Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR

Li et al. Genome Biology (2015) 16:281
DOI 10.1186/s13059-015-0843-6

Genome Biology

METHOD Open Access

Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR

Wei Li[1,2†], Johannes Köster[1,2,3†], Han Xu[4], Chen-Hao Chen[1,2], Tengfei Xiao[2,3], Jun S. Liu[5], Myles Brown[2,3,6*] and X. Shirley Liu[1,2,7*]

Abstract

High-throughput CRISPR screens have shown great promise in functional genomics. We present MAGeCK-VISPR, a comprehensive quality control (QC), analysis, and visualization workflow for CRISPR screens. MAGeCK-VISPR defines a set of QC measures to assess the quality of an experiment, and includes a maximum-likelihood algorithm to call essential genes simultaneously under multiple conditions. The algorithm uses a generalized linear model to deconvolute different effects, and employs expectation-maximization to iteratively estimate sgRNA knockout efficiency and gene essentiality. MAGeCK-VISPR also includes VISPR, a framework for the interactive visualization and exploration of QC and analysis results. MAGeCK-VISPR is freely available at http://bitbucket.org/liulab/mageck-vispr.

Keywords: CRISPR/Cas9, Screening, Maximum likelihood, Expectation-Maximization, Negative binomial, Data-driven documents, D3, Visualization, Quality control

Background

The clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system is a powerful genetic engineering technique, allowing direct modifications of genomic loci in most model organisms in a cost-effective way. Based on this system, the recent development of high-throughput CRISPR screening technology has shown great promise in functional genomics, allowing researchers to systematically identify genes associated with various phenotypes [1–4]. CRISPR screens can be performed by either direct knockout of genes using CRISPR/Cas9 [1, 2], or perturbing gene expressions using CRISPR and a dead-Cas9 (dCas9) fused with activation or repression effectors [5, 6].

While CRISPR screening is a powerful technique, it creates computational challenges that include: (1) how to evaluate the data quality; (2) how to identify gene or pathway hits from the screens and assess their statistical significance; and (3) how to visualize and explore the screening results efficiently. Until now, a comprehensive quality control (QC), data analysis, and visualization method for CRISPR screen was not available. Several algorithms are developed for screening analysis on microarray or high-throughput sequencing data, such as RIGER [7], RSA [8], HitSelect [9], as well as the MAGeCK algorithm we previously developed [10]. These algorithms are designed based on a comparison of two conditions, although many screens are conducted simultaneously across several time points, under many treatment conditions or over many cell lines. In addition, these algorithms do not consider the knockout efficiency of single guide RNAs (sgRNA) on target genes. The knockout efficiency is the ability of a sgRNA to induce cutting events that lead to the knockout of the targeted gene. It is influenced by sgRNA sequence content [11], chromatin accessibility and exon position of the targeting gene [12], and so on.

In this study, we present MAGeCK-VISPR to overcome the computational challenges of CRISPR screens. MAGeCK-VISPR (1) defines a set of QC measurements and (2) extends the MAGeCK algorithm by a maximum likelihood estimation method (MAGeCK-MLE) to call essential genes under multiple conditions while considering sgRNA knockout efficiency. Further, MAGeCK-VISPR (3) provides a web-based visualization framework (VISPR) for interactive exploration of CRISPR screen quality control and analysis results. MAGeCK-VISPR employs a

* Correspondence: myles_brown@dfci.harvard.edu; xsliu@jimmy.harvard.edu
†Equal contributors
[2]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA
Full list of author information is available at the end of the article

Snakemake [13] workflow to combine MAGeCK and VISPR in a scalable and reproducible way (Fig. 1).

Results and discussion

Quality control measurements for CRISPR screening experiments

Apart from the determination of essential genes with MAGeCK, a central purpose of MAGeCK-VISPR is to collect quality control (QC) measurements at various levels (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The proposed measurements (Table 1) can be divided into four categories: sequence level, read count level, sample level, and gene level (Fig. 2).

Sequence level QC measurements aim to detect problems with the sequencing, similar as in other next-generation sequencing (NGS) experiments. Two measurements are reported: sample GC content distribution (Fig. 2a) and the base quality distribution of sequencing reads (Fig. 2b, c). Ideally, sequencing reads should have reasonable base qualities (median value >25), and samples from the same experiment should have similar GC content distributions.

The second level of QC measurements is based on the sgRNA read counts collected from MAGeCK. Raw sequencing reads are first mapped to sgRNA sequences in the library with no mismatches tolerated. After that, the number of sequencing reads, mapped reads (and thereof the percentage of mapped reads), sgRNAs with zero read count, and the Gini index of read count distribution are reported for each sample (Fig. 2d-f). The percentage of mapped reads is a good indicator of sample quality, and low mappability could be due to sequencing error, oligonucleotide synthesis error, or sample contamination. Good statistical power of downstream analysis relies on sufficient reads (preferably over 300 reads) for each sgRNA, with low number of zero-count sgRNAs in the plasmid library or early time points. Gini index, a common measure of income inequality in economics, can measure the evenness of sgRNA read counts [14]. It is perfectly normal for later time points in positive selection experiments to have higher Gini index since a few surviving clones (a few sgRNA with extreme high counts) could dominate the final pool while most of the other cells die (more sgRNAs with zero-count). In contrast, high Gini index in plasmid library, in early time points, or in negative selection experiments may indicate CRISPR oligonucleotide synthesis unevenness, low viral transfection efficiency, and over selection, respectively.

Sample level QC (Fig. 2g-j) checks the consistency between samples. MAGeCK-VISPR reports the distributions of normalized read counts by box plots and cumulative distribution functions. It also calculates pairwise Pearson correlations of sample log read counts, and draws the samples on the first three components of a Principle Component Analysis (PCA). Biological replicates or samples with similar conditions should have similar read

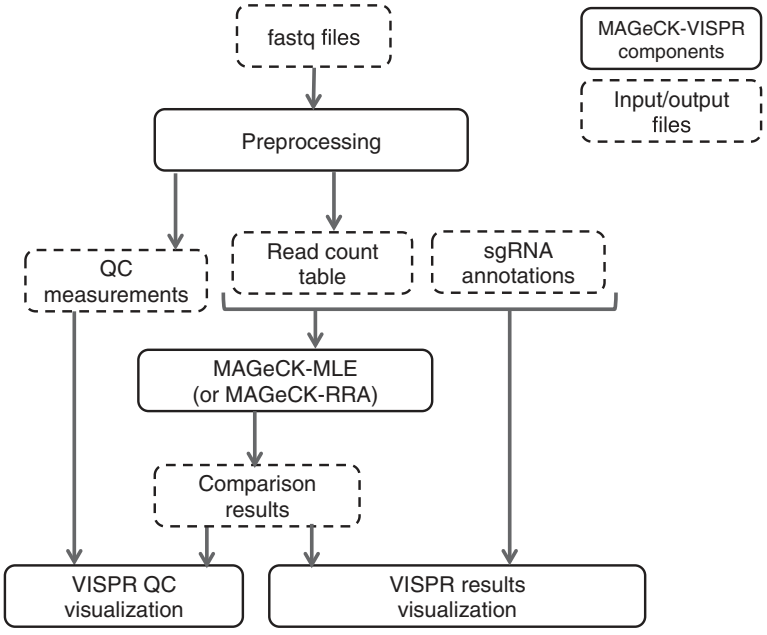

Fig. 1 An overview of the MAGeCK-VISPR workflow. Given FASTQ files and an sgRNA design, the workflow employs several preprocessing steps, including using FastQC and MAGeCK to collect quality control metrics and calculate per-sgRNA read counts. Then, either MAGeCK-RRA or MAGeCK-MLE is used to determine essential genes under user-defined conditions. Finally, results are composed for visualization and interactive exploration in VISPR

