Modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection

Katharina Imkeller [1,2,3], Giulia Ambrosi [1], Michael Boutros [1], Wolfgang Huber [2,3]

[1]German Cancer Research Center (DKFZ) and Heidelberg University, Heidelberg, Germany;
[2]European Molecular Biology Laboratory, Heidelberg, Germany;
[3]Corresponding authors. Email: k.imkeller@dkfz.de (K.I.), wolfgang.huber@embl.de (W.H.)

Abstract

Pooled CRISPR screens are a powerful tool to probe genotype-phenotype relationships at genome-wide scale. They are based on a library of target-specific guide RNAs (gRNAs) that is applied to a pool of cells, with the aim to induce a single genetic perturbation in each cell. Detection of viability phenotypes depends on statistical comparison of the frequencies of these gRNAs before and after cell proliferation. Here, we report empirical and theoretical evidence for asymmetries in gRNA count ratios, and we show how failure to consider asymmetric null distribution of the data leads to loss of detection power. We present a biology-based, generative probabilistic model to show that the asymmetry is generated during the proliferation phase of the screen and that it is mainly driven by the distribution width of the gRNA library. Based on these observations we develop a statistical test that improves hit detection at reduced experiment size. The method is implemented in the open source package `gscreend` (submission to Bioconductor pending).

Background

Pooled genetic perturbation screens are a powerful tool to systematically probe gene function as well as genotype-phenotype relationships in many different cell types. Their applications range from identification of genotype-specific vulnerabilities in human cancer cells[1–3] to discovery of genes involved in drug resistance[5–8] or virus replication[9,10].

The currently most widely employed 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)[5,11,12], which induces DNA cleavage at the genomic site it is directed to. Subsequent DNA repair via non-homologous end joining eventually leads to frame-shift mutations and premature stop-codons, nonsense-mediated RNA decay and finally gene knockout. Alternatively, it is possible to introduce more subtle perturbations such as altered splicing patterns[13,14] or quantitative modulation of gene expression[15]. To this end, modified CRISPR associated enzymes which function as epigenetic modifiers[16–18], transcriptional modulators[19–21] or single-base editors[22,23] are used.

Pooled screens enable measuring the effect of many genetic perturbations in parallel in a single experiment. To this end, a large library of gRNAs is introduced into a pool of cells at a low multiplicity of infection such that no more than one perturbation is present in each cell[24,25]. The gRNA sequence simultaneously serves as a barcode that can be used to trace which perturbation the cell carries. In the case of negative selection screens, the transduced cell pool is let grow for several divisions in the course of which the relative abundance of cells with a particular gRNA increases or decreases depending on the targeted gene's role for cell fitness. These effects are detected by amplifying, sequencing and 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[2,5,11,12,26,27]. To ensure statistical power, each gRNA needs to be represented at sufficient numbers during each step of the screen. When designing screening experiments it is commonly assumed that the abundances of the different gRNAs in the library are distributed evenly, i.e. all gRNA are present in the library at approximately same relative frequencies. Hence, the distribution is summarized to the mean gRNA abundance, also referred to as coverage or representation, and this measure is used to calculate the necessary size of an experiment[24,25]. For a library of 100,000 gRNAs and a desired coverage of 500 for example, 50 million cells (500 times 100,000) have to be kept in culture. Published recommendations on optimal coverage are poorly justified and diverge from 200[28] to 500[24]. Optimizing this choice is a major thrust of this work, since it has substantial consequences on the size and costs of an experiment.

To compare the gRNA abundances before and after the proliferation phase, a range of statistical models and computational tools are available[29–34]. 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 univariate summary statistics, the ratios of these counts, hereafter referred to as "before/after ratios". gRNAs whose data fall sufficiently outside the null distribution present evidence of a fitness effect of their target gene. Since usually each gene is targeted by multiple gRNAs, a last step of the analysis consists of aggregating gRNA-level evidence to the gene level. This can be achieved for example using Bayesian hierarchical modelling[30,34] or robust rank aggregation[29].

For the null distribution modelling and hypothesis testing, approaches derived from RNA sequencing and differential gene expression analysis have been used[35,36]. However, careful inspection of datasets shows that the distribution of the before/after ratios for negative controls is often asymmetric[37], which is not the case in RNA sequencing. Such asymmetry means that even in the absence of a fitness effect, a gRNA's relative abundance is more likely to randomly decrease rather than to increase during the screen. Failing to account for such asymmetry (as is done when using the RNA-seq based tools) leads to needlessly elevated false positive rates and 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 knockout screens conducted in human cell lines[1,2] (Fig. 1B-G). After 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 Methods). 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 not essential for cell fitness according to a study by Hart et al.[38] (Fig. 1E-G[2]). The sign of their LFCs was not reproducible between replicates, in agreement with the assumption that the observed LFCs were due to random variability and not due to target-dependent fitness effects (Fig. 1D and G). However, their distribution 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.[37].

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 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 12500 genes with 4 independent gRNAs per gene). For 10% of the genes the knockout leads to 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)[1,2,27,39], and it is unclear to what extent this choice influences the analysis outcome. Timepoint T0 is after the antibiotics selection of cells that were successfully transduced, which means one 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 (Fig. 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[38] had reduced abundance at T0 (Fig. S1B). These results strongly argue to use plasmid library rather than T0 sequencing 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 asymmetric 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 reflects in the number of cells that are seeded after every round of splitting. We found that, the smaller this cell splitting coverage is, the more pronounced the asymmetry of before/after ratios becomes (Fig. 2B-C). This higher level of asymmetry also leads to impairment of phenotype detection by MAGeCK-RRA[29], a current state-of-the-art analysis tool, which loses more than 10% of recall at 95% precision when e.g. reducing the cell splitting coverage from 400 to 100 (Fig. 2D). Similarly, the observed asymmetry increases when using faster growing cell lines, as we were able to confirm in our simulations (Fig. 2E) and in experimental datasets (Fig. S2)[39,40]. We also tested the effect of other parameters like coverage during transduction or PCR, but they only marginally

influenced the observed asymmetry of before/after ratios in our simulations (Fig. 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 only reached 3% (Fig. 2C and S3).

