nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-0837-5

Check for updates

Targeted Perturb-seq enables genome-scale genetic screens in single cells

Daniel Schraivogel[1,5], Andreas R. Gschwind[2,5], Jennifer H. Milbank[1], Daniel R. Leonce[1], Petra Jakob[1], Lukas Mathur[1], Jan O. Korbel[1], Christoph A. Merten[1], Lars Velten[1,4] ✉ and Lars M. Steinmetz[1,2,3] ✉

The transcriptome contains rich information on molecular, cellular and organismal phenotypes. However, experimental and statistical limitations constrain sensitivity and throughput of genetic screening with single-cell transcriptomics readout. To overcome these limitations, we introduce targeted Perturb-seq (TAP-seq), a sensitive, inexpensive and platform-independent method focusing single-cell RNA-seq coverage on genes of interest, thereby increasing the sensitivity and scale of genetic screens by orders of magnitude. TAP-seq permits routine analysis of thousands of CRISPR-mediated perturbations within a single experiment, detects weak effects and lowly expressed genes, and decreases sequencing requirements by up to 50-fold. We apply TAP-seq to generate perturbation-based enhancer–target gene maps for 1,778 enhancers within 2.5% of the human genome. We thereby show that enhancer–target association is jointly determined by three-dimensional contact frequency and epigenetic states, allowing accurate prediction of enhancer targets throughout the genome. In addition, we demonstrate that TAP-seq can identify cell subtypes with only 100 sequencing reads per cell.

Genetic perturbation studies have been instrumental for delineating causal relationships between genes and phenotypes[1,2]. Compared with unimodal readouts, such as growth or reporter gene expression, single-cell transcriptomics provides a greater wealth of data on molecular and cellular phenotypes. Thus, pooled CRISPR screens that couple genetic perturbations with single-cell transcriptomics (Perturb-seq, also referred to as CROP-seq) have emerged as powerful tools to characterize the consequences of genetic perturbations[3–9]. In these experiments, each cell stochastically receives one guide RNA out of a guide RNA library, enabling high numbers of perturbations to be assayed in a single experiment. Single-cell RNA-seq is then used to retrieve the identity of the gRNA in each cell along with its effect on the transcriptome, including changes in the expression of single genes[4,7–9], as well as large transcriptomic rearrangements[3–6,8]. Diverse applications have been pursued, including the characterization of genetic regulators of signaling pathway activity[6,8] and cellular differentiation[5], or the mapping of gene regulatory networks[7,9]. Moreover, measuring gene expression in single cells obviates the need for cellular assays, and can be applied to any cell type, including rare populations[5] and primary cells that cannot be cultivated extensively[3].

However, three major factors currently limit a widespread use of Perturb-seq type of experiments. First, costs are prohibitive, even for non-genome-scale screens[4]. Second, lowly expressed genes and small effects are not measured efficiently[4,7–9]. Third, data analysis suffers from a potentially insurmountable multiple-testing problem. For example, in a hypothetical genome-wide Perturb-seq screen, 20,000 hypotheses on gene expression changes need to be tested for each of 20,000 knockouts (that is, 400 million total tests). To deal with this problem, previous studies focused on hypothesis-driven analyses of just parts of the data generated in whole-transcriptome screens. For example, differential gene expression testing was restricted to pre-defined candidate genes[7,9,10], or genes were grouped into gene signatures[3,4,6,8]. Alternatively, cells were mapped to known reference cell (sub)types or states[4,5,11,12], making effective use only of cell-state- or cell-type-specific marker genes. Measuring the entire transcriptome provides no clear benefit in these cases; in turn, a targeted readout of only relevant genes could overcome limitations, enable genome-scale screens, and open new applications, such as combinatorial screens, screens in diverse genetic backgrounds or diverse cell types.

Here, we demonstrate that targeted Perturb-seq (TAP-seq) turns the necessity of restricting the hypothesis space into a virtue for overcoming the hitherto prohibitive sequencing requirements and lack of sensitivity. By amplifying genes of interest, rather than the whole transcriptome from single cells, TAP-seq enables single-cell genetic and functional genomic screens at 50-fold higher scale and lower cost. TAP-seq identifies gene expression changes more sensitively than does Perturb-seq, and robustly distinguishes cell (sub) types and differentiation stages. We demonstrate the use of TAP-seq by perturbing all putatively active enhancers in two large genomic regions and querying the effects on expressed genes in the same regions. This allowed us to sensitively identify enhancer target genes on the basis of functional evidence, rather than predictions from indirect measures such as interactions in 3D space[13,14]. We thereby provide the first high-resolution function-based enhancer–target gene map for a sizeable fraction of the human genome.

Results

We have established targeted Perturb-seq (TAP-seq) (Fig. 1a), a protocol that is designed to be compatible with various 3′ single-cell library-preparation methods, and is implemented here for 10x

[1]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. [2]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [3]Stanford Genome Technology Center, Palo Alto, CA, USA. [4]Present address: Center for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain. [5]These authors contributed equally: Daniel Schraivogel, Andreas R. Gschwind.
✉e-mail: lars.velten@crg.eu; larsms@embl.de

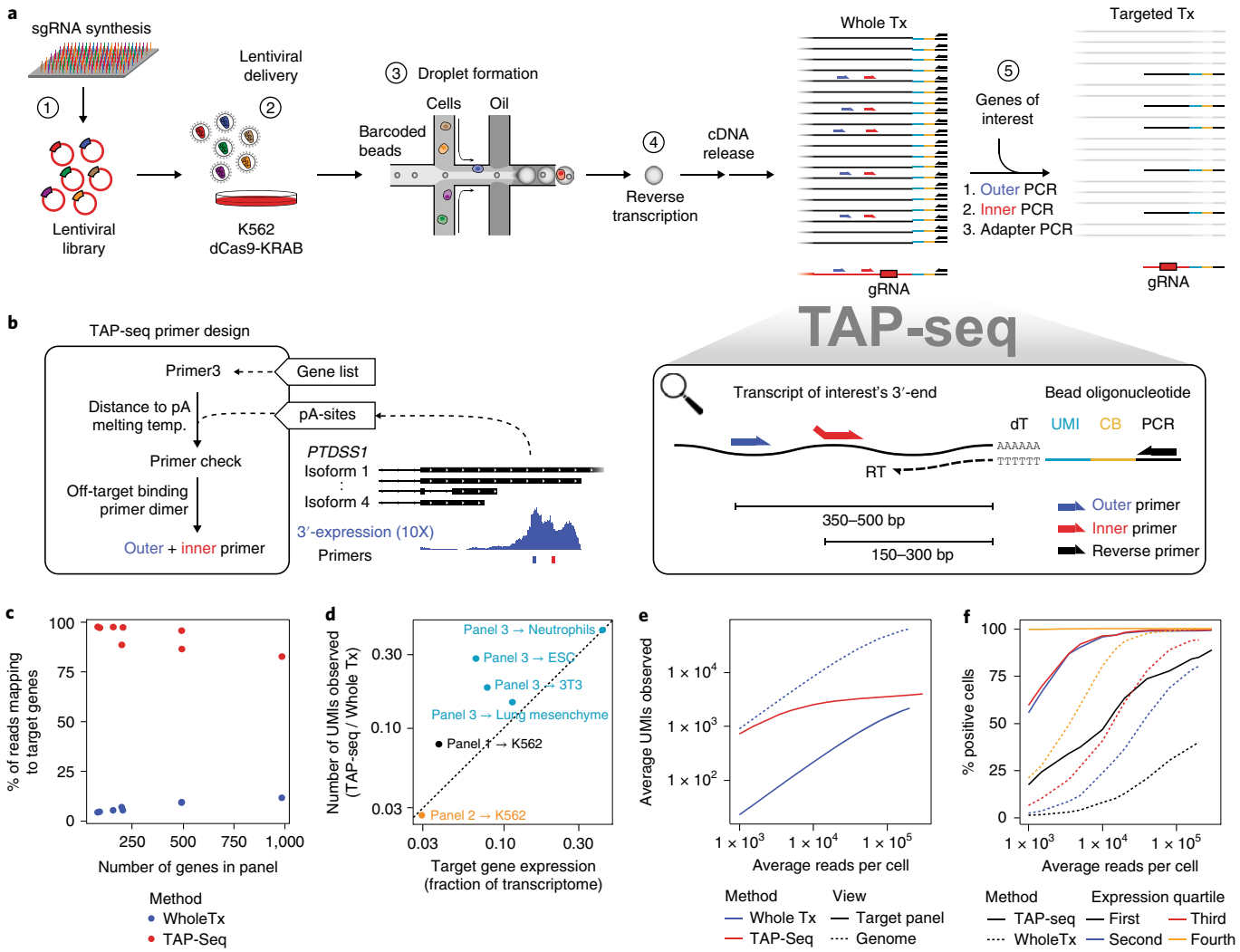


Fig. 1 | TAP-seq permits efficient expression profiling of target genes in single cells. a, Overview of the method. UMI, unique molecular identifier; CB, cell barcode; RT, reverse transcription; K562 dCas9-KRAB, K562 cells stably expressing dCas9-KRAB. **b**, Schematic presentation of the TAP-seq primer design pipeline. *PTDSS1*, phosphatidylserine synthase 1 (shown here as an example of a gene of interest). **c**, Fraction of reads mapped to target genes, comparing TAP-seq and whole transcriptome (Whole Tx) readouts for all target gene panels used in this study. **d**, For three different panels and five different cell types, library complexity was quantified as the number of UMIs observed. Complexity of TAP-seq libraries as a fraction of complexity of whole-transcriptome libraries (*y* axis) is plotted against the targeted fraction of the transcriptome (*x* axis). **e**, Deeply sequenced TAP-seq (panel 1, 74 target genes) and whole-transcriptome libraries were downsampled to a given average number of reads per cell (*x* axis). The average number of UMIs observed on the target panel (solid lines, shown for both methods) or across the entire genome (dashed line, only shown for whole-transcriptome analysis) is shown. **f**, Data were downsampled as described in **e**, and for each gene from the panel, the fraction of cells displaying expression of that gene was computed. Genes were then classified into four bins on the basis of their mean expression in whole-transcriptome data, and average positivity scores across all genes in the bin are shown.

Genomics[15] and Drop-seq[16]. In these protocols, cellular and molecular barcodes as well as a universal PCR handle are introduced during reverse transcription. Following complementary DNA retrieval, TAP-seq uses the universal PCR handle and gene-specific primers to amplify transcripts of interest in two semi-nested multiplex PCRs, and a third PCR to add Illumina sequencing adapters (Fig. 1a and Extended Data Fig. 1). The gRNA-encoding CROP-seq[6] vector is also targeted during that process. A primer design pipeline that positions primers in a specified distance from the polyadenylation site while avoiding off-targets and primer dimers (Fig. 1b) is provided as an R package on Bioconductor (http://bioconductor.org/packages/release/bioc/html/TAPseq.html).

To evaluate TAP-seq, we first designed three primer panels, targeting 74 to 198 genes of widely varying expression levels, corresponding to between 3 and 43% of the transcriptome. TAP-seq was then performed with K562 cells (panels 1 and 2), two different mouse cell lines (panel 3) or two mouse primary cell types (panel 3) (Extended Data Fig. 2a and Supplementary Table 1; see Methods for target gene selection and Extended Data Fig. 2b,c for a comparison of 10X Genomics[15] and Drop-Seq[16]). Across all panels and cell types, between 87 and 96% of sequencing reads mapped to the selected target sequences (Fig. 1c). When using a substantially larger panel of 1,000 genes (Extended Data Fig. 3), the mapping rate decreased to 81% (Fig. 1c). Thus, byproducts and off-target amplicons were rare, and the protocol amplified target genes irrespective of cell type and in primary cells.

To quantify TAP-seq's molecular sensitivity, library complexity and enrichment performance, we sequenced TAP-seq and whole

transcriptome libraries close to saturation. The number of molecules observed in TAP-seq was generally higher than expected from the whole transcriptome analysis (Fig. 1d), and the same on-panel capture efficiency was achieved by TAP-seq at a 10- to 100-fold lower sequencing depth (Fig. 1e and Extended Data Fig. 4a, b). Detection efficiency of lowly expressed genes was increased (Fig. 1f and Extended Data Fig. 4c). Together, these analyses show that the targeted PCR efficiently captures target genes without leading to a decrease in library complexity.

The reproducibility of TAP-seq between different replicate experiments with the same panel remains high (Pearson's $R = 0.98$, Extended Data Figure 5a). Pearson correlations of absolute gene quantification between TAP-seq and whole transcriptome sequencing were between 0.56 and 0.86, as expected from highly multiplexed PCR (Extended Data Fig. 5b). For comparison, bulk RNA-seq and whole transcriptome 10X Genomics data displayed a Pearson correlation between 0.60 and 0.63 on the respective genes (Extended Data Fig. 5c). As shown below (Fig. 2), TAP-seq allows highly sensitive detection of relative gene expression changes, and such measures do not depend on absolute expression level detection-accuracy. Using primer redesign and replacement experiments, we found that, given accurate annotations of polyadenylation sites, no manual optimization of primers is necessary and TAP-seq is robust with regard to variations of primer sequence (see Supplementary Note 1).

TAP-seq sensitively detects gene expression changes. To establish TAP-seq as readout for functional genomics screens, we generated a test dataset with a known ground truth. We infected K562 cells expressing nuclease-dead Cas9–Krüppel-associated box (dCas9–KRAB) with a pool of lentiviruses carrying 56 gRNA sequences[17–19] targeting 1 of 10 promoters or 4 well-described enhancers, as well as 30 control gRNAs (Fig. 2a and Supplementary Table 2, sheet 'Chr. 8 control library'). These perturbations were designed to specifically downregulate their known target genes in cis and exerted widely varying effect sizes (Extended Data Fig. 6a,b), allowing a quantitative benchmarking of differential expression tests. For TAP-seq, we used a primer panel including the known promoter/enhancer target genes, 60 presumably unrelated genes of similar expression level (panel 2), and the gRNA sequence from the CROP-seq[6] vector.

In the resulting dataset, each gRNA was covered by a median of 268.5 cells and 0.8 million reads (Fig. 2b). From the same pool of cells, we generated whole transcriptome Perturb-seq data, covering each gRNA with a median of 415.5 cells and 8 million reads (Fig. 2b). TAP-seq captured gRNA identity with an efficiency of 95%, as compared to 39% in the Perturb-seq setting or 89% in Perturb-seq with additional targeted gRNA amplification[20] (Fig. 2c and Extended Data Fig. 6c,d). Therefore, TAP-seq increased the fraction of informative cells available for differential gene expression testing over a classical Perturb-seq design.

