Bioinformatics, 32(4), 2016, 632–634
doi: 10.1093/bioinformatics/btv617
Advance Access Publication Date: 27 October 2015
Applications Note

Systems biology

caRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens

Jan Winter[1,†], Marco Breinig[1,†], Florian Heigwer[1], Dirk Brügemann[1], Svenja Leible[1], Oliver Pelz[1], Tianzuo Zhan[1,2] and Michael Boutros[1,*]

[1]German Cancer Research Center (DKFZ), Division Signaling and Functional Genomics and Heidelberg University, Heidelberg, Germany and [2]Department of Medicine II, University Hospital Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on August 19, 2015; revised on October 13, 2015; accepted on October 16, 2015

Abstract

Motivation: Genetic screens by CRISPR/Cas9-mediated genome engineering have become a powerful tool for functional genomics. However, there is currently a lack of end-to-end software pipelines to analyze CRISPR/Cas9 screens based on next generation sequencing.
Results: The CRISPR-AnalyzeR for pooled screens (caRpools) is an R package for exploratory data analysis that provides a complete workflow to analyze CRISPR/Cas9 screens. To further support the analysis of large-scale screens, caRpools integrates screening documentation and generation of standardized analysis reports.
Availability and implementation: caRpools, manuals and an open virtual appliance are available at http://github.com/boutroslab/caRpools.
Contact: m.boutros@dkfz.de

1 Introduction

CRISPR/Cas9-mediated genome engineering can be leveraged for high-throughput functional genomic screens in vertebrate cells (Koike-Yusa *et al.*, 2014; Shalem *et al.*, 2014; Wang *et al.*, 2014). For such screens, Cas9-expressing cells are infected with lentiviral short guide (sg) RNA libraries. After phenotypic selection, sgRNA sequences serve as barcodes to identify enriched or depleted mutant clones by next generation sequencing (NGS). Inter-experimental variations in sequencing quality and depth, variable sgRNA coverage per gene, on-target efficiency and off-target effects represent challenges for analysis and hit identification (Diaz *et al.*, 2015; Li *et al.*, 2014). The software package CRISPR-AnalyzeR for Pooled Screens (caRpools) performs exploratory data analysis of CRISPR/Cas9 screens combined with detailed screening documentation to enable reproducibility of analyses workflows (Boutros *et al.*, 2006; Pelz *et al.*, 2010).

2 The caRpools package

The R package caRpools is available as Source code and an open virtual appliance from a public Github repository at http://github.com/boutroslab/caRpools, which provides a convenient way to analyze screens without much prior R knowledge. Parameters relevant to caRpools' analysis options are adjusted via an Excel file. We provide a manual to guide users through the installation and running steps. To showcase caRpools' functionalities, we supply example datasets and reports for an unpublished CRISPR/Cas9 screen and caRpools reports for two published screens (Shalem *et al.*, 2014; Wang *et al.*, 2014) (see http://github.com/boutroslab/caRpools). caRpools is based on an end-to-end workflow composed of four modules (Fig. 1).

2.1 Data handling and quality control

caRpools requires a FASTA file with sgRNA identifiers and sequences, FASTQ or processed sgRNA readcount files of samples,

