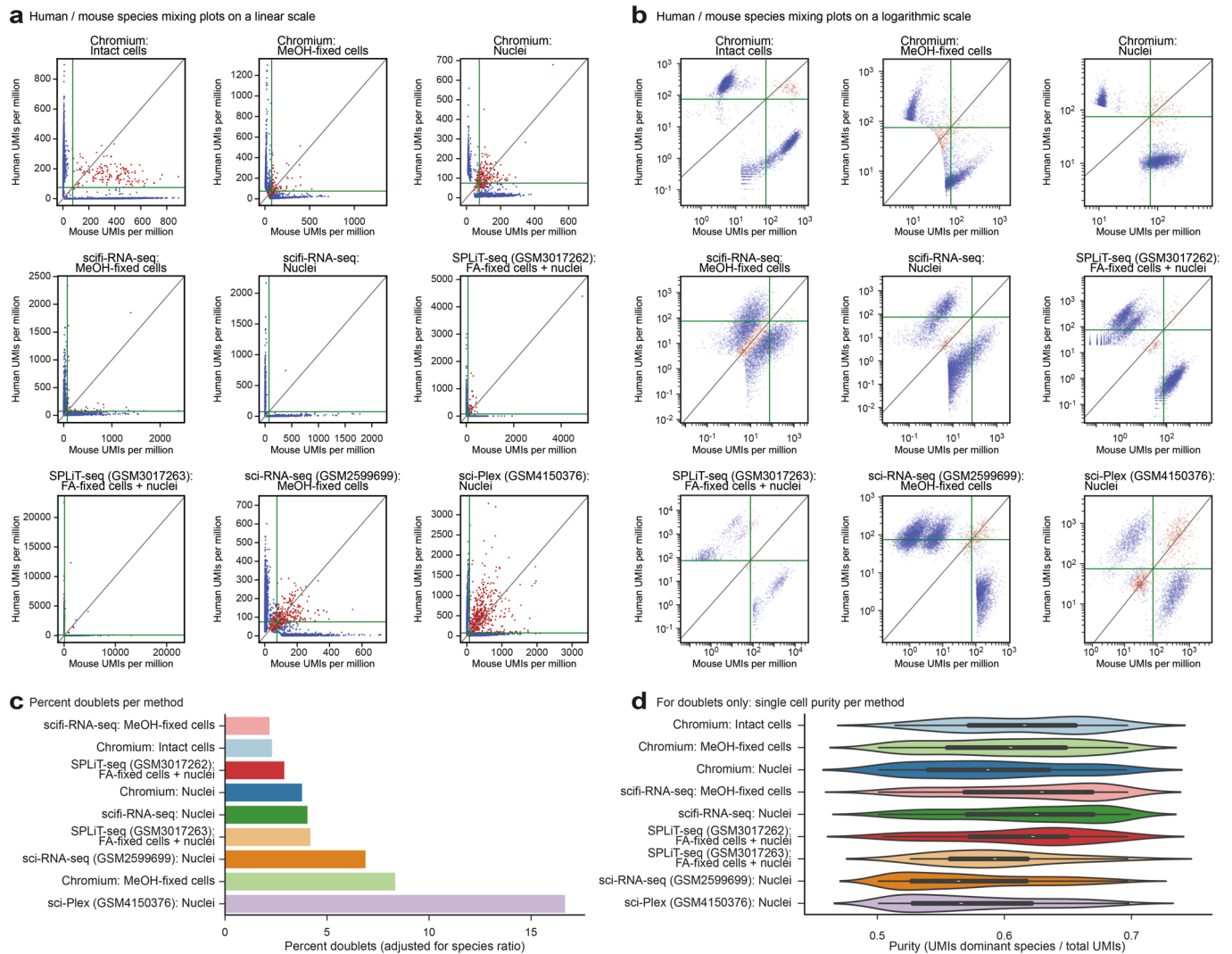
Comparison: Maxima H vs. Superscript IV



Extended Data Fig. 5 | scifi-RNA-seq performance metrics for different experimental conditions. **a**, 'Knee plot' showing the number of UMIs (y-axis) per barcode ranked by frequency (x-axis) for scifi-RNA-seq on the Chromium scATAC v1.0 chip versus the scATAC v1.1 Next GEM chip. The characteristic inflection points are indicated, which separate cells/nuclei (left, colored lines) from background noise (right, grey lines). **b**, Reads per cell plotted against UMIs per cell to assess the level of sequencing saturation for the two microfluidic chips. **c**, Reads per cell plotted against the unique read fraction per cell to assess PCR duplication and library complexity for the two microfluidic chips. **d**, Alignments to the human genome versus alignments to the mouse genome in the species mixing experiment to assess the frequency