nature biotechnology

ANALYSIS

https://doi.org/10.1038/s41587-020-0469-4

Check for updates

Benchmarking single-cell RNA-sequencing protocols for cell atlas projects

Elisabetta Mereu[1,26], Atefeh Lafzi[1,26], Catia Moutinho[1], Christoph Ziegenhain[2], Davis J. McCarthy[3,4,5], Adrián Álvarez-Varela[6], Eduard Batlle[6,7,8], Sagar[9], Dominic Grün[9], Julia K. Lau[10], Stéphane C. Boutet[10], Chad Sanada[11], Aik Ooi[11], Robert C. Jones[12], Kelly Kaihara[13], Chris Brampton[13], Yasha Talaga[13], Yohei Sasagawa[14], Kaori Tanaka[14], Tetsutaro Hayashi[14], Caroline Braeuning[15], Cornelius Fischer[15], Sascha Sauer[15], Timo Trefzer[16], Christian Conrad[16], Xian Adiconis[17,18], Lan T. Nguyen[17], Aviv Regev[17,19,20], Joshua Z. Levin[17,18], Swati Parekh[21], Aleksandar Janjic[22], Lucas E. Wange[22], Johannes W. Bagnoli[22], Wolfgang Enard[22], Marta Gut[1], Rickard Sandberg[2], Itoshi Nikaido[14,23], Ivo Gut[1,24], Oliver Stegle[3,4,25] and Holger Heyn[1,24] ✉

Single-cell RNA sequencing (scRNA-seq) is the leading technique for characterizing the transcriptomes of individual cells in a sample. The latest protocols are scalable to thousands of cells and are being used to compile cell atlases of tissues, organs and organisms. However, the protocols differ substantially with respect to their RNA capture efficiency, bias, scale and costs, and their relative advantages for different applications are unclear. In the present study, we generated benchmark datasets to systematically evaluate protocols in terms of their power to comprehensively describe cell types and states. We performed a multicenter study comparing 13 commonly used scRNA-seq and single-nucleus RNA-seq protocols applied to a heterogeneous reference sample resource. Comparative analysis revealed marked differences in protocol performance. The protocols differed in library complexity and their ability to detect cell-type markers, impacting their predictive value and suitability for integration into reference cell atlases. These results provide guidance both for individual researchers and for consortium projects such as the Human Cell Atlas.

Single-cell genomics provides an unprecedented view of the cellular makeup of complex and dynamic systems. Single-cell transcriptomic approaches in particular have led the technological advances that allow unbiased charting of cell phenotypes[1]. The latest improvements in scRNA-seq allow these technologies to scale to thousands of cells per experiment, providing comprehensive profiling of tissue composition[2,3]. This has led to the identification of new cell types[4–6] and the fine-grained description of cell plasticity in dynamic systems, such as development[7,8]. Recent large-scale efforts, such as the Human Cell Atlas (HCA) project[9], are attempting to produce cellular maps of entire cell lineages, organs and organisms[10,11] by conducting phenotyping at the single-cell level. The HCA project aims to advance our understanding of tissue function and to serve as a reference for defining variation in human health and disease. In addition to methods that capture the spatial organization of tissues[12,13], the main approach being used is scRNA-seq analysis of dissociated cells. Therefore, tissues are disaggregated and individual cells captured either by cell sorting or using microfluidic systems[1]. In sequential processing steps, cells are lysed, the RNA is reverse transcribed to complementary DNA, amplified and processed to sequencing-ready libraries.

Continuous technological development has improved the scale, accuracy and sensitivity of scRNA-seq methods, and now allows us to create tailored experimental designs by selecting from a plethora of different scRNA-seq protocols. However, there are marked differences across these methods, and it is not clear which protocols are best for different applications. For large-scale consortium projects, experience has shown that neglecting benchmarking, standardization

[1]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. [3]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK. [4]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [5]St Vincent's Institute of Medical Research, Fitzroy, Victoria, Australia. [6]Institute for Research in Biomedicine, Barcelona Institute of Science and Technology, Barcelona, Spain. [7]Catalan Institution for Research and Advanced Studies, Barcelona, Spain. [8]Centro de Investigación Biomédica en Red de Cáncer, Barcelona, Spain. [9]Max-Planck-Institute of Immunobiology and Epigenetics, Freiburg, Germany. [10]10x Genomics, Pleasanton, CA, USA. [11]Fluidigm Corporation, South San Francisco, CA, USA. [12]Department of Bioengineering, Stanford University, Stanford, CA, USA. [13]Bio-Rad, Hercules, CA, USA. [14]Laboratory for Bioinformatics Research, RIKEN Center for Biosystems, Dynamics Research, Saitama, Japan. [15]Max Delbrück Center for Molecular Medicine/Berlin Institute of Health, Berlin, Germany. [16]Digital Health Center, Berlin Institute of Health, Charité-Universitätsmedizin Berlin, Berlin, Germany. [17]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [18]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [19]Koch Institute of Integrative Cancer Research, MIT, Cambridge, MA, USA. [20]Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, MA, USA. [21]Max-Planck-Institute for Biology of Ageing, Cologne, Germany. [22]Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany. [23]School of Integrative and Global Majors, University of Tsukuba, Wako, Saitama, Japan. [24]Universitat Pompeu Fabra, Barcelona, Spain. [25]Division of Computational Genomics and Systems Genetics, German Cancer Research Center, Heidelberg, Germany. [26]These authors contributed equally: Elisabetta Mereu, Atefeh Lafzi. ✉e-mail: holger.heyn@cnag.crg.eu

and quality control at the start can lead to major problems later on in the analysis of the results[14]. Thus, success depends critically on implementing a high common standard. A comprehensive comparison of available scRNA-seq protocols will benefit both large- and small-scale applications of scRNA-seq.

The available scRNA-seq protocols vary in the efficiency of RNA-molecule capture, which results in differences in sequencing library complexity and the sensitivity of the method to identify transcripts and genes[15–17]. There has been no systematic testing of how their performance varies between cell types, and how this affects the resolution of cell phenotyping in complex samples. In the present study, we extend previous efforts to compare the molecule-capture efficiency of scRNA-seq protocols[15,16] by systematically evaluating the capability of these techniques to describe tissue complexity and their suitability for creating a cell atlas. We performed a multicenter benchmarking study to compare scRNA-seq protocols using a unified reference sample resource. Our reference sample contained: (1) a high degree of cell-type heterogeneity with various frequencies, (2) closely related subpopulations with subtle differences in gene expression, (3) a defined cell composition with trackable markers and (4) cells from different species. By analyzing human peripheral blood and mouse colon tissue, we have covered a broad range of cell types and states from cells in suspension and solid tissues, to represent common scenarios in cell atlas projects. We have also added spike-in cell lines to allow us to assess batch effects, and have combined different species to pool samples into a single reference. We performed a comprehensive comparative analysis of 13 different scRNA-seq protocols, representing the most commonly used methods. We applied a wide range of different quality control metrics to evaluate datasets from different perspectives, and to test their suitability for producing a reproducible, integrative and predictive reference cell atlas.

We observed striking differences among protocols in converting RNA molecules into sequencing libraries. Varying library complexities affected the protocol's power to quantify gene expression levels and to identify cell-type markers, a trend consistently observed across cell and tissue types. This critically impacted on the resolution of tissue profiles and the predictive value of the datasets. Protocols further differed in their capacity to be integrated into reference tissue atlases and, thus, their suitability for consortium-driven projects with flexible production designs.

Results

Reference sample and experimental design. We benchmarked current scRNA-seq protocols to inform the methodological selection process of cell atlas projects. Ideally, methods should: (1) be accurate and free of technical biases, (2) be applicable across distinct cell properties, (3) fully disclose tissue heterogeneity, including subtle differences in cell states, (4) produce reproducible expression profiles, (5) comprehensively detect population markers, (6) be integratable with other methods and (7) have predictive value with cells mapping confidently to a reference atlas.

For a systematic comparison of protocols, we designed a reference sample containing human peripheral blood mononuclear cells (PBMCs) and mouse colon, which are tissue types with highly heterogeneous cell populations, as determined by previous single-cell sequencing studies[18,19]. In addition to the well-defined cell types, the tissues contain cells in transition states (for example, colon transit-amplifying (TA) or enterocyte progenitor cells) that show transcriptional differences during their differentiation trajectory[20]. The reference sample also included a wide range of cell sizes (for example, B cells: ~7 µm; HEK293 cells: ~15 µm) and RNA content, which are key parameters that affect performance in cell capture and library preparation. Interrogation of tissues from different species allowed us to pool a large variety of cell types in a single reference sample to maximize complexity while minimizing variability introduced during sample preparation. In addition to the intra-tissue complexity, the fluorescence-labeled, spiked-in cell lines allowed us to monitor cell-type composition during sample processing, and to identify batch effects and biases introduced during cell capture and library preparation.

Specifically, the reference sample contained (estimated percentage viable cells): PBMCs (60%, human), colon cells (30%, mouse), HEK293T cells (6%, red fluorescent protein (RFP)-labeled human cell line), NIH3T3 cells (3%, green fluorescent protein (GFP)-labeled mouse cells) and MDCK cells (1%, TurboFP650-labeled dog cells) (Fig. 1). To reduce variability due to technical effects during library preparation, the reference sample was prepared in a single batch, distributed into aliquots of 250,000 cells and cryopreserved. We have previously shown that cryopreservation is suitable for single-cell transcriptomic studies of these tissue types[21]. For cell capture and library preparation, the thawed samples underwent FACS to remove damaged cells and physical doublets (see the next section for detailed analysis of cell viability sorting).

