CRISPR-based functional genomics to study gene regulatory architecture and consequences of genetic variation

Florence Chardon

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Jay Shendure, Chair

Lea Starita

Heather Mefford

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

CRISPR-based functional genomics to study gene regulatory architecture and consequences of genetic variation

Florence Chardon

Chair of the Supervisory Committee:

Jay Shendure

Department of Genome Sciences

If we divide the human genome into its noncoding and coding parts, the noncoding portion takes up ~98-99% of the genome, and the coding portion takes up ~1-2% (ENCODE Project Consortium, 2012). The coding portion of the genome contains the genetic template for RNA transcription and protein translation, and translated proteins are the functional units of the cell that carry out biological function. Despite coding genes acting as templates for RNA transcription and protein translation, the noncoding genome contains the vast majority of sequences that regulate the expression of genes (some regulatory sequences reside in coding regions). The classes of sequences that make up regulatory sequences include promoters, enhancers, silencers, insulators, and repressors, amongst others. Proper regulation of genes at the right time and in the right cell types lies at the core of proper cellular function, and consequently, healthy cells, tissues, and

organisms. Variants such as single base substitutions, deletions, and insertions, can occur in both the coding and noncoding portions of the genome. These variants are most often benign. However, some mutations are loss-of-function mutations, meaning that they interfere with and sometimes inhibit proper biological function. Understanding the functional consequences of these mutations is critical to understanding the genetic causes of diseases as well as to develop therapeutics to treat diseases that have a genetic cause. To study both coding and noncoding regions of the genome, novel methods and technologies are required. In particular, the development of methods that can perturb and assess genetic sequences at scale are needed to tackle the vast number of genes and regulatory sequences that the genome consists of, which is on the scale of tens to hundreds of thousands (ENCODE Project Consortium, 2012).

In the first chapter, I introduce the field of DNA sequencing, genomics, and genomic technology development. I also discuss the various CRISPR/Cas9-based perturbation methods and how these methods can be utilized to develop novel screening methods, two of which I developed during my PhD. In the second chapter, I describe multiplex, single-cell CRISPR activation (CRISPRa) screening, a method I developed during my PhD and utilized to identify both proximal and distal cell type-specific regulatory elements that are capable of increasing target gene expression when activated via a specific CRISPRa perturbation. In the third chapter, I introduce a CRISPR prime editing based method that allows for the identification of drug resistance mutations in a multiplex and scaled manner. In the fourth and final chapter, I describe how I envision this work advancing further in order to gain a more comprehensive understanding of gene regulatory architecture and the functional consequences of genetic variation in all noncoding and coding sequences. I conclude

by discussing how we can use this understanding to design and develop novel and effective therapies for the wide range of diseases that have genetic causes.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Graduate school went by almost in the blink of an eye, but that's often the case for periods of time in life that are fulfilling, exciting, enjoyable, and memorable. I absolutely loved my five years and nine months as a graduate student in the Department of Genome Sciences. That is not to say it was easy or fun all of the time, as it most certainly was not. What might not be clear in this dissertation is the amount of time I spent working on experiments and projects that did not amount to any published text or figures. In fact, the first three years of my PhD yielded no publishable results, and everything described in this dissertation was produced in a time period of two and a half years.

The relative lack-of-success in my first few years of graduate school were difficult, and the subsequent string of successful projects in my later years was exciting, but also stressful as I suddenly found myself with too many experiments to do and too much computational work to do with the time I had in a given day. So, this is to say, that through the difficult times that were interspersed throughout incredibly gratifying and fun years, I have a lot of people to thank.

The first two, are my advisors Jay Shendure and Lea Starita. I am not sure that either of them knew how little I knew about genetics and genomics when I came into the program. In fact, my very first PCR reaction was during a rotation in my first year in the program. This is funny to think back to now, because I think I must have done thousands of PCR reactions since then. I also had never heard of the term "enhancer" before. Coming from an undergraduate chemistry degree and having never worked in a genetics lab, I had not heard about any regulatory sequences besides a promoter. I also did not come into Genome Sciences with the plan of joining a genetics lab – I only even found out about Genome Sciences because of the mass spectrometry expertise that is here! So,

thank you, to both of you, for taking me in as a student in your labs, and trusting me to do valuable and purposeful work.

Jay – you are always right. And even though you don't do experiments yourself anymore, you somehow still have the best experimental design strategies from anyone I consult. I learned early on to trust you and to do what you suggest, because I can point to a few key suggestions you made that completely saved large experiments and saved me significant amounts of time. I cannot decide if you are a more valuable scientific advisor, or a more valuable leader. You are both. I do hope to be in a position of leadership one day, and I have been trying to absorb as many of your leadership qualities as possible over the past few years. Great leaders are calm, supportive, and cultivate great environments with a good culture. Your lab is, by far, the best environment I have ever worked in. I do not think it is a coincidence that every member in the lab is happy, supportive, collaborative, and genuinely so nice. You cultivate that and you lead by example by being helpful, supportive, and never telling anyone that you are too busy for them. Thank you for giving me the opportunity to be a graduate student in your lab, for trusting me with your time and resources, and most of all, for your never ending support especially through the early years where experiments led to one useless dataset after another.

Lea – I was excited to work with you from the day I met you in the Vista Café to talk about a potential rotation in Jay's lab. You introduced me to single-cell RNA sequencing and even found a way for me to do this work during my rotation by leveraging your connections to get a hold of clinical samples to perform single-cell RNA sequencing on. There are many things I appreciate about you, but there are a few that I am extra grateful for. The first is your genuine interest in small

experimental details. I know you love seeing things like a simple PAGE gel with a pretty band on it. The second is your prioritization of my work, even when I did not have any significant updates to share with you. You have always taken a keen interest in all my experiments, and you were always curious to know how things were going. It's easy to feel like no one cares about your failing experiments, and while you may have not been so excited about them, I knew that you always cared and helped me brainstorm ways to push projects in more successful directions. The third is your support of me as a scientist, and your efforts to provide me with opportunities that are beneficial for my career. After a multi-year hiatus from traveling to conferences due to COVID-19, you supported me in attending the CRISPR and Beyond conference in Hinxton, UK. That was my first real conference talk, and I absolutely loved it. This is just one example of your support for me as a scientist, and I am so grateful for your support and endless encouragement.

As alluded to earlier, Jay fills his lab with the best scientists in the world, both from a scientific and personality perspective. There are, in my mind, three people who the lab could not run without. Riza Daza, Beth Martin, and Charlie Lee – thank you to all of you for your massive amounts of scientific discussions and help over the past five years. Riza – thank you for teaching me how sequencing works, how to prepare a sequencing library, and an extra huge thank you for helping me prep the >100 samples from the prime editing work that is discussed in Chapter 3. Beth – thank you for helping me troubleshoot every single cloning project I've undertaken in the lab. You know more about cloning and PCR than anybody, and I cannot tell you how lucky I feel to have you seconds away whenever I have a question. I do not know if I will get there, but your intuition for molecular biology is truly inspiring.

Troy McDiarmid is a postdoctoral fellow in the lab, and he joined in February of 2021. Shortly thereafter we started the CRISPR activation screening project together (discussed in Chapter 2), and this has been the most fruitful, productive, fun, and rewarding collaboration I have ever been a part of. Troy – thank you for being so incredibly easy to work with, for your open and honest communication, for teaching me so much from neuroscience to statistics to data visualization and everything in between. I feel so grateful to have you not only as a collaborator, but also a wonderful friend. You made a true impression on me during my time in graduation school, and I think you are a big reason why the second half of graduate school was so much better than the first. I cannot wait for you to open the McDiarmid lab, because you have all the qualities of a fantastic advisor. Thank you also for all the laughs, all the fun jokes, the fun memes, the many great conversations about science and life, and a big thank you for your friendship.

The other Shendure lab postdoctoral fellows who had a huge impact on my training as a scientist are Silvia Domcke, Xiaoyi Li, Jean-Benoît Lalanne, Diego Calderon, Will Chen, and Jacob Tome (former postdoctoral fellow). Each of you helped me think about science and experiments in thoughtful, critical, and detailed ways, and each of you had major positive influences on my work in the lab. To Silvia in particular, you are an excellent role model and mentor, and you are almost always the smartest person in a room. It does not go unnoticed.

The Shendure lab group of graduate students is the best it has ever been. Thank you to all of you for the daily laughs, the camaraderie, and for always having time for me! A special thank you to Anna Minkina (former graduate student), who was my first true friend in the lab, and who I talked to about anything and everything. Anna was the first person in the lab who made me feel like I did

belong here, and that I was a good scientist. Thank you for your persistent encouragement, and for the very special friendship.

I want to thank my committee members Cole Trapnell, Doug Fowler, and Heather Mefford. Thank you for your feedback, advice, and critical thinking of my projects to help steer them in the right direction. Cole, I am particularly thankful for your expertise in statistics and data analysis, which in turn made me a better data scientist.

To my family – Mutti, Papa, Nathalie, Christian, Alessandra, my new brother-in-law Patrick, and my beautiful Lorenzo and Cassidy Rose – you are all the shining lights in my own life, and I feel your support and love for me on a daily basis. Lorenzo and Cassidy Rose – you both arrived during my time in graduate school, and being your aunt is one of the greatest gifts in my life. Nathalie – thank you for always being by my side during graduate school; I think you are the only one who really understands all the nuances and goings-on of graduate school and I can't tell you how nice it is to know that you understand everything I am going through with little explanation needed.

Thank you to my lifelong best friend, Jessie Ditmore, who is always excited for me about whatever it is I decide to pursue. It can be graduate school, a cycling race, a trip, really anything, and Jessie will show me 110% support. Thank you for having organized a surprise visit from my two sisters and so many of our friends from San Francisco to show up in Seattle during an October weekend in 2017 without me having any idea! I have always thought that everybody in the world needs a Jessie, and I am so lucky to have you.

Lastly, and most importantly, thank you to my beautiful Kathleen. Without Kathleen I would have not applied to graduate school, and I would not have found the motivation to pursue a PhD. Kathleen has had a keen and fervent interest in basic molecular, cell, and systems biology since I met her in 2014, and her interest in science coupled with her unmatched ambition has had a major impact on my own scientific trajectory. Graduate school at times felt like the hardest thing I had ever pursued and having a partner in the same boat was invaluable. Kathleen, thank you for making me believe in myself as a scientist, thank you for always caring about my work, my experiments, and all my failures and successes, and thank you for always being there for me to lean on, especially in moments of panic when no projects were going well or in any good direction. Thank you for laughing with me in the lowest moments, such as when I would come home and say that I had spent all day (or week or month) doing an experiment that did not work and worse, I did not learn anything useful. I also need to thank you for getting me into the world of cycling back in 2015, which turned out to be the most wonderful hobby to focus my mind on outside of science these past few years. The completion of this PhD is in large part because of your influence on me, and I do not think I will ever be able to thank you enough.

DEDICATION

I dedicate this dissertation to my wonderful, loving, funny, and incredibly supportive father, Alain Gerard Chardon. My “super Papa” as we (his children) like to call him, has been excited about science since the moment I recall having memories. A trained physicist, my Dad has always been and is fascinated by the way things work. My whole family is very scientifically-minded, and growing up in this environment distilled in me a desire to also understand exactly how things work. To this day, understanding how molecular systems and human cells work is what drives my own motivation. Although my Dad loves physics and mathematics, I think he would have loved to tinker around with genes and contribute to the understanding of how genes and the human genome works.

Chapter 1. INTRODUCTION

"Progress in science depends on new techniques,
new discoveries and new ideas, probably in that order."
- Sydney Brenner

This quote by Sydney Brenner eloquently describes the evolution of the field of genomics. Genomics as a field is fueled by technology development, and the rate of technology development in genomics has not slowed down since the completion of the Human Genome Project (HGP) in 2003. The two genomic technologies that are foundational to the work described in this dissertation are next-generation sequencing and CRISPR-based genome engineering.

To begin, I will go back a few decades and cover key technological inventions and advances that have led to where we are today in terms of DNA sequencing and genome engineering. I will then discuss how CRISPR-based genome engineering and perturbation methods can be coupled with next-generation sequencing technologies to perform highly scaled, multiplex screens to interrogate the relationship between genotype and phenotype as it relates to gene regulatory architecture and genetic variation.

At the start of my graduate work, I was fascinated by the idea to develop experimental methods to understand this relationship between genotype and phenotype. In particular, I was interested in how genetic alterations cause disease, and I was fervently motivated to provide some new knowledge in this area. Throughout my graduate work, I gained a true appreciation for the deep

level of understanding required both of a biological system, and a specific methodology, to enable the development of genomic methods that truly work and generate data to make novel discoveries. I gained a huge amount of respect for the noise and variability that inherently exists both in biological systems and genomic technologies. I gained an even greater appreciation for devising and utilizing data analysis strategies that enable the identification of true biological signal from a sometimes frustratingly high noise floor.

1.1 DNA SEQUENCING: FROM MAXAM-GILBERT SEQUENCING TO NEXT GENERATION SEQUENCING

The explosive growth of the field of genomics can be attributed to two key scientific and technological achievements in the twenty-first century. The first was the ability to sequence DNA in a massively parallel manner, and the second was the completion of the Human Genome Project (HGP). Prior to these two landmark achievements, decades of work in the field of genetics and DNA sequencing allowed scientists to begin to uncover the sequence of letters that made up genes and other sequences through the very first DNA sequencing technologies.

The Maxam-Gilbert sequencing method, developed in the late 1970s by Allan Maxam and Walter Gilbert, was one of the earliest techniques used for DNA sequencing. This method employed chemical modification and cleavage of DNA to determine its sequence. The process involved four key steps. First, the DNA of interest was labeled with a radioactive or fluorescent marker. Next, the DNA sample was divided into four separate reactions. In each reaction, specific chemical reactions were carried out to modify the DNA at specific nucleotide positions, resulting in the creation of unique DNA fragments. After the modification step, the DNA was denatured and

subjected to gel electrophoresis, separating the fragments based on size. The fragments were visualized through autoradiography or fluorescence imaging, revealing the sequence ladder. By comparing the fragment sizes obtained from each reaction, the sequence of the DNA molecule could be deduced (Figure 1) (Maxam & Gilbert, 1977). Although laborious and very low throughput in comparison to the technologies that followed suit, this method of sequencing DNA laid the foundation for subsequent sequencing methods that were less laborious and higher throughput.

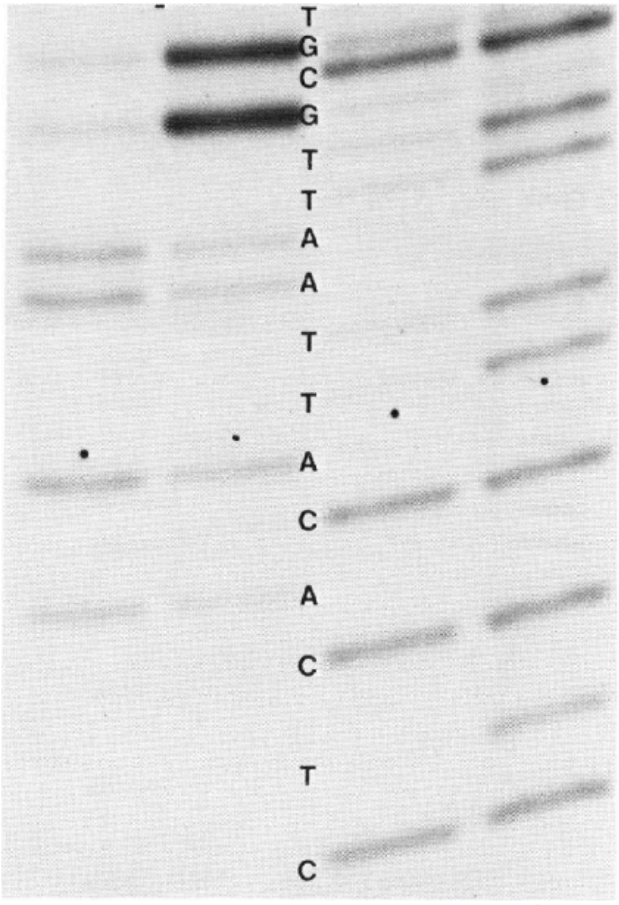

FIG. 3. Detail of the sequence gel. The four lanes are (from left to right) A > G, G > A, C, C + T; the dots show the position of the bromphenol blue dye marker, between fragments 9 and 10 long.

Figure 1. Figure 3 from Maxam, A. M., & Gilbert, W. (1977) (Maxam & Gilbert, 1977)**.**

After the Maxam-Gilbert sequencing method, the Sanger sequencing method, also known as chain-termination sequencing, emerged as the next major advancement in DNA sequencing. Developed by Frederick Sanger in the late 1970s, the Sanger sequencing method introduced a more efficient and reliable approach to DNA sequencing. Sanger sequencing relies on DNA replication using modified nucleotides called dideoxynucleotides (ddNTPs), which lack a 3' hydroxyl group necessary for further DNA chain elongation. In a sequencing reaction, a DNA template is replicated in the presence of normal nucleotides (dNTPs) and a small amount of labeled ddNTPs. The incorporation of a ddNTP at a specific position during DNA synthesis results in chain termination. By performing separate reactions with each of the four ddNTPs, fragments of varying lengths are generated. These fragments are then separated by gel electrophoresis, and the sequence is read based on the order of termination (Sanger et al., 1977).

Sanger sequencing quickly became the predominant method for DNA sequencing due to its reliability, ease of use, and ability to sequence longer DNA fragments. It played a crucial role in significant scientific breakthroughs, including the completion of the Human Genome Project. While Sanger sequencing was widely used for several decades, it was eventually superseded by next-generation sequencing (NGS) technologies, which enabled high-throughput sequencing and further advancements in genomic research.

The advent of next-generation sequencing (NGS) in the mid-2000s marked a paradigm shift in DNA sequencing technology (Levene et al., 2003; R. D. Mitra & Church, 1999; Robi D. Mitra et al., 2003; Shendure et al., 2005). NGS introduced a range of revolutionary approaches that dramatically increased sequencing capacity, speed, and cost-effectiveness. Instead of relying on

gel electrophoresis, NGS methods employ massively parallel sequencing, allowing millions of DNA fragments to be sequenced simultaneously. NGS technologies quickly transformed the field of genomics by enabling the sequencing of entire genomes, exomes (protein-coding regions), and targeted regions with unprecedented depth and breadth.

1.2 THE COMPLETION OF THE HUMAN GENOME PROJECT GIVES RISE TO THE FIELD OF GENOMICS

In 1990, the Human Genome Project (HGP) was launched, taking advantage of the advancing capabilities to sequence genes and DNA fragments. The HGP was an international collaborative effort to sequence and assemble the 3.2 billion nucleotides that make up the 23 chromosomes in the human genome and the project was completed in 2003. A few of the people who made major contributions to this project invlude Francis Collins, John Sulston, and Bob Waterston (Figure 2). This landmark achievement marked a turning point in genetics, providing a comprehensive map of the human genome and offering profound insights into our genetic makeup. Consequently, the field of genomics grew seemingly exponentially with the goal of understanding exactly how each nucleotide contributes to proper biological function in a given cell, tissue, and organ.

Figure 2. Robert (Bob) Waterston pipetting in the lab in the early 1990s.

The National Human Genome Research Institute funded his lab to sequence the worm genome, which was the first animal to ever be fully sequenced (Waterston & Sulston, 1995). His work greatly contributed to completing the Human Genome Project.

With a near-complete map of the human genome in hand as a result of the completion of the HGP, and the advent of next-generation sequencing, it was widely believed that we would soon understand the root genetic causes of diseases, and be able to finally realize the dream of "personalized medicine". The term personalized medicine refers to an approach to healthcare that tailors medical decisions and treatments to individual patients based on their unique genetic makeup. After all, with the human genome sequenced, it was initially thought that it would be easy to sequence additional genomes, of patients for example, and determine which bases in which genes differ, and consequently associate these changes with certain phenotypes, such as disease.

However, in reality, the sequencing of increasing numbers of human genomes identified a plethora of mutations (single nucleotide variants, insertions, deletions, translocations, transpositions, etc.) that had no known cause and no known phenotype. All *Homo sapiens* share 99.9% DNA sequence identity, but within that 0.1%, or 3.2 million nucleotides, of the 3.2 billion nucleotides in the human genome lies a lot of genetic variation. With increasing amounts of sequencing data came increasing amounts of unknown information and genetic sequences harboring variation within those sequences that had no annotation as to what, if any, changes in phenotype this variation resulted in. This is exemplified in the case of "variants of uncertain significance" (VUS), which are sequence variants that are not known whether they are pathogenic or benign (Figure 3). In brief, the sudden ease and availability of genetic sequencing created a mountain of sequencing data that was relatively uninterpretable due to our lack of understanding of what most genetic sequences do.

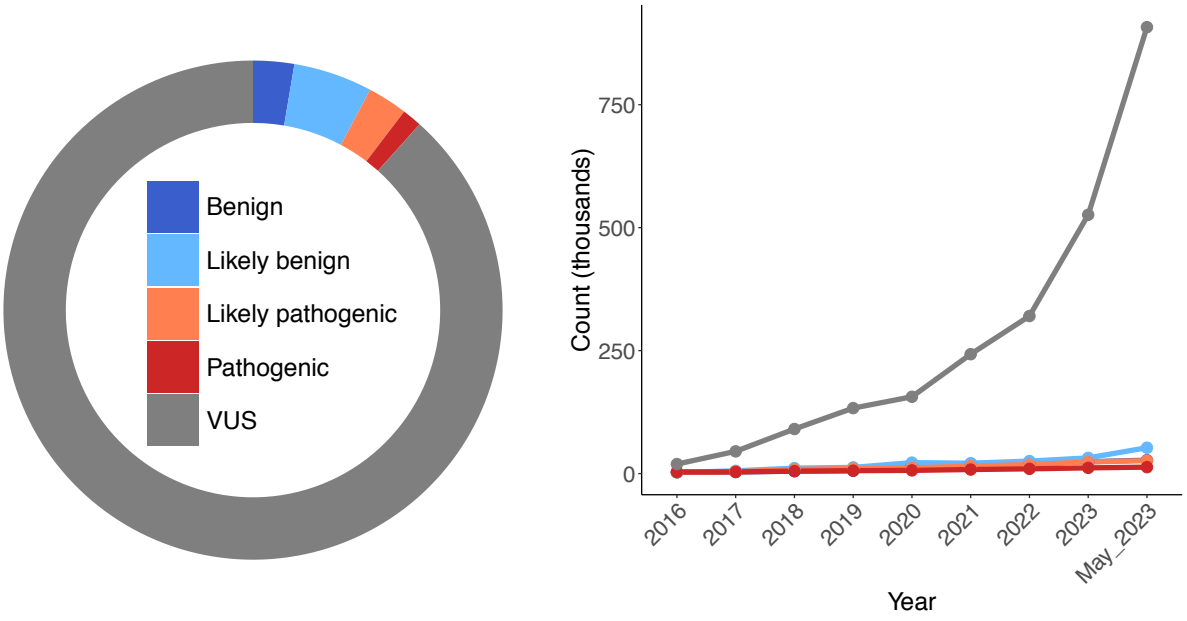

Figure 3. Updated Figure 1 from Fayer et al. (Fayer et al., 2021)

Missense variants of uncertain significance are a large and growing problem. Left: Single-nucleotide missense variants colored by ClinVar classifications (benign = 25,707; likely benign

= 16,377; VUSs = 227,365; likely pathogenic = 14,716; pathogenic = 22,489; conflicting interpretations = 20,026). ClinVar data downloaded on 10/27/2020. Right: Missense variants in ClinVar from 2015 to 2023 shown by clinical significance.

1.3 FUNCTIONAL GENOMIC TECHNOLOGIES

The increasing amount of genetic information that was becoming available due to our increasing ability to sequence genes and genomes and decreasing costs led to the development of functional genomics technologies that enable the annotation and interpretation of sequenced genomes. Functional genomics aims to uncover the functional aspects of the genome and decipher the complex interplay between genes, regulatory elements, and biological processes. In its early stages, functional genomics relied on pioneering genomic technologies and functional assays to gain initial insights into genome function. For instance, early methods like reporter gene assays and DNA microarrays played crucial roles in examining gene expression patterns and identifying regulatory elements. Reporter gene assays involved fusing a target gene to a reporter gene, allowing researchers to measure the activity of the target gene based on the expression of the reporter (Patwardhan et al., 2009). DNA microarrays enabled the simultaneous monitoring of gene expression levels for thousands of genes (Schena et al., 1995). These technologies provided the foundation for understanding gene function and regulation and facilitated a deeper understanding of the roles and interactions of genomic elements and their impact on cellular processes and disease.

Next generation sequencing-based assays allowed for interrogation of DNA sequences on a much larger scale. For instance, chromatin immunoprecipitation followed by sequencing (ChIP-seq)

allows researchers to identify protein-DNA interactions on a genome-wide scale (Johnson et al., 2007). Similarly, RNA sequencing (RNA-seq) enables comprehensive profiling of the transcriptome, allowing the quantification and characterization of gene expression levels (Bainbridge et al., 2006). Additionally, genome-wide association studies have enabled the identification of genetic variants associated with complex traits and diseases (R. J. Klein et al., 2005). Furthermore, more recent technologies like single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics provide unprecedented resolution and allow the investigation of gene expression patterns at the single-cell level and within specific tissue contexts (Cao et al., 2017; Gierahn et al., 2017; Hashimshony et al., 2012; Islam et al., 2011; A. M. Klein et al., 2015; Lubeck & Cai, 2012; Macosko et al., 2015; Srivatsan et al., 2021; Ståhl et al., 2016; Tang et al., 2009).

1.4 PERTURBING THE GENOME WITH CRISPR/CAS9-BASED GENOME EDITING TOOLS

Prior to the discovery of the CRISPR/Cas9 genome editing method, researchers used TALENs (Transcription Activator-Like Effector Nucleases) and zinc finger proteins to edit DNA. TALENs and zinc finger proteins are two classes of engineered DNA-binding proteins that can function as molecular scissors capable of targeting specific DNA sequences for modification. TALENs are created by fusing a DNA-binding domain derived from TAL effectors with a nuclease domain, such as FokI. The DNA-binding domain can be designed to recognize specific DNA sequences, enabling precise targeting (Boch et al., 2009). Similarly, zinc finger proteins are engineered by linking specific zinc finger domains, each recognizing a specific DNA triplet, with a nuclease domain. Once the engineered proteins bind to their target DNA sequence, the nuclease domain induces double-stranded breaks at the target site. These breaks can trigger DNA repair processes,

such as non-homologous end joining or homology-directed repair, leading to gene knockout or insertion of desired DNA sequences (Bibikova et al., 2002). TALENs and zinc finger proteins laid the groundwork for the development of more advanced genome editing technologies.

Shortly after the first commercially available next-generation sequencing platform (from 454 Life Sciences in 2005), two researchers, Jennifer Doudna and Emmanuelle Charpentier recognized the potential of the CRISPR-Cas immune system in bacteria as a gene-editing tool in the early 2010s (Jinek et al., 2012). They demonstrated that by reprogramming the CRISPR-associated protein 9 (Cas-9) enzyme, it could be guided by a short RNA molecule to target specific DNA sequences for cleavage (Figure 4). This breakthrough discovery laid the foundation for harnessing CRISPR/Cas-9 as a powerful genome editing system.

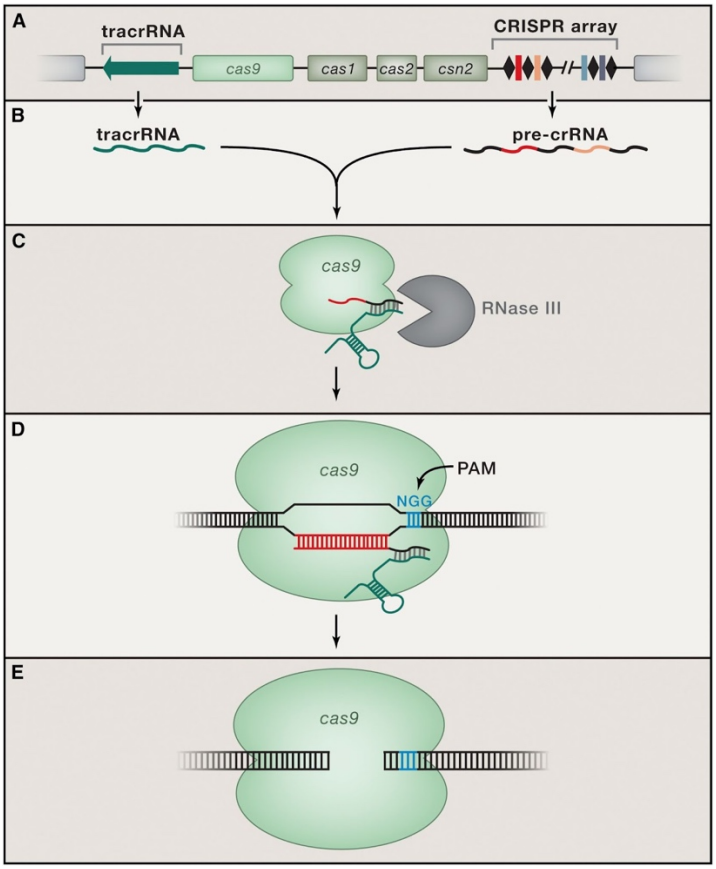

Figure 4. Figure 1 from Lander (2016) (Lander, 2016)**.**

Class 2, Type II CRISPR-Cas9 System from *Streptococcus thermophilus.* Type II systems are the simplest of the three types of CRISPR systems and have been the basis for genome editing technology. A) The locus contains a CRISPR array, four protein-coding genes (*cas9*, *cas1*, *cas2*, and *cns2*) and the tracrRNA. The CRISPR array contains repeat regions (black diamonds) separated by spacer regions (colored rectangles) derived from phage and other invading genetic elements. The *cas9* gene encodes a nuclease that confers immunity by cutting invading DNA that matches existing spacers, while the *cas1*, *cas2*, and *cns2* genes encode proteins that function in the acquisition of new spacers from invading DNA. B) The CRISPR array and the tracrRNA are transcribed, giving rise to a long pre-crRNA and a tracrRNA. C) These two RNAs hybridize via complementary sequences and are processed to shorter forms by Cas9 and RNase III. D) The resulting complex (Cas9 + tracrRNA + crRNA) then begins searching for the DNA sequences that match the spacer sequence (shown in red). Binding to the target site also requires the presence of the protospacer adjacent motif (PAM), which functions as a molecular handle for Cas9 to grab on to. E) Once Cas9 binds to a target site with a match between the crRNA and the target DNA, it cleaves the DNA three bases upstream of the PAM site. Cas9 contains two endonuclease domains, HNH and RuvC, which cleave, respectively, the complementary and non-complementary strands of the target DNA, creating blunt ends.

In its early applications, CRISPR/Cas-9 showcased its remarkable potential in diverse fields. Researchers quickly adopted this technology to edit the genomes of various organisms, from bacteria to plants and mammalian cells (Cong et al., 2013; Mali et al., 2013). It provided a simpler and more efficient alternative to previous gene-editing methods. Its versatility and ease of use

made it accessible to scientists worldwide, propelling a flurry of groundbreaking research that quickly followed. Two such research breakthroughs are CRISPR-based perturbation methods that don't directly edit target DNA, but rather change the expression of desired genes. CRISPR interference and CRISPR activation are perturbation methods that allow for the repression or activation of desired genes by fusing a nuclease-dead Cas9 (dCas9) to a transcriptional repressor or activator (Bikard et al., 2013; Maeder et al., 2013; Qi et al., 2013). This dCas9 repression or activation complex is then brought to either promoter or enhancer sequences to regulate the expression of desired genes.

The invention of CRISPR base editing marked another significant milestone in genome editing. Developed by David Liu and his lab, CRISPR base editing expanded the scope of CRISPR technology beyond the introduction of double-stranded breaks and DNA repair (Komor et al., 2016). Base editing allows for precise modifications of single DNA bases without the need for DNA cleavage. By coupling a catalytically inactive Cas-9 protein with a cytidine deaminase enzyme, researchers could convert cytosine (C) to uracil (U) at targeted genomic positions. The cell's natural DNA repair mechanisms would then convert the modified base to thymine (T), resulting in permanent base changes. This technique offered a more controlled and efficient approach for introducing point mutations.

David Liu's lab further advanced the field of CRISPR-based genome editing with the invention of CRISPR prime editing (Anzalone et al., 2019). CRISPR prime editing expands the editing capabilities beyond simple base conversions by allowing the precise insertion, deletion, or replacement of DNA sequences. It combines a catalytically impaired Cas-9 variant with a reverse

transcriptase enzyme and a prime editing guide RNA (pegRNA) (Figure 5). The pegRNA not only directs the Cas-9 complex to the target site but also carries the desired edit information as a template. The system enables precise editing to the target DNA sequence. CRISPR prime editing represents a significant step forward in genome editing, offering a lot of promise for the correction of disease-causing mutations and the engineering of specific genetic changes with enhanced precision and versatility.

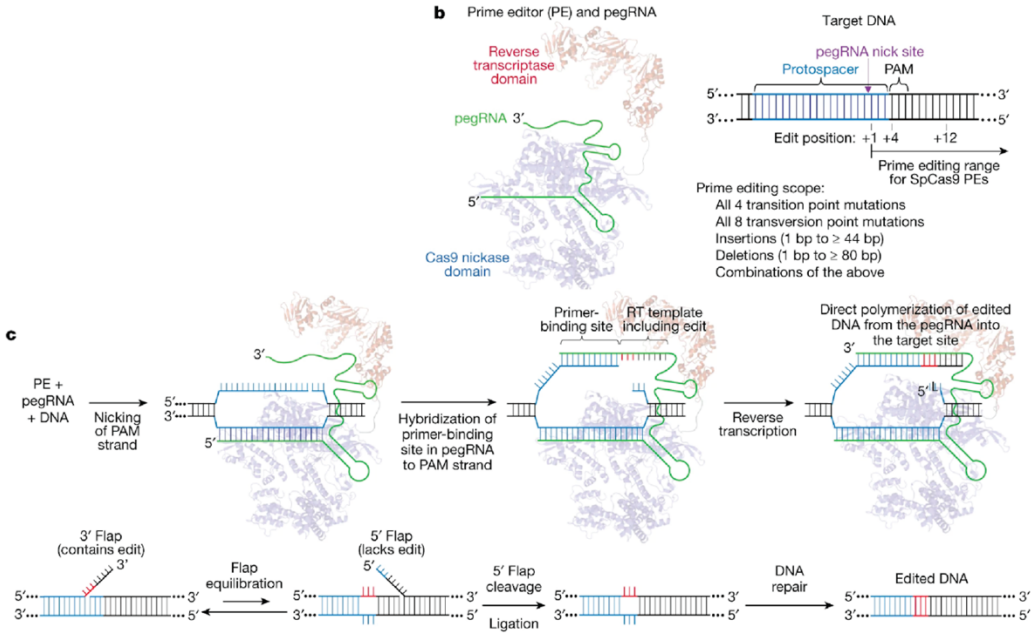

Figure 5. Figure 1b, c from Anzalone et al. (2019) (Anzalone et al., 2019)**.**

b) A prime editing complex consists of a PE protein containing an RNA-guided DNA-nicking domain, such as Cas9 nickase, fused to an RT domain and complexed with a pegRNA. The PE–pegRNA complex enables a variety of precise DNA edits at a wide range of positions. spCas9, *Streptococcus pyogenes* Cas9. c) The PE–pegRNA complex binds the target DNA and nicks the PAM-containing strand. The resulting 3′ end hybridizes to the PBS, then primes reverse transcription of new DNA containing the desired edit using the RT template of the pegRNA.

Equilibration between the edited 3′ flap and the unedited 5′ flap, cellular 5′ flap cleavage and ligation, and DNA repair results in stably edited DNA.

1.5 COUPLING CRISPR/CAS9-BASED GENOME EDITING WITH HIGH THROUGHPUT DNA SEQUENCING-BASED ASSAYS

The coupling of CRISPR-based editing methods with high-throughput sequencing assays allows us to perturb the genome and comprehensively read out its functional consequences via sequencing. CRISPR-based editing and perturbation techniques, such as CRISPR-Cas9, CRISPRi, CRISPRa, base editing, and prime editing, offer precise and efficient means to introduce targeted modifications into the genome. By leveraging these editing and perturbation methods in conjunction with high-throughput sequencing assays, we can perturb specific genetic elements and assess the resulting functional changes on a genome-wide scale. This integrative approach enables identification of regulatory elements, characterization of gene expression profiles, and assessment of the impact of genetic variants (Fulco et al., 2016; Gasperini et al., 2019, Hanna et al., 2021; Ursu et al., 2022). The combination of CRISPR-based editing and perturbation and high-throughput sequencing provides a powerful framework for dissecting the functional architecture of the genome and unraveling the complex mechanisms underlying gene regulation and cellular processes. It offers potential for advancing our understanding of the genome, diseases, discovering therapeutic targets, and developing novel precision medicine strategies.

1.6 CRISPR SCREENING FOR MEDICAL AND RARE DISEASE GENETICS

Chapter 1.6 is an excerpt from:

Christoph Bock, Paul Datlinger, **Florence Chardon**, Matthew A. Coelho, Matthew B. Dong, Keith A. Lawson, Tian Lu, Laetitia Maroc, Thomas M. Norman, Bicna Song, Geoff Stanley, Sidi Chen, Mathew Garnett, Wei Li, Jason Moffat, Lei S. Qi, Rebecca S. Shapiro, Jay Shendure, Jonathan S. Weissman, Xiaowei Zhuang. "High-content CRISPR screening." *Nature Reviews Methods Primers*, **2**, 8 (2022).

