UNIVERSITY OF CALIFORNIA SAN DIEGO

Discovery of novel functional sequences through the analysis of tiling CRISPR screens

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Patrick Christian Fiaux

Committee in charge:

Professor Graham McVicker, Chair
Professor Bing Ren, Co-Chair
Professor Chris Benner
Professor Kelly A. Frazer
Professor Nathan Lewis

2021

The Dissertation of Patrick Christian Fiaux is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

I dedicate this thesis to my parents, Lise Fiaux-Kruse and Michel Fiaux, for their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor Graham McVicker for his incredible mentorship and support. Working in his lab was an immense pleasure and it is due to his patience and guidance that I accomplished things I never thought I could. I will always think fondly of my time as a PhD student in his lab and am extremely grateful everything he has done for me.

I would like to thank my dissertation committee for all their support and guidance. Co-chair Bing Ren has been a fantastic collaborator and allowed me to significantly expand the scope of my projects. Chris Benner has provided me with insights into methods development and always animated me to think beyond my immediate area of research. Kelly Frazer had made a lasting impact on me during my interview process, was an important factor for me deciding to join UCSD and helped me think critically about the scope and the message of a research project. Nathan Lewis has always provided me with insightful questions and comments about my research, even before I joined Graham's lab and continued to do so throughout my projects.

I would like to thank all the members of the McVicker lab for all the helpful and fun discussions. Specifically, Hsiuyi Chen for all her support with my project and her patience with my ever-changing results for her *GATA3* screen; Arko Sen for his help whenever I had a question related to R or data analysis; Ishika Luthra and Jessica Zhou for all their helpful discussions about CRISPR screens, making sure my explanations made sense and for bringing life into the lab; Jessica Zhou specifically for reading my entire thesis and making sure my writing was coherent; and all the other members for making work in the McVicker lab so fun and rewarding. I would also like to thank my undergraduate mentors; Andy Clark for introducing me to computational biology and being so patient and kind; Jason Mezey for his class on genome-wide association

studies which sparked my interest for genetics and for supporting me with my work on my undergraduate thesis.

I would like to thank my parents for their unconditional love and support, for always making sure that I could focus on my academic career, for ensuring that each time I went home it was a vacation, for always encouraging me to follow my interest and for instilling a work ethic in me which allowed me to face tough challenges and still succeed.

I would like to thank my girlfriend and partner Jessica Zhou for her compassion and patience, allowing me to grow as a person in ways I never thought possible. I am incredibly grateful for her support and care which made the difficult times easier and the fun times even better. Her sense for adventure and exploration lead us to some amazing places and I look forward to see where we end up next.

I would like to thank all the colleagues and friends I gained at UCSD for their support and friendship, be that academically, advice for jobs or to just to go climbing.

Finally, I would like to thank my friends in Switzerland, which are like a second family to me. Specifically, my godfather Werner Schnorf as well as Evelyn, Sophie and Anna, who always welcomed me and made me feel at home wherever they were. Dominic Fegerl, Tim Leeman and their families for making sure that Switzerland will always be a home base for me.

The introduction is an expansion on the background as it's set in "Discovering functional sequences with RELICS, an analysis method for CRISPR screens" in *PLOS Computational Biology* by Fiaux et al. The dissertation author was a primary investigator and author of this paper.

Chapter 1 is in part taken from "Discovering functional sequences with RELICS, an analysis method for CRISPR screens" in *PLOS Computational Biology* by Fiaux et al. and from "Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+", manuscript in preparation. The dissertation author was a primary investigator and author of these papers.

Chapter 2 is in part taken from “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology* by Fiaux et al. and from “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+”, manuscript in preparation. The dissertation author was a primary investigator and author of these papers.

Chapter 3 is in part taken from “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology* by Fiaux et al., from “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+”, manuscript in preparation and “Discovery of novel, unmarked *GATA3* functional sequences”, manuscript in preparation. The dissertation author was a primary investigator and author of these papers.

VITA

2015	Cornell University *Bachelor of Arts, Biology – Computational Biology Concentration*
2021	University of California San Diego *Doctor of Philosophy, Bioinformatics and Systems Biology*

PUBLICATIONS

Patrick C. Fiaux[†], Hsiuyi V. Chen, Poshen B. Chen, Aaron R. Chen, Graham McVicker[†]. 2020. “Discovering functional sequences with RELICS, an analysis method for CRISPR screens.” *PLOS Computational Biology*. DOI: 10.1371/journal.pcbi.1008194

Patrick C. Fiaux[†], Karthik Guruvayurappan, Graham McVicker[†]. “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+” Manuscript in preparation

Hsiuyi V. Chen, **Patrick C. Fiaux**, Aaron R. Chen, Ishika Luthra, Graham McVicker. 2020. “Discovery of novel, unmarked *GATA3* functional sequences.” Manuscript in preparation

Poshen B. Chen, **Patrick C. Fiaux**, Bin Li, Kai Zhang, Naoki Kubo, Shan Jiang, Rong Hu, Sihan Wu, Mengchi Wang, Wei Wang, Graham McVicker, Paul S. Mischel, Bing Ren*. 2021. “Discovery and Therapeutic Targeting of Pro-growth Enhancers in Human Cancer Cells.” *bioRxiv*. DOI: 10.1101/2021.02.04.429675

*Corresponding author. [†]Co-corresponding authors.

ABSTRACT OF THE DISSERTATION

Discovery of novel functional sequences through the analysis of tiling CRISPR screens

by

Patrick Christian Fiaux

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2021

Professor Graham McVicker, Chair
Professor Bing Ren, Co-Chair

Precise regulation of gene expression is crucial for organismal development. However, knowledge of regulatory genomic sequences (functional sequences), their targets, and modes of activation remains limited.

Recently, tiling CRISPR screens have been developed for the unbiased interrogation of the genome within its native context. These screens leverage the CRISPR-Cas9 system to perturb putative functional sequences and examine their effects on gene expression. This approach makes it possible to identify functional sequences as well as their target genes. In this dissertation I will highlight the aspects of tiling CRISPR screens that make them both attractive to use as well as difficult to analyze and present the different analytical approaches to date. Notably, I will describe our method RELICS, which models several key components of tiling CRISPR screens to accurately identify functional sequences.

In the first chapter I describe a simulation tool, CRSsim, which I developed to systematically evaluate different analysis methods for CRISPR screens against one another. This chapter highlights the importance of simulations and shows how I statistically recreated the generative process of data from CRISPR screens to simulate realistic data sets for benchmarking.

In the second chapter I present RELICS, a method developed specifically for identifying functional sequences from tiling CRISPR screens. I will describe how RELICS models the data and demonstrate that it outperforms all other methods which are currently used for analyzing tiling CRISPR screens.

Finally, I will present the results of RELICS applied to different experimental datasets, including publicly available datasets as well as data from our in house *GATA3* tiling deletion screen. Importantly, we discovered and validated novel functional sequences that were not detected by competing methods. Some of these sequences do not exhibit canonical epigenetic marks of regulatory elements, highlighting the importance of tiling CRISPR screens as an unbiased approach for detecting functional sequences and illuminating the regulatory landscape.

INTRODUCTION

Although each cell in our body consists of the same genetic material, we have dozens of different cell types that serve different functions. This diversity in cellular phenotypes is due to the differential expression of genes encoded in the genome. Specifically, gene expression is regulated by functional sequences (FSs), which are non-coding regions in the genome that are involved in turning gene expression 'on' and 'off' at specific timepoints. Disruption of these functional sequences can have deleterious consequences such as developmental defects and cancer (Buecker & Wysocka, 2012; Dawson & Kouzarides, 2012; Mansour et al., 2014). Furthermore, it is known that a disproportionate number of genetic variants associated with human traits and diseases do not fall within protein-coding genomic regions (Grubert et al., 2015; Maurano et al., 2012). Thus, it is hypothesized that a substantial fraction of variants are located within these functional sequences. Identifying the connections between genetic variants, the functional sequences in which they are located, and the target genes that these functional sequences regulate will significantly advance our understanding of the fundamentals of gene regulation and elucidate potential therapeutic targets for a variety of genetic conditions and diseases. Unfortunately, knowledge of where functional sequences are located in the genome and what their targets are remains limited despite advancements in the field.

This knowledge gap is due in part to the fact that identifying a functional sequence is not a straightforward task. DNA is wrapped around nucleosomes and depending on how tightly the nucleosomes are packed, specific regions of the genome can become more or less accessible to binding by proteins and other factors. It is generally believed that functional sequences reside in open chromatin regions in the genome, or euchromatin. Unlike heterochromatin, which is tightly packed, the open state of euchromatin allows the transcriptional machinery to access the DNA and initiate gene expression. The transcription start site (TSS) is the region of the gene body where RNA polymerase II (RNA Pol II) initiates transcription, and the region around the TSS (+/- 50 base pairs) is referred to as the core promoter. Binding of RNA Pol II to the promoter is typically sufficient to initiate transcription; however, in most cases the expression level is low (Haberle & Stark, 2018; Shlyueva et al., 2014). It is usually through the involvement of more distal elements that the full transcription rate is achieved. These distal elements are bound by transcription factors (TFs) which can in turn recruit co-activators. Elements involved in increasing gene expression are called enhancers and tend to contain nucleosomes whose histones have been modified by acetylation of lysine (K) 27 or monomethylation of lysine (K) 4 on histone 3 (H3K27ac and H3K4me1, respectively). These enhancers are brought in close proximity to the promoter of their target gene by chromatin looping (Amano et al., 2009), prompting the transcriptional machinery to assemble. In addition to being located in open chromatin regions, it is also believed that a subset of all functional sequences are evolutionarily conserved (Visel et al., 2009). Changes to a functional sequence perturb the expression of its target gene(s) and may have deleterious consequences. Thus, there is selective pressure to preserve these functional sequences. While open chromatin, histone modifications, chromatin looping, TF binding, and sequence conservation are associated with a majority of enhancers, they are not absolute indicators of functional sequences. Thus, identifying functional sequences in the genome is challenging because there is no finite and validated set of features that define them.

Various efforts have aimed to identify and characterize functional sequences in the genome to address this knowledge gap. Some of the most common experimental approaches include massively parallel reporter assays (MPRAs) (Melnikov et al., 2012; Patwardhan et al., 2012) or self-transcribing active regulatory region sequencing (STARR-seq) (Arnold et al., 2013). In

MPRAs, thousands of candidate enhancer sequences are synthesized and placed upstream of a reporter gene to assess whether the inserted sequence has an effect on the gene’s expression. STARR-seq uses a similar approach but places the synthesized sequences downstream of the reporter gene to quantitatively assess the effects of the putative enhancer encoded by the sequence. Computational methods have also been developed to identify functional sequences from the results of experimental methods. Notable methods include ChromHMM (Ernst & Kellis, 2012) and Segway (Hoffman et al., 2012), which partition the genome into distinct categories based on epigenetic marks and subsequentially assign function (e.g. enhancer or promoter) to these categories.

All of the methods described above have advanced understanding of the genome and its functional sequences. However, they also have limitations. For example, MPRAs and STARR-seq do not evaluate sequences in their native context. This means that the sequence under examination is not subject to all the interactions that it might experience *in vivo*. Additionally, all of the methods above only address whether a sequence is functional or not and cannot identify which genes are under the influence of each functional sequence. With the exception of chromatin looping, none of the features mentioned above (H3K4me1, H3K27ac, open chromatin, sequence conservation) are informative about regulatory targets. Unfortunately, obtaining chromatin looping data is expensive, not readily available, and too low-resolution to confidently identify functional sequences. Furthermore, functional sequences for an individual gene can vary between cell types to achieve cell-type specific expression, increasing the complexity of mapping functional sequences and their targets. Finally, there is increasing evidence that not all functional sequences are delineated by canonical regulatory marks (e.g. H3K27ac, open chromatin, etc.) (Diao et al., 2017; Rajagopal et al., 2016). Studying genomic regions associated with canonical epigenetic features will certainly lead to the discovery of new functional sequences; however, limiting the search to these regions will yield an incomplete picture of the regulatory landscape. All these factors complicate the identification of functional sequences and their target genes.

Recent technological advances in gene editing with CRISPR have made it possible to overcome the challenges described above with tiling CRISPR screens. Briefly, CRISPR stands for **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats and is found in nature as part of the bacterial defense system against viruses (Jinek et al., 2012). The CRISPR locus in bacteria contains short viral sequences. The defense system loads these viral sequences into a protein with nuclease activity and looks for sequences matching the loaded viral RNA, or the guide RNA (gRNA). If a match is found it is assumed to be a foreign viral sequence. The nuclease will introduce a double-stranded break (DSB) in the matched sequence, effectively neutralizing the hostile virus. This system has been adapted and optimized to effectively target most sequences in any genome (Doudna & Charpentier, 2014). The two main components of a CRISPR experiment are 1) a gRNA containing a ~20bp sequence matching its target site, and 2) the Cas9 enzyme, which is guided by the gRNA to the target site where it introduces a DSB. These DSBs are repaired via non-homologous end joining (NHEJ) (Jinek et al., 2013), which is an error prone endogenous DNA repair mechanism that perturbs the target site by leaving the repaired region with either a small insertion or deletion (indel). These indels are effectively DNA mutations.

The CRISPR-Cas9 system is now widely used in biological experiments. However, there are two important considerations regarding gRNAs that must be made to ensure the success of an experiment. The first is guide specificity. Because the genome contains 3 billion base pairs, it is possible that guides will match not only the target of interest but also another region in the genome. This would make it very difficult to determine if the observed effects are due to successful perturbation of the intended target sequence or ‘off-target’ effects. Designing guides to uniquely match a specific region with high specificity or low off-target probability is crucial for the quality of a CRISPR experiment. The second point of consideration is guide efficiency, or the likelihood that a guide will introduce a DSB. Not all guides are equally likely to perturb their target and selecting only efficient guides will also improve the quality of the experiment. There have been different approaches taken to characterize the features that make guides efficient or specific (Doench et al., 2014, 2016; P. D. Hsu et al., 2013) and there are several computational tools which can select the highest quality guides from a set of candidate sequences (McKenna & Shendure, 2018; Perez et al., 2017). Accounting for these properties of guides is very important for designing and analyzing a successful CRISPR experiment.

Today, CRISPR-Cas9 can be used to discover genomic sequences that affect cellular phenotypes such as growth, survival or gene expression. The first CRISPR screens targeted protein-coding genes (T. Wang et al., 2014) and provided valuable insight into protein function. They were subsequently adapted to interrogate the regulatory landscape for identification of functional sequences. These functional screens tile guides across non-coding regions of the genome (Canver et al., 2015; Diao et al., 2016, 2017; Fulco et al., 2016; Gasperini et al., n.d.; Klann et al., 2017; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016; Simeonov et al., 2017), allowing for the systematic discovery of novel functional sequences in their native genomic context. This unbiased approach to functional sequence discovery examines not only candidate regions containing canonical regulatory marks but also unmarked regions that would otherwise be overlooked based on such criteria and may yield novel insights into the properties of functional sequences.

In a tiling CRISPR screen, thousands of single-guide RNAs (sgRNAs) are designed to target sequences of interest in individual cells and induce genomic perturbations, which can include mutations, transcriptional repression (CRISPR interference, CRISPRi), or transcriptional activation (CRISPR activation, CRISPRa). To target sequences for mutation, sgRNAs are introduced into cells alongside Cas9 via lentiviral infection, inducing DSBs to introduce indels. In a CRISPRi experiment, targeted sites are silenced by a deactivated Cas9 (dCas9) fused to a repressive domain such as the Krüppel-associated box (dCas9:KRAB) (Gilbert et al., 2014; Thakore et al., 2015). Similarly, in a CRISPRa experiment, dCas9 is fused to an activation domain such as VP64 or p300 (Hilton et al., 2015; Konermann et al., 2015; Perez-Pinera et al., 2013) to activate transcription of the target region.

Following genomic perturbation, cells are subsequently sorted into pools based on a cellular phenotype (e.g. high vs. low gene expression, survival, proliferation, etc.) and the distributions of sequenced sgRNAs are compared across pools to identify functional genomic sequences with an effect on the cellular phenotype used for sorting. For example, sgRNAs that disrupt activating regulatory sequences would reduce expression of the reporter gene and consequently be enriched in pools selected for low target gene expression.

Despite the potential of these tiling CRISPR screens for improving functional sequence identification, they pose numerous challenges in the analysis of their results including (i) the need to combine information across multiple sequencing pools, (ii) the noisy and overdispersed nature of genomic count data (Love et al., 2014; van de Geijn et al., 2015), (iii) the spatial organization of sgRNA target sites and functional sequences, and (iv) the ‘area of effect’ (AoE) of genomic perturbations, which often extends well beyond the genomic location targeted by the sgRNA (e.g. in CRISPRa or CRISPRi screens activating or silencing epigenetic modifications can spread over

1kb or more from target sites (Thakore et al., 2015)). Currently, no existing methods address all of these challenges and moreover, almost all analysis methods for CRISPR screens were designed for screens that target protein coding genomic regions as opposed to tiling screens non-coding sequences where the identity of functional sequences are unknown.

We have developed two methods to improve; (1) our understanding of the properties of this new data and the different sources of variation, and (2) our analysis results via design of a modeling approach that leverages several features inherent to tiling CRISPR screens.

In chapter 1 I will discuss CRSsim, a simulation framework I developed for simulating tiling CRISPR screen data. CRSsim is capable of generating realistic simulations by modeling the generative process. Simulated datasets can help inform decisions about experimental design and assess method performance for comparing different analytical approaches.

In chapter 2 I present RELICS (**R**egulatory **E**lement **L**ocation **I**dentification in **C**RISPR **S**creens), a new method for the analysis of CRISPR screens which accounts for many aspects of tiling CRISPR screens. RELICS uses a flexible Bayesian hierarchical model to jointly analyze sgRNA counts across multiple pools and accommodates overdispersion in the count distributions while also considering the collective effects of adjacent sgRNAs and the potential presence of multiple functional sequences nearby. RELICS also reports the total number of functional sequences that are supported by the data. Using data simulated with CRSsim, we demonstrate that RELICS outperforms existing analysis methods (MAGeCK (W. Li et al., 2014), CRISPR-SURF (J. Y. Hsu et al., 2018), and MAUDE (de Boer et al., 2020)) with better precision and recall across a variety of conditions.

In chapter 3 I apply RELICS and other methods to experimental datasets. RELICS detects validated hits reported in previous studies and also discovers novel regulatory sequences missed by other methods, which we experimentally validated. I will also present the results from our own in house screen for functional sequences targeting *GATA3*, where we identified regions bearing canonical regulatory marks as well as several unmarked regulatory elements, highlighting the importance of the unbiased nature of tiling CRISPR screens for *de novo* functional sequence discovery.

This introduction is, in part, based on material from “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology*, 2020 by Patrick C. Fiaux, Hsiuyi V. Chen, Poshen B. Chen, Aaron R. Chen and Graham McVicker. The dissertation author was the primary investigator and author of this paper.

Chapter 1: Simulations of tiling CRISPR screen data with CRSsim

1.1 Abstract

With the development of tiling CRISPR screens, researchers have created a new way to interrogate the genome to discover functional sequences and the genes they regulate. However, because of the recent emergence of this method there is currently no gold-standard experimental dataset to serve as ground truth. This means that the available experimental datasets cannot be relied on for an accurate comparison of the performance between different analysis methods. To enable the comprehensive benchmarking of different analytical approaches, we have developed CRSsim for the simulation of tiling CRISPR screen data. We used various statistical approaches to simulate the generative process and demonstrate the similarity of the simulated data to experimental data. CRSsim is a very flexible tool and can simulate all currently available CRISPR systems in both selection and flow-sorting screens as well as many other features of tiling CRISPR screens.

1.2 Introduction

Tiling CRISPR screens allow for the unbiased perturbation of the genome to identify functional sequences (FSs) and the gene they regulate. Several important choices must be made when designing and executing a successful CRISPR experiment. We have categorized them into the following sections: (i) gene of interest, (ii) screen type, (iii) perturbation method, and (iv) perturbation coverage.

(i) **Gene of interest**. It is important to consider not only the target gene, its function and role in the system, but also the locus around it. Specifically, it is essential to ensure that there is a sufficient number of high-quality single-guide RNAs (sgRNAs) targeting the region of interest. There are different scoring methods for measuring a guide's efficiency and specificity (Doench et al., 2014, 2016; P. D. Hsu et al., 2013). Tools such as FlashFry (McKenna & Shendure, 2018) and GuideScan (Perez et al., 2017) take user-defined target regions and return all possible complementary guide sequences and their corresponding guide scores. These guides can then be filtered for high-quality guides which are more likely to perturb only their intended target.

(ii) **Screen type.** There are generally two types of screens. The first type uses fluorescent activated cell sorting (FACS) to sort cells into pools of categorical gene expression intensities. The most common ways to measure gene expression intensity with FACS is by either tagging the protein of interest with an antibody (Simeonov et al., 2017) or by fusing GFP or another fluorophore to the end of the gene of interest (Diao et al., 2016, 2017). The cells are then sorted into pools of different expression levels based on the fluorescence intensity. The analysis of a FACS screen essentially involves detecting guides which are enriched in either the low expression pool, indicating that they reduce expression of the gene of interest, or the high expression pool, indicating that they increase expression of the gene of interest. The second type of screen is a selection screen. This could be a proliferation screen, in which guides targeting regulatory sequences will lead to either decreased or increased cell growth (Fulco et al., 2016). Another option is to place the cells under selective pressure with a drug (Gasperini et al., n.d.; Sanjana et al., 2016). In both cases there is an initial pool at time point zero (‘before’) and a second pool at a final time point (‘after’). The analysis entails looking for guides that are either enriched or depleted in the ‘after’ pool.

(iii) **Perturbation method.** There are four different perturbation methods used in tiling CRISPR screens. These include (a) Cas9, (b) CRISPRi, (c) CRISPRa and (d) dual-guide CRISPR. Each of them has advantages and disadvantages, and they all have a different ‘area of effect’ (AoE). The AoE is the number of base pairs (bp) around the target site which will likely be affected by the perturbation. (a) Cas9 perturbations use CRISPR-Cas9 to introduce double-stranded breaks (DSBs) at the target site specified by the gRNA. The break is corrected via the error prone process of non-homologous end-joining (NHEJ), resulting in a small insertion or deletion (indel) of ~20bp (van Overbeek et al., 2016). The advantage of this approach is that the perturbation introduces a mutation in the sequence and alters the function of the target. However, because the indels are so short, it is difficult to tile a larger region while ensuring that each base pair is perturbed multiple times. (b) CRISPRi perturbations inactivate parts of the genome using a catalytically dead Cas9 (dCas9) fused to a repressive domain such as a KRAB domain (dCas9:KRAB). Even though the nuclease activity is deactivated, dCas9:KRAB will still bind at the target site specified by the sgRNA. This leads to methylation of the surrounding area, effectively repressing it. The AoE can be over 1kb (Thakore et al., 2015), making this approach very desirable to tile larger regions as it will be easier to achieve a high perturbation coverage. However, because Cas9 is deactivated, perturbations by CRISPRi do not cleave the DNA and will not lead in any sequence changes. (c)

CRISPRa also uses a dCas9 but unlike CRISPRi, it is fused to an activating domain such as VP64 or p300. This results in activation of the target by acetylating the region. Similar to CRISPRi, the AoE of CRISPRa can also reach 1kb. (d) Dual-guide CRISPR screens introduce two sgRNAs instead of one into a cell. The two guides introduce cuts in the DNA spaced 1-2kb apart and delete the entire region between them. This approach is challenging because both guides need to work correctly for the deletion to occur. Otherwise, each guide will simply introduce a small indel at its target site without removing the sequence between them. This also makes analysis of the results more challenging as no method so far models the possibility of an unsuccessful deletion of the region between the guide pairs. Despite these challenges, tiling deletion screens have great potential for enabling the interrogation of very large regions with a reasonable number of guides. Furthermore, this approach removes an entire region, thereby making it more likely that an effect will be observed if a functional sequence is successfully targeted. This is in contrast to the previous approaches where it is a possibility that the indel introduced by Cas9 may not be large enough to disrupt the sequence (a), or that the epigenetic alterations may not be effective enough (b,c).

(iv) **Perturbation coverage**. When designing a tiling CRISPR screen, it is recommended that each base pair is perturbed multiple times by different sgRNAs. This increases the likelihood of identifying a true effect even if some of the guides were lost during the library preparation step, had low guide efficiency or strong off-target effects. Repeated perturbation of the same base pair by different guides can improve signal resolution, but also increases the cost of preparing the guide library because more guides must be used. Additionally, with more guides the sequencing depth must be increased to ensure adequate coverage of all the guides. The AoE mentioned in (iii) can help determine the per-bp perturbation coverage, which can be adjusted by the step size between the gRNA targets. For example, Diao et al., 2017 performed a dual-guide CRISPR deletion screen covering a 2MB region with ~11,500 guide pairs and an average deletion size of 2kb such that each genomic position had on average 20 genomic deletions. Reaching this same deletion coverage would not even be possible for a Cas9 screen (a) and would require about 100,000 guides to simply target every base pair just once.

All of the categories described above are important considerations for a tiling CRISPR screen design, as well as for developing a simulation tool to generate *in silico* data that resembles experimental data. It is equally important to use the correct tools to process and analyze the results of a tiling CRISPR screen. However, up to this point there has not been a systematic comparison of different methods to assess their performance. This makes it difficult to decide which analysis approach will yield the highest quality results.

To address the above we have developed CRSsim. CRSsim can take a wide variety of user-defined inputs describing experimental conditions such as number of guides, type of CRISPR system, type of CRISPR screen, etc. By taking into account parameters such as guide efficiency, enhancer strength and the AoE of the selected CRISPR system, CRSsim can generate realistic datasets. These simulations allow for the *in silico* evaluation of different experimental design choices. They also provide a ground truth for evaluating the performance of different analytical methods.

1.3 Material and Methods

The simulations aim to reproduce the generative process of the data. This entails correctly understanding the different components that contribute to the observed data and the variation within it. Successfully simulating data also allows for the systematic evaluation of the performances of different methods. The goal of CRSsim is to integrate the different experimental components of tiling CRISPR screens and replicate them *in-silico* using different statistical models. We have taken the four main categories above (i-iv) and broken them down into different components of CRSsim; (I) Gene and locus size, (II) Screen type, (III) Perturbation method, (IV) Perturbation coverage, (V) Signal-to-noise ratio, (VI) Simulation-specific parameters.

(I) Gene and locus size: Both gene and locus size are important in the simulations. The size of the gene and the number of exons it contains are important for methods which leverage positive controls (Fiaux et al., 2020; J. Y. Hsu et al., 2018). CRSsim also generates a set of negative controls which is required by some methods (de Boer et al., 2020). The size of the simulated region impacts the number functional sequences that can be simulated as well as the number of guides that can be placed. For instance, a larger region can hold more enhancers and more guides, but this will also coincide with longer analysis runtimes.

(II) Screen type: CRSsim can simulate either a FACS screen with a specified number of pools or a selection screen. Both are modeled by a Dirichlet multinomial (see section (V) for details).

(III) Perturbation method: These are identical to the four perturbation methods mentioned above; (III.a) Cas9, (III.b) CRISPRi, (III.c) CRISPRa and (III.d) dual-guide CRISPR. The AoE for each of them is modeled by a normal distribution where the perturbation will always happen at the target site but is less likely to occur as a position located further away from the target site. The dual-guide CRISPR perturbations are modeled by two normal distributions at the two target sites and a uniform distribution in-between to model the probability of a full deletion instead of only two indels.

(IV) Perturbation coverage: This is impacted by the total number of guides, the step-size between the guides, the size of the locus as specified in (I) and the perturbation method selected in (III). CRSsim accounts for all these parameters when generating a tiling CRISPR screen.

