A dose-response based model for statistical analysis of chemical genetic interactions in CRISPRi libraries

Sanjeevani Choudhery[1*], Michael A. DeJesus[2], Aarthi Srinivasan[1], Jeremy Rock[2], Dirk Schnappinger[3], Thomas R. Ioerger[1]

[1]Department of Computer Science and Engineering, Texas A&M University, College Station, Texas, United States of America

[2]Laboratory of Host-Pathogen Biology, The Rockefeller University, New York, New York, United States of America

[3]Department of Microbiology and Immunology, Weill Cornell Medical College, New York, New York, United States of America

* Corresponding author

E-mail: schoudhery@tamu.edu (SC)

Abstract

An important application of CRISPR interference (CRISPRi) technology is for identifying chemical-genetic interactions (CGIs). Discovery of genes that interact with exposure to antibiotics can yield insights to drug targets and mechanisms of action or resistance. The premise is to look for CRISPRi mutants whose relative abundance is suppressed (or enriched) in the presence of a drug when the target protein is depleted, reflecting synergistic behavior. One thing that is unique about CRISPRi experiments is that sgRNAs for a given target can induce a wide range of protein depletion. The effect of sgRNA strength can be partially predicted based on sequence features or empirically quantified by a passaging experiment. sgRNA strength interacts in a non-linear way with drug sensitivity, producing an effect where the concentration-dependence is maximized for sgRNAs of intermediate strength (and less so for sgRNAs that induce too much or too little target depletion). sgRNA strength has not been explicitly accounted for in previous analytical methods for CRISPRi. We propose a novel method for statistical analysis of CRISPRi CGI data called CRISPRi-DR (for Dose-Response model). CRISPRi-DR incorporates data points from measurements of abundance at multiple inhibitor concentrations using a classic dose-response equation. Importantly, the effect of sgRNA strength can be incorporated into this model in a way that mimics the non-linear interaction between the two covariates on mutant abundance. We use CRISPRi-DR to re-analyze data from a recent CGI experiment in *Mycobacterium tuberculosis* and show that genes known to interact with various anti-tubercular drugs are ranked highly. We observe similar results in MAGeCK, a related analytical method, for datasets of low variance. However, for noisier datasets, MAGeCK is more susceptible to false positives whereas CRISPRi-DR maintains higher precision, which we observed in both empirical and simulated data, due to CRISPRi-DR's integration of data over multiple concentrations and sgRNA strengths.

Author Summary

CRISPRi technology is revolutionizing research in various areas of the life sciences, including microbiology, affording the ability to partially deplete the expression of target proteins in a specific and controlled way. Among the applications of CRISPRi, it can be used to construct large (even genome-wide) libraries of knock-down mutants for profiling antibacterial inhibitors and identifying chemical-genetic interactions (CGIs), which can yield insights on drug targets and mechanisms of action and resistance. The data generated by these experiments (i.e., nucleotide barcode counts from high throughput sequencing) is voluminous and subject to various sources of noise. The goal of statistical analysis of such data is to identify significant CGIs, which are genes whose depletion sensitizes cells to an inhibitor. In this paper, we show how to incorporate both sgRNA strength and drug concentration simultaneously in a model (CRISPRi-DR) based on an extension of the classic dose-response (Hill) equation in enzymology. This model has advantages over other analytical methods for CRISPRi, which we show using empirical and simulated data.

Introduction

CRISPR interference (CRISPRi) has become popular for genome-wide profiling of the biological roles of genes in various growth conditions. By detecting growth defects caused by depletion of individual genes or operons, genes may be associated with responses to different stress conditions. The concept of gene 'vulnerability' has recently been introduced to describe the sensitivity of cells to partial depletion of individual proteins. By this definition, highly vulnerable genes are genes for which minimal depletion of protein levels causes growth impairment, which can be quantified efficiently on a genome-wide scale using high-throughput sequencing [1]. The vulnerability of a gene can be condition dependent, or strain dependent [1]. CRISPRi can be used to reveal targets of antibiotics or mechanisms

68 of resistance through chemical-genetic interactions [2, 3]. CRISPRi libraries are often designed to contain
69 multiple small guide RNAs (sgRNAs) targeting each gene, resulting in a population of thousands of
70 individual depletion mutants [1]. The abundance of each sgRNA can be quantified by amplifying the
71 sgRNA targeting sequence which functions as a molecular barcode, and then performing deep
72 sequencing to count the number of barcodes for each sgRNA in a treatment. The analysis of such
73 datasets is challenging, due to various sources of noise, which introduces variability in the counts.
74 A previously published method for analyzing CRISPRi datasets, called MAGeCK [4], fits the data
75 to a negative binomial distribution, calculates a log-fold-change (of mean counts) for each gene between
76 a treatment condition and a reference condition (control, e.g. buffer with 5% DMSO as solvent), and
77 uses a negative binomial (NB) mass function to test the differences in significance of sgRNA abundance
78 between treatments and controls. To evaluate effects at the gene level, individual sgRNAs are
79 combined in MAGeCK using Robust Rank Aggregation (RRA) to prioritize genes whose sgRNAs show
80 greater enrichment or depletion on average than other genes in the genome. MAGeCK has been used
81 for evaluating chemical-genetic interactions (CGI) with antibiotics [4].
82 However, MAGeCK has two limitations for this application. First, gene-drug interaction studies
83 are usually carried out over several drug concentrations around the MIC (minimum-inhibitory
84 concentration), since it is often difficult to anticipate what concentration will stimulate 50% growth
85 inhibition of mutants in combination with CRISPRi-induced depletion of target proteins. However,
86 MAGeCK analyzes the data for each drug concentration independently (each concentration compared to
87 a no-drug control). Knock-down mutants might exhibit depletion at one concentration but not others.
88 Results from multiple concentrations must be combined post-hoc, such as by taking the union of
89 MAGeCK hits at any concentration. Due to the noise in these CRISPRi experiments, this increases the
90 risk of detecting false positives (in the sense that non-interacting genes that might be mistakenly called
91 as hits independently at different concentrations are combined). In practice, for some datasets,

92 MAGeCK reports an unreasonably large set of significant interactions, not all of which may be
93 biologically genuine. Second, MAGeCK does not explicitly take into account differences in sgRNA
94 strength. Different sgRNAs are known to induce different degrees of depletion of their target genes.
95 This can be quantified beforehand by evaluating the growth rate of individual mutants in a passaging
96 experiment and determining how fitness correlates with target knockdown [1]. In highly vulnerable
97 genes, the strength or effectiveness of depletion by sgRNAs can span a range from no effect to severe
98 growth defect. This information was not anticipated at the time MAGeCK was developed (as the early
99 applications of CRISPRi were primarily being used to fully inactivate genes, rather than to produce
100 graded effects), and the Robust Rank Aggregation method treats all sgRNAs in a gene as "equal",
101 without differentiating them based on the expected effects due to sgRNA strength.
102 In this paper, we propose a new methodology for statistical analysis of CRISPRi libraries and
103 identification of chemical-genetic interactions. A regression model is used to integrate data over
104 multiple drug concentrations. The degree of a gene-drug interaction is reflected by the coefficient (or
105 slope) for the dependence of sgRNA abundance on drug concentration. This regression approach was
106 previously introduced for analysis of hypomorph libraries (where there is just one to three mutants
107 representing each gene) [5]. It was based on the theory that depletion of the target of a drug should
108 synergize with increasing concentrations of the drug. While exposure to sub-MIC levels of an inhibitory
109 compound will challenge the growth of all the mutants in a population (hypomorph library), mutants
110 with depletion of a gene that interacts with a drug (e.g. prototypically, an essential gene that is the drug
111 target) will exhibit excess depletion relative to others in the population due to the combined effect of
112 both the growth-inhibition due to the drug treatment in conjunction with the growth-impairment due to
113 knock-down of an essential gene, making these mutants even more sensitive to the drug. For genes
114 that genuinely interact with a given drug, this depletion effect should be exacerbated at higher drug
115 concentrations (i.e. be dose-dependent); genes of greatest relevance are those that exhibit

116 concentration-dependent effects. While the (log of) abundance of an sgRNA does not have to decrease
117 perfectly linearly with the (log of) concentration to obtain a significant negative coefficient (slope) in the
118 regression, there should be a general trend supporting that abundance decreases as concentration
119 increases. Other researchers have exploited CRISPRi in different ways to detect this synergistic behavior
120 for identifying chemical-genetic interactions. For example, the expression of an active form of dCAS9
121 was titrated to produce different levels of expression of essential proteins in *S. pyrogenes*, looking for
122 genes whose depletion shifted the MIC to inhibitors [3].

123 One of the challenges in extending this prior regression approach to CRISRPi libraries was
124 incorporating information on sgRNA strengths. Even in essential genes, some sgRNAs may produce
125 strong depletion of the target, while others might be almost completely ineffective, generally depending
126 on sequence attributes (similarity to optimal PAM sequence (protospacer-adjacent motif), length, GC
127 content, etc.) [6]. While sgRNA strength can be partially predicted (with intermediate accuracy) from
128 sequence alone, sgRNA strength can also be empirically quantified by measuring or extrapolating log2-
129 fold-changes of abundance (LFCs) in standard growth media *with* versus *without* induction of CRISPRi at
130 a fixed number of generations [1]. Although one could contemplate adding the strength of each sgRNA
131 (predicted, or empirically measured) into the regression model to predict abundances for each gene, a
132 significant problem (expanded upon below) is that sgRNAs of different strength can show different
133 concentration dependence.

134 In this paper, we propose a modified regression approach for CRISRPi data (called CRISPRi-DR)
135 that incorporates both drug concentration and sgRNA strength. The approach is based on the classic
136 dose-response (DR) model for inhibition activity of drugs; the activity of a target protein typically
137 transitions from high to low in shape of an S-curve as concentration increases (on a log scale), which can
138 be modeled with a Hill equation. The parameters of the Hill equation for a given drug can be fit by
139 performing a log-sigmoid transformation of the enzyme activity data and then using ordinary least-

140 squares regression. We show how sgRNA strength can be incorporated into this model as a
141 multiplicative effect in the Hill equation, which becomes an additive effect in the log-sigmoid
142 transformed data. The important consequence of this model is that it decouples the concentration-
143 dependence from the sgRNA strength, so they can be fit as independent (non-interacting) terms in the
144 regression. We demonstrate the value of the CRISPRi-DR analysis method by re-analyzing the data from
145 a recent paper using CRISPRi for chemical-genetic interactions to identify targets of antibiotics in *M.*
146 *tuberculosis*.

147

148 # Methods

149 CRISPRi experiments involve using high-throughput sequencing to tabulate counts of nucleotide
150 barcodes representing abundance of individual mutants in a population (or library). Each mutant has an
151 sgRNA mapping to a target gene that can reduce its expression (when induced with ATC,
152 anhydrotetracycline). In CGI applications, the library is sequenced in the presence of antibiotics or
153 inhibitors at various concentrations, along with a no-drug control. If Y_{ijk} is the abundance (i.e. count)
154 for an sgRNA i in a condition j for replicate k, normalized abundance can be given by $Y'_{ijk} = \frac{Y_{ijk}}{\sum_{x=1}^{n} Y_{xjk}}$,
155 where each count is divided by the sum of counts of the n sgRNAs observed in a given condition and
156 replicate. Let U'_i be the normalized abundance of sgRNA i in the uninduced (-ATC) library, then the
157 normalized relative abundances of an sgRNA i in all induced (+ATC) samples can be calculated as: $A_{ijk} =$
158 $\frac{Y'_{ijk}}{U'_i}$, assuming that the abundance in –ATC represents no depletion (100% full abundance). Although
159 increases greater than 1 are possible in treated conditions, these relative abundances ideally range
160 between 0 and 1 (i.e., 100% as a percentage). This absolute scale is required for the dose-response
161 model.

CRISPRi dose-response model

The CRISPRi-DR model for analyzing CRISPRi data from CGI experiments is an extension of the basic dose-response model, extended to incorporate sgRNA strengths. The dose-response effect of an inhibitor on the activity of an enzyme is traditionally modeled with the Hill-Langmuir equation.

$$\theta = \frac{1}{1 + \left(\frac{K_A}{[L]}\right)^n} \qquad [1]$$

where θ is the fraction of abundance (relative to no drug), [L] is the ligand concentration, K_A is the concentration at which there is 50% activity and n is the Hill coefficient.

Applying [1] to the CGI data, the relative abundance of sgRNAs A_{ijk} is used as the predictor variable and $[D_j]$ is the concentration of drug *j* that the *k*th replicate count of sgRNA *i* was extracted from,

$$A_{ijk} = \frac{1}{1 + \left(\frac{EC_{50}(D_j)}{[D_j]}\right)^{H_d}} \qquad [2]$$

The unknown parameters are the EC_{50} value (effective concentration that causes 50% growth inhibition) and the Hill coefficient H_d. The plot of the concentration versus relative abundance of an sgRNA (A_{ijk}) produces a sigmoidal curve, demonstrating how activity decreases as concentration increases, with the EC_{50}, representing the mid-point of the transition.

