





Deep mutational scanning of CYP2C19 in human cells reveals a substrate specificity-abundance tradeoff

Gabriel E. Boyle ¹, Katherine A. Sitko,¹ Jared G. Galloway,² Hugh K. Haddox,² Aisha Haley Bianchi ¹, Ajeya Dixon,¹ Melinda K. Wheelock,¹ Allyssa J. Vandi,¹ Ziyu R. Wang,¹ Raine E.S. Thomson,³ Riddhiman K. Garge,^{1,4} Allan E. Rettie,⁵ Alan F. Rubin,^{6,7} Renee C. Geck ¹, Elizabeth M.J. Gillam,³ William S. DeWitt,⁸ Frederick A. Matsen IV,^{1,2,9,10} Douglas M. Fowler ^{1,11,*}

¹Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

²Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

³School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD 4067, Australia

⁴Brotman Baty Institute for Precision Medicine, University of Washington, Seattle, WA 98195, USA

⁵Department of Medicinal Chemistry, University of Washington, Seattle, WA 98195, USA

⁶Bioinformatics Division, Walter and Eliza Hall Institute, Parkville, VIC 3052, Australia

⁷Department of Medical Biology, University of Melbourne, Melbourne, VIC 3052, Australia

⁸Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720, USA

⁹Howard Hughes Medical Institute, Seattle, WA 98109, USA

¹⁰Department of Statistics, University of Washington, Seattle, WA 98195, USA

¹¹Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

*Corresponding author: Department of Genome Sciences, University of Washington, William H. Foege Hall, 3720 15th Ave NE, Seattle, WA 98195, USA.

Email: dfowler@uw.edu

The cytochrome P450s enzyme family metabolizes ~80% of small molecule drugs. Variants in cytochrome P450s can substantially alter drug metabolism, leading to improper dosing and severe adverse drug reactions. Due to low sequence conservation, predicting variant effects across cytochrome P450s is challenging. Even closely related cytochrome P450s like CYP2C9 and CYP2C19, which share 92% amino acid sequence identity, display distinct phenotypic properties. Using variant abundance by massively parallel sequencing, we measured the steady-state protein abundance of 7,660 single amino acid variants in CYP2C19 expressed in cultured human cells. Our findings confirmed critical positions and structural features essential for cytochrome P450 function, and revealed how variants at conserved positions influence abundance. We jointly analyzed 4,670 variants whose abundance was measured in both CYP2C19 and CYP2C9, finding that the homologs have different variant abundances in substrate recognition sites within the hydrophobic core. We also measured the abundance of all single and some multiple wild type amino acid exchanges between CYP2C19 and CYP2C9. While most exchanges had no effect, substitutions in substrate recognition site 4 reduced abundance in CYP2C19. Double and triple mutants showed distinct interactions, highlighting a region that points to differing thermodynamic properties between the 2 homologs. These positions are known contributors to substrate specificity, suggesting an evolutionary tradeoff between stability and enzymatic function. Finally, we analyzed 368 previously unannotated human variants, finding that 43% had decreased abundance. By comparing variant effects between these homologs, we uncovered regions underlying their functional differences, advancing our understanding of this versatile family of enzymes.

Keywords: deep mutational scanning; cytochrome P450; substrate specificity; thermodynamic stability

Introduction

Nearly 20,000 cytochrome P450 (CYP) heme monooxygenases have been identified across all domains of life (Nelson 2011). CYPs catalyze a wide range of reactions with a diverse set of substrates, making them some of the most versatile enzymes in existence (Coon 2005; Munro et al. 2013). The 57 human CYP genes are grouped into 18 families with 43 subfamilies (Zhao et al. 2021), highlighting their genetic heterogeneity even within a single species. Despite their genetic and functional diversity, key structural and topological features of CYPs are highly conserved (Werck-Reichhart and Feyereisen 2000; Sirim et al. 2010). However, the relationship between CYP genetic variation, structure, and function is far from fully elucidated. For

example, within CYP family 2, subfamily C (CYP2C), CYP2C19 (MIM: 124020, 609535) and CYP2C9 (MIM: 601130) are the most closely related subfamily members, sharing 92% amino acid sequence identity. Their protein structures have nearly identical organization, with the largest deviations between their C α backbones in the substrate binding cavity being only ~3 Å (Reynald et al. 2012). Yet, the 2 homologs are functionally distinct, with largely disparate sets of substrates (Niwa and Yamazaki 2012; Wishart et al. 2018) and divergent membrane interactions (Mustafa et al. 2019). Moreover, CYP2C19's melting temperature is ~11°C higher than CYP2C9's (Thomson 2021). Thus, even between these close homolog CYPs, the 43 diverged positions drive large functional differences.

Received on 05 August 2024; accepted on 31 August 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of The Genetics Society of America. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Understanding the functional impact of variants across CYPs is particularly important because ~12 of the 57 human CYPs contribute to metabolizing 70–80% of currently prescribed drugs that are processed by enzymes for elimination. Of those 70–80% CYP-metabolized drugs, CYP2C19 and CYP2C9 account for 20–30% (Zanger and Schwab 2013). Genetic variation in CYPs can substantially alter individual drug response leading to adverse drug reactions (ADRs), which are among the leading causes of morbidity and mortality (Lazarou et al. 1998; de Vries et al. 2008), and cost an estimated \$30.1 billion annually (Sultana et al. 2013). To provide clinicians guidance for treating individuals with CYP variants, the Clinical Pharmacogenetics Implementation Consortium (CPIC) categorizes CYP genes into star (*) allele haplotypes according to enzymatic function: normal function, decreased function, no function, and increased function (Sim and Ingelman-Sundberg 2010; Relling and Klein 2011). Genetic testing and employment of CPIC Guidelines can prevent many ADRs. For example, up to 30% of the population may have a CYP2C19 variant with reduced function (Klein et al. 2018) which may result in impaired activation of the antiplatelet drug clopidogrel. Genotyping for CYP2C19 loss of function variants can avoid major adverse cardiovascular events (Galli et al. 2021; Pereira et al. 2021; Dean and Kane 2022). However, only a very small number of CYP variants have established functional consequences, and it is unknown to what degree variant effects in one CYP can be applied to others.

Measuring CYP variant function individually is laborious and low throughput, but massively multiplexed methods can be used instead. Variant abundance by massively parallel sequencing (VAMP-seq) measures steady-state protein abundance of thousands of variants in parallel (Matreyek et al. 2018). In VAMP-seq, as in other similar methods (Yen et al. 2008; Kim et al. 2013; Klesmith et al. 2017; Zutz et al. 2021), steady-state protein abundance is used as a proxy for protein stability. Steady-state protein abundance refers to the final concentration of a protein when its rates of synthesis and decay are balanced (Hargrove and Schmidt 1989). To control for variation in protein synthesis, the VAMP-seq vector uses a single promoter to express 2 fluorescent reporters from a single mRNA transcript using an internal ribosome entry site (IRES). In each cell, the fluorescent signal of the target gene fused to enhanced green fluorescent protein (eGFP) is normalized to mCherry fluorescence meaning changes in fluorescence are due to degradation. The resulting measurements correlate with changes in thermodynamic stability, indicating that reduced steady-state abundance is likely due to loss of protein stability (Matreyek et al. 2018; Suiter et al. 2020; Zutz et al. 2021; Christensen et al. 2023).

Previously, we used the VAMP-seq assay (Matreyek et al. 2018) to measure the abundance of 6,370 of 9,780 possible single amino acid variants in CYP2C9 (Amorosi et al. 2021). From the resulting variant effect map, we identified patterns of loss of abundance that revealed mutationally sensitive regions of the protein. Additionally, we revealed hundreds of variants with reduced abundance in the human population in addition to providing variant effect measurements for thousands of variants not yet observed (Amorosi et al. 2021).

Here, we used VAMP-seq to measure 7,660 of 9,780 possible single amino acid variants of CYP2C19. We identified 4,698 variants that likely result in reduced protein abundance, with 1,122 of those exhibiting complete loss of abundance equivalent to nonsense mutations. We first analyzed positions conserved across all eukaryotic CYPs, revealing that all but 6 of the 58 conserved positions were intolerant of substitutions. Four of the tolerant positions were catalytically important sites buried in the hydrophobic core where mutations are nearly always deleterious, suggesting

that some sites critical for enzyme function may not impact abundance. We jointly analyzed the CYP2C19 and CYP2C9 variant abundance dataset and found 2,366 variants the abundance of which differed between the 2 enzymes. Most differences were of small effect, though a fraction of the differences were large. While nearly all sites had at least one variant that differed, 83 of 489 (17%) sites were significantly different between the homologs. CYP2C9 had higher mutational tolerance in its hydrophobic core than CYP2C19, and variants in the structurally conserved K' helix were highly deleterious in CYP2C19, but tolerated in CYP2C9, even though all K' positions contain the same amino acid in both homologs. We analyzed wild type (WT) amino acid exchanges between CYP2C19 and CYP2C9, revealing that sequence differences in a set of diverged positions contribute to possible differences in thermodynamic stability between the homologs. These divergent positions are also important for substrate specificity, suggesting that reduced thermodynamic stability in CYP2C9 may have been evolutionarily tolerated in exchange for functional benefit (DePristo et al. 2005). Finally, we analyzed the effects of human CYP2C19 variants. Our abundance scores are largely concordant with existing functional annotations indicating that, like for many other proteins, loss of abundance accounts for the majority of loss of function alleles. We provided abundance scores for 368 out of 408 (90.2%) previously unannotated single amino acid variants in the Genome Aggregation Database (gnomAD) (Karczewski et al. 2020). Thus, by conducting the first comparative analysis of closely related CYPs using large-scale variant effect data we provide fundamental insights into common CYP structural features that differentially impact abundance between CYP2C19 and CYP2C9. We also provide functional annotations for human CYP2C19 variants which could be used to improve genotype-guided dosing of drugs metabolized by CYP2C19.

Methods

General reagents

Unless otherwise noted, all chemicals were obtained from (MilliporeSigma) and all the enzymes were obtained from New England Biolabs. All cell culture reagents were purchased from Thermo Fisher unless otherwise noted. All plasmids and oligonucleotides used in this study are listed in [Supplementary Table 3](#).

Growth media and culturing techniques

HEK293T cells (ATCC CRL-3216) and the derived landing pad cell line were cultured in Dulbecco's Modified Eagle Medium supplemented with 10% fetal bovine serum, 100 U/mL penicillin, and 0.1 mg/mL streptomycin. Landing pad expression was induced with doxycycline at a final media concentration of 2.5 µg/mL. Cells were passaged by detachment with trypsin 0.5% (w/v). All cell lines were tested for mycoplasma on a monthly basis and consistently negative.

Library mutagenesis

The CYP2C19 library was constructed using inverse PCR-based site-directed saturation mutagenesis (Jain and Varadarajan 2014). Saturation mutagenesis primers were designed for each codon of CYP2C19 across positions 2 through 490. Each forward primer contained an NNK (N: A, C, G, or T; K: G or T) at the 5' end of the sequence. Primers were obtained from Integrated DNA Technologies (IDT). Our library consisted of 7,660 of 9,291 (82.3%) possible missense substitutions represented by 147,723 unique barcodes (mean of 11.87 and median of 7 for single amino acid variants; see [Supplementary Table 2](#) for details).

CYP2C19 WT was codon-optimized for human expression in a pHSG298 backbone. We completed inverse PCRs using NNK oligos for each position excluding the methionine at position 1. Each PCR reaction contained 125 pg of template, 2 μ M of mixed primers, and 5% DMSO in a 5 μ L reaction volume of KAPA HiFi Hotstart 2 \times ReadyMix. The resulting products were confirmed by visualizing on a gel and quantified using either the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) or Qubit fluorometry (Life Technologies). The PCR products were then pooled at equimolar ratios and cleaned using the DNA Clean and Concentrator Kit (Zymo Research), followed by gel extraction. The pooled libraries were 5' phosphorylated with T4 polynucleotide kinase and subjected to intramolecular ligation overnight. Next, 8.5 μ L of phosphorylated product were combined with 1 μ L of 10 \times T4 ligase buffer and 0.5 μ L of T4 DNA ligase (NEB), incubated at 16°C overnight, and cleaned and concentrated. The ligated products were transformed into electrocompetent *Escherichia coli* cells (NEB C2989K or C3020K) with electroporation at 2 kV, and the resulting transformants were plated on LB + kanamycin. The CFUs on the plates were counted to estimate the number of unique molecules transformed and to estimate the coverage of the library. Finally, the library was subcloned into the expression and recombination vectors and barcoded.

To generate barcoded libraries, the variant library was first digested with SacII and AflIII at 37°C for 1 h, followed by heat inactivation at 65°C for 20 min. We ordered barcode oligos with 18 bp random sequences from IDT, resuspended them at 100 μ M, and annealed them by combining 1 μ L of each primer with 4 μ L of CutSmart buffer and 34 μ L of ddH₂O and running 98°C for 3 min, ramping down to 25°C at $-0.1^\circ\text{C}/\text{s}$. The annealed oligos were then filled by combining 0.8 μ L Klenow polymerase (exonuclease negative, NEB) with 1.35 μ L of 1 mM dNTPs and 40 μ L of product to fill in the barcode oligo, incubating at 25°C for 15 min, 70°C for 20 min, then ramping down to 37°C at $-0.1^\circ\text{C}/\text{s}$. The resulting products were then ligated overnight at 16°C. The barcoded library was transformed into electrocompetent *E. coli* cells (NEB C2989K), and were then midiprep (QIAGEN). The size of the barcoded library was bottlenecked and estimated by colony counts to be 67,000.