(V) Signal-to-noise ratio: In CRSsim it is possible to change the signal-to-noise ratio of multiple components; Selection strength (V.a), dispersion modeling (V.b), guide efficiency (V.c), and functional sequence strength (V.d). Selection strength T (V.a), refers to the difference in sorting probabilities for a FACS screen, or depletion for a selection screen, for guides that do/do not target functional sequences. Both cases are modeled by a Dirichlet multinomial distribution. The probabilities of sgRNAs that do not target functional sequences are given by a vector of Dirichlet parameters α, where the probability of being placed into pool j is $\frac{\alpha_j}{\sum_{i=1}^{J}\alpha_i}$. Dispersion d (V.b) is the variance in the placement probabilities α and modeled as a variable dependent on the counts of the gRNA. Specifically, for c_n cells containing guide n the dispersion would be $d_n = disp(c_n)$ and $disp()$ can be either the default exponential function $(-89.18 + 17.78 * \log{(c_n)})$ or specified by the user. The Dirichlet parameters for a given guide n are therefore $\alpha_n = \alpha * (d_n * \sum_{i=1}^{J}\alpha_i)$. Guide efficiency f_n (V.c) and functional sequence strength h_k (V.d) also contribute to the placement probabilities. Both are proportions between 0 and 1, modeled by a beta distribution. Guides targeting strong functional sequences, with a strength near $h_k = 1$, behave like positive control sgRNAs. Guides targeting non-functional sequences have $h_k = 0$. The simulation sets the efficiency of each sgRNA and the strength of each functional sequence by sampling from beta distributions with user-configurable shapes. The sgRNA efficiency, f_n, specifies the fraction of cells where the sgRNA is ‘effective’ and perturbs the sorting probabilities. We also consider the distance of the guide target site to the functional sequence to be important. For example, if c_n cells contain sgRNA n, then the number of cells that contain a effective sgRNAs is $w_n = c_n f_n p_n$ where $p = N(r_E, sd_{Cas})$ and r_E is the range between the guide perturbation site and the functional sequence E. sd_{Cas} is the standard deviation such that for a Cas9 simulation only half the cells have a perturbation extending beyond 10bp from their target site and for a CRISPRi or CRISPRa simulation only half the cells have a perturbation extending beyond 200bp from the guide target site. The w_n cells are sorted with probabilities specified by the Dirichlet vector $\alpha_n + h_k T$. The sorting vector for the $c_n - w_n$ cells with ‘ineffective’ sgRNAs, is simply α_n.

(VI) Simulation-specific parameters: These parameters consist of; (VI.a) Guide library distribution, (VI.b) Total number of input cells, and (VI.c) Sequencing depth. The guide library distribution (VI.a) is the initial distribution of gRNA counts. We assume the counts follow a zero-inflated negative binomial distribution (ZINB), $y \sim ZINB(\mu, d, \varepsilon)$, where μ is the mean, d is the dispersion and ε is the proportion of the distribution that comes from the zero point mass. The parameters of the ZINB distribution can be specified or estimated by maximum likelihood from a provided table of gRNA counts. After the gRNA counts in the gRNA library are obtained by sampling from the ZINB distribution, an input pool of cells containing gRNAs is generated by performing multinomial sampling from the gRNA library to obtain the total number of input cells (VI.b). The input cells in combination with (V.a-d) generate the counts of either the FACS pools or the selected pool. CRSsim simulates the sequencing step (VI.c) by drawing from a multivariate hypergeometric distribution (sampling without replacement) or a multinomial distribution (sampling with replacement) to obtain the counts of gRNAs in the sorted pools. Sampling without replacement is used to simulate the use of unique molecular identifiers that allow duplicate reads to be filtered.

1.4 Results

We evaluated the similarity of data simulated by CRSsim to experimental data by comparing the distributions of the guide library counts, the measures of correlation in guide counts before and after selection, the rank changes in guide counts after selection, and the quantile-quantile plots of guide counts in the simulated vs. experimental data both before and after selection. Based on these metrics, we found the output of the simulations to be highly similar to the experimental data (Fig. 1.1b-d). Thus, we believe that the simulations by CRSsim can successfully capture the effects of experimental and biological variables present in CRISPR screen experiments.

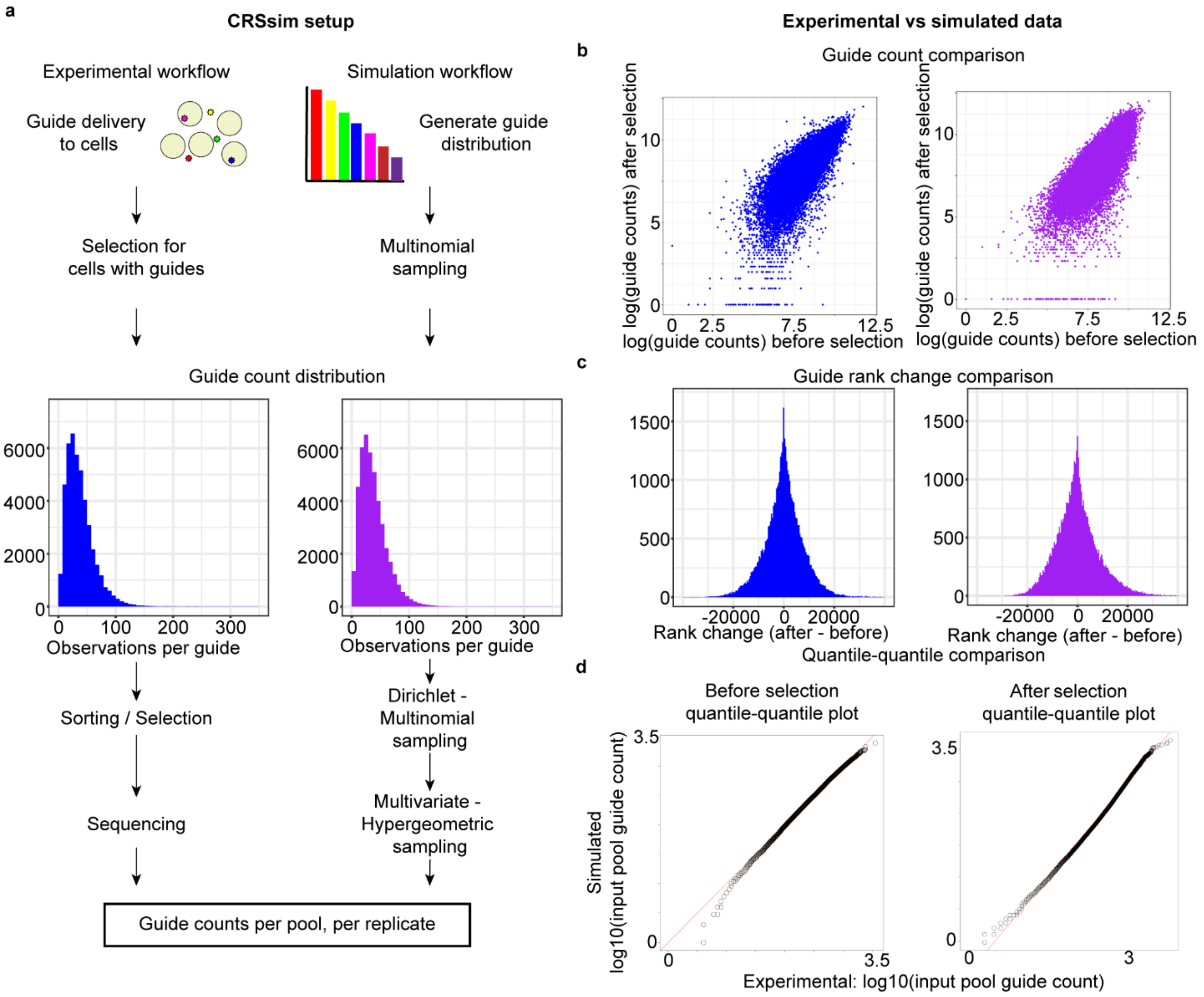


Figure 1.1. CRSsim simulation framework

CRSsim's CRISPR screen simulation framework. **(a)** Side-by-side comparison of the steps in a CRISPR screen experiment versus and the CRSsim generative process. sgRNAs are introduced into cells and only cells which receive sgRNAs are retained. The remaining cells are either sorted based on gene expression or placed under selective pressure (e.g. for survival or proliferation). In our simulations, we generate an initial sgRNA count distribution and mimic the experimental steps using several different sampling procedures. **(b-d)** Comparison between simulated and experimental data; sgRNA counts and guide ranks before and after selection are shown (b,c), as well as the quantile-quantile plots of simulated vs. experimental data both before and after selection (d).

These simulated datasets can be useful for characterizing the consequences of different experimental choices. For instance, it would allow a user to determine the tradeoff between using more guides but more shallow sequencing and vice versa. CRSsim can also be used to benchmark the performance of different analysis methods. These benchmarking analyses are described in Chapter 2.

1.5 Discussion

There is currently no substantial gold-standard set of known functional sequences where the ground truth is known. This impedes the systematic evaluation of different analysis method. With CRSsim, we provide a simulation framework that can generate realistic data sets under different experimental conditions. This simulated data can be used to compare the performance of different tools and identify the strengths and weaknesses of different methods. Furthermore, CRSsim can be used to make informed decisions about the design of a CRISPR screen experiment.

The features implemented in CRSsim make it a very flexible tool capable of generating many different scenarios. CRSsim is, to our knowledge, the only tool that simulates count data from tiling CRISPR screens, and we believe that it will be a useful resource for the community.

1.6 Data Availability

Code available from the Github repository: https://github.com/patfiaux/CRSsim

1.7 Acknowledgements

This research was supported by NIH/NIAID grant 2R01AI107027-06; by NIH/NIDDK grant 1 R01 DK122607-01; by the National Cancer Institute funded Salk Institute Cancer Center (NIH/NCI CCSG: 2 P30 014195); by a gift from the Jacobs Foundation; by a fellowship from the H.A. and Mary K. Chapman Charitable Trust to P.C.F.; by a fellowship from the Jesse and Caryl Philips Foundation to P.C.F; by a Salk Alumni Fellowship to H.V.C; and by the Frederick B. Rentschler Developmental Chair to G.M.

We thank Yarui Diao, Rongxin Fang, Ye Zheng, Zhi Liu, and members of the McVicker lab for helpful discussions about analysis methods for CRISPR regulatory screens; and Jessica Zhou and Arya Massarat for testing CRSsim.

1.8 Author information

G.M. and P.C.F. conceived of the idea for CRSsim. G.M. supervised the research. G.M. and P.C.F. wrote the manuscript, with input and edits from H.V.C. P.C.F. implemented CRSsim with the help of K.G. H.V.C. participated in many helpful discussions about CRSsim, and CRISPR screens.

Chapter 1, in part, is adapted from the material as it appears as “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology*, 2020 by Patrick C. Fiaux, Hsiuyi V. Chen, Poshen B. Chen, Aaron R. Chen, Graham McVicker and, in part, adapted from the manuscript in preparation “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+” by Patrick C. Fiaux, Karthik Guruvayurappan, Graham McVicker. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 2: RELICS (**R**egulatory **E**lement **L**ocation **I**dentification in **C**RISPR **S**creens)

2.1 Abstract

Here we describe RELICS (**R**egulatory **E**lement **L**ocation **I**dentification in **C**RISPR **S**creens), a new method for the analysis of CRISPR screens which specifically addresses the challenges described in chapter 1. RELICS uses a flexible Bayesian hierarchical model to jointly analyze sgRNA counts across multiple pools while explicitly modeling various tiling CRISPR screen processes. Using simulated data, we demonstrate that RELICS outperforms existing analysis methods with better precision and recall across a wide variety of conditions.

2.2 Introduction

The analysis of tiling CRISPR screens poses numerous challenges including (i) the need to combine information across multiple sequencing pools, (ii) the noisy and overdispersed nature of genomic count data (Love et al., 2014; van de Geijn et al., 2015), (iii) the spatial organization of sgRNA target sites and functional sequences, and (iv) the area of effect (AoE) of genomic perturbations, which often extend well beyond the sgRNA target site (e.g. in CRISPRa or CRISPRi screens, activating or silencing epigenetic modifications can spread over 1kb or more from target site (Thakore et al., 2015)). Currently, no existing methods address all of these challenges and moreover, almost all analysis methods for CRISPR screens were designed for gene-based screens, which knock out known genes, as opposed to tiling non-coding screens which aim to identify unknown functional sequences.

Our method RELICS is specifically designed to analyze tiling CRISPR screens and is, to the best of our knowledge, the first method which successfully models all of the above challenges (i-iv). We compare RELICS to three other methods; MAGeCK (W. Li et al., 2014), CRISPR-SURF (J. Y. Hsu et al., 2018) and MAUDE (de Boer et al., 2020). Briefly, MAGeCK was designed for analyzing gene-based screens and is one of the most popular methods for analyzing CRISPR screens. It has also been used to analyze tiling CRISPR screens (Diao et al., 2017); however, MAGeCK cannot combine information across multiple pools and makes a simplifying assumption that there either is or is not a perturbation and therefore does not account for AoE. CRISPR-SURF is a method designed for analyzing tiling CRISPR screens but does not model either raw counts or dispersion, and is also not able to jointly analyze multiple pools. MAUDE was also developed for analyzing tiling CRISPR screens and like CRISPR-SURF, it does not model raw counts or dispersion. Similar to MAGeCK, MAUDE makes simplifying assumptions about the AoE. While it does allow for the joint analysis of pools, it requires an even number of pools and the proportion of cells that were sorted into each pool. Unfortunately, the latter is unavailable for all but one of the publicly available data sets that we evaluated. Finally, none of the three methods are designed to model the AoE of dual guide screens. We demonstrate that the factors above (i-iv) are important to consider for the analysis of tiling CRISPR screens. By taking into account all of these factors, our innovative approach RELICS outperforms MAGeCK, CRISPR-SURF and MAUDE on evaluations of simulated data.

2.3 Material and Methods

2.3.1 RELICS

RELICS is designed to discover functional sequences from tiling CRISPR screens including those based on cell survival, cell proliferation, and gene expression. RELICS aims to determine both the number and location of functional sequences in the screened genome sequence and includes several important features: (i) it increases power by jointly modeling data from multiple pools; (ii) it models sgRNA counts appropriately without requiring transformation of the data or assuming normality; (iii) it explicitly models the guide dispersion; (iv) it accounts for guide efficiency scores; (v) it considers the spatial organization of sgRNA target sites and functional sequences; (vi) it accounts for the 'area of effect' (AoE) of sgRNAs; and (vii) it provides the credible set (CS) of segments that most likely contains the genome location of each predicted functional sequence.

One of RELICS' features is its ability to jointly analyze sgRNA count data across multiple pools, while controlling for extra variability (overdispersion) in the data. Modeling genomic count data while accounting for overdispersion increases power and reduces false positives (Love et al., 2014; Robinson et al., 2010a; van de Geijn et al., 2015). In RELICS, the counts are modeled using a Dirichlet multinomial distribution, which describes the probability that a cell containing an sgRNA will be observed in each pool. The dispersion for each guide is determined by a spline function which describes the relationship between the total counts of a guide and its dispersion (Fig 2.1a). RELICS estimates the spline from the data by sorting the guides according to total counts, binning them and estimating the dispersion with maximum likelihood estimation (MLE) for each bin. The mean of the guide counts for each bin and the estimated dispersions are used to fit a spline which in turn is then used to describe the dispersion for each guide.

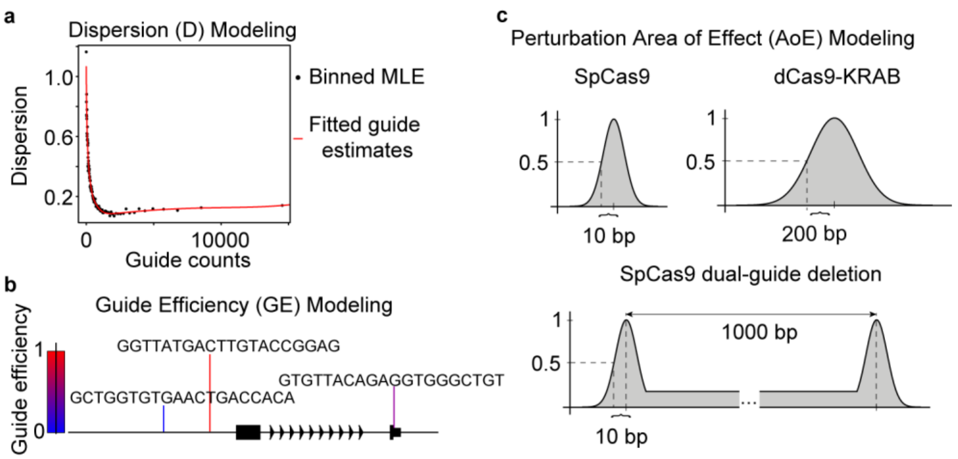


Figure 2.1. RELICS features

(a) Modeling guide count dispersion. Guides are sorted by total counts, binned and for each bin the dispersion is calculated via maximum likelihood estimation (MLE) using the guide counts. A spline is fit to the dispersion MLE estimates and the mean guide counts for each bin (black dots) to obtain per-guide dispersion estimates (red-line) **(b)** Modeling guide-efficiency scores. Each guide's efficiency can be scored based on different metrics. These scores can be integrated into the RELICS analysis model to increase result accuracy. **(c)** Modeling the area of effect (AoE). Each CRISPR system has a different AoE. For SpCas9 the probability that a position 10bp away from the target site is perturbed is reduced to only half the cells. CRISPRi's effect will extend past 200bp in half the cells and for a dual-guide deletion it's the same as for SpCas9 for each target site and a uniform probability for the entire deletion.

(iv) RELICS can also take into account guide efficiency scores. Over the last few years, it has become evident that guide efficiency plays an important role for the success of a perturbation (Doench et al., 2014, 2016; P. D. Hsu et al., 2013). Various metrics have been developed to filter for the best guides, scoring guides on efficiency, specificity and off-target effects (Fig. 2.1b). RELICS can use a provided set of scores to increase or decrease its certainty about the effect of a specific guide.

It is also important to consider the genome locations of both functional sequences (v) and sgRNA target sites as well as the area of effect (AoE) for each sgRNA (vi). The AoE of CRISPR perturbations means that a single sgRNA can potentially affect multiple functional sequences in the vicinity of the target site. Similarly, a single functional sequence can be perturbed by multiple sgRNAs with nearby target sites. To model the spatial organization of functional sequences

RELICS divides the sequence targeted by the screen into small windows called genome segments. Each genome segment is then associated with all sgRNAs that are predicted to affect it, based on their overlap with the sgRNA's AoE (Fig. 2.1c). Whether a segment is considered to be overlapped by a guide depends on the guide-AoE. For a sgRNA with Cas9 the AoE is considered to be around 20bp while with a dCas9:KRAB it's considered to be around 1kb. However, a perturbation is more likely to happen in close proximity to the target site and less likely the further away a region is located. RELICS models the AoE of sgRNAs using a normal distribution and the perturbation a dual-guide screen using sgRNA deletion probabilities at each target site and a uniform distribution to model the probability of a full deletion occurring instead of two indels.

RELICS assumes that the counts of sgRNAs are affected when the genome segment they are associated with contains a functional sequence. sgRNAs that affect functional sequences are expected to have a different count distribution across pools, and RELICS uses different Dirichlet distributions for sgRNAs that do or do not overlap functional sequences (Fig. 2.2). The hyperparameters for both distributions are estimated empirically by maximum likelihood in a supervised manner. Known functional sequences (positive controls), such as genome segments overlapping the promoter or established enhancer of the target gene are used to start this process. We refer to these positive controls as functional sequence 0 (FS0).

Using the hyperparameters and the observed sgRNA counts, RELICS calculates the posterior probability that a genome segment contains a functional sequence. We refer to the probability that a genome segment contains a functional sequence as the 'functional sequence probability' (FSP). The exact calculation of FSPs for every genome segment is impractical, however, because all combinations of functional sequence placements would have to be calculated. To overcome this problem, RELICS uses a novel Iterative Bayesian Stepwise Selection (IBSS) algorithm (G. Wang et al., 2020) to calculate approximate FSPs. Using the FSPs, RELICS computes all possible credible sets (CS), cs_i, by selecting a set of adjacent FSP, FSP_{i-j}, and identifying the segments of FSP_{i-j} that contribute the most to the signal. Each functional sequence is identified as the CS with the highest cumulative functional sequence probability.

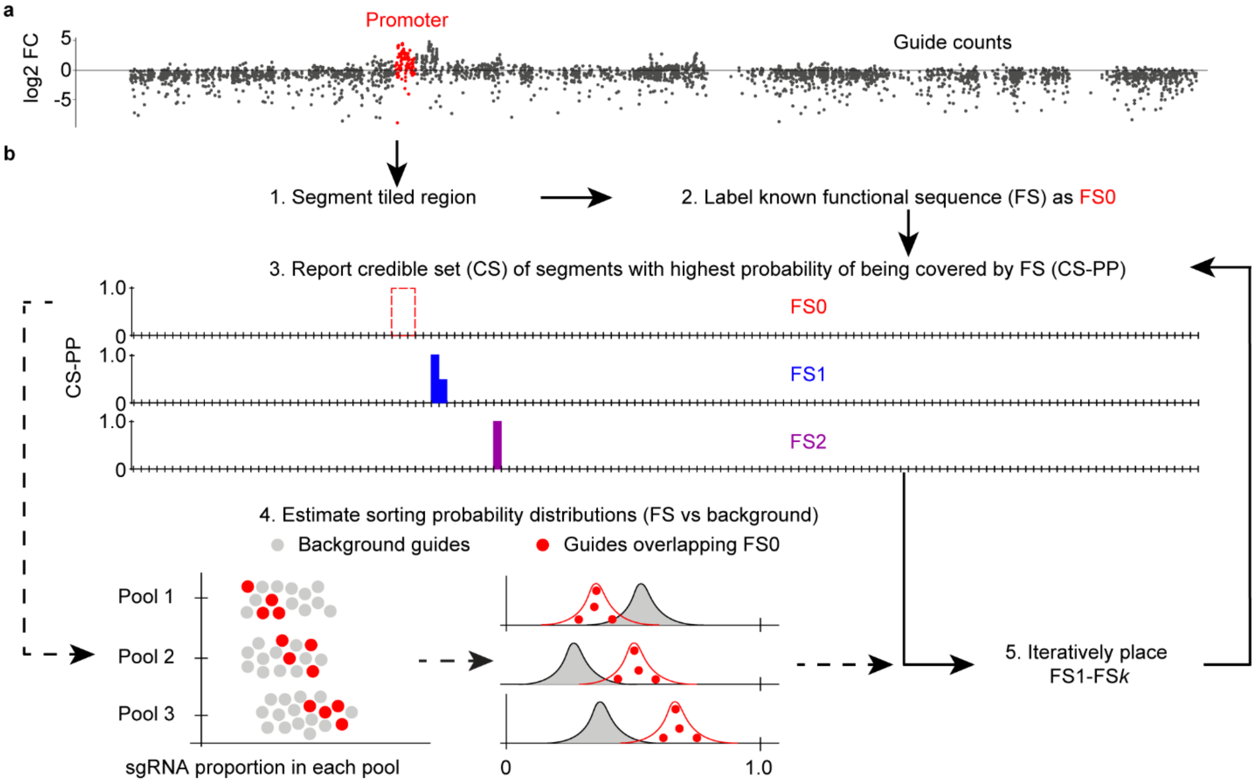

a
log2 FC
5
0
-5
Promoter
Guide counts
b
1. Segment tiled region
2. Label known functional sequence (FS) as FS0
3. Report credible set (CS) of segments with highest probability of being covered by FS (CS-PP)
CS-PP
1.0
0
FS0
FS1
FS2
4. Estimate sorting probability distributions (FS vs background)
Background guides
Guides overlapping FS0
Pool 1
Pool 2
Pool 3
sgRNA proportion in each pool
0
1.0
5. Iteratively place FS1-FS*k*

Figure 2.2. RELICS workflow

RELICS analysis workflow with example data. **(a)** Example input data with known functional sequence (FS), usually promoter or exons. Correctly identifying both the number and placement of functional sequences is non-trivial. Example data taken from a CRISPRa screen (Simeonov et al., 2017) discussed in chapter 3. **(b)** Main steps of RELICS for identifying FSs. After segmenting the tiled region and labeling FS0 (step 1, 2), the CS-PP of FS0 is placed (step 3). The sorting probability distribution is estimated by comparing guides overlapping FS0 and all other guides (background). This is done only once (step 4). Sorting probability distributions are used to iteratively identify the credible sets with the highest posterior probability of containing a functional sequence (step 5). This is done until *k* functional sequence are placed.

RELICS aims to identify a total of K functional sequences, where K is defined by the user. A prior is used to re-weight the model log-likelihood improvements with each additional functional sequence and those below a specified threshold are not reported. An important feature of RELICS is that it outputs separate credible sets for each functional sequence, providing discrete genome segments that can be used for follow-up validation experiments (Fig. 2.2b). We plot these functional sequence probabilities as separate tracks (Figs. 3.2a) or as a combined track with different colors (Figs. 3.2b, 3.3, 3.4, 3.5). In addition, the functional sequences output by RELICS are rank-ordered, with functional sequence 1 (FS1) having the strongest statistical support.

2.3.2 Formal Model

RELICS model setup and organization: A description of the variables for the RELICS model is provided in Table 2.1. RELICS divides the screened genome region into M small non-overlapping genome segments indexed as $m = 1,2, \ldots, M$. In practice we have found a segment size of 100bp to work well, and we use this as the default. As input, RELICS takes a table of observed counts for N sgRNAs (indexed as $n = 1,2, \ldots, N$) across J pools. We represent these counts as an $N \times J$ matrix, $\boldsymbol{y}$, and use $\boldsymbol{y}_n$ to denote the vector of observed counts for sgRNA n across the J pools. Let $g(n)$ be a function that maps sgRNAs to associated 'overlapping' genome segments. I.e. $g(n)$ returns the set of genome segments that are overlapped by sgRNA n. RELICS does not use non-targeting sgRNAs.

Table 2.1. Variable description for RELICS model

Variable	**Description**
J	Number of pools in screen
K	Number of functional sequences
L	Maximum length of functional sequence in genome segments
M	Number of genome segments
N	Number of sgRNAs
$\boldsymbol{y}$	Observed sgRNA counts (matrix of dimension $N \times J$)
$\boldsymbol{\delta}$	Functional sequence configuration (matrix of dimension $K \times M$). Each row, $\boldsymbol{\delta_k}$ specifies the placement (length and position) of functional sequence k. A specific placement is denoted $\delta_k[m, l]$, where m is the genome segment containing the start of the functional sequence and l is the length of the functional sequence.
$\boldsymbol{\pi}$	Probability a genome segment contains a specific functional sequence (matrix of dimension $K \times M$)
$\boldsymbol{\alpha}$	Hyperparameters for sorting probability distribution (vector of length J)
$\boldsymbol{s}$	sgRNA sorting probabilities (vector of length N)
$\boldsymbol{r}$	Number of genome segments that an sgRNA overlaps that contain a functional sequence (vector of length N)
$\boldsymbol{p}$	Probability genome segments contains any functional sequence (vector of length M)
l	Length of a functional sequence
$g(n)$	Mapping of sgRNAs to genome segments
d	Dispersion of Dirichlet distribution
P_i	Piecewise polynomial for the range t_i to t_{i+1}
e	Degrees of freedom
ge	Guide efficiency
$Dist(s\|\boldsymbol{\alpha}, AoE)$	Magnitude by which a functional sequence affects guide counts based on AoE and distance betweeb guide target site and bin position
s	Bin
AoE	Area of effect of CRISPR system

Description of variables in the RELICS model.

RELICS count model: RELICS assumes there are K functional genome sequences, indexed as $k = 1,2, \dots, K$ and that each functional sequence affects the sorting probabilities of overlapping sgRNAs. Each functional sequence has a length, l_k, which is the number of genome segments that it spans. The maximum length of a functional sequence is set to L, so that $l_k \in \{1, \dots, L\}$. By default, $L = 10$.

We describe the observed counts for an sgRNA across pools with a Dirichlet multinomial distribution and refer to the Dirichlet portion of the distribution as the sorting probability distribution. The sgRNAs that overlap a functional genome sequence have one sorting probability distribution, and the sgRNAs that do not overlap a functional genome sequence have another. Each Dirichlet distribution has J shape parameters, which define the dispersion and the expected proportion of counts in each pool. The vectors of shape parameters for the Dirichlet distributions are denoted $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_0}$ for sgRNAs that do and do not overlap a functional sequence. Then:

$$\boldsymbol{y_n}|\boldsymbol{s_n} \sim \text{Multinomial}(\boldsymbol{s_n})$$

$$\boldsymbol{s_n}|r_n = 0 \sim \text{Dirichlet}(\boldsymbol{\alpha_0}, d_{n,0})$$

$$\boldsymbol{s_n}|r_n > 0 \sim \text{Dirichlet}\left(\boldsymbol{\alpha_1}, d_{n,1}\right)$$

where r_n is the total number of genome segments containing functional sequences that are overlapped by sgRNA n. The dispersion parameter d for each guide n, d_n, is modeled via a spline and determined by the guide's total counts:

$$d_n = Spline(\boldsymbol{y_n}) = P_i\left(\sum \boldsymbol{y_n}\right), t_i \leq \sum \boldsymbol{y_n} < t_{i+1}$$

where P_i is the piecewise polynomial for the range t_i to t_{i+1} and $i = 0, \dots, e-1$ with e representing the degrees of freedom. To determine the value of e the total guide counts are sorted and binned and for each bin the dispersion is estimated using maximum likelihood while keeping the hyperparameters fixed to $\boldsymbol{\alpha_0}$ across all bins. The optimal value for e is then selected by the user.

