Building multiplexed genomic tools for editing and interpreting human variation

Nicholas Anthony Popp

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Douglas Fowler, Chair

Jill Johnsen

Lea Starita

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Building multiplexed genomic tools for editing and interpreting human variation

Nicholas Anthony Popp

Chair of the Supervisory Committee:

Douglas Fowler

Department of Genome Sciences

Advances in DNA sequencing over the past twenty years have led to an unprecedented expansion in our ability to read DNA sequences and find genetic variation in individuals. However, our ability to manipulate and interpret genetic variation has not nearly kept pace with our ability to identify variants. Despite a growing need for new genomic tools in clinical medicine, very few have been sufficiently developed and validated to be adopted in clinical settings. In particular, two tools have shown promise for clinical use– Clustered regularly interspaced short palindromic repeat (CRISPR)-Cas9 based gene editing and multiplexed assays of variant effect (MAVEs)–but remain limited in scope. In Chapter 1, I introduce each of these two methods, review how they work, how they are used in clinical settings, and discuss their current limitations. In Chapter 2, I describe a method I developed to expand the types of proteins that can be

studied using MAVEs to include secreted and extracellular proteins via library-compatible mammalian surface display. I apply this method to coagulation factor IX (FIX), the primary genetic cause of hemophilia B, to quantify variant effects on FIX secretion, identify structural constraints for folding and secretion, and apply these data to FIX variation to determine mechanisms of disease. I then expand my study of FIX beyond secretion to post-translational modifications (PTMs) of FIX, including γ-carboxylation of its GLA domain and enzymatic cleavage by coagulation factor XI_a. I find evidence for strong structural constraint in the FIX GLA domain that incompletely correlates to its level of γ-carboxylation and instead identifies distinct mechanisms for variant effects in the propeptide and GLA domains. In Chapter 3, I switch from functional characterization of genetic variants to controlled gene editing and repair using CRISPR-Cas9. I describe a new method to increase CRISPR-Cas9 based gene editing precision without compromising efficiency. By coaddition of catalytically inactive Cas9 guide RNAs (dRNAs) that bind to off-target editing sites, I significantly reduce Cas9 editing without the need for blocking mutations that alter adjacent sequences and improve editing efficiency for targeted variant knock-in. Finally, in Chapter 4, I outline future challenges for each of these technologies, suggest routes for further application with these tools, and comment on the future of clinically useful genomic tools.

Table of contents

List of figures

List of appendix figures

List of tables

List of appendix tables

Acknowledgments

My Ph.D. journey can only be described as chaotic, filled with bursts of creative productivity and excitement, then long stretches of failed experiments. Successes came early in my first project, then stalled for years in publication turmoil. I had opportunities to experience parts of the world I never dreamed of seeing back in rural Illinois that were followed by months of pandemic lockdown. My success was all but guaranteed. Yet, I was lucky enough to have built a diverse support network that would lift me when times were tough and celebrate my successes. Every person in that network, I have to thank profusely, because I would not have finished this journey without them.

There are many people who were instrumental in getting me to and through graduate school. John Bouhasin encouraged me to be curious about my hemophilia, to ask myself *how* and *why*. I am always drawn back to the difficulty of these simple questions when thinking about medicine and science. My earliest science mentors, Danaya Pakotiprapha and Chi-Chao Chan, both gave me the opportunity (and responsibility) to design projects very early on in my science career. They also taught me that a failed experiment is simply an opportunity to learn, a piece of wisdom I use more often that I care to admit. My thesis committee, Dusty Maly, Lea Starita, and Debbie Nickerson, all are tremendous scientists who have gone above and beyond to support me through my Ph.D. Debbie, in particular, had a keen eye for someone's strengths and opened the right doors for them, myself included. I could not have become the rigorous, critical, and thoughtful scientist I am today without my Ph.D. advisors, Doug Fowler and Jill Johnsen. They taught me to always question my

assumptions when looking at data and plan out what the right next experiment is. I attribute so much of my success to their guidance.

Jill and Doug have both built labs full of excellent and supportive scientists throughout the years. In particular, Rachel Powell single-handedly brought joy back into research during the pandemic and quickly became one of my closest friends and D&D bestie. This dissertation would not be anything near what it is now without her working with me. I could also always rely on Cindy Wei, Kenny Matreyek, and Melissa Chiasson to share frustrations, celebrate a win, or to think through a tough result.

Beyond science, thanks to all the people in my life from outside the lab who keep me balanced: Maham Ayaz, Evan Clamors, Matt Fritzler, Joseph Peters-Matthews, Conrad Nied, and Connor Gilroy for every conversation that let me know that I was cared for. A special thanks to my partner, Erick Lacayo, because living with a Ph.D. student is arduous on its best days. He (with our dog, Bud) have held me steady through this journey. I am also so grateful to have such supportive parents, who taught me there are no bounds to what I can do, even with my hemophilia. They have and always will be my biggest cheerleaders. I will forever hear them asking, “Did your experiment work?” in my dreams.

Dedication

This dissertation is dedicated to my grandmother, Nora Becherer. I miss you more than words can express.

1 Introduction

Next-generation DNA sequencing has greatly increased our ability to identify genetic variation in humans over the past twenty years, especially as the costs of sequencing an individual's exome and/or genome have dropped. Estimates from whole-genome, whole-exome, and phylogenetic studies suggests that the *de novo* mutation rate in humans is approximately 1-3x10^{-8} mutations per nucleotide per generation, which translates to between 60-180 unique *de novo* variants per individual.[1] Applying the *de* novo mutation rate to the entire human population suggests that every single nucleotide variant compatible with life already exists somewhere in the population.[2] These *de novo* variants occur in addition to existing rare variation that is already present within the human population. To illustrate the scale of this problem, we can look to the Genome Aggregation Database (gnomAD), wherein whole genome and whole exome sequencing of 141,456 healthy individuals identified 244.8 million high-quality variants.[3] Restricting the identified variants to those within gene exons still results in 7.1 million unique variants across nearly every gene in the human genome.

The medical impact of these 7.1 million unique variants is generally unknown, as the vast majority are difficult to interpret clinically due to their rarity. Under 2% have interpretations in ClinVar. [4–8] The rest of these variants are deemed variants of uncertain significance (VUS), meaning that their association with pathogenicity is unclear from current evidence.[9] Because so much of the promise of precision medicine relies on accurate characterization of variant effects, the utter lack thereof prevents its implementation. In order to realize the full potential of precision medicine and improve

clinical care, clear and accurate variant interpretation is required, which necessitates the development of new multiplexed methods for characterizing genetic variation.

1.1 The challenges in characterizing genetic variation

Even though rare variants represent the majority,[10] characterizing them is difficult, primarily due to the limitations of current tools in the geneticist's toolkit. are the main roadblock for implementing precision medicine. Many, like genome-wide association studies, require repeated observations in multiple individuals to identify potentially pathogenic variants. Pedigrees showing co-segregation of disease and variant can provide clear evidence, but are costly, time-consuming, reliant on voluntary assessment for unaffected individuals in the family, and only capture the variant in a single context. On the other end of the spectrum, computational variant effect predictors are scalable to millions of variants but lack precision for clinical use.[11] Even in the rare case that these approaches lead to a conclusive interpretation, little is learned about the mechanisms underlying the variant's effects. Follow up *in vitro* biochemical and cellular experiments can resolve mechanism and clarify functional effects, but are too slow and expensive to implement for each variant individually [4,12]

1.2 Multiplexed assays of variant effect (MAVEs) represent a path forward for scalable variant interpretation

Over the past decade, as next-generation sequencing costs have dropped drastically, multiplexed assays of variant effect (MAVEs) have been developed to study the effects of thousands of variants of a gene simultaneously. MAVEs have been

applied to a plethora of phenotypes including oncogenic potential, aggregation propensity, pharmacokinetics, membrane trafficking, viral evolution, and antibody escape.[4–6,9,13–27] Furthermore, these assays can be applied broadly beyond gene exons to mRNA untranslated regions (UTRs), promoters, enhancers, and splice sites.[28–33]

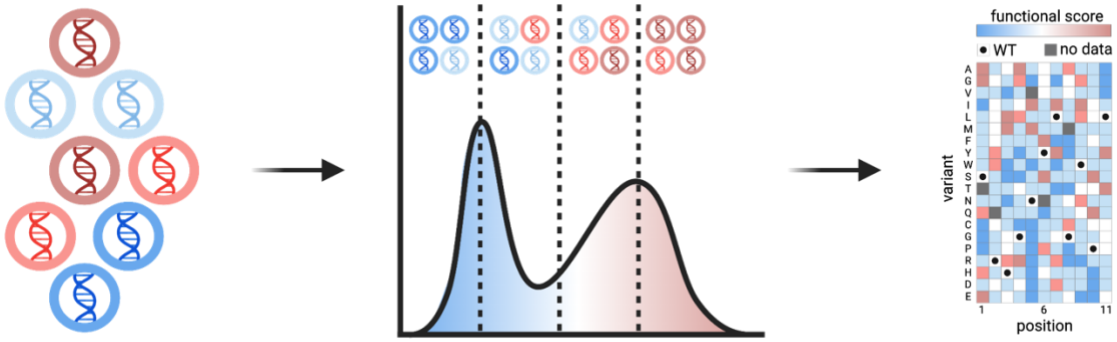

Figure 1.1: Overview of general MAVE structure

A library of genetic elements, Cas9 guide RNAs, or variants in a gene are created and introduced into cells. Only one genetic element is expressed per cell. Cells are then subjected to a selection assay, such as growth or reporter fluorescence (shown) that separates variants based on their function, which is indicated by the dotted lines in the middle panel. High-throughput sequencing of the separated cells is performed to calculate the distribution of each element across the range of the assay to generate a functional score. Functional scores are typically visualized using a heatmap of position and variant amino acid, where color indicates the functional score.

While the methods for functionally characterizing variants in each MAVE vary, a common structure exists (**Figure** 1.1). A DNA variant library containing all possible variants (typically only single amino acid changes) is introduced into a cellular system in such a way that, on average, only one variant is expressed per cell. Example methods to introduce variants include genomic integration into an inducible landing pad via

recombination,[4–8] Cas9-based delivery of homology-directed repair (HDR) templates,[16] CRISPR prime editing,[34] transient plasmid transfection,[18,19,23,27] and lentiviral transduction.[25] Cells expressing variants are then subjected to some phenotypic selection pressure and sorted based on their phenotype. Early MAVEs relied on growth phenotypes, where either functional[19] or nonfunctional[16] variants would lead to cell death and drop out of the assay over time. Later MAVEs were expanded to include fluorescent reagents that could allow variant-expressing cells to be sorted using fluorescence-assisted cell sorting (FACS).[4–6,22,24] In both cases, the selection assay stratifies variants based on their phenotypic effect. Next-generation DNA sequencing of variants or associated short DNA barcodes can then be used to calculate a variant's frequency across sorted conditions. Those frequencies are then used to compute a functional score for the given phenotype. Additionally, the functional score is often normalized to scale the range of known variant effects on the measured phenotype.[4,5,26] The result is a variant effect map that reveals the functional consequences of thousands of variants within the genetic element. In this way, MAVEs increase the scale for testing variants such that the primary limitation is the speed of sorting cells.

1.2.1 MAVEs require a tight genotype-phenotype link

In a typical MAVE for a gene of interest, the gene product, or protein, is the assayed entity. However, sequencing proteins directly and in parallel is difficult to achieve. DNA sequencing, on the other hand, is comparatively affordable, can be multiplexed to millions of individual molecules, and is easily extracted from cells. As such, the MAVE readout does not occur at the protein level but rather at the DNA level,

identifying individual variants directly via sequencing or by association with a short DNA-based barcode.[4,16] To accurately estimate variant effects in this manner requires that the protein assay and DNA sequencing readouts are highly correlated—that is to say, there is a tight genotype-phenotype link. In most cases, this assumption holds true, as the variant DNA and variant protein are both encapsulated within the confines of individual cells. The cells themselves can be imagined as a container of sorts, holding all the relevant information intact until extraction.

However, there are plenty of circumstances where this assumption of a strong genotype-phenotype link no longer holds true. A clear example of an unlinked genotype and phenotype, and the one that this dissertation is primarily concerned with, comes from secreted, extracellular proteins. These proteins are not retained by, nor is there memory of, their cell-of-origin, the functional assay measured at the level of the protein no longer correlates with the cellular DNA readout. At present, there are no methods amenable to studying secreted proteins at the MAVE scale using their endogenous cellular secretion machinery. The issue of not being able to study secreted protein variants at scale is not a small one—secreted proteins represent almost 10% of the human genome, 25% of which have been directly linked to disease in humans.[35,36]

1.2.2 Existing methods to study secreted protein variants at scale

Very few scalable methods exist that are compatible with secreted proteins, and those that are available suffer from significant limitations. For example, library scale methods for displaying proteins on the surface of *E. coli* and *S. cerevisiae* have existed for over 25 years.[37–41] However, these methods are typically unsuitable for studying

human secreted proteins, as they were first developed to allow for biochemical characterization of protein-protein interactions on the cell surface. Since researchers were most concerned with recapitulating intracellular protein-protein dynamics, the methods were developed and optimized to display obligate intracellular proteins. Typically, display was achieved by simultaneously fusing a highly-expressing signal peptide sequence (commonly from *IgK*) to the N-terminus of the protein and a transmembrane domain and cytoplasmic tail to its C-terminus to allow for intracellular reporter signaling.[38] Indeed, this fusion construct is still in use to this day.[25,38,39]

The fusion of the N-terminal leader sequence precludes accurate assessment of endogenously secreted proteins, which each contain their own signal sequence and ordered PTMs that are sequence-context-specific.[42,43] For instance, modification of the positions directly after the signal peptide have been shown to alter translocation efficiency in bacteria.[44,45] More recent work has also provided evidence that signal sequences can directly affect post-cleavage events within the endoplasmic reticulum (ER)-Golgi apparatus or can have roles in human autoimmunity.[42,46] Examples also exist where modifications in the signal peptide sequence alter protein expression sufficiently such that the resulting protein is incorrectly post-translationally modified or shuttled to the wrong cellular compartment.[43,45] Swapping out signal peptides for studying secreted protein variants represents a potentially large loss of valuable genotype-phenotype information that is integral to understanding its biological function.

Moreover, the use of *E. coli* and *S. cerevisiae* present their own issues for displaying secreted proteins. Most human secreted proteins are highly post-translationally modified, with secondary protein cleavage events, *N*- and *O*-linked

glycosylation, chaperone-assisted folding, and disulfide bond formation, among others.[47] *E. coli* and *S. cerevisiae* lack the machinery to perform a majority of these necessary PTMs, many of which are integral to the secreted protein's stability, folding, and affinity for binding partners.[47,48] To more carefully capture the functional effects of variants in secreted human proteins, a new, scalable mammalian system is needed.

1.3 A MAVE for analyzing secreted proteins at scale

Chapter 2 describes a new mammalian cell surface display system that is compatible with both MAVE libraries and secreted genes which retains their endogenous secretion machinery to function. I apply this system to interrogate the effect of variation on secretion of coagulation factor IX (FIX), the primary cause of hemophilia B.[49] I then expand the scope of the display system to show it can profile the effects of variants on FIX γ-carboxylation, a series of 12 required PTMs within the FIX GLA domain that result in the formation of a necessary three helix structure and protruding hydrophobic ω-loop.[50–54]

1.3.1 The hemostasis network, a dynamic and precarious system

For humans, a flexible, rapidly responding hemostasis system is required for health and survival, as it maintains the transport of nutrients, oxygen, and waste materials throughout the body. As such, the hemostatic system is both highly regulated and precarious. In response to vessel injury and bleeding, vasoconstriction rapidly occurs, cutting off blood flow to the damaged sites. During primary hemostasis, circulating platelets, with the assistance of von Willebrand factor (vWF), activate and

aggregate to form a platelet plug at the site of injury.[55,56] The platelet plug, however, is weak enough to be dismantled by shear stress, and requires additional support from the products of secondary hemostasis.[56] In secondary hemostasis, plasma coagulation factors, including FIX, catalyze a series of enzymatic reactions to generate thrombin, which converts fibrinogen into fibrin.[57] Fibrin is then inserted into the platelet plug and cross-linked by coagulation factor XIII (FXIII) to form the a stable clot and allow wound healing to proceed.[57,58]

1.3.2 Hemophilia and other genetic disorders of coagulation

Dysregulation of the hemostatic system can result in both thrombotic and hemorrhagic disorders. While these disorders can be acquired due to malignancy, infections, liver disease, and cardiovascular disease, frequently the causes are genetic in origin. The most common of these genetic disorders are hemophilia A and hemophilia B, for which just over 98% of patients have identifiable variants in the *F8* (coagulation factor VIII, FVIII) and *F9* (coagulation factor IX, FIX) genes.[59]

For the purposes of this dissertation, I will be focusing on FIX, for which genetic variants have been associated with hemorrhagic,[49,59,60] pro-thrombotic,[61] and pharmacogenetic phenotypes.[62,63] The most common FIX-related disorder is hemophilia B, nicknamed the Royal Disease, as it afflicted a large fraction of the European royal families in the 19th and 20th centuries. Hemophilia B affects approximately 3.8 per 100,000 males worldwide, or ~1.2 million people currently living with the disease.[64,65] As the *F9* gene is located on the X chromosome, the prevalence of hemophilia B in females is significantly lower.[66,67]. Moreover, nearly 1 in 5 patients seen at hemophilia

treatment centers (HTCs) with mild hemophilia were women, most of whom are heterozygous for a pathogenic variant in *F9*.[68] Their bleeding phenotypes often arise from skewed X-inactivation, though various deletions, transversions, and chromosomal loss have been observed.[66–68]

Hemophilia B is costly financially, with per patient healthcare costs approaching $500,000+ annually, and physically, with a significantly increased disease burden and upwards of a 35% reduction in life expectancy.[49,65] The burden is not evenly distributed, however, as hemophilia B can be subdivided into severities based on both bleeding symptoms and an *in vitro* clinical assay of FIX activity compared to pooled plasma from unaffected donors (

Table 1.1).[49,64] Disease severity is highly correlated with genotype, which, while rarely used as a diagnostic tool, is regularly used for reproductive planning, neonatal management, and to predict which patients are at high likelihood of developing an anaphylactic neutralizing antibody reaction (inhibitor) to their exogenous factor replacement therapy.[69–72] There is even preclinical evidence from mice with hemophilia that suggests that FIX variants can alter its biodistribution and affect the efficacy and clearance rate of exogenously administered wildtype FIX.[73,74]

Despite the well validated association between variants in *F9* and hemophilia B, relatively little functional work has been performed on the variants themselves to understand how things like disease severity and inhibitor risk are conferred. In fact, 48% of variants remain classified as VUS due to their lack of functional evidence.[59] Linking DNA variants to disease, protein structure, and clinical phenotypes has been used repeatedly to elucidate functional constraints on genes and to develop novel

interventions. One such example comes from FIX itself—the FIX Padua variant (R384L), found in 2009 in a family with a history of spontaneous thrombosis in Italy,[61] is a gain-of-function variant with specific activity that is 8-12 times that of wildtype FIX.[61,75] Just five years later, in 2014, the first phase I/II gene therapy trial using FIX Padua was established,[76] which is now on the market for patients. A systemic look at variant effects in FIX across multiple phenotypes, thus, has the potential to mechanistically explain how variants affect FIX function and to be repurposed to improve FIX therapeutics.

Hemophilia B severity	FIX activity level (%)	Symptoms	Median age at diagnosis
Severe	< 1%	Spontaneous hemarthrosis and soft tissue bleeding, intracranial hemorrhage, and severe and prolonged bleeding after minor trauma	1 month
Moderate	1-5%	Infrequent spontaneous soft tissue bleeding and hemarthrosis, significant bleeding after minor trauma, and prolonged bleeding after surgery	8 months
Mild	5-40%	No rare or spontaneous bleeding but prolonged bleeding after major trauma or surgery	36 months

Table 1.1: Clinical assessment of hemophilia patients

1.3.3 Variants in FIX can cause quantitative or qualitative deficits that lead to hemophilia B

One outstanding unresolved question in hemophilia B genetics is which variants lead to quantitative (amount of FIX secreted) vs. qualitative (functional activity of secreted FIX) deficits in patients and why. Many quantitative variants are so severe that

they are designated cross-reactive material negative (CRM^-), meaning there is no detectable FIX antigen or FIX enzymatic activity in the blood of patients harboring these variants.[77] CRM^- variants are nearly always associated with severe disease and increased risk of developing inhibitors.[70] However, many other quantitative variants show reduced FIX blood antigen but remain CRM^+ and retain enzymatic activity.[78] Many of these variants perform indistinguishably from wildtype FIX *in vitro*, but their reduced secretion results in a bleeding phenotype.[78] That said, clinical measurement of FIX blood antigen is rarely performed, as there is little diagnostic value that cannot be captured using FIX activity assays. As such, there is a dearth of knowledge of which variants affect FIX secretion *en masse*.

Qualitative variants (which are also CRM^+), on the other hand, show excess FIX activity loss relative to their blood antigen levels.[78,79] These variants have been sporadically described throughout the literature, but again, because FIX antigen is rarely performed outside the context of research settings, little is known about which variants lead to qualitative deficits or which of the many possible mechanistic routes lead to their dysfunction. What is known is that these qualitative variants generally affect some phenotype besides secretion.[80,81]

1.3.4 γ-carboxylation of the FIX GLA domain is required for molecular interactions and enzymatic activity

One phenotype affected by variation in FIX is γ-carboxylation of 12 glutamate residues within the GLA domain, a vitamin K-dependent modification that is common to a number of coagulation proteins including FIX, coagulation factor VII (FVII),

coagulation factor X (FX), thrombin (FII), protein C (PC), and protein S (PS).[82] γ-carboxylation of the GLA domain coordinates the binding of divalent cations, which induces a structural transition from an unstructured domain into a tightly folded three helical structure.[52,83–86] The folded GLA domain conformation in FIX creates a solvent-exposed ω-loop, which is involved in binding phospholipid membranes and type IV collagen,[53,73,74] activation of FIX into FIX_a by activated coagulation factor XI (FXI_a),[51,87] the formation of the catalytic tenase ($FVIII_a$-FIX_a-FX) complex,[52–54] and interactions with its cofactor, activated coagulation factor VIII ($FVIII_a$).[88] Interestingly, the ω-loops of γ-carboxylated coagulation proteins are variable, and research suggests that these differences in ω-loop structure confer binding partner specificity. Indeed, replacement of the FIX ω-loop with that from FVII leads to a complete loss of FXI_a-mediated binding and FIX activation, but does not affect $FVII_a$/tissue factor-mediated FIX activation.[51,89–91]

While the role of γ-carboxylation has been deciphered, it is still unclear which residues of FIX's GLA domain are integral to ω-loop formation and can lead to hemophilia if mutations arise. Evidence suggests that γ-carboxylation of wildtype FIX *in vivo* is not always complete. However, hypo-γ-carboxylated FIX, which is found in most FIX replacement therapies, can still function like wildtype FIX.[92] As such, the mechanism of action of hemophilia-causing variants in this region, particularly in the upstream propeptide, is hotly contested.[50,93–97] Additionally, *in vitro* work using synthetic disulfide-linked peptides shows that the functional ω-loop can be recapitulated without either γ-carboxylation or calcium ions so long as the peptide can take on the solvent-exposed hydrophobic orientation.[53]

With these findings coming to light, the question has arisen of whether these propeptide and GLA domain variants ablate γ-carboxylation entirely or whether they simply alter the structure of the ω-loop. For example, variants at positions 37 and 40 (legacy numbering: positions -10 and -6), have been described that reduce γ-carboxylation entirely.[63] Other variants at position 37 do not cause hemophilia, but they can lead to a warfarin-induced hemorrhagic disorder, where there is an outsized loss of FIX function relative to other γ-carboxylated coagulation proteins.[62,63] All of these variants alter the propeptide alpha-helix, which is the binding site of GGCX, the protein that catalyzes γ-carboxylation using vitamin K.[82,98,99] Warfarin and its analogues reduce GGCX function by inhibiting VKOR-based recycling of vitamin K. [82,98–100] A few positions downstream, variants at FIX residues 43, 45, and 46 (-4, -2, and -1 in legacy numbering) in the propeptide prevent its cleavage and removal.[94,101–104] Early experiments determined that these propeptide variants also led to poor γ-carboxylation, which was determined by using a γ-carboxylation-specific anti-FIX antibody.[50,94,101–104] More recent work, however, suggests that γ-carboxylation occurs before propeptide cleavage of FIX.[105,106] Related studies on the γ-carboxylation of variants that retain their propeptide suggests these variants are, in fact, fully γ-carboxylated.[97] Instead, researchers propose that retained propeptide alters or blocks the conformational shift required to form the ω-loop. However, a consensus has not yet been reached and warrants a more thorough investigation.

1.4 Limitations of using MAVEs

While the output derived from MAVEs is ever-increasing, there are still limitations to the promise of functionally characterizing every variant in the genome. The most pressing of these limitations is the need to develop phenotype-specific assays for each gene of interest. Most gene products have multiple molecular functions, making the issue even more pressing. Generalized assays, like Variant Abundance by Massively Parallel sequencing (VAMP-seq) and double-deep Protein fragment Complementation Assay (ddPCA), allow for interrogation of broadly applicable phenotypes across many genes, but fail to capture any specific functions that a gene product may have, like enzymatic activity.[4,27] When analyzing the generalized phenotype data from VAMP-seq or ddPCA to apply in clinical settings, these MAVE data often miss a large fraction of non-functional variants, because their dysfunction manifests in an unmeasured phenotype.[4–6,107,108] Additional MAVEs measuring new phenotypes can improve functional variant predictions but require time-consuming assay validation.[18,108]

In contrast are assays like Saturation Genome Editing (SGE),[16] which profile the effect of gene variants on cell growth in a haploidized line. Examples include SGE mapping for *BRCA1* and *NPC1*.[16,34] The functional scores generated from these types of assays are an amalgamation of all the molecular phenotypes for a given gene product. While this type of output is easier to translate into clinical applications and variant effect prediction, it comes at the loss of a mechanistic understanding of a variant's dysfunction. There is no simple or intuitive way to deconvolute the SGE

functional score to identify and rank the relative importance of the gene's many molecular phenotypes.

MAVEs are also limited insofar as they often remove the genetic element of interest out of its endogenous cellular and genomic context. Notably, most MAVEs are performed in systems that fail to recapitulate their endogenous biological context, which can lead to spurious conclusions on a variant's function.[109] Overexpression cDNA assays, for instance, cannot assess the effects of splicing, have been shown previously alter measurements of protein-protein binding affinities, create spurious protein interactions, and lead to ectopic sequestration in unexpected cellular compartments.[110,111] Moreover, lentiviral approaches to introducing variants can lead to off-target lentiviral integration[112] or lentiviral template swapping.[113]

Even MAVE experiments performed in endogenous contexts, like SGE, still can suffer from unexpected or unwanted issues that confound interpretation of experimental results. For example, most of these techniques rely on precise genome editing, usually via CRISPR-Cas9-based gene delivery and repair, to introduce variants within their endogenous context. However, these techniques remain limited to gene regions with compatible protospacer-adjacent motifs (PAMs) for CRISPR-Cas9 site recognition. Even if appropriate PAMs are available, experiments can suffer from off-target indel formation and chromosomal rearragnements, which can disrupt expressed genes and decouple variant effects.[114–118] SGE, in particular, is limited to essential genes, and must be performed in haploidized cells, to identify functional effects on growth.[16] Furthermore, while MAVEs are comprehensive, the data generated are often noisier than their individually measured counterparts and are heavily reliant on the distribution

of elements within the library.[119,120] As such, careful validation and replication of results with orthogonal methods is essential to interpret MAVE data.

1.5 CRISPR/Cas9-based editing is integral to modern variant interpretation and correction

As alluded to previously, genomic tools for interpreting genetic variation often rely on Cas9-based gene editing to install variants in endogenous contexts or to knock down endogenous gene expression.[5,16,34] Being able to make targeted DNA modifications in the genome reliably, efficiently, and precisely has been of great interest to better parse the relationship between genotype and phenotype. Moreover, precise editing of DNA has immense therapeutic potential via correction of pathogenic mutations that result in genetic disease.[121,122]

1.5.1 Genome editing tools introduce DSBs for cellular repair

Nuclease-targeted genome editing is performed by introducing a double strand break (DSB) that is then repaired by the cell's endogenous DNA repair machinery.[123] Typically, in mammalian cells, the two potential pathways after formation of a DSB are non-homologous end joining (NHEJ) or, less commonly, homologous recombination (e.g. HDR). Because NHEJ involves resection of mismatched end nucleotides of DNA before ligation, NHEJ often results in the formation of small indels in mammalian cells. As such, targeted nucleases have traditionally been used to introduce frameshift-causing indels to knock out a particular gene product.[5,34] The high rate of indel formation in mammalian cells was previously thought to be the result of NHEJ being

particularly error-prone.[124] However, recent work suggests that instead, NHEJ is primarily error-free, and that Cas9-based recutting of DNA results in the slow accumulation of indels.[122–127] In contrast, in HDR, 5' and 3' homologous regions of the sister chromatid or an exogenously provided fragment of DNA are used to bridge the DSB and fill in the gaps between.[128] Because HDR primarily leads to the formation of either a scarless repair (if the sister chromatid is used) or introduction of a desired edit (from the exogenous DNA fragment), it is widely applied to correct pathogenic mutations in cells derived from patients or to study the effects of variation on a particular gene.[16,34,121,129]

1.5.2 A short history of the nucleases used for genome editing

While they are the most widely used today, CRISPR/Cas9 is only the latest in a long line of engineered endonuclease systems to introduce changes to genomic DNA via introduction of a DSB. Earlier endonucleases include zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs). Both ZFNs and TALENs are fusions of an engineered DNA-binding domain and the FokI nuclease domain.[130,131] ZFNs utilize modular protein domains that each recognize 3 nucleotide sequences and can be fused together to target genomic DNA more efficiently.[130] TALENs also use a modular method, but in this case, each module recognizes a single nucleotide, allowing for more sequence targeting possibilities than ZFNs.[131] However, both tools are limited in their targeting scope and are generally considered difficult to engineer. Laborious rounds of optimization are required to ensure that a designed ZFN or TALEN will bind and edit at the target site. Moreover, redesign and re-optimization are required every

time a new target site in the genome is desired, which has prevented widespread use of these tools in both clinical and research settings.

1.5.3 Using CRISPR/Cas9 for editing mammalian genomes

CRISPR/Cas9 quickly revolutionized genome editing, as it overcomes most of the limitations surrounding ZFPs and TALENs. In nature, Cas9 uses two RNA molecules, a CRISPR-RNA (crRNA) that is usually 20 nucleotides in length, to match to and recruit Cas9 to specific genomic DNA sequences, and a second *trans*-activating RNA (tracrRNA) which pairs with the crRNA and Cas9 to form the catalytic complex.[132] In practice, these RNAs are combined into a single guide RNA (sgRNA) that can simultaneously perform both functions.[133] The only sequence requirement for *S. pyogenes* Cas9 (spCas9) RNA binding is the presence of an NGG PAM immediately 3' of the 20 nucleotides matched by the RNA molecule.[132,134] This PAM sequence requirement can change based on the species from which the Cas9 was derived, though spCas9 is the most well characterized and most widely used.[135]

1.5.4 Off-target editing is a significant problem for CRISPR/Cas9

While Cas9 has made it simpler to perform targeted genome editing, challenges still remain with its target specificity, which poses a significant problem for its use in clinical settings. Cas9's lack of target specificity comes from its 20 bp sgRNA targeting sequence, which, between the human genome's large size and highly repetitive structure, is insufficient to target a single unique locus.[136] To make the problem worse, Cas9 retains cleavage activity with up to 3 base truncations or 5-6 base mismatches

between its targeting sgRNA and genomic target.[115–117,137–140] *In vitro* and *in vivo* investigations have determined that the number and location of these mismatches are the main drivers of off-target activity. Specifically, PAM-proximal (within the 8-10 nucleotide seed region) mismatches are less tolerated than PAM-distal mismatches, as the seed region is involved in DNA binding and unwinding.[141] Even then, seed region mismatches only modestly reduce DNA binding and unwinding.[142] Recent structural work provides an explanation—because Cas9 does not directly contact the major or minor groove of its target DNA, which typically provides a steric mechanism for correct basepairing, Cas9 can accommodate noncanonical basepairing, including nucleotide skipping and shifting to make imperfect matches that maintain catalytic activity.[140]

A number of orthogonal attempts have been made to reduce Cas9-based off-target editing, for instance, by using Cas9 ribonucleoproteins (RNPs) formed *in vitro* to lower Cas9 exposure, developing truncated sgRNAs of 17-19 bases (tru-sgRNA) that reduce Cas9 target binding affinity, and engineering high-specificity Cas9 variants.[126,143–156] Each technique can reduce off-target editing, but has drawbacks in terms of ease of use or broad applicability. High-specificity variants in particular seem to reduce a majority, but not all, off-target editing at the cost of efficient on-target editing, the negative effect of which is more pronounced *in vivo* or when used as RNPs.[156–158] High-specificity Cas9 variants appear to balance on- vs off-target activity by adjusting the conformational proofreading mechanism by the REC3 domain, and as such, tend to cleave DNA at the same, refractory off-target sites.[156,159]

Off-target editing, even when rare, can have large consequences for interpreting the results of an experiment or in clinical settings. In addition to short indels and single

nucleotide variants that are typically associated with off-target editing, recent work has identified large structural variants that arise at both on- and off-target loci.[137,138,160] Indeed, DSBs induced by Cas9 editing have led to translocations, chromosomal deletions, inversions, chromothripsis, aneuploidy, and loss of entire chromosomes.[161] Underscoring the importance of identifying potentially unforeseen editing events before moving Cas9 to clinical settings, early gene therapy patients developed and died cancer that was brought about by unexpected translocations and integrations.[162] Even in a research setting, identifying off-target editing is important. Off-target variants can have bizarre, unexpected effects on phenotype. For example, a highly efficient sgRNA for neutralizing human immunodeficiency virus (HIV-1) replication in cells resulted in a 50% decrease in cell viability due to off-target editing.[163,164] When the data were re-analyzed accounting for the cell death phenotype, the viral titers for sgRNA-treated cells were indistinguishable from negative controls. This “highly efficient” sgRNA was in fact, unable to neutralize HIV and instead, its entire phenotype could be simply attributed to off-target toxicity.[163]

While off-target effects like the examples given are fairly straightforward to capture, expanding Cas9’s use into MAVEs and clinical use represents a much more laborious challenge with respect to categorizing and preventing off-target editing across many sgRNAs and loci. As such, a scalable, flexible, and orthogonal method for decreasing off-target editing is required. **Chapter 3** describes my attempts to reduce off-target editing via a new mechanism—the use of catalytically dead RNA sgRNAs (dRNAs) to self-compete away Cas9-mediated editing at specific off-target sites—that is orthogonal to and can be used in conjunction with existing off-target technologies.

2 MultiSTEP: a high-throughput method to identify sequence-function relationships in secreted proteins

A version of this chapter is being prepared for publication:

Popp, N.A., Powell, R.L., Zapp, B.D., Wheelock, M.K., Chang, A.T., Lannert, K.L., Sheehan, J.P., Johnsen, J.M., & Fowler, D.M. A massively multiplexed method to profile missense variation in extracellular proteins reveals biochemical features necessary for secretion and γ-carboxylation of coagulation factor IX.

2.1 Introduction

The sequencing of many human genomes and the emphasis on precision medicine promises personalized medical care on the basis of a patient's genetic information. However, this promise is only partially fulfilled, because DNA sequencing cannot determine the clinical consequences of most rare genetic variants. In fact, of the 7.1 million missense variants in the gnomAD database, less than 2% have clinical interpretations in ClinVar.[2,3,9,165,166] Instead, the majority of variants found in patients are variants of uncertain significance (VUS), meaning that their association with pathogenicity is unclear from current evidence.[9,12] Lack of certainty around the functional impact of variants precludes accurate genome-based diagnostics and prognostics, limits our ability to dissect the relationship between DNA sequence and protein function, and prevents the identification of biochemical mechanisms by which variation causes disease. In order to realize the full potential of precision medicine and improve clinical care, clear and accurate variant interpretation is required.

The root of the problem can, in part, be traced to current tools for understanding and characterizing DNA variation. Most common variants can be associated with disease using tools like genome-wide association studies, but these tools are underpowered for characterizing rare variation, which represents the majority of unknown variants.[10] Detailed pedigrees showing co-segregation of disease and variant, on the other hand, can provide clear evidence for disease association, but are costly, time-consuming, and can only characterize one variant in one context. Computational variant effect predictors are scalable to millions of rare variants, but currently lack precision for clinical use.[11] Even if one or more of these approaches yields a conclusive pathogenic interpretation, understanding the nuanced variant phenotypes or biochemical mechanisms of action requires difficult follow up experiments. However, functional assays are too slow and expensive to implement individually for thousands of variants.

MAVEs can be a viable tool for assessing the functional consequences of genetic variants at scale and have been successfully applied to study the effects of variation on oncogenes and tumor suppressors,[4,16,18,167,168] aggregation propensity,[20,23] pharmacogenes,[5,6,21] viral evolution and antibody escape,[22,24] and membrane trafficking.[26] However, the current technologies used in MAVEs are limited in the scope of genes that can be studied, insofar as they require the variant proteins to be expressed intracellularly or embedded within the extracellular membrane. Secreted extracellular proteins, which represent approximately 10% of known human genes and disproportionate fraction of inherited Mendelian disorders,[35] are incompatible with the DNA sequencing-based readouts of typical MAVEs.

To expand the MAVE toolkit to encompass secreted proteins, we developed Multiplexed Surface Tethering of Extracellular Proteins (MultiSTEP). MultiSTEP combines mammalian cell surface display with a genomically-encoded landing pad[8] that can express a library of gene variants with a single copy per cell. While library-scale display methods have been developed previously for *E. coli*[40,41] and *S. cerevisiae*,[38,39] use of these organisms has typically been for displaying intracellular proteins that do not require extensive PTMs. Mammalian cell display systems that allow for PTMs have also been developed, but rely on artificial secretion signals to express proteins on the surface of cells.[25,169–171]

We applied our MultiSTEP to study the fitness effects of missense variation in the gene *F9*, which encodes for coagulation factor IX (FIX). Variation in *F9* can lead to the coagulation disorder hemophilia B,[49,59] thrombosis,[61], and warfarin sensitivity.[62,63] In a recent cross-sectional study of 1,616 male hemophilia B patients, 98% had an identifiable variant in the *F9* gene, and of these, 82% were found to be missense variants.[59] However, characterization of the functional defects for the vast majority of *F9* variants is lacking, and as a result, 48% remain classified as VUS.

2.2 Results

2.2.1 MultiSTEP is compatible with diverse secreted proteins

MAVEs, in general, have a strict requirement that the measured phenotype be directly linked to a given genotype, which limits their use to intracellular and membrane-bound proteins. Secreted proteins break the genotype-phenotype link as they exit the

cell and enter the extracellular space. To expand MAVEs to encompass secreted proteins, which account for approximately 10% of known human genes,[35] we developed a mammalian cell surface display system that is compatible with variant libraries, which we call MultiSTEP (**Figure** 2.1**a,b**). We re-engineered a VAMP-seq construct to co-express any secreted protein on the surface of mammalian cells with an mCherry fluorescent transcriptional control via an internal ribosomal entry site (IRES) (**Figure** 2.1**c**).[4] We fused the secreted protein at its C-terminal end to a flexible $(GGGGS)_4$ linker and single pass transmembrane domain derived from CD28.[172] To allow for antibody-based detection of any displayed protein, we included a Strep II tag (NWSHPQFEK) in the center of the $(GGGGS)_4$ linker (**Figure** 2.1**d**).[172]

To validate MultiSTEP, we recombined a variety of secreted proteins into a genomically-integrated landing pad[8] in Freestyle 293-F cells, a HEK-293 derivative that has been adapted to produce high quantities of secreted proteins in suspension culture.[173] Cells were tested for surface expression of secreted proteins by staining with an anti-Strep II tag antibody. Flow cytometry of stained cells showed, on average, an 8.25-fold increase in anti-Strep II antibody binding for secreted proteins relative to negative controls (range: 3.8 – 15.2). As expected, removal of the endogenous signal peptide of secreted coagulation factor IX (FIX) ablated all antibody binding (**Figure** 2.1**e**). As the Strep II tag is positioned C-terminally to the secreted proteins in our constructs, we tested whether the surface-displayed proteins were folded correctly. To do so, we stained cells expressing surface-tethered FIX with five antibodies that can detect specific structural folds as well as γ-carboxylation, a required FIX post-translational modification within its GLA domain (**Figure** 2.1**f**).