To obtain more accurate library counts, we sequenced the library barcodes with Illumina sequencing. The forward and reverse reads were merged using Pear (Zhang et al. 2014), and barcode counts were estimated using Bartender (Zhao et al. 2018). Barcodes with fewer than 10 reads were filtered out, resulting in ~200,000 unique barcodes for an average of 21 \times coverage.

An important caveat to this library cloning method is underrepresentation of nonsense variants. This is because NNK oligos encode 32 unique codons with only one stop codon (i.e. TAG). Therefore, instead of 3 out of 64 (~4.7%) nonsense variants per position, this method produces 1 out of 32 (~3.1%) nonsense variants.

PacBio sequencing for barcode-variant mapping

PacBio sequencing libraries were generated with SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences) according to the manufacturer's instructions. The barcoded variant sequences were excised using restriction enzymes NheI-HF and HindIII-HF and purified with AMPure PB beads (Pacific Biosciences 100-265-900) at a 1:1 ratio of beads to DNA. Following end-repair, A-tail attachment, and ligation, the assembled product was extracted using a BluePippin instrument (Sage Science, BLU0001) using a 0.75% (w/v) agarose precast cassette (Sage Science, BLF7510). Library purity and size was confirmed by 4200 TapeStation (Agilent, G2991BA) before sequencing. Samples were submitted to University of Washington PacBio Sequencing Services and sequenced on one SMRT (Single Molecule, Real-Time) Cell in a Sequel II v2.0 run using a 15 h movie.

We filtered long reads for a minimum of 3 passes. We then analyzed the circular consensus reads using PacRAT to identify and link the gene variants with the barcode region (Yeh et al. 2022). The filtered barcode-variant library contained 12,559 unique nucleotide sequences tagged by 176,372 unique barcodes (see [Supplementary Table 2](#) for details).

FACS-based deep mutational scan (VAMP-seq)

All human cell experiments used HEK293T cells with a Bxb1 serine recombinase landing pad with an inducible Caspase 9 cassette (HEK293T-LLP-iCasp9) (Matreyek et al. 2020) that enabled expression of one variant per cell. To recombine the variant library into HEK293T cells, 3,500,000 cells were seeded in 10 cm plates (2–4 per replicate) and transfected with FuGENE 6 Transfection Reagent (Promega, E2692). In 1 tube, 7.1 μ g of barcoded library plasmid were mixed with 0.48 μ g of Bxb1 plasmid in 710 μ L of OptiMEM. In a separate tube, 28.5 μ L of Fugene were diluted into 685 μ L of OptiMEM. The Fugene and DNA tubes were then combined and incubated at room temperature for 15 min. The Fugene/DNA mixture was added to cells dropwise, and cells were incubated for a minimum of 48 h before induction with doxycycline at a final concentration of 2.5 μ g/mL. After 24 h of doxycycline treatment, we added AP1903 at a final concentration of 2 nM to induce Caspase 9 dimerization and eliminate all unrecombined cells.

Transfected HEK293T cells were sorted using a BD AriaIII sorter. Cells were gated for live, recombined singlets. In recombined cells, the ratio of GFP:mCherry fluorescence was calculated and plotted as a histogram. The histogram was split into 4 quartiles. Each quartile was sorted into separate 5 mL tubes. Cells from each bin were grown out for 1–2 days to ensure enough DNA for sequencing and to improve replicate correlation compared to immediate sequencing, as previously found (Amorosi et al. 2021). Three biological replicates from separate transfections were collected for the FACS-based deep mutational scan (see [Supplementary Table 1](#) for details).

Sorted abundance library amplification and sequencing

Sorted cells were harvested and pelleted by centrifugation, and then stored at -20°C until all replicates were collected. Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer's instructions, with the addition of a 30 min incubation step at 37°C with RNase during the resuspension step. For the first round of PCR, eight 50 μ L reactions were set up for each sample, with a final concentration of 50 ng/ μ L input genomic DNA, 1 \times Q5 High-Fidelity Master Mix, and 0.25 μ M of JS454 and JS1004 primers (see [Supplementary Table 3](#) for details). The reaction conditions were 95°C for 30 s, 98°C for 10 s, 60°C for 30 s, 72°C for 3 min, repeated 4 additional times, followed by 72°C for 2 min and a 4°C hold. The 8 reactions were then combined, bound to AMPure XP (Beckman Coulter) at 0.6 \times bead volume to sample volume, cleaned, and eluted with 38.5 μ L water. From the eluted volume, 15 μ L (40%) was mixed with Q5 High-Fidelity Master Mix, GB001, and one of the indexed reverse primers, JS385 through JS473, added at 0.25 μ M each. The PCR reaction was run with SYBR Green I on a Bio-Rad MiniOpticon. The reaction was denatured for 3 min at 95°C, cycled 18 times at 95°C for 15 s, 67°C for 30 s, and 72°C for 45 s, with a final 2-min extension at 72°C.

The indexed amplicons were then run on a TapeStation according to the manufacturer's instructions. For each sample, 1 μ L sample was mixed with 3 μ L of sample buffer, thoroughly mixed, and run on a D1000 ScreenTape (Agilent Technologies) using an internal electronic ladder. The bands were quantified using the

TapeStation analysis software. The samples were then pooled in equal amounts, loaded onto a 1% (w/v) agarose gel with SYBR Safe, and then the gel was extracted using a Freeze 'N Squeeze column (Bio-Rad). Finally, the quantification of the pooled sample was done with the Qubit 1X dsDNA Assay Kit Broad Range (Q33266).

Library sequence analysis

Barcode sequences were trimmed and filtered for a minimum base quality of Q20 using the FASTX-toolkit. These barcodes were then used to generate a FASTQ file input for Enrich2 to count variants. Variants with insertions, deletions, or multiple amino acid substitutions were excluded. Barcode counts were then collapsed to variant counts, retaining variants with a total frequency $>4 \times 10^{-5}$ across all bins (Supplementary Fig. 2). For each replicate, an abundance score was calculated using a weighted average of variant frequency across bins ($w_1 = 0.25, w_2 = 0.5, w_3 = 0.75, w_4 = 1$) (Matreyek et al. 2018). Scores were normalized to synonymous and nonsense distributions, excluding the top 20% of nonsense scores. Single amino acid variant abundance scores ranged from -0.09 to 1.5 .

Abundance classes were determined as in previous studies (Matreyek et al. 2018; Amorosi et al. 2021). To discriminate between “WT-like” and “decreased” scores, we used a synonymous score threshold. This threshold was set at the 5th percentile of synonymous scores (0.856). Variants were classified as “WT-like” if their lower confidence interval exceeded the threshold, or as “possibly WT-like” if only their score surpassed the threshold. Additionally, an upper threshold at the 95th percentile of synonymous scores (1.14) was used to differentiate between “WT-like” and “increased” scores. To distinguish between “decreased” and “nonsense-like” scores, we used a threshold at the 95th percentile of nonsense scores (0.265). Variants were categorized as “nonsense-like” if both their score and upper confidence interval were below the nonsense threshold, or as “possible nonsense-like” if only their score fell below the threshold. Out of a total of 8,480 variants, 316 were nonsense, 504 were synonymous, and 7,660 were single amino acid variants. The single amino acid variants were categorized into the following abundance classes: 2,590 WT-like, 612 possibly WT-like, 3,146 decreased, 437 possibly decreased, 340 possibly nonsense-like, and 708 nonsense-like.

VAMP-seq internal validation with individual variants

We selected 11 variants that spanned the range of abundance scores for validation. Using IVA cloning site-directed mutagenesis (García-Nafria et al. 2016), we generated 11 CYP2C19 variants into the VAMP-seq recombination vector (attB-CYP2C19-eGFP-IRES-mCherry) via primers listed in Supplementary Table 3 (HB049 through HB073, GB143, and GB144). Mutations were generated with KAPA HiFi DNA Polymerase (KAPA Biosystems KK2601) and 40 ng of CYP2C19 template plasmid attB-CYP2C19-eGFP-IRES-mCherry. After completing inverse PCR for each variant, we digested the products with DpnI to eliminate remaining WT template, and used them to transform chemically competent *E. coli* cells (NEB C2987 or Bioline BIO-85027). Bacterial clones were prepped with a midprep kit then validated by Sanger sequencing and whole plasmid nanopore sequencing. We then transfected the preps into HEK293T-LLP-iCasp9 landing pad cells in 6-well plates with 400,000 cells per well. In each transfection, 2.7 μg of plasmid were mixed with 0.300 μg of Bxb1 plasmid in 125 μL of OptiMEM and 5 μL P3000 reagent. In a separate tube, 2.25 μL of Lipofectamine was added to 125 μL of OptiMEM. The tubes were then combined and incubated at room temperature for 15 min. After incubation, the Lipofectamine/DNA mixture was added to cells dropwise and

the plates were placed in an incubator at 37°C. After 24 h, the cells were induced with doxycycline at a final concentration of 2.5 $\mu\text{g}/\text{mL}$, and at least 24 h later we selected for recombinant cells by adding the small molecule, AP1903, which causes inducible Caspase 9 in unrecombined landing pad cells to dimerize, become activated, and induce apoptosis.

Recombined cells were grown to full confluence and analyzed with a BD LSRII flow cytometer. Cells were gated for live, recombined singlets. We calculated a ratio of eGFP/mCherry fluorescence, and the geometric mean of the distribution of this ratio was reported. Flow cytometry data were collected with FACSDiva V8.0.1 (BD Biosciences) and analyzed with FlowJo V.10.8.1 (Ashland, OR, USA). Three biological replicates of each individual variant were measured.

multidms analysis of CYP2C19 and 2C9 deep mutational scans

We used an open-source software package called multidms (Haddox et al. 2023) to jointly model the deep mutational scanning (DMS) data for the CYP2C19 and CYP2C9 (Amorosi et al. 2021) homologs, using the model to estimate shifts in mutational effects between homologs. Each DMS experiment densely sampled many of the possible mutations across the 2 homologs. Variants within each library were limited to at most 1 mutation, but nearly all variants were associated with multiple unique barcodes (Supplementary Fig. 14). The input data to multidms consisted of barcode-level abundance scores from 2 biological replicate DMS experiments for both homologs. We grouped these 4 raw datasets into 2 independent training datasets, each of which consisted of 1 replicate from each CYP2C19 and CYP2C9. We then separately fit multidms models to the training sets and found the inferred parameters (see below) were well-correlated between the 2 models (Supplementary Figs. 6–9).

The models were trained to minimize a loss function that includes 2 terms: a Huber loss term that penalizes differences between predicted and experimentally measured abundance scores for each barcoded variant from each input dataset, and a lasso regularization term that penalizes nonzero shifts in mutational effects. For a given barcoded variant (v) from a given experiment (d), the model predicts the variant's abundance score ($y_{v,d}$) using the equation: $y_{v,d} = \beta_0 + \sum_m (\beta_m + A_{m,d})$, where β_0 is the inferred abundance score of the CYP2C9 wild type sequence, β_m is the inferred effect of mutation m on the protein's abundance score in the CYP2C9 experiment, $A_{m,d}$ is the inferred shift in the mutation's effect in experiment d relative to the CYP2C9 experiment ($A_{m,d}$ is fixed to zero if d is CYP2C9), and where the summation term sums over all mutations m in the variant v relative to the CYP2C9 wild type sequence. The lasso regularization term is applied to the shift ($A_{m,d}$) parameters, causing these parameters to be zero unless nonzero values are strongly supported by the data. How strongly a shift is supported by the data depends on factors like the magnitude of the shift and the number of barcoded variants with a given mutation m within each homolog's DMS dataset. To determine a reasonable penalty coefficient (λ) for the lasso regularization term, we compared the results of seven model fits—each using different coefficients ranging from $\lambda = 0$, to $\lambda = 5e - 4$. This λ sweep was performed in duplicate using each of the distinct training sets, for a total of 14 model fits evaluated. Supplementary Fig. 8 shows summary statistics of the model fits. This figure emphasizes the accuracy-simplicity tradeoff for any given value of λ . For example, as λ increases, we observe higher overall training set loss but also increased correlation between replicate parameters values. We selected a penalty coefficient of $\lambda = 1e - 5$ as a reasonable value to balance model accuracy with shift sparsity. The shifts in the main text are then reported as the averaged values

between the replicate model fits, each using the selected penalty coefficient. See <https://github.com/matsengrp/CYP-multidms> for the code used for this analysis.

We identified positions whose mean shift value was significantly different from the distribution of all shift values using a randomization test. To generate a null distribution, we randomly sampled 10 shift values, the average number of abundance scores per position, and calculated the mean of the shifts. This procedure was repeated 100,000 times. We calculated *P*-values for each position by counting the number of randomly generated mean shifts more extreme than the position mean and dividing it by 100,000, the total number of randomly generated shifts. The *P*-values were then adjusted for repeated hypothesis testing using the Benjamini–Hochberg method (Benjamini and Hochberg 1995) with a false discovery rate (FDR) of 5%. Positions with *P*-values <0.05 were considered significant.

FoldX in silico mutagenesis

Structure files for CYP2C19 (4GQS) and CYP2C9 (1OG2) were downloaded from the PDB website (rcsb.org) and Chain A was saved separately to ensure mutagenesis on a single CYP monomer. Mutations were generated using FoldX 5.1 (<https://foldxsuite.crg.eu>). FoldX command Repair PDB was used to optimize total energy of protein to the FoldX force field prior to mutagenesis. The BuildModel command was used to generate all single amino acid variants or amino acid swaps between each homolog, as listed in their corresponding individual_list.txt file. Calculations were completed 5 times for each variant and $\Delta\Delta G$ values were averaged automatically by FoldX. All configuration, mutation, and output files used in this study are available at GitHub: https://github.com/FowlerLab/cyp2c19_2c9.

