Multiplex, multimodal mapping of variant effects in secreted proteins

Nicholas A. Popp[1,2], Rachel L. Powell[1], Melinda K. Wheelock[1,3], Brendan D. Zapp[1], Kristen J. Holmes[4-6], Kathryn M. Sheldon[4-6], Shelley N. Fletcher[7], Xiaoping Wu[13], Shawn Fayer[1,3], Alan F. Rubin[8,9], Kerry W. Lannert[4-6], Alexis T. Chang[1], John P. Sheehan[10], Jill M. Johnsen[4-7,11,14], and Douglas M. Fowler[1,3,12,14]

Affiliations

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
2. Medical Scientist Training Program, University of Washington School of Medicine, Seattle, WA, USA
3. Brotman Baty Institute for Precision Medicine, Seattle, WA, USA
4. Division of Hematology and Oncology, Department of Medicine, University of Washington School of Medicine, Seattle, WA, USA
5. Center for Cardiovascular Biology, University of Washington School of Medicine, Seattle, WA, USA
6. Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA
7. Bloodworks Northwest, Seattle, WA, USA
8. Bioinformatics Division, WEHI, Parkville, VIC, AU
9. Department of Medical Biology, University of Melbourne, Melbourne, VIC, AU
10. Division of Hematology, Medical Oncology, and Palliative Care, Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin
11. Washington Center for Bleeding Disorders, Seattle, WA
12. Department of Bioengineering, University of Washington School of Medicine, Seattle, WA
13. Cell Marker Laboratory, Seattle Children's Hospital, Seattle, WA
14. Co-corresponding authors: jjohnsen@uw.edu, dfowler@uw.edu

Abstract

Despite widespread advances in DNA sequencing, the functional consequences of most genetic variants remain poorly understood. Multiplexed Assays of Variant Effect (MAVEs) can measure the function of variants at scale, and are beginning to address this problem. However, MAVEs cannot readily be applied to the ~10% of human genes encoding secreted proteins. We developed a flexible, scalable human cell surface display method, Multiplexed Surface Tethering of Extracellular Proteins (MultiSTEP), to measure secreted protein variant effects. We used MultiSTEP to study the consequences of missense variation in coagulation factor IX (FIX), a serine protease where genetic variation can cause hemophilia B. We combined MultiSTEP with a panel of antibodies to detect FIX secretion and post-translational modification, measuring a total of 45,024 effects for 9,007 variants. 49.6% of possible *F9* missense variants impacted secretion, post-translational modification, or both. We also identified functional constraints on secretion within the signal peptide and for nearly all variants that caused gain or loss of cysteine. Secretion scores correlated strongly with FIX levels in hemophilia B and revealed that loss of secretion variants are particularly likely to cause severe disease. Integration of the secretion and post-translational modification scores enabled reclassification of ~63% *F9* variants of uncertain significance in the My Life, Our Future hemophilia genotyping project. Lastly, we showed that MultiSTEP can be applied to a wide variety of secreted proteins. Thus, MultiSTEP is a multiplexed, multimodal, and generalizable method for systematically assessing variant effects in secreted proteins at scale.

Introduction

Genome sequencing and clinical genetic testing have revealed a rapidly expanding universe of human genetic variants[1–3]. However, the consequences of the vast majority of these variants with regard to human disease is unknown. Variants with insufficient evidence to interpret pathogenicity are termed Variants of Uncertain Significance (VUS)[4,5]. One way to improve interpretation of VUS is to use functional evidence derived from Multiplexed Assays of Variant Effect (MAVEs)[4,6–8]. MAVEs combine a large library of genetic variants with selection for function and high-throughput DNA sequencing to characterize the functional effects of tens of thousands of variants simultaneously[9]. Because sequencing is used as a readout, every variant protein must remain linked with its cognate variant DNA sequence. Existing yeast, phage and molecular display methods satisfy this requirement by physically linking a target protein to its cognate DNA, but generally do not work for human secreted proteins which require extensive post-translational processing for function. Secreted proteins make up approximately 10% of human genes and the number of VUS in genes encoding secreted proteins is rapidly growing, so a platform for applying MAVEs to human secreted proteins is urgently needed[10] (**Fig. 1a,b**).

One important secreted protein is coagulation factor IX (FIX), encoded by the *F9* gene. Normal FIX secretion and post-translational modification, including γ-carboxylation, are necessary to stop bleeding and achieve hemostasis[11–15]. Genetic variation leading to loss of *F9* function causes the bleeding disorder hemophilia B[16–21]. Undetectable FIX activity, defined as <0.01 IU/dL, characterizes severe hemophilia B and poses high risk of developing neutralizing anti-FIX antibodies[22–24]. *F9* clinical genotyping is recommended for all individuals who have hemophilia B or are at risk to inherit a hemophilia B-causing variant[18,20,21,25,26]. Some pathogenic FIX variants are known to impact secretion (e.g. p.G253Q), while others impact FIX γ-carboxylation (e.g. p.A37D) or activity[23,24,27–30]. However, the effect of the vast majority of the 8,759 possible FIX missense variants is unknown. This knowledge gap is evident in the *My Life, Our Future* U.S. national hemophilia clinical genotyping program, where nearly half of missense variants suspected to cause hemophilia B remain classified as VUS[20,31].

To expand MAVEs to human secreted proteins and to investigate the impact of variants on FIX secretion and post-translational modification, we developed Multiplexed Surface Tethering of Extracellular Proteins (MultiSTEP). MultiSTEP combines multiplexed human cell surface display with a genomically-encoded landing pad to measure the effects of secreted protein variants at scale[32]. We used MultiSTEP to measure both secretion and post-translational γ-carboxylation of nearly all 8,759 possible FIX missense variants using a panel of antibodies. The resulting 45,024 variant effect measurements revealed that 49.6% (n = 4,234/8,528) of FIX missense variants caused loss of FIX function, with 39.5% impacting secretion, and 10.1% impacting γ-carboxylation. Unlike for intracellular proteins, nearly all FIX missense variants resulting in loss or gain of a cysteine diminished FIX secretion, making cysteines the most damaging substitution overall. Secretion scores correlated strongly with FIX levels in hemophilia B and revealed that loss of secretion variants are particularly likely to cause severe disease. Model-based integration of the secretion and post-translational modification scores enabled reclassification of 63% of *F9* VUS in the My Life, Our Future hemophilia genotyping project.

Lastly, we demonstrated that MultiSTEP can be applied to multiple other clinically important secreted proteins including coagulation factors VII, VIII, and X; alpha-1-antitrypsin; plasma protease C1 inhibitor; and insulin.

Results

A multiplexed human cell display system compatible with secreted proteins

Multiplexed Surface Tethering of Extracellular Proteins (MultiSTEP) is a human cell protein display method designed to simultaneously measure the effects of a large number of variants in secreted proteins (**Fig. 1c**). MultiSTEP combines 293-F suspension human cells with a genomically-integrated, recombinase-based landing pad to drive the expression of a single variant protein per cell[32]. Unlike most other mammalian protein display systems, the target protein's endogenous secretion signal peptide is used, ensuring proper processing and allowing characterization of native signal peptide variants. An mCherry fluorescent transcriptional control is expressed via an internal ribosomal entry site (IRES)[32–34]. The secreted protein is fused to a C-terminal transmembrane domain via multipartite, flexible linker containing a strep II tag[35] (**Fig. 1d, Supplementary Fig. 1a,b**).

Using MultiSTEP, we expressed wild type (WT) FIX and detected cell surface-displayed FIX using antibodies directed against the strep II tag in the C-terminal linker, the FIX heavy chain, or the FIX light chain (**Fig. 1e-g**). Cell surface expression of p.C28Y, which drastically reduces signal peptide cleavage efficiency and secretion, was nearly undetectable[36]. p.A37T and p.S220T, neither of which impact FIX secretion, were present on the cell surface at levels comparable to WT[3,16,29,37] (**Fig. 1e-g**). Thus, MultiSTEP enabled FIX display and could be used to distinguish the effect of variants on secretion.

We measured the secretion of 9,007 of the 9,220 possible missense and nonsense *F9* variants using these three secretion antibodies. We computed secretion scores for each antibody by averaging replicates (mean Pearson's r = 0.94), normalizing such that a score of 1 indicated wild type-like secretion and a score of 0 indicated undetectable secretion (**Fig. 2a-c, Supplementary Fig. 1c, Supplementary Fig. 2, Supplementary Fig. 3**). Synonymous variants had secretion scores similar to WT, and missense variants were distributed bimodally, similar to previous multiplexed measurements of cytoplasmic protein variant abundance[34,38,39] (**Fig. 2d,e, Supplementary Fig. 1d**). Measurement of individual variant antibody binding strongly correlated with the MultiSTEP-derived secretion scores for seven randomly selected variants spanning the secretion score range (**Fig. 2f,g, Supplementary Fig. 1e**).

The heavy chain, light chain, and strep tag antibody-derived secretion scores were nearly identical at most positions (**Fig. 2h, Supplementary Fig. 1f,g**). A few positions showed antibody-specific differences: the median heavy chain antibody-derived secretion scores were lower than the median light chain secretion scores at 10 positions in the heavy chain, and the median light chain antibody-derived secretion scores were lower than median heavy chain secretion scores at 10 positions in the light chain (**Fig. 2h**). These two sets of positions

occupied discrete, solvent-accessible sites on FIX structure, strongly suggesting that these positions comprise each antibody's epitope (**Fig. 2i-k**). To prevent epitope effects from impacting subsequent analyses, we excluded secretion scores derived from each antibody at their respective epitope positions.

Factor IX secretion scores clarify biochemical constraints on secretion

Signal peptides direct protein secretion and vary in length and sequence[40–42]. Three distinct functional regions are conserved amongst human signal peptides: the n-region, which is weakly positively charged; the h-region, which contains the hydrophobic helix that binds to a signal recognition particle (SRP) to initiate translocation into the ER; and the c-region, which breaks the h-region helix and contains a canonical AxA signal peptide cleavage motif[41,43–45]. Variants had distinct effects in each region of the signal peptide (**Fig. 3a**). For example, positions p.A26 and p.C28 within the AxA motif tolerated substitution with small amino acids, consistent with these positions occupying shallow hydrophobic pockets near the active site of the signal peptidase[45]. Surprisingly, hydrophobic variants in the spacer of the AxA motif caused significant loss of secretion, potentially by extending the h-region hydrophobic helix and preventing cleavage[41,46,47]. 93.2% (n = 193/207) of variants predicted to be secreted by the signal peptide classifier SignalP 6.0 were also secreted in our assay, with few notable exceptions (e.g. p.E27P). However, 56.4% variants predicted to not to be secreted by SignalP were secreted normally in our assay, including p.M8T, a known benign variant[20] (n = 172/305, **Fig. 3b**).

Many secreted proteins require disulfide bonds to establish and maintain structure, and variants that remove important cysteines can cause disease[48–51]. In FIX, 22 of the 24 native cysteines are disulfide-bonded, including positions p.C178 and p.C335 which form an interdomain disulfide bond critical to maintaining activated FIX integrity after proteolytic cleavage of the activation peptide[52]. Most variants at WT cysteine positions dramatically reduced secretion, especially in the EGF1, EGF2, and serine protease domains, where most positions were completely intolerant of substitution (**Fig. 3c**). This detrimental impact of substitution on cysteines is markedly different from the much more modest effects of substitution of WT cysteines in non-secreted cytoplasmic and transmembrane proteins (**Fig. 3d, Supplementary Fig. 4a**). WT cysteine positions were overwhelmingly the most intolerant of substitution in FIX compared to other positions (**Fig. 3e**).

Missense variants that introduced new cysteines also caused loss of secretion throughout FIX (**Fig. 2b-c**). Introduction of a cysteine was the most damaging substitution, unlike cytoplasmic and membrane-associated proteins where proline or tryptophan were generally most damaging[34,38,39,53] (**Fig. 3f-g, Supplementary Fig. 4b**). Cysteine substitutions were deleterious even within the activation peptide, a highly flexible loop removed during FIX activation where other substitutions did not appreciably impact secretion (**Fig. 2a-c**). Taken together, the deleterious effect of either removing a WT cysteine or introducing a new cysteine strongly suggests that altering the number or location of cysteines in FIX is fundamentally

detrimental to FIX secretion, likely through changes in disulfide bonding patterns or protein folding differences under oxidative stress.

MultiSTEP quantifies the impact of variation on FIX γ-carboxylation

Twelve glutamates in the FIX Gla domain are γ-carboxylated, and γ-carboxylation is required for FIX activation and function. Some of these γ-carboxylated glutamates coordinate a calcium-dependent conformational change in the Gla domain, creating a three-helix structure that exposes the ω-loop[11–14,54–61]. Only 17 of the 106 hemophilia B-associated variants in either the propeptide and Gla domain have been mechanistically shown to alter function of the FIX Gla domain[16,17,23,24,29,37,62–64].

We used two different γ-carboxylation-sensitive, anti-Gla domain antibodies to interrogate the effect of FIX variants on γ-carboxylation. One is a FIX-specific antibody that recognizes the γ-carboxylation-dependent exposure of the ω-loop[58,65,66]. The other is a Gla-motif-specific antibody that interacts with γ-carboxylated glutamates in a conserved ExxxExC motif present in the FIX Gla domain[29,67,68] (**Supplementary Fig. 5a**). As expected, WT FIX-displaying cells strongly bound both antibodies. Incubation of cells with warfarin, a drug that inhibits γ-carboxylation, prior to induction of WT FIX expression eliminated binding of both γ-carboxylation-sensitive antibodies, confirming that they only detect γ-carboxylated FIX[69] (**Supplementary Fig. 5b,c**).