We thus conclude that transduction and PCR do not cause major technical bias in the data and that it is better to sequence the gRNA pool in the plasmid library rather then 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 to the mean gRNA coverage. Indeed we were able to confirm the gradual broadening of the abundance distribution of gRNA targeting non-essential genes[38] 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. The width of this distribution, i.e. the difference in abundance between low and high abundant gRNAs, can be quantified by calculating the ratio between the 90% and 10% percentiles. This ratio, elsewhere also named "skew ratio"[24], will hereafter be referred to as "distribution width". If for example the most abundant 10% gRNAs of a library have an abundance higher than 500 whereas the least abundant 10% have less than 100 counts, the library distribution width is 5.

We performed simulations starting from three different gRNA libraries with varying distribution width (Fig. 3A). Our simulations confirm that with wider gRNA abundance distribution in the plasmid library the reproducibility between replicates decreases (Fig. 3B) and at the same time the frequency of gRNAs with LFC<-1 increases (Fig. 3C). Experimental data from a screen conducted using two plasmid libraries with different distribution widths confirms this finding (Fig. 3D-E)[27]. 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 (Fig. S4). To show this we selected five datasets from published CRISPR libraries[1,27,41,42]. These libraries have different distribution widths ranging from 2.4 to 8.8 (Fig. S4A-B). To examine the sequence composition we generated probability sequence motifs for the least and most abundant gRNAs (Fig. S4C-D)[43]. Wider libraries tend to have poly-G-stretches in low abundant gRNAs and many T nucleotides in high abundant gRNAs, especially in the second half of the sequence (Fig. S4E). 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. Since it is not always possible to use plasmid libraries with low distribution width or perform screens at very high coverage, 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 Figure 4A. After scaling of gRNA counts and calculation of LFCs, the data is split into slices according the gRNA abundance in the reference sample (e.g. plasmid library). We introduce this stratification because the parameters of the null distribution depend on it. We model the overall LFC data as a mixture of a parametric null distribution, the skew normal, and an unspecified alternative distribution[44,45]. The first mixture component corresponds to gRNAs without fitness effect, the second to those with effect, and we assume that these are only a minority. `gscreend` uses least quantile of squares regression[46] to fit the null distribution to the LFCs in every data interval. Least quantile of squares regression fits a model to an empirical distribution by only taking into account a defined proportion of residuals, e.g. between 10% and 90% percentiles, thus not considering the "outliers". In the `gscreend` workflow the resulting null-models for every interval

are used to calculate p-values, which are then employed to rank individual gRNAs. In the last step, robust rank aggregation[29,47] is applied to aggregate the ranked gRNA list to the gene level.

We first tested the performance of `gscreend` in accurately ranking genes in experimental datasets. Using the published list of essential and non-essential genes[38] we calculate the recall at 99% precision (as in Fig. 2D) of our and other published tools[29,30,32]. `gscreend` outperforms MAGeCK and screenBeam when ranking genes in a CRISPR-knockout screen performed in HCT116 cells (Fig. 4B)[2]. BAGEL is the only tool that has a better precision-recall performance than `gscreend`. However, its algorithm is trained on the sets of essential and non-essential genes that we also use to calculate precision-recall statistics, which might explain the apparent superior performance. Indeed when ranking components of the ribosome, whose knockout is likely to be lethal, `gscreend` outperforms BAGEL, MAGeCK and screenBeam, especially within the 1000 lowest ranked genes (Fig. 4C). In order to illustrate in which cases `gscreend` performs better than other tools that do not take into consideration the asymmetric null distribution, we highlighted the results for five selected genes (Fig. 4D+E). *MRPL34* and *MRPS12* (components of 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 on T18 throughout 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 artifacts, because they are not reproducible between replicates (Fig. 4E). Taken together, these results indicate that `gscreend` delivers an unprecedented accuracy in correctly ranking and identifying essential genes in pooled negative selection screens.

We next also investigated whether our method is robust against different levels of the previously discussed asymmetry of before/after ratios. Similar to what we found using the experimental data, `gscreend` has a better ranking accuracy than the other tools when detecting either essential or growth suppressing genes in simulated datasets (Fig. 4F+G). When increasing the asymmetry by either reducing cell splitting coverage or increasing library distribution width, our method maintains a better ranking accuracy than MAGeCK-RRA (Fig. 4H). `gscreend` allows to reduce the coverage by approximately 50% for a library distribution width of 7.5: using 300 (`gscreend`) instead of 600 (MAGeCK) mean gRNA coverage maintains at least 95% recall at 99% precision (Fig. 4H). For libraries with larger distribution widths the gain in accuracy is even more important.

Implications for the design of screening experiments

Based on our findings concerning the asymmetry of before/after ratios we suggest to first measure the distribution width of the gRNA plasmid library when designing a screen. Based on this measure, our simulations from Fig. 4D, left panel, can be used to predict at which coverage it is best to perform cell culture (summarized in Table 2). Since PCR and transduction only mildly affect data quality, the respective coverages should be chosen in the same range as the cell splitting coverage.

Discussion

Pooled genetic perturbation screens are widely used hypothesis generating tools in many biological disciplines. Accurately detecting phenotypes in these experiments is key to deciphering biological processes at high resolution. Screens that correctly distinguish all genes that negatively or positively regulate cell fitness can not only be used to discover the strongest gene candidates, but also to measure subtle differences in growth rate and thus identify mechanisms and potentially map whole pathways.

High data quality and accurate analysis is first of all achieved by understanding how the experimental design influences the results. Although first simulations of CRISPR-based screens have been published[48], our study is the first one to systematically explore the influence of experimental design on phenotype detection in pooled screens. We show that gRNA coverage during PCR and transduction, as long as chosen in the same range as the cell splitting coverage, only marginally influences data quality and that screens are best analysed when using plasmid library sequencing as reference. We do not address the influence of the multiplicity of infection during viral transduction, but there is literature and a good model discussing this point[25]. 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.

