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

* Corresponding author

E-mail: alicia.oshlack@mcri.edu.au

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 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 of the growth of the field over time.

Author summary

In recent years single-cell RNA-sequeing 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 focued on methods specific to single-cell data and that they embrace an open-source approach with permissive licensing, sharing of code and preprint publications.

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 detect 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. As such, 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, 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.

Overview of the scRNA-tools 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 to be included 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. Various details of the tools are recorded such as the programming language or platform they use, details of any related publications, links to the source code and the associated software license. Tools are also categorised according to the analysis tasks they are able to perform. 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).

The scRNA-tools website provides a profile for each tool, with links to publications and code repositories, as well as an index by analysis category. Both of these pages can be sorted in a variety of ways, including by the number of associated publications or citations. We also provide an interactive table that allows users to filter and sort tools in more sophisticated ways to find those most relevant to their needs. An additional page shows live and up-to-date versions of some of the analysis and visualisations of the database presented below. 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.

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 [5] in November 2013 up to September 2016. In the time since then over 130 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.

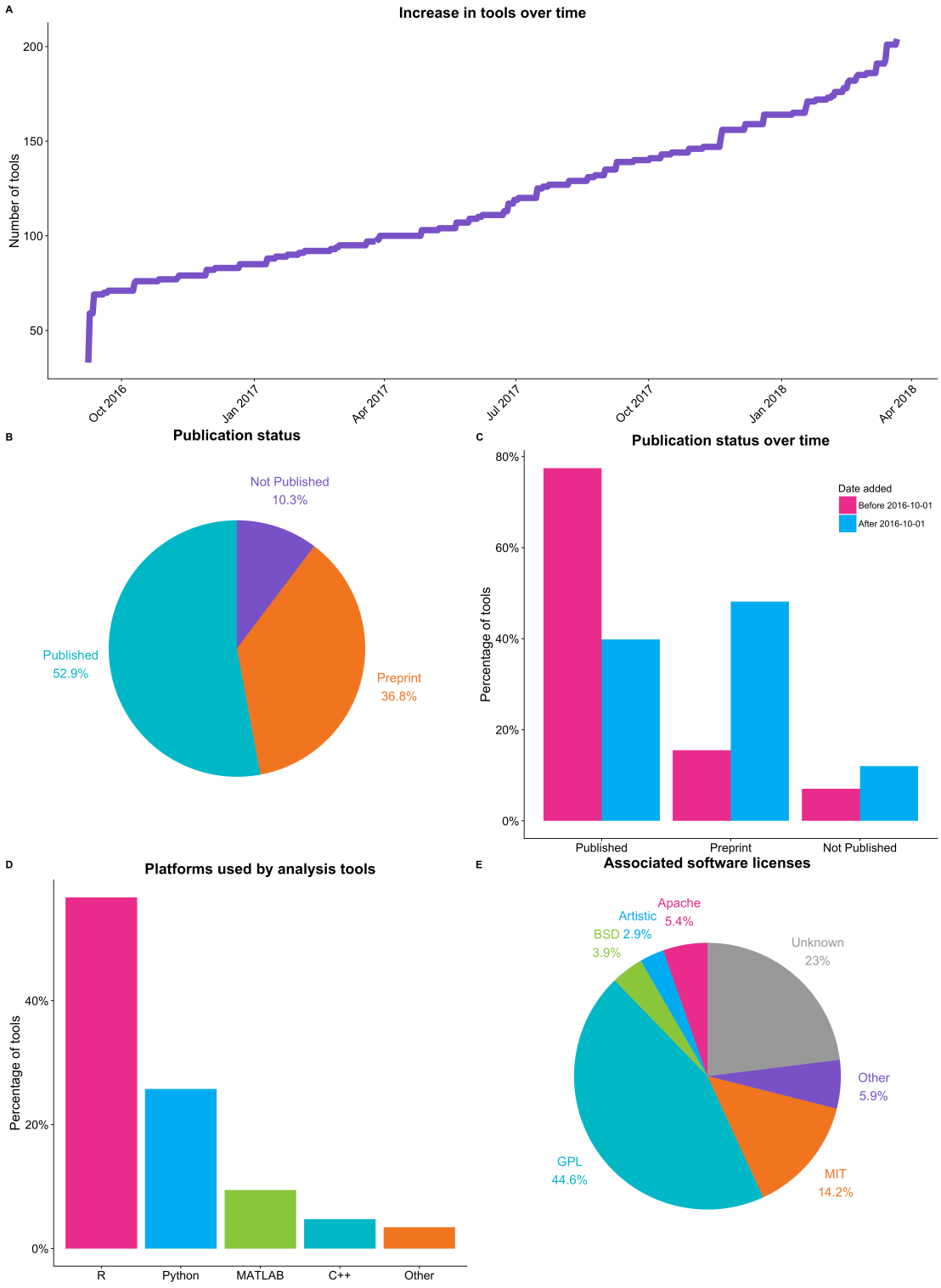

***Fig** 1 (A) Number of tools in the scRNA-tools database over time. Since the scRNA-seq tools database was started in September 2016 more than 130 new tools have been released. (B) Publication status of tools in the scRNA-tools database. Over half*

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.

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 publications 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 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 [6].

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

139 popularity of R reflects its history as the language of choice for the analysis of bulk
140 RNA-seq datasets and a range of other biological data types.

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

155 Categories of scRNA-seq analysis

156 Single-cell RNA-sequencing is often used to explore complex mixtures of cell types in
157 an unsupervised manner. As has been described in previous reviews a standard
158 scRNA-seq analysis in this setting consists of several tasks which can be completed
159 using various tools [7–11]. In the scRNA-tools database we categorise tools based on
160 the analysis tasks they perform. Here we group these tasks into four broad phases of
161 analysis: data acquisition, data cleaning, cell assignment and gene identification (Fig
162 2). The data acquisition phase (Phase 1) takes the raw nucleotide sequences from the
163 sequencing experiment and returns a matrix describing the expression of each gene
164 (rows) in each cell (columns). 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) [12]. 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 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 attention 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.