The dose-response model seen in [2] can be extended to account for sgRNA strength by incorporating a multiplicative factor in the denominator:

$$A_{ijk} = \frac{1}{1 + \left(\frac{EC_{50}(D_j)}{[D_j]}\right)^{H_d} \left(\frac{K_s}{S_i}\right)^{H_s}} \qquad [3]$$

181 sgRNA strength, S_i, is quantified by the estimate degree of growth impairment at 25 generations of
182 growth in-vitro (log2-fold-change of abundance with ATC vs without, $LFC = log2(\frac{+ATC}{-ATC})$ in the absence
183 of drug, extrapolated from a model fit to empirical data from passaging for each sgRNA [1]. K_s
184 represents the unknown intermediate sgRNA strength that causes 50% depletion of mutant abundance
185 (half-way between no depletion and full depletion), and the H_s is the unknown Hill coefficient that
186 represents how sensitive mutant abundance is to depletion of the target protein.
187

188 Relationship between drug concentration and gene depletion within
189 the CRISPRi-DR model

190 Abundance of mutants in a CRISPRi CGI experiment can be affected simultaneously by both
191 presence of an inhibitor and depletion of a vulnerable gene. However, the concentration-dependent
192 effect of a drug on mutant abundance can be different for sgRNAs of different strength. For example, a
193 strong sgRNA can cause excessive depletion, making it difficult to detect additional decreases due to
194 increasing drug concentration; weak sgRNAs might not induce enough depletion to synergize with the
195 drug; sgRNAs of intermediate strength can provide just the right amount of depletion to maximize the
196 interaction with the drug, producing the most pronounced concentration-dependent effects
197 (sensitization). Fig 1 illustrates this with sgRNAs, spanning a range of strengths, in *rpoB* (RNA polymerase
198 beta subunit, target of rifampicin) treated with rifampicin (RIF) over a range of concentrations. In Fig 1A ,
199 the sgRNA strength (extrapolated LFCs at 25 generations) is plotted versus observed depletion (log of
200 +ATC/-ATC) in the absence of any drug for each sgRNA in *rpoB* in a log-log space. Since strength is
201 measured as extrapolated LFC, the more negative the LFC, the greater the depletion and hence stronger
202 the sgRNA. The points follow the linear dashed line, demonstrating that, as sgRNA strength increases,
203 abundance decreases. The lines in Fig 1B are regression fits obtained for each sgRNA in *rpoB* in RIF (5

204 days of pre-depletion, D5) using regression of log abundances with log concentration, $\log(A_{ijk}) = C +$

205 $B \cdot \log([D_j])$, where *C* is in the intercept and *B* is the slope of the regression, representing concentration

206 dependence, and $\log(A_{ijk})$ are log relative abundances obtained as described above. The left-most side

207 of Fig 1B (log concentration = 0) shows the range of abundances with no drug concentration (ATC-

208 induced library in buffer). Regression lines have starting points at various abundances (relative to -ATC),

209 due solely to the growth impairment cause by depleting *rpoB*. As concentration of RIF increases, some of

210 the sgRNAs show very negative slopes, while other sgRNAs show slopes closer to 0. This illustrates that

211 sgRNAs within a gene in a particular condition can show vastly different concentration dependencies. A

212 parabolic-type curve emerges in Fig 1C when the slopes from the regressions performed on each sgRNA

213 seen in Fig 1B are plotted against the sgRNA strengths. The strongest sgRNAs (left on the plot) and the

214 weakest sgRNAs (right side on the plot) show slopes around 0. These regressions represent the flat lines

215 in at the top and the bottom of the graph in Fig 1B . As seen in Fig 1A, strong sgRNAs (left of plots Fig 1A

216 and Fig 1C) already have a low starting abundance, so with increasing concentration, there is little

217 depletion. With weak sgRNAs (right of plots in Fig 1A and Fig 1C), starting abundances are high, but the

218 sgRNAs are too weak to show depletion with increasing concentration. The sgRNAs surrounding the

219 minimum point of this parabolic curve (dashed line) reflect those of intermediate strength, where the

220 ability to detect synergy with the drug is maximized. Similar behavior is observed for many other genes

221 in the presence of other drug treatments. The strength where the slopes reach their extrema points can

222 be different for each gene. The variability of concentration-dependence (slope) with sgRNA strength

223 suggests a possible non-linear interaction between the variables. However, this nonlinearity is captured

224 in the multiplicative terms of the dose-response model (Eqn. 3).

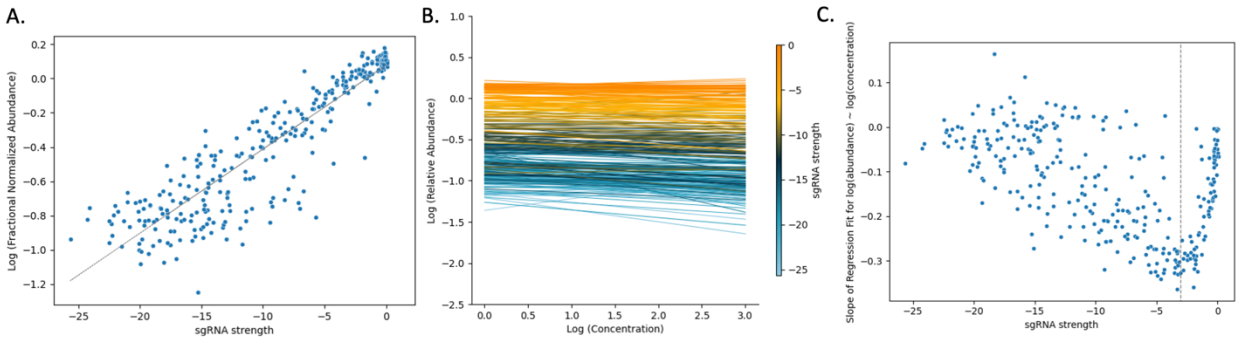


Fig 1. Effect of sgRNA strength and drug concentration on abundance of mutants in *rpoB* in a CRISPRi library treated with RIF (D5).

(A) Comparison of fractional abundances of sgRNAs in *rpoB* (+ATC / -ATC) to their strengths (in the form of extrapolated LFCs 25 generations in the future). There is a strong correlation of depletion and sgRNA strength in *rpoB* (RNA polymerase beta subunit, target of rifampicin). There is a linear relationship between these two values, evident by the line of best fit (R^2 = 0.82). Since strength is measured as extrapolated LFC, the more negative the LFC, the stronger the sgRNA. Here we see that almost linearly, as sgRNA strength increases, abundance decreases. (B) Regression lines for log(relative abundance) against log(concentration) for all sgRNAs in *rpoB* in a library treated with RIF D5. Although the starting abundance varies, the majority of the regression lines show a negative slope, demonstrating that as concentration of RIF increases, the abundance of sgRNAs in *rpoB* decrease. The lines that reflect the extremes of the sgRNA strength (orange or blue), are flat and do not show much change in abundance. Comparatively, the middle of sgRNA strength range (navy blue) show the greatest negative slopes reflecting this is the region of ideal sgRNA strength. (C) Comparison of sgRNA strength and slopes of a regression of log(relative abundance) against log(concentration) for each sgRNA in *rpoB* in a library treated with RIF D5. Each slope (one for each sgRNA) seen in Panel B versus its strength show a parabolic curve. The strongest sgRNAs (left on the plot) and the weakest sgRNAs (right side on the plot) show slopes around 0. These regressions are the flat lines in at the top and the bottom of the graph in Panel B. As seen in Panel A, with strong sgRNAs (left of plot), we already have a low starting abundance,

245 so with increasing concentration, there is little depletion. With weak sgRNAs (right of the plot), starting
246 abundances are high, but the sgRNA is too weak to show depletion with increasing concentration. The
247 minimum of the parabolic curve (dotted line) are sgRNAs of intermediate strength where the ability to
248 detect synergy with the drug is maximized
249

250 Linearization and parameter estimation

251 The dose-response model [3] can be linearized through a log-sigmoid transformation.

252 $$\log\left(\frac{A_{ijk}}{1-A_{ijk}}\right) = H_d \cdot \log([D_j]) + H_s \cdot S_i + C$$

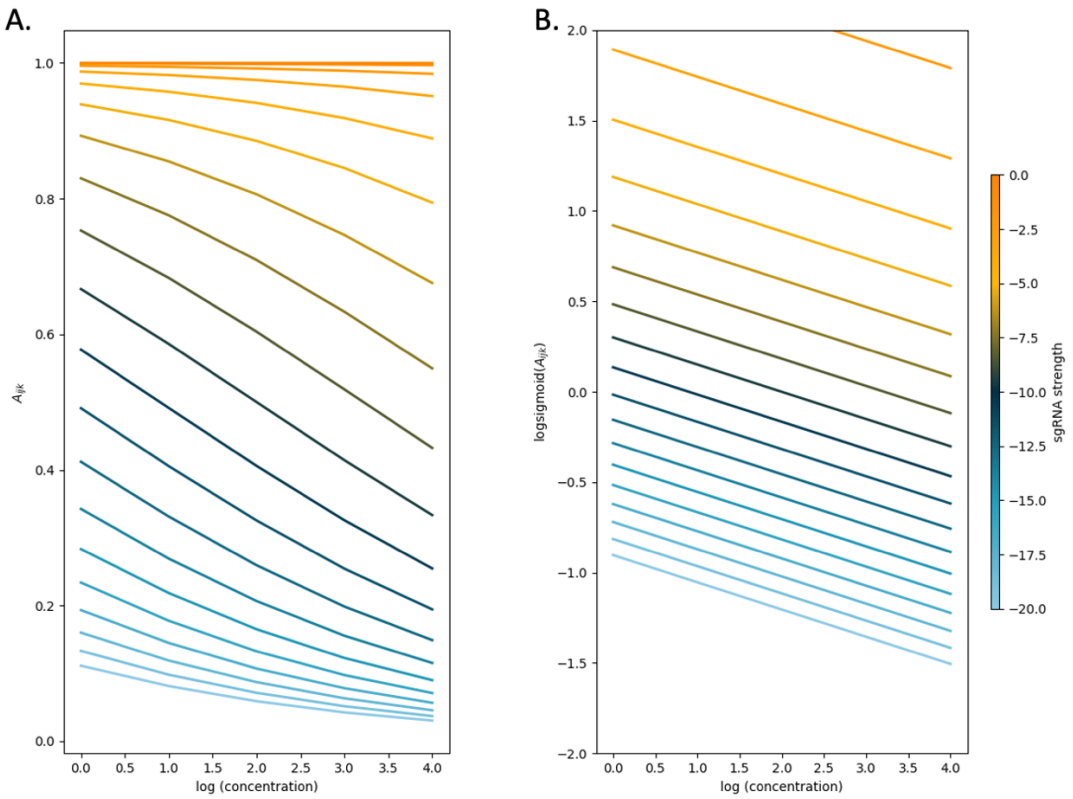
253 $$C = H_s \cdot \log(K_s) - H_d \cdot \log\left(EC_{50}(D_j)\right) \quad [4]$$

254 In this log-sigmoid transformed space, the concentration-dependence and effect of sgRNA strength have
255 been decoupled (non-interacting), and thus are independent linear terms with the Hill coefficients (H_s
256 and H_d) as the variables to solve for by a standard regression. The inflection parameters of the sigmoid
257 curve (K_s and EC_{50}) are combined as the intercept C in the model. Importantly, this model implies that
258 the effect of growth impairment due to the depletion of a vulnerable gene and growth inhibition due to
259 the drug on the overall (relative) abundance of a given mutant are independent, because the effects are
260 an "additive" in log-space. To illustrate this, the CRISPRi-DR equation is simulated by plotting idealized
261 relative abundances (in Fig 2) using parameters chosen to emulate what is seen in Fig 1B; the *rpoB* plot
262 of slopes over a systematic range of sgRNA strengths and drug concentrations. In Fig 2A, the slopes of
263 the concentrations are plotted against abundances calculated using the dose-response model. The
264 slopes change as a function of the starting depletion (left-hand side), which varies due to sgRNA
265 strength alone (colored by blue-orange gradient based on strength value). The slopes are most negative
266 for intermediate sgRNA strength, colored with a dark blue-green hue representing sgRNA strength
267 (extrapolated LFCs) around -10. Fig 2B shows the result of the linearization of the Hill equation. All the

268 individual sgRNA regression lines over concentration become parallel, eliminating the dependence on

269 sgRNA strength, and allowing them to be fit by a single common slope representing the concentration-

270 dependence averaged over all the sgRNAs.



