METHOD ARTICLE

An open RNA-Seq data analysis pipeline tutorial with an example of reprocessing data from a recent Zika virus study [version 1; referees: 3 approved]

Zichen Wang[1-3], Avi Ma'ayan[1-3]

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY, Box 1603, USA
[2]BD2K-LINCS Data Coordination and Integration Center, Icahn School of Medicine at Mount Sinai, New York, NY, Box 1603, USA
[3]Mount Sinai Knowledge Management Center for Illuminating the Druggable Genome, Icahn School of Medicine at Mount Sinai, New York, NY, Box 1603, USA

v1 **First published:** 05 Jul 2016, **5**:1574 (doi: 10.12688/f1000research.9110.1)
Latest published: 05 Jul 2016, **5**:1574 (doi: 10.12688/f1000research.9110.1)

Abstract

RNA-seq analysis is becoming a standard method for global gene expression profiling. However, open and standard pipelines to perform RNA-seq analysis by non-experts remain challenging due to the large size of the raw data files and the hardware requirements for running the alignment step. Here we introduce a reproducible open source RNA-seq pipeline delivered as an IPython notebook and a Docker image. The pipeline uses state-of-the-art tools and can run on various platforms with minimal configuration overhead. The pipeline enables the extraction of knowledge from typical RNA-seq studies by generating interactive principal component analysis (PCA) and hierarchical clustering (HC) plots, performing enrichment analyses against over 90 gene set libraries, and obtaining lists of small molecules that are predicted to either mimic or reverse the observed changes in mRNA expression. We apply the pipeline to a recently published RNA-seq dataset collected from human neuronal progenitors infected with the Zika virus (ZIKV). In addition to confirming the presence of cell cycle genes among the genes that are downregulated by ZIKV, our analysis uncovers significant overlap with upregulated genes that when knocked out in mice induce defects in brain morphology. This result potentially points to the molecular processes associated with the microcephaly phenotype observed in newborns from pregnant mothers infected with the virus. In addition, our analysis predicts small molecules that can either mimic or reverse the expression changes induced by ZIKV. The IPython notebook and Docker image are freely available at: http://nbviewer.jupyter.org/github/maayanlab/Zika-RNAseq-Pipeline/blob/master/Zika.ipynb and https://hub.docker.com/r/maayanlab/zika/.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 05 Jul 2016	✓ report	✓ report	✓ report

1 **Ravi K Madduri**, University of Chicago USA

2 **Apostolos Zaravinos**, European University Cyprus Cyprus

3 **Fredrik Pettersson**, Uppsala University Sweden, **Sarbashis Das**, Uppsala University Sweden

Discuss this article

Comments (0)

This article is included in the Zika & Arbovirus Outbreaks channel.

Corresponding author: Avi Ma'ayan (avi.maayan@mssm.edu)

How to cite this article: Wang Z and Ma'ayan A. **An open RNA-Seq data analysis pipeline tutorial with an example of reprocessing data from a recent Zika virus study [version 1; referees: 3 approved]** *F1000Research* 2016, **5**:1574 (doi: 10.12688/f1000research.9110.1)

Grant information: This work is partially supported by the National Institutes of Health (NIH) grants U54HL127624, U54CA189201, and R01GM098316.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 05 Jul 2016, **5**:1574 (doi: 10.12688/f1000research.9110.1)

Introduction

The increase in awareness about the irreproducibility of scientific research requires the development of methods that make experimental and computational protocols easily repeatable and transparent[1]. The advent of interactive notebooks for data analysis pipelines significantly enhances the recording and sharing of data, source code, and figures[2]. In a subset of recent publications, an interactive notebook was published alongside customary manuscripts[3]. Similarly, here we present an interactive IPython notebook (http://nbviewer.jupyter.org/github/maayanlab/Zika-RNAseq-Pipeline/blob/master/Zika.ipynb) that serves as a tutorial for performing a standard RNA-seq pipeline. The IPython notebook pipeline provides scripts (http://dx.doi.org/10.5281/zenodo.56311) that process the raw data into interactive figures and permits other downstream analyses that can enable others to quickly and properly repeat our analysis as well as extract knowledge from their own data. As an example, we applied the pipeline to RNA-seq data from a recent publication where human induced pluripotent stem cells were differentiated to neuronal progenitors and then infected with Zika virus (ZIKV)[4]. The aim of the study was to begin to understand the molecular mechanisms that induce the observed devastating phenotype of newborn-microcephaly from pregnant mothers infected with the virus.

