SOFTWARE

NASQAR: A web-based platform for high-throughput sequencing data analysis and visualization

Ayman Yousif[1], Nizar Drou[1], Jillian Rowe[1], Mohammed Khalfan[2] and Kristin C Gunsalus[1,2*]

*Correspondence: kcg1@nyu.edu
[1]NYU Abu Dhabi Center for Genomics & Systems Biology, Division of Biological Sciences, Abu Dhabi, United Arab Emirates
[2]Center for Genomics & Systems Biology, Department of Biology, New York University, 10003 New York, United States
Full list of author information is available at the end of the article

Abstract

Background: As high-throughput sequencing applications continue to evolve, the rapid growth in quantity and variety of sequence-based data calls for the development of new software libraries and tools for data analysis and visualization. Often, effective use of these tools requires computational skills beyond those of many researchers. To lower this computational barrier, we have created a dynamic web-based platform, NASQAR (Nucleic Acid SeQuence Analysis Resource).

Results: NASQAR offers a collection of custom and publicly available open-source web applications that make extensive use of the Shiny R package to provide interactive data visualization. The platform is publicly accessible at http://nasqar.abudhabi.nyu.edu/. NASQAR is open-source and the code is available through GitHub at https://github.com/nasqar/NASQAR. NASQAR is also available as a Docker image at https://hub.docker.com/r/aymanm/nasqarall. NASQAR is a collaboration between the core bioinformatics teams of the NYU Abu Dhabi and NYU New York Centers for Genomics and Systems Biology.

Conclusions: NASQAR provides an intuitive interface that allows users to easily and efficiently explore their data in an interactive way using popular tools for a variety of applications, including Transcriptome Data Preprocessing, RNAseq Analysis (including Single-cell RNAseq), Metagenomics, and Gene Enrichment.

Keywords: Transcriptomics; Graphical user interface; Interactive visualization; Exploratory data analysis

Background

Genomic data has experienced tremendous growth in recent years due to the rapid advancement of Next Generation Sequencing (NGS) technologies [1, 2]. Common applications include transcriptome profiling; de novo genome sequencing; metagenomics; and mapping of genomic variation, transcription factor binding sites, chromatin modifications, chromatin accessibility, and 3D chromatin conformation. Single-cell versions of these (e.g. [3]) and newer methods — such as spatial transcriptomics (e.g. [4]), CRISPR-based screens (e.g. [5]), and multi-modal profiling (simultaneous quantification of proteins and mRNAs, e.g. [6]) — are rapidly evolving, resulting in continuous churn in the field as new techniques and applications come on the scene (e.g. [7, 8]). As the volume of data and diversity of applications continue to grow, so does the number of software libraries and tools for the analysis

and visualization of these datasets. Many of the available tools for genomic data analysis require computational experience and lack a graphical user interface (GUI), making them inaccessible to many researchers whose work depends on them. Some of the common challenges include:

- Knowledge and experience in various programming/scripting languages (R, Python, Shell, etc.)
- Data munging: pre-processing and reformatting for use with specific tools
- Limited computational resources (cpu, memory, and disk storage)
- Installation of software packages and dependencies. Many required tasks can be time consuming and tedious due to issues such as satisfying software or hardware requirements and resolving software dependencies. Moreover, the rapid churn of operating system updates and hardware configurations contributes to the decline of a tool's usability and lifetime [9].

To address these challenges, we have implemented NASQAR: Nucleic Acid SeQuence Analysis Resource, a web-based platform that that wraps popular high-level analysis and visualization tools in an intuitive and appealing interface. In addition to providing the ability to explore data using various tools and visualization methods, NASQAR also enables users to download analysis results such as normalized count data and a wide array of figures (PCA plots, heatmaps, dendograms, etc.). NASQAR thus lowers the entry barrier and provides greater independence for researchers with little or no computational experience to carry out standard bioinformatics analysis, visualize their data, and produce publication-ready data files and figures.

NASQAR Platform and Implementation

The architectural framework of the NASQAR web platform is illustrated in Figure 1. NASQAR has been deployed on a cluster of virtual machines and is publicly accessible at `http://nasqar.abudhabi.nyu.edu/`. Docker [10] and Swarm provide containerization and cluster management, and the Traefik reverse proxy / load balancer (`https://traefik.io/`) manages requests and maintains sticky user sessions, which is essential for hosting Shiny applications. This framework allows access to multiple users concurrently while providing sufficient resources (RAM/CPU) for the applications. The scalable design makes it relatively easy to increase dedicated resources simply by adding more nodes to the cluster to accommodate growth in computational demand as new applications are deployed and/or the user base expands.