Results

Multiplexed measurement of CYP2C19 variant abundance

We used VAMP-seq to simultaneously measure the steady-state abundance of CYP2C19 variants in cultured human cells (Matreyek et al. 2018; Amorosi et al. 2021) (Fig. 1a). VAMP-seq relies on 2 fluorescent reporters: GFP fused to each CYP2C19 variant to read out abundance, and mCherry expressed via an IRES as a transcriptional control. Because CYP2C19 is N-terminally inserted into the ER membrane, we fused GFP onto the C-terminus, as we did for a previous VAMP-seq experiment on CYP2C9 (Amorosi et al. 2021). Expression of the WT CYP2C19 C-terminal GFP fusion led to strong fluorescent signal, and R433W, a known destabilizing CYP2C19 variant, had substantially lower signal indicating that the C-terminal GFP fusion construct was compatible with VAMP-seq (Fig. 1b).

We introduced a barcoded library of CYP2C19 variants into HEK293T cells using a recombinase-based landing pad, such that each cell expressed only one variant (Matreyek et al. 2017, 2020). Cells were sorted into quartile bins based on the ratio of GFP:mCherry fluorescence. Each bin was deeply sequenced, variant-associated barcodes were counted, and abundance scores were calculated based on weighted average of barcode frequencies across bins (Fig. 1a). Abundance scores were highly correlated between seven replicate sorting experiments arising from 3 independent library recombinations (Supplementary Fig. 1a–d; Pearson's $R = 0.82$ – 0.98). Replicate scores were averaged, filtered (Supplementary Fig. 2a–d) and normalized such that the median nonsense variant had a score of 0 and WT had a score of 1 (Matreyek et al. 2018; Amorosi et al. 2021).

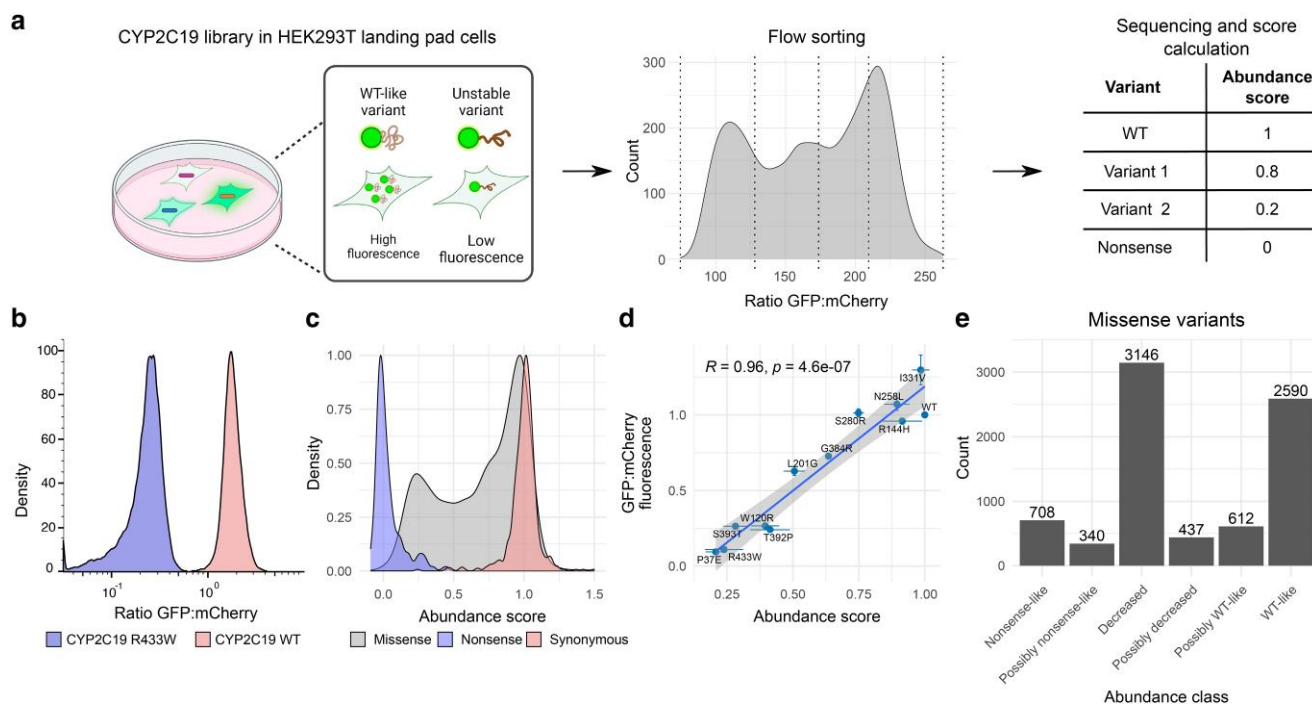


Fig. 1. Multiplexed measurement of CYP2C19 abundance. Variant assessment by massively parallel sequencing (VAMP-seq) measures variant abundance at scale. **a** In VAMP-seq, a barcoded library fused to GFP is recombined into a genomically integrated landing pad in HEK293T cells. mCherry is expressed co-transcriptionally via an IRES. Unstable variants are degraded by the proteostasis machinery of the cell, resulting in lower GFP signal compared to WT-like variants. Flow cytometry is then used to sort cells into quartile bins according to fluorescence, bins are deeply sequenced, and barcode counts are used to calculate an abundance score. Figure panel modified from Amorosi et al. (2021) **b** GFP:mCherry ratio for cells expressing either CYP2C19 WT (right) or the R433W destabilizing variant (left) ($n \sim 30,000$). **c** Abundance score distributions for synonymous ($n = 504$), nonsense ($n = 316$), and missense ($n = 7,660$) variants. **d** GFP:mCherry ratios, measured for cells using flow cytometry, for 10 individual variants plotted against their VAMP-seq derived abundance scores (Pearson's $R = 0.96$, $n = 30,000$ cells). Error bars represent the SD of abundance scores (x axis) or mean fluorescence (y axis). **e** Number of single amino acid variants in each abundance class.

Our final data set contained abundance scores for 8,480 of 10,290 (82%) possible variants, of which 7,660 were missense, 316 were nonsense, and 504 were synonymous (Supplementary Table 4). Abundance scores of synonymous and nonsense variants were well separated, with the single amino acid variant distribution spread between nonsense and synonymous variants (Fig. 1c). Individually measured GFP:mCherry ratios for 10 variants spanning the range of abundance scores were highly correlated with VAMP-seq scores (Fig. 1d; Pearson's $R=0.92$; Supplementary Table 6). We also compared our results to those from a smaller scale VAMP-seq experiment encompassing 121 variants (Zhang et al. 2020) (Supplementary Fig. 3a). We used variant weighted average calculations, scaled from 0.25–1, to ensure consistency between datasets. Weighted averages are precursors to abundance scores, which are scaled from 0 to 1 by the median nonsense weighted average (see Methods). Our results were highly consistent with the exception of four variants that diverged substantially (Zhang et al. 2020) (Supplementary Fig. 3a, Pearson's $R=0.74$). One such variant, E444Q, had low representation in our library, meaning that the disagreement could be due to sampling error (Supplementary Fig. 3b). However, the remaining 3 variants were robustly sampled in our assay (Supplementary Fig. 3b). Thus, our VAMP-seq derived abundance scores faithfully reproduced variant abundance. Lastly, we classified variants according to their abundance score relative to the range of scores from nonsense and synonymous variants (Supplementary Fig. 3b and Fig. 1c,e). The majority (58%, 4,620 variants) of single amino acid variants decreased abundance (Fig. 1e).

Mutational tolerance at conserved CYP2C19 positions reflects function

We visualized the abundance scores as a variant effect map (Fig. 2a) and projected position-averaged scores onto the CYP2C19 structure (Fig. 2b). Many of the low abundance variants occur within α -helices and β -sheets (Fig. 2a and b), especially in amino acids on interior α -helix turns and in regions closer to the protein core (Fig. 2b).

While CYPs vary widely in sequence, key structural and functional features are highly conserved (Hasemann et al. 1995; Mestres 2005). However, despite this high level of conservation, the role of some positions in human CYPs are still poorly understood because some of these positions have not been studied, and others have only been studied in evolutionarily distant, nonhuman CYPs (Gricman et al. 2015). To bridge this gap, we investigated the abundance of variants at positions that are conserved across eukaryotic CYPs, defined as positions where >80% of CYPs have the same or biophysically similar amino acids (Gricman et al. 2014). Hierarchical clustering of these eukaryotically conserved positions revealed five clusters with distinct patterns of variant abundance scores (Fig. 2c and Supplementary Fig. 4a). Overall, nearly all these conserved positions are critical for abundance. The clusters were defined by positions having similar variant effects amongst biophysically related amino acids (Gricman et al. 2014).

In clusters 1, 2, and 3 nearly all substitutions, except those of the same biophysical type, reduced abundance (Fig. 2c and Supplementary Fig. 4a). In cluster 4, substitutions caused moderate loss of abundance, with no consistent pattern across all positions. The sole exceptions were two of the three positions where glycine was the WT amino acid. These positions tolerated alanine and cysteine substitutions suggesting that amino acid size is an important factor. Cluster 5 contained M136, A297, E300, T301, K322, and I362, all of which were substantially more tolerant of mutations than the other conserved sites indicating that they are critical to CYP2C19 function but not abundance. The combined conservation

and tolerance of M136 and K322 can be explained by the fact that these positions are located on the surface of the protein and that they are likely to bind to the critical cofactor cytochrome P450 reductase (CPR), as they do in the closely related CYP2C9 (Berka et al. 2011; Lertkiatmongkol et al. 2013). However, amino acids A297, E300, T301, and I362 are buried in the hydrophobic core making their mutational tolerance more challenging to explain (Supplementary Fig. 4b). Positions 297 and 362 influence substrate specificity, and >80% eukaryotic CYPs have hydrophobic amino acids at these positions (Gricman et al. 2014). Surprisingly, while substitutions are tolerated at these positions, some hydrophobic substitutions elicit moderate reductions in abundance. T301 is a critical threonine for oxygen activation and catalysis in CYP2C19 (Altarsha et al. 2009; Haines et al. 2001; Foti et al. 2012; Reynald et al. 2012) and contains hydrogen-bonding amino acids in >80% eukaryotic CYPs, and most substitutions did not appreciably reduce abundance. Finally, E300 stabilizes a water network during proton delivery (Haines et al. 2001), and substitutions other than aspartic acid were tolerated at this position. To test the hypothesis that these positions are important for function but not abundance, we intersected CYP2C19 abundance scores with CYP2C9 variant activity scores measured with click-seq, which uses a substrate-like probe to assess variant effects (Amorosi et al. 2021). CYP2C9's VAMP-seq and click-seq scores correlated with Pearson's $R=0.76$ (Amorosi et al. 2021). CYP2C19 abundance and CYP2C9 activity scores were similarly strongly correlated (Supplementary Fig. 4c; Pearson's $R=0.682$). Most variants at positions 297 and 301 reduced CYP2C9 activity but not CYP2C19 abundance, highlighting these positions' importance for binding the click-seq probe (Supplementary Fig. 4d). At position 300, all but one variant had WT-like abundance whereas several variants had profoundly reduced activity, consistent with a functional role that does not require direct substrate binding (Supplementary Fig. 4d). CYP2C9 activity scores were unavailable for position 362.

Thus, substitutions at nearly all conserved positions caused reduced abundance. However, positions 136, 297, 300, 301, 322, and 362, which participate in catalysis or cofactor binding, were largely tolerant of substitutions despite their location in the hydrophobic core of CYP2C19. We speculate that this tolerance is a consequence of the dynamic and flexible nature of CYP active sites, making these positions important for catalytic activity but not folding and stability (Nair et al. 2016).

Comparing variant abundance effects between CYP2C19 and CYP2C9 reveals core-stabilizing regions with distinct mutational tolerance

Next, we investigated variant effect patterns in CYP2C19 compared to its closest homolog, CYP2C9. CYP2C19 and CYP2C9 share 92% protein sequence identity and nearly identical crystal structures (Supplementary Fig. 5, root mean square deviation [RMSD] = 0.596 Å). However, they have important functional differences, notably their substrate profiles and membrane interactions (Goldstein and de Moraes 1994; Niwa et al. 2002; Niwa and Yamazaki 2012; Mustafa et al. 2019). Moreover, the temperature at which they lose the ability to bind their heme cofactor, which reflects thermodynamic stability (Gumulya et al. 2018), differs by 11°C (Thomson 2021). Thus, small differences in sequence and structure translate into distinct functional and phenotypic characteristics.

