Building a high-quality Human Cell Atlas

To the Editor — Building the Human Cell Atlas (HCA) requires consistent and agile experimental designs, standardized operating protocols (SOPs), benchmarks and quality control metrics that can adapt to a rapidly evolving technological landscape. Here, the HCA Standards and Technology Working Group outlines pertinent technical challenges and their approach to defining benchmarks and quality control measures to ensure high-quality data for building a comprehensive and accurate human cell atlas and help guide other atlas projects in health and disease.

The HCA aims to create comprehensive reference maps of all human cells, the fundamental units of life, as a basis for both understanding human health and diagnosing, monitoring and treating disease[1]. By integrating single-cell resolved molecular profiles of tissues and organs, it seeks to generate cellular and spatial maps, including the identification of dynamic cell states and rare cell populations. This effort requires the generation of high-quality data in multiple laboratories around the world, first assembled as a draft atlas and then increased in resolution and breadth by continued contributions from the community over time. The HCA community is open and collaborative, sharing its data through an open-source data coordination platform (DCP; https://www.humancellatlas.org/data-coordination/) and bringing together and aligning biological, clinical, computational and engineering experts from diverse fields. Since the HCA's launch in October 2016, more than 1,700 scientists from across the globe have enthusiastically joined to help shape this effort, through scientific collaboration, planning meetings, computational jamborees, social media and funding calls. Dedicated participants have formed biological networks spanning organs and systems, established their scientific leadership, and rapidly embarked on large scale data collection and analysis to build draft atlases.

As data collection for the HCA spans labs and techniques, spanning years and many technical innovations, it requires careful experimental design to construct a cohesive atlas. The HCA community is committed to producing the highest quality data possible and establishing rigorous standards, shared openly and broadly and updated regularly, ensuring findable, accessible, interoperable and reusable (FAIR) data principles[2]. The HCA develops, adopts and shares new tools for comprehensive and multidimensional atlas production. It also maintains flexibility, so it can revise the design of the HCA production and analysis as new insights, data and technologies emerge. Design considerations include both the choice of existing data-generating technologies and efforts to develop and assess new technologies with more measurement capabilities, increased scale and/or lower cost. Thus, benchmarking HCA data and technologies is a priority both for the HCA initiative and more broadly for the single-cell genomics field — including related efforts in specific disease areas or model organisms, such as the BRAIN Initiative Cell Census Network (BICCN), the Cancer Moonshot Human Tumor Atlas Network (HTAN) and LifeTime.

To fulfill this mission, the HCA established a Standards and Technology Working Group (STWG) not only to guide and advise its members around technology choices but also to outline best practices and help coordinate and carry out scientific work to support this mission (Fig. 1). The STWG encompasses 18 members from 14 institutions across 7 countries, spanning diverse areas of expertise. The STWG initiates and leads efforts to compare, test and benchmark existing methods, the deployment of new methods and the development of analytics and quality control measures.