A reference dataset for benchmarking experimental and computational protocols. To obtain sufficient sensitivity to capture low-frequency cell types and subtle differences in the cell state, we profiled ~3,000 cells with each scRNA-seq protocol. In total, we produced datasets for five microtiter plate-based methods and seven microfluidic systems, including cell-capture technologies based on droplets (four), nanowells (one) and integrated fluidic circuits, to capture small (one) and medium (one)-sized cells (Fig. 1 and see Supplementary Table 1). We also included experiments to produce single-nucleus RNA-sequencing (snRNA-seq) libraries (one), and an experimental variant that profiled >50,000 cells to produce a reference of our complex sample. The unified sample resource and standardized sample preparation (see Methods) were designed largely to eliminate sampling effects and allow the systematic comparison of scRNA-seq protocol performance.

To compare the different protocols, and to create a resource for the benchmarking and development of computational tools (for example, batch effect correction, data integration and annotation), all datasets were processed in a uniform manner. Therefore, we designed a streamlined, primary data-processing pipeline tailored to the peculiarities of the reference sample (see Methods). Briefly, raw sequencing reads were mapped to a joint human, mouse and canine reference genome, and separately to their respective references to produce gene count matrices for subsequent analysis (accession no. GSE133549). Overall, we detected human, mouse and canine cell numbers consistent with the composition design of the reference sample (Fig. 1). However, some protocols varied markedly from the expected frequencies in human (34–95%), mouse (4–66%) and canine (0–9%) cells. Although the reference sample was prepared in a standardized way, we cannot entirely exclude the introduction of composition variability during sample handling. Thus, the subsequent evaluation of protocol performance was performed on cell types and states common to all protocols.

Notably, we observed a higher fraction of mouse colon cells in unsorted (Chromium) and the snRNA-seq datasets (Chromium (sn)). This probably results from damaging the more fragile colon cells during sample preparation, resulting in proportionally fewer colon cells when selecting for cell viability. To test whether this composition bias in scRNA-seq can be avoided by skipping viability selection, we generated matched datasets either selecting or not selecting for intact cells. After quality control the detection of mouse colon cells increased proportionally without viability selection (51% versus 19%), with good-quality cells showing comparable library complexity in both libraries (for example, numbers of detected genes; see Supplementary Figs. 1 and 2). However, considerably more cells were removed during quality filtering (44% versus 15%), and this is a source of unwanted sequencing costs that

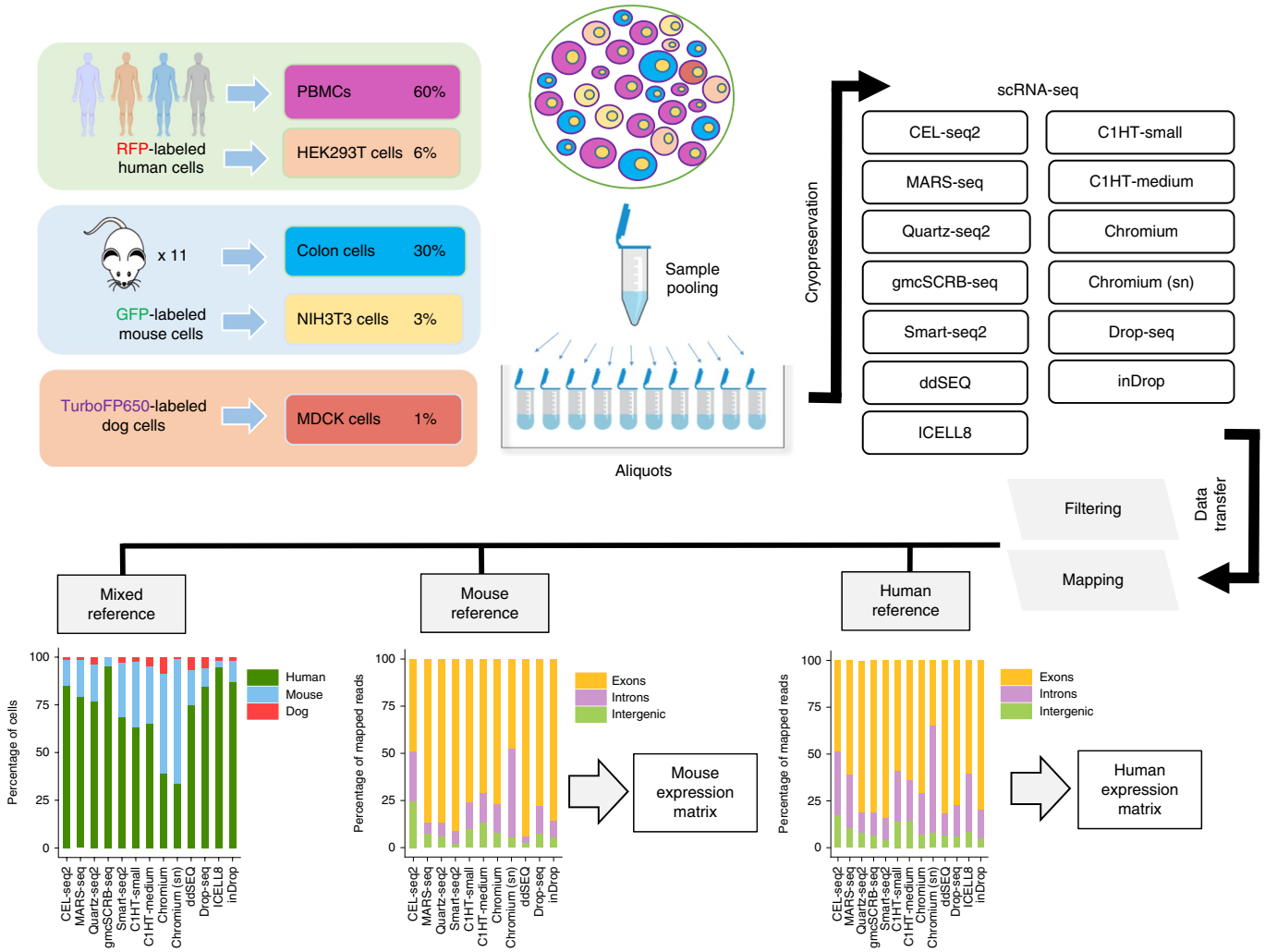

Fig. 1 | Overview of the experimental design and data processing. The reference sample consists of human PBMCs (60%), and HEK293T (6%), mouse colon (30%), NIH3T3 (3%) and dog MDCK cells (1%). The sample was prepared in one single batch, cryopreserved and sequenced by 13 different sc/snRNA-seq methods. Sequences were uniformly mapped to a joint human, mouse and canine reference, and then separately to produce gene expression counts for each sequencing method.

must be taken into account, especially for tissues with high cell damage. Consequently, replacing viability staining with thorough in silico quality filtering in cell atlas experiments might better conserve the composition of the original tissue, but result in higher sequencing costs.

The canine cells, spiked-in at a low concentration, were detected by all protocols (1–9%) except gmcSCRB-seq. Furthermore, the different methods showed notable differences in mapping statistics between different genomic locations (Fig. 1). As expected, due to the presence of unprocessed RNA in the nucleus, the snRNA-seq experiment detected the highest proportion of introns, although scRNA-seq protocols also showed high frequencies of intronic and intergenic mappings. The increased detection of unprocessed transcripts in CEL-seq2 may be due to a freezing step (−80 °C) after cell isolation and subsequent denaturation at high temperatures (95 °C), which could favor the accessibility of nuclear and chromatin-bound RNA molecules.

Molecule-capture efficiency and library complexity. We produced reference datasets by analyzing 30,807 human and 19,749 mouse cells (Chromium v.2; Fig. 2a–c). The higher cell number allowed us to annotate the major cell types in our reference sample, and to extract population-specific markers (see Supplementary Table 2). It was noteworthy that the reference samples solely provided the basis to assign cell identities and gene marker sets, and were not used to quantify the method's performance. This strategy ensured that the choice of technology for deriving the reference does not influence downstream analyses. Cell clustering and reference-based cell annotation showed high agreement (average 83%; see Supplementary Table 3), and only cells with consistent annotations were used subsequently for comparative analysis at the cell-type level. The PBMCs (human) and colon cells (mouse) represented two largely different scenarios. Although the differentiated PBMCs clearly separated into subpopulations (for example, T/B cells, monocytes; Fig. 2b, and see Supplementary Figs. 3a and 4a–d), colon cells were ordered as a continuum of cell states that differentiate from intestinal stem cells into the main functional units of the colon (that is, absorptive enterocytes and secretory cells; Fig. 2c, and see Supplementary Figs. 3b and 5a–d). Notably, the subpopulation structure of our references was largely consistent with that of published datasets for human PBMCs[18] and mouse colon cells[22] (see Supplementary Figs. 6 and 7). After identifying major subpopulations and their respective markers in our reference sample, we clustered the cells of each sc/snRNA-seq protocol and annotated cell types using matchSCore2 (see Methods). This algorithm allows a gene marker-based projection of single cells (cell by cell) on to a

