Check for updates

METHOD ARTICLE

A unified GenomeSpace recipe to identify essential genes and associated subnetworks from Genome-Scale CRISPR-Cas9 knockout screens [version 1; peer review: 2 approved with reservations]

Daniel E. Carlin , Forrest Kim, Trey Ideker, Jill P. Mesirov

Department of Medicine, University of California, San Diego, San Diego, CA, 92093, USA

V1 First published: 12 Oct 2018, 7:1636
<https://doi.org/10.12688/f1000research.16290.1>
Latest published: 12 Oct 2018, 7:1636
<https://doi.org/10.12688/f1000research.16290.1>

Abstract

We present a unified GenomeSpace recipe that combines the results of a high throughput CRISPR genetic screen and a biological network to return a subnetwork that suggests a mechanistic explanation of the screen's results. The explanatory subnetwork is found by network propagation, a popular systems biology approach. We demonstrate our pipeline on an alpha toxin screen, revealing a subnetwork that is both highly interconnected and highly enriched for hits in the screen.

Keywords

genetic screen, network biology, network propagation, CRISPR

This article is included in the **Bioinformatics** gateway.

This article is included in the **Genomics and Genetics** gateway.

This article is included in the **GenomeSpace** collection.

Open Peer Review

Approval Status

	1	2
version 1
12 Oct 2018	view	view

1. **Shouhong Guang**, University of Science and Technology of China, Hefei, China

2. **Xiaowei Wang**, Washington University School of Medicine, St Louis, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Daniel E. Carlin (dcarlin@ucsd.edu)

Author roles: **Carlin DE:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kim F:** Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Visualization, Writing – Review & Editing; **Ideker T:** Funding Acquisition, Supervision; **Mesirov JP:** Conceptualization, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: Trey Ideker is a co-founder of Data4Cure has an equity interest in this company. Data4Cure creates and markets scientific software related to computer-aided cancer diagnostics and drug design. Trey Ideker is also a consultant for and have an equity interest in Ideaya Biosciences. Ideaya is developing pharmaceuticals for cancer therapy using principles of systems biology.

Grant information: This work is supported by the National Institute of Health and National Human Genome Research Institute project number 5U41HG007517-05.

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

Copyright: © 2018 Carlin DE *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Carlin DE, Kim F, Ideker T and Mesirov JP. **A unified GenomeSpace recipe to identify essential genes and associated subnetworks from Genome-Scale CRISPR-Cas9 knockout screens [version 1; peer review: 2 approved with reservations]** F1000Research 2018, 7:1636 <https://doi.org/10.12688/f1000research.16290.1>

First published: 12 Oct 2018, 7:1636 <https://doi.org/10.12688/f1000research.16290.1>

Introduction

The rise of next generation sequencing technology and CRISPR gene editing technology has opened up new opportunities for high throughput genetic screens. Increasingly systems biology and molecular networks are becoming more important in the analysis of the mechanisms that are implicated in these screens. Here we present a GenomeSpace recipe providing a standardized pipeline for combining the analysis of a screen with networks and represents a logical next step in providing user-friendly bioinformatics workflows for these types of screens.

This recipe provides a way to process the results of a CRISPR-Cas9 genome wide knockout screen. In such a screen, single guide RNAs (sgRNAs) are designed to target and knock down genes by binding to the target gene and introducing double strand DNA breaks. (Koike-Yusa *et al.*, 2014). Those bound mRNA are subsequently digested by the Cas9 complex and thus do not yield a gene product. In a cell, if the sgRNA is introduced for a gene that is essential for the survival of the cell, that cell will die, and the sgRNA will be depleted. Thus, by sequencing the sgRNAs and looking for a depletion of the sgRNAs targeting a particular gene, we can infer the essentiality of that target gene. Since a large number of sgRNAs can be introduced in a single screen, the essentiality of many genes can be tested at once. However, there are challenges that can arise in the normalization and processing of the read counts; often more than one sgRNA corresponds to a gene but with different efficiency, and significant biases exist in sequencing different sgRNAs. For these reasons the MAGeCK (Li *et al.*, 2014) method was developed to handle data resulting from such a screen.