We used the FIX-specific and Gla-motif γ-carboxylation-dependent antibodies to generate two different γ-carboxylation scores for nearly all possible FIX missense variants (**Fig. 4a-c**). Like for secretion, missense variant γ-carboxylation scores were bimodally distributed and synonymous variants scored similarly to WT for both antibodies (**Fig. 4d-e**). γ-carboxylation scores correlated well with secretion scores except in the propeptide and Gla domain, where some variants impacted γ-carboxylation without affecting secretion (Pearson's r = 0.85, **Fig. 4f**). Thus, in our system loss of FIX γ-carboxylation did not impact secretion. Indeed, of the 1,154 missense variants in the propeptide and Gla domains that we assayed, 44.6% (n = 515) were associated with low carboxylation scores but normal secretion scores.

The FIX propeptide mediates binding of gamma-glutamyl carboxylase, the enzyme that γ-carboxylates the Gla domain prior to secretion[54–57]. Propeptide positions p.F31, p.A37, p.I40, p.L41, p.R43, p.K45, and p.R46 are important for Gla domain γ-carboxylation, though the importance of p.R43, p.K45, and p.R46 is contested[16,17,29,30,70–72]. Some substitutions at the highly conserved position p.A37 in the propeptide (e.g. p.A37D) completely ablate gamma-glutamyl carboxylase binding and thus prevent Gla domain γ-carboxylation[29,73]. In our assay, 12 of the 19 substitutions at position p.A37, including p.A37D, reduced γ-carboxylation scores for both antibodies. Serine, threonine, and glutamate substitutions at position p.I40, which is conserved across all Gla-containing coagulation proteins, also reduced γ-carboxylation scores (**Fig. 4b,c**). Thus, like position p.A37, position p.I40 appears to be important for γ-carboxylation. The propeptide furin cleavage motif (p.R43 to p.R46) harbors numerous severe hemophilia B-associated variants[23,24] (**Fig. 4a-c**). Variants in the furin cleavage motif can block

propeptide cleavage, leading to its retention[74,75]. The impact of propeptide retention on Gla domain γ-carboxylation remains controversial, with some data suggesting that γ-carboxylation itself is blocked by the retained propeptide and other data suggesting that the retained propeptide alters the conformation of the ω-loop by disrupting electrostatic interactions within the Gla domain[11,29,70,72,74–76]. In our assay, variants in the furin cleavage motif had low scores for the FIX-specific γ-carboxylation antibody but WT-like scores for the Gla-motif antibody and the secretion antibodies (**Fig. 2b,c**; **Fig. 4b,c,f**). One possible explanation for discordance between the FIX-specific and Gla-motif antibody scores at propeptide cleavage motif positions is that these substitutions perturb ω-loop conformation but permit γ-carboxylation of the glutamates in the conserved Gla-motif.

We compared the FIX-specific Gla domain and Gla-motif γ-carboxylation antibody scores for each variant within the Gla domain and found that many variants in the ω-loop (positions p.N49 to p.Q57) had WT-like scores for both antibodies (**Fig. 4b,c**). The lowest FIX-specific Gla domain antibody γ-carboxylation scores resulted from substitutions at positions involved in the coordination of calcium ions (p.N48, p.E53, p.E54, p.E67, p.E73, and p.E76) in the Gla domain core[14] (**Fig. 4g**). Variants at positions p.F87 and p.W88, which form an aromatic stack required for divalent cation-induced folding of the Gla domain, had low FIX-specific γ-carboxylation scores[77,78] but WT-like secretion scores likely due to disrupted γ-carboxylation, Gla domain conformation, or both (**Fig. 4f**). Unexpectedly, variants at half of the γ-carboxylated glutamate residues in the Gla domain (p.E66, p.E72, p.E79, p.E82, p.E86) did not impact or only modestly reduced FIX-specific Gla domain and Gla-motif antibody γ-carboxylation scores (**Fig. 4b,c,g**). These glutamate residues coordinate magnesium rather than calcium ions at physiologic concentrations[13,79–81] (**Fig. 4g**).

Secretion scores correlate with circulating FIX levels

We previously established that many pathogenic missense variants in intracellular proteins act by disrupting protein abundance[34,38,39,53]. However, for secreted proteins, the predominance of loss of secretion variants as a pathogenic mechanism remains largely unexplored. Thus, we investigated whether FIX secretion scores could predict circulating FIX antigen levels in individuals with hemophilia B. Secretion scores correlated strongly with plasma FIX antigen levels reported in the European Association for Haemophilia and Allied Disorders (EAHAD) public database[23,24] (n = 85, Pearson's r = 0.77; **Fig. 5a**). We classified variants with secretion scores less than the 5th percentile of synonymous variants as lowly secreted[34,39,53]. 90.2% (n = 37/41) of lowly secreted variants had low circulating FIX levels, defined as <40% of normal (normal usually ranges from 50-150%)[23]. The four lowly secreted variants, p.R75Q, p.R191C, p.R379G, and p.A436V, that had normal plasma FIX levels *in vivo* had secretion scores or reported FIX levels close to the assay thresholds, suggesting measurement error as the likely explanation (**Supplementary Table 1**). Conversely, five variants, p.F71S, p.D93N, p.A173V, p.Q241H, and p.N393T had markedly increased secretion scores relative to their FIX antigen levels (**Supplementary Table 1**). These variants occured throughout the protein, outside of known binding sites for extravascular compartments or clearance binding

partners[82–85]. These variants could be secreted normally but have a shortened half-life *in vivo*, which would be missed by our assay.

Another important feature of hemophilia B is disease severity, defined by the level of FIX activity in plasma[20]. Severe hemophilia B is defined by undetectable FIX activity (<1%). Individuals with severe hemophilia B receive more intensive treatment and are at higher risk for complications than individuals with non-severe hemophilia B (FIX activity 1-40%)[20,21]. To better clarify the relationship of variant effect on FIX secretion and hemophilia disease severity, we analyzed MultiSTEP secretion scores relative to disease severity reported for variants in EAHAD[23,24]. Variant secretion scores correlated strongly with hemophilia B severity (**Fig. 5b**). We found 42.6% (n = 110/258) of severe disease-associated variants had very low secretion scores (<0.05), while only 2.7% (n = 3/113) of mild disease-associated variants and 12.6% (n = 15/119) of moderate disease-associated variants had very low secretion scores. Moreover, FIX antigen levels reported in hemophilia B were markedly decreased for the majority of severe disease variants, compared to those causing mild or moderate disease (**Fig. 5c**). Thus, loss of secretion is an important mechanism by which *F9* variants cause hemophilia B, and variants that cause very low or undetectable secretion are particularly likely to cause severe disease.

Secretion and γ-carboxylation scores resolve FIX variants of uncertain significance

F9 VUS are a major problem in hemophilia B, with the VUS classification almost always resulting from insufficient evidence to inform interpretation of variant pathogenicity with certainty. This evidence is lacking in part because of the rarity of the disease and in part because of very rare and novel variants discovered during sequencing[20,21]. In the *My Life, Our Future* (MLOF) genotyping program[20], 48.6% (n = 107/220) of *F9* variants clinically suspected to cause hemophilia B were classified as VUS due to lack of evidence, of which nearly half (46.8%) were missense. To resolve *F9* VUS in hemophilia B, we curated a set of 149 *F9* missense variants of known effect (135 pathogenic/likely pathogenic, 14 benign/likely benign) from ClinVar, MLOF, and gnomAD[1,3,20]. We found that no single secretion or γ-carboxylation score set perfectly separated pathogenic and benign variants (**Fig. 5d**). Thus, we integrated the secretion and γ-carboxylation scores by training a random forest model to classify variant function as either normal or abnormal. In our test set, the model classified 61.8% (n = 21) of known pathogenic variants as abnormal function and 100% (n = 4) of known benign variants as having normal function (**Fig. 5e**). When applied to all 8,934 *F9* missense variants with secretion and γ-carboxylation scores, the model classified 43.5% (n = 3,922) as functionally abnormal (**Supplementary Table 2**).

We reassessed *F9* missense variants reported as VUS in MLOF using evidence from our predictive model, applying strong evidence strength (PS3) to variants predicted to have abnormal function[4,20]. This resulted in reclassification of 63.1% (n = 65/103) of VUS to likely pathogenic and 67.3% (n = 66/98) of likely pathogenic variants to pathogenic (**Fig. 5f, Supplementary Table 3**). An alternative Bayesian approach suggested that our model yielded moderate evidence of pathogenicity (PS3_moderate)[6,86,87]. Using moderate evidence strength, 61.2% (n = 63/103) of VUS were reclassified as likely pathogenic and 13.3% (n = 13/98) of likely

pathogenic variants were reclassified as pathogenic (**Fig. 5f, Supplementary Table 3**). Because our predictive model incorporated only secretion and γ-carboxylation functional data, it cannot account for variants that cause loss of other FIX functions and hemophilia B. Therefore, we did not use normal function predictions as evidence for benign classification. Taken together, functional evidence from MultiSTEP drastically reduced the number of *F9* VUS and lent new insight into how *F9* variants affect FIX function.

MultiSTEP can be used to study diverse secreted proteins

We applied MultiSTEP to six additional secreted proteins, coagulation factors VII, VIII, and X, alpha-1 antitrypsin, plasma protease C1 inhibitor, and proinsulin. This diverse set of proteins ranged in size from 110 to 1,670 amino acids and included enzymes, protease inhibitors, and a peptide hormone. Antibody staining of the MultiSTEP Strep II epitope tag revealed robust surface expression in each case (**Fig. 6a**). Negative controls, including the CD28 transmembrane domain alone (i.e. without a fusion protein), deletion of the FIX start codon, and deletion of the FIX signal peptide, did not yield any appreciable signal above that found in parental cell lines. To provide further evidence that proteins displayed using MultiSTEP were properly folded, we focused on factor VIII which has been successfully expressed for therapeutic use in HEK-293 cells[88,89]. Displayed factor VIII robustly bound five monoclonal antibodies directed against the five domains of FVIII (**Supplementary Fig. 6**) and two monoclonal antibodies that recognize discontinuous FVIII epitopes, confirming that displayed FVIII is folded correctly (**Fig 6b,c**).

Discussion

MultiSTEP is a generalizable, multiplexed method for measuring the effects of variants on the function of human secreted proteins. MultiSTEP can be combined with diverse antibodies to quantify secretion and post-translational modifications, which until now have been beyond the reach of MAVEs. We applied MultiSTEP to coagulation factor IX, a protein critical for hemostasis, measuring 45,024 variant effects on secretion and post-translational γ-carboxylation for 9,007 variants using a panel of five antibodies. We revealed that nearly half of FIX missense variants caused some degree of loss of function, mostly by reducing or eliminating secretion. In particular, unlike for intracellular proteins, either adding or removing cysteines profoundly impacted FIX secretion, likely reflecting the importance of correct disulfide bonding[34,38,39,53]. Most variants outside of the propeptide and Gla domain did not impact γ-carboxylation, and propeptide and Gla domain variants that reduced γ-carboxylation generally did not impact secretion. In the propeptide, our results suggested propeptide retention alters ω-loop conformation while permitting the γ-carboxylation of at least some glutamates[29,75,76]. In the Gla domain, γ-carboxylated glutamates do not appear to play equal roles, with those that coordinate calcium or comprise the C-terminal aromatic stack being markedly more sensitive to variation than those that coordinate magnesium, consistent with the importance of calcium coordination for Gla domain folding and function[23,24,77,78].

Genetic variation in *F9* is the primary cause of hemophilia B, and pathogenic missense FIX variants can disrupt protein secretion or function. Variant secretion correlated strongly with circulating FIX antigen levels in people. More than 40% of variants associated with severe disease had profound secretion defects, illustrating the importance of loss of secretion as a causal mechanism of hemophilia B. We integrated secretion and γ-carboxylation scores using a model that predicted FIX variant function. Using model predictions, we reclassified 65 of 103 *F9* missense VUS reported in the My Life, Our Future hemophilia genotyping program as likely pathogenic. We also upgraded 66 likely pathogenic variants to pathogenic. Model predictions are available for 8,528 (97.4%) FIX missense variants, meaning that newly discovered variants in hemophilia B can be more accurately classified.

MAVE-compatible *E. coli* and *S. cerevisiae* cell surface display methods have typically been applied to intracellular proteins[90–93]. Moreover, these methods are not generally suitable for human secreted proteins which require extensive post-translational modification. Mammalian cell surface display systems that allow for the study of post-translational modifications rely on non-native signal peptides[94–96]. The use of alternative or artificial signal peptides prevents accurate assessment of the functional effects of native signal peptide variants and can also disrupt protein folding, trafficking, localization, or post-translational modification[40,41,97]. Approaches to display IgG antibodies using endogenous signal peptides exist, but, unlike MultiSTEP, multiple variants per cell can be integrated, potentially obscuring variant functional effects[94,98,99]. Thus, by enabling multiplexed human cell surface display of secreted protein variants, MultiSTEP satisfies a unique and important demand.

While our FIX variant secretion and γ-carboxylation results illuminate various structural and biochemical features of FIX, identify pathogenic variant mechanisms, and improve clinical variant interpretation, MultiSTEP has limitations. For example, while all seven of the secreted proteins we displayed using MultiSTEP were expressed on the cell surface, the folding, secretion or function of some other proteins may be compromised by linkage to the C-terminal transmembrane domain fusion. Moreover, FACS-based assays like MultiSTEP have a limited dynamic range that can obscure subtle variant effects. Lastly, no functional assay fully recapitulates *in vivo* biology. For example, we did not measure FIX enzymatic activity in our system. Therefore, we discourage using our results as evidence for benign variant classification in hemophilia B.