Table 1 Quality control (QC) measures from MAGeCK-VISPR

QC term	Description	Expected
GC content	GC content distribution of the sequencing reads	Similar distribution for all samples from same library
Base quality	Base quality distribution of the sequencing reads	Single-peak distribution with median base quality at least 25
Sequencing reads	Total number of sequencing reads	Varies depending on sequencing platform
Mapped reads	Total number of reads mapped to the sgRNA library	300 * (number of sgRNAs)
% Mapped reads	Percentage of mapped reads to the total number of sequencing reads	At least 65 %
Zero sgRNAs	Number of sgRNAs with zero read counts	At most 1 % of total sgRNAs
Gini index	Gini index of log-scaled read count distributions	At most 0.1 for plasmid or initial state samples, and at most 0.2 for negative selection samples
Sample correlation	Pearson correlation coefficient between samples	At least 0.8 for replicates
Correlation clustering or PCA clustering	Hierarchical clustering of samples or first three PCA components	Samples with similar conditions should cluster together
Ribosomal gene selection	Negative selection enrichment statistics of ribosomal genes	Significant *P* values (<0.001) for ribosomal subunit (GO:0044391) in negative selection experiments

count distributions and higher correlations, and appear closer to each other in the PCA plot. PCA plots can also identify potential batch effects if the screens are conducted under different batches.

Finally, gene level QC determines the extent of negative selection in the screens. Since knocking out ribosomal genes lead to a strong negative selection phenotype [1, 2], the significance of negative selection on ribosomal genes can be evaluated in MAGeCK-VISPR by Gene Ontology (GO) enrichment analysis using GOrilla [15]. A working negative selection experiment should have a significant *P* value (<0.001), although many good experiments could have much smaller *P* values (<1e-10, see Section A of Additional file 1).

Calling essential genes under multiple conditions with MAGeCK-MLE

MAGeCK-VISPR includes a new algorithm, 'MAGeCK-MLE', to estimate the essentiality of genes in various screening conditions using a maximum likelihood estimation (MLE) approach. Compared with the original MAGeCK algorithm using Robust Rank Aggregation ('MAGeCK-RRA') that can only compare samples between two conditions, MAGeCK-MLE is able to model complex experimental designs. Furthermore, MAGeCK-MLE explicitly models the sgRNA knockout efficiency, which may vary depending on different sequence contents and chromatin structures [11, 12]. In MAGeCK-MLE, the read count of a sgRNA i targeting gene g in sample j is modeled as a Negative Binomial (NB) random variable. The mean of the NB distribution (μ_{ij}) is dependent on three factors: the sequencing depth of sample j (s_j), the knockout efficiency of sgRNA i, and a linear combination of the effects in different conditions (that is, different drug treatments) on gene g. If sgRNA i knocks out target gene g efficiently, then μ_{ij} is modeled as:

$$\mu_{ij} = s_j \, exp\left(\beta_{i0} + \sum_r d_{jr}\beta_{gr}\right)$$

The effects of r different conditions are represented as the score 'β_{gr}', a measurement of gene selections similar to the term of 'log fold change' in differential expression analysis. The presence or absence of each condition on each sample is encoded into binary elements of the *design matrix* d_{jr}, and can be obtained from experiment designs. 'β' scores reflect the extent of selection in each condition: β_{gr} >0 (or <0) means g is positively (or negatively) selected in condition r. μ_{ij} is also dependent on β_{i0}, the initial sgRNA abundance which is usually measured in plasmid or the day 0 of the experiment.

The values of β, together with the information whether an sgRNA is efficient, can be estimated by maximizing the joint log-likelihood of observing all sgRNA read counts of g on all different samples, and are optimized using an Expectation-Maximization (EM) algorithm. In the EM algorithm, MAGeCK-MLE iteratively determines the knockout efficiency of each sgRNA based on the current estimation of 'β' scores (the E step), and uses the updated knockout efficiency information to re-calculate 'β' scores (the M step). By examining the patterns of read counts of each sgRNA across all samples, the EM algorithm minimizes the effect of inefficient sgRNAs. A detailed description of the method is presented in the Methods section.

We tested MAGeCK algorithms on four public datasets. The first two datasets (the 'ESC' and 'leukemia' dataset) correspond to negative selection experiments on mouse embryonic stem cells (ESCs) and two human