271

272 **Fig 2. The log-sigmoid transformation of abundances allows the CRISPRi-DR model to factor in the**

273 **non-linear effect of sgRNA strength on concentration dependence.**

274 (A) Simulation of sgRNAs abundances for an ideal essential gene. Parameters used in simulation: H_s = -4,

275 EC_{50} = 0.25, K_s = -10 and H_d = -0.5 over a range of sgRNA strengths and drug concentrations. (B) When

276 the log-sigmoid transformation of the abundances is applied, we see all the regression fits are parallel to

277 one another, allowing to be fit by a single common slope, representing the concentration dependence

278 over all sgRNAs, regardless of sgRNA strength.

279

280 The data (sgRNA relative abundances from sequencing) are fit on a gene-by-gene basis using
281 ordinary least-square (OLS) regression by the following formula:

282 $$\log\left(\frac{A_{ijk}}{1-A_{ijk}}\right) = \beta_0 + \beta_c \cdot \log([D_j]) + \beta_s \cdot S_i \quad [5]$$

283 where *A* (relative abundance for each sgRNA at given drug concentration), S_i (sgRNA strength estimated
284 by predicted log fold depletion at 25 generations based on passaging) and $[D_j]$ (concentration of drugs)
285 are columns of a melted matrix. To include the control samples (no-drug ATC-induced controls,
286 concentration 0) in the regression, they are treated as one two-fold dilution lower than the lowest
287 available concentration tested for the drug (to avoid taking the log of 0). Since the log-sigmoid transform
288 of the relative abundances is taken, they must be within the range of (0,1) but not equal to either
289 extremum. While relative abundances are generally non-negative, they can be greater than 1.0,
290 reflecting sgRNAs that increase in abundance with drug concentration relative to the uninduced (-ATC)
291 condition. To account for this, the following squashing function is applied to adjust outlying values to be
292 within the desired range, while retaining monotonicity:

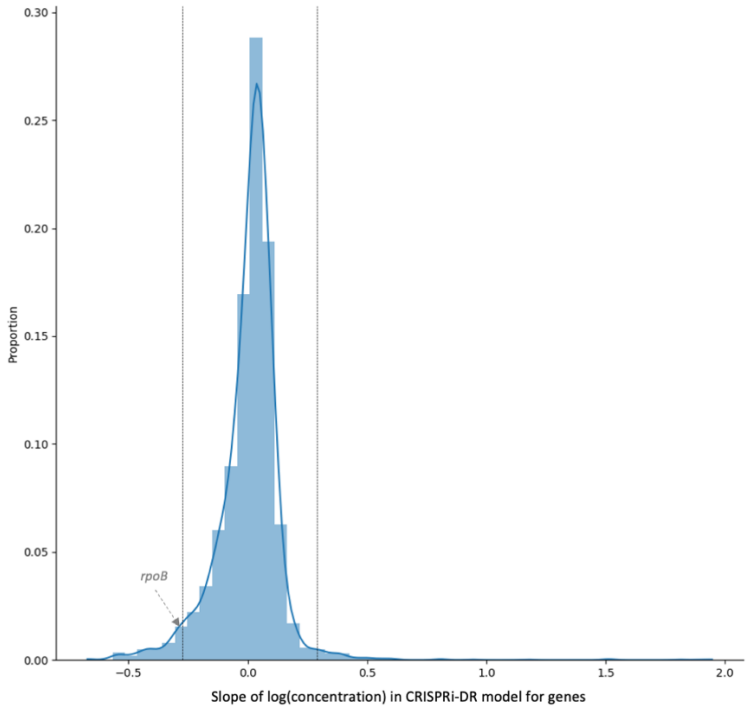
293 $$A_{ijk} = \tau + \frac{(1-\tau)(1-e^{-2A_{ijk}})}{(1+e^{-2A_{ijk}})} \quad [6]$$

294 where τ=0.01 is a pseudo count needed to make abundances non-zero for taking logarithms.
295

296 ## Significance Testing

297 The statistic that indicates the degree of interaction of each gene with a given drug is the
298 coefficient for the $log([D])$ term (i.e. slope) in the model. To determine whether the interaction is
299 statistically significant, a Wald test [7] is applied to calculate a p-value reflecting whether the coefficient
300 is significantly different than 0, adjusting for a target FDR (false discovery rate) of 5% over the whole
301 genome using the Benjamini-Hochberg procedure [8]. However, the Wald test by itself yields too many
302 hits (i.e., the genes predicted to have the greatest interaction with the drug, with adjusted p-value <

0.05). The test selects genes with slopes that are technically different than 0, but not necessarily large enough to be biologically meaningful. Therefore, genes are filtered based on the magnitude of the slopes, analogous to the criterion of |LFC|>1, used by Li et al. [2], to filter significant genes by MAGeCK. The distribution of slopes over all genes is assumed to be a normal distribution, and the Z-scores are computed for every gene g: $Z_g = \frac{\beta_{c,g} - \mu(\beta_c)}{\sigma(\beta_c)}$, where $\sigma(\beta_c)$ is the standard deviation of the slopes of log concentration dependence and $\mu(\beta_c)$ is the mean of the slopes. Genes with $|Z_g|$ < 2.0 are filtered out. This produces hits whose slopes are significant outliers (>2σ) from the rest of the population (genes in the genome). There are two groups of hits, corresponding to the two tails of the distribution: enriched hits where Z_g > 2.0, and depleted hits, Z_g < -2.0. Fig 3 shows the distribution of the slopes calculated for genes in a library treated with RIF (one day of pre-depletion, D1). The threshold for this distribution where $|Z_g|$>2.0 and adjusted p-value < 0.05, is at slope = -0.28 and slope = 0.28 (vertical bars). The 195 total genes in the tails outside the vertical lines are identified as significant genes. These genes include the target of RIF, *rpoB.*



317 **Fig 3. Coefficient of log-dependence from CRISPRi-DR model fitted for RIF D1 (1 day of pre-depletion).**
318 The distribution of the slopes of concentration dependence, extracted from the model fit for each gene.
319 The vertical lines are at slope = -0.28 and slope = 0.28. These are the slopes adjusted p-value < 0.05 and
320 the |Z-score|> 2.0. 195 genes have significant slope values, i.e., 195 genes show a significant change in
321 abundance with increasing RIF concentration while accounting for sgRNA strength. *rpoB* is significant
322 with a slope of -0.29.
323
324

325 Results

326 CRISPRi data and pre-processing

327 The data was obtained from high-throughput sequencing of a CRISPRi library of *M. tuberculosis*
328 (*Mtb*) of 96,700 sgRNAs [2]. For all 4019 genes in the Mtb H37Rv genome, there is an average of 24
329 sgRNAs per gene (range: 4-711). This library was intentionally constructed to focus on probing essential
330 genes (based on prior TnSeq analysis [9]), with a mean of 83 sgRNAs per essential gene but there are
331 some sgRNAs in each non-essential gene (mean of 10 sgRNAs per non-essential gene).
332 Samples of the library induced with ATC, in the presence of a drug were sequenced in triplicate
333 at several concentrations for each drug at 2-fold dilutions around the MIC, along with control samples
334 representing the no-drug ATC-induced samples (0 concentration). Three periods of pre-depletion (+ATC,
335 prior to antibiotic exposure) were evaluated: 1, 5, and 10 days (D1, D5, and D10). The measurements
336 reported in this library are observed barcodes counts of mutants in a culture, each with a different
337 sgRNA, representing the relative proportion of each mutant in the population (i.e., abundance).
338 However, abundance can increase or decrease if a vulnerable gene is depleted through CRISPRi
339 interference, causing a change in fitness. Although levels of a target protein are knocked down by

340 transcription interference via CRISPRi, protein levels are not directly measured. The barcodes that are
341 being counted are nucleotides amplified from plasmids in the cells. This indirectly reflects the growth
342 defect caused by depletion of a vulnerable gene. Each individual sample consisted of a vector of 96,700
343 barcode counts. Samples were normalized by dividing individual counts for each sgRNA by the sample
344 total (sum over all sgRNAs).

345 Prior estimates of sgRNA strengths are also required. These were obtained from empirical data
346 by fitting a piecewise-linear equation to fitness over multiple generations, and then inferring the
347 predicted log-fold change at 25 generations [1]. As the absolute effect of depletion solely due to the
348 sgRNA induction plays an important role in the CRISPRi-DR model (below), the analysis also requires
349 samples representing abundance of mutants in the absence of -ATC (no dCAS9 expression, and hence no
350 depletion of target transcripts by sgRNAs).

351

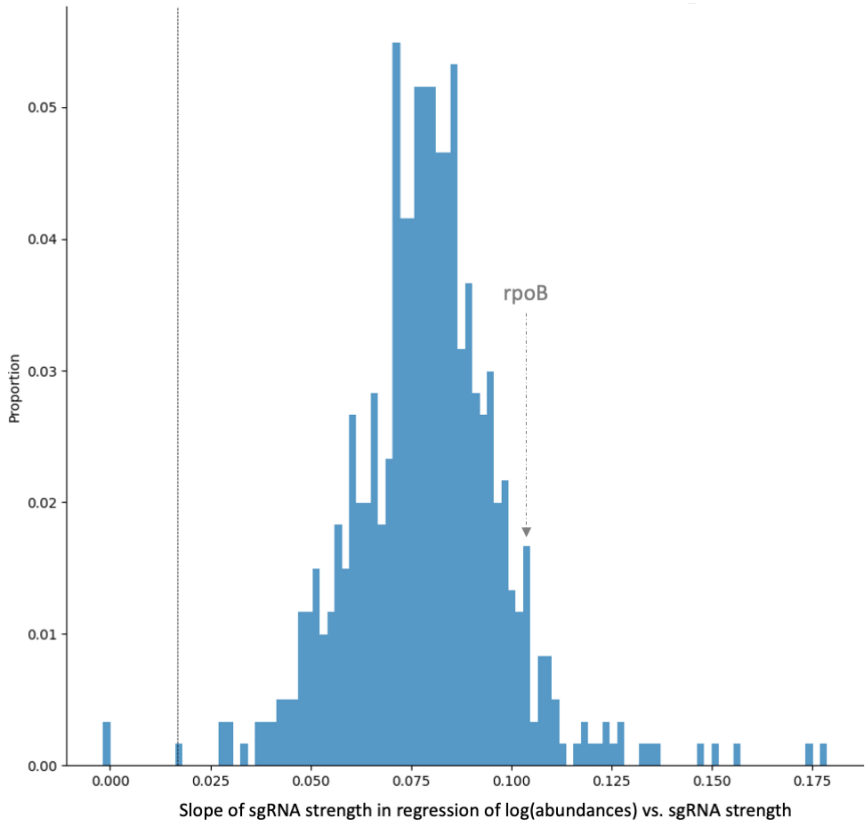
352 sgRNA strength shows a strong correlation with abundance

353 sgRNA strength shows a linear trend with log (abundances) in essential genes. For example, Fig
354 1 illustrates a strong relationship between sgRNA strength and mutant growth suppression for *rpoB*
355 (RNA polymerase). This can be quantified as the slope of the regression: $\log_{10} A_{ik} = B \cdot S_i + C$, where
356 A_{ik} is the relative log abundance of an sgRNA in replicate *k* (counts in +ATC culture divided by counts in -
357 ATC), S_i is the strength of sgRNA *i* in the form of extrapolated LFCs (calculated for the library grown in -
358 ATC in buffer), and C is the intercept. This regression was run on essential genes with at least 20
359 sgRNAs. Non-essential genes were excluded in this analysis since they have fewer sgRNAs in the library
360 and tend not to deplete regardless of concentration or sgRNA strength. As seen in the distribution in Fig
361 4, most of genes show slope greater than 0 (though not all as large as *rpoB*), and nearly all are significant
362 (Wald test, adjusted p-value < 0.05). In all the genes, as sgRNA strength increases (i.e. extrapolated LFCs
363 become more negative), abundances decrease. This demonstrates that there is a direct relationship

364 between sgRNA strength and mutant depletion extending to all essential genes in the genome.

365 Therefore, strength of the sgRNAs is an important covariate of predicting abundances and should be

366 incorporated in the model to accurately identify genes showing depletion in a condition.



367

368 **Fig 4. Distribution of slopes from regression of $\log_{10}$ (abundances) with respect to sgRNA strength, fit**

369 **for the RIF D5 dataset.**

370 For essential genes in the RIF (D5) experiment with at least 20 sgRNAs, we regressed the average log

371 normalized relative abundance at no-drug control samples against the sgRNA strengths (extrapolated

372 LFCs at 25 generations) and plotted a histogram of the coefficients. sgRNAs that are significant are those

373 with slope >= 0.024 (adjusted p-value < 0.05). Most of the slopes are greater than 0 and marked as

374 significant. As sgRNA strength increases for a mutant, abundance decreases, indicating a direct

375 relationship between sgRNA strength and mutant depletion.

