Interpreting variation in pharmacogenes using multiplex assays

Melissa A. Chiasson

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Douglas Fowler, Chair

Deborah Nickerson

Allan Rettie

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Interpreting variation in pharmacogenes using multiplex assays

Melissa A. Chiasson

Chair of the Supervisory Committee:

Douglas Fowler

Department of Genome Sciences

With the advent of genome sequencing technologies, our ability to read DNA sequences is unprecedented. However, understanding how the variation we encounter impacts humans is a formidable challenge. Technologies like multiplex assays, in which we can measure tens of thousands of variants in a high-throughput, pooled manner, allow us to tackle this problem at scale. In particular, genes that are involved in drug response, called pharmacogenes, pose a unique opportunity to apply multiplex assays. These assays would not only help us better understand pharmacogene biology, they would also provide vital evidence for tailoring drug regimens to a patient's genotype. To advance towards this goal, in Chapter 1, I give an introduction to multiplex assays and outline what pharmacogenes we should prioritize for this approach. In Chapter 2, I detail how I used multiplex assays to interrogate the biology of vitamin K epoxide reductase (VKOR), the target of warfarin. By applying assays for abundance and

activity to VKOR, I suggest a resolution for the controversy over VKOR topology, identify residues that are functionally constrained in the protein, and interpret human variants found in genomic databases. In Chapter 3, I show how applying an abundance assay to cytochrome P450 2C9 (CYP2C9) identifies highly conserved core regions of the protein and the show that 42% of missense variants present in genome databases have decreased abundance, suggesting that drug metabolism may be affected in individuals with these variants. Finally, in chapter 4, I outline challenges in designing multiplex assays for pharmacogenes, including tackling combinatorial variation, and comment on the promising future of pharmacogenomics.

Table of Contents

List of Figures and Tables

Acknowledgements

My Ph.D. ended strangely, as the coronavirus pandemic struck, and for that reason I am so grateful to have had such a strong support network. First, I want to thank Doug Fowler for building a lab where so many smart, wonderful people get to interact every day. Doug's mentorship taught me to fight for my ideas and be creative in my approaches to solving questions. Five-minute idea lab meetings will always have a place in my heart. My committee was always helpful in suggesting new avenues to pursue with my data. In particular I want to thank Allan Rettie for his VKOR and CYP2C9 acumen, and Debbie Nickerson for her unfailing enthusiasm and deep knowledge of the pharmacogenomics field. Thank you to all of the Genome Sciences support staff who make the department run smoothly: Brian Giebel, Beth Hammermeister, Sandra Pennington, and the facilities, custodial, and security staff.

Kenny Matreyek was such a bright spot of my Ph.D. I loved talking about science with him, in that he would be honest about what he knew and did not buy into hype around sexy science. His guidance transformed me from someone who would spastically do too many experiments in a fit of anxiety to someone who could fail, take a deep breath, and think, what is the most effective experiment to design next? This skill set is invaluable, and for Kenny's insight and friendship I will forever be grateful.

Molly Gasperini was the ultimate conference buddy in grad school and is one of the funniest, smartest people I know. Andrew Hill was also a steadfast friend who sent me Stella photos when I needed it most. Charlie Lee was a source of calm throughout my PhD and often would accept bad Pokemon Go trades from me. I also want to thank Greg Findlay for being my favorite dancer (don't tell Molly) and Max Dougherty for his thoughtfulness and his love of the SeaTac Qdoba.

And then there were the non-science friends. Matt Pantoja and Kirsten Cooper are my family in Seattle, making arepas for me and commenting on the dynamics of the Bon Appetit Test Kitchen. My crew of friends in Austin, who always provided me with a lovely escape from the Seattle grey: Eric Oeur, Matt Vitemb, Sarah Rutledge, and Zack Shlachter. My college roommates, who are all extremely successful, funny, kind people who lift my spirits: Liya Assefa, Elizabeth Kim, Julia Knight, and Stacey Khoury-Diaz. Anusha Alles, for her friendship and her fight for a more equitable, just world.

Finally, I want to thank my family for their unwavering support. My mom, whose resilience taught me to always get back up again, even when you want to quit. My dad, who would take me to Putt Putt Golf as a kid for a $5 Saturday deal where you could play mini-golf and then eat a foot-long hot dog. To this and this alone do I attribute my success. My brother, Christopher, and his wife, Anita, who were my biggest cheerleaders. And to my dog, Hubert, who was my constant companion through my Ph.D.

Dedication

Dedicated to Barry Piekos, whose electron microscopy course at Yale showed me how captivating science can be.

1 Introduction

Chapter 1 has been adapted with modifications from:

Chiasson, Melissa, Maitreya J. Dunham, Allan E. Rettie, and Douglas M. Fowler. 2019. "Applying Multiplex Assays to Interpret Variation in Pharmacogenes." *Clinical Pharmacology and Therapeutics* 106 (2): 290-294.

The field of pharmacogenetics studies the genetic determinants of drug response, the roots of which reach far back in time. In ancient Greece, Pythagoras described a fatal reaction to ingesting broad beans, later shown to be hemolytic anemia caused by deficiency of the enzyme glucose-6-phosphate dehydrogenase (G6PD)[1]. The study of hemolytic anemias was indeed key to the establishment of the field in the 1950s and 60s: Arno Motulsky, a pioneer of pharmacogenetics, demonstrated that a drug-induced hemolysis common in African-American men was due to inherited deficiency of G6PD[2–4]. Additional family studies in the 1960s and 1970s showed inheritance patterns for drug response. In 1968, Elliot Vesell and John Page observed that there were pharmacokinetic differences between monozygotic and dizygotic twins treated with the drug antipyrine[5]. In the 1970s, Michel Eichelbaum and Geoff Tucker posited that individuals exhibited heterogeneity in metabolism of debrisoquine[6] and sparteine[7] due to genetic polymorphism. By the 1980s, progress had been made in identifying and cloning genes responsible for differential drug response, known as pharmacogenes. This included the gene for cytochrome P450 2D6 (CYP2D6)[8], which metabolizes debrisoquine and sparteine.

Variants in pharmacogenes can affect the pharmacokinetics of a drug, which describes how a drug moves through the body, or the pharmacodynamics of a drug, which describes a drug's effect and mechanism of action. The eventual confluence of population-based studies with basic molecular biology of pharmacogenes enabled clinicians to make better informed drug choices for patients. For example, in HIV clinics it was observed that patients with the MHC class I allele HLA-B*5701 often had a life-threatening hypersensitivity reaction to the antiviral abacavir[9,10].

Clinics then began sequencing patients at this locus, and those who had HLA-B*5701 were given an alternative antiviral treatment[11]. Incidence of this hypersensitivity syndrome decreased as a result[12].

Significant progress has been made in cataloguing common variants that impact drug response. For example, many patients are prescribed warfarin, an anticoagulant with a narrow therapeutic window. Common single nucleotide variants (SNVs) in the genes *VKORC1* and *CYP2C9* (Cytochrome P450 Family 2 Subfamily C Member 9) identify haplotypes that are correlated with warfarin dose variability in the normal dose range (2 to 10 mg/day) [13]. Therefore, it was hypothesized that dosing patients based on their haplotypes may improve clinical outcomes. Indeed, prospective genotype-guided dosing in homogeneous populations leads to better clinical outcomes compared to a group that is dosed according to a clinical algorithm[14].

Despite the progress in showing the effect of common variation on warfarin dose, patients that required significantly higher warfarin doses (> 10 mg/day) were not explained by these common haplotypes. These patients were found to carry rare variants in the coding region of *VKORC1*, showing how effect sizes of these rare variants can be significant[15,16]. While some of these variants are found at positions similar to what we observed in warfarin-resistant rats[17–19], other variants are novel, and the mechanism of resistance is unclear. These variants are rare, found only in one or a few people, and the degree of warfarin resistance among these variants is also variable[20].

Today, genome sequencing has enabled the detection of unprecedented numbers of new human variants, including in pharmacogenes[21]. But, interpreting how these variants affect pharmacogene biology and ultimately drug response is difficult. The Clinical Pharmacogenetics Implementation Consortium (CPIC, https://cpicpgx.org/) was established in 2009 to address the

need for a standardized pipeline for integration of pharmacogenetic data into the clinic[22]. While CPIC is a crucial source of pharmacogenetic expertise, its guidelines focus on common variants, and the experimental data it relies on, like Western blots, cannot be scaled up easily to address the onslaught of new variants. In contrast, multiplexed assays for variant effects (MAVEs) leverage high throughput DNA sequencing to assess the functional consequences of thousands of variants simultaneously[23]. We discuss the utility of large-scale functional data in pharmacogene variant interpretation and suggest that implementing MAVEs could empower pharmacogenetics and improve patient care.

Genomes can now be sequenced with ease, but understanding the effect of the variants found therein poses a major challenge[24]. Each uninterpreted variant represents a missed opportunity to improve patient outcomes. For example, CPIC lists 358 gene-drug pairs where variation can change drug response. For 63 of these 358 pairs, CPIC has issued guidelines regarding clinical interventions that may improve patient care[25]. These guidelines focus on common variants (minor allele frequencies, MAF, typically >5%) whose clinical consequences are most clearly documented. However, understanding the effects of rare variants (MAF < 0.5-1%) is also essential, and this goal is far from realized.

The magnitude of the unmet need requires consideration of the totality of rare variation that will be identified as sequencing becomes more common. As of April 2020, the Genome Aggregation Database (gnomAD v2 and v3, https://gnomad.broadinstitute.org/)[26] contained ~125,000 exomes and ~72,000 genomes, which included 893 coding single nucleotide variants in *CYP2C9* alone, 479 of which were singletons. 55 of these variants were present in the CPIC database (accessed 4/20/20), but only 32 (58%) have been functionally annotated. Undoubtedly, as sequencing continues, many more *CYP2C9* variants will be identified. This issue is not confined

to *CYP2C9*: 731 novel non-synonymous variants in 12 CYP genes were discovered in the exomes of ~6,500 individuals[21]. ~10% of individuals carried at least one of these potentially deleterious novel variants. These results, obtained from a handful of genes in a few individuals relative to the number that will ultimately be sequenced, illustrate that an onslaught of new and potentially important variants is coming.

1.1 The challenge of variant functional analysis

Current methods for determining the impact of pharmacogene variants fall into two categories. Biochemical assays using known substrates for drug disposition genes can reveal variant functional consequences[27,28]. However, this approach is limited in scale to tens or hundreds of missense variants. For example, the traditional VKOR activity assay measures production of vitamin K epoxide using whole cell extracts treated with dithiothreitol (DTT)[15,29]. Questions remain as to the accuracy of this activity measurement, especially under warfarin treatment[30,31], but more importantly, each study using this assay only measures five to 20 variants at a time[15,32–34].

Another approach, computational predictions, can scale to all possible variants of a gene of interest, but are of limited value as they often produce incorrect or conflicting results[35]. For example, the *CYP2C9*3* variant, present in ~7% of Caucasians, confers ~90% loss of function according to experimental data[36], but is predicted computationally to be benign[37]. To overcome the limitations of biochemical assays and computational predictions, an experimental approach to assess pharmacogene variants on a massive scale is needed.

1.2 MAVEs can characterize tens of thousands of variants simultaneously

A multiplex assay for variant effects (MAVE) measures the functional consequences of a large library of genetic variants simultaneously[24,38]. MAVEs can be applied to a wide range of genetic elements including mRNA UTRs, promoters, enhancers, splice sites, and proteins[39–45]. The result of a MAVE is a variant effect map that reveals the functional consequences of all possible single variants in the genetic element.

All MAVEs share the same basic design (Figure 1.1A). First, a pooled library of variants is constructed either by PCR-based mutagenesis or synthesized oligo arrays programmed with mutations of interest[46,47]. The library is then introduced into an experimental system, typically yeast or cultured human cells. Each cell must express a single variant to maintain the link between variant sequence and phenotype. For example, in human cells, expression of a single variant is typically achieved using lentiviral transduction or recombinase-based systems[48–50]. Cells expressing the library of interest are then assayed for a phenotype of interest, like growth or reporter activation. These assays stratify variants based on their phenotypic effect. For example, in a growth assay, cells expressing wild type-like variants grow rapidly whereas cells expressing loss-of-function variants grow slowly[51,52]. In a fluorescent reporter assay, wild type-like variants drive high fluorescence whereas loss-of-function variants drive low fluorescence[53]. Cells are sorted into bins according to fluorescence. High throughput sequencing is used to measure a variant's frequency in the assay, either before and after growth or across bins. Variant frequencies are then used to compute effect scores.

MAVEs for coding and noncoding variants differ in the type of assays used. For example, noncoding MAVEs generally measure how variants affect expression, often by quantifying mRNA transcripts or using a fluorescent reporter[40,54]. Coding MAVEs measure different aspects of a

protein's function. For example, reporter assays can measure specific protein properties like abundance or substrate binding using fluorescent protein tags[53] or fluorophore-labeled antibodies[55]. Growth-based assays measure each variant's ability to drive cell growth, either in the context of a deletion of the genomic copy of the protein[50] or by using a metabolic reporter[56].

MAVEs have the power to functionally annotate variants in many, if not most, pharmacogenes. However, achieving this goal will take time and effort, requiring the implementation of existing MAVEs and the development of new assays. To illustrate these issues, we first discuss the recent application of a MAVE to thiopurine methyltransferase (TPMT) and then consider other pharmacogenes that could benefit most from MAVEs.

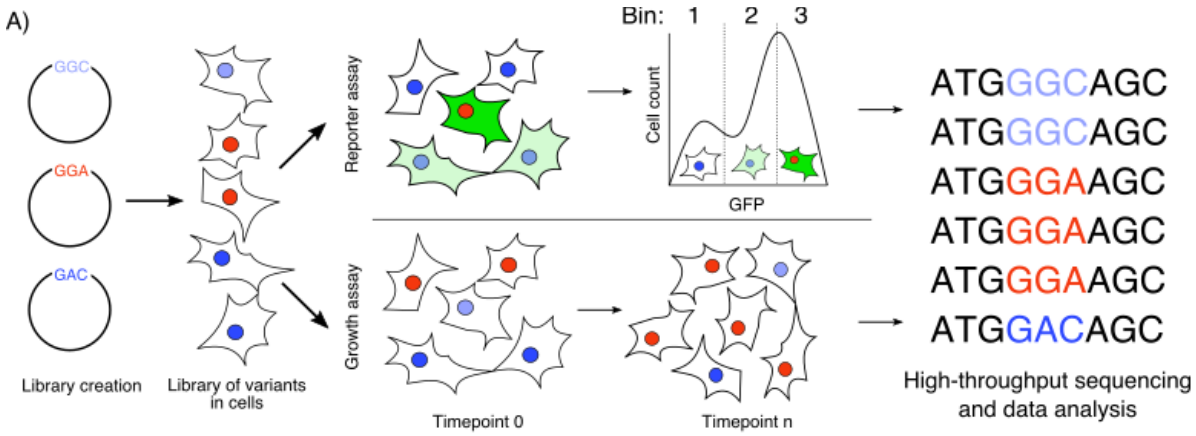
1.3 Analyzing TPMT abundance reveals new variants that confer thiopurine toxicity risk

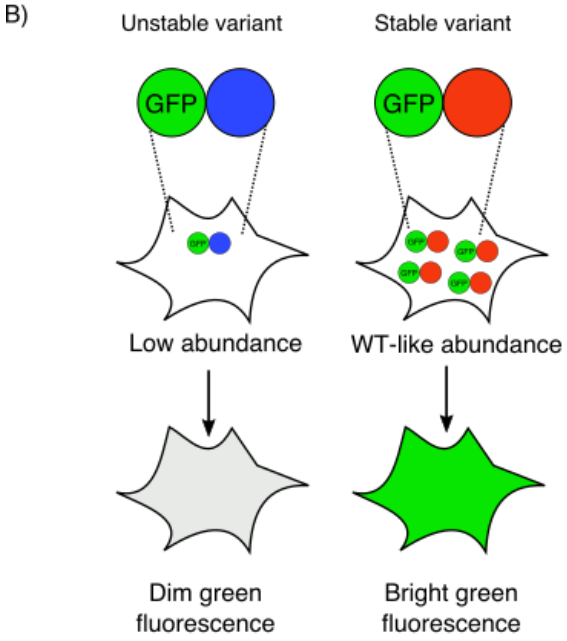
TPMT inactivates thiopurine drugs commonly used to treat cancer and autoimmune diseases, including 6-thioguanine and 6-mercaptopurine (6-MP). Thus, TPMT reduces the quantity of drug available for transformation into thioguanine nucleotides, which inhibit *de novo* purine synthesis. During routine dosing with thiopurines, TPMT deficiency results in high levels of thioguanine nucleotides and, ultimately, hematopoietic toxicity. Three variants, A80P, A154T, and Y240C, are known to lead to decreased TPMT function[57]. CPIC recommends testing for these three variants, enabling patients to be classified as normal, intermediate, or poor metabolizers based on diplotype, with doses adjusted accordingly.

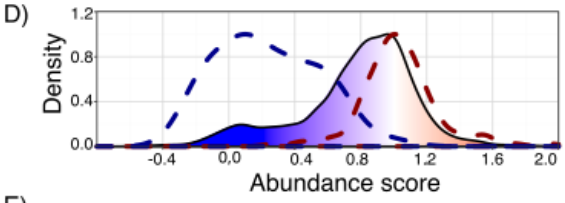
Previously, we applied Variant Abundance by Massively Parallel sequencing (VAMP-seq), a generalizable, multiplex assay for measuring protein abundance inside cells, to TPMT[53] (Figure 1.1B). We generated abundance scores for 3,689 of the 4,655 possible variants (Figure 1.1C, D, and E). A80P, A154T, and Y240C were all low abundance variants, in accordance with their poor metabolizer status. In contrast, four rare variants from a clinical study of acute

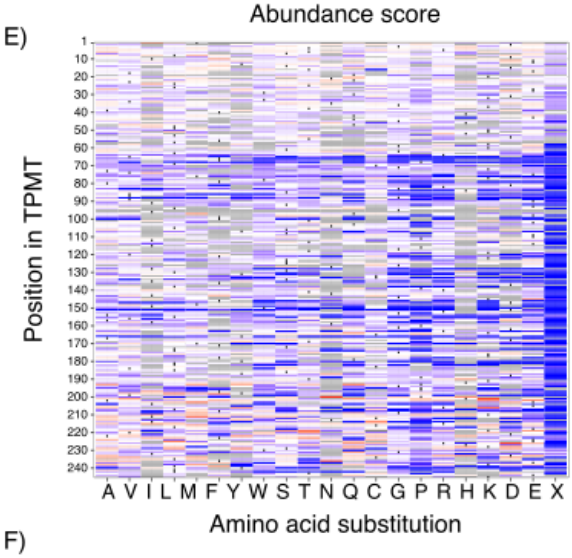
lymphoblastic leukemia (S125L, Q179H, R215H, R226Q)[58] were all wild type (WT)-like in abundance, and patients with these variants tolerated higher doses of 6-MP better than those with A80P, A154T, or Y240C. We then identified 31 reduced abundance variants in gnomAD, and suggested that patients with these variants could have increased risk for thiopurine toxicity. Since our publication of the TPMT variant abundance map, seven new TPMT variants have been added to gnomAD: K77E, W78R, G83V, L155S, P160A, K191E, and C216Y. VAMP-seq data indicate that K77E, W78R, G83V, L155S, and K191E are of low abundance relative to WT (Figure 1.1F). Accordingly, these variants might confer drug sensitivity in patients that carry them.

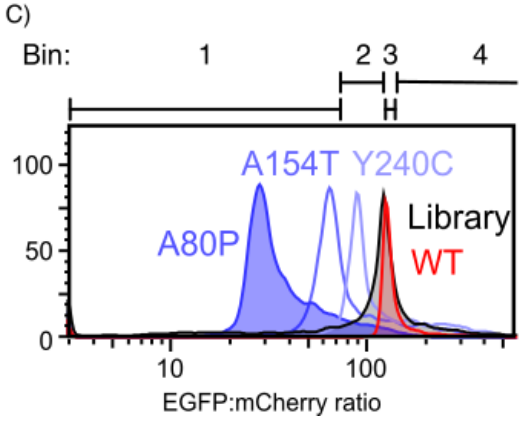
Thus, protein abundance is a useful phenotype for identifying loss-of-function variants. We also anticipate that measurement of protein activity will be necessary for many pharmacogenes. Fortunately, in some cases, existing low-throughput activity assays can be adapted. For example, a reporter cell line developed to measure vitamin K oxidoreductase (VKOR) activity[59] could be combined with a variant library to assess activity of all VKOR missense variants. Since some VKOR variants confer resistance to warfarin, cells could also be treated with warfarin to reveal the relationship between activity and resistance. Ultimately, the activity and resistance scores from such an assay could be used to help predict a patient's warfarin dose based on their *VKORC1* sequence.


A)
GGC
GGA
GAC
Library creation
Library of variants in cells
Reporter assay
Growth assay
Timepoint 0
Timepoint n
Bin: 1 2 3
Cell count
GFP
ATGGGCAGC
ATGGGCAGC
ATGGGAAGC
ATGGGAAGC
ATGGGAAGC
ATGGACAGC
High-throughput sequencing and data analysis


B)
Unstable variant
Stable variant
GFP
GFP
Low abundance
WT-like abundance
Dim green fluorescence
Bright green fluorescence


D)
Density
1.2
0.8
0.4
0.0
-0.4
0.0
0.4
0.8
1.2
1.6
2.0
Abundance score


E)
Position in TPMT
A V I L M F Y W S T N Q C G P R H K D E X
Amino acid substitution


C)
Bin: 1 2 3 4
100
50
0
A154T Y240C
A80P
Library
WT
10
100
EGFP:mCherry ratio

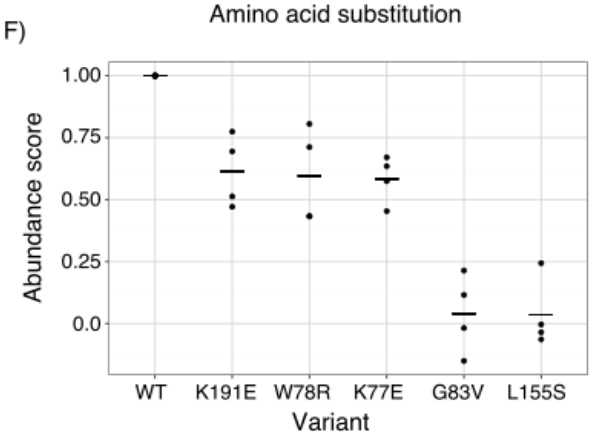

F)
Abundance score
1.00
0.75
0.50
0.25
0.0
WT
K191E
W78R
K77E
G83V
L155S
Variant

Figure 1.1. Overview of MAVEs.

