Supplementary Material and Figures—*scater:* pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R

Davis J. McCarthy [1,2,5,*], Kieran R. Campbell [2,4], Aaron T. L. Lun [6], Quin F. Wills [2,3]

1 European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), Hinxton CB10 1SD, United Kingdom;
2 Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, United Kingdom;
3 Weatherall Institute for Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom;
4 Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom;
5 St Vincent's Institute of Medical Research, 41 Victoria Parade, Fitzroy Victoria 3065, Australia; and
6 CRUK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, United Kingdom.

*** Contact: davis@ebi.ac.uk**

Abstract

Supplementary material and figures—details of package dependencies; an overview of the SCESet class; an overview of the *scater* ecosystem; examples of using the *scater* GUI.
Contact: davis@ebi.ac.uk

Details of package dependencies

The package builds on many other R packages: *Biobase* and *BiocGenerics* for core Bioconductor functionality (Huber et al., 2015); *plyr* (Wickham, 2015), *reshape2* (Wickham, 2012), *dplyr* (Wickham and Francois, 2015), *data.table* (Dowle et al., 2015) and *magrittr* (Bache and Wickham, 2014) for reading and tidying data; *ggplot2* (Wickham, 2016) for plotting; *biomaRt* (Durinck et al., 2005) for feature annotation; *edgeR* (Robinson et al., 2010) for computation of normalisation size factors and counts-per-million values; *limma* (Ritchie et al., 2015) for efficient fitting of linear models to features; *rhdf5* (Fischer and Pau, 2016), *rjson* (Couture-Beil, 2014) and *tximport* (Soneson et al., 2015) for reading in transcript-level expression values; *viridis* (Garnier, 2016) for perceptually-uniform colour maps for plotting; *parallel* for parallel computation; *matrixStats* (Bengtsson, 2016) for computation of summary statistics from matrices; *cowplot* (Wilke, 2016) for attractive plotting themes; *destiny* (Angerer et al., 2015) for producing diffusion maps; *Rtsne* (Krijthe, 2015) for producing t-SNE plots; *mvoutlier* (Filzmoser and Gschwandtner, 2015) for multivariate outlier detection from PCA of QC metrics; *roxygen2* (Wickham et al., 2015), *BiocStyle* (Huber et al., 2015), *knitr* (Xie, 2013) and *rmarkdown* (Allaire et al., 2016) for generating documentation; and *testthat* (Wickham, 2011) for unit testing. As well as functioning in the usual R

environments, *scater* also has a GUI built using *shiny* (Chang et al., 2016) and *shinydashboard* (Chang, 2015) for intuitive and interactive data visualisation. Calling the `scater_gui` function from within an R session opens up the GUI in a web browser.