Despite lower cell numbers and 10-fold lower total sequencing depth, TAP-seq outperformed Perturb-seq in differential expression testing: 48 out of 56 (86%) perturbations (that is gRNAs leading to an expression change of a promoter or enhancer target) were identified from TAP-seq data, whereas only 41 perturbations (73%) could be identified with Perturb-seq (Fig. 2d,e; see Extended Data Fig. 7 and Supplementary Note 2 for the choice of statistical test). Notably, Perturb-seq missed many enhancer perturbations that elicited weak effects (Fig. 2d and Extended Data Fig. 6b). For two genes (*CCNE2* and *PHF20L1*), statistical power in TAP-seq was lower than in Perturb-seq; these genes had disproportionally little coverage in TAP-seq, indicative of a lower amplification efficiency (Extended Data Fig. 6e and Supplementary Note 1).

To quantitatively compare the power of TAP-seq and Perturb-seq for differential expression testing, we downsampled to specific numbers of reads per cell, and numbers of cells per gRNA. We then computed the recall (or sensitivity) of the assay as the fraction of true gRNA–target gene pairs identified at a given read and cell coverage, as well as its precision, that is the fraction of positives that are true positives. Of note, precision is underestimated in these analyses, since true off-target effects are classified as false positives. We then derived precision–recall curves (Fig. 2f and Extended Data Fig. 8a,b) and computed the area under the curve (AUPRC) as a performance measure (Fig. 2g). These analyses demonstrated that the same performance in differential gene expression testing is achieved in TAP-seq at a 19- to 49-fold lower read depth than that of Perturb-seq (Fig. 2g). Importantly, for the analysis of the Perturb-seq experiment, only genes that were part of the 74 genes from the target panel were tested for differential expression. When the analysis was performed across all detected genes, precision dropped drastically owing to an excessive number of false-positive hits (Fig. 2f,g). These results emphasize the need for hypothesis-driven analysis in Perturb-seq and related experimental designs.

We quantified the absolute expression change required for an effect to be detected to be approximately 0.03 UMI per cell, corresponding to approximately 1 molecule per cell (Extended Data Fig. 8c–f and Supplementary Note 2). Further analyses of the ground truth dataset allowed us to derive recommendations for the experimental design of TAP-seq screens, provided in Supplementary Note 2.

Function-based enhancer–target gene maps for 2.5% of the human genome. To demonstrate the performance of TAP-seq at scale, we set out to generate function-based enhancer–target gene maps in K562 cells, covering 2.5% of the human genome. Since enhancers predominantly affect genes in their proximity, assigning enhancer–target gene pairs (ETPs) is a well-suited task for TAP-seq. To comprehensively test all possible interactions within genomic regions, we perturbed all 1,778 putatively active enhancers predicted on the basis of ENCODE data[21] in two regions on chromosome 8 and 11, and identified effects on expressed protein-coding genes within the same regions (Fig. 3a,b). Thus, in each cell, 68 (chromosome 8) or 79 (chromosome 11) target genes were measured.

Four gRNAs were designed per target enhancer and introduced into K562 cells at a low multiplicity of infection, followed by selection for stable expression (Extended Data Fig. 9a,b and Supplementary Table 2). An average of 37 cells per gRNA or 143 cells per enhancer were profiled, for a total of 7,055 gRNA perturbations in 231,667 cells (Fig. 3c). For each enhancer and gRNA, we identified differentially expressed genes using the statistical test established in the preceding section. We observed a total of 81 significant cis-acting ETPs, involving 24–32% of the genes profiled within the respective genomic region, and 4.4% of the tested enhancers (Fig. 3d and Supplementary Table 3). A previous study using Perturb-seq identified enhancers for only ~3.5–6% of expressed genes[9], underscoring the higher sensitivity of TAP-seq. For the analyses below, we further classified our hits into 36 'strong' ETPs supported by at least 50% of the gRNAs targeting the enhancer, and 45 'weak' ETPs (Extended Data Fig. 9c).

To gauge the ability of TAP-seq to identify bona fide enhancers, we compared our results with published enhancer–target gene associations (Fig. 3b and Extended Data Fig. 9d). For example, we validated four out of five enhancers identified in a CRISPRi screen for *HBE1* enhancers using a fluorescent reporter[22] (Fig. 3b).

In an analysis of the whole dataset, we observed several trends. First, most enhancers, and virtually all enhancers from strong ETPs, are located close (10–50 kb) to their respective target gene's transcriptional start site (TSS, Fig. 3e and Extended Data Fig. 9e). However, in 48% of cases (33% of strong ETPs), at least one other gene was located between the enhancer and target TSS (Extended Data Fig. 9f); these genes were on average expressed 1,000-fold lower than the true enhancer target (Extended Data Fig. 9g). Second, interaction frequencies from Hi-C were substantially elevated in strong and

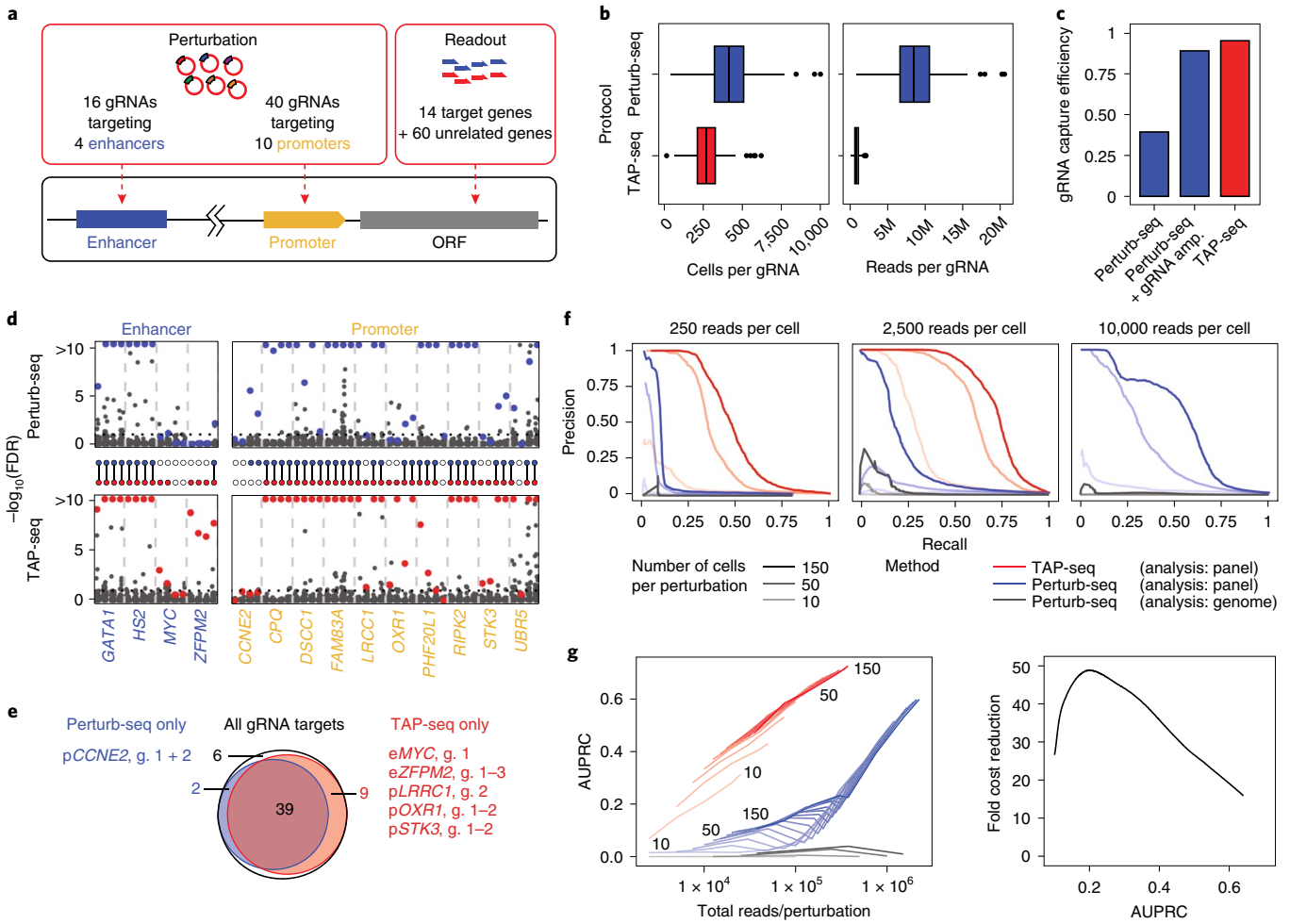


Fig. 2 | TAP-seq sensitively detects gene-expression changes. a, Illustration of the experimental design. **b**, Number of cells and reads for each gRNA in whole-transcriptome and TAP-seq experiments. $n = 86$ gRNAs per group. See 'Data visualization' in the Methods for definitions of box-plot elements. M, million. **c**, gRNA capture efficiency was computed using a generative model that takes into account multiplicity of infection[4,20]. See also Extended Data Fig. 6d. **d**, Benjamini–Hochberg-adjusted *P* values from differential expression tests, comparing cells carrying a given gRNA and cells carrying a non-targeting ('scrambled') control. For each target (*x* axis labels), four guides (columns) were analyzed separately. Colored dots correspond to target genes, dark gray dots to all other genes in the panel. *HBE1* was analyzed as the target gene for HS2 enhancer. See Supplementary Note 3 on 'Sensitivity analysis (differential expression)' for the statistical test used. A total of $n = 60{,}106$ cells were included into the tests; see **b** for the distribution of cells per gRNA. FDR, false-discovery rate. **e**, Venn diagram comparing gRNA targets identified by TAP-seq and whole-transcriptome readout. g., gRNA ID; p, promoter; e, enhancer. **f**, Data from both methods were downsampled to various average read depths per cell and 10–150 cells per gRNA (line opacity). For each sampling run, differential expression testing was performed relative to 500 cells containing scrambled gRNAs. Precision–recall curves were computed assuming that the intended gRNA targets constitute the full set of true positives. See Extended Data Fig. 8a,b for an analysis accounting for off-target effects. **g**, AUPRCs are plotted as a function of sequencing depth (left). The fold cost reduction was estimated from the difference in sequencing depth quantified as a function of desired AUPRC (right). Data from $n = 60{,}106$ cells and 9,750 sampling runs; see Supplementary Note 3 section 'Sensitivity analysis (differential expression)'. Line colors and opacity are as in **f**.

weak ETPs (Fig. 3f) and contained more information on the potential to form functional interactions than linear distance (Fig. 3h). Third, enhancers of strong ETPs were enriched for active chromatin marks H3K27ac, H3K4me1, H3K4me3 and RNA polymerase II compared with non-significant ETPs with a similar distance to TSS distribution (Fig. 3g and Supplementary Table 4).

We next integrated these results into machine-learning models to predict ETPs. In a quantitative analysis across all genes with at least one associated enhancer, a joint model (of linear distance, interaction frequency and epigenome features) identified 75% of all 61 ETPs from 3,781 potential ETPs with an enhancer–TSS distance of below 300 kb at a precision of 50% (Fig. 3h; see Extended Data Fig. 9h for an analysis of 34,493 potential ETPs across the entire data set). A model trained on data from chromosome 11 efficiently predicted ETPs on chromosome 8, and vice versa (Fig. 3i). When trained on all data from this study, our model predicted ETPs identified in a Perturb-seq study[9] (75% recall at a precision of 40%); the converse did not work effectively (Fig. 3i and Extended Data Fig. 9i). Together, these analyses demonstrate that a model based on enhancer activity and contact frequency can predict ETPs genome-wide. Furthermore, a model trained on TAP-seq data allows more accurate genome-wide predictions than those from a model trained on Perturb-seq data.

TAP-seq identifies cell types and differentiation states with shallow sequencing. Single-cell perturbation screens are also used to define changes in cell type abundance and cell state[3,5,6]. Typically, this is performed with whole-transcriptome readouts, but poten-