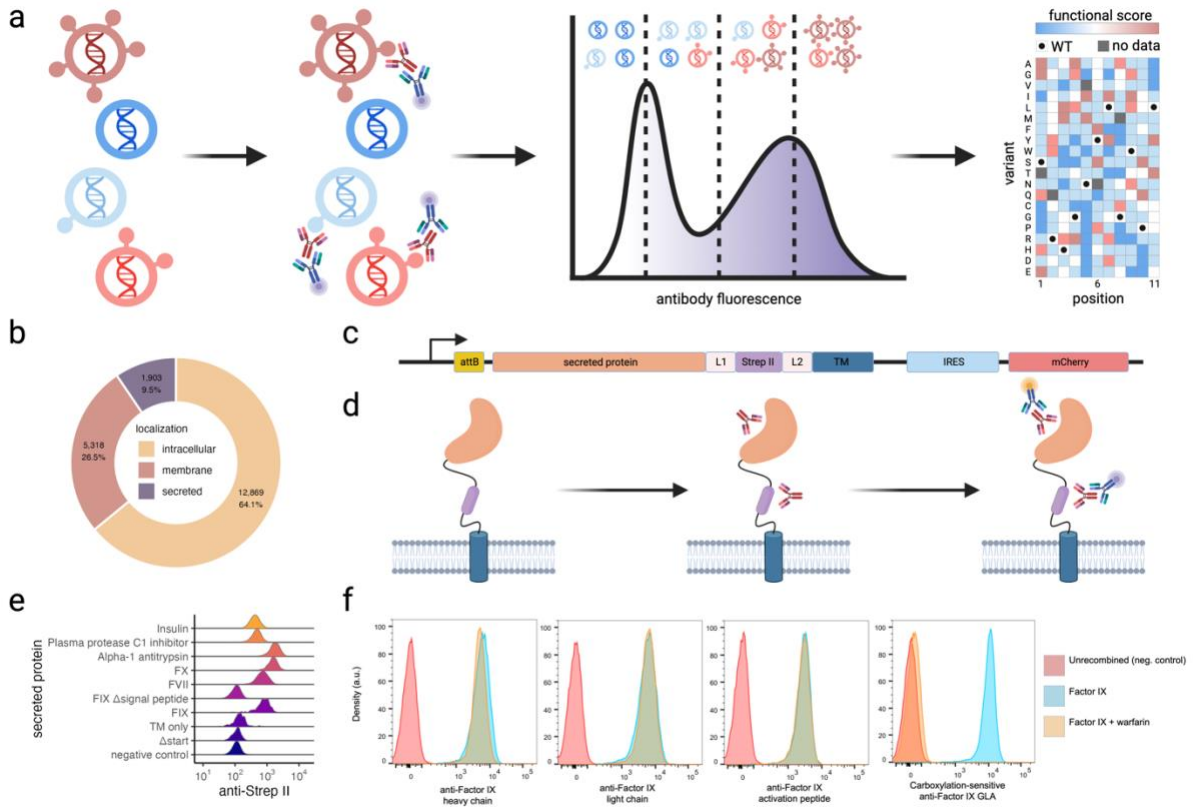

Figure 2.1: MultiSTEP for diverse secreted proteins

a. MultiSTEP allows secreted proteins to be displayed on the cell surface, re-establishing a link between genotype and phenotype. Cells can be sorted based on antibody binding phenotypes and sequenced to derive a functional score. **b.** Secreted proteins (purple) make up approximately 10% of the human proteome. **c.** MultiSTEP construct design. Secreted proteins (orange) are cloned into an attB-containing landing pad donor plasmid for genomic recombination. Secreted proteins are C-terminally fused with $(GGGGS)_2$ flexible linkers (L1 and L2, pink). In between the linkers is a strep II tag (purple). The construct also contains an IRES (blue) driving co-transcription of an mCherry fluorophore (red). **d.** Schematic of MultiSTEP cell surface display. Colors match those in c, except L1 and L2 linkers are shown as black lines. Primary and secondary antibodies directed either against the C-terminal strep II tag (purple) or secreted protein itself (orange) can be used to detect surface expression. **e.** Flow cytometry of MultiSTEP secreted protein constructs stained with an anti-Strep II tag fluorescent antibody. Δstart: removed start codon, TM only: transmembrane domain only, Δsignal peptide: coagulation factor IX with deleted signal peptide. Each density plot represents 30,000 cells. **f.** Flow cytometry of MultiSTEP-expressed FIX stained

with multiple anti-FIX antibodies, each targeting a distinct FIX domain. Cells were grown in the presence of 50 nM vitamin K_1 ± 100 μM warfarin. Each density plot represents 30,000 cells.

We then applied MultiSTEP to study the fitness effects of nearly 10,000 FIX missense variants on FIX secretion and γ-carboxylation. Three site-saturation mutagenesis FIX sublibraries encompassing positions 2 to 461 of FIX were constructed and individually barcoded. PacBio long read sequencing was used to generate a barcode-variant map to identify single amino acid variants within each sublibrary. Together, the sublibraries covered 8,532 of the 8,740 (97.6%) possible missense variants in FIX. Sublibraries were then recombined into our engineered 293-F landing pad line, and successful recombinants were enriched by treating cells with AP1903, a small molecule that kills unrecombined cells.[8] Cells were then stained with one of five antibodies and sorted into quartile bins based on the ratio of antibody fluorescence to the mCherry transcriptional control (**Figure** 2.1**a**). We deeply sequenced genomic DNA form the cells in each sorted bin to calculate the binwise frequency of each variant. Antibody binding scores were then generated using the weighted average of each variant's binwise frequency and was min-max normalized to the median of the 5th percentile of lowest scoring variants and the median of the synonymous distribution. Two to three replicate sorts from separate transfections were performed for each antibody-sublibrary pair.

After filtering out poorly represented variants (**Appendix A: Figure A.**1), we were able to score 8,488–8,512/8,740 missense variants for each of the five antibodies (mean = 8,501 (97.3%) missense variants). Scores for variants shared across experimental replicates correlated well (mean Pearson's r = 0.95, **Appendix A: Figure**

A.2) as did variants shared across sublibraries (mean Pearson's r = 0.96, **Appendix A: Figure A.**3).

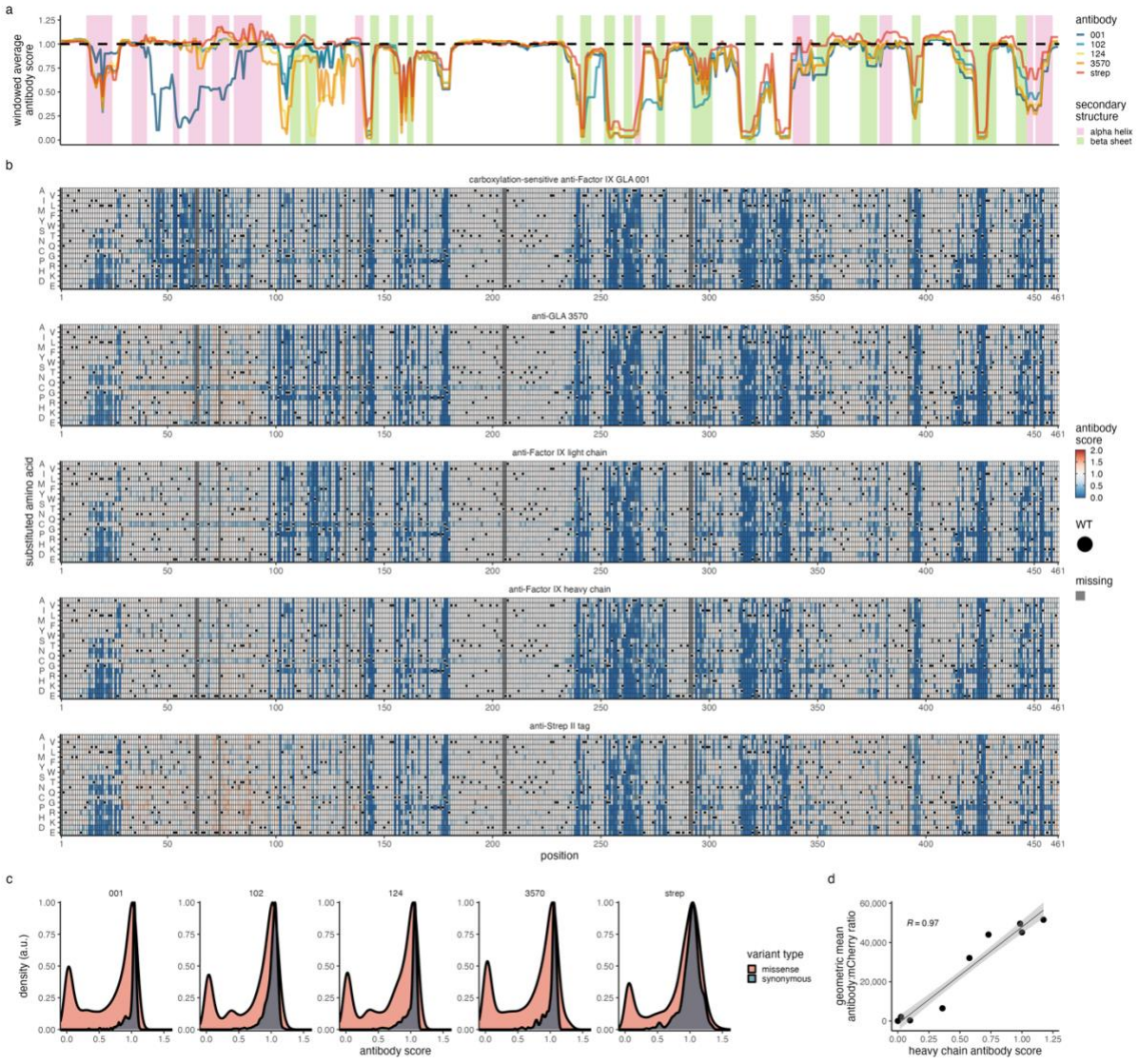

Figure 2.2: Antibody scores for FIX

a. Windowed antibody score medians (width = 7 positions) for each antibody, indicated by color. Pink and green regions identify alpha helices and beta sheets, respectively. **b.** Heatmaps showing antibody scores for nearly all missense FIX variants. Each antibody is shown as its own row. Heatmap color indicates antibody score from 0 (blue, lowest 5% of scores) to white (1, wildtype) to red (increased antibody scores). Black dots indicate wildtype (WT) residues. Missing data are grey. **c.** Density distributions for FIX

missense variants (red) and synonymous variants (blue) measured by MultiSTEP. **d.** Scatterplot comparing MultiSTEP derived variant scores and mean antibody:mCherry ratios (n = 3) measured individually using flow cytometry. Antibody shown is anti-FIX heavy chain (GMA-102). Error bars showing standard error of the mean is plotted as well, though most are smaller than points. Pearson's r shown at top left.

2.2.2 MultiSTEP can identify variant fitness effects on secretion and stability

An unresolved question in hemophilia B genetics is which variants lead to quantitative (amount of FIX secreted) vs. qualitative (functional activity of secreted FIX) deficits in patients and why. Many quantitative variants are so severe that they are designated cross-reactive material negative (CRM^-), meaning there is no detectable FIX antigen or FIX enzymatic activity in the blood of patients harboring these variants.[77] CRM^- variants are nearly always associated with severe disease and show an increased risk of developing neutralizing antibodies (inhibitors) to their replacement FIX therapy.[70] However, other quantitative variants show reduced FIX blood antigen but remain CRM^+ and retain enzymatic activity.[78] Many of these variants perform indistinguishably from wildtype FIX *in vitro*, but their reduced secretion results in a bleeding phenotype.[78] Because inhibitors can cause life-threatening anaphylactic reactions in hemophilia B patients,[69,174] and because clinical measurement of FIX blood antigen is rarely performed, there is a dearth of knowledge of which variants affect FIX secretion *en masse*.

Similar to previous multiplexed studies of variant abundance,[4–6,21] we sought to determine whether MultiSTEP could quantify how much each variant FIX protein could be secreted. To measure secretion via surface expression, we profiled each sublibrary

with antibodies directed against the C-terminal Strep II tag, FIX heavy chain, and FIX light chain (**Figure** 2.2**a-b**). In each case, the distribution of missense variant scores was bimodal with variant effects spanning the range of scores (**Figure** 2.2**c**). Cells harboring synonymous variants showed antibody scores similar to wildtype FIX, as expected (**Figure** 2.2**c**). Internal validation of 7 randomly selected variants spanning the range of antibody scores was performed using the heavy chain antibody. The individual antibody:mCherry ratios assessed by flow cytometry showed a strong correlation with antibody scores (Pearson's r = 0.97, **Figure** 2.2**d**).

2.2.3 MultiSTEP can quantify PTM status

In contrast to quantitative variants that reduce the amount of secreted FIX, qualitative variants (which are also CRM^+), show excess FIX activity loss relative to their blood antigen levels.[78,79] These variants have been sporadically described throughout the literature with a variety of phenotypic effects, including loss of binding partner affinity, reduction in enzymatic activity, or loss of required PTMs. Because FIX antigen is rarely performed outside the context of research settings, little is known about which variants lead to qualitative deficits or which of the many possible mechanistic routes lead to their dysfunction. What is known is that these qualitative variants generally affect some phenotype besides secretion.[80,81]

To expand the scope of fitness effects that can be measured using MultiSTEP, we decided to profile the effects of FIX variants on the γ-carboxylation of its GLA domain. Proper γ-carboxylation of the 12 glutamate residues in the GLA domain is required for FIX function, as γ-carboxylated residues specifically coordinate with free

Ca^{2+} and Mg^{2+} ions to form a folded three-helical structure with an exposed hydrophobic ω-loop.[52,84–86] The ω-loop is responsible for calcium-dependent interactions with phospholipid membranes and formation of the tenase complex,[52–54] activation of FIX by activated factor XI (FXIa),[51,87] and interactions with activated factor VIII (FVIIIa).[88] A small subset of FIX variants, both proximal and distal to the ω-loop, alter its structure and binding affinity,[50,62,63,95,96] but a comprehensive analysis of variant effects on ω-loop structure has not yet been performed.

FIX γ-carboxylation is vitamin K dependent and blocked by warfarin, an anticoagulant that prevents vitamin K recycling. We first sought to show that our surface-displayed FIX was properly γ-carboxylated by staining FIX-expressing cells with two γ-carboxylation-dependent antibodies in the presence of vitamin K. The conformation-sensitive antibody recognizes the exposed ω-loop structure that is formed with proper γ-carboxylation and has been shown to act as an anticoagulant.[51,103,104] The second anti-γ-carboxylation antibody reacts broadly with many known γ-carboxylated proteins and is proposed to interact with γ-carboxylated glutamates within a conserved epitope within the GLA domain[97,175] As expected, FIX-expressing cells bound both antibodies strongly compared to FIX-expressing cells treated with warfarin (**Figure** 2.1**f**).

We then used the γ-carboxylation-dependent antibodies to generate scores for nearly all possible FIX missense variants (**Figure** 2.2**b**). Similar to the secretion antibody scores, missense variant effects were bimodal, and synonymous variants scored close to wildtype (**Figure** 2.2**c**). Secretion and γ-carboxylation antibody scores were highly correlated with one another for all of FIX (except the propeptide and GLA

domains where the two γ-carboxylation antibodies show distinct mutational signatures (**Appendix A: Figure A.**4). By comparing secretion antibody scores with γ-carboxylation-dependent antibody scores, we were able to identify variants that reduce both γ-carboxylation specifically, as well as those that interfere with proper folding of the γ-carboxylated GLA domain. We hypothesized that variants that disrupted γ-carboxylated GLA domain folding and exposure of the hydrophobic ω-loop would show the largest decrease in γ-carboxylation-dependent antibody score relative to its secretion score. Indeed, variants with the lowest γ-carboxylation-dependent antibody scores and wildtype-like secretion scores included variants at positions 43 through 46 (**Figure** 2.2**a-b**). Variants at these sites remove a PACE cleavage motif and prevent the removal of the FIX propeptide, which subsequently causes the GLA domain to lose affinity for phospholipid membranes.[50,95] Other highly discordant variants occurred at position 37, a key residue in the GGCX binding motif that has previously been shown to ablate γ-carboxylation of FIX (**Figure** 2.2**b**).[62,63]

2.2.4 Secretion scores identify FIX antibody epitopes

We first compared secretion antibody scores to one another to identify any potential loss of function FIX variants that alter antibody epitopes without affecting secretion. We defined epitope-sensitive positions as those with median missense variant scores that shifted upwards or downwards by at least 0.33 and found that 20 positions (10 per antibody) met these criteria (**Appendix A: Figure A.**5**a**). Both sets of positions mapped to discrete solvent-exposed surfaces on FIX despite considerable distance in linear space (**Appendix A: Figure A.**5**b**). The identification of three-

dimensional antibody epitopes on FIX further supports our conclusion that surface-displayed FIX is conformationally stable.

2.2.5 Mutational effects in the signal peptide

We next interrogated the mutational landscape of the N-terminal secretion peptide. While there is considerable diversity in the length and sequence of signal peptides throughout the proteome, three distinct functional regions are conserved: the n-region, which is weakly positively charged; the h-region, which contains the hydrophobic helix that binds to SRP54 to initiate translocation into the ER,[176,177] and the c-region, which breaks the helix and contains the AxA cleavage motif.[43] We identified three matching mutational signatures in our secretion scores that correspond to these three regions (**Figure** 2.3**a**). The boundaries between mutational clusters mapped closely to predictions generated by SignalP 6.0 (**Figure** 2.3**a**).[178] We compared SignalP 6.0 predictions against all variants in our library and found that nearly one third of secreted variants were misclassified as non-functional (**Figure** 2.3**b**).

Mutational signatures largely mapped to expectations (**Figure** 2.2**a**). The positively charged n-region mutational cluster was characterized by mild loss of secretion for negatively charged variants but was otherwise tolerated. The hydrophobic h-region was largely tolerant to hydrophilic substitutions but not to polar or charged variants. We used GRAVY[179] to calculate the hydrophobicity of all substitutions within the h-region and found that there was a strong correlation between secretion antibody scores and hydrophobicity index (Pearson's r = 0.75, **Figure** 2.3**c**). As expected, the c-region was generally intolerant to large, polar, or charged substitutions at the -3 and -1

positions within the AxA motif. Surprisingly, hydrophilic variants in the spacer of the AxA motif showed loss of secretion, (**Figure** 2.3**a**) potentially by extending the hydrophobic helix beyond the cleavage recognition site for SRP54.[43]

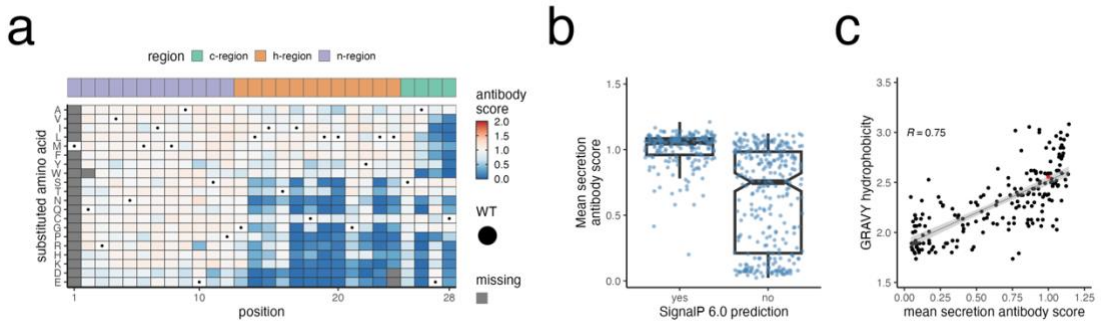

Figure 2.3: Biochemical features of secretion peptide variants

a. Top row: Predicted regions of signal peptides for wildtype FIX from SignalP 6.0, indicated by color.[178] Bottom row: Heatmap of secretion scores for the FIX signal peptide. Color indicates secretion antibody score. Black dot is wildtype. Missing is annotated in grey. **b.** Comparison of secretion antibody scores with SignalP 6.0 predictions for where FIX variants constitute a functional signal peptide. **c.** Scatterplot comparison of FIX variant secretion antibody scores with GRAVY hydropathy index. Only positions 13-24 of the h-region are included in this plot. *R* = Pearson's correlation. Linear trendline is annotated with grey shaded area denoting the 95% confidence interval. All secretion scores shown in this plot represent the mean taken across FIX heavy chain (102), FIX light chain (124), and strep II antibodies.

2.2.6 Unpaired cysteines alter secretion mechanics and likely disrupt proper folding

Because secreted proteins require disulfide bonds to maintain structural integrity in the oxidizing conditions of the extracellular space,[47,180] we hypothesized that cysteine variants would lead to significant secretion deficits. We found that most wildtype cysteine residues in the tightly packed protease domain were nearly completely

intolerant to substitution (**Figure** 2.4**a; Appendix A: Figure A.**6), though positions 407 and 435 showed only moderate effects. On the other hand, variants in the GLA, EGF1, and EGF2 domains resulted in moderate secretion deficits when measured with strep II tag but more pronounced effects with all antibodies that directly target folded FIX (**Figure** 2.4**a; Appendix A: Figure A.**6).

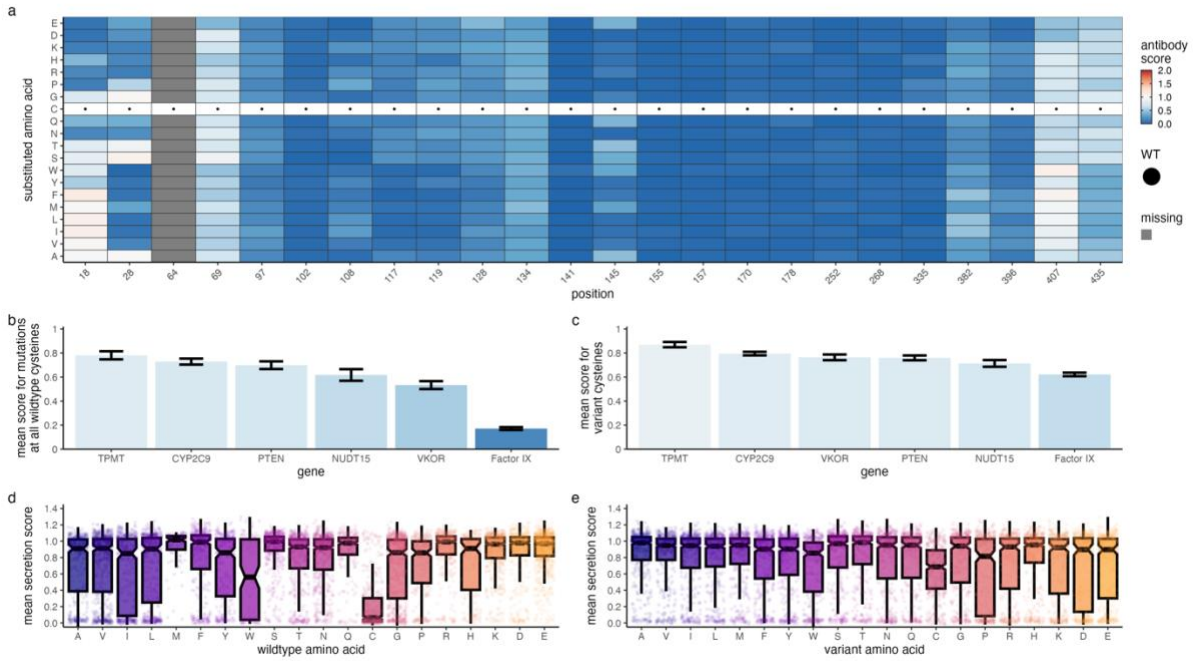

Figure 2.4: Cysteine variants show an outsized effect on FIX secretion

a. Heatmap of secretion scores for variants at wildtype cysteines in FIX. Color indicates secretion antibody score. Black dots indicate wildtype FIX amino acids, and grey are missing data. **b.** Comparison of variant scores across genes with measured abundance using VAMP-seq. Bars indicate mean abundance or secretion score for all variants at wildtype cysteine residues. Error bars show standard error of the mean. **c.** Comparison of variant cysteine scores across genes with measured abundance using VAMP-seq. Bars indicate mean abundance or secretion score for all nonsynonymous cysteine variants. Error bars show standard error of the mean. **d.** Comparison of secretion scores for all FIX variants at wildtype amino acids. x-axis is the wildtype residue and points indicate variant secretion scores at that residue. **e.** Comparison of secretion scores for all variants amino acids in FIX. x-axis is the

variant amino acid and points indicate variant secretion scores for that amino acid at all residues in FIX. Scores shown in all plots are FIX heavy chain (102), though are representative of all secretion antibodies.

Supporting our hypothesis that cysteine variants affect FIX secretion and folding, the only unpaired cysteines in FIX at positions 18 and 28 of the secretion peptide showed an alternative mutational pattern that matches more closely with their adjacent residues. Moreover, the mean functional effect of variants at wildtype cysteines in FIX was significantly reduced compared to other intracellular proteins profiled for abundance (**Figure** 2.4**b**).[4–6,21] We also identified a signature loss of function pattern for novel cysteine variants through the mature FIX protein (**Figure** 2.2**a**) that was only present for antibodies that directly target FIX. This signature is not apparent in intracellular proteins assayed for abundance (**Figure** 2.4**c**),[4–6,21] nor is it present for other amino acid substitutions in FIX (**Figure** 2.4**d-e**).

2.2.7 Hierarchical clustering identifies positions with similar functional features across assays

We expect that FIX will display unique mutational signatures for secretion and γ-carboxylation, as *in vitro* work has demonstrated that proper γ-carboxylation is not required for secretion.[105,106] We also anticipate that variants in buried regions will lead to poor secretion due to improper folding. To further explore these hypotheses, we performed hierarchical clustering on all five antibody scores by variant position. Looking at the six main clusters, we see that cluster 6 contains residues that are, in general, deeply buried in the FIX structure (**Figure** 2.5). This cluster also contains all of the disulfide-paired cysteine residues in FIX, further highlighting their importance for proper

folding and secretion. Cluster 4 identified variants that disrupted the binding of our conformation-sensitive γ-carboxylation antibody specifically, of which all residues resided in the propeptide and GLA domains, as expected. Finally, clusters 1 and 2 identified variants that were highly tolerant to variation, including all residues in the FIX activation peptide (**Figure** 2.5).

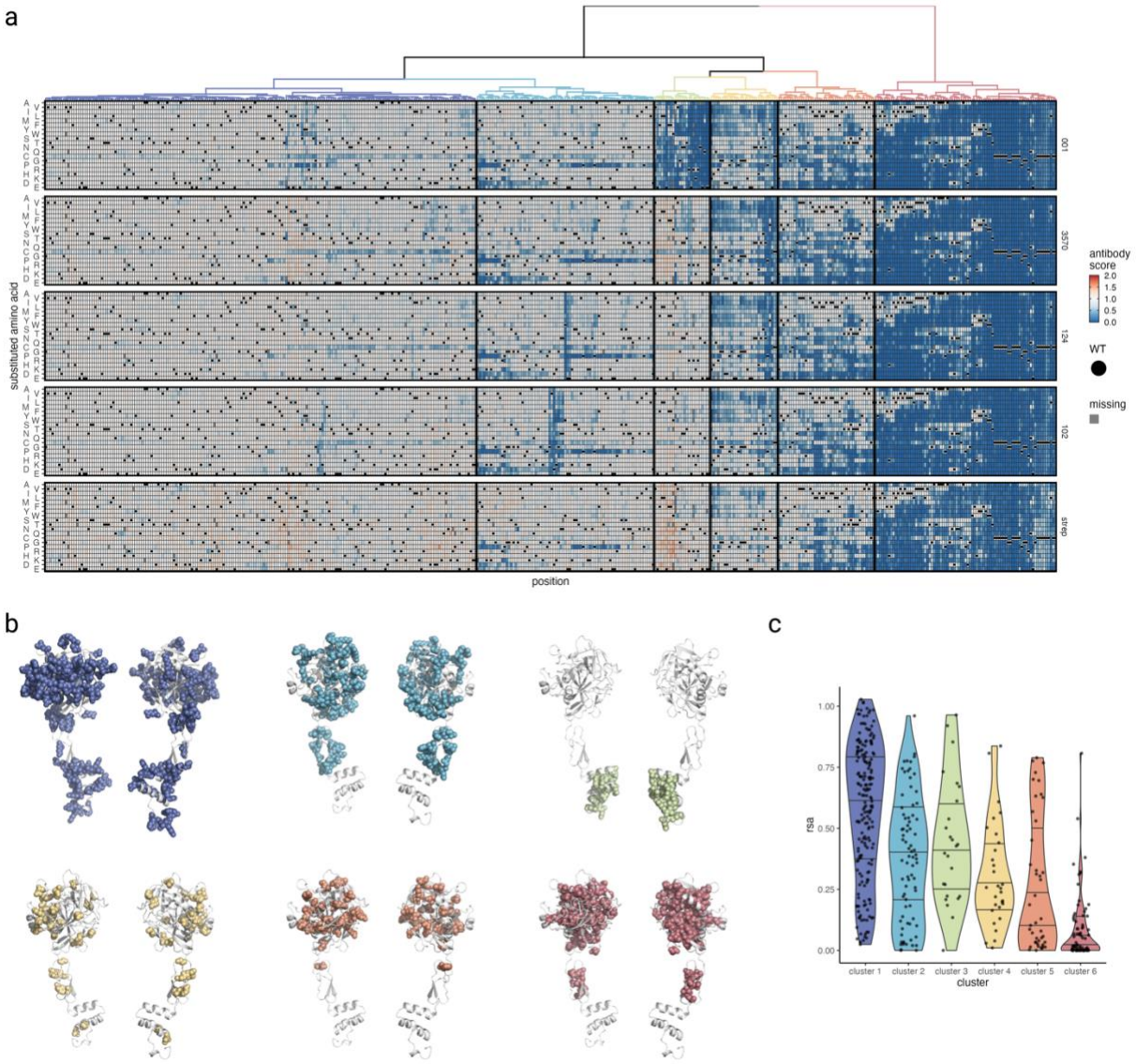

Figure 2.5: Clustering of secretion and γ-carboxylation scores

a. Dendrogram and heatmaps of FIX antibody scores by position clustered by position. Heatmaps colored by antibody score. Colors in dendrogram represent clusters. Black dots indicate wildtype residues and grey indicates missing data. Ten positions for which no data was recorded are removed from the analysis. **b.** FIX AlphaFold ribbon structure with low confidence positions removed. Side chains are shown as spheres and are colored by cluster matching those in **a.** Each structure is shown twice with 180 ° rotation. **c.** Relative solvent accessibility for each cluster shown as violin density. Horizontal lines represent the 25th, 50th, and 75th percentiles of data in each cluster.

2.2.8 Discovering potentially pathogenic variants

Because loss of secretion is a common mechanism for loss of function in FIX, we compared our secretion antibody scores to established FIX antigen levels from the European Association for Haemophilia and Allied Disorders (EAHAD) public database.[77] As expected, secretion antibody scores correlated with patient FIX antigen levels (Pearson's r = 0.70, **Figure** 2.6**a**). We then looked at hemophilia severity and found that patients with severe disease were more likely to have low secretion antibody scores than patients with moderate or mild disease (**Figure** 2.6**b**).

We attempted to predict variant function using our antibody scores on a curated set of FIX variants from ClinVar, the *MyLifeOurFuture* (MLOF) hemophilia sequencing project, and gnomAD. We trained a random forest classifier and evaluated model performance on the 25% of data that was removed before training. Because the training set was highly imbalanced (9.1% benign/likely benign and 90.1% pathogenic/likely pathogenic), we performed class-based random oversampling (ROSE).[181] The model correctly classified 100% of benign variants as having wildtype-like function and 61.7% of pathogenic variants as having loss of function, leading to an ROC-AUC of 0.934 (**Figure** 2.6**c**). When applied to all variants measured in our assays, the random forest

model predicts that 3,520 (40.9%) would be loss of function. The number of predicted loss of function variants closely mirrors that of abundance measurements.[4–6]

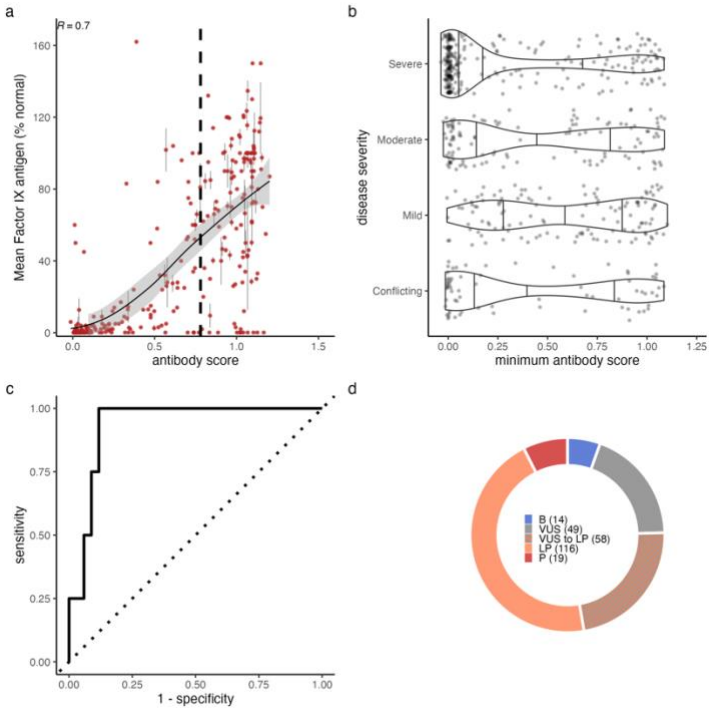

Figure 2.6: FIX antibody scores identify clinical features of hemophilia B

a. Scatterplot of secretion antibody scores and FIX antigen from hemophilia B patients in the EAHAD database.[77] Secretion antibody scores represent the mean score per variant for secretion antibodies (102, 124, strep). Vertical solid lines indicate standard error of the mean for FIX antigen levels across patients harboring the same variant. Dashed vertical line is the 5th percentile of the synonymous secretion score distribution. **b.** Comparison of EAHAD patient hemophilia B severity with antibody scores from MultiSTEP. Plotted antibody scores represent the minimum score across all five antibodies. Conflicting classification included when a single variant is associated equally with multiple disease severities across patients. Violin plot shows distribution of points with vertical lines representing the 25th, 50th, and 75th percentiles. **c.** ROC curve for random forest model for identifying loss of function FIX alleles from MultiSTEP scores. **d.** MLOF variant classifications after reinterpretation using existing evidence and MultiSTEP functional evidence. Numbers represent the number of variants with each updated classification. Dashed areas

represent variants that were reclassified from VUS to likely pathogenic. B: benign, LB: likely benign, VUS: variant of uncertain significance, LP: likely pathogenic, P: pathogenic.

We then calculated the odds of pathogenicity (OddsPath)[182,183] for using our controls (8.71 for loss of function, 0.38 for wildtype-like) which corresponds to evidence levels PS3_moderate toward loss of function and BS3_supporting for wildtype-like classifications according to ACMG criteria.[12,182] We then used these predictions to reclassify VUS from MLOF using following ACMG criteria and were able to reclassify 58/107 (54.2%) up to likely pathogenic (**Figure** 2.6**d**).

2.3 Discussion

MultiSTEP is a generalizable, multiplexed method capable of quantifying the amount of successful secretion in secreted proteins. With the use of additional antibodies, we show that MultiSTEP can also be used to quantify functional PTMs and other biochemical features of secreted proteins. In the context of inherited Mendelian diseases like hemophilia B, a validated assay showing a variant's loss of function can provide evidence for pathogenicity.[12] 64% of known pathogenic variants in *F9* show reduced secretion, reduced γ-carboxylation, or both in our assays Using a random forest model, we were able to apply PS3_moderate evidence to reclassify 58/107 VUS in *F9* up to likely pathogenic, highlighting the utility of this system for determining variant function.

We used MultiSTEP to identify and quantify variant effects throughout FIX. We show three distinct mutational signatures that map to the functional subdomains of secretion peptides.[43] We also surprisingly found an outsized effect for both novel and

removed cysteine variants that is not found in similar assays on variant abundance of intracellular and membrane-bound proteins.[4–6,21] Notably, a number of novel cysteine residues only showed loss of function effects for antibodies that directly bind to FIX but not for the C-terminal Strep II tag, suggesting that these variants permit secretion but fail to fold properly. These findings highlight the importance of disulfide bonds for the stability and activity of secreted proteins.

We also applied MultiSTEP to study γ-carboxylation of the GLA domain of FIX, where we found many variants altered binding to a conformation-specific antibody that sees the functional hydrophobic ω-loop. Though it was long thought that variants that prevent cleavage of the FIX propeptide in the ER block γ-carboxylation,[94,101] more recent work has disputed this claim.[97,184] Our data suggest that these propeptide variants alter binding of the conformation-specific antibody but not a direct γ-carboxylation antibody, supporting the latter hypothesis that retained propeptide interferes with formation of the ω-loop. Indeed, NMR structures of the γ-carboxylated GLA domain of FIX show that the α-amino group of Y47 interacts with γE53, γE67, and γE73.[50,84] Supporting this hypothesis, γE67D alters ω-loop formation via the same interactions,[96] and in our experiments, mutation of any of these three γ-carboxylated residues completely ablates conformation-sensitive antibody binding. A similar pattern is seen in our data in the aromatic stack at the C-terminal end of the GLA domain (positions 87, 88, and 91), supporting work that identified these residues as necessary for the intermediate conformational shift that is required to form the ω-loop.

While the results from our secretion and γ-carboxylation MAVEs illuminates various biochemical mechanisms for FIX dysfunction, MultiSTEP still has some

limitations. Although type I single-pass transmembrane proteins and secreted proteins utilize the same secretion peptide machinery, co-translational processing in the ER may be altered. A similar phenomenon has been described for engineered B-domain deleted coagulation factor VIII (FVIII).[185] Because of the nature of antibody binding, both assays have limited dynamic ranges that may obscure subtly hypomorphic variants. Lastly, from a clinical perspective, these data do not capture an end-stage functional effect of FIX (namely, enzymatic activity), which leaves many known pathogenic variants unaccounted for. As such, caution should be used when using these data to assign wildtype-like or benign status to FIX variants.

Generalizable assays like MultiSTEP provide a promising way to understand the effects of missense variation throughout the genome. In this work, we show that this method can be used to identify multiple functional phenotypes in FIX and provide preliminary evidence that these experiments could be applied to other clinically relevant secreted proteins. We also show that biochemical characterization of variant proteins on the surface of cells can be performed, which vastly expands the scope of this method's applications. Combining data generated in these experiments with further biochemical assays in this system could be used to systematically categorize the biological mechanisms underlying each variant's dysfunction and identify variant-specific treatments. Moreover, integrating an exogenous signal peptide sequence could allow for high-throughput biochemical characterization of intracellular proteins. Overall, our work highlights the value of using MAVEs to better understand protein function and greatly expands our ability to measure the impact of secreted protein variants.

2.4 Methods

2.4.1 General reagents

All *E. coli* were grown in Luria Broth (LB) at 37 °C and shaking at 225 rpm for 16-18 hours with 100 µg/mL carbenicillin, unless otherwise indicated. Routine cloning was performed in homemade chemically competent Top10F' *E. coli*, whereas library cloning was performed in commercially available electrocompetent NEB-10β *E. coli*.

Inverse PCR reactions, unless otherwise specified, were performed in 30 µL reactions with 2x Kapa HiFi ReadyMix (Kapa Biosystems) or Q5 2x master mix (New England Biolabs), with 40 ng starting plasmid and 0.15 µM each forward and reverse primers. Reaction conditions were 95 °C for 3 minutes, 98 °C for 20 seconds, 61 °C for 30 seconds, 72 °C for 1 minute/kb, repeating for 8 total cycles, then followed by a final 72 °C extension for 1 minute/kb, and held at 4 °C. PCR products were then digested at 37 °C with DpnI (New England Biolabs) for 2 hours to remove residual starting plasmid, followed by heat inactivation at 80 °C for 20 minutes. PCR products were then checked on a 1% agarose gel with 1x SYBR Safe (ThermoFisher Scientific) at 100 V, 45 minutes for purity and size, and gel extracted if needed.

For large modifications (>50 bp), PCR products were then Gibson assembled (for large modifications) using a 3:1 molar ratio of insert(s):backbone at 50 °C for 1 hour, after which 2 µL of product was transformed into homemade chemically competent Top10F' *E. coli* (Gibson, et al., 2009). For small modifications, *in vivo assembly* cloning

(IVA cloning) of linear products was used by transforming 5 µL of PCR product directly into Top10F' *E. coli* without recircularization.[186]

For both Gibson assembly and IVA cloning, after addition of PCR-amplified DNA, Top10F' cells were incubated on ice for 30 minutes before a 30 second heat shock at 42 °C. Cells were then returned to ice for 2 minutes before being added to 1 mL SOC to recover for 1 hour at 37 °C, shaking at 225 rpm. After recovery, 100 µL was spread on LB-ampicillin plates and grown overnight for 16-18 hours. Colonies were then screened for correct insertion by Sanger sequencing (small modifications) or colony PCR (large modifications), Sanger sequence confirmed, and miniprepped or midiprepped.

