Using high throughput technologies to understand heterogeneous drug response

Gabriel E. Boyle

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Dr. Douglas Fowler (Chair)

Dr. Emily Hatch

Dr. Brian Beliveau

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Using high throughput technologies to understand heterogeneous drug response

Gabriel E. Boyle

Chair of supervisory committee:

Douglas M. Fowler, Professor

Department of Genome Sciences

Heterogeneity is an inherent quality of biological systems, and is one of the properties that facilitates adaptation of organisms to new environments and stimuli. However, the scientific method relies on probabilistic reasoning wherein sample measurements are assumed to represent the population as a whole. In situations where this assumption is incorrect, as when a population is unknowingly made up of multiple subpopulations, the underlying biology may be obscured, resulting in faulty conclusions. Treatments developed based on these faulty conclusions may, at worst, harm people rather than help them. In Part I, I explore heterogeneity at the organismal level by investigating why drug response differs between people. Chapter 1 describes the possible consequences of assessing medical interventions on heterogeneous populations, and introduces pharmacogenomics as the first step towards personalized medicine. For the goals of pharmacogenomics to be realized, the mutational landscape of cytochrome P450 (CYP) enzymes, which catalyze the elimination or activation of most drugs currently in use, must be comprehensively assayed. In Chapter 2, I describe my contribution to

this goal in measuring the abundance of nearly all single mutations in one CYP protein, 2C19, and jointly analyzing the dataset with a similar one from its closest neighbor, 2C9. These analyses highlight regions where mutational tolerance differs between CYP2C19 and 2C9, despite their structural and sequence similarity, revealing possible mechanisms underlying their functional differences. In Part II, I investigate how morphological heterogeneity manifests in clonal populations of cells when exposed to drugs. Chapter 3 introduces how visual phenotypes are tightly linked to cell state, emphasizing the paucity of technologies that combine microscopy with mass spectrometry-based proteomics. In Chapter 4, I describe my solution to this problem by integrating Visual Cell Sorting (VCS), which separates cells based on visual phenotypes, with low-input mass spectrometry (MS). I demonstrate that this new technology, VCS-MS, separates cells with similar fidelity to fluorescence-assisted cell sorting, and achieves nearly double the proteomic depth of similar technologies. Taken together, this dissertation demonstrates the critical need to develop and employ technologies capable of investigating heterogeneity at all levels of biological systems.

Table of Contents

Acknowledgements

In writing these acknowledgements, I find myself returning to a quote that's held poignant meaning on many occasions in the past 5 years:

Life can only be understood backwards; but it must be lived forwards.

- Søren Kierkegaard

At the conclusion of living graduate school forwards, I can tell you the experience felt ambiguous, uncertain, heavy, galvanizing, grounding, inspiring. Paradoxical. In writing these acknowledgements, I seek to understand my experience, development, and accomplishments in graduate school backwards. In doing so, the major inflection points are clear. Yet the people whose influence altered the pathways quickly sprawl. I'd like to express my heartfelt appreciation to the following people, in the order that felt right on the morning of my dissertation due date.

To my advisor, Doug Fowler, who spent countless meetings with me, looking backwards at his own life and seeing how it mapped onto mine. Reflecting on new analogies and metaphors that helped me understand my own path. Empowering me and trusting me to define a path that fits who I am.

To Jennifer Smith, Molly Goodfellow, and Charlie Campbell who sit very close to the beginning threads of my scientific career. They identified my aptitude for science during my first research experiences. I was confused about a lot of things, and I still am, but I trusted you and you were right.

To Hilary Horsman, whose impact on my life cannot be overstated. We grew into our adulthood together, and taught each other what companionship could be. I never would have left Ohio, let alone moved across the country for grad school, if not for the safety and reliability that being on her team provided.

To Julia Robbins, whose love, penchant for whimsy, and undeniable panache have filled my life to the brim. Understanding life backwards, our first date in front of Agua Verdé, which was the duration of a full workday, speaks for itself.

To the late Rich Gardner, who made me feel at home in MCB. To the MCB cohort of 2018. You taught me what belonging to a community felt like. To Maia Low, our MCB mom. Your authentic warmth, groundedness, and steadfast advocacy have assured me that I always have an ally.

To my parents, who have never wanted anything more for me than my happiness and self-actualization. Mom, your care established the foundation of my life. Though morally dubious, lying to the parish was the right choice. Dad, you have always believed in me, and have helped shape me into the independent, empathetic scientist I am today. To my siblings, who have admirably navigated the re-forming of our relationships as adults. Seth, you always have a place on my pirate ship. Annie, there are few people who have supported me so tangibly and selflessly. Luke, I'm ready for another Delaware camping trip whenever you are.

To my lifelong friend and podcast co-host (check out Thru Shame on Spotify, Apple, or wherever you get your podcasts) Dr. Jonathan Reeves, who pointed out that we've both been scientists at heart since we were 10 years old. Your influence on my graduate career runs as deeply as your structural place in my life.

To the Famigas. Valentina Grillo-Alvarado, whose love knows no bounds. Whose authenticity, open honesty, and devotion to navigating what it means to be human is a rare and infectious quality. Vanessa Nguyen, whose companionship and deep friendship, with all the adjustment along the way, have catalyzed immense growth.

To the people whose teachings have touched the core of my growth, though they don't necessarily know I exist. Kemi Doll, an accomplished, tenured scientist at UW, whose Unapologetic Career (find it wherever you get your podcasts) has helped me find my compass. Beronda Montgomery, who may or may not remember me as one of the more enthusiastic nodders in the back of the auditorium. To Brené Brown, whose emotion research and passion for empathetic leadership has become a pillar of who I am.

