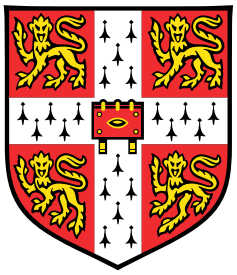
Experimental Design of Single Cell Sequencing Experiments



Alexander Baker

Supervisor: Prof. Florian Markowetz

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Hughes Hall October 2023

I would like to dedicate this thesis to my loving parents William Baker, Erica Audette-Shotwell and my wife Laurel ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Alexander Baker
October 2023

Acknowledgements

I want to express my heartfelt gratitude and acknowledge the invaluable contributions of Professor Markowetz, Mr Will Orchard, and Mr Ryan Blake in the successful completion of my thesis. Their unwavering support, guidance, and intellectual stimulation have been instrumental in shaping the trajectory of my research journey.

First and foremost, I extend my deepest appreciation to Professor Markowetz for his continuous support and mentorship throughout the entire duration of my thesis. I am grateful for his guidance, which has not only expanded my knowledge but also pushed me to explore and delve deeper into areas of my own interest. His encouragement and belief in my abilities have been paramount in my growth as a researcher.

Additionally, I would like to extend my gratitude to Mr Orchard and Mr Blake for the thought-provoking discussions and insightful debates we have engaged in. Our in-depth conversations regarding single-cell CRISPR screens, networks, and causality have greatly enriched my understanding and perspective. Will's and Ryan's intellectual curiosity, analytical thinking, and willingness to explore novel ideas have been a constant source of motivation for me. I am truly grateful for the collaborative environment we shared, which enhanced the quality and depth of my research.

Furthermore, I would like to acknowledge the support and assistance provided by Professor Markowetz, Mr Orchard, and Mr Blake in reviewing my work, providing constructive feedback, and offering valuable suggestions. Their expertise and attention to detail have undoubtedly contributed to the refinement and improvement of my thesis.

I would also like to extend my gratitude to my family, friends, and loved ones for their unwavering support, patience, and understanding. Their presence and belief in my abilities have been a constant source of motivation and strength.

Abstract

This thesis addresses limitations of current single-cell sequencing technologies and proposes alternative experimental designs to increase statistical power.

Single-cell sequencing enables high-throughput and high-dimensional studies of biological systems. This is particularly useful in functional genomics screens that introduce perturbations to investigate and reconstruct the regulatory networks within and between cells. CRISPR screens are the leading method of conducting functional genomics screens due to their specificity, precision and ease of use. Single-cell sequencing and CRISPR screens have recently been integrated to create the first generation of high-throughput and high-dimensional functional genomics screens.

However, the development of efficient experimental design is lagging behind technical advances for single cell CRISPR (scCRISPR) screens. To address this challenge, I developed two wet-lab-aware statistical simulators to compare various experimental protocols and assess their performances. I specifically studied (1) how different protocols affect the performance of scCRISPR screens and (2) how to reduce overall sparsity in the data.

First, to increase the statistical power of scCRISPR screens, new alternative experimental protocols need to be investigated. However, conducting these experiments is time consuming and expensive; therefore, I developed crisprPower, a statistical simulator capable of simulating scRNA-Seq CRISPR screens and allowing researchers to investigate alternative protocols. Simulations showed that the current experimental design of scRNA-Seq CRISPR screens is underpowered, requiring at least 600 cells to observe the effect of a perturbation compared to targeted panels that only need 100 cells.

Second, I proposed an improved single-cell experimental protocol that decreases sparsity. I developed a new simulator, Minerva, to simulate the effects that experimental enrichment protocols (e.g. antibody pulldowns or PCR) would have upon the observed counts of single-cell datasets. Using Minerva, I showed that it is possible to reduce the sparsity of single-cell datasets and measure lowly expressed genes that would have been lost otherwise. My research shows that improving the experimental design of single-cell protocols using theoretical analysis and simulation leads to concrete and easily implemented recommendations that improve scRNA-Seq CRISPR screens and single-cell sequencing methods.

Table of contents

Nomenclature

Roman Symbols

aDisp Asymmetric Dispersion

ARI Adjusted Rand Index

Cas CRISPR-associated

cDNA copy DNA

CRISPR Clustered regularly interspaced short palindromic repeats

CRISPRa CRISPR Activation

CRISPRbe CRISPR Base Editors

CRISPRi CRISPR Interference

CRISPRko CRISPR Knockout

DBS double strand break

dCas dead-CAS

E-Genes Effect Genes

eCFD empirical Cumulative Frequency Distribution

EDA Exploratory Data Analysis

eMI empirical Mutual Information

ePois Extra Poisson

eQTL expression Quantitative Trait Loci

FNH Fisher Noncentral Hypergeomtric

GBC Guide Barcode

GO Gene Ontology

gRNA guide RNA

GRNs Gene Regulatory Networks

GWAS Genome-Wide Association Studies

HVGs High Variance Genes

KDE Kernel Density Estimate

LFC Log-Fold Change

LTR Long Terminal Repeat

MCMC Markov Chain Monte Carlo

MFNH Multivariant Fisher Noncentral Hypergeometric

MI Mutual Information

MLE Maximum Likelihood Estimate

NB Negative Binomial

NHEJ Non-Homologous End Join

PCR Polymerase Chain Reaction

PMF Probability Mass Function

QC Quality Control

RNAi RNA Interference

RT-PCR Reverse Transcription PCR

scCRISPR single cell CRISPR

scRNA-Seq single cell RNA-sequencing

smRNA single molecule RNA

TAP-Seq Targeted Panel Sequencing

TF Transcription Factor

UMI Unique Molecules Identifier

WNH Wallienus Noncentral Hypergeometric

ZINB Zero-Inflated Negative Binomial

Chapter 1

The Needle in the Hay Stack

Single-cell sequencing technologies have revolutionized our understanding of biological systems and our ability to interrogate them efficiently, enabling researchers to understand the phenotypic diversity of cell populations within organisms [35, 144]. Specifically, since the onset of high-throughput droplet-based single-sequencing methods, researchers have been able to interrogate the biological context of both health and disease settings to identify novel cell types, capture interactions on both the intra- and intercellular levels, and provide a high-level characterization of these populations via expression and/or chromatin accessibility [82, 26, 120]. Owing to the improved resolution, current single-cell sequencing methods greatly increase the throughput and reduce the overall experimental burden because researchers no longer have to conduct additional experiments like isolating cell populations via flow cytometry to characterize them further. The greatest promise of single-cell sequencing is enabling greater throughput of perturbation experiments, allowing researchers to scale their experiments rapidly and interrogate biological systems to a previously impossible degree [1, 36].

1.1 Open Problems in Single Cell Sequencing Experiments

Despite the excitement and promise, single-cell sequencing technologies currently suffer from limitations, the largest of which is the relatively shallow coverage of the data modality (e.g. transcriptome or chromatin accessibility) [144, 130]. The shallow coverage results in sparsity in the observed count matrix of a single-cell experiment. Despite this limitation, single-cell technologies have been widely adopted, enabling researchers to study biological systems in great detail. Single-cell technologies are generally used in heterogeneous biological contexts, either in the form of a tissue or organ. In these contexts, the expression distributions of genes can vary dramatically between cell populations [91]. Current single-cell experiments' ability

to observe cell populations depends upon the innate difference expression distribution of the highest expressed genes.

While many methods have been developed to overcome sparsity, such as imputation, the limitations of single-cell sequencing experiments are particularly visible in scCRISPR screens [60]. scCRISPR screens hold tremendous potential, providing researchers with a high-throughput and high-resolution tool to interrogate genotype effects upon observed phenotypes. However, for this to work, you need to be able to fully characterise the expression distribution of a given cellular perturbation, as the effect of a perturbation can vary dramatically. The perturbation effect of a specific gene encompasses not only the immediate impact on the gene expression distribution but also the cumulative effects its perturbation has on its own and other gene expression distributions. For instance, essential TFs like GATA3 or ESR1 exhibit large perturbation effects. Modifying their expression not only influences their own expression distributions but also induces differential expression in a multitude of other genes. The size of a gene's perturbations effect places a fundamental limitation on what can be seen as a perturbation scale, decreasing the number of cells required to observe the perturbations increases exponentially. As such, scCRISPR screens currently have an effective limit on the type of gene that can be studied to those genes with a perturbation effect that dramatically alters cell state or transitions [114, 115]. Even under these conditions, whether the perturbation is fully characterized or is simply large enough to be observed is unclear. Ultimately, characterizing the entire transcriptome of a perturbation or a rare cell population using modern single-cell sequencing methods is equivalent to finding a needle in a haystack.

1.2 My Contributions

I have investigated the experimental design of scCRISPR screens, examined their current statistical power and determined how many cells per perturbation are required to observe an effect. I explored the utility of targeted panels as an alternative sequencing protocol to improve the statistical power of scCRISPR screens. I developed a statistical simulator called crisprPower that can simulate scCRISPR screens for CRISPR interference and knockout experiments to investigate alternative experimental designs efficiently. It can simulate off-target effects via simulating a gene regulatory network and propagate a perturbation across the network.

During the later stages of the investigation into the experimental design of scCRISPR screens, I realised that every transcript could be thought of as having an equal probability of being sequenced. However, one can alter the probability of sequencing a given transcript by using enrichment methods to increase or decrease the probability of sequencing a given

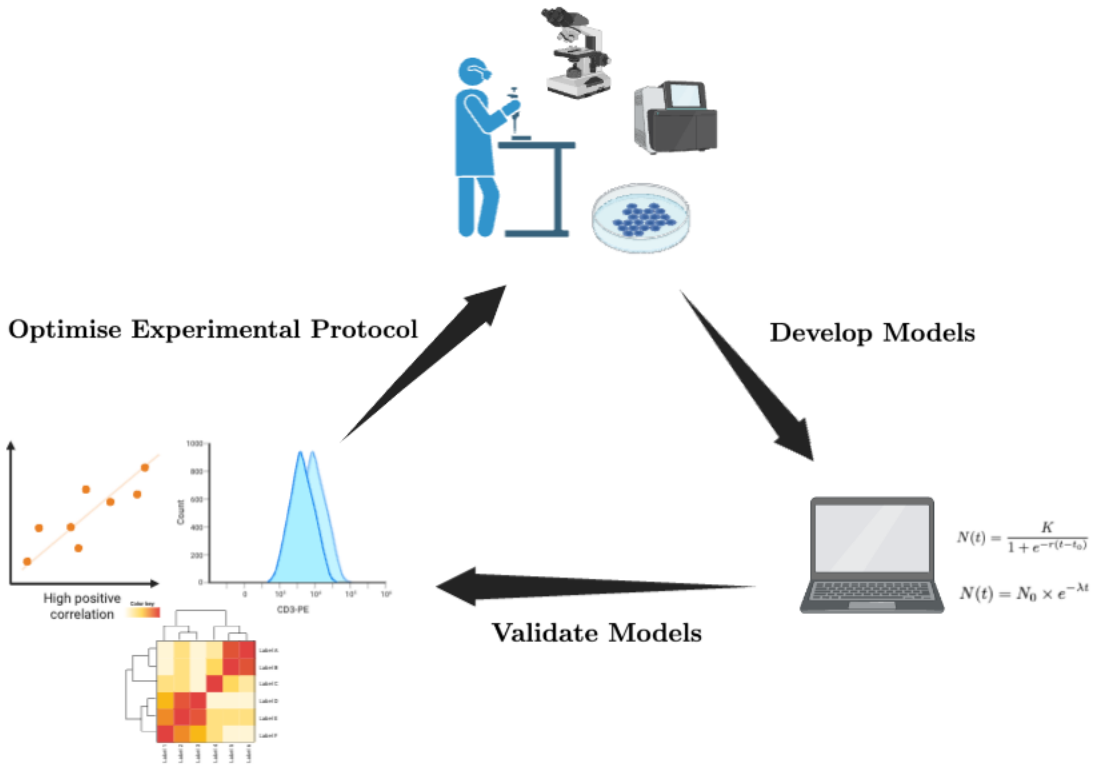


Fig. 1.1 A schematic representation of the experimental lab workflow, illustrating the iterative process of designing and validating optimal experimental protocols. In my research, I have consistently focused on the theoretical modelling of experimental protocols to enhance their statistical power. By constructing these models, I gain valuable insights into the technical intricacies of the experiments under investigation. Leveraging this understanding, I iteratively refine and develop experimental designs that maximize statistical power within the confines of the model's parameters. Subsequently, the optimized designs need to be validated in the wet lab to pave the way for their practical implementation and future utilization.

transcript using various enrichment methods. Based on this insight, I developed a new simulator called Minerva that incorporates gene-specific weights that represent potential enrichment protocols that could be used to improve the representation of gene transcripts. Allowing me to investigate how the enrichment of low to medium-expressed genes improves the characterization of the transcriptome.

Throughout this thesis, my primary objective has been to tackle the challenges currently faced by single-cell sequencing technologies and propose alternative experimental designs that enhance the statistical power of these experiments (as depicted in Figure 1.1). The next chapter begins with an in-depth review of the experimental design of CRISPR screens. Here, I delve into a step-by-step guide, unravelling the intricacies, techniques, and challenges involved in the design of CRISPR screens. Moving forward to Chapter 3, I extensively

examine the development of crisprPower, focusing on the computational and statistical methods employed. Additionally, I validate its statistical characteristics by comparing it with real biological datasets. Chapter 4 encompasses the application of crisprPower to explore alternative scCRISPR experimental designs, identifying effective ways to enhance their statistical power. Shifting to Chapter 5, I discuss the development of Minerva, providing a detailed description of its model while comparing its statistical characteristics with real data and other single-cell simulators like Splatter [139]. In Chapter 6, I investigate how manipulating the probability of transcripts being sequenced influences the observed count distribution of genes. Finally, in Chapter 7, I review the conclusions of the thesis and explore possible future work.

Chapter 2

Designing a CRISPR Screen Experiment: A Step-by-Step Guide

CRISPR screens have emerged as a powerful tool enabling researchers to efficiently perturb genes both individually and in sets with high sensitivity and specificity. In the following chapter, I will briefly review CRISPR and CRISPR screens, discuss the motivating factors to conduct a screen, describe what current CRISPR screens are currently capable of perturbing, and outline how to best design guide RNA (gRNA) to reduce noise and maximize power. Finally, I will provide an overview of scCRISPR screens and the ultimate promise of enabling high-resolution and high-throughput screens despite their limitations.

2.1 Introduction

Clustered regularly interspaced short palindromic repeats (CRISPR) were observed in the late 1980s when researchers discovered the repetitive DNA sequence array that characterizes CRISPRs in Escherichia coli [63]. While the purpose of CRISPRs was initially unknown, continued research deduced that the DNA sequences within the sequence array aligned to a diverse range of bacteriophage, indicating that these sequences provide bacteria with a targeted immune response [102]. In addition, RNA and DNA nucleases were commonly found encoded near CRISPR and referred to as CRISPR-associated (Cas) proteins. Later it was shown that CRISPRs were expressed as RNA molecules (commonly referred to as gRNAs) that form protein-RNA complexes with Cas proteins which form the core bacterial immune response to infection. Specifically, CRISPR gRNA directs an associated nuclease to induce strand breaks in the viral RNA or DNA genomes (see figure 2.1) [13, 20]. However, this was initially of little interest beyond the initial excitement in understanding bacterial immune

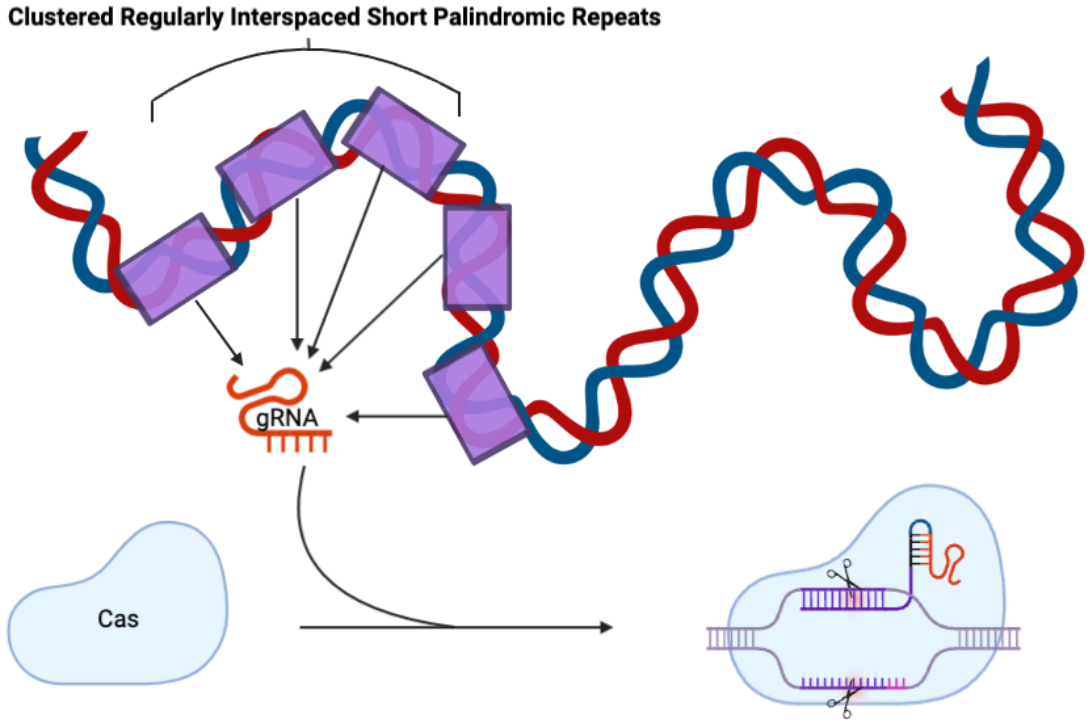


Fig. 2.1 Diagram illustrating the conversion of CRISPR DNA sequences into gRNA and the formation of a CRISPR-Cas protein-RNA complex. The CRISPR DNA sequences are transcribed and processed into precursor gRNA molecules, which are further cleaved and modified to form mature gRNA. The mature gRNA then associates with Cas proteins, forming a CRISPR-Cas protein-RNA complex capable of targeting specific DNA sequences for gene editing or regulation.

response until it was demonstrated the CRISPR-Cas complexes could be reprogrammed to target specific DNA sequences other than their original targets[67].

Before CRISPR became the mainstay of genome editing, the primary gene editing tools used Zinc finger proteins or TALENs. CRISPR's rise to prominence and eventual replacement of these methods is due to the simplicity, efficiency, and versatility it provides researchers when targeting genome sections [2, 74, 133]. Specifically, Zinc finger proteins and TALENs require the design of a custom protein for each target sequence, which is time-consuming and costly. In contrast, CRISPR only requires the design of a single RNA guide, which can be easily synthesized and modified for different targets. Another advantage of CRISPR is its high specificity, as it is less likely to make off-target edits [49]. Zinc finger proteins and TALENs can sometimes bind to similar sequences and cause unintended mutations, but CRISPR can be designed to target a particular DNA sequence more specifically.

RNA interference (RNAi) is an alternative technique to silence genes by triggering the degradation of specific mRNA molecules. While RNAi can effectively reduce the expression of target genes, it can also result in off-target effects and incomplete knockdown of gene expression. In contrast, CRISPR screens use the Cas9 enzyme to introduce targeted mutations or gene knockouts with high specificity and efficiency [2]. This allows for more precise perturbation of gene expression and avoids RNAi's off-target effects. Another advantage of CRISPR screens is their ability to target non-coding regions of the genome, such as enhancers and promoters, which play critical roles in regulating gene expression [9]. In contrast, RNAi-based screens can only target protein coding regions genes. CRISPR screens can simultaneously target multiple genes, enabling studying complex genetic interactions and identifying gene networks involved in various biological processes [89]. Overall, the higher specificity, efficiency, and versatility of CRISPR screens have made them increasingly popular for genetic studies and to interrogate biological systems. Using different enzymes, CRISPR can cut DNA, activate or repress gene expression, base pair editing, and even target the epigenome and RNA molecules, altering splicing etc. [74, 133].

Since CRISPR became the mainstay of the genomic engineering toolbox, its utility and application diversity has only increased. CRISPR's initial major use case was to conduct knockout screens to identify essential genes in organisms and particular biological contexts [135, 134]. Quickly expanding to other experimental applications such as DNA or RNA imaging enabled the targeted study of chromatin localization dynamics. However, the most interesting applications of CRISPR have been in biomedical sciences from breakthroughs in FDA-approved gene therapy treatment for patients suffering from genetic disorders such as sickle cell disease to cellular rejuvenation therapies currently in clinical trials [51, 48]. CRISPR has dramatically improved our capacity to develop GMO crops to increase their disease and stress resistance and improve their nutritional value [133]. Finally, another interesting development has been CRISPR-based diagnostic methods, dramatically increasing the sensitivity and specificity of tests [53, 23].

While the application of the CRISPR system has dramatically increased over the past decade and shows no sign of slowing down, I will discuss the process of designing a CRISPR screen experiment, the reasons why one may conduct a CRISPR screen, what can CRISPR target, and how can a target of interest be perturbed. Finally, I'll discuss the current and future impact of high-throughput and high-resolution screens utilizing single-cell sequencing technology.

2.2 Why Conduct a CRISPR Screen?

The overarching reason to conduct a CRISPR screen is to interrogate the role and function of individual genes or entire biological systems by characterising the phenotypic response of a perturbation on the molecular and/or organism level [16]. Many different types of CRISPR screens have been developed to be conducted in either an *in vivo* or *ex vivo* setting, along with methods of increasing the phenotypic readout resolution to uncover gene function better. The experiment's goals will determine the type of screen which should be conducted. The experimental context comes with trade-offs; for example, the experimental workload of *in vivo* based CRISPR screens limits the throughput capacity of the screen [31]. As a result, researchers tend to use high-resolution readouts (i.e., images or transcriptomes) to compensate for this limitation. Despite this, *in vivo* experiments can be extremely informative of a given gene's role in a particular disease state in complex organisms, allowing researchers to study the initial development and progression of diseases such as cancer. However, to maximize the efficiency of these studies, researchers need to have a reasonable understanding of what genes are particularly important to the biology they are studying.

To discover novel gene function and behaviour, pooled CRISPR screens are used in an *ex vivo* setting to maximize the number of genes perturbed [16]. While this increases the number of genes that can be perturbed, this has historically dramatically limited the resolution of readouts to one-dimensional growth-based assays [134]. Limiting the ability to characterize gene function fully. Despite this limitation, the throughput from pooled CRISPR screens allows researchers to quickly and efficiently search for genes of interest, albeit at a low resolution. This has begun to change with the development of high-resolution pooled screens utilizing imaging, single-cell transcriptomics, and single-cell chromatin accessibility as phenotypic readouts [36, 44, 111, 117, 40, 79].

Once the purpose of the experiment is well defined, the next question is to determine whether to conduct a single or combinatorial screen. Each has its advantages, but CRISPR screens targeting individual genes are more prominent due to their simplicity and ease of perturbing all of the genes within the genome compared to the exponential increase in the number of potential perturbations that can be done in a combinatorial setting. Targeting individual genes allows for the identification of essential genes in a particular biological setting (i.e. cancer cell line) [125, 143, 58]. Researchers can compare previously identified global essential genes to context-specific essential genes of a particular diseased state and identify novel drug targets with reduced chances of side effects [73]. These genes are less likely to play crucial roles in healthy tissue. Furthermore, single perturbation screens can validate a drug's mechanism of action by comparing the phenotypic readout of the drug response to that of the perturbed gene [52]. Overall, single-targeting CRISPR screens offer a

comprehensive overview of biological systems and can effectively identify essential genes and pathways within specific contexts.

However, genes rarely function in isolation and depend upon regulatory networks to control biological processes. To investigate the dependency structure of these networks and systematically investigate genetic interactions requires combinatorial perturbations [27]. A genetic interaction occurs when combinatorial effects differ from the excepted from individual effects. When the difference is positive, this combinatorial effect is called aggravating (this is an example of synthetic lethality), and if this difference is negative, it is called alleviating [85]. An aggravating perturbation is when a combinatorial perturbation induces a stronger-than-expected response to perturbation in comparison to the effects of perturbing the genes individually. While alleviating induces a weaker-than-excepted response to perturbing the genes individually. By characterizing these responses, researchers can understand redundancies within the studied biological system and better refine their drug target list.

2.3 What is Being Perturbed?

Over the past decade, the number of options and methods for introducing perturbations to biological systems has grown dramatically from the original knockout relying upon the imperfect DNA repair from the Non-Homologous End Join (NHEJ) repair pathway. The first major development in expanding the toolbox was in mutating the Cas nucleases to no longer function, creating dead-Cas (dCas) proteins [2]. Using this as the foundation, transcriptional repressors and activators were transfused with dCas to enable the inhibition or activation of genes. More proteins with alternative functional domains have been transfused with dCas proteins enabling researchers to modify cytosine methylation, histone markers, and more [16, 2]. In addition, Cas nickases have been transfused with base modification enzymes, thereby enabling the creation of targeted point mutations. From transcriptional control to point mutations and epigenetic modifications, researchers now have many options for investigating the importance of biological features and characterising their response (see figure 2.2) [9, 2].

2.3.1 Protein Coding Regions

When targeting the genome's protein-coding regions, there are three primary methods of introducing a perturbation. The first is to create a frameshift mutation via a CRISPR Knockout (CRISPRko), which utilizes Cas nuclease to induce a double-strand break (DSB)

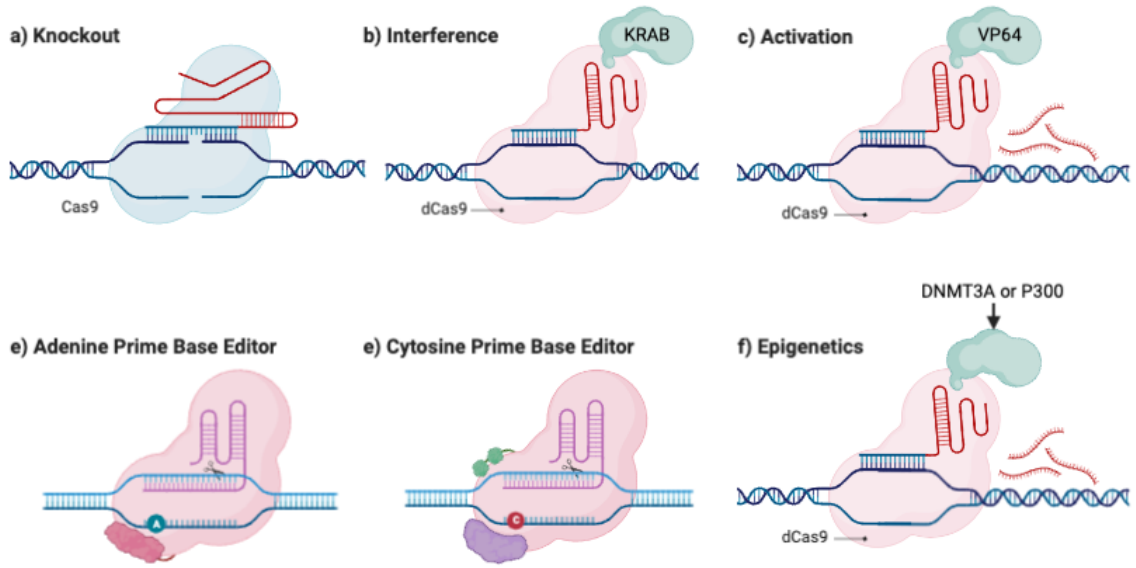


Fig. 2.2 This figure presents an overview of diverse CRISPR variants researchers employ to introduce perturbations in CRISPR screens. a) Knockouts: This approach utilizes an unmodified CRISPR-Cas protein-RNA complex, relying on a nuclease to induce double-strand breaks and subsequent faulty DNA repair mechanisms, leading to perturbation of the targeted gene. b) Interference: In this method, a dCas protein fused with a KRAB repressor domain is employed to downregulate gene expression by preventing transcription initiation or inducing chromatin modifications. c) Activation: Activation utilizes a dCas protein fused with a VP64 activator domain to induce an upregulation in gene expression by promoting transcriptional activation. d) Adenine Base Editors: These editors enable CRISPR-mediated targeted point mutations by employing an Adenine nickase, allowing precise modification of DNA bases. e) Cytosine Base Editors: Similar to Adenine Base Editors, this variant uses a Cytosine-based nickase to achieve targeted point mutations, facilitating specific DNA base alterations. f) Epigenetic Modifications: CRISPR technology also enables the introduction of epigenetic changes such as DNA methylation and histone acetylation. This is achieved by utilizing a dCas protein fused with a transfused DNMT3A or p300 domain, allowing modulation of DNA and histone modifications, respectively.

at a particular target site [67]. The second is altering gene expression through activation using a transfused transcriptional or epigenetic regulator to activate or inhibit expression [22]. When activating expression, it is called CRISPR activation (CRISPRa). When it inhibits expression, it is called CRISPR interference (CRISPRi). In addition, epigenetic modifications can be employed through CRISPR-off and CRISPR-on techniques to regulate gene expression [105]. Additionally, point mutations can be introduced at specific target sites using either CRISPR base editors (CRISPRbe) or CRISPR prime editors (CRISPRpe)

[72, 122]. All of these methods have pros and cons and should be used based on the goals of an experiment and the properties of the studied biological system.

When interested in characterizing the role of a gene in a given state, the most effective method of introducing a permanent perturbation is CRISPRko. This is because the introduced mutation shifts the gene's reading frame [143, 67]. However, despite this, multiple reports have warned about CRISPRko-induced toxicity that results in gene-independent anti-proliferation response [99, 3]. This is caused by the DSB induced by CRISPR, which forces the cell to stop and repair. If enough DSB breaks occur simultaneously, the affected cells' growth rate will significantly reduce to the point it can be observed [3]. This effect is primarily caused by either off-targets (the higher the number of off-targets, the worst the effect) or a high number of copy number amplifications affecting the target site (i.e. cancer cell lines) [99]. In addition, evidence suggests that this CRISPR-induced toxicity varies across the genome, complicating the analysis. CRISPRko is an effective method of inducing permeant perturbations; however, careful planning should be done to improve their efficiency and reduce noise.

Controlling gene expression through CRISPRi or CRISPRa provides an alternative and expands the number of perturbation options available. To induce a perturbation CRISPRi and CRISPRa both require that the target site is within 500 bp of the transcription start site of the target gene [118]. Because the perturbation is fundamentally deterministic and dependent upon a transfused repressor or activator, no DSB occurs, thereby avoiding CRISPR-induced knockout toxicity. However, the perturbations CRISPRi and CRISPRa induce are not permanent, and previous studies suggest the perturbation strength is not as significant as knockouts; this is because dCas9 has to be continuously bound to the target site to induce a perturbation [118]. While addressing the CRISPR-induced toxicity of knockouts, CRISPRi and CRISPRa have their own flaws. They provide alternative methods of perturbing the genes and can be used in biological settings with high copy number profiles (i.e. breast or ovarian cancer cell lines).

A significant limitation of CRISPRi and CRISPRa lies in the enduring stability of their perturbation effects over time. To address this challenge, the techniques CRISPRoff and its counterpart, CRISPRon, were devised to either induce or reverse epigenetically inheritable memory, thereby exerting control over gene transcription for extended durations [105]. CRISPRoff accomplishes long-term transcriptional silencing by employing a dCas9 fused with KRAB and Dnmt4A domains. The KRAB domain initiates a repressive chromatin state at the target locus, fortified through methylation, ensuring sustained silencing of the targeted gene. Importantly, since a gRNA guides Cas9, the target site need not necessarily contain a CpG island, allowing for genome-wide transcriptional silencing and the heritability

of perturbations [105]. CRISPRon functions in opposition to CRISPRoff and is typically employed to reverse the perturbations induced by CRISPRoff. Like CRISPRoff, CRISPRon employs a dCas9, integrating three transactivated protein domains: VP64, p65-AD, and Rta [105]. When binding to a methylated and repressed chromatin site CRISPRon first demethylates and then induces chromatin accessibility. The combination of CRISPRoff and CRISPRon empowers researchers to induce long-lasting and inheritable control over transcriptional perturbations, offering the opportunity to investigate the consequences of reverting such perturbations.

One of the most interesting developments to occur in the recent past has been the development of CRISPR editors. CRISPRbe works by transfusing base-modifying enzymes capable of altering base pairs without inducing a DSB [9, 16]. While base editing efficiency will vary from target site to target site, CRISPRbe presents an interesting opportunity to conduct point mutation-based CRISPR screens. CRISPRbe has been used in CRISPR screens designed to differentiate driver from passenger mutations with great success [109, 29]. Additionally, CRISPRbe has been used to interrogate the clinical implications of point mutations in their ability to alter protein function and inhibit drug-protein interactions [57]. While they are still limited and not as mature as CRIPRko, CRISPRi, and CRISPRa, as more base editors are developed, CRISPRbe will provide an interesting alternative to these mature methods.

However, CRISPRbe has inherent limitations in its ability to introduce specific point mutations, primarily restricted to cytosine-to-thymine and adenine-to-guanine mutations. CRISPRpe, on the other hand, addresses these shortcomings in existing CRISPRbe methods and expands their potential to create various point mutations and indels. The key molecular difference between CRISPRpe and CRISPRbe methods lies in two primary aspects. First, CRISPRpe employs an extended guide RNA (pegRNA) incorporating an RNA-based editing template. Additionally, it integrates an M-MLV RT with Cas9 [122]. This innovative approach harnesses the RNA editing template to induce mutations at specified target sites outlined in the template. Through these adaptations, CRISPRpe can induce a wide range of point mutations and even generate small indels, granting researchers an unprecedented level of control.

2.3.2 Non-Coding Regions

The protein-coding regions are ultimately a small fraction of the genome, with the majority of the human genome being predominately non-coding regions. The ability to perturb the non-coding region of the genome has lagged behind protein-coding regions. This is primarily because the protein-coding region has attracted more attention and is easier to perturb than

the non-coding region by introducing frameshift mutations. There is an increasing interest in perturbing non-coding regions of the genome as Genome-Wide Association Studies (GWAS) have revealed that thousands of non-coding loci are associated with disease [113]. There are two primary methods of perturbing non-coding regions: CRISPRi/a or introducing point mutations via CRISPRbe.

The most common non-coding region perturbation method is CRISPRi or CRISPRa by modifying the chromatin state surrounding the target site. Specifically, when inhibiting a dCas is transfused with a KRAB domain which induces a heterochromatin state of the DNA within 1-2 kilobases of the target site [50]. Similarly, CRISPRa transfused with MSK1 or VPR domain can be used to unwind heterochromatin by promoting histone acetylation within 1-2 kilobases of the target site [16]. Typically, CRISPRi and CRISPRa to identify enhancers and promoters and attempt to link them to the genes they regulate. One of the more interesting developments has been utilising CRIPSRi and CRISPRa to evaluate disease-associated loci in non-coding regions. Integrating association frameworks from expression Quantitative Trait Loci (eQTL) studies with CRISPR screens enables validation of associated loci [50]. Additionally, researchers can expand the search to loci that do not have frequent mutations allowing a greater understanding of the regulatory interactions between non-coding regions and the genes they regulate.

While validating the phenotypic behaviour of disease-associated loci is important, a currently understudied aspect of GWAS is determining the mechanism of action in which these mutated loci induces perturbation [16, 113]. CRISPRbe screens provide researchers with an interesting opportunity to address these questions. Specifically, CRISPRbe can introduce point mutations in validated or associated loci to unveil the regulatory mechanism. An example of this would be to saturate given loci with multiple gRNAs enabling multiple alternative mutations to occur and, using the eQTLs association framework, assess the impact each mutation has on a given disease state [113]. Such screens validate loci disease association and enable us to assess the potential perturbation responses of specific loci mutations.

2.3.3 Epigenetics

A new set of CRISPR perturbation methods currently being developed involves perturbing the epigenome. Using dCas as a foundation, an ever-increasing variety of protein domains are being transfused to take advantage of CRISPR's easy programmability and specificity to modify methylation profiles, histone acetylation, and even the splicing of mRNA molecules [84, 106, 107]. Opening the door to a new series of CRISPR screens that expands us beyond genomic engineering to epigenomic engineering. Enabling researchers to identify, validate

and interrogate epigenetic-based eQTL. While these technologies have yet to enter wide-scale use in CRISPR screens, these developments only expand their utility for exploring regulatory functions via perturbations in biological systems.

2.4 Choosing a Screens Resolution

There are three broad classes of resolution: low, medium, and high dimensional readouts. They correspond to growth, image, and single-cell-based CRISPR screens. Each screen has advantages and disadvantages and generally reflects a trade-off between ease of use, budget and resolution. Here I will go through each level of resolution, discuss their use case, and when they should be used.

2.4.1 Growth Based Assays

Growth-based CRISPR screens are the lowest dimensional readout that is available. They are one-dimensional and reflect only changes between two points in the read counts of a gRNA which is then extrapolated to infer changes in growth rates. To conduct one of these screens, first infect a cell population using lentivirus with multiple gRNAs per gene. The cells then undergo a positive selection process where cells infected with lentivirus are given an advantage over others. This can be done through fluorescent signalling and selecting for them using flow cytometry or, more commonly, through a puromycin selection where cells infected with the lentivirus gain antibiotic resistance [125, 143]. These cells are then allowed to grow for about a week when the first time point is taken, referred to as the baseline sample. Next, the cells grow for another 2-3 weeks until another sample is taken. Sometimes multiple time points samples are taken, but this is not normally the case. Next, calculate the Log-Fold Change (LFC) between the baseline and later sample read counts of gRNA. For each target gene, the individual LFC per gRNA is averaged and used to rank and order the gene list showing which perturbed genes induced the greatest change in the cell's growth rate [125, 143, 58].

These assays are highly valuable when the goal is to identify specific genes that when disrupted, impact the rate of cell population growth. These screening methods are primarily employed to identify essential genes associated with cell populations and disease states [39]. Consequently, they are well-suited for identifying potential drug targets and distinguishing between driver and passenger mutations; by assessing changes in the growth rate of cancer cell lines, these assays offer valuable insights. However, it is important to note that these assays only apply when studying phenomena that influence growth rates. A similar experimental

workflow and analysis can also conduct an assay targeting marker genes (such as receptors) [16]. In this case, instead of measuring growth, cells are sorted using flow cytometry, and the analysis focuses on differential abundance.

The primary disadvantage of these low-dimensional assays is that observing the effect of a gene perturbation requires the gene to play an essential role in the cell, having both wide and large-scale effects to be observed as an increase or decrease in cellular growth. However, despite these limitations, these screens have advantages, with the primary one being that there relatively cheap and easy to conduct in terms of both experimental workflow and analysis compared to the other higher-dimensional screens.

2.4.2 Image Based Assays

Imaged-based CRISPR screens offer a middle ground between growth and single-cell-based assays. Like the other screens, cells are first infected with gRNA via lentivirus and undergo positive selection. Once infected, cells are monitored by fluorescent or confocal microscopy [44, 40, 79]. To identify gRNAs within a given cell *in situ*, sequencing is used to identify a unique barcode associated with gRNA. Due to the enzymatically driven amplification that occurs during *in situ* sequencing, the barcodes can be easily identified at a low level of magnification [44, 40, 79]. Thereby enabling the simultaneous observation of hundreds of thousands to millions of cells.

Imaged-based assays are extremely flexible regarding the phenotypical readout and various potential experiments that can be conducted. When performing these screens, cells are monitored temporally, allowing for the continuous monitoring of cells' response to perturbation, assessing the subcellular localization of mRNA or proteins, quantifying mRNA utilizing smFISH, and extracting cell morphological features [44, 80]. With numerous options, these assays provide researchers with near-endless opportunities to analyze biological settings with a reasonably high degree of resolution and statistical power.

However, researchers often encounter challenges when performing these assays, particularly in data analysis. While analyzing single molecule RNA and subcellular localization can be relatively straightforward, it requires computational resources to capture the data and extract meaningful biological insights effectively. An additional complication arises when analyzing cell morphologies, as the features derived from such analysis may be associated with perturbations but may not directly reveal the underlying mechanisms at play. This lack of clear linkage between the extracted features and the latent biology presents a significant obstacle [80]. Despite these obstacles, image-based assays offer researchers flexibility and dynamic observations, providing a relatively high resolution for studying perturbation responses.

2.4.3 Single Cell Sequencing

scCRISPR screens work similarly to all previous methods, with the primary difference being the high-dimensional readout of biologically relevant features. Like the other methods, it infects cells with gRNAs via a lentivirus vector and undergoes a positive selection process. A key difference between single-cell and other assays is that the lentivirus vector is modified to improve cell detection. Instead of using barcodes like in image-based arrays, the gRNA is directly captured and used to identify its presence in a given cell [60]. Once selected, cells are allowed to grow for a week and are then isolated, lysed, and processed for sequencing. Two main data modalities are commonly employed in single-cell studies: scRNA-Seq and scATAC-seq. When it comes to scCRISPR screens, droplet-based microfluidics is frequently utilized, with 10x Genomics being a popular choice due to its widespread usage and high throughput capabilities [114, 115]. However, alternative single-cell methods have also been employed, such as Particle-templated instant partition sequencing [25]. Each method has its own strengths and limitations regarding single-cell sequencing and data modalities, and the choice should be made based on the specific experimental objectives.

The primary advantage of scCRISPR screens is they provide researchers with a high-throughput high-resolution method with the phenotypic readout having biologically relevant features. Depending upon the experiment's goals, will choose what modality to use, but each has its use case. For example, in scCRISPR screens, researchers can employ scATAC-seq to examine the impact of perturbing various transcription factors (TFs) and enhancers on chromatin accessibility and reconstruct the gene regulatory network of cell populations [111, 117]. While scRNA-Seq tends to be used to characterise the response to perturbation of a crucial biological pathway in a given state, monitor how a perturbation alters a biological process, and identify the mechanism of action of drugs [47]. Examples of this are the many studies focusing on immune-related subjects such as immune checkpoint and immune evasion in cancer. In addition, scRNA-Seq scCRISRP screens have been used to study the epithelial-mesenchymal transition and validate gene association to autism [97, 66].

Since their introduction in 2016, scCRISPR screens have undergone rapid development, expanding the range of data modalities and improving experimental protocols. One major challenge encountered was barcode swapping, which refers to the random swapping of guide barcodes, unique barcodes mapped to specific gRNAs, among different lentiviral vectors. This swapping occurred due to the 5' location of the guide barcodes on the lentiviral barcode during the assembly of gRNA libraries in pooled experiments.

Fortunately, the emergence of ECCITE-Seq, Direct Capture Perturb-Seq, and other associated techniques has effectively tackled these challenges. These methods achieve this by relocating the gRNA location to the 3' end of the lentiviral vector and directly

capturing the gRNA sequence during the process of identifying gRNAs within a specific cell [60, 114, 101]. However, scCRISPR screens still face a significant obstacle regarding statistical power, particularly in scRNA-Seq-based screens, when studying more nuanced perturbations beyond cell state essentials. To address this, there is a growing emphasis on developing targeted panels to enhance the screening power [114, 123]. In contrast, scATAC-Seq-based scCRISPR screens encounter fewer statistical power challenges. These screens primarily focus on the accessibility of binding sites and the impact of differential accessibility patterns on biological systems. They treat accessible regions as a binary dataset, allowing for more robust analyses and interpretations of the results [128]. Nonetheless, ongoing research aims to address these issues by increasing the number of cells and the depth of single-cell sequencing. scCRISPR screens hold tremendous promise, offering researchers unparalleled throughput and resolution to investigate biological systems and their responses to perturbations.

2.5 Designing a gRNA Library

Selecting the optimal gRNAs for an experiment is one of the most, if not the most important, steps in designing CRISPR screens. It should occur after deciding on the purpose of the screen, how perturbations will be introduced, and the screen's resolution, as all these steps influence a gRNA's effect on the screen readout. The primary aim of designing a gRNA library for an experiment is to maximize the strength of the perturbation at a given target site. While simultaneously minimizing the perturbation strength at off-target sites, if not outright eliminating them. A general rule of thumb is that the higher the resolution of the CRISPR screen, the more important it is to identify high-quality gRNAs with minimal to no off-targets. As the resolution increases, it becomes easier for off-target effects to introduce noise into the data making the design of the gRNA library even more important.

The design process varies based on whether Cas or dCas is used. Cas-driven perturbations are currently the best characterised of the two and have numerous studies designing optimal gRNA libraries that maximize perturbations at target sites while minimizing off-targets. These studies have created many excellent gRNA libraries, such as the Brunello and Gattinara libraries, from which there are three to five gRNAs to choose per target gene [37, 118]. In general, for most experiments, selecting gRNAs from these gRNA libraries is sufficient to achieve the desired objectives. However, there are situations where this is not applicable. Specifically, gRNA libraries are typically confined to the protein-coding regions of mice and humans. Therefore, if there is an interest in conducting a screen beyond these conditions, custom characterization of gRNAs becomes necessary.

2.5.1 Measuring On-Target Activity of a gRNA

The On-Target Activity of a gRNA (or guide efficiency) serves as a metric for how effectively a given guide can induce a perturbation at a specific target site. Efficiency is typically evaluated using an On-Target Activity Score. For gRNAs targeting the human and mouse genomes, these scores have already been predicted and are readily accessible via the UCSC Genome Browser [104]. The On-Target Activity Score employs a machine learning model to predict the on-target activity of a gRNA, assigning it a value between 0 and 1. Several distinct On-Target Activity Scores have been developed and use a variety of machine learning regression techniques [75].

All of these methodologies employ a consistent experimental approach to generate datasets for predicting the On-Target Activity of gRNAs. Specifically, this approach involves selecting a set of typically essential genes and designing a gRNA library that comprehensively covers these target genes using dropout screens, typically growth-based screens, to assess the gRNA's capability to induce perturbation. When I mention "tiling" a target gene, it implies that a significant portion or all of the gRNA target sites within those genes are incorporated into the screening process [37]. The primary disparities among the datasets used to train various On-Target Activity metrics generally pertain to their scale or CRISPR modulality. For instance, Azithum's (also known as the Fusi Score) On-Target Activity model was trained using a dataset derived from a tiling gRNA library targeting 15 specific genes, whereas the Vienna Bioscore employed a substantially larger tiling library encompassing 159 essential genes [37, 100].