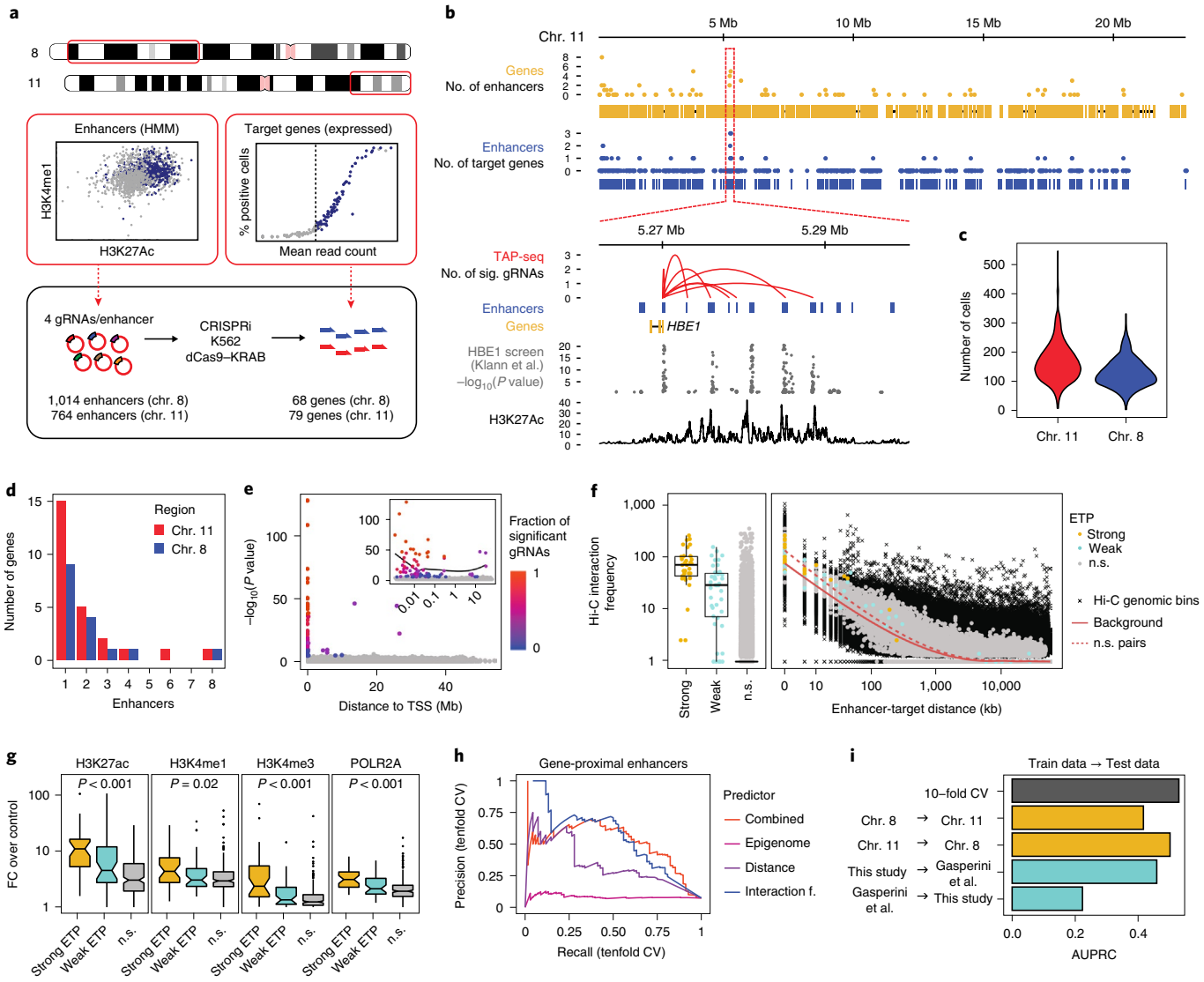


Fig. 3 | A perturbation-based screen of enhancer targets across 2.5% of the human genome. a, All enhancers in two regions of chromosomes 8 and 11, as determined by the GenoSTAN HMM[21], were selected, and four gRNAs were designed for each enhancer. All expressed genes on the same genomic regions, except *HBG1* and *HBG2*, were selected for targeted readout. Highly expressed *HBG1* and *HBG2* were omitted from the screen to achieve a higher cost efficiency. **b**, Top: the number of enhancers per gene (yellow), and the number of genes per enhancer (blue) across the selected region on chromosome 11. Bottom: zoom-in on the *HBE1* locus. Enhancers are connected to target genes via red arcs. Results are compared with *P* values from a fluorescence-activated cell sorting-based CRISPRi screen of enhancers regulating *HBE1* (ref. [22]), as well as H3K27Ac ChIP-seq signal. *P* values are from linear regression as described in ref. [22]. **c**, Number of cells profiled per enhancer perturbation depicted as a violin plot ($n = 1{,}790$ enhancer perturbations). **d**, Number of enhancers identified for the 147 target genes. For 106 genes, no candidate enhancer was identified. **e**, The distance between candidate enhancer and target gene TSS is plotted against the TAP-seq *P* value. The fraction of individual gRNAs per enhancer causing significant effects is color coded. Inset shows *x* axis in logarithmic scale. A total of $n = 231{,}667$ cells were included in the tests; see **c** for the distribution of cells per perturbation and Supplementary Note 3, 'Enhancer screen analysis' section for a description of the statistical test used. **f**, Relationship between Hi-C interaction frequency[39], linear distance of the enhancer to the TSS of the target gene and TAP-seq result. 'Strong' ETPs are supported by at least 50% of candidate gRNAs for a given enhancer (see main text section 'Function-based enhancer-target gene maps for 2.5% of the human genome'). Hi-C genomic bins are all measured interactions within the genomic regions used to estimate expected background interaction frequencies; see Supplementary Note 3, 'Hi-C and chromatin analyses' section for details. See Methods, 'Data visualization' section for definition of box plot elements. n.s., not significant. **g**, Relationship between ChIP-seq signal of various chromatin marks[40] and TAP-seq results. 'Strong' enhancers have at least one target gene supported by at least 50% of gRNAs. *P* values are from Kruskal–Wallis tests assessing overall differences between ETP classes; see Supplementary Table 4 for *P* values of pairwise comparisons and number of samples per group. **h**,**i**, Performance of machine-learning-based classifiers in predicting ETPs. Only genes with at least one enhancer were included. **h**, Random forest classifiers were trained on the indicated set of features and performance was assessed in a tenfold cross-validation (CV) scheme. **i**, Random forest classifiers were trained on a given dataset (train data), and performance of the classifier on another dataset (test data) was tested. Areas under the precision–recall curves are shown (see Extended Data Fig. 9i for curves). The dataset from Gasperini et al.[9] was used for comparison.

tially could be done with much fewer reads. Here, we applied TAP-seq to murine bone marrow (BM) to evaluate its ability to distinguish cell types, including similar progenitor cell states from immature, c-Kit$^+$ BM[23,24]. Using a whole-transcriptome reference dataset of total and immature BM cells[25], we identified 184 genes which optimally distinguished 18 different cell types (Fig. 4a, see

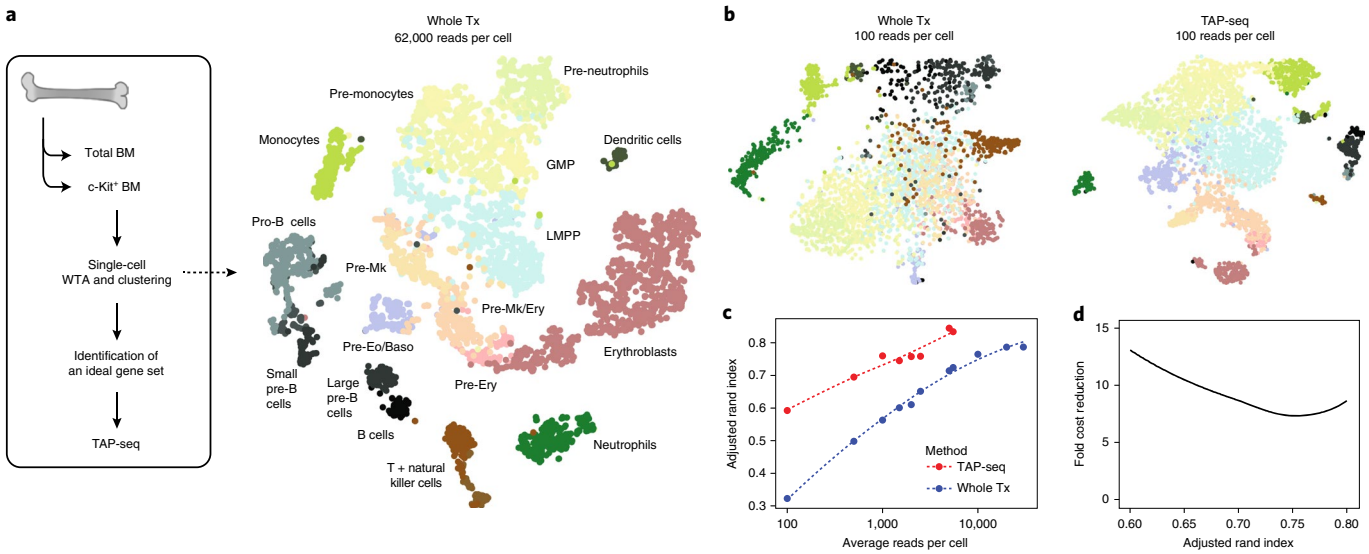


Fig. 4 | TAP-seq permits efficient identification of cell types and differentiation states at very low read depths. a, Whole-transcriptome sequencing data from mouse total BM and c-Kit+ BM[25] (GEO GSE122465) was projected using *t*-distributed stochastic neighbor embedding (t-SNE; right). Cell type annotations were taken from ref. [25]. GMP, granulocyte or monocyte precursor; LMPP, lymphoid-primed multipotent progenitor; Mk, megakaryocyte; Ery, erythroid; Pre-Eo/Baso, pre-eosinophil–basophil. A maximally informative gene set allowing to distinguish all cell types involved was then identified using LASSO (see Methods, section 'TAP-seq target gene selection'). $n = 4{,}957$ cells. **b**, t-SNE projection of whole transcriptome and TAP-seq data, downsampled to an average read depth of 100 reads per cell. Color indicates the cell type identified in non-downsampled data (see **a** for color code). $n = 4{,}957$ cells (Whole Tx) or 11,794 cells (TAP-seq). **c**, Data from both methods were downsampled to various average read depths, and unsupervised clustering was performed using the Seurat pipeline[26]. Average read depth per cell is plotted against the overlap between clusters identified in downsampled data and reference clusters, as quantified by the adjusted Rand Index. **d**, The fold difference in sequencing reads between TAP-seq and whole transcriptome is plotted as a function of Rand Index.

Methods for gene selection using a whole-transcriptome reference). TAP-seq libraries were then generated from the same BM fractions included in the reference dataset. Cell types were identified by unsupervised clustering and marker gene expression, and confirmed via label transfer[26] from the whole-transcriptome reference (Extended Data Fig. 10a,b).

To compare the ability of TAP-seq and whole-transcriptome readout to distinguish between cell states at lower sequencing depths, gene expression data were downsampled to defined average read depths. Unsupervised clustering was then applied and compared with ground truth labels obtained from our deeply sequenced reference. The results show that only 100 average reads per cell reliably distinguish cell types and differentiation stages in TAP-seq (Fig. 4b). A quantitatively similar performance is therefore achieved at a 7- to 12-fold lower sequencing depth (Fig. 4c,d). Similar decreases in sequencing requirements were obtained when using supervised methods for cell type identification (that is label transfer from a reference data set)[26] (Extended Data Fig. 10d–f). In summary, TAP-seq outperforms non-targeted approaches in cell type mapping, and reliably identifies cell types and differentiation states at an extremely low sequencing depth.

Discussion

In conclusion, targeted Perturb-seq (TAP-seq) enables genome-scale pooled CRISPR screens that use transcriptomics as a readout. By focusing sequencing coverage on genes of interest, TAP-seq lowers sequencing requirements up to 50-fold, and solves the multiple-hypothesis testing problem typically encountered in whole transcriptome screens. It increases the sensitivity towards small expression changes and lowly expressed genes, and retrieves the gRNA identity efficiently. Unlike previous high-throughput[27–29] and low-throughput[30] targeted single-cell RNA-seq methods, its modular design ensures platform-independent, robust genetic screens.

TAP-seq is applicable to a broad range of functional genomics applications, including studies of gene regulatory networks[7,19,31], signaling pathways[6,32], and combinatorial regulatory logic[7,32,33], where phenotypes of interest are represented by expression changes in small sets of genes. When applied to more complex transcriptomic signatures, for example of cellular differentiation/state or immune cell activation, informative sets of genes can be defined a priori using a whole-transcriptome reference. Alternatively, gene panels such as the L1000 panel[34] could be used (Extended Data Fig. 3), which has been shown to capture the majority of information contained in the full transcriptome across a wide range of perturbations[34]. We make L1000 primers for TAP-seq available upon request.

We applied TAP-seq to generate dense enhancer–promoter interaction maps within 2.5% of the human genome, screening over 7,000 distinct CRISPRi perturbations. Previous applications of Perturb-seq for enhancer targets[7,9] had remained limited to large effect sizes and identified enhancers only for a small fraction of candidate genes (3.5–6% in ref. [9], compared with 24–32% reported here). We identified ~80 new enhancer–target gene pairs and found that physical proximity in 3D, and to a smaller extent chromatin activity, determine the regulatory potential of an enhancer. This allowed us to derive quantitative rules of enhancer–target gene interactions.

A potential caveat of these analyses is that dCas9–KRAB does not specifically inactivate enhancers, but rather inhibits transcriptional activity both at promoters and enhancers[35]. Since targeting dCas9–KRAB to gene-proximal sites with lower levels of H3K27 acetylation did not induce a measurable effect on the gene, the CRISPRi effect does not simply spread to the promoter along linear DNA; also, enhancers within 1 kb of the target TSS were excluded. In future studies, results from CRISPRi-screens using dCas9–KRAB could be compared to dCas9–LSD[35], or expanded by potentially stronger dual-activator and dual-inhibitor constructs[36].

TAP-seq scales single-cell genetic screens to thousands of gRNA-mediated perturbations. On the basis of the analysis of the comprehensive test dataset with known ground truth, we derive recommendations on experimental design and required sequencing depth (Supplementary Note 2). In the future, a further increase in throughput can be achieved by combining TAP-seq with ongoing developments of ultra-high-throughput single-cell capture strategies[37,38]. Taken together, targeted Perturb-seq enables genetic screens at unprecedented resolution and throughput.

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/s41592-020-0837-5.

Received: 24 September 2019; Accepted: 15 April 2020;
Published online: 1 June 2020

References

1. Steinmetz, L. M. et al. Systematic screen for human disease genes in yeast. *Nat. Genet.* **31**, 400–404 (2002).
2. Winzeler, E. A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
3. Shifrut, E. et al. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell* **175**, 1958–1971.e15 (2018).
4. 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).
5. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896.e15 (2016).
6. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
7. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol. Cell* **66**, 285–299.e5 (2017).
8. 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).
9. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
10. Van Der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
11. van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
12. Cuomo, A. S. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
13. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
14. Krivega, I. & Dean, A. Enhancer and promoter interactions—long distance calls. *Curr. Opin. Genet. Dev.* **22**, 79–85 (2012).
15. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
16. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
17. Horlbeck, M. A. et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**, e19760 (2016).
18. Hilton, I. B. et al. Epigenome editing by a CRISPR–Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).
19. Fulco, C. P. et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
20. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
21. Zacher, B. et al. Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* **12**, e0169249 (2017).
22. Klann, T. S. et al. CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
23. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
24. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
25. Baccin, C. et al. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.* **22**, 38–48 (2020).
26. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
27. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
28. Shum, E. Y., Walczak, E. M., Chang, C. & Fan, H. C. in *Advances in Experimental Medicine and Biology* 63–79 (Springer, 2019).
29. Vallejo, A. F.et al. Resolving cellular systems by ultra-sensitive and economical single-cell transcriptome filtering. Preprint at *bioRxiv* https://doi.org/10.1101/800631 (2019).
30. Uzbas, F. et al. BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis. *Genome Biol.* **20**, 155 (2019).
31. Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S. & Smith, A. G. Defining an essential transcription factor program for naïve pluripotency. *Science* **344**, 1156–1160 (2014).
32. Horn, T. et al. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nat. Methods* **8**, 341–346 (2011).
33. Zetsche, B. et al. Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2016).
34. Subramanian, A. et al. A next generation connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452.e17 (2017).
35. Kearns, N. A. et al. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. *Nat. Methods* **12**, 401–403 (2015).
36. Li, K. et al. Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nat. Commun.* **11**, 485 (2020).
37. 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).
38. Datlinger, P. et al. Ultra-high throughput single-cell RNA sequencing by combinatorial fluidic indexing. Preprint at *bioRxiv* https://doi.org/10.1101/2019.12.17.879304 (2019).
39. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
40. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

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. 2020

Methods

Vectors and cloning strategies. Lentiviral packaging vectors pMD2.G and psPAX2 were a gift from D. Trono (Addgene plasmids nos. 12259 and 12260). CROPseq-Puro was a gift from C. Bock (Addgene plasmid no. 86708). UCOE–SFFV–dCas9–BFP–KRAB was a gift from J. Weissman (Addgene plasmids nos. 60955 and 85969).