of cell doublets for the two microfluidic chips. **e**, Alignment metrics for the two microfluidic chips. **f**, 'Knee plot' for the comparison of two reverse transcriptase enzymes (Maxima H Minus versus Superscript IV) in the reverse transcription step of scifi-RNA-seq (the template switching was performed with Maxima H Minus reverse transcriptase in both cases). **g**, Reads per cell plotted against UMIs per cell to assess the level of sequencing saturation for the two reverse transcriptases. **h**, Reads per cell plotted against the unique read fraction per cell to assess PCR duplication and library complexity for the two reverse transcriptases. **i**, Alignment metrics for the two reverse transcriptases.

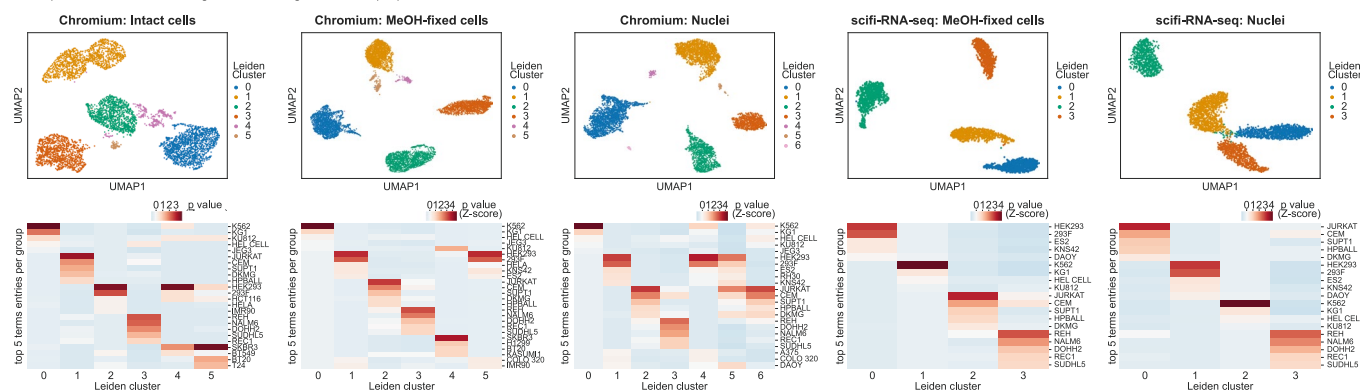
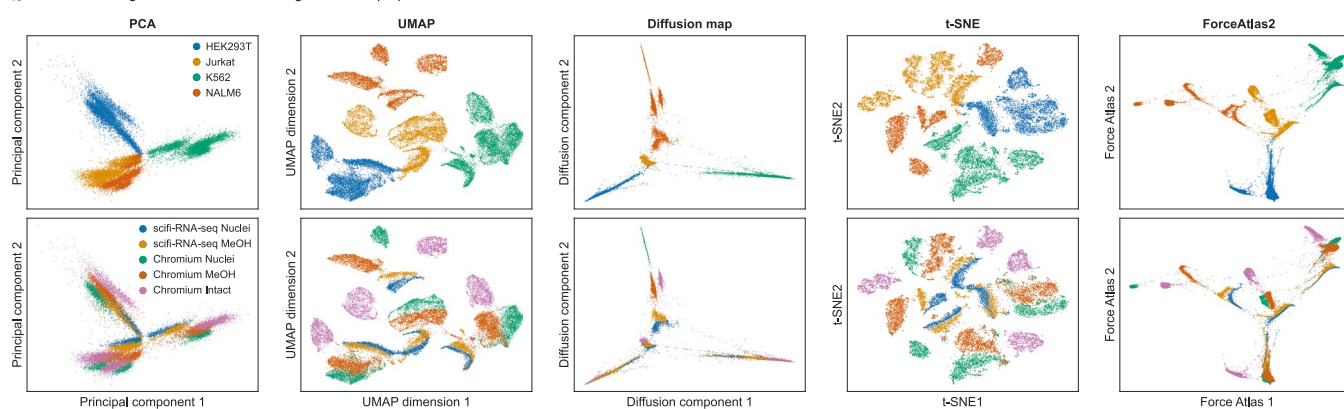
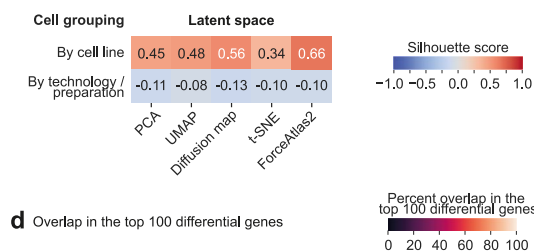
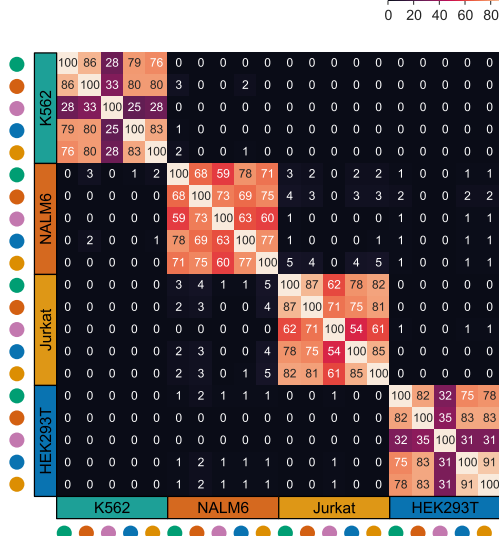
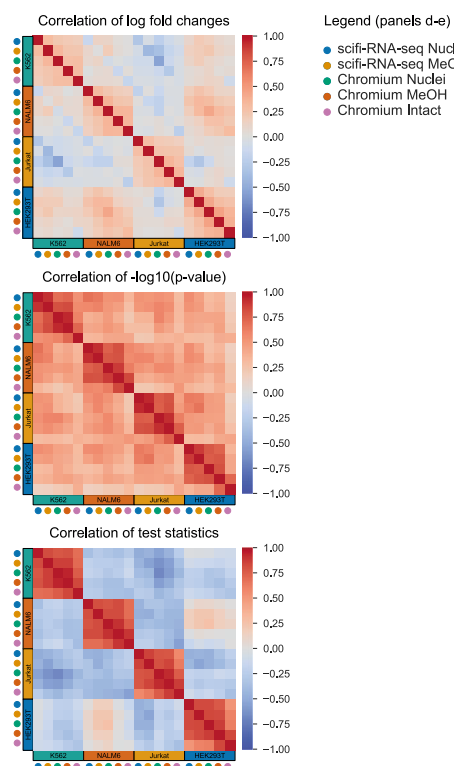