Methods and results

The first publicly available study profiling gene expression changes after ZIKV infection of human cells was deposited into NCBI's Gene Expression Omnibus (GEO) in March 2016. The raw data is available (ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP070/SRP070895/) from the Sequence Read Archive (SRA) with accession number GSE78711. In this study, gene expression was measured by RNA-seq using two platforms: MiSeq and NextSeq[4] in duplicates. The total number of samples is eight, with four untreated samples and four infected samples. We first downloaded the raw sequencing files from SRA and then converted them to FASTQ files. Quality Control (QC) for the RNA-Seq reads was assessed using FastQC[5]. The reports generated by FastQC were in HTML format and can be accessed through hyperlinks from the IPython notebook. The reads in the FASTQ files were aligned to the human genome with Spliced Transcripts Alignment to a Reference (STAR)[6]. STAR is a leading aligner that accomplishes the alignment step faster and more accurately than other current alternatives[6]. We next applied featureCounts[7] to assign reads to genes, and then applied the edgeR Bioconductor package[8] to compute counts per million (CPM) and reads per kilobase million (RPKM). The next steps are performed in Python within the IPython notebook. We first filtered out genes that are not expressed or lowly expressed. Subsequently, we performed principal component analysis (PCA) (Figure 1). The PCA plots show that the samples cluster by infected vs. control cells, but also by platform. Next, we visualized the 800 genes with the largest variance using an interactive hierarchical clustering (HC) plot (Figure 2). This analysis separates the groups of genes that are differentially expressed by infected vs. control from those that are differential by platform. The visualization of the clusters is implemented with an interactive external web-based data visualization tool called

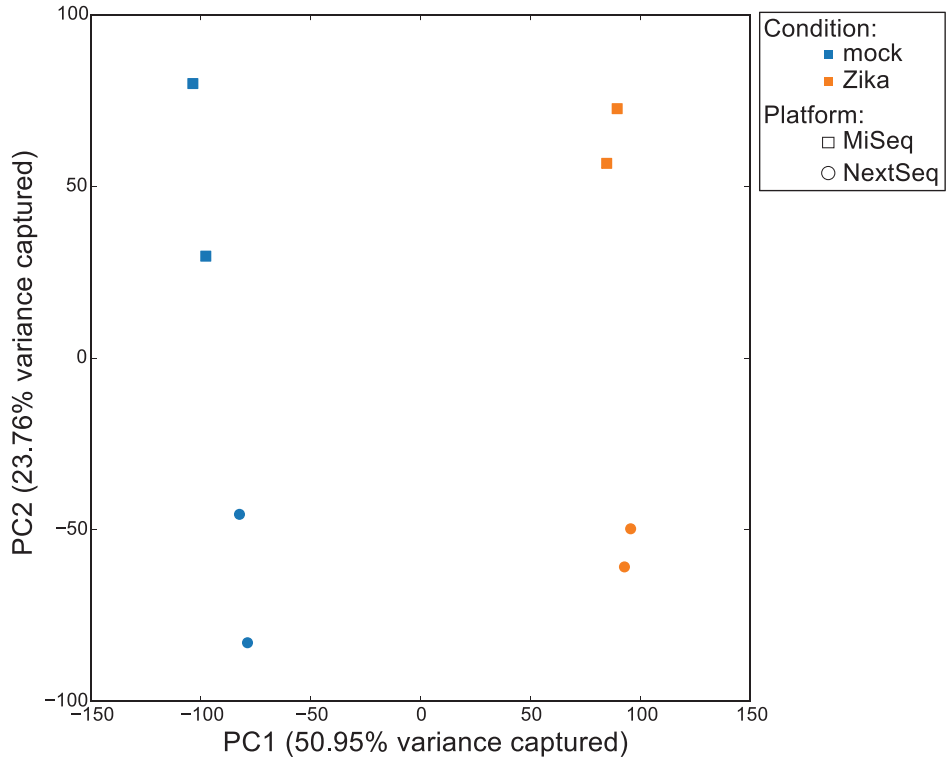

Figure 1. Principal Component Analysis (PCA) of the samples in the first two principal component space. ZIKV-infected and mock-treated cells are colored in orange and blue, respectively. The shapes of the dots indicate the sequencing platforms: MiSeq – squares, and NextSeq - circles.

