Nagy and Kampmann *BMC Bioinformatics* (2017) 18:347
DOI 10.1186/s12859-017-1759-9

BMC Bioinformatics

SOFTWARE Open Access

CRISPulator: a discrete simulation tool for pooled genetic screens

Tamas Nagy[1] and Martin Kampmann[2,3*]

Abstract

Background: The rapid adoption of CRISPR technology has enabled biomedical researchers to conduct CRISPR-based genetic screens in a pooled format. The quality of results from such screens is heavily dependent on the selection of optimal screen design parameters, which also affects cost and scalability. However, the cost and effort of implementing pooled screens prohibits experimental testing of a large number of parameters.

Results: We present CRISPulator, a Monte Carlo method-based computational tool that simulates the impact of screen parameters on the robustness of screen results, thereby enabling users to build intuition and insights that will inform their experimental strategy.
CRISPulator enables the simulation of screens relying on either CRISPR interference (CRISPRi) or CRISPR nuclease (CRISPRn). Pooled screens based on cell growth/survival, as well as fluorescence-activated cell sorting according to fluorescent reporter phenotypes are supported. CRISPulator is freely available online (http://crispulator.ucsf.edu).

Conclusions: CRISPulator facilitates the design of pooled genetic screens by enabling the exploration of a large space of experimental parameters in silico, rather than through costly experimental trial and error. We illustrate its power by deriving non-obvious rules for optimal screen design.

Keywords: CRISPR, CRISPRi, Functional genomics, Genome-wide screens, Simulation, Monte Carlo

Background

Genetic screening is a powerful discovery tool in biology that provides an important functional complement to observational genomics. Until recently, screens in mammalian cells were implemented primarily based on RNA interference (RNAi) technology. Inherent off-target effects of RNAi screens present a major challenge [1]. In principle, this problem can be overcome using optimized ultra-complex RNAi libraries [2, 3], but the resulting scale of the experiment in terms of the number of cells required to be screened can be prohibitive for some applications, such as screens in primary cells or mouse xenografts.

Recently, several platforms for mammalian cell screens have been implemented based on CRISPR technology [4]. CRISPR nuclease (CRISPRn) screens [5, 6] perturb gene function by targeting Cas9 nuclease programmed by a single guide RNA (sgRNA) to a genomic site inside the coding region of a gene of interest, followed by error-prone repair through the cellular non-homologous end-joining pathway. CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa) screens [7] repress or activate the transcription of genes by exploiting a catalytically dead Cas9 to recruit transcriptional repressors or activators to their transcription start sites, as directed by sgRNAs.

CRISPRn and CRISPRi have vastly reduced off-target effects compared with RNAi, and thus overcome a major challenge of RNAi-based screens. However, other challenges to successful screening [1] remain. The majority of CRISPRi and CRISPRn screens have been carried out as pooled screens with lentiviral sgRNA libraries. While this pooled approach has enabled rapid generation and screening of complex libraries, successful implementation of pooled screens requires careful choices of experimental parameters. Choices for many of these parameters represent a trade-off between optimal results and cost.

* Correspondence: Martin.Kampmann@ucsf.edu
[2]Department of Biochemistry and Biophysics, Institute for Neurodegenerative Diseases and California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA 94158, USA
[3]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
Full list of author information is available at the end of the article

Implementation

Code implementation and availability

CRISPulator was implemented in Julia (http://julialang.org), a high-level, high-performance language for technical computing. We have released the simulation code as a Julia package, Crispulator.jl. The software is platform-independent and is tested on Linux, OS X (macOS), and Windows. Installation details, documentation, source code, and examples are all publicly available at http://crispulator.ucsf.edu (see Availability and Requirements section for more details). CRISPulator simulates all steps of pooled screens, as visualized in Fig. 1 and explained in the Results section.

Simulated genome

A genome is defined by assigning a numerical, "true" phenotype to a number of genes, N. All simulations presented here have $N = 500$ genes. In the example shown in Fig. 2, 75% of genes were assigned a phenotype of 0 (wild-type), and 5% of genes were modeled as negative control genes, also with a phenotype of 0. 10% of genes were assigned a positive phenotype randomly drawn (unless otherwise indicated) from a Gaussian distribution with $\mu = 0.55$ and $\sigma = 0.2$ (clamped between [0.1, 1.0]), and 10% of genes were assigned a negative phenotype randomly drawn from an identical distribution except with $\mu = -0.55$ and clamping [−1.0, −0.1] (Fig. 2). Next, each gene was randomly assigned a phenotype-knockdown function (Fig. 3) to simulate different responses of genes to varying levels of knockdown. 75% of genes were assigned a linear function that linearly interpolates between 0 and the "true" phenotype from above as a function of knockdown, the remaining 25% of genes were assigned a sigmoidal function with an inflection point, p, drawn from a distribution with a mean of 0.8 and standard deviation of 0.2; the width of the inflection region, k, (over which a phenotype increased from 0 to the "true" phenotype, l) was drawn from a normal distribution with a mean of 0.1 and a standard deviation of 0.05. The function f was defined as follows:

$$f(x) = \begin{cases} 0, & x \le p{-}k \\ 1, & x \ge p+k \\ \frac{1}{2}\left(\frac{\operatorname{sign}(\delta)\cdot 1.05|\delta|}{|\delta|+1}+1\right), & \mathrm{p}{-}k < x < p+k \end{cases}$$

where $\delta = \frac{x-p}{\min\left(p,\, \min(1-p,k)\right)}$

This specific sigmoidal function was chosen over the more standard Gompertz function and the special case

