University Library



A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Zappia, L;Phipson, B;Oshlack, A

Title:
Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database

Date:
2018-06-01

Citation:
Zappia, L., Phipson, B. & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. Plos Computational Biology, 14 (6), https://doi.org/10.1371/journal.pcbi.1006245.

Persistent Link:
https://hdl.handle.net/11343/270817

License:
CC BY

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database

Luke Zappia[1,2], Belinda Phipson[1], Alicia Oshlack[1,2]*

1 Bioinformatics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia, **2** School of Biosciences, Faculty of Science, University of Melbourne, Melbourne, Victoria, Australia

* alicia.oshlack@mcri.edu.au

Check for updates

OPEN ACCESS

Citation: Zappia L, Phipson B, Oshlack A (2018) Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. PLoS Comput Biol 14(6): e1006245. https://doi.org/10.1371/journal.pcbi.1006245

Editor: Dina Schneidman, Hebrew University of Jerusalem, ISRAEL

Received: December 6, 2017

Accepted: May 30, 2018

Published: June 25, 2018

Copyright: © 2018 Zappia et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Luke Zappia is supported by an Australian Government Research Training Program (RTP) Scholarship. Alicia Oshlack is supported through a National Health and Medical Research Council Career Development Fellowship APP1126157. MCRI is supported by the Victorian Government's Operational Infrastructure Support Program. The funders had no role in study design,

Abstract

As single-cell RNA-sequencing (scRNA-seq) datasets have become more widespread the number of tools designed to analyse these data has dramatically increased. Navigating the vast sea of tools now available is becoming increasingly challenging for researchers. In order to better facilitate selection of appropriate analysis tools we have created the scRNA-tools database (www.scRNA-tools.org) to catalogue and curate analysis tools as they become available. Our database collects a range of information on each scRNA-seq analysis tool and categorises them according to the analysis tasks they perform. Exploration of this database gives insights into the areas of rapid development of analysis methods for scRNA-seq data. We see that many tools perform tasks specific to scRNA-seq analysis, particularly clustering and ordering of cells. We also find that the scRNA-seq community embraces an open-source and open-science approach, with most tools available under open-source licenses and preprints being extensively used as a means to describe methods. The scRNA-tools database provides a valuable resource for researchers embarking on scRNA-seq analysis and records the growth of the field over time.

Author summary

In recent years single-cell RNA-sequencing technologies have emerged that allow scientists to measure the activity of genes in thousands of individual cells simultaneously. This means we can start to look at what each cell in a sample is doing instead of considering an average across all cells in a sample, as was the case with older technologies. However, while access to this kind of data presents a wealth of opportunities it comes with a new set of challenges. Researchers across the world have developed new methods and software tools to make the most of these datasets but the field is moving at such a rapid pace it is difficult to keep up with what is currently available. To make this easier we have developed the scRNA-tools database and website (www.scRNA-tools.org). Our database catalogues analysis tools, recording the tasks they can be used for, where they can be downloaded from and the publications that describe how they work. By looking at this database we can see that developers have focused on methods specific to single-cell data and that they

data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare that no competing interests exist.

embrace an open-source approach with permissive licensing, sharing of code and release of preprint publications.

This is a *PLOS Computational Biology* Software paper.

Introduction

Single-cell RNA-sequencing (scRNA-seq) has rapidly gained traction as an effective tool for interrogating the transcriptome at the resolution of individual cells. Since the first protocols were published in 2009 [1] the number of cells profiled in individual scRNA-seq experiments has increased exponentially, outstripping Moore's Law [2]. This new kind of transcriptomic data brings a demand for new analysis methods. Not only is the scale of scRNA-seq datasets much greater than that of bulk experiments but there are also a variety of challenges unique to the single-cell context [3]. Specifically, scRNA-seq data is extremely sparse (there is no expression measured for many genes in most cells), it can have technical artefacts such as low-quality cells or differences between sequencing batches and the scientific questions of interest are often different to those asked of bulk RNA-seq datasets. For example many bulk RNA-seq datasets are generated to discover differentially expressed genes through a designed experiment while many scRNA-seq experiments aim to identify or classify cell types in complex tissues.

The bioinformatics community has embraced this new type of data at an astonishing rate, designing a plethora of methods for the analysis of scRNA-seq data. Keeping up with the current state of scRNA-seq analysis is now a significant challenge as the field is presented with a huge number of choices for analysing a dataset. Since September 2016 we have collated and categorised scRNA-seq analysis tools as they have become available. This database is being continually updated and is publicly available at www.scRNA-tools.org. In order to help researchers navigate the vast ocean of analysis tools we categorise tools in the database in the context of the typical phases of an scRNA-seq analysis. Through the analysis of this database we show trends in not only the analysis applications these methods address but how they are published and licensed, and the platforms they use. Based on this database we gain insight into the current state of current tools in this rapidly developing field.

Design and implementation

Database

The scRNA-tools database contains information on software tools specifically designed for the analysis of scRNA-seq data. For a tool to be eligible for inclusion in the database it must be available for download and public use. This can be from a software package repository (such as Bioconductor [4], CRAN or PyPI), a code sharing website such as GitHub or directly from a private website. When new tools come to our attention they are added to the scRNA-tools database. DOIs and publication dates are recorded for any associated publications. As preprints may be frequently updated they are marked as a preprint instead of recording a date. The platform used to build the tool, links to code repositories, associated licenses and a short description are also recorded. Each tool is categorised according to the analysis tasks it can perform, receiving a true or false for each category based on what is described in the accompanying paper or

documentation. We also record the date that each entry was added to the database and the date that it was last updated. Most tools are added after a preprint or publication becomes available but some have been added after being mentioned on social media or in similar collections such as Sean Davis' awesome-single-cell page (https://github.com/seandavi/awesome-single-cell).

