Title
Model-driven optimization of high-throughput in vivo CRISPR screen design

Author
Kim, Sandy Sung

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Model-driven optimization of high-throughput *in vivo* CRISPR screen design

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Sandy Sung Kim

2021

ABSTRACT OF THE THESIS

Model-driven optimization of
high-throughput *in vivo* CRISPR screen design

by

Sandy Sung Kim
Master of Science in Bioinformatics
University of California, Los Angeles, 2021
Professor Harold J. Pimentel, Chair

The recent developments of the CRISPR/Cas9 gene-editing system have made way for large-scale, loss-of-function genetic screens that can identify genes underlying a given phenotype, known as high-throughput CRISPR screens. By leveraging the precision of CRISPR/Cas9 and the capacity to capture millions of cells in one library preparation, these screens enrich and deplete the expression of various specific genes, identifying gene functions that help elucidate genotype-phenotype relationships. Furthermore, by modulating genetic interactions, these screens can uncover gene regulatory mechanisms, revealing genetic dependencies. Although these screens have shown to be incredibly effective, they are often prohibitively expensive. Additionally, there is a lack of information and tools to determine the optimal experimental design, such that the most informative data is produced, given experimental constraints.

Here, we introduce a statistical model that simulates high-throughput *in vivo* CRISPR screens to provide insight into optimizing the experimental protocol. We first demonstrate our model successfully simulates such screens by comparing the generated data with real ex-

perimental data. Then, we simulate screens across varying parameter inputs and investigate their influence on statistical power. Given our findings, we conclude with general guidelines and suggestions for effectively designing high-throughput *in vivo* CRISPR screens.

The thesis of Sandy Sung Kim is approved.

Bogdan Pasaniuc

Eleazar Eskin

Luis de la Torre-Ubieta

Harold J. Pimentel, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, to my thesis advisor: Harold, thank you for generosity and patience. I contacted you expressing interest in working with you before you even began at UCLA, and you still took me on, right there and then. It's been an honor to work on this project with you and to be your first student; I've grown a lot as a researcher. I'm extremely grateful to have an incredible mentor who believes in me and always has my back– I wouldn't be where I am without your guidance. It's been really exciting to see the lab grow for the past year and I can't wait for all the cool work that's to come!

To my thesis committee members: Bogdan, thank you for cheering me on and welcoming the Pimentel Lab to Bogdan Lab group meetings and socials. It's been really fun! Luis, thank you very kindly agreeing to be on my committee even though we had never met prior. The feedback and questions you provided for my thesis were invaluable; it was immensely useful to get an experimental perspective on the project. Eleazar, thank you for giving me direction last year when I was lost on what I wanted to do for my Master's thesis research– you were actually the one directed me to Harold. Also, you're a lifesaver for stepping in to be part of my committee last minute to make completing my thesis on time possible.

To my first research advisor: Eric, thank you for planting the seed by providing me with a great first research experience two years ago. You've been always been supportive of all my academic endeavors both outside and at UCLA.

To my friends: thank you for being there for me, especially in the midst of the COVID-19 pandemic. You all got me through this crazy year.

Last but not least, to my family: Mom, Dad, and my older brother Sean, thank you for loving me unconditionally. This one's for you.

This work was funded by the HHMI Hanna H. Gray Fellowship awarded to H.J. Pimentel.

CHAPTER 1

Introduction

The discovery and recent developments of the CRISPR/Cas9 gene-editing system have greatly expanded the toolbox for genetics [JCF12, HLZ14, WND16, KD18, SJ14]. In particular, the CRISPR/Cas9 system has shown to be incredibly useful in large-scale, loss-of-function screens that can identify sets of genes underlying various phenotypes and functions, known as high-throughput CRISPR screens [SSZ15, DPL16].

By using a genome-scale, single-guide RNA (sgRNA or guide) expressing lentiviral pool, a library of knockout cells is generated and screened under positive and negative selection. Each guide can then serve as a distinct DNA barcode that can be used to measure guide abundance across multiple cells via high-throughput sequencing [SSZ15].

The utility of CRISPR/Cas9 for conducting large-scale genetic screens has shown to outperform other current functional screening methods such as RNA interference (RNAi) [MDL16]. This is largely due to the precision of lentiviral transductions in such screens, enabling the ability to effectively introduce CRISPR/Cas9 into a given cell. High-throughput CRISPR knockout screens offer other powerful features such as inactivating genes at the DNA level, measuring multiple cell phenotypes at once, and relatively precise targeting with significantly less off-target effects compared to other methods [MDL16, GHA14]. As a result, CRISPR/Cas9 screens have grown increasingly popular in recent years.

Most CRISPR screens have been performed *in vitro*, using cell lines [GHA14]. However, given the complexity of physiological cues in living organisms and the growing accessibility of CRISPR technology, such screens are increasingly being conducted *in vivo* [CRC13, KLT14,

WWS14]. Many of these studies contribute primarily in identifying immuno-oncology targets and pathways [WLZ19, PSK17, PCZ14, CPZ15, KCR16, SLM17]. *In vivo* enrichment CRISPR screens are performed by infecting naïve cells from an organism with a lentiviral vector, and identifying enriched and depleted genes within the infected cells to identify clinically relevant targets [DWC19, WLZ19].

Despite the popularity of such screens, there are an overall lack of methods to optimize these experiments. While there are methods to help guide the design of the CRISPR/Cas9 complex to increase Cas9 specificity such as detecting off-target effects and selecting sgRNAs, there are very few that consider the experimental materials [DFS16, TMH16, CWL17, HSW13]. One method that aims to optimize the protocol, MAUDE, helps infer expression changes in sorting-based CRISPR screens [BRH20]. However, MAUDE strictly optimizes across varying number and sizes of expression bins, which is a feature unique to only sorting-based screens. There are also models for log fold changes across control and treatment guide read counts produced by high-throughput enrichment CRISPR screens such as those used in MAGeCK, a tool that identifies significant guides and genes from such data [LXX14]. However, MAGeCK's model is overly simplistic, relying on strong assumptions of the distribution of the guide read counts. Most importantly, neither model takes into account how the design of each step in the protocol may affect the resulting data.

To our knowledge, there exists no method that seeks to produce the most informative data in an experiment as a function of the experimental steps and parameters for high-throughput *in vivo* pooled enrichment CRISPR screens. As a result, experimentalists performing such screens often rely on heuristic processes, such as following the protocols of previous successful and similar experiments, among many other naïve procedures. This evident lack of information in designing high-throughput *in vivo* CRISPR screens ultimately result in these experiments to likely become prohibitively expensive.

In this work, we present a statistical model that simulates a high-throughput *in vivo* CRISPR/Cas9 screen. We validate that our model simulates realistic screens by comparing

the results of differential guide analysis on the generated data and real experimental data. After, we perform simulations under unobserved conditions and investigate how varying different parameters in the experimental protocol affects the statistical power of the resulting data using existing inference methods. We conclude with remarks to guide experimentalists in optimizing the design of their high-throughput *in vivo* CRISPR experiments.

CHAPTER 2

Methods

2.1 High-level description of CRISPR screen

We first give a high-level description of the experimental design of the high-throughput *in vivo* CRISPR screen we emulate in our model [DWC19]. A schematic is shown in Figure 2.1.

First, a lentiviral CRISPR vector is designed, generated, and cloned into a mouse genome-scale guide library containing both gene-specific guides and non-targeting controls. Naïve $CD8^+$ T cells (control cells) are isolated from the OT-I;Cas9 mice. The guide library lentivirus vectors are transduced into the isolated control cells and the guides of the cells are sequenced using high-throughput technology in order to measure guide abundance. The infected cells are then transferred into the tumors of tumor-bearing OT-I;Cas-9 mice. After a few days, the tumors are harvested and the guides of the cells (treatment cells) are sequenced using high-throughput technology to measure guide abundance.

The reads of the guides of the control and treatment cells are aligned to the mouse genome using read-alignment software. Using sequencing data analysis software, read counts from the guides of the treatment cells are normalized and compared relative to the normalized read counts from the guides of the control cells to identify enriched (positively selected) and depleted (negatively selected) genes and guides as a result of the screen. Enriched genes will have a significantly higher read counts for their corresponding guides in the treatment cells than that of controls cells. Conversely, depleted genes will have a significantly higher read counts in control cells than that of treatment cells.

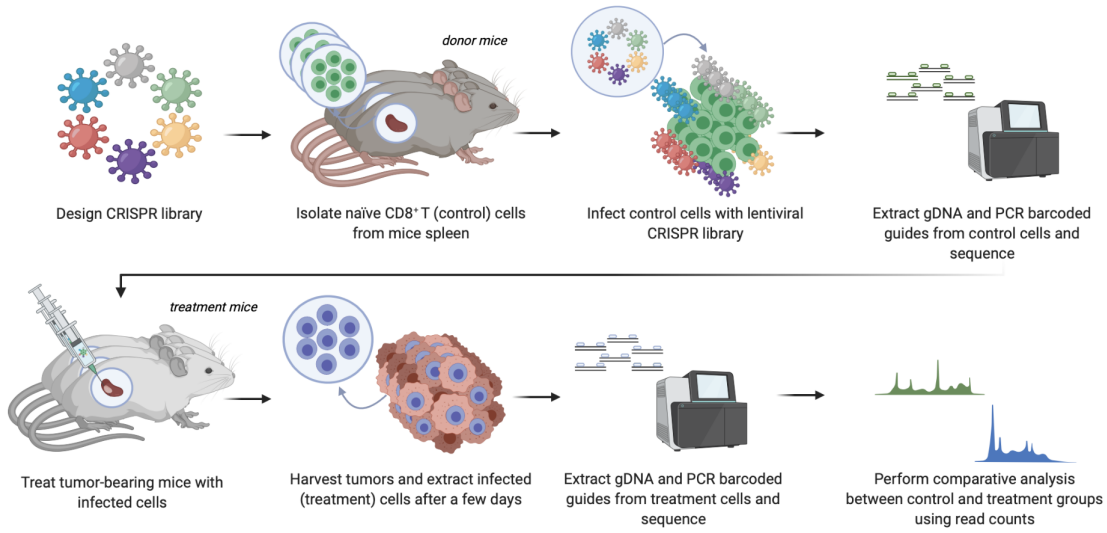


Figure 2.1: Schematic of a high-throughput *in vivo* CRISPR-Cas9 screen in OT-I;Cas9 mice (Created with BioRender.com).

