bioRxiv preprint doi: https://doi.org/10.1101/742304; this version posted August 21, 2019. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

A curated database reveals trends in single cell transcriptomics

Valentine Svensson[1,*] & Eduardo da Veiga Beltrame[1]

[1]Division of Biology and Biological Engineering, California Institute of Technology
*Address correspondence to Valentine Svensson (vale@caltech.edu)

Over 500 single cell transcriptomics studies have been published to date. Many of these have data available, but the links between data, study, and systems studied can be hard to identify through literature search. This manuscript describes a nearly exhaustive and manually curated database of single cell transcriptomics studies with descriptions of what kind of data and what biological systems have been studied. Additionally, based on the text in the listed papers information about analysis is included in the database, allowing analysis of trends in the field. As a particular example, it is demonstrated that the number of cell types identified in single cell RNA sequencing studies is directly proportional to the number of cells analyzed. Instructions to access the database are available at www.nxn.se/single-cell-studies/.

Introduction

It has been recognized that the ability to perform large scale single cell transcriptomics is opening the door to unseen views into biological variation (Klein and Treutlein 2019). This new kind of big data in biology - a large set of measurements of a large number of cells - can yield insights even after several passes of analysis of individual datasets. With hundreds of datasets available, integration of datasets becomes another avenue for exploration, highlighting the importance of standardization in how data is collected and shared, as well as curation of public data (Stuart *et al.* 2019).

As single cell transcriptomics studies become more accessible to many labs, discoverability of studies and datasets becomes a challenge, and several efforts have arisen to curate datasets. The *Human Cell Atlas* portal aims to provide uniformly processed data from all of the human body (Regev *et al.* 2017). *JingleBells* provides data, with a focus on immune cells (Ner-Gaon *et al.* 2017). The *conquer* database provides uniformly processed expression data for the sake of fair comparison of computational tools (Soneson and Robinson 2018). The *PanglaoDB* database provides count matrices from public sequencing data in the national center for biotechnology information (NCBI) sequence read archive (SRA) (Franzén, Gan, and Björkegren 2019). The *EMBL-EBI Single Cell Expression Atlas* provides uniformly processed data from data submitted to ArrayExpress. The Broad Institute offers a *Single Cell Portal* which can be used to share custom single cell RNA sequencing (scRNA-seq) data. A database called *scRNASeqDB* provides links to a number of datasets from human scRNA-seq experiments (Cao *et al.* 2017). These efforts all aim to tackle different aspects of the considerable challenge of data management in the era of big biology.

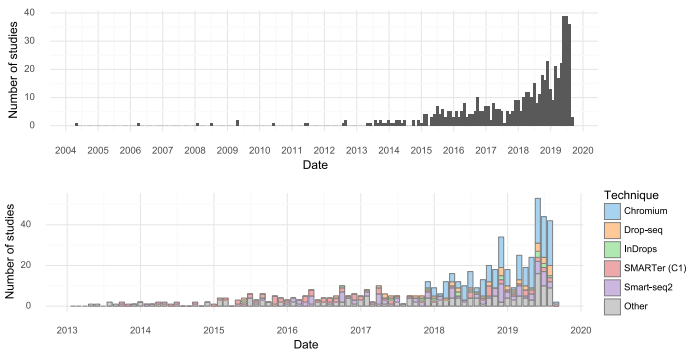

Figure 1) Studies over time.
(**upper**) The number of single cell transcriptomics studies published per month. (**lower**) The number of scRNA-seq studies published per month stratified by method.

Here we present a manually curated database of single cell transcriptomics studies rather than primary data, indexed according to publication and study authors. This resource will allow researchers to identify studies of particular tissues, together with which tissues have not been studied previously. It also aims to facilitate the citation of appropriate references when performing follow-up experiments. This database tracks metadata applicable to most studies, such as the number of cell types identified, or which protocols were used. These annotations enable analysis of trends in the field.