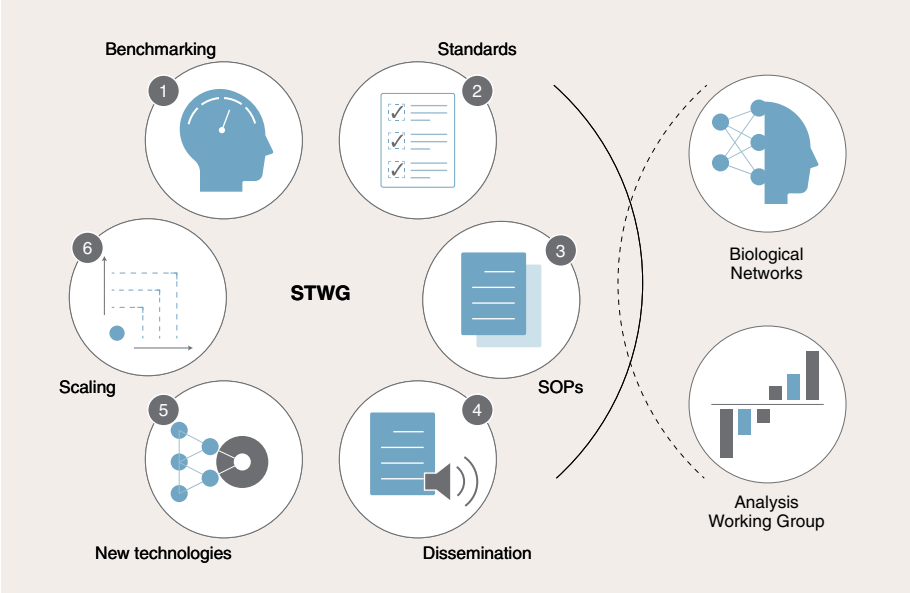

Fig. 1 | Mission of the Standards and Technology Working Group (STWG). The STWG identifies and benchmarks emerging technologies for the HCA. Before dissemination to the community, SOPs and quality control measures are defined and protocols are scaled for a cost-effective large-scale production phase. A close interaction with the Biological Networks group and Analysis Working Group ensures seamless and timely incorporation into the HCA's operations.

Data generation for the HCA

Recent advances in single-cell and spatial genomics have made it possible to build a human cell atlas[3]. This atlas relies on high-resolution measurements along two major branches: a cellular branch, based on profiling of cells or nuclei, and a spatial branch, to measure profiles in the tissue context[1].

The measurements underlying an atlas span diverse molecular aspects of cells and tissues, including the transcriptome, genome, epigenome, proteome and metabolome, as well as structural features of cells, tissues and organs. In the cellular branch, massively parallel single-cell and single-nucleus RNA sequencing (sc/snRNA-seq) allow fast and cost-effective profiling of transcriptomes of millions of individual cells and are among the primary technologies used for HCA data generation to date. In parallel, single-cell approaches for the profiling of other genomic features, such as chromatin states (for example, single-cell ATAC-seq and methylC-seq), and for joint measurement of multi-omic profiles

(for example, RNA and DNA, RNA and epigenome, and RNA and protein) have also rapidly matured and scaled[4–6]. In particular, scATAC-seq is now available at a scale comparable to that of scRNA-seq[7]. However, single-cell and single-nucleus profiling require cell dissociation or nuclei isolation, which erases critical information about spatial organization. Thus, the spatial branch is equally critical for atlas assembly; it relies on emerging technologies providing spatial information, including multiplexed in situ assays for RNA and protein (for example, imaging, sequencing, spatial coding and computational inference). Although sc/snRNA-seq techniques from dissociated cells are more consolidated and broadly disseminated ('production'), emerging methods for spatial and multi-omics profiling are progressing through refinement and dissemination ('scaling') to production[1].

The complexity of the data collection landscape highlights several core challenges for the HCA. First, the existing plethora of methods and protocols can be challenging to data generators who need to make informed choices or compare their results to those of others. Second, method developers need means to identify key areas for technical improvement and to compare their results with those of other techniques. Third, the rapid development of new methods by a highly engaged community requires strategies to adopt and disseminate new methods throughout the HCA. Finally, as is now increasingly appreciated, human tissues and organs vary widely, as do sample types (biopsy, resection and autopsy), such that a method's performance may vary substantially depending on the tissue or organ to which it is applied.

To account for a wide range of profiling methods and tissues of application and to devise effective strategies to address unavoidable biases and batch effects from multiple laboratories, technologies and protocols, we propose to use distinct methods (from sample preparation[8] to data production[9,10]) to examine the same tissues while systematically applying a smaller number of technologies with consistent SOPs across all tissues. For example, because tissue-processing protocols vary in cell recovery[11] and it is not clear which, if any, provide the full ground truth of cellular composition, generating datasets with varying experimental designs and technologies can increase the completeness of the first atlas draft. Dedicated projects have started to assess potential challenges in multi-site data generation, including the initial stages of sample preparation[12,13] and data production[9], with the goal of producing recommendations for how best to collect data for the atlas.

