SAVE: A secure cloud-based pipeline for CRISPR pooled screen deconvolution

Hyun-Hwan Jeong[1,2], Seon Young Kim[1,2], Maxime WC Rousseaux[1,2], Huda Y Zoghbi[1,2,3,4,5], Zhandong Liu[2,3]*

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA, [2]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, Texas 77030, USA, [3]Department of Pediatrics, Baylor College of Medicine, Houston, Texas 77030, USA, [4]Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA, [5]Howard Hughes Medical Institute, Baylor College of Medicine, Houston, Texas 77030, USA

*To whom correspondence should be addressed.
Contract: zhandong.liu@bcm.edu

Abstract

We present a user-friendly, cloud-based, data analysis pipeline for the deconvolution of pooled screening data. This tool, termed SAVE for Screening Analysis Visual Explorer, serves a dual purpose of extracting, clustering and analyzing raw next generation sequencing files derived from pooled screening experiments while at the same time presenting them in a user-friendly way on a secure web-based platform. Moreover, SAVE serves as a useful web-based analysis pipeline for reanalysis of pooled CRISPR screening datasets. Taken together, the framework described in this study is expected to accelerate development of web-based bioinformatics tool for handling all studies which include next generation sequencing data. SAVE is available at http://save.nrihub.org.

Introduction

Genetic screening allows for the unbiased interrogation of a genome to ask targeted questions [1]. While originally centered around random mutagenesis or arrayed RNAi [2, 3], a growing amount of studies are moving towards large-scaled, pooled approaches [4-7]. Indeed, the advent of CRISPR has heralded a dramatic increase in the number of pooled screens [8]. For instance, in the past year, the number of datasets for CRISPR screens in Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/gds) have more than tripled from the previous year (39 datasets in 2015 and 121 datasets in 2016). In these, a collection of modifiers (e.g. shRNAs or sgRNAs) are introduced to a population (e.g. in cell culture or into an organism) blindly. Following phenotypic (e.g. enriching for surviving cells following a toxic insult) or enrichment-based (e.g. FACS sorting a given fluorophore) screening, genomic DNA is extracted and enrichment of each modifier is assessed using microarray or, more commonly, next generation sequencing. Once a daunting and expensive approach to generate hypotheses in an unbiased fashion, pooled screening approaches are now more feasible and accessible than ever. Notably, cheaper sequencing costs [9], readily available libraries via non-profit entities (e.g. Addgene) and more robust gene-disruption strategies (e.g. CRISPR and CRISPRi ([10] and [11] now allow for global gene disruption-based screens in an unprecedented manner.

Nevertheless, the bioinformatics hurdle of data deconvolution following sequencing remains a roadblock for investigators with little-to-no computational knowledge. Previous tools have been generated to analyze high-throughput screening datasets but, like most bioinformatics tools, these did not always provide a good user experience (Table 1). There are several reasons to explain these shortcomings. First, these tools often require a self-installation step to use and often require the combination of several packages to work adequately. Second, raw sequence data handling is not provided and therefore users must independently manipulate the data (leading to undue errors) by uploading their data to platforms such as RIGER and ATARI [12, 13]. Third, most available tools are specialized for dropout studies such that enrichment-based studies (or any other study with more complex data involving 3 or more groups) does not fit in the analysis pipeline. Lastly, the inflexibility in statistical analysis output in these programs may lead to improper conclusions.

To overcome these issues, we generated a web-based analysis pipeline for pooled screens: SAVE. SAVE (Screening Analysis Visual Explorer, http://save.nrihub.org) is a user-friendly pipeline that lends itself to customization according to the screen at hand. In this system, the user can upload raw sequencing files confidentially, can extract valuable screening information through customizable statistical analysis and can output various end-point results set by the user.

Methods

To generate this platform, we first had to overcome the following issues: File size and confidentiality as well as multifunctionality in output formats. First, the file size of next generation sequencing datasets from pooled screening is big (about 10GB for each sample), so it would take much time to upload every dataset directly without any manipulation. Second, given that users are uploading raw data to the server, it is critical to minimize the information uploaded for analysis, to avoid unwanted leaking of user’s data. Third, we wanted to develop a web-based bioinformatics tool that was easy to handle and generate comprehensive visualizations for those performing the screens who are not necessarily well versed in bioinformatics.