Database structure

This database aims to provide a quick listing between datasets from different organs, the data location, and a citation, to make published data and results discoverable. A secondary goal is to annotate metadata about these primary studies directly which can be used to spot trends in the field.

The "*Single cell studies database*" considers the analysis of many genes at once in single cells as a "single cell transcriptomics" study. There is some ambiguity were choices had to be made. For example, multicolor fluorescence flow cytometry or mass cytometry are not considered, even though both technologies can measure dozens of analytes per cell. The main focus is on datasets where over a hundred genes are measured. Some targeted technologies measuring fewer genes such as osmFISH are also included when they are directly related to the higher throughput versions (Codeluppi *et al.* 2018; Shah *et al.* 2016; Wang *et al.* 2018).

The primary identifier of an entry is the **DOI** (digital object identifier) of a publication. Based on the DOI four entries are added using the *CrossRef* API: **Authors**, **Journal**, **Title** and **Date**. Additional fields are based on the contents of the publication and are manually annotated by investigating the text and supplement of the original publication. If the study was deposited to the bioRxiv, the **bioRxiv DOI** field gives the DOI of this, with a "-" indicating a study was not submitted to bioRxiv.

Reported cells total is the number of cells investigated in the study.

Technique is the kind of technology or protocol used to obtain the single cell gene expressions.

Panel size annotates the number of genes investigated for targeted technologies such as microarrays or multiplexed smFISH.

Measurement annotates the type of quantitative measurements, which is most cases is "RNA-seq" but can also be "In Situ" or "Microarray".

bioRxiv preprint doi: https://doi.org/10.1101/742304; this version posted August 21, 2019. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

Month	Studies	Median cells
Jan 2019	9	3,368
Feb 2019	21	11,175
Mar 2019	16	11,452
Apr 2019	21	17,725
May 2019	39	14,585
Jun 2019	39	15,000
Jul 2019	36	13,966

Tissue	Studies
Brain	64
Culture	47
Blood	16
Heart	16
Pancreas	16
Embryo	14
Lung	12

Journal	Studies
bioRxiv	63
Nature	50
Cell	49
Cell Reports	35
Science	34
Nature Communications	29
Genome Biology	19

Table 1) Single cell study trends.
(**left**) Number and size of single cell transcriptomics studies in 2019. (**middle**) Most common tissue investigated with single cell transcriptomics. ('Culture' refers to *in vitro* studies of cell lines). (**right**) Journals which have published most single cell transcriptomics studies. ('bioRxiv' means the study is so far only available on bioRxiv).

Data location provides the accession ID for the repository where the original data can be found, providing a quick reference for downloading and reanalyzing the data. A number of fields indicate what system the paper is studying.

Organism lists the species included in the study.

Tissue describes which organs single cells were collected from.

Cell Source entry provides brief notes about the cells in the study and allows straightforward searches for specific kinds of cells or sub-tissues.

Contrasts describes different experimental conditions studied, if any

Isolation describes the method used to produce the single cell suspension.

Developmental stage describes the developmental stages or ages of the animals or humans the cells were collected, as applicable.

Additionally, whether some types of analysis was performed in a paper is annotated as a "Yes" or "No" entry. It is indicated whether the paper did:

Cell clustering: Whether a study performed unsupervised clustering of cells.

Pseudotime: Whether a study investigated cellular trajectories with pseudotime methods.

RNA velocity: Whether a study investigated a vector field of cells through RNA velocity (La Manno *et al.* 2018).

PCA: Whether a study performed principal component analysis.

tSNE: Whether t-Distributed Stochastic Neighbor Embedding was used for visualization (Van der Maaten and Hinton 2008).