Another distinguishing factor in these experiments hinges on whether the On-Target Activity estimation pertains to dCas9 (Interference or Activation) or Cas9. For dCas9 models, the gRNAs chosen for the tiling dropout or growth experiment are restricted to the Transcription Start Site (TSS), with the nucleotide distance from the TSS varying depending on whether it's activation or interference [62, 118]. Specifically, for interference, optimal gRNA distance to the TSS where found to be approximately 25 to 75 nucleotides in front of the TSS, whereas for activation, it ranges from 100 to 150 nucleotides after the TSS. Additionally, the phenotypic outcome of the assay is contingent on the CRISPR modality under assessment [118]. Knockout and Interference On-Target Models employ dropouts, while Activation models use growth assessments, typically in a resistance context, to gauge a given gRNA's perturbation strength [118].

The two most prevalent machine learning models for On-Target Activity are Azithum (Fusi Score) and the Vienna Bioscore which are both scores that measure On-Target Activity of CRISPR Knockouts [37, 100]. The primary difference between these models revolves around the features considered and included in the model. Specifically, Azithum utilised a

combination of various one-hot encoding of the gRNA sequence, melting temperature, GC features, cut location on the amino acid chain, and free energy to predict gRNA On-Target Activity score. dCas On-Target Activity models also utilise the same or most of the features that Azithum established as a baseline set of features that have to be included to assess On-Target Activity of gRNAs. While, the Vienna Bioscore extended the features used to include features that measure evolution conversation of amino acid sequences a give gRNA target. By including these feature gRNAs with a high Vienna Bioscore is able to typically induce a greater perturbation response than other On-Target Activity scores.

While the features utilized by these models often overlap, it's crucial to comprehend the target variable or what they define "On-Target Activity" as the normalized rank of a gRNA. A normalized gRNA rank is computed by transforming the original rank order from 1 to N into a value between 0 and 1, with 1 indicating the best-performing gRNA [37, 100]. This is the target variable employed by Azithum and the Vienna Bioscore. At first glance, this approach may seem reasonable, but it introduces certain biases. The foremost of these biases is that you're not directly predicting the perturbation strength of a gRNA but rather the strength of the target gene's response. This implies an implicit assumption that all essential genes exhibit an equal response, which is not necessarily the case [75]. Furthermore, this restricts the model's effectiveness to only essential genes and can result in reduced prediction accuracy for genes with distinct characteristics in terms of protein structure and conservation [100].

2.5.2 Measuring Off-Target Activity of a gRNA

The Off-Target Activity of gRNAs, also known as Guide Specificity, aims to quantify the level of unintended effects inherent in a given gRNA. In typical experiments, researchers are primarily concerned with the perturbation effects on the target gene rather than the broader impact of gRNA perturbation on one or more additional genes. Measuring Off-Target Activity lags behind On-Target Activity in terms of having a well-defined measurement method and a standardized scoring system. Nevertheless, this has not discouraged attempts to experimentally measure it and develop an Off-Target Activity "score" using computational models, which can broadly be categorized into two major types: alignment-based and ML score-based models [75, 88].

Detecting Off-Target Activity remains a persistent challenge, with a multitude of methods continually evolving to address this issue. However, a couple of major methods are typically employed, including CIRCLE-Seq, GUIDE-Seq, and Discover-Seq. Each of these methods employs distinct approaches to measure off-target effects. CIRCLE-Seq identifies off-targets by pinpointing cleavage sites in vitro, utilizing a restriction-enzyme strategy that generates circularized linear fragments [131]. On the other hand, GUIDE-Seq relies on detecting and

capturing oligodeoxynucleotide strands in Double-Strand Breaks (DSBs) introduced after CRISPR induces a double-strand break [132]. In contrast, Discover-Seq employs CHIP-Seq to identify off-target sites by recognizing MRE11 binding sites. MRE11, a key protein in DSB repair, stabilizes the DNA and facilitates the repair of double-strand breaks. Additionally, Discover-Seq has been extended by incorporating DNA-Pkc inhibitors, which impede DNA repair and enable the identification of Off-Target sites that are typically repaired relatively quickly [146].

Experimental methods for detecting Off-Target activity are typically used to generate a training dataset for building models. The initial set of models developed were alignment-based, relying on setting a threshold (N, the maximum number of allowed mismatches) and then identifying all potential gRNA binding sites within N mismatches of a given gRNA. While these models are not heavily relied upon, they are often used to generate limited datasets for ML score models [75]. The current state-of-the-art Off-Target Activity score is Evelation, an extension of the CFD score [88]. Evelation was trained using a combination of alignment and On-Target Activity metrics to predict Off-Target Activity. Initially, Off-Target sites are grouped into categories based on the number of mismatches (N), with Azithum used to score On-Target activity. Subsequently, the scores of all individual off-target sites are aggregated to create a new overall score, ranging from 0 to 1. This score represents a normalized rank of gRNAs in terms of the number of Off-Targets and the On-Target activity of these Off-Targets [88].

Off-Target activity remains a dynamic research area in gRNA library design, necessitating ongoing refinement and clarification of its definition. However, despite these challenges, there are straightforward methods for mitigating these limitations. One approach is to employ multiple gRNAs per gene and subsequently regress out the Off-Target effects. Although this approach increases the number of gRNAs, leading to larger experiments and higher costs, these drawbacks are relatively minor compared to the benefits of achieving an enhanced signal-to-noise ratio through the use of multiple gRNAs [88].

2.5.3 Designing a Custom gRNA Library

When designing a custom gRNA library targeting protein-coding regions, the first thing to consider is the perturbation strength which can be quantified by estimating the frameshift probability of the target site. The frameshift probability is the probability that a frameshift mutation occurs and is the key measure of on-target activity when conducting a screen in protein-coding regions [126, 5]. The frameshift probability of a given gRNA can be estimated utilizing machine learning models such as Indelphi, Lindel, or Forecast [126, 24, 5]. While these models were originally for humans and mice, recent studies in zebrafish have shown

that despite this, these models can still predict genotype outcomes at target sites in other species [103]. The target site with the highest frameshift mutation is the 'best' target site regarding on-target activity.

However, the off-target activity of a gRNA should always be accounted for and minimized by utilizing programs such as Flashfry to search for potential off-target sites quickly. Flashfry takes an organism's genome, identifies all of the target sites in the genome by their Protospacer Ajacent Motif (PAM) site and constructs an indexed database. Users can provide the target sequences, and it will identify all off-targets with up to a user-specified number of mismatches (typically, three mismatches are used) [98]. Next, for each predicted off-target, estimate the frameshift probability as a proxy of target site activity [126, 5]. Ideally, there are no off-targets for a sequence; this is rarely the case. As such, the gRNAs should be selected based on having off-targets exclusively in non-coding regions of the genome. Moreover, if this criterion is not met, they should have minimal frameshift probability. Additionally, suppose the organism is well-characterised, and it is possible to determine the gene type of the target gene (i.e. the target gene is a TF). In this scenario, an additional selection criterion should be considered when selecting gRNAs. If encountering an off-target is unavoidable, it is preferable for the off-target to be located within a gene that belongs to a different gene type compared to the target gene. Furthermore, the selected off-target gene should be unlikely to have a substantial effect on the biological process being studied when perturbed, thereby minimizing potential undesired impacts.

Designing the gRNA library for a CRISPR screen utilizing the dCas-based perturbation method is far more difficult as there are no well-established gRNA libraries. Additionally, there are no clear methods of measuring the on-target activity and any potential variance in its activity that may occur between target sites [61]. Despite this limitation, it is advisable to minimise potential off-target effects when selecting gRNAs. This can be achieved by identifying a gRNA that is free of off-targets or by minimizing their impact through the avoidance of similar phenomena. For instance, when employing CRISPRi to perturb enhancers in the genome and possessing knowledge of the chromatin accessibility within the specific biological system, one can opt for a gRNA with an off-target in heterochromatin regions of the genome if all gRNAs exhibit off-targets. This strategic choice aims to reduce the interference injected into the phenotypic readout. To facilitate this process, tools such as Flashfry can be employed to identify off-targets, while new models based on gRNA binding kinetics can assist in predicting binding activity [98, 41].

The importance and care of selecting the right gRNAs for a CRISPR screen can not be overstated, especially as the resolution of the phenotypic readout increases. In low-dimensional readouts such as growth-based assays, off-target effects must be strong and

perturb an essential gene. As such, perturbing other types of genes while having an effect is not as large of an issue as high-dimensional screens. As the resolution of the CRISPR screen increases, so does its sensitivity to the effects off-targets. As such, researchers should go to great lengths to minimize the effect off-targets can have. If it is impossible to have no off-target, then intentionally choose gRNAs whose off-targets are orthogonal to what is being studied, if possible, target unrelated safe loci.

2.6 Conclusion

In the past decade, CRISPR screens have rapidly developed and expanded what can be perturbed and the resolution of the phenotypic readout. Providing researchers with seemingly endless choices and options when investigating biological systems. While it is tempting to pick the most recent CRISPR technologies, this is not always the best approach for conducting an experiment. Here I went back to basics and reviewed the various reasons to conduct a CRISPR screen. What options were available for introducing perturbations to a biological system? What levels of resolutions are available, and their various advantages and disadvantages. Numerous CRISPR variants have been developed, providing increased flexibility in introducing perturbations in alternative ways and across protein-coding and non-coding regions of the genome. Moreover, higher-resolution phenotypic readouts are becoming increasingly prevalent. However, significant challenges persist in accurately determining the statistical power of these high-resolution screens. In the following chapter, I will present an approach to statistically simulating scCRISPR screens to explore various experimental designs and identify those that enhance the statistical power of the screen.

Chapter 3

Statistical Simulations of Single Cell CRISPR Screens

CRISPR-Cas9 is a revolutionary genetic tool with diverse applications in biological research, from genetic engineering to drug target discovery. However, the resolution of growth-based assays, the primary use case for CRISPR in a biological research setting, is limited. scCRISPR screens were developed to address these limitations, but current scCRISPR screens are chronically underpowered. To overcome these limitations, I developed crisprPower, a scCRISPR simulator capable of simulating multiple types of perturbation and propagate perturbations across Gene Regulatory Networks (GRNs). crisprPower can simulate various experimental designs and evaluate their statistical power, allowing researchers to optimize their experiments and improve the reliability and interoperability of their results.

3.1 Introduction

Current scCRISPR screens suffer from chronic underpowering [115]. The main reason for the limited statistical power stems from the use of single-cell sequencing technologies, which often produce data with extremely low coverage of the transcriptome or genome [108]. In addition, single-cell sequencing poses unique statistical challenges due to the change in the sampling process from sampling with replacement, as observed in bulk sequencing, to sampling without replacement [56]. This change in the sampling process creates "winner takes all" effects in single-cell datasets, where the highest expressed genes are the most likely to be sequenced. Furthermore, when a low- to medium-expressed transcript is observed, the probability of observing another transcript is decreased due to the effects of sampling without the replacement [56, 114, 123].

To understand and overcome the limitations of scCRISPR screens, I developed crisprPower, a statistical scCRISPR simulator capable of simulating cell populations, multiple types of CRISPR perturbations and propagating the effects of perturbations across these cell populations' GRNs. In addition, crisprPower can simulate alternative single-cell RNA sequencing methods of whole (normal) or targeted transcriptomes. crisprPower enables researchers to investigate the statistical power of various experimental designs for scCRISPR screens through simulations.

For a given cell population, a simulated GRN is generated through a multi-step process, where the core interactions between TFs are first sampled via modular sampling from a real GRN. Once a TF network is sampled, downstream effect genes (E-Genes) are added to the network via weighted sampling. The higher the out-degree of a TF, the greater the probability of an E-gene attached to it. This process allows crisprPower to create realistic GRNs containing both power-law and modular structures, as is frequently observed in biological networks.

Next, genes are parameterized by sampling their mean expression and dispersion either from a user-provided scRNA-seq dataset or a pre-fitted distribution. Additionally, adjustments to the mean expression are made based on the user-specified sequencing depth for cells. Finally, regulatory interactions are estimated using linear regression as a part of the crisprPower framework. This enables the propagation of perturbation effects throughout the network. Through the simulation of various experimental designs and subsequent analysis of the outcomes, researchers gain a deeper understanding of the factors influencing the power of their screens. This knowledge allows them to optimize their experiments to achieve more favourable outcomes.

In summary, crisprPower is a user-friendly and powerful tool that streamlines the process of designing and evaluating the statistical power of scCRISPR screens. It enables researchers to explore various experimental designs and assess the statistical power of their screens beforehand, thereby helping them to optimize their experiments and enhance the reliability and interpretability of their results.

3.2 Methods

3.2.1 Types of Genes Simulated in crisprPower

crispPower supports two broad gene categories: Transcription Factors (TFs) and Effect genes (E-genes). TFs are genes that directly regulate gene expression by promoting or inhibiting gene expression. TFs play a central role in controlling the dynamics of the GRN

and maintaining its 'homeostasis'. While E-genes are regulated by and respond to the effects of TFs, they are genes that are not directly involved in regulating gene expression but are affected by the expression of other genes. Within the category of E-genes, there are five broad subclasses based on their Gene Ontology (GO) terms: TFs, Receptors, Ligands, Kinases, and Other Genes (a catch-all category for other genes). These subclasses allow researchers to evaluate the power of screens on a more specific and refined set of gene types.

3.2.2 Sampling Gene Regulatory Network

A network is an abstract representation consisting of nodes and edges. In the context of GRNs, genes serve as nodes, while the edges between them (typically from a TFs to another gene) represent regulatory interactions. An adjacency matrix is a matrix containing entries of 1 and 0, indicating the presence or absence of an edge connecting node i to node j. In directed networks, the orientation of the matrix is significant. Specifically, I define an edge as going from node i to node j if there is a 1 entry at row i, column j. Node degree refers to the number of edges connecting a specific node in a directed network, such as a GRN. Nodes in GRNs can have both incoming and outgoing degrees, corresponding to the number of edges directed towards or away from the node in question. When referring to the degree of a node, particularly a TF, I specifically focus on its out-degree.

The first step in constructing a simulated GRN involves sampling a core TF regulatory network using a greedy sampling algorithm known as modular sampling, which was initially developed by Marbach et al. 2009. This method, known as modular sampling, incrementally expands a subnetwork of TFs to achieve a desired size, starting from a seed node. The seed node can be randomly or manually selected from a source GRN. Prior to generating the simulated GRN, I sample from a collection of authentic GRNs obtained from GRNdb [42]. These real GRNs serve as the source network for sampling the core TF regulatory network in crisprPower.

Using the same method as utilized by Marbach et al. 2009for calculating graph modularity, I am also adopting their rationale for defining and utilizing modularity. First, modularity, as defined by Marbach et al., refers to the difference between the number of edges within a subnetwork and the expected number of edges in a randomized network. The goal of using modularity to sample nodes is to extract diverse sets of subnetworks with a user-specified size while maintaining reasonably high modularity. The formula for calculating graph modularity (Q) developed by Marbach et al. 2009is:

$$Q = \frac{1}{4m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \tag{3.1}$$

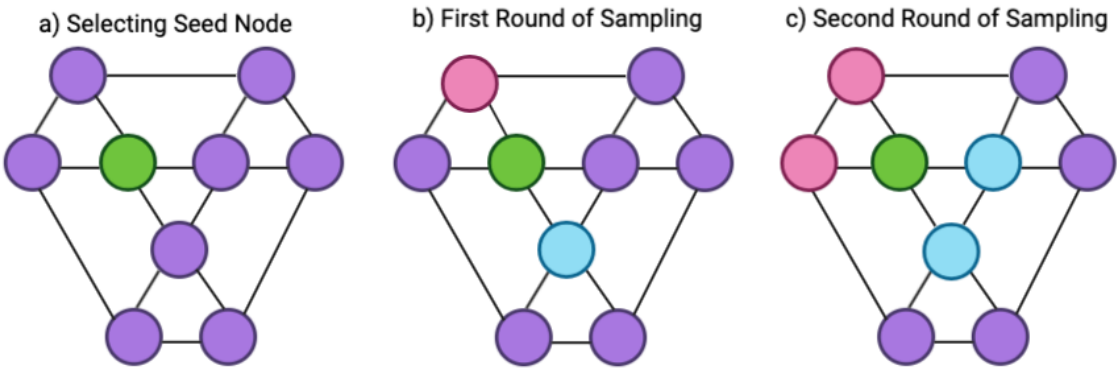


Fig. 3.1 This example demonstrates the evolution of modular sampling in successive rounds, focusing on a 3-node subnetwork. a) The whole network is displayed, and the sampled seed node is highlighted in green. b) Two potential directions of modular sampling are depicted, with light blue nodes representing one path and light pink nodes representing another. c) The final step of modular sampling illustrates the preference for nodes that increase the number of edges in the subnetwork.

The modularity of a subnetwork is calculated based on the adjacency matrix (A) of the source network and the total number of edges (m) in the source network. Here, A_{ij} represents an entry in the adjacency matrix, where $A_{ij} = 1$ indicates a directed edge from node i to node j. The degrees of nodes i and j in the source network are represented by k_i and k_j, respectively. To determine whether a given node is included in the subnetwork during a given calculation of modularity, we use the variable c. c_i and c_j represent whether node i and node j, respectively, are part of the subnetwork during this calculation of modularity. To evaluate if nodes i and j belong to the subnetwork, the function $\delta(c_i, c_j)$ is used, which returns 1 if both nodes are in the subnetwork and -1 otherwise.

To add a node to the sampled subnetwork, first, identify all neighbours of the node in the source network. A neighbour is defined as a node that is directly connected to at least one node in the current core TF regulatory network. Next, calculate the modularity (Q) of the subnetwork if a particular neighbour was to be added to the subnetwork, repeating this process for all neighbours. Then select the neighbouring node that increases the modularity of the subnetwork the most. If multiple neighbours result in the same modularity, choose one randomly and add it to the subnetwork. Finally, Repeat this process until the desired number of nodes is sampled. By utilizing modular sampling to construct the core TF regulatory network, crisprPower captures the modular regulatory structures observed in real GRNs, resulting in a simulated GRN that reflects the structure and function of a real GRN. The sampled core TF regulatory network is a subset of the source GRN utilized during modular

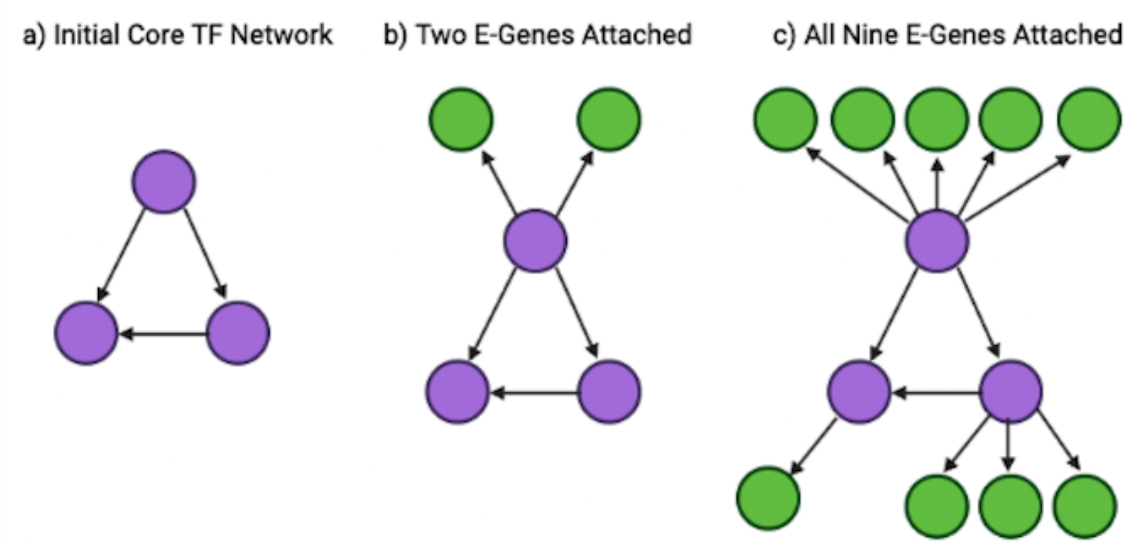


Fig. 3.2 The figure showcases the potential evolution of preferential attachment after sampling a core TF regulatory network. E-genes are attached preferentially based on the out-degree of the TFs. The illustration consists of three panels: a) The initial TF regulatory network. b) The first few rounds of preferential attachment of E-Genes. c) All nine E-genes are attached, with TFs possessing a higher out-degree receiving the majority of E-Genes.

sampling. This means that the edges connecting the TFs in the regulatory network remain consistent with those observed in the source GRN.

The second step of constructing a simulated GRN is the attachment of E-genes to the core TF regulatory network. crisprPower allows users to specify both the number of E-genes to be attached and the maximum number of regulating TFs for an E-gene (default is set to 3). To do this for each E-gene, crisprPower first samples the number of regulating TFs (t). Then it samples from the list of TFs t times using a weighted sampling without replacement procedure; the out-degree of each TF as its weight (see figure 3.2). The weighted sampling of TFs allows crisprPower to simulate the preferential attachment dynamics that are observed in real GRNs where genes are more likely to attach themselves to highly connected TFs than poorly connected TFs [11].

The utilization of a two-step process involving modular sampling to identify core TF regulatory elements and weighted sampling to attach E-genes to highly connected TFs preferentially enables crisprPower to preserve the observed modular and hierarchical structures found in real GRNs [11, 95, 119]. The capture of the GRN structure is vital for simulating how perturbations propagate throughout an entire GRN. This feature is particularly important when studying the impact on and off-target activities have on the noise profile of scCRISPR screens and the subsequent impact on the screens' results. By accurately reflecting the

modular and hierarchical structures of the GRNs, crisprPower ensures the reliable generation of representative simulated benchmarks for real biological systems (see Figure 3.1). For the complete pseudo-code implementation, please refer to Algorithm 1.

Algorithm 1: Sampling Sythentic GRNs

```
Input: Source GRN, Number of TFs, Number of E-genes, Maximum TFs per E-gene
Output: simulated GRN
Initialise Core TF Regulatory Network by selecting a node randomly or manually;
while Size of subnetwork < Number of TFs to sample do
    Identify all neighbours of the subnetwork in the source GRN;
    foreach Neighbour do
        Add the neighbour to a temporary subnetwork;
        Compute the modularity Q of the temporary subnetwork;
    end
    Select the neighbour that maximizes the modularity Q;
    Add the selected neighbour to the subnetwork;
end
Attach E-Genes to Core TF Regulatory network via weighted sampling;
foreach E-gene do
    Initialize an empty set of regulating TFs;
    for i from 1 to t do
        Sample a TF from the list of TFs using weighted sampling with replacement;
        Add the sampled TF to the set of regulating TFs;
    end
    Attach the set of regulating TFs to the E-gene in the simulated GRN;
end
return simulated GRN
```

3.2.3 Modelling Gene Expression

Parametric Assumption of Gene Expression Model

crisprPower utilizes a Negative Binomial (NB) distribution to model gene expression within a specific cell population. NBs have a long history of being used for modelling biological count data [121, 90, 116]. The NB distribution is commonly used for analyzing biological count data due to its flexibility and is equivalent to the gamma-poisson compound distribution, which combines a gamma component to capture the variability in the mean count and a Poisson component to model the random fluctuations around that mean. This compound distribution provides a flexible framework to account for a wide range of possible distributions

when analyzing gene expression or other types of biological counts [116]. The gamma-poisson model accommodates both under-dispersed and over-dispersed count data, which are frequently encountered in biological studies [90].

crisprPower utilises a particular interpretation of NBs as a steady-state distribution of a bursting transcriptional kinetic model of gene i, which was shown by Amrhein et al. that an NB is the steady-state expression distribution of bursting transcription model. Under this interpretation, it has been shown that assuming a steady state leads to a bursting transcription model, where the NB is parameterized as $NB(r_i, p_i)$, with $r_i = \frac{\text{burst}_i}{\text{deg}_i}$ representing the rate of transcription bursting over degradation and $p_i = (1 + \text{burst size}_g)^{-1}$ parameterized by the expected burst size for a given gene. In R, an NB is fitted using the 'ecological' parameterisation where μ and θ (which is the dispersion parameter) are estimated. To utilise the bursting model interpretation of NB, I need to convert μ and θ to r and p. This can be done in the following manner:

$$p_i = \frac{\theta_i}{\theta_i + \mu_i} \tag{3.2}$$

$$r_i = \theta_i \tag{3.3}$$

By making this assumption and interpreting the NB in this manner, I gain a clear set of rules for how gene expression distributions change due to alternations in regulatory interactions or are perturbed. Studies on transcriptional kinetics have shown that TFs do not appear to regulate or alter the burst size of a given gene; instead, they primarily influence the bursting rate through complex interactions with enhancers and promoters [78, 92, 93]. Therefore, the propagation of perturbations and the regulatory influence of TFs on mean expression must be interpreted in terms of modifying the burst rate of a given gene while the burst size remains constant.

Fitting Genes Mean and Dispersion

When a user provides a scRNA-Seq reference dataset of a cell line or biological setting (i.e. tissue), crisprPower utilizes this data to generate a half-normal sampling distribution for each gene class and fit a mean-dispersion curve. To accomplish this, crisprPower follows a series of steps. First, it takes a count matrix X from a single-cell dataset and normalizes it using the 'poscount' library size normalization method from the DESeq package [90]. Next, it empirically estimates the mean and variance in expression directly from the normalized counts and uses them to estimate the gene expression dispersion. Then a mean-dispersion curve is fitted using Maximum Likelihood Estimation (MLE) to estimate the asymmetric

dispersion (aDisp) and extra Poisson (ePois) properties from the observed data. This approach enables me to model dispersion as a parameter dependent on a gene's mean expression [121]. The estimated aDisps and ePois enable the estimation of gene dispersion given its mean, allowing crisprPower to sample mean expression and calculate the corresponding dispersion based on the properties of the mean-dispersion curve. This specific use case was originally developed by Schmid et al.. For a detailed implementation of each step, refer to Algorithm 2.

Once crisprPower has estimated the mean expression and fitted the mean-dispersion curve of a given dataset, a half-normal distribution is then estimated using MLE. The half-normal distribution was chosen over a gamma distribution because, after 100 random sampling and fitting mean expressions across all genes, it had the consistently lowest Akaike Information Criterion score. A half-normal distribution is a probability distribution that resembles a normal distribution with $\mu = 0$ and is truncated at zero, so $x >= 0$. It only takes positive values, with a higher concentration of values closer to zero and decreasing density as values increase. A mean expression is first sampled from a half-normal distribution for a given gene using the following notation: $\mu_g \sim \mathcal{N}^+(\mu, \sigma)$. Then, the gene dispersion can be estimated as $\text{disps}_g = \frac{\text{aDisp}+\text{ePois}}{\mu_g}$, as proposed by Schmid et al.. By default, crisprPower uses a reference sampling distribution and mean-dispersion curve fitted using the Human Cell Atlas on ICA Bone Marrow Dataset [59].

Estimating Target Panel Mean Expression

Targeted Panels represent a promising technology in single-cell sequencing, aiming to address the sparsity issues encountered in scCRISPR screens [114, 123]. However, due to the scarcity of single-cell targeted panel data, it is not possible to establish a half-normal sampling distribution for targeted panels. Recognising that targeted panels primarily influence the observed mean expression distribution, I chose to predict the mean expression of a specific gene in the targeted panel based on its mean expression in the whole transcriptome, given the scarcity of targeted panel data. Mechanistically generating targeted panel data is currently not supported by any existing models. Consequently, I had to infer the mean expression of targeted panels using paired target panel and whole transcriptome datasets from Schraivogel et al.. To achieve this, I trained a linear regression model on the log-transformed mean expression of matched targeted panel and whole transcriptome data from the 11000 gene sets, obtained from Schraivogel et al.. With this regression model, it becomes feasible to forecast a gene's mean expression in the targeted panel based on its mean expression in the whole transcriptome.

Algorithm 2: Estimate Mean-Dispersion Curve

Input: X with i genes by j cells
Result: ePois, aDisp

normalise X using 'poscount' from Love et al.;
calculate row-wise mean and variance of count matrix;
$\mu_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}$, $\sigma_i^2 = \frac{\sum_{j=1}^{n}(x_{ij}-\mu_g)^2}{n-1}$;
calculate gene dispersion using mean and variance in a row-wise manner;
$disps[i] = \frac{\sigma_i^2}{\mu_i}$;
Initialize coefficients $ePois = 0.1$, $aDisp = 1$;
for $i \leftarrow 1$ **to** 10 **do**
    for $i \leftarrow 1$ **to** n **do**
        Calculate residuals: $residuals[i] \leftarrow \frac{disps[i]}{coefs[1]+\frac{coefs[2]}{means[i]}}$;
        if $(residuals[i] > 1 \times 10^{-4})$ ***and*** $(residuals[i] < 15)$ **then**
            Add i to the set of good indices: $y \leftarrow y \cup \{i\}$;
        end
    end
    ePois, aDisp $\leftarrow$ glm($y \sim \mathrm{I}(1/x)$, family = Gamma($link =$ "$identity$"), start = coefs)
    if $ePois < 0$ *or* $aDisp < 0$ **then**
        Parametric dispersion fit failed;
    end
end
return ePois, aDisp;

Simulating Mean Expression based on Excepted Cell Library Size

To allow users to simulate different Expected Cell Library Sizes, I integrated an NB General Linear Regression (GLM) model, originally employed by Hafemeister and Satija. In their study, they established a linear relationship between the mean expression of genes ($E[x_{ij}]$) of gene i in cell j and the log-transformed cell library size (m_j) of a given dataset. This was achieved using a base 10 logarithmic transformation, where the y-intercept is estimated as β_0, and the slope between cell library size and gene mean expression is β_1 (as shown in Equation 3.4). In addition to fitting the crisprPower expression distributions to a given dataset, an NB GLM model is also fitted to model the relationship between mean expression and cell library size. This allows crisprPower to predict a new mean expression for a specific gene based on a user-provided cell library size. Once the new mean expression is estimated, crisprPower estimates a new gene dispersion using the previously fitted mean-dispersion curve.

$$\log_{10}(E[x_{ij}]) = \beta_0 + \beta_1 \times \log_{10}(m_j) \tag{3.4}$$

3.2.4 Simulating CRISPR Perturbations

CRISPR Knockouts

CRISPR knockouts' perturbation effects upon a given gene's expression are simulated using a Zero-Inflated NB (ZINB) distribution, where the probability of a zero occurrence is determined by the gRNA's on-target activity at a specific target site [126, 5]. The ZINB distribution is chosen because it effectively captures the probabilistic nature of knockouts, as the occurrence of a knockout for an individual gene in a given cell is random. Knockouts can be mathematically described as a zero-inflation process, where either a gene is knocked out and expresses zero transcripts or it retains the capability to express transcripts as usual without any changes to the underlying distribution.

Let p_{zero} represent the probability of zero occurring in the ZINB distribution. The on-target activity of the gRNA at the target site determines this probability, denoted as a. Hence, p_{zero} is a function of a, i.e., $p_{zero} = f(a)$. The remaining non-zero counts in the ZINB distribution follow an NB distribution, which accounts for the variability in gene expression due to knockouts.

Overall, the CRISPR knockout perturbation effects can be mathematically described as:

$$\text{Probability of Knockout} = \begin{cases} 0, & \text{with probability } p_{zero} \\ NB(r_g, p_g), & \text{with probability } (1 - p_{zero}), \end{cases} \tag{3.5}$$

Where the NB distribution models the non-zero expression counts in the presence of a knockout.

CRISPR Interference

CRISRPi introduces a deterministic perturbation by inducing a heterochromatin state via its transfused repressor domain at the TSS of a target gene; as such, the perturbation being introduced is fundamentally different from CRISPRko and relies upon reducing the gene burst rate [68, 7]. The reduction in bursting rate is directly linked to the binding probability of a gRNA at a particular target site, determined by the gRNA's on-target activity. A higher binding probability results in a more significant decrease in the gene's bursting rate.

When analyzing a perturbed gene, we can analytically calculate its effect on the gene expression distribution as $r_{gp} = r_{gc}(1 - a_g)$, where r_{gp} is the new perturbed r_g parameter, r_{gc}

is the original control r_g parameter, and a_g represents the on-target activity of a specific gRNA at the target site. The parameter p_g remains unchanged due to the parametric assumption of the assumed gene expression model CRISPRi does not alter the burst size. Instead, it functions by obstructing polymerase access, thereby reducing the burst rate. These values correspond to the new NB distribution's estimated mean and dispersion parameters for simulating the CRISPR interference scenario affecting the given gene.

3.2.5 Parameterize Regulatory Interactions of Transcription Factors

The regulatory effect of a given TF upon a given gene is estimated in the following manner as seen in Algorithm 3. First, the mean gene expression of the regulated gene and the TFs involved in regulation was obtained. Subsequently, the noise was introduced into the data by sampling from a normal distribution. Where the distribution's mean was set to the gene's mean expression, while the standard deviation was determined by a noise coefficient (default value is 0.001). Multiple samples (default value is 100) were drawn from the distribution, generating noisy mean expression values for both the regulated gene and the TFs. Third, linear regression was utilised to determine the strength of the regulatory interaction between the regulated gene and each TF. The β coefficients were estimated for each regulator, indicating the magnitude and direction of the regulatory effect. These coefficients could be positive or negative, signifying an activating or repressive influence of the TF on the regulated gene, respectively. Finally, a regulatory coefficient matrix is created, consisting of G rows and T columns, where each row represents a gene and each column represents a TF. Each value in the matrix contains a β coefficient value. If no regulation occurs, the index is filled with a zero.

3.2.6 Propagation of Perturbations through the GRN

Simulating the effect of perturbing a given gene relies upon crisprPowers ability to propagate the perturbation effect throughout the GRN in a realistic manner. To do this, I developed a modified version of the network signal propagation algorithm originally purposed by Kamimoto et al. (see Algorithm 4), where the effect of a perturbation is represented as the iterative updating of mean expression of downstream genes. The GRN is represented by two matrices, X and C. Matrix X is a $G \times T$ matrix, where G represents the number of genes and T represents the number of TFs in the GRN. Each element (g,t) of the matrix contains the mean expression of the TF or zero. If a TF does not regulate a particular gene, its value is set to zero. Matrix C is also a $G \times T$ matrix, where each element (g,t) represents the β

Algorithm 3: Parameterising Regulatory Interactions of GRN

Input: μ_g Gene Mean Expression, μ_t TFs Mean Expression, ε Noise Parameter
Result: Regulatory coefficient matrix C

```
foreach Gene do
    sample noise means for both regulated gene and TFs involved in regulation;
    y ~ N(μ_g, ε) ;
    foreach TF do
        X ~ N(μ_t, ε)
    end
    Perform linear regression to determine regulatory interaction strength;
    y = Xβ + e
end
Create a regulatory coefficient matrix C with dimensions G × T;
foreach gene do
    foreach TF do
        if regulation occurs then
            Set the C_gt to the corresponding β coefficient of TF;
        end
        else
            Set C_gt to zero;
        end
    end
end
```

coefficient of the TF or zero. Similarly, if a TF does not regulate a given gene, its value is set to zero. The β coefficients are estimated through linear regression in the preceding step.

To propagate a perturbation throughout the GRN, the algorithm iterates N times which will propagate the perturbation N degrees from the perturbed node (by default, N is set to 5). During each iteration, element-wise multiplication of matrices X and C is performed, representing the influence of TFs on gene expression. Subsequently, the mean expressions of TFs are updated based on any changes in the network. Row sums are computed for the rows of matrix X corresponding to TFs, and the non-zero column values in matrix X are updated with the new TF mean expression.

After completing the propagation loop, a final element-wise multiplication of matrices X and C is performed. The resulting matrix is then subjected to row summation, yielding the new mean expression μ_g of genes in the GRN. Bursting rate parameters r_g for each gene are calculated using the formula: $r_g = \frac{\mu_g \cdot p_g}{1-p_g}$, where p_g denotes the probability of bursting for the gene.

Algorithm 4: Propagation Perturbation across Simulated GRN

Input: X, C, N
Output: μ_g

for $n \leftarrow 1$ **to** N **do**
  Perform element-wise multiplication: $X \leftarrow X \odot C$;
  Calculate row sums for TFs: $S \leftarrow \text{RowSum}(X)$;
  Update non-zero column values in X with new TF mean expression;
  Perform element-wise multiplication: $X \leftarrow X \odot C$;
end
Calculate row sums for genes: $R \leftarrow \text{RowSum}(X)$;
$\mu_g \leftarrow R$;
Calculate bursting rate parameters: $r_g \leftarrow \frac{\mu_g \cdot p_g}{1-p_g}$;
return μ_g, r_g;

3.3 Results

3.3.1 Validating Statistical Characteristics of Gene Regulatory Network

Power-law dynamics are frequently observed throughout biological networks, from protein-protein interactions to metabolic networks. A common feature that defines power-law dynamics is the emergence of 'hubs' or central nodes with a disproportionate number of edges in comparison to other nodes in the network [11]. In GRNs, essential or cell state essential TFs frequently form hubs that play a greater-than-expected role in maintaining the regulatory state of the GRN in a given cell type. Networks that are controlled by these dynamics exhibit 'small world' effects where due to the network's hubs, most nodes are only a few degrees away from the furthermost node relative to its position [11, 19]. In biological networks, 'small-world' effects allow for the rapid response to environmental stimuli. In addition, in biological networks, power laws govern the existence of network motifs that carry significant functional implications for the network. These motifs introduce regulatory logic and redundancy, both of which are advantageous properties that experience positive selection during evolution. [11, 38, 6]. As such crisprPowers' ability to recapitulate these networks accurately is crucial for ensuring realistic simulations of CRISPR perturbations.

Accurately representing the modularity observed in biological networks is crucial for simulating perturbations realistically. To recreate this property, a modular sampling algorithm was implemented using GeneNetWeaver [119, 95]. This algorithm randomly samples a seed node from a true GRN with only TF-to-TF interactions. It iteratively adds nodes to the network based on which neighbours increase its modularity. To validate this, I utilized crisprPower to simulate GRNs of various sizes. Subsequently, I applied the Louvain graph

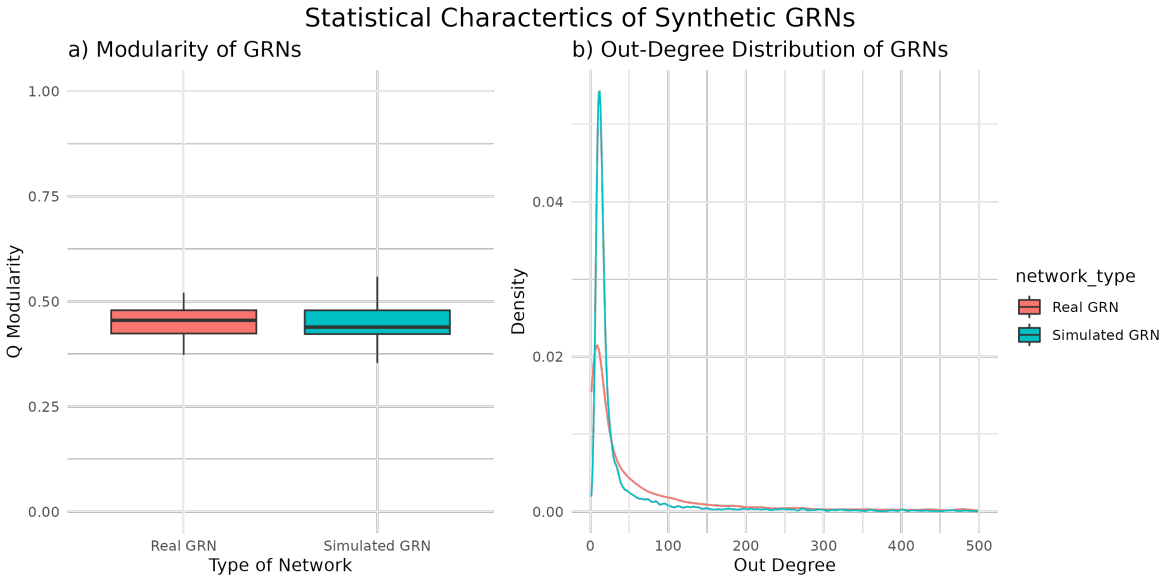


Fig. 3.3 Confirming that CrisprPower simulated GRNs exhibit similar properties of modularity observed in real GRNs.

clustering algorithm to identify communities within GRNs [15]. I then calculated the graph modularity using these communities. Furthermore, I repeated this process with real GRNs from the GRNdb for comparison [42]. Simulated GRNs generated through this process have near-identical modularity to real GRNs, with simulated GRNs having a slightly wider range and lower mean in terms of modularity see figure 3.3a. Such minimal difference in modularity indicates that crisprPower can accurately simulate GRN modularity observed in real GRNs.

While modularity allows crisprPower to replicate network motifs and redundancy within a regulatory network, power-law distributions govern the number of genes regulated by a given TF, thereby controlling the scale of first-degree perturbation effects. To assess crisprPowers simulated GRNs power-law dynamics, I simulated a hundred networks with a max number of 400 TFs, 5000 E-Genes, and 5 regulating TFs max per E-gene in order to be on par with the size of real GRNs. I then plotted the Kernel Density Estimate (KDE) of the out-degree distribution of both real and simulated GRNs in figure 3.3. While both distributions exhibit power law dynamics in their out-degree distributions, it is clear that crisprPowers differ from real GRNs. Specifically, simulated GRNs had a greater density of TFs with out-degrees between 0-50 than real GRNs. While, in the 50-200 out-degree range, simulated GRNs had fewer TFs within this out-degree range than excepted. It is not until we got to TFs of 200 out-degrees or more in the right tail of the out-degree distribution we see similar behaviour between simulated and real GRNs. This general shift to the left while exhibiting power law

characteristics differs from observed real GRNs. As such, crisprPower can simulate power law dynamics, but it differs from those observed in real GRNs.

Comparing the statistical characteristics of simulated and real GRNs showed that crisprPower can simulate characteristics of real GRNs doing particularly well with simulating modularity. While capable of simulating power-law dynamics, the simulated networks do not perfectly match the observed power-law dynamics of the source networks. Instead, the simulated networks exhibit a slightly lower power-law dynamic, which results in a smaller right tail in the power-law distribution (see Figure 3.3b). This discrepancy indicates that further refinement is needed in terms of simulating preferential attachment. These differences in power law will result in structural differences that impact crisprPowers' ability to simulate how a perturbation is propagated through a network. Specifically, crisprPower simulated GRN will contain many of the functional motifs and redundancies observed in the core regulatory network of TFs and other biological networks but will differ in terms of the perturbation scale when TFs are perturbed, which could artificially inflate or deflate the perceived importance of TF.

3.3.2 Validating CRISPR Perturbations

scCRISPR screens primarily use one of two CRISPR perturbation methods: Knockout or Interference. Each method has pros and cons, requiring their own independent model to simulate the perturbations' immediate effect upon a target gene expression distribution.

Knockouts provide a theoretically more permanent perturbation by inducing a frameshift relying upon the NHEJ pathway to induce these mutations [126]. To simulate this process, CrisprPower utilizes a ZINB where the on-target activity $f(a_g)$ is assumed to be a conservative estimate of the probability of zeros $p_{zero} = f(a_g)$ for a given target gene. At the same time, the NB component of the ZINB represents the cells where a knockout failed to occur due to the probabilistic nature of the mutations in the NHEJ pathway and is sampled $1 - p_g$. I plotted the scaled gene expression Probability Mass Function (PMF) of five TFs across four simulated on-target activities of 0, 0.25, 0.5, and 0.75 to visualise how Knockouts affect a target gene distribution. As seen in figure 3.4, as the on-target activity increases, we don't observe a shift in the gene expression distribution but rather an overall decrease in the probability of sampling from the original gene expression distribution, which corresponds with an increase in the number of zeros.

Interference perturbs a given target gene deterministically by inducing a heterochromatin state at the TSS via the transfused repressor domain [68]. Due to the differences underlying the perturbation mechanism crisprPower simulates the direct effect of interference as a modulation of the target genes' transcription burst rate r_g. Where the on-target activity a_g is

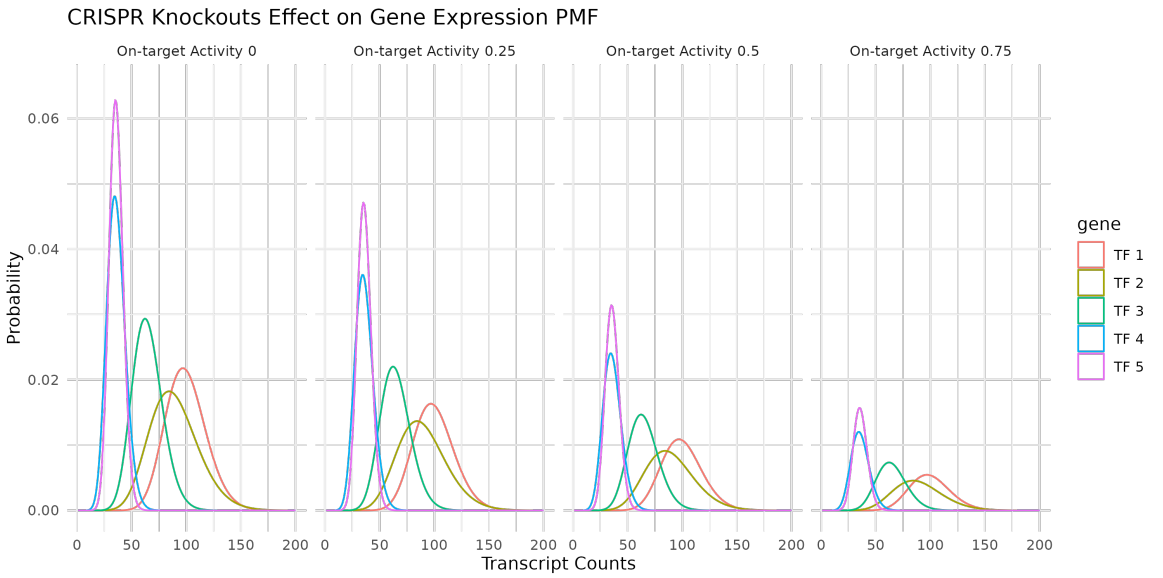


Fig. 3.4 Showing CRISPR Knockout perturbation effects on target gene expression distribution as on-target activity increases. From left to right, the simulated on-target activity is 0, 0.25, 0.5, and 0.75. As observed in the distribution above, Knockouts do not fundamentally alter the distribution of a gene expression; instead, it simply inflates the number of observed zeros.

assumed to represent that reduction of the burst rate in the following manner $r_g = r_g(1 - a_g)$. To visualise the direct effects Interference has upon a target gene expression distribution, I plotted the PMF under varying on-target activity of 0, 0.25, 0.5, 0.75 as seen in figure 3.5 as the on-target activity of an Interference based perturbation increase there is an observable shift in the gene expression distribution continuously to the left, closer to zero, with an ever greater constraint on the variance. Differing of Knockouts, these counts should be observed in all cells with a given gRNA. They should be easier to identify and infer the on-target activity of Interference compared to Knockouts.

crisprPower is a powerful tool that can simulate perturbations induced by either knockouts or interference. As illustrated above, the models employed to simulate these perturbations showcase distinct characteristics with increasing on-target activity. These dissimilarities are attributed to underlying mechanistic variations that crisprPower effectively captures. Moreover, crisprPower has the ability to model and analyze both the processes involved, as well as the disparities in statistical characteristics between these two methods.

