bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SIP: An Interchangeable Pipeline for scRNA-seq Data Processing

Sijie Chen[1,+], Zheng Wei[1,+], Yang Chen[1], Kui Hua[1], Wei Zhang[1], Changyi Liu[1], Haoxiang Gao[1], Hao Sun[1], Zhenyi Wang[1], Qijin Yin[1], Shengquan Chen[1], Shaoming Song[1], Chen Feng[1], Hairong Lu[1], Rui Jiang[1], Xiaowo Wang[1], Jin Gu[1] and Xuegong Zhang[1,2*]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division, BNRist; Department of Automation and Center for Synthetic & Systems Biology, Tsinghua University, Beijing 100084, China, [2] School of Life Sciences, Tsinghua University, Beijing 100084, China.

*To whom correspondence should be addressed.

Contact: zhangxg@tsinghua.edu.cn

Abstract

Multiple steps of bioinformatics processing are needed to convert the raw scRNA-seq data to information that can be used in downstream analyses and in building cell atlases. Dozens of software packages have been developed and different labs tend to have different preferences on choices of the workflow. Such diversity can cause difficulties in future efforts of aggregating data from multiple labs, and also difficulties for new labs to start in this field. A few pipelines have been developed to help integrating multiple steps into a whole, but the fixed software architecture makes it hard for developers to add new features or exchange parts in the pipeline.

We presented SIP, a Single-cell Interchangeable Pipeline. It is a one-stop platform for the processing of scRNA-seq data from multiple platforms, and will also support for other types of data like scATAC-seq data. SIP utilizes container technology to solve the deployment dilemma when handling multiple packages and provides an easy-to-use interface for users to conduct the complicated multi-step process from raw data to final results with a single command. It also allows advanced users to assemble different versions of the pipeline by interchanging parts or adding new modules. SIP is available at https://github.com/XuegongLab/SIP under the GPL-3.0 license.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

Analysis pipelines for single-cell RNA sequencing (scRNA-seq) gain increasing attentions as scRNA-seq is becoming more and more important in biological studies (Wang and Navin, 2015; Hedlund and Deng, 2018; Wu *et al.*, 2017). ScRNA-seq is also one of the key technologies in the collaborative effort of building the human cell atlas (HCA)(Rozenblatt-Rosen *et al.*, 2017). Automated or semi-automated bioinformatics pipelines following standardized framework are one of the fundamental elements to support the effort.

To date, a group of pipeline and software packages have been developed for analysis of single-cell data. Some tools like simpleSingleCell (Lun *et al.*, 2016) guide users to manipulate data manually and require heavy programming skills. Some packages, such as Seurat (Butler *et al.*, 2018) , Monocle 2 (Trapnell *et al.*, 2014; Qiu *et al.*, 2017) , Scater (McCarthy *et al.*, 2017), ASAP (Gardeux *et al.*, 2017), and Granatum (Zhu *et al.*, 2017), mainly focused on the downstream analyzing steps after the gene expression quantification. However, the quantification of gene expression from raw data is not a trivial task. The alignment process often consumes substantial computational resources and the computation parameters correspond to different protocols can be very tricky. Therefore, a pipeline that starts from the raw data to downstream analysis is still highly demanded, especially for labs that just start to work in this field, and for labs that need to integrate their data with those from other labs. In addition, as the sequencing throughput is quickly getting higher, a scalable pipeline capable of

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

utilizing the power of HPC (High-Performance Computing) clusters is in urgent need for labs and consortiums that generate massive single-cell data.

To meet these needs, we developed a Single-cell Interchangeable Pipeline called SIP. It is an open-source pipeline for processing from raw data to final analyses with standard framework and interchangeable components. SIP contains seven main modules, including Read QC and preprocessing, (Quasi-)alignment and transcript assembly, Gene-level abundance counting, Cluster finding, Trajectory analysis, and DEG analysis. Each module contains multiple encapsulated tools. The tools can be easily replaced with alternatives, and new tools and/or modules can also be added.

SIP is constructed with Nextflow (DI Tommaso *et al.*, 2017), which seamlessly take over outputs and inputs of the upstream and downstream modules, significantly alleviating the data manipulation efforts for users. SIP adopts the container technology to deploy the pipeline. Several useful features including e-mail notification and HPC cluster support make it more convenient to process massive amount of data using SIP. Besides, it is easy to extend SIP's function as Nextflow has a large developer community.

