Genome Biology

METHOD Open Access

gscreend: modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection

Katharina Imkeller[1,2], Giulia Ambrosi[1], Michael Boutros[1] and Wolfgang Huber[2*]

Abstract

Pooled CRISPR screens are a powerful tool to probe genotype-phenotype relationships at genome-wide scale. However, criteria for optimal design are missing, and it remains unclear how experimental parameters affect results. Here, we report that random decreases in gRNA abundance are more likely than increases due to bottle-neck effects during the cell proliferation phase. Failure to consider this asymmetry leads to loss of detection power. We provide a new statistical test that addresses this problem and improves hit detection at reduced experiment size. The method is implemented in the R package gscreend, which is available at http://bioconductor.org/packages/gscreend.

Keywords: CRISPR screen, Genetic perturbation screen, Experimental design, Biology-based model, Generative probabilistic model, gscreend package, Bioconductor

Background

Genetic perturbation screens are a powerful tool to systematically probe gene function and genotype-phenotype relationships in many different cell types. Their applications include identification of genotype-specific vulnerabilities in human cancer cells [1–5], discovery of genes involved in drug resistance [6–8], and virus replication [9, 10].

Currently, the most widespread technology to induce specific genetic perturbations is based on CRISPR (clustered regularly interspaced short palindromic repeats)-associated enzymes. In this approach, DNA constructs encoding a guide RNA (gRNA) and the CRISPR-associated enzyme are stably integrated into cells, e.g., via lentiviral transduction. The gRNA directs the CRISPR-associated enzyme to its sequence-specific target site in the genome. To generate gene knockout perturbations, a common choice of enzyme is the endonuclease Cas9 (CRISPR associated 9) [6, 7, 11], which induces DNA cleavage at the genomic site it is directed to. Subsequent DNA repair via non-homologous end joining leads to frame-shift mutations and premature stop-codons, nonsense-mediated RNA decay, and finally gene knockout (CRISPR-KO). Alternatively, it is possible to introduce more subtle perturbations such as altered splicing patterns [12, 13] or quantitative modulation of gene expression [14]. To this end, modified CRISPR-associated enzymes which function as epigenetic modifiers [15–17], transcriptional modulators [18–20], or single-base editors [21, 22] are used.

Pooled screens enable the measurement of the effects of many genetic perturbations in parallel in a single experiment. To this end, a library of gRNAs is introduced into a pool of cells at a low multiplicity of infection such that no more than one gRNA is present in the vast majority of cells [23, 24]. The gRNA sequence simultaneously serves as a barcode that is used to trace which perturbation each cell carries. In the case of negative selection screens, the transduced cell pool is allowed to grow for several divisions during which the relative abundance of cells with a particular gRNA increases or decreases depending on the extent to which the targeted gene determines cell fitness. These effects are detected by amplifying, sequencing, and

*Correspondence: wolfgang.huber@embl.org
[2]European Molecular Biology Laboratory, Heidelberg, Germany
Full list of author information is available at the end of the article

counting the gRNAs before (library or T0) and after (T1) the proliferation phase (Fig. 1a).

A typical genome-wide CRISPR screening library for a mammalian genome contains between 70,000 and 120,000 gRNAs [2, 6, 7, 11, 26, 27]. To ensure statistical power, each gRNA must be represented by a sufficient number of cells during each step of the screen. When designing screening experiments, it is convenient to assume that all gRNAs are present in the library at approximately the same relative frequencies, and the library composition is summarized by the mean gRNA abundance, also referred to as coverage or representation. This measure is then used to calculate the necessary size of an experiment [23, 24]. For a library of 100,000 gRNAs and a desired coverage of 500 for example, 50 million cells (500 times 100,000) must be cultured. Published recommendations on optimal library coverage selection range from 200 [28] to 500 [23]. Nagy et al. used computational simulations to investigate the impact of screen parameters on the robustness of screening results, highlighting how coverage and screen duration can influence signal-to-noise ratios [29]. Further optimizing such experimental choices is a major thrust of this work, since they have substantial consequences on the size, costs, and outcomes of an experiment.

To compare the gRNA abundances before and after the proliferation phase, a range of statistical models and computational tools are available [30–37]. Common approaches are to model the joint bivariate null distribution of the normalized counts before and after the proliferation phase, or the null distribution of a univariate summary statistic, the ratio of these counts, hereafter referred to as the "before/after ratio." gRNAs whose data fall sufficiently outside the null distribution present evidence of a fitness effect of their target gene. Since it is common that each gene is targeted by multiple gRNAs, a subsequent step of the analysis consists of aggregating gRNA-level evidence to the gene level. This can be achieved for example using Bayesian hierarchical modelling [31, 35, 36] or robust rank aggregation [30] (see also Additional file 1: Table S1 for a short summary of the methods compared in this work).

For the null distribution modelling and hypothesis testing, approaches derived from RNA sequencing and differential gene expression analysis have been used [38, 39]. Here, we show that the distribution of the before/after ratios for negative controls is often asymmetric in CRISPR-KO screens; a similar observation has previously been reported for CRISPRi/CRISPRa screens [35]. Such asymmetry means that even in the absence of a fitness effect, a gRNA's relative abundance x is more likely to randomly decrease to, say, x/q ($q > 1$), rather than increase to xq during the screen. Failing to account for such asymmetry (as is done when using the RNA-

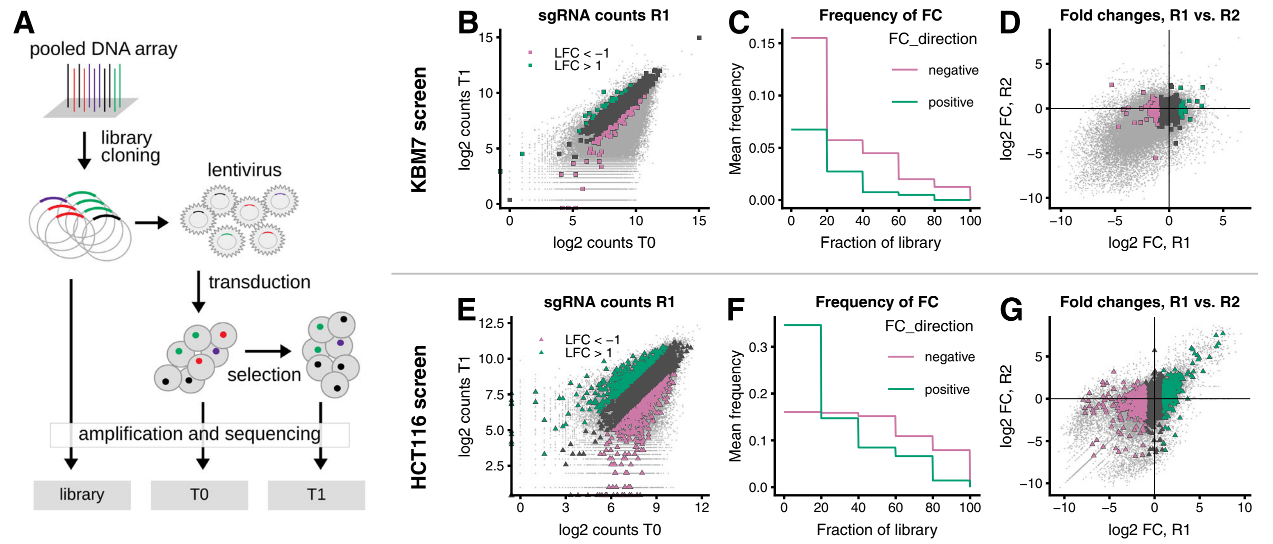

Fig. 1 Screening data show an asymmetric distribution of gRNA abundance fold changes. **a** Screen setup for measurement of gRNA effect on cell fitness. **b**–**d** KBM7 screen [1] with highlight on non-targeting controls. **e**–**g** HCT116 screen [2] with highlight on gRNAs targeting non-essential genes (defined according to Hart et al. [25]). **b**, **e** gRNA abundance at T1 compared to T0 for one of the replicates (R1). Non-targeting control gRNAs (**b**) and gRNAs targeting non-essential genes (**e**) are shown by large symbols, all other gRNAs as small grey points. LFC: logarithm of base 2 of inverted before/after ratio calculated as -log2((normalized count at T0 + 1)/(normalized count at T1 + 1)). Colors indicate LFC < -1 (pink) and LFC $> +1$ (green). **c**, **f** Fraction of non-targeting gRNAs (**c**) or gRNAs targeting non-essential genes (**f**) with LFC 1 (pink) and LFC $> +1$ (green). The gRNAs were sorted according to their abundance at T0, and the frequencies were calculated within each quintile of the abundance distribution (mean over two replicates). **d**, **g** Comparison of observed LFC for two replicates R1 and R2. Colors correspond to LFC in replicate R1 as in panels **b** and **e**

seq-based tools, which are designed for data that do not exhibit asymmetry) leads to needlessly elevated false-positive rates and/or decreased detection power.