2.2 Modeling high-throughput *in vivo* CRISPR screens

2.2.1 Generative model (base model)

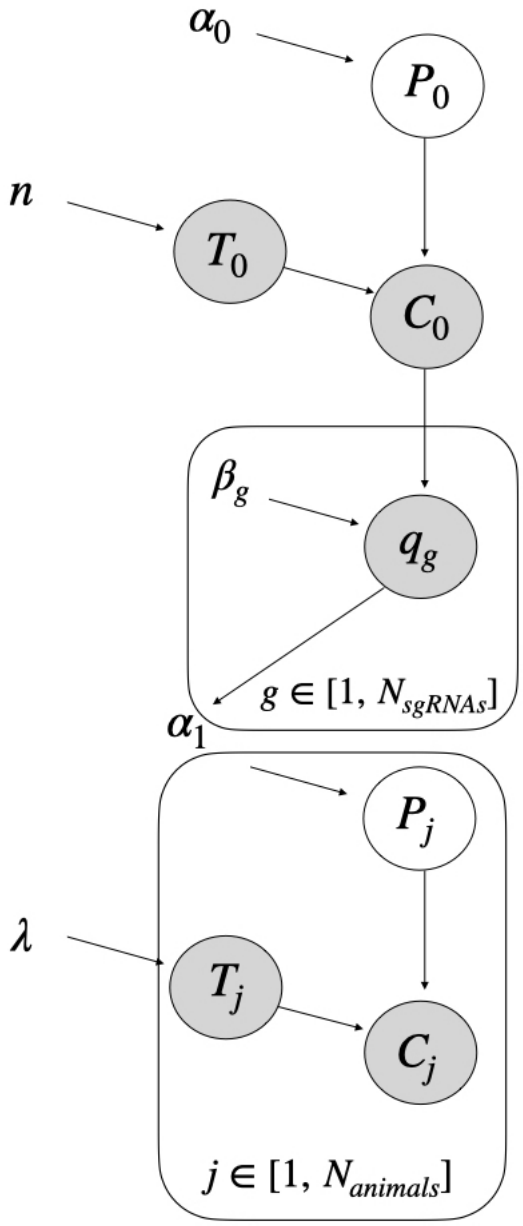
A graphical representation of the model is shown in Figure 2.2.

We introduce a generative statistical model created to simulate a high-throughput *in vivo* CRISPR screen.

We first simulate the initial library. Suppose we have N_{sgRNAs} guides, where $N_{sgRNAs} > 0$. We assume the proportion of guides in the initial library, P_0, follows a Dirichlet distribution, such that we draw a $1 \times N_{sgRNAs}$ vector from

$$P_0 \sim Dirichlet(\alpha_0), \tag{2.1}$$

where $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0k})$, a $1 \times N_{sgRNAs}$ vector containing dispersion values for each guide (guide concentration per gene), such that the distribution is symmetric; that is, $\alpha_{01} =$



$$P_0 \sim Dirichlet(\alpha_0)$$

$$T_0 \sim Poisson(n)$$

$$C_0 \sim Multinomial(P_0, T_0)$$

$$q_g = \frac{C_{0g}\beta_g}{\sum_k C_{0k}\beta_k}$$

$$a_1 = a_0 q$$

$$P_j \sim Dirichlet(\alpha_1)$$

$$T_j \sim Poisson(\lambda)$$

$$C_j \sim Multinomial(P_j, T_j)$$

Figure 2.2: Graph-based representation of generative model of a high-throughput *in vivo* CRISPR screen.

$\ldots = \alpha_{0N_{sgRNAs}}$. We define the total guide concentration to be the sum of all components in α_0.

The Dirchlet distribution generates a probability distribution that is parametrized by a vector of positive reals, α, which dictates the variance. In this model, α describes how guides are dispersed across the cells, i.e. how many cells in the pool receive a given guide, assuming each cell gets only one guide.

We assume the number of control cells extracted per mouse, T_0, follows a Poisson distribution, such that

$$T_0 \sim Poisson(n), \tag{2.2}$$

where n is the total number of control cells extracted from all mice.

Then, we draw the absolute number of cells per guide represented as a $1 \times N_{sgRNAs}$ vector, C_0, from a multinomial distribution parametrized by (2.2) and (2.1), such that

$$C_0 \sim Multinom(T_0, P_0), \tag{2.3}$$

We model the normalized population-level guide effects, q, represented as a $1 \times N_{guides}$ vector of log fold changes. Normalized population-level effect for guide $g \in [1, N_{guides}]$ is calculated by

$$q_g = \frac{C_{0g}\beta_g}{\sum_k C_{0k}\beta_k}, \tag{2.4}$$

where β_g is the effect size of guide g represented as log fold changes, C_{0g} is defined by (2.3). The summation $\sum_k$ is the sum across all guides $k \in [1, N_{sgRNAs}]$.

The effect size, β_g, is drawn from $Norm(\mu, \sigma^2)$. The estimated mean, μ, is the number of guides per gene. The estimated variance, σ^2, is centered on the guide g's targeted gene's gene-level effect size. This gene-level effect size is assumed to follow $Gamma(k, \theta)$, with

parameters k and θ are learned from the gene effects observed in Wei et al's REGNASE-1 study [WLZ19]. The direction of the effect, positive or negative, is sampled independently from q and randomly, from a point mass distribution. The probabilities in the point mass distribution are defined such that the probability of a positive effect is the number of total positive effects over the number of total effects and the probability of a negative effect is the complement.

Next, we simulate the adoptive transfer of infected cells for $N_{animals}$ mice, where $N_{animals} > 0$. We define $j \in [1, N_{animals}]$. For each mouse j, we assume the proportion of guides incorporating effects in the pooled screen, P_j follows a Dirichlet distribution, such that

$$P_j \sim Dirichlet(\alpha_1), \tag{2.5}$$

where P_j is a $1 \times k$ matrix and $\alpha_1 = \alpha_0 q$, given (2.4).

We assume the number of treatment cells that are extracted from each of the j mice, T_j, follows a Poisson distribution, such that

$$T_j \sim Poisson(\lambda), \tag{2.6}$$

where λ is the average number of treatment cells extracted per mouse.

Finally, for each mouse j, we draw the absolute number of treatment cells per guide in the pooled screen represented as a $1 \times k$ vector, C_j, from a multinomial distribution parametrized by (2.6) and (2.5), such that

$$C_j \sim Multinom(T_j, P_j), \tag{2.7}$$

Lastly, we convert the absolute cell counts, (2.3) and (2.7), into control and treatment guide read counts for guide abundance comparison. To do this, we take (2.3) to be the read counts for the control group and take (2.7) and divide the number of mice in which the screens were conducted into pools and sum the absolute cell counts of all the mice to get the

read counts for the treatment groups. Then, we perform comparative analysis between the control and treatment guide read counts.

2.2.2 Mixture model for effect size of varying guide efficiencies (alternate model)

A graphical representation of the alternate model with a mixture model for the guide effect sizes, accounting for guide efficiencies, is shown in Figure 2.3.

The population-level guide effects base model is under the assumption all guides work; that is, they operate with 100% efficiency. However, realistically, this is not the case. Many guides simply do not work due to various reasons: the guides does not target the correct DNA sequence of interest, the Cas9 doesn't snip, etc [SRJ14].

To account for the varying guide efficiencies, we developed an alternate model for the (normalized) population-level guide effect sizes, q. In this model, population effect size for guide g is calculated using the mixture model

$$q_g = \frac{C_{0g1}\beta_g + C_{0g0}}{\sum_{i=1}^{k} C_{0i1}\beta_i + C_{0i0}}, \tag{2.8}$$

where C_{0g1}, the number of cells with Cas9/guide g complexes that successfully cleave the targeted gene, is drawn from $Binom(C_0, e)$. C_0 is defined as the absolute number of control cells per guide (2.3) and e is the overall guide efficiency. C_{0g0}, the number of cells with Cas9/guide g complexes that fail to cleave the targeted gene, is simply calculated as $C_{0g} - C_{0g1}$. All other parameters are the identical and follow the same distributions as those in the base model.

2.3 Simulating experiments

Both generative models were implemented in R code [R C20]. Simulations were ran using combinations of various parameters and corresponding analyses were performed using the

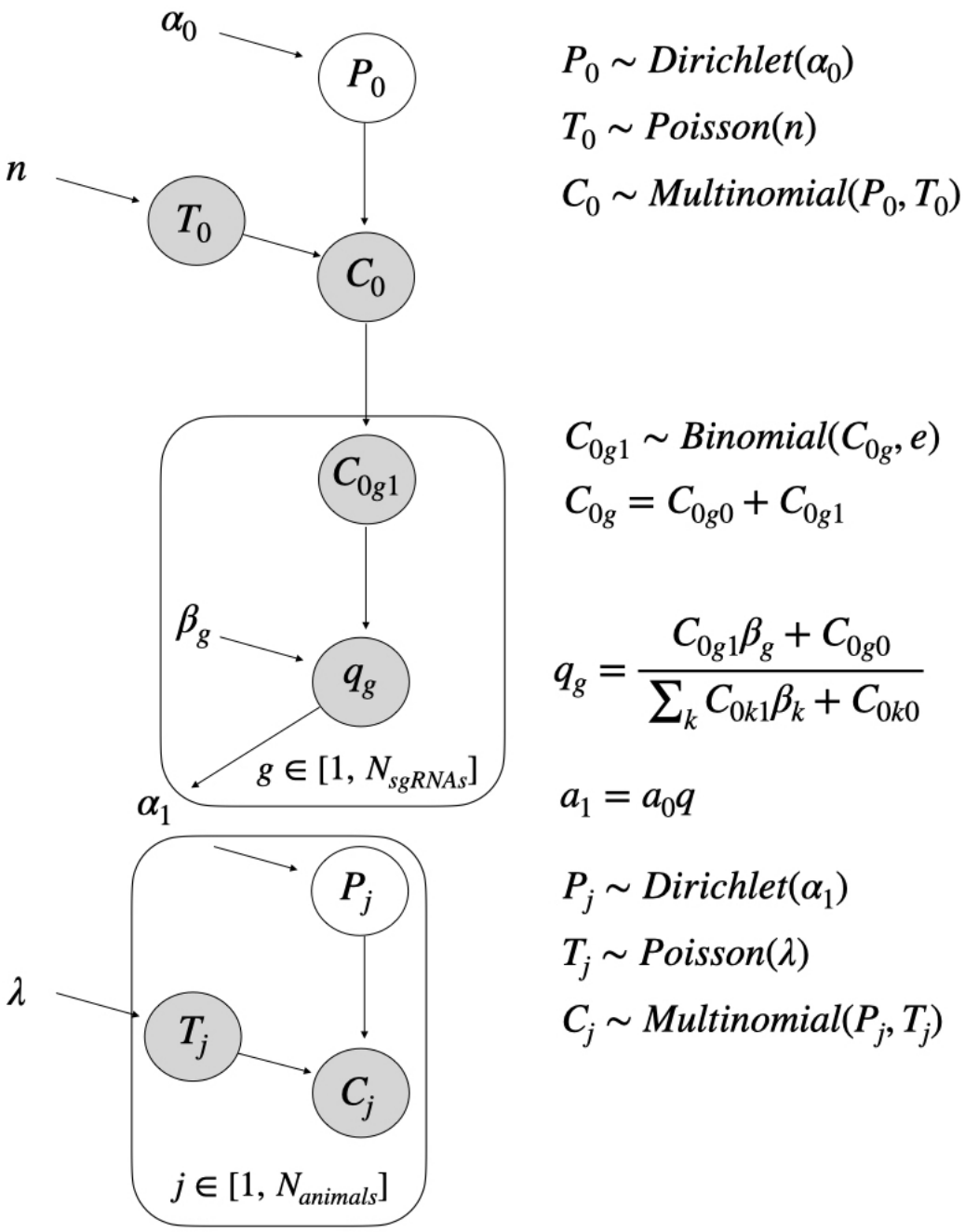


Figure 2.3: Graph-based representation of generative model with mixture model for varying guide efficiencies of a high-throughput *in vivo* CRISPR screen.