2.4.2 Cloning into the landing pad donor plasmid

To clone attB-F9-10L-StrepII-10L-CD28-IRES-mCherry (pNP0001), first, an inverse PCR was performed on attB-EGFP-PTEN-IRES-mCherry-562bgl[18] with primers NP0207 and NP0325 to remove EGFP-PTEN and create compatible Gibson overhangs using Kapa HiFi polymerase (Kapa Biosystems). A gBlock (NPg0007) containing human *F9* cDNA and a second gBlock (NPg0012) containing a GC-optimized 10 amino acid $(GGGGS)_2$ flexible linker, a strep II protein tag, and the single-pass transmembrane domain of CD28 was then assembled following the standard Gibson assembly cloning protocol above. After sequence-confirmation, a second round of inverse PCR was performed to insert a second $(GGGGS)_2$ flexible linker after the strep II protein tag using primers NP0334 and NP0356 following the standard IVA cloning protocol above.

To clone an empty tethering backbone vector attB-10L-StrepII-10L-CD28-IRES-mCherry (pNP0079), *F9* was removed by IVA cloning following the same protocol as above with primers NP0325 and NP0377. To clone missense *F9* variants for assay validation, point mutations were introduced into pNP0001 by IVA cloning following the same protocol as above.

cDNA constructs for human *F7*, *F10*, *SERPINA1*, *SERPING1*, and *INS* were ordered from the Mammalian Gene Collection (Horizon Discovery) and cloned into the landing pad donor backbone (pNP0079) using Gibson assembly. For each gene, the donor backbone was amplified using NP0295 and NP0325. To generate attB-F7-10L-StrepII-10L-CD28-IRES-mCherry (pNP0088), *F7* cDNA was amplified with NP0615 and NP0616. To generate attB-F10-10L-StrepII-10L-CD28-IRES-mCherry (pNP0089), *F10* cDNA was amplified with NP0617 and NP0618. To generate attB-SERPINA1-10L-StrepII-10L-CD28-IRES-mCherry (pNP0090), *SERPINA1* cDNA was amplified with NP0619 and NP0620. attB-SERPING1-10L-StrepII-10L-CD28-IRES-mCherry (pNP0091) was created by amplifying *SERPING1* cDNA with NP0621 and NP0622. To generate attB-INS-10L-StrepII-10L-CD28-IRES-mCherry (pNP0092), *INS* cDNA was amplified with NP0623 and NP0624.

2.4.3 Site-saturation mutagenesis library cloning

Site-saturation mutagenesis oligonucleotides were ordered from Twist Biosciences for each position in *F9*, except position 1. Each position contained one codon for each synonymous or missense variant. 50 ng of each oligonucleotide were resuspended in 10 µL water, and then pooled in equal volumes into three tiled

sublibraries encompassing the entire length of *F9*, including 20 positions of overlap between adjacent sublibraries. Tile 1 sublibrary: positions 2-164; Tile 2 sublibrary: positions 146-318; Tile 3 sublibrary: positions 299-461.

pNP0079 was inverse PCR amplified using NP0295 and NP0325 following the standard protocol, and the correct size PCR product was gel extracted. The backbone PCR product was then Gibson assembled with each of the three pooled sublibraries at a 5:1 molar ratio of insert:backbone at 50 °C for 1 hour. Gibson assembled products were then cleaned and eluted in 10 µL water (Zymo Clean and Concentrate). 1 µL of cleaned product was then added to 25 µL NEB-10β *E. coli* in pre-chilled cuvettes and allowed to rest on ice for 30 minutes. 2 electroporation replicates were performed per sublibrary tile, as well as a pUC19 control (10 pg/µL). Cells were then electroporated at 2 kV for 6 milliseconds. After electroporation, cells were immediately resuspended in 100 µL pre-warmed SOC and transferred to a culture tube. Identical replicates were pooled at this step, and pre-warmed SOC was added to a final volume of 1 mL and allowed to recover at 37 °C, shaking at 225 rpm, for 1 hour. After recovery, the entire recovery volume was added to 49 mL of LB containing 100 µg/mL carbenicillin and allowed to grow overnight. After 2-3 minutes of shaking, a 200 µL sample was taken and used for serial dilutions to estimate colony counts on LB-ampicillin plates as a proxy for the number of unique molecules transformed and to gauge coverage of the library. After 16 hours of overnight growth at 37 °C, each 50 mL midiprep culture was spun down for 30 minutes at 4,300 x *g* and midiprepped.

2.4.4 Barcoding site-saturation mutagenesis libraries

To barcode each sublibrary, 1 µg of each sublibrary plasmid was digested at 37 °C for 5 hours with NheI-HF and SacI-HF (New England Biolabs), incubated with rSAP for 30 minutes at 37 °C, then heat-inactivated at 65 °C for 20 minutes. Digested product was then run on a 1% agarose with 1x SYBR Safe (ThermoFisher Scientific) gel for 45 minutes at 100V and gel extracted (Qiagen).

A barcode ultramer (NP0490) was ordered from IDT with 18 degenerate nucleotides and resuspended at 10 µM. 1 µL NP0490 was then annealed with 1 µL of 10 µM NP0397 primer, 4 µL CutSmart buffer, and 34 µL water by running at 98 °C for 3 minutes, followed by ramping down to 25 °C at -0.1 °C per second. After annealing, 1.35 µL of 1 mM dNTPs and 0.8 µL Klenow exo- polymerase (New England Biolabs) were added to fill in the barcode oligo. The cycling conditions were 25 °C for 15 minutes, 70 °C for 20 minutes, and then ramped down to 37 °C in -0.1 °C per second increments. Once at 37 °C, 1 µL each NheI-HF and SacI-HF were added and digested for 1 hour. Digested product was then run on a 4% agarose gel with 1x SYBR Safe for 45 minutes at 100V and gel extracted (Qiagen).

Both gel extracted sublibrary and barcode oligonucleotide were cleaned and eluted in 10 µL and 30 µL water, respectively (Zymo Clean and Concentrate). A 7:1 molar ratio of barcode oligonucleotide to sublibrary was ligated overnight at 16 °C with T4 DNA ligase (New England Biolabs).

Ligated product was then cleaned and eluted in 10 µL water (Zymo). 1 µL of ligation product, ligation controls, or pUC19 was electroporated into NEB-10β *E. coli* following the same procedure as above. For sublibrary ligation products, 2 independent

replicates were pooled before recovery. After recovery, each sublibrary ligation product was bottlenecked by diluting various recovery volumes (500 µL, 250 µL, 125 µL, and 50 µL) into independent 50 mL LB-ampicillin cultures. After 2-3 minutes of shaking, a 200 µL sample from each ligation bottleneck was taken and used for serial dilutions to estimate colony counts on LB-ampicillin plates. Colony counts were then used to estimate the number of barcoded variants present in each sublibrary for each bottleneck. After 16 hours of overnight growth at 37 °C, each 50 mL midiprep culture was spun down for 30 minutes at 4,300 x *g* and midiprepped.

2.4.5 Estimation of variant coverage by Illumina sequencing

Each barcoded and bottlenecked plasmid sublibrary was diluted to 10 ng/µL and amplified for Illumina sequencing to determine the number of unique barcodes present more accurately. Briefly, to add adapter sequences, primers NP0492 and NP0493 were mixed at a final concentration of 0.5 µM with 10 ng plasmid DNA, 25 µL Q5 polymerase (New England Biolabs), and 19 µL water. Cycling conditions were an initial denaturing at 98 °C for 30 seconds, followed by 5 cycles of 98 °C for 10 seconds, 61 °C for 30 seconds, and 72 °C for 30 seconds, followed by a final 72 °C extension for 2 minutes and a hold at 4 °C. PCR products were then cleaned using 0.8x AmpureXP beads (Beckman Coulter) and eluted in 16 µL water following the manufacturer's instructions.

The entire elution volume for each sample was then mixed with 25 µL Q5 polymerase, 0.25 µL of 100x SYBR Green I (ThermoFisher Scientific) 4.75 µL water, and one uniquely indexed forward (NP0595-NP0608) and reverse (NP0551-NP0564) primer at a final concentration of 0.5 µM. Samples were run on the CFX Connect (Bio-

Rad) for a maximum of 15 cycles or until all samples were above 3,000 relative fluorescence units. Reactions were denatured at 98 °C for 30 seconds, and cycled at 98 °C for 10 seconds, 65 °C for 30 seconds, and 72 °C for 30 seconds, with a final extension at 72 °C for 2 minutes. Samples were then run on a 2% agarose gel with 1x SYBR Safe for 2 hours at 120V before gel extraction with a Freeze 'N Squeeze column (Bio-Rad). After extraction, samples were quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific), pooled in equimolar concentrations, and sequenced on a NextSeq 550 using a NextSeq 500/550 High Output v2.5 75 cycle kit (Illumina) using custom sequencing primers NP0494-NP0497. Using a custom script, sequencing reads were converted to FASTQ format and demultiplexed using bcl2fastq (v2.20), forward and reverse barcode reads were paired using PEAR (v0.9.11),[187] and unique barcodes were counted.

2.4.6 Barcode-variant mapping with PacBio sequencing

2.5 µg of each sublibrary was digested with AflII (New England Biolabs) in CutSmart buffer for 4 hours at 37 °C, followed by heat inactivation at 65 °C for 20 minutes and purified with Ampure PB beads (Pacific Biosciences, 100-265-900). All library preparation and PacBio DNA sequencing were performed at University of Washington PacBio Sequencing Services. At all steps, DNA quantity was checked with fluorometry on the DS-11 FX instrument (DeNovix) with the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific) and sizes were examined on a 2100 Bioanalyzer (Agilent Technologies) using the High Sensitivity DNA Kit. SMRTbell sequencing libraries were prepared according to the protocol "Procedure & Checklist - Preparing SMRTbell®

Libraries using PacBio® Barcoded Universal Primers for Multiplexing Amplicons" and the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, 100-938-900) with barcoded adapters (Pacific Biosciences, 101-628-400).

After library preparation, the barcoded libraries were pooled by normalizing mass to the number of constructs contained in each pool. The final library was bound with Sequencing Primer v4 and Sequel II Polymerase v2.0 and sequenced on two SMRT Cells 8M using Sequencing Plate v2.0, diffusion loading, 1.5 hour pre-extension, and 30-hour movie time. Additional data were collected after treatment with SMRTbell Cleanup Kit v2 to remove imperfect and damaged templates, using Sequel Polymerase v2.2, adaptive loading with a target of 0.85, and a 1.3 hour pre-extension time. CCS consensus and demultiplexing were calculated using SMRT Link version 10.2 with default settings and reads that passed an estimated quality filter of ≥Q20 were selected as "HiFi" reads and used to map barcodes to variants.

"HiFi" PacBio reads were first subjected to a custom analysis pipeline, AssemblyByPacBio. Each consensus CCS sequence was aligned to the wildtype *F9* cDNA sequence using BWA-MEM v0.7.10-r789,[188] generating CIGAR and MD strings, which are then used to extract barcodes and the variable region containing *F9*. The output from AssemblyByPacBio was then passed through PacRAT,[189] which takes all CCS reads containing the same barcode and performs multiple sequence alignment to improve variant calling. PacRAT was run with a variant agreement threshold of 0.6 and 3 independent CCS reads to call a barcode, resulting in 260,224 unique barcodes across all three sublibraries (**Appendix A: Figure A.**7**a-b**). A custom R script was then

used to parse the PacRAT output and generate a final barcode-variant map, 8,532 of the 8,740 possible missense variants (**Appendix A: Figure A.**7**c**).

2.4.7 General cell culture conditions

HEK-293T cells (ATCC CRL-3216) were grown at 37 °C and 5% CO_2 in Dulbecco's modified Eagle's medium (ThermoFisher Scientific) supplemented with 10% fetal bovine serum (ThermoFisher Scientific), 100 U/mL penicillin, and 100 ng/mL streptomycin (ThermoFisher Scientific). Cells were passaged every 2-3 days by detachment with 0.05% trypsin-EDTA (ThermoFisher Scientific).

Freestyle 293-F cells were grown in Freestyle 293 Expression Medium (ThermoFisher Scientific) at 37 °C and 8% CO_2 while shaking at 135 rpm. Cells were regularly counted using trypan blue staining on a Countess II FL automatic hemocytometer (ThermoFisher Scientific) and passaged by dilution to $3x10^5$ cells/mL once reaching a concentration between $1x10^6$ and $2x10^6$ cells/mL, unless otherwise stated. All Freestyle 293-F cells containing a landing pad were induced with 2 µg/mL doxycycline (Sigma-Aldrich).

2.4.8 Lentiviral transduction to generate suspension Freestyle 293-F landing pad line

To generate a landing pad lentiviral vector, $2.5x10^5$ cells HEK293T cells were passaged into 6 well plates and transfected with 500 ng pMD-VSVg (Addgene #12259), 1,750 ng psPax2 (Addgene #12260), and 1,750 ng landing pad G384A vector template using 6 µL Fugene 6 (Promega) following the manufacturer's protocol. The next day,

media was exchanged, and the supernatant was collected every 12 hours for the following 72 hours to harvest lentivirus. The supernatant was then centrifuged at 300 x *g* for 10 minutes and passed through a 0.45 µm filter, before being aliquoted, flash frozen in liquid nitrogen, and stored at -80 °C.[8]

$1x10^7$ Freestyle 293-F cells were plated in 20 mL media and then incubated with varying volumes of lentivirus-containing supernatant (1 mL to 1 µL). 24 hours later, media was removed, and cells were washed once before replating into 30 mL media. On day 4 post-transduction, 2 µg/mL doxycycline was added to the cells, which were then grown for 10 more days with regular passaging.

Cells were then washed with PBS + 1% bovine serum albumin (BSA, Sigma-Aldrich) before assessing mTagBFP2 fluorescence from the landing pad on an LSR II (BD Biosciences). Only samples with a multiplicity of infection (MOI) < 1 were then sorted by FACS on an Aria III for mTagBFP2$^+$ cells. A total of 17,114 BFP$^+$ cells were replated into a half deep 96 well plate (Applikon Biotechnology) and allowed to expand with the addition of 2 µg/mL doxycycline and 100 µg/mL blasticidin (Invivogen) to select for functional landing pad cells. Single landing pad integration was confirmed by co-transfection of EGFP and mCherry recombination vectors.

2.4.9 FACS parameters

For all experiments, the following settings and gates were used to sort individual cells. Live, cells were first identified using FSC-A vs SSC-A. Live cells were then gated for single cells using two sequential gates–the first: FSC-A vs FSC-H, and the second: SSC-A vs SSC-H. mTagBFP2 expression from the unrecombined landing pad was

excited using the 405 nm laser and captured on 450/50 nm bandpass filter. mCherry expression from the recombined landing pad was excited using the 561 nm laser and captured using a 595 nm (LSR II) or 600 nm (Aria III) long pass and 610/20 nm bandpass filters. EGFP or Alexa488 antibody expression was excited using the 488 nm laser and captured using a 505 nm long pass and 530/30 nm bandpass filters. Alexa647 antibody expression was excited using a 637 nm (LSR II) or 640 nm laser (Aria III) and captured using a 750 nm long pass and 780/60 bandpass filter (LSR II) or a 670/30 nm bandpass filter (Aria III). All flow cytometry data were collected with FACSDiva v.8.0.1 and analyzed using FlowJo v.10.7.1.

2.4.10 Recombination of Freestyle 293-F cells

Freestyle 293-F cells were transfected at $1x10^6$ cells/mL with 293Fectin (ThermoFisher Scientific) following the manufacturer's protocol with the following alterations. Briefly, for every 1 mL of cells transfected, 2 µL 293Fectin was mixed with 31.5 µL OPTI-MEM in one tube, and 1 µg of total plasmid DNA was added to OPTI-MEM for a final volume of 33.5 µL in a second tube. For recombination experiments, the 1 µg of total DNA was split in a 1:15 ratio of pCAG-NLS-Bxb1 (Addgene #51271) and recombination vector. After 5 minutes at room temperature, the two tubes were mixed together by gentle pipetting and incubated at room temperature for 20 minutes, before being added to cells. For single variants or controls, $1x10^7$ cells in 10 mL were transfected. For libraries, $3x10^7$ cells in 30 mL of cells were transfected.

48 hours after transfection, cells were split 1:9 into two separate flasks with 2 µg/mL doxycycline. The second flask containing 9 parts was additionally treated with 10

nM rimiducid to selectively kill off unrecombined cells. Two days after rimiducid treatment, cells were counted using trypan blue exclusion, and then live cells were separated from dead cells using Histopaque-1077 (Sigma-Aldrich). Cells were diluted to 35 mL total volume and then layered slowly on top of 15 mL Histopaque-1077 in a 50 mL conical. The cells were then centrifuged at 400 x *g* for 30 minutes with no acceleration and no break. The top 30 mL of media was aspirated, and the cells at the interface between Histopaque-1077 and media were removed and resuspended in 30 mL final volume of media. Cells were centrifuged again at 300 x *g* for 5 minutes, the media was removed, and the pellet was resuspended in 30 mL of fresh media and 2 µg/mL doxycycline. Cells were counted using trypan blue to determine yield. Approximately 1 week after transfection, cells were assayed on an LSR II or Symphony A3 for mCherry fluorescence to determine recombination rate and selection efficiency. Cells were maintained with 2 µg/mL doxycycline throughout.

2.4.11 Antibody staining for surface-displayed proteins

For all antibodies except GMA-001 and ab3570, cold PBS + 1% BSA was used as a staining buffer. Because proper folding of the γ-carboxylated GLA domain is dependent on calcium, a 1:10 dilution of cold PBS + Ca/Mg + 1% BSA (ThermoFisher Scientific) into PBS + 1% BSA was used as a staining buffer. See **Appendix A: Table A.**1 for primary and secondary antibody concentrations.

Cells were plated at $3x10^5$ cells/mL in either 10 mL (single variants and controls) or 30 mL (libraries) of Freestyle media with the addition of 50 nM vitamin K_1 (Sigma-Aldrich). For experiments involving GMA-001 or ab3570, an additional control flask of

wildtype FIX cells was plated and supplemented with 100 nM warfarin (Sigma-Aldrich) to reduce γ-carboxylation of the FIX GLA domain. After 24 hours, cells were induced with 2 μg/mL doxycycline and grown for an additional 48 hours.

On the day of staining, flasks of cells were split into equal 4 mL volumes (6 per 30 mL flask, 1 per 10 mL single variant or control) and spun at 300 x *g* for 3 minutes. Media was aspirated, and 3 washes of 1 mL cold staining buffer were performed, with 300 x *g* spins and supernatant aspiration between. Cells were then resuspended in 100 μL of diluted primary antibodies and incubated at room temperature for 30 minutes, with vortexing at 10 minute intervals. After primary antibody staining, cells were diluted with 1 mL staining buffer and spun at 300 x *g* for 3 minutes. This washing step was repeated twice more before secondary antibody staining. For secondary antibody staining, cells were resuspended in 100 μL of diluted secondary antibodies and incubated at room temperature for 30 minutes in the dark, with vortexing at 10 minute intervals. After secondary antibody staining, cells were again diluted with 1 mL cold staining buffer and spun at 300 x *g* for 3 minutes. This washing step was repeated twice more before a final resuspension in 500 μL staining buffer. At this point, all identical tubes were pooled.

For experiments not requiring FACS, cells were analyzed by flow cytometry for antibody fluorescence on either the LSR II or Symphony A3 cytometers. For library sorts, FACS was performed on an Aria III, dividing the library into four approximately equally sized quartile bins based on the ratio of fluorescent antibody to mCherry fluorescence. At least 2 million cells were sorted into each of the four quartile bins.

After sorting, each sorted bin of library cells was spun down at 300 x *g* for 5 minutes and then resuspended in 10 mL of Freestyle 293 Expression media and

supplemented with 100 U/mL penicillin and 100 ng/mL streptomycin. Cells were counted daily for 4-6 days, diluting to 20 mL final volume once the concentration was above $5x10^5$ cells/mL. Cells were harvested once each bin contained at least 20 million cells. To harvest, cells were spun down at 300 x *g* for 10 minutes, the supernatant was aspirated, and the pellet was flash frozen in liquid nitrogen before storage at -20 °C.

2.4.12 Genomic DNA prep, barcode amplification, and sequencing

Genomic DNA was prepared according to previously described protocols.[4–6] Briefly, genomic DNA was extracted from harvested cell pellets using the DNEasy Blood and Tissue kit (Qiagen) according to the manufacturers' protocol with the addition of a 30 minute RNase digestion at 56 °C during the resuspension step. Six DNEasy columns were used per cell pellet.

Two technical PCR replicates were performed on each bin for each sample to assess the variability in barcode counts from sequencing and amplification (**Appendix A: Figure A.**8). The protocol in **Section 2.4.5** was used to amplify each sample with the following changes: 1) For each replicate, 8 identical 50 µL first-round PCRs were prepared, and each PCR contained 2,500 ng genomic DNA, 25 µL Q5 polymerase, and 0.5 µM of NP0492 and NP0546 primers. 2) During Ampure XP cleanup, samples were eluted into 21 µL water, and 40% (8 µL) was used in the second-round PCRs. 3) During the second-round PCRs, samples were run for a maximum of 20 cycles or until all samples were above 3,000 relative fluorescence units. 4) Samples were either sequenced on a NextSeq 550 using a NextSeq 500/550 High Output v2.5 75 cycle kit or on a NextSeq 2000 using a NextSeq 1000/2000 P3 50 cycles kit (Illumina).

2.4.13 Calculating antibody scores

Using a custom script, barcode sequencing reads were first converted to FASTQ format and demultiplexed using bcl2fastq v2.20 (Illumina). Next, forward and reverse barcode reads were paired using PEAR v0.9.11,[187] and unique barcodes were counted. To reduce error associated with counting noise, we removed low frequency variants (**Appendix A: Figure A.**1) using a frequency threshold of 10^{-6}. Variants that were observed in fewer than two replicates were also removed. Sequenced barcodes were then assigned to FIX variants using the barcode-variant map generated from PacBio sequencing. Any barcode associated with insertions, deletions, or multiple amino acid substitutions in FIX were removed from the analysis.

Antibody scores were calculated using a modified VAMP-seq pipeline.[4] Briefly, for each experiment, a weighted average of every variant's frequency in each bin was calculated using the following weights: w_{bin1}: 0.25, w_{bin2}: 0.5, w_{bin3}: 0.75, w_{bin4}: 1. The weighted average for each variant was then normalized such that the median score for wildtype FIX barcodes was 1 and the median score for the lowest 5th percentile of missense variants was 0. The final antibody score for each variant was then averaged across all replicates.

2.4.14 Description of computational methods

VAMP-seq abundance scores for PTEN, TPMT, VKOR, CYP2C9, and NUDT15 were downloaded from MAVEdb.[190] Signal peptide predictions were generated by submitting variant and wildtype FIX sequences to the SignalP 6.0 server.[178]

Hydrophobicity of the h-region of FIX signal peptide variants was calculated using GRAVY. Solvent accessible surface area (SASA) was calculated from the AlphaFold wildtype FIX model using FreeSASA.[191] Relative solvent accessibility was calculated by dividing the output from FreeSASA by the maximum SASA values for wildtype residues.[192]

Hierarchical clustering was performed in R using the hclust package. Ten positions without variant data in any assays were removed from the analysis. Positions were converted into a Euclidean distance matrix based on variant scores. Ward's clustering algorithm was then applied to the distance matrix using the square of the distance between clusters.[193]

2.4.15 Clinical variant curation

To capture true pathogenic and benign FIX missense variants, we collected data from ClinVar, MLOF, and gnomAD v.3.1. ClinVar classifications were removed if there were conflicting classifications, or if evidence was not included. As the incidence of hemophilia B is approximately 3.8 per 100,000 male births,[64,65] we included any gnomAD FIX variants with minor allele frequency in hemizygotes of greater than 1 per 1,000 as benign variants.[4,194] Lastly, FIX variants observed in patients with hemophilia A from MLOF were considered benign if clinical testing of patient blood demonstrated wildtype-like FIX activity. The final curated dataset was then constructed and labeled using a consensus from all sources. The final curated dataset contained 140 pathogenic or likely pathogenic variants and 14 benign or likely benign variants, sufficient for classification model training purposes.[182,195]

Patient-level data on FIX antigen, activity, and disease severity were collected from EAHAD and collapsed based on variant.[77] Unpublished results were removed from the dataset, resulting in 635 final variants with at least one of the three clinical datapoints. Because FIX antigen and activity are often reported as "< 1%" or "undetectable" when below the limit of detection in the clinical assay, we altered these values to 0.01% to perform linear regression. Values for FIX antigen and FIX activity were averaged across patient samples. For disease severity classifications, a consensus-based approach was used to determine final disease severity. If there was a tie between two or more severities, a conflicting classification was used.

2.4.16 Random forest classifier

We then built a random forest classifier to distinguish true benign and pathogenic variants using the ranger and tidymodels packages in R. First, benign/likely benign and pathogenic/likely pathogenic classes were combined. We then split the curated variants so 75% of the data was used for training and 25% was used for testing the performance of the final model. To account for unbalanced class sizes, we implemented random oversampling (ROSE)[181] on the training set. Model hyperparameters were selected by using 5-fold cross-validation on the training set. Model performance was evaluated using the ROC curves generated with the test set. The final model was then applied to 107 VUS from MLOF that were observed in our experimental data. We used a Bayesian adaptation of ACMG rules-based guidelines to reinterpret variant classifications.[12,182,195] All other evidence codes were applied according to ACMG and ClinGen variant curation expert panel (VCEP) guidelines.[59]

3 Suppression of unwanted CRISPR-Cas9 editing by co-administration of catalytically inactivating truncated guide RNAs

A version of this chapter has been previously published as:

Rose, J.C., [§] Popp, N.A., [§] Richardson, C.D., Stephany, J.J., Mathieu, J., Wei, C.T., Corn., J.E., Maly, D.J., & Fowler, D.M. Suppression of unwanted CRISPR/Cas9 editing by co-administration of catalytically inactivating truncated guide RNAs. Nat. Comm. 11, 2697 (2020).

[§]These authors contributed equally to this work.

3.1 Introduction

The *S. pyogenes* Cas9 (spCas9) nuclease is targeted to specific sites in the genome by a single guide RNA (sgRNA) containing a 20-nucleotide target recognition sequence. The target site must also contain an NGG protospacer adjacent motif (PAM).[132] This multipartite target recognition system is imperfect, and most sgRNAs direct significant cleavage and subsequent unwanted editing at off-target sites whose sequence is similar to the target site.[114–117] Numerous approaches to reduce off-target editing have been devised, yet are hampered by various limitations.[126,143–153] For example, spCas9 variants with improved specificity have been engineered.[154–156] While useful, these high-specificity variants often decrease on-target editing[157,158] and, in most cases, do not eliminate all unwanted editing.[156] All high-specificity Cas9 variants appear to balance on- vs off-target activity via the same mechanism[156,159] and, as a consequence, often fail to suppress editing at the same obstinate off-target sites.[156,158]

Thus, new methods for off-target suppression are needed, particularly ones that preserve on-target editing, can be combined with high-specificity Cas9 variants, and require minimal expenditure of time, effort, and resources. To this end, we developed an orthogonal and general approach for suppressing off-targets that can be readily combined with existing methods, including high-specificity variants.

Our off-target suppression approach is based on the observation that sgRNAs with target sequences 16 or fewer bases in length direct Cas9 binding to DNA target sites but do not promote cleavage.[196–198] Here, we show that Cas9 bound to dRNAs with perfect complementarity to off-target sites can dramatically improve editing specificity by shielding these sites from the active Cas9-sgRNA complex (**Figure** 3.1**a**). To highlight the generality and ease of implementation of our method, which we call dRNA Off-Target Suppression (dOTS), we effectively suppress editing at 15 off-target sites, yielding up to ~40-fold increase in specificity, with minimal dRNA optimization. Furthermore, dOTS can be multiplexed to suppress several off-targets simultaneously and can be combined with other approaches for improving specificity. We also describe dRNA ReCutting Suppression (dReCS), wherein dRNAs prevent recutting of homology-directed repair (HDR)-corrected sites, eliminating the need for blocking mutations and facilitating scarless editing. Thus, we enable more precise genome editing by establishing a facile and flexible approach for suppressing unwanted editing of both off-target and HDR-corrected sites.

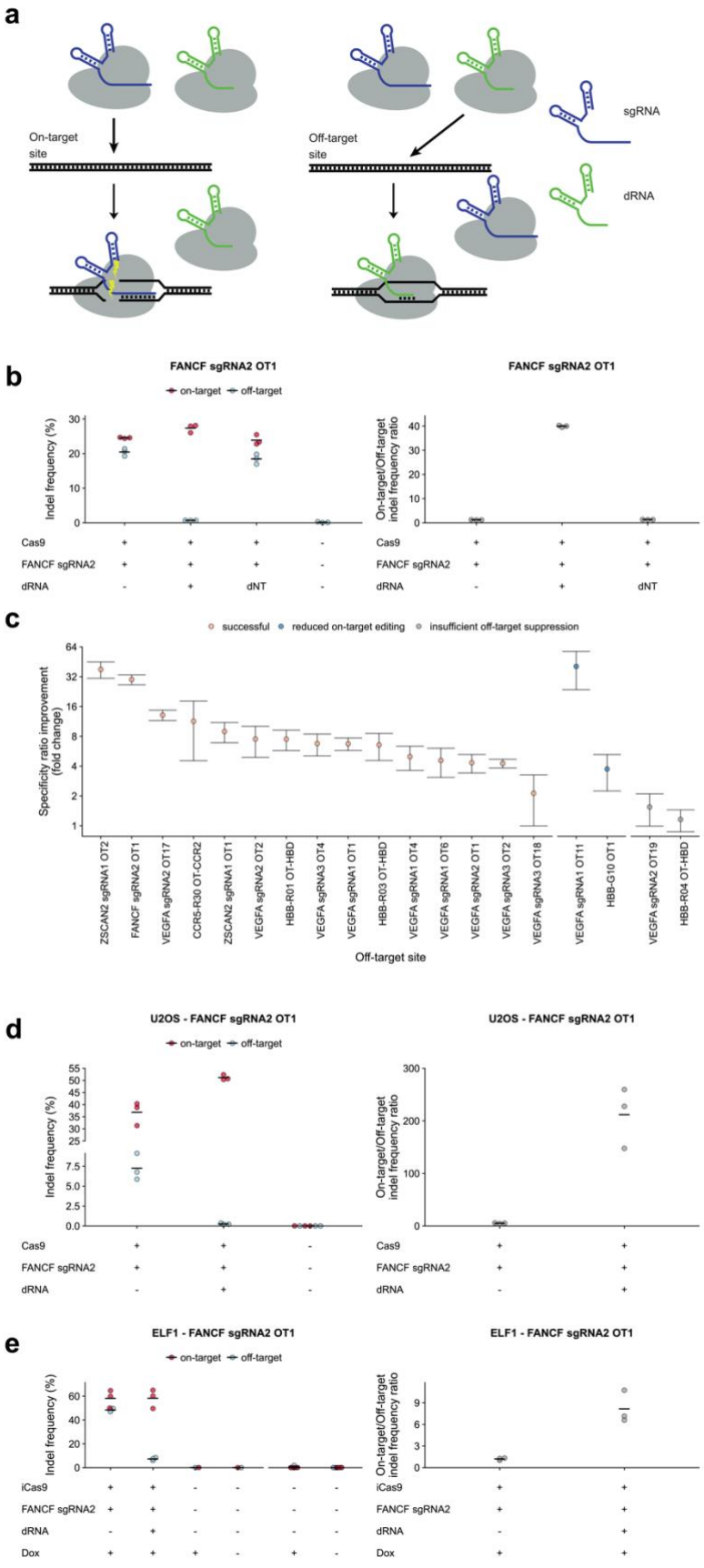
a
On-target site
Off-target site
sgRNA
dRNA
b
FANCF sgRNA2 OT1
on-target
off-target
Indel frequency (%)
On-target/Off-target indel frequency ratio
Cas9
FANCF sgRNA2
dRNA
dNT
c
successful
reduced on-target editing
insufficient off-target suppression
Specificity ratio improvement (fold change)
ZSCAN2 sgRNA1 OT2
FANCF sgRNA2 OT1
VEGFA sgRNA2 OT17
CCR5-R30 OT-CCR2
ZSCAN2 sgRNA1 OT1
VEGFA sgRNA2 OT2
HBB-R01 OT-HBD
VEGFA sgRNA3 OT4
VEGFA sgRNA1 OT1
HBB-R03 OT-HBD
VEGFA sgRNA1 OT4
VEGFA sgRNA1 OT6
VEGFA sgRNA2 OT1
VEGFA sgRNA3 OT2
VEGFA sgRNA3 OT18
VEGFA sgRNA1 OT11
HBB-G10 OT1
VEGFA sgRNA2 OT19
HBB-R04 OT-HBD
Off-target site
d
U2OS - FANCF sgRNA2 OT1
e
ELF1 - FANCF sgRNA2 OT1
iCas9
Dox

Figure 3.1: dRNA-mediated Off-Target Suppression (dOTS) effectively reduces off-target editing

a. Schematic representation of dOTS. A dRNA (green) with perfect complementarity for an off-target site directs Cas9 binding but not cleavage, protecting the site. **b.** Indel frequencies and specificity ratios (on-target/off-target indel frequency ratios) at the *FANCF*-sgRNA2 on-target site and OT1 24 h after transfection of HEK-293T cells with Cas9, sgRNA, and *FANCF*-sgRNA2 OT1 dRNA1 or a nontargeting control dRNA (dNT) that does not target genomic DNA. For conditions without dRNA, an equivalent amount of pMAX-GFP was substituted. Means of n = 3 biological replicates depicted by solid lines. **c.** Normalized specificity ratios, computed as the specificity ratio of the best dRNA condition divided by the specificity ratio of the sgRNA only condition for 19 guide/off-target pairs tested in HEK-293T cells. Points depict the mean of n = 3 biological replicates, error bars show the standard error of the mean. **d.** Indel frequencies and specificity ratios at the *FANCF*-sgRNA2 on-target site and OT1 24 h after transfection in U2OS cells, and **e.** Elf1 embryonic stem cells. Control samples to the right of the x-axis break were performed separately. iCas9 denotes stable integration of Cas9 under the control of a doxycycline-inducible promoter. Means of n = 3 cell culture replicates depicted by solid lines. OT off-target.

3.2 Results

3.2.1 Dead RNA off-target suppression increases specificity

We first determined the feasibility of using dRNAs to suppress unwanted editing at off-target site 1 (OT1) of an sgRNA (sgRNA2) targeting the *FANCF* locus.[154] We co-transfected HEK-293T cells with a plasmid encoding spCas9, along with equal amounts of plasmids encoding *FANCF*-sgRNA2 and a GFP control, or *FANCF*-sgRNA2 and one of four dRNAs with perfect complementarity to OT1 (**Appendix B: Figure B.**1**a**). Three of the four dRNAs significantly decreased off-target editing without appreciably impacting on-target editing (**Appendix B: Figure B.**1**b**). In particular, dRNA1 decreased

off-target editing from 20.44% (s.e.m. = 0.61%, n = 3) to 0.69% (s.e.m. = 0.02%, n = 3), leading to a 30-fold increase in the on-target specificity ratio (**Figure** 3.1**b**). Cas9-dRNA complexes are thought to lack cleavage activity, but a relatively small number of dRNAs have been evaluated so far.[196,197] Thus, we verified that dRNA1 did not direct any detectable Cas9 editing activity at either the on- or off-target sites (**Appendix B: Figure B.**1**c**). We further confirmed that dRNA1 showed no cleavage genome-wide using genome-wide unbiased identification of DSBs enabled by sequencing (GUIDE-seq),[117] and that it directed selective reduction of only OT1 (**Appendix B: Figure B.**2). To our knowledge, this experiment is the first to demonstrate that a dRNA leads to no detectable cleavage activity anywhere in the genome.

To demonstrate the generalizability of dOTS, we evaluated 18 additional on-target/off-target pairs in HEK-293T cells. We found at least one dRNA for 15 of the 19 pairs we tested that increased the specificity ratio by at least two-fold (mean fold-change = 10.44) while decreasing on-target editing by no more than two-fold (mean fold-change = 0.93; **Figure** 3.1**c**; **Appendix B: Figure B.**3). Across all on-target/off-target pairs, a median of six candidate dRNAs were screened, highlighting the ease of identifying effective dRNAs (**Appendix B: Figure B.**3**-4; Table B.**1). In most cases, nontargeting dRNAs had little to no impact on editing (**Appendix B: Figure B.**5). Moreover, effective dRNAs did not induce indels at either on- or off-target sites, suggesting that few, if any, Cas9-dRNA complexes are active (**Appendix B: Table B.**2**-3**). dOTS was also effective in U2OS cells and the Elf1 naïve embryonic stem cell line (**Figure** 3.1**d-e**; **Appendix B: Figure B.**6).[199] Finally, we found that dRNA-mediated

suppression of off-target editing was durable, with dRNAs effectively decreasing off-target editing for at least 72 hours post-transfection (**Appendix B: Figure B.**7).

An important application of Cas9 is editing genes containing pathogenic mutations.[200,201] For example, Cas9 has been used to target the *β*-globin locus (*HBB*) with the goal of curing sickle cell disease.[121,202] However, the *δ*-globin locus (*HBD*) is a common off-target for sgRNAs targeting *HBB* and cleavage of both on- and off-target sites can result in large chromosomal deletions at the globin locus.[137] In HEK-293T cells, dOTS decreased off-target editing at *HBD* from 1.08% (s.e.m. = 0.22%, n = 3) to 0.15% (s.e.m. = 0.03, n = 3; **Appendix B: Figure B.**3**d**). In Elf1 cells, dOTS decreased off-target editing at *HBD* from 20.72% (s.e.m. = 2.75%, n = 3) to 1.20% (s.e.m. = 0.18, n = 3), increasing the specificity ratio from 1.33 to 13.72 (**Appendix B: Figure B.**6**b**). Thus, dOTS can control unwanted editing at clinically relevant loci.

We were unable to find effective dRNAs for four off-target sites. In two cases, dRNAs strongly reduced off-target editing but also decreased on-target editing by greater than two-fold (**Figure** 3.1**c**; **Appendix B: Figure B.**3**b,i**). In two other cases, no dRNA we tested was effective in decreasing off-target editing (**Figure** 3.1**c**; **Appendix B: Figure B.**3**e,m-n**). We suspect that these ineffective dRNAs are either unstable, form unfavorable secondary structures, or have insufficient affinity for the off-target site relative to their cognate sgRNAs. However, at most off-targets we identified one or more effective dRNAs that enhanced specificity without sacrificing on-target editing, making dOTS an effective approach for off-target suppression.

3.2.2 Mechanism of off-target suppression by dRNAs

dOTS is based on our prediction that Cas9-dRNA complexes with perfect complementarity to an off-target site can directly outcompete active, imperfectly complementary Cas9-sgRNA complexes for binding. To test this Cas9 self-competition mechanism, we performed *in vitro* cleavage assays with linear DNA substrates and purified Cas9 ribonucleoprotein complexes (RNPs) containing either *FANCF*-sgRNA2 or dRNA1. Incubation of a substrate containing the *FANCF* OT1 site with a mixture of the Cas9-dRNA1 and Cas9-sgRNA2 complexes led to a robust reduction in cleavage compared to administration of the Cas9-sgRNA2 complex alone (**Figure** 3.2**a**). Consistent with our self-competition mechanism, preincubation of the substrate with the Cas9-sgRNA2 complex followed by addition of the Cas9-dRNA1 complex eliminated the reduction in cleavage (**Appendix B: Figure B.**8). Thus, Cas9-dRNA complexes can directly shield off-target loci from Cas9-sgRNA cleavage.

At low concentrations of Cas9-sgRNA2, Cas9-dRNA1 modestly reduced cleavage of the on-target *FANCF* substrate site *in vitro* (**Figure** 3.2**b**), despite this dRNA not affecting on-target editing efficiency in cells (**Figure** 3.1**b,d-e**). One possible explanation for this disparity is that, in cells, Cas9-dRNA1 mediated protection of the on-target locus decreases the rate of indel formation but editing reaches the same maximum as in cells without dRNA1 by the time of measurement. Another explanation is that cellular factors prevent Cas9-dRNA1, which should have modest affinity for the on-target site, from providing appreciable protection from cleavage by Cas9-sgRNA2. Thus, we measured rates of indel formation at *FANCF*-sgRNA2 OT1 and the on-target site in cells using a chemically inducible Cas9 (ciCas9) variant.[126,203] The activity of

ciCas9 is repressed by an intramolecular autoinhibitory switch. Addition of a small molecule, A-1155463 (A115), disrupts autoinhibition and rapidly activates ciCas9, enabling precise studies of editing kinetics.

