Probabilistic modeling of ambient noise in single-cell omics data

Authors:

Caibin Sheng[1#], Rui Lopes[1*], Gang Li[1*], Sven Schuierer[2], Annick Waldt[2], Rachel Cuttat[2], Slavica Dimitrieva[1], Audrey Kauffmann[1], Eric Durand[1], Giorgio G. Galli[1], Guglielmo Roma[2], Antoine de Weck[1#]

Affiliations:

[1]Disease area Oncology, Novartis Institute for Biomedical Research, CH-4002 Basel, Switzerland.
[2]Chemical Biology and Therapeutics, Novartis Institute for Biomedical Research, CH-4002 Basel, Switzerland.
[*]These authors contributed equally: Rui Lopes, Gang Li.
[#]Corresponding authors. Email: caibin.sheng@novartis.com (C.S.); antoine.deweck@novartis.com (A.d.).

Abstract

Droplet-based single-cell omics, including single-cell RNA sequencing (scRNAseq), single cell CRISPR perturbations (e.g., CROP-seq) and single-cell protein and transcriptomic profiling (CITE-seq) hold great promise for comprehensive cell profiling and genetic screening at the single cell resolution, yet these technologies suffer from substantial noise, among which ambient signals present in the cell suspension may be the predominant source. Current efforts to address this issue are highly specific to a certain technology, while a universal model to describe the noise across these technologies may reveal this common source thereby improving the denoising accuracy. To this end, we explicitly examined these unexpected signals and observed a predictable pattern in multiple datasets across different technologies. Based on the finding, we developed single cell Ambient Remover (scAR) which uses probabilistic deep learning to deconvolute the observed signals into native and ambient composition. scAR provides an efficient and universal solution to count denoising for multiple types of single-cell omics data, including single cell CRISPR screens, CITE-seq and scRNAseq. It will facilitate the application of single-cell omics technologies.

Introduction

Single-cell RNA sequencing (scRNAseq) enables researchers to investigate transcriptomes at single cell resolution, improving our understanding of cellular heterogeneity and interactions between single cells and the microenvironment. Recent efforts have extended scRNAseq beyond transcriptomes by encoding additional layers of information, resulting in versatile tools for single-cell omics. For instance, by combining functional screens with scRNAseq, CROP-seq has enabled the interrogation of multiple biological nodes in a single experiment[1–3]; by combining ssDNA-barcoded-antibodies with scRNAseq, CITE-seq has provided simultaneous quantification of mRNA and surface proteins in a single cell[4], which shows great potential especially in immunophenotyping in fundamental and clinical research[5]. Most recent efforts have even combined both technologies to enable multimodal profiling of transcriptome and surface proteins in response to gene perturbations in cancer cells[6]. Despite the exciting concepts and anticipated potential, applying these pioneering technologies is challenging. One outstanding reason is the wide presence of measurement noise. Various technical factors, such as ambient contamination[7,8], amplification bias[9] and index swapping[10] generate noise in single-cell omics experiments. Several methods have been proposed to correct for the background signals[7,8,11–13]. Most of these are highly specific to transcriptome data[7,8,11,12], several tools are specific to protein data in CITE-seq[13,14], while few attempt to denoise exogenous barcode counts in other extended single cell technologies, such as single cell CRISPR screens and cell indexing[15–17] technologies. Conceptually, however, there is little difference in the procedure in which these various technologies are capturing their respective information (i.e., mRNA, sgRNA, expressed barcodes and antibody counts). All relevant molecules are included in the same reaction solution during most of the involved processes, such as droplet formation, cell lysis, library construction and sequencing. Background noise likely originates in a similar (if not identical) way in each of these layers, meaning an ideal model can, in principle, describe the common sources of the artifacts in a non-technology-specific manner. To our knowledge, no such algorithm has been proposed so far.

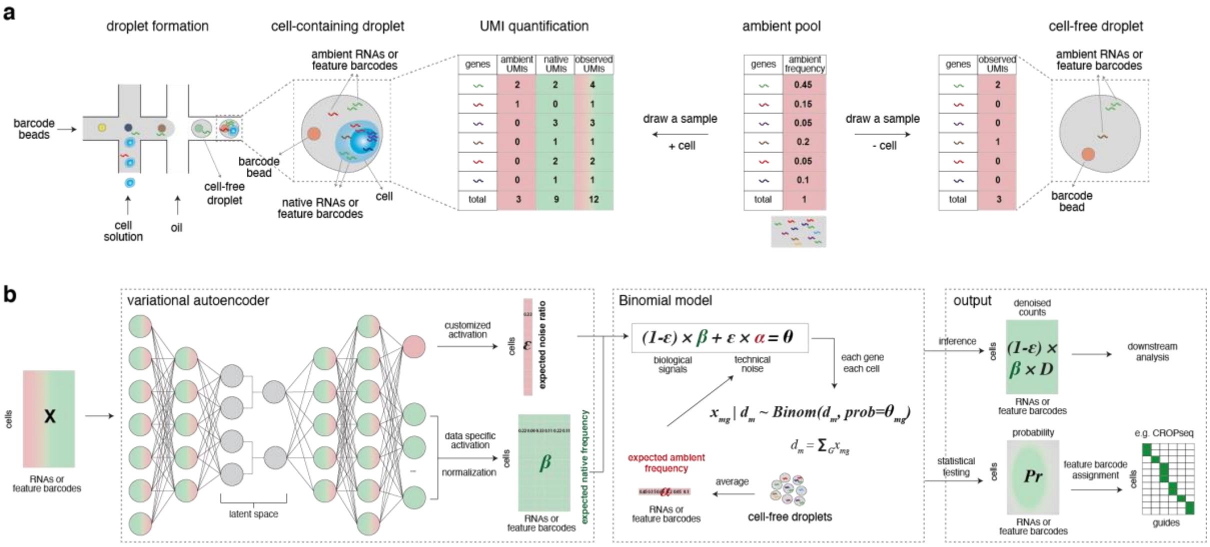
To this end, we developed single cell ambient remover (scAR), a hypothesis-driven model to identify and remove the background noise for transcriptome, protein and feature-barcode data in single-cell omics technologies. Cell-free transcripts have been observed in empty droplets[18] and are hypothesized to arise from ambient RNAs in single cell suspension, which likely originates from damaged cells caused by cell lysis[7,19,20]. This hypothesis suggests that ambient RNAs may not be completely random but deterministic signals to a certain extent. Indeed, gene frequencies are correlated between cell-containing and cell-free droplets[7]. These together motivate us to explicitly evaluate the ambient signal hypothesis in multiple single-cell omics technologies, which rationalizes a universal probabilistic model to describe this type of noise. To highlight the generality of our approach, we apply scAR to multiple datasets generated using different technologies from different sources, including an internal CROP-seq dataset and several public CITE-seq datasets and scRNAseq datasets. We also evaluate scAR with competing methods where available.

Results

The scAR model.

scAR uses a latent variable model to represent the biological and technical composition in the observed count data (Fig. 1). It is designed under ambient signal hypothesis, which assumes that ambient signals originate from broken cells during sample preparation, homogenize in single cell solution (ambient signal pool) and contaminate cell-containing and cell-free droplets (Fig. 1a). Mathematically speaking, ambient signals are drawn from Binomial distributions with a shared parameter (denoted as ambient frequencies, α) in cell-containing and cell-free droplets. This parameter therefore can be estimated using cell-free droplets. Besides, we introduce two hidden variables noise ratio (ε) and native expression frequencies (β) to represent the total contamination level per cell and normalized 'true' expression per cell respectively (Fig. 1b). scAR simultaneously infers ε and β using the variational autoencoder (VAE) framework[21–23] (Fig. 1b, Methods).

We use the optimized variables ε and β to estimate the 'theoretical' gene expression, which is considered as denoised counts for downstream analysis. In addition, in several feature barcode technologies, such as CROP-seq and CellTagging[15], the presence/absence of native signals is more critical information than the actual level. To reflect this, scAR outputs a probability matrix representing the probability whether raw observed counts contain native signals.



genes	ambient UMIs	native UMIs	observed UMIs
~	2	2	4
~	1	0	1
~	0	3	3
~	0	1	1
~	0	2	2
~	0	1	1
total	3	9	12

genes	ambient frequency
~	0.45
~	0.15
~	0.05
~	0.2
~	0.05
~	0.1
total	1

genes	observed UMIs
~	2
~	0
~	0
~	1
~	0
~	0
total	3

Fig. 1 | The overview of scAR. scAR is a hypothesis-driven model for noise reduction in droplet-based single-cell omics. **a,** Demonstration of the ambient signal hypothesis. During preparation of single cell solution, RNA or protein counts are released upon cell lysis and consequently encapsulated by droplets with or without a cell. In the case of a cell, these exogenous molecules are mixed with native ones and barcoded by the same bead, which results in mixed counts as the final output. Under this assumption, the ambient signals in both cell-containing and cell-free droplets are drawn from a same pool (meaning an identical distribution). UMI stands for unique molecular identifier. For the purpose of illustration, reddish purple, light green and their mixture hint ambient signals, native signals and observed counts, respectively. **b,** scAR takes raw count matrices of RNA or protein as input and learns two sets of parameters (ε and β) through the variational autoencoder. ε, a column vector represents noise ratios per cell and β, a matrix represents cell-wise native frequencies of RNAs or proteins. α, a row vector represents the ambient frequencies of RNAs or proteins, which is empirically estimated by averaging cell-free droplets. scAR assumes α is an experiment-specific factor thereby using a unique α for all cells from a single experiment. The observed raw counts are modelled using a Binomial model which contains known parameters α and sequencing depth D and two hidden variables ε and β. We optimized ε and β by minimizing the reconstruction errors and K-L divergence (Methods). scAR outputs two matrices, a denoised count matrix and a probability matrix. The latter represents the probability that a given observed count is not drawn from ambient sources (in other words, native signals exist). Meaning of color codes is the same as **a**.

Examination of ambient signal hypothesis.

We conducted a case study which combines CROP-seq and bulk sequencing to evaluate the ambient signal hypothesis (Fig. 2a and Methods). We designed a viral pool of 99 sgRNAs targeting 13 different genes (supplementary table I), most of them being essential in MCF7 cells[24,25] (supplementary Fig. 1). We infected MCF7 cells expressing dCas9-KRAB with the lentiviral libraries at extraordinary low multiplicity of infection (MOI=0.3) to ensure single integrations of sgRNAs. Excessive sgRNAs in a cell are supposed to be ambient signals. Cells were harvested at various time points post-transduction and split into two portions, with one portion taken for 10x scRNAseq and the other for bulk sequencing of sgRNAs, which reveals the frequencies of sgRNA libraries in the samples. In addition, the gene regulation activity of CRISPRi provides an additional way to assess the identification of ‘true’ sgRNAs.

To validate whether ambient sgRNAs are correlated with the native ones, we compared sgRNA frequencies in bulk sequencing and in cell-free droplets from CROP-seq. Results show high correlation of sgRNA frequencies at both time points (Fig. 2b). Randomly sampled subsets of cell-free droplets (from 0.5% to 5%) also show high correlation (Fig. 2c), surprisingly, as few as ~200 droplets consistently show high correlation to the bulk cells. Together, these observations suggest that ambient signals are not random noise but endogenous expression-correlated artifacts. We next examined the raw sgRNA counts in cell-containing droplets and observed presence of ambient counts (Fig. 2d). ~25 distinct sgRNAs were detected per cell on average while <=1 sgRNA is expected because of low multiplicity of infection. We also observed significant ambient contamination as well as high correlation to endogenous expression in datasets of different technologies and from various laboratories[3,26,27] (supplementary Fig. 2). Altogether, these indicate that the ambient signals are systematic noise in single-cell omics and building scAR on this basis is rational.