Here, we present a biology-based, generative model that explains the asymmetry of the before/after ratios in pooled CRISPR screens and mechanistically links it to specific steps in the screening experiment. Based on our model, we derive a statistical test that we implemented in the R package `gscreend` and that enables accurate phenotype detection at reduced experiment size compared to existing approaches. Moreover, through our model, we can calculate the minimal experiment size necessary for a given screening library and a required detection power, a point that has never been systematically addressed in any published CRISPR screening protocol.

Results

Before/after ratios from pooled genetic screens have an asymmetric null distribution

We studied the distributions of the gRNA counts at T0 and T1 in two pooled CRISPR-KO conducted in human cell lines [1, 2] (Fig. 1). After scaling normalization of the counts to the total counts at T0, we computed the logarithm of the ratio of the counts after and before the proliferation phase (logarithmic fold change, LFC, see the "Methods" section). We focused on two classes of gRNAs: (a) those that should not have a fitness effect because their sequence does not match any region in the human genome (Fig. 1b–d [1]) and (b) those that target genes that are not essential for cell fitness according to a study by Hart et al. [25] (Fig. 1e–g [2]). The sign of their LFCs was uncorrelated between replicates, in agreement with the assumption that the LFCs were due to random experimental variability and not due to target-dependent fitness effects (Fig. 1d, g). However, the distribution of LFCs was not symmetric, in particular at the tails: values of LFC < -1 (strongly decreased abundance) were approximately 5–10% more frequent than those with LFC $> +1$ (strongly increased abundance) (Fig. 1b–c and e–f). These results are qualitatively in accordance with those of Daley et al. [35].

Computational simulation of pooled CRISPR screens

To investigate the origin of this asymmetry and possible impact of experimental design parameters, we developed a quantitative model and computational simulation of pooled CRISPR screens. The state space of the model is the tuple of integer counts of the gRNAs, which the model tracks as a function of time throughout the different steps of the screen (Fig. 2a). The temporal evolution of the state is described by endomorphic functions simulating subsampling during transduction, cell splitting, and sequencing as well as exponential cell growth according to a gRNA-specific growth rate. In our simulations, we considered screens performed with a library of 50,000 gRNAs (targeting 12,500 genes with 4 independent gRNAs per gene). For 10% of the genes, the knockout leads to a growth defect, and for 1% to increased growth. Table 1 summarizes the simulation parameters. A detailed description of the simulation algorithm is provided in the "Methods" section.

Plasmid library is a better reference sample than T0 cell pool

We first investigated the effect of the choice of reference sample. Previous publications used gRNA counts from either the plasmid library or the T0 cell pool as reference (Fig. 1a) [1, 2, 27, 40, 41], and it is unclear to what extent this choice influences the analysis outcome. Time point T0 is after the antibiotics selection of cells that were successfully transduced, in other words, up to four cell doublings after transduction. Such selection is necessary because at typical multiplicities of infection, only a fraction of cells is infected. In our simulations, we observed that counts of gRNAs targeting essential genes were already decreased at T0, especially for fast growing cells (Additional file 1: Figure S1A). To confirm this experimentally, we transduced pools of Cas9 expressing HCT116 and RKO cells with a genome-wide CRISPR library and selected the successfully transduced cells for 4 days, similar to the period before T0 in a screen. We sequenced the gRNAs in the plasmid library and at T0 and compared their normalized abundances. Similar to the prediction from the simulation, gRNAs targeting essential genes [42] had reduced abundance at T0 (Additional file 1: Figure S1B). This result implies that plasmid library rather than T0 sequencing should be used as a reference to avoid premature under-representation of gRNAs targeting essential genes.

The asymmetry of before/after ratios is caused by cell splitting during the proliferation phase

We next investigated the effect of experimental parameters on the distribution of before/after ratios for gRNAs without effect on cell fitness (Fig. 2b–e). The key parameter for cell culture during a screen is the mean gRNA coverage, which is reflected in the number of cells that are seeded after every round of splitting. We found that the smaller the coverage during cell splitting, the greater the asymmetry of before/after ratios (Fig. 2b–c). Higher levels of asymmetry lead to impairment of phenotype detection by MAGeCK-RRA [30], a current state-of-the-art analysis tool, which lost 10% of recall at 95% precision when, for example, reducing the cell splitting coverage from 400 to 100 (Fig. 2d). Similarly, the asymmetry increased when using faster growing cell lines, as we observed in our simulations (Fig. 2e) and in experimental datasets (Additional file 1: Figure S2) [40, 43]. We also tested the effect of other parameters, such as coverage during transduction