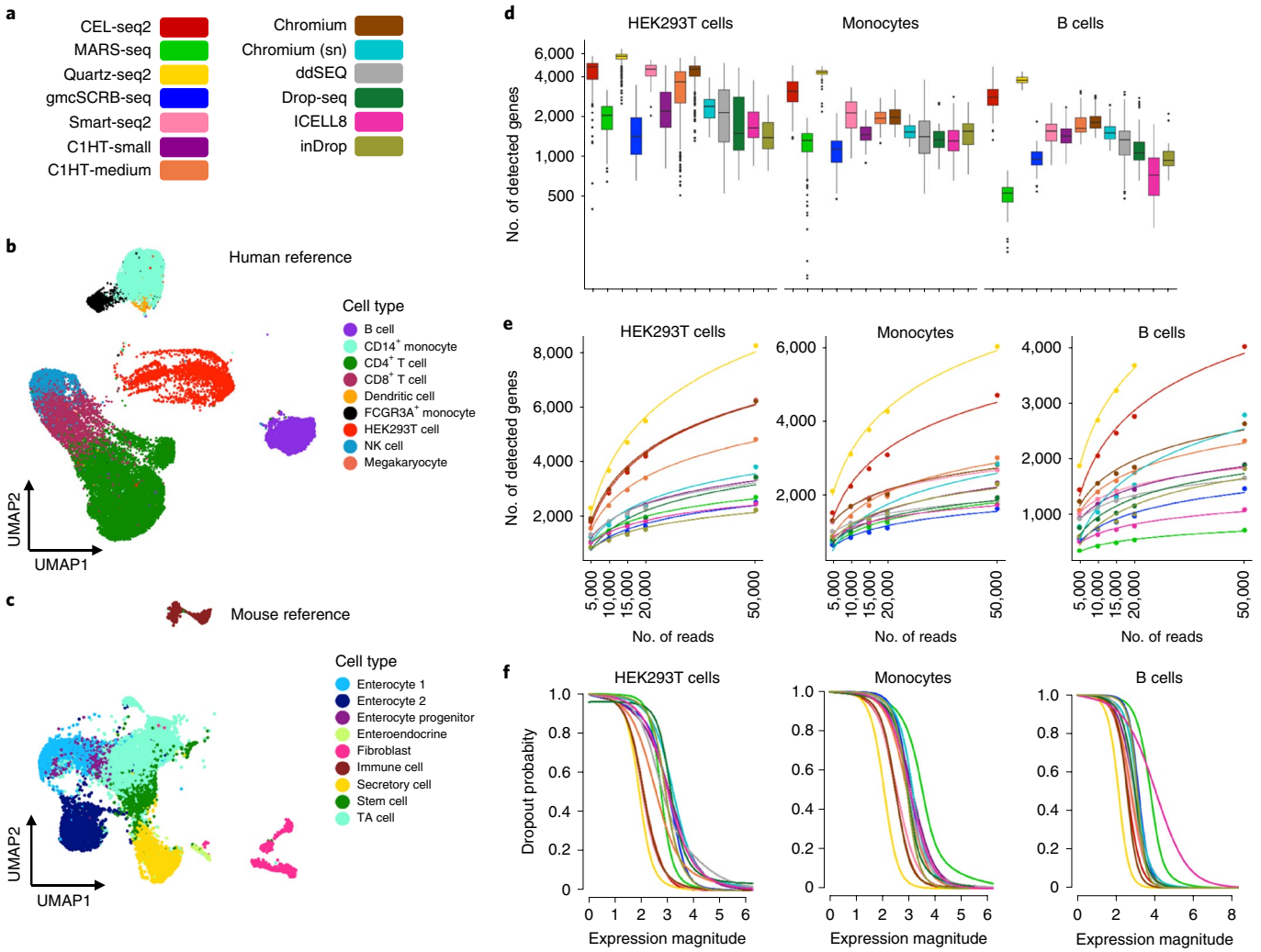

Fig. 2 | Comparison of 13 sc/snRNA-seq methods. a, Color legend of sc/snRNA-seq protocols. **b**, UMAP of 30,807 cells from the human reference sample (Chromium) colored by cell-type annotation. **c**, UMAP of 19,749 cells from the mouse reference (Chromium) colored by cell-type annotation. **d**, Boxplots displaying the minimum, the first, second and third quantiles, and the maximum number of genes detected across the protocols, in down-sampled (20,000) HEK293T cells, monocytes and B cells. Cell identities were defined by combining the clustering of each dataset and cell projection on to the reference. **e**, Number of detected genes at stepwise. down-sampled, sequencing depths. Points represent the average number of detected genes as a fraction of all cells of the corresponding cell type at the corresponding sequencing depth. **f**, Dropout probabilities as a function of expression magnitude, for each protocol and cell type, calculated on down-sampled data (20,000) for 50 randomly selected cells.

reference sample and, thus, the identification of cell types in our datasets (see Supplementary Figs. 8 and 9).

To compare the efficiency of messenger RNA capture between protocols, we down-sampled the sequencing reads per cell to a common depth and stepwise-reduced fractions. Stochasticity introduced during down-sampling did not affect the reproducibility of the results (see Supplementary Fig. 10). Library complexity was determined separately for largely homogeneous cell types with markedly different cell properties and function, namely human HEK293T cells, monocytes and B cells (Fig. 2d,e), and mouse colon secretory and TA cells (see Supplementary Fig. 11a,b). We observed large differences in the number of detected genes and molecules across the protocols, with consistent trends across cell types and gene quantification strategies (see Supplementary Fig. 11c,d). Notably, some protocols, such as Smart-seq2 and Chromium v.2, performed better with higher RNA quantities (HEK293T cells) compared with lower starting amounts (monocytes and B cells), suggesting an input-sensitive optimum. Considering the different assay versions and application types of the Chromium system, a dedicated analysis showed increased detection of molecules and genes from nuclei to intact cells and toward the latest protocol versions (see Supplementary Fig. 12). Consistent with the variable library complexity, the protocols presented large differences in dropout probabilities (Fig. 2f), with Quartz-seq2, Chromium v.2 and CEL-seq2 showing consistently lower probability. Note that, despite the considerable differences between protocols, we observed a generally high technical reproducibility within the methods (see Supplementary Fig. 13).

Technical effects and information content. We further assessed the magnitude of technical biases, and the protocol's ability to describe cell populations. To quantify the technical variation within and across protocols, we selected highly variable genes (HVGs) across all datasets, and plotted the variation in the main principal components (PCs; Fig. 3a). Using the down-sampled data for HEK293T cells, monocytes and B cells, we observed strong protocol-specific profiles, with the main source of variability being the number of genes detected per cell (Fig. 3b). Data from snRNA-seq did not show notable outliers, indicating conserved representation of the

