Guidelines for reporting single-cell RNA-seq experiments

To the Editor — Single-cell RNA-sequencing (scRNA-seq) has undergone major technological advances in recent years, enabling the conception of various organism-level cell atlas projects. With increasing numbers of datasets being deposited in public archives, there is a need to ensure the reproducibility of such datasets. To this end, we describe the minSe (Minimum Information about a Single-Cell Experiment) guidelines for a minimum set of metadata needed for robust comparative analyses of scRNA-seq.

scRNA-seq experiments have many advantages over so-called bulk RNA-sequencing and microarray experiments, as they allow researchers to study gene expression at an individual-cell rather than at tissue level. As the scRNA-seq technologies are maturing, these experiments are becoming increasingly high-throughput and widespread. Data from an estimated 3,000 scRNA-seq studies have been submitted to NCBI’s Gene Expression Omnibus (GEO), EMBL-EBI’s ArrayExpress and the European Nucleotide Archive (ENA), which aims to uncover the gene expression profiles of all human cell types; the Fly Cell Atlas (http://flycellatlas.org/), which has similar aims for Drosophila, or the Human Biomolecular Atlas Program (HuBMAP), as well as organ-specific projects, such as the BRAIN Initiative Cell Census Consortium.

Meta-analyses combining data from independent scRNA-seq studies have now been well established in the MINSEQE (Minimum Information about a SEQuencing Experiment) guidelines, but the emergence of protocols that can assay transcriptomics at single-cell resolution brings new requirements.

There is a clear need to establish minimum standards for reporting data and metadata for the various scRNA-seq assays. ArrayExpress and the HCA have published online guidelines for technical information required for scRNA-seq data submissions; however, these are not yet widely adopted community standards. Such standards would serve as the guiding principles and would ensure the usability of the submitted datasets, guide the adaptation of existing archival resources and enable reproducibility of analysis by the wider
scientific community. Here, we propose minSCe as a minimum set of single-cell metadata categories and a checklist of information that can be used to describe a single-cell assay in sufficient detail to enable analysis of the transcriptomic data. These guidelines are derived from work on building and adapting community resources to archive and add value to datasets, such as the Expression Atlas, the HCA-Data Coordination Platform and the CIRM Stem Cell Hub.

Typical designs of single-cell transcriptomic experiments include the following steps: (a) single cell isolation, (b) addition of spike-in RNAs, (c) reverse transcription, (d) amplification, (e) library construction, and (f) sequencing. Depending on the exact protocol followed, different types of metadata need to be recorded. Figure 1 shows an overview of the main steps that define the experimental workflows for scRNA-seq, with the variety of options used by different protocols.

For example, the Smart-seq2 protocol, which is frequently used for single-cell transcriptomics, involves cell isolation using fluorescence-activated cell sorting into microwell plates to separate the single cells. Barcodes are typically not used during reverse transcription; amplification is done by PCR and libraries covering the full length of the sequences are constructed. Alternatively, microdroplet-based protocols such as the commonly used single-cell controller from 10x Genomics use droplets to encapsulate individual cells and barcode individual molecules during reverse transcription. This method uses a 3’ or 5’ tag system during library preparation.

Taking the five main components of the MIAME and MINSEQE experiment model, we can describe a single-cell sequencing experiment with a few additions and changes (Fig. 2). Each component is equivalent to a main experimental step. We capture information describing each component and link it to the relevant protocols. Definitions of the individual components and a list of the single-cell-specific attributes that are introduced can be found in the Supplementary Information. For each field we recommend using terms from a suitable ontology (like NCBI taxonomy for species) or controlled vocabulary to prevent ambiguity. We refer to the Supplementary Information for examples of different implementations of the scRNA-seq metadata guidelines.

A single-cell attribute of particular importance is the “inferred cell type,” which is used to describe a cell’s classification based on a distinct gene expression signature. It is different from other experimental attributes as it is not known before the data analysis and is built on the results. Therefore, keeping record of reproducible data analysis steps is key to making the classification process transparent.

When depositing scRNA-seq data to a public archive, particular attention should be given to what is defined as a “sample” and how cell-specific metadata are recorded at the sample and the single-cell level, especially for methods that do not distinguish individual cells during the workflow. This may involve addition of extra metadata files with postanalysis information disaggregated by cell.

The adoption of MIAME guidelines by the scientific community, including the major scientific journals and public archives of functional genomics data, was an important step towards enabling a widespread reuse of these data and established an important precedent for developing similar standards for other technology and data types. We strongly believe that now is the time to discuss and adopt similar guidelines for scRNA-seq experiments, so that data generated in the growing number of these experiments are suitable for reuse and meta-analysis. The ArrayExpress database has already implemented a scRNA-seq data submission system that follows these guidelines. With this announcement we would like to
ask journals and other public resources accepting scRNA-seq data also to follow the guidelines, while remaining flexible as the technology is developing and community feedback is being received. As single-cell transcriptomics are increasingly combined with imaging of tissue sections or quantification of surface proteins\textsuperscript{17}, combined with imaging of tissue sections single-cell transcriptomics are increasingly community feedback is being received. As the technology is developing and reviewed as the technology is developing and reviewed.