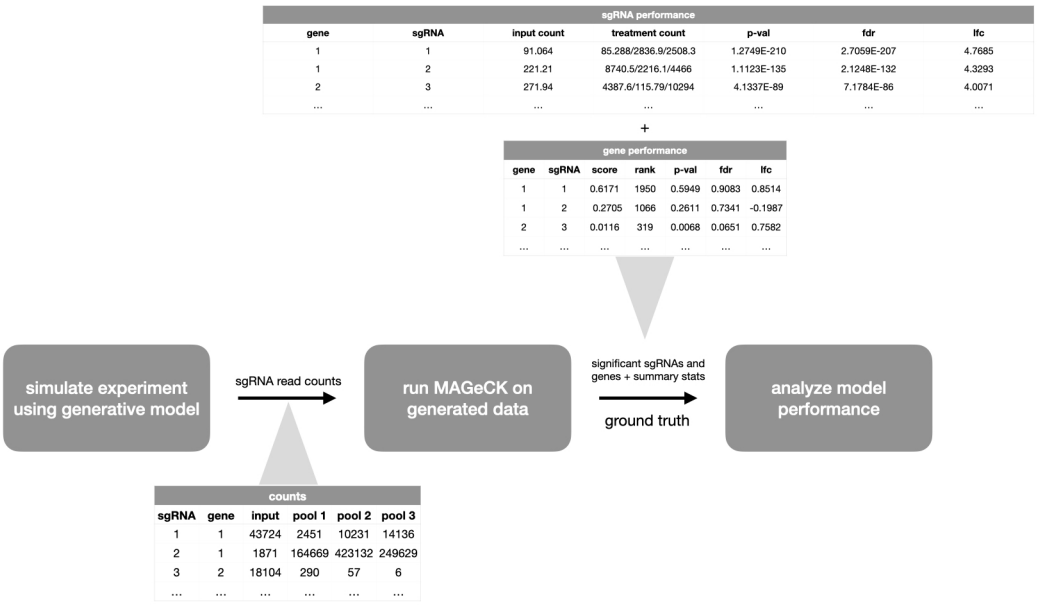


Figure 2.4: Analysis workflow with sample data sets. First, the experiment is simulated with given unobserved parameters. Then, using the generated guide read counts, significant guides and genes are identified, guide read counts are normalized, and summary statistics are calculated. Lastly, sensitivity analysis of the simulation is performed using the MAGeCK output.

snakemake workflow management system [MJL21].

2.4 Analyzing model performance

2.4.1 MAGeCK

To analyze the performance of our model, we use the state-of-the-art tool, Model-based Analysis of Genome-wide CRISPR/Cas9 Knockout (MAGeCK), a method that identifies essential genes from CRISPR knockout screens [LXX14]. We input the control and treatment read counts generated from our model into MAGeCK, which identifies positively and

negatively selected guides and genes and generates normalized read counts with summary statistics. A diagram of the analysis workflow is shown in Figure 2.4.

We distinguish summary statistics reported by MAGeCK from those we calculated in our analyses by prefacing the summary statistics reported by MAGeCK with the word 'estimated' and read counts reported by MAGeCK from those we generated by prefacing those reported by MAGeCK with the word 'normalized'.

2.4.2 Sensitivity plots

In order to identify simulations with optimal parameters, we performed power analysis by looking at sensitivity against varying parameter values. Simulations with higher sensitivity have more statistical power, and therefore, produce more informative data.

Sensitivity is defined as $\frac{TP}{\text{total } TP}$, where TP is the number of true positives.

Since we had simulated the data, the ground truth was known and we were able to calculate the sensitivity. To do so, we ordered the genes in ascending order by the estimated FDR. Sensitivity was then calculated as the cumulative number of genes in which significant effects were simulated at that given point divided by the total number of significant effects simulated in the entire screen.

2.4.2.1 Sensitivity versus false discovery rate curves

We also utilized sensitivity versus false discovery rate (FDR) curves to illustrate how well MAGeCK identified genes that our model had simulated effects on. In a sensitivity versus FDR curve, the x-axis is the true FDR and the y-axis is the sensitivity. The true FDR is calculated as $\frac{FP}{\text{total } TN}$, where FP is the number of false positives and TN is the number of true negatives. Ideally, in a sensitivity versus FDR plot, sensitivity is maximized while FDR is minimized.

To calculate the FDR, we ordered the genes in ascending order by the estimated FDR.

Then, FDR was calculated as the cumulative number of genes in which significant effects were not simulated at that given point divided by the number genes MAGeCK has classified as essential at that point.

2.4.3 MA plots

We used MA (log fold change between control and treatment groups versus average) plots to perform differential guide analysis to identify guides that are significantly enriched or depleted.

To calculate the average of the normalized read counts, we took the mean of the normalized read counts across the control and treatment groups. The log fold changes were taken from the summary statistics generated by MAGeCK.

2.4.4 Dispersion plots

We looked at dispersion plots to investigate the dispersion of simulated guide concentrations and to ensure that the simulations replicated sufficient biological noise, but not such that the output is not what our model aimed to generate. We utilized both variance versus mean (index of dispersion) and dispersion versus mean plots.

For the index of dispersion plot, the variance and mean of the normalized read counts are calculated across all pools and plotted against one another.

For the dispersion versus mean plots, the dispersion parameter ϕ and the mean of the normalized read counts are plotted against one another. To calculate ϕ, we assumed a negative binomial distribution on the normalized guide read counts. Then, we solved for ϕ, which is calculated as $\frac{\sigma^2-\mu}{\mu^2}$, where σ^2 is the variance of the guide read counts and μ is the mean of the guide read counts.

CHAPTER 3

Results

3.1 Comparison of model-generated data to experimental data

parameter	value
number of initial control cells	450000000
average number of treatment cells extracted per mouse	500000
number of genes targeted	3017
number of guides per gene	6
number of non-targeting control guides	1000
guide efficiency	0.4
proportion of guides that administer effects	0.1
proportion of effects that are positive	0.1
number of mice	120
number of pools in pooling scheme	3
total guide concentration	2000

Table 3.1: Parameters used in generative model to simulate the REGNASE-1 study.

To compare the data generated from our alternate model and real experimental data, we used similar parameters to those indicated in Wei, et al's REGNASE-1 study [WLZ19]. Parameters used in the simulation are listed in Table 3.1.

Guide efficiency was selected based on results from experiments using green fluorescent

protein and microscopy that suggest that guide efficiency in *in vivo* CRISPR screens often vary from 0.3 to 0.6 [SRJ14].

Total guide concentration was selected based on the appearance of simulating reasonable dispersion across the normalized guide read counts closely to the REGNASE-1 data.

3.1.1 Gene effect sizes

To compare the data generated from our model and real experimental data, we looked at effect sizes at both the gene-level and the guide-level of the data generated by the model and that of the REGNASE-1 study.

3.1.1.1 Gene-level

We first looked at the model's learned distribution of effect sizes at the gene-level, via maximum likelihood estimation assuming a gamma distribution. This was done by graphing a histogram of the distribution of the effect sizes of each gene in the REGNASE-1 study, determined by MAGeCK and plotting both a density curve and the learned distribution. As a comparison, we also graphed a histogram of the distribution of the generated gene effect sizes, as seen in Figure 3.1.

The overall distributions of gene-level effect sizes of the REGNASE-1 study and simulation are similar. Both have the same shape and peak at about the same density. However, the simulation's gene-level effect sizes have smaller variance that of the REGNASE-1 study. On the other hand, the gene-level effect sizes in REGNASE-1 is skewed. This is due to the fact that the REGNASE-1 study found genes with very large and significant effect sizes, a quite rare occurrence. As a result, the smaller variance on the simulated data is due to the truncation of the gene-level distribution in our model to reflect more typical experiments.

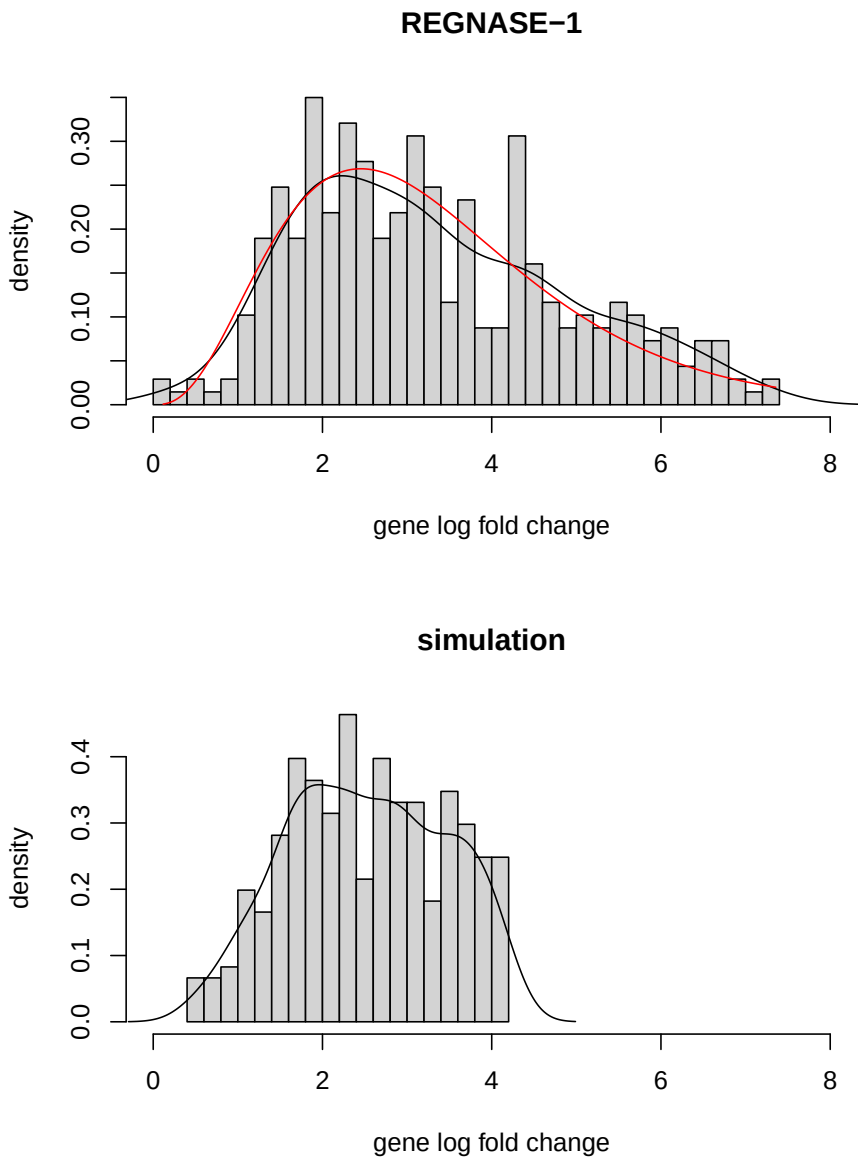


Figure 3.1: Histograms of the density of gene effect sizes of the REGNASE-1 study (above) and the generative model's simulation (below). The x-axis shows the gene log fold change and the y-axis shows the density. The black curves are the fitted densities of the distributions. The red curve in the REGNASE-1 study histogram is the learned gamma distribution using maximum likelihood estimation.