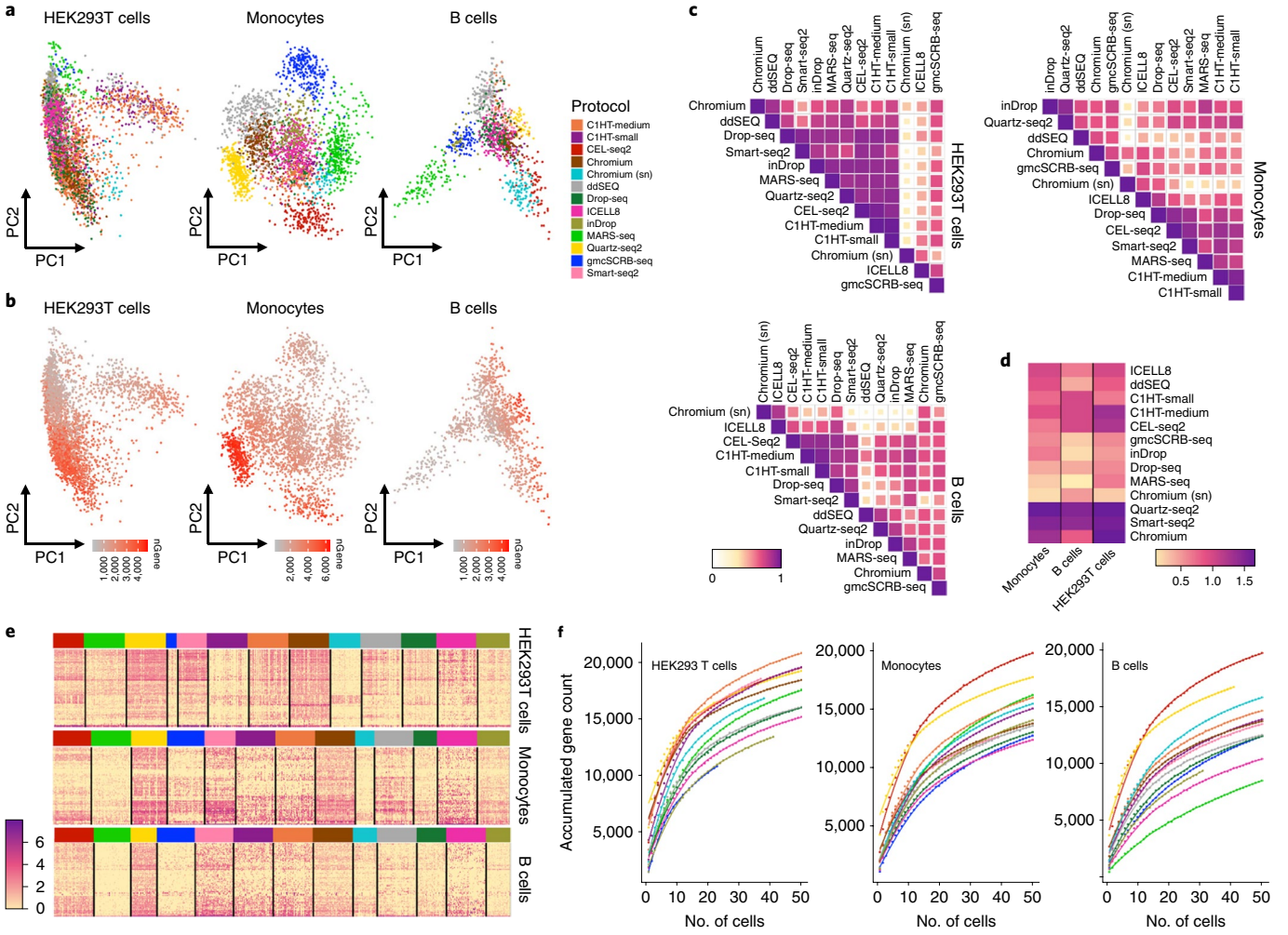

Fig. 3 | Similarity measures of sc/snRNA-seq methods. a,**b**, Principal component analysis on down-sampled data (20,000) using highly variable genes between protocols, separated into HEK293T cells, monocytes and B cells, and color coded by protocol (**a**) and number of detected genes per cell (**b**). **c**, Pearson's correlation plots across protocols using expression of common genes. For a fair comparison, cells were down-sampled to the same number for each method (B cells, $n=32$; monocytes, $n=57$; HEK293T cells, $n=55$). Protocols are ordered by agglomerative hierarchical clustering. **d**, Average log(expression) values of cell-type-specific reference markers for down-sampled (20,000) HEK293T cells, monocytes and B cells. **e**, Log(expression) values of reference markers on down-sampled data (20,000) for HEK293T cells, monocytes and B cells (maximum of 50 random cells per technique). **f**, Cumulative gene counts per protocol as the average of 100 randomly sampled HEK293T cells, monocytes and B cells, separately on down-sampled data (20,000).

transcriptome between the cytoplasm and the nucleus. To quantify the protocol-related variance, we identified the PCs that correlated with the protocol's covariates in a linear model[23]. Indeed, the variance in the data was mainly explained by the protocols (HEK293T cells = 37.3%, monocytes = 52.8% and B cells = 36.2%), a value that was reduced in HEK293T cells and monocytes when considering snRNA-seq as a specific covariate (HEK293T cells = 9.7%, monocytes = 22.2% and B cells = 48.3%; see Methods). The technical effects were also visible when using *t*-distributed stochastic neighbor embedding (tSNE) as a nonlinear, dimensionality reduction method (see Supplementary Fig. 14). By contrast, the methods largely mixed when the analysis was restricted to cell-type-specific marker genes, suggesting a conserved cell identity profile across techniques (see Supplementary Fig. 15).

Next, we quantified the similarities in information content of the protocols. Again, we used the down-sampled datasets and commonly expressed genes and calculated the correlation between methods in average transcript counts across multiple cells, thus compensating for the sparseness of single-cell transcriptome data. For the three human cell types, we observed a broad spectrum of correlation across technologies, with generally lower correlation for smaller cell types (Fig. 3c). Although the transcriptome representation was generally conserved (Fig. 3a), the snRNA-seq protocol resulted in a notable outlier when correlating the expression levels of common genes across protocols, possibly driven by decreased correlation of immature transcripts. Restricting the correlation analysis to population-specific marker genes, we observed less variation between protocols (Pearson's $r=0.5$–0.7), which underlines that the expression of these markers is largely conserved across the methods (see Supplementary Fig. 16).

To further test the suitability of protocols for describing cell types, we determined their sensitivity to detect population-specific expression signatures, and found that they had remarkably variable power to detect marker genes. Specifically, population markers were detected with different accuracies (see Supplementary Figs. 17 and 18), and the detection level varied substantially (Fig. 3d,e and see Supplementary Table 4). Quartz-seq2 and Smart-seq2 showed high expression levels for all cell-type signatures, indicating that they

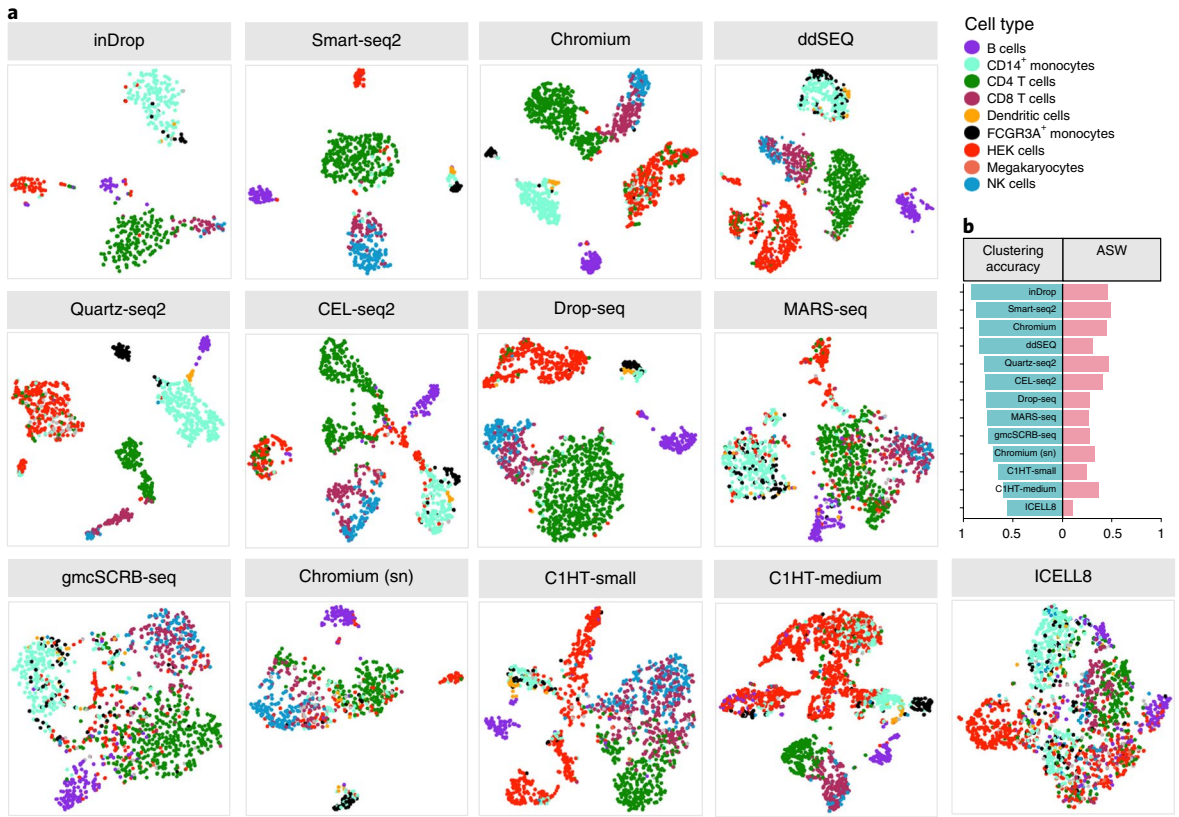

Fig. 4 | Clustering analysis of 13 sc/snRNA-seq methods on down-sampled datasets (20,000). a, The tSNE visualizations of unsupervised clustering in human samples from 13 different methods. Each dataset was analyzed separately after down-sampling to 20,000 reads per cell. Cells are colored by cell type inferred by matchSCore2 before down-sampling. Cells that did not achieve a probability score of 0.5 for any cell type were considered unclassified. **b**, Clustering accuracy and ASW for clusters in each protocol.

have higher power for cell-type identification. As marker genes are particularly important for data interpretation (for example, annotation), low marker detection levels could severely limit the interpretation of poorly explored tissues, or when trying to identify subtle differences across subpopulations. SnRNA-seq showed generally lower marker detection levels. However, gene markers were selected from intact cell experiments, which could lead to an underestimation of the performance of snRNA-seq to identify cell-type-specific signatures in this analysis approach.

The protocols also detected vastly different total numbers of genes when accumulating transcript information over multiple cells, with strong positive outliers observed for the smaller cell types (Fig. 3f). In particular, CEL-seq2 and Quartz-seq2 identified many more genes than other methods. Intriguingly, CEL-seq2 outperformed all other methods by detecting many weakly expressed genes; genes detected specifically by CEL-seq2 had significantly lower expression than the common genes detected by Quartz-seq2 ($P < 2.2 \times 10^{-16}$). The greater sensitivity to weakly expressed genes makes this protocol particularly suitable for describing cell populations in detail, an important prerequisite for creating a comprehensive cell atlas and functional interpretation.

Surprisingly, considering the increased library complexity of scRNA-seq compared with snRNA-seq, the latter protocol identified a similar number of genes when combining information across multiple cells and suggesting overall similar transcriptome complexity of the two compartments (see Supplementary Fig. 12). ScRNA-seq detected additional genes enriched in biological processes such as organelle function, including many mitochondrial genes that were largely absent in the snRNA-seq datasets (see Supplementary Table 5).