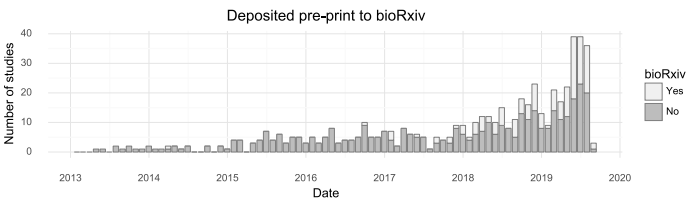

Figure 3) Pre-print usage over time.
The number of studies published in a given month stratified by whether they at some point were deposited to bioRxiv. (Including studies currently only available on bioRxiv).

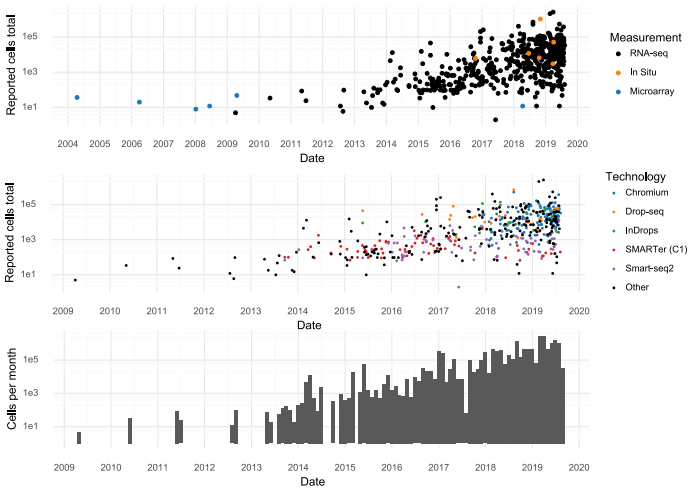

Figure 2) Scale of experiments and data over time.
(**Upper**): The number of cells measured in a study, stratified by the measurement method. (**Middle**): The number of cells measured in scRNA-seq experiments, stratified by scRNA-seq protocol. (**Lower**): The aggregate number of cells measured per month.

Finally, the number of cell types or clusters identified in the studies is annotated under **Number of reported cell types or clusters**. This is most commonly based on de novo clustering, but in some cases the number of different pre-sorted cell types.

By virtue of relying on manual curation which provide detailed and accurate annotation this database is incomplete, but is substantial enough to serve as a good starting point for a community effort to fill the gaps. Even with some missing annotations, the data available allows analysis of trends in the field.

The database can be accessed as a graphical interface through Google Sheets at www.nxn.se/single-cell-studies/gui. This view allows searching for keywords and browsing studies. Importantly, it also allows the community to contribute information to the database through comments on the individual entries of the database.

A version of the database in TSV (tab separated values) format can be downloaded from www.nxn.se/single-cell-studies/data.tsv. This enables researchers to do advanced analysis of the data.

Additional studies can be submitted through a form available at www.nxn.se/single-cell-studies/submit. Submissions require a DOI, which is the primary identifier for an entry in the database. If more information is known about the study, they can be reported through the optional fields. This facilitates annotation and addition to the database. Claims in the submissions are spot checked to be referring to the original text in the publication.

A snapshot of the database is saved (in TSV format) daily, and all snapshots are available in a public Google Storage bucket at `gs://single-cell-studies`, which can be accessed with `gsutils`. An example snapshot is provided as Supplementary Table 1, which has data on 555 studies published between 2003 and August 17 2019.

Results

The earliest single cell transcriptomics study annotated was published in 2004. Since 2013 almost every month at least one study has been published per month. The rate of studies have increased steadily, and in May, June, and July of 2019 there were over 30 single cell transcriptomics studies published per month (**Figure 1**). In 2019 the median scRNA-seq study investigates 14,000 cells (**Table 1**).

Individual studies have increased in scale, and every few months a new study is released which breaks the previous record in number of cells. During the first half of 2019 about 200,000 cells were added to the pool of public data every month (**Figure 2**).

bioRxiv preprint doi: https://doi.org/10.1101/742304; this version posted August 21, 2019. The copyright holder has placed this preprint (which was not certified by peer review) in the Public Domain. It is no longer restricted by copyright. Anyone can legally share, reuse, remix, or adapt this material for any purpose without crediting the original authors.