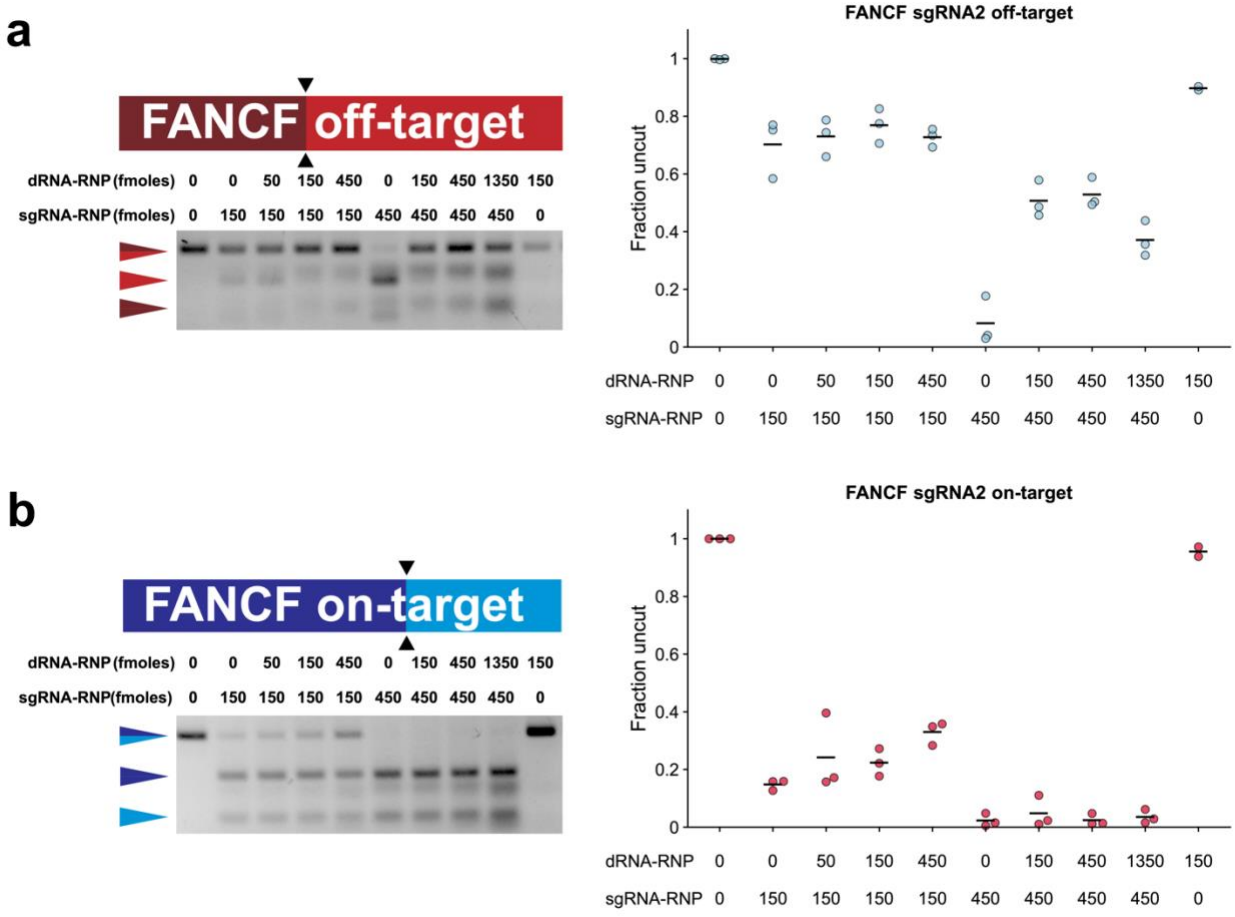

Figure 3.2: dRNAs suppress off-target editing by competing with sgRNAs for off-target sites

Representative gels of in vitro cleavage of PCR products containing either a *FANCF*-sgRNA2 OT1 or b the *FANCF*-sgRNA2 on-target site with either 150 or 450 fmoles of Cas9 *FANCF*-sgRNA2 RNP in the presence of variable amounts of the Cas9 *FANCF*-sgRNA2 OT1 dRNA1 complex. Fraction of uncut DNA determined by gel densitometry. Means of n = 3 replicates depicted by solid lines. For uncropped gels, see **Appendix B: Figure B.**14.

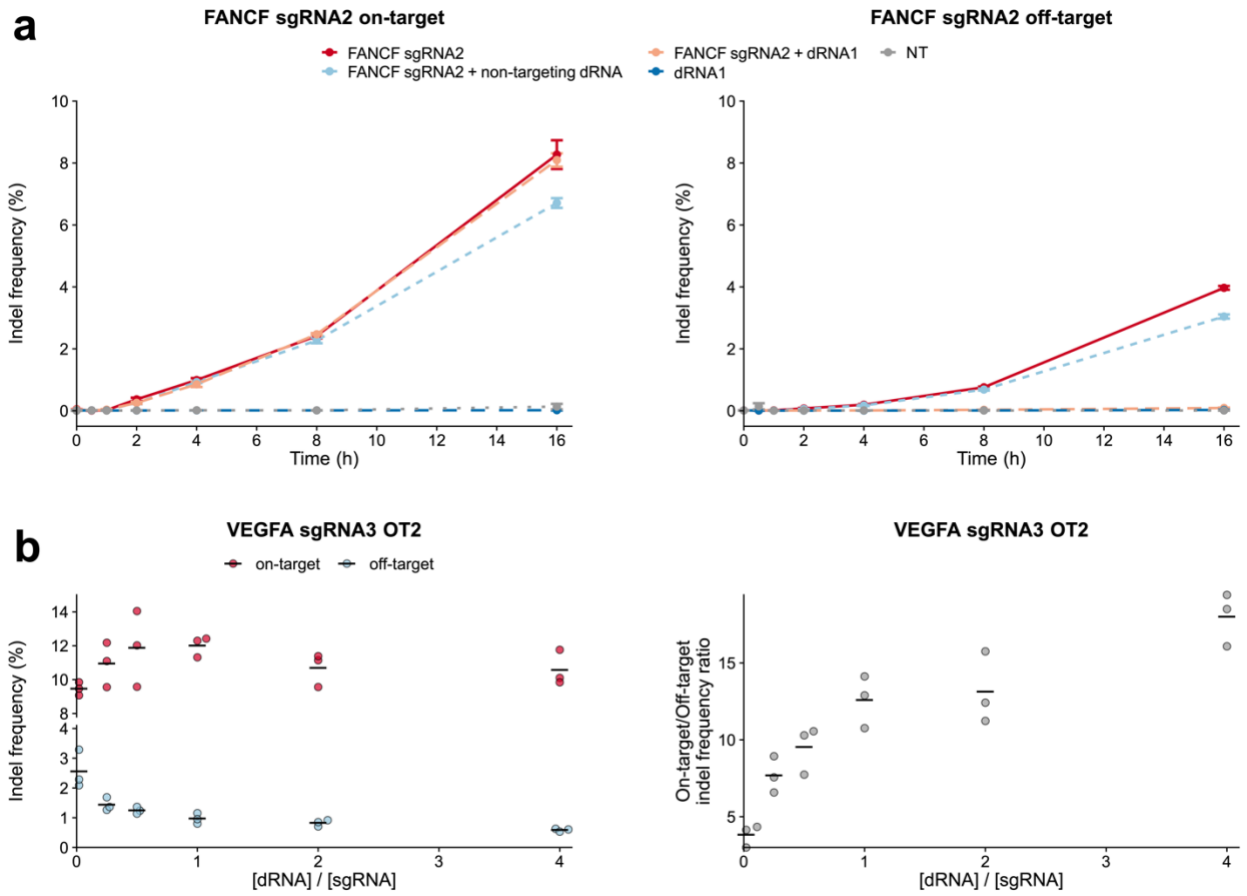

Figure 3.3: dRNAs affect off-target, but not on-target, editing kinetics and can be titrated to improve specificity

a. Editing of *FANCF*-sgRNA2 on-target and OT1 sites using chemically inducible Cas9 (ciCas9) from 0 to 16 h after activation with A115. Nontargeting dRNA is a 14-base control dRNA targeting a non-endogenous site. NT = non-transfected control. Points depict the mean of n = 3 biological replicates. Error bars show the standard error of the mean. **b.** Indel frequencies and specificity ratios at *VEGFA* sgRNA3 on-target and OT2 sites in cells transfected with plasmids encoding Cas9 and varying ratios of *VEGFA* sgRNA3 and dRNA2. dRNA untreated cells were transfected with Cas9 and a 1:1 *VEGFA* sgRNA3:GFP plasmid ratio. Error bars depict s.e.m. (n = 3 cell culture replicates). OT off-target.

As expected, activation of ciCas9 with A115 led to the rapid appearance of indels at the *FANCF*-sgRNA2 on- and off-target sites in the absence of dRNA1.

Inclusion of a plasmid encoding dRNA1 effectively eliminated ciCas9-mediated editing at the off-target site but had no measurable impact on the kinetics of on-target editing (**Figure** 3.3**a**). These results suggest that dRNAs with imperfect complementarity to an on-target site can bind to and protect that site in cell-free systems, but not in cells. The most likely explanation for this difference is that, in cells, DNA is subject to a variety of active processes that influence Cas9.[204,205] For example, the degree of complementarity between a guide and its target affects the ability of polymerases to displace dCas9 from DNA,[206] suggesting that polymerases may limit the ability of imperfectly complementary Cas9-dRNA complexes to shield on-target sites.

Our proposed Cas9 self-competition mechanism predicts that the level of off-target shielding provided by moderately effective dRNAs can be improved by manipulating the ratio of Cas9-dRNA to Cas9-sgRNA in cells. While an initial 1:1 plasmid ratio was effective for all 15 successful dRNAs, increasing the amount of dRNA relative to sgRNA further decreased off-target editing and improved the specificity ratio at each of the four sgRNA/dRNA pairs we tested (**Figure** 3.3**b**; **Appendix B: Figure B.**9). For one pair, higher dRNA:sgRNA ratios also decreased on-target editing. Thus, a trade-off between maintaining on-target editing and decreasing off-target editing exists for some sgRNA/dRNA pairs. Here the dRNA:sgRNA ratio can be tuned based on whether preserving on-target editing or suppression of a particular off-target is desired.

3.2.3 dOTS improves other approaches to increase Cas9 specificity

Other strategies to improve Cas9 specificity fail to completely suppress off-target editing and often reduce on-target efficacy. Thus, we wondered whether they could be enhanced with dOTS. One approach uses truncated sgRNAs (tru-sgRNAs) with 17-19 base target sequences to increase on-target specificity at some loci. For example, truncation of the *VEGFA* sgRNA3 target sequence (*VEGFA* tru-sgRNA3) decreases editing at some off-target sites, but editing at OT2 remains.[147] dOTS suppressed editing at this refractory off-target site without affecting on-target activity (**Figure** 3.4**a**), demonstrating that it is compatible with tru-sgRNAs.

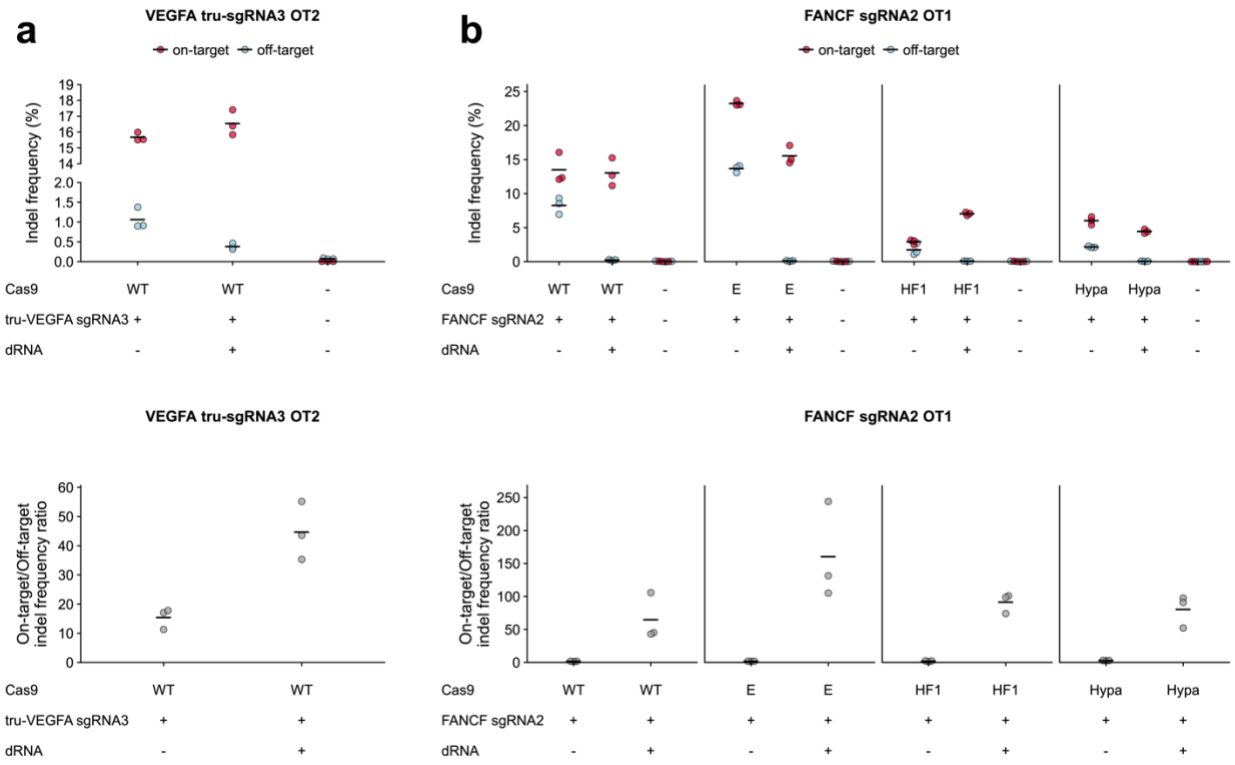

Figure 3.4: dRNAs can be combined with other approaches to improve specificity

Indel frequencies and specificity ratios 24 h after transfection with a plasmids encoding WT Cas9, **a.** dRNA targeting *VEGFA* sgRNA3 OT2 (dRNA2) and a truncated guide *VEGFA* tru-sgRNA3, or **b.** High-specificity variants of Cas9 and a dRNA targeting *FANCF*-sgRNA2 OT1 (dRNA1). Wild-type Cas9 (WT),

espCas9 (E), spCas9-HF1 (HF1), HypaCas9 (Hypa). Means of n = 3 cell culture replicates depicted by solid lines. OT off-target.

More recently, rational engineering of spCas9 has produced high-specificity variants like espCas9(1.1), spCas9-HF1, and Hypa-Cas9.[154–156] While these variants generally improve on-target specificity, they do not suppress unwanted editing at all off-target sites for all sgRNAs. For example, a recent evaluation of these three high-specificity variants revealed off-target editing by all three variants at four of the six sgRNAs tested.[156] In another example, *FANCF* sgRNA2 OT1 is still edited at high frequencies by all three high-specificity variants (**Figure** 3.4**b**).[154,156] Co-transfection of *FANCF*-sgRNA2 with an effective dRNA reduced off-target editing to levels indistinguishable from non-transfected controls for all high-specificity Cas9 variants ($p > 0.05$, one-sided *t*-test, n = 3), dramatically increasing specificity ratios (**Figure** 3.4**b**). dRNAs also effectively suppressed off-target editing by espCas9(1.1) and spCas9-HF1 at a refractory *VEGFA* sgRNA3 off-target (**Appendix B: Figure B.**10). High-specificity Cas9 variants are known to exhibit decreased on-target activity, which is sensitive to delivery method and other factors.[121,157,207] Indeed, in some cases, we observe a decrease in on-target editing when high-specificity Cas9 variants and dOTS are combined. However, this reduction in on-target editing is generally less pronounced that the efficiency loss observed comparing HypaCas9 or spCas9-HF1 to wildtype in the absence of dOTS. The reduction in on-target editing is also markedly less than the degree of suppression achieved by dOTS at the off-target site. Thus, dOTS can be combined with many other methods for improving Cas9 specificity.

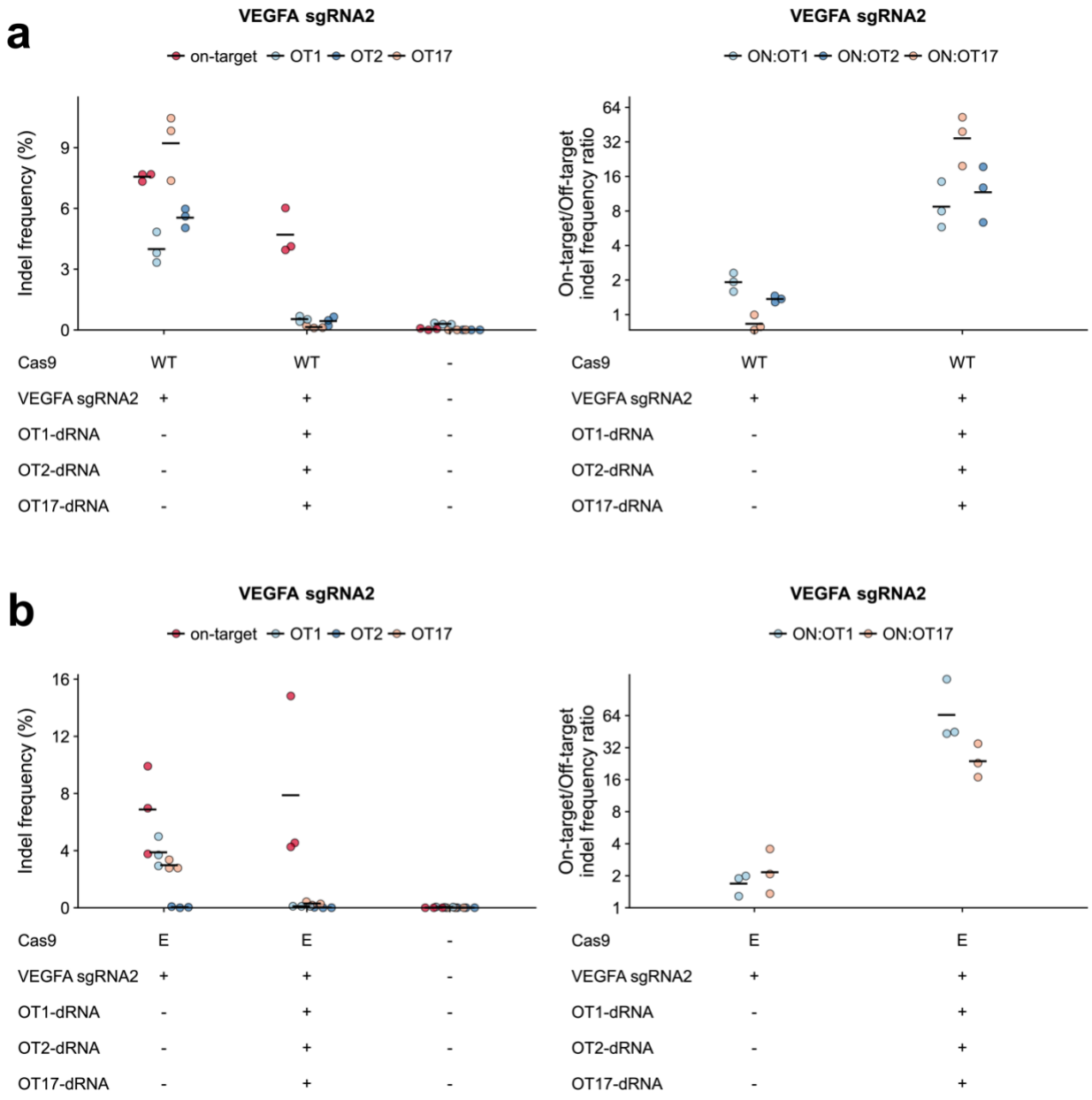

Figure 3.5: dRNAs can be multiplexed to suppress several off-target sites simultaneously

Indel frequencies and specificity ratios 24 h after transfection of plasmids encoding either **a.** wildtype (WT) or **b.** espCas9 (E), *VEGFA* sgRNA2, and dRNAs targeting one of three *VEGFA* sgRNA2 off-targets (OT1 dRNA1, OT2 dRNA8, OT17 dRNA8). Means of n = 3 cell culture replicates depicted by solid lines. OT off-target.

3.2.4 dOTS can be multiplexed to suppress multiple off-targets

Since many sgRNAs induce off-target editing at numerous sites,[116,117,208] we examined whether dOTS could be multiplexed. We selected three off-target sites for *VEGFA* sgRNA2 with individually effective dRNAs (**Figure** 3.1**c**; **Appendix B: Figure B.**3). HEK-293T cells were transfected with *VEGFA* sgRNA2 and the dRNAs individually, in duplex, or in triplex. Even when all three dRNAs were combined, editing at each off-target site was suppressed by its cognate dRNA with only small loses in on-target editing (**Figure** 3.5**a**; **Appendix B: Figure B.**11**a**). Multiplexed dOTS was also effective for two other sgRNAs (**Appendix B: Figure B.**11**b-c**) and could even suppress the off-targets of two distinct sgRNAs simultaneously (**Appendix B: Figure B.**11**d**). Notably, each dRNAs only impacted editing at its cognate off-target site, without increasing or decreasing editing at the other off-target sites of the sgRNA.

Like wildtype Cas9, high-specificity Cas9 variants can cause editing at multiple off-target sites. For example, espCas9(1.1) reportedly drives appreciable editing with *VEGFA* sgRNA2 at three different off-target sites.[156] We observed off-target editing at two of these sites and found that dRNAs could simultaneously decrease off-target editing at both sites without perturbing on-target editing (**Figure** 3.5**b**). Furthermore, multiplexed dOTS suppressed editing driven by spCas9-HF1 and HypaCas9 at these off-target sites (**Appendix B: Figure B.**12). Thus, in the context of both wildtype and variant Cas9, dRNAs can be combined to suppress multiple off-targets simultaneously.

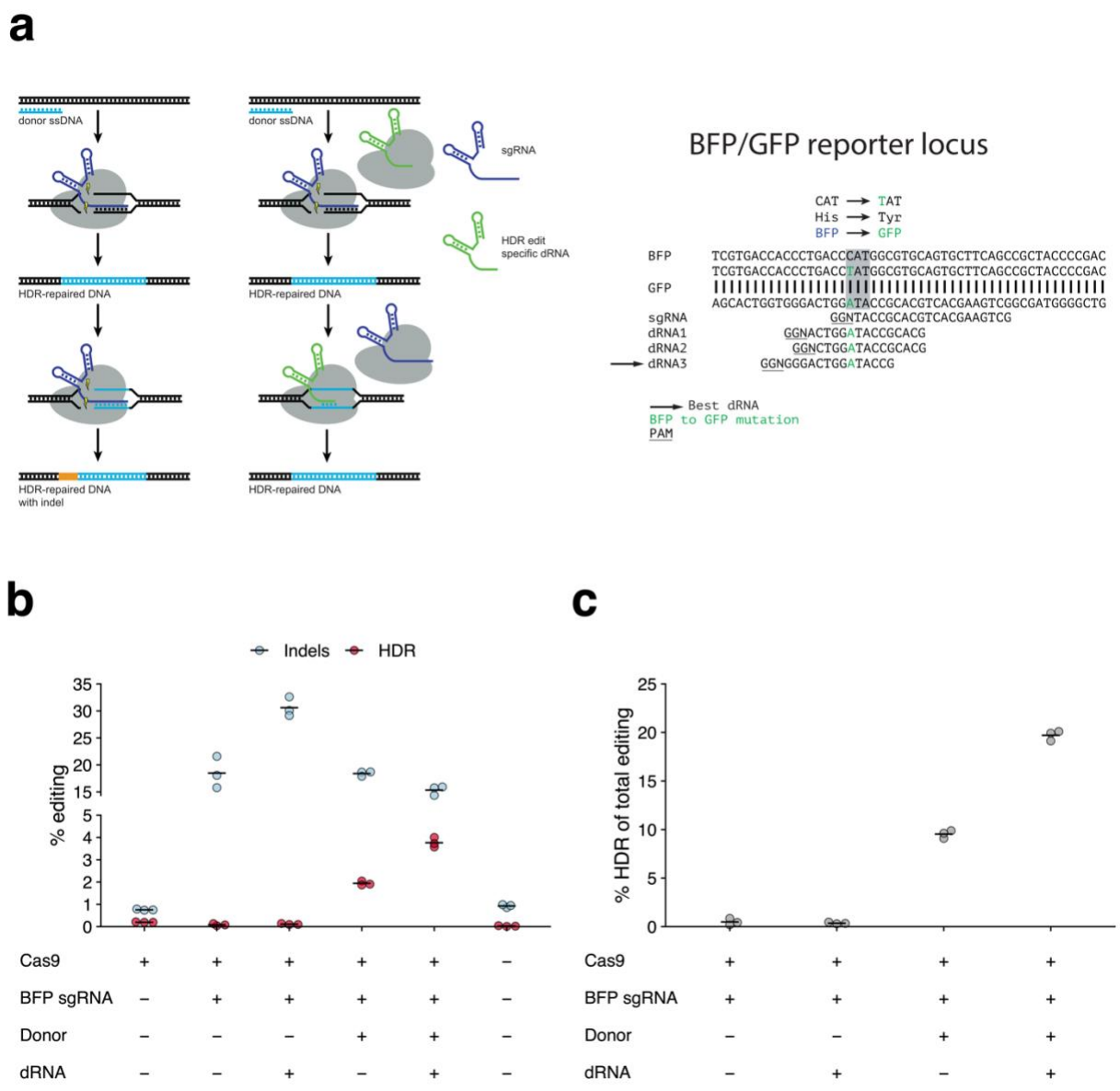

Figure 3.6: dRNAs facilitate scarless HDR

a. Schematic depicting dReCS and alignment of BFP, GFP, sgRNA, and dRNA sequences. dRNA (green) exhibiting perfect complementarity for the repaired site directs Cas9 binding but not cleavage, protecting the repaired site. Single base change to generate GFP from BFP is displayed in green with affected codon indicated by gray box. PAM sequences are underlined. Black arrow indicates best dRNA, as determined by maximal improvement in HDR yield. **b.** Indels and homology-directed repair (HDR) as assessed by flow cytometry, where indels lead to a loss of BFP signal, and HDR leads to a loss of BFP

and gain of GFP signal. **c.** HDR as a percentage of total Cas9 edits observed. Means of n = 3 cell culture replicates depicted by solid lines. dRNA: BFP sgRNA1 dRNA3 (see **Supplementary Fig. 13**).

3.2.5 dRNAs enable scarless HDR-mediated genome editing

When mutations introduced by HDR do not substantially disrupt the target sequence of Pam, as in generally the case for single nucleotide variants, Cas9 can continue to cleave the target site after repair. Continued cleavage introduces indels, substantially decreasing the frequency of loci containing the desired sequence. For example, quantification of editing outcomes at *PSEN1* revealed that up to 95% of HDR-corrected templates showed secondary indels due to recutting.[122] If a protein-coding region is being edited, synonymous blocking mutations that disrupt the sgRNA target sequence, PAM, or both are generally included in the repair template. Unfortunately, synonymous blocking mutations may alter protein expression or interfere with mRNA splicing. Furthermore, predicting functionally neutral blocking mutations in non-coding regions is extremely challenging. Base editing in some cases can make single base changes, yet its use is hindered by unwanted bystander editing within the editing window, off-target editing of RNA, and an inability to install transversion mutations, insertions, or deletions.[209–211] Thus, scarless editing, the ability to efficiently introduce single nucleotide variants and other small changes into the genome via HDR without blocking mutations or unwanted indels, would be of tremendous utility.

We predicted that dRNAs directed at a desired, HDR-corrected sequence could shield repaired sites from recutting, an approach we call dRNA ReCutting Suppression (dReCS) (**Figure** 3.6**a**). We evaluated the ability of dRNAs to improve the HDR-mediated conversion of BFP to GFP through substitution of a single amino acid.

Previously, several blocking mutations were used to prevent recutting, yet only a single nucleotide change is needed to alter the His in BFP (**C**AT) to the Tyr in GFP (**T**AT).[212] We selected a previously used sgRNA in which the permissive site within the PAM (i.e., N in NGG) for the BFP sgRNA corresponds to the mutated nucleotide. Thus, this sgRNA possesses perfect complementarity to both the native and HDR-repaired locus, representing a worst-case scenario in which Cas9-sgRNA is expected to efficiently recut HDR-repaired sites. HEK-293T cells with stably integrated BFP were transfected with a single stranded oligodeoxynucleotide (ssODN) donor template containing the single nucleotide change, the sgRNA targeting BFP, and one of three dRNAs with perfect complementarity to the GFP but not BFP sequence. After 4 days, in the absence of dRNA, scarless HDR conversion to GFP was inefficient, with 1.94% of cells expressing GFP by flow cytometry. In the presence of the best dRNA, absolute HDR efficiency increased to 3.77% (**Figure** 3.6**b**; **Appendix B: Figure B.**13), corresponding to an increase in the percentage of all edited sites exhibiting scarless HDR from 9.53% (s.e.m. = 0.40, n = 3) to 19.72% (s.e.m. = 0.52, n = 3; **Figure** 3.6**c**). Thus, dReCS can promote scarless HDR even when the sgRNA has perfect complementarity for the HDR-corrected sequence.

3.3 Discussion

Here, we describe a general approach for the targeted suppression of unwanted Cas9-mediated editing that relies on co-administration of dRNAs with perfect complementarity to the suppressed site. Our approach exploits the previously unappreciated phenomenon we refer to as Cas9 self-competition: the ability of different

Cas9-guide RNA complexes to compete for a limited number of genomic target sites. We show that catalytically inactive Cas9, in this case Cas9 bound to a dRNA, can protect sites from undesired cleavage by active Cas9-sgRNA complexes. One application of this approach, dOTS, reduced editing at 15 distinct off-target sites, in some cases below the limit of detection by high-throughput sequencing. Another application, dReCS, facilitated scarless introduction of a single base change that did not impact the PAM or target sequence. dReCS circumvents the need for blocking mutations, making it particularly useful for single nucleotide variants and small indels in non-coding regions of the genome where synonymous blocking mutations are not an option. In both cases, effective dRNAs can generally be rapidly identified with minimal screening. Moreover, dRNAs are effective in a variety of different cells lines and they can be combined to protect multiple off-target sites simultaneously.

dOTS and dReCS offer many advantages, but they are not perfect. We could not find an effective dRNA for four of the 19 target/off-target pairs we tested. In some cases, additional dRNAs could be screened, but the sequence restrictions imposed by the spCas9 NGG PAM mean that effective dRNAs may not always exist. One alternative is to improve poorly performing dRNAs by manipulating dRNA:sgRNA ratios. Another is to combine dRNAs with the recently described xCas9 or spCas9-NG variants, which have a more permissive PAM that increases the number of candidate dRNAs.[213,214] Another drawback is that some dRNAs decrease on-target editing, particularly when they are multiplexed to suppress several off-target sites simultaneously. We suspect that these losses in on-target editing likely arise due to dilution of the plasmids or competition between sgRNAs and dRNAs to complex with

Cas9. The first issue could be addressed by using a multiplex guide expression scheme,[215,216] and both could be addressed by delivering preformed ribonucleoprotein (RNP) mixtures.[217] Finally, dRNAs could yield unwanted transcriptional off-target effects. However, transcriptional repression by Cas9 in the absence of a repressive domain is modest,[218,219] and such effects would be transient unless both the Cas9 and the dRNA were integrated into the genome.

Other approaches for minimizing off-target editing are also imperfect, as they reduce on-target efficiency,[126,143–145,157,158] introduce new off-target sites,[147,150,151] limit the number of potential target sites,[147,150–153] or demand difficult Cas9 engineering.[154–158,220,221] Moreover, many of these approaches are laborious to implement in experimental models where Cas9 or a variant thereof has already been stably integrated into the genome.[126,143–145,152–158,220,221] Finally, these existing methods are generally incompatible with each other, meaning they cannot be used in concert to minimize limitations and improve performance. In contrast, dOTS and dReCS are comparatively easy to use, low-cost, and flexible. For example, dOTS could be used to address refractory off-targets o the popular engineered high-specificity Cas9 variants.[154–158,220,221] Here, we showed that dOTS could effectively suppress editing at four refractory off-target sites with three high-specificity Cas9 variants. Using dOTS to address these refractory off-targets is also far less laborious and time-intensive than further Cas9 engineering, as has been done previously.[154,220] Additionally, dReCS is simpler and less time-consuming than CORRECT,[122] a previous approach for scarless HDR editing that requires multiple rounds of HDR to introduce and subsequently remove blocking mutations. Because of their flexibility and technical simplicity, dOTS

and dReCS could be easily integrated with existing protocols and experimental systems, enabling refinement of genome editing with minimal effort.

The flexibility of dOTS and dReCS means that they have applications beyond those we demonstrated. For instance, dOTS could facilitate allele-specific editing, even when the two alleles cannot be distinguished by a Cas9-sgRNA complex alone. Based on the principle of Cas9 self-competition, electroporation of Cas9-dRNA RNPs to quench editing by the active Cas9-sgRNA RNP should allow fine tuning of editing efficiencies. Similarly, dOTS could be employed to modulate the editing rates in CRISPR lineage tracing.[222] Finally, dOTS and dReCS are likely to be effective with other CRISPR enzymes, such as saCas9 or Cpf1. Thus, dOTS and dReCS are easy to implement, effective, and complementary methods for refining genome editing in both research and clinical applications.

3.4 Methods

3.4.1 Expression plasmids

All sgRNA and dRNA target sequences, except for *VEGFA* sgRNAs, were cloned into the gRNA Cloning Vector according to the hCRISPR gRNA synthesis protocol. gRNA Cloning Vector was a gift from G. Church, Harvard (Addgene plasmid 41824). *VEGFA* site 1 (*VEGFA* sgRNA1), *VEGFA* site 2 (*VEGFA* sgRNA2), and *VEGFA* site 3 (*VEGFA* sgRNA3) were gifts from K. Joung, Massachusetts General Hospital (Addgene plasmids 47505, 47506, and 47507).

An N-terminal FLAG tag sequence was appended via Gibson Assembly Cloning (New England Biosciences) to a human codon-optimized Cas9 (subcloned from hCas9, a gift from G. Church, Harvard; Addgene plasmid 41815) with a single C-terminal NLS expressed from a pcDNA3.3-TOPO vector. This was subsequently cloned into the pcDNA5/FRT/TO backbone (ThermoFisher). High-specificity variants of Cas9—espCas9(1.1.) (gift from F. Zheng, Broad Institute; Addgene plasmid 71814) and VP12 (spCas9-HF1, gift from K. Joung, Massachusetts General Hospital; Addgene plasmid 72247) were subcloned into pcDNA5/FRT/TO backbone (ThermoFisher). HypaCas9 (BPK4410) was a gift from J. Doudna and K. Joung, University of California, Berkeley and Massachusetts General Hospital (Addgene plasmid 101178).

3.4.2 dRNA design

dRNA sequences were designed by identifying 14-16 nucleotide dRNA spacer sequences which met the following criteria: (1) the dRNA spacer sequence and/or its PAM overlaps with the off-target spacer sequence and/or its PAM, and (2) the dRNA spacer or PAM exhibits perfect complementarity to the off-target but not the on-target locus. Spacer sequences with a 5’ G were preferentially selected, but spacers containing a mismatched 5’ G were also used. Exhaustive screening of all candidate dRNAs, which met the criteria was not performed at all sites. Alignments of the on-target sites, off-target sites, and dRNAs used in this study are presented in **Appendix B: Figure B.**4.

3.4.3 Cell culture

HEK-293T cells (293T/17, ATCC) were maintained in high-glucose DMEM supplemented with 10% fetal bovine serum (FBS, Life Technologies). U2OS cells (ATCC) were maintained in McCoy's 5A (modified) medium supplemented with 10% FBS (Life Technologies). hESC Elf1 iCas9[199] were plated into Matrigel-coated 24-well plates and cultured in MEF-conditioned media supplemented with 2il-I-F (GSK3i, MEKi, LIF, IGF, bFGF). All cell lines were regularly tested and confirmed free from mycoplasma contamination.

3.4.4 Genome editing by Cas9

Unless otherwise specified, HEK-293T cells were plated in 24-well plates at 1.5 x 10^5 cells/well. The day after plating, cells were transfected with Turbofectin 8.0 (Origene). For all dOTS experiments, 1.5 µL of Turbofectin 8.0 and 500 ng of plasmid DNA were transfected. For dRNA screening experiments, the plasmid DNA mixture contained 250 ng Cas9 (spCas9, espCas9, Cas9-HF1, or HypaCas9), 125 ng sgRNA, and 125 ng dRNA. For wells without dRNA, 125 ng of pMAX-GFP was substituted for the dRNA plasmid as a transfection control. For multiplex dOTS experiments, the plasmid DNA mixture contained 250 ng Cas9, 125 ng sgRNA, and 125 ng each of 1-3 dRNAs. pMAX-GFP plasmid was used to increase total DNA transfected per well to 750 ng. U2OS cells were plated in 12-well plates at 7.5 x 10^4 cells/well. The next day, they were transfected with 3 µL of Turbofectin 8.0 and a total of 1 µg plasmid DNA (500 ng Cas9, 250 ng sgRNA, and 250 ng dRNA or pMAX-GFP plasmid). For titration experiments with all sgRNAs except *VEGFA* sgRNA3, HEK-293T cells were transfected

with 1.5 µL of Turbofectin 8.0 and 500 ng of plasmid DNA. This DNA mixture contained 250 ng Cas9. The remaining 250 ng of DNA was divided between sgRNA and dRNA at varying ratios such that the total DNA was kept constant across experiments (1:1, 125 ng each sgRNA and dRNA; 1:2, 83.3 ng sgRNA and 166.7 ng dRNA; 1:4, 50 ng sgRNA and 200 ng dRNA; 2:1, 166.7 ng sgRNA and 83.3 ng dRNA; and 4:1, 200 ng sgRNA and 50 ng dRNA). For wells without dRNA, 125 ng of pMAX-GFP plasmid was substituted for the dRNA plasmid as a transfection control. For titration experiments with *VEGFA* sgRNA3, HEK-293T cells were transfected as above, but the DNA mixture contained 166.5 ng Cas9, and the various sgRNA:dRNA ratios were as follows (1:1, 166.5 ng each sgRNA and dRNA; 1:2 111 ng sgRNA and 222 ng dRNA; 1:4, 66.6 sgRNA and 266.4 ng dRNA; 2:1, 222 ng sgRNA and 111 ng dRNA; 4:1, 266.4 ng sgRNA and 66.6 ng dRNA). For wells without dRNA, 166.4 ng of pMAX-GFP was substituted for the dRNA plasmid as a transfection control.

To harvest HEK-293T and U2OS cells for dOTS experiments, 24 hours after transfection, each well of a 24-well plate was resuspended by thorough pipetting with 400 µL ice-cold DPBS. Resuspended cells were then spun at 1500 x *g* for 10 minutes at 4 °C. DPBS was then aspirated, and cell pellets were stored at -80 °C until genomic DNA isolation. For extended timepoint experiments, the same protocol was followed, excepted cells were passaged into a new 24-well plate 24 hours after transfection and then subsequently harvested 48 hours after passaging.

Two days prior to plating, hESC Elf1 iCas9 cells were treated with 2 µg/mL doxycycline to induce Cas9 expression. At day 0, 2.5 x 10^4 cells were plated into each well of a 24-well plate with addition of fresh doxycycline (2 µg/mL) and 10 µM Rock

inhibitor to promote cell survival. After 24 hours, cells were transfected with 3 µL of GeneJuice (EMD Millipore) and 1 µg plasmid DNA. This plasmid DNA mixture contained 500 ng sgRNA and 500 ng dRNA. For wells without dRNA, 500 ng of pMAX-GFP was substituted as a transfection control.

For Elf1 cells, 48 hours after transfection, each well of a 24-well plate was rinsed once with 0.5 mL DPBS and incubated for 5 minutes to detach cells. 5 mL hESC media was added and the cells were spun down at 290 x *g* for 3 minutes. The pellet was then washed with 1 mL DPBS, spun again at 290 x *g* for 3 minutes, then flash frozen in liquid nitrogen and stored at -80 °C until genomic DNA isolation.

For GUIDE-seq experiments, U2OS cells were electroporated following previously established protocols.[117,156] Briefly, 2 x 10^5 cells per condition were transfected with 500 ng Cas9 plasmid, 250 ng sgRNA plasmid, 250 ng dRNA plasmid, and 100 pmol of an end-protected double stranded oligonucleotide (dsODN) GUIDE-seq tag. For wells without dRNA or sgRNA, pMAX-GFP plasmid was substituted as a transfection control. 20 µL transfections were performed using a Lonza 49 nucleofector X unit and SE kit using the DN-100 program. Cells were replated in 96-well plates after transfection and harvested for genomic DNA 96 hours later.

3.4.5 dReCS

For dReCS experiments, a HEK-293T cell line with a genomically encoded BFP/GFP reporter was used.[212] The BFP/GFP reporter HEK-293T cell line contains a BFP that is converted to GFP via HDR-mediated substitution of a single amino acid (His in BFP (**C**AT) to Tyr in GFP (**T**AT)). BFP/GFP reporter cells were plated at 3 x 10^5

cells/well in 12-well plates. 18 hours after plating, cells were transfected with 3 µL of Turbofectin 8.0 and 1000 ng of total DNA. The total DNA mixture contained 272.7 ng Cas9 plasmid, 54.5 ng sgRNA plasmid, 218 ng dRNA plasmid, and 454.5 ng symmetric or asymmetric single stranded donor DNA (**Appendix C, Table 4**).[212] For controls missing one or more of these DNA elements, the appropriate amount of DNA was replaced by a pKan-mCherry plasmid. Cells were maintained with standard passaging procedures for 4 days post-transfection until analysis by flow cytometry.