Website

To build the website we start with the table described above as a CSV file which is processed using an R script. The lists of packages available in the CRAN, Bioconductor, PyPI and Anaconda software repositories are downloaded and matched with tools in the database. For tools with associated publications the number of citations they have received is retrieved from the Crossref database (www.crossref.org) using the rcrossref package (v0.8.0) [5]. We also make use of the aRxiv package (v0.5.16) [6] to retrieve information about arXiv preprints. JSON files describing the complete table, tools and categories are produced and used to populate the website.

The website consists of three main pages. The home page shows an interactive table with the ability to sort, filter and download the database. The second page shows an entry for each tool, giving the description, details of publications, details of the software code and license and the associated software categories. Badges are added to tools to provide clearly visible details of any associated software or GitHub repositories. The final page describes the categories, providing easy access to the tools associated with them. Both the tools and categories pages can be sorted in a variety of ways, including by the number of associated publications or citations. An additional page shows a live and up-to-date version of some of the analysis presented here with visualisations produced using ggplot2 (v2.2.1.9000) [7] and plotly (v4.7.1) [8]. We welcome contributions to the database from the wider community via submitting an issue to the project GitHub page (https://github.com/Oshlack/scRNA-tools) or by filling in the submission form on the scRNA-tools website.

Analysis

The most recent version of the scRNA-tools database as of 6 June 2018 was used for the analysis presented in this paper. Data was manipulated in R (v3.5.0) using the dplyr package (v0.7.5) [9] and plots produced using the ggplot2 (v2.2.1.9000) and cowplot (v0.9.2) [10] packages.

Results

Overview of the scRNA-tools database

When the database was first constructed it contained 70 scRNA-seq analysis tools representing the majority of work in the field during the three years from the publication of SAMstrt [11] in November 2013 up to September 2016. In the time since then over 160 new tools have been added (Fig 1A). The almost tripling of the number of available tools in such a short time demonstrates the booming interest in scRNA-seq and its maturation from a technique requiring custom-built equipment with specialised protocols to a commercially available product.

Publication status. Most tools have been added to the scRNA-tools database after coming to our attention in a publication or preprint describing their method and use. Of all the tools in the database about half have at least one publication in a peer-reviewed journal and another third are described in preprint articles, typically on the bioRxiv preprint server (Fig 1B). Tools can be split into those that were available when the database was created and those that have been added since. We can see that the majority of older tools have been published while more recent tools are more likely to only be available as preprints (Fig 1C). This is a good