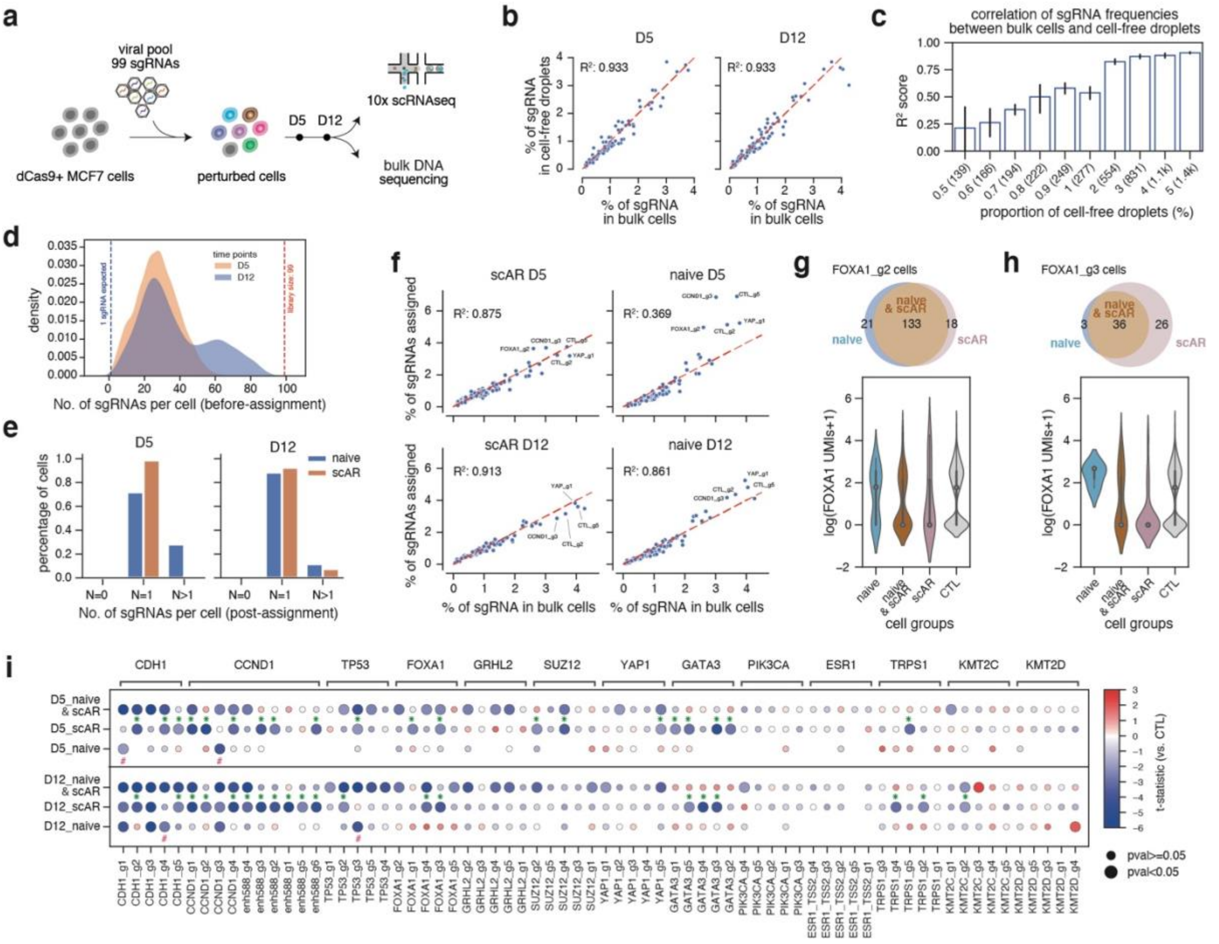


Fig. 2 | scAR identifies integrated signals in CROP-seq data. A CROP-seq experiment shows strong ambient contamination, which affects guide assignment when using the naïve approach. scAR enables more accurate guide assignment. **a,** Demonstration of the experimental design (Methods). **b,** Scatterplots of sgRNA frequencies between two sources. X-axis represents sgRNA frequencies obtained by bulk DNA sequencing, Y-axis represents sgRNA frequencies obtained by averaging RNA counts in cell-free droplets from 10x scRNAseq. Each dot represents an sgRNA, the red dashed lines represent y=x and coefficients of determination (R^2 scores) are shown. **c,** coefficients of determination (R^2 scores) show the correlation of sgRNA frequencies between bulk cells and randomly sampled cell-free droplets. X-axis shows the proportions of cell-free droplets. The droplet numbers in each group are indicated in brackets. The error bars indicate the variance of 10 samplings. **d,** Distribution of distinct sgRNAs per cell in the raw counts. X-axis represents the number (not UMI count) of distinct sgRNAs per cell, y-axis shows the density. Samples from two time points D5 and D12 are colored by light blue and orange respectively. The size of lentivirus libraries and expected number of sgRNA are highlighted by dashed lines. **e,** Cell fraction after guide assignment. scAR assignment is based on the probability matrix, each cell is assigned with the guide(s) with highest probability. Naïve assignment is based on raw count matrix, each cell is assigned with the guide(s) with highest UMI counts. **f,** Similar to **b,** scatterplots show the sgRNA frequencies in cells from 10x scRNAseq (post-assignment) and in bulk cells. Y-axis here represents cell fractions grouped by distinct sgRNAs post-assignment. The red dashed lines represent y=x. **g-h,** Two selected guide groups from D12 samples to demonstrate scAR's performance. The Venn diagrams show the number of cells assigned with FOXA1_g1 (**g**) and FOXA1_g3 (**h**). scAR assignment is marked with purple and naïve assignment is marked with blue. Cells assigned with the same guide by both approaches are labeled as 'naïve & scAR'. The below violin plots show the expression of FOXA1 in these subgroups. CTL groups represent the cells assigned with CTL sgRNAs by both approaches. Y-axis represents the log transformed UMI counts after library size normalization (Methods). **i,** the dotplot shows the overall comparison between two assignment approaches. X-axis represents guide groups. Y-axis represents subgroups of cells as exemplified in **g** and **h**, separated by two time points and assignment approaches. Target genes are shown on the top. Their expression (log transformed) in each group is compared with that in CTL group and resulting t-statistics are shown by the dot color (Methods, see t-statics of z-normalized expression in supplementary Fig. S3a). Blue color indicates down-regulation, and red indicates up-regulation. CTL group is centered at zero. The bimodal sizes of circles represent the p-values from t-test (the bigger means p<0.05, the smaller means p>=0.05). * highlights the guide groups where scAR significantly improves the accuracy and # marks the groups where scAR underperforms naïve assignment.

scAR identifies true integrated signals in CROP-seq data.

To identify the true guide, we applied scAR to this CROP-seq dataset and compared its performance with naïve assignment, which considers most highly expressed guides as the true signal. Combined with an arbitrary threshold, this naïve assignment is widely implemented in current single-cell CRISPR screens[1,2,28–30]. Here, for benchmarking purpose, we did not perform any subjective filters on either naïve assignment or scAR-based assignment. All cells that pass the default gene and cell filtering in Cellranger were included for downstream analysis (Methods). By naïve assignment, ~80% cells (20076 out of 25248, D5 and D12 combined) are assigned to unique guides and ~20% (5170 out of 25248, D5 and D12 combined) cells are assigned to multiple (>=2) guides due to equal expression (Fig. 2e). 'Multiple-infected' cells are generally filtered out before downstream analysis in CROP-seq experiments, in other words, naïve assignment causes loss of ~20% cells. scAR estimates the expected ambient counts then compares to the observed counts via hypothesis testing to evaluate the probability of presence of native signals (Fig. 1b and Methods). It assigns 96% cells (24171 out of 25248) to a single guide (Fig. 2e) despite of ~20% cells with equally expressed guides. We next examined the cell fraction grouped by distinct sgRNAs after guide assignment. This fraction is expected to be identical to sgRNA frequencies in bulk sequencing since both reflect sgRNA libraries in the cell pool. scAR-resulting cell fractions are highly correlated with sgRNA frequencies in bulk sequencing (R^2=0.875 at D5 and R^2=0.913 at D12), while naïve assignment preferably over-assigns a few sgRNAs of highest expression (R^2=0.369 at D5 and R^2=0.861 at D12, Fig. 2f).

To evaluate the accuracy of the assignment, we checked the expression levels of targeted genes in cells with certain guides assigned exclusively by either naïve or scAR (Fig. 2g-i). We consider the cells assigned by both naïve and scAR as the positive control and cells assigned with CTL sgRNAs by both naïve and scAR as the negative control. In two selected guides shown in Fig. 2g and 2h, naïve assignment assigns 154 cells to FOXA1_g2, 133 cells among which are mutually assigned to the same guide by scAR. These 133 cells show significant downregulation of FOXA1, suggesting the effectiveness of this guide. However, the remaining naïve assigned 21 cells show similar expression as in CTL cells, suggesting that these cells may not integrate FOXA1_g2. More importantly, another 18 FOXA1_g2 cells, exclusively identified by scAR, show as low expression as in the mutually assigned cells. Similarly, for the other example FOXA1_g3 (Fig. 2h), 26 cells identified by scAR but missed by naïve assignment show similar expression pattern as the mutually assigned cells (Fig. 2h). To systematically assess and visualize the difference, we perform t-test on expression of targets among these subgroups for each guide and visualize both t-statics and p-values using dotplots (Fig. 2i and supplementary Fig. 3a). In total, scAR rescues 20 sgRNA groups at each time point which are missed by naïve assignment as confirmed by statical confidences, while only two sgRNA groups are missed by scAR at each time points, compared to naïve assignment. In addition, we count cell numbers of each subgroup and visualize the difference (supplementary Fig. 3b). Clearly, naïve assignment over-assigns cells to a few guides, such as CCND1_g3 and YAP1_g1 (supplementary Fig. 3b) likely due to their stronger ambient presence than other guides, whereas the power of scAR to identify ambient sgRNAs by their distribution ensures unbiased assignment. Together, by inspecting the guide assignment in the CROP-seq dataset, we showed that scAR significantly improves the assignment accuracy in feature barcode technologies, where the presence rather than the quantity of native signals is the key information.

scAR removes the ambient protein counts in CITE-seq.

Next, we sought to investigate whether scAR precisely learns the quantity of native signals using a public CITE-seq dataset of peripheral blood mono-nuclear cells from 10x genomics[31] (PBMC5k). Prior to sequencing, PBMC cells were stained with a panel of 32 antibody-conjugated oligos which consist of surface markers of B cells, T cells, Natural killer cells (NK), monocytes and Dendritic cells (DC) (Fig. 3a). According to ambient signal hypothesis (Fig. 1a), unbound antibodies in single cell suspension will interfere with both cell-containing and cell-free droplets and antibody frequencies should be highly correlated. We calculated the antibody frequencies in randomly sampled cell-free droplets (from 0.01% to 100%) and compared them with cell-containing droplets (Fig. 3b,). As in the CROP-seq experiment (Fig. 2c), we also observed high correlation, as few as 31 randomly sampled cell-free droplets consistently show strong correlation to the cell-containing droplets.

We next discriminated cell types by clustering and annotating cells using transcriptome profiling (Fig. 3c and Methods) and examined antibody detection in cell-types. As expected, ambient protein signals are observed in cell types which do not express them (Fig. 3d-e and supplementary Fig. 4-6). For example, CD19 and CD20, which are specific markers of B cells, are detected in all other cell types (Fig. 3d and supplementary Fig. 4a), whereas T cell-specific markers like PD-1 and CD3D are detected in all other cell types (Fig. 3e and supplementary Fig. 4b). Similarly, we detected NK-specific markers CD56 and CD335 in non-NK cells (supplementary Fig. 4c-d) and monocyte specific marker CD14 in non-monocytes (supplementary Fig. 4e). By applying scAR, the ambient signals were removed and denoised protein counts correlated stronger with its corresponding RNA expression per cell type (Fig. 3d-e and supplementary Fig. 4-6). Our finding is underlined by the fact that these ambient counts are highly diversified (mean 9.6 ± STD 43.4) per marker per cell. Remarkably, scAR identifies the true signals even if the native expression is as low as background noise. For example, before denoising, B cells, NK cells, monocytes and naïve T cells show similar level of the T cell marker PD-1 (Fig. 3e and supplementary Fig. 4f); after denoising, all PD-1 counts are removed in B cells, NK cells, monocytes and DC cells but not in T cells. To test whether scAR over-corrects the signals, we calculated the scAR-estimated ambient ratios of these specific markers and compared them with a naïve approach, which simply considered all non-specific expression as ambient signals. The total ambient ratios of these markers are comparable (supplementary Fig. 4g), suggesting that scAR assigns reasonable number of UMI counts as ambient signals.

We next compared the averaged protein counts by cell type before and after denoising and aligned them with RNA expression (Fig. 3f). We observed strong background signals in almost all antibodies in raw counts, while scAR significantly reduces the background noise, leading to more specific expression of markers in subtypes. For example, denoised CD4 and CD8 are exclusively present in CD4+ T cells and CD8+ T cells, respectively (Fig. 3f and supplementary Fig. 5a-b). Naïve and memory T cells are distinguishable with CD197 and CD45RA (Fig. 3f and supplementary Fig. 5c-d). Regulatory T cells (Treg) show high CD25 and negative CD127 as reported[32] (Fig. 3f and supplementary Fig. 5e-f). Similarly, we observed high CD14, negative CD16 in CD14 monocytes, low CD14 and high CD16 in CD16 monocytes and high HLA-DR protein counts in intermediate monocytes after scAR denoising (Fig. 3f and supplementary Fig. 6a-c). In most of these cell types, scAR removes fewer than 10% of raw protein counts (supplementary Fig. 6d). As a result, Spearman correlation coefficient between RNA-protein pairs is increased in single cells (Fig. 3g). Together, scAR removes ambient signals while preserves the true signals, resulting in reliable quantification of native protein counts in CITE-seq.

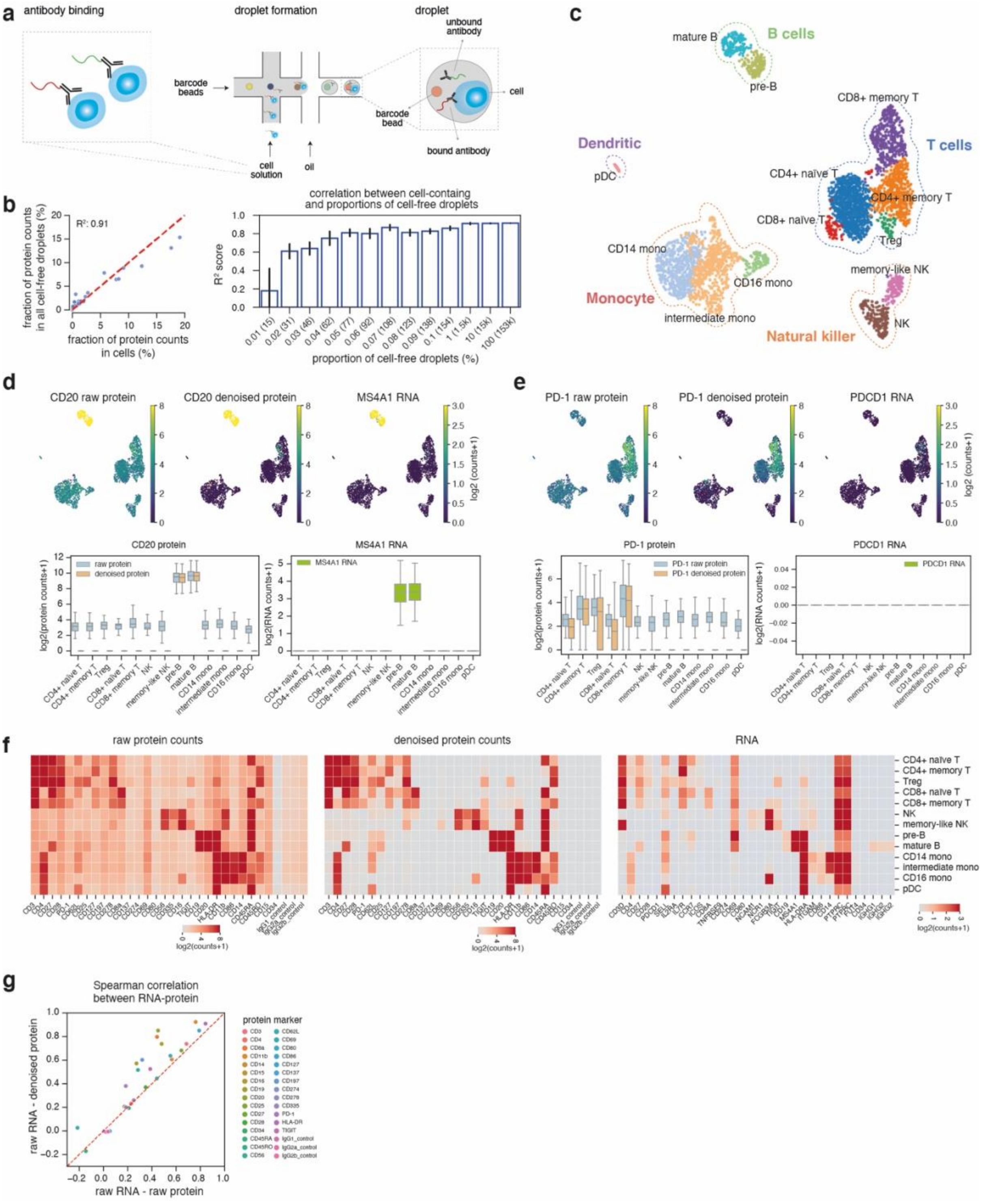


Fig. 3 | scAR removes the ambient protein counts in CITE-seq data. A public CITE-seq dataset highlights the presence of ambient contamination in protein count data and demonstrates the functionality of scAR on removing ambient protein signals. **a,** Similar to **Fig.1a**, illustration of the ambient signal hypothesis in the context of protein counts. **b,** correlation analysis of protein counts between cell-containing and -free droplets, as with **Fig. 2b-c**. The scatterplot shows the fraction of protein counts between all cell-containing (x-axis) and all cell-free droplets (y-axis). The barplot shows the R^2 scores, indicating the correlation of protein frequencies between cells and randomly sampled cell-free droplets. X-axis shows the proportions of cell-free droplets. The droplet numbers are indicated in brackets. The error bars indicate the variance of 10 samplings. **c,** UMAP of the PBMC5k dataset. Cell types are annotated using transcriptome data (Methods). **d-e,** Two selected examples, CD20 antibody (**d**) and PD-1 antibody (**e**) to demonstrate scAR's performance. UMAPs visualize raw protein counts, denoised protein counts and corresponding RNA counts, respectively. Color bars represent log2(counts +1). Barplots below plot the single cell counts grouped by cell types. **f,** Heatmaps show the average row protein counts (left), average denoised counts (middle) and average corresponding RNA counts (right) in different cell types. Columns represent the antibodies used in this dataset; rows represent cell types. **g,** scatterplots show the Spearman correlation coefficients between RNA-protein pairs before (x-axis) and after scAR-denoising of protein counts (y-axis). The red dashed line represents y=x. Dots represent antibodies.

scAR removes ambient signals for mRNA counts.

