BingleSeq: a user-friendly R package for bulk and single-cell RNA-Seq data analysis

Daniel Dimitrov and Quan Gu

MRC-University of Glasgow Centre for Virus Research, University of Glasgow, Glasgow, UK

ABSTRACT

Background: RNA sequencing is an indispensable research tool used in a broad range of transcriptome analysis studies. The most common application of RNA Sequencing is differential expression analysis and it is used to determine genetic loci with distinct expression across different conditions. An emerging field called single-cell RNA sequencing is used for transcriptome profiling at the individual cell level. The standard protocols for both of these approaches include the processing of sequencing libraries and result in the generation of count matrices. An obstacle to these analyses and the acquisition of meaningful results is that they require programing expertise. Although some effort has been directed toward the development of user-friendly RNA-Seq analysis analysis tools, few have the flexibility to explore both Bulk and single-cell RNA sequencing.

Implementation: BingleSeq was developed as an intuitive application that provides a user-friendly solution for the analysis of count matrices produced by both Bulk and Single-cell RNA-Seq experiments. This was achieved by building an interactive dashboard-like user interface which incorporates three state-of-the-art software packages for each type of the aforementioned analyses. Furthermore, BingleSeq includes additional features such as visualization techniques, extensive functional annotation analysis and rank-based consensus for differential gene analysis results. As a result, BingleSeq puts some of the best reviewed and most widely used packages and tools for RNA-Seq analyses at the fingertips of biologists with no programing experience.

Availability: BingleSeq is as an easy-to-install R package available on GitHub at https://github.com/dbdimitrov/BingleSeq/.

Subjects Bioinformatics, Computational Biology, Genomics, Molecular Biology
Keywords RNA-Seq, Single-cell RNA-Seq, R package, Functional annotation, Rank-based consensus, Differential expression

INTRODUCTION

About a decade ago, a transcriptome profiling approach, known as RNA Sequencing (RNA-Seq), was predicted to revolutionize transcriptome analyses (*Wang, Gerstein & Snyder, 2009*). Today, as a consequence of the continuous advancements and dropping costs of next-generation sequencing (NGS) technologies, differential expression (DE) analysis or Bulk RNA-Seq has established itself as a routine research tool (*Stark, Grzelak & Hadfield, 2019*).

Single-cell RNA Sequencing (scRNA-Seq) is an emerging field that enables the gene expression profiling at the individual cell level and it is believed to lead to the reconstruction of an entire human cell lineage tree (*Shapiro, Biezuner & Linnarsson, 2013*). scRNA-Seq's potential is highlighted by its broad range of applications which include the classification and discovery of cell subpopulations into putative transcriptomic profiles and novel cell types (*Jaitin et al., 2014*; *Macosko et al., 2015*; *Muraro et al., 2016*; *Campbell et al., 2017*; *Villani et al., 2017*), as well as the deconvolution of Bulk RNA-Seq results (*Schelker et al., 2017*; *Tirosh et al., 2016*).

Although there is a wide range of software tools available for both Bulk RNA-Seq and scRNA-Seq analyses, most require some proficiency in programing languages such as R. This creates a challenge for the analysis of RNA-seq data for a large portion of biologists lacking programing experience. Here we present an application, called BingleSeq, the primary goal of which is to enable the user-friendly analysis of count tables obtained by both Bulk RNA-Seq and scRNA-Seq protocols.
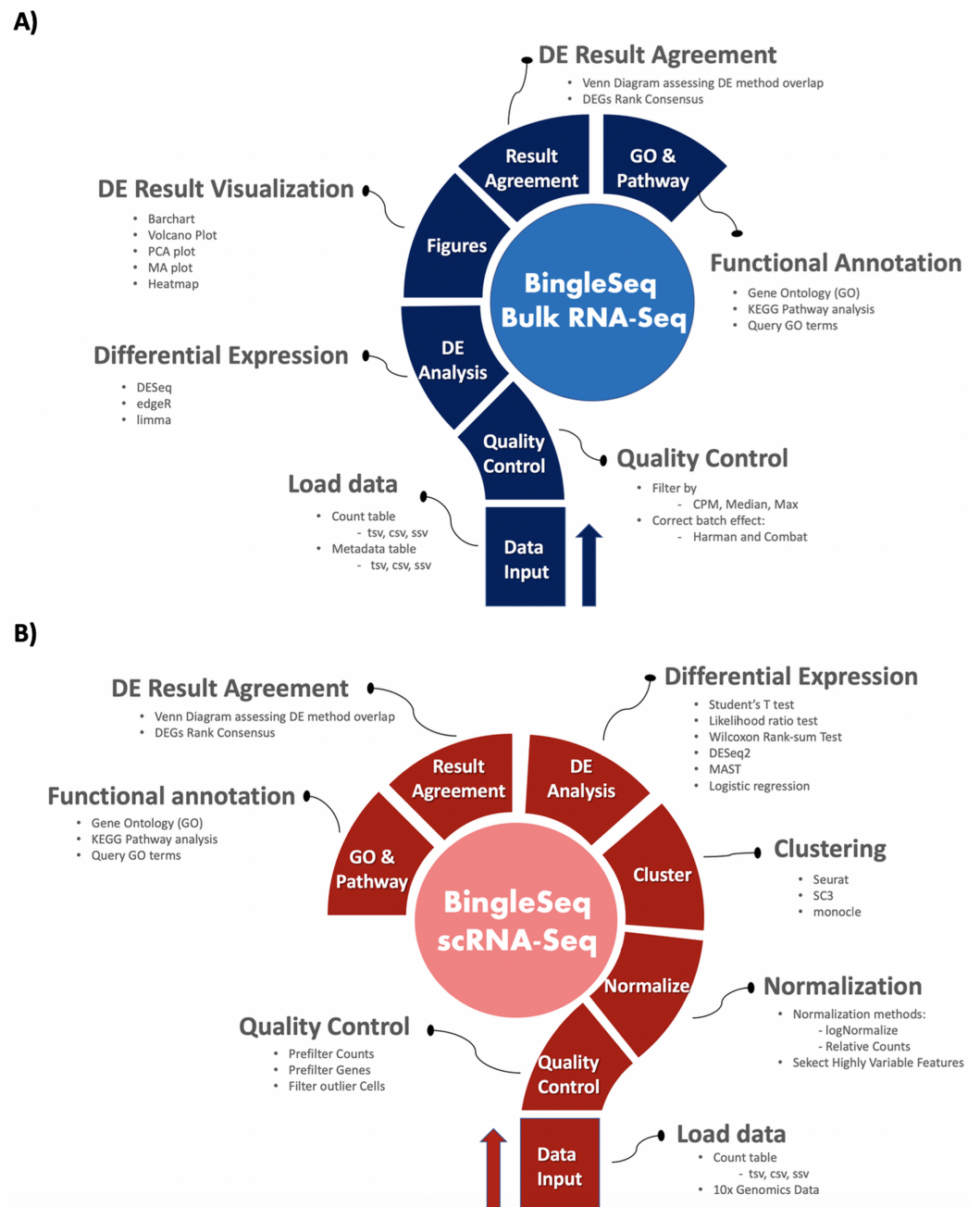
MATERIALS AND METHODS