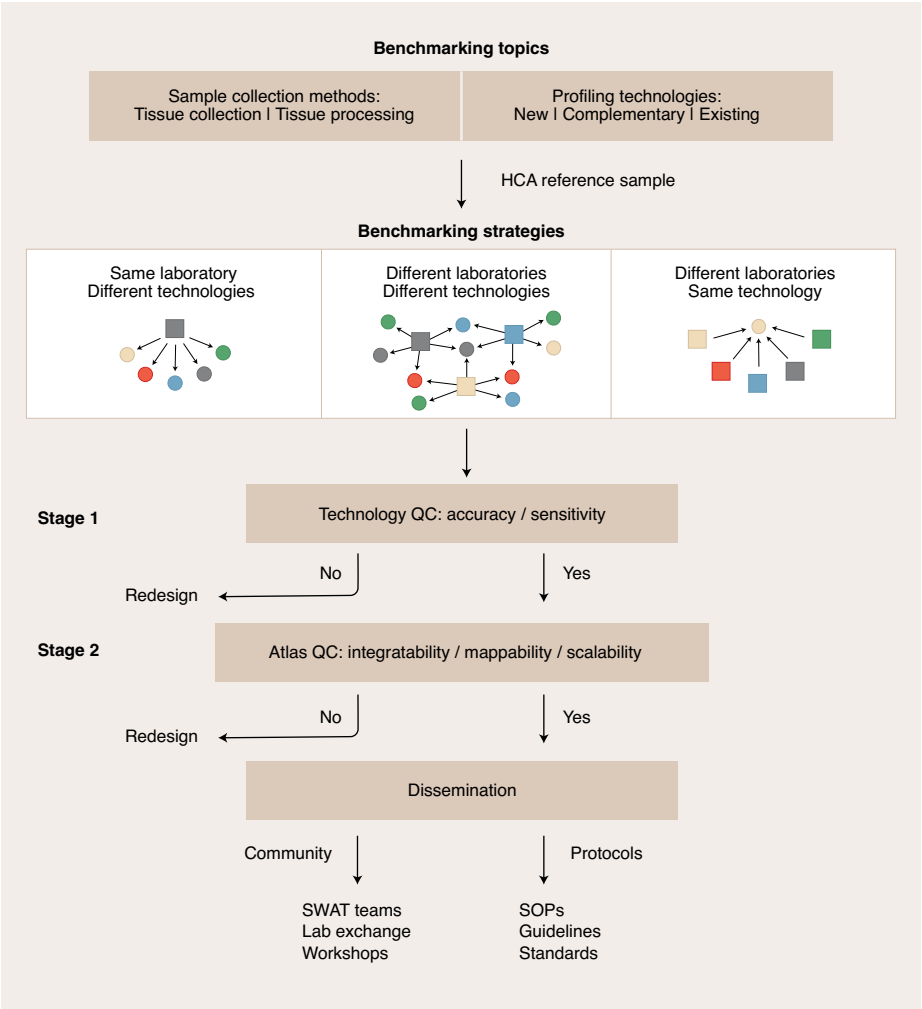

Fig. 2 | Enabling high-quality, cost-effective data generation through systematic benchmarking and decision trees. Standards and best practices are defined for sampling procedures and profiling technologies through centralized or decentralized benchmarking activities, simulating different data-generation scenarios. Protocols and technologies are evaluated on the basis of their accuracy and sensitivity to detect cell types or states (stage 1) and their suitability to be integrated with other cell atlas efforts (stage 2). Eventually, new protocols are disseminated as SOPs with related guidelines and quality control measures, combined with hands-on training activities by the community.

Validation of HCA protocols and technologies

Building the HCA will require extensive lab protocols, SOPs, benchmarks and quality-control metrics. Each technology and tissue requires careful benchmarking of protocols or validation of datasets. Benchmarking can be conducted (1) across many sites using the same technology; (2) across many sites using complementary technologies[9] and (3) at the same site using complementary technologies[10] (Fig. 2). For example, in the case of RNA applications, we should compare profiles between single-cell, single-nucleus, bulk and spatial transcriptomics methods to comprehensively identify the different cell types in tissues. Considering the diversity of tissues and questions, we argue that benchmarking experiments should aim to produce decision trees that serve to guide researchers to choose a protocol best suited to their samples and questions (Fig. 2). Such systematic testing has been performed for specific tissues[8,11] and protocols[9,10], highlighting important differences in resulting datasets, but continuous benchmarking efforts are required to broadly define applicable guidelines.