Fig. 3.5 Showing CRISPR interference perturbation effects on target gene expression distribution as on-target activity increases. From left to right, the simulated on-target activity is 0, 0.25, 0.5, and 0.75. As shown in the distribution above, interference shifts an expression distribution left while constraining variance until it becomes zero.

3.3.3 Validating Propagation of Perturbations

I demonstrated interference and knockout effects on target gene expression distribution in the previous section. However, this is the immediate perturbation; the observed scale of perturbing a target gene will vary depending upon the genes it regulates downstream. To simulate these effects, crisprPower must be able to effectively propagate the effects of the perturbation throughout the network based on regulatory interactions and the strength of these interactions. I simulated a small GRN with 10 TFs and 300 E-Genes to validate this. For visualizing the effects, I specifically plotted three E-Genes that exhibit the strongest activating regulatory influence (indicated by the largest positive β coefficient) by a given transcription factor (TF). I conducted a 'perfect' perturbation of three TFs based on the node's centrality and propagated their perturbation across the GRN (Fig. 3.6). The three chosen TFs represent varying levels of centrality within the GRN. Consequently, they exhibit different perturbation scales, which can be conceptualized in relation to the number of downstream edges away from the target gene that can be influenced. I term this phenomenon "perturbation degree," wherein the number of degrees indicates the maximum number of edges through which the perturbation on a specific target gene can be observed in terms of differential expression. For instance, a target gene with zero-degree perturbations refers to a gene where only its expression distribution is altered when perturbed. While a one-degree

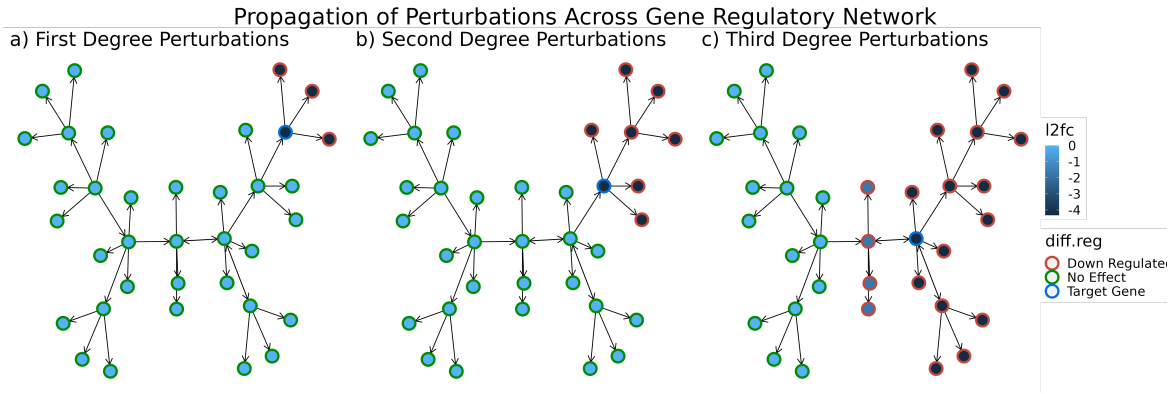


Fig. 3.6 llustrating the propagation of perturbations across a GRN based on the centrality of TFs and their impact on downstream E-Genes' expression distribution. a) Perturbing an isolated and less influential TF demonstrates limited downstream effects. b) A moderately important TF immediately induces perturbations in its regulated genes. c) Perturbing a central TF results in widespread differential expression throughout most of the GRN, showcasing the extensive influence of highly central TFs.

perturbation occurs when both the target gene and its immediately downstream regulated genes exhibit differential expression. As the number of perturbation degrees increases, the scale of perturbation expands across the GRN leading to a greater number of genes being differentially regulated as a consequence.

The simulation of 'perfect' perturbations showcases the capability of crisprPower to simulate both the perturbation of a target gene and its propagation across a GRN. In Figure 3.6, from left to right, we observe how a perturbed TF (highlighted by a red circle) immediately exhibits differential expression, subsequently perturbing the genes it regulates. Furthermore, as more central TFs are perturbed, the scale of the perturbation increases.

In Figure 3.6a, we can observe that when an edge TF is perturbed, it solely affects the genes it directly regulates. However, as we move closer to central TFs, an increasing number of downstream genes exhibit differential expression. Additionally, Figure 3.6c provides an example of how the modularity within GRNs influences the propagation of perturbations. The presence of a collider feature in the GRN causes perturbation of the central TF to impact only the right side of the GRN, while the left side shows limited to no perturbation at all.

3.4 Discussion

scCRISPR screens are a novel experimental method that provides researchers with a high-throughput and high-resolution view of cellular perturbations [1, 36]. With such a detailed view of cellular systems being perturbed, the quest to map phenotype to genotype has never

been more exciting or feasible. While significant effort has gone into improving the experimental protocol of scCRISPR screens, such as moving away from guide barcodes to direct readout of gRNAs in cells, and new single-cell sequencing methods are being developed, there remain fundamental questions about the experimental design of scCRISPR screens and, more importantly, the statistical power these screens possess [114, 101, 121]. crisprPower was developed to address these questions. crisprPower is a statistical simulator capable of simulating different experimental protocols, such as Whole or Targeted transcriptome, Interference or Knockout and how these perturbations propagate across a GRN network.

To validate crisprPower's performance, I first investigated the statistical characteristics of the simulated GRNs it generates and compared them to real GRNs. The two key characteristics of biological networks are modularity and power-law dynamics [6, 95]. Modularity is a crucial aspect of biological networks, accounting for the regulatory logic and redundancy observed in biological networks. To confirm that I observed the same characteristics in the simulated GRNs used in simulations, I calculated and compared their modularity to that observed in real GRNs. crisprPower's simulated GRNs showed similar modularity to that observed in real GRNs. However, when it came to simulating power-law dynamics, the simulated networks exhibited slightly different types of power-law behaviour than real networks. This suggests that further refinement is necessary to simulate preferential attachment in crisprPower accurately. These differences in power-law dynamics can lead to structural disparities, affecting crisprPower's ability to simulate how perturbations propagate through the network accurately. In particular, the simulated GRNs generated by crisprPower contain many of the structural motifs observed in GRN and other biological networks. However, the perturbation scale of TFs in the simulated GRNs differs from that observed in real networks, potentially leading to artificial inflation of the importance of TFs.

Next, I explored crisprPower's ability to generate Knockout or Interference perturbations as a function of on-target activity by confirming the expected shift in the target genes expression distributions. I simulated Knockouts and Interferences using two different mathematical models due to the differences in mechanisms inducing the perturbations. Knockouts induce permanent perturbations by causing frameshift mutations through the NHEJ pathway. crisprPower employs a ZINB model to simulate this process, where the on-target activity represents the probability of observing zeros in the target gene expression. This is reflective of the probabilistic nature of CRISPR Knockouts. A target gene is deterministically perturbed for interference by inducing a heterochromatin state at the TSS through a transfused repressor domain. crisprPower models the direct effect of interference by modulating the target gene's transcription burst rate. To validate that crisprPower can simulate these differences, I plotted the PMF of multiple increasing on-target activities for both perturbation methods. I observed

distinct but excepted shifts in target gene expression based upon the perturbation method, with both methods inducing greater differential expression as on-target activity increased.

Finally, I demonstrated crisprPower ability to simulate the perturbation of both a target gene and its propagation across a GRN. The simulations of 'perfect' perturbations showcased the capability of crisprPower to accurately model the immediate and downstream effects of perturbing TFs within the GRN. The results depicted in Figure 3.6 provide valuable insights into the propagation patterns of perturbations throughout the network. It is evident that the perturbation scale varies depending on the centrality of the perturbed TFs. When a less central TF is perturbed, the differential expression primarily affects the genes it directly regulates. However, as we move towards more central TFs, more downstream genes exhibit differential expression, indicating a broader impact. This phenomenon of perturbation degree, denoting the number of degrees of separation from the target gene that experiences differential expression, highlights the interconnected nature of the GRN and the potential for cascading effects. It is worth noting that the modularity within the GRN also plays a role in the propagation of perturbations. The presence of a collider feature, as observed in Figure 3.6c, can confine the perturbation to specific regions of the GRN while leaving other regions unaffected. This underscores the influence of the GRN's structural characteristics on the propagation of perturbations and raises important considerations regarding off-target effects.

In conclusion, scCRISPR screens have given researchers an unprecedented high-resolution view of cellular perturbations. crisprPower is a novel statistical simulator that enables researchers to explore the statistical power of these screens. By generating realistic GRNs that exhibit both modularity and power-law dynamics, crisprPower can simulate CRISPR perturbations such as Knockout and Interference and propagate these perturbations across the GRN. This simulator allows researchers to search for optimal experimental designs that maximize the statistical power of their scCRISPR screens. The validation studies show that crisprPower can generate realistic perturbation responses, allowing researchers to map phenotype to genotype more accurately. In summary, the crisprPower simulator provides researchers with a valuable tool to optimize the statistical power of scCRISPR screens by simulating realistic perturbations and establishing a ground truth for the development of methods to accurately map genotype to phenotype.

Chapter 4

Experimental Design of Single Cell CRISPR Screens

Single-cell CRISPR screens offer a high-throughput and high-resolution tool for researchers to perturb biological systems, allowing them to dissect and reconstruct regulatory interactions of biological systems. However, despite their tremendous promise, current scCRISPR screens suffer from a lack of statistical power. Using the crisprPower simulator, I investigated and compared how the statistical power of a scCRISPR screen varies based on alternative experimental design choices. While current scCRISPR suffers from a lack of power, simple changes such as using a targeted panel and CRISPR Knockouts can dramatically increase the statistical power of the screen for relatively little or no extra cost.

4.1 Introduction

scCRISPR screens utility has been limited to niche use cases where the biology under investigation involves essential genes, pathways, large-scale experiments with hundreds to thousands of cells per target gene, or immediate readout scenarios such as promoter and enhancer gene linkage experiments [97, 50, 115]. Although significant progress has been made in resolving many issues related to scCRISPR screens protocols, several questions still persist regarding the primary cause of their limited statistical power and how to overcome it. Is the issue primarily related to experimental design? Are we not optimizing our experiments sufficiently to maximize the likelihood of observing an effect, such as ensuring adequate sequencing depth or increasing the number of cells analyzed? Could the choice of a specific CRISPR variant for inducing perturbations yield better results compared to others? Is

it possible to modify single-cell sequencing protocols to enhance power? Or is this a fundamental limitation that is inherent to scCRISPR that cannot be overcome?

To gain deeper insights into the power issues associated with scCRISPR screens and facilitate their broader applicability in biological contexts, I employed crisprPower to conduct simulations and evaluate the statistical power of various experimental designs. By manipulating the number of cells per perturbation, ranging from 1 to 1000, I compared the performance of CRISPRko and CRISPRi. Additionally, I investigated the potential enhancement of statistical power in scCRISPR screens through the utilization of single-cell sequencing methods that employ targeted panels. While existing scCRISPR screen designs suffer from chronic underpowering, there exist practical and readily accessible means to enhance the statistical power of screens. These improvements hold the promise of enabling scCRISPR screens to deliver high-throughput and high-resolution insights into the cellular response to perturbation, unlocking their full potential.

4.2 Methods

4.2.1 Simulating CRISPR Perturbations via CrisprPower

To assess the effect of various publically available scCRISPR screen experimental protocols and novel designs, specifically experiments that utilized alternative sequencing protocols, on the statistical power of the experiment and the ability to observe a given perturbation effect, I conducted simulations comparing scCRISPR performance in terms of power using either a targeted and whole transcriptome. To conduct these simulations, I utilized crisprPower's gene expression model, a novel statistical scCRISPR simulator that can simulate CRISPR perturbations. I did not utilise simulated GRNs; instead, I directly sampled mean expression and estimate dispersions from fitted expression distribution.

4.2.2 Estimating On-target Activity Distribution

To sample the On-Target Activity of CRISPRko, I fitted the predicted on-target activity scores of all gRNAs across the Human Genome. I downloaded the University of California Santa Cruz (UCSC) genome browser data on gRNA target sites and used the Python package fitter to fit a beta distribution using MLE. The on-target activity score provided by the UCSC genome browser for a gRNA target site was the Fusi score for measuring the on-target activity of gRNAs [37, 104]. Fusi scores are no longer state of the art metric for estimating on-target activity; instead, a new score developed by DeWeirdt et al. 2022 is the current state of the art. However, the accessibility of the Fusi Score via the UCSC genome browser and its

availability for all gRNAs within protein coding regions of the genome enables an endeavor to estimate the 'global' On-Target Activity Distribution of gRNAs. However, it's crucial to acknowledge that since all current state-of-the-art On-Target activity scores are normalized ranks, they do not fully represent the probability of Knockouts or Occupancy as defined in crisprPowers' CRISPR perturbation model.

4.2.3 Calculating Statistical Power of Perturbation

Statistical power is the probability of correctly rejecting the null hypothesis in a hypothesis test or detecting an effect if it exists. It can be calculated using parametric or non-parametric methods. Non-parametric approaches, like empirical-based methods, are effective but slower as they rely on general changes in moment statistics such as the mean. In contrast, parametric methods are faster and account for the unique properties of a specific distribution.

In this particular scenario, I employed a parametric method to calculate the statistical power of CRISPR perturbation in an NB distribution, simulated using crisprPower. This approach allowed me to account for changes beyond just the mean and incorporate the variance introduced by CRISPR perturbations. The equation I utilized to compute the statistical power is derived from the work of Cundill and Alexander. Their equation was originally developed to determine the required sample size (N) for an experiment with a desired power (p), considering the mean values (μ_{ctrl}, μ_{prtb}) and dispersions (σ_{ctrl}, σ_{prtb}) of the control and perturbed samples, as well as the significance level (α) (with a default value of 0.05).Q_{ctrl} represents the percentage of samples belonging to the control condition and Q_{prtb} represents the percentage of samples belonging to the perturbed condition. To calculate the sample size, Cundill and Alexander transformed both α and p to the corresponding standard normal deviates, denoted as $Z_{1-\frac{\alpha}{2}}$ and Z_{1-p}, respectively. These standard normal deviates represent critical values associated with the chosen significance level and power. The original equation proposed by Cundill and Alexander is presented as follows:

$$\sqrt{N} = \frac{(Z_{1-\frac{\alpha}{2}} + Z_{1-p}) \times \sqrt{(\frac{1}{Q_{\text{ctrl}}}\sigma_{\text{ctrl}^2}) + (\frac{1}{Q_{\text{prtb}}}\sigma^2_{\text{prtb}})}}{\mu_{\text{ctrl}} - \mu_{\text{prtb}}} \tag{4.1}$$

While Cundill and Alexander do not provide a method for directly calculating the power p of a specific experimental condition for a given sample size (as shown in Equation 4.1), I have extended their work and derived a new formula by isolating Z_{1-p} (see Equation 4.2). This new formula enables me to estimate the statistical power for a particular experimental condition, such as CRISPR perturbation, based on the number of cells per perturbation and changes in NB distributions mean and variance between control and perturbed conditions. In

the equation as the same equation above, N represents the number of samples, μ_{ctrl} represents the control sample mean, μ_{prtb} represents the perturbed sample mean, σ^2_{ctrl} represents the control sample variance, σ^2_{prtb} represents the perturbed sample variance, Q is the percentage of samples belonging to a given condition. Where Q_{ctrl} represents the percentage of samples belonging to the control condition and Q_{prtb} represents the percentage of samples belonging to the perturbed condition, and α represents the statistical significance threshold. Finally, after calculating Z_{1-p}, it can be transformed from an estimated standard normal deviate to the cumulative distribution function (CDF) to obtain the statistical power of the experiment. By utilizing this derived formula, it is possible to determine the power of the experiment based on the sample size, as well as the control and perturbed NB parameters. The equation I derived is presented as follows:

$$Z_{1-p} = \frac{\sqrt{N} \times \left(\mu_{\text{ctrl}} - \mu_{\text{prtb}}\right)}{\sqrt{\left(\frac{1}{Q_{\text{ctrl}}}\sigma_{\text{ctrl}^2}\right) + \left(\frac{1}{Q_{\text{prtb}}}\sigma^2_{\text{prtb}}\right)}} - Z_{1-\frac{\alpha}{2}} \tag{4.2}$$

4.2.4 Datasets

For this analysis, I used the 40,000 cells from the Human Cell Atlas on ICA Bone Marrow Dataset to fit the mean expression distributions of both the overall distribution of cell populations and specific gene classes using crisprPower [59]. The advantage of utilizing this dataset lies in its substantial size, convenient accessibility, and diverse representation of biological settings. Notably, the dataset encompasses up to 35 distinct cell populations, encompassing major types such as granulocytic, monocytic, lymphoid, erythroid, megakaryocytic, and eosinophil populations. These populations originate from eight different donors spanning various age ranges, with a minimum of 9,800 cells obtained from each donor [59]. Overall, the ease of access to this publically available data (via SeuratData), excellent experimental design by the Human Cell Atlas, and diversity in a biological context provide an ideal dataset to generate simulations from.

4.3 Results

4.3.1 Quantifying the Statistical Power of scCRISPR Screens

To determine the overall statistical power of a given scCRISPR screen that has already been conducted is extremely context-specific and, while informative about the experiment, provides us with little useful information that can be used to direct further improvements to the design of scCRISPR Screens. crisprPower is a scCRISPR screen statistical simulator

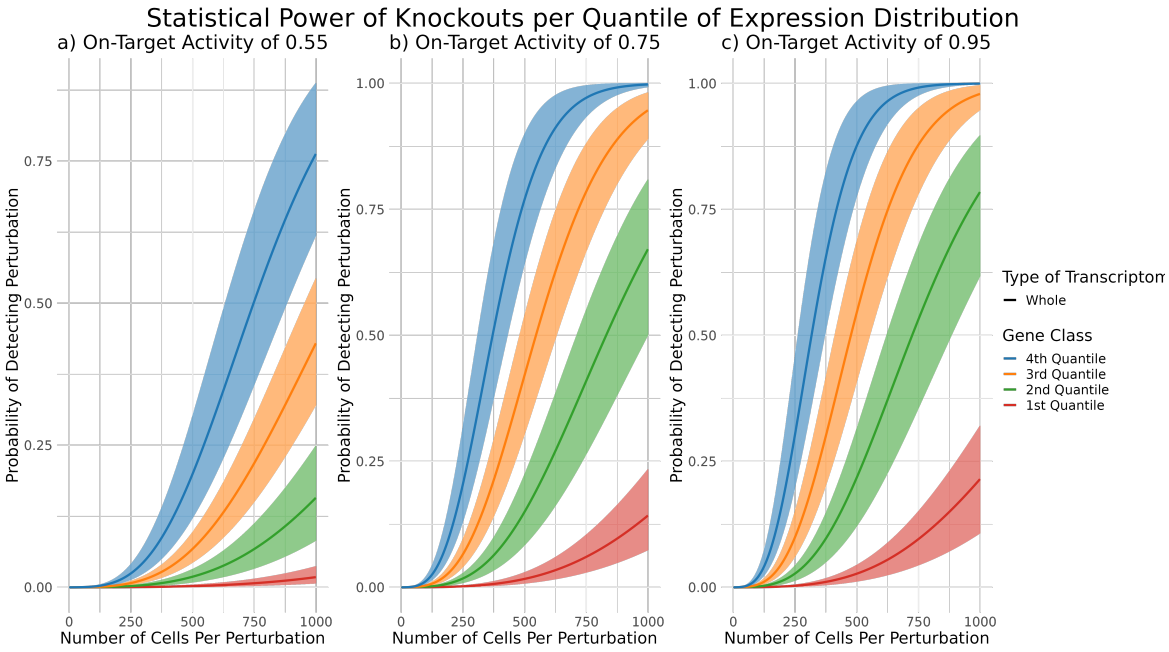


Fig. 4.1 Statistical power of CRISPRko perturbations in scCRISPR Screens as the number of cells increases from 1 to 1000. Each quantile has been coloured and is rank order from the highest mean expression in the 4th quantile to the lowest mean expression in the 1st quantile.

that offers the capability to simulate an extensive range of design options, spanning from near-perfect to the worst experimental designs, allowing for a comprehensive exploration of the entire space of potential experimental designs of scCRISPR screens. To quantify the statistical power of scCRISPR screens, I decided to start with the perfect scCRISPR screen experiment where the excepted unique molecular count per cell (also known as the cell library size) is 14000, and the number of cells perturbation varies from 1 to 1000. In addition, I simulated CRISPRko as the perturbations at min (0.55), mean (0.75), and max (0.95) values of the estimated on-target activity distribution of gRNAs and estimate the power of a knockout using Equation 4.2.

There is an overall mean expression distribution for a given cell population where genes can be ranked from highest to lowest mean expression. To understand how differences in mean expression affect the probability of observing the perturbation effect of a given gene, I split the distribution by quantile. I sampled a hundred mean expression values per quantile to be simulated across the different on-target activity levels and at the near-perfect scCRISPR screen experimental design mentioned above. Next, I estimated the statistical power as the number of cells per perturbation increased from 1 to 1000 for each sample mean expression. I plotted a series of power curves to visualize the change in statistical power as the number of cells increased, with each quantile getting a curve (see figure 4.1).

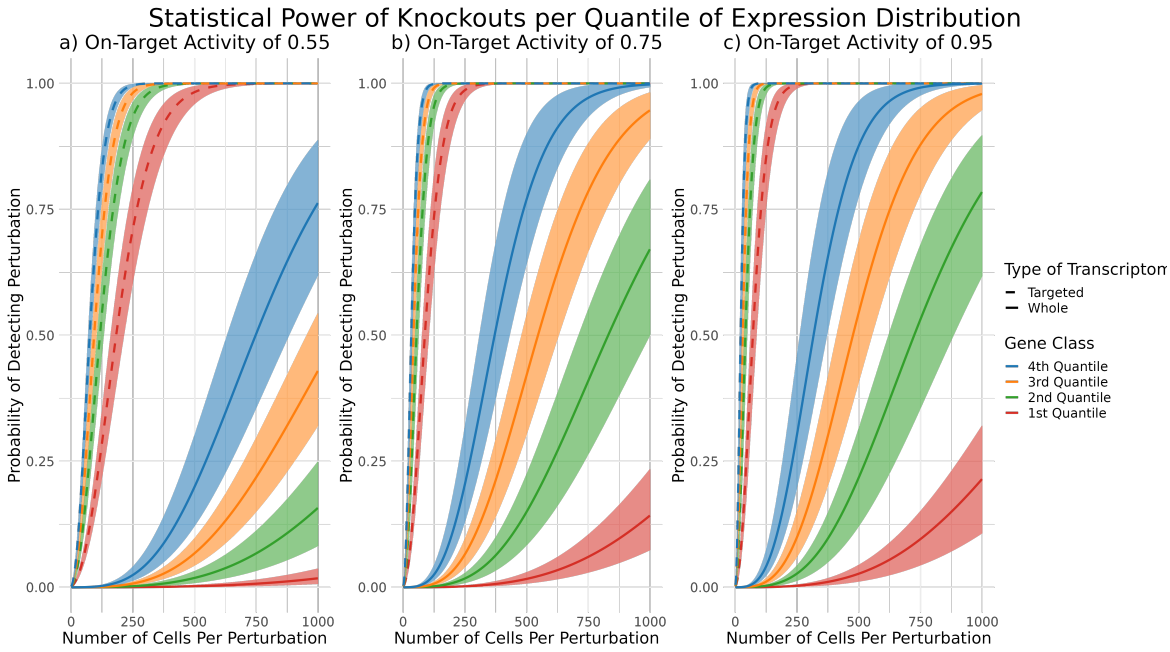


Fig. 4.2 Comparing the statistical power of targeted and whole transcriptomes as the number of cells increases from 1 to 1000. The solid line is the power curve of the whole transcriptome, and the dashed line is the power curve for a targeted transcriptome. Each quantile has a unique colour and is rank-ordered from the highest mean expression in the 4th quantile to the lowest mean expression in the 1st quantile.

The results are remarkably striking, revealing a concerning lack of statistical power in the majority of published scCRISPR screens unless they target an essential set of genes or pathways. The extent of this underpowering is quite surprising; I initially expected scCRISPR screens to be moderately underpowered, possibly requiring around 250 cells. However, the simulation results have shown a different reality, indicating that even for a perfect knockout, a minimum of 500 cells is necessary, as depicted in Figure 4.2. As anticipated, the on-target activity of a gRNA has the most significant impact on the power of a scCRISPR screen. When the average gRNA on-target activity is approximately 0.55, the experiment lacks the ability to detect any changes unless a minimum of 600 cells or more are used per perturbation, as shown in Figure 4.1a. Even with an increasing number of cells beyond 600, the experiment is likely to detect only effects in the 4th quantile, remaining unable to identify perturbation effects in genes from the other three quantiles.

On-target activity, however, appears to have diminishing returns, as is evident in minor differences in the power curves of CRISPRko with 0.75 and 0.95 on-target activity (see in figure 4.1b and 4.1c). While the increase in on-target activity exhibits a substantial difference between the on-target activities of 0.55 and 0.75, it is comparatively smaller than

the difference observed between an on-target activity of 0.55 and 0.95 (see figure 4.1b and 4.1c). Specifically, once on-target activity is greater than 0.75; it is possible to observe the perturbation effects of genes in the 4th quantile with 400 cells per perturbation. At the same time, the 3rd quantile is observable once approximately 600 to 700 cells are observed per perturbation, with the 2nd and 1st quantiles being simply unobservable even with 1000 cells (see figure 4.1b). For CRISPRko of 0.95, it is possible to observe the 2nd quantile with 800 to 900 cells per perturbation (see figure 4.1c).

This is good news, as the excepted on-target activity of most gRNAs in the genome used for CRISPRko is 0.75. Despite this good news, the power curves for on-target activity are depressingly low, indicating that it is currently not possible to see any effects until a minimum of 400 cells per perturbation for the 4th quantile, a minimum of 600 cells to see effects in the 3rd quantile, to observe the effects of genes in the 2nd quantile requires a 1000 cells. However, despite having over 1000 cells, there is no hope of seeing anything in the 4th quantile. These results highlight the significant limitations of scCRISPR screens using the current single-cell sequencing protocols, even in simulations with maximum CRISPRko on-target activity. It is evident that scCRISPR screens are severely underpowered, and this issue is further exacerbated when using gRNAs with on-target activity performance closer to the expected levels.

4.3.2 Comparing Statistical Power: Targeted vs Whole Transcriptome

While no one has quantified the statistical power of scCRISPR, it is known that, in general, the method is underpowered to address these issues. Targeted panels were merged with scRNA-Seq to enrich for specific genes of interest in perturbation screens and to reduce cost via a decrease in sequencing [123, 114]. There are two types of experimental protocols for targeted panels: custom PCR Primer and antibody pulldown-based enrichment. Each of these methods has its pros and cons; further research into their efficiency is required; at the core, however, each method attempts to overcome the experimental and statistical limitation of using the whole transcriptome as a readout in single-cell perturbation screens.

I reran my simulations using crisprPower again, assuming a near-perfect experimental design of an average cell library size of 14000 for the whole transcriptome and an average cell library size of 2000 and the number of cells per perturbation being dynamically adjustable from 1 to 1000. Again I sampled 50 genes' mean expression values from each quantile of the mean expression distribution that was simulated across the min, mean, and max (0.55, 0.75, 0.95) of on-target activity distribution. The only major difference was that I simulated both whole and targeted transcriptomes predicting the log-transformed mean expression of the targeted gene using crisprPowers linear regression model that was fitted using Schraivogel

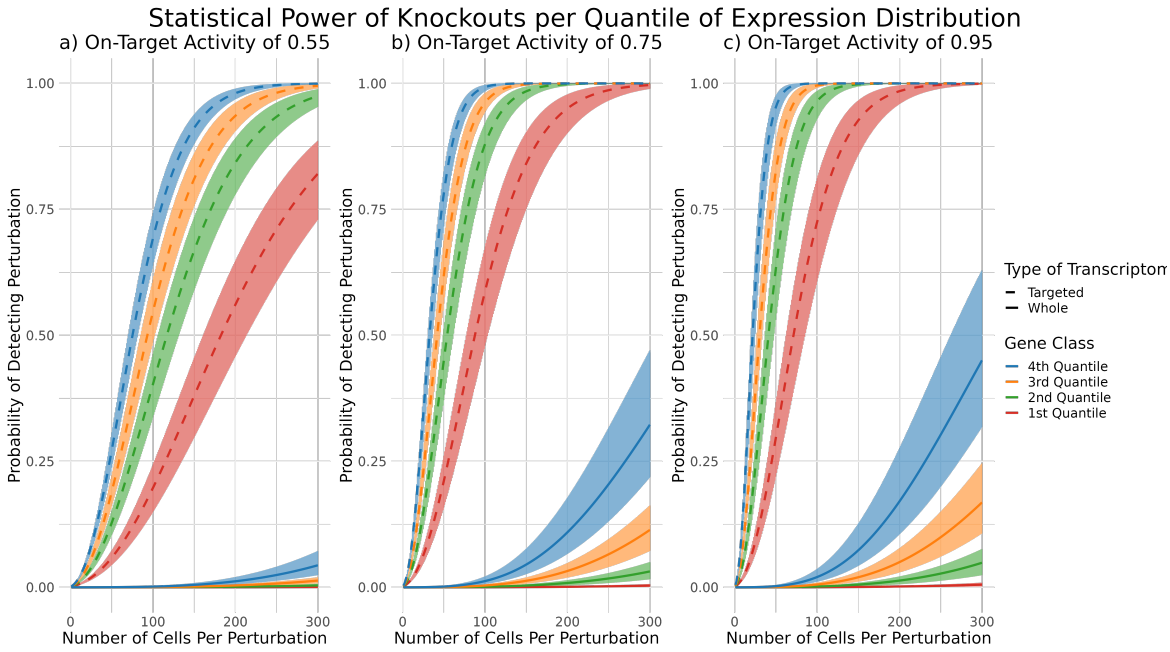


Fig. 4.3 Comparing the statistical power of targeted and whole transcriptomes as the number of cells increases from 1 to 300. The solid line is the power curve of the whole transcriptome, and the dashed line is the power curve for a targeted transcriptome. Each quantile has a unique colour and is rank-ordered from the highest mean expression in the 4th quantile to the lowest mean expression in the 1st quantile.

et al. paired targeted and whole transcriptome dataset. I then calculated the statistical power of the perturbation per on-target activity and quantile as the number of cells increased from 1 to 1000.

The difference in the statistical power between the targeted and whole transcriptome is stark (see figure 4.2). Target transcriptomes outperformed whole transcriptomes by a substantial margin with a clear increase in statistical power and are now able to observe perturbation effects across all of the quantiles in targeted transcriptomes with only 250 cells across all of the simulated on-target activities. While in comparison to the whole transcriptome, where perturbation effects in the 4th quantile a barely observable, let alone in the remaining quantiles. In fact, the difference between targeted and whole transcriptomes is so substantial that the power curves of the targeted transcriptome quickly reach the upper right corner of the power curves in 4.2 when simulating 1 to 1000 cells.

To assess the difference in performance between the quantiles, it's necessary to replot the power curves this time with increasing from 1 to 300 cells of the power curve plot (see Figure 4.3). When comparing target transcriptomes power curves in from this perspective, there are clear but excepted differences in performance driven by the quantiles of mean expression

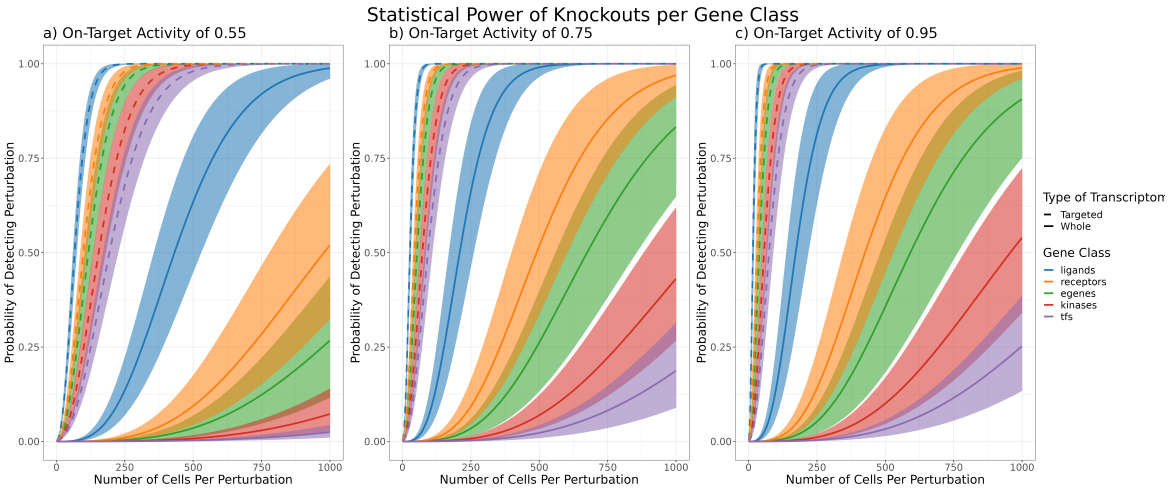


Fig. 4.4 Statistical power of CRISPR perturbations in both targeted and whole transcriptomes for the expression distribution of a gene class from 1 to 1000 cells. Similar to previous figures, solid lines are the power curve for whole transcriptome, and dashed lines represent targeted transcriptomes. gene class specific expression distributions were fitted from ICA Human Cell Atlas Data.

distribution and the on-target activity of the gRNA. For an on-target activity of 0.55, the simulations suggest that a minimum of 250 cells is needed to observe the majority of all perturbation effects in the experiment, with fewer cells required for the higher quantiles. For the on-target activities of 0.75 and 0.95, it is possible to observe the perturbation effects across all quantiles with only 150 cells. These results suggest that, on average, a scCRISPR screen using targeted transcriptomes requires approximately 150-200 cells per target gene to ensure the observability of perturbations. These findings provide compelling evidence that targeted panels greatly enhance the statistical power of scCRISPR screens and enable the detection of perturbation effects across a wide range of mean expression levels.

4.3.3 Comparing Statistical Power: Targeted vs. Whole Transcriptome by Gene Class

Previously, I have focused on measuring the statistical power across the quantiles of a given cell population mean expression distribution. While normal scCRISPR screens are chronically underpowered, targeted panels offer a solution to this issue. However, the question remains how does this apply to specific gene classes? To gain deeper insights into the practical implications of the enhanced statistical power provided by targeted transcriptomes, I conducted a new single-cell sequencing experiment. Instead of utilizing quantiles, I employed the pre-defined gene classes from crisprPowers, namely Transcription Factors,

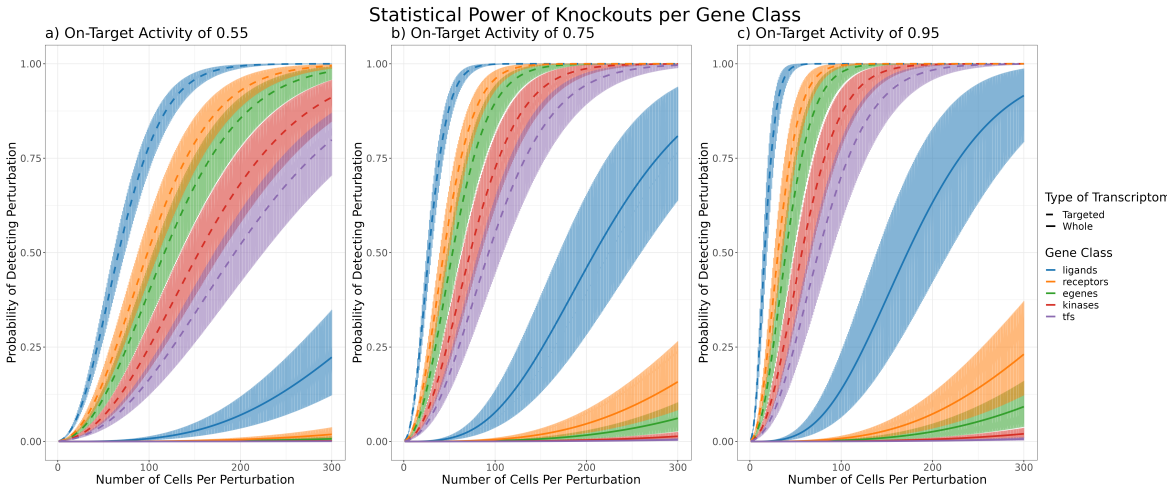


Fig. 4.5 Statistical Power of CRISPR perturbations in both targeted and whole transcriptomes for the expression distribution of a gene class from 1 to 300 cells. Similar to previous figures, solid lines are the power curve for the whole transcriptome, and dashed lines represent targeted transcriptomes. Gene class specific expression distributions were fitted from ICA Human Cell Atlas.

Kinases, Ligands, Receptors, and Other Genes, based on GO Terms. The "Other Genes" class encompasses all the genes that could not be classified into any of the other gene classes. For each gene class, I sampled 100 mean expression values from its mean expression distribution fitted from the ICA Human Cell Atlas dataset using crisprPower. I only considered the minimum, mean, and maximum on-target activity of CRISPRko (0.55, 0.75, 0.95). Finally, I plotted the power curve from the result simulations from 1 to 1000 and a zoomed-in version from 1 to 300 cells.

The results of this analysis follow a similar trend to the mean expression quantiles that were studied previously. Target panels are a more effective experimental method to improve the experiment's statistical power with substantial improvements over whole transcriptome readouts see Figure 4.4. However, when comparing the zoomed figures of 4.5 and 4.3, it is clear that there is an overall decrease in the statistical power in the gene classes when compared to the quantiles of the overall mean expression distribution. This is most likely because the quantiles of a cell population's mean expression distribution will contain a mixture of genes from the various classes there. As such, highly expressed genes will shift the overall mean expression higher when fitting the data, increasing the likelihood of observing the perturbation. When fitting gene class specific mean expression distribution, these mixtures are removed, which results in a decrease in the statistical power which is more representative of actual biology.

Regardless, the decrease in statistical power, while significant, is not dramatic. Instead of needing a min 250 cells to see all effects when accounting for gene class, a minimum of 300 cells are needed to observe all of the perturbation effects and vary with gRNA on-target activity. When perturbing gene classes with an on-target activity of 0.55, 300 cells are needed to observe all of the perturbation effects regardless of gene class (see figure 4.5a). While perturbing genes with the excepted on-target activity of 0.75 tightens the power curves substantially, suggesting that only 200 cells are required to observe perturbations across all of the gene classes. Finally, the perfect perturbation activity of 0.95 once again reflects the previously observed trend that there is no significant increase in statistical power beyond 0.75. Overall though the results reflect our previous observations that target panels are better, there is a diminishing return in on-target activity beyond the mean of 0.75.

scCRISPR screens are commonly employed to perturb a specific set of genes belonging to a particular gene class or pathway. This has practical implications in screen design, as it necessitates accounting for only one or two gene classes, potentially reducing the number of cells required substantially. Simulating scCRISPR screens using gene classes offers the advantage of extracting class-specific recommendations for experimental design. For instance, assuming an on-target activity of 0.75 or greater, a screen that is aimed at investigating ligands and receptors only needs 50 to 70 cells. While a screen focused on perturbing kinases or TFs would require far more cells per target gene, from 150 to 200 cells. It is important to note that all these class-specific recommended cell counts will vary depending on the biological context of the screen. Therefore, conducting pilot experiments or simulations is crucial to obtain an overall representation of the biological system. Using the data from the pilot experiment crisprPower can be utilised to estimate the required number of cells per class.

4.3.4 Knockouts or Interference: which is more powerful?

In the previous analyses, my focus was on determining the optimal single-cell sequencing procedure by comparing target panels to whole transcriptome outputs across a general and gene class-specific fitted mean expression distribution. Additionally, I simulated perturbations based on the assumption that CRISPRko would be the selected perturbation method. However, it's essential to acknowledge that CRISPRi, another commonly used CRISPR-based method, is also employed in scCRISPR Screens as an alternative approach. CRISPRi offers several appealing characteristics in comparison to CRISPRko, especially in its deterministic nature induced by a transfused protein with default activity. Unlike CRISPRko perturbations, CRISPRi does not rely on cellular DNA repair mechanisms to induce the perturbation. This

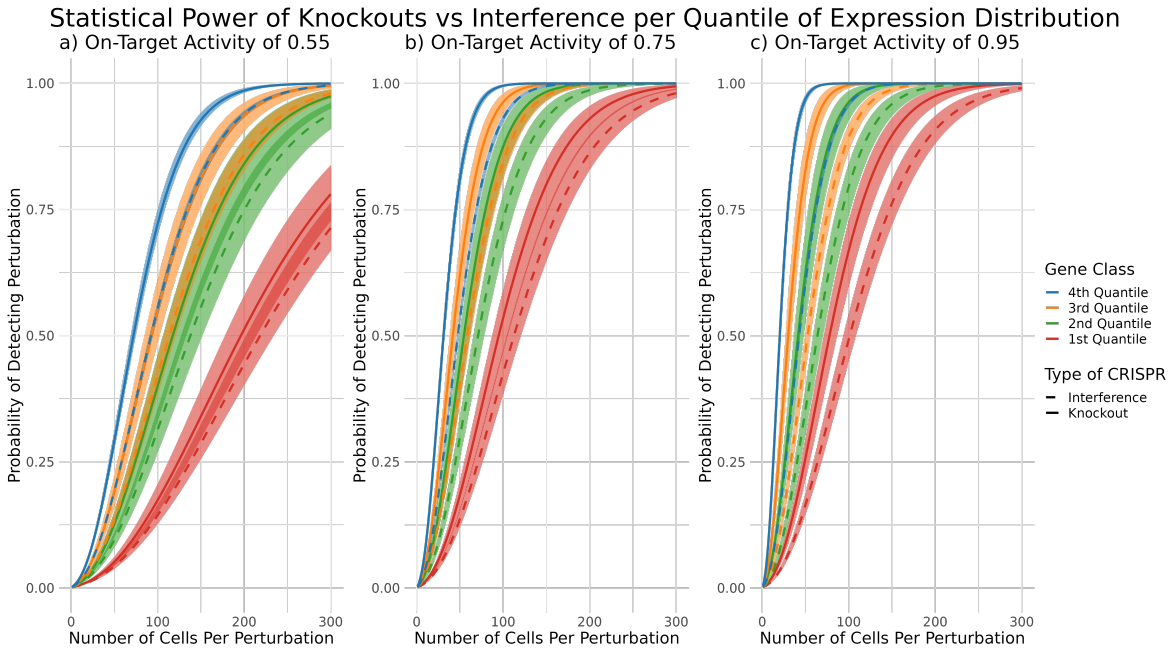


Fig. 4.6 Comparing the statistical power of CRISPRi vs CRISPRko using a targeted transcriptome as the number of cells increases from 1 to 300. In this comparison, the on-target activity is the same for both perturbation methods. The solid lines plot the power curve of CRISPRko, and the dashed line plots the power curve for CRISPR interference.

attribute allows for direct measurement of perturbation effects on both target genes and off-targets, as observed in the transcriptional readout.

While this offers researchers the appealing potential to rapidly and reliably identify off-target effects in the experiment, its superiority over CRISPRko in terms of generated statistical power remains uncertain. To investigate this, I conducted a series of simulations of CRISPRko and CRISPRi perturbations using the previously employed perfect experimental design. The focus was on comparing the differences in statistical power based on the quantile overall mean expression distributions of a cell population.

In the first round of simulations, I assumed that the on-target activity of both CRISPRko and CRISPRi were the same in order to compare the performance of underlying perturbation mechanisms. It is immediately apparent that when the on-target activity of these perturbation methods is the same CRISPRko is more potent than CRISPRi across all the quantiles across all quantiles see figure 4.6. Quantile-specific differences exist regarding the superiority of CRISPRko over CRISPRi, with the 4th quantile exhibiting the greatest disparity and the 1st quantile showing the least. This pattern is consistent across all simulated on-target activities. At an on-target activity of 0.55, both CRISPRko and CRISPRi require the same number of cells (see figure 4.6a). However, this quickly changes when the on-target activity reaches

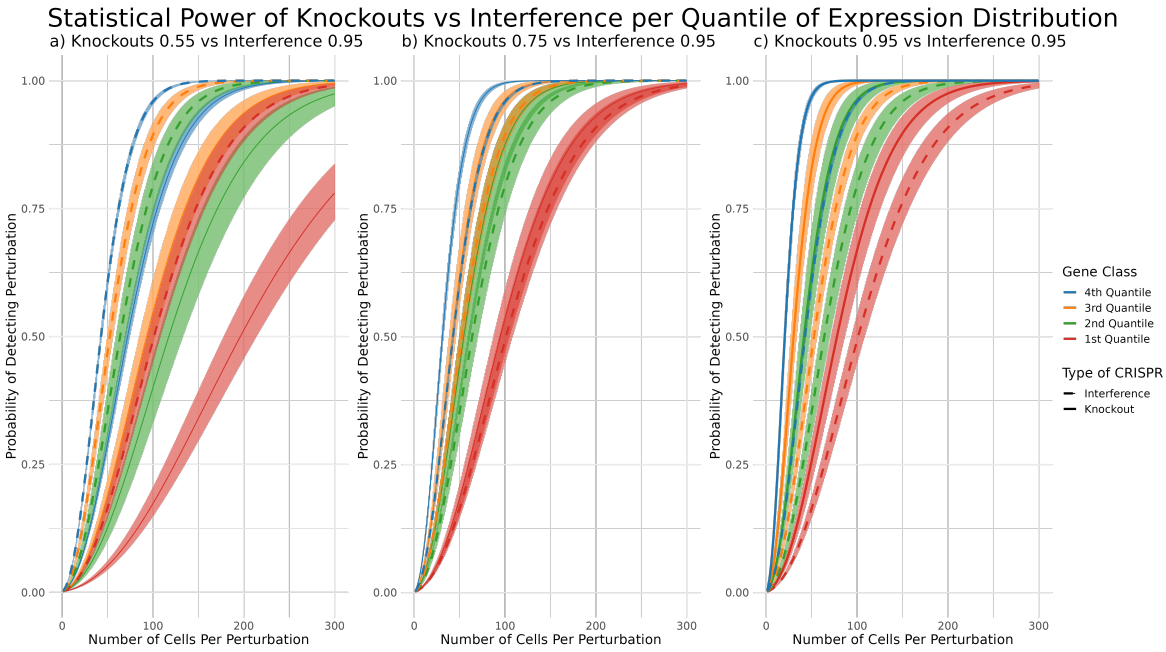


Fig. 4.7 Comparing the statistical power of CRISPRi vs CRISPRko using a targeted transcriptome as the number of cells increases from 1 to 300. In this comparison, the on-target activity differs for CRISPRko on-target activity varies from 0.55 on the right to 0.95 on the left. While CRISPRi on-target activity is kept the same at 0.95 as this is the excepted on-target activity of this perturbation method. The solid lines plot the power curve of CRISPRko, and the dashed line plots the power curve for CRISPRi.