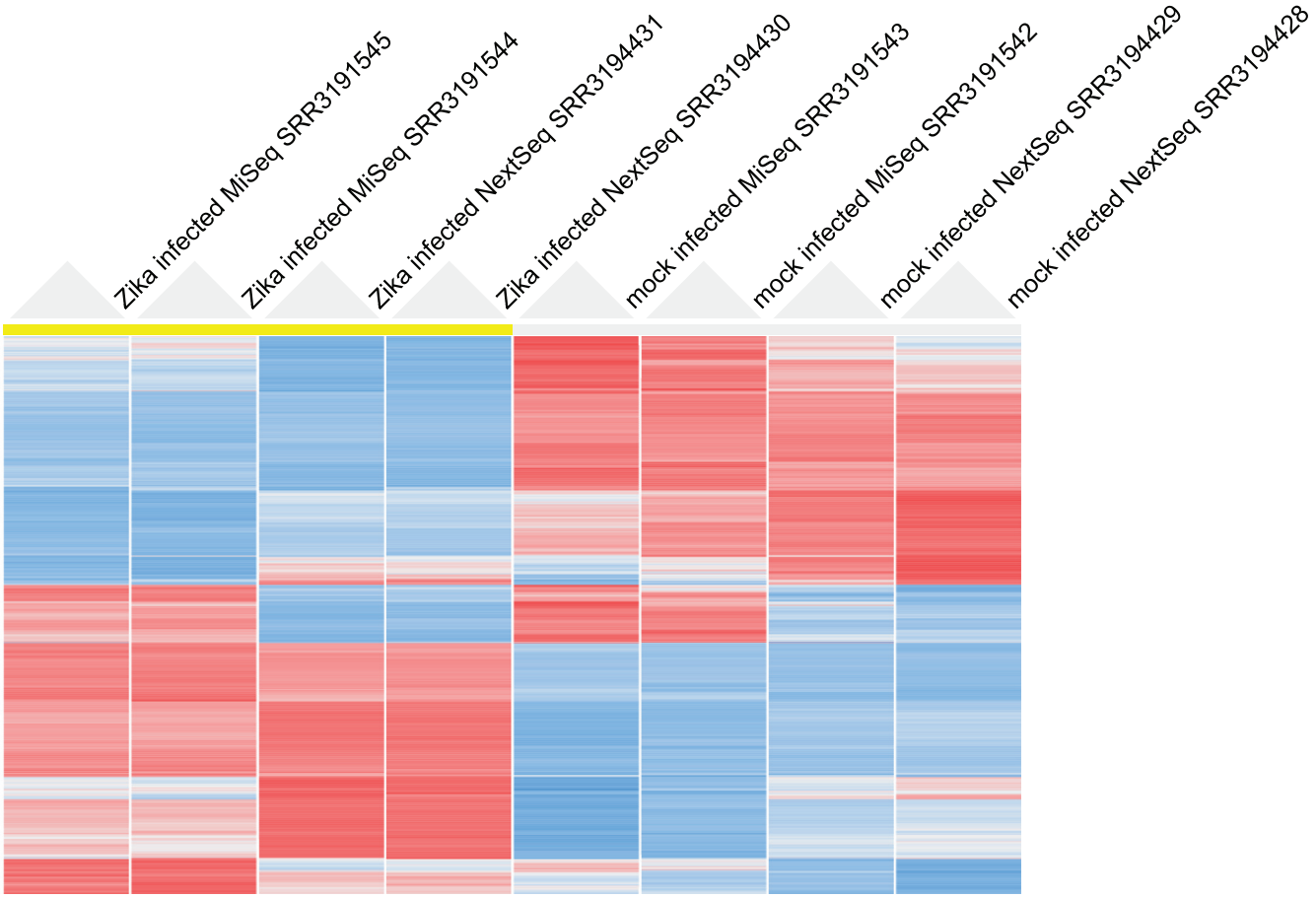

Figure 2. Hierarchical clustering heatmap of the 800 genes with the largest variance. The CPM of 800 genes with the largest variance across the eight samples were log transformed and z-score normalized across samples. Blue indicates low expression and red high.

clustergrammer (http://amp.pharm.mssm.edu/clustergrammer/). Clustergrammer provides interactive searching, sorting and zooming capabilities.

