PLOS COMPUTATIONAL BIOLOGY

METHODS

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

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

* schoudhery@tamu.edu

OPEN ACCESS

Citation: Choudhery S, DeJesus MA, Srinivasan A, Rock J, Schnappinger D, Ioerger TR (2024) A dose-response model for statistical analysis of chemical genetic interactions in CRISPRi screens. PLoS Comput Biol 20(5): e1011408. https://doi.org/10.1371/journal.pcbi.1011408

Editor: Wei Li, Children's National Hospital, George Washington University, UNITED STATES

Received: August 2, 2023

Accepted: April 22, 2024

Published: May 20, 2024

Data Availability Statement: A python-based implementation of the CRISPRi-DR method for analyzing CRISPRi data is publicly available as part of Transit2: https://transit2.readthedocs.io/en/latest/ The output files from analyses of the Mtb CRISPRi CGI screens from Li, Poulton et al. at doi: 10.1038/s41564-022-01130-y using 6 CRISPR analysis methods (including CRIPSRi-DR) are available for download at: https://orca1.tamu.edu/CRISPRi-DR/.

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 objective is to identify CRISPRi mutants whose relative abundance is suppressed (or enriched) in the presence of a drug when the target protein is depleted, reflecting synergistic behavior. Different sgRNAs for a given target can induce a wide range of protein depletion and differential effects on growth rate. The effect of sgRNA strength can be partially predicted based on sequence features. However, the actual growth phenotype depends on the sensitivity of cells to depletion of the target protein. For essential genes, sgRNA efficiency can be empirically measured by quantifying effects on growth rate. We observe that the most efficient sgRNAs are not always optimal for detecting synergies with drugs. sgRNA efficiency 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). To capture this interaction, we propose a novel statistical method called CRISPRi-DR (for Dose-Response model) that incorporates both sgRNA efficiencies and drug concentrations in a modified dose-response equation. We use CRISPRi-DR to re-analyze data from a recent CGI experiment in *Mycobacterium tuberculosis* to identify genes that interact with antibiotics. This approach can be generalized to non-CGI datasets, which we show via an CRISPRi dataset for *E. coli* growth on different carbon sources. The performance is competitive with the best of several related analytical methods. However, for noisier datasets, some of these methods generate far more significant interactions, likely including many false positives, whereas CRISPRi-DR maintains higher precision, which we observed in both empirical and simulated data.

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

Competing interests: The authors have declared that no competing interests exist.

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., sgRNA 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 efficiency 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 technology is becoming an increasingly important tool for genome-wide identification of gene functions in various environmental conditions [1–3]. For example, several different approaches have been devised to exploit CRISPR to induce depletion of target proteins. In the earlier CRISPRko approaches, a nuclease-active form of CAS9 was used to deactivate target genes by cutting the DNA at a target locus and induce DNA repair, which could introduce indels causing frameshifts or inserting novel elements, abrogating their function completely [1–3]. Another approach, CRISPRa, utilizes dCAS9 fusions with effectors that actively enhance or suppress transcription through direct interaction with the RNA polymerase (such as transcription factors that can activate transcription) [4].

In CRISPR interference (CRISPRi), a catalytically-dead CAS9 protein (dCAS9) is recruited to a chromosomal locus by a single guide RNA (sgRNA) with a short (~20 bp) complimentary sequence and physically blocks transcription [5]. dCAS9 nucleases from several different organisms are available for CRISPRi (e.g. *S*. *pyogenes*, *S*. *thermophilus*, [6]) and different promoters and chemicals have been used for dCAS9 induction. The degree of CRISPR interference can be tuned by modulating the level of dCAS9 expression [7], varying the sgRNA sequence with respect to its length, GC-content, targeting sequence complementarity, position in the gene, or similarity of targeted PAM (protospacer adjacent motif) sequence, to consensus for optimal dCAS9 recognition, [5, 6, 8–11]. While in mammalian systems, efficiency of sgRNAs can vary among multiple cell types, [9], for simplicity, our focus is on studying single defined lineages, as in bacterial strains. Tuning CRISPRi allows to deplete the targeted gene product to intermediate levels [5], which allowed the introduction of the concept of gene 'vulnerability' as describing the sensitivity of cells to partial depletion of individual proteins [12]. By this definition, highly vulnerable genes are genes for which even small depletion of the encoded protein causes growth impairment, which can be quantified efficiently on a genome-wide scale using high-throughput sequencing [12]. The vulnerability of a gene can be both condition dependent and strain or cell type dependent [12].

One interesting application of CRISPRi is to reveal targets of antibiotics or mechanisms of resistance through chemical-genetic interactions (CGI) [7, 13]. CRISPRi libraries can be designed to contain multiple sgRNAs targeting each gene, resulting in a set of thousands of individual depletion mutants [12]. In this context, 'mutant' refers to a cell line transformed

with a integrative plasmid capable of expressing the dCAS9 protein and the unique targeting sgRNA, even though it contains the wild-type gene sequence. The abundance of each mutant can be quantified by amplifying the sgRNA targeting sequence which functions as a molecular barcode, and then performing deep sequencing to count the number of barcodes for each sgRNA in a treatment [6]. The analysis of such datasets is challenging, due to various sources of noise which introduce variability in the counts.

There are several previously published methods for statistical analysis of CRISPR datasets. One, called MAGeCK [14] (originally intended for CRISPRko screens), calculates a log-fold-change (of mean counts) for each sgRNA between a treatment condition and a reference condition (control), and uses a Gaussian distribution to estimate the significance of differences in mean sgRNA abundance between treatments and controls (based on the implementation in the source code, which differs from the description in the publication). To evaluate effects at the gene level, individual sgRNAs are combined in MAGeCK using Robust Rank Aggregation (RRA) to prioritize genes whose sgRNAs show greater enrichment or depletion on average than other genes in the genome. MAGeCK has been used for evaluating chemical-genetic interactions (CGI) with antibiotics [14]. A variant called MAGeCK-MLE [15] fits a Bayesian model by Maximum Likelihood that captures changes in mean counts with increasing time or concentration, along with effectiveness of each sgRNA through posterior probabilities of a binary variable, to determine the overall probability that a gene interacts. Other approaches such as CRISPhieRmix [16] use mixture models to separate effective from ineffective sgRNAs, and thereby identify interacting genes as those containing a significant subset of effective sgRNAs. DrugZ [17] identifies significant interactions by averaging together Z-scores (assuming a Normal distribution) of log-fold-changes of sgRNAs at the gene level. DEBRA [18] utilizes DeSeq, a method for transcriptomic analysis, which employs the Negative Binomial distribution for counts and a more sophisticated method for modeling variance and using it to discriminate genes displaying significant changes in mean counts.

However, most of these methods have one of two limitations when applied to identify genes affecting drug potency. First, CGI experiments are ideally carried out with multiple drug concentrations around the MIC (minimum-inhibitory concentration), since it is often difficult to anticipate what concentration will stimulate the right amount of growth inhibition in combination with CRISPRi-induced depletion of target proteins. However, many of the existing methods analyze the data for each drug concentration independently (i.e. comparing each concentration to a no-drug control). Since knock-down mutants might exhibit depletion at one concentration but not others, results from multiple concentrations must be combined post-hoc. As an example, the authors in [13] chose to combine results from analyzing different concentrations of a given drug using MAGeCK-RRA by taking the union of significant interacting genes at each individual concentration. Due to the noise in these CRISPRi experiments, analyzing concentrations independently increases the risk of detecting false positives (in the sense that non-interacting genes might be spuriously called as hits at different concentrations).

Second, many of the analytical methods do not explicitly take into account differences in sgRNA efficiency (i.e. take sgRNA efficiencies as an input in the model). Different sgRNAs can induce different degrees of depletion of their target genes, and this in turn causes different effects on growth rate, depending on sensitivity of the cells to protein depletion [10]. In this paper, we use sgRNA efficiency to refer to the overall effect on growth rate, because we do not assess the intermediate levels of protein depletion. This can be quantified beforehand by evaluating the growth rate of individual CRISPRi mutants (with unique sgRNAs) in a growth experiment and determining the actual fitness defect caused by target knockdown [11, 12]. In highly vulnerable genes, the effect of protein depletion by sgRNAs on cell growth rate (efficiency) can span a range from no effect to severe growth defect. Early applications of CRISPR were

primarily being used to fully inactivate genes (e.g. CRISPRko), rather than to produce graded depletion effects. Therefore, at the time some of these methods were developed, this information was often not used, as methods to quantify sgRNA efficiencies were not well developed. Even in MAGeCK, the Robust Rank Aggregation method treats all sgRNAs in a gene as "equal" a priori, without differentiating them based on the expected effects on growth phenotype (i.e. sgRNA efficiency is not an input). In contrast, it has been recognized that different sgRNAs can have different efficiencies, and several papers have investigated the factors that are associated with stronger sgRNAs [19], especially sequence-based attributes such as similarity to optimal PAM sequence, length and GC content of targeting sequence, mismatches, etc. [5, 8, 10]. Mathis, Otto and Reynolds [11] exploit this to synthetically create a diverse set of sgRNAs with a range of efficiencies by mutating the guide RNA sequences, which they quantify by empirically fitting growth curves for each modified sgRNA with a logistic equation. Interacting genes are then found using differences in the fitted parameters that includes the quantified growth rates and the Hill coefficient. Among all the existing CRISPR analytical methods, MAGeCK-MLE [15] is the only other method that explicitly includes sgRNA efficiencies as an input, which are used to set the prior probabilities that each sgRNA is effective or not (because of their focus on CRISPRko) in the joint probability formula, to initialize for the Expectation Maximization iterations.

In the application to CGI data, a regression model can be used to integrate data over multiple drug concentrations [20]. The degree of a gene-drug interaction is reflected by the coefficient (or slope) for the dependence of CRISPRi mutant abundance on drug concentration. This regression approach was previously introduced in CGA-LMM for analysis of hypomorph libraries (where there is typically just one mutant representing each gene) [20]. It was based on the theory that depletion of the target of a drug should ideally synergize with increasing concentrations of the drug. While exposure to an inhibitory compound will challenge the growth of all the mutants in a hypomorph library, mutants with depletion of a gene that interacts with a drug (e.g. prototypically, an essential gene that is the drug target) will exhibit excess depletion relative to others in the library due to the combined effect of both the growth-inhibition due to the drug treatment in conjunction with the growth-impairment due to knock-down of an vulnerable gene, making these hypomorphic mutants even more sensitive to the drug. For genes that genuinely interact with a given drug, this depletion effect should be exacerbated at higher drug concentrations (i.e. be dose-dependent); thus, genes of greatest relevance would be those that exhibit concentration-dependent effects. While the (log of) abundance of a depletion mutant does not have to decrease perfectly linearly with the (log of) drug concentration to obtain a significant negative coefficient (slope) in the regression, there should be a general trend supporting that relative abundance decreases as concentration increases.

One of the challenges in extending this prior regression approach (CGA-LMM) to CRISPRi screens was incorporating information on sgRNA efficiency. Even in essential genes, some sgRNAs may produce strong depletion of the target, while others might be almost completely ineffective. While sgRNA strength can be partially predicted (with intermediate accuracy) from sequence alone [9, 12], the actual growth phenotype depends on vulnerability of the target gene (sensitivity of cells to depletion of the protein product), which is what is meant by sgRNA efficiency. Even sgRNAs that are predicted to be strong might not cause a growth defect if they are in a non-essential gene. sgRNA efficiency must be empirically quantified by measuring growth rates in standard growth media (e.g. by fitting exponential growth curves based on optical density, or using a reporter gene) with versus without induction of dCAS9, and then calculating relative fitness defects [11]. An alternative approach is to fit the abundance of depletion mutants to a piecewise linear model that allows for a preliminary lag phase, and then extrapolating the model to predicted log-fold-change (LFC) at a fixed number of

generations [12]. Any such measure of sgRNA efficiency can be incorporated as a term in the CRISPRi-DR model we present below. Although one could contemplate adding the efficiency of each sgRNA into a simple regression model to predict abundances for each gene, a significant problem (expanded upon below) is that sgRNAs of different efficiency can show different concentration dependence, resulting in non-linear interactions among variables.

In this paper, we propose a modified regression approach for CRISRPi data (called CRISPRi-DR) that incorporates both drug concentration and sgRNA efficiency. The approach is based on the classic dose-response (DR) model for inhibition activity of drugs; the activity of a target protein typically transitions from high to low in the shape of an S-curve as concentration increases (on a log scale), which can be modeled with a Hill equation. The parameters of the Hill equation for a given drug can be fit by performing a log-sigmoid transformation of the mutant abundance data and then using ordinary least-squares regression. We show how sgRNA efficiency can be incorporated into this model as a multiplicative term in the Hill equation, which becomes an additive effect in the log-sigmoid transformed data. The benefit of this model is that it decouples the concentration-dependence from the sgRNA efficiency, so they can be fit as independent (non-interacting) terms in the regression, which ultimately amplifies effects that may be apparent only for a subset of sgRNAs in an optimal efficiency range.

CRISPRi-DR is applicable to libraries where there are multiple sgRNAs representing each gene with a range of efficiencies, which can be quantified empirically as an effect on growth rate (fitness defect). The diversity of efficiencies is useful for identifying synergistic effects with treatments/conditions. Thus, the main requirements for CRISPRi-DR are that: a) there are multiple sgRNAs for each target in the library, b) the sgRNAs vary in predicted strength, and c) the actual efficiencies of the sgRNAs (i.e. growth defects due to target depletion) have been experimentally quantified in control conditions, as an input to the analysis method. The primary use case we focus on is identification of chemical-genetic interactions, with drug concentration as a covariate. We demonstrate the value of the CRISPRi-DR analysis method by re-analyzing the data from a recent paper using CRISPRi for chemical-genetic interactions to identify targets of antibiotics in *M. tuberculosis*. However, the approach can be generalized to analyze experiments with other covariates, such as time-points of a treatment, where there is a sigmoidal response in growth. We illustrate this by using CRISPRi-DR to analyze an *E. coli* CRISPRi dataset from an experiment to determine genes differentially required for growth on different carbon sources [11].

Methods

The CRISPRi-DR method applies to CRISPRi experiments that involve using high-throughput sequencing to tabulate sgRNA counts representing abundance of individual CRISPRi mutants in a population (pooled culture). Each mutant has an sgRNA (on a plasmid) mapping to a target gene that can reduce its expression (e.g. with dCAS9 induction). In CGI applications, the culture is treated with antibiotics or inhibitors at various concentrations, along with a no-drug control, and DNA is extracted, PCR-amplified, and sequenced, producing counts representing each sgRNA. If Y_{ijk} is the abundance (i.e. count) for an sgRNA i in a condition j for replicate k, normalized abundance can be calculated by $Y'_{ijk} = \frac{Y_{ijk}}{\sum_{x=1}^{n} Y_{xjk}}$, where each count is divided by the sum of counts of all the sgRNAs observed in a given condition and replicate. Let U'_i be the normalized abundance of sgRNA i in the uninduced condition, then the normalized relative abundances of an sgRNA i in all induced samples can be calculated as: $A_{ijk} = \frac{Y'_{ijk}}{U'_i}$, assuming that the counts in the uninduced condition represents full abundance of each clone (normal growth without target depletion).