MultiSTEP is a generalizable platform for measuring the effects of missense variation in secreted proteins, complementing the plethora of MAVEs already developed to study intracellular proteins. Because MultiSTEP is compatible with reagents that produce a fluorescent signal, we were able to probe multiple functional aspects of FIX using a panel of antibodies. Antibodies are available for numerous clinically and biologically important secreted proteins, as well as a variety of post-translational modifications. Moreover, other fluorescently labeled reagents such as protein binding partners or labeled covalent substrates could be used to read out other functions. The ability to characterize the diverse effects of massive numbers of secreted human protein variants on the surface of human cells creates the opportunity to understand variant effects on secreted protein structure, function, and pathogenicity.

Methods

General reagents

Chemicals were purchased from ThermoFisher Scientific and synthetic oligonucleotides were purchased from IDT unless otherwise noted. Sequences of oligonucleotides used in this work can be found in **Supplementary Table 4**.

All *E. coli* were grown in Luria Broth (LB) at 37°C with 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* (New England Biolabs).

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 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 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*[100]. 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[101].

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 isolated using miniprep or midiprep kits according to the manufacturer's instructions (Qiagen).

Cloning into the MultiSTEP 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 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. This plasmid was used as Δstart in **Fig. 6**. To remove the *F9* signal peptide (pNP0010), primers NP0295 and NP0471 were used following the standard IVA cloning protocol above. To create a transmembrane-only plasmid (pNP0009), primers NP0325 and NP0470 were used following the standard IVA protocol above.

To clone missense *F9* variants for assay validation, single amino acid variants were introduced into pNP0001 by IVA cloning following the same protocol as above. For each variant, the following primers were used: p.C28Y (pNP0011): NP0488 and NP0489; p.A37T (pNP0003): NP0293 and NP0294; p.G58E (pNP0027): NP0225 and NP0226; p.E67K (pNP0019): NP0217 and NP0218; p.C134R (pNP0080): NP0591 and NP0592; p.S220T (pNP0057): NP0478 and NP0479; p.H267L (pNP0065): NP0273 and NP0274.

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 pNP0088, *F7* cDNA was amplified with NP0615 and NP0616. To generate pNP0089, *F10* cDNA was amplified with NP0617 and NP0618. To generate pNP0090, *SERPINA1* cDNA was amplified with NP0619 and NP0620. To generate pNP0091, *SERPING1* cDNA was amplified with NP0621 and NP0622. To generate pNP0092, *INS* cDNA was amplified with NP0623 and NP0624.

pcDNA4/Full Length FVIII was a gift from Robert Peters (Addgene #41036). To generate attB-F8-10L-StrepII-10L-CD28-IRES-mCherry (pNP0084), an inverse PCR was performed on pcDNA4/Full Length FVIII using NP0398 and NP0399 to isolate *F8.* pNP0079 backbone was also inverse PCR amplified using NP0400 and NP0401 to add *F8* overhangs. The products were Gibson assembled using the standard cloning protocol above.

Because full length *F8* is poorly expressed both in cell culture and *in vivo*, we adapted a B-domain deletion version of *F8* cDNA shared by Dr. Steven Pipe (University of Michigan) that has improved coagulation factor VIII expression and retains FVIII function[102–104]. Primers NP0405 and NP0406 were used to create B-domain deleted *F8* (pNP0085) using the IVA cloning protocol above.

Site-saturation mutagenesis library cloning

Site-saturation mutagenesis oligonucleotide pools 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 p.Q2-p.K164; Tile 2 sublibrary: positions p.A146-p.I318; Tile 3 sublibrary: positions p.L299-p.T461.

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 plasmid DNA isolated using a midiprep kit according to the manufacturers instructions (Qiagen).

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 100 V and gel extracted (Qiagen).

An IDT Ultramer (NP0490) with 18 degenerate nucleotides was 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 100 V 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*g* and plasmid DNA isolated using a midiprep kit according to the manufacturer's instructions (Qiagen).

Estimation of barcoded variant coverage by Illumina sequencing

Each barcoded and bottlenecked plasmid sublibrary was diluted to 10 ng/µL and amplified for Illumina sequencing to more accurately determine the number of unique barcodes present. 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, 4.75 µL water, and one uniquely indexed forward (NP0565-NP0578) 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 120 V 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), and unique barcodes were counted and filtered[105].

PacBio sequencing of FIX sublibraries for variant-barcode mapping

2.5 µg of each library subtile 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 WT *F9* cDNA sequence using BWA-MEM v0.7.10-r789, generating CIGAR and MD strings, which are then used to extract barcodes and the variable region containing *F9*[106]. The output from AssemblyByPacBio was then passed through PacRAT, which takes all CCS reads containing the same barcode and performs multiple sequence alignment to improve variant calling[107]. Alignments with a variant agreement threshold of less than 0.6 or fewer than 3 independent CCS reads were filtered from further analysis, resulting in 260,224 unique barcodes across all three sublibraries. A custom R script was then used to parse the PacRAT output and generate a final barcode-variant map, containing 9,022 of the 9,220 expected missense, nonsense, and synonymous variants.

General cell culture conditions

HEK293T 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 3 x 10^5 cells/mL once reaching a concentration between 1 x 10^6 and 2 x 10^6 cells/mL, unless otherwise stated. All Freestyle 293-F cells containing a landing pad were induced with 2 µg/mL doxycycline (Sigma-Aldrich).

Lentiviral transduction to generate suspension HEK293 landing pad line

To generate a landing pad lentiviral vector, 2.5 x 10^5 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[32].

1 x 10^7 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.

Fluorescence-activated cell sorting 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. A 410 nm long pass filter preceded the 450/50 nm bandpass filter on the Symphony A3. mCherry expressed 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 and

Symphony A3) long pass and 610/20 nm bandpass filters. EGFP or Alexa488-labeled antibodies were excited using the 488 nm laser and captured using a 505 nm long pass and 530/30 nm bandpass filters (LSR II and Aria III) or a 515/20 bandpass filter (Symphony A3). Alexa647-labeled antibodies were excited using a 637 nm (LSR II and Symphony A3) or 640 nm laser (Aria III) and captured using a 670/30 nm bandpass filter (LSR II and Aria III) with an additional 650 nm long pass filter for the Symphony A3. All flow cytometry data were collected with FACSDiva v.8.0.1 and analyzed using FlowJo v.10.7.1.

Recombination of Freestyle 293-F cells

Freestyle 293-F cells were transfected at 1 x 10^6 cells/mL with 293Fectin 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, 1 x 10^7 cells in 10 mL were transfected. For libraries, 3 x 10^7 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 (AP1903) 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. These 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.

Antibody staining for surface-displayed proteins

For all secretion antibodies, cold PBS + 1% BSA was used as a staining and dilution 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 into PBS + 1% BSA was used as a staining and dilution buffer for γ-carboxylation antibodies. See **Supplementary Table 5** for additional antibody details, stock concentrations, and working dilutions.

Cells were plated at 3 x 10^5 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 γ-carboxylation antibodies, one control flask of WT FIX cells was additionally supplemented with 100 nM warfarin (Sigma-Aldrich) to reduce 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 bin.

After sorting, each 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 5 x 10^5 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.

Genomic DNA prep, barcode amplification, and sequencing

Genomic DNA was prepared according to previously described protocols[34,39,53]. Briefly, genomic DNA was extracted from harvested cell pellets using the DNEasy Blood and Tissue kit (Qiagen) according to the manufacturer's 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 (mean r = 0.95, **Supplementary**

Table 6). The protocol outlined above (Estimation of barcoded variant coverage by Illumina sequencing section) 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 amplified for a maximum of 20 cycles or until all samples were above 3,000 relative fluorescence units. 4) Samples were sequenced using custom sequencing primers NP0494, NP0495, NP0497, and NP0550 on either 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 cycle kit (Illumina).

Calculating scores and classifications

Using a custom script, barcode sequencing reads were first converted to FASTQ format and demultiplexed using bcl2fastq (v2.20). Next, forward and reverse barcode reads were paired using PEAR (v0.9.11), and unique barcodes were counted[105]. To reduce error associated with counting noise, we removed low frequency variants below a threshold of 1×10^{-6} as well as variants that were observed in fewer than two replicates. FIX variants were then assigned to sequenced barcodes using the barcode-variant map from PacBio sequencing. Any barcode associated with insertions, deletions, or multiple amino acid substitutions in FIX was removed from the analysis.

Secretion and γ-carboxylation scores were calculated by using a modification of a previous analysis pipeline[34,39,53]. Briefly, for each experiment, a weighted average of every variant's frequency in each bin was calculated with the following weights: $w_{bin\ 1}$: 0.25, $w_{bin\ 2}$: 0.5, $w_{bin\ 3}$: 0.75, $w_{bin\ 4}$: 1. The weighted average for each variant was then min-max normalized such that the median score of WT barcodes was 1 and the median score of the lowest 5th percentile of missense variants was 0. The final secretion and γ-carboxylation scores for each variant were then averaged across all replicates. Standard errors for each score were calculated by dividing the standard deviation of the min-max normalized values for each variant by the square root of the number of replicate experiments in which it was observed.

Description of computational methods

VAMP-seq abundance scores for PTEN (urn:mavedb:00000013-a-1), TPMT (urn:mavedb:00000013-b-1), VKOR (urn:mavedb:00000078-b-1), CYP2C9 (urn:mavedb:00000095-b-1), and NUDT15 (urn:mavedb:00000055-a-1) were downloaded from MAVEdb[108]. Gla domain protein sequences for human coagulation factor IX (P00740), human prothrombin (coagulation factor II, P00734), human coagulation factor VII (P08709), human coagulation factor X (P00742), human protein C (P04070), human protein S (P07225), human osteocalcin (bone gla-protein, P02818), bovine osteocalcin (bone Gla-protein, P02820), human growth arrest-specific protein 6 (P14393), *Pseudechis prophyriacus* venom prothrombin activator porpharin-D (P58L93), *Notechis scutatis* venom prothrombin activator notecarin-D1 (P82807), and *Oxyuranus scutellatus* venom prothrombin activator oscutarin-C (P58L96) were

downloaded from UniProt and aligned using MUSCLE[109]. Signal peptide predictions were generated by submitting variant and WT FIX sequences to the SignalP 6.0 server[110].

Clinical variant curation and re-classification

Our control set of pathogenic and benign FIX missense variants were collected from ClinVar, MLOF, and gnomAD v.3.1[1,3,20]. ClinVar variants with conflicting classifications or if evidence was not provided were removed. As the incidence of hemophilia B is approximately 3.8 to 5 per 100,000 male births, we deemed any gnomAD FIX variants with minor allele frequency in hemizygotes of greater than 1 per 1,000 as benign[34,111–113]. Lastly, *F9* variants found in individuals tested in MLOF were deemed benign if clinical testing demonstrated normal FIX activity in a hemizygous male. The final curated dataset contained 135 pathogenic or likely pathogenic variants and 14 benign or likely benign variants, sufficient for classification model training purposes[6,86,87].

Data on FIX antigen levels, FIX activity, and disease severity reported in individuals with hemophilia B were collected from EAHAD and collapsed based on variant, resulting in 674 final variants with at least one of the three clinical data points[23,24]. Because FIX antigen and activity are often reported as “<1%”, “<1 IU/dL”, or “undetectable” when below the limit of detection in clinical assays, we gave variants with these clinical assay results a value of 0.1% for the purposes of our analyses. Values for FIX antigen or FIX activity were averaged across reported individual values for each variant. 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, the variant was removed from the analysis. Further, if a variant’s antigen, activity, or severity was only reported in a single individual, that variant was removed from further analysis.

Random forest classifier

We built a random forest classifier to distinguish between true benign and pathogenic variants using the ranger and tidymodels packages in R[114]. 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 in our training dataset (8.4% benign/likely benign and 90.6% pathogenic/likely pathogenic), we performed class-based random oversampling (ROSE) and trained the model using 5-fold cross-validation[115]. Model performance was evaluated using ROC curves generated with the reserved test set. The final model was applied to 103 VUS from MLOF that were observed in our experimental data. We used ACMG rules-based guidelines to reinterpret variant classifications using *in vitro* functional data as strong evidence of pathogenicity[4,20]. As an alternative, we also employed a Bayesian framework to ACMG rules-based guidelines[6,86,87]. We calculated an odds of pathogenicity (OddsPath) and applied *in vitro* functional data as moderate evidence of pathogenicity. All other evidence codes were applied according to ACMG adapted for an X-linked inherited monogenic mendelian disorder[20]. All 103 variants and their updated classifications are provided in **Supplementary Table 3**.

Data and code availability

F9 variant scores are available in **Supplementary Table 7**. Raw Illumina sequencing, PacBio sequencing, and barcode-variant maps, and variant scores will be available upon publication in the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE242805. All other source data and code to reproduce figures and analyses presented in this work will be made available on GitHub upon publication.

.