376

The CRISPRi-DR model accurately predicts sgRNA abundances from sgRNA strength and drug concentration

For all experiments, the CRISPRi-DR model with both sgRNA strength and concentration as predictors outperforms reduced models. When the model is run on each gene in the ethambutol (EMB D5) experiment, 59.2 % of the 4032 genes show r^2 values (correlation of predicted and observed abundances) of at least 0.5. As expected, these genes include targets of EMB, *embA, embB* and *embC* as well as other cell wall related genes such as the *aft* (arabinofuranosyltransferase) genes.

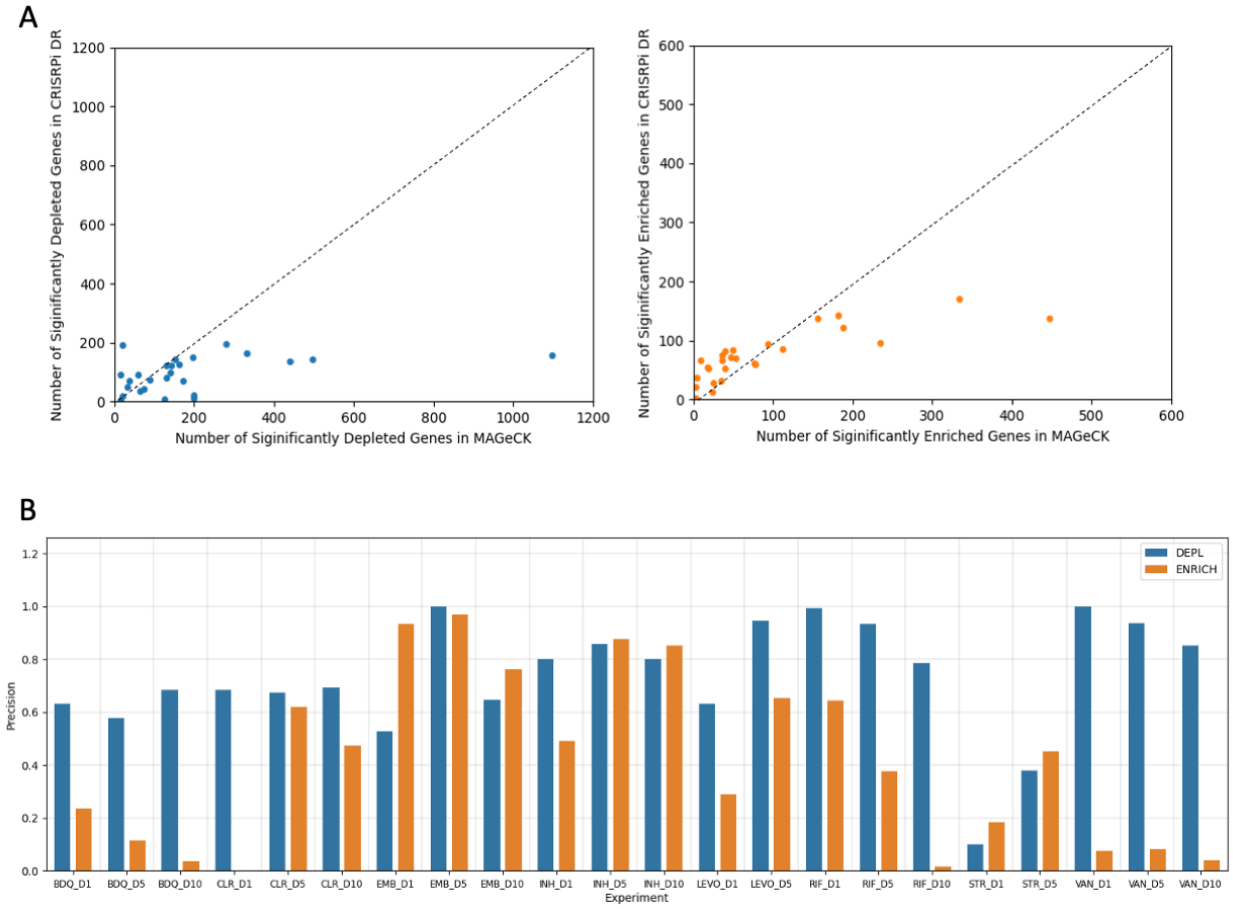
To evaluate the relative importance of the sgRNA strength and drug concentration features to the CRISPRi-DR model, each gene was run through two ablated models: M_d and M_s. The M_d model contained only log concentration as a predictor: $\log\left(\frac{A_{ijk}}{1-A_{ijk}}\right) = B \cdot \log([D_j]) + C$ and the M_s model only contained sgRNA strength as a predictor: $\log\left(\frac{A_{ijk}}{1-A_{ijk}}\right) = B \cdot S_i + C$. In the EMB D5 experiment, only 33.4% of genes fitted with M_s and 8.0% of genes fitted with M_d show r^2 values at least 0.5. *embA, embB* and *embC* do not appear in the either of these sets of significant interactors. The average log-likelihood (LL) of the full model in the EMB D5 experiment is -99.5, whereas the average log-likelihood of M_d is -245.1 and average log-likelihood of M_s is -131.4 (higher LL values represent better fit). When the log-likelihood ratio (LR) test is performed, the LR-statistics show that M_s is an improvement over M_d, and the full model is a greater improvement over both M_d than M_s. In all three models, most of the insignificant genes (adjusted p-value of LR statistic ≥ 0.05) were non-essential genes that do show much depletion regardless of concentration or sgRNA strength. For targets of EMB, *embA, embB* and *embC*, the LR statistic for M_s is higher than M_d and is the highest in the full CRISPRi-DR model. The r^2 values and results of the log-likelihood ratio test indicate the sgRNA strength contributes more strongly to the CRISPRi-DR model than the drug concentration and is the dominant feature for most genes. Additionally,

399 the full CRISPRi-DR model not only provides better fits for a greater quantity of genes than the ablated
400 models, but it also provides betters fits for targets of the drug.
401 The CRISPRi-DR model's improved performance over the reduced models for EMB extends to all
402 drugs tested, as seen in S1 Fig. The dashed line in the plot indicates $r^2 = 0.5$. In all the experiments, the
403 number of genes with fits that have $r^2 > 0.5$ is greater in the M_s model than M_d. The number of genes
404 with fits with $r^2 > 0.5$ is the greatest in the full CRISPRi-DR model. This demonstrates that in all
405 conditions, both concentration and sgRNA strength are needed to make accurate estimates of sgRNA
406 depletion.
407 Some users may not have the resources to run passaging experiments for all sgRNAs in their
408 CRISPRi library to determine sgRNAs strengths empirically, and thus may want to rely on the predicted
409 strengths based on sequence features. To evaluate how much of a difference the predicted strength in
410 place of empirical strength, we fitted the CRISPRi-DR model on all the datasets with predicted strength
411 in place of empirical strength and compared the results. The significant genes reported by the CRISPRi-
412 DR model using predicted strength (based of sequence features) were nearly identical to the significant
413 genes reported by the CRISPRi-DR model using empirical strength (based on passaging). The average
414 overlap of interacting genes detected is 93.3%, with 24 out of 26 datasets having an overlap greater
415 than 90%. Thus, using predicted sgRNA strengths is almost as good as using empirical estimates from
416 passaging.
417

418 CRISPRi-DR and MAGeCK have a high concordance of predicted gene-
419 drug interactions

420 The overall number of significant genes identified by the CRISPRi-DR model is comparable to those
421 reported by MAGeCK, but MAGeCK identifies additional genes that are not detected as significant by the

422 CRISPRi-DR model. MAGeCK and CRISPRi-DR detect about the same number of significantly enriched and
423 depleted genes, typically on the order of tens to a few hundred for any given drug, as shown in Fig 5A.
424 The number of false negatives (significant in MAGeCK but not in CRISPRi-DR) are balanced with the
425 number of false positives (significant in CRISPRi-DR but not in MAGeCK); they are both on similar scales.
426 On average, 57.5% of significant genes in CRISPRi-DR are also significant genes in MAGeCK. However, for
427 some drugs, MAGeCK predicts substantially more hits. For example, MAGeCK finds over 1066
428 significantly depleted genes for VAN (even with the filter of |LFC|>1 applied), whereas CRISPRi-DR finds
429 only 196 significant interactors.

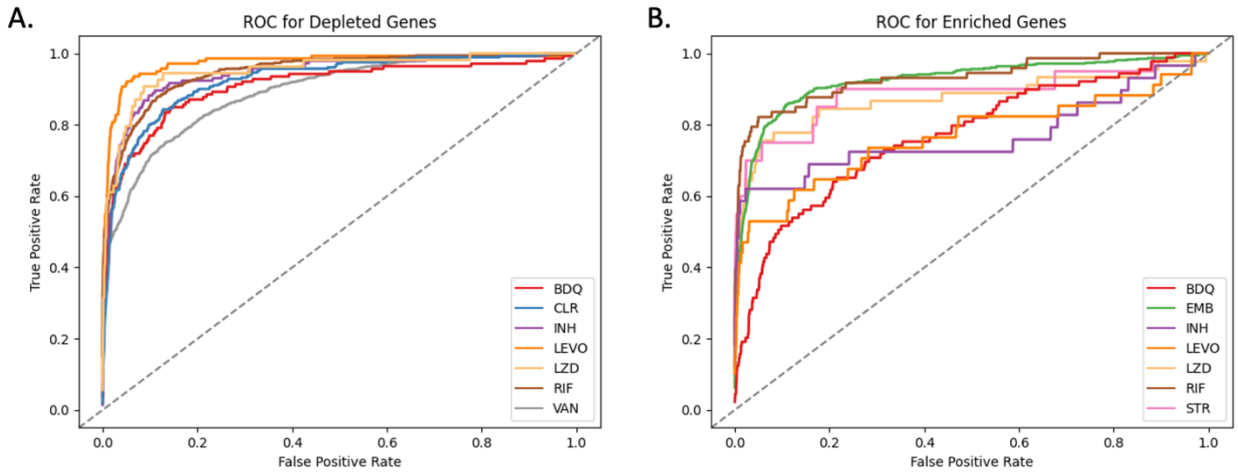


430

431 **Fig 5. Comparison of significant interactions in CRISPRi-DR and MAGeCK.**

432 (A) The number of hits (both enriched and depleted) are slightly greater in MAGeCK than in the CRISPRi-
433 DR model. However, both models produce comparable number of significant genes. The outlier point

434 seen in for the scatterplot comparing depleted genes (top) is for VAN D1. The number of genes reported
435 in the CRISPRi-DR model span a shorter range than the number of genes reported in MAGeCK. (B)
436 Precision of significant genes reported by the CRISPRi-DR model. Overall, the precision of both enriched
437 and depleted hits in the CRISPRi-DR model (compared to MAGeCK) are high. There is a greater overlap in
438 depletion hits than enriched hits. The LEVO D10 and LZD datasets had almost no hits in MAGeCK [see
439 Extended Data Fig 2 in (Li, Poulton et al. 2022)]. As a result, they were excluded from the precision
440 analysis.
441
442 The ranking of genes using the CRISPRi-DR model (using coefficient of concentration dependence, as
443 described above) correlates well with ranking of genes in MAGeCK. For each of the 9 drugs tested,
444 Receiver Operator Characteristic (ROC) curves were calculated for the D1 (1 day) pre-depletion datasets,
445 seen in Fig 6. The average areas under curves (AUC) in Fig 6A is 0.95, indicating that the genes reported
446 in MAGeCK across all concentrations are ranked highly in the CRISPRi-DR model. For instance, 70.0% of
447 the top-100 ranked depletion genes in MAGeCK appear in the top-100 ranked depletion genes in the
448 CRISPRi-DR model. The areas under the curves in Fig 6B for enriched hits are lower than of Fig 6A , with
449 an average of 0.83.