To further demonstrate the broad application of scAR, we selected another public dataset which pools equal numbers of human HEK293T cells and mouse NIH3T3 cells for single cell RNA sequencing[33]. Reads were mapped to a combined human-mouse reference genome with Cellranger and all ambiguous ones which can map to both species were excluded. We then classified the cells as human, mouse or multiplets by unsupervised clustering of mRNA (Methods and supplementary Fig. 7a). It results in 7590 HEK293 cells, 8006 NIH3T3 cells and 697 mixed droplets which contains both HEK293 and NIH3T3 cells. We detected similar number of total human and mouse specific transcripts in cell-free droplets, as expected, they are proportional to those in cells (supplementary Fig. 7b). In addition, both human cells and mouse cells exhibit low level exogenous contamination in raw counts (Fig. 4a-b). HEK293 cells contain ~1.4% mouse transcripts on average and NIH3T3 cells contain ~2.2% human transcripts on average. It should underline that contamination can also happen between the same species, e.g., a HEK293T cell-containing droplet may not only contain mouse transcripts but also human transcripts from ambient source. Namely, the actual contamination ratio should be greater than exogenous contamination.

We applied scAR to cell-containing droplets and observed that scAR removes all exogenous transcripts (Fig. 4b). In details, nearly all mouse transcripts were identified and removed in HEK293T cells and nearly all human transcripts were also identified and removed in mouse cells. On the other hand, the ratio of mouse and human transcripts remains 1:1 in the multiplets after denoising.

Interestingly, we also noticed that the ambient frequencies (α in Fig. 1b) are varying between different subpopulations of droplets. We sorted all droplets by their UMI counts and identified four subgroups of droplets by kneeplot: cell-containing, droplet I and droplet II and cell-free droplets (Fig. 4c and Methods). We next found that cell-free droplets show stronger correlation (R^2=0.85) to the cell-containing droplets than droplet I (R^2=0.33) and II (R^2=0.42) (supplementary Fig. 7c). When taking different ambient frequencies as input, scAR outputs different estimated contamination rates (Fig. 4d-g and supplementary Fig. 7d). Overall, scAR can precisely predict the cross-species ambient signals while the prediction for inter-species ambient signals depends heavily on the input ambient frequencies. Given that the global ratio of human and mouse transcripts is ~1.11 in cells (supplementary Fig. 7b), it is reasonable to expect a similar contamination ratio of human and mouse sources. However, the droplet I and II lead to too high estimation of human-source contamination, as much as ~3x of mouse source (Fig. 4e-f). This may be explained by the over-representation of human transcripts in droplet I and II (supplementary Fig. 7b). The higher human transcripts in ambient frequencies as input, the more counts to be identified as background noise by scAR. The best estimate of ambient frequencies should be drawn from population of cell-free droplets, as the estimated noise ratios are in a reasonable range in both cell lines – ambient signals from human sources are slightly stronger than mouse sources in both cell lines (Fig. 4g and supplementary Fig. 7d). These observations also suggests that compositions in these droplets are clearly different, e.g., droplet I and II may contain more human cell debris. In addition, it also suggests that a precise estimation of ambient frequencies is a key to noise reduction.

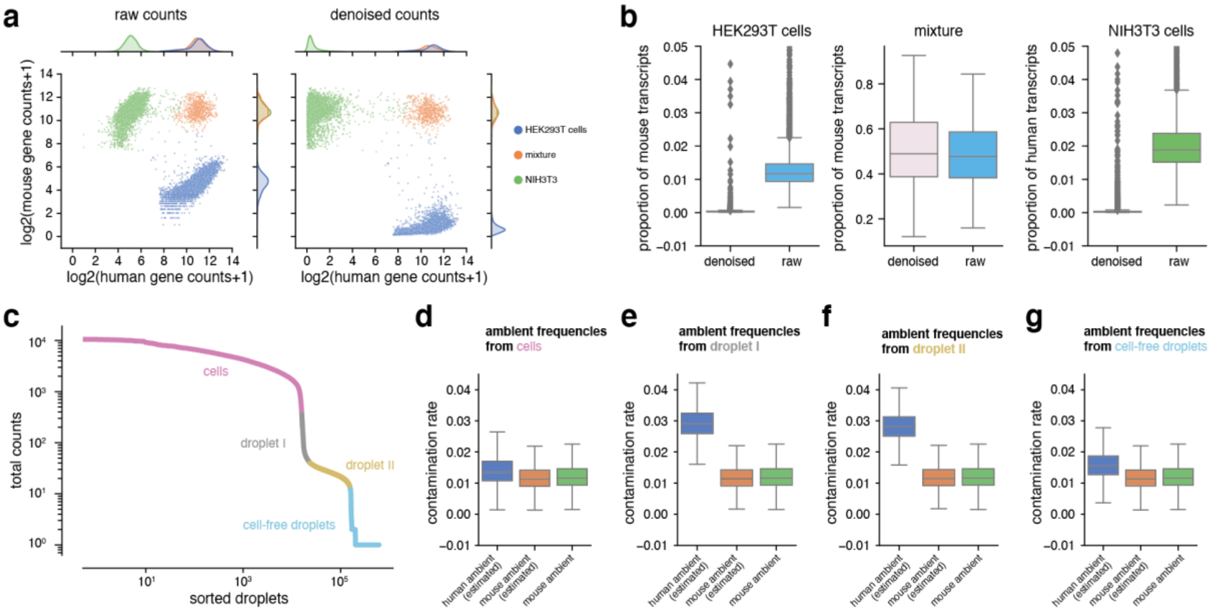


Fig. 4 | scAR reduces the noise in mRNA counts. A public scRNAseq dataset of mixed human HEK293T and mouse NIH3T3 cells (1:1) was selected to demonstrate scAR's ability in noise reduction in transcriptome data. **a,** Scatterplots show transcript composition before (left) and after (right) denoising in three populations, HEK293T, NIH3T3 and multiplets. X- and y-axis show transcripts which are exclusively mapped to human or mouse genome, respectively. **b,** Quantification of exogenous contamination before and after denoising in three populations. Y-axis represents per cell fraction of exogenous transcripts, i.e. mouse transcript rate in all HEK293T cells and human transcript rate in all NIH3T3 cells. **c,** The kneeplot shows subpopulations of droplets. **d-g,** Boxplots show the percentage of ambient signals in HEK293 cells. The green boxes represent the proportion of observed mouse transcripts. The blue and orange boxes represent scAR-estimated human and mouse ambient proportions, respectively. Ambient frequencies are averaged from cells (**d**), droplet I (**e**) or droplet II (**f**) or cell-free droplets (**g**).

Benchmarking of methods for UMI denoising.

To evaluate scAR in comparison with other state-of-the-art methods[11–13], we used the CITE-seq dataset and calculated the Spearman correlation coefficients of protein-RNA pairs as the benchmarking metric (Fig. 5 and supplementary Fig. 8, Methods). totalVI uses the module of scVI to correct for background noise for mRNA data, so we skipped scVI to avoid redundance. The results show that scAR outperforms totalVI and DCA in denoising of both or either of mRNA and protein data (Fig. 5b and supplementary Fig. 8b). In addition, totalVI trains VAE using both mRNA and protein data as input, while scAR separately denoises mRNA and protein data, meaning scAR is a more unbiased approach. DCA does not show comparative performance mainly because it is designed to denoise mRNA data.

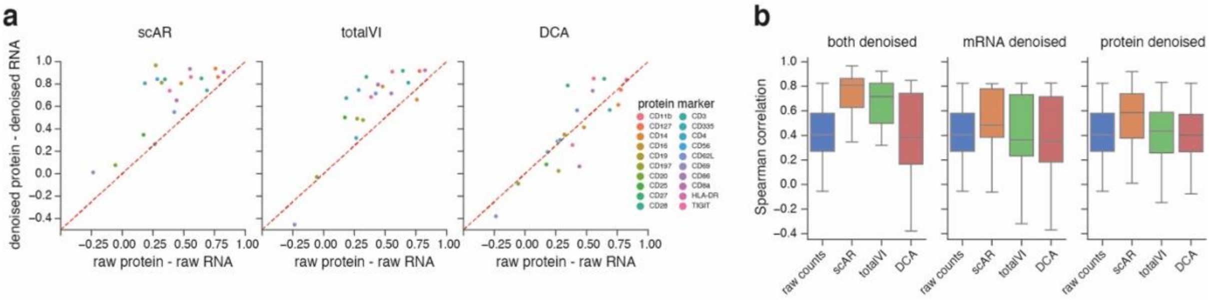


Fig. 5 | Benchmarking of protein and mRNA count denoising. a, Scatterplots show the Spearman correlation coefficients between RNA-protein pairs before (x-axis) and after denoising of both protein and mRNA counts (y-axis). The red dashed line represents y=x. Dots represent antibodies. **b,** Boxplots show Spearman correlation coefficients between RNA-protein pairs (both denoised, only mRNA denoised, and only protein denoised). Denoising methods are indicated on X-axis.

Discussion

Versatile single-cell omics technologies have expanded understanding of single cell biology and development of new technologies is constantly pushing the boundary further. In this context, we developed scAR to provide a reliable 'one-for-all' solution to UMI count denoising for multiple single-cell omics technologies.

scAR can precisely infer the native signals for protein data in CITE-seq and mRNA data in scRNAseq. Recent approaches[12,13,26,34,35] introduce deep learning technologies (such as AE and VAE) for these tasks and show great promise. These approaches proposed noise models which stick to the zero-inflation pattern of these count data and estimate all the parameters through neural networks[12,14,35]. In scAR, we constrain the noise model under the ambient signal hypothesis and empirically estimate a parameter from cell-free droplets, this roughly reduces parameters by one third and focuses the VAE on learning the biology-related native expression and the noise ratio. As a result, this hypothesis-driven modeling significantly improves the performance comparing to the ones that disregard the ambient signal hypothesis and cell-free droplets (Fig. 5 and supplementary Fig. 9). Moreover, it generalizes scAR to fit a broader range of single-cell omics datasets, independent of the sparsity of the data.

scAR can evaluate the probability of ambient contamination. This can ensure accurate assignment of identity barcode for a class of single-cell omics technologies, including single cell CRISPR screens (e.g., CROP-seq, Perturb-seq and CRISP-seq)[1,2,28,29] and cell indexing[15–17]. We tested scAR on CROP-seq, but it should fit other ones in this class as they take similar protocols to prepare, construct and sequence the libraries of feature barcodes (either sgRNA or identity barcode). Most of current studies have assigned exogenous barcodes by hard filtering approaches[28,30], which filter out cells with low depth and perform naïve assignment afterwards. This is not only inaccurate (Fig. 2) but also inefficient as it can further discard as many as >50% of cells[36]. Other approaches such as MUSIC[36] and scMAGeCK[37] propose to model single cell CRISPR data by linking transcriptome profile to sgRNA assignment. On the one hand, these methods are specific to single cell CRISPR data, and on the other, there is a risk of being misled by potentially dominant transcriptional states (e.g. cell cycle), when several nodes of the same pathway are being interrogated, or when the phenotypic effect of the perturbation is low (in the case of, e.g., low effective sgRNAs or wrong time points). scAR provides an accurate, unbiased and efficient solution to assignment of identity barcode for this class of technologies.

Nonetheless, besides ambient contamination, other technical factors[9,10] can also introduce background noise. We assume in scAR that ambient source is the most predominant artifact and in turn this hypothesis seems to be confirmed by the outstanding performance of scAR. However, further experimental validation may still be required. In addition, in CITE-seq technology, the non-specific binding of antibodies may bring in extra noise[4,14], this is not modeled in scAR as we consider it too specific (dependent of the antibody and experimental cell lines) to violate the scope of generality of scAR. Moreover, identification of this noise may require dedicated well-designed experiments (e.g., spike-in[4]), as models can hardly distinguish between specific and non-specific binding without human knowledge.

Finally, we observed different contamination levels in different datasets. scAR's ability to estimate noise ratio may allow to evaluate batch effects and guide the experimental design, such as the protocols for cell fixation and washing. Furthermore, scAR can have great potential in facilitating technology development in droplet-based single-cell omics, given the common and nonnegligible presence of ambient noise. For example, the most recent scifi-RNA-seq[38] achieves ultra-high-throughput by leveraging cell indexing technologies to encapsulate and sequence multiple cells in a droplet. scAR may have great potential in deconvoluting the cell identity in this complex setting.