CROPseq-Puro-F+E was cloned by replacing the original tracrRNA with the optimized F+E tracrRNA sequence[41] using site-directed mutagenesis and Gibson assembly. One whole CROPseq-Puro spanning PCR was done with overlapping mutagenesis primers forward 5′-TGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAA TAAGGCTAGTCCGTTATCAACTTGAAAAAG-3′ and reverse 5′-TATTTAAACTTGCTATGCTGTTTCCAGCATAGCTCTTAAACA GAGACGTACAAAAAAGAGCAAGAAG-3′ using LongAmp DNA polymerase (NEB). The PCR product was DpnI digested to deplete residual unamplified circular CROPseq-Puro template. The digested and gel-purified PCR product was then circularized with Gibson assembly Mastermix (NEB) to generate CROPseq-Puro-F+E.

Cloning of individual sgRNAs. Single guide RNA (sgRNA) sequences targeting *MYC*, *GATA1*, *ZFPM2* and HS2 enhancers were described previously[18,19] and sgRNAs targeting *ZFPM2* enhancer[42] were designed in Benchling. All protospacer sequences are shown in Supplementary Table 2. sgRNAs were cloned into CROP-seq vectors using Gibson assembly or restriction ligation.

CROPseq-Puro or CROPseq-Puro-F+E was digested with BsmBI (NEB) and the backbone fragment (8.3 kb) was gel extracted. sgRNAs were ordered as single-stranded DNA (ssDNA) forward and reverse oligonucleotide pairs with 35-nt 5′ homology (5′-GGCTTTATATATCTTGTGGAAAGGACGAAACACCG-3′, with the last G corresponding to the sgRNA transcriptional start G), 19 nt sgRNA sequence and 3′ homology to CROPseq-Puro (5′-GTTTTAGAGCTAGAAA TAGCAAGTTAAAATAAGGC-3′) or CROPseq-Puro-F+E (5′-GTTTAAG AGCTATGCTGGAAACAGCATAGCAAGTT). Oligonucleotides were mixed at an equimolar ratio and annealed by heating to 95 °C for 5 min and ramping to 4 °C over 5 min. Annealed oligonucleotides were cloned into CROPseq-Puro or CROPseq-Puro-F+E using Gibson Assembly Master Mix (NEB) or NEBuilder HiFi DNA Assembly Master Mix (NEB) at a backbone:insert ratio of 1:20. Reactions were transformed into NEB stable chemically competent *E. coli* and were grown out at 30 or 32 °C to reduce the rate of plasmid recombination.

For sgRNA cloning using restriction ligation, oligonucleotides were ordered with sticky-end overhangs matching the BsmBI digested CROPseq-Puro and CROPseq-Puro-F+E backbone, forward and reverse oligonucleotides were annealed and phosphorylated using T4 PNK and cloned into digested and dephosphorylated backbones using T4 DNA Ligase (NEB). Transformation was done as described above.

Enhancer-targeting sgRNA design. For the enhancer screen in Fig. 3, candidate enhancers were defined as DNase hypersensitive sites with a histone modification pattern indicative of active enhancers. For each candidate enhancer, we included four gRNAs in the screen (Supplementary Table 2). In detail, K562 DNase hypersensitive (HS) hotspots were obtained from the ENCODE reference epigenome[40] (experiment ENCSR921NMD) and chromatin annotations were obtained by GenoSTAN[21], a chromatin state HMM model. DNase HS hotspots that overlapped with GenoSTAN active enhancer annotations (Enh.15, EnhWF.2, EnhF.10) by at least 50 bases were included as candidate enhancers. For each enhancer, candidate gRNAs were identified based on the presence of a protospacer-adjacent motif and ranked by putative on-target activity, as determined using the activity model from ref. [17]. Candidate gRNAs were then aligned to the genome to determine off-target effects: Guides with alignment near a TSS were always excluded. Guides with alignment to other genomic positions were excluded, if a sufficient number of guides with positive activity scores exist. After these filtering steps, the four gRNAs with the highest activity scores were selected.

Cloning of sgRNA libraries. Enhancer-targeting libraries used in Fig. 3 were ordered as 89-nt ssDNA SureGuide high-fidelity libraries from Agilent, including 35-nt 5′ and 3′ homologies matching CROPseq-Puro-F+E. All protospacer sequences are shown in Supplementary Table 2. Libraries were amplified using 5′-GTATTTCGATTTCTTGGCTTTATATATCTTGTGG-3′ and 5′-GACTAGCCTTATTTAAACTTGCTATGCTGTTTC-3′ extending the homologies to 50 bp. Amplification was done with Q5 Hot Start HiFi 2× Master Mix (NEB), 0.5 µM forward primer, 0.5 µM reverse primer and 2 nM sgRNA library using one cycle at 95 °C for 3 min, then 14 cycles at [98 °C for 10 s, 51 °C for 15 s, 72 °C for 10 s], and a last cycle at 72 °C for 1 min. Optimal cycle number was determined as the highest cycle number before the appearance of concatemeric PCR products on a High Sensitivity dsDNA Bioanalyzer. Amplified sgRNA libraries were cloned into CROPseq-Puro-F+E using NEBuilder HiFi DNA Assembly Master Mix (NEB) at a backbone:insert ratio of 1:20. Assembled libraries were transformed into electrocompetent Lucigen Endura *E. coli* and grown out on plates for 16 h at 30 °C for ≥100× library coverage. Colonies were pooled and plasmid preparations were done using Zymopure II Plasmid Maxiprep kit (Zymo Research). In parallel, >20 colonies were picked per library and Sanger validated.

Libraries targeting known enhancers and promoters (see Supplementary Table 2) as used in Figs. 2 and 3 were generated from ssDNA oligonucleotides, which were annealed and amplified in separate reactions, followed by pooled Gibson cloning into CROPseq-Puro-F+E as described above. sgRNA sequences targeting promoters and non-targeting sgRNAs were taken from CRISPRi v2 libraries[17], sgRNA sequences targeting *MYC*, *GATA1*, *ZFPM2* and HS2 enhancers were described above.

Cell culture. HEK 293FT cells were maintained in DMEM (Thermo, 41965039) supplemented with 10% FBS, 2 mM L-glutamine, 1 mM sodium pyruvate, 0.1 mM NEAA, 500 µg mL^{-1} G418 and penicillin–streptomycin (P/S) (Sigma Aldrich and Thermo). K562 (from ATCC, CCL-243) and stable clones thereof were cultivated in RPMI-1640 (Thermo 21875034) supplemented with 10% FBS and P/S. NIH 3T3 cells were cultivated in DMEM (Thermo, 41965039) supplemented with 10% FBS and P/S. Mouse embryonic stem cells (mESC) were cultivated under FCS/LIF conditions as described[43].

Lentivirus production. Lentivirus was produced in HEK 293FT cells. Cells were grown to 90–95% confluency in medium without G418 in 6-well plates. Lentiviral packaging vectors pMD2.G and psPAX2 and the transfer plasmid were mixed 1:1:1 and transfected using Lipofectamine 3000 (Thermo). At 6 h post-transfection, cells were split into a 10-cm plate in medium without G418. At 3 d post-transfection, cell culture supernatant was collected, cleared through a 0.45-µm filter, and lentivirus was 20× concentrated with Lenti-X concentrator (Takara/Clontech).

Lentiviral transduction and generation of stable cell lines. For stable selection of K562 and K562 dCas9–KRAB cell lines, exponentially growing cells were diluted to 0.5×10^6 cells mL^{-1} and polybrene was added to a final concentration of 10 µg mL^{-1}. Concentrated virus was added to cells or cells were added to virus. At 24 h post infection, cells were collected and resuspended in K562 medium for antibiotic or fluorescence-activated cell sorting selection (FACS).

K562 CRISPRi polyclonal cell line expressing dCas9–KRAB was generated by transduction of K562 with UCOE–SFFV–dCas9–BFP–KRAB at high multiplicity of infection (MOI). At 4 d and again at 7 d after infection, cells were sorted for high BFP expression, and polyclonal cell line was checked regularly for the percentage of BFP-positive cells.

*I*nfection of K562 CRISPRi with CROP-seq vectors was done at low MOI to get mostly single-infected cells. To set MOIs, the timepoint of highest cell death and the maximum percentage of dead cells during Puromycin selection was determined: cell death of around 90% 4 d after infection showed a low number of multiply infected cells as determined from CROP-seq vector reads in TAP-seq and whole transcriptome Perturb-seq data. At 24 h after infection, cells were collected and diluted to 0.5×10^6 cells mL^{-1} with medium containing Puromycin (Thermo) at a final concentration of 2 ng µL^{-1}. Cell growth was monitored during selection by cell counting and viable cell staining using Trypan-blue.

qPCR-based measurement of CRISPRi effects. The CRISPRi effect on known enhancer–target gene pairs (ETPs) *MYC*, *GATA1*, HS2 enhancer and *ZFPM2* was measured using quantitative PCR. K562 dCas9–KRAB polyclonal cells infected with CROP-seq vectors were collected 10 to 14 d after infection and puromycin selection. Supernatant was removed and total RNA was prepared using NucleoSpin RNA Plus (Macherey Nagel). RT was performed using DNaseI-digested RNA and SuperScript II. qPCR was done with SYBR-Green PCR master mix (Thermo). Data were processed using the $-\Delta\Delta$Ct method, and standard deviations of the $2^{-\Delta\Delta Ct}$ values from biological replicates were calculated. qPCR primer sequences: GAPDH forward 5′-TGGTATCGTGGAAGGACTCATGAC-3′, GAPDH reverse 5′-ATGCCAGTGAGCTTCCCGTTCAGC-3′. ZFPM2 primer sequences were taken from ref. [42], *MYC, GATA1* and *HDAC* from ref. [19] and hemoglobin gene primers from ref. [18].

Primary cell samples. For BM samples and lung samples, C57BL/6 mice were bred and housed under pathogen-free conditions at the central animal facility of the German Cancer Research Center (DKFZ) or EMBL.

BM was extracted and cells were sorted as described previously[25]. In short, femurs, tibiae, hips and spines were dissected and cleaned from surrounding tissue. Bones were crushed in cell suspension medium and dissociated cells were filtered, red blood cell lysis was performed and cells were lineage-depleted using Dynabeads Untouched Mouse CD4 Cells Kit. Lineage-depleted bone and BM cells obtained following crushing and digestion were stained with FACS c-Kit antibody, and sorted as described. For the isolation of neutrophils, single viable cells were isolated using forward/sideward scatter and DAPI dye exclusion, and CD11b$^+$Ly6G$^+$ neutrophils were isolated.

Lung mesenchymal cells were provided by the group of C. Scholl, DKFZ. In short, lungs were perfused with cold PBS and instilled with collagenase I. Lungs were removed while the trachea was clamped. The lungs were cleaned from non-respiratory tissue and minced. Following depletion of erythrocytes, mesenchymal cells were sorted as DAPI$^-$Epcam$^-$CD45$^-$CD31$^-$Pdgfra$^+$Sca1$^-$Npnt$^+$.

Preparation of cells for scRNA-seq. After puromycin selection for 14 d, K562 dCas9–KRAB cells infected with CROP-seq vectors were collected and resuspended in PBS. FACS was carried out on a BD FACSAria Fusion flow cytometer (BD Biosciences) selecting single cells using forward and sideward scatter, viable cells using Draq7 dye exclusion and dCas9–BFP–KRAB intermediate- to high-expressing cells using BFP signal. Sorted cells were collected by centrifugation for 9 min at 200*g* (RT) and resuspended in PBS to 1×10^6 cells mL^{-1}, as determined by cell counting. Cells were stored on ice for up to 1 h. Cell counts were double-checked directly before loading on the 10X Genomics Chromium controller.

For NIH 3T3, mESCs, mouse lung mesenchyme and mouse neutrophils, defined numbers of viable cells, as measured using dye exclusion with DAPI or Draq7, were sorted and pooled for 10X Genomics.

scRNA-seq with whole-transcriptome readout. Whole-transcriptome single-cell 3′ RNA-seq libraries were generated using 10X Genomics Chromium 3′ reagent kit v2 according to the manufacturer's guidelines. Cell input was set to a targeted cell recovery of 8,000 cells per lane. Sample indexing was done using i7 Multiplex Kit (10X Genomics).

TAP-seq target gene selection. For the target gene panels 1 and 2 (used in Figs. 1–3), genes were selected on the basis of chromosomal location, while excluding non-expressed genes. For the large-scale enhancer screen, the highly expressed genes *HBG1* and *HBG2* were omitted from the panels to achieve higher cost efficiency. The K562 add-on panel (Fig. 1c and Extended Data Fig. 2a) was designed to target 48 genes expressed in K562 which are downstream targets of Oct4, as part of an experiment not included into this manuscript. Target genes are listed in Supplementary Table 1.

For the target gene panel 3 (used in Fig. 1 and Extended Data Figs. 2 and 3) 150 genes were selected for relatively uniform expression between NIH 3T3 cells, mESCs, lung mesenchyme and neutrophils. Average expression of all genes was calculated for each cell type using available whole-transcriptome single-cell RNA-seq data (refs. [15,25,44,45]), and used to calculate mean values and coefficients of variation across the four cell types. Mean values were stratified into 30 bins, and for each bin, 5 genes from the lowest decile of CV were randomly selected. Additionally, 12 genes with highly specific expression to each cell type were added to the panel.