Using this knowledge on asymmetric null distribution of before/after ratios, we were able to develop a new statistical test that improves phenotype detection. `gscreend` outperforms other analysis methods, which assume that the null distribution is symmetric. From the point of view of screen design, our method allows to reduce experiment size by ca. 50% compared to other tools, because it maintains high analysis accuracy throughout a broad range of experimental settings. Especially for experiments that are limited in size, because for example they use primary cell cultures[36,49], our tool may help to improve phenotype detection.

Our results also provide clear indications on how to optimize the experimental design by choosing the screening coverage according to the width of the utilized 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 library may thus be a promising technology for substantial reduction of library width and experiment size[50].

Finnaly 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[27,51].

Conclusion

We conclude that the asymmetry of the before/after ratio distribution in pooled CRISPR screens is majorly 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 new R package `gscreend` takes into account the asymmetry of the null distribution and thus improves phenotype detection at reduced experiment size.

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	50000	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.

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

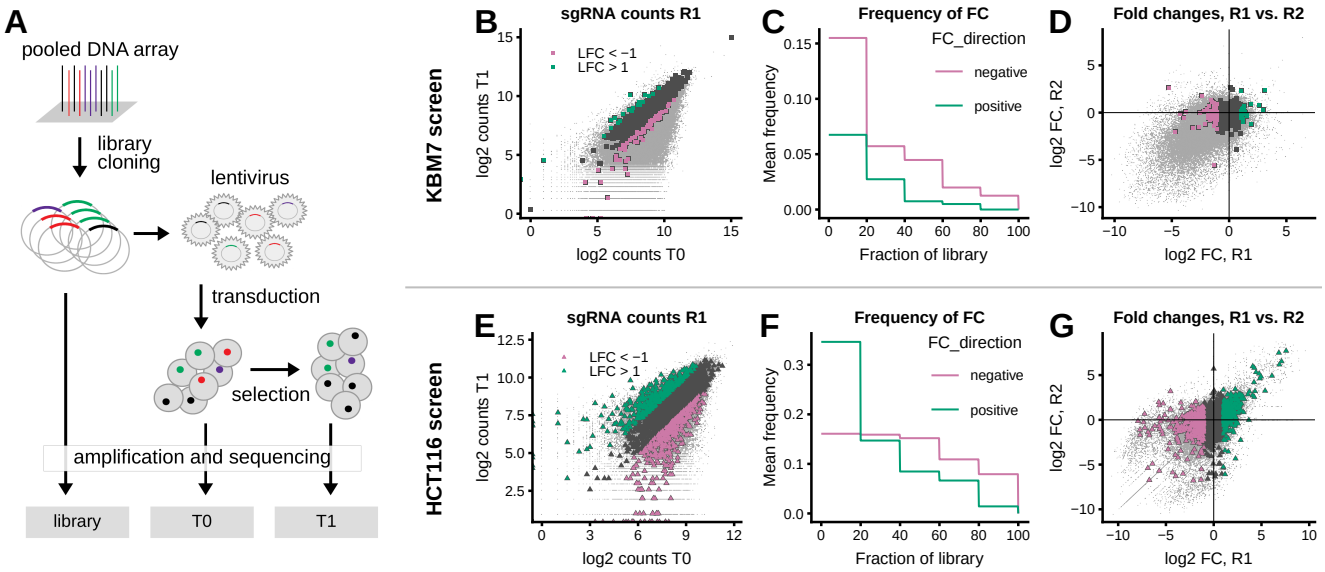


Figure 1: Screening data show asymmetric distribution of gRNA abundance fold changes. A: Screen setup for measurement of gRNA effect on cell fitness. B-D: KBM7 screen[1] with highlight on non-targeting controls. E-G: HCT116 screen[2] with highlight on gRNAs targeting non-essential genes (defined according to Hart et al.[38]). 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.