Reference

1. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
2. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
3. Lopes, R. *et al.* Systematic dissection of transcriptional regulatory networks by genome-scale and single-cell CRISPR screens. *Sci. Adv.* **7**, 1–16 (2021).
4. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
5. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
6. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
7. Young, M. D. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, 303727 (2020).
8. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
9. Smith, T., Heger, A. & Sudbery, I. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
10. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 1–6 (2018).
11. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 1–14 (2019).
12. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
13. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
14. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. (2021).
15. Biddy, B. A. *et al.* Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
16. Guo, C. *et al.* CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.* **20**, 90 (2019).
17. Gehring, J., Hwee Park, J., Chen, S., Thomson, M. & Pachter, L. Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins. *Nat. Biotechnol.* **38**, 35–38 (2020).
18. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
19. Lun, A. T. L. *et al.* EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 1–9 (2019).
20. Muskovic, W. & Powell, J. E. DropletQC: improved identification of empty droplets and damaged cells in single-cell RNA-seq data. *Genome Biol.* **22**, 329 (2021).
21. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *31st Int. Conf. Mach. Learn. ICML 2014* **4**, 3057–3070 (2014).
22. Ranganath, R., Gerrish, S. & Blei, D. M. Black box variational inference. *J. Mach. Learn. Res.* **33**, 814–822 (2014).
23. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
24. McDonald, E. R. *et al.* Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577-592.e10 (2017).
25. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).
26. Genga, R. M. J. *et al.* Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. *Cell Rep.* **27**, 708-718.e10 (2019).
27. 10x genomics. 30k A549, Lung Carcinoma Cells, Treatments Transduced with a CRISPR Pool, Multiplexed, 6 CMOs. (2021).
28. 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).

29. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e15 (2016).
30. Tian, R. *et al.* CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron* **104**, 239-255.e12 (2019).
31. 10x genomics. 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry). (2019).
32. Liu, W. *et al.* CD127 expression inversely correlates with FoxP3 and suppressive function of human CD4+ T reg cells. *J. Exp. Med.* **203**, 1701–1711 (2006).
33. 10x genomics. 20k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 cells, 3' HT v3.1. (2021).
34. Grønbech, C. H. *et al.* ScVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).
35. Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv* 791699 (2019) doi:10.1101/791699.
36. Duan, B. *et al.* Model-based understanding of single-cell CRISPR screening. *Nat. Commun.* **10**, 2233 (2019).
37. Yang, L. *et al.* ScMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biol.* **21**, 1–14 (2020).
38. Datlinger, P. *et al.* Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
39. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc.* 1–14 (2014).
40. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
41. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. *32nd Int. Conf. Mach. Learn. ICML 2015* **2**, 1613–1622 (2015).
42. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).
43. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, (2018).

Methods

The scAR model.

scAR uses hypothesis-driven probabilistic deep learning to infer the biological and technical variation in droplet-based single-cell omics experiments. The raw counts consist of biological signals and technical artifacts, which are modelled with a Binomial model.

Modeling count data using Binomial regression.

We take a generative approach to modeling the observed count matrix $X \in \mathbb{N}_0{}^{M \times G}$, which denotes *M* cells and *G* features (i.e., genes, antibodies, sgRNAs or identity barcodes). A graphic model representation of this generative model is summarized in supplementary Fig. 10. For a given cell *m*, x_m represents a G-dimensional vector of observed expression data. We assume that x_m is drawn from a Multinomial model:

$$x_m \sim Multinomial(d_m, prob = \theta_m) \qquad (1)$$

where d_m is the library size of cell *m*, θ_m is the feature frequencies. Therefore, for feature *g* in cell *m*, the observed count x_{mg} is drawn from a Binomial model:

$$x_{mg} \sim Binomial\left(d_m, prob = \theta_{mg}\right) \qquad (2)$$

where θ_{mg} represents probability of observing feature *g* in cell *m*. It determined by two factors, native expression n_{mg} and ambient signals a_{mg}, which can be modelled as,

$$\theta_{mg} = n_{mg} + a_{mg} \qquad (3)$$

$$n_{mg} = (1 - \varepsilon_m) \times \beta_{mg} \qquad (4)$$

$$a_{mg} = \varepsilon_m \times \alpha_{mg} \qquad (5)$$

where $\varepsilon_m \in [0{,}1]$ is a hidden variable, representing the fraction of total ambient counts. β_{mg} is another hidden variable, representing the feature frequency of native expression of feature *g* in cell *m*. α_{mg} represents the ambient frequency and according to ambient signal hypothesis, it is independent of cells, so we get,

$$\alpha_{mg} = \alpha_g \qquad (6)$$

Notably, the cell-free droplets can be expressed as equation (2), with the native component being zero and noise ratio being 1, so,

$$x'_{eg} \sim Binomial\left(d'_e, prob = \alpha_g\right) \qquad (7)$$

Where, x'_{eg} and d'_e represent counts for feature *g* and library size in cell-free droplet *e*, respectively. According to law of large numbers, we can approximate α_g by averaging feature counts in cell-free droplets,

$$\alpha_g = \frac{\sum_E x'_{eg}}{\sum_E \sum_G x'_{eg}} \qquad (8)$$

Put all together, we have,

$$x_{mg} \sim Binomial\left(d_m, prob = (1-\varepsilon_m) \times \beta_{mg} + \varepsilon_m \times \frac{\sum_E x'_{eg}}{\sum_E \sum_G x'_{eg}}\right) \qquad (9)$$

where only ε_m and β_{mg} are unknown parameters that need to be estimated. According to Bayes' theorem, we get the posterior probability,

$$p(\varepsilon_m, \beta_{mg} |\, x_{mg}) = \frac{p(x_{mg} \mid \varepsilon_m, \beta_{mg}) \times p(\,\varepsilon_m, \beta_{mg})}{p(x_{mg})} \qquad (10)$$

Since the prior probability $p(\,\varepsilon_m,\ \beta_{mg})$ and likelihood $p(x_{mg} \mid \varepsilon_m,\ \beta_{mg})$ both are intractable (unknown or difficult to factorize over samples), we implement variational inference[23,39] to estimate ε_m and β_{mg}, as described in the following section. To ensure flexibility, we also provide implementations of Poisson model to allow users to choose and test.

Variational inference for scAR

We apply variational autoencoders to optimize the hidden variable ε_m and β_{mg} mentioned above. The architecture of VAE is demonstrated in supplementary Fig. 11. We introduce an additional latent variable z in bottleneck layers so the marginal log-likelihood of observation x_m can then be written as,

$$\log p_\varphi(x_m) = -\log p_\varphi(z, \varepsilon_m, \beta_m |\, x_m) + \log p_\varphi(z, \varepsilon_m, \beta_m, x_m) \qquad (13)$$

where φ represents the parameter space, i.e., model weights. ε_m and β_m are calculated by deterministic neural networks (decoder),

$$\varepsilon_m,\ \beta_m = f_\eta(z) \qquad (14)$$

where, f represents neural networks and $\eta \subset \varphi$ represents the trainable weights of f. This means,

$$p_\eta(\,\varepsilon_m,\ \beta_m \mid z) = 1 \qquad (15)$$

Therefore, we can integrate out ε_m and β_m and re-write the equation (13) as

$$\log p_\varphi(x_m) = -\log p_\varphi(z \mid x_m) + \log p_\varphi(z, x_m) \qquad (16)$$

We construct variational posterior $q(\varphi \mid \omega)$ to approximate the posterior $p(\varphi \mid x_m)$. Therefore, we have,

$$\log p_\varphi(x_m) = \mathbb{E}_{z\sim q_\omega}\left[\log \frac{q_\omega(z|\,x_m)}{p_\varphi(z|x_m)}\right] + \mathbb{E}_{z\sim q_\omega}\left[\log p_\varphi(z, x_m) - \log q_\omega(z \mid x_m)\right] \quad (17)$$

where the first term on the right side is the Kullback-Leibler divergence[40,41] between distributions q and p, reflecting the difference between parameter distributions. It is non-negative, so we can get the evidence lower bound (ELBO) as follows,

$$\log p_\varphi(x_m) \geq \mathbb{E}_{z\sim q_\omega}\left[\log p_\varphi(z, x_m) - \log q_\omega(z \mid x_m)\right] =: \mathcal{L}(\varphi, \omega; x_m) \quad (18)$$

Increasing the ELBO will approximate the distribution q to p thereby ensuring the learnt variables are as close as the expectation. Therefore, the ELBO is generally used as the objective function to fit the VAE. We can further transformation equation (18) into,

$$\mathcal{L}(\varphi, \omega; x_m) = -D_{KL}(q_\omega(z|x_m) \,||\, p_\varphi(z)) + \mathbb{E}_{z\sim q_\omega}\left[\log p_\varphi(x_m|z)\right] \qquad (19)$$

The negative of ELBO is used as loss function to simultaneously optimize model weights and hidden variables in scAR. In case of M cells, the loss function is then written,

$$Loss = D_{KL}(q_\omega(z|x) \,||\, p_\varphi(z)) - \mathbb{E}_{z\sim q_\omega}\left[\textstyle\sum_i^M \log p_\varphi(x_i|z)\right] \qquad (20)$$

Minimizing the loss function requires a tradeoff between the KL divergence and expected negative log-likelihood term. On the one hand, the KL divergence between $q_\omega(z|x)$ and $p_\varphi(z)$ should be kept small, preventing the variational posterior from being too different to the prior. On the other, the variational posterior parameters should maximize the log likelihood $\log p_\varphi(x|z)$, ensuring a small reconstruction error of scAR. We use the reparameterization trick to calculate the gradients with respect to φ and ω for KL term[41]. According to equations (14) and (15), we have,

$$p_\varphi(x_m|z) = p_\varphi(x_m \mid \varepsilon_m, \beta_m) \qquad (21)$$

Since we assume x_m is drawn from Binomial distribution with latent parameters ε_m, β_m (see equation (9)), $p_\varphi(x_m \mid \varepsilon_m, \beta_m)$ also has a closed-form expression, thus the gradient descents of negative log-likelihood term in equation (20) are easy to calculate. Together, we use the gradients of the loss function to update the parameters φ and ω to determine the hidden variables noise ratio ε and expected native frequencies β.

Bayesian inference and assignment of identity barcode

We infer the expected native signals $\bar{n}_m$ and ambient signals $\bar{a}_m$ in cell *m* using the following equations,

$$\bar{n}_m = (1 - \varepsilon_m) \times d_m \times \beta_m \qquad (22)$$

$$\bar{a}_m = \varepsilon_m \times d_m \times \alpha \qquad (23)$$

where $\bar{n}_m$ is used as denoised counts for CITE-seq and scRNAseq.

For assignment of sgRNAs in CROP-seq, we use Bayesian factor as a metric to compare two hypotheses: observed counts consist of both native and ambient sgRNAs (H_1) vs observed counts contain only ambient sgRNAs (H_2). For a given sgRNA *g* in cell *m*, this can be mathematically expressed as follows,

$$H_1^{mg} := x_{mg} \sim Binomial(d_m, prob = \bar{a}_{mg} + \bar{n}_{mg})\ versus\ H_2^{mg} := x_{mg} \sim Binomial(d_m, prob = \bar{a}_{mg}) \qquad (24)$$

The Bayesian factor then is given by,

$$K_{mg} = \frac{Pr(x_{mg} | H_1^{mg})}{Pr(x_{mg} | H_2^{mg})} \qquad (25)$$

The numerator and denominator represent the probability that x_{mg} is produced under assumption of H_1^{mg} and H_2^{mg}, and we approximate them using the cumulative distribution function (stats.binom.cdf) and probability mass function (stats.binom.pmf), respectively. High K_{mg} (>=3) favors the first hypothesis, meaning the sgRNA *g* contains native signal. In the case of multiple high K, we assign the sgRNA of highest K to the cell.

Model optimization for scAR

To identify a best universal set of hyperparameters as the default setting of scAR, we perform grid search on two types of synthetic datasets (see supplementary note I), which simulate CROP-seq data type and CITE-seq/scRNAseq data type respectively. To limit the number of parameters, we fix several less important parameters. For example, the training epochs are fixed at 800. Additionally, we use the Adam optimizer[42] with exponential decay to schedule the learning rate but the decay rate is fixed at 0.97 every 5 epochs. The hyperparameters which are optimized include the number of nodes of neural networks, dimension of latent space, dropout probability of neurons, initial learning rate and KL divergence weight. As a result, we tested 6912 combinations of parameters for each dataset (supplementary Fig. 12) and identified the best set listed as follows: units of 1st layer: 150; units of 2nd layer: 100, dimension of latent space: 15; initial learning rate: 0.001, dropout probability: 0; KLD weight: 1e-

5. All experiments were performed using these optimized parameters unless otherwise specified.

It is also worth noticing that we use either ReLU or modified Softplus activation (see supplementary note II) functions to output β in decoder depending on the sparsity of expected native matrices. For example, CROP-seq or cell indexing datasets are extreme sparse as each cell is expected to have a single (or a few) native identity feature barcode, so we use ReLU as the activation function to generate sparse β; while CITE-seq datasets are generally denser, so we use a modified Softplus as the activation function to avoid too many zeros in β.

CROP-seq experiment.

The CROP-seq library was cloned into a modified pLKO-TET-ON plasmid in a pooled format by Golden Gate. The cloning reaction product was used to transform Endura electrocompetent cells, which were expanded in LB medium overnight (OD600 = 0.8) and plasmid DNA was harvested using Genopure plasmid maxi kit (Roche). We produced lentiviral particles and transduced MCF7-dCas9-KRAB cells (MOI = 0.3) with the CROP-seq library. The cells were selected with 2µg/ml puromycin (Invitrogen) and they harvested at defined time points by FACS (mCherry-positive cells). The single-cell suspensions were fixed in 90% methanol in DPBS (v/v) and stored at -80 °C prior to rehydration and further processing. The rehydration buffer was supplemented with 1% Bovine serum albumin and 0.5 U/ul RNase inhibitor (Sigma, P/N 3335399001). All samples were processed using Chromium Next GEM single-cell 3' reagents kit (10x Genomics) according to the manufacturer's protocol and the libraries were sequenced in an Illumina HiSeq 2500.

Pooled CRISPR screen.

MCF7-CRISPRi cells were transduced with independent lentiviral pools (MOI = 0.3) of the CROP-seq library. To guarantee a correct representation of all sgRNAs in the cell population we transduced ≈1000 cells per plasmid. The cells were selected using 2µg/ml puromycin (Invitrogen) at 24 hours post-transduction, after which they were expanded and harvested at indicated time points. We extracted gDNA from the cells using DNeasy kit (Qiagen) and prepared libraries for next generations sequencing.

Analysis of CROP-seq data.

Single-cell sequencing data were processed using Cell Ranger (version 3.1.0, 10x Genomics) and sgRNA count matrices were generated using KITE (https://github.com/pachterlab/kite). Human genome assembly (Ensembl GRCh38 release-98) was used as the reference to map mRNA reads. The Scanpy package[43] was used to perform quality control, cell filtering, gene filtering and differential expression analysis. We used two normalization approaches to examine the knockdown effect. The first one as shown in Fig. 2g-i is library size normalization. Sequencing depth per cell was normalized to 1.0xe^5 counts and t-test was performed on the normalized counts across cell groups using scipy.stats.ttest_ind function. The second one (supplementary Fig. 3a) is Z-normalization as reported in our previous publication[3]. For each gene, we subtracted mean value of CTL group then divided by standard deviation of CTL group.