A) A library of variants of the genetic element of interest is created and introduced into cells. The cells are subjected to a growth- or fluorescent reporter-based assay. High-throughput sequencing is used to determine the frequency of variants before and after the assay, and variant frequencies are used to calculate functional scores. B) VAMP-seq uses a GFP fusion reporter to measure steady-state variant abundance. GFP was fused N-terminally to a library of TPMT variants; mCherry was used as a transcriptional control. This library was introduced into HEK293T cells using a serine integrase landing pad system such that only one variant is expressed per cell[48]. Cells were sorted based on their fluorescence into four bins. High-throughput sequencing was used to determine the frequency of every variant in each bin. Frequencies were then converted to abundance scores. C) A FACS plot of WT TPMT (red) and three high-frequency variants known to be low abundance (blue): A80P, A154T, and Y240C. The library of TPMT variants and bins used for sorting are shown (gray). D) Density plot of abundance scores, with dotted blue line showing distribution of nonsense variants and red dotted line showing synonymous variants. The missense variant distribution is shaded from blue (low abundance) to red (high abundance). E) Heatmap of TPMT abundance scores shaded from blue (low abundance) to red (high abundance); gray indicates missing data. F) Abundance scores from four replicates for six new TPMT variants found in gnomAD.

1.4 MAVEs could aid pharmacogene variant interpretation

Including *TPMT*, CPIC lists 127 genes that have differing levels of evidence for identification as an actionable pharmacogene. 5,132,280 possible single nucleotide variants exist amongst these genes. Assaying such a large number of variants is possible, but daunting. Thus, we suggest prioritization of the most promising pharmacogenes.

We focused solely on missense variants for this analysis; however many pharmacogenes have noncoding variants that contribute to drug response and could be assayed with an appropriately designed noncoding MAVE. First, we restricted our analysis to the 31 genes that are designated as CPIC level A or B where genetic information can be used to guide drug therapy. We annotated each gene according to the localization of the protein it encodes, length, number of missense variants already in gnomAD, and number of variants registered in PharmVar (Table 1.1).

Among this list, small proteins should be given high priority, since they have fewer possible variants and are thus easier to assay. Larger proteins affecting dosing of multiple, widely-prescribed drugs should also be prioritized, as they impact many patients. For these, we suggest focusing initial efforts on functionally important domains. All the genes have tens to thousands of variants deposited in gnomAD; however, most genes do not have any variants deposited yet in PharmVar. Therefore, concentrating on the genes that have the greatest number of rare variants in gnomAD, but no information in PharmVar, would yield new insight. Two pharmacogenes, *IFN3* and *MT-RNR1*, encode secreted proteins requiring new assays that maintain the sequence-phenotype link. In addition to these factors, analyzing published CRISPR screen data will identify which of these genes cause growth defects in a relevant cell line; growth-based MAVEs would be an attractive starting point for these. For the remainder, we suggest applying reporter-based assays such as VAMP-seq.

Despite their promise, MAVEs also have limitations. MAVEs often take the genetic element of interest out of its endogenous genomic or cellular context and thus demand careful validation of results. Data generated from MAVEs, while comprehensive, can be noisy. Thus, adequate replication is required to improve measurement accuracy and facilitate error estimation. Finally, MAVEs generally focus on one or a few experimental conditions and so may not fully

capture condition-dependent effects. For pharmacogenes, therefore, it will be critical to evaluate variants in physiologically relevant concentration, time and drug contexts.

In summary, a community-wide effort to apply MAVEs to high-priority pharmacogenes would result in variant effect maps that could aid in the interpretation of variants seen in the clinic. As pharmacogene variant effect maps are produced, they will yield a better understanding of pharmacogene biology and create opportunities for more rigorous, data-driven customization of patient treatment.

Gene	Drug(s)	Length (AA)	Total possible single AA variants	Missense variants in gnomAD	Missense variants in PharmVar	Localization
MT-RNR1	aminoglycoside antibacterials	16	300	0	0	Secreted
NUDT15	azathioprine, mercaptopurine, thioguanine	164	3,260	83	12	Cytoplasm
IFNL3	peginterferon alfa-2a, peginterferon alfa-2b, ribavirin	196	3,900	152	0	Secreted
HPRT1	mycophenolic acid	218	4,340	20	0	Cytoplasm
TPMT	azathioprine, mercaptopurine, thioguanine	245	4,880	119	0	Cytoplasm
OTC	valproic acid	354	7,060	88	0	Mitochondrion matrix
HLA-B	abacavir, allopurinol, carbamazepine, oxcarbazepine	362	7,220	180	0	Membrane; Single-pass type I membrane protein
HLA-A	carbamazepine, allopurinol	365	7,280	194	0	Membrane; Single-pass type I membrane protein
ASS1	valproic acid	412	8,220	211	0	Cytoplasm
ASL	valproic acid	464	9,260	242	0	Cytoplasm, extracellular exosome

CYP2C9	phenytoin, warfarin, acenocoumarol	490	9,780	381	55	Endoplasmic reticulum membrane, peripheral membrane
CYP2C19	amitriptyline, clopidogrel, citalopram, voriconazole	490	9,780	375	5	Endoplasmic reticulum membrane, peripheral membrane
CYP2B6	efavirenz, methadone	491	9,800	331	0	Endoplasmic reticulum membrane, peripheral membrane
CYP2D6	codeine, oxycodone, tamoxifen, tramadol	497	9,920	374	30	Endoplasmic reticulum membrane, peripheral membrane
CYP3A5	tacrolimus	502	10,020	215	11	Endoplasmic reticulum membrane, peripheral membrane
G6PD	rasburicase, chloramphenicol, chloroquine, ciprofloxacin	515	10,280	171	0	Cytoplasm, extracellular exosome, nucleus
CYP4F2	warfarin, acenocoumarol	520	10,380	344	2	Endoplasmic reticulum membrane, peripheral membrane
UGT1A1	atazanavir, irinotecan, belinostat	533	10,640	308	0	Endoplasmic reticulum membrane; Single-pass membrane protein
NAGS	carglumic acid	534	10,660	216	0	Mitochondrion matrix
GBA	velaglucerase alfa	536	10,700	247	0	Lysosome membrane, peripheral membrane protein
SLCO1B1	simvastatin, cerivastatin	691	13,800	399	0	Basolateral cell membrane, Multi-pass membrane protein

DPYD	capecitabine, fluorouracil	1,025	20,480	566	0	Cytoplasm
ABL2	valproic acid	1,182	23,620	509	0	Cytoplasm, cytoskeleton
POLG	valproic acid	1,239	24,760	762	0	Mitochondrion, mitochondrion matrix, mitochondrion nucleoid
ABCB1	antidepressants, digoxin	1,280	25,580	578	0	Cell membrane, multi-pass membrane protein
CFTR	ivacaftor	1,480	29,580	991	0	Apical cell membrane
CPS1	valproic acid	1,500	29,980	679	0	Mitochondrion, nucleus, nucleolus
CACNA1S	desflurane, enflurane, isoflurane, halothane	1,873	37,440	1,071	0	Cell membrane, sarcolemma, T-tubule, multi-pass membrane protein
SCN1A	carbamazepine	2,009	40,160	587	0	Cell membrane, multi-pass membrane protein
RYR1	desflurane, enflurane, isoflurane, halothane	5,038	100,740	2,663	0	Sarcoplasmic reticulum membrane, multi-pass membrane protein

Table 1.1. 31 genes designated by CPIC as A or B level genes, along with factors to consider when designing MAVEs.

2 Multiplexed characterization of variant abundance and activity reveals vitamin K epoxide reductase topology, active site residues and human variant impact

2.1 Introduction

The enzyme vitamin K epoxide reductase (VKOR) drives the vitamin K cycle, which activates blood coagulation factors. VKOR, an endoplasmic reticulum (ER) localized transmembrane protein encoded by the gene *VKORC1*, reduces vitamin K quinone and vitamin K epoxide to vitamin K hydroquinone[15,32]. Vitamin K hydroquinone is required to enable gamma-glutamyl carboxylase (GGCX) to carboxylate Gla domains on vitamin K-dependent blood clotting factors. VKOR is inhibited by the anticoagulant drug warfarin[17,60], and *VKORC1* polymorphisms contribute to an estimated ~25% of warfarin dosing variability[61]. For example, variation in *VKORC1* noncoding and coding sequence can cause warfarin resistance (weekly warfarin dose > 105 mg) or warfarin sensitivity (weekly warfarin dose < ~10 mg)[62,63].

Though 15 million prescriptions are written for warfarin each year (https://www.clincalc.com), fundamental questions remain regarding its target, VKOR. For example, the structure of human VKOR is unsolved, though a bacterial homolog has been crystallized[64]. A homology model based on bacterial VKOR has four transmembrane domains, but the quality of the homology model is unclear, as human VKOR has only 12% sequence identity to bacterial VKOR. Moreover, experimental validation of VKOR topology yielded mixed results: similar biochemical assays suggested either three- or four- transmembrane-domain topologies[65–67].

Topology informs basic aspects of VKOR function including where vitamin K and warfarin bind, so determining the correct topology and validating the homology model is critical. In particular, VKOR has four functionally important, absolutely conserved cysteines at positions

43, 51, 132, and 135, the orientation of which differs between the two proposed topologies. In the four transmembrane domain topology, all four cysteines are located on the ER lumenal side of the enzyme. In this topology, cysteines 43 and 51 are hypothesized to be "loop cysteines" that pass electrons from an ER-anchored reductase, possibly TMX[65], to the active site[68]. However, in the three transmembrane domain topology, these cysteines are located in the cytoplasm and other pathways would be required to convey electrons to the redox center. Even for non-catalytic residues, topology plays an important role. For example, vitamin K presumably binds near the redox center, and topology dictates which residues make up the substrate binding site.

To understand the effect of human variants and to define the vitamin K and warfarin binding sites, VKOR variant activity has been extensively studied in cell-based assays[30,31,60]. In addition to activity, VKOR protein abundance has also been studied because abundance is an important driver of disease and warfarin response. For example, VKOR R98W is a decreased-abundance variant that, in homozygous carriers, causes vitamin K-dependent clotting factor deficiency 2[15]. A 5' UTR polymorphism reduces VKOR abundance and can be used to predict warfarin sensitivity[69]. However, so far, the activity and abundance of only a handful of VKOR variants has been tested.

Here, we used multiplexed, sequencing-based assays[23] to measure the effects of 2,695 VKOR missense variants on abundance and 697 variants on activity. Our analysis of the large-scale functional data supports a four transmembrane domain topology, which an orthogonal evolutionary coupling analysis confirmed. Next, we identified distinct mutational tolerance groups, which are concordant with a four transmembrane homology model. Combining this homology model with variant abundance and activity effects, we identified an active site that contains the catalytic residues C132 and C135 and shares six positions with a previously proposed vitamin K binding site[60]. We found that of four conserved cysteines putatively critical

for function, only three are absolutely required, and analyzed the mutational signatures of two putative ER retention motifs. Human *VKORC1* variants present in genetic databases and contributed by a commercial genetic testing laboratory were each classified based on abundance and activity. While most variants show wild type-like activity, 25% show low abundance or activity, which could confer warfarin sensitivity or cause disease in a homozygous context. Finally, we analyzed warfarin resistance variants and found that they span a range of abundances, indicating that increased abundance is an uncommon mechanism of warfarin resistance.

2.2 Results

2.2.1 Multiplexed measurement of VKOR variant abundance using VAMP-seq

To measure the abundance of VKOR variants, we applied Variant Abundance by Massively Parallel sequencing (VAMP-seq), an assay we recently developed[53]. In VAMP-seq, a protein variant is fused to eGFP with a short amino acid linker. If the variant is stable and properly folded, then the eGFP fusion will not be degraded, and cells will have high eGFP fluorescence. In contrast, if the variant causes the protein to misfold, protein quality control machinery will detect and degrade the eGFP fusion, leading to a decrease in eGFP signal (Fig. 2.1a). mCherry is also expressed from an internal ribosomal entry site (IRES) to control for expression. Differences in abundance are measured on a flow cytometer using the ratio of eGFP to mCherry signal. To determine whether VAMP-seq could be applied to VKOR, we fused eGFP to VKOR N- or C-terminally, and found that both orientations had high eGFP signal (Appendix 1a). We compared N-terminally tagged wild type (WT) VKOR to R98W, a variant that ablates a

putative ER retention motif and reduces abundance[70], and to TMD1Δ, a deletion of residues 10-30 which comprise the putative first transmembrane domain (TMD1; Fig. 2.1b). Both reduced-abundance variants exhibited much lower eGFP:mCherry ratios than WT, demonstrating that VAMP-seq could be applied to VKOR.

We constructed a barcoded site-saturation mutagenesis VKOR library that covered 92.5% of all 3,240 possible missense variants. To express this library in HEK293T cells we used a Bxb1 recombinase landing pad system we previously developed[48]. In this system, each cell expresses a single VKOR variant. Recombined, VKOR variant-expressing cells were then sorted into quartile bins based on their eGFP:mCherry ratios. Each bin was deeply sequenced, and abundance scores were calculated based on each variant's distribution across bins. Raw abundance scores were normalized such that WT-like variants had a score of one and total loss of abundance variants had a score of zero (Fig. 2.1c). We performed seven replicates, which were well correlated (Appendix Table 1, Appendix 1b, mean Pearson's r = 0.73; mean Spearman's ρ = 0.7). Abundance score means and confidence intervals for each variant were calculated from the replicates.

The final dataset describes the effect of 2,695 of the 3,240 possible missense VKOR variants on abundance (Fig. 2.1d and 2.1e). Validation of 10 randomly selected variants spanning the abundance score range showed high concordance between individual eGFP:mCherry ratios assessed by flow cytometry and VAMP-seq derived abundance scores (Fig. 2.1f, Pearson's r = 0.96, Spearman's ρ = 0.97).

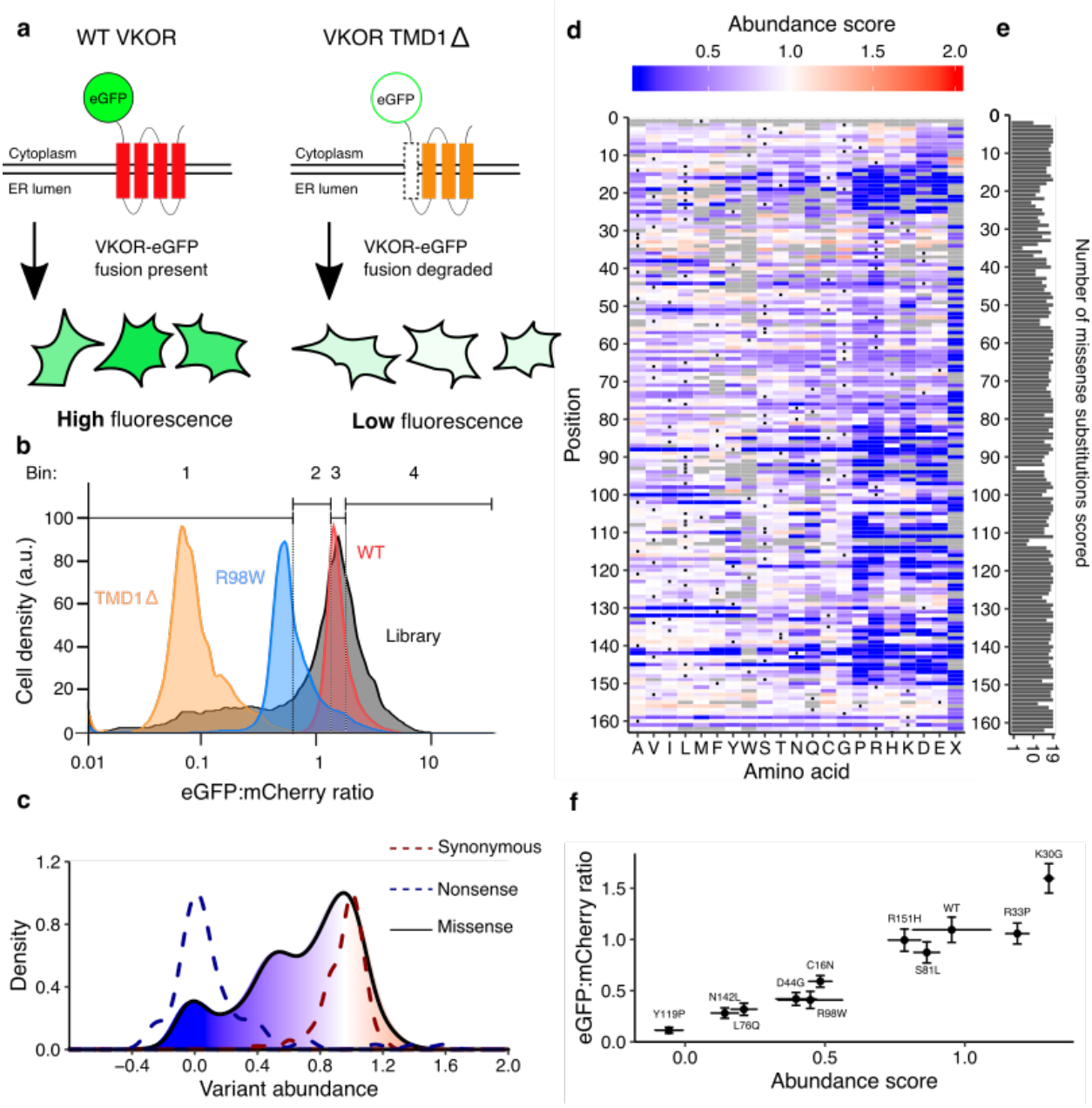


Figure 2.1. Multiplexed measurement of VKOR variant abundance using VAMP-seq.

a, To measure abundance, an eGFP reporter is fused to VKOR. eGFP-tagged WT VKOR is folded correctly, leading to high eGFP fluorescence. However, a destabilized variant is degraded by protein quality control machinery, leading to low eGFP fluorescence. **b**, Flow cytometry is

used to bin cells based on their eGFP:mCherry fluorescence intensity. Density plots of VKOR library expressing cells (grey, n = 12,109) relative to three controls: WT VKOR (red, n = 4,756), VKOR 98W (blue, n = 2,453) , and VKOR TMD1Δ (orange, n = 2,204) are shown. Quartile bins for FACS of the library are marked. **c**, Abundance score density plots of nonsense variants (dashed blue line, n = 88), synonymous variants (dashed red line, n = 127**)**, and missense variants (filled, solid line, n = 2,695). The missense variant density is colored as a gradient between the lowest 10% of abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). **d**, Heatmap showing abundance scores for each substitution at every position within VKOR. Heatmap color indicates abundance scores scaled as a gradient between the lowest 10% of abundance scores (blue), the WT abundance score (white), and abundance scores above WT (red). Grey bars indicate missing variants. Black dots indicate WT amino acids. **e**, Number of substitutions scored at each position for abundance. **f**, Scatterplot comparing VAMP-seq derived abundance scores to mean eGFP:mCherry (n = 1 replicate) ratios measured individually by flow cytometry. Variants were selected at random to span the abundance score range.

2.2.2 Multiplexed measurement of VKOR variant activity using a gamma-glutamyl carboxylation reporter

We also measured VKOR variant activity, adapting a HEK293 cell assay based on vitamin K- dependent gamma-glutamyl carboxylation of a cell-surface reporter protein[59]. In this assay, if VKOR is active, a Factor IX domain reporter is carboxylated, secreted and retained on the cell surface where it is detected with a carboxylation-specific, fluorophore-labeled antibody. However, if VKOR is inactive, the reporter is not carboxylated and the antibody cannot bind (Fig. 2.2a). We modified the HEK293 activity reporter cell line to eliminate endogenous VKOR

activity by knocking out both *VKORC1* and its paralog, VKORC1-like 1 (*VKORC1L1*)[30] (Appendix 2a). We also installed a Bxb1 landing pad to facilitate expression of individual VKOR variants or libraries (Appendix 2b,c). Recombination of WT *VKORC1* into the landing pad of the HEK293 VKOR activity reporter cell line yielded robust reporter activation, demonstrating that the reporter line could be used to assess the activity of a library of VKOR variants (Fig. 2.2b).

We recombined a library of *VKORC1* variants into the HEK293 activity reporter cell line and sorted recombinant cells into quartile bins based on carboxylation-specific antibody binding. Each bin was deeply sequenced and, as for VAMP-seq, an activity score was computed for each variant. Final activity scores and confidence intervals were computed from six replicates for a total of 697 missense variants, 21.5% of those possible (Appendix 2d, mean Pearson's $r = 0.62$ and mean Spearman's $\rho = 0.56$, Appendix Table 2). Our activity score density plot showed that most variants had WT-like activity scores (Fig. 2.2c).

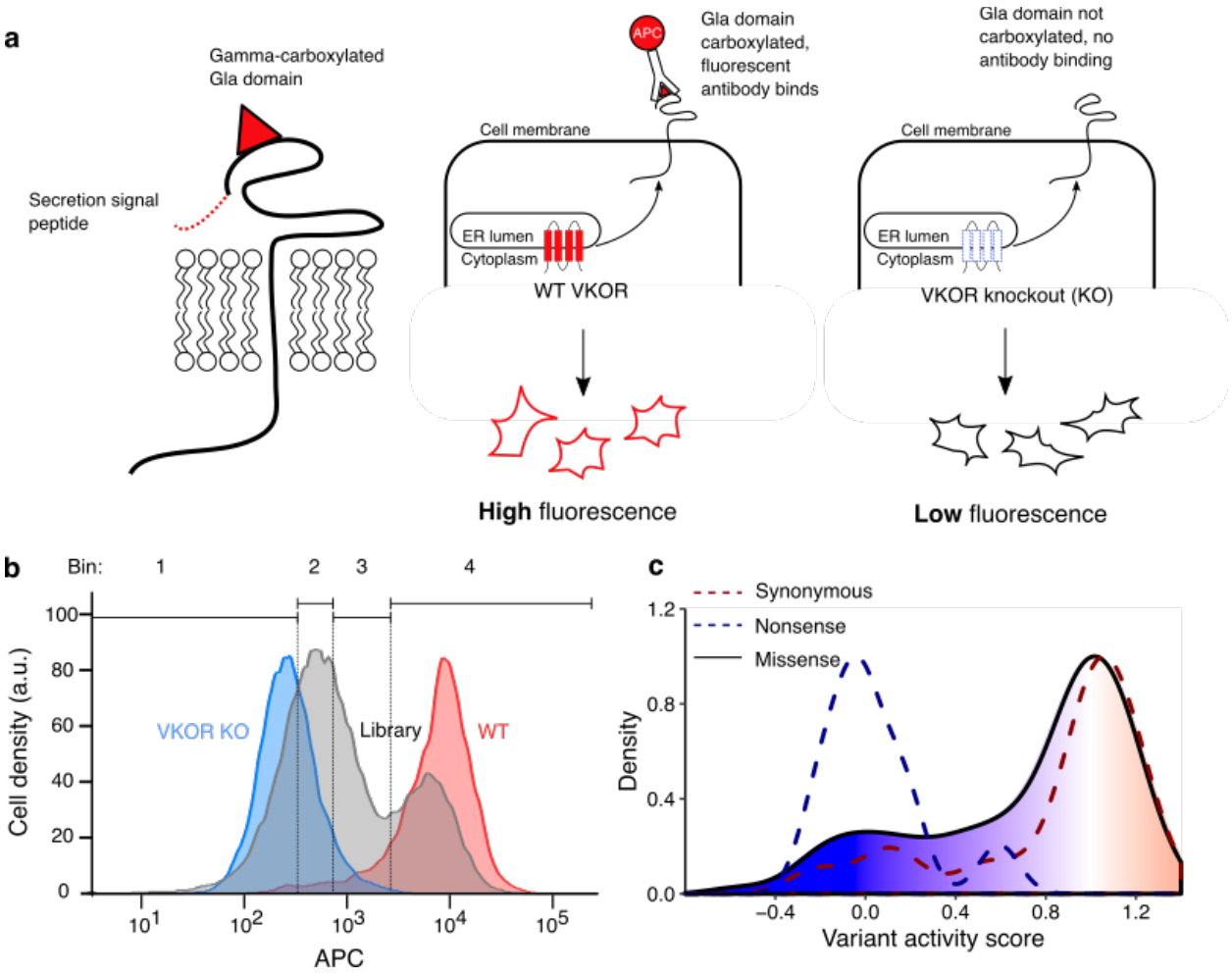


Figure 2.2. Multiplexed measurement of VKOR variant activity using a gamma-glutamyl carboxylation reporter.

a, left panel, a Factor IX Gla domain reporter is expressed inHEK293 cells and consists of a prothrombin pre-pro-peptide which allows for processing and secretion, a Factor IX Gla domain, and Proline rich Gla protein 2 (PRGP2) transmembrane and cytoplasmic domains. **middle panel,** Cells expressing WT VKOR carboxylate the reporter Gla domain, which, upon trafficking to the cell surface, can be stained using a carboxylation-specific antibody conjugated to the fluorophore APC. VKOR knockout cells do not carboxylate the reporter, so the fluorescent antibody does not bind. **b**, Density plots of HEK293 activity reporter cells stained with APC-labeled carboxylation-specific antibody expressing no VKOR (blue, n = 7,188), WT VKOR (red, n = 4,107), or the

VKOR variant library (grey, n = 41,418). Quartile bins for FACS of the library are marked. **c**, Activity score density plots of nonsense variants (dashed blue line, n = 14), synonymous variants (dashed red line, n = 35), and missense variants (filled, solid line, n = 697). The missense variant density is colored as a gradient between the lowest 10% of activity scores (blue), the WT activity score (white) and activity scores above WT (red).