Implementation

BingleSeq is based on shiny (*Chang et al., 2016*) and it is composed of a multi-tabbed UI, built as separate shiny modules with efficiency, code readability and reusability in mind. Each module (tab) corresponds to a key step in the typical Bulk and scRNA-Seq analysis pipelines (Fig. 1). Modules are generated only upon reaching a given step of the analyses which ensures efficiency and speed despite the complexity of the application. BingleSeq's UI components (e.g., plots, tables, tabs, etc.) make use of shiny's "reactivity" property and these components are automatically updated upon user input or any change related to a given component. The development of BingleSeq focused on providing a flexible and intuitive user experience. As such, BingleSeq implements three state-of-the art package options for each analysis, while it also enables users to tailor key analysis parameters according to their experiment.

User experience is complemented with customizable and interactive tables and plots as well as analysis-related tips and pop-up messages to guide the correct execution of the pipelines. Plot interactivity was achieved via conversion from ggplot to ggplotly (*Sievert, 2020*).

Bulk RNA-seq steps and features

To begin the DE analysis of Bulk RNA-Seq data, a count table and a metadata table must be loaded in the appropriate formats (Fig. S1). Genes can then be filtered according to counts per million (CPM), Max, or Median thresholds. Furthermore, batch effect correction can be performed with Harman and ComBat packages (*Leek et al., 2012*; *Oytam et al., 2016*) (Fig. S2).
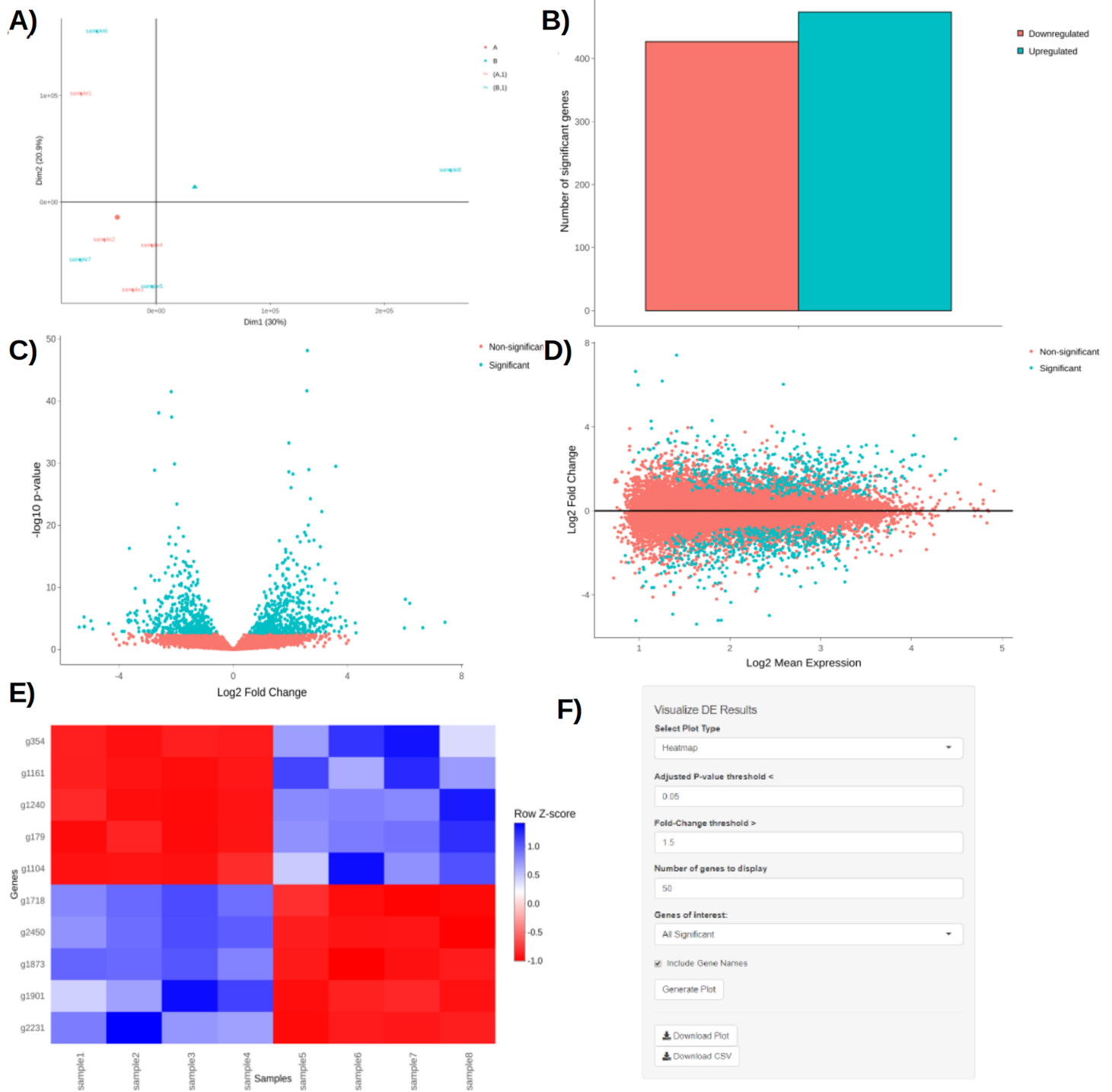
Subsequent to quality control, users can investigate differentially expressed genes (DEGs) using three state-of-the-art packages: DESeq2 (*Love, Huber & Anders, 2014*), edgeR (*Robinson, McCarthy & Smyth, 2010*), and limma (*Ritchie et al., 2015*).