To understand how these functional differences arise, we sought to estimate how much the abundance score is shifted between the CYP2C19 abundance data presented here and abundance data from a previous VAMP-seq experiment we conducted on CYP2C9 variants (Amorosi et al. 2021). The combined dataset contained 4,670

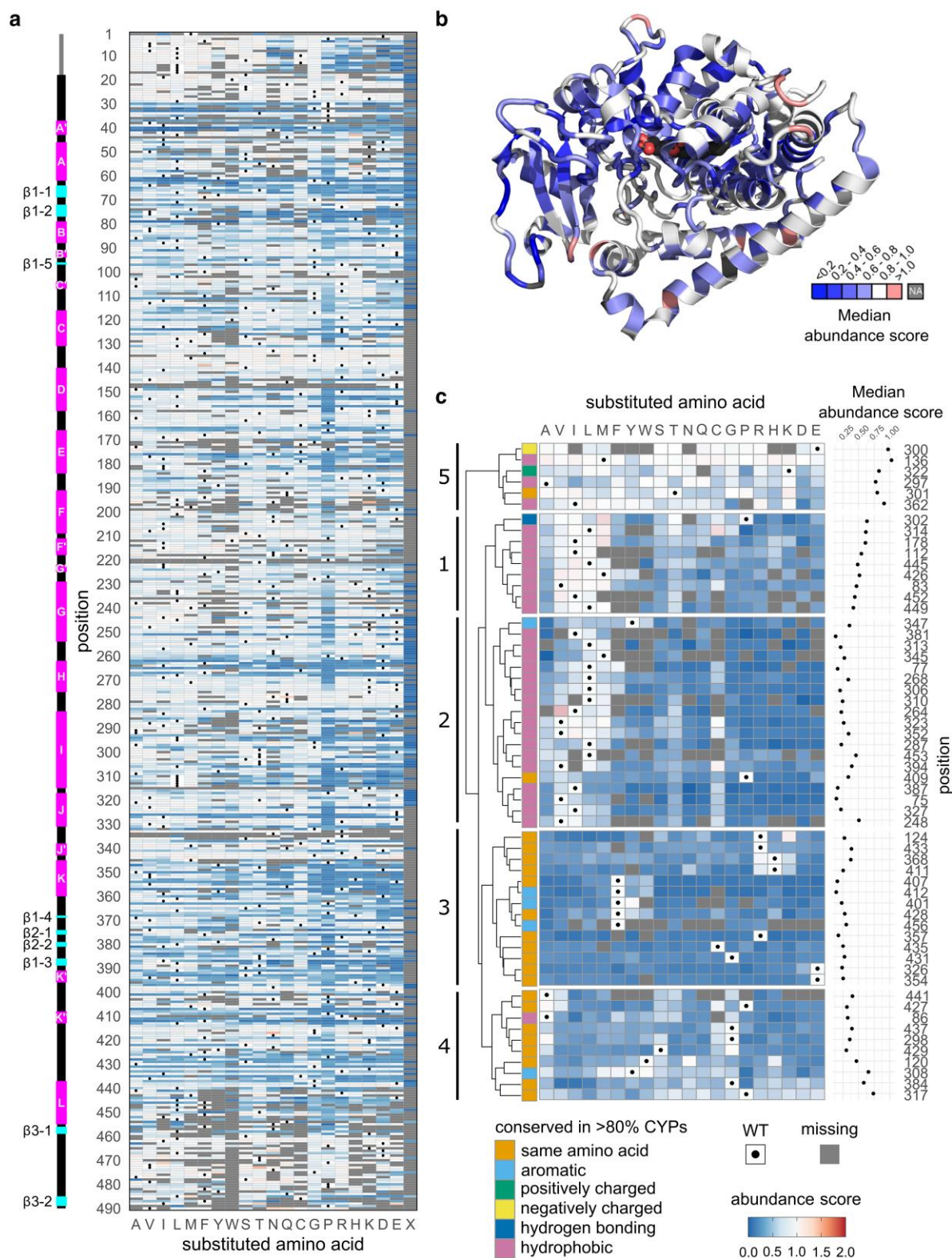


Fig. 2. CYP2C19 variant abundance scores emphasize essential roles of conserved sites. a) Heatmap of CYP2C19 abundance scores. WT amino acids are represented by black dots, and missing data are shown in gray. Substituted amino acids are represented by their single letter abbreviations with "X" denoting a premature stop. Scores range from reduced abundance (blue) to increased (red). Secondary structure of CYP2C19 represented above the heatmap with α -helices shown in magenta and β -sheets shown in cyan. b) Median abundance scores for each position projected onto the CYP2C19 crystal structure (PDB: 4GQS). Color represents the binned median score, with missing scores represented in gray. The heme is colored by element (carbon: black, nitrogen: blue, oxygen: red, iron: yellow). c) Hierarchical clustering of CYP2C19 abundance score profiles by Euclidean distance at positions where >80% of eukaryotic CYPs had the same amino acid (orange) or >80% eukaryotic CYPs had amino acids with the same biophysical property (aromatic: light blue, positively charged: green, negatively charged: yellow, hydrogen bonding: dark blue, hydrophobic: pink) (Gricman *et al.* 2014). Cluster numbers are labeled at the left of the dendrogram.

variants whose abundance was scored in both CYPs. Most variants had similar effects in both homologs, though a subset of variants had large shifts in effects (Fig. 3a, Pearson's $R = 0.77$). While some shifts are likely due to actual biological differences, others are due to the noise inherent in any high-throughput experiment. To identify which shifts are most likely due to signal, we re-estimated shifts using a joint-modeling approach called multidms (Haddox et al. 2023). This approach involved inferring shifts in variant effects between homologs, while regularizing the inferred shifts to drive them to be zero unless they were strongly supported by the abundance data from each homolog ('see Methods' for more details). We inferred shifts as the difference in abundance score in CYP2C19 relative to CYP2C9, such that positive shifts indicate a higher abundance score in CYP2C19 and vice versa. We separately fit 2 multidms models, each 1 trained on 1 experimental replicate per homolog, using different regularization weights ranging from 0.0 to $1e-4$, and selected $1e-5$ as the optimal value for subsequent analysis (Supplementary Figs. 6–8, 'see Methods'). The inferred shifts were well-correlated between replicate model fits, despite some noise, and subsequent analyses report shift values averaged between the replicate fits (Supplementary Fig. 9). A total of 2,366 variants (50.7%) had nonzero-shift values meaning that they had different effects between the 2 homologs, though most shifts were small (Fig. 3b and Supplementary Fig. 10a, Supplementary Table 7).

We calculated the mean of the shift values at each position to reveal the effect of regional and structural features (Fig. 3c and d). We identified positions with mean shift values that differed significantly from 0 using a randomization test (Supplementary Fig. 10b). The region with the largest mean shift values was in the K' helix, which is part of a region that is both highly mobile and critical for packing of the hydrophobic core (Werck-Reichhart and Feyereisen 2000; Denisov et al. 2005) (Fig. 3c and d). In this region, mean shift values were negative, meaning that substitutions were more deleterious in CYP2C19 than in CYP2C9 (Fig. 3e and Supplementary Fig. 10a).

Overall, CYP2C19 was more mutationally tolerant than CYP2C9 in the D, E, I, L, J, and J' helices (Fig. 3c and d, Supplementary Fig. 10a), which form the majority of the hydrophobic core (Werck-Reichhart and Feyereisen 2000; Denisov et al. 2005). The sites that were most differentially tolerant in these helices were on portions of the helices that sit outside of the hydrophobic core. Conversely, CYP2C9 was more mutationally tolerant than CYP2C19 at positions within the hydrophobic core near important sites for heme positioning and function (Fig. 3c). Many of these heme-associated positions reside within substrate recognition sites (SRSs) (Gotoh 1992), and CYP2C9 was more mutationally tolerant than CYP2C19 in SRSs relative to the other regions of the protein (Fig. 3d, f, and g). The mutational tolerances of positions in SRSs that were not heme-associated were similar between the homologs (Fig. 3g).

We also examined whether differences between CYP2C19 and CYP2C9 could be explained by sensitivity to variants of different biophysical types or by differences in the structures of the 2 homologs. However, we found that neither homolog is more sensitive to particular types of substitutions (Supplementary Fig. 11a), and that shift values were unrelated to the distance between positions in the CYP2C19 and CYP2C9 crystal structures (Supplementary Fig. 11b). Thus, comparison of variant effects between CYP2C19 and its closest homolog CYP2C9 revealed that CYP2C19's K' helix and, to a lesser extent, heme-associated positions in the hydrophobic core were more sensitive to mutation than CYP2C9, but that CYP2C19 was less sensitive than CYP2C9 to substitutions in other regions flanking the hydrophobic core.

Amino acid swaps reveal homolog-specific constraints on abundance at sites influencing substrate specificity

We investigated abundance shifts at all variants comparing CYP2C19 and CYP2C9. However, the phenotypic differences in substrate recognition, membrane interaction, and thermodynamic stability between the 2 homologs must be driven by divergent sites. While most divergent sites are not localized to the catalytic site, some are critical for substrate specificity, regioselectivity, and stereospecificity (Ibeanu et al. 1996; Jung et al. 1998; Klose et al. 1998; Lewis et al. 1998; Wada et al. 2008; Attia et al. 2014). In many cases, evolutionary pressures result in a protein's reduction of thermodynamic stability in exchange for new functionality (DePristo et al. 2005). We wondered whether we could link the differences in thermodynamic stability between the homologs to substrate specificity by measuring the protein abundance of the 43 divergent sites. Thus, we investigated the abundance of the variants that partially convert CYP2C19 to CYP2C9 and vice versa.

We had abundance scores for 33 of the 43 CYP2C19 variants that installed the WT CYP2C9 amino acid, (e.g. CYP2C19 → CYP2C9). For an exhaustive analysis, we individually measured GFP:mCherry fluorescence for each of the 10 CYP2C19 → CYP2C9 variants not present in our abundance data (Fig. 4a). All but three CYP2C19 → CYP2C9 substitutions were well tolerated. R261Q and L295F caused modest loss of abundance. Position 295 is critical for the specificity of CYP2C19 for S-mephenytoin and for the specificity of CYP2C9 for diclofenac (Tsao et al. 2001; Niwa et al. 2002) whereas R261Q has not been studied in either homolog. V288E caused the largest loss of abundance of all of the swaps and was classified as "nonsense-like" (Fig. 4a). Position 288 alone does not have a known functional role. However, positions 241, 288, and 289 together have been suggested to play a role in abundance and substrate specificity (Jung et al. 1998; Klose et al. 1998; Tsao et al. 2001; Niwa et al. 2002; Attia et al. 2014). In the CYP2C9 structure, K241 interacts electrostatically with E288 and hydrogen bonds with N289 to stabilize a region of the SRS4 in the I helix (Jung et al. 1998; Lewis et al. 1998). In CYP2C19, all three positions have different amino acids, E241, V288, and I289, and thus no electrostatic interaction between E241 and V288. Thus, the loss of abundance caused by V288E in CYP2C19 was likely to be due to the introduction of an electrostatic clash between E241 and V288E (Fig. 4b). Since all 3 positions contribute to functional differences between CYP2C19 and CYP2C9, we sought to understand how positions 241, 288, and 289 might interact to influence abundance.

First, we individually measured the abundance of single and double mutants at positions 241 and 288 for both CYP2C19 → CYP2C9 and CYP2C9 → CYP2C19 variants (Fig. 4b and c). When individually measured, E241K had no effect on CYP2C19 abundance and V288E profoundly reduced abundance, the same effects we measured using VAMP-seq (Fig. 4c). Combining E241K and V288E partially restored CYP2C19 abundance. CYP2C9 K241E only modestly reduced abundance, even though this variant putatively results in an electrostatic clash similar to the one that dramatically reduced CYP2C19 abundance. CYP2C9 E288V had no effect on abundance, suggesting that the native K241–E288 electrostatic interaction probably does not contribute appreciably to thermodynamic stability (Jung et al. 1998). Combining K241E and E288V fully restored CYP2C9 abundance (Fig. 4c). Thus, both homologs have a similar pattern, with installation of a second negative charge disrupting abundance. Elimination of 1 of the 2 negative charges even with variants from the other homolog restored abundance, although to differing degrees in each homolog.