Dr. Katherine Burke, who taught me a new way to use the word "and" decades after I first learned it. This new "and" represents a dialectic: Two seemingly opposite truths that create a new truth when considered together. Acceptance and change. Shifting from "I really need to rest, but I don't feel like I've been productive enough" to "I really need to rest, *and* I don't feel like I've been productive enough." The "but" negates the rest, whereas the "and" makes space for it.

For Dr. Patrick Johnson, especially for teaching me that I always have a choice. Sometimes alternative options are appallingly bad, but everything I do is a choice. He taught me that when I find myself in exasperation, asking "what is the point of this?" the critical reaction is to answer the question. What *is* the point of this? Why *am* I doing this? Is this still in line with my goals? Are there alternative viable paths with better trade-offs?

To all the people who helped me define those trade-offs and find clarity in my decisions. Beth Traxler, Steve Perlmutter, and Emily Hatch you were beacons during a time filled with fog. Atom Lesiak, you were always there when I needed you, and were a steadfast source of perspective and support rooted in humanity. You have helped me define who I want to be.

To all of my Brazilian Jiu Jitsu coaches and training partners. You have created a space for high pressure problem solving that is completely orthogonal to my intellectual pursuits, providing a unique type of “rest”. Perhaps most importantly, thank you for trying to strangle me. When fearing for one’s life, thoughts of failed experiments tend to fall to the wayside. At least for a moment.

To my physical therapists, Jenny Katz and Rob Cheng, who saw me very regularly throughout the past 5 years. Physical therapy remains one of my favorite activities. It is a rare experience in life to approach another with a tangible and debilitating problem, be carefully assessed and diagnosed, receive a little massage, do some exercises, and reliably make durable progress toward recovery. I do still question the legitimacy of some of those exercises, however. You never let me leave with my dignity. Not that I had much of that to begin with. I also want to thank Cynthia, who graciously took what little dignity I had left after a session, and always hunted me down for the debts that I owed. It’s also worth thanking the UAW Graduate Student Union for fighting for such excellent healthcare. Never before in my life had I expected to pay a certain amount and had only half requested of me.

Last, but not least, to my housemate and friend, Fink. He kept me humble, and never failed to assert his boundaries. On a daily basis, he lived the dialectic of tenacious curiosity and relentless rest. I cherished every neck scarf he was gracious enough to offer.

Part I - Heterogeneous drug response between individuals

Chapter 1: Cytochrome P450 enzymes are major contributors to pharmacogenomic heterogeneity

1.1 The problem with applying global evidence to local problems

The purpose of the evidence-based medicine movement was to de-emphasize intuition, pathophysiologic rationale, and unsystematic observations as sufficient guides for clinical care. Instead, clinical care was to be informed by reproducible and unbiased scientific evidence (Evidence-Based Medicine Working Group 1992). Ultimately, the evidence-based medicine movement drove a major shift in the standards for gathering evidence and the tools for analyzing it, leveraging the power of probabilistic reasoning to wrangle with the inherent uncertainties of clinical medicine (Davidoff, Case, and Fried 1995).

Probabilistic reasoning employs statistical estimations to discern clinically beneficial treatment effects. Typically, statistical tests assess global averages between a treatment and control group. If the treatment group's global average is unlikely to have occurred by chance, the study supports a treatment effect. This basic framework for statistical testing relies on the assumption that the population the sample data was drawn from is distributed across a single normal distribution. If this assumption is correct, treatment variability is accurately predicted and can be accounted for when the data is used to guide clinical care. However, if the true population contains multiple dominant subpopulations, each with distinct biology, treatment-effect heterogeneity can lead to the administration of ineffective treatments, denying a patient access to effective treatments, or, in the worst case, administering treatments that are actively harmful to the patient (Kravitz, Duan, and Braslow 2004; Longford 1999). The field of personalized medicine ultimately aims to resolve treatment-effect heterogeneity by identifying medically

relevant subpopulations, understanding their distinct medical needs, and tailoring clinical care to each individual patient.

1.2 Personalized medicine tailors treatment to the individual

Personalized medicine is rooted in the belief that individuals have unique characteristics that determine their specific medical needs (“Precision Health: Improving Health for Each of Us and All of Us” 2022; Carlsten et al. 2014). The factors that contribute to these medical needs span across genetics and genomics, geographic and economic access to healthcare, routine monitoring of biomarkers, environmental exposures and behaviors, and epigenetic phenomena (Goetz and Schork 2018). Ideally, when all of these factors are known they can be used to tailor treatment specifically to each individual’s holistic health status. However, many details of a patient’s life and environmental history are inaccessible. Moreover, not all patient characteristics are equally predictive of treatment outcomes. Thus, initiatives for personalized medicine are targeted towards the diseases and patient metrics that have the best prospects of success, like their genetics and genomics (Collins and Varmus 2015).

While imperfect, genetics and genomics have proven foundational to personalized medicine (Goetz and Schork 2018). Innovations in sequencing technology have made collecting genetic information substantially faster and more cost-effective (van Dijk et al. 2014). As a result, patient genotyping is now widely available. However, understanding the connection between patient genotype and disease phenotype remains a central challenge.

1.3 Pharmacogenomics brings precision medicine to drug treatment

Pharmacogenomics, one of the earliest focuses of personalized medicine, aims to identify genetic variation that drives inter-individual variability in drug response (Goetz and Schork

2018). A primary determinant of drug response is its duration and concentration at its site of action in the body, described by its pharmacokinetics. Pharmacokinetics is determined by the drug's rate of absorption, distribution, metabolism, and elimination. Since metabolism and elimination are predominantly driven by enzymatic activity, altered enzymatic abundance and function can result in ineffective drug treatment or cause potentially life-threatening adverse events (de Vries et al. 2008). Thus, understanding the consequences of genetic variation in drug-metabolizing enzymes enables genetic testing prior to drug dosing, which improves efficacy and reduces toxicity (Roden et al. 2019; Roses 2000). Clopidogrel is a prominent example of a drug that benefits from pharmacogenomic-guided care.