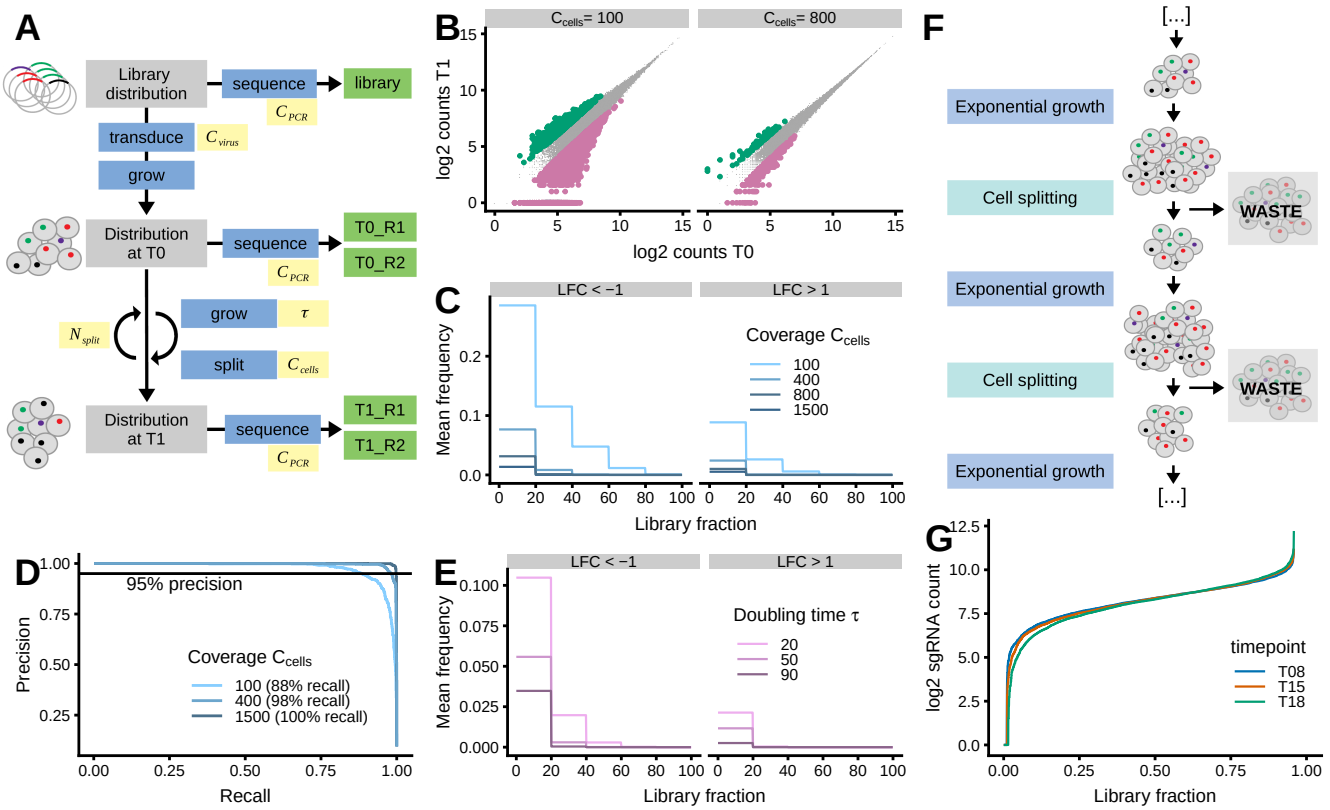


Figure 2: Computational simulation explains how cell splitting causes asymmetry of before/after ratios. A: Schematic representation of the simulation. After generation of an initial gRNA abundance distribution, different functions (blue) are applied to model transduction, cell growth, cell splitting and sequencing. The simulation outputs the gRNA counts obtained by sequencing the plasmid library as well as the cell pools at timepoints 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 hours (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.

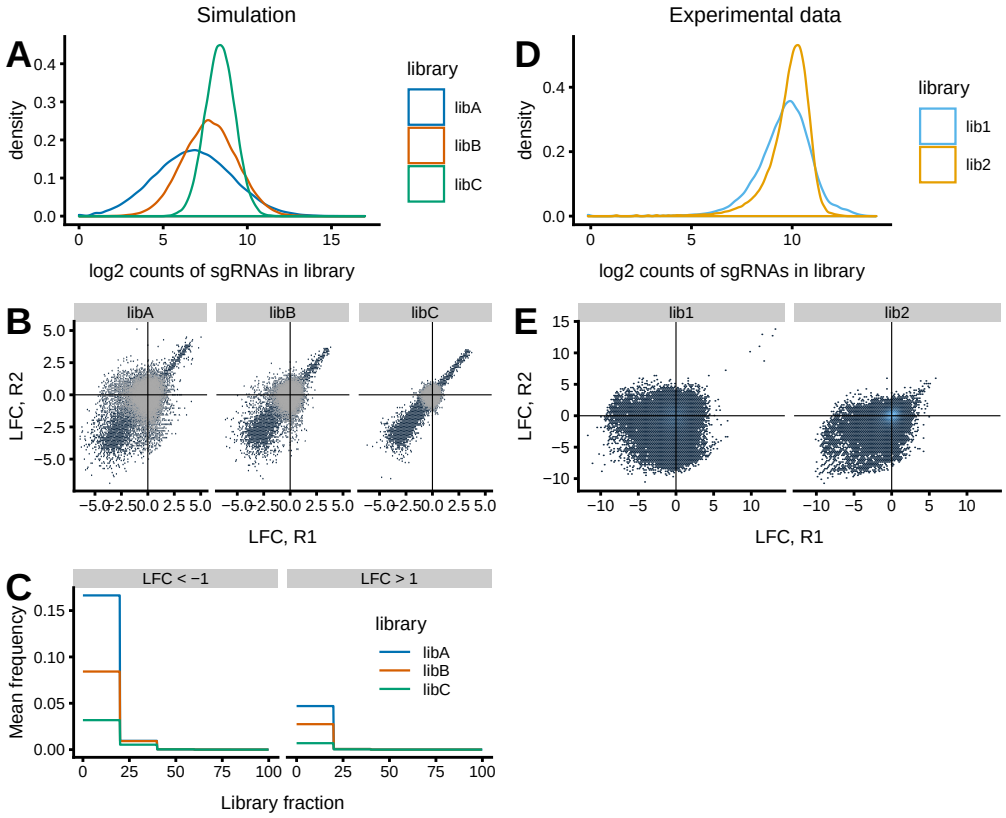


Figure 3: Wider gRNA abundance distributions in the plasmid library increase asymmetry of before/after ratios. A-C: Simulation results using three libraries with different widths of gRNA abundance distribution. A: log2 count distribution of simulated libraries libA, libB and libC. Distributions were generated as log-normal distribution with same log-mean, but differing log-sd and then size-normalized. Library distribution widths: 66.5 (libA), 17.8 (libB), 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, gRNAs without fitness effect are highlighted in grey. 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 V1 and V2 respectively from Tzelepis et al.[27]). Library distribution widths: 8.8 (lib1), 5.0 (lib2). D: log2 count distribution 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.