A Docker image of NASQAR is publicly available through DockerHub and can be used to deploy the application seamlessly on any system with Docker installed, whether a local computer or a public server. In addition, each application can be installed and launched on its own, saving users from the hassle of satisfying the different software and hardware requirements. The source code is available publicly on GitHub and is actively maintained. All applications have clear user guides with example data sets to help users get started and acclimate quickly.

NASQAR's collection of applications is primarily implemented in R, a widely used and freely available statistical programming language [11]. Most of the analysis workflows are built using R libraries for genomics and computation. The front-end

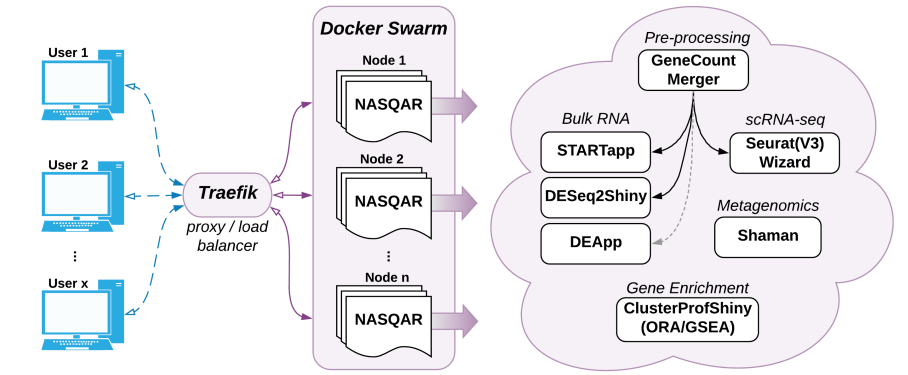

Figure 1 NASQAR Platform Architecture. A cluster of virtual machines at NYU Abu Dhabi serves NASQAR applications to multiple concurrent users. Applications are containerized and managed on the cluster using Docker and Swarm, while Traefik load-balances requests among available server nodes. Functionality includes merging gene counts, conversion of gene IDs to gene names, analysis of differential mRNA expression, metagenomics analysis, and functional enrichment analysis. Applications for bulk expression analysis include DESeq2, limma, and EdgeR. Single-cell RNAseq analysis with Seurat Wizards is built on top of the Seurat R package and includes options for filtering, normalization, dimensionality reduction (PCA), clustering, and t-SNE. Enrichment analysis includes applications for Gene Set Enrichment Analysis (GSEA) and Over-representation Analysis (ORA) built using the clusterProfiler R package.

design utilizes the R Shiny [12] library and is supported by JavaScript/CSS to enhance usability and improve overall user experience.

In addition to previously published software, we introduce here several new applications we have developed that wrap around popular analysis packages, such as DESeq2 [13] and Seurat [14, 15] for bulk and single-cell RNA-seq analysis and visualization, respectively. Since most of the analysis applications in NASQAR require a matrix of gene counts as input, we have also built a convenient tool to assist with preprocessing, GeneCountMerger. Some of the applications have been integrated to provide a seamless transition from data preprocessing to downstream analysis. This implementation gives users the option of using multiple analysis applications without having to modify/reformat the input data set, thus allowing them to easily benchmark and compare the performance of different analysis software packages.

Results

NASQAR currently hosts tools for merging gene counts; conversion of gene IDs to gene names; and analysis of differential mRNA expression, gene function enrichment, and metagenomic profiling. Packages for bulk RNA-seq analysis include DESeq2 and others, while single-cell analysis is driven by Seurat. Custom resources developed for NASQAR are summarized briefly below (see Supplementary Materials for details on available open-source and custom applications with example use cases).

