# High-throughput interrogation of genome function and cellular lineage

Gregory M. Findlay

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Jay Shendure, Chair

Robert Waterston

Stephen Tapscott

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

## Abstract

High-throughput interrogation of genome function and cellular lineage

Gregory M. Findlay

Chair of the Supervisory Committee:

Professor Jay Shendure, MD, PhD

Genome Sciences

Mutations can reveal how biological functions are encoded in our DNA, and how biological specimens relate to one another. In nature, mutations occur infrequently and are subject to natural selection. Therefore, to better learn how the DNA sequences within genomes function, methods to deliberately create mutations and study their effects have been developed and employed broadly. Recently engineered genome editing technologies constitute a means of inducing mutations at a high frequency and in a targeted fashion, allowing researchers to effectively rewrite the DNA code of a living cell's genome. One such technology called CRISPR/Cas9, has enabled genome editing at unprecedented ease and scale.

Here, I describe implementations of CRISPR/Cas9 genome editing to generate high allelic diversity at targeted loci. Experimental quantification of genome editing outcomes via next-generation sequencing is used to investigate two basic biological questions: 1.) How mutations impact the function of genomic sequences, both coding and regulatory, and 2.) How cells in the body relate to one another by way of a developmental lineage.

We investigated how mutations impact the function of DNA in two ways. First, we established and optimized a CRISPR/Cas9-mediated method to introduce all possible single nucleotide variants over a genomic region to determine the effects of each one in parallel. We employ this method, called 'saturation genome editing', to investigate thousands of variants in *BRCA1*, a gene in which loss-of-function variants cause hereditary breast and ovarian cancer predisposition. The high accuracy of the data suggests this will be a powerful method for interpreting variants encountered clinically. Second, to probe vast expanses of genomic sequence for functional effects on gene regulation, we devised a method to introduce and assay thousands of large deletions in a high throughput manner. For one gene, *HPRT1*, we use this method to show that distal regulatory elements are unlikely to be required for the gene's expression. We anticipate these two methods will be powerful and complementary tools for identifying critical regions of the genome and dissecting how they function.

Towards understanding how an entire organism develops from a single fertilized egg, we developed an approach to record relationships between individual cells. We use CRISPR/Cas9 to create diverse mutations in a short DNA barcode present within each cell of a growing organism, such that the ancestral relationship between two cells can be determined by how similar the cells' barcodes are to one another. Determining the barcode sequences of hundreds of thousands of cells sampled from grown organisms allows us to construct lineage trees that reveal how sequential cell

divisions give rise first to embryonic germ layers and then to the cell types, tissues and organs of fully formed organisms. Future use of this method, which we call ‘GESTALT’, will elucidate cell lineage in multicellular systems for normal development and disease.

Potential improvements and applications of these methods are described in a concluding section.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

Many people were incredibly helpful, supportive, generous and scientifically indispensable during my time pursuing this work. I first thank my advisor, Jay Shendure for honest mentorship, for practical guidance, for focusing on the big questions, for allowing freedom to explore, for providing the funding to do so, for serving as a role model, for thousands of his insights, and for always, always remaining excited about the work. I will remember certain high fives between us a lot more fondly than the clockwork nature of 4am paper writing hand-offs, but I am truly thankful for all of the energy and excitement. Working with Jay is easily the best single decision I have made as a scientist to date, and I am forever grateful that I have been a part of all the fun research in his lab.

I also thank my thesis committee of Bob Waterston, Stephen Tapscott, Cecilia Moens, and Alice Berger for asking questions that leave me thinking for a long time after we meet, and for always being accommodating and supportive of my academic goals.

I owe an enormous amount of gratitude to the scientists with whom I have worked most closely for making the science both exciting and enjoyable. These people include Aaron McKenna, Riza Daza, Molly Gasperini, and Evan Boyle. I also thank the many lab members who have been both world-class scientists as well as excellent friends during my time in the lab. These include Matthew Snyder, Molly Gasperini, Aaron McKenna, and Martin Kircher. All have taught me so much about what we study together, but more about life in general. I am also grateful for the people who volunteered to work on projects with me. These relationships made lab work both more manageable and more enjoyable. Particular thanks to Riza Daza, Rocío Acuña-Hidalgo, Jen Milbank, and Melissa Zhang. I also thank the entire lab for making it a great scientific environment

to work in, especially Akash Kumar, Malte Spielman and Lea Starita. I learned a tremendous amount from those in the lab before me, particularly Rupali Patwardhan and Jacob Kitzman who served as fantastic mentors. I also thank Charlie Lee, Melissa Gillies, Riza Daza and Beth Martin for their enormous efforts to keep the lab running.

I'm grateful for my previous scientific mentors who left a lasting impression on me, most notably Stephan Zweifel at Carleton and Anjana Rao at Harvard.

I also thank all of my friends in science for the many wonderful conversations and shared willingness to talk about work outside of work (and for demanding at times that we not talk about work). In particular, I thank John Lazar, Max Dougherty, Heather Machkovech, and Andrew Bogaard for providing many valuable insights and enjoyable discussions.

Lastly, I thank my family. My brother Geoff has been a tremendous scientific role model. I am truly wowed by his dedication to his own lab and to his students and it serves as an inspiration. I cannot thank my parents enough for being incredibly supportive throughout graduate school, and for long before then. It's a privilege to be a scientist and I am so grateful for their support.

# DEDICATION

This work is dedicated to my mother, Linda, my father John, and my grandparents Sue, Jack, Jean, and Lou in honor of their unwavering support of my education.

# Chapter 1. INTRODUCTION

Despite a rapidly expanding wealth of genetic data, there are fundamental limits on what we can learn through human genetics. Perhaps most importantly, there is insufficient naturally occurring genetic variation to fully close large gaps in our knowledge about which variants contribute to disease risk, and more fundamentally, exactly how the genome programs biological function on mechanistic scales spanning from single loci to entire organisms. CRISPR-based technologies have the potential to enable vast gains towards addressing these challenges, by empowering the systematic querying of genome sequence via targeted mutation. Current genome editing methods reliably enable the precise introduction of variants ranging from single nucleotide changes to megabase-scale alterations. At the same time, powerful multiplex approaches fueled by massively parallel sequencing are transforming our ability to quantify the functional consequences of mutations. In this introductory chapter, I provide a brief review of existing genome editing methods and describe how they have been used to date to improve our mechanistic understanding of how sequence dictates both coding and regulatory function.

## 1.1 EDIT THE GENOME TO REVEAL HOW IT FUNCTIONS

The central goal of human genetics is to understand how DNA sequence variation confers phenotypic consequences on the molecular, cellular and organismal level. Since the completion of the human genome project (International Human Genome Sequencing Consortium, 2001), technological advances such as microarrays and next-generation sequencing have spurred the ascertainment of detailed genetic information from millions of individuals (Shendure et al., 2017). Linking genotypes to phenotypes through methods such as genome-wide association studies (GWAS) has implicated specific variants, genes, and cellular pathways in disease, leading in turn

to major advances in healthcare (Bush and Moore, 2012). Powered by deep sampling of populations, small contributions from up to hundreds of loci can be detected for some traits (Boyle et al., 2017). Concurrently, technologies for transcriptomic and epigenomic profiling have been deployed to richly annotate genomic sequences with biochemical data measured from diverse cell and tissue types (ENCODE Project Consortium, 2012; The GTEx Consortium, 2015). Coupling the sampling of variation with such 'omics' readouts associates specific variants with molecular changes (*e.g.* expression quantitative trait loci) (Nica and Dermitzakis, 2013).

Despite the success of these approaches that leverage naturally occurring variation, immense challenges remain to delineating functional sequences across the genome — and relatedly, to predicting functional consequences of genetic variants (Shendure, 2014). For most genotype-phenotype associations, the precise mechanism of action remains unknown. Indeed, most often the causal variant(s) responsible are unresolved due to linkage (Freedman et al., 2011). Even if a single, relatively common variant can be identified, pinpointing the cell type(s), developmental stage(s), and pathological state(s) in which the variant's effects are manifested often proves a daunting task.

Rare variation poses an even greater challenge. The hundreds of thousands of whole exomes and whole genomes sequenced in the last decade have only begun to catalogue the rare variation present in our species (Lek et al., 2016; Shendure and Akey, 2015). Low allele frequencies preclude confident assessment of variant effects through phenotypic association alone, as evidenced by the tens of thousands of variants of uncertain significance (VUS) encountered in clinically sequenced genes (Landrum et al., 2016; Starita et al., 2017). Although population-scale sequencing efforts yield valuable insights (MacArthur et al., 2012), scalable perturbational methods to probe genome function are clearly needed (Gasperini et al., 2016).

Since their discovery as programmable endonucleases, bacterial CRISPR systems have been engineered to produce tools for probing genome function at scale (Cong et al., 2013; Doudna and Charpentier, 2014; Jinek et al., 2012; Mali et al., 2013). Molecular cloning (*e.g.* via plasmids) and methods such as RNA interference have been crucial tools for studying gene function, but direct perturbation of the genome itself has numerous advantages (Gibson et al., 2013). Preserving the surrounding sequence context and epigenetic status of a given element promises to yield the most accurate functional measurements. Indeed, critical questions to genetics today, such as how elements regulate gene expression from afar, are difficult to probe outside of the genome. The poor concordance observed when the same elements were tested for enhancer activity both episomally and genomically is an important cautionary example (Inoue et al., 2017). Put simply, to ascertain the rules by which the genome functions, perturbing the genome itself has unique value.

Here, we briefly review multiplex CRISPR-based methods to interrogate genome function and describe how these methods may be built upon to facilitate systematic exploration of both coding and regulatory functions of the human genome. We anticipate that CRISPR/Cas9-driven approaches to test DNA sequence function will greatly accelerate our understanding of the mechanisms underlying molecular, cellular, and organismal consequences of genetic variation.

## 1.2 CRISPR TECHNOLOGIES ALLOW GENOME EDITING AT UNPRECEDENTED SCALE

An ideal system for genome editing would allow the rapid conversion of one or many genomic sequences to one or many specified alternatives (*i.e.* 'edits'). Edits would be introduced with 100% efficiency at targeted loci but not elsewhere. The method would be cheap and scalable such that every nucleotide of the genome could be edited. Aspects of CRISPR/Cas9 genome editing systems resemble such an ideal system, but further methodological developments will prove valuable towards achieving these goals.

Numerous CRISPR-derived tools for genome editing have been engineered from different bacterial Cas protein orthologues. These include *S. pyogenes* Cas9 (SpCas9) (Cong et al., 2013; Jinek et al., 2012; Mali et al., 2013) and *S. aureus* Cas9 (Ran et al., 2015), as well as, Cpf1 proteins from multiple species (Zetsche et al., 2015). Modified versions of these proteins and their corresponding guide RNA components have been engineered for enhanced functionality, such as reduced DNA cleavage (Qi et al., 2013; Ran et al., 2013), higher target fidelity (Fu et al., 2014b; Kleinstiver et al., 2016a; Slaymaker et al., 2016), and altered specificity (Kleinstiver et al., 2015a, 2015b). The net effect of these advances is that most of the non-repetitive human genome is targetable within less than 10 base pairs (bp) by at least one, highly specific RNA-guided endonuclease (Canver et al., 2017). The rapid development also suggests these technologies will continue to improve, and researchers will be able to choose which editing enzyme is best for their experiments. The wide popularity of SpCas9 reflects that it is sufficient for many applications. The simple and well-defined rules for SpCas9 targeted cleavage, the wealth of off-target data and prediction tools for SpCas9 (Doench et al., 2014, 2016; McKenna and Shendure, 2017; Tsai et al., 2015), and the consistent performance of many existent SpCas9-containing constructs across a wide variety of systems have all contributed to the tool's popularity (Platt et al., 2014; Ran et al., 2013; Sanjana et al., 2014). Put another way, the 'infrastructure' for SpCas9 editing is already in place. However, other CRISPR enzymes have also been characterized to varying extents, and knowledge of how to best use them will keep improving as methods first used to characterize SpCas9 are deployed (Kleinstiver et al., 2016b). Rather than reviewing differences between enzymes, here we instead assume we can effectively direct editing anywhere in the genome and explore the possibilities of CRISPR editing systems.

Making a single genetic perturbation (*i.e.* 'edit') in a clonal population or model organism has become easier with CRISPR editing, and will remain a powerful approach for highly detailed study of specific variants. However, towards identifying all loci genome-wide that contribute to a given phenotype (element discovery) and towards quantifying the impact of all potential variants at a locus of importance (element dissection), multiplex approaches will be essential. Indeed, fully understanding the rules of complex biological processes — such as, how regulatory DNA elements act to coordinate gene regulation — will necessitate large experimental data sets. Here, we consolidate our thoughts on CRISPR-based multiplex assays in relation to the human genome, but these approaches are generally suited to other commonly studied organisms as well.

CRISPR-mediated editing has been used in two primary modes. Simply directing Cas9 cleavage to an element is a powerful approach for ablating function, as the insertions and deletions ('indels') that arise subsequent to double-strand break repair by non-homologous end joining (NHEJ) or microhomology-mediated end joining (MMEJ) are often sufficient to disrupt function. This is particularly true in coding regions. Indel events from a single cleavage are generally short, usually less than 10 bp (Tsai et al., 2015). However, inducing paired cleavage events can lead to a variety of repair outcomes that can effectively ablate or alter the underlying function of much larger sequences. These include large deletions, inversions, translocations, and duplications (Choi and Meyerson, 2014; Essletzbichler et al., 2014a; Guo et al., 2015; Li et al., 2015). While these events can be precisely engineered through clonal isolation of successfully edited cells, the requirement of dual cutting means they are generally made less reliably than short indels at a single cleavage site.

## 1.3 SYSTEMATICALLY IDENTIFYING GENE FUNCTION WITH CRISPR SCREENING

The scalability of CRISPR methods to disrupt function stems from the fact that all that is needed to direct cleavage at many locations is a single enzyme (*e.g.* SpCas9) and a specific gRNA for each target. Therefore, given the ease of gRNA design and the scale of modern oligonucleotide synthesis methods, directing cleavage to hundreds of thousands of sites in a single experiment is readily doable. Directly reading out indel events at each of thousands of sites with sequencing would prove difficult, but knowledge of the gRNA sequence alone can effectively inform where cutting occurred. Therefore, targeted sequencing of lentivirally delivered gRNA cassettes can serve as a quantitative proxies for the genetic alterations driving phenotypes. In this manner, CRISPR screening approaches have been developed to target hundreds of thousands of sites in a single experiment (Shalem et al., 2014; Wang et al., 2014).

Accumulating evidence suggests CRISPR-mediated screens for gene-level effects compare favorably to RNA interference screens (Evers et al., 2016; Morgens et al., 2016). This is likely a testament to the reliability of indel creation over time and the ensuing permanence of gene perturbation. The ease of delivering these libraries and using gRNA sequencing to measure effects has enabled hundreds of screens to date (Tsherniak et al., 2017). Profiling cells for which genes and pathways are essential for growth both with and without various drug treatments is one immediate application with implications for cancer biology, where new vulnerabilities can be identified. Screening multiple cell lines with diverse dependencies allows function to be assigned to unannotated genes by comparing how similarly an unknown gene scores to genes in a particular pathway across different cell lines (Wang et al., 2017). This approach will grow more powerful as more cell types are assayed. Other assays that physically separate cells of different phenotypes

(*e.g.* cell sorting, drug selection) are also readily compatible with CRISPR screens (Canver et al., 2015; Doench et al., 2014).

CRISPR editing targeted to many sites in multiplex can also be combined with single-cell RNA-seq (Datlinger et al., 2017; Dixit et al., 2016; Jaitin et al., 2016). In these experiments, the gRNA sequence of a cell is ascertained in combination with the transcriptome of a single cell via a shared molecular tag. As single-cell RNA-seq costs continue to drop precipitously (Cao et al., 2017; Macosko et al., 2015), this approach has the power to link genetic perturbations to omics phenotypes, meaning diverse effects of disrupting genes of unknown consequence can be revealed without need for informed hypotheses.

One caveat of CRISPR knockout screening is that effects of a single gene may not be revealed if there are multiple genes with redundant function expressed, each individually sufficient. CRISPR screens to target multiple genes at once, however, have been developed to enable profiling interactions between genes (Han et al., 2017; Shen et al., 2017). Such paired all-by-all screens have revealed non-additive effects, albeit infrequently. More hypothesis-driven combinations of gRNAs or pre-engineered genetic backgrounds may effectively reveal more gene interactions (Wang et al., 2017). CRISPR technology enables targeting more than two genes per construct (Xie et al., 2015), but comprehensive coverage becomes difficult as the combinatorial space increases exponentially.

Although further optimization will prove valuable for improving CRISPR screens to assay gene function (Sanjana et al., 2014), within just a few years of the first screens, it has become clear that implementing these approaches across diverse biological fields will help close remaining gaps in our knowledge of which genes contribute to which pathways and diseases.

## 1.4 TOWARDS EFFICIENT IDENTIFICATION OF NON-CODING ELEMENTS

CRISPR tools are now also being deployed to define regulatory elements and the mechanisms by which they act. While databases such as ENCODE provide omics data that can be used to define candidate gene regulatory elements, definitive proof of function must still come from sequence perturbation. Experiments towards this end will delineate the elements that contribute to each gene's regulation. A paramount challenge, however, is that the rules of coding gene structure do not apply, making it harder to predict whether indels created by editing will disrupt an element's function, and if so, to what extent. Whereas for genes, there are typically many gRNAs that can effectively induce loss-of-function mutations, it is unclear how many gRNAs can be expected to ablate a regulatory element's function (*e.g.* by disrupting a key factor binding site). This makes it more challenging to rule out off-target effects when few gRNAs (or only one) show a strong effect. Yet, similar to genome-wide screens for gene function, various CRISPR screening strategies have been employed to investigate potential regulatory elements. Generally, these approaches have shown promise, but would benefit from further technological development to more robustly perturb function.

The first CRISPR screens for regulatory function used single gRNAs to target short indels to loci predicted by a variety of features to potentially function as enhancers (Canver et al., 2015; Diao et al., 2016; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016). These features included the presence of sequence variants associated with expression differences, chromatin accessibility, ChIP-seq evidence of factor binding and enhancer-associated histone modifications, specific factor binding motifs, and gene proximity. Read-outs were used that reflected expression changes and were amenable to multiplex assays, such as isolating cells with altered gene expression by fluorescence-activated cell sorting (FACS) (Canver et al., 2015) or measuring

fitness effects stemming from expression changes (Korkmaz et al., 2016). These studies proved that CRISPR-screening could be used to identify both established and previously unrecognized regulatory elements, but were generally marked by low discovery rates, despite targeting rationally chosen candidates. Whether few new elements were identified due to biological reasons (*e.g.* few enhancers exist per gene, enhancer redundancy masks function, enhancer function is tolerant to small indels, etc.) or technical reasons (off-target effects preclude reliable measurement of small effect sizes, gRNAs do not direct editing to the key regions of active enhancers, etc.) remained undetermined.

To complement single gRNA regulatory screening approaches, we and others developed methods to express gRNA pairs to program kilobase-scale deletions (Diao et al., 2017; Gasperini et al., 2017; Zhu et al., 2016). The advantage of this strategy is that each base pair of genomic sequence can be queried for function many times over via deletions made with independent sets of gRNAs. However, the desired deletions require simultaneous target cleavage and deletion of intervening sequence, which occurs at a suboptimal rate. Initial results from this approach varied depending on the locus and the functional assay. In the 2 Mb surrounding *POU5F1*, 45 cis-regulatory elements were identified whose targeting affected the gene's expression (Diao et al., 2017). This is in contrast to what we found for *HPRT1*. In the 206 Kb encompassing *HPRT1*, not a single distal regulatory element was essential for *HPRT1* expression (Gasperini et al., 2017). Going forward, we predict that this 'tiling deletion' method will be valuable for systematically screening large amounts of sequence for regulatory function.

## 1.5 Dissection of genomic regions with multiplex editing of critical loci

The strategies discussed above center on identifying functional genes and regulatory elements, rather than dissecting them at high resolution to ask how each constituent nucleotide contributes to function. Towards answering questions such as will a rare single nucleotide variant in a tumor suppressor gene increase cancer risk, or how will a specific GWAS-identified variant located in a transcription factor binding motif affect enhancer activity, more precise mutations need to be engineered. In many cell types, such mutations can be engineered through homology-directed DNA repair (HDR) by providing a donor template containing the desired allele (Liang et al., 1998). Much work has centered on increasing the HDR rate within cells to more effectively engineer specific alleles with CRISPR systems (Liang et al., 2017; Lin et al., 2014; Maruyama et al., 2015; Song et al., 2016; Zhang et al., 2017). Furthermore, alternative editing strategies have been developed that obviate the requirement of HDR, instead relying on end-joining to make programmed edits (Suzuki et al., 2016). Despite this progress, due to the requirement of specific donor templates and the fact HDR often occurs less frequently than competing repair pathways, methods for editing specific variants into the genome in multiplex have been less widely used.

Towards the goal of dissecting critical genomic elements at scale, we developed a method called ‘saturation genome editing’ in which a library of hundreds to thousands of programmed variants is introduced to a specific locus via HDR and assayed in multiplex via targeted deep sequencing (Findlay et al., 2014). We employed this approach in human cell lines to assay all single nucleotide variants across ~100 bp of genomic sequence for effects on splicing or protein function. A similar approach was also used to aid experimental evolution in yeast (Ryan et al., 2014) and to assay chromatin accessibility of synthetic sequences introduced to mouse embryonic

stem cells (Hashimoto et al., 2016). More recently, we've illustrated the potential of saturation genome editing for functionally characterizing variants of uncertain significance by testing 96.5% of all possible SNVs (~4,000 SNVs total) across 13 exons *BRCA1* (detailed in Chapter 3). Critically, by engineering SNVs in their native context, we achieve near-perfect accuracy (~98%) when predicting clinical interpretations of pathogenicity from our data, regardless of the mechanism by which a SNV impacts function (*e.g.* protein folding, splicing, etc.). High quality data such as this will be vital for providing more definitive results to people for whom genetic testing reveals a variant of unknown consequence (Starita et al., 2017).

One limit of our approach is that the genomic region that can be deeply mutagenized is limited by the length of gene conversion tracts arising during HDR (Findlay et al., 2014; Paquet et al., 2016). To induce single nucleotide substitutions over larger regions, fusion proteins that combine the highly specific and programmable targeting of SpCas9 with enzymes that induce single nucleotide substitutions (*e.g.* cytidine deaminases) have been used to assay phenotypes such as drug resistance across multiple exons (Hess et al., 2016; Ma et al., 2016). High allelic diversity at many sites can be generated simply by providing cells with a new gRNA for each site, this making this approach more scalable. A caveat, though, is that measuring functional selection on each edit in a diverse population still requires targeted sequencing.

Collectively, these methods are well suited for applications to regions where SNVs are known to be of high functional impact. As we highlight in subsequent chapters, there is great benefit to exploring sequence-function relationships at the resolution of single nucleotides.

## 1.6 TOPICS IN THIS DISSERTATION

The ensuing chapters of this dissertation stem from four projects I've worked on that used CRISPR/Cas9 genome editing to generate high amounts of DNA sequence variation at specific

sites, albeit to ask fundamentally different biological questions. Chapter 2 describes the development of saturation genome editing (SGE), a method to create and functionally assay all possible single nucleotide changes at a single region in the genome. Chapter 3 describes how we next optimized SGE and used it to assay thousands of variants in *BRCA1*, a gene linked to hereditary cancer predisposition in which thousands of people harbor variants of unknown consequence to their health. Chapter 4 stems from experiments Molly Gasperini and I jointly led to use CRISPR editing to search for regulatory function in the human genome. Chapter 5 is the most dissimilar to the others, in that we use CRISPR editing to record new information in the genome, instead of to reveal information already present. Aaron McKenna and I, in collaboration with others, show CRISPR editing can be used to record lineage relationships between cells of developing organisms. All chapters have been modified from manuscripts either in preparation (Chapter 3) or already published (Chapters 2, 4, and 5).

# Chapter 2. SATURATION EDITING OF GENOMIC REGIONS BY MULTIPLEX HOMOLOGY-DIRECTED REPAIR

Chapter 2 is adapted with minimal modifications from:

> Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature *513*, 120–123.
>
> First authorship is shared between GMF and EAB.

## 2.1 ABSTRACT

Saturation mutagenesis – coupled to an appropriate biological assay – represents a fundamental means of achieving a high-resolution understanding of regulatory and protein-coding nucleic acid sequences of interest (Cunningham and Wells, 1989; Fowler et al., 2010; Myers et al., 1986; Patwardhan et al., 2009). However, mutagenized sequences introduced in trans on episomes or via random or "safe-harbor" integration fail to capture the native context of the endogenous chromosomal locus (Botstein and Shortle, 1985). This shortcoming markedly limits the interpretability of the resulting measurements of mutational impact. Here, we couple CRISPR/Cas9 RNA-guided cleavage (Jinek et al., 2012) with multiplex homology-directed repair (HDR) using a complex library of donor templates to demonstrate saturation editing of genomic regions. In exon 18 of *BRCA1*, we replace a six base-pair (bp) genomic region with all possible hexamers, or the full exon with all possible single nucleotide variants (SNVs), and measure strong effects on transcript abundance attributable to nonsense-mediated decay and exonic splicing elements. We similarly perform saturation genome editing of a well-conserved coding region of an essential gene, DBR1, and measure relative effects on growth that correlate with functional

impact. Measurement of the functional consequences of large numbers of mutations with saturation genome editing will potentially facilitate high-resolution functional dissection of both cis-regulatory elements and trans-acting factors, as well as the interpretation of variants of uncertain significance observed in clinical sequencing.

## 2.2 MAIN TEXT

Functional consequences of genetic variants are best studied by manipulating the endogenous locus, which provides the native chromosomal context with respect to DNA sequence and epigenetic milieu, and for proteins, endogenous levels and patterns of expression (Gibson et al., 2013). Programmable endonucleases, e.g. zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) or clustered regularly interspaced short palindromic repeat (CRISPR)/Cas-based RNA-guided DNA endonucleases, enable direct genome editing with increasing practicality (Gaj et al., 2013). However, genome editing has primarily been applied to introduce single changes to one or a few genomic loci (Wang et al., 2013), rather than many programmed changes to a single genomic locus.

We sought to leverage CRISPR/Cas9 (Cong et al., 2013; Jinek et al., 2012; Mali et al., 2013) to introduce saturating sets of programmed edits to a specific locus via multiplex HDR. We first targeted six bases of a *BRCA1* exon (Mazoyer et al., 1998). We cloned an HDR library containing random hexamers substituted at positions +5 to +10 of *BRCA1* exon 18 and fixed, nonsynonymous changes at positions +17 to +23 (as a 'handle' for selective PCR and to prevent re-cutting (Sander and Joung, 2014) by destroying the protospacer adjacent motif (PAM) (**Figure 2.1**). We co-transfected pCas9-sgBRCA1x18 and the HDR library into ~800,000 HEK293T cells, achieving 3.33% HDR efficiency. We performed two independent transfections with the same HDR library ('biological replicates' 1, 2), and cells were split on day 3 ('D3 replicates' a, b).

We prepared genomic DNA (gDNA) and cDNA from bulk cells on D5. PCR reactions were primed on the 'handle' uniquely present within successfully edited genomes. Amplification was observed in HDR library/pCas9-sgBRCA1x18 transfected samples, but not in HDR library-only controls. Amplicons derived from gDNA and cDNA were deeply sequenced (**Figure 2.1a**). The relative abundances of hexamers within replicates and the correlation between the HDR library and edited gDNA were consistent with limited 'bottlenecking' during transfection and minimal influence of hexamer identity on HDR efficiency (**Figure 2.2**, **Figure 2.3**).

We estimated the effect of introducing each hexamer to these genomic coordinates on transcript abundance by calculating enrichment scores (cDNA divided by gDNA counts, calibrated to wildtype). These enrichment scores were well correlated between biological replicates (**Figure 2.1b**, 1a vs. 2a: R = 0.659) and between D3 replicates (**Figure 2.3c**; 1a vs. 1b: R = 0.662). When we pooled read counts from D3 replicates, correlation between biological replicates improved (**Figure 2.3d**; 1 vs. 2: R = 0.706).

To maximize precision, we merged data across all four replicates for 4,048 hexamers (**Figure 2.1c**). Several results support the biological validity of the resulting enrichment scores. First, as anticipated by nonsense-mediated decay (NMD), hexamers introducing stop codons were associated with markedly reduced mRNA levels (**Figure 2.1c**; Wilcoxon rank sum test (WRST) $P = 9.7 \times 10^{-84}$; median for nonsense hexamers 12-fold below overall median). Second, previous studies measured hexamer influence on splicing at analogous coordinates of different exons via a plasmid minigene assay (Ke et al., 2011). Despite these contextual differences, the strongest exonic splicing silencers (ESSs) (bottom 2% in the minigene assay) scored 9-fold below median (**Figure 2.1c**; WRST $P = 2.0 \times 10^{-24}$), the strongest exonic splicing enhancers (ESEs) (top 2% in the minigene assay) scored 1.5-fold above median (**Figure 2.1c**; WRST $P = 2.4 \times 10^{-11}$), and the

complete datasets correlated reasonably well (**Figure 2.4a**; ρ = 0.524). We also observed correlation between GC content and enrichment scores (**Figure 2.4b**), strongest for bases most proximal to the splice junction, consistent with a posited role for GC content in the stability of splicing structures (Zhang et al., 2011).

We next sought to assay the effects of SNVs across the full 78 bp *BRCA1* exon 18 (**Figure 2.5**). We cloned three HDR libraries with selective PCR sites in either the 5' or 3' region and 3% doping (Patwardhan et al., 2012) (97(wt):1:1:1) in the other half of the exon (L: 5' degeneracy, 3' nonsynonymous selective PCR site; R: 3' degeneracy, 5' nonsynonymous selective PCR site; R2: 3' degeneracy, 5' synonymous selective PCR site). Five days post-transfection with pCas9-sgBRCA1x18 (1.02-1.29% HDR efficiency), we selectively amplified and deeply sequenced gDNA and cDNA. Using data from all edited exons with ≥1 mutation and ≥10 gDNA counts, we estimated effect sizes by (beta values) of all possible SNVs using a weighted linear model. Estimated effect sizes were reproducible (R = 0.846 (R), 0.853 (R2), and 0.686 (L); **Figure 2.6a**, **Figure 2.7**,**Figure 2.8**). Effect sizes for the same SNVs interrogated with different selective PCR strategies (R vs. R2) were also well correlated (R = 0.847; **Figure 2.6b**).

The estimated effect sizes reflect empirically measured changes in transcript abundance resulting from programmed edits (**Figure 2.6c**). As expected with NMD, nonsense mutations reduced transcript abundance (WRST $P = 1.4 \times 10^{-203}$; 5.6-fold below median). Additionally, several missense and synonymous SNVs reproducibly resulted in large reductions in transcript abundance, and SNV effect sizes correlated with a predictive model for exonic variants that disrupt splicing (Mort et al., 2014) (ρ = 0.322; **Figure 2.9a**). Because library L does not destroy the PAM, we calculated enrichment scores for indels from non-homologous end-joining (NHEJ). As

expected with NMD, only frameshifting indels were associated with large depletions (**Figure 2.9b**,**c**).

To further demonstrate this method, we targeted a well-conserved region of *DBR1*, the RNA lariat debranching enzyme, which scored highly in a genome-wide screen for essentiality (Wang et al., 2014) (**Figure 2.10**). We used array-synthesized oligonucleotides to program a *DBR1* HDR library to include the wild-type sequence and every possible SNV across 75 bp (73 3'-most bases of exon 2 and first two bases of intron 2), and also all 63 possible codon substitutions at three residues (388 genome edits were programmed; single base deletions were abundant from synthesis errors). The HDR library also introduced two fixed synonymous changes (to disrupt the PAM and prevent re-cutting) and a selective PCR site in intron 2.

An optimized sgRNA sequence (Hsu et al., 2013; Ran et al., 2013) was cloned into a bicistronic sgRNA/Cas9-2A-EGFP vector (pCas9-EGFP-sgDbr1x2). Five million haploid human cells (HAP1; Carette et al., 2009) were co-transfected with the *DBR1* HDR library and pCas9-EGFP-sgDbr1x2. On D2, ~250,000 EGFP+ cells were FACS-sorted and further cultured, taking samples on D5, D8 and D11 (1.14% HDR efficiency, estimated on D8). Following gDNA isolation and selective PCR, deep sequencing was performed to quantify the relative abundance of edited haplotypes in each sample.

We first examined the relative proportions of mutation classes at each time point (**Figure 2.11a**). The strong enrichment of synonymous mutations and depletion of nonsense and frameshifting mutations over time indicated that selection was acting on edited cells in culture, consistent with *DBR1* essentiality. We calculated enrichment scores (D8 or D11 counts divided by D5 counts) for 365 of the 388 (94%) programmed edits and 12 single base deletions (the subset with relative abundance $>5 \times 10^{-5}$ on D5) (**Figure 2.11b**, **Figure 2.12**). Enrichment scores strongly

correlated with functional consequence. The median enrichment score for synonymous edits was nearly identical to wild-type (1.006-fold lower), but 73-fold lower for missense edits ($P$ = 1.7 x $10^{-8}$; WRST against synonymous edits), 207-fold lower for nonsense edits ($P$ = 1.9 x $10^{-9}$), and 211-fold lower for frameshifting single base deletion edits ($P$ = 1.5 x $10^{-8}$). Furthermore, enrichment scores for SNVs were inversely correlated with metrics of predicted deleteriousness like CADD (Kircher et al., 2014) ($\rho$ = -0.295; $P$ = 1.2 x $10^{-5}$; **Figure 2.13a**,**b**). Residues N84, H85 and E86 of DBR1 were edited to all 63 possible non-wild-type codons. Consistent with their predicted role in the active site of an essential enzyme (Khalid et al., 2005), only synonymous mutations and a few missense substitutions were tolerated (**Figure 2.13c**).

Amino-acid level enrichment scores were well correlated between D11 biological replicates (R = 0.752; $P$ = 2.6 x $10^{-40}$; **Figure 2.13c**), and were bimodally distributed in each replicate, allowing broad classification of changes as tolerated or deleterious. The small proportion of discordantly classified variants might be explained by HAP1 reversion to diploidy or off-target effects, highlighting the importance of biological replicates for this experimental design. Notably, there were no reproducibly tolerated nonsense or frameshifting edits. Overall, these data support the conclusion that our empirically derived enrichment scores reflect true biological effects of specific genomic point mutations within *DBR1*.

We demonstrate that it is feasible to generate and functionally analyze hundreds to thousands of programmed genome edits at a single locus in a single experiment. We emphasize three major limitations of the method as it stands. First, we only introduced programmed edits to the immediate vicinity of coordinates targeted by the endonuclease (**Figure 2.7a**, **Figure 2.12a**), and the narrow window associated with HDR mechanisms in mammalian cells (Elliott et al., 1998) may fundamentally limit the size of the region that can be subjected to multiplex editing in one

experiment. Saturation genome editing of a full gene – *e.g.* to measure functional consequences of all possible variants of uncertain significance – will require multiple experiments tiling along its exons.

Second, only a small proportion of cells were successfully edited in each experiment, bottlenecking complexity, limiting reproducibility, and necessitating the selective PCR site. Looking forward, a variety of techniques, *e.g.* transient hypothermia (Doyon et al., 2010) or oligonucleotide-based HDR (Chen et al., 2011), can be used to improve editing efficiency. Consistent with this, we note that ZFNs and TALENs have demonstrated efficiencies up to 50% in some studies (Carroll, 2014; Reyon et al., 2012). Also, although the low editing efficiency necessitated using haploid cells for DBR1 mutagenesis, this could potentially have been performed in diploid cells by knocking out one allele via NHEJ and then knocking in the HDR library to the other allele.

Finally, the development of functional assays that are biologically relevant and technically viable remains a challenge. Here, we exploited strategies that directly linked genotype to phenotype – *e.g.* targeted RNA sequencing to measure transcript abundance or targeted DNA sequencing to measure reduced cellular fitness. Analogous approaches can be taken in other contexts – *e.g.* targeted ChIP-seq of co-activators to assay enhancers, increased cellular growth rate to assay cancer drivers or drug resistance (Smurnyy et al., 2014), or FACS-based phenotypic sorting for cellular assays more generally (Kinney et al., 2010).

There is a strong demand for techniques that accurately and scalably measure mutational consequences, and a dearth of experimental data measuring distributions of effect sizes or corresponding to direct manipulation of the genome. By multiplexing both the introduction and

assaying of mutations in their native context, we anticipate that saturation genome editing will accelerate our ability to measure and interpret the functional consequences of genetic variation.

## 2.3 METHODS

### 2.3.1 *BRCA1 Experimental Design*

As a proof-of-principle experiment, we chose to target an exon in a clinically relevant gene in which known mutations cause aberrant splicing. Previous molecular studies of a G to T nonsense mutation occurring naturally in cancer patients at chr17:41,215,963 suggested exon skipping (Mazoyer et al., 1998) was secondary to the creation of an exonic splicing silencer site (Goina et al., 2008). From this, we hypothesized that saturation genome editing of this exon could result in a wide range of splicing outcomes.

A chief consideration when performing parallel functional analysis of complex allelic series is the challenge of associating each of many mutations with the biological effects they produce. This task is more difficult when attempting such approaches at the endogenous genomic locus, and with limited editing efficiencies. By performing these experiments in an exon and focusing on the effects of mutations on transcript abundance, we directly link genotype and phenotype by observing the frequency of each genome edit in the transcript pool, relative to its frequency in genomic DNA. This design is advantageous because it requires no specialized (*i.e.* gene-specific) functional assay, thus making it amenable to interrogation of transcribed variants' effects on splicing/transcript abundance in any gene.

### 2.3.2 *Rationale for Including Selective PCR Sites*

Given the modest proportion of HDR-edited loci in a given experiment and the high number of variants that we set out to interrogate (*i.e.* hundreds to thousands), it would require a

large amount of sequencing to sufficiently sample every variant in gDNA and cDNA pools from a population of cells that are predominantly unedited or harboring products of NHEJ. Furthermore, at such efficiencies, the rate of error in high-throughput sequencing is high enough to obscure signal from single nucleotide variants (SNVs) (unpublished observations). Therefore, until better methods exist to isolate populations of cells successfully edited with HDR, techniques to selectively sequence molecules derived from edited cells are likely to be advantageous. To implement this, we designed our HDR libraries to include short, fixed edits to serve as unique priming sites in genomes that successfully undergo HDR. PCR reactions primed at this site, therefore, should only amplify material from edited cells, thus reducing both the noise associated with error from sequencing unedited material and the cost of sequencing in each experiment.

Additionally, we predicted that selective PCR sites that mutate the PAM and protospacer sequences would prevent Cas9 from re-cutting HDR-edited genomes. This should have the effect of increasing the proportion of cells bearing experimentally informative edits, and given the bottleneck imposed by limitations on how many successfully edited cells can be sampled, should result in more robust experimental signal.

### 2.3.3 *DBR1 Experimental Design*

To demonstrate that saturation genome editing can be used to explore effects of mutations on protein function and cellular fitness, we targeted *DBR1*, a well-conserved gene that scored highly in a human haploid cell genome-wide loss-of-function screen for essentiality (Wang et al., 2014). Using haploid cells prevents gene compensation from an unedited copy. Not knowing how sensitive the cells would be to mutations, we chose to target a region of exon 2 that was highly conserved, included in all transcript annotations on the UCSC Genome Browser, and coded for at least 2 residues (N84, H85) predicted to participate at the enzyme's active site (Khalid et al., 2005).

Selection against edited cells in culture allows phenotype to be linked to genotype from sequencing of the gDNA pool over a series of timepoints. During HDR library construction, we designed a selective PCR site in a downstream intron to minimize any effect on gene function, and used two synonymous mutations to abrogate Cas9 re-cutting.

Given the lower transfection efficiency of HAP1 cells (~4% for the plasmids used here), we cloned a *DBR1*-targeting CRISPR construct that expressed EGFP with Cas9 and used FACS to sort a population of successfully transfected cells. The sgRNA was designed using the Zhang Lab tool (http://crispr.mit.edu/), and selected to minimize off-target effects that could potentially impair cellular fitness (Hsu et al., 2013).

### 2.3.4 *HDR Library and Cas9-sgRNA Cloning*

A homology-directed repair (HDR) library containing all possible 4,096 DNA hexamers substituted at positions +5 to +10 of *BRCA1* exon 18 (chr17:41,215,962-41,215,967; CCDS11453.1) was constructed using a partially degenerate oligonucleotide (IDT DNA; "BRCA1ex18NNNNNN5_10selPCR") containing a 7 bp selective PCR site / EcoRV restriction digest site at position +17 to +23. The oligonucleotide was PCR amplified and cloned via the In-Fusion reaction (Clontech) into a PCR-linearized pUC19- BRCA1ex18 vector containing a pre-inserted 1,573 bp fragment amplified from the surrounding BRCA1ex18 locus in HEK293T cells (chr17:41,215,127-41,216,699) to serve as homologous arms. Additional libraries from a second degenerate oligonucleotide that was synthesized with a 3% mutation rate (97% wt, 1% each non-wt base) across the 78 bp exon were cloned similarly, such that one end of the exon would be fixed and contain either missense (as above) or synonymous mutations for selective PCR. All PCR reactions were performed with the KAPA HiFi HotStart ReadyMix PCR Kit.

The *DBR1* HDR library was cloned as above except with the following differences. HDR library variants were derived from 388 oligonucleotides synthesized on a microarray (CustomArray) to include all possible single base pair changes in a 75 bp region comprising part of *DBR1* exon 2 (chr3:137,892,342-137,892,416), all codon variants at the first three residues of the 75 bp region (chr3:137,892,408-137,892,416), and the reference 75 bp sequence. All *DBR1* HDR library sequences also included two synonymous mutations designed to prevent re-cutting of edited genomes by disrupting PAM and protospacer sequences (chr3:137,892,424 and chr3:137,892,421), and a 6 bp selective PCR site in intron 2 of *DBR1* (chr3:137,892,331-137,892,336). The library was cloned into a pUC19-DBR1ex2 backbone, a vector containing the surrounding *DBR1* sequence cloned from HAP1 gDNA (chr3:137,891,573-137,893,293). A bicistronic Cas9-sgRNA vector designed to cleave within *BRCA1* exon 18 ("pCas9-sgBRCA1x18") was cloned according to a published protocol (Ran et al., 2013) by ligating annealed oligonucleotides into a human codon-optimized S. pyogenous Cas9-sgRNA vector from the lab of Feng Zhang (pX330-U6-Chimeric_BB-CBh-hSpCas9; Addgene plasmid #42230). The same protocol was followed to create pCas9-EGFP-sgDbr1x2 from a similar vector that allows for fluorescent identification of Cas9-expressing cells (pSpCas9(BB)-2A-GFP (pX458); Addgene plasmid #48138).

### 2.3.5 *Cell Culture and Transfection*

For *BRCA1* experiments, HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (Life Technologies) supplemented with 10% FBS (AATC) and 100 U/ml penicillin + 100 µg/ml streptomycin (Life Technologies). One day prior to transfection, cells were split to ~40% confluency in 12-well plates with antibiotic-free media. The next day, 0.5-1.0 µg of each library was co-transfected (Lipofectamine 2000, Invitrogen) with an equivalent amount of pCas9-

sgBRCA1x18. Cells were expanded to 6-well plates, then split 1:4 on day 3 into two pools, and DNA and RNA were harvested on D5 (AllPrep DNA/RNA Mini Kit, Qiagen). Biological replicates of each transfection and negative control transfections of each library without pCas9-sgBRCA1x18 were also performed.

For the *DBR1* experiment, HAP1 cells (Haplogen) were cultured in Iscove's Modified Dulbecco's Medium supplemented with 10% FBS and 100 U/ml penicillin + 100 µg/ml streptomycin. ~3 x $10^6$ HAP1 cells were passaged to a 60 mm dish in antibiotic-free media one day prior to cotransfection with 3 µg each of pCas9-EGFP-sgDbr1X2 and the *DBR1* HDR library via Turbofectin 8.0 (OriGene) according to protocol. On D2, FACS was performed (BD FACSAria III) to isolate ~250,000 $EGFP^+$ cells, which were then expanded in culture with samples taken of ~1 x $10^6$ cells on D5, and 4-8 x $10^6$ on D8 and D11. gDNA was isolated according to protocol with the QiaAmp Kit (Qiagen). A biological replicate was performed, as well as negative controls in which the HDR library was transfected with the empty pSpCas9(BB)-2A-GFP construct (to enable FACS of transfected cells without editing).

### 2.3.6 *RT, Selective PCR and Sequencing*

For *BRCA1* experiments, reverse transcription (RT) was performed using SuperScriptIII (Invitrogen) with a gene-specific primer located in either *BRCA1* exon 19 (hexamer experiments) or exon 21 (whole exon experiments). Initial rounds of PCR were performed on large quantities of sample gDNA (8-12 µg gDNA, 100-150 ng/reaction) and cDNA (25 µg total RNA reverse transcribed and split into 45-47 reactions) using the KAPA HiFi HotStart ReadyMix PCR kit. In the first gDNA PCR, a primer external to the HDR library was used to prevent amplification of plasmid DNA. cDNA reactions were either primed from exons 16 and 18 (hexamer experiment; Library L) or exons 18 and 20 (Libraries R, R2). After the initial gDNA and cDNA reactions, all

PCR products from a single sample were pooled and purified using the QIAquick PCR Purification Kit (Qiagen).

For both cDNA and gDNA reactions, a primer designed to selectively amplify edited molecules bearing the selective PCR site was used either in the first or second reaction. Optimal annealing temperatures for each primer pair were determined via gradient PCR, and negative control reactions were performed using input from HDR library-only transfections to ensure products were derived from edited genomes as opposed to the HDR library. Negative controls failed to amplify for all experiments. Two subsequent PCRs were performed to add sequencing adaptors ("PU1L" and "PU1R"), sample indices, and flow cell adaptors.

For the *DBR1* experiment, 30 cycles of selective PCR were performed on gDNA (300 ng per reaction) from D5 (3 μg), D8 and D11 (27 μg each). Wells from each sample were pooled, PCR purified, and then re-amplified for 15 additional cycles. The 1,055 bp product was gel-purified (QIAquick Gel Extraction Kit, Qiagen), and two subsequent PCRs were performed to incorporate sequencing and flow cell adaptors prior to sequencing as above. After final reactions were purified (AMPure XP beads, Agencourt), paired-end sequencing was performed on all samples with the Illumina MiSeq to quantify gDNA and/or cDNA abundances for each edited haplotype.

HDR efficiencies were estimated for all experiments via deep sequencing of target loci by performing PCR on 150-300 ng of gDNA using primers external to the region of editing and the selective PCR site. Reported HDR efficiencies were conservatively calculated as the fraction of sequencing reads containing the selective PCR site and bearing at least one variant represented in the HDR library.

### 2.3.7 *Analysis of Sequencing Data*

For quality control, fully overlapping paired-end reads were merged with PEAR (Paired-End reAd mergeR) (Zhang et al., 2014) and discordant pairs were eliminated. By design, the mutagenized region is covered by both the forward and reverse reads on the Illumina platform, resulting in high-confidence calls per base.

For *BRCA1* hexamer reads to be included, the six bases on either side of the hexamer were required to match the reference sequence, and every base call in the hexamer required a quality score of at least Q30. For *BRCA1* whole-exon mutagenesis, the full read was required to be the correct length and match the library consensus sequence outside of the mutagenized region, every base quality score inside the mutagenized region was required to be at least Q30, and no indels were tolerated in alignment with BWA-MEM (Li and Durbin, 2009). cDNA reads not matching any gDNA haplotype with at least 10 reads were eliminated. After normalizing for sequencing coverage, enrichment scores were calculated as cDNA read counts incremented by one pseudocount divided by gDNA reads, calibrated to the wild-type hexamer.

For *DBR1* mutagenesis, reads were subjected to the same requirements of the sequence outside the mutagenized bases matching the consensus and every quality score in the mutagenized region exceeding Q30. Only reads matching programmed haplotypes were analyzed, and haplotypes below a D5 relative abundance of $5 \times 10^{-5}$ were excluded from analysis. After incrementing all read counts by one pseudocount and dividing by the total number of reads, the abundance of each haplotype on D8 or D11 was divided by the corresponding abundance on D5, and the fold-change relative to the wild-type sequence was taken to calculate an enrichment score. Based on the bimodal distribution observed in each replicate, mutations with log2-transformed enrichment scores less than -2 were considered “deleterious”; otherwise, mutations were

considered “tolerated”. Discordant effects between replicates were defined as mutations “tolerated” in one replicate but “deleterious” in the other. Amino acid level enrichment scores were calculated as the median of SNV enrichment scores for programmed edits resulting in the same change (or lack of change, for synonymous edits).

### 2.3.8 *SNV Effect Size Linear Modeling and Replicate Pooling*

To determine effects of SNVs in the *BRCA1* whole-exon experiments, cDNA and gDNA read counts were converted into percentages (number of reads for a given haplotype divided by the total number of reads for a given replicate) after discarding haplotypes with fewer than 10 gDNA Reads. To predict single nucleotide effect size across exon 18 of *BRCA1*, we then fit the weighted linear model. Regression analyses were performed in R 3.0.0 using the lm() function. The resulting coefficients of the model adjusted for the model intercepts were taken as effect sizes of the individual SNVs on exon splicing/stability. To merge data across replicates, effect sizes were averaged (including across overlapping bases between libraries L and R).

### 2.3.9 *Comparisons to other metrics of functional impact*

For comparison to plasmid studies, ESR-seq scores were taken from (Ke et al., 2011). Hexamers with positive ESR-seq scores are deemed exonic splicing enhancers, whereas negative ESR-seq scores denote exonic splicing silencers. For comparison of *BRCA1* exon 18’s SNV effect sizes to an *in silico* method, all SNVs were queried on MutPredSplice’s web server. MutPredSplice reports a single score estimating the likelihood that a variant will disrupt splicing at any genomic locus. Absolute values of *BRCA1* exon 18 splicing effect sizes were then correlated with MutPredSplice scores to determine concordance between our data and predicted effects on splicing.

For *DBR1*, calculated enrichment scores were compared to BLOSUM62 substitution scores (Henikoff and Henikoff, 1992) (obtained from NCBI), PolyPhen-2 (Adzhubei et al., 2010), and CADD (Kircher et al., 2014) (PolyPhen-2 and CADD scores obtained from querying genomic coordinates from CADD's precomputed genomic annotations. Whereas BLOSUM62 is derived from evolutionary conservation and PolyPhen-2 predicts changes in protein function, CADD is an integrated measure of deleteriousness that incorporates many functional annotations (including PolyPhen-2).

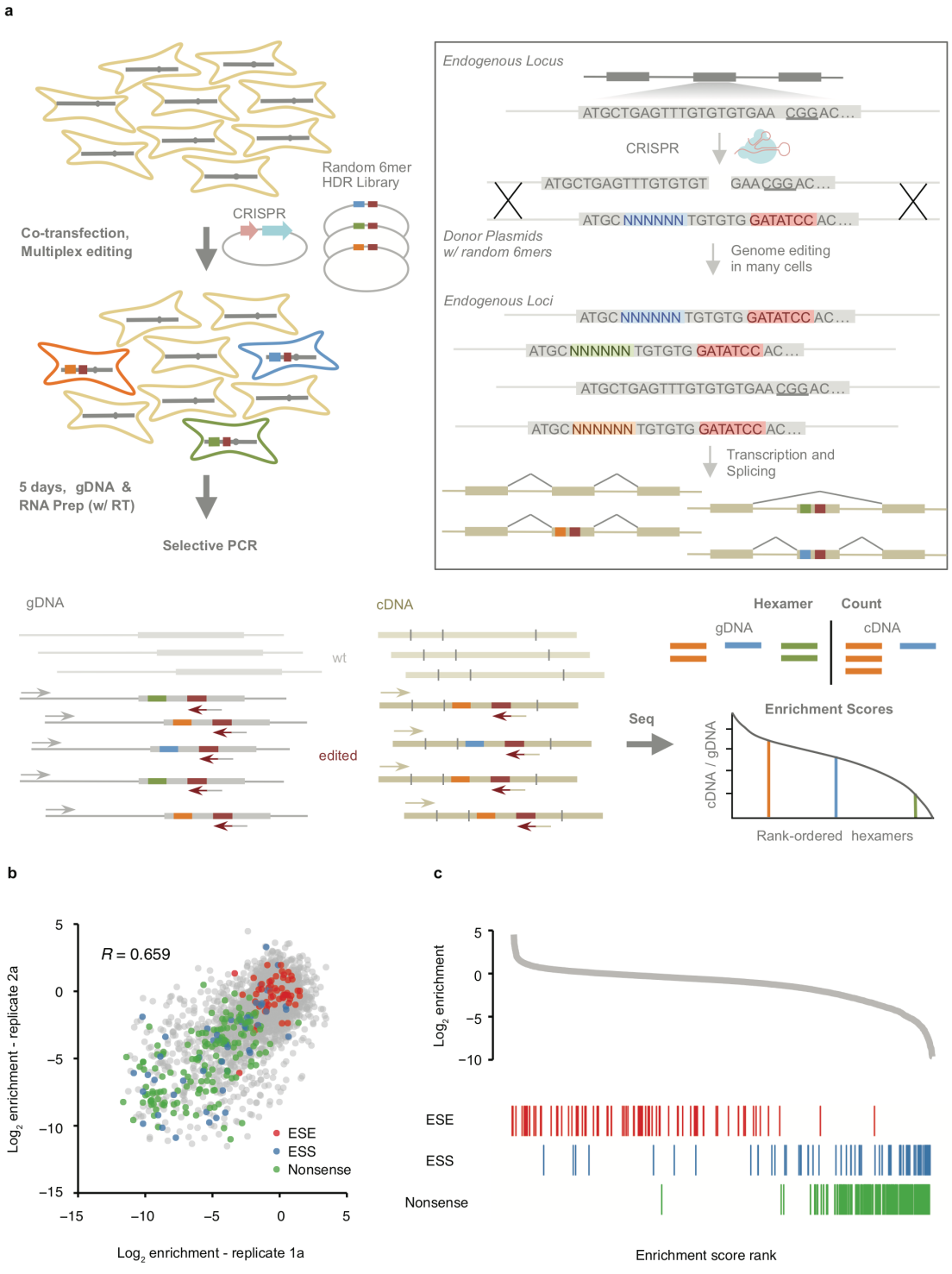



**Figure 2.1. Saturation genome editing and multiplex functional analysis of a hexamer region influencing *BRCA1* splicing.**

**a**, Experimental schematic. Cultured cells were co-transfected with a single Cas9-sgRNA construct

(CRISPR) and a complex homology-directed repair (HDR) library containing an edited exon that harbors a random hexamer (blue, green, orange) and a fixed selective PCR site (red). CRISPR-induced cutting stimulated homologous recombination with the HDR library, inserting mutant exons into the genomes of many cells. At five days post- transfection, cells were harvested for gDNA and RNA. After reverse transcription, selective PCR was performed followed by sequencing of gDNA and cDNA derived amplicons. Hexamer enrichment scores were calculated by dividing cDNA counts normalized by gDNA counts. **b**, Correlation of enrichment scores between biological replicates for hexamers observed in each experiment with positions of previously identified exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs) and stop codons indicated. **c**, Rank-ordered plot of enrichment scores with positions of ESEs, ESSs, and stop codons indicated.

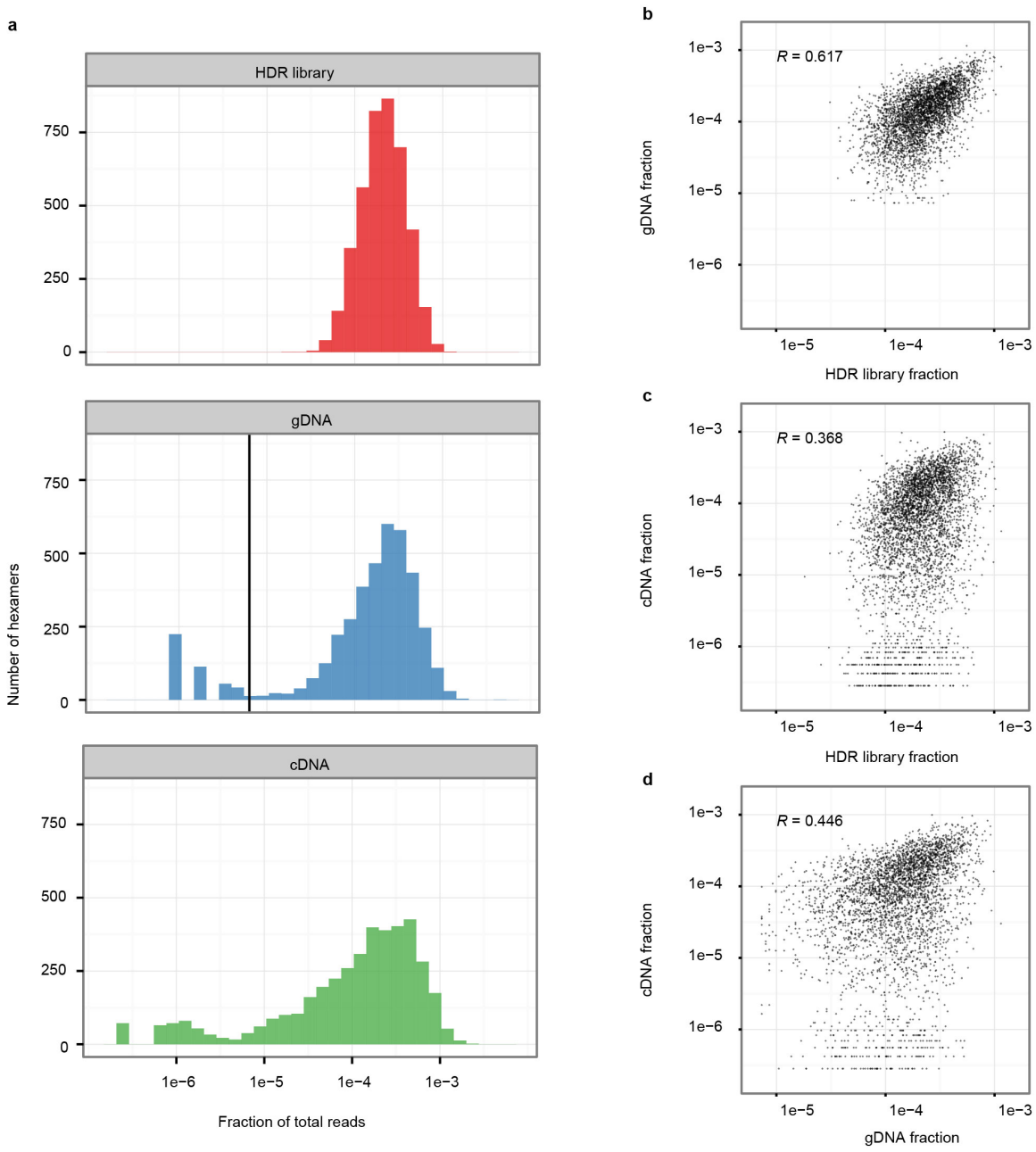



**Figure 2.2. Distributions and pair-wise correlations of hexamer abundances.**

**a**, The relative abundance of hexamers within the HDR library (red), gDNA (blue), cDNA data (green) are shown for a single experiment. The vertical black line represents our threshold of 10 gDNA reads. **b-d**, Scatterplots from a single replicate show pair-wise correlations between sequencing counts for the HDR library, gDNA, and cDNA for hexamers with at least 10 observations in the gDNA library, excluding wild type and control hexamers (n = 3,633). The HDR library and the gDNA data are most highly correlated (R 95% CI: 0.596-0.636), followed by the gDNA and cDNA (R 95% CI: 0.419-0.471) and the HDR library and cDNA (R 95% CI: 0.341-0.394).

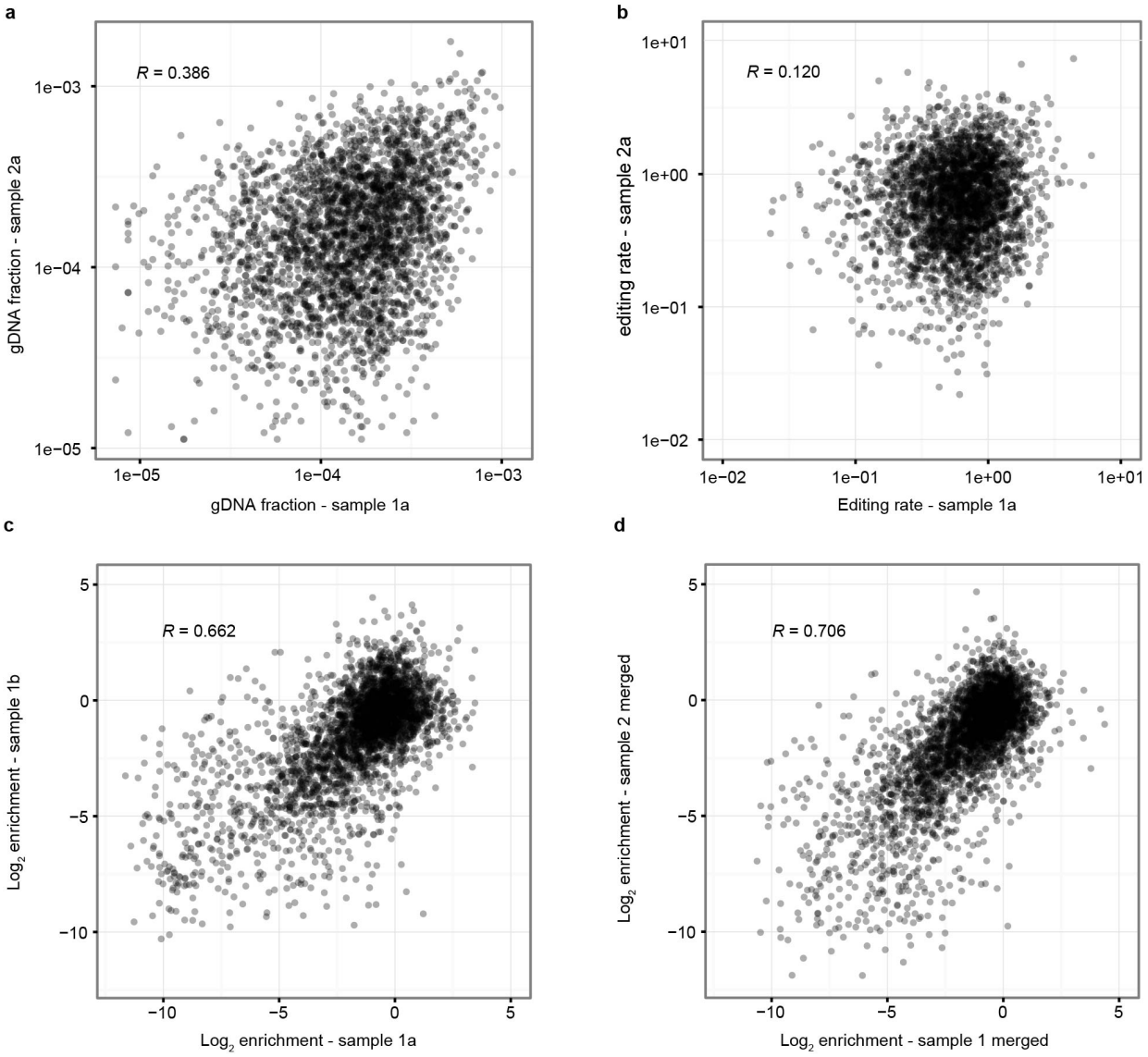



**Figure 2.3. Correlations for hexamer genome editing efficiency and enrichment scores between biological replicates.**

**a**, gDNA counts for all hexamers with at least ten reads in each of two gDNA preps from separate transfections with the same HDR library (n = 2,980) exhibited moderate correlation (R 95% CI: 0.355-0.416). **b**, However, hexamer editing rates, defined as gDNA counts normalized to HDR library counts, were substantially less correlated (*R* 95% CI: 0.084-0.155), consistent with a hexamer's HDR library abundance contributing more to its gDNA abundance than systematic differences in HDR efficiency secondary to the hexamer sequence itself. **c**, Hexamer enrichment scores for two pools of cells from a single transfection split on D3 were well-correlated (*R* 95% CI: 0.643- 0.681). **d**, Pooling data from cells split on D3 replicates from a single transfection yielded an improved correlation between biological replicates (i.e. independent transfections; *R* 95% CI: 0.690-0.722).