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 efficiencies. The dose-response effect of an inhibitor on the activity of an enzyme is traditionally modeled with the Hill-Langmuir equation.

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

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 Eq (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 kth replicate count of sgRNA i was extracted from,

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

The unknown parameters are the IC_{50} value (inhibitory 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 IC_{50}, representing the mid-point of the transition.

The dose-response model seen in Eq (2) can be extended to account for sgRNA efficiency by incorporating a multiplicative factor in the denominator:

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

sgRNA efficiency, S_i, is an empirical measure of the degree of growth impairment resulting from target depletion. This can be assessed in several ways, such as estimating change in exponential growth rate in a reference condition in a growth experiment [21]. Alternatively, Bosch et al [12] use estimated log-fold change of abundance (induced vs uninduced) at a fixed number of generations of growth in-vitro in the absence of drug, extrapolated from a model fit to empirical data (passaging experiment) that allows for a lag phase. K_s represents the unknown intermediate sgRNA efficiency that causes 50% depletion of mutant abundance (half-way between no depletion and full depletion), and the H_s is the unknown Hill coefficient that represents how sensitive mutant abundance is to depletion of the target protein.

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

Abundance of mutants in a CRISPRi CGI experiment can be affected simultaneously by both presence of an inhibitor and depletion of an interacting gene. However, the concentration-dependent effect of a drug on mutant abundance can be different for sgRNAs of different efficiency. Fig 1 illustrates the interaction between these two effects for *rpoB* (RNA polymerase beta chain) in an *Mtb* CRISPRi library treated with rifampicin with 5 days pre-depletion. The lines in Fig 1A are regression fits obtained for each sgRNA in *rpoB* using regression of log abundances against log concentration of rifampicin, $\log(A_{ijk}) = C + B \cdot \log([D_j])$, where C is in the intercept and B is the slope of the regression, representing concentration dependence, and

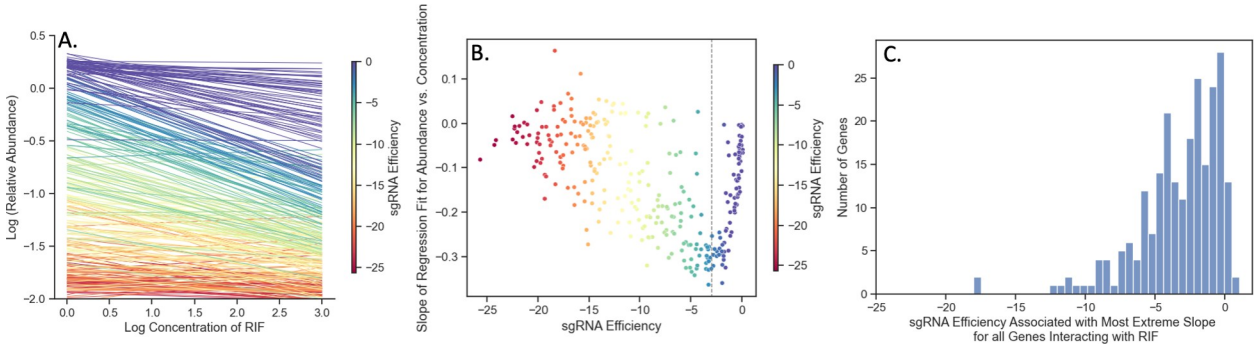

Fig 1. Effect of sgRNA efficiency on concentration dependence for sgRNAs in *rpoB* in a CRISPRi library treated with RIF (D5). (A) Regression lines for log(relative abundance) against log(concentration) for all sgRNAs in *rpoB* in a library treated with RIF for 5 days pre-depletion. The lines that reflect the extremes of the sgRNA efficiency (red or purple), are flat and do not show much change in abundance. Comparatively, intermediate sgRNA efficiency (dark green to indigo) shows the most negative slopes, reflecting maximum synergy with drug. (B) Comparison of sgRNA efficiency and slopes of the regressions seen in Panel A for each sgRNA. Each point is an sgRNA colored by its efficiency. The most efficient sgRNAs (purple) and the least efficient sgRNAs (red) show concentration slopes around 0. The dotted line reflects the minimum of the parabolic curve. (C) Histogram of sgRNA efficiencies where the slopes reach their most extreme (positive or negative) for 236 interacting genes in RIF D5. The distribution shows that most of the extrema sgRNAs for interacting genes fall in the range of -5 to 0 (note: not the strongest sgRNAs, which would have efficiencies around -25).

https://doi.org/10.1371/journal.pcbi.1011408.g001

$\log(A_{ijk})$ are log relative abundances obtained as described above. The left-most side of Fig 1A shows the range of abundances in the no-drug control (induced library in media without rifampicin). These differences in abundances (dispersion along Y-axis) are due solely to the growth impairment caused by depleting RpoB. As concentration of RIF increases, some of the sgRNAs show very negative slopes, while other sgRNAs show slopes closer to 0. A parabolic-type curve emerges in Fig 1B when the slopes B from the regressions are plotted against the sgRNA efficiencies. Both the most efficient sgRNAs (colored red) and the least efficient sgRNAs (purple) have slopes around 0 (no concentration dependence). Highly efficient sgRNAs (red) can cause excessive depletion (even without drug), making it difficult to detect additional decreases due to increasing drug concentration. Comparatively, sgRNAs with very low efficiency (purple) might not induce enough depletion to synergize with the drug. The sgRNAs surrounding the minimum point of the parabolic curve (dashed line) in Fig 1B reflect those of intermediate efficiency where the ability to detect synergy with the drug is maximized. These are the sgRNAs in Fig 1B that show the most negative slope with increasing concentration (dark green-indigo). As Fig 1C shows, the efficiency where the slopes reach their extremes (most negative; or most positive for those showing enrichment) can be different for each gene but tend to fall in an intermediate region of sgRNA efficiency (0 to -5). The histogram shows that sgRNA efficiency at which the most extreme (largest or smallest) concentration-dependent slope is achieved over all interacting genes (236 for RIF D5). Hence, the sgRNAs that are optimal for detecting CGIs are not necessarily the strongest (most efficient). The variability of concentration-dependence (slope) with sgRNA efficiency suggests a possible non-linear interaction between the variables. This nonlinearity is captured in the multiplicative terms of the dose-response model (Eq (3)).

Linearization and parameter estimation

The dose-response model Eq (3) can be linearized through a log-sigmoid transformation.

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

$$C = H_s \cdot \log(K_s) - H_d \cdot \log(IC_{50}(D_j)) \quad (4)$$

In this log-sigmoid transformed space, the concentration-dependence and effect of sgRNA efficiency have been decoupled, appearing as independent linear terms with the Hill coefficients (H_s and H_d) as the variables to solve for by a standard regression. The inflection parameters of the sigmoid curve (K_s and IC_{50}) are combined in the intercept C in the model. Importantly, this model implies that the effects of growth impairment due to the depletion of a vulnerable gene and growth inhibition due to the drug on the overall (relative) abundance of a given mutant become additive in this log-sigmoid-transformed space. To illustrate this, the CRISPRi-DR equation is simulated by plotting idealized relative abundances (in Fig 2) using parameters chosen to emulate what is seen in Fig 1A, the plot of slopes over a systematic range of sgRNA efficiencies and drug concentrations for *rpoB*. In Fig 2A, the slopes of the concentrations are plotted against abundances calculated using the dose-response model. The slopes vary as a function of the starting depletion (left-hand side), which is due to sgRNA efficiency alone (colored gradient based on sgRNA efficiency value). The slopes are most negative for intermediate sgRNA efficiency, colored with a dark blue-green hue representing sgRNA efficiency around -10. Fig 2B illustrates the result of the linearization (log-sigmoid transformation) of the Hill equation. All the individual sgRNA regression lines over concentration become parallel, eliminating the dependence on sgRNA efficiency, and allowing them to be fit by a single common slope representing the concentration-dependence averaged over all the sgRNAs.

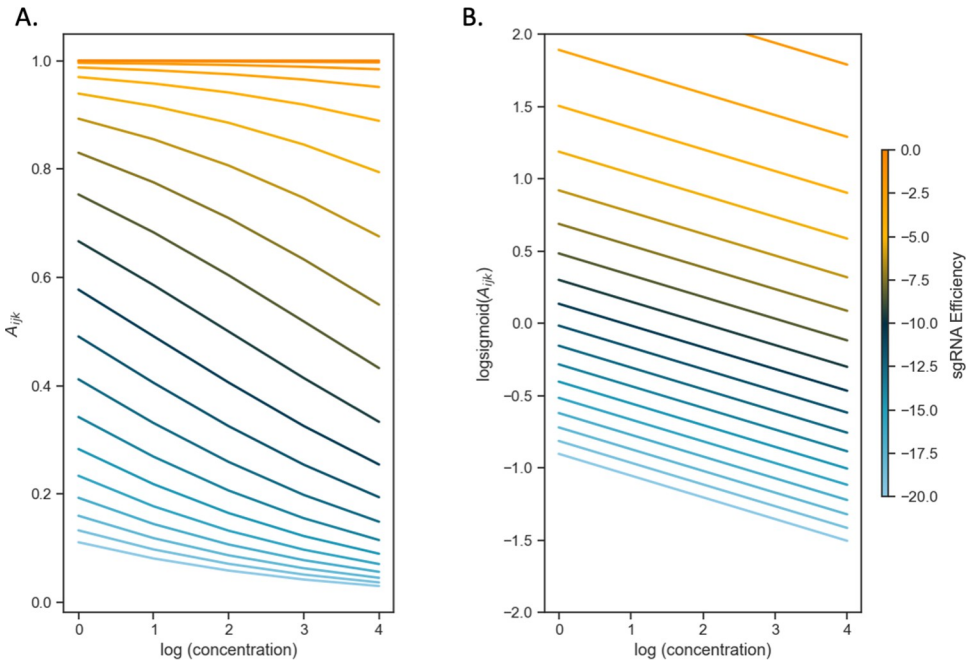

Fig 2. The log-sigmoid transformation of abundances allows the CRISPRi-DR model to factor in the non-linear effect of sgRNA efficiency on concentration dependence. (A) Simulation of sgRNAs abundances for an ideal essential gene. Parameters used in simulation: H_s = -4, IC_{50} = 0.25, K_s = -10 and H_d = -0.5 over a range of sgRNA efficiencies and drug concentrations. (B) When the log-sigmoid transformation of the abundances is applied, we see all the regression fits are parallel to one another, allowing to be fit by a single common slope, representing the concentration dependence over all sgRNAs, regardless of sgRNA efficiency.

https://doi.org/10.1371/journal.pcbi.1011408.g002

Experimental data (i.e. counts from sequencing, converted to relative abundances for mutants with each sgRNA) are fit on a gene-by-gene basis using ordinary least-square (OLS) regression by the following formula:

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

where A (relative abundance for each CRISPRi mutant at given drug concentration), S_i (sgRNA efficiency) and $[D_j]$ (concentration of drugs) are columns of a melted matrix. To include the control samples (no-drug, dCAS9-induced controls) in the regression, they are treated as one two-fold dilution lower than the lowest available concentration tested for the drug (to avoid taking the log of 0). Since the log-sigmoid transform of the relative abundances is taken, they must be within the range of (0,1). Although relative abundances greater than 1.0 are possible in treated conditions (relative to uninduced, no-drug controls), especially in cases where target depletion confers a growth advantage and consequent enrichment, we use a squashing function to ensure the relative abundances range between 0 and 1, which is required to take the log-sigmoid transform.

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

where $\tau = 0.01$ is a pseudo count needed to make abundances non-zero for taking logarithms. Relative abundances that are greater than 1.0 are mapped to just below 1.0, though the mapping is monotonic, so the order among sgRNAs is still preserved (higher abundances become exponentially closer to 1.0).

Significance testing

The statistic that indicates the degree of interaction of each gene with a given drug is the coefficient for the $log([D])$ term (i.e. slope) in the model. To determine whether the interaction is statistically significant, a Wald test [22] is applied to calculate a P-value reflecting whether the coefficient is significantly different than 0, adjusting for a target FDR (false discovery rate) of 5% over the whole genome using the Benjamini-Hochberg procedure [23]. However, the Wald test by itself yields many genes predicted to interact with the drug (often thousands), even 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 relevant to the drug mechanism. Our assumption is that most of genes in the genome do not interact with a given drug (at least not directly involved in the mechanism of action or resistance). Therefore, we apply a secondary filtering step using empirical Bayes FDR analysis to the distribution of slopes to identify genes whose slopes are outliers with respect to the rest of the distribution, similar to its use for analysis of log-fold-changes in differential gene expression [24]. Empirical Bayes FDR approach treats the entire genome as a mixture of null genes and genes showing a significant effect. It models slopes for most genes with a null distribution centered around 0, and then estimates parameters including proportion of null genes, p_0, to calculate the posterior probability that each gene satisfies the null hypothesis (local false discovery rate).

Let $\{s_i\}$ be the set of coefficients of concentration dependence (slopes) for each gene, which is approximately normally-distributed (bell-curve-shaped), with non-null genes in the tails. We start by computing null distribution Z-scores $\{z_i^0\}$ using the parameters from the *empirical null distribution* [25]. To eliminate the effect of outliers (e.g. non-null genes, with large slopes in the tails of the distribution), the parameters of the null distribution (m_0, s_0) are estimated by

fitting the central peak with a Normal distribution using numerical methods. Let $z_i^0 = \frac{(s_i - m_0)}{s_0}$. We model the density of the whole distribution of Z-scores with a two-component mixture model: $f(z_i^0) = p_0 f_0(z_i^0) + (1 - p_0) f_1(z_i^0)$, where p_0 is the probability that gene i is null, f_0 is the estimated probability density from the empirical null distribution, and f_1 is the unspecified probability of the alternative hypothesis [24]. While the probability density for null genes is assumed to be $f_0(z_i^0) \sim N(0,1)$ (standardized Gaussian pdf), the values of the mixture coefficients are initially unknown. Polynomial fitting to the peak (i.e. central matching) is used to empirically estimate the proportion of null genes, p_0 [25]. This can then be used to solve for the maximum-likelihood estimate of the posterior probability of each gene being null: $p(null|z_i^0) = \frac{p_0 f_0(z_i^0)}{f(z_i^0)}$, where $f(z_i^0)$ is approximated locally from the sample in bins. This is the local probability density at point, z_i^0. Finally, to calculate the empirical Bayes FDR (*ebFDR*), the genes are sorted based on their z values, $z_{(1)}^0 .. z_{(n)}^0$, and for each possible threshold z^*, we estimate the probability of committing type I errors (rejecting null hypothesis for null genes, false positives) via the cumulative null probability, which can be approximated by the average null probability over genes with $z_{(i)}^0 < z^*$:

$$ebFDR = p(null\ for\ some\ i | z_i^0 < z^*) = \int_{-\infty}^{z^*} f_0(Z) dZ \cong \frac{1}{k} \sum\nolimits_{z_i^0 < z^*} p(null|z_i^0)$$

where k is the number of genes with $z_i^0 < z^*$.