After 4 days, cells were washed with 2 mL DPBS, trypsinized with 0.5 mL 0.25% trypsin-EDTA (Life Technologies) for 2-4 minutes and quenched with DMEM supplemented with 10% FBS. Cells were then spun down at 290 x *g* for 4 minutes, aspirated, and resuspended in DPBS supplemented with 1% FBS. Cells were run through a 35 µm filter and analyzed by flow cytometry on an LSR-II flow cytometer. After gating for live cells (FSC-A vs SSC-A) and single cells (FSC-A vs SSC-W), cells were analyzed for their BFP and GFP fluorescence. Gates for BFP and GFP positivity were determined by comparison to an untransfected BFP cell line. BFP+ GFP– cells were considered wildtype (WT). BFP– GFP– cells were considered to have undergone non-homologous end joining (NHEJ) but not HDR, as indels in this region of BFP lead to loss of fluorescence. Any cell that was GFP+ (regardless of residual BFP fluorescence) was considered to have undergone successful HDR. Percentages for each result (WT, HDR, and NHEJ) were calculated as a fraction of total cells that passed singlet gating. Percent HDR of total editing was determined as the fraction of cells with successful HDR divided by the total number of cells that underwent either HDR or NHEJ.

3.4.6 *In vitro* Cas9-RNP nuclease assays

Cas9-2NLS in a pMJ915 vector (Addgene plasmid 69090) was expressed in *E. coli* and purified by a combination of affinity, ion exchange, and size exclusion chromatography as previously described,[223] except the final purified protein was eluted into a buffer containing 20 mM HEPES KOH pH 7.5, 5% glycerol, 150 mM KCl, 1 mM DTT at a final concentration of 40 μM of Cas9-2NLS. *FANCF*-sgRNA2 and *FANCF*-dRNA1 were generated by HiScribe (NEB E2050S) T7 *in vitro* transcription using PCR-generated DNA as a template.[223]

A 463 basepair fragment containing the on-target cut site of *FANCF*-sgRNA2 (*FANCF* target sit) was PCR amplified from a custom *FANCF*-sgRNA2 target sit substrate gBlock (IDT) using primers oCR1711 and oCR1712. A 329 basepair fragment containing the cut site for off-target 1 of *FANCF*-sgRNA2 (*FANCF* off-target) was PCR amplified from a custom *FANCF*-sgRNA2 off-target substrate gBlock (IDT) using oCR1713 and oCR1714. Prior to nuclease experiments, sgRNA and dRNA-RNP complexes were generated by incubating purified Cas9-2NLS and *FANCF*-sgRNA or dRNA1 in equimolar amounts for 10 minutes. For dRNA-RNP titration experiments, 150 or 450 fmoles of *FANCF*-sgRNA2-RNP complex and 0, 50, 150, or 450 fmoles of dRNA-RNP Cas9-sgRNA complex were co-added to 150 fmoles of *FANCF* target site or *FANCF* off-target substrate DNA. Reaction mixtures were incubated at 37 °C for 20 minutes in 20 mM Tris, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 0.01% Tween, and 50 μg/mL heparin. Reactions were stopped by the addition of a 1:4 volume of STOP solution (8 mM Tris, 0.025% BPB, 0.025% XC, 50% glycerol, 110 mM EDTA, 1% SDS, and 3 mg/mL proteinase K), followed by incubation at 55 °C for 5 minutes to liberate cut

DNA fragments. Each digestion reaction was run on a 2% TAE agarose gel, post-stained with ethidium bromide, and resolved on a Gel-Doc (BioRad).

For pre-incubation experiments, *FANCF*-sgRNA2 or dRNA1 RNP complexes were generated as described above. 450 fmoles of a single RNP complex was added to 150 fmoles of *FANCF* target site or *FANCF* off-target substrate DNA and incubated at 37 °C for 10 minutes. After 10 minutes, 450 fmoles of the other Cas9-RNP complex was added and allowed to incubate at 37 °C for an additional 10 minutes. Reactions were quenched, incubated, and run on a gel in an identical manner to the above experiments.

Gel densitometry analysis was performed in ImageJ. For each lane, background density was subtracted from the quantification of each band. The density of the uncut band was then divided by the total intensity of all bands in the lane to determine the uncut DNA fraction.

3.4.7 Genomic editing by ciCas9

HEK-293T cells were treated according to previous methods.[126] Briefly, HEK-293T cells were plated in 12-well plates at 3 x 10^5 cells/well. The day after plating, cells were transfected with 1.5 µL Turbofectin 8.0 and 500 ng of plasmid DNA. The plasmid DNA mixture contained 250 ng Cas9, 125 ng *FANCF*-sgRNA2 and 125 ng dRNA. For wells without dRNA, the 125 ng of dRNA plasmid was replaced by pMAX-GFP as a transfection control.

24 hours after transfection, cells were treated with 10 µM A115 dissolved in DMSO to induce ciCas9 activity. 24 hours after treatment with A115, cells were harvested after washing with 600 µL DPBS to remove excess A115 and resuspended in

600 µL ice-cold DPBS. Resuspended cells were then spun at 1500 x *g* for 10 minutes at 4 °C. DPBS was aspirated, and the cell pellets were stored at -80 °C until genomic DNA isolation.

3.4.8 Indel detection by high-throughput sequencing

Genomic DNA isolation, sequencing, and analysis were performed as previously described.[126] Briefly, genomic DNA was isolated using the DNEasy Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions, except that the proteinase K digested was conducted for 1 hours at 56 °C. 15 cycles of primary PCR to amplify the region of interest was performed using 2 µL of DNEasy eluate (~100-300 ng template) in a 5 µL Kapa HiFi HotStart polymerase reaction (Kapa Biosystems). The PCR reaction was diluted with 35 µL DNAse-free water (Ambion). Illumina adapters and indexing sequences were added via 20 cycles of secondary PCR with 3 µL of diluted primary PCR product in a 10 µL Kapa Robust HotStart polymerase reaction (Kapa Biosystems). The final amplicons were run on a 1.5% TBE-agarose gel, and the product band was excised and extracted using the Freeze and Squeeze kit according to the manufacturer's instructions (BioRad). Gel-purified amplicons were quantified using Qubit dsDNA HS Assay Kit (Invitrogen). Then, up to 1200 indexed amplicons were pooled, quantified by Kapa Library Quantification (Kapa Biosystems) and sequenced on a NextSeq (NextSeq 150/300 Mid v2 kit, Illumina).

Indels were quantified as previously described.[126] Briefly, after demultiplexing of reads (bcl2fastq/2.18, Illumina), indels were quantified with a custom Python script that is freely available upon request. 8-mer sequences were identified in the reference

sequence located 20 bp upstream and downstream of the target sequence. Sequence distal to these 8-mers was trimmed. Reads lacking these 8-mers were discarded. For the *VEGFA* sgRNA3 OT2 locus, the process was the same, except 20-mer sequences located 10 bp upstream and downstream of the target sequence were used. For the *VEGFA* sgRNA3 OT4 locus, 8-mer sequences located 10 bp upstream and downstream of the target sequence were used. The trimmed reads were then evaluated for indels using the Python difflib package. Indels were defined as trimmed reads, which differed in length from the trimmed reference and for which an insertion or deletion operation spanning or within 1 bp of the predicted Cas9 cleavage site was present. For dRNA only experiments, indels were quantified using both the sgRNA and dRNA predicted cut sites. Specificity ratios were calculated by dividing the indel percentage at the on-target locus by the indel percentage at the off-target locus for each sgRNA. For quantification of off-target editing for one of the *VEGFA* tru-sgRNA3 dRNA replicates (**Figure** 3.4**a**), reads were acquired from multiple sequencing runs.

3.4.9 GUIDE-seq

Calculation of indels was performed at the *FANCF*-sgRNA2 ON and OT1 loci as described above. To determine the percentage of reads containing a dsODN tag, the same Python script as above was used and modified to count integration of the full length dsODN within 1 bp of the predicted Cas9 cleavage site. A ratio of dsODN-containing reads to indel-containing reads was calculated. To perform GUIDE-seq analysis, samples were prepared according to established protocols.[117] Briefly, genomic DNA was isolated using the DNEasy Blood and Tissue Kit according to the

manufacturer's instructions except that the proteinase K digestion was conducted for 1 hour at 56 °C. DNA was sheared using a Covaris LE220 to an average size of 500 bp and cleaned using Ampure XPRI beads according to the manufacturer's protocol. DNA was then end-repaired, A-tailed, and ligated to adapters containing an 8 bp unique molecular identifier. Samples were then amplified with two rounds of nested PCR with primers that complement the oligo tag. Sample libraries were prepared as described above and sequenced on an Illumina MiSeq. Data were analyzed with the GUIDE-seq software[224] allowing for up to eight mismatches with a modification of a 35 bp window for detecting off-target alignments to the reference sequence. The frequency of dsODN-containing reads genome-wide was calculated per sample.

3.4.10 Statistical analysis

Statistical analysis of indel frequency and specificity ratios were performed using a one-side two sample Student's *t*-test.

3.4.11 Data availability

Raw sequencing data have been uploaded to the SRA with BioProject accession number PRJNA629634. Custom Python scripts for indel quantification and R scripts for figure generation are available on GitHub.

4 The future of technology development in genomics

The scientific environment is ripe for developing new genomic technologies, as is evidenced by the explosion of new methods being published. Even though much work has been done to address a variety of needs in the scientific community, there is plenty more to develop. The following are a few possible directions.

4.1 Resolving missing variants in patients

The bulk of my work has focused on the effects of missense variants in hemophilia B, as this represents the largest fraction of affected individuals. However, more complex variation within *F9* exists, and a number of patients remain genetically undiagnosed, even after targeted gene sequencing with Illumina.[59] That said, Illumina's short read lengths make it difficult to resolve larger structural rearrangements, identify copy number variants, and phase genetic variants to detect compound heterozygotes. In order to capture these types of variants and bring genetic diagnostics to 100% of patients, additional methods are required.

Long read sequencing methods, like PacBio SMRT technology and Oxford Nanopore sequencing, have been used to identify structural variants, but are typically cost-prohibitive for genome-wide sequencing.[225] To reduce costs, targeted capture of specific DNA regions can be used when clinical suspicion is high and disease mechanism is known, but the methods used to select specific regions of DNA can lead to unintended alterations that can be confused with true variants.[226] To combat these issues, computational methods have recently been developed that allow for real time acceptance or rejection of DNA molecules during Oxford Nanopore sequencing without

the need for targeted capture.[226–229] Moreover, PacBio and Oxford Nanopore both have read accuracy rates in the low 90% range, which calls into question their utility for calling single base changes.[225] Consensus calling for PacBio sequencing has raised accuracy to >99%, but comes with a significant reduction in read lengths.[225] In either case, applying long read sequencing to undiagnosed patients with hemophilia B represents a logical next step for identifying potentially causative variants that can then be assayed for function. That said, for patients that remain undiagnosed after targeted long read sequencing, the only remaining recourse is whole genome sequencing. As such, more technology development to make long read sequencing tools more accurate and less costly are still needed.

4.2 Expanding the MAVE toolkit

MAVEs are themselves a powerful tool for characterizing variation within genetic elements, allowing for the assessment of thousands to tens of thousands of variants simultaneously. However, MAVEs are not without their limitations. Most MAVEs can only profile variants in a single context for a single phenotype, and as such, often paint an incomplete picture of how a variant affects function. They are also often performed in particularly hardy cell lines with exogenous, overexpressed constructs, which loses any contextual information and can lead to false conclusions. In this section, I identify areas for development in MAVE technology and some potential solutions.

4.2.1 Expanding cell surface display to additional secreted proteins and phenotypes

In **Chapter 2**, I measured the effect of variation on FIX secretion, and in doing so, found a number of biochemical features that correlate with whether or not a variant FIX molecule will be successfully secreted. Many of these same features, like solvent accessibility and secondary structure elements, correlate in orthogonal assays on the abundance of intracellular proteins.[4–6] These findings emphasize that the ability for a variant to fold properly determines whether said variant will be capable of being expressed *in vivo*. However, I also identified unique features that only seem to be predictive for FIX secretion and are otherwise uncorrelated in measures of abundance in intracellular proteins. Of these, some are expected, like the effect seen in secretion peptide variants. Others, like the outsized role of cysteine residues, are unexpected and warrant further investigation.

Specifically, a question arises about the generalizability of both of the secretion peptide and cysteine residue findings. With regards to signal peptides, much is known about their general structure,[43] but their diverse nature makes understanding variation within them challenging. Algorithms to predict signal peptide presence cannot predict whether a given variant within a secretion peptide will alter its function.[42,178] Indeed, alignments of secretion peptides has led to the longstanding hypothesis that the primary requirement is simply a stretch of otherwise random hydrophobic amino acids.[230] However, this does not explain the fact that secretion peptides typically are species-specific,[231,232] and that optimization of signal peptide sequences can lead to unforeseen consequences like protein misfolding and improper post-translational modification.[43]

More recent work suggests that there are potentially species-specific cryptic motifs embedded within the hydrophobic h-region that permit successful secretion.[233] As such, there is a surplus of missing knowledge about the structure of secretion peptides and the role of variation within them. Expanding the MAVE described in **Chapter 2** to many diverse, secreted proteins, would be instrumental in creating better algorithms for understanding how variation impacts secretion. Moreover, when used in conjunction with additional assays, like the γ-carboxylation assay described in **Chapter 2**, these algorithms can then be used to optimize secretion peptides while maintaining overall protein function.[45,234,235]

Similarly, the effect of cysteine residues on FIX is not seen in intracellular assays but has been described in a limited number of extracellular-facing membrane-embedded and secreted proteins.[236–238] Intracellular proteins have fewer disulfide bonds compared to secreted proteins, which is thought to maintain protein folding in the harsher oxidizing environment of the extracellular space, and is supported by evidence that unpaired cysteines are rarely observed in secreted proteins.[239] Intracellular protein disulfide bonds that do form are typically maintained as a redox-sensing mechanism.[240,241] Challenging the idea that disulfide bonds in secreted proteins are solely structural in nature, recent work suggests that extracellular protein disulfide bonds can function as allosteric activators or inhibitors, induce proteolysis, and alter membrane permeability.[242–244] Which of these possibilities is happening for cysteine variants in secreted proteins is currently unknown but could be answered by using MultiSTEP. Because display methods are conducive to biochemical measurements of protein affinity and modification,[37] I envision developing novel assays to use with

MultiSTEP that can differentiate between the many roles cysteines seem to play in secreted proteins.

While measuring secretion and γ-carboxylation in FIX is informative, it does not give a complete picture of how missense variation affects FIX function. When comparing my antibody scores to clinical data from patients with hemophilia, many known pathogenic variants show little to no phenotypic effect. Scoring variants on additional phenotypes, such as drug response, enzymatic activity, and protein-protein interactions can paint a rich protein atlas and will likely identify the causative mechanism for many, if not all, of these missed pathogenic variants. As I have shown through a pilot FXI_a activation experiment (**Fig. X**), MultiSTEP can be used for assaying these phenotypes and represents a straightforward next step to understanding the complexity of FIX variant function *in vivo*. Moreover, it reasons that these expanded assays can be used on other secreted proteins. That said, assay optimization and validation will be paramount to ensure that the generated results are applicable to clinical settings.

4.2.2 Interrogating synonymous and noncoding variant effects

In addition to coding variation, understanding the effects of variation in noncoding or synonymous contexts is valuable, especially in FIX. As an example, we can look to hemophilia B Leyden, a subtype of hemophilia characterized by a coagulation disorder that resolves during puberty. Hemophilia B Leyden is caused by single nucleotide variants within the promoter region of *F9*. There are two overlapping transcription factor binding motifs within the -26 to -18 positions of the promoter

region—one for hepatocyte nuclear factor 4 (HNF-4) and another for androgen receptor (AR).[245,246] Variants at position -26 lead to lifelong hemophilia B due to alteration of both the AR and HNF-4 motifs, whereas variants at positions -20 and -21 alter only the HNF-4 motif. Because the AR motif is still active in patients with -20 or -21 variants, their bleeding phenotype resolves at puberty when androgen levels rise significantly.[245] Similarly, synonymous and intronic variants that alter or ablate FIX splicing, 5' UTR variants that alter transcription, and large structural variants that eliminate FIX expression have been described, but are not profiled in my assays.[59,77]

As it is currently built, MultiSTEP does not have the ability to profile variants of these types. Massively parallel reporter assays (MPRAs), like STARR-seq can detect noncoding variant effects, but typically rely on plasmid reporters that are separated from their native genomic context.[247,248] How translatable the findings from plasmid-based MPRAs will be is still in question. Recent work suggests that testing putative regulatory DNA outside of its endogenous context can lead to different conclusions about functional effects than assays performed in endogenous contexts.[249]

Existing MAVEs that assay variant effects in their native contexts, like SGE, cannot profile secreted protein variant effects, since they cannot maintain the link between genotype and phenotype. To study FIX or other secreted proteins in their native context in a massively parallel way will require development of new tools. A simple possibility would be to append the linker and transmembrane domain sequences used in my display system to endogenous FIX by using Cas9-based gene editing, so that expression can be measured on the surface rather than on the secreted protein itself. Variants could then be introduced and measured using SGE,[16] though adaptation

to a non-growth-based assay would be required. Because of the concern about Cas9 off-target editing raised in **Chapter 1** and **Chapter 4**, I could envision using other Cas9-based effectors that install variants without forming DSBs like prime editing,[250] which has been used to functionally characterize variation in *NPC1* in its endogenous context.[34]

4.2.3 Developing MAVEs using appropriate cellular contexts

In **Chapter 2**, I performed secretion and γ-carboxylation assays on FIX in a suspension HEK-293 line. In humans, however, FIX is produced, modified, and stored in the liver before secretion. Recent work suggests that the PTMs placed on FIX are dependent on their mode of production, including cell-of-origin.[251] In other genes, variant effects can only be identified in the appropriate cell types, as typical lines do not express the gene itself or required binding partners. For instance, in *MYBPC3*, a gene in which variants lead to an autosomal dominant hypertrophic cardiomyopathy, the characteristic contractile defect could only be elicited in cardiomyocytes derived from induced pluripotent stem cells (iPSC).[252] Both examples identify the need for new cellular contexts to identify variant effects.

The problem with non-workhorse cell lines is that they tend to be difficult to work with, whether that be through their transfectability, growing conditions, or ability to be engineered. Moreover, many phenotypes may not be measurable in a commercial cell line and require iPSC differentiation to be observed.[252] Since MAVEs measure thousands of variants simultaneously, efficient introduction and expression are necessary to be able to adequately measure phenotypic effects for all variants. The

introduction of landing pads[8] for recombination-based variant integration have improved throughput significantly, but these approaches have had variable successes in other systems, primarily through genetic silencing of the landing pad itself. In fact, because iPSCs have such fluid epigenetic profiles, we have found in our lab that landing pads completely silence upon differentiation. A simple solution could be using histone deacetylase inhibitors during differentiation, though there is high likelihood that these inhibitors would alter differentiation itself. A more conceivable direction forward, though laborious, could involve engineering a landing pad that maintains its own epigenetic active state. For instance, integration of a Cas9-fused histone acetyltransferase that can maintain chromatin state in the landing pad promoter could circumvent any epigenetic silencing that would otherwise occur.[253]

4.3 Improving Cas9-based editing and off-target identification

In the past decade since its discovery, Cas9-based approaches have seemingly overtaken other methods for introducing gene edits, interrogating functional effects, and modifying transcriptomes.[32,161] However, Cas9 is not without its limitations, most notably off-target effects. Much work, including my own, has strived to reduce or eliminate all off-target effects, with major, but incomplete, success.[147,152,154,156] While work continues to be done to improve off-target editing through a variety of methods, I will focus the rest of this section on improvements to the detection of off-targets and the potential applications of the work I developed in **Chapter 3**.

4.3.1 Improved methods for detecting off-targets genome-wide

As shown in **Chapter 3**, Cas9-based effectors have significant off-target effects that can confound MAVE results, especially when multiple rounds of editing with various sgRNAs are required. In order to reliably use Cas9-based effectors in MAVEs, simple and unbiased methods for detection of off-target effects genome-wide are required. GUIDE-seq, for instance, is unbiased and can be performed genome-wide; however, it is reliant on intact end-joining cellular machinery, and my work suggests that less than 30% of edits successfully integrate the dsODN (**Appendix X, Fig. X**). Because of these two features of GUIDE-seq, it often fails to capture all DSBs, as it cannot capture edits that have been resolved without the integrated dsODN.[117] On the other hand, breaks labeling *in situ* and sequencing (BLISS) is both unbiased and does not require DNA repair machinery, as it is performed on fixed cells prior to ligation of DSB adapters.[254] However, the blunt end ligation of DSB adapters is inefficient and the fixation steps can lead to artificial DSBs. Moreover, neither approach allows for the simultaneous identification of DSBs, indels, and base changes.

Even when DSBs are not introduced in the process of installing base changes, off-target effects occur, both within the gene of interest and distally.[255,256] To detect off-target base editor effects requires whole genome off-target sequencing, as many of the existing tools rely on incorporation of a label at the DSB site.[117,254,257] One possible way to reduce the need for off-target editing is to specifically label edited genomic DNA with biotin, which can easily be captured with streptavidin. A similar approach has been used to purify 5-hydroxymethylcytosine residues, though that method relied on an inefficient multi-enzyme reaction.[258] Instead, the biotinylated modification could be engineered into

the repair template itself, allowing for time-resolved capture of editing events. In the event that biotinylated DNA is unstable or toxic to living cells, 5-methylcytosine or ferrocene labels, which can both be identified using Oxford Nanopore sequencing,[259,260] could be used.

4.3.2 Using dRNAs for allele-specific editing

A common need in biological research and in clinical settings is to identify or modify only one allele in a given gene in a diploid organism. Standard Cas9-based editing approaches can rarely differentiate two alleles within an organism due to limited sequence dissimilarity. The dRNA method I developed in **Chapter 3**, however, can be adapted to this end. For instance, dRNAs can be used to block the allele that is to be left unedited, if sequence differences between the two exist. It can also be used during Cas9-based gene deletion to generate a haploid cell line by protecting only one of the alleles.[34] Unlike existing allele-specific Cas9 editing approaches, dRNAs do not require the variant to either be immediately adjacent to or within the PAM sequence.[261] Simultaneously, additional dRNAs can be introduced that protect the edited site, preventing reversion or indel formation.

4.4 Final thoughts on the promise of genomic medicine

Throughout this dissertation, I have described new genomic tools that allow us to explore new ways of interrogating the microscopic world at scale. Through the entire process, I have kept clinical implementation in mind, with the hope that something I have done here will have a measurable impact on someone's life. This is the promise of

precision medicine—that our DNA can teach us how to heal each other—and I believe that we are on the precipice of fulfilling that promise. My hope is that this work represents a significant step towards the goals of precision medicine and that those who come after me can use, build upon, and improve what I have created, so that one day, we will know the effects of every possible variant and how to alter them in ourselves.

Appendix A. Chapter 2 supplementary material

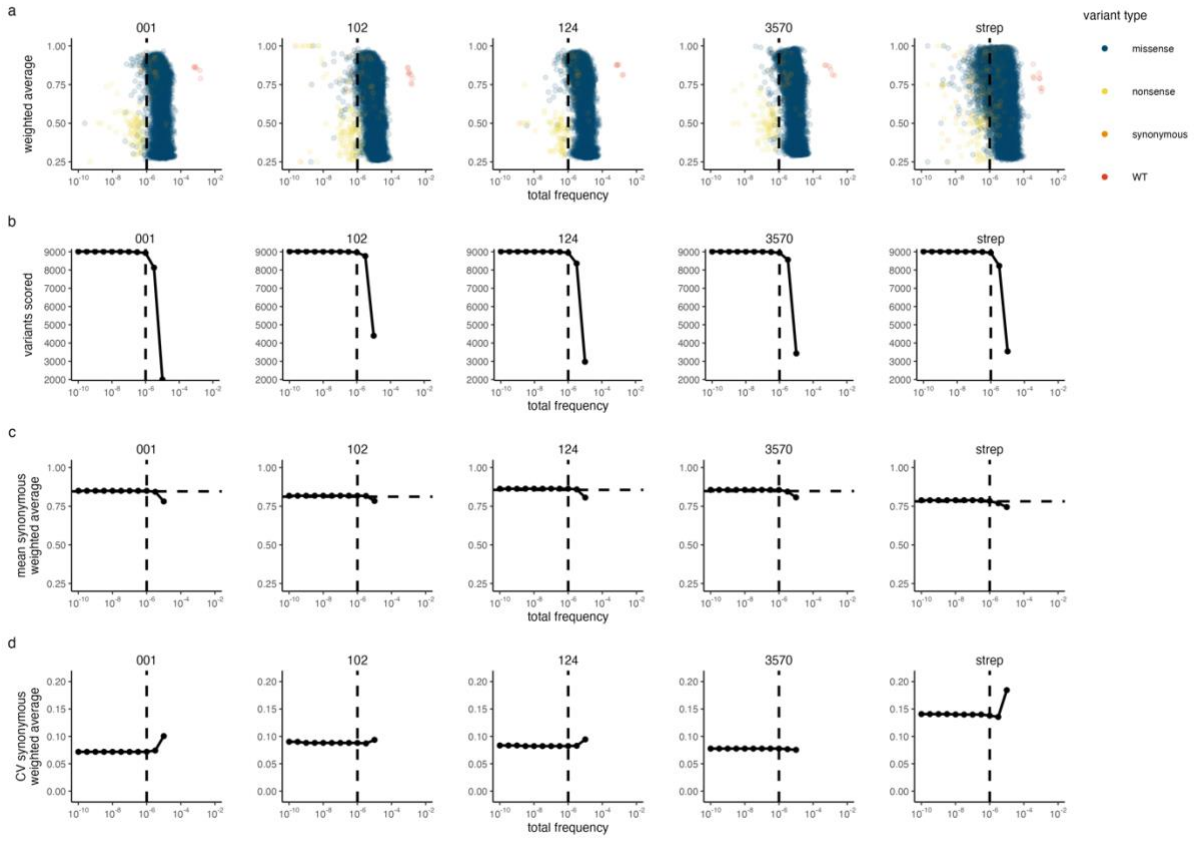

Figure A.1: Variant frequency filtering

Panels describing variant frequency filtering metrics for each antibody sort. **a.** Scatterplots showing the total frequency of variants across all bins in each library sort and their respective weighted average value. Colors indicate variant type. **b.** Number of variants retained at each threshold frequency value. **c.** Mean synonymous variant distribution of weighted averages at each threshold frequency. Dashed horizontal line indicates the weighted average of wildtype FIX **d.** Coefficient of variation of the synonymous variant distribution of weighted averages at each threshold frequency. The final frequency filter of 10^{-6} is shown as a dashed vertical line.

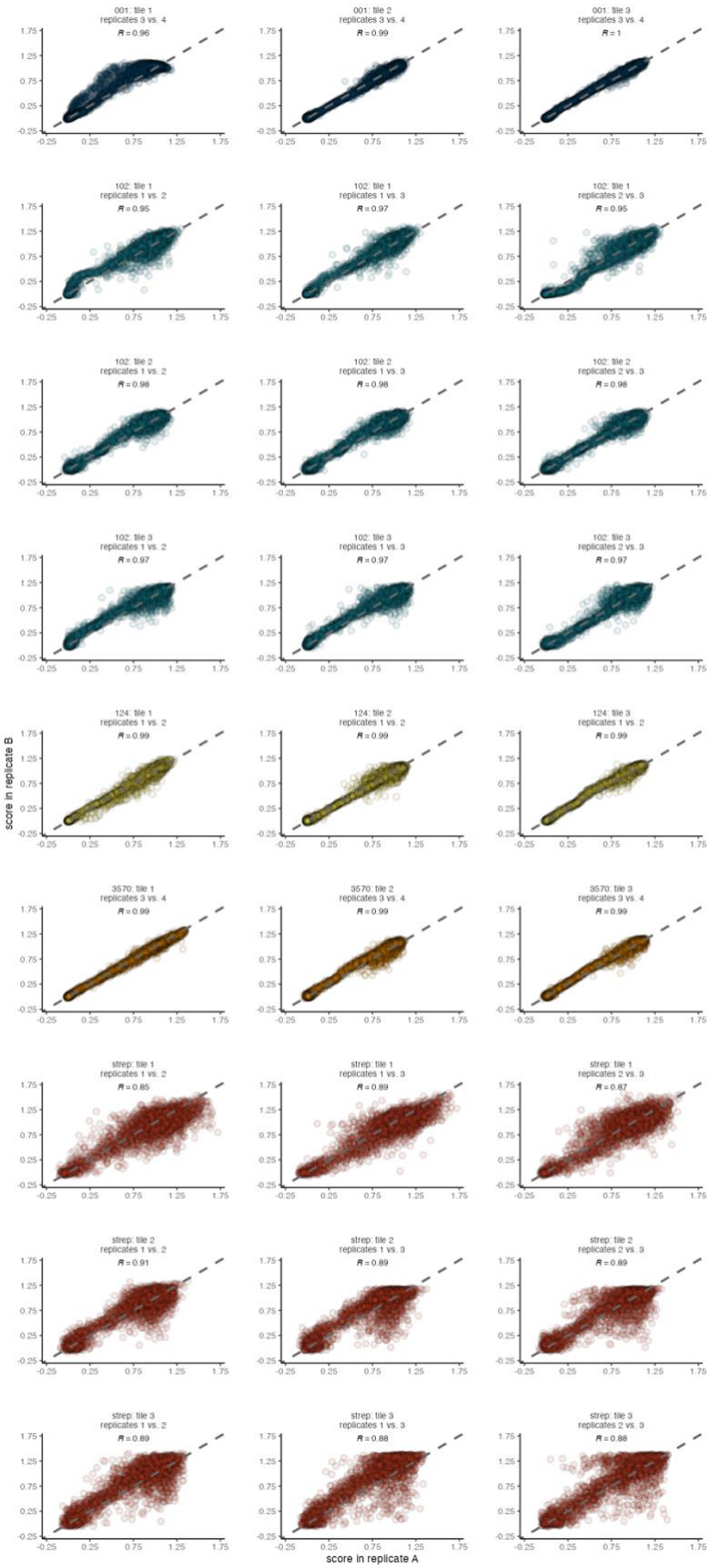
001: tile 1
replicates 3 vs. 4
R = 0.96
001: tile 2
replicates 3 vs. 4
R = 0.99
001: tile 3
replicates 3 vs. 4
R = 1
102: tile 1
replicates 1 vs. 2
R = 0.95
102: tile 1
replicates 1 vs. 3
R = 0.97
102: tile 1
replicates 2 vs. 3
R = 0.95
102: tile 2
replicates 1 vs. 2
R = 0.98
102: tile 2
replicates 1 vs. 3
R = 0.98
102: tile 2
replicates 2 vs. 3
R = 0.98
102: tile 3
replicates 1 vs. 2
R = 0.97
102: tile 3
replicates 1 vs. 3
R = 0.97
102: tile 3
replicates 2 vs. 3
R = 0.97
124: tile 1
replicates 1 vs. 2
R = 0.99
124: tile 2
replicates 1 vs. 2
R = 0.99
124: tile 3
replicates 1 vs. 2
R = 0.99
3570: tile 1
replicates 3 vs. 4
R = 0.99
3570: tile 2
replicates 3 vs. 4
R = 0.99
3570: tile 3
replicates 3 vs. 4
R = 0.99
strep: tile 1
replicates 1 vs. 2
R = 0.85
strep: tile 1
replicates 1 vs. 3
R = 0.89
strep: tile 1
replicates 2 vs. 3
R = 0.87
strep: tile 2
replicates 1 vs. 2
R = 0.91
strep: tile 2
replicates 1 vs. 3
R = 0.89
strep: tile 2
replicates 2 vs. 3
R = 0.89
strep: tile 3
replicates 1 vs. 2
R = 0.89
strep: tile 3
replicates 1 vs. 3
R = 0.88
strep: tile 3
replicates 2 vs. 3
R = 0.88
score in replicate B
score in replicate A
-0.25
0.25
0.75
1.25
1.75

Figure A.2: Variant-level replicate correlations across tiles and antibodies

Variant scores across replicates for each antibody. All combinations of experimental replicates are shown as individual panels. Colors indicate antibody. *R* = pairwise Pearson's correlation.

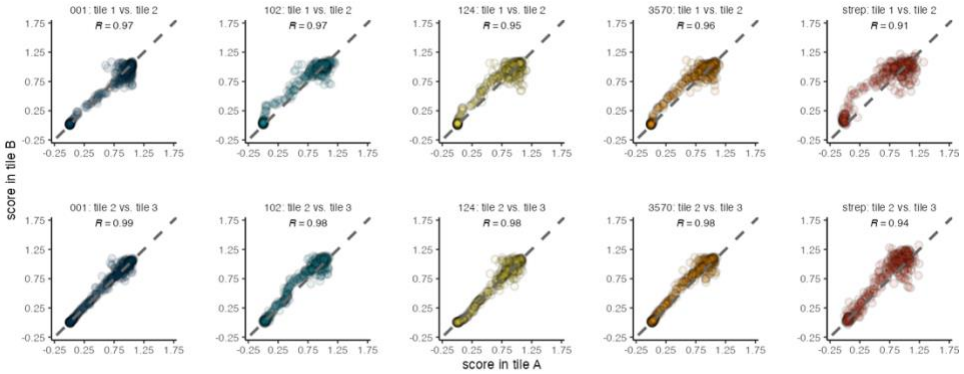

Figure A.3: Tile-level replicate correlations for shared variants

Scatterplot showing the score of the 20 library positions that were shared between each adjacent tile for each antibody. Colors indicate antibody. *R* = pairwise Pearson's correlation.

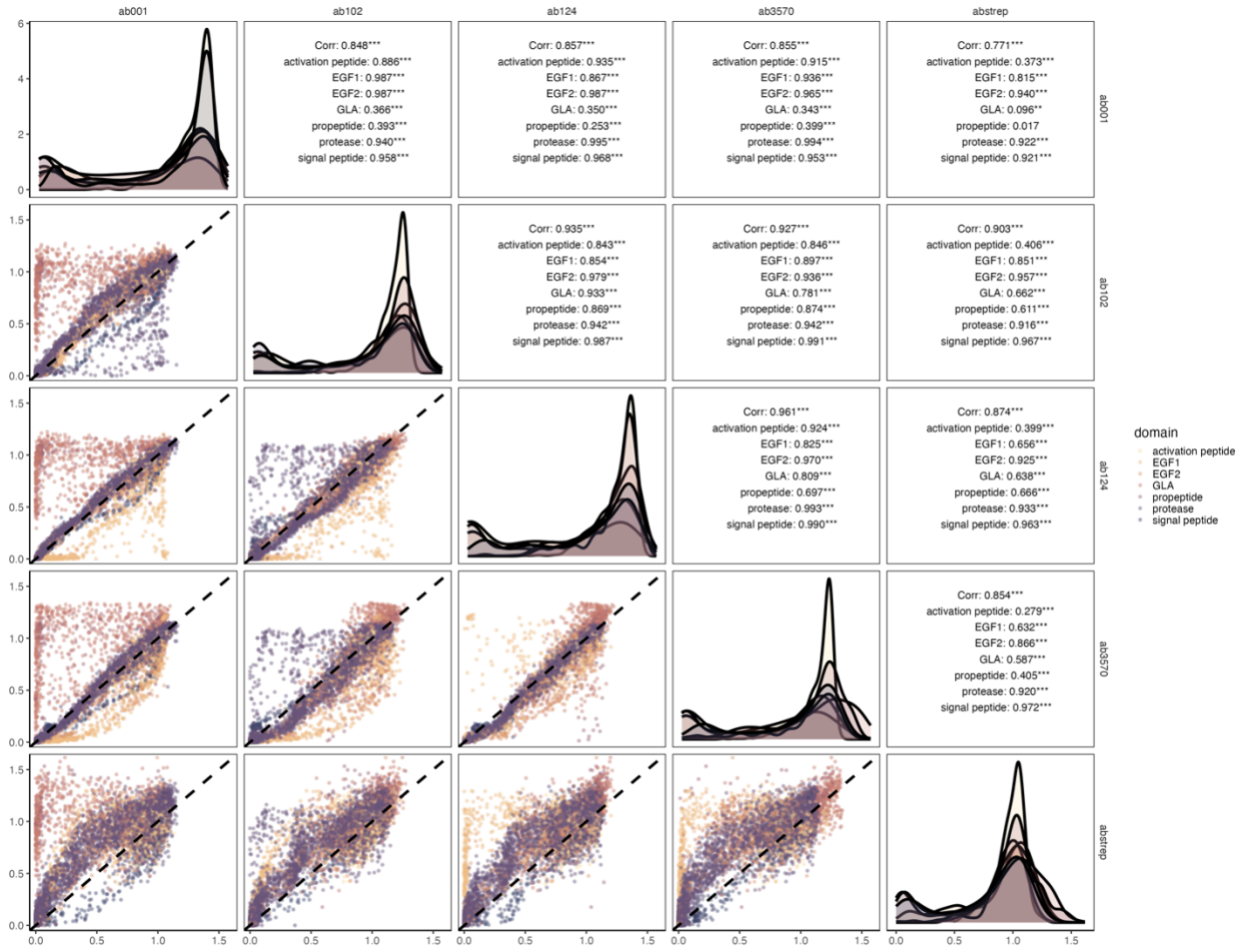

Figure A.4: Antibody score correlations

Pairwise plot of mean variant antibody scores against one another. Lower triangle: scatterplot, colored by FIX domain. Diagonal: density plots colored by FIX domain. Upper triangle: Pearson's r correlations, split by domain and overall.

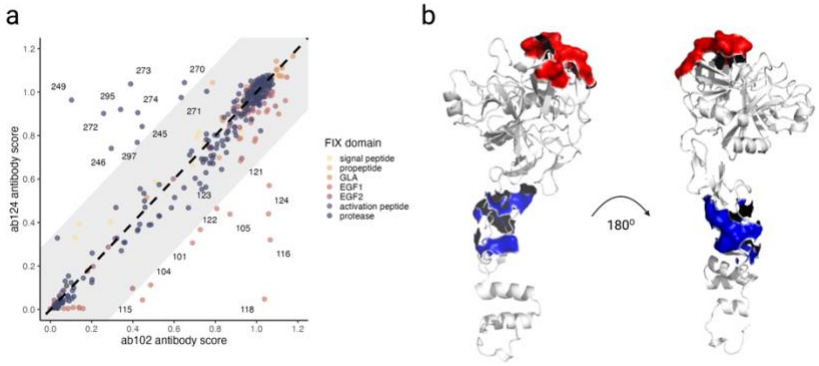

Figure A.5: Epitope mapping using MultiSTEP

a. Scatterplot of mean antibody scores at each FIX position, comparing FIX heavy chain (ab102) and FIX light chain (ab124). Colors indicate FIX domain. Light chain: GLA, EGF1, and EGF2 domains. Heavy chain: protease domain. Dashed line indicates perfect comparison. Shaded grey area represents window range of a score deviation of 30% or more from perfect correlation. Labeled points outside grey band indicate positions that are differentially bound to heavy chain vs. light chain antibodies. **b.** FIX structure based on EAHAD homology model, with two views rotated by 180°.[77] Red surface indicates labeled positions identified in **a** in upper left quadrant. Blue surface indicates labeled positions identified in **a** in lower right quadrant.

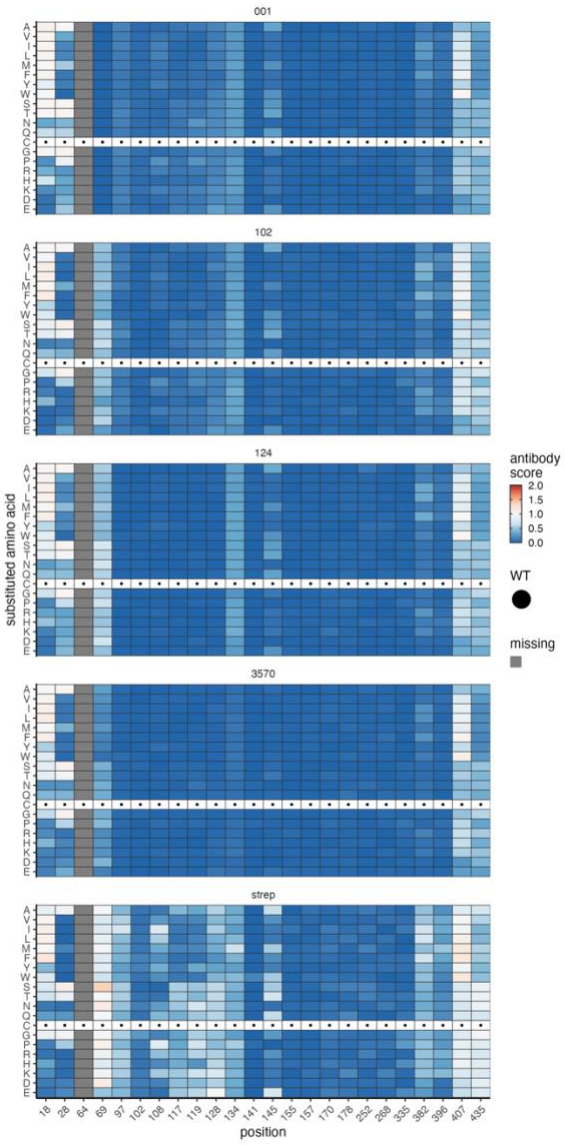

Figure A.6: Wildtype cysteines show significant loss of function with anti-FIX antibodies

Heatmaps of wildtype cysteine residues in FIX for all five measured antibodies. Scores are shown as colored boxes, where blue is loss of function and white is wildtype-like. Black dots indicate wildtype residues, and grey boxes indicate missing data.