Figure 1 Overview of BingleSeq's Bulk RNA-Seq (A) and scRNA-Seq (B) analysis pipelines.
Full-size ☑ DOI: 10.7717/peerj.10469/fig-1

Upon obtaining DE results, users can visualize them using interactive plotting techniques (Figs. 2A–2E). These include: a PCA plot used to provide insights about the relationship between samples; Barchart plot supplemented with a summary table used to summarize the up- and downregulated DEGs; Volcano and MA plots that assess the relationship between fold change (FC) versus significance and average expression; and Heatmaps which are arguably the most versatile and informative type of visualization technique when looking at DE results.

Figure 2 PCA plot (A), barchart (B), volcano plot (C), MA plot (D), and a heatmap (E) with its corresponding interactive control panel (F) as generated by BingleSeq. Full-size ☑ DOI: 10.7717/peerj.10469/fig-2

BingleSeq's visualization techniques were implemented with customization in mind and users can specify parameters such as *p*-value threshold and fold-change threshold, among others. Due to their versatility, heatmaps were designed as BingleSeq's most customizable plotting component (Fig. 2F).

To assess BingleSeq's Bulk RNA-Seq pipeline, we used a synthetic dataset generated with the compcodeR package (*Soneson, 2014*). We also tested it on a real data set looking at DEGs between HSV-1 infected control and interferon B treatment (*McFarlane et al., 2019*)—see Article S1.
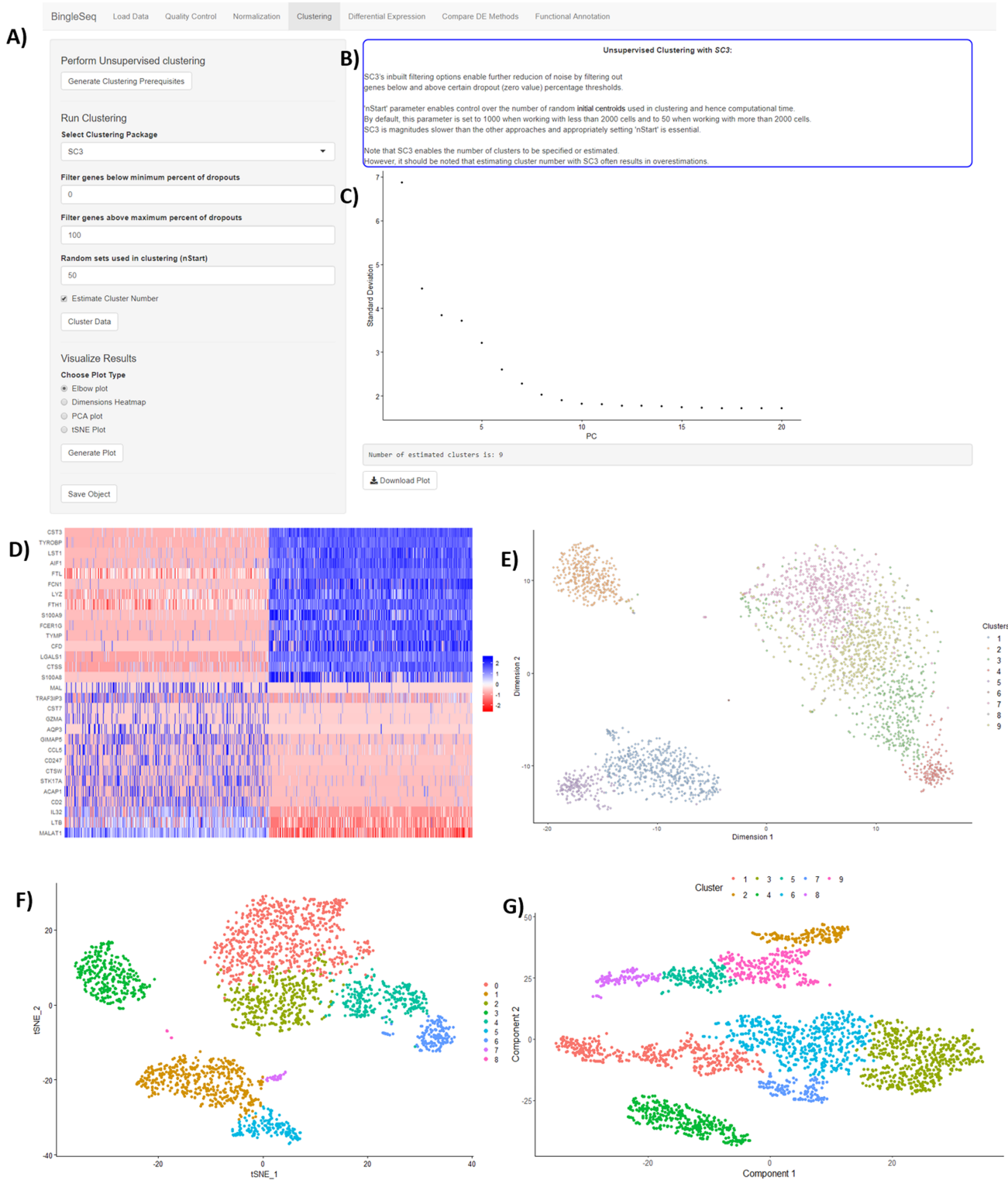
scRNA-seq pipeline steps and features