0.75 or higher. In comparison to CRISPRi, CRISPRko only requires 150 cells to observe all perturbation effects as seen in 4.6b and 4.6c. As the on-target activity increases, the performance gap between CRISPRko and CRISPRi widens further. It is important to note that this comparison is unrealistic, as CRISPRi perturbation is deterministic and depends on the proximity of the target site to the Transcription Start Site of a given gene. Typically, CRISPRi exhibits an on-target activity of 0.95.

To compare a more realistic CRISPRi on-target activity, I plotted the power curves for both CRISPRko for the 0.55, 0.75, and 0.95 vs 0.95 for CRISPRi (see figure 4.7). Even in this more realistic comparison regarding the excepted on-target activity of these methods, CRISPRko consistently outperforms CRISPRi in all cases except in the minimum gRNA on-target activity of 0.55 (4.7a). Once the on-target activity of CRISPRko surpasses 0.75, a significant and consistent trend emerges, reaffirming previous observations. Specifically, screens utilizing CRISPRko exhibit a greater amount of statistical power compared to CRISPRi-based screens when the on-target activity exceeds 0.75 across all quantiles, with the 4th quantile exhibiting a slight advantage (4.7b). This superior performance of CRISPRko becomes more pronounced as the on-target activity increases to 0.95, with even

the 4th quantile demonstrating a significant increase in statistical power over CRISPRi (4.7c). Overall, these results demonstrate that even in a more realistic scenario with near-perfect perturbations using CRISPRi, CRISPRko-based screens remain inherently more powerful due to the permeant nature of CRISPRko's, in comparison to CRISPRi perturbations where the perturbation only occurs while CRISPR is bound to the target site.

While these results are initially shocking, it becomes clear once the underlying perturbation mechanisms are considered in greater detail. CRISPRko induces a permanent alteration to the DNA sequence of a gene that will always prevent it from being expressed; despite the probabilistic nature of CRISPRko, a permanent perturbation will consistently outperform an induced perturbation. CRISPRi is an induced perturbation that only occurs so long as it's bound to the target site as a transfused transcriptional repressor mediates it. These results indicate that CRISPRko induces a more significant perturbation effect than CRISPRi, reflected in the greater power of CRISPRko simulations. However, there are still use cases for CRISPRi, such as concerns for off-target effects and non-protein coding genomic regions.

4.4 Discussion

scCRISPR screens are a novel high-throughput high-resolution experimental protocol with tremendous promise in empowering researchers to investigate biological systems. However, despite the initial excitement and developments, it is clear that these screens suffered from numerous issues in their experimental protocols and were underpowered. Much of the research done in scCRISPR screens has previously focused on improving experimental protocols by switching from gRNA barcodes to direct capture of gRNA and improving the lentiviral backbone used. Little research has been done on determining the overall statistical power of scCRISPR screens and how various experimental design choices in single-cell sequencing experiments affect these screens' statistical power. Finally, while these developments have dramatically improved scCRISPR screens, they are still chronically underpowered.

In this study, I quantified the statistical power of scCRISRP screens using the analytical simulator crisprPower under varying experimental conditions and highly optimal experimental protocol based on previously published experimental methodology. My initial simulations focused on a normal single-cell sequencing experiment that generates a whole transcriptome readout under idealistic circumstances. Specifically, I simulated a near-perfect single-cell sequencing experiment where up to 1000 cells per perturbation and an average cell library size of 14000. Even under this idealistic and costly experimental design, I observed that it was impossible to observe the effects of perturbations for most genes without a minimum of

600 cells for a given target gene. Indicating that scCRISPR screens are underpowered and practically useless in their current experimental workflow. Any proposed results observed are noise and can not be definitively said to mean anything unless an essential gene/pathway was being investigated or the study had at least 600 cells per perturbation.

Targeted transcriptomes for single-cell sequencing experiments are recent developments initially aimed at reducing the cost of the experiments through reducing sequencing. There are two primary methods of creating a targeted panel; the first uses PCR primers called Targeted Panel Sequencing (TAP-Seq), and the second is based on antibody pulldown of hybridized probes. TAP-Seq works by modifying the 10x Genomics workflow at the cDNA amplification step. Instead of amplifying the cDNA using 10x protocol, custom PCR primers are designed using the TAPseq R package to amplify only a specific subset of the desired transcriptome [123]. While the antibody pulldown method works by adding a step into the 10x protocol after cDNA amplification, hybridization probes are added to the amplified cDNA designed to target a specific subset of the transcriptome. After 24 hours, the hybridized probes are pulled using antibodies, and this pulled pool of DNA is then sequenced, thereby enriching the desired subset of the transcriptome. Although these are two distinct methods for this study, I treated them as the same. In simulations, comparing Targeted to Whole transcriptomes demonstrated that targeted panels could dramatically reduce the number of cells required to observe a perturbation. Specifically, when simulating a targeted panel of 1000 genes, I observed that only 75-100 cells, on average, were required to observe a perturbation effect. Through careful choice of a transcriptome subset (for instance, targeting the L1000), researchers can observe statistically significant results in scCRISPR screens.

In previous simulations, I have explored the statistical power of experimental designs in terms of the overall mean expression distribution of a given cell population. However, this doesn't relate to specific gene classes. As such, I reran by Targeted vs Whole Transcriptome fitting individual mean expression distributions using Human Cell Atlas on ICA Bone Marrow Dataset for interesting sets of gene classes: TFs, Ligands, Receptors, Kinases, and E-Genes. The results of this simulation followed a general trend similar to previous simulations that Targeted transcriptomes outperformed Whole transcriptomes substantially. Still, they did suggest that around 150 to 200 cells are needed to observe effects across all gene classes. However, suppose only all of the targets come from a particular gene class, such as Ligands or Receptors in this dataset. In that case, the overall number of cells required decreases to 30-50 cells per perturbation. Indicating that experimental design has some flexibility depending upon the target genes being perturbed when using a targeted panel. These simulations' results suggest that a Targeted Transcriptome should always be used over a Whole Transcriptome.

Targeted transcriptomes are essential to have a powered scCRISPR screen. Previous simulations perturbations simulated CRISPRko, but how does using CRISPRi affect the statistical power of the screen? I simulated and compared the statistical power of both CRISPRko and CRISPRi and observed that CRISPRko were more potent than CRISPRi under all circumstances. The performance difference reflects the underlying method of introducing the perturbation. CRISPRko induces a frameshift mutation which knocks the target gene out of frame, thereby preventing gene expression. While CRISPRi is binding at the target's TSS and repressing the gene's expression via a transfused repressor. This results in a more deterministic but fundamentally leaky perturbation as CRISPRi perturbation only occurs so long as it binds to its target site. As it cycles through binding and unbinding, transcripts can be expressed in the interlude. Yet despite CRISPRko's superior performance over CRISPRi, this doesn't necessarily mean it should be used in all experiments. CRISPRko only outperforms CRISPRi in the protein-coding sequences of the genome. If scCRISPR screens are being used to determine promoters-gene linkage, enhancers-gene linkage, or another non-coding section of the genome, CRISPRi is still an invaluable tool. When utilizing CRISPRi, it is still necessary to have a Targeted transcriptome to ensure that the screen has sufficient statistical power.

scCRISPR screens are an exciting new experimental protocol with tremendous promise in helping researchers bridge the gap and map genotype to phenotype in biological systems. However, despite recent improvements in technology, there still underpowered. In this study, I first quantified the statistical power of various scCRISPR Screen experimental protocols and methods of introducing perturbation. My results indicate a statistical power and cost-effective scCRISPR screen is possible if a Targeted Transcriptome is used. CRISPRko should be used whenever possible over CRISPRi, but there are still cases in the non-protein coding part of the genome.

4.5 Caveats and Limitations

Several limitations are inherent in crisprPower, although most are minor. For instance, using linear regression to model regulatory interactions assumes that these interactions are fundamentally linear and disregards non-linearity, which is an obvious oversimplification. Nonetheless, these approximations serve as reasonable representations to simulate the broader effects of CRISPR perturbations in scCRISPR. However, there are two implicit assumptions that are significantly related to crisprPower's ability to simulate targeted panels, which arise from using linear regression to predict the mean expression of genes within such panels.

Firstly, it should be noted that the current version of crisprPower assumes a targeted panel consisting of 978 genes, which corresponds to the number of genes in the l1000 gene set. I initially developed crisprPowers target transcriptome model in this manner because it was easy to estimate using Schraivogel et al. data. This limitation means that the tool cannot simulate a targeted panel of any other size, rendering it non-generalizable to different target panel sizes, such as a 100-gene panel or any other size.

Second, crisprPower assumes that the gene expression composition of the panel is the same as the l1000 gene set. Target panels are designed to address the issue of sparsity in datasets; they are able to address this by reducing the number of genes being sequenced to a specific set of genes. However, even after selecting a specific gene set, variations in mean expression can still result in sparsity if the composition of the gene transcripts varies significantly. This is because the underlying sampling process of single-cell experiments remains unchanged.

For instance, consider a target panel comprised of one gene with low expression and four genes with high expression. In this scenario, sparsity would be observed in the low-expressed gene due to the larger number of available transcripts for sequencing highly expressed genes. Conversely, a target panel consisting of five low-expression genes would exhibit minimal sparsity since all genes have an equal chance of being sequenced. The statistical power of the targeted panel will vary depending on the alterations in the number of genes and the composition of the expressed transcripts within those genes.

I do not believe these limitations of crisprPower fundamentally alter the results that targeted transcriptomes have greater statistical power than whole transcriptomes. Rather they affect the general usability of crisprPower as a software package to design new targeted panels for experiments in other biological settings. Due to these limitations, users cannot simulate a custom selection in terms of the number of genes or the composition of gene expression in the targeted panel for their experiment.

The use of target transcriptomes is currently limited, and it is unlikely researchers will conduct pilot experiments using this technology as it limits the overall resolution of the experiment. This presents a significant challenge that must be addressed to ensure the generalisability of crisprPower and improved experimental design of scCRISPR screens. In order to overcome this limitation, a mechanistic simulator that can flexibly incorporate wet lab methods for manipulating the transcriptome, even in the absence of a reference dataset, is needed. By developing such a simulator, these limitations can be effectively tackled, leading to broader applicability and increased usability of crisprPower.

Chapter 5

Mechanistic Simulations for Improved Single Cell Sequencing Experimental Design

A diverse array of specialized single-cell sequencing protocols is being continually developed to tackle specific challenges and cater to various applications. One such method is targeted transcriptomes, initially devised to address the limitations of scCRISPR screens by enhancing their statistical power. Nevertheless, the limited availability of datasets and their restricted use in a particular set of biological and experimental contexts pose a challenge to the design of optimal experiments across different settings. Moreover, the design of targeted transcriptomes involves multiple experimental parameters to consider, including the selection of genes for the panel and the composition of gene expressions. Existing simulators lack the capability to dynamically simulate these diverse conditions and their effect on the statistical properties of an experiment's observed counts. Consequently, a new mechanistic framework for simulating these emerging protocols is imperative. Minerva is a new wet-lab-aware mechanistic single-cell simulator capable of simulating various experimental designs of target transcriptomes and more.

5.1 Introduction

Minerva was designed from first principles to simulate single-cell datasets based on the sampling process of each single-cell experiment: capture chemistry, RT-PCR, cDNA Amplification, and Sequencing. The capture chemistry phase includes the use of capture probes designed to specifically target the Poly(A) tails of mRNA molecules. Once captured, the

molecules undergo conversion into copy DNA (cDNA) via Reverse Transcription PCR (RT-PCR), where due to the inherent inefficiencies of the reaction, up to 50% the captured mRNA molecules can be destroyed, effectively reducing the pool of captured molecules by half [70, 64, 142]. Following RT-PCR, the cDNA is amplified through PCR and subsequently sequenced. Throughout these stages, the inherent differences in the relative composition of transcripts among genes within a cell are further exacerbated as each step applies a sampling without replacement process. Consequently, a feedback loop arises, amplifying the imbalances between highly expressed and lowly expressed genes [110, 71]. This phenomenon is responsible for technical dropouts in single-cell sequencing experiments and presents challenges in accurately detecting low counts of specific genes.

Targeted transcriptomes have been developed to enhance the effectiveness of scCRISPR screens by focusing amplification and sequencing upon a specific set of genes. This is typically achieved through one of two methods either using custom PCR primers during cDNA amplification to selectively amplify the panel genes or employing a probe-based antibody pulldown after cDNA amplification to enrich the panel genes [123, 114]. Regardless of the chosen method, targeted transcriptomes are fundamentally manipulate the probabilities of sequencing a particular gene's transcripts in a binary manner based on whether the gene is enriched or not. One of the main challenges with target transcriptomes lies in the composition of the panel genes, as variations in their expression can significantly impact the statistical properties of the experiment, rendering the manipulation ineffective. However, if binary manipulation is possible, there is no inherent reason why a continuous manipulation of the sequencing probability of a gene's transcripts can not be done. By manipulating gene weights continuously, it may be possible to resolve the issues encountered with targeted transcriptomes and has the potential to eliminate technical sparsity from single-cell datasets. Transcriptomes whose probability of sequencing a given transcript is manipulated in a continuous manner I referred to as 'weighted transcriptomes'.

Minerva improves upon existing simulations by incorporating experimental manipulations such as amplification or enrichment protocols and modelling degradation during RT-PCR and is capable of simulating whole, targeted, and weighted transcriptomes. In this chapter, I validate Minerva's accuracy by comparing its output to both state-of-the-art simulators, SPARsim and Splatter, as well as real world data. This analysis highlights Minerva's capabilities in accurately simulating typical single-cell experiments. Furthermore, I demonstrate Minerva's effectiveness in simulating targeted panels, commonly used for enhancing sensitivity in single-cell experiments. Minerva also serves as a valuable tool for exploring theoretical experiments, allowing users to manipulate parameters and explore the statistical characteristics of these experimental protocols, such as weighted transcriptomes.

5.2 Methods

5.2.1 Properties of Noncentral Hypergeometric distributions

Minerva utilizes Noncentral Hypergeometric distributions to model the sampling process of single-cell experiments, which involves weighted sampling without replacement. Specifically, Minerva employs the Fisher Noncentral Hypergeometric (FNH) distribution to model the sampling processes for capture chemistry and cDNA amplification. To provide an intuition of the behaviour of this distribution and highlight the differences from its sibling distribution, the Wallienus Noncentral Hypergeometric (WNH), I will explain what Noncentral Hypergeometric distributions are, how they extend the Hypergeometric distribution, and present the mean, variance, and PMF of the FNH.

Differences between the Wallienus and Fisher Noncentral Hypergeometric distributions

The Hypergeometric distribution models a specific sampling process known as sampling without replacement. When discussing the Hypergeometric distribution and its related noncentral distributions, it is common to use a visual analogy of drawing balls from an urn, where the urn contains all the available balls for sampling. In this analogy, the number of balls in the urn is represented by N, and there are K populations whose individual populations, denoted as m, add up to N. Unlike the binomial distribution and most statistical distributions, when a ball is drawn, it is removed from the urn and not replaced. Noncentral Hypergeometric distributions are an extension of the Hypergeometric distribution that can accommodate weighted sampling scenarios. In these scenarios, each population has a weight ω that alters the probability of a class being sampled, with an increasing weight leading to a higher probability of selection [81, 86, 10].

There are two primary types of Noncentral Hypergeometric distributions: the WNH distribution and the FNH distribution [45]. The main distinction between the WNH and FNH distributions lies in the sampling process of the model. In the case of the WNH distribution, sampling is done sequentially, meaning that the ordering of sample draws affects the probability of sampling the next ball drawn based on the ball drawn and its assigned weight. In addition, because the WNH sampling process is sequential, it is possible to determine a desired sample size or the number of balls to be drawn in advance [45, 46]. The FNH distribution assumes no dependence between draws. Therefore, when sampling from an FNH distribution, all balls are drawn simultaneously from the urn without any prior knowledge of how many balls will be sampled [45, 46].

An insightful metaphor to understand these differences was proposed by Fog in his work on Biased Urn Theory. He likened the sampling process of the WNH distribution to fishing with a fishing rod, where only one fish can be caught at a time. The probability of catching a particular fish species increases with its weight. In this scenario, the desired number of fish to catch can be decided in advance. While the FNH distribution models a process where a fishing net is used. The fishing net is cast into a lake, left for a period of time, and then pulled back in. The probability of catching a specific fish species improves with its weight, but it is impossible to know how many fish will be caught in advance. Based on the differences between these two distributions, I chose to utilize the Multivariant version of the FNH (MFNH) due to no dependence between samples, as the fishing net metaphor fits the single-cell sequencing scenario to a greater extent than the WNH.

Parameters and Functions of the Fisher Noncentral Hypergeometric Distribution

The FNH distribution is parameterized as follows: for a simple univariate FNH distribution, it is characterized by two populations, denoted as m_1 and m_2, both of which are non-negative integers ($m_1, m_2 \in \mathbb{N}$). The total number of elements in the populations is represented by N, which is the sum of m_1 and m_2, i.e., $N = m_1 + m_2$. Each population has a weight, denoted as ω, which is a positive real number ($\omega \in \mathbb{R}_+$). Where an increase in the weight increases the probability of drawing a specific population. When the weights are close to one, the FNH distribution collapses into the hypergeometric distribution. Finally, there are n samples to be drawn x is the number of balls from the m_1 population sampled, where n is an integer within the range $[0, N)$.

$$P(x) = \frac{\binom{m_1}{x}\binom{m_2}{n-x}\omega^x}{P_0} \tag{5.1}$$

The terms P_0, P_1, and P_2 are used in the calculation of the mean and variance of the FNH distribution. They are defined as follows:

$$P_k = \sum_{y=x_{\min}}^{x_{\max}} \binom{m_1}{y}\binom{m_2}{n-y}\omega^y y^k \tag{5.2}$$

The mean of the FNH distribution, denoted as $\frac{P_1}{P_0}$, is calculated using P_1 divided by P_0. The variance of the FNH distribution, denoted as $\frac{P_2}{P_0} - \left(\frac{P_1}{P_0}\right)^2$, is calculated using P_2 divided by P_0 minus the square of $\frac{P_1}{P_0}$. Please note that $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of x in the summation.

The Multivariant Fisher Noncentral Hypergeometric distribution (MFNH) is a generalization of the Fisher Noncentral Hypergeometric distribution to the case where $K > 2$, with K being the number of populations in the Biased Urn. The parameters of the MFNH distribution are denoted by $\mathbf{m} = (m_1, \ldots, m_k) \in \mathbb{N}^k$, where m_k represents the number of balls in population k. Additionally, the total number of balls in the Biased Urn is given by $N = \sum_{k=1}^{K} m_k$, and the number of balls sampled from the Biased Urn is denoted by n where $n \in [0, N)$. The weight of each population is represented by $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k) \in \mathbb{R}_+^k$.

The Probability Mass Function (PMF) for the MFNH is defined as follows:

$$P_0 = \sum_{(y_1,\ldots,y_k)\in \mathrm{S}} \prod_{k=1}^{K} \binom{m_k}{y_k} \omega_k^{y_k} \tag{5.3}$$

Where $P(x_1, \ldots, x_k)$ is the joint probability of observing x_k balls from population k, and is given by:

$$P(x_1, \ldots, x_k) = \frac{1}{P_0} \prod_{k=1}^{K} \binom{m_k}{x_k} \omega_k^{x_k} \tag{5.4}$$

The mean of the MFNH distribution, denoted by μ_k, is calculated as follows:

$$\mu_k = \frac{m_k r \omega_k}{r\omega_k + 1} \tag{5.5}$$

where r is the unique positive solution to the equation:

$$\sum_{k=1}^{K} \mu_k = n \tag{5.6}$$

These equations define the MFNH distribution and allow for the analysis of biased sampling processes involving multiple populations.

5.2.2 Modelling Cell-Specific Parameters

To model the various sources of variance within a single-cell experiment requires a ground truth to start from, which in Minerva starts with estimating cell specific parameters. Within the modelling framework, there are five key parameters for cell j: the excepted mRNA content of the cell population p that cell j belongs to, capture chemistry efficiency c_j, RT-PCR efficiency r_j, its sequencing saturation s_j, μ_{ip} and Φ_{ip} which is, given gene i, mean expression and dispersion, for a given cell population p.

Estimating Excepted mRNA Size of Cell Populations

Minerva treats cells as samples from a cell population and that the mRNA content of cells from a given cell population will (the total number of transcripts within a cell) randomly fluctuates around the expected mRNA content of the population. This biological characteristic is of significant importance as it can have a profound impact on the observed counts and subsequent processes in a single-cell experiment [136, 18]. To identify cell populations from a given dataset, Minerva employs the state-of-the-art single-cell quality control (QC), normalization, and cell population identification pipeline described by Luecken and Theis.

Briefly, QC is performed by filtering out cells with a mitochondrial percentage of UMIs greater than 10% and cells with library size larger than three median absolute deviations. Once the filtered cells' library sizes are normalized, Minerva utilizes the pooled size factor normalization method developed by L. Lun et al.. Next, the genes exhibiting the top 10% of variance are selected to identify cell populations and undergo PCR to reduce the dimensionality to 50 dimensions. Once compressed, a K-nearest neighbours (KNN) graph is constructed based on the Euclidean distance between cells in the reduced space. Finally, the Louvain community detection algorithm is applied to identify the cell populations [91, 15].

Once the cell populations have been identified, I employ a slightly modified version of the heuristic developed by Ye et al. to estimate the expected mRNA content of each cell population. Instead of estimating cell-specific sampling efficiencies, I estimate the expected mRNA content of the cell population itself.

To estimate the expected cell library size of a population (L_p), which serves as a proxy for the mRNA content of that population relative to others, I make the following assumptions. Suppose a cell library is smaller or larger than other cell populations. In that case, it indicates that the mRNA content of the cell population is expected to be smaller or larger compared to the others. Based on this assumption, I estimate the simulated mRNA content of a cell population using the user-defined minimum mRNA content (C_{min}) and maximum mRNA content (C_{max}) in conjunction with the library size.

First, I perform a logarithmic transformation (base 10) on all expected library sizes and determine the minimum observed library size (O_{min}) and the maximum observed library size (O_{max}). Then, I calculate a library size weight using the equation shown in equation 5.7 to obtain the library weight (l_w) for each cell population. Finally, I estimate the expected mRNA content ($\mathbf{E}[C_p]$) of a cell population using equation 5.8, as demonstrated in the equations below:

$$l_w = \frac{E[L_p] - O_{min}}{O_{max} - O_{min}} \tag{5.7}$$

$$\mathbf{E}[C_p] = (1 - l_w) * C_{min} + l_w * C_{max} \tag{5.8}$$

Sampling Cell Specific Capture Chemistry efficiency

The capture chemistry efficiency of a cell is considered to be a random variable that is not specific to a particular cell population. Instead, it is a technical variable that exhibits random fluctuations around the expected capture chemistry efficiency of a specific single-cell sequencing protocol. In order to simulate this process, Minerva utilizes sampling from a beta distribution. The parameters α_C and β_C are user-defined and determine the shape of the beta distribution. The expectation of the beta distribution should be the reported or desired value of the expected capture chemistry efficiency of the experiment.

$$c_j \sim Beta(\alpha_C, \beta_C) \tag{5.9}$$

Sampling Cell Specific RT-PCR

The RT-PCR efficiency of a cell is considered to be a random variable that is not specific to a particular cell population. Instead, it is a technical variable that exhibits random fluctuations around the expected RT-PCR efficiency of the RT-Polymerase used in the experiment. To simulate this process, Minerva employs sampling from a beta distribution, which is parameterized by user-defined parameters α_R and β_R. By default, these parameters are set to 18, 18, resulting in a beta distribution with an expectation of 0.5. While users have the flexibility to modify these parameters, it is worth noting that manipulating RT-PCR efficiency is not recommended unless you are simulating an improved RT-PCR.

$$r_j \sim Beta(\alpha_R, \beta_R) \tag{5.10}$$

Sampling Cell Specific Sequence Saturation

For a given cell, cell-specific sequencing saturation occurs, where sequencing saturation represents the percentage of unique transcripts observed in the counts obtained from the sequencing. I consider cell-specific sequencing saturation as a randomly fluctuating variable. However, unlike capture chemistry efficiency and RT-PCR, I simulate it based on the sampling efficiency of the cell. The sampling efficiency (ρ_j) of cell j is calculated by dividing the cell's library size (L_j) by its population's expected mRNA content (C_p), as shown in equation 5.11. Using the previously sampled values of c_j and r_j, I calculate a cell's sequencing saturation (s_j) as the ratio of ρ_j over the product of c_j and r_j, as demonstrated in equation 5.12.

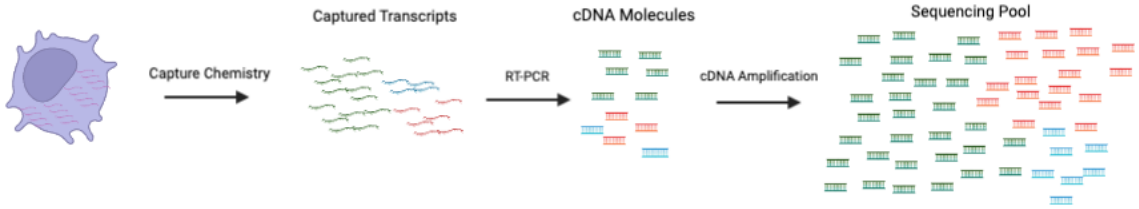


Fig. 5.1 This illustration depicts a step-by-step cell workflow, starting from cell lysis and capturing mRNA transcripts, leading to the creation of the final end sequencing pool. Throughout the process, there is a progressive reduction in the amount of information as the number of transcripts decreases at each step. Notably, RT-PCR, a common technique, typically results in approximately 50% loss of captured transcripts. Following that, the successfully converted cDNA molecules are labelled with UMIs (Unique Molecular Identifiers) and amplified using PCR, generating the sequencing pool from which the reads are ultimately sequenced.

$$\rho_j = \frac{L_j}{C_p} \tag{5.11}$$

$$s_j = \frac{\rho_j}{c_j * r_j} \tag{5.12}$$

Estimating Gene-Specific Parameters for a Given Cell Population

To estimate the gene expression distributions for a given cell population, Minerva employs methods that have been previously utilized by crisprPower in Chapter 3. For more detailed information on this, please refer to the 'Fitting Genes Mean and Dispersion' subsection of Chapter 3's methods and Algorithm 2.

Once the mean expression and dispersion values are estimated, Minerva proceeds to scale the mean expression so that the sum total of all genes' mean expression is equal to the expected mRNA content. This scaling enables the biological variance samples to fluctuate randomly around the expected mRNA content cell population, thereby providing a more realistic simulation of transcriptional behaviour in a single-cell experiment.

5.2.3 Theoretical Model of Single Cell Experiments

Single-cell experiments are an invaluable approach for analyzing gene expression of cell populations in heterogeneous biological settings and processes. However, the outcomes of single-cell experiments can be affected by a diverse set of sources of variability [142, 91] (see Figure 5.1). This requires the use of a hierarchical model that effectively captures the different levels of biological and technical variance. In addition, to variability, it is important

to model the specific type of sampling without replacement that occurs in these experiments, where all transcripts are sampled simultaneously without any dependence between draws [139, 18, 136]. To develop a precise theoretical model for single-cell experiments, it is essential to thoroughly consider and address each source of variability and sampling bias within Minerva's model.

The first crucial aspect to consider is biological variability, which encompasses the inherent stochastic differences between cells and represents the true expression distribution Iaim to estimate in an experiment. To model this variability, I utilized a Gamma distribution, its extensive use in modelling biological variance in both differential analysis and simulator methods, as demonstrated in previous studies [8, 90, 96, 139, 14]. The Gamma distribution offers significant flexibility, allowing it to effectively capture variations in both highly and lowly expressed genes, regardless of whether they have large or small variances. Once the gamma distribution parameters for all genes are estimated from single-cell data, Minerva scales them to a given cell population's expected mRNA content, thereby enabling these gamma distributions to simulate true transcript counts.

The second crucial factor to consider is the technical variability originating from the capture chemistry step in single-cell experiments. Capture chemistry is the step of a single-cell experiment that is continuously being improved upon. For example, previous versions like 10xv1 chemistry captured only 10% of the mRNA molecules, while the current 10xv3 chemistry captures approximately 30% [142]. Capture chemistry involves a sampling without replacement process within the experimental protocol, where only a fraction of the mRNA molecules within a given cell is captured. Furthermore, it serves as the initial step in the experiment where enrichment or amplification techniques can be applied, potentially introducing a weighted sampling process.

To accurately model the sampling process in the capture chemistry step of single-cell experiments, the chosen sampling distribution must accommodate sampling without replacement while allowing for the potential application of weights. While the Hypergeometric distribution has been used previously, it does not precisely reflect the sampling process that occurs during capture chemistry and lacks the flexibility of weighted samplings [14]. However, the MFNH distribution fits Minerva's use case as it captures the sampling without replacement process in which the draws are independent, and it can account for variations in the probability of sampling different classes within the specified distribution by utilizing a 'weight' parameter. Where increasing the weight increases the probability of sampling the corresponding transcript [45, 46]. By utilizing the MFNH distribution, Minerva can effectively model both the sampling without replacement process in the experiment and the manipulation of a gene's probability of being sequenced.

The next significant source of variation to consider is the RT-PCR step, where captured mRNA molecules are converted into cDNA for downstream amplification in single-cell experiments. RT-PCR is a critical and pivotal step in the experiment, and it is the primary cause of sparsity in the resulting data. This inefficiency leads to the degradation of approximately 50% of the captured molecules, resulting in a significant reduction in the counts, which affects the detection of low-expressed genes. To model this step, Ican employ a binomial downsampling approach [18, 136]. Where n of represents the count of captured molecules for a given gene, and p represents the RT-PCR efficiency. By sampling from a binomial distribution with these parameters to simulate the sequencing pool count, where the sampled value indicates the number of unique molecular identifiers (UMIs) that remain for cDNA amplification and subsequent sequencing.

Finally, cDNA amplification bias is another source of variability that arises from biases introduced during the PCR amplification step, leading to differences in cDNA production and gene expression measurements. To simulate the impact of cDNA amplification and sequencing, it is important to consider the nature of the core reaction driving both of these steps is PCR. PCR exhibits stochastic exponential behaviour, which introduces variability in its efficiency on a per-gene basis, influenced by factors such as GC content and temperatures [71]. In the context of single-cell sequencing experiments, the presence of unique molecular identifiers (UMIs) attached to mRNA molecules ensures the deduplication of reads and the identification of unique transcripts [64, 94]. However, issues arise when trying to detect poorly amplified transcripts due to the Polya process PCR follows, which can result in a decrease in the number of UMIs observed for the affected transcripts, making them more challenging to detect within a given cell [110]. Another important aspect to consider during cDNA amplification which is the second opportunity in a single cell experiment to introduce customized amplification and enrichment protocols, similar to the capture chemistry step discussed earlier. To capture these effects and potential experimental alterations, I once again utilise the MFNH distribution. Using the MFNH distribution, Minerva can incorporate weights to represent experimental alterations applied during cDNA amplification. Higher primer concentrations, for example, can increase the odds of a transcript being sequenced, providing a more accurate representation of the single-cell experimental process.

The resulting hierarchical model utilized by Minerva is a Gamma-MFNH-Binomial-MFNH model. Each individual distribution captures both biological and technical processes occurring at various steps of a single-cell experiment:

$$
\begin{aligned}
X_j &\sim \Gamma\left(\Phi_i^{-1}, \mu_i \cdot \Phi_i\right) \\
M_j &\sim MFNH(n = P_j, m = X_j, o = M_w) \\
S_j &\sim \text{Binom}(n = M_j, p = r_j) \\
Y_j &\sim MFNH(n = L_j, m = S_j, o = S_w)
\end{aligned}
$$

This model accounts for the estimation of biological variance that researchers aim to quantify, as well as the potential experimental manipulations that can take place during different stages of the experiment.

Modelling Biological Variability

Let N be the number of genes and M be the number of cells to simulate, describing the entire count matrix or single experimental condition. Let X_{ij} be a random variable representing the expression level of gene i in cell j $(i = 1,...,N; j = 1,...,M)$. Let Y_{ij} be a random variable representing the count value (read or UMI count) of gene i in cell j. In a real scenario, only y_{ij} (the observed value of Y_{ij}) is known, while x_{ij} (the realization of X_{ij}) is unknown and to find it is often the primary objective in a scRNA-seq experiment.

In Minerva, a given gene i, the expression levels X_{ij} are modelled using a gamma distribution. Where μ_i is the "average" expression level of gene i and Φ_i is a parameter describing the biological variability in the expression level of gene i:

$$X_{ij} \sim \Gamma(\frac{1}{\Phi_i}, \mu_i \cdot \Phi_i) \tag{5.13}$$

Modelling Capture Chemistry

Next, the previously sampled biological variance for cell j is sampled to create its capture molecule pool M_j using the MFNH parameterized by the X_j gene expression vector, the size of cell j capture molecule pool P_j which is calculated in the following manner $P_j = \sum X_j * c_j$, and M_w, which is a vector containing weights that represent any potential experimental enrichment of a particular set of transcripts being applied at the capture chemistry step of the simulation. By default, M_w contains all one, which simulates no biasing of the sampling process. Sampling the MFNH creates the vector M_j, which is the Captured Pool of X_j. This can be expressed formally as:

$$M_j \sim MFNH(n = P_j, m = X_j, o = M_w) \tag{5.14}$$

Modelling Reverse Transcription of mRNA

The captured transcripts in M_j are now transformed into a cell j sequencing pool, denoted as S_j. This pool represents the number of transcripts successfully converted into cDNA molecules. To simulate this conversion process, I sample cell j capture pool vector M_j using a Binomial distribution. Here, the number of trials denoted as n, corresponds to the individual gene's molecule count M_{ij}, while the success probability, denoted as p, is parameterized by the cell j RT-PCR efficiency r_j. Itake one sample per gene, and the resulting sampled value represents the number of mRNA molecules for a specific gene creating the sequence pool vector S_j which represents mRNA converted to cDNA, as shown in the following equation:

$$S_j \sim Binom(n = M_j, p = r_j) \tag{5.15}$$

Modelling cDNA Amplification

Finally, the cDNA molecules in the S_j are sampled to create cell j observed counts Y_j. To do this sample, an MFNH distribution that is parameterised so that S_j are the populations, L_j is the number of samples to draw, which is calculated by $L_j = \sum S_j * s_j$, and S_w is a weight vector representing any potential experimental enrichment of a particular set of transcripts being applied during cDNA amplification. By default, S_w contains all ones, which simulates no biasing of the sampling process. This is expressed formally as:

$$Y_j \sim MFNH(n = L_j, m = S_j, o = S_w) \tag{5.16}$$

Simulating Target Transcriptomes

Targeted transcriptomes are simulated by modifying the S_w during the cDNA amplification step in Minerva. When a gene is part of a panel, its weight is set to a default value of 30. If a gene is not part of the targeted panel, its weight is set to 1. This approach allows for flexibility and enables Minerva to simulate slight off-target effects, although these effects may differ and vary for each gene from the weights currently used. The default weights of 30 for target panel genes were selected utilising Mean Square Error to assess the distance between the log-transformed observed means of Minerva's simulated transcriptomes and the observed mean expression of targeted panel data from Schraivogel et al. (see Fig 5.2).

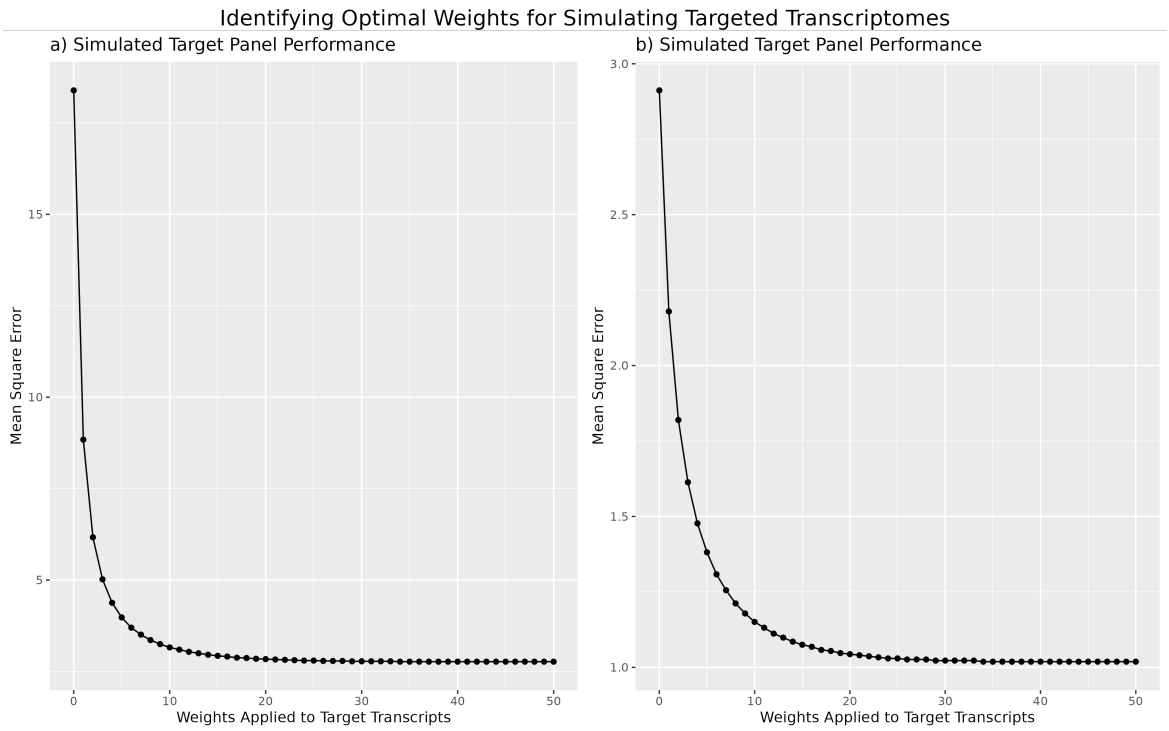


Fig. 5.2 Identifying the Optimal Weights for Simulating Target Transcriptomes via Mean Square Error Comparison between Minerva's Observed Mean Counts and Schraivogel et al. l1000 Target Panels Observed Mean Counts. a) Displays all of the weights tested from 1 to 50 and b) focuses along the y-axis to show the elbow stops between the weights 25 and 30.

Simulating Weighted Transcriptomes

In Minvera's model, there are two steps in the single-cell sequencing experiment where enrichment protocols can be applied: capture chemistry and cDNA amplification. Depending on the type of experiment being planned to simulate, it will alter how and when gene-specific odds are applied. To compute the weights for a weighted transcriptome, it is necessary to have either a reference dataset or a pilot experiment from which an observed relative frequency can be obtained. Currently, the weighted transcriptomes are calculated using a target frequency, denoted as t. This target frequency is determined as the reciprocal of the number of non-zero expressed genes ($t = \frac{1}{nGenes}$) in the pilot experiment or reference dataset.

To simulate a weighted transcriptome where only one of the steps is to conduct either weighted chemistry or library experiment. In this case, the weight for gene w_g is calculated by dividing the desired target relative frequency by the observed relative frequency of gene f_g. The relative frequency of a gene is calculated by dividing a given gene's mean expression μ_g by the sum of all genes' mean. This can be calculated rather straightforwardly using the following equations:

$$f_g = \frac{\mu_g}{\sum_{i=1}^{n} \mu_i} \tag{5.17}$$

$$w_g = \frac{t}{f_g} \tag{5.18}$$

To calculate gene-specific weight for an experiment that weighs All Steps (Capture Chemistry and cDNA Amplification) of a single-cell experiment, I first calculate the weighted mean (M_w) as described above. Next, in order to determine the weights during cDNA amplification (S_w), an estimate of the expected number of captured molecules ($\mathbf{E}[S_j]$) is required. To calculate this, multiply the expected number of captured molecules in $\mathbf{E}[M_j]$ by 0.5, which represents the expected RT-PCR efficiency. Subsequently, use the formulas 5.17 and 5.18 to calculate S_w.

5.2.4 Evaluation of Model Performance

Datasets

To assess the influence of alternative single-cell experimental protocols, particularly those involving weighted transcriptomes, on the observed count data and our ability to capture the true expression distribution, simulations were performed on four experimental protocols using seven single-cell datasets from diverse biological contexts. For the purposes of this chapter and Chapter 6, the datasets in Table **??** were used.

Dataset Name	Biological Setting	Protocol	Cell Count	Chapter
BaronData [12]	Pancreatic islets	inDrop	8569	5 & 6
MacoskoData [94]	Retina cells	Drop-Seq	49300	5 & 6
JessaData [65]	Pediatric brain tumors	10x	61595	6
KotliarovData [76]	Immune system	10x v2	58654	6
WuData [137]	Adult kidney	snDrop-Seq	17542	5 & 6
ZhaoData [141]	Immune cells	10x	68100	6
ZeiselData [140]	Nervous system	10x v1	160796	6
ZilionisData [145]	Tumor-infiltrating myeloids	inDrop	173954	6

Table 5.1 Datasets Used for Simulating Data: Protocol, Biological Setting, Cell Count, and Chapter (5 or 6).

Assessment Metrics

To evaluate the performance of Minvera and compare it to count matrices generated by other simulators and real datasets, I employed comparison methods similar to those used in Baruzzo et al.. The performance was assessed by comparing the statistical properties of simulated counts with real count matrices in terms of gene sparsity, dispersion, and log-transformed mean expression on a per dataset and cell population basis.

Gene sparsity was measured as the fraction of zeros observed in cells belonging to a specific cell population. Gene dispersion was calculated as the variance of the normalized counts for each gene within a cell population. Finally, genes' mean expression values were calculated and subsequently log-transformed.

For the analysis of gene sparsity and dispersion, I estimated the distributions using KDE and visualised them using violin plots. These plots effectively illustrate the variations in these statistical characteristics across different simulators and real datasets. Additionally, I generated a two-dimensional KDE estimate to explore the relationship between sparsity and log mean expression.

Evaluating Minerva Against the State-of-the-Art

I compared the performance of Minerva with two other state-of-the-art simulators: SPARSim [14] and Splatter [139]. Both of these simulators were used with their default parameters.

Minerva was directly inspired by SPARSim, as it was developed to improve the limitations of SPARSim by simulating more of the technical steps of single cell experiments and allowing for experimental manipulation [14]. SPARSim is a simulator that uses a Gamma-Multivariant Hypergeometric Distribution to model single-cell data, with a gamma distribution for biological variance and a multivariant hypergeometric distribution for technical variability.

Splatter, a well-known and widely used scRNA-Seq simulator based on a Gamma-Poisson hierarchical model [139], has been widely employed in the field. However, a notable limitation of Splatter is its inability to account for the dynamics of a sampling process, which both SPARSim and Minerva can naturally accommodate. In Splatter, gene mean expression levels are simulated from a Gamma distribution, and the corresponding count values are generated using a Poisson distribution. The simulator also incorporates high-expression outlier genes and enforces a mean-variance trend, previously utilized in bulk RNA-Seq simulations [139]. Additionally, Splatter employs a logistic function to describe the relationship between gene expression level and sparsity per gene, enabling the simulation of dropout events.