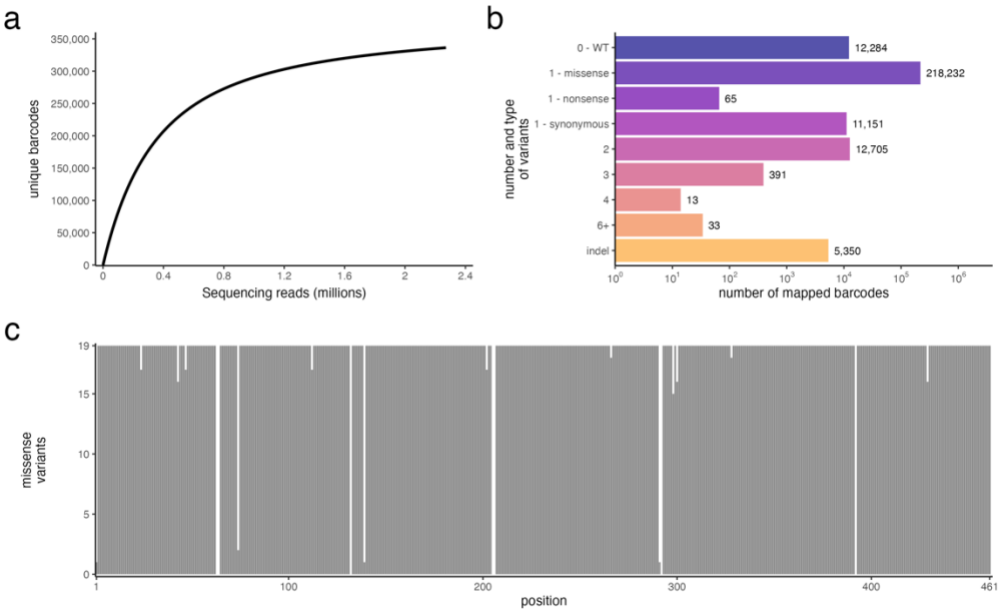

Figure A.7: Barcode sequencing analysis

a. Collector's curve for PacBio sequencing. Line shown is the average of 100 iterations of random sequence draws without replacement. **b.** Mutation type analysis for PacBio barcode sequencing. Shown are the number of unique barcodes associated with various variant types. Text indicates exact numbers. **c.** Number of barcoded missense variants at each position in FIX.

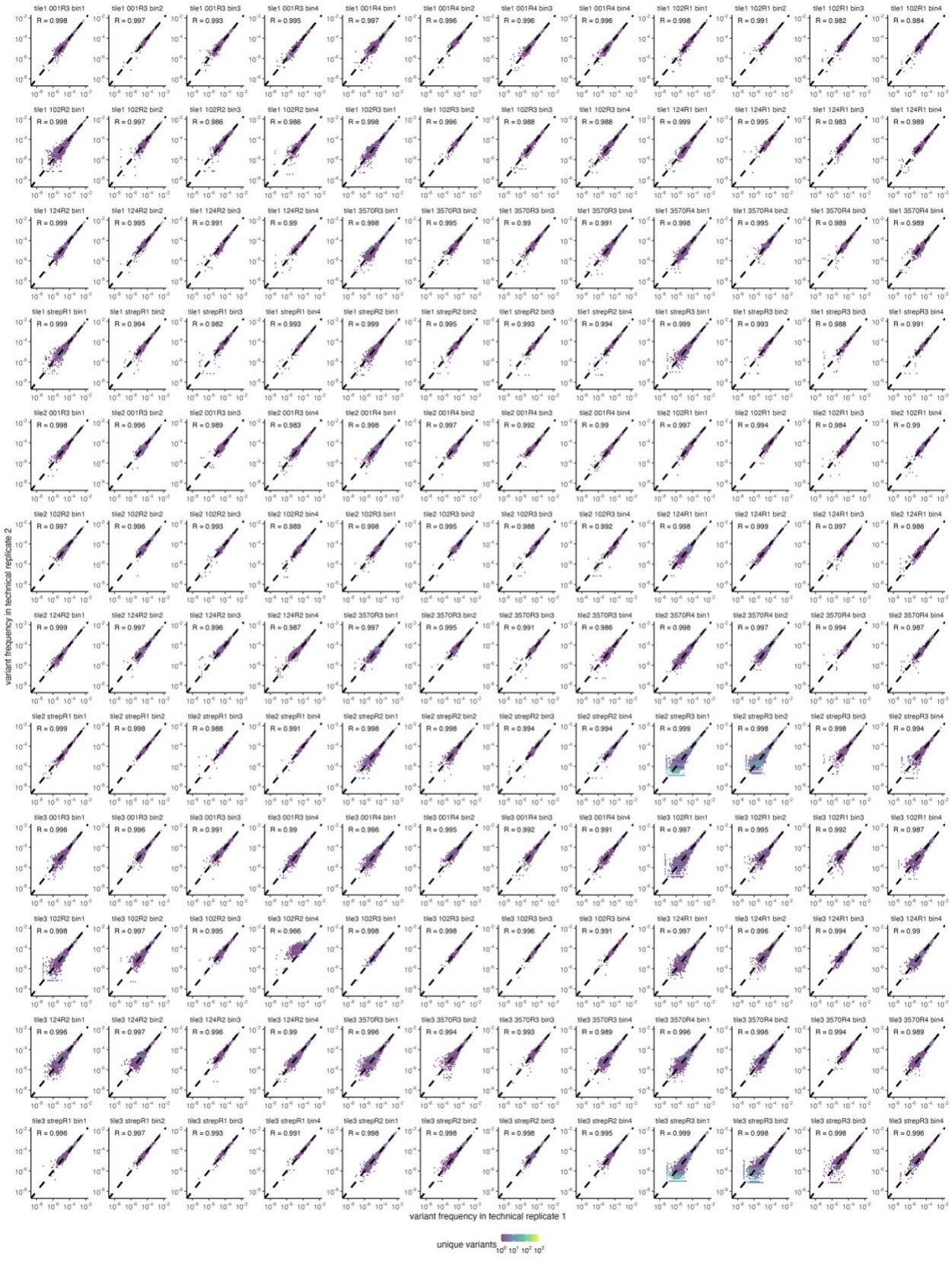
tile1 001R3 bin1 R = 0.998
tile1 001R3 bin2 R = 0.997
tile1 001R3 bin3 R = 0.993
tile1 001R3 bin4 R = 0.995
tile1 001R4 bin1 R = 0.997
tile1 001R4 bin2 R = 0.996
tile1 001R4 bin3 R = 0.996
tile1 001R4 bin4 R = 0.996
tile1 102R1 bin1 R = 0.998
tile1 102R1 bin2 R = 0.991
tile1 102R1 bin3 R = 0.982
tile1 102R1 bin4 R = 0.984
tile1 102R2 bin1 R = 0.998
tile1 102R2 bin2 R = 0.997
tile1 102R2 bin3 R = 0.986
tile1 102R2 bin4 R = 0.986
tile1 102R3 bin1 R = 0.998
tile1 102R3 bin2 R = 0.996
tile1 102R3 bin3 R = 0.988
tile1 102R3 bin4 R = 0.988
tile1 124R1 bin1 R = 0.999
tile1 124R1 bin2 R = 0.995
tile1 124R1 bin3 R = 0.983
tile1 124R1 bin4 R = 0.989
tile1 124R2 bin1 R = 0.999
tile1 124R2 bin2 R = 0.995
tile1 124R2 bin3 R = 0.991
tile1 124R2 bin4 R = 0.99
tile1 3570R3 bin1 R = 0.998
tile1 3570R3 bin2 R = 0.995
tile1 3570R3 bin3 R = 0.99
tile1 3570R3 bin4 R = 0.991
tile1 3570R4 bin1 R = 0.998
tile1 3570R4 bin2 R = 0.995
tile1 3570R4 bin3 R = 0.989
tile1 3570R4 bin4 R = 0.989
tile1 strepR1 bin1 R = 0.999
tile1 strepR1 bin2 R = 0.994
tile1 strepR1 bin3 R = 0.982
tile1 strepR1 bin4 R = 0.993
tile1 strepR2 bin1 R = 0.999
tile1 strepR2 bin2 R = 0.995
tile1 strepR2 bin3 R = 0.993
tile1 strepR2 bin4 R = 0.994
tile1 strepR3 bin1 R = 0.999
tile1 strepR3 bin2 R = 0.993
tile1 strepR3 bin3 R = 0.988
tile1 strepR3 bin4 R = 0.991
tile2 001R3 bin1 R = 0.998
tile2 001R3 bin2 R = 0.996
tile2 001R3 bin3 R = 0.989
tile2 001R3 bin4 R = 0.983
tile2 001R4 bin1 R = 0.998
tile2 001R4 bin2 R = 0.997
tile2 001R4 bin3 R = 0.992
tile2 001R4 bin4 R = 0.99
tile2 102R1 bin1 R = 0.997
tile2 102R1 bin2 R = 0.994
tile2 102R1 bin3 R = 0.984
tile2 102R1 bin4 R = 0.99
tile2 102R2 bin1 R = 0.997
tile2 102R2 bin2 R = 0.996
tile2 102R2 bin3 R = 0.993
tile2 102R2 bin4 R = 0.989
tile2 102R3 bin1 R = 0.998
tile2 102R3 bin2 R = 0.995
tile2 102R3 bin3 R = 0.988
tile2 102R3 bin4 R = 0.992
tile2 124R1 bin1 R = 0.998
tile2 124R1 bin2 R = 0.999
tile2 124R1 bin3 R = 0.997
tile2 124R1 bin4 R = 0.986
tile2 124R2 bin1 R = 0.999
tile2 124R2 bin2 R = 0.997
tile2 124R2 bin3 R = 0.996
tile2 124R2 bin4 R = 0.987
tile2 3570R3 bin1 R = 0.997
tile2 3570R3 bin2 R = 0.995
tile2 3570R3 bin3 R = 0.991
tile2 3570R3 bin4 R = 0.986
tile2 3570R4 bin1 R = 0.998
tile2 3570R4 bin2 R = 0.997
tile2 3570R4 bin3 R = 0.994
tile2 3570R4 bin4 R = 0.987
tile2 strepR1 bin1 R = 0.999
tile2 strepR1 bin2 R = 0.998
tile2 strepR1 bin3 R = 0.988
tile2 strepR1 bin4 R = 0.991
tile2 strepR2 bin1 R = 0.998
tile2 strepR2 bin2 R = 0.998
tile2 strepR2 bin3 R = 0.994
tile2 strepR2 bin4 R = 0.994
tile2 strepR3 bin1 R = 0.999
tile2 strepR3 bin2 R = 0.998
tile2 strepR3 bin3 R = 0.998
tile2 strepR3 bin4 R = 0.994
tile3 001R3 bin1 R = 0.996
tile3 001R3 bin2 R = 0.996
tile3 001R3 bin3 R = 0.991
tile3 001R3 bin4 R = 0.99
tile3 001R4 bin1 R = 0.996
tile3 001R4 bin2 R = 0.995
tile3 001R4 bin3 R = 0.992
tile3 001R4 bin4 R = 0.991
tile3 102R1 bin1 R = 0.997
tile3 102R1 bin2 R = 0.995
tile3 102R1 bin3 R = 0.992
tile3 102R1 bin4 R = 0.987
tile3 102R2 bin1 R = 0.998
tile3 102R2 bin2 R = 0.997
tile3 102R2 bin3 R = 0.995
tile3 102R2 bin4 R = 0.966
tile3 102R3 bin1 R = 0.998
tile3 102R3 bin2 R = 0.998
tile3 102R3 bin3 R = 0.996
tile3 102R3 bin4 R = 0.991
tile3 124R1 bin1 R = 0.997
tile3 124R1 bin2 R = 0.996
tile3 124R1 bin3 R = 0.994
tile3 124R1 bin4 R = 0.99
tile3 124R2 bin1 R = 0.996
tile3 124R2 bin2 R = 0.997
tile3 124R2 bin3 R = 0.996
tile3 124R2 bin4 R = 0.99
tile3 3570R3 bin1 R = 0.996
tile3 3570R3 bin2 R = 0.994
tile3 3570R3 bin3 R = 0.993
tile3 3570R3 bin4 R = 0.989
tile3 3570R4 bin1 R = 0.996
tile3 3570R4 bin2 R = 0.996
tile3 3570R4 bin3 R = 0.994
tile3 3570R4 bin4 R = 0.989
tile3 strepR1 bin1 R = 0.996
tile3 strepR1 bin2 R = 0.997
tile3 strepR1 bin3 R = 0.993
tile3 strepR1 bin4 R = 0.991
tile3 strepR2 bin1 R = 0.998
tile3 strepR2 bin2 R = 0.998
tile3 strepR2 bin3 R = 0.998
tile3 strepR2 bin4 R = 0.995
tile3 strepR3 bin1 R = 0.999
tile3 strepR3 bin2 R = 0.998
tile3 strepR3 bin3 R = 0.998
tile3 strepR3 bin4 R = 0.996
variant frequency in technical replicate 2
variant frequency in technical replicate 1
unique variants

Figure A.8: Amplification and sequencing technical replicates

Scatterplots comparing variant frequency in each sequencing sample across all experiments. Points represent binned densities where the color indicates the number of variants that fall within that bin. Dashed line indicates perfect correlation. R = Pearson's r.).

Antibody	Catalog number	Target	[Stock] (mg/mL)	Dilution
GMA-001	GMA-001	FIX GLA domain, conformation and γ-carboxylation sensitive	1	100
GMA-102	GMA-102	FIX heavy chain	1	100
GMA-124	GMA-124	FIX light chain	1	100
GMA-134	GMA-134	FIX activation peptide	1	1000
3570	3570	γ-carboxylation	1	500
Strep	76950	Strep II tag	1	100
Goat anti-mouse Alexa 647	ab150115	Mouse IgG	2	200
Donkey anti-rabbit Alexa 488 preabsorbed	ab150061	Rabbit IgG	2	100

Table A.1: Antibody concentration for MultiSTEP experiments

Appendix B. Chapter 3 supplementary material

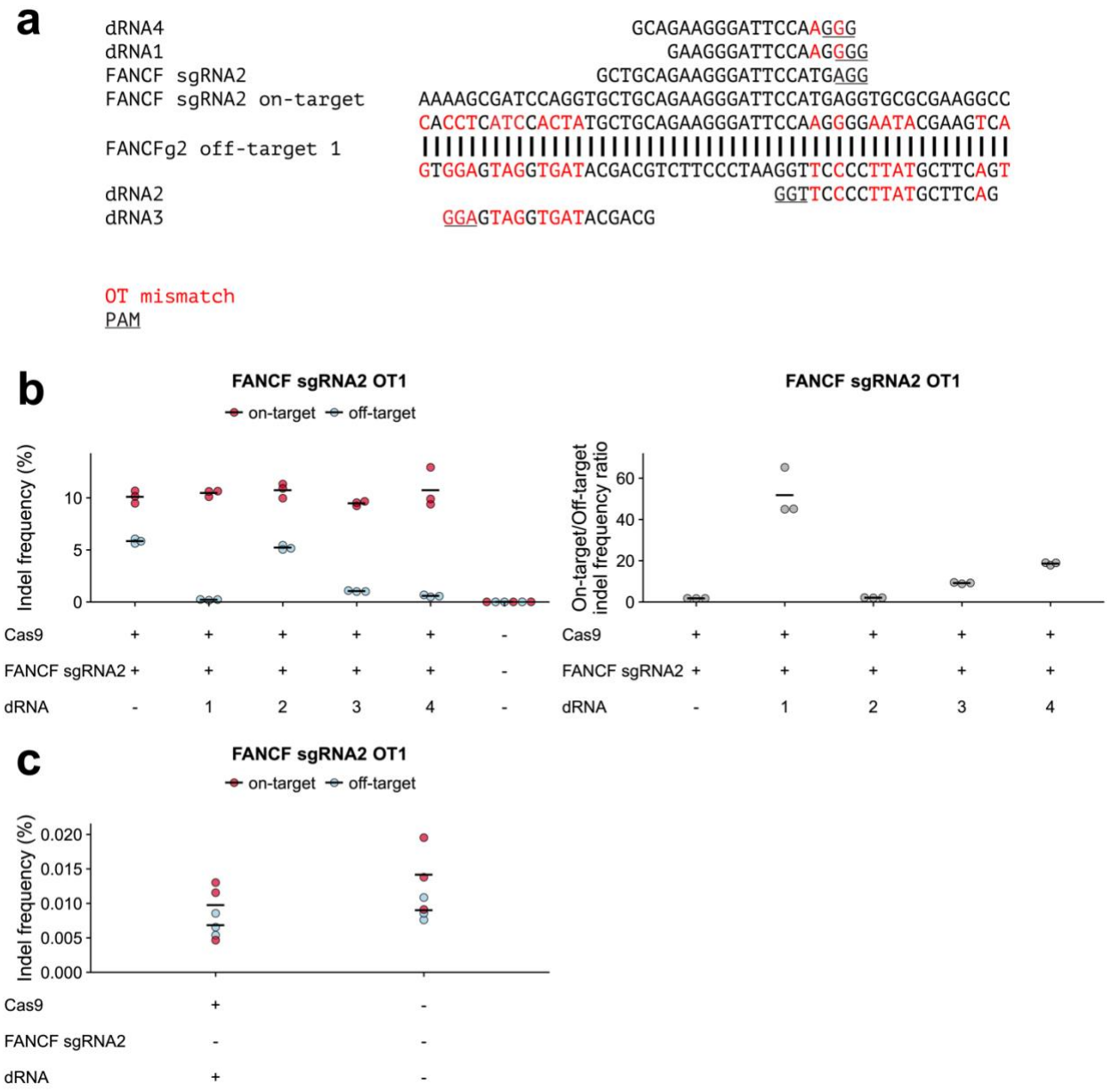

Figure B.1: *FANCF* dRNA1 does not promote Cas9-mediated editing

a. Sequence alignment of *FANCF* sgRNA2, its on-target site, the most prominent off-target, off-target site 1 (OT1), and multiple dRNAs complementary to OT1. Black arrows indicate best dRNA, as determined by maximal off-target editing suppression with minimal on-target editing suppression. **b.** Indel frequencies and specificity ratios (on-target/off-target indel frequency ratios) at the *FANCF* sgRNA2 on-target site and OT1 24 hours after transfection with Cas9, sgRNA, and various dRNAs. For conditions without dRNA, an

equivalent amount of pMAX-GFP was substituted. **c.** Indel frequencies at the *FANCF* sgRNA2 on-target and OT1 sites 24 hours after transfection with Cas9 and dRNA1 but no sgRNA. The predicted cut sites of dRNA1 are the same as *FANCF* sgRNA2. Indel frequencies for untransfected cells are shown as a control. Numbers denote dRNA identity. Solid lines denote the mean of n = 3 biological replicates.

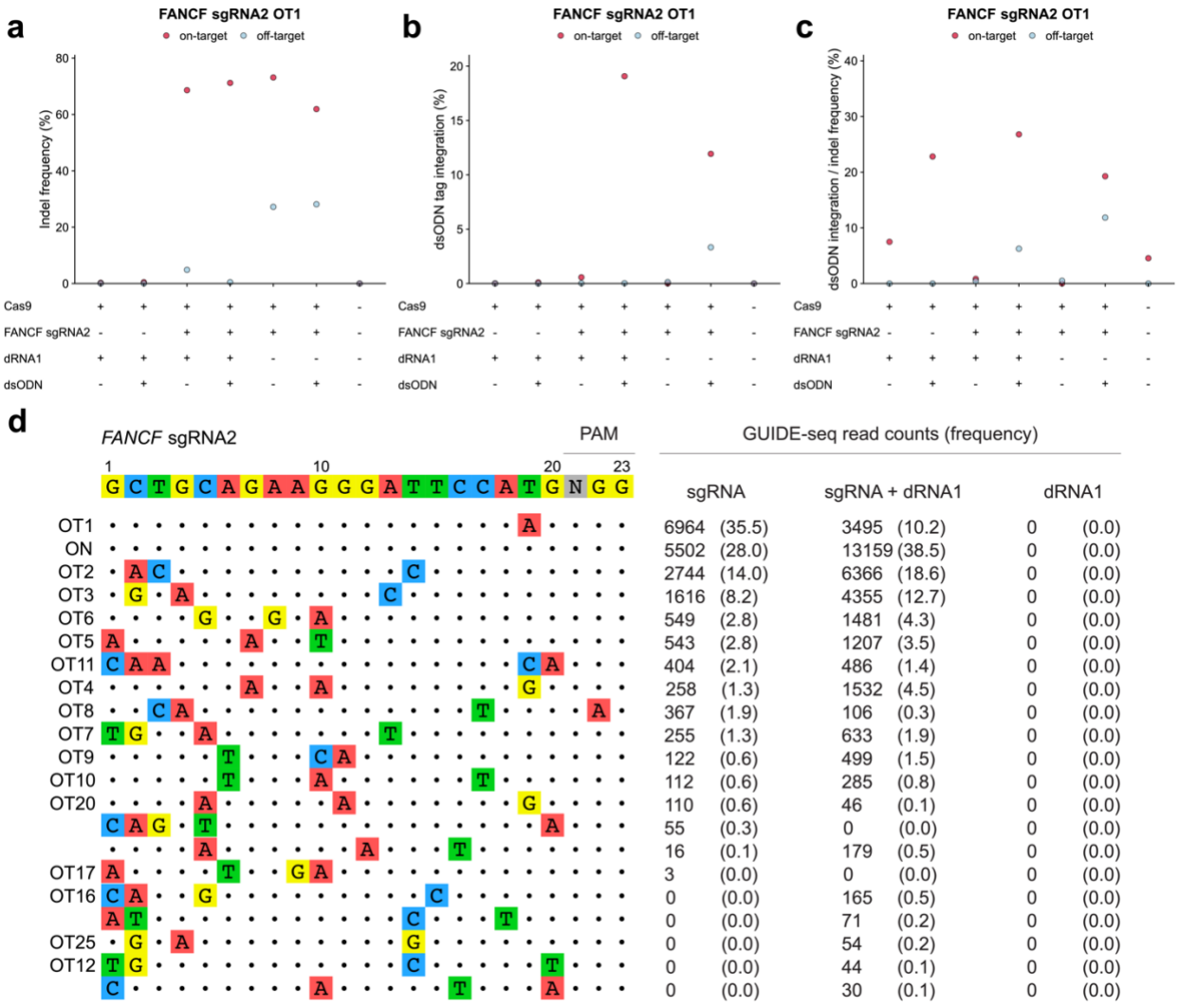

Figure B.2: Genome-wide assessment of DNA cleavage using dRNAs

a. Indel frequency at *FANCF* sgRNA2 on-target site and OT1 96 hours after electroporation with plasmids encoding Cas9, sgRNA and dRNA1 in U2OS cells. **b.** Integration of an end-protected double-stranded oligonucleotide (dsODN) 96 hours after electroporation with Cas9, sgRNA and dRNA1 in U2OS cells. **c.** dsODN tag integration efficiency ratio (integration:indel) 96 hours after electroporation. Indel frequencies

and dsODN tag integration for untransfected cells are shown as a control. **d.** GUIDE-seq genome-wide specificity profiles for Cas9 paired with *FANCF* sgRNA2, FANCF dRNA1 or both allowing for up to 8 mismatches from on-target sequence. Mismatched positions in off-target sites are highlighted by color. Black dots indicate matched nucleotide to on-target site. GUIDE-seq counts and frequencies are shown to the right of the sequences. Off-target numbering system to the left of sequences adapted from Kleinstiver, et al.[154] and **Figure** 3.1.

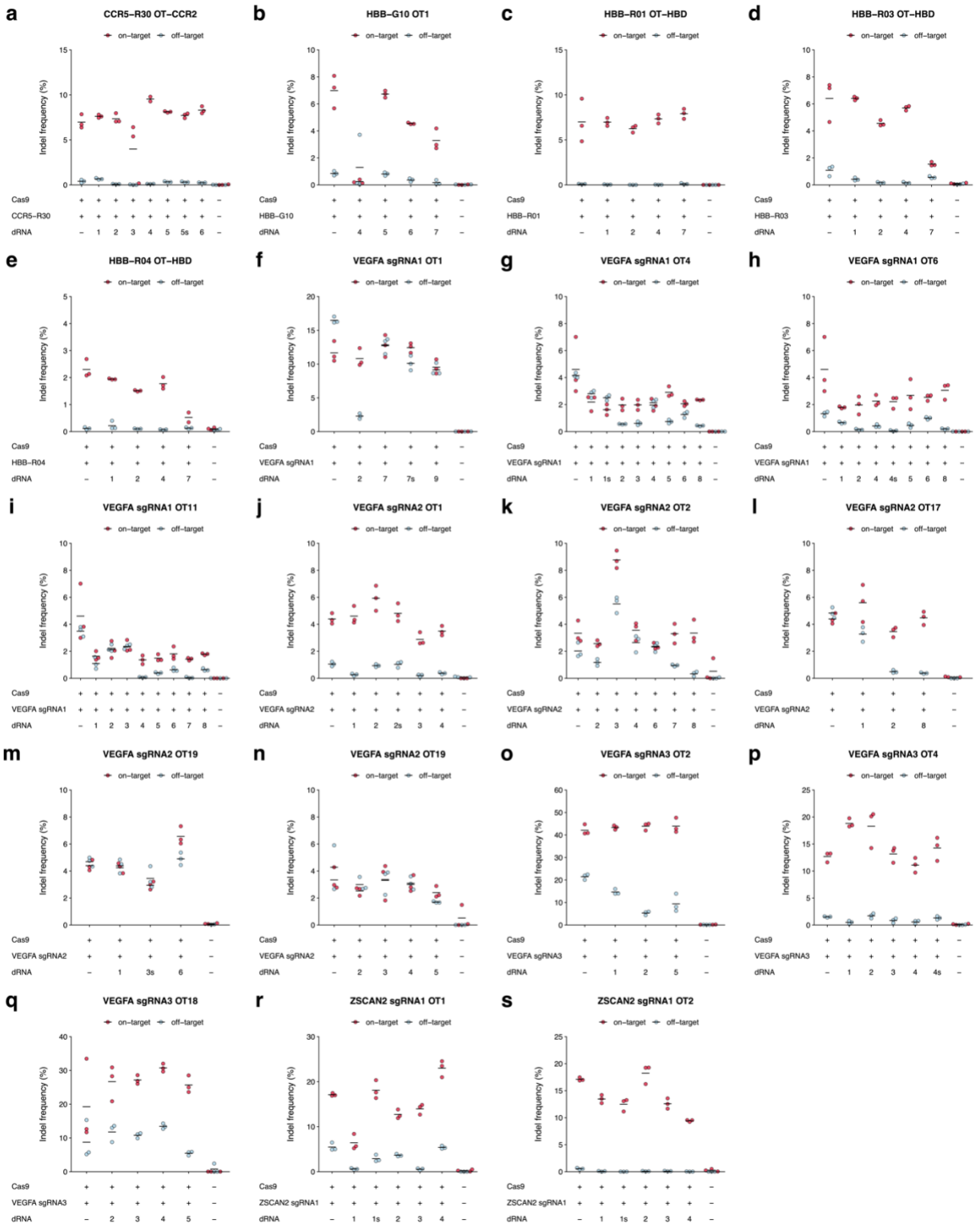
a
CCR5–R30 OT–CCR2
b
HBB–G10 OT1
c
HBB–R01 OT–HBD
d
HBB–R03 OT–HBD
e
HBB–R04 OT–HBD
f
VEGFA sgRNA1 OT1
g
VEGFA sgRNA1 OT4
h
VEGFA sgRNA1 OT6
i
VEGFA sgRNA1 OT11
j
VEGFA sgRNA2 OT1
k
VEGFA sgRNA2 OT2
l
VEGFA sgRNA2 OT17
m
VEGFA sgRNA2 OT19
n
VEGFA sgRNA2 OT19
o
VEGFA sgRNA3 OT2
p
VEGFA sgRNA3 OT4
q
VEGFA sgRNA3 OT18
r
ZSCAN2 sgRNA1 OT1
s
ZSCAN2 sgRNA1 OT2
on–target
off–target
Indel frequency (%)
Cas9
CCR5–R30
HBB–G10
HBB–R01
HBB–R03
HBB–R04
VEGFA sgRNA1
VEGFA sgRNA2
VEGFA sgRNA3
ZSCAN2 sgRNA1
dRNA

Figure B.3: dRNAs screened to increase the specificity ratios of 18 additional on-target/off-target pairs

On and off-target indel frequencies 24 hours after transfection with Cas9, sgRNA, and off-target specific dRNAs in HEK293T cells **a.** *CCR5*-R30 OT (*CCR2*). **b.** *HBB*-G10 OT1. 4 additional dRNAs were screened, which are not shown here. **c.** *HBB*-R01 OT (*HBD*). **d.** *HBB*-R03 OT (*HBD*). **e.** *HBB*-R04 OT (*HBD*). **f.** *VEGFA* sgRNA1 OT1. **g.** *VEGFA* sgRNA1 OT4. **h.** *VEGFA* sgRNA1 OT6. **i.** *VEGFA* sgRNA1 OT11. **j.** *VEGFA* sgRNA2 OT1. **k.** *VEGFA* sgRNA2 OT2. **l.** *VEGFA* sgRNA2 OT17. **m.** *VEGFA* sgRNA2 OT19 (dRNAs 1, 3s, and 6). **n.** *VEGFA* sgRNA2 OT19 (dRNAs 2-5). **o.** *VEGFA* sgRNA3 OT2. **p.** *VEGFA* sgRNA3 OT4. **q.** *VEGFA* sgRNA3 OT18. **r.** *ZSCAN2* sgRNA1 OT1. **s.** *ZSCAN2* sgRNA1 OT2. Indel frequencies for untransfected cells are shown as a control. Numbers denote dRNA identity. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

Figure B.4: Alignments of on-target sites and dRNAs to off-target sites

Sequence alignments for all off-targets used in this study (For *FANCF* sgRNA2, see **Figure** 3.1) and their corresponding dRNAs and on-targets. Mismatches in the off-target and dRNA sequences relative to the on-target sequence are displayed in red. Mismatched 5' guanines in dRNAs are displayed in cyan. PAM sequences are underlined. Black arrows indicate best dRNA, as determined by maximal off-target editing suppression with minimal on-target editing suppression.

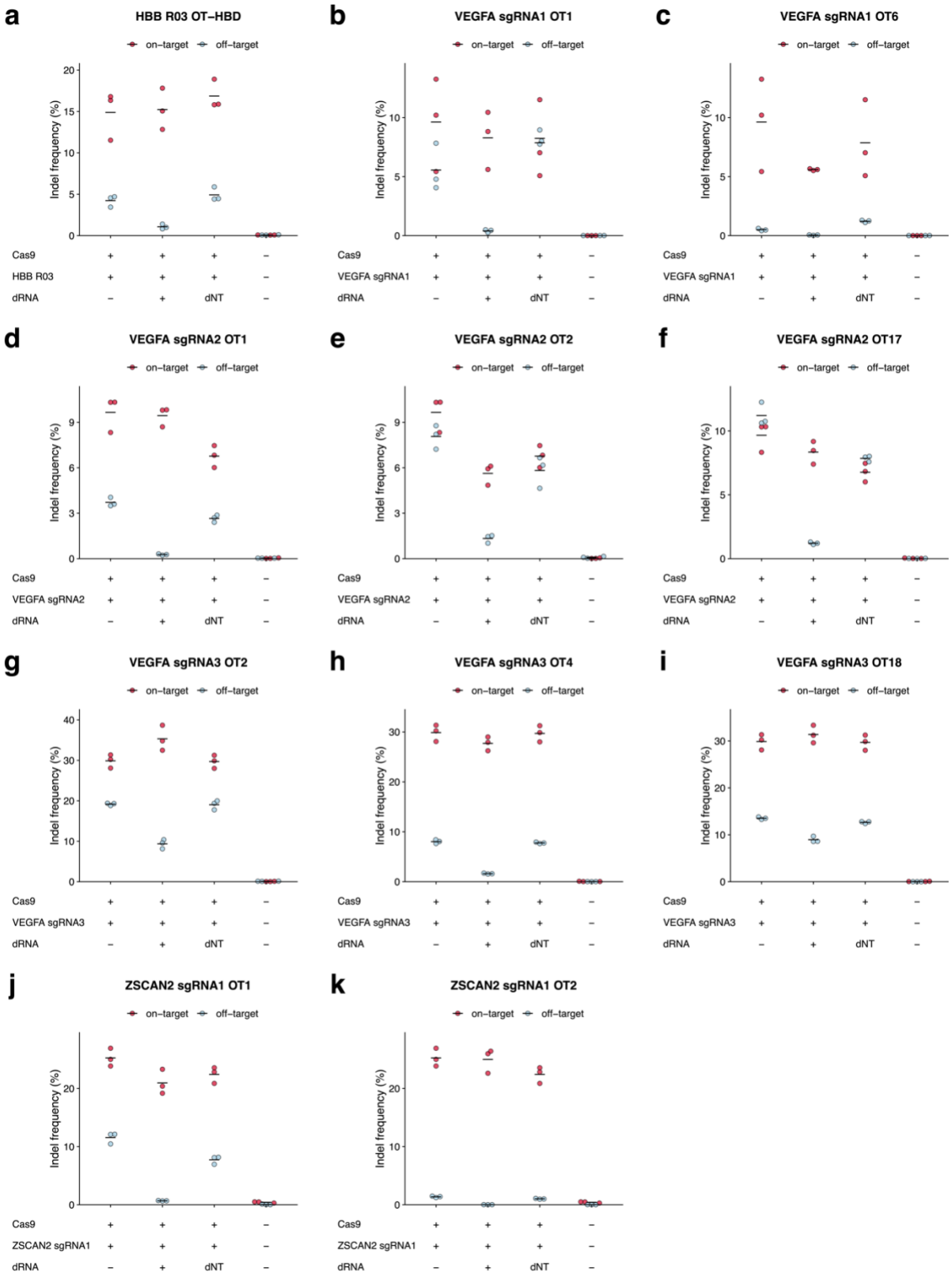
a
HBB R03 OT–HBD
b
VEGFA sgRNA1 OT1
c
VEGFA sgRNA1 OT6
d
VEGFA sgRNA2 OT1
e
VEGFA sgRNA2 OT2
f
VEGFA sgRNA2 OT17
g
VEGFA sgRNA3 OT2
h
VEGFA sgRNA3 OT4
i
VEGFA sgRNA3 OT18
j
ZSCAN2 sgRNA1 OT1
k
ZSCAN2 sgRNA1 OT2
on–target
off–target
Indel frequency (%)
Cas9
HBB R03
VEGFA sgRNA1
VEGFA sgRNA2
VEGFA sgRNA3
ZSCAN2 sgRNA1
dRNA
dNT

Figure B.5: Nontargeting dRNAs have minimal effects on on-target and off-target editing

Comparison of most effective dRNA for 12 different off-target loci with a nontargeting dRNA (dNT). Indel frequency of on-target and off-target loci 24 hours after transfection with Cas9, sgRNA, ± dRNA or nontargeting dRNA in HEK-293T cells. Indel frequencies for untransfected cells are shown as a control. **a.** *HBB* R03 OT (*HBD*). **b.** *VEGFA* sgRNA1 OT1. **c.** *VEGFA* sgRNA1 OT6. **d.** *VEGFA* sgRNA2 OT1. **e.** *VEGFA* sgRNA2 OT2. **f.** *VEGFA* sgRNA2 OT17. **g.** *VEGFA* sgRNA3 OT2. **h.** *VEGFA* sgRNA3 OT4. **i.** *VEGFA* sgRNA3 OT18. **j.** *ZSCAN2* sgRNA1 OT1. **k.** *ZSCAN2* sgRNA1 OT2. Numbers denote dRNA identity. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

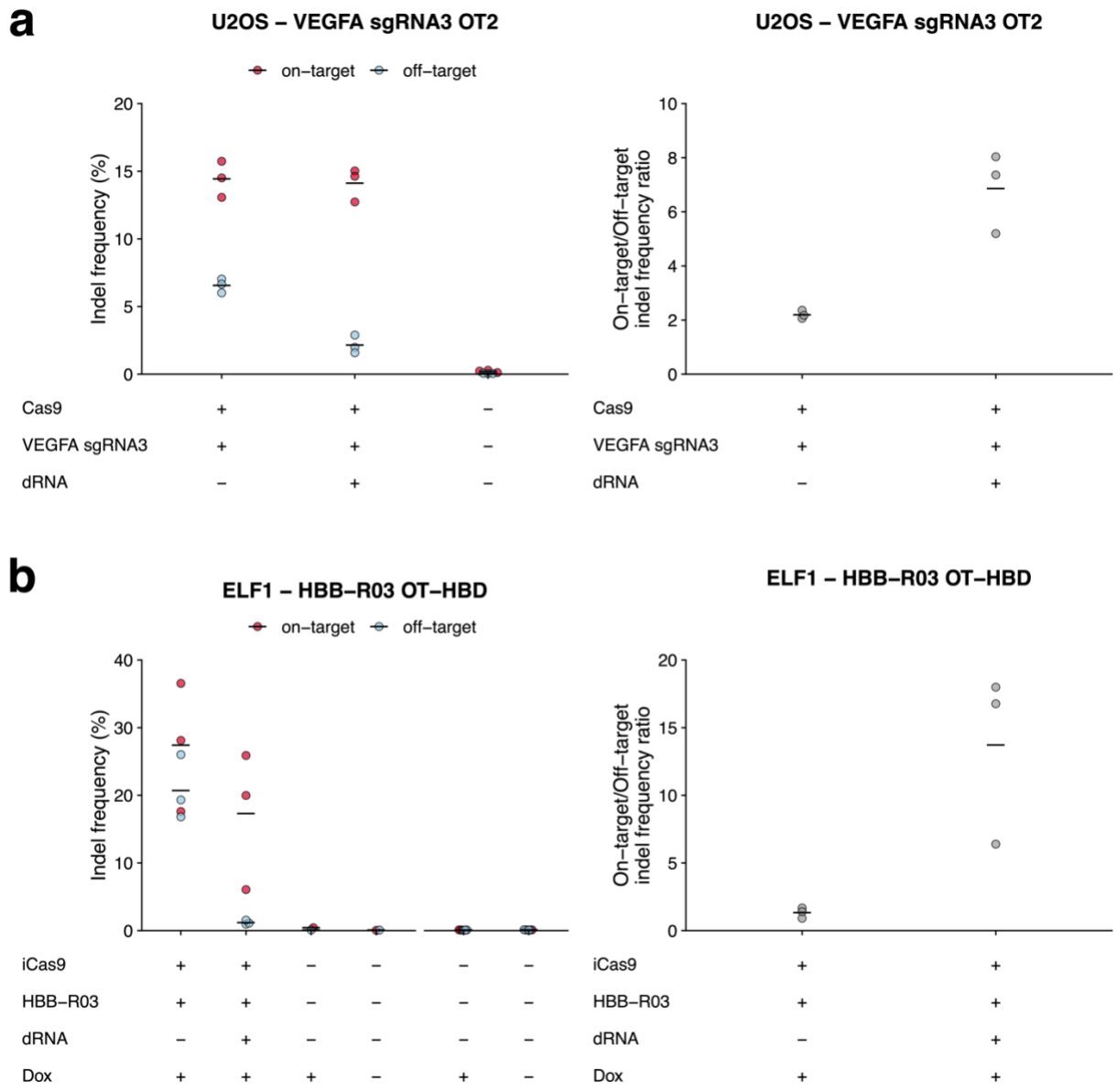

Figure B.6: dOTS is effective in multiple cell types

On-target and off-target indel frequencies and specificity ratios 24 hours after transfection with Cas9, sgRNA, and off-target specific dRNA. iCas9 denotes stable integration of Cas9 under the control of a doxycycline-inducible promoter. **a.** *VEGFA* sgRNA3 OT2 in U2OS cells. **b.** *HBB* R03 OT (*HBD*) in Elf1 cells. Indel frequencies for untransfected cells are shown as a control. Control samples to the right of the x-axis break were performed separately. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

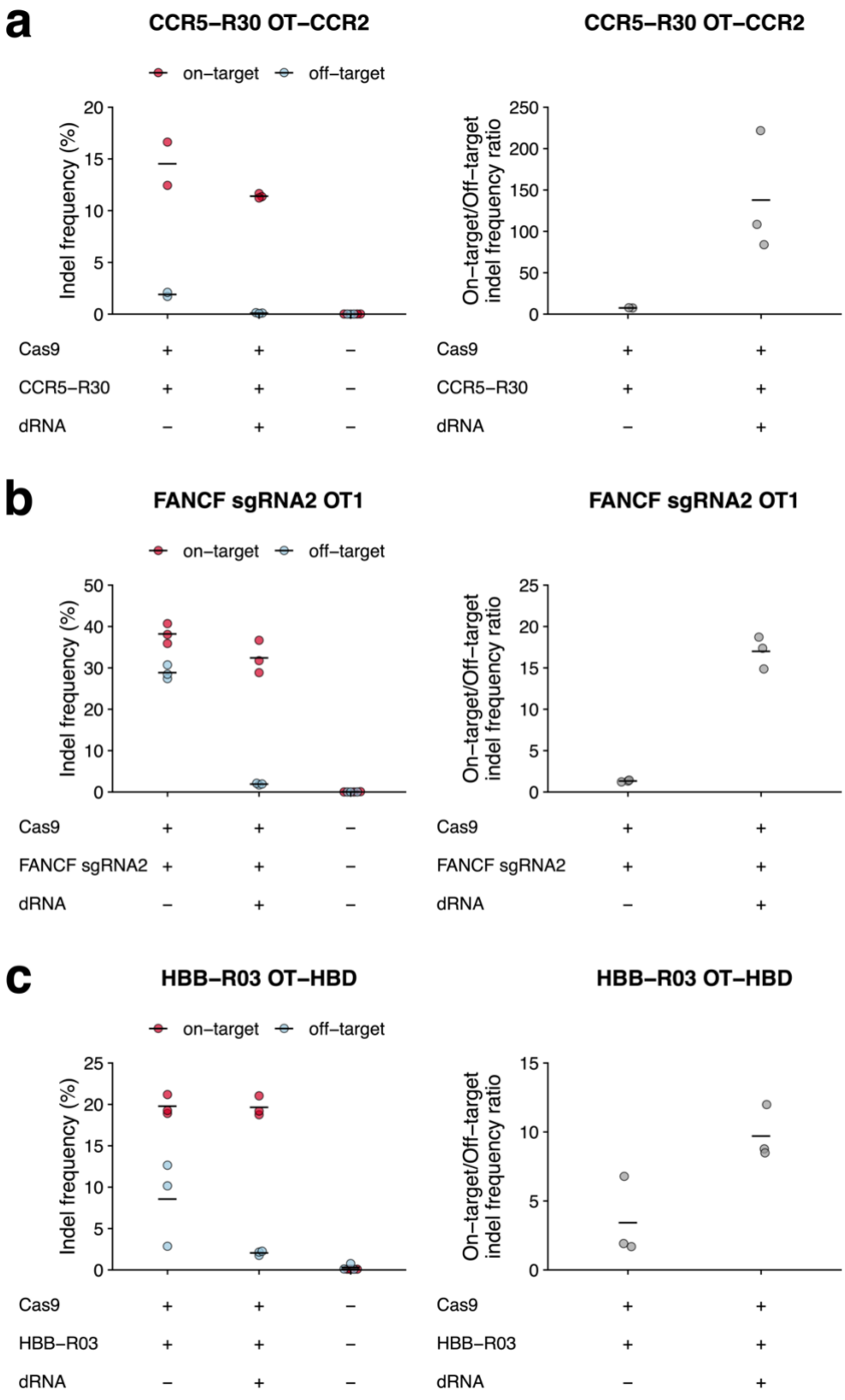
a
CCR5–R30 OT–CCR2
on–target
off–target
Indel frequency (%)
20
15
10
5
0
Cas9 + + –
CCR5–R30 + + –
dRNA – + –
CCR5–R30 OT–CCR2
On–target/Off–target indel frequency ratio
250
200
150
100
50
0
Cas9 + +
CCR5–R30 + +
dRNA – +
b
FANCF sgRNA2 OT1
on–target
off–target
Indel frequency (%)
50
40
30
20
10
0
Cas9 + + –
FANCF sgRNA2 + + –
dRNA – + –
FANCF sgRNA2 OT1
On–target/Off–target indel frequency ratio
25
20
15
10
5
0
Cas9 + +
FANCF sgRNA2 + +
dRNA – +
c
HBB–R03 OT–HBD
on–target
off–target
Indel frequency (%)
25
20
15
10
5
0
Cas9 + + –
HBB–R03 + + –
dRNA – + –
HBB–R03 OT–HBD
On–target/Off–target indel frequency ratio
15
10
5
0
Cas9 + +
HBB–R03 + +
dRNA – +

Figure B.7: dRNA-mediated off-target suppression is durable

On-target and off-target indel frequencies and specificity ratios 72 hours after transfection with Cas9, sgRNA, and off-target specific dRNAs in HEK-293T cells **a.** *CCR5*-R30 OT (*CCR2*). **b.** *FANCF* sgRNA2 OT1. **c.** *HBB*-R03 OT (*HBD*). Indel frequencies for untransfected cells are shown as a control. Numbers denote dRNA identity. Solid lines denote the mean of n = 3 biological replicates, except *CCR5*-R30 without dRNA where n = 2. 24 hour comparison shown in **Appendix B: Figure B.**3.