Clopidogrel is an antiplatelet medicine that prevents blood clots. As a prodrug, it requires activation in the liver by the heme monooxygenase cytochrome P450 (CYP) 2C19 before it can have a therapeutic effect. The Clinical Pharmacogenetics Implementation Consortium (CPIC) centralizes pharmacogene variant information and provides guidelines for drug treatments. CPIC establishes CYP star (*) alleles and categorizes haplotypes according to enzymatic activity: normal function, decreased function, no function, or increased function (Relling and Klein 2011; Sim and Ingelman-Sundberg 2010). Patients with one inactive allele, like CYP2C19*2, have reduced activation of clopidogrel and require a higher dose. Patients homozygous for CYP2C19*2 receive no benefit from clopidogrel and require a different drug (Mega et al. 2011; Saydam et al. 2017). Thus, guiding treatment according to CYP2C19 genotype can improve patient outcomes (Beitelshees et al. 2022).

The variant CYP2C19*2 is now a well studied loss-of-function allele (Lee et al. 2022). However, the functional consequences of most CYP2C19 variants are unknown, and therefore cannot be used for genotype-guided treatment. Moreover, CYP2C19 is only one of ~12 CYP enzymes that cumulatively metabolize 70-80% of drugs that are eliminated through oxidation (Zanger and

Schwab 2013). Thus, deeply characterizing natural variation in CYPs is a central goal in pharmacogenomics.

1.4 Cytochrome P450 overview

Originating as a single protein, the CYP family now consists of ~20,000 proteins across all domains of life (Nelson 2011). Collectively, CYPs catalyze more than 60 unique biotransformations on various substrates including xenobiotics, steroids, and fatty acids (Esteves, Rueff, and Kranendonk 2021). Mammals express 18 CYP families which encode the 57 human CYP genes (Nebert, Wikvall, and Miller 2013). CYP family 1 (CYP1), CYP2, and CYP3 are unique in that they contain many more genes than the remaining 15. The ancestral expansion of CYP families 1, 2, and 3 coincides with increased organismal exposure to xenobiotics from diet, drugs, chemical inducers, and pheromones, suggesting that evolution favored organisms with CYPs capable of transforming diverse compounds (Nebert, Wikvall, and Miller 2013).

In humans, just 12 CYPs are involved in the biotransformation of the majority of environmental chemicals including 66% of carcinogens and 75% of drugs that undergo metabolism (Rendic and Guengerich 2012; Zanger and Schwab 2013). As heme monooxygenases, CYPs catalyze a reaction that adds an oxygen molecule to make a lipophilic compound more hydrophilic. This reaction requires electrons from cofactors cytochrome P450 reductase (CPR), and cytochrome *b5* (Ionescu and Caira 2005; Bernhardt 2006). The resulting hydrophilic metabolite is then further processed or eliminated through the kidneys.

Despite their extremely low sequence conservation, the general topology and structural folds of CYPs are highly conserved (Hasemann et al. 1995; Mestres 2005; Sirim et al. 2010; Dorner et

al. 2015). To facilitate comparison across CYPs, their amino acid positions have been standardized by combining structural and sequence conservation, and can be queried through the Cytochrome P450 Engineering Database (CYPED) (Gricman, Vogel, and Pleiss 2014; Fischer et al. 2007).

Microsomal CYPs are embedded in the membrane of the endoplasmic reticulum via their N-terminal transmembrane domain. The globular domain is mostly composed of α-helices, assigned letters A-L. The proximal side of the globular domain faces away from the membrane toward the aqueous phase, and interacts with cofactors CPR and *b5* to transfer electrons for catalysis (Baylon et al. 2013). The distal side faces the membrane, and has a highly mobile portion that immerses itself into the lipid bilayer, providing substrate access into the inner region of the enzyme. Catalysis occurs within the hydrophobic core, which is formed by an assembly of structures.

The walls of the hydrophobic core are made up of a four-helix bundle composed of D, E, I, and L running through the protein (**Fig. 1.1**). Helices J, J', K, a coil called the 'meander' (**Fig. 1.1**), and two sets of β sheets are smaller structures interspersed with loop regions that complete the core. These regions contain several characteristic P450 consensus sequences. First, the K helix harbors the critical EXXR motif which stabilizes the core structure and is involved in heme-binding (Hasemann et al. 1995; Sirim et al. 2010; Werck-Reichhart and Feyereisen 2000). Second, the heme-binding loop contains the FXXGXRXCXG (abbreviated CXG) motif, which is located on the proximal side of the heme group and harbors the absolutely conserved cysteine that serves as the fifth ligand to the heme iron (Sirim et al. 2010; Hasemann et al. 1995; Werck-Reichhart and Feyereisen 2000). Finally, the central part of the I helix contains another P450 signature A/G-G-X-N/Q-T-T/S which forms the proton transfer groove on the distal side of the heme (Sirim et al. 2010; Hasemann et al. 1995; Werck-Reichhart and Feyereisen 2000).

The precise mechanisms of a ligand's passage through a CYP active site is not yet elucidated, and has only been indirectly measured experimentally (Oguri, Yamada, and Yoshimura 1994). However, theoretical calculations and molecular dynamics (MD) simulations have produced hypothetical access channels, mediated by several highly dynamic structures (Šrejber et al. 2018; Scott et al. 2016). The F-G domain, composed of the F, F', G', and G helices and the F/G loop, is perpendicular to the I helix and forms a lid to the active site (Otyepka et al. 2007). The F-G domain, together with the B/C-loop, controls substrate access to the active site through membrane contacts, with the F/G loop partially immersed in the lipid bilayer (Otyepka et al. 2007; Baylon et al. 2013). Amphiphilic substrates within the lipid bilayer interior enter the access channel where the F/G loop serves as a gatekeeper (**Fig. 1.2**). Substrates with chemical compositions that allow movement through the access channel with the reactive position oriented towards the heme are transformed, and the product egresses via the solvent channel leading to the membrane/water interface (Šrejber et al. 2018; Paloncýová et al. 2016; Scott et al. 2016). The flexibility of the catalytic cavity correlates with the size of the substrate pools of CYPs 3A4, 2C9, and 2A6, indicating that a more flexible core corresponds to a larger substrate pool (Skopalík, Anzenbacher, and Otyepka 2008). Thus, structural conformation is a prominent mechanism through which CYPs can accommodate a wide variety of substrates while retaining notable substrate specificity, regiospecificity, and stereospecificity.