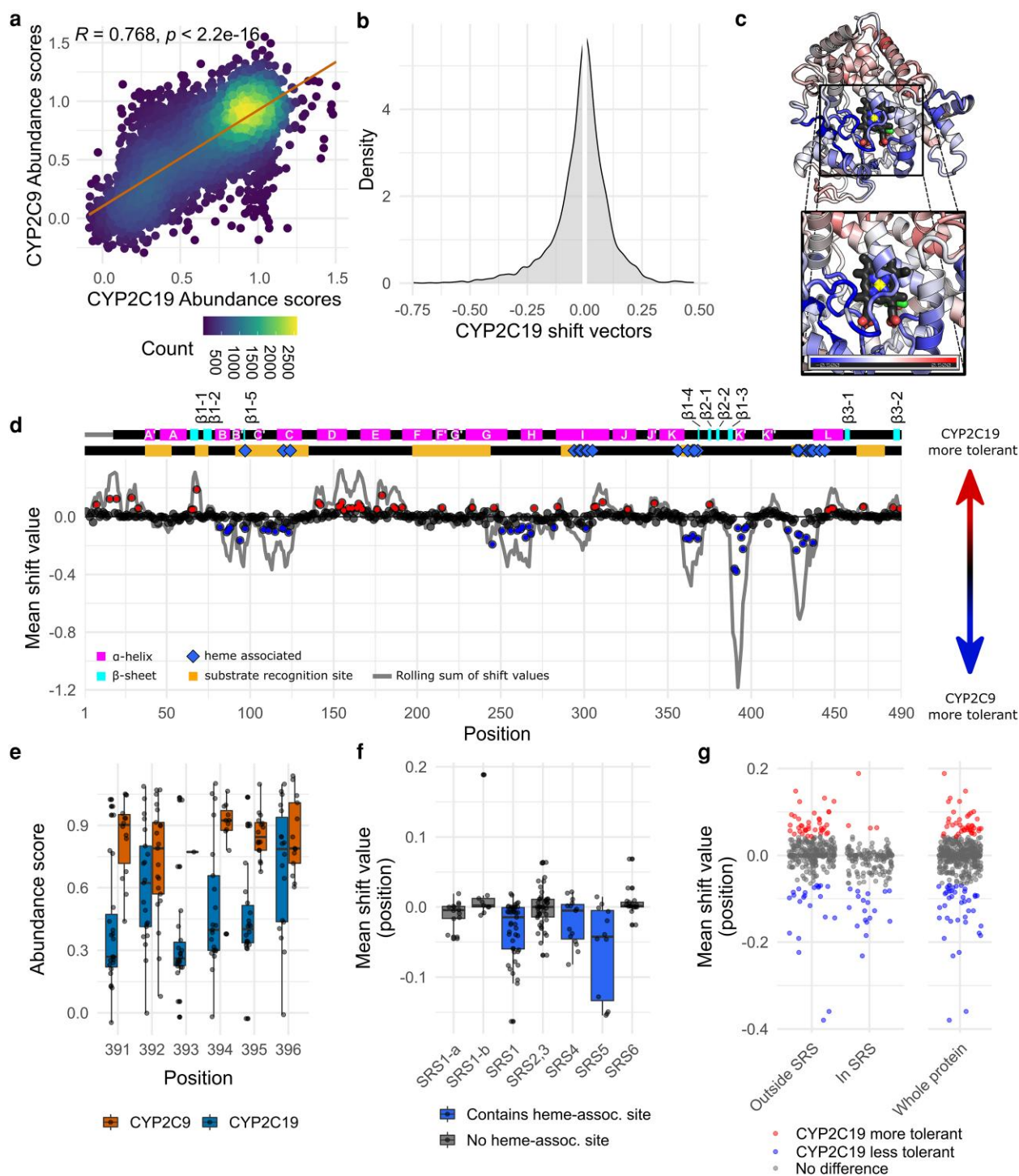


Fig. 3. Comparison of VAMP-seq mutational tolerance of CYP2C19 and CYP2C9. a) Scatterplot of 5,979 abundance scores present in CYP2C19 and our previous CYP2C9 VAMP-seq experiment (Amorosi et al. 2021). b) Distribution of nonzero-shift values calculated using multidms (Haddox et al. 2023). Shift values are shown only for variants present in both datasets. c) CYP2C19 structure (PDB: 4GQS) colored by the rolling sum of position mean shift values shown in d (tiled window size $k = 5$, color gradient -0.5 to 0.5). The heme is colored by element (carbon: black, nitrogen: blue, oxygen: red, iron: yellow), and PDB chemical 0XV (4-hydroxy-3,5-dimethylphenyl)(2-methyl-1-benzofuran-3-yl)methanone is shown in green. The inset box highlights the zoomed view of the K helix adjacent to the heme. d) Top: Secondary structure of CYP2C19 represented with α -helices shown in magenta and β -sheets shown in cyan. Middle: Substrate recognition regions are shown in orange, and sites that interact with the heme are shown with blue diamonds. Bottom: Scatter plot of the mean shift in variant abundance scores between homologs across all variants at a given position in the primary sequence. Filled dots represent positions that are significantly more tolerant of mutations in CYP2C19 (red) or more tolerant of mutations in CYP2C9 (blue) with false discovery rate (FDR) controlled P-values < 0.05 using a randomization test. The rolling sum of the mean shift values is depicted by the gray line. The trends in this plot were highly reproducible across the two replicate model fits (Supplementary Fig. 9). e) Boxplot of variant abundance scores for CYP2C19 (blue) and CYP2C9 (orange) across positions in the K' helix. Dots represent variant abundance scores. f) Boxplot of shift values in SRSs with or without a heme-associated site. Dots represent mean shift values at each position within the substrate recognition site (SRS). g) Dot plot of mean shift values for each position separated by whether or not the position is in an SRS. Colors represent mean shift values that are significantly more tolerant in CYP2C19 (red), more tolerant in CYP2C9 (blue), or are not significantly different (gray) by randomization test.

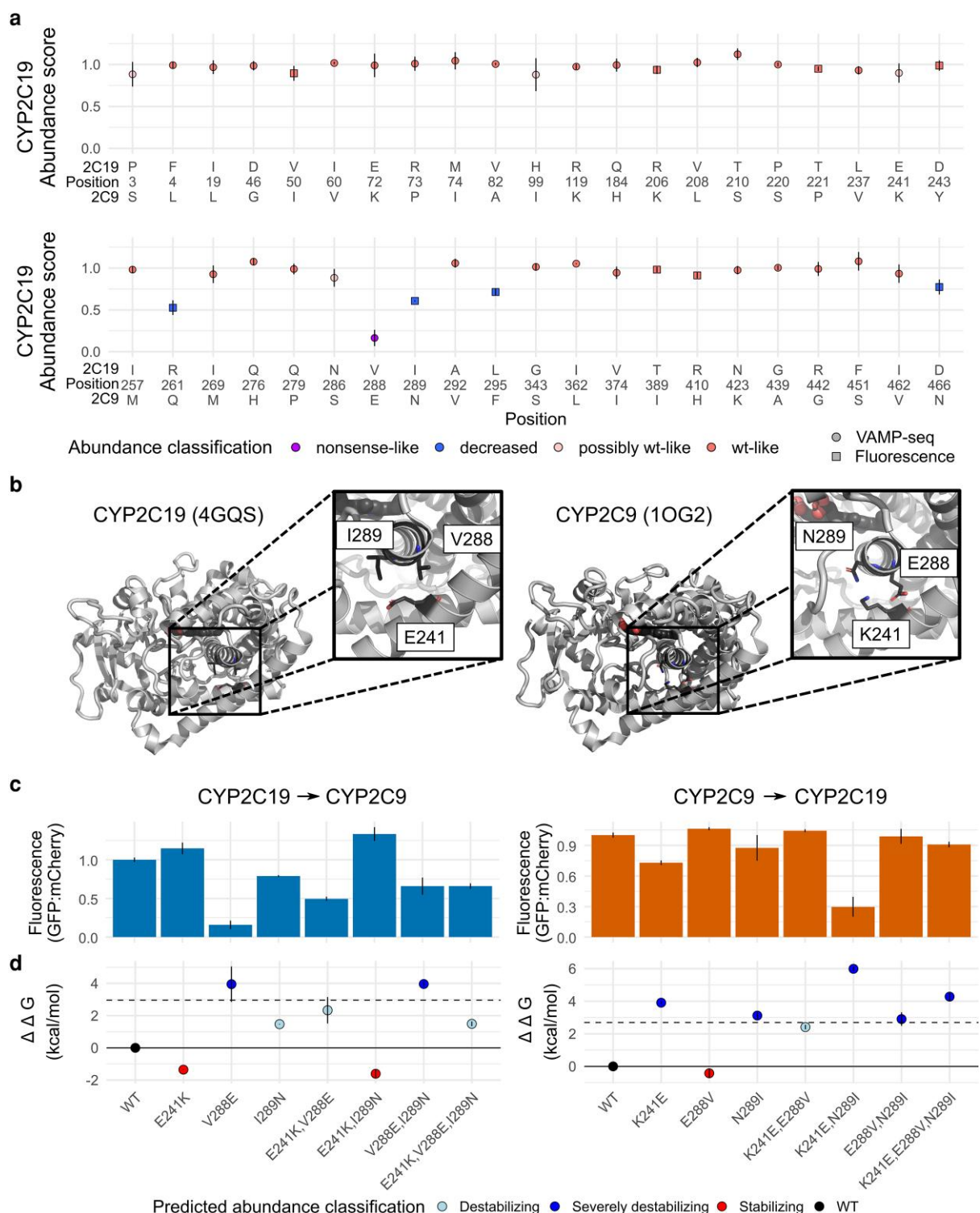


Fig. 4. Abundance of CYP2C19 to CYP2C9 WT amino acid swaps. a) Dot plot of CYP2C19 abundance scores at positions that differ between CYP2C19 and 2C9. Each variant represents the abundance of CYP2C19 with the CYP2C9 WT amino acid installed. The WT amino acids for CYP2C19 and CYP2C9 are shown above and below the position. Dots are colored by CYP2C19 abundance classification as shown in the legend. Circles represent abundance scores derived from the VAMP-seq, and squares are individual GFP/mCherry fluorescence measurements normalized to CYP2C19 WT. Error bars show the SD of VAMP-seq abundance scores or the SD of the geometric means of GFP/mCherry fluorescence across 3–4 technical replicates ($n = 50,000$ cells per experiment). All points have error bars, but some are smaller than the points. b) CYP2C19 (PDB: 4GQS) and CYP2C9 (PDB: 1OG2) crystal structures. Positions 241 and 288 are shown as sticks and elements are colored (carbon: black, nitrogen: blue, oxygen: red). c) Bar plot of individually measured GFP/mCherry fluorescence for CYP2C19 (blue) and CYP2C9 (orange) variants. Each sample represents the geometric mean of GFP/mCherry fluorescence. Error bars show the SD across 3–4 technical replicates ($n > 50,000$ cells per experiment). Fluorescence of each variant is normalized to its respective WT CYP. d) Dot plot showing FoldX predicted $\Delta\Delta G$ values for CYP2C19 → CYP2C9 (left) and CYP2C9 → CYP2C19 (right) amino acid swaps. The dotted line represents the predicted $\Delta\Delta G$ at 0.5 abundance for CYP2C19 (2.96 kcal/mol) and CYP2C9 (2.70 kcal/mol) based on a linear model fit to predicted $\Delta\Delta G \sim$ abundance score for all single amino acid variants in each homolog. Color denotes whether the predicted $\Delta\Delta G$ value was negative (stabilizing), positive and below the predicted $\Delta\Delta G$ value at 0.5 abundance (destabilizing) or above $\Delta\Delta G$ at 0.5 abundance (severely destabilizing).

Next, to incorporate position 289 into our analysis, we measured the abundance of the CYP2C19 → CYP2C9 and CYP2C9 → CYP2C19 241, 288, 289 triple mutants (Fig. 4c). The CYP2C19 → CYP2C9 triple mutant had a modestly reduced abundance relative to CYP2C19 WT, largely restoring the low abundance of the E241K, V288E double mutant. The CYP2C9 → CYP2C19 triple mutant had an abundance equivalent to CYP2C9 WT and to each of the 2 double mutants. Notably, none of the CYP2C9 → CYP2C19 swaps increased abundance. Thus, the interaction between these 3 positions is complex and would require direct measurements of thermodynamic stability to fully elucidate.

To shed light on how our measurements of steady-state abundance relate to predicted protein stability, we employed FoldX to calculate Gibbs free energy ($\Delta\Delta G$) of single amino acid variants relative to WT in CYP2C19 and CYP2C9 (Delgado et al. 2019; Schymkowitz et al. 2005). Then, we compared each variant's predicted $\Delta\Delta G$ to its corresponding abundance score (Supplementary Fig. 12; CYP2C19 Pearson's $R = -0.46$, CYP2C9 Pearson's $R = -0.42$). While modest, the correlations between predicted $\Delta\Delta G$ values and abundance scores were similar to other comparisons of *in silico* predicted and experimentally determined variant effects (Gerasimavicius et al. 2023). Next, we fit linear regression models and used each model to predict $\Delta\Delta G$ at an abundance score of 0.5 for CYP2C19 and CYP2C9 (Supplementary Fig. 12). Both values were similar (CYP2C19: 2.96 kcal/mol, CYP2C9: 2.70 kcal/mol; Supplementary Fig. 12a and b), making it difficult to draw general conclusions about each protein's thermodynamic stability, given the modest correlation between predicted $\Delta\Delta G$ s and measured abundances.

Next, we used FoldX $\Delta\Delta G$ s of CYP2C19 → CYP2C9 and CYP2C9 → CYP2C19 swaps to investigate the interaction of positions 241, 288, and 289 in CYP2C19 vs CYP2C9 (Fig. 4d). The pattern of predicted $\Delta\Delta G$ values was similar to the CYP2C19 → CYP2C9 variant abundance scores (Fig. 4c and d, left panels). The abundance scores of V288E-I289N and E241K-V288E-I289N were reduced, but not as severely as predicted by $\Delta\Delta G$ (Fig. 4c and d, left panels). However, the pattern of predicted $\Delta\Delta G$ values was not similar to the CYP2C9 → CYP2C19 variant abundance scores. Here, the $\Delta\Delta G$ values of the single amino acid swaps and the K241E-N289I double mutant were in agreement with abundance scores (Fig. 4c and d, right panels), but the remaining double and triple mutants had highly destabilizing $\Delta\Delta G$ predictions despite WT-like abundance scores (Fig. 4c and d, right panels). These results could arise from inaccuracies in the FoldX predictions or they may suggest that CYP2C9 can conformationally accommodate the sterically unfavorable interactions introduced by the CYP2C9 → CYP2C19 swaps. Thus, while our analysis did not reveal a clear explanation for the structural role of this region in each homolog, the sequence changes in this region of the protein seem likely to contribute to the thermodynamic stability of CYP2C9 and CYP2C19.