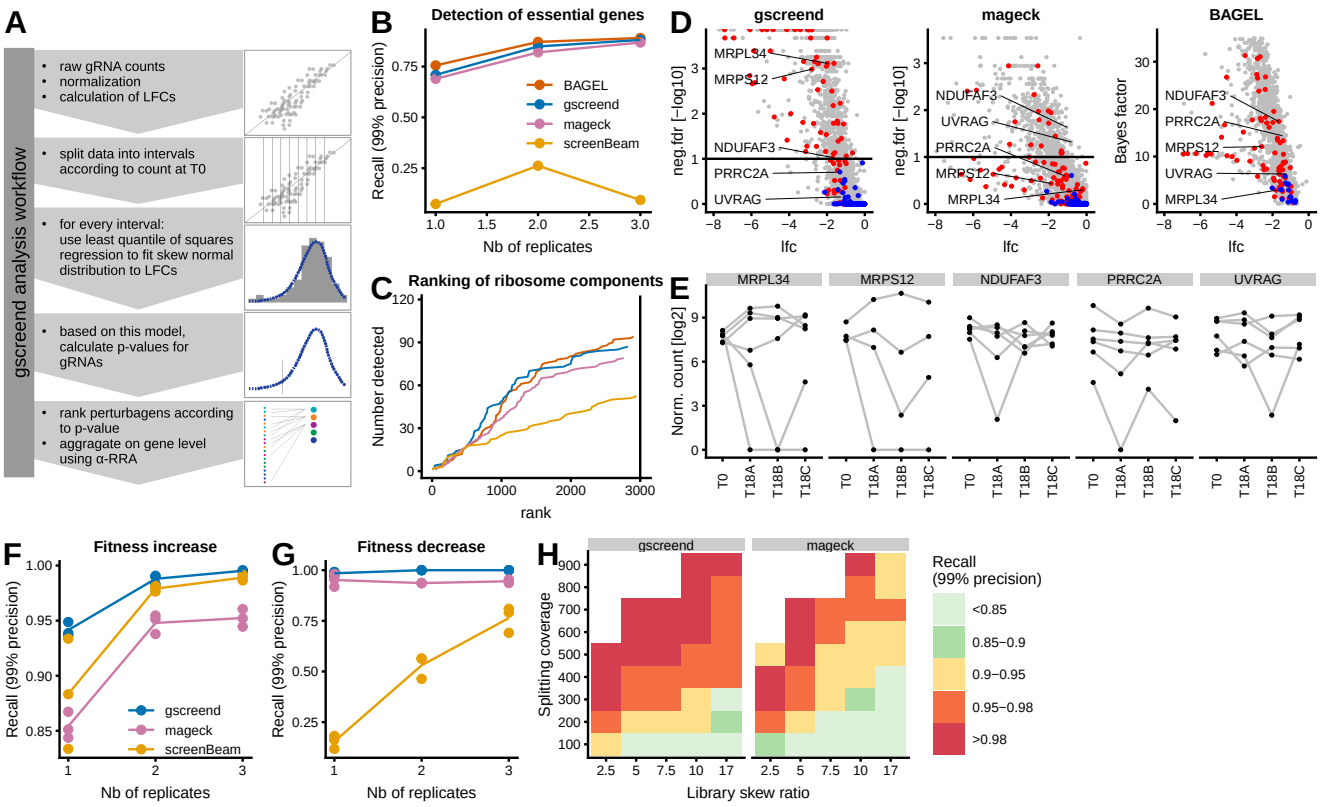


Figure 4: gscreend increases phenotype detection accuracy. A: gscreend analysis workflow. B-E: Comparison of gene ranking by BAGEL, MAGeCK, screenBeam and gscreend for CRISPR knockout screen performed in HCT116 cells[2]. B: Recall at 99% precision (as in Fig. 1D) for analysis of 1 to 3 biological replicates of timepoint T18. Essential and non-essential genes were defined according to Hart et al.[38]. C: Ranking of genes encoding ribosome components (GO: structural constituent of ribosome) by the four different analysis tools and using all 3 biological replicates. D: Volcano plots illustrating analysis results for genes encoding ribosomal components (red) and non-essential genes[38] (blue). Horizontal lines indicate FDR thresholds of 1%. E: Log2 normalized abundances of gRNAs targeting the five selected genes at timepoint T0 and three replicates of timepoint T18. F-G: Recall at 99% precision by 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 30h. 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).

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, timepoints T08, T15, T18, raw data file version 1[2] (Fig. 1, 2, 4); screen in mouse ESC using mouse genome wide libraries V1 and V2[27] (Fig. 3). gRNA counts data from the DepMap project[39,40] was downloaded together with a dataset of cell doubling times[3] (Fig. S2).

Data from library and T0 sequencing used in Fig. 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 presence of 8 μg/ml polybrene (Merck Millipore) with the 90k TKO gRNAs library at a 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 hours. On day four after transduction, a portion of cells were harvested as T0 timepoint. 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 1μg 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 & 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 75bp 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[29].

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 2).

In a first step, the abundance $n_{\mathsf{lib},g}$ of every gRNA g (where $g = 1, \ldots, N_{\mathsf{tot}}$) 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 $n^{\mathsf{seq}}_{\mathsf{lib},g}$ are obtained by making $C_{\mathsf{PCR}} N_{\mathsf{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\mathsf{lib},g} / \sum_g n_{\mathsf{lib},g}$. This is repeated N_{libpcr} times, to model the technical replicates.

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

The pool of gRNAs of total size N_{tot} is partitioned into three sets: gRNAs without effect on cell fitness ($\Gamma_{\mathsf{neutral}}$), gRNAs increasing cell fitness (Γ_{pos}) and gRNAs decreasing cell fitness (Γ_{neg}). The sets $\Gamma_{\mathsf{neutral}}$, Γ_{pos} and Γ_{neg} have respective sizes N_{neutral}, N_{pos} and N_{neg} such that $N_{\mathsf{neg}} = \phi_{\mathsf{neg}} N_{\mathsf{tot}}$, $N_{\mathsf{pos}} = \phi_{\mathsf{pos}} N_{\mathsf{tot}}$ and $N_{\mathsf{neg}} + N_{\mathsf{pos}} + N_{\mathsf{neutal}} = N_{\mathsf{tot}}$. gRNAs from the different categories are assigned to essential, non-essential or growthsupressing 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 $\beta = log(2^{\frac{\Delta_t}{\tau}})$.

A gRNA specific growth rate β_g is then derived from $\beta_{\mathsf{baseline}}$ such that:

$\beta_g = \beta$ for every $g \in \Gamma_{\mathsf{neutral}}$,

$\beta_g = \beta(1 + \epsilon)$ for every $g \in \Gamma_{\mathsf{pos}}$,

$\beta_g = \beta(1 - \epsilon)$ for every $g \in \Gamma_{\mathsf{neg}}$,

where ϵ is randomly chosen from $\{0, .01, 0.02, \ldots ,0.2\}$.

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

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

The sequencing counts from T0 $n^{\mathsf{seq}}_{\mathsf{t0},g}$ are obtained by making $C_{\mathsf{PCR}} N_{\mathsf{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\mathsf{t0},g} / \sum_g n_{\mathsf{t0},g}$. 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 $\mathsf{i} = 1 \ldots \mathsf{N_{split}}$, the gRNAs abundances after cell splitting $n_{\mathsf{i,split},g}$ are obtained by making $C_{\mathsf{cells}} N_{\mathsf{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{\mathsf{i},g} / \sum_g n_{\mathsf{i},g}$. This random sampling step is followed by an exponential growth step $n_{\mathsf{i+1},g} = e^{\beta_g} n_{\mathsf{i,split},g}$. After completion of all cell splittings the gRNAs in all biological replicates (timepoint T1) are sequenced by making $C_{\mathsf{PCR}} N_{\mathsf{tot}}$ draws from the multivariate hypergeometric distribution with probabilities $p_g = n_{N_{\mathsf{split}},g} / \sum_g n_{N_{\mathsf{split}},g}$. This is repeated N_{biopcr} times, to model the technical replicates.