The following step is to identify the differentially expressed genes (DEG) between the two conditions. This is achieved with a unique method we developed called the Characteristic Direction (CD)[9]. The CD method is a multivariate method that we have previously demonstrated to outperform other leading methods that compute differential expression between two conditions[9]. Once we have ranked the lists of DEG, we submit these for signature analysis using two tools: Enrichr[10] and L1000CDS2[11]. Enrichr queries the up and down gene sets against over 180,000 annotated gene sets belonging to 90 gene set libraries covering pathway databases, ontologies, disease databases, and more[10]. The results from this enrichment analysis confirm that the downregulated genes after ZIKV infection are enriched for genes involved in cell cycle-related processes (Figure 3a). These genes are enriched for targets of the transcription factors E2F4 and FOXM1 (Figure 3b). Both transcription factors are known to regulate cell proliferation and play central role in many cancers. The downregulation of cell cycle genes was already reported in the original publication; nevertheless, we obtained more interesting results for the enriched terms that appeared most significant for the upregulated genes. Particularly, the top two terms from the mouse genome informatics (MGI) Mammalian Phenotype Level 4 library are abnormal nervous system (MP0003861) and abnormal brain morphology (MP0002152) (Table S1). This library associates gene knockouts in mice with mammalian phenotypes. These enriched terms enlist a short set of genes that potentially link ZIKV infection with the concerning observed microcephaly phenotype. Finally, to identify small molecules that can potentially either reverse or mimic ZIKV-induced gene expression changes, we query the ZIKV-induced signatures against the LINCS L1000 data. For this, we utilize L1000CDS2[11], a search engine that prioritize small molecules given a gene expression signature as input. L1000CDS2 contains 30,000 significant signatures that were processed from the LINCS L1000 data with the CD method. The results suggest small molecules that could be tested in follow-up studies in human cells for potential efficacy against ZIKV (Table S2).

To ensure the reproducibility of the computational environment used for the whole RNA-Seq pipeline, we packaged all the software components used in this tutorial, including the command line tools, R packages, and Python packages into a Docker image. This Docker image is made publically available at https://hub.docker.com/r/maayanlab/zika/. The Docker image was created