Annotating human CYP2C19 variants

CYP2C19 variants can increase, decrease, or eliminate an individual's ability to metabolize many important drugs, and knowing variant function can help avoid severe and expensive adverse events (Lazarou et al. 1998; de Vries et al. 2008; Sultana et al. 2013; Goulding et al. 2015; Schmiedl et al. 2018). For example, the antiplatelet drug clopidogrel is activated by CYP2C19. Thus, individuals with deleterious CYP2C19 variants experience reduced or nonexistent benefit from clopidogrel, requiring higher doses or alternative drugs. Genetic testing for CYP2C19 variants prior to clopidogrel treatment is important for avoiding major adverse cardiovascular events (Galli et al. 2021; Pereira et al. 2021). PharmVar is a repository for pharmacogene allelic variation and functional information, including

CYP2C19. Alleles in PharmVar are known as “star alleles,” and annotated using star notation (Sim and Ingelman-Sundberg 2010). For example, CYP2C19*5 refers to R433W. Despite decades of study, 10 of the 39 CYP2C19 star alleles are of uncertain function. Thus, we analyzed the functional effects of CYP2C19 alleles in PharmVar (Supplementary Table 5). All four PharmVar “normal function” alleles had WT-like abundance scores (Fig. 5a). Of the 8 “decreased” and “no function” alleles, 6 were low abundance. The remaining 2 decreased/no function alleles, CYP2C19*6 and CYP2C19*9, were WT-like in abundance. PharmVar lists the CYP2C19*6 allele (R132Q) as “no function” with “definitive” evidence; however, we measured a WT-like abundance score of 1.06 (95% CI 1.09–1.03). This strongly suggests that the CYP2C19*6 allele's loss of function results from disrupted enzymatic activity rather than loss of abundance. Consistent with this interpretation, the CYP2C19*6 allele is intact enough to bind its heme cofactor but has a decreased ability to metabolize substrates and disrupted electron flow from CPR (Ibeanu et al. 1998; Derayea et al. 2020). Allele CYP2C19*9 (R144H) had an abundance score of 0.914 (95% CI 0.956–0.872). Consistent with these results, CYP2C19*9 has WT-like affinity for CPR (Blaisdell et al. 2002), suggesting that it is at least partially folded. However, our abundance results conflict with another, smaller scale VAMP-seq experiment in which CYP2C19*9 was identified as “decreased” abundance (Zhang et al. 2020) with $\sim 50\%$ of WT abundance. We therefore individually validated this variant and reaffirmed its WT-like abundance in our hands (Supplementary Fig. 13). In light of the moderately reduced activity of CYP2C19*9 against mephenytoin, its ability to bind CPR normally, and WT-like abundance in our assay, we suggest that, like CYP2C19*6, CYP2C19*9 has normal abundance but decreased catalytic activity. Overall, 6 of the 8 known loss of function alleles had reduced abundance, and all normal function alleles had WT-like abundance. While variants with WT-like abundance could have low or no function, reflecting the many ways function can be compromised, low abundance variants were always low or no function alleles. Thus, abundance is a powerful method for identifying loss of function variants. Of the 10 PharmVar alleles with uncertain function, 4 were present in our VAMP-seq library, and we found that CYP2C19*30 and CYP2C19*23 had decreased abundance strongly suggesting that they would disrupt drug metabolism (Fig. 5a).

As sequencing and genetic testing are more widely deployed, rare variants with unknown clinical consequences are being identified at an exponentially increasing rate (Fayer et al. 2021). Reflecting this reality, the annotated alleles in the PharmVar database are only a fraction of all of the CYP2C19 variants discovered so far. There are 408 unique CYP2C19 single amino acid variants in the exome database gnomAD v2.1, 390 of which have no CPIC annotation or functional information. We annotated 368 (90.2%) of the variants in gnomAD v2.1 (Fig. 5b) and identified 131 (35.6%) variants with “decreased” abundance and 29 (7.88%) with “nonsense-like” or “possibly nonsense-like” abundance relative to WT, strongly suggesting that these variants have decreased or no function. We annotated 210 (57.0%) variants as “WT-like” or “possibly WT-like,” indicating that these variants may have normal function. However, assessment of enzymatic activity would be needed to definitively determine if these “WT-like” or “possibly WT-like” variants have normal function since variants can eliminate activity without affecting protein stability. These results are broadly consistent with a study that genotyped 2.29 million participants for CYP2C19*2, CYP2C19*3, and CYP2C19*17 alleles. The study discovered that CYP2C19*2 was present in 15.2%, CYP2C19*3 in 0.3%, and CYP2C19*17 in 20.4% of individuals, and nearly 60% had at least 1 of these star alleles (Ionova et al. 2020). Thus, CYP2C19 variants with reduced abundance appear common in the population.

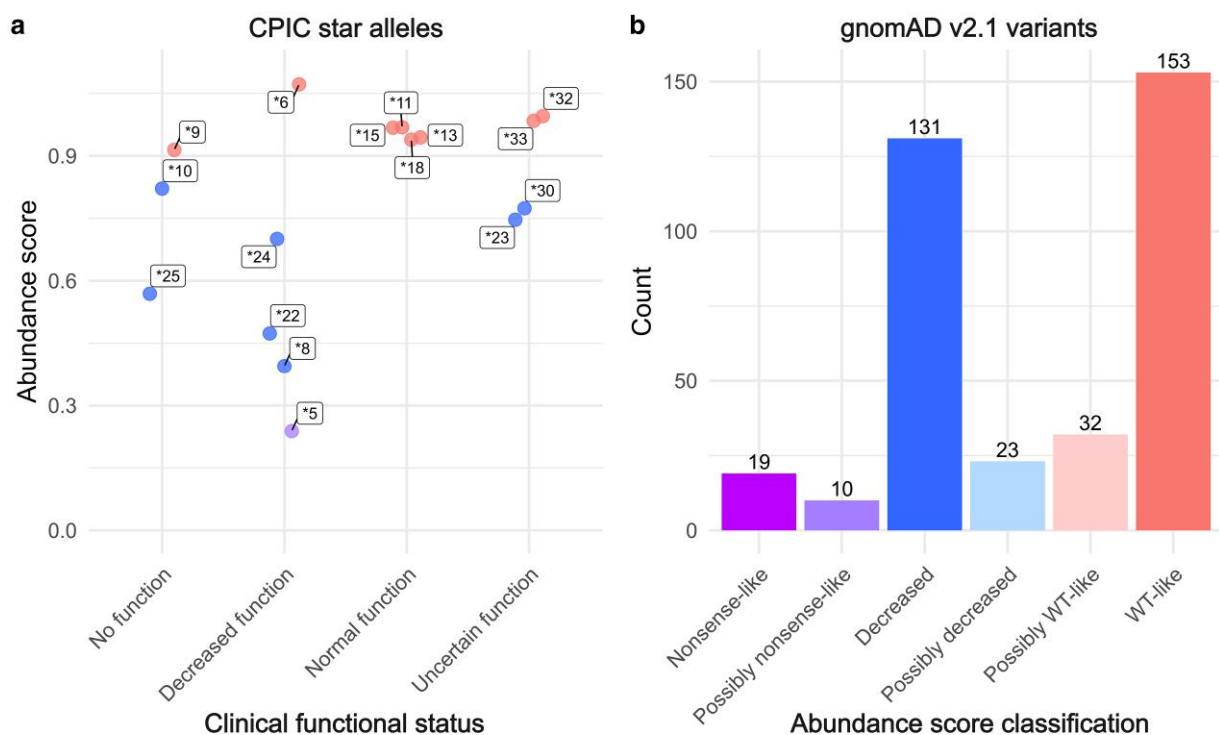


Fig. 5. CYP2C19 abundance scores for variants found in humans. a) Scatter plot of CYP2C19 abundance scores of star (*) alleles with clinical functional status according to the CPIC database (accessed 18 May 2022) (Supplementary Table 5). Dots are colored by abundance score classification and labeled by their star allele designation. b) Bar plot representing abundance score classification of single amino acid variants in gnomAD v2.1 (accessed 18 May 2022).

Discussion

The CYP family tree spans all animal kingdoms and comprises an exceptionally versatile set of enzymes. Understanding the phenotypic consequences of natural variation in human CYPs is particularly important since they catalyze the metabolism of most drugs currently in use. However, even closely related CYPs, like CYP2C19 and CYP2C9, are functionally distinct, and the underlying causes of these distinctions are largely unknown. We used VAMP-seq to measure the abundance of 7,660 CYP2C19 single amino acid variants. In addition to confirming positions known to be critical for CYPs structure and function, we revealed that variants at 4 conserved positions in the hydrophobic core do not impact CYP2C19 abundance. By jointly analyzing 4,670 shared CYP2C19 and CYP2C9 abundance scores, we discovered regions where the 2 homologs have different mutational tolerances. CYP2C9 has a more tolerant hydrophobic core, whereas CYP2C19 is more tolerant in regions surrounding the core. We measured the abundance of WT amino acid swaps between CYP2C19 and CYP2C9, discovering a region likely responsible for at least some of the thermodynamic stability difference between the homologs. Finally, our abundance scores identify known reduced activity CYP2C19 variants with high fidelity, and indicates that 2 star alleles of unknown function, CYP2C19*30 and CYP2C19*23, are likely to have reduced abundance. We also evaluated 368 of the 408 human CYP2C19 variants with no prior annotation. Notably, 43% of these variants are low abundance, warranting follow-up studies to measure their impacts on drug metabolism.

Of the 58 positions conserved across eukaryotic CYPs, 52 had more than 65% reduced abundance variants when substituted with amino acids of a different biophysical type. The remaining 6 were surprisingly mutationally tolerant. The conservation and tolerance of positions 136 and 322 can be explained by their location on the surface of the protein and are likely involved in binding

cofactor CPR, as they do in the closely related CYP2C9 (Berka et al. 2011; Lertkiatmongkol et al. 2013). However, positions, 297, 300, 301, and 362, were tolerant to mutations despite being in the hydrophobic core where mutations are nearly always deleterious. These positions impact substrate specificity of many drugs in CYP2C9 including warfarin, flurbiprofen, and acetaminophen, and positions 297 and 301 have low mutational tolerance as measured by click-seq (Polgár et al. 2007; Peng et al. 2008; Reynald et al. 2012; Amorosi et al. 2021). The significance of these positions in CYP2C19 lies not only in their impact on substrate specificity but also in their specialized role, as they primarily influence specific functions rather than overall protein abundance.

We also jointly analyzed CYP2C19 and CYP2C9 (Amorosi et al. 2021) abundance scans using multidms (Haddox et al. 2023) to find variants with different abundances. Variants in the K'-helix reduced abundance in CYP2C19 but were tolerated in CYP2C9, suggesting a markedly different mutational tolerance of K' in the enzymes despite having identical WT amino acids. Moreover, CYP2C9 had a more tolerant hydrophobic core than CYP2C19, especially in SRSs that contain heme-associated positions. The higher mutational tolerance of CYP2C9 in its core may indicate more flexibility. We speculate that, since the flexibility of CYP active sites is correlated with its promiscuity (Skopalik et al. 2008; Nair et al. 2016), this may allow CYP2C9 to bind more substrates (Wishart et al. 2018).

Variants capable of imparting novel function, like those that alter substrate specificity, often reduce thermodynamic stability (DePristo et al. 2005). To determine if the altered substrate profile and lower thermodynamic stability of CYP2C9 relative to CYP2C19 (Thomson 2021) constituted such a tradeoff, we analyzed the 43 divergent positions between CYP2C19 and CYP2C9. We found that positions 241, 288, and 289 are a likely locus of such a tradeoff because these 3 positions impact substrate specificity (Jung et al. 1998; Klose et al. 1998; Attia et al. 2014), and they are

also adjacent in the structure of both enzymes. Position 288 was the only position in CYP2C19 where installing the CYP2C9 amino acid caused profound loss of abundance, and combining it with swaps at 241 and 289 revealed that these sites have distinct interactions in each homolog. Thus, we hypothesize that these positions are partially responsible for the difference in thermodynamic stability. The CYP2C19 V288E substitution likely causes loss of abundance because it places a negative charge adjacent to E241, reflected by its high predicted $\Delta\Delta G$. Likewise, CYP2C9 K241E introduces the same opposing negative charge adjacent to E288, increasing its predicted $\Delta\Delta G$ and reducing its abundance. This pattern of destabilization and loss of abundance suggests that CYP2C9 may have evolved from CYP2C19. This is because CYP2C9 is the only enzyme in the family that has a negatively charged amino acid at 288, meaning that the ancestral sequence had valine at position 288 (Lewis et al. 1998). Thus, the ancestral CYP2C9 likely acquired E241K or I289N first, both of which partially ameliorate the loss of abundance induced by V288E. We did not find that any combination of swaps at these 3 positions could fully restore CYP2C19 V288E abundance. One possibility is that swaps at other sites, which by themselves do not affect CYP2C19 abundance, could fully rescue V288E. However, the reduced abundance of the CYP2C19 E241K-V288E-I289N variant is in line with the reduced thermodynamic stability of CYP2C9 (Thomson 2021). The new substrate binding capabilities of CYP2C9 apparently made this loss of thermodynamic stability evolutionarily tolerable.

Our interpretation of the evolutionary substrate specificity-abundance tradeoff suggested by our study is partially based on the homologs having different thermodynamic stabilities (Thomson 2021). However, some of our findings are inconsistent with CYP2C19 having a higher thermodynamic stability. Higher thermodynamic stability often confers increased mutational tolerance (Tokuriki and Tawfik 2009; Hormoz 2013; Starr and Thornton 2016). In our multidms analysis, we found that the same variants were more deleterious in CYP2C19 than in CYP2C9. Moreover, our *in silico* analysis suggested that both proteins had similar thermodynamic stabilities. Since the thermodynamic stability difference between CYP2C19 and CYP2C9 was measured in bacteria and required small alterations in amino acid sequence for expression, it is possible that the finding does not apply to WT CYP2C19 and CYP2C9 expressed in mammalian cells (Thomson 2021). Measurements of the thermodynamic stability of CYP2C19 and CYP2C9 in their native cellular context will be required to definitively resolve these conflicting results.