2.2.3 Human VKOR has four transmembrane domains

Two different domain models, one with three transmembrane domains and another with four, have been proposed for human VKOR[64,66](Fig. 2.3a). Because charged amino acids occur infrequently in transmembrane domains and should be less tolerated, we reasoned we could discriminate between these two models using a sliding window average of the effect of charged substitutions on VKOR abundance[71,72]. We found four clearly demarcated regions where charged substitutions profoundly reduced VKOR abundance, relative to aliphatic substitutions (Fig. 2.3b). To exclude the possibility that the eGFP tag used in our VAMP-seq assay somehow affected topology, we also analyzed the activity score data. The activity data, derived using native, untagged VKOR, revealed the same four minima as the abundance data (Fig. 2.3c). In addition to these four minima, we also observed an activity score minimum at position 57, corresponding to a conserved serine at this position. This serine occurs at the end of the lumenal half-helix hypothesized to shield the active site from non-specific oxidation, so it is likely this signal is the result of disruption of that half helix. Together, these results strongly support the hypothesis that, like its distant bacterial homolog, human VKOR has four transmembrane domains.

To validate these findings, we performed evolutionary coupling analysis to infer the three-dimensional structure suggested by co-evolution. We aligned 2,770 VKOR sequences from

both eukaryotes and prokaryotes and identified coupled residues using the EVcouplings software[73,74]. Local patterns of evolutionary couplings (i.e. between nearby positions, *i* to *i*+4) supported a four-helix topology. The helices predicted by these local evolutionary couplings overlapped 70 of the 82 residues in alpha-helices of the bacterial structure (PDB 4NV5)[31] and present in our alignment (hyper-geometric test p-value = 3.26^{-23}, Fig. 2.3d).

We identified non-local evolutionary coupling patterns characteristic of three-dimensional contacts, which also strongly supported the four transmembrane domain model. Using these contacts, we computationally folded human VKOR, yielding a modeled structure similar to the bacterial structure (RMSD = 2.58 Å over 97/143 C_{alpha}, Appendix 3a,b). The predicted tertiary structure had a four-helix topology, with antiparallel contacts between transmembrane domains 1 and 2 (Fig. 2.3e, Fig. 2.3f) and between transmembrane domains 1 and 4 (Fig. 2.3e, Fig. 2.3g), which are only possible in a four-helix topology.

Comparison of our abundance data to the energy required to insert different amino acids into the membrane yields additional evidence for the four transmembrane domain model. The apparent change in free energy ($\Delta\Delta G_{app}$) of insertion relative to wild type for every amino acid has been determined experimentally using deep mutational scanning of bacterial membrane proteins[75]. Median abundance score and $\Delta\Delta G_{app}$ for each amino acid are correlated (Fig. 2.3h). In particular, the large energetic cost of insertion of transmembrane domains with charged amino acids is apparent, including within the second transmembrane domain TMD2. Beyond insertion energies of individual amino acids, the overall hydrophobicity of transmembrane helices contributes to membrane protein insertion[75], as well as topology[72] and degradation[76]. To determine whether overall helix hydrophobicity was a large factor contributing to abundance scores, we calculated the free energy for insertion (ΔG_{helix}) of each helix in the four transmembrane domain model using the ΔG prediction server v1.0[77]. The four helices of VKOR

have different ΔG_{helix}, with only transmembrane domain 3 having favorable ΔG_{helix} for insertion (TMD1: 0.435, TMD2: 1.551, TMD3: -1.749, and TMD4: 1.734). Interestingly, we observed that TMD3 has a high density of substitutions with WT-like scores (Appendix 3c), suggesting that TMD3's favorable insertion energy might explain its mutational tolerance.

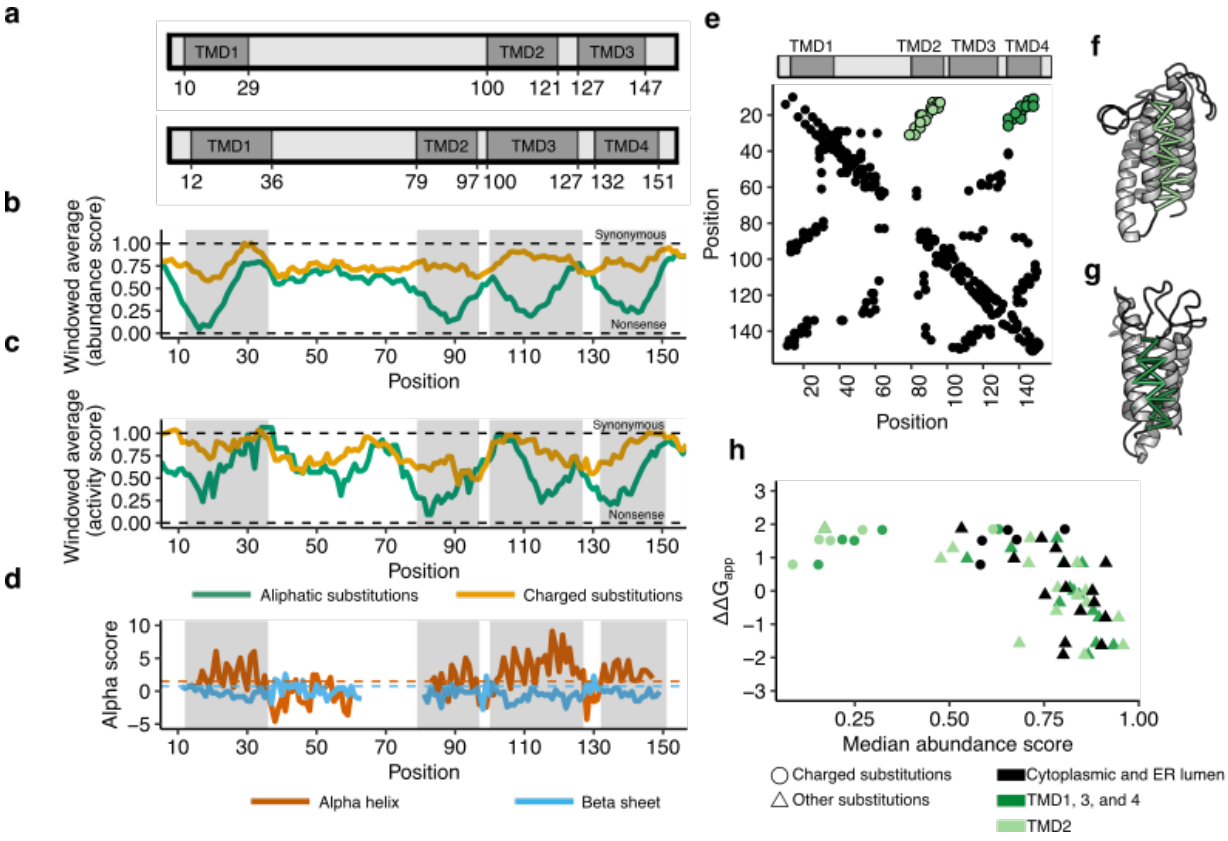


Figure 2.3. Abundance, activity, and evolutionary data support four transmembrane domains.

a, Three and four transmembrane domain (TMD) models of VKOR, with TMDs in dark grey[64,66]. **b**, Windowed abundance score means (width = 10 positions) for charged substitutions (green) and aliphatic substitutions (gold). Dark grey boxes correspond to TMDs proposed in the four domain model. Dashed lines show median synonymous and the nonsense abundance scores. **c**, Windowed activity score means (width = 10 positions) for charged substitutions (green) and

aliphatic substitutions (gold). Boxes and dashed lines as described in **b**. **d**, Secondary structure classification from local evolutionary couplings shown as alpha scores calculated for alpha helices (red) and beta sheets (blue). Dashed lines show significance cut-offs for alpha helices (1.5, red) and beta sheets (0.75, blue)[78]. **e**, A contact map derived from evolutionary couplings. Black points show pairs of positions with significant coupling. Light green points show predicted contacts between TMD1 and TMD2. Dark green points show predicted contacts between TMD1 and TMD4. **f**, Predicted tertiary contacts between TMD1-TMD2 (shown in light green in **e**) and **g**, TMD1-TMD4 (shown in dark green in **e**) shown on the evolutionary couplings-derived hVKOR structural model. **h**, Scatterplot comparing change in free energy for membrane insertion[75] ($\Delta\Delta G_{app}$) to median abundance score for each amino acid substitution. Cytoplasmic and lumenal positions shown in black, TMD2 in light green, and TMDs 1, 3, and 4 in dark green. Charged substitutions shown as circles, all other substitutions as triangles.

2.2.4 Detailed structural context of VKOR variant abundance effects

Having confirmed that human VKOR has four transmembrane domains, we next explored the detailed pattern of mutational effects we observed in the context of a four transmembrane domain homology model. We generated a homology model of human VKOR with I-TASSER using the bacterial VKOR structure[31,79]. We performed hierarchical clustering of positions based on abundance scores, which yielded four groups of positions with characteristic mutational patterns (Fig. 2.4a). In Group 1, most substitutions were neutral or increased abundance; in Group 2, charged amino acid and proline substitutions decreased abundance; in Group 3, all substitutions decreased abundance; and in Group 4, all substitutions decreased abundance profoundly. Each group corresponded to a spatially distinct region of the homology model structure (Fig. 2.4b).

Group 1 positions were located in or adjacent to cytoplasmic and ER lumenal loops, which were more tolerant of substitutions than the transmembrane domains. At four Group 1 positions, K30, R33, R35, and R37, almost every substitution increased abundance. These positively charged positions are positioned either at the edge of TMD1 (K30) or in the ER lumen directly abutting the top of TMD1 (R33, R35, and R37). The "positive inside rule"[80], suggests that positive charges in membrane proteins generally reside in the cytoplasm, and this phenomenon is important for driving topology and membrane insertion[72,80,81]. K30, R33, R35, and R37 violate the positive inside rule, and substitutions at these positions may increase abundance by reducing charge inside the ER, reducing topological frustration or increasing membrane insertion efficiency. Compared to the other 12 arginine and lysine positions in WT VKOR, K30, R33, R35, and R37 are the only ones where substitutions generally increased abundance (Appendix 4a)**.** Our observations are consistent with a screen of rat VKOR variants intended to improve protein expression in *E. coli* where deletion of positions 31 to 33 increased protein levels[82].

In Group 2, charged amino acids or proline substitutions generally decreased abundance. Group 2 consisted mostly of transmembrane positions that had side chains projecting into the lipid bilayer. Such transmembrane positions usually have hydrophobic, nonpolar side chains[83]. Proline has poor helix forming propensity, explaining why proline substitutions decreased abundance at these positions. Group 3 consisted of a mixture of cytoplasmic, ER lumenal and transmembrane positions where most substitutions decreased abundance. The cytoplasmic positions in this group included the putative dilysine ER localization motif at positions 159 and 161. Also in this group were R98, part of another putative ER retention motif at positions 98 and 100, and a glycine adjacent to TMD1 at position nine. The transmembrane positions had side

chains projecting towards neighboring transmembrane helices, suggesting that, as for other membrane proteins[84,85], intramolecular sidechain packing is important for abundance.

Finally, substitutions in Group 4, consisting of positions G19, Y88, I141, and L145, resulted in catastrophic loss of abundance. These positions are all in transmembrane domains with side chains projecting into the interior of the protein. On the basis of strict mutational intolerance of these positions, we hypothesized that their coordinated side chain packing comprises the core of the VKOR four helix bundle. Indeed, Group 4 residues had dramatically lower relative solvent accessibility than Groups 1-3 (Fig. 2.4c).

The four transmembrane domain homology models also allowed us to explain VKOR's unusual trimodal distribution of variant abundance scores. Previous VAMP-seq derived abundance score distributions for the cytosolic proteins TPMT and PTEN were bimodal (Appendix 4b)[53], and 15 of 16 deep mutational scans of other soluble proteins using a variety of other assays also exhibited bimodal functional score distributions[86]. Because VKOR is an ER resident, transmembrane protein, we hypothesized that its unusual trimodal abundance score distribution resulted from transmembrane domain substitutions. Indeed, the lowest mode of the distribution was composed almost exclusively of deleterious transmembrane domain substitutions (Fig. 2.4d). In contrast, the intermediate mode consisted of substitutions in the ER lumen, cytoplasm, and transmembrane domains. Similarly, substitutions that profoundly decreased activity occurred in transmembrane domains (Fig. 2.4e).

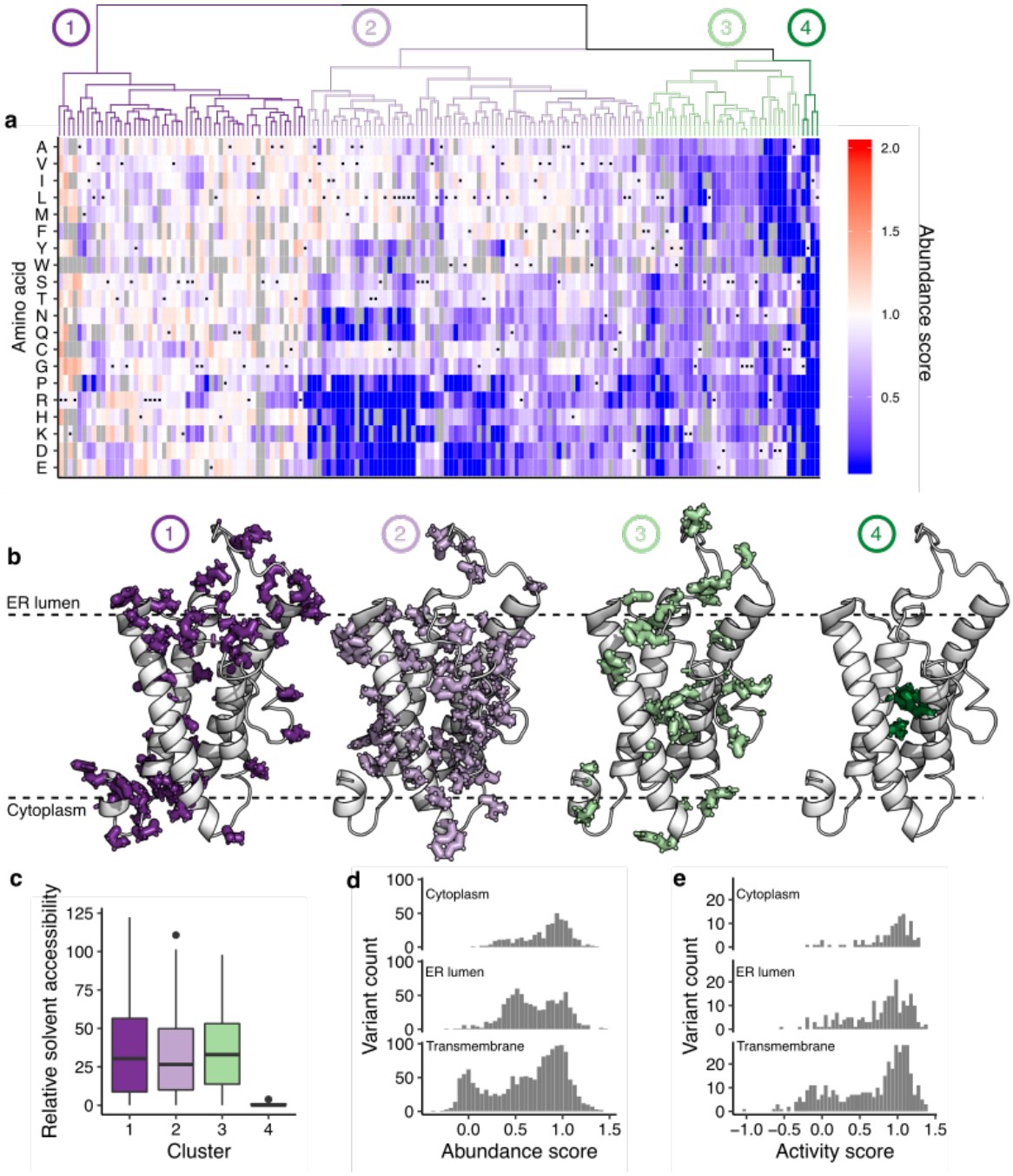


Figure 2.4. Hierarchical clustering of abundance scores and distributions of abundance and activity scores by domain.

a, A heatmap showing hierarchical clustering of positions based on abundance score vectors, with the dendrogram above. Groups of positions, chosen based on the dendrogram, are numbered and colored. Heatmap color indicates abundance scores scaled as a gradient between the lowest

10% of abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). Grey bars indicate missing variants. Black dots indicate WT amino acids. **b**, Positions in groups 1-4 shown on the VKOR homology model, with numbers and colors corresponding to panel **a**. **c**, Boxplot showing relative solvent accessibility of positions in each cluster determined using DSSP[87,88] and colored as in **b**. Bold black line shows median, box shows 25th and 75th percentile. Line shows 1.5 interquartile range above and below percentiles, and outliers are shown as black points. **d**, Histograms of abundance scores for missense variants in the cytoplasmic, ER lumenal, or transmembrane domains. **e**, Histograms of activity scores for missense variants in the cytoplasmic, ER lumenal, or transmembrane domains.

2.2.5 Variant activity and abundance identify functionally constrained regions of VKOR

We reasoned that our activity and abundance data could reveal the location of functionally important positions in VKOR, including the active site, since functionally important positions should have many loss-of-activity but few loss-of-abundance variants. Thus, we calculated the specific activity for each variant by taking the ratio of its rescaled activity score and abundance score (see Methods). We computed the median specific activity for each position; substitutions at positions with low median specific activity generally have low activity relative to their abundance. We set a specific activity threshold based on two absolutely conserved cysteines that form VKOR's redox center, C132 and C135. Using this threshold, positions with the lowest 12.5% of specific activity scores and with at least four variants scored for activity were deemed functionally constrained and mapped on the homology model of VKOR (Fig. 2.5a, Appendix 5a). These 11 functionally constrained positions are organized around C132 and C135 and define, at least in part, the VKOR active site (Fig. 2.5b,c, Appendix 5b,c). Among the functionally constrained positions are six positions previously identified in vitamin K docking

simulations[60] (Appendix 5d), including F55, which is hypothesized to bind vitamin K. Three functionally constrained positions, G60, R61, and A121, did not match any position in the active site predicted by docking, but were immediate neighbors of W59 and L120, positions that were.

Besides C132 and C135, VKOR has two additional absolutely conserved cysteines, C43 and C51. In the four transmembrane domain model, C43 and C51 are postulated to be loop cysteines that relay electrons to the C132/C135 redox center[89]. We classified C43 as having low specific activity, but we only observed one variant at this position, so it was not included in our set of functionally constrained positions (Appendix 5e). In contrast, substitutions at C51 resulted in only modest activity loss, a phenomenon that has been observed previously[31]. Interestingly, every substitution at C51 and 15 of 19 at C132 decreased VKOR abundance (Appendix 5f). Inside cells, the majority of VKOR molecules have a C51-C132 disulfide bond, and warfarin binds to this redox state of VKOR[31]. Since disruption of this disulfide bond apparently impacts abundance as well as activity, this bond may be important for VKOR folding and stability.

VKOR is thought to contain two sequences important for ER localization. The first is a diarginine motif (RxR) at positions 98-100, and the second is a dilysine motif (KXKXX) at positions 159-163. While we did not directly measure localization, we found that only six of 19 R98 variants and seven of 14 R100 variants resulted in low abundance (Appendix 5g). In contrast, nearly all variants at K159 (14 of 18) and K161 (17 of 19) resulted in low abundance (Appendix 5h). A histidine substitution was tolerated at position 161, which mimics the KXHXX motif commonly found in coronaviruses and a small number of human proteins[90]. Because protein localization and degradation are coupled[91], we suggest that the reductions in abundance we observe are the result of degradation caused by mislocalization, and that the dilysine motif at positions 159-163 is essential for VKOR ER localization. Overall, comparison of VKOR variant

activity and abundance revealed functionally important regions, refining our understanding of the active site, redox-active cysteines, and ER retention motifs.