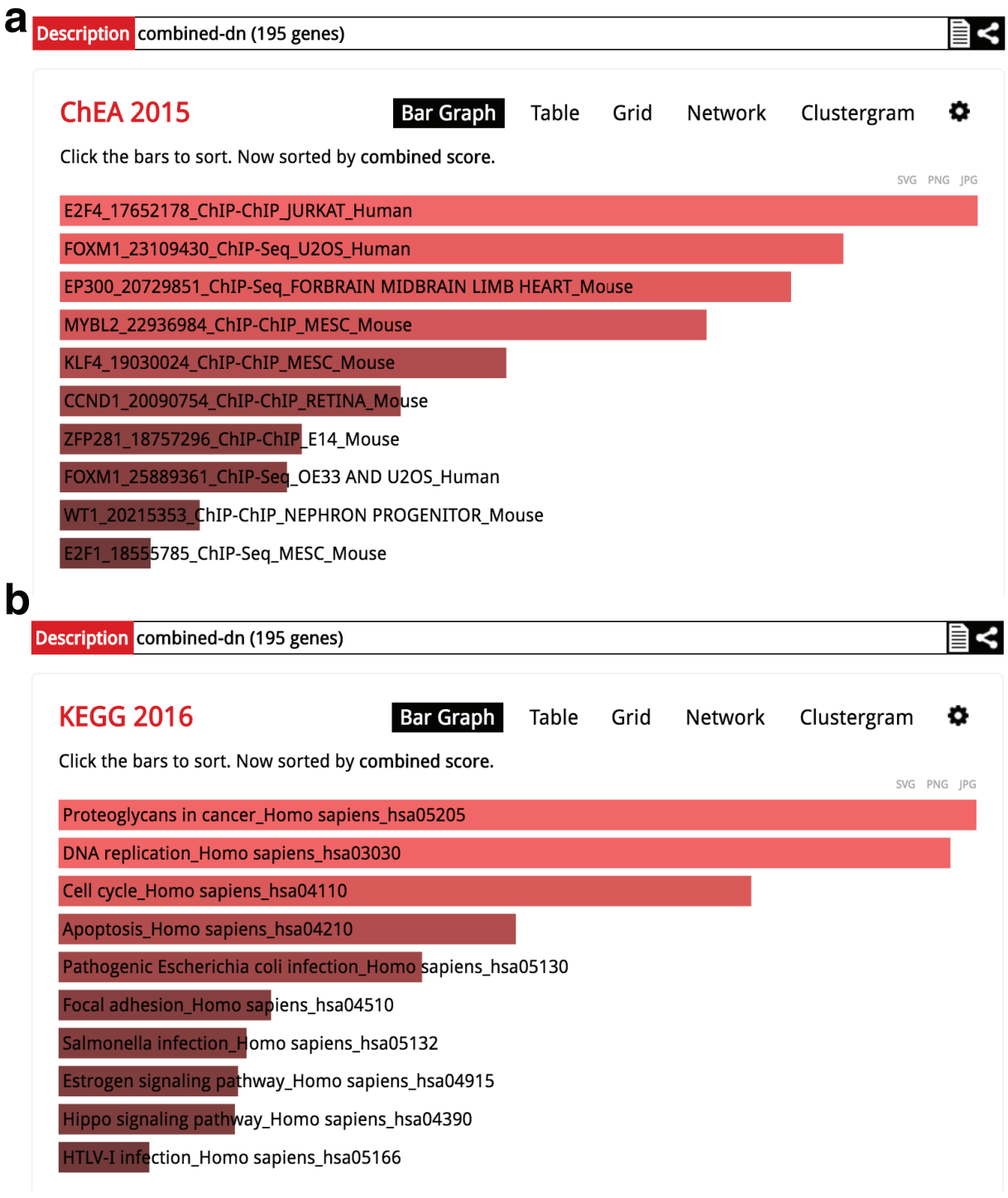

Figure 3. Bar plots of the top enriched gene sets from the (**a**) ChEA and (**b**) KEGG libraries for the downregulated genes after ZIKV infection.

based on the specifications outlined on the official IPython's Scipy Stack image (https://hub.docker.com/r/ipython/scipystack/). The additional command line tools, R scripts, and Python packages together with their dependencies were compiled and installed into the Docker image. The RNA-Seq pipeline Docker image was deployed onto our Mesos cluster, which allows users to run the IPython notebook interactively. The Docker image can also be downloaded and executed on local computers and servers, or deployed in the cloud if users have access to cloud provider services with a Docker Toolbox installed (https://www.docker.com/products/docker-toolbox). We also provide detailed instructions on how to download and execute the Docker image (https://hub.docker.com/r/maayanlab/zika/).

The 'Dockerization' of the RNA-Seq pipeline facilitates reproducibility of the pipeline at the software level because the Docker image ensures that all versions of the software components are consistent and static. Dockerization also helps users to handle the complex installation of many dependencies required for the computational pipeline. Moreover, the Docker image can be executed on a single computer, clusters/servers and on the cloud. The only limitation of having a Docker image is that it prevents users from adding or altering the various steps which require additional software components and packages. However, advanced users can build their own Docker images based on our initial image to customize it for their needs.

Discussion and conclusions

In summary, we provide an open source RNA-seq processing pipeline (Figure 4) that can be used to extract knowledge from any study that profiled gene expression using RNA-seq applied to mammalian cells, comparing two conditions. The advantage of providing the pipeline in the IPython notebook format and as a Docker container is that it enables others to quickly reproduce our results with minimal overhead and potentially apply similar methodology for the analysis of other similar datasets. Advanced