Supplementary Figures

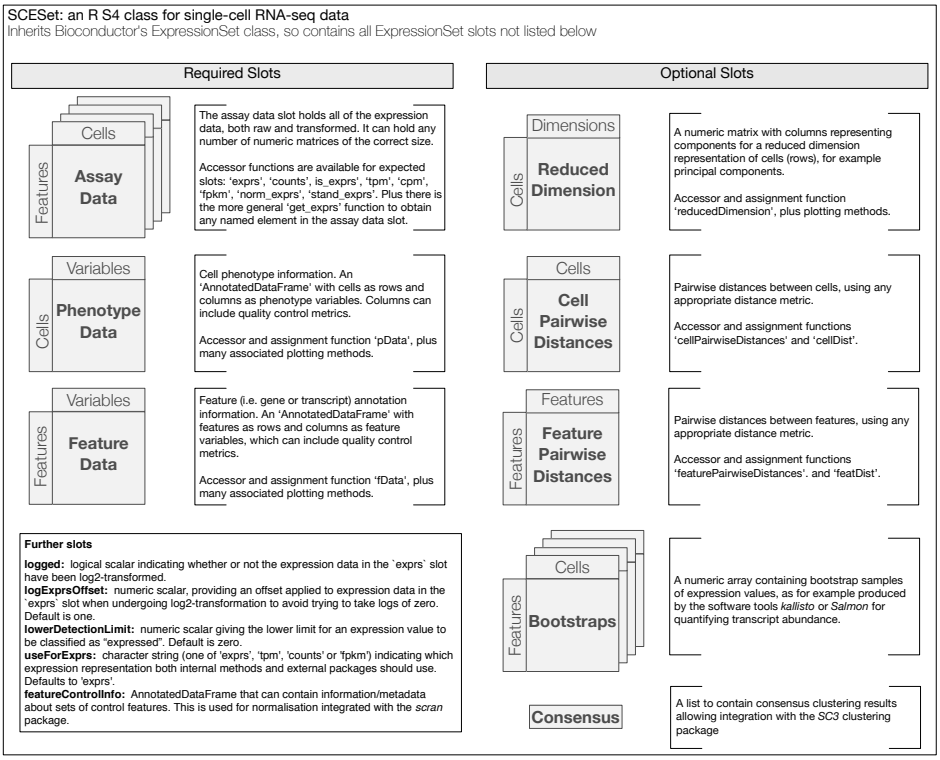

Figure 1: An overview of the SCESet class that underpins the *scater* package. Building on Bioconductor's ExpressionSet class, it is a fully-featured, sophisticated and flexible data class tailored to scRNA-seq data.

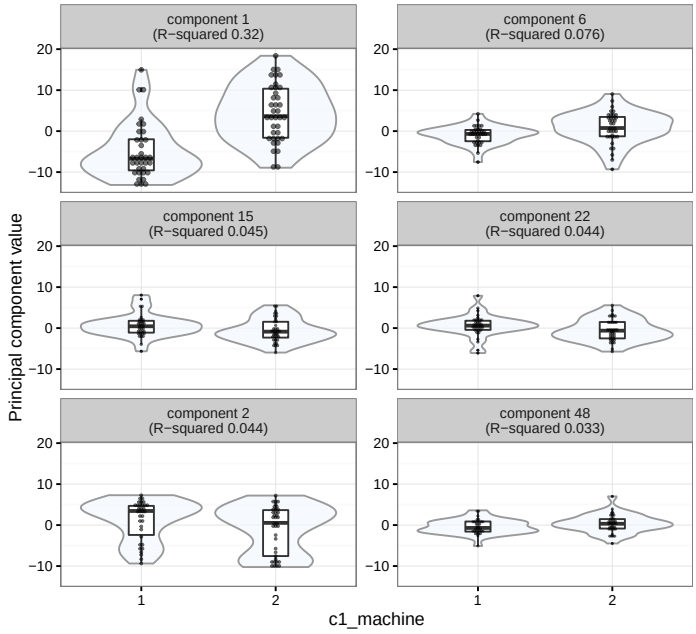

Figure 2: A QC plot produced by the plotQC function in *scater* showing violin, scatter- and boxplots of principal component values against the C1 machine used for each cell for the six principal components most strongly correlated with C1 machine used.