Guide efficiency is modeled with a logistic function that adjusts the sorting proportions between $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_0}$ such that $\boldsymbol{\alpha_{1,n}} = \boldsymbol{\alpha_0} - (\boldsymbol{\alpha_0} - \boldsymbol{\alpha_1}) * ge_n$ where ge_n is the guide efficiency score for guide n. RELICS can also adjust the weighting on $\boldsymbol{ge}$ such that $\boldsymbol{ge_w} =$

$\frac{1}{1+\exp\left(-(\beta_0+\beta_1*\boldsymbol{ge})\right)}$ where β_0 and β_1 are determined using maximum likelihood estimation. This format is very flexible and allows for the combination of different sets of guide scores $\boldsymbol{ge_w} = \frac{1}{1+\exp\left(-(\beta_0+\beta_1*\boldsymbol{ge_1}+\beta_2*\boldsymbol{ge_2}\dots)\right)}$ where $\boldsymbol{ge_1}$ and $\boldsymbol{ge_2}$ are two different guide scores.

<u>RELICS functional sequence configurations</u>: We define a configuration to be the positions and lengths of all of the functional sequences. We specify a configuration with a matrix, $\boldsymbol{\delta}$**,** of dimension $K \times M$, where an element, $\delta_{\boldsymbol{k},\boldsymbol{m}}$**,** is 1 if genome segment m contains functional sequence k, and is 0 otherwise. We call a single row vector of the configuration matrix, $\boldsymbol{\delta_k}$, the placement of a functional sequence. In other words, a configuration is a collection of functional sequence placements.

We want to estimate the probability, p_m, that a given genome segment, m, contains a functional sequence. To compute p_m we could sum the posterior probabilities of all configurations that have a functional sequence in genome segment m. However, exact calculation of p_m is intractable because the likelihoods of all possible configurations must be computed. For example, even in a simple case where there are $M = 10{,}000$ genome segments and $K = 5$ regulatory sequences of length $L = 1$, the number of possible $\boldsymbol{\delta}$ configurations is $\binom{10{,}000}{5} = 8.3 \times 10^{17}$.

To overcome this problem, we developed an approximate inference algorithm known as Iterative Bayesian Stepwise Selection (IBSS) (G. Wang et al., 2020). Our version of the IBSS algorithm includes extensions that are specific to RELICS including allowing for functional sequences of variable length and the use of non-normal count data with a Dirichlet multinomial error distribution. Our IBSS algorithm performs stepwise placement of a single functional sequence at a time, while accounting for the (uncertain) placements of all of the other functional sequences.

To implement the IBSS algorithm, and to account for uncertainty in functional sequence placements, we introduce a functional sequence probability matrix, $\boldsymbol{\pi}$. This is like the $\boldsymbol{\delta}$ matrix, but rather than binary values, it contains probabilities. Specifically, each element is the probability that a genome segment contains a specific functional sequence: $\pi_{k,m} = \Pr(\delta_{k,m} = 1)$.

To allow the positions of known functional sequences (positive controls) to be specified, we add an additional row to the functional sequence probability matrix, which we index as row 0 and denote $\boldsymbol{\pi_0}$. The probability that a genome segment contains *any* functional sequence is then: $p_m = \sum_{k=0}^{K} \pi_{k,m}$.

The number of genome segments that are overlapped by sgRNA n and contain a functional sequence follows a Poisson binomial distribution. In other words, the Poisson binomial is used to calculate the probability that an sgRNA overlaps r_n genome segments containing functional sequences:

$$r_n \sim \text{PoissonBinomial}(\boldsymbol{p}_{g(n)})$$

where $\boldsymbol{p}_{g(n)}$ is the vector of probabilities for all genome segments associated with sgRNA n. To incorporate the AoE when calculating the probability that sgRNA n has r overlapping functional sequences (computed using the Poisson binomial distribution);

$$\sum_{s} \Pr_{\text{PB}}(r|\boldsymbol{p_s}) * Dist(s|\boldsymbol{\alpha}, AoE)$$

where $Dist(s|\boldsymbol{\alpha}, AoE) = \begin{cases} N(b - sgT, \mu = 0, \sigma_{AoE}), if\ \boldsymbol{\alpha} = \ \boldsymbol{\alpha_1}, AoE = normal \\ 1 - N(b - sgT, \mu = 0, \sigma_{AoE}), if\ \boldsymbol{\alpha} = \ \boldsymbol{\alpha_0}, AoE = normal \end{cases}$ and s is the bin, b is the position of the bin, sgT is the target position of the guide and σ_{AoE} is the standard deviation corresponding to the Cas9 model. In the case of regular Cas9 $\sigma_{AoE} = 8.5$ and for both CRISPRi and CRISPRa $\sigma_{AoE} = 170$. If $AoE = uniform$, corresponding to RELICS.1, then $Dist(s|\boldsymbol{\alpha}, AoE) = 1$ for all $\boldsymbol{\alpha} = \ \boldsymbol{\alpha_1}$ and $Dist(s|\boldsymbol{\alpha}, AoE) = 0$ for all $\boldsymbol{\alpha} = \ \boldsymbol{\alpha_0}$.

RELICS IBSS algorithm**:** The following is a description of the IBSS algorithm that is used to estimate the functional sequence probability matrix $\boldsymbol{\pi}$.

```
Initialize:
    • Set $K$ to number of functional sequences.
    • Set known functional sequences (positive controls) to 1.0 in row $\boldsymbol{\pi_0}$.
    • Set all other elements of $\boldsymbol{\pi}$ to 0.0.
    • Estimate sgRNA sorting hyperparameters $(\alpha_0, \alpha_1)$ by maximum likelihood.

  # Estimate configuration probabilities
  • For $k$ in $1 \ldots K$:
        o Set elements of row $\boldsymbol{\pi_k}$ to 0.
        o Compute $\boldsymbol{p}$, the probability that each genome segment contains one of the
          other functional sequences
        o Set elements of $\boldsymbol{\pi_k}$ by calculating posterior probability of every possible
          placement of functional sequence $k$, conditional on $\boldsymbol{p}$.
        o Compute all possible $CS_k$ and select the one with highest PP as final CS
```

RELICS posterior probabilities of functional sequence placements: To compute the posterior probability of each functional sequence placement, we first compute the likelihood of all possible placements of functional sequence k, taking into account the placements of all of the other functional sequences. We set the elements of $\boldsymbol{\pi_k}$ to 0 and also set all elements of $\boldsymbol{\delta_k}$ to 0, except those corresponding to the functional sequence placement being considered, which are set to 1. We denote a specific placement as $\boldsymbol{\delta_k}^{<m,l>}$**,** where m is the genome segment containing the start of the functional sequence, and l is the length of the functional sequence. That is, $\boldsymbol{\delta_k}^{<m,l>}$, is a vector of 0s except for elements $m..(m + l - 1)$, which are set to 1. We can compute the probability that a given genome segment contains a functional sequence as $p_m = \delta_{k,m} + (1 - \delta_{k,m}) \sum_{k=0}^{K} \pi_{k,m}$. The likelihood of a specific functional sequence placement is then:

$$\mathcal{L}\left(\boldsymbol{\delta_k}^{<m,l>} \middle| \boldsymbol{y}, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_0}, \boldsymbol{\pi}\right)$$

$$= \prod_n \left[\sum\nolimits_s \Pr_{\mathrm{PB}}\left(r_{\boldsymbol{n}} > 0 | \boldsymbol{p_{g(n)}}\right) * Dist(s|\boldsymbol{\alpha}_1, AoE) \Pr_{\mathrm{DMN}}(\boldsymbol{y_n}|\boldsymbol{\alpha_1})\right.$$

$$\left. + \sum\nolimits_s \Pr_{\mathrm{PB}}\left(r_n = 0 | \boldsymbol{p_{g(n)}}\right) * Dist(s|\boldsymbol{\alpha_0}, AoE) \Pr_{DMN}(\boldsymbol{y_n}|\boldsymbol{\alpha_0})\right]$$

where the product is over all sgRNAs with observed counts, $\Pr_{\mathrm{DMN}}(\boldsymbol{y}|\boldsymbol{\alpha})$ is the probability of the observed counts (computed using the Dirichlet multinomial distribution).

The posterior probability of a specific functional sequence placement is:

$$PP_{\boldsymbol{\delta_k}^{<m,l>}} = \frac{\Pr\left(\boldsymbol{\delta_k}^{<m,l>}\right) \mathcal{L}\left(\boldsymbol{\delta_k}^{<m,l>} \middle| \boldsymbol{y}, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_0}, \boldsymbol{\pi}\right)}{\sum_{i,j} \Pr\left(\boldsymbol{\delta_k}^{<i,j>}\right) \mathcal{L}\left(\boldsymbol{\delta_k}^{<i,j>} \middle| \boldsymbol{y}, \boldsymbol{\alpha_1}, \boldsymbol{\alpha_0}, \boldsymbol{\pi}\right)}$$

where the denominator is the sum over all possible placements of this functional sequence and $\Pr\left(\boldsymbol{\delta_k}^{<m,l>}\right)$ is the prior probability of a specific functional sequence placement.

We can compute the posterior probability that a genome segment contains functional sequence k, by summing the probabilities from all of the possible placements overlapping the segment. We use these posteriors to set the elements of $\boldsymbol{\pi_k}$:

$$\pi_{k,m} = \sum_{l=1}^{L} \sum_{i=m-l+1}^{m} PP_{\boldsymbol{\delta_k}^{<i,l>}}$$

RELICS prior probabilities on length of functional sequences**:** As prior probabilities for each functional sequence placement, we use a weighting that favors shorter functional sequences. Specifically, we use a geometric distribution, truncated at a maximum of L, to weight each possible length:

$$w(l) = \frac{(1-\lambda)^{l-1}\lambda}{\sum_{i=1}^{L}(1-\lambda)^{i-1}\lambda}$$

where λ is a constant between 0 and 1 that controls the weighting. We make the prior uniform for all placements with the same value of l.

RELICS calculation of Credible Sets. For each FS, RELICS uses the functional sequence specific functional sequence probability placements, $\boldsymbol{\pi_k}$, to compute a 90% credible set for 10 adjacent segments, for all possible CS. The segments of the CS with the highest cumulative posterior probability are then identified as part of the configuration matrix, $\boldsymbol{\delta_k}$.

RELICS empirical estimation of hyperparameters**:** The RELICS model has hyperparameters, $\boldsymbol{\alpha_0}$ and $\boldsymbol{\alpha_1}$, which control the sorting probabilities and dispersion of sgRNA counts across pools. RELICS performs MLE of these parameters each iteration of the IBSS algorithm. This estimation is performed using the full dataset of sgRNA counts, and keeping the functional sequence probabilities fixed to their current estimates, $\boldsymbol{\hat{\pi}}$:

$$\alpha_0, \alpha_1 = \text{argmax}_{\alpha_0,\alpha_1} \mathcal{L}(\boldsymbol{\alpha_1}, \boldsymbol{\alpha_0} | \boldsymbol{y}, \boldsymbol{\hat{\pi}})$$

We perform MLE by numerical optimization using the L-BFGS-B algorithm (Byrd et al., 1995).

RELICS prior on number of functional sequences (K)**:** RELICS computes the iterative placement of K functional sequences using the IBSS algorithm described above. The log-likelihood contribution of each functional sequence to the model is weighted by a user-defined prior, FS_e, where $prior \sim binomial(K, FS_e)$. RELICS will report all K functional sequence that are significant in a chi-square test with one degree of freedom. Based on current screens, RELICS sets $FS_e = \frac{total\ nr.\ bp\ covered}{50{,}000}$ and $K = FS_e + \max{(3, 0.3 * FS_e)}$ as default.

RELICS Input Format. RELICS takes a text-based .csv file as input (for examples see https://github.com/patfiaux/RELICS). The file contains the sgRNA information as well as the observed sgRNA counts in different pools. We have made the public datasets we analyzed available in this format (https://figshare.com/projects/RELICS_2_data/74376,). Alternatively, RELICS can also take in two .csv files, where one file contains the sgRNA information and the other file contains the observed sgRNA counts in different pools.

2.3.3 Other tiling CRISPR screen methods

We have compared the performance of RELICS against three other methods. The first one is MAGeCK (W. Li et al., 2014) which has been developed for analyzing gene knockout screens. MAGeCK (version 0.5.9.2) is designed to be run on sgRNAs that are grouped into functional units (i.e. genes). To run MAGeCK, we therefore grouped sgRNAs into non-overlapping genome windows containing 10 sgRNAs per window (note that the GeCKO v2 sgRNA library that MAGeCK is commonly applied to uses 6 sgRNAs per gene). To run MAGeCK we used the following command line:

```
mageck test -k MAGeCK_Input.txt -t 0,1 -c 2,3 -n mageckOut
```

where the MAGeCK_Input.txt file contains the observed counts for each sgRNA.

We next compared against CRISPR-SURF (J. Y. Hsu et al., 2018). CRISPR-SURF (downloaded 02/13/2019 from GitHub) was run using the procedure specified in the supplemental materials of Hsu et al. 2018. CRISPR-SURF analyzes count data in two steps. The first step filters out guides of low quality and computes log2 fold change (Command ‘SURF_count’). The second step performs the deconvolution (Command ‘SURF_ deconvolution) to identify functional sequences. During the first step, all guides which have a count less than 50 in any of the pools get removed. While this works with data sets that have been sequenced deeply, this can become an issue when the sequencing depth is not as high. For this reason, to run CRISPR-SURF on the simulated datasets, we set the guide count filter to 15 instead of 50. For the experimental datasets we used the default guide count filter of 50. CRISPR-SURF accepts positive control as input, which can improve the output. We provided CRISPR-SURF with the same positive controls that RELICS used for training. The SURF_count command was run as follows:

```
docker run -v $Path_to_file/:$Path_to_file -w $Path_to_file pinellolab/crisprsurf SURF_count -f CRISPR_SURF_count.csv -nuclease cas9 -pert crispri
```

The SURF_deconvolution command was run as follows:

```
docker run -v $Path_to_file/:$Path_to_file -w $Path_to_file pinellolab/crisprsurf SURF_deconvolution -f CRISPR_SURF_Input.csv -pert crispri
```

Lastly, we also looked at the performance of MAUDE (de Boer et al., 2020) (downloaded 11/16/2020 from GitHub). We ran MAUDE by following the steps from the *CD69* tutorial and computing the combined Stouffer Z-score.

Performance assessment in simulations. We compared the performance of analysis methods for CRISPR screens using average precision, which summarizes a precision-recall curve, using scores provided by each method. Specifically, as scores we used the CS functional sequence probability (CS-PP) for RELICS, the -log10(FDR) for MAGeCK, the -log10(p-value) for CRISPR-SURF, and the Z-scores for MAUDE.

2.4 Results

Here we have used RELICS to analyze different simulated datasets generated with CRSsim (Fiaux et al., 2020) under eight different scenarios with 15 simulations for each. In total, we simulated four paired gRNA deletion screens and four CRISPRi screens with single gRNAs (Fig. 2.3). For both types of screens, we varied both guide efficiency and enhancer strength between the low and high pools (Table 2.2).

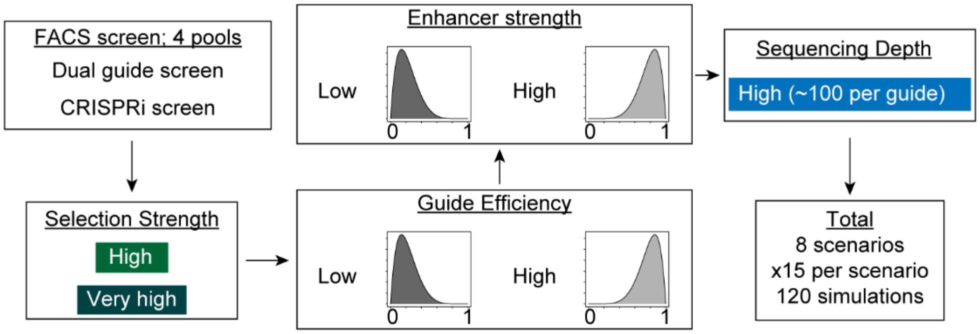


Figure 2.3. Simulated tiling CRISPR screen data

Scenarios simulated with CRSsim. A total of eight scenarios were simulated. For very high selection strength, only high guide efficiency and high enhancer strength were used. For high selection strength, both high and low guide efficiency and enhancer strength were used (see Table 2.2 for details).

Table 2.2. Simulation parameters

Simulation name	Screen Type	Selection Strength	Guide Efficiency	Enhancer Strength
Scenario 1	Dual-Guide	Very high	High	High
Scenario 2	Dual-Guide	High	High	High
Scenario 3	Dual-Guide	High	Low	High
Scenario 4	Dual-Guide	High	High	Low
Scenario 5	CRISPRi	Very high	High	High
Scenario 6	CRISPRi	High	High	High
Scenario 7	CRISPRi	High	Low	High
Scenario 8	CRISPRi	High	High	Low

Description of simulation parameters used for each scenario.

We quantified performance by summarizing the precision recall curve with the average precision (AP) for each method. Overall, RELICS demonstrated the best performance with the highest mean AP in all scenarios (Fig. 2.4-7).

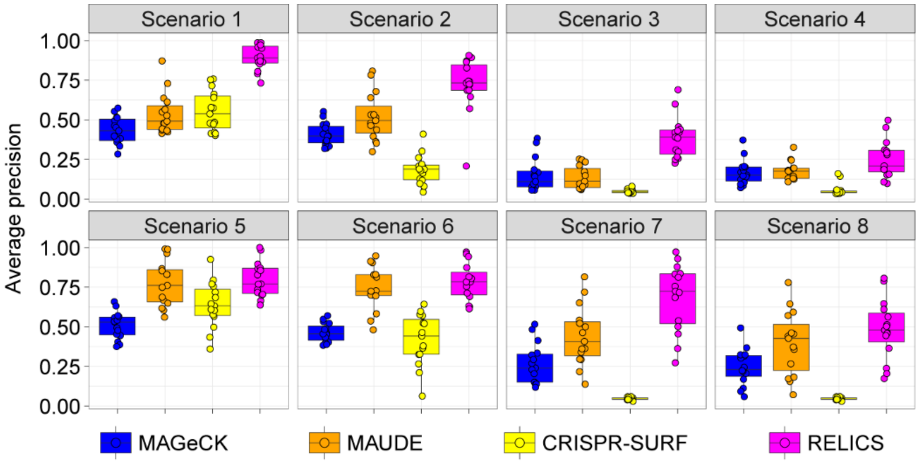


Figure 2.4. Average precision results for all simulated scenarios
Scenarios 1-4 are dual-guide simulations; scenarios 5-8 are CRISPRi simulations (see Table 2.2 for details). Performance was compared between MAGeCK, MAUDE, CRISPR-SURF and RELICS.

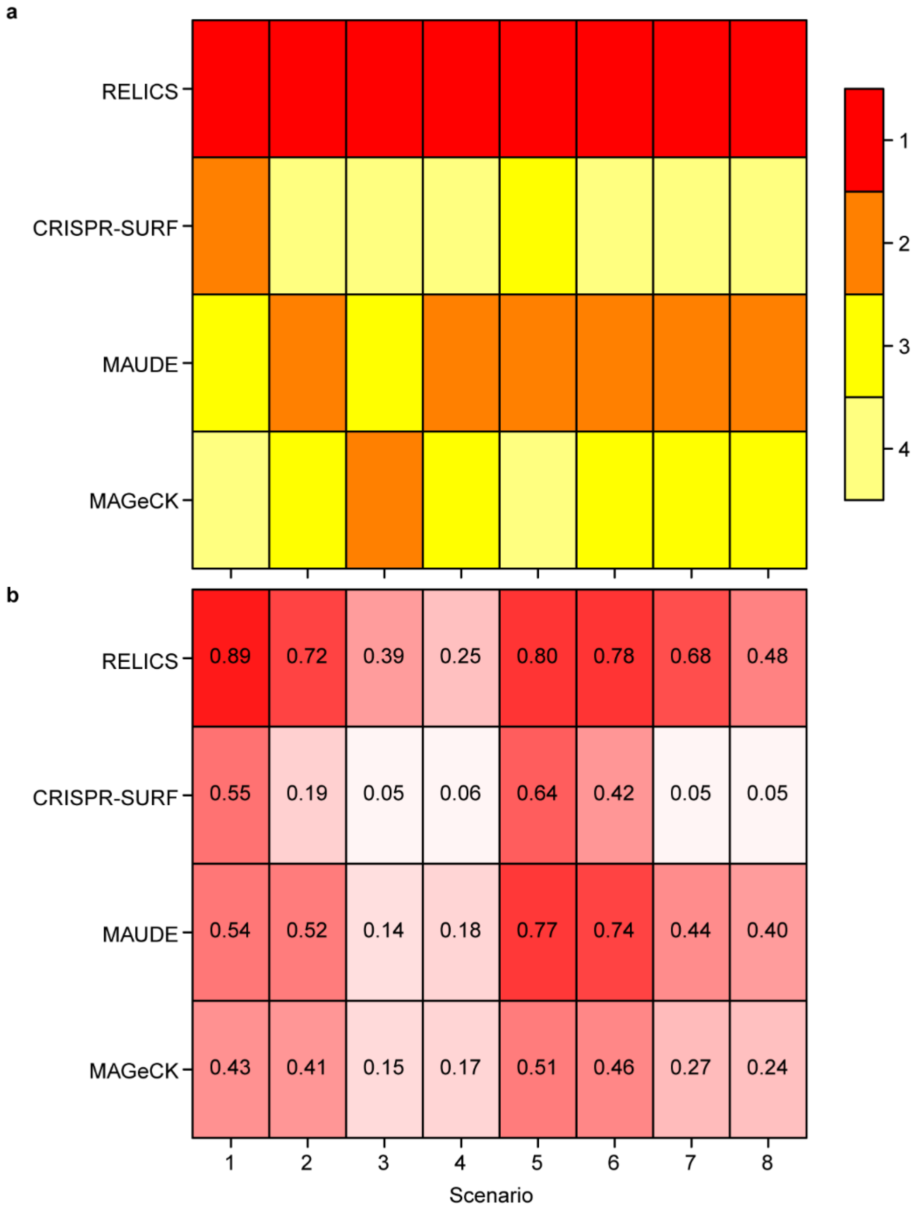


Figure 2.5. Heatmap of method performance

(a) Mean average precision rank. **(b)** Mean average precision scores.

We also assessed how individual features of RELICS affected method performance by iteratively dropping each feature from the model (Fig. 2.6-7). The feature with the largest impact on performance was the modeling of the relationship between dispersion (D) and total gRNA count, which dramatically improved performance across all scenarios when included.

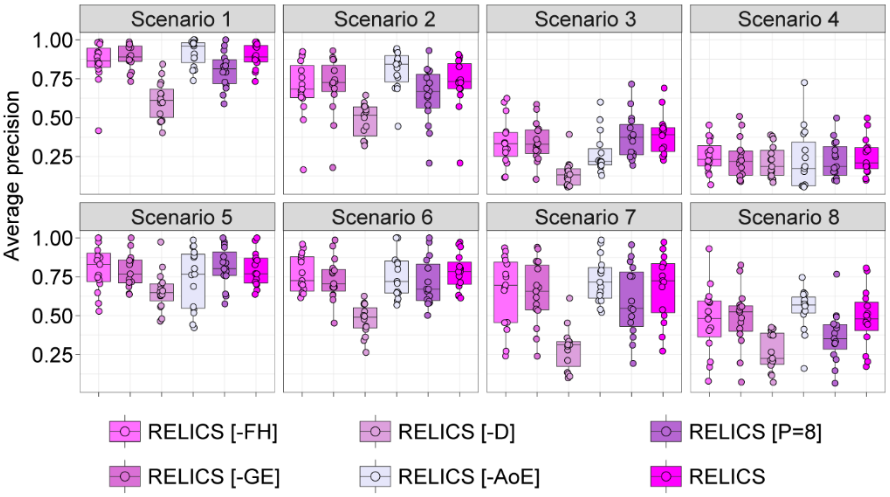


Figure 2.6. RELICS feature performance

Scenarios 1-4 are dual-guide simulations; scenarios 5-8 are CRISPRi simulations (see Table 2.2 for details). Performance was compared between RELICS with all features included and with individual features dropped from the model. RELICS [-FH]: modeled without fixing hyperparameters and instead recomputing them after each functional sequence placement; RELICS [-GE]: modeled without guide efficiency; RELICS [-D] : modeled without dispersion; RELICS [-AoE]: modeled without area of effect; RELICS [P=8]: modeled with priors set to 8 instead of 4.

Modeling AoE was particularly important for RELICS performance in the paired guide screen with more challenging settings (scenarios 3 and 4), highlighting the importance of including this feature when analyzing dual-guide screens.

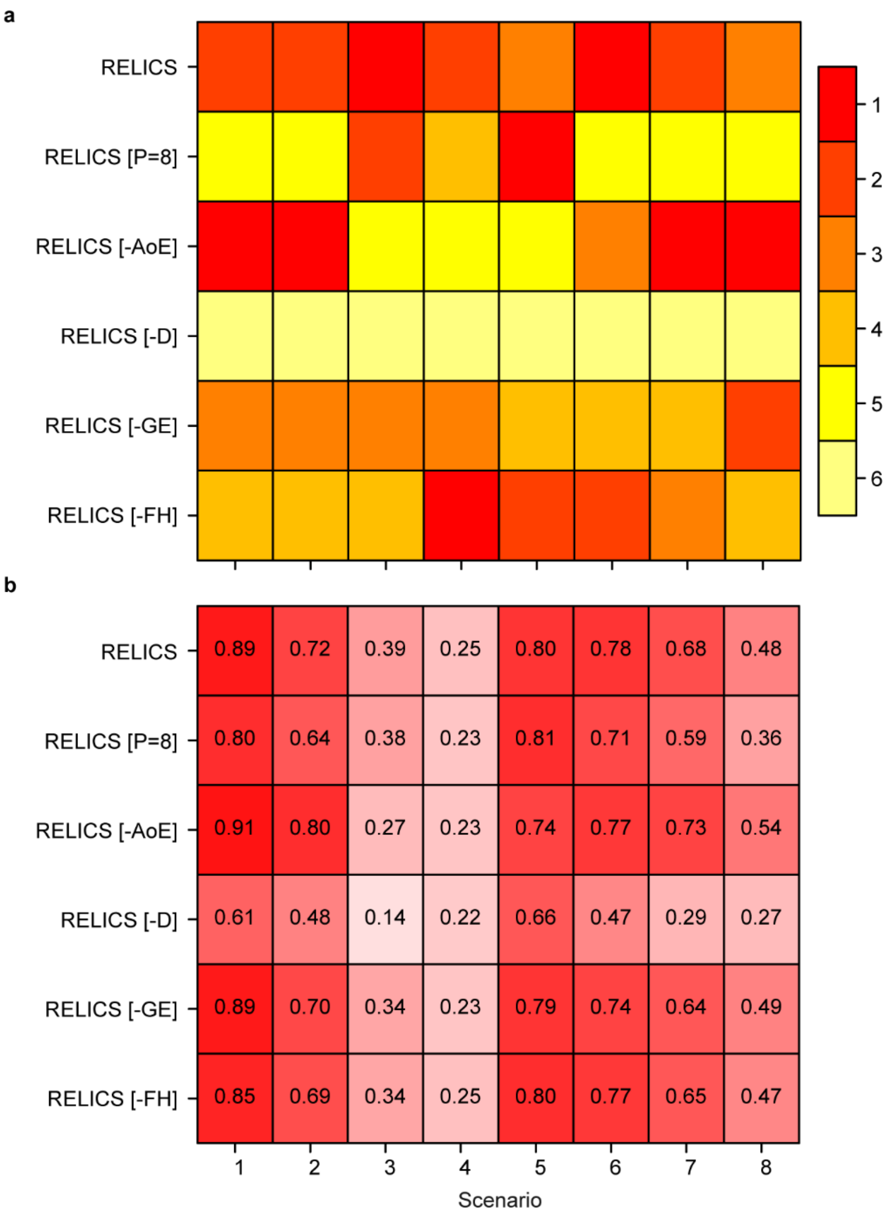


Figure 2.7. Heatmap of RELICS feature performance

. **(a)** Mean average precision rank. **(b)** Mean average precision scores.

We also examined the effects of changing the prior (P) on the number of functional sequences and found that setting the prior to the same number as the true number of functional sequence resulted in the same or slightly diminished performance due to the detection of false positive functional sequences. Therefore, we recommend setting the prior at or slightly below the expected number of functional sequences. When modeling guide efficiency (GE), we found that although it only improved RELICS performance slightly, it was robust against noisy guide efficiency estimates (Fig. 2.8a), indicating that even scores which inaccurately capture guide efficiency (Fig. 2.8b) still provide useful information to the model. Therefore, we suggest including guide efficiency in the analysis wherever possible.