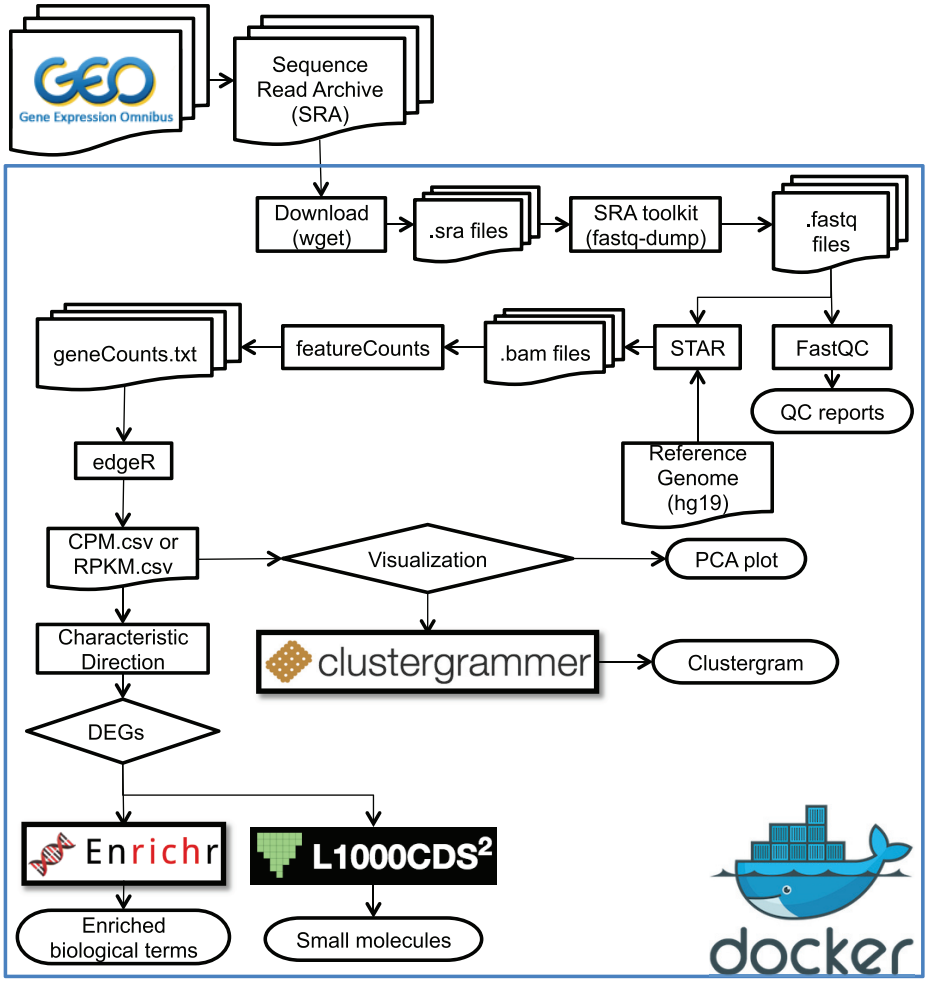

Figure 4. Workflow of the different steps carried out in the pipeline.

users can add, improve and customize the pipeline by forking it on GitHub. The results that we obtained for ZIKV are consistent with the results published in the original study, but also enhance those findings by discovering a link between the upregulated genes and genes that, when knocked out in mice, induce morphological brain defects. Some of these genes could be the causal genes of the microcephaly phenotype observed in newborns of mothers infected with the virus. Nevertheless, caution should be used when interpreting these results because they may simply indicate a reduction in cell cycle activity and an increase in neuronal differentiation of the type of cells used in the original study.

Data and software availability

The IPython notebook, as well as other scripts and data files for this tutorial are available on GitHub at: https://github.com/MaayanLab/Zika-RNAseq-Pipeline, doi: http://dx.doi.org/10.5281/zenodo.56311[12].

The Docker image for this tutorial is available on DockerHub at: https://hub.docker.com/r/maayanlab/zika/.

Author contributions

AM conceived and lead the study. ZW developed the software and performed the analysis. AM interpreted the results. Both authors wrote the paper and tutorials.

Competing interests

No competing interests were disclosed.

Grant information

This work is partially supported by the National Institutes of Health (NIH) grants U54HL127624, U54CA189201, and R01GM098316.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank Dr. Ajay Pillai from NHGRI for useful suggestions and Kathleen Jagodnik from NASA for copyediting an early version of the manuscript.

Supplementary material

Table S1.

Top enriched gene sets from the MGI Mammalian Phenotype Level 4 gene set library for the upregulated genes after ZIKV infection.

Click here to access the data.

Table S2.

(a) Top scoring small molecules that are potential mimickers of the ZIKV infection signatures. (b) Top scoring small molecules that are potential reversers of the ZIKV infection signatures.

Click here to access the data.