BingleSeq's scRNA-Seq part is based on Seurat's scRNA-Seq pipeline and visualizations (*Satija et al., 2015*). Nonetheless, clustering can also be performed with monocle (*Trapnell et al., 2014*) and SC3 (*Kiselev et al., 2017*) packages.

To begin scRNA-Seq analysis, data can be supplied in two formats—"Cell Ranger 10× Genomics" data and a count table in a predefined format (Fig. S3). Once the data is loaded, users can filter unwanted cells and features (Fig. S4). The next step is to normalize the data and BingleSeq provides two Seurat-supplied normalization methods—"LogNormalize" and "Relative counts". Simultaneously with normalization and scaling, the highly variable features within the dataset are identified and selected for clustering with Seurat as a way to minimize noise. The highly variable features are visualized using an interactive scatter plot (Fig. S4).
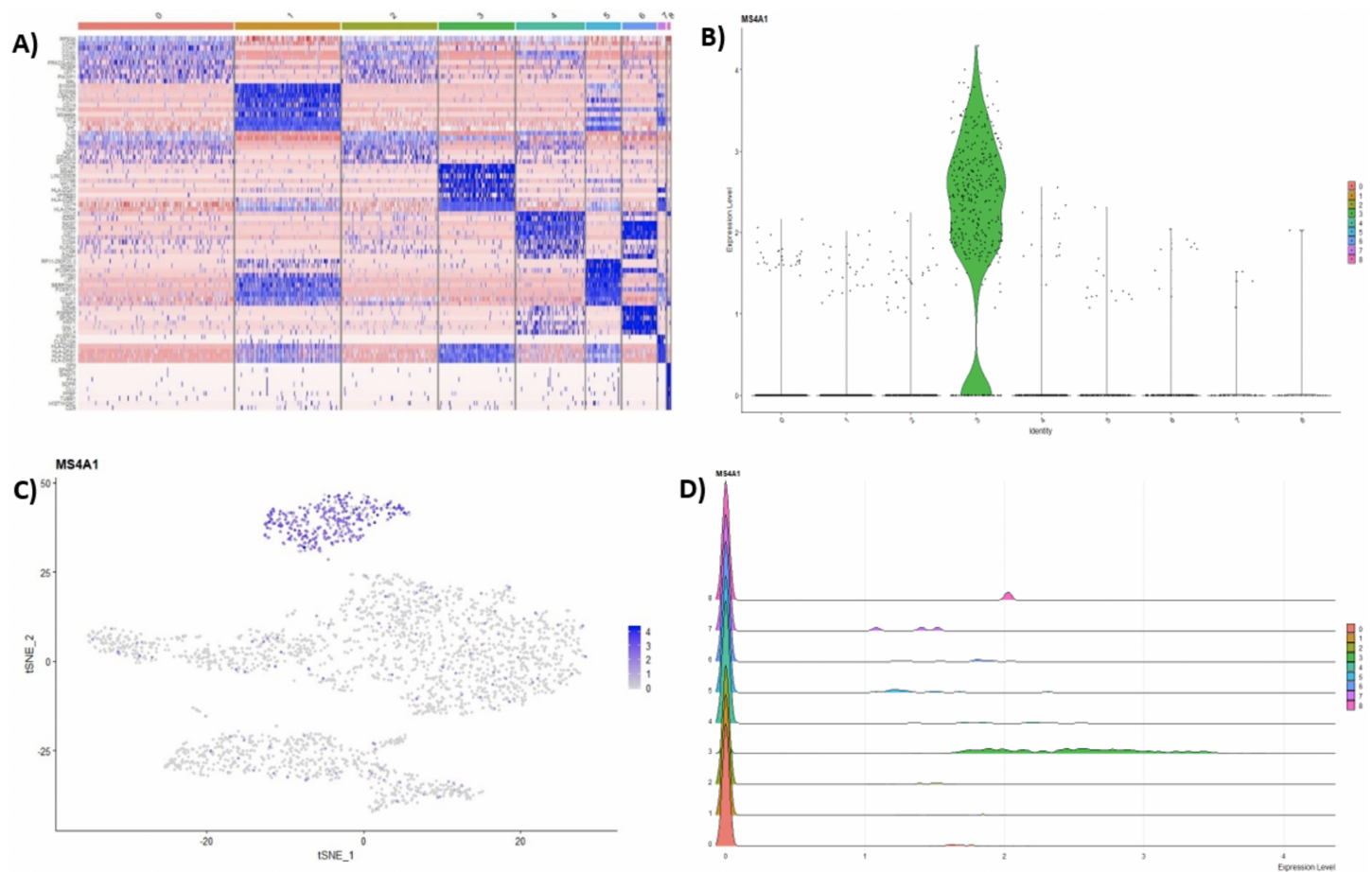
Following normalization, the "Clustering" tab is generated (Fig. 3) which provides a high degree of control over the different steps of the analysis (Fig. 3A). Additionally, general tips are provided, such as clustering advice provided for each package (Fig. 3B). Moreover, each of the visualization techniques in this tab are interactive to further facilitate the interpretation of results. First, the clustering tab requires the generation of pre-clustering prerequisites such as scaling the data and dimensionality reduction with Principal Component Analysis (PCA). Then an elbow plot is returned which is used to determine the dimensionality of the dataset (Fig. 3C)—this is essential for excluding noise when clustering with Seurat and monocle. PC heatmaps (Fig. 3D) are also available as a further tool for PC Selection. Once the data is filtered and transformed, users can proceed to unsupervised clustering with Seurat, SC3, and monocle. The primary way to visualize clustering results is via t-distributed stochastic neighbor embedding (tSNE) plots (Figs. 3E–3G)—a method designed for the purpose of visualizing high dimensional datasets (*Van der Maaten & Hinton, 2008*).