Extended Data Fig. 6 | Comparison of scifi-RNA-seq, droplet-based scRNA-seq and multiround combinatorial indexing in terms of library complexity, read duplication and barcode sequencing efficiency. a, 'Knee plot' for the comparison of scifi-RNA-seq (this study), droplet-based scRNA-seq using the Chromium system (this study), and multi-round combinatorial indexing (published data) on mouse 3T3 cells. **b**, Box plot showing UMI counts for each assay. Box plots in this figure summarize a total number of single cell profiles $n = 2,994$ (Chromium: Intact cells); 2,878 (Chromium: MeOH-fixed cells); 3,523 (Chromium: Nuclei); 4,305 (scifi-RNA-seq: MeOH-fixed cells); 4,945 (scifi-RNA-seq: Nuclei); 3,443 (SPLiT-seq [GSM3017262](#)); 526 (SPLiT-seq [GSM3017263](#)); 8,874 (sci-RNA-seq [GSM2599699](#)); 1,944 (sci-Plex [GSM4150376](#)). Box plots depict the interquartile range with marked median and whiskers extending to 1.5 times the interquartile range. **c**, Reads per cell plotted against UMIs per cell to assess the level of sequencing saturation for each assay. **d**, Box plot showing the UMIs per read ratio. **e**, Reads per cell plotted against the unique read fraction per cell to assess PCR duplication and library complexity for each assay. **f**, Box plot showing the unique read fraction for each assay. **g**, Barcoding combinations in the largest published experiment against the total number of sequencing cycles used in that experiment. The grey line shows the total number of 138 sequencing cycles (including index cycles) available in the NovaSeq 100-cycle kits. **h**, Sequencing cycles used for reading the composite cell barcode (excluding the UMI). Uninformative sequencing cycles from ligation overhangs, primer binding sites, and transposase mosaic ends are depicted in gray, and the fraction of uninformative sequencing cycles is shown as a percentage value.