3.1.1.2 Guide-level

In our model, gene-level effect sizes influence guide-level effect sizes and the effect of treatment in a CRISPR screen is quantified by relative guide abundance. Therefore, we performed differential guide analysis on the normalized counts of each data set, as generated by MAGeCK.

We began with looking at MA plots, which are shown in Figure 3.2. We defined significant guides to be those with an estimated false discovery rate less than 10%.

The qualitative nature of the two plots are similar; both have the same approximate shape and effect sizes. However, the REGNASE-1 data has more significant guides compared to that of the simulated data. In particular, there are many more guides that are negatively selected. Additionally, the variance of the mean of simulation normalized guide read counts is much smaller in the REGNASE-1 data compared to the simulated data.

3.1.2 Dispersion

We also looked at the dispersion of the normalized guide read counts using two different plots: the index of dispersion and the dispersion versus mean plot.

As seen in Figure 3.3, overdispersion is much more prevalent in the simulation than that in REGNASE-1. We can see in both variation of dispersion plots: In the index of dispersion plot, there larger distance between the overall fitted curve and line of slope one for the simulation. In the dispersion versus mean plot, we can see that there is more variation dispersion parameters for the simulation.

3.1.3 Additional comparative analysis

Additional differential guide analysis was performed using the DESeq2 library [LHA14]. Results are reported in the Supplement (Chapter 5).

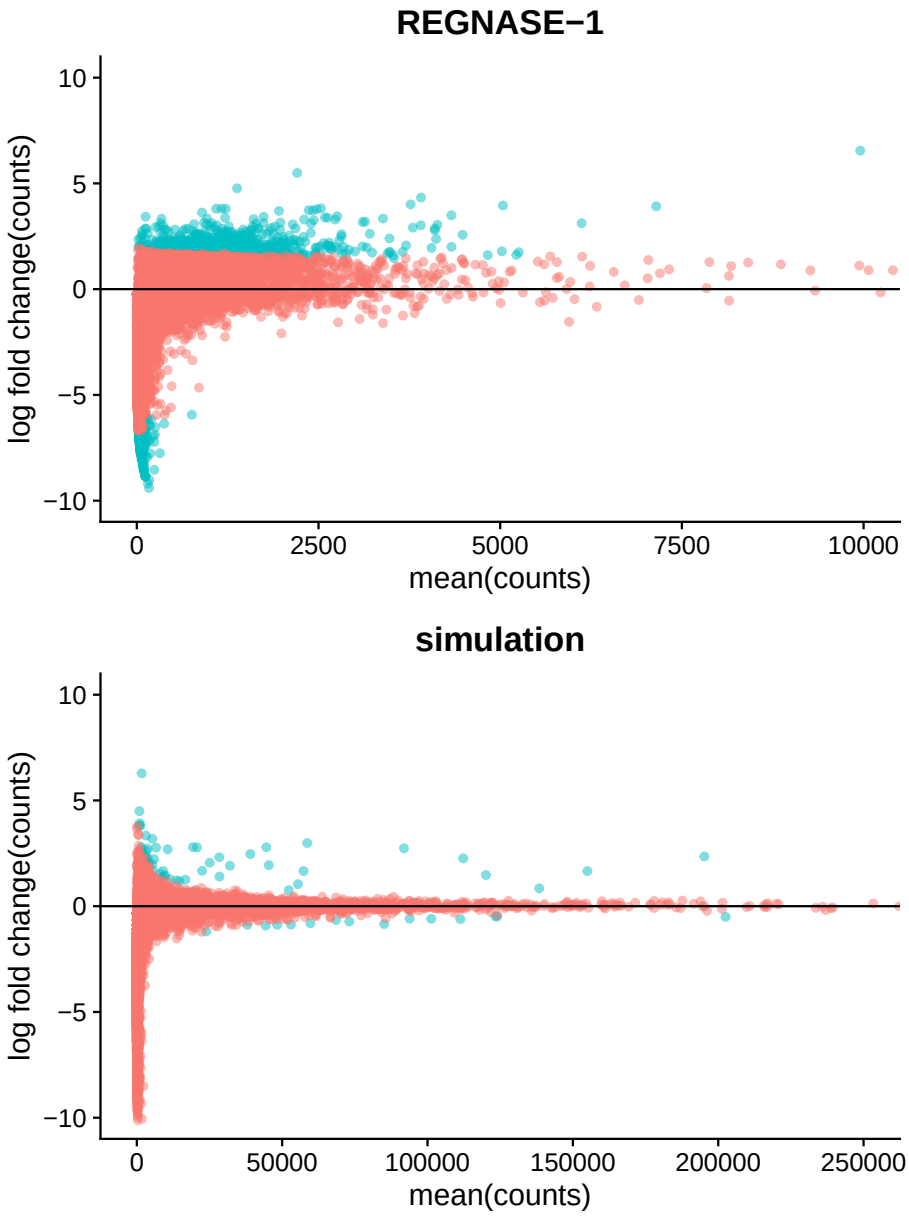


Figure 3.2: MA plots of guide read counts from the REGNASE-1 study (above) and the generative model's simulation (below). The x-axis shows the mean of the guide read counts and the y-axis shows the log fold change of the read counts between the control and treatment groups. Each point is an guide; blue points indicate significant guides (FDR < 0.1).

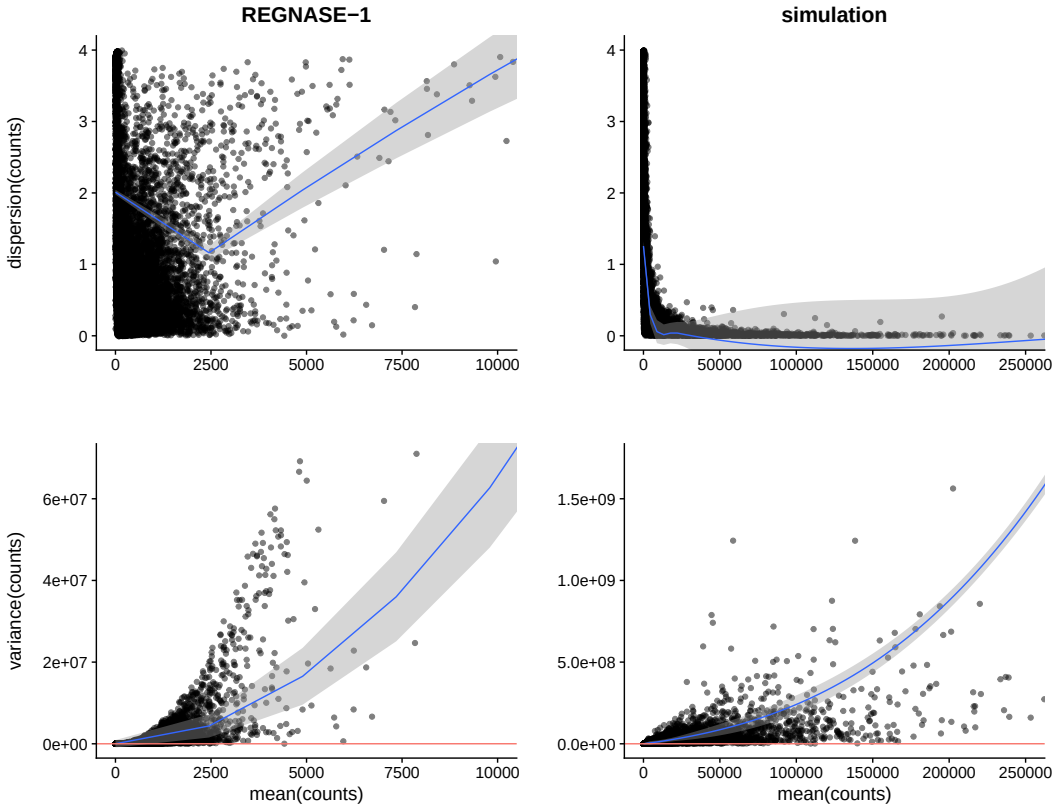


Figure 3.3: Dispersion plots of normalized guide read counts from the REGNASE-1 study (left) and the generative model's simulation (right) with blue fitted curves. The plots above are dispersion versus mean plots. The x-axis shows the mean of the read counts and the y-axis shows the dispersion parameter of the read counts. The plots below represent the index of dispersion. The x-axis shows the mean of the read counts and the y-axis shows the variance of the read counts. The red line has slope 1, indicating an index of dispersion of 1.

3.2 Base model performance

parameter	value
number of initial control cells	35000000
average number of treatment cells extracted per mouse	10000
number of genes targeted	500
number of guides per gene	3
number of non-targeting control guides	1000
proportion of guides that administer effects	0.1
proportion of effects that are positive	0.1
number of mice	60
number of pools in pooling scheme	3, 6, 9
total guide concentration	2000

Table 3.2: Parameters used the base model performance analysis.

We first introduce analysis of the base model to give some intuitive information on high-throughput *in vivo* CRISPR screens. Parameters used in the model to simulate experiments in the analysis are indicated Table 3.2, unless stated otherwise.

3.2.1 Effect of varying the number of treatment cells extracted from each mouse

We investigated the statistical power as a function of the average number of treatment cells extracted from each mouse. The number of cells simulated were of values 5000, 10000, 25000, 40000, and 50000.

We can see in Figure 3.4 that as the number of cells increases, the sensitivity eventually plateaus; this is especially evident when increasing from 40000 and 50000 cells. Therefore, in most cases, there are diminishing returns when one extracts more treatment cells per mouse. Additionally, as the number of pools increases, the sensitivity decreases. In the case of 3 and

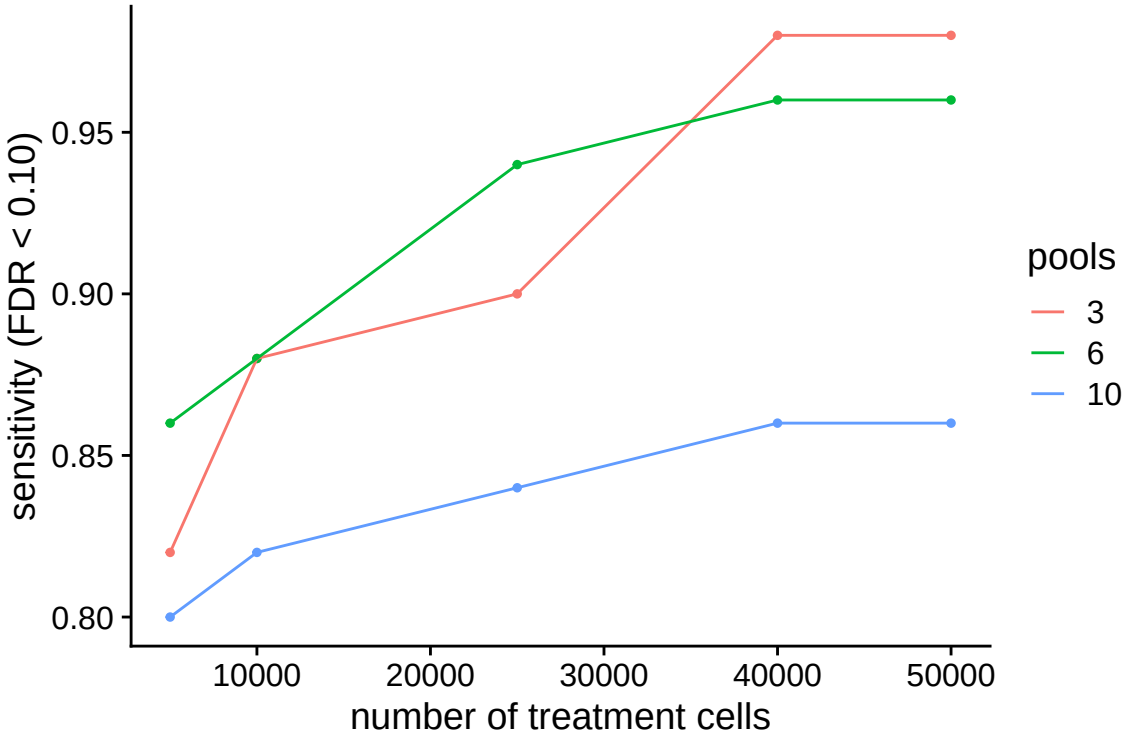


Figure 3.4: The effect of the average number of treatment cells extracted from each mouse on the base model's performance, stratified by the number of pools in the pooling scheme. The x-axis shows the average number of treatment cells extracted from each mouse in a given simulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$.