Following clustering, DE analysis can be conducted using Seurat's inbuilt testing methods to identify marker genes. The implemented Seurat DE methods include: Student's *T* test, Wilcoxon Rank Sum test, DESEq2 (*Love, Huber & Anders, 2014*), and MAST package (*Finak et al., 2015*). DE results and specific marker genes can then be visualized using Seurat's inbuilt plots. These plots include a cluster heatmap and visualizations for the exploration of specific genes via Violin, Feature, and Ridge plots (Fig. 4).

Figure 3 An overview of BingleSeq's "Clustering" tab with (A) clustering customization options, (B) general tips and advice for the selected package, (C) PC elbow plot, (D) PC heatmap. tSNE plots from Seurat (E), SC3 (F), and monocle (G).

Figure 4 (A) Heatmap showing the top 10 genes for each cluster in the 2,700 PBMCs dataset, while Violin (B), feature (C), and ridge (D) plots are shown for MS4A1 gene—a biomarker of B lymphocytes. Full-size ⬚ DOI: 10.7717/peerj.10469/fig-4

An evaluation of BingleSeq's scRNA-Seq pipeline was performed by reproducing and extending the results of Seurat's online tutorial (https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html)—see Article S2. The tutorial is based on a 10× Genomics dataset of 2,700 Peripheral Blood Mononuclear Cells (PBMCs) with ~69,000 reads per cell. The data set is available at https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k.

Functional annotation

For both Bulk and scRNA-Seq pipelines, BingleSeq enables the functional annotation of results using GOseq (*Young et al., 2010*) and transcription factor (TF) and pathway activity inference tools DoRothEA (*Garcia-Alonso et al., 2019*) and PROGENy (*Schubert et al., 2018; Holland, Szalai & Saez-Rodriguez, 2019; Holland et al., 2020*).

GOSeq is implemented in the "Gene Ontology" tab and it enables users to obtain results from KEGG pathway analysis and three types of GO categories, including "Cellular Component", "Molecular Function", and "Biological Function" (Fig. S5). The "Gene Ontology" tab can also be used to generate top 10 GO term histograms and to obtain

additional information about a given GO term using the "GO.db" package (*Carlson et al., 2019*) (Fig. S5). Note that BingleSeq supports human, mouse, drosophila, zebrafish, and *Escherichia coli* K12 strain genomes (*Carlson, 2019a*, *2019b*, *2019c*, *2019d*, *2019e*) for analysis with GOSeq.

In BingleSeq, the signed TF-target interactions from DoRothEA are coupled to the analytic Rank-based Enrichment Analysis (aREA) method from the viper package (*Alvarez et al., 2016*) and together they enable TF activity estimation (Figs. S6A and S6B). On the other hand, PROGENy uses a linear model to estimate the activity of 14 signaling pathways, built on downstream expression patterns from a compendium of perturbation experiments (Figs. S6C and S6D). DoRothEA and PROGENy are footprint-based tools and they infer activity from the expression of molecules considered to be downstream of a given TF or pathway, respectively. Both of these tools are implemented in the "Footprint Analysis" tab and are available for human and mouse data.

DE package comparison and rank-based consensus

BingleSeq supplies an option to assess the agreement between the different DE analysis methods in the form of a Venn diagram (Fig. S7A). Moreover, BingleSeq provides a rank-based consensus approach to improve confidence in DE results (*Costa-Silva, Domingues & Lopes, 2017*; *Guo et al., 2014*; *Moulos & Hatzis, 2015*)—see Fig. S7B. In the context of the Bulk RNA-Seq pipeline the overlap in DEGs is assessed on the results obtained using DESeq2, edgeR, and limma packages. In the case of scRNA-Seq, this is done using three of Seurat's inbuilt DE methods—MAST, Wilcoxon Rank Sum Test, and Student's *T* test.