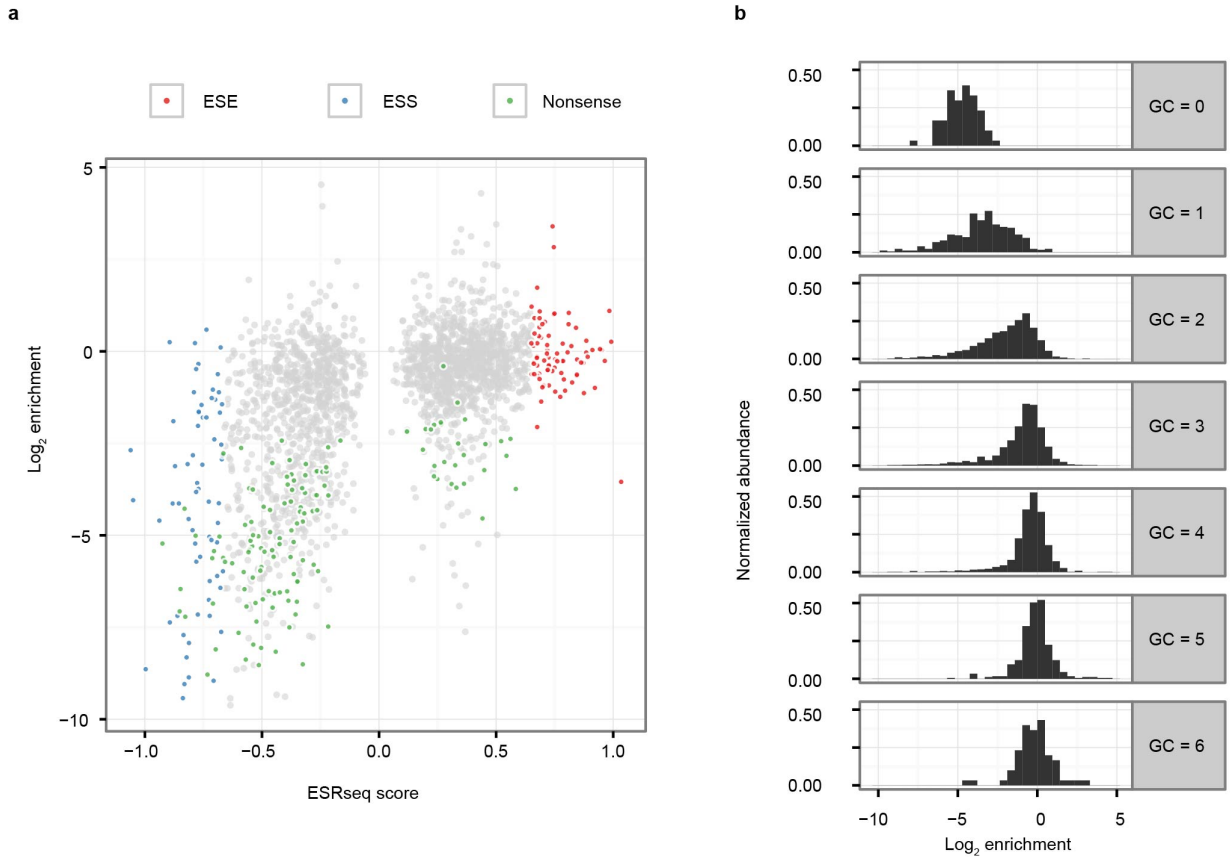



**Figure 2.4. Comparison of genome-based hexamer enrichment scores to plasmid-based hexamer scores.**

**a**, There was a modest correlation between ESS and ESE hexamers defined by a previous study (Ke et al., 2011) (x-axis) and the enrichment scores calculated here (y- axis; Spearman = 0.524). The previous study also interrogated hexamers positioned +5 to +10 nucleotides relative to a splice junction, but was plasmid-based rather than genome-based and in the context of different exons. **b**, To reveal effects of GC content on hexamer abundance, histograms display the distribution of enrichment scores for each possible GC level (0-6). Hexamers containing two or fewer GC base pairs exhibited broadly lower enrichment scores than hexamers containing three or more GC base pairs.

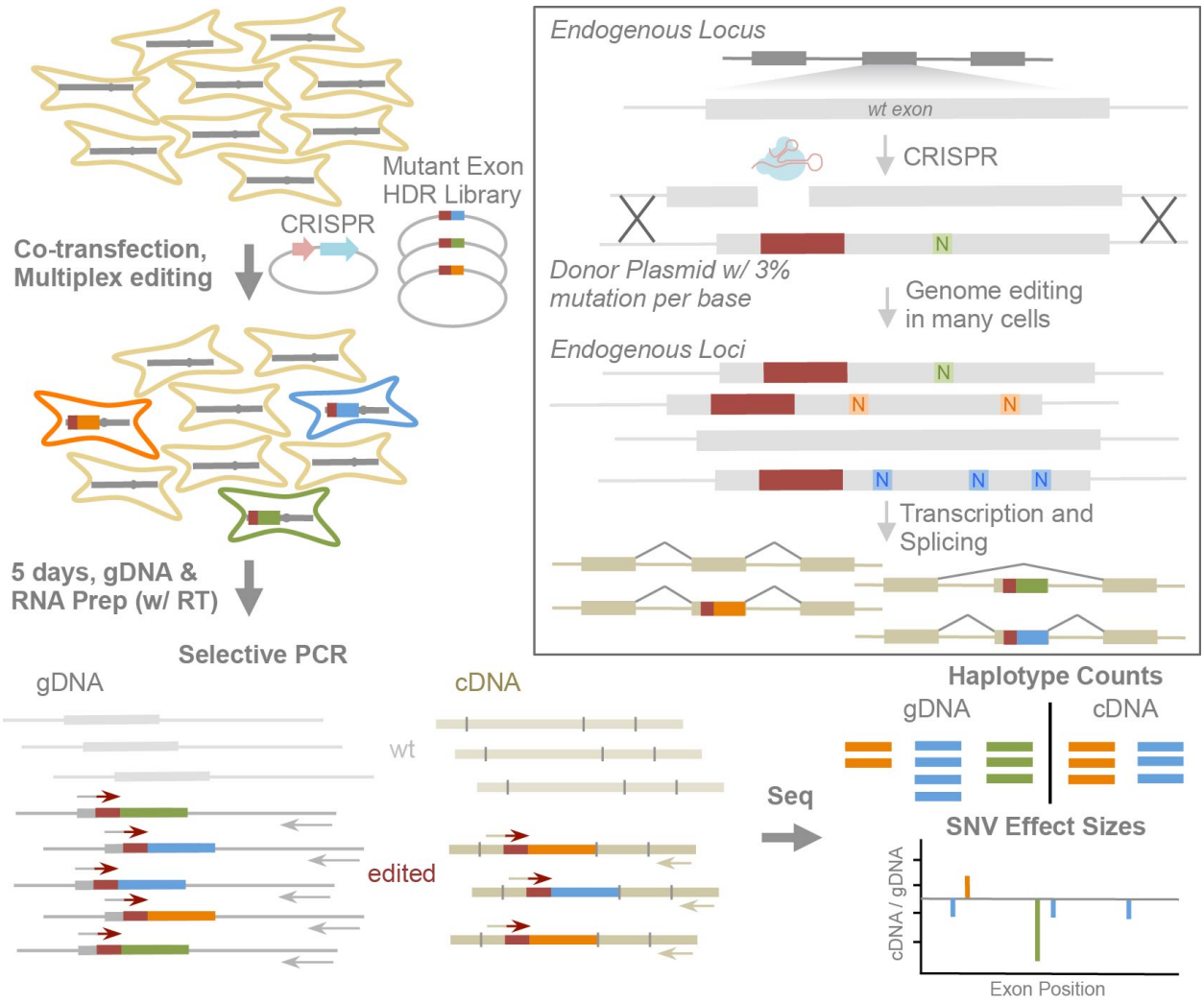



**Figure 2.5. Experimental schematic for genome editing and functional analysis of *BRCA1* exon 18.**

Cultured cells were co-transfected with a single Cas9-sgRNA construct (CRISPR) and an HDR library. Each HDR library was generated from cloning of an oligonucleotide synthesized with 3% nucleotide degeneracy (97wt:1:1:1) for approximately half of the exon and a selective PCR site introduced to the other (fixed) half of the exon (red). CRISPR-induced HDR integrates mutant exons into the genome. Cells were cultured for five days post-transfection, and then harvested for gDNA and total RNA. After reverse transcription, selective PCR was performed prior to sequencing the edited pools of gDNA and cDNA. Each exon haplotype's enrichment score was measured by dividing cDNA reads by gDNA reads, and effect sizes for each SNV were calculated via weighted linear regression.

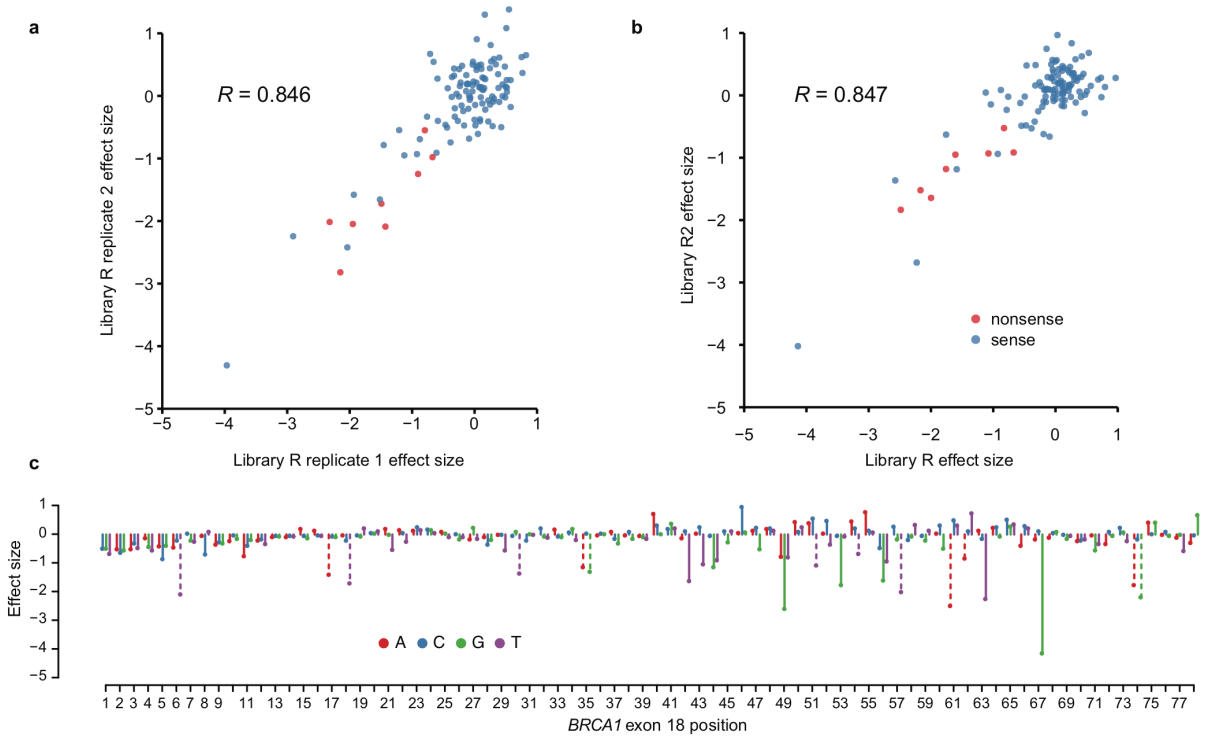



**Figure 2.6. Multiplex homology-directed repair reveals effects of single nucleotide variants on transcript abundance.**

Three separate HDR libraries (R, R2, and L) containing a 3% mutation rate (97% wt, 1% each non-wt base) in either half of *BRCA1* exon 18 were introduced to the genome via co-transfection with pCas9-sgBRCA1x18. Enrichment scores were calculated for each haplotype observed at least 10 times in the gDNA, and effect sizes of SNVs were determined by weighted linear regression modeling. 'Sense' includes both missense and synonymous SNVs. **a,** Effect sizes calculated from replicate transfections of HDR library R, consisting of a 3% per-nucleotide mutation rate in the 3'-most 39 bases, were highly correlated ($R$ = 0.846). **b**, Library R2 harbored a selective PCR site composed of 5 synonymous changes, none of which are present in Library R. When effect sizes derived from experiments with library R2 were plotted against those from library R, there was a strong correlation ($R$ = 0.847), indicating reproducibility and demonstrating that differences between selective PCR sites did not strongly influence scores. **c**, Effect sizes for SNVs across the exon are displayed. Datasets from libraries R and L were combined to span the entire exon. Dashed lines represent SNVs that introduce nonsense codons.

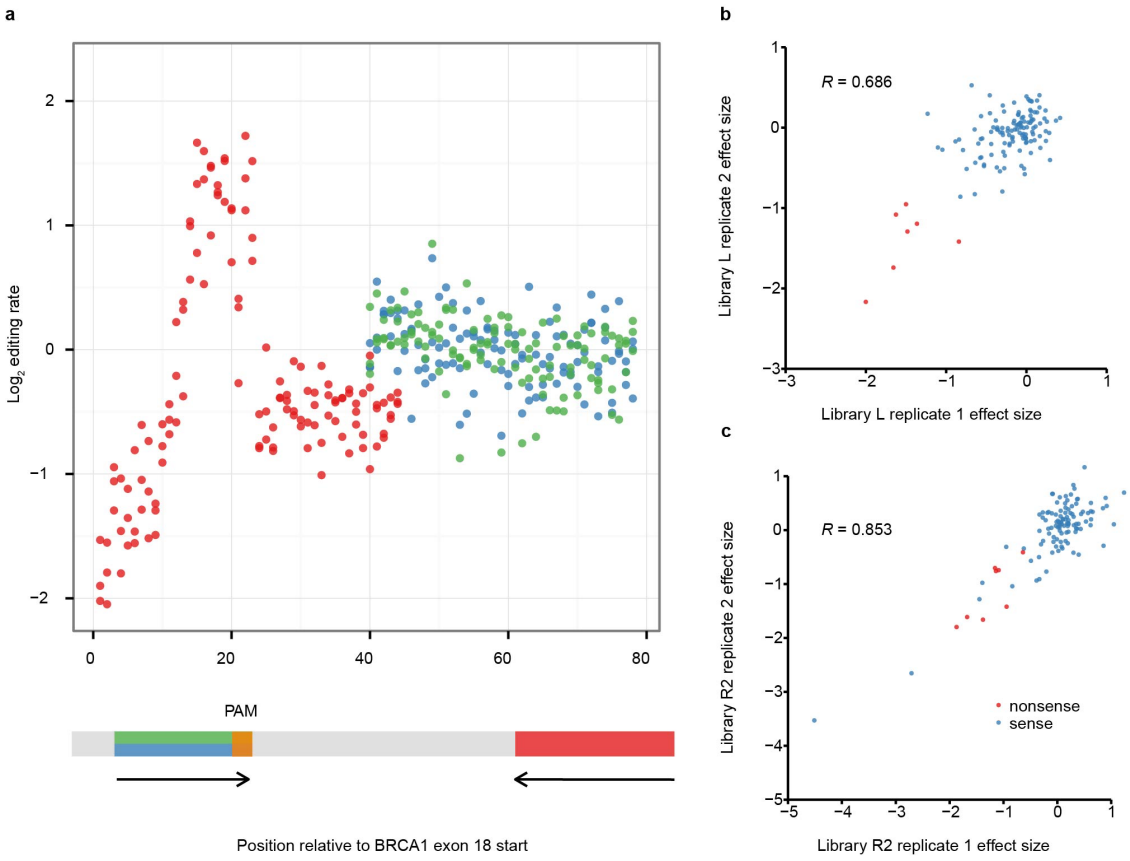



**Figure 2.7. Positional SNV editing rates and replication of effect sizes.**

**a**, Editing rates for each SNV in *BRCA1* exon 18 were calculated by dividing each SNV's gDNA sequencing abundance by its HDR library abundance. Editing rates were then plotted across the exon for each library (red = L, blue = R, green = R2) with locations of their selective PCR sites and the CRISPR-targeted PAM illustrated below. For HDR libraries R and R2, there was a subtle decrease in editing rate with increasing distance from the Cas9 cleavage site (R: rho = -0.264, $P = 4.1 \times 10^{-3}$; R2: rho = -0.361, $P = 4.8 \times 10^{-5}$). For library L, which allowed re-cutting by not destroying the PAM, there was a sharp peak centered on the Cas9 cleavage site, and a rapid decline in efficiencies in the 5' direction (further from the 3' selective PCR handle). **b**,**c**, SNV effect sizes were concordant across biological replicates for libraries R2 (**b**) and L (**c**). Notably, variants of high effect size scored similarly across independent transfections.

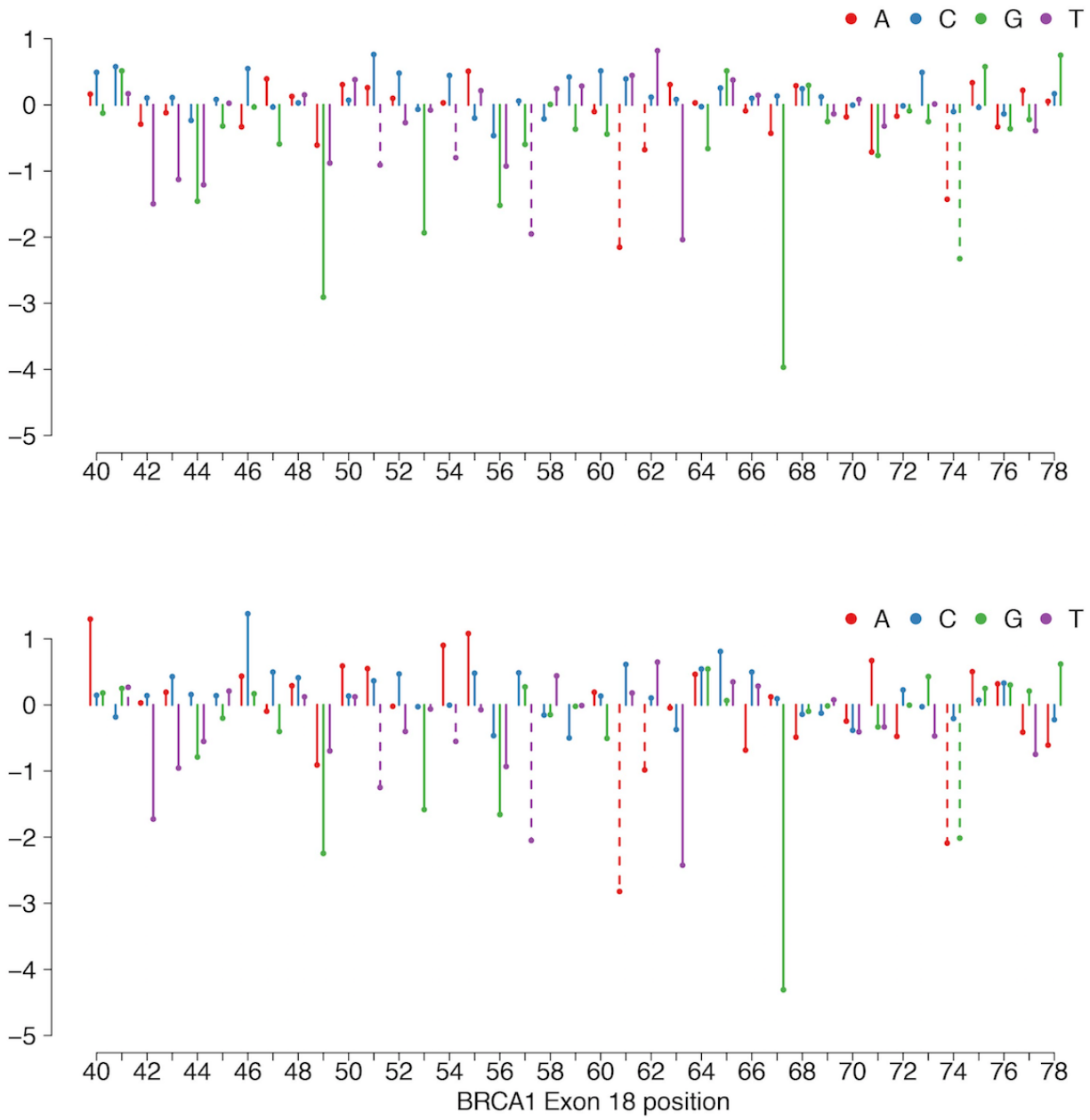



**Figure 2.8. Biological replicate effect size reproducibility for library R.**

Effect sizes of individual variants for libraries R were well correlated between biological replicates. Dashed lines represent SNVs that introduce nonsense codons

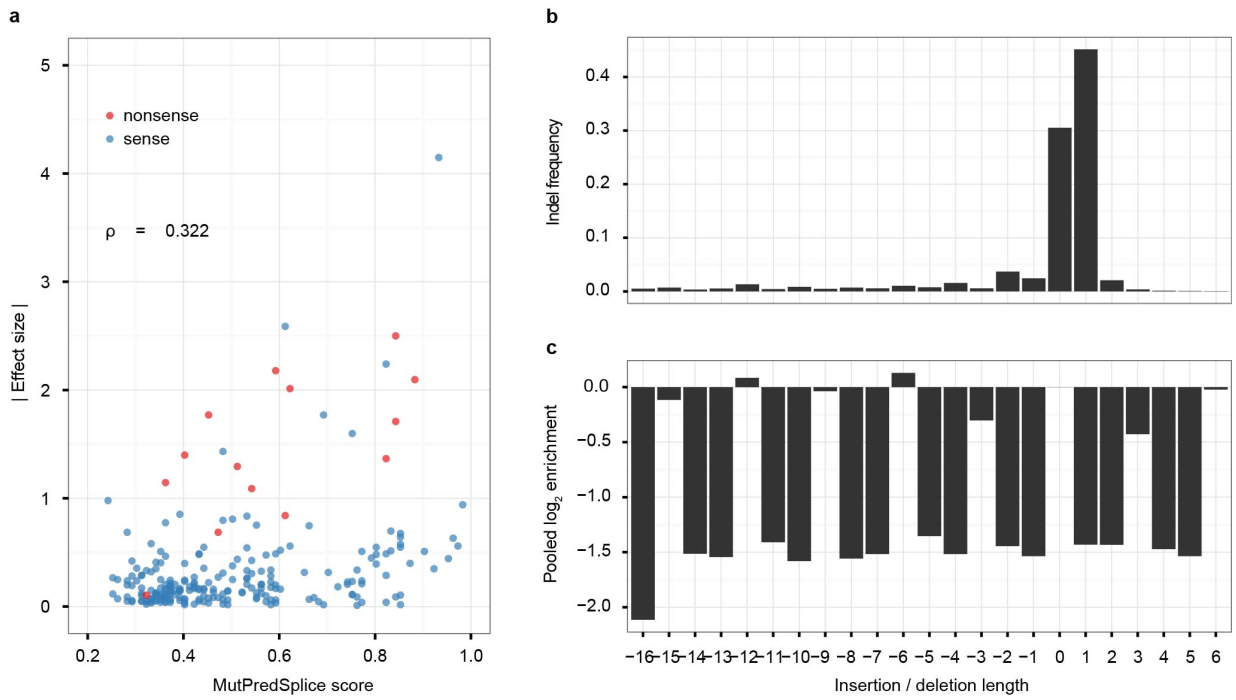



**Figure 2.9. Correlation between effect sizes and predicted disruption of splicing motifs and indel effects.**

**a**, MutPred Splice was used to predict the functional impact of all 234 single nucleotide substitutions on splicing in *BRCA1* exon 18 (x-axis), and these scores were compared to absolute values of our empirically measured effect sizes (y-axis; = 0.322). Although nonsense variants contributed to this trend, the sense variants with the largest effect sizes generally had high MutPred Splice scores. **b**, For indels observed in gDNA from library 2 (virtually all of which occur at the Cas9 cleavage site), size frequencies are plotted. Indel size = 0 includes all haplotypes with wild type length. **c**, For each indel size, enrichment scores were calculated and normalized to that of the average full length exon. As predicted by nonsense-mediated decay, indels that shift the coding frame were associated with low transcript abundance.

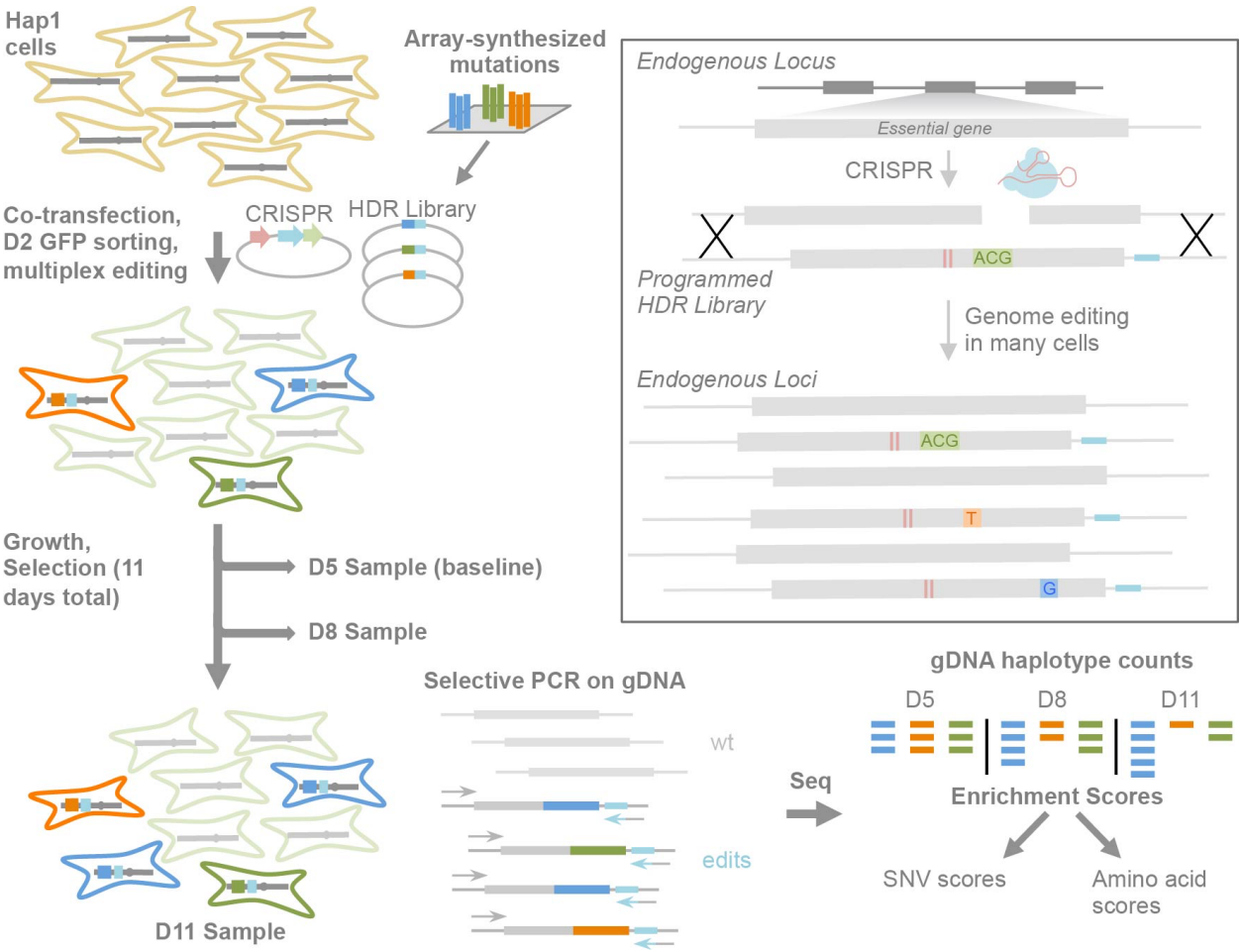



**Figure 2.10. Experimental schematic for saturation genome editing and multiplex functional analysis of *DBR1* exon 2.**

HAP1 cells were co-transfected with a single Cas9-2A-EGFP-sgRNA construct (CRISPR) and an HDR library cloned from array-synthesized oligonucleotides containing programmed SNVs (orange, blue) and active site codon substitutions (green). The HDR library exon haplotypes also included two synonymous mutations (red) to disrupt PAM and protospacer sequences to prevent Cas9 re-cutting, and a 6 bp selective PCR site (light blue) substituted in the downstream intron. Successfully transfected cells (EGFP+) were selected on D2 by FACS, and cultured. On D5, D8, and D11, samples of cells were taken and selective PCR was performed prior to targeted sequencing of gDNA. Each haplotype's enrichment score, a measure of the haplotype's fitness in cell culture, was calculated by dividing D8 or D11 abundance by D5 abundance.

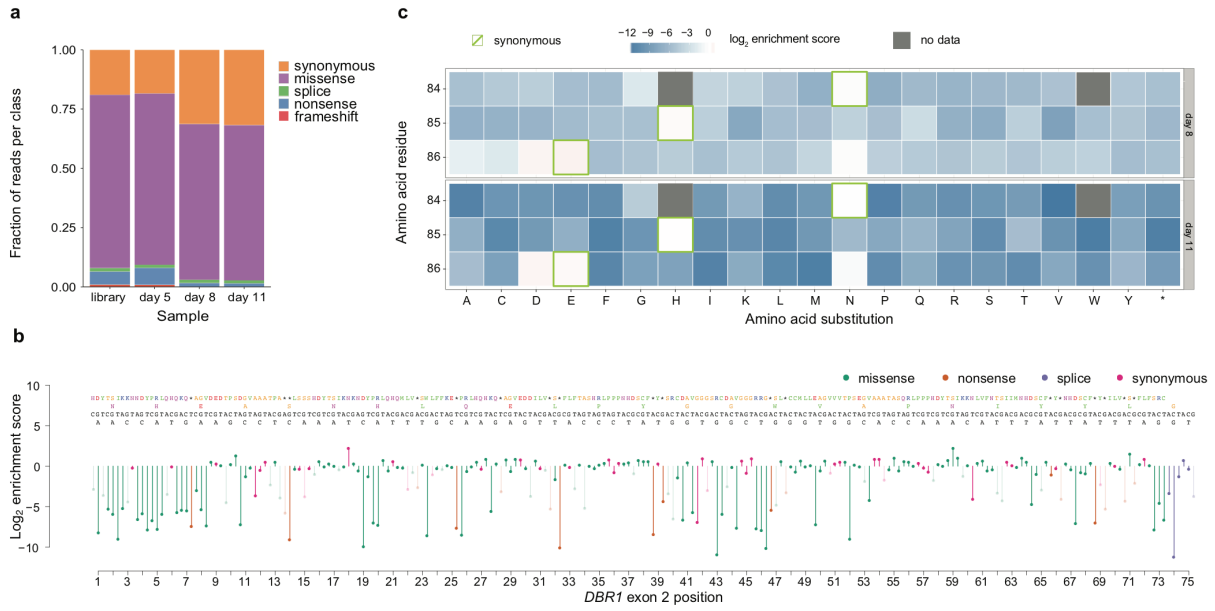



**Figure 2.11. Saturation genome editing and multiplex functional analysis at an essential gene, *DBR1*, in HAP1 cells.**

An HDR library targeting a highly conserved region of *DBR1* exon 2 was used with pCas9-EGFP-sgDbr1x2 to introduce point mutations across 75 bp and all possible codon substitutions at three residues believed to participate at the enzyme's active site. **a**, Sequencing of gDNA from the HDR library and populations of edited cells at D5, D8, and D11 reveals selection for synonymous mutations, and depletion of frameshift, nonsense, and missense variants. **b**, Mean D11 enrichment scores are plotted as line segments for SNVs in the 3'-most 73 bases of exon 2 and two bases of intron 2. Above the enrichment scores in ascending order are the wt nucleotide at each position, each one bp genome edit, the wild-type amino acid (AA), and the AA derived from each genome edit (asterisk indicates a stop codon). Segment color indicates mutation type, faded segments indicate discordant effects between replicates, and AAs are colored according to the Lesk color scheme (small nonpolar – orange, hydrophobic – green, polar – magenta, negatively-charged – red, and positively charged – blue). The first nine bases shown correspond to the active site residues. **c**, D8 and D11 amino acid level enrichment scores were calculated for active site residues N84, H85, E86 after excluding discordant observations between replicates. On both D8 and D11 we observe strong selective effects and tolerance of only synonymous (green boxes) and a few missense variants.

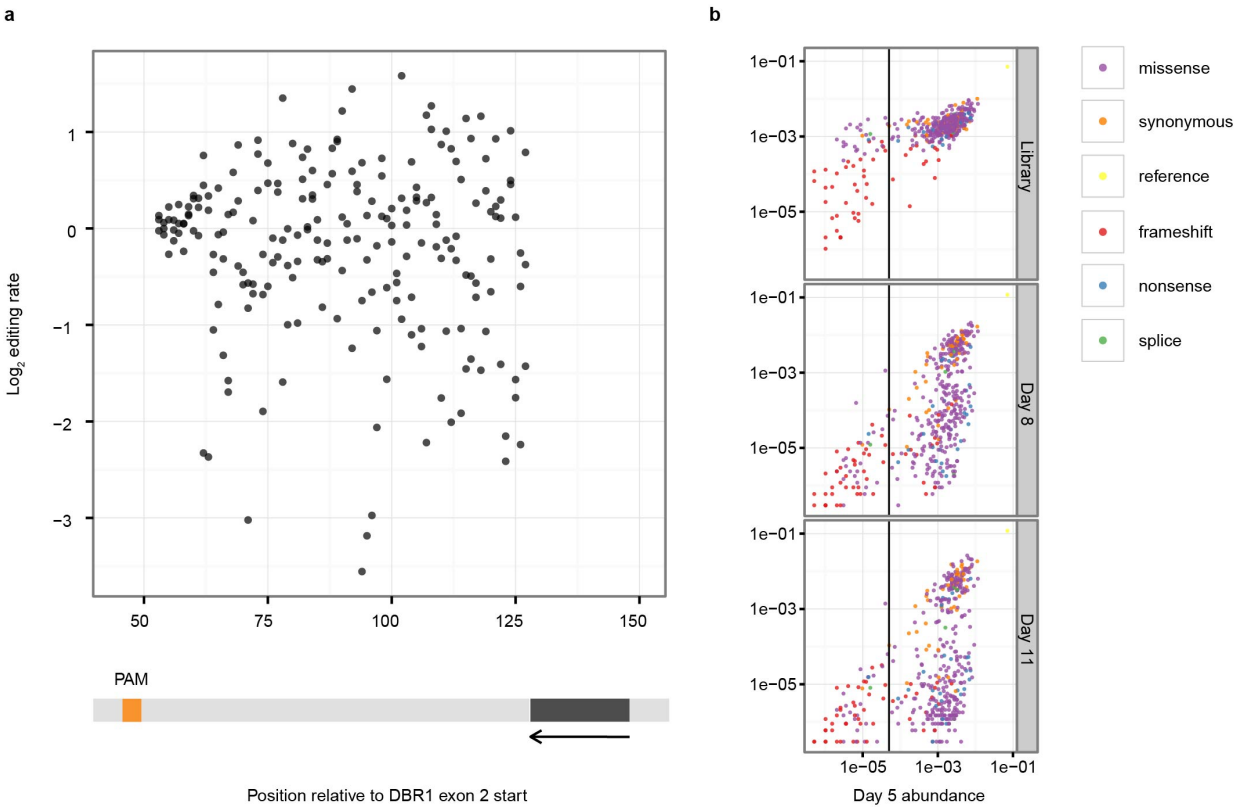



**Figure 2.12. *DBR1* editing rates by position and comparison of haplotype abundances between D5 and the HDR library, D8, and D11.**

**a**, Editing rates for programmed SNVs represented in the *DBR1* gDNA library above threshold (n = 216) were calculated by normalizing each SNV's gDNA abundance by its HDR library abundance. Rates are plotted by position, with the locations of the targeted PAM (orange) and selective PCR site (purple) indicated below. The editing rate did not significantly change with position ($P > 0.05$), consistent with positional effects being negated by eliminating re-cutting and performing selective PCR from a distal site. **b**, Scatterplots display the frequencies at which each haplotype was observed in the D5 sample vs. the HDR library, D8, and D11 samples. To account for bottlenecking from editing of a limited number of cells in this representative experiment, analysis of individual haplotypes was restricted to those present at frequencies above 5E-5 in the D5 sample (n = 377; represented by the vertical line). Selection was evident by the depletion of many haplotypes in D8 and D11 samples.

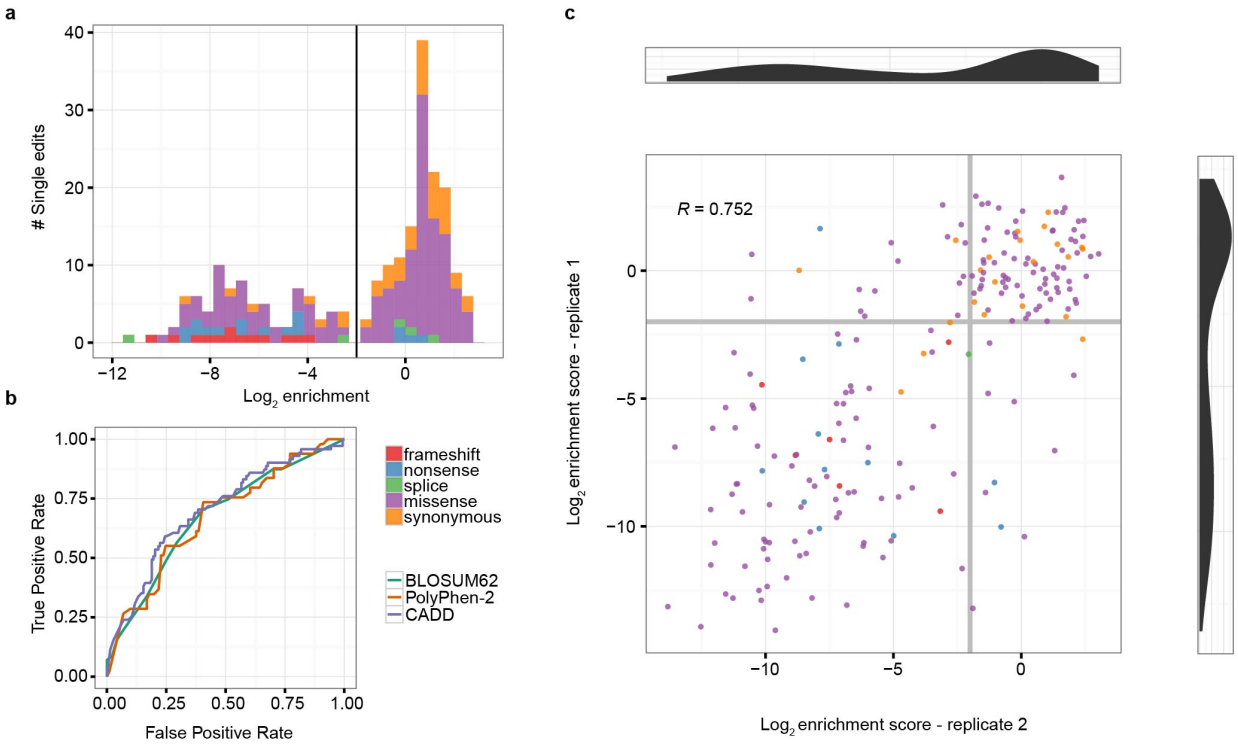



**Figure 2.13. Performance of computational predictions of deleterious *DBR1* mutations and reproducibility between biological replicates.**

**a**, D11 enrichment scores from a single experiment were used to empirically define deleterious mutations as those with scores four-fold below wild type (vertical line). **b**, Three *in silico* metrics of functional impairment were tested for their ability to anticipate the deleteriousness of these mutations as indicated by the area under the receiver operating characteristic curve (AUC): BLOSUM62 (AUC = 0.672, 214 SNVs), PolyPhen-2 (AUC = 0.671, 155 non-synonymous SNVs), and CADD (AUC = 0.701, 214 SNVs). Despite the different approaches of these algorithms, all three exhibited comparably moderate predictive power. **c**, A biological replicate of the *DBR1* experiment was performed and D11 enrichment scores for amino acid substitutions were well correlated (gray lines on scatterplot indicate the "deleteriousness" threshold of four-fold depletion). The distribution of amino-acid level enrichment scores for each experiment is displayed along each axis, reflecting bimodality. Notably, unexpected effects (i.e. nonsense mutations scoring as tolerated) were among the relatively small percentage of effects not consistent between replicates.

# Chapter 3. ACCURATE FUNCTIONAL CLASSIFICATION OF THOUSANDS OF *BRCA1* VARIANTS WITH SATURATION GENOME EDITING

Chapter 3 is adapted from an unpublished manuscript under review as of March, 2018:

> Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Starita, L.M., Shendure, J. (2018). Accurate functional classification of thousands of *BRCA1* variants with saturation genome editing. *Submitted*.

Upon publication, raw data and function scores will be made freely available for nonprofit uses, as well as by nonexclusive license under reasonable terms to commercial entities that have committed to open sharing of *BRCA1* sequence variants. Data and scores are owned by the University of Washington and are not to be used commercially without licensing agreement.

## 3.1 ABSTRACT

Variants of uncertain significance (VUS) fundamentally limit the utility of genetic information in a clinical setting. The challenge of VUS is epitomized by *BRCA1*, a tumor suppressor gene integral to DNA repair and genomic stability. Germline *BRCA1* loss-of-function (LOF) variants predispose women to early-onset breast and ovarian cancers. Although *BRCA1* has been sequenced in millions of women, the risk associated with most newly observed variants cannot be definitively assigned. Data sharing attenuates this problem but it is unlikely to solve it, as most newly observed variants are exceedingly rare. In lieu of genetic evidence, experimental approaches can be used to functionally characterize VUS. However, to date, functional studies of *BRCA1* VUS have been conducted in a *post hoc,* piecemeal fashion. Here we employ saturation genome editing to assay 96.5% of all possible single nucleotide variants (SNVs) in 13 exons that

encode functionally critical domains of BRCA1. Our assay measures cellular fitness in a haploid human cell line whose survival is dependent on intact *BRCA1* function. The resulting function scores for nearly 4,000 SNVs are bimodally distributed and almost perfectly concordant with established assessments of pathogenicity. Sequence-function maps enhanced by parallel measurements of variant effects on mRNA levels reveal mechanisms by which loss-of-function SNVs arise. Hundreds of missense SNVs critical for protein function are identified, as well as dozens of exonic and intronic SNVs that compromise *BRCA1* function by disrupting splicing or transcript stability. We predict that these function scores will be directly useful for the clinical interpretation of cancer risk based on *BRCA1* sequencing. Furthermore, we propose that this paradigm can be extended to overcome the challenge of VUS in other genes in which genetic variation is clinically actionable.

## 3.2 INTRODUCTION

Despite our rapidly advancing knowledge of the genetic underpinnings of human disease, our ability to predict the phenotypic consequences of an arbitrary genetic variant in a human genome remains poor. This problem manifests most poignantly in the large numbers of 'variants of uncertain significance' (VUS) identified in 'clinically actionable' genes, *i.e.* genes that are already etiologically linked with a specific disease, and for which a definitive interpretation of the variant as benign or pathogenic would significantly impact clinical care (Cooper, 2015; Rehm et al., 2015).

The gene that perhaps best highlights the challenge of VUS is *BRCA1*. Germline variants that disrupt *BRCA1* function are associated with a hereditary predisposition to breast and ovarian cancer (Friedman et al., 1994; Hall et al., 1990; Kuchenbaecker et al., 2017; Miki et al., 1994). Functionally disruptive germline variants in *BRCA1* are clinically actionable, *e.g.* by more

aggressive screening or prophylactic surgery, interventions which lead to improved outcomes (Olopade and Artioli, 2004; Rebbeck et al., 2004). Furthermore, functionally disruptive somatic *BRCA1* mutations influence how tumors respond to specific therapeutic agents, *e.g.* PARP inhibitors (Chan and Mok, 2010; Farmer et al., 2005; Hollis et al., 2017). Clinical sequencing of *BRCA1*, as well as many other genes linked to cancer predisposition such as *BRCA2, PALB2, BARD1, ATM*, etc., has the potential to implicate specific variants in disease (Easton et al., 2015). Documented pathogenic *BRCA1* variants in the ClinVar database include complete or partial gene deletions, frameshifting insertions and deletions (indels), nonsense SNVs, missense variants detrimental to protein stability and function, and both intronic and exonic variants that perturb splicing (Landrum et al., 2016). However, as of January 2018, over half of *BRCA1* SNVs in ClinVar are classified as VUS. VUS are typified by rare missense SNVs, but also include variants potentially affecting mRNA production, such as SNVs near splice junctions. Further indicative of the challenge of variant interpretation, ClinVar is replete with *BRCA1* variants that have received conflicting interpretations from different experts. Of 3,936 germline *BRCA1* SNVs currently represented in ClinVar, only 983 are classified by an expert panel as 'benign' or 'pathogenic' without conflicting interpretations.

There are two major approaches for resolving VUS. The first approach, data sharing, relies on the expectation that as *BRCA1* is sequenced in increasing numbers of individuals (Cook-Deegan et al., 2013), the recurrent observation of a specific variant in multiple individuals who either have or have not developed breast and/or ovarian cancer will enable the definitive interpretation of that variant. However, although this may be possible for some variants, given that the vast majority of potential SNVs in *BRCA1* are exceedingly rare (Lek et al., 2016; Yang et al., 2017) and that the phenotype is incompletely penetrant, it may be decades or centuries before sufficient numbers of

humans are included in genotype-phenotype studies to accurately quantify cancer risk for each individual rare variant.

The second approach, functional assessment, has spurred the development of diverse *in vitro* assays for *BRCA1 (Millot et al., 2012)*. As the homology-directed DNA repair (HDR) function of BRCA1 is key for tumor suppression, one commonly used assay involves expressing a BRCA1 variant in cells and assessing the integrity of the cells' HDR pathway via inducing repair of a double strand DNA break in a fluorescent reporter construct (Pierce et al., 2001; Ransburgh et al., 2010). Other approaches include assays for embryonic stem cell viability (Bouwman et al., 2013), cell sensitivity to chemotherapeutic drugs (Bouwman et al., 2013), binding to known partners such as BARD1 (Ransburgh et al., 2010; Starita et al., 2015), and minigene-based splicing assays (de la Hoya et al., 2016; Steffensen et al., 2014). Computational tools can predict variant effects based on features such as amino acid conservation. However, although many such metrics correlate with pathogenicity, at present no computational tool is sufficiently accurate to be used for the clinical interpretation of newly observed *BRCA1* variants in the absence of genetic or experimental evidence (Ghosh et al., 2017; Richards et al., 2015).

Functional assessment of *BRCA1* variants has historically been limited in several ways. Chiefly, experimental studies are *post hoc* and have not kept pace with the scaling of *BRCA1* sequencing and the accumulation of VUS. Additionally, assays that express variants as cDNA-based transgenes removed from their genomic context (Ransburgh et al., 2010; Starita et al., 2015) fail to assess effects on splicing or transcript stability, as well as potential artifacts of overexpression (Gibson et al., 2013). Genome editing technologies provide a means to overcome these challenges. Yet to our knowledge, genome editing has not yet been applied to functionally characterize VUS in *BRCA1* or other genes similarly linked to cancer predisposition.

Here we set out to apply genome editing to measure the functional consequences of all possible SNVs in *BRCA1*, regardless of whether they have been previously observed in a human. Given *BRCA1*'s immense size, this initial study focuses on 13 exons that encode the functionally critical RING and BRCT domains. In each experiment, a single exon is subjected to 'saturation genome editing' (Findlay et al., 2014), wherein all possible SNVs are simultaneously introduced to a haploid human cell line in which *BRCA1* is essential. Consequently, *BRCA1* variants that result in nonfunctional alleles are depleted over time, a selection that is quantified by deep targeted sequencing. We optimized this method to obtain function scores for 3,893 SNVs, comprising 96.5% of all possible SNVs in the targeted exons. These function scores are bimodally distributed and nearly perfectly concordant with expert-based assessments of pathogenicity. We predict that our functional classifications will be of immediate clinical utility, and argue that the scaling of this approach to additional clinically actionable genes will substantially enhance the utility of genetic testing.

## 3.3 RESULTS

### 3.3.1 *Saturation genome editing of BRCA1 exons*

Many genes in the HDR pathway, including those associated with hereditary cancer predisposition such as *BRCA1*, *BRCA2*, *PALB2* and *BARD1 (Easton et al., 2015)*, were recently identified in a gene trap screen as being essential in the human haploid cell line HAP1 (Blomen et al., 2015) (**Figure 3.1a**). To validate this finding, we designed guide RNAs (gRNAs) to target exons of each of these genes and assessed HAP1 cell viability after transfecting each gRNA on a plasmid co-expressing Cas9 and a puromycin resistance cassette (Ran et al., 2013). High cell death was evident by light microscopy (**Figure 3.1b**), and a luminescence-based survival assay established that targeting any of these genes substantially reduces viability of HAP1 cells within

one week (**Figure 3.2**). Deep sequencing of the edited loci of *BRCA1*-targeted cells confirmed that cell death was consequent to mutations, as there was widespread selection against frameshifting indels in favor of unedited loci and some in-frame indels (**Figure 3.1c**). Overall, these results confirm the essentiality of HDR pathway components in HAP1 cells and establish targeted sequencing as a strategy to distinguish functional vs. non-functional *BRCA1* variants in a population of edited HAP1 cells.

We next designed and optimized experiments for saturation genome editing (SGE) (Findlay et al., 2014) (**Figure 3.1d**). We chose to focus on the thirteen exons of *BRCA1* encoding the RING (exons 2-5) and BRCT domains (exons 15-23) because these domains are essential for the protein's role as a tumor suppressor (Drost et al., 2011; Moynahan et al., 1999; Shakya et al., 2011) and harbor missense variants known to be pathogenic or benign, as well as ~400 VUS or variants with conflicting reports of pathogenicity (Easton et al., 2007; Landrum et al., 2016; Vega et al., 2001). To create a library of repair templates, we used array-synthesized oligo pools containing all possible SNVs spanning each exon and ~10 bp of adjacent intronic sequence. Oligo pools for each exon were PCR-amplified and cloned into plasmids with homology arms to mediate genomic integration and make 'SNV libraries'. Each SNV library molecule also included a fixed synonymous substitution at the target site to reduce re-cutting by Cas9 after successful HDR (Findlay et al., 2014). Each SGE experiment targeted a single exon. In brief, a population of 20 million HAP1 cells was co-transfected on day 0 with the exon's corresponding SNV library and Cas9/gRNA plasmid. Successfully transfected cells were selected with puromycin (days 1-4), expanded, and sampled on day 5 and day 11. Variant frequencies were quantified by targeted amplification and sequencing of the edited exon from genomic DNA (gDNA) harvested on day 5

and day 11. Negative controls were used to confirm that PCR amplicons were not derived from the plasmid DNA of the SNV library.

We initially performed SGE experiments in replicate for each exon in wild-type (WT) HAP1 cells. In each of the 13 exons, we observed depletion of frameshifting indels, confirming intolerance to loss of *BRCA1* function (**Figure 3.3**). However, towards achieving more robust data, we optimized SGE in HAP1 cells in two ways. First, to increase HDR rates in HAP1 cells, we generated a monoclonal *LIG4* knockout HAP1 line (HAP1-Lig4KO) (**Figure 3.4a-b**). LIG4 acts in the non-homologous end joining (NHEJ) pathway, and its depletion can increase the proportion of cells with HDR-mediated repair of double-stranded breaks (Beumer et al., 2008; Ma et al., 2016). We observed a median 3.6-fold increase in HDR rates on day 5 in HAP1-Lig4KO relative to WT HAP1 (**Figure 3.5a**). Second, it is known that HAP1 cells can spontaneously revert to diploidy (Essletzbichler et al., 2014a). Simply sorting HAP1 cells for 1N ploidy prior to editing improved reproducibility (**Figure 3.4c-e**).

We next performed optimized SGE experiments for each of the 13 targeted exons in 1N-sorted HAP1-Lig4KO cells, testing nearly every possible SNV per exon in replicate (**Figure 3.5b**). Functional effects of SNVs on survival were determined by targeted DNA sequencing of each SNV library as well as the edited exon in gDNA harvested on day 5 and day 11 (**Figure 3.5c-e**). Additionally, targeted RNA sequencing of day 5 samples was used to determine how abundant exonic SNVs were in *BRCA1* mRNA (**Figure 3.5f**). Because these optimizations resulted in greater reproducibility (**Figure 3.6**), we moved forward with data from the 1N-sorted HAP1-Lig4KO cells only.

### 3.3.2 *Function scores for 3,893 BRCA1 SNVs*

We sought to calculate function scores for each SNV in a way that accurately quantified selection throughout the experiment while also minimizing experimental biases. First, we calculated the log2 ratio of the SNV's frequency on day 11 vs. its frequency in the original plasmid library. Second, positional biases in editing rates were modeled (using day 5 SNV frequencies) and subtracted (**Figure 3.7**). Third, to enable comparisons between exons, we normalized function scores such that each experiment's median synonymous and nonsense SNV matched global medians. Finally, a small number of SNVs were filtered out that could not confidently be scored (*e.g.* SNVs poorly represented on day 5; **Figure 3.8**). Altogether, we obtained function scores for 3,893 SNVs within or immediately intronic to these exons (**Figure 3.5e**). This corresponds to 96.5% of all possible SNVs in these regions.

Function scores for SNVs in these 13 *BRCA1* exons were bimodally distributed (**Figure 3.5g**). All nonsense SNVs scored below –1.25 (N = 138, median = -2.12), whereas 98.7% of synonymous SNVs >3 bp from splice junctions scored above -1.25 (N = 544, median = 0.00). We classified all SNVs as 'functional', 'non-functional', or 'intermediate' by fitting a two-component Gaussian mixture model in which the parameters of the 'non-functional' distribution were based on all nonsense SNVs and the 'functional' distribution based on synonymous SNVs not depleted in RNA (**Figure 3.9**). We then used this model to estimate the probability of each SNV's score being drawn from the non-functional distribution ($P_{nf}$). SNVs with $P_{nf} < 0.01$ were categorized as functional (72.5%); SNVs with $P_{nf} > 0.99$ were categorized as non-functional (21.1%); and SNVs with $0.01 < P_{nf} < 0.99$ (6.4%) were categorized as intermediate.

Rare missense variants in *BRCA1* are particularly challenging to interpret clinically. Of the missense SNVs that we scored here, 21.1% (441 of 2,086) scored as non-functional (**Figure 3.5h**).

Although most of the remaining missense SNVs were functional (70.6%), there was an enrichment for missense SNVs with intermediate effects (8.1%, compared to 4.4% of all other SNVs; Fisher's exact $P = 2.7 \times 10^{-6}$).

An advantage of assaying variants by genome editing is that their impact on native regulatory mechanisms such as RNA splicing can be ascertained (Findlay et al., 2014). Whereas SNVs disrupting canonical splice sites (the two intronic positions immediately flanking each exon) overwhelmingly scored as non-functional (89.5%) or intermediate (5.5%) ('CS' in **Figure 3.5h**). SNVs positioned 1-3 bp into the exon or 3-8 bp into the intron had variable effects. We defined SNVs in these regions that did not alter the amino acid sequence as 'splice region' variants, of which 22.9% were non-functional ('SR' in **Figure 3.5h**), on par with missense SNVs (21.2% non-functional). SNVs positioned more deeply in introns or in the 5' UTR were similar to non-splice-region synonymous SNVs, in that they were much less likely to score as non-functional (intronic: 1.8% non-functional; 5' UTR: 0.0% non-functional; synonymous: 1.3% non-functional).

### 3.3.3 *Function scores are nearly perfectly concordant with ClinVar*

We next asked how well our function scores agreed with expert-based clinical variant interpretations, where available in ClinVar. Of 169 SNVs deemed 'pathogenic' in ClinVar that overlapped with our classifications, 162 were designated 'non-functional', 2 'functional', and the remaining 5 'intermediate'. In contrast, of 22 SNVs deemed 'benign' in ClinVar that overlapped with our classifications, 1 was designated 'non-functional', 1 'intermediate', and 20 'functional' (**Figure 3.10a**). The three SNVs for which our function scores are unambiguously discordant with ClinVar are discussed further below. A ROC curve showed a sensitivity of 96.7% at 98.2% specificity when we treat 'likely pathogenic' and 'likely benign' ClinVar annotations as pathogenic and benign, respectively (**Figure 3.10b**). Importantly, our assay accurately predicts

ClinVar interpretations independent of mutational consequence; sensitivity and specificity are high for both missense and splice site SNVs when these are considered separately from nonsense SNVs (**Figure 3.9f**). We find 64 of 256 (25.0%) VUS and 60 of 122 (49.2%) SNVs with conflicting interpretations to be non-functional in our assay (**Figure 3.10c**). Missense VUS from ClinVar were significantly more likely to score as non-functional compared to missense SNVs absent from ClinVar (25.9% vs. 17.2%, $P = 0.002$). Apart from largely corroborating established ClinVar annotations, our scores also provide functional classifications for an additional 3,140 SNVs, the vast majority of which have yet to be publicly reported in clinical sequencing. Of these SNVs, 498 (15.9%) are classified as non-functional.

We also investigated the relationship between our function scores and SNV frequencies in large-scale databases of human genetic variation. Of 302 assayed SNVs that overlap with the Genome Aggregation Database (gnomAD) (Lek et al., 2016), higher allele frequencies were associated with higher function scores (**Figure 3.10d**). For instance, 33 of 166 (19.9%) of singleton gnomAD variants were non-functional, whereas only 8 of 136 SNVs (5.9%) seen in multiple individuals were non-functional (Fisher's exact $P = 3 \times 10^{-4}$). A similar trend was observed with the Bravo database (**Figure 3.11a**). The FLOSSIES database contains *BRCA1* variants observed in women over seventy years old who have not developed breast or ovarian cancer ([CSL STYLE ERROR: reference with no printed form.]). Of 39 intersecting SNVs, only one scored as non-functional (**Figure 3.11b**). Collectively, these observations show that *BRCA1* SNVs with higher allele frequencies are more likely to be functional, as expected. However, the fact that >70% of ClinVar variants and >95% of non-ClinVar variants that we assayed here have not been observed even once in sequencing of >120,000 humans illustrates the challenges facing observational approaches to variant interpretation.

Several computational metrics are currently used to the assess deleteriousness of variants and often included in genetic testing reports. Although our function scores correlate with metrics such as CADD (Kircher et al., 2014), phyloP (Pollard et al., 2010), and Align-GVGD (Tavtigian et al., 2008), which are largely based on evolutionary conservation and biochemical properties of missense variants, the modesty of these correlations underscores the value of functional assays (**Figure 3.10e,Figure 3.12a-g**). ROC curve analysis restricted to missense variants reveals that SGE-based function scores outperform these metrics at predicting pathogenicity status in ClinVar (**Figure 3.12h-l**). This outperformance is likely underestimated because some of these metrics (*e.g.* Align-GVGD) or their correlates (*e.g.* evolutionary conservation) informed the ClinVar classifications of pathogenicity in the first place.

### 3.3.4 *Mechanisms of BRCA1 loss-of-function*

To gain insights into the various mechanisms by which SNVs compromise function, we performed targeted RNA sequencing of *BRCA1* transcripts from day 5 cells. We normalized SNV frequencies in cDNA to their frequency in gDNA to produce mRNA expression scores ('RNA scores') for 96% of the functionally characterized exonic SNVs. Together with function scores, RNA scores enable fine mapping of molecular consequences of SNVs (**Figure 3.13**). For instance, regions of exons 2 and 15 that respectively code for RING and BRCT domain residues contain numerous loss-of-function missense variants. This contrasts with coding sequence in the same exons that fall outside of the boundaries of these protein domains. Overall, 89% of non-functional missense SNVs did not reduce RNA levels substantially, suggesting that their effects are likely mediated at the protein level (**Figure 3.14a**). Many residues that are sensitive to missense SNVs *not* impacting RNA levels map to buried hydrophobic residues or to the zinc-coordinating loops that are required for proper RING domain folding (**Figure 3.14b-c**). However, 11% of non-

functional missense SNVs are depleted from RNA by 4-fold or more. Many of these SNVs map outside of key protein-protein interfaces and rather in unstructured loops, suggesting that they cause loss-of-function by lowering mRNA expression levels. Consistent with this, the 12 synonymous SNVs classified as non-functional also tended to markedly reduce mRNA levels (median 5.4-fold reduction).

How do these exonic SNVs cause reductions in mRNA levels? Although other mechanisms cannot be ruled out, many of the variants depleted in mRNA are likely impacting RNA splicing. This is evidenced by an overrepresentation of non-functional SNVs near splice junctions, including low scores for many SNVs at terminal G nucleotides of exons (**Figure 3.13**), non-functional exonic SNVs with low mRNA levels that create new acceptor or donor sequences (SNVs annotated with asterisks in **Figure 3.14d**), and the presence of short regions (~6-8 bp) in which many SNVs have moderate-to-strong effects on RNA levels, suggestive of exonic splice enhancers(Desmet et al., 2009) (**Figure 3.14e**). Certain exons appeared particularly prone to harbor non-functional SNVs with low RNA scores. In exon 16, for instance, 46 of 244 SNVs (excluding nonsense) were non-functional (**Figure 3.14e**). Of these, more than half (n = 26) reduced RNA levels by more than 2-fold, and nearly a third (n = 15) by more than 4-fold. In contrast, in exon 19, of 55 of 234 SNVs (excluding nonsense) that were non-functional, none lowered expression by more than 2-fold (**Figure 3.14f**). Exon 19 also completely lacks non-functional SNVs in its flanking intronic regions (apart from the acceptor and donor sites), suggesting the exon is robustly spliced compared to other exons.

### 3.3.5 *Discordances with ClinVar Interpretations*

We leveraged sequence-function maps in reviewing the evidence around the three SNVs for which our classifications were clearly discordant with ClinVar. Discordant SNVs assayed in

our preliminary experiments in WT HAP1 cells had similar scores, suggesting their classifications are not secondary to noise in our assay (**Figure 3.15**). One missense SNV designated 'pathogenic' in ClinVar that we scored as functional, c.5359T>A (C1787S), was identified through segregation with disease. However, in each case, it was seen in *cis* with a second SNV at the neighboring amino acid position (Goldgar et al., 2004). Our data as well as data from other functional assays (Woods et al., 2016) suggest c.5359T>A on its own is functional. The linked SNV c.5363G>T (G1788D), however, scored as non-functional, calling into question the ClinVar annotation (**Figure 3.15c**).

A second disagreement was identified in the exon 2 splice acceptor, c.-19-2A>G. This SNV was annotated as 'pathogenic' in ClinVar based on its occurrence at a splice acceptor site (Spurdle et al., 2012a), rather than from having been associated with disease. Exon 2 contains the *BRCA1* translation initiation codon, meaning that alternate splice forms may preserve the complete open reading frame. Of note, CADD scores for SNVs across the exon 2 acceptor site were much lower than for SNVs in other canonical splice sites (**Figure 3.15d**), and none of the 6 SNVs that we introduced here scored as non-functional. Further supporting that this splice site is not essential for *BRCA1* function, RNA sequencing from breast and ovarian tissue in the GTEx database (GTEx Consortium, 2013) shows this exon junction is poorly represented among *BRCA1* transcripts (**Figure 3.15e**). This suggests that this acceptor site is likely dispensable both in our assay and in tissues relevant to disease, again calling the ClinVar annotation into question.