1.5 The most similar human CYPs, 2C19 and 2C9, have phenotypic differences

The versatility of CYPs is evident (Bernhardt 2006; Zanger and Schwab 2013; Rendic and Guengerich 2012; Munro et al. 2013). Yet, the molecular mechanisms underlying such diverse enzymatic activity remain elusive. Even small differences in wild-type (WT) sequence result in substantial functional differences. For example, CYP2C19 and 2C9 are the closest related

human CYPs, sharing 92% sequence homology and near identical structures. Yet 2C19 and 2C9 have largely disparate substrate pools, distinct membrane interactions, heme angles, and melting temperatures that differ by ~11 ℃ (Niwa and Yamazaki 2012; Wishart et al. 2018; Mustafa et al. 2019; Reynald et al. 2012; Thomson 2021).

The expert-curated database DrugBank illustrates the scale of 2C19 and 2C9's distinct substrate specificity. DrugBank contains 3,859 drugs with known metabolizing enzymes (Wishart et al. 2018) (Accessed March 13th, 2023). Of those, 253 drugs are metabolized by CYP2C9 and 205 by CYP2C19. The homologs have 123 overlapping substrates, but 2C19 has 83 distinct substrates and 2C9 has 130 (**Fig. 2.4I**). Substrate recognition sites (SRSs) were originally inferred based on sequence conservation and location relative to the catalytic core (Gotoh 1992). However, despite being studied for decades, only a small number of sites have been mapped to enzyme-specific substrate metabolism.

Mechanisms of substrate specificity have predominantly been characterized using chimeric CYPs. These studies typically use restriction enzyme cloning to exchange large portions of the CYPs or site-directed mutagenesis to exchange single amino acids. Then, the metabolic activity of the chimera is measured across known enzyme-specific drugs. Between 2C19 and 2C9, chimera studies have identified positions 72, 99, 220, 221, 241, 286, 288, 289, 292, 295, and 331 as sites that distinguish substrate specificity and metabolic activity (Ibeanu et al. 1996; Tsao et al. 2001; Klose et al. 1998; Jung et al. 1998; Wada et al. 2008; Niwa et al. 2002; Attia et al. 2014; Huang et al. 2007; Krenc and Na-Bangchang 2022).

Some sites that contribute to substrate specificity are located in or near the catalytic core with most residing within the predicted SRSs. For example, positions 286, 289, 292, and 295 in SRS4 on the I-helix, which runs through the catalytic core, play a role in determining

4′-hydroxylation of diclofenac, and ibuprofen specificity in CYP2C9 (Klose et al. 1998; Niwa et al. 2002). Additionally, SRS4 sites 241, 286, 288, 289 affect binding and metabolism of warfarin and sulfaphenazole binding (Jung et al. 1998).

Elucidating how sites that are far from SRSs or the catalytic core determine substrate specificity is more challenging. Positions 99, 220, and 221 are critical determinants of omeprazole 5-hydroxylase activity, fully restoring 2C19-like metabolic activity when I99H, S220T, and P221T are substituted into 2C9 (Ibeanu et al. 1996; Wada et al. 2008). These positions also partially contribute to (S)-mephenytoin 4′-hydroxylation specificity in 2C19, in addition to 286, 292, and 295 in SRS4 (Ibeanu et al. 1996; Wada et al. 2008). Since sites 99, 220, and 221 are far from the active site, they must affect substrate specificity through more complex indirect mechanisms (Tsao et al. 2001; Niwa et al. 2002).

MD simulations support long-range conformational changes affecting substrate specificity (Seifert et al. 2006; Mustafa et al. 2020, 2019; Skopalík, Anzenbacher, and Otyepka 2008; Yang, Wong, and Lightstone 2014; Berka et al. 2011; Nair, McKinnon, and Miners 2016; Cojocaru, Winn, and Wade 2007). For example, amino acids K72, P73, and I99 in CYP2C9, and E72, R73, and H99 are located at the protein-membrane interface, far from the catalytic core. These sites, along with site 241, alter substrate specificity and metabolism of tricyclic antidepressants amitriptyline, imipramine, and dothiepin (Attia et al. 2014; Tsao et al. 2001; Niwa et al. 2002). Moreover, differing membrane interactions at sites 72, 73, and 99 result in altered orientation and structure of the substrate access tunnels in the homologs, providing a possible structure-function relationship that explains how variants far from the active site affect substrate specificity (Mustafa et al. 2019, 2020).

[put Fig 5 from mustafa 2019]

Taken together, several overlapping positions affect substrate specificity in 2C19 and 2C9, likely through a combination of direct or local substrate interactions and indirect conformational changes. More information is needed to understand how small changes in sequence result in large functional differences across CYPs like those found between 2C19 and 2C9.