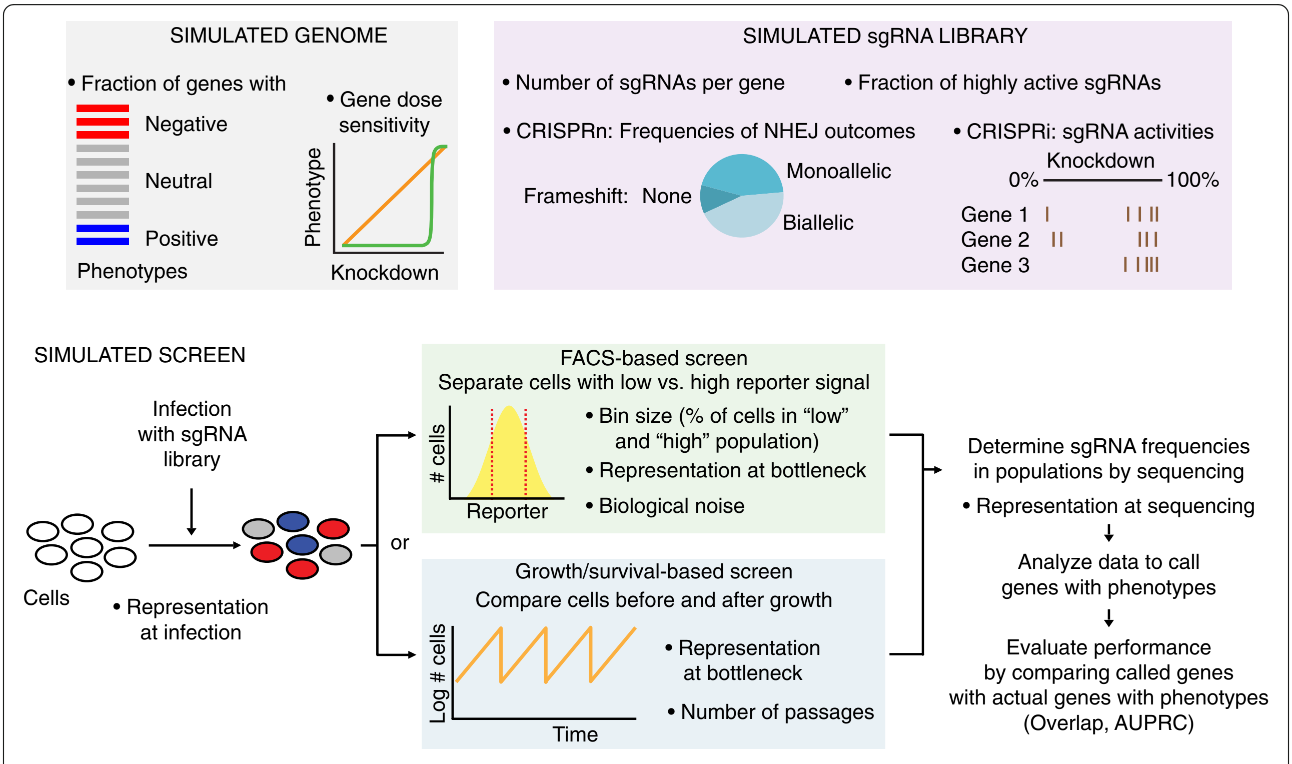
SIMULATED GENOME
• Fraction of genes with
Negative
Neutral
Positive
Phenotypes
• Gene dose sensitivity
Phenotype
Knockdown
SIMULATED sgRNA LIBRARY
• Number of sgRNAs per gene
• Fraction of highly active sgRNAs
• CRISPRn: Frequencies of NHEJ outcomes
Frameshift: None
Monoallelic
Biallelic
• CRISPRi: sgRNA activities
Knockdown
0%
100%
Gene 1
Gene 2
Gene 3
SIMULATED SCREEN
Infection with sgRNA library
Cells
• Representation at infection
or
FACS-based screen
Separate cells with low vs. high reporter signal
# cells
Reporter
• Bin size (% of cells in "low" and "high" population)
• Representation at bottleneck
• Biological noise
Growth/survival-based screen
Compare cells before and after growth
Log # cells
Time
• Representation at bottleneck
• Number of passages
Determine sgRNA frequencies in populations by sequencing
• Representation at sequencing
Analyze data to call genes with phenotypes
Evaluate performance by comparing called genes with actual genes with phenotypes (Overlap, AUPRC)

Fig. 1 CRISPulator simulates pooled genetic screens to evaluate the effect of experimental parameters on screen performance. Overview of simulation steps: Parameters listed with bullet points can be varied to examine consequences on the performance of the screen, which is evaluated as the detection of genes with phenotypes (quantified as overlap or area under the precision-recall curve, AUPRC). Details are given in the Implementation section

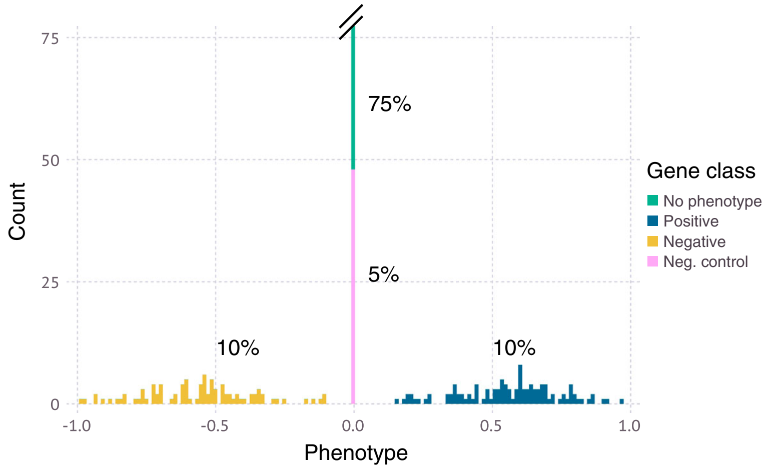

Fig. 2 Phenotype distribution in an example simulated genome. A typical distribution is shown, which includes 75% of genes without phenotype (*green*), 5% of negative control genes (*pink*), 10% of genes with a positive phenotype (*blue*), and 10% of genes with a negative phenotype (*yellow*). The frequencies of each category and strengths of the phenotypes are set by the user and are library specific (see text for more details). *N* genes are randomly given phenotypes from this artificial genome and used in later steps of the simulation

of the logistic function because it is highly tunable and has a range between 0 and l on a domain of [0, 1].

Simulated sgRNA libraries

CRISPRn and CRISPRi sgRNA libraries are generated to target the simulated genome. For the results featured here, each gene was targeted by $m = 5$ independent sgRNAs. For CRISPRi screens, each sgRNA was randomly assigned a knockdown efficiency from a bimodal distribution (Fig. 4): 10% of sgRNAs had low activity with a knockdown drawn from a Gaussian (μ = 0.05, σ = 0.07), 90% of guides had high activity drawn from a Gaussian (μ = 0.90, σ = 0.1). We assumed such a high rate of active sgRNAs based on our recently developed highly active CRISPRi sgRNA libraries [8]. For CRISPRn screens, high-quality guides all had a maximal knockdown efficiency of 1.0 and were 90% of the population (the 10% low-activity CRISPRn guides were drawn from the same Gaussian (μ = 0.05, σ = 0.07) as above). The initial frequency distribution of sgRNAs in the library was modeled as a log-normal distribution such that a guide in the 95th percentile of frequencies is 10 times as frequent as one in the 5th percentile (Fig. 5), which is typical of high-quality libraries in our hands [7].

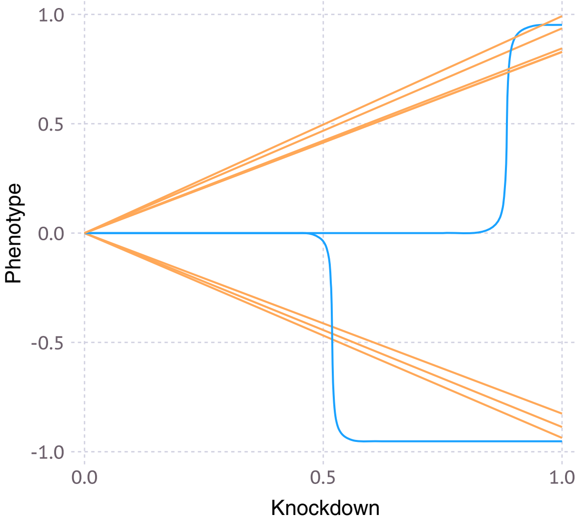

Fig. 3 Relationship between gene knockdown level and resulting phenotype for CRISPRi simulations. This relationship is defined for each gene, and represents either a linear function (*orange*) or a sigmoidal function (*blue*), as defined in the Implementation section

Simulated screens

Every step of the pooled screening process is simulated discretely. Infections are modeled as a Poisson process with a given multiplicity of infection, λ. The initial pool of cells is randomly infected by sgRNAs based on the frequency of each sgRNA in the library. A λ = 0.25 is used unless otherwise noted, which is commonly used to approximate single-copy infection [9]. Only cells with a single sgRNA are then used in subsequent steps, which is $P(x = 1; \text{Poisson}(\lambda = 0.25)) \approx 19.5\%$ of the initial pool.