CRISPR screening facilitates the annotation of disease-linked genes and genetic variants, specifically by assessing the biological function of variants of uncertain significance in disease-causing genes (Shendure & Fields, 2016). Deleterious genetic variants in *BRCA1* are a risk factor of breast cancer with high clinical relevance, yet many variants are too rare to assess with confidence based on medical genetics data whether the variant is pathogenic or benign. This challenge has been tackled by saturation genome editing**,** using CRISPR and homology-directed repair to introduce several thousand single-nucleotide variants into the *BRCA1* locus and measuring their effect in vitro (Findlay et al., 2018). Further, cytosine base editors have been used to assess genetic variants in *BRCA1* and *BRCA2* (Hanna et al., 2021; Kweon et al., 2020) and a recent study extended this approach to 86 genes involved in the DNA damage response (Cuella-Martin et al., 2021).

Not all genetic variants can be targeted with base editors owing to their sequence specificity including the need for a Cas-specific PAM sequence close to the target site. Nevertheless, a gRNA

library has been developed that targets more than 50,000 ClinVar variants with base editing (Hanna et al., 2021), and CRISPR prime editing promises to provide even more flexibility to engineer a wide range of genetic variants across the genome (Anzalone et al., 2019).

CRISPR screening is useful for the functional analysis of genetic variants associated with polygenic diseases, including risk alleles identified through genome-wide association studies (GWAS) and population genome sequencing. Such studies have statistically linked thousands of genomic regions to a wide range of diseases and human phenotypes, although their rate of pinpointing causal variants and underlying mechanisms has been low. CRISPR screens can complement genetic association studies by testing the biological function of a large number of genetic variants in parallel before labor-intensive investigation of individual variants. In one such screen, gRNAs were tiled across the *BCL11A* enhancer to map which parts of this enhancer are associated with fetal hemoglobin expression, which may be therapeutically relevant for β-thalassemia and sickle cell anemia (Canver et al., 2015). Another study applied CRISPRi screens with a FISH readout of target gene expression to quantify gene-regulatory effects for thousands of candidate enhancer-gene pairs (Fulco et al., 2019).

A challenge of using CRISPR screens for assaying genetic variants is the need to develop and validate specific reporter assays for each gene or phenotype of interest. This can be avoided with scCRISPR-seq, exploiting the versatility of transcriptional profiles as correlates of diverse cellular phenotypes. For example, CRISPRi followed by a single-cell RNA-seq readout was used to measure the effect of enhancer silencing on the transcriptome (Xie et al., 2017) and to obtain single-cell profiles for the effect of several thousand enhancers in a single experiment (Gasperini

et al., 2019). This approach is capable of linking genetic variants to the genes they regulate, which is an important task given that most disease-linked genetic variants identified by GWAS lie in non-coding genomic regions.

~ ~ ~

In this dissertation work, I will describe two novel approaches that utilize the combination of CRISPR screening and next-generation sequencing to 1) study gene regulatory architecture (including both promoter and enhancer elements) controlling the expression of haploinsufficient genes (Chapter 2) and 2) study mutations that give rise to drug resistance in the context of lung cancer (Chapter 3). I will conclude this dissertation work by reflecting on the state of high throughput functional genomics technologies today, and where I envision this field going in the next few years and decades, both in terms of technology development, and in terms of biological discovery and therapeutic applications (Chapter 4).

Chapter 2. MULTIPLEX, SINGLE-CELL CRISPR ACTIVATION SCREENING FOR THE IDENTIFICATION OF CELL TYPE SPECIFIC REGULATORY ELEMENTS

Chapter 2 has been adapted with minimal modification from:

Florence M. Chardon, Troy A. McDiarmid, Nicholas F. Page, Beth Martin, Silvia Domcke, Samuel G. Regalado, Jean-Benoît Lalanne, Diego Calderon, Lea M. Starita, Stephan J. Sanders, Nadav Ahituv, and Jay Shendure. "Multiplex, single-cell CRISPRa screening for cell type specific regulatory elements." *Nature Biotechnology (in revision)*, (2023).

2.1 ABSTRACT

CRISPR-based gene activation (CRISPRa) is a promising therapeutic approach for gene therapy, upregulating gene expression by targeting promoters or enhancers in a tissue/cell-type specific manner. Here, we describe an experimental framework that combines highly multiplexed perturbations with single-cell RNA sequencing (sc-RNA-seq) to identify cell-type-specific, CRISPRa-responsive *cis*-regulatory elements and the gene(s) they regulate. Random combinations of many gRNAs are introduced to each of many cells, which are then profiled and partitioned into test and control groups to test for effect(s) of CRISPRa perturbations of both enhancers and promoters on the expression of neighboring genes. Applying this method to candidate *cis*-regulatory elements in both K562 cells and iPSC-derived excitatory neurons, we identify gRNAs capable of specifically and potently upregulating target genes, including autism spectrum disorder

(ASD) and neurodevelopmental disorder (NDD) risk genes. A consistent pattern is that the responsiveness of individual enhancers to CRISPRa is restricted by cell type, implying a dependency on either chromatin landscape and/or additional *trans*-acting factors for successful gene activation. The approach outlined here may facilitate large-scale screens for gRNAs that activate therapeutically relevant genes in a cell type-specific manner.

2.2 INTRODUCTION

There are millions of candidate *cis*-regulatory elements (cCREs) in the human genome, yet only a handful have been functionally validated and confidently linked to their target gene(s) (Gasperini et al., 2020). Recently, we and others have combined CRISPR-interference (CRISPRi) and sc-RNA-seq to scalably validate distal cCREs, while also linking them to the gene(s) that they regulate (Fulco et al., 2016; Gasperini et al., 2019, 2020; Xie et al., 2017). However, to date, the vast majority of work in the field has focused on screening candidate regulatory elements for *necessity*, with only a few studies screening for *sufficiency* in the endogenous context.

CRISPR-activation (CRISPRa) is a versatile approach that allows one to test for the sufficiency of cCRE activity (Gilbert et al., 2014; Konermann et al., 2014; Schmidt et al., 2022; Tian et al., 2021). CRISPRa screens of noncoding regulatory elements have at least four potential advantages over CRISPRi screens. First, as noted above, CRISPRa can identify cCREs that are sufficient even if not singularly necessary to drive target gene expression. Second, CRISPRa can identify elements that, when targeted, may upregulate already active genes above their baseline levels. Third, CRISPRa has the potential to discover inactive regions that, when transcriptional activation machinery is recruited to them, can act as active enhancers and increase expression of nearby genes

(Simeonov et al., 2017). Finally, CRISPRa has the potential to identify cCRE-targeting gRNAs whose activity is cell type-specific, opening the door to "*cis* regulatory therapy" (CRT) for haploinsufficient and other low-dosage associated disorders, as recently demonstrated for monogenic forms of obesity and autism spectrum disorder (Matharu et al., 2019; Tamura et al., 2022). However, despite these potential advantages, CRISPRa targeting of noncoding regulatory elements has mostly been deployed in an *ad hoc* manner (Dai et al., 2021; Joung et al., 2017; Simeonov et al., 2017; Tak et al., 2021), and typically in workhorse cancer cell lines rather than more therapeutically relevant *in vitro* models.

Here, we present a scalable framework in which we introduce multiple, random combinations of CRISPRa perturbations to each of many cells followed by sc-RNA-seq (Figure 6), analogous to an approach that we previously developed for CRISPRi screening (Gasperini et al., 2019; Xie et al., 2017). Computational partitioning of cells into test and control groups based on detected gRNAs enables greater power than single-plex CRISPRa screens, as any given single-cell transcriptome is informative with respect to multiple perturbations (Gasperini et al., 2019). In this proof-of-concept study, we performed two screens in which the same set of cCREs was targeted, first in K562 cells and then in human iPSC-derived excitatory neurons. We discover both enhancer and promoter-targeting gRNAs capable of mediating upregulation of target gene(s). For enhancers in particular, the upregulatory potential of individual gRNAs was consistently restricted to one cell type, implying a dependency on either the *cis* chromatin landscape and/or additional *trans*-acting factors for successful gene activation.

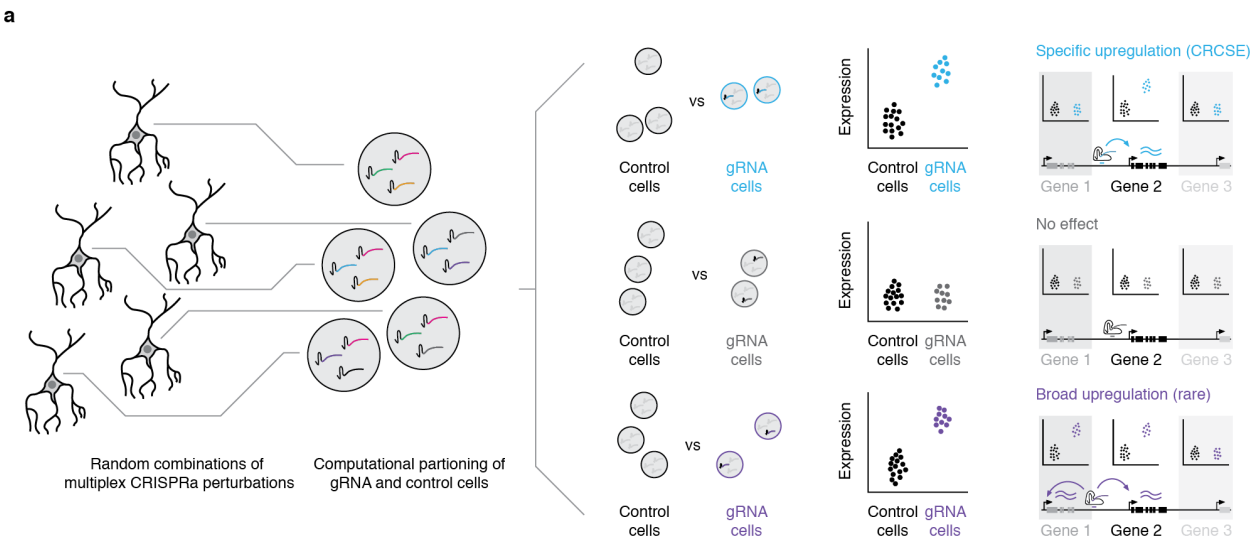

Figure 6. Multiplex, single cell CRISPRa screening for cell type-specific regulatory elements.

(Left) A library of gRNAs targeting candidate cis-regulatory elements (cCREs) is introduced in a multiplex fashion to a population of cells expressing CRISPRa machinery, such that each cell contains a random combination of multiple CRISPRa-mediated perturbations. (Middle) Following single cell transcriptional profiling and gRNA assignment, cells are systematically computationally partitioned into those with or without a given gRNA and tested for upregulation of neighboring genes. (Right) CRISPRa perturbations can either result in target-specific upregulation, no detectable effect (*e.g.*, for non-targeting controls) or, at least theoretically, broad *cis*-upregulation of multiple genes in the vicinity of the gRNA/CRISPRa machinery. Furthermore, patterns of upregulation can either be general or cell type-specific.

2.3 MULTIPLEX SINGLE-CELL CRISPRA SCREENING OF REGULATORY ELEMENTS IN K562 CELLS

As a proof of principle, we first sought to implement multiplex single-cell CRISPRa screening in the chronic myelogenous leukemia cell line K562, an ENCODE Tier 1 cell line (Zhou et al., 2019)

in which we had previously performed a multiplex CRISPRi screen (Gasperini et al., 2019). Our proof-of-concept library included gRNAs targeting transcription start site (TSS) positive controls (30 gRNAs), candidate promoters (313 gRNAs), candidate enhancers (100 gRNAs) and non-targeting controls (NTCs; 50 gRNAs). The 30 TSS positive control gRNAs were selected from a previously reported hCRISPRa-v2 library (Horlbeck et al., 2016), while the 313 candidate promoter-targeting gRNAs were designed to 50 annotated TSSs of 9 high-confidence haploinsufficient risk genes associated with ASD and NDD (*BCL11A*, *TCF4*, *ANK2*, *CHD8*, *TBR1*, *SCN2A*, *SYNGAP1*, *FOXP1*, and *SHANK3*) that are potential therapeutic targets for CRT (Matharu & Ahituv, 2020). The candidate enhancer-targeting guides included 50 gRNAs designed to target 25 enhancer hits previously validated by CRISPRi (Gasperini et al., 2019), as well as 50 gRNAs designed to target 25 enhancer "non-hits" (*i.e.* sequences with biochemical markers strongly predictive of enhancer activity in K562 cells that did not alter gene expression when targeted with CRISPRi (Gasperini et al., 2019)) (Figure 7a, b; Methods). We cloned this gRNA library (n=493) into piggyFlex, a piggyBac transposon-based gRNA expression vector, to allow for genomic integration and stable expression of gRNAs (Lalanne et al., 2022). The piggyFlex vector has both antibiotic (puromycin) and fluorophore (GFP) markers, enabling flexibly stringent selection for cells with higher numbers of gRNA integrants. Additionally, this vector design allows for gRNA transcript capture during single-cell library preparation (Lalanne et al., 2022) (Figure 7c).

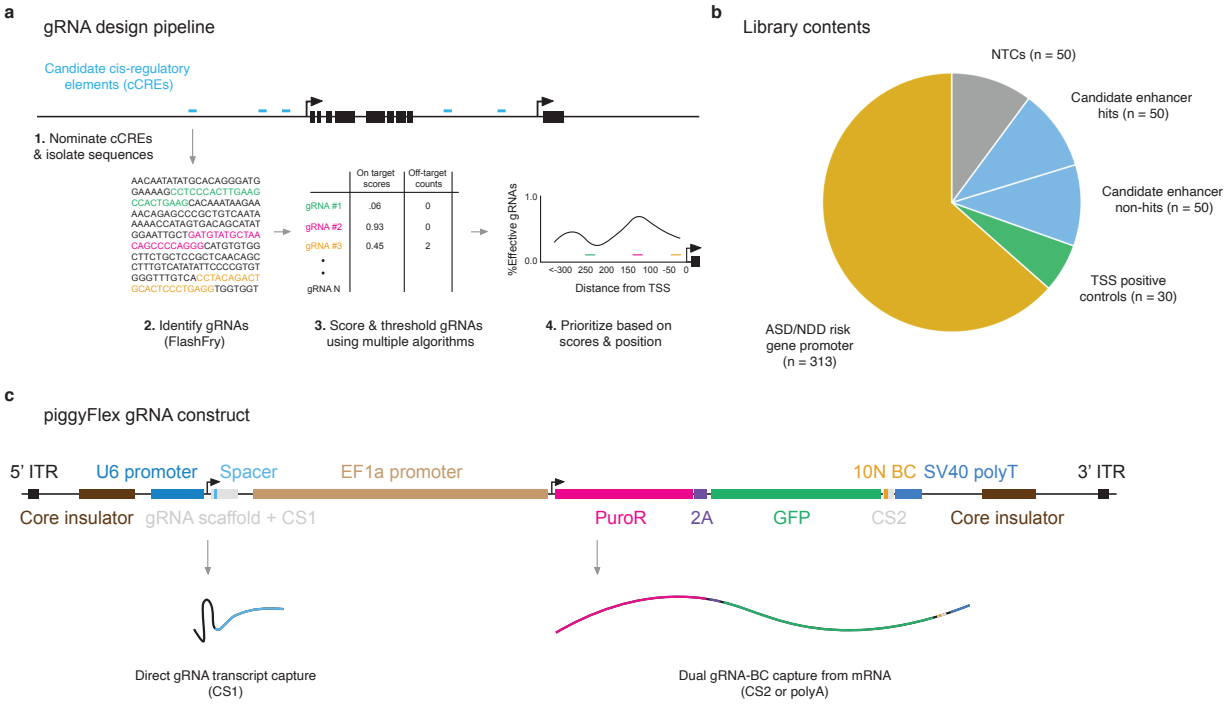

Figure 7. gRNA design pipeline, library contents, and piggyFlex gRNA delivery construct.

a) gRNA design pipeline. First, candidate Cis-Regulatory Elements (cCREs) surrounding a gene of interest were identified based on biochemical marks of regulatory activity (*e.g.*, accessibility, active transcription, etc.). Next, candidate gRNAs targeting each cCRE were generated using FlashFry (McKenna & Shendure, 2018). Then, gRNAs were scored and prioritized using multiple algorithms. Finally, in the case of promoters where systematic CRISPRa design rules are available, gRNAs were prioritized based on optimal position relative to the TSS (Sanson et al., 2018). b) PiggyFlex gRNA library contents by target category. c) PiggyFlex construct design. PiggyFlex is a piggyBac transposon-based gRNA delivery vector equipped with a dual antibiotic (puromycin) and fluorophore (GFP) selection cassette that enables enrichment for cells with many integrated gRNAs (Lalanne et al., 2022). PiggyFlex enables direct capture of gRNA transcripts or optional capture of gRNA-associated barcodes from GFP mRNA via CS2 or polydT capture.

There is no consensus on which CRISPRa activation complex is best suited for broad and scalable targeting of enhancers (Tak et al., 2021). We therefore tested both the VP64 activation complex, which consists of four copies of the VP16 effector, and the VPR activation complex, which consists of the VP64 effector fused to the p65 and Rta effectors (Chavez et al., 2015; Maeder et al., 2013). The VPR complex has been shown to lead to higher levels of transcriptional activation than that of the VP64 complex. However, this increased upregulation could achieve higher than therapeutically needed expression levels and being much larger than VP64 could impinge on packaging and delivery of gene therapy vectors such as adeno associated virus (AAV) (Chavez et al., 2015). We generated a monoclonal, stably VP64-expressing K562 cell line, purchased a polyclonal, stably VPR-expressing K562 cell line (Figure 9a; Methods), and validated the capacity of these lines for CRISPRa with a minimal cytomegalovirus (CMV) promoter-tdTomato reporter expression assay (Esvelt et al., 2013) (Figure 8).

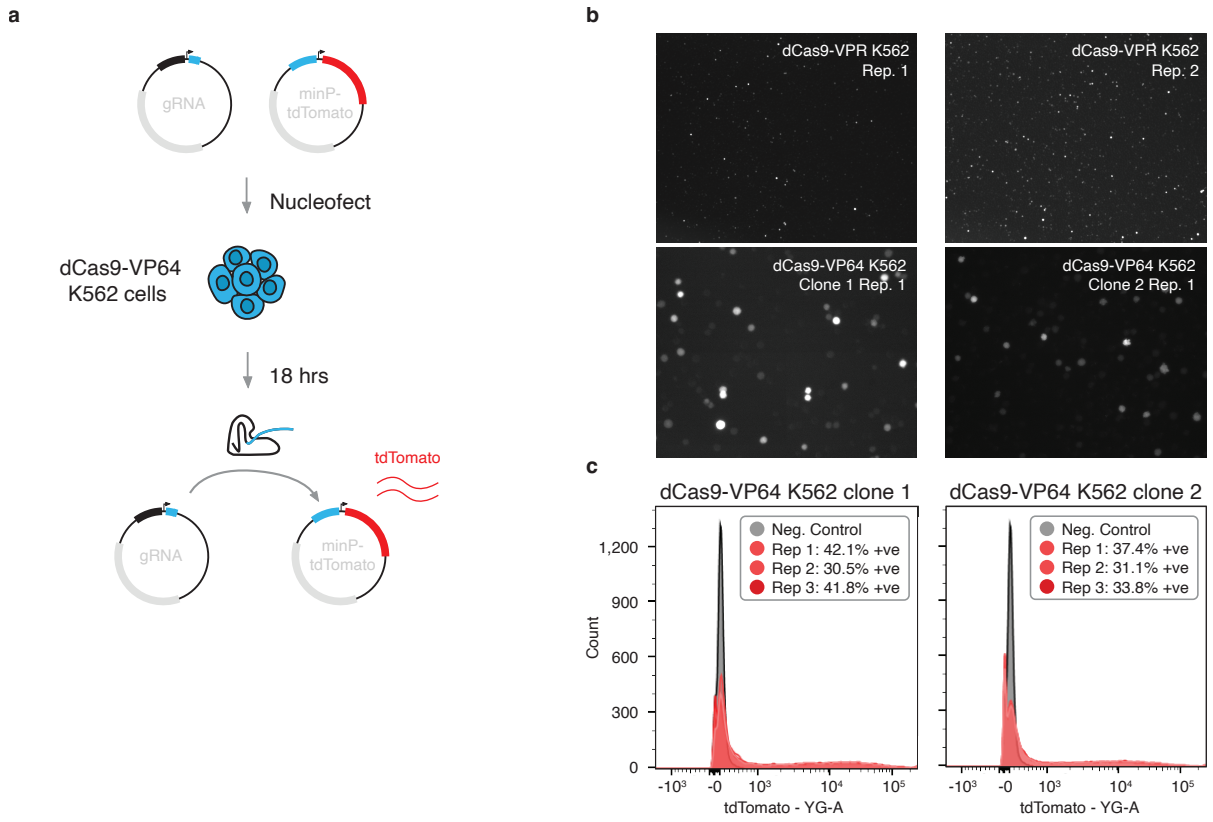

Figure 8. Functional validation of CRISPRa K562 cell lines.

a) Schematic of the minP-tdTomato functional assay used to validate CRISPRa cell lines. Two plasmids, one encoding a minP-tdTomato and another encoding a gRNA targeting a sequence immediately upstream of minP were co-nucleofected into K562 cell lines with either dCas9-VP64 or dCas9-VPR constructs integrated (only dCas9-VP64 is illustrated for simplicity). b) Following nucleofection, both dCas9-VPR (top, lower magnification) and dCas9-VP64 (bottom, higher magnification) K562 cell lines drove strong tdTomato expression, confirming the presence of functional CRISPRa machinery in these cell lines. dCas9-VPR images represent two replicate transfections into a single monoclonal line, while dCas9-VP64 images each represent one transfection replicate from two monoclonal lines. Note these are transient transfections without selection, so not all cells are expected to have been successfully transfected and fluoresce under

these conditions. c) Example FACS analysis of tdTomato fluorescence in individual dCas9-VP64 transfection replicates of two monoclonal lines.

We then transfected the gRNA library and piggyBac transposase into each cell line at a 20:1 library-to-transposase ratio to achieve high multiplicity of integration (MOI), and selected cells with puromycin. Cells were cultured for nine days before harvesting for sc-RNA-seq to capture and assign gRNAs to single cell transcriptomes (Figure 7; Figure 9a). After QC filtering, we recovered 33,944 high-quality single-cell transcriptomes across the two cell lines, with 79% of cells having one or more detected gRNAs. We recovered a mean of 2.5 gRNAs per cell (Figure 9b) and 178 cells per gRNA (Figure 9c). Transcriptome quality, MOI, gRNA assignment rate, and gRNA coverage were similar across all four sc-RNA-seq batches (10x Genomics lanes) as well as the two cell lines tested (Figure 10).

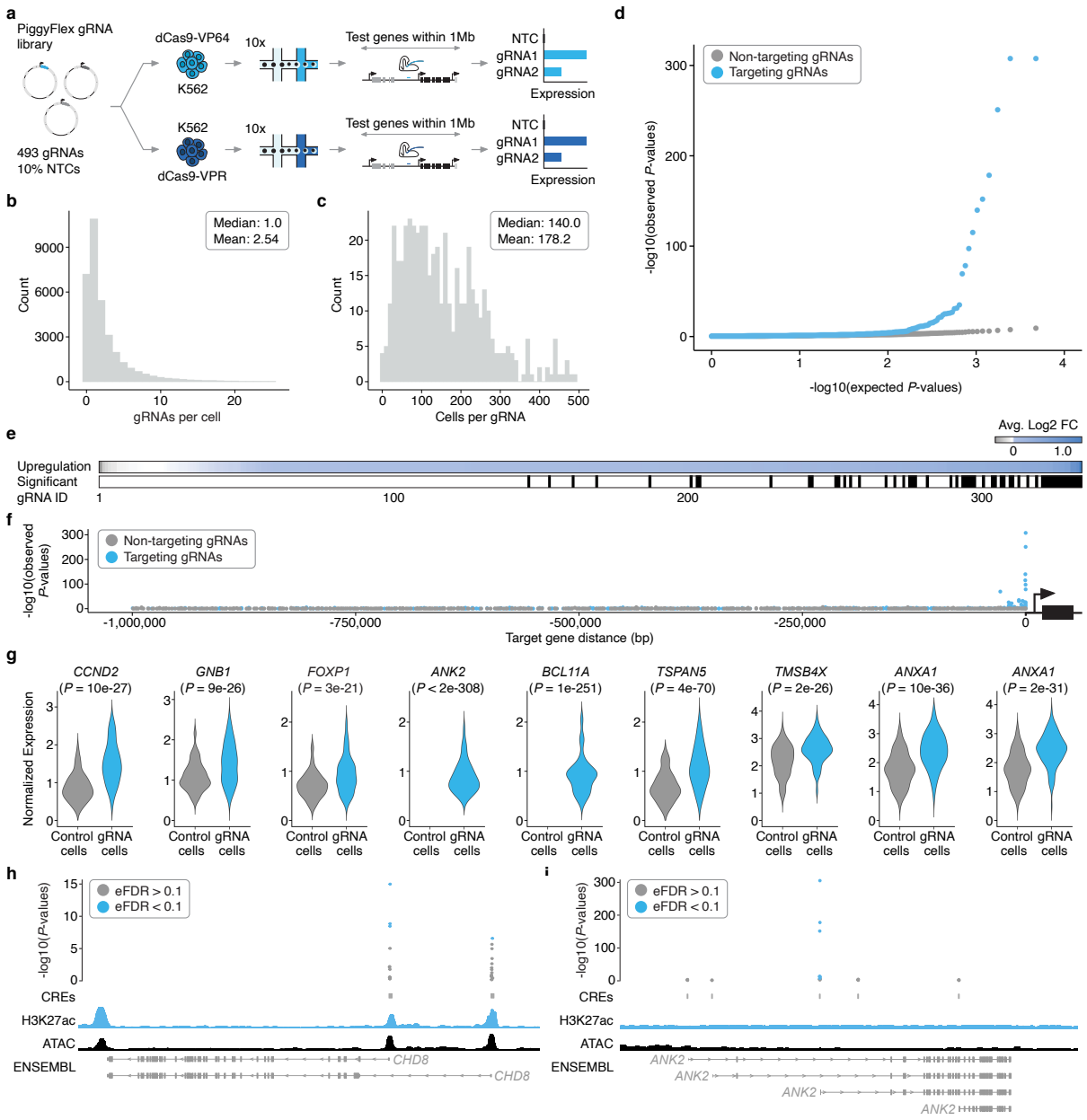

Figure 9. Multiplex single cell CRISPRa screening of regulatory elements in K562 cells.

a) A piggyFlex library containing gRNAs targeting candidate promoters and distal CREs, TSS positive controls, and 10% NTCs was introduced via nucleofection to two K562 cell lines expressing integrated CRISPRa machinery: 1) K562 CRISPRa-VP64 and 2) K562 CRISPRa-VPR. Following selection, 20,000 cells per CRISPRa K562 line (40,000 total) were harvested and profiled using sc-RNA-seq to capture and assign gRNAs to single cell transcriptomes (see Figure

7 and methods for details on piggyFlex design and gRNA capture). b) Following QC and gRNA assignment, we identified an average of 2.54 gRNAs/cell (median 1.0 gRNAs/cell). c) Multiplexing more than one perturbation per cell enabled an average of 178.2 cells/gRNA (median 140.0 cells/gRNA). d) Quantile-quantile plot showing the distribution of expected vs. observed *P*-values for targeting (blue) and non-targeting (gray, downsampled) differential expression tests. e) (Top) Heatmap showing the average log2 fold change in expression between cells with each targeting gRNA vs. controls for each of the primary/programmed target genes. Tests are sorted left-to-right by increasing log2 fold change. (Bottom) Categorical heatmap showing which of the perturbations drove significant upregulation using an Empirical FDR approach (EFDR < 0.1). f) Targeting gRNAs yielding significant upregulation are enriched for proximity to their target gene. We observe no such enrichment for NTCs tested for associations with target genes randomly selected from the same set. g) Example violin plots showing the average log2 fold change between cells with a given gRNA and controls for select hit gRNAs. Hits include TSS positive controls (*CCND2*, *GNB1*), candidate promoters of genes rarely or not expressed in K562 cells (*ANK2*, *BCL11A*) and candidate K562 enhancers (*TSPAN5*, *TMSB4X*, and *ANXA1*). Control cells are downsampled to have the same number of cells as the average number of cells detected per gRNA (n = 178) for visualization. h) Hits included multiple gRNAs targeting isoform-specific promoters of *CHD8*. Empirical *P*-values are visualized alongside tracks for K562 ATAC-seq (ENCODE), H3K27ac signal (ENCODE), and RefSeq validated transcripts (ENSEMBL/NCBI) i) The strongest hit gRNAs for ANK2 target the same promoter that is not prioritized by biochemical marks (*e.g.*, accessibility or H3K27ac). Genomic tracks are the same as in panel h. Abbreviations: NTC: Non-targeting controls.

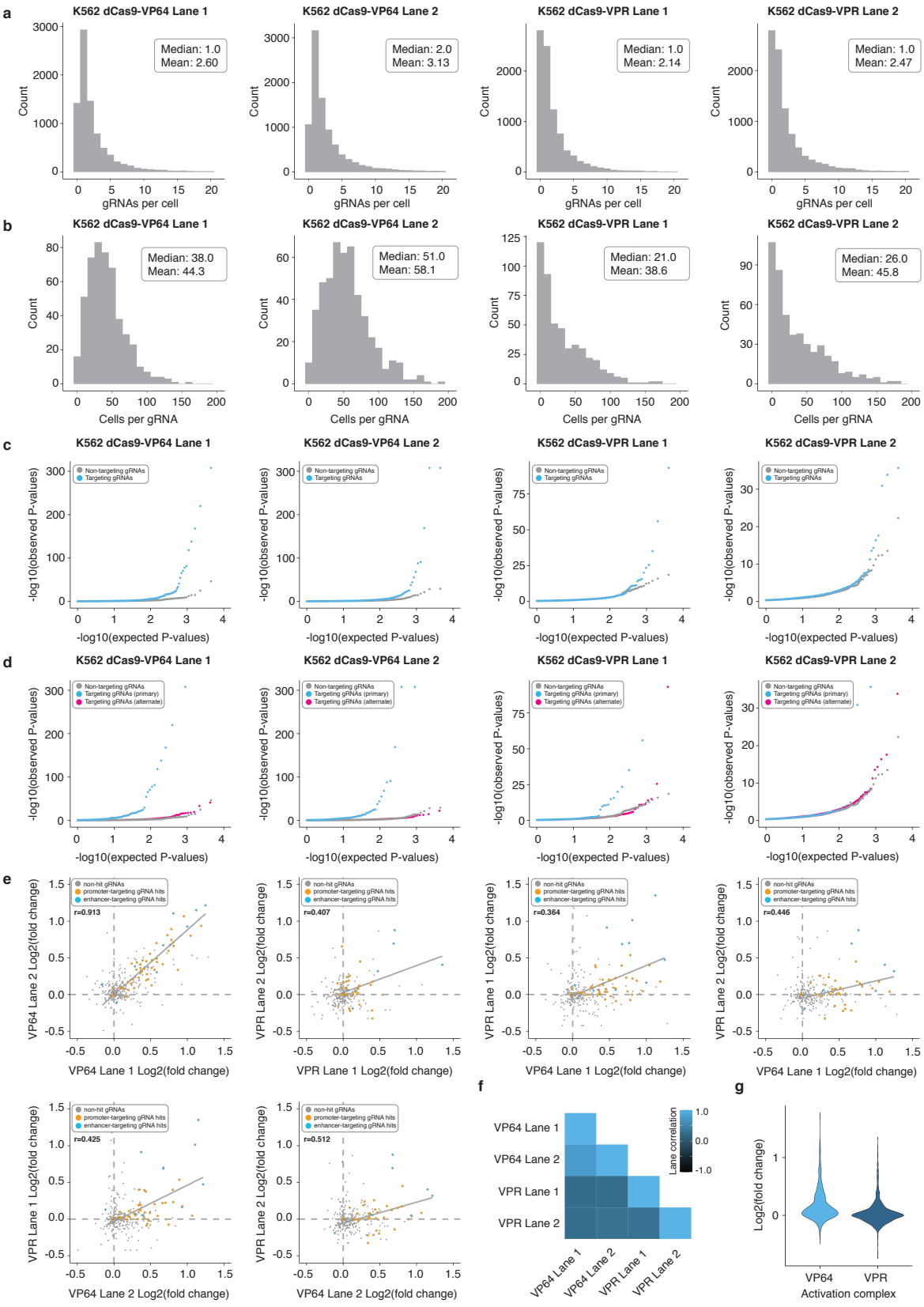
a
K562 dCas9-VP64 Lane 1
Median: 1.0
Mean: 2.60
K562 dCas9-VP64 Lane 2
Median: 2.0
Mean: 3.13
K562 dCas9-VPR Lane 1
Median: 1.0
Mean: 2.14
K562 dCas9-VPR Lane 2
Median: 1.0
Mean: 2.47
Count
gRNAs per cell
b
K562 dCas9-VP64 Lane 1
Median: 38.0
Mean: 44.3
K562 dCas9-VP64 Lane 2
Median: 51.0
Mean: 58.1
K562 dCas9-VPR Lane 1
Median: 21.0
Mean: 38.6
K562 dCas9-VPR Lane 2
Median: 26.0
Mean: 45.8
Cells per gRNA
c
K562 dCas9-VP64 Lane 1
K562 dCas9-VP64 Lane 2
K562 dCas9-VPR Lane 1
K562 dCas9-VPR Lane 2
Non-targeting gRNAs
Targeting gRNAs
-log10(observed P-values)
-log10(expected P-values)
d
K562 dCas9-VP64 Lane 1
K562 dCas9-VP64 Lane 2
K562 dCas9-VPR Lane 1
K562 dCas9-VPR Lane 2
Non-targeting gRNAs
Targeting gRNAs (primary)
Targeting gRNAs (alternate)
-log10(observed P-values)
-log10(expected P-values)
e
non-hit gRNAs
promoter-targeting gRNA hits
enhancer-targeting gRNA hits
r=0.913
r=0.407
r=0.364
r=0.446
r=0.425
r=0.512
VP64 Lane 2 Log2(fold change)
VP64 Lane 1 Log2(fold change)
VPR Lane 2 Log2(fold change)
VPR Lane 1 Log2(fold change)
f
VP64 Lane 1
VP64 Lane 2
VPR Lane 1
VPR Lane 2
Lane correlation
g
Log2(fold change)
VP64
VPR
Activation complex

Figure 10. Results for four independent 10x Genomics lanes from K562 screen.

a) The four 10x Genomics lanes profiled included two lanes with dCas9-VP64 K562 cells and two lanes with dCas9-VPR K562 cells. Following QC and gRNA assignment we identified an average of 2.60, 3.13, 2.14, and 2.47 gRNAs/cell for the four different 10x Genomics lanes profiled (median 2.60, 3.13, 2.14, and 2.47 gRNAs per cell). PiggyBac integrations per cell distribution is not well-modeled by a standard Poisson distribution and is better approximated by an exponential function. b) Multiplexing more than one perturbation per cell yielded an average of 38.0, 51.0, 21.0, and 26.0 cells/gRNA for the four different 10x Genomics lanes profiled (median 44.3, 58.1, 38.6, and 45.8 cells/gRNA). c) QQ-plots displaying observed vs. expected *P*-value distributions for targeting (blue) and NTC (downsampled) populations across the four different 10x Genomics lanes profiled. d) QQ-plots for targeting tests against their intended/programmed target (blue) compared to targeting tests of all other genes with 1Mb of each gRNA (pink) and NTCs (gray downsampled) across the four different 10x Genomics lanes profiled. e) Correlation plots of log2 fold changes of gRNAs across the two K562 cell lines (dCas9-VP64 and dCas9-VPR) for all four 10x Genomics lanes profiled. Pearson correlations of gRNA hits are shown. f) Matrix correlation plot displaying the Pearson correlations of the log2 (fold change) of target gene expression values for programmed targets across the four different 10x Genomics lanes profiled. g) Violin plot displaying the log2 (fold change) of target gene expression values for programmed targets for K562 cells harboring the dCas9-VP64 activation complex and the dCas9-VPR activation complex.

To systematically assess the effect of each CRISPR perturbation on target gene expression, we adapted an iterative differential expression testing strategy in which all single cell transcriptomes are computationally partitioned into cells with or without a given gRNA (Gasperini et al., 2019).

These two groups are then tested for differential expression of all genes within 1 megabase (Mb) (upper estimate of topologically associated domain size in mammalian genomes (Dixon et al., 2012)) upstream and downstream of the gRNA target site (Figure 6; Figure 9a; Methods). In both VP64- and VPR-mediated CRISPRa screening experiments, we observed robust upregulation from both promoter and enhancer-targeting gRNAs (276/391 $\log_2$FC>0, 70.6%, $p<2.2x10^{-16}$, Fisher's Exact Test; Figure 9d-e). The presence of an excess of highly significant *P*-values for cells harboring targeting gRNAs versus non-targeting controls (NTCs) also indicates that this multiplex framework successfully detects upregulation of genes from CRISPRa perturbations (Figure 9d). Effects were consistently much stronger and more significant in the dCas9-VP64 cell line as compared to the dCas9-VPR line (Figure 10). This may be due to differences between the VP64 and VPR effectors, site-of-integration effects (VP64 line is monoclonal while VPR line is polyclonal), MOI differences of the integrated effectors, power differences (more cells were recovered per perturbation for the VP64 line than the VPR line), or a combination of these factors.