1.6 Mutational scanning of CYPs is needed to understand functional differences between homologs

Deep mutational scanning (DMS) was developed to measure variant effects of all possible single mutants in a protein simultaneously (Fowler and Fields 2014; Araya and Fowler 2011). First, a barcoded library containing all possible single amino acid variants is created. Next, the library is transfected into cells such that each cell contains only one variant, and a selection is applied to discern functional variants from non-functional. In growth-based assays, a functional score is calculated by counting the number of variant copies in initial conditions compared with the variant count after selection (**Fig. 1.4**). Over the years, DMS has proven to be a powerful method to link protein sequence to function, making it an ideal tool to understand how natural sequence variation affects CYPs (Wei and Li 2023). Since its creation, DMS has been applied to several pharmacogenes to determine variant effects and discover protein regions that are sensitive to loss-of-function mutations (Matreyek et al. 2018; Amorosi et al. 2021; Chiasson et al. 2019; Geck et al. 2022; Suiter et al. 2020; Cagiada et al. 2021; L. Zhang et al. 2020; Jones et al. 2020; L. Zhang et al. 2021; Kircher et al. 2019). CYP2C19 and 2C9 were scanned previously, but only ~1% of all 9,310 possible single amino acid mutations were measured (L. Zhang et al. 2020). Thus, the recent scan of CYP2C9 currently stands as the only CYP scan that measures variants at scale (Amorosi et al. 2021).

The CYP2C9 scan reaffirms many critical elements of CYP structure (Amorosi et al. 2021). In this study, variant abundance and variant activity was measured using two distinct assays. First, variant abundance was measured with Variant Abundance by Massively Parallel Sequencing (VAMP-seq) (Matreyek et al. 2018). In VAMP-seq, barcoded CYP2C9 variants are fused to a green fluorescent protein (GFP) reporter, and steady-state variant abundance is measured by sorting cells into quartile bins by GFP fluorescence. Each bin is then deeply sequenced, the number of variant-associated barcodes are counted across bins, and an abundance score is calculated (Amorosi et al. 2021; Matreyek et al. 2018). Variant activity was measured using click-seq (Amorosi et al. 2021). Click-seq measures CYP2C9 activity using a small molecule probe that binds the active site and uses click chemistry to attach a fluorophore reporter. Variants are then quantified using the same schema as VAMP-seq (Amorosi et al. 2021). Using VAMP-seq and Click-seq, Amorosi et al. measured the abundance of 6,370 and the activity of 6,142 missense variants. By intersecting the results from the two assays, the authors identified variants that reduced activity, but not abundance, indicating that those positions are important for CYP function but not stability. They also found that low solvent accessibility was associated with loss of abundance, indicating that variants at buried sites are more likely to be deleterious. Using hierarchical clustering, Amorosi et al. identified that loss-of-function variants clustered in core-facing amino acids on helices D, E, I, J, K, and L, which form the highly conserved heme-binding core (Werck-Reichhart and Feyereisen 2000; Sirim et al. 2010; Amorosi et al. 2021). This DMS of CYP2C9 is one of the first to describe thousands of variant effects across a CYP. However, it is unknown to what degree variant effects can be generalized across the other 12 human CYPs important for drug metabolism.

Deep mutational scans in other homologs reveal that even small deviations in WT sequence can result in unpredictable differences in variant effects. For example, Faber et al. completed three growth-based mutational scan assays: one on WT amidase AmiE, and two on variants of

AmiE that each differ from WT by a single amino acid (Faber et al. 2019). The two variants were chosen such that they had the same enzymatic activity but different probabilities of folding (Faber et al. 2019). The authors then asked whether beneficial variants were shared across all homologs. Out of 190 beneficial mutations, 105 (55.3%) were beneficial in all three AmiE proteins. However, 49 out of 190 (25.8%) of the variants were beneficial in one or both of the variant AmiEs but not in the WT, meaning the single amino acid substitutions in the variant AmiEs enabled epistatic interactions that were unavailable in the WT AmiE (Faber et al. 2019). Thus, three amidase AmiE enzymes that differed by only a single amino acid from WT had dozens of distinct variant effects (Faber et al. 2019).

CYP2C19 and 2C9 share ~92% sequence homology, differing by only 43 amino acids. However, even single amino acids can alter variant effects (Faber et al. 2019). Therefore, I measured the abundance of nearly all single amino acid variants in CYP2C19 using VAMP-seq, and analyzed them with the scores from the 2C9 DMS using a new method called *multidms* (Amorosi et al. 2021; Haddox et al. 2023). This joint analysis marks the first comprehensive comparison of variant abundance between CYP homologs, and may pave a path towards mapping variant effects throughout the CYP family. Moreover, this work provides a wealth of molecular information that may help inform pharmacogenomic applications.

Chapter 2: Understanding the CYP family tree through deep mutational scanning: A joint analysis of CYP2C19 and 2C9 variant abundance

Abstract

Cytochrome P450s (CYPs) are a family of enzymes responsible for metabolizing nearly 80% of small molecule drugs. Variants in CYPs can substantially alter drug metabolism, which may result in improper dosing and severe adverse drug reactions. CYPs have low sequence conservation, making it difficult to anticipate whether variant effects measured in one CYP may extend to others based on sequence alone. Even closely related CYPs, like CYP2C9 and its closest homolog 2C19, have distinct phenotypic properties despite sharing 92% of their sequences. Thus, we used Variant Abundance by Massively Parallel sequencing to measure the protein abundance of 7,660 missense variants in CYP2C19 expressed in cultured human cells. Our results confirmed positions and structural features critical for CYP function, and revealed how variants at positions conserved across all eukaryotic CYPs influence abundance. We jointly analyzed 5,979 variants whose abundance was measured in both CYP2C19 and 2C9, finding that the homologs have different variant abundances in substrate recognition sites within the hydrophobic core, and that substitutions in some regions reduced abundance in CYP2C19 but not 2C9. We also measured the abundance of all single and some multiple WT amino acid exchanges between CYP2C19 and 2C9. While most exchanges had no effect, substitutions in substrate recognition site 4 (SRS4) reduced abundance in CYP2C19. When nearby amino acids were exchanged in double and triple mutants, we found distinct interactions between the sites in 2C19 and 2C9, revealing a region that is at least partially responsible for the difference in thermodynamic stability between the two homologs. Since these positions are also important for determining substrate specificity, there may be an evolutionary tradeoff between stability and altered enzymatic function. Finally, we used our data to analyze 368 previously unannotated human variants, finding that 43% had decreased abundance. Thus, by comparing variant effects

between two closely related, important human genes, we have uncovered regions underlying their functional differences and paved the way for a more complete understanding of one of the most versatile families of enzymes.