To further illustrate the power of the different protocols to chart the heterogeneity of complex samples, we clustered and plotted down-sampled datasets in two-dimensional space (Fig. 4a) and then calculated the cluster accuracy and average silhouette width (ASW[24], Fig. 4b), a commonly used measure for assessing the quality of data partitioning into communities. Consistent with the assumption that library complexity and sensitive marker detection provide greater power to describe complexity, methods that performed well for these two attributes showed better separation of subpopulations, and greater ASW and cluster accuracy. This is illustrated in the monocytes, for which accurate clustering protocols separated the major subpopulations (CD14$^+$ and FCGR3A$^+$), whereas methods with low ASW did not distinguish between them. Similarly, several methods were able to distinguish between CD8$^+$ and natural killer (NK) cells, whereas others were not.

Joint analysis across datasets. A common scenario for cell atlas projects is that data are produced at different sites using different scRNA-seq protocols. However, the final atlas is created from a combination of datasets, which requires that the technologies used be compatible. To assess how suitable it is to combine the results from our protocols into a joint analysis, we used down-sampled human and mouse datasets to produce a joint quantification matrix for all techniques[25]. Importantly, single cells grouped themselves by cell type, suggesting that cell phenotypes are the main driver of heterogeneity in the joint datasets (Fig. 5a–d, and see Supplementary Figs. 19a,b and 20). Indeed, the combined data showed a clear separation of cell states (for example, T cell and enterocyte subpopulations) and rarer cell types, such as dendritic cells. However, within these populations, differences between the protocols pointed to the

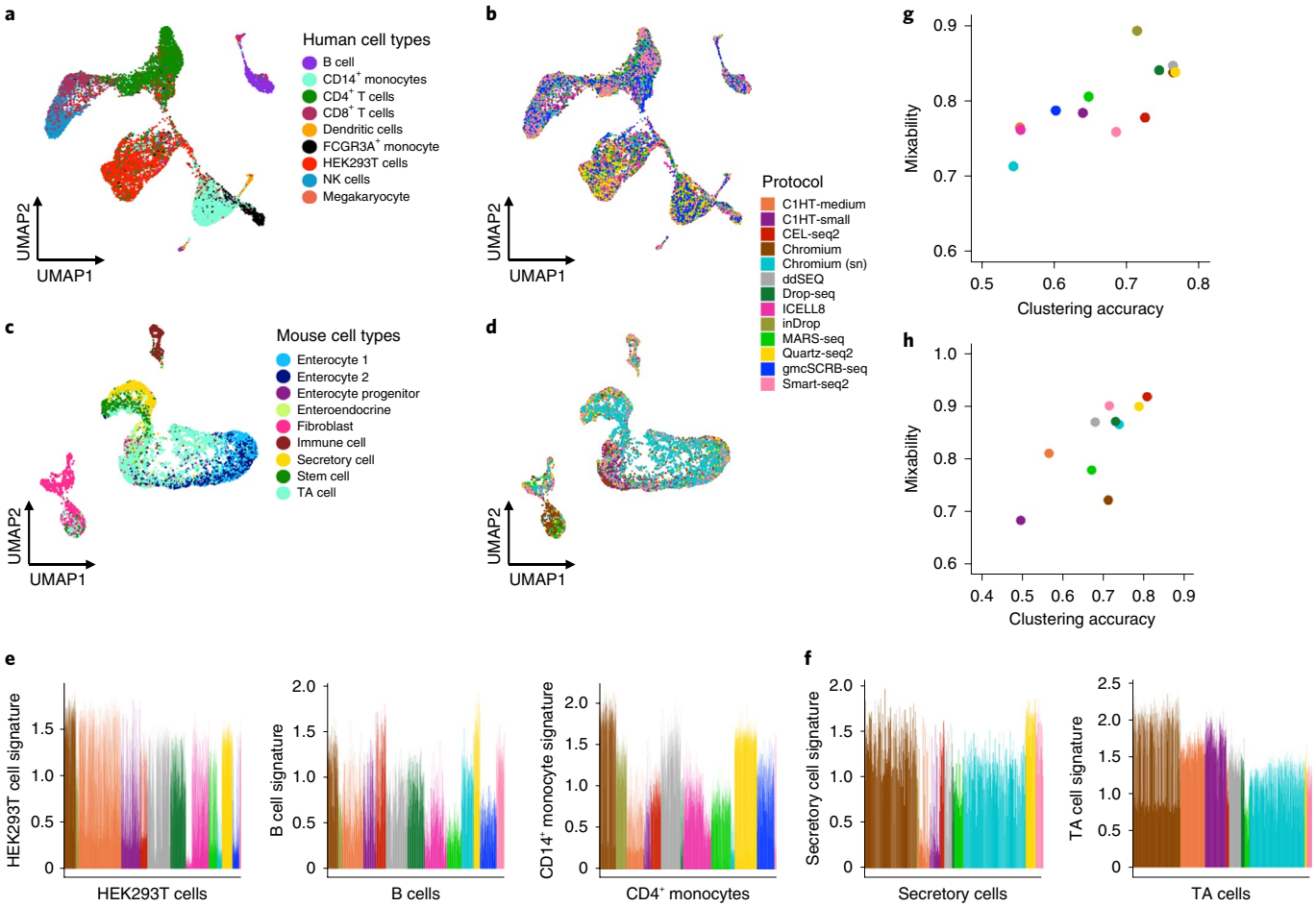

Fig. 5 | Integration of sc/snRNA-seq methods. a–d, UMAP visualization of cells after integrating technologies for 18,034 human (**a**,**b**) and 7,902 mouse (**c**,**d**) cells. Cells are colored by cell type (**a**,**c**) and sc/snRNA-seq protocol (**b**,**d**). **e**,**f**, Barplots showing normalized and method-corrected (integrated) expression scores of cell-type-specific signatures for human HEK293T cells, monocytes, B cells (**e**), and mouse secretory and TA cells (**f**). Bars represent cells and colors methods. **g**,**h**, Evaluation of method integratability in human (**g**) and mouse (**h**) cells. Protocols are compared according to their ability to group cell types into clusters (after integration) and mix with other technologies within the same clusters. Points are colored by sequencing method.

presence of technical effects that could not be entirely removed with down-sampling to equal read depth and different merging tools (Fig. 5e,f, and see Supplementary Figs. 19c,d, 21a,b and 22a,b). To formally assess the capacity of the methods to be combined, we calculated the degree to which technologies mix in the merged datasets (Fig. 5g,h, and see Supplementary Figs. 21c,d and 22c,d). The suitability of protocols to be combined (mixability) was directly correlated with their power to discriminate between cell types (clustering accuracy). Thus, well-performing protocols result in high-resolution cellular maps and are suitable for consortium-driven projects that include different data sources. When integrating further down-sampled datasets, we observed a drop in mixing ability (see Supplementary Fig. 19e). Consequently, quality standard guidelines for consortia might define minimum coverage thresholds to ensure the subsequent option of data integration. A separate analysis of the single-nucleus and single-cell Chromium datasets resulted in well-integrated profiles, further supporting the potential to integrate cell atlases from cells and nuclei (see Supplementary Figs. 23 and 24).

Cell atlas datasets will serve as a reference for annotating cell types and states in future experiments. Therefore, we assessed cells' ability to be projected on to our reference sample (Fig. 2b,c). We used the population signature model defined by matchSCore2 and evaluated the protocols based on their cell-by-cell mapping probability, which reflects the confidence of cell annotation (see Supplementary Fig. 25a–c). Although there were some differences in the projection probabilities of the protocols, and a potential bias due to the selection of the reference protocol, a confident annotation was observed for most cells with inDrop and ddSEQ reporting the highest probabilities. Notably, high probability scores were also observed in further down-sampled datasets (see Supplementary Fig. 25b). This has practical consequences, because data derived from less well-performing methods (from a cell atlas perspective), or from poorly sequenced experiments, could be identifiable and thus suitable for specific analysis types, such as tissue composition profiling.

Discussion

Systematic benchmarking of available technologies is a crucial prerequisite for large-scale projects. In the present study, we evaluated scRNA-seq protocols for their power to produce a cellular map of complex tissues. Our reference sample simulated common scenarios in cell atlas projects, including differentiated cell types and dynamic cell states. We defined the strengths and weaknesses of key features that are relevant for cell atlas studies, such as comprehensiveness, integratability and predictive value. The methods revealed a broad spectrum of performance, which should be considered when defining guidelines and standards for international consortia (Fig. 6).