Analysis of CITE-seq data.

The cellranger outputs of PBMCs5k[31] dataset were downloaded from 10x genomics. Cells with extreme counts (<1500 counts or >15000 counts) were discarded. Stressed cells with high presence of mitochondrial genes (>=0.2) were also discarded. The cell clustering was performed using Scanpy and annotated based on expression of a panel of marker genes.

Correlation of RNA-protein pairs. Both raw and denoised RNA counts were library size normalized. Raw and denoised protein counts were used without any normalization as library sizes of them represent cell type variance. Spearman's correlation was performed between RNA and protein counts using scipy.stats.spearmanr function. Control antibodies were removed for this correlation analysis. CD45RA and CD45RO, which are encoded by an identical gene PTPRC were also removed due to the difficulty of identifying isoform transcripts. In addition, several markers (CD15, CD34, CD80, CD137, CD274, CD278, PD-1) were removed due to extremely low counts of either protein or corresponding RNAs. On the other hand, a version of full antibodies was also plotted in supplementary Fig. 8.

Species-mixing experiment.

The cellranger outputs of species-mixing dataset[33] were downloaded from 10x genomics. Scanpy was used to perform quality control, gene filtering, cell filtering and species identification. In brief, we first took the 'filtered_feature_bc_matrix' from cellranger output and further filtered out genes with extreme counts (<200 or >6000 in total) and cells with low gene counts (<200). We then performed library size normalization, log transformation, clustering and UMAP. By checking the differently expressed genes, we identified 7590 HEK293T cells, 8005 NIH3T3 cells as well as 697 multiplets mixed with both HEK293T and NIH3T3.

Examination of droplets. To identify the best representation of ambient signals, we examined subpopulations of droplets in the unfiltered matrix – namely, 'raw_feature_bc_matrix'. All droplets were ranked by their total UMI counts and split into four subgroups through kneeplot: 1) droplets in 'filtered_feature_bc_matrix' were marked as cells, 2) droplets with high counts (>40) were marked as 'droplet I', 3) droplets with intermediate counts (>12 and <=40) were marked as 'droplet II', 4) droplets with low counts (<=12 and >0) were marked as 'cell-free droplets'. We took the total gene frequencies in each subpopulation as the ambient frequencies and run scAR to compare the estimated noise ratio.

Data availability.

The CROP-seq data discussed in this manuscript have been deposited to the Sequence Read Archive and are accessible through BioProject accession number: PRJNA794328. All other datasets are public. The CITE-seq datasets (PBMCs5k) and HEK293T and NIH3T3 pooled scRNAseq (20k_hgmm dataset) were downloaded from 10x genomics datasets. Other datasets were downloaded from Sequence Read Archive.

Code availability.

The package of scAR and codes to reproduce the results in this manuscript is available at Github (https://github.com/CaibinSh/scAR).

Acknowledgements.

We thank Joel Wagner and Joshua Korn for inspiring discussions, Mathias Eder and Esther Uijttewaal for additional technical support.

Author contributions.

C.S. and A.d. conceived and designed the study. C.S., G.L. and A.d. designed the statistical model and performed analysis. R.L. and G.G.G. designed CROP-seq and bulk sequencing experiments and R.L. conducted the experiments. A.W. and R.C. performed scRNAseq and bulk sequencing and S.S. performed preprocessing of CROP-seq data. S.D., A.K., E.D., G.G.G, G.R. and A.d. supervised the study. C.S. wrote the original draft. C.S., G.L., R.L., S.D., E.D. and A.d. reviewed and edited the draft.

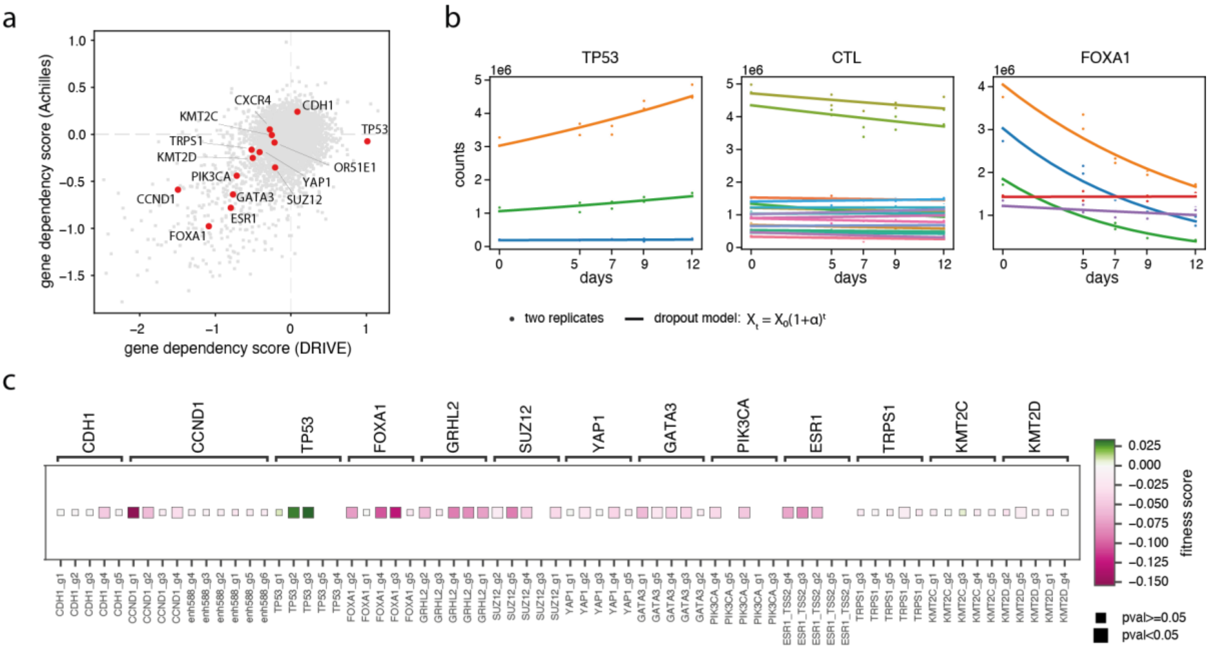
Supplementary Tables

sgRNA			Target	Group	
number	**sgRNA name**	**Guide sequence**	**genes**	**names**	**Library name**
1	CDH1_g1	AGTTCCGACGCCACTGAGAG	CDH1	CDH1	CROP-seq_pilot
2	CDH1_g2	GACTTGCGAGGGACGCATTC	CDH1	CDH1	CROP-seq_pilot
3	CDH1_g3	CCGAGAGGCTGCGGCTCCAA	CDH1	CDH1	CROP-seq_pilot
4	CDH1_g4	CGGTGACGACGGGAGAGGAA	CDH1	CDH1	CROP-seq_pilot
5	CDH1_g5	CCTCAGGACCCGAACTTTCT	CDH1	CDH1	CROP-seq_pilot
6	ESR1_TSS1_g1	TCCGTTCTGAGTCGGTAGAC	ESR1	ESR1_TSS1	CROP-seq_pilot
7	ESR1_TSS1_g2	AGCTCTTTAACAGGCTCGAA	ESR1	ESR1_TSS1	CROP-seq_pilot
8	ESR1_TSS1_g3	CTATAGAATGGGCAGGAGAA	ESR1	ESR1_TSS1	CROP-seq_pilot
9	ESR1_TSS1_g4	GTATGTTATCTGGAACAGAC	ESR1	ESR1_TSS1	CROP-seq_pilot
10	ESR1_TSS1_g5	GGTGTGCTGTACTAAGAAAA	ESR1	ESR1_TSS1	CROP-seq_pilot
11	FOXA1_g1	GCCGCCCGTCGCTTCGCACA	FOXA1	FOXA1	CROP-seq_pilot
12	FOXA1_g2	CCAACGCCACCCGGGCGAAG	FOXA1	FOXA1	CROP-seq_pilot
13	FOXA1_g3	CGCCTCCGCGGGAAGTGAGC	FOXA1	FOXA1	CROP-seq_pilot
14	FOXA1_g4	CAACTGCACTTGCCTCGCAG	FOXA1	FOXA1	CROP-seq_pilot
15	FOXA1_g5	GTGAGCGGGCTGCCTCTGCG	FOXA1	FOXA1	CROP-seq_pilot
16	GATA3_g1	CTGTGGCGCGACGCAACTTA	GATA3	GATA3	CROP-seq_pilot
17	GATA3_g2	GCAACGCAATCTGACCGAGC	GATA3	GATA3	CROP-seq_pilot
18	GATA3_g3	GCGGCGGCGTACGACCTGCT	GATA3	GATA3	CROP-seq_pilot
19	GATA3_g4	TTCGCTACCCAGGTTGGTAC	GATA3	GATA3	CROP-seq_pilot
20	GATA3_g5	TTAGGTCCTCCCAAGTGGTT	GATA3	GATA3	CROP-seq_pilot
21	GRHL2_g1	ACTAAAGGGTACAAGCCCGA	GRHL2	GRHL2	CROP-seq_pilot
22	GRHL2_g2	CGCGGAGTCCTCCTGGATCG	GRHL2	GRHL2	CROP-seq_pilot
23	GRHL2_g3	CCTCACCTAGCCGGAAAGGT	GRHL2	GRHL2	CROP-seq_pilot
24	GRHL2_g4	GTGTGTGAGAGCGCCCGAGA	GRHL2	GRHL2	CROP-seq_pilot
25	GRHL2_g5	CCTTGCGAGAAAGTTACCTG	GRHL2	GRHL2	CROP-seq_pilot
26	KMT2D_g1	AACAGACGAGATGCCTCCGG	KMT2D	KMT2D	CROP-seq_pilot
27	KMT2D_g2	GATAGAGGCGTCTCAAGTGC	KMT2D	KMT2D	CROP-seq_pilot
28	KMT2D_g3	GACAAGGGCGACTCCTCCAG	KMT2D	KMT2D	CROP-seq_pilot
29	KMT2D_g4	GGGCAATTCCTCAGGTGGCG	KMT2D	KMT2D	CROP-seq_pilot
30	KMT2D_g5	GGGCGATGCTTCAGGTGGTG	KMT2D	KMT2D	CROP-seq_pilot
31	KMT2C_g1	GACTAGGATGTCGTCGGAGG	KMT2C	KMT2C	CROP-seq_pilot
32	KMT2C_g2	CGCACTCACACACATCGGCG	KMT2C	KMT2C	CROP-seq_pilot
33	KMT2C_g3	GGATCCCGGTCCTCCTCCTG	KMT2C	KMT2C	CROP-seq_pilot
34	KMT2C_g4	AAATGCGAGAGGCTGAGCCG	KMT2C	KMT2C	CROP-seq_pilot
35	KMT2C_g5	TCTCGCATTTCCCGCAGCCC	KMT2C	KMT2C	CROP-seq_pilot
36	CTRL_g1	TCTCGTCTGATACCTCGGTC	OR2L13	CTL	CROP-seq_pilot
37	CTRL_g2	CTCATCGTGGTCGGCGGTCG	OR2L13	CTL	CROP-seq_pilot
38	CTRL_g3	GCGGCGTCTTTGGCAGTAGT	OR2L13	CTL	CROP-seq_pilot
39	CTRL_g4	GGCGTGCTTGCGGGTCCAGG	OR2L13	CTL	CROP-seq_pilot
40	CTRL_g5	CGCTGCTGCGAGACCAGCCG	OR2L13	CTL	CROP-seq_pilot
41	CTRL_g6	ACTCACCTCAACCGTATGGA	CTL	CTL	CROP-seq_pilot
42	CTRL_g7	CTGCAAGTAACCCATGCACC	CTL	CTL	CROP-seq_pilot
43	CTRL_g8	ATGCACTCAGCAAGTCTAAC	CTL	CTL	CROP-seq_pilot
44	CTRL_g9	GGCTGTGAAGAACCAGAAGT	CTL	CTL	CROP-seq_pilot
45	CTRL_g10	GCTGCCTGTCCTTTGAGTCA	CTL	CTL	CROP-seq_pilot
46	CTRL_g11	CCGCAGCAATATCTTGGCTC	CTL	CTL	CROP-seq_pilot
47	CTRL_g12	GGGCTCTCCAACTCACCAGG	CTL	CTL	CROP-seq_pilot
48	CTRL_g13	TGCTCAGCAGACTAGGCAGC	CTL	CTL	CROP-seq_pilot
49	CTRL_g14	GAAGCTCTGCTCAGCAGACT	CTL	CTL	CROP-seq_pilot
50	CTRL_g15	TCTGTCTCTGAGCTAGACTT	CTL	CTL	CROP-seq_pilot
51	TRPS1_g1	GACGTAATGCGCGGAGACTG	TRPS1	TRPS1	CROP-seq_pilot
52	TRPS1_g2	CTTGAAACTGACGTAATGCG	TRPS1	TRPS1	CROP-seq_pilot
53	TRPS1_g3	AGAGCAATCGAGAGGACGCG	TRPS1	TRPS1	CROP-seq_pilot
54	TRPS1_g4	AAGGCGAGAGAGCAATCGAG	TRPS1	TRPS1	CROP-seq_pilot
55	TRPS1_g5	GGATGTGCCCGGTGCCGGGT	TRPS1	TRPS1	CROP-seq_pilot
56	YAP1_g1	CCGCCAGACCAGTGGAGCCG	YAP1	YAP1	CROP-seq_pilot
57	YAP1_g2	CCTCCGTCAAGGGAGTTGGA	YAP1	YAP1	CROP-seq_pilot
58	YAP1_g3	CGGCGCTGTCCTCGCTCTCA	YAP1	YAP1	CROP-seq_pilot
59	YAP1_g4	GGCGAGTTTCTGTCTCAGTC	YAP1	YAP1	CROP-seq_pilot
60	YAP1_g5	CTGCGAGGCACTCGGACCTG	YAP1	YAP1	CROP-seq_pilot
61	CTL _g16	TAGATCTGAAAGGCTGGGAT	CTL	CTL	CROP-seq_pilot
62	CTL _g17	TGTCTCCTACTGCGTGTTGA	CTL	CTL	CROP-seq_pilot
63	CTL _g18	TCTTAATGATAGAATCTTCC	CTL	CTL	CROP-seq_pilot
64	CTL _g19	GCTCCCAGTGTCCTGTGATA	CTL	CTL	CROP-seq_pilot
65	CTL _g20	AAGCACCCAGTAGTAAAACA	CTL	CTL	CROP-seq_pilot