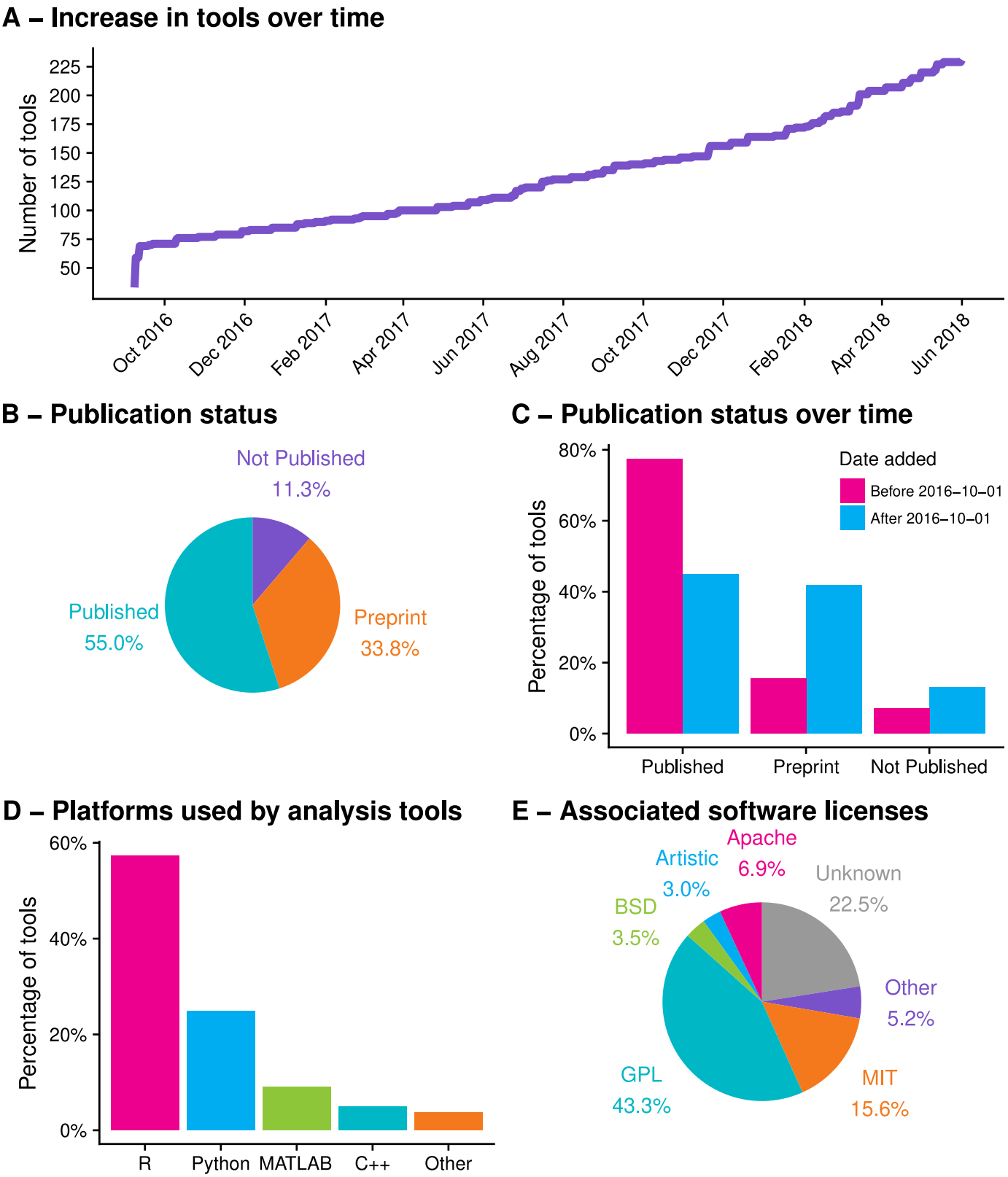


Fig 1. (A) Number of tools in the scRNA-tools database over time. Since the scRNA-seq tools database was started in September 2016 more than 160 new tools have been released. (B) Publication status of tools in the scRNA-tools database. Over half of the tools in the full database have at least one published peer-revirew paper while another third are described in preprints. (C) When stratified by the date tools were added to the database we see that the majority of tools added before

October 2016 are published, while around half of newer tools are available only as preprints. Newer tools are also more likely to be unpublished in any form. (D) The majority of tools are available using either the R or Python programming languages. (E) Most tools are released under a standard open-source software license, with variants of the GNU Public License (GPL) being the most common. However licenses could not be found for a large proportion of tools. Up-to-date versions of these plots (with the exception of C) are available on the analysis page of the scRNA-tools website (https://www.scrna-tools.org/analysis).

https://doi.org/10.1371/journal.pcbi.1006245.g001

demonstration of the delay imposed by the traditional publication process. By publishing preprints and releasing software via repositories such as GitHub, scRNA-seq tool developers make their tools available to the community much earlier, allowing them to be used for analysis and their methods improved prior to formal publication [12].

Platforms and licensing. Developers of scRNA-seq analysis tools have choices to make about what platforms they use to create their tools, how they make them available to the community and whether they share the source code. We find that the most commonly used platform for creating scRNA-seq analysis tools is the R statistical programming language, with many tools made available through the Bioconductor or CRAN repositories (Fig 1D). Python is the second most popular language, followed by MATLAB, a proprietary programming language, and the lower-level C++. The use of R and Python is consistent with their popularity across a range of data science fields. In particular the popularity of R reflects its history as the language of choice for the analysis of bulk RNA-seq datasets and a range of other biological data types.

The majority of tools in the scRNA-tools database have been developed with an open-source approach, making their code available under permissive software licenses (Fig 1E). We feel this reflects the general underlying sentiment and willingness of the bioinformatics community to share and build upon the work of others. Variations of the GNU Public License (GPL) are the most common, covering almost half of tools. This license allows free use, modification and distribution of source code, but also has a "copyleft" nature which requires any derivatives to disclose their source code and use the same license. The MIT license is the second most popular which also allows use of code for any purpose but without any restrictions on distribution or licensing. The appropriate license could not be identified for almost a quarter of tools. This is problematic as fellow developers must assume that source code cannot be reused, potentially limiting the usefulness of the methods in those tools. We strongly encourage tool developers to clearly display their license in source code and documentation to provide certainty to the community as to any restrictions on the use of their work.

Categories of scRNA-seq analysis

Single-cell RNA-sequencing is often used to explore complex mixtures of cell types in an unsupervised manner. As has been described in previous reviews a standard scRNA-seq analysis in this setting consists of several tasks which can be completed using various tools [13–17]. In the scRNA-tools database we categorise tools based on the analysis tasks they perform. Here we group these tasks into four broad phases of analysis: data acquisition, data cleaning, cell assignment and gene identification (Fig 2). The data acquisition phase (Phase 1) takes the raw nucleotide sequences from the sequencing experiment and returns a matrix describing the expression of each gene in each cell. This phase consists of tasks common to bulk RNA-seq experiments, such as alignment to a reference genome or transcriptome and quantification of expression, but is often extended to handle Unique Molecular Identifiers (UMIs) [18]. Once an expression matrix has been obtained it is vital to make sure the resulting data is of high enough quality. In the data cleaning phase (Phase 2) quality control of cells is performed as well as filtering of uninformative genes. Additional tasks may be performed to normalise the