Inbuilt bulk and single cell RNA-seq example datasets

BingleSeq also provides inbuilt Bulk RNA-Seq and scRNA-Seq test data. Bulk RNA-Seq data is represented by a 3-sample contrast between HSV-1 infected control and interferon B treatment (*McFarlane et al., 2019*). The data files are available from European Nucleotide Archive (ENA) under accession number PRJEB27501. The single-cell RNA-Seq example is a Cell Ranger 10× Genomics dataset looking at filtered data of 2,700 human peripheral blood mononuclear cells.

RESULTS

Some effort has already been directed towards lowering the entry requirements to RNA-Seq analyses as there are some software tools which implement UI components. However, many of these applications are limited to only some key features or particular parts of RNA-Seq analysis (*DeTomaso & Yosef, 2016*; *Kiselev et al., 2017*).

From the available software packages providing comprehensive solutions for Bulk RNA-Seq, NetworkAnalyst (*Zhou et al., 2019*), DEapp (*Li & Andrade, 2017*), DEBrowser (*Kucukural et al., 2019*), and Omics Playground (*Akhmedov et al., 2020*) were thought to provide the most extensive RNA-Seq analysis options. As seen in Tables S1 and S2, the functionality implemented BingleSeq's Bulk RNA-Seq can be argued to put our package close to even the best available Bulk RNA-Seq solutions.

When looking at similar applications providing solutions to scRNA-Seq analysis, BingleSeq is most analogous to SeuratWizard (*Yousif et al., 2020*) as both are based on Seurat's pipeline. However, by implementing SC3 and monocle, BingleSeq provides solutions to some of Seurat's inherent limitations. For instance, SeuratWizard does not implement functionality to explicitly specify the number of clusters, nor a way to estimate the number of clusters, while BingleSeq provides two distinct approaches to achieve that. Another major functionality that sets our application apart is that it enables functional annotation analysis and a way to compare and provide a consensus for Seurat's inbuilt DE methods. Consequently, BingleSeq can be argued to match even the most comprehensive scRNA-Seq applications, such as ASAP (*Gardeux et al., 2017*) and singleCellTK (*Jenkins et al., 2020*)—see Tables S2 and S3.

In terms of providing a solution to both Bulk and single-cell RNA-Seq analyses, BingleSeq's features and comprehensiveness are only contested by those of Omics Playground. However, what sets BingleSeq apart is that it provides multiple clustering packages and algorithms.

DISCUSSION

As solutions to Bulk RNA-Seq, BingleSeq implements DESeq2, edgeR, and limma. These packages are well-tested and regarded as being among the best performing ones (*Schurch et al., 2016*; *Seyednasrollah, Laiho & Elo, 2015*; *Soneson & Delorenzi, 2013*). Despite being accepted as being among the best DE analysis solutions, different studies often present contrasting conclusions. Hence, there is little consensus regarding which DE algorithm has the best performance. This stems from the fact that there is no optimal package under all circumstances and different variables are known to affect package performance, with sample size in particular (*Schurch et al., 2016*; *Soneson & Delorenzi, 2013*). As such, the method of choice depends on the dataset being analyzed.

BingleSeq's scRNA-Seq pipeline includes three unsupervised clustering solutions provided by monocle, Seurat, and SC3 packages. The latter two packages are regarded as having the best overall clustering performance (*Duò, Robinson & Soneson, 2018*; *Freytag et al., 2018*). However, similarly to packages used in the DE analysis of Bulk RNA-Seq data, there seems to be little consensus on which package provides the best-performing clustering approach. This is largely due to the inherent limitations of the different algorithms used in clustering, as a result no algorithm performs well in every circumstance (*Wiwie, Baumbach & Röttger, 2015*). *Kiselev, Andrews & Hemberg (2019)* suggest that Seurat may be inappropriate for small scRNA-Seq datasets, due to the inherent limitations of the Louvain algorithm. On the contrary, as a way to amend for the limitations of k-means clustering algorithm used in SC3, the authors implemented an extensive iterative-consensus approach, which makes SC3 magnitudes slower than Seurat and downgrades its scalability (*Duò, Robinson & Soneson, 2018*; *Kiselev, Andrews & Hemberg, 2019*). Another difference between these two packages is that Seurat does not include functionality to estimate or explicitly specify cluster number, while SC3 does.

Thus, by implementing multiple DE and clustering packages BingleSeq enables users to explore and pick the method that is most suitable for their experiment.

CONCLUSIONS

BingleSeq is a comprehensive and intuitive solution that enables users to choose from multiple state-of-the-art DE analysis and unsupervised clustering packages according to their preferences or the dataset in question. In terms of Bulk RNA-Seq analyses, BingleSeq implements functionality that puts it close to, what are to our understanding, the best available similar applications. In terms of scRNA-Seq, BingleSeq could be argued to be among the most exhaustive applications, such as ASAP (*Gardeux et al., 2017*) and singleCellTK (*Jenkins et al., 2020*). Thus, the implementation of multiple state-of-the-art packages for both types of analysis, alongside functionality that bolsters the confidence and subsequent interpretation of DE results, makes BingleSeq a solid choice for the analysis of both Bulk and scRNA-Seq.