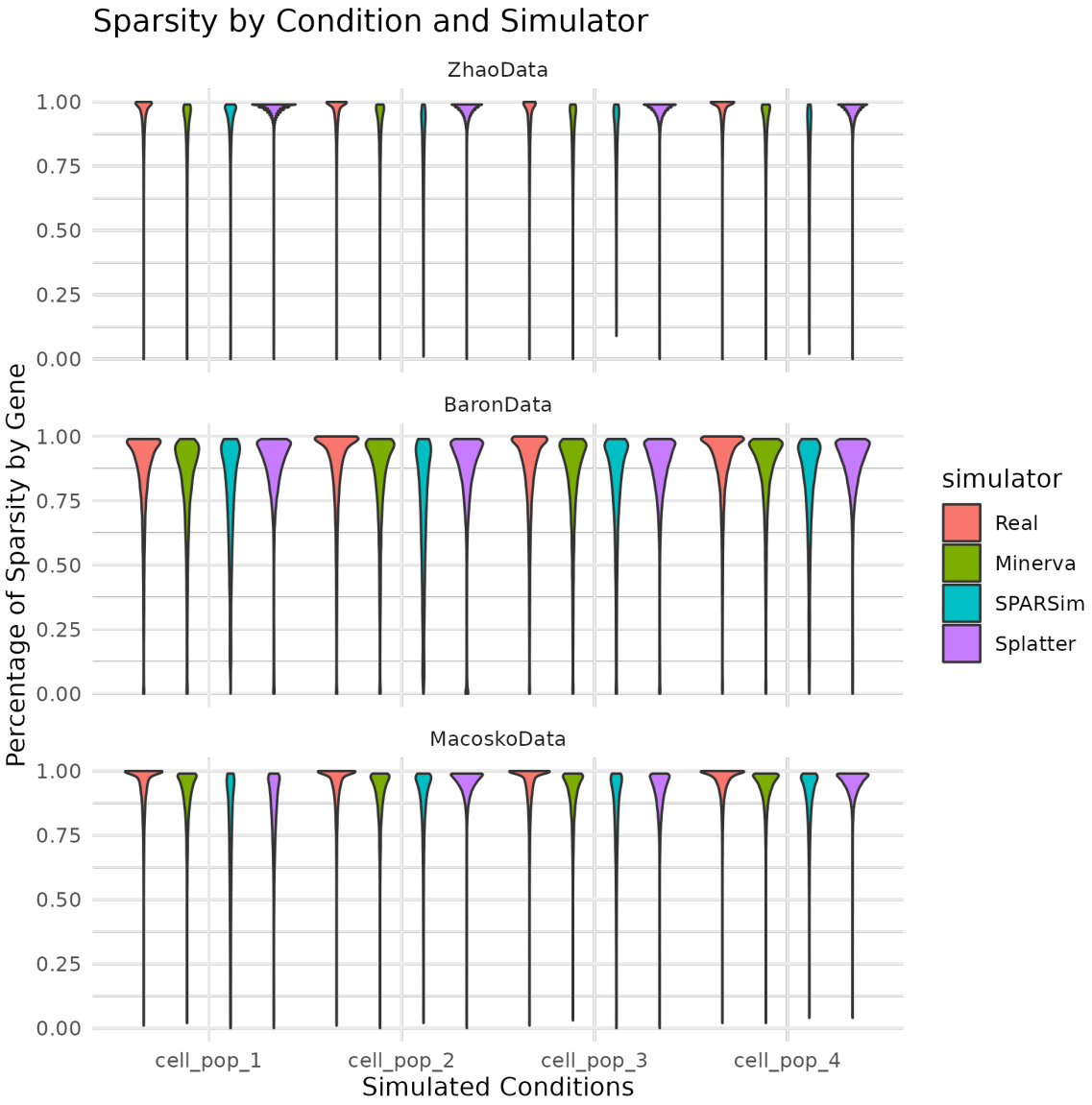


Fig. 5.3 Comparison of gene sparsity between simulators (Minerva, SPARSim, and Splatter) and real scRNA-Seq datasets across multiple datasets and cell populations. Violin plots show the distribution of sparsity for each dataset and cell population.

5.3 Results

5.3.1 Statistical Comparison of Minerva to Datasets and Other Simulators

scRNA-Seq simulations are a valuable tool for researchers studying gene expression at the single-cell level. However, accurately capturing key features of real scRNA-Seq data is

essential for simulations to be useful. In this study, I performed a statistical comparison of Minerva, a novel scRNA-Seq simulator that I developed, to two other simulators: SPARSim and Splatter. My primary objective was to assess the performance of these simulators in capturing three key relationships found in real scRNA-Seq data: the distribution of sparsity, the distribution of the genes dispersion, and the relationship between a gene's sparsity and mean expression (which decreases in sparsity as mean expression increases). The specific metrics used for comparison were a series of violin plots, which quantified and visualized the representations of the relationships.

Minerva consistently captures the distribution of sparsity by gene per cell population across various datasets, as evidenced by the near-identical violin plots and substantial overlap with real datasets. This ability to accurately model gene sparsity is also observed in SPARsim and Splatter. Nevertheless, the fidelity of simulated data sparsity is influenced by factors such as sequencing depth, capture chemistry, and biological context. Notably, when utilizing the Zhao et al. data as input, none of the simulators perform exceptionally well when compared to real data. SPARsim and Minerva tend to underestimate sparsity, whereas Splatter tends to overestimate it (see Figure 5.3, first row). Conversely, in the Baron et al. dataset, all simulators accurately reflect gene sparsity (see Figure 5.3, third row). These findings collectively demonstrate that Minerva is capable of simulating gene sparsity across diverse biological settings; however, the quality of the dataset significantly impacts its simulation capability.

The second comparison focused on gene dispersion, which quantifies the biological variability of genes. In this comparison, both Minerva and SPARSim outperformed Splatter significantly. Splatter consistently displayed an overestimation of gene dispersion or exhibited a distinct bimodal distribution with a cluster of high dispersion and another cluster with no gene dispersion. This is evident in its distribution, which consistently deviated from the real data and the other simulators across all datasets and cell populations (see Figure 5.4). While Minerva and SPARSim exhibited comparable performance overall, there were instances where Minerva's performance appeared slightly superior to SPARSim's. Specifically, across most cell populations of Macosko et al., SPARSim consistently underestimated gene dispersion (third row of Figure 5.4). Conversely, Minerva did not exhibit any pronounced discrepancies in modelling gene dispersion compared to other simulators, although minor differences were observed between the simulated gene dispersion of Minerva and the real data in Zhao et al. (see the first row of Figure 5.4). These findings suggest that Minerva accurately captures gene dispersion compared to real datasets and other simulators.

The final comparison delves into the relationship between gene sparsity and mean expression, which is a crucial aspect of scRNA-Seq data. To evaluate the performance of the

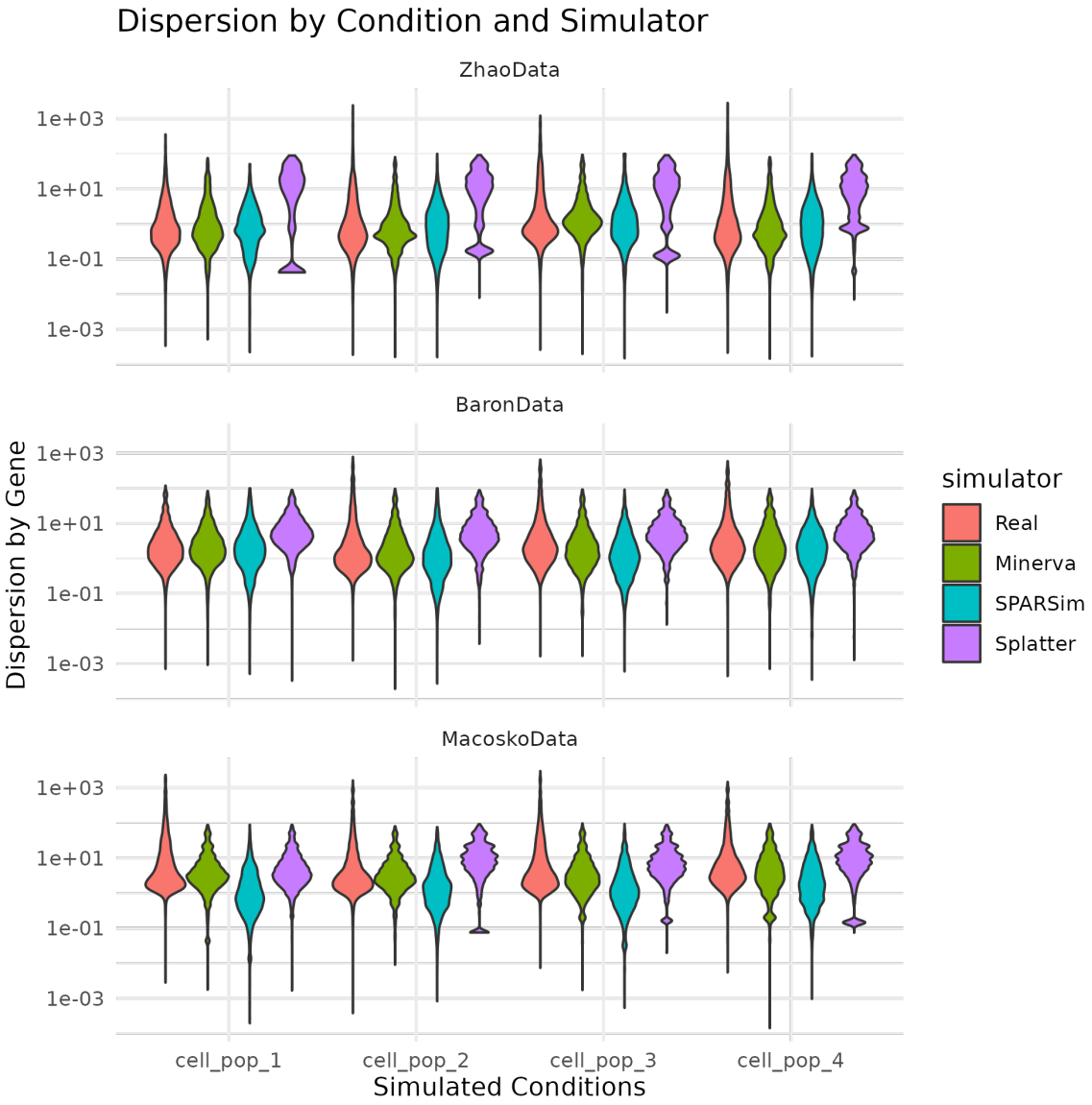


Fig. 5.4 Comparison of gene dispersion between simulators (Minerva, SPARSim, and Splatter) and real scRNA-Seq datasets across multiple datasets and cell populations. Violin plots show the distribution of gene dispersion for each dataset and cell population.

three simulators, I utilized kernel density estimates for each dataset and cell population, representing the probability of a given value through a series of topologically shaded colours (as depicted in figure 5.5). This analysis reveals that, once again, while Splatter roughly captures the overall characteristics, it does not reproduce them perfectly or as effectively as Minerva. Splatter's kernel density is concentrated and shifted towards higher sparsity and lower mean expression per gene compared to real datasets. In contrast, Minerva and SPARSim closely resemble the kernel distribution of real data. However, SPARSim's kernel

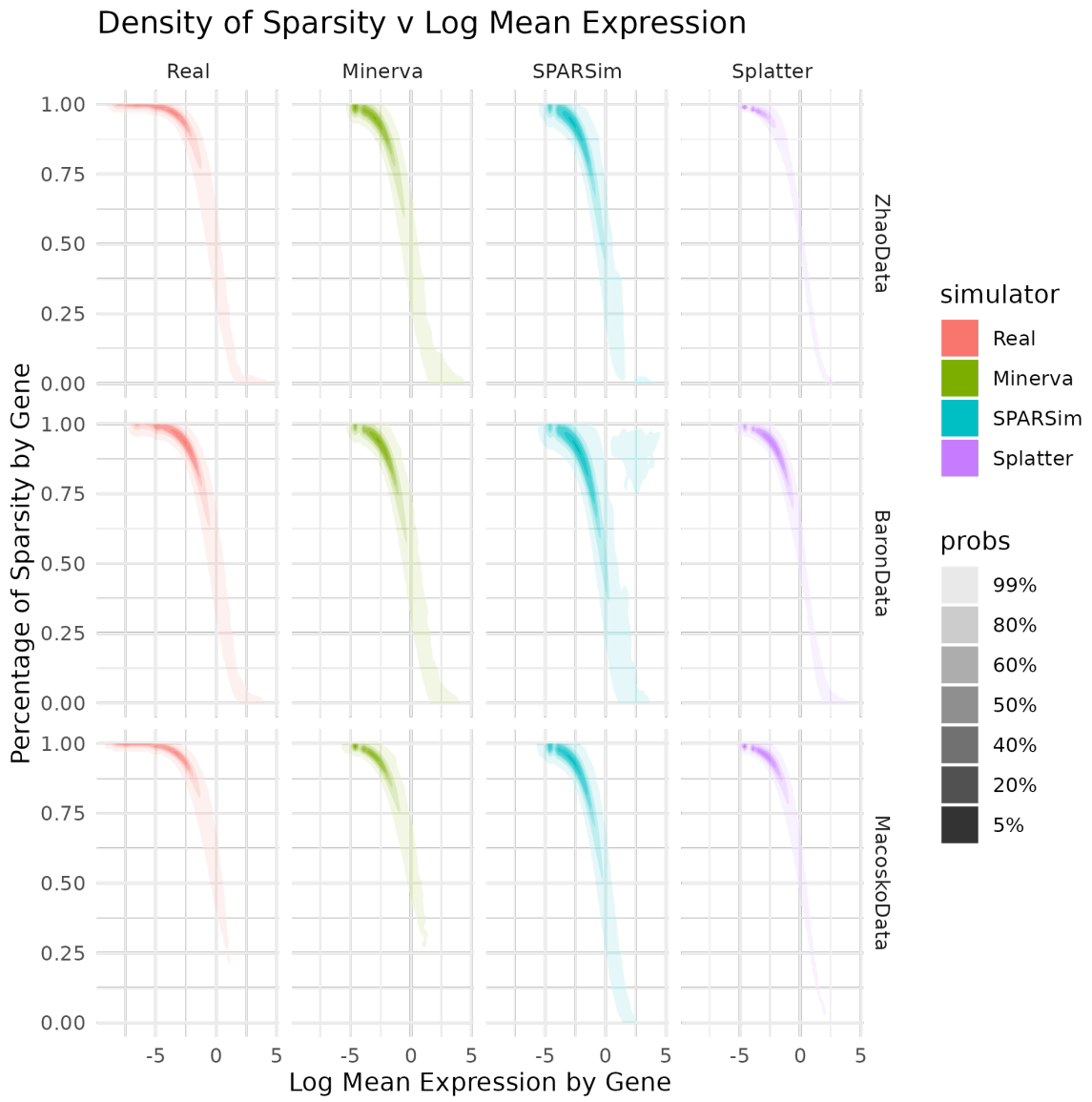


Fig. 5.5 Comparison of the relationship between gene sparsity and mean expression between simulators (Minerva, SPARSim, and Splatter) and real scRNA-Seq datasets across multiple and cell populations. 2-dimensional kernel density plots show the distribution of these variables for each dataset and cell population.

exhibits a set of unexplained outliers in the Baron et al. Baron et al. and Macosko et al. datasets, while Minerva does not have such outliers (see the second row of figure 5.5). Moreover, all simulators overestimate the majority of the mean expression distribution in comparison to real data for Zhao et al.. However, each simulator demonstrates different behaviour: Splatter significantly underestimates mean expression, while both Minerva and SPARSim overestimate it, although Minerva's overestimation is the closest to the real data

among the three (see the first row of figure 5.5). These findings affirm that Minerva adeptly models the relationship between gene sparsity and mean expression across diverse biological settings and experimental contexts.

These results demonstrate that Minerva is a highly effective simulator for scRNA-Seq data, accurately capturing key features such as sparsity distribution, gene dispersion, and the relationship between gene sparsity and mean expression. Minerva performed on par with its primary rival SPARSim, and surpassed the state-of-the-art simulator Splatter across these major metrics. These findings suggest that Minerva is a reliable and accurate tool for researchers who wish to simulate gene expression at the single-cell level.

5.3.2 Statistical Comparison of Minerva Targeted Transcriptome to Datasets

Targeting panel sequencing is an innovative scRNA-Seq technique developed to enhance the representation of genes expressed at low levels. By concentrating cDNA amplification and sequencing on a specific gene panel, this technique enables comprehensive characterization of their expression distribution. To the best of my knowledge, Minerva stands as the sole simulator capable of effectively simulating this data type. Here I examine the ability of Minerva to simulate gene sparsity, dispersion, and the relationship between sparsity and mean expression using whole transcriptome data obtained from Schraivogel et al.. In this study, targeted transcriptome data was paired with whole transcriptome data, which enabled me to compare Minerva's simulated targeted transcriptomes to a ground truth. The primary objective is to validate Minerva's capability to accurately reproduce the statistical characteristics specific to targeted transcriptome data derived from whole transcriptome datasets.

Initially, I conducted a comparison between the gene dispersion distribution of real targeted transcriptome data and Minerva's simulated targeted transcriptome using the paired whole transcriptome as a reference (see Figure 5.6a). The results depicted in Figure 5.6 indicate that Minerva's simulated targeted transcriptome successfully generates realistic gene sparsity when compared to the real targeted transcriptome data. Both the Minerva simulated and targeted transcriptome data display a significant reduction in sparsity, which aligns with the primary objective of targeting panel sequencing. However, there are discernible differences. Specifically, Minerva's simulated targeted transcriptome exhibits an elongated violin plot with a lower density of genes exhibiting near 100% sparsity. This observation suggests that Minerva's approach to simulating targeted transcriptomes, particularly in terms of probe efficiency and off-target behaviour, may differ from real targeted transcriptomes.

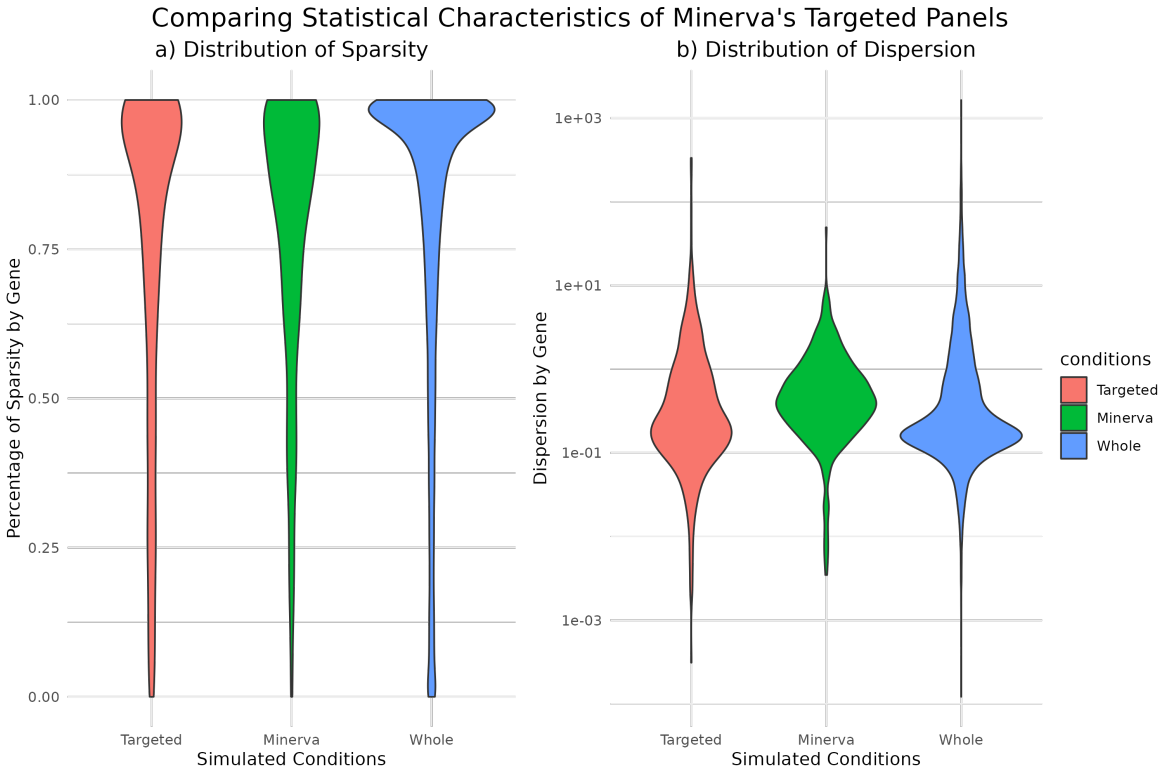


Fig. 5.6 Comparison of gene dispersion between real TAP-Seq data, whole transcriptome scRNA-Seq data, and simulated data from Minerva. Violin plots show the distribution of gene dispersion for each condition.

Next, I conducted a comparison between the gene dispersion in real targeted transcriptome data, whole transcriptome data, and the simulated targeted transcriptome generated by Minerva. The results of this comparison revealed excellent performance from Minerva, as the simulated data nearly perfectly matched the targeted transcriptome data. The violin plots in Figure 5.6b show that Minerva's simulated data exhibit a slightly higher dispersion estimate than the real targeted data, which is likely due to miscalibration in the off-target effects and probe efficiency during the simulation process with Minerva. Despite these slight differences, both the simulated and targeted panels exhibited an overall increase in gene dispersion. This increase can be attributed to a more accurate estimation of a gene's mean expression, usually resulting from a better representation of the gene due to an increase in the number of detected UMIs. Consequently, more of the expression distribution can be observed. These results demonstrate the effectiveness of Minerva in simulating targeted panel sequencing data and highlight the potential benefits of this technique for improving the accuracy of gene expression measurements.

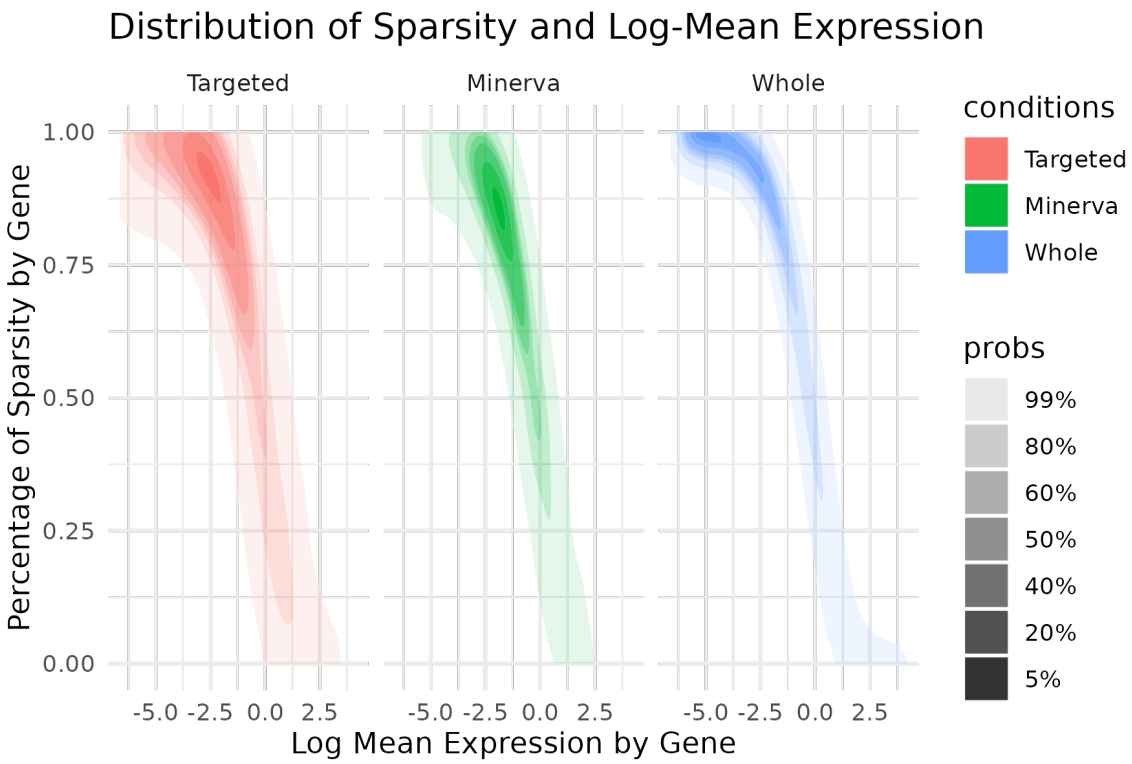


Fig. 5.7 Comparison of the relationship between gene sparsity and mean expression between real TAP-Seq data, whole transcriptome scRNA-Seq data, and simulated data from Minerva. 2-dimensional kernel density plots show the distribution of these variables for each condition.

Finally, I examined the relationship between gene sparsity and mean expression in the targeted transcriptome, both real and simulated, in comparison to the whole transcriptome. To accomplish this, I generated 2-dimensional kernel density plots. The results of this analysis, illustrated in Figure 5.7, reveal a substantial overlap in density between the Minerva simulated data and the targeted transcriptome data. Furthermore, significant differences are observed between both the real and simulated targeted transcriptomes when compared to the whole transcriptome. It is important to note, however, that Minerva tends to overestimate mean expression, and the reduction in sparsity is only slightly apparent, as indicated by the slight shift to the left and downward when compared to the real targeted transcriptome data.

Overall, our comparisons in this subsection demonstrate the effectiveness of Minerva in simulating targeted panel sequencing data and the potential benefits of this technique for scRNA-Seq studies. Minerva can accurately capture the relationships between gene targeting, sparsity, dispersion, and mean expression, making it a valuable tool for researchers studying gene expression at the single-cell level.

5.3.3 Weighted Sampling of Gene Transcripts

Now that Minerva has been validated and proven to accurately replicate the statistical properties observed in current single-cell experiments and binary manipulations of sequencing probabilities in targeted transcriptomes, the next feature of Minerva to explore are the various theoretically possible single-cell experiments involving continuous manipulations of sequencing probabilities. One of the significant challenges faced in current single-cell experiments is the technically-induced gene sparsity, which disproportionately affects genes expressed at low to medium levels. Unfortunately, many biologically interesting gene types, such as transcription factors, kinases, receptors, and others, fall into this category. As a result of their low expression, these genes often remain concealed, making it arduous for single-cell experiments to accurately characterize them. The hierarchical model used by Minerva for simulating each technical step of the experiment reveals two significant stages: Capture Chemistry and cDNA Amplification, during which the sequencing probabilities can be altered. These stages give rise to a range of theoretically possible experimental protocols, namely Non-Weighted, Weighted Library, Weighted Chemistry, and Both Steps Weighted.

A Non-Weighted experiment corresponds to a standard single-cell sequencing approach that is presently in use. Weighted Library involves manipulating the ratio of PCR amplified reads, achieved through antibody pull-down or custom PCR probes. This method focuses on adjusting the read-to-UMI ratio to increase or decrease the probability of sequencing a specific UMI-labeled transcript. In a Weighted Chemistry experiment, the manipulation relies on probe-based modifications of mRNA molecule capture probabilities. Lastly, All Steps of experiments involve manipulating both the probability of capturing a molecule and the read-to-UMI ratio. To explore the potential of these experimental protocols, I employed a simple heuristic for calculating transcript weights, as previously described in the methods section of this chapter. The weights are derived based on pre-set target frequencies, which I set to $\frac{1}{nGenes}$ as simple default heuristic to use for calculating new gene-specific weights.

Weighted Transcriptomes reduces gene sparsity in single-cell datasets

In the first study, I compared the impact of continuously weighted experimental protocols on the distribution of gene sparsity observed in simulated counts in comparison to Non-Weighted single-cell experiments. Notably, all experimental protocols that continuously manipulated sequencing probabilities exhibited significant reductions in gene sparsity compared to the standard Non-Weighted protocols (refer to Figure 5.8). Among these protocols, Weighted Chemistry and Both Steps Weighted simulated counts displayed the most substantial decreases in observed sparsity, in contrast to Weighted Library counts. This finding aligns with

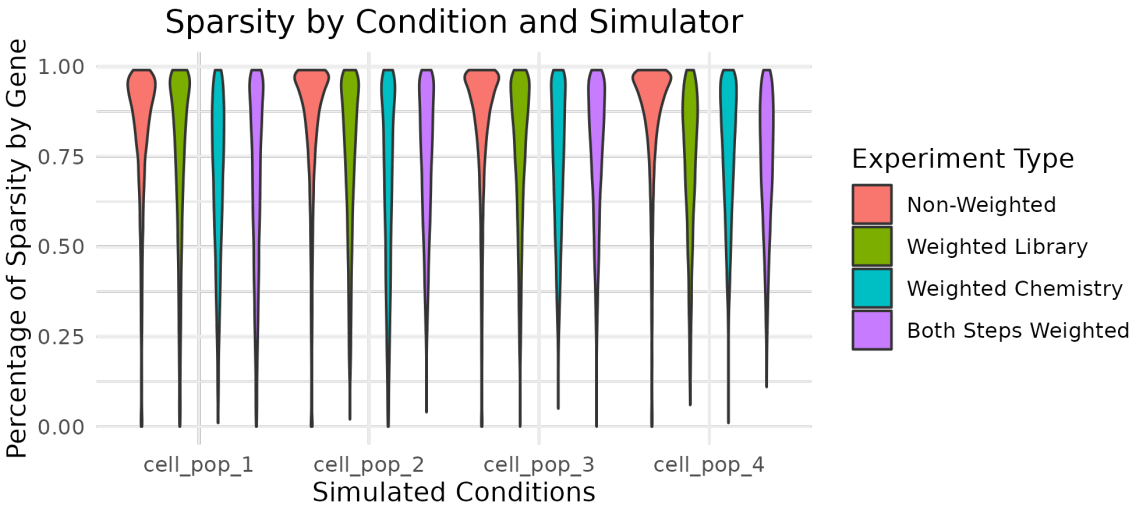


Fig. 5.8 Exploring how different Experimental Protocols alter the sparsity in Single Cell Data across multiple cell populations. The current state of scRNA-Seq experimental protocols involves the Non-Weighted Experiment Type. The other experimental protocols are proposed as theoretically possible experimental protocols, wherein gene transcripts are weighed to capture chemistry, cDNA amplification, or both.

the fact that these methods augment the number of gene transcripts subjected to RT-PCR. It is important to note that RT-PCR is recognized as the most inefficient aspect of a Single Cell Experiment. As such, any increase in the number of transcripts per gene enhances the probability of being sequenced. On average, RT-PCR exhibits on average only 50% efficiency, thereby limiting the capacity to bias cDNA amplification steps during or after PCR to minimize sparsity relative to the number of transcripts that passed through RT-PCR [124]. Consequently, it is evident that the most effective approach to achieve the greatest reduction in gene sparsity is to focus primarily on enhancing or manipulating the Capture Chemistry and RT-PCR steps within single-cell experiments.

Weighted Transcriptomes preserve gene dispersion in Single Cell Datasets

In a second study, I explored how the weighted experimental protocols may warp the gene dispersion compared to current Non-Weighted protocols (see Figure 5.9). Surprisingly, there appears to be little to no substantial general warping of the gene dispersion in continuously manipulated protocols. This is most likely due to the increased representation of low and medium expressed genes and the nature of sampling without a replacement process. However, it is still possible that highly expressed genes may still suffer from an artificially reduced gene dispersion. Such potential warping of gene dispersion will also vary based on when the weighting is applied in the experiment. Despite these concerns, the results are extremely

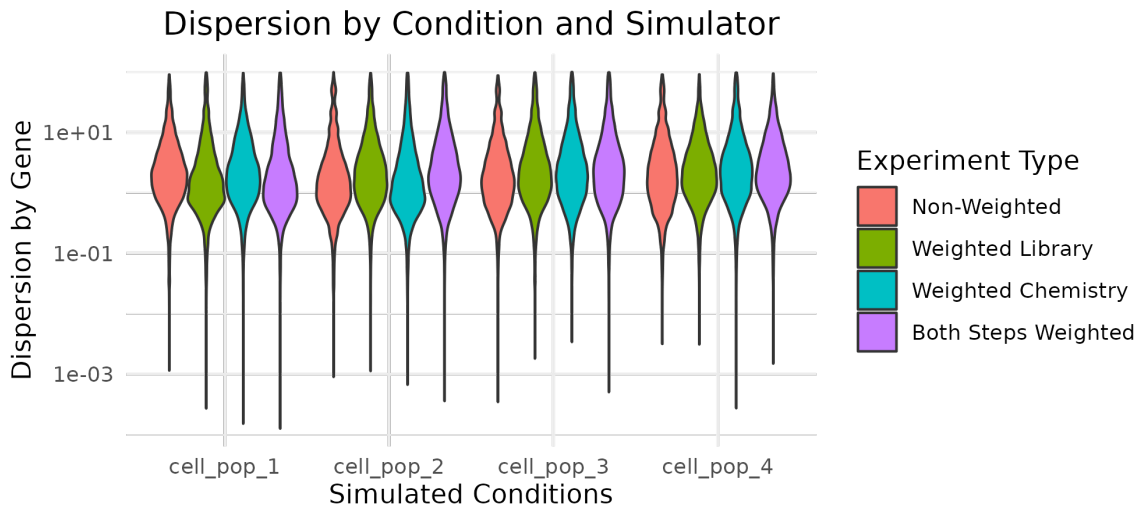


Fig. 5.9 Investigating the impact of various experimental protocols on Gene Dispersion Estimates in Single Cell Data across multiple cell populations. This figure demonstrates that the gene dispersion distributions remain consistent across the different experimental protocols, suggesting that the application of a weighing protocol in the experiment does not significantly affect the overall gene dispersion distribution across multiple simulated cell types.

promising and suggest that weighted methodologies can enrich low and medium expressed genes while preserving their dispersion.

Weighted Transcriptomes reduces sparsity and increases gene mean expression in Single Cell Datasets

For the third comparison, I conducted an investigation into how continuous manipulations of sequencing probability impact the relationship between gene sparsity and mean expression. Notably, all weighted methods exhibited simultaneous increases in mean expression and reductions in sparsity (refer to Figure 5.10). Particularly in the lower range of mean expression for all weighted protocols, there was a noticeable shift towards the left and downward direction. However, the degree of improvement varied among the weighting protocols. While the Weighted Library protocol did elevate the mean expression of genes and decrease sparsity, the overall pattern of the mean sparsity distribution observed in the Non-Weighted Experimental protocol remained unchanged, especially in terms of the distribution tail of highly expressed genes (see Figure 5.10).

The Weighted Chemistry protocols exhibited more significant changes compared to the Non-Weighted and Weighted Library distribution. The Weighted Chemistry protocols demonstrated substantial reductions in sparsity and a narrower distribution of mean expression,

effectively removing the tail of highly expressed genes observed in Weighted Library and Non-Weighted data. This result can be attributed to the weighting applied to the Capture Chemistry stage (see Figure 5.10). Importantly, despite the sparsity reduction, gene dispersion was preserved. These findings suggest that the Biasing of Capture Chemistry step in single-cell experimental protocols can enrich low-expressed genes while maintaining gene dispersion. However, it is important to note that these protocols inherently impose limitations and degradation on the detectable mean expression of highly expressed genes, which in turn may influence gene dispersion as well.

After manipulating the weights for both Capture Chemistry and cDNA amplification, which I refer to as "Both Steps Weighted," I observed a significant increase in the mean counts of transcripts and a substantial reduction in gene sparsity to a greater degree than Weighted Chemistry protocols. This is clearly illustrated by the decrease/elimination of the lower observed mean counts tail in Fig 5.10. However, this enhancement comes at the highest cost compared to other weighted protocols, significantly constraining highly expressed genes, which enables a more pronounced reduction in observed mean expression in the most highly expressed genes compared to other protocols due to the applied weighting across multiple steps. Consequently, Both Steps Weighted Single Cell Experiments effectively minimize gene sparsity but may restrict the observed mean counts of highly expressed genes.

Weighted Transcriptomes increase the overall observable number of genes in Single Cell Datasets

A significant limitation of current single-cell experiments is that the number of observable genes is primarily limited to genes with a high mean expression. This functional exclusion hampers proper characterization of gene distributions of medium to low expressed genes due to the inherent "winner takes all" effects of sampling without replacement. My hypothesis was that continuous manipulation of the probability of sequencing a transcript should increase the number of observed genes compared to Non-Weighted experiments. To test this, I examined how the number of genes with a non-zero count increased as a function of experimental sequencing saturation. I found that, across all weighted protocols, there was a general augmentation in the number of genes with non-zero counts, regardless of sequencing saturation, compared to Non-Weighted experiments (refer to Figure 5.11). Biasing cDNA Amplification demonstrated an initial significant increase in the number of observable genes, particularly in lower sequencing saturations. However, as sequencing saturation increased, the observed count of Weighted libraries eventually plateaued and approached that of the Non-Weighted experiment. This suggests that there is an effective limit to the ability of cDNA

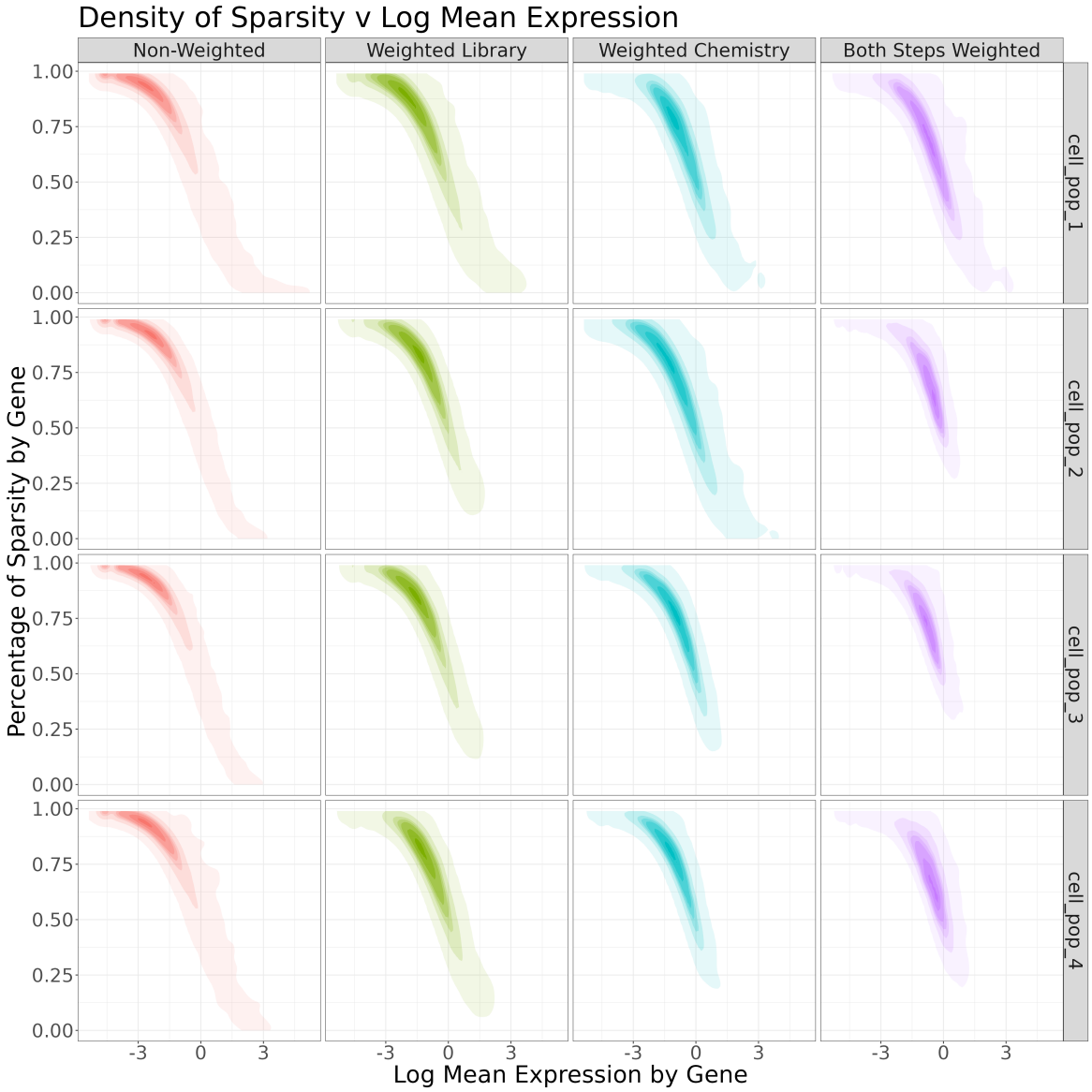


Fig. 5.10 Exploring how different experimental protocols alter the sparsity and the log-transformed mean gene expression in Single Cell Data across multiple cell populations. The figure demonstrates that each of the weighted experimental protocols leads to a decrease in gene sparsity and an increase in mean expression. However, there are distinctions in terms of the overall distribution location among the weighted transcriptomes. Specifically, the weighted library protocol appears to have the least impact compared to both the Weighted Chemistry and Both Steps Weighted protocols.

Amplification Biasing protocols to enhance the observability of low and medium-expressed genes effectively (see Figure 5.11).

While Weighted Chemistry and Both Steps Weighted protocols exhibited a substantial increase in the number of detectable genes compared to both Weighted Library and Non-

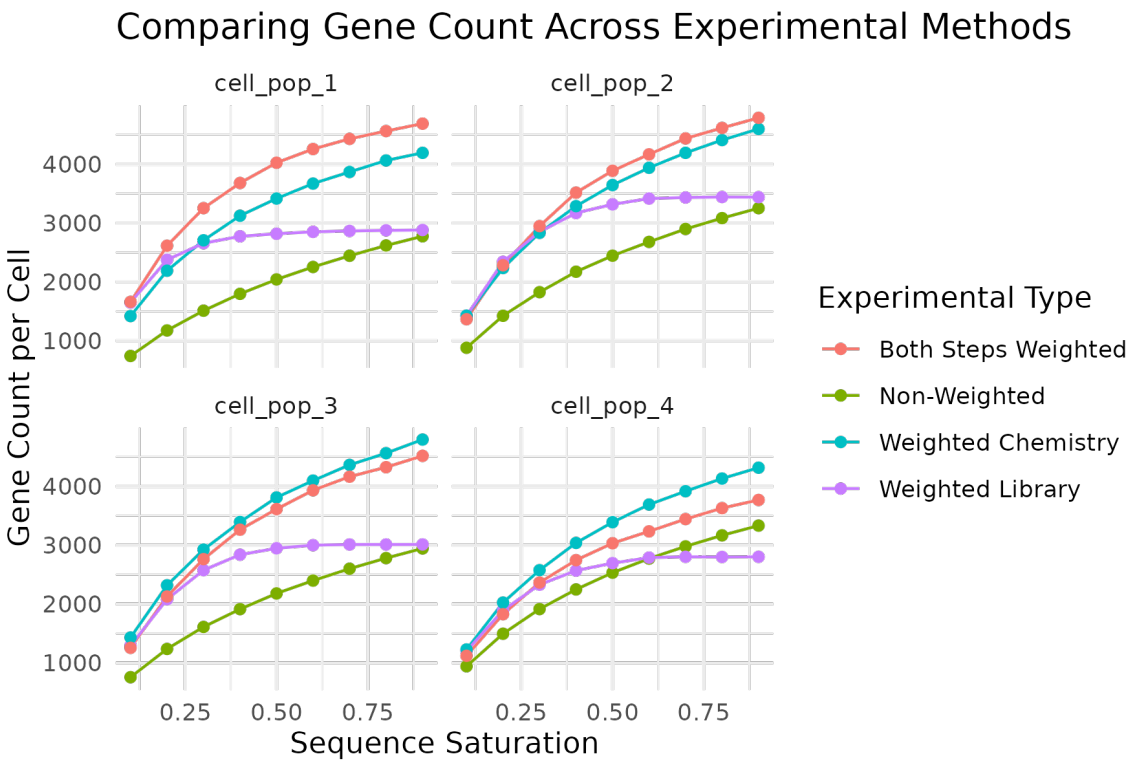


Fig. 5.11 Investigating the impact of different experimental protocols on the number of observable genes in a dataset across multiple cell populations as a function of sequencing saturation. This figure illustrates that using weighted experimental protocols enhances the number of observable genes, particularly as the sequencing saturation of the experiment increases. Significant differences are observed in the number of genes observed when employing various weighted protocols. The Weighted Library protocol exhibits the least increase, eventually reaching a plateau. Comparatively, the Both Steps Weighted protocol performs better than the Weighted Library but is not as effective as the Weighted Chemistry protocol. There could be several reasons for these discrepancies, including the possibility of utilizing an inappropriate method for calculating transcript weights.

Weighted protocols from the outset, this increase persisted as sequencing saturation increased, eventually plateauing closer to a sequencing saturation of 0.9 and 1 (see Figure 5.11). There is a significant difference that can be observed between Both Steps Weighted protocols and Weighted Chemistry protocols in terms of the number of detectable genes as a function of sequencing saturation, particularly in the first two cell populations. However, in cell populations three and four, Weighted Chemistry appears to have outperformed Both Steps Weighted in terms of the number of observed non-zero genes; it remains unclear why this is the case and whether the issues are related to inherited differences in the cell populations or if there may be an issue with the heuristic used for calculating the weights of the protocols

(see Figure 5.11). Despite this, the overall results indicate that single-cell experiments employing Both Steps Weighted protocols can achieve the most significant reduction in sparsity, increased observed mean expression, and the largest increase in the number of non-zero expressing genes. However, it's worth noting that fine-tuning the optimal weights may be necessary to harness these benefits fully.

5.4 Discussion

Single Cell Biology has a variety of simulators available that aim to simulate different statistical and biological phenomena, creating a rich ecosystem for the field. Although they share the purpose of benchmarking, these simulators differ in their focus. For instance, Splatter and SPARsim are designed to benchmark normalization and clustering methods. Other simulators address specific problems, such as benchmarking cellular trajectories using Dyngen or incorporating a range of data modalities (scRNA-Seq, scATAC-Seq, spatial transcriptomics) and biological phenomena (gene regulatory networks, cellular trajectories) using scMultiSim [139, 14, 21, 83].

Minerva was specifically developed with a distinct purpose compared to other simulators. Its primary objective is to enable the simulation of both binary and continuous experimental manipulations of sequencing probability. It achieves this by utilizing a wet lab-aware hierarchical model capable of accurately capturing the statistical characteristics of single-cell data while taking into account potential experimental manipulations. This unique feature empowers users to explore the theoretical landscape of single-cell experiments and swiftly identify experimental protocols that generate the statistical properties they desire. Minerva's wet lab-aware hierarchical model allows for the simulation of various types of single-cell experiments, including targeted or weighted transcriptomes, using normal scRNA-Seq as input. This capability enables efficient exploration and search for experimental protocols that optimise single-cell data by reducing sparsity and enhancing the observed mean count of genes.

To validate Minerva's performance, I conducted a series of computational experiments, assessing its ability to simulate single-cell data across different biological contexts. First, I compared its simulation of normal whole transcriptome scRNA-Seq data against Splatter and SPARsim simulators and real scRNA-Seq data. Where I observed that Minerva was able to faithfully recapitulate the statistical characteristics of normal scRNA-Seq data. However, it is important to note that the quality of the simulated datasets can vary depending on the quality of the real data provided as input. Next, I evaluated Minerva's capability to simulate a novel single-cell experimental method of targeted transcriptomes. I found that it could do

so without any priors beyond the target gene list. Once validated, I further explored the space of theoretically possible single-cell experiments to search for experimental protocols that could reduce sparsity and increase detectable mean expression in genes with low expression.

Using Minerva, I conducted an investigation into three weighted experimental protocols: Weighted Library, Weighted Chemistry, and Both Steps Weighted. These protocols all involve some form of enrichment during the experiment, but the timing and impact of the enrichment on the data can vary significantly. Across all three methods, sparsity was effectively reduced, and detectable mean expression was increased. In order of performance from least to the greatest increase, in reduction of sparsity and observed mean count Weighted Library, Weighted Chemistry, and Both Steps Weighted with each of these protocols outperforming the previous one significantly. This discrepancy can be attributed to the timing of the enrichment during the experiment, specifically before or after the RT-PCR stage. Depending on when the enrichment is applied, there are dramatic differences in the performance of these protocols in terms of reducing sparsity and increasing detectable mean expression of low-expressed genes.