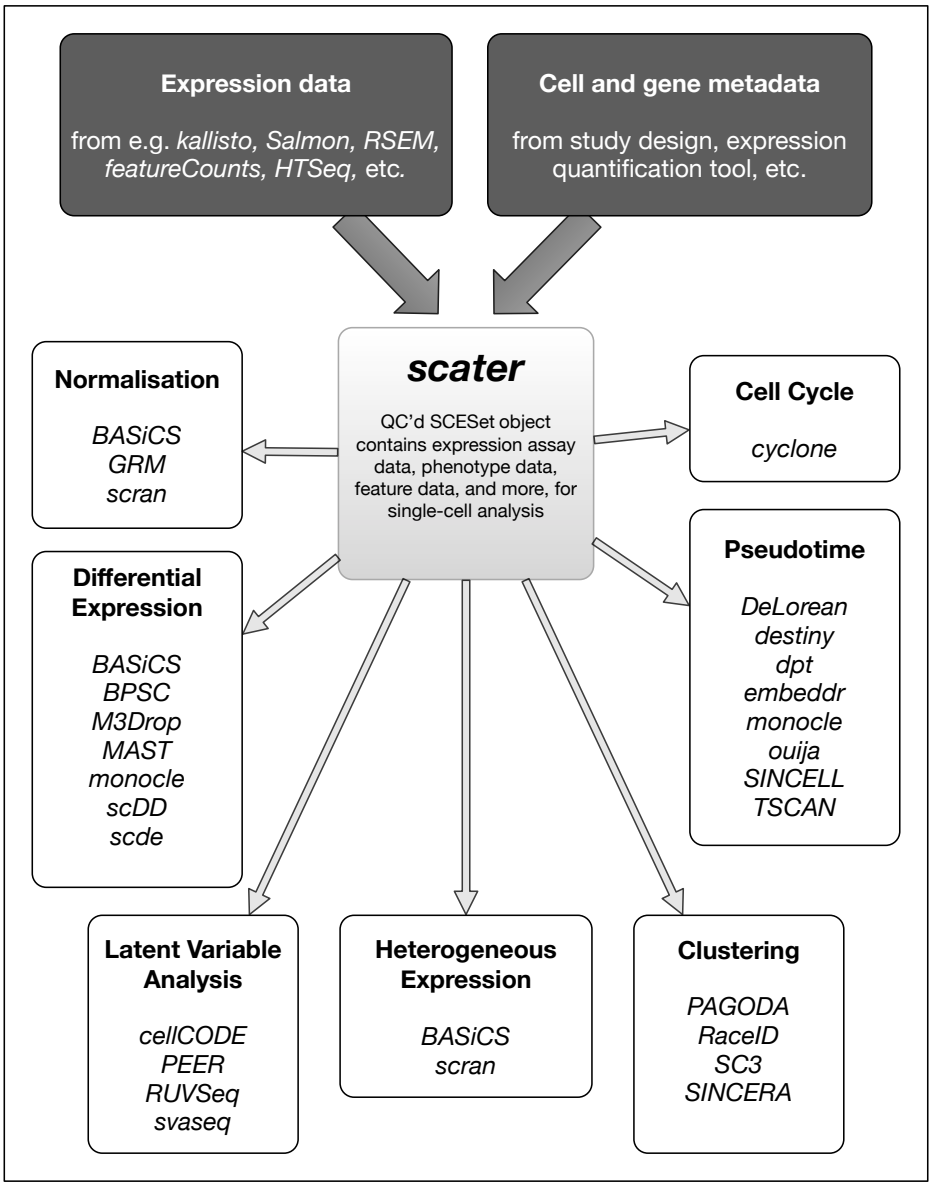

Figure 3: An overview of the *scater* ecosystem. The SCESet class in *scater* acts as a convenient hub for datasets so that many other methods and tools implemented in R can be applied.

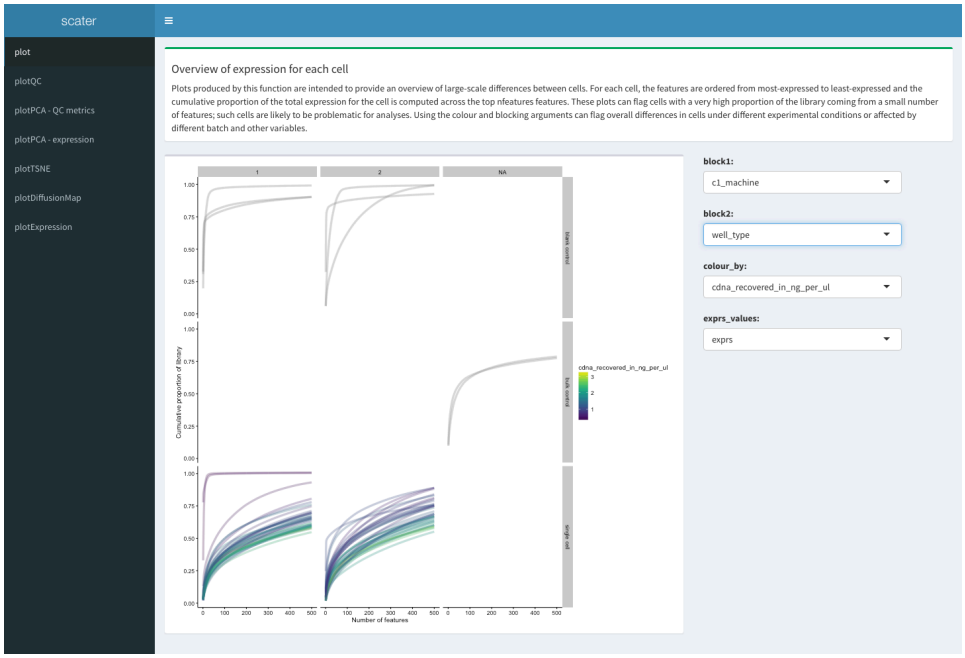

Figure 4: The landing page for the *scater* graphical user interface (GUI).