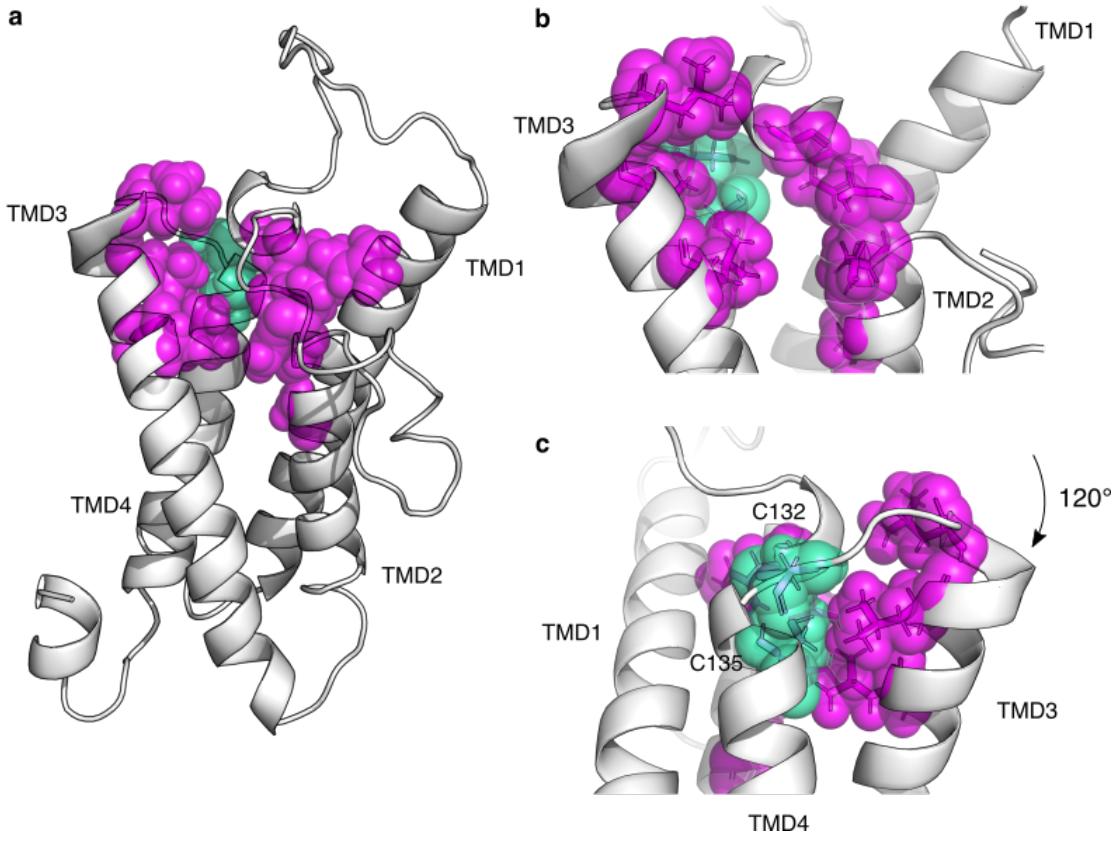


Figure 2.5. Functionally constrained positions reveal VKOR active site and critical cysteines. **a**, Positions with the lowest 12.5% of median specific activity scores and at least four variants scored for activity are shown as magenta spheres on the VKOR homology model. Cysteines C132, and C135, also in the bottom 12.5% of median specific activity scores, are shown in green spheres. **b**, Magnified view of the redox center cysteines (positions 132, and 135, green spheres) and surrounding residues that define the active site (magenta spheres). Residues shown in transparent spheres, with side chains also shown in sticks. **c**, panel **b** rotated 120°C.

2.2.6 Functional consequences of VKOR variants observed in humans

Variation in VKOR is linked to both disease and warfarin response, but the overwhelming majority of VKOR variants found in humans so far have unknown effects. Thus, we curated a total of 215 variants that had either been previously reported in the literature as affecting warfarin response, were in ClinVar, were in gnomAD v2 or v3, or were present in individuals whose healthcare provider had ordered a multi-gene panel test from a commercial testing laboratory (Color Genomics). Of eight variants present in ClinVar, we included only one (D36Y) in our analysis as it was the only variant reviewed by an expert panel[92]. 159 variants were present in gnomAD, and all but one missense variant (D36Y) had population frequencies less than 0.2%. 28 were literature-curated warfarin response variants, only 12 variants of which were in one of the databases surveyed. D36Y was the only warfarin response variant present in all databases, ClinVar, gnomAD, and Color (Appendix 6).

We classified 193 of the 215 variants we curated according to their abundance. All synonymous variants with the exception of two were WT-like or possibly WT-like, while the three nonsense variants scored as having low abundance (Fig. 2.6a). Missense variants spanned all abundance categories, with 129 (60%) having WT-like or possibly WT-like abundance. 30 missense variants were low abundance, and 12 were high abundance. The single known pathogenic variant R98W was low abundance (Fig. 2.6b). We also classified 54 variants according to their activity. Only one variant, A115V, exhibited low activity. It had WT-like abundance, indicating that the loss of activity is not due to loss of abundance.

We examined warfarin response variants including W5X, the only variant observed so far linked to human warfarin sensitivity[93]. As expected, W5X was low abundance, reinforcing that heterozygous loss of VKOR is the cause of warfarin sensitivity in carriers of this variant.

Warfarin resistance variants, on the other hand, are predicted to abrogate warfarin binding[64], but it is unclear whether these variants have appreciable effects on abundance or activity. We found that warfarin resistance variants span a range of abundances and that the distribution of warfarin resistant variant abundance was not different from missense variants generally (Fig. 2.6c, two-sided Kolmogorov-Smirnov test p= 0.438). Five warfarin-resistance variants had low abundance, suggesting that these variants must block drug binding or increase activity to confer resistance. One variant, A26T, had high abundance, a possible mechanism of warfarin resistance. The five warfarin resistance variants, R58G, W59L, V66M, G71A, and N77S, whose activity we scored, were all WT-like. Thus, our abundance and activity data are consistent with warfarin resistance arising largely from variants that block warfarin binding.

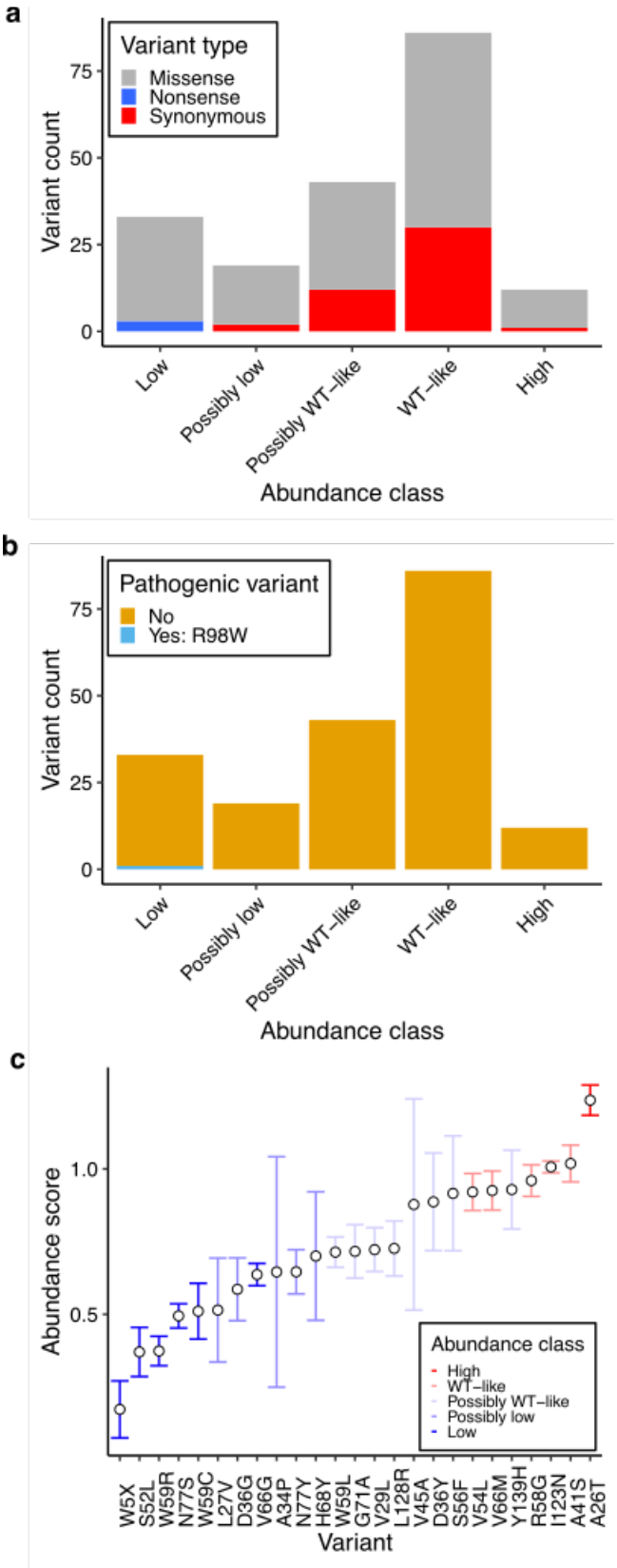

a
Variant type
Missense
Nonsense
Synonymous
Variant count
0
25
50
75
Low
Possibly low
Possibly WT−like
WT−like
High
Abundance class
b
Pathogenic variant
No
Yes: R98W
Variant count
0
25
50
75
Low
Possibly low
Possibly WT−like
WT−like
High
Abundance class
c
Abundance score
0.5
1.0
Abundance class
High
WT−like
Possibly WT−like
Possibly low
Low
W5X
S52L
W59R
N77S
W59C
L27V
D36G
V66G
A34P
N77Y
H68Y
W59L
G71A
V29L
L128R
V45A
D36Y
S56F
V54L
V66M
Y139H
R58G
I123N
A41S
A26T
Variant

Figure 2.6. Characterization of human variants using abundance and activity data.

a, Histogram of abundance classifications for variants from gnomAD, ClinVar, and Color Genomics. Nonsense variants colored in blue, synonymous in red, and missense in grey. **b**, Histogram of abundance classifications for same variants in **a**, colored by pathogenicity. The only variant known to cause disease, R98W, is colored in blue. All other variants shown in yellow. **c**, Scatterplot showing abundance scores for literature-curated warfarin resistance variants. Bars show standard error and are colored by abundance class. Variants are arranged in order of abundance score.

2.3 Discussion

We conducted multiplexed assays to measure the effects of 2,695 VKOR variants on abundance and 697 variants on activity. Both abundance and activity data provided evidence for a four transmembrane topology, which was further supported by evolutionary couplings analysis. We evaluated a VKOR homology model in the context of the patterns of variant effects on abundance we measured and found that the homology model could explain these patterns. Low specific activity residues mapped onto this homology model identify, at least in part, the active site, which largely overlaps with the results of a vitamin K docking simulation[60]. Our active site is shallower than what the docking simulation predicts; this is the result of low abundance scores at some of the deeper, transmembrane positions predicted by docking to bind the isoprenoid chain of vitamin K (F87, Y88), and poor coverage of activity scores for other positions (V112, S113). In light of the fact that substitutions at F87 and Y88 resulted in low abundance, we note that the modeled vitamin K binding mode would disrupt packing of VKOR core residues and require repacking of helices to maintain protein stability[94]. In addition to the active site,

substitutions at the dilysine and, to a lesser extent, the diarginine ER localization motifs caused abundance loss.

We also used our large-scale functional data to analyze 215 VKOR variants found in humans. 16% of these variants affect neither activity nor abundance; we identified 54 previously uncharacterized low abundance or low activity variants that could be pathogenic or alter warfarin response. We found that only one warfarin resistance variant had increased abundance, indicating that increased abundance is not a pervasive warfarin resistance mechanism. All five of the warfarin resistance variants whose activity we scored were WT-like. Taken together these data support the notion that warfarin resistance generally involves alterations to warfarin binding rather than abundance or activity. We analyzed one known warfarin sensitivity variant, W5X, and found that it is low abundance, suggesting that one should not exclude the possibility that any of the 52 other low abundance variants, if found in a person, also confer warfarin sensitivity.

While our VKOR variant abundance and activity data illuminates various aspects of VKOR's structure and function, the data have limitations. For example, neither assay captures variant effects on mRNA splicing. Both assays have limited dynamic ranges, meaning that subtle effects on abundance or activity cannot be discerned. In addition, both assays have inherent noise, largely arising from the limited number of cells we can sample due to the bottleneck of cell sorting. We account for this noise by filtering each dataset based on variant frequency and presenting a confidence interval for each abundance and activity score.

In the future, we envision that the assays we used could be employed to better understand VKOR's interaction with warfarin. Here, we could measure warfarin's effect on both variant abundance and activity, mapping the warfarin binding site more finely. In addition, we could identify warfarin resistance mutations that have not yet been observed in the clinic and group variants by their putative resistance mechanism. Overall, our work highlights the value of

multiplexed assays of variant effect for better understanding protein structure, function and human variant effects.

2.4 Methods

General reagents, DNA oligonucleotides, and plasmids.

Details on general reagents can be found in Supplementary Table 5. Unless otherwise noted, all chemicals were obtained from Sigma and all enzymes were obtained from New England Biolabs. *E. coli* were cultured at 37°C in Luria broth. All cell culture reagents were purchased from ThermoFisher Scientific unless otherwise noted. HEK 293T cells (ATCC CRL-3216) and derivatives thereof were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U ml^{-1} penicillin, and 0.1 mg ml^{-1} streptomycin. Cells were induced with 2.5 ug mL^{-1} doxycycline. Cells were passaged by detachment with trypsin-EDTA 0.25%, and cells were prepared for sorting by detachment with versene. All cell lines tested negative for mycoplasma. Because our activity assay is vitamin-K dependent, all activity assays were done with the same lot of FBS to ensure similar concentrations of vitamin K in each replicate.

All synthetic oligonucleotides were obtained from IDT and can be found in Supplementary Table 6. All non-library-related plasmid modifications were performed with Gibson assembly[95].

Library construction

A gBLOCK encoding the sequence for human VKOR was ordered from IDT. It was then cloned into the vector pHSG298 (Clontech). Saturation mutagenesis primers were designed for

each codon in VKOR from positions 2 to 163[46] and ordered resuspended from IDT. Forward and reverse primers for each position were mixed at 2.5 mM, and used in a PCR reaction with 125 pg of pHSG298-VKOR, 5% DMSO, and 5 uL of KAPA Hifi Hotstart 2X ReadyMix. PCR products were visualized on a 0.7% agarose gel to confirm amplification of the correct product.

PCR products were then quantified using the Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) using DNA control curves done in triplicate. To pool, a total amount of DNA for each reaction was calculated that maximized the volume to be drawn from the lowest concentration PCR product. Pooled PCR products were cleaned and concentrated using Zymogen Clean and Concentrate kit and then gel extracted. The pooled library was phosphorylated with T4 PNK (NEB), incubated at 37°C for 30 minutes, 65°C for 20 minutes, and then 4° indefinitely. 8.5 uL of this phosphorylated product was combined with 1 uL of 10X T4 ligase buffer (NEB) and 0.5 uL of T4 DNA ligase (NEB) to make a 10 uL overnight ligation reaction. This reaction was incubated at 16°C overnight.

The overnight ligation was then cleaned and concentrated (Zymogen) and eluted in 6 uL of ddH2O. 1 uL of this ligation was then transformed into high efficiency *E. coli* using electroporation at 2 kV. Each reaction contained 1 uL of ligation (or ligation control or pUC19 10 pg/uL) and 25 uL of E. coli. 975 uL of pre-warmed SOC media was added to each cuvette after electroporation, transferred to a culture tube, and recovered at 37°C, shaking for 1 hour. At 1 hour, 1 and 10 uL samples from all cultures were taken and plated on appropriate media (LB + kanamycin for ligation and ligation control; LB + ampicillin for pUC19), the remaining 989 uL was used to inoculate a 50 mL culture (+ kanamycin). Plates and 50 mL culture were incubated at 37°C overnight (shaking for 50 mL culture). Colonies on plates were then counted, and counts

were used to calculate how many unique molecules were transformed to gauge coverage of the library. 50 mL culture was spun down and midiprepped.

To transfer the library from pKan to the recombination vector, the pKan library and recombination vector were digested with XbaI and AflII for 1 hour at 65°C. The library and cut vector were then gel extracted. The library was then ligated with the cut vector at 5:1 using T4 ligase, overnight at 16°C. The ligation was heat inactivated the next morning, clean and concentrated. Another high efficiency transformation was performed the same as described above, except this ligation was plated on LB + ampicillin (antibiotic switching strategy). Plates and 50 mL culture were incubated at 37°C overnight (shaking for 50 mL culture). Colonies on plates were then counted, and counts were used to calculate how many unique molecules had been transformed to gauge coverage of the library. A 50 mL culture was spun down and midiprepped.

To barcode individual variants, plasmid library harvested from midiprep was digested with EcoRI-HF and NdeI at 37°C for 1 hour, 65°C for 20 minutes. Barcode oligos were ordered from IDT, resuspended at 100 uM, and then annealed by combining 1 uL each of primer with 4 uL CutSmart Buffer and 34 uL ddH2O and running at 98°C for 3 minutes followed by ramping down to 25°C at -0.1°C/second. After annealing, 0.8 uL of Klenow polymerase (exonuclease negative, NEB) and 1.35 uL of 1 mM dNTPS was then combined with the 40 uL of product to fill in the barcode oligo (cycling conditions: 25°C for 15:00, 70°C for 20:00, ramp down to 37°C at -0.1°C/s). Digested vector and barcode oligo were then ligated overnight at 16°C.

The overnight ligation was then cleaned and concentrated and eluted in 6 uL of ddH2O. 1 uL of this ligation was then transformed into high efficiency *E. coli* using electroporation at 2 kV. Each reaction contained 1 uL of ligation (or ligation control or pUC19 10 pg/uL) and 25 uL of *E. coli*. 975 uL of pre-warmed SOC media was added to each cuvette after electroporation, transferred to a culture tube, and recovered at 37°C, shaking for 1 hour. At 1 hour, 1 and 10 uL samples from water and pUC19 cultures were taken and plated on LB supplemented with ampicillin. For ligation and ligation control, four flasks were prepared with 50 mLs of LB and ampicillin, and then 500 uL, 250 uL, 125 uL, 62.5 uL was sample from the 1 mL of recovery and transferred into a corresponding flask. From those flasks, 1 uL, 10 uL, and 100 uL, were sampled and plated onto LB ampicillin plates. Plates and 50 mL culture were incubated at 37°C overnight. Colonies on plates were then counted, and counts were used to calculate how many unique molecules were transformed to gauge number of barcodes. Flask with the target number of barcodes was then spun down and midiprepped.

Cell line description

VAMP-seq assay cell line

HEK293T cells with a serine integrase landing pad integrated at the AAVS1 locus were used[48].

Activity assay cell line

We used a previously published reporter cell line[59] and inserted a recombinase-based landing pad at the *AAVS1* safe harbor locus using a previously published strategy[48]. Single cell clones were transfected with TALENs for AAVS1 and the landing pad plasmid, and single cell

clones were sorted. Presence of one landing pad was confirmed by 1) barcode sequencing of the landing pad and 2) co-transfection experiment with GFP and mCherry. From this, we moved forward with one clone demonstrated to have only one landing pad present (clone 45).

gRNAs to delete portions of the first exon of both *VKORC1* and *VKORC1L1* were ordered and cloned into pSpCas9(BB)-2A-GFP (PX458), which was a gift from Feng Zhang (Addgene plasmid #48138 ; http://n2t.net/addgene:48138 ; RRID:Addgene_48138). Clone 45 was then transfected with these four plasmids, and single cells were sorted based on GFP positivity. Disruption of *VKORC1* and *VKORC1L1* was confirmed by performing nested PCR, TA cloning, and then sequencing of products. We detected three alleles for both *VKORC1* and *VKORC1L1*, indicating that these loci are triploid in HEK293 cells.

A Western blot was also used to confirm absence of VKOR protein product in our activity reporter cell line. Protein lysates were harvested from ~1 million cells using 100 uL NP40 lysis buffer with freshly prepared protease inhibitor cocktail and 1 mM PMSF. Protein lysates were Qubited for concentration, and 20 ug of each protein lysate was loaded. 4-12% BisTris NuPage gel (Thermo Fisher) was used with MES buffer + 500µl of antioxidant added to the inner chamber. The gel ran at 150V for 90 min. Gel was then transferred to Nitrocellulose using 1X transfer buffer 20% EtOH at 24V for 1 hour on ice. The blot was washed for 5 minutes with 1X TBS-T 0.1% Tween 3 times. Blot was then blocked for overnight 1X TBS-T 0.1% Tween + 5% Milk. Blot was then washed for 5 minutes with 1X TBS-T 0.1% Tween 3 times. Blot was then cut in half at the between the 25kDa and 35kDa molecular weight markers. The bottom blot was incubated with: αVKOR 1:1000 + 1X TBS-T 0.1% Tween + 5% Milk. The top blot was incubated with αbeta-actin dHRP 1:1000+ 1X TBS-T 0.1% Tween + 5% Milk. Both blots were incubated with their primary antibodies overnight at 4ºC. The αVKOR blot was washed for 5 minutes with 1X TBS-T 0.1% Tween 3 times. The αVKOR blot was then incubated

with 1:10,000 secondary anti-mouse-HRP (GE Healthcare NA931V) + 1X TBS-T 0.1% Tween + 5% Milk for one hour. The αbeta-actin dHRP blot remained in primary antibody during this time, as no secondary antibody is needed for a direct HRP conjugate. Both blots were then washed for 5 minutes with 1X TBS-T 0.1% Tween 3 times. Blots were then incubated with Supersignal West Dura Extended Duration Substrate (Thermo Fisher). 500µl of both substrates incubated on blot for 5 min. Blots were then dried by kimwipe and exposed using the colorimetric and chemiluminescence functions on the BioRad ChemiDoc MP (Biorad).

Recombination of variants in cell lines: Abundance assay

Cells were transfected in six well plates, 250,000 cells per well (12-24 wells transfected total for each experiment). Sequential transfections were performed. On day 1, 3 ug of pCAG-NLS-Bxb1 was diluted in 250 uL of OptiMEM and 6 uL of Fugene (Promega). On day 2, 3 ug of barcoded library was diluted in 250 uL of OptiMEM and 6 uL of Fugene6 and transfected. 48 hours after this second transfection, cells were induced with doxycycline at a final concentration of 2.5 ug/mL.

Recombination of variants in cell lines: Activity assay

Cells were transfected in six- well plates, 500,000 cells per well (18-24 wells transfected total for each experiment). 272 ng of pCAG-NLS-Bxb1 was diluted in 125 uL of OptiMEM with 2.7 ug of barcoded library. 2.25 uL of Lipofectamine 3000 (Thermo Fisher) was diluted in 125 uL of OptiMEM in a separate tube. The DNA mixture was then added to the Lipofectamine 3000 mixture and incubated at room temperature for 15 minutes. Transfection mixture was then added dropwise to one six-well plate. Cells were induced with doxycycline 48 hours after transfection, with a final concentration of 2.5 ug/mL doxycycline .

Enrichment sorting for recombined cells

Cells were washed once with PBS, then dissociated with versene. Media was added to dilute EDTA, and cells transferred to 15 mL conical and spun down at 300 x g for 4 minutes. Media was aspirated off, and cells were resuspended in PBS, then filtered through a 35 um nylon mesh filter. Cells were sorted on a BD Aria III FACS machine. mTagBFP2, expressed from the unrecombined landing pad, was excited with a 405 nM laser. Recombined cells either expressed mCherry (abundance) or eGFP (activity), and these were excited by a 561 nm laser and a 488 nm laser, respectively. Samples were gated for live cells using FSC-A and SSC-A, then singlets using SSC-H vs. SSC-W, FSC-H vs. FSC-W. For activity assay reporter cell line, cells were then sorted for DsRed positivity to ensure robust expression of reporter. Cells that had successfully recombined a single VKOR variant were gated on recombinant mTagBFP2 negativity and either mCherry positivity (abundance) or eGFP positivity (activity) (see Supplementary Fig. 2b for gating example). Recombined cells were sorted on "Yield" mode in the BD Diva software and grown out for 3-5 days.

Abundance assay quartile sorting

Recombined cells were run on a BD Aria III FACS machine. Cells were prepared for sorting as described above, and were then gated for live, recombined singlets. A ratio of eGFP/mCherry was created using the BD Diva software as a unique parameter, and the histogram of this ratio was divided into four equal bins. Each quartile was sorted into a 5 mL tube on "4-Way Purity" mode. Sorted cells were grown out for 2-4 days post sorting to ensure enough DNA for sequencing. The details of replicate sorts for activity assay are in Supplementary Table 1.