450

451 **Fig 6. ROC curves comparing gene rankings in MAGeCK and CRISPRi-DR for enriched hits (A) and**
452 **depleted hits (B) in 1 day pre-depletion experiments.**
453 The recovery of the depleted hits outperforms the recovery of enriched hits, showing that MAGeCK and
454 the CRISPRi-DR model rank depleted genes similarly. EMB and STR are excluded in the ROC analysis of
455 depleted genes and CLR and VAN are excluded in the analysis of enriched genes. These libraries had too
456 few significant genes reported by MAGeCK in their respective categories to yield meaningful ROC curves.
457 The lower performance of the enrichment gene rankings may be due to a few reasons, including noise.
458
459 The discrepancy between interactions detected by MAGeCK and CRISPRi-DR for enriched hits can be
460 observed as an imbalance between false negatives and false positives in the confusion matrices (see S2
461 Table). Many genes with significant enrichment by MAGeCK are not called significant by CRISPRi-DR.
462 This imbalance can be quantified as *precision* (calculated as TP/(TP+FP), or fraction of true positives
463 (defined by MAGeCK) vs all positives (predicted by CRISPRi-DR). The precision of these CRISPRi-DR calls
464 can be seen in Fig 5B. The average overlap of significantly depleted genes is 73.3%, whereas the average
465 of significantly enriched genes is nearly half that, at 41.7%. The significant genes reported using the
466 CRISPRi-DR model are largely a subset of the genes reported by MAGeCK, with a smaller overlap of
467 significant enriched genes than significant depleted genes. This lower concordance of the two models
468 for *enriched* hits shows that MAGeCK may be selecting genes with large variations, deceptively seeming
469 to be significant interactions, that the CRISPRi-DR model does not. This might be attributable to the
470 greater susceptibility of MAGeCK to noise in barcode counts, which is higher for some enriched genes
471 (discussed below).
472

473 **CRISPRi-DR model correctly detects genes known to interact with anti-**

474 **tubercular drugs.**

475 When genes are ordered by coefficients of the slope representing the dependence of abundance on

476 drug concentration from the CRISPRi-DR model, genes for existing anti-mycobacterial drugs are ranked

477 highly, as expected (Table 1). The more positive a gene's coefficient is, the higher the gene's enrichment

478 ranking, and the more negative a gene's coefficient is, the higher it's depletion ranking.

479

480 **Table 1 : Ranking of Select Genes using the CRISPRi-DR model in 1 Day pre-depletion of treated**

481 **libraries.**

Drug	Gene	D1 Depletion Ranking	D1 Enrichment Ranking
BDQ	*atpA*	11	4022
BDQ	*atpB*	6	4027
BDQ	*atpC*	35	3998
BDQ	*atpD*	12	4021
BDQ	*atpE*	23	4010
BDQ	*atpF*	7	4026
BDQ	*atpG*	9	4024
BDQ	*atpH*	8	4025
BDQ	*mmpL5*	2	4031
CLR	*RVBD3579c*	35	3998
CLR	*erm(37)*	1	4032
INH	*inhA*	6	4027
INH	*ahpC*	2	4031
INH	*katG*	4031	2
INH	*ndh*	4029	4
EMB	*embA*	4	4029
EMB	*embB*	5	4028
EMB	*embC*	12	4021
LEVO	*gyrA*	3834	199
LEVO	*gyrB*	3967	66
LZD	*erm(37)*	3994	39
LZD	*tsnR*	4032	1
RIF	*rpoB*	108	3925
RIF	*rpoC*	148	3885
STR	*ettA*	4023	10
STR	*gidB*	4022	11

482 For each drug, the CRISPRi-DR model is run on each gene (using data from D1). The coefficient for the

483 slope of concentration dependence (β_c) is extracted from the fitted regression and used to rank the

484 genes in both increasing order (for depletion) and inversely (for enrichment). Green reflects results

485 consistent with expectations based on knowledge of known gene-drug interactions

486

487 Genes that are known to be involved in the target mechanism of a drug should have a high
488 depletion rank, i.e., show a negative slope, indicating that as concentration increases, abundance for the
489 given depletion-mutant decreases. This can be seen in S1 Table, in the ranking for genes using the
490 CRISPRi-DR model. *embA, embB,* and *embC* (subunits of the arabinosyltransferase, target of ethambutol,
491 EMB) rank within the top 100 depleted genes for all three pre-depletion conditions for EMB. They rank
492 the highest in D1 and the lowest in D10. This can be attributed to the fact that by D10 genes are already
493 quite depleted, even at concentration 0, increasing noise, and making it difficult to pick up on depletion
494 signals over increasing concentration. Therefore, the ranking of relevant genes in D1 was assessed in this
495 analysis (Table 1). In RIF, target genes *rpoB, rpoC* are ranked within the top 150 genes. Significant
496 negative interacting genes for RIF also include many cell wall related genes such as *ponA2, rodA, ripA,*
497 *aftABCD, embABC,* etc., consistent with recent studies that show RIF exposure (or mutations in *rpoB*)
498 leads to various cell wall phenotypes [10-12]. Similarly, the targets of bedaquiline (BDQ), the 8 ATP
499 synthase genes (*atpA-atpH*, subunits of F0F1 ATP synthase), along with efflux pump *mmpL5,* are ranked
500 within the top 40 depleted genes in BDQ. In levofloxacin (LEVO), *gyrA* and *gyrB* (subunits of the DNA
501 gyrase, the target of fluoroquinolones) are observed to be enriched. The reason that depletion of this
502 drug target leads to enrichment of mutants (hence a growth advantage, rather than the expected
503 growth impairment) is likely due to reduced generation of double-stranded breaks in the DNA and other
504 toxic intermediates as a side-effect of inhibiting the gyrase, an effect that has been observed in *E. coli*
505 [13]. The significantly depleted genes in vancomycin (VAN) show significant enrichment for the cell
506 wall/membrane/envelope biogenesis pathway (as defined by in COG pathways [14]) using Fischer's
507 Exact Test This follows previous studies that show cell wall genes are targets of vancomycin [15, 16],
508 which binds to peptidoglycan in the cell wall. For clarithromycin (CLR), an inhibitor of translation,
509 *Rv3579c* and *erm(37)* are observed as hits. *Erm(37)* adds a methyl group on the A2058/G2099 nucleotide
510 in the 23S component of the ribosome, the same position to which CLR attempts to bind [17]. This

511 natively increases tolerance to CLR in *Mtb*. As this gene is depleted, CLR has greater opportunity to bind,
512 reducing the bacillus' natural tolerance to the drug. Following this observation, e*rm(37)* has a depletion
513 rank of #1 in the CLR D1 condition. *Rv3579c* is another methyltransferase with a similar function that
514 ranks highly (#35) in CLR.

515 In contrast to methylation inhibiting the binding of CLR, there are ribosome methyltransferases
516 where methylation facilitates binding of a drug. Mutants for these genes would be expected to show a
517 high enrichment rank in presence of drug. For instance, streptomycin (STR) interferes with ribosomal
518 peptide/protein synthesis by binding near the interaction of the large and small subunits of the
519 ribosome [18]. Two relevant genes that influence the binding of STR include *gidB* and *Rv2477c/ettA*.
520 *gidB* is an rRNA methyltransferase that methylates the ribosome at nucleotide G518 of the 16S rRNA,
521 the position at which STR interacts [19], increasing native affinity for STR. This is consistent with the
522 observation that one of the most common mutations in STR-resistant clinical isolates is loss of function
523 mutations in *gidB* [20]. *Rv2477c* is a ribosome accessory factor, also known as *ettA*, which is an ATPase
524 that enhances translation efficiency. It has also recently been shown to bind the ribosome near the P-
525 site (peptidyl transfer center), potentially interfering with binding of aminoglycosides [21], and loss-of-
526 function mutations observed in drug-resistant clinical isolates of *M. tuberculosis* have shown to confer
527 resistance to STR [2]. The ranking of both genes using the CRISPRi-DR model are within the top 12
528 enriched genes in STR. For linezolid (LZD), relevant genes identified are *erm(37)* and *tsnR. tsnR* is an
529 rRNA methyltransferase, analogous to *gidB* and results in tolerance to LZD in a similar manner as *gidB*
530 does for STR [2]. Following this expectation, *tsnR* has an enrichment ranking of #1 in LZD. Whereas
531 depletion of *erm(37)* gives tolerance to CLR, it increases sensitivity to LZD. The nucleotides that *erm(37)*
532 methylates in the 23S RNA are proximal in 3D space to where mutations conferring LZD-resistance are
533 found, which both lie in the PTC (peptidyl-transfer center) of the ribosome [22].

534 For isoniazid (INH), there are multiple relevant genes identified by CRISRPi-DR, including *inhA, ahpC*,
535 *ndh* [23], and *katG* [24]. *inhA* (enoyl-ACP reductase, in mycolic acid pathway) is an essential gene that is
536 the target of INH, and *ahpC* (alkyl hydroperoxide reductase) responds to the oxidative effects of
537 isonicotinic radicals in the cells. Therefore, as dosage of the drug increases, the abundances of the
538 mutants of these genes should decrease. These genes are in the top 10 highest ranked depletion genes
539 for INH (see Table 1). In contrast, *katG* and *ndh* are found among the top 5 enriched hits, exhibiting
540 increased survival when the proteins are depleted. KatG (catalase) is the activator of INH, and the most
541 common mutations in INH-resistant strains occur in *katG*, decreasing activity [25]. *Ndh* (type II NADH
542 reductase) mutants have also been shown to decrease sensitivity to INH by shifting intracellular NADH
543 levels (needed for INH-NADH adduct formation), and mutations in *ndh* have been shown to be defective
544 in target enzyme (NdhII) activity [23], which is consistent with the observation in the CRISPRi data that
545 depletion of *ndh* leads to increase survival in the presence of INH.
546

547 The CRISPRi-DR model is less sensitive to noise than MAGeCK

548 MAGeCK's greater sensitivity to noise could be a reason that the CRISPRi-DR model shows lower
549 consistency with MAGeCK for enriched hits (e.g. lower AUC in Fig 6B than Fig 6A). There is some noise in
550 these experiments due to variability in sequencing barcode counts across replicates. This can
551 differentially affect the accuracy of predictions of gene-drug interaction made by these models. Three
552 replicates were available for each measurement, i.e., 3 different counts estimating the relative
553 abundance of each sgRNA in the presence of a drug at a given concentration. Coefficient of variation
554 (CV) can be used to measure relative consistency across these observations for each measurement,
555 which in turn can be used to evaluate MAGeCK and the CRISPRi-DR model's sensitivity to noise in the
556 raw data.

557 For each sgRNA s_i the coefficient of variation (CV) was calculated across the relative abundances for

558 the 3 replicates for each concentration © in drug (D) ($CV_{D,C,i} = \frac{\sigma(i)}{\mu(i)}$), where $\sigma(i)$ is the standard

559 deviation of the 3 relative abundances in concentration C and $\mu(i)$ is the mean. In Fig 7A, the

560 $CV_{D=DMSO,C=0,i}$ (CV of abundances in concentration 0) for a random subset of sgRNAs (~5%) in an ATC-

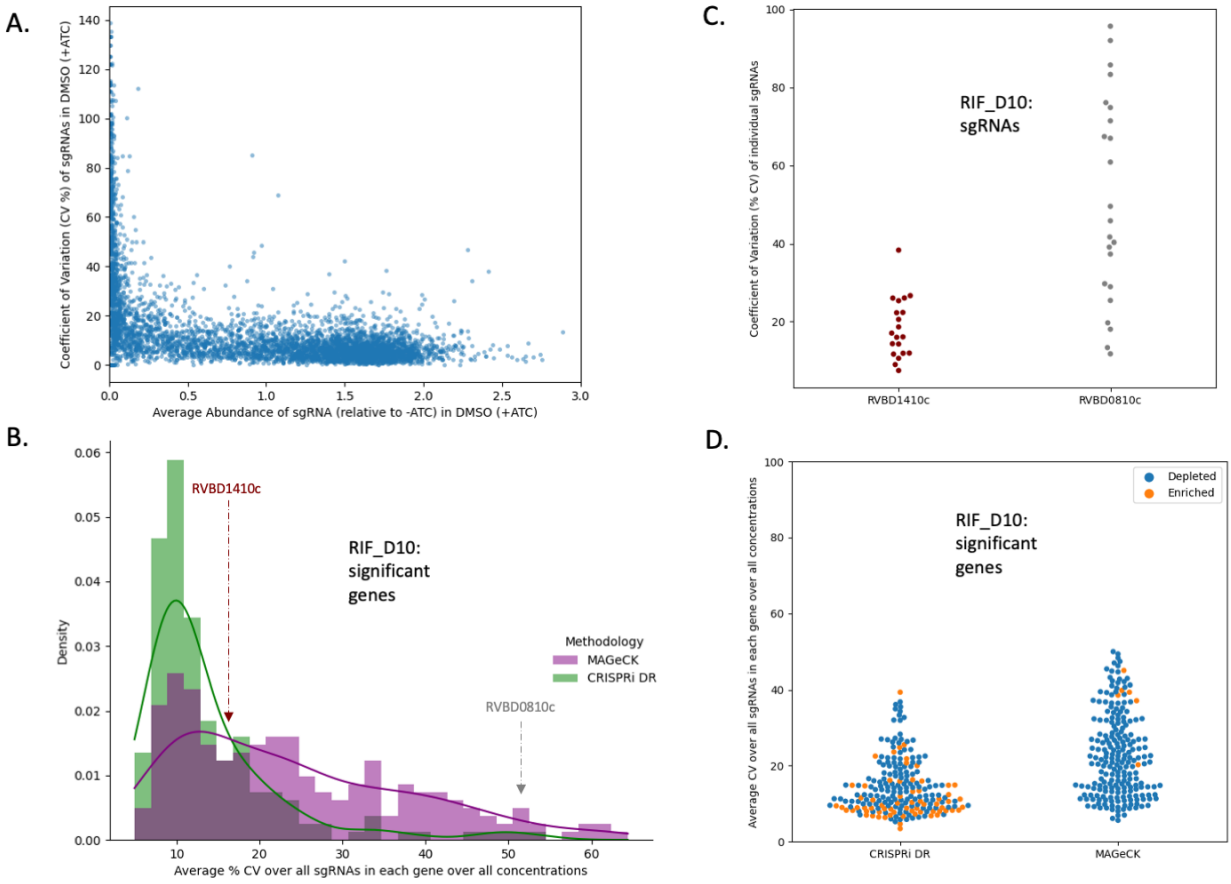
561 induced no-drug condition is compared to the average abundance. For sgRNAs of medium to high

562 abundance, the CV is fairly constant at approximately 10%. However, as the average abundance

563 decreases (below relative abundance of 0.1), CV value increases substantially to 140%. If a gene contains

564 multiple such sgRNAs with high CV values, then the variation may be misconstrued as a genetic

565 interaction by a noise-susceptible methodology.