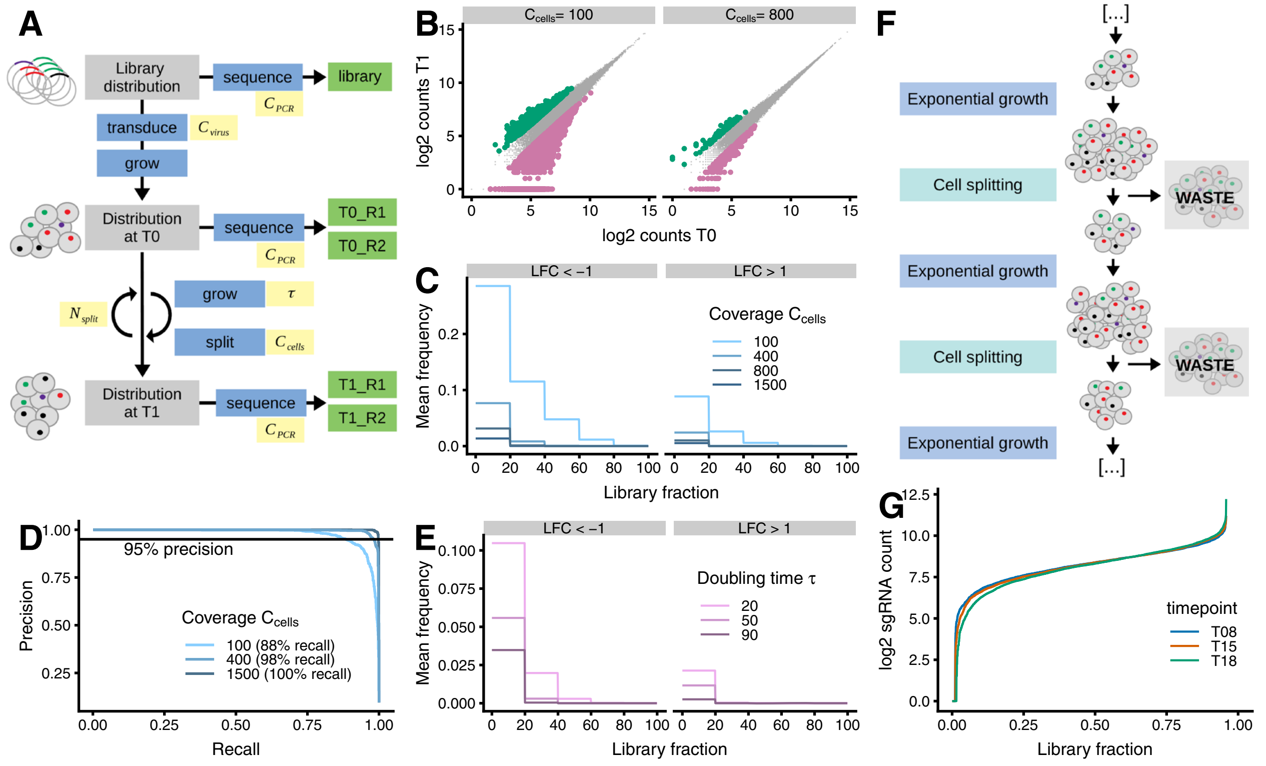

Fig. 2 Computational simulation explains how cell splitting causes asymmetry of before/after ratios. **a** Schematic representation of the simulation. After generation of an initial gRNA abundance distribution, different functions (blue) are applied to model transduction, cell growth, cell splitting, and sequencing. The simulation outputs the gRNA counts obtained by sequencing the plasmid library as well as the cell pools at time points T0 and T1 (green, R1 and R2 are technical replicates). The simulation depends on a set of user-defined parameters (yellow, see Table 1). **b**–**e** Simulation results for different values of cell splitting coverage C_{cells} and cell doubling time τ, while other parameters remain fixed. Only gRNAs without fitness effects are shown. **b** gRNA abundance at T1 compared to T0 for simulation with C_{cells} of 100 and 800. gRNAs with large observed fold changes are colored (LFC < -1 in pink, LFC $> +1$ in green). **c**, **e** Fraction of gRNAs with LFC < -1 (left) and LFC $> +1$ (right) for C_{cells} ranging from 100 to 1500 (**c**) and τ ranging from 20 to 90 h (**e**). Mean over 5 simulations is depicted. **d** MAGeCK-RRA precision-recall curves on data simulated using different values for C_{cells} (100, 400, and 1500). The recall at 95% precision is indicated. **f** Schematic representation of cell splitting during the proliferation phase of screen, which consists of multiple rounds of exponential growth and random sampling. **g** Count distribution of gRNAs targeting non-essential genes at T08, T15, and T18 of the screen in HCT116 cells [2]. gRNAs were ranked according to their abundance and the resulting ranks normalized to [0;1] (library fraction, *x*-axis). On the *y*-axis, the counts per gRNA are shown

or polymerase chain reaction (PCR). These, however, only marginally influenced the asymmetry of before/after ratios in our simulations (Additional file 1: Figure S3). Decreasing the cell splitting coverage led to up to 20% of gRNAs with LFC < -1, whereas for similar changes in PCR or transduction coverage, this fraction was only 3% (Fig. 2c and Additional file 1: Figure S3).

We conclude that transduction and PCR do not cause major technical biases in the data and that it is better to sequence the gRNA pool in the plasmid library rather than at T0. The observed asymmetry however can be mechanistically explained by multiple rounds of cell splitting bottlenecks and exponential growth (Fig. 2f). With every round of exponential growth followed by random sampling of cells, the distribution of gRNA abundances gets wider, i.e., there are more and more gRNAs that are underrepresented in comparison with the mean gRNA coverage. We confirmed the gradual broadening of the abundance distribution of gRNA targeting non-essential genes [25] in a published dataset from a CRISPR-knockout screen performed in HCT116 cells (Fig. 2g) [2].

Wide initial gRNA abundance distributions increase asymmetry of before/after ratios

Since the observed asymmetry is caused by broadening of the gRNA abundance distribution, we hypothesized that the width of the gRNA abundance distribution in the plasmid library also influences the data quality of the screen. A measure of the width of this distribution, i.e., the difference in abundance between low and high abundant gRNAs, is the ratio between the 90 and 10% percentiles. This measure, elsewhere also named “skew ratio” [23], will hereafter be referred to as “distribution width.” If, for example, the most abundant 10% gRNAs of a library have

Table 1 Parameters of CRISPR screen simulation

Symbol	Variable in software	Default value	Description
C_{virus}	cov_virus	400	Coverage during viral transduction.
C_{cells}	cov_cells	400	Coverage during cell culture.
C_{PCR}	cov_pcr	400	Coverage for PCR amplification.
L	lib_width	7.5	Library distribution width.
τ	dupl_time	30	Cell duplication time in hours.
ϕ_{neg}	freq_negfc	0.1	Fraction of gRNAs with negative fitness effect.
ϕ_{pos}	freq_posfc	0.01	Fraction of gRNAs with positive fitness effect.
N_{tot}	n_sgrnas	50,000	Number of gRNAs in total.
N_{gRNA}	n_sgrnas_per_gene	4	Number of gRNAs per gene.
N_{libpcr}	n_repl_lib_pcr	2	Number of replicates for library sequencing.
N_{bio}	n_repl_sel	10	Number of biological replicates.
N_{biopcr}	n_repl_pcr	3	Number of sequencing replicates per biological replicate.
N_{split}	n_splittings	7	Number of cell splittings.
Δ_t	-	72	Time between cell splittings in hours.

an abundance higher than 500 whereas the least abundant 10% have less than 100 counts, the distribution width is 5.

We performed simulations starting from three different gRNA libraries with varying distribution width (Fig. 3a). Our simulations showed that with higher width, the reproducibility between replicates decreased (Fig. 3b) and at the same time the frequency of gRNAs with LFC < -1 increased (Fig. 3c). Experimental data from screens conducted using plasmid libraries with different distribution widths confirmed this finding (Fig. 3d–f) [2, 27, 40, 41, 43, 44]. Using plasmid libraries with narrower gRNA abundance distributions thus increases data quality by reducing the asymmetry of the distribution of before/after ratios.

Furthermore, we also found that the gRNA sequence composition of a library correlates with its width and that gRNAs with specific sequence properties are more likely to be over- or underrepresented (Additional file 1: Figure S4). To show this, we selected five datasets from published CRISPR libraries [1, 27, 40, 43–45]. These libraries have different distribution widths ranging from 2.4 to 8.8 (Additional file 1: Figure S4A-B). To examine the sequence composition, we generated probability sequence motifs for the least and most abundant gRNAs (Additional file 1: Figure S4C) [46]. Wider libraries tend to have poly-G-stretches in low abundant gRNAs and poly-T-stretches high abundant gRNAs (Additional file 1: Figure S4D). This is probably due to sequence-specific biases during the generation of the plasmid library, for example during synthesis or PCR amplification of gRNAs.

New statistical method for improved phenotype detection

We showed that before/after ratio distributions in pooled CRISPR screens are asymmetric due to technical artifacts arising during the cell proliferation phase. This asymmetry is influenced not only by cell splitting parameters but also by the width of the gRNA abundance distribution in the plasmid library. In principle, it would be possible to eliminate this asymmetry by using plasmid libraries with minimal distribution width and to perform the screen at very high coverage. However, since this is generally neither feasible nor economically reasonable, we developed a new statistical test that accounts for the asymmetric null distribution. The underlying idea of our method is to use a skew normal distribution to model the LFC null distribution.

The workflow of our new analysis method `gscreend` is depicted in Fig. 4a. After scaling of gRNA counts and calculation of LFCs, the data is split into slices according to the gRNA abundance in the reference sample (e.g., plasmid library). We introduce this stratification since it allows the parameters of the null distribution to be different for gRNAs with low and high abundance, consistent with what we observed in datasets. We model the LFCs in each stratum as a mixture of a parametric null distribution, the skew normal, and an unspecified alternative distribution [47, 48]. The first mixture component corresponds to gRNAs without fitness effect, the second to those with effect, where we assume that these are only a minority. `gscreend` uses least quantile of squares regression [49] to fit the null distribution to the LFCs in each stratum. Least quantile of squares regression fits a model by only taking into account a defined proportion of residuals, e.g., those between the 10 and 90% percentiles. In contrast to the commonly used least sum of squares regression, it is thus more robust to outliers. In the `gscreend` workflow, the resulting null models for every stratum are used to calculate p values, which are then employed to rank the gRNAs. Subsequently, robust rank aggregation [30, 50] is applied to aggregate the ranked gRNA list to the gene level.