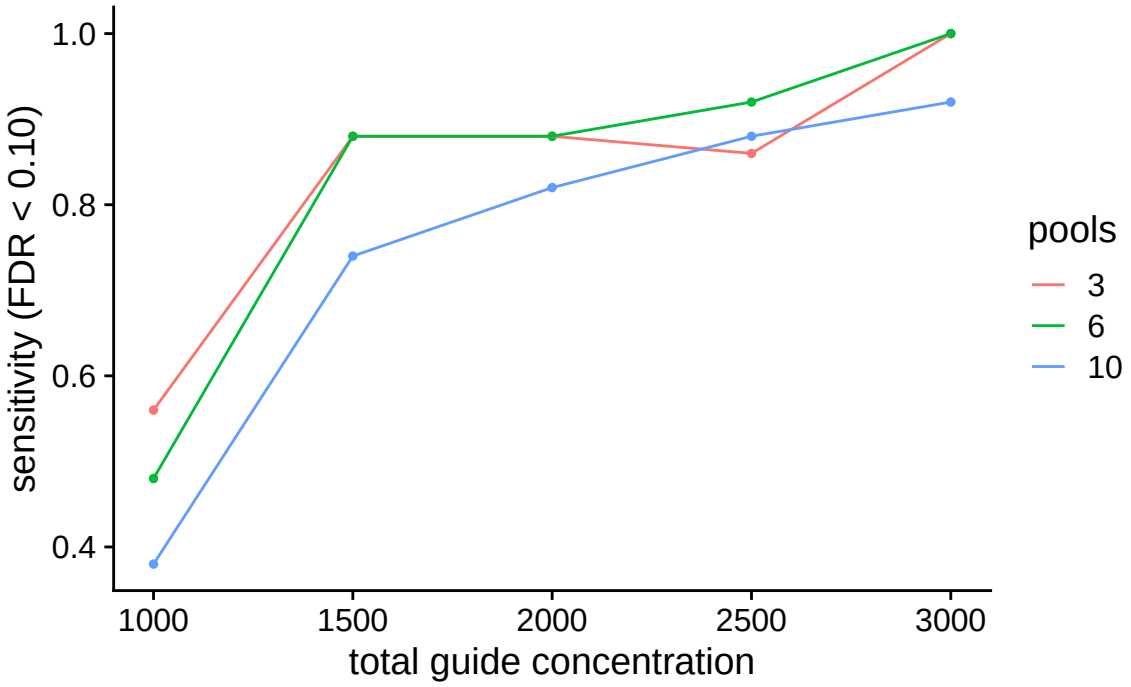


Figure 3.5: The effect of guide concentration on the base model's performance, stratified by the number of pools in the pooling scheme. The x-axis shows the guide concentration in a given simulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$.

6 pools in this particular simulation, the performance of 3 pools makes larger returns as the number of cells increases.

3.2.2 Effect of varying the total guide concentration

In addition, we explored how the total guide concentration effects the statistical power of the data generated by the simulation across different pooling schemes. The total guide concentration values simulated were 1000, 1500, 2000, 2500, and 3000.

Note that in Figure 3.5, that the statistical power gives a very large return when increasing from 1000 to 1500 total guide concentration. However, the rate at which statistical power increases rapidly declines as the total guide concentration increases, and sometimes even

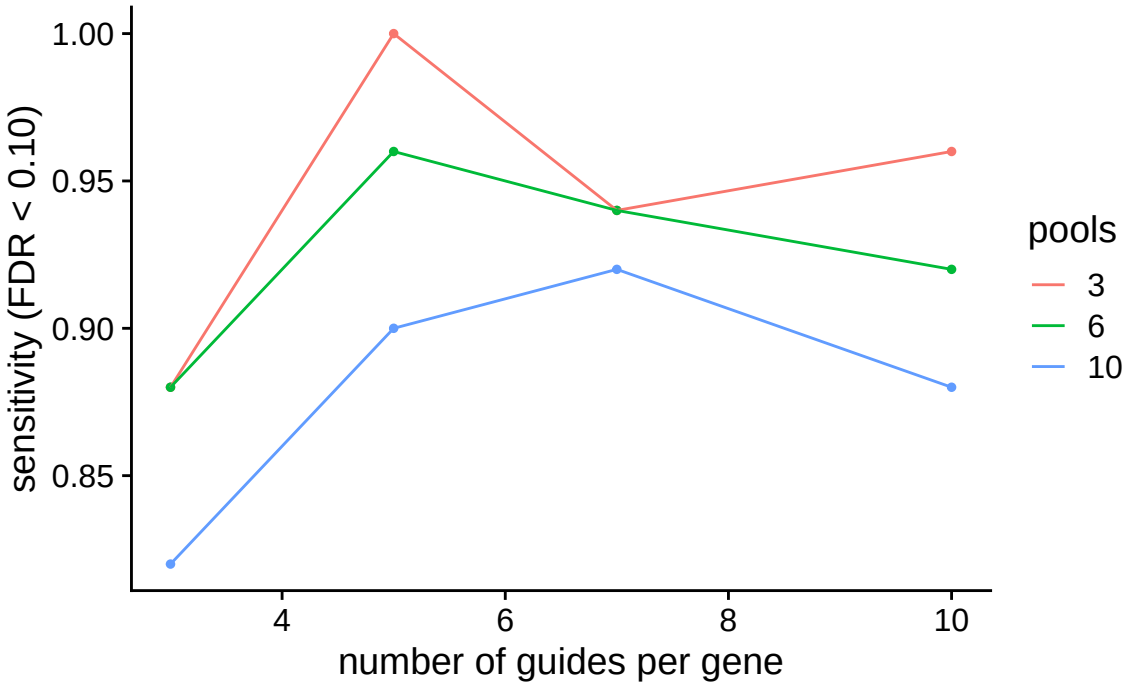


Figure 3.6: The effect of the number of guides targeting a gene on the base model's performance, stratified by the number of pools in the pooling scheme. The x-axis shows the number of guides per gene in a given simulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$.

gives a negative return. However, we suspect overregularization in MAGeCK due to the slow convergence to 1. Statistical power is similar between 3 pools and 6 pools in the pooling scheme and is lowest in 10 pools.

3.2.3 Effect of varying the number of guides targeting each gene

We also observed the model's statistical power as a function of the the number of guides targeting a given gene across different pooling schemes. The number of guides per gene simulated were 3, 5, 7, and 10.

We can see in Figure 3.6, across the number of guides per gene, there are definite optimal

parameter	value
number of initial control cells	35000000
average number of treatment cells extracted per mouse	10000
number of genes targeted	500
number of guides per gene	3
number of non-targeting control guides	1000
guide efficiency	0, 0.2, 0.4, 0.6, 0.8, 1
proportion of guides that administer effects	0.1
proportion of effects that are positive	0.1
number of mice	60
number of pools in pooling scheme	3, 6, 10
total guide concentration	2000

Table 3.3: Parameters used the alternate model analysis.

choices in each pooling schemes. With 3 and 6 pools, the largest sensitivity is produced when using 5 guides to target a gene. In 10 pools, it is produced when using 7 guides to target a gene. Overall, the statistical power is inversely related to the number of pools; as the number of pools decrease, the sensitivity increases.

3.3 Alternate population-level guide effects model performance

Next, we investigated the model performance in our alternate model. We repeat our analysis, but focus on lower guide efficiencies which reflect more realistic experimental conditions. Parameters used in the model to simulate experiments in the analysis are indicated in Table 3.3, unless stated otherwise. Note that the model at a guide efficiency of 1 is equivalent to the base model, and serves as a control. A guide efficiency of 0 serves as a base line.

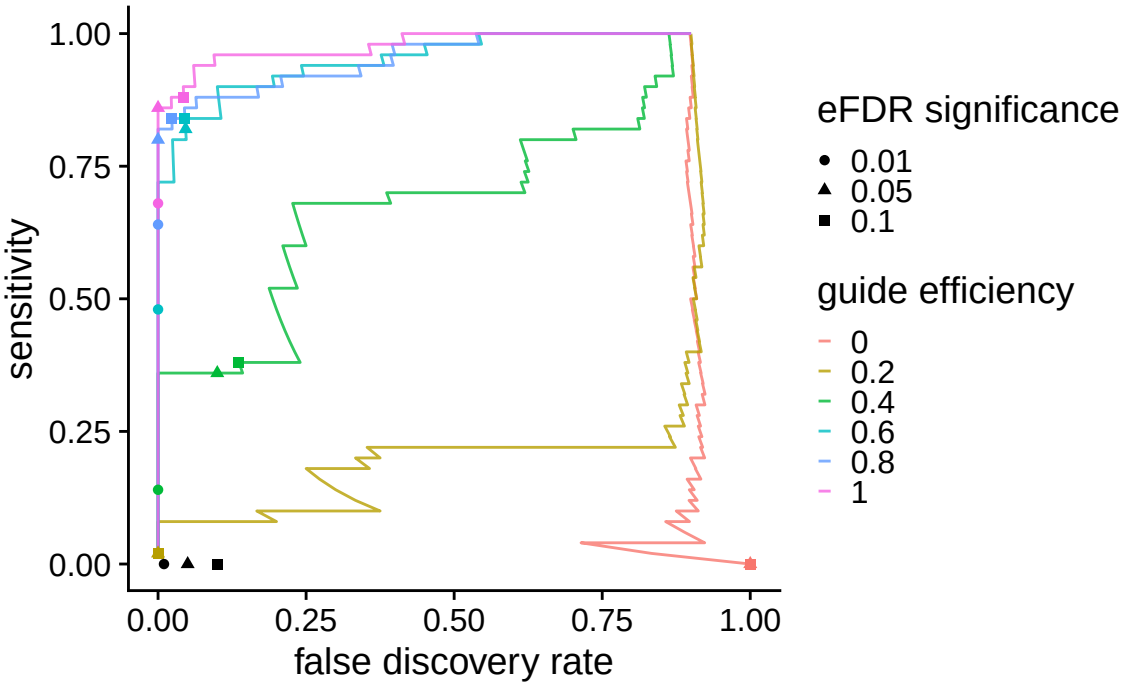


Figure 3.7: Statistical power stratified by guide efficiencies. The x-axis shows the false discovery rate and the y-axis shows the sensitivity. Marked on the bottom of the graph are the shapes of the significance levels indicated in the legend at the location of the true false discovery rate.

