



SOFTWARE TOOL ARTICLE

UPDATE

edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens [v2; ref status: indexed, http://f1000r.es/4kw]

Previously titled: shRNA-seq data analysis with *edgeR*

Zhiyin Dai¹, Julie M. Sheridan^{2,3}, Linden J. Gearing^{1,3}, Darcy L. Moore^{1,3}, Shian Su¹, Sam Wormald^{3,4}, Stephen Wilcox^{3,4}, Liam O'Connor^{3,4}, Ross A. Dickins^{1,3}, Marnie E. Blewitt^{1,3}, Matthew E. Ritchie^{1,3}

¹Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia

²Stem Cells and Cancer Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia

³Department of Medical Biology, The University of Melbourne, Parkville, Victoria, 3010, Australia

⁴Systems Biology and Personalised Medicine Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, 3052, Australia

v2 First published: 24 Apr 2014, 3:95 (doi: [10.12688/f1000research.3928.1](https://doi.org/10.12688/f1000research.3928.1))

Latest published: 21 Oct 2014, 3:95 (doi: [10.12688/f1000research.3928.2](https://doi.org/10.12688/f1000research.3928.2))

Abstract

Pooled library sequencing screens that perturb gene function in a high-throughput manner are becoming increasingly popular in functional genomics research. Irrespective of the mechanism by which loss of function is achieved, via either RNA interference using short hairpin RNAs (shRNAs) or genetic mutation using single guide RNAs (sgRNAs) with the CRISPR-Cas9 system, there is a need to establish optimal analysis tools to handle such data. Our open-source processing pipeline in edgeR provides a complete analysis solution for screen data, that begins with the raw sequence reads and ends with a ranked list of candidate genes for downstream biological validation. We first summarize the raw data contained in a fastq file into a matrix of counts (samples in the columns, genes in the rows) with options for allowing mismatches and small shifts in sequence position. Diagnostic plots, normalization and differential representation analysis can then be performed using established methods to prioritize results in a statistically rigorous way, with the choice of either the classic exact testing methodology or generalized linear modeling that can handle complex experimental designs. A detailed users' guide that demonstrates how to analyze screen data in edgeR along with a point-and-click implementation of this workflow in Galaxy are also provided. The edgeR package is freely available from <http://www.bioconductor.org>.

Open Peer Review

Referee Status:

Invited Referees

1 2 3

UPDATE	version 2	report
published 21 Oct 2014		
version 1	report	report
published 24 Apr 2014		

1 James W. MacDonald, University of Washington USA

2 Ross Lazarus, Baker IDI Heart and Diabetes Institute Australia

3 Sumit Deswal, The Research Institute of Molecular Pathology Austria

Discuss this article

Comments (0)

Corresponding author: Matthew E. Ritchie (mritchie@wehi.edu.au)

How to cite this article: Dai Z, Sheridan JM, Gearing LJ *et al.* **edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens [v2; ref status: indexed, <http://f1000r.es/4kw>]** *F1000Research* 2014, 3:95 (doi: [10.12688/f1000research.3928.2](https://doi.org/10.12688/f1000research.3928.2))

Copyright: © 2014 Dai Z *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This research was supported by NHMRC Project grants 1050661 (MER) and 1059622 (MER and MEB), Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRISS.

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: 24 Apr 2014, 3:95 (doi: [10.12688/f1000research.3928.1](https://doi.org/10.12688/f1000research.3928.1))

First indexed: 08 May 2014, 3:95 (doi: [10.12688/f1000research.3928.1](https://doi.org/10.12688/f1000research.3928.1))

UPDATE Amendments from Version 1

In this revised version of our article, we have extended our software to accommodate data from pooled genetic sequencing screens that make use of CRISPR-Cas9 technology. On the software side, the major change is the new `processAmplicons` function (available in edgeR version 3.8.0) that replaces the `processHairpinReads` function. It handles both shRNA-seq and sgRNA-seq data generated with either single or dual sample indexing strategies. These changes have been incorporated in our Galaxy tool along with further refinements to allow filtering of samples with low representation and experimental designs with up to two factors (the original Galaxy tool only accommodated single factor experiments). We provide data and example analyses of two CRISPR-Cas9 screens in the user guide (available from <http://bioinf.wehi.edu.au/shRNaseq/>) to demonstrate this new capability. To incorporate these changes in the main text, the title and introduction have been broadened to reflect the expanded scope of our tool and the author list has been updated to include the additional contributions required to make these incremental improvements to our software. We trust that other researchers will also find these changes useful.