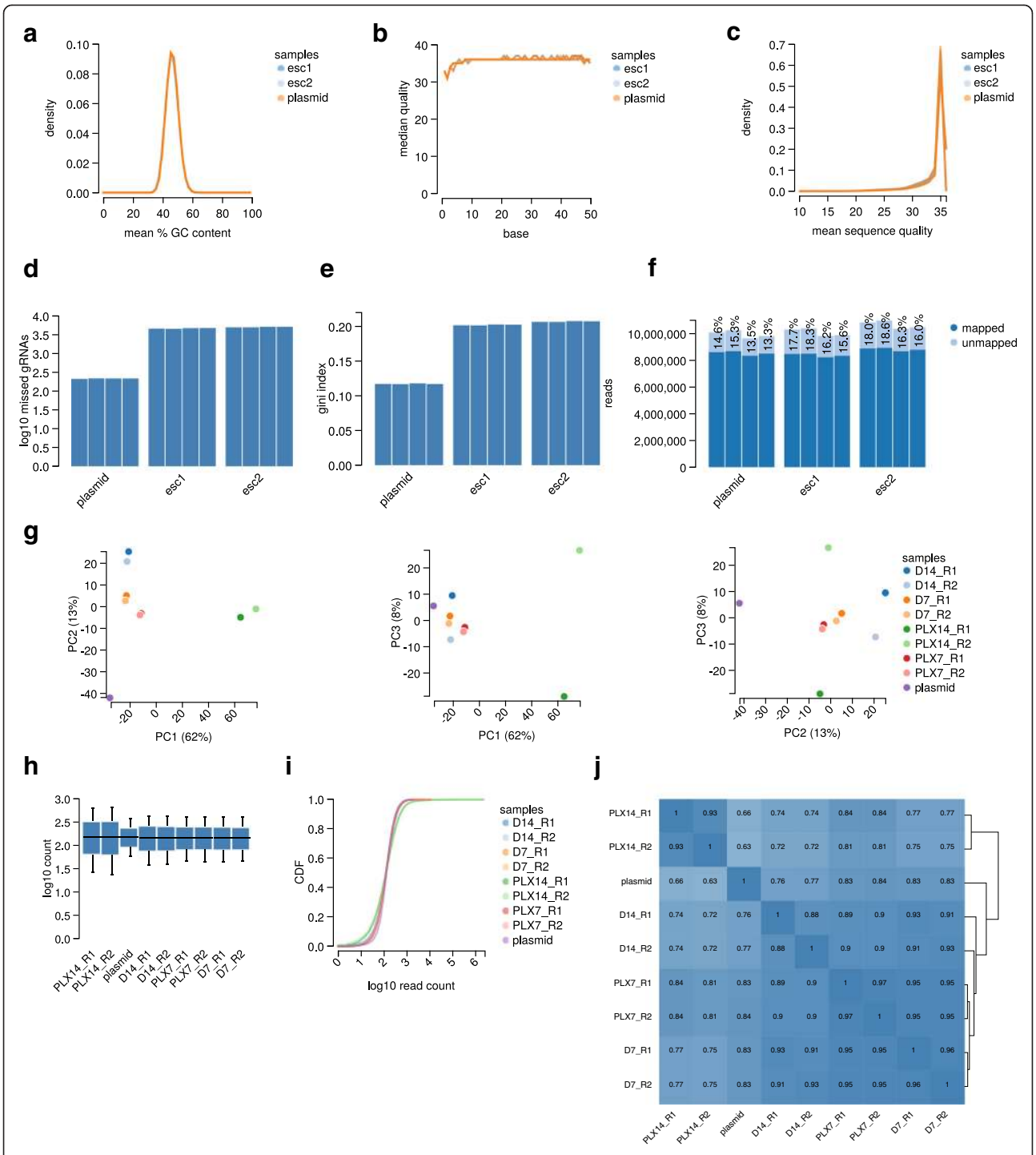

Fig. 2 The quality control (QC) view of VISPR, the visualization framework of MAGeCK-VISPR. The measurements include the distribution of GC content (**a**), median base quality (**b**), the distribution of mean sequence quality (**c**), the number of zero-count sgRNAs (**d**), Gini-index (**e**), total number of reads and the percentage of mapped reads (**f**), Principle Component Analysis (PCA) plot (**g**), normalized read count distribution (**h**, **i**), and pairwise sample correlations (**j**). Shown results are from ESC (**a**-**f**) and melanoma dataset (**g**-**j**)

leukemia cell lines (KBM7 and HL-60), respectively (Fig. 3a and b) [1, 4]. In both datasets, cells were grown with their natural growing condition and negative selections occurred in cells after CRISPR/Cas9 is activated. The other two datasets ('melanoma' knockout and activation dataset) are different CRISPR screens on the human melanoma cell line A375 that harbors a BRAF V600E mutation (Figs. 4 and 5). The cells were treated with BRAF

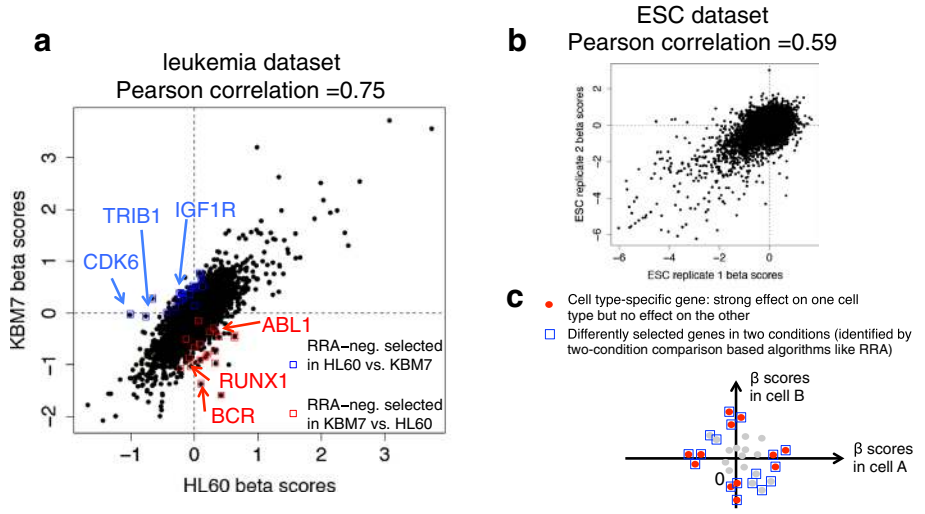

Fig. 3 The gene essentiality scores (*β* scores) reported from MAGeCK-MLE on two conditions. **a**, **b** the *β* scores of two leukemia cell lines in the leukemia dataset (**a**), and two biological replicates of mouse ESC cells in the ESC dataset (**b**). In (**a**), some well-known driver genes and cell type-specific genes are also labeled. These genes may play distinct roles in two different leukemia subtypes (HL60: acute myeloid leukemia; KBM7: chronic myeloid leukemia), including CDK6 and TRIB1 for HL60, and RUNX1 in KBM7. CDK6 is required in AML growth [17] and TRIB1 over-expression is observed in AML patients compared with CML patients [18]. On the other hand, the frequent RUNX1 loss-of-function mutations are observed in CML to AML transformations [19]. **c** An illustration of differentially selected genes identified by two-condition comparison algorithms (like RRA, blue rectangles). MAGeCK-MLE can further distinguish cell type-specific genes (red dots) from other genes. Cell type-specific genes are genes having no essentiality in one condition but strong essentiality in the other, and are usually more biologically interesting

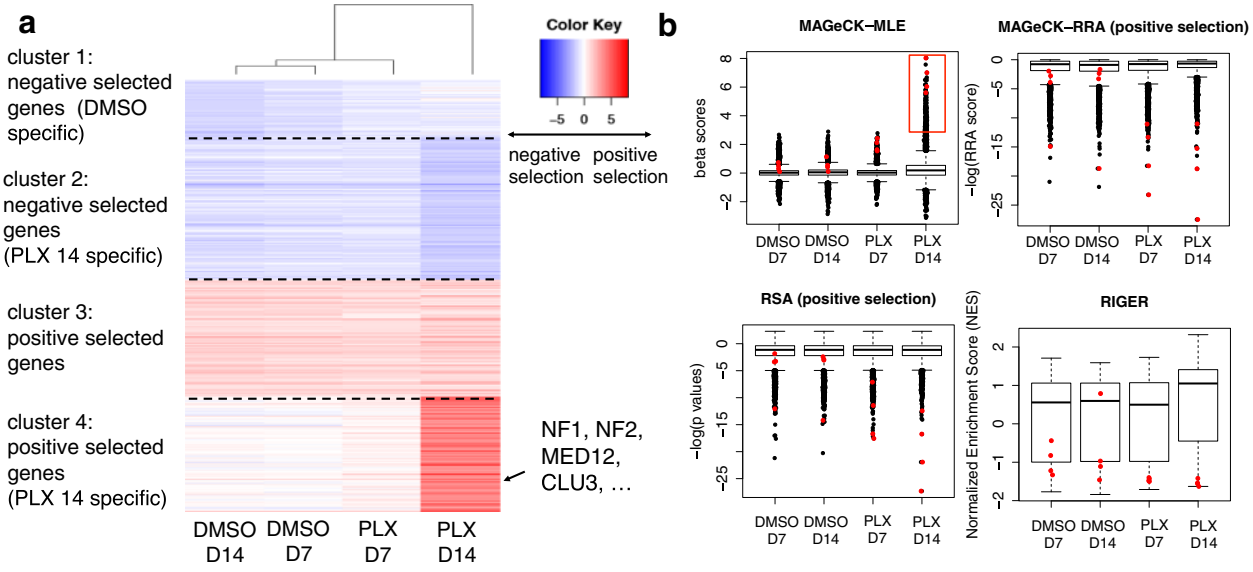

Fig. 4 The *β* scores of MAGeCK-MLE on the melanoma knockout dataset. **a** A k-means clustering view of *β* scores of all conditions from top selected genes (k = 4). Only genes with the highest or lowest 1 % *β* scores in DMSO or PLX 14-day treatment conditions are shown. **b** The distribution of scores across four conditions using different algorithms. The red rectangle in MAGeCK-MLE indicates genes in cluster 4 in Fig. 4a, or genes that are strongly positively selected in PLX 14-day condition. Some validated genes in the original study are marked as red dots, including NF1, NF2, MED12, and CUL3