We first tested how well `gscreend` performed in accurately ranking genes in experimental datasets. Using a

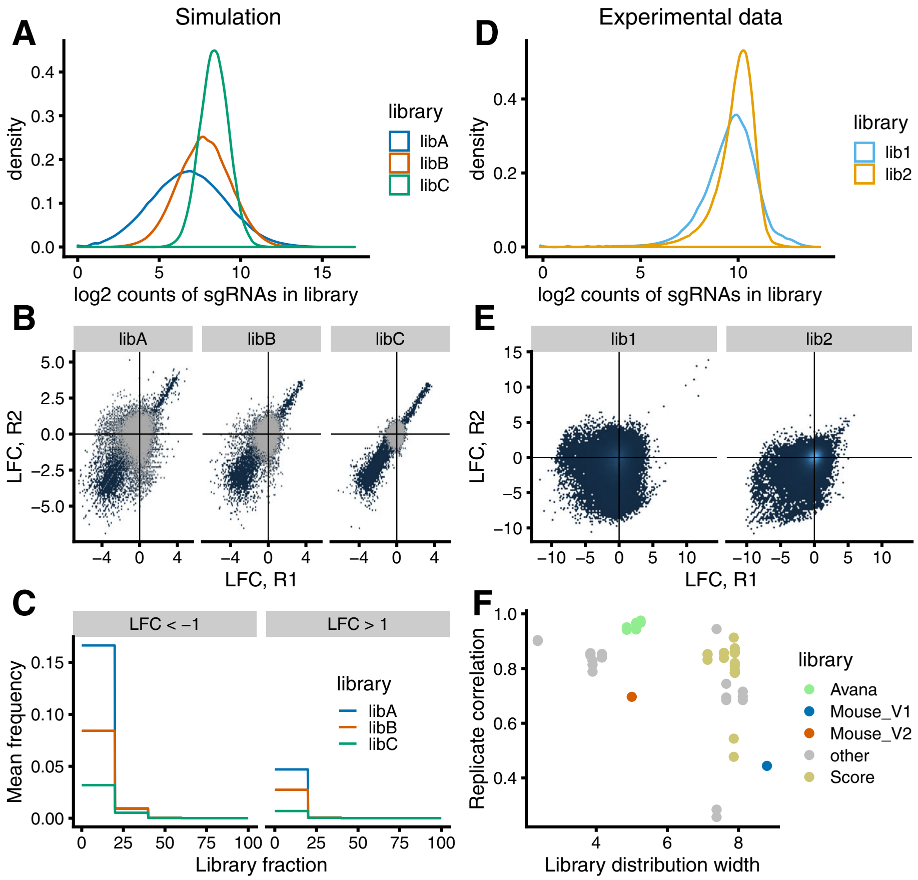

Fig. 3 Wider gRNA abundance distributions in the plasmid library increase asymmetry of before/after ratios. **a**–**c** Simulation results using three libraries with different widths of gRNA abundance distribution. **a** log2 count distribution of simulated libraries libA, libB, and libC. Distributions were generated as log-normal distributions with same log-mean, but differing log-sd and then size-normalized. Library distribution widths were as follows: 66.5 (libA), 17.8 (libB), and 4.8 (libC). **b** Reproducibility of LFC between two replicates. The simulations were performed with libA, libB, or libC. All gRNAs are shown in blue, and gRNAs without fitness effect are highlighted in gray. **c** Fraction of gRNAs with LFC < -1 (left) and LFC $> +1$ (right) for simulations with libA, libB, or libC. gRNAs used for frequency calculation do not have fitness effects. **d**–**e** Dataset from screen performed in mESCs using two different libraries lib1 and lib2 (libraries Mouse_V1 and Mouse_V2 respectively from Tzelepis et al. [27]). Library distribution widths were as follows: 8.8 (lib1) and 5.0 (lib2). **d** log2 count distributions of lib1 and 2, normalized to total count. **e** Reproducibility of LFC between two replicates in the screens performed with lib1 or lib2. All gRNAs are shown in blue. **f** Replicate correlation as a function of library distribution width for different published screens [2, 27, 40, 41, 43, 44]. Data for Avana, Human_V1 (also known as Score), Mouse_V1, and Mouse_V2 is highlighted in color

published list of essential and non-essential genes [25, 42], we calculated the recall at 95% precision (as in Fig. 2d) of our and other tools [30, 31, 33, 35, 37]. Additional file 1: Table S1 summarizes the different statistical concepts underlying the six methods. `gscreend` outperformed MAGeCK, ScreenBEAM, CRISPhieRmix, and CRISPRBetaBinom when ranking genes in a CRISPR-knockout screen performed in HCT116 cells (Fig. 4b) [2]. BAGEL was the only tool that had a better precision-recall performance than `gscreend` on these data. However, its algorithm was trained on the same benchmark set of essential and non-essential genes that we used here to calculate precision-recall statistics, which might explain some of this performance. Indeed, when ranking components of the ribosome, whose knockout is likely to be lethal, `gscreend` outperformed BAGEL, MAGeCK, ScreenBEAM, CRISPhieRmix, and CRISPRBetaBinom, especially within the 1000 lowest ranked genes (Fig. 4c). In order to illustrate some examples, we highlighted the results for five selected genes (Fig. 4d, e). *MRPL34* and *MRPS12* (components of the mitochondrial ribosome) are detected with low rank only by `gscreend`, although their gRNA abundance profile indicates that they are truly essential, because two of the corresponding gRNAs are strongly depleted at T18 in all three replicates (Fig. 4e). The other three genes *NDUFAF3*, *PRRC2A*, and *UVRAG* are assigned low ranks in BAGEL or MAGeCK but remain above the 1% false discovery rate (FDR) threshold in the `gscreend` results (Fig. 4d). Their gRNA abundance profile indicates that the observed negative LFCs are technical