See referee reports

Introduction

Pooled library sequencing screens couple gene knock-down/editing technology with second generation sequencing to allow researchers to elucidate gene function in an unbiased, high-throughput manner^{1,2}. Several recent high impact studies have exploited this approach to discover novel genes involved in processes including cell fate decisions of normal and cancer cells, drug resistance, and to generate genetic interaction maps in mammalian cells using RNA interference (RNAi)^{3–5} and sgRNAs with the clustered regularly interspaced palindromic repeats-Cas9 (CRISPR-Cas9) genome editing system^{2,6}.

Pooled screening relies on the stable genomic integration (often by viral transduction) of a library of uniquely identifiable expression constructs within a population of cells. Each construct expresses an RNA transcript that targets nuclelease machinery to a specific nucleotide sequence. This is currently achieved in two main ways: shRNAs can be designed to target specific mRNA transcripts for degradation via the DICER/RISC pathway⁷ or sgRNAs can be designed to target a co-expressed Cas9 nuclease to a specific sequence in the genome⁸. By targeting constitutive exons at the 5' region of a gene, Cas9-mediated double-stranded breaks are repeatedly repaired by non-homologous end joining until a mutation is introduced that renders the site unrecognisable by the sgRNA. Such mutations typically comprise an insertion or deletion and can give rise to altered coding sequences, disrupted splice sites, frame shifts and/or premature stop codons in the target gene⁸.

Depending on the biological question of interest, typically two or more cell populations are compared either in the presence or absence of a selective pressure, or as a time-course before and after a selective pressure is applied. Gain of shRNA/sgRNA representation within a pool suggests that disrupting target gene function confers some sort of advantage to a cell. Similarly, genes whose knockdown/knockout is disadvantageous may be identified through loss of shRNA/sgRNA representation. Screening requires a library

of constructs in a lentiviral or retroviral vector backbone that is used to generate a pool of virus for transducing cells of interest. The relative abundance of these constructs in transduced cells is then quantified by PCR amplification of proviral integrants from genomic DNA using primers designed to amplify all cassettes (shRNA/sgRNA) equally, followed by second-generation amplicon sequencing (Figure 1A). Sample-specific primer indexing allows many different conditions to be analyzed in parallel.

As the popularity of these approaches grows, there is a need to develop suitable analysis pipelines to handle the large volumes of raw data that each screen generates. The major steps in an analysis involve processing the raw sequence reads, assessing the data quality and determining representational differences in the screen in a statistically rigorous way.

Two pipelines are currently available for this task that have been tailored for data from shRNA-seq screens. The `shALIGN` program⁹ is a custom Perl script that trims the sequence reads to the pre-defined base positions and then matches these to a library of hairpin sequences. Mismatch bases are permitted, and any ambiguous matches are ignored from the final hairpin count. Statistical analysis of the data is then performed using the `shRNaseq R package`⁹, which calculates log-ratios of the counts from each screen replicate, normalizes these values and ranks hairpins by their median, mean or *t*-statistic. Another solution is the BiNGS!SL-seq program¹⁰ that uses Bowtie to perform sequence mapping followed by statistical analysis in `edgeR`¹¹.