GeneCountMerger

This preprocessing tool is used to merge individual raw gene count files produced from software such as htseq-count [16] or featurecounts [17] (Figure 2). Options include:

- Merge individual sample count files into one matrix
- Merge multiple raw count matrices
- Convert Ensembl gene IDs to gene names
- Select from available genomes / versions
- Add pseudocounts
- Rename sample column headers
- Download merged counts file in .csv format
- Seamless transcriptome analysis following count merger (Seurat Wizard for single-cell RNA analysis; DESeq2Shiny or START [18] for bulk RNA analysis)

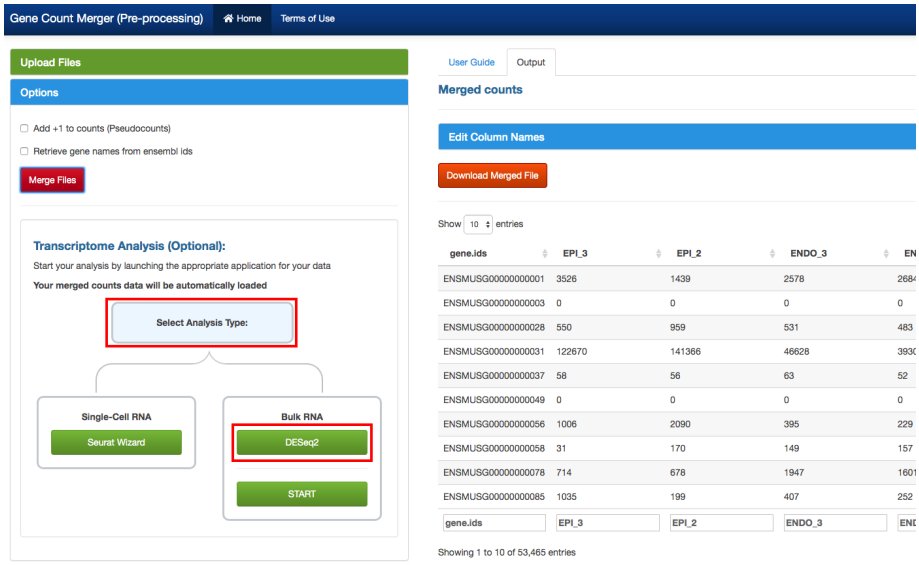

Figure 2 GeneCountMerger screenshot. a preprocessing utility to generate the gene count matrices required as input to many analysis tools. It can merge individual raw gene count files from htseq-count and other similar applications. Convenient features include conversion of Ensembl gene IDs to gene names for reference genomes and seamless launching of downstream analysis applications.

Seurat Wizards

Seurat Wizards are wizard-style web-based interactive applications to perform guided single-cell RNA-seq data analysis and visualization (Figure 3). They are based on Seurat, a popular R package designed for QC, analysis, and exploration of single-cell RNAseq data. The wizard style makes it intuitive to go back and forth between steps and adjust parameters based on the results/feedback of different outputs/plots/steps, giving the user the ability to interactively tune the analysis. SeuratWizard and SeuratV3Wizard implementations provide support for Seurat versions 2 and 3, respectively.

Another web-based tool for scRNA-seq analysis, IS-CellR [19], has recently been described that also utilizes Seurat v2. The SeuratWizard and SeuratV3Wizard take a different approach to design and implementation and follow closely the Seurat Guided Clustering Tutorials devised by the authors (https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html). Users can follow the tutorials while using