Exon 16 harbored the third discordantly classified SNV, the 'benign' c.5044G>A (E1682K) variant, which scored as non-functional in our assay. Of note, c.5044G>A resides in a predicted exonic splice enhancer (ESE) (Desmet et al., 2009), and its low function score was substantiated by a reduction in RNA levels of over 90% (**Figure 3.14e**). Neighboring SNVs in

the predicted ESE also reduced RNA expression, corroborating the element's importance. Although this missense SNV is rare (absent from gnomAD and Bravo), reports indicate it was designated as benign based on being observed in *trans* with a variant considered pathogenic (Easton et al., 2007), as biallelic *BRCA1* loss-of-function mutations are thought to be embryonic lethal. The underlying data supporting this finding are not publically available, and previous assays of this variant did not measure splicing consequences (Woods et al., 2016).

## 3.4 DISCUSSION

Here we applied saturation genome editing to the 13 exons that encode functionally critical domains of the cancer risk gene, *BRCA1*, characterizing the functional consequences of nearly 4,000 SNVs in their native genomic context. Specifically, we used CRISPR/Cas9 to introduce hundreds of SNVs per experiment, followed by deep sequencing to measure the functional consequences of each SNV in parallel. Because we measured cell survival, the effects of SNVs on multiple layers of gene function (*e.g.* RNA splicing, translation, protein function, protein stability) are effectively integrated. The approach is validated by nearly perfect concordance of function scores with available evidence for clinical pathogenicity.

Our experimental approach has several caveats. First, the exact requirements for *BRCA1* function essential to maintaining *in vitro* viability and growth of HAP1 cells, as opposed to mediating *in vivo* tumor suppression, are not known. For instance, we cannot rule out, differences in splicing or dosage requirements between our *in vitro* model vs. *in vivo* physiology. Second, we are not currently able to interrogate every possible SNV. Of note, most of the 3.5% of SNVs for which we do not provide function scores were excluded by factors related to genome editing, rather than because of sampling (**Figure 3.8**). Lastly, as these experiments were designed to measure

loss-of-function in a haploid cell line, we are unable to detect all types of functional effects (*e.g.* dominant negative variants).

Notwithstanding these limitations, we achieved nearly comprehensive coverage of the targeted regions and our functional classifications are nearly perfectly concordant with current clinical interpretations. As such, we anticipate that our results will be clinically useful, both for adjudicating hundreds of observed variants whose interpretation is currently ambiguous, as well as for providing immediate functional assessments for variants newly observed. Therefore, the pressing question becomes how to best to integrate this functional data within existing clinical variant classification schemes (Starita et al., 2017).

A benefit of functional data is that measurements are systematically derived, independent of prior expectation (Gasperini et al., 2016). As such, function scores add an additional layer of evidence to support interpretations of variants made through segregation with disease. However, for the large number of VUS for which genetic evidence is insufficient, the predictive power demonstrated here suggests function scores can be used to classify variants with >95% accuracy. As current standards for defining 'likely pathogenic' and 'likely benign' variants accept a comparable level of uncertainty (Plon et al., 2008), we argue that a failure to use appropriately validated functional data to inform clinical care would be a missed opportunity. There is precedent for incorporating functional data in interpretation guidelines (Richards et al., 2015), but the breadth and predictive power demonstrated by SGE calls for an increased role. Indeed, given the low likelihood that observational approaches will ever be sufficient to classify variants not yet seen once in humans, we believe that there is a strong argument to be made for using highly predictive function scores, where available, to inform initial interpretations of newly observed variants.

The orthologous nature of SGE data also presents an opportunity for integration with other data sources. For example, a multiplex reporter assay for HDR activity strengthens the functional evidence presented here for *BRCA1* missense variants. Integration and optimal weighting of experimental and computational approaches may also further improve classification of variants lacking genetic evidence. In cases where evidence is contradictory, functional data may yield specific hypotheses to test. For example, c.5044G>A, for which our data contradicts the ClinVar interpretation (**Figure 3.14e**), would be disambiguated by testing *BRCA1* mRNA levels in individuals harboring this SNV. Similar approaches should be taken to more confidently resolve unlikely functional classifications, such as synonymous SNVs with low function scores and canonical splice SNVs deemed functional. Furthermore, the ~6% of SNVs exhibiting intermediate function scores remain beyond definitive interpretation. The fact that we observe an excess of missense SNVs with intermediate scores suggests that some of these may be hypomorphic *BRCA1* alleles (Domchek et al., 2013; Lovelock et al., 2007; Spurdle et al., 2012b). Further studies will be necessary to quantify the penetrance of intermediately functional variants.

Moving forward, our study provides a blueprint for comprehensive functional analysis of all potential SNVs in clinically actionable genes for which appropriate assays can be developed. Here, we prioritized *BRCA1* exons encoding the RING and BRCT domains, but SGE of the entire coding sequence and promoter are also well motivated. Furthermore, the essentiality of *BRCA2*, *PALB2*, *BARD1*, and *RAD51C* in HAP1 suggests that these genes are assayable by the same method. For genes in other pathways, assays that are compatible with saturation genome editing (*e.g.* drug selection, FACS on phenotypic markers, etc.) may need to be developed and validated. For any gene tested, it is critical that functional measurements be calibrated to clinical evidence of pathogenicity. Given that SGE tests variants in their endogenous genomic context, the scaling of

SGE to many loci promises to improve our understanding of how diverse biological functions are encoded by the genome.

Delivering on the promise of genomic medicine requires that we not only be able to cost-effectively ascertain genetic variation, but also accurately and definitively interpret it. Presently, interpretation is the rate limiting step. As a potential path forward, we show that saturation genome editing is a viable strategy for functionally classifying thousands of variants in a clinically actionable gene, most of which have yet to be observed in a human. With further scaling, we anticipate that this paradigm will substantially improve the utility of genetic information in clinical decision making.

## 3.5 METHODS

### 3.5.1 *HDR pathway essentiality analysis in HAP1 cells*

HAP1 cells were derived from KBM7 cells (a near-haploid immortalized chronic myelogenous leukemia line) by introduction of induced pluripotent stem cell factors (Carette et al., 2011). HAP1 gene essentiality scores were obtained (Blomen et al., 2015) and filtered on genes with greater than 20 mapped gene-trap insertions (N = 14,306). Of 78 HDR genes defined by the GO term 'double-strand break repair via homologous recombination' (GO:0000724), 66 were among the 14,306 genes included in analysis. To rank genes by essentiality, they were first ordered by q-value (low to high) and second by the proportion of gene-trap insertions in the sense orientation (low to high). HDR pathway genes implicated in cancer (labelled in **Figure 3.1**) were defined as those included on the University of Washington BROCA sequencing panel (Walsh et al., 2010).

### 3.5.2 *gRNA design and cloning*

All CRISPR gRNAs used in SGE and essentiality experiments were cloned into pX459 (Ran et al., 2013). This plasmid expresses the gRNA from a U6 promoter, as well as a Cas9-2A-puromycin resistance (-puroR) cassette. *S. pyogenes* Cas9 target sites were chosen for SGE experiments on multiple criteria, assessed in the following order: 1.) To induce cleavage within *BRCA1* coding sequence, 2.) To target a genomic site permissive to synonymous substitution within the guanine dinucleotide of the PAM or the protospacer, 3.) To have minimal predicted off-target activity (Hsu et al., 2013), 4.) To have maximal predicted on-target activity (Doench et al., 2016).

Complementary oligos ordered from Integrated DNA Technologies (IDT) were annealed, phosphorylated, diluted and ligated into BbsI-digested and gel-purified pX459, as described (Ran et al., 2013). Ligation reactions were transformed into *E. coli* (Stellar competent cells, Takara), which were plated on ampicillin. Colonies were cultured and Sanger sequenced to confirm correct gRNA sequences. Purification of sequence-verified plasmids for transfection was performed with the ZymoPure Maxiprep kit (ZymoResearch). For targeting *LIG4* in HAP1 cells, pX458 (Ran et al., 2013) was used instead of pX459, which expresses EGFP in lieu of puroR.

### 3.5.3 *HDR library design and cloning*

Array-synthesized oligos were designed as follows for each saturation genome editing region (*i.e.* a *BRCA1* exon). The sequence to be mutated (~100bp) was obtained from the human genome (hg19) and a synonymous substitution was introduced at the chosen Cas9 target site (*e.g.* a substitution at the PAM site). This 'fixed' substitution in the library was included in design to serve multiple purposes: 1.) plasmid library molecules harboring the substitution are predicted to be cleaved less frequently by Cas9:gRNA complexes, 2.) SNVs introduced to cells are predicted

to be depleted via Cas9 re-cutting less frequently as a consequence of the fixed substitution, and 3.) sequencing reads can be filtered on the fixed substitution to distinguish true SNVs introduced via HDR from sequencing errors. A second synonymous substitution at an alternative CRISPR target site was introduced to the sequence as well, such that each exon's SNV library would be compatible with multiple gRNAs. Next, a sequence was created for every possible single nucleotide substitution on this template (**Table 3.1**). For all sequences, adapters were added to both ends to enable PCR amplification from the oligo pool. For each SGE region, the total number of oligos designed was three times the length of the region, plus the oligo template without any SNV (*e.g.* for a 100 bp SGE region, 301 total oligos were designed).

Pooled oligos were synthesized (Agilent Technologies). Primers designed to amplify the subset of oligos corresponding to a single exon's region were used to perform PCR with Kapa HiFi Hot-start Ready Mix ('Kapa HiFi', Kapa Biosystems). PCR products were purified with Ampure beads (Agencourt) to be used in subsequent library cloning reactions.

Homology arms were cloned into pUC19 by PCR-amplifying (Kapa HiFi) regions surrounding each targeted exon from HAP1 gDNA. Primers for these reactions were designed such that homology arms would be between 600 and 1,000 bp on both sides of the targeted region. Adapters homologous to pUC19 were added to primers to facilitate NEBuilder HiFi Assembly cloning (NEB) into a linearized pUC19 vector. Cloning reactions were transformed into Stellar competent cells and selected with ampicillin. Plasmid DNA was isolated from colonies (Qiagen MiniPrep kit) and sequence-verified.

To make the HDR library, homology arm plasmids were linearized via PCR using primers that conferred 15-20 bp of terminal overlap with the adapter sequences flanking each PCR-amplified oligo pool. This sequence overlap enabled cloning via the NEBuilder HiFi Assembly

Cloning Kit (NEB). Cloning reactions were transformed into Stellar competent cells, and a small proportion (1%) of the transformation was plated on ampicillin-containing plates to assess efficiency. All remaining transformed cells were grown directly in 100 ml of media with ampicillin for 16-18 hours, and plasmid DNA from the culture was isolated (ZymoPure Maxiprep kit) to produce each final HDR library.

### 3.5.4 *HAP1 cell culture*

Quality-controlled WT HAP1 cells were purchased (Haplogen/Horizon Discovery) and cultured in media comprising Iscove's Modified Dulbecco's Medium (IMDM) with L-glutamine and 25 mM HEPES (GIBCO) supplemented with 10% fetal bovine serum (Rocky Mountain Biologicals) and 1% penicillin-streptomycin (GIBCO). Cells were grown on plates at 37C with 5% CO2, and passaged prior to becoming confluent. For routine passaging, cells were washed once with 1x phosphate buffered saline (PBS, Gibco), trypsinized with 0.25% trypsin with EDTA (Gibco), resuspended in media, centrifuged for 5 min at 300 rcf, and then resuspended and plated.

A monoclonal *LIG4* knock-out HAP1 line (HAP1-Lig4KO) was generated by transfecting a plasmid expressing a Cas9-2A-GFP cassette and a gRNA targeting the human *LIG4* coding sequence (gRNA sequence: 5'-GCATAATGTCACTACAGATC) into WT HAP1 cells. Single GFP-expressing HAP1 cells were sorted into wells of a 96-well plate and cultured. After two weeks, gDNA was harvested and Sanger sequencing was performed to assess *LIG4* editing. A clone with a 4bp deletion was identified and expanded further for use in saturation genome editing experiments.

HAP1 cells can spontaneously revert to a diploid state in cell culture. Therefore, to sort a 1N-enriched population of cells prior to transfection, cells were stained for DNA content with Hoechst 34580 (BD Biosciences) at 5 μg/ml media for 1h at 37C. FACS was performed to isolate

1-2 x $10^6$ cells from the lowest intensity Hoechst peak, corresponding to 1N ploidy. These cells were expanded for seven days prior to transfection.

### 3.5.5 *Transfection of HAP1 cells*

For all experiments, HAP1 cells were transfected using TurboFectin 8.0 (Origene) according to manufacturer's protocol. A 2.5x volume of Turbofectin was added to the transfection mix for each µg of plasmid DNA in Opti-Mem (Life Technologies). For each SGE transfection, 10 million cells were passaged to a 10 cm dish. The next day (day 0), cells were co-transfected with 12 µg of the Cas9/gRNA plasmid (pX459) and 3 µg of the SGE library corresponding to a single exon. For negative control transfections, a pX459 vector targeting *HPRT1* was used instead. On day 1, cells were passaged into media supplemented with puromycin (1 µg/ml) to select for successfully transfected cells. On day 4, cells were washed twice and passaged to 6 cm plates in regular media.

Cell populations were sampled on day 5 and day 11 for all SGE experiments. On day 5, half of the cells were pelleted and frozen and the other half passaged. The cells were passaged on day 8 into 15 cm dishes and then harvested on day 11. Negative control transfections were harvested on day 5.

For the luminescence-based viability assay, HAP1 cells were plated at ~35-40% confluency in a 6-well dish (approximately 1.2 million cells per well per target) then transfected with 1.5 µg Cas9/gRNA plasmid targeting coding exons of HDR genes or controls the following day. 24 hours after transfection the cells were plated in time-point triplicates at 20,000 cells per well in 96-well clear bottom plates in media with and without puromycin. Cells without puromycin were assessed 4 hours after plating to establish baseline absorbance for each target. Cell survival

was assessed at day 2, day 5, and day 7 post-transfection using the CellTiterGlow reagent (Promega, 1:10 dilution of suggested reagent). Luminescence at 135 nm absorbance was measured using a Synergy plate reader (Biotek Instruments).

### 3.5.6 *Nucleic acid sampling and sequencing library production*

For obtaining WT HAP1 genomic DNA for cloning homology arms and for genotyping the HAP1-Lig4KO cell line, DNA was isolated using the DNeasy kit (Qiagen). For each SGE experiment, DNA and total RNA were purified using the AllPrep kit (Qiagen). DNA samples were quantified with the Qubit dsDNA Broad Range kit (Thermo Fisher) and RNA samples by UV spectrometry (Nanodrop). PCR primers for genomic DNA were designed such that one primer would anneal outside of the homology arm sequence, thereby selecting for amplicons derived from gDNA and not plasmid DNA. PCR conditions were optimized using gradient qPCR on WT HAP1 gDNA.

All gDNA harvested from the population of day 5 cells was sampled by performing many PCR reactions in parallel on a 96-well plate, using 250 ng of gDNA per 50 µl reaction such that all day 5 gDNA was used in PCR (Kapa HiFi). At least as many PCR reactions were performed for day 11 samples (which yielded more gDNA) to ensure adequate sampling. PCRs were performed for the minimal number of cycles needed to complete amplification, with cycling conditions as specified in the Kapa HiFi protocol. An additional PCR was performed using day 5 gDNA from negative control transfections for each exon.

After PCR, multiple wells of amplicons from the same sample were pooled and purified using Ampure beads. Next, a nested qPCR was performed using the first reaction as template to produce a smaller amplicon with custom sequencing adapters (‘PU1L’ and ‘PU1R’), which was likewise purified with Ampure beads. The SGE libraries were also PCR-amplified at this step,

starting from 50 ng of plasmid DNA. Lastly, a final qPCR was performed using purified products from the second reaction as template to add dual sample indexes and flow cell adapters.

RNA was sampled from day 5 HAP1-Lig4KO cells (AllPrep, Qiagen). Reverse transcription followed by RNase H treatment was performed on all RNA harvested or a maximum of 5 μg per sample (Superscript IV Kit, Life Technologies). This reaction was primed with a gene-specific primer complementary to the 3' UTR in exon 23 of *BRCA1*. Primers were designed for each exon to amplify across exon junctions, and reaction conditions were optimized using gradient PCR. cDNA was distributed into 5 equal PCR reactions, which were run on a qPCR machine and then pooled in equal ratios. Flow cell adapters and sample indexes were added in an additional reaction (as for gDNA samples).

All sequencing libraries were purified with Ampure beads, quantified with the Qubit dsDNA High Sensitivity kit (Life Technologies), diluted and denatured for sequencing in accordance with protocols for the Illumina NextSeq or MiSeq machines.

### 3.5.7 *Sequencing and data analysis*

Sequencing was performed on an Illumina NextSeq or MiSeq instrument, allocating about 3 million reads to each gDNA and cDNA sample, 1 million reads for each HDR library, and 500,000 reads for each negative control sample. gDNA samples for individual exons were sequenced on the same run. 300 cycle kits were used, with 150 cycles for read 1 and read 2 each, and 19 cycles for dual index reads. Custom sequencing primers and indexing primers are available upon request. Illumina PhiX control DNA was added to each sequencing run (~10% MiSeq, ~30-40% NextSeq) to improve base calling.

Illumina's bcl2fastq 2.16 was used to call bases and perform sample demultiplexing and fastqc 0.11.3 was run on all samples to assess sequencing quality. SeqPrep was used with the

following parameters to perform adapter trimming and to merge perfectly matched overlapping read pairs:

'-A GGTTTGGAGCGAGATTGATAAAGT -B CTGAGCTCTCTCACAGCCATTTAG -M 0.1 -m 0.001 -q 20 -o 20'.

Merged reads containing 'N' bases were removed. Reads from cDNA samples were removed if they contained indels or did not perfectly match transcript sequence flanking each targeted exon. Remaining cDNA reads were processed to match genomic DNA amplicons by removing flanking exonic sequence and replacing it with the exon's corresponding intronic sequence. All reads were then aligned to reference gDNA amplicons for each exon using the needleall command in the EMBOSS 6.4.0 package with the following parameters:

'-gapopen 10 -gapextend 0.5 -aformat sam'.

Reads not aligning to the reference amplicon (alignment score < 300) were removed from analysis. To analyze indels, unique cigar counts were quantified from day 5 and day 11 samples using a custom Python script. Reads were classified as HDR events for rate calculations if the programmed edit or edits to the PAM or protospacer (HDR marker edits) were observed in the alignment. Variants without identifiable markers of HDR were not used. Abundances of SNVs were quantified only from aligned reads that had no other mismatches or indels, except for the HDR markers. SNV reads with only the cut-site proximal HDR marker were summed with reads that had both HDR markers to get total abundances for each SNV in each sample, to which a pseudocount of 1 was added to all variants present in either the library, day 5 or day 11 sample. Frequencies for each SNV were calculated as SNV reads over total reads. SNV measurements from WT HAP1 cells and HAP1-Lig4KO cells were processed separately at all steps.

### 3.5.8 *Modeling positional biases of library integration*

Positional biases in editing rates were modeled for each SNV by using a LOESS regression to fit the log2 day 5 over library ratios as a function of chromosomal position. To avoid modeling biological effects instead of positional effects, the model was fit only on the subset of SNVs that were not substantially depleted between any two timepoints in the experiment (*i.e.* SNVs with day 5 over library ratios > 0.5 and day 11 over day 5 ratios > 0.8.). The regression was performed for each exon replicate, using the 'loess' function in R with span = 0.15. Each model was extended flatly outward to include any positions not fit (a total of 22 nucleotides of sequence on the edges of the edited regions). We subtracted each SNV's positional fit (*e.g.* the model's output) from the SNV's log2 day 11 over library ratio to get position-adjusted ratios for each SNV.

### 3.5.9 *Normalizing scores within and across exons*

Position-adjusted log2 day 11 over library ratios were normalized first across exon replicates, and then across all exons assayed. Scores from within each replicate were linearly scaled such that the median synonymous and median nonsense SNVs within the replicate were set to the median synonymous and median nonsense SNV values averaged across replicate experiments. The ensuing SNV scores for each replicate were then normalized across exons in the same way by again using median synonymous and median nonsense SNVs.

### 3.5.10 *SNV functional class assignment*

Function scores were averaged across replicates and a mixture model was used to estimate the probability that each SNV's score was drawn from the non-functional distribution of scores. The non-functional distribution was defined as nonsense SNVs across all exons. The functional distribution was defined as exonic synonymous SNVs not within 3 bp of splice junctions and with

RNA scores within 1 standard deviation of the median synonymous SNV. This definition does not fully guarantee that these SNVs have no functional consequence. The means and variances of the 'non-functional' and 'functional' groups were fixed and a model was fit using the normalmixEM function of the mixtools package in R, with starting component proportions set to 0.5. The posterior probabilities generated from the model were used as point estimates of the probability of drawing each SNVs score from the non-functional distribution ($P_{nf}$). Functional classifications were made by setting thresholds for $P_{nf}$ as follows: $P_{nf} > 0.99$ = 'non-functional', $0.01 < P_{nf} < 0.99$ = 'intermediate', $P_{nf} < 0.01$ = 'functional'.

Independent of mixture modelling, ROC curves were used to assess performance of SGE data and other metrics' ability to predict assigned ClinVar classifications. These analyses were performed with the plotROC package in R, and Youden's J-statistic was calculated (sensitivity plus specificity minus 1) to determine optimal values reported in text.

### 3.5.11 *Variant filtering*

A small minority of SNVs that could not be accurately scored were removed from analysis. If a SNV was not present in the HDR library at a frequency over 1 in $10^4$, it was presumed to have been lost in oligo synthesis or cloning and was removed. Additionally, if a SNV was not observed with complete HDR markers at a frequency over over 1 in $10^5$ in day 5 genomic DNA samples from both replicate experiments, it was removed. SNVs introduced near the CRISPR recognition site have the potential to facilitate Cas9 re-cutting of the locus (*e.g.* by replacing the PAM edit or introducing an alternative PAM site). Because these SNVs are likely to score lower consequent to Cas9 editing biases and not their effects on gene function, SNVs were filtered that created increased potential for re-cutting as follows: When an HDR marker mutation used to disrupt editing occurred at position 2 of the PAM (*e.g.* 'NGG' to 'NCG'), SNVs that replaced this marker

with an alternate base were removed to prevent biases introduced by re-cutting non-canonical *S. pyogenes* Cas9 PAMs (*e.g.* 'NAG', 'NTG'). Additionally, variants that created a new PAM 1 bp 3' of the mutated PAM were excluded due to the potential for re-cutting (*e.g.* unedited PAM: 5'-NGGA, edited PAM with HDR marker: 5'-NCGA, filtered out **SNV** that creates *new PAM* +1bp 3': 5'-NCG***G***). (**Figure 3.8** describes re-cutting observed at alternative PAMs.) To prevent misinterpretation, we also removed SNVs that created amino acid changes specific to the context of the library's fixed edits (*e.g.* if in the unedited background, the SNV causes an X to Y change, but with a fixed edit in the same codon, the SNV causes an X to Z change). We also applied this logic to remove SNVs that introduced splice donor sites only in the context of the edited PAM, and SNVs that create splice donor sites in the unedited context but not in the context of the edited PAM.

The RNA scores for exon 18 samples were neither well correlated across replicates nor with SNV abundances in genomic DNA, indicating likely bottlenecking in library preparation. Therefore, RNA data from exon 18 was excluded. WT HAP1 function scores from exon 22 were excluded because there was an unusually high correlation between SNV frequencies sampled from the plasmid library and from day 5 gDNA, suggesting plasmid contamination in gDNA sequencing. This problem was fixed by designing a new primer to prepare gDNA sequencing samples from HAP1-Lig4KO cells.

### 3.5.12 *External data sources and software*

Variant annotations were downloaded from CADD (Kircher et al., 2014) version 1.3 (http://cadd.gs.washington.edu/download). This included the following scores: mammalian phyloP, Grantham deviation, SIFT, Polyphen-2, and CADD. Align-GVGD scores were obtained by running the Align-GVGD program on BRCA1 sequences conserved to sea urchin. ClinVar data

were downloaded on 1/2/2018 for all germline SNVs with at least a 1-star annotation. SNVs annotated as 'Benign/Likely benign' were grouped with 'Likely benign' SNVs and SNVs classified 'Pathogenic/Likely pathogenic' were grouped with 'Likely pathogenic' SNVs. SNV allele frequencies were obtained from http://gnomad.broadinstitute.org/ on 12/26/2017 for gnomAD (Lek et al., 2016), from https://bravo.sph.umich.edu/freeze5/hg38/ on 11/19/2017 for Bravo, and from https://whi.color.com/ on 10/9/2017 for FLOSSIES data. Transcript data was obtained from GTEx on 1/3/2018. Throughout this study, *BRCA1* exons, coding nucleotide positions, and amino acid positions are referenced by the ClinVar transcript annotation for *BRCA1*, transcript NM_007294.3 (NCBI).

All statistical tests described were performed as two-tailed tests using the R software package. Custom scripts for analyzing sequencing data were written in Python and R.

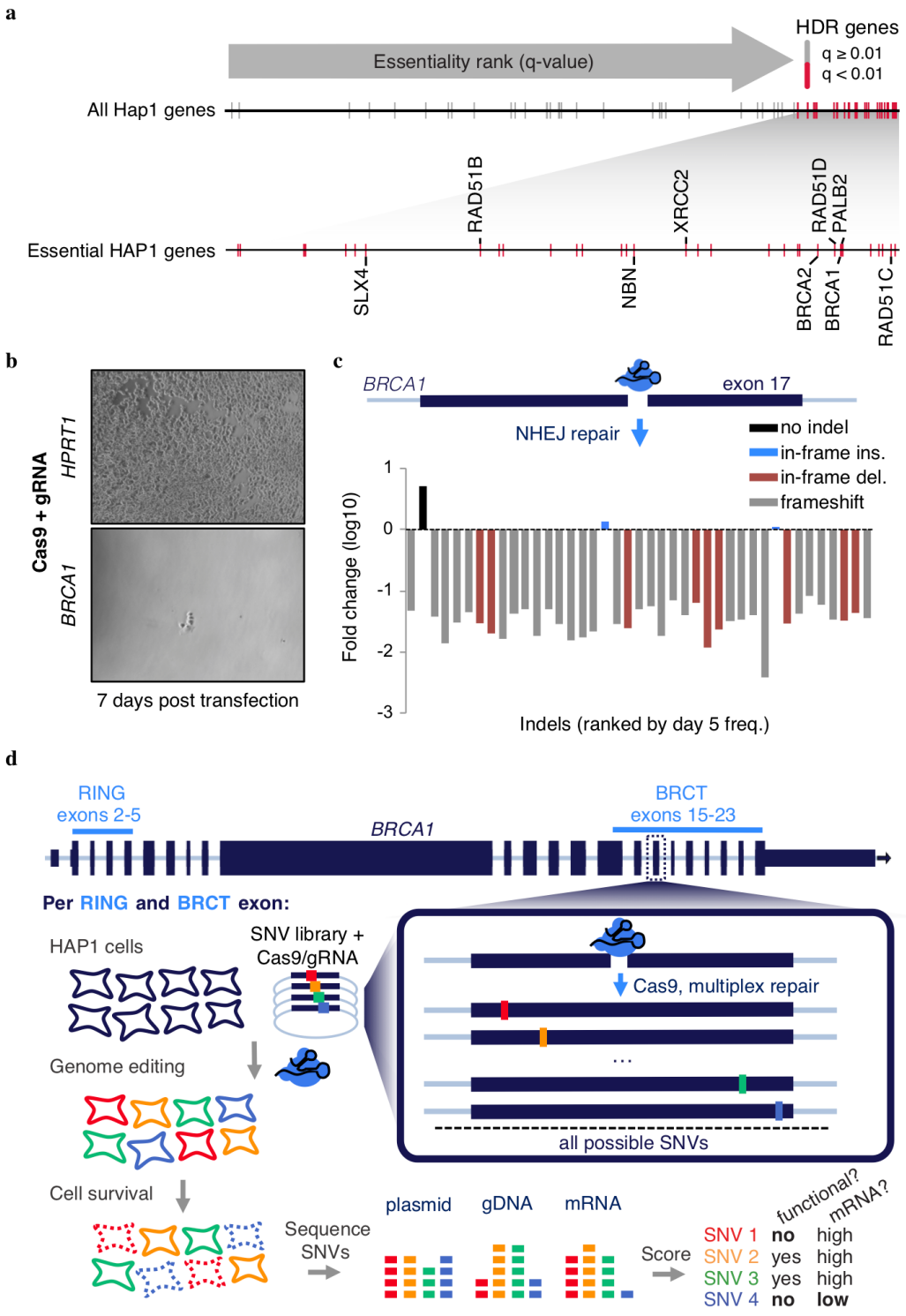



**Figure 3.1. *BRCA1* and other HDR pathway genes are essential in HAP1 cells.**

**a**, The q-value rankings of HDR pathway genes (N = 66, defined by Gene Ontology) among 14,306 genes scored in a HAP1 gene trap screen for essentiality (Blomen et al., 2015) are indicated with

tick marks. Essential HDR genes are colored red and those implicated in cancer predisposition are labelled in the enlargement below. Of the 66 HDR pathway genes scored, 34 including *BRCA1* were 'essential', a 3.4-fold enrichment compared to non-HDR genes (Fisher's exact test $P = 6.1 \times 10^{-12}$). **b**, HAP1 cell populations were transfected with a Cas9/gRNA plasmid either targeting the non-essential gene *HPRT1* (control) or exon 17 of *BRCA1* on day 0. Successfully transfected cells were selected with puromycin (days 1-4) and cultured until day 7, at which point cells were washed prior to imaging. Images are representative of two transfection replicates. **c**, The targeted *BRCA1* exon 17 locus was deeply sequenced from a population of transfected cells sampled on day 5 and day 11. The fold-change from day 5 to day 11 for each editing outcome observed at a frequency over 0.001 in day 5 sequencing reads is plotted. All alleles but indel-free sequences and two in-frame insertions were depleted. **d**, Saturation genome editing experiments were designed to introduce all possible SNVs across thirteen *BRCA1* exons encoding the protein's RING (exons 2-5) and BRCT domains (exons 15-23). For each exon, a Cas9/gRNA construct was designed to be transfected with a library of plasmids containing all SNVs across ~100 bp of genomic sequence (the 'SNV library'). SNV libraries were designed to saturate a total of 1,345 bp of genomic sequence, spanning BRCT and RING domain coding regions and adjacent intronic sequences. SNV library plasmids contain homology arms to mediate genomic integration, as well as fixed synonymous variants within the CRISPR target site to prevent Cas9 re-cutting. Upon HAP1 cell transfection of each Cas9/gRNA plasmid / SNV library pair, successfully edited cells harbor a single *BRCA1* SNV from the library. Cells are sampled 5 and 11 days after transfection and targeted gDNA and RNA sequencing is performed to quantify SNV abundances. SNVs compromising *BRCA1* function are selected against, manifesting in reduced gDNA representation, and SNVs impacting mRNA production are depleted in RNA samples relative to gDNA.

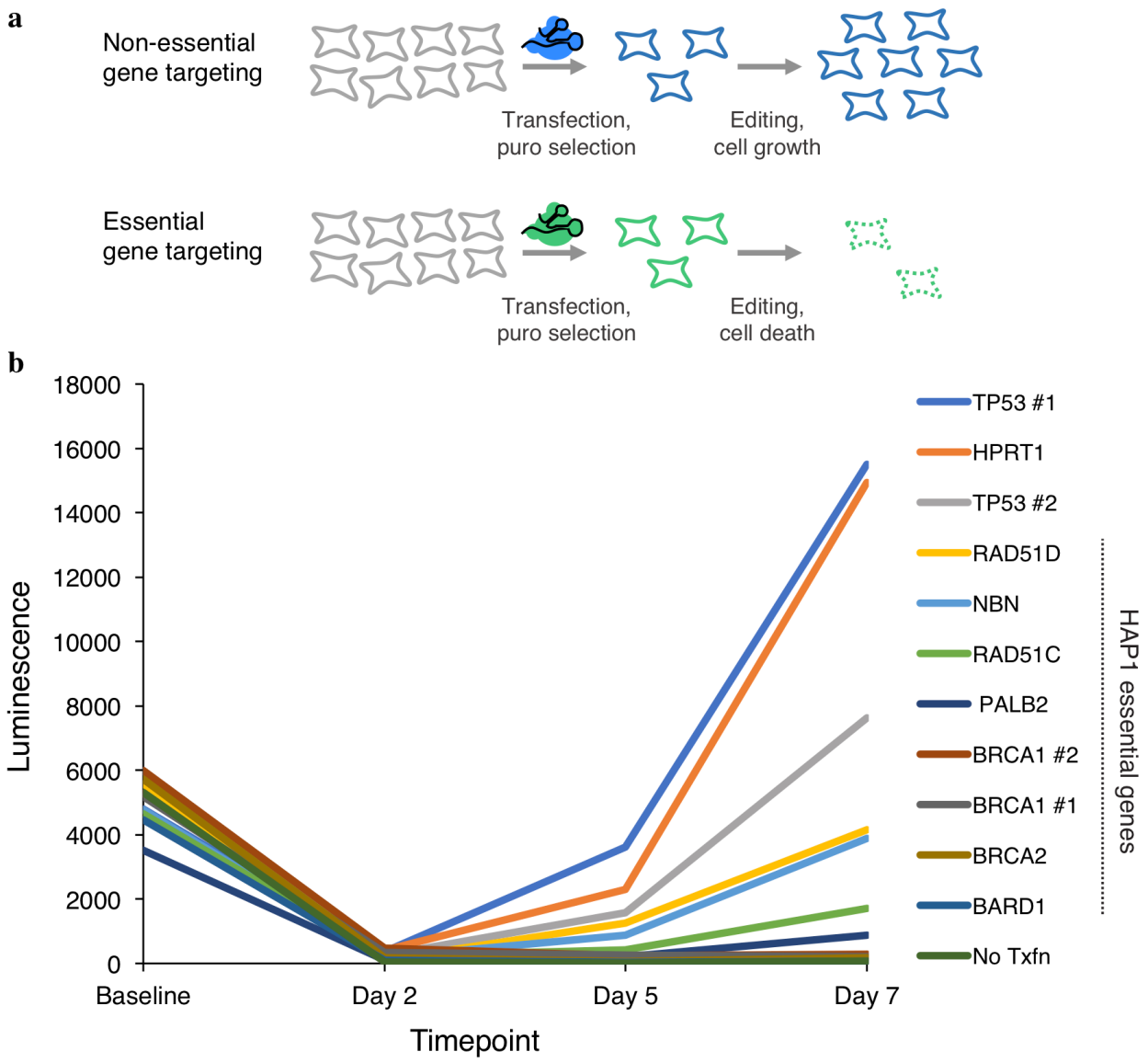



**Figure 3.2. CRISPR targeting of HDR pathway genes to confirm essentiality in HAP1 cells.**

**a**, Schematic; HAP1 cells are transfected with a plasmid expressing a gRNA and a Cas9-2A-puromycin cassette (Ran et al., 2013) . Due to low transfection rates for HAP1 cells, puromycin selection reduces viable cells in all transfections. Over time, however, CRISPR targeting of non-essential genes leads to increased cell growth compared to CRISPR targeting of essential genes. **b**, Cell viability of HAP1 cells transfected with Cas9/gRNA constructs targeting different HDR genes and controls (*HPRT1*, *TP53*) was measured using the CellTiterGlow assay. Luminescence is proportional to the number of living cells in each well when the assay is performed. Triplicate wells for each gRNA at each time point were processed, quantified on a plate reader and averaged.

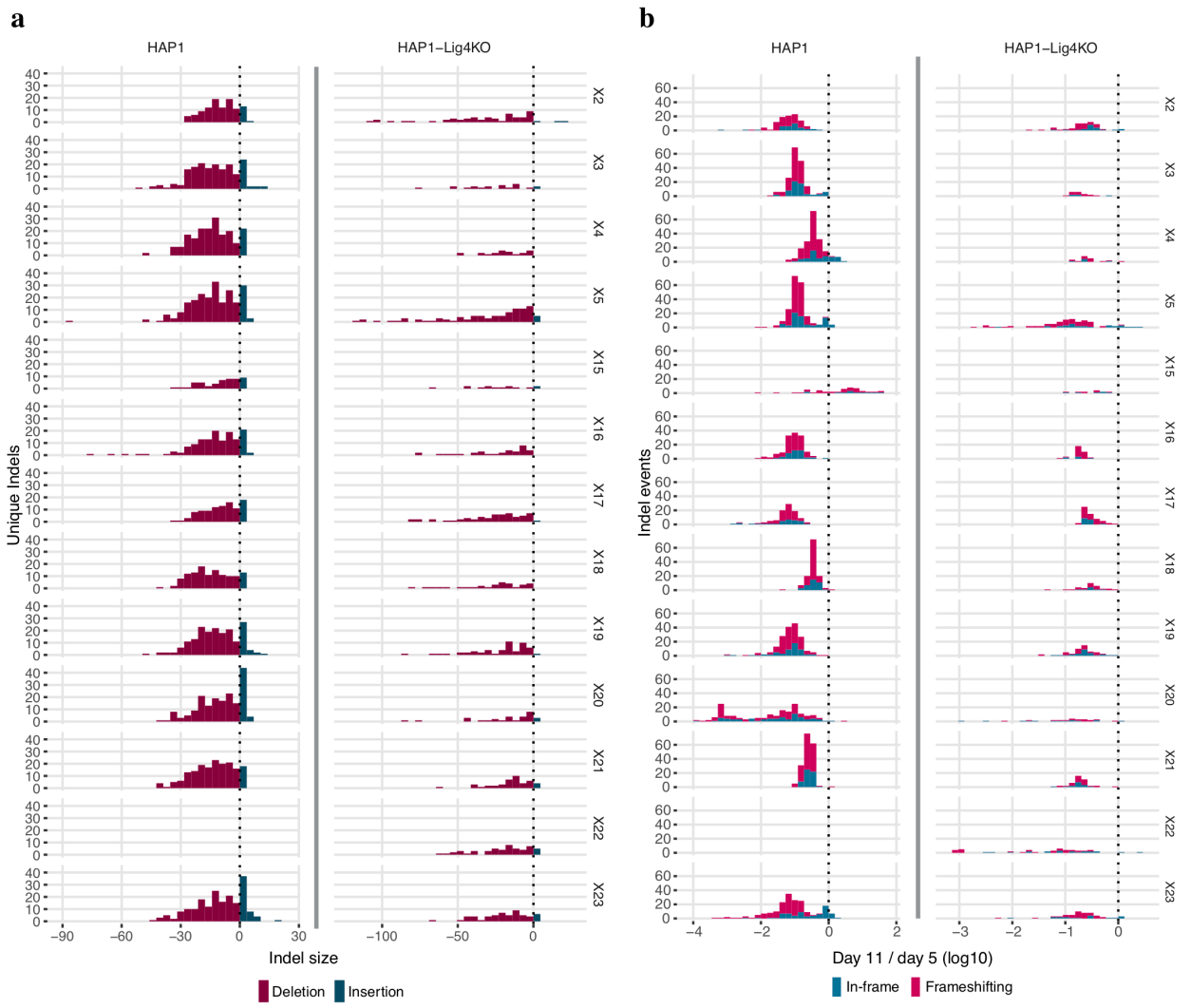



**Figure 3.3. Analysis of Cas9-induced indels observed in *BRCA1* SGE experiments.**

Variants observed in gDNA sequencing were included in this analysis if i) they aligned to the reference with either a single insertion or deletion within 15 bp of the predicted Cas9 cleavage site and ii) were observed at a frequency greater than 1 in 10,000 reads in both replicates. **a**, Histograms show the number of unique indels observed of each size, with negative sizes corresponding to deletions. More unique indels were observed in WT HAP1 cells compared to HAP1-Lig4KO cells for exons compared (WT data for exon 22 was excluded). **b**, Day 11 over day 5 indel frequencies were normalized to the median synonymous SNV in each replicate and then averaged across replicates to measure selection on each indel. The distribution of selective effects is shown for each experiment as a histogram, in which indels are colored by whether their size was divisible by 3 (*i.e.* 'in-frame' vs. 'frameshifting'). Whereas frameshifting variants were consistently depleted, some exons were tolerant to in-frame indels.

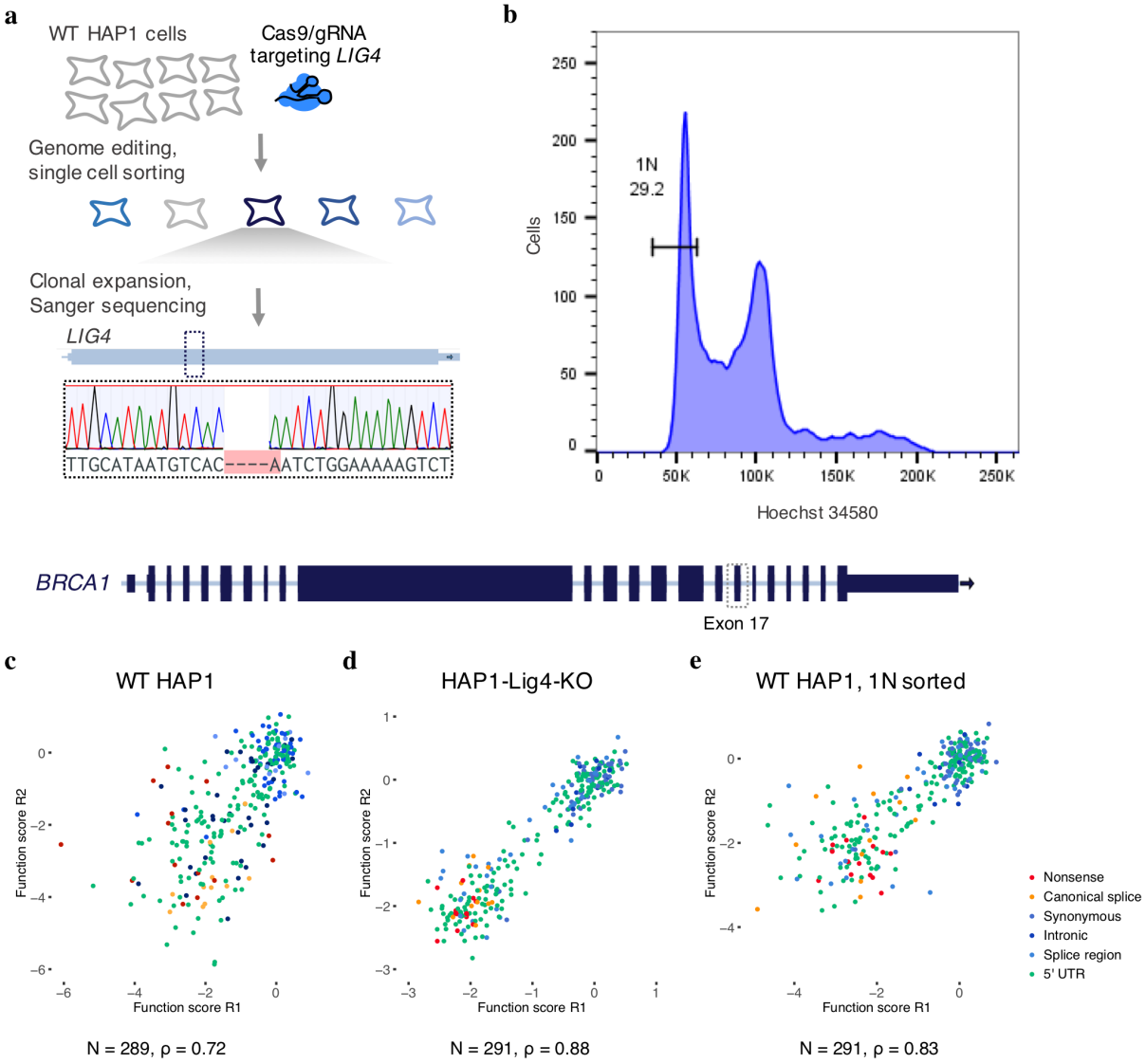



**Figure 3.4. HAP1 cell line optimizations for saturation genome editing to assay essential genes.**

**a**, A gRNA targeting Cas9 to the coding sequence of *LIG4*, a gene integral to the non-homologous end-joining pathway, was cloned into a vector co-expressing Cas9-2A-GFP (Ran et al., 2013). WT HAP1 cells were transfected, and single GFP-expressing cells were sorted into wells of a 96-well plate. Eight monoclonal lines were grown out over a period of three weeks and screened using Sanger sequencing for frameshifting indels in *LIG4*. The Sanger trace shows the frameshifting deletion present in the clonal line chosen for subsequent experiments, referred to as 'HAP1-Lig4KO'. **b**, To purify HAP1 cells for haploid cells, live cells were stained for DNA content with Hoechst 34580 and sorted using a gate to select cells with the lowest DNA content, corresponding to 1N cells in G1. **c-e**, Plots comparing SNV function scores across replicate experiments for exon 17 saturation genome editing experiments performed in unsorted WT HAP1 cells (**c**), HAP1-Lig4KO cells (**d**), and WT HAP1 cells sorted on 1N ploidy (**e**). Both *LIG4* knockout and 1N-sorting improved replicate correlations.

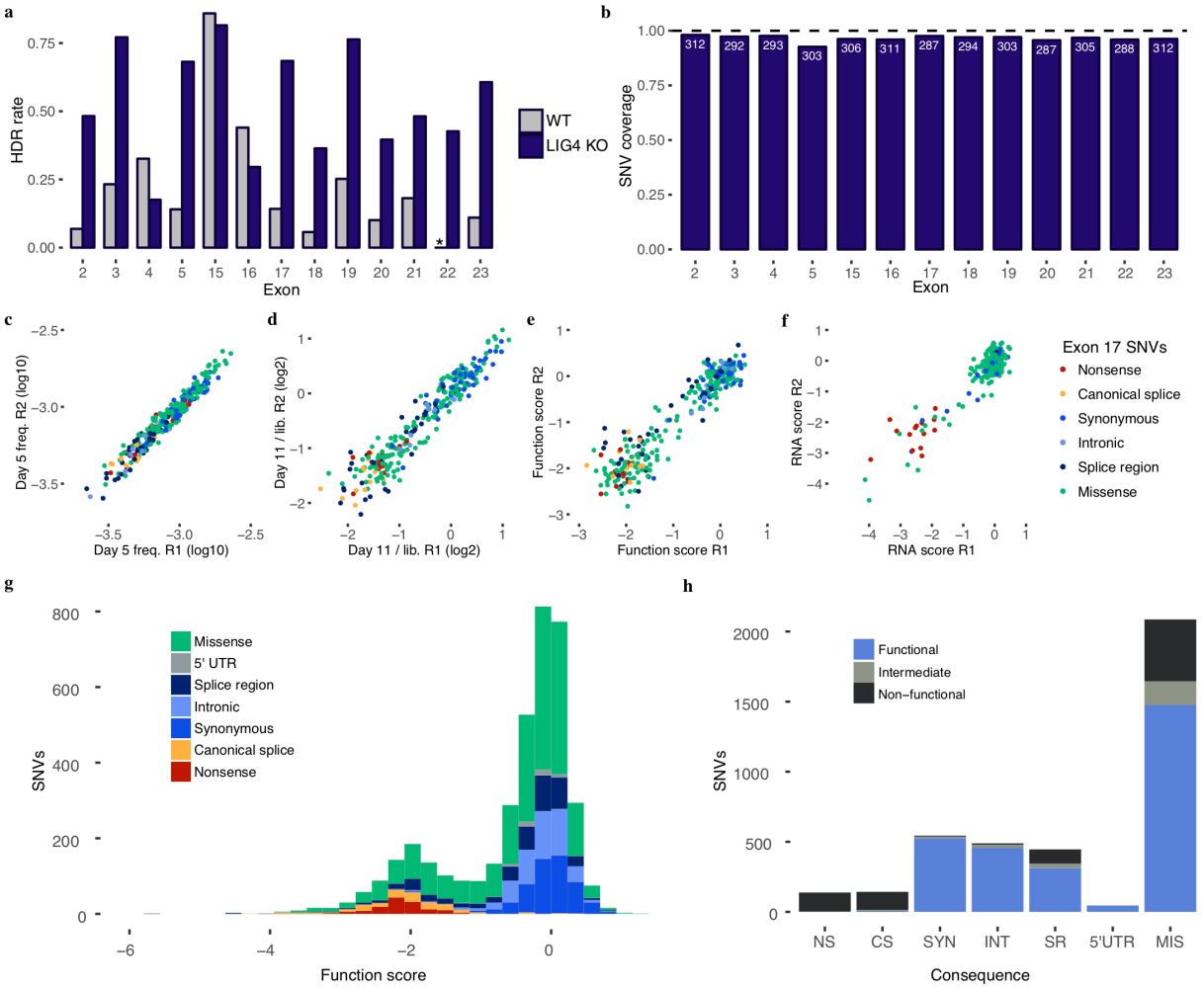



**Figure 3.5. Saturation genome editing enables functional classification of 3,893 *BRCA1* SNVs.**

**a**, HDR editing rates were calculated for each exon as the fraction of day 5 reads containing the SNV library's fixed synonymous variant (*i.e.* an 'HDR marker' edit). The average of two WT HAP1 replicates and two HAP1-Lig4KO replicates is plotted for comparison. (Asterisk denotes missing exon 22 data.) **b**, The fraction of all possible SNVs scored is shown for each exon. SNVs were excluded mainly due to proximity to the HDR marker and/or poor sampling (**Figure 3.8** and **Methods**). **c-f**, Reproducibility was assessed across all exon replicates (**Figure 3.6**). Measurements for exon 17 SNVs assayed in HAP1-Lig4KO cells are plotted to show correlations of day 5 frequencies (**c**, ρ = 0.97), day 11 over library ratios (**d**, ρ = 0.95), function scores (**e**, ρ = 0.88), and RNA expression scores (**f**, ρ = 0.61). **g**, A histogram of 3,893 SNV function scores (averaged across replicates and normalized across exons) shows how each category of mutation compares to the overall distribution. **h**, The number of SNVs within each category of mutation is plotted and colored by functional classification determined by SGE. (NS = nonsense, CS = canonical splice, SYN = synonymous, INT = intronic, SR = splice region, 5'UTR = 5' untranslated region, MIS = missense.

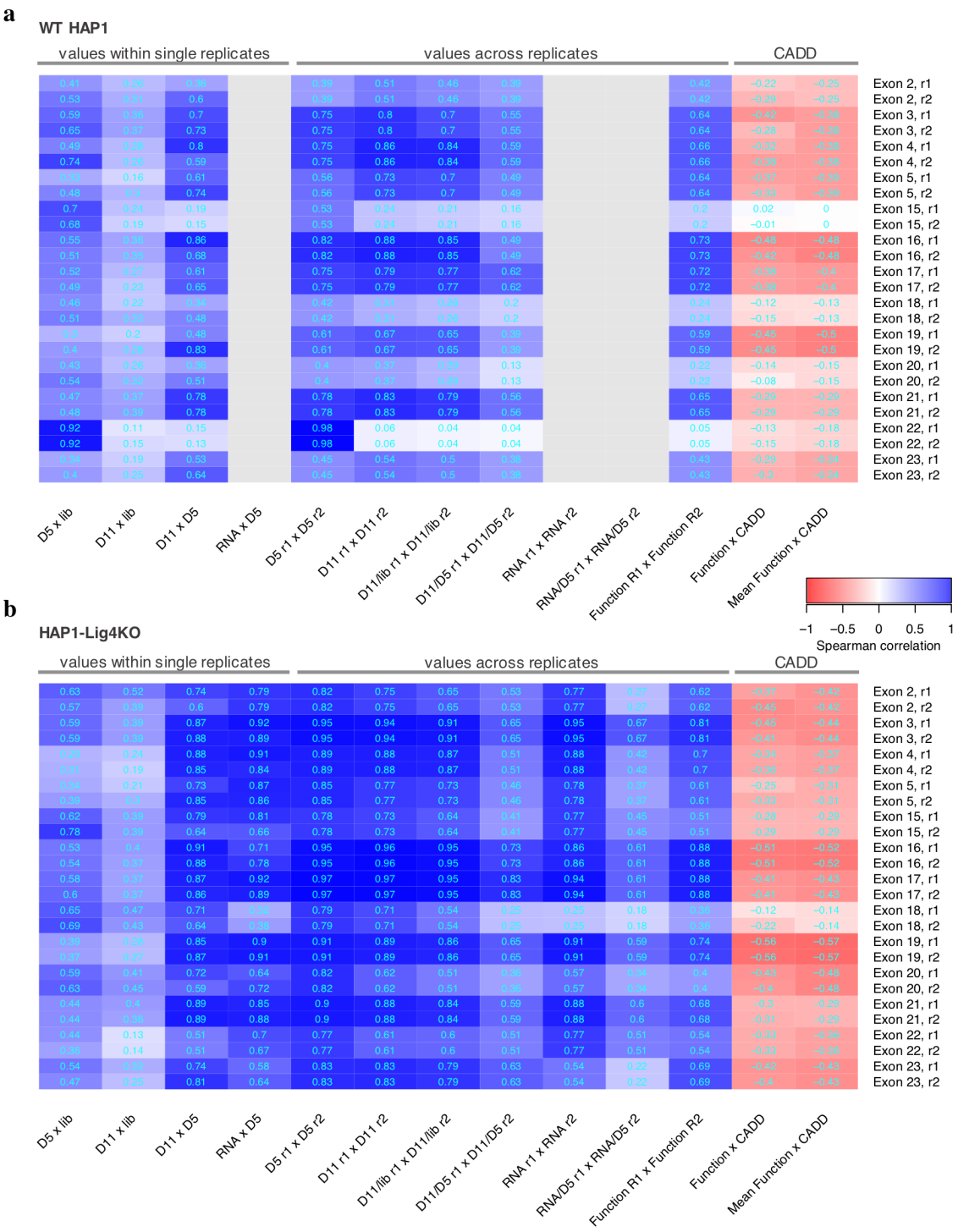



**Figure 3.6. Correlations for SNV measurements within single experiments, across transfection replicates, and to CADD scores for all SGE experiments.**

Heatmaps indicate Spearman correlation coefficients for SNV measurements from experiments in WT HAP1 cells (**a**) and in HAP1-Lig4KO cells (**b**). Gray boxes indicate absent RNA data from

WT HAP1 cells. The four leftmost columns show how SNV frequencies correlate between samples from within a single replicate experiment. The unusually high correlations between exon 22 SNV frequencies in the plasmid library and in day 5 gDNA samples from WT HAP1 cells suggests plasmid contamination in gDNA. Indeed, primer homology to a repetitive element in the exon 22 library was identified. Consequently, the WT HAP1 exon 22 data was removed from analysis and a different primer specific to gDNA was used to prepare exon 22 sequencing amplicons from HAP1-Lig4KO cells. The low HAP1-Lig4KO correlations between exon 18 SNV frequencies in day 5 gDNA and RNA and between RNA replicates suggests RNA sample bottlenecking consequential to low RNA yields. Therefore, exon 18 RNA was also excluded from analysis. Consistent with the higher rates of HDR-mediated genome editing (**Figure 3.5a**), replicate correlations (middle columns) were generally higher in HAP1-Lig4KO cells than WT HAP1 cells. CADD scores predict the deleteriousness of each SNV, and are therefore negatively correlated with function scores (rightmost columns).

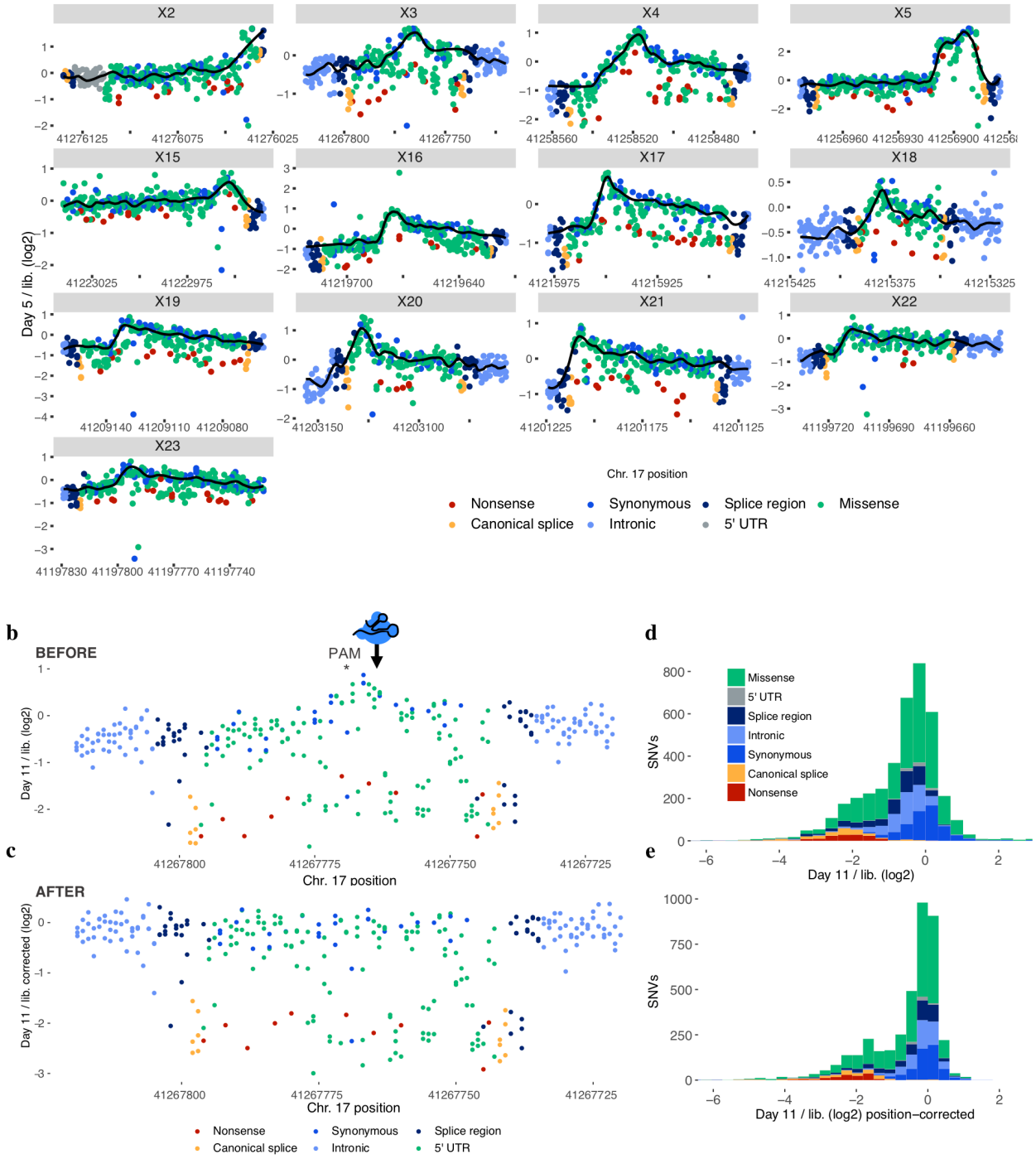



**Figure 3.7. Models of SNV editing rates across *BRCA1* exons account for positional biases.**

**a**, Gene conversion tracts arising during HDR in human cells are short such that library SNVs are introduced to the genome more frequently near the CRISPR target site. We modelled this positional effect in our data using a LOESS regression fit on day 5 over library SNV ratios. Plots shown here are of the average of two replicate experiments per exon, with the black line indicating

the LOESS regression. By day 5 sampling, selective effects on gene function are evidenced by nonsense SNVs (red) appearing at lower frequencies compared to neighbouring SNVs. Therefore, to best approximate the SNV editing rate as a function of position alone (*i.e.* the 'baseline'), the regression excluded SNVs that were selected against between day 11 and day 5 (see Methods). **b**,**c**, Day 11 over library SNV ratios were adjusted by the positional fit for each experiment in calculating function scores. This adjustment is illustrated here for an exon 3 replicate by plotting the ratio as a function of position before (**b**) and after (**c**) adjustment. The elevated day 11 over library ratios for SNVs near the CRISPR target site are corrected to achieve a more uniform baseline across the mutagenized region. **d**,**e**, The distributions of SNV day 11 over library ratios before and after accounting for positional effects are shown, colored by mutational consequence (pre-filtering, N = 4,002).

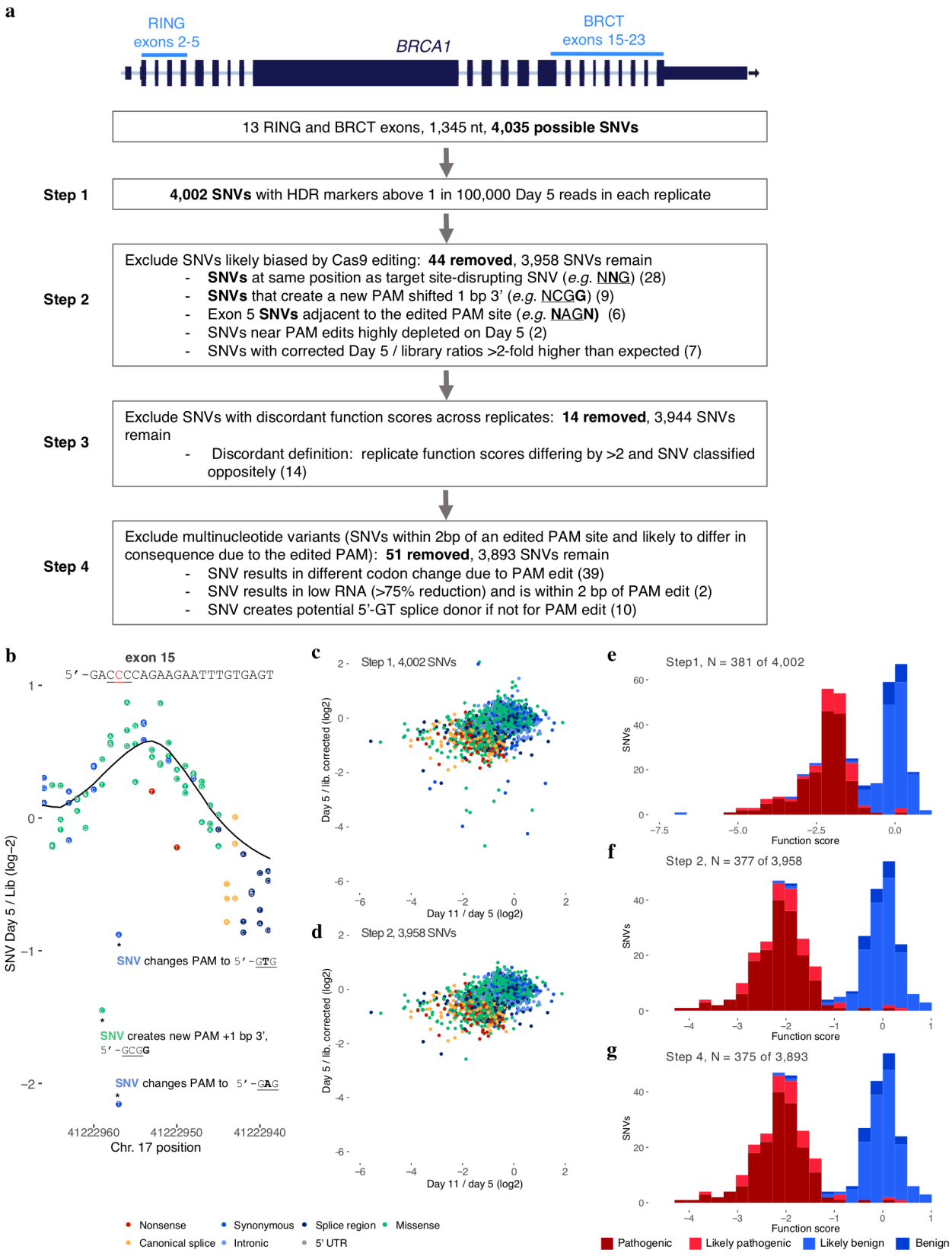



**Figure 3.8. SNV filtering to prevent erroneous functional classification.**

**a**, The flow chart describes filters used to produce the final SNV data set and shows how many SNVs were removed at each step. **b**, Raw day 5 over library SNV ratios are shown for a portion of exon 15 to illustrate how re-editing biases necessitate filtering. The three depleted SNVs marked with asterisks create alternative PAM sequences that likely allow the Cas9:gRNA complex to re-cut the locus and cause their removal. For other SNVs, the fixed PAM edit (a GGG to GCG synonymous change) minimalizes re-editing. The location of the target PAM is underlined and each indicated SNV is bolded in the annotations. The LOESS regression curve in shown in black. **c**,**d**, Plots show the relationship between day 5 over library and day 11 over day 5 ratios before (**c**) and after (**d**) filtering steps 1 and 2. Filtering removes outliers because editing biases primarily affect the day 5 over library ratio. **e-g**, Histograms show the distributions of function scores for SNVs deemed 'pathogenic' or 'benign' in ClinVar at different stages of filtering. Scores in **e** are derived prior to normalization across exons.