For the analyses shown in Figures 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 total read counts of the reference sample. This was not necessary for simulated datasets, because these already had same read counts. For a given gRNA with count n_{lib} in the library and n_1 at timepoint T1, the log fold change was calculated as $\mathsf{LFC} = \mathsf{log2}(\frac{\mathsf{n_1+1}}{\mathsf{n_{lib}+1}})$. 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 90% and 10% percentile of the distribution[24]: $\mathsf{library\ width} = \frac{\mathsf{percentile_{90}(n_{lib,g})}}{\mathsf{percentile_{10}(n_{lib,g})}}$. $n_{\mathsf{lib,g}}$ is the distribution of gRNA abundances in the library.

gRNA sequence composition

The sequence logos in Fig. S4 were generated using the plogo online tool[43]. gRNAs were ranked according to their abundance and the sequence logos generated for the lower and upper 1% 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 timepoint (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[44,45]. 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 squares regression[46] 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 gRNA 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 algorithm with a α cutoff of 5% to aggregate the data to the gene level[29,47]. Gene level LFCs are calculated by averaging the LFCs over all gRNAs belonging to the gene.

Comparison of analysis tools

Results from analysis ofimulated 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 same p-value according to their mean LFC overall gRNAs.
- MAGeCK analysis was performed using the RRA algorithm, without normalization to controls. Genes were ranked according to the rank provided by MAGeCK.
- BAGEL analysis was performed only on experimental data because the algorithms needs 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 done based on Bayes factors.
- screenBeam analysis was performed without removal of low counts. Genes were ranked according to p-values.

List of abbreviations

LFC: logarithmic fold change (base 2); gRNA: guide RNA; RNA: ribonucleic acid; CRISPR: clustered regularly interspaced short palindromic repeats; Cas9: CRISPR associated 9; DNA: deoxyribonucleic acid; PCR: polymerase chain reaction; GO: gene ontology; RRA: robust rank aggregation.

Availability of data and material

The simulation is available at https://github.com/imkeller/simulate_pooledscreen

`gscreend` (submission to Bioconductor pending) is available at https://github.com/imkeller/gscreend.

Acknowledgements

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

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

Authors contribution

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. The authors declare that they have no conflict of interest.

Supplement

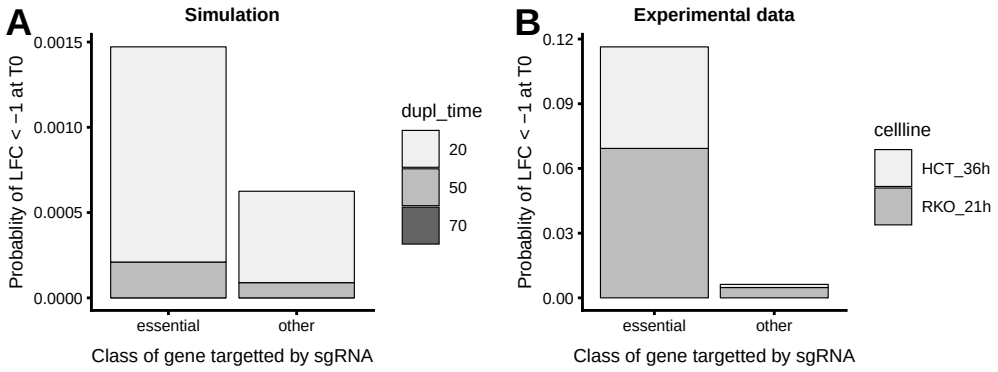


Figure S1: Fraction of gRNAs with reduced abundance at T0. The class of gene targeted by each category of gRNAs is indicated. Essential genes in (A) according to Hart et al.[38]. In (B) essential genes are all those that belonged to the set of growth rate reducing genes. A: Simulation with doubling times of 10, 50 and 70 hours. B: Experimental data from RKO (doubling time 21h) and HCT116 cells (doubling time 36h).

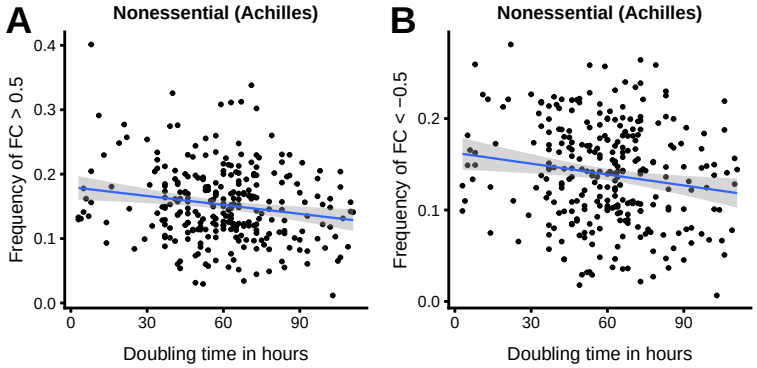


Figure S2: Screening data from the DepMap project shows higher frequency of LFC at higher cell growth rate. Analyses were performed only for gRNAs targeting non-essential genes. A: Positive LFC>+0.5. B: Negative LFC<-0.5.[3,39,40]