Acknowledgements

We thank A.P. Leith, C. Lee, D.E. Prunkard, and A. Silvestroni of the UW Foege Flow Lab and the UW Pathology Flow Cytometry Core for their assistance with cell analysis, staining, and sorting; K.M. Munson of the UW PacBio Sequencing Service for assistance with long-read PacBio sequencing; D.A. Nickerson, L.M. Starita, D.J. Maly, S. Nariya, J.J. Stephany, and A.E. McEwen for advice on analyzing data and feedback on the manuscript. We thank S.W. Pipe and A. Scheller of the University of Michigan Department of Pediatrics and Department of Hematology for providing FVIII constructs and advice on FVIII expression. We thank J. Kulman for discussions of FIX carboxylation. We thank and acknowledge B.A. Konkle the PI of MLOF, the MLOF partners at Bloodworks, the American Thrombosis and Hemostasis Network, the National Hemophilia Foundation, funding from Biogen/Bioverativ, providers and staff at HTC sites, and the 11,341 participants who made MLOF a success. This work was supported by the National Heart, Lung, and Blood Institute (5R01HL152066 to J.M.J. and D.M.F., 5F30HL151075 to N.A.P., and 5R01HL149855 to J.P.S.), the National Human Genome Research Institute (1RM1HG010461 to D.M.F.), the National Institute of General Medical Sciences (5R01GM109110 to D.M.F.), and the Washington Center for Bleeding Disorders (to J.M.J).

Author contributions

N.A.P., K.W.L., J.P.S., J.M.J., and D.M.F. conceived of the work. N.A.P., J.P.S., J.M.J., and D.M.F. wrote the manuscript. N.A.P., R.L.P., M.K.W, B.D.Z., K.J.H., K.M.S., X.W., K.W.L., and A.T.C. performed experiments. N.A.P. performed the statistical and computational analysis. N.A.P., S.N.F., and J.M.J. manually curated ClinVar, gnomAD, and MLOF for variants. N.A.P. and A.F.R. prepared results for distribution. N.A.P., S.N.F., S.F., and J.M.J. performed variant reinterpretation. All authors approved the final manuscript.

Figures and Tables

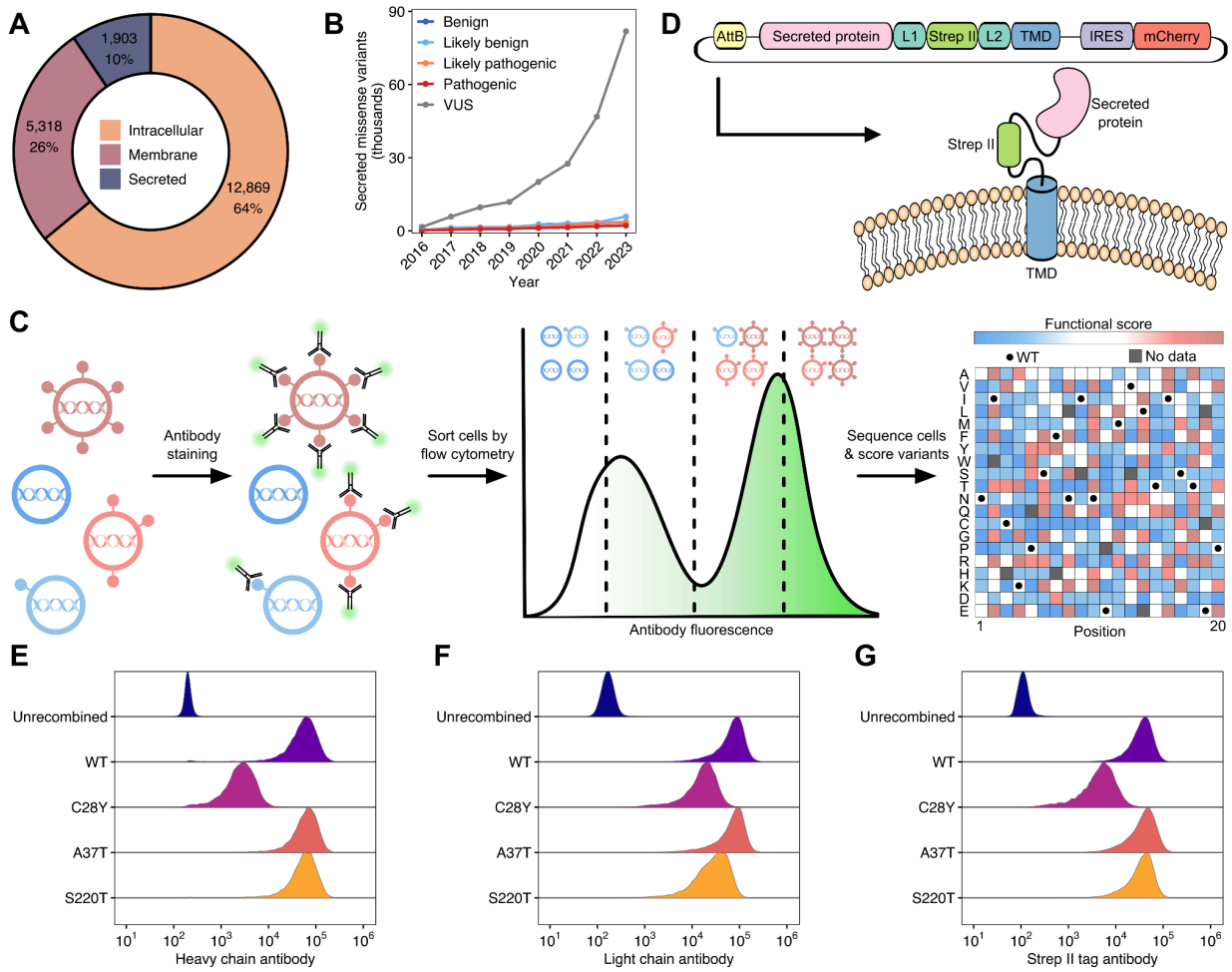


Figure 1: MultiSTEP enables at-scale measurement of variant effects in secreted proteins.

A. Secreted proteins (purple) make up approximately 10% of the human proteome[10]. **B.** Missense variants in secreted proteins found in ClinVar from 2016 to 2023 colored by clinical significance. **C.** MultiSTEP retains secreted proteins on the cell surface, establishing a physical link between genotype and phenotype (left panel). Cells expressing a library of variants of the target protein are sorted into bins based upon intensity of fluorescent antibody binding, followed by deep sequencing to derive a functional score for each individual variant (middle panels). The result is a variant effect map (right panel). **D.** MultiSTEP design. Secreted protein coding sequences (pink) are cloned into an attB-containing landing pad donor plasmid. Secreted proteins are engineered to have C-terminally fused $(GGGGS)_2$ flexible linkers (L1 and L2, teal) attached to a single pass transmembrane domain (TMD, blue). In between the

linkers is a strep II epitope tag for surface detection (green). The construct contains an IRES (purple) driving co-transcription of an mCherry fluorophore (red) that serves as a transcriptional control. **E-G.** Flow cytometry of known well-secreted (p.A37T, p.S220T, WT) and poorly-secreted (p.C28Y) FIX variants displayed using MultiSTEP (n ~30,000 cells per variant). Unrecombined cells do not display FIX and serve as a negative control. Fluorescent signal was generated by staining the library with either a mouse monoclonal anti-FIX heavy chain antibody (**E**), a mouse monoclonal anti-FIX light chain antibody (**F**), or a mouse monoclonal anti-strep II tag antibody (**G**), followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody.

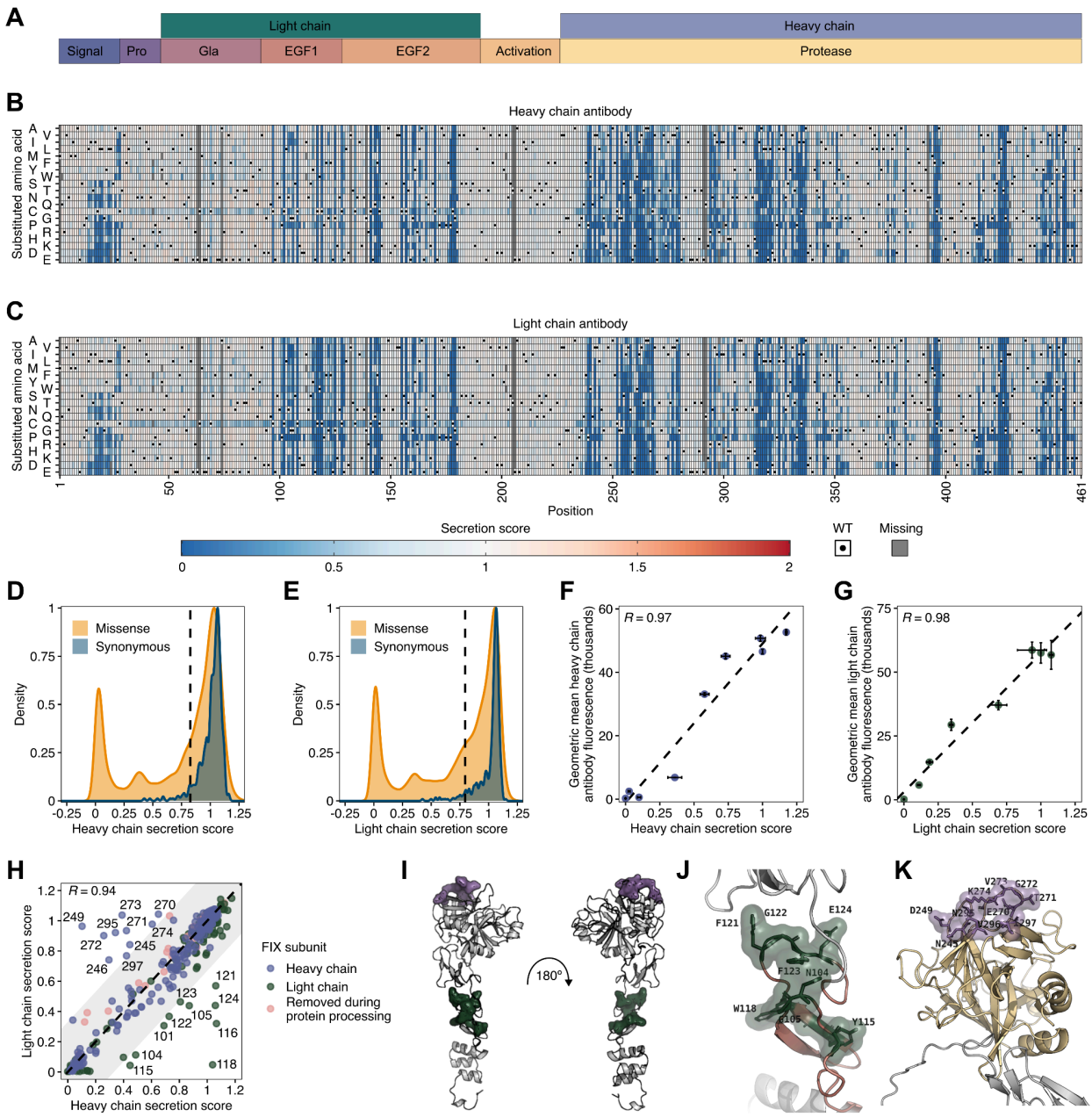


Figure 2: 18,009 MultiSTEP-derived secretion scores for 9,007 factor IX variants.

A. Factor IX domain and chain architecture. Signal: Signal peptide. Pro: Propeptide. Gla: Gla domain. EGF1: Epidermal growth-like factor 1 domain. EGF2: Epidermal growth-like factor 2 domain. Activation: Activation peptide. Protease: Serine protease domain. **B-C.** Heatmaps showing FIX heavy chain secretion scores (**B**) or FIX light chain secretion scores (**C**) for nearly all missense FIX variants. Heatmap color indicates antibody score from 0 (blue, lowest 5% of scores) to white (1, WT) to red (increased scores).

Black dots indicate the WT amino acid. Missing data scores are colored gray. **D-E.** Density distributions of heavy chain (**D**) or light chain (**E**) secretion scores for FIX missense variants (orange) and synonymous variants (blue). Dashed line denotes the 5th percentile of the synonymous variant distribution. **F-G.** Scatterplots comparing MultiSTEP-derived heavy chain (**F**) or light chain (**G**) secretion scores for seven different variants to the geometric mean of Alexa Fluor-647 fluorescence measured using flow cytometry on cells expressing each variant individually. Error bars show standard error of the mean (n = 3; >10,000 cells per replicate). Black dashed line indicates the line of best fit. Pearson's correlation coefficients are shown. **H.** Scatterplot of median MultiSTEP-derived heavy chain and light chain secretion scores at each position in FIX. Points are colored by chain architecture, using the same color scheme as **A**. Black dashed line indicates the line of perfect correlation between secretion scores. Pearson's correlation coefficients are shown. Gray background indicates <30% deviation from perfect correlation. Points with median positional scores outside gray background are labeled with their corresponding FIX position. **I.** AlphaFold2 model of mature, two-chain FIX (positions 47-191 and 227-461). Positions labeled in (**H**) are shown as colored surfaces where color corresponds to the FIX heavy chain (magenta) or light chain (green). **J.** Magnified view of FIX EGF1 domain in the light chain (orange). Putative epitope positions with discordant light and heavy chain antibody scores (**H-I**) are shown as a green colored surface with visible amino acids labeled. **K.** Magnified view of FIX serine protease domain in the heavy chain (yellow). Putative epitope positions (**H-I**) are shown as a purple colored surface with visible amino acids labeled.