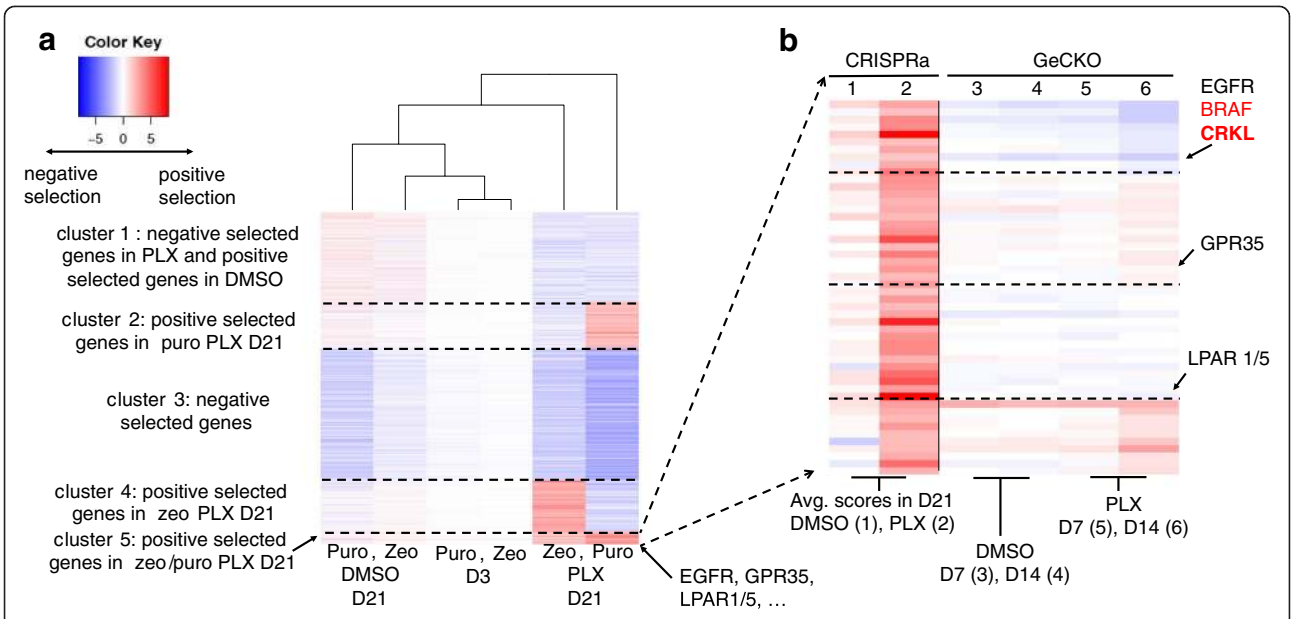

Fig. 5 The β scores of MAGeCK-MLE on the melanoma activation dataset. **a** A k-means clustering view of β scores of all conditions from top selected genes (k = 5). Only genes with the highest or lowest 1 % β scores in DMSO or PLX 21-day treatment conditions are shown. **b** the average β scores of genes in cluster 5 of A (consistently positive selected genes in both zeocin and puromycin conditions), as well as the β scores of these genes in melanoma knockout dataset. Similar to (**a**), the k-means clustering algorithm is applied to the selected genes (k = 4)

inhibitor vemurafenib (PLX) or dimethyl sulfoxide (DMSO) control, and screened either with GeCKO [2] or with CRISPR/dCas9 Synergistic Activation Mediator (SAM) libraries [5]. These two datasets include multiple experimental conditions that are difficult to compare directly using the original MAGeCK-RRA algorithm. In the melanoma knockout dataset, cells were under 7-day or 14-day selection [2]. In the melanoma activation dataset, two different drugs (puromycin and zeocin) were used to select cells with lentiviral infection, and both DMSO and PLX treatments were profiled under 3-day or 21-day selection [5].

In two-condition comparisons, MAGeCK-MLE gives similar results with existing methods such as MAGeCK-RRA, RSA, and RIGER. All the algorithms identified genes that are commonly essential to different cell types [16], as well as known positively selected genes in PLX treated conditions in two melanoma datasets (Fig. 3; also see Section A and B of Additional file 1). In the leukemia dataset, two-condition comparison algorithms (like MAGeCK-RRA) identified genes that are differentially selected in two cell lines by a direct comparison of HL60 and KBM7 (Fig. 3a) [10]. However, not all of these genes are equally biologically interesting, as MAGeCK-MLE further distinguished them into two groups: genes having little effect in one (β scores close to zero) but strong selection effect in the other cell line (large absolute β scores), and genes having weak and opposite effects in two cell lines (Fig. 3c). The first group of genes are often more biologically interesting as they are cell type-specific genes. This includes some well-known driver genes (like BCR in KBM7) as well as genes that may be functional in only one cell type: CDK6 and TRIB1 in HL60 [17, 18], and RUNX1 in KBM7 [19].

One of the advantages of MAGeCK-MLE over other methods is that it enables accurate comparisons of gene essentialities across multiple conditions and experiments in one run (Fig. 4 and Section C of Additional file 1). In the melanoma knockout dataset, a k-means clustering of the β scores of top selected genes demonstrated that these genes have various essentialities across conditions (Fig. 4a). Some of the genes are universally positively or negatively selected in all conditions (cluster 3), while others have different essentiality across different conditions (clusters 1, 2, and 4). Genes in cluster 4 are particularly interesting as they show strong positive selection in 14-day PLX treated condition. Indeed, genes whose knockout leads to strong positive selection in PLX-treated cells are in cluster 4, including NF1, NF2, MED12, CUL3 [2]. In contrast, the k-means clusters of measurements from other algorithms did not reveal the strong effect of genes in cluster 4 (Section C of Additional file 1). This is because their score distributions are similar across different conditions (Fig. 4b), and do not reflect the fact that the one condition (PLX 14-day treatment) induces much stronger positive selection than other conditions [2]. This is partly because MAGeCK-RRA, RIGER, and RSA all use a rank-based method to compare sgRNA between two conditions, which may lose quantitative information.

Another example of using MAGeCK-MLE on multiple conditions is demonstrated in the melanoma activation

dataset, where cells underwent different selection methods (using puromycin or zeocin), drug treatments (DMSO or PLX), and durations (3-day or 21-day treatment) (Fig. 5). Similar to the melanoma knockout dataset, we performed k-means clustering of the top-selected gene β scores. Many positively selected genes are dependent on the selection method, which might not be biologically interesting. For example, genes in clusters 2 and 4 correspond to positively selected genes that are specific to puromycin or zeocin selection, respectively. A small set of genes (cluster 5) are consistently selected in both zeocin and puromycin, including genes that are validated in the original study, for example, EGFR, GPR35, LPAR1/5 [5]. We further examined the genes in cluster 5 (Fig. 5b), and focused on genes positively selected in the CRISPR activation experiment but strongly negatively selected in the knockout experiment. These genes include EGFR and BRAF, two known kinases that drive melanoma progression and PLX resistance [20, 21], and CRKL, a protein kinase that activates RAS and JUN pathway. CRKL amplification is reported to lead to drug resistance against EGFR inhibitors by activating EGFR downstream pathways [22], implying its potential role in PLX drug resistance.