Fortunately, the client-side web technology like JavaScript and HTML5 have rapidly developed in the recent year, and those technologies have enabled the high-performance computation for large sized file handling using modern web browsers such as Google Chrome, Mozilla Firefox, and Microsoft Edge [14]. Thus, we can perform the initial steps of data analysis at the *client-side* with the technologies, allowing for rapid and safe data minimization. Since a web-based analysis system with a centralized server may run slower than the expectation of the user, one must overcome this limitation to allow for proper web-based integration. Cloud computing offers such a solution which provides powerful and scalable computational resources, such as CPU and storage, to meet users’ expected running time at an economical efficient manner.

To address the aforementioned problems, we developed SAVE (Screening Analysis Visual Explorer) - a secure cloud-based pipeline analysis and data management for high-throughput pooled CRISPR screening - and have started to provide the public service (Figure 1). This service gives a user-friendly graphical interface (Figure 2) and it minimizes the data sent to the server by preprocessing FASTQ files on the user's computer. The only data the server gets is the counts matrix. Also, SAVE provides a scalable service through the infrastructure provided by Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). It supports two different types of study (dropout and enrichment based experiment) with various proper statistics (a Student's t-test with log-transformation of counts [15], DESeq2 [16], and the inverted beta-binomial test -- a test used for paired datasets [17]). Also, users can get a visualized web report for their data and export the results via an address-encrypted link (Figure 3).

Discussion

The decreasing costs and time-frames associated with pooled genome-wide screening approaches is incentivizing an increasing number of investigators. Nevertheless, the sequencing-to-data gap remains discouraging. We therefore developed SAVE as a solution to this issue. The multifunctionality of SAVE is clear with current established technologies such as shRNA and CRISPR but can easily be extended to any technology where short sequence barcoding is used (e.g. ORF screens and CRISPRi/a) [10] [11] [18]. Moreover, this user-friendly system allows for re-analysis of other established datasets in the context of one's statistical analysis parameters (i.e. if you filter using certain cut-offs and want to confirm with someone else's dataset, you can do so with SAVE). Lastly, the implications of this software extent passed screening approaches and are potentially applicable to any datasets concerning multiple sequencing files that require rapid deconvolution. One example of these would be single-cell RNA-seq experiments where massive amounts of sequencing results must be aligned, quantified and presented in a user-friendly manner.

Taken together, this study highlights a web-based platform upon which investigators can securely deposit raw sequencing files, extract the valuable data and perform statistical analyses on them, generate and prioritize hit lists and export datasets for downstream validation. Thus, this software passes the last hurdle for this type of project, enabling screen-based biological analysis for any interested user.

Reference

1. Miles, L.A., R.J. Garippa, and J.T. Poirier, *Design, execution, and analysis of pooled in vitro CRISPR/Cas9 screens.* FEBS J, 2016. **283**(17): p. 3170-80.
2. Simon, M.M., et al., *Current strategies for mutation detection in phenotype-driven screens utilising next generation sequencing.* Mamm Genome, 2015. **26**(9-10): p. 486-500.
3. Mohr, S.E., et al., *RNAi screening comes of age: improved techniques and complementary approaches.* Nat Rev Mol Cell Biol, 2014. **15**(9): p. 591-600.
4. Kampmann, M., M.C. Bassik, and J.S. Weissman, *Functional genomics platform for pooled screening and generation of mammalian genetic interaction maps.* Nat Protoc, 2014. **9**(8): p. 1825-47.
5. Schlabach, M.R., et al., *Cancer proliferation gene discovery through functional genomics.* Science, 2008. **319**(5863): p. 620-4.
6. Shalem, O., et al., *Genome-scale CRISPR-Cas9 knockout screening in human cells.* Science, 2014. **343**(6166): p. 84-7.
7. Wang, T., et al., *Genetic screens in human cells using the CRISPR-Cas9 system.* Science, 2014. **343**(6166): p. 80-4.
8. Shalem, O., N.E. Sanjana, and F. Zhang, *High-throughput functional genomics using CRISPR-Cas9.* Nat Rev Genet, 2015. **16**(5): p. 299-311.
9. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
10. Gilbert, L.A., et al., *Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation.* Cell, 2014. **159**(3): p. 647-61.
11. Horlbeck, M.A., et al., *Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation.* Elife, 2016. **5**.
12. Luo, B., et al., *Highly parallel identification of essential genes in cancer cells.* Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20380-5.
13. Shao, D.D., et al., *ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens.* Genome Res, 2013. **23**(4): p. 665-78.
14. *Node.js vs C++ g++ (64-bit Ubuntu quad core)*. 2017; Available from: http://benchmarksgame.alioth.debian.org/u64q/measurements.php?lang=node.
15. Wu, Z.J., et al., *Gene expression profiling of human breast tissue samples using SAGE-Seq.* Genome research, 2010. **20**(12): p. 1730-1739.
16. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
17. Pham, T.V. and C.R. Jimenez, *An accurate paired sample test for count data.* Bioinformatics, 2012. **28**(18): p. i596-i602.
18. Konermann, S., et al., *Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex.* Nature, 2015. **517**(7536): p. 583-8.
19. Li, W., et al., *Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR.* Genome Biol, 2015. **16**: p. 281.
20. *ScreenProcessing*. 2016; Available from: https://github.com/mhorlbeck/ScreenProcessing.
21. Winter, J., et al., *caRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens.* Bioinformatics, 2016. **32**(4): p. 632-4.