To identify significant associations between cCRE-targeting gRNAs and their target genes, which we term "hit gRNAs", we set an empirical false discovery rate (FDR) threshold based on the *P*-values from the NTC gRNA differential expression tests, which are subject to the same sources of noise and error as the targeting gRNA tests. Using an empirical FDR cutoff of 0.1 (Methods), we identified 60 activating gRNA hits, including 8 TSS-targeting positive control gRNAs, 40 candidate promoter-targeting gRNAs, 9 distal enhancer hit gRNAs, 2 distal enhancer hit gRNAs wherein the target gene of CRISPRa vs. CRISPRi differed, and 1 distal enhancer non-hit gRNA (in the last three contexts, hit vs. non-hit refers to whether they were "hits" in the previous CRIPSRi-based screen with the same guides and cell line (Gasperini et al., 2019)) (Figure 9e;

Figure 11). Successfully activating gRNAs were strongly enriched for targeting regions proximal to the genes that they upregulated (Figure 9f) and were specific to their predicted target (45/48 promoter-targeting gRNA hits and 9/12 successful enhancer-targeting gRNAs exclusively upregulated the predicted target and no other gene within 1 Mb; Figure 11). The gRNAs that upregulated a gene other than the predicted target are discussed further below. Of note, we also observed no instances where targeting a regulatory element, whether a promoter or enhancer, caused significant upregulation of >1 gene.

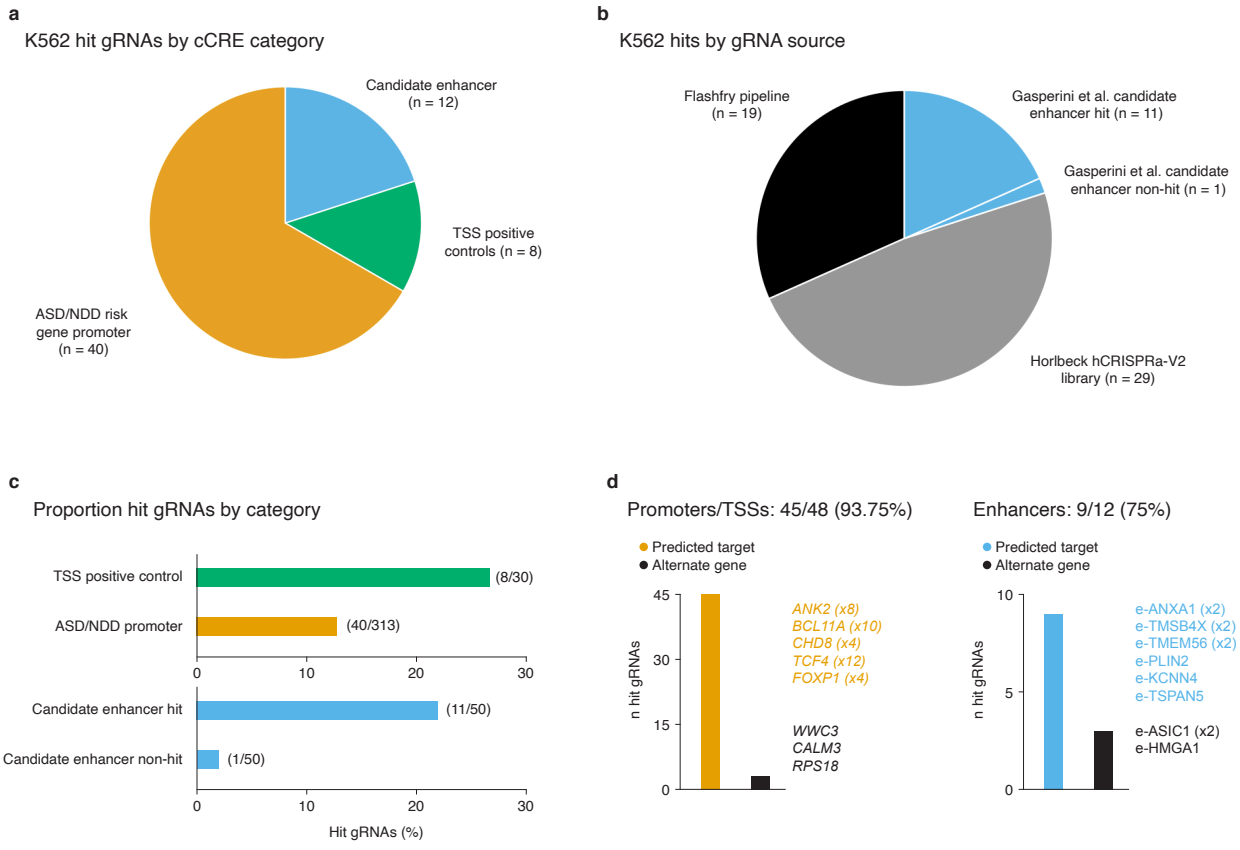

Figure 11. Hit breakdown for screen conducted in K562 cells.

a) K562 hit gRNAs by cCRE category. b) K562 hit gRNAs by gRNA source library or design pipeline. c) Proportion of hit gRNAs by cCRE category. d) Proportion of hit gRNAs yielding upregulation of their intended/expected target gene or an alternate gene for candidate

promoters/TSSs (left) or enhancers (right). Example hits targeting candidate NDD risk gene promoters (left) and K562 enhancers (right) are listed. Bracketed numbers denote the number of independent hit gRNAs targeting the same cCRE.

Taken together, these results demonstrate the potential of this framework to efficiently identify promoter- or enhancer-targeting gRNAs that drive potent, specific upregulation of their target genes in a cell type of interest. Of note, the promoters that were successfully targeted with CRISPRa included genes that were already well-expressed (*e.g., CCND2*, *GNB1*), including two that are haploinsufficient neurodevelopmental disease genes (*FOXP1*, *CHD8*) (Figure 9g-h; Figure 11). For *CHD8*, in which variants leading to haploinsufficiency are important risk factors for ASD and NDD (Fu et al., 2022; Satterstrom et al., 2020), we identified multiple CRISPRa-potent gRNAs targeting distinct isoform-specific promoters, providing an inroad to isoform-specific CRT (Figure 9h; Figure 8; Figure 10; Figure 11).

Our strongest hits were at the promoters of genes with very low or undetectable expression in K562 (*e.g., ANK2*, *BCL11A*; Figure 9g). For example, we identified multiple CRISPRa-potent gRNAs targeting *ANK2*, an ASD/NDD risk gene with a complex isoform structure (Fu et al., 2022; Satterstrom et al., 2020) that is very lowly expressed in K562 cells (Figure 9i). Interestingly, the strongest hits for *ANK2* all targeted a TSS that is not prioritized by biochemical marks (*i.e.*, it is relatively inaccessible and displays a low degree of H3K27ac in K562 cells compared to candidate TSSs of other genes in our library; Figure 9i). On the other hand, for many targeted TSSs or promoters, only one gRNA, if any, potently activated their target gene when coupled to CRISPRa. More specifically, out of the 313 candidate promoter-targeting gRNAs designed to 50 annotated

TSSs of 9 genes, only 38 gRNAs, targeting 10 TSSs and 5 genes, successfully mediated upregulation. An additional 2 gRNAs upregulated different genes (*RPS18* and *WWC3*) than their intended targets (*SYNGAP1* and *FOXP1*). These results underscore the value of inclusive, empirical screens to identify both CRISPRa-competent promoters as well as gRNAs that can successfully activate them.

At the outset of this work, it was unclear if targeting CRISPRa perturbations to enhancers alone (without co-targeting putatively associated promoters) could reliably increase target gene expression to an extent detectable with conventional sc-RNA-seq (Dai et al., 2021; Simeonov et al., 2017; Tak et al., 2021). To determine if CRISPRa targeted to a single enhancer alone could effectively upregulate target gene expression, we analyzed our 50 targeted candidate enhancers, 25 of which were previously validated by multiplex CRISPRi in K562 cells (Gasperini et al., 2019). We observed target gene upregulation for 8 of these 50 targeted candidate enhancers (as noted above, mediated by 12 gRNAs; Figure 9g; Figure 10). Six of the 8 enhancers come from the set of 25 enhancer-gene pairs that we also identified with CRISPRi (Gasperini et al., 2019), including several cases where distinct gRNAs targeting the same enhancer are both successful, *e.g.* two CRISPRa-competent enhancers of *ANXA1* (Figure 9g; Figure 11). In addition, we identified: (1) an enhancer-targeting gRNA that was not a hit in the CRISPRi screen, but here led to upregulation of *HMGA1*; and (2) two enhancer-targeting gRNAs that mediate downregulation of *TUBA1A* when coupled to CRISPRi, but upregulation of *ASIC1* when coupled to CRISPRa. Taken together, these results show that multiplex CRISPRa screens leveraging sc-RNA-seq can identify enhancer-targeting gRNAs that can mediate potent upregulation of specific genes without co-targeting of the corresponding promoters (Figure 9g; Figure 11). Furthermore, differences in

activity and target-choice despite using the same gRNAs hint at potential differences between CRISPRi and CRISPRa that warrant further exploration.

2.4 MULTIPLEX SINGLE-CELL CRISPRA SCREENING OF REGULATORY ELEMENTS IN POST-MITOTIC IPSC-DERIVED NEURONS

We next sought to extend this framework beyond K562 cells to a model that is more relevant for native biology as well as CRT, post-mitotic human induced pluripotent stem cell (iPSC)-derived neurons (Figure 13a) (Zhang et al., 2013). For this, we used a WTC11 iPSC line equipped with a doxycycline-inducible *NGN2* transgene expressed from the *AAVS1* safe-harbor locus to drive neural differentiation, as well as a ecDHFR-dCas9-VPH construct, expressed from the *CLYBL* safe-harbor locus, to drive CRISPRa (Figure 12a-b) (Tian et al., 2021). In this line, addition of doxycycline to induce *NGN2* expression and trimethoprim (TMP) to inhibit the ecDHFR degrons drives neural differentiation and initiates CRISPRa (Tian et al., 2021). Expression of NGN2 in iPSCs commits these cells to a neuronal fate, and post-mitotic neurons with neuronal morphology develop within days (C. Wang et al., 2017).

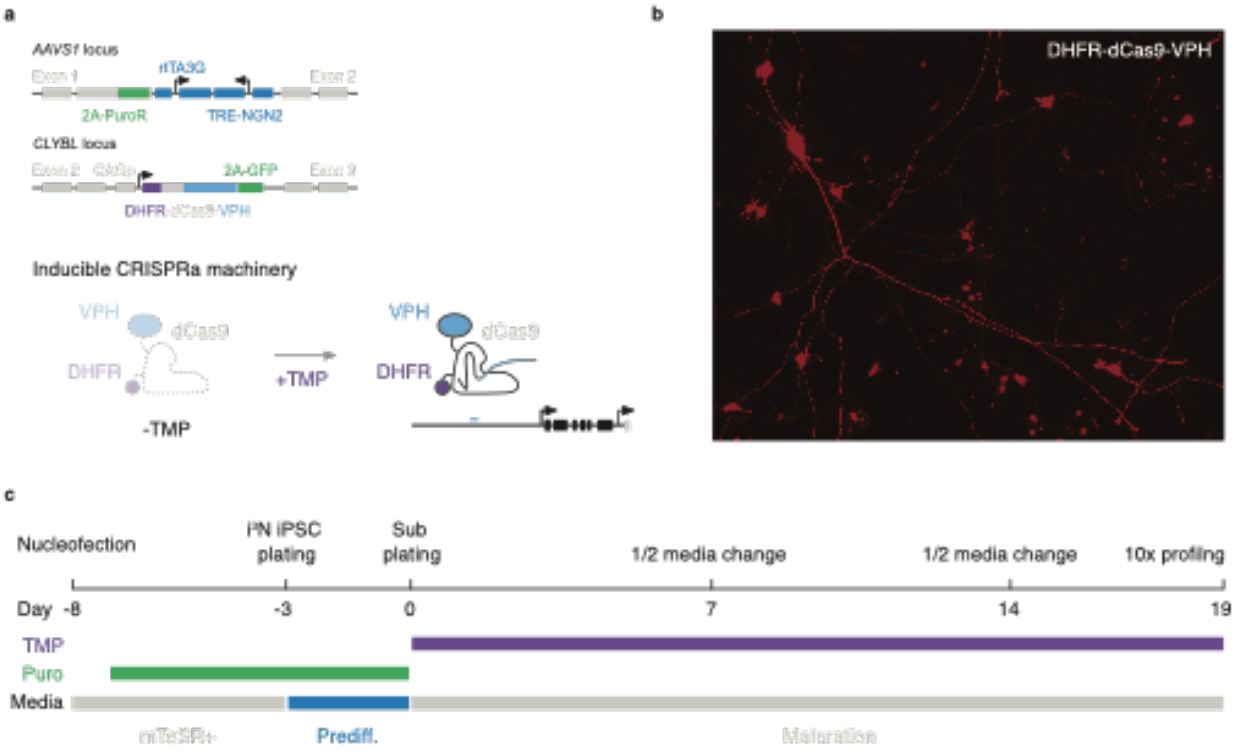

Figure 12. Inducible CRISPRa iPSC-derived neuron line functional validation, selection, and differentiation timeline.

a) (Top) iPSCs equipped with a Dox-inducible *NGN2* transcription factor to drive neural differentiation (integrated at the *AAVS1* safe harbor locus) and TMP-inducible CRISPRa-VPH machinery (integrated at the *CLYBL* locus) were used for all iPSC-derived neuron experiments. (Bottom) In the absence of TMP, CRISPRa-VPH machinery is degraded via a DHFR degron. In the presence of TMP, the CRISPR-VPH machinery is stabilized, enabling perturbation. b) Functional validation of CRISPRa machinery in iPSC-derived neurons. Neurons were lipofected with a minP-tdTomato reporter and sgRNA that targets the minimal promoter. CRISPRa machinery drove clear tdTomato expression in differentiated neurons. c) Nucleofection, selection, and differentiation timeline. iPSCs were nucleofected with piggyFlex gRNA constructs at a high MOI and selected with puromycin to enrich cells for with multiple integrated gRNAs. Following

differentiation induction neurons were subplated in maturation media with TMP to induce CRISPRa machinery. Neurons were single cell profiled following 19 days of differentiation (10x Genomics V3.1 chemistry with direct gRNA capture).

After optimizing iPSC transfection conditions to achieve high numbers of integrated gRNAs per cell via nucleofection, we integrated the same gRNA library (at a 5:1 gRNA-library:transposase ratio) into iPSCs as we did for the K562 screen (Figure 13a). Following integration, we confirmed functional CRISPRa activity in these neurons via the same tdTomato expression assay used in our K562 CRISPRa validation (Figure 12b). In addition to optimizing transfection conditions, we sought to further boost the multiplicity of gRNA integrations per cell by selecting the cells with a high concentration of puromycin (Figure 13a). After differentiating to neurons over 19 days, we proceeded to sc-RNA-seq. Half of the neurons went directly into sc-RNA-seq (10x Genomics), while the other half were dissociated and flow sorted based on GFP expression (top 40%) prior to sc-RNA-seq, again with the goal of boosting the multiplicity of gRNA integrations (Figure 13a). After quality control filtering, we retained 51,183 single-cell transcriptomes, of which we recovered 1+ associated gRNAs for 89%. With our optimized transfection protocol, we identified a mean of 6.14 gRNAs/cell (Figure 13b) and a mean of 638 cells that harbored each individual gRNA (Figure 13c). Sorting on GFP expression prior to sc-RNA-seq resulted in a 2-fold increase in the number of gRNAs identified per cell (Figure 14).

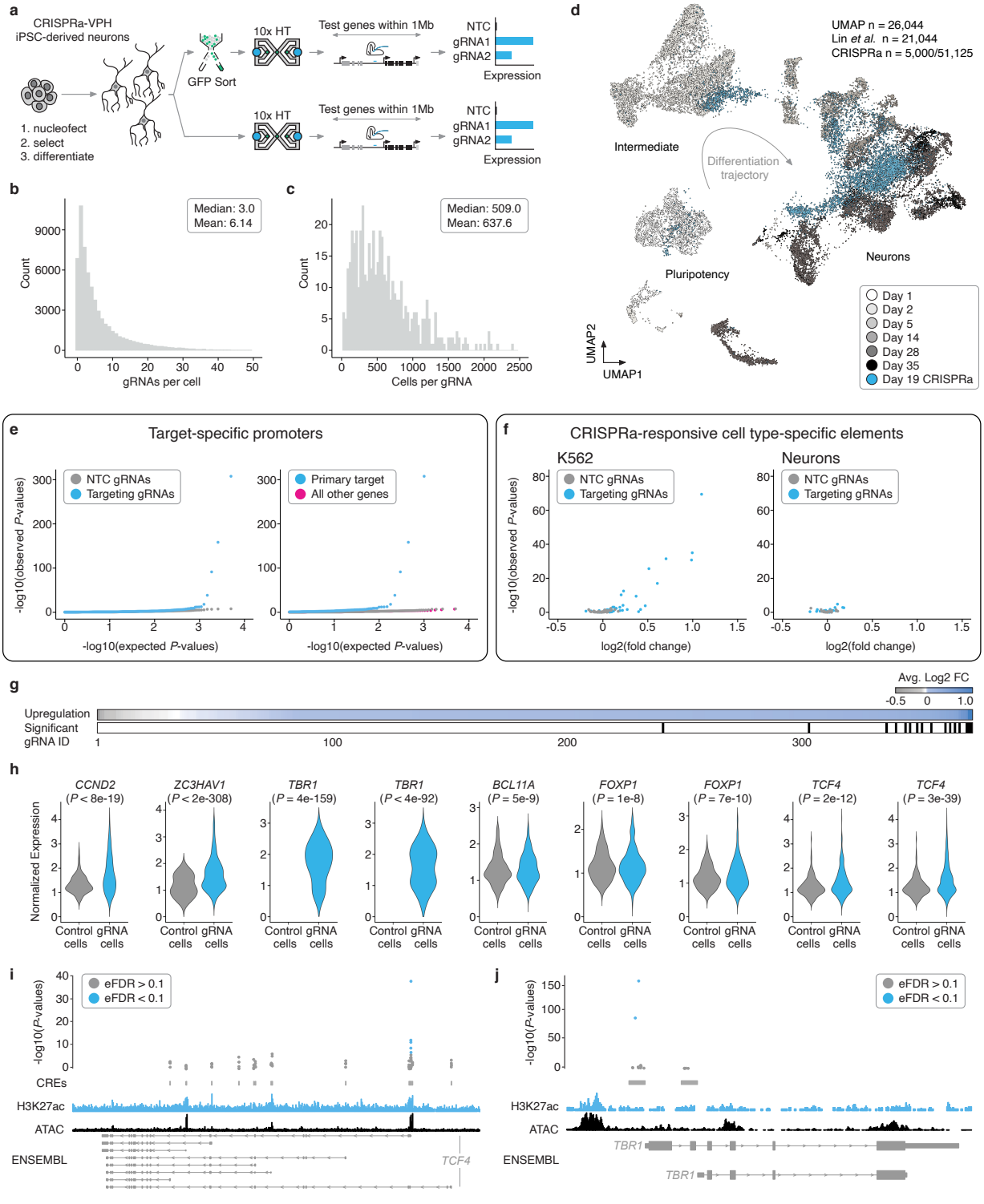

Figure 13. Multiplex single cell CRISPRa screening of regulatory elements in post-mitotic iPSC-derived neurons.

a) The same piggyFlex library as used in K562 experiments was introduced to a human WTC11 iPSC line harboring TMP-inducible CRISPRa machinery and a Dox-inducible NGN2 transgene to drive neural differentiation. Following selection and differentiation, cells were harvested and profiled with sc-RNA-seq to capture gRNAs and assign them to single cell transcriptomes. Half of the neurons were sorted on GFP immediately prior to sc-RNA-seq to increase the multiplicity of gRNA integrations. b) Following QC and gRNA assignment, we identified an average of 6.14 gRNAs/cell (median 3.0). c) Neuron gRNA coverage: each gRNA was identified in an average of 637.6 cells (median 509.0). d) UMAP projection of the neuron dataset from this study (blue, 51,183 cells downsampled to 5,000 cells to aid with visualization) onto a sc-RNA-seq differentiation time-course from a similar differentiation protocol and NGN2 iPSC line (21,044 cells)(H.-C. Lin et al., 2021). This reference time-course dataset is coloured from white to black based on differentiation day. e) (Left) QQ-plot displaying observed vs. expected *P*-value distributions for targeting (blue) and NTC (downsampled) populations. (Right) QQ-plot for targeting tests against their intended/programmed target (blue) compared to targeting tests of all other genes with 1Mb of each gRNA (pink) and NTCs (gray downsampled). There is a clear excess of highly significant *P*-values for programmed targets compared to targeting tests of neighboring genes (pink) or NTCs (gray). f) Volcano plot showing the average log2 fold change and *P*-values exclusively for gRNAs that target putative enhancers in K562 cells (left) and iPSC-derived neurons (right). g) (Top) Heatmap showing the average log2 fold change in expression between cells with each targeting gRNA vs. controls for each of the primary/programmed target genes. (Bottom) Categorical heatmap showing which of the perturbations produced significant upregulation using an Empirical FDR approach (EFDR < 0.1). h) Example violin plots showing the average log2 fold change between cells with a given gRNA and controls for select hit gRNAs. Hits include TSS positive controls (*CCND2*,

ZC3HAV1), candidate promoters of genes rarely or not expressed NGN2, including the cortical neuron marker *TBR1*, and candidate promoters of genes with native expression in iPSC-derived neurons that could be further upregulated (*BCL11A*, *FOXP1*, and *TCF4*). Control cells are downsampled to have the same number of cells as the average number of cells detected per gRNA (n = 638) for visualization. i) Of 14 targeted candidate promoters, five hit gRNAs for TCF4 target the same candidate promoter that aligns with biochemical marks of regulatory activity (ATAC-Seq and H3K27ac). Empirical *P*-values are visualized alongside tracks for iPSC-derived neuron ATAC-seq (accessibility) (Song et al., 2019), and H3K27ac (Song et al., 2019), and RefSeq validated transcripts (ENSEMBL/NCBI). j) Hits included multiple gRNAs targeting *TBR1*. Genomic tracks are the same as in panel i.

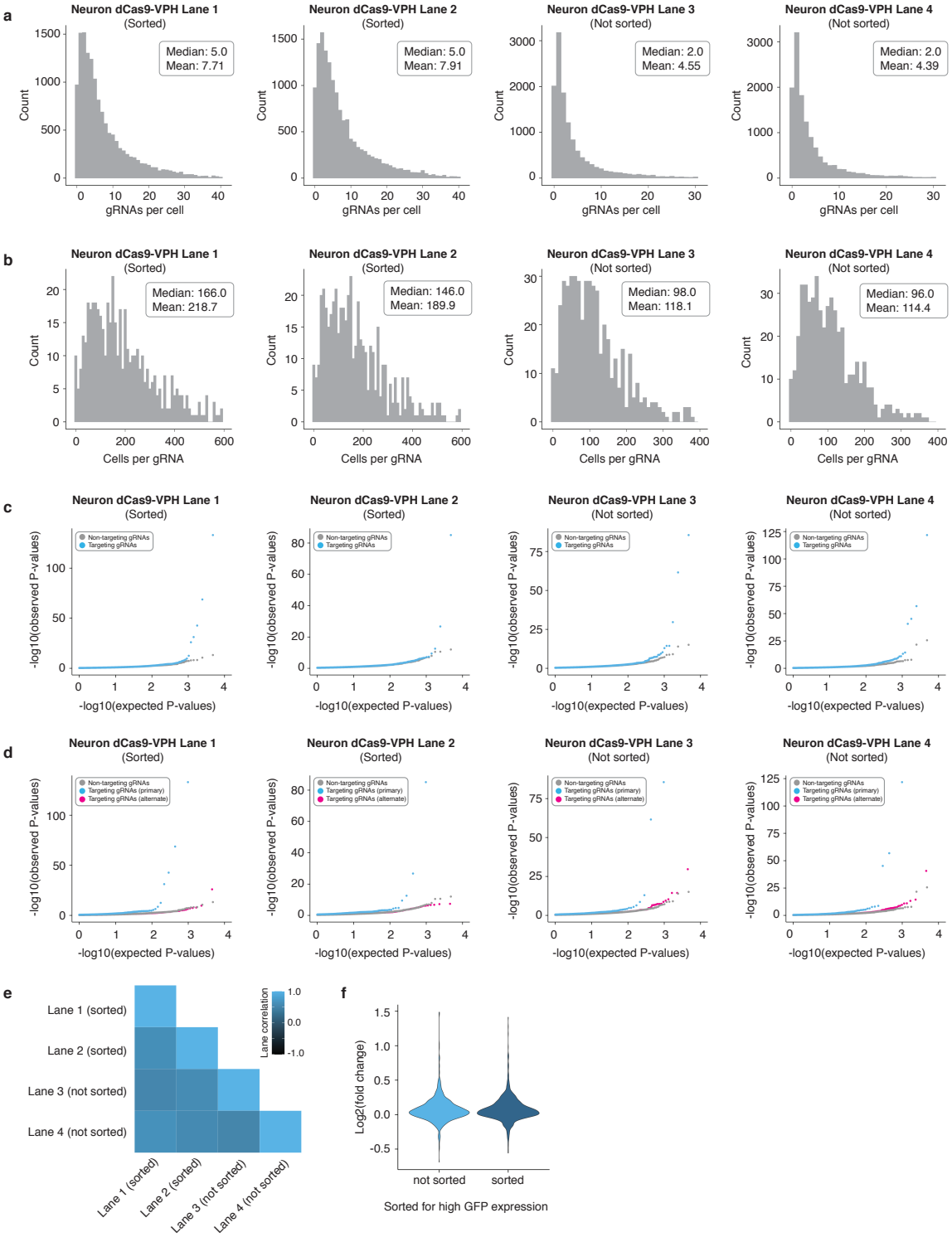

Figure 14. Results for four independent 10x Genomics lanes from iPSC-derived neuron screen.

a) The four 10x Genomics lanes profiled consisted of two lanes with dCas9-VPH neurons that were sorted on the top 40% of GFP expression in these cells, and two lanes that were not on the top 40% of GFP expression in these cells. The cells that were not sorted were still 100% GFP+. Following QC and gRNA assignment we identified an average of 7.71, 7.91, 4.55, and 4.39 gRNAs/cell for the four different 10x Genomics lanes profiled (median 7.71, 7.91, 4.55, and 4.39 gRNAs per cell). Note that sorting neurons on the top 40% of GFP expression boosted the median and mean gRNAs/cell ~2 fold. PiggyBac integrations per cell distribution is not well-modeled by a standard Poisson distribution and is better approximated by an exponential function. b) Multiplexing multiple perturbations per cell yielded an average of 218.7, 189.9, 118.1, and 114.4 cells/gRNA for the four different 10x Genomics lanes profiled (median 166, 146, 98, and 96 cells/gRNA). c) QQ-plots displaying observed vs. expected *P*-value distributions for targeting (blue) and NTC (downsampled) populations across the four different 10x Genomics lanes profiled. d) QQ-plots for targeting tests against their intended/programmed target (blue) compared to targeting tests of all other genes with 1Mb of each gRNA (pink) and NTCs (gray downsampled) across the four different 10x Genomics lanes profiled. e) Matrix correlation plot displaying the Pearson correlations of the log2 (fold change) of target gene expression values for programmed targets across the four different 10x Genomics lanes profiled. f) Violin plot displaying the log2 (fold change) of target gene expression values for programmed targets for neurons that were sorted on the top 40% GFP expression (sorted) and neurons that were not sorted (not sorted).

Our differentiated neurons most closely resemble 14- to 35-day differentiated neurons obtained via *NGN2* induction in iPSCs by an independent group (H.-C. Lin et al., 2021) (inferred by integration of these sc-RNA-seq datasets; Figure 13d; Figure 15). A minority of the neurons

transcriptionally resemble an intermediate neuronal fate, a difference that we tentatively attribute to the absence of co-cultured glia in our differentiation protocol. Although glia are known to promote maturation of NGN2-induced neurons (and were used in generating the dataset we are comparing to (H.-C. Lin et al., 2021)), we excluded them because they can also introduce culture variability due to batch effects introduced by primary glia (C. Wang et al., 2017).

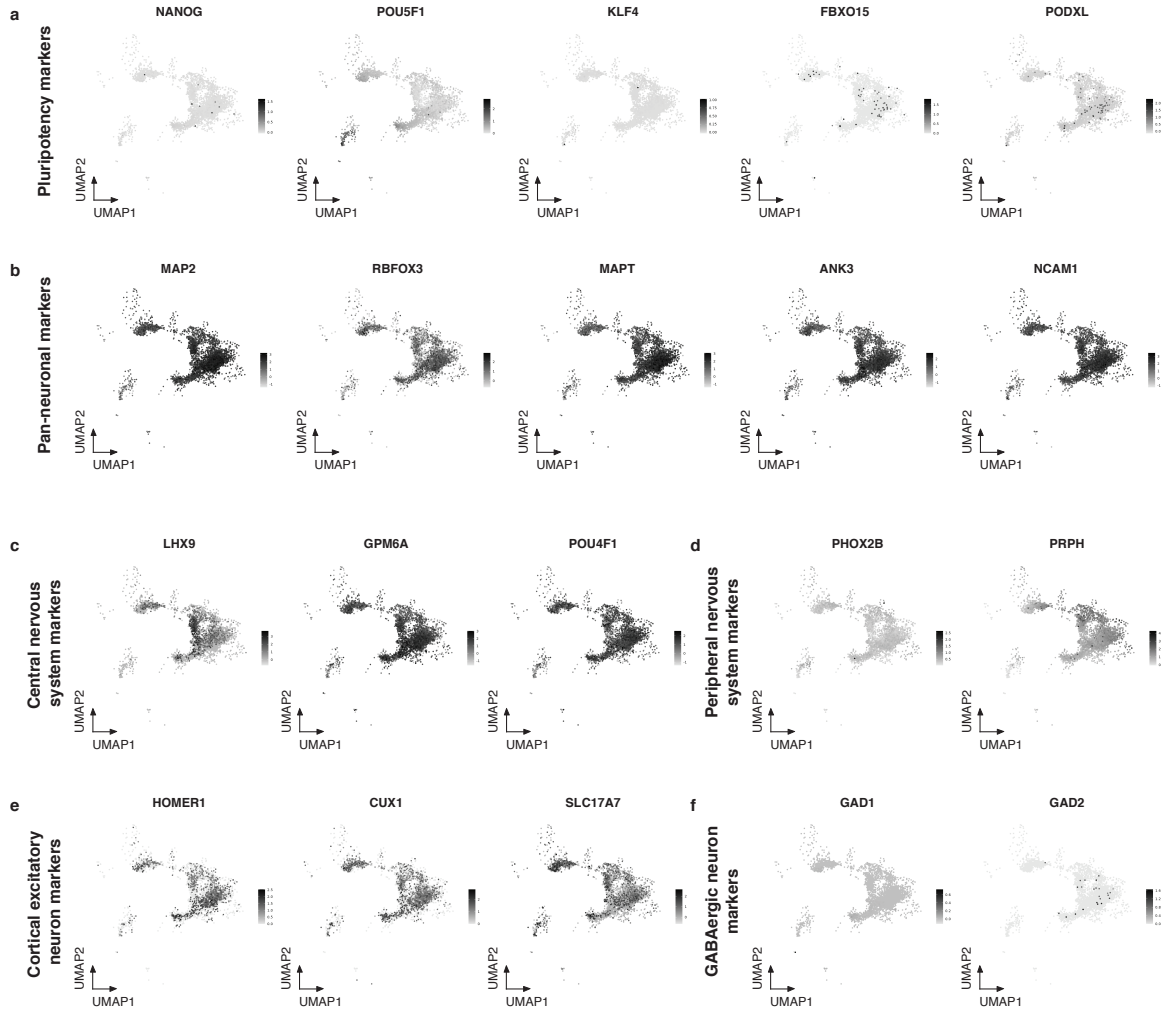

Figure 15. Single-cell transcriptomic characterization of iPSC-derived neurons used in screen.

a) Expression feature plots of canonical pluripotency markers *NANOG*, *POU5F1*, *KLF4*, *FBXO15*, and *PODXL*. b) Expression feature plots of pan-neuronal markers *MAP2*, *RBFOX3*, *MAPT*, *ANK3*,

and *NCAM1*. c) Expression feature plots of central nervous system marker genes *LHX9*, *GPM6A*, and *POU4F1*. d) Expression feature plots of peripheral nervous system marker genes *PHOX2B* and *PRPH*. e) Expression feature plots of cortical excitatory neuron markers *HOMER1*, *CUX1*, and *SLC17A7*. f) Expression feature plots of GABAergic neuron marker genes *GAD1* and *GAD2*.

We confirmed that the neurons had progressed beyond a pluripotent state and were committed to a post-mitotic neuronal fate by the expression of the pan-neuronal marker *MAP2* and the lack of expression of the pluripotency marker *NANOG (*Figure 15). These neurons also express *LHX9* and *GPM6A* -- markers of central nervous system (CNS) neurons (Figure 15c); and *CUX1* and *SLC17A7*, but not GABAergic markers *GAD1* and *GAD2,* supporting their assignment as excitatory rather than inhibitory neurons (Figure 15f) (Zhang et al., 2013). Consistent with this, when we co-embedded our transcriptome data onto data from Lin et al. (H.-C. Lin et al., 2021), they overlay with “Fate 2” and “Fate 3” cells, which transcriptionally resemble CNS neurons (Figure 13d; Methods). Of note, there was no readily apparent enrichment of specific gRNAs within particular clusters (Figure 16), which is consistent with the specificity of the observed instances of upregulation (Figure 18).

Figure 16. Distribution of CRISPRa gRNAs in single-cell neuron transcriptome data.

Cells harboring specific CRISPRa gRNAs (dark blue) overlaid onto *NGN2*-induced neuron differentiation transcriptome data (H.-C. Lin et al., 2021). No readily apparent spatial enrichment of gRNAs is observed in UMAP plots. Note that the CRISPRa dataset was randomly downsampled to 5000 cells for all UMAP comparison analyses.

We applied the same differential expression testing strategy as used for the K562 screen to the iPSC-derived neuron screen data, with an empirical FDR cutoff of 0.1 to call significant hits. Similarly to the K562 screen, we observed robust upregulation from targeting gRNAs (281/383 $\log_2$FC>0, 73.4%, p<2.2x10^{-16}, Fisher's Exact Test) and an excess of highly significant *P*-values for targeting gRNA tests compared to NTCs (Figure 13e), confirming that this overall framework is transferable to more physiologically and clinically relevant models such as iPSC-derived neurons. As with the K562 screen, we observed strong enrichment for genomic proximity between successful gRNAs and their target genes, but no such enrichment for NTCs tested for associations with target genes randomly selected from the same set (Figure 17).

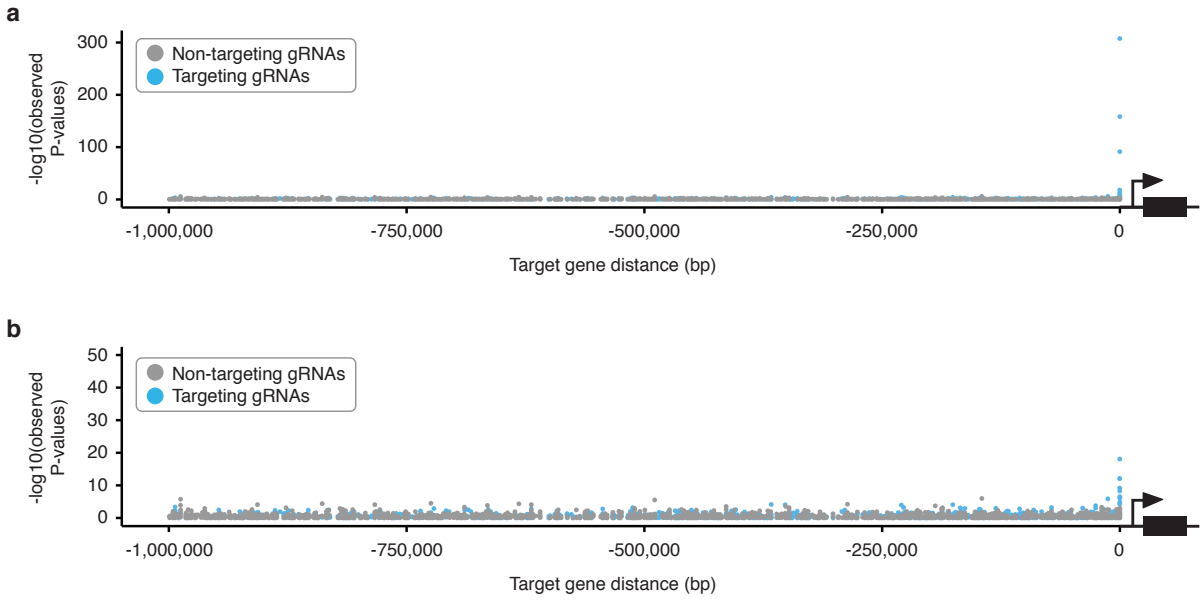

Figure 17. Successful targeting gRNAs are enriched for genomic proximity to their paired target gene scores near target genes in the iPSC-derived neurons.

a) Targeting gRNAs yielding significant upregulation are enriched for proximity to their target gene, while NTCs are not. b) Same plot as in a, with the y-axis clipped at 50.