For CRISPRi screens, phenotypes for each cell were determined based on the sgRNA knockdown efficiency (from above) and based on both the phenotype and the knockdown-phenotype relationship of the targeted gene. For CRISPRn screens, phenotypes for each cell were set using sgRNA knockdown efficiency (specific for CRISPRn screens, see previous section) and the gene phenotype. Our setup was such that if a cell was infected with a low-quality CRISPRn guide, it behaved similarly to one infected with a low-quality CRISPRi guide, i.e. mostly indistinguishable from WT. All cells with high-quality

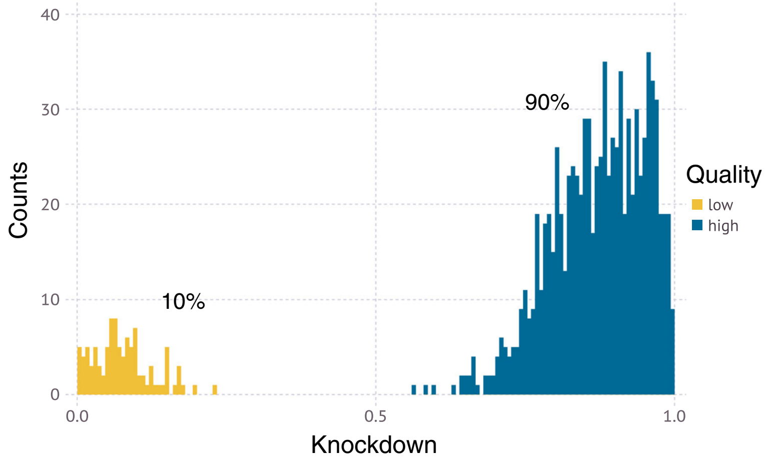

Fig. 4 An example sgRNA activity distribution for a simulated CRISPRi library. The 80–90% high quality guides is typical for second-generation CRISPRi [8] libraries. We define high quality sgRNAs as sgRNAs that have high activity and lead to a > 60% knockdown. Low quality sgRNAs are essentially indistinguishable from the negative controls and will lead to minimal effects on phenotype as they cause <20% knockdown of a given gene

guides CRISPRn guides had a 1/9, 4/9, or 4/9 chance of having 0%, 50%, or 100% knockdown efficiency, respectively (see Results for the underlying rationale). This knockdown efficiency was then used with the knockdown-phenotype relationship and true phenotype of the gene to calculate the observed phenotype.

FACS sorting was simulated by convolving the theoretical phenotypes of each cell independently with a Gaussian ($\mu = 0, \sigma$) where σ is a tunable "noise" parameter, reflecting biological variance in fluorescence intensity of isogenic cells. Populations of cells in FACS can be identified by the fitting of Gaussian mixture models [10], giving support for this approach. The number of cells prior to this step is termed the bottleneck representation and is tunable. Post-convolution, cells were sorted according to their new, "observed" phenotype and then the bottom *X* percentile and top 1 – X percentile (where X is a real value between 0 and 50) were taken as the two comparison bins.

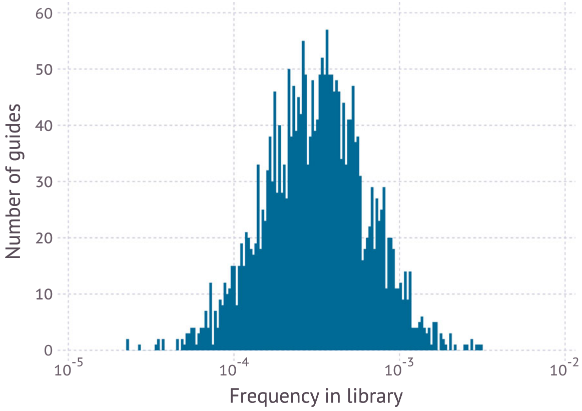

Fig. 5 Initial frequency distribution of plasmids encoding each sgRNAs in the library. An example of a typical distribution (in our experience) is shown, in terms of the spread of frequencies. During the chemical synthesis of oligos encoding each sgRNA in the library, there is variation in the initial frequency of each oligo and this is library-specific. The frequency distribution of a library used by a specific researcher can be determined empirically by next-generation sequencing of the plasmid library prior to conducting the screen

Growth experiments were simulated as follows: (1) in the time frame that WT cells (true phenotype = 0) divide once, cells with the maximal negative phenotype, −1, do not divide, and cells with maximal positive phenotype divide twice. For cells with phenotypes in between 0 and ±1, cells randomly pick whether they behave like WT cells or maximal phenotype cells weighted by their phenotype (i.e. cells with phenotypes close to 0 behave mostly like WT cells). (2) After one timestep where WT cells double once, a random subsample of the cells is taken. The size of the bottleneck is tunable. (3) This is repeated *n* number of times. Finally, the samples of cells at $t = 0$ and $t = n$ are taken as the two populations for comparison.

Sample preparation was simulated by taking the frequencies of each guide in the cells after selection and constructing a categorical distribution with the frequencies as the weights. Next-generation sequencing was then simulated by sampling from this categorical distribution up to the number of total reads. This approach for modelling next-generation sequencing of pooled libraries has been used successfully in earlier Monte Carlo simulations [11].

Evaluation of screen performance

Based on the simulated sequencing read counts, *P* values and gene-level phenotypes were calculated for each gene essentially as previously described [3, 7]. Briefly, observed sgRNA phenotypes were calculated as $\log_2$ ratios of sgRNA frequencies in two cell populations. Gene-level phenotypes were calculated by averaging the sgRNA phenotypes. *P* values were calculated based on the Mann–Whitney rank-sum test by comparing the phenotypes of sgRNAs targeting a given gene with the phenotypes of negative control sgRNAs. Genes were ranked by the product of the absolute gene-level phenotype and their $-\log_{10}P$ value to call hit genes. Screen performance was quantified in two ways (Fig. 6): As the overlap of the top 50 called hit genes with the top 50 actual hit genes (based on true phenotype), or as the area under the precision-recall curve (AUPRC). AUPRC was chosen over the more common area under the receiver operator characteristic (AUROC) due to the highly-skewed nature of the generated dataset (<20% of dataset is made up of true hits, based on the typical number of hits detected by CRISPR screens [5–7]). AUPRC is better able to distinguish performance differences between approaches on highly skewed datasets as compared to AUROC [12]. The AUPRC was calculated using a lower trapezoidal estimator, which had been previously shown to be a robust estimator of the metric [13]. The "signal" of an experiment was defined as the median signal for true hit genes (ones initially labeled as having a positive or negative phenotype). The true hit gene signal was calculated as the average ratio of the $\log_2$ fold change over the theoretical phenotype of all guides targeting that gene. Guides that dropped out of the analysis were excluded from the signal calculation. "Noise" was quantified as the standard deviation of negative-control sgRNA phenotypes, and the "signal-to-noise" ratio was the ratio of these two metrics. For display purposes, all are normalized in each graph.

Results

Here, we present a Monte Carlo method-based computational tool, termed CRISPulator, which simulates how experimental parameters will affect the detection of different types of gene phenotypes in pooled CRISPR-based screens. CRISPulator is freely available online (http://crispulator.ucsf.edu) to enable researchers to develop an intuition for the impact of experimental parameters on pooled screening results, and to optimize the design of pooled screens for specific applications. A previously published simulation tool, Power Decoder [11], addresses some of the parameters of interest for RNAi-based, growth-based screens. Our goal in developing CRISPulator was to enable the simulation of CRISPRi and CRISPRn screens for additional modes of pooled screening, such as FACS-based screens or multiple-round growth based screens, and to enable the exploration of more experimental parameters. Instead of measuring screen performance in terms of the power of identifying individual active shRNAs, we focus instead on the correct identification of hit genes, which is the primary goal of experimental genetic screens.