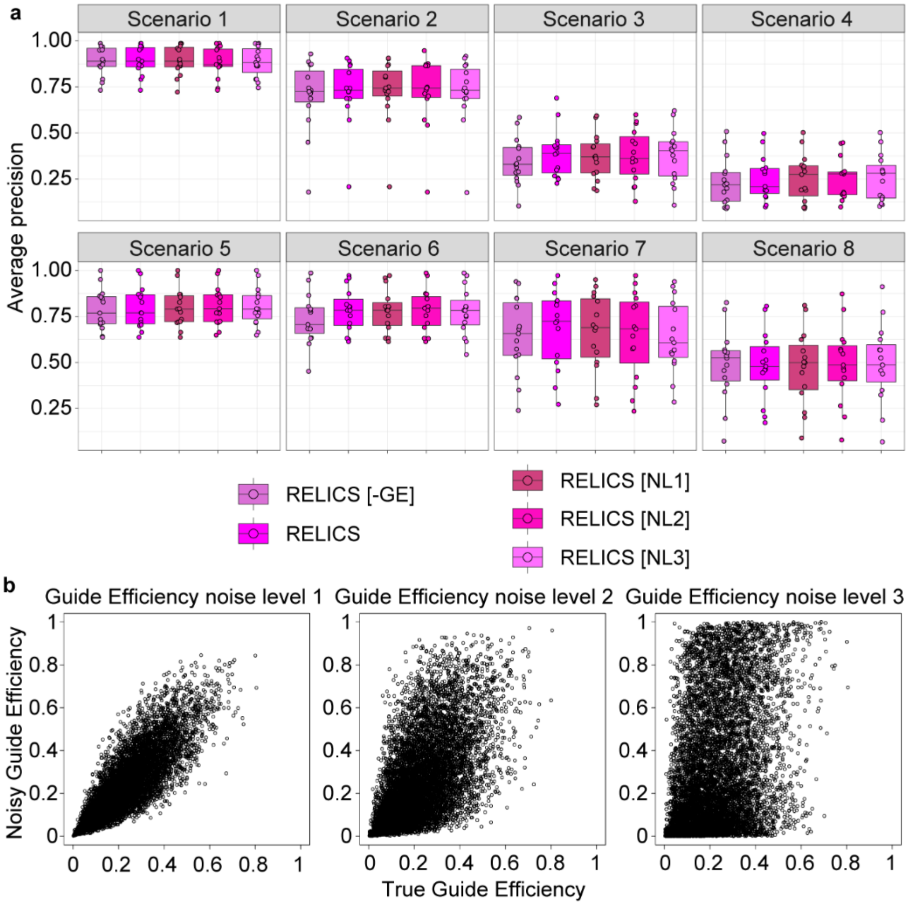


Figure 2.8. RELICS guide efficiency modeling

(a) Average precision of RELICS in different scenarios with different noise levels (NL). **(b)** Scatterplot of true guide efficiency against guide efficiency noise levels 1-3.

Lastly, we looked at the effect of recomputing the hyperparameters after the placement of each functional sequence rather than keeping them fixed. Doing so resulted in either equal or slightly worse performance. This could be due to RELICS including false positive functional sequences in the hyperparameter estimates. Additionally, keeping the hyperparameters fixed speeds up the runtime. Both aspects indicate that RELICS will work better with just the initial hyperparameter estimates from FS0. Overall, our simulations demonstrate that RELICS outperforms all other methods in our simulations.

2.5 Discussion

We have developed RELICS, a Bayesian hierarchical model for the analysis of CRISPR screens. Unlike gene-based analysis methods, RELICS is specifically designed to analyze CRISPR screens where the locations of functional sequences are not known. The RELICS model provides numerous advantages. First, it considers the collective effects of multiple nearby sgRNA target sites and functional sequences. Second, it provides interpretable probabilistic outputs for each functional sequence that can be used to delineate small genomic regions that contain each functional sequence with confidence. Third, it models sgRNA counts appropriately without requiring transformation of the data or assuming normality. Fourth, it increases power for functional sequence discovery by jointly modeling data from multiple sgRNA pools. Fifth, it specifically models the relationship between gRNA counts and dispersion. Sixth, it considers the AoE of each CRISPR perturbation method in both single guide and dual guide scenarios. Seventh, it can use guide efficiency scores to improve the detection of functional sequences. While other methods include some of these features (e.g. CRISPR-SURF deconvolves the effects of multiple sgRNAs and MAGeCK models overdispersed count data with a negative binomial distribution), only RELICS combines all of them into a single model.

RELICS leverages known functional sequences, labeled positive controls, as well as unlabeled sequences to learn model hyperparameters. CRISPR-SURF similarly makes use of known positive control sequences while MAUDE requires negative control sequences. A strength of the RELICS approach is that it learns the behavior of sgRNAs across different pools from the data. A limitation is that it may be difficult to apply RELICS to datasets that do not contain a known positive control region or have a low number of sgRNAs overlapping positive control regions, although it is generally advisable to include positive controls in all screen designs. In addition, positive controls may not adequately represent all types of functional sequences, such as repressive or weak regulatory elements. As an alternative approach, the hyperparameters can be specified by the user. It may also be possible to develop an unsupervised learning approach where positive control labels are not provided and instead sequences with sgRNAs are identified by clustering on similar behavior. Ideally, the categories identified by such an approach would represent different types of sequences (e.g. strong regulatory elements, weak regulatory elements, silencers, non-regulatory elements). Future versions of RELICS that can account for this spectrum of functional sequence behavior will likely achieve even better results.

In summary, RELICS is able to outperform all other methods used for analyzing tiling CRISPR screens by modeling numerous biologically and technically relevant features. Thus, we believe that RELICS is an extremely useful tool for the discovery of functional sequences from CRISPR screens.

2.6 Data Availability

RELICS is available on GitHub at https://github.com/patfiaux/RELICS. The simulated data sets can be found at https://figshare.com/projects/RELICS_data/98219.

2.7 Acknowledgements

This research was supported by NIH/NIAID grant 2R01AI107027-06; by NIH/NIDDK grant 1 R01 DK122607-01; by the National Cancer Institute funded Salk Institute Cancer Center (NIH/NCI CCSG: 2 P30 014195); by a gift from the Jacobs Foundation; by a fellowship from the H.A. and Mary K. Chapman Charitable Trust to P.C.F.; by a fellowship from the Jesse and Caryl Philips Foundation to P.C.F; by a Salk Alumni Fellowship to H.V.C; and by the Frederick B. Rentschler Developmental Chair to G.M.

We thank Yarui Diao, Rongxin Fang, Ye Zheng, Zhi Liu, and members of the McVicker lab for helpful discussions about analysis methods for CRISPR regulatory screens; and Jessica Zhou and Arya Massarat for testing RELICS.

2.8 Author information

G.M. and P.C.F. conceived of the idea for RELICS. G.M. supervised the research. G.M. and P.C.F. wrote the manuscript, with input and edits from H.V.C. P.C.F. implemented RELICS and analyzed the data with the help of K.G. H.V.C. participated in many helpful discussions about RELICS.

Chapter 2, in part, is adapted from the material as it appears as “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology*, 2020 by Patrick C. Fiaux, Hsiuyi V. Chen, Poshen B. Chen, Aaron R. Chen, Graham McVicker and, in part, adapted from the manuscript in preparation “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+” by Patrick C. Fiaux, Karthik Guruvayurappan, Graham McVicker. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 3: Analysis of tiling CRISPR screens

3.1 Abstract

Tiling CRISPR screens provide an unbiased approach to interrogate the genome and discover gene-specific functional sequences to give insight into the mechanisms that govern the regulation of gene expression. However, the analysis of these screens is challenging and previous methods may have missed some relevant functional sequences, yielding an incomplete picture of the regulatory landscape. Here we have performed a systematic analysis of seven publicly available CRISPR screen data sets with RELICS, comparing the results to what has previously been reported and what other analysis methods detected in the same datasets. We also report the results from the analysis of our own unpublished tiling deletion screen for *GATA3*. Altogether, we demonstrate that RELICS has an improved performance over previous methods and can successfully analyze both single-guide and dual-guide screens to identify novel functional sequences corroborated by experimental validation.

3.2 Introduction

Tiling CRISPR screens have the potential to interrogate the region surrounding a gene of interest in an unbiased fashion, allowing for the discovery of not only canonical functional sequences but also novel sequences which might not display characteristic epigenetic marks such as H3K27ac. However, these screens have only recently been developed and the analysis of the data they generate poses numerous challenges. Additionally, every study that has performed tiling CRISPR screens thus far has used a different analysis method. This is in part due to fact that robust tiling CRISPR screen analysis methods have only recently been developed (de Boer et al., 2020; Fiaux et al., 2020; J. Y. Hsu et al., 2018). Additionally, there has not been a systematic comparison between existing methods, making the choice of analysis more complicated. However, using simulations from CRSsim I have shown in the previous chapter that RELICS outperforms all other methods. This leads to the intriguing question of whether RELICS can discover additional previously unreported reported functional sequences when reanalyzing public data sets. Here I will focus on the analysis of six genes from previous CRISPR screen studies (Diao et al., 2017; Fulco et al., 2016; Gasperini et al., n.d.; Simeonov et al., 2017) and demonstrate that RELICS finds not only previously reported sequences but also additional putative functional sequences which we experimentally validate. We also show that previous methods (J. Y. Hsu et al., 2018; W. Li et al., 2014) have difficulties analyzing dual-guide screen while RELICS can successfully recover validated sequences and finds additional putative functional sequences supported by sequence conservation and H3K27ac.

Lastly, I will also present the results from an unpublished tiling deletion screen for *GATA3* from the McVicker lab. In addition to canonical functional sequences, RELICS also identifies unmarked functional sequences which we confirm through validations, highlighting the importance of using an unbiased approach when searching for gene-specific functional sequences. One of these unmarked functional sequences is overlapped by potentially causal variants for allergic diseases and asthma, allowing us to better understand immune disorders associated with the disruption of specific *GATA3* functional sequences.

3.3 Material and Methods

3.3.1 CRISPRa screens by Simeonov et al.

Here we use two datasets from CRISPRa screens for functional sequences that affect the expression of *CD69* and *IL2RA* in Jurkat T cells (Simeonov et al., 2017). For both genes, cells were flow sorted into four pools based on expression (negative, low, medium, high) and putative functional sequences were identified by computing log fold change of target gene expression between pairs of pools. We downloaded the sgRNA counts from the supplemental data of the study and converted them to the RELICS input format. The promoter regions of the target genes were used as positive control regions for RELICS and CRISPR-SURF and were defined as the region $+/-$ 1kb around the transcription start site (TSS). RELICS was used to jointly analyze all pools for the analysis. For MAGeCK and CRISPR-SURF, the input pool (all the cells in the experiment prior to sorting) was used as the control pool and the high-expression pool was used as the treatment pool. RELICS and CRISPR-SURF used the same sgRNAs for positive controls. For RELICS, we used priors of 5 and 15 for evaluating *CD69* and *IL2RA* respectively, based on results from our previous analysis (Fiaux et al., 2020). For the validation experiments we used FlashFry (McKenna & Shendure, 2018) to design two sgRNAs per target site. FlashFry calculates guide efficiency and off-target effects for each guide by reporting the Doench 2014 (Doench et al., 2014), Doench 2016 (Doench et al., 2016), and Hsu 2013 scores (P. D. Hsu et al., 2013). For each of the target sites, we selected guides with low estimated off-target effects and high estimated efficiency relative to the other possible guides in the region (Table 3.1). The synthesized sgRNAs were cloned into pCRISPRiaV2 plasmids (Addgene #84832). Lentiviruses carrying the sgRNAs were generated and transduced into Jurkat cells expressing dCas9-VP64 (Simeonov et al., 2017), obtained from the Berkeley Cell Culture Facility. 7-10 days after transduction, cells from different treatments (i.e. transduced with sgRNAs for distinct target sites) were stained with PE anti-human *CD69* antibody (Biolegend #310906) and the expression of *CD69* was measured by flow cytometry using a BD FACSCanto II system.

Table 3.1. *CD69* validation guides

Target gene	region	gRNA	hg19 coordinate (chr12)
*CD69*_tss_sg1	TSS	CAATGTATAGTGTGTTGTTG	9913617-9913636
*CD69*_tss_sg2	TSS	TCAAGCAAGTAGGCGGCAAG	9913557 -9913576
*CD69*_FS1_sg2	FS1	AGGTAACCATGAGTAAACGG	9917833-9917852
*CD69*_TNC_sg1	'Neg. Target'	CCATTTCCCTCCACAAGCCC	9919830-9919849
*CD69*_FS4_sg1	FS4	GCATAGAATTGATATCACCA	9925936-9925955
*CD69*_FS4_sg2	FS4	GGATTCGTCTTCTGAGTTAC	9925975-9925994
*CD69*_FS3_sg1	FS3	TATTCCTGCTGTATAACAGA	9950460-9950479
*CD69*_FS3_sg2	FS3	GTTAAATAATAGAGGGCACA	9950613-9950632
NTC		CTGAAAAAGGAAGGAGTTGA	

CD69 CRISPRa validation sgRNAs. TSS sgRNAs target the transcription start site of *CD69*. FSs, and 'Neg. Target' sgRNAs correspond to labels in Fig. 4d. NTC is a non-targeting control sgRNA.

3.3.2 CRISPRi screens by Fulco et al.

Fulco et al., 2016 performed a CRISPRi proliferation screen surrounding both the *MYC* and the *GATA1* loci in K562 cells. We downloaded the sgRNA counts from the paper supplement and converted them to RELICS input format. For detecting *MYC* regulatory elements, all sgRNAs labelled as '*MYC* Tiling' sgRNAs as well as 'Protein Coding Gene Promoters' sgRNAs were used. For detecting *GATA1* regulatory elements, all sgRNAs labelled as '*GATA1* Tiling' sgRNAs as well as 'Protein Coding Gene Promoters' sgRNAs were used. For MAGeCK and CRISPR-SURF, the pool at T0 (before) was used as the treatment pool and the pool at T14 (after) was used as the control pool in order to look for enrichment instead of depletion. RELICS and CRISPR-SURF used the same positive control sgRNAs. RELICS priors were set to 15 for *MYC* and 5 for *GATA1* based on results from our previous analysis (Fiaux et al., 2020). For the validations K562 cells were transduced with dCas9-KRAB-GFP lentivirus (Addgene #71237) carrying different sgRNAs (Table 3.2). The percentage of GFP positive cells was recorded by FACS 3 days after transduction (D0) and again after an additional 6 (D6) and 14 (D14) days of culture. Two biological replicates (from separate cultures) were performed at each time point.

Table 3.2. *MYC* CRISPRi validation sgRNAs

gRNA target	sgRNA	d0_r1	d0_r2	d6_r1	d6_r2	d14_r1	d14_r2
FS12	GATAAGAAAACGGAGCCATC	54.2	56.22	48.79	50.16	48.21	47.51
FS12	GACGTGATCAGCAGCCATAG	25.64	25.31	18.46	19.28	18.23	18.81
e2 (*MYC*)	CCTGGAAAGACAACAGCTTG	50.86	49.65	39.14	40.48	34.17	33.44
Neg. control	GACAAGCTGCAAGGTGTAAAT	47.25	46.63	46.03	45.27	41.01	41.71
dCas9:KRAB only		26.72	27.79	27.95	27.15	25.44	26.14
K562 wildtype		0.14	0.14	0.13	0.13	0.38	0.38

Guide sequences are given for each target. The positive control targets a known functional sequence (e2) identified by Fulco et al. As negative controls a dCas9:KRAB only or an sgRNA targeting a non-functional safe harbor region on chromosome 8 were used.

3.3.3 Tiling screens by Gasperini et al.

Gasperini et al. performed 2 screens on the *HPRT1* locus in HAP-1 cells. In addition to a dual-guide screen they also performed a sgRNA screen, tiling single guides along the same locus. Both screens were selection screens with two replicates, each resulting in a pool before and after selection. We removed guides according to their filtering criteria and subsequently used GuideScan (Perez et al., 2017) to remove additional guide (see below). For RELICS' FS0 and CRISPR-SURF positive control all *HPRT1* exon overlapping guides were used. Additionally, for RELICS we used a prior of 2 for both data sets as no functional sequence were discovered in the original study.

We filtered all guides from both the CRISPRi and CRISPRa as well as the Gasperini experimental screens using GuideScan. All possible guides for the regions targeted in each screen were obtained from the online version of the GuideScan tool (http://www.guidescan.com/, used 05/08/2020). GuideScan eliminates all guides with either a perfect match or a 1bp mismatch at any other site in the genome. All remaining guides receive a specificity score. We used the specificity score cutoff of 0.2 recommended by Tycko et al. (Tycko et al., 2019). All remaining guide sequences were then matched to the guide sequences from the studies described above. Filtering sgRNAs with GuideScan reduced the total number of guides in each study substantially (Table 3.3). For the *HPRT1* dual-guide screen we removed guide pairs where at least one of the guides had a specificity score below 0.2.

Table 3.3. GuideScan filtering

Data set	Initial nr. Guides	Final nr. Guides
CD69	10650	3025
IL2RA	19751	7869
GATA1	6034	988
MYC	73241	10718
HPRT1 single-guide	26324	6438
HPRT1 dual-guide	11365	5969

Number of sgRNAs present in each data set before and after filtering with GuideScan.

3.3.4 Tiling deletion screen by Diao et al.

Diao et al. 2017 performed a tiling deletion screen on the *OCT4* locus in H1-hESC. They flow-sorted cells from input pools (Ctrl1, Ctrl2) into pools of high and low (Cis1-5) expression. The published putative functional sequences were identified by using a modified version of MAGeCK. We have downloaded their data from their supplementary section and used the set of guides which passed their filtering criteria. For RELICS' FS0 and CRISPR-SURF positive control all *OCT4* overlapping guide-pairs were used. We used the same pools they did for their study

(Cis1-5, Ctrl1-2). However, because it was not clear which pools were from the same replicate we analyzed them all together with RELICS and specified a prior with 45 expected functional sequence (they found 45 in their study). For MAGeCK we used the 5 low pools and the 2 input pools and for CRISPR-SURF we used both input pools and Cis2 and Cis3.

All screens above were analyzed with MAGeCK and CRISPR-SURF as described in chapter 2. All default settings were used for RELICS with the addition of also using guide specificity information as provided by GuideScan for *MYC*, *GATA1*, *CD69, IL2RA* and *HPRT1*. Because MAUDE also requires the proportion of cells sorted into each pool it was not possible to run it for any of the data sets except for the *CD69* data for which they reported estimates of the *CD69* sorting proportions.

3.3.5 Tiling deletion screen for *GATA3*

We performed a tiling deletion screen to identify functional sequences regulating *GATA3* expression. The guides were selected for a Hsu score above 75 and were obtained with an in house script. Guides were paired such that the average deletion size was ~1k and the step size at ~65 bp. As was described previously (Gioia et al., 2018), Jurkat cells contain various chromosomal rearrangements. We removed guide-pairs that were located in regions containing homozygous deletions, translocation, inversions, long deletions, long insertions, short deletions and short insertions. This resulted in 13,253 guides, targeting a 2MB region around *GATA3* with a median step size of ~100bp (Fig. 3.1). Following the experimental procedure from Diao et al. 2017 we generated a total of 4 replicates out of which two were flow-sorted into high, medium and low

GATA3 expression pools (replicates 1 and 2) and the other two into high-high, high, high-low, medium-high, medium, medium-low and low *GATA3* expression pools (replicates 3 and 4).

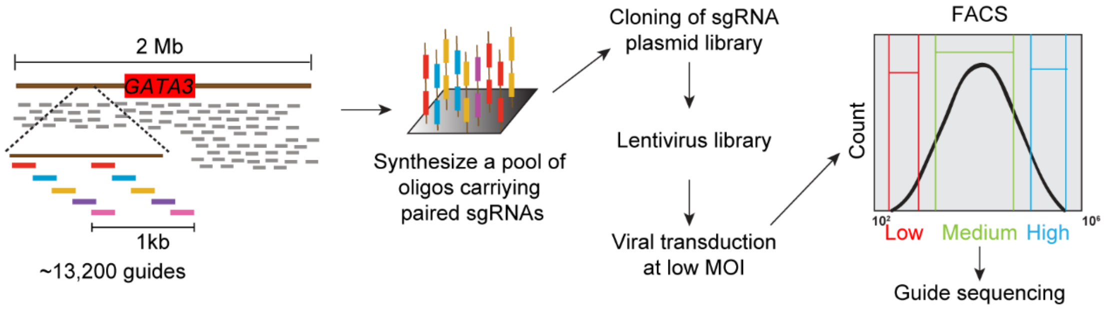


Figure 3.1. ***GATA3*** **tiling deletion experimental design**
A 2 MB region around the *GATA3* locus was tiled with 13,253 guides, deleting ~1kb per pair with a step size of ~100 bp. A lentiviral library was used to transduce the guides at a multiplicity of infection (MOI) of 0.3 into Jurkat cells. The input pool was flow-sorted into different pools of based on *GATA3* expression.

For RELICS' FS0 we used all guide-pairs overlapping *GATA3* exons and set the prior to 45 as we expected a similar number of functional sequences as was reported by Diao et al. since we were also targeting a 2MB region.

We also obtained summary statistics from an allergic diseases and asthma cross-trait meta-analysis (Z. Zhu et al., 2018), obtained at (http://lianglab.rc.fas.harvard.edu/AsthmaAllergyHeritability/). Fine mapping to identify putative causal variants was done with SuSiE (G. Wang et al., 2020).

Hi-C data was obtained from (Lucic et al., 2019).

3.3.6 Epigenetic data

Jurkat data. Jurkat H3K27ac data from Mansour et al. (Mansour et al., 2014) was downloaded from GEO (GSM1296384). The reads were aligned with BWA-MEM (H. Li, 2013) using default parameters and filtered for duplicates and low mapping quality (MAPQ < 30) using samtools (H. Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, & Durbin, 2009). The reads were then converted to reads per kilobase per million in 200bp bins using deepTools2 (Ramírez et al., 2016). ENCODE H3K27ac ChIP-seq data for the K562 cell line was downloaded in bedgraph format from the UCSC genome browser. H3K27ac peaks were called with macs2 using the 'bdgpekcall' function: `macs2 bdgpeakcall -i input.bedGraph -c 2 -l 245 -g 100 -o output_peaks.bed`. H3K4me3 peaks were downloaded from ENCODE (ENCFF400IIQ). DNase hypersensitive sites were also downloaded from ENCODE (NCFF688ZSR and ENCFF304GVP). We used bedtools to intersect the overlapping peaks: bedtools intersect -a dhs.el1 – b dhs.repl2 -wa

K562 data. We downloaded peak calls for H3K27ac (ENCFF044JNJ), H3K4me1 (ENCFF183UQD) and H3K4me3 (ENCFF737AMS) from ENCODE (Sloan et al., 2016) (https://www.encodeproject.org/).

HAP-1 data. We downloaded peak calls for H3K27ac (ENCFF646CAB), H3K4me1 (ENCFF049JIN) and H3K4me3 (ENCFF962XKU) from ENCODE.

H1-ESC data. We downloaded peak calls for H3K27ac (ENCFF045CUG), H3K4me1 (ENCFF429INQ) and H3K4me3 (ENCFF277AOQ) from ENCODE.

Sequence conservation. The sequence conservation was obtained from 46-way phastCons tracks from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate/). The files were converted to bedgraph format for analysis purposes.

OCT4 and *GATA3* permutation tests. To perform the permutation tests we binned the interrogated region into 100bp bins. For each permutation we moved the last 100 bins to the front to preserve the feature structure of the genome. For the observed enrichment of in conserved sequences we computed the median conservation score of all base pairs per bin. For promoter, H3K27ac, H3K4me1 and H3K4me3 we computed whether there was an overlap.

3.4 Results

We ran RELICS, CRISPR-SURF, MAGeCK and MAUDE on the data from the screen for *CD69*. RELICS identified $K = 11$ functional sequences for *CD69* (Fig. 3.2). In contrast, MAGeCK predicted a large number ($K = 23$) of significant regions (FDR = 5%) while CRISPR-SURF predicted multiple large functional sequences at and downstream of the *CD69* promoter. MAUDE identified similar regions to RELICS with the addition of several regions between FS11 and FS2 which was not reported by any of the other methods (Fig. 3.2b).

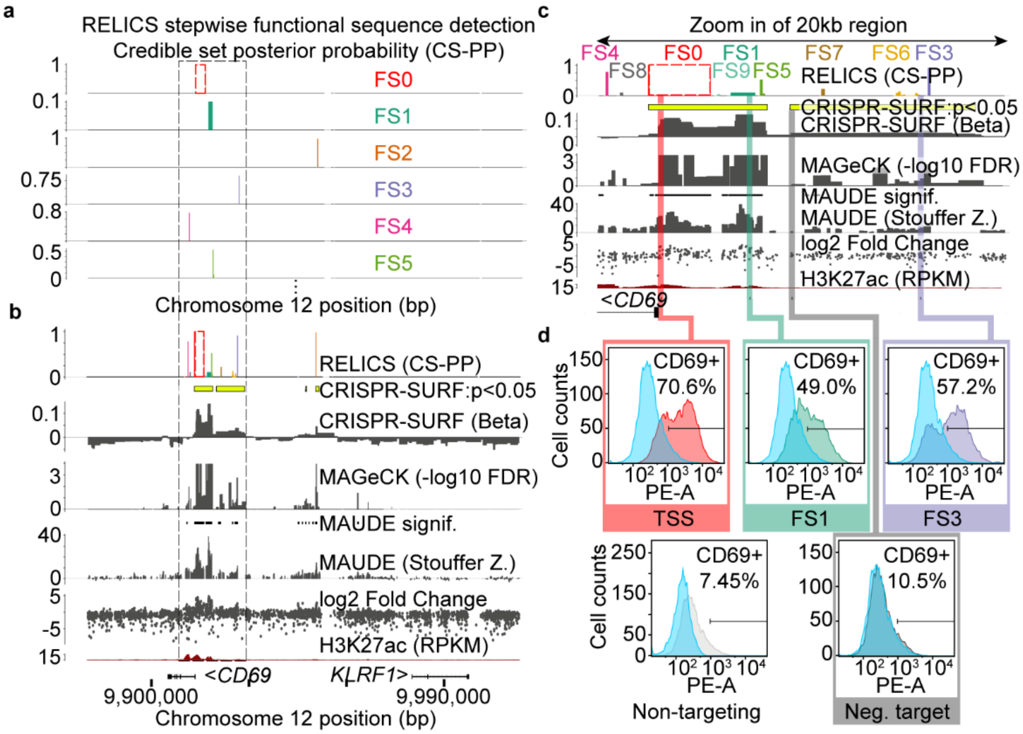


Figure 3.2. *CD69* screen analysis

Analysis of a published CRISPR activation (CRISPRa) screen for *CD69* expression in Jurkat T cells. **(a)** RELICS detects 11 functional sequences (FS) of which we show the first 5, labeled FS1-FS5. FS0 is a known positive control sequence (the *CD69* promoter) provided as input to RELICS and CRISPR-SURF. **(b)** Analysis of the *CD69* screen by RELICS, CRISPR-SURF, MAGeCK, MAUDE and log2 fold change of *CD69* expression. The RELICS credible set posterior probabilities for each functional sequence are collapsed into a single track (RELICS (CS-PP)). An H3K27ac ChIP-seq track for Jurkat cells is included (H3K27ac (RPKM)). **(c)** Zoom in of a 20kb region (indicated by dashed box in (a) and (b)). Experimentally validated regions are indicated by colored bars. **(d)** Experimental validation results. Lentiviruses carrying sgRNAs targeting different sites were transduced into Jurkat cells expressing dCas9:VP64 and *CD69* expression was measured by flow cytometry using PE-conjugated anti-human *CD69* antibody. The results from each experiment are overlaid atop those from a non-targeting negative control sgRNA (blue). sgRNAs were chosen for their high specificity and high predicted efficiency (relative to all possible sgRNAs in the target site region) and in some cases are adjacent to the predicted functional sequence rather than within the FS.

To test a subset of the *CD69* functional sequence predictions, we designed sgRNAs to target the putative functional sequences, cloned the sgRNAs into lentiviral vectors, and transduced them into Jurkat T cells expressing dCas9:VP64. We used a non-targeting sgRNA as a negative control and a sgRNA targeting a site near the *CD69* TSS as a positive control (Fig. 3.2c,d). The region including FS1 was previously reported to activate *CD69* expression ;we confirmed that targeting this region with CRISPRa increased the number of *CD69* expressing cells. Similarly, we validated that FS3 also increases *CD69* expression. All computational methods applied to this dataset detected a signal at FS2; however, we were unable to confirm that targeting this region affects *CD69* expression.