In the CRISPRi-DR implementation, we use the *locfdr* package in R to determine the empirical null distribution and posterior probabilities for each gene. To perform the last step symmetrically on both tails, the genes are sorted by probability instead of Z-score, which combines hits in the left and right tails, and the optimal threshold θ is determined such that $\frac{1}{k} \sum\nolimits_{p_i < \theta} p_i < 0.05$. The empirical Bayes procedure effectively identified genes with slopes that are outliers compared to the rest of the distribution. Since the Wald test and the empirical Bayes FDR analysis capture different aspects of significance, significant interactions are defined by a joint criterion as genes with slopes of concentration dependence that have both an adjusted P-value by Wald test < 0.05 and $ebFDR < 0.05$.

Results

CRISPRi Dataset and Pre-processing

A chemical-genomics dataset was obtained from high-throughput sequencing of a CRISPRi library of *M. tuberculosis* (*Mtb*) that had been treated with several antibiotics. The library consists of 96,700 sgRNAs targeting all 4019 genes in the *Mtb* H37Rv genome [13]. This library was intentionally constructed to focus on probing essential genes (based on prior TnSeq analysis [26]), with a mean of 83 sgRNAs per essential gene, but there are some sgRNAs in each non-essential gene too (mean of 10 sgRNAs per non-essential gene).

The library was individually treated with 9 anti-TB drugs (rifampicin, RIF; isoniazid, INH, ethambutol, EMB; vancomycin, VAN; levofloxacin, LEVO; linezolid, LZD; streptomycin, STR; clarithromycin, CLR; bedaquiline, BDQ) to evaluate and validate the CRISPRi system in preparation for target identification for novel inhibitors (from high-throughput screens). These drugs were selected because certain genes are expected to interact for each (based on known mechanisms of action), although additional genes might also exhibit interactions, which could extend our knowledge. We note that some drug targets are members of a complexes; although a drug may bind directly to one subunit, other subunits in those complexes often show similar CRISPRi phenotypes. RIF binds RpoB (RNA polymerase subunit) inhibiting transcription and

compensatory mutations are often found in *rpoC* [27], BDQ binds and inhibits AtpE (subunit of the ATP synthase) [28] and *mmpL5* effluxes the drug [29], GyrA and GyrB (subunits of DNA gyrase) would be expected to interact with fluoroquinolones like LEVO [30], EMB targets *embABC* in the arabinogalactan pathway [31, 32], CLR, LZD and STR bind to the ribosome and inhibit translation, which can be protected by rRNA methyltransferases [33–35], VAN binds to peptidoglycan and is expected to interact with genes in the peptidoglycan synthesis pathways [36, 37], and genes such as *inhA*, *katG*, *ahpC*, *ndh*, *mshA* and *cinA* are implicated in the mechanism of action or resistance for isoniazid, an inhibitor of mycolic acid synthesis [38–41]. These define selected interactions that would be expected to be observed in a CRISPRi CGI experiment.

Samples of the library (pooled cultures) were treated with each of the drugs, with induction of the Sth1 dCAS9 by ATC (anhydrotetracycline), and were sequenced in triplicate at several concentrations for each drug at 2-fold dilutions around the MIC, along with control samples representing the no-drug samples (0 concentration). Three periods of pre-depletion were evaluated: 1, 5, and 10 days (D1, D5, and D10), since it was initially unknown how many days would be optimal for reducing protein expression after induction of CRISPRi. The measurements reported in this experiment are observed counts of sgRNAs, representing the relative proportion of each mutant in the population (pooled culture of CRISPRi mutants). Abundance of a mutant increases or decreases if silencing of the targeted gene causes a change in fitness. Although target proteins are knocked down by inhibiting transcription via CRISPRi, intracellular protein levels are not directly measured in the experiment. Instead, unique nucleotide barcodes representing each sgRNA are amplified from (integrated) plasmids in the cells, sequenced, and counted. The counts reflect the relative abundance of each CRISPRi mutant. Samples were normalized by dividing individual counts for each sgRNA by the sample total (sum over all sgRNAs).

In this dataset, prior estimates of sgRNA efficiency were obtained from empirical data by fitting a piecewise-linear equation to fitness over multiple generations, and then using the model for to extrapolate the predicted log-fold change (LFC) each sgRNA at 25 generations [12]. The scale for these efficiencies ranged between -25 (highest depletion) and 0 (no depletion). To determine the effect of depletion solely due to the sgRNA (without drug), uninduced samples (in the absence of dCAS9 induction, -ATC) were also sequenced, to provide counts representing mutant abundances in the absence of depletion of targets as an input to the model.

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

The CRISPRi-DR model was fitted for all chemical-genetic interaction datasets from Li, Poulton [13], which included nine drugs tested at three different concentration levels (after 1, 5, and 10-days of pre-depletion without drug). The analyses by CRISPRi-DR found a range of tens to hundreds of significant genes for each dataset. Table 1 show a more detailed account of the significant genes founds in these CRISPRi screens by CRISPRi-DR, categorized into depleted (mutant abundance decreases with drug concentration) and enriched (mutant abundance increases with drug concentration).

The significant genes identified by CRISPRi-DR generally have coefficients of concentration dependence that are outliers with respect to the rest of the genes. Fig 3 shows the slopes calculated for genes in a library treated with EMB (one day of pre-depletion, D1). Panel A shows a volcano plot with vertical dashed where ebFDR = 0.05 and a horizontal dashed line where adjusted P-value = 0.05. The 568 Genes with ebFDR $<$ 0.05 and adjusted P-value $<$ 0.05

Table 1. Number of Significant Genes found by CRISPRi-DR across the nine drugs CRISPRi screen for each of pre-depletion days.

Drug	D1		D5		D10	
	Depleted	Enriched	Depleted	Enriched	Depleted	Enriched
BDQ	91	139	159	31	196	0
CLR	447	23	129	252	172	176
EMB	28	540	45	415	54	208
INH	92	99	34	139	62	189
LEVO	139	116	103	250	15	0
LZD	122	323	47	323	54	131
RIF	456	186	399	140	305	102
STR	83	147	115	88	-	-
VAN	366	0	324	47	213	120

https://doi.org/10.1371/journal.pcbi.1011408.t001

are significant and colored orange. These genes include the targets of EMB: *embA*, *embB* and *embC* [31, 32], which have slopes -0.45, -0.43 and -0.32, respectively. The calculation of ebFDR is shown in Panel B, where the baseline distribution shown is that of the calculated Z-scores from the coefficients of concentration dependence. The green curve fits the overall density, and the blue dashed curve models the central null hypothesis density. The yellow triangles indicate the thresholds at which the cumulative posterior probability of a gene being a null gene is < 0.05 (i.e. where ebFDR < 0.05).

To evaluate the relative importance of the sgRNA efficiency and drug concentration features to the CRISPRi-DR model, each gene was fit with 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

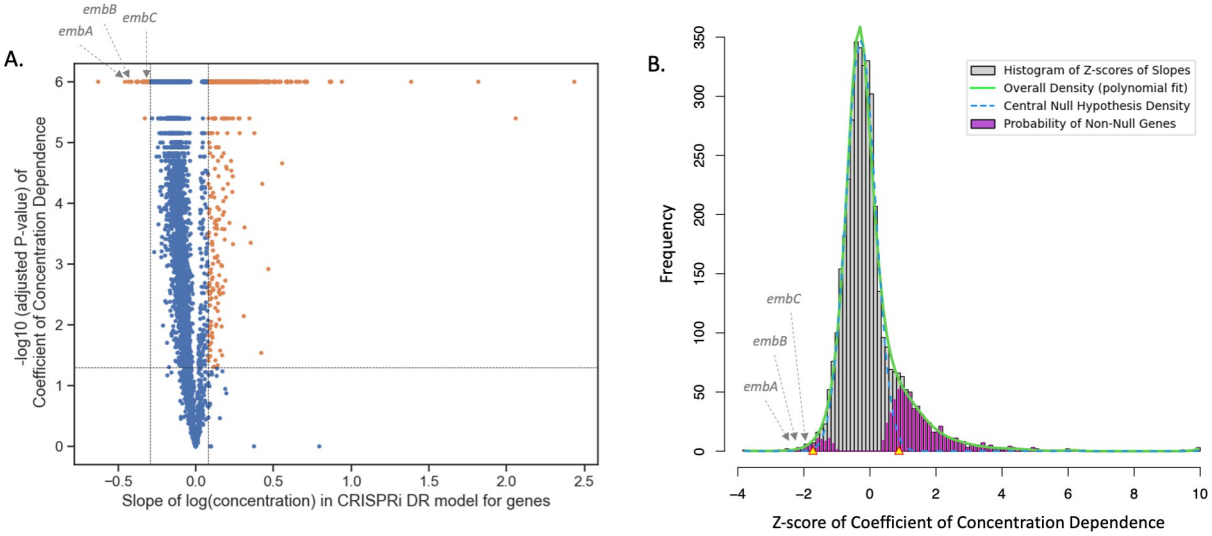

Fig 3. Coefficients of concentration-dependence from CRISPRi-DR model fitted for EMB D1 (1 day of pre-depletion). A) Volcano Plot of the coefficient of concentration dependence of each gene versus their -log10 of the adjusted P-value, extracted from the model fit. The horizontal line is where adjusted P-value = 0.05. The vertical lines are at slope = -0.29 and slope = 0.08, where ebFDR = 0.05. The orange points are significant genes, i.e. genes with slopes of concentration dependence with adjusted P-value < 0.05 and ebFDR < 0.05. B) The histogram of Z-scores calculated from the coefficients of concentration dependence for all genes for EMB D1 (generated by *locfdr* R package). The green curve fits the overall density, and the blue dashed curve models the central null hypothesis density. The magenta areas are the posterior probability of non-null genes. The yellow triangles indicate the thresholds at which the cumulative posterior probability of a gene being a null gene is < 0.05 (i.e. where ebFDR < 0.05). These triangles correspond to the vertical lines in Panel A.

https://doi.org/10.1371/journal.pcbi.1011408.g003

the M_s model only contained sgRNA efficiency as a predictor: $\log\left(\frac{A_{ijk}}{1-A_{ijk}}\right) = B \cdot S_i + C$. In the EMB D1 experiment, the average r^2 (% variance explained) across all genes in full CRISPRi-DR model is 0.43. Comparatively, the average r^2 is 0.29 for M_s and 0.13 for M_d. *embA* appears as one of the genes in the M_d set of significant interactors, but the other targets of the drug, *embB* and *embC* do not appear in the sets of significant interactors for either of these ablated models. As a measure of the model quality (goodness of fit), the Akaike Information Criterion (AIC) for the full model in the EMB D1 experiment is 87.6, whereas the AIC of M_d is 300.7 and AIC of M_s is 124.7. The full model has the lowest AIC, indicating it is the best fitting model of the three. The AIC for the model incorporating only drug concentrations but not sgRNA efficiency (M_d) is highest (worst), suggesting that sgRNA efficiency encodes critical information needed for predicting mutant abundance. A Likelihood Ratio Test shows that the differences between these models is highly significant (P-value $<< 0.05$; χ^2 distribution using one degree of freedom, since the ablated models each have one parameter less than full model). The r^2 values and results of the AIC-based likelihood comparison indicate that sgRNA efficiency contributes strongly to accuracy of the model, and reinforces the importance of including sgRNA efficiency as a term in the CRISPRi-DR model.

The improved performance of CRISPRi-DR over the reduced models for EMB extends to the other drugs tested, as seen in S1 Fig. In all the experiments, the number of genes with fits with $r^2 > 0.5$ is the greatest in the full CRISPRi-DR model, and the number of genes with fits that have $r^2 > 0.5$ is greater in model M_s than M_d. This demonstrates that in all conditions, both concentration and sgRNA efficiency are needed to make accurate estimates of mutant abundance.

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

Significant CGIs are identified on the same order of magnitude by both the CRISPRi-DR model and MAGeCK (MAGeCK-RRA) as reported in Li, Poulton [13]. Since MAGeCK is evaluated at a single concentration at a time, we combined P-values across concentrations for each of the drug treated samples using Fisher's method (see Supplemental for details). Analogous to the policy in Li, Poulton [13] that augments the original MAGeCK definition for significant genes, an additional filter of |LFC|>1 is applied along with the adjusted P-value < 0.05 to find the significant number of genes. As seen Fig 4A and the Extended S2 Fig from Li, Poulton [13], MAGeCK with additional constraints and CRISPRi-DR find on the order of the same number of significant genes. For both depleted genes and enriched genes, the number of significant genes fall along the diagonal dashed line. Additionally, the Venn diagrams in Fig 4B show there is high overlap of calls made by the two methodologies (enriched and depleted combined). Across all the datasets, an average of 68.8% of genes identified as significant by CRISPRi-DR are also found to be significant as per the constrained MAGeCK output. Additional details of the overlap of significant interacting genes in the filtered MAGeCK outputs and CRISPRi-DR can be found in S3 Table.