3.3.1 Guide efficiency validation

To confirm that the alternate model is behaving well, we observed how statistical power changes across every gene for each level of guide efficiency. The number of pools in the pooling scheme was fixed at 3.

We can see in Figure 3.7, simulations with higher guide efficiencies maximizing sensitivity while minimizing FDR. As expected, this indicates that statistical power increased with guide efficiency.

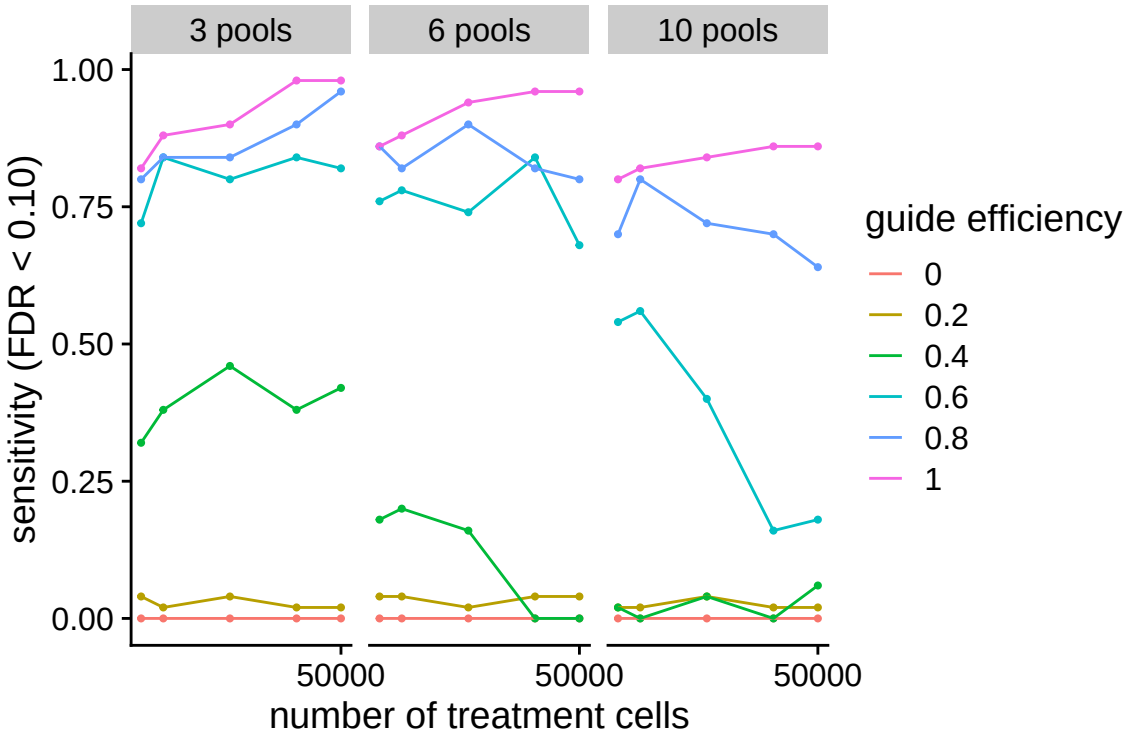


Figure 3.8: The effect of the average number of treatment cells extracted from each mouse on statistical power, stratified by guide efficiencies. The x-axis shows the average number of treatment cells extracted per mouse in a given simulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$. From left to right, the number of pools in the pooling scheme increases from 3 to 6 to 10.

3.3.2 Effect of varying the average number of treatment cells extracted from each mouse

We examined how statistical power is affected by the average number of treatment cells extracted per mouse across different guide efficiencies and pooling schemes. The number of cells simulated were of values 5000, 10000, 25000, 40000, and 50000.

In Figure 3.8, we can see that as the number of cells increases, as guide efficiencies lower, the statistical power actually fluctuates and often decreases after some threshold. Generally,

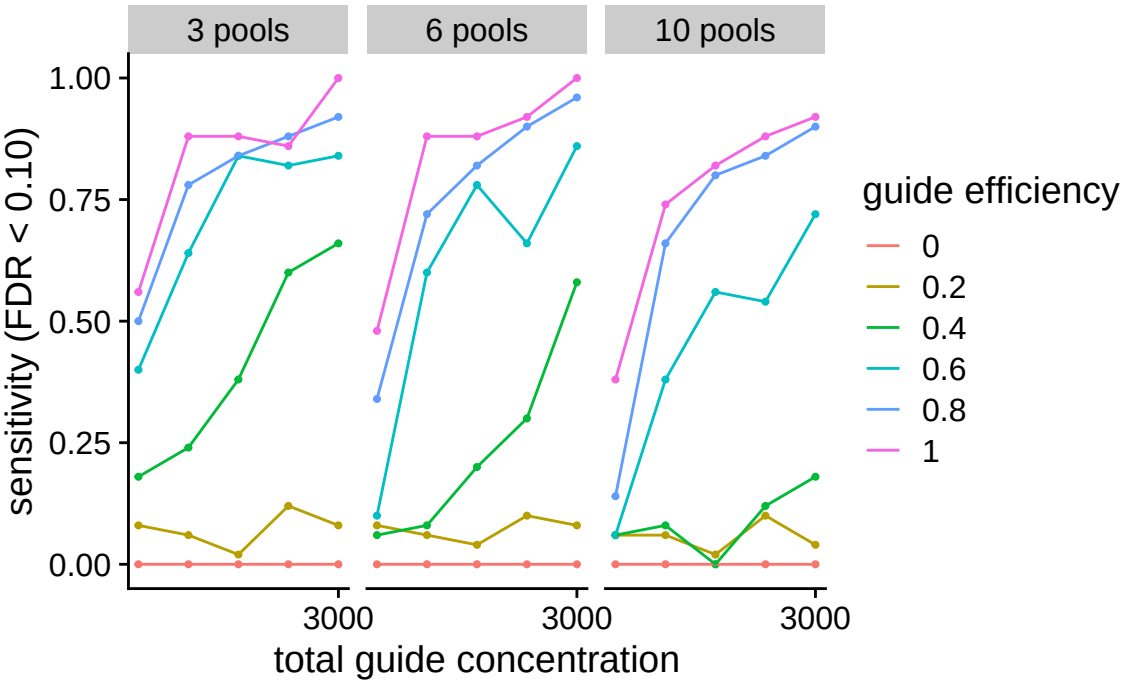


Figure 3.9: The effect of the total guide concentration on statistical power, stratified by guide efficiencies. The x-axis shows the total guide concentration in a given simulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$. From left to right, the number of pools in the pooling scheme increases from 3 to 6 to 10.

sensitivity is higher when there is a smaller number of pools in the pooling scheme.

3.3.3 Effect of varying the guide concentration

Then, we observed how guide concentration influences statistical power across different guide efficiencies and pooling schemes. The total guide concentration values simulated were 1000, 1500, 2000, 2500, and 3000.

Figure 3.9 shows that across all guide efficiencies, as the total guide concentration grows larger, the sensitivity will increase until it approaches 1. Also, the statistical power is similar in pooling schemes of 3 and 6 pools, but is smaller in 10 pools.

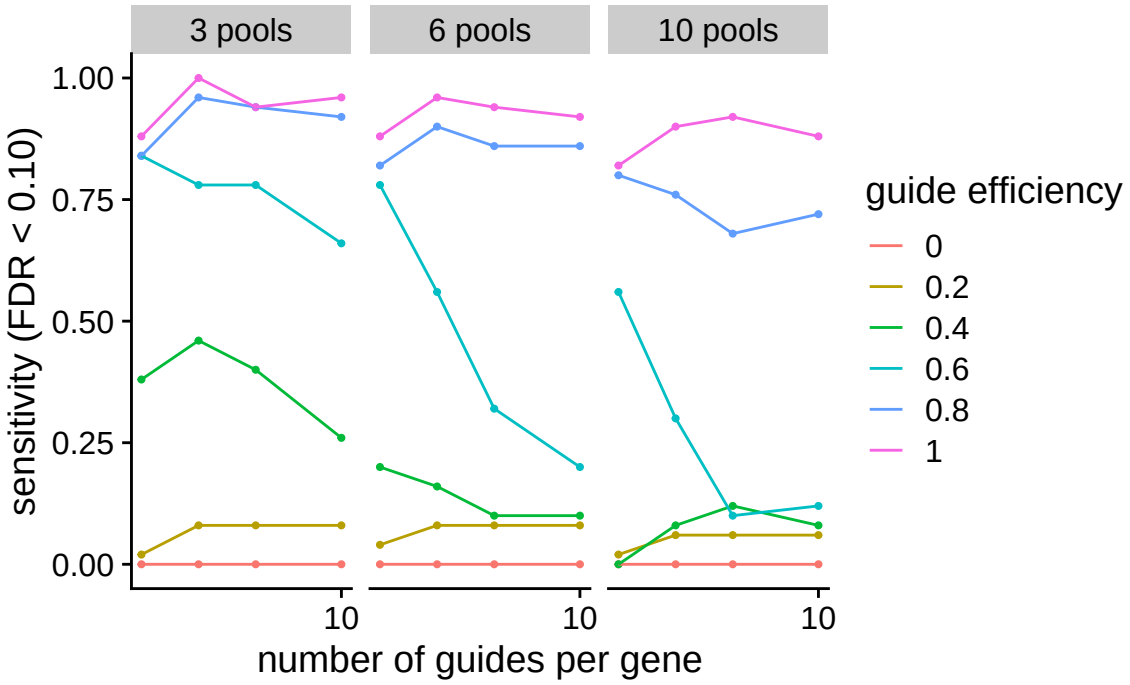


Figure 3.10: The effect of the number of guides per gene on statistical power, stratified by guide efficiencies. The x-axis shows the number of guides per gene in a given simmulation and the y-axis shows the sensitivity, fixed at FDR $< 10\%$. From left to right, the number of pools in the pooling scheme increases from 3 to 6 to 10.

3.3.4 Effect of varying the number of guides targeting each gene

We looked at how the number of guides per gene affects statistical power across different guide efficiencies and pooling schemes. The number of guides per gene simulated were 3, 5, 7, and 10.

As seen in Figure 3.10, there is a point at which sensitivity peaks in all guide efficiencies and across all pooling schemes. For most simulations, 5 guides per gene produces the highest statistical power. However, at low efficiencies the rate at which sensitivity decreases is overall much higher that that of higher efficiencies.

CHAPTER 4

Discussion

Our statistical model is able to simulate high-throughput *in vivo* CRISPR screens relatively well. Due to the the learned parameters from the REGNASE-1 study's gene-level effect sizes, the generated data was able to emulate similar gene-level effect sizes under the same conditions. Differential guide analysis also uncovered similar structures in positively and negatively selected guides as determined by MAGeCK. In addition, although dispersion of the simulated normalized guide read counts was much higher than that of the REGNASE-1 study, both figures beared similarities in their shapes and trajectories. We believe the discrepancy in the mean of read counts between the simulation and REGNASE-1 study may be attributed to the normalization techniques used in MAGeCK. Due to this, we tested additional methods that normalize read counts as part of our analysis to ensure that these similarities are maintained; results using DESeq2 are reported in the Supplement. Accordingly, we argue that these simulations provide valuable information to guide the design of high-throughput CRISPR screen experiments.