There were 17 hit gRNAs in neurons (FDR < 0.1; Figure 13g), all of which were TSS-targeting positive controls (n = 6) or candidate promoters of ASD/NDD risk genes (n = 11) (Figure 18). Of these 17 hit gRNAs, 12 were also hits in the K562 screen while 5 were specific to iPSC-derived neurons (Figure 19). The screen in iPSC-derived neurons was strikingly target-specific: 16 of 17 of our identified hits, all promoter-targeting gRNAs, upregulated their anticipated target gene and no other genes within the 1-Mb window tested. The only gRNA hit in iPSC-derived neurons resulting in upregulation of an unintended gene was a gRNA targeting the TSS of the pseudogene *PPP5D1* that led to upregulation of the calmodulin gene *CALM3 (*Figure 18d), but this is presumably due to these two genes sharing a bidirectional, outward-oriented core promoter. This gRNA also drove upregulation of *CALM3* in the CRISPRa screen of K562 cells (Figure 11d). We observed no significant differences across several characteristics (*e.g.*, GC content, baseline target gene expression level, the number of cells harboring each gRNA) between gRNAs yielding successful activation and those not in K562 cells and neurons, with the exception that K562 enhancer hit gRNAs tended to have more cells (Figure 20).

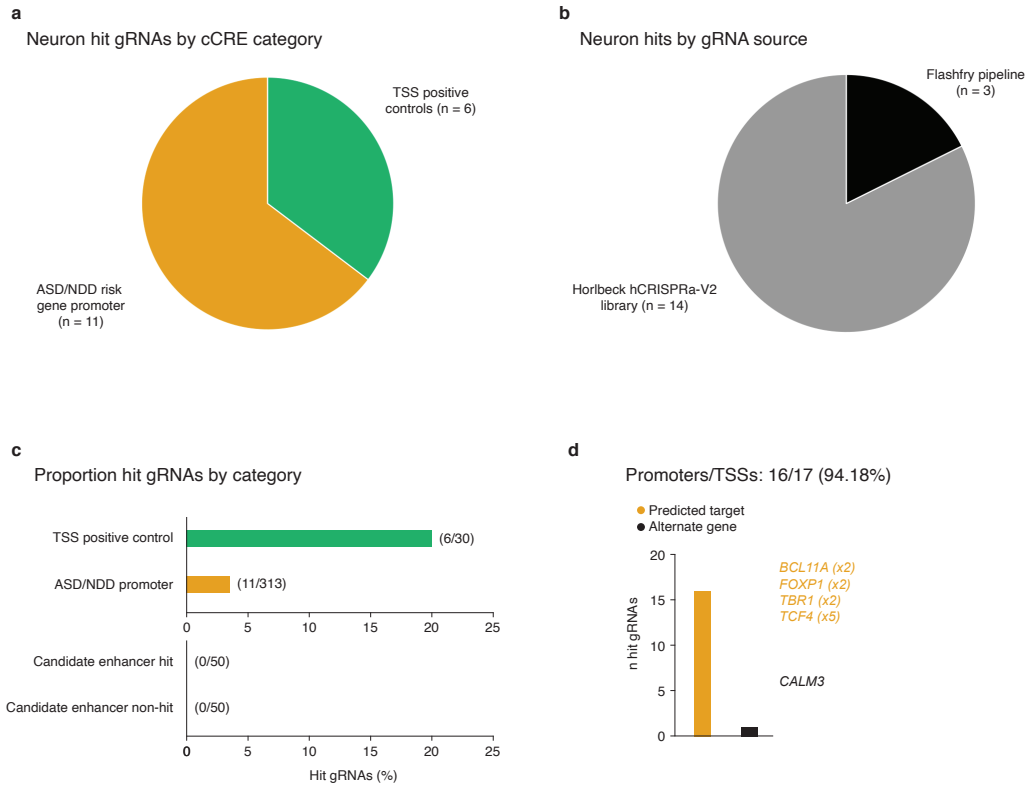

Figure 18. Hit breakdown for screen conducted in iPSC-derived neurons.

a) Neuron hit gRNAs by cCRE category. b) Neuron hit gRNAs by gRNA source library or design pipeline. c) Proportion of hit gRNAs by cCRE category. d) Proportion of hit gRNAs yielding upregulation of their intended/expected target gene or an alternate gene for candidate promoters/TSSs. Example hits targeting candidate NDD risk gene promoters are listed. Bracketed numbers denote the number of independent hit gRNAs targeting the same cCRE.

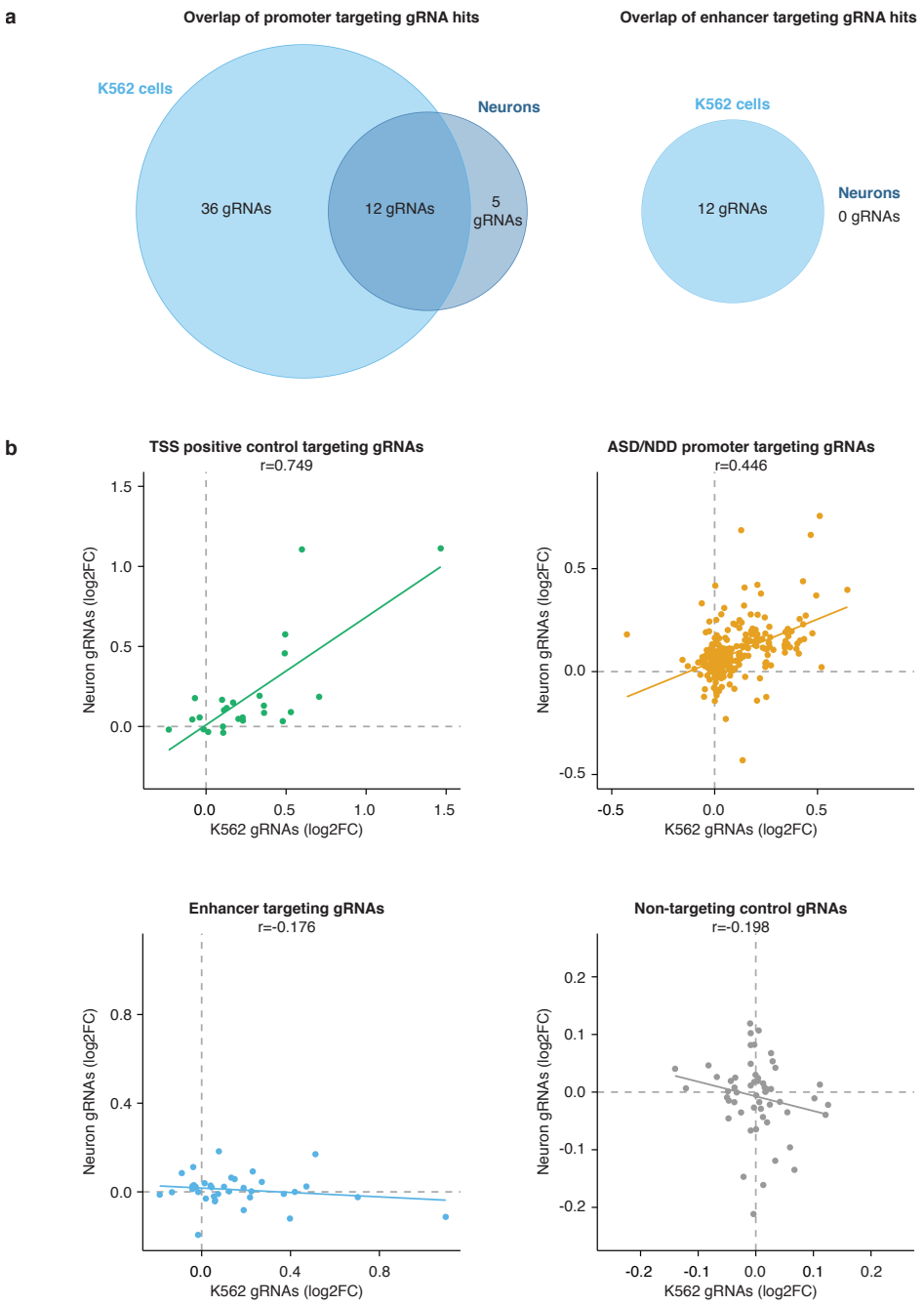

Figure 19. Comparison of K562 vs. neuronal CRISPRa screening hits.

a) Venn diagram showing number of overlapping promoter-targeting gRNA hits (left) and enhancer-targeting gRNA hits (right) between the K562 and neuron CRISPRa screens. b) Correlation plots of log2 fold changes of TSS positive control targeting gRNAs (top left),

ASD/NDD promoter targeting gRNAs (top right), enhancer targeting gRNAs (bottom left), and NTC gRNAs (bottom right) between the K562 and neuron CRISPRa screens.

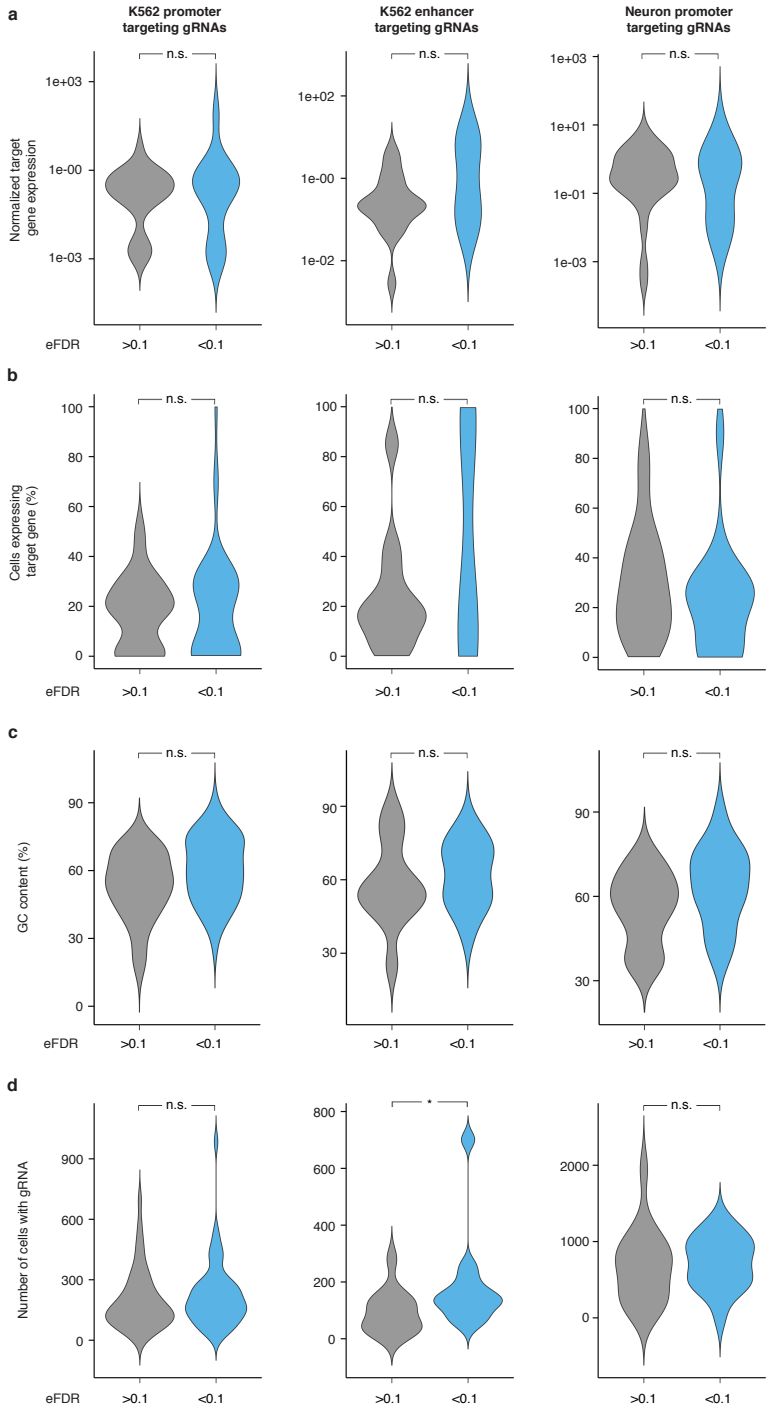

Figure 20. Characteristics of gRNAs leading to upregulation at EFDR<0.1 vs. EFDR>0.1.

a) Comparison of normalized gene expression values of targeted genes of gRNAs that resulted in an EFDR<0.1 (designated as “hit” gRNAs) versus gRNAs that resulted in an EFDR>0.1 (not designated as “hit” gRNAs). b) Comparison of the percentage of cells expressing the target gene of gRNAs that resulted in an EFDR<0.1 versus gRNAs that resulted in an EFDR>0.1. c) GC content (in percent) of gRNAs that resulted in an EFDR<0.1 versus gRNAs that resulted in an EFDR>0.1. d) Number of cells harboring each gRNA for gRNAs that resulted in an EFDR<0.1 versus gRNAs that resulted in an EFDR>0.1. For all panels, K562 promoter-targeting gRNAs (left), K562 enhancer-targeting gRNAs (middle), and neuron promoter-targeting gRNAs (right) are shown. Abbreviations: n.s.: not significant (p>0.05, Wilcoxon rank sum test), *: p<0.05 (Wilcoxon rank sum test).

Similar to K562 cells, we observed several instances where a specific TSS was most amenable to activation (Figure 21). One such example is *TCF4*, an ASD/NDD risk gene (Fu et al., 2022; Satterstrom et al., 2020)that is a strong candidate for CRT due to its large cDNA size (precluding it from fitting into an AAV) and complex locus architecture. We tested 14 candidate TSSs of *TCF4* and identified 5 gRNAs capable of driving upregulation of *TCF4* in neurons, all of which target the same candidate TSS that resides in open chromatin with strong H3K27ac signal (Figure 13h-i; Figure 21a). Our hits also included examples of cell type-specific promoters. Among these were several gRNAs targeting candidate promoters of ASD/NDD risk genes capable of upregulating genes that are not expressed or rarely expressed in iPSC-derived *NGN2*-differentiated neurons (Figure 13h). For example, gRNAs targeting the promoter of *TBR1*, a transcription factor expressed in forebrain cortical neurons but known not to be expressed in NGN2-differentiated iPSC-derived neurons (Zhang et al., 2013) led to *TBR1* upregulation (Figure 13j; Figure 21b). Of

note, these same gRNAs did not result in upregulation of *TBR1* in K562 cells. This suggests that these neurons are in a permissive state for CRISPRa to activate *TBR1*, despite a lack of highly accessible chromatin in the region targeted by the *TBR1* gRNA (Figure 13h, j; Figure 21b). Whether these differences in "*TBR1* activatability" are due to differences in the chromatin environment at this locus between K562 cells and iPSC-derived neurons, or alternatively differences in the milieu of *trans*-acting factors, remains an open question.

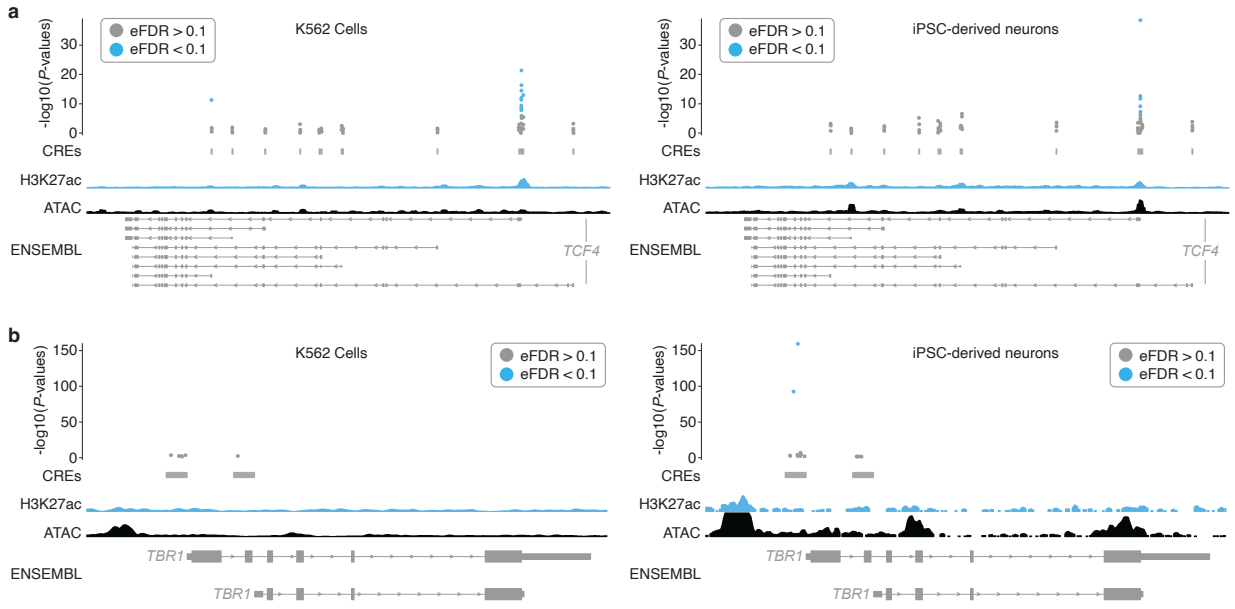

Figure 21. TSS and cell-type specific promoters.

a) The majority of hit gRNAs for *TCF4* target the same TSS in K562 cells and iPSC-derived neurons. Empirical *P*-values are visualized alongside tracks for K562 ATAC-seq (ENCODE), K562 H3K27ac signal (ENCODE), iPSC-derived neuron ATAC-seq (accessibility) (Song et al., 2019), iPSC-derived neuron H3K27ac (Song et al., 2019) and RefSeq validated transcripts (ENSEMBL/NCBI). b) Two hit gRNAs targeting the same TSS of *TBR1* drive upregulation specifically in iPSC-derived neurons. Genomic tracks are the same as in panel a.

However, in contrast to the cell type-specific promoter examples noted above, we more often observed consistent upregulation across promoter targets and TSS-targeting controls between the two cell types (Figure 19). Specifically, 12 out of 17 of the promoter- and TSS-targeting hit gRNAs in neurons were also hits in K562 cells, and upregulation was correlated across cellular contexts (Pearson's correlation coefficient = 0.75; Figure 19). In contrast, we observed striking cell type-specificity for targeted enhancers that were successfully upregulated. While 20% (12/60) of our K562 screening hits were enhancer-targeting gRNAs (Figure 11), none of these were also hits in neurons (Figure 18; Figure 22). Even putting aside significance, the fold-effects on the anticipated target genes of K562-competent activating gRNAs were not well-correlated between cell types (Figure 13f; Figure 19b, Pearson's correlation coefficient = -0.18). Overall, these results show that it is possible to drive cell type-specific upregulation of a gene of interest by targeting CRISPRa to a cell type-specific distal enhancer, without co-targeting of the corresponding promoter.

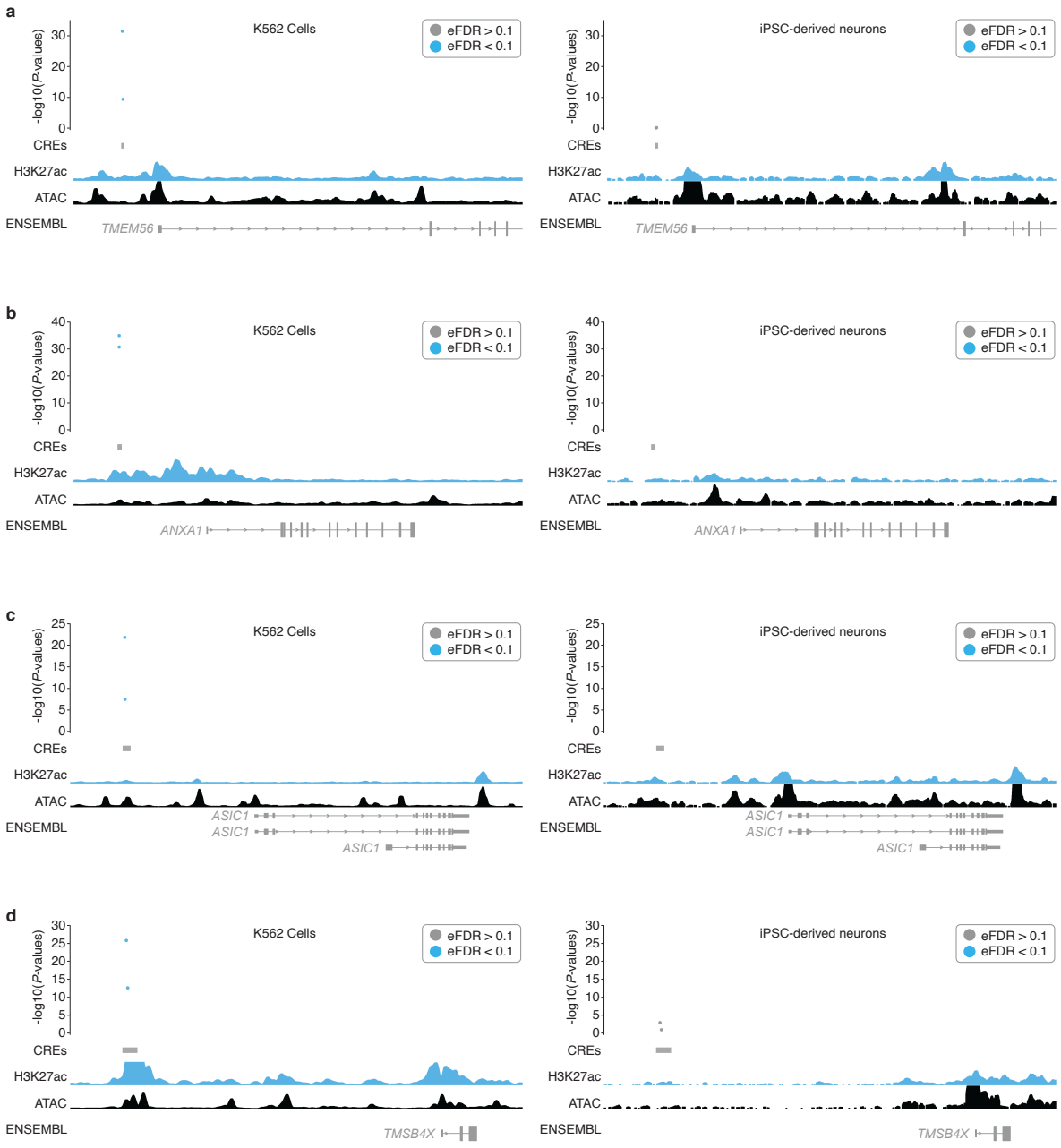

Figure 22. Cell-type specific enhancers.

a-d) Empirical *P*-values are visualized alongside tracks for K562 ATAC-seq (ENCODE), K562 H3K27ac signal (ENCODE), iPSC-derived neuron ATAC-seq (accessibility) (Song et al., 2019), iPSC-derived neuron H3K27ac (Song et al., 2019) and RefSeq validated transcripts

(ENSEMBL/NCBI). All K562 enhancer hits were cell type specific. Enhancers with multiple hit gRNAs are shown.

2.5 DISCUSSION

Here, we describe a scalable framework for identifying cell-type-specific regulatory elements which when targeted with CRISPRa can drive the upregulation of specific target genes. In applying this framework, we identified gRNAs functionally and cell type-specifically targeting promoters of haploinsufficient genes in K562 cells and iPSC-derived excitatory neurons. We identified a novel candidate enhancer-gene pair that is CRISPRa- but not CRISPRi-sensitive, as well as an instance in which a single enhancer, targeted by the same gRNAs, modulated different genes when coupled to CRISPRa vs. CRISPRi. Our approach holds potential to massively scale the screening for gRNAs and cell-type-specific CREs capable of upregulating remaining functional copies of the roughly 660 genes known to cause disease or disorders when haploinsufficient.

Several of our strongest gRNA hits were not prioritized by typical predictors of enhancer function, such as chromatin accessibility or H3K27ac histone modifications. For example, we are able to upregulate *TBR1* in iPSC-derived neurons by targeting a promoter region that is largely within closed chromatin in this cellular context. Indeed, while measures of proximity, accessibility, and enhancer-related biochemical marks are all strong predictors, none are conclusive or deterministic predictors of regulatory sequence function, either alone or in combination. This underscores the importance of empirical, systematic screens for CRISPRa-responsive regulatory sequences with approaches such as the one described here. Ultimately, a variety of factors including chromatin accessibility and epigenetic modifications, gRNA design quality, and target-specific nuances

around CRISPRa-responsiveness, may play a role in determining the success of a CRISPRa perturbation in a given cellular context. Future scaling of this technology and its application to additional, clinically relevant cell types, will provide rich training sets that may enable derivation of rational CRISPRa gRNA design rules for distal, cell-type-specific gene activation, which, in contrast to promoters and CRISPRi (Horlbeck et al., 2016; Sanson et al., 2018; Yao et al., 2022), are quite lacking at present. Further, these results illustrate the unique potential of noncoding CRISPRa screens to identify regulatory elements that can mediate upregulation of target genes, regardless of whether or not the gene is natively expressed in the cell type of interest or not.

A major question that we sought to answer through these experiments was whether one can target candidate enhancer sequences with a CRISPRa perturbation and observe upregulation of an intended target gene via scRNA-seq. There have been relatively few efforts to apply CRISPRa to enhancers to date, and most have focused on a handful of enhancer regions and measuring expression of only nearby genes of interest as a readout (Dai et al., 2021; Simeonov et al., 2017; Tak et al., 2021). Recent literature suggests that co-targeting a promoter and the candidate enhancer in question can make the enhancer CRISPRa perturbations more efficient and reliable (Tak et al., 2021). Although feasible, co-targeting an enhancer and promoter is less likely to yield cell-type-specific upregulation of target genes -- a likely requirement for effective CRT. Delivery of multiple gRNAs also complicates therapeutic delivery and increases the chances of effects on off-target genes (not to mention off-target cell types). Despite using gRNAs that were optimized for CRISPRi screening in our CRISPRa screen, we observed target gene upregulation for 8 of 25 enhancers that we targeted (32%), showing that one can reliably increase target gene expression

by targeting enhancers alone. We imagine that this success rate can be improved via a combination of brute force, *i.e.* testing more gRNAs, and better CRISPRa-specific gRNA design.

Multiplex, single-cell CRISPRa screening is a scalable approach to identifying functional CRISPRa gRNAs that can upregulate intended target genes in either a general or cell-type-specific manner. We introduced multiple perturbations per cell, which increased the power of our assay (*i.e.* a mean of 1 gRNA per cell would have required sc-RNA-seq of >400,000 cells to achieve the same power). Given the ease of generating large numbers of differentiated neurons with *in vitro* human neural cultures, sorting on the GFP-positive gRNA expression vector prior to single-cell transcriptome profiling offers a straightforward way to further boost the number of gRNAs captured per cell. In addition, improvements in methods to capture specific transcripts (in this case, gRNAs) with more cost-effective and scalable transcriptional profiling methods such as sci-RNA-seq (Cao et al., 2019; Xu et al., 2023) may enable considerably larger screens for a given cost.

CRT is a promising, next-generation therapeutic approach that harnesses endogenous gene regulatory circuits to treat genetic disorders (Matharu & Ahituv, 2020; Matharu et al., 2019; Tamura et al., 2022). However, CRT requires an intricate knowledge of the regulatory elements capable of driving target gene upregulation at physiologically relevant levels specifically in affected tissues. We envision that the framework described here can be deployed in increasingly sophisticated *in vitro* and *in vivo* models of human development to discover reagents capable of treating the hundreds of disorders associated with low gene dosage.

2.6 METHODS

Cell Lines and Culture

K562 cell culture

K562s cells are a pseudotriploid ENCODE Tier I erythroleukemia cell line derived from a female (age 53) with chronic myelogenous leukemia (Zhou et al., 2019). All K562 cells were grown at 37°C, and cultured in RPMI 1640 + L-Glutamine (GIBCO, Cat. No. 11-875-093) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals, Cat No. FBS-BSC) and 1% penicillin-streptomycin (GIBCO/ Thermo Fisher Scientific; Cat. No. 15140122).

Induced pluripotent stem cell (iPSC) culture

Human WTC11 iPSCs equipped with a doxycycline-inducible *NGN2* transgene expressed from the *AAVS1* safe-harbor locus as well as an ecDHFR-dCas9-VPH construct (VPH consists of 12 copies of VP16, fused with a P65-HSF1 activator domain) expressed from the CLYBL safe-harbor locus were a gift from the Kampmann lab (Tian et al., 2021). These iPSCs were cultured in mTeSR Plus Basal Medium (Stemcell technologies; Cat. No. 100-0276) on Greiner Cellstar plates (Sigma-Aldrich; assorted Cat. Nos.) coated with Geltrex™ LDEV-Free, hESC-Qualified, Reduced Growth Factor Basement Membrane Matrix (Gibco; Cat. No. A1413302) diluted 1:100 in Knockout DMEM (GIBCO/Thermo Fisher Scientific; Cat. No. 10829018). mTeSR Plus Basal Medium was replaced every other day. When 70–80% confluent, cells were passaged by aspirating media, washing with DPBS (GIBCO/Thermo Fisher Scientific; Cat. No. 14190144), incubating with StemPro Accutase Cell Dissociation Reagent (GIBCO/Thermo Fisher Scientific; Cat. No. A1110501) at 37 °C for 5 min, diluting Accutase 1:1 in mTeSR Plus Basal Medium, collecting

cells in conical tubes, centrifuging at 800*g* for 3 min, aspirating supernatant, resuspending cell pellet in mTeSR Plus Basal Medium supplemented with 0.1% dihydrochloride ROCK Inhibitor (Stemcell technologies; Cat. No. Y-27632), counting and plating onto Geltrex-coated plates at the desired number.

Human iPSC-derived neuronal cell culture, differentiation, and CRISPRa induction

The iPSCs described above were used for the differentiation protocol below. On day -3, iPSCs were dissociated and centrifuged as above, and pelleted cells were resuspended in Pre-Differentiation Medium containing the following: Knockout DMEM/F-12 (GIBCO/Thermo Fisher Scientific; Cat. No. 12660012) as the base, 1X MEM Non-Essential Amino Acids (GIBCO/Thermo Fisher Scientific; Cat. No. 11140050), 1X N-2 Supplement (GIBCO/ Thermo Fisher Scientific; Cat. No. 17502048), 10 ng/mL NT-3 (PeproTech; Cat. No. 450-03), 10ng/mL BDNF (PeproTech; Cat. No. 450-02), 1 ug/mL Laminin mouse protein (Thermo Fisher Scientific; Cat. No. 23017015), 10 nM ROCK inhibitor, and 2 mg/mL doxycycline hyclate (Sigma-Aldrich; Cat. No. D9891) to induce expression of *NGN2*. iPSCs were counted and plated at 800K cells per Geltrex-coated well of a 12-well plate in 1 mL of Pre-Differentiation Medium, for three days. At day -2 and day -1, media changes were performed using pre-differentiation medium without ROCK inhibitor. On day -1, 12-well plates for differentiation were coated with 15 ug/mL Poly-L-Ornithine (Sigma-Aldrich; Cat. No. P3655) in DPBS, and incubated overnight at 37 degrees Celsius. On day 0, the Poly-L-Ornithine coated plates were washed three times using DPBS, and the plates were air dried in a 37 degree Celsius incubator until all the DPBS evaporated. Pre-differentiated cells were dissociated and centrifuged as above, and pelleted cells were resuspended in Maturation Medium containing the following: 50% Neurobasal-A medium (GIBCO/Thermo

Fisher Scientific; Cat. No. 10888022) and 50% DMEM/F-12 (GIBCO/Thermo Fisher Scientific; Cat. No. 11320033) as the base, 1X MEM Non-Essential Amino Acids, 0.5X GlutaMAX Supplement (GIBCO/Thermo Fisher Scientific; Cat. No. 35050061), 0.5X N-2 Supplement, 0.5X B-27 Supplement (GIBCO/Thermo Fisher Scientific; Cat. No. 17504044), 10 ng/mL NT-3, 10 ng/mL BDNF, 1 ug/mL Laminin mouse protein, and 2 ug/mL doxycycline hyclate. Pre-differentiated cells were subsequently counted and plated at 400,000-450,000 cells per well of a 12-well plate coated with Poly-L-Ornithine in 1 mL of Maturation medium with 20 uM trimethoprim (TMP) (Sigma-Aldrich, Cat No. 92131) to activate the CRISPRa machinery in these cells (TMP stabilizes the degron-tagged CRISPRa machinery). On day 7, half of the medium was removed and an equal volume of fresh Maturation medium without doxycycline was added. On day 14, half of the medium was removed and twice that volume of fresh medium without doxycycline was added. On day 19, neurons were harvested for sc-RNA-seq.

Cell line generation and CRISPRa validation

K562 cells

K562 cells expressing dCas9-VP64 were generated in-house via lentiviral integration of a dCas9-VP64-blast construct (Konermann et al., 2014) (Addgene Plasmid #61422) into K562 cells. Cells were selected with 10 ug/mL blasticidin, and polyclonal cells were single-cell sorted into 96-well plates to grow up clonal cell lines expressing dCas9-VP64. Clonal cell lines were tested for CRISPRa activity by testing the ability of a CRISPRa gRNA to activate a minP-tdTomato construct (Chavez et al., 2015), and the highest tdTomato expressing cell line was used for experiments. K562 cells expressing dCas9-VPR were purchased from Horizon Discovery/Perkin

Elmer (catalog ID: HD dCas9-VPR-005), and these cells were tested for CRISPRa activity using the same tdTomato expression assay described above.

iPSC-derived neurons

Human WTC11 iPSCs equipped with a doxycycline-inducible *NGN2* transgene expressed from the *AAVS1* safe-harbor locus as well as an ecDHFR-dCas9-VPH construct expressed from the CLYBL safe-harbor locus were a gift from the Kampmann lab (Tian et al., 2021). These cells were tested for CRISPRa activity using the same tdTomato expression assay that was used to validate the K562 cell lines, which is described above.

gRNA selection and design

A complete breakdown of gRNA library contents and overview of the gRNA design pipeline is illustrated in Figure S1. Briefly, enhancer-targeting gRNAs were selected from our CRISPRi library (Gasperini et al., 2019; McKenna & Shendure, 2018). Specifically, 50 spacer sequences (2 per candidate enhancer) were randomly selected from the list of 664 significant "hit" enhancer-gene pairs in the at-scale library. Another 50 spacer sequences targeting an additional 25 candidate enhancers (again 2 per candidate enhancer) were randomly selected from candidate enhancer non-hits (*i.e.*, gRNAs from the at-scale library targeting candidate enhancer regions with strong biochemical marks predictive of regulatory activity that did not yield significant downregulation of any neighboring genes in our previous CRISPRi study). An additional 30 TSS-positive control gRNAs were randomly sampled from the top quartile of gRNAs recommended by Horlbeck *et al*. (hCRISPRa-v2 library) (Horlbeck et al., 2016). 50 NTC negative control spacer sequences were

also selected from the hCRISPRa-v2 library (Horlbeck et al., 2016). The 313 candidate promoter targeting gRNAs were either selected from the Horlbeck *et al.* library (Horlbeck et al., 2016) or designed using FlashFry (McKenna & Shendure, 2018) (Figure S1). Briefly, 50 candidate promoters of 9 NDD risk genes (*TCF4*, *FOXP1*, *SCN2A*, *CHD8*, *BCL11A*, *TBR1*, *SHANK3*, *SYNGAP1*, *ANK2*) (Fu et al., 2022; Satterstrom et al., 2020) were pulled from Basic GENCODE annotations (Frankish et al., 2019) and were filtered for "type" == "transcript" and "transcript_type" == "protein coding". Separate bed files were generated for all promoter regions defined as the 500bp upstream of each protein coding transcript. Careful attention was paid to the strand orientation of each transcript when annotating promoter regions. Bed files were sorted and merged to combine multiple promoters with >1bp overlap into a single promoter annotation. Transcript bounds provided for each merged promoter begin +1bp from the end of the promoter and end at the position corresponding to the longest transcript mapping to that promoter. NGG-protospacer within these candidate promoters were identified using FlashFry and subsequently scored using default parameters (see FlashFry manuscript and user guide for a complete description of scoring metrics/algorithms) (McKenna & Shendure, 2018). A TSS-distance metric was then calculated for each gRNA using human fetal brain 5' Capped Analysis of Gene Expression (CAGE) data (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014; Lizio et al., 2019) obtained from FANTOM (https://fantom.gsc.riken.jp/5/sstar/FF:10085-102B4; CTSS, hg38). First, the strongest FANTOM annotated TSS was identified within each +/-500 bp region up and downstream of each hg38 Gencode Basic protein coding transcript TSS. For regions with a tie between the highest scoring FANTOM TSSs, the TSS position closest to Gencode annotated TSS position was prioritized. Each candidate sgRNA from FlashFry was annotated with the distance to the nearest FANTOM TSS using the command "bedtools closest -a

sgRNAs_with_fantom_tss -b strongest_fantom_tss_within_gencode_promoter -D b -t first." For Gencode Basic protein coding transcripts without a human fetal brain FANTOM peak within 500 +/- bp, the distance of each sgRNA to the nearest Gencode TSS was reported instead. A distance of zero indicates that an sgRNA overlaps with the nearest annotated TSS. Multiple rounds of successively relaxing score and distance thresholds were then iterated until the top 4 gRNAs for each candidate promoter were selected (five selection rounds in total). Optimal TSS-distances were approximated using genome-wide CRISPRa design rules (Sanson et al., 2018). gRNAs flagged for potentially problematic polythymidine tracks or GC content were excluded. The gRNA selection criteria used in each round were as follows:

Round 1: 1. TSS Distance between -150 and -75 BP 2. Doench2014OnTarget >= 0.2 3. Dangerous_in_genome <= 1 4. Hsu2013 > 80.