Antibody conjugation

Factor IX Gla domain antibody specific for carboxylation was conjugated to APC following LYNX Rapid APC Antibody Conjugation Kit instructions. Antibody was resuspended at 1 mg/mL in nuclease-free water. 1 uL of Modifier reagent was then added for every 10 uL of antibody and mixed by pipetting. That mixture was then pipetted directly onto the LYNX lyophilized mix and gently mixed by pipetting up and down twice. The conjugation mixture was then capped and incubated in the dark at room temperature overnight. After overnight incubation, 1 uL of Quencher reagent was added for every 10 uL of antibody used and left to incubate for 30 minutes. At that point, antibody was divided into 20 uL aliquots to be used for replicate experiments and stored at -20°C.

Activity assay antibody staining and quartile sorting

Cells were plated in six-well plates at 500,000 cells per well with D10 media with no doxycycline. All replicates were performed with 18-24 wells of cells total. After 24 hours, doxycycline was added to cells to induce expression of reporter and VKOR variant. Cells were then incubated with doxycycline for 48 hours. On day of cell sorting, each six well was washed with cold PBS, dissociated with 200 uL of versene, and then resuspended in 1 mL of phenol red-free DMEM + 1% FBS and transferred to a 5 mL FACS tube. Cells were spun at 300 x g, then washed once with 1 mL of phenol red-free DMEM + 1% FBS. Cells were spun at 300 x g, and after aspirating supernatant, cell pellet was resuspended in 100 uL of antibody diluted 1:100 in phenol red-free DMEM + 1% FBS. Cells were incubated in antibody for 20 minutes at 4°C in the dark, with vortexing at five minute intervals to ensure staining. After 20 minutes, 1 mL of staining buffer was added to each tube to dilute out antibody. Cells were spun at 300 x g, washed

twice more similarly with staining buffer, then resuspended in 200 uL. At this point, all tubes were pooled and filtered to remove clumps. Cells were then sorted using a FACSAria III (BD Biosciences) into bins based on their APC intensity. First, live, single, recombinant cells were selected as described above. A histogram of APC was created and gates dividing the library into four equally populated bins based on APC fluorescence intensity were drawn. The details of replicate sorts for activity assay are in Supplementary Table 2.

gDNA prep, barcode amplification, and sequencing

Cells were then collected, pelleted by centrifugation and stored at −20 °C. Genomic DNA was prepared using a DNEasy kit, according to the manufacturer's instructions (Qiagen), with the addition of a 30 min incubation at 37 °C with RNAse in the re-suspension step. Eight 50 µl first-round PCR reactions were each prepared with a final concentration of ~50 ng µl−1 input genomic DNA, 1 × Q5 High-Fidelity Master Mix and 0.25 µM of the KAM499/VKORampR 1.1 primers. The reaction conditions were 98 °C for 30 s, 98°C for 10 s, 65°C for 20 s, 72°C for 60 s, repeat 5 times, 72°C for 2 min, 4°C hold. Eight 50 µl reactions were combined, bound to AMPure XP (Beckman Coulter), cleaned and eluted with 21 µl water. Forty percent of the eluted volume was mixed with Q5 High-Fidelity Master Mix; VKOR_indexF_1.1 and one of the indexed reverse primers, PTEN_seq_R1a through PTEN_seq_R2a, were added at 0.25 µM each. These reactions were run with Sybr Green I on a BioRad MiniOpticon; reactions were denatured for 3 minutes at 95°C and cycled 20 times at 95°C for 15s, 60°C for 15s, 72°C for 15s with a final 3 min extension at 72°C. The indexed amplicons were mixed based in relative fluorescence units and run on a 1% agarose gel with Sybr Safe and gel extracted using a freeze and squeeze column (Bio-Rad). The product was quantified using Kapa Illumina Quant kit.

Subassembly

Barcoded VKOR library was subassembled using a MiSeq 600 kit (Illumina). Two amplicons were generated, one forward, one reverse. PCR reactions were each prepared with ~500 ng input plasmid DNA, 1 × KAPA High-Fidelity Master Mix and 0.25 μM of the VKOR_SA_amp_F/VKOR_SA_amp_R or VKOR_SA_for_amp_R2.0/VKOR_SA_rev_amp_F2.0 primers. PCR reactions were run at 95°C for five minutes, then cycled 15 times at 98°C for 0:20, 60°C for 0:15, 72°C for 0:30, with a final extension at 72°C for 2:00. Amplicons (741 bp) were gel extracted on a 1.0% gel run at 130V for 35 mins. The product was quantified using Qubit and Kapa Illumina quant kit. Read lengths were as follows: 289 bp forward read, 18 bp index1, 18 bp index 2 (index = barcode forward and reverse). All reads sharing a common barcode sequence were collapsed to form the consensus variant sequence, resulting in 175,052 barcodes after filtering.

Barcode counting and variant calling

Enrich2 was used to quantify barcodes from bin sequencing, using a minimum quality filter of 20[96]. FASTQ files containing barcodes and the barcode map for VKOR were used as input for Enrich2. Enrich2 configuration files for each experiment are available on the GitHub repository. Barcodes assigned to variants containing insertion, deletions, or multiple amino-acid alterations were removed from the analysis.

Calculating scores and classifications

Scores and classifications were assigned using previously published analysis pipeline[53]. Briefly, for each protein variant, frequencies in each bin were calculated by dividing counts by total counts. From there, we filtered variants based on the number of experiments in which it was

observed ($F_{expt} = 2$) and their frequency ($F_{freq} = 10^{-4}$), after noticing that low frequency variants introduced noise to the analysis. These frequencies were then each weighted by multiplying by 0.25, 0.5, 0.75, and 1 in a bin-wise fashion. We generated a replicate score for each variant by using min-max normalization: normalizing to the median weighted average of the nonsense distribution set at 0 and the median weighted average of the synonymous distribution set at 1. We then averaged those scores for a final, experiment-wide variant score. Standard deviation and standard error were also calculated for each variant, and 95% confidence intervals were estimated using standard error, assuming a normal distribution. Abundance and activity classifications were assigned by assessing variant score and confidence intervals in relation to synonymous variant distribution. To do this, we established a cut-off that separated the 5% of synonymous variants with the lowest abundance (or activity) scores from the 95% of synonymous variants with higher abundance (or activity) scores. Variants with both scores and upper confidence intervals below this threshold were classified as "low," while those with scores below but upper confidence above were classified as "possibly low." Variants with scores and lower confidence intervals above the threshold were classified as "WT-like", while those with scores above lower confidence intervals below the threshold were classified as "possibly WT-like." Finally, another threshold was set that separated the 5% of synonymous variants with the highest scores from the rest of the synonymous distribution. Variants that had scores above this threshold, with lower confidence intervals above the lower threshold were classified as "high."

Windowed abundance and activity analysis

Windowed averages of abundance and activity scores were calculated using a window length of 10 positions with center alignment. Scores were calculated for both charged amino acids (R, K, H, D, E,) and aliphatic amino acids (G, A, V, I, L).

Evolutionary couplings analysis

EVcouplings extracts the constraints between pairs of residues, as evidenced in alignments of homologous sequences: first homologous sequences must be collected and aligned, and then a model of statistical energy costs and benefits between residues is fit to explain the sequence variation in the alignment. We collected an alignment of 2770 sequences using jackhammer (http://hmmer.org/) to query the human VKOR sequence against UniRef100 (https://www.uniprot.org/uniref/), with a bitscore per residue cutoff of 0.4 and 7 search iterations. We predicted secondary structures where the summed strength of couplings at would-be alpha helix and beta strand contacts scored above 1.5 and 0.75 respectively, for two or more consecutive residues.[42] We extended the called helices and strands by one residue on each side for a minimum structure size of four residues. All methods used for building alignments, training the model, folding, and predicting secondary structure are part of the EVcouplings software (https://evcouplings.org/)[97].

Homology modeling

A homology model of human VKOR was made by accessing I-TASSER[79] and using PDB structure 4NV5 as a template for threading. Model1 from results was used for all figures in this paper.

Hierarchical clustering

Hierarchical clustering was performed on abundance score vectors for each position using the hclust function in R. Dendrogram for hierarchically clustered heatmap was drawn using dendextend package (version 1.12.0).

Active site residue analysis

Activity and abundance scores were rescaled so that the lowest score present in the dataset was set at 0, and the highest score at 1. A ratio of rescaled activity to rescaled abundance (specific activity) was then calculated for every variant. Using variant specific activity scores, median specific activity was calculated for each position. Threshold for classification as an active site position was drawn based on scores of known redox cysteines at positions 132 and 135, resulting in lowest 12.5% of median specific activity scores being classified as active site residues. We additionally required that any position within this group had been scored for at least four variants to eliminate noise from poor sampling.

Data availability

All raw sequence data and function scores are freely available for all academic users by non-exclusive license under reasonable terms to commercial entities that have committed to open sharing of VKOR sequence variants and under a free non-exclusive license to non-profit entities. The Illumina raw sequencing files and barcode–variant maps can be accessed at the NCBI Gene Expression Omnibus (GEO) repository under accession number GSEXXXXXX. The data presented in the manuscript are available as Supplementary Data files.

Code availability

Code for analysis is available at http://github.com/FowlerLab/VKOR_DMS. The code used to train the evolutionary couplings model is available at the EVcouplings GitHub repository (https://github.com/debbiemarkslab/EVcouplings). The data used to train the model is publicly available at uniprot (https://uniprot.org).

3 Measuring CYP2C9 abundance in multiplex identifies conserved structural elements and determines human variant effect

3.1 Abstract

CYP2C9 is a pharmacogene responsible for ~15% of phase I drug metabolism in humans. Common variation in CYP2C9 has been curated and classified according to activity. These common variants can impact enzyme activity, but they are also known to affect CYP2C9 abundance. In addition to common variants, a deluge of rare variants from large-scale sequencing projects are being discovered. We have no functional annotations for these rare variants, which makes it difficult to tailor drug treatment to a patient's unique genotype and increases the risk of an adverse drug event. With this in mind, we measured the abundance of 6,370 missense variants of CYP2C9 using a sequencing-based, multiplex assay. From this data, we find that positions with the lowest 20% of median abundance scores are clustered in the hydrophobic internal lobes of the protein, and identify heme coordination as a likely modifier of protein abundance. In addition, we find that 42% of variants found in human genetic databases have decreased abundance, suggesting that patients who carry these variants may have poor metabolism status. Our data shows the promise of using protein abundance maps in an integrated pharmacogenomic approach to treat patients.

3.2 Introduction

Cytochrome P450s are hemoproteins that oxidize xenobiotic compounds, including pharmaceutical drugs. Humans have 57 putatively functional CYP450 genes, a small subset of which metabolize 70-80% of pharmaceutics[98]. Among these is CYP2C9, a member of the CYP2 family that metabolizes warfarin and flurbiprofen, among other drugs. CYP2C9 is highly

expressed, making up approximately 20% of P450s present in human microsomes[99] , and in phase I metabolism is responsible for ~15% of total drugs metabolized[36].

CYP2C9 is highly polymorphic, which can affect the protein's function and ability to metabolize drugs. Common CYP2C9 alleles (minor allele frequency > 5%) are assigned unique names based on the star (*) system[100]. The database PharmVar (https://www.pharmvar.org/) currently lists 61 star alleles, 57 of which are single missense variants. Some of these variants, like CYP2C9 R335W (*11) are known to have low abundance phenotypes relative to WT[101], suggesting that low CYP2C9 abundance may be one cause of poor metabolizer status.

While the star system is useful for curating common variation found in humans, more rare variants will be discovered as we sequence more genomes. How these variants affect CYP2C9 abundance and overall drug metabolism is unknown. Testing each rare variant as we encounter it is time-consuming and will not scale readily as we encounter more and more variants. To solve this problem, we performed a multiplex assay for variant abundance in CYP2C9. We measured abundance phenotypes for 6,370 of the possible 9,800 single amino acid CYP2C9 variants. We show a strong membrane protein signature for the N-terminal helix, further show that the heme center is important for stability, and classify existing human variants according to abundance, data that could inform drug choice and dose in the clinic.

3.3 Results

3.3.1 Measuring CYP2C9 abundance using VAMP-seq

We recently developed a method, VAMP-seq, that allows us to measure protein abundance using fluorescent proteins. We fused eGFP C-terminally to CYP2C9, and in the same

construct express mCherry from behind an internal ribosomal entry site (IRES) to control for expression. We then conducted a pilot experiment, in which we measured the ratio of eGFP to mCherry of both WT CYP2C9 and CYP2C9 R335W, a known destabilized variant (Figure 1a). Because CYP2C9 R335W is destabilized, the protein misfolds, and protein quality control degrades both CYP2C9 and the fused eGFP, leading to lower eGFP intensity. After validating that R335W indeed had lower eGFP signal, we constructed a barcoded, site-saturation mutagenesis library of CYP2C9, from positions 2 to 490. This library covers 8,113 of the 9,780 possible single amino acid variants (83%).

We expressed this library in cells using a lentiviral serine integrase landing pad[49]. Recombinant cells were selected using the small molecule AP1903, which causes inducible Caspase 9 in unrecombined landing pads to dimerize and activate. After selection, recombinant cells were assayed based on eGFP:mCherry ratio and sorted into four quartile bins (Fig. 3.1a). Bins were deeply sequenced, and data analysis leveraged variants frequency across bins, along with normalizing to nonsense and synonymous distributions, to assign each variant a score. Variant abundance scores showed distinct, separable distributions of synonymous and nonsense variants, with missense variants spanning the range of scores (Fig. 3.1b). After filtering, we assigned variant scores and classifications to 6,370 missense variants (Fig. 3.1c), and the three replicates correlated well (Pearson's $r = 0.79$, Spearman's $\rho = 0.75$).

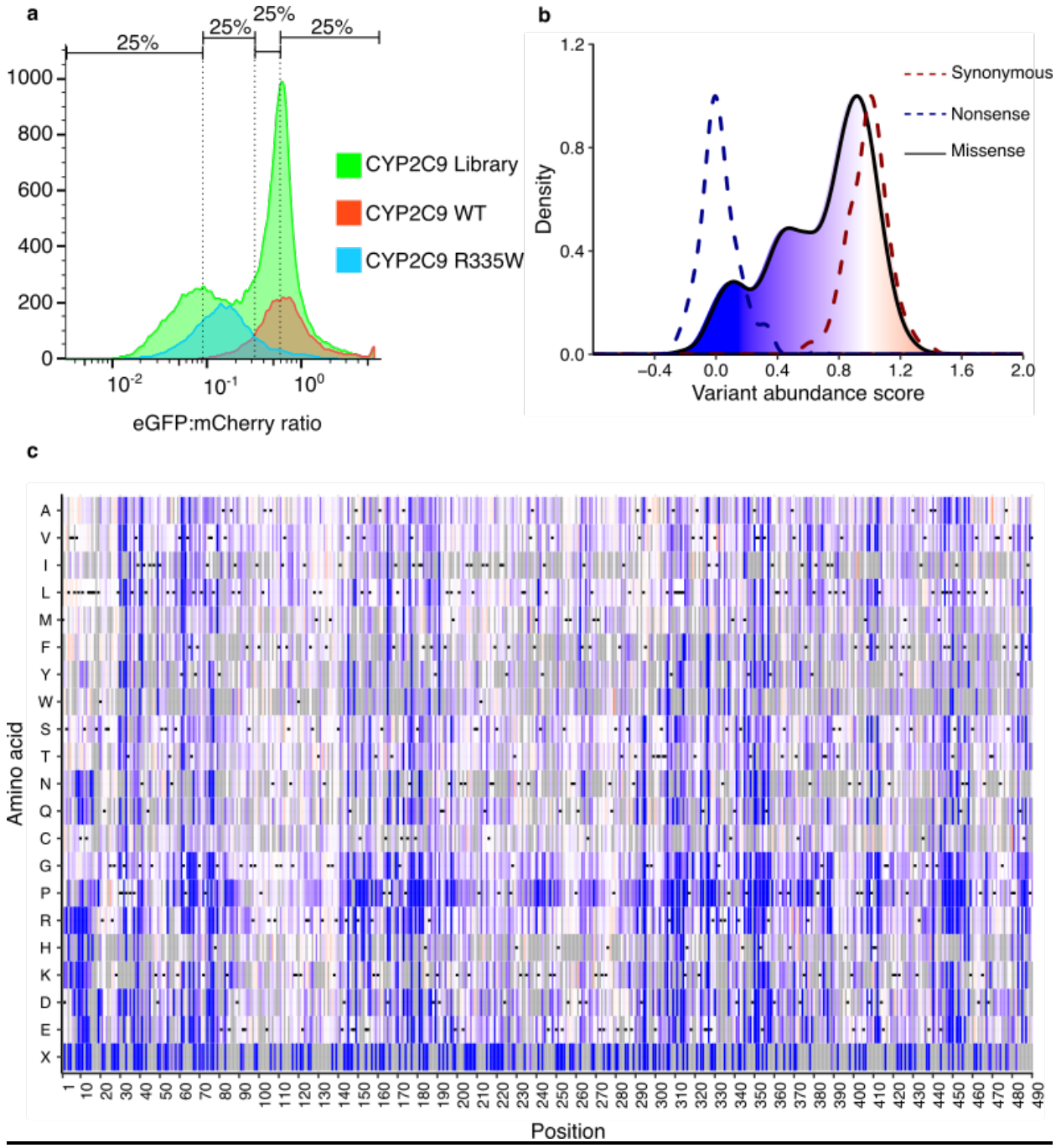


Figure 3.1. Multiplexed measurement of CYP2C9 abundance.

a Flow cytometry histogram showing WT CYP2C9 (red), CYP2C9 R335W (blue), and the library of CYP2C9 variants (green). Library was sorted into quartile bins (shown). **b** Density plot of variant abundance scores. Synonymous distribution shown as a red dashed line (n = 261 variants), nonsense shown as a blue dashed line (n = 189 variants), and missense variation in a

gradient (n = 6,370 variants). **c** Heatmap of abundance scores, with position in protein on x-axis and amino acid substitution on y. Colors correspond to the density plot in **b**.

3.3.2 Highly conserved regions of CYP2C9 show decreased abundance phenotypes

All eukaryotic CYP450s share the same highly conserved globular fold consisting of an N-terminal membrane-associated domain and a C-terminal catalytic domain[102]. CYP2C9 has been crystallized, both unliganded and with two different substrates, warfarin[103] and flurbiprofen[104]. While the structures do differ in terms of substrate-specific interactions, they largely agree in terms of structure, with 12 helices, labeled A through L, and four beta sheets, labeled β1 through β4.

We first mapped median positional abundance scores onto the CYP2C9 structure (Fig. 3.2a,b). We focused in particular on positions with the lowest 20% of abundance scores and found these positions cluster into two distinct regions: positions in and directly abutting β sheet-1 and core-facing positions in helices D, E, I, J, K, and L. Both of these regions are highly conserved across CYP450s and are comprised of buried, hydrophobic residues [105], substitution of which leads to destabilization and degradation. In addition, substitutions in beta-sheet 1 may disrupt distal sidechains that coordinate with the central heme iron[106].

We also found abundance signatures at smaller, highly conserved regions hypothesized to be important for protein stability: the cysteine that coordinates with the heme molecule and a stabilizing motif in the K helix which interacts with arginine and histidine in a motif known as the ERR-triad[105]. The hydrophobic active site is encased in the core of the protein, with a cysteine at position 435 penta-coordinating the heme molecule. Of nine substitutions observed at this position, six were classified as low or possibly low abundance. This suggests that coordination with heme is a crucial aspect of maintaining protein abundance in CYP2C9. The

ERR triad consists of E354 and R357 in the K helix and H411 in the following loop known as the meander region. The triad forms a network of salt bridges and hydrogen bonds N-terminally of the heme-binding region. We saw 39 of 40 missense mutations at these positions resulted in decreased abundance. The triad motif has been hypothesized to "lock" C435 into place for optimal heme coordination, so disruption of this structure likely disrupts heme coordination.

Next, we calculated a sliding window average of abundance scores across CYP2C9 for both aliphatic and proline substitutions (Figure 3.2c). We would expect proline substitutions to disrupt helices as proline has poor helix-forming propensity. Indeed, we see score minimas for almost all helices with proline compared to aliphatic substitutions, and we also see a minimum at the beginning of the protein, confirming the presence of an N-terminal transmembrane domain that inserts into the ER membrane. Interestingly, we see weak signal for helices A and C, and we also noticed that proline substitutions in some helices decreased abundance more significantly than others (e.g., helices F vs. K). The F-G loop of CYP2C9 is embedded in the membrane and is flexible, possibly forming a lid for the substrate-binding region. While proline substitutions in this region do not impact abundance significantly, they have been shown to decrease substrate affinity [106].

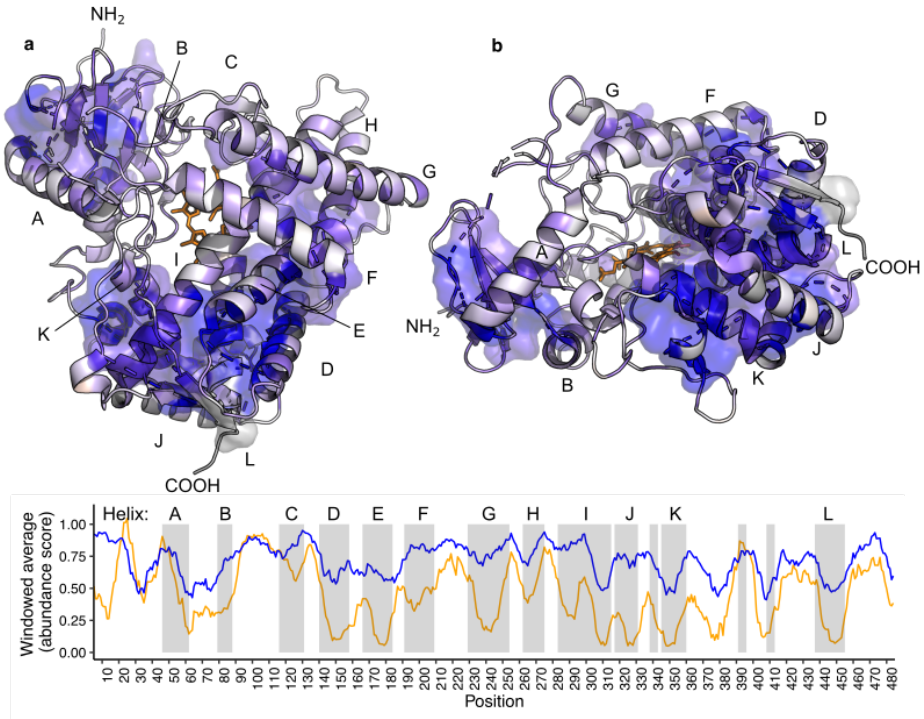


Figure 3.2. High-throughput mutagenesis reveals most impacted helices in CYP2C9 structure.