We also demonstrated that we can simulate experiments under different combinations of parameter inputs used in a typical protocol of an *in vivo* CRISPR screen. We also show that our alternate model is consistent with the expectation; that is, statistical power will increase with guide efficiency. With the generated data from our simulations, we performed statistical power analysis to gain insight on optimizing experimental design and general guiding practices.

From the simulations performed, by varying the average number of treatment cells ex-

tracted from each mouse, as the number of treatment cells increase, there was a threshold in which statistical power either plateaus or even decreases. That is, at some point, increasing the number of cells yields diminishing returns. We believe that this is a result of saturation; that is, as the number of cells approaches infinity, that the statistical power will eventually converge to a fixed value, determined by the total guide concentration.

The case is also similar with the total guide concentration, as when the total guide concentration increases, the rate at which statistical power increases becomes smaller. This is consistent with the fact that with a larger total guide concentration, there is less of variance across the guides read counts. At more realistic guide efficiencies, performance drops significantly. We believe that this is due to larger amounts of dropout at lower efficiencies.

However, we observed that when the number of guides per gene are varied, there is a clear and distinct point at which sensitivity is maximized. Beyond this point, sensitivity either plateaus or decreases. At more realistic guide efficiencies, the latter behavior is often exhibited. Therefore, we presume that the number of guides per gene is one of the most crucial parameters that affect the output of the screen. It is counter-intuitive as if there were infinite cells extracted, the more guides per gene should give more power. But since the number of cells extracted and the guide concentration is fixed, increasing the number of guides per genes results in fewer reads corresponding to each guide, resulting in weak estimates for each guide abundance due to the lack of coverage.

In all simulations, in most cases, statistical power is inversely related to the number of pools used in the pooling scheme. It is important to note that this is only the case when there is a restriction on the number of cells extracted, a scenario consistent with *in vivo* screens, as the number of treatment cells that can be extracted is limited by the number of cells one can get from an animal. This behavior stems from the bias-variance trade-off. As the number of pools in the pooling scheme increases, the number of mice in each pool decreases. This smaller region leads to higher variance in the statistical power amongst the mice, as a fluctuation in just one mouse can generate results in the pool that are not representative of

the biological variance nor the experiment at hand.

Also, in all simulations, there is a large amount of variance in performance across the guide efficiencies. Our findings suggest the importance of assessing guide efficiency of the CRISPR library prior to performing the screen. The performance in higher guide efficiencies such as 0.8 are high regardless of how parameters may vary. However, at more realistic guide efficiencies such as 0.4 there is significant effects on the performance by varying parameters, which is where our model serves the most helpful. At lower guide efficiencies like 0.2, we would suggest that the CRISPR screen should not be performed as regardless of whether or not the parameters are optimally selected, the return on power is still not enough.

While our model has been able to provide insight on better designing high-throughput *in vivo* CRISPR screens, many assumptions and methods used to make our model that may not generalize to all *in vivo* CRISPR screens. For instance, there are many ways to perform screens, beyond the protocol that we had emulated. We chose to select the specific protocol we had emulated due to the fact that it has been done before in multiple studies. However, in principle, our model can extended and multiple adaptations can be made easily. A few examples include: if there were no pooling schemes, setting the number of pools to 1 would suffice and if the control cells were never extracted from the animal prior to infection and rather cells were directly infected in the animals, setting the number of control cells equal to the number of treatment cells extracted per mouse would suffice. Also, we do not incorporate multiplicity of infection as a parameter in our model as our model assumes only one guide is expressed in a given cell. This would require a model at the cell-level with the ability to infer cell-level effects. Additionally, the gene effect sizes are learned from the REGNASE-1 study using maximum likelihood estimation; however, we must assume distributions of all random variables due to the variability across experiments. Lastly, we hold our suspicions on different methods used in MAGeCK in their analysis such as their read normalization and other summary statistic calculations.

In our future work, we hope to implement an model that can be used to infer different

variables in the design. Using sampling techniques such as Markov chain Monte Carlo, we can infer the posterior distributions of various random variables such as the guide dispersion. We believe that learning the distribution of these variables rather than using maximum likelihood estimation to estimate fixed parameters of distributions will significantly improve our model, as it will take into account the fact that each experiment will be subject to different effects. Moreover, we hope to implement our model as a tool with an easy-to-use interface, where experimentalists can input their own parameters to optimally design their experiments with the resources they have.

CHAPTER 5

Supplement

5.1 REGNASE-1 and simulation differential guide analysis comparison using DESeq2

For additional differential guide analysis, we use DESeq2 [LHA14]. We assume the read counts generated in our model are from paired-end reads to remain consistent with the analysis performed in the REGNASE-1 study [WLZ19]. Prior to the analysis, the log fold changes were shrunk using the native DESeq2 apeglm log fold change shrinkage estimator to prepare the data for visualization [ZIL19].

We first looked at the generated MA-plots from the REGNASE-1 study's data and the simulation's data. As seen in Figure 5.1, most guides in both of the plots follow similar qualitative patterns. First, the mass of the guides lie against log fold change off. Secondly, at the apex of the mass, the guides follow a striped pattern, creating relatively parallel lines. But, we do note that these stripes may be due to normalization issues. However, the REGNASE-1 data's MA-plot shows there are more positively-selected significant guides identified than that of the simulation data. Also, the scale of mean of normalized counts and log fold changes in the simulation are slightly larger than that of REGNASE-1. These findings from our analysis using DESeq2 are consistent with our analysis reported in the main text.

Next, we looked at the dispersion and fitted estimates of the guide counts. Figure 5.2 shows that there is much more variation in dispersion across the guide read counts of the

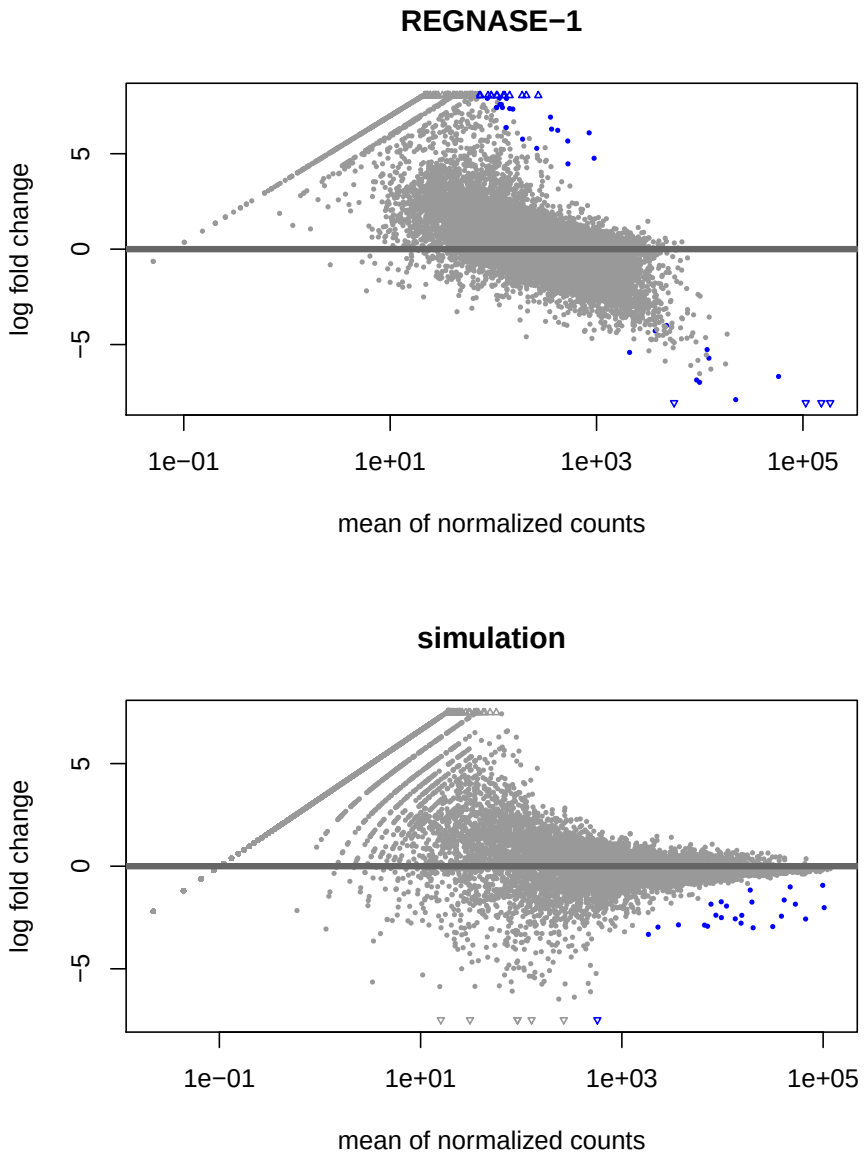


Figure 5.1: DESeq2-generated MA plots of guide read counts from the REGNASE-1 study (above) and the generative model's simulation (below). The x-axis shows the mean of the normalized guide read counts and the y-axis shows the log fold change of the read counts between the control and treatment groups. Each point is an guide; points in a triangular shape lie outside the plotted window, blue points indicate significant guides ($p < 0.1$).