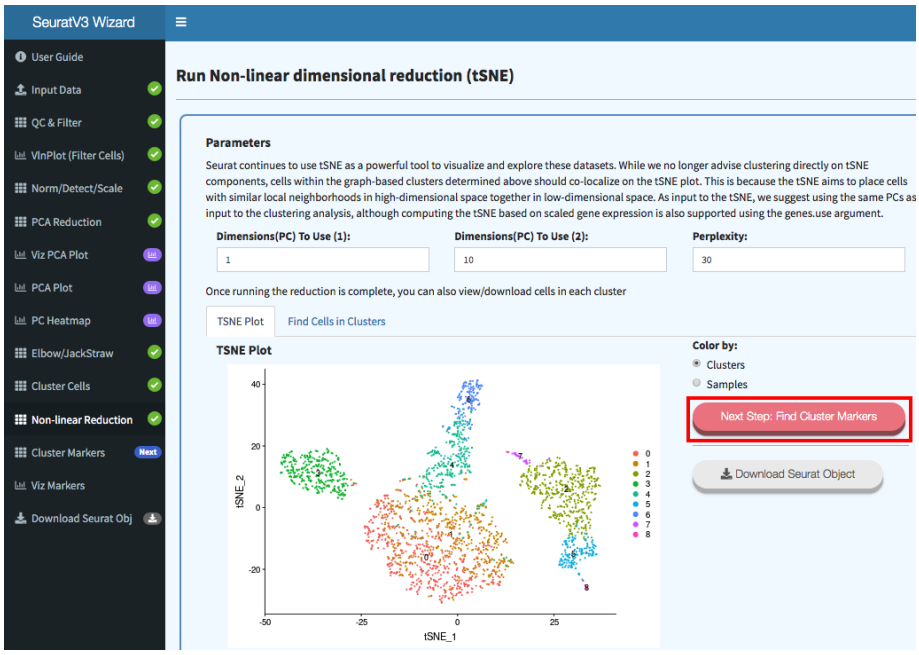

Figure 3 Seurat Wizard screenshot. Wizard-style web-based interactive applications based on Seurat, a popular R package designed for QC, analysis,and exploration of single-cell RNA-seq data. The wizards guide users through single-cell RNA-seq data analysis and visualization and provide an intuitive way to fine-tune parameters using feedback from results at each stage of the analysis. Functionality includes filtering, normalization, dimensionality reduction (PCA), clustering, and t-SNE

the Wizards and can edit parameters at almost every step, which is instrumental in producing accurate results. A unique feature of the Seurat Wizards is that they can accept as input processed 10X Genomics data files in place of a matrix of gene counts, which eliminates the need for this additional pre-processing step. SeuratV3Wizard integrates several additional features like the UCSC Cell Browser (https://github.com/maximilianh/cellBrowser), enabling users to interactively visualize clusters and gene markers, and the newly published sctransform method [20], which gives users the ability to run the analysis using two slightly different workflows and compare the results. These differences in features and design give the Seurat Wizards more versatility and improve usability in comparison with other publicly available implementations of Seurat.

DESeq2Shiny

The DESeq2Shiny app is a Shiny wrapper around DESeq2, a popular R package for performing differential mRNA expression analysis of RNA-seq data (Figure 4). This web-based application implements the standard default workflow outlined in the DESeq2 Bioconductor tutorial (https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html). This includes normalization, data transformation (e.g., rlog and vst for clustering), and estimation for dispersion and log fold-change. This app follows the same implementation as other apps on NASQAR, whereby users can fine tune the analysis parameters interactively.

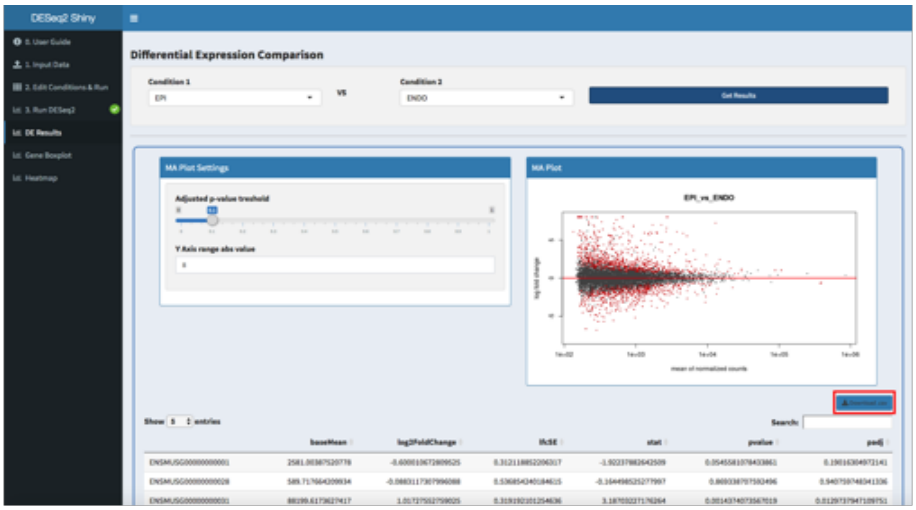

Figure 4 DESeq2Shiny screenshot. A web-based shiny wrapper around DESeq2, a popular R package for performing differential mRNA expression analysis of RNA-seq data.

ClusterProfShiny

The ClusterProfilerShiny apps wrap the popular clusterProfiler [21] package, which implements methods to analyze and visualize functional profiles of genomic coordinates, genes, and gene clusters (Figure 5). Users can upload their own data from DESeq2 or import data from the upstream Deseq2Shiny app. These apps allow for quick and easy over-representation analysis (ORA) and gene set enrichment analysis (GSEA) of GO terms and KEGG pathways. Visuals produced include dot plots, word clouds, category net plots, enrichment map plots, GO induced graphs, GSEA plots, and enriched KEGG pathway plots using the Pathview [22] package.