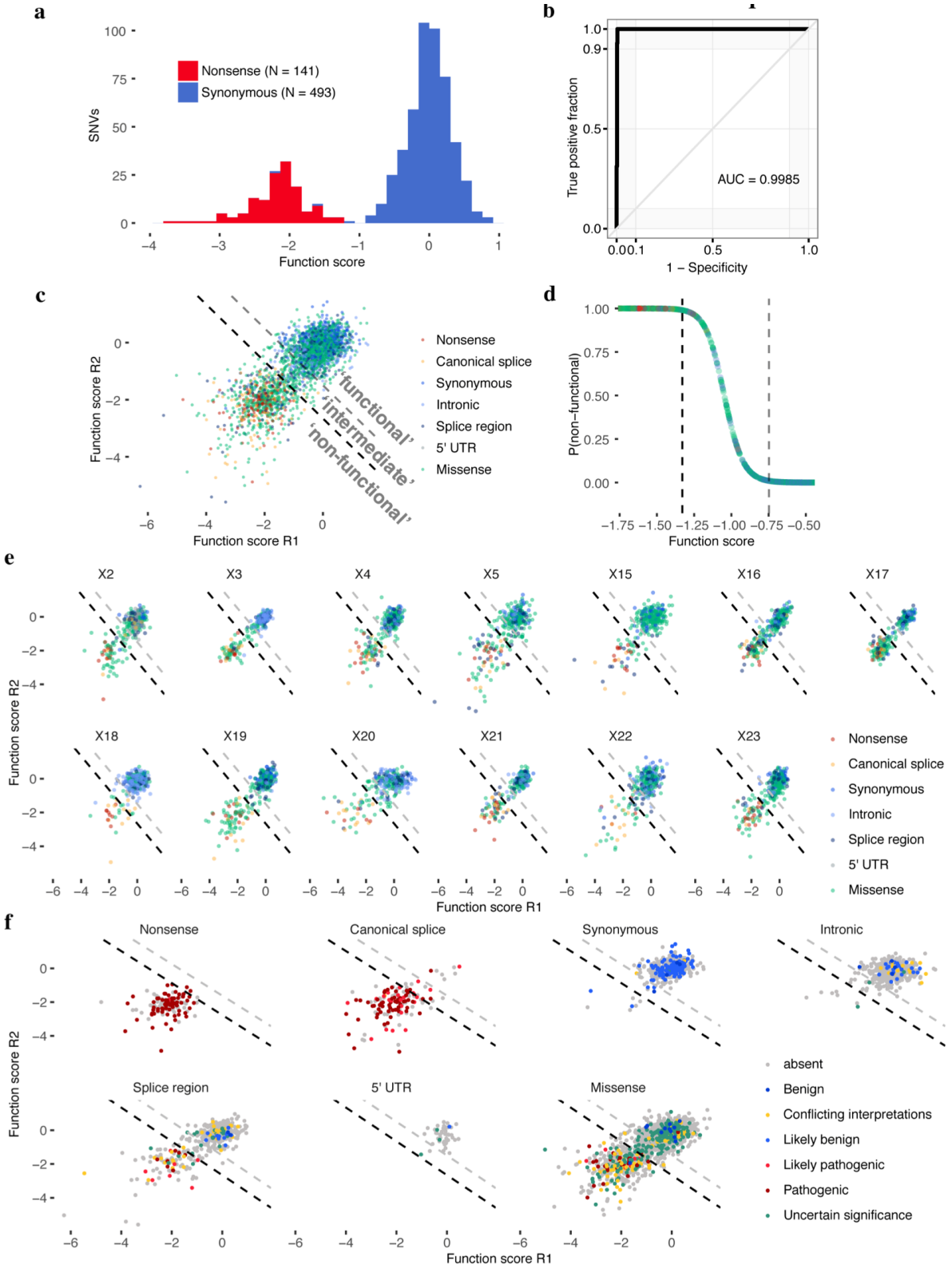



**Figure 3.9. Mixture modeling of scores to classify SNVs by functional effect.**

**a**, Distributions of 'non-functional' and 'functional' SNVs plotted here were defined respectively

as all nonsense SNVs and all synonymous SNVs with RNA scores within 1 SD of the median synonymous SNV. **b**, An ROC curve was generated using SGE function scores to distinguish the 634 'functional' and 'non-functional' SNVs defined in **a**. **c**, A two-component Gaussian mixture model was used to produce point estimates of the probability that each SNV was 'non-functional', *P*(nf), given its average function score across replicates. These P-values are plotted in **d** against function scores for a subset of the data. Thresholds were set such that $P(\text{nf}) < 0.01$ corresponds to 'functional', and $P(\text{nf}) > 0.99$ corresponds to 'non-functional', and $0.01 < P(\text{nf}) < 0.99$ corresponds 'intermediate' classification. Functional classification thresholds are drawn as dashed lines; black denotes the non-functional threshold and gray the intermediate threshold. **e**,**f**, SNV function scores across replicates are plotted for each exon with SNVs colored by mutational consequence (**e**), and for each type of mutational consequence with SNVs colored by ClinVar status (**f**). Using the optimal function score cutoff for all SNVs tested (**Figure 3.10b**), sensitivities and specificities for distinguishing 'Pathogenic'/'Likely pathogenic' from 'Benign'/'Likely benign' ClinVar annotations for each type of mutation are as follows: 92.7% and 92.9% for missense SNVs (N = 55), 100% and 100% for splice region SNVs (N = 23), and 95.2% sensitivity for canonical splice site SNVs (N = 83; specificity not calculable).

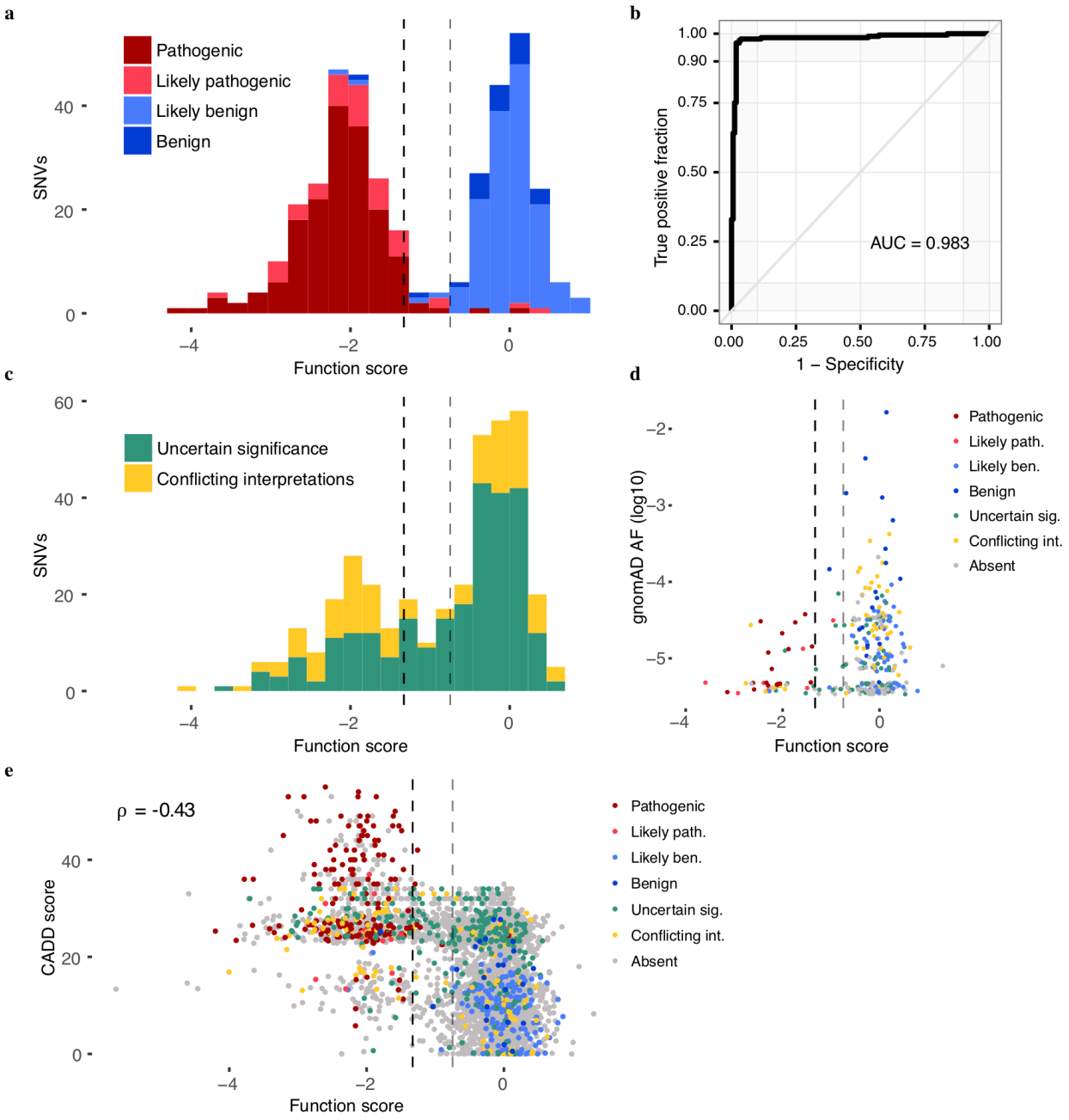



**Figure 3.10. SGE function scores are highly accurate at predicting clinical interpretations of *BRCA1* SNVs.**

**a**, The distribution of SNV function scores colored by ClinVar interpretation. Scores are shown for the 375 SNVs with at least a '1-star' review status in ClinVar and either a 'pathogenic' or 'benign' interpretation (including 'likely'). The dashed lines indicate the functional classification thresholds determined by mixture modeling (gray = intermediate, black = non-functional). **b**, An ROC curve reveals optimal sensitivity and specificity for classifying the same 375 SNVs in **a** at SGE function score cutoffs from -1.03 to -1.22. **c**, The distribution of scores plotted as in **a** for the 378 SNVs annotated as variants of uncertain significance or with conflicting interpretations. 91.3%

of such variants are classified as ‘functional’ or ‘non-functional’. **d**,**e**, SNVs are colored by ClinVar annotation. **d**, Among the 302 SNVs assayed also present in gnomAD, higher allele frequencies associated with higher function scores (Wilcoxon Signed Rank Test, $P = 3.7 \times 10^{-12}$). **e**, CADD scores (which predict deleteriousness) inversely correlate with function scores.

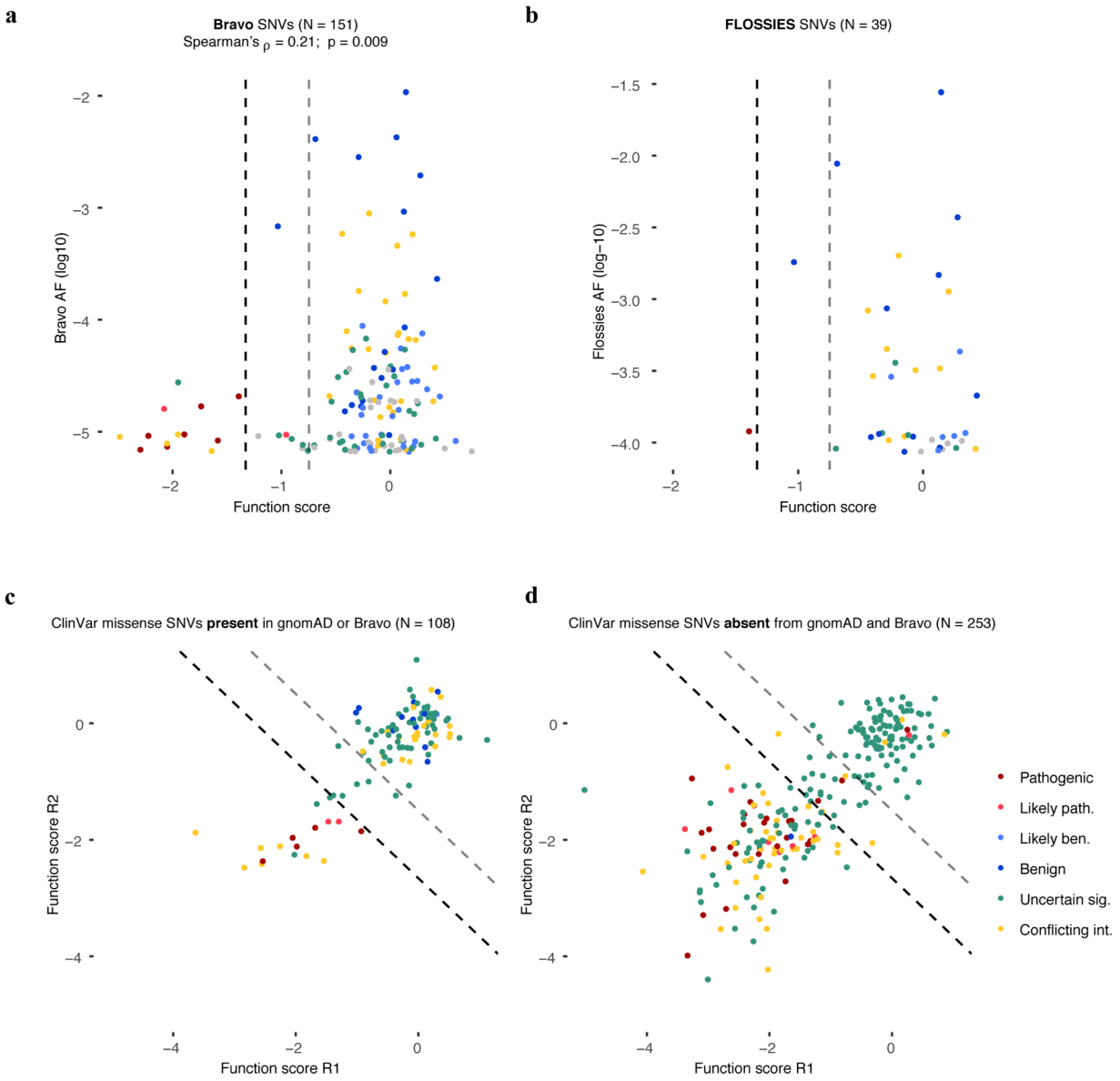



**Figure 3.11. *BRCA1* SNVs observed more frequently in large-scale population sequencing are more likely to score as functional.**

SNV function scores are plotted against Bravo allele frequencies (**a**) and FLOSSIES allele frequencies (**b**). **a**, Bravo is a collection of whole genome sequences ascertained from 62,784 individuals through the NHLBI TOPMed program. Similarly to SNVs present in gnomAD (**Figure 3.10d**), higher allele frequencies of SNVs in Bravo correlate with higher function scores. **b**, FLOSSIES is a database of variants seen in targeted sequencing of breast cancer genes sampled from approximately 10,000 cancer-free women at least 70 years old. Only 1 of 39 SNVs observed in FLOSSIES scored as non-functional. **c**,**d**, Missense SNVs in ClinVar are separated by whether they have (**c**) or have not (**d**) been seen in either gnomAD or Bravo and function scores across replicates are plotted, with dashed lines demarcating functional classes. A higher proportion of ClinVar missense SNVs absent from gnomAD and Bravo score as non-functional (50.6% vs. 15.7%, Fisher's exact $P = 1.80 \times 10^{-17}$).

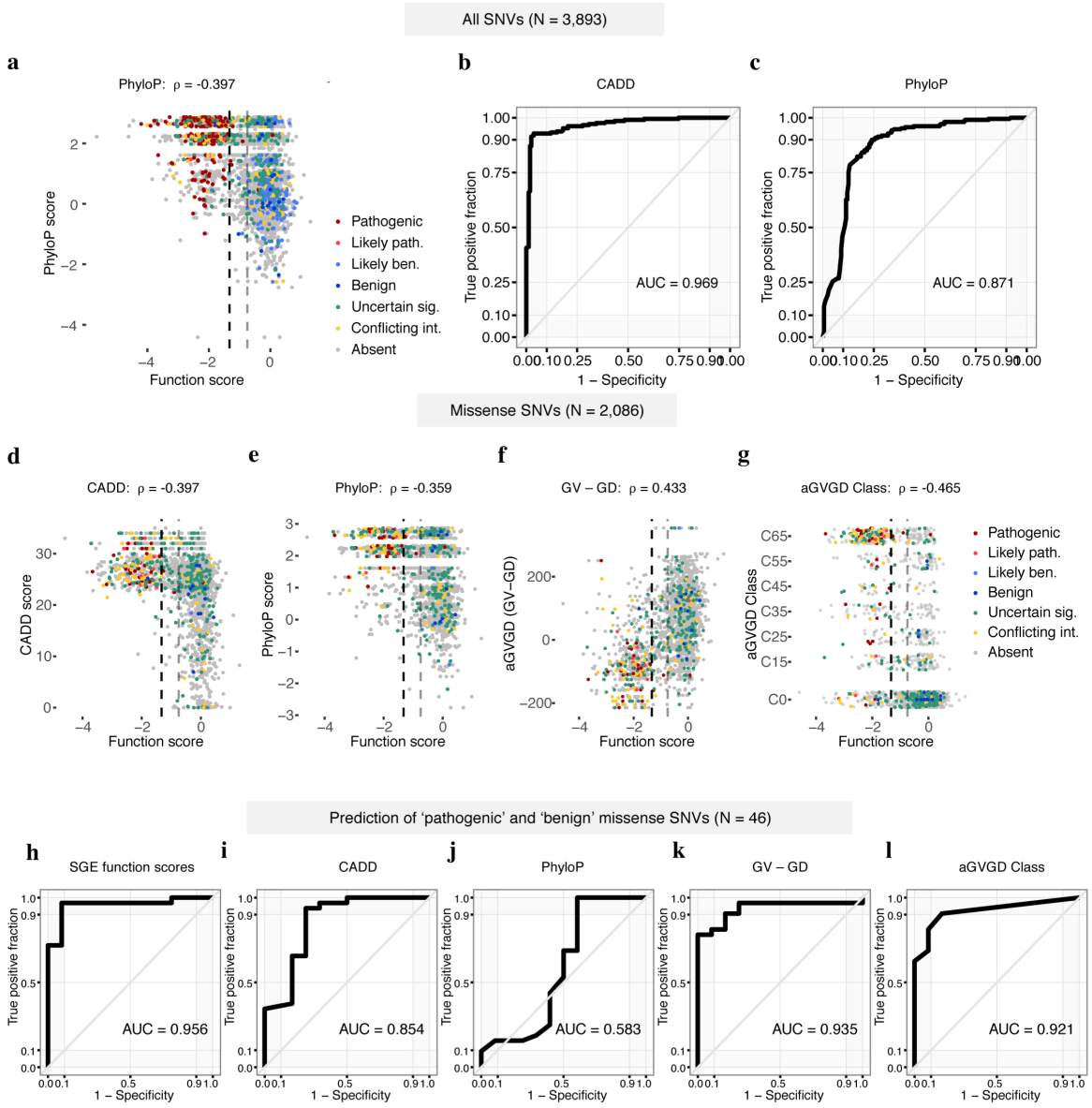



**Figure 3.12. SGE function scores correlate with computational metrics and perform favorably at predicting ClinVar annotations.**

**a**, SNV function scores are plotted against mammalian phyloP scores, with colors indicative of ClinVar status. **b**,**c**, ROC curves show the performance of CADD scores and phyloP scores for discriminating ClinVar 'pathogenic' and 'benign' SNVs (including 'likely'), as described in **Figure 3.10b** for SGE data. **d-g** Plots as in **a**, but for missense SNVs only, showing correlations between SGE function scores and CADD (Kircher et al., 2014) scores, phyloP scores (Pollard et al., 2010), Grantham differences (Grantham amino acid variation minus Grantham amino acid deviation; GV - GD), and align-GVGD classifications (Tavtigian et al., 2006). Missense SNV function scores also correlate with SIFT scores (Kumar et al., 2009) (ρ = 0.363) and PolyPhen-2 scores (Adzhubei and Jordan, 2013) (ρ = -0.277). ($P < 1 \times 10^{-37}$ for all correlations.) **h-l**, ROC curves assess the performance of SGE function scores and each indicated metric at distinguishing firmly 'pathogenic' and 'benign' missense SNVs. (*i.e.* not including 'likely').

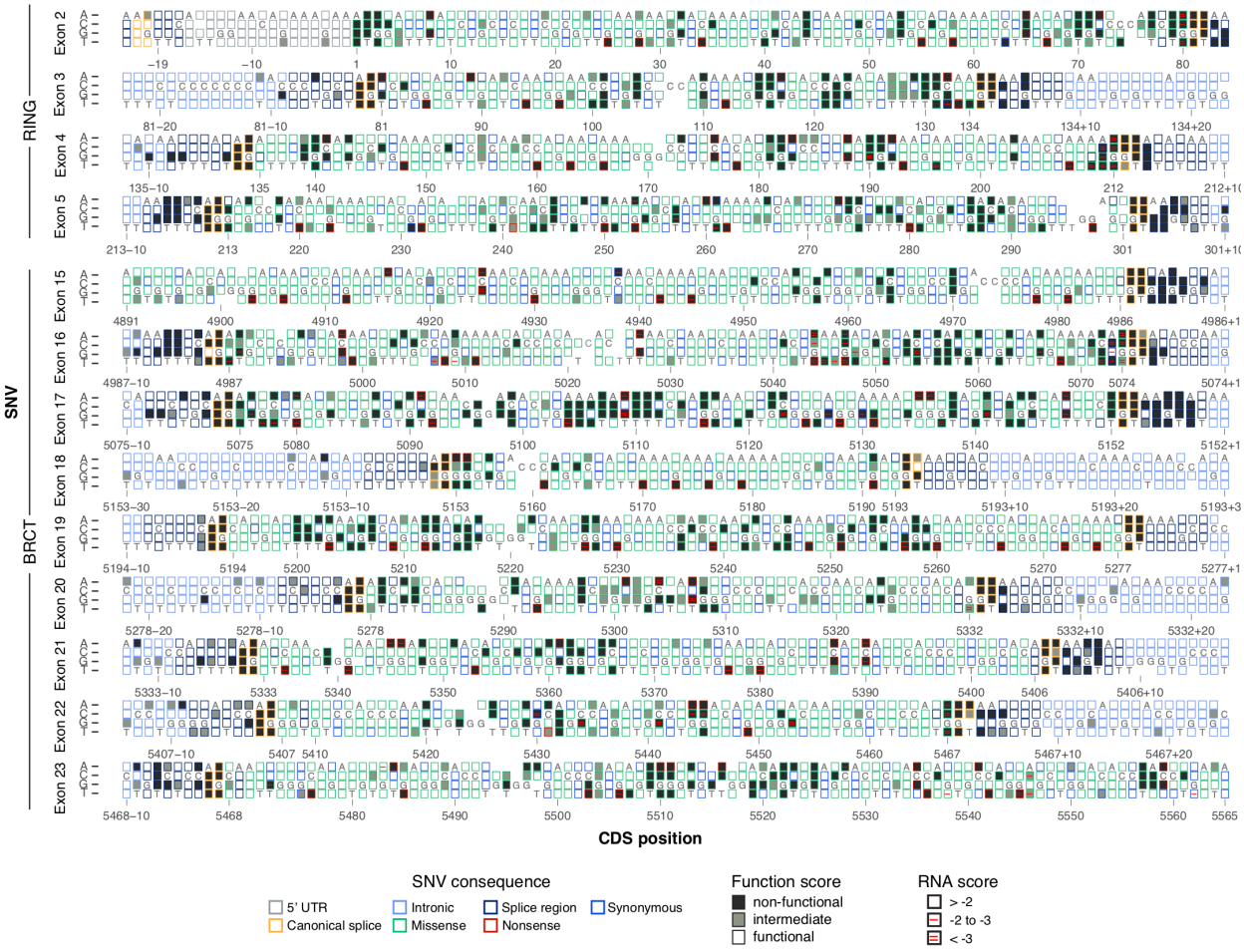



**Figure 3.13. Sequence-function maps for 13 *BRCA1* exons.**

The 3,893 SNVs scored with SGE are each represented by a box corresponding to coding sequence position (NCBI transcript ID: NM_007294.3) and nucleotide identity. Boxes are filled corresponding to functional class, and outlined corresponding to the SNV's mutational consequence. Red lines within boxes mark SNVs depleted in RNA; one line indicates an RNA score between -2 and -3 (log2 scale) and two lines indicate a score below -3. RNA measurements were determined only for exonic SNVs, excluding exon 18. Reference nucleotides are indicated by dark gray letters; blank boxes indicate missing data.

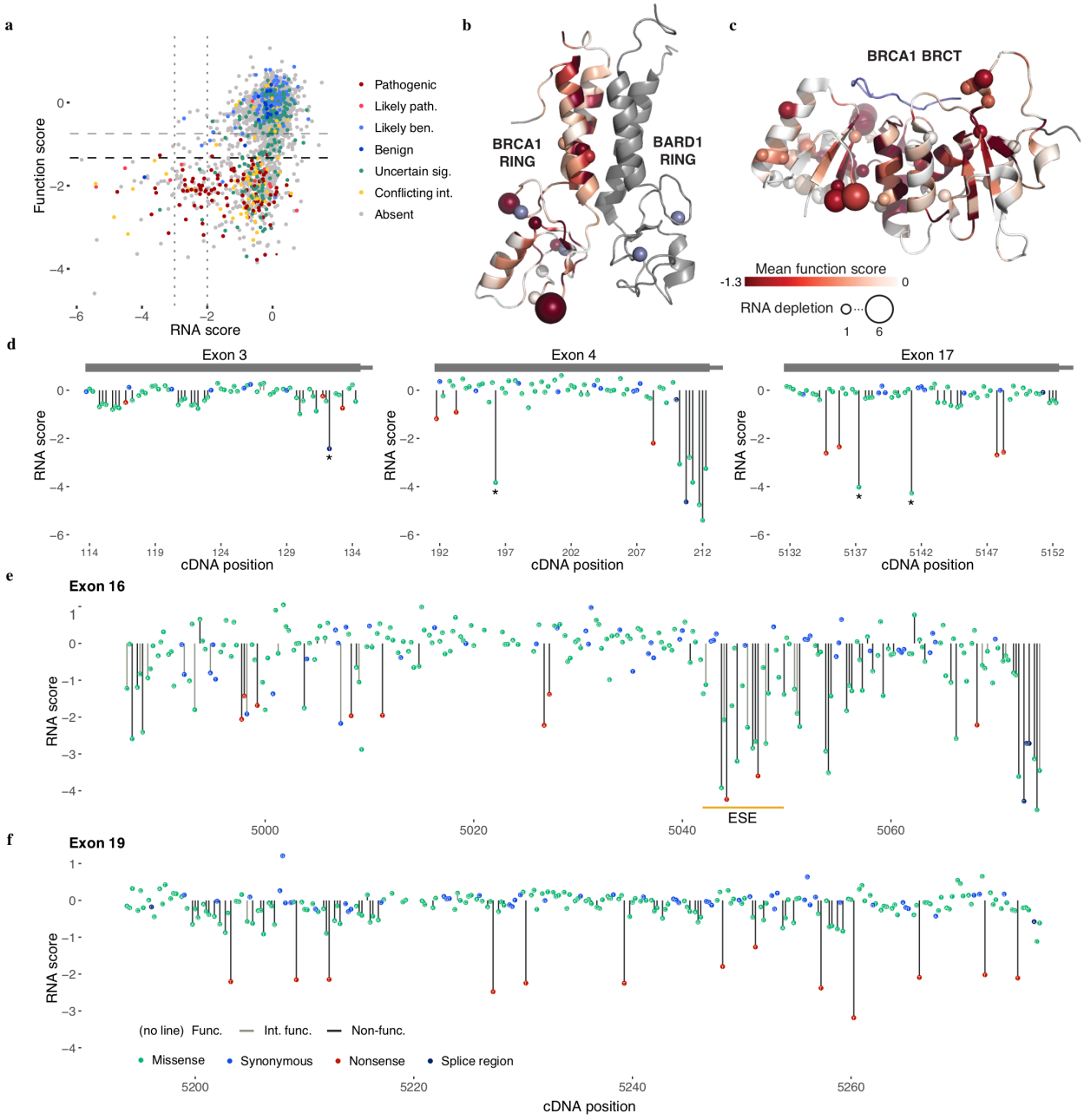



**Figure 3.14. Measuring SNV mRNA abundance and function in parallel delineates mechanisms of variant effect.**

**a**, Function scores are plotted against RNA scores for all exonic synonymous and missense SNVs scored (N = 2,646). Horizontal dashed lines indicate functional thresholds, and vertical dotted lines mark RNA scores of -2 and-3. **b,c**, Function scores for all SNVs were mapped onto the structures of the RING (**b**, pdb 1JM7) and BRCT (**c**, pdb 1T29) domains in shades of red by averaging missense SNV scores at each amino acid position. The number of SNVs that cause >75% reduction in RNA levels at each amino acid position is represented by the size of the sphere at the alpha-carbon at each residue. Grey denotes residues not assayed and the BACH1 peptide bound to the BRCT structure is colored slate blue. **d**,**e**,**f**, SNV RNA scores are plotted by transcript position,

with lines denoting SNV functional classification. **d**, Examples of non-functional SNVs with low RNA scores that create new 5'-GU splice donor motifs are shown. Complete maps of RNA scores for exons 16 (**e**) and exon 19 (**f**) reveal highly variable sensitivity to RNA depletion. The location of the strongest predicted exonic splice enhancer in exon 16 is indicated by the orange line (Desmet et al., 2009) (**e**).

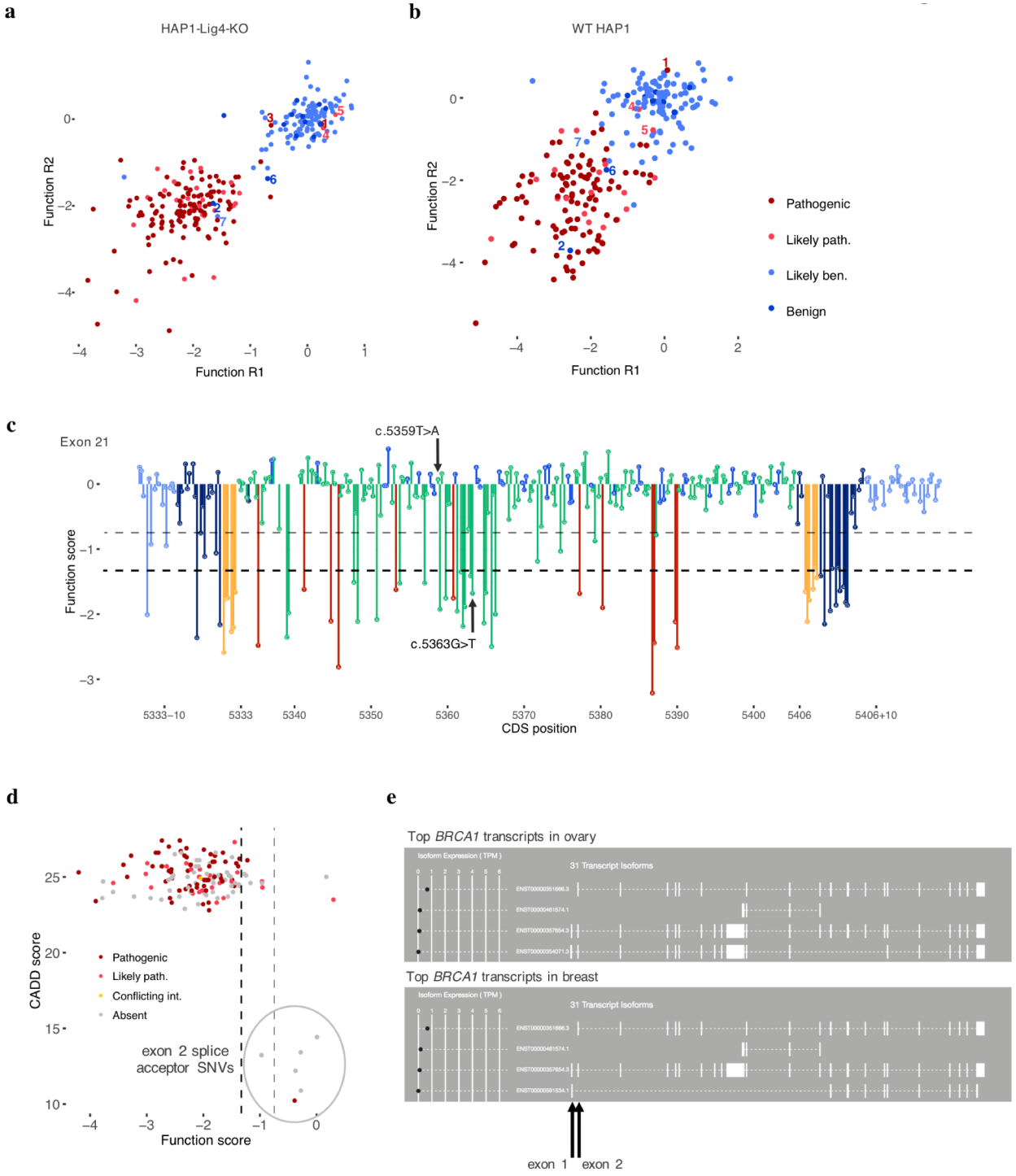



**Figure 3.15. Evidence supporting SNV scores in discordance with ClinVar classifications.**

Function scores of SNVs classified as 'benign' or 'pathogenic' (including likely's) are shown across replicates for experiments using HAP1-Lig4KO cells (**a**) and for preliminary experiments using WT HAP1 cells (**b**). Plots exclude exons with low overall reproducibility in WT HAP1 cells (replicate correlations < 0.4: exons 15, 18, 20 and 22). The three SNVs firmly discordant with ClinVar are labelled 1-3 in **a**, corresponding to c.5359T>A (dark red 1), c.5044G>A (dark blue 2),

and c.-19-2A>G (dark red 3), respectively. The same filtering criteria were applied to both sets of experiments, which led to the removal of SNV 3 from the WT HAP1 data due to disagreement of scores between replicates. Discordant 'likely pathogenic' SNVs (4,5), an intermediate scoring 'benign' SNV (6) and a discordant 'likely benign' SNV (7) are also labelled for comparison. **c**, The sequence-function map of exon 21 is shown with the function scores for the two 'pathogenic' SNVs observed in linkage indicated. Dashed lines demarcate functional classifications. **d**, Function scores are plotted against CADD scores for all canonical splice SNVs assayed, colored by ClinVar status. The six possible exon 2 splice acceptor SNVs (circled) have the lowest CADD scores among all canonical splice SNVs assayed, and none score as 'non-functional'. **e**, GTEx browser shots show that many of the most common *BRCA1* transcripts mapped from ovarian and breast tissues lack the exon 1 / exon 2 junction.

**Table 3.1. DNA sequences used to program SNVs in *BRCA1* saturation genome editing experiments.**

| Name | Sequence (5' to 3') |
|---|---|
| BRCA1x2_SGE | AAGTTCAcTGGAACAGAAAGAAATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAATGTCATTA ATGCTATGCAGAAAATCTTAGAGTGTCCgATCTGGTAA |
| BRCA1x3_SGE | TTTCTCCCCCCCTACCCTGCTAGTCTGGAGTTGATCAAGGAACCTGTCTCgACAAAGTGTGACCACATA TTTTGCAAGTAAGTTTGAATGTGTTATGTGc |
| BRCA1x4_SGE | TATAATTTATAGATTTTGCATGCTGAAACTTCTCAACCAGAAGAAAGGcCCTTCACAGTGTCCTTTATG TAAGAATGATATAACCAAAAGGTATATAATT |
| BRCA1x5_SGE | TTAATTTCAGGAGCCTACAgGAAAGTACGAGATTTAGTCAACTTGTTGAAGAGCTATTGAAAATCATTT GTGCTTTTCAGCTTGACACAGGTTTaGAGTGTAAGTGTTG |
| BRCA1x15_SGE | AGTGTGAGCAGaGAGAAGCCAGAATTGACAGCTTCAACAGAAAGGGTCAACAAAAGAATGTCCATGG TGGTGTCTGGCCTGACgCCAGAAGAATTTGTGAGTGTAT |
| BRCA1x16_SGE | TTAATTTCAGATGCTCGTGTACAAGTTTGCCAGAAAACACCACATaACcTTAACTAATCTAATTACTGA AGAGACTACTCATGTTGTTATGAAAACAGGTATACCAAG |
| BRCA1x17_SGE | CATTCTGCAGATGCTGAGTTTGTGTGTGAACGcACACTGAAATATTTTCTAGGAATTGCGGGAGGAAAA TGGGTAGTTAGCTATTTCTGTAAGTATAA |
| BRCA1x18_SGE | TGTAACCTGTCTTTTCTATGATCTCTTTAGGGGTGACgCAGTCTATTAAAGAAAGAAAAATGCTGAATG AGGTAAGTACTTGATGTTACAAACTAACtAGA |
| BRCA1x19_SGE | TTTCTTTCAGCATGATTTTGAAGTCAGAGGAGATGTcGTCAATGGAAGAAACCACCAAGGTCCAAAGC GAGCAAGAGAATCCCAGGACAGAAAGGTAAAGCTCa |
| BRCA1x20_SGE | CTCTCTTCCTCTCTTCTTCCAGATCTTCAGGGGcCTAGAAATCTGTTGCTATGGGCCCTTCACCAACATG CCCACAGGTAAGAGCCTGcGAGAACCCCAG |
| BRCA1x21_SGE | ATGTCCATTTTAGATCAACTcGAATGGATGGTACAGCTGTGTGGTGCTTCTGTGGTGAAGGAGCTTTCA TCATTCACCCTTGGCACAGTAAGTATTtGGTGCCCT |
| BRCA1x22_SGE | TCCTGGGGATCCAGGGTGTCCACCCAATTGTcGTTGTGCAGCCAGATGCCTGGACAGAGGACAATGGC TTCCATGGTAAGGTGtCTGCATGTACCTGTGC |
| BRCA1x23_SGE | CTGTCTCCAGCAATTGGGCAGATGTGTGAGGCACCTGTcGTGACCCGAGAGTGGGTGTTaGACAGTGTA GCACTCTACCAGTGCCAGGAGCTGGACACCTACCTGATA |

# Chapter 4. CRISPR/CAS9-MEDIATED SCANNING FOR REGULATORY ELEMENTS REQUIRED FOR *HPRT1* EXPRESSION VIA THOUSANDS OF LARGE, PROGRAMMED GENOMIC DELETIONS

Chapter 4 is adapted with minimal modifications from:

> Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.
>
> First authorship is shared between MG and GMF.

## 4.1 ABSTRACT

The extent to which non-coding mutations contribute to Mendelian disease is a major unknown in human genetics. Relatedly, the vast majority of candidate regulatory elements have yet to be functionally validated. Here we describe a CRISPR-based system that uses pairs of guide-RNAs (gRNAs) to program thousands of kilobase-scale deletions that deeply scan across a targeted region in a tiling fashion ("ScanDel"). We applied ScanDel to *HPRT1*, the housekeeping gene underlying Lesch-Nyhan syndrome, an X-linked recessive disorder. Altogether, we programmed 4,342 overlapping 1- and 2- kilobase (Kb) deletions that tiled 206 Kb centered on *HPRT1* (including 87 Kb upstream and 79 Kb downstream), with median 27-fold redundancy per base. Programmed deletions were functionally assayed in parallel by selecting for loss of HPRT function with 6-thioguanine. As expected, sequencing gRNA pairs before and after selection confirmed all *HPRT1* exons are needed. However, *HPRT1* function was robust to deletion of any intergenic or deeply intronic non-coding region, indicating proximal regulatory sequences are sufficient for

*HPRT1* expression. Although our screen did identify the disruption of exon-proximal non-coding sequences (*e.g.* the promoter) as functionally consequential, long-read sequencing revealed this signal was driven by rare, imprecise deletions that extended into exons. Our results suggest no singular distal regulatory element is required for *HPRT1* expression, and that distal mutations are unlikely to contribute substantially to Lesch-Nyhan syndrome burden. Further application of ScanDel may shed light on the role of regulatory mutations in disease at other loci, while also facilitating a deeper understanding of endogenous gene regulation.

## 4.2 INTRODUCTION

The success of human genetics in identifying the genes and mutations underlying Mendelian diseases has been facilitated by the fact that the majority of causal mutations lie in protein-coding sequences or splice junctions. Indeed, this assumption is explicit in both classic and contemporary practices in genetics (*e.g.* exome sequencing). However, it is clear that distal non-coding mutations make *some* contribution to Mendelian disease. Understanding how often non-coding mutations play a causal role, as well as developing best practices for pinpointing those that do, are critical challenges for the field. For example, in the clinic, even if a person is diagnosed with a monogenic Mendelian disorder on the basis of phenotype, clinical sequencing mainly of coding regions fails to identify a causal mutation ~10% of the time (Chong et al., 2015). However, possible explanations include not only distal regulatory mutations, but also misdiagnosis, somatic mutation, technical false negatives, and others. Furthermore, non-coding loci could contribute to the estimated ~25-50% of undiagnosed but apparently Mendelian cases in which the underlying gene is unknown (Chong et al., 2015; Yang et al., 2013).

The picture is very different for the genetics of common disease, where over 90% of disease-associated SNPs fall in non-coding regions (Maurano et al., 2012). Many resources have

been developed to predict the location of putative regulatory elements and the effects of regulatory mutations (Ernst and Kellis, 2012; Hoffman et al., 2012; Kircher et al., 2014), with ~88% of all protein-coding genes tied to a *cis*-expression quantitative locus (eQTL) (Aguet et al., 2016), ~80% of the genome annotated with biochemical function (ENCODE Project Consortium, 2012), and numerous tools to link regulatory elements to their target genes (Boyle et al., 2012; Coetzee et al., 2012; Li et al., 2013; Ward and Kellis, 2012). However, the vast majority of these predictions are either confounded (*e.g.* for *cis*-eQTLs, by linkage disequilibrium) or lack functional validation. Indeed, there are few distal non-coding regulatory elements that we can confidently assign to a target gene, or for which we understand the consequences of disruption.

Large-scale functional experiments are clearly an important next step for both common disease genetics (to facilitate the identification of causal regulatory variants and their target genes) and rare disease genetics (to identify distal regulatory elements for Mendelian disease genes where causal non-coding mutations might be found). Many important studies have undertaken functional work to identify and characterize causal or risk contributory non-coding variants for specific rare and common diseases (Claussnitzer et al., 2015; Wakabayashi et al., 2016; Weedon et al., 2014), but by approaches that are not easily scalable. Within the last year, several studies have used CRISPR/Cas9 genome editing in cell-based screens to introduce and functionally assay large numbers of non-coding mutations at an unprecedented scale (Canver et al., 2015; Chen et al., 2015; Diao et al., 2016; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2014). The common approach of these studies is to introduce complex libraries of guide RNAs (gRNAs) via lentiviral infection to a population of cells at a low multiplicity of infection (MOI), followed by an assay that queries the function or expression of a gene of interest. CRISPR/Cas9 mediates double-

stranded breaks at sites specified by the gRNA in each cell, eventually resulting in a mutation at each targeted site via imperfect non-homologous end joining (NHEJ).

A fundamental limitation of these singleton gRNA screens is that because of design constraints (*e.g.* the uneven distribution of protospacer adjacent motif (PAM) sequences, the variable efficiency of gRNAs, and others), the resulting coverage of regions of interest is incomplete and uneven. As the majority of bases will be perturbed by zero or only one gRNA, these studies rely on the aggregate behavior of clusters of target sites within potential regulatory elements or arbitrarily sized windows (*e.g.* 500 base-pairs) (Sanjana et al., 2016), rather than redundant targeting of each base-pair (bp) by independent gRNAs. Furthermore, it is possible that the mutations introduced by NHEJ at single sites, which are highly heterogeneous but mainly dominated by small 1-10 bp deletions (McKenna et al., 2016; Tsai et al., 2015), are insufficient to fully disrupt many regulatory elements. Several recent studies have employed an inhibitory domain guided by nuclease-inactive Cas9 to screen non-coding regulatory regions, *i.e.* CRISPRi (Fulco et al., 2016; Klann et al., 2017). Epigenetic modifications mediated by these domains can spread to regions on the order of ~200 bp to 4.5 Kb (Horlbeck et al., 2016; Thakore et al., 2015), and thus mitigate the challenges related to redundancy and coverage of individual gRNA screens. However, CRISPRi screens may be less precise because of this spreading effect and furthermore, do not directly test the consequences of alterations in primary sequence.

Here we sought to overcome these weaknesses by introducing *pairs* of gRNAs to each cell, with the goal of inducing a kilobase-scale deletion of the intervening DNA between two programmed cuts. A principal advantage of this method is that by tiling deletions across a region, each targeted base-pair can be covered with high redundancy (scanning deletion or "ScanDel"). Furthermore, kilobase-scale deletions are much more likely to eliminate the function of an

overlapping or fully contained regulatory element, relative to small indels resulting from NHEJ at a single target site. Our approach is analogous to classic deletion scanning experiments (Reid et al., 1990; Rincón-Limas et al., 1991), but with advantages in throughput and of targeting much larger regions in the endogenous genome rather than sequences cloned to a plasmid. Similar strategies have recently been described for the interrogation of lncRNA genes (Zhu et al., 2016) and non-coding sequences (Diao et al., 2017). Critically, these implementations (and indeed, all CRISPR genetic screens) rely on indirectly genotyping the lentivirally inserted gRNA sequences, instead of using direct sequencing of edited loci to confirm exactly which CRISPR-induced genotypes are driving effects.

Here we applied ScanDel to survey the genomic locus encompassing *HPRT1*, which encodes the enzyme hypoxanthine(-guanine) phosphoribosyltransferase (HPRT). *HPRT1* is a housekeeping gene, a class of genes primarily defined by their broad expression and for which the underlying regulatory architecture remains unclear (Zabidi et al., 2014). Loss-of-function mutations in *HPRT1* result in the X-linked Lesch-Nyhan syndrome (Lesch and Nyhan, 1964), in which a minority of individuals present with reduced HPRT enzymatic activity despite the absence of identifiable coding mutations (Fu et al., 2014a). Such individuals could carry non-coding mutations that result in reduced *HPRT1* expression. Reduced HPRT activity also causes resistance to the drug 6-thioguanine (6TG), a purine analog and chemotherapeutic agent. Thus, it is straightforward to assay cell populations for loss of *HPRT1* function, as only cells with highly reduced expression of functional HPRT will survive selection by 6TG (**Figure 4.1c**). Although there are no known distal regulatory elements of *HPRT1*, its nine exons serve as internal controls.

Adopting the framework of genome-wide CRISPR/Cas9 screens, we synthesized, cloned, and lentivirally delivered thousands of programmed gRNA pairs to cells at a low MOI. Each gRNA

pair targets nearby sites, effectively leveraging CRISPR/Cas9's ability to generate kilobase-scale deletions when NHEJ-mediated repair of two double-stranded breaks results in excision of the intervening DNA segment. In total, we designed and introduced gRNA pairs programming 4,342 overlapping ~1 and ~2 kilobase (Kb) deletions that tiled a 206 Kb region centered on *HPRT1*. 6TG was used to select for cells that had lost *HPRT1* function. By quantifying gRNA pairs both before and after 6TG selection and then directly genotyping putatively important deletions by long-read sequencing, we were able to identify programmed deletions that significantly compromised *HPRT1* expression and function.

## 4.3 RESULTS

### 4.3.1 *Development of ScanDel*

In genome-wide CRISPR/Cas9 screens, a gRNA library is lentivirally delivered to a large pool of cells at a low MOI, such that each infected cell is likely to receive only one gRNA (Shalem et al., 2014; Wang et al., 2014; Zhou et al., 2014). Each gRNA induces NHEJ-mediated indels centered at the Cas9-mediated cleavage position within the target sequence, with the goal of perturbing the function of the targeted locus. However, given the small and variable length of indels, the robustness of perturbation is inherently limited, particularly when targeting non-coding sequences in which frameshifts are irrelevant. To instead program a kilobase-scale deletion in each cell, we devised the following approach (**Figure 4.1**). First, gRNA pairs are designed to program specific deletions (with each gRNA specifying one of the deletion's boundaries, **Figure 4.1a**), and the corresponding pairs of 20 bp spacers are synthesized *in cis* on a microarray (**Figure 4.1b**). Second, the paired spacers are inserted into the lentiGuide-Puro plasmid between the U6 promoter and the gRNA backbone. Third, a second gRNA backbone and a second RNA Polymerase (Pol) III promoter (H1 or U6) are inserted between the paired spacers. Fourth, libraries of "gRNA pairs"

are lentivirally delivered to a large pool of cells at a low MOI, such that each cell receives a pair of gRNAs that programs a single deletion (**Figure 4.1c**). Finally, analogous to conventional genome-wide CRISPR/Cas9 screens, deep sequencing of the integrated gRNA pairs is used as a surrogate measure of the prevalence of each programmed deletion in a population of cells (*e.g.* before and after the cells have been subjected to functional selection) thus capturing the phenotypic consequences of individual deletions.

As an initial test of our paired guide system, we compared the efficacy of using two different promoters for the two guides (a 'U6-H1' system) versus using two copies of the same promoter ('U6-U6'). We tested these lentiviral gRNA pair expression constructs by targeting the same genomic site for deletion with each system (**Figure 4.2**). PCR amplification of the site was performed with UMIs in order to minimize biases related to amplicon size. The U6-H1 system induced more programmed deletions than the U6-U6 system (20% vs. 10% of reads from cells one week after transduction). The U6-H1 system has several advantages (*e.g.* avoiding recombination between the two U6 promoters during cloning; unique primer design for deep sequencing of each gRNA), and we therefore proceeded with it.

An important caveat for ScanDel, relative to conventional gRNA cell-based screens, is that deletions programmed by gRNA pairs only occur in a minority of cells (Byrne et al., 2015; Canver et al., 2014), with the other major outcomes being small NHEJ-mediated indels at one or both gRNA-targeted sites. For example, in our test of the U6-H1 system, the programmed deletion was found in 32% of cells that had any edit, while the remaining edited cells were mutated at one or both gRNA-targeted sites but retained the intervening sequence. While this complicates interpretation, the problem can be overcome by using a robust functional assay in conjunction with multiple, independent gRNA pairs that query the same genomic region, as well as by including

unpaired gRNA controls to ensure that observed effects do not occur with the individual gRNAs that comprise each pair (but rather are dependent on the presence of both gRNAs).

### 4.3.2 *Application of ScanDel to survey the 206 Kb region surrounding HPRT1*

With the goal of investigating the potential of non-coding mutations to compromise its function, we applied ScanDel to a 206 Kb region on the X chromosome centered on the *HPRT1* gene (**Figure 4.1a**, **Figure 4.3a**). We designed pairs of gRNAs that programmed deletions tiling across the 206 Kb region, including tiles that overlapped *HPRT1* exons in order to allow coding regions to serve as positive controls. As deletion length has been shown to affect deletion rate (Canver et al., 2014), deletions were programmed to be consistently either ~1 or ~2 Kb in length (**Figure 4.1a**). This design resulted in 4,342-programmed deletions that tiled across the region, collectively covering each base-pair a median of 27 times (**Figure 4.3b**). Testing each base-pair with numerous independently programmed, tiling deletions is expected to reduce noise and also increase resolution (as all successfully made deletions tiling a critical regulatory element should exhibit positive selection). However, to guard against the possibility that individual gRNAs' effects could confound analysis (*e.g.* via off-target mutations, or on-target small ~10 bp indels), we also included all spacers in the library as pairs with themselves ('self-pairs'; **Figure 4.1b** inset, **Figure 4.4**). Additionally, we included 330 negative control gRNA pairs not expected to survive 6TG selection, as they program deletions in non-genic regions far from *HPRT1* or use spacers made of random sequence not present in the reference genome (hg19).

The gRNA pair library was array-synthesized, cloned, and delivered via lentiviral infection to HAP1 cells in replicate (**Figure 4.1b**,**c**). Cell populations were sampled before and after one week of the 5 μM 6TG selection, with PCR amplification and deep sequencing of gRNA pairs to quantify abundance at each time-point. The functional selection score was calculated as the log10

ratio of normalized read counts after selection relative to before selection ("selection score"). Positively scoring self-paired spacers were flagged, and gRNA pairs that used these flagged spacers were excluded from further analysis (11% of pairs in replicate 1 and 3% of pairs in replicate 2). To integrate signal from overlapping programmed deletions, we calculated a "per base-pair" metric as the mean of selection scores of all deletions overlapping a given base (**Figure 4.3d**, **Figure 4.5**). This per base-pair score across the *HPRT1* locus was well-correlated between biological replicates (Pearson: 0.708; **Figure 4.6**). Importantly, none of the negative-control gRNA pairs that were sampled in each of the two replicates were positively selected in both experiments (**Figure 4.7**).

Crucially, all nine *HPRT1* exons exhibited strong functional scores, confirming the sensitivity of ScanDel as applied here to detect sequences essential to *HPRT1* function (**Figure 4.8**). However, all of the reproducibly positive non-coding signal across the 206 Kb region was immediately proximal to an *HPRT1* exon. This result suggests that there is no distal regulatory element in the 206 Kb region that is essential to *HPRT1* expression in HAP1 cells.

Near exons, non-coding regions exhibiting positive signal did so even when deletions that also overlapped the exons themselves were excluded from the analysis (**Figure 4.8d**). This suggested the presence of essential, proximal regulatory sequences. We noted that the positively scoring regions immediately upstream and downstream of the first exon overlapped with a region of open chromatin identified by performing ATAC-seq in HAP1 cells, supporting the region's role in gene regulation (**Figure 4.3c, Figure 4.8a, Figure 4.9**). Together, these observations motivated us to attempt validation experiments for this region, with the goal of directly confirming which deletions of putative regulatory elements were impairing *HPRT1* function (**Figure 4.10a,e**).

### 4.3.3 *Direct genotyping of deletions that survive functional selection*

With the goal of validating the positive signal upstream of the first exon, we repeated the experiment with a small pool of 4 gRNA pairs targeting the putative *HPRT1* promoter (**Figure 4.10b**). We then amplified 3 Kb of this region by PCR and performed long-read sequencing of the amplicons (Pacific Biosciences). As expected, before 6TG selection, the programmed deletions were all well-represented in the population, although deletions with boundaries deviating from Cas9 cut sites (*i.e.* 'unprogrammed') were also detected (**Figure 4.10c**). However, after selection with 6TG, deletions with unprogrammed boundaries predominated, including those unseen before 6TG, and those that extend beyond the transcriptional start site (TSS) (**Figure 4.10d**). The fact that these initially rare deletions were strongly selected (while 2 Kb promoter deletions that did not cross the TSS were not) suggests that even relatively proximal sequences upstream of the *HPRT1* TSS are not strictly essential for expression. Based on the results of these validation experiments, we conclude that only a narrow window of non-coding sequence immediately upstream of the TSS and 5'UTR is required for *HPRT1* expression.

We next sought to validate the positive signal downstream of the first exon. To do so, we again repeated the experiment with a small pool of just 5 gRNA pairs targeting the first ~2.7 Kb of intron 1 (**Figure 4.10f**). We then amplified the region and again performed long-read sequencing of the amplicons (Pacific Biosciences). As with the promoter, the programmed deletions were all well-represented before 6TG selection, although deletions with unprogrammed boundaries are also detected at a low rate (**Figure 4.10g**). After selection, deletions with unprogrammed boundaries predominated again, particularly those that extended into the first exon, thereby disrupting coding sequences (**Figure 4.10h**). A low rate of non-exonic deletions survived post-6TG, but these were present at the same level as unedited reads, implying that there may be some other explanation for

6TG resistance in these cells. Thus, as with the promoter, the positive signals that we originally observed for deletions in the first intron were likely consequent to the positive selection of rare 'on-target-but-with-incorrect-boundaries' deletions that extend into the first *HPRT1* exon.

### 4.3.4 *An individual gRNA screen of the same region for comparison to ScanDel*

We next compared our ScanDel results against a more conventional screen relying on only individual gRNAs (**Figure 4.3e**). For this, we cloned a second lentiviral library consisting of 12,151 individual gRNAs targeting the same 206 Kb region and assayed HPRT function in HAP1 cells as previously. Under the assumption that each individual gRNA potentially disrupts a ~10 bp region, this experiment at best interrogates ~70% of bases within the 206 Kb region due to the sparsity of PAM sites (as compared to our coverage of the entire locus at median ~27-fold redundancy per base-pair with ScanDel). 86% of exon-targeting gRNAs were positively selected and exonic selection scores were well correlated between biological replicates (Pearson: 0.781). Of 612 negative control gRNAs, none that were sampled in each replicate were positively selected in both experiments (**Figure 4.11**). In non-coding sequence, scores were poorly correlated between biological replicates, with a paucity of reproducible, positively selected signal (Pearson: 0.156, **Figure 4.12**).

Notably, we did observe a greater proportion of positively scoring gRNAs in the vicinity of exons – *i.e.* whereas only 2% of intergenic gRNAs were positively selected, 7.5% of deep intronic (>2 Kb away from an exon boundary) and 20.5% of proximal intronic (<2 Kb from an exon boundary) gRNAs were positively selected (**Figure 4.13a**). Given our earlier observation with ScanDel of rare, 'on-target-but-with-incorrect-boundaries' that were confounding when targeting near exon boundaries, we next performed similar validation experiments on individual gRNAs that targeted non-coding sequences nearby exons (**Figure 4.13b**). We chose 10 gRNAs in

the *HPRT1* promoter region (**Figure 4.13c**), and repeated the individual gRNA experiment with a small pool of just these 10 gRNAs, again using long reads (Pacific Biosciences) to sequence the locus before (**Figure 4.13d**) and after 6TG selection (**Figure 4.13e**). Similar to our results with ScanDel in this region, the only mutations that survived 6TG selection were initially rare deletions whose boundaries extended past the TSS and into the 5' UTR and/or coding sequence (**Figure 4.13d**). This result strongly underscores that caution should be exercised in the interpretation of results from CRISPR-based screens of non-coding regions, whether performed with individual gRNAs or gRNA pairs, and the importance of sequencing-based validation of edited regions in the context of such screens.

## 4.4 DISCUSSION

We developed a method that uses CRISPR/Cas9 and pairs of gRNAs to experimentally test the functional consequences of thousands of programmed, kilobase-scale genomic deletions in a single experiment. We applied this method to perform the systematic investigation of the regulatory architecture of a housekeeping gene via editing of the endogenous genome. Upon introducing a set of densely tiling deletions spanning a 206 Kb region centered on the gene *HPRT1*, we found no evidence for any distal regulatory element that is critical for its activity, as measured by 6TG sensitivity in HAP1 cells. A screen of this same region with individual gRNAs supported this finding. The dearth of positive selection from disruption of non-coding regions contrasts with the strong positive selection observed from disruption of any exon of *HPRT1*, either by programmed deletions or individual guides.

*HPRT1* is a widely expressed housekeeping gene with no eQTLs identified by the Genotype-Tissue Expression Project (The GTEx Consortium, 2015), and thus may not require multiple (or any) distal regulatory regions for its expression. The simplest explanation of our

results is that sequences immediately proximal to the *HPRT1* transcriptional start site may be sufficient to confer the level of expression that provides sensitivity to 6TG, such that even if we disrupt distal regulatory elements that subtly modulate expression, they would go undetected by our strong selection. For future applications of ScanDel, implementing more quantitative readouts will be critical. For example, ScanDel is compatible with any functional selection that reliably separates cells on the basis of gene expression (*e.g.* knocking in GFP to a locus of interest, and then using FACS to stratify ScanDel-edited cells on the basis of expression). Such quantitative readouts may facilitate validation of the many candidate regulatory elements (and cognate target gene assignments) nominated by eQTL and functional genomics studies (Kumasaka et al., 2016; Won et al., 2016). We anticipate that the application of ScanDel to non-housekeeping genes coupled to a more quantitative readout will likely identify more regulatory elements than found for *HPRT1*, especially for genes that play key roles in development and cell fate determination.

Another possibility, albeit an unlikely one, is that critical regulatory elements for *HPRT1* lie outside of the 206 Kb window that we surveyed. For example, the gene resides at the terminus of a ~300 Kb topologically associated domain identified in HAP1 cells that spans ~185 Kb beyond our interrogated region (Sanborn et al., 2015). This could potentially be addressed by increasing the complexity of the library of programmed deletions in order to densely tile a larger region, or by simply increasing the size of each programmed deletion to interrogate more sequence per gRNA pair.

We note that the paucity of regulatory sequences discovered by CRISPR/Cas9-based screening is not exclusive to this study. Collectively, individual gRNA CRISPR/Cas9 screens have surveyed over a megabase of prioritized non-coding sequences, but only a handful of gRNAs tested have robust phenotypic effects that validate (Canver et al., 2015; Chen et al., 2015; Diao et al.,

2016; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2014). One explanation is that the assays being used are insufficiently sensitive and fail to detect modest regulatory effects. This could be addressed through the implementation of more quantitative assays.

A second explanation is that as implemented, genome editing has poor sensitivity due to redundancy in mammalian gene regulation. Redundancy of transcription factor binding sites within enhancers could prevent ~1-10 bp indels introduced by individual gRNAs from sufficiently disrupting function. Indeed, this was part of the motivation for developing ScanDel, whose programmable kilobase-scale deletions exceed the size of enhancers. Although we did not identify distal enhancers, the essentiality of the TSS and portions of the 5'UTR in our assay was detected primarily by deletions substantially larger than 1-10 bp (**Figure 4.13d,e**), suggesting paired gRNA libraries will be effective for enhancing sensitivity. However, there may also be redundancy amongst sets of distal regulatory elements, a question which can only be fully addressed by combinatorial perturbations.

A third explanation is that gene expression levels depend in part on historical events, such that disruption of an enhancer in a differentiated cell line would not result in the same outcome as disrupting the same enhancer prior to differentiation. This could be potentially addressed by performing lentivirally-mediated genome editing steps in stem cells, followed by differentiation to a cell type of interest. Any differences in functional consequences that are dependent on the timing of mutation would be of great interest.

Our results also provide a cautionary example of the importance of validation by direct genotyping in the context of CRISPR/Cas9-based screens of non-coding sequences. NHEJ generates a wide assortment of mutations, and strong selections may recover rare editing outcomes. For example, whereas targeting regions adjacent to exons might have been interpreted to reflect

the presence of critical proximal regulatory elements, validation experiments using a long-read sequencer showed that this signal was caused by rare deletions that extended into exonic sequence. Many of these unexpected events would have been difficult to detect had we been relying solely on a short-read sequencing platform to genotype editing outcomes. Additionally, validating CRISPR/Cas9-based screens by assessing selection for specific edited haplotypes adds biological information. Here, with long-read genotyping we were able to identify a set of variable deletions that either did or did not drive selection, thus enabling greater resolution (**Figure 4.10c,d**).

We also note that in experiments relying on pairs (or more) of gRNAs to program deletions, it is critical to include controls that quantify the effects of the individual gRNAs comprising these pairs, as these can have direct effects or off-target effects that might be misinterpreted as being consequent to the programmed deletion. While this work was being completed, a study was published that similarly used gRNA pairs to program deletion of a large number of lncRNAs, followed by phenotyping for cellular growth. Although the results are of great interest, these important controls were not included for the vast majority of spacers used. It will also be important to confirm the validity of each of this screen's findings through direct genotyping.