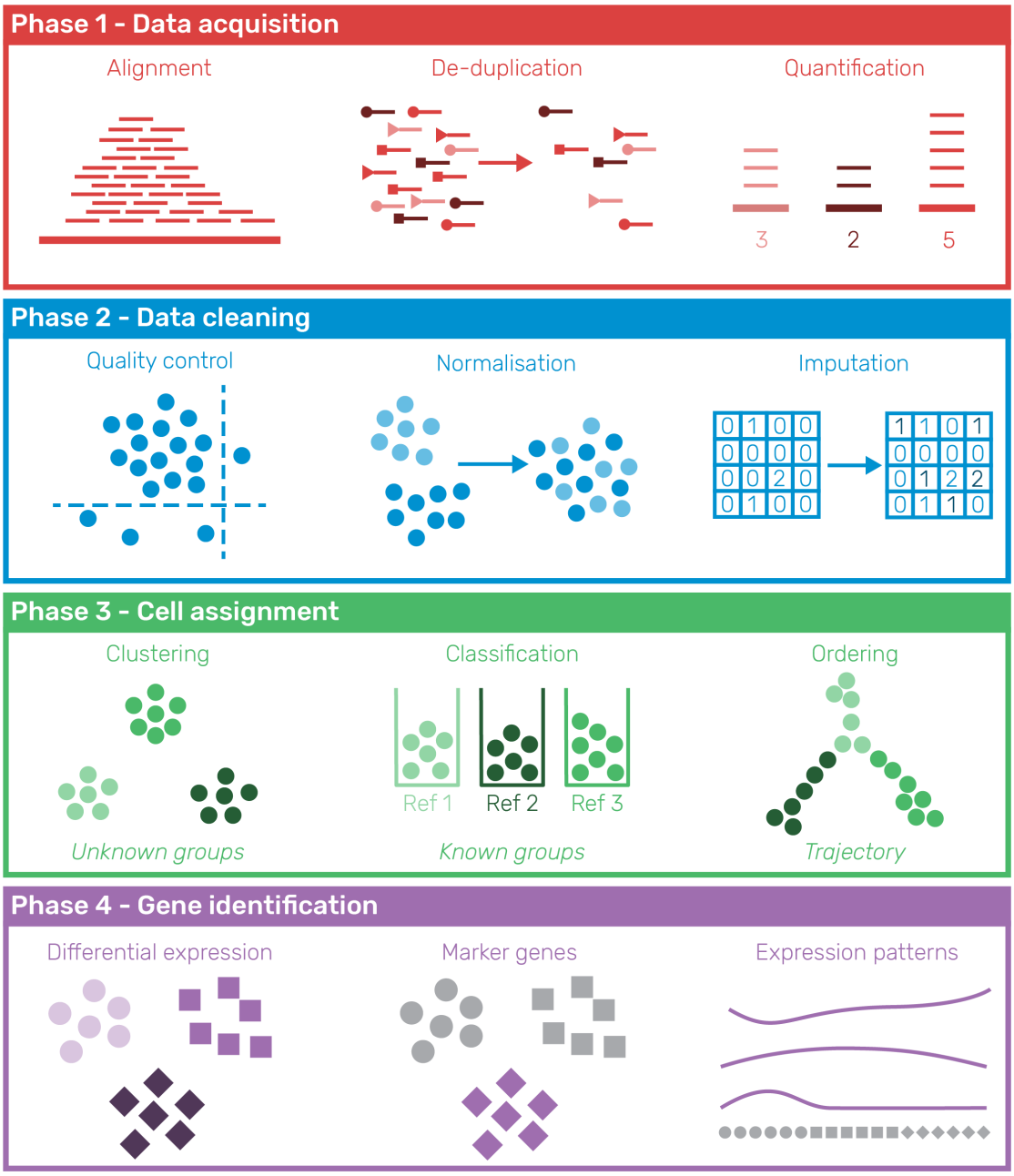

***Fig** 2 Phases of a typical unsupervised scRNA-seq analysis process. In Phase 1 (data acquisition) raw sequencing reads are converted into a gene by cell expression matrix. For many protocols this requires the alignment of genes to a reference genome and the assignment and de-duplication of Unique 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).*

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

198 biological insight applicable to many studies. An exception to this may be
199 experiments designed to test a specific hypothesis where cell populations may have
200 been sorted or the interest lies in differences between experimental conditions rather
201 than cell types. In this case Phase 3 may not be required and slightly different tools or
202 approaches may be used by many of the same challenges will apply. In addition, as
203 the field expands and develops it is likely that data will be used in new ways to answer
204 other biological questions, requiring new analysis techniques. Descriptions of the
205 categories in the scRNA-tools database are given in Table 1, along with the associated
206 analysis phases.

207 *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

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 [[13]; Butler2017-lb], SC3 [14] and BackSPIN [15] being used to identify cell types in a sample and trajectory analysis tools (for example Monocle [[16]; Qiu2017-bi; Qiu2017-mq], Wishbone [17] and DPT [18]) 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.