CRISPRi-DR model correctly detects genes known to interact with anti-tubercular drugs

When genes are ordered by coefficients of the slope representing the dependence of abundance on drug concentration from the CRISPRi-DR model, genes known to affect the potency of the anti-mycobacterial drug tested are ranked highly, as expected (Table 2). The more

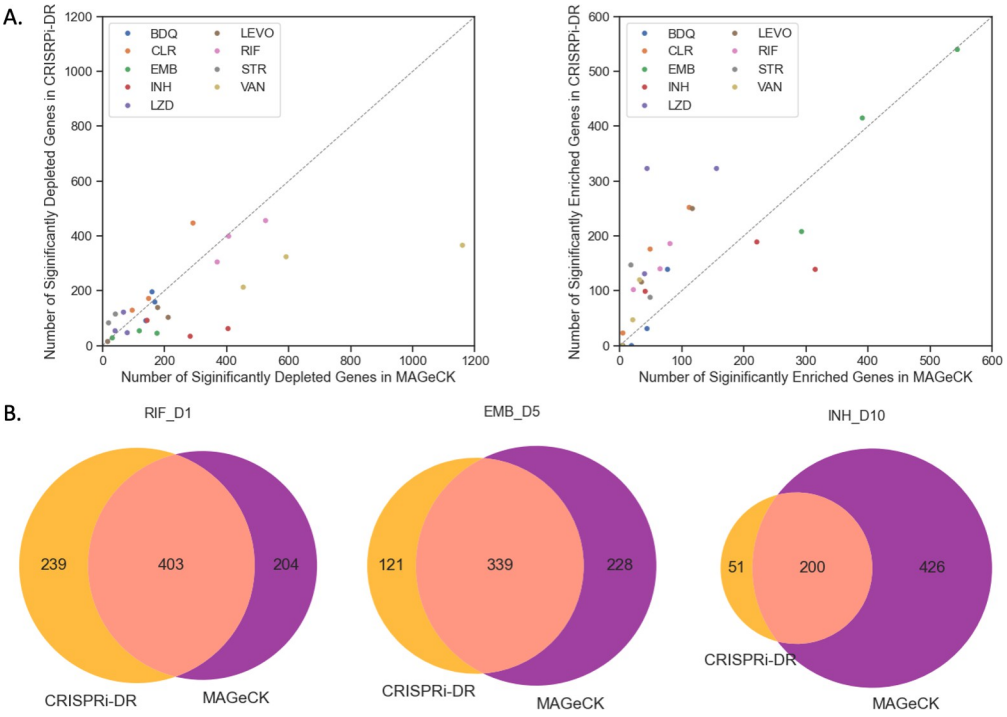

Fig 4. Comparison of significant interactions found by CRISPRi-DR and MAGeCK. (A) The points in the plots are the analyses of CRISPRi screens by both MAGeCK and CRISPRi-DR, colored by drug treatment. The left plot compares the depleted hits called by the two methodologies and the right plot compares the enriched hits called by the two methodologies. (B) Venn Diagram of significant genes, both depleted and enriched, found by CRISPRi-DR and MAGeCK for select drug-treated libraries. The genes identified by CRISPRi-DR highly overlap with the genes found by MAGeCK.

https://doi.org/10.1371/journal.pcbi.1011408.g004

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

For each drug, the CRISPRi-DR model is run on each gene (using data from D1). The coefficient for the slope of concentration dependence (β_c) is extracted from the fitted regression and used to rank the genes both in increasing order (for depletion) and inversely (for enrichment). Green reflects results consistent with expectations based on knowledge of known gene-drug interactions

Genes that encode the target of a drug would typically be expected to have a high depletion rank, i.e., show a negative slope, indicating that as concentration increases, abundance for the given depletion-mutant decreases. This can be seen in S1 Table, in the ranking for genes using the CRISPRi-DR model. These genes rank the highest in D1 and not as well in D10. This can be attributed to the fact that, after 10 days of pre-depletion, these mutants (especially for essential genes) are already quite depleted, even at concentration 0, increasing noise, and making it difficult to pick up on concentration-dependent signals (further depletion). Therefore, the ranking of relevant genes in D1 was assessed in this analysis (Table 2).

For isoniazid (INH), there are multiple relevant genes identified by CRISRPi-DR, including *inhA*, *ahpC*, *ndh* [42], and *katG* [43]. *inhA* (enoyl-ACP reductase) is an essential gene in mycolic acid pathway that is the target of INH, and AhpC (alkyl hydroperoxide reductase) responds to the oxidative effects of isonicotinic radicals in the cells, MshA is a protein involved in synthesis of mycothiol, which helps maintain redox balance [41], and CinA is a NADH metabolizing protein that can hydrolyze the isoniazid-NAD adduct [40]. Therefore, as dosage of the

Table 2. Ranking of Select Genes using the CRISPRi-DR model in 1 Day pre-depletion of treated libraries.

Drug	Gene	*D1* Depletion Ranking	*D1* Enrichment Ranking
BDQ	*atpA*	11	4022
BDQ	*atpB*	6	4027
BDQ	*atpC*	51	3982
BDQ	*atpD*	14	4019
BDQ	*atpE*	25	4008
BDQ	*atpF*	9	4024
BDQ	*atpG*	12	4021
BDQ	*atpH*	8	4025
BDQ	*mmpL5*	2	4031
CLR	*RVBD3579c*	40	3993
CLR	*erm(37)*	1	4021
EMB	*embA*	2	4031
EMB	*embB*	3	4030
EMB	*embC*	19	4014
INH	*inhA*	3	4030
INH	*ahpC*	2	4031
INH	*cinA*	5	4028
INH	*katG*	4031	2
INH	*ndh*	4028	5
INH	*mshA*	4025	8
LEVO	*gyrA*	4012	21
LEVO	*gyrB*	4021	12
LZD	*erm(37)*	3865	168
LZD	*tsnR*	4032	1
RIF	*rpoB*	94	3939
RIF	*rpoC*	147	3886
STR	*RVBD2477c*	4021	12
STR	*gidB*	4022	11

https://doi.org/10.1371/journal.pcbi.1011408.t002

drug increases, the abundances of the mutants of these genes should decrease. These genes are in the top 10 highest ranked depletion genes for INH (see Table 2). In contrast, *katG* and *ndh* are found among the top 5 enriched hits, exhibiting increased survival when the proteins are depleted. KatG (catalase) is the activator of INH, and the most common mutations in INH-resistant strains occur in *katG*, decreasing activity [44]. *Ndh* (type II NADH reductase) mutants have also been shown to decrease sensitivity to INH by shifting intracellular NADH levels (needed for INH-NADH adduct formation), and mutations in *ndh* have been shown to be defective in target enzyme (NdhII) activity [42], which is consistent with the observation in the CRISPRi data that depletion of *ndh* leads to increase survival in the presence of INH. Similarly, *mshA* is highly enriched, consistent with mutations found in resistant mutants.

For EMB, *embA*, *embB*, and *embC* (subunits of the arabinosyltransferase, target of ethambutol, EMB) rank within the top 100 depleted genes for all three pre-depletion conditions [31, 32]. However, interactions with the other genes in the arabinogalactan pathway, like *ubiA* (which sometimes acquires resistance mutations [45]), were not observed.

In RIF, *rpoB* and *rpoC*, subunits of the core RNA polymerase, are ranked within the top 150 genes. Significant negative interacting genes for RIF also include many cell wall related genes such as *ponA2*, *rodA*, *ripA*, *aftABCD*, *embABC*, etc., consistent with recent studies that show RIF exposure (or mutations in *rpoB*) leads to various cell wall phenotypes [46–48]. Similarly,

the target of bedaquiline (BDQ), the F0F1 ATP synthase (which includes 8 subunits encoded by *atpA-atpH*, of which AtpE is the one bound by BDQ) [28], and *mmpL5*, which can efflux the drug [29], are ranked within the top 40 depleted genes in BDQ.

The significantly interacting genes in vancomycin (VAN) involve many genes in the cell wall/membrane/envelope biogenesis pathway (as defined by in COG pathways [49]) (adjusted P-value for pathway enrichment = 0.0004 using Fisher's Exact Test). This follows previous studies that show cell wall genes are targets of vancomycin [50, 51], which binds to peptidoglycan in the cell wall.

In levofloxacin (LEVO), CRISPRi mutants of *gyrA* and *gyrB* (subunits of the DNA gyrase, the target of fluoroquinolones) are also observed to be enriched. The reason that depletion of this drug target leads to enrichment of mutants (hence a growth advantage, rather than the expected growth impairment) is likely due to reduced generation of double-stranded breaks in the DNA and other toxic intermediates as a side-effect of inhibiting the gyrase, an effect that has been observed in *E. coli* [52].

For clarithromycin (CLR), an inhibitor of translation, *Rv3579c* and *erm(37)* are observed as hits. *Erm(37)* adds a methyl group on the A2058/G2099 nucleotide in the 23S component of the ribosome, the same site in which clarithromycin binds [53]. This natively increases tolerance to CLR in *Mtb*. As this gene is depleted, CLR has greater opportunity to bind, reducing the cells' natural tolerance to the drug. Consistent with this observation, e*rm(37)* has a depletion rank of #1 in the CLR D1 condition. Rv3579c is another methyltransferase with a similar function that ranks highly (#35) in CLR.

In contrast to methylation inhibiting the binding of CLR, there are ribosome methyltransferases in *Mtb*, where methylation facilitates binding of a drug. Mutants for these genes would be expected to show a high enrichment rank in presence of drug. For instance, streptomycin (STR) interferes with ribosomal peptide/protein synthesis by binding near the interaction of the large and small subunits of the ribosome [54]. Two relevant genes that influence the binding of STR include *gidB* and *Rv2477c/ettA*. GidB is an rRNA methyltransferase that methylates the ribosome at nucleotide G518 of the 16S rRNA, the position at which STR interacts [35], increasing native affinity for STR. This is consistent with the observation that one of the most common mutations in STR-resistant clinical isolates is loss of function mutations in *gidB* [55]. Rv2477c is a ribosome accessory factor, also known as EttA, which is an ATPase that enhances translation efficiency. It has also recently been shown to bind the ribosome near the P-site (peptidyl transfer center), potentially interfering with binding of aminoglycosides [56], and loss-of-function mutations observed in drug-resistant clinical isolates of *M. tuberculosis* have shown to confer resistance to STR [13]. The ranking of both genes using the CRISPRi-DR model are within the top 12 enriched genes in STR. For linezolid (LZD), relevant genes identified are *erm(37)* and *tsnR*. TsnR is an rRNA methyltransferase, analogous to GidB, and results in tolerance to LZD in a similar manner as GidB does for STR [13]. Following this expectation, *tsnR* has an enrichment ranking of #1 in LZD. Whereas depletion of Erm(37) gives tolerance to CLR, it increases sensitivity to LZD. The nucleotides that Erm(37) methylates in the 23S RNA are proximal in 3D space to where mutations conferring LZD-resistance are found, which both lie in the PTC (peptidyl-transfer center) of the ribosome [57].

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

A reason that the CRISPRi-DR model shows lower consistency with MAGeCK (RRA) in some datasets could be due to different sensitivity to noise. There is some noise in these experiments due to variability in sequencing sgRNA counts across multiple concentrations and replicates. This can differentially affect the accuracy of predictions of gene-drug interaction made by

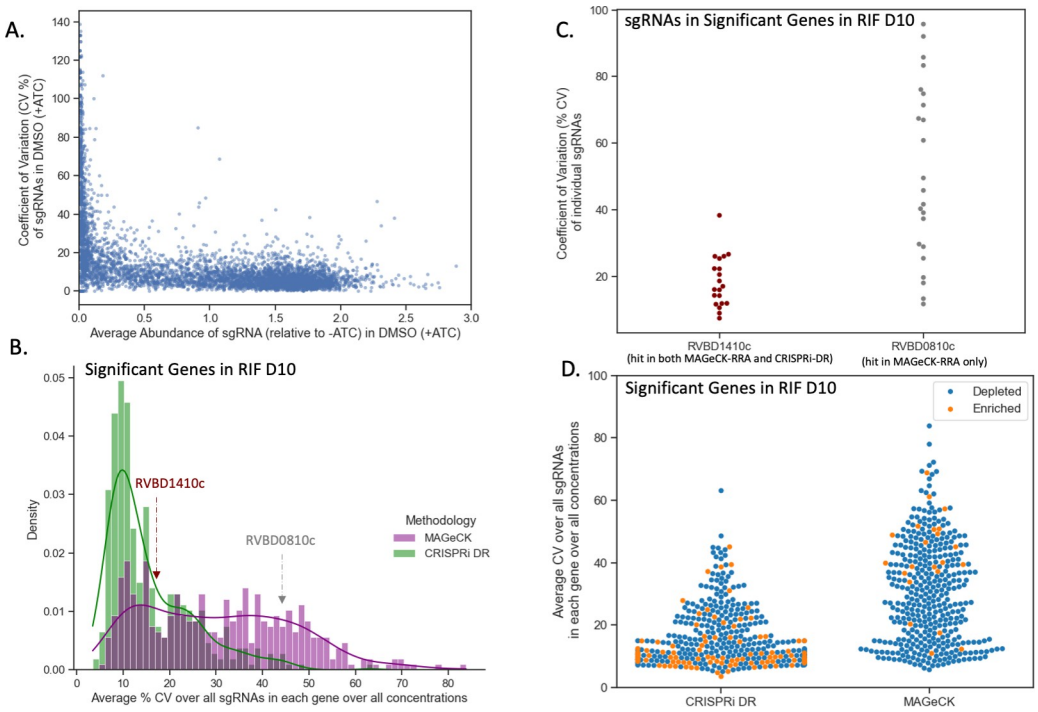

Fig 5. CRISPRi-DR model shows less sensitivity to noise than MAGeCK. (A) Comparison of average relative abundance and average CV across replicates in no-drug control samples for a sample of sgRNAs: For each sgRNA, we looked at the average CV of sgRNAs in the 3 control replicates against the average abundance of the sgRNA across those replicates. The lower the average abundance, the greater the noise present for the sgRNA. (B) Distribution of average CV of gene for significant genes in MAGeCK and significant genes in CRISPRi-DR in RIF D10: The distribution of average CV of significant genes in CRISPRi-DR model is more skewed and has a peak at CV ≈ 10%. Although most significant genes in MAGeCK show an average CV around 15%, there are quite a few genes with higher average CVs not found significant by the CRISPRi-DR model. (C) Coefficient of Variation (CV) of each sgRNA in two genes with similar number of sgRNAs for a library treated with RIF D10: *Rv1410c* is significant in both methodologies and *Rv0810c* significant in MAGeCK but not in CRISPRi-DR. The majority of CV values for sgRNAs in *Rv1410c* is around 20%. Although both genes have about 20 sgRNAs, *Rv0810c* shows 8 sgRNAs whose CV values exceed 60.5%, which is the maximum CV present in *Rv1410c*. (D) Distribution of average CV for enriched and depleted significant genes in MAGeCK and CRISPRi-DR in a RIF D10 library. This plot shows the distribution plot of Panel B, separated by depletion, and enriched significant genes. The average CV values for significant genes in the CRISPRi-DR model are low for both enriched and depleted genes. As seen in Panel B, significant genes in MAGeCK show low average CV, but they also show high average CV. Although there is a substantially lower number of significantly enriched in MAGeCK, they still show a large amount of noise compared the significantly enriched genes in CRISPRi-DR model.

https://doi.org/10.1371/journal.pcbi.1011408.g005

these models. Three replicate counts were collected for estimating the relative abundance of each CRISPRi mutant (with a unique sgRNA) in the presence of a drug at a given concentration. The coefficient of variation (CV) can be used to measure the relative consistency of measurements across these observations, which in turn can be used to evaluate the sensitivities of CRISPRi-DR and MAGeCK to noise in the raw data.

For each sgRNA s_i the coefficient of variation (CV) was calculated across the relative abundances for the 3 replicates for each concentration (C) in drug (D) ($CV_{D,C,i} = \frac{\sigma(i)}{\mu(i)}$), where $\sigma(i)$ is the standard deviation of the 3 relative abundances in concentration C and $\mu(i)$ is the mean. In Fig 5A, the $CV_{D=DMSO,C=0,i}$ (C of abundances for a random subset of sgRNAs (~5%) in a dCAS9-induced, no-drug condition (concentration 0) is compared to the average abundance. For sgRNAs of medium to high relative abundance (i.e., less depletion), the CV is fairly constant at approximately 10%. However, at low relative (to uninduced) abundances (i.e. higher depletion), CV value increases substantially to over 100%. If a gene contains multiple such sgRNAs

with high CV values, then the variation may be misconstrued as a genetic interaction by a methodology that is susceptible to noise.