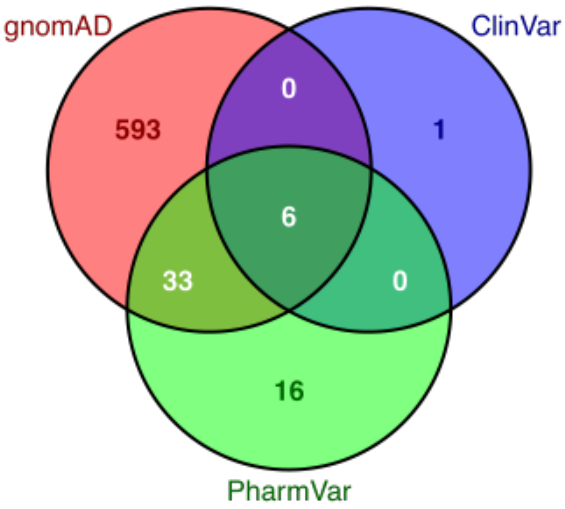
a, Crystal structure of CYP2C9 (PDB: 1R9O), positions with lowest 20% of median abundance scores shown as a surface in blue. Chains are colored according to positional median abundance scores using a gradient between the lowest 10% of positional median abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). **b,** Alternate view of crystal structure. **c**, Windowed average of CYP2C9 abundance score (window length = 10 amino acids, center alignment), colored by amino acid substitution class. Aliphatic substitutions shown in blue, proline substitutions in orange. Helices within CYP2C9 are shown in grey boxes and annotated.

3.3.3 Human variants show range of abundance phenotypes

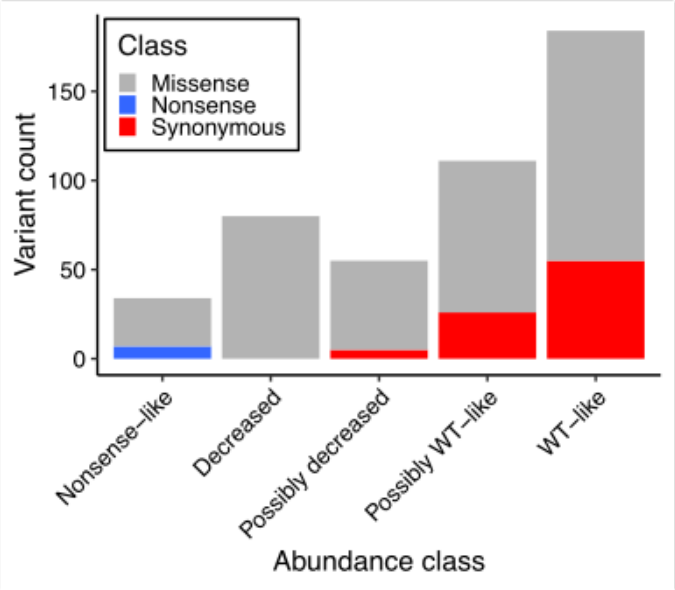
To classify human variants based on abundance phenotypes, we curated data from PharmVar[107,108], gnomAD[26], and ClinVar[109]. In total, we curated 649 variants, the majority of which come from gnomAD, where variants are not functionally annotated (Fig. 3.3a). Three nonsense variants from these databases were classified as nonsense-like abundance, while the majority of synonymous variants were classified as possibly WT-like or WT-like (Fig. 3.3b). We classified 371 missense variants and found that 157 variants (42%) were classified as decreased or possibly decreased abundance.

The Clinical Pharmacogenetics Implementation Consortium (CPIC) reviews evidence for CYP2C9 variants and based on that evidence, functionally annotates each variant as "no function", "decreased function", and "unknown function"[100]. We compared our abundance data to these classifications. First, we see enrichment of no function variants in low abundance classification classes, with 6 of 10 variants being classified as nonsense-like, decreased or possibly decreased abundance (Fig 3.3c). For unknown function variants, we classified 11 of 24 (46%) as low or possibly low abundance, suggesting that low abundance is a common route for decreased CYP2C9 metabolism.

a


gnomAD
ClinVar
PharmVar
593
0
1
6
33
0
16

b


Class
Missense
Nonsense
Synonymous
Variant count
150
100
50
0
Nonsense−like
Decreased
Possibly decreased
Possibly WT−like
WT−like
Abundance class

c

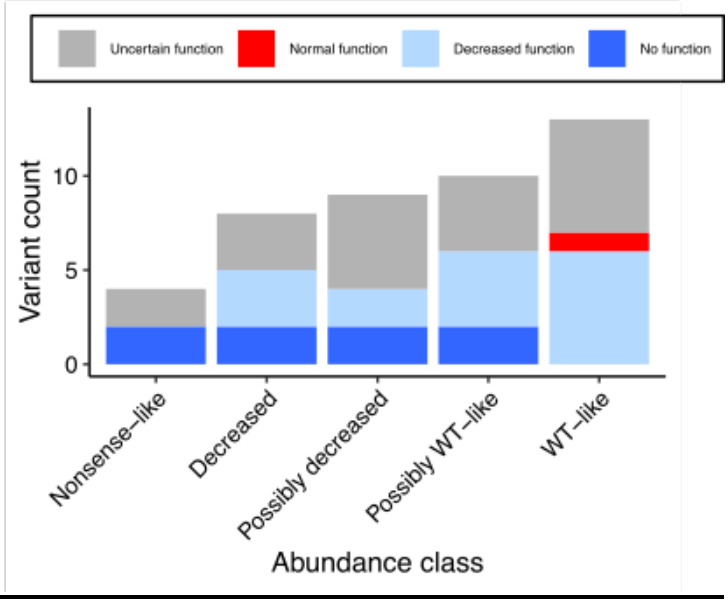

Uncertain function
Normal function
Decreased function
No function
Variant count
10
5
0
Nonsense−like
Decreased
Possibly decreased
Possibly WT−like
WT−like
Abundance class

Figure 3.3. Abundance classifications of human CYP2C9 variants.

a Venn diagram showing where CYP2C9 variants are catalogued across three databases: gnomAD, PharmVar, and ClinVar. **b** Bar plot of variants classified by abundance class and colored by class of variant: nonsense (blue), synonymous (red), and missense (grey). **c** Bar plot of variants classified by abundance class and colored by class of CPIC classification: uncertain function (grey), decreased function (light blue), and no function (dark blue).

3.4 Discussion

We have shown that our method VAMP-seq can be applied to CYP2C9 to measure protein abundance of a library of variants. We scored 6,370 missense variants for abundance, and the missense distribution of scores spanned nonsense and synonymous variants. We found that positions with the lowest 20% of median abundance scores were found in hydrophobic core regions of the protein, both in the N-terminal membrane association domain and in the C-terminal catalytic domain. Zooming in more closely, we suggest a pivotal role for C435 in maintaining protein abundance, presumably through its role coordinating the heme molecule in the central cavity of the holoprotein. In addition, we show that a highly conserved folding motif, the ERR triad, shows pervasive low abundance, suggesting its role in positioning C435 is crucial for heme coordination and ultimately protein abundance.

While we apply the data to better understand CYP2C9 biology, the data has limitations in terms of what it can detect. First, CYP2C9 interacts with many other proteins *in vivo*: cytochrome P450 reductase[110], other CYPs[111], and may form dimers with other CYP2C9 molecules[112]. However, our data will be unable to detect any disruption of protein-protein interactions that does not also impact abundance. Second, we express CYP2C9 as an open

reading frame, so we will not detect effects of any variant that affects splicing or mRNA stability[113]. Finally, the data is noisy, most likely due to the bottleneck of cells that can be sorted in a reasonable time frame. We combat this by calculating standard error and a 95% confidence interval for each variant.

This work could be extended in multiple ways to better understand CYP2C9 biology and pharmacogenomics. First, this method could be used with known CYP2C9 substrates. If a subset stabilizes an otherwise unstable variant of CYP2C9, this could help us map structures of CYP2C9 in complex with the substrate of interest. Comparison of these results across substrates will help reveal positions that are key in substrate-specific channeling and binding. Next, we can use activity-based protein probes[114] to measure CYP2C9 activity in a similar high-throughput manner. These probes mimic natural substrates of CYP2C9, so would yield not only substrate-specific insights, but also would allow us to generate an abundance-activity map of the protein. Finally, these results combined with electronic health record analysis can be used to build predictive models of how patients' unique genotypes may affect drug metabolism for particular drugs.

3.5 Methods

General reagents, DNA oligonucleotides, and plasmids

Unless otherwise noted, all chemicals were obtained from Sigma and all enzymes were obtained from New England Biolabs. *E. coli* were cultured at 37°C in Luria broth. All cell culture reagents were purchased from ThermoFisher Scientific unless otherwise noted. HEK 293T cells (ATCC CRL-3216) and derivatives thereof were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U ml^{-1} penicillin, and 0.1 mg ml^{-1} streptomycin.

Cells were induced with 2.5 ug mL^{-1} doxycycline. Cells were passaged by detachment with trypsin-EDTA 0.25%, and cells were prepared for sorting by detachment with versene. All cell lines tested negative for mycoplasma.

All synthetic oligonucleotides were obtained from IDT and can be found in Supplementary Table 6. All non-library-related plasmid modifications were performed with Gibson assembly[95].

Library construction

A gBLOCK with an optimized sequence for human CYP2C9 was ordered from IDT. It was then cloned into the vector pHSG298 (Clontech). Saturation mutagenesis primers were designed for each codon in CYP2C9 from positions 2 to 490 and ordered resuspended from IDT. Forward and reverse primers for each position were mixed at 2.5 mM, and used in a PCR reaction with 125 pg of pHSG298-CYP2C9, 5% DMSO, and 5 uL of KAPA Hifi Hotstart 2X ReadyMix. PCR products were visualized on a 0.7% agarose gel to confirm amplification of the correct product.

PCR products were then quantified using the Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) using DNA control curves done in triplicate. To pool, a total amount of DNA for each reaction was calculated that maximized the volume to be drawn from the lowest concentration PCR product. Pooled PCR products were cleaned and concentrated using Zymogen Clean and Concentrate kit and then gel extracted. The pooled library was phosphorylated with T4 PNK (NEB), incubated at 37°C for 30 minutes, 65°C for 20 minutes, and then 4° indefinitely. 8.5 uL of this phosphorylated product was combined with 1 uL of 10X T4 ligase buffer (NEB) and 0.5 uL of T4 DNA ligase (NEB) to make a 10 uL overnight ligation reaction. This reaction was incubated at 16°C overnight.

The overnight ligation was then cleaned and concentrated (Zymogen) and eluted in 6 uL of ddH2O. 1 uL of this ligation was then transformed into high efficiency E. coli (NEB C3020K) using electroporation (settings: 2 kV). Each reaction contained 1 uL of ligation (or ligation control or pUC19 10 pg./uL) and 25 uL of E. coli. 975 uL of pre-warmed SOC media was added to each cuvette after electroporation, transferred to a culture tube, and recovered at 37°C, shaking for 1 hour. At 1 hour, 1 and 10 uL samples from all cultures were taken and plated on appropriate media (LB + kanamycin for ligation and ligation control; LB + ampicillin for pUC19), the remaining 989 uL was used to inoculate a 50 mL culture (+ kanamycin). Plates and 50 mL culture were incubated at 37°C overnight (shaking for 50 mL culture). Colonies on plates were then counted, and counts were used to calculate how many unique molecules were transformed to gauge coverage of the library. 50 mL culture was spun down and midiprepped.

To transfer the library from pKan to the recombination vector, the pKan library and recombination vector were digested with MluI and SphI for 1 hour at 65°C. The library and cut vector were then gel extracted. The library was then ligated with the cut vector at 5:1 using NEB T4 ligase, overnight at 16°C. The ligation was heat inactivated the next morning, clean and concentrated with the Zymo kit. Another high efficiency transformation was performed the same as described above, except this ligation was plated on LB + ampicillin (antibiotic switching strategy). Plates and 50 mL culture were incubated at 37°C overnight (shaking for 50 mL culture). Colonies on plates were then counted, and counts were used to calculate how many unique molecules were transformed to gauge coverage of the library. 50 mL culture was spun down and midiprepped.

To barcode individual variants, plasmid library harvested from midiprep was digested with AgeI-HF at 37°C for 1 hour, 65°C for 20 minutes. Barcode oligos were ordered from IDT, resuspended at 100 uM, and then annealed by combining 1 uL each of primer with 4 uL CutSmart Buffer and 34 uL ddH2O and running at 98°C for 3 minutes followed by ramping down to 25°C at -0.1°C/second. After annealing, 0.8 uL of Klenow polymerase (exonuclease negative, NEB) and 1.35 uL of 1 mM dNTPS was then combined with the 40 uL of product to fill in the barcode oligo (cycling conditions: 25°C for 15:00, 70°C for 20:00, ramp down to 37°C at -0.1°C/s). Digested vector and barcode oligo were then ligated overnight at 16°C.

The overnight ligation was then cleaned and concentrated and eluted in 6 uL of ddH2O. 1 uL of this ligation was then transformed into high efficiency *E. coli* using electroporation at 2 kV. Each reaction contained 1 uL of ligation (or ligation control or pUC19 10 pg/uL) and 25 uL of *E. coli*. 975 uL of pre-warmed SOC media was added to each cuvette after electroporation, transferred to a culture tube, and recovered at 37°C, shaking for 1 hour. At 1 hour, 1 and 10 uL samples from water and pUC19 cultures were taken and plated on LB supplemented with ampicillin. For ligation and ligation control, four flasks were prepared with 50 mLs of LB and ampicillin, and then 500 uL, 250 uL, 125 uL, 62.5 uL was sample from the 1 mL of recovery and transferred into a corresponding flask. From those flasks, 1 uL, 10 uL, and 100 uL, were sampled and plated onto LB ampicillin plates. Plates and 50 mL culture were incubated at 37°C overnight. Colonies on plates were then counted, and counts were used to calculate how many unique molecules were transformed to gauge number of barcodes. Flask with the target number of barcodes was then spun down and midiprepped.

PacBio Subassembly

To associate barcode with variants, we used PacBio SMRT sequencing to capture both the open reading frame and barcode. Barcoded plasmid library was digested with NheI and SmaI for one hour at 37°C. Ampure beads PB were brought to room temperature. 30 uL of Ampure beads were added to each 50 uL restriction digest reaction. DNA fragments were then allowed to bind to beads for 5 minutes at room temperature. Tubes were then left on magnet for two minutes, and supernatant was then aspirated to remove contaminants. Beads with DNA fragments attached were then washed twice with 50 uL of 70% ethanol. Ethanol was removed after washing and bead pellet was allowed to dry. After removing any residual ethanol, DNA fragments were eluted in 30 uL of PacBio elution buffer.

To blunt the ends of the restriction digest fragments, 500 ng of DNA was combined with 5 uL of 10X DNA Damage Repair Buffer, 0.5 uL of 100X NAD+, 5 uL of 10mM ATP, 0.5 uL of 10 mM dNTPs, 2 uL DNA Damage Repair Mix. Water was added to achieve final volume of 50 uL. This mixture was then incubated at 37°C for 20 minutes, then 4°C for 1 min. At this point, 2.5 uL End Repair Mix was added to the reaction mixture and incubated at 25°C for 5 minutes, then 4°C. The reaction mixture was then cleaned with Ampure beads PB as described above and eluted in 30 uL of PacBio elution buffer.

To append SMRTbells with blunt end ligation, 30 uL of end repaired DNA was mixed with 1 uL 20 mM Blunt End Adaptor, 4 uL of 10X Template Prep Buffer, 2 uL of 1 mM ATP, and 1 uL Ligase. Water was added to achieve final volume of 40 uL. This reaction was incubated at 25°C for 15 minutes, and then inactivated at 65°C for 10 minutes. 10 uL of CutSmart buffer was added to each reaction, the total volume of 50 uL was then split into two reactions of 25 uL each. 10 uL

of CutSmart buffer and 1 uL of BamHI-HF was added to each of these reactions, and then they were incubated at 37°C for 1 hour. After this digestion, 1 uL of ExoIII (100U/uL) and 1 uL of Exo VII (10 U/uL) were added to the reaction, and incubated at 37°C for 1 hour, then 4°C until purification. Ampure bead purification was carried out as described above, and eluted in 10 uL of elution buffer. Samples were then submitted to University of Washington PacBio Sequencing Services and sequenced on two SMRT cells in a Sequel run.

Cell line description

VAMP-seq assay cell line

HEK293T cells with a serine integrase landing pad integrated via lentivirus with a selectable inducible Caspase 9 cassette[49] were used for all experiments.

Recombination of variants in cell line

Cells were transfected in 10 cm plates, 3,500,000 cells per plate (4 plates per replicate). 7.1 ug of library plasmid was mixed with 0.48 ug of Bxb1 plasmid in 710 uL of OptiMEM. In a separate tube, 28.5 uL of Fugene was diluted in 685 uL of OptiMEM. The tubes were then combined and incubated at room temperature for 15 minutes. After incubation period, Fugene/DNA mixture was added to cells dropwise, and plates were placed in incubator at 37°C. A minimum of 48 hours after transfection, cells were induced with doxycycline at a final concentration of 2.5 ug/mL. 24 hours after induction with doxycycline, small molecule AP1903 was added to select from recombinant cells.

VAMP-Seq quartile sorting

Recombined cells were run on a BD Aria sorter. Cells were then gated for live, recombined singlets. For this population, a ratio of eGFP/mCherry was calculated, and the histogram of this ratio was divided into four quartiles. Each quartile was sorted into a 5 mL tube. Sorted cells were grown out for 2-4 days post sorting to ensure enough DNA for sequencing.

gDNA prep, barcode amplification, and sequencing

Cells were then collected, pelleted by centrifugation and stored at −20 °C. Genomic DNA was prepared using a DNEasy kit, according to the manufacturer's instructions (Qiagen), with the addition of a 30 min incubation at 37 °C with RNAse in the re-suspension step. Eight 50 µl first-round PCR reactions were each prepared with a final concentration of ~50 ng µl−1 input genomic DNA, 1 × Q5 High-Fidelity Master Mix and 0.25 µM of the KAM499/VKORampR 1.1 primers. The reaction conditions were 98 °C for 30 s, 98°C for 10 s, 65°C for 20 s, 72°C for 60 s, repeat 5 times, 72°C for 2 min, 4°C hold. Eight 50 µl reactions were combined, bound to AMPure XP (Beckman Coulter), cleaned and eluted with 21 µl water. Forty percent of the eluted volume was mixed with Q5 High-Fidelity Master Mix; VKOR_indexF_1.1 and one of the indexed reverse primers, PTEN_seq_R1a through PTEN_seq_R2a, were added at 0.25 µM each. These reactions were run with Sybr Green I on a BioRad MiniOpticon; reactions were denatured for 3 minutes at 95°C and cycled 20 times at 95°C for 15s, 60°C for 15s, 72°C for 15s with a final 3 min extension at 72°C. The indexed amplicons were mixed based in relative fluorescence units and run on a 1% agarose gel with Sybr Safe and gel extracted using a freeze and squeeze column (Bio-Rad). The product was quantified using Kapa Illumina quant kit.

Barcode counting and variant calling

Enrich2 was used to quantify barcodes from bin sequencing, using a minimum quality filter of 20^{96}. FASTQ files containing barcodes and the barcode map for VKOR were used as input for Enrich2. Enrich2 configuration files for each experiment are available on the GitHub repository. Barcodes assigned to variants containing inseriton, deletions, or multiple amino-acid alterations were removed from the analysis.

Calculating scores and classifications

For each protein variant, frequencies in each bin were calculated by dividing counts by total counts. These frequencies were then each weighted by multiplying by 0.25, 0.5, 0.75, and 1 in a bin-wise fashion.

4 The path forward for multiplex assays in pharmacogenomics

4.1 Combining MAVE data with pharmacogenetic and clinical data: another level of evidence

While the data generated by multiplex assays is powerful in and of itself, comparison to CPIC classifications and clinical data yields further insight. For example, in chapter 3, of CYP2C9 variants classified by CPIC as "No function"[25] and scored in our assay, six of eight (75%) had decreased abundance. In comparison, for those variants classified by CPIC as "decreased function" and scored by our assay, only five of 16 (31%) had decreased abundance. Our findings for "No function" variants matches estimates that 75% of pathogenic variants in monogenic disease disrupt thermostability and ultimately alter abundance[115,116], and also mirror findings for *TPMT*, another pharmacogene in which three high-frequency, non-functional variants were found to be low abundance [53].

I envision MAVE data will serve as another data source that organizations such as CPIC can use to make decisions about functional classification and levels of evidence (https://cpicpgx.org/levels-of-evidence/). CPIC currently uses three levels, weak, moderate, and high, to grade evidence based on the design and statistical power of studies, and evidence is also designated by type, in vitro or clinical. More explicit definitions for evidence, like those used by the American College of Medical Genetics to classify pathogenic and benign variants[117], may prove useful in the future to aid pharmacogenetic interpretation and implementation[118,119]. MAVE results would therefore serve as another source of in vitro data and identify variants that could be prioritized for more in-depth studies to further develop guidelines.

To further apply MAVE data, I anticipate coordinating with the Electronic Medical Records and Genomics Network (eMERGE, https://emerge-network.org/) to compare our results with patient drug metabolism phenotypes. These data would include warfarin dose and response

and whole genome sequencing, so we would be able to see, for example, if low abundance variants in patients translates to a lower stable warfarin dose.

4.2 Engineering cell assays for an exhaustive pharmacogene atlas

In chapters 2 and 3, I measured VKOR abundance and activity and CYP2C9 abundance, respectively. While these are informative datasets on their own, having datasets for multiple protein phenotypes would yield rich protein atlas from which we could get a more complete idea of how these proteins fold, localize, function, respond to drugs, and interact with other proteins. Technologies to interrogate some of these phenotypes will need to be developed or further refined to be protein-specific, and developing engineering pipelines that allow for modular construction and quick clonal expansion are key to making swift progress.

As a case in point, I attempted to measure warfarin response for VKOR variants. Rare variants in *VKORC1* drive warfarin resistance phenotypes outside the normal dosing range[13], so I sought to develop a cell assay that could measure warfarin resistance phenotypes. However, while conducting these experiments, I observed after engineering this assay that the reporter protein was still being carboxylated even under high warfarin treatment concentrations. In follow-up experiments, I found that the high expression of *VKORC1* off the tetracycline-inducible promoter of the landing pad eliminated the reporter protein's response to warfarin. I fixed this problem in monocultures by using a less efficient Kozak sequence[120]. Unfortunately, when I transitioned to library experiments with the less efficient Kozak, the warfarin treatment data was noisy and did not exhibit a wide dynamic range.

While ultimately a failure, this experience was useful because it highlights the challenges in developing multiplex assays for pharmacogenetic applications. Moving forward, the best

practices should include designing direct readouts of protein phenotypes when possible. If the protein phenotype can be genetically encoded, as opposed to antibody-based, that is often a more robust strategy, as antibodies can vary in quality and the antibody staining process stresses and kills cells. Noise in these experiments were caused by low sampling numbers of cells, caused by the bottleneck of cell sorting. Development of alternative assays, including growth-based assays[49], would not be constrained by flow sorting and would most likely yield higher precision data. Finally, sorting longer for maximal recovery of cells, as opposed to many short sorts, may also result in more consistent, precise data.

4.3 Combinatorial scans to assay more complex pharmacogenomic interactions

While the experiments I executed looked at VKOR and CYP2C9 singly, the reality is that a patient's warfarin response will largely be determined by the genotypes of both *VKORC1* and *CYP2C9*, along with other environmental covariates (e.g., vitamin K consumption). 18% of people carry common variants in both *VKORC1* and *CYP2C9*[121]. These common variants are non-coding SNPs that were discovered in GWAS studies and identify haplotypes with characteristic levels of expression for the gene of interest[13,122]. Combinations of rare and common variants in *VKORC1* and *CYP2C9* exist in patients, including a case report of an individual with both a *VKORC1* D36Y variant and a *CYP2C9**11 variant[123]. Therefore, either finding cell lines with these haplotypes already present or tuning expression to mimic what we find in humans are two ways to approach this problem. A natural extension of my work would be to test single missense variants on a background of these common variants to better understand how these variants would interact to produce a cell-level phenotype.