Excellent computational tools constantly appear in the field of single-cell data analysis. To keep pace with the development of new tools, a good pipeline should have an open software architecture for changes and be endorsed by a development community so that new tools can be integrated efficiently. The framework used in SIP allows very easy maintaining

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and updating the pipeline, and more potential Nextflow developers can join the improvement of SIP in a collaborative manner.

2 Methods

SIP prepares a curated list of single-cell RNA-seq data processing tools. The input and output of each tool are parsed by the Nextflow framework and transferred automatically. As shown in Fig. 1, users can choose the interchangeable parts for a specific step. With different combination of components, workflows of different steps are established to process the data. Users can easily explore proper tool sets for different datasets, and make a performance comparison on different datasets.

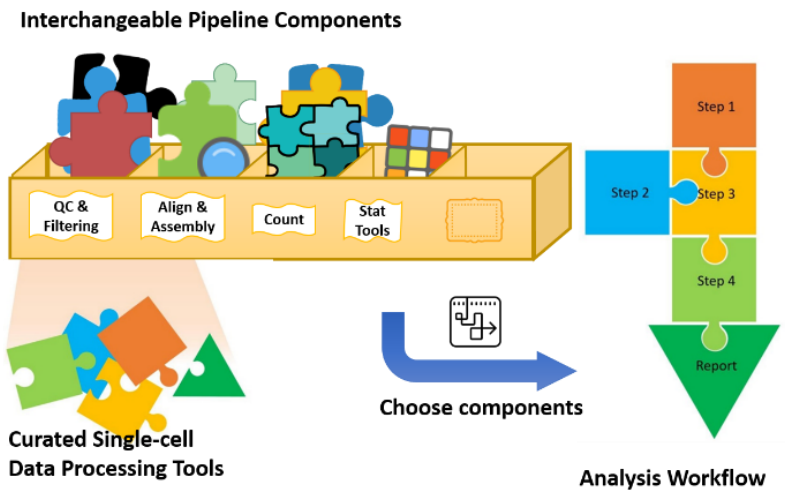


Fig. 1. Concept map of SIP. A group of software and packages are evaluated and selected to build up interchangeable parts in the pipeline toolbox. Users can freely choose the tools to process the data without concerning the tedious intermediate data manipulation operations.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.1 The SIP workflow

2.1.0 Feed raw sequencing data to pipeline

SIP accepts raw sequencing data (fastq files) as input. Read files should be named after identifiers of the cells and placed in the cell home folder. The pipeline detects all files in the folder and group the files by name. The layout (single-end, paired-end) of sequencing library will be automatically determined and all layout-associated parameters will be set by the SIP. Please check the online documentation for more details of read arrangements.

2.1.1 Read quality control and preprocessing

Various quality defects (e.g. adapter sequences, low-quality bases, poly-T ends etc.) may exist in the raw sequencing reads. Directly processing the raw reads would lower down the quality of results too (Del Fabbro et al., 2013), hence it is necessary to check the quality of reads before further steps and filter out defective reads and produce clean reads for succeeding steps. SIP integrates serveral tools for read quality checking and filtering, including FastQC (Andrews, 2010), fastp (Chen et al., 2018), AfterQC (Chen et al., 2017), and applies the quality control and preprocessing step to each sample or cell. Separated results for each cell are further summarized with MultiQC (Ewels et al., 2016).

2.1.2 (Quasi-)Alignment of reads and transcript assembly

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Alignment (or quasi-alignment) or short sequencing reads is the essential step in the reference-based transcriptome studies. SIP automatically detects the read layout and the "strandness" of the RNA-seq experiment with RSeQC (Wang et al., 2012) and set the corresponding parameters for users in the (quasi-)alignment steps.

SIP by default uses a quasi-alignment tool, salmon (Patro et al., 2017), for transcript quantification because it is much faster than alignment-based tools. For alignment-based transcript assembly, we include several classical packages in SIP: STAR (Dobin et al., 2013) and Hisat2 (Kim et al., 2015, 2017) for read alignment, StringTie (Pertea et al., 2015) for transcript assembly. Like the previous Read QC and Preprocessing step, the (quasi-)alignment qualities (e.g. read mapping rate, aligned read count, fragment length distribution) are also summarized and reported by MultiQC.