Future work will focus on including more functional annotation options and on extending user control over the packages implemented in BingleSeq. Moreover, the implementation of both Bulk RNA-Seq and scRNA-Seq pipelines puts BingleSeq in a particularly good position to implement user-friendly deconvolution of Bulk RNA-Seq results using scRNA-Seq data. Hence, an excellent and practical conclusion to the development of BingleSeq would be to include state-of-the-art deconvolution methods such as Cell Population Mapping (CPM) (*Frishberg et al., 2019*) or MUlti-Subject SIngle Cell deconvolution (MuSiC) (*Wang et al., 2019*).

BingleSeq is as an easy-to-install R package available on GitHub at https://github.com/dbdimitrov/BingleSeq/. The application's GitHub page provides an easy guide on how to install the application as well as examples of its general applicability and an extensive description of typical workflows when working with Bulk RNA-Seq and scRNA-Seq data.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Daniel Dimitrov performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Quan Gu conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Bulk RNA-Seq data is represented by a 3-sample contrast between HSV-1 infected control and interferon B treatment (*McFarlane et al., 2019*).

The data files are available from European Nucleotide Archive: PRJEB27501.

The single-cell RNA-Seq example is a Cell Ranger 10× Genomics public dataset looking at filtered data of 2,700 Peripheral Blood Mononuclear Cells (PBMCs) with ~69,900 reads per cell (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k) and requires user registration to access. It is also available at GitHub: https://github.com/dbdimitrov/BingleSeq/tree/master/inst/extdata.

The R package source code is also available at GitHub: https://github.com/dbdimitrov/BingleSeq.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10469#supplemental-information.

REFERENCES

Akhmedov M, Martinelli A, Geiger R, Kwee I. 2020. Omics playground: a comprehensive self-service platform for visualization, analytics and exploration of big omics data. *NAR Genomics and Bioinformatics* **2(1)**:6 DOI 10.1093/nargab/lqz019.

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A. 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics* **48(8)**:838–847 DOI 10.1038/ng.3593.

Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, Goldman M, Verstegen AMJ, Resch JM, McCarroll SA, Rosen ED, Lowell BB, Tsai LT. 2017. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature Neuroscience* **20(3)**:484–496 DOI 10.1038/nn.4495.

Carlson M. 2019a. org. Hs. eg. db: genome wide annotation for E coli strain K12. R package version 3.8.2. DOI 10.18129/B9.bioc.org.EcK12.eg.db.

Carlson M. 2019b. org. Hs. eg. db: genome wide annotation for fly. R package version 3.8.2. DOI 10.18129/B9.bioc.org.Dm.eg.db.

Carlson M. 2019c. org. Hs. eg. db: genome wide annotation for human. R package version 3.8.2. DOI 10.18129/B9.bioc.org.Hs.eg.db.

Carlson M. 2019d. n.d. org. Mm. eg. db: genome wide annotation for mouse. R package version 3.8.2. Bioconductor. DOI 10.18129/B9.bioc.org.Mm.eg.db.

Carlson M. 2019e. org. Hs. eg. db: genome wide annotation for zebrafish. R package version 3.8.2. DOI 10.18129/B9.org.Dr.eg.db.

Carlson M, Falcon S, Pages H, Li N. 2019. A set of annotation maps describing the entire gene ontology. R package version 3. DOI 10.18129/B9.bioc.GO.db.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. 2016. *Shiny: web application framework for R.* R package version 0.13.2. *Available at http://CRAN.R-project.org/package=shiny.*

Costa-Silva J, Domingues D, Lopes FM. 2017. RNA-Seq differential expression analysis: an extended review and a software tool. *PLOS ONE* **12(12)**:e0190152 DOI 10.1371/journal.pone.0190152.

DeTomaso D, Yosef N. 2016. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* **17(1)**:315 DOI 10.1186/s12859-016-1176-5.

Duò A, Robinson MD, Soneson C. 2018. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research* **7**:1141 DOI 10.12688/f1000research.15666.2.

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16(1)**:278 DOI 10.1186/s13059-015-0844-5.

Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. 2018. Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data. *F1000Research* **7**:1297 DOI 10.12688/f1000research.15809.2.

Frishberg A, Peshes-Yaloz N, Cohn O, Rosentul D, Steuerman Y, Valadarsky L, Yankovitz G, Mandelboim M, Iraqi FA, Amit I, Mayo L, Bacharach E, Gat-Viks I. 2019. Cell composition analysis of bulk genomics using single-cell data. *Nature Methods* **16(4)**:327–332 DOI 10.1038/s41592-019-0355-5.

Garcia-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research* **29(8)**:1363–1375 DOI 10.1101/gr.240663.118.

Gardeux V, David FPA, Shajkofci A, Schwalie PC, Deplancke B. 2017. ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* **33(19)**:3123–3125 DOI 10.1093/bioinformatics/btx337.

Guo Y, Zhao S, Ye F, Sheng Q, Shyr Y. 2014. MultiRankSeq: multiperspective approach for RNAseq differential expression analysis and quality control. *BioMed Research International* **2014(1)**:1–8 DOI 10.1155/2014/248090.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, Amit I. 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343(6172)**:776–779 DOI 10.1126/science.1247651.

Jenkins D, Akavoor V, Alabdullatif S, Bandyadka S, Briars E, Cao X, Carrasco Pro S, Faits T, Hong R, Khan MM, Koga Y, Leshchyk A, Sarfraz I, Wang Y, Wang Z, Johnson WE, Campbell JD. 2020. singleCellTK: comprehensive and interactive analysis of single cell RNA-Seq Data. R package version 2.0.0. *Available at https://compbiomed.github.io/sctk_docs/.*

Holland C, Szalai B, Saez-Rodriguez J. 2019. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. *Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms* **1863(6)**:194431 DOI 10.1016/j.bbagrm.2019.194431.