Introduction

Nearly 20,000 cytochrome P450s (CYPs) 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 (Munro et al. 2013; Coon 2005). The 57 human CYP genes are grouped into 18 families with 43 subfamilies (M. 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 *2C9* (MIM: 601130) are the most closely related subfamily members, sharing 92% sequence homology. Their protein structures have nearly identical organization, with the largest deviations between their Cα backbones in the substrate binding cavity being only ~3 Å (Reynald et al. 2012). Yet, the two 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℃ higher than 2C9's (Thomson 2021). Thus, even between these close homolog CYPs, the 43 sequence changes drive large functional differences.

Understanding the functional impact of variants across CYPs is particularly important because ~12 of the 57 human CYPs are responsible for contributing to the metabolism of 70-80% of currently prescribed drugs with CYP2C19 and 2C9, together, accounting for 15-20% (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, Pomeranz, and Corey 1998; de Vries et al. 2008), and cost an

estimated $30.1 billion annually (Sultana, Cutroneo, and Trifirò 2013). In order to provide clinicians guidance for treating patients 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 (Relling and Klein 2011; Sim and Ingelman-Sundberg 2010). 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, Lee, and Stouffer 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.

Previously, we used the VAMP-seq assay (Matreyek et al. 2018) to measure the abundance of 6,370 of 9,780 possible missense 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 missense variants of CYP2C19. We identified 4,698 variants that likely result in reduced protein abundance, with 1,122 of those exhibiting complete abundance loss equivalent to nonsense mutations. We first analyzed positions conserved across all eukaryotic CYPs, revealing that all but six of the 58 of 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 2C9 variant abundance datasets, and found 1,912

variants whose abundance differed between the two enzymes. While nearly all sites had at least one variant that differed, 103 of 489 (21%) sites were significantly different between the homologs, substantially more than others have found in mutational scans of influenza (14 of 497; 2.3%) and HIV (30 of 662; 4.5%) homologs (Doud, Ashenberg, and Bloom 2015; Haddox et al. 2018). CYP2C9 had higher mutational tolerance in its hydrophobic core than 2C19, and variants in the structurally conserved K' helix are highly deleterious in CYP2C19, but tolerated in 2C9 even though all K' positions contain the same amino acid in both homologs. We analyzed WT amino acid exchanges between CYP2C19 and 2C9, revealing that sequence differences in a set of diverged positions partially drive the differences in thermodynamic stability between the homologs. These diverged 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, Weinreich, and Hartl 2005). Finally, we analyzed the effects of human CYP2C19 variants. Here, 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 missense variants in the gnomAD database. 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 2C9, some of which have identical sequences. We also provide functional annotations for human CYP2C19 variants which could be used to improve genotype-guided dosing of drugs metabolized by CYP2C19.

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. 2.1A**). VAMP-seq relies on two fluorescent reporters: GFP fused to each CYP2C19 variant to read out abundance, and mCherry expressed via an internal ribosome entry site (IRES) as a transcriptional control. Because CYP2C19 is N-terminally inserted into the endoplasmic reticulum 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 wild type (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. 2.1B**).

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, Stephany, and Fowler 2017; Matreyek et al. 2020). Cells were sorted into quartile bins based on the ratio of GFP:mCherry fluorescence. Each bin was deeply sequenced, variant-associated barcodes counted, and abundance scores calculated based on weighted average of barcode frequencies across bins (**Fig. 2.1A**). Abundance scores were highly correlated between seven replicate sorting experiments arising from three independent library recombinations (**Fig. 2.2A-D;** Pearson's R = 0.82 - 0.98). Replicate scores were averaged, filtered (**Fig 2.3A-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).

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. Abundance scores of synonymous and nonsense variants were well separated, with the missense variant distribution spread between nonsense and synonymous variants (**Fig. 2.1C**). Individually measured GFP:mCherry ratios for 10 variants spanning the range of abundance scores were highly correlated with VAMP-seq scores (**Fig. 2.1D;** Pearson's R= 0.92). Our

results are also highly consistent with a smaller scale VAMP-seq experiment encompassing 121 variants (L. Zhang et al. 2020) (**Fig. 2.4A**, Pearson's R = 0.74). 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 (**Fig. 2.4B, Fig. 2.1C,E**). The majority (58%, 4,620 variants) of missense variants decreased abundance (**Fig. 2.1E**).

Mutational tolerance at conserved CYP2C19 positions reflects function

We visualized the abundance scores as a variant effect map (**Fig. 2.5A**) and projected position-averaged scores onto the CYP2C19 structure (**Fig. 2.5B**). Many of the low abundance variants occur within α-helices and β-sheets (**Fig. 2.5A-B**), especially in amino acids on interior α-helix turns and in regions closer to the protein core (**Fig. 2.5B**).

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 non-human CYPs (Gricman, Vogel, and Pleiss 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, Vogel, and Pleiss 2014). Hierarchical clustering of these eukaryotically conserved positions revealed five clusters with distinct patterns of variant abundance scores (**Fig. 2.5C, Fig. 2.6A**). 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, Vogel, and Pleiss 2014).