Figure 5: The `plotQC` page for the *scater* graphical user interface (GUI).

Figure 6: The `plotPCA - QC` page for the *scater* graphical user interface (GUI).

References

J. J. Allaire, J. Cheng, Y. Xie, J. McPherson, W. Chang, J. Allen, H. Wickham, A. Atkins, and R. Hyndman. rmarkdown: Dynamic Documents for R, 2016. URL https://CRAN.R-project.org/package=rmarkdown.

P. Angerer, L. Haghverdi, M. Büttner, F. J. Theis, C. Marr, and F. Buettner. destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, 14 Dec. 2015. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btv715. URL http://dx.doi.org/10.1093/bioinformatics/btv715.

S. M. Bache and H. Wickham. Magrittr: A forward-pipe operator for R. *R package version*, 2014.

H. Bengtsson. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors), 2016. URL https://CRAN.R-project.org/package=matrixStats.

W. Chang. shinydashboard: Create Dashboards with 'Shiny', 2015. URL https://CRAN.R-project.org/package=shinydashboard.

W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. shiny: Web Application Framework for R, 2016. URL https://CRAN.R-project.org/package=shiny.

A. Couture-Beil. rjson: JSON for R, 2014. URL https://CRAN.R-project.org/package=rjson.

M. Dowle, A. Srinivasan, T. Short, and S. Lianoglou. data.table: Extension of Data.frame, 2015. URL https://CRAN.R-project.org/package=data.table.

S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 15 Aug. 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti525. URL http://dx.doi.org/10.1093/bioinformatics/bti525.

P. Filzmoser and M. Gschwandtner. mvoutlier: Multivariate outlier detection based on robust methods, 2015. URL https://CRAN.R-project.org/package=mvoutlier.

B. Fischer and G. Pau. rhdf5: HDF5 interface to R, 2016.

S. Garnier. viridis: Default Color Maps from 'matplotlib', 2016. URL https://CRAN.R-project.org/package=viridis.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2):115–121, Feb. 2015. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3252. URL http://dx.doi.org/10.1038/nmeth.3252.

J. Krijthe. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation. *0.10, URL http://CRAN. R-project. org/package= Rtsne*, 2015.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47, 20 Apr. 2015. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv007. URL http://dx.doi.org/10.1093/nar/gkv007.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan. 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp616. URL http://dx.doi.org/10.1093/bioinformatics/btp616.

C. Soneson, M. I. Love, and M. D. Robinson. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4:1521, 30 Dec. 2015. ISSN 2046-1402. doi: 10.12688/f1000research.7563.1. URL http://dx.doi.org/10.12688/f1000research.7563.1.

H. Wickham. testthat: Get started with testing. *The R journal*, 3(1):5–10, 2011. URL http://www.academia.edu/download/5759380/rjournal_2011-1.pdf#page=5.

H. Wickham. reshape2: Flexibly reshape data: a reboot of the reshape package. *R package version*, 2012. URL http://cran.ms.unimelb.edu.au/web/packages/reshape2/.

H. Wickham. plyr: Tools for splitting, applying and combining data. R package version 1.8. 1. *R Found. Stat. Comput., Vienna*, 2015.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 8 June 2016. ISBN 9783319242774. URL http://books.google.co.uk/books/about/ggplot2.html?hl=&id=XgFkDAAAQBAJ.

H. Wickham and R. Francois. dplyr: A grammar of data manipulation. *R package version 0. 4*, 1:20, 2015. URL http://user2014.stat.ucla.edu/abstracts/talks/45_Wickham.pdf.

H. Wickham, P. Danenberg, and M. Eugster. roxygen2: In-Source Documentation for R, 2015. URL https://CRAN.R-project.org/package=roxygen2.

C. O. Wilke. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2', 2016. URL https://CRAN.R-project.org/package=cowplot.

Y. Xie. *Dynamic Documents with R and knitr*, volume 29. CRC Press, 2013. URL https://books.google.co.uk/books?hl=en&lr=&id=QZwAAAAAQBAJ&oi=fnd&pg=PP1&dq=knitr&ots=4Wyup9mudf&sig=mB4kbKQyrej7lHbQ0dk75j2QT4g.