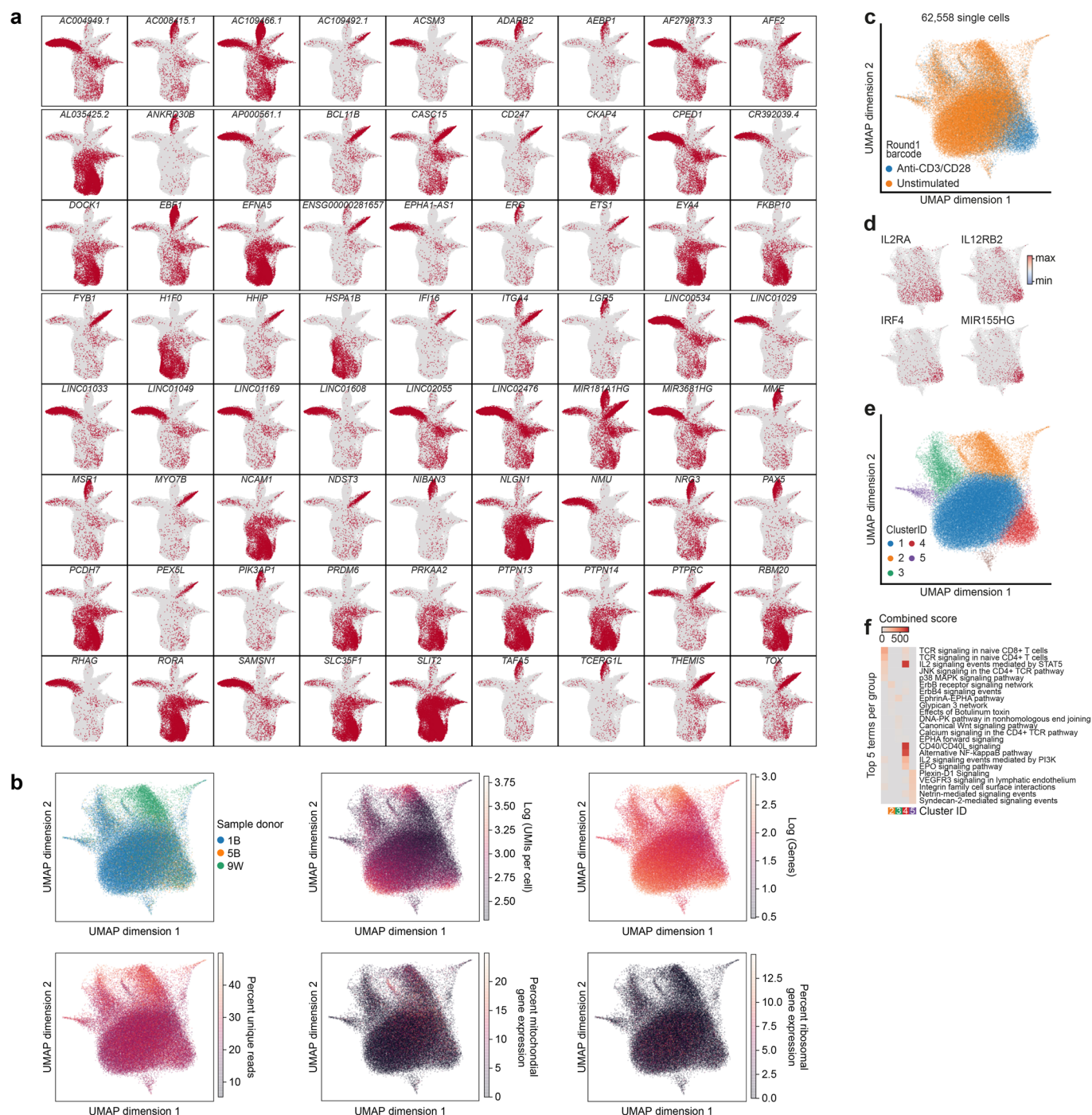
Extended Data Fig. 7 | Comparison of sciRNA-seq, droplet-based scRNA-seq and multi-round combinatorial indexing in terms of cell doublet rates.