For the BM experiment in Fig. 4, target genes were identified using a machine-learning-based workflow. Whole-transcriptome data from total BM and c-Kit$^+$ BM from mouse was downloaded from NCBI GEO (GSE122465, ref. [25]), and the original cell type annotation from ref. [25] was used. For each cell type, we then identified the 20 differentially expressed (DE) genes with the highest log-fold change compared with all remaining cell types to create a list of 267 candidate genes. We then used a generalized linear model of the multinomial family with a LASSO penalty to select an optimal set of genes whose expression distinguished all populations. Four final gene panels were compared in their ability to distinguish cell types, using a cross-validation scheme: (a) 238 genes selected using an optimal lambda penalty parameter, as determined by tenfold CV; (b) 126 genes selected using a more stringent parameter; (c3) a list of 92 genes provided by an expert hematologist (S. Haas, DKFZ Heidelberg); and (d) the union of (b) and (c) (192 genes). While (a), (b) and (d) provided a similar well type classification performance (Cohen's Kappa 0.8–0.85 in tenfold CV), (c) performed worse (Cohen's kappa, 0.65). (d) was selected because it contained several genes considered informative for the annotation of cell types (Fig. 4b). Primers were successfully designed for 184 of the 192 target genes. Target gene selection for cell types using a whole-transcriptome reference is implemented in the TAP-seq R package for primer design.

All primer panels used in this study are available from the corresponding authors upon request.

TAP-seq primer panel design. A customized pipeline based on Primer3 (refs. [46,47]) was devised to design targeted PCR primers. Primers were designed based on protein-coding exon annotations for the target genes (Gencode GRCh38.89). Available Drop-seq or 10X Genomics scRNA-seq data were used to identify expressed isoforms and likely polyadenylation (polyA) sites for every target gene of interest. If multiple polyA sites were found, the most downstream site was chosen. The most downstream annotated 3′ end was used if no polyA site could be identified. If several transcript isoforms overlapped with the inferred polyA sites or no polyA sites were found, their annotations were merged (union) to form a consensus annotation for primer design. Transcript sequences were extracted (hg38) for all target genes and nested forward primers were designed around the 3′ end: outer forward primers were selected to result in 350- to 500-bp-long amplicons, which were then used to design inner forward primers that yield 150- to 300-bp-long amplicons. The Drop-seq reverse primer used for all PCR reactions was provided to optimize compatibility. Ten forward and reverse primers were designed per target gene and aligned against a human genome and transcriptome database using blast to identify potential off-target annealing. The inner and outer primers with the fewest exonic, intronic and intergenic off-target alignments (in that order) were selected for every gene, and all primers were assessed for multiplex suitability with Primer3's check_primers functionality. Primers were synthesized as separate oligonucleotides and purified by desalting (Sigma Aldrich). All TAP-seq gene panel primer sequences used in this study are listed in Supplementary Table 1.

TAP-seq library preparation using 10X Genomics Chromium. The protocol described here was optimized using 10X Genomics 3′ reagent kit v2, but can be used for v3 chemistry, since the oligonucleotide handle sequence (PartialRead1) is identical between v2 and v3. All oligonucleotide sequences described here are shown in Supplementary Table 6. A step-by-step protocol for TAP-seq is provided as Supplementary Protocol, and was made available at Protocol Exchange[48].

Cell barcoding and reverse transcription. 10X Genomics Chromium run for TAP-seq was done with cells sorted as described above and following 10X Genomics user guidelines for 3′ reagent kit v2 with the following modifications: for GEM generation and barcoding, no 10X RT primer (containing the template-switch oligonucleotide) was added to the single-cell master mix (10X v2 protocol step 1.1). To correct for the missing volume, 3.8 µL H_2O or low-TE was added. For large-scale experiments in Fig. 3, the 10X Chromium run was performed with a modified chip loading protocol (adapted from ref. [49]): 28 µl of beads were loaded per 10X lane and 10X RT Enzyme was replaced by 10 µl Superscript IV (Thermo, 200 U μL^{-1}). We then continued with the 10X 3′ reagent kit v2 protocol until GEM-RT incubation (10X v2 Protocol Step 1.5). After GEM-RT incubation, cDNA was cleaned up according to the manufacturer's protocol, eluted using 35 µL Elution Solution I and stored at 4 °C as input for TAP-seq PCR1.

PCR1 and PCR2 with gene-specific outer and inner primers. Pooled gene-specific nested primers were used as forward primers in PCR1 and PCR2. A primer PartialRead1 annealing to the PCR handle on 10X Genomics 3′ beads served as reverse primer for both PCRs. Forward outer and inner primer panels were generated by pooling all single oligos (100 µM in H_2O) in equimolar ratio. CROP-seq vector-derived polymerase II transcripts were amplified with CROPouter and CROPinner primers, which were added in 8× excess relative to each single primer in the outer and inner primer mix. CROPouter and CROPinner primers were not added in mouse BM TAP-seq experiments.

PCR1 was done with 35 µL purified cDNA from 10X GEM-RT, 2.5 µl 100 µM gene-specific outer primer mix, 2 µL 12 µM CROPouter, 4 µL 10 µM PartialRead1 and 50 µl KAPA HiFi Hotstart Readymix in 100 µL total volume. Per 10X Genomics lane, one 100 µL reaction was set up. Cycling conditions were one cycle at 95 °C for 3 min; 11 cycles at [98 °C for 20 s, 67 °C for 60 s, 72 °C for 60 s], one cycle at 72 °C for 5 min, and cooling to 4 °C. Cycle number was increased to 12 cycles for the large-scale enhancer screen in Fig. 3. PCR product was purified using 1.5× AMPure XP (Beckman) with two 80% ethanol washes, elution was done in 30 µL Elution Buffer (Qiagen). Typical yield from PCR1 was 50 to 250 ng per reaction.

PCR2 was done with 10 ng PCR1 product, 2.5 µL 100 µM gene-specific inner primer mix, 2 µL 12 µM CROPinner primer and 4 µL 10 µM PartialRead1 in a 100 µL reaction using KAPA HiFi Hotstart Readymix. Cycling conditions were 95 °C for 3 min, 8 cycles at [98 °C for 20 s, 67 °C for 60 s, 72 °C for 60 s], one cycle at 72 °C for 5 min, and cooling to 4 °C. Product was purified as above and eluted in 30-µL elution buffer. Typical yield from PCR2 was 50 to 100 ng per reaction.

PCR3 with Illumina primers. PCR3 adds Illumina adapters to generate a sequencing-ready library. PCR3 was done using 10 ng PCR2 purified product, 4 µL 10 µM Targeted10X and 2.5 µl 10 µM Illumina reverse primer N7XX in a 100 µL reaction with KAPA HiFi Hotstart Readymix. Cycling conditions were 95 °C at 3 min, 8 cycles at [98 °C for 20 s, 60 °C for 15 s, 72 °C for 45 s], one cycle at 72 °C for 5 min, and cooling to 4 °C. Product was purified as above and eluted in 30 µL elution buffer. Typical yield from PCR3 was 300 to 600 ng per reaction. Product was checked on a High Sensitivity DNA Bioanalyzer (Agilent). Sample bioanalyzer traces are shown in Extended Data Fig. 1.

For Illumina sequencing, TAP-seq produces Illumina sequencing-ready libraries identical to 10X Genomics single-cell 3′ libraries using standard Illumina Read 1 and Read 2 primers. The cell barcode and UMI are encoded in read 1 (26 cycles), and read 2 contains the cDNA fragment (~58 cycles). Sample index sequences are sequenced as i7 index read (8 cycles).

TAP-seq library preparation from Drop-seq. The first steps of the Drop-seq protocol until Exonuclease I treatment were implemented as described[16]. The Drop-seq template-switch oligonucleotide (TSO) was left out to avoid full-length cDNA amplification in parallel to targeted amplification. Starting with 2,000 beads from Exonuclease I treatment, TAP-seq was performed as described above, with the following modification: as reverse primer for PCR1 and PCR2, an oligonucleotide ISPCR was added instead of PartialRead1. As reverse primer for PCR3, Targeted10X was replaced by New-P5 smart PCR hybrid oligo. All oligonucleotide sequences are described in Supplementary Table 6.

Methods for computational data analysis. All computational analysis methods are available in Supplementary Note 3.

Data visualization. All plots were generated using the ggplot2 (v. 3.1.0), gviz (1.24.0) and pheatmap (v. 1.0.10) packages in R 3.5.1. Boxplots are defined as follows: the middle line corresponds to the median; lower and upper hinges correspond to first and third quartiles; the upper whisker extends from the hinge to the largest value no further than 1.5 × IQR from the hinge (where IQR is the interquartile range, or distance between the first and third quartiles); and the lower whisker extends from the hinge to the smallest value, at most 1.5 × IQR of the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually[50].

Statistics. Statistical analyses were performed using R and scripts available at https://github.com/argschwind/TAPseq_manuscript. Statistical details for each experiment are also provided in the figure legends.

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

Data availability

All data are available from GEO GSE135497. A summary of all genomics data used for each figure is provided in Supplementary Table 5. ENCODE bulk RNA- and ChIP-seq data is available from encodeproject.org (experiment IDs: ENCSR545DKY, ENCSR000AKP, ENCSR000EWC, ENCSR000EWA, ENCSR000EWB, ENCSR388QZF, ENCSR921NMD). Hi-C data from ref. [39] are available from GEO GSE63525. Mouse BM single-cell RNA-seq data are available from GEO GSE122465.

Code availability

The TAP-seq R package for primer design is available through Bioconductor (http://bioconductor.org/packages/release/bioc/html/TAPseq.html). Code required to reproduce the analyses of this paper, as well as a pipeline for TAP-seq data processing, is available at https://github.com/argschwind/TAPseq_manuscript and https://github.com/argschwind/TAPseq_workflow.

References

41. Chen, S. et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).
42. Mendenhall, E. M. et al. Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotechnol.* **31**, 1133–1136 (2013).
43. Deliu, E. et al. Haploinsufficiency of the intellectual disability gene SETD5 disturbs developmental gene expression and cognition. *Nat. Neurosci.* **21**, 1717–1727 (2018).
44. Velten, L. et al. Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. *Mol. Syst. Biol.* **11**, 812 (2015).
45. Xie, T. et al. Single-cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis article Single-Cell deconvolution of fibroblast heterogeneity in mouse pulmonary fibrosis. *Cell Rep.* **22**, 3625–3640 (2018).
46. Untergasser, A. et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
47. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
48. Schraivogel, D., Velten, L., Gschwind, A. R. & Steinmetz, L. M. A protocol for Targeted Perturb (TAP)-seq and targeted single-cell RNA-seq. *Protoc. Exch.* https://doi.org/10.21203/rs.3.pex-864/v1 (2020).
49. Tilgner, H. et al. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**, 231–242 (2018).
50. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009).

Acknowledgements

This work was supported by grants from the European Research Council Advanced Investigator Grant (AdG-294542 and AdG-742804 to L.M.S.) and the Emerson Collective (award 643577 to L.V. and L.M.S.). D.S. was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EIPOD) program under Marie Sklodowska-Curie Actions COFUND (grant agreement number 664726). A.R.G. was supported by an Early Postdoc.Mobility fellowship (project number P2LAP3_171806) from the Swiss National Science Foundation (SNSF). We thank P. Collier, I. Gupta, H. Tilgner and D. Pavlinic for advice on 10X Chromium; S. Vonesch, K. Roy and J. Smith for advice on gRNA library cloning; V. Benes and D. Pavlinic for Illumina sequencing; M. Paulsen and team for flow cytometry service; S. Haas, J. Al-Sabah, C. Baccin, L. Ballenberger, L. Martins, C. Scholl and K.-M. Noh for providing cell samples; and A. Rabinowitz for advice on the analysis of Hi-C data.

Author contributions

D.S., L.V., A.R.G. and L.M.S. conceptualized the project, with contributions by J.O.K. D.S., D.R.L., P.J. and J.H.M. performed experiments. A.R.G. developed primer design pipeline. A.R.G. and L.V. performed data analysis. L.M. and C.M. implemented Drop-seq. D.S., A.R.G., L.V. and L.M.S. wrote the manuscript. All authors commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

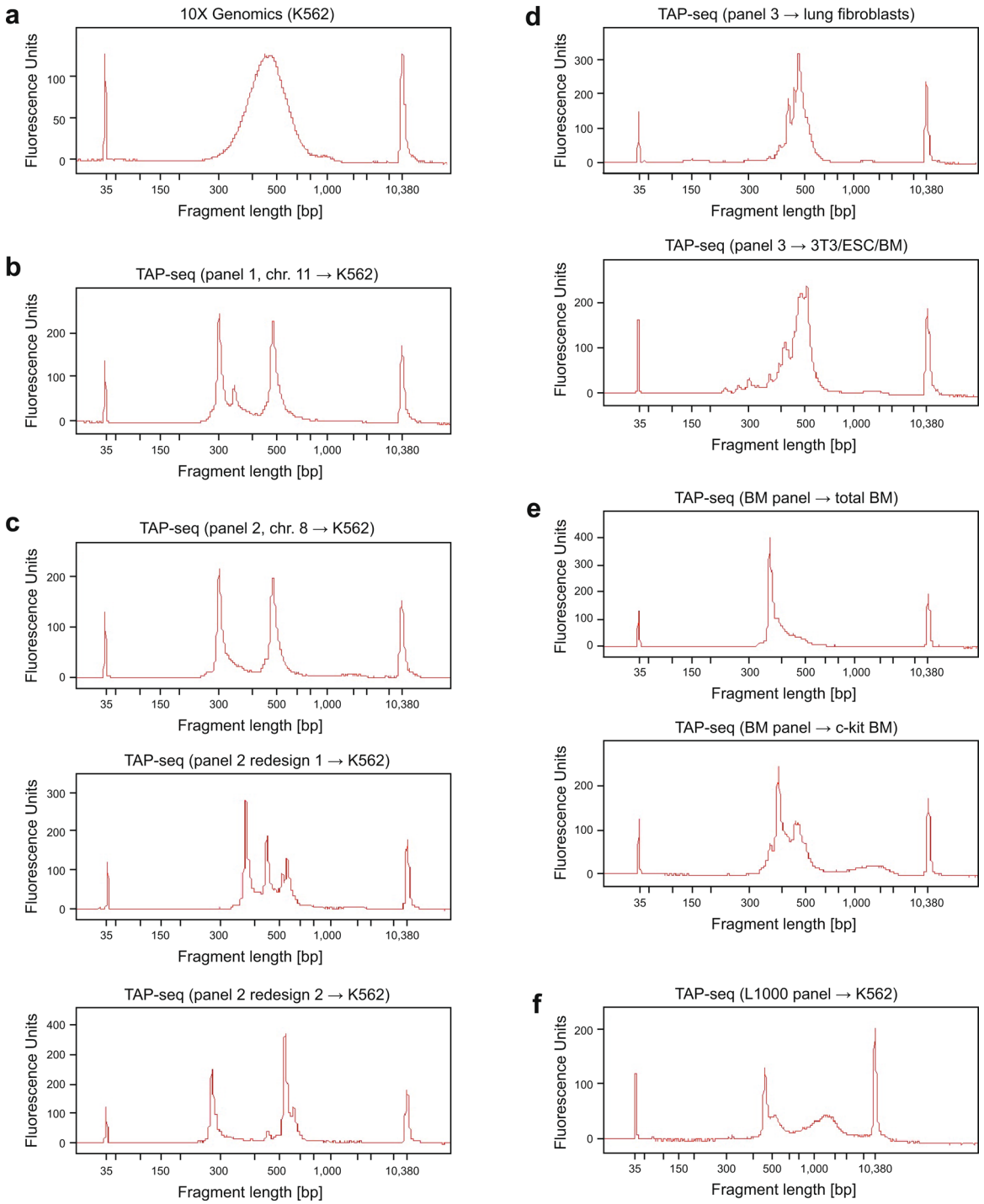
Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0837-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0837-5.