References

1. Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature.* 2012; **483**(7391): 531–533.
PubMed Abstract | Publisher Full Text
2. Shen H: **Interactive notebooks: Sharing the code.** *Nature.* 2014; **515**(7525): 151–152.
PubMed Abstract | Publisher Full Text
3. https://IPython.org. **A gallery of interesting IPython Notebooks.** [cited 2016 25 April]; 2016.
Reference Source
4. Tang H, Hammack C, Ogden SC, *et al.*: **Zika Virus Infects Human Cortical Neural Progenitors and Attenuates Their Growth.** *Cell Stem Cell.* 2016; **18**(5): 587–590.
PubMed Abstract | Publisher Full Text
5. Andrews S: **FastQC: A quality control tool for high throughput sequence data**. 2010.
Reference Source
6. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**(7): 923–930.
PubMed Abstract | Publisher Full Text
8. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Clark NR, Hu KS, Feldmann AS, *et al.*: **The characteristic direction: a geometrical approach to identify differentially expressed genes.** *BMC Bioinformatics.* 2014; **15**(1): 79.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Chen EY, Tan CM, Kou Y, *et al.*: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics.* 2013; **14**(1): 128.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Duan Q, Reid SP, Clark NR, *et al.*: **L1000CDS^2: LINCS L1000 Characteristic Direction Signature Search Engine.** NPJ Systems Biology and Applications, 2016.
Reference Source
12. Wang Z, Ma'ayan A: **Zika-RNAseq-Pipeline v0.1.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 15 August 2016

doi:10.5256/f1000research.9804.r15289

Fredrik Pettersson, **Sarbashis Das**
Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

Wang and Ma'ayan have developed an open source RNAseq processing pipeline using standard methods as well as integrated tools for visualization and data analysis of differentially expressed genes between two sets of data. They test the pipeline by reanalyzing a previously published study of Zika virus infected human cells. They replicated the main result of the original study where genes related to the cell cycle were downregulated after infection. They also find a potential link to genes involved in brain morphology and a normal functioning of the nervous system in mice, something the original study missed. These genes are upregulated after Zika virus infection. This pipeline should be useful in any type of study where two conditions are compared, e.g. infected vs uninfected cells or treated vs untreated condition.

Minor comments:

1. Figure 2. Have the different conditions/figure labels been mixed up? The 2 Zika infected MiSeq conditions look complementary to the 2 mock infected NextSeq controls, while the 2 Zika infected NextSeq conditions look complementary to the 2 mock infected MiSeq controls.

2. Figure 3a and 3b have been mixed up. 3b should be 3a and vice versa.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 29 July 2016

doi:10.5256/f1000research.9804.r15290

Apostolos Zaravinos
The School of Sciences, European University Cyprus, Engomi, Cyprus

The authors provide an open source RNA-seq processing pipeline that can be used to extract differential expression data between two conditions from an RNA-seq experiment. To test this, data from a recent publication where human iPSCs were differentiated to neuronal progenitors and then infected with Zika virus (ZIKV), were analysed with their pipeline, and their results were consistent with the original ones. Of interest, the authors discovered a link between the upregulated genes of this study and genes that, when

knocked out could cause microcephaly observed in newborns of mothers infected with ZIKV. Overall, this is a novel and exciting computational protocol that promotes reproducibility and transparency of the results, and it is definately worth to be tested in other conditions, as well (eg, cancer).

Minor comment: Perhaps its a matter of zoom-in/out of the clustergrammer, but it would be nice to show the names of the top up-/down-regulated genes (or their categories) as depicted in Figure 2.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 21 July 2016

doi:10.5256/f1000research.9804.r14924

Ravi K Madduri
Computation Institute, University of Chicago, Chicago, IL, USA

Wang and Ma'ayan introduced an RNA-seq pipeline tutorial using IPython notebook and a Docker image. Specifically, the authors applied the pipeline to analyze data from a recent Zika virus study. The authors found that their pipeline not only confirms the down-regulated cell cycle genes, but also uncovers a set of genes with a biological function that potentially associated with a particular phenotype. While the work and context sound interesting, there are several concerns that need to be addressed or discussed:

1. This reviewer really liked the approach that the authors have taken to showcase analysis. I wish more researchers adopt this methodology.

2. This could be a great way to do additional analysis easily.. I wonder if authors can look into additional RNASeq pipelines and compare/contrast how Jupiter-friendly they are.

Minor comments:

The clustering (in Figure 1) is based on z-score and the 800 genes serve well to cluster the samples into two different groups. Was z-scores close to zero excluded as they are uninformative? Was FDR applied?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.