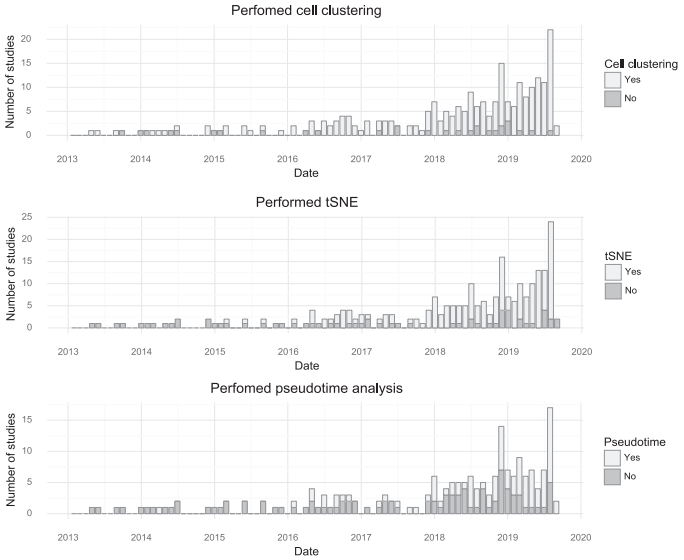

Figure 4) Popularity of forms of analysis over time.
(**Top**) The number of studies doing clustering per month. (**Middle**) The number of studies using tSNE per month. (**Bottom**) The number of studies doing pseudotime analysis per month.

Many tissues have been investigated by single cell transcriptomics, but the brain is the most popular with 65 associated citations out of 550. Another trend observed from this database is that authors of single cell transcriptomics papers are increasingly making use of bioRxiv. In total 145 of 555 studies were deposited to bioRxiv (26 %). The fraction is now about 41% in a given month **(Figure 3)**. Single cell studies are published in many different journals, with Nature and Cell having published the most. The increasing popularity of these kinds of studies means the field has grown, with 5,823 unique authors of single cell transcriptomics studies.

By tracking what forms of analysis is performed on single cell transcriptomics data it is possible to see what the community is aiming to learn from the assays. The most common application is to survey molecular "cell types" by clustering cells based on gene expression. Almost every study performs clustering at some point (87%). An interesting case is the use of tSNE which allows researchers to visualize which cells are in the same cluster. After it was used for the first time in 2015 it became a near universal visualization technique. The fraction of studies per month using it has decreased slightly in the last year, perhaps due to the introduction of UMAP (McInnes and Healy 2018). Analysis of "pseudotime" is less common but still very popular, with about half of published studies investigating pseudotime trajectories **(Figure 4)**.

Since de novo clustering and cell type discovery is a nearly universal single cell transcriptomics analysis, the number of clusters of cell types identified in the studies was annotated. This revealed a clear relation between the number of cell types identified and the total number of cells investigated. For small to medium sized studies on average one cell type is identified per 100 cells studied. For massive studies with hundreds of thousands of cells, the rate is closer to one cell type per 1,000 cells investigated **(Figure 5)**.

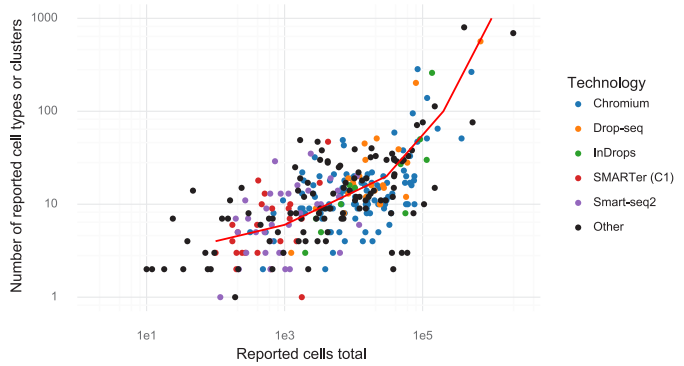

Figure 5) Cluster and cell numbers.
The number of cells studied vs the number of clusters or cell types reported in a study.