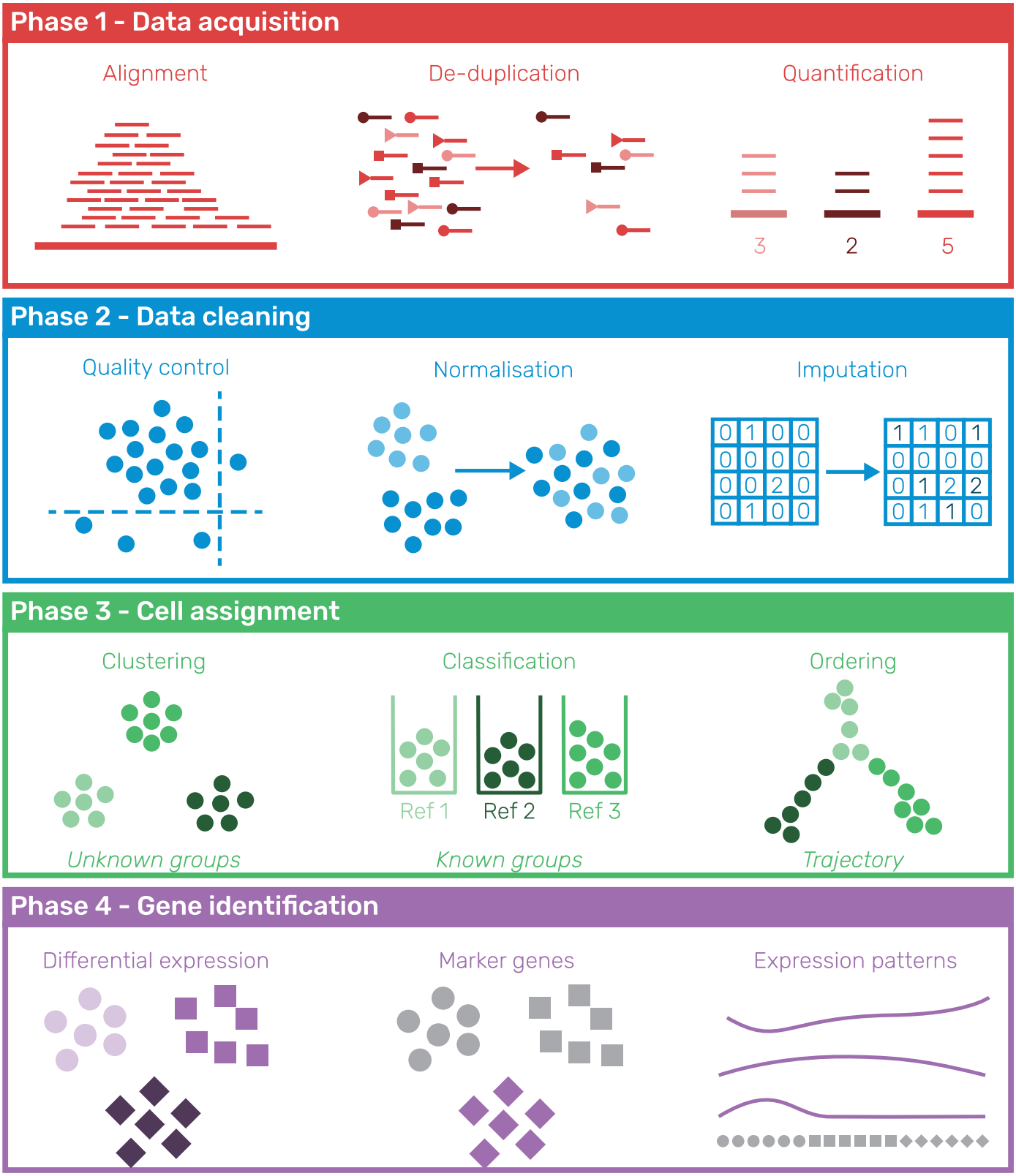


Fig 2. Phases of a typical unsupervised scRNA-seq analysis process. In Phase 1 (data acquisition) raw sequencing reads are converted into a gene by cell expression matrix. For many protocols this requires the alignment of genes to a reference genome and the assignment and de-duplication of Unique Molecular Identifiers (UMIs). The data is then cleaned (Phase 2) to remove low-quality cells and uninformative genes, resulting in a high-quality dataset for further analysis.

The data can also be normalised and missing values imputed during this phase. Phase 3 assigns cells, either in a discrete manner to known (classification) or unknown (clustering) groups or to a position on a continuous trajectory. Interesting genes (eg. differentially expressed, markers, specific patterns of expression) are then identified to explain these groups or trajectories (Phase 4).

https://doi.org/10.1371/journal.pcbi.1006245.g002

data or impute missing values. Exploratory data analysis tasks are often performed in this phase, such as viewing the datasets in reduced dimensions to look for underlying structure.

The high-quality expression matrix is the focus of the next phases of analysis. In Phase 3 cells are assigned, either to discrete groups via clustering or along a continuous trajectory from one cell type to another. As high-quality reference datasets become available it will also become feasible to classify cells directly into different cell types. Once cells have been assigned the focus of analysis turns to interpreting what those assignments mean. Identifying interesting genes (Phase 4), such as those that are differentially expressed across groups, marker genes expressed in a single group or genes that change expression along a trajectory, is the typical way to do this. The biological significance of those genes can then be interpreted to give meaning to the experiment, either by investigating the genes themselves or by getting a higher-level view through techniques such as gene set testing.

While there are other approaches that could be taken to analyse scRNA-seq data these phases represent the most common path from raw sequencing reads to biological insight applicable to many studies. An exception to this may be experiments designed to test a specific hypothesis where cell populations may have been sorted or the interest lies in differences between experimental conditions rather than cell types. In this case Phase 3 may not be required, and slightly different tools or approaches may be used, but many of the same challenges will apply. In addition, as the field expands and develops it is likely that data will be used in new ways to answer other biological questions, requiring new analysis techniques. Descriptions of the categories in the scRNA-tools database are given in Table 1, along with the associated analysis phases.

Trends in scRNA-seq analysis tasks. Each of the tools in the database is assigned to one or more analysis categories. We investigated these categories in further detail to give insight into the trends in scRNA-seq analysis. Fig 3A shows the frequency of tools performing each of the analysis tasks. Visualisation is the most commonly included task and is important across all stages of analysis for exploring and displaying data and results. Tasks for assigning cells (ordering and clustering) are the next most common. This has been the biggest area of development in single-cell analysis with clustering tools such as Seurat [19,20], SC3 [21] and BackSPIN [22] being used to identify cell types in a sample and trajectory analysis tools (for example Monocle [23–25], Wishbone [26] and DPT [27]) being used to investigate how genes change across developmental processes. These areas reflect the new opportunities for analysis provided by single-cell data that are not possible with bulk RNA-seq experiments.