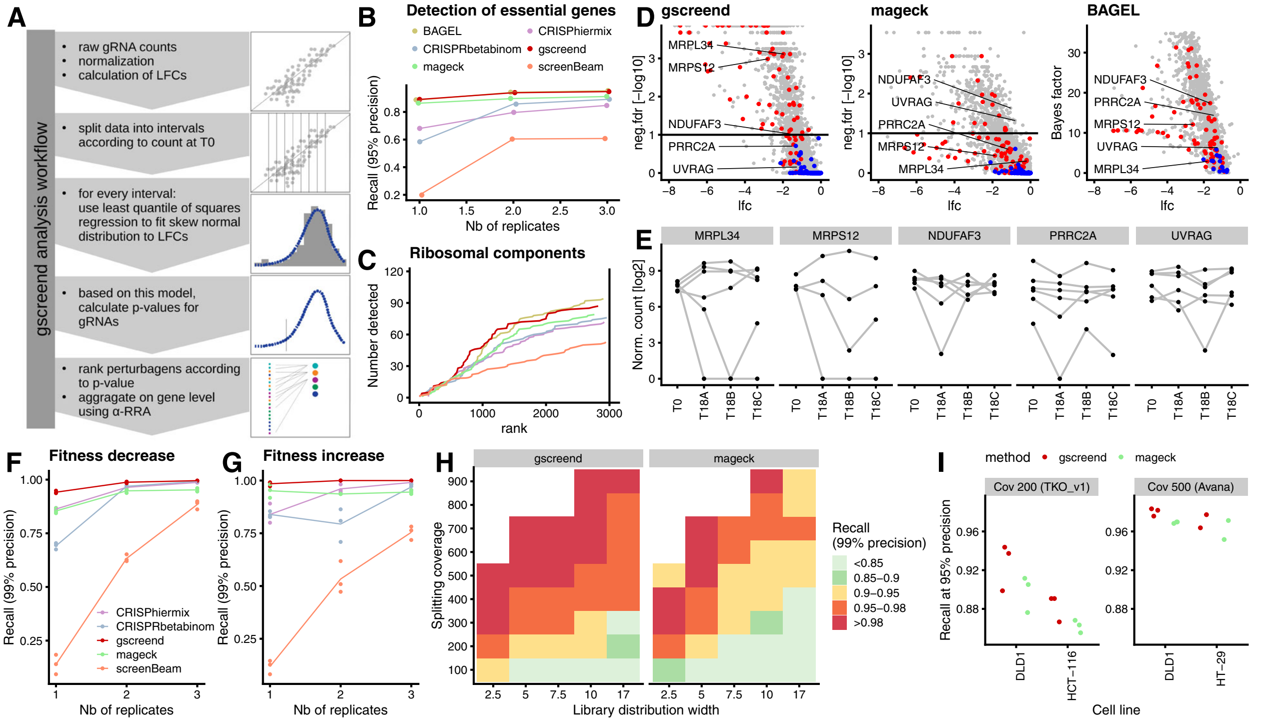

Fig. 4 gscreend increases phenotype detection accuracy. **a** `gscreend` analysis workflow. **b**–**e** Comparison of gene ranking by BAGEL, CRISPhieRmix, CRISPRBetaBinomial, MAGeCK, ScreenBEAM, and `gscreend` for CRISPR knockout screen performed in HCT116 cells [2]. **b** Recall at 95% precision (analysis as in Fig. 1d) for 1 to 3 biological replicates of timepoint T18. Essential and non-essential genes were defined according to Hart et al. [25, 42]. **c** Ranking of gene encoding ribosome components (structural constituent of ribosome - GO:0003735) by the six different analysis tools and using all 3 biological replicates. **d** Volcano plots illustrating `gscreend`, mageck, and BAGEL analysis results for gene encoding ribosomal components (red) and non-essential genes [25] (blue). Horizontal lines indicate FDR thresholds of 1%. **e** Log2 normalized abundances of gRNAs targeting the 5 selected genes at time point T0 and 3 replicates of time point T18. **f**–**g** Recall at 99% precision by CRISPhieRmix, CRISPRBetaBinomial, `gscreend`, MAGeCK, gscreend, and ScreenBEAM for simulated data of 1 to 3 biological replicates. Other simulation parameters: library distribution width 7.5, cell splitting coverage 200, doubling time 30 h. Precision-recall curves were calculated for detection of essential (**f**) and growth-suppressing (**g**) genes. **h** Recall at 99% precision of essential genes for simulations with different library width and cell splitting coverage. Genes were ranked using `gscreend` (left) or MAGeCK (right). **i** Recall at 95% precision for gscreend and MAGeCK analysis of screens performed either using a cell splitting coverage of 200 and the TKO_v1 library (library distribution width ca. 7.8) [2] or a cell splitting coverage of 500 and the Avana library (library distribution width ca. 4.6) [40]. Essential and non-essential genes were defined according to Hart et al. [25, 42]

artifacts, because they were not reproduced between replicates (Fig. 4e). Taken together, these results indicate that `gscreend` delivers superior accuracy in ranking and identifying essential genes in pooled negative selection screens.

We also investigated whether our method is robust against different levels of the asymmetry of before/after ratios by simulations. Similar to what we found using the experimental data, `gscreend` had a better ranking accuracy than the other tools when detecting genes that either increase or decrease cell fitness (Fig. 4f, g). When increasing the asymmetry by reducing cell splitting coverage or increasing library distribution width, the method maintained a better ranking accuracy than MAGeCK-RRA (Fig. 4h). `gscreend` enables reduction of the cell splitting coverage by approximately 50% for a library distribution width of 7.5: using 300 (`gscreend`) instead of 600 (MAGeCK) mean gRNA coverage maintained at least 95% recall at 99% precision (Fig. 4h). For libraries with larger distribution widths, the gain in accuracy is even more substantial.

To further assess the relevance of these findings for the analysis of experimental data, we compared the performance of gscreend and MAGeCK on a series of datasets with highly different experimental setups. The screens performed by Hart et al. [2] were conducted with a coverage of 200 and a library distribution width of around 7.8; according to our analysis, these parameters lead to high asymmetry and high levels of noise. In contrast, screens from the DepMap consortium [40, 43] were conducted at a coverage of 500 and a library width of around 4.6, which represents a more favorable experimental setup. When comparing precision-recall performance, and more specifically recall at 95% precision, `gscreend` outperforms MAGeCK on both datasets (Fig. 4i). However, the difference was small for the DepMap data, while it was more substantial for the Hart et al. dataset. These results show that gscreend can provide tangible improvements on real data, although the improvements are less pronounced for data produced from optimal experimental designs.

Implications for the design of screening experiments

Based on the findings reported above, we suggest that when designing a screen, the distribution width of the gRNA plasmid library should first be measured. Based on this measure, our simulation tool (Fig. 4h, left panel), can then be used to predict the corresponding optimal coverage (summarized in Table 2). Libraries with a narrow width can be screened at lower coverage than wide ones to achieve the same signal-to-noise ratio, since the impact of asymmetric loss is smaller. This, in turn, may have a significant impact on the costs and effort associated with the experiment.

Discussion

Accurately detecting phenotypes in pooled genetic perturbation screens is key to generating hypotheses that justify follow-up. Screens that correctly distinguish all genes that negatively or positively regulate cell fitness can be used not only to identify the strongest "hits," but also to measure subtle differences in growth rate and thus map whole pathways and potentially identify mechanisms.

To achieve high data quality and accurate analysis, we need to understand how the experimental design influences the results. Previously reported simulations of CRISPR-based screens highlighted the importance of coverage for reducing the signal-to-noise ratio [29]. Our study is the first to systematically explore the influence of experimental design, including the quality of the gRNA library—as measured by the library distribution width—on phenotype detection in pooled screens. Given a certain gRNA library distribution width, our method provides a quantitative prescription for the choice of cell splitting coverage. We show that gRNA coverage during PCR and transduction, provided it is in the same range as the cell splitting coverage, only marginally influences data quality. We also find that screens are best analyzed when using plasmid library sequencing as reference. We do not discuss the influence of the multiplicity of infection during viral transduction, as there is already literature and a good model available to address this point [24]. Our most important novel finding is that the asymmetry of the distribution of before/after ratios is caused during the proliferation phase of pooled negative selection screens. Multiple consecutive rounds of cell splitting and exponential growth gradually lead to random loss of low abundant gRNAs.

Table 2 Recommended screening coverage for different library distributions

Library width	Screening coverage
2.5	200
5	300
7.5	300
10	400
17	400

Using this understanding of the asymmetric null distribution of before/after ratios, we developed a new statistical test that improves phenotype detection. gscreend outperforms existing analysis methods, which rely on the assumption that the null distribution is symmetric. From the point of view of screen design, our method enables reduction of experiment size by up to 50% compared to other tools, because it maintains high analysis accuracy throughout a broad range of experimental settings. Especially for experiments that are limited by their size because of limited supply of cells, for example in primary cell cultures [39, 51], our method may help to improve phenotype detection.

Our results also provide indications on how to optimize the experimental design by choosing the screening coverage according to the width of the plasmid library (Table 2). Intriguingly, the width of the library distribution is the limiting parameter that dictates the minimal size of a screening experiment. It would thus be possible to strongly reduce the experiment size by using a library with a narrower distribution. Our analyses indicate that gRNAs with specific sequence characteristics are likely to be over- or underrepresented in gRNA libraries obtained using arrayed synthesis approaches and cloning. We hypothesize that the broadening of library distribution is due to sequence-specific differences in synthesis or amplification efficiency. A recently published approach to synthesize covalently-closed-circular-synthesized (3Cs) gRNA libraries may thus be a promising technology for substantial reduction of library width and experiment size [52].

Finally, the discovery of sequence specific representation differences of gRNAs in a library also has important implications for the evaluation of their gene knockout efficiency. gRNAs with specific sequence properties might seem more efficient than others simply because they are less abundant in the library and thus more likely to suffer from the here described asymmetric loss phenomenon [27, 53].

Conclusion

We conclude that the asymmetry of the before/after ratio distribution in pooled CRISPR screens is primarily caused by insufficient coverage of gRNAs during the cellular growth phase of a screen. Our results can be used to predict necessary experiment sizes, which are most importantly dictated by the width of the plasmid library. Our R package `gscreend` takes into account the asymmetry of the null distribution and improves phenotype detection at reduced experiment size.