Holland C, Tanevski J, Perales-Patón J, Gleixner J, Kumar M, Mereu E, Joughin B, Stegle O, Lauffenburger D, Heyn H, Szalai B, Saez-Rodriguez J. 2020. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biology* **21(1)**:e120 DOI 10.1186/s13059-020-1949-z.

Kiselev VY, Andrews TS, Hemberg M. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20(5)**:273–282 DOI 10.1038/s41576-018-0088-9.

Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, Hemberg M. 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods* **14(5)**:483–486 DOI 10.1038/nmeth.4236.

Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. 2019. DEBrowser: interactive differential expression analysis and visualization tool for count data. *BMC Genomics* **20(1)**:6 DOI 10.1186/s12864-018-5362-x.

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28(6)**:882–883 DOI 10.1093/bioinformatics/bts034.

Li Y, Andrade J. 2017. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code for Biology and Medicine* **12(1)**:2 DOI 10.1186/s13029-017-0063-4.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15(12)**:550 DOI 10.1186/s13059-014-0550-8.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161(5)**:1202–1214 DOI 10.1016/j.cell.2015.05.002.

McFarlane S, Orr A, Roberts APE, Conn KL, Iliev V, Loney C, Da Silva Filipe A, Smollett K, Gu Q, Robertson N, Adams PD, Rai TS, Boutell C. 2019. The histone chaperone HIRA promotes the induction of host innate immune defences in response to HSV-1 infection. *PLOS Pathogens* **15(3)**:e1007667 DOI 10.1371/journal.ppat.1007667.

Moulos P, Hatzis P. 2015. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Research* **43(4)**:e25 DOI 10.1093/nar/gku1273.

Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, Van Gurp L, Engelse MA, Carlotti F, De Koning EJP, Van Oudenaarden A. 2016. A single-cell transcriptome atlas of the human pancreas. *Cell Systems* **3(4)**:385–394.e3 DOI 10.1016/j.cels.2016.09.002.

Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. 2016. Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* **17(1)**:332 DOI 10.1186/s12859-016-1212-5.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43(7)**:e47 DOI 10.1093/nar/gkv007.

Robinson MD, McCarthy DJ, Smyth GK. 2010. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26(1)**:139–140 DOI 10.1093/bioinformatics/btp616.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33(5)**:495–502 DOI 10.1038/nbt.3192.

Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, Schoeberl B, Raue A. 2017. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Communications* **8(1)**:1–12 DOI 10.1038/s41467-017-02289-3.

Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N, Saez-Rodriguez J. 2018. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications* **9(20)**:57 DOI 10.1038/s41467-017-02391-6.

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22(6)**:839–851 DOI 10.1261/rna.053959.115.

Seyednasrollah F, Laiho A, Elo LL. 2015. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* **16(1)**:59–70 DOI 10.1093/bib/bbt086.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14(9)**:618–630 DOI 10.1038/nrg3542.

Sievert C. 2020. *Interactive web-based data visualization with R, plotly, and shiny*. London: Chapman and Hall/CRC.

Soneson C. 2014. CompcodeR—an R package for benchmarking differential expression methods for RNA-seq data. *Bioinformatics* **30(17)**:2517–2518 DOI 10.1093/bioinformatics/btu324.

Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14(1)**:91 DOI 10.1186/1471-2105-14-91.

Stark R, Grzelak M, Hadfield J. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20(11)**:631–656 DOI 10.1038/s41576-019-0150-2.

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin J-R, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CGK, Kazer SW, Gaillard A, Kolb KE, Villani A-C, Johannessen CM, Andreev AY, Allen EMV, Bertagnolli M, Sorger PK, Sullivan RJ, Flaherty KT, Frederick DT, Jané-Valbuena J, Yoon CH, Rozenblatt-Rosen O, Shalek AK, Regev A, Garraway LA. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352(6282)**:189–196 DOI 10.1126/science.aad0501.

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32(4)**:381–386 DOI 10.1038/nbt.2859.

Van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**:2579–2605.

Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, Griesbeck M, Butler A, Zheng S, Lazo S, Jardine L, Dixon D, Stephenson E, Nilsson E, Grundberg I, McDonald D, Filby A, Li W, Jager PLD, Rozenblatt-Rosen O, Lane AA, Haniffa M, Regev A, Hacohen N. 2017. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356(6335)**:eaah4573 DOI 10.1126/science.aah4573.

Wang X, Park J, Susztak K, Zhang NR, Li M. 2019. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* **10(1)**:1–9 DOI 10.1038/s41467-018-07882-8.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10(1)**:57–63 DOI 10.1038/nrg2484.

Wiwie C, Baumbach J, Röttger R. 2015. Comparing the performance of biomedical clustering methods. *Nature Methods* **12(11)**:1033–1038 DOI 10.1038/nmeth.3583.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11(2)**:R14 DOI 10.1186/gb-2010-11-2-r14.

Yousif A, Drou N, Rowe J, Khalfan M, Gunsalus KC. 2020. NASQAR: a web-based platform for high-throughput sequencing data analysis and visualization. *BMC Bioinformatics* **21**:267 DOI 10.1186/s12859-020-03577-4.

Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. 2019. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research* **47(W1)**:W234–W241 DOI 10.1093/nar/gkz240.