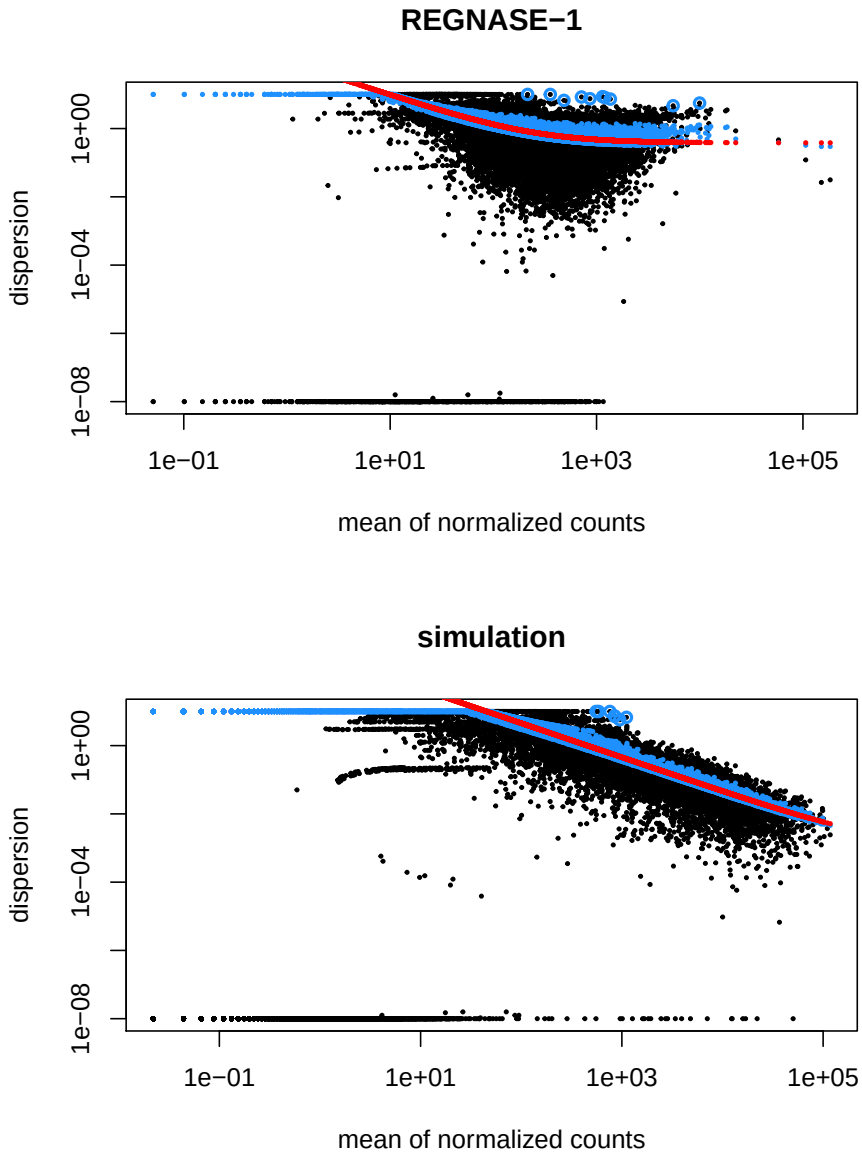


Figure 5.2: DESeq2-generated dispersion plots of guide read counts from REGNASE-1 study (above) and the generative model's simulation (below). The x-axis shows the mean of normalized guide read counts and the y-axis shows the dispersion of the read counts. Each point is a guide; the red line is an estimated fitted curve, blue points are shrunk towards the fitted value, black points are outliers.

REGNASE-1 study, relative to that of the simulation's. However, there is more variation in the mean of guide read counts in the simulation's data than in that of the REGNASE-1 study's. Also, there is more curvature in the fitted curve of the REGNASE-1's dispersion than that of the simulation's. But, overall, the mass of dispersion plots are similar. These findings are also consistent with those reported in the results of the main text.

REFERENCES

[BRH20] Carl G. de Boer, John P. Ray, Nir Hacohen, and Aviv Regev. "MAUDE: inferring expression changes in sorting-based CRISPR screens." *Genome Biology*, **21**(1):134, June 2020.

[CPZ15] David Benjamin Turitz Cox, Randall Jeffrey Platt, and Feng Zhang. "Therapeutic genome editing: prospects and challenges." *Nature Medicine*, **21**(2):121–131, February 2015. Number: 2 Publisher: Nature Publishing Group.

[CRC13] Le Cong, F. Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D. Hsu, Xuebing Wu, Wenyan Jiang, Luciano A. Marraffini, and Feng Zhang. "Multiplex Genome Engineering Using CRISPR/Cas Systems." *Science*, **339**(6121):819–823, February 2013. Publisher: American Association for the Advancement of Science Section: Report.

[CWL17] Guo-hui Chuai, Qi-Long Wang, and Qi Liu. "In Silico Meets In Vivo: Towards Computational CRISPR-Based sgRNA Design." *Trends in Biotechnology*, **35**(1):12–21, January 2017.

[DFS16] John G. Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, Herbert W. Virgin, Jennifer Listgarten, and David E. Root. "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9." *Nature Biotechnology*, **34**(2):184–191, February 2016. Number: 2 Publisher: Nature Publishing Group.

[DPL16] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens." *Cell*, **167**(7):1853–1866.e17, December 2016. Publisher: Elsevier.

[DWC19] Matthew B. Dong, Guangchuan Wang, Ryan D. Chow, Lupeng Ye, Lvyun Zhu, Xiaoyun Dai, Jonathan J. Park, Hyunu R. Kim, Youssef Errami, Christopher D. Guzman, Xiaoyu Zhou, Krista Y. Chen, Paul A. Renauer, Yaying Du, Johanna Shen, Stanley Z. Lam, Jingjia J. Zhou, Donald R. Lannin, Roy S. Herbst, and Sidi Chen. "Systematic Immunotherapy Target Discovery Using Genome-Scale In Vivo CRISPR Screens in CD8 T Cells." *Cell*, **178**(5):1189–1204.e23, August 2019.

[GHA14] Luke A. Gilbert, Max A. Horlbeck, Britt Adamson, Jacqueline E. Villalta, Yuwen Chen, Evan H. Whitehead, Carla Guimaraes, Barbara Panning, Hidde L. Ploegh,

Michael C. Bassik, Lei S. Qi, Martin Kampmann, and Jonathan S. Weissman. "Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation." *Cell*, **159**(3):647–661, October 2014.

[HLZ14] Patrick D. Hsu, Eric S. Lander, and Feng Zhang. "Development and Applications of CRISPR-Cas9 for Genome Engineering." *Cell*, **157**(6):1262–1278, June 2014.

[HSW13] Patrick D. Hsu, David A. Scott, Joshua A. Weinstein, F. Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J. Fine, Xuebing Wu, Ophir Shalem, Thomas J. Cradick, Luciano A. Marraffini, Gang Bao, and Feng Zhang. "DNA targeting specificity of RNA-guided Cas9 nucleases." *Nature Biotechnology*, **31**(9):827–832, September 2013. Number: 9 Publisher: Nature Publishing Group.

[JCF12] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science*, **337**(6096):816–821, August 2012. Publisher: American Association for the Advancement of Science Section: Research Article.

[KCR16] Alexandra Katigbak, Regina Cencic, Francis Robert, Patrick Sénécha, Claudio Scuoppo, and Jerry Pelletier. "A CRISPR/Cas9 Functional Screen Identifies Rare Tumor Suppressors." *Scientific Reports*, **6**(1):38968, December 2016. Number: 1 Publisher: Nature Publishing Group.

[KD18] Gavin J. Knott and Jennifer A. Doudna. "CRISPR-Cas guides the future of genetic engineering." *Science*, **361**(6405):866–869, August 2018. Publisher: American Association for the Advancement of Science Section: Review.

[KLT14] Hiroko Koike-Yusa, Yilong Li, E.-Pien Tan, Martin Del Castillo Velasco-Herrera, and Kosuke Yusa. "Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library." *Nature Biotechnology*, **32**(3):267–273, March 2014. Number: 3 Publisher: Nature Publishing Group.

[LHA14] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**(12):550, December 2014.

[LXX14] Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I. Love, Feng Zhang, Rafael A. Irizarry, Jun S. Liu, Myles Brown, and X. Shirley Liu. "MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens." *Genome Biology*, **15**(12):554, December 2014.

[MDL16] David W. Morgens, Richard M. Deans, Amy Li, and Michael C. Bassik. "Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes." *Nature*

Biotechnology, **34**(6):634–636, June 2016. Number: 6 Publisher: Nature Publishing Group.

[MJL21] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. "Sustainable data analysis with Snakemake." *F1000Research*, **10**:33, April 2021.

[PCZ14] Randall J. Platt, Sidi Chen, Yang Zhou, Michael J. Yim, Lukasz Swiech, Hannah R. Kempton, James E. Dahlman, Oren Parnas, Thomas M. Eisenhaure, Marko Jovanovic, Daniel B. Graham, Siddharth Jhunjhunwala, Matthias Heidenreich, Ramnik J. Xavier, Robert Langer, Daniel G. Anderson, Nir Hacohen, Aviv Regev, Guoping Feng, Phillip A. Sharp, and Feng Zhang. "CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling." *Cell*, **159**(2):440–455, October 2014.

[PSK17] Shashank J. Patel, Neville E. Sanjana, Rigel J. Kishton, Arash Eidizadeh, Suman K. Vodnala, Maggie Cam, Jared J. Gartner, Li Jia, Seth M. Steinberg, Tori N. Yamamoto, Anand S. Merchant, Gautam U. Mehta, Anna Chichura, Ophir Shalem, Eric Tran, Robert Eil, Madhusudhanan Sukumar, Eva Perez Guijarro, Chi-Ping Day, Paul Robbins, Steve Feldman, Glenn Merlino, Feng Zhang, and Nicholas P. Restifo. "Identification of essential genes for cancer immunotherapy." *Nature*, **548**(7669):537–542, August 2017. Number: 7669 Publisher: Nature Publishing Group.

[R C20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[SJ14] Jeffry D. Sander and J. Keith Joung. "CRISPR-Cas systems for editing, regulating and targeting genomes." *Nature Biotechnology*, **32**(4):347–355, April 2014. Number: 4 Publisher: Nature Publishing Group.

[SLM17] Chun-Qing Song, Yingxiang Li, Haiwei Mou, Jill Moore, Angela Park, Yotsawat Pomyen, Soren Hough, Zachary Kennedy, Andrew Fischer, Hao Yin, Daniel G. Anderson, Darryl Conte, Lars Zender, Xin Wei Wang, Snorri Thorgeirsson, Zhiping Weng, and Wen Xue. "Genome-Wide CRISPR Screen Identifies Regulators of Mitogen-Activated Protein Kinase as Suppressors of Liver Tumors in Mice." *Gastroenterology*, **152**(5):1161–1173.e1, April 2017.

[SRJ14] Samuel H. Sternberg, Sy Redding, Martin Jinek, Eric C. Greene, and Jennifer A. Doudna. "DNA interrogation by the CRISPR RNA-guided endonuclease Cas9." *Nature*, **507**(7490):62–67, March 2014. Number: 7490 Publisher: Nature Publishing Group.

[SSZ15] Ophir Shalem, Neville E. Sanjana, and Feng Zhang. "High-throughput functional genomics using CRISPR–Cas9." *Nature Reviews Genetics*, **16**(5):299–311, May 2015. Number: 5 Publisher: Nature Publishing Group.

[TMH16] Josh Tycko, Vic E. Myer, and Patrick D. Hsu. "Methods for Optimizing CRISPR-Cas9 Genome Editing Specificity." *Molecular Cell*, **63**(3):355–370, August 2016. Publisher: Elsevier.

[WLZ19] Jun Wei, Lingyun Long, Wenting Zheng, Yogesh Dhungana, Seon Ah Lim, Cliff Guy, Yanyan Wang, Yong-Dong Wang, Chenxi Qian, Beisi Xu, Anil Kc, Jordy Saravia, Hongling Huang, Jiyang Yu, John G. Doench, Terrence L. Geiger, and Hongbo Chi. "Targeting REGNASE-1 programs long-lived effector T cells for cancer therapy." *Nature*, **576**(7787):471–476, December 2019. Number: 7787 Publisher: Nature Publishing Group.

[WND16] Addison V. Wright, James K. Nuñez, and Jennifer A. Doudna. "Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering." *Cell*, **164**(1):29–44, January 2016. Publisher: Elsevier.

[WWS14] Tim Wang, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. "Genetic Screens in Human Cells Using the CRISPR-Cas9 System." *Science*, **343**(6166):80–84, January 2014. Publisher: American Association for the Advancement of Science Section: Report.

[ZIL19] Anqi Zhu, Joseph G Ibrahim, and Michael I Love. "Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences." *Bioinformatics*, **35**(12):2084–2092, June 2019.