Round 2: 1. TSS Distance between -400 and -50 BP 2. Doench2014OnTarget >= 0.2 3. Dangerous_in_genome <= 1 4. Hsu2013 > 80.

Round 3: 1. TSS Distance between -400 and -50 BP 2. Doench2014OnTarget >= 0.2 3. Dangerous_in_genome <= 1 4. Hsu2013 > 50.

Round 4: 1. TSS Distance between -400 and -50 BP 2. Doench2014OnTarget >= 0.2 3. Dangerous_in_genome <= 2 4. Hsu2013 > 50.

Round 5: 1. Doench2014OnTarget >= 0.2 2. Dangerous_in_genome <= 2 3. Hsu2013 > 10 4. DoenchCFD_maxOT < 0.95

Complete oligo sequences with gRNA spacers and additional sequences for cloning into piggyFlex are listed in Table S1 (all tables are available here:

https://krishna.gs.washington.edu/content/members/CRISPRa_QTL_website/public/). Note all gRNAs in our library are designed/modified to start with a G followed by the 19 base pair spacer to facilitate Pol III transcription.

gRNA library cloning into piggyFlex vector

The 493 gRNAs with associated 10N random barcodes were ordered as an IDT oPool and PCR amplified with Q5 High-Fidelity polymerase (NEB, Cat. No. M0491S) to make double stranded DNA. The piggyFlex backbone vector was digested with SalI (NEB, Cat. No. R3138S) and BbsI (NEB, Cat. No. R0539S) in 10X NEBuffer r2 at 37 degrees Celsius overnight to ensure complete digestion of the backbone. This digestion cuts out the EF1a-puro-GFP cassette of the vector which is then added back in a later cloning step. The digestion product was run on a 1% agarose gel in TAE buffer, and the linear backbone vector (5098 base pairs in size) was gel extracted using a gel extraction kit (NEB, Cat. No. T1020S). The second product from the digestion (2878 base pairs) which contains the EF1a-puro-GFP cassette was saved for a later assembly reaction in the final cloning step (described below). The PCR amplified IDT oPool gRNAs with associated 10N random barcodes were cloned into the linear backbone using NEBuilder HiFi DNA Assembly (NEB, Cat. No. E2621S) using 0.15 pmol of the insert (gRNA library) and 0.02 pmol of the linear backbone. Assembled product was transformed into electrocompetent cells (NEB, Cat. No. C3020K) and plasmid DNA was extracted with a midiprep kit (Zymo Research, Cat. No. D4200). The resulting vector was then digested with SapI (NEB, Cat. No. R0569S), for one hour at 37 degrees Celsius. Digested product was cleaned with 0.5X AMPure beads (Beckman Coulter, Cat. No. A63880) and cleaned digested linear backbone was used for a subsequent assembly reaction to add the EF1a-puro-GFP cassette back into the final piggyFlex vector between the gRNA

sequence and the 10N random barcode sequences. 0.014 pmol of the linear backbone was assembled with 0.056 pmol of the insert sequence and the assembly reaction was cleaned with a 0.5X AMPure step. The assembled product was transformed into electrocompetent cells and plasmid DNA was extracted with a midiprep kit. The final plasmid library was subsequently PCR amplified and sequenced to ensure that all 493 gRNAs were successfully cloned into the piggyFlex vector. Note: The 10N barcode is an additional gRNA identification strategy that can be used to assign gRNAs to cells, however, we used directly sequenced gRNAs (from the 10x Genomics capture sequence) to identify gRNAs in this work as this more accurately assigns gRNA transcripts to cells (Replogle et al., 2020).

Transfection of the gRNA library and selection for transfected cells

K562 cells

16 million K562 cells (8 million K562-VP64 cells and 8 million K562-VPR cells) were transfected with the gRNA library and the piggyBac transposase (System Biosciences, Cat. No. PB210PA-1) at a 20:1 molar ratio of library:transposase using a Lonza 4D nucleofector and the Lonza nucleofector protocol for K562 cells. The 16 million cells were split across 8 100 uL nucleofection cartridges, with each individual nucleofection cartridge receiving 2 million cells and 2 ug of total DNA. After nucleofection, cells were transferred to pre-warmed RPMI media in a cell culture flask and incubated at 37 degrees Celsius. One day after transfection, cells were selected with 2 ug/mL puromycin (GIBCO/Thermo Fisher Scientific; Cat. No. A1113803). After 9 days, cells were harvested for single-cell transcriptome profiling.

Induced pluripotent stem cells

6 million dCas9-VPH iPSCs (same cells as described above) were transfected with the gRNA library and the piggyBac transposase at a 5:1 molar ratio of library:transposase using the Lonza nucleofector and the Lonza nucleofector CB-150 program. The 6 million cells were split across 6 100 uL nucleofection cartridges, with each individual nucleofection cartridge receiving 1 million cells and 17.5 ug of total DNA. After nucleofection, cells were transferred to pre-warmed mTeSr Plus basaI medium with ROCK inhibitor in a cell culture flask and incubated at 37 degrees Celsius. One day after transfection, cells were selected with 20 ug/mL puromycin (note: the AAVS1-NGN2 construct has a puromycin resistance cassette on it, so a higher dose of puromycin was used to successfully select for cells that received a gRNA in the presence of an existing puromycin resistance cassette). Media changes were performed daily (mTeSr Plus basaI medium with ROCK inhibitor and 10 ug/mL puromycin) for seven days prior to initiating neuron differentiation (described in "Human iPSC-derived neuron cell culture and differentiation" methods section).

10x Genomics sc-RNA-seq with associated gRNA transcript capture

K562 screen

Cells were harvested and prepared into single-cell suspensions following the 10x Genomics Single Cell Protocols Cell Preparation Guide (Manual part number CG00053, Rev C). Four lanes were used for the single-cell transcriptome profiling, with two lanes containing cells from the K562-VP64 cell line, and two lanes containing cells from the K562-VPR cell line. Roughly 10,000 cells were captured per lane of a 10x Chromium chip (Next GEM Chip G) using Chromium Next GEM

Single Cell 3′ Reagent Kits v3.1 with Feature Barcoding technology for CRISPR Screening (10x Genomics, Inc, Document number CG000205, Rev D).

iPSC-derived neuron screen

iPSC-derived neurons were harvested and prepared into single-cell suspensions following a published protocol (Jerber et al., n.d.). Cells were split into two batches, with one batch going through a fluorescence-activated cell sorting (FACS) step to sort on the top 40% of green fluorescent protein (GFP) expression to enrich for neurons with greater numbers of gRNAs integrated, and the second batch going directly into the 10x Genomics single-cell library preparation protocol. Sorting on the top 40% of GFP expression resulted in a two-fold increase in the mean number of gRNAs integrated in those cells as compared to unsorted cells. Four lanes were used for the single-cell transcriptome profiling, with two lanes containing GFP-positive sorted cells, and two lanes containing unsorted cells. Roughly 13,000 cells were captured per lane of a 10x Chromium high-throughput chip (Next GEM Chip M) using Chromium Next GEM Single Cell 3' HT Reagent Kits v3.1 (Dual Index) with Feature Barcode technology for CRISPR Screening (10x Genomics, Inc, Document number CG000418, Rev C).

Sequencing of scRNA-seq libraries

Final libraries were sequenced on a NextSeq 2000 P3 100 cycle kit (R1:28 I1:10, I2:10, R2:90) for each screen (K562 and iPSC-derived neuron screens). Gene expression and gRNA transcript libraries were pooled at a 4:1 ratio for sequencing.

Transcriptome data processing and quality control filtering for K562 and iPSC-derived neuron screens

CellRanger version 6.0.1 was used to perform bcl2fastq and count matrix generation. CellRanger mkfastq was run using default parameters, and CellRanger count was run using the GRCh38-3.0.0 reference transcriptome from 10x Genomics and default parameters. For the K562 screen, cells with greater than 10% mitochondrial reads and less than 4000 UMIs were filtered out. For the iPSC-derived neuron screen, cells with greater than 17% mitochondrial reads and less than 1500 unique molecular identifiers (UMIs) were filtered out. After quality control filtering, 33,944 cells were retained in the K562 screen, and 51,183 cells were retained in the iPSC-derived neuron screen. The resulting count matrix output after this filtering was used for all downstream analyses.

Neuron differentiation transcriptome projection

Single-cell transcriptome data from a time course study of iPSC-derived neurons (C. Wang et al., 2017) was downloaded from https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10632 (Accession No. E-MTAB-10632, matrices_timecourse.tar.gz), and integrated with the neuron CRISPRa screening dataset described here. Seurat v4 was used for all data analyses (Hao et al., 2021). The CRISPRa dataset was randomly downsampled to 5,000 cells for this analysis. Count matrices from both matrices were filtered to include only shared genes from the two datasets (n=14,777 genes). SelectIntegrationFeatures() and FindIntegrationAnchors() were run using default parameters to identify anchors for integration. 20,606 anchors were found and 2,953 anchors were retained for data integration. IntegrateData() was run using the retained 2,953 anchors to integrate the two datasets. After integration, standard Seurat single-cell analysis was performed to scale the data, and run the PCA and UMAP algorithms.

gRNA assignment and differential gene expression testing

Genomic coordinates (hg38) for final gRNA spacers were isolated using a loop built around the matchPattern() function from the BSgenome package (Pagès, n.d.). A 2Mb window (1Mb upstream and downstream) around each gRNA was then calculated and all genes within the 2Mb window were isolated using a loop built around ENSEMBL biomaRt getBM() function (Durinck et al., 2005, 2009). These 1Mb neighboring gene sets were then filtered to unique entries (unique HGNC symbols) for compatibility with the Seurat FindMarkers() function used in DE testing.

A global UMI filter of 5 gRNA UMIs/cell was used to assign gRNAs to single cell transcriptomes for both K562 and iPSC-derived neuron datasets (note this heuristic threshold was chosen based on manual inspection of the UMI count distributions for each gRNA and prior work) (Gasperini et al., 2019). gRNA UMI counts for each cell were derived from the count matrix of passing cells output by CellRanger (which applies an automatic total UMI threshold to cells) and which also passed QC.

Expression of a given gene within 1Mb of the gRNA of interest was compared between all cells with a given gRNA and all other cells as control. log2() fold changes in expression for a given gene were calculated using the Seurat FindMarkers() function with the following arguments: ident.1 = gRNA_Cells, ident.2 = Control_Cells, min.pct = 0, min.cells.feature = 0, min.cells.group = 0, features = target_gene, logfc.threshold = 0. A Wilcoxon rank-sum test was used to generate raw differential expression *P*-values. This process was then iterated for all genes within 1Mb of all gRNAs. NTCs were tested against all genes within 1Mb of any targeting gRNA. Only tests involving genes detected in >0.2% of test gRNA and control cells were carried forward.

These raw differential expression *P*-values were then used to calculate empirical P-values to call EFDR < 0.1 sets (Gasperini et al., 2019). Specifically, an empirical *P*-value was calculated for each gRNA-gene test as:

[(the number of NTCs with a P-value lower than that test's raw P-value) + 1] /

[the total number of NTCs tests + 1]

Empirical *P*-values were then Benjamini-Hochberg corrected, and those < 0.1 were kept for 10% EFDR sets.

Log2 fold changes between gRNA and control cells were visualized using the gviz package (Hahne & Ivanek, 2016) along with tracks for RefSeq transcripts (ENSEMBL biomaRt), H3K27ac, and ATAC seq peaks. The K562 ATAC and H3K27ac data were downloaded from ENCODE (Rosenbloom et al., 2013). ATAC-seq and H3K27ac CUT&RUN data from 7-8 week old NGN2-iPSC inducible excitatory neurons was obtained from Song et al. 2019 (Song et al., 2019). As previously described, ATAC-seq and CUT&RUN reads were trimmed to 50bp using TrimGalore with the command --hardtrim5 50 before alignment (https://github.com/FelixKrueger/TrimGalore). ATAC-seq reads were realigned to hg38 using the standard Encode Consortium ATAC-seq and ChIP-seq pipelines respectively with default settings and pseudo replicate generation turned off. Trimmed, sorted, duplicate and chrM removed ATAC-seq bam files from multiple biological replicates were combined into a single bam file using samtools merge v1.10 (H. Li et al., 2009). Trimmed CUT&RUN reads were realigned to hg38

using Bowtie2 v2.3.5.1 with the following settings --local --very-sensitive-local --no-mixed --no-discordant -I 10 -X 700 and output sam files were convert to bam format using samtools view (Langmead & Salzberg, 2012; H. Li et al., 2009). Duplicated reads were removed from the CUT&RUN bam file using Picard MarkDuplicates v2.26.0 with the --REMOVE_DUPLICATES =true and --ASSUME_SORTED=true options (http://broadinstitute.github.io/picard/). Finally, bam files were converted using the bedtools genomecov followed by the UCSC bedGraphToBigWig utility.

Chapter 3. A MULTIPLEX, PRIME EDITING FRAMEWORK FOR IDENTIFYING DRUG RESISTANCE VARIANTS AT SCALE

Chapter 3 has been adapted with minimal modification from:

Florence M. Chardon, Chase C. Suiter, Riza Daza, Nahum T. Smith, Phoebe Parrish, Troy A. McDiarmid, Jean-Benoît Lalanne, Beth Martin, Diego Calderon, Amira Ellison, Alice H. Berger, Jay Shendure, and Lea M. Starita. "A multiplex, prime editing framework for identifying drug resistance variants at scale." *Nature Biotechnology (submitted), (*2023).

2.1 ABSTRACT

CRISPR-based genome editing has revolutionized functional genomics, enabling screens in which thousands of perturbations of either gene expression or primary genome sequence can be competitively assayed in single experiments. However, for libraries of specific mutations, a challenge of CRISPR-based screening methods such as saturation genome editing is that only one region (*e.g.* one exon) can be studied per experiment. Here we describe prime-SGE ("prime saturation genome editing"), a new framework based on prime editing, in which libraries of specific mutations can be installed into genes throughout the genome and functionally assessed in a single, multiplex experiment. Prime-SGE is based on quantifying the abundance of prime editing guide RNAs (pegRNAs) in the context of a functional selection, rather than quantifying the mutations themselves. We apply prime-SGE to assay thousands of single nucleotide changes in eight oncogenes for their ability to confer drug resistance to three EGFR tyrosine kinase inhibitors.

Although currently restricted to positive selection screens by the limited efficiency of prime editing, our strategy opens the door to the possibility of functionally assaying vast numbers of precise mutations at locations throughout the genome.

2.2 INTRODUCTION

Resistance to targeted therapies is a major barrier to the successful treatment of many cancers, including non-small cell lung cancer (Glickman & Sawyers, 2012; Konieczkowski et al., 2018; Vasan et al., 2019). Both *de novo* and acquired resistance to therapeutic small molecules can be caused by mutations in target genes that inhibit drug binding, such as with *EGFR* resistance mutations to covalent EGFR inhibitors. Alternatively, resistance can develop by activating bypass signaling through mutation or amplification of other oncogenes (Reita et al., 2021). Pinpointing resistance mutations and developing novel drugs that circumvent them is crucial for continued progress in the treatment of cancer (Vasan et al., 2019). Mutations associated with resistance can be identified retrospectively in clinical specimens from repeat biopsies or circulating tumor DNA (Clark et al., 2018; Frampton et al., 2013; Garcia-Murillas et al., 2015; Zehir et al., 2017). However, obtaining these specimens can be challenging (Overman et al., 2013) and causation of resistance must still be verified with functional experiments (Awad et al., 2021).

Ideally, resistance mutations would be identified prospectively as part of the clinical development of each drug (Vasan et al., 2019). Three complementary approaches have been taken to model drug resistance *in vitro*. First, cells can be cultured in the presence of a drug to allow resistant mutations to arise spontaneously, outcompete other cells until they rise to an appreciable frequency, and be identified by whole genome or targeted sequencing (Ramirez et al., 2016). Second, wild type and

mutant versions of oncogenes or tumor suppressors can be overexpressed to measure their capacity to provide resistance (Berger et al., 2017). Finally, deep mutational scans of open reading frames can identify all potential resistance mutations in candidate genes (Awad et al., 2021; Persky et al., 2020; Wagenaar et al., 2014). However, these approaches are limited in various ways. For example, evolutionary selections are subject to chance, and will likely only identify a subset of potential resistance mutations, while the latter approaches query transgenes rather than mutations in the endogenous genome. An ideal approach would concurrently introduce and characterize vast numbers of potential resistance mutations in genes of interest in their endogenous genomic context. This would allow for upfront identification of resistance mutations to drugs during early stage development, and possible repurposing of agents that can overcome drug resistant variants.

To functionally assess single nucleotide variants or specific small insertions or deletions at scale via CRISPR, a powerful approach is to leverage libraries of homology-directed repair (HDR) templates to install mutations of interest to a short region of interest, *e.g.* an exon. However, the major limitation of this strategy (also known as saturation genome editing (SGE) (Buckley et al., 2023; Findlay et al., 2014, 2018; Meitlis et al., 2020; Radford et al., 2022)) is that only one region can be studied per experiment, both because the HDR template library relies on a locus-directing gRNA, and because deciphering which mutation was installed in which cell requires sequencing of the edited locus itself. As such, evaluating candidate resistance mutations in many exons of many genes in response to a panel of compounds via HDR-based saturation genome editing would be highly labor intensive. Saturating scans of multiple exons or genes are possible with base editing, but these lack precision (and range) with respect to which mutations are (or can be) introduced (Hanna et al., 2021; Martin-Rufino et al., 2023).

Prime editing (PE) is a genome editing method that allows for the precise installation of insertions, deletions, and single base substitutions using a nicking Cas9 fused to a reverse transcriptase (prime editor) and a single prime editing gRNA (pegRNA) (Anzalone et al., 2019). Using PE to assay the effects of single nucleotide variants has two major advantages over other CRISPR-based genome editing methods. First, it allows for multiplex experiments in which any number of mutations in any number of genes can be programmed (Ren et al., 2023). This is due to the unique design of the pegRNA, which couples both the target site and programmed edit in a single molecule (Anzalone et al., 2019). Second, this approach takes advantage of the precision of PE, which does not rely on disruptive DNA double strand breaks and exhibits much lower rates of both on-target, intended edits as well as off-target edits, relative to other approaches such as CRISPR/Cas9-mediated HDR or base editing (Anzalone et al., 2019; Hanna et al., 2021). However, in a recent study deploying PE for multiplex genome editing, the PE outcomes were read out from the edited locus itself (Erwood et al., 2022). This undercuts the first advantage of PE, as targeting of multiple exons or genes would require each to be serially amplified and sequenced, which increases input requirements and is impractical to scale genome wide.

Here, we present prime-SGE, a scalable, multiplex prime editing framework in which we introduce and assess thousands of precise edits simultaneously in PC-9 lung adenocarcinoma cells. We devised a positive selection strategy to overcome the low editing rate of PE and select for edits that provide resistance to various EGFR inhibitors. In this framework, each cell is engineered to harbor, on average, a single pegRNA encoding a precise edit. This pool of cells is then subjected

to tyrosine kinase inhibitor (TKI) drug treatment or a vehicle control. Cells are harvested over a time course of two to three weeks. A key point is that in contrast with other genome editing-based mutational scans (Buckley et al., 2023; Findlay et al., 2014, 2018; Hanna et al., 2021; Meitlis et al., 2020; Radford et al., 2022; Ren et al., 2023), prime-SGE achieves readout of the functional selection by quantifying the abundance of prime editing guide RNAs (pegRNAs), rather than the mutated loci. In this proof-of-concept, we found that deep sequencing of the integrated pegRNAs allows for the identification of programmed mutations that give rise to drug resistant cells (Figure 23). Specifically, prime-SGE was able to resolve well-characterized resistance mutations, such as EGFR C797S and KRAS G12C (Reita et al., 2021), and also identified potentially novel, previously uncharacterized resistance mutations.

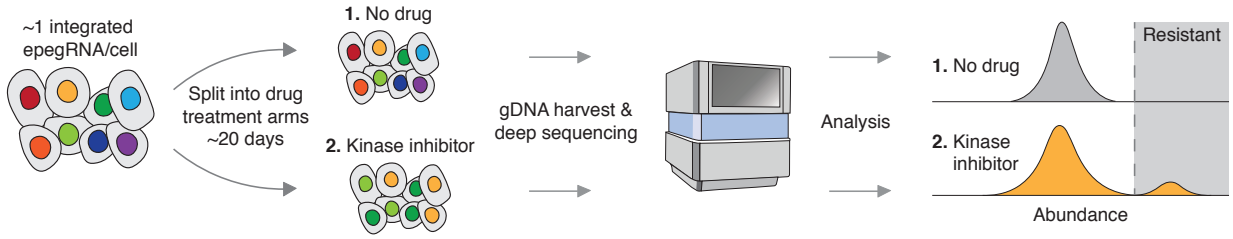

Figure 23. Prime-SGE identifies drug resistance variants at scale.

A library of prime editing gRNAs (pegRNAs) are lentivirally transduced into a pool of cells at a low multiplicity of infection (MOI) such that the majority of cells harbor a single pegRNA. This pool of cells is then subjected to treatment with DMSO (no drug) or one of three kinase inhibitors, and cultured for a period of ~20 days (~10 cell doublings). Integrated pegRNAs are amplified from genomic DNA and sequenced. Read counts are analyzed to determine the distribution of variant abundances across different treatment conditions.

2.2 PRIME EDITING OF THE OSIMERTINIB RESISTANCE MUTATION C797S IN EGFR

As a first experiment, we sought to introduce a single, well-characterized mutation to the *EGFR* gene that confers resistance to the small molecule TKI Osimertinib (Thress et al., 2015) via prime editing, and then ask whether we could detect its selection during an *in vitro* resistance screen. This mutation, a T to A base change at the first position of amino acid residue 797 of the EGFR open reading frame, changes the wild-type cysteine residue to a mutant serine residue. Osimertinib is a third-generation TKI that targets the ATP binding pocket of EGFR by covalently binding the C797 residue (Cross et al., 2014) and is the current standard-of-care therapy for advanced stage EGFR-mutant lung cancer (Ettinger et al., 2023). The cysteine-to-serine change at position 797 blocks the binding of osimertinib and leads to drug resistance and poor survival outcomes (Thress et al., 2015). We performed all experiments in PC-9 cells, which are both addicted to EGFR signaling (M.-Y. Park et al., 2017) and sensitive to the TKI osimertinib, providing a model for identification of secondary mutations that confer resistance.

For this initial experiment, we designed three different pegRNAs programming the T to A base change at the first base of residue 797 in *EGFR*. We performed an arrayed experiment in which we transiently co-transfected plasmids expressing each of these pegRNAs and the PE2 prime editor (Anzalone et al., 2019), into wild type PC-9 cells. Two days later, the cells were treated with osimertinib, and 24 days after drug treatment, cells were harvested and the *EGFR* locus amplified and sequenced (Figure 24a). We observed a 16.8 to 25.3% frequency of T to A edits from reads overlapping the *EGFR* locus across the three pegRNAs tested in contrast with a 0.31% frequency of T to A edits in the wild type, untransfected control, presumably sequencing errors (Figure 24b).

As PC-9 cells harbor eight copies of *EGFR,* we estimate that on average, resistant cells had one to two mutated copies of *EGFR* after selection. This experiment confirmed the ability for three different pegRNAs, each encoding the same T to A mutation to 1) successfully program this mutation at a low but appreciable frequency, sufficient for subsequent selection; and 2) confer resistance to osimertinib.

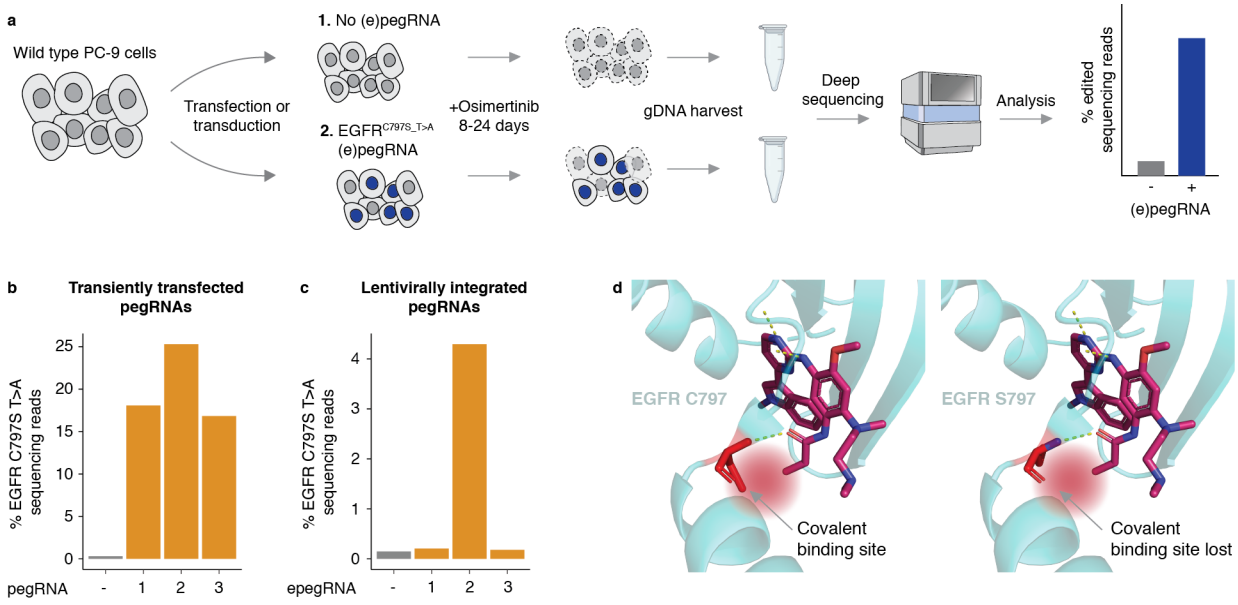

Figure 24. Proof of concept prime editing of EGFR C797S.

a) Wild type PC-9 cells are transiently transfected or lentivirally transduced with a pegRNA or epegRNA programming the EGFR C797S T>A mutation, along with the prime editor enzyme. Osimertinib is added to the cells for 24 days (transient transfection) or 8 days (lentiviral transduction) and genomic DNA is harvested and the EGFR locus is amplified and sequenced. b) Transient transfection of 3 pegRNAs into PC-9 cells leads to edited cells proliferating in the presence of osimertinib. c) Lentiviral transduction of 3 epegRNAs into PC-9 cells leads to successful editing and proliferation for 1 of 3 transduced epegRNAs. d) Left: Crystal structure of

EGFR covalently bound with osimertinib when residue 797 is a cysteine (wild type, Protein Data Bank identifier: 4ZAU) (. Right: Crystal structure of EGFR no longer covalently binding with osimertinib when residue 797 is mutated to a serine (C797S mutation).

Next, we sought to develop an experimental screening framework in which cells stably express pegRNAs, such that pegRNA identities can be read out by directly sequencing integrated pegRNAs in place of sequencing the edited locus. We modified the LentiGuide-Puro-P2A-EGFP vector (Panda et al., 2020) to allow for cloning of pegRNAs (Figure 25). In addition to this lentiviral integration strategy, we also employed the engineered pegRNA ("epegRNA") construct design that incorporates an RNA stabilizing motif (Nelson et al., 2022). We also switched from using the PE2 prime editor to using the PEmax prime editor to enable higher rates of editing (Chen et al., 2021). We cloned the same three pegRNAs that we used in the transient transfection experiment (Figure 24b) into this modified lentiviral vector, and individually transduced these lentiviral epegRNA constructs into wild-type PC-9 cells. Eight days after osimertinib treatment, cells were harvested and the *EGFR* locus was amplified and sequenced. One of the three virally delivered epegRNAs, which was also the pegRNA that resulted in the highest editing rate in our transient transfection experiment (Figure 24b), successfully edited the EGFR locus and conferred resistance to osimertinib treatment (Figure 24c-d). This experiment confirmed our ability to perform prime editing experiments using integrated epegRNAs, but also highlighted a key challenge, which is that some guides may fail to edit at appreciable frequencies. For all future experiments, we designed up to four epegRNAs per intended mutation.

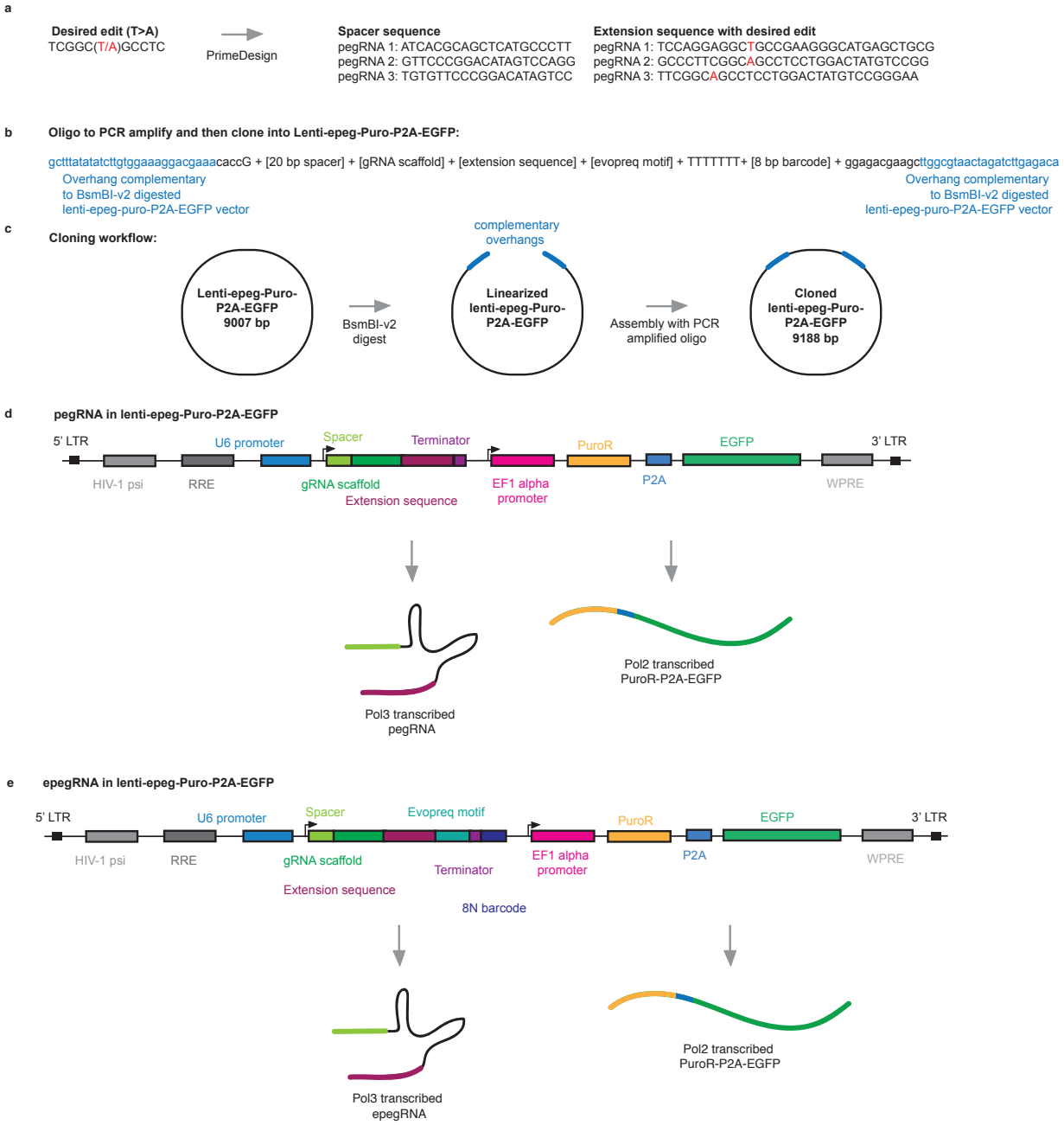

Figure 25. pegRNA design, cloning, and expression vector components.

a) Schematic of workflow to design pegRNAs with PrimeDesign(Hsu et al., 2021). Desired edit is in red. b) DNA sequence (oligo) to amplify and clone into lenti-epeg-Puro-P2A-EGFP. c) Schematic of cloning workflow. Lenti-epeg-Puro-P2A-EGFP is digested with BsmBI-v2, and the PCR amplified oligo from b) is assembled with the linearized vector via a Gibson assembly

reaction. d) Schematic of the lenti-epeg-Puro-P2A-EGFP vector with a pegRNA cloned into it. e) Schematic of the lenti-epeg-Puro-P2A-EGFP vector with an epegRNA cloned into it (the epegRNA contains an evopreq RNA stabilizing motif and an 8N barcode sequence 3' of the terminator sequence).

2.3 MULTIPLEX PRIME EDITING RESOLVES WELL-CHARACTERIZED RESISTANCE MUTATIONS IN BRAF, KRAS, EGFR, RIT1, MET, AND PIK3CA

Motivated by our ability to prime edit and select for cells harboring the EGFR C797S T>A mutation, we next performed a pilot screen to assess our ability to install mutations in multiple genes for drug resistance in a single pooled experiment, and to then detect these via guide sequencing (Figure 26a, b). We engineered *MLH1*ko-PEmax-PC-9, a PC-9 cell line with an *MLH1* knockout and integrated PEmax to increase the prime editing rate (Chen et al., 2021) (Figure 29). We then designed a library of 121 epegRNAs programming 35 mutations in six oncogenes (*EGFR*, *KRAS*, *PIK3CA*, *RIT1*, *BRAF*, and *MET*, Figure 26a). Nearly all of these mutations have been previously hypothesized to confer resistance to osimertinib, except for the EGFR T790M "gatekeeper" mutation, which was included as a control as it is a known sensitive mutation (Cross et al., 2014). We transduced *MLH1*ko-PEmax-PC-9 cells with this epegRNA library. After selecting for cells containing an integrated epegRNA with puromycin, we treated cells with osimertinib. We harvested cells at 21 and 26 days after the initiation of drug treatment. We sequenced both the integrated epegRNA lentiviral construct, as well as the endogenous loci for 31 of 35 programmed edits, to determine whether the frequency of sequenced epegRNAs matched the frequency of endogenous edits. This dual sequencing approach confirmed that the edits programmed by the various epegRNAs were creating the intended edits in the genome in the

untreated (DMSO) arm for 23 of the 31 mutations programmed for which we successfully amplified the endogenous locus, albeit at a low editing rate (between 0.1 and 0.9% per mutation). Further, this approach showed that the epegRNAs that were increasing in frequency in the drug screen were also increasing at the endogenous loci for eight of the 23 successfully introduced edits (Figure 26c, d), suggesting that sequencing the epegRNA can be a proxy for sequencing the edited locus. For the other programmed mutations, we did not observe such consistent increases between the DMSO control vs. osimertinib-treated cells. This is potentially due to the fact that some programmed prime edits were unsuccessful (as evidenced by the lack of detectable editing for 12 of the 35 programmed mutations) and/or because not every mutation included in this library may be a bonafide drug resistance mutation. Of note, for four of the programmed mutations (PIK3CA E453K G>A, MET 1010 splice site variant G>A, MET 1010 splice site variant T>C, and MET 1010 splice site AGGT deletion), we observed increases in epegRNA frequency between the DMSO control and drug conditions, but we did not successfully amplify these targets via PCR so do not have corroborating locus sequencing data.