566

567 **Fig 7. CRISPRi-DR model shows less sensitivity to noise than MAGeCK.** (A) Comparison of average

568 relative abundance and average CV across replicates in no-drug control samples (+ ATC) for a sample of

569 sgRNAs : For each sgRNA, we looked at the average CV of sgRNAs in the 3 control replicates against the
570 average abundance of the sgRNA across those replicates. The lower the average abundance, the greater
571 the noise present for the sgRNA. (B) Distribution of average CV of gene for significant genes in MAGeCK
572 and significant genes in CRISPRi-DR in RIF D10 : The distribution of average CV of significant genes in
573 CRISPRi-DR model is more skewed and has a peak at CV ≈ 10%. Although most significant genes in
574 MAGeCK show an average CV around 15%, there are quite a few genes with higher average CVs not
575 found significant by the CRISPRi-DR model. (C) Coefficient of Variation (CV) of each sgRNA in two genes
576 with similar number of sgRNAs for a library treated with RIF D10 : *Rv1410c* is significant in both
577 methodologies and *Rv0810c* significant in MAGeCK but not in CRISPRi-DR. The majority of CV values for
578 sgRNAs in *Rv1410c* is around 20%. Although both genes have about 20 sgRNAs, *Rv0810c* shows 8 sgRNAs
579 whose CV values exceed 60.5%, which is the maximum CV present in *Rv1410c.* (D) Distribution of
580 average CV for enriched and depleted significant genes in MAGeCK and CRISPRi-DR in a RIF D10 library.
581 This plot shows the distribution plot of Panel B, separated by depletion and enriched significant genes.
582 The average CV values for significant genes in the CRISPRi-DR model are low for both enriched and
583 depleted genes. As seen in Panel B, significant genes in MAGeCK show low average CV, but they also
584 show high average CV. Although there is a substantially lower number of significantly enriched in
585 MAGeCK, they still show a large amount of noise compared the significantly enriched genes in CRISPRi-
586 DR model.
587
588 The average noise in a gene *g* for a given drug D can be quantified as the average $CV_{D,C,i}$, for all
589 concentrations C and all sgRNA*s* in the gene ($\overline{CV_D}(g)$). Therefore, $\overline{CV_D}(g)$ reflects the measure of
590 overall noise present in a gene in a drug D. The distribution of $\overline{CV_D}(g)$ in RIF D10 for the 215 total
591 significant genes (enriched and depleted combined) in the CRISPRi-DR model and in 218 total significant
592 genes (enriched and depleted combined over all concentrations) in MAGeCK can be seen in Fig 7B. The

593 distributions for both methodologies share a mode at about $\overline{CV_D}(g) \approx 10\%$. The distribution of $\overline{CV_D}(g)$
594 for significant genes in MAGeCK has a fatter tail than the distribution of $\overline{CV_D}(g)$ for significant genes in
595 the CRISPRi-DR model. This trend is seen not only in RIF D10, but across all the experiments conducted
596 (See S2 Fig). This indicates that although MAGeCK is identifying genes with low noise (like the CRISPRi-
597 DR model), it is also detecting many genes with high noise that the CRISPRi-DR model is not.

598 An example of such a gene is *Rv0810c*. The gene has 22 sgRNAs and has a $\overline{CV_D}(g)$ value (average CV
599 over sgRNAs in a gene) of 51.4%, one of the highest measures in the RIF D10 experiment. In RIF D10, it is
600 reported to be significantly depleted only in MAGeCK and not in the CRISPRi-DR model. The distributions
601 of the CV values for each sgRNA are compared to those of *Rv1410c* in Fig 7C. *Rv1410c* has 20 sgRNAs, an
602 $\overline{CV_D}(g)$ of 16.3% and is reported to be significantly depleted in both MAGeCK and the CRISPRi-DR
603 model. Although both genes have some sgRNAs with low CVs (below 40%), *Rv0810c* shows 8 sgRNAs
604 with CVs of at least 60.5%, which is the maximum CV of sgRNAs in *Rv1410c*. The CRISPRi-DR model
605 considers the abundances at all concentrations, whereas MAGeCK compares each concentration to the
606 baseline independently. Therefore, if sgRNAs have a high CV value at a particular concentration, they
607 can be picked up as a significant genetic interaction by MAGeCK. The average relative abundance for the
608 3 replicates at concentration 0 for all sgRNAs in *Rv0810c* is 0.19, whereas the average relative
609 abundance in *Rv1410c* for the same is 1.08. As Fig 7A shows, *Rv0810c* falls in the low abundance/high
610 noise section of the graph, with an average sgRNA no-drug CV of 47.9%, whereas *Rv1410c* falls in the
611 low noise section of the graph, with an average sgRNA no-drug CV of 11.2%. This demonstrates that
612 MAGeCK reports genes such as *Rv0810c* with low abundances resulting in large $\overline{CV_D}(g)$, which the
613 CRISPRi-DR model does not, i.e., MAGeCK is more suspectable to noise than the CRISPRi-DR model.

614 Furthermore, the $\overline{CV_D}(g)$ for significantly enriched genes in MAGeCK is higher than the $\overline{CV_D}(g)$ for
615 significantly depleted genes. As seen in Fig 7B, both methodologies detect genes with $\overline{CV_D}(g) \approx 10\%$ in
616 RIF D10. The $\overline{CV_D}(g)$ values for both significantly depleted and enriched genes in the CRISPRi-DR model

617 are close to this value (Fig 7D). MAGeCK detects significantly depleted genes at around this value, but
618 also detects genes with much larger $\overline{CV_D}(g)$ values. Although there are fewer significantly enriched
619 genes reported in MAGeCK than CRISPRi-DR, they show a larger amount of noise compared the
620 significantly enriched genes detected by CRISPRi-DR. Since the significantly enriched genes in MAGeCK
621 show higher noise than either significantly enriched or significantly depleted genes in the CRISPRi-DR
622 model, it might partially explain the lower levels of overlap (AUC) seen in the ROC curves for enriched
623 genes in Fig 6B.
624

625 # Simulation

626 The sensitivity and accuracy of the CRISPRi-DR model and MAGeCK was assessed under different
627 sources of noise using simulated barcode counts sampled from the negative binomial distribution [26],
628 with means at different concentrations determined by the dose-response model (Eqn. 3). sgRNAs and
629 their empirical strength estimates from a previous study [2] were used to simulate the combined effects
630 of CRISPRi depletion and exposure to a virtual inhibitor at four concentrations (1uM, 2uM, 4uM, and
631 8uM), with three replicates each. The aim was to determine how noise within and between
632 concentrations affects the performance of each method. Detailed information on the simulation is
633 provided in the S1 File.
634 Four datasets (LL, LH, HL, and HH) were simulated by varying two noise parameters: σ_B
635 (variability of abundances between concentrations) and *p* (variability of replicates within a
636 concentration, parameter of the negative binomial distribution). 50 genes were randomly selected for
637 negative interactions (consistent depletion effects) and another set of 50 genes for positive interactions
638 (positive biased trend). The negative interactions were simulated using the dose-response formula (Eqn.
639 3) above, whereas the positive interactions and non-interacting sgRNAs were simulated using small
640 random slopes to reflect concentration dependent effects. CRISPRi-DR and MAGeCK were run ten times

641 each on these 4 scenarios. MAGeCK was run independently for each drug concentration (2uM, 4uM,
642 8uM, compared to a no-drug control), while CRISPRi-DR was performed on all four concentrations
643 simultaneously.

644 Both methods displayed high recall in the LL scenario (lowest noise) (CRISPRi-DR : 95.4%,
645 MAGeCK : 84.6%) but their recall rates are slightly degraded in the HH scenario (highest noise) (CRISPRi-
646 DR : 59.7%, MAGeCK : 70.5%). The difference in sensitivity to noise is more apparent in the *precision* of
647 the two methods. In the HH scenario, MAGeCK generates nearly four times as many false-positive
648 predictions (463.3), leading to a very low precision of approximately 13.3%, whereas CRISPRi-DR's
649 precision is 36.5%, with 104.2 false positives. This indicates that MAGeCK is prone to classifying non-
650 interacting genes as hits when noise is high, likely due to stochastic count fluctuations at individual drug
651 concentrations that may not be observed at other concentrations. Comparatively, CRISPRi-DR relies
652 more on consistent trends in abundance across concentrations, and thus makes less erroneous false
653 positive predictions. Notably, the consistent trends in abundance detected by this regression-based
654 model are not required to change perfectly linearly with increasing $\log_2$ drug concentration. Rather, as
655 long as, there is a general trend (increasing or decreasing) across concentrations, then the gene's slope
656 coefficient (concentration dependence) can still be significant. For example, abundances for some
657 sgRNAs may drop off sharply at either end of the concentration range. Several examples of sgRNAs with
658 these patterns are shown in S1 File.

659 To assess the impact of profiling a CRISPRi library at multiple concentrations on the performance
660 of CRISPRi-DR and MAGeCK, we conducted the simulation above with high-noise settings (HH) and
661 varying numbers of drug concentrations (1, 2, or 3) for 10 iterations each. The recall of both methods
662 held fairly constant as concentrations were added. However, increasing the number of concentration
663 points caused a significant decrease in the precision of MAGeCK from 21.2% to 13.2%. While MAGeCK
664 shows susceptibility to false positives when evaluating only a single concentration point, this effect was

665 amplified with more concentrations. This accumulation of errors explains the decrease in precision with
666 additional concentration points. In contrast, CRISPRi-DR is more robust with respect to false-positive
667 errors. By incorporating data from all available concentrations and identifying significant trends,
668 CRISPRi-DR maintained higher precision that did not diminish with the addition of more concentration
669 points.
670

671 # Discussion

672 CRISPRi can be used to conduct CGI experiments through several approaches. One approach is
673 to modulate expression of dCAS9 (with an active nuclease function) to control expression of the target
674 gene at various levels. This allows for the quantification of phenotype (e.g. growth rate in presence of
675 inhibitor) as a function of expression level of a target gene. Typically, sgRNAs are selected that are
676 validated to strongly bind their target genes and provide strong depletion [3]. Another strategy to
677 generate mutants with graded phenotypes is by using parent sgRNAs that are progressively weakened
678 through mutations [27]. Mutants with knock-down of a particular gene that exhibit a statistically
679 significant depletion-dependent shift in MIC are deemed interactions. Alternatively, one can use a
680 catalytically-dead dCAS9 (since binding to gene targets is sufficient to block transcription), and rely
681 instead on a range of sgRNAs with varying strength (which can be barcoded separately and quantified
682 independently) to evaluate depletion-dependent fitness effects [1]. In these CRISPRi libraries, stronger
683 sgRNAs better inhibit expression of targets genes and cause greater protein depletion, which can better
684 reveal interactions with drug treatment (through synergies). Inclusion of multiple sgRNAs with different
685 strengths for each target gene can be used to test for expression-dependent sensitization to inhibitors.
686 The availability of CRISPRi data for multiple sgRNAs of different strengths for each target gene
687 presents new challenges for statistical analysis for CGI experiments. In previous work [5], we showed