On the systems biology side of the analysis, we have chosen to employ network propagation as a method of identifying subnetworks representing inferred mechanisms that are implicated by the CRISPR screen. Network propagation has become an essential tool in many network applications; it has been used to identify mechanisms of cancer (Leiserson *et al.*, 2015), to implicate genes in GWAS studies (Qian *et al.*, 2014), and to find functional modules (Vanunu *et al.*, 2010). Network propagation considers genes as nodes on the graph of a biological network. It performs a random walk along the edges of the graph from a set of query nodes. We expect that genes that are implicated in a phenotype will occur in regions of the network that represent mechanisms that are relevant to the screen conditions, and so the random walk will be likely to land on relevant genes. Genes that are near query nodes are therefore implicated by association. For a review of the many flavors and applications of network propagation, see (Cowen *et al.*, 2017).

Methods

An overview of the pipeline appears in Figure 1. The recipe begins using the raw read counts of sgRNAs as input to the MAGeCK module in GenePattern (www.genepattern.org). After normalizing the data, MAGeCK detects differential read counts for each sgRNA using an over-dispersed Poisson model. Next, it detects statistical underrepresentation of the sgRNAs corresponding to particular genes to infer that a gene is essential for the survival of a cell. The reasoning behind this is that if

a sgRNA targets an essential gene, the cells that contain that sgRNA will not replicate and the sgRNA will be underrepresented compared to other genes.

After we determine a set of essential genes, we pass their identity via GenomeSpace to Cytoscape (Shannon *et al.*, 2003). All analysis in Cytoscape is based on networks, thus a relevant reference molecular network must be imported; the recipe uses the NDEx database (Pratt *et al.*, 2015) to identify such a network. In this case, we choose the National Cancer Institute's Pathway Interaction Database (Schaefer *et al.*, 2009). The set of essential genes, i.e., the hits from the genetic screen, are imported from GenomeSpace as a table, then used as the seed nodes for network propagation.

The propagation process starts with a single unit of “heat” on each of the nodes that represent the genes that are found to be underrepresented in the screen, and therefore essential for the growth of the cell. We use a heat diffusion process, treating the network as an unweighted, undirected graph. Heat diffusion smooths the original signal over the network, iteratively passing the signal on each node to its neighbor. It identifies regions of the network that have a high concentration of hits. Here we use a time parameter (which represents the amount of time that the heat is allowed to diffuse over the graph) of 0.1. This is a common choice for the time parameter (see Paull *et al.*, 2013). The recipe employs the network diffusion service built into Cytoscape natively (Carlin *et al.*, 2017).

Next, applying a cutoff of the top 200 genes with the most heat after diffusion, we choose a subnetwork that has a high concentration of hits. Finally, in order to understand the composition of the subnetwork, we apply the GeneMANIA Cytoscape plugin. For any network, GeneMANIA (Montejo *et al.*, 2010) shows what functional Gene Ontology categories are enriched in that network.

Use case

We used a previously published CRISPR study (Koike-Yusa *et al.*, 2014) to illustrate the use of this pipeline. In this study, the authors use mouse embryonic stem cells grown in the presence of alpha-toxin. This screen was therefore designed to expose the genes involved in the mechanism of resistance to the toxin. The largest connected network component of the top 200 genes after propagation appears in Figure 2.

The results of the GeneMania enrichment suggest that DNA repair is the single most important gene set in handling alpha-toxin. This is consistent with the findings in (Bantel *et al.*, 2001) that alpha-toxin causes an influx of monovalent ions that can cause DNA fragmentation. In the absence of DNA repair machinery, the cells cannot recover from this stress and therefore die. The complete table of the Gene Ontology terms that were significantly enriched in the subnetwork appear in Table S1.