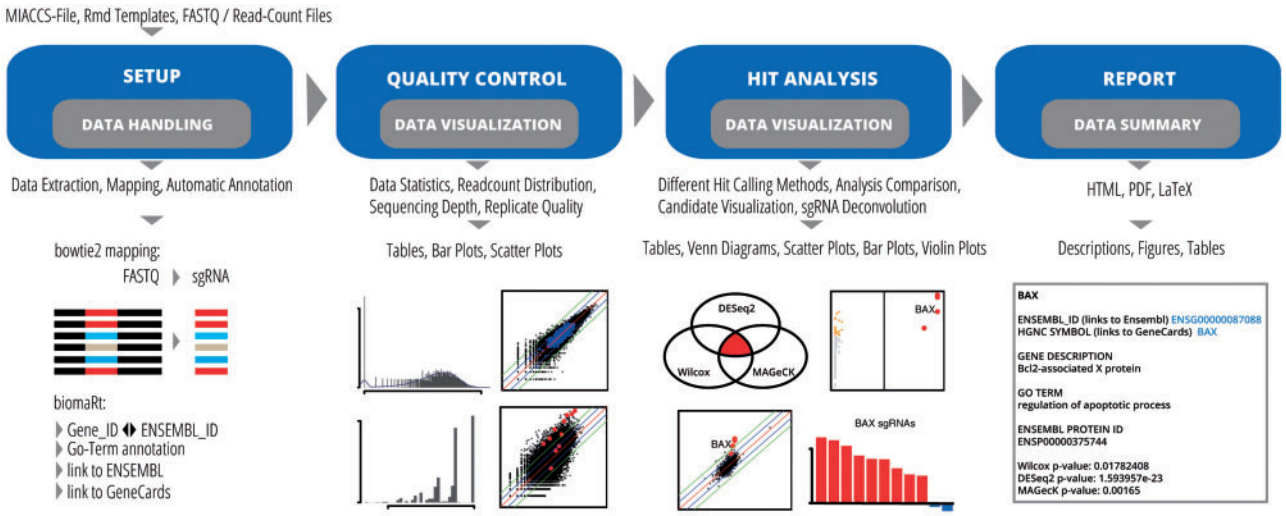

Fig. 1. caRpools workflow and functionalities. Schematic representation of main analysis steps for the analysis of CRISPR/Cas9 screens using caRpools are shown

template files for report generation, and a parameter and screen description file.

caRpools retrieves analysis parameters from a configuration file, which has been adapted from MIARE (http://miare.sourceforge.net/) and supports documentation of screens and analysis steps. File paths for datasets, and parameters required for sgRNA target sequence retrieval from FASTQ files and hit calling must be entered. Further, this file includes general information about screening protocols and settings to e.g. adjust visualizations or to annotate genes with information retrieved from biomaRt. Upon execution, caRpools provides diagnostic plots to evaluate screening performance, including read-count distribution and sequencing depth plots for each sample, as well as sgRNA coverage and evaluation of included control sgRNAs.

2.2 Hit calling and report generation

caRpools offers three methods to rank candidate genes based on the phenotypic effect of all sgRNAs per gene with orthogonal approaches (see CaRpools-Manual.pdf). For the 'Wilcox' approach, read counts are median normalized. The fold change of each population of sgRNAs for a gene is tested against the population of non-targeting control sgRNAs or randomly picked sgRNAs, using a two-sided Mann-Whitney test. For the 'DESeq2'-approach, read counts of all sgRNAs for a given gene are aggregated to generate gene-level read counts. DESeq2 analysis includes size-factor estimations, variance stabilization using a parametric fit and a Wald-Test to determine the difference in $\log_2$ fold changes between the untreated and treated data, essentially as described (Li *et al.*, 2014; Love *et al.*, 2014). For the 'MAGeCK'-approach, a rank-based model is used to test for a change in sgRNA abundance after median normalization of the dataset (Li *et al.*, 2014). Users are asked to familiarize themselves with the specifications and parameter settings of each method by consulting the caRpools user manual and the original literature (Love *et al.*, 2014; Li *et al.*, 2014).

Venn diagrams containing enriched or depleted genes are generated for direct comparison and can reveal differences between hit calling methods. Potential pitfalls, e.g. no identified hits with a defined threshold, or no overlap between methods are reported. Hits are visualized with annotations as well as links to external databases. Since sgRNA efficiency can vary (Diaz *et al.*, 2015; Li *et al.*, 2014; Shalem *et al.*, 2014; Wang *et al.*, 2014), caRpools visualizes the performance of all sgRNAs/gene and lists respective target sequences for all candidates so that users can choose efficient sgRNAs. caRpools supports standardized report generation including ready-to-use plots and tables in a single step.

2.3 Summary and outlook

caRpools is designed to be user-friendly for novice and expert users: caRpools' open virtual appliance allows analysis without prior programming knowledge. For every hit, caRpools provides biological information and links to external databases. Finally, caRpools incorporates detailed screening documentation into the analysis process and generates comprehensive reports. caRpools can be extended to e.g. include novel hit calling algorithms or to export efficient sgRNA designs to external databases such as Protospacer Workbench (MacPherson and Scherf, 2015). caRpools' transparent analysis reports support the establishment of repositories for CRISPR/Cas9 screens and will facilitate meta-analyses of datasets.

Acknowledgement

We thank F. Zhang and T. Wang for reagents and data.

Funding

This work was supported in part by an ERC Advanced Grant, an iMED grant and a DKFZ Postdoc Fellowship to T.Z.

Conflict of Interest: none declared.

References

Boutros,M. *et al.* (2006) Analysis of cell-based RNAi screens. *Genome Biol.*, **7**, R66.

Diaz,A.A. *et al.* (2015) HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Res.*, **43**, e16.

Koike-Yusa,H. *et al.* (2014) Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.*, **32**, 267–73.

Li,W. *et al*. (2014) MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, **15**, 554.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

MacPherson,C.R. and Scherf,A. (2015) Flexible guide-RNA design for CRISPR applications using Protospacer Workbench. *Nat. Biotechnol.*, **33**, 805–806.

Pelz,O. *et al*. (2010) web cellHTS2: a web-application for the analysis of high-throughput screening data. *BMC Bioinformatics*, **11**, 185.

Shalem,O. *et al*. (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.

Wang,T. *et al*. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.