We expect that our results will guide informed decision-making processes for designing sc/snRNA-seq studies. There are several features to consider when selecting protocols to produce a

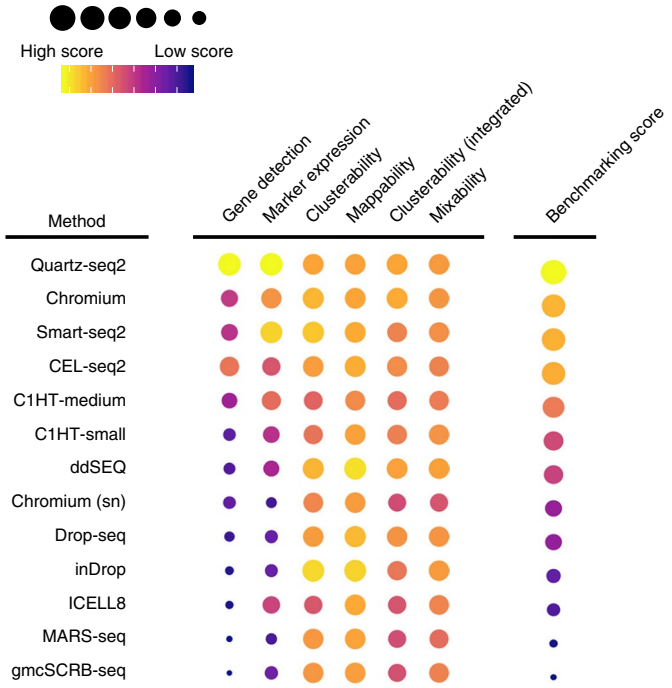

Fig. 6 | Benchmarking summary of 13 sc/snRNA-seq methods. Methods are scored by key analytical metrics, characterizing protocols according to their ability to recapitulate the original structure of complex tissues, and their suitability for cell atlas projects. The methods are ordered by their overall benchmarking score, which is computed by averaging the scores across metrics assessed from the human datasets.

reproducible, integrative and predictive reference cell atlas. At a given sequencing depth, the number and complexity of detected RNA molecules define the power to describe cell phenotypes and infer their function. There are also additional essential features for cell atlas projects and their interpretation, such as population marker identification. Improved versions of plate-based methods, including Quartz-seq2, CEL-seq2 and Smart-seq2, generate such high-resolution transcriptome profiles. Also, microfluidic systems showed excellent performance in our comparison, particularly the Chromium system. Although the scale of plate-based experiments is limited by the lower throughput of their individual processing units, microfluidic systems, especially droplet-based methods, can be easily applied to thousands of cells simultaneously. Protocol modification scales up throughput even further, and allows more cost-effective experiments[26–29]. Generally, late multiplexing methods, such as Smart-seq2, are more costly, but costs can be reduced by miniaturization[30] and use of noncommercial enzymes[31]. Custom droplet-based protocols have lower costs than their commercialized counterparts, but the optimized chemistry in commercial systems resulted in improved performance in this comparison. Nevertheless, existing platforms are undergoing continued development in both the private (see Supplementary Fig. 12) and the academic sectors, so updated protocol versions promise to improve performance further. For consortium-driven projects, it is important to consider the integratability of data. We have shown that several protocols, including those with reduced library complexity and snRNA-seq, were readily integratable with other methods.

The use of PBMCs is ideal for multicenter benchmarking efforts; blood cells are easy to isolate and show a high recovery rate after freezing. We also included mouse colon, a solid tissue requiring dissociation before scRNA-seq. Tissue digestion and cryopreservation of colon cells present additional challenges (for example, increased rate of damaged cells), which we addressed by focusing on commonly detected cell types. Although we observed differences in the frequencies of cells from mice and humans, the composition of cell subtypes within tissues was conserved, reassuring the consistent capture of major cell types across all methods. Accordingly, subsequent analyses could be stratified by cell type, avoiding the need for a ground truth in sample composition. Furthermore, viability sorting with minimal mechanical forces (low speed and wide nozzle size) was applied to remove damaged cells and benchmark protocols with high-quality samples. This work standardized sample processing to limit technical variance in the library preparation steps, a crucial requisite for the multicenter benchmarking design. Nevertheless, on-site differences introduced during sample thawing or viability sorting could not be entirely excluded. However, our analysis also showed that viable cells selected by sorting or through thorough data quality control generate highly similar library complexity, suggesting that potential differences in sample processing have minor impacts on the data quality and supporting the robustness of our results. Processing time presents another variable related to sample and data quality. Although cells are directly sorted into their respective reaction volumes for plate-based methods, processing times can vary across microfluidic systems. However, this was considered to be an inherent feature of the library preparation workflow of the protocols that contributes to the overall performance.

Across sample origins and cell types, all tested features pointed to consistent protocol performance. In addition to the differences in protocol performance, it was the cells' RNA content and complexity that dominated the molecule and gene detection rates, which we have seen through the stratified analysis of vastly different cell types. As such, we expect the conclusions to be valid beyond the human and mouse tissues tested in the present study.

Several additional steps are crucial for the success of single-cell projects, especially sample preparation. Optimization of sample procurement and tissue-processing conditions is of crucial importance to avoid composition biases and gene expression artifacts[32–35] that could limit the value of a cell atlas. Therefore, dedicated studies are required to define optimal conditions for tissue and organ preparation in healthy and disease contexts.

From a technical perspective, multiple steps of a protocol are critical for generating complex sequencing libraries. All sc/snRNA-seq methods require multi-step, whole-transcriptome amplification, including reverse transcription, conversion to amplifiable cDNA and amplification[1]. Theoretically, the multiplicative reaction efficiency of respective steps determines a method's power to detect RNA molecules, and in this sense Quartz-Seq2 was particularly efficient. We specifically tested for potential advantages of the Quartz-seq2 column-based over bead-based purification, but did not detect differences in cDNA yield (see Supplementary Fig. 26). However, we observed that bead concentration critically affected the yield of amplified cDNA. Moreover, performance was more stable for purification with columns compared with beads, which should be taken into account when implementing existing or developing new sc/snRNA-seq methods.

A further essential step toward complex libraries is the conversion of first-strand cDNA to amplifiable cDNA. Three main strategies are used for this conversion: (1) template switching, (2) RNaseH/DNA polymerase I-mediated, second-strand synthesis for in vitro transcription and (3) poly(A) tagging[1]. Improvement of the three strategies led to better quantitative performance of scRNA-seq[36–39]. For Quartz-Seq2 (ref. [37]), improved poly(A) tagging was most important to increase the amplified cDNA yield compared with Quartz-Seq[40], and probably explains the excellent result in this benchmarking exercise. However, optimization of the cDNA conversion still has the potential to improve scRNA-seq methods.

Within the cDNA amplification step, increased PCR cycle numbers lead to PCR biases within the sequencing libraries. Early pooling increases the number of cDNA molecules in the amplification

step and reduces PCR bias. This especially favors early pooling methods at low sequencing depth (as performed in the present study), as previously shown for bulk RNA-seq[41]. Similarly, in vitro transcription linearly amplifies cDNA with fewer biases than PCR-based methods, and partly explains the good performance of CEL-seq2. Furthermore, early multiplexing of different cell numbers leads to different PCR cycle requirements (Quartz-Seq2 with 768 cells and 10 cycles versus gmcSCRB-seq with 96 cells and 19 cycles, using the same DNA polymerase for amplification). The number of cells per amplification pool depends on the amount of amplifiable cDNA, implying that the good performance of Quartz-Seq2 was mainly due to efficient conversion of amplifiable cDNA from RNA with poly(A) tagging.

It is equally important to benchmark computational pipelines for data analysis and interpretation[23,42–44]. We envision the datasets provided by our study serving as a valuable resource for the single-cell community to develop and evaluate new strategies for an informative and interpretable cell atlas. Moreover, the multicenter benchmarking framework presented in the present study can readily be transferred to other organs where common tissue/cell types are analyzed using different scRNA-seq protocols (for example, brain atlas projects).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0469-4.

Received: 7 May 2019; Accepted: 26 February 2020;
Published online: 06 April 2020

References

1. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
2. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
3. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
4. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
5. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
6. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
7. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
8. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
9. Regev, A. et al. Science forum: the human cell atlas. *eLife* **6**, e27041 (2017).
10. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
11. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
12. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **113**, 11046–11051 (2016).
13. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
14. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
15. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
16. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
17. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
18. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
20. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
21. Guillaumet-Adkins, A. et al. Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol.* **18**, 45 (2017).
22. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
23. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
24. Azuaje, F. A cluster validity framework for genome expression data. *Bioinforma* **18**, 319–320 (2002).
25. Lin, Y. et al. scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl Acad. Sci. USA* **116**, 9775–9784 (2019).
26. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
27. Stoeckius, M. et al. Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
28. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
29. Gaublomme, J. T. et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 1–8 (2019).
30. Mora-Castilla, S. et al. Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* **21**, 557–567 (2016).
31. Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
32. Brink, S. Cvanden et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
33. Wohnhaas, C. T. et al. DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing. *Sci. Rep.* **9**, 1–14 (2019).
34. Tosti, L. et al. Single nucleus RNA sequencing maps acinar cell states in a human pancreas cell atlas. Preprint at *bioRxiv* https://doi.org/10.1101/733964 (2019).
35. Massoni-Badosa, R. et al. Sampling artifacts in single-cell genomics cohort studies. Preprint at *bioRxiv* https://doi.org/10.1101/2020.01.15.897066 (2020).
36. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
37. Sasagawa, Y. et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
38. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
39. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
40. Sasagawa, Y. et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.* **14**, 3097 (2013).
41. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
42. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
43. Saelens, W. et al. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
44. Holland, C. H. et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36 (2020).

Methods

Ethical statement. The present study was approved by the Parc de Salut MAR Research Ethics Committee (reference no. 2017/7585/I) to H.H. We adhered to ethical and legal protection guidelines for human participants, including informed consent.

Reference sample. *Cell lines.* NIH3T3-GFP, MDCK-TurboFP650 and HEK293-RFP cells were cultured at 37 °C in an atmosphere of 5% (v:v) carbon dioxide in Dulbecco's modified Eagle's medium, supplemented with 10% (w:v) fetal bovine serum (FBC), 100 U penicillin, and 100 µgl^{-1} of streptomycin (Invitrogen). On the reference sample preparation day, the culture medium was removed and the cells were washed with 1× phosphate-buffered saline (PBS). Afterwards, cells were trypsinized (trypsin 100×), pelleted at 800*g* for 5 min, washed in 1× PBS, resuspended in PBS + ethylenediaminetetraacetic acid (EDTA) (2 mM) and stored on ice.