CRISPulator simulates all steps of pooled screens (Fig. 1). Briefly, a theoretical genome is generated in which genes are assigned quantitative phenotypes (Fig. 2). The user can set the size of the "genome", N, which corresponds to the number of genes targeted by the CRISPR library, e.g. a genome-wide human library would have $N \approx 20{,}000$. Additionally, the user can set the magnitude of both negative and positive phenotypes and their frequency in the genome. These values should be set based on the expected strength of the selection process and expected frequency of "hits." For example, for growth-based screens under standard culture conditions, mostly negative phenotypes are expected [5–7], whereas a comparable number of genes with positive phenotypes can be observed in screens in the presence of selective pressures, such as toxins [7] or drugs [5, 6, 14, 15].

Independently, the quantitative relationship between gene knockdown level and resulting phenotype is defined for each gene (Fig. 3). We will refer to a gene as a "linear gene" if the relationship between knockdown and phenotype is linear. Such linear genes are routinely observed in CRISPRi screens [7, 16]. A different class of

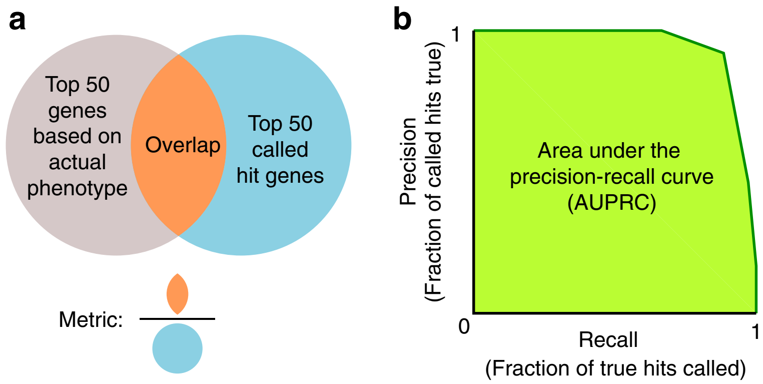
a
Top 50 genes based on actual phenotype
Overlap
Top 50 called hit genes
Metric:
b
Precision (Fraction of called hits true)
1
0
Area under the precision-recall curve (AUPRC)
Recall
(Fraction of true hits called)
1

Fig. 6 Metrics to evaluate screen performance. **a** "Venn diagram" overlap between the 50 genes with the strongest actual phenotypes, and the top 50 hit genes called based on the screen results – expressed as the ratio of the number of genes in the overlap over the number of called top hit genes, i.e. 50. **b** Area under the precision-recall curve (AUPRC)

genes, which we will refer to as “sigmoidal genes” displays a more switch-like behavior, where a phenotype is only observed above a certain level of knockdown [1]. As described in the Implementation section, the simulated genes contains both linear and sigmoidal genes, as observed for actual screens.

Next, a sgRNA library targeting this genome is defined. Each gene is targeted by a number of independent sgRNAs, m, that is set by the user and depends on the CRISPR library that they choose to use. Major libraries such as hGeCKOv2 [5] and hCRISPRi-v2 [8] have $m = 6$ and $m = 5$, respectively. For CRISPRi, the technical performance of each sgRNA is randomly assigned based on a user-defined distribution of sgRNA activities. A typical distribution, based on second-generation CRISPRi libraries [8] is shown in Fig. 4. For CRISPRn, 90% of sgRNAs are assumed to be highly active; however, the outcome of the DNA repair process resulting from sgRNA-directed DNA cleavage is stochastic. We assume that 2/3 of repair events at a given locus lead to a frameshift, and that the screen is carried out in diploid cells. All cells with active CRISPRn guides had a 1/9, 4/9, or 4/9 chance of having 0%, 50%, or 100% knockdown efficiency, respectively. The assumption that only bi-allelic frame-shift mutations lead to a phenotype in CRISPRn screens for most sgRNAs is supported by the empirical finding that in-frame deletions mostly do not show strong phenotypes, unless they occur in regions encoding conserved residues or domains [17]. To mitigate this issue, some CRISPRn screens have been conducted in quasi-haploid cell lines [6]. Future CRISPRn libraries may be designed to specifically target conserved residues, or incorporate algorithms that maximize the chance of frame-shift repair events. Once such libraries are validated, the stochastic outcomes for an active CRISPRn sgRNA can be updated to reflect the improved libraries.

Lastly, the initial frequency distribution of lentiviral plasmids encoding each sgRNA is specified (Fig. 5). These values are again library-specific and have to be set by the user. The frequency distribution can be determined empirically by next-generation sequencing of the library, and the distribution shown in Fig. 5 approximates distributions we routinely observe for our libraries generated in our laboratory.

Simulation of the screen itself discretely models infection of cells with the pooled sgRNA library, phenotypic selection of cells and quantification of sgRNA frequencies in selected cell populations by next-generation sequencing. Based on the resulting data (Fig. 7), hit genes are called using our previously described quantitative framework [3], as detailed in the Implementation section. The performance of the screen with a specific set of experimental parameters is evaluated by comparing the called hit genes to the actual genes with phenotypes

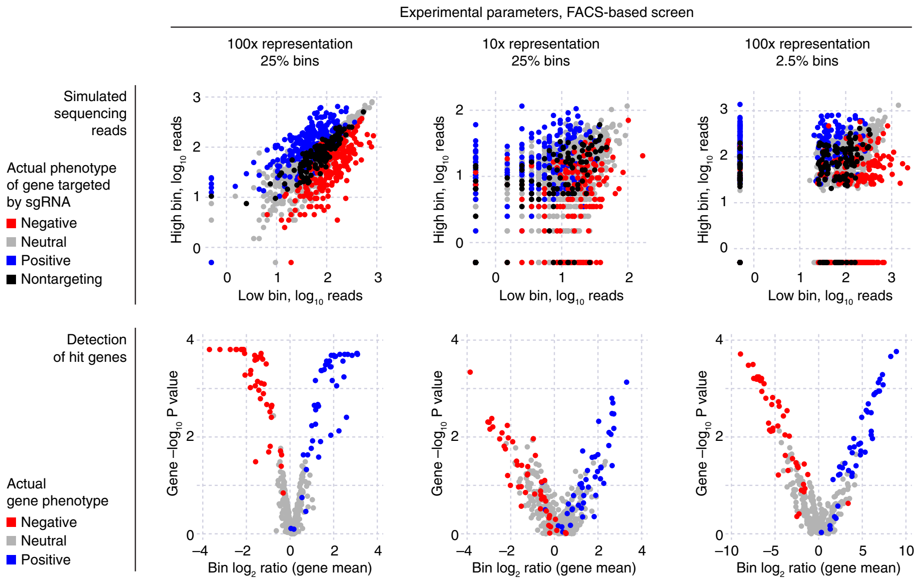

Fig. 7 Sample results from a CRISPulator simulation of a CRISPRi FACS-based screen. Top row: Each point represents and individual sgRNA, plotting its read numbers in the simulated deep sequencing run for the “low reporter signal” bin and the “high reporter signal” bin. sgRNAs are color-coded to indicate whether they target a gene with a positive phenotype (knockdown increases reporter signal, *blue*), a gene with a negative phenotype (knockdown decreases reporter signal, *red*), a gene without phenotype (*grey*), or whether they are non-targeting control sgRNAs (*black*). Bottom row: Based on the observed sgRNA phenotypes, gene phenotypes are calculated (mean $\log_2$ ratio of read frequencies in “high” over “low” bins), and a gene P value is calculated to express statistical significance of deviation from wild-type. These are visualized in volcano plots in which each dot represents a gene. Genes are color-coded to indicate the actual phenotype: positive, *blue*; negative, red; no phenotype, *grey*