4.4 Interrogating non-coding regulation of pharmacogenes

In addition to coding variation, understanding how variation in non-coding regions of the genome contributes to warfarin response is important. 5' UTR and intronic SNPs at the *VKORC1* locus are associated with warfarin sensitivity[13]. These SNPs are correlated with both reduced *VKORC1* mRNA [13]and VKOR protein abundance[69]. While the 5' UTR SNP is not causal, studies have shown that the alternative allele creates a suppressor E-box site, presumably allowing repressive transcription factors to bind and inhibit transcription[124].

As it stands, our understanding of the non-coding genome is incomplete, but recent advances in technology allow us to perturb regulatory regions in a massively parallel fashion. Early MAVE approaches, like STARR-seq[125], used a reporter plasmid, into which one could then clone a putative promoter or enhancer.. Expression level was captured through an identifying sequence on the reporter gene transcript and normalized to the DNA-level representation. However, these approaches test putative regulatory DNA outside of their endogenous context in the genome, which can lead to conflicting results[126].

More recently, MAVEs that detect regulatory changes in their native context have been developed, like using tiled CRISPR-Cas9 deletions across an endogenous locus to identifiy cis-regulatory regions[52], or multiplexing CRISPR-Cas9 transcriptional repressors to identify trans-regulatory regions[54]. To apply these technologies to *VKORC1*, we could use a VKOR activity cell line with *VKORC1* intact and use paired gRNAs and Cas9 to delete putative cis-regulatory regions around the *VKORC1* locus. Cells would be sorted based on reporter staining, and gRNAs that were enriched in cells that did not have WT activity would define candidate regulatory regions. To detect trans-interactions, we could introduce a library of gRNAs along with KRAB domain-fused, catalytically dead, Cas9. Those gRNAs combinations that were shown to repress *VKORC1* expression would identify putative regulatory regions. To move beyond VKOR to

other pharmacogenes, we should prioritize functional studies of regulatory DNA based on what we know about the regulatory landscape of each pharmacogene, variation that exists in putative regulatory regions, and cell assays that would allow us to read out the effect of regulatory changes.

4.5 Single cell RNA-sequencing will allow more precise measurement of transcriptional variation

Understanding the transcriptional landscape of pharmacogenes would illuminate drug response. RNA-sequencing of individuals showed variation of expression both across tissues and across individuals for 389 pharmacogenes[127]. In addition to expression level, this study also observed differential alternative splicing patterns, which can influence drug response. For example, alternative splicing of *CYP2D6* can lower enzyme activity[128].

While bulk RNA-sequencing gives us a population-level measure of RNA abundance and alternative splicing, these studies cannot detect cellular sub-populations with distinct transcriptional signatures which may inform pharmacogenetic biology. In addition, these studies were also not designed to measure how rare variation in pharmacogenes impacts transcriptional response. In that vein, developing cell assays where we can introduce a variant and then measure transcriptomes in single cells would allow us to directly connect these phenotypes. With the advent of single cell RNA-sequencing technologies[129,130], it is possible to assay hundreds of thousands of cells, which would allow for sufficient coverage of variant libraries.

4.6 Single molecule, real-time sequencing (SMRT) can resolve complex variants in pharmacogenes

While the focus of my work has been on single amino acid variants in pharmacogenes, more complex variation exists in pharmacogenes. For example, *CYP2D6*, which is the pharmacogene associated with the most drugs recommended for pharmacogenetic-guided drug dosing, has over 100 variants curated by CPIC[108], the majority of which are copy number variants, splicing variants, small insertions and deletions, and larger scale genome rearrangements, like gene hybrid fusions with a neighboring non-functional pseudogene, *CYP2D7*. While Illumina technology is used widely to sequence genomes, its short read lengths make it difficult to resolve copy number and hybrid gene rearrangements. Algorithms like Stargazer for *CYP2D6*[131] can be used to interpolate structural variation from Illumina data, but finding a more general method to detect complex genome architecture for all pharmacogenes is critical.

Fortunately, an orthogonal sequencing technology, single molecule, real-time sequencing, allows iterative sequencing of a single DNA molecule using a processive polymerase and nanopore technology[132,133]. SMRT sequencing is therefore able to sequence much larger amplicons multiple times, and can detect diverse classes of variation, including single nucleotide variants, small indels (<50 bp), and structural variation (>50 bp). Efforts to integrate SMRT sequencing into genomic medicine workflows has begun[134] and will better catalog the complex variation present in pharmacogenes. To functionally test the complex variants we uncover, with a focus on CYP2D6, I would use a CRISPR-Cas9 approach. I would program a library of HDR templates with observed variants from SMRT sequencing along with gRNAs targeting the *CYP2D6* locus. For the functional assay, I can then use click chemistry probes to label active

CYP2D6 in whole cells[114]. Cells can then be sorted based on fluorescence intensity, and deeply sequenced to determine variant-level activities.

4.7 The promise of pharmacogenomic multiplex assays

Pharmacogenomics is a pillar of personalized medicine. The strides I made in my PhD to illuminate variants that may influence drug dosing will hopefully motivate others to engineer multiplex assays for the many pharmacogenes that remain. This field is such an interesting blend of biology and technology development, and it's an exciting time to be working on problems that have direct application to the clinic. Together we can build towards a pharmacogenomic functional atlas, uncovering biological insight and improving patients' lives along the way.

References

1. Nebert, D. W. Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? *Clin. Genet.* **56**, 247–258 (1999).
2. Motulsky AG. DRug reactions, enzymes, and biochemical genetics. *JAMA* **165**, 835–837 (1957).
3. Motulsky, A. G. PHARMACOGENETICS. *Prog. Med. Genet.* **23**, 49–74 (1964).
4. Motulsky, A. G. The Genetics of Abnormal Drug Responses. *Ann. N. Y. Acad. Sci.* **123**, 167–177 (1965).
5. Vesell, E. S. & Page, J. G. Genetic control of drug levels in man: antipyrine. *Science* **161**, 72–73 (1968).
6. Tucker, G. T., Silas, J. H., Iyun, A. O., Lennard, M. S. & Smith, A. J. Polymorphic hydroxylation of debrisoquine. *Lancet* **2**, 718 (1977).
7. Eichelbaum, M., Spannbrucker, N., Steincke, B. & Dengler, H. J. Defective N-oxidation of sparteine in man: a new pharmacogenetic defect. *Eur. J. Clin. Pharmacol.* **16**, 183–187 (1979).
8. Gonzalez, F. J. *et al.* Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* **331**, 442–446 (1988).
9. Mallal, S. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
10. Hetherington, S. *et al.* Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *Lancet* **359**, 1121–1122 (2002).
11. Martin, M. A. *et al.* Clinical pharmacogenetics implementation consortium guidelines for

HLA-B genotype and abacavir dosing. *Clin. Pharmacol. Ther.* **91**, 734–738 (2012).

12. Mallal, S. *et al.* HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**, 568–579 (2008).
13. Rieder, M. J. *et al.* Effect of VKORC1 Haplotypes on Transcriptional Regulation and Warfarin Dose. *N. Engl. J. Med.* **352**, 2285–2293 (2005).
14. Pirmohamed, M. *et al.* A Randomized Trial of Genotype-Guided Dosing of Warfarin. *N. Engl. J. Med.* **369**, 2294–2303 (2013).
15. Rost, S. *et al.* Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature* **427**, 537–541 (2004).
16. Loebstein, R. *et al.* A coding VKORC1 Asp36Tyr polymorphism predisposes to warfarin resistance. *Blood* **109**, 2477–2480 (2007).
17. Zimmermann, A. & Matschiner, J. T. Biochemical basis of hereditary resistance to warfarin in the rat. *Biochem. Pharmacol.* **23**, 1033–1040 (1974).
18. Pelz, H.-J. *et al.* The Genetic Basis of Resistance to Anticoagulants in Rodents. *Genetics* **170**, 1839–1847 (2005).
19. Oldenburg, J., Müller, C. R., Rost, S., Watzka, M. & Bevans, C. G. Comparative genetics of warfarin resistance: *Hämostaseologie* **34**, 143–159 (2013).
20. Watzka, M. *et al.* Thirteen novel VKORC1 mutations associated with oral anticoagulant resistance: insights into improved patient diagnosis and treatment. *J. Thromb. Haemost.* **9**, 109–118 (2011).
21. Gordon, A. S. *et al.* Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum. Mol. Genet.* **23**, 1957–1963 (2014).

22. Relling, M. V. & Klein, T. E. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin. Pharmacol. Ther.* **89**, 464–467 (2011).

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

24. Starita, L. M. *et al.* Variant Interpretation: Functional Assays to the Rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

25. Johnson, J. A. *et al.* Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin Dosing: 2017 Update. *Clin. Pharmacol. Ther.* **102**, 397–404 (2017).

26. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019) doi:10.1101/531210.

27. Niinuma, Y. *et al.* Functional characterization of 32 CYP2C9 allelic variants. *Pharmacogenomics J.* **14**, 107–114 (2014).

28. Hiratsuka, M. Genetic Polymorphisms and *in Vitro* Functional Characterization of CYP2C8, CYP2C9, and CYP2C19 Allelic Variants. *Biol. Pharm. Bull.* **39**, 1748–1759 (2016).

29. 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).

30. Tie, J.-K., Jin, D.-Y., Tie, K. & Stafford, D. W. Evaluation of warfarin resistance using transcription activator-like effector nucleases-mediated vitamin K epoxide reductase knockout HEK293 cells. *J. Thromb. Haemost.* **11**, 1556–1564 (2013).

31. Shen, G. *et al.* Warfarin traps human vitamin K epoxide reductase in an intermediate state during electron transfer. *Nat. Struct. Mol. Biol.* **24**, 69 (2017).
32. Li, T. *et al.* Identification of the gene for vitamin K epoxide reductase. *Nature* **427**, 541–544 (2004).
33. Matagrin, B. *et al.* New insights into the catalytic mechanism of vitamin K epoxide reductase (VKORC1) – The catalytic properties of the major mutations of rVKORC1 explain the biological cost associated to mutations. *FEBS Open Bio* **3**, 144–150 (2013).
34. Rost, S. *et al.* Site-directed mutagenesis of coumarin-type anticoagulant-sensitive VKORCI. Evidence that highly conserved amino acids define structural requirements for enzymatic activity and inhibition by warfarin. *Thromb. Haemost.* (2005) doi:10.1160/TH05-02-0082.
35. Nair, P. C., McKinnon, R. A. & Miners, J. O. Cytochrome P450 structure-function: insights from molecular dynamics simulations. *Drug Metab. Rev.* **48**, 434–452 (2016).
36. Rettie, A. E. & Jones, J. P. Clinical and toxicological relevance of CYP2C9: drug-drug interactions and pharmacogenetics. *Annu. Rev. Pharmacol. Toxicol.* **45**, 477–494 (2005).
37. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
38. Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum. Genet.* **137**, 665–678 (2018).
39. Cao, J. *et al.* High-Throughput 5' UTR Engineering for Enhanced Protein Production in Non-Viral Gene Therapies. *bioRxiv* 2020.03.24.006486 (2020) doi:10.1101/2020.03.24.006486.
40. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

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

42. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **advance online publication**, (2014).

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

44. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E1263–E1272 (2013).

45. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* genetics.115.175802 (2015).

46. Jain, P. C. & Varadarajan, R. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal. Biochem.* **449**, 90–98 (2014).

47. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42**, e112–e112 (2014).

48. Matreyek, K. A., Stephany, J. J. & Fowler, D. M. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* doi:10.1093/nar/gkx183.

49. 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.* (2019) doi:10.1093/nar/gkz910.

50. Kotler, E. *et al.* A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. *Mol. Cell* **71**, 178–190.e8 (2018).

51. Suiter, C. C. *et al.* Massive parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. *bioRxiv* 740837 (2019) doi:10.1101/740837.

52. Gasperini, M. *et al.* CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017).

53. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* **50**, 874–882 (2018).

54. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377–390.e19 (2019).

55. Adams, R. M., Mora, T., Walczak, A. M. & Kinney, J. B. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife* **5**, (2016).

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

57. Relling, M. V. *et al.* Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin. Pharmacol. Ther.* **93**, 324–325 (2013).

58. Liu, C. *et al.* Genomewide Approach Validates Thiopurine Methyltransferase Activity Is a Monogenic Pharmacogenomic Trait. *Clin. Pharmacol. Ther.* **101**, 373–381 (2017).

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

60. Czogalla, K. J. *et al.* Warfarin and vitamin K compete for binding to Phe55 in human VKOR. *Nat. Struct. Mol. Biol.* (2016) doi:10.1038/nsmb.3338.

61. Owen, R. P., Gong, L., Sagreiya, H., Klein, T. E. & Altman, R. B. VKORC1 Pharmacogenomics Summary. *Pharmacogenet. Genomics* **20**, 642–644 (2010).

62. Osinbowale, O., Al Malki, M., Schade, A. & Bartholomew, J. R. An algorithm for managing warfarin resistance. *Cleve. Clin. J. Med.* **76**, 724–730 (2009).

63. Yuan, H.-Y. *et al.* A novel functional VKORC1 promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. *Hum. Mol. Genet.* **14**, 1745–1751 (2005).

64. Li, W. *et al.* Structure of a bacterial homologue of vitamin K epoxide reductase. *Nature* **463**, 507–512 (2010).

65. Schulman, S., Wang, B., Li, W. & Rapoport, T. A. Vitamin K epoxide reductase prefers ER membrane-anchored thioredoxin-like redox partners. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15027–15032 (2010).

66. Tie, J.-K., Jin, D.-Y. & Stafford, D. W. Human Vitamin K Epoxide Reductase and Its Bacterial Homologue Have Different Membrane Topologies and Reaction Mechanisms. *J. Biol. Chem.* **287**, 33945–33955 (2012).

67. Wu, S. *et al.* Warfarin and vitamin K epoxide reductase: a molecular accounting for observed inhibition. *Blood* (2018) doi:10.1182/blood-2018-01-830901.

68. Rishavy, M. A., Usubalieva, A., Hallgren, K. W. & Berkner, K. L. Novel Insight into the Mechanism of the Vitamin K Oxidoreductase (VKOR) ELECTRON RELAY THROUGH Cys43 AND Cys51 REDUCES VKOR TO ALLOW VITAMIN K REDUCTION AND FACILITATION OF VITAMIN K-DEPENDENT PROTEIN CARBOXYLATION. *J. Biol.*

Chem. **286**, 7267–7278 (2011).

69. Gong, I. Y. *et al.* Clinical and genetic determinants of warfarin pharmacokinetics and pharmacodynamics during treatment initiation. *PLoS One* **6**, e27808 (2011).
70. Czogalla, K. J., Biswas, A., Rost, S., Watzka, M. & Oldenburg, J. The Arg98Trp mutation in human VKORC1 causing VKCFD2 disrupts a di-Arginine-based ER retention motif. *Blood* (2014) doi:10.1182/blood-2013-12-545988.
71. Sharpe, H. J., Stevens, T. J. & Munro, S. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* **142**, 158–169 (2010).
72. Elazar, A., Weinstein, J. J., Prilusky, J. & Fleishman, S. J. Interplay between hydrophobicity and the positive-inside rule in determining membrane-protein topology. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10340–10345 (2016).
73. Hopf, T. A. *et al.* Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **149**, 1607–1621 (2012).
74. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
75. Elazar, A. A. *et al.* Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. *eLife Sciences* e12125 (2016).
76. Guerriero, C. J. *et al.* Transmembrane helix hydrophobicity is an energetic barrier during the retrotranslocation of integral membrane ERAD substrates. *Mol. Biol. Cell* **28**, 2076–2090 (2017).
77. Hessa, T. *et al.* Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450**, 1026–1030 (2007).
78. Toth-Petroczy, A. *et al.* Structured States of Disordered Proteins from Genomic Sequences.

Cell **167**, 158–170.e12 (2016).

79. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
80. vonHeijne, G. Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341**, 456–458 (1989).
81. Nilsson, I. & von Heijne, G. Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. *Cell* **62**, 1135–1141 (1990).
82. Hatahet, F. *et al.* Altered Escherichia coli membrane protein assembly machinery allows proper membrane assembly of eukaryotic protein vitamin K epoxide reductase. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15184–15189 (2015).
83. Ulmschneider, M. B. & Sansom, M. S. Amino acid distributions in integral membrane protein structures. *Biochim. Biophys. Acta* **1512**, 1–14 (2001).
84. Fleming, K. G. & Engelman, D. M. Specificity in transmembrane helix-helix interactions can define a hierarchy of stability for sequence variants. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14340–14344 (2001).
85. Mravic, M. *et al.* Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science* **363**, 1418–1423 (2019).
86. Gray, V. E., Hause, R. J. & Fowler, D. M. Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics* **207**, 53–61 (2017).
87. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
88. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–8 (2015).

89. Liu, S., Cheng, W., Fowle Grider, R., Shen, G. & Li, W. Structures of an intramembrane vitamin K epoxide reductase homolog reveal control mechanisms for electron transfer. *Nat. Commun.* **5**, (2014).

90. Ma, W. & Goldberg, J. Rules for the recognition of dilysine retrieval motifs by coatomer. *EMBO J.* **32**, 926–937 (2013).

91. Hessa, T. *et al.* Protein targeting and degradation are coupled for elimination of mislocalized proteins. *Nature* **475**, 394–397 (2011).

92. Kurnik, D. *et al.* Effect of the VKORC1 D36Y variant on warfarin dose requirement and pharmacogenetic dose prediction. *Thromb. Haemost.* **108**, 781–788 (2012).

93. Oldenburg, J. *et al.* Mutations in the VKORC1 Gene Cause Warfarin Resistance, Warfarin Sensitivity and Combined Deficiency of Vitamin K Dependent Coagulation Factors. *Blood* **104**, 277–277 (2004).

94. Merkle, P. S. *et al.* Substrate-modulated unwinding of transmembrane helices in the NSS transporter LeuT. *Sci Adv* **4**, eaar6179 (2018).

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

96. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).

97. Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019).

98. Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138**, 103–141 (2013).

99. Zhang, H.-F. *et al.* Physiological Content and Intrinsic Activities of 10 Cytochrome P450 Isoforms in Human Normal Liver Microsomes. *J. Pharmacol. Exp. Ther.* **358**, 83–93 (2016).

100. Johnson, J. A. *et al.* Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 Genotypes and Warfarin Dosing. *Clin. Pharmacol. Ther.* **90**, 625–629 (2011).

101. Tai, G. *et al.* In-vitro and in-vivo effects of the CYP2C9*11 polymorphism on warfarin metabolism and dose. *Pharmacogenet. Genomics* **15**, 475–481 (2005).

102. Poulos, T. L., Finzel, B. C., Gunsalus, I. C., Wagner, G. C. & Kraut, J. The 2.6-A crystal structure of Pseudomonas putida cytochrome P-450. *J. Biol. Chem.* **260**, 16122–16130 (1985).

103. Williams, P. A. *et al.* Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* **424**, 464–468 (2003).

104. Wester, M. R. *et al.* The Structure of Human Cytochrome P450 2C9 Complexed with Flurbiprofen at 2.0-Å Resolution. *J. Biol. Chem.* **279**, 35630–35637 (2004).

105. Hasemann, C. A., Kurumbail, R. G., Boddupalli, S. S., Peterson, J. A. & Deisenhofer, J. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* **3**, 41–62 (1995).

106. Arendse, L. B. & Blackburn, J. M. Effects of polymorphic variation on the thermostability of heterogenous populations of CYP3A4 and CYP2C9 enzymes in solution. *Sci. Rep.* **8**, 11876 (2018).

107. Gaedigk, A. *et al.* The Evolution of PharmVar. *Clin. Pharmacol. Ther.* **105**, 29–32 (2019).

108. Gaedigk, A., Whirl-Carrillo, M., Pratt, V. M., Miller, N. A. & Klein, T. E. PharmVar and

the Landscape of Pharmacogenetic Resources. *Clin. Pharmacol. Ther.* **107**, 43–46 (2020).

109. Landrum, M. J. & Kattman, B. L. ClinVar at five years: Delivering on the promise. *Hum. Mutat.* **39**, 1623–1630 (2018).

110. Guengerich, F. P. & Johnson, W. W. Kinetics of ferric cytochrome P450 reduction by NADPH-cytochrome P450 reductase: rapid reduction in the absence of substrate and variations among cytochrome P450 systems. *Biochemistry* **36**, 14741–14750 (1997).

111. Subramanian, M., Tam, H., Zheng, H. & Tracy, T. S. CYP2C9-CYP3A4 protein-protein interactions: role of the hydrophobic N terminus. *Drug Metab. Dispos.* **38**, 1003–1009 (2010).

112. Hu, G., Johnson, E. F. & Kemper, B. CYP2C8 exists as a dimer in natural membranes. *Drug Metab. Dispos.* **38**, 1976–1983 (2010).

113. Xiong, Y. *et al.* A systematic genetic polymorphism analysis of the CYP2C9 gene in four different geographical Han populations in mainland China. *Genomics* **97**, 277–281 (2011).

114. Wright, A. T., Song, J. D. & Cravatt, B. F. A Suite of Activity-Based Probes for Human Cytochrome P450 Enzymes. *J. Am. Chem. Soc.* **131**, 10692–10700 (2009).

115. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).

116. Redler, R. L., Das, J., Diaz, J. R. & Dokholyan, N. V. Protein Destabilization as a Common Factor in Diverse Inherited Disorders. *J. Mol. Evol.* **82**, 11–16 (2016).

117. 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.* (2015) doi:10.1038/gim.2015.30.

118. Flockhart, D. A. *et al.* Pharmacogenetic testing of CYP2C9 and VKORC1 alleles for warfarin. *Genet. Med.* **10**, 139–150 (2008).

119. Caudle, K. E. *et al.* Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genet. Med.* (2016) doi:10.1038/gim.2016.87.

120. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.* **10**, 748 (2014).

121. Ruaño, G. *et al.* High carrier prevalence of combinatorial CYP2C9 and VKORC1 genotypes affecting warfarin dosing. *Per. Med.* **5**, 225–232 (2008).

122. Cooper, G. M. *et al.* A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112**, 1022–1027 (2008).

123. D'ambrosio, R. L., D'andrea, G., Cafolla, A., Faillace, F. & Margaglione, M. A new vitamin K epoxide reductase complex subunit-1 (VKORC1) mutation in a patient with decreased stability of CYP2C9 enzyme: Letters to the Editor. *J. Thromb. Haemost.* **5**, 191–193 (2007).

124. Wang, D. *et al.* Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. *Blood* **112**, 1013–1021 (2008).

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

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

127. Chhibber, A. *et al.* Transcriptomic variation of pharmacogenes in multiple human tissues

and lymphoblastoid cell lines. *Pharmacogenomics J.* **17**, 137–145 (2017).

128. Toscano, C. *et al.* Impaired expression of CYP2D6 in intermediate metabolizers carrying the *41 allele caused by the intronic SNP 2988G>A: evidence for modulation of splicing events. *Pharmacogenet. Genomics* **16**, 755–766 (2006).

129. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).