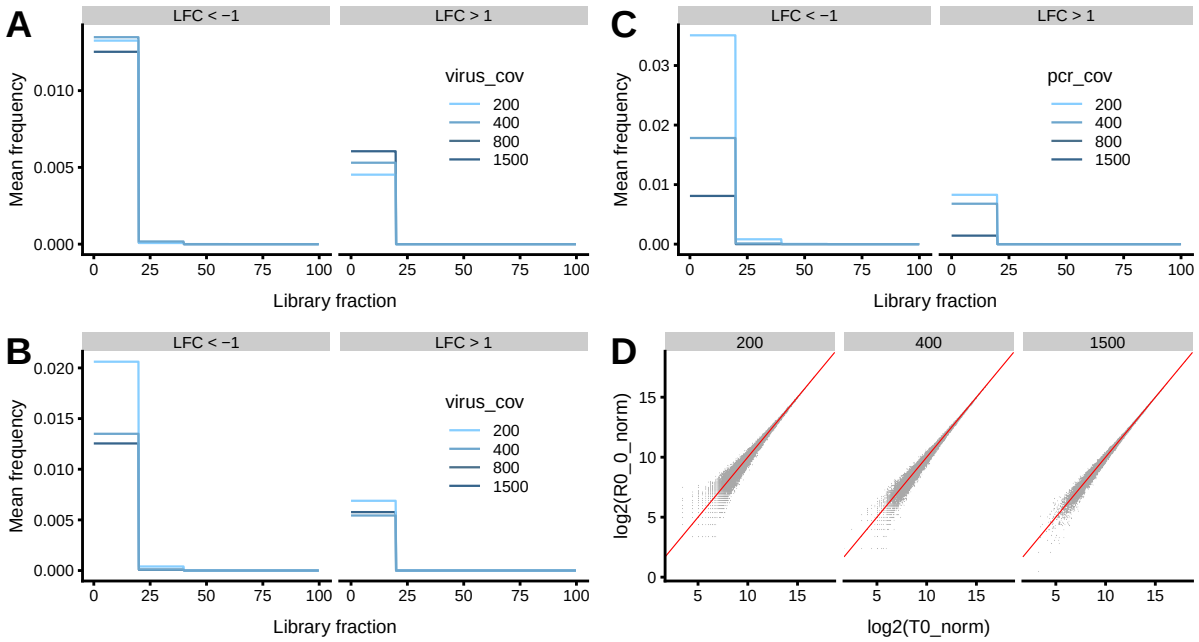
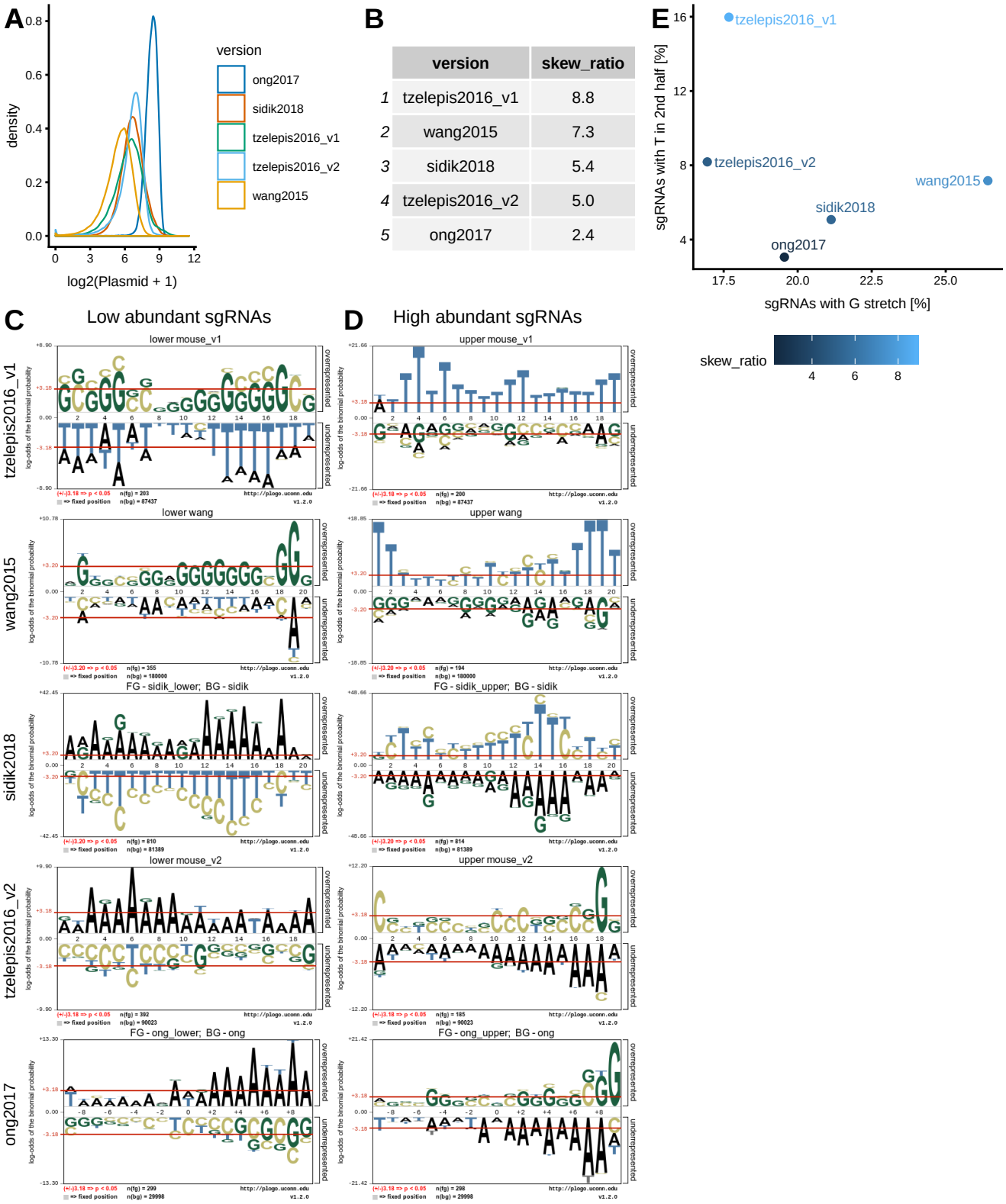


Figure S3: Effect of changing transduction or PCR coverage is small compared to the effects described in Figure 2. A: Frequency of extreme LFC between library sequencing and T1 for different values of transduction coverage. B: Frequency of extreme LFC between T0 and T1 for different values of transduction coverage. C: Frequency of extreme LFC between T0 and T1 for different values of PCR coverage. D: log2 gRNA counts at T1 vs. T0 for different values of PCR coverage.



	version	skew_ratio
1	tzelepis2016_v1	8.8
2	wang2015	7.3
3	sidik2018	5.4
4	tzelepis2016_v2	5.0
5	ong2017	2.4

Figure S4: gRNA sequence features influence homogeneity of library distribution. A+B: Selection of 5 different whole-genome CRISPR-KO libraries with different log2 gRNA count distributions (A, normalized to total counts) and library distribution widths (B)[1,27,41,42]. C+D: Probability sequence motifs for low (B) and high (C) abundant gRNAs in the five libraries (ordered by decreasing distribution width). E: Percentage of gRNAs with more than 3 Ts in the 2nd half of the sequence (y-axis) and percentage of gRNAs with at least 3 Gs in a row (x-axis).