Methods

Experimental datasets

The following datasets from published CRISPR knockout screens were used: screen in KBM7 cells [1] (Fig. 1); screen with TKO library in HCT116 cells, time points T08, T15, and T18 [2] (Figs. 1, 2, 4); screen in mouse ESC using mouse genome-wide libraries V1 and V2 [27] (Fig. 3). gRNA count data from the DepMap project [40, 43] was downloaded together with a dataset of cell doubling times [3] (Fig. 3f, 4i, Additional file 1: Figure S2). Data from project Score was downloaded from the project data repository [41] (Fig. 3f).

Data from library and T0 sequencing used in Additional file 1: Figure S1 was collected during a CRISPR screen in HCT116 and RKO cells. The 90k Toronto human Knockout pooled library (TKO) was a gift from Dr. Jason Moffat (1000000069, Addgene). Plasmid library was amplified using ElectroMAXTM Stbl4TM cells (Invitrogen) accordingly to the manufacturer's protocol. Library vector was transfected into HEK293T cells (ATCC) with TransIT-LT1 (Mirus Bio) transfection reagent along with psPAX2 (12260, Addgene) and pMD2.G (12259, Addgene) packaging plasmids to produce lentivirus. HCT116 and RKO cells (ATCC) stably expressing Cas9 (73310, Addgene) were infected in the presence of 8 μg/ml polybrene (Merck Millipore) with the 90k TKO gRNAs library at a multiplicity of infection (MOI) equal to 0.3 such that each gRNA was present in 500 cells on average. The day after, puromycin-containing medium was added to the infected cells for 48 h. On day 4 after transduction, a portion of cells were harvested as T0 time point. Genomic DNA from cell pellets was extracted using QIAamp DNA Blood Maxi kit (Qiagen). To amplify the gRNA sequences, a total of 140 PCR reactions were performed using 1 μg of genomic or plasmid library DNA each (250-fold coverage), Q5 Hot Start HF polymerase (NEB), and primers harboring the Illumina TruSeq adapter sequences. PCR products were purified using DNA Clean and Concentrator TM-100 (Zymo Research) and MagSi-NGSprep Plus beads (Steinbrenner). Sample concentrations were measured using Qubit HS DNA Assay (Thermo Fisher). Library amplicon size was verified using DNA High Sensitivity Assay on a BioAnalyzer 2100 (Agilent) and then sequenced on a NextSeq (Illumina) by 75 bp single-end sequencing and addition of 25% PhiX control v3 (Illumina). gRNAs were counted using the count function with automatic sequence trimming provided by MAGeCK [30].

Simulation of pooled CRISPR screens

We simulate a complete pooled CRISPR-knockout screen, providing output files that represent gRNA counts after sequencing of the plasmid library and T0 and T1 samples (see also Fig. 2a). The simulation depends on several parameters that reflect the experimental setup (see Table 1).

In a first step, the abundance $n_{\text{lib},g}$ of every gRNA g (where $g = 1, \ldots, N_{\text{tot}}$) in the plasmid library is sampled from a lognormal distribution $LN(\mu, \sigma)$, where $\mu = 5$ and σ is chosen to match the user-specified library distribution width L. We chose $\mu = 5$ because resulting distributions resemble those seen in experimental data. The sequencing counts $n^{\text{seq}}_{\text{lib},g}$ are obtained by making $C_{\text{PCR}}N_{\text{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\text{lib},g} / \sum_g n_{\text{lib},g}$. This is repeated N_{libpcr} times, to model the technical replicates.

In the next step, the abundance of gRNAs in the transduced cell pool $n_{\text{trans},g}$ is obtained by making $C_{\text{PCR}}N_{\text{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\text{lib},g} / \sum_g n_{\text{lib},g}$.

The pool of gRNAs of total size N_{tot} is partitioned into three sets: gRNAs without effect on cell fitness (Γ_{neutral}), gRNAs increasing cell fitness (Γ_{pos}), and gRNAs decreasing cell fitness (Γ_{neg}). The sets Γ_{neutral}, Γ_{pos}, and Γ_{neg} have respective sizes N_{neutral}, N_{pos}, and N_{neg} such that $N_{\text{neg}} = \phi_{\text{neg}} N_{\text{tot}}$, $N_{\text{pos}} = \phi_{\text{pos}} N_{\text{tot}}$ and $N_{\text{neg}} + N_{\text{pos}} + N_{\text{neutal}} = N_{\text{tot}}$. gRNAs from the different categories are assigned to essential, non-essential, or growth-suppressing genes according to N_{gRNA}.

In general, the cell proliferation-induced change in abundance of gRNA g between times t and $t + \Delta_t$ can be modeled as $n_g(t + \Delta_t) = e^{\beta} n_g(t)$, where β is the baseline cellular growth factor between two splittings and Δ_t the time between two splittings. β for a specific cell doubling time τ can thus be calculated as $\beta = \log\left(2^{\frac{\Delta_t}{\tau}}\right)$.

A gRNA specific growth rate β_g is then derived from β_{baseline} such that:

$\beta_g = \beta$ for every $g \in \Gamma_{\text{neutral}}$,
$\beta_g = \beta(1 + \epsilon)$ for every $g \in \Gamma_{\text{pos}}$,
$\beta_g = \beta(1 - \epsilon)$ for every $g \in \Gamma_{\text{neg}}$,

where ϵ is randomly chosen from 0, 0.01, 0.02, ... ,0.2.

The gRNA abundances at time t0 are calculated from the abundances in the transduced cell pool as:

$n_{\text{t0},g} = e^{\beta_g} n_{\text{trans},g}$ (real numbers are converted to integer by only taking the integer part).

The sequencing counts from T0 $n^{\text{seq}}_{\text{t0},g}$ are obtained by making $C_{\text{PCR}}N_{\text{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\text{t0},g} / \sum_g n_{\text{t0},g}$. This is repeated N_{biopcr} times, to model the technical replicates.

Next, the proliferation phase of the screen is simulated N_{bio} independent times to model the biological replicates. For $i = 1 \ldots N_{\text{split}}$, the gRNA abundances after cell splitting $n_{\text{i,split},g}$ are obtained by making $C_{\text{cells}}N_{\text{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\text{i},g}/\sum_g n_{\text{i},g}$. This random sampling step is followed by an exponential growth step $n_{\text{i+1},g} = e^{\beta_g} n_{\text{i,split},g}$. After completion of all cell splittings, the gRNAs in all biological replicates (time point T1) are sequenced by making $C_{\text{PCR}}N_{\text{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{N_{\text{split}},g}/\sum_g n_{N_{\text{split}},g}$. This is repeated N_{biopcr} times, to model the technical replicates.

For the analyses shown in Figs. 2b–e, 3a–c, and 4f–h, C_{PCR} and C_{virus} are chosen as indicated in the following table. The values of C_{PCR} and C_{virus} are chosen so that the 10% percentile of low abundant gRNAs in the library have a coverage of 100 fold.

Library width	C_{PCR} and C_{virus}
2.5	120
5	250
7.5	380
10	500
17	850

Normalization and LFC calculation

Counts from experimental data were normalized using size normalization to the total read counts of the reference sample. This was not necessary for simulated datasets, because these already had the same read counts. For a given gRNA with count n_{lib} in the library and n_1 at time point T1, the log fold change was calculated as $\text{LFC} = \log2\left(\frac{n_1+1}{n_{lib}+1}\right)$. Pseudo-counts had to be added for division and log transformation since some of the low abundant gRNAs had 0 counts in one or more of the replicates.

Library width calculation

The width of a distribution of gRNA abundances can be quantified by calculating the ratio between the 90 and 10% percentile of the distribution [23]: $\text{library width} = \frac{\text{percentile}_{90}(n_{\text{lib,g}})}{\text{percentile}_{10}(n_{\text{lib,g}})}$. $n_{\text{lib,g}}$ is the distribution of gRNA abundances in the library.

gRNA sequence composition

The sequence probability logos in Additional file 1: Figure S4 were generated using the output of the plogo online tool [46] and R. gRNAs were ranked according to their abundance and the sequence logos generated for the lower and upper 1% and 5% of gRNAs.