Figures and Tables

Table 1. A comparison of SAVE with existing method for the multifunctionality

Feature	RIGER	ATARIS	MAGeCK-VISPR	ScreenProcessing	caRpools	**SAVE**
Web interface	✓	✓				✓
Cloud service						✓
FASTQ file handling			✓	✓	✓	✓
GUI		✓			✓	✓
Multiple group studies			✓	✓	✓	✓
Paired data analysis						✓
Quality control			✓		✓	✓
Reference	[12]	[13]	[19]	[20]	[21]	

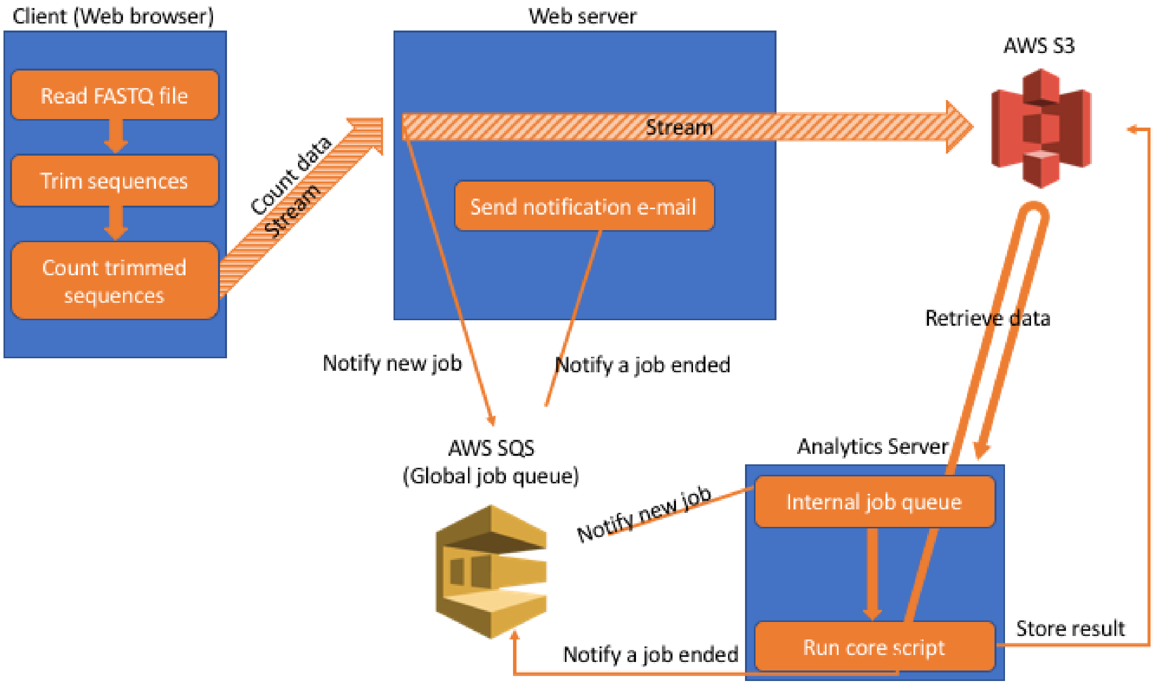
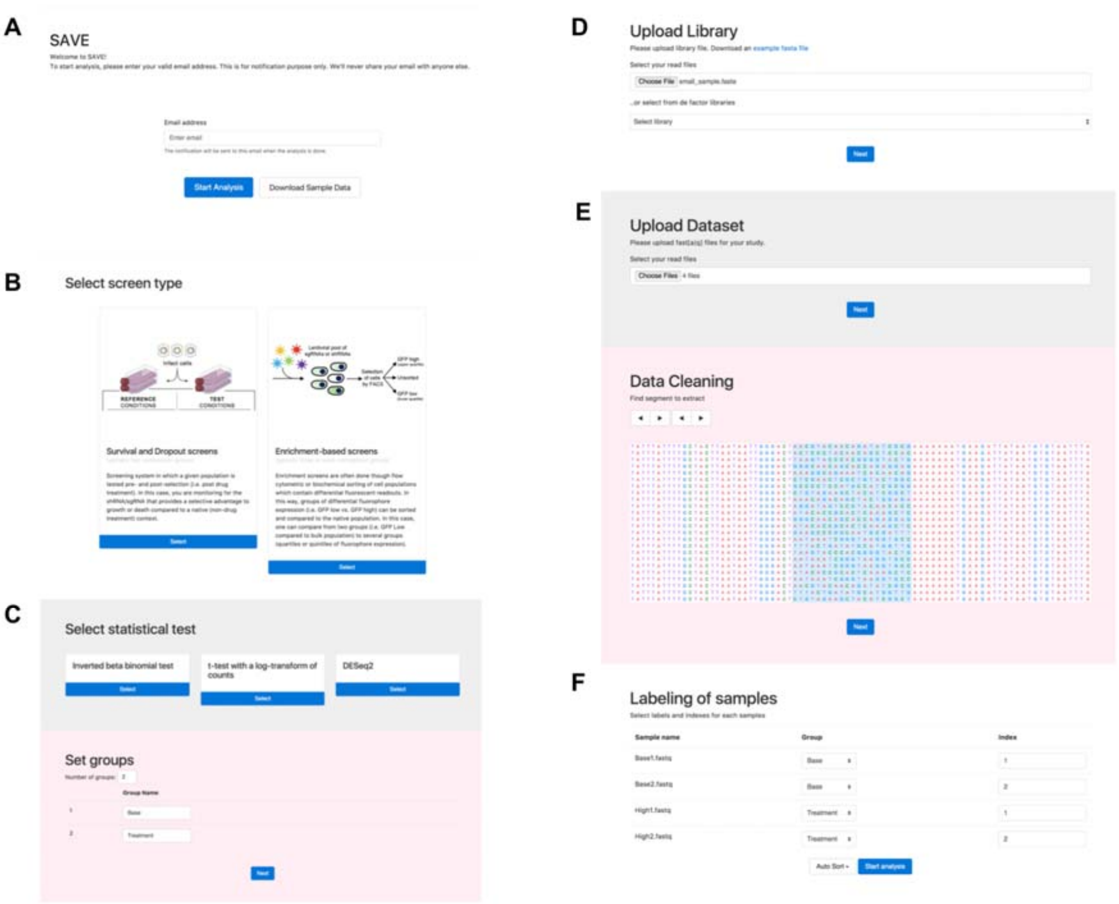


Figure 1. A detailed illustration of the pipeline of SAVE



SAVE

Welcome to SAVE!

To start analysis, please enter your valid email address. This is for notification purpose only. We'll never share your email with anyone else.

Email address: Enter email

The notification will be sent to this email when the analysis is done.

Select screen type

Survival and Dropout screens

Screening system in which a given population is tested pre- and post-selection (i.e. post drug treatment). In this case, you are monitoring for the shRNA/sgRNA that provides a selective advantage to growth or death compared to a native (non-drug treatment) context.