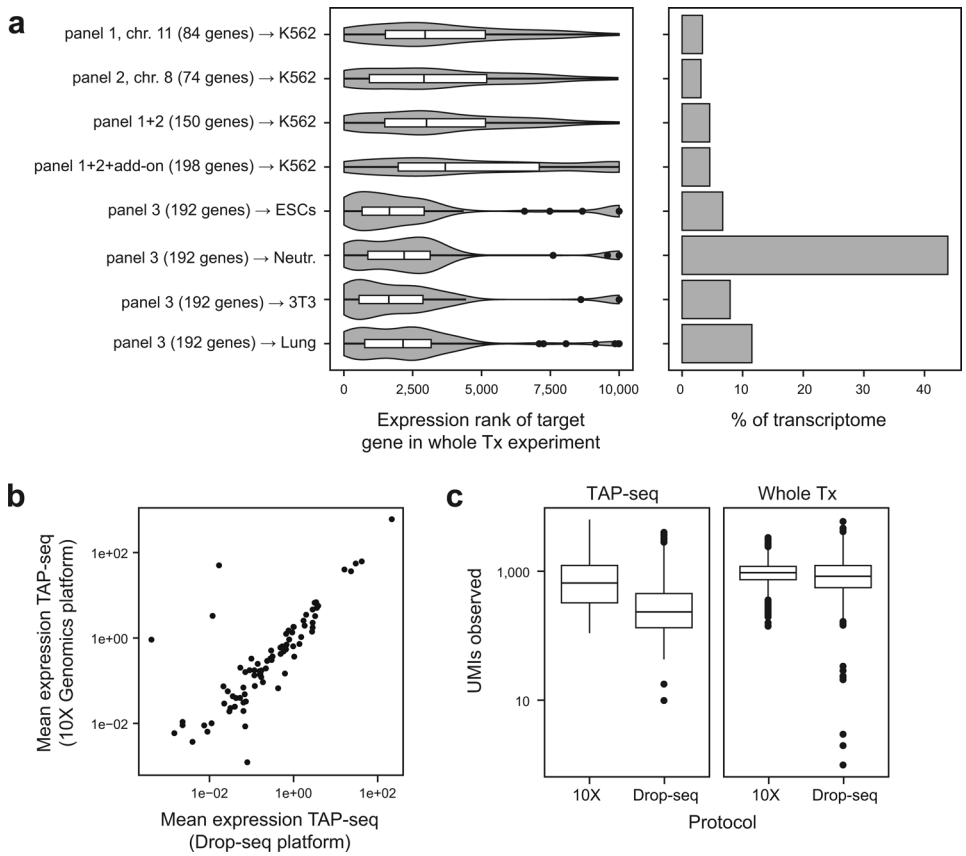
Correspondence and requests for materials should be addressed to L.V. or L.M.S.

Peer review information 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.

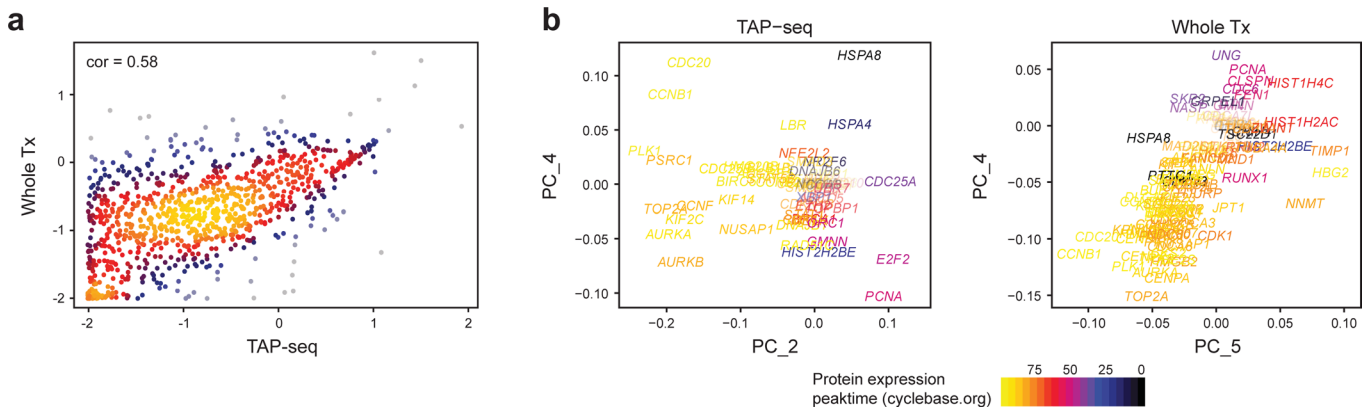
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Sample bioanalyzer traces for libraries from TAP-seq and 10X Genomics. a, Standard 10X Genomics protocol. **b**, TAP-seq library using panel 1 and cDNA from 10X Genomics K562 cells as input material. Strong peaks in TAP-seq profile correspond to highly expressed genes in the primer panel (*HBG1*, *HBG2*, *HBE1*), as validated by sub-cloning of bands and Sanger sequencing (not shown). **c-f**, Remaining target gene panels applied to different cell types and cell lines.

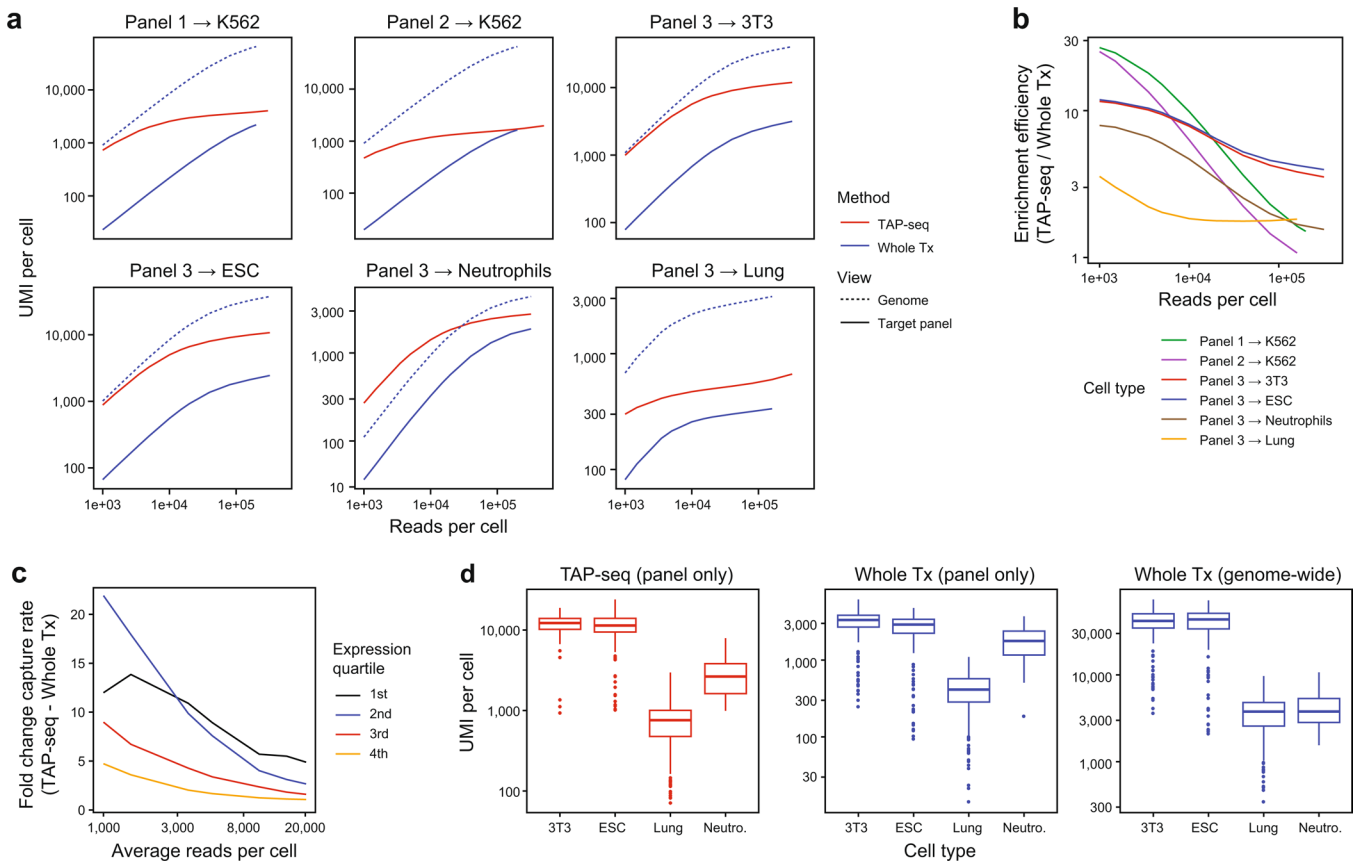


Extended Data Fig. 2 | Choice of target genes and single cell capture platform. a, Left panel: Expression ranks of genes used for the three test panels. An *x* axis value of 1 refers to highest expression. See *y* axis labels for number of target genes, and refer to Methods section on 'data visualization' for a description of violin plot elements. Right panel: Fraction of the transcriptome covered by these panels, computed from a whole transcriptome reference data set from the same cell type (*y* axis labels). For panels 1+2, K562 cells were used, panel 3 was applied to mouse embryonic stem cells (ESCs), mouse 3T3 cells, mouse neutrophils (Neutr.), and mouse lung mesenchymal cells (Lung). **b**, Mean gene expression levels for 10X Genomics and Drop-seq based TAP-seq. Panel 1 was used in both cases. n = 684 genes are shown. **c**, Number of UMIs observed per cell for both TAP-seq and whole transcriptome readout (Whole Tx), using 10X Genomics or Drop-seq for RNA capture and reverse transcription. Experiments were downsampled to an average of 1,000 reads per cell.

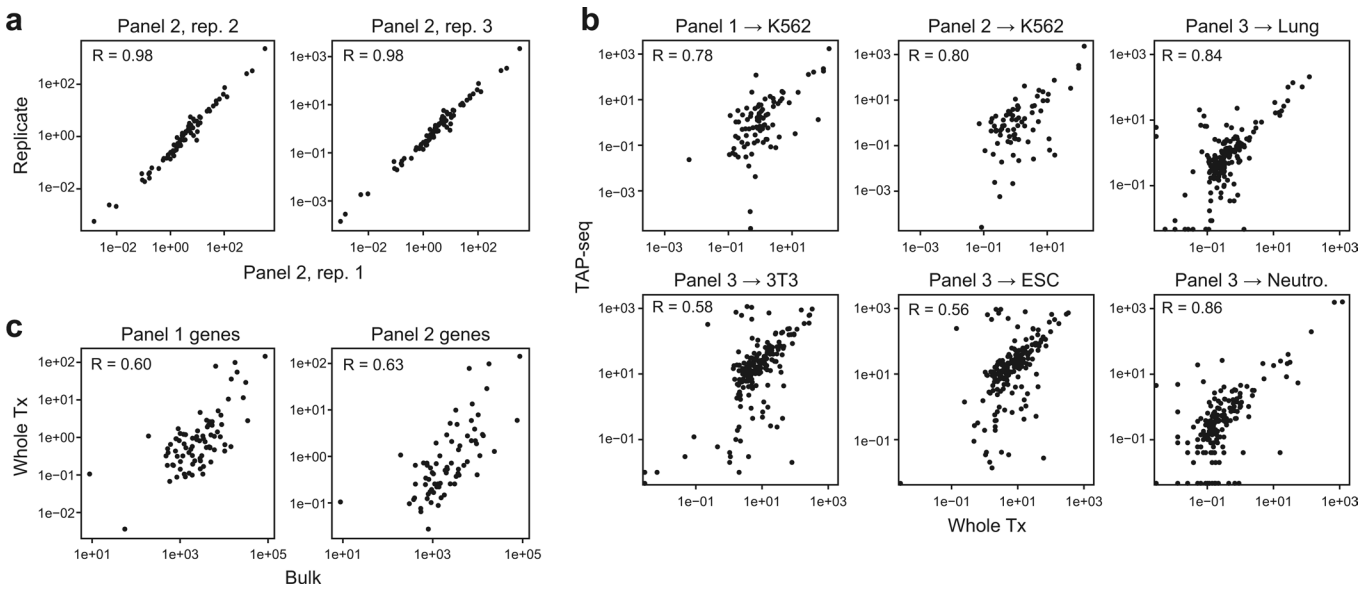


Extended Data Fig. 3 | Analysis of the L1000 panel. a, Spearman correlation in mean gene expression levels between TAP-seq and whole transcriptome readout (Whole Tx) for a panel targeting the L1000 gene set[34]. n = 6,963 genes were covered in TAP-seq and are included in the plot. **b**, Principal component analysis of the TAP-seq dataset and the whole transcriptome dataset. Principal component loadings of all genes annotated in cyclebase v3[51] are shown, with the peak-time of expression color-coded. In the whole transcriptome dataset, PC1-3 were not significantly associated with GO-terms (not shown). n = 68,734 cells (TAP-seq) and 8,282 cells (Whole Tx).