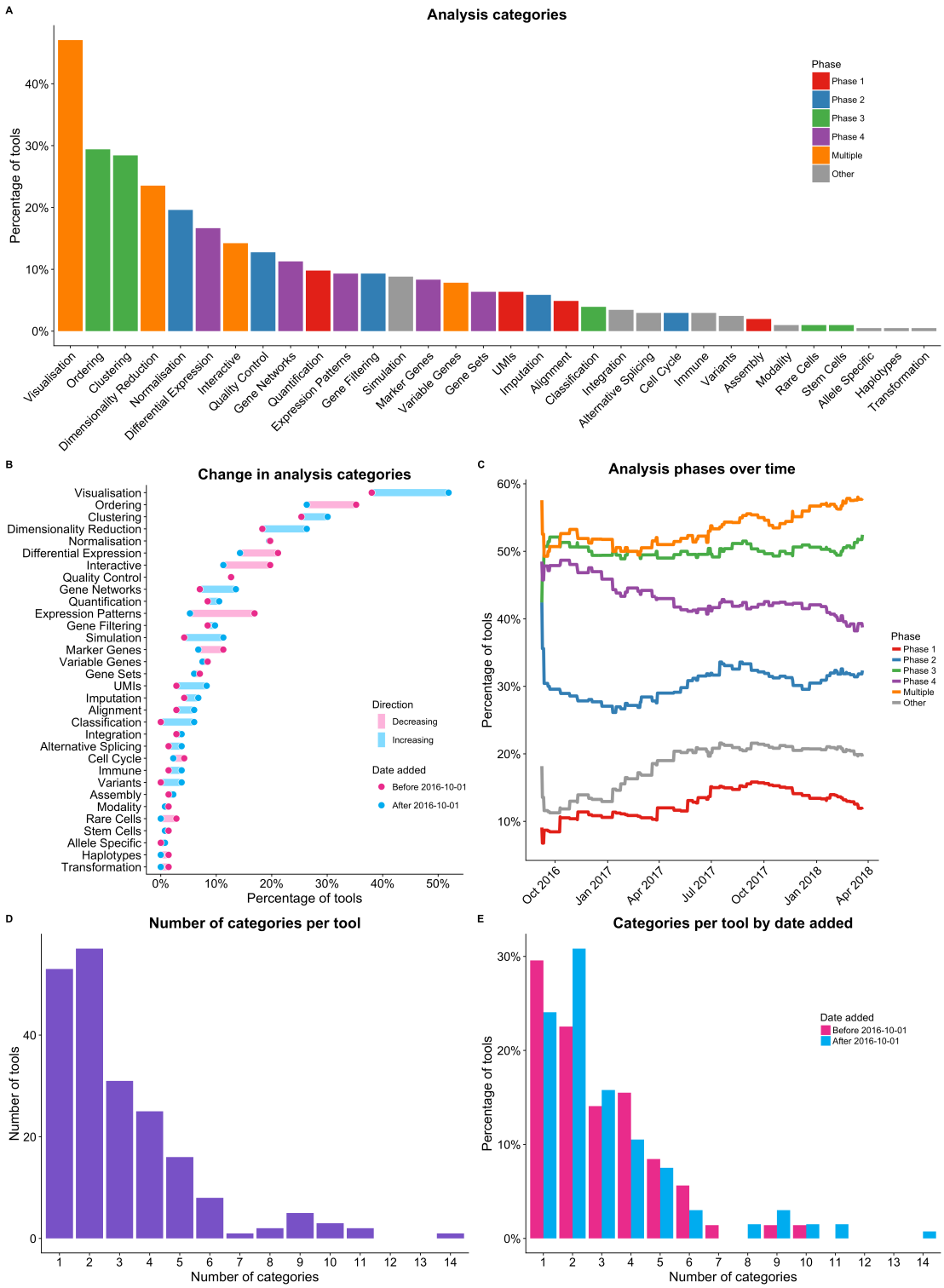

***Fig** 3 (A) Categories of tools in the scRNA-tools database. Each tool can be assigned to multiple categories based on the tasks it can complete. Categories associated with multiple analysis phases (visualisation, dimensionality reduction) are among the most common, as are categories associated with 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.

Dimensionality reduction is also a common task and has applications in visualisation (via techniques such as t-SNE [19]), 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 cluster 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 [20,21], DESeq2 [22] and limma [23], 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 [24], and scDD [25]) it is still unclear whether they improve on methods that have been established for bulk data [26–28], with the most comprehensive comparison to date finding that bulk methods do not perform significantly worse than those designed for scRNA-seq data [29].

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 published 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

262 (e.g. UMI-tools [30], umis [31], zUMIs [32]). Simulation is a valuable technique for
263 developing, testing and validating scRNA-seq tools. More packages are now including
264 their simulation functions and some tools have been developed for the specific
265 purpose of generating realistic synthetic scRNA-seq datasets (e.g. powsimR [33],
266 Splatter [34]). Classification of cells into known groups has also increased as
267 reference datasets become available and more tools are identifying or making use of
268 co-regulated gene networks.