Mouse colon tissue. The colons from 11 mice (7 *LGR5/GFP* and 4 wild-type) were dissected and removed. For single-cell separation the colons were treated separately. The colon was sliced, opened and washed twice in cold 1× Hank's balanced salt solution (HBSS). It was then placed on a Petri dish on ice and minced with razor blades until disintegration. The minced tissue was transferred to a 15-ml tube containing 5 ml of 1× HBSS and 83 µl of collagenase IV (final concentration 166 U ml^{-1}). The solution was incubated for 15 min at 37 °C (vortexed for 10 s every 5 min). To inactivate the collagenase IV, 1 ml of FBS was added and it was vortexed for 10 s. The solution was filtered through a 70-µm nylon mesh (changed when clogged). Finally, all samples were combined, and the cells pelleted for 5 min at 400*g* and 4 °C. The supernatant was removed and the cells resuspended in 20 ml of 1× HBSS and stored on ice.

Isolation of PBMCs. Whole blood was obtained from four donors (two female, two male). The extracted blood was collected in heparin tubes (GP Supplies) and processed immediately. For each donor, PBMCs were isolated according to the manufacturer's instructions for Ficoll extraction (pluriSelect). Briefly, blood from two heparin tubes (approximately 8 ml) was combined, diluted in 1× PBS and carefully added to a 50-ml tube containing 15 ml of Ficoll. The tubes were centrifuged for 30 min at 500*g* (minimum acceleration and deceleration). The interphase was carefully collected and diluted with 1× PBS + 2 mM EDTA. After a second centrifugation, the supernatant was discarded and the pellet resuspended in 2 ml of 1× PBS + 2 mM EDTA and stored on ice.

Preparation of the reference sample. Cell counting was performed using an automated cell counter (TC20 Automated Cell Counter, Bio-Rad Laboratories). The reference sample was calculated to include human PBMCs (60%), mouse colon cells (30%), and HEK293T (6%, RFP-labeled human cell line), NIH3T3 (3%, GFP-labeled mouse cells) and MDCK (1%, TurboFP650-labeled dog cells) cells. To adjust for cell integrity loss during sample processing, we measured the viability during cell counting and accounted for an expected viability loss after cryopreservation (10% for cell lines and PBMCs; 50% for colon cells[21]). All single-cell solutions were combined in the proportions mentioned above and diluted to 250,000 viable cells per 0.5 ml. For cryopreservation, 0.5 ml of cell suspension was aliquoted into cryotubes and gently mixed with a freezing solution (final concentration 10% dimethylsulfoxide; 10% heat-inactivated FBS). Cells were then frozen by gradually decreasing the temperature (1 °C min^{-1}) to −80 °C (cryopreserved), and stored in liquid nitrogen. MARS-Seq and Smart-Seq2 experiments were performed to validate sample quality and composition before distributing aliquots to the partners.

Sample processing. Samples were stored at −80 °C on arrival. Before processing, samples were de-frozen in a water bath (37 °C) with continuous agitation until the material was almost thawed. The entire volume was transferred to a 15-ml Falcon tube using a 1,000-µl tip (wide-bored or cut tip) without mixing by pipetting; 1,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample. The sample was then rested for 1 min. An additional 2,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample. The sample was again rested for 1 min. Another 2,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise while gently swirling the sample and the sample was rested for 1 min. Then, 3,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise and the Falcon tube inverted six times. The sample was rested for 1 min. An additional 5,000 µl of prewarmed (37 °C) Hibernate-A was added drop-wise and the Falcon tube inverted six times. The sample was rested for 1 min. It was then centrifuged at 400*g* for 5 min at 4 °C (pellet clearly visible). The supernatant was removed until 500 µl remained in the tube. The pellet was resuspended by gentle pipetting. Then 3,500 µl of 1× PBS + 2 mM EDTA was added and the sample stored on ice until processing. Before FACS isolation, cells were filtered through a nylon mesh and 3 µl DAPI was added before gentle mixing. During FACS isolation, DAPI-positive cells were excluded to remove dead and damaged cells. Furthermore, the exclusion of GFP-positive cells simulated the removal of a cell type from a complex sample. Supplementary Fig. 27 shows representative FACS plots and gating strategies.

ScRNA-seq library preparation. For a detailed sample processing description, see Supplementary Notes.

Data analysis. For primary data preprocessing, clustering, sample deconvolution and annotation, and reference datasets, see Supplementary Notes.

MatchSCore2. To systematically assign cell identities to unannotated cells coming from different protocols, we used matchSCore2, a mathematical framework for classifying cell types based on reference data (https://github.com/elimereu/matchSCore2). The reference data consist of a matrix of gene expression counts in individual cells, the identity of which is known. The main steps of the matchSCore2 annotation are the following:

(1) Normalization of the reference data. Gene expression counts are log(normalized) for each cell using the natural logarithm of 1 + counts per 10,000. Genes are then scaled and centered using the ScaleData function in the Seurat package.
(2) Definition of signatures and their relative scores. For each of the cell types in the reference data, positive markers were computed using Wilcoxon's rank-sum test. The top 100 ranked markers in each cell type were used as the signature for that type. To each cell, we assigned a vector $\mathbf{x} = (\mathbf{x}_1, .., \mathbf{x}_n)$ of signature scores, where n is the number of cell types in the reference data. The ith signature score for the kth cell is computed as follows:

$$\text{Score}_k = \sum\nolimits_{j \text{ in } J} z_{jk}$$

where J is the set of genes in signature i, and z_{jk} represents the z-score of gene j in the kth cell.
(3) Training of the probabilistic model on the reference data.

We proposed a supervised multinomial logistic regression model, which uses enrichment of the signature of each reference cell type in each cell to assign identity to that cell. In other words, for each cell k and signature i, we calculate the ith cell-type signature score $\mathbf{x}_i$ in the kth cell as described in point 2. The distribution of the signature scores is preserved, independent of which protocol is used (see Supplementary Figs. 28 and 29). More specifically, we defined the variables $\mathbf{x}_1$, …, $\mathbf{x}_n$, where $\mathbf{x}_i$ is the vector in which the scores for signature i of all cells are contained. Then we used $\mathbf{x}_i$ as the predictor of a multinomial logistic regression.

The model assumes that the number of cells from each type in the training reference data T_1, T_2, …,T_n are random variables and that the variable $T = (T_1, T_2, …,T_n)$ follows a multinomial distribution $M(N, \pi = (\pi_1, …, \pi_n))$, where π_i is the proportion of the ith cell type and N is the total number of cells.

To test the performance of the model, training and test sets were created by subsampling the reference into two datasets, maintaining the original proportions of cell types in both sets. The model was trained by using the multinom function from the nnet R package (decay = 1×10^{-4}, maxit = 500). To improve the convergence of the model function, $\mathbf{x}_i$ variables were scaled to the interval [0,1].

Cell classification. For each cell, model predictions consisted of a set of probability values per identity class, and the highest probability was used to annotate the cell if it was >0.5; otherwise the cell remained unclassified.

Model accuracy. To evaluate the fitted model using our reference datasets, we assessed the prediction accuracy in the test set, which was around 0.9 for human and 0.85 for mouse reference. We further assessed matchSCore2 classifications in datasets from other sequencing methods by looking at the agreement between clusters and classification. Notably, the resulting average agreement was 80% (range: from 58% in gmcSCRB-seq to 92% in Quartz-Seq2), whereas the rate for unclassified cells was <2%.

Down-sampling. To decide on a common down-sampling threshold for sequencing depth per cell, we inspected the distribution of the total number of reads per cell for each technique, and chose the lowest first quartile (fixed to 20,000 reads per cell). We then performed stepwise down-sampling (25%, 50% and 75%) using the zUMIs down-sampling function. We omitted cells that did not achieve the required minimum depth (see Supplementary Table 6). Notably, stochasticity introduced during down-sampling did not affect the results of the present study, as exemplified by the consistent numbers of detected molecules across different down-sampling iterations (see Supplementary Fig. 10).

Estimation of dropout probabilities. We investigated the impact of dropout events in HEK293T cells, monocytes and B cells extracted for each technique on down-sampled data (20,000 reads per cell). For datasets with >50 cells from the selected populations, we randomly sampled 50 cells to eliminate the effect of differing cell number. The dropout probability was computed using the SCDE R package[45]. SCDE models the measurements of each cell as a mixture of a negative binomial process to account for the correlation between amplification and detection of a transcript and its abundance, and a Poisson process to account for the background signal. We then used estimated individual error models for each cell as a function of expression magnitude to compute dropout probabilities using

SCDE's scde.failure.probability function. Next, we calculated the average estimated dropout probability for each cell type and technique. To integrate dropout measures into the final benchmarking score, we calculated the area under the curve of the expression prior and failure probabilities (see Fig. 2f and also Supplementary Table 7). We expected that protocols resulting in fewer dropouts would have smaller areas under the curve.

Quantification of variance introduced by batches. To quantify the amount of variance that is introduced by batches (protocols, processing units or experiments), we used the top 20 PCs and the s.d. of each PC, previously calculated on HVGs. Next, using the pcRegression function of kBET R package[23], we regressed the batch covariate (protocols/processing units/experiments as categories defined in the kBET model) and each PC to obtain the coefficient of determination as an approximation of the variance explained by batches, and the proportions of explained variance in each PC. We either reported the percentage of the variance that correlates significantly with the batch in the first 20 PCs, or R-squared measures of the model for each PC.