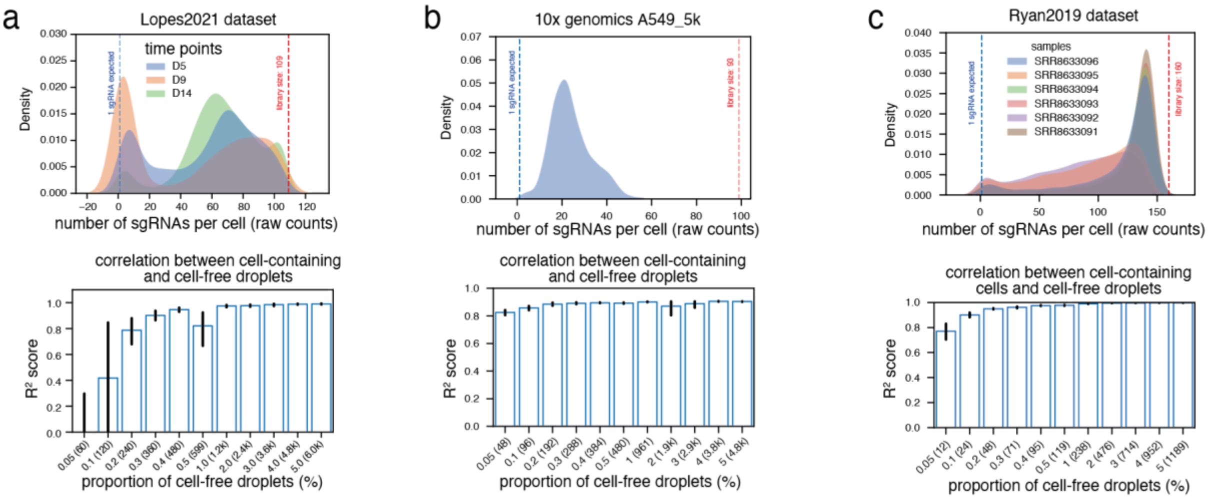
66	CTRL_g21	CTGAAAAAGGAAGGAGTTGA	CTL	CTL	CROP-seq_pilot
67	CTRL_g22	AAGATGAAAGGAAAGGCGTT	CTL	CTL	CROP-seq_pilot
68	CTRL_g23	TGCGCGGCTTGGGAAGCCCA	CTL	CTL	CROP-seq_pilot
69	CTRL_g24	GACGCGAGGAAGGAGGGCGC	CTL	CTL	CROP-seq_pilot
70	enh588_g1	GGATCTGCAGGCCCAAGGTC	CCND1	enh588	CROP-seq_pilot
71	enh588_g2	CTCTCAGTCATCCTTGACCTT	CCND1	enh588	CROP-seq_pilot
72	enh588_g3	GCTCTCAGTCATCCCTGACCT	CCND1	enh588	CROP-seq_pilot
73	enh588_g4	TCCTCTAGCAGACGGCCCTG	CCND1	enh588	CROP-seq_pilot
74	enh588_g5	TCTGCTAGAGGATCACTCCT	CCND1	enh588	CROP-seq_pilot
75	enh588_g6	GGCGGAGTCATGCCAGCTCA	CCND1	enh588	CROP-seq_pilot
76	CCND1_g1	GCAGCAGAGTCCGCACGCTC	CCND1	CCND1	CROP-seq_pilot
77	CCND1_g2	GGTGAGTAGCAAAGAAACGT	CCND1	CCND1	CROP-seq_pilot
78	CCND1_g3	ACTCCGCCGCAGGGCAGGCG	CCND1	CCND1	CROP-seq_pilot
79	CCND1_g4	CTATGAAAACCGGACTACAG	CCND1	CCND1	CROP-seq_pilot
80	ESR1_TSS2_g1	AAGCCGGGCGACCCGAC	ESR1	ESR1_TSS2	CROP-seq_pilot
81	ESR1_TSS2_g2	GGCGCACGAGGATCTGCTAA	ESR1	ESR1_TSS2	CROP-seq_pilot
82	ESR1_TSS2_g3	GGAGCCCAGGAGCTGGCGGA	ESR1	ESR1_TSS2	CROP-seq_pilot
83	ESR1_TSS2_g4	TCAGGGCAAGGCAACAGTCCC	ESR1	ESR1_TSS2	CROP-seq_pilot
84	ESR1_TSS2_g5	GGAGACCAGTACTTAAAGT	ESR1	ESR1_TSS2	CROP-seq_pilot
85	TP53_g1	ATGAGTCCTCTCTGAGTCAC	TP53	TP53	CROP-seq_pilot
86	TP53_g2	TCAGGAGCTTACCCAATCCA	TP53	TP53	CROP-seq_pilot
87	TP53_g3	CCGAGAGCCCGTGACTCAGAG	TP53	TP53	CROP-seq_pilot
88	TP53_g4	TGGGGACTTAGCGAGTTT	TP53	TP53	CROP-seq_pilot
89	TP53_g5	GGAAGCGTGTCACCGTCG	TP53	TP53	CROP-seq_pilot
90	PIK3CA_g1	TCTCCCAGCGTCGGCCCG	PIK3CA	PIK3CA	CROP-seq_pilot
91	PIK3CA_g2	AGCGTGAGTAGAGCGCGGAC	PIK3CA	PIK3CA	CROP-seq_pilot
92	PIK3CA_g3	GAGAGGGTGCGGCGATCGC	PIK3CA	PIK3CA	CROP-seq_pilot
93	PIK3CA_g4	CCCCGAGCGTGAGTAGAGCG	PIK3CA	PIK3CA	CROP-seq_pilot
94	PIK3CA_g5	GGAGTCTCCGGCACCCACC	PIK3CA	PIK3CA	CROP-seq_pilot
95	SUZ12_g1	GGGCCGCCCGGCGGGTAGCTGG	SUZ12	SUZ12	CROP-seq_pilot
96	SUZ12_g2	CTCCGGCGGACCGAGGGGGGA	SUZ12	SUZ12	CROP-seq_pilot
97	SUZ12_g3	CGGAGCGAGGCCAGGGTA	SUZ12	SUZ12	CROP-seq_pilot
98	SUZ12_g4	CAGGCTCCGGCGGACCGAGG	SUZ12	SUZ12	CROP-seq_pilot
99	SUZ12_g5	TATTGCAGGCGCTTGCTCTC	SUZ12	SUZ12	CROP-seq_pilot

Supplementary Table 1 | CROP-seq libraries.

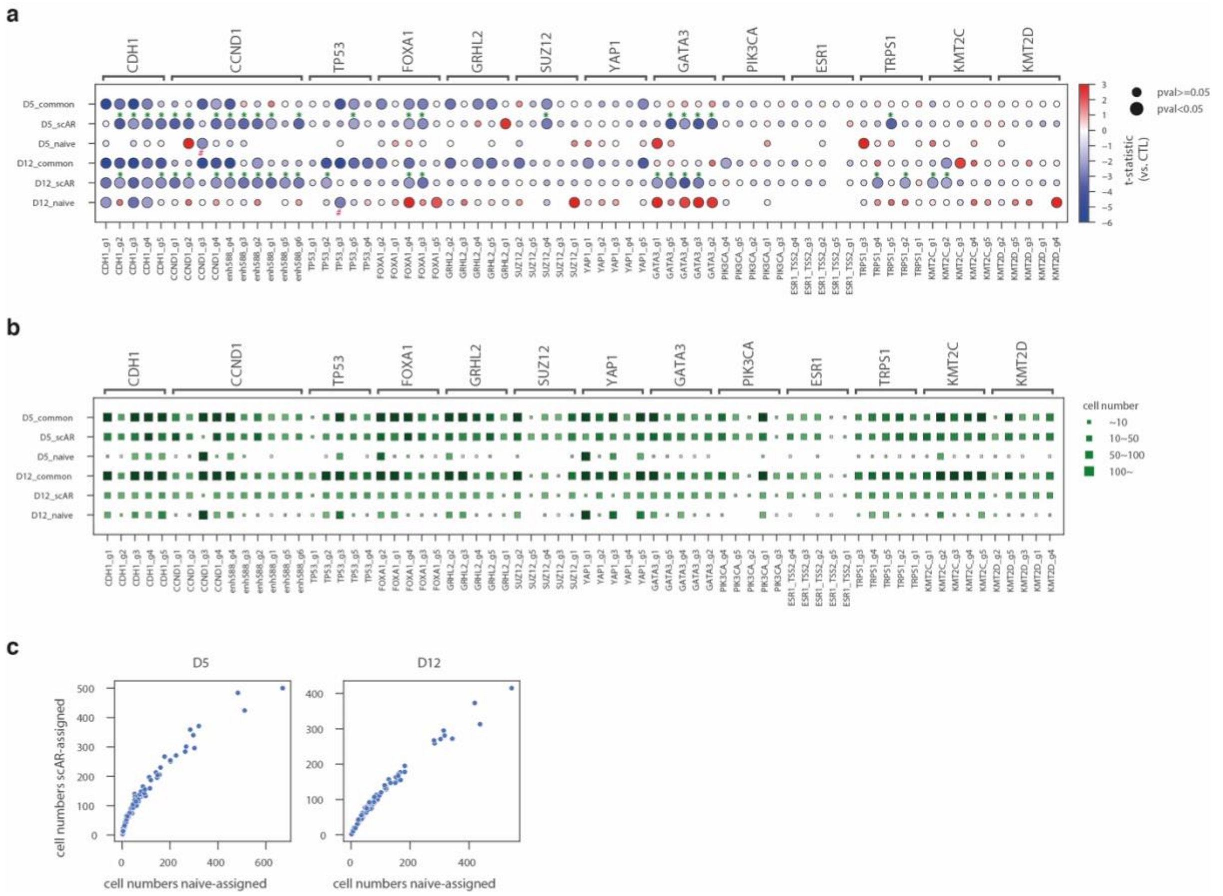
Supplementary Figures



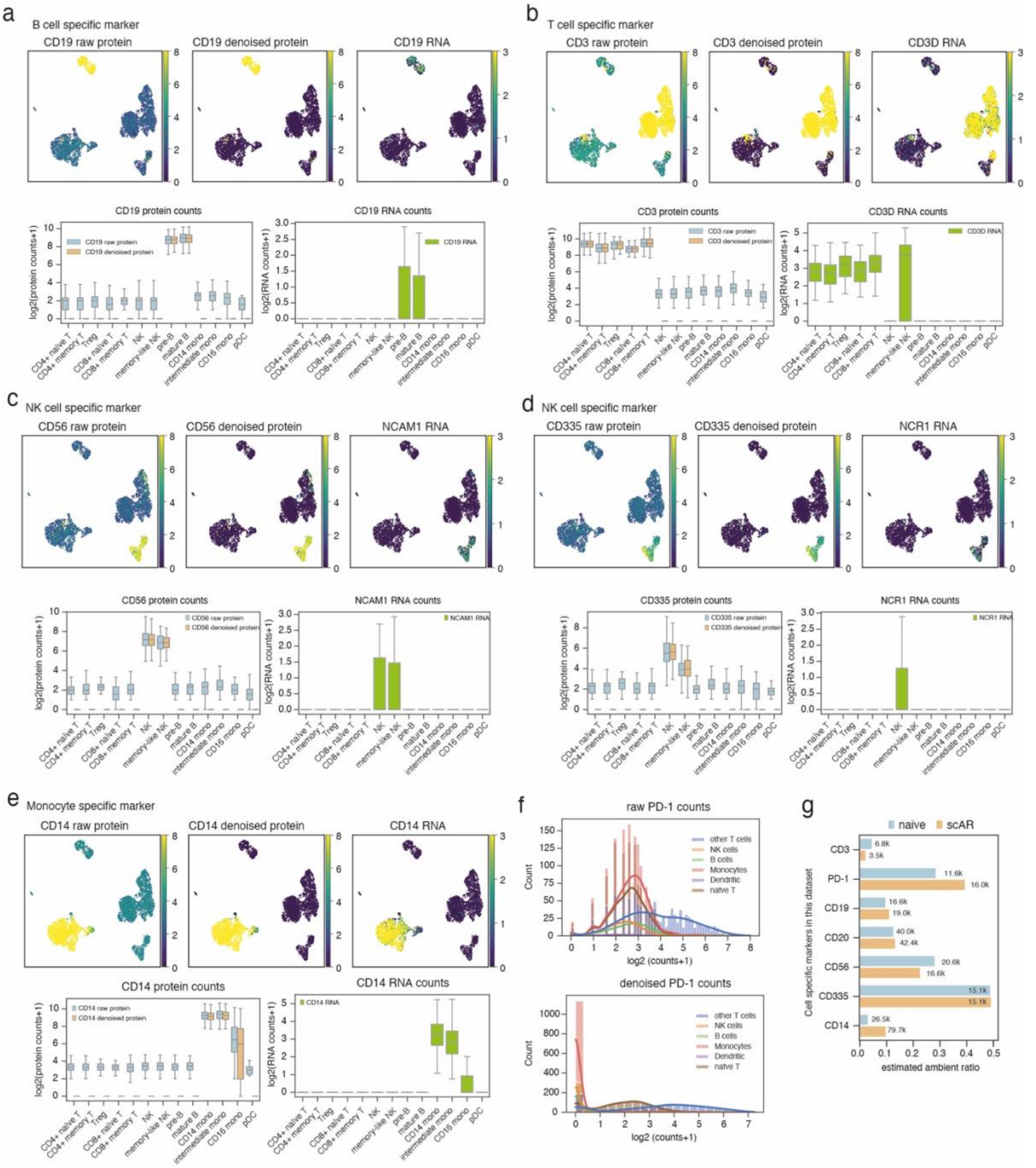
Supplementary Fig. 1 | Validation of CROP-seq libraries. a, The gene dependency scores of most target genes show negative effects on cell growth in MCF7. Data are originally from two pooled screening projects (DRIVE and Achilles) and obtained from https://depmap.org/. **b,** Three selected examples to illustrate an exponential model to evaluate the dropout rate of sgRNAs in cell pool. Dots represent biology replicates. Lines represent the exponential dropout model, as indicated by the equation. **c,** The fitness scores (the parameter alpha in **b**) of each guide. Green (positive alpha) indicates cancer cell promotion and red indicates cancer cell inhibition. Sizes of squares indicate p values.