Dimensionality reduction is also a common task and has applications in visualisation (via techniques such as t-SNE [28]), quality control and as a starting point for analysis. Testing for differential expression (DE) is perhaps the most common analysis performed on bulk RNA-seq datasets and it is also commonly applied by many scRNA-seq analysis tools, typically to identify genes that are different in one group of cells compared to the rest. However it should be noted that the DE testing applied by scRNA-seq tools is often not as sophisticated as the rigorous statistical frameworks of tools developed for bulk RNA-seq such as edgeR [29,30], DESeq2 [31] and limma [32], often using simple statistical tests such as the likelihood ratio test. While methods designed to test DE specifically in single-cell datasets do exist (such as SCDE [33], and scDD [34]) it is still unclear whether they improve on methods that have been established for bulk data [35–37], with the most comprehensive comparison to date finding

Table 1. Descriptions of categories for tools in the scRNA-tools database.

Phase	Category	Description
Phase 1	Alignment	Alignment of sequencing reads to a reference
Phase 1	Assembly	Tools that perform assembly of scRNA-seq reads
Phase 1	UMIs	Processing of Unique Molecular Identifiers
Phase 1	Quantification	Quantification of expression from reads
Phase 2	Quality Control	Removal of low-quality cells
Phase 2	Gene Filtering	Removal of lowly expressed or otherwise uninformative genes
Phase 2	Imputation	Estimation of expression where zeros have been observed
Phase 2	Normalisation	Removal of unwanted variation that may affect results
Phase 2	Cell Cycle	Assignment or correction of stages of the cell cycle, or other uses of cell cycle genes, or genes associated with similar processes
Phase 3	Classification	Assignment of cell types based on a reference dataset
Phase 3	Clustering	Unsupervised grouping of cells based on expression profiles
Phase 3	Ordering	Ordering of cells along a trajectory
Phase 3	Rare Cells	Identification of rare cell populations
Phase 3	Stem Cells	Identification of cells with stem-like characteristics
Phase 4	Differential Expression	Testing of differential expression across groups of cells
Phase 4	Expression Patterns	Detection of genes that change expression across a trajectory
Phase 4	Gene Networks	Identification or use of co-regulated gene networks
Phase 4	Gene Sets	Testing for over representation or other uses of annotated gene sets
Phase 4	Marker Genes	Identification or use of genes that mark cell populations
Multiple	Dimensionality Reduction	Projection of cells into a lower dimensional space
Multiple	Interactive	Tools with an interactive component or a graphical user interface
Multiple	Variable Genes	Identifcation or use of highly (or lowly) variable genes
Multiple	Visualisation	Functions for visualising some aspect of scRNA-seq data or analysis
Other	Allele Specific	Detection of allele-specific expression
Other	Alternative Splicing	Detection of alternative splicing
Other	Haplotypes	Use or assignment of haplotypes
Other	Immune	Assignment of receptor sequences and immune cell clonality
Other	Integration	Combining of scRNA-seq datasets or integration with other single-cell data types
Other	Modality	Identification or use of modality in gene expression
Other	Simulation	Generation of synthetic scRNA-seq datasets
Other	Transformation	Transformation between expression levels and some other measure
Other	Variants	Detection or use of nucleotide variants

https://doi.org/10.1371/journal.pcbi.1006245.t001

that bulk methods do not perform significantly worse than those designed for scRNA-seq data [38].

To investigate how the focus of scRNA-seq tool development has changed over time we again divided the scRNA-tools database into tools added before and after October 2016. This allowed us to see which analysis tasks are more common in recently released tools. We looked at the percentage of tools in each time period that performed tasks in the different analysis categories (Fig 3B). Some categories show little change in the proportion of tools that perform them while other areas have changed significantly. Specifically, both visualisation and dimensionality reduction are more commonly addressed by recent tools. The UMIs category has also seen a big increase recently as UMI based protocols have become commonly used and tools designed to handle the extra processing steps required have been developed (e.g. UMI-