RT-PCR is only about 50% efficient in converting mRNA transcripts into cDNA, making it functionally similar to a filter or shedder, which removes nearly half of the captured transcripts from being sequenced in a single-cell experiment. This has a greater impact on the lowest expressed genes. In addition, sampling without the replacement process that occurs in the Capture Chemistry step is the greatest challenge facing single-cell experiments and is largely responsible for the winner takes all effects observed. This step follows a Hypergeometric Distribution and is disadvantageous for low-expressed genes as their transcripts are less likely to be sampled again. However, by manipulating the probability of transcripts being sampled again via experimental protocols, Capture Chemistry follows an MFNH distribution. The Weighted Chemistry and Both Steps Weighted experimental protocols can ensure and increase the number of transcripts for a given gene that is captured, increasing the number of transcripts that can be sequenced. This is in contrast to the Weighted Library protocol, which can only work with transcripts that have passed the RT-PCR step.

The Weighted Library Experimental protocol is fundamentally at a disadvantage compared to Weighted Chemistry, and Both Steps Weighted because its experimental manipulation occurs after RT-PCR. Instead of enriching mRNA transcripts, the Weighted Library step alters the ratio of reads to UMI during or after PCR, shifting the nature of the enrichment from sampling without replacement to sampling with addition, also known as the Polya urn problem. This subtle shift is because I am biasing a PCR reaction with uniquely labelled transcripts, which means I am altering the sequencing library's composition. This increases the probability of observing low-expressed gene transcripts that have made it through RT-

PCR but does not enrich the transcripts that were not captured. Despite these limitations, Weighted Library protocols have their uses as they are easier to implement in a Single Cell Experiment than weighting Capture Chemistry.

The field of Single Cell Biology has a rich ecosystem of simulators that simulate various statistical and biological phenomena. Minerva stands out as a wet-lab-aware simulator, taking into account experimental manipulations and how they would alter the statistical characteristics of single-cell data. Computational experiments validated Minerva's ability to faithfully recapitulate the statistical characteristics of normal and targeted transcriptome single-cell experiments. I then explored the theoretical space of single-cell experiments to search for protocols that could reduce sparsity and increase detectable mean expression in low-expression genes. This revealed that biasing the Capture Chemistry step was the greatest factor in improving experimental performance. While biasing the library preparation step also improved experimental performance, it was to a lesser degree. Regardless, biasing the library preparation is simpler than biasing the Capture Chemistry.

Minerva proves to be a powerful tool for efficient exploration of the theoretical Single Cell Experiment space, enabling the evaluation of the resulting data's statistical characteristics and facilitating the identification of experimental protocols with the potential to enhance overall performance. However, there remain numerous questions surrounding the impact of weighted transcriptomes on the statistical properties of gene expression distributions. A significant concern is their potential to alter the observed gene dispersion of genes with high mean expression, potentially leading to the loss of vital information. Furthermore, there may exist various less obvious and subtle manipulations and alterations that could occur in low to medium-expressed genes within weighted transcripts. To mitigate these potential issues, a thorough Exploratory Data Analysis (EDA) of weighted transcriptomes becomes essential to ensure the integrity of the data.

Chapter 6

Examining the Statistical Properties of Weighted Single Cell Data

Single-cell experiments provide a valuable means to examine biological systems at a high resolution, enabling the identification and study of rare cell populations and their responses to perturbations. However, the characterization of individual cell transcriptomes is limited by inefficiencies in experimental protocols and shallow sequencing. To address this, I developed Minerva, a wet-lab-aware single-cell simulator that facilitates the exploration of continuous manipulation of sequencing probabilities. In the previous chapter, I introduced a new class of weighted single-cell experiments specifically designed to enrich low and medium expressed genes. Now I will explore how these experimental manipulations impact the resulting datasets' statistical properties and what biases they might introduce. To address this question, I used Minerva to simulate weighted transcriptomes and investigated the statistical properties of three different protocols. The results showed no evidence of significant alterations in the statistical characteristics of observed counts nor a degradation in the ability of regular single-cell bioinformatic methods to identify cell populations. In addition, I observed that weighted datasets contained more information compared to Non-Weighted single-cell experiments, enabling a more comprehensive characterization of low and medium expressed genes. This demonstrates the potential of weighted transcriptomes to enhance our understanding of cellular heterogeneity and gene expression dynamics at the single-cell level.

6.1 Introduction

Weighted single-cell experiments represent a theoretical, experimental protocol that involves continuous manipulation of sequencing probability. These experiments are specifically

designed to enrich low and medium expressed genes, thereby enhancing our ability to characterize and observe these genes more effectively. In Chapter 5, I have shown that these types of experimental protocols effectively remove sparsity from single-cell datasets and increase the observed mean expression. However, a more comprehensive examination of the statistical properties of the weighted single-cell datasets is necessary. Specifically, I need to investigate whether the heteroscedasticity of the mean-variance relationship of gene expression remains preserved. This is crucial because enrichment protocols have the potential to alter the statistical properties and information content of these datasets. Confirming or disproving that weighted single-cell experiments' statistical characteristics are like normal single-cell datasets will determine whether novel normalization methods are needed and whether the statistical assumptions of current single-cell bioinformatics methods are being violated.

Here I will show that in weighted single-cell experiments, most genes across the expression distribution exhibit no alteration or effects of manipulating the mean-variance relation. In addition, weighted datasets contain more information than non-weighted state-of-the-art single-cell experiments, allowing us to better characterize the true expression distribution from the observed counts. Finally, I validated that current single-cell normalization methods' can remove technical variance and identify cell populations in weighted datasets. No degradation in performance was observed. These results suggest that weighted single-cell experiments do not alter the statistical characteristics of genes observed counts and provide more information on the transcriptome than the current state-of-the-art Non-Weighted single-cell experimental protocols.

6.2 Methods

6.2.1 Simulating Data

I utilized Minerva, a Single Cell Simulator incorporating wet-lab parameters to simulate the statistical characteristics of single-cell data. Through this simulation, I examined how different experimental protocols, including targeted or weighted transcriptome approaches, can influence the resulting data. For the specific datasets I used, please refer to Table 5.1 in Chapter 5.

6.2.2 Describing the Observed Count Distributions

Moment-Based Statistics

To assess the effects of weighted transcripts and identify any potential distortions they may apply to observed count distributions, I calculated the first (mean) and second (variance) moments of each gene's observed count distribution for a given cell population within a dataset. Single-cell count matrices follow a gene-by-cell format, where the counts of a given row represents the expression count of a particular gene in all cells, and the count in a given column represents the expression count of all genes in a particular cell. To calculate the mean and variance, I first subset the count matrix for a given cell population, creating a cell population count matrix. A cell population is a set of J cells that belong to a given population. Using this cell population count matrix, I estimated the empirical mean μ and variance σ^2 for a given gene i by iterating through the rows of the expression matrix X, where $X_{i,k}$ is the expression count of gene i in cell k.

$$\mu_i = \frac{1}{J}\sum_{j=1}^{J} X_{i,j} \tag{6.1}$$

$$\sigma_i^2 = \frac{1}{J-1}\sum_{j=1}^{J} (X_{i,j} - \mu_i)^2 \tag{6.2}$$

Once all the genes' mean expressions were calculated, I determined which quantile a given gene belonged to. To do this is first removed all genes with a mean expression less than 0.001. A small subset of extreme outliers was on the low end of the estimated mean expression. Preventing them from skewing the quantiles to be lower than they were. Next, I ranked the order to mean expression from lowest to highest and calculated the empirical Cumulative Frequency Distribution (eCFD) [33]. I determined with 25th, 50th, and 75th percentiles where and assigned genes to a given quantile. Determining the quantile a gene belongs to allows me to explore the effects weighted transcriptomes have upon the gene mean and variance as a greater level of resolution, as their effects may not be the same across all of the quantiles of the mean expression distribution.

Calculating the Log Fold Change in Weighted Transcriptomes

To compare how the different weighting protocols increased and decreased mean gene expression and variance, I calculated the log-fold change of both the mean and variance between a given weighted transcriptome and the Non-Weighted transcriptome using the following formula:

$$\text{log fold change of stat} = \log(\text{weighted stat}) - \log(\text{non-weighted stat}) \tag{6.3}$$

This log-fold change was calculated on a per-gene, cell population, and dataset basis. Once calculated, I then fitted a Kernel Density Estimate (KDE) to visualize the changes that occurred and determine the overall directionality of weighted transcriptomes' effects on gene expression and variance.

Measuring the Strength of the Relationship Between Observed and True Count

While comparing moments can be informative and provide observational evidence of how the observed count distribution has changed, it does not indicate the amount of information the observed count distribution contains about the true count distribution between the observed and true data. Therefore, I also calculated the mutual information between the observed counts of each experimental method and the true gene expression distribution per cell populating in a given dataset. The formula for calculating mutual information between two random variables A and B is:

$$I(A,B) = \sum_{a \in A} \sum_{b \in B} p(a,b) \log_2 \frac{p(a,b)}{p(a)p(b)} \tag{6.4}$$

Where A is the observed counts of each experimental method, and B is the true gene expression for a given cell population in a given dataset. $p(a,b)$ is the joint probability distribution of A and B, and $p(a)$ and $p(b)$ are the marginal probability distributions of A and B.

Calculating mutual information allowed me to explore the mutual dependency between the simulated observed and biological variance counts and better understand how different experimental protocols influence the observed counts of single-cell experiments [28]. Mutual Information is a non-linear measurement of the relationship between the observed and biological variance count distributions; this allows me to capture a wider range of the relations between the two [28]. Although Mutual Information cannot indicate whether the relationship is positive or negative, this is of little consequence for this analysis. I am most interested in whether weighted transcriptomes have more information about the biological variance of a gene's expression distribution than non-weighted transcriptomes from current single-cell protocols.

6.2.3 Assessing Performance of Normalization Methods

Normalization Methods

Normalization is a critical step in the downstream analysis of Single Cell Data Analysis. There are various approaches for normalizing single-cell data, including delta, residual, and count-based normalization [17, 4]. The delta approach is a log-transformation of count data that addresses the issue of zero counts by adding a small pseudo count (e.g., 0.001) to the count matrix, allowing for the application of the log transformation [17, 4]. The optimal use case for applying the delta transformation is when the variance primarily depends on the mean. Within a biological context, it provides a quick and easy way of removing heteroskedasticity. Mathematically, the delta transformation can be written as:

$$x'_{i,j} = log_2(x_{i,j} * s_j + \alpha) \tag{6.5}$$

Where x_{ij} is the count for gene i in cell j, s_j is the cell specific scaling factor to adjust x_{ij} by, α is the pseudo count, and x'_{ij} is the transformed value used for downstream analysis. Here, we tested variations of the delta approach, including the Sum Pooled Factor and 10k scaling log transformation.

Residual-based normalization methods stabilize variance by fitting a null model on a gene-specific level that was originally proposed by Hafemeister and Satija. The motivation for developing these methods comes from the inability of the delta approach to deal with genes with an extremely low mean expression (typically with a mean less than 0.01) [55, 4]. Once the null model is fit, the observed count data is transformed into a residual, typically a Pearson residual, which is then used for downstream analysis. The Pearson residual is calculated as the difference between the observed count and the expected count based on the null model, normalized by the variance of the expected count where θ is gene dispersion. Mathematically, the Pearson residual can be calculated as follows:

$$r_{i,j} = \frac{x_{ij} - \mu_i}{\sqrt{\mu_i + \theta \cdot \mu_i^2}} \tag{6.6}$$

Where x_{ij} is the observed count for gene i in cell j, $\hat{x}_{ij}$ is the expected count based on the null model, and $Var(\hat{x}_{ij})$ is the variance of the expected count. I utilized the SCTransform package developed by Hafemeister and Satija to conduct residual-based normalization.

Finally, count-based normalization, developed by Townes et al., involves taking the observed counts and directly applying GLM-PCA to uncover latent structures for subsequent downstream analysis. Through a comparison of these orthogonal methods for normalizing weighted single-cell datasets and an assessment of their effectiveness in accurately identi-

fying differentiating features between cell populations, I aim to determine the capability of current single-cell bioinformatic methods in removing technical variance from weighted transcriptomes.

Adjusted Rand Index

scRNA-Seq is primarily utilized to discover novel cell populations and subpopulations, heavily reliant on the ability of current single-cell bioinformatics methods to identify genes with the highest variance. These high-variance genes are then used as features for dimensionality reduction and population visualization. However, the introduction of weighted transcriptomes has the potential to modify the statistical properties, possibly leading to a decrease in the performance of existing single-cell methods. To investigate whether weighted transcriptomes alter or diminish the performance of these methods, I employed the Adjusted Rand Index (ARI). The ARI is a statistical measure that quantifies the similarity between two clustering results. It takes into account all pairs of data points and calculates the proportion of pairs assigned to the same cluster in both results, normalized by the maximum possible agreement based on chance [112]. In this case, the two clusters being compared are the clusters identified by a normalization method and the ground truth cell population simulated by Minerva.

The ARI can be calculated using the contingency table, where n_{ij} is the number of data points simultaneously assigned to cluster C_i in the first clustering result and to cluster C_j in the second clustering result. Let a_i be the total number of data points assigned to cluster C_i in the first clustering result, and b_j be the total number of data points assigned to cluster C_j in the second clustering result. Then, the ARI is given by the following formula:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{6.7}$$

Where n is the total number of data points, the ARI ranges between -1 and 1, where a value of 1 indicates perfect agreement between the two clustering results, 0 indicates agreement no better than chance, and a negative value indicates disagreement worse than chance [112].

ARI is an established method for evaluating the performance of clustering algorithms and their results. In previous studies, ARI has been used to assess the performance of single-cell normalization methods in unweighted single-cell experiments [130]. ARI has been used previously to evaluate the performance of normalization methods because single-cell normalisation aims to remove the technical variance introduced during the experiment, enabling users to identify clusters driven by biological variance. Here I am using ARI to

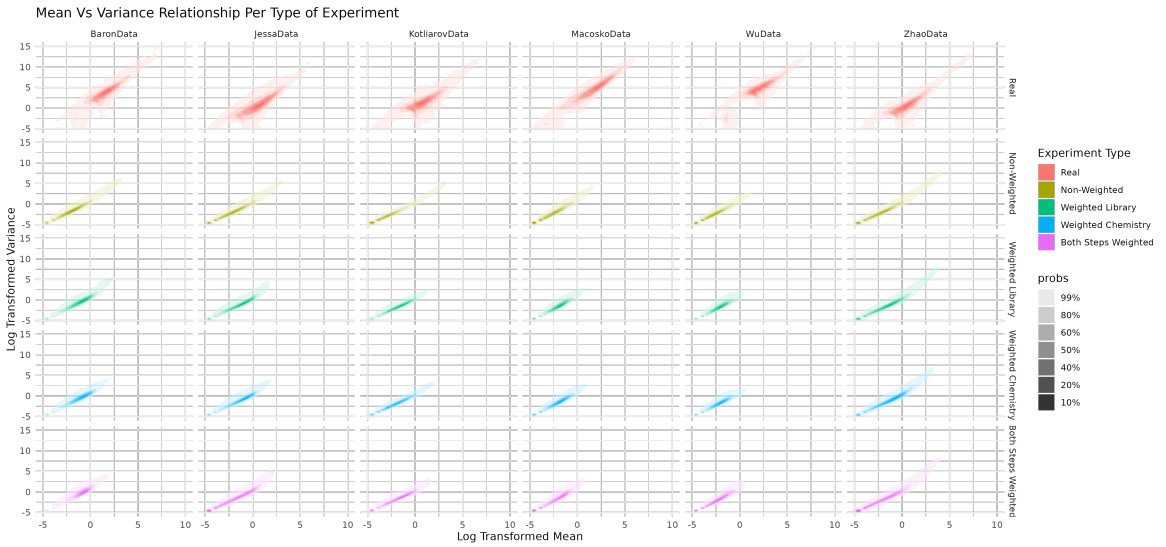


Fig. 6.1 Two-Dimensional KDE of Mean-Variance by Dataset and Type of Transcriptome. The Real transcriptome is the mean and variance of the gamma distribution used by Minerva to simulate biological variance. While all of the others are different types of single-cell experiments. Non-Weighted is a normal single-cell experiment and the rest apply a different weighting protocol to enrich for low expressed genes.

assess the effect a weighted transcriptome has on the ability of current standard Single Cell Normalization methods to remove the technical variance and identify biological clusters.

6.3 Results

6.3.1 Assessing the Effects of Weighted Transcriptomes on Gene Expression Mean and Variance

Weighted Single Cell Experiments is a theoretically plausible experimental protocol I previously developed where the probability of sequencing a gene transcript is manipulated depending on its relative frequency in the transcriptome. This results in highly expressed genes being unenriched while low and medium expressed genes are enriched, which improves the overall visibility of the transcriptome in Single Cell Experiments. Moreover, this approach increases the number of observable genes and the observed mean expression of the enriched low to medium expressed genes while potentially constraining the mean expression and variance of highly-expressed genes. However, the extent of these effects is unclear. Do they affect only a few or a lot of genes? Are there varying degrees to which genes are affected?

To address these questions, I used Minerva, a wet lab-aware simulator that allowed me to simulate normal and weighted single-cell experiments across eight publicly available datasets. These datasets spanned various biological contexts, from pediatric brain cancer to the pancreas, and utilized different droplet-based single-cell isolation techniques (i.e., Drop-seq, 10x, and inDrop). Simulating across various biological and technical contexts allowed me to determine the true effects of weighted transcriptomes on observed gene expression by searching for general changes across these settings. The mean-variance relationship is of utmost importance and requires thorough examination, as it characterizes the gene expression distribution key aspect for downstream bioinformatic methods in both single-cell and bulk RNA-Seq data analysis. These methods heavily rely on assumptions related to this relationship. Notably, it is assumed that as the mean expression increases, so does the variance, with the variance often surpassing the mean (known as heteroskedasticity). Additionally, for identifying cell populations, genes with high variance are believed to contain multiple unique distributions, each associated with specific cell populations. In biological settings, it is widely acknowledged that variance is frequently larger than the mean and tends to increase as the mean expression rises. If the observed count statistics in weighted transcriptomes deviate from this relationship, it would undermine the validity of most statistical assumptions used in current techniques for analyzing single-cell data [8].

To explore the effects of weighting a transcriptome, I first estimated the mean and variance of the simulated, Non-Weighted, and weighted transcriptomes on a per cell population and dataset basis. I then visualized these relationships using a two-dimensional KDE, creating a density estimation of the joint distribution of gene expression's mean and variance. The first observation is that the Non-Weighted single-cell data does not follow the mean-variance relationship observed in the simulated biological variance mean-variance (see Figure 6.1), suggesting that Non-Weighted single-cell data does not fully reflect the true statistical behaviours of genes' biological variance but rather follows a more constrained shadow of the true transcriptome. This is most likely attributed to the technical noise introduced during single-cell experiments, where transcripts undergo multiple rounds of downsampling. As a result, the statistical properties of low to medium-expressed genes are altered, leading to reductions in both the mean and variance of gene expression. Despite this, the overall heteroskedasticity of the mean-variance relation observed in the simulated biological variance was observed in Non-Weighted transcriptomes.

When comparing the various weighted transcriptomes (Weighted Library, Weighted Chemistry, and Both Steps Weighted), it was observable that their mean-variance relationship all followed the same general trend as the Non-Weighted transcriptome. However, in this relationship, there appeared to be certain constraints, with the upper right outliers demonstrating

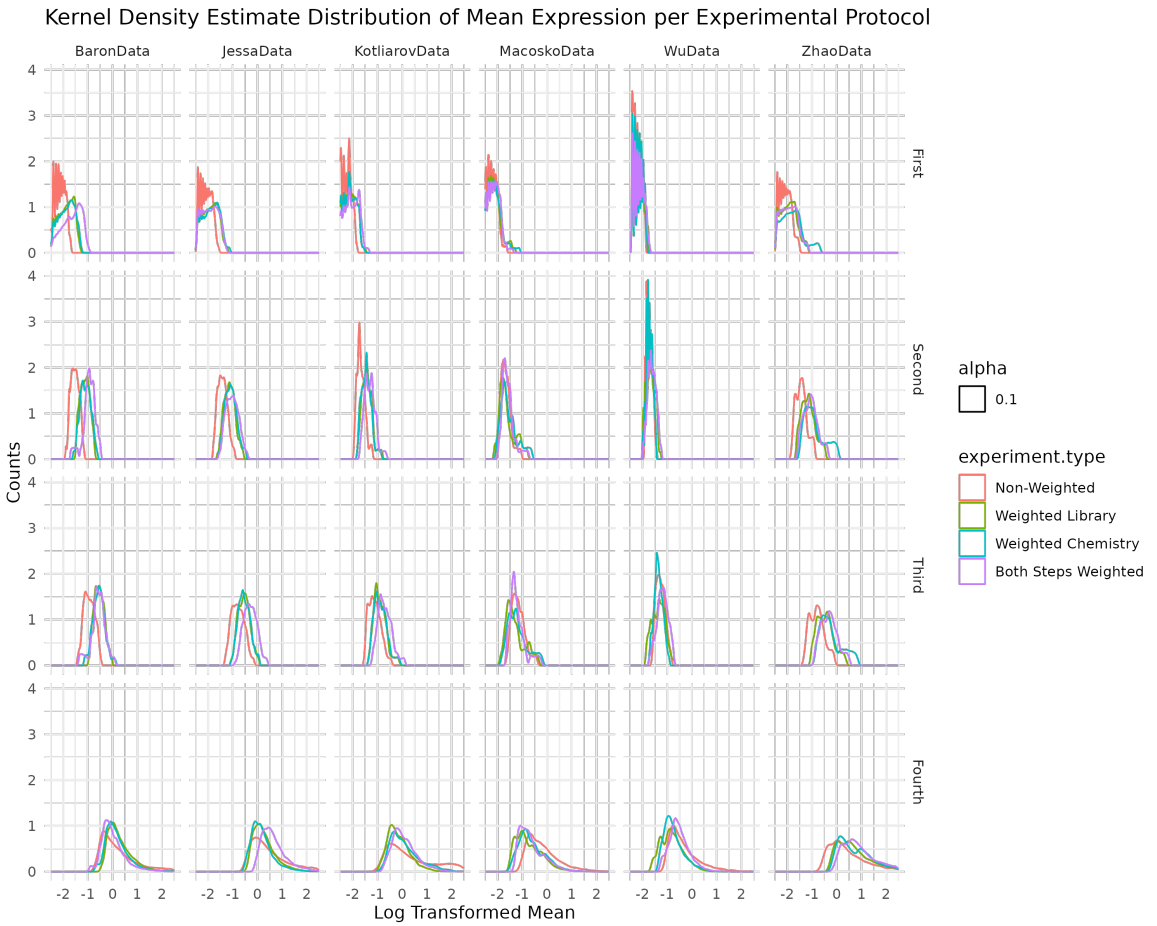


Fig. 6.2 One-Dimensional KDE of Mean Expression by Dataset and Quantile. Only the Observed Mean expression distribution was plotted to compare weighted transcriptomes to the Non-Weighted transcriptome.

a regression to the mean effect. This observation aligns with the expectation that weighted transcriptomes would limit the count variance of genes with high mean expression (see figure 6.1). The impact of different experimental protocols for weighting the transcriptome on the count variance of highly expressed genes varied, primarily depending on whether the weighting step occurred before or after the RT-PCR stage in the single-cell experiment. Weighted Library transcriptomes had the largest outliers among the weighted transcriptomes and were the closest to the Non-Weighted transcriptomes compared to Weighted Chemistry and Both Steps Weighted. Whereas the Weighted Chemistry and Both Steps Weighted outliers were far more constrained. Despite this, the area with the greatest density appears to follow the same trend as the Non-Weighted transcriptome. This suggests that genes in the upper percentile are constrained, but the bulk of genes with a high mean expression appear to be unaffected.

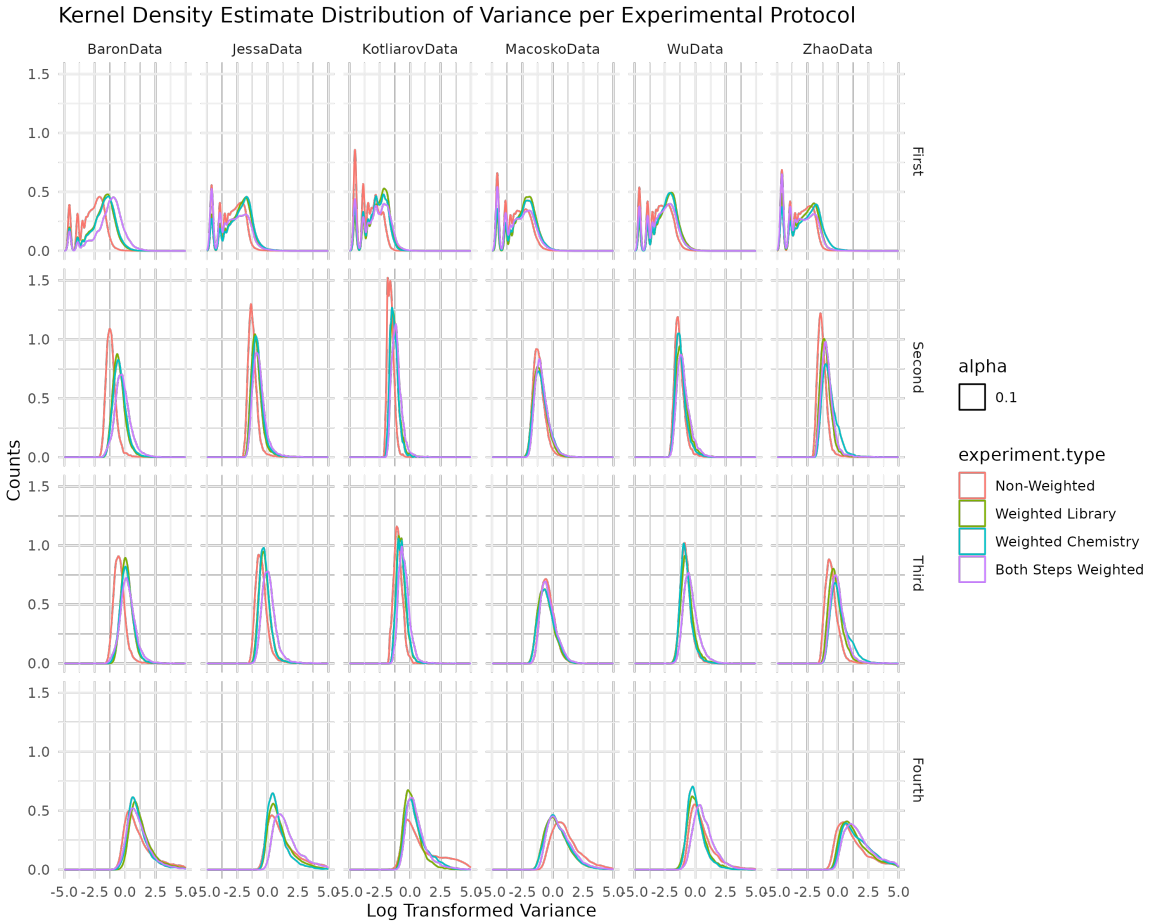


Fig. 6.3 One-Dimensional KDE of Variance by Dataset and Quantile. Only the Observed Mean expression distribution was plotted to compare weighted transcriptomes to the Non-Weighted transcriptome.

To better understand how the effects of weighted transcriptomes may constrain a gene's mean and variance, I explored the distribution at a higher resolution. To do this, I estimated the eCDF of the gene expression distribution of a given cell population in a dataset for an experimental protocol. Using the eCDF I grouped genes into quantiles by converting their mean expression distribution into an empirical percentile. Next, I estimated one-dimensional KDEs for both the mean and variance per quantile and dataset (see figures 6.2 and 6.3). Both the mean and variance of all experimental protocols, including both weighted and Non-Weighted transcriptomes, exhibited a similar pattern. The only notable distinction observed in the weighted transcriptomes was a higher mean and variance in the first and second quantiles compared to Non-Weighted transcriptomes. In the third quantile, the expression distribution of all weighted transcriptomes closely resembled, if not identical to, that of the Non-Weighted transcriptome in terms of both mean and variance. Additionally, in the fourth

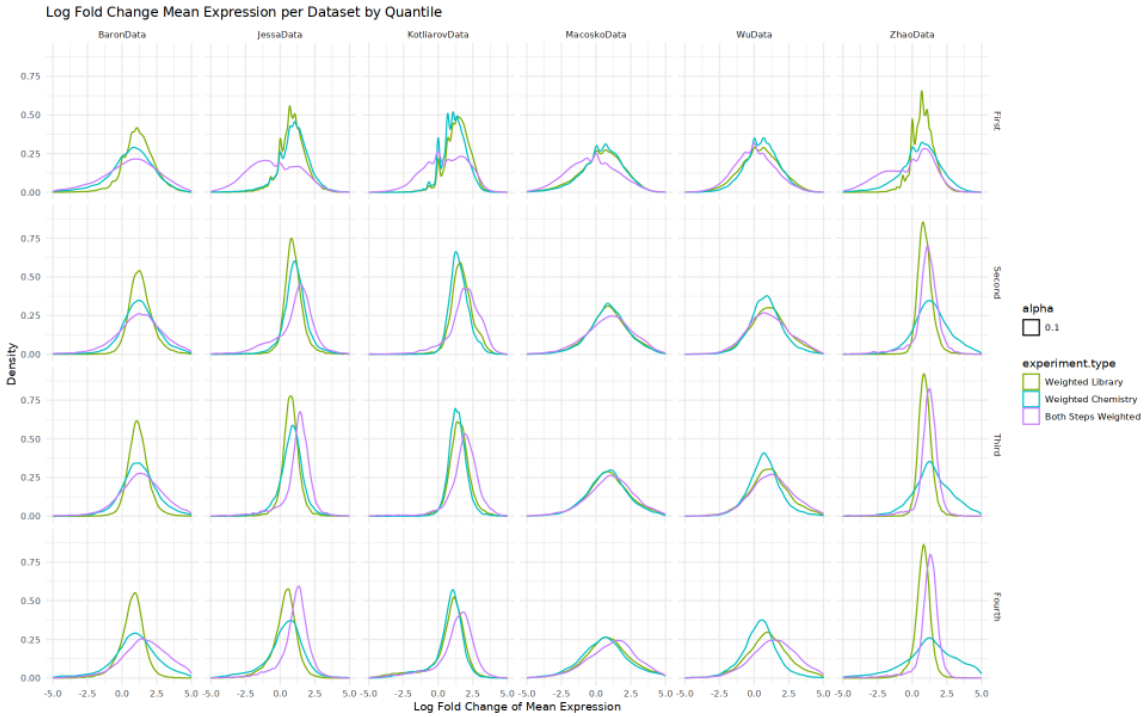


Fig. 6.4 One-Dimensional KDE of the genes' mean expression Log Fold Change by Dataset and Quantile. The Log Fold Change in mean expression was computed using the Non-Weighted gene mean expression as the baseline. The KDE estimates displayed in the figure illustrate the distribution representing the overall alteration in gene mean expression compared to the Non-Weighted Transcriptomes.

quantile, the mean and variance of weighted transcriptomes exhibited greater constraints, resulting in a thinner right tail compared to the Non-Weighted transcriptome. These findings suggest that the mean-variance relationship is largely preserved across most genes in the weighted transcriptomes.

It is important to highlight that the KDE distributions depicted in figures 6.2 and 6.3 indicate that weighting the transcriptome does not significantly alter the moments of the distribution. However, it should be noted that the purpose of weighting methods is to enrich genes that were previously unobserved. In order to confirm that weighted transcriptomes effectively enrich genes, I computed the LFC of both moments in the weighted transcriptomes compared to the Non-Weighted transcriptome. I then visualized the distribution of LFC for both moments by fitting a one-dimensional KDE. An expected outcome of weighting transcriptomes is a notable increase in the mean expression of genes within the 1st and 2nd quantiles. Simultaneously, the variance in gene expression is anticipated to increase as well, owing to the inherent heteroscedasticity property of gene expression distributions. As shown in figures 6.4 and 6.5, the mean and variance are both positively skewed, indicating

that weighted transcriptomes enrich the observed counts of genes. What was unexpected was the observation that the increase in both the mean and variance of gene expression occurred across all quantiles. This finding suggests that cell populations' transcriptomes are predominantly influenced by a small number of highly expressed genes.

In the third and fourth quantiles, the left tail (representing a decrease in the LFC of a gene's mean and/or variance) of LFC KDE is most pronounced, showing that some genes are constrained when weighting is applied. However, these characteristics differ based on the experimental protocol used to weight the transcriptome. Specifically, Weighted Chemistry and, to a lesser degree, the Weighted Library protocols have a thicker left tail in the third and fourth quantiles than Both Steps Weighted. While Both Steps Weighted transcriptomes appear to have a mostly positive LFC in the upper quantiles of the expression distribution (see Figures 6.2 and 6.3). By applying the enrichment protocols during capture chemistry, there is an increase in the number of constrained genes, which is expected as it was previously shown that manipulating the capture chemistry has the greatest effects on single-cell data, as enrichment before RT-PCR. When employing the Weighted Chemistry protocol, it becomes evident that a substantial number of genes undergo excessive compression of their mean and variance. Nevertheless, by adjusting the cDNA amplification process, Both Steps Weighted are able to mitigate the risk of over-correction that might have arisen during the Capture Chemistry step of the protocol.

6.3.2 Determining the Amount of Information in Weighted Transcriptomes

The primary objective of a single-cell experiment is to acquire a comprehensive and efficient understanding of biological systems by obtaining high-resolution and high-throughput data. In such experiments, the observed counts of genes reflect the true expression distribution, thereby offering valuable insights into the biological processes at play. The greater the extent to which the observed counts distribution reveals the true expression distribution of genes, the more valuable the obtained data becomes. Quantifying the amount of information the observed counts contain about the true expression distribution in a biological setting is currently impossible. Nevertheless, assessing the extent to which observed counts reflect biological variance is feasible when conducting simulations, as both the observed and true counts are known. This enables the quantification of the information contained within observed counts concerning biological variance in simulation. To measure this information, a commonly used statistical measure is Mutual Information (MI), which is based on information theory and quantifies the amount of information that variable X provides about variable Y. When variable

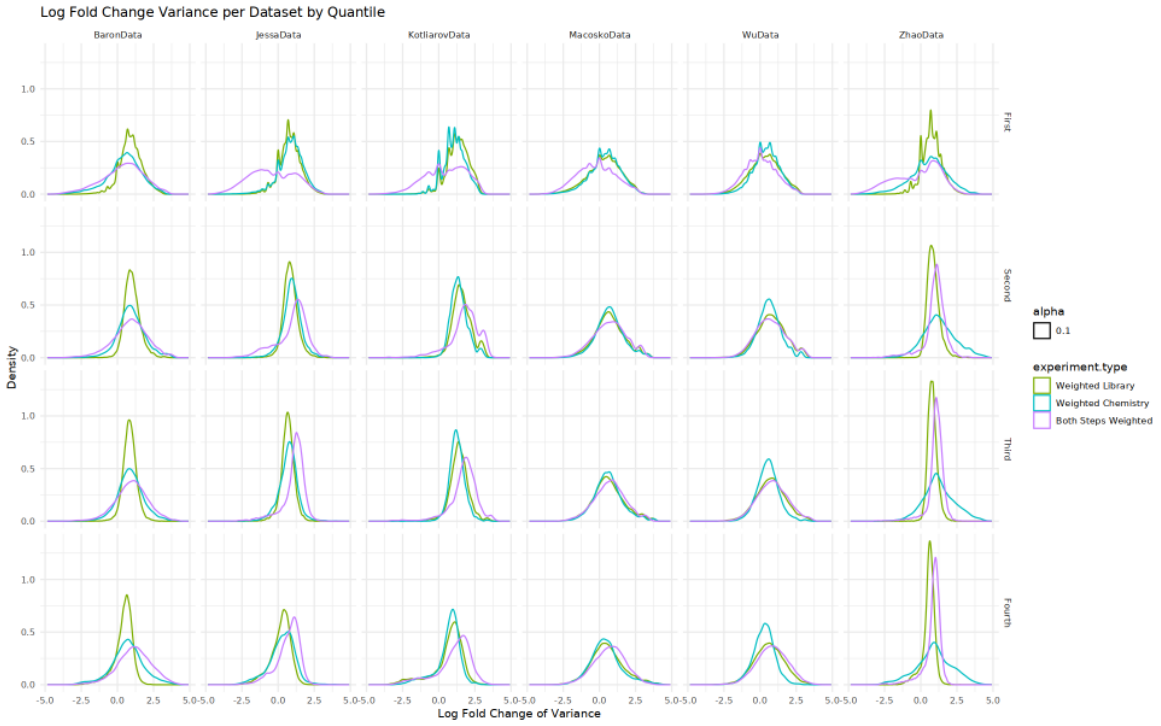


Fig. 6.5 One-Dimensional KDE of the genes' variance Log Fold Change by Dataset and Quantile. The Log Fold Change in variance was computed using the Non-Weighted gene variance as the baseline. The KDE estimates depicted in the figure illustrate the distribution representing the overall change in gene variance compared to the Non-Weighted Transcriptomes.

X provides us with no information about Y MI is 0 if X provides perfect information about Y then MI is 1. The ideal experiment protocol aims to maximize the information that can be extracted from the observed counts about the true expression distribution. Using Minerva's simulated counts, I calculated how much information differently weighted transcriptomes contained about the simulated ground truth by calculating the empirical MI (eMI) between the observed and simulated true counts.

I first calculated the eMI between the observed and simulated true counts per gene for a given cell population and dataset to assess the overall amount of information a given weighted transcriptome contains. Next, I fitted the KDE across the calculated eMI to visualize the overall distribution of MI for a given type of transcriptome for each dataset. As shown in Figure 6.6, the eMI distribution of weighted transcriptomes consistently contains more information than Non-Weighted single-cell transcriptomes. This is observed in both a positive shift in the overall density of the distribution and the increase in tail thickness of the right tail of the distribution. There are differences in the amount of skew and increase in tail thickness depending on how the transcriptome was weighted. For instance, Weighted Library

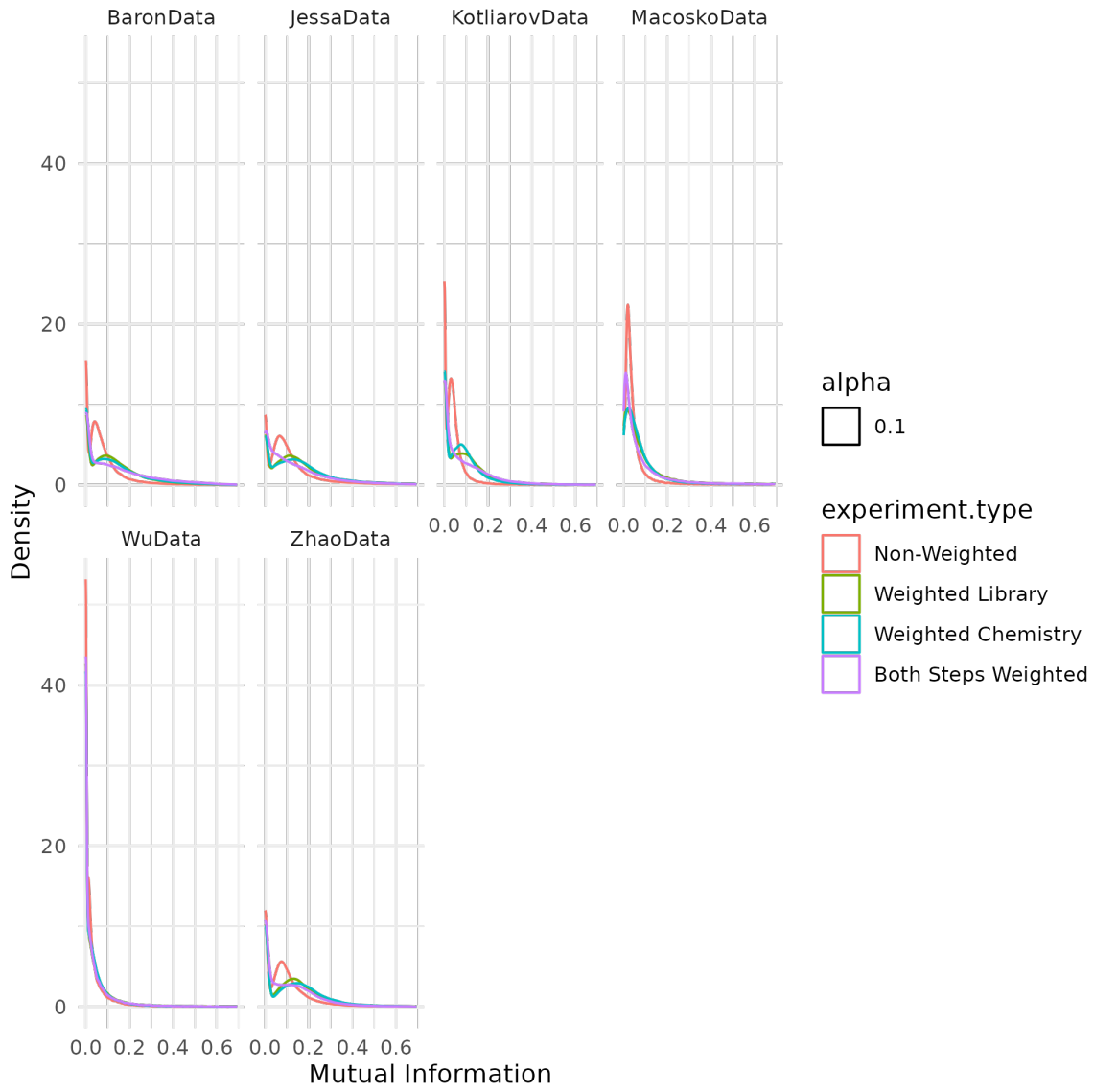


Fig. 6.6 One-dimensional KDE of the eMI of the transcriptome. eMI of an individual gene was estimated with the simulated observed counts of a gene vs their simulated true expression. With a simulated true mean expression, I can quantify how much information simulated observed counts contain about the original true. In this particular figure, I estimated KDE by Dataset

experiments show less of an increase in the overall informational content than Weighted Chemistry, and Weighted Library and Chemistry contain less information than Both Steps Weighted experiments. Both Steps Weighted transcriptomes eMI distribution varied from the other weighted protocols in terms of the shape of its distribution consistently forming a parteo-like distribution with a thick right tail across multiple datasets (see figure 6.6).

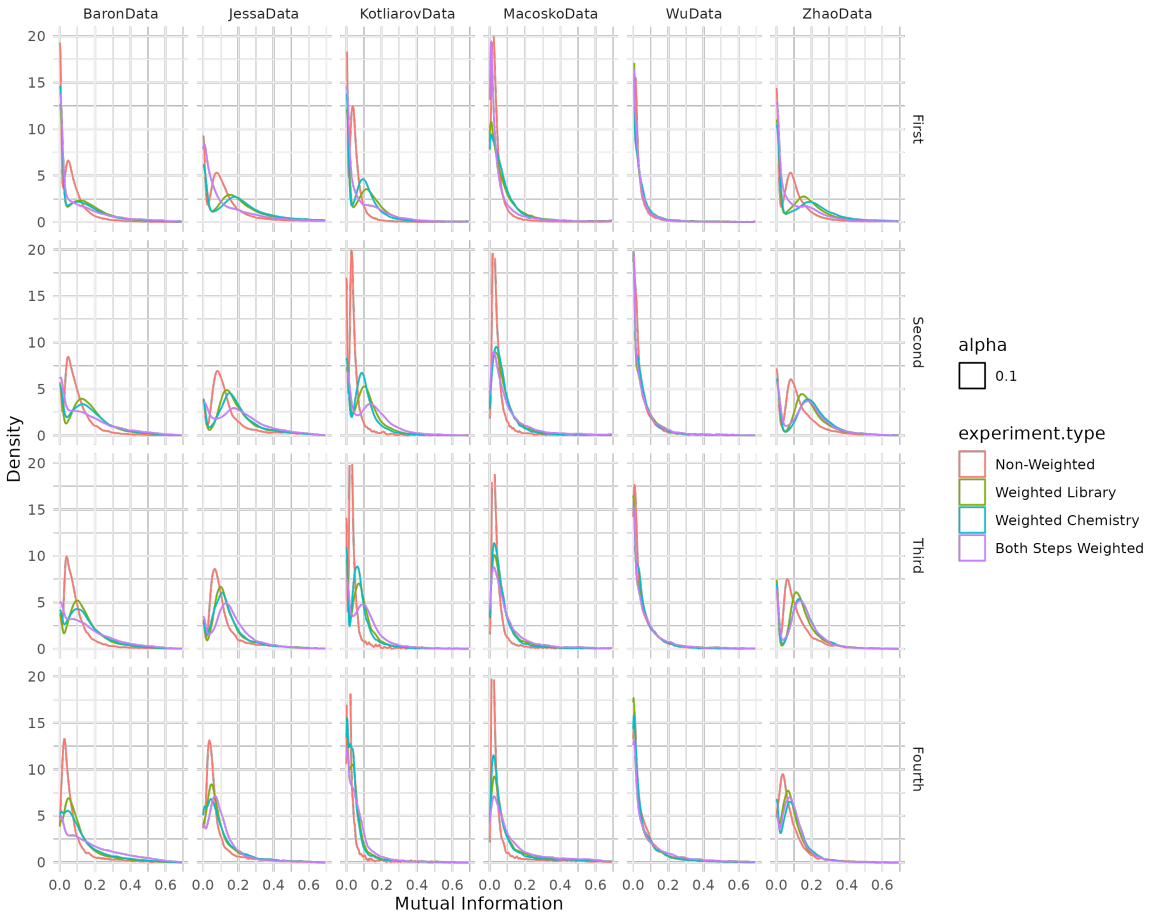


Fig. 6.7 One-dimensional KDE of the eMI of the transcriptome. eMI of an individual gene was estimated with the simulated observed counts of a gene vs their simulated true counts. With simulated true counts, I can quantify how much information simulated observed counts contain about the original true. In this particular figure, I estimated KDE by Dataset and Quantile.