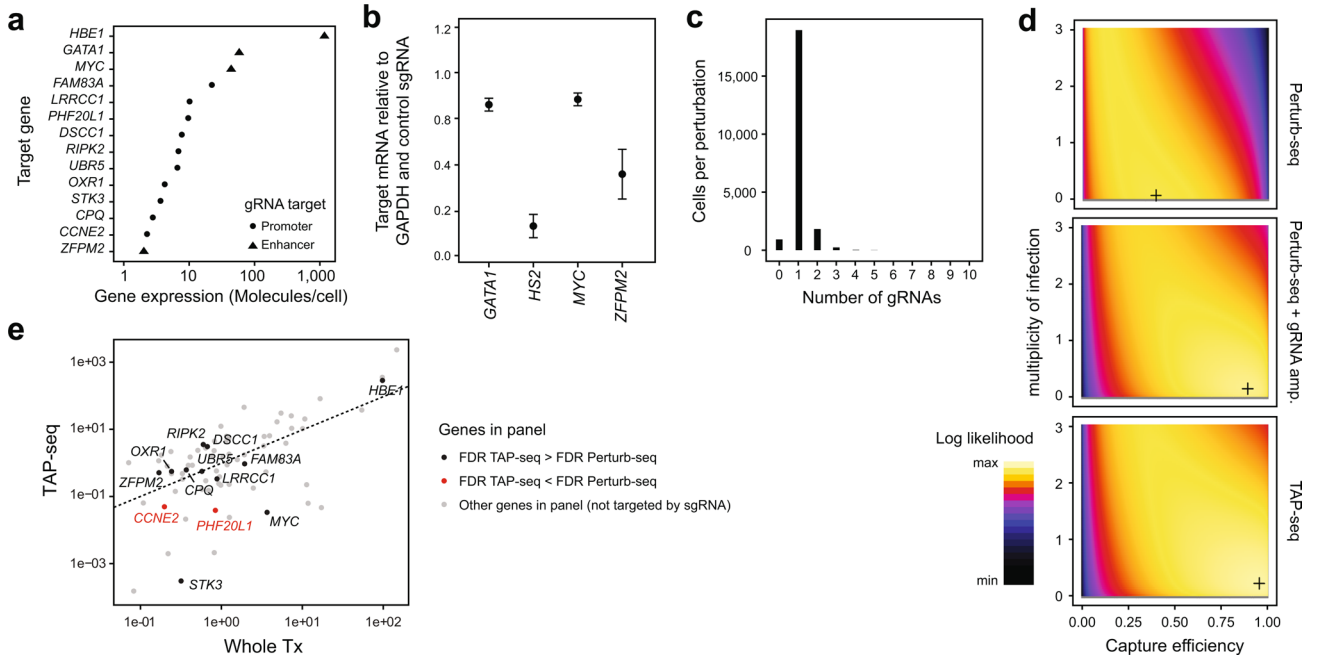
51. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).



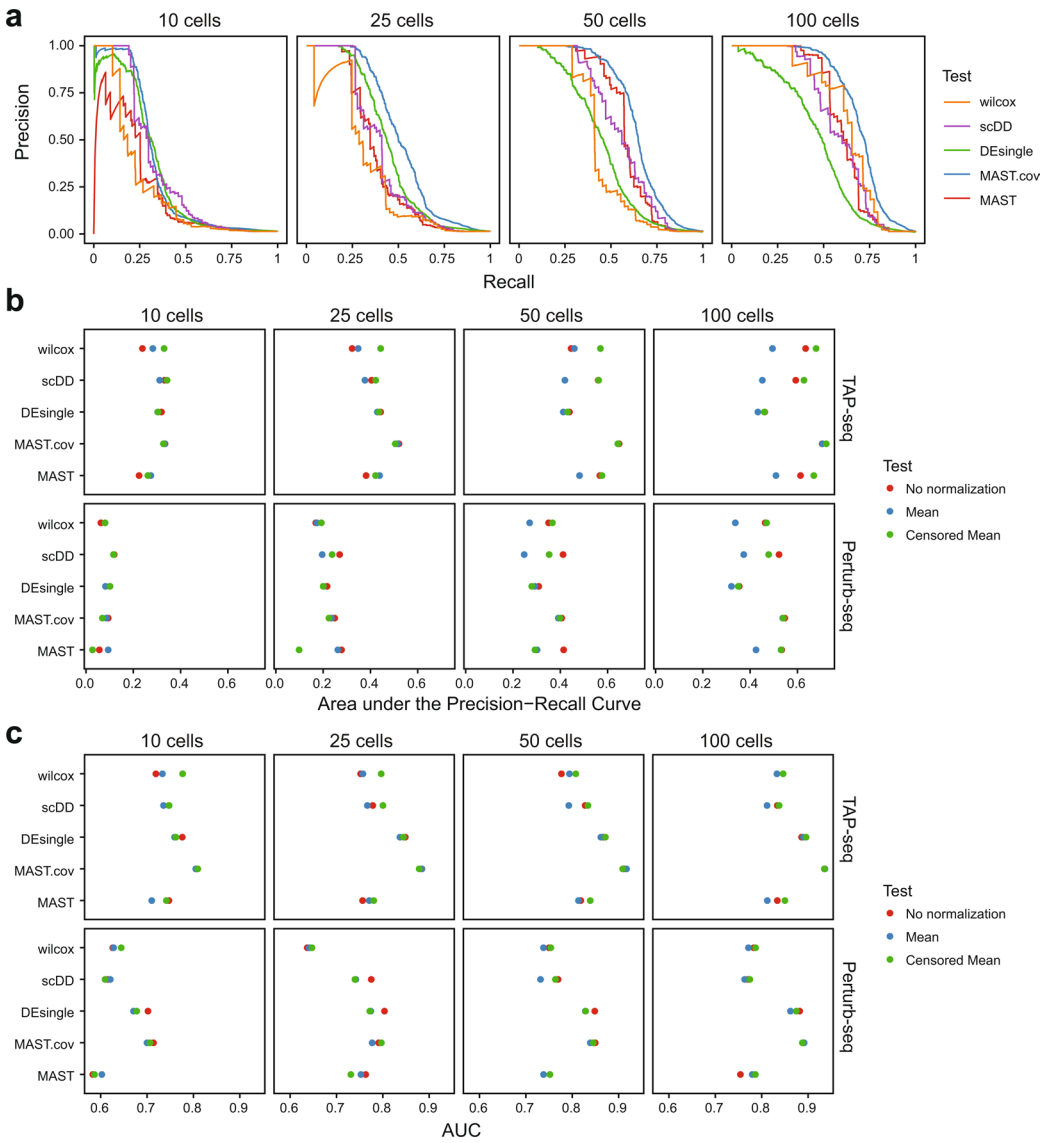
Extended Data Fig. 4 | Analysis of library complexity in TAP-seq and whole transcriptome 10x Genomics. a, Deeply sequenced TAP-seq and whole transcriptome (Whole Tx) libraries were downsampled to a given average number of reads per cell (*x* axis). The average number of UMIs observed on the target panel (solid lines, shown for both methods) or across the entire genome (dashed line, only shown for whole transcriptome readout) is shown. See also Fig. 1e. **b**, Deeply sequenced TAP-seq and whole transcriptome libraries were down-sampled to a given average number of reads per cell (*x* axis). The ratio in UMIs observed on the target gene panel between TAP-seq and whole transcriptome sequencing is plotted as a measure of enrichment efficiency **c**, For K562 cells and panel 1, gene detection levels were compared between genes of different expression levels. See also Fig. 1f. **d**, Number of molecules observed per cell in different cell types at 160,000 reads per cell. n = 6,109 3T3 cells, 160 ESCs, 130 Lung cells and 55 Neutrophils. See Methods section on 'data visualization' for a description of box plot elements.



Extended Data Fig. 5 | Analysis of reproducibility in TAP-seq and whole transcriptome 10X Genomics. a, Pearson correlation in mean gene expression levels across all genes of panel 2 (n = 674 genes) between three biological replicates. **b**, Pearson correlation between whole transcriptome 10X Genomics and TAP-seq for various panels and cell lines/cell types (see Extended Data Fig. 2a for number of genes per panel). **c**, Pearson correlation between whole transcriptome 10X Genomics and bulk RNA-seq (GEO: GSM2343836), across the n = 684 genes of panel 1 and the n = 674 genes of panel 2.

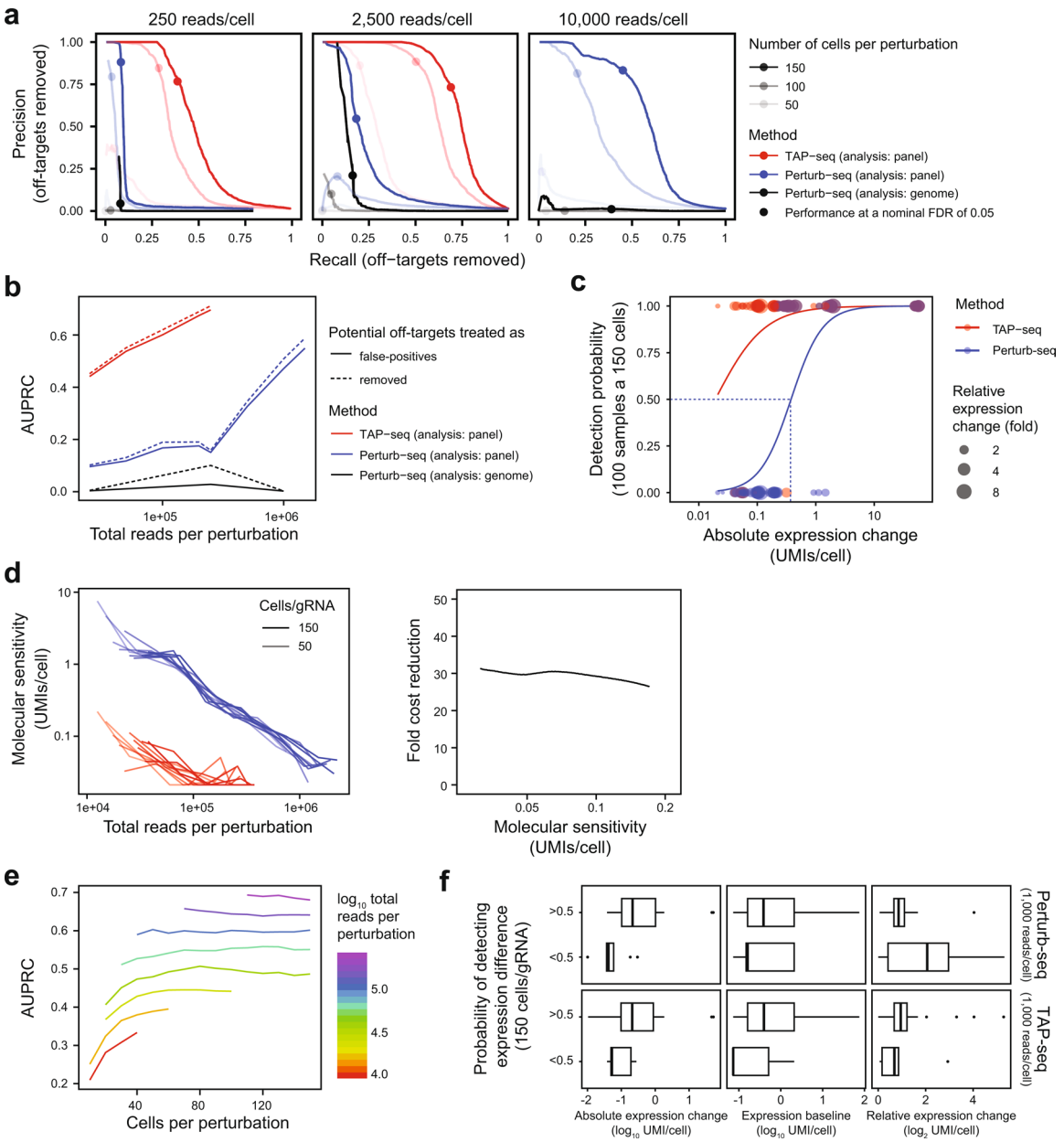


Extended Data Fig. 6 | Technical properties of the ground truth perturbation experiment. a, Gene expression level in K562 cells of the various gRNA target genes used. **b**, Enhancer gRNAs were validated by pooled transduction of K562 dCas9-KRAB cells with all four enhancer-targeting guides, and the effect on target gene expression was quantified by qPCR. *HBE1* was analyzed as target gene for the HS2 enhancer. n = 63 replicates. **c**, Histogram of the number of gRNAs identified per cell in the TAP-seq experiment of Fig. 2. **d**, The number of gRNAs observed per cell (see also in c) was fitted with a generative model of gRNA capture efficiency and multiplicity of infection[4,20]. Log-likelihood is plotted as a function of the parameters; the maximum likelihood estimate is marked by a cross. Data from n = 621,977 (TAP-seq), n = 67,994 (Perturb-Seq) or n = 637,971 cells (Perturb-seq + gRNA amp.) was used. **e**, Mean expression per gene for whole transcriptome 10X Genomics compared to TAP-seq, with perturbation target genes highlighted. n = 674 genes from panel 2 are shown. Two genes for which perturbation effects were detected with a lower efficiency in TAP-seq are highlighted in red.