Variations

There are several variations that can be used depending on preferences of the user. For example, different tools such as DESeq

Figure 1. Workflow for the recipe. This recipe shows how using GenomeSpace seamlessly integrates multiple bioinformatics tools into a single, easily reproducible pipeline. The publicly available preprocessed knockout screen files and sgRNA library from Koike-Yusa *et al.* can be transferred directly from GenomeSpace to the ported MAGECK module in GenePattern (Li *et al.*, 2014). By exporting the resulting list of significant genes to GenomeSpace, the data can be imported directly to Cytoscape without having to download the files locally. Cytoscape's integrated plugins for NDEx (Pratt *et al.*, 2015), Network Diffusion (Carlin *et al.*, 2017; Cowen *et al.*, 2017), and GeneMANIA (Montejo *et al.*, 2010) allow for the remainder of the recipe to be completed within its user-friendly environment.

Figure 2. The final subnetwork associated with alpha toxin resistance. The black nodes represent genes that are significantly depleted in a CRISPR screen. Grey nodes represent genes that are closely associated with the hits by the network and are scaled by the strength of their association.

(Anders & Huber, 2010) and edgeR (Robinson *et al.*, 2010) can be used to identify the hits. Also, the final biological interpretation of the subnetworks was performed with the GeneMania plugin, but the gene list can also be exported and interpreted by another annotation tool. Another approach is to export the gene list and corresponding heats using GenomeSpace and use the Molecular Signature Database gene set overlap tool (<http://software.broadinstitute.org/gsea/msigdb/index.jsp>), (Liberzon *et al.*, 2011) applied to the genes in the identified subnetwork.

Data and software availability

The recipe and Koike-Yusa *et al.* datasets are publicly available at <http://recipes.genomespace.org/view/75>. GenomeSpace, an open-source bioinformatics tool, serves as the data highway allowing for seamless transfer of information between tools, and can be found at <http://www.genomespace.org/>. The MAGeCK

algorithm has been wrapped as a GenePattern module, which can be run locally or on the public GenePattern servers at <http://genepattern.org>. Additionally, GenePattern has added Jupyter Notebook compatibility through GenePattern Notebook (<http://genepattern-notebook.org/>). Finally, Cytoscape and all associated plugins (ie. GeneMANIA and NDEx) can be found at <http://www.cytoscape.org/>.

Grant information

This work is supported by the National Institute of Health and National Human Genome Research Institute project number 5U41HG007517-05.

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

Supplementary material

Table S1: Enriched Gene Ontology categories for the network appearing in Figure 2.

[Click here to access the data](#)

References

- Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol.* 2010; **11**(10): R106.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bantel H, Sinha B, Domschke W, *et al.*: **alpha-Toxin is a mediator of *Staphylococcus aureus*-induced cell death and activates caspases via the intrinsic death pathway independently of death receptor signaling.** *J Cell Biol.* 2001; **155**(4): 637–48.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Carlin DE, Demchak B, Pratt D, *et al.*: **Network propagation in the cytoscape cyberinfrastructure.** *PLoS Comput Biol.* 2017; **13**(10): e1005598.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cowen L, Ideker T, Raphael BJ, *et al.*: **Network propagation: a universal amplifier of genetic associations.** *Nat Rev Genet.* 2017; **18**(9): 551–562.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Koike-Yusa H, Li Y, Tan EP, *et al.*: **Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library.** *Nat Biotechnol.* 2014; **32**(3): 267–73.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Leiserson MD, Vandin F, Wu HT, *et al.*: **Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.** *Nat Genet.* 2015; **47**(2): 106–14.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li W, Xu H, Xiao T, *et al.*: **MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens.** *Genome Biol.* 2014; **15**(12): 554.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liberzon A, Subramanian A, Pinchback R, *et al.*: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics.* 2011; **27**(12): 1739–40.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Montejo J, Zuberi K, Rodriguez H, *et al.*: **GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop.** *Bioinformatics.* 2010; **26**(22): 2927–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Paull EO, Carlin DE, Niepel M, *et al.*: **Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).** *Bioinformatics.* 2013; **29**(21): 2757–64.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pratt D, Chen J, Welker D, *et al.*: **NDEx, the Network Data Exchange.** *Cell Syst.* 2015; **1**(4): 302–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Qian Y, Besenbacher S, Mailund T, *et al.*: **Identifying disease associated genes by network propagation.** *BMC Syst Biol.* BioMed Central, 2014; **8** Suppl 1: S6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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](#)
- Schaefer CF, Anthony K, Krupa S, *et al.*: **PID: the pathway interaction database.** *Nucleic Acids Res.* 2009; **37**(Database issue): D674–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–504.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vanunu O, Magger O, Ruppin E, *et al.*: **Associating genes and protein complexes with disease via network propagation.** *PLoS Comput Biol.* 2010; **6**(1): e1000641.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 03 September 2019