The HCA's STWG and Analysis Working Group will facilitate this process by developing broadly agreed-upon experimental and computational metrics and guidelines for these comparisons.

The STWG will also receive feedback from the HCA Biological Networks on the application of these guidelines to specific organs, tissues and systems. Below, we outline the considerations for the construction of robust and useful SOPs and benchmarking datasets for each stage of the process, including sample collection, sample processing, sample profiling and data analysis.

Sample collection. HCA labs obtain and process human specimens from healthy living donors, clinical biopsies and surgical resections on living patients, deceased transplant organ donors and rapid autopsies. It is important to maximize biospecimen quality early in the sample collection process by rapid sample processing or preservation in clinical settings[12,13] and by minimizing the post mortem interval (PMI) for deceased donor samples. We emphasize three key preanalysis quality metrics: first, pathology review with careful recording of the precise anatomical location of each specimen (ideally following a common coordinate framework allowing mapping and comparison of the sample to a reference template[14]); second, review and collection of associated donor metadata, including health and disease states; recording of sample metadata (for example, PMI measurements, freezing and/or fixation times); and third, scoring of biomolecule quality and integrity, if possible, and recording of quality control (QC) data for downstream assays (for example, viability, Bioanalyzer for scRNA-seq).

Molecular profiling of dissociated cells or nuclei. Although sc/snRNA-seq is already one of the main profiling methods in the HCA, two key challenges remain. First, each tissue type typically requires at least some optimization for successful cell dissociation or nuclei extraction. Cell dissociation depends on the cell type and extracellular matrix composition of each tissue, and its process directly impacts the atlas's quality as a result of transcriptional responses and/or RNA degradation during extended incubation[8], as well as biases in cell viability and recovery[15]. snRNA-seq instead isolates nuclei from snap-frozen or lightly fixed tissue, tackling archived (frozen)[11] and hard-to-dissociate tissues (for example, brain)[16], but different buffers, detergents and physical forces can affect the recovery of nuclei from tissues, fewer genes and transcripts are detected by snRNA-seq[17], and cell type enrichment is challenging. Both approaches recover cells with similar profiles, but sometimes at different proportions, with immune cells often more prevalent in scRNA-seq and many parenchymal cells more prevalent in snRNA-seq[8,11]. To assess such biases, we can use computational QC to determine cell composition[11] or the presence of ambient RNA[18], as well as auxiliary experimental data to determine the ground truth of cellular composition, including bulk RNA-seq (also providing a tissue-specific reference transcriptome) and spatial profiling. For example, in a lung dataset, bulk RNA-seq identified the depletion of fibroblast and endothelial cells and the enrichment of immune cell types in scRNA-seq datasets as a result of dissociation[15]. Efforts of the STWG involve the comparison of different sn/snRNA-seq modalities (3′, 5′, full-length and total RNA), multi-omics protocols, scATAC-seq, and spatial RNA and protein measures from donor-matched kidney samples. This framework can be readily extended to other tissue types in health and disease.

In situ and spatial profiling. To build an atlas, it is essential to characterize cells in their spatial context in tissues and whole organs. Benchmarking these methods, many of them not yet as broadly adopted, spans several challenges, including testing and sharing reagents — in particular, for spatial methods relying on RNA probes (for example, MERFISH or Seq-FISH) and antibodies (for example, MIBI, CODEX, or tCy-CIF); testing protocol-specific optimizations for specific tissues; testing equipment, particularly for methods that rely on highly specialized equipment that is not yet broadly available to other labs and that poses a cost barrier; and comparing to complementary methods like single-molecule FISH and immunohistochemistry of individual RNA and proteins, respectively. One key strategy is comparing different technologies on the same tissue. Given the highly specialized nature of many of these techniques, this often involves a collaborative effort whereby different expert labs apply different technologies to the same sample (Fig. 2; for example, using consecutive sections; see SpaceTx project below). In addition, applying both spatial profiling and molecular profiling of dissociated cells from the same tissue, as has been used for the atlas of the developing human heart[19], can help assess the congruence of the two methods. As spatial technologies mature, they will require systematic evaluation to ensure a high-quality dataset for HCA, and we believe that their robustness and reproducibility will continue to progress in the near future.