688 that regressing the relative abundances of mutants in hypomorph libraries over concentrations (on log-
689 scale) can be used to improve detection of CGIs. This regression approach captured dose-dependent
690 behavior, i.e. genes whose decreased expression caused either suppressed or enhanced fitness that
691 increases in magnitude with drug concentration (i.e. exhibits a trend, which is important for statistical
692 robustness). The CRISPRi-DR method described in this paper extends this previous work by showing
693 how effects of both drug concentration and sgRNA strength can be accommodated in the same model.
694 What we are looking for, ideally, is genes that exhibit synergistic behavior with a drug, where depletion
695 of a target protein induces excess depletion (or enrichment) of the mutants grown in the presence of an
696 inhibitor, and this effect is concentration-dependent (exhibits dose-response behavior).
697 In theory, both CRISPRi depletion of essential genes and exposure to antibiotics should impair
698 growth of CRISPRi mutants (at least for depletion of essential genes). One might expect to observe a
699 depletion effect due to either increasing sgRNA strength, or drug concentration, each producing
700 regression "slopes" (in log-transformed space), with slopes for sgRNAs targeting non-essential genes
701 being expected to be flat, regardless of sgRNA strength. However, we observed that sgRNA strength
702 and concentration effects are not independent - they interact in a non-linear way. sgRNAs that are too
703 weak do not produce enough depletion of a drug target to cause sensitization (MIC shift), and sgRNAs
704 that are too strong deplete a mutant to such low abundances that concentration-dependent effects are
705 difficult to quantify. Often, there is a "sweet spot", or an intermediate sgRNA strength which maximizes
706 the concentration-dependent effect (which could be different for each gene). Mathis et al. [27]
707 suggested that dose-response behavior could be modeled with a classic Hill equation, where the
708 number of mutations between the sgRNA sequence and target gene was used as a proxy for strength in
709 a logistic function fitted to growth rate. However, this covariate was not explicitly combined with
710 environmental variables (such as drug concentration) in their model. Our CRISPRi-DR model
711 incorporates both sgRNA strength and drug concentration as parameters, and reproduces the non-linear

712 interaction between them, where the "slopes" for the effect of drug concentration on relative
713 abundance of mutants can be larger in magnitude for sgRNAs of intermediate strength, while being
714 flatter (slopes closer to 0) for sgRNAs of high or low strength.

715 The strength with which different sgRNAs cause a growth phenotype depends on various factors
716 affecting how well they bind to and suppress transcription of their genomic targets. First, the strength
717 depends on how well the guide RNA matches the optimal PAM sequence, in order to be recognized by
718 and recruit the dCAS9 nuclease [6]. Second, it depends on the length (typically 17-24 bp) and GC
719 content of the complementary region that hybridizes with the chromosome. These sequence factors
720 can be combined to make a predictive model of the effect on expression of target proteins, which has
721 been shown to predict sgRNA strength with moderate accuracy (R^2=0.74) (see Fig 2C in [1]). For greater
722 accuracy, sgRNA strength can also be empirically quantified by conducting a passaging experiment. By
723 inducing expression of the dCAS9 and measuring growth-rate over several generations, the strength of
724 each sgRNA can be fit using a piecewise linear model and extrapolated to an implied depletion at a
725 constant number of generations (e.g. estimated log2-fold-change of abundance in +ATC vs -ATC at 25
726 generations) [1]. However, for some labs that might prefer to use predicted strengths instead of running
727 passaging experiments, we showed that using predicted strengths from sequence features with CRISPRi-
728 DR in place of empirical strength produces results that are nearly as good.

729 In this paper, we showed that this non-linear interaction between sgRNA strength and drug
730 concentration can be modeled using an augmented Dose-Response equation, in which terms for both
731 effects are included. By fitting the parameters in this equation to CRISPRi data from a CGI experiment
732 (normalized barcode counts), one can estimate the degree to which depletion of a given gene sensitizes
733 cells to an inhibitor, and thereby identify CGIs. While various computational methods exist for fitting
734 non-linear equations, such as the Levenberg–Marquardt algorithm [28], we chose to linearize the
735 modified Hill equation by applying a log-sigmoid transform. The transformation enables us to express

736 the equation in a linear form, where the parameters (EC_{50}, Hill slopes, etc.) appear as coefficients of
737 linear terms or constants. Consequently, we can use ordinary least-squares regression (OLS) to fit the
738 model to the CRISPRi dataset.

739 An alternative approach for analyzing CRISPRi data is MAGeCK, which is a based on the DeSeq2
740 method for analyzing RNA-seq data [29]. It calculates LFCs for each sgRNA at each individual drug
741 concentration and combines them using RRA (robust rank aggregation) to identify significant CGIs.
742 When MAGeCK was developed, exploiting the spectrum of sgRNA strengths was not anticipated, so the
743 sgRNAs in a gene are not treated differentially, and the RRA relies on the expectation that at least a
744 subset of sgRNAs will be strong enough to elicit suppression of the target gene and produce a consistent
745 effect on fitness (enrichment or depletion of mutant abundance), which will be detected as a signal
746 through rank aggregation, i.e. several sgRNAs for a gene having exceptionally high or low LFCs.

747 In principle, one could imagine incorporating the number of days of pre-depletion into the
748 regression approach of CRISPRi-DR. It is often observed that a longer pre-depletion period increases the
749 sensitivity of the experiment and synergy with drug. However, we elected to treat the days of pre-
750 depletion independently, to facilitate the comparison with the analysis in Li, et al [2]. In retrospect, a
751 single day of pre-depletion (D1) has proven adequate for detecting known interactions in most CGI
752 experiments conducted thus far. MAGeCK-MLE is an extension of MAGeCK that can incorporate
753 additional covariates such as days of pre-depletion into the generalized linear model [30]. However, the
754 maximum likelihood parameter estimation process used by MAGeCK-MLE can be time-consuming.
755 CRISPRi-DR provides several advantages over MAGeCK. First, it explicitly incorporates sgRNA strengths
756 as a covariate in the model, taking advantage of this useful information. Second, CRISPRi-DR integrates
757 data over multiple concentrations via regression. This provides enhanced statistical robustness. In
758 contrast, MAGeCK analyzes each drug concentration independently, comparing them to a no-drug
759 control to compute LFCs. But with any single concentration point, there is a risk of detecting false

760 positives (due to noise), which could cause spurious fluctuations in barcode counts, making LFCs possibly
761 appear significant. The susceptibility to noise was evident in the experimental data as predictions made
762 by CRISPRi-DR differed from MAGeCK more on datasets with higher coefficients of variation (S2 Fig).
763 Ideally, it is better to collect data over multiple concentrations for CGI experiments, because it is difficult
764 to know ahead of time what concentration will be optimal to test for each drug. While choosing the MIC
765 for single-point assays might sound reasonable, the actual potency in the CRISPRi experiment could shift
766 due to expression of the dCAS9, inoculation effects, etc. Hence, CGI data is usually collected over a
767 range of concentrations, with the hope that one or more of them will be near the inhibition-transition
768 point. Furthermore, it is not always the case that the highest concentration should be the most
769 informative one for detecting CGIs, as it might cause too much growth inhibition, making it difficult to
770 assess dose-dependent behavior.
771 A simplistic way to use MAGeCK with CGI data collected over multiple drug concentrations is to
772 evaluate each concentration independently, and then combine selected hits (significant genes) using a
773 policy such as taking the union [2]. However, our simulation results showed that this strategy is
774 susceptible to accumulating false positive hits (i.e. non-interacting genes that achieve statistical
775 significance), resulting in low precision. In fact, in previous experiments with a CRISPRi library in *Mtb*,
776 MAGeCK often identified hundreds of genes (and in some cases, up to one-quarter of the genome) as
777 potential interactions for certain antibiotics. While it is true that a variety of genes could interact with a
778 drug directly or indirectly (not just the drug target), revealing multiple complex drug-tolerance and
779 stress-response pathways, it is implausible that there will be hundreds of genuine interactions for most
780 inhibitors. The CRISPRi-DR approach addresses this issue by requiring that apparent interactions
781 (depletion or enrichment) at one concentration be consistent with trends in abundance at other
782 concentrations. The abundance does not have to change in a perfectly linear way over the
783 concentration range (which is helpful, because sometimes the largest effect occurs at the edge of the

784 range, like dropping off a cliff, due to uncertainty about the optimal concentration), but large
785 fluctuations in abundance in the middle of the range, or in opposite directions at different
786 concentrations, will generally get filtered out as insignificant by CRISPRi-DR. Thus, incorporating data
787 from sgRNAs of different strength over multiple concentrations via the modified Dose-Response model
788 make CRISPRi-DR more noise-tolerant and robust for detecting chemical-genetic interactions.
789

790 # Acknowledgments

791 This work was supported by NIH grant P01 AI143575 (TRI, JR, and DS) and by grant INV-004761
792 from the Bill and Melinda Gates Foundation (DS and TRI). The funders had no role in study design, data
793 collection and analysis, decision to publish, or preparation of the manuscript.

794

795 # References

796 1. Bosch B, DeJesus MA, Poulton NC, Zhang W, Engelhart CA, Zaveri A, et al. Genome-wide gene expression
797 tuning reveals diverse vulnerabilities of M. tuberculosis. Cell. 2021;184(17):4579-92 e24. Epub 20210722.
798 doi: 10.1016/j.cell.2021.06.033. PubMed PMID: 34297925; PubMed Central PMCID: PMCPMC8382161.
799 2. Li S, Poulton NC, Chang JS, Azadian ZA, DeJesus MA, Ruecker N, et al. CRISPRi chemical genetics and
800 comparative genomics identify genes mediating drug potency in Mycobacterium tuberculosis. Nat
801 Microbiol. 2022;7(6):766-79. Epub 20220530. doi: 10.1038/s41564-022-01130-y. PubMed PMID:
802 35637331; PubMed Central PMCID: PMCPMC9159947.
803 3. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al. A Comprehensive, CRISPR-based
804 Functional Analysis of Essential Genes in Bacteria. Cell. 2016;165(6):1493-506. Epub 20160526. doi:
805 10.1016/j.cell.2016.05.003. PubMed PMID: 27238023; PubMed Central PMCID: PMCPMC4894308.

806 4. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes
807 from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15(12):554. doi: 10.1186/s13059-
808 014-0554-4. PubMed PMID: 25476604; PubMed Central PMCID: PMCPMC4290824.
809 5. Dutta E, DeJesus MA, Ruecker N, Zaveri A, Koh EI, Sassetti CM, et al. An improved statistical method to
810 identify chemical-genetic interactions by exploiting concentration-dependence. PLoS One.
811 2021;16(10):e0257911. Epub 20211001. doi: 10.1371/journal.pone.0257911. PubMed PMID: 34597304;
812 PubMed Central PMCID: PMCPMC8486102.
813 6. Rock JM, Hopkins FF, Chavez A, Diallo M, Chase MR, Gerrick ER, et al. Programmable transcriptional
814 repression in mycobacteria using an orthogonal CRISPR interference platform. Nature Microbiology.
815 2017;2(4):16274. doi: 10.1038/nmicrobiol.2016.274.
816 7. Wald A. The Fitting of Straight Lines if Both Variables are Subject to Error. The Annals of Mathematical
817 Statistics. 1940;11(3):284-300.
818 8. Benjamini Y, Krieger AM, Yekutieli D. Adaptive Linear Step-up Procedures That Control the False Discovery
819 Rate. Biometrika. 2006;93(3):491-507.
820 9. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive Essentiality Analysis of
821 the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. mBio. 2017;8(1). Epub
822 2017/01/18. doi: 10.1128/mBio.02133-16. PubMed PMID: 28096490; PubMed Central PMCID:
823 PMCPMC5241402.
824 10. McNeil MB, Chettiar S, Awasthi D, Parish T. Cell wall inhibitors increase the accumulation of rifampicin in
825 Mycobacterium tuberculosis. Access Microbiol. 2019;1(1):e000006. Epub 20190320. doi:
826 10.1099/acmi.0.000006. PubMed PMID: 32974492; PubMed Central PMCID: PMCPMC7470358.
827 11. Patel Y, Soni V, Rhee KY, Helmann JD. Mutations in rpoB That Confer Rifampicin Resistance Can Alter Levels
828 of Peptidoglycan Precursors and Affect β-Lactam Susceptibility. mBio. 2023;14(2):e0316822. Epub
829 20230213. doi: 10.1128/mbio.03168-22. PubMed PMID: 36779708; PubMed Central PMCID:
830 PMCPMC10128067.
831 12. Campodonico VL, Rifat D, Chuang YM, Ioerger TR, Karakousis PC. Altered Mycobacterium tuberculosis Cell
832 Wall Metabolism and Physiology Associated With RpoB Mutation H526D. Front Microbiol. 2018;9:494.