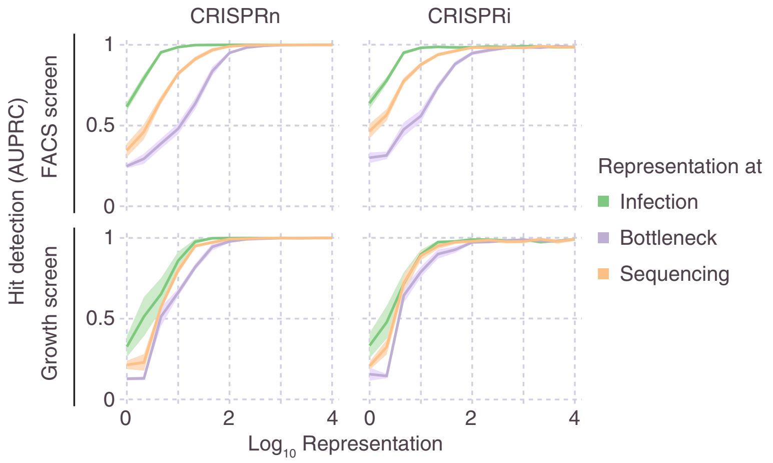

Fig. 8 Importance of representation of library elements at different stages of the screen. CRISPulator simulations reveal the effect of library representation at different screen stages (Transfection, bottlenecks, sequencing) on hit detection. Simulations were run for FACS-based screens (*top row*) and growth-based screens (*bottom row*). Lines and light margins represent means and 95% confidence intervals, respectively, for 10 independent simulation runs

defined by the theoretical genome. It is quantified either as overlap of the list of top called hits with the actual list of top hits, or as area under the precision-recall curve (AUPRC), a metric commonly used in machine learning [18] (Fig. 6).

A central consideration for all pooled screens is the number of cells used relative to the number of different sgRNAs in the library. We refer to this parameter as representation, and distinguish representation at the time of infection, representation at times during phenotypic selection, and – by extension – representation at the sequencing stage (where it is defined as the number of sequencing reads relative to the relative to the number of different sgRNAs). From first principles, higher representation is desirable to reduce Poisson sampling noise ("jackpot effects"), and has been shown empirically to improve results of pooled screens [3, 11, 19, 20]. In practical terms, higher representation is also more costly and difficult to achieve, for example when working with non-dividing cell types such as neurons [21]. A major application of CRISPulator is the exploration of parameters to guide the choice of suitable representation at each step of the screen to enable researchers to strike the desired balance between screening cost and performance.

CRISPulator implements two distinct strategies for phenotypic selection. In fluorescence-activated cells sorting (FACS)-based screens, cell populations are separated based on a fluorescent reporter signal that is a function of the phenotype. We [22] and others [23] have successfully implemented such screens by isolating and comparing cell populations with the highest and the lowest reporter levels. More commonly, pooled screens are conducted to detect genes with growth or survival phenotypes [5–7] by comparing cell populations at an early time point with cells grown in the absence or presence of selective pressures, such as drugs or toxins.

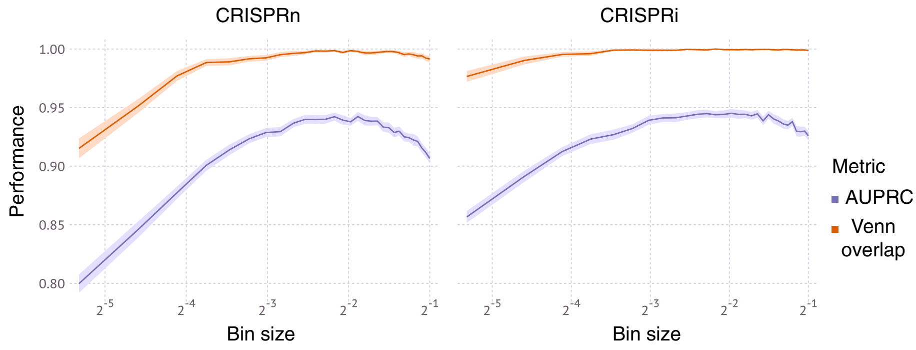

Fig. 9 Effect of bin size on performance of FACS-based screens. Simulations were run for 100× representation at the transfection, bottleneck and sequencing stages. Lines and light margins represent means and 99% confidence intervals, respectively, for 100 independent simulation runs

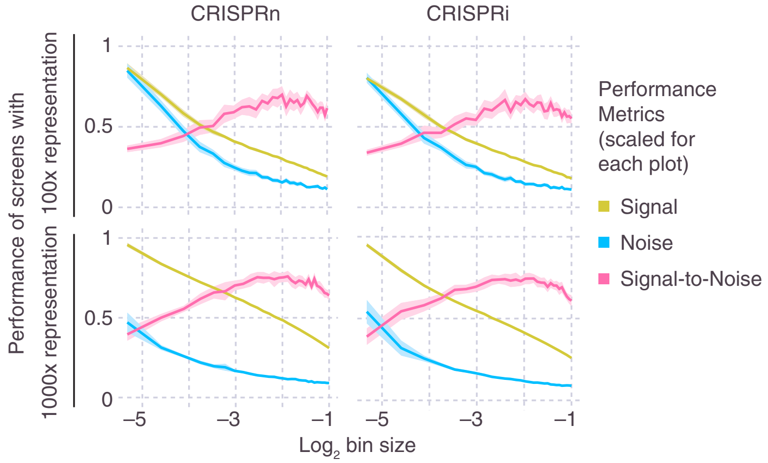

Fig. 10 Effect of bin size on signal and noise of FACS-based screens. For FACS-based screens, the effect of the size of the sorted bins (see Fig. 1) on metrics for signal, noise, and signal-to-noise ratio (scaled within each plot) is shown. Metrics are defined in the Implementation section. Simulations were run for 100× representation (top row) or 1000× representation (bottom row) at the transfection, bottleneck and sequencing stages. Lines and light margins represent means and 99% confidence intervals, respectively, for 25 independent simulation runs

We first asked how representation at the infection, selection and sequencing stages affects FACS- and growth-based screens (Fig. 8). The performance of FACS-based screens was most sensitive to the representation at the selection bottleneck, and least sensitive to representation at the infection stage, highlighting the importance of collecting a sufficient number of cells for each population during FACS sorting, ideally more than 100-fold the number of different library elements. By contrast, the performance of growth-based screens was similarly sensitive to representation at all stages.

For FACS screens using a given number of cells, an important decision is how extreme the cutoffs defining the “high-reporter” and “low-reporter” bins should be. CRISPulator simulation suggests that separating and comparing the cells with the top quartile and bottom quartile reporter activity results in the optimal detection of hit genes (Fig. 9). Closer inspection revealed that while both signal (sgRNA frequency differences between the two populations) and the noise (due to lower representation in the sorted population) decrease with larger bin sizes, the signal-to-noise ratio reaches a local maximum around 25% (Fig. 10), close to the bin size chosen fortuitously in published studies [22, 23].