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

276 By grouping categories based on their associated analysis phases we see similar
277 trends over time (Fig 3C). We see increases in the percentage of tools performing
278 tasks in Phase 1 (quantification), across multiple phases (such as visualisation and
279 dimensionality reduction) and alternative analysis tasks. In contrast the percentage
280 of tools that perform gene identification tasks (Phase 4) has decreased and the
281 percentage assigning cells (Phase 3) has remained steady. Phase 2 (quality control
282 and filtering) has flucuated over time but currently sits at a similar level to when the
283 database was first created. This too indicates a maturation of the analysis space as
284 developers shift away from the tasks that were the focus of bulk RNA-seq analysis
285 and continue to focus on those specific to scRNA-seq while working on methods for
286 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 [35], DrSeq2 [36], scPipe [37]) are pre-processing pipelines, taking raw sequencing reads and producing an expression matrix. Others, such as Scanpy [38], SCell [39], Seurat, Monocle and scater [40] 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 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 functionalities to continue to represent the state of the art in all of them. Support for common data formats, such as the recently released SingleCellExperiment object in R [41], 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 alternative splicing (BRIE [42], Outrigger [43], SingleSplice [44]), single nucleotide variants (SSrGE [45]), copy number variants (inferCNV [46]) and allele-specific expression (SCALE [47]). Reconstruction of immune cell receptors is another area that has received considerable attention from tools such as BASIC [48], TraCeR [49] and TRAPeS [50]. 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.

Discussion

Since October 2016 we have seen the number of software tools for analysing single-cell RNA-seq data almost triple, with more than 200 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 continue 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, with be another area of growth and projects such as the Human Cell Atlas [51] 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.

Methods

Database

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.

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) [52]. We also make use of the aRxiv package (v0.5.16) [53] to retrieve information about arXiv preprints. JSON files describing the complete table, tools and categories are outputted 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. An additional page shows a up-to-date version of some of the analysis presented here with visulisations produced using ggplot2 (v2.2.1.9000) [54] and plotly (v4.7.1) [55].

Analysis

The most recent version of the scRNA-tools database as of 2018-03-22 was used for the analysis presented in this paper. Data was manipulated in R (v3.4.3) using the dplyr package (v0.7.4) [56] and plots produced using the ggplot2 (v2.2.1.9000) and cowplot (v0.9.2) [57] packages.

Declarations

Ethics

Not applicable.

Availability of data and materials

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.

Competing interests

The authors declare no competing interests.

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.

Authors' contributions

LZ developed the scRNA-tools database and website. BP contributed to supervision of the project. AO oversaw all aspects of the project. All authors contributed to drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

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.

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. doi:10.1038/nmeth.1315

2. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13: 599. doi:10.1038/nprot.2017.149

3. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16: 133–145. doi:10.1038/nrg3833

416 4. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al.
417 Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods.
418 2015;12: 115–121. doi:10.1038/nmeth.3252

419 5. Katayama S, Töhönen V, Linnarsson S, Kere J. SAMstrt: statistical test for
420 differential expression in single-cell transcriptome with spike-in normalization.
421 Bioinformatics. 2013;29: 2943–2945. doi:10.1093/bioinformatics/btt511

422 6. Bourne PE, Polka JK, Vale RD, Kiley R. Ten simple rules to consider regarding
423 preprint submission. PLoS Comput Biol. 2017;13: e1005473.
424 doi:10.1371/journal.pcbi.1005473

425 7. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-
426 sequencing experiments. Genome Biol. 2016;17: 63. doi:10.1186/s13059-016-0927-y

427 8. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-
428 cell genomics. Nat Biotechnol. 2016;34: 1145–1160. doi:10.1038/nbt.3711

429 9. Miragaia RJ, Teichmann SA, Hagai T. Single-cell insights into transcriptomic
430 diversity in immunity. Current Opinion in Systems Biology. 2017;5: 63–71.
431 doi:10.1016/j.coisb.2017.08.003

432 10. Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics
433 and computational challenges. Front Genet. 2016;7. doi:10.3389/fgene.2016.00163

434 11. Rostom R, Svensson V, Teichmann SA, Kar G. Computational approaches for
435 interpreting scRNA-seq data. FEBS Lett. 2017; doi:10.1002/1873-3468.12684