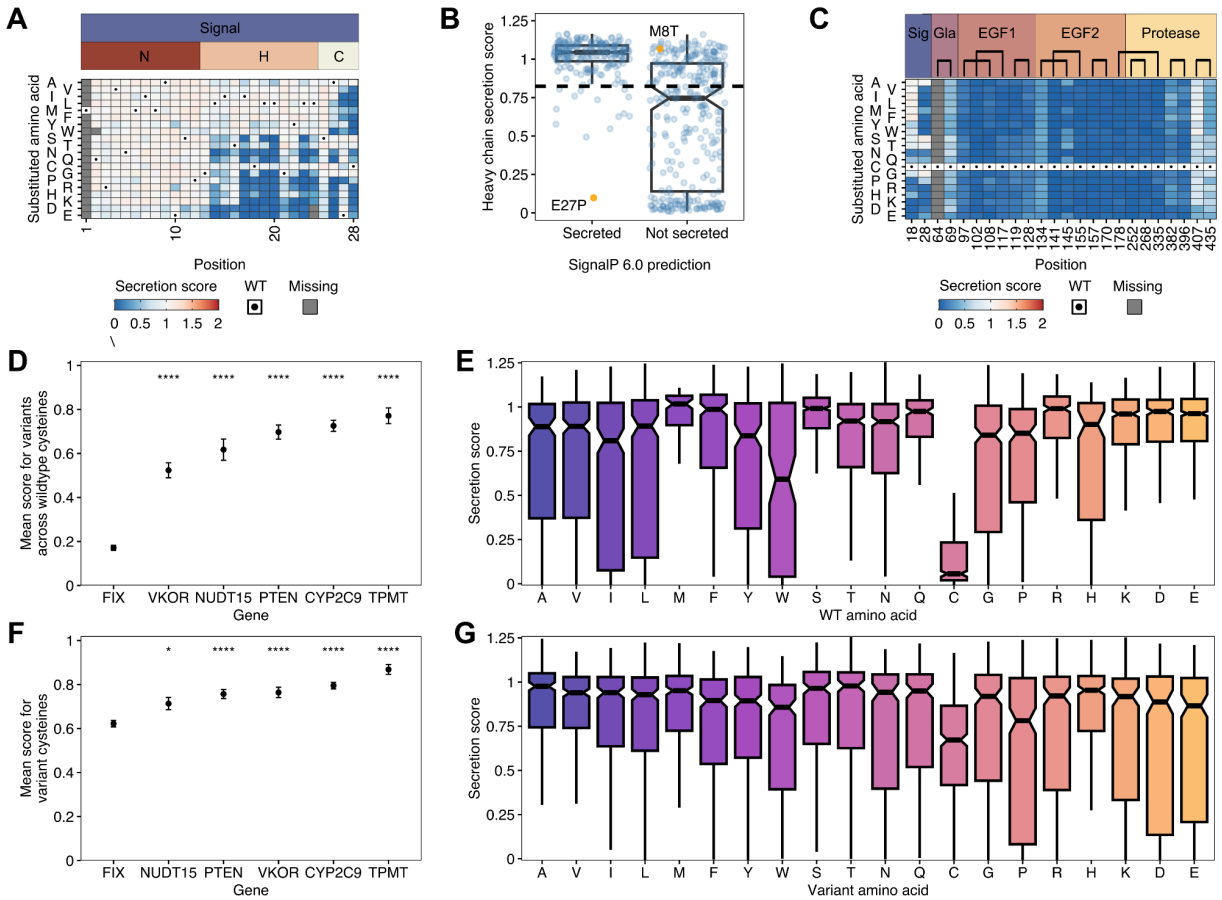


Figure 3: MultiSTEP reveals biochemical constraints on secretion.

A. Predicted signal peptide regions for WT FIX from SignalP 6.0, indicated by color (top). Heatmap showing FIX heavy chain secretion scores for signal peptide variants (bottom). Heatmap color indicates antibody score from 0 (blue, lowest 5% of scores) to white (1, WT) to red (increased antibody scores). Black dots indicate the WT amino acid. Missing scores are gray. N: n-region; H: h-region; C: c-region. **B.** Secretion antibody scores for variants predicted by SignalP 6.0 to be either secreted or not secreted. **C.** FIX cysteine variants colored by domain architecture (top). Sig: Signal peptide. Gla: Gla domain. EGF1: Epidermal growth-like factor 1 domain. EGF2: Epidermal growth-like factor 2 domain. Protease: Serine protease domain. Disulfide bridges in WT FIX are denoted by black connecting lines[14,116,117]. Heatmap of FIX heavy chain secretion scores for variants at positions with a cysteine in the WT sequence, colored as in (**B**) (bottom). **D.** Mean variant effect scores at all positions with a WT cysteine for different proteins. For FIX, the average secretion score determined by MultiSTEP is shown. For all other proteins, the average abundance score measured using VAMP-seq is shown[34,38,39,53]. Points indicate mean abundance or

secretion score for all scored variants at all WT cysteine positions. Error bars show standard error of the mean. Asterisks indicate level of statistical significance as determined by Bonferroni-corrected pairwise two-sided t-test to FIX variant scores. **** = p <0.0001. **E.** Boxplots of secretion scores for all missense variants across all positions with the indicated WT amino acid. **F.** Mean variant effect scores for all cysteine substitutions for the same set of proteins as in (**D**). Points indicate mean abundance or secretion score for all cysteine variants. Error bars show standard error of the mean. Asterisks indicate level of statistical significance as determined by Bonferroni-corrected pairwise two-sided t-test to FIX variant scores. * = p <0.05, **** = p <0.0001.**G.** Boxplots of secretion scores for all missense substitutions of the indicated variant amino acid across all positions.

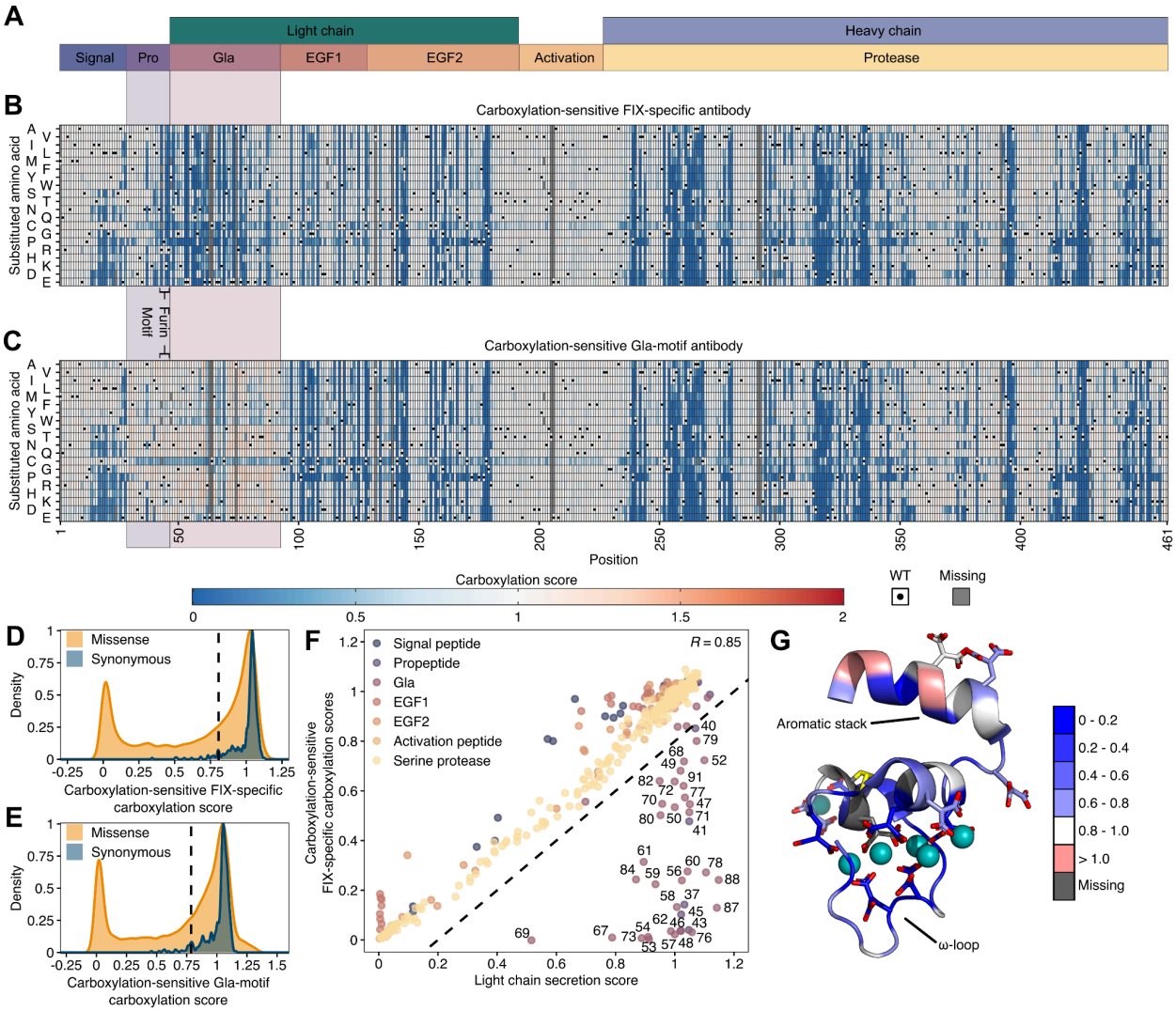


Figure 4: MultiSTEP enables measurement of variant effects on FIX post-translational modification.

A. Factor IX domain and chain architecture. Signal: Signal peptide. Pro: Propeptide. Gla: Gla domain. EGF1: Epidermal growth-like factor 1 domain. EGF2: Epidermal growth-like factor 2 domain. Activation: Activation peptide. Protease: Serine protease domain. **B-C.** Heatmaps showing carboxylation-sensitive FIX-specific carboxylation scores (**B**) or carboxylation-sensitive Gla-motif carboxylation scores (**C**) for nearly all missense FIX variants. Heatmap color indicates antibody score from 0 (blue, lowest 5% of scores) to white (1, WT) to red (increased antibody scores). Black dots indicate the WT amino acid. Missing data scores are colored gray. **D-E.** Density distributions of carboxylation-sensitive FIX-specific (**D**) or carboxylation-sensitive Gla-motif (**E**) carboxylation scores for FIX missense variants (orange) and synonymous variants (blue). Dashed line denotes the 5th percentile of the synonymous variant

distribution. **F.** Scatterplot of median MultiSTEP-derived carboxylation-sensitive FIX-specific carboxylation scores and light chain secretion scores at each position in FIX. Points are colored by domain architecture, using the same color scheme as **A**. Black dashed line indicates >0.2 point deviation threshold from perfect correlation between carboxylation and secretion scores. Points with deviation greater than this threshold are labeled with their corresponding FIX position. Pearson's correlation coefficient is shown. **G.** Crystal structure of FIX Gla domain (positions 47-92)[14]. Disulfide bridges and γ-carboxylated glutamates are shown as sticks. Calcium ions are shown as teal spheres. Structure is colored according to the ratio of median positional carboxylation-sensitive FIX-specific carboxylation score to median positional FIX light chain secretion score. Missing positions are colored gray. Disulfide bridge side chains are colored yellow.

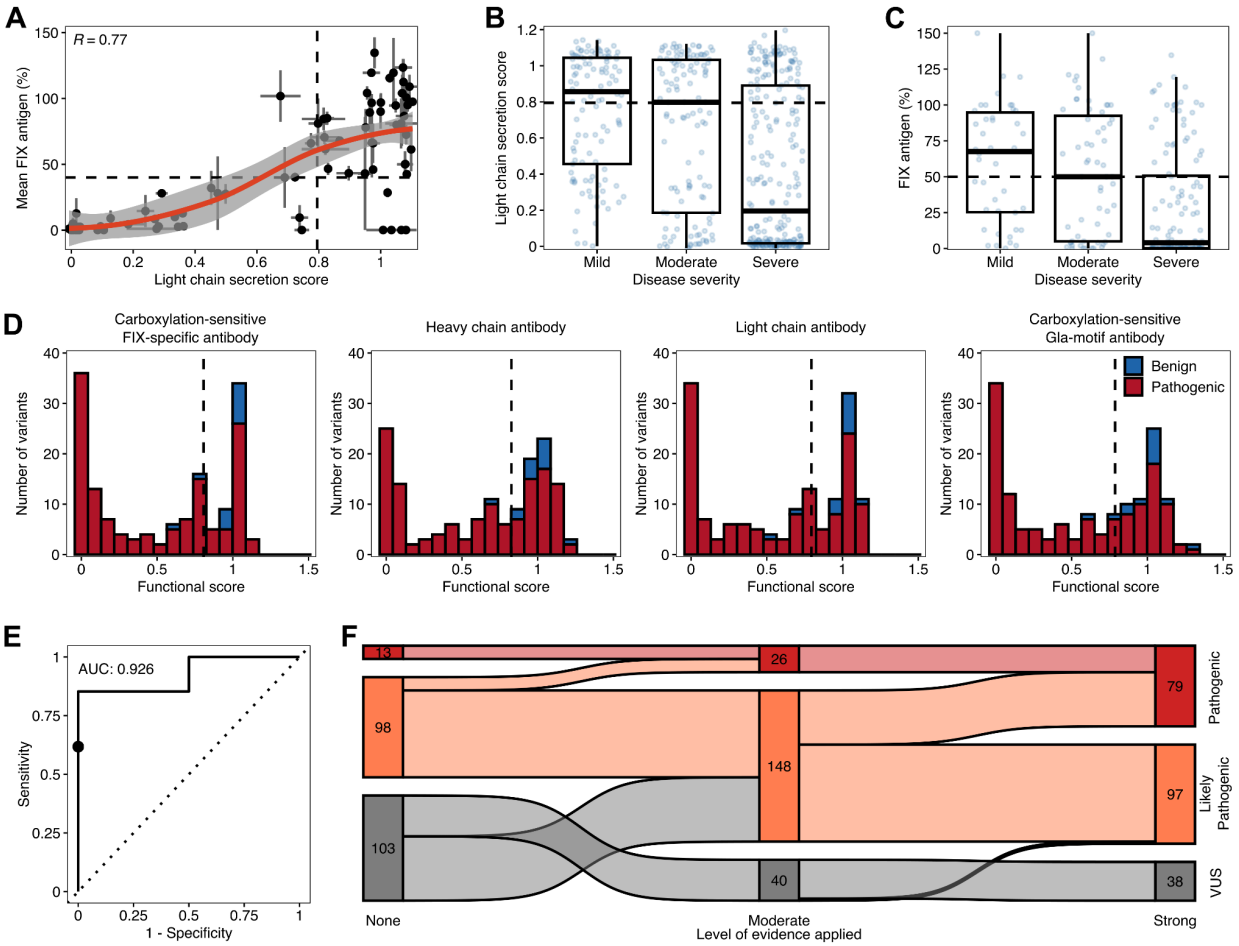


Figure 5: Secretion and gamma-carboxylation scores reveal clinical features of hemophilia B and enable variant reinterpretation.