Aside from benchmarking, there are also key opportunities for further development through concerted and collaborative efforts. Among these are improvements in the signal-to-background ratio and the resolution for approaches based on an imaging readout (through preparation approaches like tissue clearing[20] or expansion microscopy[21]), as well as enhancing the resolution or deconvolution[22] of approaches that are not single-cell or imaging based (for example, spatial transcriptomics, Slide-Seq and HDST). There are also efforts to improve the throughput for imaging-based strategies (for example, MIBI, FISSEQ, MERFISH, SeqFISH and STARmap), which require substantial imaging or processing time.

Minimizing and addressing confounding variables in data generation

HCA data should be reproducible (recovering cells with the same profiles and features across experiments), comprehensive (capturing cell proportions and rare cells within a tissue), and of predictive value (mapping molecular profiles and spatial features to predict a new entity). Inevitable technical and biological confounders pose a barrier to achieving these objectives, but several strategies can minimize such confounders in the HCA dataset.

SOPs. Sample collection, lab protocol and organizational SOPs can all reduce technical confounders, facilitate comparison and streamline HCA operations. Sample collection SOPs provide a clear set of operation and sampling features (donor information and metadata, site, time, size and preservation) for obtaining biospecimens with reduced inter-individual variability and maximal quality, as demonstrated by other large efforts with excellent SOP collections, such as the GTEx project[23]. Lab protocol SOPs describing each pipeline from tissue type to data type will be summarized in a STWG virtual handbook linked to detailed open access workflows in protocols.io. Finally, organizational SOPs help address sample tracking systems and necessary equipment, setting labs up for success.

Reference toolkit. A toolkit will be made available including samples, reagents and computational pipelines to share across the HCA community when testing and evaluating methods across labs and new methods, further helping to minimize experimental batch effects. Banked reference samples enable comparison of protocols across the community and monitoring of performance to avoid drift. For example, the HCA reference sample used for benchmarking scRNA-seq technologies[9]

can monitor assay performance over time and is also available to extend the effort to future technologies (for example, used in Smart-seq3[24]) and modalities (for example, sc/snATAC-seq). In particular, peripheral blood monocytes (PBMCs)[10], which are commercially available in large batches and are stably frozen in aliquots, can be easily shared across labs.

Extensive metadata recording. Adding metadata on clinical, epidemiological, collection, histological and technical features allows the identification of potential factors driving batch effects and helps to correct or minimize variations in data. Metadata collection is performed for all HCA datasets in the DCP (https://data.humancellatlas.org/metadata).

Systematic QC metrics. In-process experimental QC processes and post hoc computational QC processes can help guarantee the retrieval of consistent data with appropriate quality. They also are key to detecting sample mislabeling or swapping and assessing viability, library quality and quantity metrics. Computational QC procedures that will be applied to flag low-quality samples in the DCP are developed by working closely with the Analysis Working Group.

Early sample multiplexing. Multiplexing helps reduce batch effects or randomize batches. For example, for sc/snRNA-seq, multiplexing can now be achieved by genetics[25] or by DNA-barcoded lipids[26], chemicals[27] or antibodies[28]. Although sample pooling reduces technical variance across donors, caution must be taken when selecting the labeling strategy to avoid biases in cell type composition[29].