436 12. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al.
437 Counting absolute numbers of molecules using unique molecular identifiers. Nat
438 Methods. 2012;9: 72–74. doi:10.1038/nmeth.1778

439 13. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of
440 single-cell gene expression data. Nat Biotechnol. 2015;33: 495–502.
441 doi:10.1038/nbt.3192

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

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

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

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

454 18. Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ. Diffusion pseudotime
455 robustly reconstructs lineage branching. Nat Methods. 2016;
456 doi:10.1038/nmeth.3971

457 19. Maaten L van der, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res.
458 2008;9: 2579–2605. Available:
459 http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

460 20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for
461 differential expression analysis of digital gene expression data. Bioinformatics.
462 2010;26: 139–140. doi:10.1093/bioinformatics/btp616

463 21. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor
464 RNA-Seq experiments with respect to biological variation. Nucleic Acids Res.
465 2012;40: 4288–4297. doi:10.1093/nar/gks042

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

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

472 24. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell
473 differential expression analysis. Nat Methods. 2014;11: 740–742.
474 doi:10.1038/nmeth.2967

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

478 26. Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to
479 detect differentially expressed genes between single-cell populations. Brief
480 Bioinform. 2016; doi:10.1093/bib/bbw057

481 27. Miao Z, Zhang X. Differential expression analyses for single-cell RNA-Seq: old
482 questions on new data. Quant Biol. 2016;4: 243–260. doi:10.1007/s40484-016-
483 0089-7

484 28. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: assessment of
485 differential expression analysis methods. Front Genet. 2017;8: 62.
486 doi:10.3389/fgene.2017.00062

487 29. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell
488 differential expression analysis. Nat Methods. 2018; doi:10.1038/nmeth.4612

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

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

495 32. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs: A fast and flexible
496 pipeline to process RNA sequencing data with UMIs. bioRxiv. 2017. p. 153940.
497 doi:10.1101/153940

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

501 34. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA
502 sequencing data. Genome Biol. 2017;18: 174. doi:10.1186/s13059-017-1305-0

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

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

509 37. Tian L, Su S, Amann-Zalcenstein D, Biben C, Naik SH, Ritchie ME. scPipe: a
510 flexible data preprocessing pipeline for single-cell RNA-sequencing data. bioRxiv.
511 2017. p. 175927. doi:10.1101/175927

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

514 39. Diaz A, Liu SJ, Sandoval C, Pollen A, Nowakowski TJ, Lim DA, et al. SCell:
515 integrated analysis of single-cell RNA-seq data. Bioinformatics. 2016;
516 doi:10.1093/bioinformatics/btw201

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

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

521 42. Huang Y, Sanguinetti G. BRIE: transcriptome-wide splicing quantification in
522 single cells. Genome Biol. 2017;18: 123. doi:10.1186/s13059-017-1248-5

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

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

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

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

534 47. Jiang Y, Zhang NR, Li M. SCALE: modeling allele-specific gene expression by
535 single-cell RNA sequencing. Genome Biol. 2017;18: 74. doi:10.1186/s13059-017-
536 1200-8

537 48. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from
538 single cells. Bioinformatics. 2017;33: 425–427. doi:10.1093/bioinformatics/btw631

539 49. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T
540 cell fate and clonality inference from single-cell transcriptomes. Nat Methods.
541 2016;13: 329–332. doi:10.1038/nmeth.3800

50. 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; doi:10.1093/nar/gkx615

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

52. Chamberlain S, Boettiger C, Hart T, Ram K. rcrossref: Client for Various 'CrossRef' 'APIs'. 2017. Available: https://CRAN.R-project.org/package=rcrossref

53. Ram K, Broman K. aRxiv: Interface to the arXiv API. 2017. Available: https://CRAN.R-project.org/package=aRxiv

54. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2010. Available: https://market.android.com/details?id=book-rhRqtQAACAAJ

55. Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, et al. plotly: Create Interactive Web Graphics via 'plotly.js'. 2017. Available: https://CRAN.R-project.org/package=plotly

56. Wickham H, Francois R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2017. Available: https://CRAN.R-project.org/package=dplyr

57. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. 2017. Available: https://CRAN.R-project.org/package=cowplot