In clusters 1, 2, and 3 nearly all substitutions, except those of the same biophysical type, reduced abundance (**Fig. 2.5C, Fig. 2.6A**). 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 is the WT amino acid, which 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 likely 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 (**Fig 2.6B**). Positions 297 and 362 influence substrate specificity, and >80% eukaryotic CYPs have hydrophobic amino acids at these positions (Gricman, Vogel, and Pleiss 2014). Surprisingly, while various 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 2C19 (Reynald et al. 2012; Foti et al. 2012; Altarsha et al. 2009; Haines et al. 2001) and contains hydrogen-bonding amino acids in >80% eukaryotic CYPs at this position, and most substitutions did not appreciably reduce abundance. Finally, E300 stabilizes a water network during proton delivery (Haines et al. 2001), and tolerated substitutions other than aspartic acid.

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, McKinnon, and Miners 2016).

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

We analyzed variant effect patterns at positions conserved across eukaryotes, where, except for a subset of catalytically important sites, most substitutions reduced CYP2C19 abundance. Next, we investigated variant effect patterns in CYP2C19 compared to its closest homolog, CYP2C9. CYP2C19 and CYP2C9 share 92% protein sequence homology and nearly identical crystal structures (**Fig 2.8A,** RMSD = 0.596 Å). However, they have important functional differences, notably their substrate profiles and membrane interactions (Niwa et al. 2002; Niwa and Yamazaki 2012; Goldstein and de Morais 1994; 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(Amorosi et al. 2021). The combined dataset contained 5,979 variants whose abundance was scored in both CYPs. Most variants had similar effects in both homologs, though there is evidence of both noise and bias in each dataset (**Fig. 2.9A**, Pearson's r = 0.77). A simple approach would be to compute the difference in abundance scores between homologs. However, it would be difficult to tell whether shifts were due to signal or experimental noise. Instead, we estimated shifts using a joint-modeling approach called *multidms* that uses L1 regularization to drive the

estimated shifts to zero unless they are strongly supported by the DMS data for each homolog (Haddox et al. 2023). How strongly a shift is supported by the data depends both on its magnitude and the number of times a given variant was observed in the DMS data. We inferred shifts as the difference in abundance score in CYP2C19 relative to 2C9, such that positive shifts indicate a higher abundance score in CYP2C19 and negative shifts indicate a lower abundance score. We applied regularization weights ranging from 0.0 to 1e-4, and selected 1e-5 as the optimal value for subsequent analysis (**Fig. 2.7**). 1,912 variants (32.0%) had non-zero-shift values meaning that they had different effects between the two homologs (**Fig. 2.8B, Fig. 2.9B**).

We calculated the mean of the shift values at each position to reveal the effect of regional and structural features (**Fig. 2.9C**). We identified positions with mean shift values that differed significantly from 0 using a randomization test (**Fig. 2.8C**). 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. 2.9C**). In this region, mean shift values were negative meaning that substitutions were more deleterious in CYP2C19 than in 2C9 (**Fig. 2.8D, Fig. 2.9D**).

Overall, CYP2C19 was more mutationally tolerant than 2C9 in the D, E, I, L, J, and J' helices (**Fig. 2.8B, Ei-ii, Fig. 2.9C**), which form the majority of the hydrophobic core(Denisov et al. 2005; Werck-Reichhart and Feyereisen 2000). The sites that were most differentially tolerant in these helices were on portions of the helices that sit outside of the hydrophobic core (**Fig. 2.8Ei-ii**). Conversely, CYP2C9 was more mutationally tolerant than 2C19 at positions within the hydrophobic core near important sites for heme positioning and function (**Fig. 2.8D**). Many of these heme-associated positions reside within substrate recognition sites (SRSs) (Gotoh 1992), and CYP2C9 was more mutationally tolerant than 2C19 in SRSs relative to the other regions of the protein (**Fig. 2.9C, E-F**). The mutational tolerance of positions in SRSs that were not heme-associated were similar between the homologs (**Fig. 2.9F**).

We also examined whether differences between CYP2C19 and 2C9 could be explained by sensitivity to variants of different biophysical types or by differences in the structures of the two homologs. However, we found that neither homolog is more sensitive to particular types of substitutions **(Fig. 2.8F**), and that shift values were unrelated to the distance between positions in the CYP2C19 and 2C9 crystal structures (**Fig. 2.8G**) 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 2C9, but that CYP2C19 was less sensitive than CYP2C9 to substitutions in other regions flanking the hydrophobic core.

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

We investigated abundance shifts at all variants comparing 2C19 and 2C9. However, the phenotypic differences in substrate recognition, membrane interaction, and thermodynamic stability between the two homologs must be driven by divergent sites. While most divergent sites are not localized to the catalytic site, some are critical for substrate specificity, regiospecificity, and stereospecificity(Jung et al. 1998; Attia et al. 2014; Wada et al. 2008; Klose et al. 1998; Ibeanu et al. 1996; Lewis et al. 1998). In many cases, evolutionary pressures result in a protein's reduction of thermodynamic stability in exchange for new functionality(DePristo, Weinreich, and Hartl 2005). We wondered whether we could link the differences in thermodynamic stability between the homologs to substrate specificity by measuring the abundance of the 43 divergent sites. Thus, we investigated the abundance of the variants that partially convert CYP2C19 to 2C9 and vice versa.

We had abundance scores for 33 of the 43 2C19 variants that install the WT CYP2C9 amino acid, (e.g. CYP2C19→CYP2C9). To enable an exhaustive analysis, we individually

measured GFP:mCherry fluorescence for each of the 10 CYP2C19→CYP2C9 variants not present in our abundance data (**Fig. 2.10A**). 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 diclonefac(Tsao et al. 2001; Niwa et al. 2002) whereas R261Q has not been studied in either homolog. V288E caused profound loss of abundance and was classified as “nonsense-like” (**Fig. 2.10A**). 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 due to the introduction of an electrostatic clash between E241 and V288E (**Fig. 2.10B**). In support of our hypothesis, V288, K241, and I289 have indeed been suggested to play a role in abundance and substrate specificity(Jung et al. 1998; Klose et al. 1998; Attia et al. 2014; Tsao et al. 2001; Niwa et al. 2002). Thus, we sought to understand how positions 241, 288, and 289 might interact to influence abundance in CYP2C19 and CYP2C9.