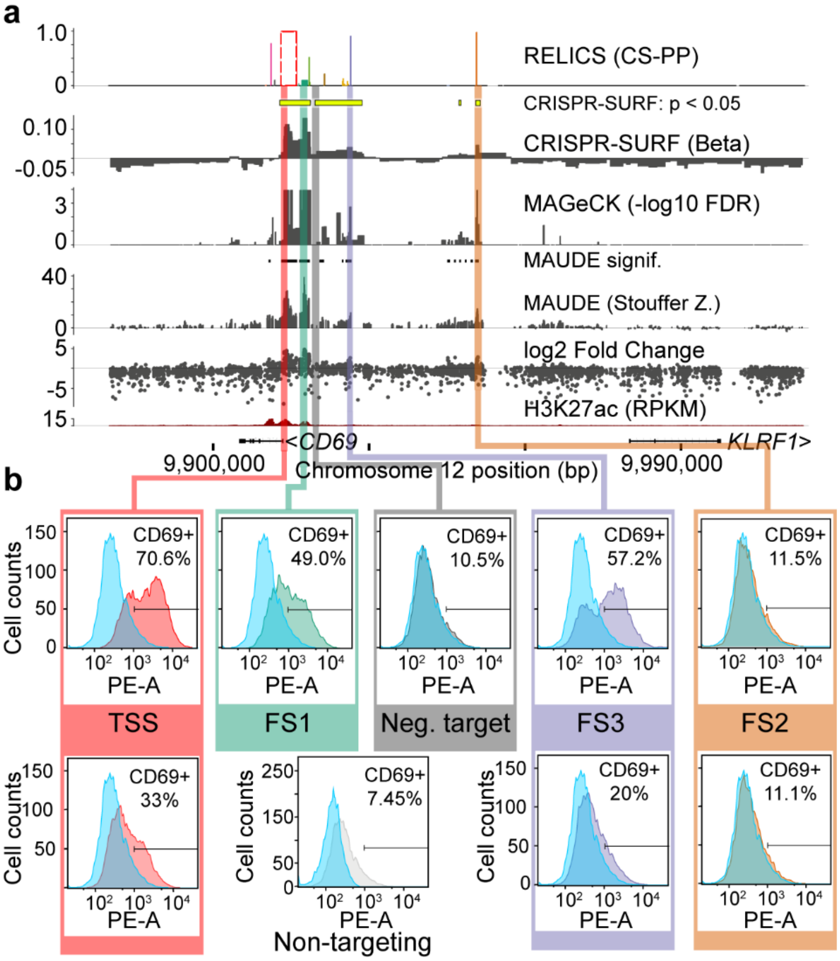


Figure 3.3. *CD69* screen analysis

Analysis of a published CRISPRa screen for *CD69* expression in Jurkat T cells. **(a)** Analysis of the *CD69* screen by RELICS, CRISPR-SURF, MAGeCK, MAUDE and log2 fold change. An H3K27ac ChIP-seq track for Jurkat cells is included (H3K27ac (RPKM)). **(b)** Experimental validation results. Lentiviruses carrying sgRNAs targeting different sites were transduced into Jurkat cells expressing dCas9:VP64, and *CD69* expression was measured by flow cytometry using PE-conjugated anti-human *CD69* antibody. The results from each experiment are overlaid atop those from a non-targeting negative control sgRNA (blue). sgRNAs were chosen for their high specificity and high predicted efficiency (relative to all possible sgRNAs in the target site region) and in some cases are adjacent to the predicted functional sequence rather than within the FS.

Finally, targeting a “negative sequence” that was not predicted by RELICS but is located on the edge of a large significant putative functional sequence reported by CRISPR-SURF had no effect on *CD69* expression. These results confirm that the predictions made by RELICS are accurate. Notably, the output from RELICS is easier to interpret compared to the output of other methods because RELICS provides ranked, cleanly-delineated predictions for each functional sequence (Fig. 3.3a).

Table 3.4. *CD69* CRISPRa validation sgRNAs

Target gene	region	gRNA	hg19 coordinate (chr12)
*CD69*_tss_sg1	TSS	CAATGTATAGTGTGTTGTTG	9913617-9913636
*CD69*_tss_sg2	TSS	TCAAGCAAGTAGGCGGCAAG	9913557 -9913576
*CD69*_FS1_sg2	FS1	AGGTAACCATGAGTAAACGG	9917833-9917852
*CD69*_TNC_sg1	‘Neg. Target’	CCATTTCCCTCCACAAGCCC	9919830-9919849
*CD69*_FS3_sg1	FS3	GCATAGAATTGATATCACCA	9925936-9925955
*CD69*_FS3_sg2	FS3	GGATTCGTCTTCTGAGTTAC	9925975-9925994
*CD69*_FS2_sg1	FS2	TATTCCTGCTGTATAACAGA	9950460-9950479
*CD69*_FS2_sg2	FS2	GTTAAATAATAGAGGGCACA	9950613-9950632
NTC		CTGAAAAAGGAAGGAGTTGA	

TSS sgRNAs target the transcription start sites of *CD69*. functional sequence and ‘Neg. Target’ sgRNAs correspond to labels in Fig. 4d. NTC is a non-targeting control sgRNA.

RELICS identified 17 functional sequences for *IL2RA*, of which all are located within or very close to the six regions identified in the original study (Fig 3.4). Since RELICS predictions are higher resolution than other methods, the multiple smaller regions predicted by RELICS may reflect the true presence of multiple functional sequences. Alternatively, some large functional sequences may be mistakenly split into smaller sequences by RELICS, particularly if some of the sgRNAs targeting the middle of the sequence have very low efficiency.

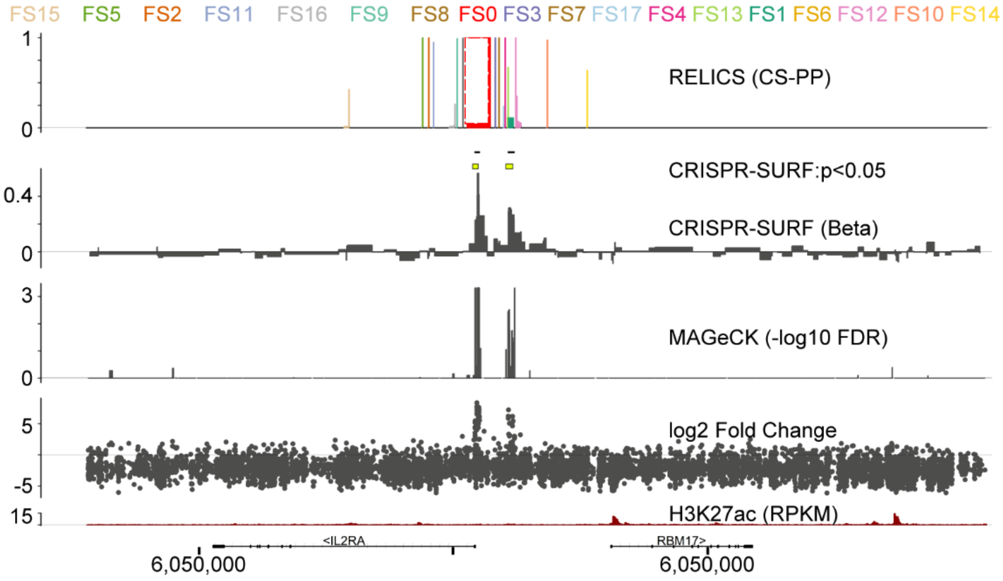


Figure 3.4. *IL2RA* screen analysis

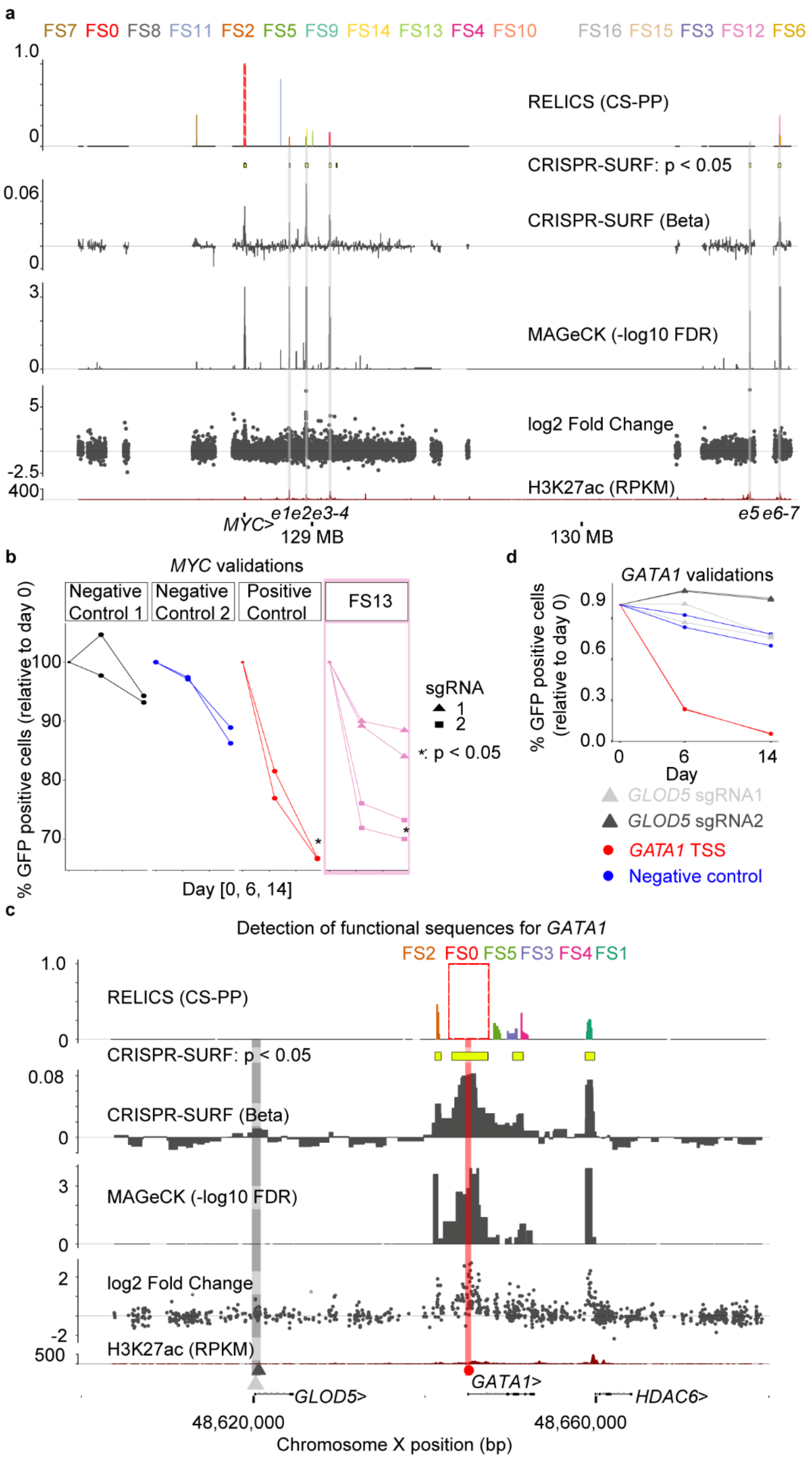
Analysis of a CRISPRa screen for *IL2RA* expression by Simeonov et al. 2017. Output of RELICS and other analysis methods. Each functional sequence predicted by RELICS is assigned a different color and the labels are ordered by genomic position.

Next, we applied RELICS to a CRISPRi proliferation screen surrounding the *MYC* locus in K562 cells (Fulco et al., 2016) (Fig. 3.5) and discovered 16 functional sequences. RELICS detected all of the previously reported signals but, as with *IL2RA*, split some of them into smaller putative functional sequences. Interestingly, RELICS also identified three regions that have not been previously reported and were undetected by both CRISPR-SURF and MAGeCK (FS7, FS11, FS13). To test whether these sequences are functional, we targeted FS13 with CRISPRi for validation. As a positive control we targeted a previously-detected functional sequence reported by Fulco et al. (referred to as e2). For negative controls, we used dCas9:KRAB alone or dCas9:KRAB with an sgRNA targeting a safe harbor sequence on chromosome 8 with no known function. Of the two guides we used to target FS13, one showed a substantial decrease in proliferation (specificity of 0.94) while the other one showed only a small decrease in proliferation (specificity of 0.21), suggesting that the latter guide might not have worked properly (Fig. 3.5b).

While additional validations are needed for the other functional sequences discovered by CRISPR, these results do confirm that RELICS has discovered novel functional sequences missed by other methods.

Figure 3.5. *MYC* and *GATA1* screen analysis

Analysis of a *MYC* CRISPRi cellular proliferation screen by Fulco et al. 2016. **(a)** Output of RELICS and other analysis methods. Each functional sequence predicted by RELICS is assigned a different color and the labels are ordered by genomic position. **(b)** Experimental validations. Each validation experiment is a cellular proliferation assay in which the percent of GFP-positive cells (i.e. cells that received the sgRNA) are measured at day 0, day 6 and day 14. As negative controls we used dCas9:KRAB alone (no sgRNA) as well as sgRNAs targeting a 'safe harbor' non-functional region on chromosome 8. As positive controls we used sgRNAs targeting a known regulatory region (e2, identified by Fulco et al., corresponding to FS2 and FS3). Each functional sequence was targeted with either one, two, or three different sgRNAs with two replicates each. sgRNAs that resulted in a significant reduction in % GFP compared to negative controls (dCas9 only, Neg. Ctrl.) at day 14 are denoted with an asterisk (Student's one-sided t-test, $p < 0.05$). **(c)** Results from RELICS and other analysis methods. RELICS detects five functional sequences (FS1-5). FS1 and FS2 have previously been validated; FS3, FS4 and FS5 fall within *GATA1*. **(d)** Results from validation experiments (2 replicates) using sgRNAs targeting sites indicated in panel (c). Each validation experiment is a cellular proliferation assay in which the percent of GFP-positive cells are measured at day 0, day 6 and day 14. While targeting the *GATA1* promoter greatly reduces proliferation, targeting the *GLOD5* region does not change proliferation compared to a negative control sgRNA, which targets a non-functional 'safe harbor' region on chromosome 8.


a
FS7 FS0 FS8 FS11 FS2 FS5 FS9 FS14 FS13 FS4 FS10 FS16 FS15 FS3 FS12 FS6
1.0
0
RELICS (CS-PP)
CRISPR-SURF: p < 0.05
0.06
0
CRISPR-SURF (Beta)
3
0
MAGeCK (-log10 FDR)
5
-2.5
log2 Fold Change
400
H3K27ac (RPKM)
MYC>
e1 e2 e3-4
129 MB
130 MB
e5 e6-7
b
MYC validations
Negative Control 1
Negative Control 2
Positive Control
FS13
% GFP positive cells (relative to day 0)
100
90
80
70
Day [0, 6, 14]
sgRNA
1
2
*: p < 0.05
*
*
d
GATA1 validations
% GFP positive cells (relative to day 0)
0.9
0.6
0.3
0.0
0
6
14
Day
GLOD5 sgRNA1
GLOD5 sgRNA2
GATA1 TSS
Negative control
c
Detection of functional sequences for GATA1
FS2 FS0 FS5 FS3 FS4 FS1
1.0
0
RELICS (CS-PP)
CRISPR-SURF: p < 0.05
0.08
0
CRISPR-SURF (Beta)
3
0
MAGeCK (-log10 FDR)
2
-2
log2 Fold Change
500
H3K27ac (RPKM)
GATA1>
GLOD5>
HDAC6>
48,620,000
48,660,000
Chromosome X position (bp)

The same study (Fulco et al., 2016) included a CRISPRi proliferation screen interrogating the region around *GATA1*. We applied RELICS to this dataset which predicted five functional sequences, two of which (FS1 and FS2) have been previously validated, and three of which fall within *GATA1* (FS3, FS4, FS5) (Fig. 3.5c). The original study also predicted a functional region near *GLOD5*. However, neither RELICS nor the other two analysis methods identified a functional sequence at this location. This suggests that the region near *GLOD5* may have been a false positive detected in the original study. To test this hypothesis, we used CRISPRi to target the *GLOD5* region, the *GATA1* promoter (positive control), and a sequence on chromosome 8 (negative control). While the sgRNAs targeting the *GATA1* promoter decreased cellular proliferation as expected, targeting of the *GLOD5* region did not affect proliferation relative to the negative control (Fig. 3.5d). Thus, this region is unlikely to be a functional sequence and it is appropriately not reported by RELICS.

Table 3.5. *GATA1* CRISPRi validation sgRNAs

% of GFP cells	**D0**	**D0**	**D6**	**D6**	**D14**	**D14**
NS	21.98	24.21	20.32	20.21	17.26	16.93
GATA1-tss	55.08	54.74	12.78	12.91	3.02	2.8
Glod5-1	20.88	19.06	18.18	19.19	15.86	14.76
Glod5-2	31.41	31.87	34.71	35.03	32.88	33.03
hg38	**chr**	**start**	**end**	**gRNA sequence**		
*GATA1*_TSS	chrX	48786605	48786625	GGTTCGGCCGCCTTGGGGATG		
Glod5-1	chrX	48761964	48761983	GCTTGTCTCTGAAAGAGAAA		
Glod5-2	chrX	48762201	48762220	GCTTTAGGAGAGGAATTCAG		
NS	chr8	128176215	128176234	GACAAGCTGCAAGGTGTAAAT		
hg19	**chr**	**start**	**end**	**gRNA sequence**		
*GATA1*_TSS	chrX	48645013	48645033	GGTTCGGCCGCCTTGGGGATG		
Glod5-1	chrX	48620368	48620387	GCTTGTCTCTGAAAGAGAAA		
Glod5-2	chrX	48620605	48620624	GCTTTAGGAGAGGAATTCAG		
NS	chr8	129188461	129188480	GACAAGCTGCAAGGTGTAAAT		

Coordinates are given in both hg38 and hg19 genome assemblies. NS is a negative control sgRNA targeting a safe harbor region on chromosome 8. *GATA1*_TSS sgRNA targets the transcription start site of *GATA1*. The Glod5-1 / Glod5-2 sgRNAs target the putative functional sequence near *GLOD5* identified by Fulco et al. 2016.

We proceeded to analyze the two currently publicly available dual guide screens. The first screen by Gasperini et al. aimed to identify functional sequences around the *HPRT1* locus using a selection screen with two replicates. In addition to a dual-guide screen, they also performed a sgRNA screen tiling single guides along the same locus. In their study they observed that all *HPRT1* exons contributed to survival, but no functional sequences were found. We analyzed the same data and similarly observed that the majority of the signal was around the exons (Fig. 3.6, 3.7). However, RELICS did identify three functional sequence in the dual-guide screen that were located in the intron between exon 1 and exon, suggesting that this region is important for *HPRT1* expression. Because the intronic signal was only observed in the dual-guide screen, it implies that the function of this region can only be repressed by larger deletions and is robust against smaller indels. However, further studies are necessary to determine the role of the region detected.

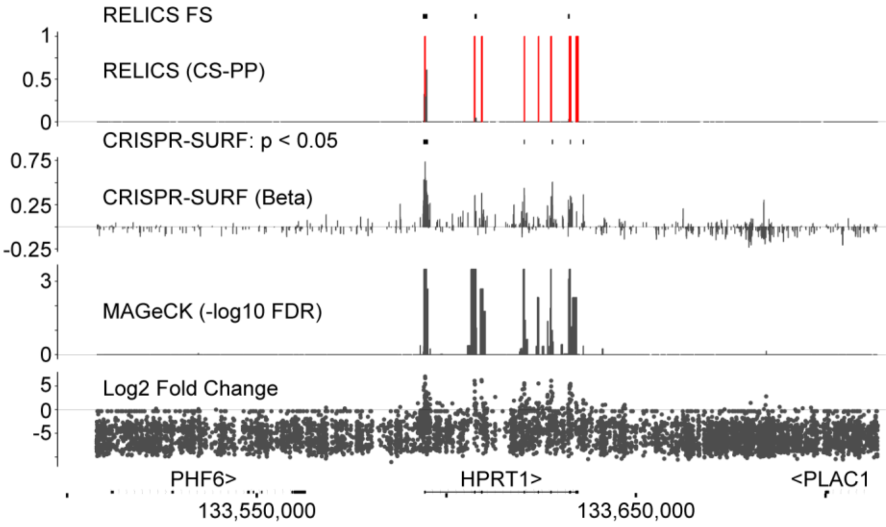


Figure 3.6. *HPRT1* single-guide screen analysis

Analysis of a single guide screen for *HPRT1* in HAP-1 cells by RELICS, CRISPR-SURF, MAGeCK and log2 fold change. All methods only identified the exons and areas immediately adjacent to the exons.

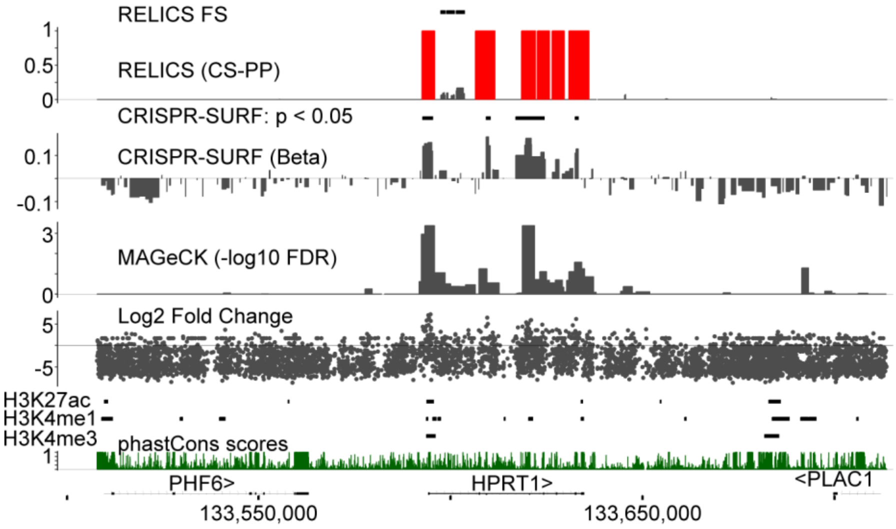


Figure 3.7. *HPRT1* dual-guide screen analysis

Analysis of a tiling deletion screen for *HPRT1* in HAP-1 cells by RELICS, CRISPR-SURF, MAGeCK and log2 fold change. Shown below are the ENCODE narrow peaks for H3K27ac, H3K4me1 and H3K4me3 as well as the 46-way phastCons conservation track. RELICS identifies 3 functional sequence all between exon 1 and exon 2.

The second dual-guide screen we analyzed used flow sorting to bin cells into pools of high and low expression to identify *OCT4*-specific functional sequences (Diao et al., 2017). The initial study included 45 reported putative elements (RPEs) (one overlapping with FS0). Diao et al. 2017 validated six of these, as well as a negative target. When analyzing their data with RELICS we discovered 51 FS, including five of the six previously validated sequences (Fig. 3.8a). The sixth validated sequence was located in a non-coding region next to FS36 which overlapped *HSPA1B*, and it is possible that that they are actually the same regulatory element. RELICS also did not label the negative-control region as significant. CRISPR-SURF only detected one region as significant, which overlaps both FS1 and an RPE. MAGeCK did not identify any regions passing a threshold of FDR < 0.05 apart from *OCT4*. Almost all of the 51 functional sequence detected overlapped a promoter region or an ENCODE peak for either H3K27ac, H3K4me1 or H3K4me3 (Fig. 3.8b). Using permutation testing, we confirmed that the 51 functional sequence identified by RELICS were enriched for overlapping promoters, all three epigenetic marks and high sequence conservation scores (Fig. 3.8c). Interestingly, we observed that the 18 RELICS-specific functional sequence were especially enriched for high sequence conservation as well as H3K27ac and H3K4me3, while this is not the case for RPE (Fig. 3.8d). Given the importance of these marks in the regulatory landscape, it is very likely that RELICS managed to detect additional functional sequences for *OCT4*.

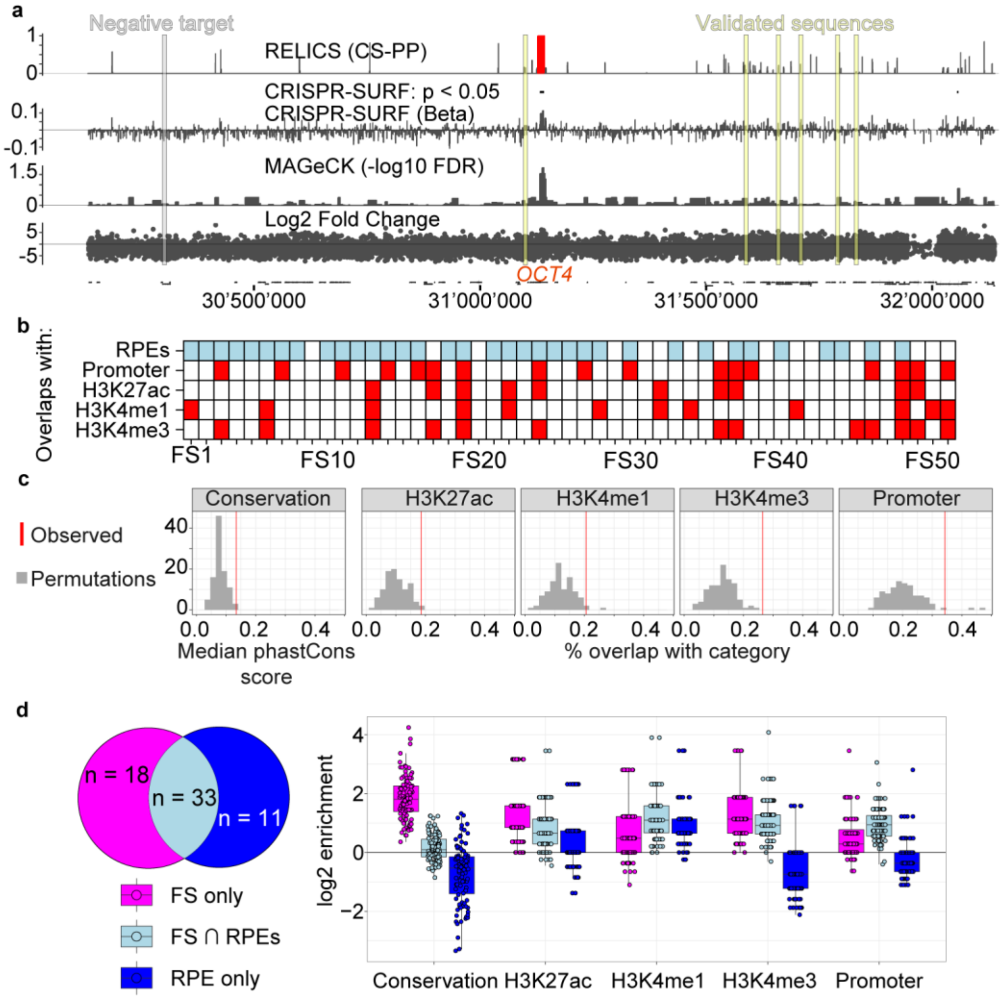


Figure 3.8. ***OCT4*** **dual-guide screen analysis**

Analysis of a dual guide screen for *OCT4* in H1-ESC by RELICS, CRISPR-SURF, MAGeCK and log2 fold change. **(a)** Results for each method. Highlighted in yellow are the 6 validated sequences. Negative targets are highlighted in grey **(b)** Heatmap of overlaps between functional sequences (FSs) identified by RELICS, the 45 RPEs reported by Diao et al. 2017, promoters of other genes, H3K27ac, H3K4me1 and H3K4me3. **(c)** Permutations for determining enrichment of the 51 functional sequence identified by RELICS. 100 permutation were made for each category. Conservation scores for each functional sequence were calculated from the average of the median 46-way phastCons. For all other features we report the percent of overlapping FSs. Vertical red lines indicate observed values. **(d)** Log2 enrichment of each subset of putative elements. For each subset, 100 permutations were performed to calculate the log2 enrichment relative to the observed value.

Lastly, we analyzed our in-house tiling deletion screen for *GATA3* using all pools of cells available. Log2 fold change between the high and low expression pools showed a slight enrichment of guides overlapping *GATA3* exons compared to background (Fig. 3.9a), indicating that using

GATA3 exons for FS0 is appropriate. Bootstrap sampling the data for FS0 confirmed that the RELICS estimates where withing the expected range. We compared RELICS hyperparameter estimates (black dots, Fig. 3.9b) against estimates from the bootstrap samples and observed that both the low and the medium-low pools were enriched for guides reducing *GATA3* expression. This suggests that increasing the number of pools provides additional information which can be leveraged by RELICS to produce higher resolution findings.