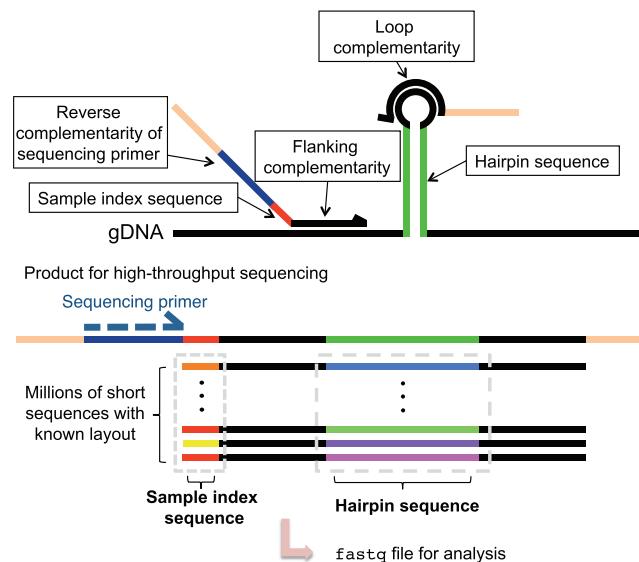
In this article, we describe a complete analysis solution for shRNA/sgRNA-seq screens accessible from within the `edgeR` package available from `Bioconductor`¹².

Implementation

A summary of the main steps in a typical shRNA/sgRNA-seq analysis alongside the functions in `edgeR` that perform each task is given in Figure 1B.

Sequence pre-processing

Our sequence counting procedure has been tailored for screens where PCR amplified shRNA/sgRNA constructs of known structure are sequenced using second generation sequencing technology (Figure 1A). The location of each index and hairpin/guide sequence is used to determine matches between a specified list of index and hairpin/guide sequences expected in the screen with the sequences in the fastq file. Mismatches in the hairpin/guide sequence are allowed to accommodate sequencing errors, as are small shifts in the position of these sequences within the read. Analysis of unpublished in-house data reveals that allowing for mismatches can yield up to 4.4% additional reads, and shifting an extra 2.6%. This simple searching strategy is implemented in C, with the user interface provided by the `processAmplicons` function in `edgeR`. Input to this function consists of a fastq file/s, a second file containing sample IDs and their index sequences and a third file listing hairpin/guide IDs and their respective sequences (the latter files are tab-delimited). A screen with 100 million reads (one lane from an Illumina HiSeq 2000) can be processed in 2–15 minutes depending on the processing parameters. Fastq processing requires minimal

(A) Overview of pooled amplicon sequencing**(B) Summary of edgeR workflow**

Step	Function
Data Import	processAmplicons
Quality Assessment, Plotting & Normalization	plotMDS calcNormFactors plotBCV plotSmear
Differential Representation Analysis	estimateDisp exactTest glmFit glmLRT topTags
Gene-level Analysis	camera roast

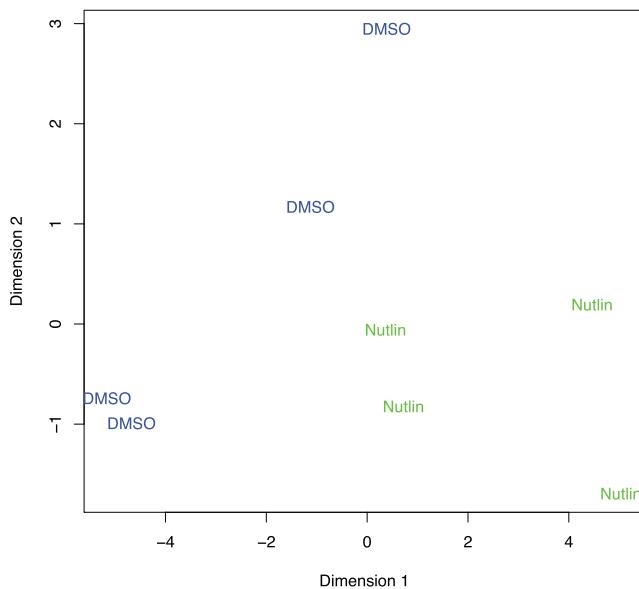
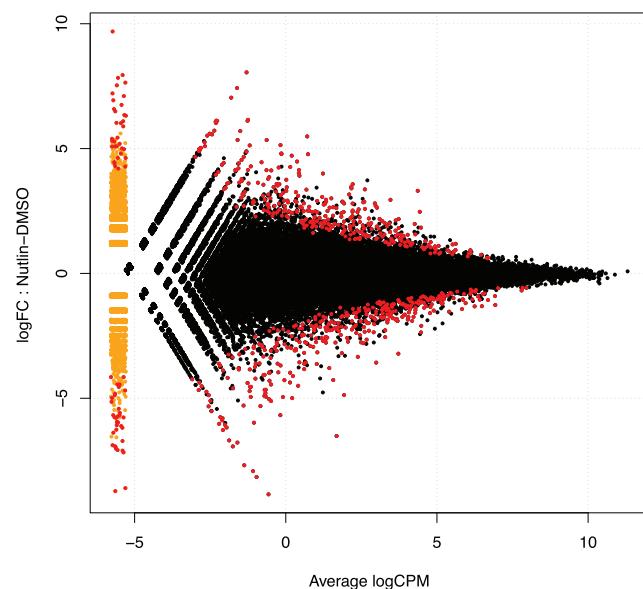
(C) MDS plot of screen replicates**(D) Fold-change versus abundance**