References

1. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).

2. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* (2015). doi:10.1016/j.cell.2015.11.015

3. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564–576.e16 (2017).

4. Wang, T. *et al.* Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. *Cell* **168**, 890–903.e15 (2017).

5. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80–84 (2014).

6. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nature Reviews Genetics* **16**, 299–311 (2015).

7. Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nature Biotechnology* **33**, 661–667 (2015).

8. Steinhart, Z. *et al.* Genome-wide CRISPR screens reveal a Wnt–FZD5 signaling circuit as a druggable vulnerability of RNF43-mutant pancreatic tumors. *Nature Medicine* **23**, 60–68 (2017).

9. Kim, H. S. *et al.* CRISPR/Cas9-mediated gene knockout screens and target identification via whole-genome sequencing uncover host genes required for picornavirus infection. *Journal of Biological Chemistry* **292**, 10664–10671 (2017).

10. Han, J. *et al.* Genome-wide CRISPR/Cas9 Screen Identifies Host Factors Essential for Influenza Virus Replication. *Cell reports* **23**, 596–607 (2018).

11. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnology* **32**, 267–273 (2014).

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

13. Charton, K. *et al.* Exploiting the CRISPR/Cas9 system to study alternative splicing in vivo : application to titin. *Human Molecular Genetics* **25**, ddw280 (2016).

14. Gapinske, M. *et al.* CRISPR-SKIP: programmable gene splicing with single base editors. *Genome Biology* **19**, 107 (2018).

15. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife* **5**, e19760 (2016).

16. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology* **33**, 510–517 (2015).

17. Liu, X. S. *et al.* Editing DNA Methylation in the Mammalian Genome. *Cell* **167**, 233–247.e17 (2016).

18. Morita, S. *et al.* Targeted DNA demethylation in vivo using dCas9–peptide repeat and scFv–TET1 catalytic domain fusions. *Nature Biotechnology* **34**, 1060–1065 (2016).

19. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

20. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

21. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nature Methods* **12**, 326–328 (2015).

22. Nishida, K. *et al.* Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems.

Science (New York, N.Y.) **353**, aaf8729 (2016).

23. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

24. Joung, J. *et al.* Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nature Protocols* **12**, 828–863 (2017).

25. Doench, J. G. Am i ready for CRISPR? A user's guide to genetic screens. *Nature Reviews Genetics* **19**, 67–80 (2018).

26. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nature Methods* **11**, 783–784 (2014).

27. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell reports* **17**, 1193–1205 (2016).

28. Aregger, M., Chandrashekhar, M., Tong, A. H. Y., Chan, K. & Moffat, J. Pooled Lentiviral CRISPR-Cas9 Screens for Functional Genomics in Mammalian Cells. in *Methods in molecular biology* **1869**, 169–188 (Humana Press, New York, NY, 2019).

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

30. Yu, J., Silva, J. & Califano, A. ScreenBEAM: a novel meta-analysis algorithm for functional genomics screens via Bayesian hierarchical modeling. *Bioinformatics* **32**, btv556 (2015).

31. Diaz, A. A., Qin, H., Ramalho-Santos, M. & Song, J. S. HiTSelect: a comprehensive tool for high-complexity-pooled screen analysis. *Nucleic Acids Research* **43**, e16–e16 (2015).

32. Hart, T. & Moffat, J. BAGEL: A computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 1–7 (2016).

33. Jia, G., Wang, X. & Xiao, G. A permutation-based non-parametric analysis of CRISPR screen data. *BMC Genomics* **18**, 545 (2017).

34. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knock-out screens. *bioRxiv* 285114 (2018). doi:10.1101/285114

35. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487–491 (2014).

36. Parnas, O. *et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675–686 (2015).

37. Daley, T. P. *et al.* CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. *Genome Biology* **19**, 159 (2018).

38. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3: Genes|Genomes|Genetics* **7**, 2719–2727 (2017).

39. DepMap; Broad. DepMap Achilles 19Q1 Public. *figshare* **Fileset**, (2019).

40. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779–1784 (2017).

41. Sidik, S. M. *et al.* A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. *Cell* **166**, 1423–1435.e12 (2016).

42. Ong, S. H., Li, Y., Koike-Yusa, H. & Yusa, K. Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. *Scientific Reports* **7**, 1–10 (2017).

43. O'Shea, J. P. *et al.* PLogo: A probabilistic approach to visualizing sequence motifs. *Nature Methods* **10**, 1211–1212 (2013).

44. Efron, B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the*

American Statistical Association **99**, 96–104 (2004).

45. Strimmer, K. A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303 (2008).

46. Rousseeuw, P. J. & Leroy, A. M. Robust regression and outlier detection. *Wiley Series in Probability and Statistics* 329 (1987).

47. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics (Oxford, England)* **28**, 573–80 (2012).

48. Nagy, T. & Kampmann, M. CRISPulator: A discrete simulation tool for pooled genetic screens. *BMC Bioinformatics* **18**, 1–12 (2017).

49. Shifrut, E. *et al.* Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell* **175**, 1958–1971.e15 (2018).

50. Wegner, M. *et al.* Circular synthesized CRISPR/Cas gRNAs for functional interrogations in the coding and noncoding genome. *eLife* **8**, (2019).

51. Chen, C.-H. *et al.* Improved design and analysis of CRISPR knockout screens. *Bioinformatics* 1–7 (2018). doi:10.1093/bioinformatics/bty450