`gscreend` method

`gscreend` is designed to account for asymmetric distribution of before/after ratios in pooled genetic perturbation screens (see also Fig. 4a).

`gscreend` takes (non-normalized) gRNA counts from several samples as its input. One of these is the reference sample (e.g., the library or T0); the others are one or several replicates of a post-screen time point (e.g. T1). The counts are scaled (a.k.a. normalized) to the total counts of the reference sample. Log2 fold changes are calculated as described above. The data are split into slices according to the gRNA abundance in the reference sample; the current implementation uses 10 slices split at the 10%, 20%, …quantiles. We use this stratification because the null distributions of the fold changes depend on it and are fit separately in each stratum.

We model the overall LFC data as a mixture of a parametric null distribution, the skew normal, and an unspecified alternative distribution [47, 48]. The first mixture component corresponds to gRNAs without fitness effect, the second to those with effect, and we will assume that these are only a minority. We use the R package fGarch for computations involving the skew normal distribution and use least quantile of square regression [49] on a 10–90% percentile of the log-likelihood to infer the model parameters from the distribution of LFCs (function lbfgs from R package nloptr).

In the next step, for every stratum, p values are calculated for every gRNA. The gRNAs are ranked based on their p values (if there are multiple replicates, each gRNA gets as many ranks). On this ranking, `gscreend` uses an α-RRA (robust rank aggregation) algorithm with an α cutoff of 5% to aggregate the data to the gene level [30, 50]. Gene level LFCs are calculated by averaging the LFCs over all gRNAs belonging to the gene.

Comparison of analysis tools

Results from the analysis of simulated and experimental data using the different analysis tools were compared as follows:

- `gscreend` analysis was performed with 10–90% percentile for least quantile of square method and 5% threshold for α-RRA algorithm. Genes were ranked according to the p value and for genes with the same p value according to their mean LFC over all corresponding gRNAs.
- MAGeCK [30] analysis was performed using the RRA algorithm, without normalization to controls. Genes were ranked according to the rank provided by MAGeCK.
- BAGEL [33] analysis was performed only on experimental data because the algorithms need a list of essential and non-essential genes as training sets. Creating this type of list on a set of simulated data would be arbitrary, since its quality cannot be compared to the currently available lists of essential and non-essential genes. BAGEL analysis was run

without removal of low counts. Ranking of genes was performed based on Bayes factors.

- ScreenBEAM [31] analysis was performed without removal of low counts. Genes were ranked according to *p* values.
- CRISPhieRmix [35] analysis was performed according to the software default settings. The packages take LFC data, which was calculated as described in the above LFC calculation section. Genes were ranked based on the genescore returned by the CRISPhieRmix analysis.
- CRISPRBetaBinomial [37] results were ranked according to the fdr_pa parameter, which corresponds to enrichment in the after-screen time point at the gene level. For genes with identical fdr_pa, rank was attributed according to LFC.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-1939-1.

Additional file 1: Table S1 and Figures S1–S4.

Additional file 2: Review history.

Acknowledgements

We would like to thank Sabine Ottinger for the help with library preparation and acknowledge the support provided by the Genomics Core Facility of the European Molecular Biology Laboratory (Heidelberg, Germany).

Peer review information

Yixin Yao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 2.

Authors' contributions

KI implemented the model and performed the analyses. GA performed the screening experiment. MB and WH designed and supervised the research. KI and WH wrote the paper. All authors read and approved the final manuscript.

Authors' information

Twitter handles: @K_Imkeller (Katharina Imkeller), @Michael_Boutros (Michael Boutros), @wolfgangkhuber (Wolfgang Huber).

Funding

This work was supported by the NCT 3.0 Integrative Projects in Basic Cancer Research Program and the DFG Sonderforschungsbereich 1324.

Availability of data and materials

The simulation is available under the GPL-3.0 license at https://github.com/imkeller/simulate_pooled_screen [54].
`gscreend` is available under the GPL-3.0 license as part of Bioconductor and at https://github.com/imkeller/gscreend [55].

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]German Cancer Research Center (DKFZ) and Heidelberg University, Heidelberg, Germany. [2]European Molecular Biology Laboratory, Heidelberg, Germany.

Received: 18 August 2019 Accepted: 19 January 2020
Published online: 02 March 2020

References

1. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096–101. https://doi.org/10.1126/science.aac7041.
2. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell. 2015;163(6):1515–26. https://doi.org/10.1016/j.cell.2015.11.015.
3. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a cancer dependency map. Cell. 2017;170(3):564–57616. https://doi.org/10.1016/J.CELL.2017.06.010.
4. Wang T, Yu H, Hughes NW, Liu B, Kendirli A, Klein K, Chen WW, Lander ES, Sabatini DM. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. Cell. 2017;168(5): 890–90315. https://doi.org/10.1016/j.cell.2017.01.013.
5. Steinhart Z, Pavlovic Z, Chandrashekhar M, Hart T, Wang X, Zhang X, Robitaille M, Brown KR, Jaksani S, Overmeer R, Boj SF, Adams J, Pan J, Clevers H, Sidhu S, Moffat J, Angers S. Genome-wide CRISPR screens reveal a Wnt–FZD5 signaling circuit as a druggable vulnerability of RNF43-mutant pancreatic tumors. Nat Med. 2017;23(1):60–8. https://doi.org/10.1038/nm.4219.
6. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343(6166):80–4. https://doi.org/10.1126/SCIENCE.1246981.
7. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343(6166):84–7. https://doi.org/10.1126/science.1247005.
8. 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(6):661–7. https://doi.org/10.1038/nbt.3235.
9. Kim HS, Lee K, Bae S, Park J, Lee C-K, Kim M, Kim E, Kim M, Kim S, Kim C, Kim J-S. CRISPR/Cas9-mediated gene knockout screens and target identification via whole-genome sequencing uncover host genes required for picornavirus infection. J Biol Chem. 2017;292(25):10664–71. https://doi.org/10.1074/jbc.M117.782425.
10. Han J, Perez JT, Chen C, Li Y, Benitez A, Kandasamy M, Lee Y, Andrade J, TenOever B, Manicassamy B. Genome-wide CRISPR/Cas9 screen identifies host factors essential for influenza virus replication. Cell Rep. 2018;23(2):596–607. https://doi.org/10.1016/j.celrep.2018.03.045.
11. 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(3):267–73. https://doi.org/10.1038/nbt.2800.
12. Charton K, Suel L, Henriques SF, Moussu J-P, Bovolenta M, Taillepierre M, Becker C, Lipson K, Richard I. Exploiting the CRISPR/Cas9 system to study alternative splicing in vivo: application to titin. Hum Mol Genet. 2016;25(20):280. https://doi.org/10.1093/hmg/ddw280.
13. Gapinske M, Luu A, Winter J, Woods WS, Kostan KA, Shiva N, Song JS, Perez-Pinera P. CRISPR-SKIP: programmable gene splicing with single base editors. Genome Biol. 2018;19(1):107. https://doi.org/10.1186/s13059-018-1482-5.
14. 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;5:19760. https://doi.org/10.7554/eLife.19760.
15. Hilton IB, D'Ippolito AM, Vockley CM, Thakore PI, Crawford GE, Reddy TE, Gersbach CA. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol. 2015;33(5):510–7. https://doi.org/10.1038/nbt.3199.