**Editorial note:** This article has been peer reviewed.

Anja Füllgrabe\textsuperscript{1,2,3,4,5}, Nancy George\textsuperscript{1,6},
Matthew Green\textsuperscript{1,7}, Parisa Nejad\textsuperscript{1,8,9,10},
Laura Clarke\textsuperscript{1,9,10}

\textsuperscript{1}European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. \textsuperscript{2}University of California, Santa Cruz, Santa Cruz, CA, USA. \textsuperscript{3}Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA. \textsuperscript{4}J. Craig Venter Institute, Department of Pathology, University of California, San Diego, La Jolla, CA, USA. \textsuperscript{5}Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia, Philadelphia, PA, USA. \textsuperscript{6}Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. \textsuperscript{7}Oregon Health & Science University, Portland, OR, USA. \textsuperscript{8}Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. \textsuperscript{9}Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. \textsuperscript{10}Cavendish Laboratory, University of Cambridge, Cambridge, UK. \textsuperscript{11}Present address: Dotmatics Limited, Bishop’s Stortford, UK.

e-mail: anjaf@ebi.ac.uk; irene@ebi.ac.uk

Published online: 13 November 2020
https://doi.org/10.1038/s41587-020-00744-z

**References**

1. Trapnell, C. Genome Res. 25, 1491–1498 (2015).
2. Barrett, T. et al. Nucleic Acids Res. 41, D991–D995 (2013).
3. ATHAN, A. et al. Nucleic Acids Res. 47D1, D711–D713 (2019).
4. Harrison, P. W. et al. Nucleic Acids Res. 47D1, D84–D88 (2019).
5. Regev, A. et al. Preprint at arXiv https://arxiv.org/abs/1810.05192 (2018).
6. HuBMAP Consortium. Nature 574, 187–192 (2019).
7. Ecker, J. R. et al. Neuron 96, 542–557 (2017).
8. Barkas, N. et al. Nat. Methods 16, 695–698 (2019).
9. Hie, B., Bryan, B. & Berger, B. Nat. Biotechnol. 37, 685–691 (2019).
10. Braiman, A. et al. Nat. Genet. 29, 365–371 (2001).

**Correspondence**

**Jim Kent**\textsuperscript{2}, **John Marioni** and **Irene Papatheodorou**

**Richard H. Scheuermann**

Matthew Green\textsuperscript{1, Parisa Nejad}
Laura Clarke

To the Editor — I read your recent editorial entitled “Thank you for sharing” with appreciation for its spotlight on materials collections. As executive director of Addgene — a non-profit repository created to help scientists share plasmids — I believe that the impact of materials sharing on science acceleration and reproducibility is enormous, particularly at this time of vital research and product development to rapidly discover diagnostics, vaccines and treatments for COVID-19.

The most expensive and demanding part of fulfilling Addgene’s mission is to solicit deposits to keep the collection current and growing. It would be invaluable to the community if journals and granting agencies put more emphasis on materials sharing. For example, although *Nature Biotechnology* and other *Nature* journals instruct authors that they are “required to make unique materials promptly available to others without undue qualifications” and “strongly encourage” this, it is not enforced and especially not so for plasmids (which are free to deposit at Addgene).

It would be a public good for all research if editors (and grant officers) had a checklist for authors to complete, confirming they have initiated deposit of all materials (including plasmids) associated with a paper. In addition, I would advocate that journals include links to all materials and reagents used in the research described in their publications and that journals have clear minimal requirements (rather than recommendations) for materials deposition.

We wish ’strong encouragement’ was enough, but alas, once the paper is out the door, authors often need more than encouragement to take the time to execute on materials sharing. An extra push could be a game changer.

**Joanne Kamens**\textsuperscript{1,2}

Addgene, Watertown, MA, USA.

e-mail: joanne.kamens@addgene.org

Published online: 30 November 2020
https://doi.org/10.1038/s41587-020-00752-z

**References**

1. Anonymous. Nat. Biotechnol. 38, 1005 (2020).

**Competing interests**

Joanne Kamens is executive director of Addgene and is an advisor to Protocol.io.

**Acknowledgements**

This work was supported by the European Molecular Biology Laboratory; the Wellcome Trust Biomedical Resources grant Single Cell Gene Expression Atlas (108437/Z/15/Z); the California Institute for Regenerative Medicine (GCIR-06673-B); the Chan Zuckerberg Initiative DAF; an advised fund of the Silicon Valley Community Foundation (2018–182730); and the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director (OT2OD026677).

**Author contributions**

A.F., N. George, M.G., P.N., B.A., S.K.F., C.F., M.A.F., L.H., D.T., N.V. and N.M. curated data and developed metadata standards; L.C., J.K., N.M., R.H.S., A.B. and I.P. supervised the project and engaged with the larger community; S.T. and J.M. provided support and evaluation of the standards and contributed to the manuscript; A.F., N. George, M.G., P.N., B.A., S.K.F., C.F., M.A.F., L.C., N. Gehlenborg, L.H., R.H.S., D.T., N.V., A.B. and I.P. wrote the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00744-z.