The average noise in a gene *g* for a given drug *D* can be quantified as the average $CV_{D,C,i}$, for all concentrations *C* and all sgRNA*s* in the gene ($\bar{CV}_D(g)$). Therefore, $\bar{CV}_D(g)$ reflects the measure of overall noise present in a gene in a drug D. The distribution of $\bar{CV}_D(g)$ in RIF D10 for the 407 total significant genes (enriched and depleted combined) in the CRISPRi-DR model and in 379 total significant genes (enriched and depleted combined over all concentrations) in MAGeCK can be seen in Fig 5B. The distributions for both methodologies share a peak at about $\bar{CV}_D(g) \approx 10\%$. The distribution of $\bar{CV}_D(g)$ for significant genes in MAGeCK has a fatter tail than the distribution of $\bar{CV}_D(g)$ for significant genes in the CRISPRi-DR model. Fig 5D also shows that the average CV of significant genes found by MAGeCK is much higher than CRISPRi-DR (colored by depleted and enriched) for the RIF D10 screen. This trend of higher noise in MAGeCK hits is seen not only in RIF D10, but across all the experiments conducted (See S2 Fig). This indicates that although MAGeCK is identifying genes with low noise (like the CRISPRi-DR model), it is also detecting many genes with high noise that the CRISPRi-DR model is not.

An example of such a gene is *Rv0810c* (unknown function). The gene has 22 sgRNAs and has a $\bar{CV}_D(g)$ value (average CV over sgRNAs in a gene) of 51.4%, one of the highest measures in the RIF D10 experiment. In RIF D10, it is reported to be significantly depleted only in MAGeCK and not in the CRISPRi-DR model. The dispersion of the CV values of the sgRNAs in *Rv0810* are compared to those of *Rv1410c* in Fig 5C. *Rv1410c* (ABC transporter/efflux pump) has 20 sgRNAs, an $\bar{CV}_D(g)$ of 16.3% and is reported to be significantly depleted in both MAGeCK and the CRISPRi-DR model. Although both genes have some sgRNAs with low CVs (below 40%), *Rv0810c* shows 8 sgRNAs with CVs of at least 60.5%, which is the maximum CV of sgRNAs in *Rv1410c*. The CRISPRi-DR model considers the abundances at all concentrations, whereas MAGeCK compares each concentration to the baseline independently. Therefore, if sgRNAs have a high CV value at a particular concentration, they can be picked up as a significant genetic interaction by MAGeCK. The average relative abundance for the 3 replicates at concentration 0 for all sgRNAs in *Rv0810c* is 0.19, whereas the average relative abundance in *Rv1410c* for the same is 1.08. As Fig 5A shows, *Rv0810c* falls in the low abundance/high noise section of the graph, with an average sgRNA no-drug CV of 47.9%, whereas *Rv1410c* falls in the low noise section of the graph, with an average sgRNA no-drug CV of 11.2%. This demonstrates that MAGeCK reports genes such as *Rv0810c* with low abundances resulting in a large $\bar{CV}_D(g)$, which the CRISPRi-DR model does not, i.e., MAGeCK is more suspectable to noise than the CRISPRi-DR model.

Effects of noise on model performance using simulated CRISPRi data

The sensitivity and accuracy of the CRISPRi-DR model, MAGeCK-RRA and MAGeCK-MLE was assessed under different sources of noise using simulated sgRNA counts sampled from the Negative Binomial distribution [58], with means at different concentrations determined by the dose-response model (Eq (3)). sgRNAs with empirical efficiencies sampled from a uniform distribution from -25 to 0 were used to simulate the combined effects of CRISPRi depletion and exposure to a virtual inhibitor at four concentrations (1μM, 2μM, 4μM, and 8μM), with three replicates each. The aim was to determine how noise within and between concentrations affects the performance of each method. Detailed information on the simulation is provided in the S1 Text.

Nine datasets (LL, LM, LH, ML, MM, MH, HL, HM and HH) were simulated by varying two noise parameters: variability of abundances *between* concentrations (σ_C), and variability

among replicates *within* a concentration (P_{nb}, probability parameter of the Negative Binomial distribution), each with low (L), medium (M), and high (H) setting. A total of 1000 genes was simulated with 20 sgRNAs each. The first 50 genes are chosen as true negative interactions (with a virtual drug), the second 50 as positive interactions, and the last 50 as negative controls (for MAGeCK-RRA and MAGeCK-MLE). For interacting genes, slopes are chosen from a Normal distribution around +0.8 or -0.8, with a standard deviation of 0.2. For non-interacting genes, slopes are chosen from a Normal distribution around 0, with a standard deviation of 0.2. CRISPRi-DR, MAGeCK-RRA and MAGeCL-MLE were run ten times each on these 4 scenarios. MAGeCK was run independently for each drug concentration (2uM, 4uM, 8uM, compared to a no-drug control) and combined using Fisher's method post-hoc, while CRISPRi-DR and MAGeCK-MLE were run on all four concentrations simultaneously.

In lowest noise scenario (LL = low noise between concentrations and low noise among replicates), CRISPRi-DR identified 79.8% of the simulated interacting genes, MAGeCK-RRA identifies 56.2% and MAGeCK-MLE identifies 99.9%. As noise increases, the recall rate of MAGeCK-MLE remains quite high at 96.8% in the highest noise scenario (HH), and MAGeCK-RRA increases to 86.3%. The recall rate of CRISPRi-DR drops down to 21.0%. However, the false positive rate of CRISPRi-DR remains low at 0.2% in this HH scenario, and the false positive rates of MAGeCK-MLE and MAGeCK increase substantially (MLE = 27.5%, RRA = 33.7%), diluting the sets of predicted enriched and depleted genes with non-interacting genes (false positives). Therefore, although CRISPRi-DR identifies less of the true interacting genes in higher noise, it maintains its ability to keep the set of reported interacting genes from being diluted with non-interacting. Across the low and medium noise scenarios, CRISPRi-DR has higher F1-scores than the other two methods, where $F1\ score = 2 \times \frac{recall \times precision}{recall + precision}$, reflecting a better tradeoff between recall and precision (see Supplemental for more details). For the higher noise scenarios, MAGeCK-RRA shows instances with greater F1-scores, than both CRISPRi-DR and MAGeCK-MLE. Note for MAGeCK-RRA, we used Fisher's method to combine the outputs from the 3 individual concentrations, as well as applied a filter of $|LFC|>1$ on the significant genes. Without these additional constraints, the performance of MAGeCK-RRA is the worst of the three methods in all the noise scenarios.

The effect of noise on the true and false positive calls made by the methods can be seen in Fig 6, where number of significant genes is plotted for each of the adjusted noise parameters. For MAGeCK-MLE, significant genes were identified as those with adjusted P-value (based on a Wald test) less than 0.05. For MAGeCK-RRA, significant genes were identified as those with adjusted combined P-value less than 0.05 and an |LFC| greater than 1. MAGeCK-RRA is more affected by noise among replicates than between concentrations, as evident by the orange bar for $P_{nb} = 0.1$. This is likely a result of stochastic fluctuations of counts at individual drug concentrations that are not necessarily supported at other concentrations. Thus, we can speculate that the drug treated screens where MAGeCK-RRA identifies large numbers of significant genes (e.g. VAN D1) are those with high noise resulting in a possibly higher number of false positives. Comparatively, CRISPRi-DR and MAGeCK-MLE seem to be more affected by noise between concentrations than noise between replicates, showing lower precision as σ_C increases. Since these methods rely more on increasing or decreasing trends in abundance that must be (at least somewhat) consistent across concentrations, noise between concentrations may make these trends more difficult to identify.

To assess the impact of performing a CRISPRi screen at multiple drug concentrations on the performance of CRISPRi-DR, MAGeCK and MAGeCK-RRA, we conducted the simulation above with high-noise settings (HH) and varying numbers of drug concentrations (1, 2, or 3) for 10 iterations each (see Fig 6 in Supplement). The recall of the methods held fairly

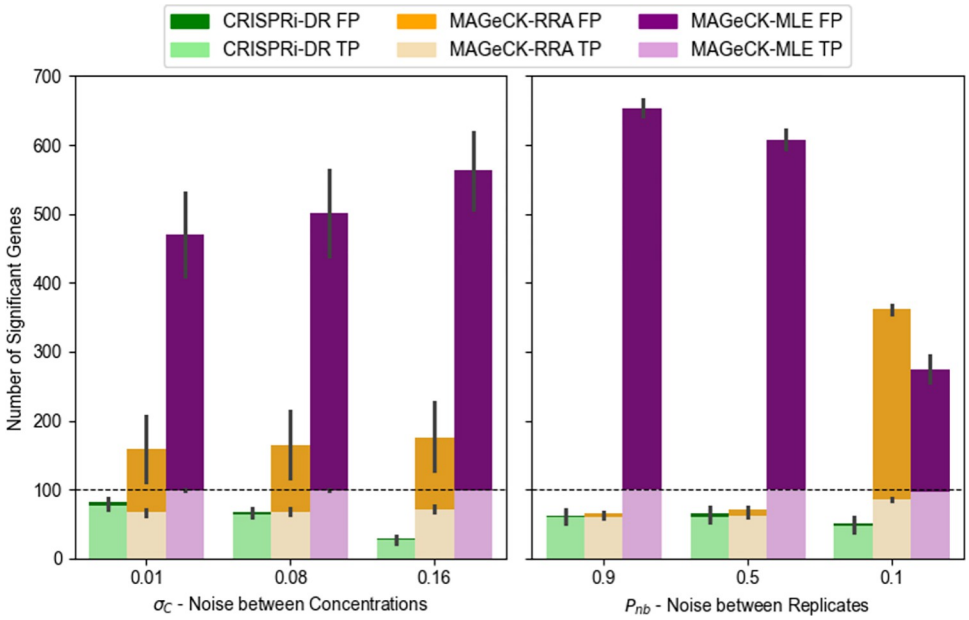

Fig 6. Average True Positives (TP) and False Positives (FP) found by CRIPSRi-DR, MAGeCK-RRA and MAGeCK-MLE as Simulated Noise Increases. The horizonal dashed line in both panels is the number of total simulated interacting genes (100 total). The parameters in the x-axis are ordered to reflect increasing noise. The leftmost bars of the two plots are the lowest noise and the rightmost bars are the highest noise. MAGeCK-MLE produces a high false positive rate for all scenarios and MAGeCK-RRA is more sensitive to noise among replicates as seen by the orange bar for $P_{nb} = 0.1$.

https://doi.org/10.1371/journal.pcbi.1011408.g006

constant as concentrations were added. However, increasing the number of concentration points caused a significant increase in false positive calls by MAGeCK-RRA from 200 to 400. While MAGeCK-RRA shows susceptibility to false positives when evaluating only a single concentration point, this effect was amplified with more concentrations. This accumulation of errors explains the decrease in precision with additional concentration points. In contrast, CRISPRi-DR is more robust with respect to false-positive errors. By incorporating data from all available concentrations and identifying significant trends, CRISPRi-DR maintains higher precision that does not diminish with the addition of more concentration points. Although MAGeCK-MLE makes many more calls, including false positives, the number of false positives did not increase as concentrations were added, because, like CRISPRi-DR, MAGeCK-MLE incorporates data from all available concentrations.

Comparison of CRISPRi-DR to alternative methods for CRISPRi analysis

To understand how well CRISPRi-DR performs relative to other CRISPR analysis methods, we applied the following methods on the *M. tuberculosis* CGI data from [13] described above: CGA-LMM [20], MAGeCK-RRA [14], MAGeCK-MLE [15], DrugZ [17], DEBRA [18], and CRISPhieRmix [16]. Each method offers a unique approach to analyzing CRISPRi data. Some of these methods, such as CGA-LMM do not explicitly incorporate multiple sgRNAs per gene or account for differences in sgRNA efficiency. Other methods, such as DEBRA, MAGeCK-RRA and drugZ, do not explicitly account for different drug concentrations in a CGI experiment, and so they must be run independently on each concentration and the results combined. Only CRISPRi-DR and MAGeCK-MLE incorporate both of these factors in their statistical analysis.

The details of applying each method, including parameter settings, handling of negative controls, and merging of results, are described in the Supplement. Several of the methods,

including MAGeCK-MLE, produced more significant interactions (in the thousands, in some cases), whereas other methods, like CRISPRi-DR, produced much more focused lists of significant hits for each drug (often less than 100) (see details in the Supplement).

To evaluate the accuracy of the predictions by each method, we ranked the genes by significance (usually based on P-value, for most methods) and then generated ROC (Receiver-Operator Characteristic) curves. To define a list of expected hits (i.e. interacting genes) for isoniazid (INH D1, with one day of pre-depletion), we obtained a list of 90 conditionally essential genes from a previously published TnSeq study of *M*. *tuberculosis* H37Rv exposed to sub-MIC concentrations of antibiotics [37]. While changes in essentiality due to knock-out of a gene by transposon insertion are not technically the same as fitness defects resulting from CRISPRi depletion of a target gene, there is substantial overlap between essentiality and vulnerability [12]. Many genes known to play a role in INH resistance (*fabG1*, *katG*, *ndh*, *ahpC*, *cinA*, etc.) are highly interacting (enriched or depleted) in both experiments. Thus, the list of TnSeq conditional essentials serves as a proxy for the genes that are expected to exhibit an interaction effect in the CRISPRi screen (even though, admittedly, not all necessarily will). Importantly, conditional essentiality in this context includes genes whose disruption causes either a growth defect or growth advantage (hypothetically corresponding to depletion or enrichment in a CRISPRi experiment). Similarly, to define a list of expected hits for rifampicin, we used a list of 75 conditionally essential genes based on exposure of the TnSeq library to rifampicin, which does not include subunits of the RNA polymerase because they are essential, but includes conditionally essential genes that might play a biological role in tolerating inhibition of transcription [37]. For levofloxacin (LEVO), we used 83 genes in the DNA damage-response pathway (based on the KEGG annotation [59]), plus *pafABC* (recently shown to be involved in DNA damage signaling [60]). Levofloxacin binds to the DNA gyrase (*gyrAB*), which produces a variety of types of damage to DNA, including double-stranded breaks, and requires several DNA replication and repair mechanisms to survive, such as recombination and the SOS response [61, 62]. The genes that will exhibit a chemical-genetic interaction with LEVO are likely to overlap substantially with some of the genes in this DNA damage-response pathway.