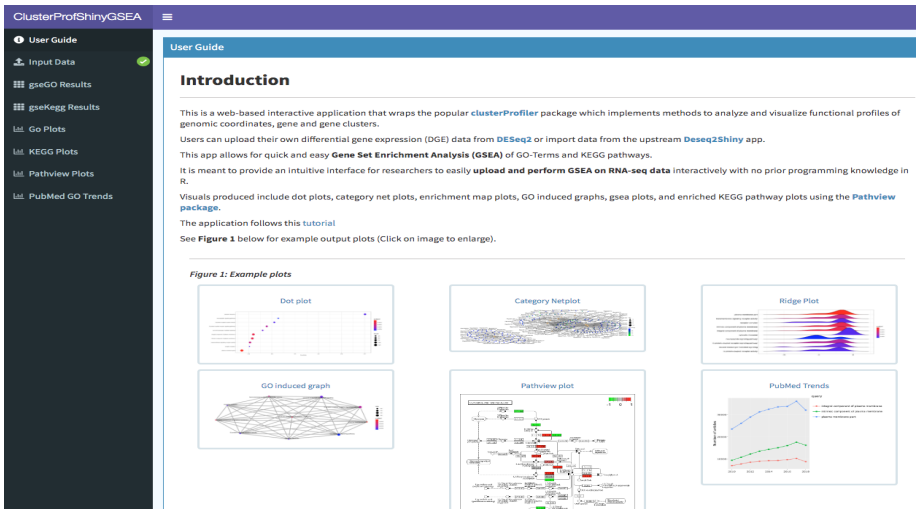

Figure 5 ClusterProfShinyGSEA screenshot. Web-based apps wrap the popular R package clusterProfiler for the analysis and visualization of functional themes and enrichment among gene clusters, using data from either DESeq2 or DESeq2Shiny. Both Gene Set Enrichment Analysis (GSEA) and Over-Representation Analysis (ORA) are implemented.

Other open-source apps

- **START**: a web-based RNA-seq analysis and visualization resource. We have modified this application slightly from the published version to add options to some plots. We have also integrated it with GeneCountMerger so that once merging gene counts is complete, users can launch the START app and have their merged matrix data loaded automatically.
- **DEApp** [23]: an interactive web application for differential expression analysis.
- **Shaman** [24]: a Shiny application that enables the identification of differentially abundant genera within metagenomic datasets. It wraps around the Generalized Linear Model implemented in DESeq2. It includes multiple visualizations, and is compatible with common metagenomic file formats.

Conclusion

The NASQAR platform offers a publicly available, comprehensive suite of interactive bioinformatics analysis and visualization tools that is accessible to all researchers with or without computational experience. Our aim is to continue expanding this toolbox to include new analysis and visualization categories. NASQAR is actively developed and will continue to offer user support. Future work will also focus on deploying NASQAR on cloud computing platforms including Kubernetes on AWS.

Availability and requirements

Project name: NASQAR
Project home page: https://github.com/nasqar/NASQAR
Operating system(s): Platform independent
Programming language: R, JavaScript
Other requirements: Docker (version >= 17.03.0-ce)
License: GNU GPL.
Any restrictions to use by non-academics: none

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
NASQAR is publicly accessible at http://nasqar.abudhabi.nyu.edu/. The platform is available as a Docker image at https://hub.docker.com/r/aymanm/nasqarall. NASQAR is open-source and the code is available through GitHub:
NASQAR (main page): https://github.com/nasqar/NASQAR
SeuratV3Wizard (scRNA): https://github.com/nasqar/seuratv3wizard
SeuratWizard (scRNA): https://github.com/nasqar/SeuratWizard
deseq2shiny (Bulk RNA): https://github.com/nasqar/deseq2shiny
GeneCountMerger (Pre-processing): https://github.com/nasqar/GeneCountMerger
ClusterProfShinyGSEA (Enrichment): https://github.com/nasqar/ClusterProfShinyGSEA
ClusterProfShinyORA (Enrichment): https://github.com/nasqar/ClusterProfShinyORA

Competing interests
The authors declare that they have no competing interests.

Author's contributions
AY carried out the interface design and software development. ND defined platform requirements, contributed scripts and extensive software testing. JR contributed to the platform architecture design. MK contributed to the development of enrichment applications and provided guidance and extensive software testing. KCG supervised the project. All authors contributed to writing the manuscript. All authors approved the final version of the manuscript.