130. Cao, J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

131. Lee, S.-B. *et al.* Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).

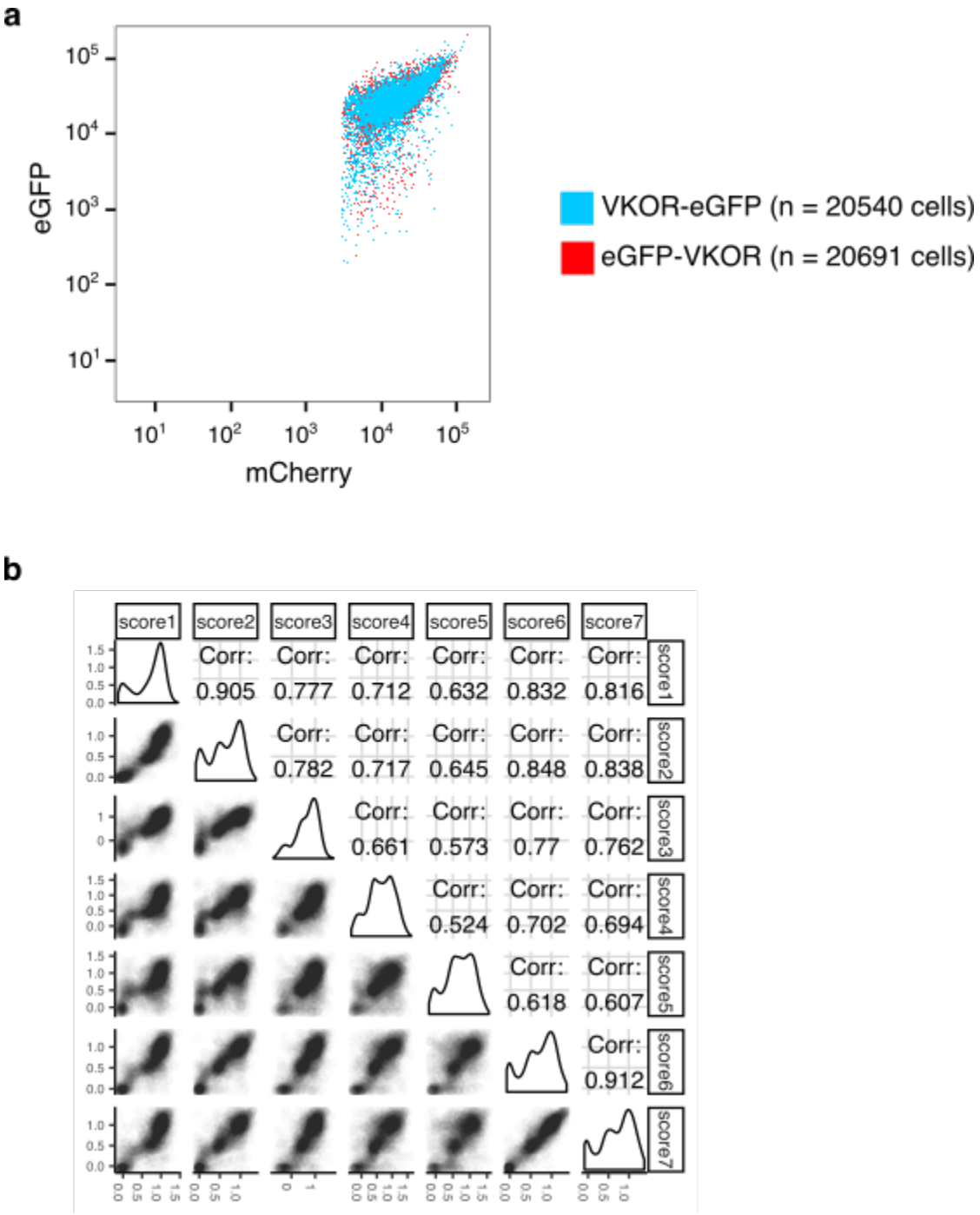
132. Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. *J. Hum. Genet.* **65**, 11–19 (2020).

133. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

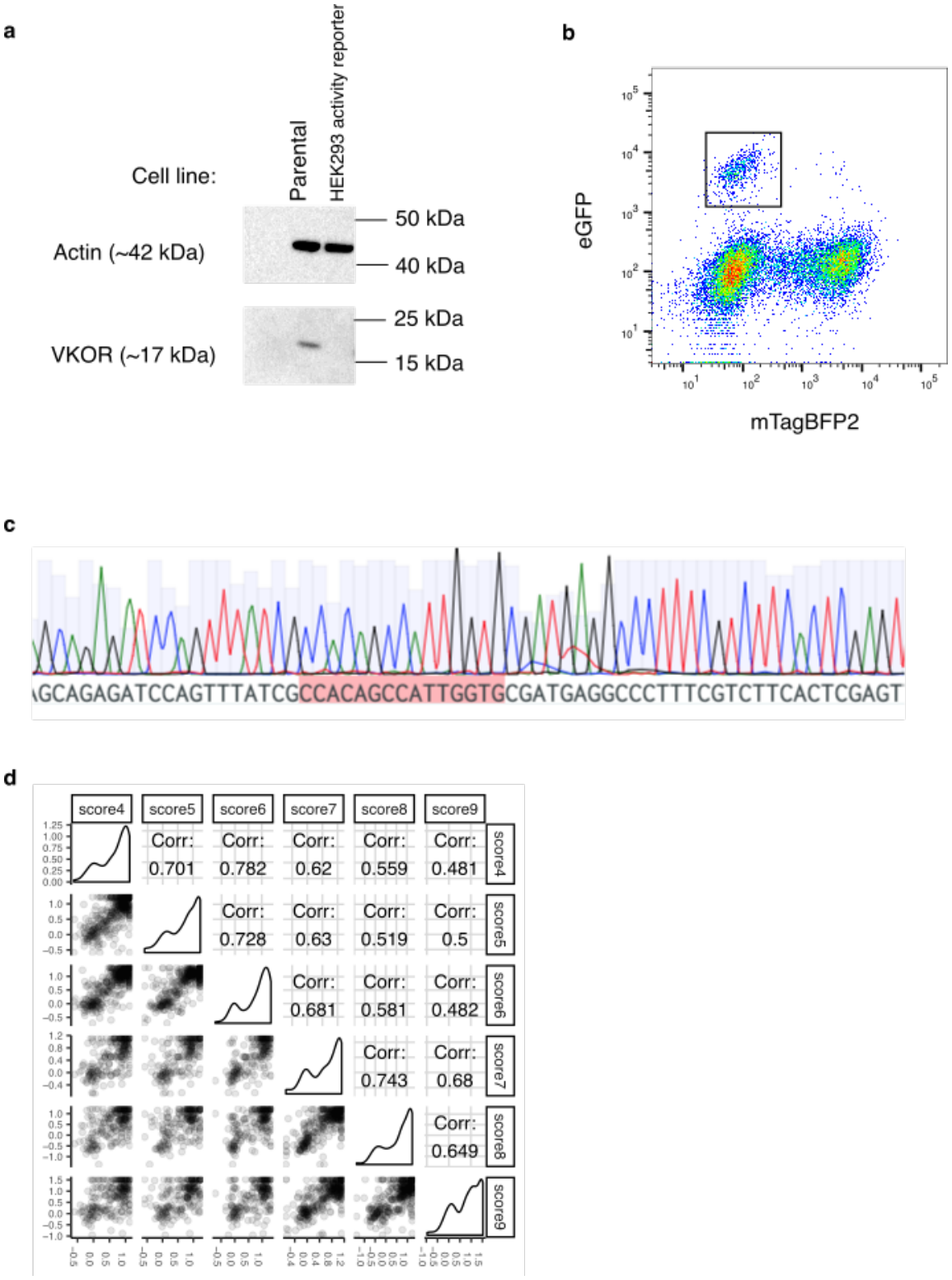
134. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168 (2018).

135. Hallgren, K. W., Qian, W., Yakubenko, A. V., Runge, K. W. & Berkner, K. L. r-VKORC1 Expression in Factor IX BHK Cells Increases the Extent of Factor IX Carboxylation but Is Limited by Saturation of Another Carboxylation Component or by a Shift in the Rate-Limiting Step†. *Biochemistry* **45**, 5587–5598 (2006).

Appendix

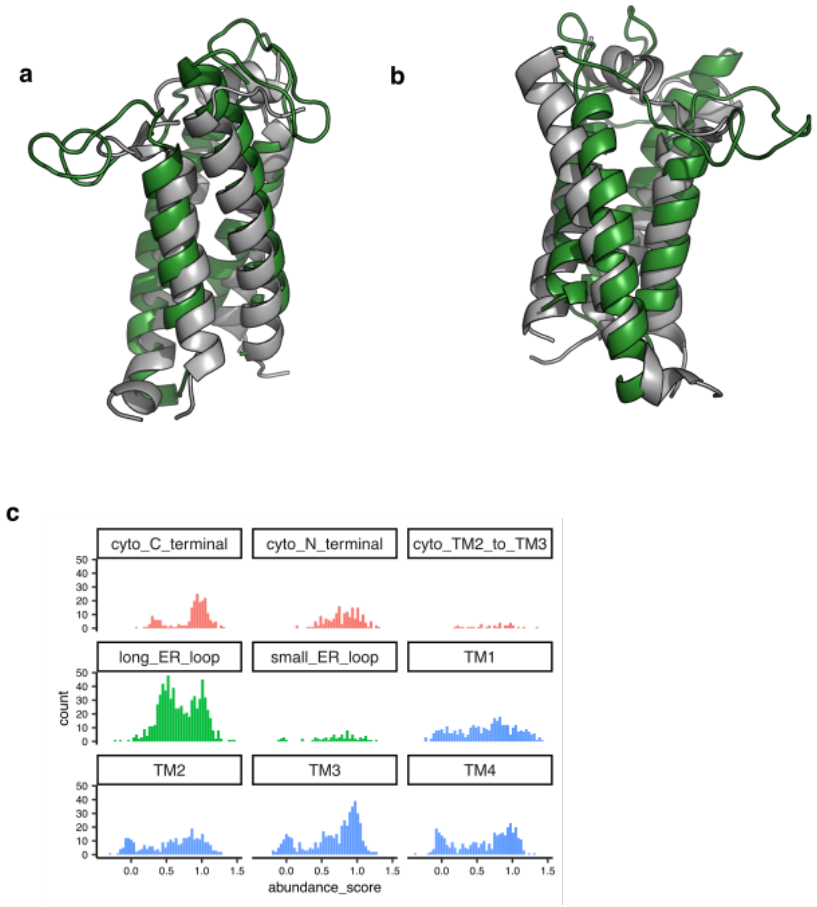


Appendix 1. Scatterplot of eGFP vs. mCherry fluorescence for cells expressing either C-terminally eGFP-tagged VKOR (VKOR-eGFP, blue) or N-terminally eGFP-tagged VKOR (eGFP-VKOR, red). **b**, Pairwise abundance score correlations between replicate sorting experiments. Seven VAMP-seq replicates were performed. Pearson's correlation coefficients are shown. Score numbers in this figure correspond to replicate numbers shown in Appendix Table 1.

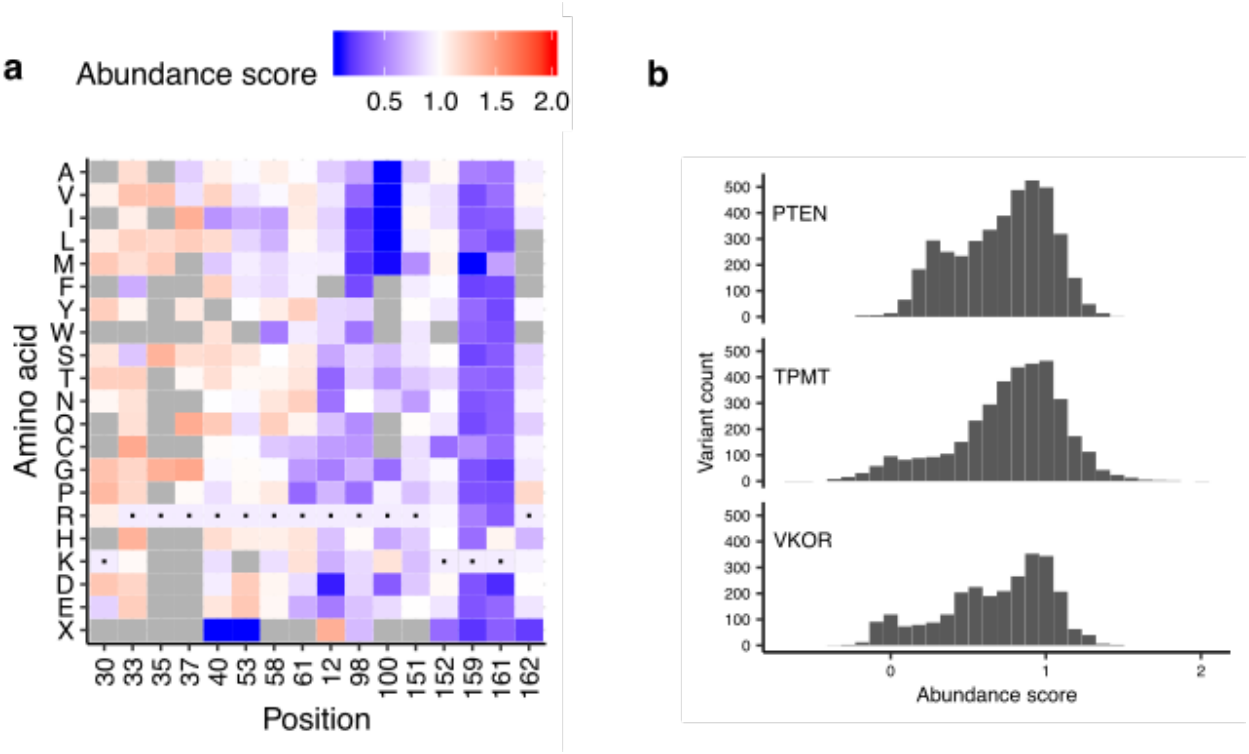


Appendix 2. **a**, Western blot of parental cell line vs. HEK293 activity reporter cell line. Loading control is actin (42 kDa). VKOR was probed using an antibody generated against a peptide from the C-terminal of VKOR (FRKVQEPQGKAKRH)[135]. The band for VKOR at 17 kDA is visible in the parental cell line but is not present in the HEK293 activity reporter cell line. **b**, Scatterplot showing

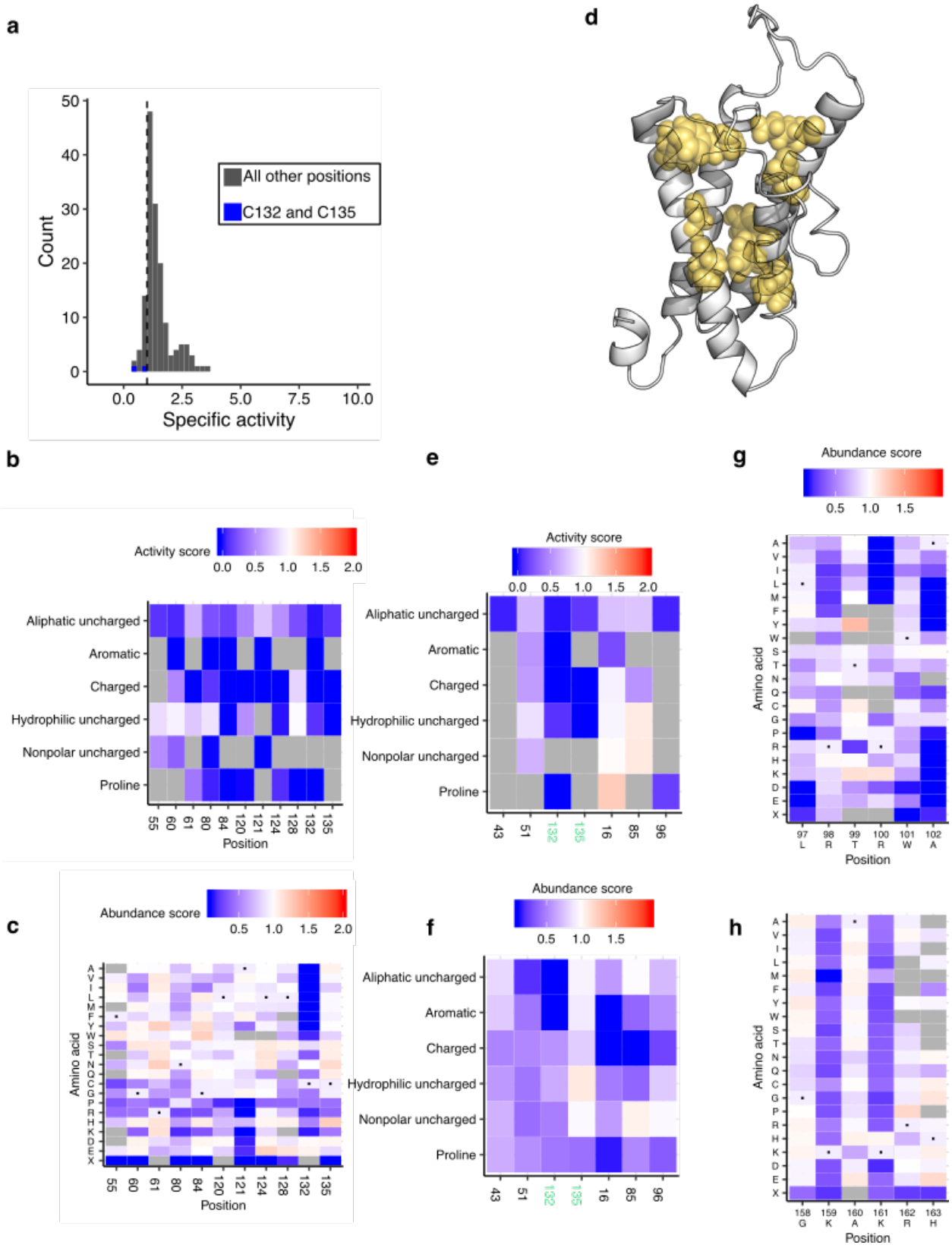
mTagBFP2 vs. eGFP mean fluorescence intensities for HEK293 activity reporter cells recombined with a construct encoding WT VKOR followed by internal ribosomal entry sequence and eGFP. The emergence of a distinct recombined population that is eGFP positive and mTagBFP2 negative (black outline, n = 768 cells) supports the presence of a single landing pad into the cell genome, and not multiple insertions. **c**, A chromatogram showing the barcode sequence of the landing pad inserted at the *AAVS1* locus in the HEK293 activity reporter cell line. The presence of a single barcode, highlighted in red, instead of mixed peaks, supports insertion of one landing pad rather than multiple landing pads. **d**, Pairwise score correlations between replicate sorting experiments of VKOR activity. Six replicates of the activity assay were performed. Pearson's correlation coefficients are shown. Score numbers in this panel correspond to replicate numbers shown in Appendix Table 2**.**



Appendix 3. **a**, Pymol graphic showing overlap between EVcouplings-folded model of VKOR (shown as a cartoon in green) compared to the bacterial structure (PDB: 4NV5, shown as a cartoon in grey). **b**, shows the same two structures, rotated 120°C. **c**, Histograms of abundance scores for missense variants, grouped by domain and colored by cytoplasmic, ER lumenal, or transmembrane localization.

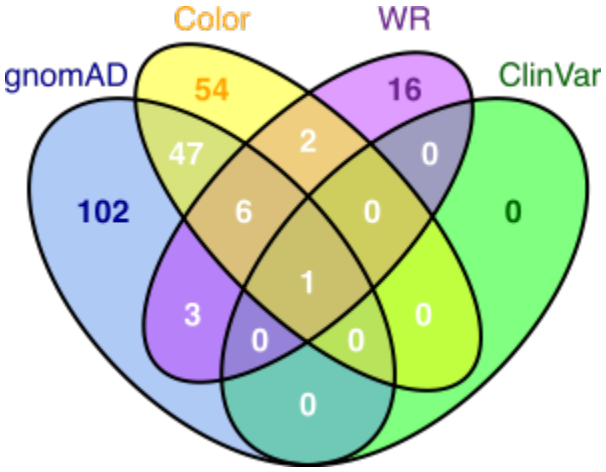


Appendix 4. **a**, Heatmap of abundance scores for all arginines and lysines in VKOR. First four positions (K30, K33, K35, K37) are in or proximal to transmembrane domain 1. Heatmap color indicates abundance scores scaled as a gradient between the lowest 10% of abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). Grey bars indicate missing variants. Black dots indicate WT amino acids. **b**, Histograms of abundance scores for missense variants for three proteins: PTEN, TPMT, and VKOR.


a
Count
Specific activity
All other positions
C132 and C135
d
b
Activity score
Aliphatic uncharged
Aromatic
Charged
Hydrophilic uncharged
Nonpolar uncharged
Proline
Position
e
Activity score
g
Abundance score
Amino acid
Position
c
Abundance score
Amino acid
Position
f
Abundance score
h
Amino acid
Position

Appendix 5. **a**, Histogram of specific activity, with catalytic cysteines C132 and C135 labeled in blue. Dashed line demarcates bottom 12.5%. **b**, Heatmap of activity scores for residues with lowest 12.5% of specific activity scores, collapsed by amino acid class. Color indicates abundance scores scaled as a gradient between the lowest 10% of abundance scores (blue), the WT abundance score (white) and abundance scores above WT (red). Grey indicates missing data **c**, Heatmap of abundance scores for residues with lowest 12.5% of specific activity scores. Color legend same as in **b**. **d**, Active site positions as defined by computational docking, shown on the homology model as yellow spheres[60]. **e**, Heatmap of activity scores for cysteines. Catalytic cysteines C132 and C135 labeled in green. Color legend same as in **b**. **f**, Heatmap of abundance scores for cysteines. Catalytic cysteines C132 and C135 labeled in green. Color legend same as in **b**. **g**, Heatmap of abundance scores for diarginine ER retention motif. X-axis shows residues and position. Color legend same as in **b**. **h**, Heatmap of abundance score for dilysine ER retention motif. X-axis shows residues and position. Color legend same as in **b**.

SUPPLEMENTARY FIGURE 6



Appendix 6. Venn diagram of VKOR missense variants present in gnomAD v2 and v3, ClinVar, Color Genomics, a commercial genetic testing company, and literature-reported warfarin resistant variants.

Replicate	Cells recombined	Cells sorted in four-way sort
1	256,324	200,000
2	256,324	200,000
3	111,570	183,000
4	155,169	100,000
5	105,000	103,000
6	54,045	120,000
7	54,045	125,000

Appendix Table 1. The seven replicates of VAMP-seq performed with cells recombined and sorted for each.

Replicate	Cells recombined	Cells sorted in four-way sort
1	85,492	200,000
2	85,492	150,000
3	85,492	100,000
4	165,000	90,000
5	165,000	90,000
6	165,000	100,000

Appendix Table 2. The six replicates of the activity assay performed with cells recombined and sorted for each.

VITA

Melissa was born and grew up in Arlington, Texas. She attended Yale University, earning a B.S. in Biology in 2011 and conducting undergraduate research with Paul Turner. She then worked for two years on *Toxoplasma gondii* genetics in Michael Grigg's group at the Laboratory of Parasitic Disease, NIAID, NIH. She started graduate school in 2013 at the University of Washington in the Department of Genome Sciences.