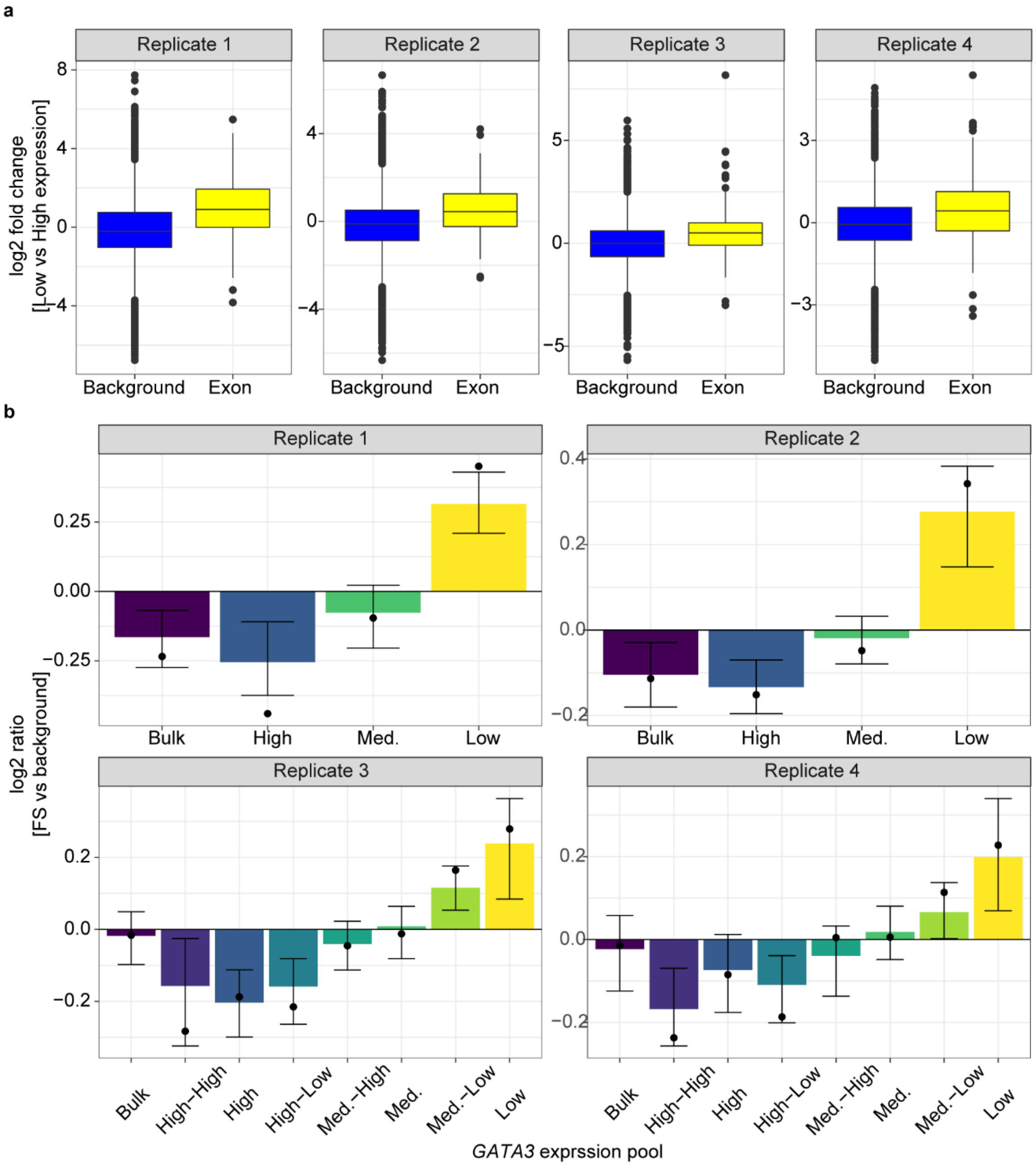


Figure 3.9. *GATA3* data quality

(a) Log2 fold change of the number of guides overlapping the *GATA3* exons and all other guides (background) between the of low and high expression pools. **(b)** Log2 ratio of RELICS hyperparameter estimates for functional sequence vs. background. Ratio for guides overlapping *GATA3* exons are shown with black dots; 95% bootstrap confidence interval from 100 bootstrap iterations shown with bars.

RELICS discovered 48 FS, most of which were downstream of *GATA3* with a majority of them landing within the topological associated domain (TAD) of *GATA3* (Fig. 3.10).

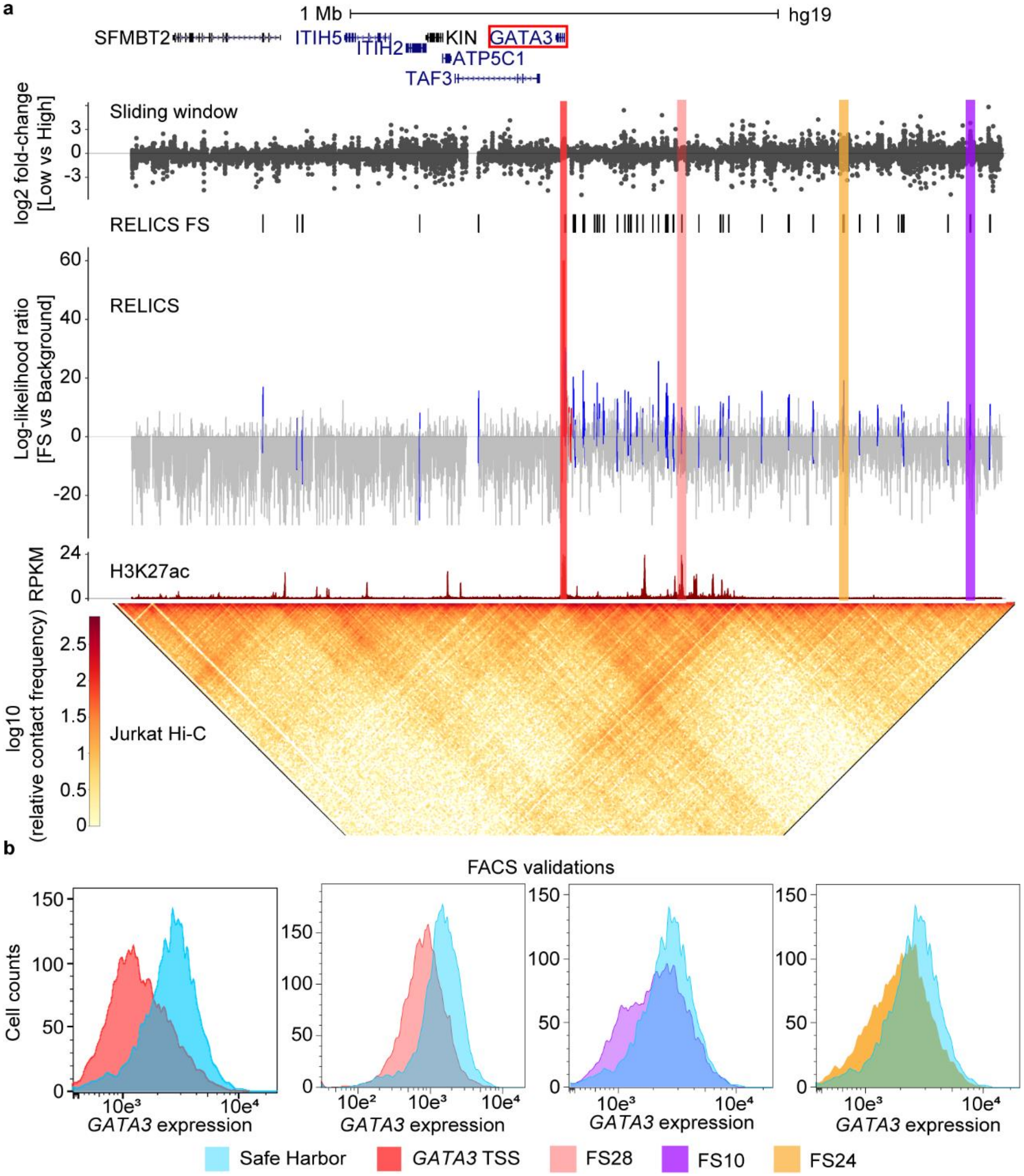


Figure 3.10. ***GATA3*** **dual-guide screen analysis**

(a) *GATA3* locus targeted by tiling deletion screen. Sliding window average of the log2 fold change across replicates gives an overview of the data. RELICS functional sequence locations are plotted over the log-likelihood ratio of functional sequence vs. background (i.e. likelihood that sequence is functional). The canonical enhancer mark H3K27ac is shown above the Hi-C map, which shows genome interactions. Lighter colors indicate a higher number of interactions. **(b)** Targeting the *GATA3* locus for validation decreases *GATA3* expression. Similar effects are seen when targeting FS28, which resembles a canonical enhancer. Surprisingly, targeting both unmarked regions FS10 and FS24 for validations also resulted in decreased *GATA3* expression.

While some functional sequence resembled canonical enhancers with high H3K27ac signal, others landed in regions completely devoid of signals characteristic of enhancers and outside of the *GATA3* TAD. Performing a permutation analysis did not show any enrichment in high phastCons conservation scores, open chromatin or H3K27ac (Fig. 3.11).

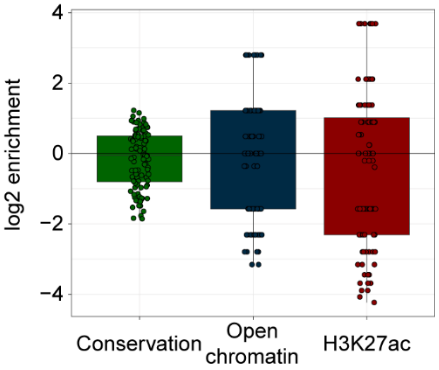


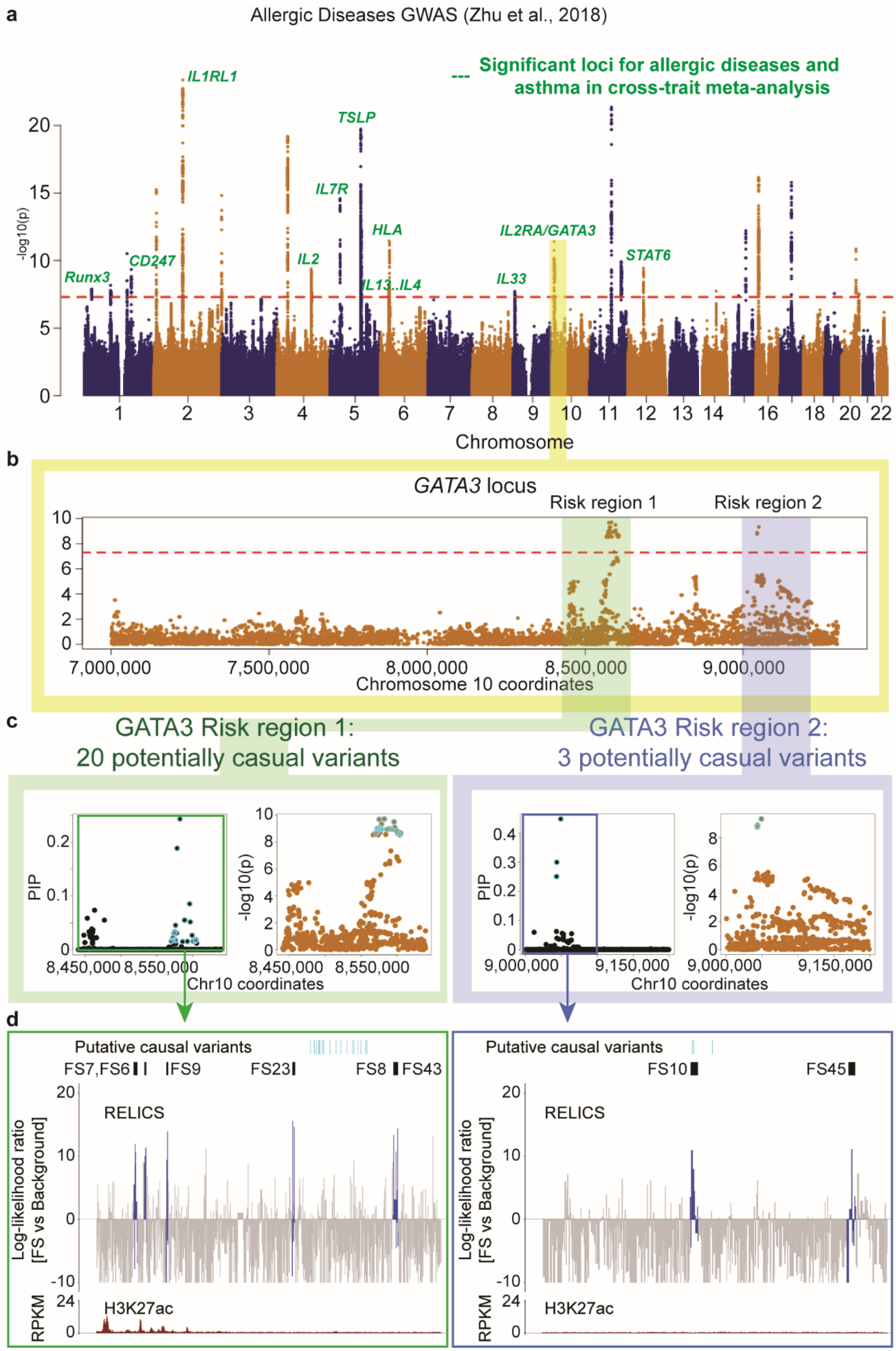
Figure 3.11. *GATA3* functional sequence enrichment
Log2 enrichment of functional sequences in high phastCons scores (Conservation), open chromatin delineated by DNase hypersensitive peaks, and H3K27ac peaks. For each category, 100 permutations were performed to calculate the log2 enrichment relative to the observed value.

We followed up on several regions identified by RELICS, including one resembling canonical enhancers (FS28) and two that were unmarked and located outside of the *GATA3* TAD (FS10, FS24). In all three cases, we confirmed that they have a significantly reduce *GATA3* expression, thereby confirming the RELICS findings.

Furthermore, we discovered that FS10 also contained two out of three putative causal SNPs from a genome wide association study (GWAS) for allergic diseases and asthma (Z. Zhu et al., 2018) (Fig. 3.12), which informed our hypothesis that FS10 regulates *GATA3* and disruption of the sequence can lead to immune disorders via reduced *GATA3* expression.

Figure 3.12. *GATA3* FS10 overlaps putative causal variants for allergic diseases

(a) Allergic diseases GWAS by Zhu et al. The *GATA3* locus is highlighted in yellow. **(b)** Tiled deletion screen for *GATA3* contains two risk regions. **(c)** Fine mapping the 200kb risk loci with SuSiE reveals 20 potentially causal variants for risk region 1 and 3 potentially causal variants for risk region 2 (cyan circles) (PIP = Posterior Inclusion Probability computed by SuSiE). **(d)** Two putative causal variants from risk region 1 overlap FS10.


a
Allergic Diseases GWAS (Zhu et al., 2018)
--- Significant loci for allergic diseases and asthma in cross-trait meta-analysis
IL1RL1
TSLP
IL7R
HLA
IL2RA/GATA3
STAT6
Runx3
CD247
IL2
IL13..IL4
IL33
-log10(p)
0
5
10
15
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22
Chromosome
b
GATA3 locus
Risk region 1
Risk region 2
0
2
4
6
8
10
7,000,000
7,500,000
8,000,000
8,500,000
9,000,000
Chromosome 10 coordinates
c
GATA3 Risk region 1:
20 potentially casual variants
GATA3 Risk region 2:
3 potentially casual variants
PIP
0
0.1
0.2
-log10(p)
0
2
4
6
8
10
8,450,000 8,550,000
8,450,000 8,550,000
Chr10 coordinates
PIP
0
0.1
0.2
0.3
0.4
-log10(p)
0
2
4
6
8
9,000,000
9,150,000
9,000,000
9,150,000
Chr10 coordinates
d
Putative causal variants
FS7,FS6
FS9
FS23
FS8
FS43
RELICS
Log-likelihood ratio
[FS vs Background]
20
10
0
-10
RPKM
24
0
H3K27ac
Putative causal variants
FS10
FS45
RELICS
Log-likelihood ratio
[FS vs Background]
20
10
0
-10
RPKM
24
0
H3K27ac

3.5 Discussion

In chapter 2 I demonstrated that RELICS outperformed all other comparable method when used to analyze simulated CRISPR screen data. Here I have shown that RELICS also outperforms other methods in the analysis of experimental data. Across the publicly available data sets, RELICS identified all functional sequences reported in the original studies with the exception of one RPE for *OCT4* reported by Diao et al. 2017. However, for this case it is likely that RELICS picked up the same signal but at a different location. Furthermore, RELICS correctly did not report two experimentally validated negative sequences as well as a negative region tested in the *OCT4* screen. RELICS also discovered additional putative functional sequences that we confirmed with experimental validations but which were not detected by any of the other methods. Lastly, RELICS identified 18 new functional sequences in the *OCT4* screen which were enriched for high sequence conservation scores and H3K27ac, indicating that these sequences may be regulators of *OCT4*. When analyzing our in house tiling deletion screen for *GATA3,* we detected many functional sequences that did not resemble canonical functional sequences. However, our validations demonstrate that both marked and unmarked regions affect *GATA3* expression. Interestingly, we found that FS10 overlapped 2 out of 3 SNPs credible risk SNPs for asthma and allergic diseases. This suggests that variants in FS10 repress *GATA3* expression and increase the risk for immune-related diseases. None of the functional sequences identified by RELICS for a second region in the *GATA3* locus overlapped any putative causal variants. This could be due to false negatives which our tiling deletion screen did not detect. However, it is also possible that the variants in this region do not act on allergic diseases via the expression of *GATA3*. This would be consistent with studies reporting that functional sequences do not always regulate the closest gene (Visel et al., 2009).

In summary, RELICS, discovers both previously reported functional sequences as well as novel ones. Thus, RELICS is an extremely useful tool for mapping the regulatory landscape using CRISPR screens.

3.6 Data Availability

All results above can be found at https://figshare.com/projects/RELICS_2_data/74376 and https://figshare.com/projects/RELICS_data/98219.

3.7 Acknowledgements

This research was supported by NIH/NIAID grant 2R01AI107027-06; by NIH/NIDDK grant 1 R01 DK122607-01; by the National Cancer Institute funded Salk Institute Cancer Center (NIH/NCI CCSG: 2 P30 014195); by a gift from the Jacobs Foundation; by a fellowship from the H.A. and Mary K. Chapman Charitable Trust to P.C.F.; by a fellowship from the Jesse and Caryl Philips Foundation to P.C.F; by a Salk Alumni Fellowship to H.V.C; and by the Frederick B. Rentschler Developmental Chair to G.M.

We thank Yarui Diao, Rongxin Fang, Ye Zheng, Zhi Liu, and members of the McVicker lab for helpful discussions about analysis methods for CRISPR regulatory screens; Arko Sen for guidance in obtaining and processing public datasets; Bing Ren, for his laboratory's assistance with CRISPRi validation experiments; and Jessica Zhou and Arya Massarat for testing RELICS and CRSsim.

3.8 Author Contributions

G.M. supervised the research. G.M. and P.C.F. wrote the two manuscript describing RELICS, with input and edits from H.V.C. P.C.F. obtained the public data sets from Fulco et al., 2016, Simeonov et al., 2017, and analyzed them. H.V.C. and A.R.C. performed CRISPRa validation experiments in Jurkat T cells. P.B.C. performed CRISPRi validation experiments in K562 cells. I.L. and H.V.C. designed the *GATA3* guides. H.V.C. performed the *GATA3* tiling deletion screen. P.C.F. processed the *GATA3* data and analyzed it. H.V.C. obtained the GWAS data and performed fine mapping.

Chapter 3, in part, is adapted from the material as it appears as “Discovering functional sequences with RELICS, an analysis method for CRISPR screens” in *PLOS Computational Biology*, 2020 by Patrick C. Fiaux, Hsiuyi V. Chen, Poshen B. Chen, Aaron R. Chen, Graham McVicker, in part, adapted from the manuscript in preparation “Modeling of dispersion and perturbation effects in tiling CRISPR screens with RELICS+” by Patrick C. Fiaux, Karthik Guruvayurappan, Graham McVicker, and in part, adapted from the manuscript in preparation “Discovery of novel, unmarked *GATA3* functional sequences” by Hsiuyi V. Chen, Patrick C. Fiaux, Aaron R. Chen, Ishika Luthra, Graham McVicker. The dissertation author was a primary investigator and author of these paper.

References

Abraham, B. J., Hnisz, D., Weintraub, A. S., Kwiatkowski, N., Li, C. H., Li, Z., Weichert-Leahey, N., Rahman, S., Liu, Y., Etchin, J., Li, B., Shen, S., Lee, T. I., Zhang, J., Look, A. T., Mansour, M. R., & Young, R. A. (2017). Small genomic insertions form enhancers that misregulate oncogenes. *Nat. Commun.*, *8*, 14385.

Adiconis, X., Haber, A. L., Simmons, S. K., Levy Moonshine, A., Ji, Z., Busby, M. A., Shi, X., Jacques, J., Lancaster, M. A., Pan, J. Q., Regev, A., & Levin, J. Z. (2018). Comprehensive comparative analysis of 5'-end RNA-sequencing methods. *Nat. Methods*. https://doi.org/10.1038/s41592-018-0014-2

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891

Allen, F., Behan, F., Khodak, A., Iorio, F., Yusa, K., Garnett, M., & Parts, L. (2019). JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.*, *29*(3), 464–471. https://doi.org/10.1101/gr.238923.118

Altman, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *J. Am. Stat. Assoc.*, *102*(477), 201–210.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: Limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, *16*(1), 47–57. https://doi.org/10.1016/j.devcel.2008.11.011

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays*, *37*(3), 314–323. https://doi.org/10.1002/bies.201400162

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. https://doi.org/10.1038/nature12787

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, *339*(6123), 1074–1077. https://doi.org/10.1126/science.1232542

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti,

R. C., Kadaveru, K., Mazutis, L., Rudensky, A. Y., & Pe’er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, *174*(5), 1293-1308.e36.

Baggerly, K. A., Deng, L., Morris, J. S., & Aldaz, C. M. (2003). Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics*, *19*(12), 1477–1483. https://doi.org/10.1093/bioinformatics/btg173

Baggerly, K. A., Deng, L., Morris, J. S., & Aldaz, C. M. (2004). Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics*, *5*, 144. https://doi.org/10.1186/1471-2105-5-144

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308.

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., & Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, *10*(9), e1004663. https://doi.org/10.1371/journal.pgen.1004663

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709–1712. https://doi.org/10.1126/science.1138140

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., Shao, Z., Canver, M. C., Smith, E. C., Pinello, L., Sabo, P. J., Vierstra, J., Voit, R. A., Yuan, G.-C., Porteus, M. H., Stamatoyannopoulos, J. A., Lettre, G., & Orkin, S. H. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science*, *342*(6155), 253–257. https://doi.org/10.1126/science.1242088

Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C. A., Chotalia, M., Xie, S. Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M. R., Fraser, J., Dostie, J., Game, L., Dillon, N., Edwards, P. A. W., Nicodemi, M., & Pombo, A. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*. https://doi.org/10.1038/nature21411

Beer, M. A. (2017). Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.*, *38*(9), 1251–1258.

Beer, M. A., & Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, *117*(2), 185–198.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, 289–300.

Benton, M. L., Talipineni, S. C., Kostka, D., & Capra, J. A. (2017). Genome-wide Enhancer Maps Differ Significantly in Genomic Distribution, Evolution, and Function. In *BioRxiv*. https://doi.org/10.1101/176610

Berisa, T., & Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, *32*(2), 283–285. https://doi.org/10.1093/bioinformatics/btv546

Bodapati, S., Daley, T. P., Lin, X., Zou, J., & Qi, L. S. (2020). A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biol.*, *21*(1), 62. https://doi.org/10.1186/s13059-020-01972-x

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, *151*(Pt 8), 2551–2561. https://doi.org/10.1099/mic.0.28048-0

Bonnal, R. J. P., Ranzani, V., Arrigoni, A., Curti, S., Panzeri, I., Gruarin, P., Abrignani, S., Rossetti, G., & Pagani, M. (2015). De novo transcriptome profiling of highly purified human lymphocytes primary cells. *Sci Data*, *2*, 150051. https://doi.org/10.1038/sdata.2015.51

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Boyle, E. A., Pritchard, J. K., & Greenleaf, W. J. (2018). High-resolution mapping of cancer cell networks using co-functional interactions. In *BioRxiv*. https://doi.org/10.1101/369751

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.*, *9*(2), 378–400. https://doi.org/10.3929/ethz-b-000240890

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., & van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, *321*(5891), 960–964.

Buecker, C., & Wysocka, J. (2012). Enhancers as information integration hubs in development: Lessons from genomics. *Trends Genet.*, *28*(6), 276–284. https://doi.org/10.1016/j.tig.2012.02.008

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, *109*, 21.29.1-9. https://doi.org/10.1002/0471142727.mb2129s109

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, *144*(3), 327–339. https://doi.org/10.1016/j.cell.2011.01.024

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM J. Sci. Comput.*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Campbell, K. R., & Yau, C. (2017). Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers. *Wellcome Open Res*, *2*, 19. https://doi.org/10.12688/wellcomeopenres.11087.1

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H., & Bauer, D. E. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, *527*(7577), 192–197. https://doi.org/10.1038/nature15521

Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M. T. S., Cheng, C., Fan, X., Gerstein, M., Cheng, A. S. L., & Yip, K. Y. (2017). Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.*, *49*, 1428. https://doi.org/10.1038/ng.3950

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biol.*, *16*, 195. https://doi.org/10.1186/s13059-015-0762-6

Chae, M., Danko, C. G., & Kraus, W. L. (2015). GroHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics*, *16*, 222. https://doi.org/10.1186/s12859-015-0656-3

Chen, L., Liu, P., Evans, T. C., & Ettwiller, L. M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, *355*(6326), 752–756. https://doi.org/10.1126/science.aai8690

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., & Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*, *168*(3), 442-459.e20. https://doi.org/10.1016/j.cell.2016.12.016

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P.-A., Hsu, Y.-H., Drucker, D. J., Mellgren, G., Hui, C.-C., Hauner, H., & Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.*, *373*(10), 895–907. https://doi.org/10.1056/NEJMoa1502214

Cohen, A. J., Saiakhova, A., Corradin, O., Luppino, J. M., Lovrenert, K., Bartels, C. F., Morrow, J. J., Mack, S. C., Dhillon, G., Beard, L., Myeroff, L., Kalady, M. F., Willis, J., Bradner, J. E., Keri, R. A., Berger, N. A., Pruett-Miller, S. M., Markowitz, S. D., & Scacheri, P. C. (2017). Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat. Commun.*, *8*, 14400. https://doi.org/10.1038/ncomms14400

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science*, *339*(6121), 819–823. https://doi.org/10.1126/science.1231143

Cook, K. D., & Miller, J. (2010). TCR-dependent translational control of GATA-3 enhances Th2 differentiation. *J. Immunol.*, *185*(6), 3209–3216.

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A. A., Berman, B. P., Staudt, L. M., … Chang, H. Y. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, *362*(6413).

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, *322*(5909), 1845–1848. https://doi.org/10.1126/science.1162228

Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G., & Collins, F. S. (2006). DNase-chip: A high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods*, *3*(7), 503–509. https://doi.org/10.1038/nmeth888

CRISPR and Beyond: Perturbations at Scale to Understand Genomes—P9—Discovery of novel unmarked regulatory sequences with RELICS, an analysis method for tiling CRISPR screens. (n.d.). Retrieved September 23, 2020, from https://stream.venue-av.com/e/CRISPR/poster/245

Cui, J., Yao, Q., Li, S., Ding, X., Lu, Q., Mao, H., Liu, L., Zheng, N., Chen, S., & Shao, F. (2010). Glutamine deamidation and dysfunction of ubiquitin/NEDD8 induced by a bacterial effector family. *Science*, *329*(5996), 1215–1218. https://doi.org/10.1126/science.1193844

Daley, T. P., Lin, Z., Lin, X., Liu, Y., Wong, W. H., & Qi, L. S. (2018). CRISPhieRmix: A hierarchical mixture model for CRISPR pooled screens. *Genome Biol.*, *19*(1), 159. https://doi.org/10.1186/s13059-018-1538-6

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., van Helden, J., Puthier, D., & Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* https://doi.org/10.1038/ng.3884

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, *14*(3), 297–301. https://doi.org/10.1038/nmeth.4177

Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., & Aerts, S. (2015). Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.*, *11*(2), e1004994. https://doi.org/10.1371/journal.pgen.1004994

Dawson, M. A., & Kouzarides, T. (2012). Cancer epigenetics: From mechanism to therapy. *Cell*, *150*(1), 12–27. https://doi.org/10.1016/j.cell.2012.06.013

de Boer, C. G., Ray, J. P., Hacohen, N., & Regev, A. (2020). MAUDE: inferring expression changes in sorting-based CRISPR screens. *Genome Biol.*, *21*(1), 134. https://doi.org/10.1186/s13059-020-02046-8

de Kleer, I. M., Kool, M., de Bruijn, M. J. W., Willart, M., van Moorleghem, J., Schuijs, M. J., Plantinga, M., Beyaert, R., Hams, E., Fallon, P. G., Hammad, H., Hendriks, R. W., & Lambrecht, B. N. (2016). Perinatal Activation of the Interleukin-33 Pathway Promotes Type 2 Immunity in the Developing Lung. *Immunity*, *45*(6), 1285–1298. https://doi.org/10.1016/j.immuni.2016.10.031

De Obaldia, M. E., & Bhandoola, A. (2015). Transcriptional regulation of innate and adaptive lymphocyte lineages. *Annu. Rev. Immunol.*, *33*, 607–642.