2.1.3 Gene-level abundance counting

For alignment-based quantification, the generated Sequence Alignment Map files (e.g. SAM, BAM files) can be either counted with featureCounts (Liao et al., 2014) or HTSeq (Anders et al., 2015) at read level, or counted based on normalized results (TPM, RPKM, or FPKM) produced by StringTie.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

For the quasi-alignment method, abundance levels of transcripts are directly estimated without explicitly assembling the transcriptome. Normalized results or read counts are supplied by salmon in SIP.

As accurate quantification of transcriptional isoforms at single-cell level is still challenging (Vu et al., 2018; Arzalluz-Luque ángeles and Conesa, 2018), SIP currently imports all cells' transcript-level quantification results and merge them into gene-level expression matrix where each column depicts an expression profile of a single cell with the help of the R package tximport (Soneson et al., 2016).

2.1.4 Post-quantification downstream analysis

SIP can take the gene-level expression matrix generated by the previous abundance counting step, and continue to conduct downstream analyses like clustering, identifying differentially expressed genes, and building cell trajectories, etc. A number of packages have been developed to accomplish such statistical analysis. SIP has currently integrated Seurat, Monocle 2, and Scater.

2.2 Other useful features

By utilizing the advantages of the Nextflow framework, SIP has the following useful features:

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

- **User-friendly reports**: Important intermediate results and summarized final results are well-organized in the output directory along with multiple user-friendly HTML and PDF graphical reports.
- **Cluster environment**: Nextflow supports the deployment on cluster environment, therefore the pipeline is capable of processing very big data with the slurm HPC cluster management system.
- **Parallel processing**: Processes can be computed in a parallel manner, maximizing the usage of computational resources.
- **Breakpoint resume**: SIP caches accomplished work with Nextflow. Finished processes will not be wasted when a very long pipeline is interrupted in the middle.
- **Container support**: SIP provides Docker images with all software and packages installed inside. Users can plug and play the pipeline directly without pre-install the numerous dependencies.
- **E-mail notification**: SIP can be configured to send information automatically when a certain process or the whole workflow is finished.
- **Developer community**: Nextflow has a large developer community. This will help the future development and extension of SIP.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3 Results

We applied SIP on several single-cell RNA-seq datasets to test the usability and performance of the pipeline. We selected a subset of the single-cell dataset generated by SMART-seq with accession number GSE38495 (Ramsköld et al., 2012). The subset contains 8 hESCs (human embryonic stem cells), and 19 LNCaP cells (androgen-sensitive human prostate adenocarcinoma cells). The whole analysis workflow is finished in 46 minutes on a workstation with 2 Intel® Xeon® E5-2670 CPU and 96 GB memory. The real memory requirement for the package depends on the particular set of modules uses choose in the pipeline. The actually memory consumption in this example experiment is <40GB, and we have also checked that SIP can be installed and ran on personal computers or laptops with memory as low as only 8GB.

The raw sequencing reads are checked and filtered with fastp using the "--readqc=fastp" option. Pre-processed clean reads are quantified in quasi-alignment mode with salmon using the "--quant=salmon" option. Fig. 2 (a) shows snapshots of read quality control results and quasi-alignment results. In general, the read qualities of 27 samples are good enough to perform succeeding steps: 97.5%~99.5% of the reads passes the read filtering; adapter sequences are detected in 12 out of the 27 samples, and the percentage of adapter-trimmed reads are $\leqslant$ 1.0% . More details about GC-content, sequence quality, duplication rates, N contents, etc. are also suppled in the report.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

After quantification, the transcript-level expression profiles of multiple cells are automatically transformed into gene-level and then merged into a matrix of all cells to feed the downstream statistical analysis. Clustering identification and differentially expressed gene identification are conducted using Seurat with the "--using_seurat" option for this dataset. We used a typical Seurat workflow:

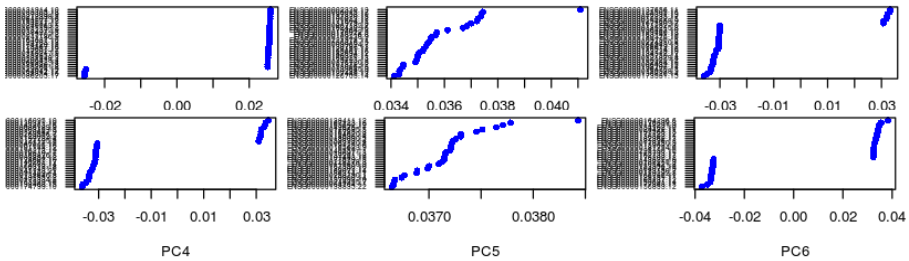
1) Initialize a Seurat object;
2) Filter cells based on built-in metrics;
3) Normalize the data;
4) Select variable genes across cells;
5) Removing confounding factors by regression.
6) Linear dimensionality reduction.
7) t-SNE visualization
8) PCA-based clustering
9) Identify differentially expressed genes

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

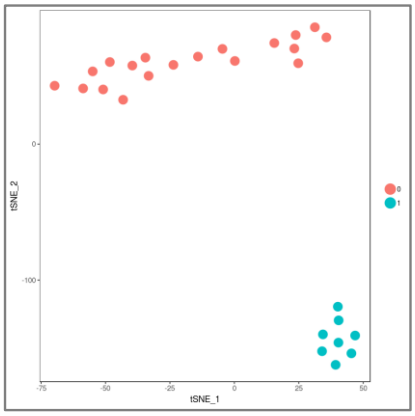
(a)

	% Duplication	GC content	% PF	% Adapter	% Aligned	M Aligned
SRR522081_fastp	50.7%	43.9%	99.2%		72.9%	23.1
SRR522082_fastp	55.0%	44.3%	99.2%		72.5%	23.0
SRR522083_fastp	59.8%	44.3%	99.2%		77.9%	25.2
SRR522084_fastp	60.0%	44.9%	97.5%		74.9%	19.6
SRR522085_fastp	59.3%	45.0%	99.2%		78.1%	22.5
SRR522086_fastp	59.3%	44.6%	99.2%			

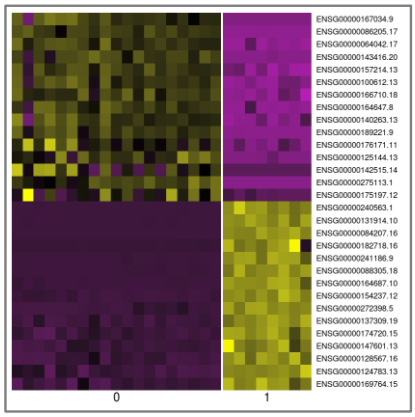
(b)



(c)



(d)



(e)

p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
2.603158e-24	2.958067	1	0.750	5.137072e-20	0	ENSG00000167034.9
5.290915e-13	2.758801	1	1.000	1.044109e-08	0	ENSG00000100612.13
5.740315e-13	2.668593	1	1.000	1.132794e-08	0	ENSG00000166710.18
3.523598e-07	2.754762	1	0.000	6.953468e-03	0	ENSG00000142515.14
2.272058e-43	3.101922	1	0.632	4.483680e-39	1	ENSG00000084207.16
4.793175e-29	3.277976	1	1.000	9.458851e-25	1	ENSG00000154237.12
4.011684e-25	3.006538	1	1.000	7.916658e-21	1	ENSG00000272398.5
2.813670e-20	2.589402	1	1.000	5.552496e-16	1	ENSG00000174720.15

Fig. 2. Selected results of experiment 1. (a) read quality check report and quasi-alignment report; (b) Visualizations of PCA's first 6 components; (c) t-SNE plot of the 2 types of cells; (d) Heatmap of the cells across part of selected marker genes; (e) Selected differentially expressed marker genes.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fig. 2 (b) visualizes the result of PCA linear dimensionality reduction with SIP. The first 6 principle components are listed, where x-axis represents the PC's value, and y-axis represents the corresponding feature genes. Fig. 2 (c) is the t-SNE plot of the PCA-transformed data reported by SIP. The PCA-based clustering algorithm obtains two cluster and they exactly correspond to the two real cell types (19 LNCap cells and 8 hESCs). The colors in the t-SNE plot are inferred from cluster information automatically by SIP. The t-SNE is derived with perplexity parameter = 5 and first two PCA dimension. The clustering algorithm used the first 10 principle components and a resolution parameter = 1.04. Differentially expressed marker genes are then identified between the two computed groups. A heatmap with an obvious pattern shows top 15 genes with high fold changes in Fig. 2 (d). Fig. 2 (e) lists the most significant differentially expressed marker genes in two cell types.