In the gnomAD v2.1 database, 408 single amino acid variants with no prior clinical annotation were identified in the population. We produced abundance scores for 368 of these variants, finding that 43% had reduced abundance. Since CYP abundance is rarely measured in clinical studies, determining precise abundance score thresholds that cause an observable phenotype in patients is not yet possible. The large proportion of reduced abundance variants we identified among single amino acid variants in gnomAD highlights the need for population-level studies that relate CYP2C19 abundance to clinical phenotypes.

Importantly, our VAMP-seq derived abundance data have limitations. First, VAMP-seq has limited resolution for detecting increased abundance variants. Since our analysis focused on detecting deleterious variant effects, we designed our FACS gating using a quartile scheme. As a result, variants with increased abundance may have been grouped with WT-like variants in the highest fluorescence bins. Moreover, variants that increased stability in the already-stable CYP2C19 protein may not have produced an increase in its abundance. Thus, we cannot make conclusions about the lack of variants with increased abundance in our findings.

Other limitations to VAMP-seq arise from the expression system itself. First, the abundance of each variant arises from the balance between protein synthesis and degradation driven by cellular protein quality control systems. All variants are expressed transgenically from an inducible promoter, and an mCherry reporter expressed via an IRES is used to control for cell-to-cell variation in expression. Thus, VAMP-seq derived abundance scores largely reflect changes in degradation which are, in turn, generally driven by stability-related changes in protein folding (Matreyek et al. 2018; Suiter et al. 2020; Zutz et al. 2021; Christensen et al. 2023). VAMP-seq derived abundance scores have a linear relationship to protein abundance measured by western blot (Matreyek et al. 2018; Chiasson et al. 2020). However, abundance changes can result from other mechanisms such as changes to degron sequences or protein localization. Moreover, since we analyzed single amino acid variants, we did not investigate abundance changes caused across codons encoding the same synonymous, nonsense, or missense variant. As a result, some features of our dataset are confusing. For example, the synonymous variant distribution in VAMP-seq experiments sometimes has median abundance slightly different than WT, perhaps due to the effect of codon optimality (Angov 2011; Al-Hawash et al. 2017; Matreyek et al. 2018; Chiasson et al. 2020; Liu et al. 2021). Another unexpected feature is that some nonsense variants at the C-terminus of CYP2C19 have profound but incomplete loss of GFP signal. This is puzzling since GFP is downstream of CYP2C19 in our construct. Alternate start sites or transcriptional readthrough may have driven low, but detectable, GFP expression in these nonsense variants (Reyes and Huber 2018; Caldas et al. 2024). Second, we expressed the CYP2C19 cDNA from an inducible promoter, meaning we cannot detect variants that induce splicing defects or affect transcriptional regulation. Third, variants can affect function without affecting abundance. For example, variants may disrupt a critical substrate binding position or prohibit binding to critical cofactors like CPR or cytochrome b_5 . Therefore, while variants that we identified with reduced abundance are likely to alter drug metabolism, variants with WT-like abundance may not necessarily have normal function. Finally, the VAMP-seq assay depends on fluorescent reporters and fluorescence activated cell sorting. As a result, subtle changes in abundance are difficult to discern.

In the future, we envision intersecting mutational scans from important CYPs in other subfamilies such as CYP2D6 and CYP3A4. Since multidms is capable of jointly analyzing more than two scans, it will be a powerful tool to compare mutational effects across additional CYPs, helping us understand the extent to which variant effects are conserved across the family. In addition to improving personalized drug dosing, such comprehensive profiling could expand our overall understanding of CYP function.

Data availability

The accession number for the sequencing data reported in this article is NCBI GEO: GSE244489. The CYP2C19 VAMP-seq abundance score sets are available on MaveDB under accession number: urn:mavedb:00001199-a. The CYP2C9 VAMP-seq abundance score sets are available on MaveDB under accession number: urn:mavedb:00000095-b. Code and processed variant scores generated during this study are available at GitHub: https://github.com/FowlerLab/cyp2c19_2c9. multidms analyses and data are available at GitHub: <https://github.com/matsengrp/CYP-multidms>.

Supplemental material available at GENETICS online.

Funding

This work was supported by the NIH (5R01GM132162-04, 5RM1HG010461-05, 3UM1HG011969-03). G.E.B. was supported by the National Human Genome Research Institute Interdisciplinary Training in Genome Sciences (T32HG000035). W.S.D. was supported by a Fellowship in Understanding Dynamic and Multi-scale Systems from the James S. McDonnell Foundation. R.C.G. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award F32 GM143852 and by the Momental Foundation. A.E.R. was also supported by the NIH under award number P01GM116691. R.K.G. was supported by the NIH under award number 5RM1HG010461-05 and by the Washington Research Foundation Postdoctoral Fellowship. A.F.R. was supported by the NIH under award number 5RM1HG010461-05 and 5RM1HG010461-05, and this project received grant funding from the Australian Government. F.A.M., J.G.G., and H.K.H. were supported by the NIH under award number R01 AI146028. F.A.M. is an investigator of the Howard Hughes Medical Institute. We thank C. Lee of the UW Foegen Flow Lab, and X. Wu, A. Silvestroni, and J. Fredrickson of the UW Pathology Flow Cytometry Core Facility for assistance with cell sorting and all members of the Fowler lab for helpful feedback on figures.

Conflicts of interest

The authors declare no conflict of interests.

Literature cited

- Al-Hawash AB, Zhang X, Ma F. 2017. Strategies of codon optimization for high-level heterologous protein expression in microbial expression systems. *Gene Rep.* 9:4653. doi:10.1016/j.genrep.2017.08.006.
- Altarsha M, Benighaus T, Kumar D, Thiel W. 2009. How is the reactivity of cytochrome P450cam affected by Thr252X mutation? A QM/MM study for X = serine, valine, alanine, glycine. *J Am Chem Soc.* 131(13):4755–4763. doi:10.1021/ja808744k.
- Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, Kowalski JP, Rettie AE, Fowler DM, Dunham MJ. 2021. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. *Am J Hum Genet.* 108(9):1735–1751. doi:10.1016/j.ajhg.2021.07.001.
- Angov E. 2011. Codon usage: nature's roadmap to expression and folding of proteins. *Biotechnol J.* 6(6):650–659. doi:10.1002/biot.201000332.
- Attia TZ, Yamashita T, Hammad MA, Hayasaki A, Sato T, Miyamoto M, Yasuhara Y, Nakamura T, Kagawa Y, Tsujino H, et al. 2014. Effect of cytochrome P450 2C19 and 2C9 amino acid residues 72 and 241 on metabolism of tricyclic antidepressant drugs. *Chem Pharm Bull (Tokyo).* 62(2):176–181. doi:10.1248/cpb.c13-00800.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
- Berka K, Hendrychová T, Anzenbacher P, Otyepka M. 2011. Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *J Phys Chem A.* 115(41):11248–11255. doi:10.1021/jp204488j.
- Blaisdell J, Mohrenweiser H, Jackson J, Ferguson S, Coulter S, Chanas B, Xi T, Ghanayem B, Goldstein JA. 2002. Identification and functional characterization of new potentially defective alleles of human CYP2C19. *Pharmacogenetics.* 12(9):703–711. doi:10.1097/00008571-200212000-00004.
- Caldas P, Luz M, Baseggio S, Andrade R, Sobral D, Grosso AR. 2024. Transcription readthrough is prevalent in healthy human tissues and associated with inherent genomic features. *Commun Biol.* 7(1):100. doi:10.1038/s42003-024-05779-5.
- Chiasson MA, Rollins NJ, Stephany JJ, Sitko KA, Matreyek KA, Verby M, Sun S, Roth FP, DeSloover D, Marks DS, et al. 2020. Multiplexed measurement of variant abundance and activity reveals VKOR topology, active site and human variant impact. *eLife.* 9:e58026. doi:10.7554/eLife.58026.
- Christensen S, Wernersson C, André I. 2023. Facile method for high-throughput identification of stabilizing mutations. *J Mol Biol.* 435(18):168209. doi:10.1016/j.jmb.2023.168209.
- Coon MJ. 2005. Cytochrome P450: nature's most versatile biological catalyst. *Annu Rev Pharmacol Toxicol.* 45(1):1–25. doi:10.1146/annurev.pharmtox.45.120403.100030.
- Dean L, Kane M. 2022. Clopidogrel Therapy and CYP2C19 Genotype. Bethesda (MD): National Center for Biotechnology Information (US).
- Delgado J, Radusky LG, Cianferoni D, Serrano L. 2019. Foldx 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics.* 35(20):4168–4169. doi:10.1093/bioinformatics/btz184.
- Denisov IG, Makris TM, Sligar SG, Schlichting I. 2005. Structure and chemistry of cytochrome P450. *Chem Rev.* 105(6):2253–2277. doi:10.1021/cr0307143.
- DePristo MA, Weinreich DM, Hartl DL. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet.* 6(9):678–687. doi:10.1038/nrg1672.
- Derayea SM, Tsujino H, Oyama Y, Ishikawa Y, Yamashita T, Uno T. 2020. Impact of single nucleotide polymorphisms (R132Q and W120R) on the binding affinity and metabolic activity of CYP2C19 toward some therapeutically important substrates. *Xenobiotica.* 50(12):1510–1519. doi:10.1080/00498254.2020.1786189.
- de Vries EN, Ramrattan MA, Smorenburg SM, Gouma DJ, Boermeester MA. 2008. The incidence and nature of in-hospital adverse events: a systematic review. *Qual Saf Health Care.* 17(3):216–223. doi:10.1136/qshc.2007.023622.
- Fayer S, Horton C, Dines JN, Rubin AF, Richardson ME, McGoldrick K, Hernandez F, Pesaran T, Karam R, Shirts BH, et al. 2021. Closing the gap: systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN. *Am J Hum Genet.* 108(12):2248–2258. doi:10.1016/j.ajhg.2021.11.001.
- Foti RS, Rock DA, Han X, Flowers RA, Wienkers LC, Wahlstrom JL. 2012. Ligand-based design of a potent and selective inhibitor of cytochrome P450 2C19. *J Med Chem.* 55(3):1205–1214. doi:10.1021/jm201346g.
- Galli M, Benenati S, Capodanno D, Franchi F, Rollini F, D'Amario D, Porto I, Angiolillo DJ. 2021. Guided versus standard antiplatelet therapy in patients undergoing percutaneous coronary intervention: a systematic review and meta-analysis. *Lancet.* 397(10283):1470–1483. doi:10.1016/S0140-6736(21)00533-X.
- García-Nafria J, Watson JF, Greger IH. 2016. IVA cloning: a single-tube universal cloning system exploiting bacterial In Vivo Assembly. *Sci Rep.* 6(1):27459. doi:10.1038/srep27459.
- Gerasimavicius L, Livesey BJ, Marsh JA. 2023. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. *Protein Sci.* 32(7):e4688. doi:10.1002/pro.4688.
- Goldstein JA, de Morais SM. 1994. Biochemistry and molecular biology of the human CYP2C subfamily. *Pharmacogenetics.* 4(6):285–299. doi:10.1097/00008571-199412000-00001.