A. Scatterplot of light chain secretion scores and FIX plasma antigen from individuals with hemophilia B in the EAHAD database[23,24]. Horizontal solid lines indicate standard error of the mean for light chain secretion scores. Vertical solid lines indicate standard error of the mean for FIX plasma antigen levels across individuals with hemophilia B harboring the same variant. Dashed horizontal line is the lower limit of WT FIX plasma antigen. Dashed vertical line is the 5th percentile of the synonymous secretion score distribution. LOESS line of best fit depicted in red with 95% confidence interval shaded in gray. **B.** Comparison of EAHAD individual hemophilia B severity with light chain secretion scores from MultiSTEP. Boxplot shows distribution of points and horizontal lines representing the 25th, 50th, and 75th percentiles.

Dashed horizontal line is the 5th percentile of the synonymous secretion score distribution. **C.** Comparison of EAHAD individual hemophilia B severity with FIX plasma antigen levels. Dashed horizontal line is the lower limit of WT FIX plasma antigen. **D.** Histograms of multiplexed functional scores for *F9* missense variants of known effect curated from ClinVar, gnomAD, and MLOF. Color indicates clinical variant interpretation. Data from four antibodies are shown. Dashed vertical line indicates the 5th percentile of synonymous variants used as a threshold for loss of function. **E.** Receiver-operator curve for random forest classifier for identifying loss of function FIX alleles from MultiSTEP scores. Dot indicates final classifier performance metrics. **F.** Sankey diagram of *F9* variant reinterpretation using moderate and strong levels of evidence for functional data. Labeled nodes represent the number of variants of each class.

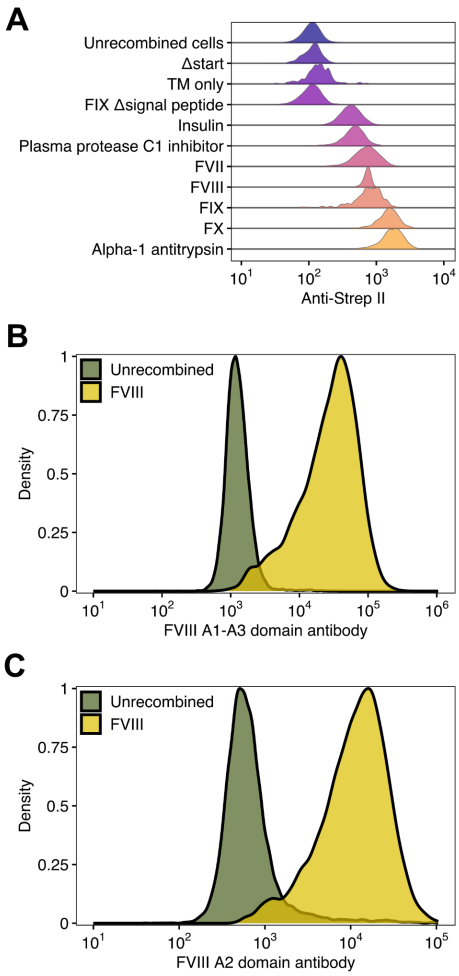
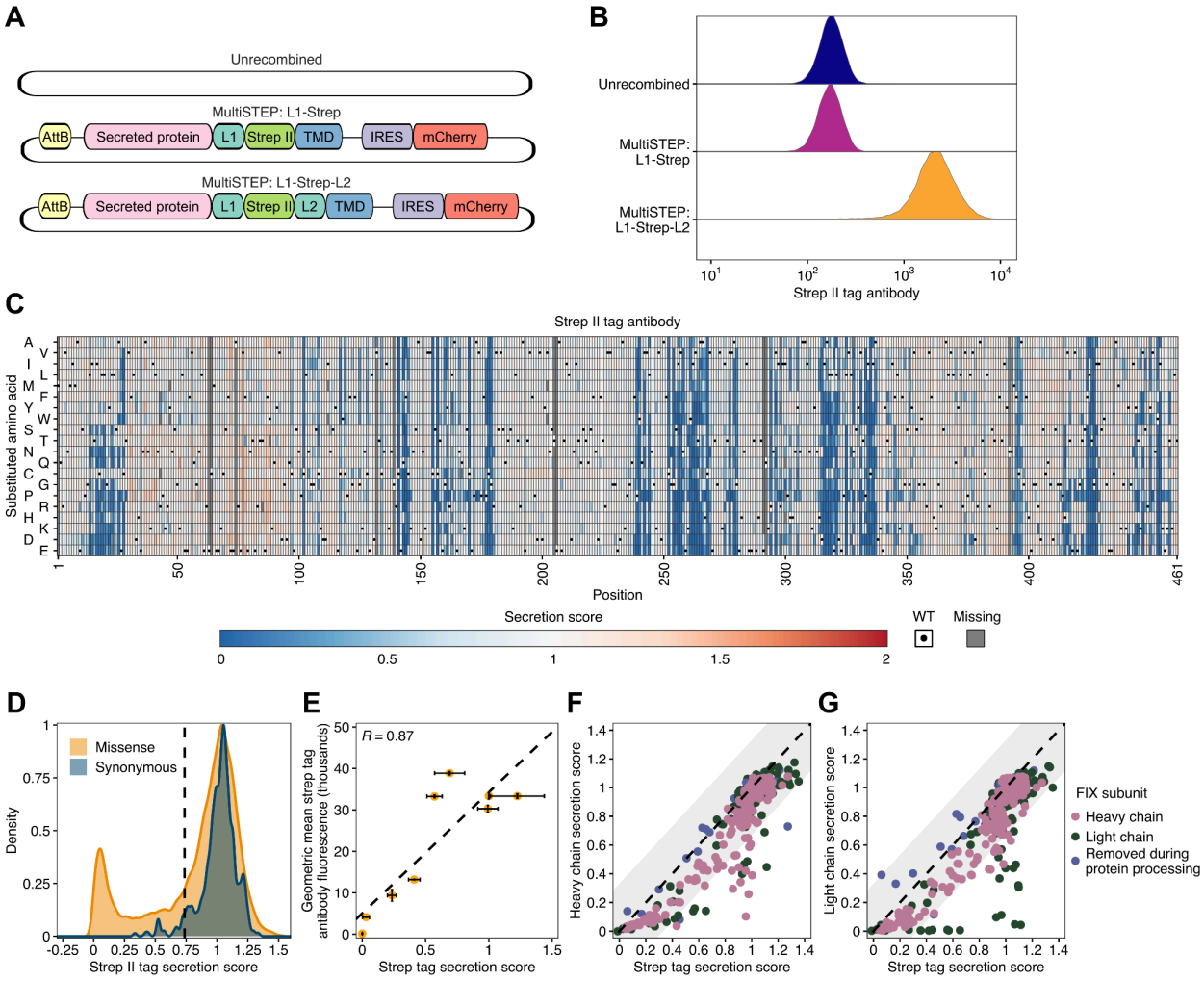


Figure 6: MultiSTEP can be applied to diverse secreted proteins.

A. Experimental flow cytometry of various genes and control constructs in the MultiSTEP backbone (n ~30,000 cells per variant). Unrecombined cells do not display FIX and serve as a negative control. All other constructs contain the MultiSTEP flexible linker, strep II tag, and transmembrane domain. Δstart does not contain a start codon. TM only does not contain a secreted protein of interest. FIX Δsignal peptide expresses a FIX molecule without its secretion-targeting signal peptide. Fluorescent signal was generated by staining the library with a rabbit polyclonal anti-strep II tag antibody followed by staining with an Alexa Fluor-488-labeled donkey anti-rabbit secondary antibody. **B.** Experimental flow cytometry of B-domain deleted coagulation factor VIII (FVIII) in the MultiSTEP backbone (n ~30,000 cells per variant). Fluorescent signal was generated by staining the library with a mouse monoclonal anti-FVIII A1-A3 antibody, which targets the discontinuous epitope at the interface of the A1 and A3 domains, followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody. **C.** Experimental flow

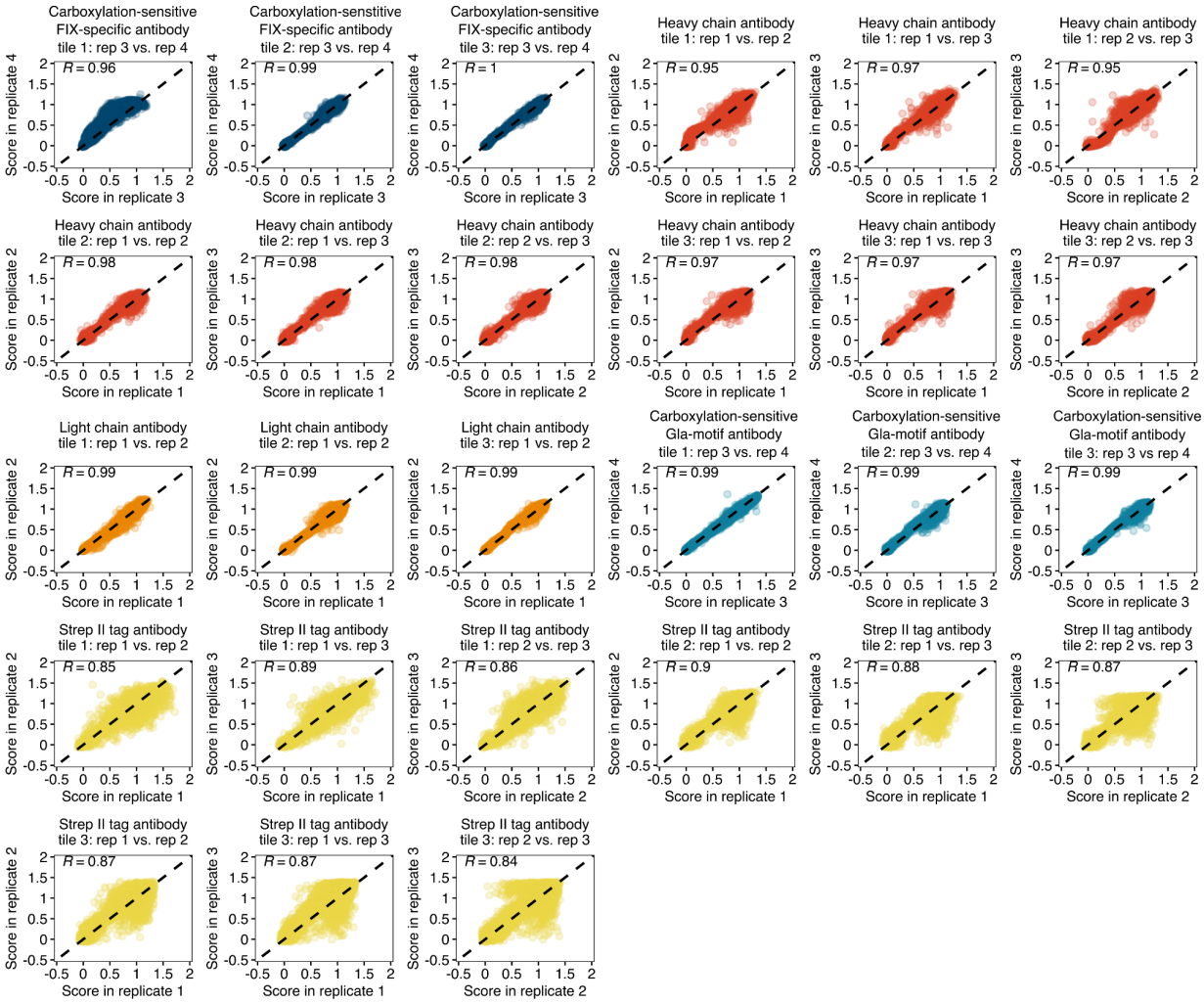
cytometry of B-domain deleted coagulation factor VIII (FVIII) in the MultiSTEP backbone (n ~30,000 cells per variant). Fluorescent signal was generated by staining the library with a mouse monoclonal anti-FVIII A2 antibody, which targets a discontinuous epitope (positions 497-510 and 584-593) within the A2 domain, followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody. Experiments with a panel of monoclonal antibodies to all FVIII domains provide further evidence that the FVIII protein is displayed correctly (**Supplementary Fig. 6**).



Supplementary Figure 1: A flexible tag-based approach to assessing variant effects on secretion.