For growth-based screens, the duration of the screen influences the signal (by amplifying differences in frequency due to different growth phenotypes) but also the noise (by increasing the number of Poisson sampling bottlenecks generated by cell passaging or repeated applications of selective pressure). Interestingly, CRISPulator suggests that the effect of screen duration on optimal performance is different for genes with positive and negative phenotypes, and strongly depends on the presence of genes with positive phenotypes (Fig. 11). While genes with positive phenotypes (increased growth/survival) were detected more reliably after longer screens, genes with negative phenotypes (decreased growth/survival) were optimally detected in screens of intermediate duration, and their detection in longer screens rapidly declined if genes with stronger positive phenotypes were present in the simulated genome. While genes with positive

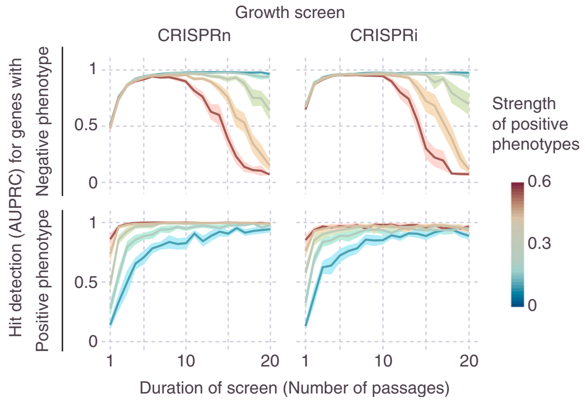

Fig. 11 Effect of positive phenotypes on growth-based screens. For growth-based screens, the presence of genes with positive phenotypes (fitter than wild type) strongly influences hit detection as a function of screen duration. Screens were simulated for a set of genes in which 10% of all genes had negative phenotypes (less fit than wild type), and 2% of genes had positive phenotypes. The strength of positive phenotypes was varied, as encoded by the heat map. Hit detection was quantified separately for genes with negative phenotypes (top row) and genes with positive phenotypes (bottom row). Simulations were carried out for screens with different durations, as measured by the number of passages. Lines and light margins represent means and 95% confidence intervals, respectively, for 25 independent simulation runs. In **a** and **c**, hit detection is measured as Area under the Precision-Recall curve (AUPRC), as detailed in the Implementation section

phenotypes are rare in screens based on growth in standard conditions [5–7], selective pressures, such as growth in the presence of toxin, can reveal strong positive phenotypes for genes conferring resistance to the selective pressure [7]. The optimal screen length for growth-based screens was dictated by a local maximum of the signal-to-noise ratio, which itself depended on the representation: screens with lower representation were performing better at shorter duration (Fig. 12). Our results therefore predict that especially for growth-based screens using selective pressures, and screens implemented with low representation, short durations are preferable.

A question that is vigorously debated in the CRISPR screening field is whether CRISPRn or CRISPRi based screens perform better. As both technologies are rapidly evolving, this question has not been settled. For example, in a side-by-side test of early implementations of these technologies, CRISPRn outperformed CRISPRi [24]. However, the second version of the genome-wide CRISPRi screening platform performed comparably to the best current CRISPRn platforms [8]. CRISPulator is not suitable to compare CRISPRi performance to CRISPRn performance – instead, it is suitable to simulate the impact of experimental parameters within one of these

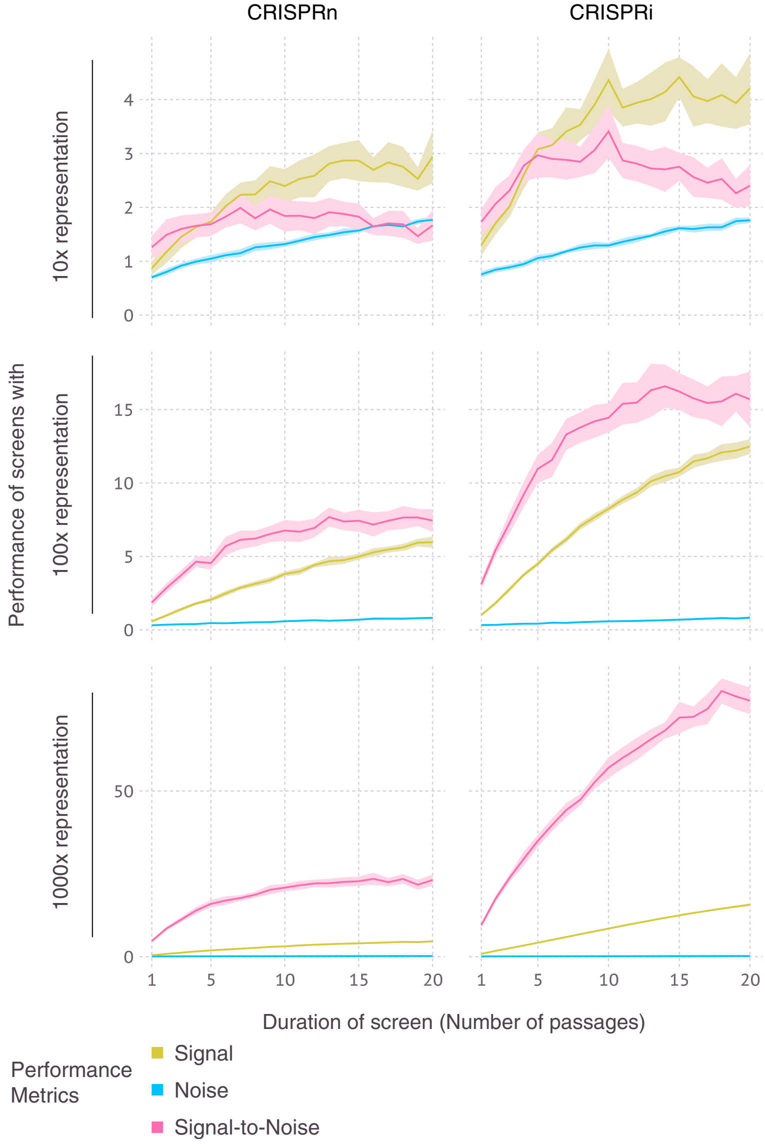

Fig. 12 Effect of duration of growth-based screens on performance. Screens were simulated for a set of genes in which 10% of all genes had negative phenotypes (less fit than wild type). Simulations were carried out for screens with different durations, as measured by the number of passages, and for different representations at the transfection, bottleneck and sequencing stages. Metrics for signal, noise, and signal-to-noise ratio are defined in the Implementation section. Lines and light margins represent means and 95% confidence intervals, respectively, for 25 independent simulation runs

screening modes. We were, however, able to make a prediction regarding the relative performance of CRISPRi and CRISPRn for different types of genes. While CRISPRn and CRISPRi screens performed similarly overall in the simulations described above (Figs. 8, 9, 10 and 11), separate evaluation of genes with linear versus sigmoidal phenotype-knockdown relationship revealed that CRISPRn outperforms CRISPRi for the detection of sigmoidal genes (which require very stringent knockdown to result in a phenotype), whereas CRISPRi performs relatively better for genes with a linear knockdown-phenotype relationship (Fig. 13).

Discussion

CRISPulator recapitulated rules for pooled screen design previously articulated for RNAi-based screens based on experimental and simulated data [11, 19, 20]. CRISPulator also revealed several non-obvious rules for the design of pooled genetic screens, illustrating its usefulness. Varying of several parameters in combination reveals areas in the multidimensional parameter space that are relatively robust, while in other areas, screen performance is highly sensitive to parameter changes (Figs. 11 and 12). Of particular practical importance to researchers designing or optimizing pooled screens are the following novel predictions:

(1) For FACS-based screens in which 2 cell populations are collected based on a continuous fluorescence phenotype, the best binning strategy is to collect the top quartile and bottom quartile of the population based on fluorescence (Fig. 9). This optimum is robust with respect to variation in other parameters we tested (Fig. 9).