Even with the aforementioned open questions and remaining technical hurdles, it is critical that we continue to advance and apply methods for multiplex perturbation of the regulatory landscape with genome editing. The importance of experimental perturbation is highlighted by our results. The non-coding region surrounding *HPRT1's* first exon resides in open chromatin in this cell line (**Figure 4.3**, **Figure 4.8**), yet our results with ScanDel and subsequent validation experiments indicate the essential regulatory region is only a small part of the broader ATAC-seq peak. Perturbing the endogenous genome represents a highly complementary approach to the more classic strategy of reporter assays (Banerji et al., 1981; Patwardhan et al., 2009), in which short

sequences are tested for their regulatory potential on an episomal vector. Of note, the results of early reporter assay-based tests of potential regulatory sequences flanking *HPRT1* are largely consistent with our findings but also identify three sequences immediately proximal to the first or second exons that are critical for episomal *HPRT1* expression (Reid et al., 1990; Rincón-Limas et al., 1991). Though this discrepancy could be due to cell type or species differences (as two of these elements were required only in mouse embryonic stem cells but not human cells (Reid et al., 1990), and the remaining one was only tested in Chinese hamster fibroblasts (Rincón-Limas et al., 1991)), it could also be due to differences in regulatory element activity when assayed via episomes versus genome editing. For example, elements necessary to drive expression of a gene on a plasmid may not be required in the genome, where redundancy is more likely. This underscores the ongoing challenge that genome editing can address: understanding how short sequences with regulatory potential coordinate with one another across endogenous loci to give rise to specific levels of expression.

In summary, ScanDel enables the multiplex characterization of the functional consequences of thousands of programmed, kilobase-scale deletions to the endogenous genome in a single experiment. We applied ScanDel to *HPRT1,* a housekeeping gene in which disruptive mutations cause Lesch-Nyhan syndrome, introducing densely tiled 1-2 Kb deletions across a 206 Kb region encompassing the gene, covering each base-pair with median ~27-fold redundancy. Our results demonstrate the absence of distal *cis*-regulatory elements in this region that are critical for *HPRT1* expression. In the future, we anticipate that large-scale perturbation of putative regulatory elements in their endogenous context with methods such as ScanDel will provide further insights into gene regulation and the contribution of non-coding mutations to human disease.

## 4.5 METHODS

### 4.5.1 *Tissue culture*

HAP1 cells were purchased from Horizon Discovery and cultured in Iscove's Modified Dulbecco's Medium with L-glutamine and 25 mM HEPES (Gibco). The HAP1 cell line was derived from the near-haploid KBM7 line (male cells of chronic myelogenous leukemia origin) by introduction of induced pluripotent stem cell factors. Despite the cell line's male origin, HAP1 cells no longer hold a Y chromosome (Essletzbichler et al., 2014b). HEK293T cells were purchased from ATCC and cultured in Dulbecco's Modified Eagle's Medium with high glucose and sodium pyruvate (LifeTechnologies). Both media were supplemented with 10% Fetal Bovine Serum (Rocky Mountain Biologicals) and 1% Penicillin-Streptomycin (Gibco), and grown with 5% $CO_2$ at 37° C.

### 4.5.2 *gRNA library design*

To generate a list of gRNAs, we identified all 20 bp protospacers followed by a 5'-NGG PAM sequence from chrX:133,507,694-133,713,798 (hg19). We then excluded protospacers that had a perfect sequence match elsewhere in the genome, and scored the remaining gRNAs for both on-target and off-target activity. We considered off-target sequences that had five or fewer mismatches to the putative gRNA, and calculated an aggregate off-target score using the method of (Hsu et al., 2013). In addition we scored each site for on-target efficiency (Doench et al., 2014). Final deletion pairs were matched using spacers that did not contain BsmBI restriction sites, were not predicted to have off-target hits in other 6TG resistance genes or in KBM7 essential genes (the HAP1 parental cell line), were greater than 25 bp apart, further than 50 bp from an exon, and passed on-target (above 10) and off-target (above 25) thresholds. Contrastingly, the individual

gRNA library included all of the spacers targeting the same region, excluding those predicted to have 2,000 or more off-targets or to have off-targets with 4 or fewer mismatches within the targeted *HPRT1* region.

### 4.5.3 *Building the gRNA pair library*

This library cloning method was developed in parallel to similar recently published methods (Aparicio-Prat et al., 2015) and is modified from the GeCKO single gRNA cloning scheme (Sanjana et al., 2014; Shalem et al., 2014). First, the lentiGuide-Puro backbone (Addgene #52963) is digested with BsmBI (FastDigest Esp3I, Thermo) and gel purified. The paired spacers (flanked with lentiGuide-Puro overlap sequences) are synthesized twice on a microarray (CustomArray, Inc.) such that each pairing is represented in both possible orders (**Figure 4.4**).

To ensure quality of array synthesis, 1 ng of the oligo pool was amplified with Kapa HiFi Hotstart ReadyMix (KHF, Kapa Biosystems) and run on a gel to confirm oligos are of the expected 108 bp length. After PCR purification with Agencourt AMPure XP beads (Beckman Coulter), the amplicon is cloned into lentiGuide-Puro using In-Fusion HD Cloning Plus (Clontech) and transformed into Stable Competent *E. coli* (NEB C3040H) to minimize repeat-based recombination of the lentivirus. This ensuing library (lentiGuide-Puro-2xSpacers) now contains each pair of spacers, but is still missing the additional gRNA backbone and PolIII promoter.

We next cloned in the additional gRNA backbone and H1 promoter between each spacer pairing to enable expression of the two independent gRNAs. The gRNA backbone-H1 promoter fragment was ordered as a gBlock (IDT) with flanking BsmBI sites to allow ligation into the BsmBI-digested lentiGuide-Puro-2xSpacers library. The gBlock and the lentiGuide-Puro-2xSpacers are each digested with BsmBI, purified, ligated together with Quick Ligase (NEB

M2200S), and transformed into Stable Competent *E. coli* to create a final lentiGuide-Puro-2xgRNA library.

To prevent bottlenecking of the library, these cloning steps are performed with enough replicates at high efficiency to maintain a minimum of 20x average library coverage (relative to the expected library complexity). Sequencing of the lentiGuide-Puro-2xgRNA library revealed 97.8% retention of diversity from the designed paired spacers. However, 16% of library reads held unprogrammed, interswapped pairs. 88.5% of these swaps are only seen in a single read, implying a more likely cause is template switching during either PCR or cluster generation. For all experimental analysis, only reads of gRNA pairs that perfect matched programmed pairs were considered.

### 4.5.4 *Building the individual gRNA library*

The spacers of this library were similarly synthesized on an array, amplified, and purified as above. The lentiGuide-Puro backbone was linearized as above, and the library cloned into it using the NEBuilder HiFi DNA Assembly Master Mix (NEB). This plasmid was transformed into Stable Competent *E. coli*, generating enough transformants for 30x average coverage. This method produced 98.5% retention of complexity from the designed array.

### 4.5.5 *Lentiviral library production, delivery, and 6-thioguanine selection*

Lentivirus was produced using Lipofectamine 3000 (Life Technologies) to transfect HEK293T with the lentiviral vector libraries made above and 3rd generation packaging plasmids (pMDLg/pRRE Addgene 12251, pRSV-Rev Addgene 12253, pMD2.G Addgene 12259). Supernatant was collected 72 hours after transfection, centrifuged at 300 rcf for 5 minutes to remove cell debris, and passed through a 0.45 μm syringe filter.

To create a monoclonal HAP1 cell line stably expressing Cas9, HAP1 cells were transduced with lentivirus produced using lentiCas9-Blast (Addgene 52962), selected with 5 μg/mL Blasticidin (Thermo Fisher Scientific), and single-cell sorted via FACS.

HAP1-Cas9-Blast monoclonal cells were plated to be at 30% confluency on the day of lentiviral gRNA/pair transduction. To transduce, 5% of the recipient cells' media was replaced with filtered virus, limiting the MOI to < 0.3. Media was changed after 24 hours, and selection for transduced cells began 48 hours post-transduction. Puromycin was added at 2 μg/mL for two days to assess the percentage of cells transduced, and then cells were maintained in 1 μg/mL for 5 more days.

After puromycin treatment, an initial population of cells was collected. Selection for loss of HPRT function was performed by applying 5 μM 6TG to the remaining cells at <50% confluency for 7 days. An additional concern is that minor changes in gene expression caused by ScanDel-mediated mutations in regulatory elements will not be strong enough to confer resistance. To mitigate this, we used the lowest dosage of 6TG that completed HAP1 selection after seven days. 6TG concentrations of 6-60 μM are reported in the literature to achieve effective selection in this timeframe, depending on cell type (Jacobs and DeMars, 1984; Monnat et al., 1992). We tested our monoclonal HAP1-lenti-Cas9-Blast line at concentrations just below this range (1 μM, 2.5 μM, and 5 μM 6TG). After 7 days, the 5 μM treatment had no readily identifiable surviving cells, whereas the 2.5 μM treatment retained a sparse population, and the 1 μM treatment produced appreciably more outgrowing colonies. Based on these results, we proceeded with selections using 6TG at 5 μM for seven days. Enough cells were transduced and sampled at each timepoint to maintain minimum 2,000x average coverage of the library in each population.

Sequencing of the baseline (*i.e.* pre-6TG) population revealed 98.4% of diversity of the lentiGuide-Puro-2xgRNA library was preserved from replicate 1, and replicate 2 retained 78.8%. As our deletions are highly overlapping, we proceeded with replicate 2 as all base-pairs are interrogated despite the lower diversity. We observed 95.6% retention of programmed library diversity in replicate 1 of single gRNA plasmid library and 71.2% of replicate 2.

Interswapped gRNA pairs were observed in 35.5% of reads from the baseline pre-6TG sample. This is an increase from the 16% observed in reads from the lentiGuide-Puro-2xgRNA plasmid library. This suggests additional template switching during the library's amplification from gDNA, which requires more cycles of PCR. However, since we are directly sequencing each gRNA spacer as a read out opposed to using barcoded libraries and only taking exact sequence matches, this does not pose a problem.

### 4.5.6 *gRNA library amplification and sequencing from HAP1 cells*

gDNA was extracted from the cells sampled before and after 6TG selection using the DNeasy Blood & Tissue kit (QIAGEN). KHF was used for all amplification steps. The libraries were initially amplified from a minimum of 6 µg of gDNA divided across thirty 50 µL reactions, ensuring sampling of ~2 million haploid genome equivalents at each timepoint. Two additional PCRs were performed to add sequencing adapters and sample indices to the amplicon, with AMPure bead purification between each reaction. Amplification conditions were optimized using qPCR to minimize overamplification of the construct.

Sequencing was performed on an Illumina Miseq using a 50-cycle kit. Read 1 and the Illumina Index read were used to sequence the two gRNAs in the paired gRNA construct prior to paired-end turnaround, and Read 2 was used to sequence the 9 bp sample index.

### 4.5.7 *Calculation of a selection score assignment per base-pair*

Custom Python scripts counted tallies of gRNAs (for individual gRNA library experiments) or gRNA pairs before and after selection. These counts were normalized to the total number of reads per sample. An enrichment ratio was calculated for each gRNA/pair by dividing its normalized read count after selection by its before selection read count. A selection score is the log10 of the enrichment ratio (log10(after/before)). If a gRNA or gRNA pair was absent before selection, it was excluded from further analysis. Any gRNA pairs that used a self-paired gRNA with an independent selection ratio > 0 were also excluded from further analysis.

If a gRNA/pair is absent after 6TG selection, its selection score as calculated will be a negative number relatively large in magnitude that is somewhat arbitrarily determined by the number of pre-selection reads. Thus, to limit the contribution of these scores to average measurements derived from many independent deletions, we set a minimum selection score equal to the middle of the bimodal distribution between the positively and negatively selected deletions of each replicate (**Figure 4.5**). For example, in ScanDel replicate 1, if the log10-value of a selection score was less than -0.35, that gRNA pair's score was set to -0.35. Each individual base-pair was assigned a per base-pair selection score by taking the mean of all deletions programmed to cover that base-pair. The per base-pair score was normalized to the median score for all positive scores in that replicate. The per base-pair selection score of each replicate was averaged to get the final selection score per base-pair. Per base-pair scores were uploaded as a bedgraph for visualization on the UCSC Genome Browser.

For the individual gRNA mutagenesis screen, we calculated selection scores per base-pair similarly, assuming a 10 bp deletion was made by each gRNA queried. If a base-pair was scored

at the minimum negative threshold in one screen, it was given that value for the consensus selection score of the two replicates.

### 4.5.8 *Bulk ATAC-seq of HAP1 cells*

Two biological replicates were separately maintained (on 10cm dishes, split 1:10 three times per week) and processed separately. Chromatin accessibility in the HAP1 cell line was profiled with the ATAC-seq protocol (Buenrostro et al., 2013) with slight modifications. The media for 10cm plates of confluent HAP1 cells was aspirated and replaced with 2 mL of ice cold lysis buffer ('CLB+'; made as described in the original paper, but supplemented with protease inhibitors (Sigma cat. no. P8340)). Cells were incubated on ice for 10 minutes in CLB+ and then were dislodged with a cell scraper and transferred to a 15 mL conical tube and pelleted at 500 rcf for 5 min at 4° C. Nuclei were re-suspended in 1 mL of CLB+ and counted on a hemocytometer. 50,000 nuclei in 22.5ul of CLB+ were combined with 2.5 μL of TDE1 enzyme and 25ul of TD buffer (Illumina). Tagmentation conditions were as described in the original paper (37° C for 30 min). After MinElute purification into 10 μL EB buffer (Qiagen), 5 μL of tagmented DNA was amplified in 25 μL reactions for 12 cycles using the NEBNext Master Mix (NEB). Reactions were monitored with SYBR Green to ensure that samples were not overamplified. PCR products were cleaned once with a QiaQuick PCR Cleanup Kit (Qiagen) and once with 1x AMPure beads (Agencourt). The quality of the library was assessed on a 6% TBE gel and the yield was measured by Qubit (1.0) fluorometer (Invitrogen).

Samples were sequenced on two paired-end Illumina NextSeq 500 runs. Read lengths were 2x75 bp for the first run and 2x151 bp for the second run, so the second run was truncated to 75 bp. Sequencing reads were also trimmed for read-through of adapter sequences and quality with Trimmomatic (Bolger et al., 2014) ('NexteraPE-PE.fa:2:30:10:1:true TRAILING:3

SLIDINGWINDOW:4:10 MINLEN:20' parameters) and then mapped to the 1000 genomes integrated reference genome 'hs37d5' with bowtie2 (Langmead and Salzberg, 2012), using the '-X 2000 -3 1' parameters. Only properly paired and uniquely mapped reads with a mapping quality above 10 were retained ('samtools -f3 -F12 -q10'). Reads mapping to the mitochondrial genome and non-chromosomal contigs were also filtered out. In addition, duplicate reads were removed with Picard. After checking QC metrics on the individual replicates, reads from the two libraries were combined for downstream analysis. Hypersensitive sites were called (at a 1% false discovery rate) with the Hotspot algorithm (John et al., 2011).

### 4.5.9 *Validation and direct genotyping of positive signal from the screens*

gRNA pairs that drove the ScanDel signal surrounding *HPRT1*'s first exon were cloned into simple lentiGuide-Puro-2xgRNA libraries. The TSS ScanDel validation library contained four pairs and the intron 1 library contained five. For the individual gRNA screen TSS library, ten gRNAs were cloned into lentiGuide-Puro. These constructs were lentivirally delivered to HAP1-Cas9-Blast cells, selected with 6TG, and gDNA extracted as described above.

As the expected deletions could remove up to two Kb, the loci were sequenced with a Pacific Biosciences RSII (University of Washington PacBio Sequencing Services, P6C4 chemistry, RSII platform). To prepare libraries for PacBio sequencing, the TSS- or intron 1-targeted regions were amplified from 800 ng of gDNA each, using four 50 μL KHF reactions with primers adding sample indices and SbfI or NotI cut sites. The purified amplicons (Zymo Research DNA Clean & Concentrator-5) were digested with SbfI-HF (NEB) and NotI-HF (NEB), leaving sticky ends. 5'-phosphorylated SMRT-bell hairpin oligos (IDT) containing the PacBio priming site, hairpin-forming sequence, and resulting sticky ends for either SbfI or NotI were annealed by heating to 85C and snap frozen in 10mM Tris 8.5, 0.1mM EDTA, 100 mM NaCl. These were

ligated at 10x molar excess to the digested amplicons, destroying the restriction site once attached. To remove undigested amplicons and primers, this ligation was performed in the presence of further SbfI and NotI, and followed by treatment with Exo7 (Affymetrix) and Exo3 (Enzymatics).

Only reads with over five circular consensus sequence passes and containing the expected first twelve 5' and 3' base-pairs of the amplicon were used for further analysis. Reads positive for complex inversions ( >99 bp) were removed from the library using the Waterman-Eggert algorithm with match, mismatch, gap open, and gap extend scores of 2, 10, 10, and 5, respectively (Döring et al., 2008). The resulting reads were then were aligned to the amplicon reference using the NEEDLEALL (Rice et al., 2000) aligner with a gap open penalty of 10 and a gap extension penalty of 0.5. Insertions were required to start within a window of five bases up or downstream of the putative cut site. Deletions were required to either start or end within the same 10 bp window or span the window. Reads that carried the same edit pattern were collapsed into haplotypes, and figures were generated using a custom D3 script.

### 4.5.10 *Comparing deletion rate of U6-H1 versus U6-U6*

Two protospacers were chosen to program a 365 bp deletion within the second intron of *HPRT1* and their spacers were cloned into a U6-H1 construct and U6-U6 construct (**Figure 4.2**). Virus was produced and delivered to cells, which were selected with puromycin, and gDNA extracted as described above. The locus was amplified in four successive rounds of nested PCR. The first reaction was only 3 cycles and included a forward primer with a 10-bp unique molecular index (UMI). The second reaction amplified any UMI-tagged fragments. The third and fourth reactions added sample indices and Illumina flow cell adapters. The products were AMPure cleaned between each reaction at a concentration that would lose primer dimer but retain the smaller deletion-holding fragments, and sequenced on a MiSeq. Any reads that

contained the same UMI or edit pattern were collapsed using custom scripts and their alignments were visualized with the same D3 script as above.

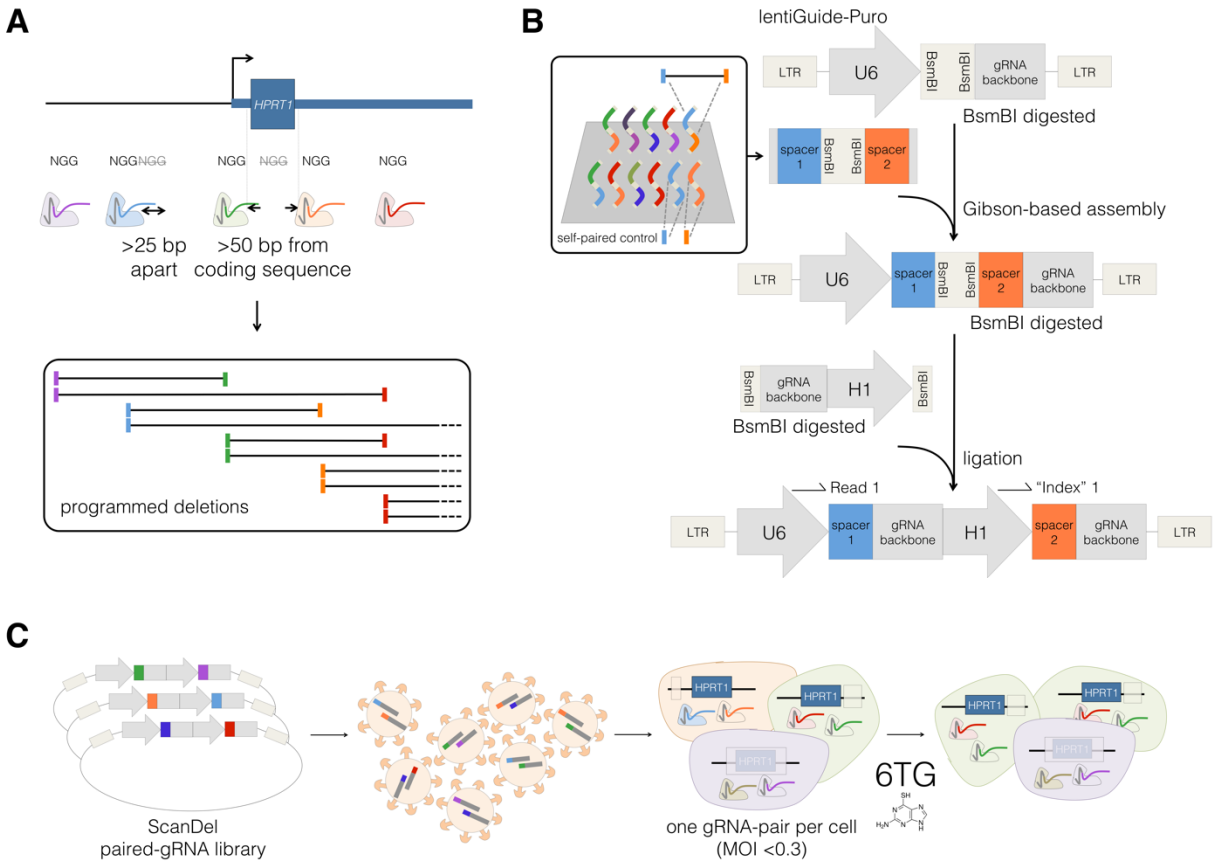



**Figure 4.1. Design, delivery and selection of ScanDel library of CRISPR/Cas9 programmed deletions for identification of non-coding regulatory elements.**

**a**, gRNA pairs are designed from a filtered set of protospacers from all Cas9 PAM sequences (5'-NGGs) in the *HPRT1* locus. Sites that are >25 bp apart and >50 bp away from exons are kept. Tiles are designed by pairing each remaining spacer to two downstream spacers targeting sequence ~1 Kb away and ~2 Kb away. This results in high redundancy of independently programmed, overlapping deletions across the locus. **b**, All spacer pairs that correspond to programmed deletions are synthesized on a microarray (*inset*). Each spacer is also synthesized as a self-pair as a control for its independent effects. If a self-paired spacer scores positively in the screen, any pairs that use that spacer are removed from analysis. U6 and gRNA backbone sequence flank the spacer pairs for Gibson-mediated cloning into lentiGuide-Puro, and mirrored BsmBI cut sites separate the spacer pairs to facilitate insertion of a second gRNA backbone and the H1 promoter (*beige*). In the final library, each gRNA is expressed from its own PolIII promoter. This design facilitates PCR and direct sequencing-based quantification of gRNA pair abundances. **c**, The lentiviral library of gRNA pairs is cloned at a minimum of 20x coverage (relative to library complexity) and transduced into HAP1 cells stably expressing Cas9 (via lentiCas9-Blast) at low MOI. After a week

of puromycin selection, the cells are sampled to measure the baseline abundance of each gRNA pair. The final cell population is harvested after a week of 6-thioguanine (6TG) treatment, which selects for cells that have lost HPRT enzymatic function. The prevalence of each programmed deletion is quantified by PCR and deep sequencing of the gRNA pairs before and after selection.

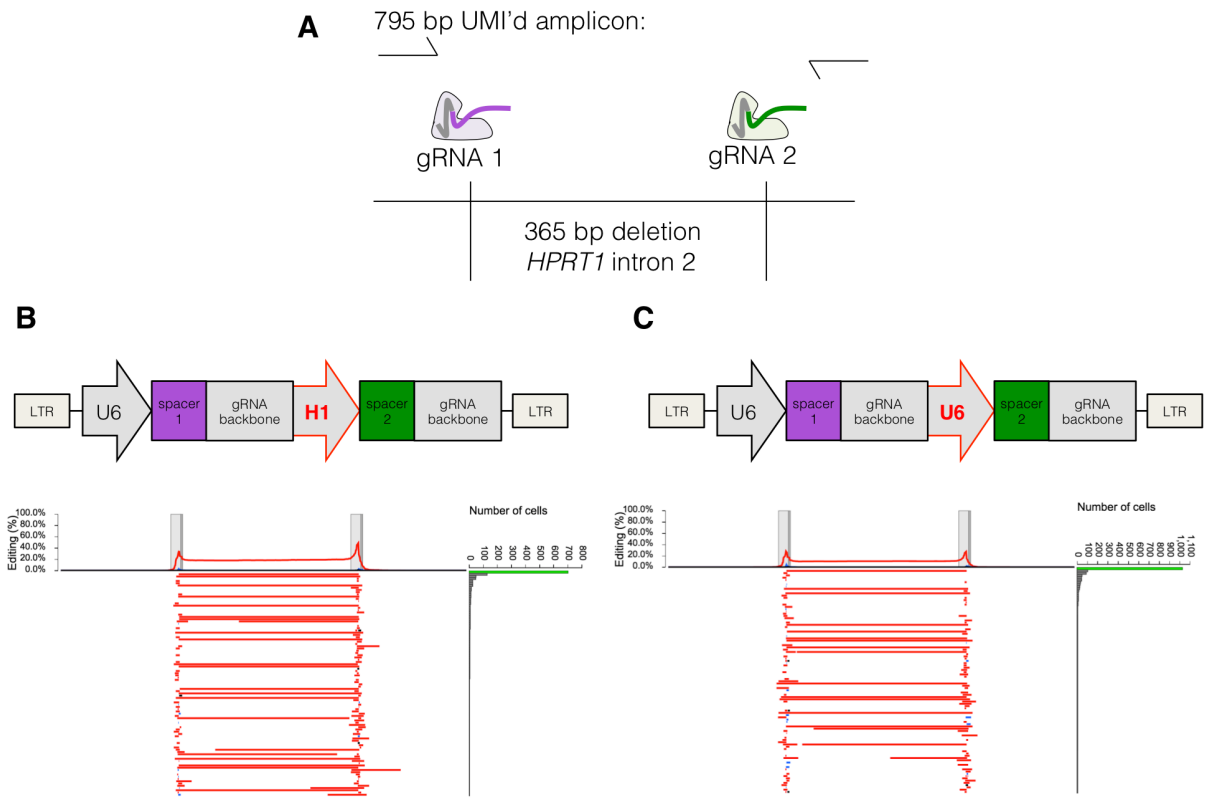



**Figure 4.2. The U6-H1 gRNA pair expression construct induces a higher deletion rate.**

**a**, Two spacers were chosen to program a 365 bp deletion within the second intron of HPRT1. To test deletion efficiency, virus was made from the constructs depicted in **b** and **c**, and separately transduced into HAP1 at MOI < 0.3. Following 1 week of puromycin selection, gDNA was extracted and the targeted region amplified. The first 3 cycles of this PCR contained a forward primer with a unique molecular tag (UMI) to track reads from the same original cell. Sequencing was performed on a MiSeq. Of note, PCR bias for smaller deletion-holding amplicons was reduced by collapsing reads with the same UMI, but the potential remains for higher clustering efficiency of the shorter amplicons. **b**, The spacers for the deletion in **a** were placed behind either a U6 or H1 PolIII promoter. 20% of sampled haplotypes contained the programmed deletion, but 36% of sampled haplotypes remained unedited, implying longer editing time could result in a higher deletion rate. The per base-pair editing rate summed across all sampled haplotypes is charted as a percentage at top, and the top 100 most prevalent haplotypes are displayed below it. Red indicates deletions and blue insertions. **c**, The spacers for the deletion in **a** were each placed behind a U6 PolIII promoter, and delivered, sampled, and visualized as above. With this expression construct, 10% of sampled haplotypes contained the programmed deletion.

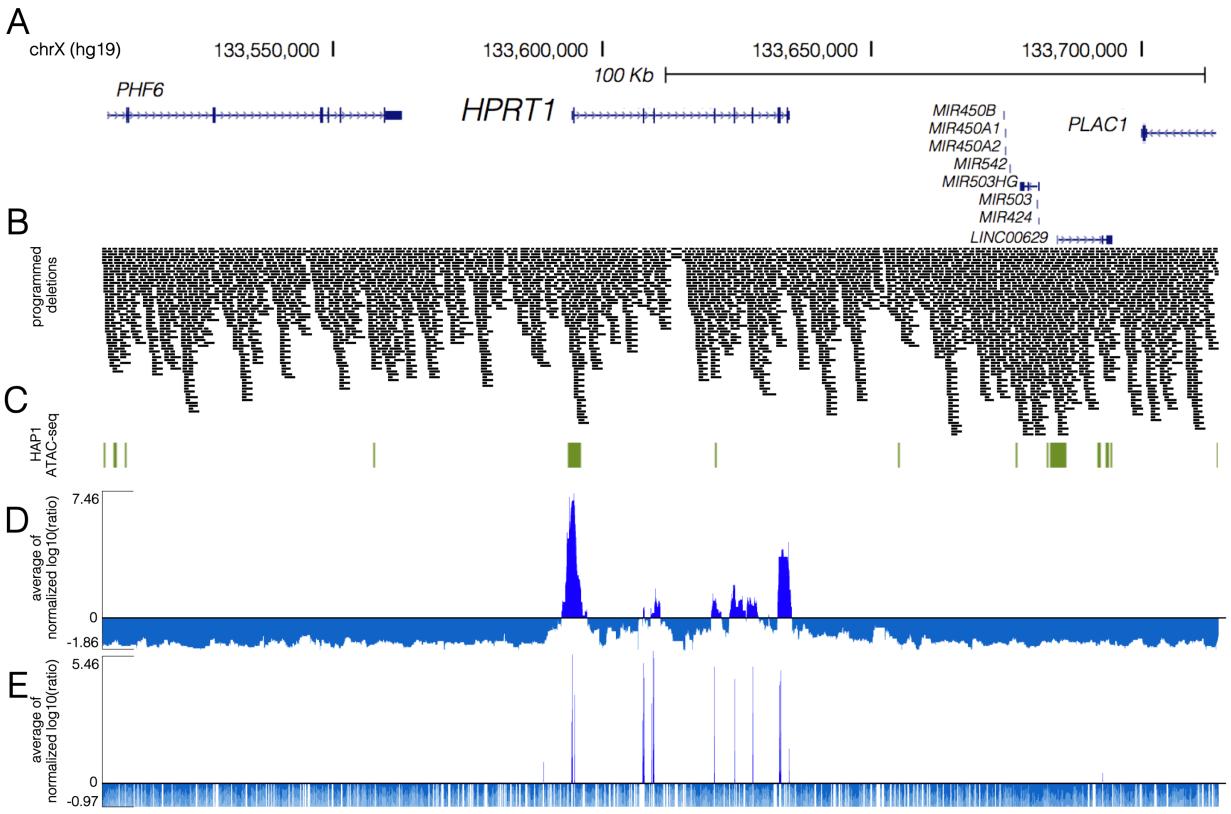



**Figure 4.3. High coverage ScanDel library across the *HPRT1* locus reveals a paucity of critical distal regulatory elements.**

**a,** Deletions were programmed across 206.1 Kb of the *HPRT1* locus and its surrounding sequence (chrX:133,507,694-133,713,798, hg19, UCSC Genes track in blue). **b,** A total of 4,342 x 1 Kb or 2 Kb deletions were programmed, tiling across the locus such that each base-pair was interrogated by a median of 27 independently programmed deletions. A high density of repeat elements results in reduced coverage of a region within *HPRT1*'s intron 3. Deletions are visualized as black bars spanning the gRNA pair's programmed cut sites. **c,** HAP1 ATAC-seq hotspots (green) indicate regions of open chromatin in the cell line. Of note, a hotspot extends 600 bp upstream and 1.6 Kb downstream of exon 1. **d,** ScanDel scores were assigned to each base-pair as the average of all selection scores log10(after/before) for gRNA pairs that programmed deletions that span that base-pair. If a gRNA pair used a spacer that was positively selected on its own as a self-pair, the gRNA pair was removed from analysis. To avoid over-weighting negative values, a minimum score was determined from each replicate's gRNA pair score distribution, and scores below it were set at this minimum. For each biological replicate, the base-pair's score was normalized to the replicate's median of positive scores. The average of the two biological replicates' normalized scores for that base-pair is displayed, with positive scores in royal blue and negative scores in blue-grey. **e,** An individual gRNA mutagenesis screen of the same region was also performed covering only ~70% of bases in the region due to the sparsity of high-quality designable spacers. Individual base-pairs

were scored based on nearby cut-sites, under the assumption that each gRNA queries a ~10 bp region. The plotted scores were calculated as in **d**, with positive scores in royal blue and negative scores in blue-grey.

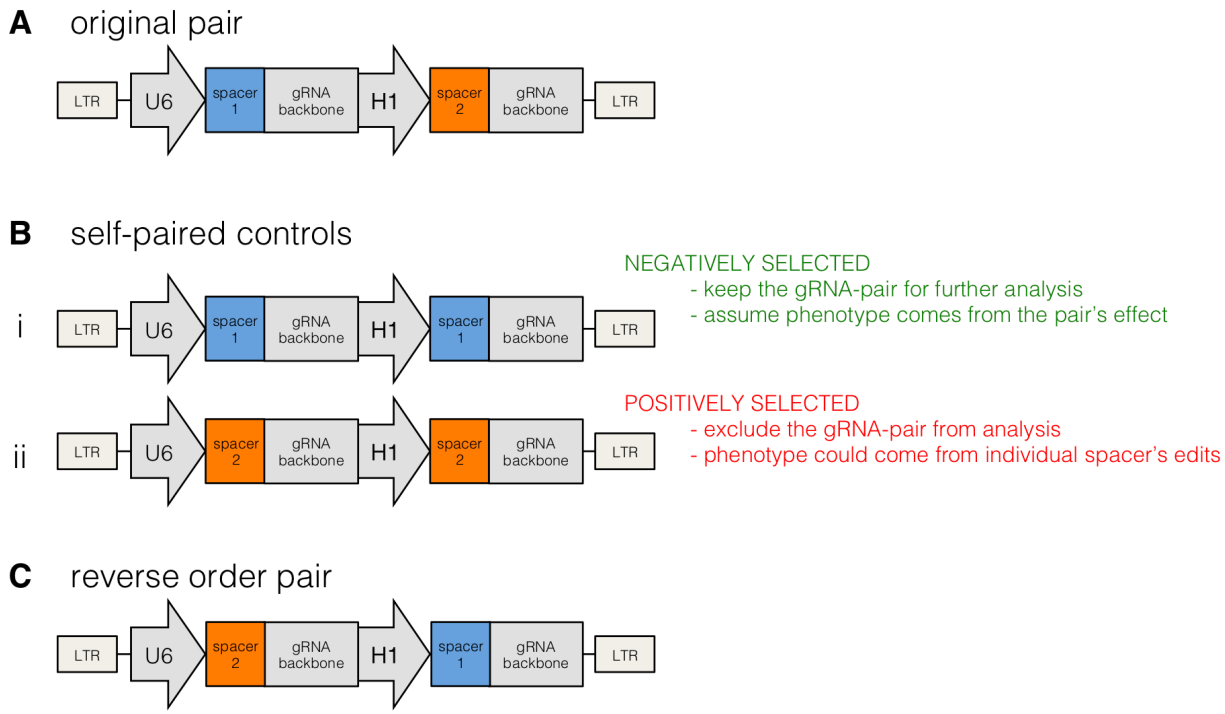



**Figure 4.4. Self-paired spacers in the ScanDel library reveal phenotypes independently created by individual spacers.**

**a**, The spacers used in every designed gRNA pair had their own self-paired control included in the programmed gRNA pair library. **b**, The self-paired controls consisted of the exact same spacer included behind each promoter in the expression construct (two for each pair; (i) and (ii)). If a self-paired spacer was positively selected, any gRNA pairs that included that spacer were excluded from further analysis. This avoided any confounding effects of alternative repair outcomes that result from an individual gRNA's edit that could cause 6TG resistance (e.g. a ~10 bp indel disrupting a transcription factor binding site, or disrupting an off-target locus that affects 6TG resistance, or an individual gRNA inducing translocations of HPRT1 at a high rate). By excluding these gRNAs, we can more confidently attribute observed phenotypes to programmed deletion induced by the gRNA pairs. **c**, Each gRNA pair was included in both possible orderings on the microarray. This was intended to minimize the impact of differences between the promoters, as well as to increase the chance that each deletion will be represented in the library, as synthesizing each pair twice reduces loss due to synthesis errors and cloning bottlenecks.

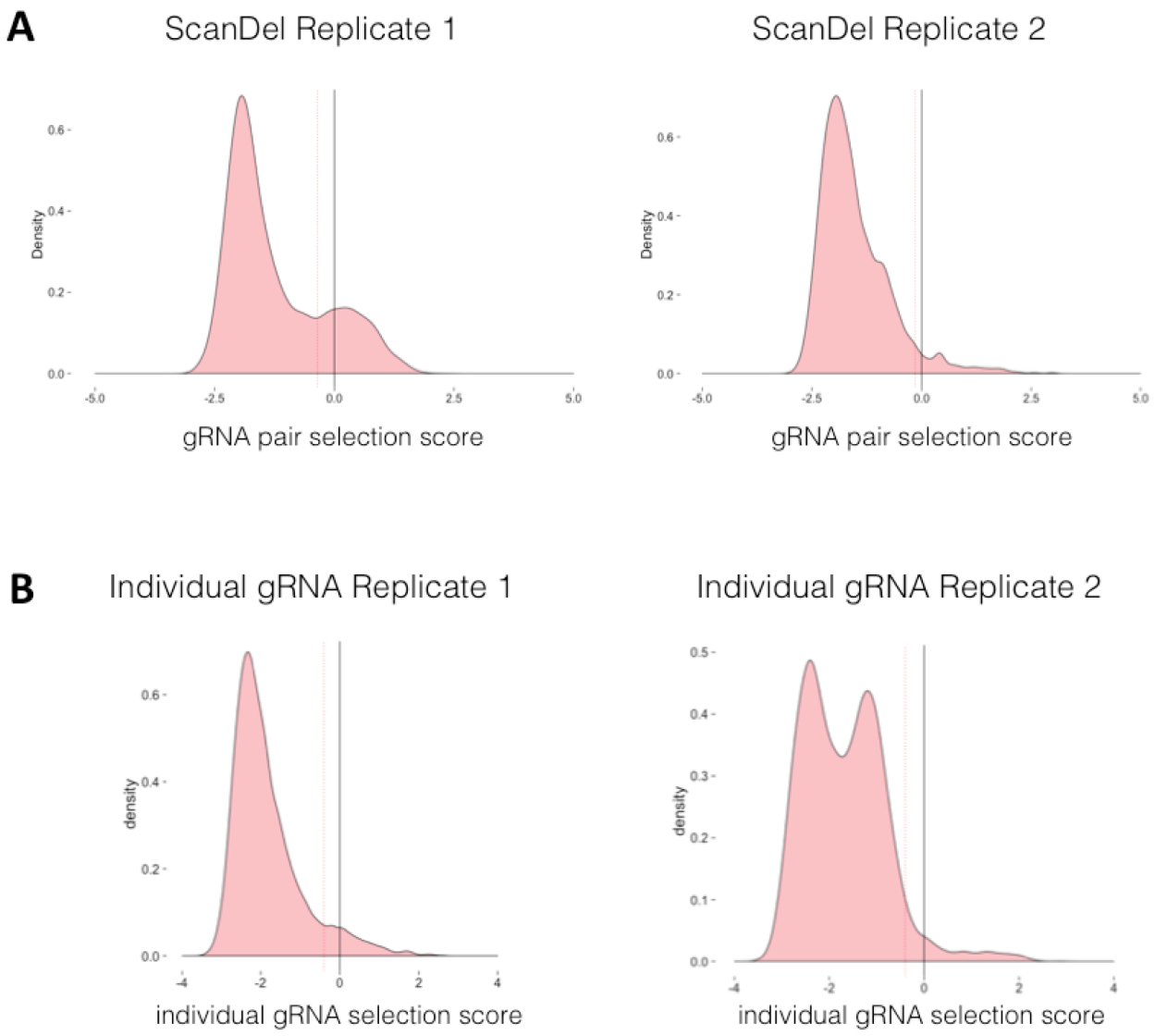



**Figure 4.5. Distribution of selection scores across biological replicates for ScanDel gRNA pairs or individual gRNAs.**

**a**, Each gRNA pair in the ScanDel screens was assigned a selection score equal to log10(after/before 6TG). The minimum selection score threshold described in Methods (-0.35 for replicate 1, -0.15 for replicate 2) is drawn with a dotted red line. **b**, Each gRNA in the individual gRNA screen was assigned a selection score as in **a**, for each replicate. The minimum negative selection score threshold (-0.4 for both replicates) is drawn with a dotted red line.

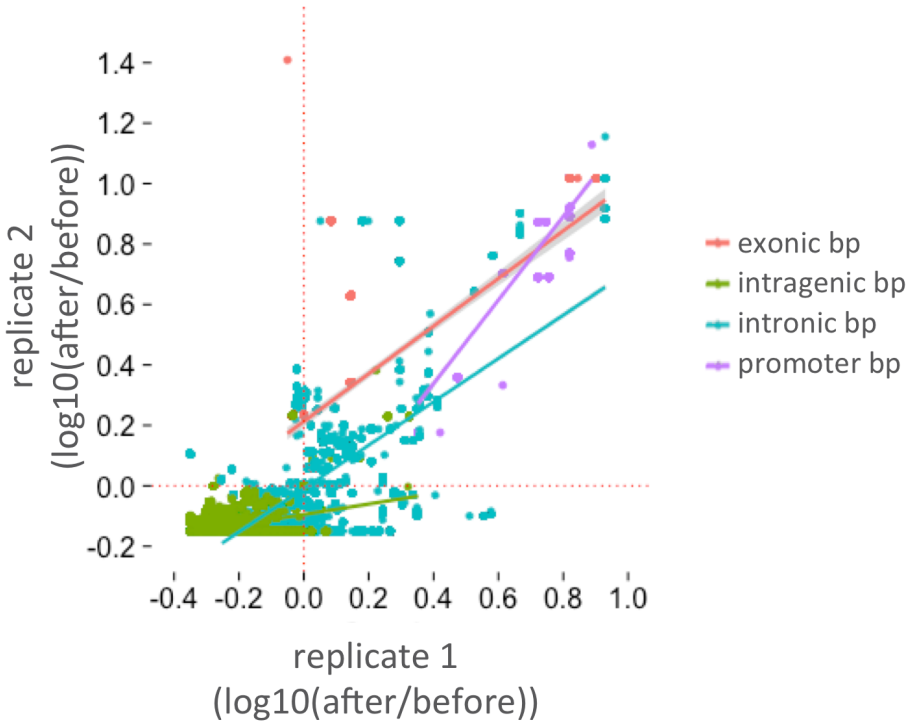
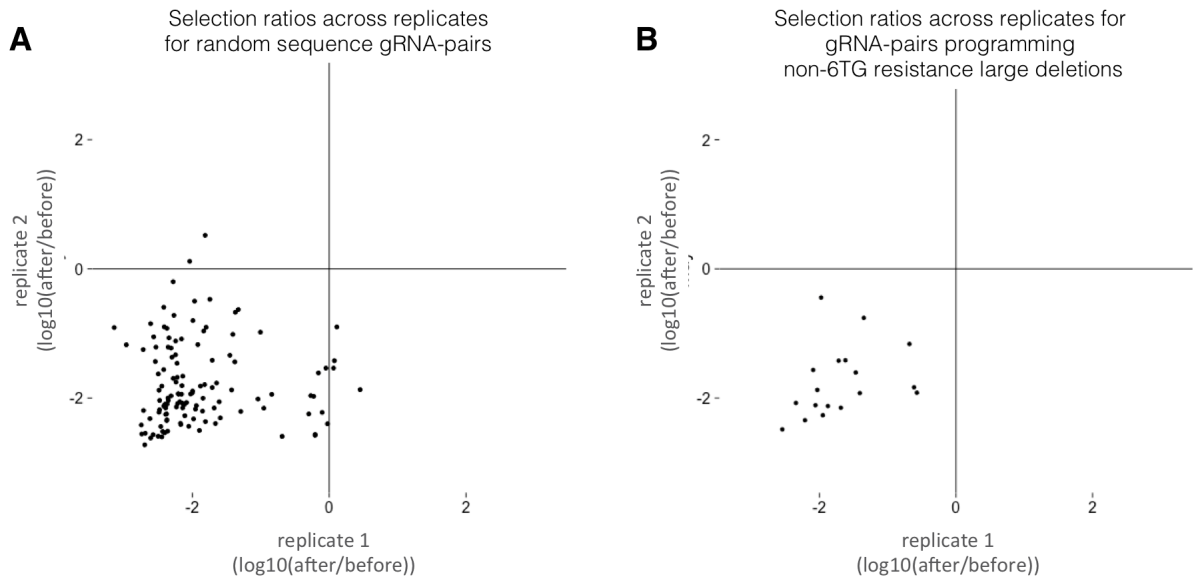



**Figure 4.6. ScanDel scores correlate across two biological replicates.**

The ScanDel selection scores for each biological replicate were calculated per base-pair by averaging the log10(after/before 6TG) for every programmed deletion that covers that base-pair. Least squares lines and points are colored by sequence content category. The stronger correlation for the 'intronic' category is driven by sequences proximal to the exons. Red corresponds to exons (Pearson: 0.736); green to intragenic regions (Pearson: 0.417); blue to intronic regions (within 2 Kb of an exon, Pearson: 0.628; deeply intronic, Pearson: -0.0194); and purple is the promoter (1 Kb upstream of the TSS, Pearson: 0.905).




| C | positively selected in both | negatively selected in both | positively selected in rep 1 | negatively selected in rep 1 | positively selected in rep 2 | negatively selected in rep 2 |
|---|---|---|---|---|---|---|
| random sequence not found in hg19 (**A**) | 0 | 236 | 8 | 256 | 4 | 244 |
| deletions targeting non-6TG resistance regions (**B**) | 0 | 18 | 1 | 31 | 0 | 18 |

**Figure 4.7. None of the negative control gRNA pairs were positively selected by 6TG in both ScanDel replicates.**

**a**, Negative control gRNA pairs targeting random sequences not found in hg19 were given a selection score of log10(after/before 6TG). Only gRNA pairs sampled in both replicates are plotted. **b**, Additional negative control gRNA pairs were programmed to create 1 and 2 Kb deletions in regions not expected to cause 6TG resistance. Selection scores were calculated for each gRNA pair as in **a**, and plotted for gRNA pairs found in both replicates. These region's coordinates were randomly generated from poorly conserved sequence 1 not within 10 Kb of any gene and far from HPRT1 (chr8:23768553-23771053, chr4:25697737-25700237, chr9:41022164-41024664, chr5:12539119-12541619, chr6:23837183-23839683, chr8:11072736-11075236). **c**, Table showing counts of positively and negatively selected negative control gRNA pairs across experiments.

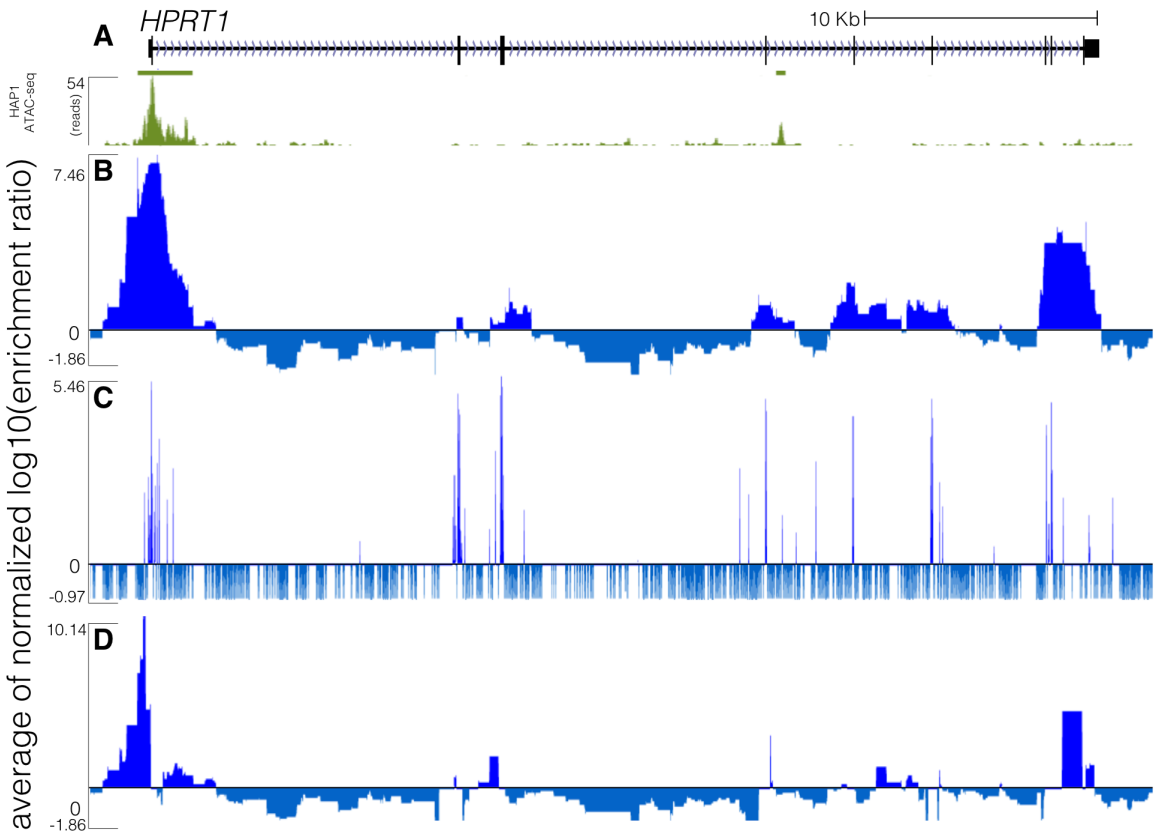



**Figure 4.8. All exons and some exon-proximal non-coding regions score strongly in both the ScanDel gRNA pair screen and the individual gRNA screen.**

**a**, ATAC-seq data (green) from the HAP1 cell line displayed for the HPRT1 locus (chrX:133,591,675-133,637,198, hg19). Bars depict hotspots 2 and beneath is the pile-up representation of ATAC-seq reads. **b**, The same ScanDel data is displayed as in **Figure 4.3c** but zoomed-in on the HPRT1 locus. Each base-pair's score is the mean of the log10(after/before 6TG) values for all the programmed deletions that cover that base-pair. These scores are normalized to the median positive score from the replicate. The average of the two replicates' scores for each base-pair is displayed. **c,** The same individual gRNA data is displayed as in **Figure 4.3d** but zoomed in on HPRT1. Each base-pair score is the mean of the log10(after/before 6TG) values for all the inferred 10 bp deletions that remove that base-pair, normalized as above. **d**, The same ScanDel track as in **a** but with per base-pair scores calculated after excluding any deletions programmed to disrupt an exon.

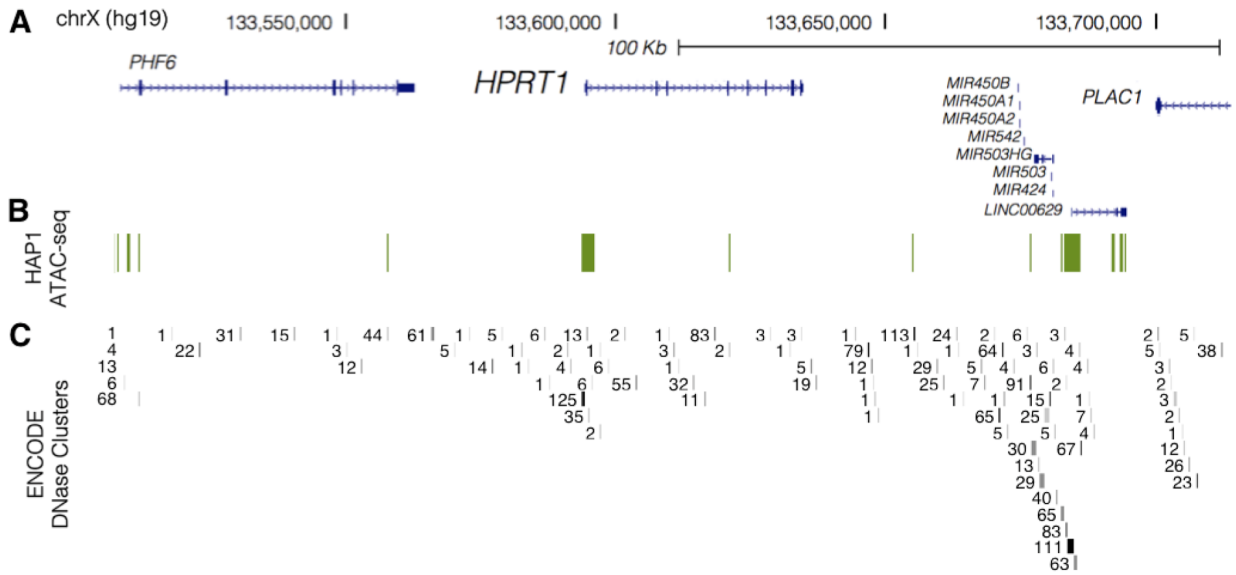



**Figure 4.9. HAP1 chromatin accessibility near *HPRT1*.**

**a**, To assess HAP1's suitability as a model in which to study the ubiquitously expressed HPRT1, regions of accessibility were compared across HAP1 and 125 ENCODE cell types. The 206.1 Kb encompassing HPRT1 and its surrounding sequence interrogated by this screen (chrX:133,507,694-133,713,798, hg19, UCSC Genes track in blue). **b**, Regions of open chromatin in HAP1 cells (green) as profiled by ATAC-seq. **c**, Clusters of DNase accessibility peaks across 125 cell lines assayed by the ENCODE. Each accessible region is labeled with the number of cell lines in which it is detected. Though there are many cell-type specific peaks, the HAP1 open chromatin regions match sites commonly accessible across many cell lines.

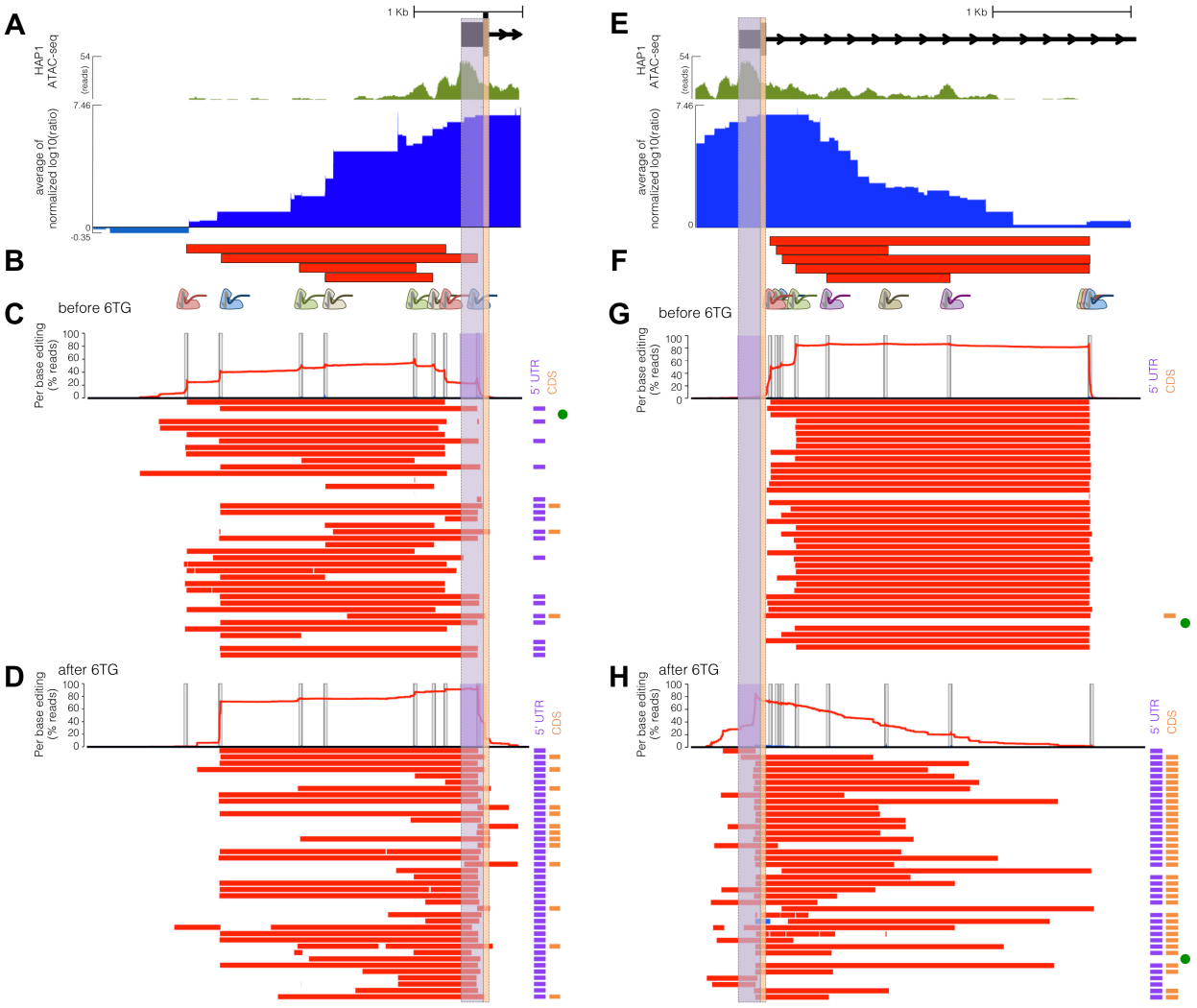
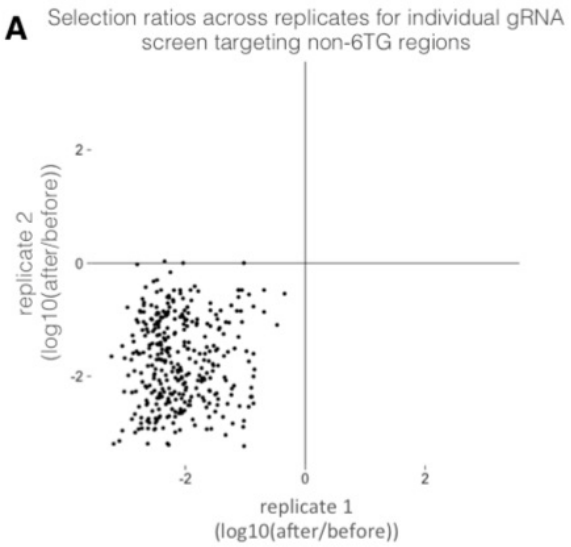



**Figure 4.10. Long-read sequencing reveals rare, unprogrammed, exon-interrupting deletions that drive selective effects.**

**a,** A putative promoter is implicated by open chromatin (HAP1 ATAC-seq broad peaks, green) surrounding *HPRT1*'s first exon (UCSC Genes, black). ScanDel signal in the 2 Kb upstream of *HPRT1* also suggests the possibility of critical regulatory sequences in this region (blue**,** chrX:133,591,603-133,594,626, hg19**)**. The 5' UTR and coding regions of exon 1 are highlighted in purple and orange, respectively. **b,** Four gRNA pairs targeting the promoter were cloned as a small pool, delivered, and selected with 6TG to enable sequencing of the edited locus (programmed deletions displayed as red bars). A 3 Kb region was amplified and sequenced with long reads (Pacific Biosciences). **c,** The chart at the top displays per-base percentages for deletions (red) or insertions (blue), with target sites indicated by vertical gray bars. Horizontal bars show the edits found on each haplotype ranked by decreasing prevalence. All programmed deletions are abundant before 6TG treatment, in addition to rare, unexpected deletions. The notations to the right indicate if the edits interrupt the TSS/5'UTR (purple bar) and/or coding sequence (orange bar). The unedited haplotype is marked with a green dot. Of note, PCR and sequencing on the PacBio

RSII are biased towards smaller fragments, limiting accurate quantitative comparison of read counts from differently sized edits. **d,** Haplotypes from 6TG-selected cells are plotted as in **c** revealing that only edits that interrupt the TSS/5'UTR survive selection, with no programmed or 'promoter only' deletions surviving selection. **e,** Open chromatin (green) and ScanDel signal suggests the presence of critical non-coding regulatory sequences in the first ~2.7 Kb of intron 1 (chrX:133,593,871-133,596,998, hg19). **f,** 5 gRNA pairs that drove the signal in this intronic region were cloned and 6TG selected as a small pool, as in **c**. **g,** A 3.1 Kb region spanning the 5'-most part of intron 1 was amplified and sequenced from cells sampled before 6TG selection. Haplotypes and per-base editing rates are diagrammed as in **c**. **h,** Post-6TG selection haplotypes from the intron 1-targeted cells are plotted as in **g** revealing that the vast majority of surviving edits disrupt the exon. Two edited haplotypes do not interfere with the exon, but these are present at approximately the level of unedited haplotypes, suggesting 6TG resistance in these cells is caused by mutations elsewhere.




| B | positively selected in both | negatively selected in both | positively selected in rep 1 | negatively selected in rep 1 | positively selected in rep 2 | negatively selected in rep 2 |
|---|---|---|---|---|---|---|
| gRNA targeting non-6TG resistance regions (**A**) | 0 | 336 | 2 | 520 | 3 | 344 |
| random sequence not found in hg19 | 0 | 9 | 0 | 12 | 0 | 9 |

**Figure 4.11. None of the negative control random-sequence gRNAs were positively selected in both individual gRNA screen replicates.**

**a**, Selection scores across replicates for individual gRNAs that target regions not expected to induce 6TG resistance (as described in **Figure 4.7**). Only gRNAs sampled in both replicates are plotted. **b**, Table of the negative control gRNAs selected in both, either, or neither biological replicate.

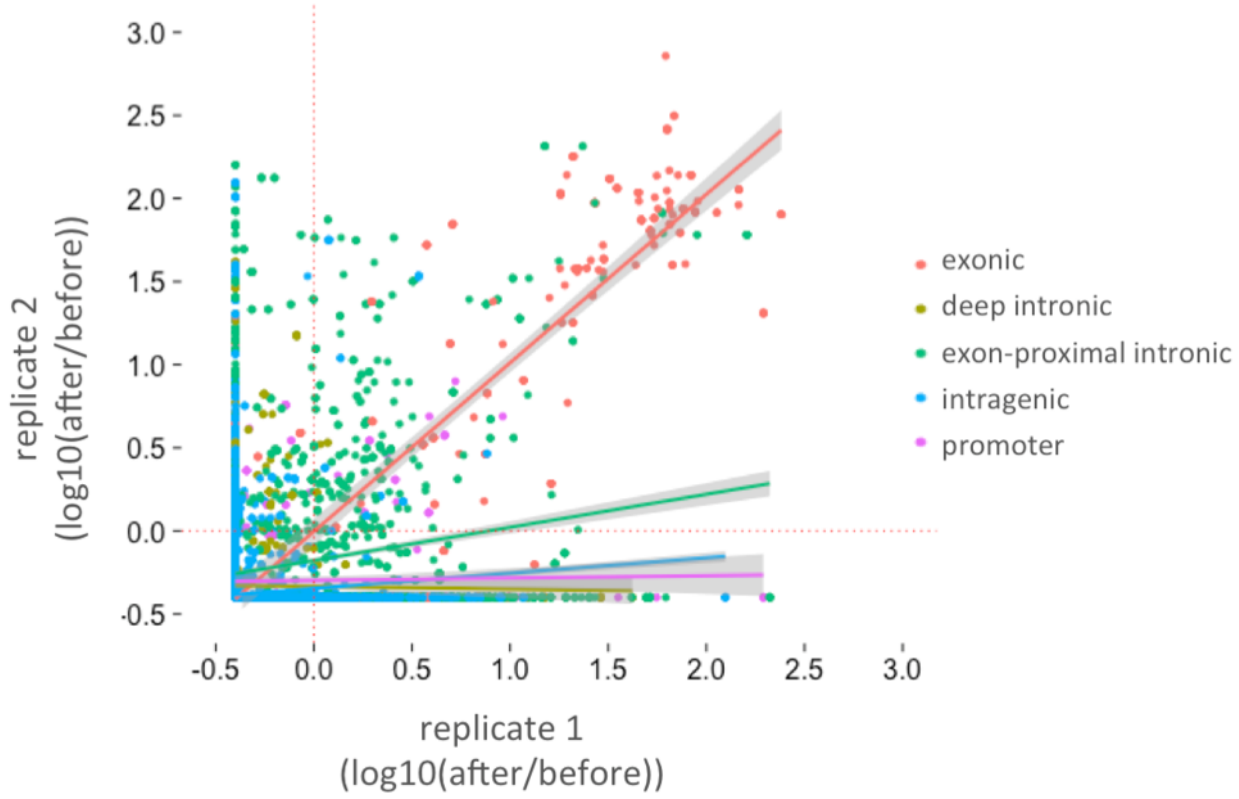



**Figure 4.12. Correlation of the individual gRNA screen scores across two biological replicates.**

The individual gRNA scores for each biological replicate were calculated per base-pair and presented as mean of log10(after/before 6TG) between replicates. Least squares lines and points are colored by sequence content category. Specifically, intronic sequence within 2 Kb of an exon is colored in green (Pearson: 0.176); exons are red (Pearson: 0.818); deep intronic is yellow (Pearson: -0.14); intragenic sequences are blue (Pearson: 0.070; and promoter sequence (2 Kb upstream of the TSS) is purple (Pearson: 0.022).