Figure 1. Summary of the raw data, workflow and diagnostic plots from edgeR. (A) Structure of the amplicons sequenced in a typical shRNA-seq screen. Each amplicon will contain sample and hairpin specific sequences at predetermined locations. In sgRNA-seq screens, the amplicon sequences have a similar structure, with the sgRNA sequence replacing the hairpin. After sequencing, the raw data is available in a fastq file. (B) The main steps and functions used in an analysis of shRNA/sgrNA-seq screen data in edgeR are shown. (C) Example of a multidimensional scaling (MDS) plot showing the relationships between replicate dimethyl sulfoxide (DMSO) and Nutlin treated samples (data from Sullivan *et al.* (2012)⁴). MDS plots provide a quick display of overall variability in the screen and can highlight inconsistent samples. (D) Plot of log₂-fold-change versus hairpin abundance (log₂CPM) for the same data. Hairpins with a false discovery rate < 0.05 from an exact test analysis in edgeR (highlighted in red) may be prioritized for further validation.

RAM, allowing analysis to be completed on any standard computer with R¹³ installed.

The matrix of counts returned by the `processAmplicons` function, which contains genes in the rows and samples in the columns, is stored as a `DGEList` object so that it is fully interoperable with the downstream analysis options available in edgeR. Such an object can also be created directly by the user in the event that these counts have been summarized by alternate means.

Next, the data quality of a screen can be assessed conveniently using multidimensional scaling (MDS) plots via `plotMDS` (Figure 1C) and access to a range of normalization options is available through the `calcNormFactors` function.

Differential representation analysis

The shRNaseq software⁹ assumes simple experimental set-ups (e.g. comparing two conditions) that are unsuitable in more complicated situations, such as time-course designs. In edgeR, screens can be analyzed using either the classic method¹⁴, ideal for simple two-group comparisons, or generalized linear models (GLMs)¹⁵ for more complex screens with multiple conditions (using the `glmFit` function). This framework can accommodate hairpin/guide-specific variation of both a technical and biological nature as estimated via the `estimateDisp` function and visualized using `plotBCV`, which plots biological variability as a function of average hairpin/guide abundance. Robust regression is also possible via the use of observation weights that are estimated using the `estimateGLMRobustDisp` function¹⁶. Statistical testing for changes in

shRNA/sgRNA abundance between conditions of interest (typically over time) is carried out using exact (see `exactTest` function) or likelihood ratio (`glmLRT`) tests that allow results to be ranked by significance using the `topTags` function and plotted using the `plotSmear` function (Figure 1D).

Gene set analysis tools available via `roast`¹⁷ and `camera`¹⁸ allow researchers to further test and prioritize screen results. This capability can be used to obtain a gene-by-gene ranking, rather than a hairpin/guide-specific one, which can be helpful when shRNA or sgRNA libraries contain multiple hairpins or guides targeting each gene.