833 Epub 20180319. doi: 10.3389/fmicb.2018.00494. PubMed PMID: 29616007; PubMed Central PMCID:
834 PMCPMC5867343.
835 13. Palmer AC, Kishony R. Opposing effects of target overexpression reveal drug mechanisms. Nat Commun.
836 2014;5:4296. Epub 20140701. doi: 10.1038/ncomms5296. PubMed PMID: 24980690; PubMed Central
837 PMCID: PMCPMC4408919.
838 14. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus
839 on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021;49(D1):D274-
840 D81. doi: 10.1093/nar/gkaa1018. PubMed PMID: 33167031; PubMed Central PMCID: PMCPMC7778934.
841 15. Provvedi R, Boldrin F, Falciani F, Palu G, Manganelli R. Global transcriptional response to vancomycin in
842 Mycobacterium tuberculosis. Microbiology (Reading). 2009;155(Pt 4):1093-102. doi:
843 10.1099/mic.0.024802-0. PubMed PMID: 19332811.
844 16. Soetaert K, Rens C, Wang XM, De Bruyn J, Laneelle MA, Laval F, et al. Increased Vancomycin Susceptibility
845 in Mycobacteria: a New Approach To Identify Synergistic Activity against Multidrug-Resistant
846 Mycobacteria. Antimicrob Agents Chemother. 2015;59(8):5057-60. Epub 20150601. doi:
847 10.1128/AAC.04856-14. PubMed PMID: 26033733; PubMed Central PMCID: PMCPMC4505240.
848 17. Hansen JL, Ippolito JA, Ban N, Nissen P, Moore PB, Steitz TA. The structures of four macrolide antibiotics
849 bound to the large ribosomal subunit. Mol Cell. 2002;10(1):117-28. doi: 10.1016/s1097-2765(02)00570-1.
850 PubMed PMID: 12150912.
851 18. Chulluncuy R, Espiche C, Nakamoto JA, Fabbretti A, Milón P. Conformational Response of 30S-bound IF3 to
852 A-Site Binders Streptomycin and Kanamycin. Antibiotics (Basel). 2016;5(4). Epub 20161213. doi:
853 10.3390/antibiotics5040038. PubMed PMID: 27983590; PubMed Central PMCID: PMCPMC5187519.
854 19. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE, 3rd. Mutations in gidB confer low-level
855 streptomycin resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2011;55(6):2515-
856 22. Epub 20110328. doi: 10.1128/AAC.01814-10. PubMed PMID: 21444711; PubMed Central PMCID:
857 PMCPMC3101441.
858 20. Spies FS, Ribeiro AW, Ramos DF, Ribeiro MO, Martin A, Palomino JC, et al. Streptomycin resistance and
859 lineage-specific polymorphisms in Mycobacterium tuberculosis gidB gene. J Clin Microbiol.

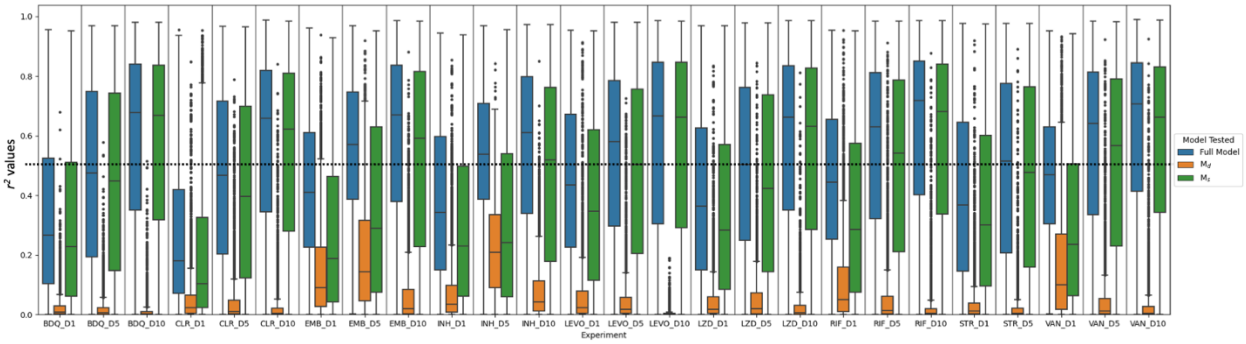
860 2011;49(7):2625-30. Epub 20110518. doi: 10.1128/JCM.00168-11. PubMed PMID: 21593257; PubMed
861 Central PMCID: PMCPMC3147840.
862 21. Cui ZL, Xiaojun ; Shin, Joonyoung ; Gamper, Howard ; Hou, Ya-Ming ; Sacchettini , James C ; Zhang, Junjie
863 Interplay between an ATP-binding cassette F protein and the ribosome from Mycobacterium tuberculosis.
864 Nature Communications. 2022. PubMed Central PMCID: PMC35064151.
865 22. Madsen CT, Jakobsen L, Buriankova K, Doucet-Populaire F, Pernodet JL, Douthwaite S. Methyltransferase
866 Erm(37) slips on rRNA to confer atypical resistance in Mycobacterium tuberculosis. J Biol Chem.
867 2005;280(47):38942-7. Epub 20050920. doi: 10.1074/jbc.M505727200. PubMed PMID: 16174779.
868 23. Vilcheze C, Weisbrod TR, Chen B, Kremer L, Hazbon MH, Wang F, et al. Altered NADH/NAD+ ratio mediates
869 coresistance to isoniazid and ethionamide in mycobacteria. Antimicrob Agents Chemother. 2005;49(2):708-
870 20. doi: 10.1128/AAC.49.2.708-720.2005. PubMed PMID: 15673755; PubMed Central PMCID:
871 PMCPMC547332.
872 24. Hazbón MH, Brimacombe M, Bobadilla del Valle M, Cavatore M, Guerrero MI, Varma-Basil M, et al.
873 Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant
874 Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2006;50(8):2640-9. doi: 10.1128/aac.00112-
875 06. PubMed PMID: 16870753; PubMed Central PMCID: PMCPMC1538650.
876 25. Bollela VR, Namburete EI, Feliciano CS, Macheque D, Harrison LH, Caminero JA. Detection of katG and inhA
877 mutations to guide isoniazid and ethionamide use for drug-resistant tuberculosis. Int J Tuberc Lung Dis.
878 2016;20(8):1099-104. doi: 10.5588/ijtld.15.0864. PubMed PMID: 27393546; PubMed Central PMCID:
879 PMCPMC5310937.
880 26. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential
881 transcript expression. Bioinformatics. 2015;31(17):2778-84. Epub 20150428. doi:
882 10.1093/bioinformatics/btv272. PubMed PMID: 25926345; PubMed Central PMCID: PMCPMC4635655.
883 27. Mathis AD, Otto RM, Reynolds KA. A simplified strategy for titrating gene expression reveals new
884 relationships between genotype, environment, and bacterial growth. Nucleic Acids Research.
885 2020;49(1):e6-e. doi: 10.1093/nar/gkaa1073.

28. Helgesson P, Sjostrand H. Fitting a defect non-linear model with or without prior, distinguishing nuclear reaction products as an example. Rev Sci Instrum. 2017;88(11):115114. doi: 10.1063/1.4993697. PubMed PMID: 29195386.

29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PubMed Central PMCID: PMCPMC4302049.

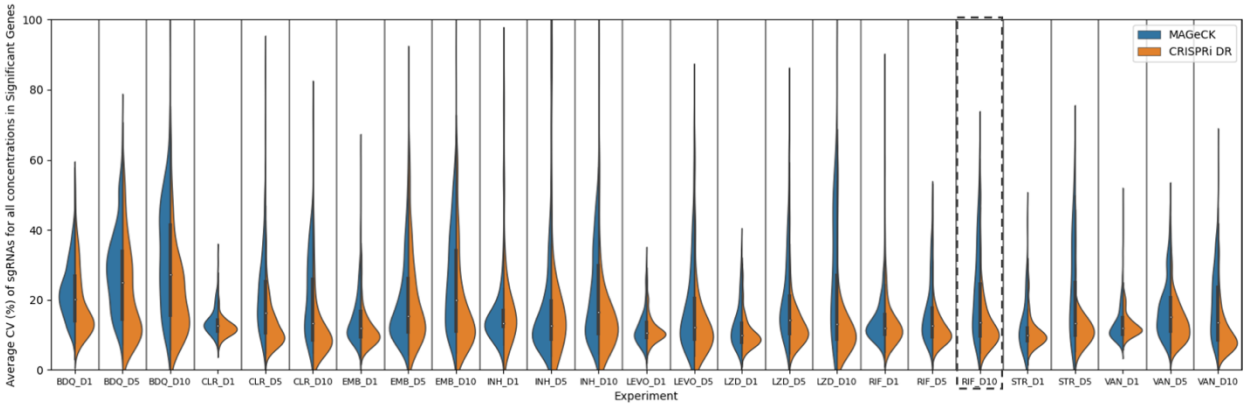
30. Li W, Köster J, Xu H, Chen C-H, Xiao T, Liu JS, et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biology. 2015;16(1):281. doi: 10.1186/s13059-015-0843-6.

Supporting Information



S1 Fig. Evaluation sgRNA strength and log concentration as predictors of CRISPRi-DR model through comparison of distribution of r^2 values of full (CRISPRi-DR) and ablated (M_s and M_d) models for each gene in each experiment.

The horizontal line is where r^2 = 0.5. The average r^2 M_s model for all genes across all the experiments is 0.42, the average r^2 for the M_d model is 0.07. This alongside the Log-likelihood tests indicate sgRNA strength is the more significant predictor. However, the full CRISPRi-DR model outperforms both M_d and M_s (average r^2 is 0.50) indicating the inclusion of both sgRNA strength and log concentration is needed for accurate assessment of significant sgRNA depletion in a gene in a condition.



S2 Fig. Distribution of average CV of sgRNAs in significant genes (depleted and enriched) in the CRISPRi-DR model and MAGeCK.

In this Fig, we see all the noise distributions for hits in MAGeCK and the CRISPRi-DR model for all experiments. The dashed panel is that of RIF D10. The same distribution of noise of hits can be seen in Fig 7. The trend seen with RIF D10 is present with all the experiments except LEVO D10. We see that the CRISPRi-DR model is unimodal with a low CV as the mode, whereas MAGeCK shows significant genes with low average CV values but also a significant amount of genes with high average CV values. LEVO D10 was left out of this plot due to the low number of hits in either model.

S1 Table. Ranking of Select Genes using the CRISPRi-DR model in 1 Day, 5 day and 10 Day pre-depletion of treated libraries.

An extended version of Table 1, where the CRISPRi-DR model is run on each gene for each drug and pre-depletion day. The coefficient for the slope of concentration dependence (β_c) is extracted from the fitted regressions and used to rank the genes in both increasing order (for depletion) and inversely (for enrichment). Green reflects results consistent with expectations based on knowledge of known gene-drug interactions.

924 **S2 Table. Comparison of significant interactions Identified by CRISPRi-DR and MAGeCK for each drug**
925 **and pre-depletion day.**
926 For each drug and pre-depletion day, both CRISPRi-DR and MAGeCK are run on data. MAGeCK is run
927 separately for each concentration and the overall significant interactions are determined as the union of
928 the individual runs. CRISPRi-DR is run is run once using data from all three concentrations (and sgRNA
929 strengths) together. The comparison of the significant interactions identified by the models is evaluated
930 using true positives, true negatives, false positives and false negatives. The results from MAGeCK are
931 used as the "ground truth" against which the other model's results are compared. Cells with red font in
932 the "tp" column represent low overlaps between the interactions found by the two models, and cell
933 with red font in the "Number of …" columns highlight low number of interactions found in the relative
934 model.
935
936 **S3 Table. Matrices for comparison of significant interactions Identified by CRISPRi-DR and MAGeCK for**
937 **each drug and pre-depletion day.**
938 The table presents the results of CRISPRi-DR and MAGeCK analyses for different drugs and pre-depletion
939 days. Significant interactions are compared in matrix form. Cells with red font indicate low overlaps
940 between the interactions found by the two models, while cells with green font represent high overlaps.
941
942 **S1 File. Evaluating performance differences between CRISPRi-DR and MAGeCK using a simulated**
943 **sgRNA barcodes.**
944 To better understand the differences in performance between CRISPRi-DR and MAGeCK, and to evaluate
945 the sensitivity of these methods to different sources of noise, we developed a simulation model to
946 generate artificial datasets of sgRNA barcode counts. In this experiment, we used the same set of
947 ~99,000 sgRNAs and empirical measurements of sgRNA strengths for genes in the *Mtb* genome as in the

948 CRISPRi library in the paper by (Li, Poulton et al. 2022), and simulated exposure to a virtual inhibitor over
949 4 concentrations (1μM, 2μM, 4μM, and 8μM), 3 replicates each. Our objective was to quantify how
950 much noise in the counts, both within concentrations and between concentrations, affects the precision
951 and recall of each method.