However, it's important to point out that Boths Steps Weighted density at the lower end of the eMI distribution is significantly lower than all of the other weighted protocols. Indicating that Both Steps Weighted substantially increase the overall informational content compared to other weighted transcriptomes. However it's important to note that both Weighted Library and Chemistry also significantly increased the information they contain compared to Non-Weighted single-cell experiments as well.

Finally, I assessed the amount of information in weighted transcriptomes on a quantile basis, as seen in Figure 6.7. While the overall pattern for each quantile shows that weighted transcriptomes contain more information than Non-Weighted transcriptomes, with an overall shift in the density of the MI distribution and thicker tails, it is clear that the largest gains in

terms of information for weighted transcriptomes come from the first and second quantile. These quantiles have a larger shift in the distribution and thicker right tails. In contrast, in the third and fourth quantiles, the information gain primarily comes from an increase in the thickness of the right tail rather than a large shift in the density of the distribution. This indicates that most information gains come from genes of low to medium expressed genes.

6.3.3 Performance of Single Cell Normalization Methods in Weighted Transcriptomes

Normalization plays a crucial role in the analysis of Single-Cell experiments, aiming to eliminate technical variance while preserving biological variance. Technical sources of variance may arise from variations in sequencing depth, batch effects, and other factors. By removing variance attributed to technical sources, we can effectively identify and characterize cell populations, allowing the remaining observed variance to be driven by biological factors. Previous studies have assessed the performance of normalization techniques by quantifying their ability to eliminate technical variance and applying clustering algorithms to the normalized data for cell population identification [130, 4]. scRNA-Seq is primarily utilized to discover novel cell populations and subpopulations, heavily reliant on the ability of current single-cell bioinformatics methods to identify genes with the highest variance. These high-variance genes are then used as features for dimensionality reduction and population visualization. In this study, I evaluated the impact of different normalization techniques on the identification of pre-characterized cell populations using ARI as a performance metric. Specifically, I tested the performance of four normalization methods (SCTransform, GLM-PCA, Delta, and Sum Pooled Factors) across datasets and different transcriptome types, including both Non-Weighted and Weighted Transcriptomes.

The analysis results can be seen in figure 6.8, which shows no substantial differences between the performance of the various methods. The only noticeable difference is that SCTransform appears more consistent than the other normalization methods. However, these results are not surprising given the previous observations that there were no substantial alterations to the statistical characteristics of the weighted transcriptomes, except for the most highly expressed genes. Therefore, the use of current single-cell bioinformatics methods to identify cell populations should not significantly impact the accuracy or reliability of these methods' ability to accurately identify cell populations. These encouraging findings suggest that standard single-cell normalization and downstream analysis techniques can be effectively applied to weighted transcriptomes.

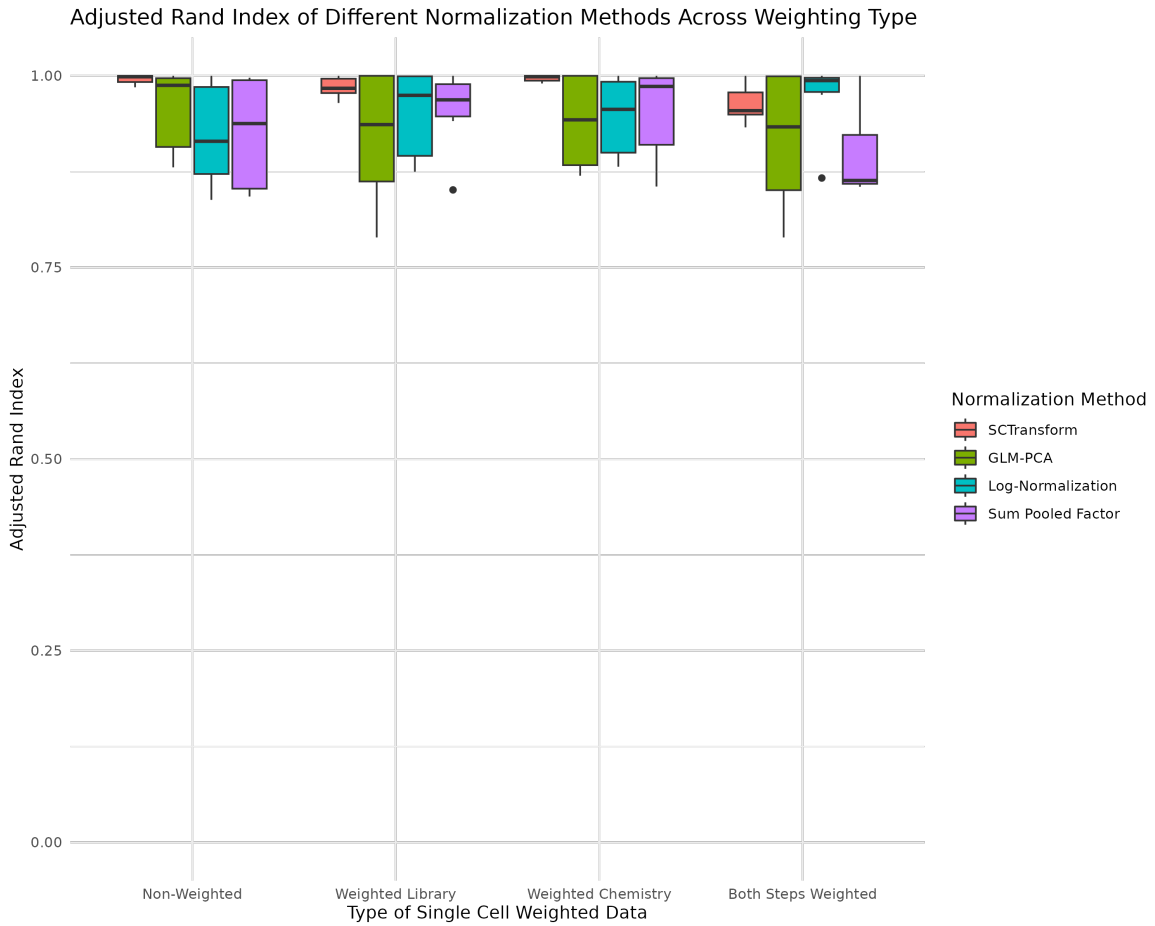


Fig. 6.8 Boxplots of the Adjusted Rand Index of Single Cell Normalization Methods by Type of Transcriptome.

6.4 Discussion

In this study, I aimed to investigate the impact of weighted single-cell experiments on the statistical characteristics of observed gene count distributions. Weighted transcriptomes represent a novel experimental approach where enrichment techniques are employed to modify the probability of sequencing a gene's transcriptome in a continuous manner. In contrast to targeted panel-based single-cell experiments that manipulate sequencing probability in a binary manner (i.e., a gene is either part of the panel or not), weighted transcriptomes enable a more gradual manipulation of sequencing probability. This allows for the enrichment of low to medium expressed genes that may not be readily observable in previous Non-Weighted single-cell experiments due to the winner-take-all effects of sampling without replacement.

Currently, three proposed weighting protocols are Weighted Library, Weighted Chemistry, and Both Steps Weighted. In the Weights Library protocol, enrichment occurs during

the cDNA Amplification step, altering the composition of PCR-amplified reads. In the Weighted Chemistry protocol, enrichment techniques are applied during or prior to the capture chemistry step of the Single-Cell Experiment, thereby manipulating the number of transcripts from a given gene that passes through the RT-PCR process. Lastly, the Both Steps Weighted protocol involves enrichment at both the Capture Chemistry and cDNA Amplification steps. When weighted transcriptomes were introduced in the previous chapter, it remained unclear whether manipulating the experimental protocol to increase transcriptome representation might also impact or compromise the information pertaining to the true expression distribution of a gene.

To explore the effects of these weighting protocols on the observed counts of the transcriptome, I conducted simulations using Minerva, a wet-lab-aware simulator introduced in the previous chapter. Each of the weighting protocols (Weighted Library, Weighted Chemistry, and Both Steps Weighted) was simulated, along with a Non-Weighted single-cell experiment, to assess their impact. To capture diverse biological contexts, eight publicly available datasets ranging from pediatric brain cancer to pancreatic islets were utilized. The results of the simulations revealed that, regardless of the weighting protocol employed, the statistical relationship between the mean and variance of gene expression was generally preserved for the majority of genes, including those with high expression levels. This indicates that the weighted protocols do not disrupt the statistical characteristics of the data but rather enhance the representation of gene expression distribution. However, it was observed that the weighted protocols tended to constrain outliers among highly expressed genes, which were more pronounced in the Non-Weighted data. This suggests that the increased representation of low and medium-expressed genes in the weighted protocols comes at the expense of limiting the expression range of the most highly expressed genes.

In addition to exploring the statistical characteristics of the observed counts for weighted transcriptomes, I also assessed the amount of information these datasets contained. I utilized MI, an information theory-based dependency measurement, to quantify the amount of information on a per-gene, cell population, and dataset level. I then estimated the distribution using KDE. The results showed that Weighted transcriptomes contained a clear increase in the information the observed count data contains about the true expression distribution used for simulations. This indicates that Weighted datasets provide an overall gain in information, suggesting that the trade-off of potentially destroying information on the most highly expressed genes is worth it, as it greatly improves the overall amount of information available.

In conclusion, this chapter investigated the effects of weighted transcriptomes on the statistical characteristics of observed gene expression counts. The results indicate that the

novel experimental protocol of weighted transcriptomes enhances the representation of gene expression distribution while not seriously disrupting the statistical characteristics of most genes. Weighted datasets exhibit a notable increase in information regarding transcriptomes, indicating that while there may be a trade-off in terms of potentially diminishing information on the most highly expressed genes, there is an overall gain in information primarily driven by improved characterization of low and medium expressed genes. This highlights the importance of considering the balance between preserving information on highly expressed genes and acquiring a more comprehensive understanding of gene expression across the entire transcriptome. These findings suggest that weighted transcriptomes offer a promising approach to enhance the overall representation of the transcriptome with minimal trade-offs. Furthermore, current single-cell normalization methods appear to be unaffected by the incorporation of weighted transcriptomes. However, additional research is necessary to investigate the potential applications of weighted transcriptomes in biological studies and verify whether the observed characteristics from simulations can be validated experimentally in a biological setting. These experimental validations will provide crucial insights into the practical utility and reliability of weighted transcriptomes in real-world biological scenarios.

6.5 Caveats and Limitations

There are two limitations in the current version of Minerva regarding its ability to simulate targeted panels and weighted transcriptomes. The first limitation is that weights used to simulate targeted panels were selected ad hoc, where all of the weights of a targeted gene is set to 30, and all other genes' weight is set to 1 in order to attempt to account for off-targets. It is unlikely that gene-specific probes in PCR or antibody pulldowns would behave in a uniform fashion, and they will definitely have off-targets. Due to this uniformity in the gene weights, Minerva's targeted panels likely represent an ideal or near-perfect targeted panel. However, it is unclear how to address this limitation until more publicly available single-cell targeted panel datasets exist. Once more data is available, I plan to develop a regression model, study probe off-target, and make changes in targeted panel gene weights to better reflect the data.

The second limitation is how I calculate gene-specific weights for weighted transcriptomes, utilising a target frequency and a genes' relative frequency estimated from a pilot experiment. This limitation is two-fold. First, all of my EDA into weighted transcriptome assumes that gene-specific weights are calculated based on relative frequency. If the weights change, there is a chance that the weighted transcriptomes' statistical characteristics will most likely change as well. Second, preliminary results from pilot experiments trying to

replicate my simulation conducted by Ryan Blake have indicated that gene-specific weights above one have little to no utility, placing an effect range of zero to one where gene weights can be effective. This is intuitive when considering the behaviour of PCR and antibody pulldown probes since increasing the concentration of a probe or primer does not result in an increase in the number of reads per transcript copy. On the other hand, decreasing their concentration lowers the maximum number of observable transcripts. As a result, reducing concentration has a more significant impact on mitigating sparsity in single-cell datasets compared to increasing concentrations.

To address these limitations in regard to calculating gene-specific weights for weighted transcriptomes, I plan first to develop an optimization function that'll allow me to calculate optimised gene weights between zero and one. Next, with a new set of gene weights, I will reconduct the EDA of weighted transcriptomes and confirm whether there are any major changes in its statistical characteristics.

Chapter 7

Conclusion

7.1 The Importance of Theoretical Analysis of Experimental Protocols

Experimental methods and protocols are typically developed through an iterative developmental process of trial and error. Relying upon the knowledge and expertise of biologists and chemists on what experiment they are trying to conduct. While this process has created a set of well-calibrated and rigorously tested protocols, this process is time and resource-consuming. In addition, as the complexity and scale of the experimental protocols increase, so too does the time, resources, and expertise required to develop novel methodologies. Mathematical models have been used in physics for centuries to understand phenomena' theoretical limits and design experimental protocols to validate hypotheses. Merging models with experimental designs enables physicists to conduct a theoretical analysis to identify optimal experimental protocols.

Biology has yet to develop an extensive theoretical framework to understand biological systems. However, this does not prevent us from developing mathematical descriptions of experimental protocols through intimate dialogue between biologists and computational biologists. While simulators are nothing new in the literature, their utility is frequently limited to bench-marking computational methods for normalizing or analysing novel data types. They are rarely used to optimize experimental protocols, if at all. Yet I would argue constructing theoretical models of experimental protocols offers us greater insight into the technology we depend upon and allows us to understand the true limit of the technology. By merging the practical knowledge of biologists and the mathematical skills of computational biologists, optimal experiments can be designed, and significant improvements in the technology can be identified and developed in a fraction of the time and resources.

A theoretical model of any technology or experimental protocol requires a firm mathematical description and an intuitive or explicit understanding of rules determining how they interact. Once a preliminary draft of this model is obtained, the validity of the model must be interrogated through conversations with biologists, in-depth searches through the literature, and constantly comparing the model's statistical characteristics in relation to known ground truth. However, once confidence is gained in the theoretical model of an experimental protocol, a powerful system representation is obtained that allows manipulation and exploration of varying parameters and experimental designs. This enables the identification of superior ways of conducting the experiment of interest or even key ways of improving it.

An example of where lack of developing a firm theoretical understanding of the experimental protocol, the various tools available, and how they can interact were GBCs in the first generation of scCRISPR Screens. In the first set of scCRISPR studies, GBCs were used as expressed labels that contained polyadenylated tails that could be captured by the 3' scRNA-Seq chemistry. At the time, 3' scRNA-seq was the only viable chemistry. They were located near the lentiviral vector's 3' long terminal repeat (LTR), which exposed the GBC to the risk of template switching during pooled lentiviral production. This was a previously known issue with the lentiviral reverse transcriptase template switching occurring every kilobase and the probability of switching events increasing to a maximum of 50% as a sequence nears the 3' LTR of the vector. To avoid this issue, Adamson et al., Dixit et al. utilised arrayed lentiviral production rendering the swapping mute as each gRNA-GBC combo was assembled separately. However, this wasn't entirely clear, and many screens were published that assembled the vectors using the pooled lentiviral production-based approach, exposing this to template swapping.

In the end, Hill et al. demonstrated through a combination of the theoretical analysis and experimental validation that the template swapping issue could be avoided by utilizing an alternative assembly method where the GBC inserted near the 3' LTR was copied and inserted in front of the 5' LTR, which reduced the intermolecular swapping rates to negligible levels [60]. This lentiviral assembly method used by CROP-seq improved the statistical power of scCRISPR screens and dramatically simplified the screens' experimental workflow [32, 60]. Throughout this thesis, I have greatly utilised this idea to identify ways of removing sparsity from single-cell sequencing experiments and determine the optimal design of scCRISPR Screens.

7.2 Contextualisation of Results

7.2.1 Optimizing scCRISPR Screens Statistical Power

scCRISPR screens hold the extraordinary potential to empower researchers by enabling them to investigate and understand the regulatory interactions of biological systems. Allowing for a high-resolution view of healthy and diseased tissues allowing for the potential development of in-depth computational models of entire tissues under varying conditions for the first time. However, despite this tremendous promise of scCRISPR screens, they are chronically underpowered. To better understand the underlying issues in the experimental design of these screens, I developed a simulator called crisprPower, capable of simulating every major aspect of the scCRISPR screen. From designing gRNAs, the method of sequencing, the number of off-targets, and the targeted genes. In addition, it can propagate perturbations across a synthetic GRN, enabling it to simulate perturbation effects. These factors enable me to efficiently search for various experimental designs to identify ways to improve the statistical power of scCRISPR Screens.

Simulations showed that the greatest hindrance to powered scCRISPR screens was the sparsity of scRNA-Seq. On-target or off-target activity of a gRNA, the various CRISPR perturbation methods barely affected the statistical power in any meaningful manner. The only way of improving the statistical power of scCRISPR screens was to use a targeted panel. This resulted in multiple orders of magnitude increase in statistical power, moving the number of cells per perturbation required to see an effect from 600 only to see the fourth quantile to 75 cells to see a perturbation response across the mean expression distribution. Once using a targeted panel, other choices in the experimental design begin to matter. CRISPR Knockouts are more effective at inducing perturbations than interference, assuming a protein-coding sequence is being targeted. Of course, the higher the on-target activity of a gRNA, the better.

To summarize, current scCRISPR screens suffer from being chronically underpowered; this result is backed by simulation. Significant improvements in terms of statistical power can be achieved through the use of a targeted panel. Once using a targeted panel, CRISPR should be used when targeting protein-coding sequences. By making these simple adjustments to the experimental protocol, it is possible to reduce the number of cells required per perturbation from 600-1000 to 75, dramatically improving the feasibility of the utility of scCRISPR Screens.

7.2.2 Removing Sparsity from Single Cell Sequencing Experiments

To better understand Single Cell Sequencing experiments, I developed a novel wet-lab-aware simulator called Minerva to conduct a theoretical analysis of single-cell sequencing experiments. Specifically, I designed Minerva to simulate the three primary phases of single-cell sequencing: capture chemistry, reverse transcription, and cDNA amplification allowing for potential experimental manipulation in each phase. Using Minerva, I could rapidly search through the various experimental designs and came across three methodologies for removing sparsity from scRNA-Seq data: Library Biased, Chemistry Biased, and All-Steps Biased. Each method utilizes an enrichment protocol to alter gene transcripts' probability of being sequenced through careful calculations of individual gene transcripts' probabilities and be altered by adjusting the concentrations of enrichment protocol on a per gene basis. To calculate this, either a pilot experiment or a reference dataset of the biological system being investigated is needed. Each of these methods. The primary difference between these methods is in which phase of the single-cell sequencing experiment the enrichment protocol is applied. I refer to the set of these experimental protocols as weighted transcriptomes.

Altering the sequencing probability of gene transcripts introduces concerns about potential artificial manipulation or altering the statistical properties of the observed counts data generated from a weighted transcriptome. To determine what effects weighing the transcriptome has on gene mean expression and variance, I conducted exploratory data analysis. I observed that most genes across the first, second, and third quantiles showed little to no evidence of any alteration in their statistical properties. While a few genes exhibited changes in the statistical properties, these genes were on the highest end of the quantile. To assess the impact these changes could have on downstream stream single-cell analysis, we assessed the performance of various normalization methods and the ability to identify simulated cell populations accurately. I observed that regardless of the weighted protocol and normalization methods, simulated cell populations were accurately identified, indicating that any alterations in the statistical characteristics of highly expressed genes had a minimal impact on downstream analysis.

These results show how a theoretical analysis of single-cell sequencing experiments can significantly improve experimental protocols. Sparsity is frequently a confounder in analysing single-cell datasets. It introduces uncertainty and inhibits researchers' ability to investigate biologically relevant genes that are frequently lowly expressed, such as transcription factors. In addition, exploratory data analysis revealed that the introduction of experimental weights had a minimal effect on the statistical characteristics of gene expression mean and variance relation while also increasing the transcriptomes resolution.

7.3 Future Directions

7.3.1 Improving scCRISPR Screens

There are numerous opportunities to continue and expand upon my work. In particular, there are three major opportunities that I am going to continue working on. The first opportunity is to refactor crisprPower from Python into R, allowing me to integrate my work with scCRISPR with Minerva and investigate the potential of using weighted transcriptomes. Additionally, it makes my simulator easier for researchers to access and use to design their own scCRISPR screens. Second, I need to validate the results of my simulation experimentally. Finally, Off-targets are seldom explored aspects of CRISPR screens, but scCRISPR screens pose a particular challenge. However, while most view the effects of off-targets as primarily noise or bias to be removed, it may be possible to exploit them to extract even more information about biological systems from the screen.

Traditional methods of assessing On-Target activity have historically been measured in terms of observable perturbations, such as growth or dropout, that must significantly affect essential genes and exert a noticeable impact on growth rates. In a Single-Cell Experiment, a given perturbation can now theoretically be observed for any potentially perturbed gene, thereby increasing the amount of observable noise. Traditionally, this noise is addressed by employing multiple gRNAs to regress out its effects, as commonly done in bulk RNA-Seq experiments. However, this perspective assumes that off-target effects are inherently 'bad' and need to be minimized. Instead, I would suggest that for the purpose of studying biological phenomena, off-target effects, if known in advance or detectable, can offer valuable insights into the function of individual genes and biological systems as a whole [127]. From this perspective and within the experimental context, off-target effects represent a potential treasure trove of useful information that researchers can exploit to gain further insights and answer their research questions.

Refactoring crisprPower from Python to R

Experimental design is an important question for any biological experiment typically overlooked by most researchers. This is primarily because, for most researchers, it is unclear how to define an optimal experimental design. I would argue there is no such thing as an optimal experiment, and it's principally defined by a bivariant relationship of statistical power and cost, with the goal being to optimize the statistical power of an experiment at a reasonable cost. For scCRISPR finding the optimal experimental design is even more important due to the current high costs of these experiments. To help researchers maximise their resources for

the greatest impact possible, I developed crisprPower, a statistical simulator of scCRISPR screens that enables users to assess the statistical power of their screens and how manipulating experimental parameters alters their power.

crisprPower was originally developed before Minerva. I initially wrote most of my code using Python, which is why crisprPower was originally developed using it. However, as I made a breakthrough in removing sparsity from single-cell sequencing experiments, I had to write the code for Minerva in R due to the package dependencies. With crisprPower already written, I decided to keep it in Python. However, this is no longer feasible with the release of the crisprVerse and the development of weighted transcriptomes [61]. The need to refactor crisprPower from Python to R has become necessary.

The crisprVerse is a software ecosystem that integrates a multitude of computational tools for the design of gRNAs. In particular, it marries all of the major on-target and off-target tools for all major CRISPR modalities (i.e. nucleases, dead-Cas9, and base editing), allowing researchers to quickly access all these tools through a convenient interface within a single R package. Originally developed by Hoberecht et al., it aims to improve the overall quality of gene perturbation studies by identifying high-quality gRNAs that these studies depend upon. Currently, the package has extremely broad applicability as it only focuses on gRNA design in terms of on and off-target. I wish to integrate an R implementation of crisprPower to allow researchers also to explore various experimental protocols with their gRNAs and determine the statistical power of their experiment. Addressing a key area that the current crisprVerse ecosystem lacks.

A second benefit of refactoring crisprPower in R is that I can integrate it with Minerva and rerun my power analysis to include weighted transcriptomes. Weighted transcriptomes are particularly suited to scCRISPR screens as, unlike observation studies, it is normal in perturbation studies to have a fairly good understanding of the steady state of the biological system of interest. Utilizing such an understanding, we calculate gene-specific odds-weighted transcriptomes that may enable whole transcriptome-wide readouts without a loss of statistical power that was previously observed occurs in normal scCRISPR screens. Presenting us with an exciting opportunity of removing the need to conduct a targeted transcriptome altogether and fulfilling the true promise of scCRISPR screens of high throughput and high resolution.

Mostly, the refactor will not significantly change how crisprPower works. There are only three areas I plan to change regarding the parameterising of genes and the statistical expression model. I plan on using an alternative statistical simulator equivalent to Minerva's numerical simulator. Specifically for parameterising gene expression, instead of sampling from predefined distributions I have previously fitted, I plan to utilize reference datasets 100% of the time. Specifically, I utilise Minerva to estimate gene-specific parameters of the

reference dataset. Then I will sample a given cell population, and within this cell population, I will sample gene parameters based on the number of TFs, E-Genes, and HKs being simulated. Finally, I will utilize sampled GRNs to assess the influence of off-target effects on Differential Expression and Network Reconstruction methods. The goal is to develop approaches capable of detecting and leveraging off-target effects of a gRNA, enabling us to extract additional information from the GRN.

In crisprPower, I currently utilise a statistical simulator similar to the one developed for scPower [121]. It estimates the gene expression distribution from a reference dataset similar to Minerva. The primary difference between the two is Minerva scales the gene expression distribution to simulate biological variance, allowing for alternative single-cell sequencing methods. For the Minerva equivalent statistical simulator, I will calculate a gene's observed mean expression based on the mean of the biological variance and the effects different single-cell sequencing methods have upon it. Specifically, I will start by calculating $E[CM_i]$, the mean capture molecules per gene i as a mean from the MFNH distribution $meanMFNH$ for a cell excepted capture pool $E[CP_j]$. Next, I will reduce these mean captured molecules by half to simulate the effects of RT-PCR on counts giving us the $E[MSP_i]$ (excepted molecules in sequencing pool per gene i). Finally, the mean observed counts per gene i are calculated by finding the mean of the observed molecules sequenced by calculating the mean from the MFNH for a cell excepted cell library $E[CL_j]$. By calculating each of these steps separately, I can simulate the effects of the weighted transcriptomes statistically on the mean expression of genes. Combining these current capabilities of crisprPower will allow me to reconduct power analysis of scCRISPR screens and include in this new analysis weighted transcriptomes. For greater details in the statistical simulator of Minerva, see the equations below:

$$E[CM_i] = meanMFNH(n = E[CP_j], m = X_i, o = CMP_{odds}) \tag{7.1}$$

$$E[MSP_i] = E[CM_i] * \text{rt.pcr.constant} \tag{7.2}$$

$$E[Y_i] = meanMFNH(n = E[CL_j], m = E[MSP_i], o = SP_{odds}) \tag{7.3}$$

Exploiting Off-Targets for Information

CRISPR off-targets are a chronically understudied aspect of the technology and face numerous technical challenges. Primarily methods follow two broad camps: computation and wet lab-based approaches [61]. Computational methods represent the high-throughput method of exploring CRISPR off-target relying upon a combination of alignment rules (i.e. identify all

off-target sites with up to two mismatches) and machine learning to predict off-target effects. Examples of these scores are the MIT or CFD score [54, 37]. Computational models are typically trained on a subset of gRNAs where off-target effects can be measured using flow or sequencing-based experiments. A key assumption is these gRNAs generalise to the rest of the genome's target sites and, therefore, can be used to make predictions. However, it is known that computational models, while useful for initial filtering, do not fully characterise the behaviour of gRNAs as such wet-lab methods as CIRCLE-Seq and Discover-seq must be conducted to validate predicted off-target behaviour of gRNAs, especially in clinical application [146, 131].

From a data analysis perspective, CRISPR off-targets are traditionally considered sources of noise and reduce how informative a given experiment can be; therefore, it has become increasingly important to remove their effects from the analysis of the experiment. As the resolution of the experiment increases, it is easier for off-targets to inject noise into the data of an experiment. When conducting a dropout screen, the number of off-targets that can affect the growth of cells is relatively small and is restricted to the essential gene set of the cells of an organism. Any observation of an off-target effect has to be large in itself. However, increasing the screen's resolution with images or transcriptomes makes it easier for off-target inject noise into the experiment. As such scCRISPR screens are extremely prone to noise being injected by off-target effects of gRNAs, as all an off-target has to do is alter the expression of a gene. While altering the expression of one gene is not a huge deal altering 10s or 100s of genes is a different story. The traditional view of an off-target effect is centred on the perspective of the target site. However, is this view the only way of viewing off-target effects?

Instead of centring our perspective on the target site, it should be centred on the gRNA itself. Once viewed from this perspective, the readouts of gRNAs can be thought of as a set of mixed signals, with each gRNA providing us with a partial view of the source signals (i.e. target genes, enhancers, promoters, etc.). If this is the case, we can utilize Blind Source Separation algorithms such as Independent Component Analysis to identify and separate the source signals from the gRNAs mixed signals. Further research is required to determine what algorithms could be used or whether a new method must be developed. Despite shifting our perspective from the traditional view to this one, it may be possible to extract even more information from scCRISPR Screens than previously thought possible. Designing less than one gRNA per gene experiment may be possible by accounting for overlapping perturbation effects so that functionally, multiple gRNAs perturb the same genes. Still, they target so many other genes simultaneously the total number of gRNAs required is less than the number of genes being perturbed.

7.3.2 Improving Weighted Transcriptomes

There are three primary areas of opportunities for future work on weighted transcriptomes. The first opportunity is to conduct validation experiments to confirm Minerva's predicted statistical behaviour of weighted transcriptomes. The next challenge is removing all the technical biases introduced during the experiment. Despite the previous chapter showing that altering the sequencing probability of transcripts is minimal, there is some effect. Removing the last effect would alleviate any concerns researchers may have. Finally, generalising gene odds would be another great area to work on. Weighted transcriptomes require a pilot experiment or a reference dataset to calculate gene-specific odds for an experiment, adding additional costs. These opportunities are worth pursuing, and I have developed initial plans to achieve them.

Experimental Validation of Theoretical Results

Three key characteristics must be confirmed to validate the theoretically predicted behaviour of weighted transcriptomes. First, before anything else, I must demonstrate that altering the enrichment protocols chemistry by increasing or decreasing its concentration shifts experimental measurement. Second, Minerva's model predicts that for a homogeneous cell population, weighted transcriptomes whose odds were calculated for this particular population should generate a mean of one for all observable transcripts. Finally, the previous chapter demonstrated that when gene odds are calculated, bulk data represents a mixture of heterogeneous cell populations it was still possible to identify individual cell populations consistently. To validate these predictions, I have developed three experiments to validate a particular prediction of Minvera.

The first experiment will be a qPCR study and aims to validate that altering the enrichment protocol chemistry results in alterations of experimental measurements. This experiment will utilize 20 target genes, and our reference housekeeping gene will use GAPDH. In addition, to the normal three replicates of a qPCR experiment, we plan to repeat the experiment in two different cell lines, MCF7 and T47D. To manipulate the qPCR primer concentration of the target gene and a reference housekeeping gene to move their CT values closer together and, if possible, roughly equal. To calculate the target and reference gene genes in the same manner described in Chapter 2, the only difference is that it's two instead of 18-20 thousand genes. If the experimental data observed confirms that manipulation of PCR primer chemistry is possible and ideally behaves as expected will proceed to the second experiment.

In the second experiment, I will utilize the same cell lines from the qPCR experiment; however, this experiment will conduct a TAP-Seq experiment developed by Schraivogel et al..

The cell lines will be running separately to create homogenous cell populations. As described in Chapter 5, for each cell line will calculate the gene-specific odds for all the genes in the l1000 gene list. The estimated gene-specific odds are excepted to generate a mean of one for all l1000 genes observed counts. Important limitations to this experiment are that we only use the l1000 gene list and not the whole transcriptome. We've made this decision due to the cost of the order of the inner and outer PCR primers required to run TAP-Seq. We are currently not testing alternative enrichment protocols like the one developed by Replogle et al., which uses an antibody pulldown of hybridized probes. Finally, the last major limitation of this experiment is that we're only conducting a library-biased weighted transcriptome experiment. We cannot experimentally validate the biasing of capture chemistry or all steps weighted transcriptomes with the currently available wet-lab methods.

Finally, the third and final experiment we plan to conduct will be to repeat the same l1000-based TAP-Seq experiment, but instead of using homogenous cell lines will use Peripheral Blood Mononuclear Cells (PBMCs) to validate that we can identify known cell populations. In addition, I will utilize publically available PBMC datasets to compare the observed gene expression distributions and validate the results of the EDA conducted in Chapter 6. The ideal result of this experiment would be that we can both accurately identify cell populations observed in previous experiments and do not observe any significant alteration of observed count distributions except in the highest outliers of the mean expression distribution.

I anticipate observing some discrepancies between these experiments and Minerva's models, primarily due to variations in enrichment chemistry's impact on both on-target and off-target effects. As we carry out these experiments, particularly after the second one, it will be necessary to refine the model further to incorporate off-target behavior. Another avenue for enhancing Single Cell Experiments is the improvement of RT-PCR, which is an active area of research. However, despite ongoing efforts, there has been limited progress to date. Moreover, while enhancing RT-PCR could lead to an expected performance boost, akin to optimizing capture chemistry weighting, it still doesn't address the challenges posed by cDNA amplification and the imbalances introduced by varying gene transcript abundances. Therefore, some form of weighting cDNA amplification will likely remain essential to fully optimize Single Cell Experiments.

Finding Optimal Gene Weights and Generalizing beyond a particular biological context

In this thesis, I have proposed a novel single-cell sequencing method of weighted transcriptomes, which holds tremendous promise to dramatically reduce the overall sparsity in single-cell datasets, enabling us to achieve the high resolution and throughput the technology promises. Despite this promise, weighted transcriptomes face two major challenges, and the

first is determining the optimal weights for a transcriptome for a given biological context. Second is the requirement of a pilot experiment or reference dataset to calculate gene-specific odds for every experiment, which increases the time required and the overall cost.

I calculate gene-specific weights using a simple heuristic dividing the observed gene frequency over a desired target frequency. This serves as a reasonable starting point for exploring the utility of weighted transcriptomes. However, this heuristic is fundamentally limited and sub-optimal; therefore, I need to develop a new method of calculating gene weights w_g is needed by reframing gene-specific weights as a budget optimization problem where I try to minimize the percentage of molecules not sequenced on a per-gene basis q_g allowing low and medium-expressed genes to have equal importance as high-expressed genes instead of minimising the percentage of molecules as a whole. By summing all of the q_g genes together, we get Q, providing us with the basis for an optimization function to minimize (see equation 7.7). Specifically, to minimize Q, two sets of weights need to be optimized O_g and w_g regarding the number of reads B that will be sequenced. W_g represents the weights at the capture chemistry step of a single-cell sequencing experiment; this process is modelled using an MFNCH distribution. To optimise, I calculate the excepted molecular count m_g as the weight changes for n sample size of N mRNA in the cell with X_g mRNA for gene g as seen in equation 7.4).

$$m_g = \text{meanMFNCH}(n, N, X_g, W_g) \tag{7.4}$$

Once estimated, m_g, the relative frequency of the molecules does not change as manipulating the cDNA step of the single-cell sequencing experiment only alters the reads to umi ratio for a given gene. To account for this, I calculate the excepted umi count y_g as a function of the number of reads that belong to a given gene g, which can be manipulated via the weight w_g as seen in equation 7.5.

$$y_g = m_g \left(1 - \frac{m_g}{\left(B \frac{m_g w_g}{\sum m_g} \right)} \right) \tag{7.5}$$

Once estimated, calculate q_g as the percentage of molecules not sequenced based upon the known number of molecules and the excepted observed umi count as seen in equation 7.6. Using this new optimization method, I plan to search for better gene weights and explore different limits being applied to them, such as weights only being between one and zero.

$$q_g = 1 - \frac{y_g}{X_g} \tag{7.6}$$

$$Q = \sum q_g \tag{7.7}$$

Using the optimization function described above, I will identify and develop a series of new generalized sets of gene-specific weights that can be used in multiple biological settings. To identify these sets of weights, I plan to conduct an exploratory analysis to identify biological settings where similar weights can be reused. To do this, I will fit a meta-distribution of the mean and variance in expression for each gene from various biological settings within a particular organism. This analysis requires a large dataset with extremely diverse biological settings. Fortunately, such datasets exist with the Human and Mouse Cell Atlas, which I plan to use for this analysis [87, 43].

Before I can fit the meta-distribution of genes mean and variance, I must first fit a cell type-specific gene expression distribution. To do this, I'll first normalize the data using the normalization method described above to provide me with estimates and uncertainties of the true transcript count in a given cell population. Next, I'll fit the gene expression distribution using a NB, a distribution traditionally used to describe biological count datasets. Recent studies in transcriptional kinetics have indicated the NB distribution is best first for the steady distribution under all currently proposed transcriptional kinetics models [7]. From the fitted NB, I'll take each cell type's mean and dispersion parameter for a given gene to fit the meta-distribution. Once I have a gene's meta-distribution, I will identify the sub-populations within it using infinite mixture modelling, thereby automatically identifying the optimal number of peaks. Multiple peaks would indicate fundamentally different regulatory processes influencing its behaviour, thereby requiring more than one unique gene weight for a given gene.

Finally, I will conduct validation experiments for every set of gene-specific odds created. Following a similar pattern to the third experiment in the experimental validation subsection above. The primary difference here, however two-fold using the generalized gene-specific odds and utilizing samples from the particular biological context a given set of odds should be used for. This experiment would aim to validate we can accurately identify known cell populations and again assess the effects these odds have on observed counts of genes and whether they inject any technical bias into the dataset.

Differential Expression Analysis of Weighted Transcriptome

Differential expression analysis of single-cell datasets remains an open question, with new methods being continuously developed to account for the technical bias introduced by single-cell sequencing experiments. While my exploratory analysis of weighted transcriptomes

in Chapter 6 indicates that the technical biased introduced by weighted transcriptome is minimum, a new method differential expression method with a model to account for any potential bias that may exist explicitly is needed. This new method aims to accurately estimate the distribution of transcript counts X_i given the observed counts y_i and then to be able to compare the distribution between the cells of one or populations or conditions and identify any differences between the distribution. To achieve this, I assume the following assumptions of a cell expression distribution:

1. A given gene's expression distribution is at a steady state within a given cell population or condition
2. The number of mRNA transcripts of a given gene randomly fluctuates according to a given gene's expression distribution
3. The number of mRNA molecules within a cell is the sum of all the randomly fluctuating transcripts. Which I refer to as the cell size
4. A given cells library size is the product of the sum of all mRNA molecules in the cell and the capture chemistry efficiency
5. The observed counts of a given gene per cell are the result of both cell and gene-specific capture chemistry

Based on these assumptions, the differential expression method must account for the technical bias introduced on both the cell and gene levels and remove these effects. Cell-specific bias is currently viewed in terms of sequencing depths (i.e. the difference in library size); to account for this during normalization, a size factor is estimated and all cells are then adjusted to the same sequencing depth. Gene-specific biases are currently viewed in terms of the heteroskedasticity of high expressed genes; to account for this, a log transformation is applied to the size factor adjusted counts [91]. However, this view of normalization only partly addresses the problem and fails to see the immediately observed UMI counts as a source of technical bias themselves. For weighted transcriptomes, I need to account for the technique bias introduced by gene-specific weights upon the observed counts for a group of cells from a given condition. The easiest way to account for and remove this bias is to infer a given gene's true transcript count distribution, which requires an estimate of the excepted cell size for a group of cells.

During single-cell sequencing experiments, a number of mRNA from the true transcript counts are sampled randomly based on the experiment's capture chemistry efficiency. The sum of sampled counts is referred to as the cell library size, which can be thought of as a

sample from a conjugate distribution where the percent of molecules captured reflect capture efficiency from a cell's randomly fluctuating cell size. Previously, this has been described using a Beta-Binomial where the Beta distribution represents capture chemistry efficiency, and the binomial distribution represents the downsampling based on a sampled p from n cell size [138]. However, the Beta-Binomial distribution implicitly limits the cell size of a cell population. In keeping with my assumption, the cell size is a random variable; a Beta-Poisson distribution would better fit biological differences in cell size and technical differences in capture chemistry efficiency. In this model, the lambda of the Beta-Poisson would be the excepted cell size for a given cell population, and the alpha and beta would describe the capture chemistry efficiency of the single-cell sequencing protocol. In addition, a Beta-Poisson's parameters are easier to infer than a Beta-Binomial. Multiple options are also available to infer the parameters, including method of moment (see equations **??**), maximum likelihood estimate (see equations 7.18), and approximate Bayesian computations (see algorithm 14) [129].

Weighting the transcript introduces a large potential set of challenges combined with the issues sampling without replacement nature of observed counts, making it unclear how to adjust them to remove gene-specific bias. Due to the difficulties of removing gene-specific bias on the observed counts, it would be easier to infer the latent transcript count distribution of a given cell population conditions on the observed counts and gene-specific weights used in the experiment. To do this, an initial estimate of the transcript counts given the observed accounts is needed; this can be achieved using Cornfield's Approximation [81]. To approximate transcript counts, I need an estimate of the excepted cell size lambda parameter from the estimated Beta-Poisson and the sample mean of the observed counts. This will provide me with a reasonable starting point to then use Markov Chain Monte Carlo (MCMC) (see Algorithm 15) or Variational Inference (see Algorithm 7) to infer the latent transcript count distribution of a given gene. The primary difference between these two methods is that Variational Inference requires a conjugate prior distribution assumed for the transcript count distribution, which I chose the gamma distribution.

Beta-Poisson Method of Moments:

$$M_1 = \frac{1}{Q}\sum_{i=1}^{Q} cl_i \tag{7.8}$$

$$M_2 = \frac{1}{Q}\sum_{i=1}^{Q} cl_i(cl_i - 1) \tag{7.9}$$

$$M_3 = \frac{1}{Q}\sum_{i=1}^{Q} cl_i(cl_i - 1)(cl_i - 2) \tag{7.10}$$

$$r_1 = M_1, r_2 = \frac{M_2}{M_1}, r_3 = \frac{M_3}{M_2}, \tag{7.11}$$

$$\lambda = \frac{2r_1(r_3 - r_2)}{r_1 r_2 - 2r_1 r_3 + r_2 r_3} \tag{7.12}$$

$$\alpha = \frac{2(r_3 - r_2)(r_1 - r_3)(r_2 - r_1)}{(r_1 r_2 - 2r_1 r_3 + r_2 r_3)(r_1 - 2r_2 + r_3)} \tag{7.13}$$

$$\beta = \frac{2r_1 r_3 - r_1 r_2 - r_2 r_3}{r_1 - 2r_2 + r_3} \tag{7.14}$$

Beta-Poisson MLE:

$$dPB(x;\alpha,\beta,\lambda) = \begin{cases} K(\lambda, \alpha + x, \beta + \alpha + x) & \text{if } x \leq 0 \\ \frac{\prod_{i=0}^{\lfloor x-1 \rfloor}(\alpha + \text{sign}(x-1)\cdot i)}{\Gamma(x+1)} \cdot \lambda^x \cdot K(-\lambda, \alpha + x, \beta + \alpha + x) & \text{if } x > 0 \end{cases} \tag{7.15}$$

$$\mathscr{L}(\alpha,\beta,c) = \prod_{i=1}^{n} \text{dpb}(x_i;\alpha,\beta,c) \tag{7.16}$$

$$\ell(\alpha,\beta,c) = \sum_{i=1}^{n} \log(\text{dpb}(x_i;\alpha,\beta,c)) \tag{7.17}$$

$$\mathscr{L}(\hat{\alpha},\hat{\beta},\hat{c}) = \arg\max_{\alpha,\beta,c} \left[\sum_{i=1}^{n} \log(\text{dpb}(x_i;\alpha,\beta,c)) \right] \tag{7.18}$$

7.4 Future Prospective of Single Cell Sequencing Experiments

Single-cell sequencing experiments have significantly enhanced researchers' ability to examine organs, tissues, and various other biological entities. These experiments offer high

Algorithm 5: ABC inference of Beta-Poisson via Method of Moment

Input: Cell Library Size x ;
Output: Point estimates cell population excepted cell size and capture chemistry θ ;
Initialization: Set $M5\%$ as the 5th percentile and $M95\%$ as the 95th percentile of 1000 bootstrapped MME estimates;
for $i = 1$ **to** N **do**
    Draw Beta-Poisson parameters set $\theta^* \sim \pi(\theta)$ **if** $M1(\theta^*) \in \{M5\%, M95\%\}$ **then**
        Simulate data cl^* by sampling cell library size from Beta-Poisson using θ^* ;
        Calculate Hellinger distance $H = d(x, x^*)$;
        Add H to vector of distances $H = (H_1, H_2, \ldots, H_{10000})$;
    end
end
Sort vector H in ascending order ;
Select the parameter sets within the lowest 5% of H as accepted ;
Compute medians of the accepted parameter sets as point estimates for kinetic parameters θ ;
Return: θ

resolution and increased throughput, enabling not only observational studies but also large-scale genome-wide perturbation experiments through scCRISPR screens. By leveraging the high throughput nature of single-cell sequencing, scCRISPR screens have opened the door to exploring cell-specific regulatory networks and facilitating network reconstruction. Through the analysis of these reconstructed networks, it becomes possible to interrogate biological systems and identify crucial regulatory genes that control network dynamics, serving as potential targets for drugs or other manipulations. However, the statistical power of current scCRISPR screens and single-cell sequencing experiments is limited due to implicit experimental protocols.

The work in this thesis addressed these challenges and proposed alternative experimental protocols to overcome the limitations of statistical power in scCRISPR screens. Currently, overcoming the statistical limitations of scCRISPR screens and observing the effect of a perturbation requires the use of a targeted panel. However, the use of a targeted panel restricts the experiment's resolution, potentially discarding valuable information. To overcome this limitation, I propose an alternative experimental protocol in which the manipulation probability of sequencing a transcriptome is weighted, allowing for a more continuous manipulation and resulting in the weighting of gene transcripts.

By integrating scCRISPR screens with weighted transcriptomes, researchers will be able to investigate any biological system, reconstruct its regulatory network, and model further manipulations with a greater degree of precision than previously possible. This expansion

Algorithm 6: Metropolis-within-Gibbs, unknown M_1 and N

1 Choose initial values $N^{(0)}, M_0^{(1)}$;
2 **for** $t \leftarrow 1$ *to* T **do**
3 Draw M_1^* from a proposal distribution $q_t(M_1^*|M_{t-1}^{(1)})$, e.g. $M_1^* \sim \text{Pois}(M_{t-1}^{(1)})$;
4 Compute the acceptance ratio ;
5

$$\gamma_{M_1} = \min\left\{1, \frac{\pi(M_1^*|N_{t-1}, w, n, x_1)}{\pi(M_{t-1}^{(1)}|N_{t-1}, w, n, x_1)} \frac{q_t(M_{t-1}^{(1)}|M_1^*)}{q_t(M_1^*|M_{t-1}^{(1)})}\right\}$$

Draw $u \sim \text{Unif}(0,1)$;
6 **if** $\gamma_{M_1} > u$ **then**
7 Set $M_t^{(1)} = M_1^*$;
8 **end**
9 **else**
10 Set $M_t^{(1)} = M_{t-1}^{(1)}$;
11 **end**
12 Draw N^* from a proposal distribution $q_t(N^*|N_{t-1})$, e.g. $N^* \sim \text{Pois}(N_{t-1})$;
13 Compute the acceptance ratio ;
14

$$\gamma_N = \min\left\{1, \frac{\pi(N^*|M_t^{(1)}, w, n, x_1)}{\pi(N_{t-1}|M_t^{(1)}, w, n, x_1)} \frac{q_t(N_{t-1}|N^*)}{q_t(N^*|N_{t-1})}\right\}$$

Repeat lines 6-11 for N using γ_N ;
15 **end**

of single-cell experiments transcends their traditional use for observational studies focused on identifying cell populations and monitoring changes in cell type composition. Instead, it enables a more proactive experimental setting, facilitating the identification of novel drug targets for therapies, determination of their mechanism of action, and even the assessment of off-target effects through comprehensive characterization of biological systems' responses to perturbations.