Visualization of QC measurements and gene essentiality with VISPR

VISPR (VISualization of crisPR screens) is a web-based frontend for interactive visualization of CRISPR screen QC and comparison results. Interactive access is provided by an HTML5 based browser interface, while visualizations are realized with Vega [23], a declarative visualization grammar on top of Data-Driven Documents (D3) [24]. VISPR provides three types of views for interactive exploration of CRISPR screening: a quality control view, a result view, and an experiment comparison view. The quality control view shows the QC measurements described before (Fig. 2).

In the result view, screening results can be interactively explored. It contains a table showing the comparison results of each gene (Fig. 6a). The table can be sorted by different columns and filtered (from 'Search') via gene names or regular expressions. Further, the distribution of *P* values is displayed as cumulative distribution function (CDF) (Fig. 6b) and as a histogram (Fig. 6c). For each gene, the normalized sgRNA counts in all samples can be displayed in a parallel coordinate visualization (Fig. 6d). If available, knockout efficiency predictions [11] and gene coordinates of each sgRNA are displayed as separate axes. Axes can be reordered or toggled on or off, and sgRNAs can be highlighted by selecting ranges on each axis. Genes selected in the table are highlighted in the CDF, allowing to assess their occurrence within the *P* value distribution of all genes.

VISPR provides various ways to further explore the analysis results. Individual genes can be viewed in Ensembl [25] and IGV [26]. Selected genes can be visualized in terms of their interaction network and function via GeneMANIA [27]. Functional analysis can be performed with GOrilla [15], an online Gene Ontology (GO) enrichment analysis tool. GOrilla takes a ranked list of genes (here based on the *P* values reported by MAGeCK) to perform a threshold-free enrichment analysis. The resulting GO term enrichments can be further used for gene-level quality control.

The comparison view of VISPR can compare different experiments by visualizing the common and exclusive significant genes via Euler diagrams (Fig. 6e). Clicking on segments of the Euler diagram opens the result views of the corresponding experiments. For example, clicking on the intersection between two experiments will open 'restricted' result views for each experiment, where only the common significant genes are displayed. These views provide the same features as the unconstrained result views described above. However, in this case, GO enrichment analysis with GOrilla is performed with the shown genes (that is, the genes from the intersection) as foreground and the other genes of the experiment as background.

The visualizations displayed in VISPR can be downloaded as publication-ready SVG files. In addition, a command line interface is provided to store visualizations as Vega specifications. This format allows users to modify and style the output of VISPR programmatically.

Implementation of the MAGeCK-VISPR workflow with Snakemake

We implemented the MAGeCK-VISPR workflow with the workflow management system Snakemake [13], allowing an automatic execution of some or all of the MAGeCK-VISPR functions: quality control, essential gene analysis, and visualization. Choosing a workflow management system like Snakemake has several advantages. First, the workflow steps can be automatically parallelized and executed on workstations, servers, and compute clusters without the modification of the workflow. Second, Snakemake tracks metadata (like creation date, input, and log files) for all generated result and intermediate files. This way, used data, methods, and parameters are documented comprehensively for each analysis (also called data provenance), an important requirement of reproducible science. MAGeCK-VISPR provides a command line interface to initialize the workflow in a given work directory. This installs the workflow definition as a so-called *Snakefile*, along with a configuration file and documentation. The configuration file is used to define locations of raw data and additional parameters for MAGeCK-VISPR. Once configured, the Snakefile can be executed with Snakemake. Since the Snakefile is installed into the given work directory, it can be easily modified or extended by the user.

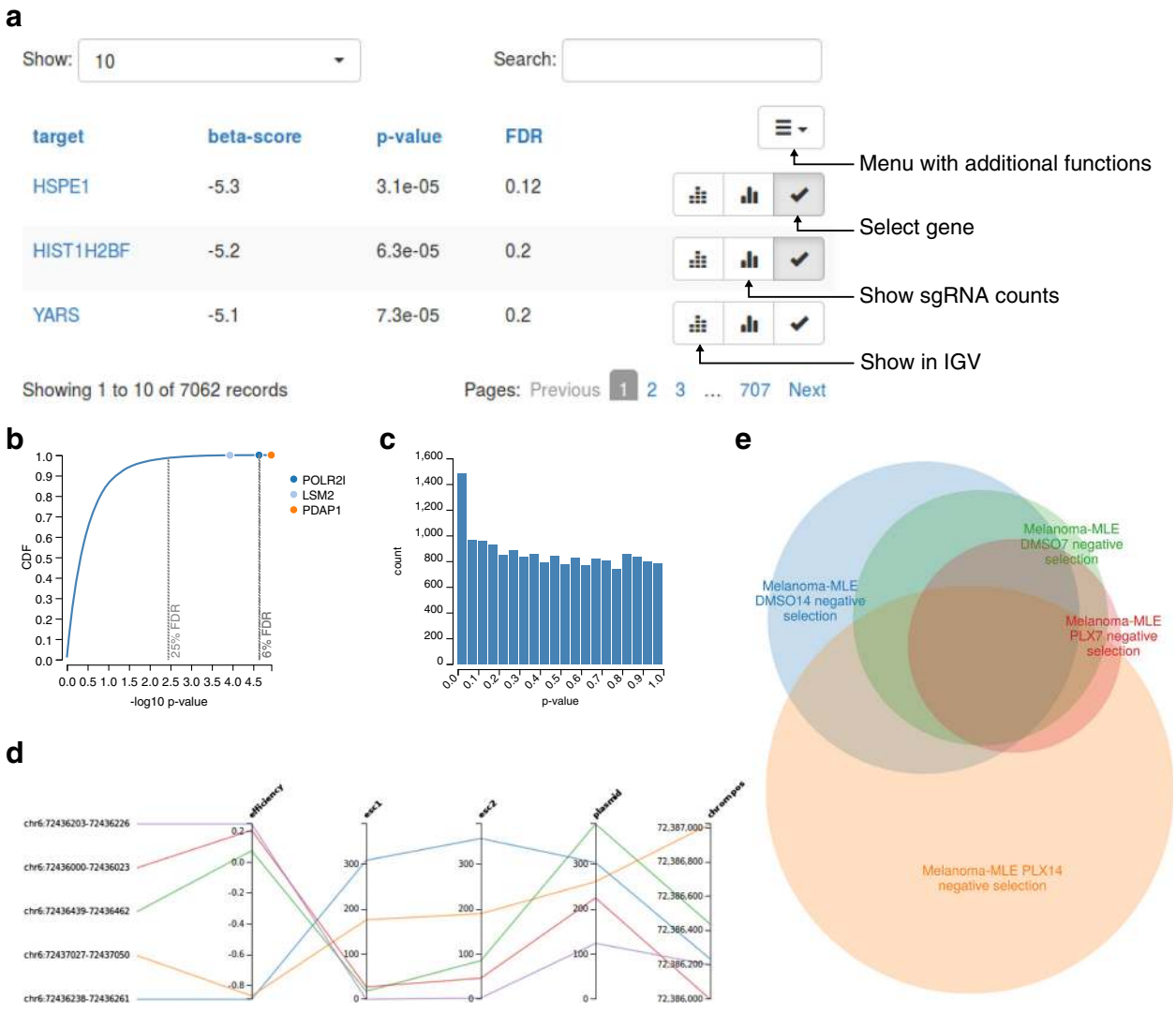

Fig. 6 The result and comparison views of VISPR. The result view includes a gene comparison table (**a**), the distribution of *P* values as CDF (**b**) and histogram (**c**), and the normalized sgRNA counts in all samples of selected genes along with chromosome positions and predicted efficiency (**d**). The comparison view of VISPR (**e**) displays the overlap between significant genes under different selectable conditions and experiments as an Euler diagram. Shown results are from ESC (**a**, **d**) and melanoma dataset (**b**, **c**, **e**)

We provide all components of the workflow as Conda packages [28], such that MAGeCK-VISPR can be installed with a single command. Optionally, the Conda package manager can create isolated environments for the workflow to, for example, freeze or compare different software versions or publish snapshots of a MAGeCK-VISPR workflow instance along with all data and used software. This further increases the reproducibility of the generated results.