(2) Optimal parameter choices for growth-based screens, in particular the number of passages, depend strongly on the genes with positive phenotypes (Fig. 11). While genes with positive phenotypes are rare in growth-based screens of cancer cell lines under standard culture conditions [5–7], a large number of genes with strongly positive phenotypes can be observed in screens in which cells are cultured in the presence of selective pressures, such as toxins [7] or drugs [5, 6, 14, 15]. Therefore, these seemingly similar modes of screening will require different parameters for optimal performance.

(3) Optimal passage number for growth-based screens also depends on the representation at bottleneck. Signal-to-noise reaches an optimum for lower passage numbers for screens with lower representation (Fig. 12), indicating that if high representation is not achievable (e.g. due to a limitation in available cells numbers), passage number should be reduced, relative to screens in which high representation can be achieved.

The simulated sequencing reads generated by CRISPulator (Fig. 7) recapitulate patterns observed in experimental data (Fig. 14), thereby facilitating the interpretation of suboptimal experimental data and providing a tool to predict which experimental parameters need to be changed to obtain data more suitable for robust hit detection.

Since certain parameters used by CRISPulator (such as the quality of sgRNA libraries or the signal-to-noise of FACS-based phenotypes) are estimates informed by published data, but not directly known, the predicted screen performance does not represent absolute performance metrics. Rather, the goal is to predict the relative performance of screens conducted with different experimental parameters to enable researchers to optimize those parameters.

While the simulations presented here focus on CRISPRn and CRISPRa, CRISPulator can also be used to

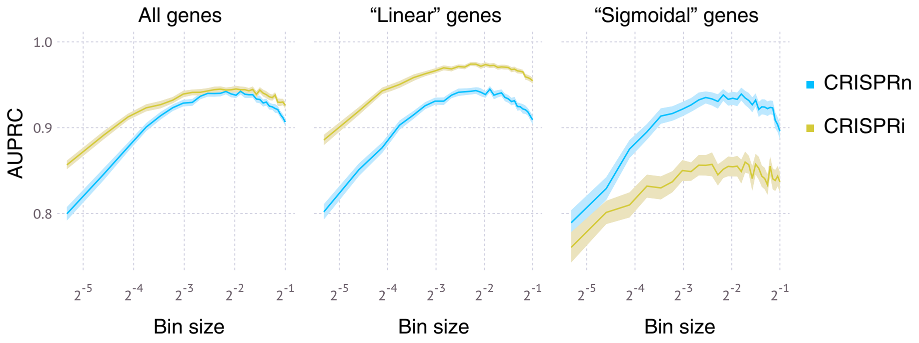

Fig. 13 Comparison of CRISPRn and CRISPRi screen performance for genes with different knockdown-phenotype relationships. Simulations of FACS-based screens were run for 100× representation at the transfection, bottleneck and sequencing stages. The simulated genome contained 75% of genes with a linear knockdown-phenotype relationship and 25% of genes with a sigmoidal knockdown-phenotype relationship, as defined in the Implementation section. Performance in hit detection was quantified as AUPRC either for all genes, or only for linear or sigmoidal genes. Lines and light margins represent means and 99% confidence intervals, respectively, for 100 independent simulation runs

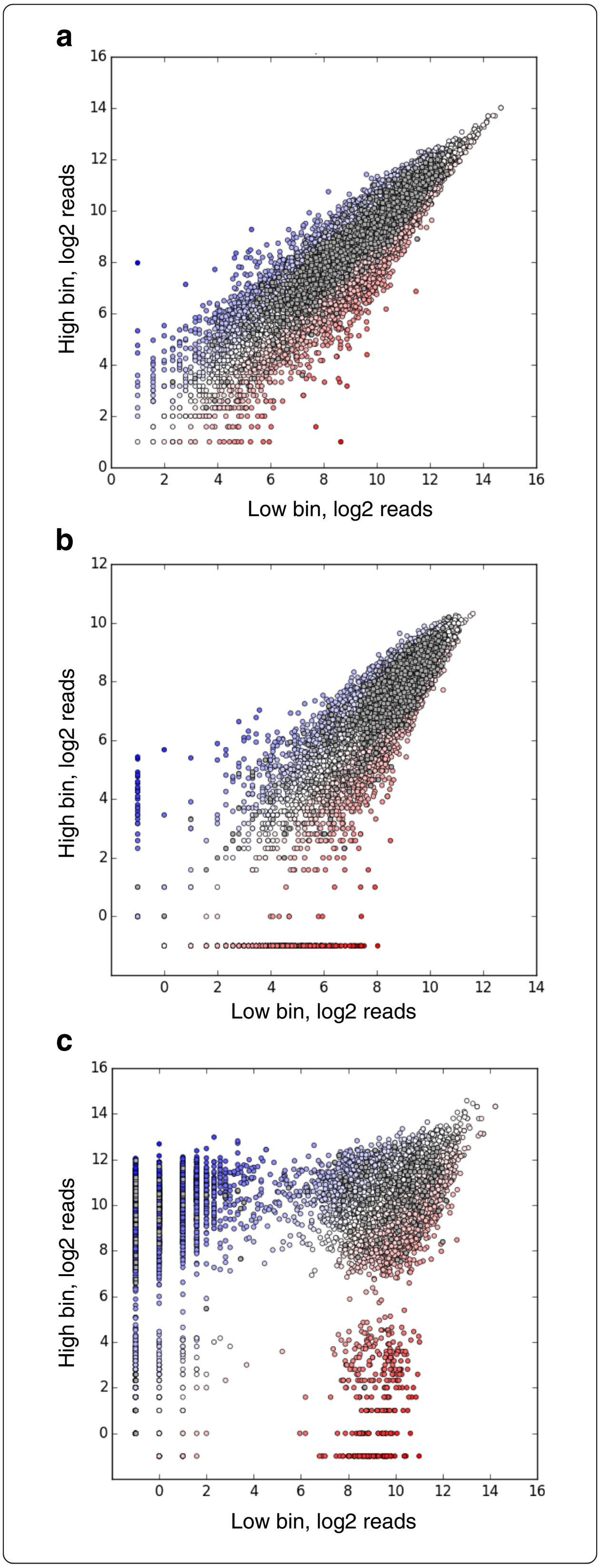

Fig. 14 Experimental data from FACS-based CRISPRi screens resembles simulated data shown in Fig. 6. Grey dots: non-targeting sgRNAs, dots on a red-white-blue color scale: targeting sgRNAs. Number of deep sequencing reads for each sgRNA in two populations separated based on a fluorescent reporter signal are shown. **a** Screen carried out with high representation at all stages. **b** Screen with low representation at the infection stage. **c** Screen with low representation at the selection stage

simulate CRISPRa gain-of-function screens, by reinterpreting the knockdown/phenotype relationship of genes (Fig. 3) as overexpression/phenotype relationships. Once more datasets from CRISPRa screens have been published, the empirical data can inform realistic choices of parameters for CRISPRa screen simulations.

Many of the lessons from CRISPulator should in principle also apply to RNAi-based screens. However, RNAi-based screens are notorious for off-target effects [1], which are difficult to predict, and which we chose not to model in CRISPulator.