Case studies and further extensions

We provide example data sets and a complete analysis script that demonstrate how to use the edgeR package to prioritize data from four different shRNA-seq screens and two sgRNA-seq screens¹⁹. These examples were chosen to showcase edgeR's ability to deal with experiments of varying size (from tens to thousands of genes) and complexity, from two-group situations, to settings with four groups, or a time-course design, where a GLM with a slope and intercept term is most appropriate. We have also developed a Galaxy tool^{20–22} that implements this workflow as a point-and-click application to improve accessibility for researchers who are unfamiliar with the R programming environment (Figure 2).

Discussion

Although the major functionality of edgeR has been developed with RNA-seq data in mind, the analysis of numerous in-house data sets¹⁹ and the results of others⁴ have demonstrated its utility for count data

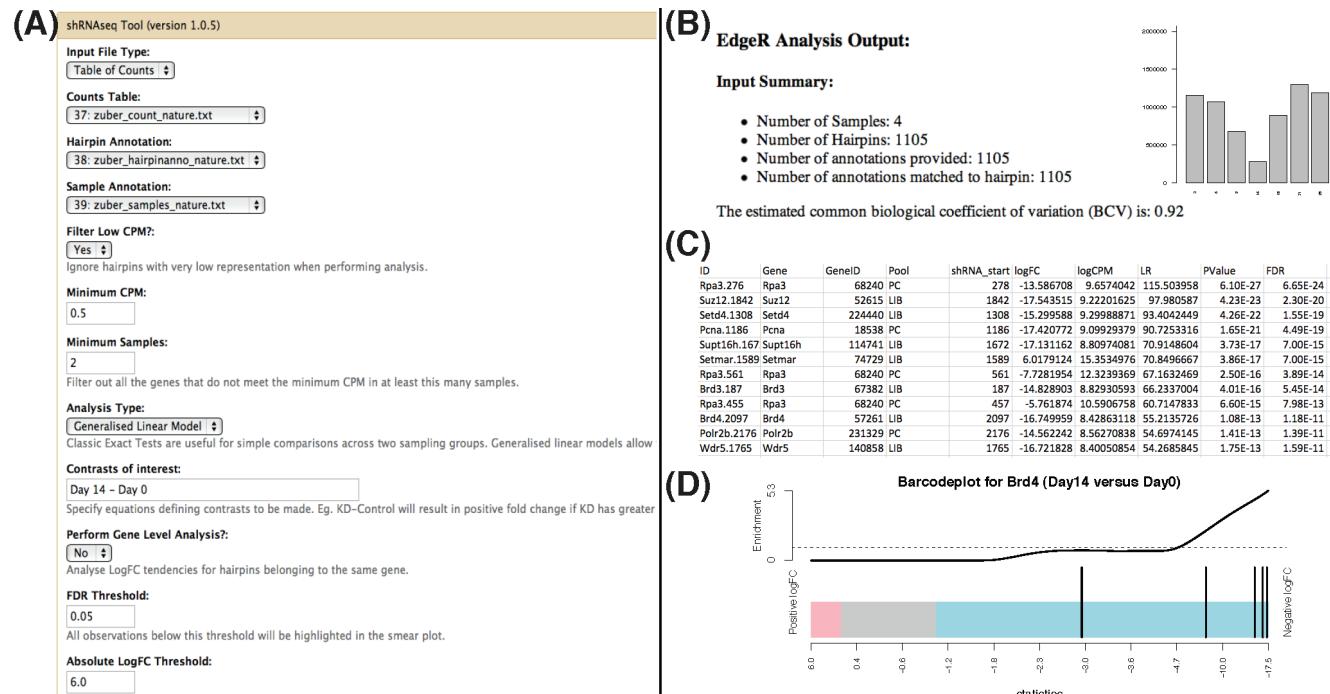


Figure 2. Screenshots of the Galaxy tool for analyzing pooled genetic sequencing screens using edgeR. (A) From the main screen, the user selects the appropriate input files and analysis options. (B) The results of an analysis are summarized in an HTML page that includes various diagnostic plots. (C) Output also includes a table of ranked results at the hairpin/guide and gene-level (where appropriate) as well as barcode plots (D) that highlight the ranks of hairpins/guides targeting a specific gene relative to all other hairpins/guides in the data set.

derived from pooled amplicon sequencing screens, edgeR provides users with a unique tool for the analysis of data from this emerging application of second generation sequencing technology, that is capable of handling both the biological variability and experimental complexity inherent in these screens. Provision of a Galaxy module puts these powerful statistical methods within reach of experimentalists. Future work will be focused on the use of a suitable control data set to compare this analysis pipeline with other approaches such as shRNAsq⁹.