Enrichment-based screens

Enrichment screens are often done though flow cytometric or biochemical sorting of cell populations which contain differential fluorescent readouts. In this way, groups of differential fluorophore expression (i.e. GFP low vs. GFP high) can be sorted and compared to the native population. In this case, one can compare from two groups (i.e. GFP Low compared to bulk population) to several groups (quartiles or quintiles of fluorophore expression).

Select statistical test

Inverted beta binomial test

t-test with a log-transform of counts

DESeq2

Set groups

Number of groups: 2

	Group Name
1	Base
2	Treatment

Upload Library

Please upload library file. Download an example fasta file

Select your read files: email_sample.fasta

...or select from de factor libraries: Select library

Upload Dataset

Please upload fast(a/q) files for your study.

Select your read files: 4 files

Data Cleaning

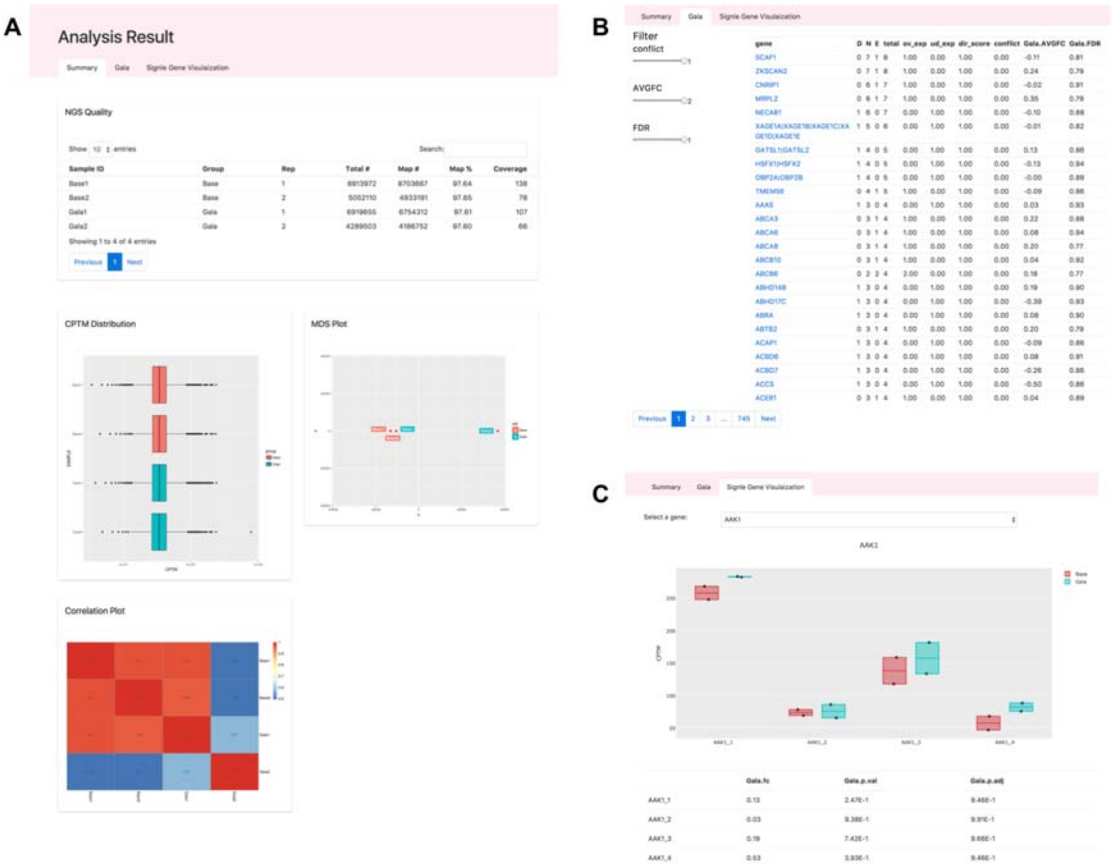
Find segment to extract

Labeling of samples

Select labels and indexes for each samples

Sample name	Group	Index
Base1.fastq	Base	1
Base2.fastq	Base	2
High1.fastq	Treatment	1
High2.fastq	Treatment	2

Figure 2. Screenshots of GUI to input user data in the SAVE website: (A) The first page to input user email address. (B) A page for the selection of a type of the experiment. (C) choice of statistical test and add labels for each group. (D) A page for the upload sgRNA Library. (E) A page for the upload FASTQ file and page to choose a region for each read of deep sequencing data to extract the sgRNA for the input sequences. (F) a page for labeling of samples to design the study.