Establishing and disseminating HCA benchmarks and standards

The STWG has already initiated guides and analytical benchmarking projects. Its first effort was to tackle the ~20 different scRNA-seq methods, through systematic comparisons with two complementary benchmark studies[9,10]. In the first approach, a complex mixture of cells (human PBMC, mouse colon and different cell lines) was sent to labs across the world to test different sc/snRNA-seq methods[9]. In a second approach, several sc/snRNA-seq technologies were tested on the same samples (PBMC, NIH3T3/HEK293 cell mix, T cells and mouse brain) in one location[10]. Both studies pointed to differences in protocol performance as evaluated by sensitivity, throughput and cost, by their power to detect genes and cell type markers, and in their successful merger into a joint atlas or projection onto a reference. The computational pipelines scumi and matchSCore2, which were developed to analyze, compare and integrate the benchmarking data of these efforts, will be helpful as new or improved methods emerge. Furthermore, the benchmarking data and remaining banked samples are an excellent resource for computational tool developers tackling batch-effect correction and data integration. Further comparisons can be made between laboratories using the same protocol to assess robustness and reproducibility.

A similar effort, SpaceTx, tackled spatial transcriptomics techniques, with labs expert in each technique analyzing the same brain sample, followed by a community-wide analysis effort through 'SpaceJam' Jamborees. As the scRNA-seq and SpaceJam efforts mature, STWG has now turned its attention to benchmarking different methods for scATAC-seq across labs and technologies, and to developing guidelines for sample handling across tissues for the HCA community[9,10] through decision trees that help researchers choose the most suitable protocol for their research goals and guidelines on how to optimize protocols (Fig. 2).

A final key role of STWG is to disseminate protocols and best practices for high-quality data production for the atlas. To ensure that technologies are disseminated and standardized across laboratories, Specialized Work Acquisition Teams (SWATs) provide personnel exchanges through short visits for hands-on in-person training, learning and troubleshooting. Moreover, in the spirit of open research, protocols and SOPs are shared using protocols.io. The STWG leverages these efforts to ensure experimental methods are applied in a coordinated fashion across HCA sites using agreed metrics.

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/s41587-020-00812-4. ❐

Orit Rozenblatt-Rosen[1,25], Jay W. Shin[2,25], Jennifer E. Rood[1], Anna Hupalowska[1], Human Cell Atlas Standards and Technology Working Group*, Aviv Regev[1,3,4] and Holger Heyn[5,6]✉

*[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Laboratory for Advanced Genomics Circuit, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan. [3]Koch Institute of Integrative Cancer Research, Cambridge, MA, USA. [4]Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, MA, USA. [5]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [6]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [25]These authors contributed equally: Orit Rozenblatt-Rosen, Jay W. Shin. *A list of authors and their affiliations appears at the end of the paper.*
✉*e-mail: holger.heyn@cnag.crg.eu*

Published online: 26 January 2021
https://doi.org/10.1038/s41587-020-00812-4

References

1. Regev, A. et al. *eLife* **6**, e27041 (2017).
2. Wilkinson, M. D. et al. *Sci. Data* **3**, 1–9 (2016).
3. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. *Nature* **550**, 451–453 (2017).
4. Angermueller, C. et al. *Nat. Methods* **13**, 229–232 (2016).
5. Macaulay, I. C. et al. *Nat. Methods* **12**, 519–522 (2015).
6. Stoeckius, M. et al. *Nat. Methods* **14**, 865–868 (2017).
7. Satpathy, A. T. et al. *Nat. Biotechnol.* **37**, 925–936 (2019).
8. Denisenko, E. et al. *Genome Biol.* https://doi.org/10.1186/s13059-020-02048-6 (2020).
9. Mereu, E. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0469-4 (2020).
10. Ding, J. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0465-8 (2020).
11. Slyper, M. et al. *Nat. Med.* **26**, 792–802 (2020).
12. Massoni-Badosa, R. et al. *Genome Biol.* **21**, 112 (2020).
13. Madissoon, E. et al. *Genome Biol.* **21**, 1 (2019).
14. Rood, J. E. et al. *Cell* **179**, 1455–1467 (2019).
15. Lambrechts, D. et al. *Nat. Med.* **24**, 1277–1289 (2018).
16. Habib, N. et al. *Science* **353**, 925–928 (2016).
17. Bakken, T. E. et al. *PLoS One* **13**, e0209648 (2018).
18. Fleming, S.J., Marioni, J.C. & Babadi, M. Preprint at *bioRxiv* https://doi.org/10.1101/791699 (2019).
19. Asp, M. et al. *Cell* **179**, 1647–1660.e19 (2019).
20. Moffitt, J. R. et al. *Proc. Natl Acad. Sci. USA* **113**, 14456–14461 (2016).
21. Chen, F., Tillberg, P. W. & Boyden, E. S. *Science* **347**, 543–548 (2015).
22. Elosua, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.03.131334 (2020).
23. GTEx Consortium. *Nat. Genet.* **45**, 580–585 (2013).
24. Hagemann-Jensen, M. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0497-0 (2020).
25. Kang, H. M. et al. *Nat. Biotechnol.* **36**, 89–94 (2018).
26. McGinnis, C. S. et al. *Nat. Methods* **16**, 619–626 (2019).
27. Gehring, J., Hwee Park, J., Chen, S., Thomson, M. & Pachter, L. *Nat. Biotechnol.* **38**, 35–38 (2020).
28. Stoeckius, M. et al. *Genome Biol.* **19**, 224 (2018).
29. McGinnis, C. S. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2020.02.12.946509 (2020).