4 Conclusion

We have developed a one-stop pipeline SIP for processing and analyzing single-cell RNA-seq data from the raw fastq data to downstream analyses and automatic report generation. The pipeline is interchangeable and expansible, and can be scaled up to be ran on HPC clusters. The current version of SIP supports data generated from all major RNA-seq platforms. And the future upgrades will soon support other types of single-cell sequencing data such as scATAC-seq data. With this handy pipeline, biological labs that generating single-cell sequencing data can be freed from the tedious and challenging labors of installing and running multiple software packages and setting computer environments to make everything compatible. This allows them to better focus on the scientific questions to be

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

answered from the data. Besides, the pipeline also allows biological labs to be able to share bioinformatics pipelines including detailed parameter settings with other labs, which makes it easier for integrating data from multiple labs in big collaborative projects such as the Human Cell Atlas (HCA) program. SIP is available at https://github.com/chansigit/SIP under the GPL-3.0 license.

Acknowledgements

We thank Zhun Miao, Xiangyu Li, Dongfang Wang and Dongyang Luo for their helpful discussions.

Funding

This work is supported by CZI HCA pilot project, the National Key R&D Program of China grant 2018YFC0910400 and the NSFC grant 61721003.

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2015) HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Andrews,S. (2010) FastQC - A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arzalluz-Luque ángeles and Conesa,A. (2018) Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biol.*, **19**.

Butler,A. *et al.* (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.

Chen,S. *et al.* (2017) AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics*, **18**.

Chen,S. *et al.* (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Dobin,A. *et al.* (2013) {STAR}: ultrafast universal {RNA}-seq aligner. *Bioinformatics*, **29**, 15–21.

Ewels,P. *et al.* (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Del Fabbro,C. *et al.* (2013) An extensive evaluation of read trimming effects on illumina NGS data analysis. *PLoS One*, **8**.

Gardeux,V. *et al.* (2017) ASAP: A web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics*, **33**, 3123–3125.

Hedlund,E. and Deng,Q. (2018) Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Aspects Med.*, **59**, 36–46.

Kim,D. *et al.* (2015) HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Kim,D. *et al.* (2017) HISAT2: Graph-based alignment of next-generation sequencing reads to a population of genomes.

Liao,Y. *et al.* (2014) FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.

Lun,A.T.L. *et al.* (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, **5**, 2122.

McCarthy,D.J. *et al.* (2017) Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186.

Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

Qiu,X. *et al.* (2017) Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*, **14**, 309–315.

Ramsköld,D. *et al.* (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**, 777–782.

Rozenblatt-Rosen,O. *et al.* (2017) The Human Cell Atlas: From vision to reality. *Nature*, **550**, 451–453.

Soneson,C. *et al.* (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.

DI Tommaso,P. *et al.* (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.

Trapnell,C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.

Vu,T.N. *et al.* (2018) Isoform-level gene expression patterns in single-cell RNA-sequencing data. *Bioinformatics*.

Wang,L. *et al.* (2012) RSeQC: quality control of RNA-seq experiments. *Bioinforma. Oxford Engl.*, **28**, 2184–2185.

Wang,Y. and Navin,N.E. (2015) Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cell*, **58**, 598–609.

bioRxiv preprint doi: https://doi.org/10.1101/456772; this version posted November 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Wu,A.R. *et al.* (2017) Single-Cell Transcriptional Analysis. *Annu. Rev. Anal. Chem.*, **10**, 439–462.

Zhu,X. *et al.* (2017) Granatum: A graphical single-cell RNA-Seq analysis pipeline for genomics scientists. *Genome Med.*, **9**, 1–12.