- Gotoh O. 1992. Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences. *J Biol Chem.* 267(1): 83–90. doi:10.1016/S0021-9258(18)48462-1.
- Goulding R, Dawes D, Price M, Wilkie S, Dawes M. 2015. Genotype-guided drug prescribing: a systematic review and meta-analysis of randomized control trials. *Br J Clin Pharmacol.* 80(4):868–877. doi:10.1111/bcp.12475.
- Gricman Ł, Vogel C, Pleiss J. 2014. Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. *Proteins.* 82(3):491–504. doi:10.1002/prot.24415.
- Gricman Ł, Vogel C, Pleiss J. 2015. Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. *Proteins.* 83(9): 1593–1603. doi:10.1002/prot.24840.
- Gumulya Y, Baek J-M, Wun S-J, Thomson RES, Harris KL, Hunter DJB, Behrendorff JBYH, Kulig J, Zheng S, Wu X, et al. 2018. Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat Catal.* 1(11):878–888. doi:10.1038/s41929-018-0159-5.
- Haddox HK, Galloway JG, Dadonaite B, Bloom JD, Matsen FA, DeWitt WS. 2023. Jointly modeling deep mutational scans identifies shifted mutational effects among SARS-CoV-2 spike homologs. *bioRxiv.* <https://doi.org/10.1101/2023.07.31.551037>, preprint: not peer reviewed.
- Haines DC, Tomchick DR, Machius M, Peterson JA. 2001. Pivotal role of water in the mechanism of P450BM-3. *Biochemistry.* 40(45): 13456–13465. doi:10.1021/bi011197q.
- Hargrove JL, Schmidt FH. 1989. The role of mRNA and protein stability in gene expression. *FASEB J.* 3(12):2360–2370. doi:10.1096/fasebj.3.12.2676679.
- Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J. 1995. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure.* 3(1):41–62. doi:10.1016/S0969-2126(01)00134-4.
- Hormoz S. 2013. Amino acid composition of proteins reduces deleterious impact of mutations. *Sci Rep.* 3(1):2919. doi:10.1038/srep02919.
- Ibeanu GC, Ghanayem BI, Linko P, Li L, Pedersen LG, Goldstein JA. 1996. Identification of residues 99, 220, and 221 of human cytochrome P450 2C19 as key determinants of omeprazole activity. *J Biol Chem.* 271(21):12496–12501. doi:10.1074/jbc.271.21.12496.
- Ibeanu GC, Goldstein JA, Meyer U, Benhamou S, Bouchardy C, Dayer P, Ghanayem BI, Blaisdell J. 1998. Identification of new human CYP2C19 alleles (CYP2C19*6 and CYP2C19*2B) in a Caucasian poor metabolizer of mephenytoin. *J Pharmacol Exp Ther.* 286(3): 1490–1495.
- Ionova Y, Ashenhurst J, Zhan J, Nhan H, Kosinski C, Tamraz B, Chubb A. 2020. CYP2C19 allele frequencies in over 2.2 million direct-to-consumer genetics research participants and the potential implication for prescriptions in a large health system. *Clin Transl Sci.* 13(6):1298–1306. doi:10.1111/cts.12830.
- Jain PC, Varadarajan R. 2014. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Anal Biochem.* 449:90–98. doi:10.1016/j.ab.2013.12.002.
- Jung F, Griffin KJ, Song W, Richardson TH, Yang M, Johnson EF. 1998. Identification of amino acid substitutions that confer a high affinity for sulfaphenazole binding and a high catalytic efficiency for warfarin metabolism to P450 2C19. *Biochemistry.* 37(46): 16270–16279. doi:10.1021/bi981704c.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birbaum DP. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 581(7809):434–877.
- Kim I, Miller CR, Young DL, Fields S. 2013. High-throughput analysis of in vivo protein stability. *Mol Cell Proteomics.* 12(11):3370–3378. doi:10.1074/mcp.O113.031708.
- Klein MD, Lee CR, Stouffer GA. 2018. Clinical outcomes of CYP2C19 genotype-guided antiplatelet therapy: existing evidence and future directions. *Pharmacogenomics.* 19(13):1039–1046. doi:10.2217/pgs-2018-0072.
- Klesmith JR, Bacik J-P, Wrenbeck EE, Michalczyk R, Whitehead TA. 2017. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci U S A.* 114(9):2265–2270. doi:10.1073/pnas.1614437114.
- Klose TS, Ibeanu GC, Ghanayem BI, Pedersen LG, Li L, Hall SD, Goldstein JA. 1998. Identification of residues 286 and 289 as critical for conferring substrate specificity of human CYP2C9 for diclofenac and ibuprofen. *Arch Biochem Biophys.* 357(2):240–248. doi:10.1006/abbi.1998.0826.
- Lazarou J, Pomeranz BH, Corey PN. 1998. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *JAMA.* 279(15):1200–1205. doi:10.1001/jama.279.15.1200.
- Lertkiatmongkol P, Assawamakin A, White G, Chopra G, Rongnoparut P, Samudrala R, Tongsima S. 2013. Distal effect of amino acid substitutions in CYP2C9 polymorphic variants causes differences in interatomic interactions against (S)-warfarin. *PLoS One.* 8(9):e74053. doi:10.1371/journal.pone.0074053.
- Lewis DF, Dickins M, Weaver RJ, Eddershaw PJ, Goldfarb PS, Tarbit MH. 1998. Molecular modelling of human CYP2C subfamily enzymes CYP2C9 and CYP2C19: rationalization of substrate specificity and site-directed mutagenesis experiments in the CYP2C subfamily. *Xenobiotica.* 28(3):235–268. doi:10.1080/004982598239542.
- Liu Y, Yang Q, Zhao F. 2021. Synonymous but not silent: the codon usage code for gene expression and protein folding. *Annu Rev Biochem.* 90(1):375–401. doi:10.1146/annurev-biochem-071320-112701.
- Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, Kircher M, Khechaduri A, Dines JN, Hause RJ, et al. 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat Genet.* 50(6):874–882. doi:10.1038/s41588-018-0122-z.
- Matreyek KA, Stephany JJ, Chiasson MA, Hasle N, Fowler DM. 2020. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* 48(1):e1. doi:10.1093/nar/gkz910.
- Matreyek KA, Stephany JJ, Fowler DM. 2017. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* 45(11):e102. doi:10.1093/nar/gkx183.
- Mestres J. 2005. Structure conservation in cytochromes P450. *Proteins.* 58(3):596–609. doi:10.1002/prot.20354.
- Munro AW, Girvan HM, Mason AE, Dunford AJ, McLean KJ. 2013. What makes a P450 tick? *Trends Biochem Sci.* 38(3):140–150. doi:10.1016/j.tibs.2012.11.006.
- Mustafa G, Nandekar PP, Bruce NJ, Wade RC. 2019. Differing membrane interactions of two highly similar drug-metabolizing cytochrome P450 isoforms: CYP 2C9 and CYP 2C19. *Int J Mol Sci.* 20(18):4328. doi:10.3390/ijms20184328.
- Nair PC, McKinnon RA, Miners JO. 2016. Cytochrome P450 structure-function: insights from molecular dynamics simulations. *Drug Metab Rev.* 48(3):434–452. doi:10.1080/03602532.2016.1178771.

- Nelson DR. 2011. Progress in tracing the evolutionary paths of cytochrome P450. *Biochim Biophys Acta*. 1814(1):14–18. doi:[10.1016/j.bbapap.2010.08.008](https://doi.org/10.1016/j.bbapap.2010.08.008).
- Niwa T, Kageyama A, Kishimoto K, Yabusaki Y, Ishibashi F, Katagiri M. 2002. Amino acid residues affecting the activities of human cytochrome P450 2C9 and 2C19. *Drug Metab Dispos*. 30(8):931–936. doi:[10.1124/dmd.30.8.931](https://doi.org/10.1124/dmd.30.8.931).
- Niwa T, Yamazaki H. 2012. Comparison of cytochrome P450 2C subfamily members in terms of drug oxidation rates and substrate inhibition. *Curr Drug Metab*. 13(8):1145–1159. doi:[10.2174/138920012802850092](https://doi.org/10.2174/138920012802850092).
- Peng C-C, Cape JL, Rushmore T, Crouch GJ, Jones JP. 2008. Cytochrome P450 2C9 type II binding studies on quinoline-4-carboxamide analogues. *J Med Chem*. 51(24):8000–8011. doi:[10.1021/jm8011257](https://doi.org/10.1021/jm8011257).
- Pereira NL, Rihal C, Lennon R, Marcus G, Shrivastava S, Bell MR, So D, Geller N, Goodman SG, Hasan A, et al. 2021. Effect of CYP2C19 genotype on ischemic outcomes during oral P2Y12 inhibitor therapy: a meta-analysis. *JACC Cardiovasc Interv*. 14(7):739–750. doi:[10.1016/j.jcin.2021.01.024](https://doi.org/10.1016/j.jcin.2021.01.024).
- Polgár T, Menyhárd DK, Keserű GM. 2007. Effective virtual screening protocol for CYP2C9 ligands using a screening site constructed from flurbiprofen and S-warfarin pockets. *J Comput Aided Mol Des*. 21(9):539–548. doi:[10.1007/s10822-007-9137-8](https://doi.org/10.1007/s10822-007-9137-8).
- Relling MV, Klein TE. 2011. CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clin Pharmacol Ther*. 89(3):464–467. doi:[10.1038/clpt.2010.279](https://doi.org/10.1038/clpt.2010.279).
- Reyes A, Huber W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res*. 46(2):582–592. doi:[10.1093/nar/gkx1165](https://doi.org/10.1093/nar/gkx1165).
- Reynald RL, Sansen S, Stout CD, Johnson EF. 2012. Structural characterization of human cytochrome P450 2C19: active site differences between P450s 2C8, 2C9, and 2C19. *J Biol Chem*. 287(53):44581–44591. doi:[10.1074/jbc.M112.424895](https://doi.org/10.1074/jbc.M112.424895).
- Schmiedl S, Rottenkolber M, Szymanski J, Drewelow B, Siegmund W, Hippus M, Farker K, Guenther IR, Hasford J, Thuermann PA, et al. 2018. Preventable ADRs leading to hospitalization—results of a long-term prospective safety study with 6,427 ADR cases focusing on elderly patients. *Expert Opin Drug Saf*. 17(2):125–137. doi:[10.1080/14740338.2018.1415322](https://doi.org/10.1080/14740338.2018.1415322).
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res*. 33(Web Server):W382–W388. doi:[10.1093/nar/gki387](https://doi.org/10.1093/nar/gki387).
- Sim SC, Ingelman-Sundberg M. 2010. The human cytochrome P450 (CYP) allele nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum Genomics*. 4(4):278–281. doi:[10.1186/1479-7364-4-4-278](https://doi.org/10.1186/1479-7364-4-4-278).
- Sirim D, Widmann M, Wagner F, Pleiss J. 2010. Prediction and analysis of the modular structure of cytochrome P450 monooxygenases. *BMC Struct Biol*. 10(1):34. doi:[10.1186/1472-6807-10-34](https://doi.org/10.1186/1472-6807-10-34).
- Skopalík J, Anzenbacher P, Otyepka M. 2008. Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J Phys Chem B*. 112(27):8165–8173. doi:[10.1021/jp800311c](https://doi.org/10.1021/jp800311c).
- Starr TN, Thornton JW. 2016. Epistasis in protein evolution. *Protein Sci*. 25(7):1204–1218. doi:[10.1002/pro.2897](https://doi.org/10.1002/pro.2897).
- Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, Yang W, Hoshitsuki K, Singh M, Trehan A, et al. 2020. Massively parallel variant characterization identifies *NUDT15* alleles associated with thiopurine toxicity. *Proc Natl Acad Sci U S A*. 117(10):5394–5401. doi:[10.1073/pnas.1915680117](https://doi.org/10.1073/pnas.1915680117).
- Sultana J, Cutroneo P, Trifirò G. 2013. Clinical and economic burden of adverse drug reactions. *J Pharmacol Pharmacother*. 4(Suppl 1):S73–S77. doi:[10.4103/0976-500X.120957](https://doi.org/10.4103/0976-500X.120957).
- Thomson R. 2021. Structural and functional characterisation of ancestral cytochromes P450 from family 2 in tetrapods [PhD thesis]. The University of Queensland: School of Chemistry and Molecular Biosciences. doi:[10.14264/a159633](https://doi.org/10.14264/a159633).
- Tokuriki N, Tawfik DS. 2009. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 19(5):596–604. doi:[10.1016/j.sbi.2009.08.003](https://doi.org/10.1016/j.sbi.2009.08.003).
- Tsao CC, Wester MR, Ghanayem B, Coulter SJ, Chanas B, Johnson EF, Goldstein JA. 2001. Identification of human CYP2C19 residues that confer S-mephenytoin 4'-hydroxylation activity to CYP2C9. *Biochemistry*. 40(7):1937–1944. doi:[10.1021/bi001678u](https://doi.org/10.1021/bi001678u).
- Wada Y, Mitsuda M, Ishihara Y, Watanabe M, Iwasaki M, Asahi S. 2008. Important amino acid residues that confer CYP2C19 selective activity to CYP2C9. *J Biochem*. 144(3):323–333. doi:[10.1093/jb/mvn070](https://doi.org/10.1093/jb/mvn070).
- Werck-Reichhart D, Feyereisen R. 2000. Cytochromes P450: a success story. *Genome Biol*. 1(6):REVIEWS3003. doi:[10.1186/gb-2000-1-6-reviews3003](https://doi.org/10.1186/gb-2000-1-6-reviews3003).
- Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 46(D1):D1074–D1082. doi:[10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037).
- Yeh C-LC, Amorosi CJ, Showman S, Dunham MJ. 2022. PacRAT: a program to improve barcode-variant mapping from PacBio long reads using multiple sequence alignment. *Bioinformatics*. 38(10):2927–2929. doi:[10.1093/bioinformatics/btac165](https://doi.org/10.1093/bioinformatics/btac165).
- Yen H-CS, Xu Q, Chou DM, Zhao Z, Elledge SJ. 2008. Global protein stability profiling in mammalian cells. *Science* (1979). 322(5903):918–923. doi:[10.1126/science.1160489](https://doi.org/10.1126/science.1160489).
- Zanger UM, Schwab M. 2013. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther*. 138(1):103–141. doi:[10.1016/j.pharmthera.2012.12.007](https://doi.org/10.1016/j.pharmthera.2012.12.007).
- Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate illumina paired-end reAd mergeR. *Bioinformatics*. 30(5):614–620. doi:[10.1093/bioinformatics/btt593](https://doi.org/10.1093/bioinformatics/btt593).
- Zhang L, Sarangi V, Moon I, Yu J, Liu D, Devarajan S, Reid JM, Kalari KR, Wang L, Weinshilboum R. 2020. CYP2C9 and CYP2C19: deep mutational scanning and functional characterization of genomic missense variants. *Clin Transl Sci*. 13(4):727–742. doi:[10.1111/cts.12758](https://doi.org/10.1111/cts.12758).
- Zhao L, Liu Z, Levy SF, Wu S. 2018. Bartender: a fast and accurate clustering algorithm to count barcode reads. *Bioinformatics*. 34(5):739–747. doi:[10.1093/bioinformatics/btx655](https://doi.org/10.1093/bioinformatics/btx655).
- Zhao M, Ma J, Li M, Zhang Y, Jiang B, Zhao X, Huai C, Shen L, Zhang N, He L, et al. 2021. Cytochrome P450 enzymes and drug metabolism in humans. *Int J Mol Sci*. 22(23):12808. doi:[10.3390/ijms222312808](https://doi.org/10.3390/ijms222312808).
- Zutz A, Hamborg L, Pedersen LE, Kassem MM, Papaleo E, Koza A, Herrgård MJ, Jensen SI, Teilum K, Lindorff-Larsen K, et al. 2021. A dual-reporter system for investigating and optimizing protein translation and folding in *E. coli*. *Nat Commun*. 12(1):6093. doi:[10.1038/s41467-021-26337-1](https://doi.org/10.1038/s41467-021-26337-1).