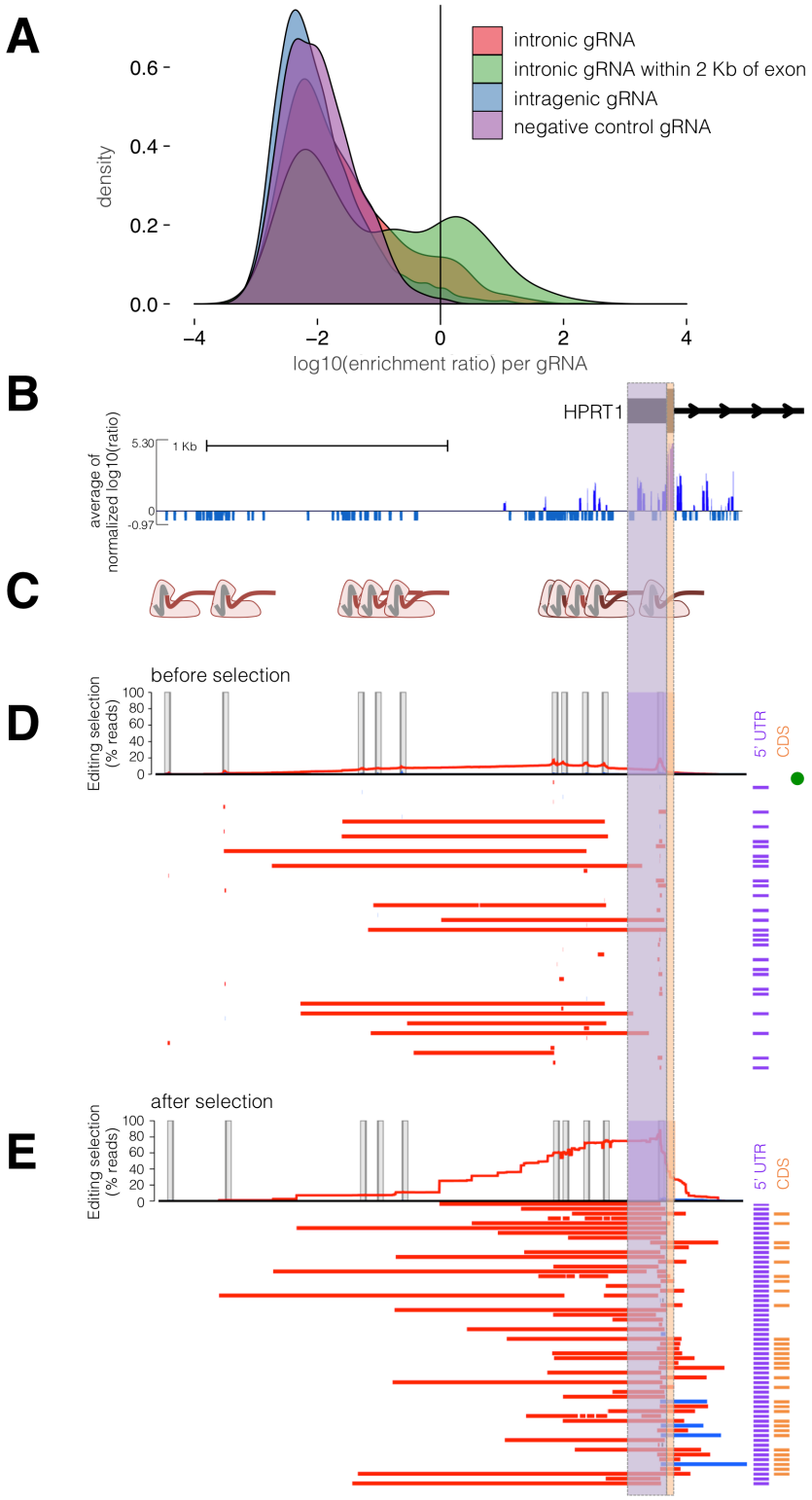



**Figure 4.13. Direct genotyping of edits from an individual gRNA mutagenesis screen also reveals rare, unexpected edits that disrupt *HPRT1*'s exon 1.**

**a,** A greater proportion of gRNAs targeting non-coding sequence within 2 Kb of exons were positively selected in an individual gRNA screen across the *HPRT1* locus (**Figure 4.3e**; data shown from replicate 1). Each gRNA was assigned a score equal to the log10(after/before 6TG). **b,** gRNAs that target upstream of the transcriptional start site are positively selected. The 2.4 Kb region sequenced for genotype validation (chrX:133,592,240-133,594,646, hg19) is shown. **c,** For validation, 10 gRNAs in this 2.4 Kb promoter region were cloned into a low complexity library, delivered to HAP1 cells expressing Cas9, and selected with 6TG. After selection, the 2.4 Kb promoter region was amplified for long-read sequencing. **d,** Before 6TG selection reads are plotted as in **Figure 4.10c**. Briefly, the per-base percentage of haplotypes that carried a deletion (red) or insertion (blue) is charted. The edits of the most-prevalent haplotypes from long-read sequencing are drawn as colored bars, and the notations to the right indicate if the edits interrupt the TSS/5'UTR (purple) or coding sequence (orange) of exon 1. A green dot signifies the unedited haplotype. Target site programmed edits are observed and are mainly comprised of the expected small indels, in addition to rarely occurring larger deletions. PCR and sequencing on the PacBio RSII are biased towards smaller fragments, limiting accuracy of quantitative comparison of the read-count prevalence of different sized edits. **e,** The most abundant haplotypes from cells after 6TG selection are visualized as in **d**. Only mutations that interrupt exon 1 survive 6TG selection.

# Chapter 5. WHOLE-ORGANISM LINEAGE TRACING BY COMBINATORIAL AND CUMULATIVE GENOME EDITING

Chapter 5 is adapted with minimal modifications from:

> McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.
>
> First authorship is shared between AM, GMF, and JAG.

## 5.1 ABSTRACT

Multicellular systems develop from single cells through distinct lineages. However, current lineage tracing approaches scale poorly to whole, complex organisms. Here we use genome editing to progressively introduce and accumulate diverse mutations in a DNA barcode over multiple rounds of cell division. The barcode, an array of CRISPR/Cas9 target sites, marks cells and enables the elucidation of lineage relationships via the patterns of mutations shared between cells. In cell culture and zebrafish, we show that rates and patterns of editing are tunable, and that thousands of lineage-informative barcode alleles can be generated. By sampling hundreds of thousands of cells from individual zebrafish, we find that most cells in adult organs derive from relatively few embryonic progenitors. In future analyses, genome editing of synthetic target arrays for lineage tracing (GESTALT) can be used to generate large-scale maps of cell lineage in multicellular systems for normal development and disease.

## 5.2 INTRODUCTION

The tracing of cell lineages was pioneered in nematodes by Charles Whitman in the 1870s, at a time of controversy surrounding Ernst Haeckel's theory of recapitulation, which argued that embryological development paralleled evolutionary history (Stent, 2002). This line of work culminated a century later in the complete description of mitotic divisions in the roundworm *C. elegans* - a tour de force facilitated by its visual transparency as well as the modest size and invariant nature of this nematode's cell lineage (Sulston et al., 1983).

Over the past century, a variety of creative methods have been developed for tracing cell lineage in developmentally complex organisms (Kretzschmar and Watt, 2012). In general, subsets of cells are marked and their descendants followed as development progresses. The ways in which cell marking has been achieved include dyes and enzymes (Keller, 1975; Kimmel and Law, 1985; Weisblat et al., 1978), cross-species transplantation (Le Douarin and Teillet, 1974), recombinase-mediated activation of reporter gene expression (Dymecki and Tomasiewicz, 1998; Zinyk et al., 1998), insertion of foreign DNA (Lu et al., 2011; Porter et al., 2014; Walsh and Cepko, 1992), and naturally occurring somatic mutations (Behjati et al., 2014; Lodato et al., 2015; Salipante and Horwitz, 2006). However, despite many powerful applications, these methods have limitations for the large-scale reconstruction of cell lineages in multicellular systems. For example, dye and reporter gene-based cell marking are uninformative with respect to the lineage relationships *between* descendent cells. Furthermore, when two or more cells are independently but equivalently marked, the resulting multitude of clades cannot be readily distinguished from one another. Although these limitations can be overcome in part with combinatorial labeling systems (Livet et al., 2007) or through the introduction of diverse DNA barcodes (Lu et al., 2011; Porter et al., 2014; Walsh and Cepko, 1992), these strategies fall short of a system for inferring lineage

relationships throughout an organism and across developmental time. In contrast, methods based on somatic mutations have this potential, as they can identify lineages and sub-lineages within single organisms (Carlson et al., 2011; Salipante and Horwitz, 2006). However, somatic mutations are distributed throughout the genome, necessitating whole genome sequencing, which is expensive to scale beyond small numbers of cells and not readily compatible with *in situ* readouts (Ke et al., 2013; Lee et al., 2014).

What are the requirements for a system for comprehensively tracing cell lineages in a complex multicellular system? First, it must uniquely and incrementally mark cells and their descendants over many divisions and in a way that does not interfere with normal development. Second, these unique marks must accumulate irreversibly over time, allowing the reconstruction of lineage trees. Finally, the full set of marks must be easily read out in each of many single cells.

We hypothesized that genome editing, which introduces diverse, irreversible edits in a highly programmable fashion (Doudna and Charpentier, 2014), could be repurposed for cell lineage tracing in a way that realizes these requirements. To this end, we developed genome editing of synthetic target arrays for lineage tracing (GESTALT), a method that uses CRISPR/Cas9 genome editing to accumulate combinatorial sequence diversity to a compact, multi-target, densely informative barcode. Edited barcodes can be efficiently queried by a single sequencing read from each of many single cells (**Figure 5.1A**). In both cell culture and in the zebrafish *Danio rerio*, we demonstrate the generation of thousands of uniquely edited barcodes that can be related to one another to reconstruct cell lineage relationships. In adult zebrafish, we observe that the majority of cells of each organ are derived from a small number of progenitor cells. Furthermore, ancestral progenitors, inferred on the basis of shared edits amongst subsets of derived alleles, make highly non-uniform contributions to germ layers and organ systems.

## 5.3 RESULTS

### 5.3.1 *Combinatorial editing of a compact genomic barcode in cultured cells*

To test whether genome editing can be used to generate a combinatorial diversity of mutations within a compact region, we synthesized a contiguous array of ten CRISPR/Cas9 targets separated by 3 base-pair (bp) linkers (total length of 257 bp). The first target perfectly matched one single guide RNA (sgRNA), whereas the remainder were off-target sites for the same sgRNA, ordered from highest to lowest activity (Tsai et al., 2015). This array of targets ('v1 barcode') was cloned downstream of an EGFP reporter in a lentiviral construct (Sancak et al., 2008). We then transduced HEK293T cells with lentivirus and used FACS to purify an EGFP-v1 positive population. To edit the barcode, we co-transfected these cells with a plasmid expressing Cas9 and the sgRNA and a vector expressing Discosoma red fluorescent protein (DsRed). Cells were sorted three days post-transfection for high DsRed expression, and genomic DNA (gDNA) was harvested on day 7. The v1 barcode was PCR amplified and the resulting amplicons subjected to deep sequencing.

To minimize confounding sequencing errors, which are primarily substitutions, we analyzed edited barcodes for only insertion-deletion changes relative to the 'wild-type' v1 barcode. In this first experiment, we observed 1,650 uniquely edited barcodes (each observed in ≥25 reads) with diverse edits concentrated at the expected Cas9 cleavage sites, predominantly inter-target deletions involving sites 1, 3 and 5, or focal edits of sites 1 and 3 (**Figure 5.1B,C**). These results show that combinatorial editing of the barcode can give rise to a large number of unique sequences, *i.e.* "alleles".

To evaluate reproducibility, we transfected the same editing reagents to cultures expanded from three independent EGFP-v1 positive clones. Targeted RT-PCR and sequencing of EGFP-

v1 RNA showed similar distributions of edits to the v1 barcode in the transcript pool, between replicates as well as in comparison to the previous experiment (**Figure 5.2**). These results show that the observed editing patterns are largely independent of the site of integration and that edited barcodes can be queried from either RNA or DNA.

To evaluate how editing outcomes vary as a function of Cas9 expression, we co-transfected EGFP-v1 positive cells with a plasmid expressing Cas9 and the sgRNA as well as an DsRed vector, and after four days sorted cells into low, medium, and high DsRed bins and harvested gDNA. Overall editing rates matched DsRed expression (frequency of non-wild-type barcodes: low DsRed = 40%; medium DsRed = 69%; high DsRed = 91%). The profile of edits observed remained similar, but there were fewer inter-target deletions in the lower DsRed bins (**Figure 5.3**). These results show that adjusting expression levels of editing reagents can be used to modify the rates and patterns of barcode editing.

We also synthesized and tested three barcodes (v2-v4) with nine or ten weaker off-target sites for the same sgRNA as used for v1 (Tsai et al., 2015). Genome editing resulted in derivative barcodes with substantially fewer edits than seen with the v1 barcode, but a much greater proportion of these edits were to a single target site, *i.e.* fewer inter-target deletions were observed (**Figure 5.1D**,**E**, **Figure 5.4A,B**). As only a few targets were substantially edited in designs v1-v4, we combined the most highly active targets to a new, twelve target barcode (v5). This barcode exhibited more uniform usage of constituent targets, but with relative activities still ranging over two orders of magnitude (**Figure 5.4C**). These results illustrate the potential value of iterative barcode design.

To determine whether the means of editing reagent delivery influences patterns of barcode editing, we introduced a lentiviral vector expressing Cas9 and the same sgRNA to cells containing

the v5 barcode (Sanjana et al., 2014). After two weeks of culturing a population bottlenecked to 200 cells by FACS, we observed diverse barcode alleles but with substantially fewer inter-target deletions than with episomal delivery of editing reagents (**Figure 5.4D**). This finding demonstrates that the allelic spectrum can also be modulated by the delivery mode of editing reagents.

Taken together, these results show that editing multiple target sites within a compact barcode can generate a combinatorial diversity of alleles, and also that these alleles can be read out by single sequencing reads derived from either DNA or RNA. Rates and patterns of barcode editing are tunable by using targets with different activities and/or off-target sequences, by iteratively recombining targets to new barcode designs, and by modulating the concentration and means of delivery of editing reagents.

### 5.3.2 *Reconstruction of lineage relationships in cultured cells*

To determine whether GESTALT could be used to reconstruct lineage relationships, we applied it to a designed lineage in cell culture (**Figure 5.5**). A monoclonal population of EGFP-v1 positive cells was transfected with editing reagents to induce a first round of mutations in the v1 barcode. Clones derived from single cells were expanded, sampled, split, and re-transfected with editing reagents to induce a second round of mutations of the v1 barcode. For each clonal population, two 100-cell samples of the re-edited populations were expanded and harvested for gDNA. In these experiments, we began incorporating unique molecular identifiers (UMIs; 10 bp) during amplification of barcodes by a single round of polymerase extension (**Figure 5.6A**). Each UMI tags the single barcode present within each single cell, thereby allowing for correction of subsequent PCR amplification bias and enabling each UMI-barcode combination to be interpreted as deriving from a single cell (Miner et al., 2004).

Seven of twelve clonal populations we isolated contained mutations in the v1 barcode that were unambiguously introduced during the first round of editing (**Figure 5.5A**). Additional edits accumulated in re-edited cells but generally did not disrupt the early edits (**Figure 5.5B**, **Figure 5.7**). We next sought to reconstruct the lineage relationships between all alleles observed in the experiment using a maximum parsimony approach (**Figure 5.6B**) (Felsenstein, 1989). The resultant tree contained major clades that were defined by the early edits present in each lineage (**Figure 5.5C**). Four clonal populations (#3, #5, #7 and #8) were cleanly separated upon lineage reconstruction, with >99.7% of cells accurately placed into each lineage's major clade. Two lineages (#1 and #6) were mixed because they shared identical mutations from the first round of editing. These most likely represent the recurrence of the same editing event across multiple lineages, but could also have been daughter cells subsequent to a single, early editing event prior to isolating clones. Consequently, 99.9% of cells of these two lineages were assigned to a single clade (**Figure 5.5C**, blue). One clonal population (#4) appears to have derived from two independent cells, one of which harbored an unedited barcode. Later editing of these barcodes confounded the assignment of this lineage on the tree. Overall, however, these results demonstrate that GESTALT can be used to capture and reconstruct cell lineage relationships in cultured cells.

### 5.3.3 *Combinatorial and cumulative editing of a genomic barcode in zebrafish*

To test the potential of GESTALT for *in vivo* lineage tracing in a complex multicellular organism, we turned to the zebrafish *Danio rerio.* We designed two new barcodes, v6 and v7, each with ten sgRNA target sites that are absent from the zebrafish genome and predicted to be highly editable. In contrast to v1-v5, in which the target sites are variably editable by one sgRNA, the targets within v6 or v7 are designed to be edited by distinct sgRNAs. We generated transgenic zebrafish that harbor each barcode in the 3' UTR of DsRed driven by the ubiquitin promoter

(Kawakami, 2007; Porter et al., 2014) and a GFP marker that is expressed in the cardiomyocytes of the heart (**Figure 5.8**) (Pan et al., 2013). To evaluate whether diverse alleles could be generated by *in vivo* genome editing, we injected Cas9 and ten different sgRNAs with perfect complementarity to the barcode target sites into single-cell v6 embryos (**Figure 5.9A**). Editing of integrated barcodes had no noticeable effects on development (**Figure 5.10**). To characterize barcode editing *in vivo*, we extracted gDNA from a series of single 30 hours post fertilization (hpf) embryos, and UMI-tagged, amplified and sequenced the v6 barcode. In control embryos (Cas9-; n = 2), all 4,488 captured barcodes were unedited. In contrast, in edited embryos (Cas9+; n = 8), fewer than 1% of captured barcodes were unedited. We recovered barcodes from hundreds of cells per embryo (median 943; range 257-2,832) and identified dozens to hundreds of alleles per embryo (median 225; range 86-1,323). 41% +/- 10% of alleles were observed recurrently within single embryos, most likely reflecting alleles that were generated in a progenitor of two or more cells. Fewer than 0.01% of alleles were shared in pairwise comparisons of embryos, revealing the highly stochastic nature of editing in different embryos. These results demonstrate that GESTALT can generate very high allelic diversity *in vivo*.

### 5.3.4 *Reconstruction of lineage relationships in embryos*

To evaluate whether lineage relationships can be reconstructed using edited barcodes, we focused on the v6 embryo with the lowest rates of inter-target deletions and edited target sites (**Figure 5.9**; avg. 58% +/- 27% of target sites no longer a perfect match to the unedited target, compared to 87% +/- 21% for all other 30 hpf v6 embryos). Application of our parsimony approach (**Figure 5.6B**) to the 1,961 cells in which we observed 1,323 distinct alleles generated the large tree shown in **Figure 5.11**. 1,307 of the 1,323 (98%) alleles could be related to at least one other allele by one or more shared edits, 85% by two or more shared edits, and 56% by three

or more shared edits. These results illustrate the principle of using patterns of shared edits between distinct barcode alleles to reconstruct their lineage relationships *in vivo*.

### 5.3.5 *Developmental timing of barcode editing*

To determine the developmental timing of barcode editing, we injected Cas9 and ten sgRNAs into one-cell stage v7 transgenic embryos and harvested genomic DNA before gastrulation (dome stage, 4.3 hpf; n = 10 animals), after gastrulation (90% epiboly / bud stage, 9 hpf; n = 11 animals), at pharyngula stage (30 hpf; n = 12 animals), and from early larvae (72 hpf; n = 12 animals) (**Figure 5.9A**). We recovered barcode sequences from a median of 8,785 cells per embryo (range 461-31,640; total of 45 embryos), comprising a median of 1,223 alleles per embryo (range 15-4,195) (**Figure 5.9C**). Within single embryos, 65% +/- 6% of alleles were observed recurrently, whereas in pairwise comparisons of embryos only 2% +/- 5% of alleles were observed recurrently. The abundances of alleles were well-correlated between technical replicates for each of two 72 hpf embryos (**Figure 5.12A**,**B**), and alleles containing many edits were more likely to be unique to an embryo than those with few edits (**Figure 5.12C**). To assess when editing begins, we analyzed the proportions of the most common editing events across all barcodes sequenced in a given embryo, reasoning that the earliest edits would be the most frequent. Across eight v6 and 45 v7 embryos, we never observed an edit that was present in 100% of cells. This observation indicates that no permanent edits were introduced at the one-cell stage. In nearly all embryos, we observe that the most common edit is present in >10% of cells, and in some cases in ~50% of cells (**Figure 5.9D, Figure 5.13**). This observation also holds in ~4,000-cell dome stage embryos, which result from approximately 12 rounds of largely synchronous division unaccompanied by cell death. Most of these edits are rare or absent in other embryos, suggesting they are unlikely to have arisen recurrently within each lineage. These results suggest

that the edits present in ~50% of cells were introduced at the two-cell stage and that the edits present in >10% of cells were introduced before the 16-cell stage.

How long does barcode editing persist? Two aspects of the data suggest that it tapers relatively early in development. First, in dome stage embryos (4.3 hpf), we captured barcodes from a median of 2,086 cells, in which a median of 4.8 targets were edited. Although the number of cells and alleles that we were able to sample increased at the later developmental stages, the proportion of edited sites appeared relatively stable (**Figure 5.9C**). If editing were occurring throughout this time course, we would instead expect the proportion of edited sites to increase substantially. Second, the number of unique alleles appears to saturate early, never exceeding 4,200 (**Figure 5.9E**). For example, only 4,195 alleles were observed in a 72 hpf embryo in which we sampled the highest number of cells (n = 31,639). These results suggest that the majority of editing events occurred before dome stage.

### 5.3.6 *Editing diversity in adult organs*

To evaluate whether barcodes edited during embryogenesis can be recovered in adults, we dissected two edited 4-month old v7 transgenic zebrafish (ADR1 and ADR2) (**Figure 5.14A**). We collected organs representing all germ layers - the brain and both eyes (ectodermal), the intestinal bulb and posterior intestine (endodermal), the heart and blood (mesodermal), and the gills (neural crest, with contributions from other germ layers). We further divided the heart into four samples – a piece of heart tissue, dissociated unsorted cells (DHCs), FACS-sorted GFP+ cardiomyocytes, and non-cardiomyocyte heart cells (NCs) (**Figure 5.15**). We isolated genomic DNA from each sample, amplified and sequenced edited barcodes with high technical reproducibility (**Figure 5.16**), and observed barcode editing rates akin to those in embryos (**Figure 5.17**). For zebrafish ADR1, we captured barcodes from between 776 and 44,239 cells

from each tissue sample (median 17,335), corresponding to a total of 197,461 cells and 1,138 alleles. For zebrafish ADR2, we captured barcodes from between 84 and 52,984 cells from each tissue sample (median 20,973), corresponding to a total of 217,763 cells and 2,016 alleles. These results show that edits introduced to the barcode during embryogenesis are inherited through development and tissue homeostasis and can be detected in adult organs.

### 5.3.7 *Differential contribution of embryonic progenitors to adult organs*

To analyze the contribution of diverse alleles to different organs, we compared the frequency of edited barcodes within and between organs. We first examined blood (of note, zebrafish erythrocytes are nucleated (Thisse and Zon, 2002)). Only 5 alleles defined over 98% of cells in the ADR1 blood sample (**Figure 5.14B**), suggesting highly clonal origins of the adult zebrafish blood system from a few embryonic progenitors. Consistent with the presence of blood in all dissected organs, these common blood alleles were also observed in all organs (10-40%; **Figure 5.14C**) but largely absent from cardiomyocytes isolated by flow sorting (0.5%). Furthermore, the relative proportions of these five alleles remained constant in all dissected organs, suggesting that they primarily mark the blood and do not substantially contribute to non-blood lineages (**Figure 5.14D**). In performing similar analyses of clonality across all organs (while excluding the five most common blood alleles), we observed that a small subset of alleles dominates each organ (**Figure 5.14E**). Indeed, for all dissected organs, fewer than 7 alleles comprised >50% of cells (median 4, range 2-6), and, with the exception of the brain, fewer than 25 alleles comprised >90% of cells (median 19, range 4-38). Most of these dominant alleles were organ-specific, *i.e*. although they were found rarely in other organs, they tended to be dominant in only one organ (**Figure 5.14F**). For example, the most frequent allele observed in the intestinal bulb comprised 13.6% of captured non-blood cells observed in that organ, but <0.01% of cells

observed in any other organ. There are exceptions, however. For example, one allele is observed in 24.7% of sorted cardiomyocytes, 13.4% of the intestinal bulb, and at lower abundances in all other organs. Similar results were observed in ADR2 (**Figure 5.18**). These results indicate that the majority of cells in diverse adult organs are descended from a few differentially edited embryonic precursors.

### 5.3.8 *Reconstructing lineage relationships in adult organs*

To reconstruct the lineage relationships between cells both within and across organs on the basis of shared edits, we again relied on maximum parsimony methods (**Figure 5.6B**). The resulting trees for ADR1 and ADR2 are shown in **Figure 5.19** and **Figure 5.20**, respectively. We observed clades of alleles that shared specific edits. For example, ADR1 had 8 major clades, each defined by 'ancestral' edits that are shared by all captured cells assigned to that clade (**Figure 5.21A**; also indicated by colors in the tree shown in **Figure 5.19**). Collectively, these clades comprised 49% of alleles and 90% of the 197,461 cells sampled from ADR1 (**Figure 5.21A**). Blood was contributed to by 3 major clades (#3, #6, #7) (**Figure 5.21B**). After re-allocating the 5 dominant blood alleles from the composition of individual organs back to blood (**Figure 5.14B** and **Figure 5.22**), we observed that all major clades made highly non-uniform contributions across organs. For example, clade #3 contributed almost exclusively to mesodermal and endodermal organs, while clade #5 contributed almost exclusively to ectodermal organs. These results reveal that GESTALT can be used to infer the contributions of inferred ancestral progenitors to adult organs.

Although some ancestral clades appear to contribute to all germ layers, we find that subclades, defined by additional shared edits within a clade, exhibit greater specificity. For example, although clade #1 contributes substantially to all organs except blood, additional edits

divide clade #1 into three subclades with greater tissue restriction (**Figure 5.21C**,**D**). The #1+A subclade primarily contributes to mesendodermal organs (heart, both gastrointestinal organs) whereas the #1+C subclade primarily contributes to neuroectodermal organs (brain, left eye, and gills). Similar patterns are observed for clade #2 (**Figure 5.21E,F**), where the #2+A subclade contributes primarily to mesendodermal organs, the #2+B subclade to the heart, and the #2+C clade to neuroectodermal organs. Additional edits divide these subclades into further tissue-specific sub-subclades. For example, whereas the #2+A subclade is predominantly mesendoderm, additional edits define #2+A+D (heart, primarily cardiomyocytes), #2+A+E (heart and posterior intestine), and #2+A+F (intestinal bulb). All of the major clades exhibit similar patterns of increasing restriction with additional edits (**Figure 5.21C-F** and **Figure 5.23**). Similar observations were made in fish ADR2 (**Figure 5.24**). These results indicate that GESTALT can record lineage relationships across many cell divisions and capture information both before and during tissue restriction.

## 5.4 Discussion

We describe a new method, GESTALT, which uses combinatorial and cumulative genome editing to record cell lineage information in a highly multiplexed fashion. We successfully applied this method to both artificial lineages (cell culture) as well as to a whole organism (zebrafish). Full tree reconstructions for cell culture, zebrafish embryo, and zebrafish adult experiments are provided at http://gestalt.gs.washington.edu/.

The strengths of GESTALT include: 1) the combinatorial diversity of mutations that can be generated within a dense array of CRISPR/Cas9 target sites; 2) the potential for informative mutations to accumulate across many cell divisions and throughout an organism's developmental history; 3) the ability to scalably query lineage information from at least hundreds of thousands of

cells and with a single sequencing read per single cell; 4) the likely applicability of GESTALT to any organism, from bacteria and plants to vertebrates, that allows genome editing, as well as human cells (*e.g.* tumor xenografts). Even in organisms in which transgenesis is not established, lineage tracing by genome editing may be feasible by expressing editing reagents to densely mutate an endogenous, non-essential genomic sequence.

Our experiments also highlight several remaining technical challenges. Chief amongst these are: 1) the chance recurrence of identical edits or similar patterns of edits in distantly related cells can confound lineage inference; 2) non-uniform editing efficiencies and inter-target deletions within the barcode contribute to suboptimal sequence diversity and loss of information, respectively; 3) the transient means by which Cas9 and sgRNAs are introduced likely restrict editing to early embryogenesis; 4) the computational challenge of precisely defining the multiple editing events that give rise to different alleles complicates the unequivocal reconstruction of lineage trees; and 5) the difficulty of isolating tissues without contamination by blood and other cells can hinder the assignment of alleles to specific organs. A broader set of challenges includes the lack of information about the precise anatomical location and exact cell type of each queried cell, the fact that genome editing events are not directly coupled to the cell cycle, and the failure to recover all cells. These challenges currently hinder the reconstruction of a lineage tree as complete and precise as the one that Sulston and colleagues described for *C. elegans*. Despite these limitations, our proof-of-principle study shows that GESTALT can inform developmental biology by richly defining lineage relationships among vast numbers of cells recovered from an organism.

The current challenges highlight the need for further optimization of the design of targets and arrays, as well as the delivery of editing reagents. For example, an array containing twice as many targets as used here could fit within a single read on contemporary sequencing platforms,

thus yielding more lineage information per cell without sacrificing throughput. Also, as we have shown, adjustments to the target sequences and dosages of editing reagents can be used to fine-tune mutation rates and to minimize undesirable inter-target deletions. Finally, sgRNA sequences and lengths (Fu et al., 2014b), Cas9 cleavage activity and target preferences (Kleinstiver et al., 2015a; Slaymaker et al., 2016), and the means by which Cas9 and sgRNA(s) are expressed (e.g. transient, constitutive (Platt et al., 2014), or induced (Ablain et al., 2015; Yin et al., 2015)), can be altered to control the pace, temporal window and tissue(s) at which the barcodes are mutated. For example, coupling editing to cell cycle progression might enable higher resolution reconstruction of lineage relationships throughout development.

Our application of GESTALT to a vertebrate model organism, zebrafish, demonstrates its potential to yield insights into developmental biology. First, our results suggest that relatively few embryonic progenitor cells give rise to the majority of cells of many adult zebrafish organs, reminiscent of clonal dominance (Gupta and Poss, 2012; Snippert et al., 2010). For example, only 5 of the 1,138 alleles observed in ADR1 gave rise to >98% of blood cells, and for all dissected organs, fewer than 7 alleles comprised >50% of cells. There are several mechanisms by which such dominance can emerge, *e.g.* by uneven starting populations in the embryo, drift, competition, interference, unequal cell proliferation or death, or a combination of these mechanisms (Blanpain and Simons, 2013; Henson and Hume, 2006; Klein and Simons, 2011; Pellettieri and Sánchez Alvarado, 2007). Controlling the temporal and spatial induction of edits and isolating defined cell types from diverse organs should help resolve the mechanisms by which different embryonic progenitors come to dominate different adult organs.

Second, we show that GESTALT can inform the lineage relationships amongst thousands of differentiated cells. For example, following the accumulation of edits from ancestral to more

complex reveals the progressive restriction of progenitors to germ layers and then organs. Cells within an organ can both share and differ in their alleles, revealing additional information about organ development. Future studies will need to determine whether such lineages reflect distinct cell fates (*e.g.*, blood sub-lineages or neuronal subpopulations), because the anatomical resolution at which we queried alleles was restricted to grossly dissected organs and tissues. Because edited barcodes are expressed as RNA, we envision that combining our system with other platforms will permit much greater levels of anatomical resolution without sacrificing throughput. For example, *in situ* RNA sequencing of barcodes would provide explicit spatial and histological context to lineage reconstructions (Ke et al., 2013; Lee et al., 2014). Also, capturing richly informative lineage markers in single cell RNA-seq or ATAC-seq datasets may inform the interpretation of those molecular phenotypes, while also adding cell type resolution to studies of lineage (Cusanovich et al., 2015; Satija et al., 2015). Such integration may be particularly relevant to efforts to build comprehensive atlases of cell types. Because these single cell methods generate many reads per single cell, this would also facilitate using multiple, unlinked target arrays. In principle, the combined diversity of the barcodes queried from single cells could be engineered to uniquely identify every cell in a complex organism. In addition, orthogonal imaging-based lineage tracing approaches in fixed and live samples (*e.g.*, Brainbow and related methods (Livet et al., 2007; Pan et al., 2013)) and longitudinal whole animal imaging approaches (Liu and Keller, 2016; Megason and Fraser, 2007) might be leveraged in parallel to validate and complement lineages resolved by GESTALT.

Although further work is required to optimize GESTALT towards enabling spatiotemporally complete maps of cell lineage, our proof-of-principle experiments show that using multiplex *in vivo* genome editing to record lineage information to a compact barcode at an

organism-wide scale will be a powerful tool for developmental biology. This approach is not limited to normal development but can also be applied to animal models of developmental disorders, as well as to investigate the origins and progression of cancer. Our study also supports the notion that whereas its most widespread application has been to modify endogenous biological circuits, genome editing can also be used to stably record biological information (Farzadfard and Lu, 2014), analogous to recombinase-based memories but with considerably greater flexibility and scalability. For example, coupling editing activity to external stimuli or physiological changes could record the history of exposure to intrinsic or extrinsic signals. In the long term, we envision that rich, systematically generated maps of organismal development, wherein lineage, epigenetic, transcriptional and positional information are concurrently captured at single cell resolution, will advance our understanding of normal development, inherited diseases, and cancer.

## 5.5 METHODS

### 5.5.1 *Design of synthetic target arrays*

Barcodes were designed as arrays of nine to twelve sense-oriented CRISPR/Cas9 target sites (23 bp, protospacer plus PAM sequences) separated by 3-5 bp linker sequences. Four initial designs (barcodes v1-v4) comprised of target sites for the sgRNA spacer sequence: 5'-GGCACTGCGGCTGGAGGTGG. The v1 barcode was comprised of ten targets arrayed in order of decreasing activity as measured with the GUIDE-seq assay performed in human cells (*22*), starting with the target perfectly matching the sgRNA spacer sequence. The v2-v4 barcodes comprised of nine to ten non-overlapping target sets, all with activities less than half the perfectly matching target in the GUIDE-seq assay. To reduce repetitive subsequences within each barcode, protospacers were chosen such that no 8 bp sequence was present in more than one protospacer within each barcode. After testing activities of targets in the v1-v4 barcodes in cell culture, the v5

barcode was designed to contain twelve targets that showed greater than ~1% editing activity, including v1 targets 1-6, v3 target 1, v2 targets 1, 2 and 5, and v4 targets 1 and 3.

Two new barcodes, v6 and v7, were designed for use in zebrafish, each with ten CRISPR target sites not found in the *D. rerio* genome. Candidate target sequences were screened to remove any homopolymer runs, outside of the NGG of the protospacer, and were selected for editing activity [http://crispr.mit.edu]. The v6 and v7 barcodes were constructed as a series of 10 protospacer sequences meeting these criteria, with 4 bp linkers.

Each barcode was ordered as a gBlock (IDT) with ends compatible for In-Fusion cloning (ClonTech) into the 3' UTR of the EGFP gene in the lentiviral construct pLJM1-EGFP (Addgene #19319).

### 5.5.2 *Generation of cell lines containing synthetic target arrays*

To generate cell lines harboring single copies of barcodes, lentiviral particles were produced in HEK 293T cells transfected with lentivirus V2 packaging plasmids and barcode constructs. Viral supernatant harvested three days post transfection was used at low MOI to transduce 293T cells (MOI < 0.2). Successfully transduced cells were selected using puromycin (2 μg/ml), yielding polyclonal, barcode$^+$ populations for barcodes v1-v5. Three monoclonal lines each harboring barcode v1 were generated by single-cell FACS, and used experimentally to compare editing rates across different integration sites. One of these was used as the parent line for cell culture lineages derived using barcode v1.

### 5.5.3 *Editing of barcodes in cell lines*

293T populations bearing barcodes v1-v5 were grown to 50-90% confluency in a 6-well dish. Cells were co-transfected using Lipofectamine 3000 (Life Technologies) according to protocol with 2

µg pX330-v1 and 0.5 µg pDsRed in a 6-well dish. One to three days post transfection, the cells were sorted on an Aria III FACS machine for DsRed fluorescence (as a marker transfection). As indicated, either DsRed low, DsRed high, or total DsRed populations were sorted and cultured. At 7 days post-transfection, cells were harvested for gDNA preparation using the Qiagen DNeasy kit. To stably deliver Cas9 and the sgRNA via lentivirus, the spacer sequence was cloned into the plasmid LentiCrispr v2 (Zhang lab, Addgene #52961) and virus was produced in 293T cells in the same manner described above. Wild-type 293T cells were transduced with pLenti-Crispr-V2-HMID.v1 and selected with puromycin, and then transduced with lentivirus bearing barcode v5. To impose a bottleneck, 200 GFP+ cells were sorted from this population and expanded under puromycin selection for two weeks prior to sampling gDNA.

### 5.5.4 *Cell culture lineage experiments*

Twelve lineages were established from a monoclonal barcode v1 293T cell line by transfecting cells as described above, and sorting single DsRed-low cells into a 96-well plate (DsRed low cells were used to limit Cas9 delivery and thus potential saturation of possible edits in this initial editing round). Cell sorting was performed seven days post-transfection, to reduce the likelihood that additional edits would arise after lineages were separated. Single cell-derived populations were expanded in culture for 3 weeks. A sample of cells from each lineage was pelleted and frozen. Next, each of the twelve lineages were transfected a second time, to induce another round of editing. Two 100-cell DsRed-low populations from each lineage were sorted 4-days post-transfection, and cultured to confluence in 96-well plates before harvesting gDNA.

Four additional monoclonal populations bearing v5 barcodes edited via transfection of pX330-v1 were also isolated by single-cell sorting. Re-editing of each population was achieved by two

successive rounds of transfection with pX330-v1 (3 days apart). Cells were harvested for gDNA one week after the second transfection.

### 5.5.5 *Barcode amplification and sequencing protocols*

Kapa High Fidelity Polymerase was used for all barcode amplification steps. Gradient PCRs were performed to optimize annealing temperatures for amplification from gDNA. For experiments performed without UMIs, up to 250 ng of gDNA was loaded into a single 50 µl PCR reaction and amplified using primers immediately flanking the barcode. If there was less than 250 ng from a sample, all of it was used in a single reaction. For experiments performed with UMIs, a primer with a sequencing adapter and 10 nt of fully degenerate sequence 5' to the barcode-flanking sequence was used for a single prolonged extension step, in which the temperature was ramped between annealing and extending for five cycles (without a denaturing step to prevent re-sampling of gDNA barcodes). All cell culture experiments and v6 zebrafish embryos received a single extension to incorporate UMIs, whereas v7 embryo time-course experiments and all ADR1 tissues (also v7) received 2 UMI incorporation cycles due to having low gDNA consequent to fewer cells being present in early embryo and sorted heart samples. To minimize repetitive amplification of the same barcode, no reverse primer was included in UMI-tagging reactions. DNA was then purified using AMPure beads (Agencourt), and loaded into a PCR primed from the sequencing adapter flanking each UMI and a site immediately 3' of the barcode.

For all experiments, two ensuing qPCRs were performed prior to sequencing to incorporate sequencing adapters, sample indexes, and flow cell adapters. AMPure beads were used to purify PCR products after each reaction.

Paired-end sequencing was performed on an Illumina MiSeq using 500- or 600-cycle kits for all cell culture experiments. Zebrafish experiments were sequenced on an Illumina NextSeq using

300-cycle kits. All sequencing generated adequate depth to sample each barcode present in a given sample to an average of greater than 10x coverage. To minimize contributions from sequencing error a read threshold was used for calling unique barcodes. This was conservatively set by dividing the number of reads from a sample by the number of expected barcode copies to be present in the amount of gDNA loaded into each PCR based on the assumption that each cell contributed a single barcode.

Sequencing data for all samples was processed in a custom pipeline available on GitHub. Briefly, amplicon sequencing reads were first processed with the Trimmomatic software package to remove low quality bases (**Figure 5.6A**) (Bolger et al., 2014). The resulting reads were then grouped by their UMI tag. A raw read count threshold was set for each experiment based on sequencing depth, such that only UMIs observed in at least that many reads were analyzed to minimize contributions from sequencing error. For each UMI, a consensus sequence was called by jointly aligning all UMI-matched reads using the MAFFT multiple sequence aligner (Rice et al., 2000). These reads were merged using the FLASH read merging tool (Magoč and Salzberg, 2011), and both merged and unmerged reads were aligned to the amplicon reference using the NEEDLEALL aligner (Rice et al., 2000) with a gap open penalty of 10 and a gap extension penalty of 0.5. To capture read-through, UMI degenerate bases and adapter sequences were included in the reference amplicon sequence, and mismatches to Ns in the degenerate bases were set to a penalty of 0. To eliminate off-target sequencing reads, aligned sequences were required to match greater than 85% of bases at non-indel positions, to have correct PCR primer sequences on both the 5' and 3' ends, and to match at least 50 bases of the reference sequence (including primer sequences). Target sites were deemed edited if there was an insertion or deletion event present within 3 bases of the predicted Cas9 cut site (3 nucleotides 5' of each PAM), or if a deletion

spanned the site entirely. Sites were marked as disrupted if there was not perfect alignment of the barcode over the entirety of the reference target sequence. An edited barcode was then defined as the complete list of insertion and deletion events (*i.e.* 'editing events') within the consensus sequence for a given UMI.

### 5.5.6 *Maximum parsimony lineage reconstruction*

For lineage reconstruction (**Figure 5.6B**), recurrently observed barcode alleles within a single organ or cell population were reduced to a single representative entry. We then used Camin-Sokal maximum parsimony to reconstruct lineages, as implemented in the PHYLIP Mix software package (Felsenstein, 1989). Camin-Sokal maximum parsimony assumes that the initial cell or zygote is unedited, and that editing is irreversible. To run Mix, a matrix was created where each row corresponded to an allele, and each column corresponds to a unique editing event. Each entry in this matrix is an indicator variable of presence or absence of a specific edit in that allele (1 or 0). Events were also weighted by their log-abundance and scaled to the range allowed in Mix (0-Z). Mix was run with both the indicator data matrix as well as the weights file (selecting run options P, W, 4, and 5), and the output was parsed to recover the edit state of ancestral (internal) nodes. When Mix discovered multiple equal-scoring trees, we took the tree in the highest proportion. If two trees tied for highest proportion, we took the last highest scoring tree. To eliminate unsupported internal branching, we pruned internal parent-child nodes that had identical alleles. When a parent node and child node share the same allele, and neither node was a leaf, the grandchildren nodes were transferred to the parent and the child node was removed, creating multifurcating parent nodes. The resulting tree was converted to an annotated JSON tree compatible with our visualization tools.

### 5.5.7 *Zebrafish husbandry*

All vertebrate animal work was performed at the facilities of Harvard University, Faculty of Arts & Sciences (HU/FAS). This study was approved by the Harvard University/Faculty of Arts & Sciences Standing Committee on the Use of Animals in Research & Teaching under Protocol No. 25–08. The HU/FAS animal care and use program maintains full AAALAC accreditation, is assured with OLAW (A3593-01), and is currently registered with the USDA.

### 5.5.8 *Cloning transgenesis vector*

The transgenesis vectors pTol2-DRv6 and pTol2-DRv7 were constructed as follows. The v6 or v7 array was cloned into the 3' UTR of a DsRed coding sequence under control of the ubiquitin promoter (*51*). This cassette was placed in a Tol2 transgenesis vector containing a cmlc2:GFP marker, which drives expression of GFP in the cardiomyocytes of the heart from 24 hpf to adulthood (*52*). Plasmids are available from Addgene.

### 5.5.9 *Generating transgenic zebrafish*

To generate founder fish, 1-cell embryos were injected with zebrafish codon optimized Tol2 mRNA and pTol2-DR1v6 or pTol2-DR1v7 vector. Potential founder fish were screened for heart GFP expression at 30 hpf and grown to adulthood. Adult founder transgenic fish were identified by outcrossing to wild type and screening clutches of embryos for heart GFP expression at 30 hpf.

### 5.5.10 *Transgene copy number quantification*

To identify single copy Tol2 transgenics, copy number was quantified using qPCR (Pan et al., 2013). Briefly, genomic DNA was extracted from candidate embryos or fin-clips of adult fish

using the HotSHOT method (Meeker et al., 2007) and subjected to qPCR using a set of primers targeting DsRed and a set targeting a diploid conserved region of the genome and compared to reference non-transgenic, 1-copy and 2-copy transgenic animals using the ddCt method.

### 5.5.11 *Generation and delivery of editing reagents*

sgRNAs specific to each site of the v6 or v7 array were generated as previously described (Gagnon et al., 2014), except that sgRNAs were isolated after transcription by column purification (Zymogen). 1-cell embryos resulting from an outcross of a transgenic founder were injected with two different volumes (0.5 nl, 1/3x or 1.5nl, 1x) of Cas9 protein (NEB) and sgRNAs in salt solution (8 µM Cas9, 100 ng/µl pooled sgRNAs, 50 mM KCl, 3 mM $MgCl_2$, 5 mM Tris HCl pH 8.0, 0.05% phenol red). Transgenic embryos were collected at the time points indicated in the text and genomic DNA extracted as described below. To confirm editing, PCR was conducted on a subset of samples using primers flanking the v6 or v7 array, and amplicons were loaded on a 2% agarose gel for electrophoresis.

### 5.5.12 *Imaging*

Embryos were anaesthetized and manually dechorionated in MS222, mounted in methylcellulose and imaged using a Leica upright fluorescence microscope.

### 5.5.13 *Organ Dissection*

Adult edited single copy transgenic fish were isolated without food for one day to reduce food particles in the gastrointestinal system, then anaesthetized in MS222 and euthanized on ice. Before dissection, blood was collected using a centrifugation method (Babaei et al., 2013). This collection method greatly enriches for blood cells, particularly red blood cells, but also results in contamination from skin or other tissues. The fish were pinned on a silicon mat and surgery was

conducted using sterile tools to remove organs as in (Gupta and Mullins, 2010). Organs were washed in PBS and, with the exception of the heart, frozen in tubes on dry ice. A piece of heart tissue was collected before the remainder of the heart was dissociated following manufacturer's instructions (Miltenyi #130-098-373). After dissociation, a sample of dissociated heart cells was collected (DHCs), and the remaining cells sorted using a Beckman Coulter MoFlo XDP Cell Sorter through a series of three gates to minimize debris and cell doublets, and then split into two additional populations: GFP+ cardiomyocytes and GFP- non-cardiomyocyte heart cells (NCs).

### 5.5.14 *Genomic DNA preparation from zebrafish embryos and organs*

Zebrafish embryo and adult organ gDNA was prepared using the Qiagen DNeasy kit. For heart samples from cell sorting experiments, 1 µl of poly-dT carrier DNA (25 µM) was added prior to gDNA preparation. Digestion with proteinase K at 56$^{o}$ C was performed overnight for intact organs (brain, eyes, gills, intestinal bulb, posterior intestine, and piece of heart) and for 30 minutes for blood samples, dissociated heart cells and embryos. gDNA was eluted in 100 µl, then concentrated using an Eppendorf Vacufuge for samples yielding less than 1 µg.

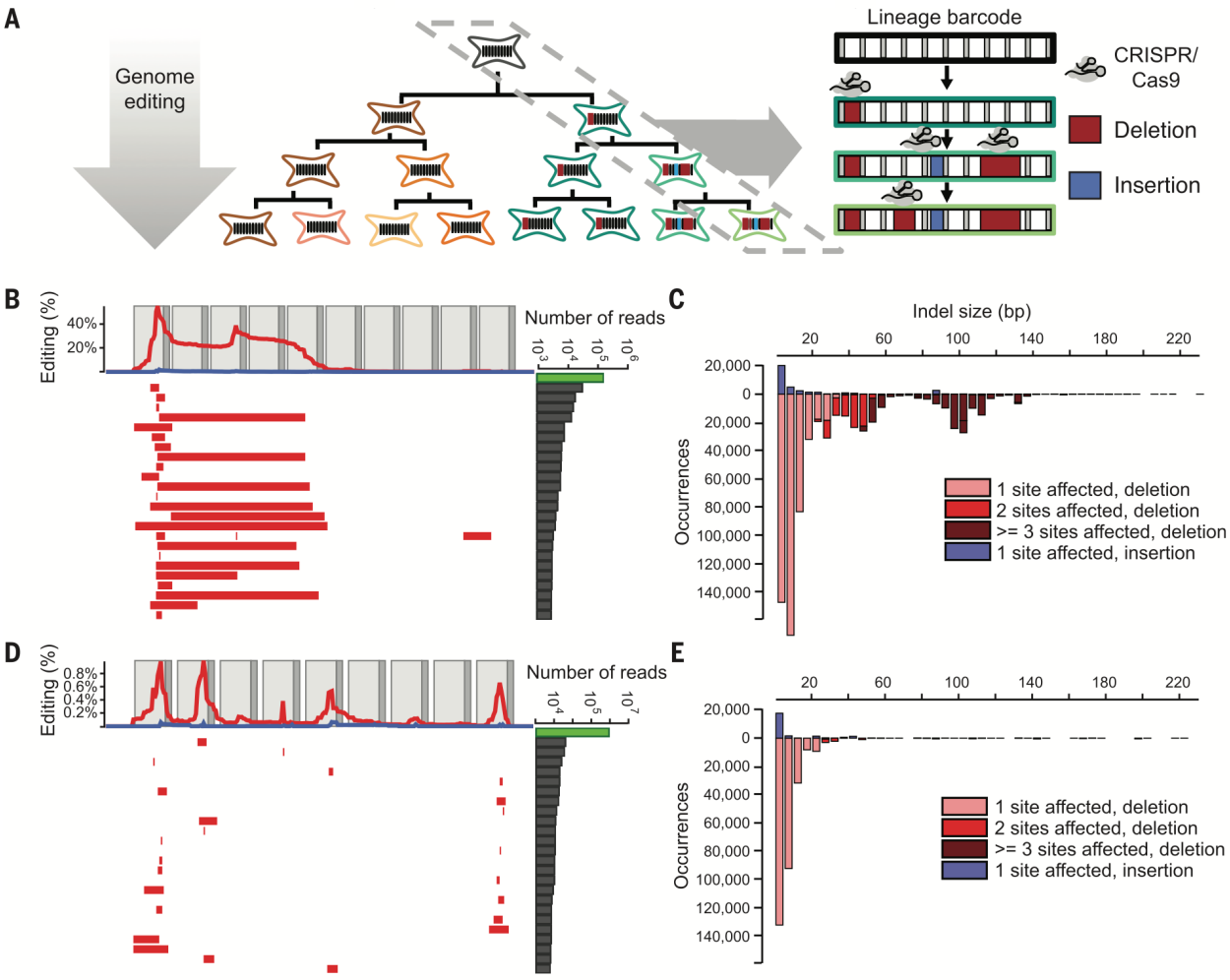



**Figure 5.1. Genome editing of synthetic target arrays for lineage tracing (GESTALT).**

**(A)** An unmodified array of CRISPR/Cas9 target sites (*i.e.*, a barcode) is engineered into a genome (gray cell). Editing reagents are introduced during expansion of cell culture or *in vivo* development of an organism, resulting in a unique pattern of insertions and deletions (right), and are stably accumulated in specific lineages (green cell lineage). The lineage relationships of alleles that differ in seque-------nce can often be inferred on the basis of these accumulated edits. **(B)** The 25 most frequent alleles from the edited v1 barcode are shown. Each row corresponds to a unique sequence, with red bars indicating deleted regions and blue bars indicating insertion positions. Blue bars begin at the insertion site, with their width proportional to the size of the insertion, which will rarely obscure immediately adjacent deletions. The number of reads observed for each allele is plotted at the right (log10 scale; the green bar corresponds to the unedited allele). The frequency at which each base is deleted (red) or flanks an insertion (blue) is plotted at the top. Light gray boxes indicate the location of CRISPR protospacers while dark gray boxes indicate PAM sites. For the v1 array, inter-target deletions involving sites 1, 3 and 5, or focal (single target) edits of sites 1 and 3 were observed predominantly. **(C)** A histogram of the size distribution of insertion (top) and deletion (bottom) edits to the v1 array is shown. The colors indicate the number of target sites impacted. Although most edits are short and impact a single target, a substantial proportion

of edits are inter-target deletions. **(D)** We tested three array designs in addition to v1, each comprising nine to ten weaker off-target sites for the same sgRNA (v2-v4). Editing of the v2 array is shown with layout as described in panel (B). Editing of the v3 and v4 array are shown in **Figure 5.4**. The weaker sites within these alternative designs exhibit lower rates of editing than the v1 array, but also a much lower proportion of inter-target deletions. **(E)** A histogram of the size distribution of insertion (top) and deletion (bottom) edits to the v2 array is shown. In contrast with the v1 array, almost all edits impact only a single target.

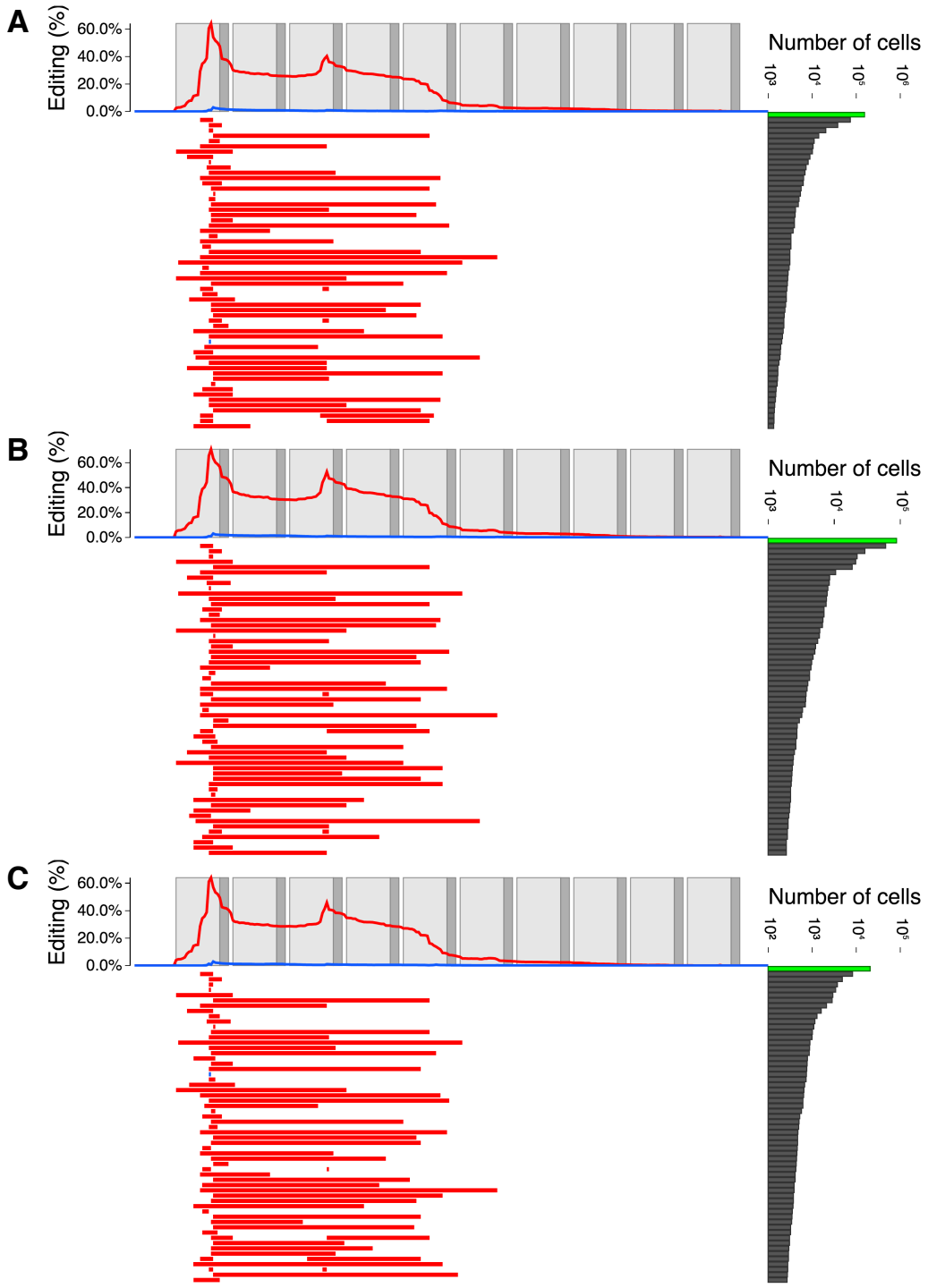



**Figure 5.2. RNA-based readout of v1 barcode editing**.

(**A-C**) The 50 most frequent alleles of the v1 barcode are depicted, based on reverse transcription, amplification and sequencing from mRNA. Three biological replicates are shown for comparison, wherein each experiment was performed on a culture expanded from an independent v1+ clone. Layout is as described in the **Figure 5.1B** legend.

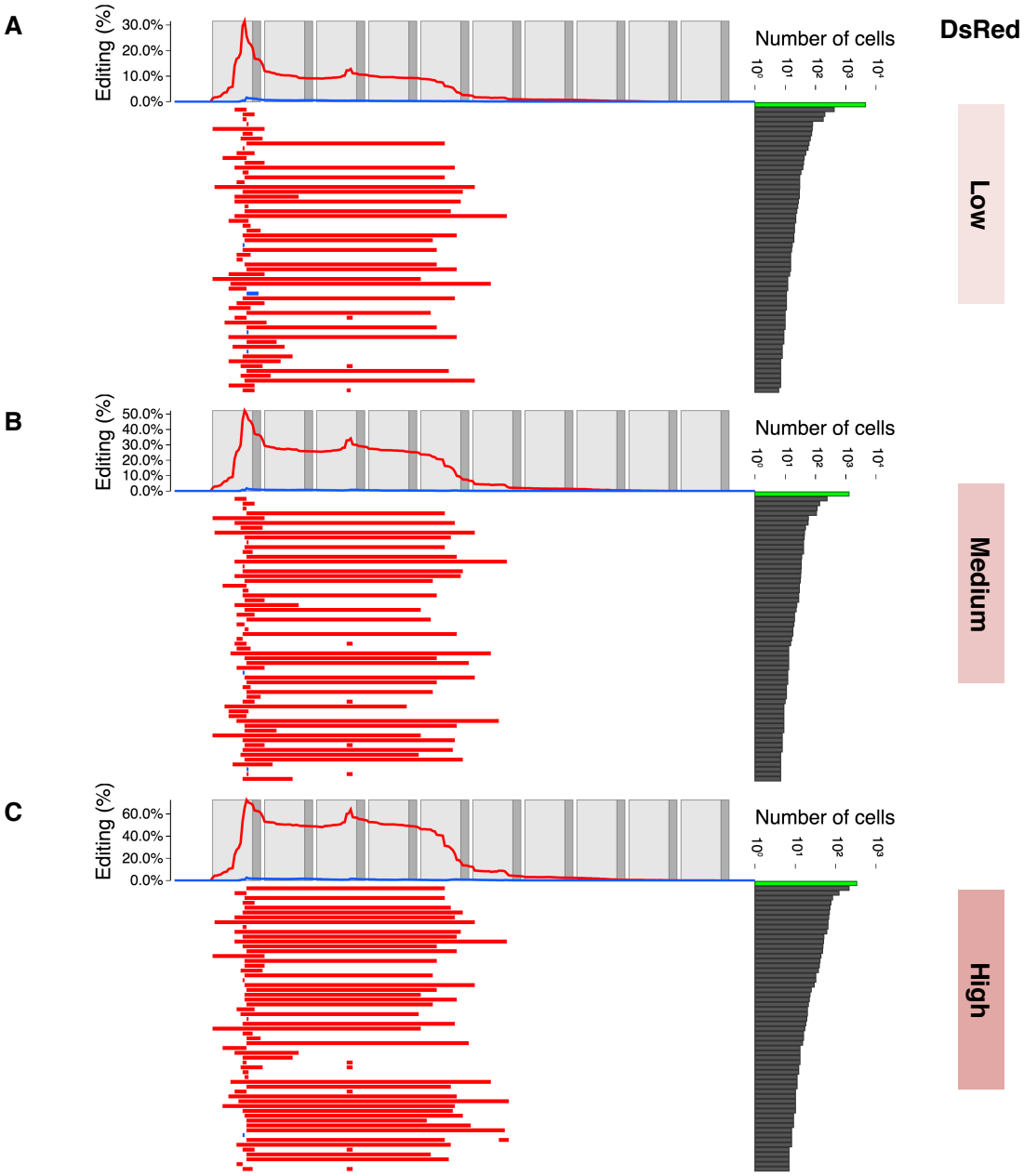



**Figure 5.3. Editing rates of the v1 barcode correlate with transfection efficiency.**

To evaluate how editing outcomes vary as a function of Cas9 expression, v1+ cells were co-transfected with pX330-v1 and a DsRed vector four days prior to sorting into bins based on DsRed expression and harvesting gDNA. The observed patterns of editing are shown for low (**A**), medium (**B**) and high (**C**) DsRed expression bins. Layout is as described in the **Figure 5.1B** legend, and top 60 alleles in each experiment are shown. Overall editing rates correlated with DsRed expression (low: 40%; medium: 69%; high: 91%), presumably reflecting transfection efficiency. The overall profile of edits observed remained approximately similar, although the proportion of inter-target deletions correlated with DsRed expression.