Cumulative number of genes. The cumulative number of detected genes in the down-sampled data was calculated separately for each cell type. For cell types with >50 cells annotated, we randomly selected 50 cells and calculated the average number of detected genes per cell after 50 permutations over n sampled cells, where n is an increasing sequence of integers from 1 to 50.

GO enrichment analysis. To compare functional gene sets between single-cell and single-nucleus datasets, we performed Gene Ontology (GO) enrichment analysis on the set of protocol-specific genes using simpleGO (https://github.com/iaconogi/simpleGO). For each cell type (HEK293T cells, monocytes and B cells), we selected two gene sets extracted from the cumulated genes and using the maximum number of detected cells common to all three Chromium versions: (1) genes that were uniquely detected in the intersection of Chromium (v.2) and (v.3), but not in Chromium (sn), and (2) genes that were uniquely identified with Chromium (sn). For each of the gene sets, we identified the union over cell types before applying simpleGO.

Correlation analysis. Pearson's correlations across protocols were computed independently for B cells, monocytes and HEK293T cells. For each cell type, cells were down-sampled to the maximum common number of cells across all protocols. Gene counts of commonly expressed genes (from datasets down-sampled to 20,000 reads) were averaged across cells before computing their Pearson's correlations. The corplot library was then used to plot the resulting correlations. Protocols were ordered by agglomerative hierarchical clustering.

Silhouette scores. To measure the strength of the clusters, we calculated the ASW[24]. The down-sampled data (20,000 reads per cell) were clustered by Seurat[46], using graph-based clustering with the first eight PCs and a resolution of 0.6. We then computed an ASW for the clusters using a Euclidean distance matrix (based on PCs 1–8). We reported the ASW for each technique separately.

Dataset merging. Dataset integration across protocols is challenging and we applied different tools to assess the integratability of the sc/snRNA-seq methods, while conserving biological variability. To integrate datasets, we used Seurat[46], harmony[47] and scMerge[25], evaluated the results separately and averaged the integration capacity of the protocols into a joint score. We combined down-sampled count matrices using the sce_cbind function in scMerge, which includes the union of genes from different batches. Although both harmony and Seurat integration apply similar preprocessing steps (log(normalization), scaling and HVG identification), as implemented in the Seurat tool, scMerge uses a set of genes with stable expression levels across different cell types, and then creates pseudo-replicates across datasets, allowing the estimation and correction for undesired sources of variability. However, for all three alignment methods, Seurat was applied to perform clustering and Uniform Manifold Approximation and Projection (UMAP) after the protocol correction, to minimize the variability related to the downstream analysis. The clustering accuracy metric was used together with the mixability score to quantify the success of the integration. Omitting the cell integration step before visualizing the datasets together in a single tSNE/UMAP resulted in a protocol-specific distribution with cell types scattered to multiple clusters (see Supplementary Fig. 30).

Clustering accuracy. To determine the clusterability of methods to identify cell types, we measured the probability of cells being clustered with cells of the same type. Let C_k, $k\in\{1,\ldots,N\}$ represent the cluster of cells corresponding to a unique cell type (based on the highest agreement between clusters and cell types), and T_j, $j\in\{1,\ldots,S\}$ represent the set of different cell types, where $C\subseteq T$. For each cell type T_j, we compute the proportion p_{jk} of T_j cells that cluster in their correct cluster C_k. We define the cell-type separation accuracy as the average of these proportions.

Mixability. To account for the level of mixing of each technology, we used kBet[23] to quantify batch effects by measuring the rejection rate of Pearson's χ^2 test for random neighborhoods. To make a fair comparison, kBet was applied to the common cell types separately by subsampling batches to the minimum number of cells in each cell type. Due to the reduced number of cells, the option heuristic was set to 'False', and the testSize was increased to ensure a minimum number of cells.

Mixability was calculated by averaging cell-type-specific rejection rates.

Benchmarking score. To create an overall benchmarking score against which to compare technologies, we considered six key metrics: gene detection, overall level of expression in transcriptional signatures, cluster accuracy, classification probability, cluster accuracy after integration and mixability. Each metric was scaled to the interval [0,1], then, to equalize the weight of each metric score, the harmonic mean across these metrics was calculated to obtain the final benchmarking scores. Gene detection, overall expression in cell-type signatures and classification probabilities were computed separately for B cells, HEK293T cells and monocytes, and then aggregated by the arithmetic mean across cell types. Notably, the choice of protocol to create the reference dataset (Chromium) for initial cell annotation had no impact on the outcome of the present study (see Supplementary Fig. 31).

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

Data availability

All raw sequencing data and processed gene expression files are freely available through the Gene Expression Omnibus (accession no. GSE133549).

Code availability

All code for the analysis is provided as supplementary material. All code is also available under https://github.com/ati-lz/HCA_Benchmarking and https://github.com/elimereu/matchSCore2.

References

45. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
46. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
47. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

Acknowledgements

This project has been made possible in part by grant no. 2018-182827 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII). C.M. is supported by an AECC postdoctoral fellowship. This work has received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2015-675752 (Singek), and the Ministerio de Ciencia, Innovación y Universidades (SAF2017-89109-P; AEI/FEDER, UE). S. was supported by the German Research Foundation's (DFG's) (GR4980) Behrens-Weise-Foundation. D.G. and S. are supported by the Max Planck Society. C.Z. was supported by the European Molecular Biology Organization through the long-term fellowship ALTF 673-2017. The snRNA-seq data were generated with support from the National Institute of Allergy and Infectious Diseases (grant no. U24AI118672), the Manton Foundation and the Klarman Cell Observatory (to A.R.). I.N. was supported by JST CREST (grant no. JPMJCR16G3), Japan, and the Projects for Technological Development, Research Center Network for Realization of Regenerative Medicine by Japan, the Japan Agency for Medical Research and Development. A.J., L.E.W., J.W.B. and W.E. were supported by funding from the DFG (EN 1093/2-1 and SFB1243 TP A14). We thank ThePaperMill for critical reading and scientific editing services and the Eukaryotic Single Cell Genomics Facility at Scilifelab (Stockholm, Sweden) for support. This publication is part of a project (BCLLATLAS) that received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 810287). Core funding was from the ISCIII and the Generalitat de Catalunya.

Author contributions

H.H. designed the study. E.M. and A.L. performed all data analyses. C.M., A.A.V. and E.B. prepared the reference sample. C.Z., D.J.M., S.P. and O.S. supported the data analysis. M.G. and I.G. provided technical and sequencing support. S., D.G., J.K.L., S.C.B., C.S., A.O., R.C.J., K.K., C.B., Y.T., Y.S., K.T., T.H., C.B., C.F., S.S., T.T., C.C., X.A., L.T.N., A.R., J.Z.L., A.J., L.E.W., J.W.B., W.E., R.S. and I.N. provided sequencing-ready single-cell libraries or sequencing raw data. H.H., E.M. and A.L. wrote the manuscript with contributions from the co-authors. All authors read and approved the final manuscript.

Competing interests

A.R. is a co-founder and equity holder of Celsius Therapeutics, and an SAB member of Thermo Fisher Scientific and Syros Pharmaceuticals. He is also a co-inventor on patent applications to numerous advances in single-cell genomics, including droplet-based

sequencing technologies, as in PCT/US2015/0949178, and methods for expression and analysis, as in PCT/US2016/059233 and PCT/US2016/059239. K.K., C.B. and Y.T. are employed by Bio-Rad Laboratories. J.K.L. and S.C.B. are employees and shareholders at 10x Genomics, Inc. S.C.B. is a former employee and shareholder of Fluidigm Corporation. C.S. and A.O. are employed by Fluidigm. All other authors declare no conflicts of interest associated with this manuscript.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0469-4.

Correspondence and requests for materials should be addressed to H.H.

nature research

Corresponding author(s): Holger Heyn

Last updated by author(s): Jan 20, 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	The data collection code is available under https://github.com/ati-lz/HCA_Benchmarking
Data analysis	The analysis code is available under https://github.com/ati-lz/HCA_Benchmarking AND https://github.com/elimereu/matchSCore2. Seurat package (version 2) Monocle package (version 2.8.0)

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 raw sequencing data and processed gene expression files are available through the Gene Expression Omnibus (GEO; GSE133549).

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	Not applicable.
Data exclusions	Single-cell datasets were quality filtered as described in the Online Methods.
Replication	Single-cell sequencing of the same reference sample was replicated with 13 protocols for comparative analysis.
Randomization	Not applicable.
Blinding	All partners had knowledge of the reference sample design.

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)	HEK-293T_RFP: AMSBIO; MDCK_FP650: Cell Trend; NIH3T_GFP: Cell biolabs
Authentication	Cell lines authenticated by commercial providers.
Mycoplasma contamination	The cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	Not applicable. All lines were ordered from commercial sources.

Animals and other organisms

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

Laboratory animals	Mus musculus, C57BL/6 (LGR5-gfp-creERT2), 2 month, 5 male and 6 female.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected in the field.
Ethics oversight	No ethical approval was required.

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Four healthy donors (2 male / 2 female) of Caucasian ancestry (age 30-40 years).
Recruitment	Donor recruitment within the local work environment of the leading institute.
Ethics oversight	This study was approved by the Parc de Salut MAR Research Ethics Committee (reference number: 2017/7585/I).

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