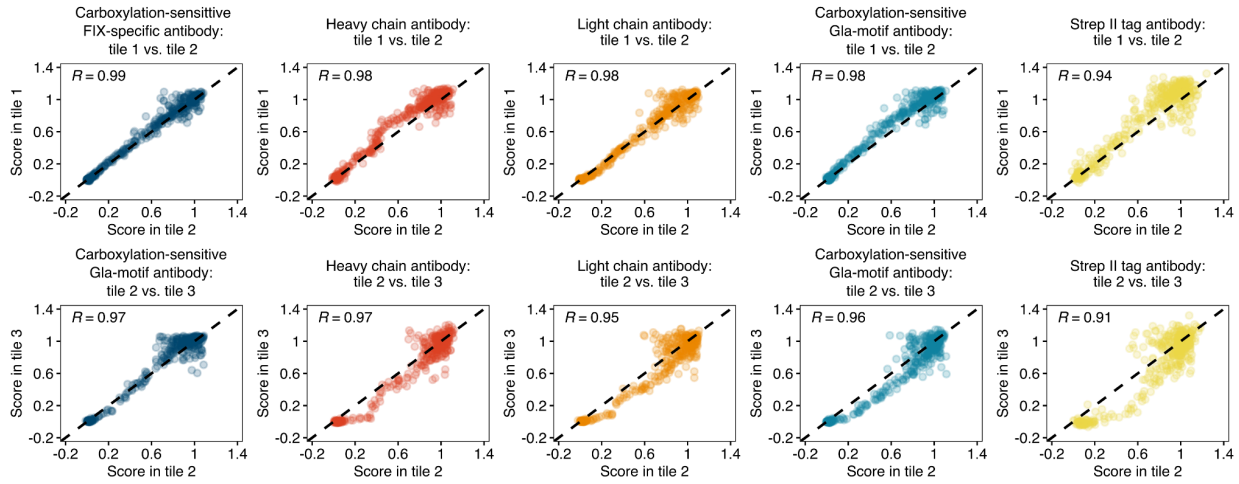
A. Various MultiSTEP donor plasmid constructs. Secreted protein coding sequences (pink) are cloned into an attB-containing landing pad donor plasmid. Secreted proteins are engineered to have C-terminally fused $(GGGGS)_2$ flexible linkers (L1 and/or L2, teal) attached to a single pass transmembrane domain (TMD, blue). In between the linkers is a strep II epitope tag for surface detection (green). The construct contains an IRES (purple) driving co-transcription of an mCherry fluorophore (red) that serves as a transcriptional control. MultiSTEP: L1-Strep does not contain an L2 linker, whereas MultiSTEP: L1-Strep-L2 does. **B.** Experimental flow cytometry of known well-secreted (A37T, S220T, WT) and poorly-secreted (C28Y) FIX variants displayed using various MultiSTEP constructs depicted in **A** (n ~30,000 cells per variant). Unrecombined cells do not display FIX and serve as a negative control.

Fluorescent signal was generated by staining the library with a mouse monoclonal anti-strep II tag antibody followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody. **C.** Heatmaps showing strep tag secretion scores for nearly all missense FIX variants. Heatmap color indicates antibody score from 0 (blue, lowest 5% of scores) to white (1, WT) to red (increased antibody scores). Black dots indicate the WT amino acid. Missing data scores are colored gray. **D.** Density distributions of strep tag secretion scores for FIX missense variants (orange) and synonymous variants (blue). Dashed line denotes the 5th percentile of the synonymous variant distribution. **E.** Scatterplot comparing MultiSTEP-derived strep tag secretion scores for seven different variants to the geometric mean of Alexa Fluor-647 fluorescence measured using flow cytometry on cells expressing each variant individually. Error bars show standard error of the mean (n = 3; >10,000 cells per replicate). Black dashed line indicates the line of best fit. Pearson's correlation coefficient is shown. **F-G.** Scatterplots of median MultiSTEP-derived strep tag secretion scores vs. heavy chain (**F**) or light chain (**G**) at each position in FIX. Points are colored by chain architecture, using the same color scheme as **Fig. 2A**. Black dashed line indicates the line of perfect correlation between secretion scores. Gray background indicates <30% deviation from perfect correlation.



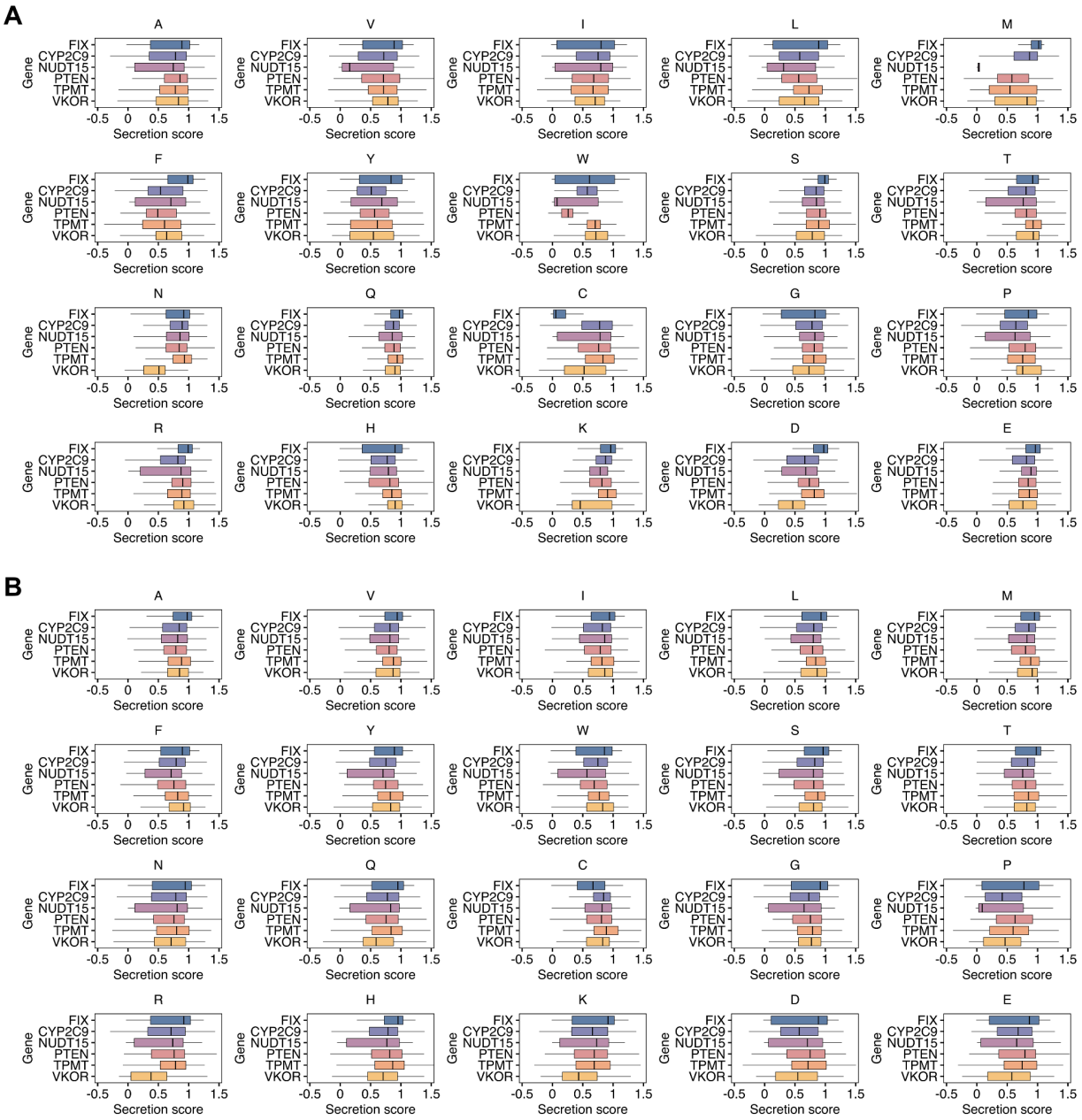
Supplementary Figure 2: Correlations between FIX sublibrary tile experimental replicates for all antibodies.

Pairwise MultiSTEP secretion and γ-carboxylation functional score correlations between replicate sorting experiments for FIX sublibrary tiles (n ≥ 3,144). Pearson's correlation coefficients are shown.



Supplementary Figure 3: Correlations between FIX variants shared across multiple sublibrary tiles for all antibodies.

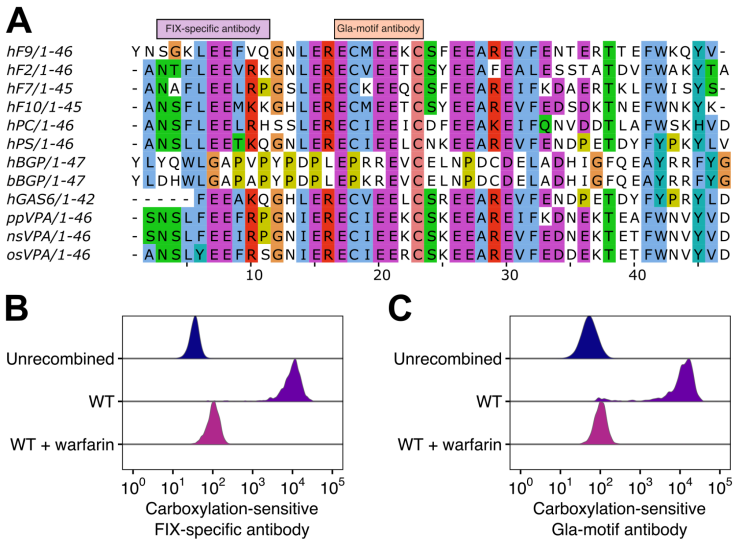
Pairwise MultiSTEP secretion and γ-carboxylation functional score correlations for variants that are shared between sublibrary tiles (top row: n = 397, bottom row: n = 380). Pearson's correlation coefficients are shown.



Supplementary Figure 4: Effect of missense FIX variation on secretion compared to missense variant effects on abundance in cytosolic or transmembrane proteins.

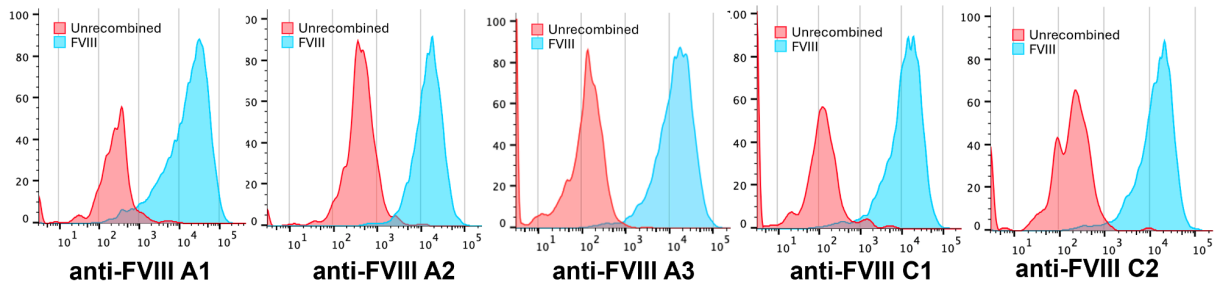
A. Boxplots of secretion or abundance scores for all nonsynonymous variants across all positions with the indicated WT amino acid for six different proteins. For FIX, the average secretion score determined by MultiSTEP is shown. For all other proteins, the average abundance score measured using VAMP-seq is shown[34,38,39,53]. **B.** Boxplots of secretion or abundance scores for all nonsynonymous substitutions of the

indicated variant amino acid across all positions for six different proteins. Average scores were determined as in (**A**).



Supplementary Figure 5: Carboxylation-sensitive antibodies bind specific epitopes.

A. Multiple sequence alignment of Gla-domain containing proteins that bind the carboxylation-sensitive pan-Gla antibody using MUSCLE[109]. Sequences were retrieved from UniProt. Antibody epitopes for both the carboxylation-sensitive FIX-specific antibody (ω-loop) and the carboxylation-sensitive Gla-motif antibody (ExxxExC) are shown. hF9: human coagulation factor IX (P00740); hF2: human prothrombin (coagulation factor II, P00734); hF7: human coagulation factor VII (P08709); hF10: human coagulation factor X (P00742); hPC: human protein C (P04070); hPS: human protein S (P07225); hBGP: human osteocalcin (bone gla-protein, P02818); bBGP: bovine osteocalcin (bone Gla-protein, P02820); hGAS6: human growth arrest-specific protein 6 (P14393); ppVPA: *Pseudechis prophyriacus* venom prothrombin activator porpharin-D (P58L93); nsVPA: *Notechis scutatis* venom prothrombin activator notecarin-D1 (P82807); osVPA: *Oxyuranus scutellatus* venom prothrombin activator oscutarin-C (P58L96). **B-C.** Flow cytometry of WT FIX with and without pretreatment with warfarin in the presence of 50 nM vitamin K (n ~30,000 cells per variant). Warfarin inhibits the activity of vitamin K epoxide reductase complex 1 (VKORC1), an enzyme essential in the steps leading to γ-carboxylation of FIX. Unrecombined cells do not display FIX and serve as a negative control. Fluorescent signal was generated by staining the library with either a mouse monoclonal carboxylation-sensitive FIX-specific antibody (**B**) or a mouse monoclonal carboxylation-sensitive Gla-motif antibody that binds the conserved motif in (**A**), followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody.



Supplementary Figure 6. Detection of cell-surface displayed FVIII. Experimental flow cytometry of B-domain deleted coagulation factor VIII (FVIII) in the MultiSTEP backbone (n = ~30,000 cells per variant). Unrecombined cells do not display FVIII and serve as a negative control. Fluorescent signal was generated by staining cells with mouse monoclonal anti-FVIII antibodies specific to each of the five FVIII domains in the heavy chain (A1 and A2) and light chain (A3, C1, and C2), followed by staining with an Alexa Fluor-647-labeled goat anti-mouse secondary antibody.

Supplementary Table 1: Variants with discordant circulating antigen and secretion scores.