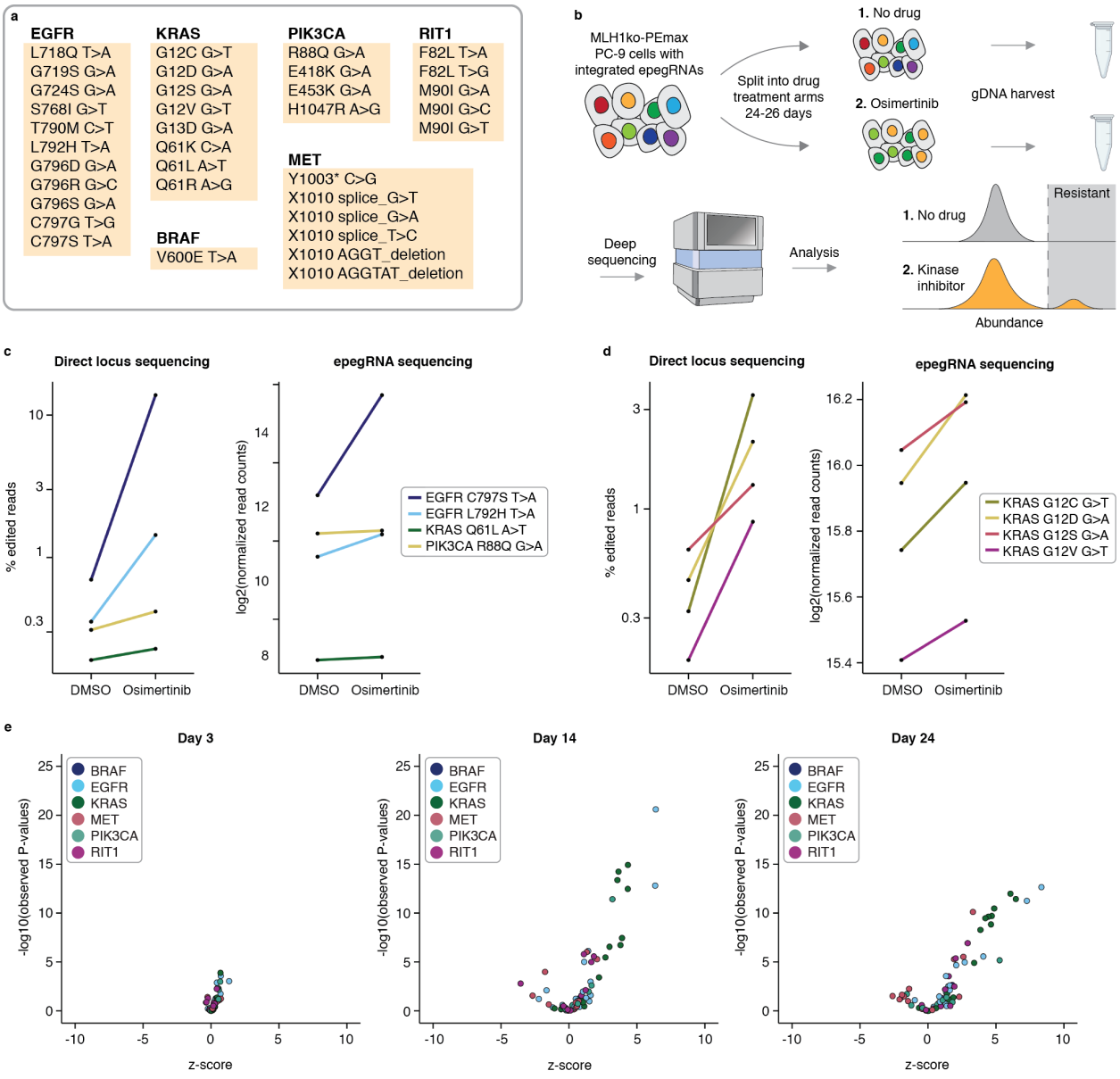

Figure 26. Proof of concept prime editing of EGFR, KRAS, BRAF, PIK3CA, MET, and RIT1.

a) List of 35 programmed edits in the 121 epegRNA pooled library, split by target gene. b) A pool of 121 epegRNAs was lentivirally integrated into *MLH1*ko-PEmax PC-9 cells and split into a no drug and an osimertinib drug treatment arm for 24 days. Genomic DNA is harvested and integrated epegRNAs are amplified and sequenced. c) Left: Direct locus sequencing results of targeted loci

in the epegRNA pooled library screen in the no drug (DMSO) and osimertinib drug treatment arms of selected loci. Right: epegRNA sequencing results of epegRNAs in the epegRNA pooled library screen in the no drug (DMSO) and osimertinib drug treatment arms for the epegRNAs that programmed the specified edits. d) Same as c) but for a different set of targets. e) Volcano plots showing the z-scores and p-values (unpaired two-sided t-test) for epegRNAs in the 121 epegRNA pooled library screen at days 3, 14, and 24.

In summary, from this screen, we identified 12 mutations across *EGFR*, *KRAS, PIK3CA* and *MET* that were enriched in the osimertinib-treated cells (log2FC>0 in the drug condition at day 21), including EGFR C797S, EGFR L792H, KRAS G12C, KRAS G12D, KRAS G12S, KRAS G12V, KRAS Q61L, PIK3CA E453K G>A, PIK3CA R88Q G>A, MET 1010 splice site variant G>A, MET 1010 splice site variant T>C, and a MET 1010 splice site AGGT deletion. The EGFR T790M control mutation did not confer resistance to osimertinib, as expected (log2FC=-0.06 in drug condition at day 21).

The results from this screen also exhibited the differential resistance phenotypes of these 12 resistant variants. EGFR C797S is the most well-documented and well-characterized osimertinib resistance mutation, and we observed that it vastly outcompetes all the other identified resistant variants in our screen. From the day 21 to day 26 timepoint, the frequency of the C797S variant rose from 12% to 56% in the epegRNA pool. As such, this differential resistance phenotype indicates that this screening framework can possibly rank variants by their degree of resistance by quantifying the relative fitness of a variant within a pool of cells of many variants, similar to

growth-based deep mutational scans (Fowler & Fields, 2014) and possibly even reflecting clonal competition that happens during tumor growth.

Using this same library of 121 epegRNAs, we performed a second screen to identify optimal drug concentrations and timepoints to select for edited cells. Cells were treated with three different concentrations of osimertinib (100, 300, and 500 nM), and we harvested cells at 3, 7, 10, 14, 17, 21, and 24 days after drug treatment to profile the rate at which variants were proliferating throughout the timecourse. Replicates were well correlated in this screen across all timepoints (Pearson's r = 0.73, 0.76, and 0.77 for 100, 300, and 500 nM screens, respectively Figure 28b-d). From this second screen, we identified 18 statistically significant drug resistant variants (log2FC>0, $p<0.05$, unpaired two-sided t-test between variants and EGFR T790M at days 14-24), including BRAF V600E, EGFR S768I, EGFR G796D, EGFR L792H, EGFR G796R, EGFR C797S, EGFR C797G, KRAS G12C, KRAS G12S, KRAS G12V, KRAS G12D, KRAS Q61L, RIT1 M90I, PIK3CA R88Q, PIK3CA E453K, a MET 1010 splice site G>T variant, a MET 1010 splice site T>C variant, and MET Y1003* early stop codon variant (Figure 26e, Figure 28a**)**. As expected, cells with the EGFR T790M mutation did not proliferate in the presence of osimertinib. We concluded that a 300 nM osimertinib treatment and a timepoint of 10 to 14 days, which corresponds roughly to 7 to 10 doublings (Figure 27), resulted in high signal in this screen (strictly standardized mean difference between EGFR C797S T>A and EGFR T790M = 14.1-14.4). Taken together, the results of these two pilot screens demonstrated that this experimental framework is capable of identifying mutations that confer resistance to TKIs and that sequencing epegRNAs can serve as an effective proxy for sequencing of the edited locus itself (Figure 26c-e).

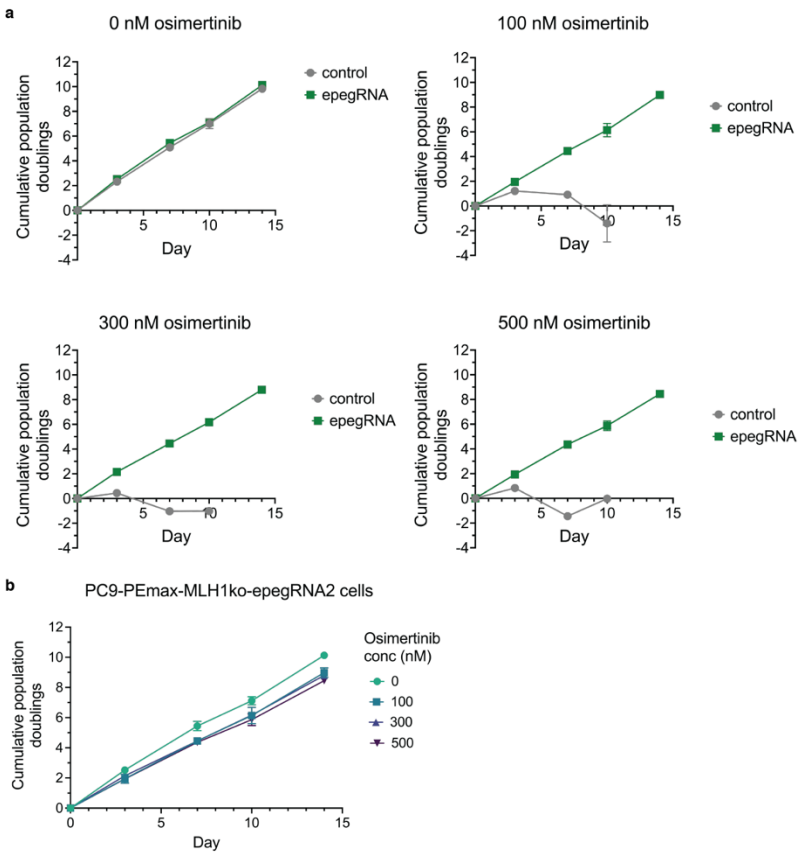

Figure 27. PC-9 cell drug dosing experiments with varying concentrations of osimertinib.

a) PC-9 cells with no integrated pegRNA ("control") or a pegRNA programming the EGFR C797S T>A osimertinib resistance mutation ("epegRNA"). Cells were treated with 0, 100, 300, and 500 nM osimertinib and cell population doublings were tracked over a period of 14 days. Data shown are the mean ± standard deviation of two biological replicates. b) Same data as in a). Cell population doublings of osimertinib resistant, EGFR C797S T>A harboring cells in 0, 100, 300, and 500 nM osimertinib.

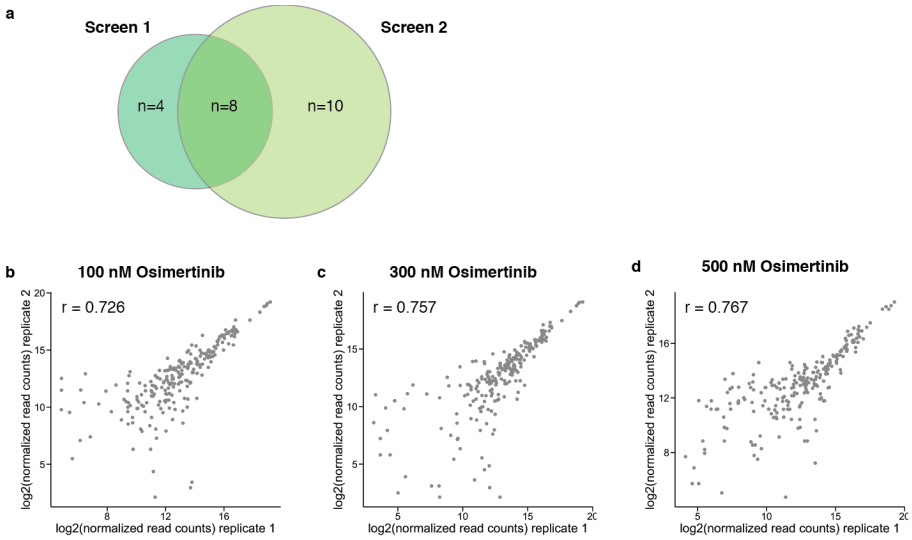

Figure 28. Overlap of resistant variants between 121 epegRNA screens and replicate correlation of second 121 epegRNA screen.

a) Overlap of hits between the first and second 121 epegRNA screens. b) Replicate correlation shown for 121 epegRNA screen at 100 nM osimertinib harvested at seven timepoints (days 3, 7, 10, 14, 17, 21, and 24). Each data point represents a single epegRNA at a single timepoint. c) Replicate correlation shown for 121 epegRNA screen at 300 nM osimertinib harvested at seven timepoints (days 3, 7, 10, 14, 17, 21, and 24). Each data point represents a single epegRNA at a single timepoint. d) Replicate correlation shown for 121 epegRNA screen at 500 nM osimertinib harvested at seven timepoints (days 3, 7, 10, 14, 17, 21, and 24). Each data point represents a single epegRNA at a single timepoint.

2.4 LARGE-SCALE TESTING OF DRUG RESISTANCE MUTATIONS WITH THREE INHIBITORS ACROSS SEVEN ONCOGENES

Osimertinib is the current standard therapy for non-small cell lung cancer patients harboring an EGFR T790M mutation. However, resistance to osimertinib typically develops, on average, within 10 months of treatment due to histological transformation or the acquisition of oncogene

amplifications or other resistance mutations (Jänne et al., 2015). Because of this, newer fourth-generation tyrosine kinase inhibitors have been developed to treat osimertinib-resistant non-small cell lung cancers. Two of these newer inhibitors include sunvozertinib and CH7233163. Sunvozertinib, which is currently in Phase 2 clinical trials and irreversibly and covalently binds EGFR at the C797 residue, is able to treat tumors that have an *EGFR* exon 20 insertion that renders these cells resistant to Osimertinib (M. Wang et al., 2022). CH7233163 is a next-generation *EGFR* inhibitor that is a non-covalent, competitive binder of the ATP binding pocket of EGFR (Kashima et al., 2020) and is currently in preclinical development for treatment of non-small cell lung cancers that are resistant to third-generation inhibitors (Du et al., 2021). Mechanistically, osimertinib and sunvozertinib are similar in that they irreversibly and covalently bind EGFR C797, and CH7233163 differs in that it is a non-covalent, competitive binder of the ATP binding pocket.

Recognizing the unique potential of prime-SGE to scale without the need to sequence each targeted locus, we next asked whether we could apply it to saturate exons and splice site regions of various known oncogenes to screen thousands of single nucleotide variants for resistance to multiple different TKIs in a single experiment. We designed 3,825 epegRNAs programming 1,220 single nucleotide mutations, both missense and synonymous, or deletions in seven different genes (*EGFR*, *KRAS*, *MET*, *RIT1*, *BRAF*, *MEK1*, and *AKT*, Figure 32a, b). In designing these epegRNAs, we also included a randomized eight nucleotide barcode directly 3' of the epegRNA terminator sequence (Figure 25e). The rationale for this barcode was to understand whether resistant cells were clonally derived or whether independent introductions of the same mutation recurrently resulted in resistance.

After generating lentivirus, cells were transduced in triplicate at >2,000X coverage into *MLH1*ko-PEmax-PC-9 cells at an MOI of 0.35 (Figure 30). After puromycin selection, cells were subjected to one of four treatment conditions: DMSO, CH7233163, osimertinib, or sunvozertinib (Figure 32a). Cells were harvested at days 3, 7, 11, 15, and 19, and genomically integrated epegRNAs were amplified and sequenced (Figure 32a). epegRNA read counts were correlated across replicates, timepoints and drug treatments (Figure 31).

We employed DESeq2 (Love et al., 2014) to identify hits from this large screen by identifying variants that are differentially abundant between the DMSO control and the three different drug treatments (CH7233163, osimertinib, and sunvozertinib) over the timecourse of 19 days. A likelihood ratio test (LRT) was performed between a reduced model that includes the variable time, and a full model that includes the variables time, drug treatment, and the interaction of drug treatment and time. Synonymous variants were used as controls for false discovery rate (FDR) testing (Figure 34b). From the results of this likelihood ratio test, we identified 17 differentially abundant variants in the osimertinib screen, 31 differentially abundant variants in the sunvozertinib screen, and 37 differentially abundant variants in the CH7233163 screen (FDR<0.01, log2FC>0, Figure 32c-e, Figure 37g).

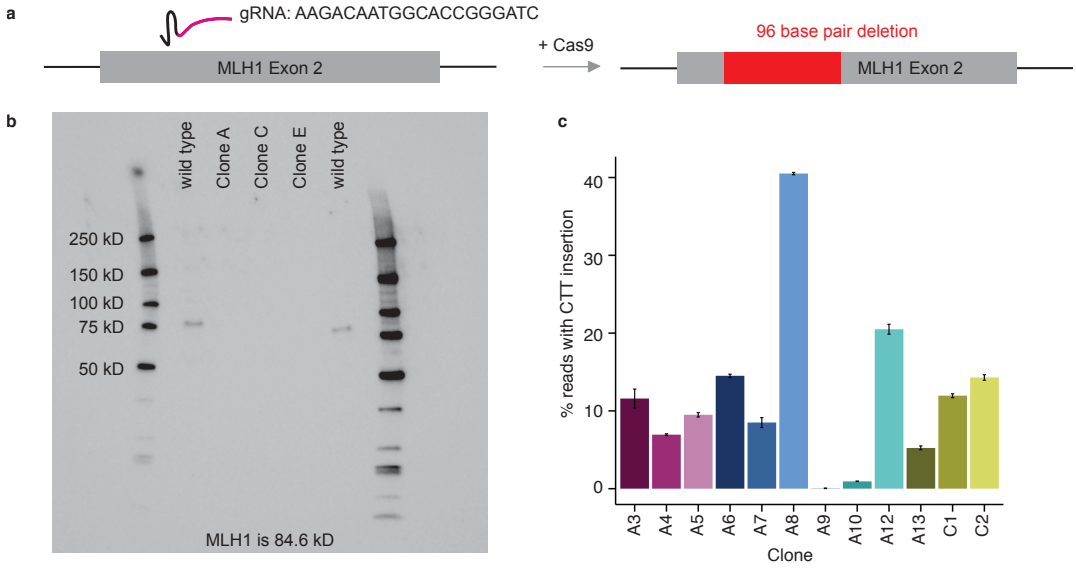

Figure 29. Improvements to prime editing efficiency via an *MLH1* knockout.

a) A single gRNA (in the pSpCas9 (BB)-2A-Puro vector) targeting exon 2 in *MLH1* was transfected into PC-9. This knockout led to a 96 base pair deletion in both copies of *MLH1* in PC-9 cells. b) Western blot analysis of three PC-9 knockout clones. All three clones (clones A, C, and E) show complete loss of the MLH1 protein. Wild type cells were run in parallel (lanes 1 and 5) and show presence of the MLH1 protein. c) 12 monoclonal *MLH1* knockout PC-9 cell lines were tested for insertion efficiency of a trinucleotide CTT insertion at the *HEK3* locus.

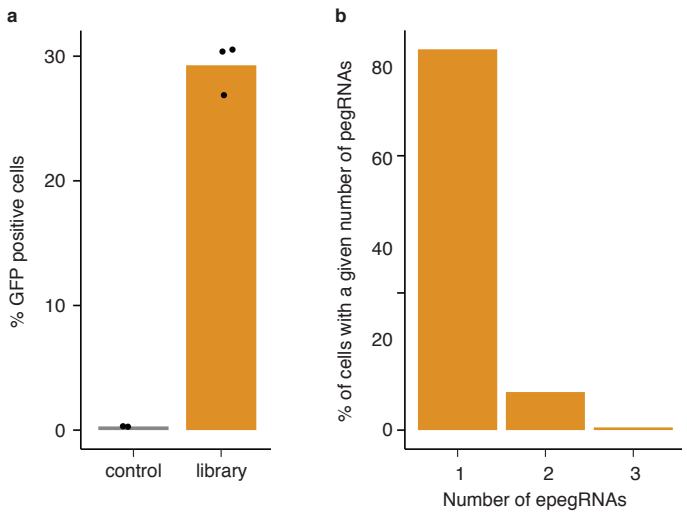

Figure 30. Lentiviral transduction of 3,825 epegRNA library for scaled screen.

a) PC-9-MLH1ko cells were analyzed by fluorescence activated cells sorting (FACS) to analyze the percentage of GFP+ cells following lentiviral transduction of the epegRNA library (GFP is expressed off the lentiviral epegRNA vector). Control cells were not transduced, and the library cells were transduced with the epegRNA library. The three data points represent three independent transduction replicates. b) Plot showing the percentage of cells harboring 1, 2, and 3 epegRNAs based off of the achieved MOI (~0.35).

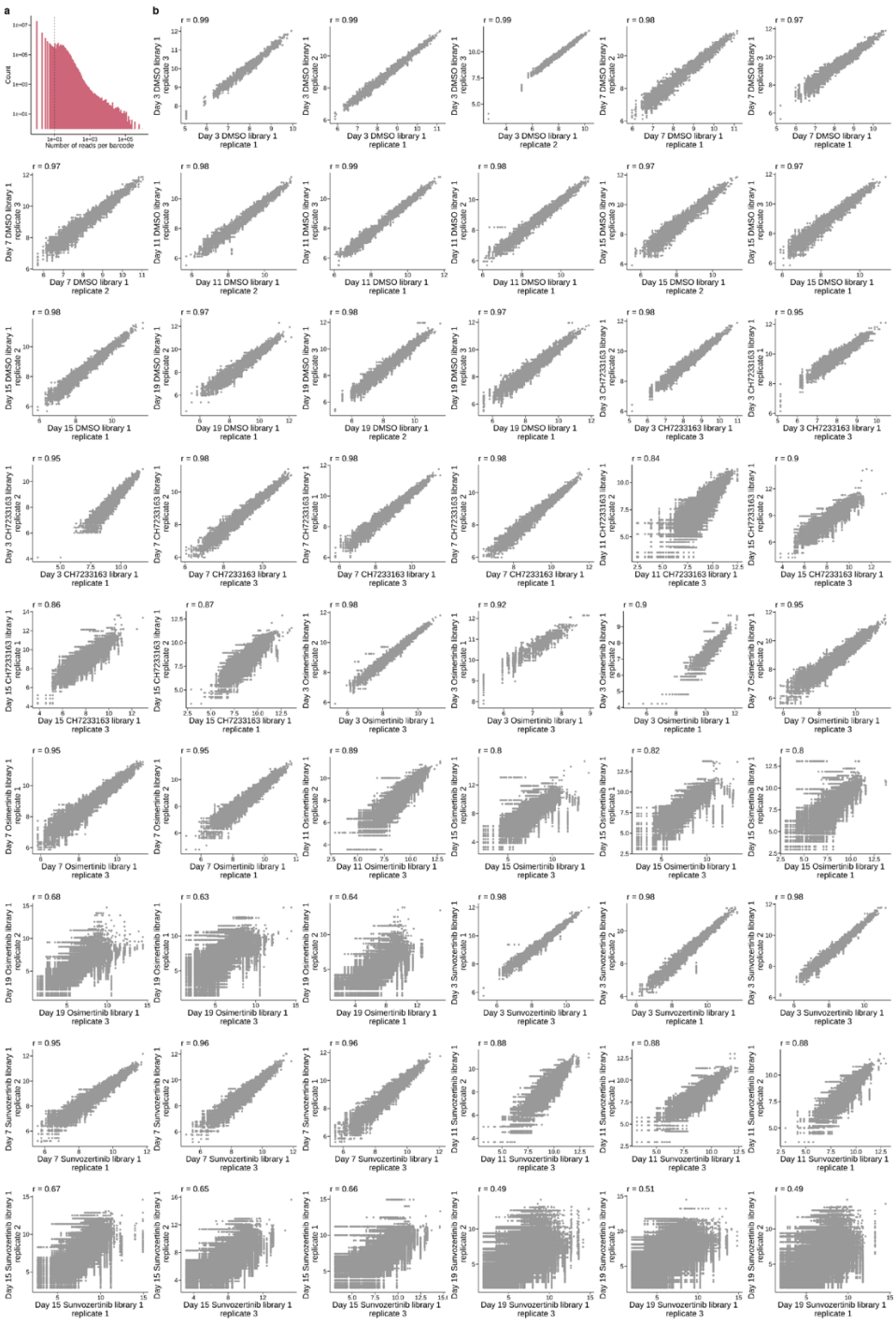

Figure 31. Barcode read count cutoff and replicate correlations in scaled screen.

a) Histogram showing the number of sequencing reads per barcode. A read cutoff of 10 was used for all analyses. b) Replicate correlation plots of log2 (normalized read counts) (Methods) of three independent transduction replicates in the three drug screens.

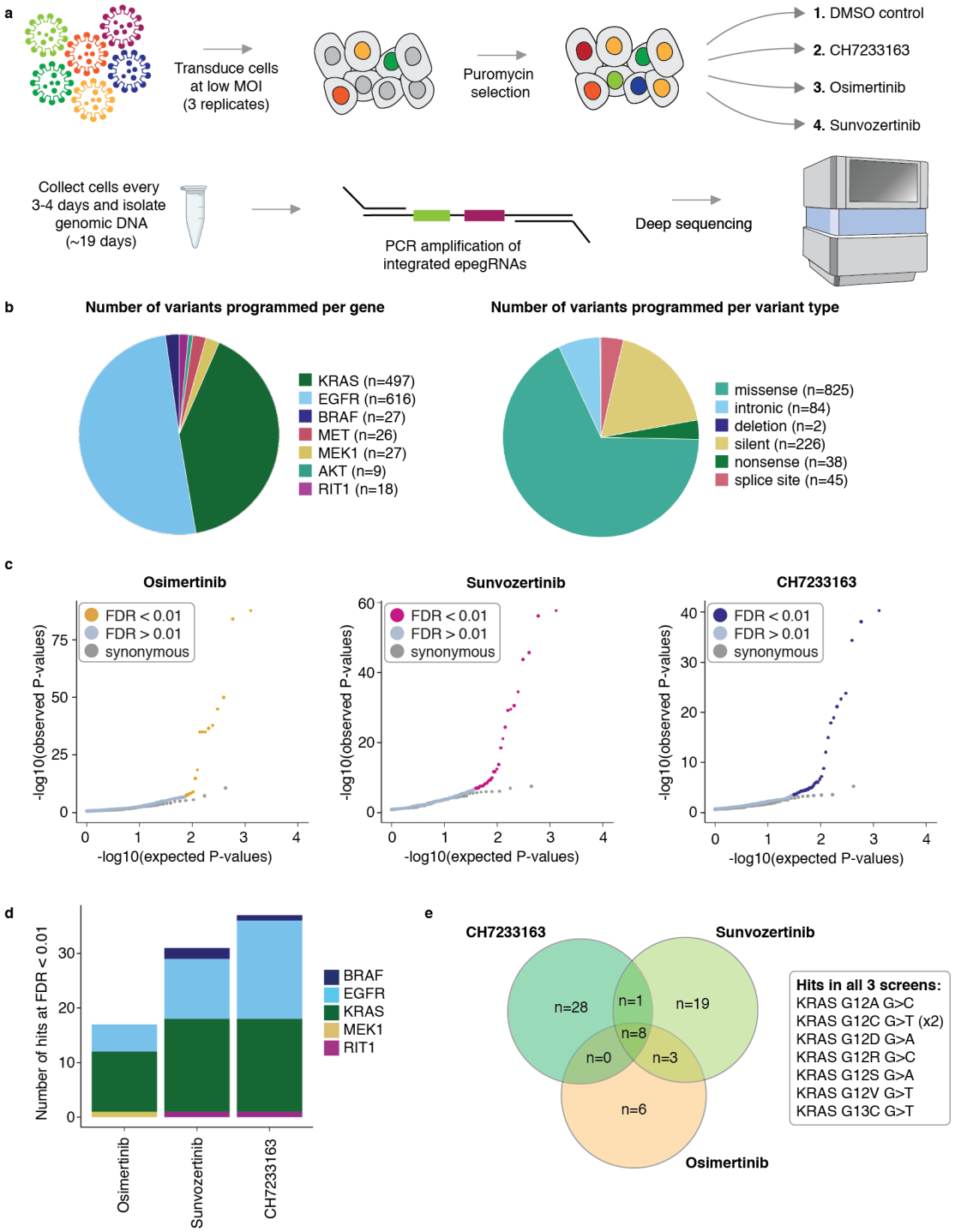

Figure 32. Drug resistance screen of 1,220 mutations against 3 different EGFR inhibitors.

a) A pool of 3,825 epegRNAs was lentivirally transduced into *MLH1*ko-PEmax PC-9 cells at a low MOI. Cells were selected with puromycin and split into one of four treatment arms (DMSO, CH7233163, osimertinib, and sunvozertinib). Cells were harvested every 3-4 days for over 19 days

and integrated epegRNAs were amplified and sequenced from genomic DNA. b) Left: Pie chart showing the number of variants programmed for each of seven genes. Right: Pie chart showing the number of variants programmed for each type of mutation. c) Quantile-quantile plots showing the distribution of expected versus observed p-values for the three different drug screens. Each point represents a unique genetic variant encoded by 1-4 epegRNAs. d) Stacked bar plot showing the number of resistant hits in each screen, colored by gene. Hits are variants that are at an FDR<0.01 and a log2FC>0 by DESeq2 differential epegRNA abundance analysis. e) Venn diagram showing the overlap of resistant hits between the three different drug treatment screens.

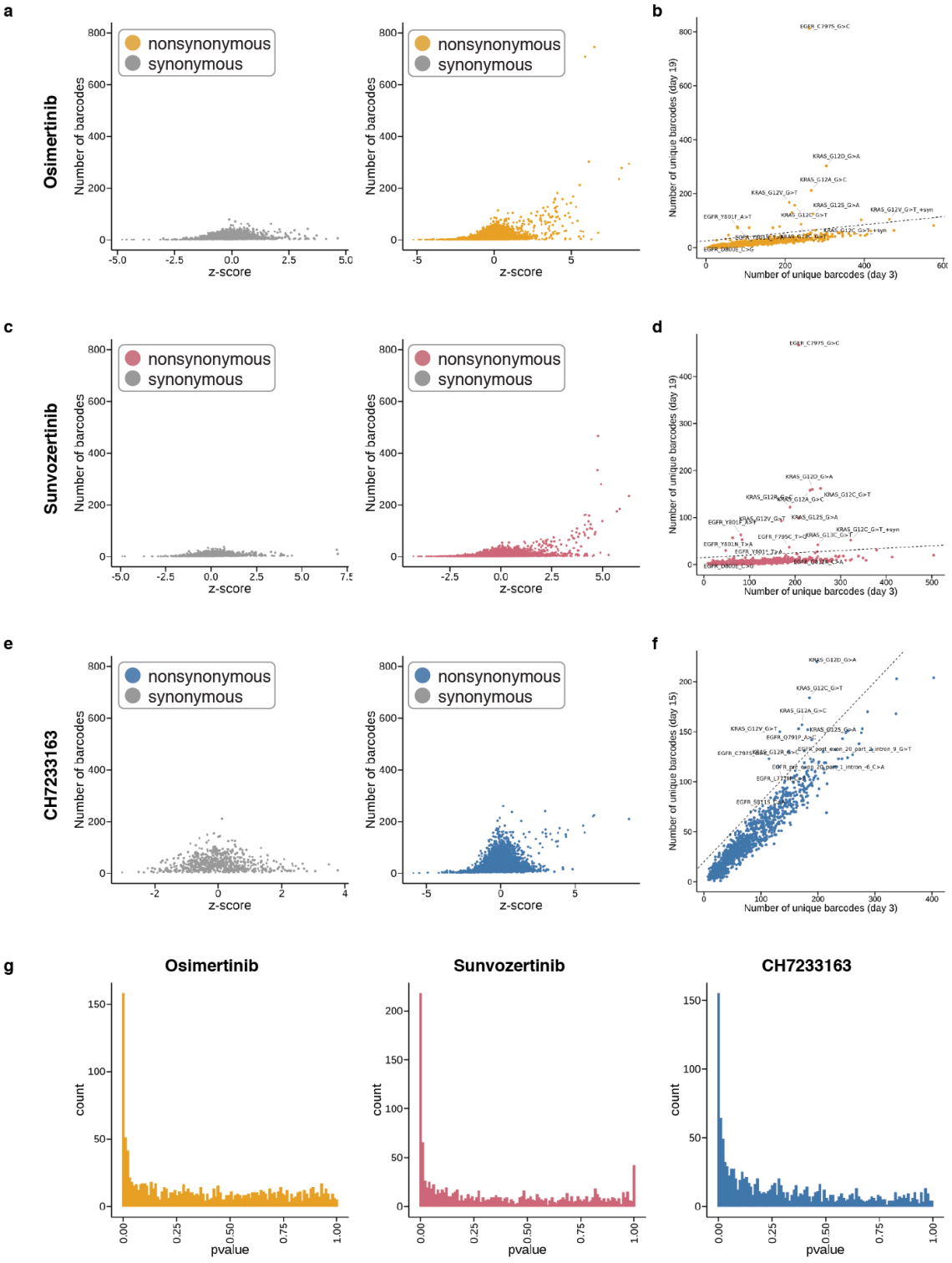

Figure 33. Z-score, barcode count, and p-value statistics from 3,825 epegRNA drug screens.

a) Z-score and barcode counts plotted for day 15 and day 19 data (combined) for all three replicates for the osimertinib screen. Left: synonymous variants, right: nonsynonymous variants. b) Unique

barcode count correlation plot between day 3 and day 19 of the osimertinib screen. Variants that fall above the diagonal (y = 0.15x + 25) are labeled. c) Z-score and barcode counts plotted for day 15 and day 19 data (combined) for all three replicates for the sunvozertinib screen. Left: synonymous variants, right: nonsynonymous variants. d) Unique barcode count correlation plot between day 3 and day 19 of the sunvozertinib screen. Variants that fall above the diagonal (y = 0.05x + 15) are labeled. e) Z-score and barcode counts plotted for day 15 and day 19 data (combined) for all three replicates for the CH7233163 screen. Left: synonymous variants, right: nonsynonymous variants. f) Unique barcode count correlation plot between day 3 and day 15 of the CH7233163 screen. Variants that fall above the diagonal (y = 0.6x + 20) are labeled. g) p-value distributions from DESeq2 differential pegRNA abundance testing for the three drug screens.

2.5. 3,825 EPEGRNA SCREEN IDENTIFIES DRUG RESISTANCE MUTATIONS IN EGFR AND KRAS

There were seven resistant variants that were hits in all three screens, and all seven of these are variants at the 12th and 13th residues of the *KRAS* oncogene (Figure 32e). The *KRAS* gene, which is part of the RAS family of proteins, is the most frequently mutated gene in cancer. Mutations in *KRAS* are a major driver of lung cancers (Huang et al., 2021), particularly the *KRAS* G12C mutation. *KRAS*, until the recent approval of sotorasib (Skoulidis et al., 2021), was considered to be undruggable due to four decades of failed drug discovery attempts to target this oncogene (Huang et al., 2021). Mutations at the 12th and 13th residues in the phosphate-binding loop of *KRAS* are known oncogenic drivers as these mutations directly impair the ability of KRAS to hydrolyze guanosine triphosphate (GTP). This causes the protein to remain in an active GTP-bound state, leading to continued cell growth and the development of cancer. Mutations in *KRAS*

do not cause drug resistance by directly inhibiting a drug from binding its target site, but rather by re-activating oncogenic signaling pathways that lead to constitutive signaling and cell growth. All programmed G12 missense mutations were hits in all three screens ($p<2.27x10^{-7}$, LRT), but only a single G13 (G13C) mutation was a hit in all three screens ($p<1.58x10^{-9}$ in all three screens, LRT). KRAS G13D is a known oncogenic mutation (Hunter et al., 2015), suggesting that the KRAS G13D variant may not have been successfully installed into the genomic DNA via prime editing in this screen.

Another variant which we would expect to give rise to resistant cells in the osimertinib and sunvozertinib screens, but not in the CH7233163 screen, is EGFR C797S, and this is exactly what we observe (Figure 34c). EGFR C797S was a strong hit in the osimertinib and sunvozertinib screens ($p=1.14x10^{-84}$ and $p=2.28x10^{-55}$, respectively; LRT), whereas it was not a hit in the CH7233163 screen ($p=0.34$, LRT) (Figure 34c). CH7233163 overcomes the EGFR L858R/T790M/C797S triple mutation (Kashima et al., 2020), and we observe that EGFR C797S was not a hit when cells are treated with CH7233163, suggesting that EGFR C797S does not confer resistance to cells against this inhibitor. CH7233163 differs from both osimertinib and sunvozertinib in that it is not a covalent binder to the EGFR ATP binding site, but rather, is a noncovalent ATP-competitive inhibitor of EGFR. This difference in binding mechanism could explain the differential resistance profiles we observe in cells treated with this inhibitor, such as the lack of EGFR C797S as an identified resistance mutation. Because EGFR C797S is well-edited in this and previous experiments, and because cells in the different drug arms came from the same starting pool of edited cells, it is unlikely to be a false negative in the CH7233163 treatment arm.

In addition to identifying well-characterized resistance mutations, we also identified numerous missense mutations that confer resistance to TKI treatment that are less well-characterized. Examples of such mutations include EGFR Q791P (Figure 34c) and EGFR Q791L, which were both hits in the CH7233163 screen ($p=4.35x10^{-35}$ and $p=8.44x10^{-13}$, respectively). EGFR Q791 missense mutations are not extensively documented as being drivers of TKI resistance, and documented cases of EGFR Q791 in the context of lung cancer, while present, are sparse (Guibert et al., 2018; C. Li et al., 2021). Mutations in Q791 are predicted to reduce the binding affinity of EGFR to Osimertinib (L. Lin et al., 2020). EGFR Y801 is another example of a residue that when mutated has been identified in single cases of lung cancer (Boldrini et al., 2009), malignant peritoneal mesothelioma (Foster et al., 2010), gastric carcinoma (Z. Liu et al., 2011) and two cases of squamous cell carcinoma (Y. Liu et al., 2013; Na et al., 2007), but it is not known for certain whether a mutation at this residue is a primary driver of resistance to TKIs. EGFR Y801 is a well conserved residue (Foster et al., 2010) and lies within the activation loop of EGFR. EGFR Y801F (Figure 34c, d) and Y801N are hits in both the osimertinib and sunvozertinib screens ($p<5.96x10^{-8}$, Figure 34c) but not in the CH7233163 screen. Although our screening framework is not able to conclusively identify non-resistant variants due to the fact that we cannot be certain that prime edits were made, the fact that these EGFR Y801 variants are resistant in two of the three screens is highly suggestive that EGFR Y801 missense variants are successfully being installed by the prime editing machinery, and that they are likely sensitive to CH7233163. Furthermore, the difference in mechanism of EGFR binding between the two covalent inhibitors (osimertinib and sunvozertinib) and the non-covalent inhibitor (CH7233163) plausibly underlies the difference in resistance mutations we observe between these two classes of inhibitors, such as is the case for EGFR C797S. The ability of our screening method to identify rare resistance mutations makes it

useful for identifying unknown resistance mutations that have not yet been documented in cancer sequencing databases.