Software availability

Software access

The edgeR software is an R¹³ package distributed as part of the Bioconductor project¹² (<http://www.bioconductor.org>). The Galaxy tool that implements this workflow is available from <http://toolshed.g2.bx.psu.edu/view/shians/shrnaseq>.

Latest source code

<http://www.bioconductor.org/packages/release/bioc/html/edgeR.html>

Archived source code as at the time of publication

<http://dx.doi.org/10.5281/zenodo.12267>²³

Software license

GNU GPL version 2.

Author contributions

ZD and MER developed the sequence processing software and SS developed the Galaxy tool. JMS, LJJ, DLM, SW, SW, LO'C and MEB generated the screen data analyzed in the user guide that accompanies this article and RAD developed the hairpin technology. All authors wrote and approved the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This research was supported by NHMRC Project grants 1050661 (MER) and 1059622 (MER and MEB), Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRISS.

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

Acknowledgements

We thank Matthew Wakefield and Gordon Smyth for advice on data analysis, Cynthia Liu for code testing, Ophir Shalem and Feng Zhang for providing summarized counts from their sgRNA-seq screen and our many collaborators at the WEHI whose research has motivated this work.

References

- Bassik MC, Lebbink RJ, Churchman LS, et al.: **Rapid creation and quantitative monitoring of high coverage shRNA libraries.** *Nat Methods.* 2009; 6(6): 443–5. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang T, Wei JJ, Sabatini DM, et al.: **Genetic screens in human cells using the CRISPR/Cas9 system.** *Science.* 2014; 343(6166): 80–4. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zuber J, Shi J, Wang E, et al.: **RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia.** *Nature.* 2011; 478(7370): 524–8. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sullivan KD, Padilla-Just N, Henry RE, et al.: **ATM and MET kinases are synthetic lethal with nongenotoxic activation of p53.** *Nat Chem Biol.* 2012; 8(7): 646–54. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bassik MC, Kampmann M, Lebbink RJ, et al.: **A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility.** *Cell.* 2013; 152(4): 909–22. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shalem O, Sanjana NE, Hartenian E, et al.: **Genome-scale CRISPR-Cas9 knockout screening in human cells.** *Science.* 2014; 343(6166): 84–7. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tijsterman M, Plasterk RH: **Dicers at RISC; the mechanism of RNai.** *Cell.* 2004; 117(1): 1–3. [PubMed Abstract](#) | [Publisher Full Text](#)
- Mali P, Esvelt KM, Church GM: **Cas9 as a versatile tool for engineering biology.** *Nat Methods.* 2013; 10(10): 957–63. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sims D, Mendes-Pereira AM, Frankum J, et al.: **High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing.** *Genome Biol.* 2011; 12(10): R104. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kim J, Tan AC: **BINGSISL-seq: a bioinformatics pipeline for the analysis and interpretation of deep sequencing genome-wide synthetic lethal screen.** *Methods Mol Biol.* 2012; 802: 389–98. [PubMed Abstract](#) | [Publisher Full Text](#)
- Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; 26(1): 139–40. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gentleman RC, Carey VJ, Bates DM, et al.: **Bioconductor: open software development** for computational biology and bioinformatics. *Genome Biol.* 2004; 5(10): R80. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- R Core Team. **R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing, Vienna, Austria, 2014. [Reference Source](#)
- Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to sage data.** *Biostatistics.* 2008; 9(2): 321–32. [PubMed Abstract](#) | [Publisher Full Text](#)
- McCarthy DJ, Chen Y, Smyth GK: **Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.** *Nucleic Acids Res.* 2012; 40(10): 4288–97. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhou X, Lindsay H, Robinson MD: **Robustly detecting differential expression in RNA sequencing data using observation weights.** *Nucleic Acids Res.* 2014; 42(11): e91. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wu D, Lim E, Vaillant F, et al.: **ROAST: rotation gene set tests for complex microarray experiments.** *Bioinformatics.* 2010; 26(17): 2176–82. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wu D, Smyth GK: **Camera: a competitive gene set test accounting for inter-gene correlation.** *Nucleic Acids Res.* 2012; 40(17): e133. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ritchie ME: **Analysing shRNA-seq data using edgeR supplementary data and documentation.** 2014.
- Giardine B, Riemer C, Hardison RC, et al.: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome Res.* 2005; 15(10): 1451–5. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goecks J, Nekrutenko A, Taylor J: The Galaxy Team. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol.* 2010; 11(8): R86. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Blankenberg D, Von Kuster G, Coraor N, et al.: **Galaxy: a web-based genome analysis tool for experimentalists.** *Curr Protoc Mol Biol.* 2010; Chapter 19: Unit 19.10.1–21. [PubMed Abstract](#) | [Publisher Full Text](#)
- Dai Z, Sheridan JM, Gearing LJ, et al.: **edgeR version 3.8.** Zenodo. 2014. [Data Source](#)