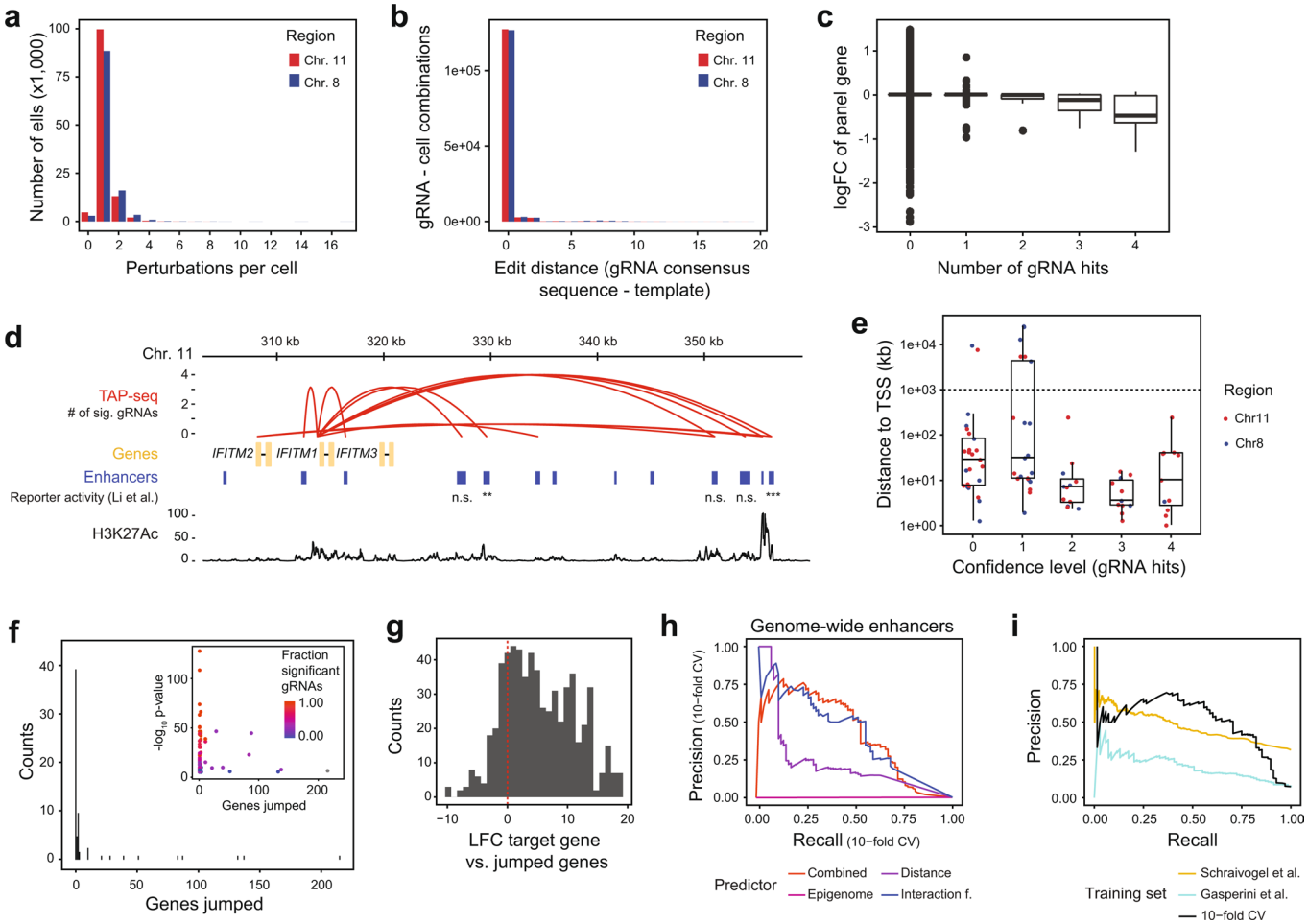


Extended Data Fig. 7 | Comparison of differential expression testing methods. a, Comparison using Precision-Recall curves, as in Fig. 2f. TAP-seq data were downsampled to 10, 25, 50 or 100 cells per gRNA. For each sampling run, differential expression testing was performed using a simple (two-sided) Wilcoxon test, MAST[52], DEsingle[53] and scDD[54], as well as MAST with the number of genes observed as an additional covariate. Precision-Recall curves were computed assuming that the intended gRNA targets constitute the full set of true positives. Data were normalized across cells using the censored mean, that is division with the mean expression of all genes not part of the highest decile. **b**, Performance comparison in terms of area under the Precision-Recall curve for different data normalization strategies and tests. **c**, Performance comparison in terms of area under the ROC curve.

52. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
53. Miao, Z., Deng, K., Wang, X. & Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* **34**, 3223–3224 (2018).
54. Korthauer, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17**, 222 (2016).

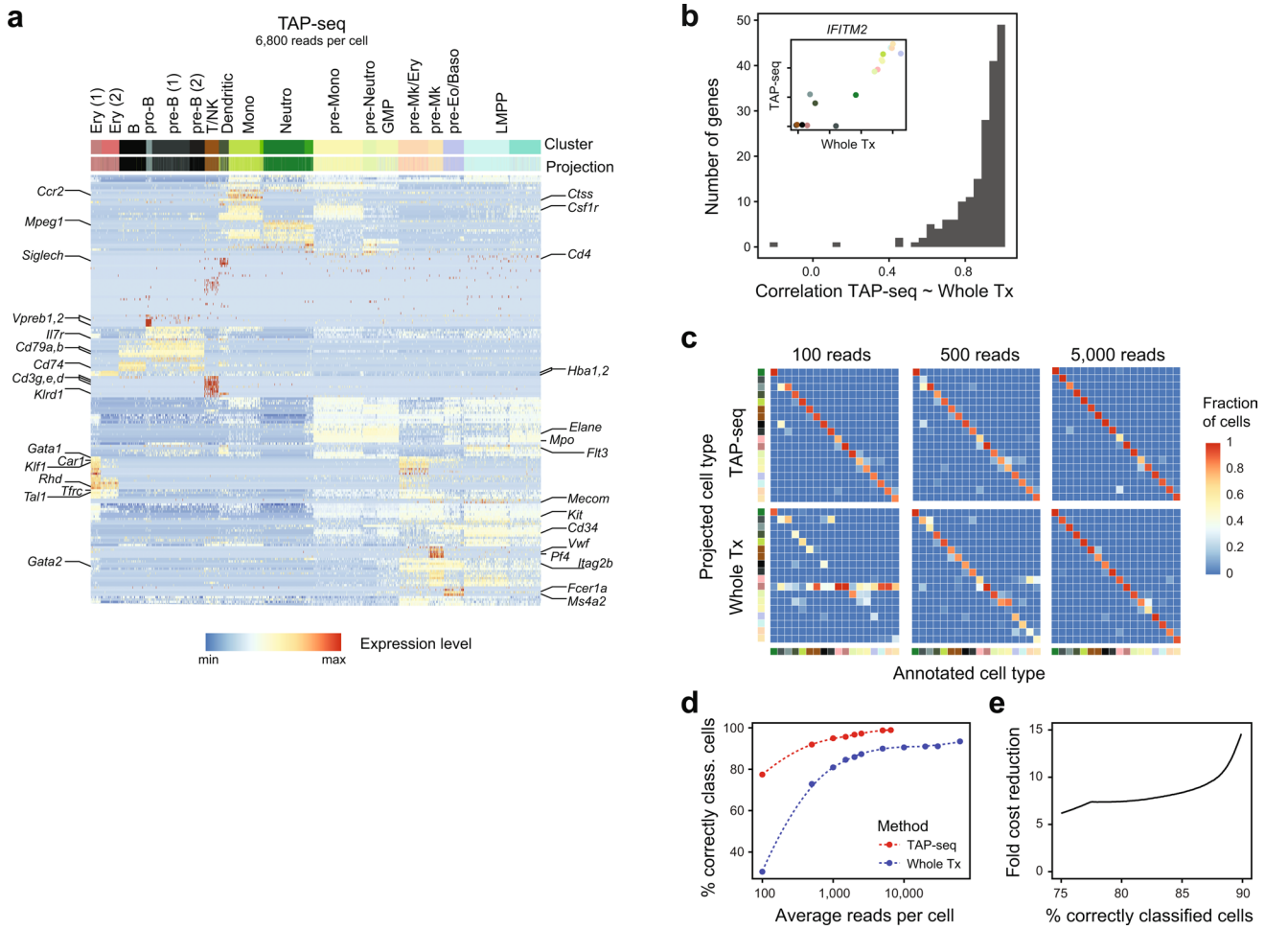


Extended Data Fig. 8 | Additional analyses of the ground truth perturbation dataset. a, Precision-Recall curves, as in Fig. 2f. Potentially true gRNA off-target or downstream effects were identified by differential expression testing across all cells, and then excluded from the analysis. Points indicate performance at a nominal FDR of 0.05. See Note S3 section 'Sensitivity analysis (differential expression)' for detail on the statistical test used. **b**, Comparison of Area under the precision-recall curves (AUPRC) for n = 6,100 cells per perturbation, sampled to various read depths. Potential gRNA off-target and downstream effects were treated as false positives (solid lines, same as in Fig. 2g) or excluded (dashed lines). **c**, The absolute effect of a gRNA-mediated perturbation in UMIs/cell was quantified from non-downsampled whole transcriptome data (*x* Axis). The probability of observing these effects as significant was the quantified by drawing 100 samples using 150 cells per sample and 1,000 average reads per cell (*y* Axis). Lines derive from a logistic regression. The UMI difference required for achieving a 50% detection probability was used as a measure of molecular sensitivity (dotted line). Data from n = 660,106 cells and 9,750 sampling runs. **d**, Like Fig. 2g, but using molecular sensitivity as defined in panel c as the measure of sensitivity. Down-sampling was restricted to 50–150 cells per perturbation, since estimates of molecular sensitivity were otherwise driven by excessive sampling noise. Data from n = 660,106 cells and 7,150 sampling runs. **e**, AUPRC plotted in relationship to number of cells per perturbation and total number of reads (data from Fig. 2g). **f**, For of n = 656 each gRNA targets, the absolute and relative expression change elicited by the perturbation, as well as the expression baseline, were computed from whole transcriptome data without subsampling (*x* axis). Data from both methods were then downsampled repeatedly to 150 cells per perturbation and 10,000 (Perturb-seq) or 1,000 (TAP-seq) reads per cell to determine the probability of detecting a change (*y* axis). Refer to methods section on 'data visualization' for a definition of box plot elements.



Extended Data Fig. 9 | Additional analyses of the enhancer screen. a, Number of detected gRNAs/perturbations per cell were plotted. **b**, Levenshtein edit distance between the consensus sequence of a gRNA in a given cell, and the template sequence, showing that in 93 % (chr. 8) or 95 % (chr. 11) of cases, there were no mismatches between consensus and template. **c**, Fold change in gene expression of enhancer targets is plotted in relation to the number of gRNAs supporting an enhancer-target gene pair (ETP). Number of ETPs per confidence level: 0 = 61, 1 = 621, 2 = 612, 3 = 611, 4 = 611. **d**, Zoom-in on a region surrounding the IFITM locus shows identification previously known enhancers[55]. **e**, Distance to transcription start site (TSS) was plotted against confidence level, as calculated from the number of individual gRNAs with a detected effect on the target gene. Number of ETPs per confidence level: 0 = 61, 1 = 621, 2 = 612, 3 = 611, 4 = 611. See Methods section on 'data visualization' for a definition of boxplot elements. **f**, Number of genes jumped between an enhancer and the identified target gene was plotted (main panel). Inset shows association strength, calculated from the proportion of gRNAs that support the ETP, plotted against the number of jumped genes. **g**, Histogram of log-fold expression differences between jumped genes and the respective true enhancer target. **h**, Like Fig. 3h, but including all 34,493 potential ETPs across the whole dataset, instead of just gene-proximal ETPs. **i**, Precision-Recall curves for classifiers trained on the dataset generated in this study, and applied to the dataset from ref. [9] (orange line), or classifiers trained on the dataset from ref. [9] and applied to this dataset.

55. Li, P. et al. Coordinated regulation of IFITM1, 2 and 3 genes by an IFN-responsive enhancer through long-range chromatin interactions. *Biochim. Biophys. Acta Gene Regul. Mech.* **1860**, 885–893 (2017).



Extended Data Fig. 10 | Additional analyses of the mouse bone marrow experiment. a, Heatmap depicting the expression of all 182 target genes across 11,794 cells, as measured by TAP-seq. Top row ('Cluster') depicts the result of unsupervised clustering, second row ('Projection') depicts the result of transferring labels[26] from the whole transcriptome reference data set (see Fig. 4a for color code). **b**, Gene expression correlations across populations. Mean gene expression for each gene in the mouse bone marrow panel was computed for each cell type, and the Pearson correlation between TAP-seq and whole transcriptome readout (Whole Tx) across n = 618 cell types was computed. Main panel shows Pearson correlation coefficients for all tested genes across cell types. Inset shows expression of *IFITM2* as measured by TAP-seq and whole transcriptome readout for each cell type (color code as described in Fig. 4a). **c**, Data from both methods were downsampled to various average read depths and an identical number of cells, and labels were transferred[26] from the non-downsampled reference. For each cell type plotted on the *x* axis, the fraction of cells projected to the cell types plotted on the *y* axis was quantified (color code as described in Fig. 4a). **d**, Average read depth per cell is plotted against the fraction of cells correctly classified. **e**, The fold difference in sequencing reads between TAP-seq and whole transcriptome is plotted as a function of the fraction of cells correctly classified.

nature research

Corresponding author(s): Lars Velten
Lars Steinmetz

Last updated by author(s): Jan 16, 2020

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	For primer design, a custom pipeline was used. The pipeline is made available on Github (https://github.com/argschwind/TAPseq), and will be made available on Bioconductor upon manuscript acceptance.
Data analysis	All scripts for data processing and data analysis are available at https://github.com/argschwind/tapseq_manuscript . These scripts were run in R version 3.5.1 and make use of the following packages: ggplot2 (v. 3.1.0), gviz (1.24.0), pheatmap (v. 1.0.10), Seurat (3.4.1), MAST (1.12.0), ROCR (1.0), LSD (4.0), plyr (1.8.6), reshape2 (1.4.3), Gviz (1.30.3) and GenomicInteractions (1.20.1)

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 data are available from GEO GSE135497 (reviewer access token mdojumiybjetjkv). A summary of all genomics data used for each figure is provided in Table S5.

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

Life sciences study design

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

Sample size	Sample size for the screen (Fig. 3) was determined using a deeply samples control experiment, as discussed in Figure 2. For the analysis underlying Figure 1+4, no statistical procedures were used to select a sample size, but rather, cell numbers typically used in the field were used.
Data exclusions	For the cell type experiment underlying Fig. 4, single cell transcriptomes were excluded if less than 500 (Whole Tx, pre-established as a standard in the field) or 10 (TAP, not pre-established) genes were observed as expressed. For the data underlying Fig. 1-3, the standard DropSeq analysis workflow was used to distinguish between true cells and background events. All events classified as cells were included into the analysis.
Replication	For large-scale screens or deeply sequenced datasets, no replication was performed for reasons of cost. Reproducibly of TAP-seq is assessed in figures S4a and Note S1, figure 1
Randomization	Genetic perturbations were randomized by means of pooled lentiviral transduction. No other randomization strategies were applied.
Blinding	Not applied. Blinding was not relevant since sample identities were encoded into the gRNA perturbations themselves, and read out using an automated processing pipeline.

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

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	K562 cells were newly obtained from ATCC (CCL-243). HEK293 FT cells used for lentiviral production were a kind gift of J. Korbel and B. Mardin. NIH 3T3 cells were a kind gift from Matthias Hentze. Mouse embryonic stem cells were a kind gift of Kyung-Min Noh.
Authentication	Cells were not authenticated
Mycoplasma contamination	Cells were tested negative for mycoplasms
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	C57BL/6J, female, 8 to 12 weeks of age
Wild animals	The study did not involve wild animals
Field-collected samples	The study did not involve field-collected samples
Ethics oversight	Killing of animals for tissue extraction was done under EMBL Policy on the Protection and Welfare of Animals Used for Scientific Purposes or DKFZ policies for work with laboratory animals. Animals were solely used for collection of material, no animal experiments were performed.

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