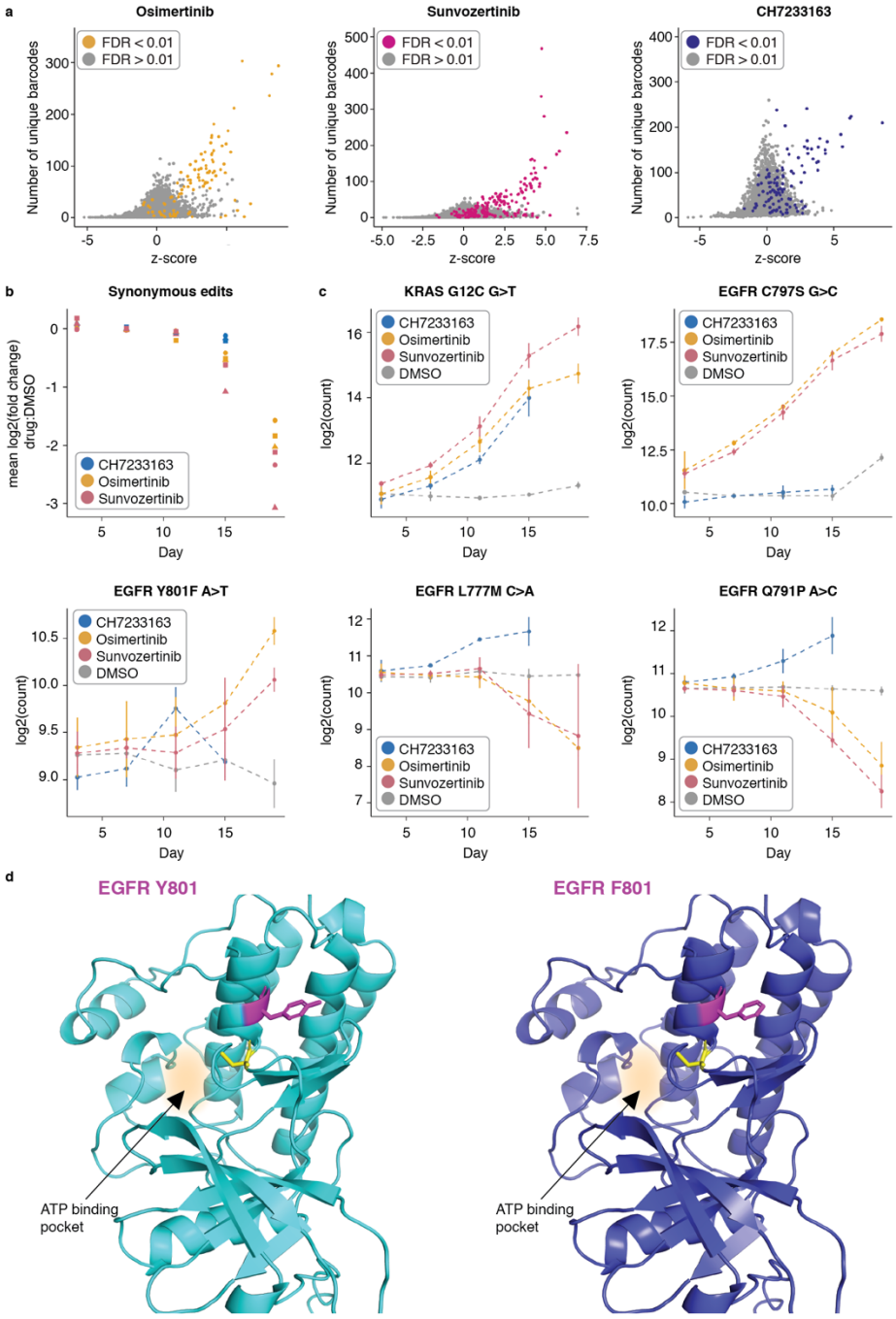

Figure 34. Barcode analysis and individual variant trends across the three scaled drug screens.

a) Relationship between the z-score and the number of unique barcodes recovered for a given epegRNA. Plots shown include data from two timepoints (day 15 and day 19) and three replicates. Each point represents a unique genetic variant encoded by 1-4 epegRNAs at a single timepoint and a single replicate. Points are colored by FDR (Figure 32c). b) Mean of the log2 fold change of synonymous variants in the three drug screens when compared to the DMSO control. Shapes represent three biological replicates. c) Individual variant trends in drug treated cells and DMSO treated cells. Points represent the mean of three replicates, error bars represent one standard deviation from the mean of the three replicates. d) Left: AlphaFold predicted structure of EGFR Y801. Right: AlphaFold predicted structure of EGFR F801. Residue 801 is in magenta, C797 residue is in yellow, and the ATP binding pocket is highlighted.

2.6 BARCODED EPEGRNAS ELUCIDATE CLONALITY OF RESISTANT CELL POPULATIONS AND THEIR GROWTH TRAJECTORIES

Two features of prime-SGE have the potential to yield unique insights into the emergence and growth behavior of resistant cells. The first of these is the inclusion of a pegRNA-specific barcode, that is directly 3' of the epegRNA terminator sequence (Figure 25e). The placement of this barcode does not affect epegRNA binding to its target site, as the barcode is not transcribed, but it allows us to amplify and sequence the barcode identity alongside the epegRNA from genomic DNA. The inclusion of this barcode in the epegRNA design allows us to determine whether resistant cells arose from a single versus multiple editing events (and if multiple, exactly how many). To leverage this feature, we first calculated a z-score for each variant to determine its enrichment in the drug

treatment conditions vs. control. Next, we examined the relationship between z-scores and the number of underlying barcodes (Figure 34a). This analysis shows that many resistant variants (as identified by DESeq2) are characterized by a high number of unique barcodes (Figure 34a, Figure 33a, c, e). This suggests that, for the most part, resistant cells arose from multiple independent editing events that introduced the underlying resistance mutation. As DESeq2 did not leverage these guide-embedded barcodes, this provides orthogonal support for the classification of many of these hits as resistant. Finally, although this warrants further investigation, this result also suggests that the low efficiency of prime editing is highly target specific - *i.e.* some targets edit well, whereas others are rarely or not at all edited.

A second feature of prime-SGE experimental design is that it allows us to observe the trajectory of resistant cells at high resolution via multiple harvests over the 19 day timecourse (Figure 34b, c). Consistent trajectories over the timecourse adds further confidence to hits - if a variant is increasing in abundance over the timecourse of 19 days in a drug treatment condition as compared to a DMSO control, we can be more certain it is a bonafide resistance mutation. This is useful in a prime editing screen, as the presence of a pegRNA does not guarantee the presence of the programmed mutation. Particularly when identifying potential novel drug resistance mutations, such as EGFR Y801F, observing a consistent increase in the frequency of this variant over the timecourse gives further support to the identification of this missense variant as a statistically significant hit in this screen (Figure 34c).

2.7 DISCUSSION

We describe prime-SGE, a multiplex, prime editing-based screening framework to identify drug resistant variants. A key feature of prime-SGE is that it installs genomic edits via prime editing, which uses a single prime editing gRNA to both encode the target site and the programmed edit. The identity of the pegRNA, and its unique barcode, is read out by sequencing amplified pegRNAs from genomic DNA, which allows for increased scaling of this method to many variants in many exons in many genes in a single screen. Furthermore, individually barcoded epegRNAs enable discrimination and tracking of independently originating editing events, something that is not possible with current saturation genome editing methods. In applying this method to several oncogenes and several tyrosine kinase inhibitors in a lung adenocarcinoma cell line, we were able to resolve well-characterized resistance mutations and oncogenic driver mutations, such as EGFR C797S and KRAS G12 missense variants, as well as less well-characterized resistance mutations, such as EGFR Q791 and Y801 missense variants. We also observed differential resistance phenotypes between covalent and non-covalent binders of EGFR, providing a means to directly compare the resistance behavior of programmed variants across different classes of inhibitors. Looking forward, prime-SGE holds the potential to screen increasing numbers of genetic variants, throughout the genome, for resistance to any number of inhibitors.

Although prime-SGE is able to resolve several drug resistance mutations, a major caveat of this screening framework is that it is unable to conclusively identify all drug-sensitive variants. Despite considerable effort put into designing effective pegRNAs (Hsu et al., 2021; Koeppel et al., 2023; Nelson et al., 2022), for any given programmed edit, there remains a high degree of uncertainty that an edit occurred. For example, the KRAS G13 residue is a known oncogenic driver when

mutated. However, in our three drug screens, we only identified a single KRAS G13 missense mutation as a hit (KRAS G13C), despite the fact that five other epegRNAs programming G13 missense mutations were present in the library, one of which is a known oncogenic mutation (KRAS G13D) (Hunter et al., 2015). This suggests that at least the G13D missense variant was not successfully edited into the genomic DNA of cells, rendering our screening framework unable to identify this variant as resistant. This false negative identification of the KRAS G13D mutation surely extends to the other variants we intended to screen, and suggests that the hits we identified are fewer in number than the true number of resistant variants. From screening 1,220 mutations, we identified 17, 31, and 37 drug resistant hits in the osimertinib, sunvozertinib, and CH7233163 screens, respectively. This represents roughly a 2.3% hit rate. It is challenging to determine the false negative rate of this screen, but we hypothesize that we are missing out on the identification of numerous drug resistant variants, as is exemplified by the identification of only one of two expected KRAS G13 missense drug resistant variants.

The prime editing field is advancing at a rapid pace, and we were able to incorporate numerous improvements that were developed over the course of this work. These include knocking out *MLH1* which has been shown to improve efficiency (Chen et al., 2021), utilizing PEmax, a further engineered nCas9-RT, and incorporating an improved enhanced pegRNA structural design that includes a 3' RNA stabilizing motif (Nelson et al., 2022). With further improvements in editing efficiency, we expect the potential of prime-SGE to grow. With continued improvements, it is also plausible to use this screening framework as a way to identify drug-sensitive variants, to assay drug resistance behavior to combination therapies, to envision prime editing-based screens being used for the large-scale identification of loss-of-function mutations, as has been recently

demonstrated (Ren et al., 2023), and other functional screening applications. Looking forward, we also envision prime-SGE being applied in increasingly complex cell types and model systems. Particularly if the efficiency issues are addressed, prime-SGE has the potential to greatly accelerate the functional annotation of the extensive list of coding and non-coding variants that underlie risk for both Mendelian and complex human diseases.

2.8 METHODS

Cell Lines and Culture

PC-9 cell culture

PC-9 cells were originally derived from a metastatic lung adenocarcinoma from a 45 year old male patient. All PC-9 cells were grown at 37°C, and cultured in RPMI 1640 + L-Glutamine (GIBCO, Cat. No. 11-875-093) supplemented with 10% fetal bovine serum (Fisher Scientific, Cat No. SH3039603) and 1% penicillin-streptomycin (Thermo Fisher Scientific, Cat. No. 15070063).

Piggybac-PEmax vector cloning

The PB-EFS-PEmax vector was constructed as follows. The PEmax coding sequence from the T7 promoter to downstream of the bGH-PolyA tail was amplified out of the pCMV-PEmax plasmid (Addgene #174820) with the following primers CGCCAGAACACAGGACCGGTTAATACGACTCACTATAGGGAGAG (forward primer) and AGCGATCGCAGATCCTTCGCTAATGTGAGTTAGCTCACTCATT (reverse primer). An inverse PCR amplification was done of the backbone of the PiggyBac-PE2-Blast plasmid (generated by exchanging the puromycin resistance cassette in the Piggybac-PE2-puro plasmid[52] with a blasticidin resistance cassette) with the following primers: CCCTATAGTGAGTCGTATTAGGTGGCAGCGCTCTAGAACC (forward primer) and GAGTGAGCTAACTCACATTACTTCTGAGGCGGAAAGAACC (reverse primer). These two amplification products were then assembled into a single vector (PB-EFS-PEmax) using NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs, Cat. No. E2621S) using the

standard protocol for a 2-3 fragment assembly. 1 uL of the 20 uL assembly reaction was transformed into 50 uL of stable competent E. coli cells (New England Biolabs, Cat. No. C3040H) using the NEB 5 minute transformation protocol. 100 uL of transformed E. coli cells was plated on an LB agar plate containing ampicillin, and single colonies were picked 1 day later to grow up and extract plasmid DNA using a Monarch Plasmid Miniprep Kit (New England Biolabs, Cat. No. T1010L). Extracted plasmid DNA was sequence confirmed via long-read Nanopore sequencing (Primordium Labs) and DNA from a single clone harboring the correct assembled sequence was used for all experiments.

MLH1 **knockout-PEmax cell line generation and validation**

MLH1 **knockout, selection and single-cell sorting**

MLH1 was knocked out of a population of PC-9 cells using a single gRNA targeting exon 2 of *MLH1*. The sequence of this gRNA is AAGACAATGGCACCGGGATC. This gRNA was cloned into pSpCas9 (BB)-2A-Puro (PX459) V2.0 (Addgene #62988) via the Zhang lab protocol (https://media.addgene.org/data/plasmids/62/62988/62988-attachment_KsK1asO9w4owD8K6wp8.pdf). 2.5 ug of assembled vector was transiently transfected into 250,000 wild type PC-9 cells using the transIT-LT1 transfection reagent (Mirius Bio, Cat. No. MIR 2300). 2 days after transfection, 1 ug/mL concentration of puromycin (GIBCO/Thermo Fisher Scientific, Cat. No. A1113803) was added to cells to select for successfully transfected cells over a period of 4 days.

After puromycin selection, this population of cells was single-cell sorted into 96-well plates to grow up clonal cell lines. 12 clonal lines were expanded, split into two sets of parallel cultures (i.e.

24 wells with 2 wells per clonal line), and one set was treated with 1.5 uM 6-TG (Sigma Aldrich, Cat. No. A4882) for 4 days to screen for cells with *MLH1* successfully knocked out. 5 of 12 treated wells survived 6-TG treatment (denoted clones A, B, C, D, and E). PCR primers targeting *MLH1 (*forward primer: TGTATGAGCCTGTAAGACAAAGGAA, reverse primer: CATCCATATTGAAGCCTTCCTGAAC were used on extracted gDNA from these 5 clonal lines to amplify the *MLH1* locus and confirm knockout via sanger sequencing. A western blot was performed on 3 of the monoclonal lines to confirm knockout (primary antibody used: MLH1 Monoclonal Antibody; Invitrogen, Cat No. MA5-15431, secondary antibody used: Goat anti-Mouse IgG (H+L) Secondary Antibody, HRP; Invitrogen, Cat. No. 31430), and complete loss of the MLH1 protein was confirmed for two clones (A and E) (**Fig. S4**). Clones A and E were chosen for further cell line engineering. Clone E, was ultimately used for all experiments.

Prime Editor-max (PEmax) transfection, selection and single-cell sorting

Clones A and E were transfected with a Piggybac-PEmax plasmid with the Piggybac transposase (System Biosciences, Cat. No. PB210PA-1) at a 10:1 molar ratio using the transIT-LT1 transfection reagent (Mirius Bio, Cat. No. MIR 2300). 2.5 ug of total DNA (Piggybac-PEmax plasmid, and the transposase plasmid), along with 5 uL of trans-IT reagent was reverse-transfected into 100,000 cells for each well of a 12-well plate. Cells were selected with 10 ug/mL of blasticidin for 10 days to select for cells that successfully integrated the PEmax construct. These polyclonal cells were single-cell sorted into 96-well plates to grow up clonal cell lines to generate a MLH1ko-PEMax-PC9 cell line.

Prime editing experiment to assess editing efficiencies of 15 monoclonal lines

15 monoclonal lines were expanded, and prime editing efficiency of 12 of these lines were tested by performing an arrayed experiment in which we tested for the ability of these cells to insert a

trinucleotide sequence (CTT) at the *HEK3* locus when transfected with a pegRNA that programs this insertion. From this experiment, we chose to use Clone A6 for all further prime editing experiments. While Clones A8 and A12 exhibited higher editing efficiency (Fig. S4), these clones grew poorly and were not considered for future experiments.

pegRNA selection and design

All pegRNAs were designed using either the PrimeDesign[49] web or command line interface, using default parameters. Up to four pegRNAs were designed for each individual programmed edit. Input files used for designing pegRNAs are in **Tables S3 and S7**.

pegRNA cloning into transient and lentiviral vectors

EGFR C797S T>A transiently transfected editing experiments

For the EGFR transient C797S T>A editing experiments, the following three pegRNA containing oligos (denoted pegRNA_1, pegRNA_2, and pegRNA_3) were ordered as three separate oPools from Integrated DNA Technologies (IDT):

1) pegRNA_1:

CACCG**ATCACGCAGCTCATGCCCTT**GTTTTAGAGCTAGAAATAGCAAGTTA

AAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGGACCGAGTCGGTCC**TC**

CAGGAGGC*T*GCCGAAGGGCATGAGCTGCTTTT

2) pegRNA_2:

CACC**GTTCCCGGACATAGTCCAGG**GTTTTAGAGCTAGAAATAGCAAGTTAA

AATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGGACCGAGTCGGTCC**TTC**

GGC*A*GCCTCCTGGACTATGTCCGGTTTT

3) pegRNA_3:

CACCG**TGTGTTCCCGGACATAGTCC**GTTTTAGAGCTAGAAATAGCAAGTTA

AAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGGACCGAGTCGGTCC**TT**

CGGC*A*GCCTCCTGGACTATGTCCGGGAATTTT

20 base pair spacer sequences and variable length extension sequences are in bold. Programmed single nucleotide edit are in bold and italics.

The pU6-pegRNA-GG-acceptor (Addgene plasmid #132777) was digested with BsaI-HFv2 (New England Biolabs, Cat. No. R3733S) in 10X rCutSmart Buffer at 37 degrees Celsius overnight to ensure complete digestion of the backbone. This digestion cuts out the mRFP1 cassette (821 base pairs). The linear backbone vector (2,183 base pairs in size) was gel extracted using a gel extraction kit (NEB, Cat. No. T1020S) and assembled with the pegRNA oligos (listed above) via Golden Gate assembly using the following amounts: 30 ng of linearized backbone, 1 uL of 1 uM pegRNA oligo, 0.25 uL of BsaI-HFv2 (New England Biolabs, Cat. No. R3733S), 0.5 uL of T4 DNA ligase (New England Biolabs, Cat. No. M0202S) and 1 uL of 10X T4 DNA ligase reaction buffer (New England Biolabs, Cat. No. B0202S) in a final volume of 10 uL. 1 uL of the assembly reaction was transformed into 50 uL of stable competent E. coli cells (New England Biolabs, Cat. No. C3040H) and plated on an LB agar plate containing ampicillin, and single colonies miniprepped and used for transfection (Zymo Research, Cat. No. D4208T).

EGFR C797S T>A arrayed and 121 epegRNA pooled lentiviral editing screens

For the EGFR lentiviral C797S T>A editing experiments, the following three pegRNA-containing oligos (denoted lenti_epegRNA_1, lenti_epegRNA_2, and lenti_epegRNA _3) were ordered as three separate oPools from Integrated DNA Technologies (IDT):

1) lenti_epegRNA_1:

gctttatatatcttgtggaaaggacgaaacacc**GATCACGCAGCTCATGCCCTT**gttttagagctagaaata gcaagttaaaataaggctagtccgttatcaacttgaaaaagtggGaccgagtcggtCc**TCCAGGAGGC*T*GCCG AAGGGCATGAGCTGC**TTGACGCGGTTCTATCTAGTTACGCGTTAAACCAACT AGAAAttttttttNNNNNNNNggagacgaagcttggcg

2) lenti_epegRNA_2:

gctttatatatcttgtggaaaggacgaaacacc**GGTTCCCGGACATAGTCCAGG**gttttagagctagaaat agcaagttaaaataaggctagtccgttatcaacttgaaaaagtggGaccgagtcggtCc**TTCGGC*A*GCCTCCT GGACTATGTCCGG**TTGACGCGGTTCTATCTAGTTACGCGTTAAACCAACTAG AAAttttttttNNNNNNNNggagacgaagcttggcg

3) lenti_epegRNA_3:

gctttatatatcttgtggaaaggacgaaacacc**GTGTGTTCCCGGACATAGTCC**gttttagagctagaaata gcaagttaaaataaggctagtccgttatcaacttgaaaaagtggGaccgagtcggtCc**TTCGGC*A*GCCTCCTG GACTATGTCCGGGAA**TTGACGCGGTTCTATCTAGTTACGCGTTAAACCAACT AGAAAttttttttNNNNNNNNggagacgaagcttggcg

20 base pair spacer sequences and variable length extension sequences are in bold. Programmed single nucleotide edit are in bold and italics.

The LentiGuide-Puro-P2A-EGFP (Addgene plasmid #137729) was modified to enable cloning of epegRNAs via Gibson assembly. The modified plasmid, which we term Lenti-epeg-Puro-P2A-

EGFP, was used for cloning all epegRNAs. Lenti-epeg-Puro-P2A-EGFP was digested with BsmBI-v2 (New England Biolabs, Cat. No. R0739S) to create a single cut and produce a linearized backbone vector of 9007 base pairs. The three lenti_epegRNA oligos were PCR amplified with the following primers: gctttatatatcttgtggaaaggacg (forward primer) and cgccaagcttcgtctcc (reverse primer) to create double stranded DNA. The double-stranded lenti_epegRNA oligos were then assembled with the linearized Lenti-epeg-Puro-P2A-EGFP backbone using NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs, Cat. No. E2621S) using the standard protocol. 1 uL of the 20 uL assembly reaction was transformed into 50 uL of stable competent E. coli cells (New England Biolabs, Cat. No. C3040H), plated on an LB agar plate containing ampicillin, and single colonies were miniprepped (Zymo Research, Cat. No. D4208T) and used used for lentivirus generation.

1,220 variant lentiviral pooled editing screen

epegRNA-containing oligos were ordered as four separate sub-libraries in one single oligo pool from Twist biosciences. Sequences in this oligo pool are in **Table S2.** The four sub-libraries were cloned separately into the Lenti-epeg-Puro-P2A-EGFP vector, following the steps described above for the 121 epegRNA pool, with the only difference being that after assembly, 2 uL of each library was transformed into 50 uL of electrocompetent E. coli cells (New England Biolabs, Cat. No. C3020K) via electroporation. 990 uL of transformed cells were cultured in 50 mL of LB media + 100 ug/mL ampicillin at 31C. 10 uL (out of a total volume of 1000 uL) of the transformed E. coli cells was plated on LB agar plates containing ampicillin, and colonies were counted the next day to estimate the number of clones in each library. Transformations were performed for each library until each library had a minimum of 1000X coverage of each epegRNA. After overnight culture

at 31C, the epegRNA lentiviral plasmid libraries were extracted using a Qiagen Plasmid Midi Kit (Qiagen, Cat. No. 12143). Extracted plasmid DNA from the four libraries was pooled to generate a pool with 3,825 epegRNAs and used for lentivirus generation.

Lentivirus generation

To generate lentivirus, HEK293T cells (ATCC, Cat. No. CRL-3216) were either plated the day before transfection at 0.7 $x10^6$ cells in a T-25 cell culture flask, or the day of transfection at 1.4 $x10^6$ cells in a T-25 cell culture flask. Virapower lentiviral packaging mix (ThermoFisher Scientific, Cat. No. K497500) was used with Lipofectamine 3000 (ThermoFisher Scientific, Cat. No. L3000001) for transfection into HEK293T cells. 1,500 uL of Opti-MEM reduced serum media (Cat. No. 31985062) was mixed with 42 uL of Lipofectamine 3000 in one tube. 13.5 ug of Virapower lentiviral packaging mix, 4.5 ug of the epegRNA lentiviral plasmid or plasmid library, and 36 uL of P3000 reagent was added to a second tube. The two tubes were combined into a single tube and incubated for 10-20 minutes at room temperature. 50% of the media was removed from the T-25 flask containing the HEK293T cells (unless the cells were plated the day of, then they were only plated in half the amount of media), and the lipid complex from the single tube was added to the cells. The cells were incubated overnight, and media harvests were taken at 24, 48, and 72 hours post-transfection. The lentivirus-containing media was mixed at a 1:4 ratio of PEG-it virus precipitation solution (System Biosciences, Cat. No. LV810A-1) to media, and refrigerated at 4C to concentrate the lentiviral particles. After 96 hours, all three harvests (24, 48, and 72 hours) were spun down at 1,500xg for 30 minutes at 4C. The lentivirus was a visible white pellet after this spin. The media above the pellet was aspirated, and the lentivirus was resuspended in 400 uL

ice cold 1X PBS (Invitrogen, Cat. No. 14190-144), aliquoted into 100 uL aliquots, and frozen at -80 for later use. Each lentiviral aliquot was only thawed a single time prior to use to avoid multiple freeze-thaw cycles. For the larger scale screens, all amounts were increased in scale to make more lentivirus in larger cell culture dishes (T-75 cell culture flasks).

Lentivirus titration experiment in 3,825 epegRNA lentiviral pooled editing screens

A titration experiment was performed to determine lentivirus amounts to achieve the desired multiplicity of infection (MOI) for both screens. 100,000 PC-9 cells were seeded into each well of a 12-well plate in 1 mL of RPMI 1640 + L-Glutamine media (GIBCO, Cat. No. 11-875-093). 0, 0.5, 1, 2, 4, and 8 uL of virus was added to each of 2 wells in the 12-well plate (2 replicates per lentivirus amount). 48 hours after transduction, varying amounts of GFP were observed in the different conditions (GFP is expressed off the epegRNA vector), indicating successful transduction. MOI was determined by flow cytometry (**Fig. S4**). 3 uL of virus for every 100,000 cells was the condition used for the large scale screen to achieve ~30% GFP positivity, which represents an MOI of ~0.35.

Osimertinib dose titration curves

*MLH1*ko-PEmax PC-9 cells expressing either no epegRNA (termed control cells) or an epegRNA coding (lenti_epegRNA_2) for an EGFR C797S mutation (termed $EGFR^{C797S}$ cells) were seeded in duplicate in 12-well dishes at a density of 50,000 cells per well (approximately 14,300 cells/cm^2) in a total volume of 2 mL. Cells were treated with either vehicle (DMSO, Sigma-Aldrich Cat. No.

D2650) at a concentration of 500 nM or osimertinib (AZD9291, SelleckChem Cat. No. S7297) at a concentration of 100, 300, or 500 nM. From this point, cells were passaged every 3 days and replated at a density of 50,000 cells per well in the appropriate concentration of either vehicle or osimertinib. To measure growth rate and cumulative population doublings of control versus $EGFR^{C797S}$ cells in vehicle and osimertinib, cells were counted at each passage using a Vi-CELL XR analyzer (Beckman Coulter). The 300 nM osimertinib dose, which led to an appreciable growth rate difference between control and $EGFR^{C797S}$ cells was used for all further experiments (**Fig. S2**).

pegRNA transient transfection or lentiviral transduction into PC-9 cells

Transient transfections

For the EGFR C797S T>A proof of concept experiment, the SF Cell Line 4D-Nucleofector X Kit S (Lonza, Cat. No. V4XC-2032) was used to transfect the three pegRNAs and one no DNA control. 2.5 ug of DNA (the pegRNA plasmid and the pCMV-PE2 plasmid (Addgene #132775) at a 1:2 ratio by mass of pegRNA plasmid:PE2), was transfected into 100,000 PC-9 cells that went into four wells of a 12-well plate. 2 days later, 400 nM osimertinib was added to the cells for all conditions (no drug, and the three separate pegRNAs). 24 days later, cells were harvested and the EGFR C797 locus was amplified and sequenced.

Lentiviral transductions

A ratio of 3 uL of lentivirus was added for every 100,000 cells transduced. Cell culture amounts were scaled as necessary to transduce cells at 2,000X coverage of epegRNAs. One day after transduction, media was aspirated and replenished with fresh media. Two days after lentiviral

transduction, 2 ug/mL puromycin (GIBCO/Thermo Fisher Scientific, Cat. No. A1113803) was added to the cells for 4 days to select for successfully transduced cells. The cells were replenished with fresh media with 2 ug/mL puromycin each day during these 4 days. For the arrayed *EGFR* C797S T>A experiment, after puromycin selection, we transiently transfected cells with PB-EFS-PEmax twice on two consecutive days. On the second transfection day, cells were treated with 200 nM osimertinib. Eight days later, cells were harvested and the *EGFR* locus was amplified and sequenced. For the 3,825 epegRNA lentiviral pooled editing screens, two parallel plates were cultured without puromycin selection to enable a FACS analysis one week later to determine the MOI of this screen (**Fig. S5**). After 4 days of puromycin selection, 300 nM of drug (for all other lentiviral editing experiments) was added to each treatment condition (CH7233163, osimertinib, or sunvozertinib), and the same volume of DMSO was added to the no treatment control condition. The day that the drug treatments were added marks Day 0 of the screens.

Drug treatments used in screens

Cells were treated with DMSO (SigmaAldrich, Cat. No. D2438) or 100, 300, or 500 nM of CH7233163 (Selleckchem, Cat. No. S9711), osimertinib (Selleckchem, Cat. No. S7297), or sunvozertinib (Selleckchem, Cat. No. E0368), depending on the screen. Drugs were initially diluted to 1 uM in DMSO, and all further dilutions were done in PBS (Invitrogen, Cat. No. 14190-144) to reach the desired concentrations.

Cell harvests and genomic DNA extraction for the 3,825 epegRNA lentiviral pooled editing screens

Cells were harvested on day 8 for the lentiviral EGFR C797S T>A proof of concept experiment. Cells were harvested on days 3, 7, 11, 15, and 19 for both of the large scale screens. Cells were harvested such that a minimum of 500X coverage of epegRNAs was retained in the remaining culture. At each harvest, fresh media and drug (DMSO, CH7233163, osimertinib, or sunvozertinib) was added to the passaged cells. The harvested cells were spun down at 400 x g for 5 minutes, aspirated, and cell pellets were stored at -20C. The Puregene Cell Kit ($8x10^8$) (Qiagen, Cat. No. 158767) was used for all genomic DNA (gDNA) extractions for the large scale screens using the standard protocol, with a single modification to add 2 uL of GlycoBlue coprecipitant (ThermoFisher Scientific, Cat. No. AM9515) to the DNA pellet.

pegRNA amplicon amplification and sequencing

Amplification of the EGFR C797 locus

For the EGFR C797S T>A proof of concept transient and lentiviral editing experiments, the EGFR locus was PCR amplified with the following primers: GCGTCAGATGTGTATAAGAGACAG**CATCTGCCTCACCTCCACCGTG** (forward primer) and GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**ACCAGTTGAGCAGGTACTGGGAGC** (reverse primer). The sequences in bold are the locus-binding part of the primer, and the sequences not in bold contain Nextera and Truseq adapter sequences. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for amplification using the recommended protocol with a 60C annealing temperature and a 30 second extension time. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No.

A63880), and eluted in 20 uL of water. 1 uL of cleaned PCR product was used for a second PCR which adds the Illumina P5 and P7 adapter sequences and sample-specific indices. This second PCR was done using the following primers: AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG (forward primer) and CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT (reverse primer). 10N sequences denote unique indices for each sample. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for amplification using the recommended protocol with a 60C annealing temperature and a 30 second extension time for this second PCR. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No. A63880), and eluted in 20 uL of water. Amplicon concentration and size was determined using the Qubit (ThermoFisher Scientific) and a Tapestation (Agilent) instrument.

Amplification of the integrated genomic DNA pegRNA construct

For the small (121 epegRNAs) and large (3,825 epegRNAs) scale screens, the genomically integrated epegRNAs were amplified via PCR. This is a two-step PCR, with the first PCR amplifying the epegRNA construct and adding partial Illumina read adapter sequences, and with the second PCR adding the Illumina P5 and P7 adapter sequences and sample-specific indices. The first PCR uses the following primers: GCGTCAGATGTGTATAAGAGACAG**cttgtggaaaGGACGAAACACC** (forward primer) and ACGTGTGCTCTTCCGATCT**tctcaagatctagttacgccaagc** (reverse primer). The sequences in bold are the locus-binding part of the primer, and the sequences not in bold contain Nextera and Truseq

adapter sequences. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for the 121 epegRNA small scale screen and KAPA HiFi HotStart ReadyMix (Roche Diagnostics, Cat. No. KK2602) was used for amplification in the large scale screens using the recommended protocol with a 65C annealing temperature for the KAPA2G Robust PCR and the KAPA HiFi PCR, and a 30 second extension time for both protocols. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No. A63880), and eluted in 20 uL of water. 5 uL of cleaned PCR product was used for a second PCR which adds the Illumina P5 and P7 adapter sequences and sample-specific indices. This second PCR was done using the following primers: AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNTCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG (forward primer) and CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGT GCTCTTCCGATCT (reverse primer). 10N sequences denote unique indices for each sample. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for the 121 epegRNA small scale screen and KAPA HiFi HotStart ReadyMix (Roche Diagnostics, Cat. No. KK2602) was used for amplification in the large scale screen using the recommended protocol with a 65C annealing temperature for the KAPA2G Robust PCR and the KAPA HiFi PCR, and a 30 second extension time was used for this second PCR for both protocols. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No. A63880), and eluted in 20 uL of water. Amplicon concentration and size was determined using the Qubit (ThermoFisher Scientific) and a Tapestation (Agilent) instrument.

Amplification of endogenous loci targeted with prime editing gRNAs

For the 121 epegRNA lentiviral pooled experiment, the endogenous locus of each target was PCR amplified and sequenced. Each locus was amplified with locus-specific primers which are listed in **Table S5**. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for amplification using the recommended protocol with a 60C annealing temperature and a 30 second extension time. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No. A63880), and eluted in 20 uL of water. 1 uL of cleaned PCR product was used for a second PCR which adds the Illumina P5 and P7 adapter sequences and sample-specific indices. This second PCR was done using the following primers: AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNTCGTCGGCAGCGTCAGA TGTGTATAAGAGACAG (forward primer) and CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGT GCTCTTCCGATCT (reverse primer). 10N sequences denote unique indices for each sample. KAPA2G Robust HotStart ReadyMix (Roche Diagnostics, Cat. No. KK5702) was used for amplification using the recommended protocol with a 60C annealing temperature and a 30 second extension time for this second PCR. The product of this PCR was cleaned with a 1X AMpure XP bead cleanup (Beckman Coulter, Cat. No. A63880), and eluted in 20 uL of water. Amplicon concentration and size was determined using the Qubit (ThermoFisher Scientific) and a Tapestation (Agilent) instrument.

Sequencing of prime editing screening libraries

Final libraries were sequenced on either a Miseq 300 cycle kit or a NextSeq 2000 P3 200 cycle kit with standard Illumina Nextera and Truseq adapter sequences. The pegRNA spacer was sequenced

on Read 1, and the prime editing gRNA extension sequence was sequenced on Read 2. 10 bp index sequences were sequenced with 10 cycle index 1 and index 2 reads.

Raw data processing and quality control filtering of sequencing data

Fastq file generation, pegRNA read counting, and z-score calculation

Bcl2fastq version 2.20 was used to generate fastq files using default parameters. Fastq files were then input into count_reads.py via a Snakemake pipeline to count the occurrence of each pegRNA in each day, drug treatment, and replicate condition. A pseudocount of 1 was added to all pegRNA read counts, and then read counts were log2 normalized by condition to normalize for variable sequencing depths across samples. This log2 normalized count matrix was used to calculate the log2 fold change of each pegRNA between each treatment condition and the DMSO control treatment condition. Finally, a z-score was calculated for each pegRNA, which is equal to:

$$\text{z-score} = \log_2(\text{fold change}) - \text{mean}(\log_2(\text{fold change})_{\text{controls}}) / \text{standard deviation}(\log_2(\text{fold change})_{\text{controls}})$$

Controls are pegRNAs that encode for synonymous edits.

DESeq2 for differential pegRNA abundance analysis

DESeq2 was used to identify differential pegRNA abundances in the large screen that programmed 1,220 variants with 3,825 epegRNAs in ten oncogenes. The count matrix that was used as an input to DESeq2 was generated from the raw read counts (i.e. not log2 normalized) from the sequencing data. DESeq2 was run for each of the three drug screens separately (osimertinib, sunvozertinib,

and CH7233163), using the likelihood ratio test (LRT) for statistical testing. The full model includes the variables time, drug treatment, and drug:time (the latter interaction term tests for the effect of drug as a result of time), and the reduced model contains only the variable time. The likelihood ratio between the full and reduced model tests if the increased likelihood of the data using the full model is more than expected (i.e. more than 1) and if the parameters in the full model fit the data better than the parameters in the reduced model. DESeq2 was run using default parameters, except for when running the DESeq () function, the parameter minReplicatesForReplace=Inf was used, and in the results () function, the additional parameters cooksCutoff=FALSE and independentFiltering=FALSE were used. These additional parameters reduce the filtering of the data to include outlier data (i.e. data with higher read counts). This data was not filtered out because the high read counts are coming from variants with high resistance phenotypes (i.e. EGFR C797S and KRAS G12 variants). Because DESeq2 is typically used for gene expression analyses, these parameters were appropriately changed for this drug resistance screening data type.

All further downstream analyses were done using custom Python and R scripts, which will soon be accessible on the following website: https://krishna.gs.washington.edu/content/members/multiplex_PE_screening/public/.

Chapter 4. LOOKING FORWARD: HOW WE CAN COMPREHENSIVELY UNDERSTAND GENE REGULATORY ARCHITECTURE AND FUNCTIONAL CONSEQUENCES OF GENETIC VARIATION

In the preceding two chapters, I discuss the development and application of 1) a CRISPRa screening framework by which we can identify *cis*-regulatory sequences that, when perturbed with CRISPRa, lead to upregulation of target genes and 2) a prime editing-based screening framework by which we can identify mutations that render cells resistant to tyrosine kinase inhibition. These two pieces of work are largely methodological and have been applied in their infancy - i.e. both of these methods can and are being applied to further genetic sequences and variants to enable additional biological discovery in these two areas. In this last chapter, I will discuss three areas that I perceive to be critical to render these types of highly scaled, multiplex perturbation strategies as impactful as I truly believe they can be. I highlight multiple angles by which these advances can be achieved, and I am certain that they will be.