<https://doi.org/10.5256/f1000research.17795.r52690>

© 2019 Wang X. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Xiaowei Wang

Department of Radiation Oncology, Washington University School of Medicine, St Louis, MO, USA

The authors present a bioinformatic recipe for network-based functional analysis of CRISPR screening data. CRISPR screening has quickly become a mainstream tool for functional genomics studies. However, bioinformatics tools are lacking in this emerging field. This study provides a timely bioinformatic framework for functional analysis of CRISPR screening data. This new pipeline combines multiple existing bioinformatics tools and provides a streamlined data analysis process for general users, especially for those who are not experts in bioinformatics. To demonstrate the utility of the pipeline, the authors present a specific example by analyzing a public dataset. Overall, this study is well described. The website also provides very detailed step-by-step instruction for users to follow the recipe. I have a couple of suggestions for further improvement.

1. It would be helpful to provide additional details on individual components of the recipe, such as the rationale behind adopting the widely-used MAGeCK tool (i.e. summarizing its advantages) for robust identification of sgRNA hits. Similarly, there are multiple tools available for network propagation, and it is not clear why a heat diffusion process is adopted by this recipe (i.e. summarizing its advantages).
2. It would be helpful to implement version control for this presented recipe. It is likely that new bioinformatics tools, such as new versions of MAGeCK (or similar tools) or improved strategies for network propagation, will be available in the near future. Accordingly, this recipe needs to be updated to accommodate the latest progress in the bioinformatics field.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics; Genomics; miRNA; CRISPR;

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 24 June 2019

<https://doi.org/10.5256/f1000research.17795.r49975>

© 2019 Guang S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shouhong Guang

School of Life Sciences, University of Science and Technology of China, Hefei, China

The CRISPR gene editing technology followed by next generation sequencing technology has brought up new means for high throughput genetic screening. However, a streamlined and systematic method for downstream data analysis is required to examine these high throughput screenings. In this work, the authors provide a pipeline for combining the analysis of a screen with network illustrations that facilitate the assay of this type of screens. The work is timely and applicable for a lot of related research.

I am not an expert in bioinformatics. Therefore, I will provide several comments from the biological side.

1. In a CRISPR-mediated high throughput genetic screening, people frequently apply several sgRNAs to target a single gene, which exhibit different knocking out efficiencies. In addition, these sgRNAs may target distinct isoforms of a gene, which may lead to different knocking out phenotypes. How the pipeline deals with these sophisticated cases requires elaboration.
2. The author used one example, the alpha toxin resistance screen, to validate their method. The analysis of another independent screen is needed to test its validity.
3. During the analysis of alpha-toxin screen, the authors conclude that the result is consistent with previous findings. This argument requires more detailed comparison with previous

work. In addition, if possible, at least some of the new hits coming out of the work could be validated by wet lab experiments.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genetics, epigenetics, small RNAs, C. elegans.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research