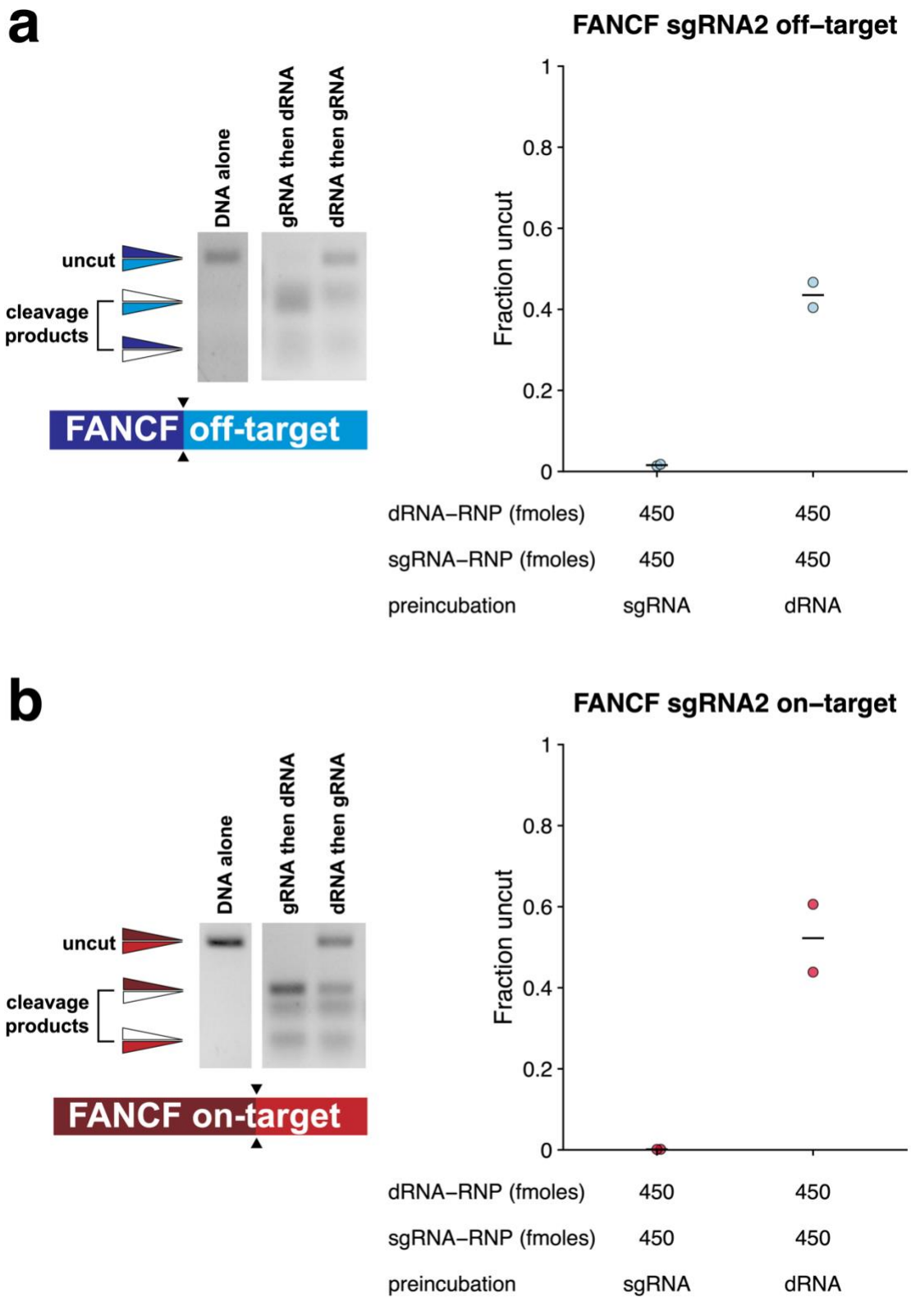
a
DNA alone
gRNA then dRNA
dRNA then gRNA
uncut
cleavage products
FANCF off-target
FANCF sgRNA2 off–target
Fraction uncut
1
0.8
0.6
0.4
0.2
0
dRNA–RNP (fmoles) 450 450
sgRNA–RNP (fmoles) 450 450
preincubation sgRNA dRNA
b
DNA alone
gRNA then dRNA
dRNA then gRNA
uncut
cleavage products
FANCF on-target
FANCF sgRNA2 on–target
Fraction uncut
1
0.8
0.6
0.4
0.2
0
dRNA–RNP (fmoles) 450 450
sgRNA–RNP (fmoles) 450 450
preincubation sgRNA dRNA

Figure B.8: dRNAs and sgRNAs compete for target site occupancy

Representative gels of *in vitro* Cas9 *FANCF* sgRNA2 RNP cleavage of linear PCR products containing either **a.** the *FANCF* sgRNA2 off-target site (OT1) or **b.** the *FANCF* sgRNA2 on-target site. PCR products were either preincubated with the sgRNA-RNP complex for 10 minutes prior to addition of the dRNA-RNP complex (sgRNA then dRNA) or were preincubated with the dRNA-RNP complex for 10 minutes prior to addition of the sgRNA-RNP complex (dRNA then sgRNA). ImageJ was used to quantify the intensity of the uncut and all cut bands in each lane. Fraction uncut was determined by dividing uncut intensity by sum of all band intensities in each lane. Solid lines denote the mean of n = 2 biological replicates.

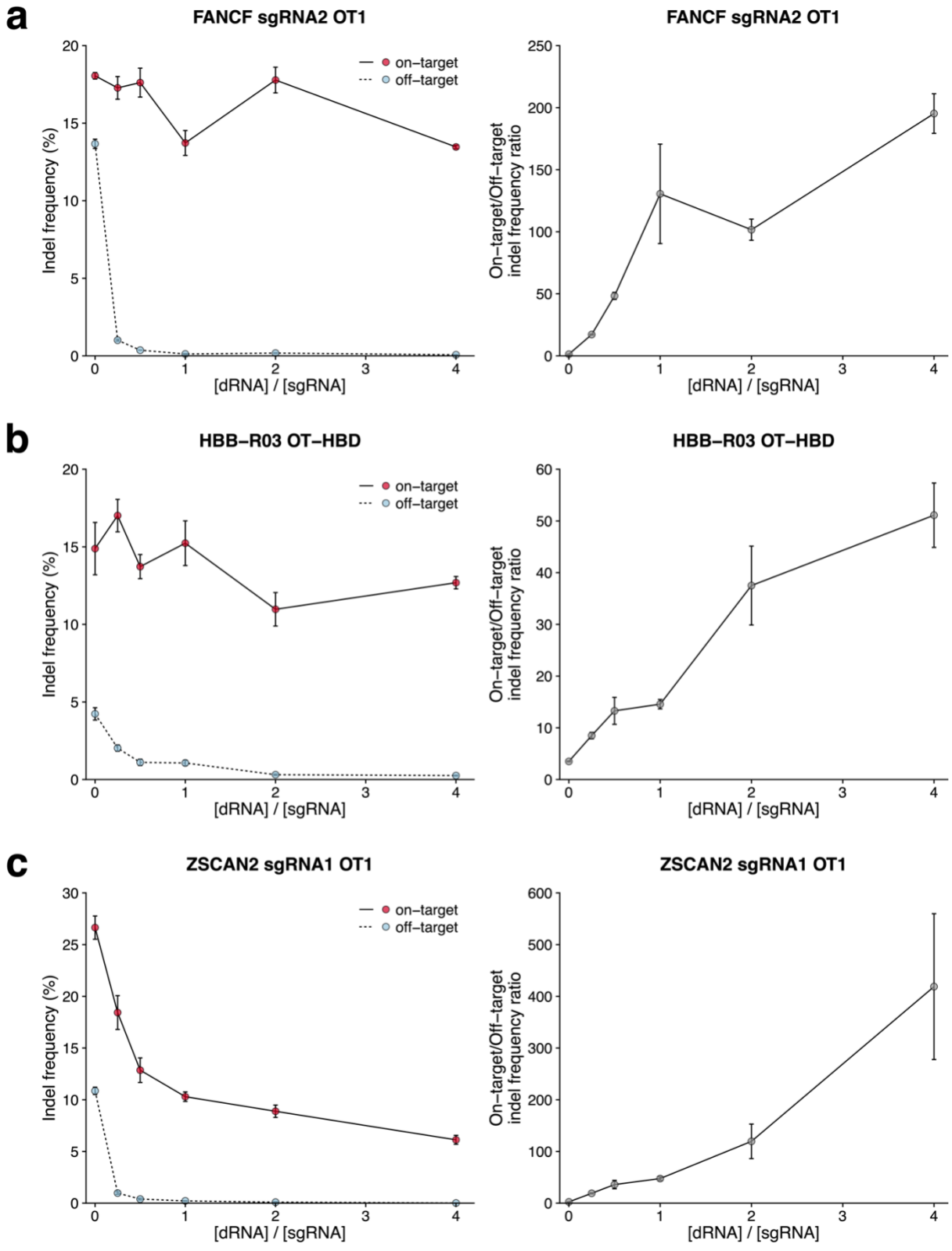
a
FANCF sgRNA2 OT1
on−target
off−target
Indel frequency (%)
[dRNA] / [sgRNA]
FANCF sgRNA2 OT1
On−target/Off−target indel frequency ratio
[dRNA] / [sgRNA]
b
HBB−R03 OT−HBD
on−target
off−target
Indel frequency (%)
[dRNA] / [sgRNA]
HBB−R03 OT−HBD
On−target/Off−target indel frequency ratio
[dRNA] / [sgRNA]
c
ZSCAN2 sgRNA1 OT1
on−target
off−target
Indel frequency (%)
[dRNA] / [sgRNA]
ZSCAN2 sgRNA1 OT1
On−target/Off−target indel frequency ratio
[dRNA] / [sgRNA]

Figure B.9: Titration of dRNAs further reduces unwanted off-target editing at additional sites

Target and off-target indel frequencies and specificity ratios 24 hours after transfection of various dRNA/sgRNA plasmid ratios for **a.** *FANCF* sgRNA 2 and dRNA1; **b.** *HBB* R03 and dRNA4; **c.** *ZSCAN2* sgRNA1 and dRNA3. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

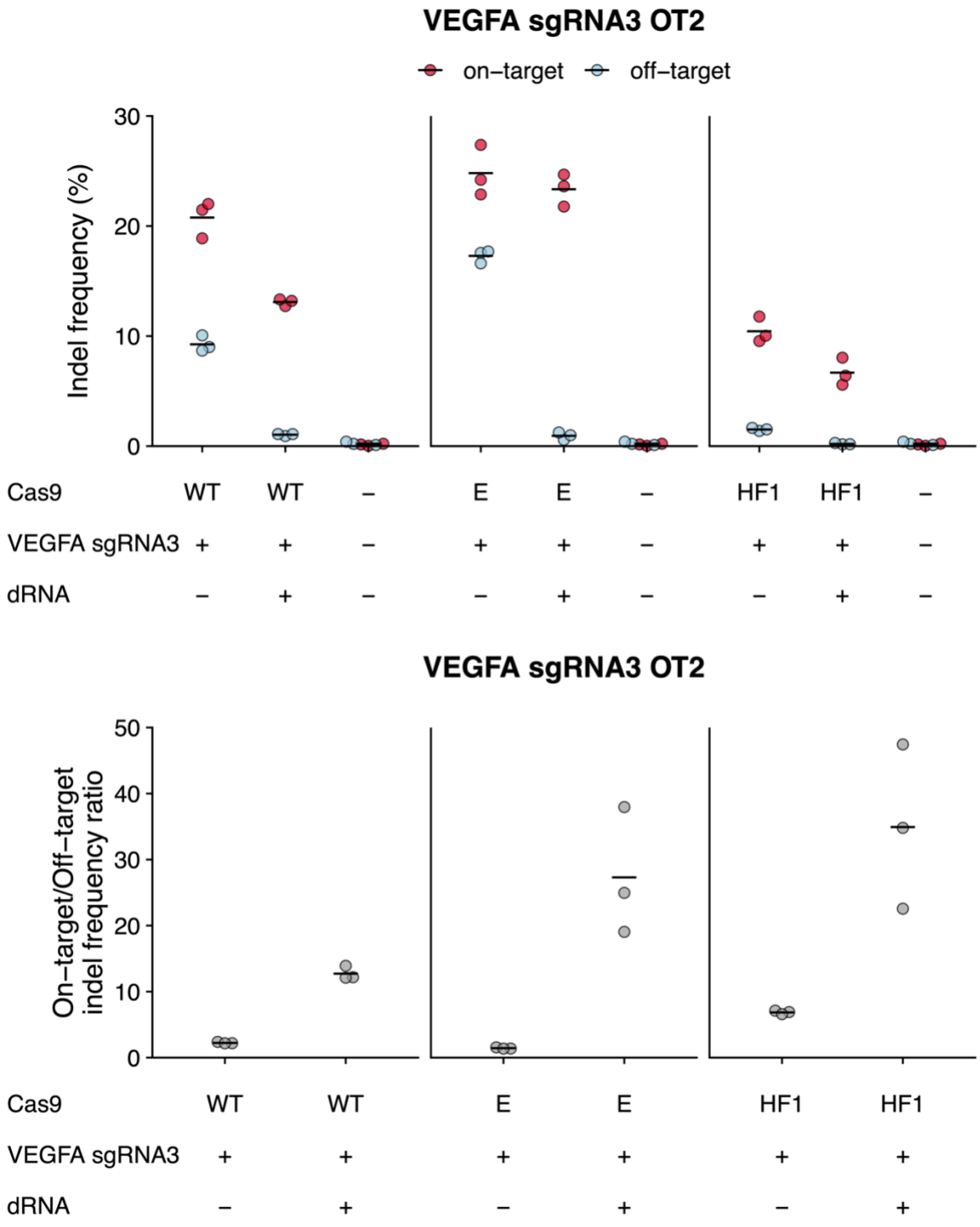

Figure B.10: dOTS can suppress refractory off-target editing of high-specificity Cas9 variants

On-target and off-target indel frequencies and specificity ratios 24 hours after transfection of plasmids encoding *VEGFA* sgRNA3, dRNA and either wildtype Cas9 (WT), espCas9 (E), or spCas9-HF1 (HF1). Indel frequencies for untransfected cells are shown as a control. Numbers denote dRNA identity. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

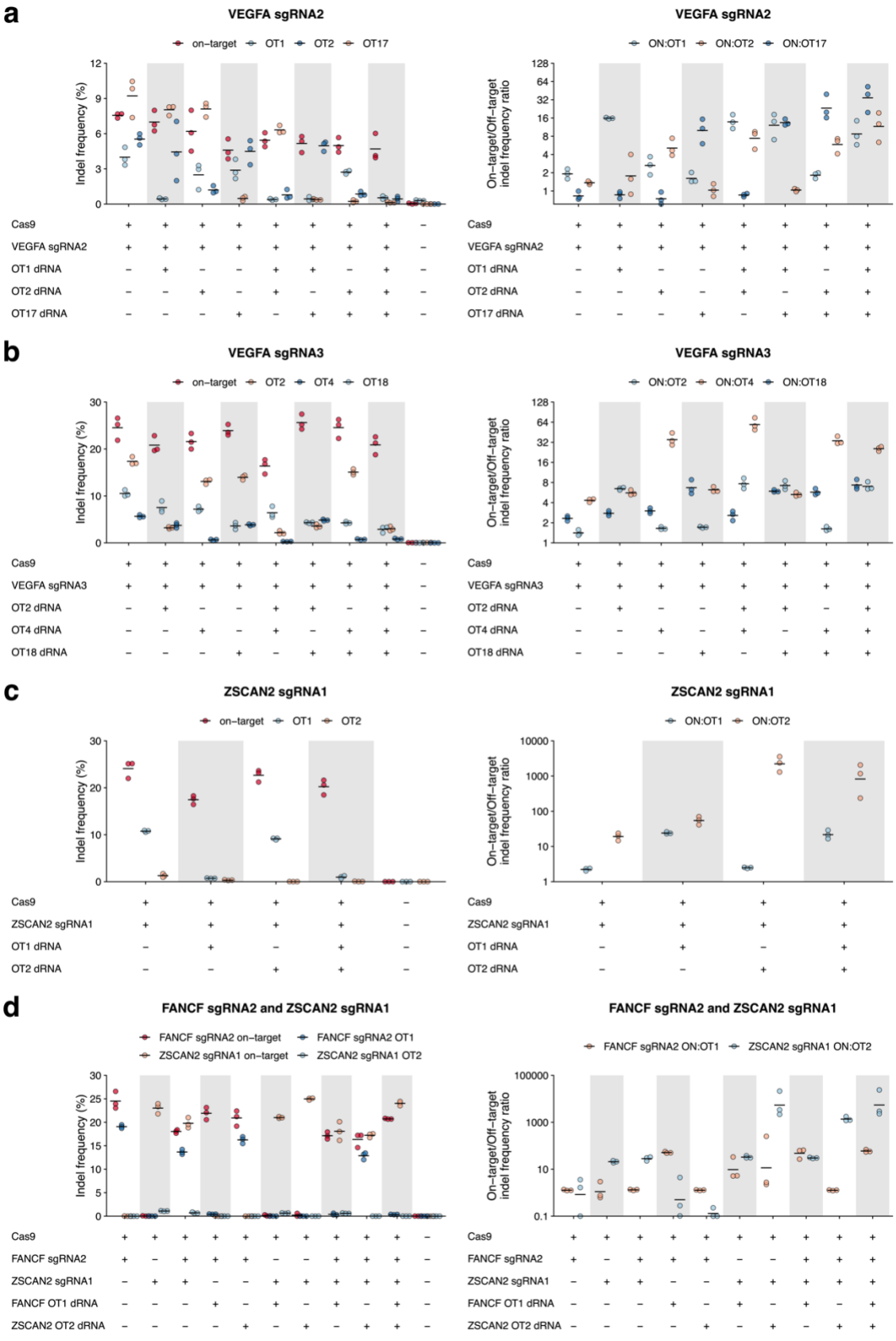
a
VEGFA sgRNA2
on−target
OT1
OT2
OT17
Indel frequency (%)
Cas9
VEGFA sgRNA2
OT1 dRNA
OT2 dRNA
OT17 dRNA
VEGFA sgRNA2
ON:OT1
ON:OT2
ON:OT17
On−target/Off−target indel frequency ratio
b
VEGFA sgRNA3
on−target
OT2
OT4
OT18
Indel frequency (%)
Cas9
VEGFA sgRNA3
OT2 dRNA
OT4 dRNA
OT18 dRNA
VEGFA sgRNA3
ON:OT2
ON:OT4
ON:OT18
On−target/Off−target indel frequency ratio
c
ZSCAN2 sgRNA1
on−target
OT1
OT2
Indel frequency (%)
Cas9
ZSCAN2 sgRNA1
OT1 dRNA
OT2 dRNA
ZSCAN2 sgRNA1
ON:OT1
ON:OT2
On−target/Off−target indel frequency ratio
d
FANCF sgRNA2 and ZSCAN2 sgRNA1
FANCF sgRNA2 on−target
FANCF sgRNA2 OT1
ZSCAN2 sgRNA1 on−target
ZSCAN2 sgRNA1 OT2
Indel frequency (%)
Cas9
FANCF sgRNA2
ZSCAN2 sgRNA1
FANCF OT1 dRNA
ZSCAN2 OT2 dRNA
FANCF sgRNA2 and ZSCAN2 sgRNA1
FANCF sgRNA2 ON:OT1
ZSCAN2 sgRNA1 ON:OT2
On−target/Off−target indel frequency ratio

Figure B.11: dRNAs can be combined to suppress unwanted off-target editing at a variety of sites

On-target and off-target indel frequencies and specificity ratios 24 hours after transfection with plasmids encoding Cas9 and various combinations of sgRNAs and dRNAs at **a.** *VEGFA* sgRNA2 OT1, OT2, and OT17; **b.** *VEGFA* sgRNA3 OT2, OT4, and OT18; **c.** *ZSCAN2* sgRNA1 OT1 and OT2; **d.** *FANCF* sgRNA2 OT1 and *ZSCAN2* sgRNA1 OT2. Indel frequencies for untransfected cells are shown as a control. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

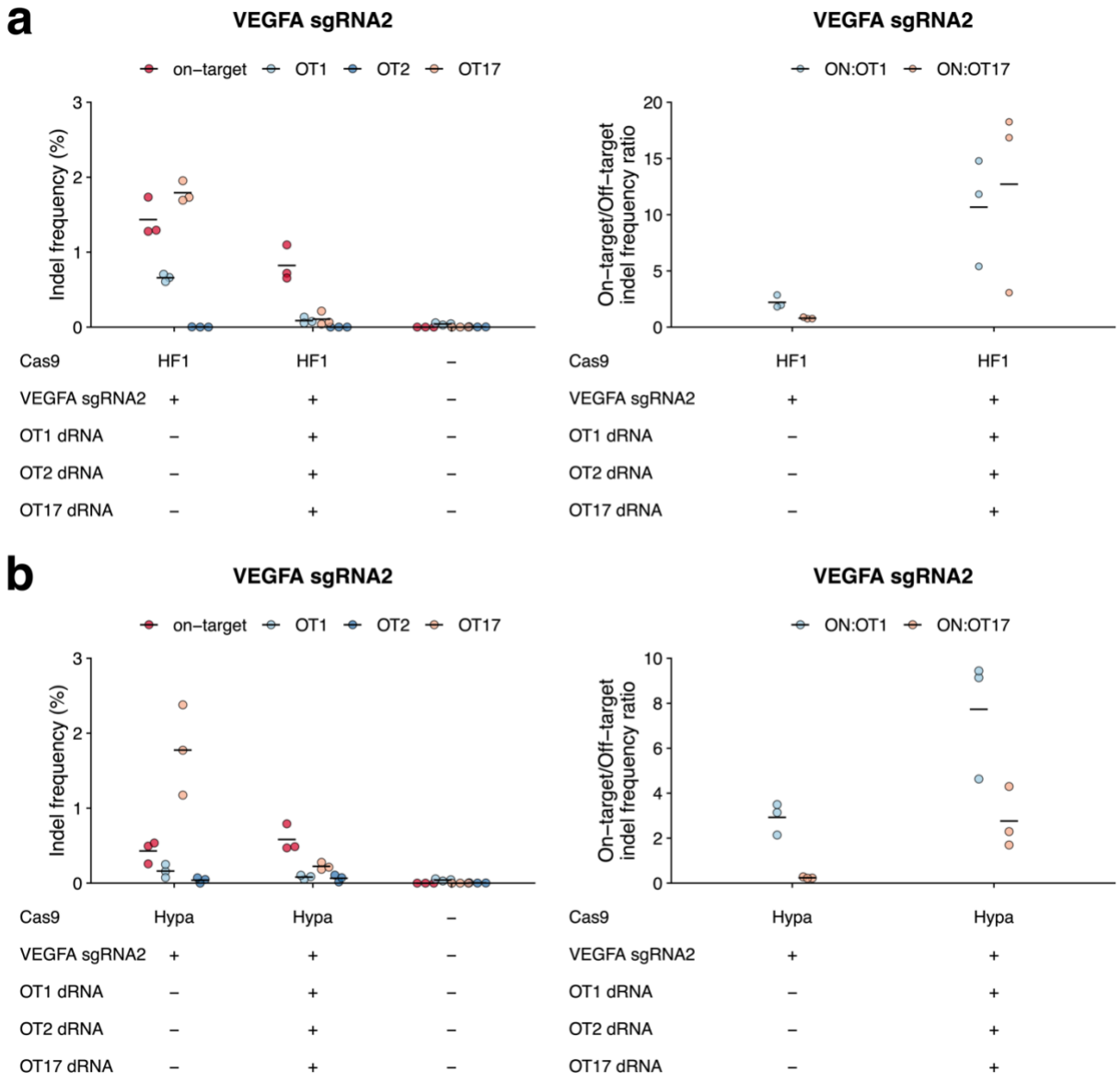

Figure B.12: Multiple dRNAs can be combined to reduce unwanted off-target editing at multiple refractory off-target sites of high-fidelity Cas9 variants

On-target and off-target indel frequencies and specificity ratios 24 hours after transfection with plasmids encoding *VEGFA* sgRNA2, a combination of three dRNAs, and either **a.** spCas9-HF1 (HF1) or **b.** HypaCas9 (Hypa). Despite being reported previously,[157] indels were not observed at OT2, so specificity ratios were not plotted. Indel frequencies for untransfected cells are shown as a control. Solid lines denote the mean of n = 3 biological replicates. OT = off-target.

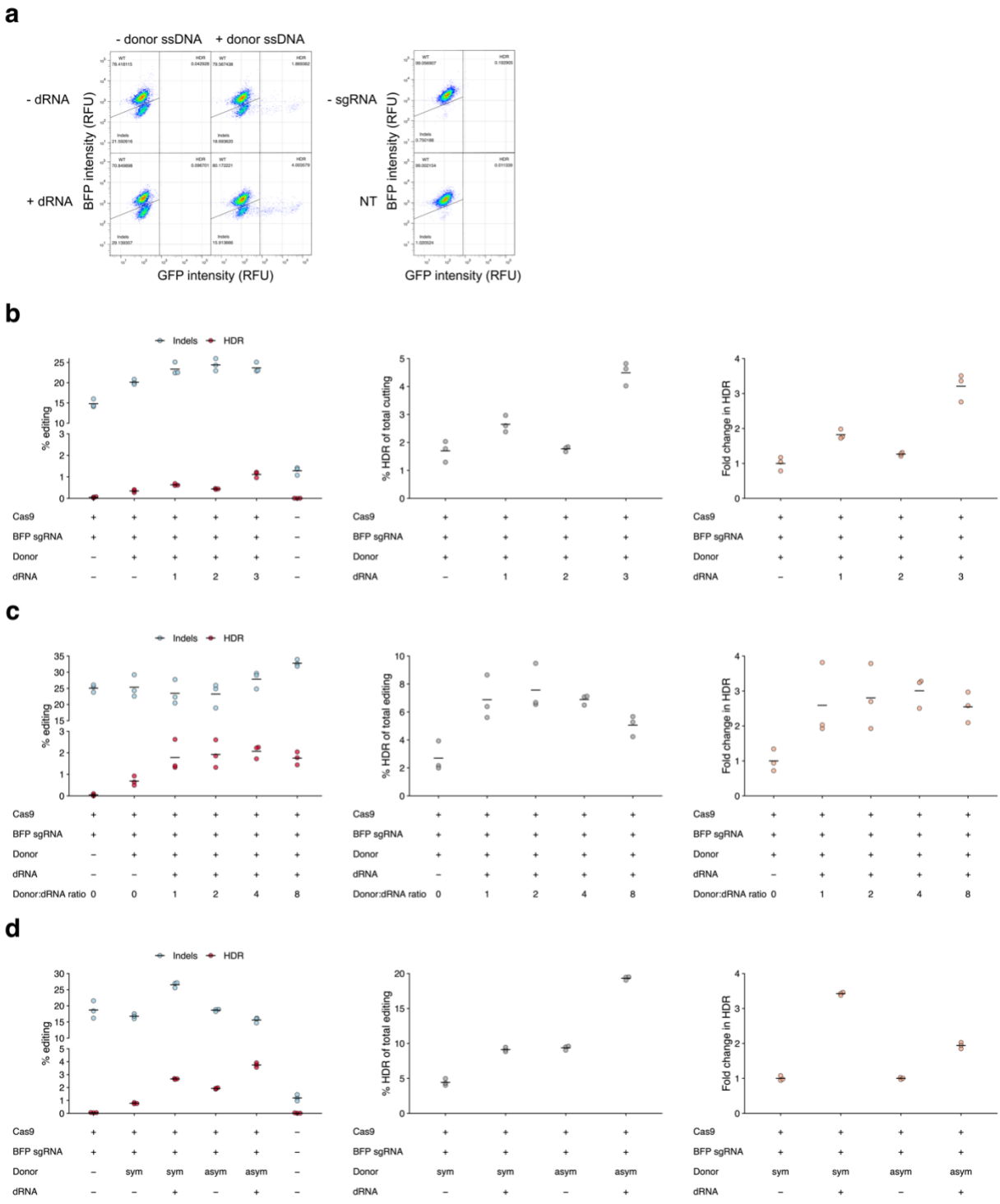

Figure B.13: Screening guides, donors, and dRNAs for scarless HDR in a fluorescent reporter system

a. Representative flow cytometry plots illustrating alteration of a synthetic BFP locus to GFP after Cas9•sgRNA editing. **b.** Screening of three dRNAs for indels or HDR events, percent HDR of total Cas9 editing observed, and fold change in HDR observed. **c.** Screening of various ratios of dRNA3 to sgRNA for indels or HDR events, percent HDR of total Cas9 editing observed, and fold change in HDR observed. **d.** Comparison of symmetric donor and asymmetric donor for indels or HDR events, percent HDR of total Cas9 editing observed, and fold change in HDR observed. HDR donors do not contain blocking mutations. Indel frequencies for untransfected cells are shown as a control. Numbers denote sgRNA and dRNA identities. Solid lines denote the mean of n = 3 biological replicates.

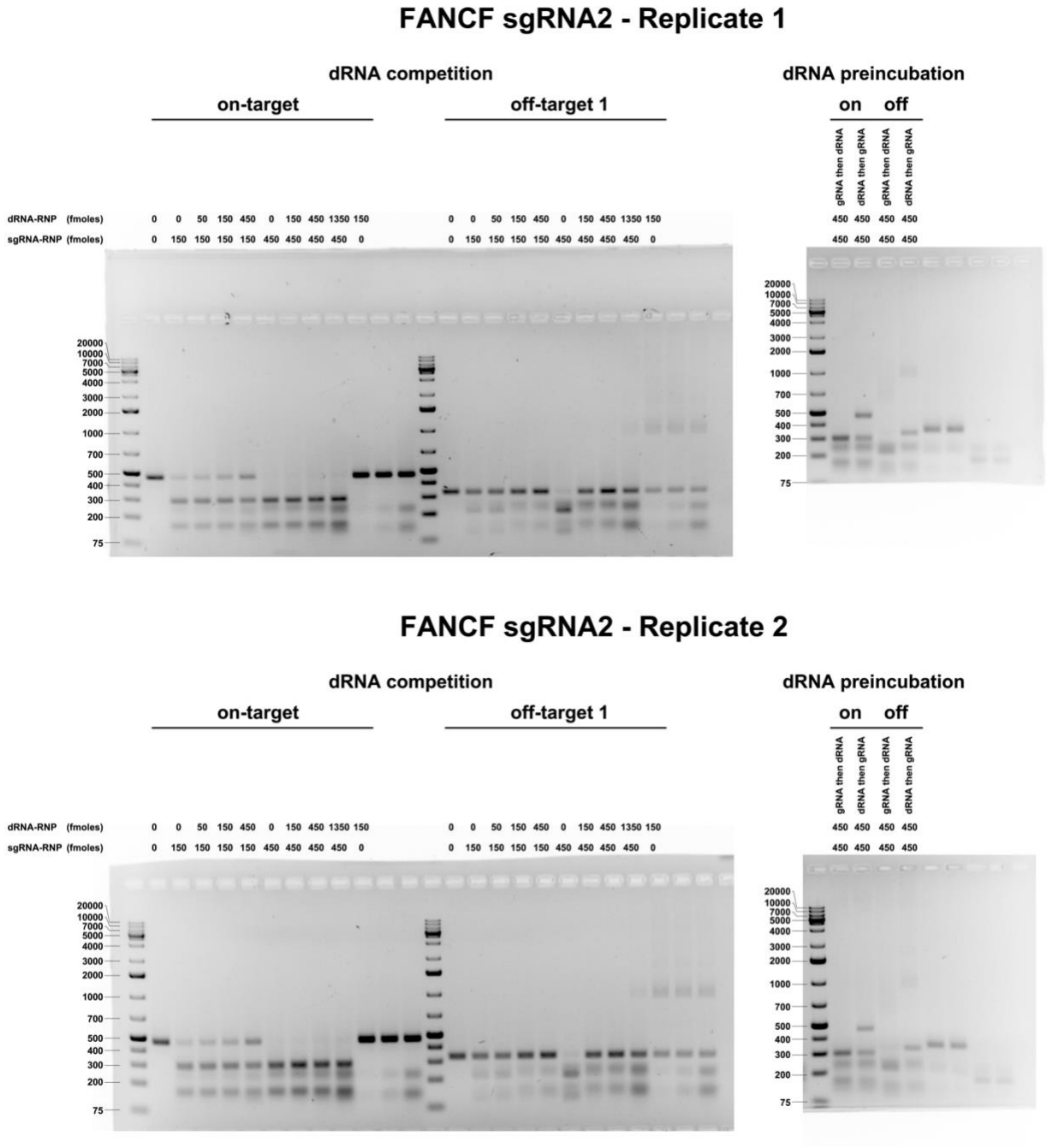

Figure B.14: Uncropped gel images for Figure 3.2 and Appendix B: Figure B.8

Full gel images of in vitro Cas9 *FANCF* sgRNA2 RNP cleavage of linear PCR products containing either the *FANCF* sgRNA2 on- or off-target site (OT1).

Site	Best dRNA	n	Normalized specificity ratio (mean)	On-target ratio (mean)	Normalized specificity ratio (s.e.m.)	On-target ratio (s.e.m.)
ZSCAN2 sgRNA1 OT2	1s	3	37.93	0.73	7.07	0.04
FANCF sgRNA2 OT1	1	3	29.96	1.04	3.42	0.04
VEGFA sgRNA2 OT17	8	3	13.11	1.02	1.57	0.08
CCR5-R30 OT-*CCR2*	3	3	11.34	0.57	6.81	0.28
ZSCAN2 sgRNA1 OT1	3	3	8.95	0.82	2.07	0.04
VEGFA sgRNA2 OT2	8	3	7.50	1.00	2.60	0.21
HBB-R01 OT-*HBD*	2	3	7.48	0.89	1.74	0.18
VEGFA sgRNA3 OT4	1	3	6.75	1.48	1.68	0.07
VEGFA sgRNA1 OT1	2	3	6.72	0.93	0.97	0.10
HBB-R03 OT-*HBD*	4	3	6.55	0.89	2.01	0.12
VEGFA sgRNA1 OT6	8	3	4.57	0.66	1.49	0.19
VEGFA sgRNA2 OT1	1	3	4.32	1.05	0.92	0.10
VEGFA sgRNA3 OT2	2	3	4.26	1.04	0.43	0.04
VEGFA sgRNA3 OT18	5	3	2.13	1.33	1.13	0.50
VEGFA sgRNA1 OT11	73	3	40.60	0.31	16.94	0.08
HBB-G10 OT1	7	3	3.74	0.47	1.49	0.08
VEGFA sgRNA2 OT19	5	3	1.55	0.72	0.55	0.13
HBB-R04 OT-*HBD*	4	3	1.16	0.77	0.29	0.09

Table B.1: dRNAs designed for a variety of sites increase specificity ratio with minimal effects on on-target editing

Normalized specificity ratios, computed as the specificity ratio in the presence of the best dRNA at a site divided by the specificity ratio in the absence of the dRNA, and on-target ratios, computed as the ratio of on-target editing in the presence of the best dRNA at a site divided by the on-target editing in the absence of the dRNA, for the best dRNA for 19 sgRNA/off-target pairs. n = 3 biological replicates, error measured as the standard error of the mean (s.e.m.).

Site	Δ (On)	p (On)	p_{adj} (On)	Δ (OT)	p (OT)	p_{adj} (OT)
FANCF sgRNA2 OT1	-0.004	0.835	1	-0.002	0.910	1
HBB R03 OT-*HBD*	-0.014	0.845	1	-0.009	0.761	1
VEGFA sgRNA1 OT1	0.0006	0.403	1	0.0007	0.094	1
VEGFA sgRNA1 OT6	-0.0002	0.524	1	0.025	0.209	1
VEGFA sgRNA2 OT1	-0.045	0.912	1	0.083	0.124	1
VEGFA sgRNA2 OT2	-0.018	0.750	1	-0.0009	0.683	1
VEGFA sgRNA2 OT17	0.015	0.306	1	0.008	0.218	1
VEGFA sgRNA3 OT4	-0.007	0.907	1	0	1	1
VEGFA sgRNA3 OT18	-0.0003	0.513	1	0.002	0.319	1
ZSCAN2 sgRNA1 OT1	-0.0004	0.789	1	0	1	1
ZSCAN2 sgRNA1 OT2	0.0003	0.479	1	0.001	0.092	1
VEGFA sgRNA3 OT2	0.040	0.406	1	0.080	0.050	1

Table B.2: dRNAs alone do not promote editing at sgRNA target sites

Difference between indel frequencies at on- and off-target (OT) sites for the best dRNA compared to a negative control at 12 different on/off-target pairs (Δ). p: p-value, based on two-sided Student's t-test. p_{adj}: Bonferroni-adjusted p-value. n = 3 biological replicates at on- and off-target sites, except for *VEGFA* sgRNA3 OT2 (n = 9) and *VEGFA* sgRNA3 OT18 (n = 3 at on-target, n = 2 at off-target due to failed sequencing reactions).