Each of the CRISPR analysis methods was evaluated using these approximate lists of expected hits for each drug. Since some of the methods were not designed to integrate information from multiple concentrations, the methods were initially evaluated by analyzing each concentration (LOW, MED, HIGH) of a given drug independently. Unsurprisingly, the ROC curves showed considerable dispersion of performance (Fig 7A), which was a consequence of both the method and concentration used (expected interactions were often not well-detected at low drug concentrations). Therefore, to make fairer comparisons to methods like CRISPRi-DR, CGA-LMM, and MAGeCK-MLE, we combined the results of each of the other methods over multiple concentrations by using Fisher's method [63] to combine P-values of genes at each concentration (by summing the logs of the P-values, which is similar to taking the geometric mean) and using this to re-rank the genes. This strategy for combining results from multiple concentrations produced more uniform ROC curves for all the methods, as illustrated in Fig 7B. For methods which required a single set of counts per gene, like DEBRA and CGA-LMM, the most efficient sgRNA was chosen per gene.

When the results for different concentrations were combined using Fisher's method, many of the methods exhibited reasonably good performance, ranking expected hits highly (Fig 8B–8D). For example, for INH, 50% of the expected interactions were ranked in roughly the top 20% of all genes by most of the methods, and for RIF, the identification of expected interactions (based on TnSeq) was even better (producing higher rankings of expected hits). For LEVO, the ROC curves show lower AUCs for all of the methods, probably due to the fact that not all the genes in the DNA damage response pathway are required to tolerate exposure to

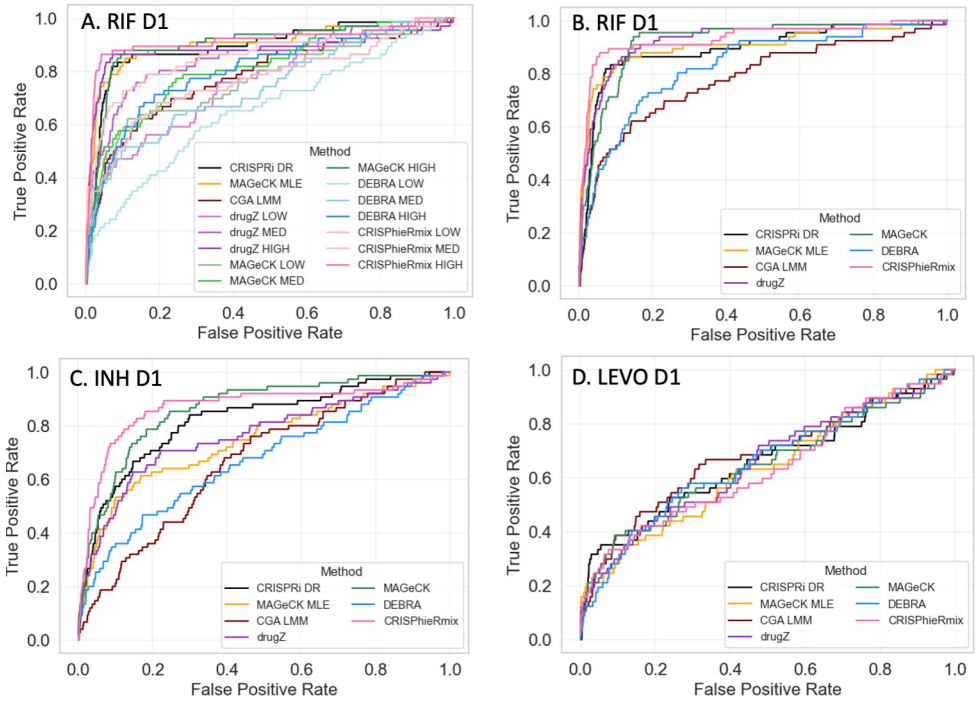

Fig 7. ROC Curves for RIF, INH and LEVO with 1 day pre-depletion. Using expected interactions derived from TnSeq studies [37] (INH and RIF) and the DNA-damage pathway (for LEVO), ROC Curves are plotted for CRISPRi-DR and 6 other CRISPR analysis methods. A) For methods that do not take concentration into account (MAGeCK, drugZ, DEBRA and CRISPhieRmix), each concentration (LOW, MED, HIGH) was analyzed independently, producing distinct ROC curves. B-D). For methods that do not take concentration into account, results of the 3 concentrations were combined using Fisher's method for combining P-values. Essential genes were filtered out, because they are not assessed by TnSeq experiments.

https://doi.org/10.1371/journal.pcbi.1011408.g007

fluoroquinolones. Though there were some variations in performance from drug to drug, indicating that differences in performance were drug-specific, the overall performance was matched fairly well, as quantified by the AUC values in Table 3. In particular, the performance of CRISPRi-DR, while not uniformly the best, was comparable to that of the other methods evaluated. It is notable methods such as CGA-LMM and DEBRA that do not account for multiple sgRNAs often had the worst performance (lowest AUC values). The similarity in performance suggests that genes that exhibited CGIs (enrichment or depletion, at least at some concentration) in this experiment were easily detected by all the methods evaluated, despite

Table 3. AUC values for 7 CRISPR analysis methods, showing comparative performance on 3 datasets (drug treatments, with 1 day of pre-depletion), based on the ROC curves in Fig 7.

	INH D1 AUCs	RIF D1 AUCs	LEVO D1 AUCs
Definition of Hits:	90 TnSeq conditional essentials (Xu et al, 2017) [37]	75 TnSeq conditional essentials (Xu et al., 2017) [37]	83 genes in DNA damage response pathway (KEGG)
CRISPRi-DR	0.821	0.894	0.655
CGA-LMM	0.665	0.782	0.679
MAGeCK-RRA	0.861	0.906	0.644
MAGeCK-MLE	0.747	0.865	0.635
drugZ	0.763	0.926	0.651
DEBRA	0.670	0.831	0.655
CRISPhieRmix	0.869	0.934	0.632

https://doi.org/10.1371/journal.pcbi.1011408.t003

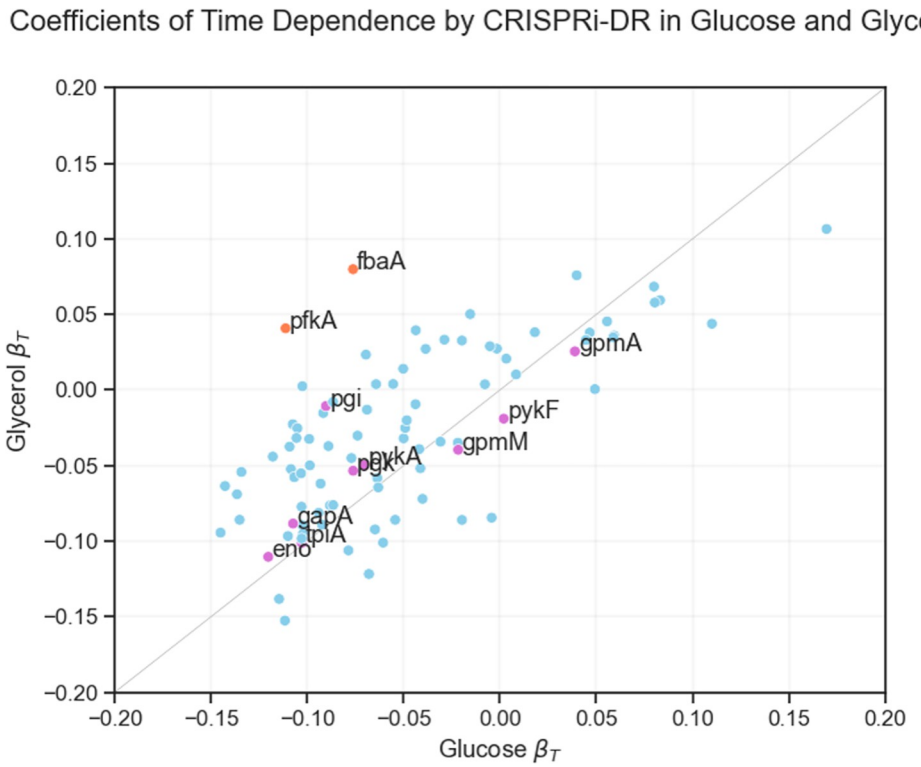

Fig 8. Coefficients of time dependence from CRISPRi-DR models fit for glucose and glycerol *E. coli* datasets. Each point in the scatterplot represents the coefficients of time dependence of a gene from the fit of the two models (glucose and glycerol). Individually, the gene show a range of growth defect over time, but the coefficients for most genes are equally negative for both conditions, except for a few outliers. The genes colored fuchsia are involved in both gluconeogenesis and glycolysis, hence, as expected, have similar time dependence coefficients in both carbon sources. The points farther away from this line, the orange labeled points (*pfkA* and *fbaA)*, are genes involved in glycolysis but not gluconeogenesis and, as expected, they have more negative coefficients in glucose than in glycerol.

https://doi.org/10.1371/journal.pcbi.1011408.g008

their different analytical frameworks. Although the AUC values for all the methods were comparable, the other methods often reported many more false positives than CRISPRi-DR. CRIPSRi-DR tends to have slightly lower recall but much higher precision than the other methods (see S2 Table), suggesting it makes more conservative calls (see Supplement). However, it has the highest F1-scores in nearly all drug screens evaluated, which reflects the best tradeoff of recall and precision.

Analysis of CRISPRi Data for *E. coli* genes required for growth on different carbon sources

To illustrate the application of the CRISPRi-DR method to other datasets, we re-analyzed the data from a CRISPRi library in *E. coli* that was used to investigate differential requirements for growth on glycerol versus glucose as a carbon source [11]. While this is not technically a chemical-genetics experiment, the data included multiple time points. The growth curves of CRISPRi knock-down mutants (depletion over time) follows sigmoidal behavior very analogous to dose-response curves for antibiotic exposure (depletion with increasing concentration). Furthermore, while only 88 genes were analyzed instead of a whole-genome screen, this dataset is suitable for analysis by CRISPRi-DR because multiple unique sgRNAs were synthesized for each gene (68 per gene on average), spanning a range of efficiencies (which were quantified by fitting growth data to a logistic curve).

We ran CRISPRi-DR on this data for each carbon source independently (fitting the model to 7 timepoints for glucose, 5 for glycerol) (see Supplemental Material for additional details). Many genes exhibited significant depletion effects (reduced fitness), because many of the 88 genes were essential for growth (on either carbon source). However, when the coefficients of the time parameter from the CRISPRi-DR analysis were plotted as a scatter plot between the carbon sources, two genes stood out as being preferentially required for growth on glucose (highlighted in orange in Fig 8, most divergent from the diagonal): *fbaA* (fructose bisphosphate aldolase) and *pfkA* (phosphofructokinase). These genes are well-known examples required for preliminary steps in glycolysis but not for incorporation of glycerol, and were identified in the analysis by [11]. Additional metabolic genes needed for growth on both carbon sources are observed to lie along the diagonal. This demonstrates that the CRISPRi-DR method can be applied to other datasets, including those not explicitly designed for chemical-genetics. The modified dose-response model nicely incorporates the simultaneous effects of time and the variable efficiency of sgRNAs on mutant abundance.

Discussion

There are a variety of ways to use CRISPRi technology for probing the biological roles of genes by modulating their expression levels in-situ. While early experiments utilized the intrinsic nuclease activity of the CAS9 to knock-out genes entirely [1–3], more recent approaches have enabled partial knock-down of targets, generally using an inactive CAS9 (dCAS9) to bind to target genes and block transcription [5]. One way of controlling the level of depletion is through manipulating the expression of the dCAS9 itself. However, a second approach to creating variability in levels of target depletion is to utilize multiple sgRNAs of different efficiency. The nucleotide sequence of both the PAM and target-specific parts of the guide RNA can impact the hybridization and recruitment of the dCAS9 [9, 10]. This variability can be useful for gauging or titrating phenotypic effects. Rather than all-or-none responses, one can look for genes whose level of depletion correlates with the phenotype of interest.

While CRISPRi libraries can be constructed with multiple sgRNAs per target, most CRISPR analytical methods do not explicitly handle such, and those that do (such as MAGeCK-RRA and CRISPhieRmix) are essentially designed to identify significant genes by focusing on a subset of apparently effective sgRNAs (i.e. allowing for ineffective sgRNAs, which are filtered out for each target). However, sgRNA efficiency can be quantified a priori, such as by running a growth experiment to determine the fitness effect of inducing the depletion of the target gene. If this information is available (collected beforehand), then it can be incorporated into the analysis as a "covariate", to enable comparison of the impact of treatment conditions on the expected magnitude of the phenotypic effect. We note that sgRNA efficiency is different than predicted strength, because it also depends on the vulnerability of the gene. In an essential gene, some sgRNAs might be more efficient than others. In contrast, typically, all the sgRNAs targeting a non-essential will turn out to be non-efficient (i.e. have 0 growth defect, or relative fitness of around 1), at least under control conditions, since the cells are unaffected by depletion of these proteins and continue to grow at the same rate. However, they might cause growth impairment if expressed in certain stress conditions where they might play a role in survival/tolerance. In fact, in chemical-genetic interaction experiments, variable sgRNA efficiency can be further exploited to identify genes whose level of depletion synergizes with increasing drug concentration. We developed the CRISPRi-DR model with this use case in mind, extending the Hill equation, which quantifies dose-response behavior of a growth inhibitor, to incorporate an extra term representing the relative efficiency of each of the sgRNAs targeting a gene. This approach, however, is not limited to CGI experiments. It can be applied to

other treatments that induce a sigmoidal response. For example, in re-analysis of data from the Mathis, Otto and Reynolds [11] paper, we showed the same equation could be adapted for modeling the effect of *E. coli* cultures grown on medium with different carbon sources; the time parameter could be substituted for the concentration, since depletion of essential genes caused a gradual killing with an S-curve shape over time.

Therefore, the CRISPRi-DR approach we developed has 3 main requirements. First, the CRISPRi library should contain multiple sgRNAs per target gene. Anecdotal evidence suggests that at least 5 sgRNAs per gene are necessary to maintain overall sensitivity for detecting expected interactions and maximizing AUC (based on experiments where we subsampled a limited number of sgRNAs per screen; see Supplement). Fewer sgRNAs per gene reduced the stability of the regression and increased variance of the fitted parameters (specifically the slope of concentration dependence). Second, ideally, sgRNAs of differing strength should be included. Strength can be predicted from sequence features using various types of trained models [9, 12]. This covers both essential and non-essential genes. For essential (or vulnerable) genes, sgRNA efficiency correlates with predicted strength, so this is equivalent to choosing sgRNAs with a range of efficiencies (that create varying growth defects). For non-essential genes, one could choose a set of sgRNAs with a range of predicted strengths, even though they might all turn out to be ineffective experimentally in standard growth conditions. This diversity could be created by selecting sgRNAs that deviate from the optimal PAM sequence [6], choosing hybridizing sequences of different length or GC content [5, 8], or adding random nucleotide substitutions [10]. Third, the actual efficiency of each sgRNA must be empirically quantified a priori, such as by running a growth experiment and comparing growth rates with and without induction of the dCAS9 (hence, with and without depletion of target genes). These quantities become inputs to the model. The CRISPRi-DR method can be applied to any CRISPRi dataset that meets these requirements. The methodology works best when treatment produces a sigmoidal effect on mutant abundances.