Acknowledgements

The authors would like to acknowledge all the faculty and researchers in the NYU Abu Dhabi CGSB and Division of Biology for their excellent feedback, which has motivated the development of NASQAR. The authors would also like to acknowledge David Gresham and Siyu Sun (NYU New York CGSB) for their guidance during the development of the enrichment applications.

This research was carried out on the High Performance Computing resources at New York University Abu Dhabi. We extend special thanks to Fayizal Kunhi, NYU Abu Dhabi HPC.

Funding

This work was supported by a grant from the NYU Abu Dhabi Research Institute to the NYU Abu Dhabi Center for Genomics and Systems Biology (CGSB).

Author details

[1]NYU Abu Dhabi Center for Genomics & Systems Biology, Division of Biological Sciences, Abu Dhabi, United Arab Emirates. [2]Center for Genomics & Systems Biology, Department of Biology, New York University, 10003 New York, United States.

References

1. Goodwin, S., McPherson, J.D., McCombie, W.R.: Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics **17**, 333 (2016)
2. Wetterstrand, K.: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Accessed on 07.08.2019. https://www.genome.gov/sequencingcostsdata
3. Zheng, M., Tian, S.Z., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J.J., Li, Z., Wong, C.H., Ngan, C.Y., Wang, P., Ruan, X., Wei, C.-L., Ruan, Y.: Multiplex chromatin interactions with single-molecule precision. Nature **566**(7745), 558–562 (2019)
4. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén, J.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
5. Canver, M.C., Haeussler, M., Bauer, D.E., Orkin, S.H., Sanjana, N.E., Shalem, O., Yuan, G.-C., Zhang, F., Concordet, J.-P., Pinello, L.: Integrated design, execution, and analysis of arrayed and pooled crispr genome-editing experiments. Nature Protocols **13**, 946 (2018)
6. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P.: Simultaneous epitope and transcriptome measurement in single cells. Nature Methods **14**, 865 (2017)
7. Stuart, T., Satija, R.: Integrative single-cell analysis. Nature Reviews Genetics **20**(5), 257–272 (2019)
8. Mimitou, E.P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., Satija, R., Sanjana, N.E., Koralov, S.B., Smibert, P.: Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. Nature Methods **16**(5), 409–412 (2019)
9. Mangul, S., Martin, L.S., Eskin, E., Blekhman, R.: Improving the usability and archival stability of bioinformatics software. Genome Biology **20**(1), 47 (2019)
10. Merkel, D.: Docker: Lightweight linux containers for consistent development and deployment. Linux J. **2014**(239) (2014)
11. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). R Foundation for Statistical Computing. https://www.R-project.org/
12. Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J.: Shiny: Web Application Framework for R. (2018). R package version 1.1.0. https://CRAN.R-project.org/package=shiny
13. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**, 550 (2014)
14. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology **36**, 411 (2018)
15. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M.I., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. Cell **177**(7), 1888–1902 (2019)
16. Anders, S., Pyl, P.T., Huber, W.: HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics **31**(2), 166–169 (2014)
17. Smyth, G.K., Shi, W., Liao, Y.: featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**(7), 923–930 (2013)
18. Sklenar, J., Nelson, J.W., Minnier, J., Barnes, A.P.: The START App: a web-based RNAseq analysis and visualization resource. Bioinformatics **33**(3), 447–449 (2016)
19. Patel, M.V.: iS-CellR: a user-friendly tool for analyzing and visualizing single-cell RNA sequencing data. Bioinformatics **34**(24), 4305–4306 (2018)
20. Hafemeister, C., Satija, R.: Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. bioRxiv (2019)
21. Yu, G., Wang, L.-G., Han, Y., He, Q.-Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology **16**(5), 284–287 (2012)
22. Luo, W., Brouwer, C.: Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics **29**(14), 1830–1831 (2013)
23. Li, Y., Andrade, J.: Deapp: an interactive web interface for differential expression analysis of next generation sequence data. Source Code for Biology and Medicine **12**(1), 2 (2017)
24. Quereda, J.J., Dussurget, O., Nahori, M.-A., Ghozlane, A., Volant, S., Dillies, M.-A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B., Cossart, P., Pizarro-Cerda, J.: Bacteriocin from epidemic listeria strains alters the

host intestinal microbiota to favor infection. Proceedings of the National Academy of Sciences **113**(20), 5706–5711 (2016)

Additional Files

Additional file 1 — SupplementaryMaterials-NASQAR-final.pdf
This file includes supplementary materials such as instructions on how to launch NASQAR and example use cases on data analysis and visualization.