Site	Δ (On_{pred})	p (On_{pred})	p_{adj} (On_{pred})	Δ (OT_{pred})	p (OT_{pred})	p_{adj} (OT_{pred})
FANCF sgRNA2 OT1	-0.004	0.835	1	-0.002	0.910	1
HBB R03 OT-*HBD*	-0.056	0.699	1	-0.006	0.828	1
VEGFA sgRNA1 OT1	-0.004	0.730	1	0.001	0.312	1
VEGFA sgRNA1 OT6	0.0008	0.278	1	0.027	0.204	1
VEGFA sgRNA2 OT1	0.316	0.220	1	0.083	0.124	1
VEGFA sgRNA2 OT2	0.028	0.253	1	0.001	0.302	1
VEGFA sgRNA2 OT17	0.092	0.115	1	0.074	0.260	1
VEGFA sgRNA3 OT4	-0.015	0.908	1	0	1	1
VEGFA sgRNA3 OT18	0.0007	0.471	1	-0.013	0.622	1
ZSCAN2 sgRNA1 OT1	0.002	0.089	1	-0.0007	0.539	1
ZSCAN2 sgRNA1 OT2	0.0003	0.479	1	0.001	0.092	1
VEGFA sgRNA3 OT2	0.061	0.094	1	0.073	0.071	1

Table B.3: dRNAs alone do not promote editing at predicted dRNA target sites

Difference between indel frequencies at on- and off-target (OT) sites for the best dRNA compared to a negative control at 12 different on/off-target pairs (Δ). Predicted indel locations (pred) are the location of expected indels if the dRNA were a full length sgRNA. p: p-value, based on two-sided Student's t-test. p_{adj}: Bonferroni-adjusted p-value. n = 3 biological replicates at on- and off-target sites, except for *VEGFA* sgRNA3 OT2 (n = 9) *VEGFA* sgRNA2 OT17 (n = 3 at on-target, n = 2 at off-target due to failed sequencing reactions), and *VEGFA* sgRNA3 OT18 (n = 3 at on-target, n = 2 at off-target due to failed sequencing reactions).

List of abbreviations

A115	A-1155463
ACMG	American College of Medical Genetics
AR	androgen receptor
AUC	area under the curve
BFP	blue fluorescent protein
BLISS	breaks labeling *in situ* and sequencing
CRISPR	clustered regularly interspaced short palindromic repeat
CRM	cross-reactive material
crRNA	CRISPR-RNA
ddPCA	double-deep protein fragment complementation assay
DNA	deoxyribonucleic acid
dOTS	dRNA off-target suppression
dReCS	dRNA recutting suppression
dRNA	dead single guide RNA
DSB	double strand break
dsODN	double stranded oligodeoxynucleotide
EAHAD	European association for haemophilia and allied disorders
ER	endoplasmic reticulum
FACS	fluorescence assisted cell sorting
FII	thrombin (coagulation factor II)
FIX	coagulation factor IX
FIX_a	coagulation factor IX, activated

FVII	coagulation factor VII
$FVII_a$	coagulation factor VII, activated
FVIII	coagulation factor VIII
$FVIII_a$	coagulation factor VIII, activated
FX	coagulation factor X
FX_a	coagulation factor X, activated
FXI	coagulation factor XI
FXI_a	coagulation factor XI, activated
GFP	green fluorescent protein
gnomAD	genome aggregation database
GUIDE-seq	genome-wide unbiased identification of DSBs enabled by sequencing
HDR	homology-directed repair
HNF-4	hepatocyte nuclear factor 4
HIV-1	human immunodeficiency virus
iPSC	induced pluripotent stem cell
IRES	internal ribosomal entry site
MAVE	multiplexed assays of variant effect
MPRA	massively parallel reporter assay
MultiSTEP	multiplexed surface tethering of extracellular proteins
NHEJ	non-homologous end joining
OddsPath	odds of pathogenicity
OT	off-target site
PAM	protospacer adjacent motif

PC	protein C
PCR	polymerase chain reaction
PS	protein S
PTM	post-translational modification
RNA	ribonucleic acid
ROC	receiver operating characteristic
ROSE	random oversampling examples
RSA	relative solvent accessibility
SASA	solvent accessible surface area
SGE	saturation genome editing
ssODN	single stranded oligodeoxynucleotide
TALEN	transcription activator-like effector nuclease
TF	tissue factor
tru-sgRNA	truncated sgRNA
UTR	untranslated region
VAMP-seq	variant abundance by massively parallel sequencing
VCEP	ClinGen variant curation expert panel
VUS	variant of uncertain significance
vWF	von Willebrand factor

References

1. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
2. Shirts, B. H., Pritchard, C. C. & Walsh, T. Family-specific variants and the limits of human genetics. *Trends Mol. Med.* **22**, 925–934 (2016).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
5. Chiasson, M. A. *et al.* Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**, (2020).
6. Amorosi, C. J. *et al.* Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).
7. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* e102 (2017).
8. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* **48**, e1–e1 (2020).
9. Starita, L. M. *et al.* Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

10. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455–E464 (2014).

11. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences* **112**, E5189–E5198 (2015).

12. Richards, S. *et al.* 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**, 405–423 (2015).

13. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).

14. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).

15. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* **13**, 957 (2017).

16. Findlay, G. M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217 (2018).

17. Lee, J. M. *et al.* Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. *Proceedings of the National Academy of Sciences* **115**, E8276–E8285 (2018).

18. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A saturation mutagenesis approach to understanding PTEN lipid phosphatase activity and genotype-phenotype relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).

19. Ahler, E. *et al.* A combined approach reveals a regulatory mechanism coupling Src's kinase activity, localization, and phosphotransferase-independent functions. *Mol. Cell* **74**, 393-408.e20 (2019).

20. Gray, V. E. *et al.* Elucidating the Molecular Determinants of Aβ Aggregation with Deep Mutational Scanning. *G3* **9**, 3683–3689 (2019).

21. Suiter, C. C. *et al.* Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5394–5401 (2020).

22. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).

23. Seuma, M., Faure, A. J., Badia, M., Lehner, B. & Bolognesi, B. The genetic landscape for amyloid beta fibril nucleation accurately discriminates familial Alzheimer's disease mutations. *Elife* **10**, (2021).

24. Starr, T. N. *et al.* Prospective mapping of viral mutations that escape antibodies used to treat COVID-19. *Science* **371**, 850–854 (2021).

25. Frank, F. *et al.* Deep mutational scanning identifies SARS-CoV-2 Nucleocapsid escape mutations of currently available rapid antigen tests. *Cell* **0**, (2022).

26. Coyote-Maestas, W., Nedrud, D., He, Y. & Schmidt, D. Determinants of trafficking, conduction, and disease within a K+ channel revealed through multiparametric deep mutational scanning. *Elife* **11**, (2022).

27. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

28. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).

29. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

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

31. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).

32. Gasperini, M. *et al.* 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 (2017).

33. Cao, J. *et al.* High-throughput 5’ UTR engineering for enhanced protein production in non-viral gene therapies. *Nat. Commun.* **12**, 4138 (2021).

34. Erwood, S. *et al.* Saturation variant interpretation using CRISPR prime editing. *Nat. Biotechnol.* **40**, 885–895 (2022).

35. Uhlén, M. *et al.* The human secretome. *Sci. Signal.* **12**, (2019).

36. Robinson, J. L., Feizi, A., Uhlén, M. & Nielsen, J. A Systematic Investigation of the Malignant Functions and Diagnostic Potential of the Cancer Secretome. *Cell Rep.* **26**, 2622-2635.e5 (2019).

37. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).

38. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).

39. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17**, 467–473 (2007).

40. van Bloois, E., Winter, R. T., Kolmar, H. & Fraaije, M. W. Decorating microbes: surface display of proteins on Escherichia coli. *Trends Biotechnol.* **29**, 79–86 (2011).

41. Salema, V. & Fernández, L. Á. Escherichia coli surface display for the selection of nanobodies. *Microb. Biotechnol.* **10**, 1468–1484 (2017).

42. Hegde, R. S. & Bernstein, H. D. The surprising complexity of signal sequences. *Trends Biochem. Sci.* **31**, 563–571 (2006).

43. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).

44. Geukens, N. *et al.* Analysis of type I signal peptidase affinity and specificity for preprotein substrates. *Biochem. Biophys. Res. Commun.* **314**, 459–467 (2004).

45. Low, K. O., Muhammad Mahadi, N. & Md Illias, R. Optimisation of signal peptide for recombinant protein secretion in bacterial hosts. *Appl. Microbiol. Biotechnol.* **97**, 3811–3826 (2013).

46. Braud, V., Jones, E. Y. & McMichael, A. The human major histocompatibility complex class Ib molecule HLA-E binds signal sequence-derived peptides with primary anchor residues at positions 2 and 9. *Eur. J. Immunol.* **27**, 1164–1169 (1997).

47. Braakman, I. & Hebert, D. N. Protein folding in the endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **5**, a013201 (2013).

48. Hanson, S. R. *et al.* The core trisaccharide of an N-linked glycoprotein intrinsically accelerates folding and enhances stability. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3131–3136 (2009).

49. Konkle, B. A., Josephson, N. C. & Nakaya Fletcher, S. Hemophilia B. in *GeneReviews* (eds. Pagon, R. A. et al.) (University of Washington, Seattle, 2014).

50. Wojcik, E. G., Van Den Berg, M., Poort, S. R. & Bertina, R. M. Modification of the N-terminus of human factor IX by defective propeptide cleavage or acetylation results in a destabilized calcium-induced conformation: effects on phospholipid binding and activation by factor XIa. *Biochem. J* **323 (Pt 3)**, 629–636 (1997).

51. Aktimur, A., Gabriel, M. A., Gailani, D. & Toomey, J. R. The Factor IX γ-Carboxyglutamic Acid (Gla) Domain Is Involved in Interactions between Factor IX and Factor XIa. *J. Biol. Chem.* **278**, 7981–7987 (2003).

52. Huang, M. *et al.* Structural basis of membrane binding by Gla domains of vitamin K-dependent proteins. *Nat. Struct. Biol.* **10**, 751–756 (2003).

53. Grant, M. A., Baikeev, R. F., Gilbert, G. E. & Rigby, A. C. Lysine 5 and phenylalanine 9 of the factor IX omega-loop interact with phosphatidylserine in a membrane-mimetic environment. *Biochemistry* **43**, 15367–15378 (2004).

54. Huang, M., Furie, B. C. & Furie, B. Crystal Structure of the Calcium-stabilized Human Factor IX Gla Domain Bound to a Conformation-specific Anti-factor IX Antibody. *J. Biol. Chem.* **279**, 14338–14346 (2004).

55. Puri, R. N. & Colman, R. W. ADP-induced platelet activation. *Crit. Rev. Biochem. Mol. Biol.* **32**, 437–502 (1997).

56. Periayah, M. H., Halim, A. S. & Mat Saad, A. Z. Mechanism Action of Platelets and Crucial Blood Coagulation Pathways in Hemostasis. *Int J Hematol Oncol Stem Cell Res* **11**, 319–327 (2017).

57. Gale, A. J. Continuing education course #2: current understanding of hemostasis. *Toxicol. Pathol.* **39**, 273–280 (2011).

58. Malkhassian, D., Sabir, S. & Sharma, S. *Physiology, Factor XIII*. (StatPearls Publishing, 2022).

59. Johnsen, J. M. *et al.* Results of genetic analysis of 11 341 participants enrolled in the My Life, Our Future hemophilia genotyping initiative in the United States. *J. Thromb. Haemost.* **20**, 2022–2034 (2022).

60. Ludwig, M. *et al.* Hemophilia B caused by five different nondeletion mutations in the protease domain of factor IX. *Blood* **79**, 1225–1232 (1992).

61. Simioni, P. *et al.* X-linked thrombophilia with a mutant Factor IX (Factor IX Padua). *N. Engl. J. Med.* **361**, 1671–1675 (2009).

62. Chu, K., Wu, S. M., Stanley, T., Stafford, D. W. & High, K. A. A mutation in the propeptide of Factor IX leads to warfarin sensitivity by a novel mechanism. *J. Clin. Invest.* **98**, 1619–1625 (1996).

63. Oldenburg, J. *et al.* Missense mutations at ALA-10 in the factor IX propeptide: an insignificant variant in normal life but a decisive cause of bleeding during oral anticoagulant therapy. *Br. J. Haematol.* **98**, 240–244 (1997).

64. Kulkarni, R. *et al.* Sites of initial bleeding episodes, mode of delivery and age of diagnosis in babies with haemophilia diagnosed before the age of 2 years: A report from the Centers for Disease Control and Prevention's (CDC) Universal Data Collection (UDC) project. *Haemophilia* **15**, 1281–1290 (2009).

65. Iorio, A. *et al.* Establishing the Prevalence and Prevalence at Birth of Hemophilia in Males: A Meta-analytic Approach Using National Registries. *Ann. Intern. Med.* **171**, 540–546 (2019).

66. DiMichele, D. M. *et al.* Severe and moderate haemophilia A and B in US females. *Haemophilia* **20**, e136-43 (2014).

67. Miller, C. H. & Bean, C. J. Genetic causes of haemophilia in women and girls. *Haemophilia* **27**, e164–e179 (2021).

68. Miller, C. H. *et al.* Women and girls with haemophilia receiving care at specialized haemophilia treatment centres in the United States. *Haemophilia* **27**, 1037–1044 (2021).

69. Warrier, I. *et al.* Factor IX Inhibitors and Anaphylaxis in Hemophilia B. *J. Pediatr. Hematol. Oncol.* **19**, 23 (1997).

70. Oldenburg, J. & Pavlova, A. Genetic risk factors for inhibitors to factors VIII and IX. *Haemophilia* **12 Suppl 6**, 15–22 (2006).

71. Chalmers, E. *et al.* Guideline on the management of haemophilia in the fetus and neonate. *Br. J. Haematol.* **154**, 208–215 (2011).

72. Konkle, B. A. *et al.* Genotypes, phenotypes and whole genome sequence: Approaches from the My Life Our Future haemophilia project. *Haemophilia* **24**, 87–94 (2018).

73. Cooley, B. *et al.* Dysfunctional endogenous FIX impairs prophylaxis in a mouse hemophilia B model. *Blood* **133**, 2445–2451 (2019).

74. Mann, D. M., Stafford, K. A., Poon, M.-C., Matino, D. & Stafford, D. W. The Function of extravascular coagulation factor IX in haemostasis. *Haemophilia* **27**, 332–339 (2021).

75. Crudele, J. M. *et al.* AAV liver expression of FIX-Padua prevents and eradicates FIX inhibitor without increasing thrombogenicity in hemophilia B dogs and mice. *Blood* **125**, 1553–1561 (2015).

76. George, L. A. *et al.* Hemophilia B Gene Therapy with a High-Specific-Activity Factor IX Variant. *N. Engl. J. Med.* **377**, 2215–2227 (2017).

77. Rallapalli, P. M., Kemball-Cook, G., Tuddenham, E. G., Gomez, K. & Perkins, S. J. An interactive mutation database for human coagulation Factor IX provides novel insights into the phenotypes and genetics of hemophilia B. *J. Thromb. Haemost.* **11**, 1329–1340 (2013).

78. Kurachi, S., Pantazatos, D. P. & Kurachi, K. The carboxyl-terminal region of Factor IX is essential for its secretion. *Biochemistry* **36**, 4337–4344 (1997).

79. Wulff, K., Bykowska, K., Lopaciuk, S. & Herrmann, F. H. Molecular analysis of hemophilia B in Poland: 12 novel mutations of the factor IX gene. *Acta Biochim. Pol.* **46**, 721–726 (1999).

80. Green, P. M., Bentley, D. R., Mibashan, R. S., Nilsson, I. M. & Giannelli, F. Molecular pathology of haemophilia B. *EMBO J.* **8**, 1067–1072 (1989).

81. Goodeve, A. C. Hemophilia B: Molecular pathogenesis and mutation analysis. *J. Thromb. Haemost.* **13**, 1184–1195 (2015).

82. Bandyopadhyay, P. K. *et al.* γ-Glutamyl carboxylation: An extracellular posttranslational modification that antedates the divergence of molluscs, arthropods, and chordates. *Proceedings of the National Academy of Sciences* **99**, 1264–1269 (2002).

83. Sunnerhagen, M. *et al.* Structure of the Ca(2+)-free Gla domain sheds light on membrane binding of blood coagulation proteins. *Nat. Struct. Biol.* **2**, 504–509 (1995).

84. Freedman, S. J., Furie, B. C., Furie, B. & Baleja, J. D. Structure of the Metal-free γ-Carboxyglutamic Acid-rich Membrane Binding Region of Factor IX by Two-dimensional NMR Spectroscopy (∗). *J. Biol. Chem.* **270**, 7980–7987 (1995).

85. Freedman, S. J. *et al.* Identification of the phospholipid binding site in the vitamin K-dependent blood coagulation protein factor IX. *J. Biol. Chem.* **271**, 16227–16236 (1996).

86. Shikamoto, Y., Morita, T., Fujimoto, Z. & Mizuno, H. Crystal structure of Mg2+- and Ca2+-bound Gla domain of factor IX complexed with binding protein. *J. Biol. Chem.* **278**, 24090–24094 (2003).

87. Gailani, D. *et al.* The mechanism underlying activation of Factor IX by Factor XIa. *Thromb. Res.* **133**, S48–S51 (2014).

88. Larson, P. J. *et al.* Structural integrity of the γ-carboxyglutamic acid domain of human blood coagulation Factor IXa is required for its binding to cofactor VIIIa. *J. Biol. Chem.* **271**, 3869–3876 (1996).

89. Ndonwi, M., Broze, G. J., Agah, S., Schmidt, A. E. & Bajaj, S. P. Substitution of the Gla Domain in Factor X with That of Protein C Impairs Its Interaction with Factor VIIa/Tissue Factor: LACK OF COMPARABLE EFFECT BY SIMILAR SUBSTITUTION IN FACTOR IX*. *J. Biol. Chem.* **282**, 15632–15644 (2007).

90. Geng, Y. *et al.* A sequential mechanism for exosite-mediated Factor IX activation by Factor XIa. *J. Biol. Chem.* **287**, 38200–38209 (2012).

91. Geng, Y. *et al.* Analysis of the factor XI variant Arg184Gly suggests a structural basis for factor IX binding to factor XIa. *J. Thromb. Haemost.* **11**, 1374–1384 (2013).

92. Gillis, S. *et al.* γ-Carboxyglutamic acids 36 and 40 do not contribute to human Factor IX function. *Protein Sci.* **6**, 185–196 (1997).

93. Bentley, A. K., Rees, D. J. G., Rizza, C. & Brownlee, G. G. Defective propeptide processing of blood clotting factor IX caused by mutation of arginine to glutamine at position −4. *Cell* **45**, 343–348 (1986).

94. Ware, J. *et al.* Factor IX San Dimas. Substitution of glutamine for Arg-4 in the propeptide leads to incomplete γ-carboxylation and altered phospholipid binding properties. *J. Biol. Chem.* **264**, 11401–11406 (1989).

95. Bristol, J. A., Freedman, S. J., Furie, B. C. & Furie, B. Profactor IX: The propeptide inhibits binding to membrane surfaces and activation by factor XIA. *Biochemistry* **33**, 14136–14143 (1994).

96. Wolberg, A. S. *et al.* Characterization of γ-carboxyglutamic acid residue 21 of human Factor IX. *Biochemistry* **35**, 10321–10327 (1996).

97. Gao, W. *et al.* Characterization of missense mutations in the signal peptide and propeptide of FIX in hemophilia B by a cell-based assay. *Blood Adv* **4**, 3659–3667 (2020).

98. Furie, B. & Furie, B. C. Molecular basis of vitamin K-dependent gamma-carboxylation. *Blood* **75**, 1753–1762 (1990).

99. Higgins-Gruber, S. L. *et al.* Effect of vitamin K-dependent protein precursor propeptide, vitamin K hydroquinone, and glutamate substrate binding on the structure and function of γ-glutamyl carboxylase. *J. Biol. Chem.* **285**, 31502–31508 (2010).

100. Haque, J. A., McDonald, M. G., Kulman, J. D. & Rettie, A. E. A cellular system for quantitation of vitamin K cycle activity: Structure-activity effects on vitamin K antagonism by warfarin metabolites. *Blood* **123**, 582–589 (2014).

101. Diuguid, D. L., Rabiet, M. J., Furie, B. C., Liebman, H. A. & Furie, B. Molecular basis of hemophilia B: A defective enzyme due to an unprocessed propeptide is

caused by a point mutation in the Factor IX precursor. *Proceedings of the National Academy of Sciences* **83**, 5803–5807 (1986).

102. Rabiet, M. J., Jorgensen, M. J., Furie, B. & Furie, B. C. Effect of propeptide mutations on post-translational processing of Factor IX. Evidence that beta-hydroxylation and gamma-carboxylation are independent events. *J. Biol. Chem.* **262**, 14895–14898 (1987).

103. Feuerstein Giora Z. *et al.* Antithrombotic efficacy of a novel murine antihuman factor IX antibody in rats. *Arterioscler. Thromb. Vasc. Biol.* **19**, 2554–2562 (1999).

104. Toomey, J. R. *et al.* Comparing the antithrombotic efficacy of a humanized anti-factor IX(a) monoclonal antibody (SB 249417) to the low molecular weight heparin enoxaparin in a rat model of arterial thrombosis. *Thromb. Res.* **100**, 73–79 (2000).

105. Stenina, O., Pudota, B. N., McNally, B. A., Hommema, E. L. & Berkner, K. L. Tethered processivity of the vitamin K-dependent carboxylase: factor IX is efficiently modified in a mechanism which distinguishes Gla's from Glu's and which accounts for comprehensive carboxylation in vivo. *Biochemistry* **40**, 10301–10309 (2001).

106. Hallgren, K. W., Hommema, E. L., McNally, B. A. & Berkner, K. L. Carboxylase overexpression effects full carboxylation but poor release and secretion of Factor IX: Implications for the release of vitamin K-dependent proteins. *Biochemistry* **41**, 15045–15055 (2002).

107. Hasle, N., Matreyek, K. A. & Fowler, D. M. The Impact of Genetic Variants on PTEN Molecular Functions and Cellular Phenotypes. *Cold Spring Harb. Perspect. Med.* **9**, (2019).

108. Matreyek, K. A., Stephany, J. J., Ahler, E. & Fowler, D. M. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med.* **13**, 165 (2021).

109. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nat. Methods* **10**, 715–721 (2013).

110. Rizzo, M. A., Davidson, M. W. & Piston, D. W. Fluorescent protein tracking and detection: applications using fluorescent proteins in living cells. *Cold Spring Harb. Protoc.* **2009**, db.top64 (2009).

111. Calabrese, E. J. & Baldwin, L. A. Hormesis: U-shaped dose responses and their centrality in toxicology. *Trends Pharmacol. Sci.* **22**, 285–291 (2001).

112. Kim, H. S. *et al.* CReVIS-Seq: A highly accurate and multiplexable method for genome-wide mapping of lentiviral integration sites. *Mol Ther Methods Clin Dev* **20**, 792–800 (2021).

113. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).

114. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

115. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).

116. Cameron, P. *et al.* Mapping the genomic landscape of CRISPR–Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).

117. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).

118. Rose, J. C. *et al.* Suppression of unwanted CRISPR-Cas9 editing by co-administration of catalytically inactivating truncated guide RNAs. *Nat. Commun.* **11**, 2697 (2020).

119. Haendiges, J., Jinneman, K. & Gonzalez-Escalona, N. Choice of library preparation affects sequence quality, genome assembly, and precise in silico prediction of virulence genes in shiga toxin-producing Escherichia coli. *PLoS One* **16**, e0242294 (2021).

120. Cawley, S. *et al.* A framework for evaluating edited cell libraries created by massively parallel genome engineering. *bioRxiv* 2021.09.23.458228 (2022) doi:10.1101/2021.09.23.458228.

121. DeWitt, M. A. *et al.* Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Sci. Transl. Med.* **8**, 360ra134-360ra134 (2016).

122. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).

123. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).

124. Bétermier, M., Bertrand, P. & Lopez, B. S. Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.* **10**, e1004086 (2014).

125. Chakraborty, A. *et al.* Classical non-homologous end-joining pathway utilizes nascent RNA for error-free double-strand break repair of transcribed genes. *Nat. Commun.* **7**, 13049 (2016).

126. Rose, J. C. *et al.* Rapidly inducible Cas9 and DSB-ddPCR to probe editing kinetics. *Nat. Methods* **14**, 891–896 (2017).

127. Song, B., Yang, S., Hwang, G.-H., Yu, J. & Bae, S. Analysis of NHEJ-Based DNA Repair after CRISPR-Mediated DNA Cleavage. *Int. J. Mol. Sci.* **22**, (2021).

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

129. Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat. Genet.* **49**, 625–634 (2017).

130. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636–646 (2010).

131. Joung, J. K. & Sander, J. D. TALENs: a widely applicable technology for targeted genome editing. *Nat. Rev. Mol. Cell Biol.* **14**, 49–55 (2013).

132. Jiang, F. & Doudna, J. A. CRISPR–Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).

133. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

134. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).

135. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37**, 67–78 (2017).

136. Liu, Z., Venkatesh, S. S. & Maley, C. C. Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples. *BMC Genomics* **9**, 509 (2008).

137. Cradick, T. J., Fine, E. J., Antico, C. J. & Bao, G. CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41**, 9584–9592 (2013).

138. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).

139. Zhang, L. *et al.* Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. *Nucleic Acids Res.* **48**, 5037–5053 (2020).

140. Pacesa, M. *et al.* Structural basis for Cas9 off-target activity. *Cell* **185**, 4067-4081.e21 (2022).

141. Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat. Commun.* **7**, 12778 (2016).

142. Boyle, E. A. *et al.* Quantification of Cas9 binding and cleavage across diverse guide sequences maps landscapes of target engagement. *Sci Adv* **7**, (2021).

143. Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. Small molecule–triggered Cas9 protein with improved genome-editing specificity. *Nat. Chem. Biol.* **11**, 316–318 (2015).

144. Zetsche, B., Volz, S. E. & Zhang, F. A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nat. Biotechnol.* **33**, 139–142 (2015).

145. Maji, B. *et al.* Multidimensional chemical control of CRISPR–Cas9. *Nat. Chem. Biol.* **13**, 9–11 (2017).

146. Tycko, J., Myer, V. E. & Hsu, P. D. Methods for optimizing CRISPR-Cas9 genome editing specificity. *Mol. Cell* **63**, 355–370 (2016).

147. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).

148. Yin, H. *et al.* Partial DNA-guided Cas9 enables genome editing with reduced off-target activity. *Nat. Chem. Biol.* **14**, 311–316 (2018).

149. Ryan, D. E. *et al.* Improving CRISPR–Cas specificity with chemical modifications in single-guide RNAs. *Nucleic Acids Res.* **46**, 792–803 (2018).

150. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).

151. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).

152. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32**, 577–582 (2014).

153. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).

154. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

155. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

156. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

157. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).

158. Lee, J. K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, 1–10 (2018).

159. Singh, D. *et al.* Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. *Nat. Struct. Mol. Biol.* **25**, 347–354 (2018).

160. Höijer, I. *et al.* CRISPR-Cas9 induces large structural variants at on-target and off-target sites in vivo that segregate across generations. *Nat. Commun.* **13**, 627 (2022).

161. Tao, J., Bauer, D. E. & Chiarle, R. Assessing and advancing the safety of CRISPR-Cas tools: from DNA to RNA editing. *Nat. Commun.* **14**, 212 (2023).

162. Hacein-Bey-Abina, S. *et al.* A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **348**, 255–256 (2003).

163. Lebbink, R. J. *et al.* A combinational CRISPR/Cas9 gene-editing approach can halt HIV replication and prevent viral escape. *Sci. Rep.* **7**, 41968 (2017).

164. Atkins, A. *et al.* Off-Target Analysis in Gene Editing and Applications for Clinical Translation of CRISPR/Cas9 in HIV-1 Therapy. *Front Genome Ed* **3**, 673022 (2021).

165. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).

166. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

167. Giacomelli, A. O. *et al.* Mutational processes shape the landscape of TP53 mutations in human cancer. *Nat. Genet.* **50**, 1381–1387 (2018).

168. Boettcher, S. *et al.* A dominant-negative effect drives selection of *TP53* missense mutations in myeloid malignancies. *Science* **365**, 599–604 (2019).

169. Ho, M. & Pastan, I. Mammalian cell display for antibody engineering. in *Therapeutic Antibodies* (ed. Dimitrov, A. S.) 337–352 (Humana Press, 2009).

170. Bandaranayake, A. D. *et al.* Daedalus: a robust, turnkey platform for rapid production of decigram quantities of active recombinant proteins in human cell lines using novel lentiviral vectors. *Nucleic Acids Res.* **39**, e143 (2011).

171. Crook, Z. R. *et al.* Mammalian display screening of diverse cystine-dense peptides for difficult to drug targets. *Nat. Commun.* **8**, 2244 (2017).

172. Liu, L. *et al.* Inclusion of Strep-tag II in design of antigen receptors for T-cell immunotherapy. *Nat. Biotechnol.* **34**, 430–434 (2016).

173. Vink, T., Oudshoorn-Dickmann, M., Roza, M., Reitsma, J.-J. & de Jong, R. N. A simple, robust and highly efficient transient expression system for producing antibodies. *Methods* **65**, 5–10 (2014).

174. DiMichele, D. Inhibitor development in haemophilia B: An orphan disease in need of attention. *Br. J. Haematol.* **138**, 305–315 (2007).

175. Brown, M. A., Stenberg, L. M., Persson, U. & Stenflo, J. Identification and Purification of Vitamin K-dependent Proteins and Peptides with Monoclonal Antibodies Specific for γ-Carboxyglutamyl (Gla) Residues *. *J. Biol. Chem.* **275**, 19795–19802 (2000).

176. Luirink, J. & Sinning, I. SRP-mediated protein targeting: structure and function revisited. *Biochim. Biophys. Acta* **1694**, 17–35 (2004).

177. Halic, M. & Beckmann, R. The signal recognition particle and its interactions during protein targeting. *Curr. Opin. Struct. Biol.* **15**, 116–125 (2005).

178. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).

179. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

180. Robinson, P. J. & Bulleid, N. J. Mechanisms of Disulfide Bond Formation in Nascent Polypeptides Entering the Secretory Pathway. *Cells* **9**, (2020).

181. Menardi, G. & Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **28**, 92–122 (2014).

182. Brnich, S. E. *et al.* Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.* **12**, 1–12 (2020).

183. Fayer, S. *et al.* Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *Am. J. Hum. Genet.* **108**, 2248–2258 (2021).

184. Hao, Z., Jin, D.-Y., Stafford, D. W. & Tie, J.-K. Vitamin K-dependent carboxylation of coagulation factors: insights from a cell-based functional study. *Haematologica* **105**, 2164–2173 (2020).

185. Moussalli, M. *et al.* Mannose-dependent Endoplasmic Reticulum (ER)-Golgi Intermediate Compartment-53-mediated ER to Golgi Trafficking of Coagulation Factors V and VIII*. *J. Biol. Chem.* **274**, 32539–32542 (1999).

186. García-Nafría, J., Watson, J. F. & Greger, I. H. IVA cloning: A single-tube universal cloning system exploiting bacterial In Vivo Assembly. *Sci. Rep.* **6**, 1–12 (2016).

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

188. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

189. Yeh, C.-L. C., Amorosi, C. J., Showman, S. & Dunham, M. J. PacRAT: a program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignment. *Bioinformatics* **38**, 2927–2929 (2022).

190. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).

191. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res.* **5**, 189 (2016).

192. Miller, S., Lesk, A. M., Janin, J. & Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834–836 (1987).

193. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

194. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).

195. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).

196. Kiani, S. *et al.* Cas9 gRNA engineering for genome editing, activation and repression. *Nat. Methods* **12**, 1051–1054 (2015).

197. Dahlman, J. E. *et al.* Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nat. Biotechnol.* **33**, 1159–1161 (2015).

198. Ye, L. *et al.* Programmable DNA repair with CRISPRa/i enhanced homology-directed repair efficiency with a single Cas9. *Cell Discovery* **4**, 1–12 (2018).

199. Ferreccio, A. *et al.* Inducible CRISPR genome editing platform in naive human embryonic stem cells reveals JARID2 function in self-renewal. *Cell Cycle* **17**, 535–549 (2018).

200. Ma, H. *et al.* Correction of a pathogenic gene mutation in human embryos. *Nature* **548**, 413–419 (2017).

201. De Ravin, S. S. *et al.* CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci. Transl. Med.* **9**, eaah3480 (2017).

202. Dever, D. P. *et al.* CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. *Nature* **539**, 384–389 (2016).

203. Rose, J. C., Stephany, J. J., Wei, C. T., Fowler, D. M. & Maly, D. J. Rheostatic control of Cas9-mediated DNA double strand break (DSB) generation and genome editing. *ACS Chem. Biol.* **13**, 438–442 (2018).

204. Clarke, R. *et al.* Enhanced bacterial immunity and mammalian genome editing via RNA-polymerase-mediated dislodging of Cas9 from double-strand DNA breaks. *Mol. Cell* **71**, 42-55.e8 (2018).

205. Isaac, R. S. *et al.* Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *Elife* **5**, e13450 (2016).

206. Vigouroux, A., Oldewurtel, E., Cui, L., Bikard, D. & van Teeffelen, S. Tuning dCas9's ability to block transcription enables robust, noiseless knockdown of bacterial genes. *Mol. Syst. Biol.* **14**, e7899 (2018).

207. Choi, G. C. G. *et al.* Combinatorial mutagenesis en masse optimizes the genome editing activities of SpCas9. *Nat. Methods* **16**, 722–730 (2019).

208. Tsai, S. Q. *et al.* CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).

209. Grünewald, J. *et al.* Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).

210. Zhou, C. *et al.* Off-target RNA mutation induced by DNA base editing and its elimination by mutagenesis. *Nature* **571**, 275–278 (2019).

211. Rees, H. A. & Liu, D. R. Base editing: Precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).

212. Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).

213. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).

214. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

215. Kabadi, A. M., Ousterout, D. G., Hilton, I. B. & Gersbach, C. A. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res.* **42**, e147–e147 (2014).

216. Gu, B. *et al.* Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **359**, 1050–1055 (2018).

217. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).

218. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

219. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

220. Kulcsár, P. I. *et al.* Crossing enhanced and high fidelity SpCas9 nucleases to optimize specificity and cleavage. *Genome Biol.* **18**, 190 (2017).

221. Casini, A. *et al.* A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).

222. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, (2016).

223. Anders, C. & Jinek, M. In vitro enzymology of Cas9. in *Methods in Enzymology* (eds. Doudna, J. A. & Sontheimer, E. J.) vol. 546 1–20 (Academic Press, 2014).

224. Tsai, S. Q., Topkar, V. V., Joung, J. K. & Aryee, M. J. Open-source guideseq software for analysis of GUIDE-seq data. *Nat. Biotechnol.* **34**, 483–483 (2016).

225. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

226. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).

227. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754 (2016).

228. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* **39**, 442–450 (2021).

229. Miller, D. E. *et al.* Targeted Long-Read Sequencing Identifies a Retrotransposon Insertion as a Cause of Altered GNAS Exon A/B Methylation in a Family With Autosomal Dominant Pseudohypoparathyroidism Type 1b (PHP1B). *J. Bone Miner. Res.* **37**, 1711–1719 (2022).

230. Kaiser, C. A., Preuss, D., Grisafi, P. & Botstein, D. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science* **235**, 312–317 (1987).

231. Bird, P., Gething, M. J. & Sambrook, J. Translocation in yeast and mammalian cells: not all signal sequences are functionally equivalent. *J. Cell Biol.* **105**, 2905–2914 (1987).

232. Zheng, T. & Nicchitta, C. V. Structural Determinants for Signal Sequence Function in the Mammalian Endoplasmic Reticulum*. *J. Biol. Chem.* **274**, 36623–36630 (1999).

233. Duffy, J., Patham, B. & Mensa-Wilmot, K. Discovery of functional motifs in h-regions of trypanosome signal sequences. *Biochem. J* **426**, 135–145 (2010).

234. Wu, Z. *et al.* Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).

235. Kober, L., Zehe, C. & Bode, J. Optimized signal peptides for the development of high expressing CHO cell lines. *Biotechnol. Bioeng.* **110**, 1164–1173 (2013).

236. Haataja, L. *et al.* Disulfide Mispairing During Proinsulin Folding in the Endoplasmic Reticulum. *Diabetes* **65**, 1050–1060 (2016).

237. Zhang, H. *et al.* Unpaired Extracellular Cysteine Mutations of CSF3R Mediate Gain or Loss of Function. *Cancer Res.* **77**, 4258–4267 (2017).

238. Woodard, D. R. *et al.* A loss-of-function cysteine mutant in fibulin-3 (EFEMP1) forms aberrant extracellular disulfide-linked homodimers and alters extracellular matrix composition. *Hum. Mutat.* **43**, 1945–1955 (2022).

239. Thornton, J. M. Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287 (1981).

240. Jakob, U., Muse, W., Eser, M. & Bardwell, J. C. Chaperone activity with a redox switch. *Cell* **96**, 341–352 (1999).

241. Choi, H. *et al.* Structural basis of the redox switch in the OxyR transcription factor. *Cell* **105**, 103–113 (2001).

242. Düsterhöft, S. *et al.* Membrane-proximal domain of a disintegrin and metalloprotease-17 represents the putative molecular switch of its shedding activity operated by protein-disulfide isomerase. *J. Am. Chem. Soc.* **135**, 5776–5781 (2013).

243. Chiu, J. & Hogg, P. J. Allosteric disulfides: Sophisticated molecular structures enabling flexible protein regulation. *J. Biol. Chem.* **294**, 2949–2960 (2019).

244. Xu, X., Chiu, J., Chen, S. & Fang, C. Pathophysiological roles of cell surface and extracellular protein disulfide isomerase and their molecular mechanisms. *Br. J. Pharmacol.* **178**, 2911–2930 (2021).

245. Reitsma, P. H., Bertina, R. M., van Amstel, J. K. P., Riemens, A. & Briet, E. The putative Factor IX gene promoter in hemophilia B Leyden. *Blood* **72**, 1074–1076 (1988).

246. Reijnen, M. J., Peerlinck, K., Maasdam, D., Bertina, R. M. & Reitsma, P. H. Hemophilia B Leyden: substitution of thymine for guanine at position -21 results in a disruption of a hepatocyte nuclear factor 4 binding site in the factor IX promoter. *Blood* **82**, 151–158 (1993).

247. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

248. Griesemer, D. *et al.* Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell* **184**, 5247-5260.e19 (2021).

249. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

250. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

251. Zacchi, L. F. *et al.* Coagulation factor IX analysis in bioreactor cell culture supernatant predicts quality of the purified product. *Commun Biol* **4**, 390 (2021).

252. Birket, M. J. *et al.* Contractile Defect Caused by Mutation in MYBPC3 Revealed under Conditions Optimized for Human PSC-Cardiomyocyte Function. *Cell Rep.* **13**, 733–745 (2015).

253. Chen, H. P., Zhao, Y. T. & Zhao, T. C. Histone deacetylases and mechanisms of regulation of gene expression. *Crit. Rev. Oncog.* **20**, 35–47 (2015).

254. Yan, W. X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).

255. Wang, Q. *et al.* A general theoretical framework to design base editors with reduced bystander effects. *Nat. Commun.* **12**, 6529 (2021).

256. Wei, C. T., Peleg, O., Borenstein, E., Maly, D. J. & Fowler, D. M. A versatile, chemically-controlled DNA binding switch enables temporal modulation of Cas9-based effectors. *bioRxiv* 2022.05.10.491425 (2022) doi:10.1101/2022.05.10.491425.

257. Crosetto, N. *et al.* Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365 (2013).

258. Zahid, O. K., Zhao, B. S., He, C. & Hall, A. R. Quantifying mammalian genomic DNA hydroxymethylcytosine content using solid-state nanopores. *Sci. Rep.* **6**, 29565 (2016).

259. Gilboa, T. *et al.* Single-Molecule DNA Methylation Quantification Using Electro-optical Sensing in Solid-State Nanopores. *ACS Nano* **10**, 8861–8870 (2016).

260. Simonova, A. *et al.* Tuning of Oxidation Potential of Ferrocene for Ratiometric Redox Labeling and Coding of Nucleotides and DNA. *Chemistry* **26**, 1286–1291 (2020).

261. Christie, K. A. *et al.* Mutation-Independent Allele-Specific Editing by CRISPR-Cas9, a Novel Approach to Treat Autosomal Dominant Disease. *Mol. Ther.* **28**, 1846–1857 (2020).