a, Alignments to the mouse genome versus alignments to the human genome for each species mixing experiment, assessing the frequency of cell doublets for the different methods. Data are shown on a linear scale, normalized to UMIs per million to allow the use of common thresholds across experiments. Cells were classified as doublets if (i) there was less than twice as many UMIs aligning to one species' genome than to the other species' genome, or (ii) more than 75 UMIs per million were detected in both species. Doublet cells are highlighted in red. The grey line indicates $x=y$ after accounting for species bias. The green lines indicate the threshold value of 75 UMIs per million. **b**, Same visualization as in panel **a**, but plotted on a logarithmic scale. **c**, Percentage of cell doublets for each method, corresponding to the red cells in panels **a** and **b**. **d**, Single-cell purity plotted for doublet cells only. Purity was calculated as the cell's number of UMIs for the dominant species divided by its total number of UMIs.

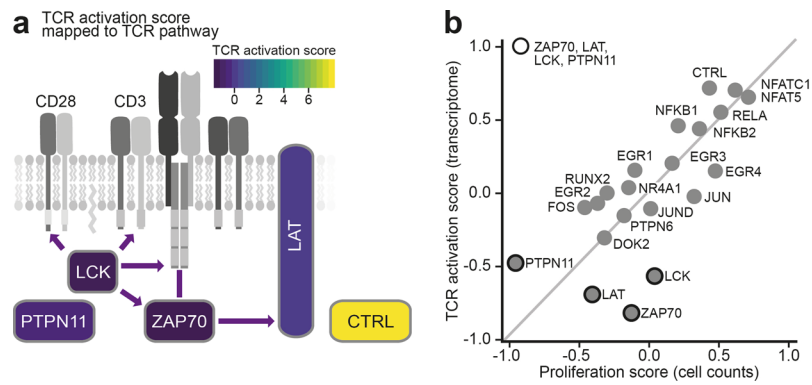
a Separate UMAP embeddings for technologies and cell preparation methods**b** Joint embeddings of cells across technologies and cell preparation methods**c** Quantification of cellular separation with the silhouette score**d** Overlap in the top 100 differential genes**e** Quantification of cellular separation with the silhouette score

Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Single-cell embeddings and transcriptome comparisons of scifi-RNA-seq and droplet-based scRNA-seq. An equal mixture of four human cell lines (HEK293T, Jurkat, K562, NALM-6) was processed in parallel with scifi-RNA-seq and with the Chromium 3' v3.1 Single Cell Gene Expression kit. **a**, Single-cell transcriptomes displayed in a two-dimensional UMAP projection, with cluster IDs identified by the Leiden algorithm mapped on top. Enrichment of cell line signatures obtained from the ARCHS4 database for the identified Leiden clusters. These results can be used to assign the respective cell line for each cluster, and to identify spurious clusters of doublet cells. **b**, Joint embeddings combining data across methods (scifi-RNA-seq, standard droplet-based scRNA-seq) and sample preparation methods (intact cells, nuclei, methanol-fixed cells), using dimensionality reduction by principal component analysis (PCA), uniform manifold approximation and projection (UMAP), diffusion maps, t-distributed stochastic neighbor embedding (t-SNE), and the ForceAtlas2 algorithm. Individual cells are colored by cell line (top panel) or sample preparation method (bottom panel). The grouping by cell line (rather than by assay or sample preparation method) was observed without batch effect correction. **c**, The separation of cells in the latent spaces was quantified using the silhouette score. **d**, Overlap in the top-100 differential genes between cell lines. **e**, Correlation matrices of log fold changes, p-values, and test statistics across assays and sample preparation methods.