Open Peer Review

Current Referee Status:



Version 2

Referee Report 21 October 2014

doi:[10.5256/f1000research.5936.r6477](https://doi.org/10.5256/f1000research.5936.r6477)



Sumit Deswal

Department of Molecular Immunology, The Research Institute of Molecular Pathology, Vienna, Austria

The authors have addressed the concerns raised in first version of article and even improved the manuscript by incorporating CRISPR-Cas9 screening data analysis in the software tool. So in my opinion manuscript is now ready for indexation.

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.

Version 1

Referee Report 12 May 2014

doi:[10.5256/f1000research.4204.r4548](https://doi.org/10.5256/f1000research.4204.r4548)



Sumit Deswal

Department of Molecular Immunology, The Research Institute of Molecular Pathology, Vienna, Austria

The authors introduce the multiplex shRNA screening approach as commonly used currently to find new drug targets or other similar applications. They also provide a literature overview on the relevant bioinformatics tools used in this context. With rapid improvements in the shRNA technology, data analysis methods for this are high in demand. The manuscript provides a good framework for analysis of data on pooled shRNA screens. There is a strong need to have such user friendly, open-source programs dedicated to analysis of shRNA screens that can easily be used by experimental biologists. The manuscript provides a coherent platform that nicely incorporates analytic tools for the current requirements of the RNAi or other similar functional genetic screens. As part of the review process, we have analyzed our own data using the software and found it fully functional and very easy to use. Based on our experience, we recommend the manuscript for indexation.

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 08 May 2014

doi:[10.5256/f1000research.4204.r4552](https://doi.org/10.5256/f1000research.4204.r4552)



Ross Lazarus

Computational Biology, Baker IDI Heart and Diabetes Institute, Melbourne, Vic., Australia

This is a well written paper describing a potentially useful application for the analysis shRNA-seq screening data. It includes a brief description of this relatively novel technique and some competing methods as well as a description of the method used in this implementation. The authors are to be commended for providing a wrapper for Galaxy which will make it very easy for biologists to access the method in a reproducible analysis environment and this seems likely to improve the real availability and eventual impact of their work.

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 01 May 2014

doi:[10.5256/f1000research.4204.r4546](https://doi.org/10.5256/f1000research.4204.r4546)



James W. MacDonald

Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, USA

This is a useful manuscript, detailing the author's extensions of existing functionality within the Bioconductor edgeR package to include shRNA screen data.

The manuscript itself provides an overview of shRNA screens, competing analysis pipelines, and the methods available in edgeR. In addition, the authors provide a link to a vignette that gives example analyses of four different shRNA experiments that vary in depth of sequencing and complexity, along with the data so potential users can recapitulate the analyses provided, before making an attempt to analyze their own data.

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.