Conclusions

CRISPulator facilitates the design of pooled genetic screens by enabling the exploration of a large space of experimental parameters in silico, rather than through costly experimental trial and error. For pooled genetic screens in animal models, such as mice, choices of experimental parameters can also have ethical implications, namely the numbers of animals required to power the study. As larger numbers of pooled genetic screens are published, we will further refine the assumptions underlying the simulation using empirical data.

Availability and requirements

Project name: Crispulator

Project home page: http://crispulator.ucsf.edu/

Repository page: https://github.com/tlnagy/Crispulator.jl

The repository provides the package, Jupyter notebooks to generate key figures, unit tests, and a command-line interface. The package can also be installed using the Julia Package Manager by running “Pkg.update(); Pkg.clone ("https://github.com/tlnagy/Crispulator.jl.git"); Pkg.build ("Crispulator")” inside the Julia command-line interface (REPL)

Operating system(s): any supporting Julia 0.5+ (tested on Linux, Mac, Windows)

Programming language: Julia >=0.5

Other requirements: Gadfly>=2.0.0, StatsBase, Distributions, DataFrames, HypothesisTests, Iterators, ColorBrewer, Gadfly, ArgParse, Compat >= 0.17.0, YAML

These packages are all installed automatically when installing Crispulator using the Julia package manager.

Licence: The Apache License 2.0 (http://www.apache.org/licenses/LICENSE-2.0)

Abbreviations

AUPRC: Area under the precision-recall curve; CRISPRa: CRISPR activation; CRISPRi: CRISPR interference; CRISPRn: CRISPR nuclease; FACS: Fluorescence-activated cell sorting; RNAi: RNA interference; SgRNA: Single guide RNA

Acknowledgments

We thank Ruilin Tian and Daniel Asarnow for feedback on this manuscript, and Diane Nathaniel, John Chen, Kathleen Keough and Xiaoyan Guo for sharing unpublished experimental data.

Funding

TN was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1650113. MK was supported by NIH/NIGMS New Innovator Award DP2 GM119139. MK is a Chan Zuckerberg Biohub investigator. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The software, CRISPulator, described in this study, which was also used to generate the data, is publicly available at http://crispulator.ucsf.edu (an archived version is available at https://doi.org/10.5281/zenodo.345188). It is released under Apache License 2.0 and there are no additional restrictions for commercial use.

Authors' contributions

TN and MK designed and analyzed the research and wrote the manuscript. TN developed the CRISPulator software. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

MK is an inventor on a patent application related to CRISPRi and CRISPRa screening (PCT/US15/40449).

Author details

[1]Graduate program in Bioinformatics, University of California, San Francisco, CA 94158, USA. [2]Department of Biochemistry and Biophysics, Institute for Neurodegenerative Diseases and California Institute for Quantitative Biomedical Research, University of California, San Francisco, CA 94158, USA. [3]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA.

Received: 17 March 2017 Accepted: 13 July 2017
Published online: 21 July 2017

References

1. Kaelin WG Jr. Molecular biology. Use and abuse of RNAi to study mammalian gene function. Science. 2012;337:421–2.
2. Kampmann M, Horlbeck MA, Chen Y, Tsai JC, Bassik MC, Gilbert LA, Villalta JE, Kwon SC, Chang H, Kim VN, Weissman JS. Next-generation libraries for robust RNA interference-based genome-wide screens. Proc Natl Acad Sci U S A. 2015;112:E3384–91.
3. Kampmann M, Bassik MC, Weissman JS. Integrated platform for genome-wide screening and construction of high-density genetic interaction maps in mammalian cells. Proc Natl Acad Sci U S A. 2013;110:E2317–26.
4. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015;16:299–311.
5. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343:84–7.
6. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343:80–4.
7. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159:647–61.
8. Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. elife. 2016;5
9. Fellmann C, Zuber J, McJunkin K, Chang K, Malone CD, Dickins RA, Xu Q, Hengartner MO, Elledge SJ, Hannon GJ, Lowe SW. Functional identification of optimized RNAi triggers using a massively parallel sensor assay. Mol Cell. 2011;41:733–46.
10. Chan C, Feng F, Ottinger J, Foster D, West M, Kepler TB. Statistical mixture modeling for cell subtype identification in flow cytometry. Cytometry Part A. 2008;73A:693–701.
11. Stombaugh J, Licon A, Strezoska Z, Stahl J, Anderson SB, Banos M, van Brabant SA, Birmingham A, Vermeulen A. The power decoder simulator for the evaluation of pooled shRNA screen performance. J Biomol Screen. 2015;20:965–75.
12. Goadrich M, Oliphant L, Shavlik J. Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction. Inductive Logic Programming, Proceedings. 2004;3194:98–115.
13. Boyd K, Eng KH, Page CD: In Machine Learning and Knowledge Discovery in Databases. Edited by Blockeel H, Kersting K, Nijssen S: Springer; 2013: 451–466.
14. Acosta-Alvear D, Cho MY, Wild T, Buchholz TJ, Lerner AG, Simakova O, Hahn J, Korde N, Landgren O, Maric I, et al. Paradoxical resistance of multiple myeloma to proteasome inhibitors by decreased levels of 19S proteasomal subunits. elife. 2015;4:e08153.
15. Kruth KA, Fang M, Shelton DN, Abu-Halawa O, Mahling R, Yang H, Weissman JS, Loh ML, Muschen M, Tasian SK, et al. Suppression of B-cell development genes is key to glucocorticoid efficacy in treatment of acute lymphoblastic leukemia. Blood. 2017;129:3000–8.
16. Anderson DJ, Le Moigne R, Djakovic S, Kumar B, Rice J, Wong S, Wang J, Yao B, Valle E, Kiss von Soly S, et al. targeting the AAA ATPase p97 as an approach to treat cancer through disruption of protein homeostasis. Cancer Cell. 2015;28:653–65.
17. Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol. 2015;33:661–7.
18. Davis J, Goadrich M: The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning 2006:233–240.
19. Sims D, Mendes-Pereira AM, Frankum J, Burgess D, Cerone MA, Lombardelli C, Mitsopoulos C, Hakas J, Murugaesu N, Isacke CM, et al. High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. Genome Biol. 2011;12:R104.
20. Strezoska Z, Licon A, Haimes J, Spayd KJ, Patel KM, Sullivan K, Jastrzebski K, Simpson KJ, Leake D, van Brabant SA, Vermeulen A. Optimized PCR conditions and increased shRNA fold representation improve reproducibility of pooled shRNA screens. PLoS One. 2012;7:e42341.
21. Kampmann M. A CRISPR approach to neurodegenerative diseases. Trends Mol Med. 2017;23:483–5.
22. Sidrauski C, Tsai JC, Kampmann M, Hearn BR, Vedantham P, Jaishankar P, Sokabe M, Mendez AS, Newton BW, Tang EL, et al. Pharmacological dimerization and activation of the exchange factor eIF2B antagonizes the integrated stress response. elife. 2015;4:e07314.
23. DeJesus R, Moretti F, McAllister G, Wang Z, Bergman P, Liu S, Frias E, Alford J, Reece-Hoyes JS, Lindeman A, et al. Functional CRISPR screening identifies the ufmylation pathway as a regulator of SQSTM1/p62. elife. 2016;5
24. Evers B, Jastrzebski K, Heijmans JP, Grernrum W, Beijersbergen RL, Bernards R. CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nat Biotechnol. 2016;34:631–3.