Extended Data Fig. 9 | Large-scale scifi-RNA-seq profiling for a mixture of four human cell lines and for primary human T cells with and without TCR stimulation. **a**, Gene expression levels obtained with scifi-RNA-seq for a mixture of four human cell lines. The expression levels of 72 cell-line specific genes were mapped on top of the UMAP projection from Fig. 2j. **b**, UMAP projections for 62,558 single-cell transcriptomes of human primary T cells, with additional variables mapped on top of the projections: Donor ID, logarithm of UMIs per cell, logarithm of detected genes per cell, percent unique reads per cell, percent mitochondrial expression, and percent ribosomal expression. **c**, UMAP projection for the single-cell transcriptomes, with T cell receptor stimulation status mapped on top of the projection. **d**, Expression levels of four genes induced by TCR stimulation mapped on top of the UMAP projection. **e**, UMAP projection for the single-cell transcriptomes, with single cells colored according to the clusters assigned by graph-based clustering using the Leiden algorithm. **f**, Gene set enrichment analysis for the differentially expressed genes in each cluster.



Extended Data Fig. 10 | Arrayed CRISPR screen for TCR activation with multiplexed scifi-RNA-seq readout. a, TCR activation signature as defined in Fig. 3c, mapped on top of a schematic of cell signaling in TCR pathway activation. **b**, TCR activation score derived from the transcriptome data plotted against a proliferation score derived from the cell counts. Key regulators of the TCR pathway are highlighted.

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](#).

Statistics

For all statistical analyses, confirm that the following items are present in the 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
- ☐ ☒ A statement on 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 statistical parameters 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

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No specialized software was used for data collection

Data analysis

A detailed description of the data analysis is provided in the "Online Methods" section, specifically in the subsections entitled: "Modeling of microfluidic device loading", "Processing of scifi-RNA-seq data", "Bioinformatic analysis of scifi-RNA-seq data", "Comparative analysis of single-cell RNA-seq datasets", and "Analysis of scifi-RNA-seq data for the arrayed CRISPR screen". Two versions of the source code are provided as separate GitHub repositories: (i) analysis code for reproducing the results in this paper (https://github.com/epigen/scifiRNA-seq_publication); (ii) pipeline code for processing new scifi-RNA-seq datasets (<https://github.com/epigen/scifiRNA-seq>).

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. 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

All datasets from this study are available as raw and processed data from the NCBI GEO database (accession number GSE168620).

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

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

Sample size	Given the study's focus on methods development, no biological sample size calculations were performed. The sample size and cell numbers were chosen such that it enabled an adequate technical validation and benchmarking of the scifi-RNA-seq method.
Data exclusions	One blood donor was excluded due to low quality of the primary material; one cell line sample (HL-60) was excluded due to incompatibility with the nuclei preparation.
Replication	Technical replication and benchmarking supported the reproducibility and performance of the of the scifi-RNA-seq method. The arrayed CRISPR screen for TCR activation replicated our previous study (Datlinger et al., 2017). There were no findings that we failed to replicate.
Randomization	Sample groups such as cell line and T cell receptor stimulation state were not randomized. Batch effects were mitigated by running all reverse transcription reactions with the same master mix in the same plate, and packaging all cells in the same channel of the microfluidic device.
Blinding	For 10x Genomics scRNA-seq, the operator was blinded as to which sample types were run in a particular channel.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

Antibodies

Antibodies used	anti-human CD3 antibody (clone OKT3, ThermoFisher Scientific #16-0037-81) anti-human CD28 antibody (clone CD28.2, ThermoFisher Scientific #16-0289-81)
Validation	The used antibodies are commercially available, with detailed validation provided on the vendor's website. Concentrations and timing for TCR activation was validated as part of a previous study (Datlinger et al., 2017)

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	K-562 (DSMZ), HEK293T (DSMZ), Jurkat E6.1 (DSMZ), NALM-6 (DSMZ)
Authentication	Identity of the cell lines was assessed by transcriptome analysis (Fig. 2I)
Mycoplasma contamination	Cell lines tested negative for mycoplasma contamination at regular intervals, using the following assay: MycoAlert Mycoplasma Detection Kit, Lonza #LT07-118
Commonly misidentified lines (See ICLAC register)	No such cell lines were used

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	The study used samples from anonymous blood donors, hence no population characteristics were available
Recruitment	Blood samples from anonymous blood donors were purchased from the Austrian Red Cross
Ethics oversight	The study complied with all relevant ethical regulations for working with human primary samples. Informed consent was obtained from all sample donors. The use of blood samples was approved by the ethical committees of the Austrian Red Cross and the Medical University of Vienna.

Note that full information on the approval of the study protocol must also be provided in the manuscript.