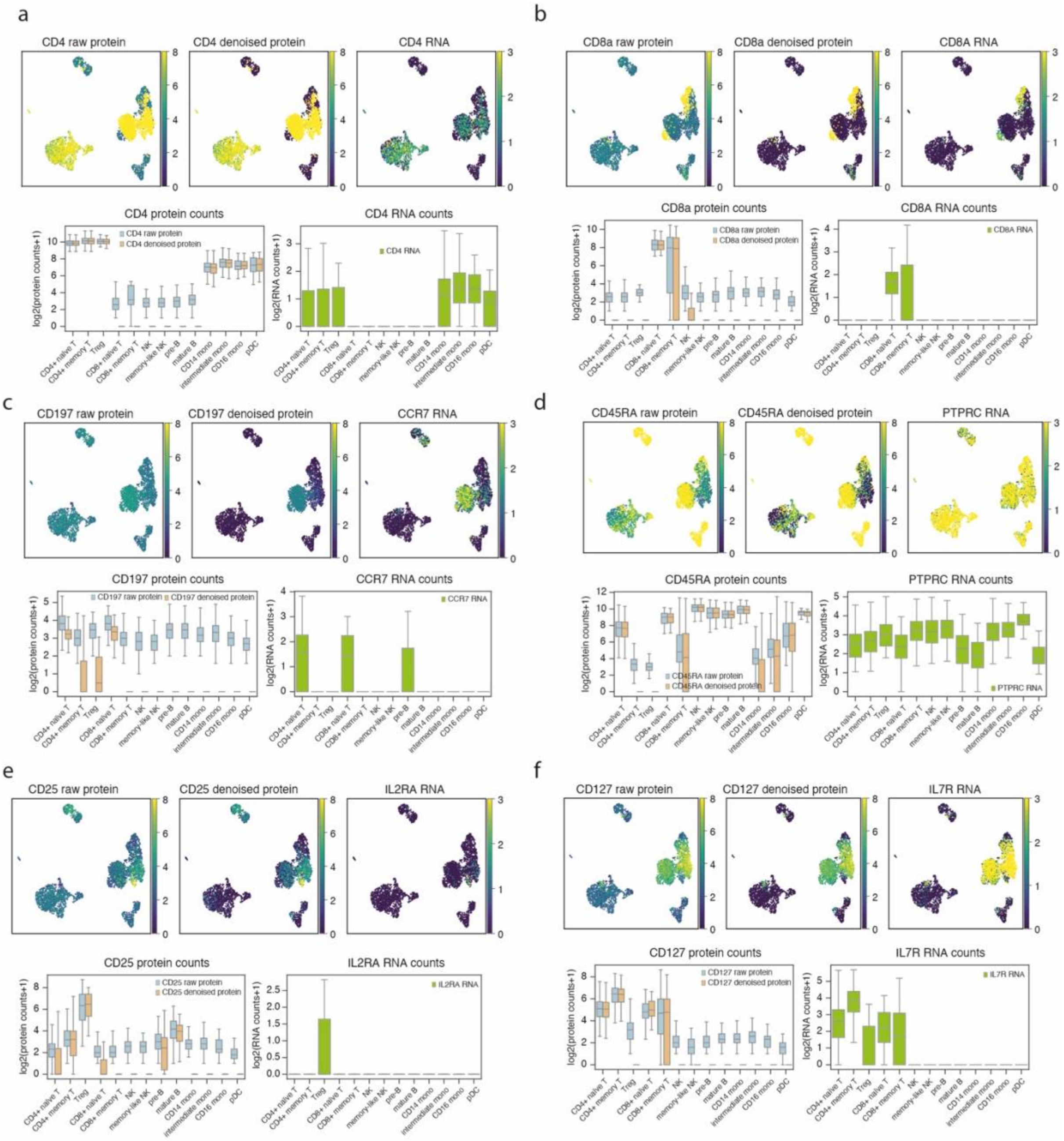
Supplementary Fig. 2 | Ambient noise widely exist in single-cell omics technologies. Ambient contamination is observed in several public datasets and the background noise is highly correlated with endogenous signal. **a,** Lopes2021 dataset. **b,** The A549_5k dataset from 10x genomics. **c,** The Ryan2019 dataset.



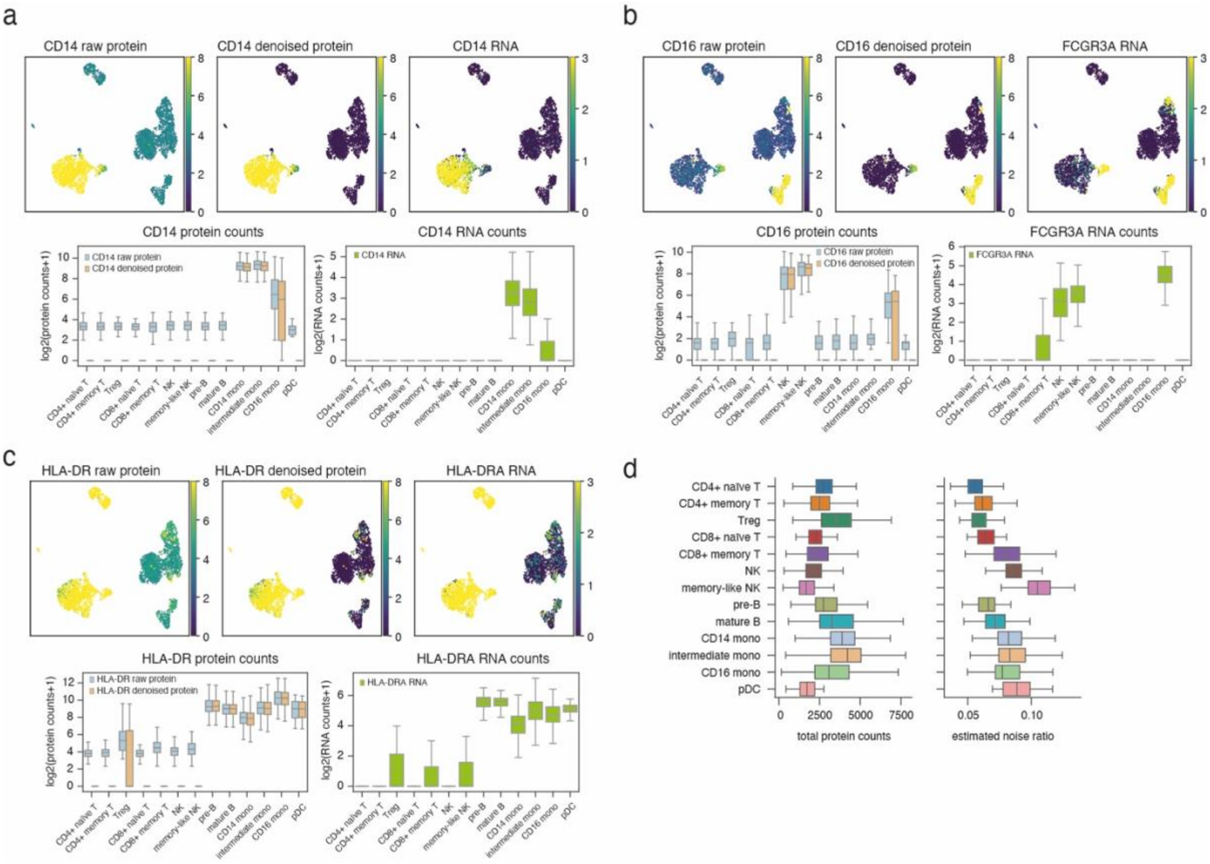
Supplementary Fig. 3 | Supplementary to Fig. 2. Evaluation of scAR on CROP-seq. **a,** Similar to **Fig. 2i**, the dotplot shows the overall comparison between two assignment approaches on t-statics of z-normalized expression. X-axis represents guide groups. Y-axis represents subgroups of cells as exemplified in **Fig. 2g** and **Fig. 2h**, separated by two time points and assignment approaches. Target genes are shown on the top. Their expression (log transformed) in each group is compared with that in CTL group and resulting t-statistics are shown by the dot color. Blue color indicates down-regulation, and red indicates up-regulation. CTL group is centered at zero. The bimodal sizes of circles represent the p-values from t-test (the bigger means p<0.05, the smaller means p>=0.05). * highlights the guide groups where scAR significantly improves the accuracy and # marks the groups where scAR underperforms naïve assignment. **b,** The comparison between two assignment approaches on cell number after assignment. Sizes of squares represent cell numbers of each assignment. **c,** The overall comparisons of cell number. Each dot represents an sgRNA.



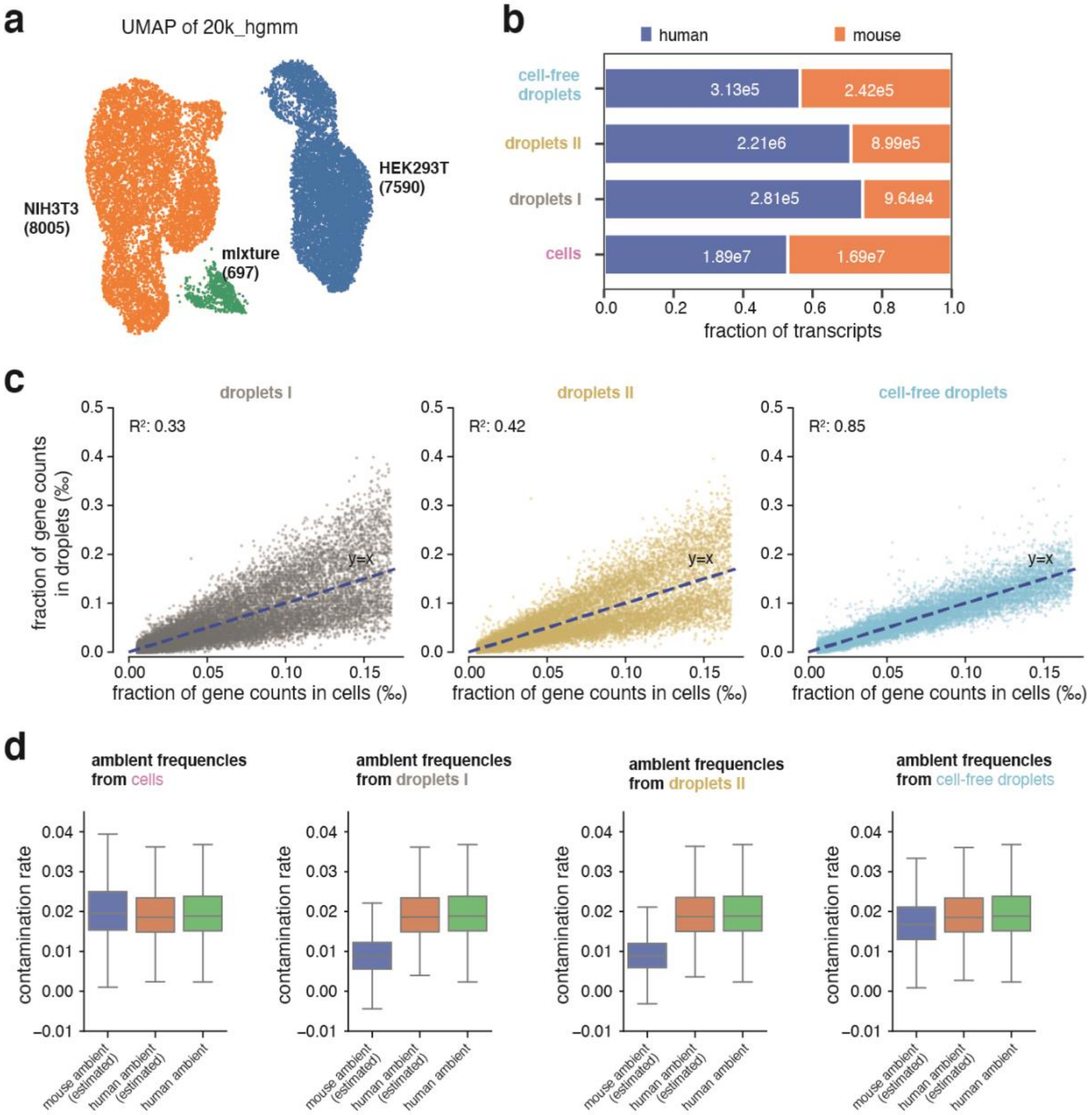
Supplementary Fig. 4 | scAR reduces the non-specific ambient antibodies.



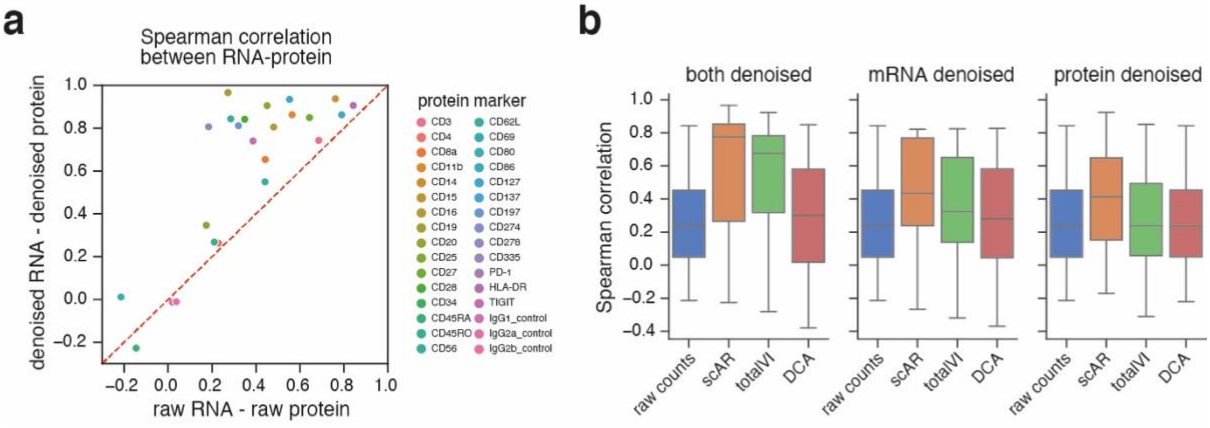
Supplementary Fig. 5 | scAR identifies markers for subtypes of T cells.