De Ravin, S. S., Li, L., Wu, X., Choi, U., Allen, C., Koontz, S., Lee, J., Theobald-Whiting, N., Chu, J., Garofalo, M., Sweeney, C., Kardava, L., Moir, S., Viley, A., Natarajan, P., Su, L., Kuhns, D., Zarember, K. A., Peshwa, M. V., & Malech, H. L. (2017). CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci. Transl. Med.*, *9*(372). https://doi.org/10.1126/scitranslmed.aah3480

de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes Dev.*, *26*(1), 11–24.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., & Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, *482*(7385), 390–394. https://doi.org/10.1038/nature10808

Dekker, J., & Misteli, T. (2015). Long-Range Chromatin Interactions. *Cold Spring Harb. Perspect. Biol.*, *7*(10), a019356. https://doi.org/10.1101/cshperspect.a019356

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J., & Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607. https://doi.org/10.1038/nature09886

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K.-L., & Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*.

Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y., & Ren, B. (2016). A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res.*, *26*(3), 397–405.

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., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A., & Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. https://doi.org/10.1038/nature14222

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082

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., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, *34*(2), 184–191. https://doi.org/10.1038/nbt.3437

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., & Root, D. E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, *32*(12), 1262–1267. https://doi.org/10.1038/nbt.3026

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, *16*(10), 1299–1309. https://doi.org/10.1101/gr.5571506

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096. https://doi.org/10.1126/science.1258096

Du, D., Roguev, A., Gordon, D. E., Chen, M., Chen, S.-H., Shales, M., Shen, J. P., Ideker, T., Mali, P., Qi, L. S., & Krogan, N. J. (2017). Genetic interaction mapping in mammalian cells using CRISPR interference. *Nat. Methods*, *14*(6), 577–580. https://doi.org/10.1038/nmeth.4286

Ehlers, M. D. (2016). Lessons from a Recovering Academic. *Cell*, *165*(5), 1043–1048. https://doi.org/10.1016/j.cell.2016.05.005

Emanuelsson, O., Nagalakshmi, U., Zheng, D., Rozowsky, J. S., Urban, A. E., Du, J., Lian, Z., Stolc, V., Weissman, S., Snyder, M., & Gerstein, M. B. (2007). Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res.*, *17*(6), 886–897. https://doi.org/10.1101/gr.5014606

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, *28*(8), 817–825. https://doi.org/10.1038/nbt.1662

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, *9*(3), 215–216. https://doi.org/10.1038/nmeth.1906

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., & Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.*, *34*(11), 1180–1190. https://doi.org/10.1038/nbt.3678

Fan, X., Wang, D., Burgmaier, J. E., Teng, Y., Romano, R.-A., Sinha, S., & Yi, R. (2018). Single Cell and Open Chromatin Analysis Reveals Molecular Origin of Epidermal Cells of the Skin. *Dev. Cell*.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343. https://doi.org/10.1038/nature13835

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., & Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science*, *350*(6258), 325–328.

Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S., & Levine, M. S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(23), 6508–6513.

Fiaux, P. C., Chen, H. V., Chen, P. B., Chen, A. R., & McVicker, G. (2020). Discovering functional sequences with RELICS, an analysis method for CRISPR screens. *PLoS Comput. Biol.*, *16*(9), e1008194. https://doi.org/10.1371/journal.pcbi.1008194

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L., & Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*(7584), 110–114. https://doi.org/10.1038/nature16490

Flutre, T., Wen, X., Pritchard, J., & Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.*, *9*(5), e1003486.

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: Simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, *31*(17), 2778–2784. https://doi.org/10.1093/bioinformatics/btv272

Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., & Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, *354*(6313), 769–773.

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, *51*(12), 1664–1669. https://doi.org/10.1038/s41588-019-0538-0

Gaffney, D. J., Veyrieras, J.-B., Degner, J. F., Pique-Regi, R., Pai, A. A., Crawford, G. E., Stephens, M., Gilad, Y., & Pritchard, J. K. (2012). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.*, *13*(1), R7. https://doi.org/10.1186/gb-2012-13-1-r7

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., Quan, J., GTEx Consortium, Nicolae, D. L., Eskin, E., Kellis, M., Getz, G., McCarthy, M. I., Dermitzakis, E. T., Cox, N. J., & Ardlie, K. G. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.*, *50*(7), 956–967.

García-Martínez, J., Aranda, A., & Pérez-Ortín, J. E. (2004). Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol. Cell*, *15*(2), 303–313. https://doi.org/10.1016/j.molcel.2004.06.004

Gasperini, M., Findlay, G. M., McKenna, A., Milbank, J. H., Lee, C., Zhang, M. D., Cusanovich, D. A., & Shendure, J. (n.d.). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* https://doi.org/10.1016/j.ajhg.2017.06.010

Gasperini, M., Hill, A., McFaline-Figueroa, J. L., Martin, B., Trapnell, C., Ahituv, N., & Shendure, J. (2018). CrisprQTL mapping as a genome-wide association framework for cellular genetic screens. In *BioRxiv*. https://doi.org/10.1101/314344

Gate, R E, Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., & others. (2018). Genetic determinants of co-accessible chromatin regions in T cell activation across humans. *BioRxiv*. https://www.biorxiv.org/content/early/2018/02/07/090241.abstract

Gate, Rachel E, Cheng, C. S., Aiden, A. P., Siba, A., Tabaka, M., Lituiev, D., Machol, I., Gordon, M. G., Subramaniam, M., Shamim, M., Hougen, K. L., Wortman, I., Huang, S.-C., Durand, N. C., Feng, T., De Jager, P. L., Chang, H. Y., Aiden, E. L., Benoist, C., … Regev, A. (2018). Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.*, *50*(8), 1140–1150. https://doi.org/10.1038/s41588-018-0156-2

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://market.android.com/details?id=book-lV3DIdV0F9AC

Genest, C., & Mackay, J. (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *Am. Stat.*, *40*(4), 280–283. https://doi.org/10.1080/00031305.1986.10475414

Gifford, C. A., Ziller, M. J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A. K., Kelley, D. R., Shishkin, A. A., Issner, R., Zhang, X., Coyne, M., Fostel, J. L., Holmes, L., Meldrim, J., Guttman, M., Epstein, C., Park, H., Kohlbacher, O., … Meissner, A. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, *153*(5), 1149–1163.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M., & Weissman, J. S. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, *159*(3), 647–661. https://doi.org/10.1016/j.cell.2014.09.029

Gioia, L., Siddique, A., Head, S. R., Salomon, D. R., & Su, A. I. (2018). A genome-wide survey of mutations in the Jurkat cell line. *BMC Genomics*, *19*(1), 334.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, *17*(6), 877–885. https://doi.org/10.1101/gr.5533506

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., & Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, *518*(7539), 365–369. https://doi.org/10.1038/nature14252

Gong, L., Wong, C.-H., Cheng, W.-C., Tjong, H., Menghi, F., Ngan, C. Y., Liu, E. T., & Wei, C.-L. (2018). Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat. Methods*, *15*(6), 455–460. https://doi.org/10.1038/s41592-018-0002-6

Gorkin, D. U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S. L., Loftus, S. K., Beer, M. A., Pavan, W. J., & McCallion, A. S. (2012). Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.*, *22*(11), 2290–2301. https://doi.org/10.1101/gr.139360.112

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., Greenside, P., Srivas, R., Phanstiel, D. H., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J. K., Bustamante, C. D., Steinmetz, L. M., Kundaje, A., & Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, *162*(5), 1051–1065. https://doi.org/10.1016/j.cell.2015.07.048

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., Yuan, Y., Shi, J., Sun, Y., Han, M.-R., Wang, J., Liu, Q., Wen, W., Li, B., Long, J., Chen, J., & Zheng, W. (2018). A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *Am. J. Hum. Genet.*, *102*(5), 890–903.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, *48*(3), 245–252. https://doi.org/10.1038/ng.3506

Haapaniemi, E., Botla, S., Persson, J., Schmierer, B., & Taipale, J. (2018). CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* https://doi.org/10.1038/s41591-018-0049-z

Haarhuis, J. H. I., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., Krijger, P. H. L., Teunissen, H., Medema, R. H., van Steensel, B., Brummelkamp, T. R., de Wit, E., & Rowland, B. D. (2017). The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell*, *169*(4), 693-707.e14.

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/s41580-018-0028-8

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M., & Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, *139*(5), 945–956. https://doi.org/10.1016/j.cell.2009.07.040

Hart, T., & Moffat, J. (2016). BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics*, *17*, 164. https://doi.org/10.1186/s12859-016-1015-8

Harting, M. T., Cox, C. S., Day, M.-C., Walker, P., Gee, A., Brenneman, M. M., Grotta, J. C., & Savitz, S. I. (2009). Bone marrow-derived mononuclear cell populations in pediatric and adult patients. *Cytotherapy*, *11*(4), 480–484. https://doi.org/10.1080/14653240902960452

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., Wen, X., Luca, F., & Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics*, *31*(8), 1235–1242. https://doi.org/10.1093/bioinformatics/btu802

Hawkins, R. D., Larjo, A., Tripathi, S. K., Wagner, U., Luu, Y., Lönnberg, T., Raghav, S. K., Lee, L. K., Lund, R., Ren, B., Lähdesmäki, H., & Lahesmaa, R. (2013). Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity*, *38*(6), 1271–1284. https://doi.org/10.1016/j.immuni.2013.05.011

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., & Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, *39*(3), 311–318. https://doi.org/10.1038/ng1966

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, *38*(4), 576–589.

Heinz, S., Texari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., Rialdi, A., White, K. M., Albrecht, R. A., Pache, L., Marazzi, I., García-Sastre, A., Shaw, M. L., & Benner, C. (2018). Transcription Elongation Can Affect Genome 3D Structure. *Cell*, *174*(6), 1522-1536.e22.

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.*, *33*(5), 510–517. https://doi.org/10.1038/nbt.3199

Hnisz, D., Schuijers, J., Lin, C. Y., Weintraub, A. S., Abraham, B. J., Lee, T. I., Bradner, J. E., & Young, R. A. (2015). Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol. Cell*, *58*(2), 362–370. https://doi.org/10.1016/j.molcel.2015.02.014

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, *169*(1), 13–23.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, *9*(5), 473–476.

Hon, G., Ren, B., & Wang, W. (2008). ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.*, *4*(10), e1000201. https://doi.org/10.1371/journal.pcbi.1000201

Hon, G., Wang, W., & Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.*, *5*(11), e1000566. https://doi.org/10.1371/journal.pcbi.1000566

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., & Weissman, J. S. (2016). Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife*, *5*. https://doi.org/10.7554/eLife.12677

Hsu, J. Y., Fulco, C. P., Cole, M. A., Canver, M. C., Pellin, D., Sher, F., Farouni, R., Clement, K., Guo, J. A., Biasco, L., Orkin, S. H., Engreitz, J. M., Lander, E. S., Joung, J. K., Bauer, D. E., & Pinello, L. (2018). CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. *Nat. Methods*, *15*(12), 992–993. https://doi.org/10.1038/s41592-018-0225-6

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*(6), 1262–1278. https://doi.org/10.1016/j.cell.2014.05.010

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013).

DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, *31*(9), 827–832. https://doi.org/10.1038/nbt.2647

Huang, W.-C., Ferris, E., Cheng, T., Hörndli, C. S., Gleason, K., Tamminga, C., Wagner, J. D., Boucher, K. M., Christian, J. L., & Gregg, C. (2017). Diverse Non-genetic, Allele-Specific Expression Effects Shape Genetic Architecture at the Cellular Level in the Mammalian Brain. *Neuron*, *93*(5), 1094-1109.e7. https://doi.org/10.1016/j.neuron.2017.01.033

Huang, Y. F., Gulko, B., & Siepel, A. (2016). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *BioRxiv*. http://biorxiv.org/content/early/2016/08/15/069682.abstract

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., & Others. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, *41*(2), 178–186.

Isaac, R. S., Jiang, F., Doudna, J. A., Lim, W. A., Narlikar, G. J., & Almeida, R. (2016). Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *Elife*, *5*. https://doi.org/10.7554/eLife.13450

Itzhak, D. N., Tyanova, S., Cox, J., & Borner, G. H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. *Elife*, *5*. https://doi.org/10.7554/eLife.16950

Jansen, Rund, van Embden, J. D. A., Gaastra, W., & Schouls, L. M. (2002). Identification of a novel family of sequence repeats among prokaryotes. *OMICS*, *6*(1), 23–33. https://doi.org/10.1089/15362310252780816

Jansen, Ruud, Embden, J. D. A. van, Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, *43*(6), 1565–1575.

Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science*, *293*(5532), 1074–1080.

Jeong, H.-H., Kim, S. Y., Rousseaux, M. W. C., Zoghbi, H. Y., & Liu, Z. (2019). Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome Res.*, *29*(6), 999–1008. https://doi.org/10.1101/gr.245571.118

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science, 337*(6096), 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., & Doudna, J. (2013). RNA-programmed genome editing in human cells. *Elife*, *2*, e00471. https://doi.org/10.7554/eLife.00471

Johnson, S. A., & Hunter, T. (2005). Kinomics: Methods for deciphering the kinome. *Nat. Methods*, *2*(1), 17–25. https://doi.org/10.1038/nmeth731

Kabadi, A. M., Ousterout, D. G., Hilton, I. B., & Gersbach, C. A. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res.*, *42*(19), e147. https://doi.org/10.1093/nar/gku749

Kain, M. P., Bolker, B. M., & McCoy, M. W. (2015). A practical guide and power analysis for GLMMs: Detecting among treatment variation in random effects. *PeerJ*, *3*, e1226.

Kanhere, A., Hertweck, A., Bhatia, U., Gökmen, M. R., Perucha, E., Jackson, I., Lord, G. M., & Jenner, R. G. (2012). T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nat. Commun.*, *3*, 1268.

Karnik, R., & Beer, M. A. (2015). Identification of Predictive Cis-Regulatory Elements Using a Discriminative Objective Function and a Dynamic Search Space. *PLoS One*, *10*(10), e0140557. https://doi.org/10.1371/journal.pone.0140557

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., Li, J., Xie, D., Olarerin-George, A., Steinmetz, L. M., Hogenesch, J. B., Kellis, M., Batzoglou, S., & Snyder, M. (2013). Extensive variation in chromatin states across humans. *Science*, *342*(6159), 750–752. https://doi.org/10.1126/science.1242510

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., & Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.*, *22*(12), 2497–2506. https://doi.org/10.1101/gr.143008.112

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, and D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.*, *10*(10), e1004722. https://doi.org/10.1371/journal.pgen.1004722

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., … Dermitzakis, E. T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, *342*(6159), 744–747. https://doi.org/10.1126/science.1242463

Kim, H., & Kim, J.-S. (2014). A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.*, *15*(5), 321–334.

Kim, H. S., Tan, Y., Ma, W., Merkurjev, D., Destici, E., Ma, Q., Suter, T., Ohgi, K., Friedman, M., Skowronska-Krawczyk, D., & Rosenfeld, M. G. (2018). Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature*, *556*(7702), 510–514. https://doi.org/10.1038/s41586-018-0048-8

Kim, H.-J., Barnitz, R. A., Kreslavsky, T., Brown, F. D., Moffett, H., Lemieux, M. E., Kaygusuz, Y., Meissner, T., Holderried, T. A. W., Chan, S., Kastner, P., Haining, W. N., & Cantor, H. (2015). Stable inhibitory activity of regulatory T cells requires the transcription factor Helios. *Science*, *350*(6258), 334–339.

Kim, J.-H., Ebersole, T., Kouprina, N., Noskov, V. N., Ohzeki, J.-I., Masumoto, H., Mravinac, B., Sullivan, B. A., Pavlicek, A., Dovat, S., Pack, S. D., Kwon, Y.-W., Flanagan, P. T., Loukinov, D., Lobanenkov, V., & Larionov, V. (2009). Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res.*, *19*(4), 533–544. https://doi.org/10.1101/gr.086496.108

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.3853

Knowles, D. A., Davis, J. R., Edgington, H., Raj, A., Favé, M.-J., Zhu, X., Potash, J. B., Weissman, M. M., Shi, J., Levinson, D. F., Awadalla, P., Mostafavi, S., Montgomery, S. B., & Battle, A. (2017). Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods*. https://doi.org/10.1038/nmeth.4298

Kolde, R., Laur, S., Adler, P., & Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, *28*(4), 573–580. https://doi.org/10.1093/bioinformatics/btr709

Komander, D., & Rape, M. (2012). The ubiquitin code. *Annu. Rev. Biochem.*, *81*, 203–229. https://doi.org/10.1146/annurev-biochem-060310-170328

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2015).

Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588. https://doi.org/10.1038/nature14136

Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., & Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.*, *34*(2), 192–198. https://doi.org/10.1038/nbt.3450

Kosicki, M., Tomberg, K., & Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4192

Kramer, A. C., Kothari, A., Wilson, W. C., Celik, H., Nikitas, J., Mallaney, C., Ostrander, E. L., Eultgen, E., Martens, A., Valentine, M. C., Young, A. L., Druley, T. E., Figueroa, M. E., Zhang, B., & Challen, G. A. (2017). Dnmt3a regulates T-cell development and suppresses T-ALL transformation. *Leukemia*. https://doi.org/10.1038/leu.2017.89

Kreimer, A., Zeng, H., Edwards, M. D., Guo, Y., Tian, K., Shin, S., Welch, R., Wainberg, M., Mohan, R., Sinnott-Armstrong, N. A., Li, Y., Eraslan, G., Amin, T. B., Tewhey, R., Sabeti, P. C., Goke, J., Mueller, N. S., Kellis, M., Kundaje, A., … Yosef, N. (2017). Predicting gene expression in massively parallel reporter assays: A comparative study. *Hum. Mutat.*, *38*(9), 1240–1250. https://doi.org/10.1002/humu.23197

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., & Troyanskaya, O. G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.*, *19*(11), 1454–1462. https://doi.org/10.1038/nn.4353

Kukurba, K. R., Parsana, P., Balliu, B., Smith, K. S., Zappala, Z., Knowles, D. A., Favé, M.-J., Davis, J. R., Li, X., Zhu, X., Potash, J. B., Weissman, M. M., Shi, J., Kundaje, A., Levinson, D. F., Awadalla, P., Mostafavi, S., Battle, A., & Montgomery, S. B. (2016). Impact of the X Chromosome and sex on regulatory variation. *Genome Res.*, *26*(6), 768–777. https://doi.org/10.1101/gr.197897.115

Kumasaka, N., Knights, A., & Gaffney, D. (2017). High resolution genetic mapping of causal regulatory interactions in the human genome. In *BioRxiv*. https://doi.org/10.1101/227389

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.*, *48*(2), 206–213. https://doi.org/10.1038/ng.3467

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, *24*(10), 1595–1602. https://doi.org/10.1101/gr.173518.114

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, *11*(3), 191–203. https://doi.org/10.1038/nrg2732

Le Gras, S., Keime, C., Anthony, A., Lotz, C., De Longprez, L., Brouillet, E., Cassel, J.-C., Boutillier, A.-L., & Merienne, K. (2017). Altered enhancer transcription underlies Huntington's disease striatal transcriptional signature. *Sci. Rep.*, *7,* 42875. https://doi.org/10.1038/srep42875

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, *47*(8), 955–961.

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, *13*(4), 233–245. https://doi.org/10.1038/nrg3163

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., … Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, *518*(7539), 350–354. https://doi.org/10.1038/nature14217

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1–2), 84–98. https://doi.org/10.1016/j.cell.2011.12.014

Li, H. (2013a). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv:1303.3997 [q-Bio]*. http://arxiv.org/abs/1303.3997

Li, H. (2013b). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *ArXiv [q-bio.GN]*. http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, P., Mitra, S., Spolski, R., Oh, J., Liao, W., Tang, Z., Mo, F., Li, X., West, E. E., Gromer, D., Lin, J.-X., Liu, C., Ruan, Y., & Leonard, W. J. (2017). STAT5-mediated chromatin interactions in superenhancers activate IL-2 highly inducible genes: Functional dissection of the Il2ra gene locus. *Proc. Natl. Acad. Sci. U. S. A.*

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., & Liu, X. S. (2014). MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, *15*(12), 554. https://doi.org/10.1186/s13059-014-0554-4

Li, Xin, Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Hess, G. T., Zappala, Z., Strober, B. J., Scott, A. J., Li, A., Ganna, A., Bassik, M. C., Merker, J. D., GTEx Consortium, Laboratory, D. A. &Coordinating C. (LDACC)—Analysis W. G., Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, … Montgomery, S. B. (2017). The impact of rare variation on gene expression across tissues. *Nature*, *550*(7675), 239–243. https://doi.org/10.1038/nature24267

Li, Xinrui, Wu, J., Ptacek, T., Redden, D. T., Brown, E. E., Alarcón, G. S., Ramsey-Goldman, R., Petri, M. A., Reveille, J. D., Kaslow, R. A., Kimberly, R. P., & Edberg, J. C. (2013). Allelic-dependent expression of an activating Fc receptor on B cells enhances humoral immune responses. *Sci. Transl. Med.*, *5*(216), 216ra175. https://doi.org/10.1126/scitranslmed.3007097

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y., & Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, *352*(6285), 600–604. https://doi.org/10.1126/science.aad9417

Liang, Y., Tsoi, L. C., Xing, X., Beamer, M. A., Swindell, W. R., Sarkar, M. K., Berthier, C. C., Stuart, P. E., Harms, P. W., Nair, R. P., Elder, J. T., Voorhees, J. J., Kahlenberg, J. M., & Gudjonsson, J. E. (2017). A gene network regulated by the transcription factor VGLL3 as a promoter of sex-biased autoimmune diseases. *Nat. Immunol.*, *18*(2), 152–160.

Liao, H.-K., Hatanaka, F., Araoka, T., Reddy, P., Wu, M.-Z., Sui, Y., Yamauchi, T., Sakurai, M., O'Keefe, D. D., Núñez-Delicado, E., Guillen, P., Campistol, J. M., Wu, C.-J., Lu, L.-F., Esteban, C. R., & Izpisua Belmonte, J. C. (2017). In Vivo Target Gene Activation via CRISPR/Cas9-Mediated Trans-epigenetic Modulation. *Cell*, *171*(7), 1495-1507.e15. https://doi.org/10.1016/j.cell.2017.10.025

Liau, W. S., Tan, S. H., Ngoc, P. C. T., Wang, C. Q., Tergaonkar, V., Feng, H., Gong, Z., Osato, M., Look, A. T., & Sanda, T. (2017). Aberrant activation of the GIMAP enhancer by oncogenic transcription factors in T-cell acute lymphoblastic leukemia. *Leukemia*. https://doi.org/10.1038/leu.2016.392

Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., Abe, Y., Skola, D., Romanoski, C. E., Tao, J., Fonseca, G. J., Troutman, T. D., Spann, N. J., Strid, T., Sakai, M., Yu, M., Hu, R., Fang, R., Metzler, D., … Glass, C. K. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell*, *173*(7), 1796-1809.e17. https://doi.org/10.1016/j.cell.2018.04.018

Liu, S. J., Horlbeck, M. A., Cho, S. W., Birk, H. S., Malatesta, M., He, D., Attenello, F. J., Villalta, J. E., Cho, M. Y., Chen, Y., Mandegar, M. A., Olvera, M. P., Gilbert, L. A., Conklin, B. R., Chang, H. Y., Weissman, J. S., & Lim, D. A. (2017). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science, 355*(6320). https://doi.org/10.1126/science.aah7111

Liu, Y., Easton, J., Shao, Y., Maciaszek, J., Wang, Z., Wilkinson, M. R., McCastlain, K., Edmonson, M., Pounds, S. B., Shi, L., Zhou, X., Ma, X., Sioson, E., Li, Y., Rusch, M., Gupta, P., Pei, D., Cheng, C., Smith, M. A., … Mullighan, C. G. (2017). The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.*, *49*(8), 1211–1218.

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, *167*(5), 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Love, M. I., Myšičková, A., Sun, R., Kalscheuer, V., Vingron, M., & Haas, S. A. (2011). Modeling read counts for CNV detection in exome sequencing data. *Stat. Appl. Genet. Mol. Biol.*, *10*(1). https://doi.org/10.2202/1544-6115.1732

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., & Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, *153*(2), 320–334. https://doi.org/10.1016/j.cell.2013.03.036

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., Levens, D. L., Lee, T. I., & Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, *151*(3), 476–482. https://doi.org/10.1016/j.cell.2012.10.012

Lu, J., Tomfohr, J. K., & Kepler, T. B. (2005). Identifying differential expression in multiple SAGE libraries: An overdispersed log-linear model approach. *BMC Bioinformatics*, *6*, 165. https://doi.org/10.1186/1471-2105-6-165

Lucic, B., Chen, H.-C., Kuzman, M., Zorita, E., Wegner, J., Minneker, V., Wang, W., Fronza, R., Laufs, S., Schmidt, M., Stadhouders, R., Roukos, V., Vlahovicek, K., Filion, G. J., & Lusic, M. (2019). Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12046-3

Luo, C., Sidote, D. J., Zhang, Y., Kerstetter, R. A., Michael, T. P., & Lam, E. (2013). Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant J.*, *73*(1), 77–90. https://doi.org/10.1111/tpj.12017

Ma, J., Köster, J., Qin, Q., Hu, S., Li, W., Chen, C., Cao, Q., Wang, J., Mei, S., Liu, Q., Xu, H., & Liu, X. S. (2016). CRISPR-DO for genome-wide CRISPR design and optimization. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw476

Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., Yang, L., & Church, G. M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, *31*(9), 833–838. https://doi.org/10.1038/nbt.2675

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., & Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science*, *339*(6121), 823–826. https://doi.org/10.1126/science.1232033

Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., Loh, M. L., Hunger, S. P., Sanda, T., Young, R. A., & Look, A. T. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, *346*(6215), 1373–1377. https://doi.org/10.1126/science.1259037

Marigorta, U. M., Denson, L. A., Hyams, J. S., Mondal, K., Prince, J., Walters, T. D., Griffiths, A., Noe, J. D., Crandall, W. V., Rosh, J. R., Mack, D. R., Kellermayer, R., Heyman, M. B., Baker, S. S., Stephens, M. C., Baldassano, R. N., Markowitz, J. F., Kim, M.-O., Dubinsky, M. C., … Gibson, G. (2017). Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn’s disease. *Nat. Genet.*, *49*(10), 1517–1521.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, *18*(9), 1509. https://doi.org/10.1101/gr.079558.108

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., Cooper, N. J., Barton, A., Wallace, C., Fraser, P., Worthington, J., & Eyre, S. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.*, *6*, 10069. https://doi.org/10.1038/ncomms10069

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, *7*, 29–59.

Matharu, N., Rattanasopha, S., Maliskova, L., Wang, Y., Hardin, A., Vaisse, C., & Ahituv, N. (2017). Promoter Or Enhancer Activation By CRISPRa Rescues Haploinsufficiency Caused Obesity. In *BioRxiv*. https://doi.org/10.1101/140426

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in

Regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *J. Am. Stat. Assoc.*, *92*(437), 162–170. https://doi.org/10.1080/01621459.1997.10473613

McFarland, J. M., Ho, Z. V., Kugener, G., Dempster, J. M., Montgomery, P. G., Bryan, J. G., Krill-Burger, J. M., Green, T. M., Vazquez, F., Boehm, J. S., Golub, T. R., Hahn, W. C., Root, D. E., & Tsherniak, A. (2018). Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat. Commun.*, *9*(1), 4610. https://doi.org/10.1038/s41467-018-06916-5

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, *20*(9), 1297–1303.

McKenna, A., & Shendure, J. (2018). FlashFry: A fast and flexible tool for large-scale CRISPR target design. *BMC Biol.*, *16*(1), 74. https://doi.org/10.1186/s12915-018-0545-0

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., Painter, J. N., Martin, N. G., Visscher, P. M., & Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.*, *15*(5), R73. https://doi.org/10.1186/gb-2014-15-5-r73

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., & Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, *342*(6159), 747–749. https://doi.org/10.1126/science.1242429

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Jr, Kinney, J. B., Kellis, M., Lander, E. S., & Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, *30*(3), 271–277.

Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K., & Bernstein, B. E. (2013). Locus-specific editing of histone modifications at endogenous enhancers. *Nat. Biotechnol.*, *31*(12), 1133–1136. https://doi.org/10.1038/nbt.2701

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., … Tsherniak, A. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.*, *49*(12), 1779–1784. https://doi.org/10.1038/ng.3984

Michlits, G., Hubmann, M., Wu, S.-H., Vainorius, G., Budusan, E., Zhuk, S., Burkard, T. R., Novatchkova, M., Aichinger, M., Lu, Y., Reece-Hoyes, J., Nitsch, R., Schramek, D., Hoepfner, D., & Elling, U. (2017). CRISPR-UMI: single-cell lineage tracing of pooled CRISPR–Cas9 screens. *Nat. Methods*, *14*, 1191. https://doi.org/10.1038/nmeth.4466

Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S., Urich, M. A., Nery, J. R., Sejnowski, T. J., Lister, R., Eddy, S. R., Ecker, J. R., & Nathans, J. (2015). Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. *Neuron*, *86*(6), 1369–1384. https://doi.org/10.1016/j.neuron.2015.05.018

Morgens, D. W., Deans, R. M., Li, A., & Bassik, M. C. (2016). Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat. Biotechnol.*, *34*(6), 634–636.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, *13*(11), 919–922. https://doi.org/10.1038/nmeth.3999

Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Ryan Corces, M., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., … Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* https://doi.org/10.1038/ng.3963

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, *320*(5881), 1344–1349. https://doi.org/10.1126/science.1158441

Navarro, J.-M., Touzart, A., Pradel, L. C., Loosveld, M., Koubi, M., Fenouil, R., Le Noir, S., Maqbool, M. A., Morgado, E., Gregoire, C., Jaeger, S., Mamessier, E., Pignon, C., Hacein-Bey-Abina, S., Malissen, B., Gut, M., Gut, I. G., Dombret, H., Macintyre, E. A., … Nadel, B. (2015). Site- and allele-specific polycomb dysregulation in T-cell leukaemia. *Nat. Commun.*, *6*, 6094.

Northcott, P. A., Lee, C., Zichner, T., Stütz, A. M., Erkek, S., Kawauchi, D., Shih, D. J. H., Hovestadt, V., Zapatka, M., Sturm, D., Jones, D. T. W., Kool, M., Remke, M., Cavalli, F. M. G., Zuyderduyn, S., Bader, G. D., VandenBerg, S., Esparza, L. A., Ryzhova, M., … Pfister, S. M. (2014). Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, *511*(7510), 428–434. https://doi.org/10.1038/nature13379

Nuzzo, R. (2015). How scientists fool themselves—And how they can stop. *Nature*, *526*(7572), 182–185. https://doi.org/10.1038/526182a

Olshen, A. B., Venkatraman, E. S., Lucito, R., & Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, *5*(4), 557–572. https://doi.org/10.1093/biostatistics/kxh008

Ong, C.-T., & Corces, V. G. (2011). Enhancer function: New insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, *12*(4), 283–293. https://doi.org/10.1038/nrg2957

Orlando, G., Law, P. J., Cornish, A. J., Dobbins, S. E., Chubb, D., Broderick, P., Litchfield, K., Hariri, F., Pastinen, T., Osborne, C. S., Taipale, J., & Houlston, R. S. (2018). Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nat. Genet.*, 1. https://doi.org/10.1038/s41588-018-0211-z

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., & Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, *143*(1), 46–58. https://doi.org/10.1016/j.cell.2010.09.001

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*, *554*(7691), 239–243. https://doi.org/10.1038/nature25461

Pan, J.-X., & Thompson, R. (2000). *Generalized linear mixed models: An improved estimating procedure*. 373–378. https://doi.org/10.1007/978-3-642-57678-2_49

Park, Y., Yoon, P., Hyung-seung, J., Daisuke, A., Jeeho, L., & Yun-Cai, L. (2014). The Ubiquitin System in Immune Regulation. In *Advances in Immunology* (pp. 17–66). https://doi.org/10.1016/b978-0-12-800147-9.00002-9

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S.-I., Cooper, G. M., Ahituv, N., Pennacchio, L. A., & Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, *30*(3), 265–270. https://doi.org/10.1038/nbt.2136

Perez, A. R., Pritykin, Y., Vidigal, J. A., Chhangawala, S., Zamparo, L., Leslie, C. S., & Ventura, A. (2017). GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.*, *35*(4), 347–349. https://doi.org/10.1038/nbt.3804

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nat. Methods*, *10*(10), 973–976. https://doi.org/10.1038/nmeth.2600

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.*, *20*(17), 1562–1567. https://doi.org/10.1016/j.cub.2010.07.043

Philip, M., Fairchild, L., Sun, L., Horste, E. L., Camara, S., Shakiba, M., Scott, A. C., Viale, A., Lauer, P., Merghoub, T., Hellmann, M. D., Wolchok, J. D., Leslie, C. S., & Schietinger, A. (2017). Chromatin states define tumour-specific T cell dysfunction and reprogramming. *Nature*, *545*(7655), 452–456. https://doi.org/10.1038/nature22367

Piao, Y., Lee, S. K., Lee, E.-J., Robertson, K. D., Shi, H., Ryu, K. H., & Choi, J.-H. (2016). CAME: identification of chromatin accessibility from nucleosome occupancy and methylome sequencing. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw785

Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, *94*(4), 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–772.

Pinello, L., Canver, M. C., Hoban, M. D., Orkin, S. H., Kohn, D. B., Bauer, D. E., & Yuan, G.-C. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.*, *34*(7), 695–697. https://doi.org/10.1038/nbt.3583

Pinheiro, J. C., & Chao, E. C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. *J. Comput. Graph. Stat.*, *15*(1), 58–81. https://doi.org/10.1198/106186006X96962

Pirinen, M., Lappalainen, T., Zaitlen, N. A., GTEx Consortium, Dermitzakis, E. T., Donnelly, P., McCarthy, M. I., & Rivas, M. A. (2015). Assessing allele-specific expression across multiple tissues from RNA-seq read data. *Bioinformatics*, *31*(15), 2497–2504. https://doi.org/10.1093/bioinformatics/btv074

Poplin, R., Newburger, D., Dijamco, J., Nguyen, N., Loy, D., Gross, S. S., McLean, C. Y., & DePristo, M. A. (2016). Creating a universal SNP and small indel variant caller with deep neural networks. In *BioRxiv*. https://doi.org/10.1101/092890

Pott, S., & Lieb, J. D. (2014). What are super-enhancers? *Nat. Genet.*, *47*(1), 8–12.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K., & Ren, B. (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.*, *21*(3), 432–439.

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F. H., Swigut, T., & Wysocka, J. (2015). Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell*, *163*(1), 68–83. https://doi.org/10.1016/j.cell.2015.08.036

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Pulido-Quetglas, C., Aparicio-Prat, E., Arnan, C., Polidori, T., Hermoso, T., Palumbo, E., Ponomarenko, J., Guigo, R., & Johnson, R. (2017). Scalable Design of Paired CRISPR Guide RNAs for Genomic Deletion. *PLoS Comput. Biol.*, *13*(3), e1005341. https://doi.org/10.1371/journal.pcbi.1005341

Qiu, J., Sheedlo, M. J., Yu, K., Tan, Y., Nakayasu, E. S., Das, C., Liu, X., & Luo, Z.-Q. (2016). Ubiquitination independent of E1 and E2 enzymes by bacterial effectors. *Nature*, *533*(7601), 120–124.

Raffa, J. D., & Dubin, J. A. (2015). Multivariate longitudinal data analysis with mixed effects hidden Markov models. *Biometrics*, *71*(3), 821–831. https://doi.org/10.1111/biom.12296

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K., & Sherwood, R. I. (2016). High-throughput mapping of regulatory DNA. *Nat. Biotechnol.*, *34*(2), 167–174. https://doi.org/10.1038/nbt.3468

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M., & Ren, B. (2013). RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.*, *9*(3), e1002968. https://doi.org/10.1371/journal.pcbi.1002968

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.*, *12*(8), 529–541. https://doi.org/10.1038/nrg3000

Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., & Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nat. Methods*, *14*(3), 263–266. https://doi.org/10.1038/nmeth.4155

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, *44*(W1), W160-5. https://doi.org/10.1093/nar/gkw257

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Raudenbush, S. W., Yang, M.-L., & Yosef, M. (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *J. Comput. Graph. Stat.*, *9*(1), 141–157. https://doi.org/10.1080/10618600.2000.10474870

Reeme, A. E., Claeys, T. A., Aggarwal, P., Turner, A. J., Routes, J. M., Broeckel, U., & Robinson, R. T. (2018). Human IL12RB1 expression is allele-biased and produces a novel IL12 response regulator. *Genes Immun.* https://doi.org/10.1038/s41435-018-0023-2

Reilly, S. K., Gosai, S. J., Gutierrez, A., Ulirsch, J. C., Kanai, M., Berenzy, D., Kales, S., Butler, G. B., Gladden-Young, A., Finucane, H. K., Sabeti, P. C., & Tewhey, R. (2020). HCR-FlowFISH: A flexible CRISPR screening method to identify cis-regulatory elements and their target genes. In *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.05.11.078675

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

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

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010b). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, *23*(21), 2881–2887. https://doi.org/10.1093/bioinformatics/btm453

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Rusu, R. B., Blodow, N., & Beetz, M. (2009). Fast point feature histograms (FPFH) for 3D registration. *Robotics and Automation, 2009.* http://ieeexplore.ieee.org/abstract/document/5152473/

Sadhu, M. J., Bloom, J. S., Day, L., & Kruglyak, L. (2016). CRISPR-directed mitotic recombination enables genetic mapping without crosses. *Science, 352*(6289), 1113–1116. https://doi.org/10.1126/science.aaf5124

Samarakoon, P. S., Sorte, H. S., Stray-Pedersen, A., Rødningen, O. K., Rognes, T., & Lyle, R. (2016). cnvScan: A CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data. *BMC Genomics, 17,* 51. https://doi.org/10.1186/s12864-016-2374-2

Sanjana, N. E., Shalem, O., & Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods, 11*(8), 783–784. https://doi.org/10.1038/nmeth.3047

Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., & Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science, 353*(6307), 1545–1549. https://doi.org/10.1126/science.aaf7613

Santoni, F. A., Stamoulis, G., Garieri, M., Falconnet, E., Ribaux, P., Borel, C., & Antonarakis, S. E. (2017). Detection of Imprinted Genes by Single-Cell Allele-Specific Gene Expression. *Am. J. Hum. Genet., 100*(3), 444–453. https://doi.org/10.1016/j.ajhg.2017.01.028

Sauls, J. T., & Buescher, J. M. (2014). Assimilating genome-scale metabolic reconstructions with modelBorgifier. *Bioinformatics, 30*(7), 1036–1038. https://doi.org/10.1093/bioinformatics/btt747

Savas, P., Virassamy, B., Ye, C., Salim, A., Mintoff, C. P., Caramia, F., Salgado, R., Byrne, D. J., Teo, Z. L., Dushyanthen, S., Byrne, A., Wein, L., Luen, S. J., Poliness, C., Nightingale, S. S., Skandarajah, A. S., Gyorki, D. E., Thornton, C. M., Beavis, P. A., … Loi, S. (2018). Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med., 24*(7), 986–993.

Schreiber, J., Durham, T. J., Bilmes, J., & Noble, W. S. (2018). Multi-scale deep tensor factorization learns a latent representation of the human epigenome. In *BioRxiv*. https://doi.org/10.1101/364976

Schrode, N., Ho, S.-M., Yamamuro, K., Dobbyn, A., Huckins, L., Matos, M. R., Cheng, E., Deans, P. J. M., Flaherty, E., Barretto, N., Topol, A., Alganem, K., Abadali, S., Gregory, J., Hoelzli, E., Phatnani, H., Singh, V., Girish, D., Aronow, B., … Brennand, K. J. (2019). Synergistic effects of common schizophrenia risk variants. *Nat. Genet., 51*(10), 1475–1485. https://doi.org/10.1038/s41588-019-0497-5

Servin, B., & Stephens, M. (2005). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics, preprint*(2007), e114. https://doi.org/10.1371/journal.pgen.0030114.eor

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., & Zhang, F. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, *343*(6166), 84–87. https://doi.org/10.1126/science.1247005

Shalem, O., Sanjana, N. E., & Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.*, *16*(5), 299–311. https://doi.org/10.1038/nrg3899

Shema, E., Jones, D., Shoresh, N., Donohue, L., Ram, O., & Bernstein, B. E. (2016). Single-molecule decoding of combinatorially modified nucleosomes. *Science*, *352*(6286), 717–721.

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C.-C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., … Mali, P. (2017). Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods*. https://doi.org/10.1038/nmeth.4225

Shi, J., Wang, E., Milazzo, J. P., Wang, Z., Kinney, J. B., & Vakoc, C. R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.*, *33*(6), 661–667. https://doi.org/10.1038/nbt.3235

Shih, H.-Y., Sciumè, G., Mikami, Y., Guo, L., Sun, H.-W., Brooks, S. R., Urban, J. F., Jr, Davis, F. P., Kanno, Y., & O’Shea, J. J. (2016). Developmental Acquisition of Regulomes Underlies Innate Lymphoid Cell Functionality. *Cell*, *165*(5), 1120–1133. https://doi.org/10.1016/j.cell.2016.04.029

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P., & Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, *100*(26), 15776–15781. https://doi.org/10.1073/pnas.2136655100

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.*, *15*(4), 272–286.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Simeonov, D. R., Gowen, B. G., Boontanrart, M., Roth, T. L., Gagnon, J. D., Mumbach, M. R., Satpathy, A. T., Lee, Y., Bray, N. L., Chan, A. Y., Lituiev, D. S., Nguyen, M. L., Gate, R. E., Subramaniam, M., Li, Z., Woo, J. M., Mitros, T., Ray, G. J., Curie, G. L., …

Marson, A. (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*, *549*(7670), 111–115. https://doi.org/10.1038/nature23875

Simon, N. C., Aktories, K., & Barbieri, J. T. (2014). Novel bacterial ADP-ribosylating toxins: Structure and function. *Nat. Rev. Microbiol.*, *12*(9), 599–611. https://doi.org/10.1038/nrmicro3310

Skaar, J. R., Pagan, J. K., & Pagano, M. (2014). SCF ubiquitin ligase-targeted therapies. *Nat. Rev. Drug Discov.*, *13*(12), 889–903. https://doi.org/10.1038/nrd4432

Sloan, C. A., Chan, E. T., Davidson, J. M., Malladi, V. S., Strattan, J. S., Hitz, B. C., Gabdank, I., Narayanan, A. K., Ho, M., Lee, B. T., Rowe, L. D., Dreszer, T. R., Roe, G., Podduturi, N. R., Tanaka, F., Hong, E. L., & Cherry, J. M. (2016). ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, *44*(D1), D726-32. https://doi.org/10.1093/nar/gkv1160

Smith, J. D., Suresh, S., Schlecht, U., Wu, M., Wagih, O., Peltz, G., Davis, R. W., Steinmetz, L. M., Parts, L., & St Onge, R. P. (2016). Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. *Genome Biol.*, *17*, 45. https://doi.org/10.1186/s13059-016-0900-9

Smyth, G. K. (n.d.). *Limma: Linear models for microarray data. Bioinformatics and computational biology solutions using R and bioconductor. Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. 2005*.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397–420). Springer New York. https://doi.org/10.1007/0-387-29362-0_23

Song, L., & Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, *2010*(2), db.prot5384. https://doi.org/10.1101/pdb.prot5384

Speed, D., Cai, N., UCLEB Consortium, Johnson, M. R., Nejentsev, S., & Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, *49*(7), 986–992. https://doi.org/10.1038/ng.3865

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nat. Rev. Genet.*, *13*(9), 613–626.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., & Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, *480*(7378), 490–495. https://doi.org/10.1038/nature10716

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature, 403*(6765), 41–45. https://doi.org/10.1038/47412

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., & Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nat. Genet.*, *39*(10), 1217–1224. https://doi.org/10.1038/ng2142

Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, *9*, 303. https://doi.org/10.1186/1471-2105-9-303

Su, J., Yan, H., Wei, Y., Liu, H., Liu, H., Wang, F., Lv, J., Wu, Q., & Zhang, Y. (2013). CpG_MPs: Identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Res.*, *41*(1), e4. https://doi.org/10.1093/nar/gks829

Sun, J. H., Zhou, L., Emerson, D. J., Phyo, S. A., Titus, K. R., Gong, W., Gilgenast, T. G., Beagan, J. A., Davidson, B. L., Tassone, F., & Phillips-Cremins, J. E. (2018). Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell*, *175*(1), 224-238.e15. https://doi.org/10.1016/j.cell.2018.08.005

Suzuki, H. I., Young, R. A., & Sharp, P. A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. *Cell*, *168*(6), 1000-1014.e15.

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*, *9*(6), 465–476. https://doi.org/10.1038/nrg2341

Takaku, M., Grimm, S. A., Roberts, J. D., Chrysovergis, K., Bennett, B. D., Myers, P., Perera, L., Tucker, C. J., Perou, C. M., & Wade, P. A. (2018). GATA3 zinc finger 2 mutations reprogram the breast cancer transcriptional network. *Nat. Commun.*, *9*(1), 1059. https://doi.org/10.1038/s41467-018-03478-4

Taudt, A., Colomé-Tatché, M., & Johannes, F. (2016). Genetic sources of population epigenomic variation. *Nat. Rev. Genet.*, *17*(6), 319–332. https://doi.org/10.1038/nrg.2016.45

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, *165*(6), 1519–1529. https://doi.org/10.1016/j.cell.2016.04.027

Thakore, P. I., D’Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., & Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82.

Tycko, J., Wainberg, M., Marinov, G. K., Ursu, O., Hess, G. T., Ego, B. K., Aradhana, Li, A., Truong, A., Trevino, A. E., Spees, K., Yao, D., Kaplow, I. M., Greenside, P. G., Morgens, D. W., Phanstiel, D. H., Snyder, M. P., Bintu, L., Greenleaf, W. J., … Bassik, M. C. (2019). Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat. Commun.*, *10*(1), 4063. https://doi.org/10.1038/s41467-019-11955-7

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T. S., & Sankaran, V. G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*, *165*(6), 1530–1545.

Vahedi, G., Kanno, Y., Furumoto, Y., Jiang, K., Parker, S. C. J., Erdos, M. R., Davis, S. R., Roychoudhuri, R., Restifo, N. P., Gadina, M., Tang, Z., Ruan, Y., Collins, F. S., Sartorelli, V., & O’Shea, J. J. (2015). Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature*, *520*(7548), 558–562.

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, *12*(11), 1061–1063.

van Duin, R. A., Sarah Rennie, Maria Dalby, Lucas. (n.d.). *Transcriptional decomposition reveals activechromatin architectures and cell specific regulatoryinteractions*. https://doi.org/10.1101/130070

van Overbeek, M., Capurso, D., Carter, M. M., Thompson, M. S., Frias, E., Russ, C., Reece-Hoyes, J. S., Nye, C., Gradia, S., Vidal, B., Zheng, J., Hoffman, G. R., Fuller, C. K., & May, A. P. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Molecular Cell*, *63*(4), 633–646. https://doi.org/10.1016/j.molcel.2016.06.037

Venegas, C., Kumar, S., Franklin, B. S., Dierkes, T., Brinkschulte, R., Tejera, D., Vieira-Saecker, A., Schwartz, S., Santarelli, F., Kummer, M. P., Griep, A., Gelpi, E., Beilharz, M., Riedel, D., Golenbock, D. T., Geyer, M., Walter, J., Latz, E., & Heneka, M. T. (2017). Microglia-derived ASC specks cross-seed amyloid-β in Alzheimer’s disease. *Nature*, *552*(7685), 355–361. https://doi.org/10.1038/nature25158

Ventham, N. T., Kennedy, N. A., Adams, A. T., Kalla, R., Heath, S., O’Leary, K. R., Drummond, H., IBD BIOM consortium, IBD CHARACTER consortium, Wilson, D. C., Gut, I. G., Nimmo, E. R., & Satsangi, J. (2016). Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. *Nat. Commun.*, *7*, 13507. https://doi.org/10.1038/ncomms13507

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., & Pritchard, J. K. (2008). High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genetics*, *4*(10), e1000214. https://doi.org/10.1371/journal.pgen.1000214

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Schizophrenia Working Group of the Psychiatric Genomics Consortium, D., Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Visel, A., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., & Pennacchio, L. A. (2009). Functional autonomy of distant-acting human enhancers. *Genomics*, *93*(6), 509–513. https://doi.org/10.1016/j.ygeno.2009.02.002

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., & Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.*, *51*(4), 592–599. https://doi.org/10.1038/s41588-019-0385-z

Wan, Y. Y. (2014). GATA3: A master of many trades in immune regulation. *Trends Immunol.*, *35*(6), 233–242. https://doi.org/10.1016/j.it.2014.04.002

Wang, G., Sarkar, A., Carbonetto, P., & Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.*, *82*(5), 1273–1300. https://doi.org/10.1111/rssb.12388

Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2010). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, *26*(1), 136–138. https://doi.org/10.1093/bioinformatics/btp612

Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, *343*(6166), 80–84. https://doi.org/10.1126/science.1246981

Wang, X., Tucker, N. R., Rizki, G., Mills, R., Krijger, P. H., de Wit, E., Subramanian, V., Bartell, E., Nguyen, X.-X., Ye, J., Leyton-Mange, J., Dolmatova, E. V., van der Harst, P., de Laat, W., Ellinor, P. T., Newton-Cheh, C., Milan, D. J., Kellis, M., & Boyer, L. A. (2016). Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *Elife*, *5*. https://doi.org/10.7554/eLife.10557

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q., & Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, *40*(7), 897–903. https://doi.org/10.1038/ng.154

Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., Orioli, A., Wiederkehr, M., Panousis, N. I., Yurovsky, A., Romano-Palumbo, L., Planchon, A., Bielser, D., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Hernandez, N., Reymond, A., … Dermitzakis, E. T. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, *162*(5), 1039–1050. https://doi.org/10.1016/j.cell.2015.08.001

Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S., & Young, R. A. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, *171*(7), 1573-1588.e28.

Wellcome Trust Case Control Consortium, Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.*, *44*(12), 1294–1301. https://doi.org/10.1038/ng.2435

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, *42*(Database issue), D1001-6. https://doi.org/10.1093/nar/gkt1229

Wen, X. (2015). Effective QTL Discovery Incorporating Genomic Annotations. In *BioRxiv*. https://doi.org/10.1101/032003

Wen, X., Lee, Y., Luca, F., & Pique-Regi, R. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.*, *98*(6), 1114–1129. https://doi.org/10.1016/j.ajhg.2016.03.029

Wen, X., Luca, F., & Pique-Regi, R. (2015). Cross-population joint analysis of eQTLs: Fine mapping and functional annotation. *PLoS Genet.*, *11*(4), e1005176. https://doi.org/10.1371/journal.pgen.1005176

Wessels, H.-H., Méndez-Mancilla, A., Guo, X., Legut, M., Daniloski, Z., & Sanjana, N. E. (2020). Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0456-9

Westra, H.-J., Martínez-Bonet, M., Onengut-Gumuscu, S., Lee, A., Luo, Y., Teslovich, N., Worthington, J., Martin, J., Huizinga, T., Klareskog, L., Rantapaa-Dahlqvist, S., Chen, W.-M., Quinlan, A., Todd, J. A., Eyre, S., Nigrovic, P. A., Gregersen, P. K., Rich, S. S., & Raychaudhuri, S. (2018). Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet.* https://doi.org/10.1038/s41588-018-0216-7

Whalen, S., Truty, R. M., & Pollard, K. S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, *48*(5), 488–496. https://doi.org/10.1038/ng.3539

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., & Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035

Wienert, B., Wyman, S. K., Richardson, C. D., Yeh, C. D., Akcakaya, P., Porritt, M. J., Morlock, M., Vu, J. T., Kazane, K. R., Watry, H. L., Judge, L. M., Conklin, B. R., Maresca, M., & Corn, J. E. (2019). Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science*, *364*(6437), 286–289. https://doi.org/10.1126/science.aav9023

Wilczynski, B., Liu, Y.-H., Yeo, Z. X., & Furlong, E. E. M. (2012). Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput. Biol.*, *8*(12), e1002798. https://doi.org/10.1371/journal.pcbi.1002798

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). BAYESIAN MODEL SEARCH AND MULTILEVEL INFERENCE FOR SNP ASSOCIATION STUDIES. *Ann. Appl. Stat.*, *4*(3), 1342–1364.

Winter, J., Breinig, M., Heigwer, F., Brügemann, D., Leible, S., Pelz, O., Zhan, T., & Boutros, M. (2016). caRpools: An R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. *Bioinformatics*, *32*(4), 632–634. https://doi.org/10.1093/bioinformatics/btv617

Wiśniewski, J. R., Hein, M. Y., Cox, J., & Mann, M. (2014). A “proteomic ruler” for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics*, *13*(12), 3497–3506. https://doi.org/10.1074/mcp.M113.037309

Wu, H., Nord, A. S., Akiyama, J. A., Shoukry, M., Afzal, V., Rubin, E. M., Pennacchio, L. A., & Visel, A. (2014). Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet.*, *10*(9), e1004610. https://doi.org/10.1371/journal.pgen.1004610

Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., …

Xie, W. (2016). The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature*, *534*(7609), 652–657. https://doi.org/10.1038/nature18606

Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M. K., Shu, X.-O., Lu, Y., Cai, Q., Al-Ejeh, F., Rozali, E., Wang, Q., Dennis, J., Li, B., Zeng, C., Feng, H., Gusev, A., Barfield, R. T., … Zheng, W. (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.*

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P. D., Dadon, D. B., Cheng, A. W., Trevino, A. E., Konermann, S., Chen, S., Jaenisch, R., Zhang, F., & Sharp, P. A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.*, *32*(7), 670–676. https://doi.org/10.1038/nbt.2889

Xie, S., Duan, J., Li, B., Zhou, P., & Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell*, *66*(2), 285-299.e5. https://doi.org/10.1016/j.molcel.2017.03.007

Xinchen Wang, L. H., Sarah Goggin, Alham Saadat, Li Wang, Melina Claussnitzer, Manolis Kellis. (n.d.). *High-resolution genome-wide functional dissection of transcriptional regulatory regions in human*. https://doi.org/10.1101/193136

Xu, H., Xiao, T., Chen, C.-H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., Brown, M., & Liu, X. S. (2015). Sequence determinants of improved CRISPR sgRNA design. *Genome Res.*, *25*(8), 1147–1157. https://doi.org/10.1101/gr.191452.115

Xu, J., Carter, A. C., Gendrel, A.-V., Attia, M., Loftus, J., Greenleaf, W. J., Tibshirani, R., Heard, E., & Chang, H. Y. (2017). Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. *Nat. Genet.*, *49*(3), 377–386. https://doi.org/10.1038/ng.3769

Xu, M., Kladde, M. P., Van Etten, J. L., & Simpson, R. T. (1998). Cloning, characterization and expression of the gene coding for a cytosine-5-DNA methyltransferase recognizing GpC. *Nucleic Acids Res.*, *26*(17), 3961–3966.

Yan, B., Guan, D., Wang, C., Wang, J., He, B., Qin, J., Boheler, K. R., Lu, A., Zhang, G., & Zhu, H. (2017). An integrative method to decode regulatory logics in gene transcription. *Nat. Commun.*, *8*(1), 1044. https://doi.org/10.1038/s41467-017-01193-0

Yang, L., Zhu, Y., Yu, H., Chen, S., Chu, Y., Huang, H., Zhang, J., & Li, W. (2019). Linking genotypes with multiple phenotypes in single-cell CRISPR screens. In *BioRxiv*. https://doi.org/10.1101/658146

Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Liu, C., Freedman, J. E., Munson, P. J., Hill, D. E., Vidal, M., & Levy, D. (2017). Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *Am. J. Hum. Genet.*, *100*(4), 571–580. https://doi.org/10.1016/j.ajhg.2017.02.003

Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., Lim, J., Tai, E., Poh, H. M., Wong, E., Mulawadi, F. H., Sung, W.-K., Nicolis, S., Ahituv, N., Ruan, Y., & Wei, C.-L. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, *504*(7479), 306–310. https://doi.org/10.1038/nature12716

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., Packer, A., Darnell, R. B., & Troyanskaya, O. G. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, *51*(6), 973–980. https://doi.org/10.1038/s41588-019-0420-0

Zhou, Yan, & Zhu, Y. (2015). Diversity of bacterial manipulation of the host ubiquitin pathways. *Cell. Microbiol.*, *17*(1), 26–34. https://doi.org/10.1111/cmi.12384

Zhou, Yuexin, & Wei, W. (2016). Mapping regulatory elements. *Nat. Biotechnol.*, *34*(2), 151–152. https://doi.org/10.1038/nbt.3477

Zhou, Yuexin, Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., & Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature*, *509*(7501), 487–491. https://doi.org/10.1038/nature13166

Zhu, J., Yamane, H., Cote-Sierra, J., Guo, L., & Paul, W. E. (2006). GATA-3 promotes Th2 responses through three different mechanisms: Induction of Th2 cytokine production, selective growth of Th2 cells and inhibition of Th1 cell-specific factors. *Cell Res.*, *16*(1), 3–10. https://doi.org/10.1038/sj.cr.7310002

Zhu, Z., Lee, P. H., Chaffin, M. D., Chung, W., Loh, P.-R., Lu, Q., Christiani, D. C., & Liang, L. (2018). A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.*, *50*(6), 857–864. https://doi.org/10.1038/s41588-018-0121-0

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., & Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, *500*(7463), 477–481. https://doi.org/10.1038/nature12433

Zinngrebe, J., Montinaro, A., Peltzer, N., & Walczak, H. (2014). Ubiquitin in the immune system. *EMBO Rep.*, *15*(1), 28–45.