16. Liu XS, Wu H, Ji X, Stelzer Y, Wu X, Czauderna S, Shu J, Dadon D, Young RA, Jaenisch R. Editing DNA methylation in the mammalian genome. Cell. 2016;167(1):233–24717. https://doi.org/10.1016/J.CELL.2016.08.056.
17. Morita S, Noguchi H, Horii T, Nakabayashi K, Kimura M, Okamura K, Sakai A, Nakashima H, Hata K, Nakashima K, Hatada I. Targeted DNA demethylation in vivo using dCas9–peptide repeat and scFv–TET1 catalytic domain fusions. Nat Biotechnol. 2016;34(10):1060–5. https://doi.org/10.1038/nbt.3658.
18. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, Qi LS, Kampmann M, Weissman JS. Genome-scale CRISPR-mediated control of gene repression and activation. Cell. 2014;159(3):647–61. https://doi.org/10.1016/J.CELL.2014.09.029.
19. Konermann S, Brigham MD, Trevino AE, Joung J, Abudayyeh OO, Barcena C, Hsu PD, Habib N, Gootenberg JS, Nishimasu H, Nureki O, Zhang F. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature. 2015;517(7536):583–8. https://doi.org/10.1038/nature14136.
20. Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, P R Iyer E, Lin S, Kiani S, Guzman CD, Wiegand DJ, Ter-Ovanesyan D, Braff JL, Davidsohn N, Housden BE, Perrimon N, Weiss R, Aach J, Collins JJ, Church GM. Highly efficient Cas9-mediated transcriptional programming. Nat Methods. 2015;12(4):326–8. https://doi.org/10.1038/nmeth.3312.
21. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, Mochizuki M, Miyabe A, Araki M, Hara KY, Shimatani Z, Kondo A. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science. 2016;353(6305):8729. https://doi.org/10.1126/science.aaf8729.
22. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016;533(7603):420–4. https://doi.org/10.1038/nature17946.
23. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017;12(4):828–63. https://doi.org/10.1038/nprot.2017.016.
24. Doench JG. Am I ready for CRISPR? A user's guide to genetic screens. Nat Rev Genet. 2018;19(2):67–80. https://doi.org/10.1038/nrg.2017.97.
25. 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(7):733. https://doi.org/10.15252/msb.20145216.
26. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014;11(8):783–4. https://doi.org/10.1038/nmeth.3047.
27. Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, Mupo A, Grinkevich V, Li M, Mazan M, Gozdecka M, Ohnishi S, Cooper J, Patel M, McKerrell T, Chen B, Domingues AF, Gallipoli P, Teichmann S, Ponstingl H, McDermott U, Saez-Rodriguez J, Huntly BJP, Iorio F, Pina C, Vassiliou GS, Yusa K. A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. Cell Rep. 2016;17(4):1193–205. https://doi.org/10.1016/j.celrep.2016.09.079.
28. Aregger M, Chandrashekhar M, Tong AHY, Chan K, Moffat J. Pooled lentiviral CRISPR-Cas9 screens for functional genomics in mammalian cells. In: Methods in Molecular Biology, vol 1869. New York: Humana Press; 2019. p. 169–88. https://doi.org/10.1007/978-1-4939-8805-1_15.
29. Nagy T, Kampmann M. CRISPulator: A discrete simulation tool for pooled genetic screens. BMC Bioinformatics. 2017;18(1):1–12. https://doi.org/10.1186/s12859-017-1759-9.
30. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15(12):554. https://doi.org/10.1186/s13059-014-0554-4.
31. Yu J, Silva J, Califano A. ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. Bioinformatics. 2015;32(2):556. https://doi.org/10.1093/bioinformatics/btv556.
32. Diaz AA, Qin H, Ramalho-Santos M, Song JS. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. Nucleic Acids Res. 2015;43(3):16. https://doi.org/10.1093/nar/gku1197.
33. Hart T, Moffat J. BAGEL: A computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics. 2016;17(1):1–7. https://doi.org/10.1186/s12859-016-1015-8.
34. Jia G, Wang X, Xiao G. A permutation-based non-parametric analysis of CRISPR screen data. BMC Genomics. 2017;18(1):545. https://doi.org/10.1186/s12864-017-3938-5.
35. Daley TP, Lin Z, Lin X, Liu Y, Wong WH, Qi LS. CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. Genome Biol. 2018;19(1):159. https://doi.org/10.1186/s13059-018-1538-6.
36. Allen F, Khodak A, Behan F, Iorio F, Yusa K, Garnett M, Parts L. JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Res. 2019;29:464–71. https://doi.org/10.1101/gr.238923.118.
37. Jeong H-H, Kim SY, Rousseaux MWC, Zoghbi HY, Liu Z. Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives. Genome Res. 2019;29(6):999–1008. https://doi.org/10.1101/gr.245571.118.
38. Zhou Y, Zhu S, Cai C, Yuan P, Li C, Huang Y, Wei W. High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature. 2014;509(7501):487–91. https://doi.org/10.1038/nature13166.
39. Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, Shalem O, Satija R, Raychowdhury R, Mertins P, Carr SA, Zhang F, Hacohen N, Regev A. A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. Cell. 2015;162(3):675–86. https://doi.org/10.1016/J.CELL.2015.06.059.
40. DepMap; Broad. DepMap Achilles 19Q1 Public. figshare. 2019;Fileset:. https://doi.org/10.6084/m9.figshare.7655150.
41. Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, Santos R, Rao Y, Sassi F, Pinnelli M, Ansari R, Harper S, Jackson DA, McRae R, Pooley R, Wilkinson P, van der Meer D, Dow D, Buser-Doepner C, Bertotti A, Trusolino L, Stronach EA, Saez-Rodriguez J, Yusa K, Garnett MJ. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature. 2019;568(7753):511–6. https://doi.org/10.1038/s41586-019-1103-9.
42. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Costanzo M, Myers CL, Andrews BJ, Boone C, Durocher D, Moffat J. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3 Genes|Genomes|Genetics. 2017;7(8):2719–27. https://doi.org/10.1534/g3.117.041277.
43. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC, Tsherniak A. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84. https://doi.org/10.1038/ng.3984.
44. Ong SH, Li Y, Koike-Yusa H, Yusa K. Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. Sci Rep. 2017;7(1):1–10. https://doi.org/10.1038/s41598-017-07827-z.
45. Sidik SM, Huet D, Ganesan SM, Huynh MH, Wang T, Nasamu AS, Thiru P, Saeij JPJ, Carruthers VB, Niles JC, Lourido S. A genome-wide CRISPR screen in toxoplasma identifies essential apicomplexan genes. Cell. 2016;166(6):1423–143512. https://doi.org/10.1016/j.cell.2016.08.019.
46. O'Shea J. P, Chou MF, Quader SA, Ryan JK, Church GM, Schwartz D. PLogo: a probabilistic approach to visualizing sequence motifs. Nat Methods. 2013;10(12):1211–2. https://doi.org/10.1038/nmeth.2646.
47. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc. 2004;99:96–104.
48. Strimmer K. A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008;9(1):303. https://doi.org/10.1186/1471-2105-9-303.
49. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. Wiley Ser Probab Stat; 1987, p. 329. https://doi.org/10.1002/0471725382.
50. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28(4):573–80. https://doi.org/10.1093/bioinformatics/btr709.
51. Shifrut E, Carnevale J, Tobin V, Roth TL, Woo JM, Bui CT, Li PJ, Diolaiti ME, Ashworth A, Marson A. Genome-wide CRISPR screens in primary

human T cells reveal key regulators of immune function. Cell. 2018;175(7): 1958–197115. https://doi.org/10.1016/j.cell.2018.10.024.

52. Wegner M, Diehl V, Bittl V, de Bruyn R, Wiechmann S, Matthess Y, Hebel M, Hayes MG, Schaubeck S, Benner C, Heinz S, Bremm A, Dikic I, Ernst A, Kaulich M. Circular synthesized CRISPR/Cas gRNAs for functional interrogations in the coding and noncoding genome. eLife. 2019;8:. https://doi.org/10.7554/eLife.42549.
53. Chen C-H, Xiao T, Xu H, Jiang P, Meyer CA, Li W, Brown M, Liu XS. Improved design and analysis of CRISPR knockout screens. Bioinformatics (June). 20181–7. https://doi.org/10.1093/bioinformatics/bty450.
54. Imkeller K. Simulation of pooled screens. Github. 2019. https://github.com/imkeller/simulate_pooled_screen. Accessed 31 Jan 2020.
55. Imkeller K, Huber W. gscreend - analysis of pooled CRISPR screens. Bioconductor. 2019. http://bioconductor.org/s/gscreend. Accessed 31 Jan 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.