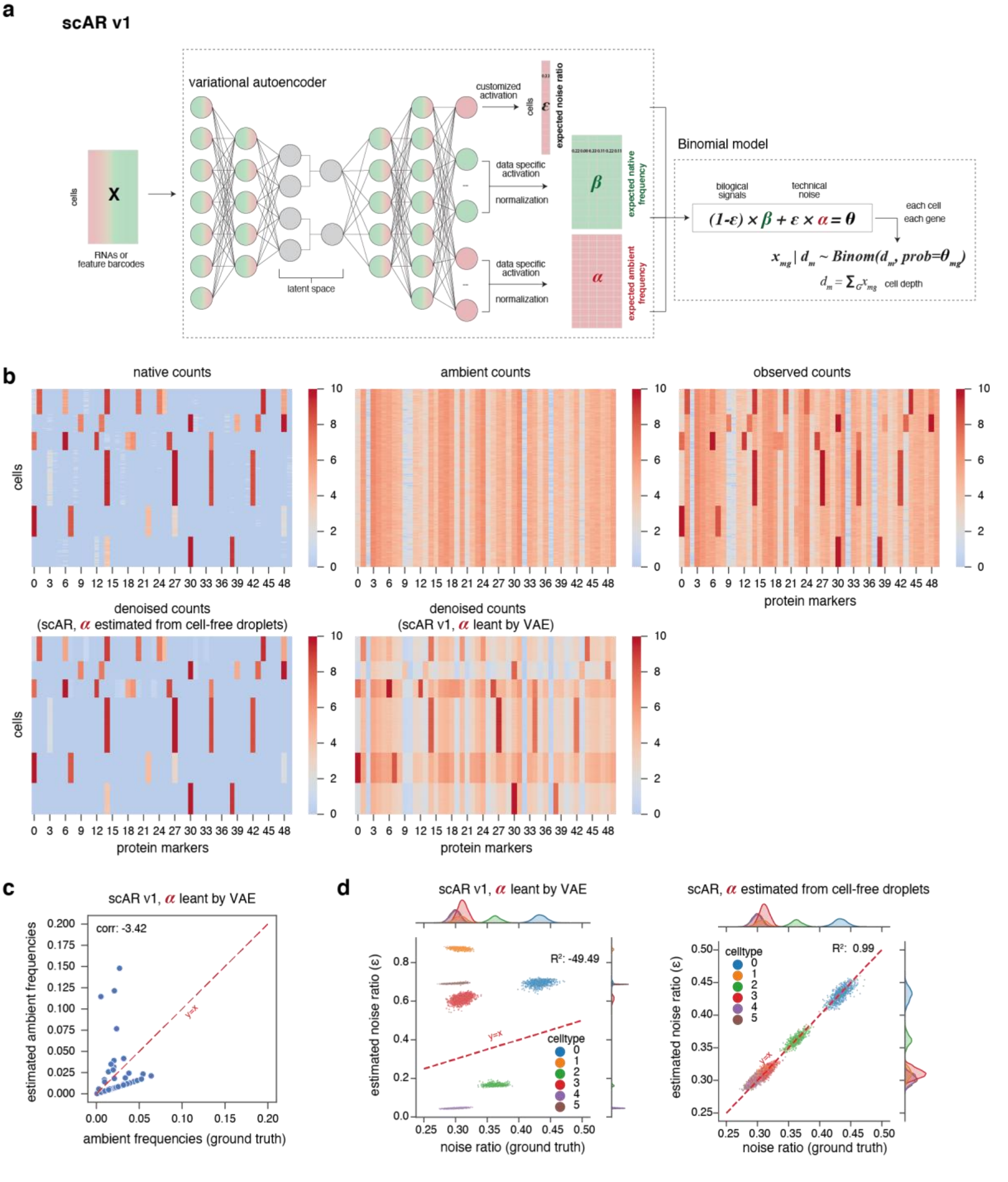
Supplementary Fig. 6 | scAR identifies marker for subtypes of monocytes.



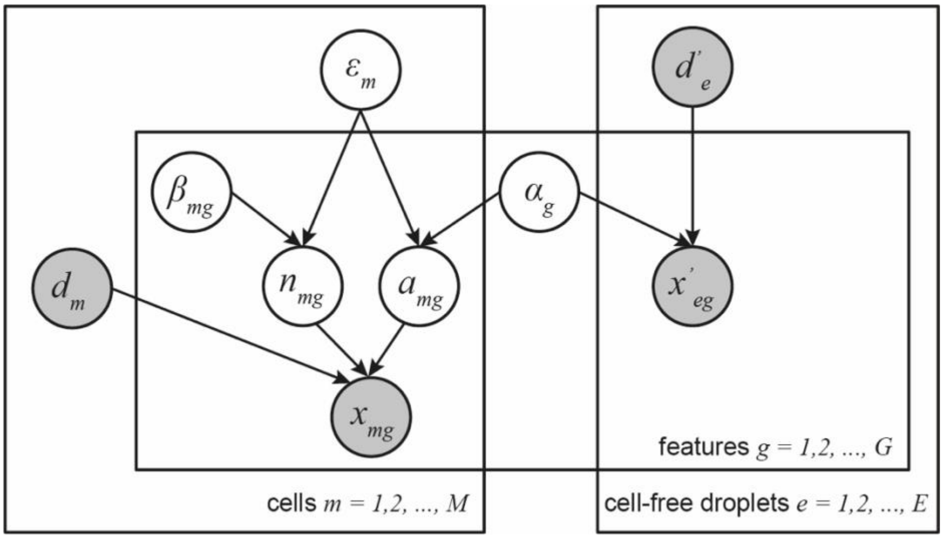
Supplementary Fig. 7 | Supplementary to Fig. 4. A public scRNAseq dataset of mixed human HEK293T and mouse NIH3T3 cells (1:1) was selected to demonstrate scAR's ability in noise reduction in transcriptome data. **a,** UMAP shows three populations of cell-containing droplets, HEK293T, NIH3T3 and multiplets. **b,** Fraction of transcripts in subpopulations of droplets. **c,** Correlation of gene frequencies between subpopulations of droplets and cell-containing droplets. **d,** Boxplots show the percentage of ambient signal in NIH3T3 cells. The green boxes represent the proportion of observed human transcripts. The blue and orange boxes represent scAR-estimated mouse and human ambient proportions, respectively.



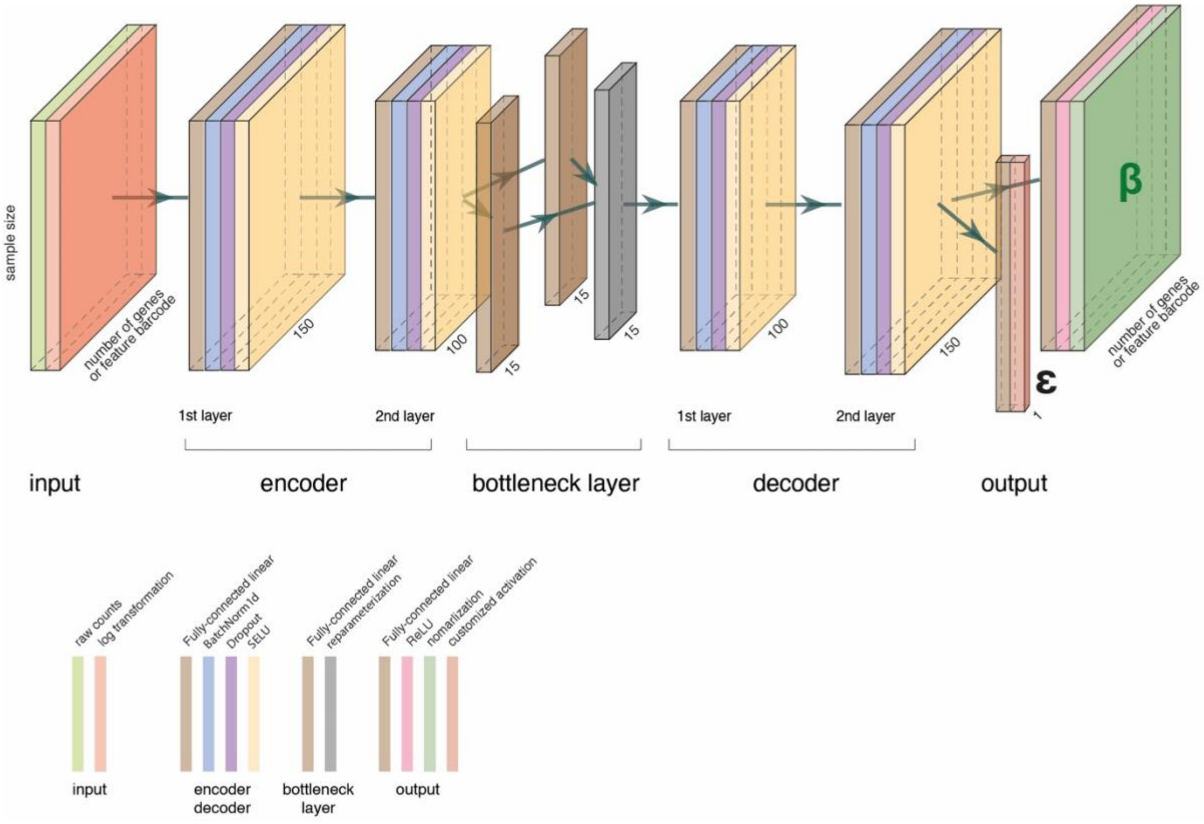
Supplementary Fig. 8 | Supplementary to Fig. 5 with full list of markers. Benchmarking of protein and mRNA count denoising. **a,** The scatterplot shows the Spearman correlation coefficients between RNA-protein pairs before (x-axis) and after scAR-denoising of both protein and mRNA counts (y-axis). The red dashed line represents y=x. Dots represent antibodies. **b**, Boxplots show Spearman correlation coefficients between RNA-protein pairs (both denoised, only mRNA denoised, and only protein denoised). Denoising methods are indicated on X-axis.



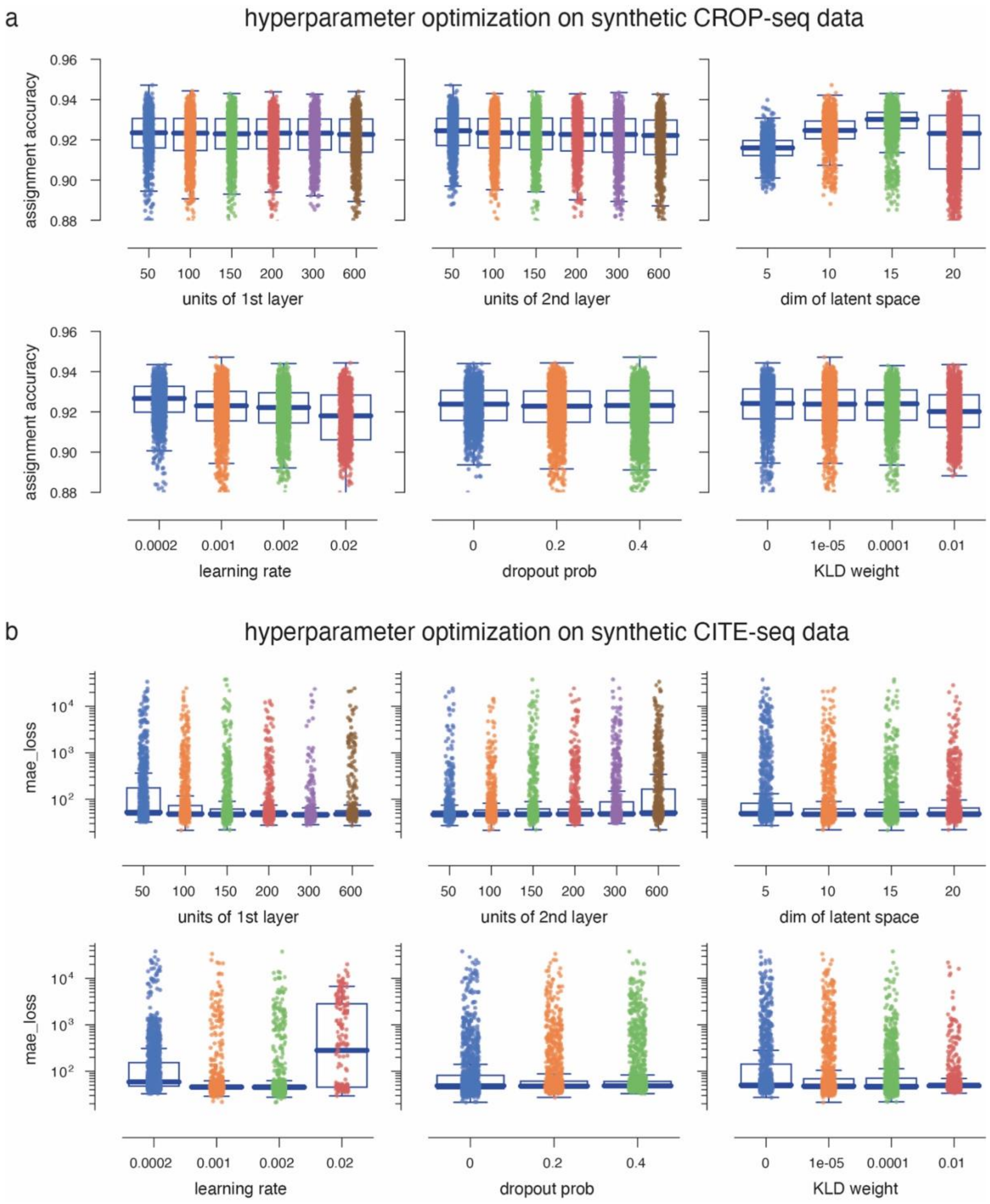
Supplementary Fig.9 | The performance of two versions of scAR. We compared two versions of scAR to demonstrate the necessary of using cell-free droplets. **a,** scAR v1 is an early version of scAR in which we fully relied on VAE to learn the ambient frequencies. **b,** Heatmaps of synthetic CITE-seq data (supplementary Note I). Native counts represent the ground truth, which are the signals we aimed to recover from observed counts. scAR refers to the version in **Fig. 1**, which we use cell-free droplets to estimate the ambient frequencies. scAR v1 refers to the version in (**a**). **c,** scAR v1 fails to learn the real ambient frequencies. Each dot represents a protein marker. **d,** The noise ratios estimated by two versions of scAR. Each dot represents a cell, colors represent cell type.



Supplementary Fig. 10 | Graphic model of scAR. The plates (three rectangles) represent independent replication, here, meaning individual cells, features and cell-free droplets, respectively. Grey circles represent observed random variables, e.g., d_m represents total counts in cell m and x_{mg} represents the observed count of feature g in cell m. Open circles represent latent random variables. Edges denote conditional dependencies among the variables.



Supplementary Fig.11 | The architecture of VAE in scAR. The optimized dimension numbers (supplementary Note II) of neural network layers are indicated and used as default parameters in scAR. They can also be modified by assigning optional arguments in the scAR command line tool (supplementary Note III).



Supplementary Fig.12 | Hyperparameter optimization of scAR. Two synthetic datasets (supplementary Note I), simulating CROP-seq data type and CITE-seq data type, were used to optimize scAR to allow generalized performance. We performed grid search to identify the optimal parameter set. In total, there are 6912 sets of parameters in each dataset. **a,** Hyperparameter optimization on a synthetic CROP-seq dataset. In this class of single-cell omics technologies, assignment of identify barcodes is key information (classification problem), so we used assignment accuracy as a metric to compare performance among parameters. **b,** Hyperparameter optimization on a synthetic CITE-seq dataset. As with scRNAseq, levels of feature barcodes are important information (regression problem). So, we use Mean Absolute Error (MAE loss) as metric in this case.