Doench, Fusi [9] have proposed several systems for design/optimization of CRISPRi libraries. These were more focused on minimizing off-target effects while maximizing sensitivity for detecting of genuine interactions. They do not give a specific recommendation about how many sgRNAs per gene to select. Their library design guidance is to prefer more efficient sgRNAs (e.g. Rule Set 1 selects top 20% of sgRNAs by empirical efficiency and uses these to build a model to predict sgRNA strength; Rule Set 2 extends this with a machine learning model based on additional sequence features to predict sgRNA strength, and prefers sgRNAs with highest score [9]). This contrasts with our approach, where we advocate selecting sgRNAs with a diversity of efficiencies, since we observed that the sgRNAs that exhibited the most synergy with drug treatments were not always the strongest or weakest, but somewhere in the middle of the range.

For application to CGI experiments, the availability of CRISPRi data for multiple sgRNAs of varying strengths for each target gene presents new challenges for statistical analysis. In previous work [20], we showed that regressing the relative abundances of mutants in hypomorph libraries over multiple concentrations of a drug (on log-scale) can be used to improve detection of CGIs. This regression approach captured dose-dependent behavior, i.e. genes whose decreased expression caused either suppressed or enhanced fitness that increases in magnitude with drug concentration (i.e. exhibits a trend, which is important for statistical robustness). The CRISPRi-DR method described in this paper extends this previous work by showing how effects of both drug concentration and sgRNA efficiency can be accommodated in the same model. Ideally, interacting genes would be expected to exhibit synergistic behavior with a drug, where depletion of a target protein induces excess depletion (or enrichment) of the

mutants grown in the presence of an inhibitor, and this effect is concentration-dependent (exhibits dose-response behavior).

In theory, both CRISPRi depletion of essential genes and exposure to antibiotics should impair growth of CRISPRi mutants (at least for depletion of essential genes). One might expect to observe a depletion effect due to either increasing sgRNA efficiency, or drug concentration, each producing regression "slopes" (in log-transformed space), with slopes for sgRNAs targeting non-essential genes being expected to be flat, regardless of predicted sgRNA strength. However, we observed that sgRNA efficiency and concentration effects are not independent—they interact in a non-linear way. sgRNAs that are too weak do not produce enough depletion of a drug target to cause sensitization, and sgRNAs that are too strong deplete a mutant to such low abundances that concentration-dependent effects are difficult to quantify. Often, there is a "sweet spot", or an intermediate sgRNA efficiency (effect on growth phenotype) which maximizes the concentration-dependent effect (which could be different for each gene). Our CRISPRi-DR model incorporates both sgRNA efficiency and drug concentration as parameters, and captures the non-linear interaction between them, where the "slopes" for the effect of drug concentration on relative abundance of mutants can be larger in magnitude for sgRNAs of intermediate efficiency, while being flatter (slopes closer to 0) for sgRNAs of high or low efficiency. MAGeCK-MLE is the only other analytical method that take sgRNA efficiencies as an input; in that method, the empirical measures of efficiency are used to initialize the prior probability that each sgRNA is effective (assuming each gene is represented by a subset of sgRNAs that are effective and others that are not), which is combined with other conditional probabilities in a Bayesian framework to determine the posterior probability of interaction for each gene. However, we observed that MAGeCK-MLE often reports far more significant interactions that CRISPRi-DR or several other methods and has lower precision.

In this paper, we showed that this non-linear interaction between sgRNA efficiency and drug concentration can be modeled using an augmented dose-response equation, in which terms for both effects are included. By fitting the parameters in this equation to CRISPRi data from a CGI experiment (normalized mutant abundances from sgRNA counts), one can estimate the degree to which depletion of a given gene sensitizes cells to an inhibitor, and thereby identify CGIs. While various computational methods exist for fitting non-linear equations, such as the Levenberg–Marquardt algorithm [64], we chose to linearize the modified Hill equation by applying a log-sigmoid transform. The transformation enables us to express the equation in a linear form, where the parameters (IC_{50}, Hill slopes, etc.) appear as coefficients of linear terms or constants. Consequently, we can use ordinary least-squares regression (OLS) to fit the model to the CRISPRi dataset.

Sometimes, positive and/or negative controls are included in a CRISPRi experiment [8]. While negative controls can be used in methods like MAGeCK-RRA, CRISPRi-DR is not designed to use controls explicitly in the statistical analysis of CGIs. Hypothetically, negative controls could be used in the final filtering step to calculate Z-scores for each gene. Instead of determining significance of genes based on the null distribution from the empirical Bayes FDR analysis, they could be based on the distribution of slope coefficients from the negative controls. While we tested this idea (using 1750 non-targeting sgRNAs included in the *Mtb* CRISPRi dataset as negative controls), it resulted in many more genes being labeled as interactions (up to half the genome). It appears that unrelated genes (not involved in the mechanism of action or resistance to a drug) often have slightly positive or negative random slopes, due to some source of noise in the experiment that is unaccounted for. Some genes could exhibit weak phenotypic effects, conferring slight growth defects or advantages under antibiotic stress, even though they do not play any direct role in the mechanism of action or resistance to the drug. This is the reason that we advocate identifying genes that are outliers with respect to the

rest of the population of genes, achieved through the empirical Bayes filtering step at the end, instead of just reporting all genes with slope coefficient statistically different from 0.

We compared CRISPRi-DR to several other analytical methods, including MAGeCK-RRA, MAGeCK-MLE, DEBRA, CRISPhieRMix, CGA-LMM, and drugZ. Some of these methods incorporate multiple drug concentrations, while other incorporate sgRNA efficiency as an input to their models. However, only MAGeCK-MLE incorporates both types of input. The importance of incorporating both inputs in CRISPRi-DR was demonstrated via an experiment with ablated models; the model fits (AICs) for each gene were significantly worse for models that regressed abundances against either drug concentration or sgRNA efficiency alone. For those methods that do not explicitly combine data from multiple drug concentrations and must be run on each concentration independently, we employed Fisher's method of combining P-values to create a merged ranking of genes. Using ROC curves to compare ranking of expected interactions, CRISPRi-DR performed comparably to the best of these methods, though the method with the highest AUC differed depending on the drug. This evaluation was facilitated by using lists of conditionally essential genes from TnSeq experiments (exposure to same drugs) to define an objective list of expected interactions for each drug for making fair comparisons of performance. However, a major difference observed among the methods was in the number of significant interactions detected. Methods like CRISPRhieRMix, DEBRA, MAGeCK-RRA, and MAGeCK-MLE produced hundreds to thousands of hits for each drug, whereas CRISPRi-DR reported a more conservative list of typically less than a hundred interacting genes. It is likely that many of the interactions detected by the former methods could be false positives. This was borne out in simulation experiments, where MAGECK-RRA, and MAGeCK-MLE exhibited substantially lower precision than CRISPRi-DR. In both the simulated data and real drug screen datasets, CRISPRi-DR had the highest F1-scores, reflecting the best tradeoff between precision and recall compared to other methods. Reducing false positives is important because experimental validation of hits can be expensive, and follow-up is usually only applied to a handful of top-ranked genes. Furthermore, we used simulated datasets to explore how noise within or between drug concentrations could affect both the recall and precision of CRISPRi-DR, MAGECK-RRA, and MAGeCK-MLE. Both types of noise increasingly degrade the recall of all methods, but noise within concentrations (i.e. sgRNA counts among replicates) seemed to cause the greatest decrease in precision, especially for MAGeCK-RRA. Using the joint criterion for significance based on the empirical Bayes FDR analysis and the Wald test in CRISPRi-DR partially helps to mitigate this, producing a more focused list of candidate interactions and eliminating genes with small random slopes of concentration dependence that are not likely to be genuine interactions (i.e. false positives).

Supporting information

S1 Fig. Evaluation sgRNA efficiency 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, along with the AIC comparisons and Log-likelihood tests, indicate sgRNA efficiency 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 efficiency and log concentration is needed for accurate assessment of significant sgRNA depletion in a gene in a condition.
(TIFF)

S2 Fig. Distribution of average CV of sgRNAs in significant genes (depleted and enriched) in the CRISPRi-DR model and MAGeCK. Comparisons of distributions of coefficient of variation (noise) for significant genes found by MAGeCK and the CRISPRi-DR model for all experiments. The dashed panel is the noise distributions for RIF D10, seen in Fig 5. The trend seen in RIF D10 is present with all the experiments except LEVO D10. The distribution of noise for hits found by the CRISPRi-DR model is unimodal with a low CV as the mode, whereas MAGeCK finds significant genes with low average CV values but also a substantial 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.
(TIFF)

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 2, 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.
(XLSX)

S2 Table. Comparison of significant interactions identified by CRISPR analysis methods of EMB, INH, LEVO, VAN and RIF CRISPRi screens. For each drug and pre-depletion day of the selected datasets, all 7 CRIPSR methods were run. For methods that do not account for multiple concentrations, they were run separately for each concentration and the overall significant interactions are also addressed post-combination of the individual runs using Fisher's method. The comparison of the significant interactions identified by the models was evaluated using an objectively defined list of true positives. The conditionally essential genes identified by Xu, DeJesus (37) were used as the "ground truth" against which the other model's results were compared. For LEVO, genes in the DNA Damage Response pathway are used. Recall, Precision and F1-score columns are colored such that higher values are darker green.
(XLSX)

S3 Table. Matrices for comparison of significant interactions identified by CRISPRi-DR and MAGeCK for each drug and pre-depletion day. The table presents the results of CRISPRi-DR and MAGeCK analyses for different drugs and pre-depletion days. Significant interactions are compared in matrix form. Cells with red font indicate low overlaps between the interactions found by the two models, while cells with green font represent high overlaps.
(XLSX)

S1 Text. We expand on the following four topics from the main text in this document: 1) An assessment of CRISPRi-DR, MAGeCK and MAGeCK-MLE on datasets with simulated noise, 2) Comparison of CRISPRi-DR to other analysis methods using CGI datasets, 3) Analysis of *E. coli* CRISPRi screens using CRISPRi-DR and, 4) The minimum number of sgRNAs recommended per gene in CRISPRi-DR.
(PDF)

Author Contributions

Conceptualization: Thomas R. Ioerger.

Funding acquisition: Dirk Schnappinger, Thomas R. Ioerger.

Investigation: Sanjeevani Choudhery.

Methodology: Sanjeevani Choudhery, Thomas R. Ioerger.

Software: Sanjeevani Choudhery.

Visualization: Aarthi Srinivasan.

Writing – original draft: Sanjeevani Choudhery, Thomas R. Ioerger.

Writing – review & editing: Sanjeevani Choudhery, Michael A. DeJesus, Aarthi Srinivasan, Jeremy Rock, Dirk Schnappinger, Thomas R. Ioerger.

References

1. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012; 337(6096):816–21. Epub 20120628. https://doi.org/10.1126/science.1225829 PMID: 22745249; PubMed Central PMCID: PMC6286148.
2. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339(6121):823–6. Epub 20130103. https://doi.org/10.1126/science.1232033 PMID: 23287722; PubMed Central PMCID: PMC3712628.
3. Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. Cell. 2013; 154 (6):1370–9. Epub 20130829. https://doi.org/10.1016/j.cell.2013.08.022 PMID: 23992847; PubMed Central PMCID: PMC3961003.
4. Jensen TI, Mikkelsen NS, Gao Z, Foßelteder J, Pabst G, Axelgaard E, et al. Targeted regulation of transcription in primary cells using CRISPRa and CRISPRi. Genome Res. 2021; 31(11):2120–30. Epub 20210818. https://doi.org/10.1101/gr.275607.121 PMID: 34407984; PubMed Central PMCID: PMC8559706.
5. Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell. 2013; 152(5):1173–83. https://doi.org/10.1016/j.cell.2013.02.022 PMID: 23452860; PubMed Central PMCID: PMC3664290.
6. Rock JM, Hopkins FF, Chavez A, Diallo M, Chase MR, Gerrick ER, et al. Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. Nat Microbiol. 2017; 2:16274. Epub 20170206. https://doi.org/10.1038/nmicrobiol.2016.274 PMID: 28165460; PubMed Central PMCID: PMC5302332.
7. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, et al. A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. Cell. 2016; 165(6):1493–506. Epub 20160526. https://doi.org/10.1016/j.cell.2016.05.003 PMID: 27238023; PubMed Central PMCID: PMC4894308.
8. Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014; 159(3):647–61. Epub 20141009. https://doi.org/10.1016/j.cell.2014.09.029 PMID: 25307932; PubMed Central PMCID: PMC4253859.
9. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016; 34(2):184–91. Epub 20160118. https://doi.org/10.1038/nbt.3437 PMID: 26780180; PubMed Central PMCID: PMC4744125.
10. Hawkins JS, Silvis MR, Koo BM, Peters JM, Osadnik H, Jost M, et al. Mismatch-CRISPRi Reveals the Co-varying Expression-Fitness Relationships of Essential Genes in Escherichia coli and Bacillus subtilis. Cell Syst. 2020; 11(5):523–35.e9. Epub 20201019. https://doi.org/10.1016/j.cels.2020.09.009 PMID: 33080209; PubMed Central PMCID: PMC7704046.
11. Mathis AD, Otto RM, Reynolds KA. A simplified strategy for titrating gene expression reveals new relationships between genotype, environment, and bacterial growth. Nucleic Acids Research. 2020; 49(1): e6–e. https://doi.org/10.1093/nar/gkaa1073 PMID: 33221881
12. Bosch B, DeJesus MA, Poulton NC, Zhang W, Engelhart CA, Zaveri A, et al. Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. Cell. 2021; 184(17):4579–92 e24. Epub 20210722. https://doi.org/10.1016/j.cell.2021.06.033 PMID: 34297925; PubMed Central PMCID: PMC8382161.
13. Li S, Poulton NC, Chang JS, Azadian ZA, DeJesus MA, Ruecker N, et al. CRISPRi chemical genetics and comparative genomics identify genes mediating drug potency in Mycobacterium tuberculosis. Nat

Microbiol. 2022; 7(6):766–79. Epub 20220530. https://doi.org/10.1038/s41564-022-01130-y PMID: 35637331; PubMed Central PMCID: PMC9159947.

14. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014; 15(12):554. https://doi.org/10.1186/s13059-014-0554-4 PMID: 25476604; PubMed Central PMCID: PMC4290824.
15. Li W, Koster J, Xu H, Chen CH, Xiao T, Liu JS, et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biol. 2015; 16:281. Epub 20151216. https://doi.org/10.1186/s13059-015-0843-6 PMID: 26673418; PubMed Central PMCID: PMC4699372.
16. Daley TP, Lin Z, Lin X, Liu Y, Wong WH, Qi LS. CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. Genome Biol. 2018; 19(1):159. Epub 20181008. https://doi.org/10.1186/s13059-018-1538-6 PMID: 30296940; PubMed Central PMCID: PMC6176515.
17. Colic M, Wang G, Zimmermann M, Mascall K, McLaughlin M, Bertolet L, et al. Identifying chemogenetic interactions from CRISPR screens with drugZ. Genome Med. 2019; 11(1):52. Epub 20190822. https://doi.org/10.1186/s13073-019-0665-3 PMID: 31439014; PubMed Central PMCID: PMC6706933.
18. Akimov Y, Bulanova D, Timonen S, Wennerberg K, Aittokallio T. Improved detection of differentially represented DNA barcodes for high-throughput clonal phenomics. Mol Syst Biol. 2020; 16(3):e9195. https://doi.org/10.15252/msb.20199195 PMID: 32187448; PubMed Central PMCID: PMC7080434.
19. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014; 32(12):1262–7. Epub 20140903. https://doi.org/10.1038/nbt.3026 PMID: 25184501; PubMed Central PMCID: PMC4262738.
20. Dutta E, DeJesus MA, Ruecker N, Zaveri A, Koh EI, Sassetti CM, et al. An improved statistical method to identify chemical-genetic interactions by exploiting concentration-dependence. PLoS One. 2021; 16(10):e0257911. Epub 20211001. https://doi.org/10.1371/journal.pone.0257911 PMID: 34597304; PubMed Central PMCID: PMC8486102.
21. Wiser MJ, Lenski RE. A Comparison of Methods to Measure Fitness in Escherichia coli. PLoS One. 2015; 10(5):e0126210. Epub 20150511. https://doi.org/10.1371/journal.pone.0126210 PMID: 25961572; PubMed Central PMCID: PMC4427439.
22. Wald A. The Fitting of Straight Lines if Both Variables are Subject to Error. The Annals of Mathematical Statistics. 1940; 11(3):284–300.
23. Benjamini Y, Krieger AM, Yekutieli D. Adaptive Linear Step-up Procedures That Control the False Discovery Rate. Biometrika. 2006; 93(3):491–507.
24. Efron B. Size, power and false discovery rates. The Annals of Statistics. 2007; 35(4):1351–77, 27.
25. Efron B. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. Journal of the American Statistical Association. 2004; 99(465):96–104.
26. DeJesus MA, Gerrick ER, Xu W, Park SW, Long JE, Boutte CC, et al. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. mBio. 2017; 8(1). Epub 20170117. https://doi.org/10.1128/mBio.02133-16 PMID: 28096490; PubMed Central PMCID: PMC5241402.
27. de Vos M, Muller B, Borrell S, Black PA, van Helden PD, Warren RM, et al. Putative compensatory mutations in the rpoC gene of rifampin-resistant Mycobacterium tuberculosis are associated with ongoing transmission. Antimicrob Agents Chemother. 2013; 57(2):827–32. Epub 20121203. https://doi.org/10.1128/AAC.01541-12 PMID: 23208709; PubMed Central PMCID: PMC3553702.
28. Guo H, Courbon GM, Bueler SA, Mai J, Liu J, Rubinstein JL. Structure of mycobacterial ATP synthase bound to the tuberculosis drug bedaquiline. Nature. 2021; 589(7840):143–7. Epub 20201209. https://doi.org/10.1038/s41586-020-3004-3 PMID: 33299175.
29. Kaniga K, Lounis N, Zhuo S, Bakare N, Andries K. Impact of Rv0678 mutations on patients with drug-resistant TB treated with bedaquiline. Int J Tuberc Lung Dis. 2022; 26(6):571–3. https://doi.org/10.5588/ijtld.21.0670 PMID: 35650698; PubMed Central PMCID: PMC9165736.
30. Mayer C, Takiff H. The Molecular Genetics of Fluoroquinolone Resistance in Mycobacterium tuberculosis. Microbiol Spectr. 2014; 2(4):MGM2-0009-2013. https://doi.org/10.1128/microbiolspec.MGM2-0009-2013 PMID: 26104201.
31. Cui Z, Li Y, Cheng S, Yang H, Lu J, Hu Z, Ge B. Mutations in the embC-embA intergenic region contribute to Mycobacterium tuberculosis resistance to ethambutol. Antimicrob Agents Chemother. 2014; 58(11):6837–43. Epub 20140902. https://doi.org/10.1128/AAC.03285-14 PMID: 25182646; PubMed Central PMCID: PMC4249443.
32. Zhang L, Zhao Y, Gao Y, Wu L, Gao R, Zhang Q, et al. Structures of cell wall arabinosyltransferases with the anti-tuberculosis drug ethambutol. Science. 2020; 368(6496):1211–9. Epub 20200423. https://doi.org/10.1126/science.aba9102 PMID: 32327601.

33. Mougari F, Bouziane F, Crockett F, Nessar R, Chau F, Veziris N, et al. Selection of Resistance to Clarithromycin in Mycobacterium abscessus Subspecies. Antimicrob Agents Chemother. 2017; 61(1). Epub 20161227. https://doi.org/10.1128/AAC.00943-16 PMID: 27799212; PubMed Central PMCID: PMC5192163.

34. Gan WC, Ng HF, Ngeow YF. Mechanisms of Linezolid Resistance in Mycobacteria. Pharmaceuticals (Basel). 2023; 16(6). Epub 20230524. https://doi.org/10.3390/ph16060784 PMID: 37375732; PubMed Central PMCID: PMC10303974.

35. Wong SY, Lee JS, Kwak HK, Via LE, Boshoff HI, Barry CE 3rd. Mutations in gidB confer low-level streptomycin resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2011; 55(6):2515–22. Epub 20110328. https://doi.org/10.1128/AAC.01814-10 PMID: 21444711; PubMed Central PMCID: PMC3101441.

36. Alam MT, Petit RA, 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014; 6(5):1174–85. Epub 20140430. https://doi.org/10.1093/gbe/evu092 PMID: 24787619; PubMed Central PMCID: PMC4040999.

37. Xu W, DeJesus MA, Rucker N, Engelhart CA, Wright MG, Healy C, et al. Chemical Genetic Interaction Profiling Reveals Determinants of Intrinsic Antibiotic Resistance in Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2017; 61(12). Epub 20171122. https://doi.org/10.1128/AAC.01334-17 PMID: 28893793; PubMed Central PMCID: PMC5700314.

38. Palomino JC, Martin A. Drug Resistance Mechanisms in Mycobacterium tuberculosis. Antibiotics (Basel). 2014; 3(3):317–40. Epub 20140702. https://doi.org/10.3390/antibiotics3030317 PMID: 27025748; PubMed Central PMCID: PMC4790366.

39. Vilcheze C, Jacobs WR Jr. The mechanism of isoniazid killing: clarity through the scope of genetics. Annu Rev Microbiol. 2007; 61:35–50. https://doi.org/10.1146/annurev.micro.61.111606.122346 PMID: 18035606.

40. Kreutzfeldt KM, Jansen RS, Hartman TE, Gouzy A, Wang R, Krieger IV, et al. CinA mediates multidrug tolerance in Mycobacterium tuberculosis. Nat Commun. 2022; 13(1):2203. Epub 20220422. https://doi.org/10.1038/s41467-022-29832-1 PMID: 35459278; PubMed Central PMCID: PMC9033802.

41. Vilcheze C, Av-Gay Y, Barnes SW, Larsen MH, Walker JR, Glynne RJ, Jacobs WR Jr. Coresistance to isoniazid and ethionamide maps to mycothiol biosynthetic genes in Mycobacterium bovis. Antimicrob Agents Chemother. 2011; 55(9):4422–3. Epub 20110627. https://doi.org/10.1128/AAC.00564-11 PMID: 21709101; PubMed Central PMCID: PMC3165297.

42. Vilcheze C, Weisbrod TR, Chen B, Kremer L, Hazbon MH, Wang F, et al. Altered NADH/NAD+ ratio mediates coresistance to isoniazid and ethionamide in mycobacteria. Antimicrob Agents Chemother. 2005; 49(2):708–20. https://doi.org/10.1128/AAC.49.2.708-720.2005 PMID: 15673755; PubMed Central PMCID: PMC547332.

43. Hazbón MH, Brimacombe M, Bobadilla del Valle M, Cavatore M, Guerrero MI, Varma-Basil M, et al. Population genetics study of isoniazid resistance mutations and evolution of multidrug-resistant Mycobacterium tuberculosis. Antimicrob Agents Chemother. 2006; 50(8):2640–9. https://doi.org/10.1128/AAC.00112-06 PMID: 16870753; PubMed Central PMCID: PMC1538650.

44. Bollela VR, Namburete EI, Feliciano CS, Macheque D, Harrison LH, Caminero JA. Detection of katG and inhA mutations to guide isoniazid and ethionamide use for drug-resistant tuberculosis. Int J Tuberc Lung Dis. 2016; 20(8):1099–104. https://doi.org/10.5588/ijtld.15.0864 PMID: 27393546; PubMed Central PMCID: PMC5310937.

45. Giri A, Gupta S, Safi H, Narang A, Shrivastava K, Kumar Sharma N, et al. Polymorphisms in Rv3806c (ubiA) and the upstream region of embA in relation to ethambutol resistance in clinical isolates of Mycobacterium tuberculosis from North India. Tuberculosis (Edinb). 2018; 108:41–6. Epub 20171012. https://doi.org/10.1016/j.tube.2017.10.003 PMID: 29523326.

46. McNeil MB, Chettiar S, Awasthi D, Parish T. Cell wall inhibitors increase the accumulation of rifampicin in Mycobacterium tuberculosis. Access Microbiol. 2019; 1(1):e000006. Epub 20190320. https://doi.org/10.1099/acmi.0.000006 PMID: 32974492; PubMed Central PMCID: PMC7470358.

47. Patel Y, Soni V, Rhee KY, Helmann JD. Mutations in rpoB That Confer Rifampicin Resistance Can Alter Levels of Peptidoglycan Precursors and Affect β-Lactam Susceptibility. mBio. 2023; 14(2):e0316822. Epub 20230213. https://doi.org/10.1128/mbio.03168-22 PMID: 36779708; PubMed Central PMCID: PMC10128067.

48. Campodonico VL, Rifat D, Chuang YM, Ioerger TR, Karakousis PC. Altered Mycobacterium tuberculosis Cell Wall Metabolism and Physiology Associated With RpoB Mutation H526D. Front Microbiol. 2018; 9:494. Epub 20180319. https://doi.org/10.3389/fmicb.2018.00494 PMID: 29616007; PubMed Central PMCID: PMC5867343.

49. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021; 49 (D1):D274–D81. https://doi.org/10.1093/nar/gkaa1018 PMID: 33167031; PubMed Central PMCID: PMC7778934.

50. Provvedi R, Boldrin F, Falciani F, Palu G, Manganelli R. Global transcriptional response to vancomycin in Mycobacterium tuberculosis. Microbiology (Reading). 2009; 155(Pt 4):1093–102. https://doi.org/10.1099/mic.0.024802-0 PMID: 19332811.

51. Soetaert K, Rens C, Wang XM, De Bruyn J, Laneelle MA, Laval F, et al. Increased Vancomycin Susceptibility in Mycobacteria: a New Approach To Identify Synergistic Activity against Multidrug-Resistant Mycobacteria. Antimicrob Agents Chemother. 2015; 59(8):5057–60. Epub 20150601. https://doi.org/10.1128/AAC.04856-14 PMID: 26033733; PubMed Central PMCID: PMC4505240.

52. Palmer AC, Kishony R. Opposing effects of target overexpression reveal drug mechanisms. Nat Commun. 2014; 5:4296. Epub 20140701. https://doi.org/10.1038/ncomms5296 PMID: 24980690; PubMed Central PMCID: PMC4408919.

53. Hansen JL, Ippolito JA, Ban N, Nissen P, Moore PB, Steitz TA. The structures of four macrolide antibiotics bound to the large ribosomal subunit. Mol Cell. 2002; 10(1):117–28. https://doi.org/10.1016/s1097-2765(02)00570-1 PMID: 12150912.

54. Chulluncuy R, Espiche C, Nakamoto JA, Fabbretti A, Milón P. Conformational Response of 30S-bound IF3 to A-Site Binders Streptomycin and Kanamycin. Antibiotics (Basel). 2016; 5(4). Epub 20161213. https://doi.org/10.3390/antibiotics5040038 PMID: 27983590; PubMed Central PMCID: PMC5187519.

55. Spies FS, Ribeiro AW, Ramos DF, Ribeiro MO, Martin A, Palomino JC, et al. Streptomycin resistance and lineage-specific polymorphisms in Mycobacterium tuberculosis gidB gene. J Clin Microbiol. 2011; 49(7):2625–30. Epub 20110518. https://doi.org/10.1128/JCM.00168-11 PMID: 21593257; PubMed Central PMCID: PMC3147840.

56. Cui ZL, Xiaojun; Shin Joonyoung; Gamper Howard; Hou Ya-Ming; Sacchettini, James C; Zhang, Junjie Interplay between an ATP-binding cassette F protein and the ribosome from Mycobacterium tuberculosis. Nature Communications. 2022. PubMed Central PMCID: PMC35064151.

57. Madsen CT, Jakobsen L, Buriankova K, Doucet-Populaire F, Pernodet JL, Douthwaite S. Methyltransferase Erm(37) slips on rRNA to confer atypical resistance in Mycobacterium tuberculosis. J Biol Chem. 2005; 280(47):38942–7. Epub 20050920. https://doi.org/10.1074/jbc.M505727200 PMID: 16174779.

58. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015; 31(17):2778–84. Epub 20150428. https://doi.org/10.1093/bioinformatics/btv272 PMID: 25926345; PubMed Central PMCID: PMC4635655.

59. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023; 51(D1):D587–D92. https://doi.org/10.1093/nar/gkac963 PMID: 36300620; PubMed Central PMCID: PMC9825424.

60. Fudrini Olivencia B, Muller AU, Roschitzki B, Burger S, Weber-Ban E, Imkamp F. Mycobacterium smegmatis PafBC is involved in regulation of DNA damage response. Sci Rep. 2017; 7(1):13987. Epub 20171025. https://doi.org/10.1038/s41598-017-14410-z PMID: 29070902; PubMed Central PMCID: PMC5656591.

61. Diaz-Diaz S, Recacha E, Machuca J, Garcia-Duque A, Docobo-Perez F, Blazquez J, et al. Synergistic Quinolone Sensitization by Targeting the recA SOS Response Gene and Oxidative Stress. Antimicrob Agents Chemother. 2021; 65(4). Epub 20210318. https://doi.org/10.1128/AAC.02004-20 PMID: 33526493; PubMed Central PMCID: PMC8097469.

62. Tran T, Ran Q, Ostrer L, Khodursky A. De Novo Characterization of Genes That Contribute to High-Level Ciprofloxacin Resistance in Escherichia coli. Antimicrob Agents Chemother. 2016; 60(10):6353–5. Epub 20160923. https://doi.org/10.1128/AAC.00889-16 PMID: 27431218; PubMed Central PMCID: PMC5038283.

63. Mosteller F, Fisher RA. Questions and Answers. The American Statistician. 1948; 2(5):30–1. https://doi.org/10.2307/2681650

64. 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. https://doi.org/10.1063/1.4993697 PMID: 29195386.