Conclusion

The recently developed CRISPR screening is a powerful technology in functional studies with different foci, including tumor progression and metastasis [29], drug resistance [3], immune response [30], and stem cell differentiation [4]. To our knowledge, MAGeCK-VISPR is the first comprehensive pipeline developed for quality control, analysis, and visualization of CRISPR screens, and highlights new features compared with existing screening analysis algorithms. For example, a typical CRISPR screening experiment usually includes complex designs that are difficult to analyze using existing algorithms, as they are all designed for two-condition comparisons. To address this challenge, MAGeCK-VISPR uses a maximum likelihood approach to estimate the effect of different conditions using a generalized linear model (GLM). It also incorporates sgRNA knockout efficiency information by using a probabilistic mixture model. We demonstrated that MAGeCK-MLE provides additional insights into cell type-specific essential genes and is able to compare

gene essentiality scores across conditions or even experiments. Also, MAGeCK-VISPR is able to handle screens of different types including CRISPR knockout and CRISPR activation screens, and can be potentially applied to high-throughput sequencing datasets of traditional RNA interference (RNAi) screens.

The MAGeCK-MLE approach is able to estimate sgRNA knockout efficiencies from CRISPR screens besides gene essentiality. We previously reported that sequence-specific features learned from CRISPR screening data helped the design of efficient sgRNAs [11]. With more CRISPR screen data becoming available, the algorithm will help us identify sgRNAs with the best behavior and learn patterns of 'good' sgRNAs. The information will further guide the design of optimized sgRNAs for CRISPR screens and individual gene knockouts.

One potential limitation of MAGeCK-MLE is that its EM algorithm uses an iterative process involving matrix operations, making it slower than our previous MAGeCK RRA method and other competing algorithms. Future approaches to speed up MAGeCK-MLE include improving parametric tests for *P* value estimation (instead of using permutation) and implementing the algorithm in Cython instead of Python. Another potential limitation of MAGeCK-VISPR on the quality control assessment is that the current QC thresholds for 'successful' experiments are determined heuristically due to limited number of publicly available CRISPR screening datasets. We and other researchers have previously reported that bigger collections of ChIP-seq datasets provide better criteria on ChIP-seq quality control [31, 32]. As more public CRISPR screening datasets become available, the QC metrics (and other parts of MAGeCK-VISPR) can be further refined.

As CRISPR screens become more popular, complications in the data such as batch effects will be unavoidable which need proper correction for meaningful downstream analysis. Existing batch removal algorithms, including ComBat [33] and RUVseq [34], have been widely used to remove batch effects in gene expression analysis. In the future, these algorithms can be integrated into MAGeCK-VISPR pipeline. After that, MAGeCK-VISPR will be able to identify cancer- and disease-specific essential genes by a direct comparison between different datasets or experiments, providing potentially new therapeutic insights into the mechanisms of diseases and cancers.

Methods

MAGeCK-MLE: a maximum likelihood approach for essential gene detection

The Negative Binomial model for high-throughput CRISPR screening read counts

After read mapping, the sequencing results of CRISPR screening are presented as a read count table, where rows correspond to sgRNAs and columns correspond to samples. Read counts generated from high-throughput sequencing data have higher variances when a high number of read counts are observed (also called 'over-dispersion'). This is usually modeled using Negative Binomial (NB) distribution, such as in the statistical models used in many RNA-seq differential expression analysis algorithms: edgeR, DESeq/DESeq2, and so on [35–37]. MAGeCK-MLE uses a similar model; briefly, the read count of sgRNA i in sample j, or x_{ij}, is modeled as:

$$x_{ij} \sim NB\left(\mu_{ij}, \alpha_i\right)$$

Where μ_{ij} and α_i are the mean and over-dispersion factor of the NB distribution, respectively. The mean value μ_{ij} is further modeled as:

$$\mu_{ij} = s_j q_{ij} \quad (1)$$

Where s_j is the size factor of sample j for adjusting sequencing depths of the samples, and q_{ij} is a variable modeling the behavior of sgRNA i in sample j that will be discussed in later sections. s_j is calculated by the 'median ratio method' in MAGeCK and DESeq2 [10, 37]:

$$s_j = median_i\left\{\frac{x_{ij}}{\hat{x}_{i\cdot}}\right\}$$

Here, $\hat{x}_i$ is the geometric mean of the read counts of sgRNA i across all J samples: $\hat{x}_i = \left(\prod_{k=1}^{J} x_{ik}\right)^{1/J}$. s_j can also be calculated based on a set of predefined 'control' sgRNAs instead of all sgRNAs. This is particularly useful when a majority of the genes in the library are supposed to be essential; in such cases it is not suitable to calculate s_j based on all sgRNAs. Both methods are implemented in MAGeCK-VISPR and users can specify which method to use.

The over-dispersion factor α_i is calculated based on the regression residual and will be discussed in more details in the last Methods section.

Modeling sgRNA knockout efficiency and complex experimental settings

Different studies demonstrated that sgRNAs have various DNA cutting efficiencies [11, 38], but such information is not considered in most essential gene calling algorithms (including MAGeCK). In MAGeCK-MLE, we use a binary variable π_i to model whether sgRNA i is efficient or not: $\pi_i = 1$ corresponds to an efficient sgRNA i and vice versa. Since π_i is unknown, the probability of

observing a read count x from x_{ij} is a mixture of two distributions:

$$P(x_{ij}=x)=p(x_{ij}=x|\pi_i=1)p(\pi_i=1)+p(x_{ij}=x|\pi_i=0)p(\pi_i=0)$$

In CRISPR screening experiments, it is common to have cells treated with different conditions. For example in melanoma activation dataset [5], cell lines underwent different sgRNA expression selection methods (cells are first selected using puromycin or zeocin), duration of treatment (3-day or 21-day treatment) and drug treatments (DMSO or PLX). For an efficient sgRNA i ($\pi_i = 1$), MAGeCK-MLE uses a generalized linear model (GLM) to model the effect of q_{ij} as a linear combination of effects from different sources:

$$\begin{aligned} P(x_{ij}=x|\pi_i=1) &\sim NB\left(x; s_j q_{ij}, \alpha_i\right) \\ log\left(q_{ij}\right) &= \beta_{i0}+\sum_r d_{jr}\beta_{gr} \end{aligned} \quad (2)$$

Here, β_{i0} is the baseline abundance of sgRNA i, corresponding to its abundance in an initial state (in plasmid or day 0). d_{jr} is an element of a *design matrix* given by the user (explained later), and β_{gr} is the (unknown) coefficient that we would like to estimate.