Variants with discordant secretion scores and FIX plasma antigen from individuals with hemophilia B in the EAHAD database[23,24]. Variants with undetectable FIX antigen are labeled as <1, as reported by the clinical laboratory in EAHAD. Variants are classified as low antigen if they have a mean circulating FIX antigen of <40. Variants are classified as low secretion if they have a mean secretion score that is less than 0.795, which is the 5th percentile of the synonymous secretion score distribution. SE: standard error of the mean.

Supplementary Table 2: Random forest classifier functional predictions for 9,007 F9 variants.

Mean secretion scores, γ-carboxylation scores, and functional variant predictions using a random forest classifier trained on known pathogenic and benign FIX variants (see Methods). Variants without functional scores for all antibodies were removed prior to classifier implementation and do not have associated functional predictions.

Supplementary Table 3: Classification criteria and reclassification results for My Life, Our Future *F9* variants.

Classification criteria, random forest classifier predictions, and resultant pathogenicity classifications for 214 *F9* variants from *My Life, Our Future.* Classification criteria were curated and assigned by experts in hemophilia genetics based on associated clinical data and literature review. Original interpretations as well as reinterpretations using random forest classifier predictions as either moderate or strong evidence are included.

Supplementary Table 4: Oligonucleotides.

Oligonucleotides used in this study.

Supplementary Table 5: Detailed antibody information.

Properties, stock concentrations, and experimental conditions for each assayed antibody. Epitopes, when known, are provided.

Supplementary Table 6: Technical replicate correlations.

Technical replicate PCR correlations for each genomic DNA-derived barcode library amplification.

Supplementary Table 7: MultiSTEP variant scores.

Table of scores for each variant derived from each of the five antibodies presented in this work.

References

1. Landrum, M. J. *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
2. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. 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).
5. Starita, L. M. *et al.* Variant interpretation: Functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).
6. 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).
7. Scott, A. *et al.* Saturation-scale functional evidence supports clinical variant interpretation in Lynch syndrome. *Genome Biol.* **23**, 266 (2022).
8. Radford, E. J. *et al.* Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation. *Nat. Commun.* **14**, 7702 (2023).
9. Tabet, D., Parikh, V., Mali, P., Roth, F. P. & Claussnitzer, M. Scalable Functional Assays for the Interpretation of Human Genetic Variation. *Annu. Rev. Genet.* **56**, 441–465 (2022).
10. Uhlén, M. *et al.* The human secretome. *Sci. Signal.* **12**, (2019).
11. 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).

12. 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).
13. 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).
14. 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).
15. 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).
16. 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).
17. 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).
18. DiMichele, D. M. *et al.* Severe and moderate haemophilia A and B in US females. *Haemophilia* **20**, e136–43 (2014).
19. Goodeve, A. C. Hemophilia B: Molecular pathogenesis and mutation analysis. *J. Thromb. Haemost.* **13**, 1184–1195 (2015).
20. 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).
21. Konkle, B. A., Josephson, N. C. & Nakaya Fletcher, S. Hemophilia B. in *GeneReviews* (eds. Pagon, R. A. et al.) (University of Washington, Seattle, Seattle (WA), 2023).
22. Oldenburg, J. & Pavlova, A. Genetic risk factors for inhibitors to factors VIII and IX.

Haemophilia **12 Suppl 6**, 15–22 (2006).

23. 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).
24. Xu, Z., Spencer, H. J., Harris, V. A. & Perkins, S. J. An updated interactive database for 1692 genetic variants in coagulation factor IX provides detailed insights into hemophilia B. *J. Thromb. Haemost.* **21**, 1164–1176 (2023).
25. Plug, I. *et al.* Bleeding in carriers of hemophilia. *Blood* **108**, 52–56 (2006).
26. MASAC Document 273 - Recommendations on Genotyping for Persons with Hemophilia. *National Hemophilia Foundation* https://www.hemophilia.org/healthcare-professionals/guidelines-on-care/masac-documents/masac-document-273-recommendations-on-genotyping-for-persons-with-hemophilia.
27. Green, P. M. *et al.* The incidence and distribution of CpG----TpG transitions in the coagulation factor IX gene. A fresh look at CpG mutational hotspots. *Nucleic Acids Res.* **18**, 3227–3231 (1990).
28. 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).
29. 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).
30. 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).
31. Johnsen, J. M. *et al.* Novel approach to genetic analysis and results in 3000 hemophilia patients enrolled in the My Life, Our Future initiative. *Blood Advances* **1**, 824–834 (2017).
32. 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).

33. 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).
34. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).
35. Liu, L. *et al.* Inclusion of Strep-tag II in design of antigen receptors for T-cell immunotherapy. *Nat. Biotechnol.* **34**, 430–434 (2016).
36. Bicocchi, M. P. *et al.* Insight into molecular changes of the FIX protein in a series of Italian patients with haemophilia B. *Haemophilia* **12**, 263–270 (2006).
37. Pezeshkpoor, B. *et al.* Variants in FIX propeptide associated with vitamin K antagonist hypersensitivity: functional analysis and additional data confirming the common founder mutations. *Ann. Hematol.* **97**, 1061–1069 (2018).
38. 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).
39. Amorosi, C. J. *et al.* Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am. J. Hum. Genet.* **108**, 1735–1751 (2021).
40. Kurys, G., Tagaya, Y., Bamford, R., Hanover, J. A. & Waldmann, T. A. The long signal peptide isoform and its alternative processing direct the intracellular trafficking of interleukin-15. *J. Biol. Chem.* **275**, 30653–30659 (2000).
41. 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).
42. Tikhonova, E. B., Karamysheva, Z. N., von Heijne, G. & Karamyshev, A. L. Silencing of Aberrant Secretory Protein Expression by Disease-Associated Mutations. *J. Mol. Biol.* **431**, 2567–2580 (2019).
43. Luirink, J. & Sinning, I. SRP-mediated protein targeting: structure and function revisited.

Biochim. Biophys. Acta **1694**, 17–35 (2004).

44. Halic, M. & Beckmann, R. The signal recognition particle and its interactions during protein targeting. *Curr. Opin. Struct. Biol.* **15**, 116–125 (2005).
45. Liaci, A. M. *et al.* Structure of the human signal peptidase complex reveals the determinants for signal peptide cleavage. *Mol. Cell* **81**, 3934–3948.e11 (2021).
46. von Heijne, G. How signal sequences maintain cleavage specificity. *J. Mol. Biol.* **173**, 243–251 (1984).
47. Nilsson, I., Whitley, P. & von Heijne, G. The COOH-terminal ends of internal signal and signal-anchor sequences are positioned differently in the ER translocase. *J. Cell Biol.* **126**, 1127–1132 (1994).
48. Braakman, I. & Hebert, D. N. Protein folding in the endoplasmic reticulum. *Cold Spring Harb. Perspect. Biol.* **5**, a013201 (2013).
49. Zhang, H. *et al.* Unpaired Extracellular Cysteine Mutations of CSF3R Mediate Gain or Loss of Function. *Cancer Res.* **77**, 4258–4267 (2017).
50. Robinson, P. J. & Bulleid, N. J. Mechanisms of Disulfide Bond Formation in Nascent Polypeptides Entering the Secretory Pathway. *Cells* **9**, (2020).
51. 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).
52. Di Scipio, R. G., Kurachi, K. & Davie, E. W. Activation of human Factor IX (Christmas Factor). *J. Clin. Invest.* **61**, 1528–1538 (1978).
53. Chiasson, M. A. *et al.* Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *Elife* **9**, (2020).
54. 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).

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

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

57. 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).

58. 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).

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

60. 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).

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

62. Mayhew, M., Handford, P. & Brownlee, G. G. The binding of natural variants of human factor IX to endothelial cells. *FEBS Lett.* **341**, 74–78 (1994).

63. Wojcik, E. G. *et al.* Factor IX Zutphen: a Cys18-->Arg mutation results in formation of a heterodimer with alpha 1-microglobulin and the inability to form a calcium-induced conformation. *Biochem. J* **311 (Pt 3)**, 753–759 (1995).

64. 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).

65. 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).

66. 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).

67. Price, P. A., Fraser, J. D. & Metz-Virca, G. Molecular cloning of matrix Gla protein: implications for substrate recognition by the vitamin K-dependent gamma-carboxylase. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 8335–8339 (1987).

68. 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).

69. Whitlon, D. S., Sadowski, J. A. & Suttie, J. W. Mechanism of coumarin action: significance of vitamin K epoxide reductase inhibition. *Biochemistry* **17**, 1371–1377 (1978).

70. 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).

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

72. 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).

73. Belvini, D. *et al.* Molecular genotyping of the Italian cohort of patients with hemophilia B. *Haematologica* **90**, 635–642 (2005).

74. 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).

75. 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).
76. 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).
77. Liebman, H. A. The metal-dependent conformational changes in factor IX associated with phospholipid binding. Studies using antibodies against a synthetic peptide and chemical modification of factor IX. *Eur. J. Biochem.* **212**, 339–345 (1993).
78. Jacobs, M., Freedman, S. J., Furie, B. C. & Furie, B. Membrane binding properties of the factor IX gamma-carboxyglutamic acid-rich domain prepared by chemical synthesis. *J. Biol. Chem.* **269**, 25494–25501 (1994).
79. Sekiya, F., Yoshida, M., Yamashita, T. & Morita, T. Localization of the specific binding site for magnesium(II) ions in factor IX. *FEBS Lett.* **392**, 205–208 (1996).
80. Agah, S. & Bajaj, S. P. Role of magnesium in factor XIa catalyzed activation of factor IX: calcium binding to factor IX under physiologic magnesium. *J. Thromb. Haemost.* **7**, 1426–1428 (2009).
81. Vadivel, K., Schmidt, A. E., Marder, V. J., Krishnaswamy, S. & Bajaj, S. P. Structure and function of Vitamin K-Dependent Coagulant and Anticoagulant Proteins. in *Hemostasis and Thrombosis : Basic Principles and Clinical Practice* (eds. Marder, V. J., Aird, W. C., Bennett, J. S., Schulman, S. & Gilbert C. White, I. I.) 208–232 (Wolters Kluwer Health, Philadelphia, UNITED STATES, 2012).
82. Rohlena, J., Kolkman, J. A., Boertjes, R. C., Mertens, K. & Lenting, P. J. Residues Phe342-Asn346 of activated coagulation factor IX contribute to the interaction with low

density lipoprotein receptor-related protein. *J. Biol. Chem.* **278**, 9394–9401 (2003).

83. Westmark, P. R., Tanratana, P. & Sheehan, J. P. Selective disruption of heparin and antithrombin-mediated regulation of human factor IX. *J. Thromb. Haemost.* **13**, 1053–1063 (2015).
84. Plautz, W. E. *et al.* Anticoagulant Protein S Targets the Factor IXa Heparin-Binding Exosite to Prevent Thrombosis. *Arterioscler. Thromb. Vasc. Biol.* **38**, 816–828 (2018).
85. Cooley, B. *et al.* Dysfunctional endogenous FIX impairs prophylaxis in a mouse hemophilia B model. *Blood* **133**, 2445–2451 (2019).
86. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054–1060 (2018).
87. 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).
88. Casademunt, E. *et al.* The first recombinant human coagulation factor VIII of human origin: human cell line and manufacturing characteristics. *Eur. J. Haematol.* **89**, 165–176 (2012).
89. Vatandoost, J. & Dolatabadi, B. Stable and transient expression of human coagulation factor IX in mammalian expression systems; CHO versus HEK cells. *Gene Cell Tissue* **4**, (2017).
90. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
91. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17**, 467–473 (2007).
92. 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).
93. Salema, V. & Fernández, L. Á. Escherichia coli surface display for the selection of nanobodies. *Microb. Biotechnol.* **10**, 1468–1484 (2017).

94. Ho, M. & Pastan, I. Mammalian cell display for antibody engineering. in *Therapeutic Antibodies: Methods and Protocols* (ed. Dimitrov, A. S.) 337–352 (Humana Press, Totowa, NJ, 2009).

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

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

97. Szabady, R. L., Peterson, J. H., Skillman, K. M. & Bernstein, H. D. An unusual signal peptide facilitates late steps in the biogenesis of a bacterial autotransporter. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 221–226 (2005).

98. Mason, D. M. *et al.* High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. *Nucleic Acids Res.* **46**, 7436–7449 (2018).

99. Parthiban, K. *et al.* A comprehensive search of functional sequence space using large mammalian display libraries created by gene editing. *MAbs* **11**, 884–898 (2019).

100. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

101. 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).

102. Miao, H. Z. *et al.* Bioengineering of coagulation factor VIII for improved secretion. *Blood* **103**, 3412–3419 (2004).

103. Kessler, C. M. *et al.* B-domain deleted recombinant factor VIII preparations are bioequivalent to a monoclonal antibody purified plasma-derived factor VIII concentrate: a randomized, three-way crossover study. *Haemophilia* **11**, 84–91 (2005).

104. Ward, N. J. *et al.* Codon optimization of human factor VIII cDNAs leads to high-level expression. *Blood* **117**, 798–807 (2011).

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

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

107. 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).

108. 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).

109. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

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

111. 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).

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

113. 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).

114. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).

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

116. Baron, M. *et al.* The three-dimensional structure of the first EGF-like module of human factor IX: comparison with EGF and TGF-alpha. *Protein Sci.* **1**, 81–90 (1992).

117. Johnson, D. J. D., Langdown, J. & Huntington, J. A. Molecular basis of factor IXa recognition by heparin-activated antithrombin revealed by a 1.7-Å structure of the ternary complex. *Proceedings of the National Academy of Sciences* **107**, 645–650 (2010).