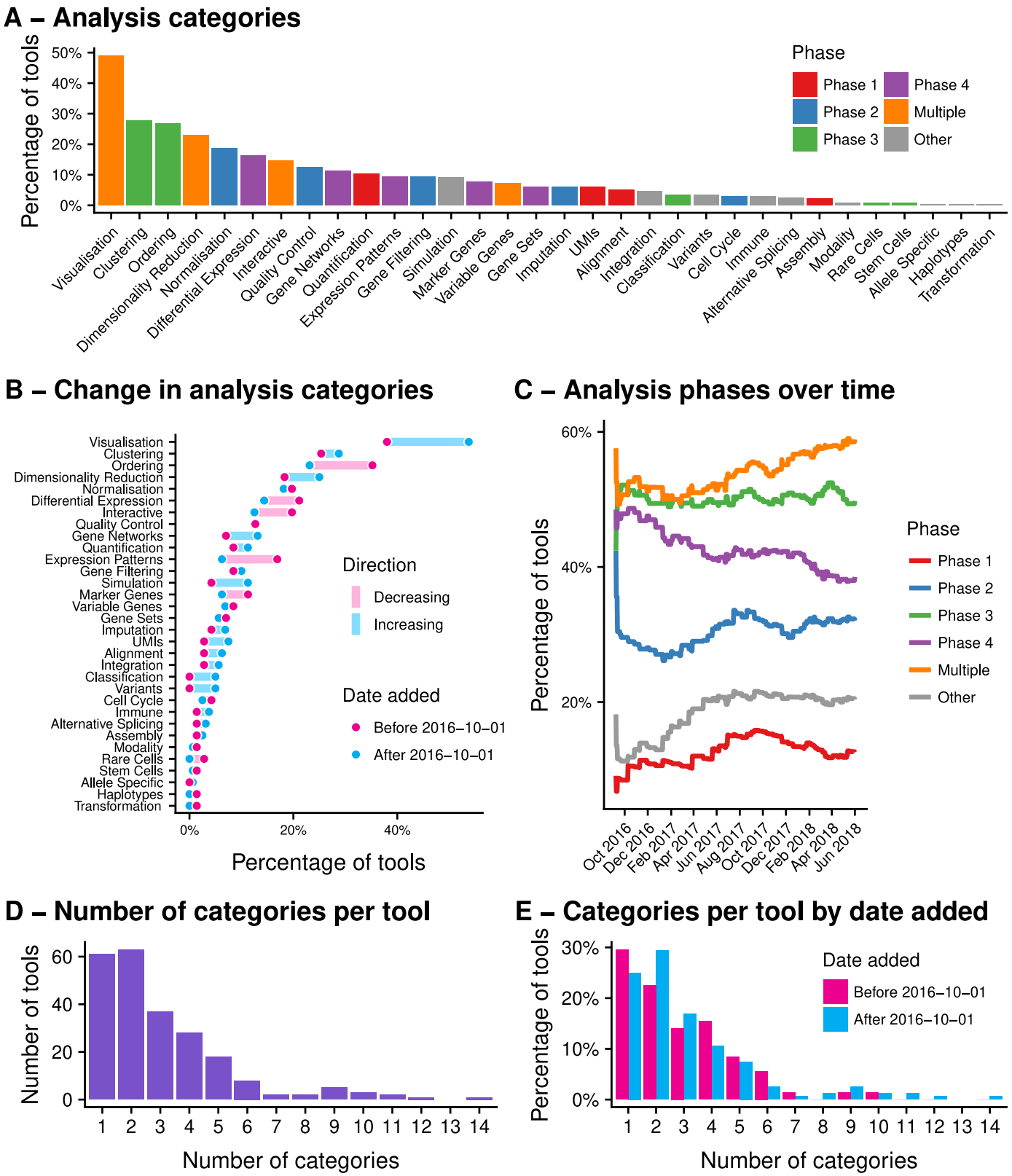


Fig 3. (A) Categories of tools in the scRNA-tools database. Each tool can be assigned to multiple categories based on the tasks it can complete. Categories associated with multiple analysis phases (visualisation, dimensionality reduction) are among the most common, as are categories associated with the cell assignment phase (ordering, clustering). (B) Changes in analysis categories over time, comparing tools added before and after October 2016. There have been

significant increases in the percentage of tools associated with visualisation, dimensionality reduction, gene networks and simulation. Categories including expression patterns, ordering and interactivity have seen relative decreases. (C) Changes in the percentage of tools associated with analysis phases over time. The percentage of tools involved in the data acquisition and data cleaning phases have increased, as have tools designed for alternative analysis tasks. The gene identification phase has seen a relative decrease in the number of tools. (D) The number of categories associated with each tools in the scRNA-tools database. The majority of tools perform few tasks. (E) Most tools that complete many tasks are relatively recent.

https://doi.org/10.1371/journal.pcbi.1006245.g003

tools [39], umis [40], zUMIs [41]). Simulation is a valuable technique for developing, testing and validating scRNA-seq tools. More packages are now including their simulation functions and some tools have been developed for the specific purpose of generating realistic synthetic scRNA-seq datasets (e.g. powsimR [42], Splatter [43]). Classification of cells into known groups has also increased as reference datasets become available and more tools are identifying or making use of co-regulated gene networks.

Some categories have seen a decrease in the proportion of tools they represent, most strikingly testing for expression patterns along a trajectory. This is likely related to the change in cell ordering analysis, which is the focus of a lower percentage of tools added after October 2016. The ordering of cells along a trajectory was one of the first developments in scRNA-seq analysis and a decrease in the development of these tools could indicate that researchers have moved on to other techniques or that use has converged on a set of mature tools.

By grouping categories based on their associated analysis phases we see similar trends over time (Fig 3C). We see increases in the percentage of tools performing tasks in Phase 1 (quantification), across multiple phases (such as visualisation and dimensionality reduction) and alternative analysis tasks. In contrast the percentage of tools that perform gene identification tasks (Phase 4) has decreased and the percentage assigning cells (Phase 3) has remained steady. Phase 2 (quality control and filtering) has fluctuated over time but currently sits at a level slightly above when the database was first created. This also indicates a maturation of the analysis space as developers shift away from the tasks that were the focus of bulk RNA-seq analysis and continue to focus on those specific to scRNA-seq while working on methods for handling data from new protocols and performing alternative analysis tasks.

Pipelines and toolboxes. While there are a considerable number of scRNA-seq tools that only perform a single analysis task, many perform at least two (Fig 3D). Some tools (dropEst [44], DrSeq2 [45], scPipe [46]) are pre-processing pipelines, taking raw sequencing reads and producing an expression matrix. Others, such as Scanpy [47], SCell [48], Seurat, Monocle and scater [49] can be thought of as analysis toolboxes, able to complete a range of complex analyses starting with a gene expression matrix. Most of the tools that complete many tasks are relatively more recent (Fig 3E). Being able to complete multiple tasks using a single tool can simplify analysis as problems with converting between different data formats can be avoided. However it is important to remember that it is difficult for a tool with many functions to continue to represent the state of the art in all of them. Support for common data formats, such as the recently released SingleCellExperiment [50], anndata [47] or loom (http://loompy.org) objects provides another way for developers to allow easy use of their tools and for users to build custom workflows from specialised tools.