In conclusion, the resolution of statistical limitations in scCRISPR screens might be the beginning of a second revolution propelled by single-cell technologies. The approaches I have presented throughout this thesis can empower researchers to adopt a more proactive approach to investigating biological systems. The potential applications of this technology will grow as new CRISPR variants continue to expand the range of aspects in biological

Algorithm 7: Variational Inference for Gamma-Fisher Noncentral Hypergeometric unknown M_1 and N

Initialize variational parameters: $\alpha_N^q, \beta_N^q, \alpha_M^q, \beta_M^q$;
for $t \leftarrow 1$ *to* T **do**
    Sample $\theta_N \sim \text{Gamma}(\alpha_N^q, \beta_N^q)$;
    Sample $\theta_M \sim \text{Gamma}(\alpha_M^q, \beta_M^q)$;
    Compute ELBO:
    $\text{ELBO} = \mathbb{E}_q[\log p(\text{obs}|\theta_N, \theta_M)] - \mathbb{E}_q[\log q(\theta_N, \theta_M)] + \mathbb{E}_q[\log p(\theta_N, \theta_M)]$;
    Update variational parameters: $\alpha_N^q, \beta_N^q, \alpha_M^q, \beta_M^q \leftarrow \text{UpdateRule}(\alpha_N^q, \beta_N^q, \alpha_M^q, \beta_M^q)$;
end

systems that can be perturbed. The opportunity to elucidate and map the genotypic effects on observed phenotypes, spanning from cell populations to whole organisms, has never been more promising. The future holds immense possibilities limited only by our imagination.

References

[1] Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., NuÃśez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., and Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, 167(7):1867–1882.e21.

[2] Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1):1911. Number: 1 Publisher: Nature Publishing Group.

[3] Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C.-Z., Ben-David, U., Cook, A., Ha, G., Harrington, W. F., Doshi, M. B., Kost-Alimova, M., Gill, S., Xu, H., Ali, L. D., Jiang, G., Pantel, S., Lee, Y., Goodale, A., Cherniack, A. D., Oh, C., Kryukov, G., Cowley, G. S., Garraway, L. A., Stegmaier, K., Roberts, C. W., Golub, T. R., Meyerson, M., Root, D. E., Tsherniak, A., and Hahn, W. C. (2016). Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discovery*, 6(8):914–929.

[4] Ahlmann-Eltze, C. and Huber, W. (2023). Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, pages 1–8. Publisher: Nature Publishing Group.

[5] Allen, F., Crepaldi, L., Alsinet, C., Strong, A. J., Kleshchevnikov, V., De Angeli, P., PÃąlenÃ■kovÃą, P., Khodak, A., Kiselev, V., Kosicki, M., Bassett, A. R., Harding, H., Galanty, Y., MuÃśoz-MartÃ■nez, F., Metzakopian, E., Jackson, S. P., and Parts, L. (2019). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature Biotechnology*, 37(1):64–72. Number: 1 Publisher: Nature Publishing Group.

[6] Almaas, E. and BarabÃąsi, A.-L. (2006). Power Laws in Biological Networks. In Koonin, E. V., Wolf, Y. I., and Karev, G. P., editors, *Power Laws, Scale-Free Networks and Genome Biology*, Molecular Biology Intelligence Unit, pages 1–11. Springer US, Boston, MA.

[7] Amrhein, L., Harsha, K., and Fuchs, C. (2019). A mechanistic model for the negative binomial distribution of single-cell mRNA counts.

[8] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

[9] Anzalone, A. V., Koblan, L. W., and Liu, D. R. (2020). Genome editing with CRISPRâĂŞ-Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, 38(7):824–844. Number: 7 Publisher: Nature Publishing Group.

[10] Ballerini, V. and Liseo, B. (2022). Fisher's Noncentral Hypergeometric Distribution for Population Size Estimation. arXiv:2210.08346 [stat].

[11] BarabÃąsi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113. Number: 2 Publisher: Nature Publishing Group.

[12] Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–360.e4.

[13] Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science*, 315(5819):1709–1712. Publisher: American Association for the Advancement of Science.

[14] Baruzzo, G., Patuzzi, I., and Di Camillo, B. (2020). SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics*, 36(5):1468–1475.

[15] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[16] Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., Stanley, G., Chen, S., Garnett, M., Li, W., Moffat, J., Qi, L. S., Shapiro, R. S., Shendure, J., Weissman, J. S., and Zhuang, X. (2022). High-content CRISPR screening. *Nature Reviews Methods Primers*, 2(1):1–23. Number: 1 Publisher: Nature Publishing Group.

[17] Booeshaghi, A. S., HallgrÃ■msdÃşttir, I. B., GÃąlvez-MerchÃąn, Ã., and Pachter, L. (2022). Depth normalization for single-cell genomics count data. Pages: 2022.05.06.490859 Section: New Results.

[18] Breda, J., Zavolan, M., and van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 39(8):1008–1016. Number: 8 Publisher: Nature Publishing Group.

[19] Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1):1017. Number: 1 Publisher: Nature Publishing Group.

[20] Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V., and van der Oost, J. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, 321(5891):960–964. Publisher: American Association for the Advancement of Science.

[21] Cannoodt, R., Saelens, W., Deconinck, L., and Saeys, Y. (2021). Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*, 12(1):3942. Number: 1 Publisher: Nature Publishing Group.

[22] Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C. D., Wiegand, D. J., Ter-Ovanesyan, D., Braff, J. L., Davidsohn, N., Housden, B. E., Perrimon, N., Weiss, R., Aach, J., Collins, J. J., and Church, G. M. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, 12(4):326–328. Number: 4 Publisher: Nature Publishing Group.

[23] Chen, J. S., Ma, E., Harrington, L. B., Da Costa, M., Tian, X., Palefsky, J. M., and Doudna, J. A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science*, 360(6387):436–439. Publisher: American Association for the Advancement of Science.

[24] Chen, W., McKenna, A., Schreiber, J., Haeussler, M., Yin, Y., Agarwal, V., Noble, W. S., and Shendure, J. (2019). Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Research*, 47(15):7989–8003.

[25] Clark, I. C., Fontanez, K. M., Meltzer, R. H., Xue, Y., Hayford, C., May-Zhang, A., DâĂŹAmato, C., Osman, A., Zhang, J. Q., Hettige, P., Ishibashi, J. S. A., Delley, C. L., Weisgerber, D. W., Replogle, J. M., Jost, M., Phong, K. T., Kennedy, V. E., Peretz, C. A. C., Kim, E. A., Song, S., Karlon, W., Weissman, J. S., Smith, C. C., Gartner, Z. J., and Abate, A. R. (2023). Microfluidics-free single-cell genomics with templated emulsification. *Nature Biotechnology*, pages 1–10. Publisher: Nature Publishing Group.

[26] Consortium*, T. T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., Hemenez, M., Ponnusamy, R., Salehi, A., Sanagavarapu, B. A., Spallino, E., Aaron, K. A., Concepcion, W., Gardner, J. M., Kelly, B., Neidlinger, N., Wang, Z., Crasta, S., Kolluru, S., Morri, M., Pisco, A. O., Tan, S. Y., Travaglini, K. J., Xu, C., AlcÃąntara-HernÃąndez, M., Almanzar, N., Antony, J., Beyersdorf, B., Burhan, D., Calcuttawala, K., Carter, M. M., Chan, C. K. F., Chang, C. A., Chang, S., Colville, A., Crasta, S., Culver, R. N., CvijoviÄĞ, I., DâĂŹAmato, G., Ezran, C., Galdos, F. X., Gillich, A., Goodyer, W. R., Hang, Y., Hayashi, A., Houshdaran, S., Huang, X., Irwin, J. C., Jang, S., Juanico, J. V., Kershner, A. M., Kim, S., Kiss, B., Kolluru, S., Kong, W., Kumar, M. E., Kuo, A. H., Leylek, R., Li, B., Loeb, G. B., Lu, W.-J., Mantri, S., Markovic, M., McAlpine, P. L., Morree, A. d., Morri, M., Mrouj, K., Mukherjee, S., Muser, T., NeuhÃűfer, P., Nguyen, T. D., Perez, K., Phansalkar, R., Pisco, A. O., Puluca, N., Qi, Z., Rao, P., Raquer-McKay, H., Schaum, N., Scott, B., Seddighzadeh, B., Segal, J., Sen, S., Sikandar, S., Spencer, S. P., Steffes, L. C., Subramaniam, V. R., Swarup, A., Swift, M., Travaglini, K. J., Treuren, W. V., Trimm, E., Veizades, S., Vijayakumar, S., Vo, K. C., Vorperian, S. K., Wang, W., Weinstein, H. N. W., Winkler, J., Wu, T. T. H., Xie, J., Yung, A. R., Zhang, Y., Detweiler, A. M., Mekonen, H., Neff, N. F., Sit, R. V., Tan, M., Yan, J., Bean, G. R., Charu, V., ForgÃş, E., Martin, B. A., Ozawa, M. G., Silva, O., Tan, S. Y., Toland, A., Vemuri, V. N. P., Afik, S., Awayan, K., Botvinnik, O. B., Byrne, A., Chen, M., Dehghannasiri, R., Detweiler, A. M., Gayoso, A., Granados, A. A., Li, Q., Mahmoudabadi, G., McGeever, A., Morree, A. d., Olivieri, J. E., Park, M., Pisco, A. O., Ravikumar, N., Salzman, J., Stanley, G., Swift, M., Tan, M., Tan, W., Tarashansky, A. J., Vanheusden, R., Vorperian, S. K., Wang, P., Wang, S., Xing, G., Xu, C., Yosef, N., AlcÃąntara-HernÃąndez, M., Antony, J., Chan, C. K. F., Chang, C. A., Colville, A., Crasta, S., Culver, R., Dethlefsen, L., Ezran, C., Gillich, A., Hang, Y., Ho, P.-Y., Irwin, J. C., Jang, S., Kershner, A. M., Kong, W., Kumar, M. E.,

Kuo, A. H., Leylek, R., Liu, S., Loeb, G. B., Lu, W.-J., Maltzman, J. S., Metzger, R. J., Morree, A. d., NeuhÃűfer, P., Perez, K., Phansalkar, R., Qi, Z., Rao, P., Raquer-McKay, H., Sasagawa, K., Scott, B., Sinha, R., Song, H., Spencer, S. P., Swarup, A., Swift, M., Travaglini, K. J., Trimm, E., Veizades, S., Vijayakumar, S., Wang, B., Wang, W., Winkler, J., Xie, J., Yung, A. R., Artandi, S. E., Beachy, P. A., Clarke, M. F., Giudice, L. C., Huang, F. W., Huang, K. C., Idoyaga, J., Kim, S. K., Krasnow, M., Kuo, C. S., Nguyen, P., Quake, S. R., Rando, T. A., Red-Horse, K., Reiter, J., Relman, D. A., Sonnenburg, J. L., Wang, B., Wu, A., Wu, S. M., and Wyss-Coray, T. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*. Publisher: American Association for the Advancement of Science.

[27] Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B., PÃąl, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The Genetic Landscape of a Cell. *Science*, 327(5964):425–431.

[28] Cover, T. M. (2006). *Elements of information theory Thomas M. Cover, Joy A. Thomas.* Wiley-Interscience, Hoboken, N.J., 2nd ed. edition. Publication Title: Elements of information theory.

[29] Cuella-Martin, R., Hayward, S. B., Fan, X., Chen, X., Huang, J.-W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C., Shen, Y., and Ciccia, A. (2021). Functional interrogation of DNA damage response variants with base editing screens. *Cell*, 184(4):1081–1097.e19.

[30] Cundill, B. and Alexander, N. D. (2015). Sample size calculations for skewed distributions. *BMC Medical Research Methodology*, 15(1):28.

[31] Dai, M., Yan, G., Wang, N., Daliah, G., Edick, A. M., Poulet, S., Boudreault, J., Ali, S., Burgos, S. A., and Lebrun, J.-J. (2021). In vivo genome-wide CRISPR screen reveals breast cancer vulnerabilities and synergistic mTOR/Hippo targeted combination therapy. *Nature Communications*, 12(1):3055. Number: 1 Publisher: Nature Publishing Group.

[32] Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301. Number: 3 Publisher: Nature Publishing Group.

[33] Dekking, M. (2005). *A modern introduction to probability and statistics : understanding why and how / F.M. Dekking [and others].* Springer texts in statistics. Springer, London. Publication Title: A modern introduction to probability and statistics : understanding why and how.

[34] DeWeirdt, P. C., McGee, A. V., Zheng, F., Nwolah, I., Hegde, M., and Doench, J. G. (2022). Accounting for small variations in the tracrRNA sequence improves sgRNA activity predictions for CRISPR screening. *Nature Communications*, 13(1):5255.

[35] Ding, J., Adiconis, X., Simmons, S. K., Kowalczyk, M. S., Hession, C. C., Marjanovic, N. D., Hughes, T. K., Wadsworth, M. H., Burks, T., Nguyen, L. T., Kwon, J. Y. H., Barak, B., Ge, W., Kedaigle, A. J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A. K., Villani, A.-C., Regev, A., and Levin, J. Z. (2020). Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746. Number: 6 Publisher: Nature Publishing Group.

[36] Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17. Publisher: Elsevier.

[37] Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., and Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2):184–191. Number: 2 Publisher: Nature Publishing Group.

[38] D'Souza, R. M., Borgs, C., Chayes, J. T., Berger, N., and Kleinberg, R. D. (2007). Emergence of tempered preferential attachment from optimization. *Proceedings of the National Academy of Sciences*, 104(15):6112–6117. Publisher: Proceedings of the National Academy of Sciences.

[39] Dvir, E., Shohat, S., and Shifman, S. (2022). Genetic mechanisms for tissue-specific essential genes. Pages: 2021.04.09.438977 Section: New Results.

[40] Emanuel, G., Moffitt, J. R., and Zhuang, X. (2017). High-throughput, image-based screening of pooled genetic-variant libraries. *Nature Methods*, 14(12):1159–1162. Number: 12 Publisher: Nature Publishing Group.

[41] Eslami-Mossallam, B., Klein, M., Smagt, C. V. D., Sanden, K. V. D., Jones, S. K., Hawkins, J. A., Finkelstein, I. J., and Depken, M. (2022). A kinetic model predicts SpCas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. *Nature Communications*, 13(1):1367. Number: 1 Publisher: Nature Publishing Group.

[42] Fang, L., Li, Y., Ma, L., Xu, Q., Tan, F., and Chen, G. (2021). GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Research*, 49(D1):D97–D103.

[43] Fei, L., Chen, H., Ma, L., E, W., Wang, R., Fang, X., Zhou, Z., Sun, H., Wang, J., Jiang, M., Wang, X., Yu, C., Mei, Y., Jia, D., Zhang, T., Han, X., and Guo, G. (2022). Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development. *Nature Genetics*, 54(7):1051–1061. Number: 7 Publisher: Nature Publishing Group.

[44] Feldman, D., Singh, A., Schmid-Burgk, J. L., Carlson, R. J., Mezger, A., Garrity, A. J., Zhang, F., and Blainey, P. C. (2019). Optical Pooled Screens in Human Cells. *Cell*, 179(3):787–799.e17.

[45] Fog, A. (2008). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. *Communications in Statistics - Simulation and Computation*, 37(2):241–257. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03610910701790236.

[46] Fog, A. (2023). Biased Urn Theory.

[47] Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M. S., Zhao, M., Ager, C. R., Rogava, M., Hovey, L., Rotem, A., Bernatchez, C., Wucherpfennig, K. W., Johnson, B. E., Rozenblatt-Rosen, O., Schadendorf, D., Regev, A., and Izar, B. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3):332–341. Number: 3 Publisher: Nature Publishing Group.

[48] Frangoul, H., Altshuler, D., Cappellini, M. D., Chen, Y.-S., Domm, J., Eustace, B. K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., Ho, T. W., Kattamis, A., Kernytsky, A., Lekstrom-Himes, J., Li, A. M., Locatelli, F., Mapara, M. Y., de Montalembert, M., Rondelli, D., Sharma, A., Sheth, S., Soni, S., Steinberg, M. H., Wall, D., Yen, A., and Corbacioglu, S. (2021). CRISPR-Cas9 Gene Editing for Sickle Cell Disease and Îš-Thalassemia. *New England Journal of Medicine*, 384(3):252–260. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa2031054.

[49] Gaj, T., Gersbach, C. A., and Barbas, C. F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, 31(7):397–405.

[50] Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1):377–390.e19. Publisher: Elsevier.

[51] Gillmore, J. D., Gane, E., Taubel, J., Kao, J., Fontana, M., Maitland, M. L., Seitzer, J., OâĂŹConnell, D., Walsh, K. R., Wood, K., Phillips, J., Xu, Y., Amaral, A., Boyd, A. P., Cehelsky, J. E., McKee, M. D., Schiermeier, A., Harari, O., Murphy, A., Kyratsous, C. A., Zambrowicz, B., Soltys, R., Gutstein, D. E., Leonard, J., Sepp-Lorenzino, L., and Lebwohl, D. (2021). CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *New England Journal of Medicine*, 385(6):493–502. Publisher: Massachusetts Medical Society.

[52] GonÃğalves, E., Segura-Cabrera, A., Pacini, C., Picco, G., Behan, F. M., Jaaks, P., Coker, E. A., van der Meer, D., Barthorpe, A., Lightfoot, H., Mironenko, T., Beck, A., Richardson, L., Yang, W., Lleshi, E., Hall, J., Tolley, C., Hall, C., Mali, I., Thomas, F., Morris, J., Leach, A. R., Lynch, J. T., Sidders, B., Crafter, C., Iorio, F., Fawell, S., and Garnett, M. J. (2020). Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. *Molecular Systems Biology*, 16(7):e9405. Publisher: John Wiley & Sons, Ltd.

[53] Gootenberg, J. S., Abudayyeh, O. O., Kellner, M. J., Joung, J., Collins, J. J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science*, 360(6387):439–444. Publisher: American Association for the Advancement of Science.

[54] Haeussler, M., SchÃűnig, K., Eckert, H., Eschstruth, A., MiannÃl', J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., Joly, J.-S., and Concordet, J.-P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, 17(1):148.

[55] Hafemeister, C. and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296.

[56] Hagemann-Jensen, M., Ziegenhain, C., and Sandberg, R. (2022). Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nature Biotechnology*, 40(10):1452–1457. Number: 10 Publisher: Nature Publishing Group.

[57] Hanna, R. E., Hegde, M., Fagre, C. R., DeWeirdt, P. C., Sangree, A. K., Szegletes, Z., Griffith, A., Feeley, M. N., Sanson, K. R., Baidi, Y., Koblan, L. W., Liu, D. R., Neal, J. T., and Doench, J. G. (2021). Massively parallel assessment of human variants with base editor screens. *Cell*, 184(4):1064–1080.e20.

[58] Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., and Moffat, J. (2015). High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6):1515–1526.

[59] Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L., and Salomonis, N. (2018). The Human Cell Atlas bone marrow single-cell interactive web portal. *Experimental Hematology*, 68:51–61. Publisher: Elsevier.

[60] Hill, A. J., McFaline-Figueroa, J. L., Starita, L. M., Gasperini, M. J., Matreyek, K. A., Packer, J., Jackson, D., Shendure, J., and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. *Nature Methods*, 15(4):271–274. Number: 4 Publisher: Nature Publishing Group.

[61] Hoberecht, L., Perampalam, P., Lun, A., and Fortin, J.-P. (2022). A comprehensive Bioconductor ecosystem for the design of CRISPR guide RNAs across nucleases and technologies. *Nature Communications*, 13(1):6568. Number: 1 Publisher: Nature Publishing Group.

[62] Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M., and Weissman, J. S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *eLife*, 5:e19760.

[63] Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of Bacteriology*, 169(12):5429–5433. Publisher: American Society for Microbiology.

[64] Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., LÃűnnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166. Number: 2 Publisher: Nature Publishing Group.

[65] Jessa, S., Blanchet-Cohen, A., Krug, B., Vladoiu, M., Coutelier, M., Faury, D., Poreau, B., De Jay, N., HÃl'bert, S., Monlong, J., Farmer, W. T., Donovan, L. K., Hu, Y., McConechy, M. K., Cavalli, F. M. G., Mikael, L. G., Ellezam, B., Richer, M., Allaire, A., Weil, A. G., Atkinson, J., Farmer, J.-P., Dudley, R. W. R., Larouche, V., Crevier, L., Albrecht, S., Filbin, M. G., Sartelet, H., Lutz, P.-E., Nagy, C., Turecki, G., Costantino, S., Dirks, P. B., Murai, K. K., Bourque, G., Ragoussis, J., Garzia, L., Taylor, M. D., Jabado, N., and Kleinman, C. L. (2019). Stalled developmental programs at the root of pediatric brain tumors. *Nature Genetics*, 51(12):1702–1713. Number: 12 Publisher: Nature Publishing Group.

[66] Jin, X., Simmons, S. K., Guo, A., Shetty, A. S., Ko, M., Nguyen, L., Jokhi, V., Robinson, E., Oyler, P., Curry, N., Deangeli, G., Lodato, S., Levin, J. Z., Regev, A., Zhang, F., and Arlotta, P. (2020). In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes.

[67] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A Programmable Dual-RNAâĂŞGuided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*. Publisher: American Association for the Advancement of Science.

[68] Jost, M., Santos, D. A., Saunders, R. A., Horlbeck, M. A., Hawkins, J. S., Scaria, S. M., Norman, T. M., Hussmann, J. A., Liem, C. R., Gross, C. A., and Weissman, J. S. (2020). Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nature Biotechnology*, 38(3):355–364. Number: 3 Publisher: Nature Publishing Group.

[69] Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751. Number: 7949 Publisher: Nature Publishing Group.

[70] Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., and Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, 52(9):1419–1427. Number: 9 Publisher: Nature Publishing Group.

[71] Kebschull, J. M. and Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research*, 43(21):e143.

[72] Kim, D., Kim, D.-e., Lee, G., Cho, S.-I., and Kim, J.-S. (2019). Genome-wide target specificity of CRISPR RNA-guided adenine base editors. *Nature Biotechnology*, 37(4):430–435. Number: 4 Publisher: Nature Publishing Group.

[73] Kim, E. and Hart, T. (2021). Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. *Genome Medicine*, 13(1):2.

[74] Knott, G. J. and Doudna, J. A. (2018). CRISPR-Cas guides the future of genetic engineering. *Science*, 361(6405):866–869. Publisher: American Association for the Advancement of Science.

[75] Konstantakos, V., Nentidis, A., Krithara, A., and Paliouras, G. (2022). CRISPRâĂŞCas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Research*, 50(7):3616–3637.

[76] Kotliarov, Y., Sparks, R., Martins, A. J., MulÃÍ, M. P., Lu, Y., Goswami, M., Kardava, L., Banchereau, R., Pascual, V., Biancotto, A., Chen, J., Schwartzberg, P. L., Bansal, N., Liu, C. C., Cheung, F., Moir, S., and Tsang, J. S. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, 26(4):618–629. Number: 4 Publisher: Nature Publishing Group.

[77] L. Lun, A. T., Bach, K., and Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1):75.

[78] Larsson, A. J. M., Johnsson, P., Hagemann-Jensen, M., Hartmanis, L., Faridani, O. R., Reinius, B., Segerstolpe, Ã., Rivera, C. M., Ren, B., and Sandberg, R. (2019). Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254. Number: 7738 Publisher: Nature Publishing Group.

[79] Lawson, M. J., Camsund, D., Larsson, J., Baltekin, Ã., Fange, D., and Elf, J. (2017). In situ genotyping of a pooled strain library after characterizing complex phenotypes. *Molecular Systems Biology*, 13(10):947. Publisher: John Wiley & Sons, Ltd.

[80] Lazar, N. H., Celik, S., Chen, L., Fay, M., Irish, J. C., Jensen, J., Tillinghast, C. A., Urbanik, J., Bone, W. P., Roberts, G. H. L., Gibson, C. C., and Haque, I. S. (2023). High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by CRISPR-Cas9 editing. Pages: 2023.04.15.537038 Section: New Results.

[81] Levin, B. (1984). Simple Improvements on Cornfield's Approximation to the Mean of a Noncentral Hypergeometric Random Variable. *Biometrika*, 71(3):630–632. Publisher: [Oxford University Press, Biometrika Trust].

[82] Li, H., Janssens, J., De Waegeneer, M., Kolluru, S. S., Davie, K., Gardeux, V., Saelens, W., David, F. P. A., BrbiÄĞ, M., Spanier, K., Leskovec, J., McLaughlin, C. N., Xie, Q., Jones, R. C., Brueckner, K., Shim, J., Tattikota, S. G., Schnorrer, F., Rust, K., Nystul, T. G., Carvalho-Santos, Z., Ribeiro, C., Pal, S., Mahadevaraju, S., Przytycka, T. M., Allen, A. M., Goodwin, S. F., Berry, C. W., Fuller, M. T., White-Cooper, H., Matunis, E. L., DiNardo, S., Galenza, A., OâĂŹBrien, L. E., Dow, J. A. T., FCA ConsortiumÂğ, Jasper, H., Oliver, B., Perrimon, N., Deplancke, B., Quake, S. R., Luo, L., Aerts, S., Agarwal, D., Ahmed-Braimah, Y., Arbeitman, M., Ariss, M. M., Augsburger, J., Ayush, K., Baker, C. C., Banisch, T., Birker, K., Bodmer, R., Bolival, B., Brantley, S. E., Brill, J. A., Brown, N. C., Buehner, N. A., Cai, X. T., Cardoso-Figueiredo, R., Casares, F., Chang, A., Clandinin, T. R., Crasta, S., Desplan, C., Detweiler, A. M., Dhakan, D. B., DonÃă, E., Engert, S., FlocâĂŹhlay, S., George, N., GonzÃąlez-Segarra, A. J., Groves, A. K., Gumbin, S., Guo, Y., Harris, D. E., Heifetz, Y., Holtz, S. L., Horns, F., Hudry, B., Hung, R.-J., Jan, Y. N., Jaszczak, J. S., Jefferis, G. S. X. E., Karkanias, J., Karr, T. L., Katheder, N. S., Kezos, J., Kim, A. A., Kim, S. K., Kockel, L., Konstantinides, N., Kornberg, T. B., Krause, H. M., Labott, A. T., Laturney, M., Lehmann, R., Leinwand, S., Li, J., Li, J. S. S., Li, K., Li, K., Li, L., Li, T., Litovchenko, M., Liu, H.-H., Liu, Y., Lu, T.-C., Manning, J., Mase, A., Matera-Vatnick, M., Matias, N. R., McDonough-Goldstein, C. E., McGeever, A., McLachlan, A. D., Moreno-Roman, P., Neff, N., Neville, M., Ngo, S., Nielsen, T., OâĂŹBrien, C. E., Osumi-Sutherland, D., ÃŰzel, M. N., Papatheodorou, I., Petkovic, M., Pilgrim, C., Pisco, A. O., Reisenman, C., Sanders, E. N., Dos Santos,

G., Scott, K., Sherlekar, A., Shiu, P., Sims, D., Sit, R. V., Slaidina, M., Smith, H. E., Sterne, G., Su, Y.-H., Sutton, D., Tamayo, M., Tan, M., Tastekin, I., Treiber, C., Vacek, D., Vogler, G., Waddell, S., Wang, W., Wilson, R. I., Wolfner, M. F., Wong, Y.-C. E., Xie, A., Xu, J., Yamamoto, S., Yan, J., Yao, Z., Yoda, K., Zhu, R., and Zinzen, R. P. (2022a). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science*, 375(6584):eabk2432.

[83] Li, H., Zhang, Z., Squires, M., Chen, X., and Zhang, X. (2023). scMultiSim: simulation of multi-modality single cell data guided by cell-cell interactions and gene regulatory networks. Pages: 2022.10.15.512320 Section: New Results.

[84] Li, J., Mahata, B., Escobar, M., Goell, J., Wang, K., Khemka, P., and Hilton, I. B. (2021). Programmable human histone phosphorylation and gene activation using a CRISPR/Cas9-based chromatin kinase. *Nature Communications*, 12(1):896. Number: 1 Publisher: Nature Publishing Group.

[85] Li, R., Klingbeil, O., Monducci, D., Young, M. J., Rodriguez, D. J., Bayyat, Z., Dempster, J. M., Kesar, D., Yang, X., Zamanighomi, M., Vakoc, C. R., Ito, T., and Sellers, W. R. (2022b). Comparative optimization of combinatorial CRISPR screens. *Nature Communications*, 13(1):2469. Number: 1 Publisher: Nature Publishing Group.

[86] Liao, J. (1992). An Algorithm for the Mean and Variance of the Noncentral Hypergeometric Distribution. *Biometrics*, 48(3):889–892. Publisher: [Wiley, International Biometric Society].

[87] Lindeboom, R. G. H., Regev, A., and Teichmann, S. A. (2021). Towards a Human Cell Atlas: Taking Notes from the Past. *Trends in Genetics*, 37(7):625–630. Publisher: Elsevier.

[88] Listgarten, J., Weinstein, M., Kleinstiver, B. P., Sousa, A. A., Joung, J. K., Crawford, J., Gao, K., Hoang, L., Elibol, M., Doench, J. G., and Fusi, N. (2018). Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nature Biomedical Engineering*, 2(1):38–47. Number: 1 Publisher: Nature Publishing Group.

[89] Lopez-Obando, M., Hoffmann, B., GÃl'ry, C., Guyon-Debast, A., TÃl'oulÃl', E., Rameau, C., Bonhomme, S., and NoguÃl', F. (2016). Simple and Efficient Targeting of Multiple Genes Through CRISPR-Cas9 in Physcomitrella patens. *G3 Genes|Genomes|Genetics*, 6(11):3647–3653.

[90] Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

[91] Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746. Publisher: John Wiley & Sons, Ltd.

[92] Luo, S., Wang, Z., Zhang, Z., Zhou, T., and Zhang, J. (2023a). Genome-wide inference reveals that feedback regulations constrain promoter-dependent transcriptional burst kinetics. *Nucleic Acids Research*, 51(1):68–83.

[93] Luo, S., Zhang, Z., Wang, Z., Yang, X., Chen, X., Zhou, T., and Zhang, J. (2023b). Inferring transcriptional bursting kinetics from single-cell snapshot data using a generalized telegraph model. *Royal Society Open Science*, 10(4):221057.

[94] Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214.

[95] Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. *Journal of Computational Biology*, 16(2):229–239. Publisher: Mary Ann Liebert, Inc., publishers.

[96] McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Wills, Q. F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33(8):1179–1186.

[97] McFaline-Figueroa, J. L., Hill, A. J., Qiu, X., Jackson, D., Shendure, J., and Trapnell, C. (2019). A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nature Genetics*, 51(9):1389–1398. Number: 9 Publisher: Nature Publishing Group.

[98] McKenna, A. and Shendure, J. (2018). FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biology*, 16(1):74.

[99] Meyers, R. M., Bryan, J. G., McFarland, J. M., Weir, B. A., Sizemore, A. E., Xu, H., Dharia, N. V., Montgomery, P. G., Cowley, G. S., Pantel, S., Goodale, A., Lee, Y., Ali, L. D., Jiang, G., Lubonja, R., Harrington, W. F., Strickland, M., Wu, T., Hawes, D. C., Zhivich, V. A., Wyatt, M. R., Kalani, Z., Chang, J. J., Okamoto, M., Stegmaier, K., Golub, T. R., Boehm, J. S., Vazquez, F., Root, D. E., Hahn, W. C., and Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPRâĂŞCas9 essentiality screens in cancer cells. *Nature Genetics*, 49(12):1779–1784. Number: 12 Publisher: Nature Publishing Group.

[100] Michlits, G., Jude, J., Hinterndorfer, M., de Almeida, M., Vainorius, G., Hubmann, M., Neumann, T., Schleiffer, A., Burkard, T. R., Fellner, M., Gijsbertsen, M., Traunbauer, A., Zuber, J., and Elling, U. (2020). Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nature Methods*, 17(7):708–716. Number: 7 Publisher: Nature Publishing Group.

[101] Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., Satija, R., Sanjana, N. E., Koralov, S. B., and Smibert, P. (2019). Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5):409–412. Number: 5 Publisher: Nature Publishing Group.

[102] Mojica, F. J., DÃ■ez-VillaseÃśor, C., GarcÃ■a-MartÃ■nez, J., and Soria, E. (2005). Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 60(2):174–182.

[103] Naert, T., Tulkens, D., Edwards, N. A., Carron, M., Shaidani, N.-I., Wlizla, M., Boel, A., Demuynck, S., Horb, M. E., Coucke, P., Willaert, A., Zorn, A. M., and Vleminckx, K. (2020). Maximizing CRISPR/Cas9 phenotype penetrance applying predictive modeling of editing outcomes in Xenopus and zebrafish embryos. *Scientific Reports*, 10(1):14662. Number: 1 Publisher: Nature Publishing Group.

[104] Nassar, L. R., Barber, G. P., Benet-PagÃĺs, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A., Lee, B., Lee, C., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B., Schmelter, D., Speir, M., Wick, B., Zweig, A., Haussler, D., Kuhn, R., Haeussler, M., and Kent, W. (2023). The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research*, 51(D1):D1188–D1195.

[105] NuÃśez, J. K., Chen, J., Pommier, G. C., Cogan, J. Z., Replogle, J. M., Adriaens, C., Ramadoss, G. N., Shi, Q., Hung, K. L., Samelson, A. J., Pogson, A. N., Kim, J. Y. S., Chung, A., Leonetti, M. D., Chang, H. Y., Kampmann, M., Bernstein, B. E., Hovestadt, V., Gilbert, L. A., and Weissman, J. S. (2021). Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, 184(9):2503–2519.e17.

[106] Okada, M., Kanamori, M., Someya, K., Nakatsukasa, H., and Yoshimura, A. (2017). Stabilization of Foxp3 expression by CRISPR-dCas9-based epigenome editing in mouse primary T cells. *Epigenetics & Chromatin*, 10(1):24.

[107] OâĂŹGeen, H., Ren, C., Nicolet, C. M., Perez, A. A., Halmai, J., Le, V. M., Mackay, J. P., Farnham, P. J., and Segal, D. J. (2017). dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression. *Nucleic Acids Research*, 45(17):9901–9916.

[108] Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., and Satija, R. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature Genetics*, 53(3):322–331. Number: 3 Publisher: Nature Publishing Group.

[109] Parrish, P. C. R. and Berger, A. H. (2021). CRISPR base editor screens identify variant function at scale. *Molecular Cell*, 81(4):647–648.

[110] Peng, X. and Dorman, K. S. (2023). Accurate estimation of molecular counts from amplicon sequence data with unique molecular identifiers. *Bioinformatics*, 39(1):btad002.

[111] Pierce, S. E., Granja, J. M., and Greenleaf, W. J. (2021). High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nature Communications*, 12(1):2969. Number: 1 Publisher: Nature Publishing Group.

[112] Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1971.10482356.

[113] Rao, S., Yao, Y., and Bauer, D. E. (2021). Editing GWAS: experimental approaches to dissect and exploit disease-associated genetic variation. *Genome Medicine*, 13(1):41.

[114] Replogle, J. M., Norman, T. M., Xu, A., Hussmann, J. A., Chen, J., Cogan, J. Z., Meer, E. J., Terry, J. M., Riordan, D. P., Srinivas, N., Fiddes, I. T., Arthur, J. G., Alvarado, L. J., Pfeiffer, K. A., Mikkelsen, T. S., Weissman, J. S., and Adamson, B. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology*, 38(8):954–961. Number: 8 Publisher: Nature Publishing Group.

[115] Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., and Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28. Publisher: Elsevier.

[116] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[117] Rubin, A. J., Parker, K. R., Satpathy, A. T., Qi, Y., Wu, B., Ong, A. J., Mumbach, M. R., Ji, A. L., Kim, D. S., Cho, S. W., Zarnegar, B. J., Greenleaf, W. J., Chang, H. Y., and Khavari, P. A. (2019). Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*, 176(1):361–376.e17.

[118] Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Vaimberg, E. W., Goodale, A., Root, D. E., Piccioni, F., and Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nature Communications*, 9(1):5416. Number: 1 Publisher: Nature Publishing Group.

[119] Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270.

[120] Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., Chen, S., Green, F., Jones, R. C., Maynard, A., Penland, L., Pisco, A. O., Sit, R. V., Stanley, G. M., Webber, J. T., Zanini, F., Baghel, A. S., Bakerman, I., Bansal, I., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chen, M. B., Chen, S., Cho, M., Cirolia, G., Conley, S. D., Darmanis, S., Demers, A., Demir, K., de Morree, A., Divita, T., du Bois, H., Dulgeroff, L. B. T., Ebadi, H., Espinoza, F. H., Fish, M., Gan, Q., George, B. M., Gillich, A., Green, F., Genetiano, G., Gu, X., Gulati, G. S., Hang, Y., Hosseinzadeh, S., Huang, A., Iram, T., Isobe, T., Ives, F., Jones, R. C., Kao, K. S., Karnam, G., Kershner, A. M., Kiss, B. M., Kong, W., Kumar, M. E., Lam, J. Y., Lee, D. P., Lee, S. E., Li, G., Li, Q., Liu, L., Lo, A., Lu, W.-J., Manjunath, A., May, A. P., May, K. L., May, O. L., Maynard, A., McKay, M., Metzger, R. J., Mignardi, M., Min, D., Nabhan, A. N., Neff, N. F., Ng, K. M., Noh, J., Patkar, R., Peng, W. C., Penland, L., Puccinelli, R., Rulifson, E. J., Schaum, N., Sikandar, S. S., Sinha, R., Sit, R. V., Szade, K., Tan, W., Tato, C., Tellez, K., Travaglini, K. J., Tropini, C., Waldburger, L., van Weele, L. J., Wosczyna, M. N., Xiang, J., Xue, S., Youngyunpipatkul, J., Zanini, F., Zardeneta, M. E., Zhang, F., Zhou, L., Bansal, I., Chen, S., Cho, M., Cirolia, G., Darmanis, S., Demers, A., Divita, T., Ebadi, H., Genetiano, G., Green, F., Hosseinzadeh, S., Ives, F., Lo, A., May, A. P., Maynard, A., McKay, M., Neff, N. F., Penland, L., Sit, R. V., Tan, W., Waldburger, L., Youngyunpipatkul, J., Batson, J., Botvinnik, O., Castro, P., Croote, D., Darmanis, S., DeRisi, J. L., Karkanias, J., Pisco, A. O., Stanley, G. M., Webber, J. T.,

Zanini, F., Baghel, A. S., Bakerman, I., Batson, J., Bilen, B., Botvinnik, O., Brownfield, D., Chen, M. B., Darmanis, S., Demir, K., de Morree, A., Ebadi, H., Espinoza, F. H., Fish, M., Gan, Q., George, B. M., Gillich, A., Gu, X., Gulati, G. S., Hang, Y., Huang, A., Iram, T., Isobe, T., Karnam, G., Kershner, A. M., Kiss, B. M., Kong, W., Kuo, C. S., Lam, J. Y., Lehallier, B., Li, G., Li, Q., Liu, L., Lu, W.-J., Min, D., Nabhan, A. N., Ng, K. M., Nguyen, P. K., Patkar, R., Peng, W. C., Penland, L., Rulifson, E. J., Schaum, N., Sikandar, S. S., Sinha, R., Szade, K., Tan, S. Y., Tellez, K., Travaglini, K. J., Tropini, C., van Weele, L. J., Wang, B. M., Wosczyna, M. N., Xiang, J., Yousef, H., Zhou, L., Batson, J., Botvinnik, O., Chen, S., Darmanis, S., Green, F., May, A. P., Maynard, A., Pisco, A. O., Quake, S. R., Schaum, N., Stanley, G. M., Webber, J. T., Wyss-Coray, T., Zanini, F., Beachy, P. A., Chan, C. K. F., de Morree, A., George, B. M., Gulati, G. S., Hang, Y., Huang, K. C., Iram, T., Isobe, T., Kershner, A. M., Kiss, B. M., Kong, W., Li, G., Li, Q., Liu, L., Lu, W.-J., Nabhan, A. N., Ng, K. M., Nguyen, P. K., Peng, W. C., Rulifson, E. J., Schaum, N., Sikandar, S. S., Sinha, R., Szade, K., Travaglini, K. J., Tropini, C., Wang, B. M., Weinberg, K., Wosczyna, M. N., Wu, S. M., Yousef, H., Barres, B. A., Beachy, P. A., Chan, C. K. F., Clarke, M. F., Darmanis, S., Huang, K. C., Karkanias, J., Kim, S. K., Krasnow, M. A., Kumar, M. E., Kuo, C. S., May, A. P., Metzger, R. J., Neff, N. F., Nusse, R., Nguyen, P. K., Rando, T. A., Sonnenburg, J., Wang, B. M., Weinberg, K., Weissman, I. L., Wu, S. M., Quake, S. R., Wyss-Coray, T., The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372. Number: 7727 Publisher: Nature Publishing Group.

[121] Schmid, K. T., HÃűllbacher, B., Cruceanu, C., BÃűttcher, A., Lickert, H., Binder, E. B., Theis, F. J., and Heinig, M. (2021). scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nature Communications*, 12(1):6625. Number: 1 Publisher: Nature Publishing Group.

[122] Scholefield, J. and Harrison, P. T. (2021). Prime editing âĂŞ an update on the field. *Gene Therapy*, 28(7):396–401. Number: 7 Publisher: Nature Publishing Group.

[123] Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., and Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, 17(6):629–635. Number: 6 Publisher: Nature Publishing Group.

[124] Schwaber, J., Andersen, S., and Nielsen, L. (2019). Shedding light: The importance of reverse transcription efficiency standards in data interpretation. *Biomolecular Detection and Quantification*, 17:100077.

[125] Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., and Zhang, F. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*, 343(6166):84–87.

[126] Shen, M. W., Arbab, M., Hsu, J. Y., Worstell, D., Culbertson, S. J., Krabbe, O., Cassa, C. A., Liu, D. R., Gifford, D. K., and Sherwood, R. I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563(7733):646–651. Number: 7733 Publisher: Nature Publishing Group.

[127] Srivatsa, S., Kuipers, J., Schmich, F., Eicher, S., Emmenlauer, M., Dehio, C., and Beerenwinkel, N. (2018). Improved pathway reconstruction from RNA interference screens by exploiting off-target effects. *Bioinformatics*, 34(13):i519–i527.

[128] Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. *Nature Methods*, 18(11):1333–1341. Number: 11 Publisher: Nature Publishing Group.

[129] Tang, W., JÃÿrgensen, A. C. S., Marguerat, S., Thomas, P., and Shahrezaei, V. (2023). Modelling capture efficiency of single cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. Pages: 2023.03.06.531327 Section: New Results.

[130] Townes, F. W., Hicks, S. C., Aryee, M. J., and Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):295.

[131] Tsai, S. Q., Nguyen, N. T., Malagon-Lopez, J., Topkar, V. V., Aryee, M. J., and Joung, J. K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPRâĂŞCas9 nuclease off-targets. *Nature Methods*, 14(6):607–614. Number: 6 Publisher: Nature Publishing Group.

[132] Tsai, S. Q., Zheng, Z., Nguyen, N. T., Liebers, M., Topkar, V. V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A. J., Le, L. P., Aryee, M. J., and Joung, J. K. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, 33(2):187–197. Number: 2 Publisher: Nature Publishing Group.

[133] Wang, J. Y. and Doudna, J. A. (2023). CRISPR technology: A decade of genome editing is only the beginning. *Science*, 379(6629):eadd8643. Publisher: American Association for the Advancement of Science.

[134] Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101. Publisher: American Association for the Advancement of Science.

[135] Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, 343(6166):80–84.

[136] wenhao, t. (2019). bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data | Bioinformatics | Oxford Academic.

[137] Wu, H., Kirita, Y., Donnelly, E. L., and Humphreys, B. D. (2019). Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *Journal of the American Society of Nephrology*, 30(1):23.

[138] Ye, C., Speed, T. P., and Salim, A. (2019). DECENT: differential expression with capture efficiency adjustmeNT for single-cell RNA-seq data. *Bioinformatics*, 35(24):5155–5162.

[139] Zappia, L., Phipson, B., and Oshlack, A. (2017). Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174.

[140] Zeisel, A., Hochgerner, H., LÃűnnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., HÃd'ring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., and Linnarsson, S. (2018). Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4):999–1014.e22.

[141] Zhao, J., Zhang, S., Liu, Y., He, X., Qu, M., Xu, G., Wang, H., Huang, M., Pan, J., Liu, Z., Li, Z., Liu, L., and Zhang, Z. (2020). Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discovery*, 6(1):1–19. Number: 1 Publisher: Nature Publishing Group.

[142] Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049. Number: 1 Publisher: Nature Publishing Group.

[143] Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, 509(7501):487–491. Number: 7501 Publisher: Nature Publishing Group.

[144] Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643.e4.

[145] Zilionis, R., Engblom, C., Pfirschke, C., Savova, V., Zemmour, D., Saatcioglu, H. D., Krishnan, I., Maroni, G., Meyerovitz, C. V., Kerwin, C. M., Choi, S., Richards, W. G., De Rienzo, A., Tenen, D. G., Bueno, R., Levantini, E., Pittet, M. J., and Klein, A. M. (2019). Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity*, 50(5):1317–1334.e10.

[146] Zou, R. S., Liu, Y., Gaido, O. E. R., Konig, M. F., Mog, B. J., Shen, L. L., Aviles-Vazquez, F., Marin-Gonzalez, A., and Ha, T. (2023). Improving the sensitivity of in vivo CRISPR off-target detection with DISCOVER-Seq+. *Nature Methods*, pages 1–8. Publisher: Nature Publishing Group.