NGS Quality

Show 10 entries

Search:

Sample ID	Group	Rep	Total #	Map #	Map %	Coverage
Base1	Base	1	8913972	8703667	97.64	138
Base2	Base	2	5052110	4933191	97.65	78
Gala1	Gala	1	6919655	6754312	97.61	107
Gala2	Gala	2	4289503	4186752	97.60	66

Showing 1 to 4 of 4 entries

Previous 1 Next

gene	D	N	E	total	ov_exp	ud_exp	dir_score	conflict	Gala.AVGFC	Gala.FDR
SCAF1	0	7	1	8	1.00	0.00	1.00	0.00	-0.11	0.81
ZKSCAN2	0	7	1	8	1.00	0.00	1.00	0.00	0.24	0.79
CNRIP1	0	6	1	7	1.00	0.00	1.00	0.00	-0.02	0.91
MRPL2	0	6	1	7	1.00	0.00	1.00	0.00	0.35	0.79
NECAB1	1	6	0	7	0.00	1.00	1.00	0.00	-0.10	0.88
XAGE1A|XAGE1B|XAGE1C|XAGE1D|XAGE1E	1	5	0	6	0.00	1.00	1.00	0.00	-0.01	0.82
GATSL1|GATSL2	1	4	0	5	0.00	1.00	1.00	0.00	0.13	0.86
HSFX1|HSFX2	1	4	0	5	0.00	1.00	1.00	0.00	-0.13	0.94
OBP2A|OBP2B	1	4	0	5	0.00	1.00	1.00	0.00	-0.00	0.89
TMEM56	0	4	1	5	1.00	0.00	1.00	0.00	-0.09	0.86
AAAS	1	3	0	4	0.00	1.00	1.00	0.00	0.03	0.93
ABCA3	0	3	1	4	1.00	0.00	1.00	0.00	0.22	0.88
ABCA6	0	3	1	4	1.00	0.00	1.00	0.00	0.08	0.94
ABCA8	0	3	1	4	1.00	0.00	1.00	0.00	0.20	0.77
ABCB10	0	3	1	4	1.00	0.00	1.00	0.00	0.04	0.92
ABCB6	0	2	2	4	2.00	0.00	1.00	0.00	0.18	0.77
ABHD14B	1	3	0	4	0.00	1.00	1.00	0.00	0.19	0.90
ABHD17C	1	3	0	4	0.00	1.00	1.00	0.00	-0.39	0.93
ABRA	1	3	0	4	0.00	1.00	1.00	0.00	0.08	0.90
ABTB2	0	3	1	4	1.00	0.00	1.00	0.00	0.20	0.79
ACAP1	1	3	0	4	0.00	1.00	1.00	0.00	-0.09	0.86
ACBD6	1	3	0	4	0.00	1.00	1.00	0.00	0.08	0.91
ACBD7	1	3	0	4	0.00	1.00	1.00	0.00	-0.26	0.86
ACCS	1	3	0	4	0.00	1.00	1.00	0.00	-0.50	0.86
ACER1	0	3	1	4	1.00	0.00	1.00	0.00	0.04	0.89

Previous 1 2 3 … 745 Next

Select a gene: AAK1

	Gala.fc	Gala.p.val	Gala.p.adj
AAK1_1	0.13	2.47E-1	9.46E-1
AAK1_2	0.03	8.38E-1	9.91E-1
AAK1_3	0.19	7.42E-1	9.66E-1
AAK1_4	0.53	3.93E-1	9.46E-1

Figure 3. An example of the visualized report of SAVE: (A) Overview of quality of user's deep sequencing data. (B) Statistics page to select hit genes (C) A boxplot of read count of sgRNAs for a gene