Discussion

The curated database described here is hosted at www.nxn.se/single-cell-studies/. It has been designed for easy access to the underlying data for in depth analysis in Python or R. The focus of the database is to expose researchers to published papers, so that for example a researcher can find all single cell studies of pancreas and explore the results and perhaps reanalyze public data. By also tracking other aspects of the studies mentioned in the papers, such as protocol, number of cells, or the number of clusters identified, trends in the field can be revealed. As an example, it was shown here that the vast majority of studies perform clustering, and in general the number of clusters identified is directly proportional to the number of cells analysed.

A notebook with the analsysis and generation of the figures here is available on GitHub as a Jyputer Notebook: https://github.com/vals/single-cell-studies.

The database is also designed to be expanded by the community suggesting additions to it be leaving comments at www.nxn.se/single-cell-studies/gui, and by adding data through the submission form at www.nxn.se/single-cell-studies/submit. The analysis notebook and data snapshot have also been deposited to CaltechDATA with accession 10.22002/D1.1267.

Acknowledgements

We thank Carlos Talavera-López for helpful feedback on the manuscript. Cloud infrastructure was funded through the *Google Cloud Platform research credits program*.

References

Cao, Yuan, Junjie Zhu, Guangchun Han, Peilin Jia, and Zhongming Zhao. 2017. "**scRNASeqDB: A Database for Gene Expression Profiling in Human Single Cell by RNA-Seq**." bioRxiv. https://doi.org/10.1101/104810.

Codeluppi, Simone, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. 2018. "**Spatial Organization of the Somatosensory Cortex Revealed by osmFISH**." Nature Methods 15 (11): 932–35.

Franzén, Oscar, Li-Ming Gan, and Johan L. M. Björkegren. 2019. "**PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data**." Database: The Journal of Biological Databases and Curation 2019 (January). https://doi.org/10.1093/database/baz046.

Klein, Allon M., and Barbara Treutlein. 2019. "**Single Cell Analyses of Development in the Modern Era**." Development 146 (12). https://doi.org/10.1242/dev.181396.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. "**RNA Velocity of Single Cells**." Nature 560 (7719): 494–98.

McInnes, Leland, and John Healy. 2018. "**UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**." arXiv [stat.ML]. arXiv. http://arxiv.org/abs/1802.03426.

Ner-Gaon, Hadas, Ariel Melchior, Nili Golan, Yael Ben-Haim, and Tal Shay. 2017. "**JingleBells: A Repository of Immune-Related Single-Cell RNA-Sequencing Datasets**." Journal of Immunology 198 (9): 3375–79.

Regev, Aviv, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. "**The Human Cell Atlas**." eLife 6 (December). https://doi.org/10.7554/eLife.27041.

Shah, Sheel, Eric Lubeck, Wen Zhou, and Long Cai. 2016. "**In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus**." Neuron 92 (2): 342–57.

Soneson, Charlotte, and Mark D. Robinson. 2018. "**Bias, Robustness and Scalability in Single-Cell Differential Expression Analysis**." Nature Methods, February. https://doi.org/10.1038/nmeth.4612.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. "**Comprehensive Integration of Single-Cell Data**." Cell 177 (7): 1888–1902. e21.

Van der Maaten, Laurens, and Geoffrey Hinton. 2008. "**Visualizing Data Using T-SNE**." Journal of Machine Learning Research: JMLR 9 (2579-2605): 85.

Wang, Xiao, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, et al. 2018. "**Three-Dimensional Intact-Tissue Sequencing of Single-Cell Transcriptional States**." Science 361 (6400). https://doi.org/10.1126/science.aat5691.