Acknowledgements

We thank A. K. Shalek for insightful comments on the manuscript. This work has received funding from the Ministerio de Ciencia, Innovación y Universidades of Spain (SAF2017-89109-P; AEI/FEDER, UE). J.W.S. received a research grant for the RIKEN Center for Integrative Medical Sciences from MEXT. This project has been made possible in part by grant number 2018-182827 and HCA-A-1704-01742 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. Work was supported by the Manton Foundation, the Klarman Cell Observatory and HHMI (A.R.). This publication is also supported as part of a project (BCLLATLAS and ESPACE) that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant agreement No 810287 and 874710). We acknowledge support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de

Excelencia Severo Ochoa and the CERCA Programme / Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III, the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement, and co-financing by the Spanish Ministry of Ministry of Science and Innovation with funds from the European Regional Development Fund (ERDF) corresponding to the 2014–2020 Smart Growth Operating Program.

Competing interests

A.R. is a cofounder and equity holder of Celsius Therapeutics, an equity holder in Immunitas and an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Asimov and Neogene Therapeutics. A.R. is a co-inventor on patent applications to 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. O.R.R. is a co-inventor on patent applications filed by the Broad Institute for inventions relating to single-cell genomics, such as in PCT/US2018/060860 and US Provisional Application No. 62/745259.

Human Cell Atlas Standards and Technology Working Group

Kristin Ardlie[7], Menna Clatworthy[8,9], Piero Carninci[10], Wolfgang Enard[11], William Greenleaf[12], Holger Heyn[5], Edward Lein[13], Joshua Z. Levin[1,14], Sten Linnarsson[15,16], Emma Lundberg[17], Kerstin Meyer[18], Nicholas Navin[19,20], Garry Nolan[21], Aviv Regev[1,3,4], Orit Rozenblatt-Rosen[1,25], Sarah Teichmann[18], Thierry Voet[22,23] and Xiaowei Zhuang[24]

[7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [9]Molecular Immunity Unit, Department of Medicine, University of Cambridge, MRC Laboratory of Molecular Biology, Cambridge, UK. [10]Division of Genomic Medicine, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [11]Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians-University, Martinsried, Germany. [12]Department of Genetics, Stanford University, Stanford, CA, USA. [13]Allen Institute for Brain Science, Seattle, WA, USA. [14]Stanley Center, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]Division of Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [16]Science for Life Laboratory, Stockholm, Sweden. [17]Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH - Royal Institute of Technology, Stockholm, Sweden. [18]Wellcome Sanger Institute, Hinxton, UK. [19]Department of Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [20]Department of Genetics, University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [21]Baxter Laboratory in Stem Cell Biology, Department of Microbiology and Immunology, Stanford University, Stanford, CA, USA. [22]Department of Human Genetics, University of Leuven, KU Leuven, Leuven, Belgium. [23]Sanger Institute-EBI Single-Cell Genomics Centre, Wellcome Trust Sanger Institute, Hinxton, UK. [24]Howard Hughes Medical Institute, Department of Chemistry and Chemical Biology, and Department of Physics, Harvard University, Cambridge, MA, USA.