4.1 FURTHER IMPROVEMENTS IN CRISPR-BASED GENOME ENGINEERING EFFICIENCY

One of the largest challenges we face in our experimental methods is the suboptimal editing efficiencies or perturbations with CRISPR-based systems. This is exemplified in the case of CRISPR prime editing, where editing efficiencies are often in the single digit percentages. CRISPR prime editing relies on four distinct molecular steps that have to occur in succession for an edit to be incorporated into the genome. The first is the nicking of the top strand of DNA by the Cas9

nuclease, the second is the reverse transcription of the pegRNA extension sequence, the third is the hybridization of the top and bottom DNA strands and the cleavage of the 5' flap to retain the edited DNA, and the last and final step is the ligation of the newly synthesized piece of DNA to the endogenous DNA along with a mismatch repair step that renders the bottom strand of DNA complementary to the edited top strand. The last two of these steps are highly unfavorable for the cell and endogenous DNA repair machinery often inhibits the successful completion of these two steps. To this end, numerous efforts have aimed to address the inefficiency of prime editing including identifying specific mismatch repair proteins that inhibit editing (Chen et al., 2021), and adding a 3' stabilizing motif to the pegRNA structure to prevent degradation of the pegRNA by endogenous exonucleases (Nelson et al., 2022). Both of these improvements are incorporated into the work described in Chapter 3. These insights have already led to substantial improvements in the efficiency of prime editing, and it is reasonable to think that future improvements such as these will continue to increase the efficiency of prime editing. For CRISPR-based perturbation methods, such as CRISPRi and CRISPRa, I envision that additional improvements in effector efficacy, and the ability to specifically tune expression levels via effector choice and gRNA design will enable further fine tuning of these perturbation strategies. This will be particularly important for the application of these methods in *cis*-regulatory therapies as discussed in Chapter 3, where appropriate levels of the expression of genes is critical to rescue the desired phenotype and avoid undesirable side effects of therapy.

4.2 EXTENDING CRISPR-BASED FUNCTIONAL GENOMIC METHODS TO COMPLEX MODEL SYSTEMS

Genomic technologies are often developed and applied in *in vitro* mammalian cell culture systems, and often these cell culture systems are composed of immortalized cancer cells. In Chapters 2 and 3, I describe two such technologies that were developed in an erythroleukemia cell line (K562) and a lung adenocarcinoma cell line (PC-9), respectively. Immortalized cancer cell lines are often used for the development of new genomic technologies primarily for two reasons. The first is that they proliferate and divide indefinitely making them relatively easy to use for cell culture experiments, and the second is that there exists both robust protocols for their use and extensive biochemical characterization for these cell lines due to their extensive use. Despite the advantages of using these cell lines for the development of new genomic technologies, these cell lines do not always accurately reflect the cell types that are most affected by certain diseases that we are interested in studying.

In Chapter 2, I present the development of the CRISPRa screening method and proof-of-concept screen in K562 cells, and then I discuss how we extended this method to iPSC-derived neurons to better understand how these *cis*-regulatory regions regulate genes in a neuronal context. Studying these regions in a neuronal context is critical due to the relevance of the sequences that are being perturbed, which are sequences that regulate the expression of neurodevelopmental disorder risk genes. The results of the iPSC-derived neuron screen highlight the importance of performing these functional genomics screens in relevant cell types, because as was most highlighted by the enhancer targeting results, different cell contexts lead to different results and outcomes. For the CRISPRa screening methodology, the ultimate goal of the method is to nominate gRNAs and

specific genomic loci that can be targeted to upregulate the many genes that are known to cause disease when haploinsufficient. These *cis*-regulatory reagents would then be tested *in vivo* for successful activation of target genes, and ultimately would be candidates for therapeutic development. Extending our CRISPRa screening framework to iSPC-derived neurons represents a step in the right direction for CRISPR-based genomic functional screens. A further extension of this methodology to an even more relevant, and more complex, model system would be even more powerful to more accurately reflect neurological development. Such work has been in the context of brain organoids (Papes et al., 2022), and is a next possible step for this screening framework.

In Chapter 3, I introduce the development of a prime editing screening framework by which I install single nucleotide variants and measure whether these variants proliferate in the presence of tyrosine kinase inhibition. This work was done in PC-9 lung adenocarcinoma cells which were derived from differentiated lung tissue and harbor the activating exon 19 deletion of five amino acids (Glu746-Ala750) in the *EGFR* gene. In this case, the use of this lung adenocarcinoma cell line is a relevant cell type to study these mutations as we are particularly interested in understanding how cancer cells respond to treatment. However, the caveat is that this cell line was derived from a single individual and does not necessarily reflect the genetic landscape of all lung adenocarcinomas, most notably because not all lung adenocarcinomas harbor the activating *EGFR* exon 19 deletion. For this drug resistance screening method, a natural extension of this work would be to perform analogous screens in other cancer cell lines to study these mutations in different cancer types. A further extension into more complex systems would be to study the effects of these mutations in tumors *in vivo*, such as in a mouse model.

4.3 APPLYING KNOWLEDGE GAINED TO THERAPEUTIC STRATEGIES

As mentioned in opening remarks in Chapter 1, I have a particular interest in understanding how genetic alterations cause disease, and further, how the tools and technologies we develop to discover which genetic sequences are playing a causal role in disease initiation and progression can aid in therapeutic discovery and development. We already have a good understanding as to what genetic alterations cause specific diseases. For example, it is known that a single A to T base pair change in the *HBB* gene, which encodes the beta-globin protein, is the cause of sickle-cell anemia. In the case of sickle-cell anemia, CRISPR-based gene therapy approaches are being pursued to correct this mutation and revert the sickle-cell phenotype (S. H. Park & Bao, 2021). However, many diseases do not have a known causal genetic alteration, as is the case for the hundreds of genes that are believed to cause disease when haploinsufficient, such as is the case for numerous neurodevelopmental disorders (Tamura et al., 2022). Some haploinsufficiencies arise from a heterozygous loss of function mutation in the coding region of a gene, however, it is also known that mutations in distal enhancer regions that alter the function of these enhancers can also be the cause of haploinsufficiency.

The CRISPRa screening framework I cover in Chapter 2 describes a framework by which we can identify both proximal and distal regulatory elements (i.e. promoters and enhancers) that can upregulate genes that when haploinsufficient, cause disease. This is a particularly exciting piece of work because there already exists a strategy by which CRISPRa perturbations can be harnessed to revert disease phenotypes. This strategy, *cis*-regulatory therapy, described in Chapter 3, has already shown enormous promise in its ability to revert certain disease phenotypes, such as obesity caused by *Sim1* haploinsufficiency (Matharu et al., 2019), and electrophysiology deficits caused

by *SCN2A* haploinsufficiency (Tamura et al., 2022). The translation of discoveries gained from genomic technologies to actual gene therapies is exciting from a scientific perspective and holds enormous promise and potential. Particularly with the continued development and evolution of precise genome editing methods, it is plausible to think that the number of genetic disorders that can be rescued via a gene therapy approach will only grow in number over the next few decades.

4.4 CLOSING REMARKS

Molecular and cell biology, which is encoded by genomic information, is an intricate and complex process whereby many different factors play a crucial role in orchestrating the many biological pathways that need to function properly and at the right time to render a healthy cell, tissue, and organism. The experiments and tools we as biologists and genomicists develop aim to provide insight into the vast complexity of the cell, and despite our incomplete understanding of molecular biology and genomics, we can learn a lot by perturbing the genome and observing the outcome.

I envision that as a field, genomics will continue to move towards using more complex model systems to bring us closer to studying genomics in its true native context that got us so interested in this field in the first place - and that is the context of living organisms. In addition, I have no doubt that the development of new genomic technologies, including new genome engineering methods, will continue to push our understanding of genes and genomes and will greatly contribute to the larger effort of using genomic knowledge to pursue novel therapeutic strategies that have a genetic cause.

It has been a true honor to be a graduate student in the Department of Genome Sciences and to have two incredible PhD advisors who gave me the freedom to pursue interesting ideas, risky projects, complex experiments, and who allowed me to dive head first into the field of genomic technology development with little knowledge of the field beforehand. I am incredibly proud to have worked on two genomic technologies that I fervently believe provide invaluable understanding to how genes are regulated in neuronal cell types, and which genetic variants lead to drug resistance in the context of lung cancer. Through doing this work, my appreciation and respect for genomics and molecular and cell biology has grown exponentially larger. The human cell and the human body is so intricate, so complex, and at times so difficult to understand, that any discovery and understanding of how the various pieces of this amazing molecular machine work is truly incredible. I hope my efforts described in this dissertation have contributed a clarifying piece to the vast and complex molecular puzzle that is cell biology.

BIBLIOGRAPHY

Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149–157.

Awad, M. M., Liu, S., Rybkin, I. I., Arbour, K. C., Dilly, J., Zhu, V. W., Johnson, M. L., Heist, R. S., Patil, T., Riely, G. J., Jacobson, J. O., Yang, X., Persky, N. S., Root, D. E., Lowder, K. E., Feng, H., Zhang, S. S., Haigis, K. M., Hung, Y. P., … Aguirre, A. J. (2021). Acquired Resistance to KRASG12C Inhibition in Cancer. *The New England Journal of Medicine*, *384*(25), 2382–2393.

Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A., & Jones, S. J. M. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, *7*, 246.

Berger, A. H., Brooks, A. N., Wu, X., Shrestha, Y., Chouinard, C., Piccioni, F., Bagul, M., Kamburov, A., Imielinski, M., Hogstrom, L., Zhu, C., Yang, X., Pantel, S., Sakai, R., Watson, J., Kaplan, N., Campbell, J. D., Singh, S., Root, D. E., … Boehm, J. S. (2017). High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell*, *32*(6), 884.

Bibikova, M., Golic, M., Golic, K. G., & Carroll, D. (2002). Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics*, *161*(3), 1169–1175.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., & Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Research*, *41*(15), 7429–7437.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., & Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, *326*(5959), 1509–1512.

Boldrini, L., Alì, G., Gisfredi, S., Ursino, S., Baldini, E., Melfi, F., Lucchi, M., Comin, C. E., Maddau, C., Tibaldi, C., Camacci, T., Servadio, A., Mussi, A., & Fontanini, G. (2009). Epidermal growth factor receptor and K-RAS mutations in 411 lung adenocarcinoma: a population-based prospective study. *Oncology Reports*, *22*(4), 683–691.

Buckley, M., Kajba, C. M., Forrester, N., Terwagne, C., Sawyer, C., Shepherd, S. T. C., De Jonghe, J., Dace, P., Turajlic, S., & Findlay, G. M. (2023). Saturation Genome Editing Resolves the Functional Spectrum of Pathogenic VHL Alleles. In *bioRxiv* (p. 2023.06.10.542698). https://doi.org/10.1101/2023.06.10.542698

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.

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., Adey, A., Waterston, R. H., Trapnell, C., & Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

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

Chen, P. J., Hussmann, J. A., Yan, J., Knipping, F., Ravisankar, P., Chen, P.-F., Chen, C., Nelson, J. W., Newby, G. A., Sahin, M., Osborn, M. J., Weissman, J. S., Adamson, B., & Liu, D. R. (2021). Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell*, *184*(22), 5635-5652.e29.

Clark, T. A., Chung, J. H., Kennedy, M., Hughes, J. D., Chennagiri, N., Lieber, D. S., Fendler, B., Young, L., Zhao, M., Coyne, M., Breese, V., Young, G., Donahue, A., Pavlick, D., Tsiros, A., Brennan, T., Zhong, S., Mughal, T., Bailey, M., … Lipson, D. (2018). Analytical Validation of a Hybrid Capture–Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA. *The Journal of Molecular Diagnostics: JMD*, *20*(5), 686–702.

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.

Cross, D. A. E., Ashton, S. E., Ghiorghiu, S., Eberlein, C., Nebhan, C. A., Spitzler, P. J., Orme, J. P., Finlay, M. R. V., Ward, R. A., Mellor, M. J., Hughes, G., Rahi, A., Jacobs, V. N., Red Brewer, M., Ichihara, E., Sun, J., Jin, H., Ballard, P., Al-Kadhimi, K., … Pao, W. (2014). AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. *Cancer Discovery*, *4*(9), 1046–1061.

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

Dai, Z., Li, R., Hou, Y., Li, Q., Zhao, K., Li, T., Li, M. J., & Wu, X. (2021). Inducible CRISPRa screen identifies putative enhancers. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, *48*(10), 917–927.

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. In *Nature* (Vol. 485, Issue 7398, pp. 376–380). https://doi.org/10.1038/nature11082

Du, X., Yang, B., An, Q., Assaraf, Y. G., Cao, X., & Xia, J. (2021). Acquired resistance to third-generation EGFR-TKIs and emerging next-generation EGFR inhibitors. *Innovation (Cambridge (Mass.))*, *2*(2), 100103.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* , *21*(16), 3439–3440.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Erwood, S., Bily, T. M. I., Lequyer, J., Yan, J., Gulati, N., Brewer, R. A., Zhou, L., Pelletier, L., Ivakine, E. A., & Cohn, R. D. (2022). Saturation variant interpretation using CRISPR prime editing. *Nature Biotechnology*, *40*(6), 885–895.

Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., & Church, G. M. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods*, *10*(11), 1116–1121.

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., Bruno, D. S., Chang, J. Y., Chirieac, L. R., DeCamp, M., Dilling, T. J., Dowell, J., Durm, G. A., Gettinger, S., Grotz, T. E., Gubens, M. A., Hegde, A., Lackner, R. P., Lanuti, M., … Hughes, M. (2023). NCCN Guidelines® Insights: Non–Small Cell Lung Cancer, Version 2.2023: Featured Updates to the NCCN Guidelines. *Journal of the National Comprehensive Cancer Network: JNCCN*, *21*(4), 340–350.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470.

Fayer, S., Horton, C., Dines, J. N., Rubin, A. F., Richardson, M. E., McGoldrick, K., Hernandez, F., Pesaran, T., Karam, R., Shirts, B. H., Fowler, D. M., & Starita, L. M. (2021). Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *The American Journal of Human Genetics*, *108*(12), 2248–2258.

Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., & Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, *513*(7516), 120–123.

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., Janizek, J. D., Huang, X., Starita, L. M., & Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, *562*(7726), 217–222.

Foster, J. M., Radhakrishna, U., Govindarajan, V., Carreau, J. H., Gatalica, Z., Sharma, P., Nath, S. K., & Loggie, B. W. (2010). Clinical implications of novel activating EGFR mutations in malignant peritoneal mesothelioma. *World Journal of Surgical Oncology*, *8*, 88.

Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature Methods*, *11*(8), 801–807.

Frampton, G. M., Fichtenholtz, A., Otto, G. A., Wang, K., Downing, S. R., He, J., Schnall-Levin, M., White, J., Sanford, E. M., An, P., Sun, J., Juhn, F., Brennan, K., Iwanik, K., Maillet, A., Buell, J., White, E., Zhao, M., Balasubramanian, S., … Yelensky, R. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, *31*(11), 1023–1031.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773.

Fu, J. M., Satterstrom, F. K., Peng, M., Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P., Stevens, C. R., Cusick, C., Babadi, M., Banks, E., Collins, B., Dodge, S., Gabriel, S. B., Gauthier, L., Lee, S. K., … Talkowski, M. E. (2022). Rare coding

variation provides insight into the genetic architecture and phenotypic context of autism. *Nature Genetics*, *54*(9), 1320–1331.

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. *Nature Genetics*, *51*(12), 1664–1669.

Garcia-Murillas, I., Schiavon, G., Weigelt, B., Ng, C., Hrebien, S., Cutts, R. J., Cheang, M., Osin, P., Nerurkar, A., Kozarewa, I., Garrido, J. A., Dowsett, M., Reis-Filho, J. S., Smith, I. E., & Turner, N. C. (2015). Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science Translational Medicine*, *7*(302), 302ra133.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, *176*(1–2), 377-390.e19.

Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews. Genetics*, *21*(5), 292–310.

Gierahn, T. M., Wadsworth, M. H., 2nd, Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., Fortune, S., Love, J. C., & Shalek, A. K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature Methods*, *14*(4), 395–398.

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.

Glickman, M. S., & Sawyers, C. L. (2012). Converting cancer therapies into cures: lessons from infectious diseases. *Cell*, *148*(6), 1089–1098.

Guibert, N., Hu, Y., Feeney, N., Kuang, Y., Plagnol, V., Jones, G., Howarth, K., Beeler, J. F., Paweletz, C. P., & Oxnard, G. R. (2018). Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer. *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO*, *29*(4), 1049–1055.

Hahne, F., & Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor. In *Methods in Molecular Biology* (pp. 335–351). https://doi.org/10.1007/978-1-4939-3578-9_16

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

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29.

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, *2*(3), 666–673.

Horlbeck, M. A., Gilbert, L. A., Villalta, J. E., Adamson, B., Pak, R. A., Chen, Y., Fields, A. P., Park, C. Y., Corn, J. E., Kampmann, M., & Weissman, J. S. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *ELife*, *5*. https://doi.org/10.7554/eLife.19760

Hsu, J. Y., Grünewald, J., Szalay, R., Shih, J., Anzalone, A. V., Lam, K. C., Shen, M. W., Petri, K., Liu, D. R., Joung, J. K., & Pinello, L. (2021). PrimeDesign software for rapid and simplified design of prime editing guide RNAs. *Nature Communications*, *12*(1), 1034.

Huang, L., Guo, Z., Wang, F., & Fu, L. (2021). KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy*, *6*(1), 386.

Hunter, J. C., Manandhar, A., Carrasco, M. A., Gurbani, D., Gondi, S., & Westover, K. D. (2015). Biochemical and Structural Analysis of Common Cancer-Associated KRAS Mutations. *Molecular Cancer Research: MCR*, *13*(9), 1325–1335.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167.

Jänne, P. A., Yang, J. C.-H., Kim, D.-W., Planchard, D., Ohe, Y., Ramalingam, S. S., Ahn, M.-J., Kim, S.-W., Su, W.-C., Horn, L., Haggstrom, D., Felip, E., Kim, J.-H., Frewer, P., Cantarini, M., Brown, K. H., Dickinson, P. A., Ghiorghiu, S., & Ranson, M. (2015). AZD9291 in EGFR Inhibitor–Resistant Non–Small-Cell Lung Cancer. *The New England Journal of Medicine*, *372*(18), 1689–1699.

Jerber, J., Haldane, J., Steer, J., Pearce, D., & Patel, M. (n.d.). *Dissociation of neuronal culture to single cells for scRNA-seq (10x Genomics) v1*. https://doi.org/10.17504/protocols.io.bh32j8qe

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.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502.

Joung, J., Engreitz, J. M., Konermann, S., Abudayyeh, O. O., Verdine, V. K., Aguet, F., Gootenberg, J. S., Sanjana, N. E., Wright, J. B., Fulco, C. P., Tseng, Y.-Y., Yoon, C. H., Boehm, J. S., Lander, E. S., & Zhang, F. (2017). Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*, *548*(7667), 343–346.

Kashima, K., Kawauchi, H., Tanimura, H., Tachibana, Y., Chiba, T., Torizawa, T., & Sakamoto, H. (2020). CH7233163 Overcomes Osimertinib-Resistant EGFR-Del19/T790M/C797S Mutation. *Molecular Cancer Therapeutics*, *19*(11), 2288–2297.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389.

Koeppel, J., Weller, J., Peets, E. M., Pallaseni, A., Kuzmin, I., Raudvere, U., Peterson, H., Liberante, F. G., & Parts, L. (2023). Prediction of prime editing insertion efficiencies using sequence features and DNA repair determinants. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01678-y

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), 420–424.

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. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588.

Konieczkowski, D. J., Johannessen, C. M., & Garraway, L. A. (2018). A Convergence-Based Framework for Cancer Drug Resistance. *Cancer Cell*, *33*(5), 801–815.

Kweon, J., Jang, A.-H., Shin, H. R., See, J.-E., Lee, W., Lee, J. W., Chang, S., Kim, K., & Kim, Y. (2020). A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants. *Oncogene*, *39*(1), 30–35.

Lalanne, J.-B., Regalado, S. G., Domcke, S., Calderon, D., Martin, B., Li, T., Suiter, C. C., Lee, C., Trapnell, C., & Shendure, J. (2022). Multiplex profiling of developmental enhancers with quantitative, single-cell expression reporters. In *bioRxiv* (p. 2022.12.10.519236). https://doi.org/10.1101/2022.12.10.519236

Lander, E. S. (2016). The Heroes of CRISPR. *Cell*, *164*(1–2), 18–28.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, *299*(5607), 682–686.

Li, C., Wang, Y., Su, K., Liu, Y., Wang, L., Zheng, B., Yan, N., Yuan, D., Zhang, Y., Xue, L., Gao, S., & He, J. (2021). Presentation of EGFR mutations in 162 family probands with multiple primary lung cancer. *Translational Lung Cancer Research*, *10*(4), 1734–1746.

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.

Lin, H.-C., He, Z., Ebert, S., Schörnig, M., Santel, M., Nikolova, M. T., Weigert, A., Hevers, W., Kasri, N. N., Taverna, E., Camp, J. G., & Treutlein, B. (2021). NGN2 induces diverse neuron types from human pluripotency. *Stem Cell Reports*, *16*(9), 2118–2127.

Lin, L., Lu, Q., Cao, R., Ou, Q., Ma, Y., Bao, H., Wu, X., Shao, Y., Wang, Z., & Shen, B. (2020). Acquired rare recurrent EGFR mutations as mechanisms of resistance to Osimertinib in lung cancer and in silico structural modelling. *American Journal of Cancer Research*, *10*(11), 4005–4015.

Liu, Y., Wu, B.-Q., Zhong, H.-H., Hui, P., & Fang, W.-G. (2013). Screening for EGFR and KRAS mutations in non-small cell lung carcinomas using DNA extraction by hydrothermal pressure coupled with PCR-based direct sequencing. *International Journal of Clinical and Experimental Pathology*, *6*(9), 1880–1889.

Liu, Z., Liu, L., Li, M., Wang, Z., Feng, L., Zhang, Q., Cheng, S., & Lu, S. (2011). Epidermal growth factor receptor mutation in gastric cancer. *Pathology*, *43*(3), 234–238.

Lizio, M., Abugessaisa, I., Noguchi, S., Kondo, A., Hasegawa, A., Hon, C. C., de Hoon, M., Severin, J., Oki, S., Hayashizaki, Y., Carninci, P., Kasukawa, T., & Kawaji, H. (2019). Update of the FANTOM web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Research*, *47*(D1), D752–D758.

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

Lubeck, E., & Cai, L. (2012). Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods*, *9*(7), 743–748.

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

Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., & Joung, J. K. (2013). CRISPR RNA–guided activation of endogenous human genes. *Nature Methods*, *10*(10), 977–979.

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.

Martin-Rufino, J. D., Castano, N., Pang, M., Grody, E. I., Joubran, S., Caulier, A., Wahlster, L., Li, T., Qiu, X., Riera-Escandell, A. M., Newby, G. A., Al'Khafaji, A., Chaudhary, S., Black, S., Weng, C., Munson, G., Liu, D. R., Wlodarski, M. W., Sims, K., … Sankaran, V. G. (2023). Massively parallel base editing to map variant effects in human hematopoiesis. *Cell*, *186*(11), 2456-2474.e24.

Matharu, N., & Ahituv, N. (2020). Modulating gene regulation to treat genetic disorders. *Nature Reviews. Drug Discovery*, *19*(11), 757–775.

Matharu, N., Rattanasopha, S., Tamura, S., Maliskova, L., Wang, Y., Bernard, A., Hardin, A., Eckalbar, W. L., Vaisse, C., & Ahituv, N. (2019). CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science*, *363*(6424). https://doi.org/10.1126/science.aau0629

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(2), 560–564.

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

Meitlis, I., Allenspach, E. J., Bauman, B. M., Phan, I. Q., Dabbah, G., Schmitt, E. G., Camp, N. D., Torgerson, T. R., Nickerson, D. A., Bamshad, M. J., Hagin, D., Luthers, C. R., Stinson, J. R., Gray, J., Lundgren, I., Church, J. A., Butte, M. J., Jordan, M. B., Aceves, S. S., … James, R. G. (2020). Multiplexed Functional Assessment of Genetic Variants in CARD11. *American Journal of Human Genetics*, *107*(6), 1029–1043.

Mitra, R. D., & Church, G. M. (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research*, *27*(24), e34.

Mitra, Robi D., Shendure, J., Olejnik, J., Edyta-Krzymanska-Olejnik, & Church, G. M. (2003). Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, *320*(1), 55–65.

Na, I. I., Kang, H. J., Cho, S. Y., Koh, J. S., Lee, J. K., Lee, B. C., Lee, G. H., Lee, Y. S., Yoo, H. J., Ryoo, B.-Y., Yang, S. H., & Shim, Y. S. (2007). EGFR mutations and human

papillomavirus in squamous cell carcinoma of tongue and tonsil. *European Journal of Cancer*, *43*(3), 520–526.

Nelson, J. W., Randolph, P. B., Shen, S. P., Everette, K. A., Chen, P. J., Anzalone, A. V., An, M., Newby, G. A., Chen, J. C., Hsu, A., & Liu, D. R. (2022). Engineered pegRNAs improve prime editing efficiency. *Nature Biotechnology*, *40*(3), 402–410.

Overman, M. J., Modak, J., Kopetz, S., Murthy, R., Yao, J. C., Hicks, M. E., Abbruzzese, J. L., & Tam, A. L. (2013). Use of research biopsies in clinical trials: are risks and benefits adequately discussed? *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, *31*(1), 17–22.

Pagès, H. (n.d.). BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. *R Package Version*.

Panda, S. K., Wigerblad, G., Jiang, L., Jiménez-Andrade, Y., Iyer, V. S., Shen, Y., Boddul, S. V., Guerreiro-Cacais, A. O., Raposo, B., Kasza, Z., & Wermeling, F. (2020). IL-4 controls activated neutrophil FcγR2b expression and migration into inflamed joints. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(6), 3103–3113.

Papes, F., Camargo, A. P., de Souza, J. S., Carvalho, V. M. A., Szeto, R. A., LaMontagne, E., Teixeira, J. R., Avansini, S. H., Sánchez-Sánchez, S. M., Nakahara, T. S., Santo, C. N., Wu, W., Yao, H., Araújo, B. M. P., Velho, P. E. N. F., Haddad, G. G., & Muotri, A. R. (2022). Transcription Factor 4 loss-of-function is associated with deficits in progenitor proliferation and cortical neuron content. *Nature Communications*, *13*(1), 2387.

Park, M.-Y., Jung, M. H., Eo, E. Y., Kim, S., Lee, S. H., Lee, Y. J., Park, J. S., Cho, Y. J., Chung, J. H., Kim, C. H., Yoon, H. I., Lee, J. H., & Lee, C.-T. (2017). Generation of lung cancer

cell lines harboring EGFR T790M mutation by CRISPR/Cas9-mediated genome editing. *Oncotarget*, *8*(22), 36331–36338.

Park, S. H., & Bao, G. (2021). CRISPR/Cas9 gene editing for curing sickle cell disease. *Transfusion and Apheresis Science: Official Journal of the World Apheresis Association: Official Journal of the European Society for Haemapheresis*, *60*(1), 103060.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, *27*(12), 1173–1175.

Persky, N. S., Hernandez, D., Do Carmo, M., Brenan, L., Cohen, O., Kitajima, S., Nayar, U., Walker, A., Pantel, S., Lee, Y., Cordova, J., Sathappa, M., Zhu, C., Hayes, T. K., Ram, P., Pancholi, P., Mikkelsen, T. S., Barbie, D. A., Yang, X., … Johannessen, C. M. (2020). Defining the landscape of ATP-competitive inhibitor resistance residues in protein kinases. *Nature Structural & Molecular Biology*, *27*(1), 92–104.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *152*(5), 1173–1183.

Radford, E. J., Tan, H. K., Andersson, M. H. L., Stephenson, J. D., Gardner, E. J., Ironfield, H., Waters, A. J., Gitterman, D., Lindsay, S., Abascal, F., Martincorena, I., Kolesnik, A., Ng-Cordell, E., Firth, H. V., Baker, K., Perry, J. R. B., Adams, D. J., Gerety, S. S., & Hurles, M. E. (2022). Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. In *bioRxiv*. https://doi.org/10.1101/2022.06.10.22276179

Ramirez, M., Rajaram, S., Steininger, R. J., Osipchuk, D., Roth, M. A., Morinishi, L. S., Evans, L., Ji, W., Hsu, C.-H., Thurley, K., Wei, S., Zhou, A., Koduru, P. R., Posner, B. A., Wu,

L. F., & Altschuler, S. J. (2016). Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells. *Nature Communications*, *7*, 10690.

Reita, D., Pabst, L., Pencreach, E., Guérin, E., Dano, L., Rimelen, V., Voegeli, A.-C., Vallat, L., Mascaux, C., & Beau-Faller, M. (2021). Molecular Mechanism of EGFR-TKI Resistance in EGFR-Mutated Non-Small Cell Lung Cancer: Application to Biological Diagnostic and Monitoring. *Cancers*, *13*(19). https://doi.org/10.3390/cancers13194926

Ren, X., Yang, H., Nierenberg, J. L., Sun, Y., Chen, J., Beaman, C., Pham, T., Nobuhara, M., Takagi, M. A., Narayan, V., Li, Y., Ziv, E., & Shen, Y. (2023). High throughput PRIME editing screens identify functional DNA variants in the human genome. In *bioRxiv* (p. 2023.07.12.548736). https://doi.org/10.1101/2023.07.12.548736

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

Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., Wong, M. C., Maddren, M., Fang, R., Heitner, S. G., Lee, B. T., Barber, G. P., Harte, R. A., Diekhans, M., Long, J. C., Wilder, S. P., Zweig, A. S., Karolchik, D., Kuhn, R. M., … Kent, W. J. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research*, *41*(Database issue), D56-63.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

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

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568-584.e23.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470.

Schmidt, R., Steinhart, Z., Layeghi, M., Freimer, J. W., Bueno, R., Nguyen, V. Q., Blaeschke, F., Ye, C. J., & Marson, A. (2022). CRISPR activation and interference screens decode stimulation responses in primary human T cells. *Science*, *375*(6580), eabj4008.

Shendure, J., & Fields, S. (2016). Massively Parallel Genetics. *Genetics*, *203*(2), 617–619.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., & Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728–1732.

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.

Skoulidis, F., Li, B. T., Dy, G. K., Price, T. J., Falchook, G. S., Wolf, J., Italiano, A., Schuler, M., Borghaei, H., Barlesi, F., Kato, T., Curioni-Fontecedro, A., Sacher, A., Spira, A., Ramalingam, S. S., Takahashi, T., Besse, B., Anderson, A., Ang, A., … Govindan, R. (2021). Sotorasib for Lung Cancers with KRAS p.G12C Mutation. *The New England Journal of Medicine*, *384*(25), 2371–2381.

Song, M., Yang, X., Ren, X., Maliskova, L., Li, B., Jones, I. R., Wang, C., Jacob, F., Wu, K., Traglia, M., Tam, T. W., Jamieson, K., Lu, S.-Y., Ming, G.-L., Li, Y., Yao, J., Weiss, L. A., Dixon, J. R., Judge, L. M., … Shen, Y. (2019). Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nature Genetics*, *51*(8), 1252–1262.

Srivatsan, S. R., Regier, M. C., Barkan, E., Franks, J. M., Packer, J. S., Grosjean, P., Duran, M., Saxton, S., Ladd, J. J., Spielmann, M., Lois, C., Lampe, P. D., Shendure, J., Stevens, K. R., & Trapnell, C. (2021). Embryo-scale, single-cell spatial transcriptomics. *Science*, *373*(6550), 111–117.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78–82.

Tak, Y. E., Horng, J. E., Perry, N. T., Schultz, H. T., Iyer, S., Yao, Q., Zou, L. S., Aryee, M. J., Pinello, L., & Joung, J. K. (2021). Augmenting and directing long-range CRISPR-mediated activation in human cells. *Nature Methods*, *18*(9), 1075–1081.

Tamura, S., Nelson, A. D., Spratt, P. W. E., Kyoung, H., Zhou, X., Li, Z., Zhao, J., Holden, S. S., Sahagun, A., Keeshen, C. M., Lu, C., Hamada, E. C., Ben-Shalom, R., Pan, J. Q., Paz, J. T., Sanders, S. J., Matharu, N., Ahituv, N., & Bender, K. J. (2022). CRISPR activation rescues abnormalities in SCN2A haploinsufficiency-associated autism spectrum disorder. In *bioRxiv* (p. 2022.03.30.486483). https://doi.org/10.1101/2022.03.30.486483

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382.

Thress, K. S., Paweletz, C. P., Felip, E., Cho, B. C., Stetson, D., Dougherty, B., Lai, Z., Markovets, A., Vivancos, A., Kuang, Y., Ercan, D., Matthews, S. E., Cantarini, M., Barrett, J. C., Jänne, P. A., & Oxnard, G. R. (2015). Acquired EGFR C797S mutation mediates resistance to AZD9291 in non–small cell lung cancer harboring EGFR T790M. *Nature Medicine*, *21*(6), 560–562.

Tian, R., Abarientos, A., Hong, J., Hashemi, S. H., Yan, R., Dräger, N., Leng, K., Nalls, M. A., Singleton, A. B., Xu, K., Faghri, F., & Kampmann, M. (2021). Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, *24*(7), 1020–1034.

Ursu, O., Neal, J. T., Shea, E., Thakore, P. I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M. M., Fagre, C., Lo, A., McSharry, M., Giacomelli, A. O., Ly, S. H., Rozenblatt-Rosen, O., Hahn, W. C., Aguirre, A. J., Berger, A. H., … Boehm, J. S. (2022). Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nature Biotechnology*, *40*(6), 896–905.

Vasan, N., Baselga, J., & Hyman, D. M. (2019). A view on drug resistance in cancer. *Nature*, *575*(7782), 299–309.

Wagenaar, T. R., Ma, L., Roscoe, B., Park, S. M., Bolon, D. N., & Green, M. R. (2014). Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell & Melanoma Research*, *27*(1), 124–133.

Wang, C., Ward, M. E., Chen, R., Liu, K., Tracy, T. E., Chen, X., Xie, M., Sohn, P. D., Ludwig, C., Meyer-Franke, A., Karch, C. M., Ding, S., & Gan, L. (2017). Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening. *Stem Cell Reports*, *9*(4), 1221–1233.

Wang, M., Yang, J. C.-H., Mitchell, P. L., Fang, J., Camidge, D. R., Nian, W., Chiu, C.-H., Zhou, J., Zhao, Y., Su, W.-C., Yang, T.-Y., Zhu, V. W., Millward, M., Fan, Y., Huang, W.-T., Cheng, Y., Jiang, L., Brungs, D., Bazhenova, L., … Jänne, P. A. (2022). Sunvozertinib, a Selective EGFR Inhibitor for Previously Treated Non-Small Cell Lung Cancer with EGFR Exon 20 Insertion Mutations. *Cancer Discovery*, *12*(7), 1676–1689.

Waterston, R., & Sulston, J. (1995). The genome of Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(24), 10836–10840.

Xie, S., Duan, J., Li, B., Zhou, P., & Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, *66*(2), 285-299.e5.

Xu, Z., Sziraki, A., Lee, J., Zhou, W., & Cao, J. (2023). PerturbSci-Kinetics: Dissecting key regulators of transcriptome kinetics through scalable single-cell RNA profiling of pooled CRISPR screens. In *bioRxiv* (p. 2023.01.29.526143). https://doi.org/10.1101/2023.01.29.526143

Yao, D., Tycko, J., Oh, J. W., Bounds, L. R., Gosai, S. J., Lataniotis, L., Mackay-Smith, A., Doughty, B. R., Gabdank, I., Schmidt, H., Youngworth, I., Andreeva, K., Ren, X., Barrera, A., Luo, Y., Siklenka, K., Yardımcı, G. G., The ENCODE4 Consortium, Tewhey, R., … Reilly, S. K. (2022). Multi-center integrated analysis of non-coding CRISPR screens. In *bioRxiv* (p. 2022.12.21.520137). https://doi.org/10.1101/2022.12.21.520137

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S. M., Hellmann, M. D., Barron, D. A., Schram, A. M., Hameed, M., Dogan, S., Ross, D. S., Hechtman, J. F., DeLair, D. F., Yao, J., … Berger, M. F. (2017). Erratum: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature Medicine*, *23*(8), 1004.

Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., Xu, W., Yang, N., Danko, T., Chen, L., Wernig, M., & Südhof, T. C. (2013). Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron*, *78*(5), 785–798.

Zhou, B., Ho, S. S., Greer, S. U., Zhu, X., Bell, J. M., Arthur, J. G., Spies, N., Zhang, X., Byeon, S., Pattni, R., Ben-Efraim, N., Haney, M. S., Haraksingh, R. R., Song, G., Ji, H. P., Perrin, D., Wong, W. H., Abyzov, A., & Urban, A. E. (2019). Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Research*, *29*(3), 472–484.

VITA

Florence Marie Chardon was born in Versailles, France. A few months after her birth, her family moved to Wolfratshausen, a small town in the Bavarian province of Germany. Her family then moved to Santa Barbara, California when she was five, and that is where Florence underwent her primary and secondary education. She received her undergraduate Bachelor's in Science degree in Chemistry from the University of California, Berkeley. She then worked as a research associate for three years at Genentech, Inc. in South San Francisco before starting graduate school at the University of Washington in the Fall of 2017.