Alternative analyses. Some tools perform analyses that lie outside the common tasks performed on scRNA-seq data described above. Simulation is one alternative task that has already been mentioned but there is also a group of tools designed to detect biological signals in scRNA-seq data apart from changes in expression. For example identifying alternative splicing (BRIE [51], Outrigger [52], SingleSplice [53]), single nucleotide variants (SSrGE [54]), copy number variants (inferCNV [55]) and allele-specific expression (SCALE [56]). Reconstruction of immune cell receptors is another area that has received considerable attention from tools such as BASIC [57], TraCeR [58] and TRAPeS [59]. While tools that complete these tasks are

unlikely to ever dominate scRNA-seq analysis we expect to see an increase in methods for tackling specialised analyses as researchers continue to push the boundaries of what can be observed using scRNA-seq data.

Availability and future directions

Since October 2016 we have seen the number of software tools for analysing single-cell RNA-seq data more than triple, with more than 230 analysis tools now available. As new tools have become available we have curated and catalogued them in the scRNA-tools database where we record the analysis tasks that they can complete, along with additional information such as any associated publications. By analysing this database we have found that tool developers have focused much of their efforts on methods for handling new problems specific to scRNA-seq data, in particular clustering cells into groups or ordering them along a trajectory. We have also seen that the scRNA-seq community is generally open and willing to share their methods which are often described in preprints prior to peer-reviewed publication and released under permissive open-source licenses for other researchers to re-use.

The next few years promise to produce significant new developments in scRNA-seq analysis. New tools will continue to be produced, becoming increasingly sophisticated and aiming to address more of the questions made possible by scRNA-seq data. We anticipate that some existing tools will continue to improve and expand their functionality while others will cease to be updated and maintained. Detailed benchmarking and comparisons will show how tools perform in different situations and those that perform well, continue to be developed and provide a good user experience will become preferred for standard analyses. As single-cell capture and sequencing technology continues to improve analysis tools will have to adapt to significantly larger datasets (in the millions of cells) which may require specialised data structures and algorithms. Methods for combining multiple scRNA-seq datasets as well as integration of scRNA-seq data with other single-cell data types, such as DNA-seq, ATAC-seq or methylation, will be another area of growth. In addition, projects such as the Human Cell Atlas [60] will provide comprehensive cell type references which will open up new avenues for analysis.

As the field expands the scRNA-tools database will continue to be updated with support from the community. We hope that it provides a resource for researchers to explore when approaching scRNA-seq analyses as well as providing a record of the analysis landscape and how it changes over time.

Availability

The scRNA-tools databases is publicly accessible via the website at www.scRNA-tools.org. Suggestions for additions, updates and improvements are warmly welcomed at the associated GitHub repository (https://github.com/Oshlack/scRNA-tools) or via the submission form on the website. The code and datasets used for the analysis in this paper are available from https://github.com/Oshlack/scRNAtools-paper.

Acknowledgments

We would like to acknowledge Sean Davis' work in managing the awesome-single-cell page and producing a prototype of the script used to process the database. Daniel Wells had the idea for recording software licenses and provided licenses for the tools in the database at that time. Breon Schmidt designed a prototype of the scRNA-tools website and answered many questions about HTML and Javascript. Our thanks also to Matt Ritchie for his thoughts on early versions of the manuscript.

Author Contributions

Conceptualization: Luke Zappia.

Formal analysis: Luke Zappia.

Investigation: Luke Zappia.

Methodology: Luke Zappia.

Software: Luke Zappia.

Supervision: Belinda Phipson, Alicia Oshlack.

Writing – original draft: Luke Zappia.

Writing – review & editing: Luke Zappia, Belinda Phipson, Alicia Oshlack.

References

1. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009; 6: 377–382. https://doi.org/10.1038/nmeth.1315 PMID: 19349980
2. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018; 13: 599. https://doi.org/10.1038/nprot.2017.149 PMID: 29494575
3. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015; 16: 133–145. https://doi.org/10.1038/nrg3833 PMID: 25628217
4. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015; 12: 115–121. https://doi.org/10.1038/nmeth.3252 PMID: 25633503
5. Chamberlain S, Boettiger C, Hart T, Ram K. rcrossref: Client for Various 'CrossRef' 'APIs'. 2017. https://CRAN.R-project.org/package=rcrossref
6. Ram K, Broman K. aRxiv: Interface to the arXiv API. 2017. https://CRAN.R-project.org/package=aRxiv
7. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2010.
8. Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, et al. plotly: Create Interactive Web Graphics via 'plotly.js'. 2017. https://CRAN.R-project.org/package=plotly
9. Wickham H, Francois R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2017. https://CRAN.R-project.org/package=dplyr
10. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. 2017. https://CRAN.R-project.org/package=cowplot
11. Katayama S, Töhönen V, Linnarsson S, Kere J. SAMstrt: statistical test for differential expression in single-cell transcriptome with spike-in normalization. Bioinformatics. 2013; 29: 2943–2945. https://doi.org/10.1093/bioinformatics/btt511 PMID: 23995393
12. Bourne PE, Polka JK, Vale RD, Kiley R. Ten simple rules to consider regarding preprint submission. PLoS Comput Biol. 2017; 13: e1005473. https://doi.org/10.1371/journal.pcbi.1005473 PMID: 28472041
13. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. 2016; 17: 63. https://doi.org/10.1186/s13059-016-0927-y PMID: 27052890
14. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016; 34: 1145–1160. https://doi.org/10.1038/nbt.3711 PMID: 27824854
15. Miragaia RJ, Teichmann SA, Hagai T. Single-cell insights into transcriptomic diversity in immunity. Current Opinion in Systems Biology. 2017; 5: 63–71. https://doi.org/10.1016/j.coisb.2017.08.003
16. Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics and computational challenges. Front Genet. 2016; 7. https://doi.org/10.3389/fgene.2016.00163 PMID: 27708664
17. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for interpreting scRNA-seq data. FEBS Lett. 2017; https://doi.org/10.1002/1873-3468.12684 PMID: 28524227
18. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012; 9: 72–74. https://doi.org/10.1038/nmeth.1778 PMID: 22101854
19. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33: 495–502. https://doi.org/10.1038/nbt.3192 PMID: 25867923

20. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; https://doi.org/10.1038/nbt.4096 PMID: 29608179

21. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017; 14: 483–486. https://doi.org/10.1038/nmeth.4236 PMID: 28346451

22. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015; 347: 1138–1142. https://doi.org/10.1126/science.aaa1934 PMID: 25700174

23. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32: 381–386. https://doi.org/10.1038/nbt.2859 PMID: 24658644

24. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017; https://doi.org/10.1038/nmeth.4150 PMID: 28114287

25. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017; https://doi.org/10.1038/nmeth.4402 PMID: 28825705

26. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016; 34: 637–645. https://doi.org/10.1038/nbt.3569 PMID: 27136076

27. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods. 2016; https://doi.org/10.1038/nmeth.3971 PMID: 27571553

28. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008; 9: 2579–2605. Available: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

29. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308

30. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012; 40: 4288–4297. https://doi.org/10.1093/nar/gks042 PMID: 22287627

31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15: 550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

32. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43: e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

33. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11: 740–742. https://doi.org/10.1038/nmeth.2967 PMID: 24836921

34. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016; 17: 222. https://doi.org/10.1186/s13059-016-1077-y PMID: 27782827

35. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. Brief Bioinform. 2016; https://doi.org/10.1093/bib/bbw057 PMID: 27373736

36. Miao Z, Zhang X. Differential expression analyses for single-cell RNA-Seq: old questions on new data. Quant Biol. 2016; 4: 243–260. https://doi.org/10.1007/s40484-016-0089-7

37. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: assessment of differential expression analysis methods. Front Genet. 2017; 8: 62. https://doi.org/10.3389/fgene.2017.00062 PMID: 28588607

38. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018; https://doi.org/10.1038/nmeth.4612 PMID: 29481549

39. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017; 27: 491–499. https://doi.org/10.1101/gr.209601.116 PMID: 28100584

40. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017; https://doi.org/10.1038/nmeth.4220 PMID: 28263961

41. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs—A fast and flexible pipeline to process RNA sequencing data with UMIs. Gigascience. 2018; https://doi.org/10.1093/gigascience/giy059 PMID: 29846586

42. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Bioinformatics. 2017; 33: 3486–3488. https://doi.org/10.1093/bioinformatics/btx435 PMID: 29036287

43. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017; 18: 174. https://doi.org/10.1186/s13059-017-1305-0 PMID: 28899397

44. Petukhov V, Guo J, Baryawno N, Severe N, Scadden D, Kharchenko PV. Accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. bioRxiv. 2017. p. 171496. 10.1101/171496

45. Zhao C, Hu S, Huo X, Zhang Y. Dr.seq2: A quality control and analysis pipeline for parallel single cell transcriptome and epigenome data. PLoS One. 2017; 12: e0180583. https://doi.org/10.1371/journal.pone.0180583 PMID: 28671995

46. Tian L, Su S, Amann-Zalcenstein D, Biben C, Naik SH, Ritchie ME. scPipe: a flexible data preprocessing pipeline for single-cell RNA-sequencing data. bioRxiv. 2017. p. 175927. 10.1101/175927

47. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018; 19: 15. https://doi.org/10.1186/s13059-017-1382-0 PMID: 29409532

48. Diaz A, Liu SJ, Sandoval C, Pollen A, Nowakowski TJ, Lim DA, et al. SCell: integrated analysis of single-cell RNA-seq data. Bioinformatics. 2016; https://doi.org/10.1093/bioinformatics/btw201 PMID: 27153637

49. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017; 33: 1179–1186. https://doi.org/10.1093/bioinformatics/btw777 PMID: 28088763

50. Lun A, Risso D. SingleCellExperiment: S4 Classes for Single Cell Data. 2017.

51. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in single cells. Genome Biol. 2017; 18: 123. https://doi.org/10.1186/s13059-017-1248-5 PMID: 28655331

52. Song Y, Botvinnik OB, Lovci MT, Kakaradov B, Liu P, Xu JL, et al. Single-cell alternative splicing analysis with Expedition reveals splicing dynamics during neuron differentiation. Mol Cell. 2017; https://doi.org/10.1016/j.molcel.2017.06.003 PMID: 28673540

53. Welch JD, Hu Y, Prins JF. Robust detection of alternative splicing in a population of single cells. Nucleic Acids Res. 2016; 44: e73. https://doi.org/10.1093/nar/gkv1525 PMID: 26740580

54. Poirion OB, Zhu X, Ching T, Garmire LX. Using single nucleotide variations in cancer single-cell RNA-seq data for subpopulation identification and genotype-phenotype linkage analysis. bioRxiv. 2016. p. 095810. 10.1101/095810

55. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014; 344: 1396–1401. https://doi.org/10.1126/science.1254257 PMID: 24925914

56. Jiang Y, Zhang NR, Li M. SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol. 2017; 18: 74. https://doi.org/10.1186/s13059-017-1200-8 PMID: 28446220

57. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinformatics. 2017; 33: 425–427. https://doi.org/10.1093/bioinformatics/btw631 PMID: 28172415

58. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016; 13: 329–332. https://doi.org/10.1038/nmeth.3800 PMID: 26950746

59. Afik S, Yates KB, Bi K, Darko S, Godec J, Gerdemann U, et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. Nucleic Acids Res. 2017; https://doi.org/10.1093/nar/gkx615 PMID: 28934479

60. Regev A, Teichmann S, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. bioRxiv. 2017. p. 121202. 10.1101/121202