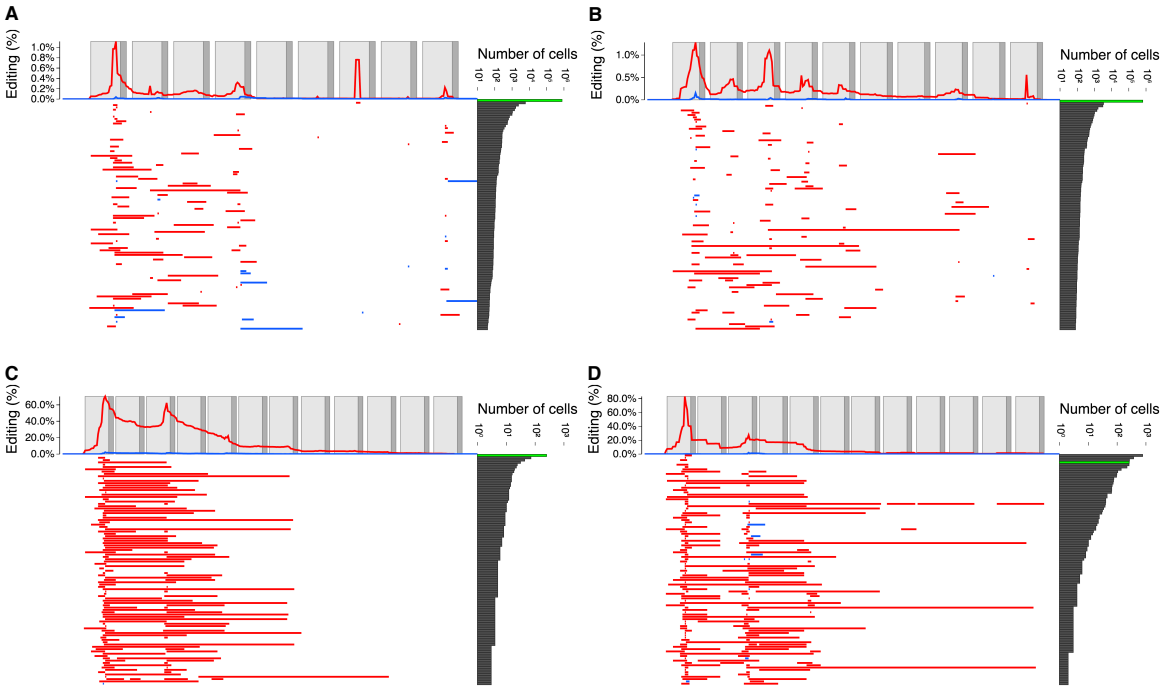



**Figure 5.4. Genome editing of alternative barcode designs.**

(**A, B**) In addition to v1, three additional barcode designs were tested, each with nine or ten weaker off-target sites for the same sgRNA (v2-v4). Editing observed in the v3 and v4 barcodes is shown above (panels (**A**) and (**B**), respectively; 100 alleles in each). Layout is as described in the **Figure 5.1B** legend. The weaker off-targets within these alternative barcode designs exhibit lower rates of editing than the v1 barcode, but also a much lower proportion of inter-target deletions. (**C**) As only a subset of targets were substantially edited in designs v1-v4, the most highly active targets were combined to a new, twelve target design (v5), consisting, in order, of v1 targets 1-6; v3 target 1; v2 targets 1, 2 and 5; and v4 targets 1 and 3. Editing observed in the v5 barcode is shown in panel (**C**), with layout as described in the **Figure 5.1B** legend. (**D**) To evaluate whether the means by which editing reagents are introduced impacts the rate and pattern of edits to the barcode, a lentiviral vector expressing Cas9 and the same sgRNA was introduced to cells prior to integration of the v5 barcode. After two weeks of culturing a population bottlenecked to 200 cells by FACS, diverse barcodes were observed but with substantially fewer inter-site deletions than with episomal delivery of editing reagents, i.e. (**C**): episomal expression of Cas9 and sgRNA vs. (**D**): lentiviral expression of Cas9 and sgRNA, both editing the v5 barcode.

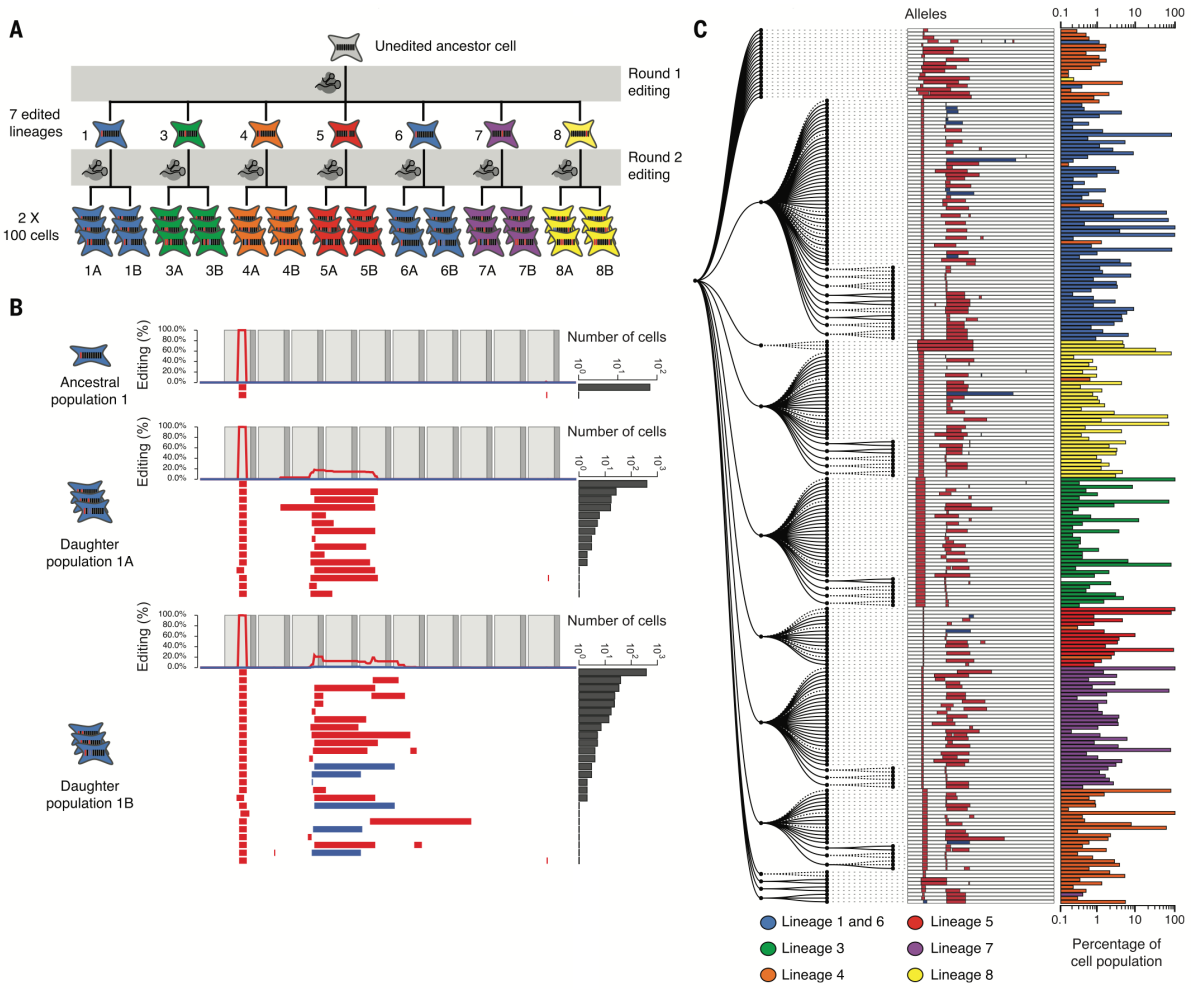



**Figure 5.5. Reconstruction of a synthetic lineage based on genome editing and targeted sequencing of edited barcodes.**

**(A)** A monoclonal population of cells was subjected to editing of the v1 array. Single cells were expanded, sampled (#1 to #12), re-transfected to induce a second round of barcode editing, and then expanded and sampled from 100-cell subpopulations (#1a, 1b to #12a, 12b). For clarity, the five clones where the original population was unedited are not shown. **(B)** Alleles observed in the synthetic lineage experiment are shown, with layout as described in the **Figure 5.1B** legend. Cell population #1 represents sampling of cells that had been subjected to only the first round of editing; virtually all cells contain a shared edit to the first target. Populations #1a and #1b are derived from #1 but subjected to a second round of editing prior to sampling. These retain the edit to the first target, but subpopulations bear additional edits to other targets. **(C)** Maximum parsimony reconstruction using PHYLIP Mix (**Figure 5.6B**) from alleles seen two or more times in the seven cell lineages represented in panel (**A**). Lineage membership and abundance of each allele are shown on the right. Progenitor cell lineage #4 (orange) appears to be derived from two cells, one edited and the other wild-type: only 62% of lineage #4 falls into a single clade, consistent with the proportion (64%) of the lineage edited after the first round. We assume that cells unedited in the first round either accrued edits matching other lineages (thus causing mixing), or accrued different edits (thus remaining outside the major clades).

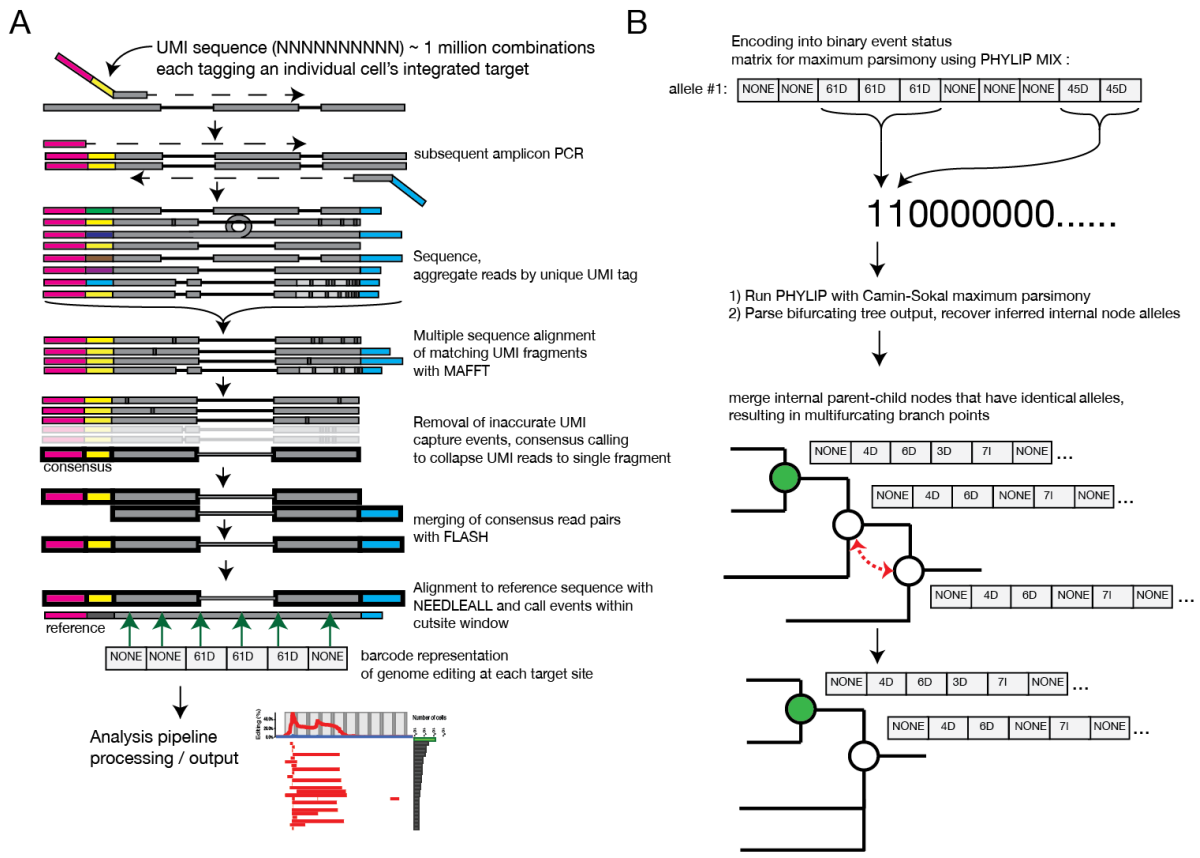



**Figure 5.6. Counting edited barcode alleles with unique molecular identifiers (UMIs) and building lineage trees by maximum parsimony.**

(**A**) Single genomic copies of barcodes, each derived from a single cell, were tagged by performing either one or two polymerase extension cycles using a single primer with ten degenerate bases, *i.e.* a unique molecular identifier, or UMI. Amplicon sequencing reads were initially cleaned to remove low quality bases using the Trimmomatic software package. Sequencing reads were then aggregated using this UMI tag, and a consensus read was created by aligning matched UMI reads using the MAFFT aligner. These consensus reads were then merged using the FLASH bioinformatics tool. Both the consensus merged reads as well as any unmerged read pairs were aligned to the reference sequence, and insertions and deletions over target sites were called for each UMI-specific barcode. These barcode calls were used for downstream analysis. (**B**) Maximum parsimony was performed as described in the Methods. Briefly, individual alleles were converted to an indicator matrix, and maximum parsimony reconstruction was performed with the PHYLIP Mix program. The output was parsed to recover inferred ancestral alleles. To reduce the number of bifurcations in the tree, internal nodes with identical alleles were merged. Parent / child pairs internal to the tree that shared an inferred allele were first identified. When such a pair was found, the grandchildren nodes were moved to the parent node, and the child node was removed. This was repeated until no such pairs could be found. The resulting multifurcating tree was then visualized with custom scripts.

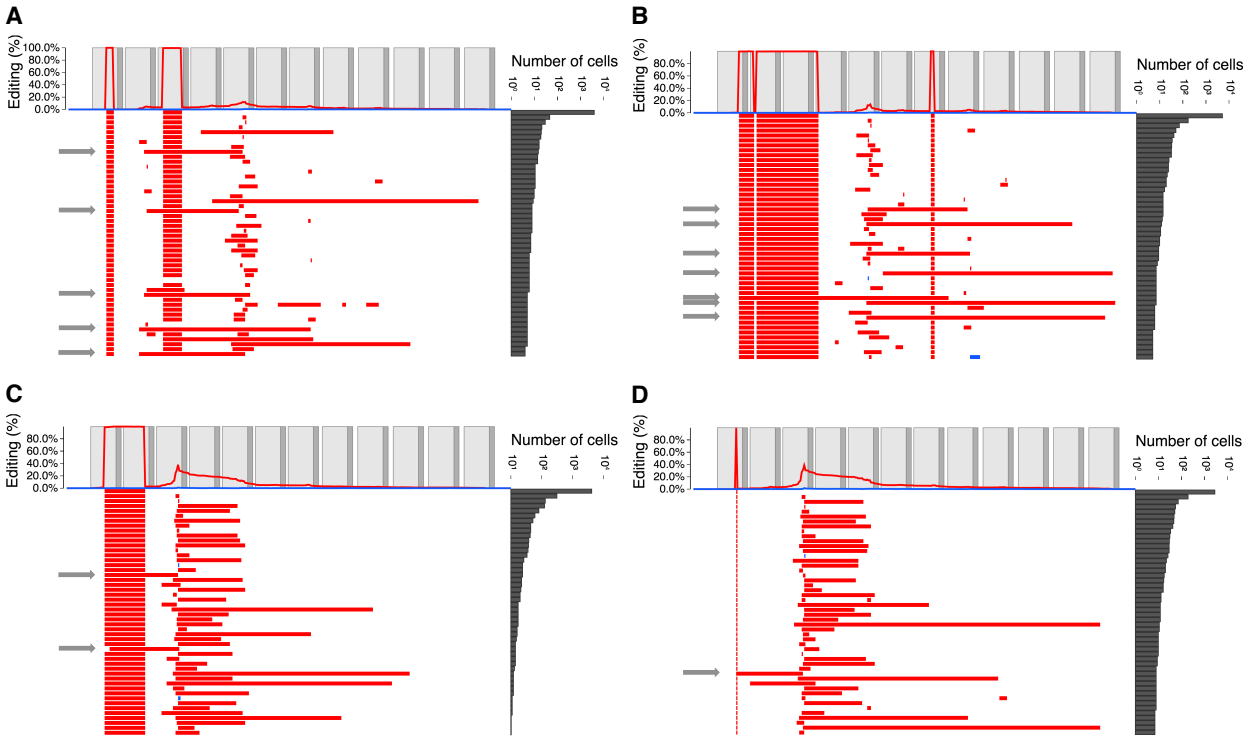



**Figure 5.7. Low frequency elimination of lineage-specific edits by re-editing of the v5 barcode in cell culture.**

A population of HEK293T cells bearing the unedited v5 barcode was subjected to initial editing by transfection with pX330-v1. Monoclonal populations containing edited v5 barcodes were then cultured and re-transfected twice to induce additional editing. The outcomes of re-editing are shown for each population with barcode editing plots (as in **Figure 5.1B**) showing the top 50 alleles. In each, the top allele corresponds to the parental allele verified by sequencing each monoclonal line. We observed a variety of mechanisms in which an established edit was lost, examples of which are highlighted with gray arrows. (**A**) Loss of an existing deletion at site 3 occurred in 4.2% of cells (16.7% of cells with barcodes that were re-edited). Loss of this edit appears to have arisen from simultaneous Cas9 cleavage at site 2 and any of sites 4 through 10, thus forming a larger deletion spanning the original site 3 deletion. (**B**) An initial site 7 deletion was removed by re-editing in 3.0% of cells (12.7% of re-edited barcodes), likely due to the same mechanism describe in A. (**C**) A single deletion event spanning sites 1 and 2 was disrupted in 3.3% of cells (7.6% of re-edited barcodes). These alleles likely formed by deletions at sites 3 that extended as far as or beyond the ancestral deletion in sites 1 and 2. (**D**) Re-editing of a target site bearing a single 1 bp deletion occurred in 0.9% of cells (1.9% of re-edited barcodes), presumably consequential to residual Cas9 activity at the edited site. These experiments highlight examples of information loss in the v5 barcode, but understanding how often this occurs in different barcode designs and editing systems remains to be determined.

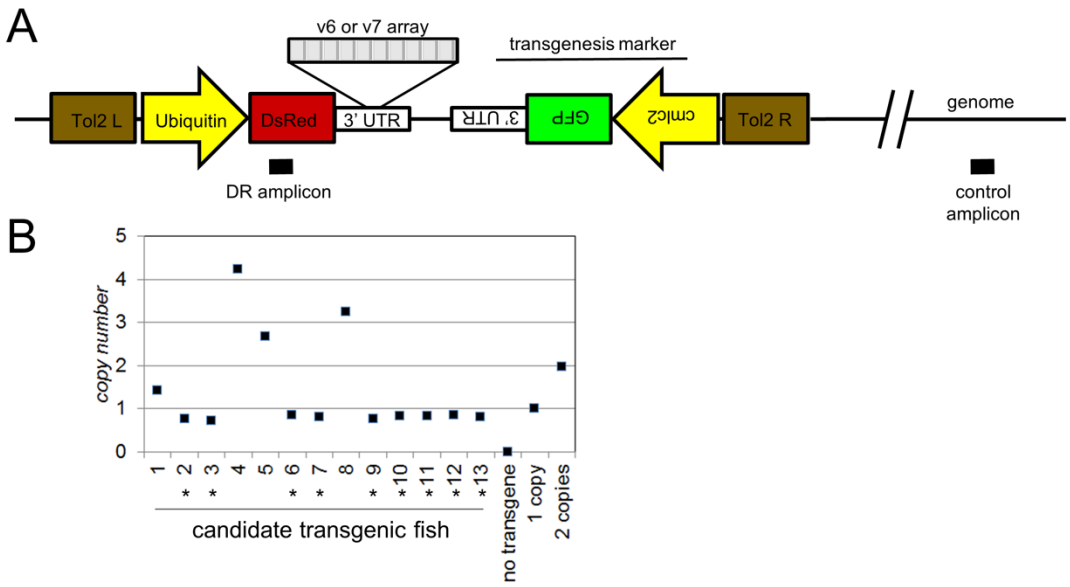



**Figure 5.8. Generation of single copy transgenic v6 or v7 zebrafish.**

(**A**) Diagram of the barcode transgene with Tol2 integration arms, a Ubiquitin promoter upstream of DsRed with either the v6 or v7 barcode embedded in the 3' UTR, and a cmlc2:GFP transgenesis marker. (**B**) Quantification of transgene copy number by qPCR using DR amplicon and control amplicon as indicated in (A). Copy number was determined using the ddCt method and reference non-transgenic, 1-copy and 2-copy transgenic animals. Putative single copy individuals are indicated by asterisks (*).

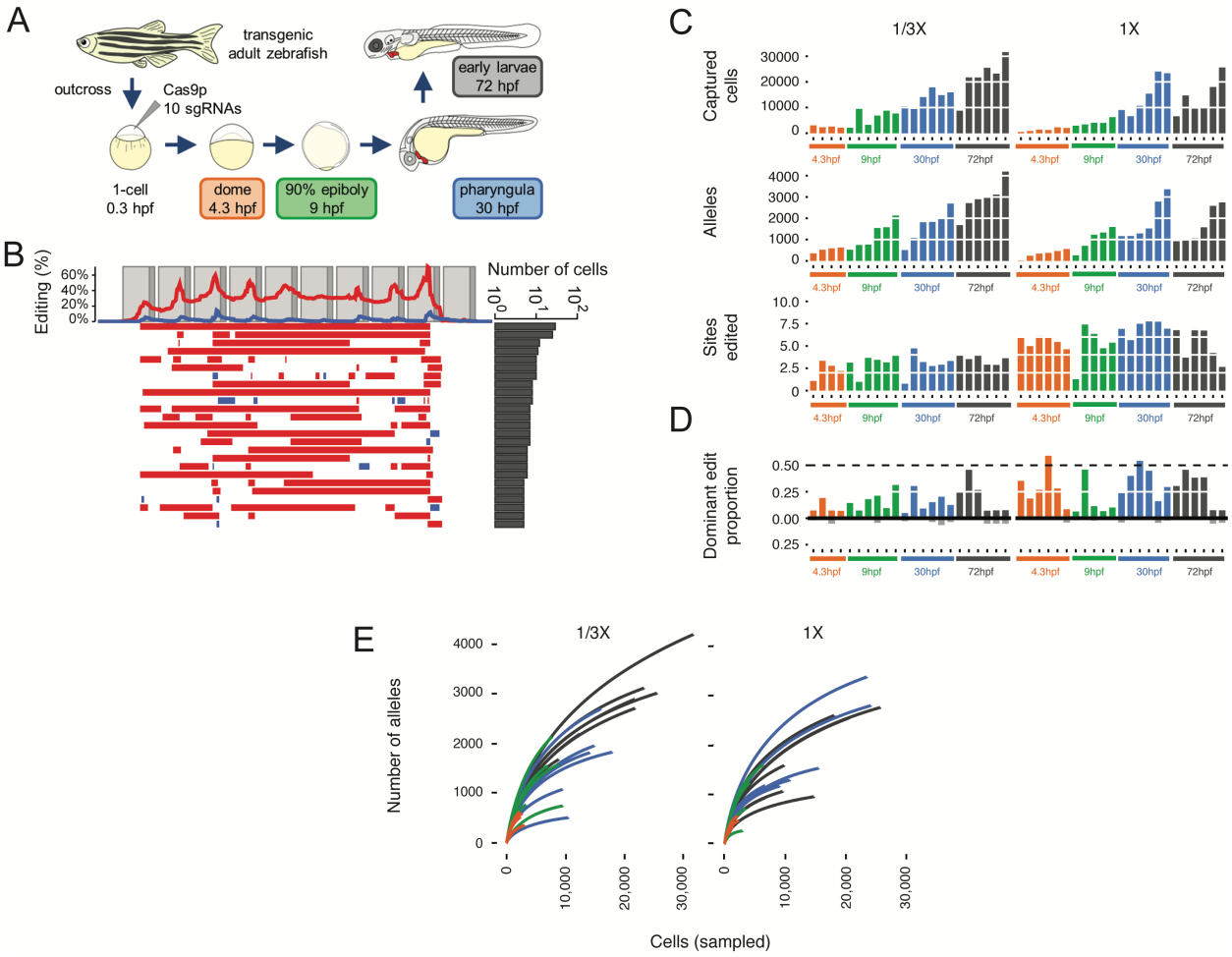



**Figure 5.9. Generating combinatorial barcode diversity in transgenic zebrafish.**

**(A)** One-cell zebrafish embryos were injected with complexed Cas9 ribonucleoproteins (RNPs) containing sgRNAs that matched each of the 10 targets in the array (v6 or v7). Embryos were collected at time points indicated. UMI-tagged barcodes were amplified and sequenced from genomic DNA. **(B)** Patterns of editing in alleles recovered from a 30 hpf v6 embryo, with layout as described in the **Figure 5.1B** legend. **(C)** Bar plots show the number of cells sampled (top), unique alleles observed (middle) and proportion of sites edited (bottom) for 45 v7 embryos collected at four developmental time-points and two levels of Cas9 RNP (1/3x, 1x). Colors correspond to stages shown in panel (A). Although more alleles are observed with sampling of larger numbers of cells at later time points, the proportion of target sites edited remains relatively constant. **(D)** Bar plots show the proportion of edited barcodes containing the most common editing event in a given embryo. Six of 45 embryos had the most common edit in approximately 50% of cells (dashed line), consistent with this edit having occurred at the two-cell stage. Colors correspond to stages shown in panel (A). These same edits are rarer or absent in other embryos (black bars below). **(E)** For each of the 45 v7 embryos, all barcodes observed were sampled without replacement. The cumulative number of unique alleles observed as a function of the number of cells sampled is shown (average of the 500 iterations shown per embryo; two levels of Cas9 RNP: 1/3x on left, 1x on right). The number of unique alleles observed, even in later developmental

stages where we are sampling much larger numbers of cells, appears to saturate, and there is no consistent pattern supporting substantially greater diversity in later time-points, consistent with the bottom row of panel (C) in supporting the conclusion that the majority of editing occurs before dome stage.

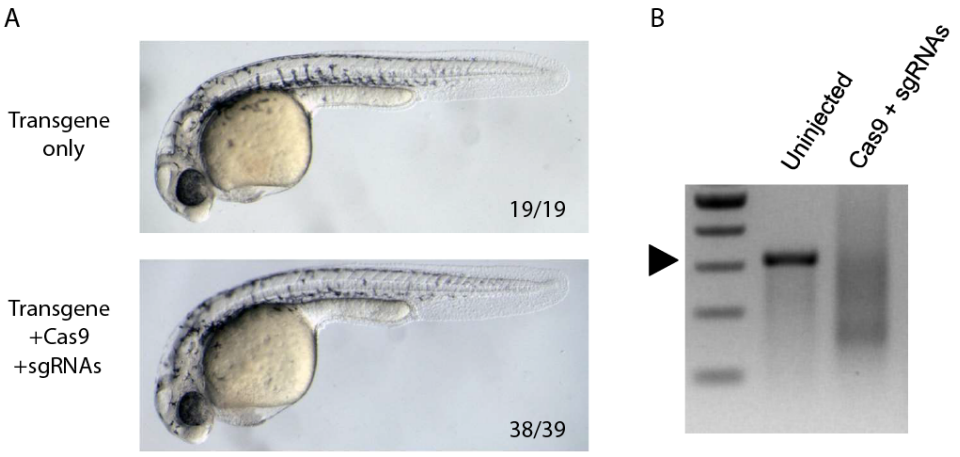



**Figure 5.10. Barcode editing in transgenic zebrafish embryos is robust and does not affect development.**

(**A**) Representative v6 transgenic embryos uninjected or injected with Cas9 protein pre-complexed with a mix of 10 sgRNAs are shown, with phenotypic penetrance indicated. (**B**) Gel electrophoresis of v6 array PCR products from uninjected and injected embryos show extensive barcode editing as a smear of smaller molecules running predominantly below the expected size of the unedited barcode (arrowhead, 311 bp).

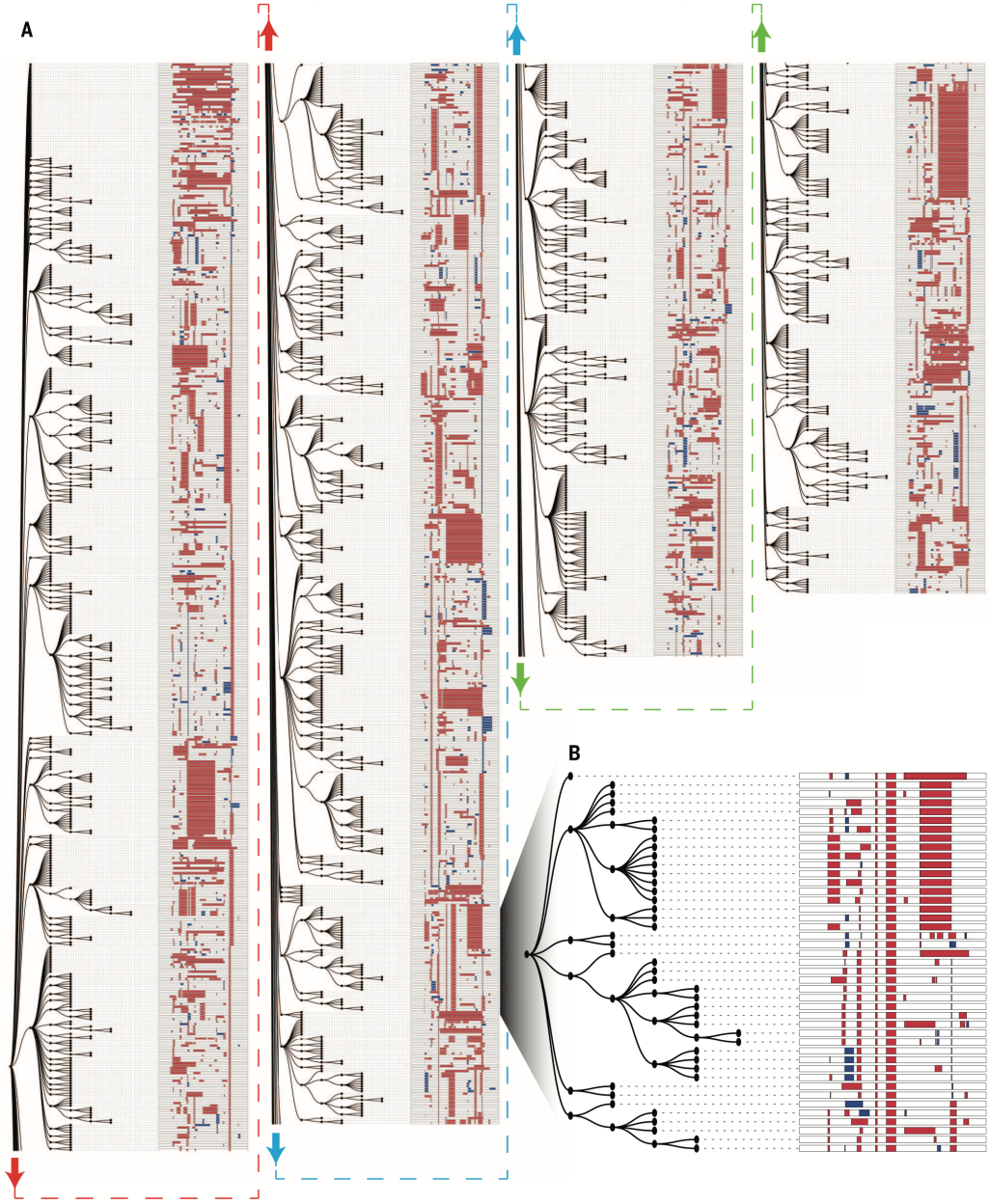



**Figure 5.11. Lineage reconstruction of an edited zebrafish embryo.**

**(A)** A lineage reconstruction of 1,323 alleles recovered from the v6 embryo also represented in **Figure 5.9B**, generated by a maximum parsimony approach implemented in the PHYLIP Mix package (see **Methods**). A dendrogram to the left of each column represents the lineage relationships, and the alleles are represented on the right. Each row represents a unique allele. Matched colored arrows and dashed lines connect subsections of the tree together. There are many large clades of alleles sharing specific edits, as well as sub-clades defined by 'dependent' edits. These dependent edits occur within a clade defined by a more frequent edit but are rare or absent elsewhere in the tree. **(B)** A portion of the tree is shown at higher resolution. Two edits are shared by all alleles in this clade. Six independent edits define descendent sub-clades within this clade, and further edits define additional sub-sub-clades within the clade.

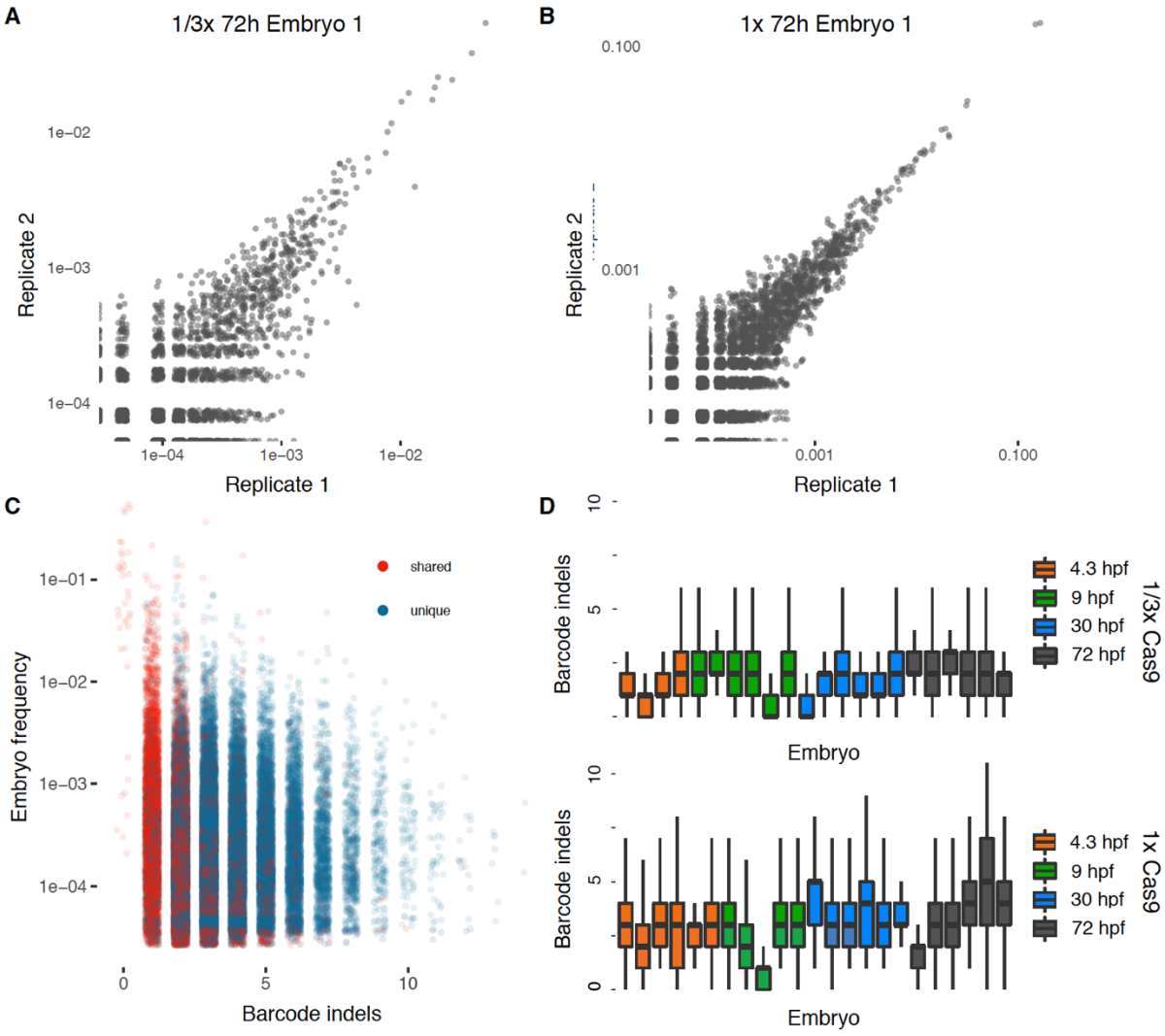



**Figure 5.12. Characteristics of Cas9-mediated barcode editing in zebrafish embryos.**

After isolation of genomic DNA from each of two 72 hpf embryos injected with either **(A)** 1/3x volume or **(B)** 1x volume, the material was split and two separate amplification reactions performed, replicates 1 and 2. Unique Molecular Identifiers (UMIs) were used to tag genomic copies of embryo barcodes, such that each UMI's consensus sequencing call corresponds to a single cell's edited barcode. For each embryo, allele frequencies from UMI-tagging technical replicates are plotted against each other (*i.e.* technical replicates of one embryo in panel (**A**), and technical replicates of another embryo in panel (**B**)). Each point corresponds to a single barcode present in the union of the replicates. Pearson correlations: (**A**) = 0.96, (**B**) = 0.998. Spearman correlations: (**A**) = 0.42, (**B**) = 0.64. (**C**) To compare barcodes that were shared between embryos to those that were unique to a single embryo, we plotted each allele's indel count against its proportion on a per embryo basis (n = 45). Alleles that were shared had significantly fewer indels than those that were seen in only one embryo (2.01 mean indels per shared allele vs. 3.52 mean indels per embryo-specific allele (Wilcoxon Rank Sum (WRS); $P << 0.00001$)). The mean frequency within an embryo of a shared allele was modestly higher than a unique allele: 0.12%

vs. 0.052%, respectively (WRS; $P = 3.8x10^{-25}$). (**D**) Boxplots show the distribution of barcode indel events per embryo. (Bold line; median, box; 25th to 75th percentile; whiskers extend to the furthest point within 1.5x the interquartile range (IQR) of the box; outliers not shown).

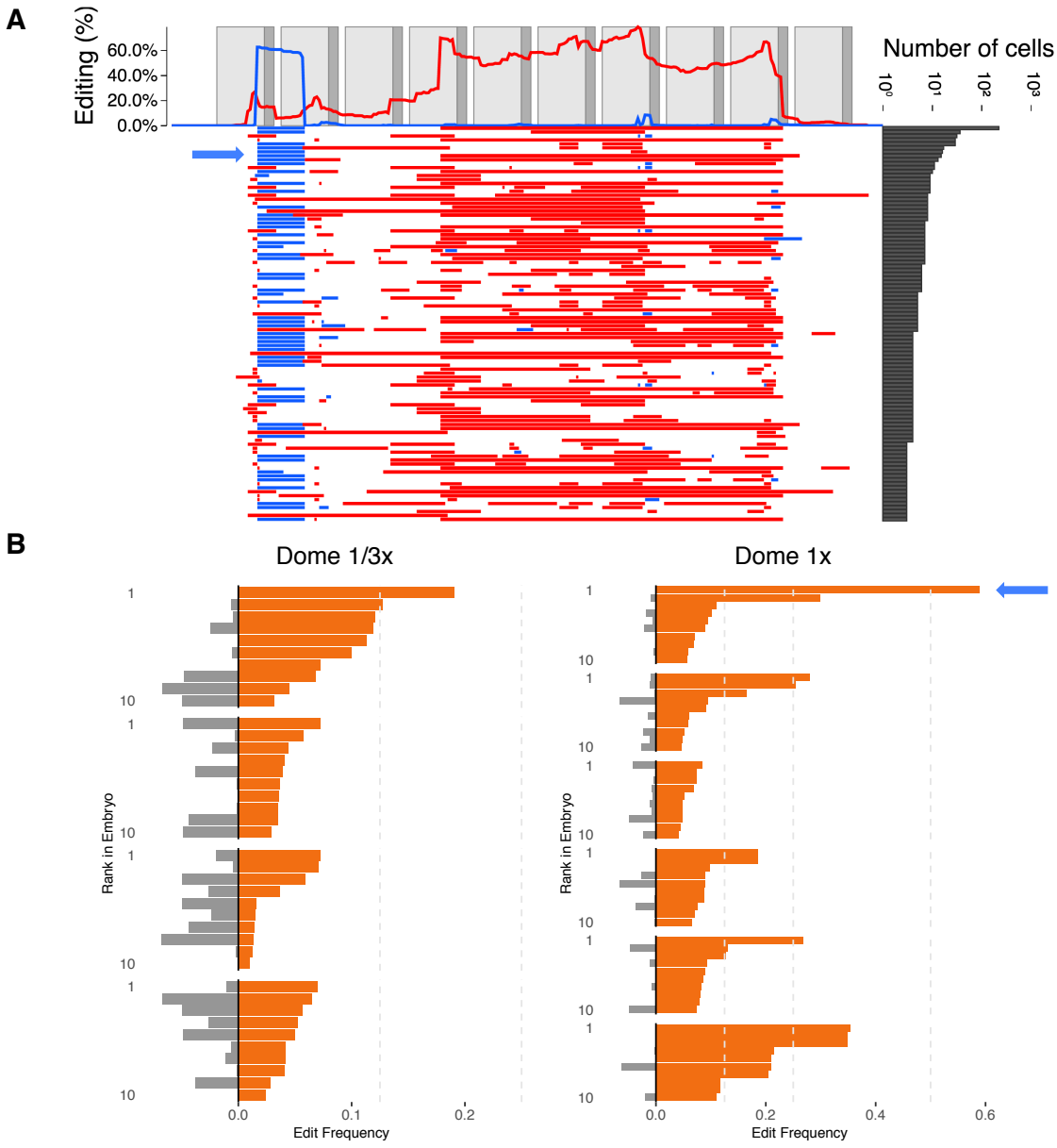



**Figure 5.13. Abundances of the most common editing events in each embryo often reflect the onset of editing.**

(**A**) A barcode editing plot showing the top 50 alleles from one dome stage embryo (1x #1) exemplifies a high frequency editing event – in this case a 20 bp insertion at the first target site (blue bar indicated by arrow). The event is seen in 59.0% of barcodes and was absent from all other embryos, suggesting it derived from editing at the two-cell stage. (**B**) The frequencies of the top ten indels from each of ten dome stage embryos are shown in (orange), plotted next to the average frequency of that indel in all other embryos (gray). Cases in which the edit is common in one embryo and rare in others strongly suggest the edit's abundance resulted from occurring early in development, and not from stereotypical double-strand break repair outcomes during to barcode editing. The arrow indicates the event corresponding to the common insertion shown in (**A**).

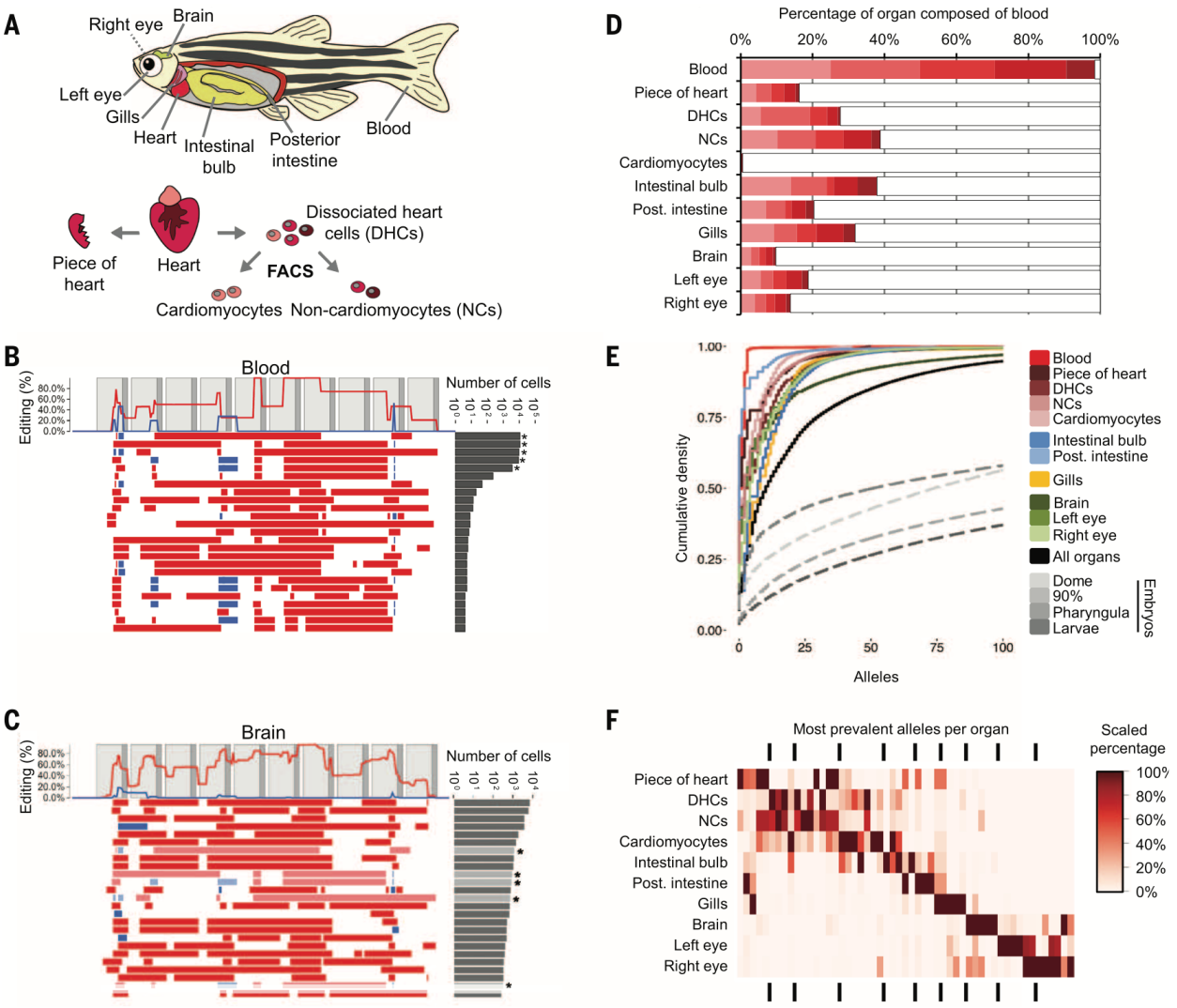



**Figure 5.14. Organ-specific progenitor cell dominance.**

**(A)** The indicated organs were dissected from a single adult v7 transgenic edited zebrafish (ADR1). A blood sample was collected as described in the Methods. The heart was further split into the four samples shown (**Figure 5.15**). **(B)** Patterns of editing in the most prevalent 25 alleles (out of 135 total) recovered from the blood sample. Layout as described in the **Figure 5.1B** legend. The most prevalent 5 alleles (indicated by asterisks) comprise >98% of observed cells. **(C)** Patterns of editing in the most prevalent 25 alleles (out of 399 total) recovered from brain. Layout as described in the **Figure 5.1B** legend. Alleles that have identical editing patterns compared to the most prevalent blood alleles are indicated by asterisks and light shading. **(D)** The five dominant blood alleles (shades of red) are present in varying proportions (10-40%) in all intact organs except the FACS-sorted cardiomyocyte population (0.5%). All other alleles are summed in grey. **(E)** The cumulative proportion of cells (y-axis) represented by the most frequent alleles (x-axis) for each adult organ of ADR1 is shown, as well as the adult organs in aggregate. In all adult organs except blood, the five dominant blood alleles are excluded. All organs exhibit dominance of sampled cells by a small number of progenitors, with fewer than 7 alleles comprising the majority of cells. For comparison, a similar plot for the median embryo (dashed) from each time-point of the

developmental time course experiment is also shown. **(F)** The distribution of the most prevalent alleles for each organ, after removal of the five dominant blood alleles, across all organs. The most prevalent alleles were defined as being at >5% abundance in a given organ (median 5 alleles, range 4-7). Organ proportions were normalized by column and colored as shown in legend.

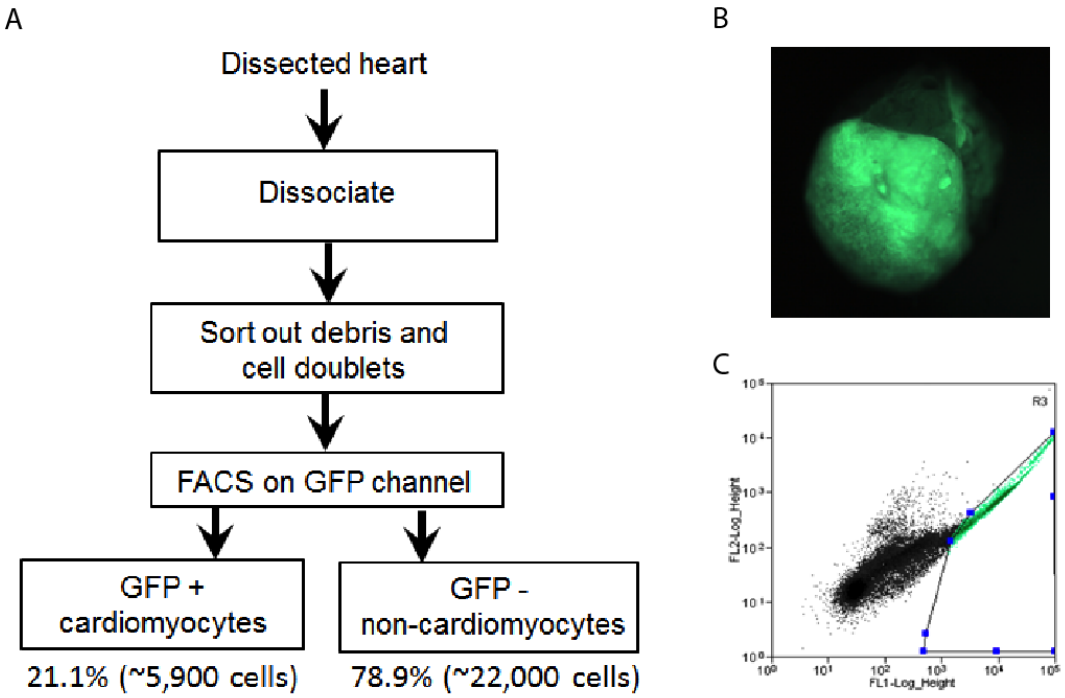



**Figure 5.15. FACS sorting of cardiomyocytes and non-cardiomyocyte heart cells.**

(**A**) Schematic. Adult transgenic hearts (example in **B**) were dissected, dissociated, and sorted via FACS. Gates were applied to remove cellular debris (exclude high SSC-H, low FSC-H, keep 15.6% of events as cells) and cell doublets (exclude high SSC-W, keep 97.2% of events as single cells) before gating on the ratio of GFP:RFP fluorescence (**C**) to sort GFP+ cardiomyocytes from GFP- non-cardiomyocyte heart cells (21.1% GFP+, 78.9% $GFP^-$). Percentages provided are for ADR1.

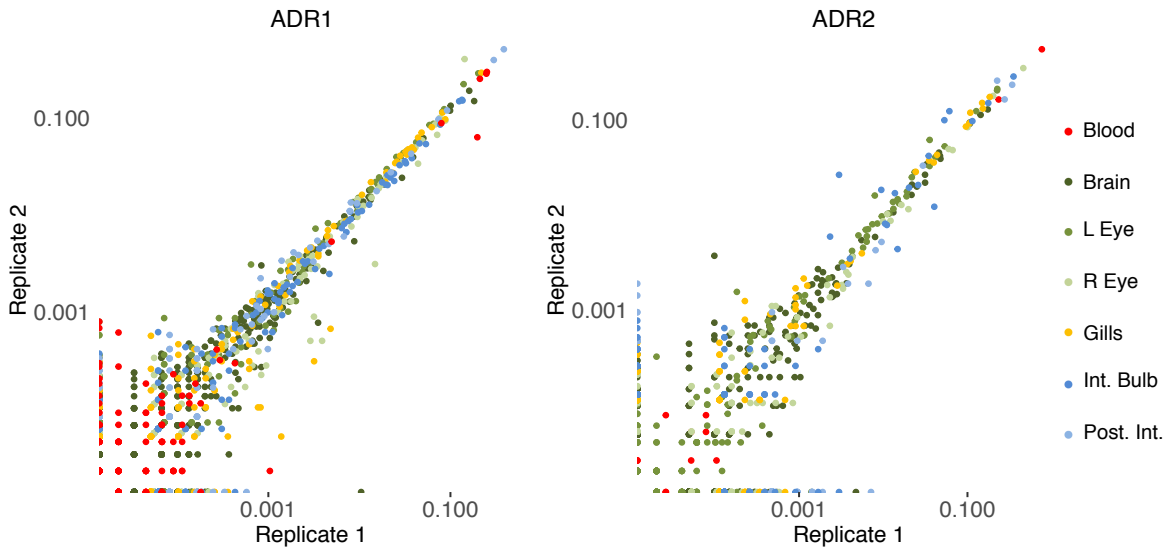



**Figure 5.16. Reproducibility of barcode sampling from adult zebrafish organs.**

After isolation of genomic DNA, two separate amplification reactions (Replicate 1 and Replicate 2) were performed for each of seven organs from ADR1 and ADR2. UMIs were used to tag genomic copies of the barcodes. Replicate samples were sequenced in separate runs, and UMI consensus read thresholds were set proportional to sequencing depth. For each organ in each fish, allele frequencies for each replicate are plotted against each other. Each point corresponds to a single barcode present in the union of the two replicate samplings, colored by organ. Pearson correlations calculated from log10-transformed values are: ADR1 = 0.90, ADR2 = 0.85. For this analysis, the top five (ADR1) or two (ADR2) blood alleles were computationally removed from non-blood samples.

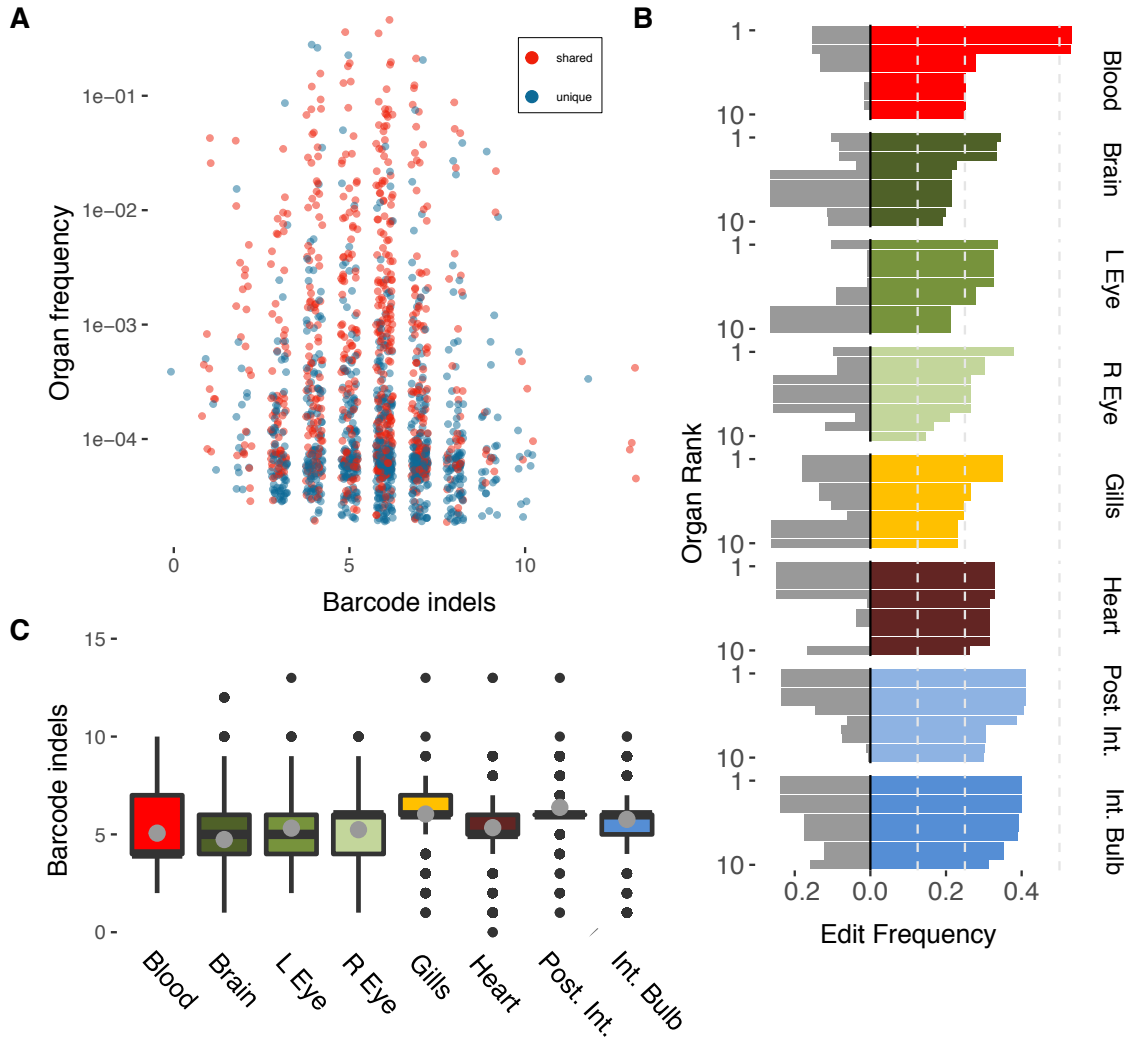



**Figure 5.17. Barcode editing characteristics in organs from adult zebrafish ADR1.**

For these analyses, the top five blood alleles were computationally removed from non-blood samples. (**A**) Sharing of alleles across organs as a function of frequency within organs and the number of indels. Compared to barcodes shared between unrelated embryos by chance recurrence (**Figure 5.12**C), alleles contributing to multiple organs (red) harbor more indels and are seen at higher proportions compared to alleles restricted to a single organ (blue), suggesting sharing is largely explained by way of developmental lineages as opposed to by chance. (**B**) Frequencies of the top ten indel events per organ (colored bars on right) plotted against the average frequency of that edit in all other organs (gray bars on left). The most common edits within organs are also seen frequently in other organs. (**C**) Boxplots show the distribution of barcode indel events within each organ from ADR1. (Bold line; median, box; 25th to 75th percentile; whiskers extend to the furthest point within 1.5x the IQR from the box, gray dot; mean value, outliers not shown)

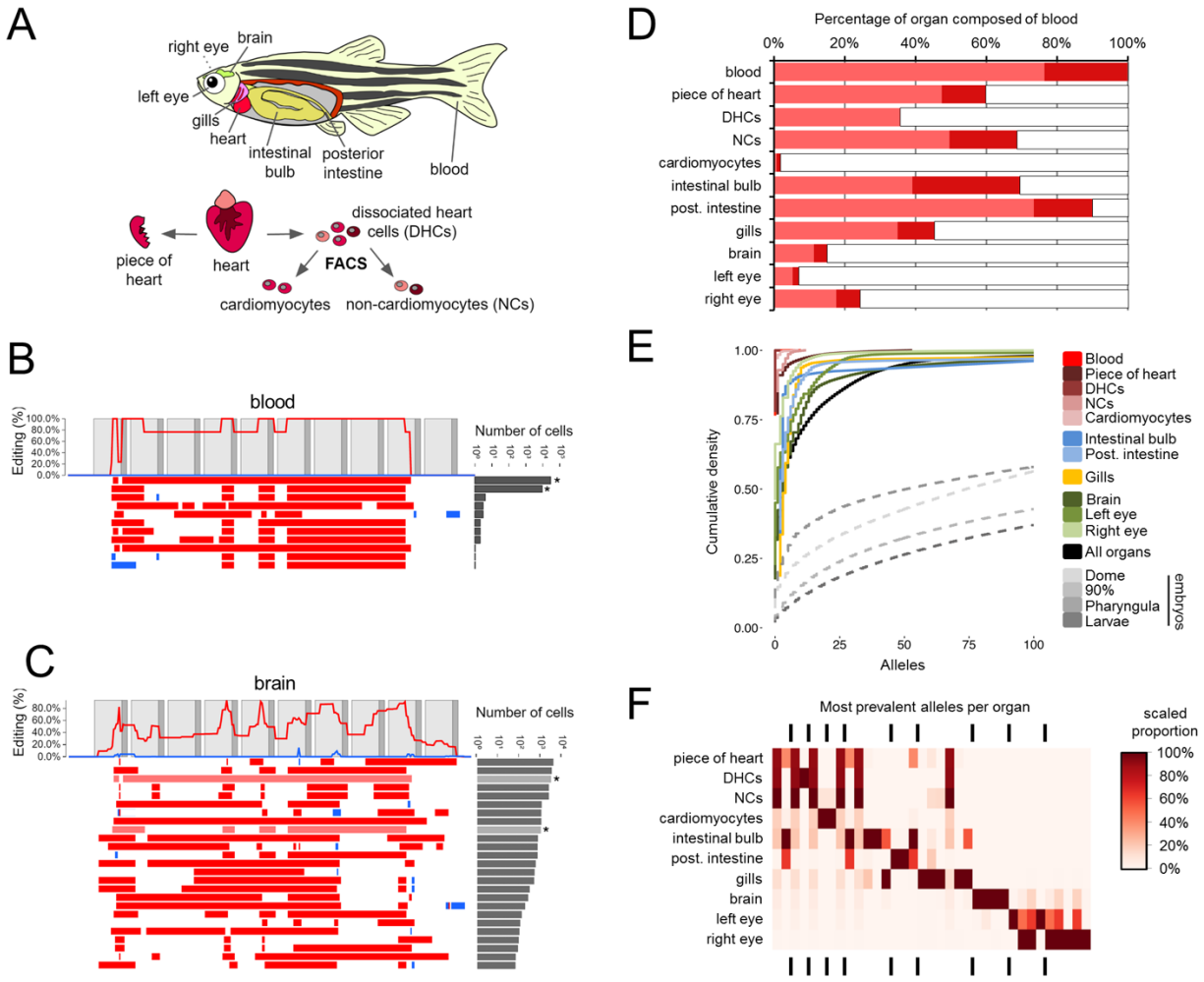



**Figure 5.18. Organ-specific progenitor cell dominance in ADR2.**

**(A)** The indicated organs were dissected from a single adult v7 transgenic edited zebrafish (ADR1). A blood sample was collected as described in **Methods**. The heart was further split into the four samples shown (**Figure 5.17**). **(B)** Patterns of editing in the 11 alleles recovered from the blood sample. Layout as described in the **Figure 5.1B** legend. The most prevalent 2 alleles (indicated by asterisks) comprise >98% of observed cells. **(C)** Patterns of editing in the most prevalent 25 alleles (out of 699 total) recovered from brain. Layout as described in the **Figure 5.1B** legend. Alleles that are identical in sequence to the most prevalent blood alleles are indicated by asterisks and light shading. **(D)** The five dominant blood alleles (shades of red) are present in varying proportions (7-90%) in all intact organs except the FACS-sorted cardiomyocyte population (2%). All other alleles are summed in grey. **(E)** The cumulative proportion of cells (y-axis) represented by the most frequent alleles (x-axis) for each adult organ of ADR1 is shown, as well as the adult organs in aggregate. In all adult organs except blood, the five dominant blood alleles are excluded. All organs exhibit dominance of sampled cells by a small number of progenitors, with fewer than 5 alleles comprising the majority of cells. For comparison, a similar plot for the median embryo (dashed) from each time-point of the developmental time course experiment is also shown. **(F)** The distribution of the most prevalent alleles for each organ, after

removal of the five dominant blood alleles, across all organs. The most prevalent alleles were defined as being at >5% abundance in a given organ (median 4 alleles, range 4-6). Organ proportions were normalized by column and colored as shown in legend.