If sgRNA i is inefficient ($\pi_i = 0$), then its read counts in all samples are not determined by any experimental conditions except the baseline abundance:

$$\begin{aligned} P(x_{ij}=x|\pi_i=0) &\sim NB\left(x; s_j q_{ij}, \alpha_i\right) \\ log q_{ij} &= \beta_{i0} \end{aligned} \quad (3)$$

The design matrix

Design matrices have been used in many gene expression analysis algorithms for modeling complex experimental designs, including LIMMA [39], VOOM [40], DESeq2 [37], and so on. The design matrix D models the combination of effects of different conditions. For J samples that are affected by R conditions, D is a $J * R$ binary matrix with element $d_{jr} = 1$ if sample j is affected by condition R, and 0 otherwise. An example of the design matrix is presented in Additional file 1.

Based on the design matrix, the equations in (2) and (3) can be written in a matrix form. For a gene g with N sgRNAs in J samples, let $\vec{q}_g$ be the vector of q values of all sgRNAs in all samples in gene g:

$$\vec{q}_g = \left(q_{11}, q_{21}, \ldots, q_{N1}, \ldots, q_{1J}, q_{2J}, \ldots, q_{NJ}\right)^T$$

It can be written as:

$$\log\left(\vec{q}_g\right) = D'^{*}\vec{\beta}_g$$

Where $\vec{\beta}_g$ is a $N + r$ vector of β values in Equations (2) and (3). The first N elements of $\vec{\beta}_g$ are the baseline abundances of N sgRNAs, and the following R elements of $\vec{\beta}_g$ are the coefficients corresponding to R columns in the design matrix:

$$\vec{\beta}_g = \left(\beta_{00}, \beta_{10}, \ldots, \beta_{N0}, \beta_1, \ldots, \beta_r\right)^T.$$

The binary *extended design matrix D'* is used to set up the linear relationship between $\vec{\beta}_g$ and $\vec{q}_g$, and can be derived directly from the design matrix. See Additional file 1 for the definition and an example of D'.

The EM approach

MAGeCK-MLE uses a maximum likelihood estimation (MLE) approach to find the values of $\vec{\beta}^*{}_g$ The objective function of MAGeCK-MLE is:

$$\left(\vec{\beta}_g^*, \pi_i^*\right) = arg\max_{\beta_g,\ \pi_i}\left(\sum_{\substack{i\in g,\\ j=1,\ldots J}} \log p(x_{ij})\right)$$

Similar to DESeq2 [37], MAGeCK-MLE also adds a prior $p\left(\vec{\beta}_g\right)$ that follows a normal distribution centered on zero in the objective function. Adding this prior makes sure $\vec{\beta}^*{}_g$ does not become arbitrarily large, when the sgRNA knockout efficiency is low and the differences of read counts between samples are high.

The objective function can be maximized using expectation maximization (EM). At the beginning, we have an initial guess of $p(\pi_i = 1)$. Subsequently, we iteratively update the values of $p(\pi_i = 1)$ and $\vec{\beta}$ in the E step and the M step, respectively.

The initial guess of sgRNA knockout efficiency

We demonstrated that the SSC (Spacer Scoring of CRISPR) algorithm accurately predicts sgRNA knockout efficiency from genomic sequence content [11]. For each sgRNA, SSC generates an efficiency score in the range (−2,2). We scale the score linearly to the range (0,1) as an initial guess of $p(\pi_i = 1)$. If no initial estimates are given, MAGeCK-MLE starts with $p(\pi_i = 1) = 1$ for all sgRNAs.

The expectation step

In the E step, we re-estimate the posterior probability $p(\pi_i = 1)$ and the current estimation of $\vec{\beta}_g$:

$$p\left(\pi_i = 1|x_{ij}, \vec{\beta}_g\right) = \frac{\prod_j p\left(x_{ij}|\pi_i = 1, \vec{\beta}_g\right) p(\pi_i = 1|\vec{\beta}_g)}{\prod_j p(x_{ij}|\pi_i = 1, \vec{\beta}_g) p\left(\pi_i = 1|\vec{\beta}_g\right) + \prod_j p\left(x_{ij}|\pi_i = 0, \vec{\beta}_g\right) p(\pi_i = 0|\vec{\beta}_g)}$$

The maximization step

In the M step, we maximize the values of $\vec{\beta}_g$ based on the values of $p(\pi_i = 1)$. To derive the formula for updating $\vec{\beta}_g$, we write the probability of observing a read count x of x_{ij} as:

$$P(x_{ij} = x) = P(x_{ij} = x|\pi_i = 1)^{I(\pi_i=1)} * P(x_{ij} = x|\pi_i = 0)^{I(\pi_i=0)}$$

where $I(.)$ is an indicator function. Taking the logarithm on both sides of the equation, we get

$$\log P(x_{ij} = x) = I(\pi_i = 1)\log P(x_{ij} = x|\pi_i = 1) + I(\pi_i = 0)\log P(x_{ij} = x|\pi_i = 0)$$

In the EM algorithm, it can be approximated by replacing the indicator function $I(\pi_i = 1)$ and $I(\pi_i = 0)$ with the posterior probability of $P(\pi_i = 1)$ and $P(\pi_i = 0)$, respectively [41], using the results from the E step. Therefore, the log likelihood function from the mixture model can be written as:

$$\sum_{i,j} \log P(x_{ij} = x) = \sum_{i,j} P\left(\pi_i = 1|x_{ij}, \vec{\beta}_g\right) logP(x_{ij} = x|\ \pi_i = 1) + P\left(\pi_i = 0|x_{ij}, \vec{\beta}_g\right) logP(x_{ij} = x|\pi_i = 0)$$

Since NB distribution belongs to exponential family distributions, a fast algorithm exists for the maximum likelihood estimation of generalized linear models [42]. Taking the prior of $\vec{\beta}_g$ into consideration, the objective function can be maximized using iteratively reweighted ridge regression, or weighted updates, the same the algorithm used in DESeq2 [37]. The update rule for calculating $\vec{\beta}^t_g$ at step t of the iteration can be written as:

$$\vec{\beta}^t_g = (\mathrm{D}'^{\mathrm{T}} W \mathrm{D}' + \lambda I)^{-1} \mathrm{D}'^{\mathrm{T}} W \vec{z}^t$$

Here, W is the diagonal matrix with its values given by $w_{ii} = e^t_i/(1/\mu_i + \alpha_i)$, where e^t_i is the current estimate of the efficiency of sgRNA $i : e^t_i = P(\pi_i = 1 \left| x_{ij}, \beta^{t-1\rightarrow}_g \right.)$, λ is the regularization parameter in the ridge regression, and μ_i is the current estimate of the mean of the NB variable:

$$\begin{aligned} \mu^{t\rightarrow} &= s_j exp(\, h^{t\rightarrow}) \\ \mathrm{h}^{\mathrm{t-1}\rightarrow} &= \mathrm{D}'\, \beta^{t-1\rightarrow}_g \end{aligned} \qquad (4)$$

$\vec{\mathrm{z}}^t$ is the residue vector of the current estimate, with its ith element:

$$z^t_i = \mathrm{h}^{t-1}_{\mathrm{i}} + e^t_i(\mathrm{x_i} - \mu^t_i)/\mu^t_i$$

Here, x_i is the read count of sgRNA i.

Convergence

The EM approach iterates the E step and the M step until it converges or reaches a predefined maximum number of iteration.