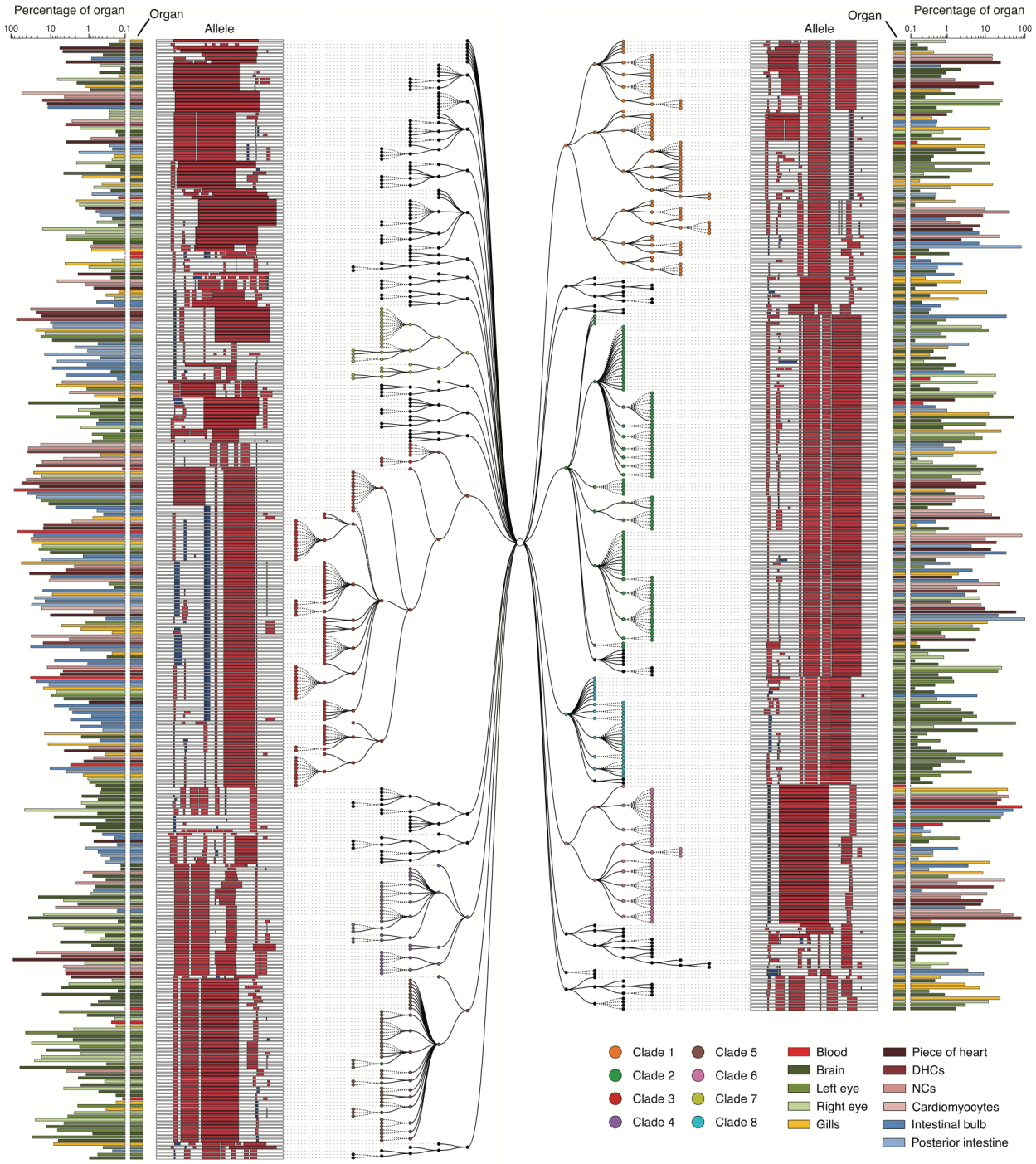



**Figure 5.19. Lineage reconstruction for adult zebrafish ADR1.**

Unique alleles sequenced from adult zebrafish organs can be related to one another using a maximum parsimony approach implemented in the PHYLIP Mix package (see **Figure 5.6B**). For reasons of space, we show a tree reconstructed from the 601 ADR1 alleles observed at least five times in individual organs. Eight major clades are displayed with colored nodes, each defined by 'ancestral' edits that are shared by all alleles assigned to that clade. Editing patterns in individual alleles are represented as shown previously. Alleles observed in multiple organs are plotted on

separate lines per organ and are connected with stippled branches. Two sets of bars outside the alleles identify the organ in which the allele was observed and the proportion of cells in that organ represented by that allele (log scale).

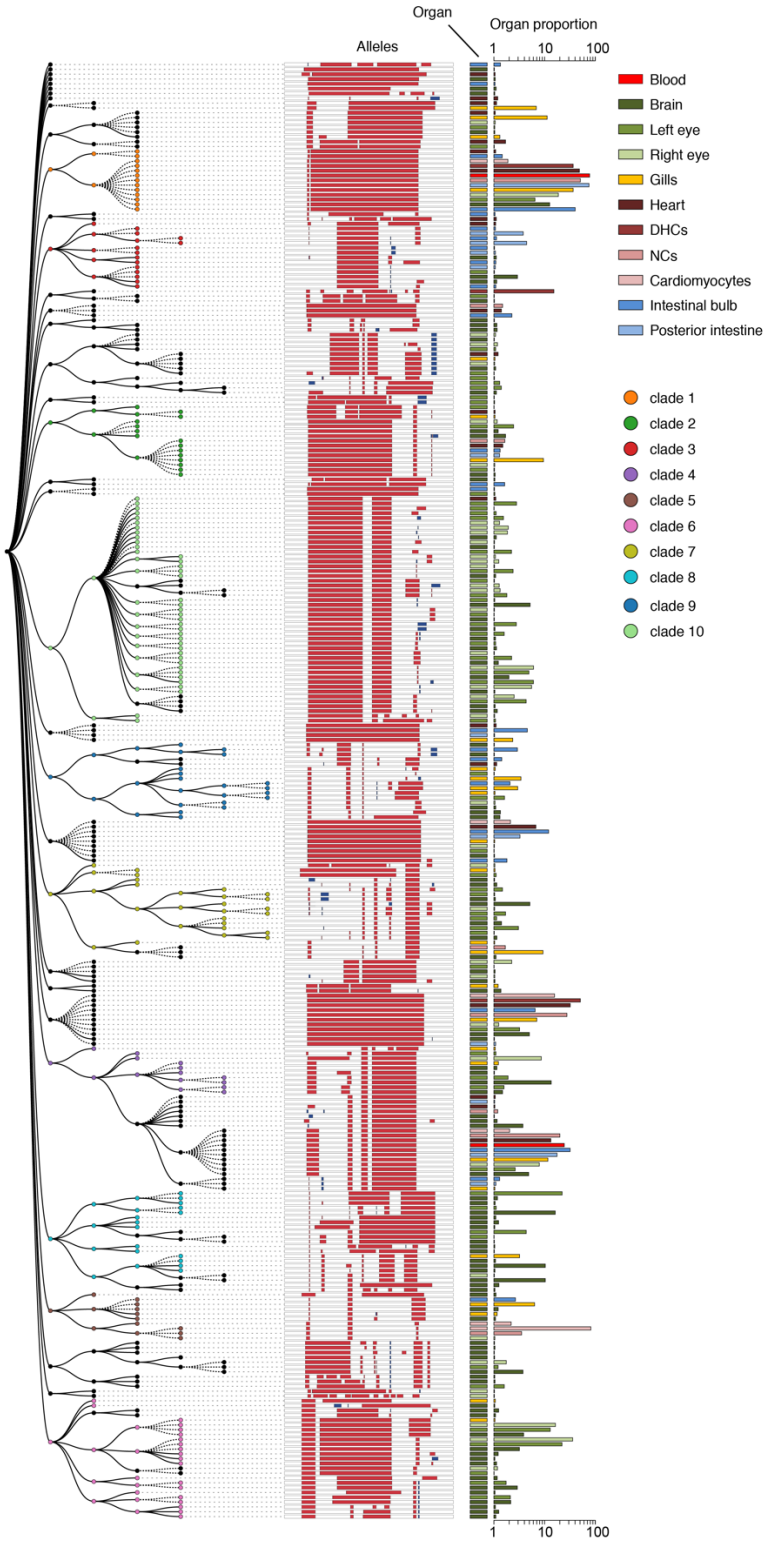



**Figure 5.20. Lineage reconstruction for adult zebrafish ADR2.**

Unique alleles sequenced from adult zebrafish organs can be related to one another using a

maximum parsimony approach into a multifurcating lineage tree. For reasons of space, we show a tree reconstructed from the 302 ADR2 alleles observed at least 5 times in individual organs. Ten major clades are displayed with colored nodes, each defined by 'ancestral' edits that are shared by all alleles assigned to that clade. Editing patterns in individual alleles are represented as shown previously. Alleles in multiple organs are plotted on separate lines per organ and these nodes connected with stippled branches. Two sets of bars outside the alleles identify the organ in which the allele was observed and the proportion of cells in that organ represented by that allele (log scale).

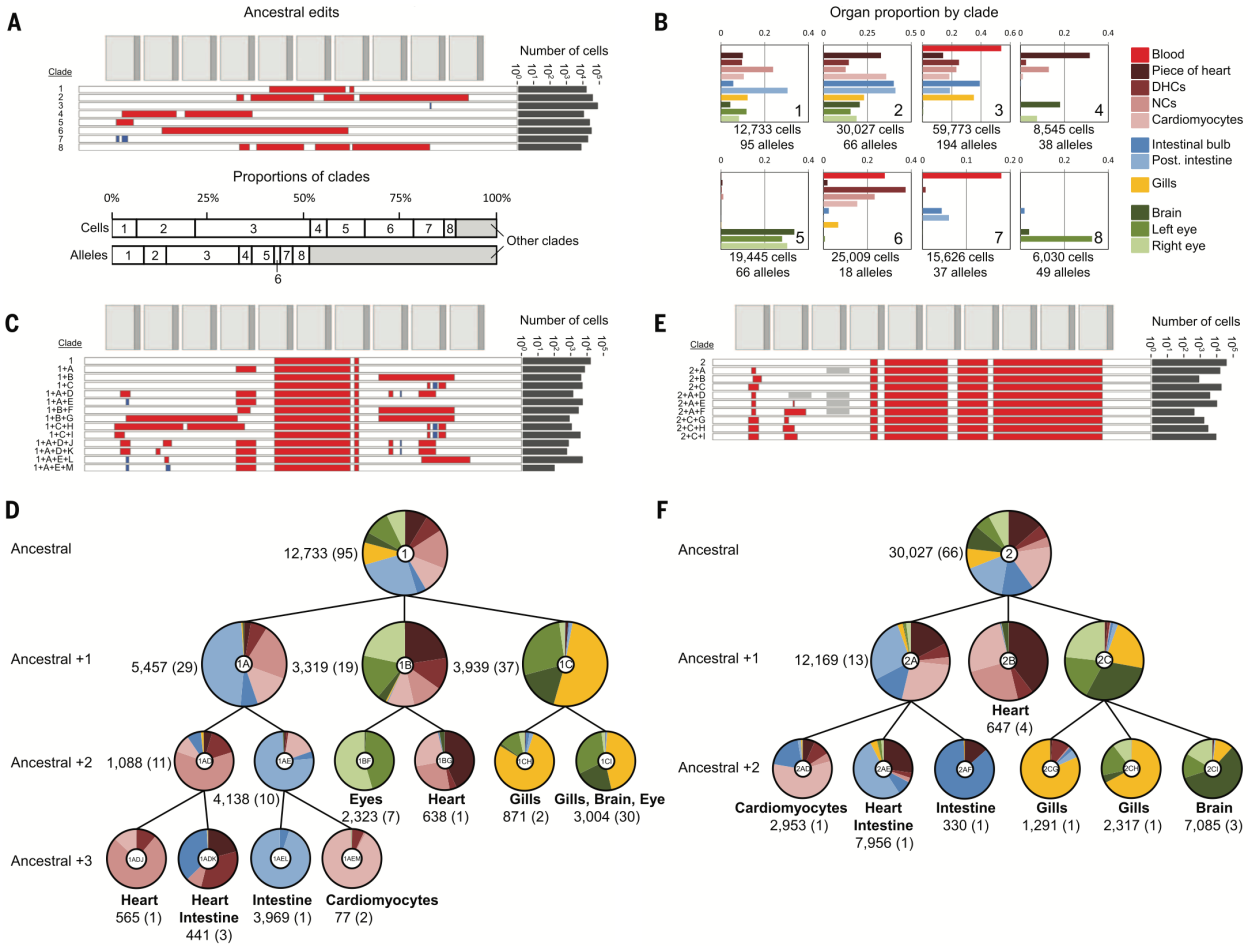



**Figure 5.21. Clades and subclades corresponding to inferred progenitors exhibit increasing levels of organ restriction.**

**(A)** Top panel: The parsimony inferred ancestral edits that define eight major clades of ADR1 are shown, with the total number of cells in which these are observed indicated on the right. Bottom panel: Contributions of the eight major clades to all cells or all alleles. 19 alleles (out of 1,138 total) that contained ancestral edits from more than one clade were excluded from assignment to any clade, and any further lineage analysis. **(B)** Contributions of each of the eight major clades to each organ, displayed as a proportion of each organ. To accurately display the contributions of the eight major clades to each organ, we first re-assigned the five dominant blood alleles from other organs back to the blood. The total number of cells and alleles within a given major clade are listed below. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCs' = non-cardiomyocyte heart cells. **(C,E)** Edits that define subclades of clade #1 (**C**) and clade #2 (**E**), with the total number of cells in which these are observed indicated on the right. A grey box indicates an unedited site or sites, distinguishing it from related alleles that contain an edit at this location. **(D,F)** Lineage trees corresponding to subclades of clade #1 (**D**) and clade #2 (**F**) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization that accounts for differential depth of sampling between organs. Labels in the center

of each pie chart correspond to the subclade labels in (**C**) and (**E**). Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), *i.e.* those comprising >25% of a subclade by proportion.

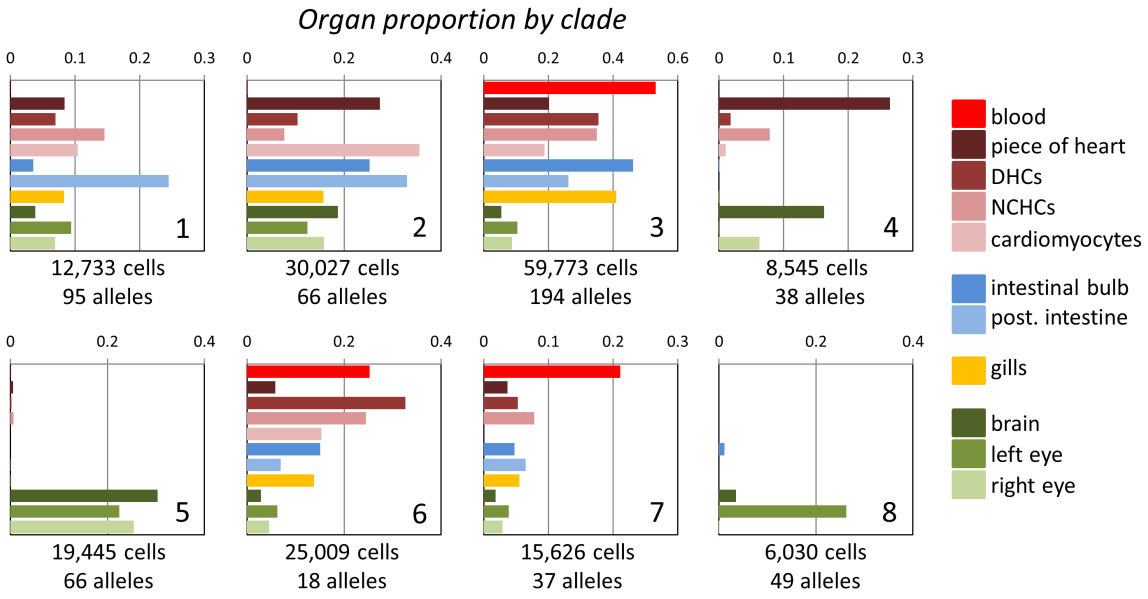



**Figure 5.22. Contributions of the eight major clades within ADR1 to each organ, prior to the reassignment of the most prevalent blood alleles.**

The total number of cells and unique alleles within a given major clade are listed below. **Figure 5.21B** shows similar information, but after the reassignment of the dominant blood alleles from all organs to blood. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCHCs' = non-cardiomyocyte heart cells.

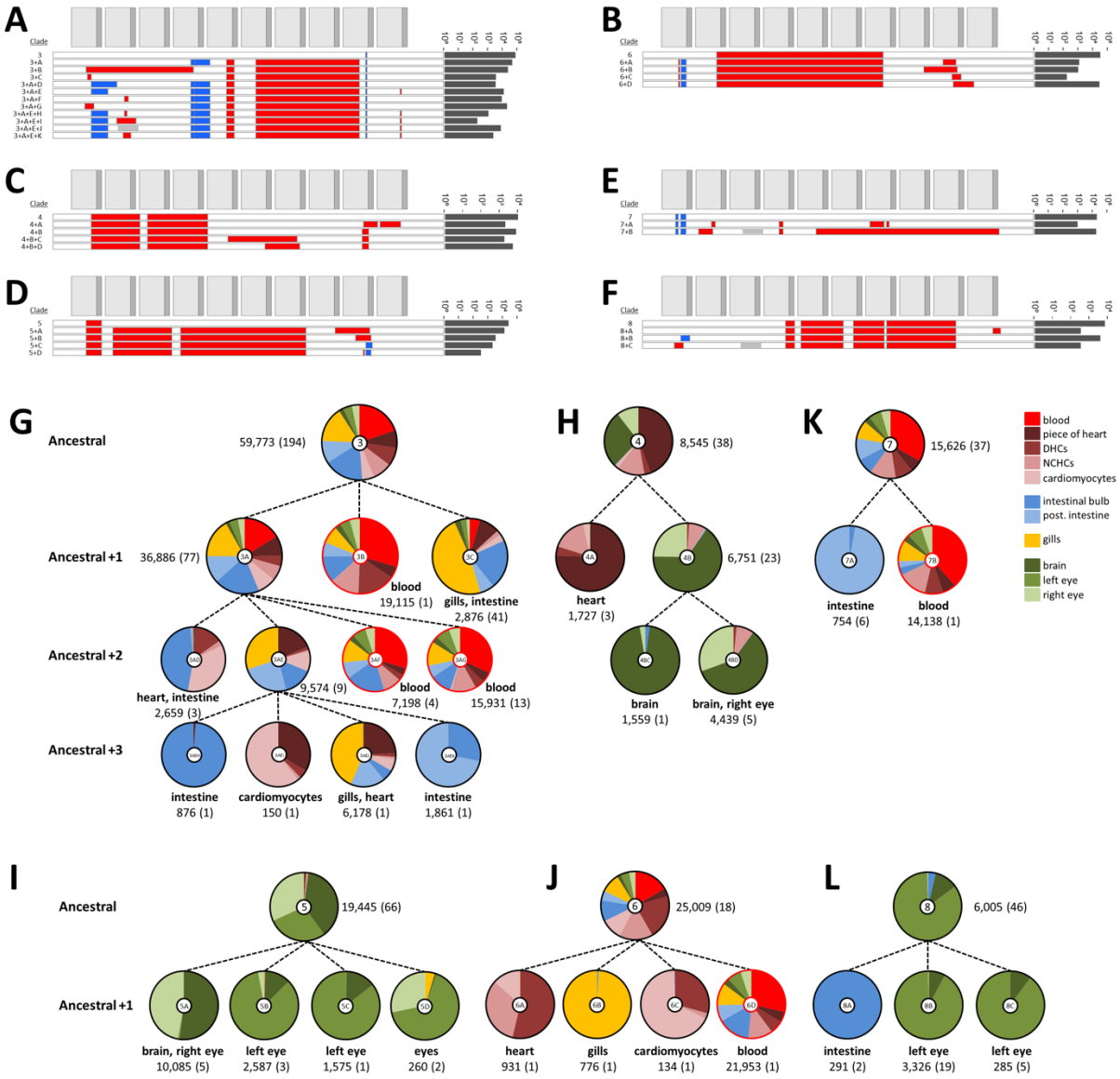



**Figure 5.23. Tracing lineage through editing patterns within additional ADR1 clades.**

(**A**-**E**) Edits that define subclades of clades #3 (**A**), #4 (**B**), #5 (**C**), #6 (**D**), #7 (**E**) and #8 (**F**), with the total number of cells in which these are observed indicated on the right. A grey box indicates an unedited site or sites, distinguishing it from related alleles that contain an edit at this location. **(G-L)** Lineage trees corresponding to subclades of #3 (**G**), #4 (**H**), #5 (**I**), #6 (**J**), #7 (**K**) and #8 (**L**) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization which accounts for differential depth of sampling between organs. Labels in the center of each pie chart correspond to the subclade labels. Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), *i.e.* those comprising >25% of a subclade by proportion.

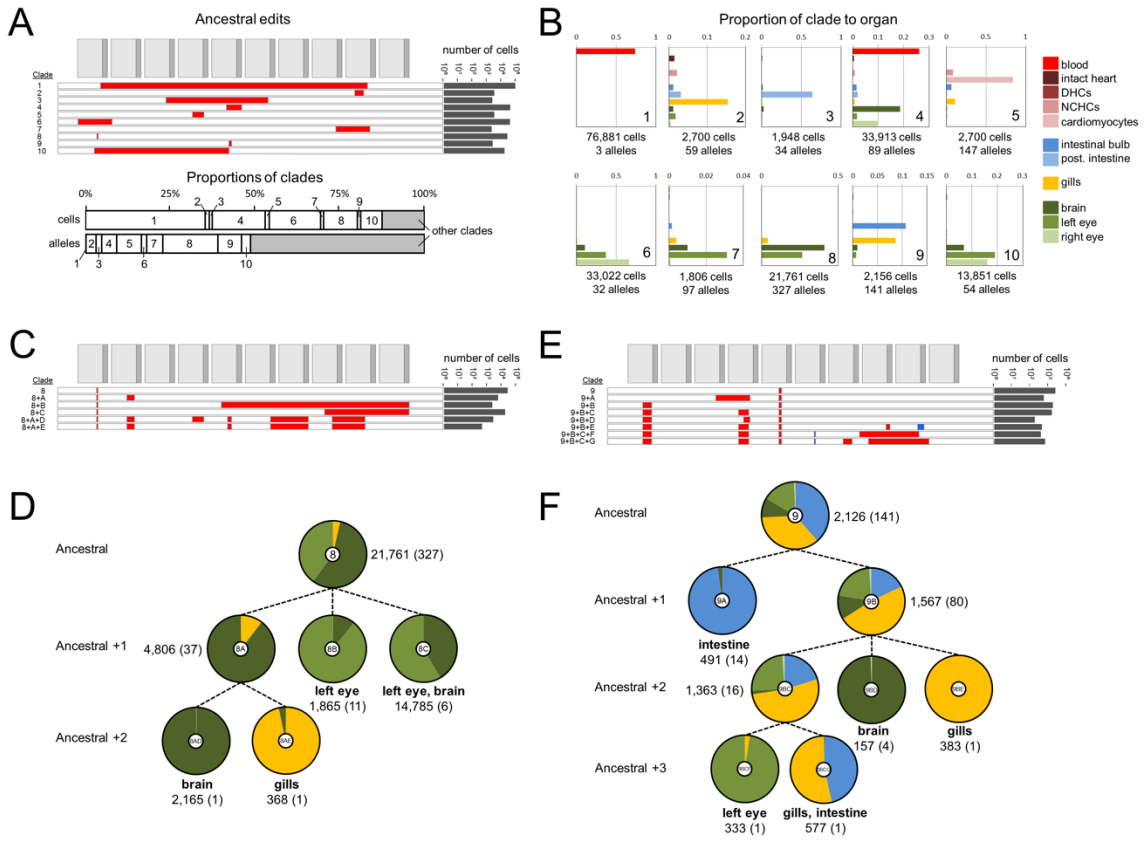



**Figure 5.24. Clades and subclades corresponding to inferred progenitors exhibit increasing levels of organ restriction in ADR2.**

**(A)** Top panel: The inferred ancestral edits that define ten major clades of ADR1, as determined by parsimony, are shown, with the total number of cells in which these are observed indicated on the right. Bottom panel: Contributions of the ten major clades to all cells or all alleles. 126 alleles (out of 2,016 total) that contained ancestral edits from more than one clade were excluded from assignment to any clade, and any further lineage analysis. **(B)** Contributions of each of the ten major clades to each organ, displayed as a proportion of each organ, as described in **Figure 5.21**. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCHCs' = non-cardiomyocyte heart cells. **(C,E)** Edits that define subclades of clade #8 (**C**) and clade #9 (**E**), with the total number of cells in which these are observed indicated on the right. **(D,F)** Lineage trees corresponding to subclades of clade #8 (**D**) and clade #9 (**F**) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization which accounts for differential depth of sampling between organs. Labels in the center of each pie chart correspond to the subclade labels in (**C**/**E**). Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), *i.e.* those comprising >25% of a subclade by proportion.

# Chapter 6. CONCLUSIONS AND FUTURE DIRECTIONS

In this concluding section, I reflect on how CRISPR genome editing can be employed in the future to greatly expand our knowledge of how genomes function, and importantly, how variants impact our susceptibility to disease. Additionally, I highlight future technological capabilities that will be important to develop to harness the full potential of genome editing. Over the next decade, I anticipate the unprecedented capabilities that CRISPR-based tools provide will transform our understanding of how sequence dictates function.

## 6.1 IMPLEMENTING CURRENT TECHNOLOGY TOWARDS COMPREHENSIVE CHARACTERIZATION OF GENOMIC REGIONS

Extension of current CRISPR-based methods to address how variation impacts function in regions already implicated in disease has potential for immediate clinical impact, for instance, resolving variant effects in clinically actionable genes or predicting how somatic variants influence drug susceptibility in cancer (Starita et al., 2017). Saturation genome editing provides a route forward so long as the multiplex assay used accurately reflects *in vivo* variant effects. Developing such assays remains a critical challenge, but a lesson from editing *BRCA1* was that loss-of-function variants in HAP1 cells (a haploid human cell line) are almost always loss-of-function variants in the relevant tissues *in vivo* (as judged by their pathogenicity assessments) even for missense and splice SNVs. Therefore, we predict our assay may be directly applicable for other 'essential genes' whose functional loss predisposes individuals to cancer, including commonly sequenced genes such as *BRCA2*, *PALB2*, *RAD51C*, and *RAD51D* (Walsh et al., 2010).

To understand where critical domains lie within less well-studied proteins, or where mutations may confer drug resistance, Cas9-cytodine deaminase targeting methods (Han et al.,

2017; Hess et al., 2016; Kim et al., 2017; Komor et al., 2016) will be powerful due to their higher scalability. Comprehensive views of gene dependencies gained from CRISPR screening approaches likely will reveal new targets for drug development. Additionally, CRISPR screening data from diverse cell lines analyzed in light of expression and epigenetic data may improve our ability to predict functional dependencies of tumors profiled by DNA and RNA sequencing (Shen et al., 2017; Wang et al., 2017).

To better understand gene regulation, it will be important to first delineate the regions critical to function in an unbiased manner and then to use more precise editing methods to dissect how they work. Scanning the same regions with both single gRNAs and paired gRNAs may prove beneficial for increasing the signal in these experiments (Gasperini et al., 2017). Additional approaches for regulatory element discovery have been developed that involve recruiting either repressive or activating domains to loci using inactive Cas9 (Fulco et al., 2016; Simeonov et al., 2017; Xie et al., 2017). While these perturbations are not genetic in nature, they can still uncover functional regulatory elements. However, follow-up mutagenesis of regions identified with these tools is critically important to confirm observed effects are driven by the underlying DNA sequences at these sites.

Once more regulatory elements are identified, increasing the density of cleavage sites in implicated regions will help pinpoint specific factor binding sites (Canver et al., 2015; Vierstra et al., 2015). Deeply sequencing edits at putative regulatory sites identified from screening is vitally important because it both validates that editing at the specific site is driving the phenotype (and opposed to off-target editing) and enables higher resolution demarcation of functional sequences (Gasperini et al., 2017; Korkmaz et al., 2016). Re-screening smaller libraries of gRNAs showing functional effects in larger screens may enable targeted sequencing of edited sites in multiplex (for

instance, via targeted capture probes or multiplex PCR). Combinatorial screens in which multiple gRNAs targeting specific factor binding sites are queried together will help elucidate redundancy built into regulatory networks.

Methods that use genome editing to introduce programmed allelic series (Findlay et al., 2014; Ryan et al., 2014) will be powerful when applied to non-coding elements of defined function. For instance, we are now performing saturation genome editing in promoters of genes implicated in disease to ask whether the underlying regulatory logic is susceptible to disease-causing variation. Similar dissection of enhancers will reveal the spectrum of mutational effects in these regions. In addition to SNVs, short deletions and insertions can also be synthesized and assayed, as well as synthetic elements with altered arrangements of factor motifs. Well-designed libraries of perturbations will, therefore, not only reveal mutational effects as they pertain to human variation, but may shed light on the underlying rules by which elements function.

## 6.2 Future improvements to multiplex editing technology

Several potential improvements to CRISPR editing technologies would aid efforts to interrogate genome function. One top priority is to increase the percentage of cells in each experiment that receive a desired edit. In an ideal scenario, entry of a single vector into a cell would virtually guarantee a specific edit, such that barcodes on vectors could be used akin to gRNA sequencing in CRISPR screens to inform what edit occurred. Along these lines, Cas9-derived 'base editors' are an intriguing avenue for further development, as it has been shown their specificity and base conversion preferences can be highly engineered (Gaudelli et al., 2017; Kim et al., 2017; Komor et al., 2016). They have the advantage of not inducing double-strand breaks, therefore preventing indels from accruing and thus allowing desired single base changes to accumulate over time. If their efficiency and specificity reach the point where expression of editing components

ensures a specific change is made, then multiplex screening approaches in which base editors are used to engineer hundreds of thousands of mutations across the genome could be implemented and read out simply by sequencing gRNAs (akin to current CRISPR screening strategies).

Methods are also needed that can increase the efficiency of engineering larger edits, such as large deletions, insertions of exogenous sequences, programmed translocations, and inversions. If efficiencies of these types of edits approached 100%, assaying them in multiplex would be readily doable. Dual-selection markers that can be excised after use may be one potential avenue forward (Yusa, 2013). Development of multi-step editing protocols in which cells are pre-engineered to allow more robust editing will likely help, too. Lastly, more knowledge of how the sequence at cleavage sites can influence repair outcome may help to improve efficiencies (van Overbeek et al., 2016).

CRISPR reagents' capability to create diverse mutations calls for better tools for genotyping each cell in edited populations, and linking these genotypes to phenotypic effects. Current methods such as deep sequencing of a targeted locus do not work for ascertaining variation if edits are either large (>1 Kb) or scattered at many sites across the genome. As sequencing technologies improve and allow for longer reads (Feng et al., 2015), more complex editing outcomes over larger swaths of genomic space may be more readily quantified from a mixture of cells. Monoclonal cell line isolation is laborious and can be subject to stochasticity associated with clonal generation, but technical advances that would allow genotyping and assaying hundreds of thousands of clonal populations in a high-throughput manner would make such concerns trivial. Ultimately, coupling highly scalable single-cell omics readouts (Cao et al., 2017; Cusanovich et al., 2015) with methods to precisely ascertain all editing events in a given cell would be a powerful approach for high-throughput and comprehensive analysis of mutational effect.

Perhaps the biggest challenge in implementing these approaches is finding the most relevant biological context. The more portable these methods are across cell lines and model organisms, the better positioned individual labs will be to ask how sequence confers function in areas of biology where they have expertise. Continued engineering of more cell lines and organisms for the purposes of CRISPR-based editing will lead to further gains. While cell lines are naturally amenable to multiplex assays, being able to introduce diverse edits into model organisms has the potential to better capture relevant biology. While *in vivo* CRISPR knockout screening approaches have been used (Chen et al., 2015), efforts to introduce programmed allelic series *in vivo* have been limited (Winters et al., 2017). Therefore, designing delivery systems capable of introducing specific edits to specific tissues and cell types would be highly valuable for performing multiplex experiments *in vivo*. Harnessing temporal control over multiple editing events by using orthologous CRISPR systems (Najm et al., 2018) could reveal effects of mutations acquired sequentially. In short, with continued technological development, labs studying diverse aspects of biology will be equipped to find creative and informative ways to use CRISPR editing in their specific systems.

## 6.3 Final Remarks

In summary, multiplex CRISPR methods for genome editing are already transforming how we study genome function. Moreover, the rapid expansion of the CRISPR suite of tools suggests more powerful methods are still on the way. Geneticists should leverage these tools to produce large functional data sets. Doing so will immediately advance our mechanistic understanding of regions of the genome where the clinical demand for interpreting variation is highest. More fundamentally, however, such data sets can also further our understanding of how complex processes are coded in our DNA. In enabling the pursuit of these goals, CRISPR technologies are

ushering in a new era in genetics — one in which genome editing extends our knowledge far beyond what can be learned through natural variation alone.

# BIBLIOGRAPHY


Ablain, J., Durand, E.M., Yang, S., Zhou, Y., and Zon, L.I. (2015). A CRISPR/Cas9 vector system for tissue-specific gene disruption in zebrafish. Dev. Cell *32*, 756–764.

Adzhubei, I., and Jordan, D.M. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics *76*, 7.20.1–7.20.41.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248.

Aguet, F., Brown, A.A., Castel, S., Davis, J.R., Mohammadi, P., Segre, A.V., Zappala, Z., Abell, N.S., Fresard, L., Gamazon, E.R., et al. (2016). Local genetic effects on gene expression across 44 human tissues.

Aparicio-Prat, E., Arnan, C., Sala, I., Bosch, N., Guigó, R., and Johnson, R. (2015). DECKO: Single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. BMC Genomics *16*, 846.

Babaei, F., Ramalingam, R., Tavendale, A., Liang, Y., Yan, L.S.K., Ajuh, P., Cheng, S.H., and Lam, Y.W. (2013). Novel blood collection method allows plasma proteome analysis from single zebrafish. J. Proteome Res. *12*, 1580–1590.

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

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422–425.

Beumer, K.J., Trautman, J.K., Bozas, A., Liu, J.-L., Rutter, J., Gall, J.G., and Carroll, D. (2008). Efficient gene targeting in Drosophila by direct embryo injection with zinc-finger nucleases. Proc. Natl. Acad. Sci. U. S. A. *105*, 19821–19826.

Blanpain, C., and Simons, B.D. (2013). Unravelling stem cell dynamics by lineage tracing. Nat. Rev. Mol. Cell Biol. *14*, 489–502.

Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Botstein, D., and Shortle, D. (1985). Strategies and applications of in vitro mutagenesis. Science *229*, 1193–1201.

Bouwman, P., van der Gulden, H., van der Heijden, I., Drost, R., Klijn, C.N., Prasetyanti, P., Pieterse, M., Wientjens, E., Seibler, J., Hogervorst, F.B.L., et al. (2013). A high-throughput functional complementation assay for classification of BRCA1 missense variants. Cancer Discov. *3*, 1142–1155.

Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. *22*, 1790–1797.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS Comput. Biol. *8*, e1002822.

Byrne, S.M., Ortiz, L., Mali, P., Aach, J., and Church, G.M. (2015). Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells. Nucleic Acids Res. *43*, e21.

Canver, M.C., Bauer, D.E., Dass, A., Yien, Y.Y., Chung, J., Masuda, T., Maeda, T., Paw, B.H., and Orkin, S.H. (2014). Characterization of Genomic Deletion Efficiency Mediated by Clustered Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells. J. Biol. Chem. *289*, 21312–21324.

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., et al. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192–197.

Canver, M.C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E.N., Needleman, A.J., Galactéros, F., Brugnara, C., Kutlar, A., et al. (2017). Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. Nat. Genet. *49*, 625–634.

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

Carette, J.E., Guimaraes, C.P., Varadarajan, M., Park, A.S., Wuethrich, I., Godarova, A., Kotecki, M., Cochran, B.H., Spooner, E., Ploegh, H.L., et al. (2009). Haploid genetic screens in human cells identify host factors used by pathogens. Science *326*, 1231–1235.

Carette, J.E., Raaben, M., Wong, A.C., Herbert, A.S., Obernosterer, G., Mulherkar, N., Kuehne, A.I., Kranzusch, P.J., Griffin, A.M., Ruthel, G., et al. (2011). Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. Nature *477*, 340–343.

Carlson, C.A., Kas, A., Kirkwood, R., Hays, L.E., Preston, B.D., Salipante, S.J., and Horwitz, M.S. (2011). Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nat. Methods *9*, 78–80.

Carroll, D. (2014). Genome engineering with targetable nucleases. Annu. Rev. Biochem. *83*, 409–439.

Chan, S.L., and Mok, T. (2010). PARP inhibition in BRCA-mutated breast and ovarian cancers. Lancet *376*, 211–213.

Chen, F., Pruett-Miller, S.M., Huang, Y., Gjoka, M., Duda, K., Taunton, J., Collingwood, T.N., Frodin, M., and Davis, G.D. (2011). High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. Nat. Methods *8*, 753–755.

Chen, S., Sanjana, N.E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D.A., Song, J., Pan, J.Q., Weissleder, R., et al. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell *160*, 1246–1260.

Choi, P.S., and Meyerson, M. (2014). Targeted genomic rearrangements using CRISPR/Cas technology. Nat. Commun. *5*, 3728.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., Smith, J.D., Harrell, T.M., McMillin, M.J., Wiszniewski, W., Gambin, T., et al. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am. J. Hum. Genet. *97*, 199–215.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A., and Noushmehr, H. (2012). FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucleic Acids Res. *40*, e139.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cook-Deegan, R., Conley, J.M., Evans, J.P., and Vorhaus, D. (2013). The next controversy in genetic testing: clinical data as trade secrets? Eur. J. Hum. Genet. *21*, 585–588.

Cooper, G.M. (2015). Parlez-vous VUS? Genome Res. *25*, 1423–1426.

Cunningham, B.C., and Wells, J.A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science *244*, 1081–1085.

Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single cell profiling of chromatin

accessibility by combinatorial cellular indexing. Science *348*, 910–914.

Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. *37*, e67.

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

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K.C., Huang, H., Liu, T., Marina, R.J., et al. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat. Methods *14*, 629–635.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell *167*, 1853–1866.e17.

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

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184.

Domchek, S.M., Tang, J., Stopfer, J., Lilli, D.R., Hamel, N., Tischkowitz, M., Monteiro, A.N.A., Messick, T.E., Powers, J., Yonker, A., et al. (2013). Biallelic Deleterious BRCA1 Mutations in a Woman with Early-Onset Ovarian Cancer. Cancer Discov. *3*, 399–405.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics *9*, 11.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096.

Doyon, Y., Choi, V.M., Xia, D.F., Vo, T.D., Gregory, P.D., and Holmes, M.C. (2010). Transient cold shock enhances zinc-finger nuclease–mediated gene disruption. Nat. Methods *7*, 459.

Drost, R., Bouwman, P., Rottenberg, S., Boon, U., Schut, E., Klarenbeek, S., Klijn, C., van der Heijden, I., van der Gulden, H., Wientjens, E., et al. (2011). BRCA1 RING function is essential for tumor suppression but dispensable for therapy resistance. Cancer Cell *20*, 797–809.

Dymecki, S.M., and Tomasiewicz, H. (1998). Using Flp-Recombinase to Characterize

Expansion ofWnt1-Expressing Neural Progenitors in the Mouse. Dev. Biol. *201*, 57–65.

Easton, D.F., Deffenbaugh, A.M., Pruss, D., Frye, C., Wenstrup, R.J., Allen-Brady, K., Tavtigian, S.V., Monteiro, A.N.A., Iversen, E.S., Couch, F.J., et al. (2007). A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer–Predisposition Genes. Am. J. Hum. Genet. *81*, 873–883.

Easton, D.F., Pharoah, P.D.P., Antoniou, A.C., Tischkowitz, M., Tavtigian, S.V., Nathanson, K.L., Devilee, P., Meindl, A., Couch, F.J., Southey, M., et al. (2015). Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. N. Engl. J. Med. *372*, 2243–2257.

Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J.A., and Jasin, M. (1998). Gene conversion tracts from double-strand break repair in mammalian cells. Mol. Cell. Biol. *18*, 93–101.

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

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Essletzbichler, P., Konopka, T., Santoro, F., Chen, D., Gapp, B.V., Kralovics, R., Brummelkamp, T.R., Nijman, S.M.B., and Bürckstümmer, T. (2014a). Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. Genome Res. *24*, 2059–2065.

Essletzbichler, P., Konopka, T., Santoro, F., Chen, D., Gapp, B.V., Kralovics, R., Brummelkamp, T.R., Nijman, S.M.B., and Bürckstümmer, T. (2014b). Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. Genome Res. *24*, 2059–2065.

Evers, B., Jastrzebski, K., Heijmans, J.P.M., Grernrum, W., Beijersbergen, R.L., and Bernards, R. (2016). CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nat. Biotechnol. *34*, 631–633.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917–921.

Farzadfard, F., and Lu, T.K. (2014). Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. Science *346*, 1256272.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164–166.

Feng, Y., Zhang, Y., Ying, C., Wang, D., and Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics *13*, 4–16.

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

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741–746.

Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. *43*, 513–518.

Friedman, L.S., Ostermeyer, E.A., Szabo, C.I., Dowd, P., Lynch, E.D., Rowell, S.E., and King, M.C. (1994). Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. Nat. Genet. *8*, 399–404.

Fu, R., Ceballos-Picot, I., Torres, R.J., Larovere, L.E., Yamada, Y., Nguyen, K.V., Hegde, M., Visser, J.E., Schretlen, D.J., Nyhan, W.L., et al. (2014a). Genotype–phenotype correlations in neurogenetics: Lesch-Nyhan disease as a model disorder. Brain *137*, 1282–1303.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014b). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat. Biotechnol. *32*, 279.

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

Gagnon, J.A., Valen, E., Thyme, S.B., Huang, P., Akhmetova, L., Ahkmetova, L., Pauli, A., Montague, T.G., Zimmerman, S., Richter, C., et al. (2014). Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. PLoS One *9*, e98186.

Gaj, T., Gersbach, C.A., and Barbas, C.F., 3rd (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. *31*, 397–405.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nat. Protoc. *11*, 1782–1787.

Gasperini, M., Findlay, G.M., McKenna, A., Milbank, J.H., Lee, C., Zhang, M.D., Cusanovich, D.A., and Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. Am. J. Hum. Genet. *101*, 192–205.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464.

Ghosh, R., Oak, N., and Plon, S.E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. *18*, 225.

Gibson, T.J., Seiler, M., and Veitia, R.A. (2013). The transience of transient overexpression. Nat. Methods *10*, 715.

Goina, E., Skoko, N., and Pagani, F. (2008). Binding of DAZAP1 and hnRNPA1/A2 to an exonic splicing silencer in a natural BRCA1 exon 18 mutant. Mol. Cell. Biol. *28*, 3850–3860.

Goldgar, D.E., Easton, D.F., Deffenbaugh, A.M., Monteiro, A.N.A., Tavtigian, S.V., and Couch, F.J. (2004). Integrated Evaluation of DNA Sequence Variants of Unknown Clinical Significance: Application to BRCA1 and BRCA2. Am. J. Hum. Genet. *75*, 535–544.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Gupta, T., and Mullins, M.C. (2010). Dissection of organs from the adult zebrafish. J. Vis. Exp.

Gupta, V., and Poss, K.D. (2012). Clonally dominant cardiomyocytes direct heart morphogenesis. Nature *484*, 479–484.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684–1689.

Han, K., Jeng, E.E., Hess, G.T., Morgens, D.W., Li, A., and Bassik, M.C. (2017). Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat. Biotechnol. *35*, 463–474.

Hashimoto, T., Sherwood, R.I., Kang, D.D., Rajagopal, N., Barkal, A.A., Zeng, H., Emons, B.J.M., Srinivasan, S., Jaakkola, T., and Gifford, D.K. (2016). A synergistic DNA logic predicts genome-wide chromatin accessibility. Genome Res. *26*, 1430–1440.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. *89*, 10915–10919.

Henson, P.M., and Hume, D.A. (2006). Apoptotic cell removal in development and tissue homeostasis. Trends Immunol. *27*, 244–250.

Hess, G.T., Frésard, L., Han, K., Lee, C.H., Li, A., Cimprich, K.A., Montgomery, S.B., and Bassik, M.C. (2016). Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat. Methods *13*, 1036–1042.

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

Hollis, R.L., Churchman, M., and Gourley, C. (2017). Distinct implications of different BRCA mutations: efficacy of cytotoxic chemotherapy, PARP inhibition and clinical outcome in ovarian cancer. Onco. Targets. Ther. *10*, 2539–2551.

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., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife *5*.

de la Hoya, M., Soukarieh, O., López-Perolio, I., Vega, A., Walker, L.C., van Ierland, Y., Baralle, D., Santamariña, M., Lattimore, V., Wijnen, J., et al. (2016). Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. Hum. Mol. Genet. *25*, 2256–2268.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Inoue, F., Kircher, M., Martin, B., Cooper, G.M., Witten, D.M., McManus, M.T., Ahituv, N., and Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. *27*, 38–52.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860.

Jacobs, L., and DeMars, R. (1984). Chemical mutagenesis with diploid human fibroblasts. In Handbook of Mutagenicity Test Procedures (Second Edition), (Elsevier), pp. 321–356.

Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. Cell *167*, 1883–1896.e15.

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

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

Kawakami, K. (2007). Tol2: a versatile gene transfer vector in vertebrates. Genome Biol. *8 Suppl 1*, S7.

Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. Nat. Methods *10*, 857–860.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360–1374.

Keller, R.E. (1975). Vital dye mapping of the gastrula and neurula of Xenopus laevis. I.

Prospective areas and morphogenetic movements of the superficial layer. Dev. Biol. *42*, 222–241.

Khalid, M.F., Damha, M.J., Shuman, S., and Schwer, B. (2005). Structure–function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. Nucleic Acids Res. *33*, 6349–6360.

Kim, Y.B., Komor, A.C., Levy, J.M., Packer, M.S., Zhao, K.T., and Liu, D.R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat. Biotechnol. *35*, 371–376.

Kimmel, C.B., and Law, R.D. (1985). Cell lineage of zebrafish blastomeres. III. Clonal analyses of the blastula and gastrula stages. Dev. Biol. *108*, 94–101.

Kinney, J.B., Murugan, A., Callan, C.G., Jr, and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. U. S. A. *107*, 9158–9163.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol. *35*, 561.

Klein, A.M., and Simons, B.D. (2011). Universal patterns of stem cell fate in cycling adult tissues. Development *138*, 3103–3111.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V.V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., et al. (2015a). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V.V., Zheng, Z., and Joung, J.K. (2015b). Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. *33*, 1293–1298.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016a). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.

Kleinstiver, B.P., Tsai, S.Q., Prew, M.S., Nguyen, N.T., Welch, M.M., Lopez, J.M., McCaw, Z.R., Aryee, M.J., and Joung, J.K. (2016b). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. Nat. Biotechnol. *34*, 869–874.

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

Korkmaz, G., Lopes, R., Ugalde, A.P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R., and Agami, R. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nat. Biotechnol. *34*, 192–198.

Kretzschmar, K., and Watt, F.M. (2012). Lineage tracing. Cell *148*, 33–45.

Kuchenbaecker, K.B., Hopper, J.L., Barnes, D.R., Phillips, K.-A., Mooij, T.M., Roos-Blom, M.-J., Jervis, S., van Leeuwen, F.E., Milne, R.L., Andrieu, N., et al. (2017). Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. JAMA *317*, 2402–2416.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868.

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

Le Douarin, N.M., and Teillet, M.-A.M. (1974). Experimental analysis of the migration and differentiation of neuroblasts of the autonomic nervous system and of neurectodermal mesenchymal derivatives, using a biological cell marking technique. Dev. Biol. *41*, 162–184.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., et al. (2014). Highly multiplexed subcellular RNA sequencing in situ. Science *343*, 1360–1363.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lesch, M., and Nyhan, W.L. (1964). A FAMILIAL DISORDER OF URIC ACID METABOLISM AND CENTRAL NERVOUS SYSTEM FUNCTION. Am. J. Med. *36*, 561–570.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, J., Shou, J., Guo, Y., Tang, Y., Wu, Y., Jia, Z., Zhai, Y., Chen, Z., Xu, Q., and Wu, Q. (2015). Efficient inversions and duplications of mammalian regulatory DNA elements and gene clusters by CRISPR/Cas9. J. Mol. Cell Biol. *7*, 284–298.

Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C., and Wang, J. (2013). GWAS3D: Detecting human

regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic Acids Res. *41*, W150–W158.

Liang, F., Han, M., Romanienko, P.J., and Jasin, M. (1998). Homology-directed repair is a major double-strand break repair pathway in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *95*, 5172–5177.

Liang, X., Potter, J., Kumar, S., Ravinder, N., and Chesnut, J.D. (2017). Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. J. Biotechnol. *241*, 136–146.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife Sciences *3*, e04766.

Liu, Z., and Keller, P.J. (2016). Emerging Imaging and Genomic Tools for Developmental Systems Biology. Dev. Cell *36*, 597–610.

Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56–62.

Lodato, M.A., Woodworth, M.B., Lee, S., Evrony, G.D., Mehta, B.K., Karger, A., Lee, S., Chittenden, T.W., D'Gama, A.M., Cai, X., et al. (2015). Somatic mutation in single human neurons tracks developmental and transcriptional history. Science *350*, 94–98.

Lovelock, P.K., Spurdle, A.B., Mok, M.T.S., Farrugia, D.J., Lakhani, S.R., Healey, S., Arnold, S., Buchanan, D., kConFab Investigators, Couch, F.J., et al. (2007). Identification of BRCA1 missense substitutions that confer partial functional activity: potential moderate risk variants? Breast Cancer Res. *9*, R82.

Lu, R., Neff, N.F., Quake, S.R., and Weissman, I.L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol. *29*, 928–933.

Ma, Y., Zhang, J., Yin, W., Zhang, Z., Song, Y., and Chang, X. (2016). Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat. Methods *13*, 1029–1035.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Maruyama, T., Dougan, S.K., Truttmann, M.C., Bilate, A.M., Ingram, J.R., and Ploegh, H.L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat. Biotechnol. *33*, 538.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Mazoyer, S., Puget, N., Perrin-Vidoz, L., Lynch, H.T., Serova-Sinilnikova, O.M., and Lenoir, G.M. (1998). A BRCA1 nonsense mutation causes exon skipping. Am. J. Hum. Genet. *62*, 713–715.

McKenna, A., and Shendure, J. (2017). FlashFry: a fast and flexible tool for large-scale CRISPR target design.

McKenna, A., Findlay, G.M., Gagnon, J.A., Horwitz, M.S., Schier, A.F., and Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science *353*, aaf7907.

Meeker, N.D., Hutchinson, S.A., Ho, L., and Trede, N.S. (2007). Method for isolation of PCR-ready genomic DNA from zebrafish tissues. Biotechniques *43*, 610, 612, 614.

Megason, S.G., and Fraser, S.E. (2007). Imaging in systems biology. Cell *130*, 784–795.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.

Millot, G.A., Carvalho, M.A., Caputo, S.M., Vreeswijk, M.P.G., Brown, M.A., Webb, M., Rouleau, E., Neuhausen, S.L., Hansen, T. v. O., Galli, A., et al. (2012). A guide for functional analysis of BRCA1 variants of uncertain significance. Hum. Mutat. *33*, 1526–1537.

Miner, B.E., Stöger, R.J., Burden, A.F., Laird, C.D., and Hansen, R.S. (2004). Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. *32*, e135.

Monnat, R.J., Jr, Hackmann, A.F., and Chiaverotti, T.A. (1992). Nucleotide sequence analysis of human hypoxanthine phosphoribosyltransferase (HPRT) gene deletions. Genomics *13*, 777–787.

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

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., and Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol. *15*, R19.

Moynahan, M.E., Chiu, J.W., Koller, B.H., and Jasin, M. (1999). Brca1 controls homology-directed DNA repair. Mol. Cell *4*, 511–518.

Myers, R.M., Tilly, K., and Maniatis, T. (1986). Fine structure genetic analysis of a beta-globin promoter. Science.

Najm, F.J., Strand, C., Donovan, K.F., Hegde, M., Sanson, K.R., Vaimberg, E.W., Sullender, M.E., Hartenian, E., Kalani, Z., Fusi, N., et al. (2018). Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. Nat. Biotechnol. *36*, 179–189.

Nica, A.C., and Dermitzakis, E.T. (2013). Expression quantitative trait loci: present and future. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20120362.

Olopade, O.I., and Artioli, G. (2004). Efficacy of risk-reducing salpingo-oophorectomy in women with BRCA-1 and BRCA-2 mutations. Breast J. *10 Suppl 1*, S5–S9.

van Overbeek, M., Capurso, D., Carter, M.M., Thompson, M.S., Frias, E., Russ, C., Reece-Hoyes, J.S., Nye, C., Gradia, S., Vidal, B., et al. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol. Cell *63*, 633–646.

Pan, Y.A., Freundlich, T., Weissman, T.A., Schoppik, D., Wang, X.C., Zimmerman, S., Ciruna, B., Sanes, J.R., Lichtman, J.W., and Schier, A.F. (2013). Zebrabow: multispectral cell labeling for cell tracing and lineage analysis in zebrafish. Development *140*, 2835–2846.

Paquet, D., Kwart, D., Chen, A., Sproul, A., Jacob, S., Teo, S., Olsen, K.M., Gregg, A., Noggle, S., and Tessier-Lavigne, M. (2016). Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. Nature *533*, 125.

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

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Pellettieri, J., and Sánchez Alvarado, A. (2007). Cell turnover and adult tissue homeostasis: from humans to planarians. Annu. Rev. Genet. *41*, 83–105.

Pierce, A.J., Hu, P., Han, M., Ellis, N., and Jasin, M. (2001). Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. Genes Dev. *15*, 3237–3242.

Platt, R.J., Chen, S., Zhou, Y., Yim, M.J., Swiech, L., Kempton, H.R., Dahlman, J.E., Parnas, O., Eisenhaure, T.M., Jovanovic, M., et al. (2014). CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling. Cell *159*, 440–455.

Plon, S.E., Eccles, D.M., Easton, D., Foulkes, W.D., Genuardi, M., Greenblatt, M.S.,

Hogervorst, F.B.L., Hoogerbrugge, N., Spurdle, A.B., Tavtigian, S.V., et al. (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum. Mutat. *29*, 1282–1291.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Porter, S.N., Baker, L.C., Mittelman, D., and Porteus, M.H. (2014). Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. Genome Biol. *15*, R75.

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

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., Syed, T., Emons, B.J.M., Gifford, D.K., and Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. Nat. Biotechnol. *34*, 167–174.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186.

Ransburgh, D.J.R., Chiba, N., Ishioka, C., Toland, A.E., and Parvin, J.D. (2010). Identification of breast tumor mutations in BRCA1 that abolish its function in homologous DNA recombination. Cancer Res. *70*, 988–995.

Rebbeck, T.R., Friebel, T., Lynch, H.T., Neuhausen, S.L., van ’t Veer, L., Garber, J.E., Evans, G.R., Narod, S.A., Isaacs, C., Matloff, E., et al. (2004). Bilateral prophylactic mastectomy reduces breast cancer risk in BRCA1 and BRCA2 mutation carriers: the PROSE Study Group. J. Clin. Oncol. *22*, 1055–1062.

Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen — The Clinical Genome Resource. N. Engl. J. Med. *372*, 2235–2242.

Reid, L.H., Gregg, R.G., Smithies, O., and Koller, B.H. (1990). Regulatory elements in the introns of the human HPRT gene are necessary for its expression in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *87*, 4299–4303.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nat. Biotechnol. *30*, 460–465.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

Rincón-Limas, D.E., Krueger, D.A., and Patel, P.I. (1991). Functional characterization of the human hypoxanthine phosphoribosyltransferase gene promoter: evidence for a negative regulatory element. Mol. Cell. Biol. *11*, 4157–4164.

Ryan, O.W., Skerker, J.M., Maurer, M.J., Li, X., Tsai, J.C., Poddar, S., Lee, M.E., DeLoache, W., Dueber, J.E., Arkin, A.P., et al. (2014). Selection of chromosomal DNA libraries using a multiplex CRISPR system. eLife Sciences *3*, e03703.

Salipante, S.J., and Horwitz, M.S. (2006). Phylogenetic fate mapping. Proc. Natl. Acad. Sci. U. S. A. *103*, 5448–5453.

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A. *112*, E6456–E6465.

Sancak, Y., Peterson, T.R., Shaul, Y.D., Lindquist, R.A., Thoreen, C.C., Bar-Peled, L., and Sabatini, D.M. (2008). The Rag GTPases bind raptor and mediate amino acid signaling to mTORC1. Science *320*, 1496–1501.

Sander, J.D., and Joung, J.K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. *32*, 347–355.

Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784.

Sanjana, N.E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A., and Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. Science *353*, 1545–1549.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Shakya, R., Reid, L.J., Reczek, C.R., Cole, F., Egli, D., Lin, C.-S., deRooij, D.G., Hirsch, S., Ravi, K., Hicks, J.B., et al. (2011). BRCA1 Tumor Suppression Depends on BRCT Phosphoprotein Binding, But Not Its E3 Ligase Activity. Science *334*, 525–528.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelson, T., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

Shen, J.P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A.N., et al. (2017). Combinatorial CRISPR–Cas9 screens for de

novo mapping of genetic interactions. Nat. Methods *14*, 573.

Shendure, J. (2014). Life after genetics. Genome Med. *6*, 86.

Shendure, J., and Akey, J.M. (2015). The origins, determinants, and consequences of human mutations. Science *349*, 1478–1483.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., and Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345.

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., et al. (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature *549*, 111–115.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Smurnyy, Y., Cai, M., Wu, H., McWhinnie, E., Tallarico, J.A., Yang, Y., and Feng, Y. (2014). DNA sequencing and CRISPR-Cas9 gene editing for target validation in mammalian cells. Nat. Chem. Biol. *10*, 623–625.

Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D., et al. (2010). Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell *143*, 134–144.

Song, J., Yang, D., Xu, J., Zhu, T., Chen, Y.E., and Zhang, J. (2016). RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. Nat. Commun. *7*, 10548.

Spurdle, A.B., Healey, S., Devereau, A., Hogervorst, F.B.L., Monteiro, A.N.A., Nathanson, K.L., Radice, P., Stoppa-Lyonnet, D., Tavtigian, S., Wappenschmidt, B., et al. (2012a). ENIGMA–evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum. Mutat. *33*, 2–7.

Spurdle, A.B., Whiley, P.J., Thompson, B., Feng, B., Healey, S., Brown, M.A., Pettigrew, C., kConFab, Van Asperen, C.J., Ausems, M.G.E.M., et al. (2012b). BRCA1 R1699Q variant displaying ambiguous functional abrogation confers intermediate breast and ovarian cancer risk. J. Med. Genet. *49*, 525–532.

Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M., Parvin, J.D., Shendure, J., and Fields, S. (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics *200*, 413–422.

Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am. J. Hum. Genet. *101*, 315–325.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T. vO (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. Eur. J. Hum. Genet. *22*, 1362–1368.

Stent, G.S. (2002). Developmental cell lineage. Int. J. Dev. Biol. *42*, 237–241.

Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.

Suzuki, K., Tsunekawa, Y., Hernandez-Benitez, R., Wu, J., Zhu, J., Kim, E.J., Hatanaka, F., Yamamoto, M., Araoka, T., Li, Z., et al. (2016). In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. Nature *540*, 144.

Tavtigian, S.V., Deffenbaugh, A.M., Yin, L., Judkins, T., Scholl, T., Samollow, P.B., de Silva, D., Zharkikh, A., and Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. *43*, 295–305.

Tavtigian, S.V., Byrnes, G.B., Goldgar, D.E., and Thomas, A. (2008). Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum. Mutat. *29*, 1342–1354.

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

The GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Thisse, C., and Zon, L.I. (2002). Organogenesis–heart and blood formation from the zebrafish point of view. Science *295*, 457–462.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat. Biotechnol. *33*, 187–197.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564–576.e16.

Vega, A., Campos, B., Bressac-De-Paillerets, B., Bond, P.M., Janin, N., Douglas, F.S., Domènech, M., Baena, M., Pericay, C., Alonso, C., et al. (2001). The R71G BRCA1 is a founder Spanish mutation and leads to aberrant splicing of the transcript. Hum. Mutat. *17*, 520–521.

Vierstra, J., Reik, A., Chang, K.-H., Stehling-Sun, S., Zhou, Y., Hinkley, S.J., Paschon, D.E., Zhang, L., Psatha, N., Bendana, Y.R., et al. (2015). Functional footprinting of regulatory DNA. Nat. Methods *12*, 927–930.

Wakabayashi, A., Ulirsch, J.C., Ludwig, L.S., Fiorini, C., Yasuda, M., Choudhuri, A., McDonel, P., Zon, L.I., and Sankaran, V.G. (2016). Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. Proc. Natl. Acad. Sci. U. S. A. *113*, 4434–4439.

Walsh, C., and Cepko, C.L. (1992). Widespread dispersion of neuronal clones across functional regions of the cerebral cortex. Science *255*, 434–440.

Walsh, T., Lee, M.K., Casadei, S., Thornton, A.M., Stray, S.M., Pennil, C., Nord, A.S., Mandell, J.B., Swisher, E.M., and King, M.-C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc. Natl. Acad. Sci. U. S. A. *107*, 12629–12633.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell *153*, 910–918.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell *168*, 890–903.e15.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. *40*, D930–D934.

Weedon, M.N., Cebola, I., Patch, A.-M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H.-H., Allen, H.L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat. Genet. *46*, 61–64.

Weisblat, D.A., Sawyer, R.T., and Stent, G.S. (1978). Cell lineage analysis by intracellular injection of a tracer enzyme. Science *202*, 1295–1298.

Winters, I.P., Chiou, S.-H., Paulk, N.K., McFarland, C.D., Lalgudi, P.V., Ma, R.K., Lisowski, L., Connolly, A.J., Petrov, D.A., Kay, M.A., et al. (2017). Multiplexed in vivo homology-directed repair and tumor barcoding enables parallel quantification of Kras variant oncogenicity. Nat. Commun. *8*, 2053.

Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

Woods, N.T., Baskin, R., Golubeva, V., Jhuraney, A., De-Gregoriis, G., Vaclova, T., Goldgar, D.E., Couch, F.J., Carvalho, M.A., Iversen, E.S., et al. (2016). Functional assays provide a robust tool for the clinical annotation of genetic variants of uncertain significance. Npj Genomic Medicine *1*, 16001.

Xie, K., Minkenberg, B., and Yang, Y. (2015). Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. Proc. Natl. Acad. Sci. U. S. A. *112*, 3570–3575.

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

Yang, S., Cline, M., Zhang, C., Paten, B., and Lincoln, S.E. (2017). DATA SHARING AND REPRODUCIBLE CLINICAL GENETIC TESTING: SUCCESSES AND CHALLENGES. Pac. Symp. Biocomput. *22*, 166–176.

Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. *369*, 1502–1511.

Yin, L., Maddison, L.A., Li, M., Kara, N., LaFave, M.C., Varshney, G.K., Burgess, S.M., Patton, J.G., and Chen, W. (2015). Multiplex Conditional Mutagenesis Using Transgenic Expression of Cas9 and sgRNAs. Genetics *200*, 431–441.

Yusa, K. (2013). Seamless genome editing in human pluripotent stem cells using custom endonuclease-based gene targeting and the piggyBac transposon. Nat. Protoc. *8*, 2061–2078.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2014). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.

Zhang, J., Kuo, C.C.J., and Chen, L. (2011). GC content around splice sites affects splicing through pre-mRNA secondary structures. BMC Genomics *12*, 90.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

Zhang, J.-P., Li, X.-L., Li, G.-H., Chen, W., Arakaki, C., Botimer, G.D., Baylink, D., Zhang, L., Wen, W., Fu, Y.-W., et al. (2017). Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. Genome Biol. *18*, 35.

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

Zhu, S., Li, W., Liu, J., Chen, C.-H., Liao, Q., Xu, P., Xu, H., Xiao, T., Cao, Z., Peng, J., et al. (2016). Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat. Biotechnol. *34*, 1279–1286.

Zinyk, D.L., Mercer, E.H., Harris, E., Anderson, D.J., and Joyner, A.L. (1998). Fate mapping of the mouse midbrain–hindbrain constriction using a site-specific recombination system. Curr. Biol. *8*, 665–672.

# VITA

Gregory M. Findlay (born 1986, Berkeley, CA) grew up in Seattle, WA and attended Garfield High School. Greg worked in the labs of Nina Salama and Barry Stoddard at the Fred Hutchinson Cancer Research Center during college. He graduated with a B.A. in Biology from Carleton College, in Northfield, MN in 2009 before moving to Boston to pursue research in the lab of Dr. Anjana Rao, where he used RNA interference methods to study calcium signaling in the context of T cell activation. Before leaving Boston, he used TALEN-mediated genome engineering to fluorescently tag endogenously expressed proteins in the lab of Tomas Kirchhausen. He entered the University of Washington's Medical Scientist Training Program in 2012 where he has remained since. He joined Jay Shendure's lab in the winter of 2013 while still a 'full-time' student in medical school. That summer he began to work on the research presented in this dissertation. He has officially been a graduate student in the Genome Sciences department since the autumn of 2014. Upon completion of his PhD, Greg plans to return to medical school.