Statistical significance

The statistical significance of $\vec{\beta}_g$ is calculated in both permutation and Wald test. In permutation test, MAGeCK-MLE shuffles all sgRNAs in a gene to generate empirical null distribution of $\vec{\beta}_g$. The number of shufflings is a parameter specified by the user, and the default value is set to be 2*(total number of genes). In the Wald test, MAGeCK-MLE compares the value of $\vec{\beta}_g/SE\left(\vec{\beta}_g\right)$ to the standard Normal distribution, where $SE\left(\vec{\beta}_g\right)$ is the standard error of $\vec{\beta}_g$:

$$SE\left(\vec{\beta}_g\right) = \sqrt{diag\left(Cov\left(\left(\vec{\beta}_g\right)\right)\right)}$$
$$Cov\left(\left(\vec{\beta}_g\right)\right) = \left(D'^T W D' + \lambda I\right)^{-1}\left(D'^T W D'\right)\left(D'^T W D' + \lambda I\right)^{-1}$$

Here, $diag\left(Cov\left(\vec{\beta}_g\right)\right)$ are the diagonal elements of the covariance matrix of $\vec{\beta}_g$.

Calculating the over-dispersion factor

The over-dispersion factor, α_i, is calculated based on the mean and variance estimation algorithm used in MAGeCK [10] and VOOM [40]. We first calculate the fitted values of $\vec{\beta}_g$, or $\hat{\beta}_g$, using the EM algorithm proposed before, with the over-dispersion factor set to a fixed value (for example, 0.01). Then the fitted means $\hat{\mu}_i$ are calculated using Equation (4), and the residual variances are calculated using the following equation:

$$\hat{\sigma}_i{}^2 = (x_i - \hat{\mu}_i)^2$$

MAGeCK-MLE then models the sample residual variance $\hat{\sigma}^2$ and fitted mean $\hat{\mu}$ using the same model as in MAGeCK [10]:

$$\hat{\sigma}^2 = \hat{\mu} + k\hat{\mu}^b$$

Where k and b are learned from the fitted means and residual variances of all sgRNA read counts. The values of α_i are then calculated based on the fitted values of sample residual variance $\hat{\sigma}_f{}^2$ from this model:

$$\alpha_i = \frac{\hat{\sigma}_f^2 - \hat{\mu}_i}{\hat{\mu}_i{}^2}$$

Availability

The MAGeCK-VISPR workflow is available open source at http://bitbucket.org/liulab/mageck-vispr under the MIT license.

Additional file

Additional file 1: Supplementary materials. (PDF 2045 kb)

Abbreviations

AML: acute myeloid leukemia; CDF: cumulative distribution function; CML: chronic myeloid leukemia; CRISPR: clustered regularly interspaced short palindromic repeats; D3: Data-Driven Documents; dCas9: dead Cas9; DMSO: dimethyl sulfoxide; EM: expectation-maximization; GeCKO: genome-scale CRISPR/Cas9 knockout; GLM: generalized linear model; GO: gene ontology; MAGeCK: Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout; MLE: maximum-likelihood estimation; NB: negative binomial; NGS: next-generation sequencing; PCA: principle component analysis; QC: quality control; RNAi: RNA interference; RRA: robust rank aggregation; SAM: Synergistic Activation Mediator; sgRNA: single-guide RNA; VISPR: VISualization of crisPR screens.

Competing interests

The authors declare no competing financial interests.

Authors' contributions

WL, JSL, and XSL designed the statistical model. WL and JK developed the algorithm, designed and performed the analysis. XH and CHC performed sgRNA efficiency prediction analysis. WL, JK, and XSL wrote the manuscript with help from all other authors. XSL and MB supervised the whole project. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank Michael I. Love, Clifford Meyer, Peng Jiang, Bo Li, and Graham McVicker for helpful discussions.

Funding

The project was supported by the NIH grant U01 CA180980 (to XSL), R01 HG008728 (to MB and XSL), Department of Defense Synergistic Idea Development Award PC140817 (to MB and XSL), R01 GM113242-01 (to JSL), NSF grant DMS-1120368 (to JSL), and the Claudia Adams Barr Award in Innovative Basic Cancer Research from the Dana-Farber Cancer Institute.

Author details

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA. [2]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [3]Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [4]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA. [5]Department of Statistics, Harvard University, Science Center 715, 1 Oxford Street, Cambridge, MA 02138, USA. [6]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02215, USA. [7]School of Life Science and Technology, Tongji University, Shanghai 200092, China.

Received: 31 July 2015 Accepted: 23 November 2015
Published online: 16 December 2015

References

1. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343:80–4.
2. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343:84–7.
3. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, et al. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature. 2014;509:487–91.
4. Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32:267–73.
5. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015;517:583–8.
6. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159:647–61.
7. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci U S A. 2008;105:20380–5.
8. König R, Chiang C-Y, Tu BP, Yan SF, DeJesus PD, Romero A, et al. A probability-based approach for the analysis of large-scale RNAi screens. Nat Methods. 2007;4:847–9.
9. Diaz AA, Qin H, Ramalho-Santos M, Song JS. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. Nucleic Acids Res. 2015;43:e16–6.
10. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15:554.
11. Xu H, Xiao T, Chen C-H, Li W, Meyer C, Wu Q, et al. Sequence determinants of improved CRISPR sgRNA design. Genome Res. 2015;25:1147–57.
12. Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol. 2014;32:670–6.
13. Köster J, Rahmann S. Snakemake - a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
14. Wittebolle L, Marzorati M, Clement L, Balloi A, Daffonchio D, Heylen K, et al. Initial community evenness favours functionality under selective stress. Nature. 2009;458:623–6.
15. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.

16. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10:733–3.
17. Placke T, Faber K, Nonami A, Putwain SL, Salih HR, Heidel FH, et al. Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. Blood. 2014;124:13–23.
18. Röthlisberger B, Heizmann M, Bargetzi MJ, Huber AR. TRIB1 overexpression in acute myeloid leukemia. Cancer Genet Cytogenet. 2007;176:58–60.
19. Zhao L-J, Wang Y-Y, Li G, Ma L-Y, Xiong S-M, Weng X-Q, et al. Functional features of RUNX1 mutants in acute transformation of chronic myeloid leukemia and their contribution to inducing murine full-blown leukemia. Blood. 2012;119:2873–82.
20. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002;417:949–54.
21. Prahallad A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature. 2012;483:100–3.
22. Cheung HW, Du J, Boehm JS, He F, Weir BA, Wang X, et al. Amplification of CRKL induces transformation and epidermal growth factor receptor inhibitor resistance in human non-small cell lung cancers. Cancer Discov. 2011;1:608–25.
23. VEGA. A Visualization Grammar. [https://vega.github.io].
24. Bostock M, Ogievetsky V, Heer J. D^3: Data-Driven Documents. IEEE Trans Vis Comput Graph. 2011;17:2301–9.
25. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662–9.
26. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics. 2013;14:178–92.
27. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214–20.
28. The Conda project [https://anaconda.org].
29. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell. 2015;160:1246–60.
30. Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, et al. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell. 2015;162:675–86.
31. Wang Q, Huang J, Sun H, Liu J, Wang J, Wang Q, et al. CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. Nucleic Acids Res. 2014;42(Database issue):D450–8.
32. Diaz A, Nellore A, Song JS. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol. 2012;13:R98.
33. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
34. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32:896–902.
35. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.
36. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
38. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32:1262–7.
39. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3–25.
40. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
41. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977;39:1–38.
42. Fox J. Applied Regression Analysis and Generalized Linear Models. London: SAGE Publications; 2015.