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. 2.10B, 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. 2.10C**). 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 likely does not contribute appreciably to thermodynamic stability (Jung et al. 1998). Combining K241E and E288V fully restored CYP2C9 abundance (**Fig. 2.10C**). Thus, both homologs have a similar pattern, with installation of a second negative charge disrupting abundance. Elimination of one

of the two negative charges even with variants from the other homolog restored abundance, although to differing degrees in each homolog.

Next, to incorporate 289 into our analysis, we measured the abundance of the CYP2C19→CYP2C9 and CYP2C9→CYP2C19 241, 288, 289 triple mutants (**Fig. 2.10C**). 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 two double mutants. Thus, while the interaction between these three positions is complex, it seems likely that sequence changes in this region of the protein contribute to the increased thermodynamic stability of CYP2C19.

Annotating human *CYP2C19* variants.

CYP2C19 variants can increase, decrease, or eliminate a patient's ability to metabolize many important drugs, and knowing variant function can help avoid severe and expensive adverse events (Lazarou, Pomeranz, and Corey 1998; de Vries et al. 2008; Goulding et al. 2015; Schmiedl et al. 2018; Sultana, Cutroneo, and Trifirò 2013). For example, the anti-platelet drug clopidogrel is activated by CYP2C19. Thus, patients with deleterious CYP2C19 variants experience reduced or non-existent 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 (Pereira et al. 2021; Galli 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 unknown function. Thus, we analyzed the functional effects of *CYP2C19* alleles in PharmVar. All four PharmVar "normal

function" alleles had WT-like abundance scores (**Fig. 2.11A**). Of the eight "decreased" and "no function" alleles, six were low abundance. The remaining two decreased/no function alleles, *6 and *9, were WT-like abundance. PharmVar lists the *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 *6 allele's loss of function results from disrupted enzymatic activity rather than loss of abundance. Consistent with this interpretation, the *6 allele is intact enough to bind its heme cofactor, but has a decreased ability to metabolize substrates and disrupted electron flow from CYP reductase (Ibeanu et al. 1998; Derayea et al. 2020). *9 (R144H) had an abundance score of 0.914 (95% CI 0.956-0.872). Consistent with these results, *9 has WT-like affinity for cytochrome P450 reductase(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(L. Zhang et al. 2020) with less than ~50% of WT abundance. We therefore individually validated this variant and reaffirmed its WT-like abundance in our hands (**2.12**). In light of *9's moderately reduced activity against mephenytoin, ability to bind cytochrome P450 reductase normally and WT-like abundance in our assay, we suggest that, like *6, *9 has normal abundance but decreased catalytic activity. Overall, the six of eight 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 four PharmVar alleles with uncertain function, we found that *30 and *23 had decreased abundance strongly suggesting that they would disrupt drug metabolism (**Fig. 2.11A**).

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 PharmVar database represents only a tiny fraction of the

CYP2C19 variants discovered so far. For example, there are 408 unique *CYP2C19* missense variants in the genome and exome database gnomAD, 390 of which have no CPIC annotation or functional information. We were able to annotate 368 (90.2%) of the variants in gnomAD (**Fig. 2.11B**), 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 are of decreased or no function. We annotated 210 (57.0%) variants as “WT-like” or “possibly WT-like,” providing some evidence that these variants 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, *3, and *17 alleles. The study discovered that *2 was present in 15.2%, *3 in 0.3%, and *17 in 20.4% of individuals, and nearly 60% had at least one of these star alleles (Ionova et al. 2020). Thus, *CYP2C19* variants with reduced abundance appear common in the population.

Overall, our abundance data agree with variant functional annotations for *CYP2C19* variants in PharmVar, and we provide evidence for most variants lacking functional information (Lee et al. 2022). We also provide evidence for the functional consequences of 90.2% of the variants found in the gnomAD database, highlighting that ~40% of variants are likely to be of no or reduced function. Finally, we measured the abundance of 7,660 missense variants representing 82% of possible *CYP2C19* single missense variants. As genome sequencing and genetic testing continue, many of these variants will be discovered. Thus, we substantially expand the currently available functional evidence for *CYP2C19* variants.

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 2C19 and 2C9, are functionally distinct, and the underlying causes of these distinctions are largely unknown. Thus, we used VAMP-seq to measure the abundance of 7,660 CYP2C19 missense variants. In addition to confirming positions known to be critical for CYPs structure and function, we revealed that variants at four conserved positions in the hydrophobic core do not impact CYP2C19 abundance. By jointly analyzing 5,979 shared CYP2C19 and 2C9 abundance scores, we discovered regions where the two homologs have different mutational tolerances. CYP2C9 has a more tolerant hydrophobic core, whereas 2C19 is more tolerant in regions surrounding the core. We measured the abundance of WT amino acid swaps between CYP2C19 and 2C9, 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 two star alleles of unknown function, *30 and *23, may 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, suggesting that they would impact drug metabolism.

Of 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 six 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 they likely bind cofactor cytochrome P450 reductase (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 (Peng et al. 2008; Reynald et al. 2012; Polgár, Menyhárd, and Keseru 2007). That these positions are important for substrate specificity but not abundance highlights their specialized role in CYP2C19.

We also jointly analyzed CYP2C19 and 2C9 (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 2C19 but were tolerated in 2C9, suggesting a markedly different functional role of K' in the enzymes despite having identical WT amino acids. Moreover, CYP2C9 had a more tolerant hydrophobic core than 2C19, especially in SRSs that contain heme-associated positions. CYP2C9's higher mutational tolerance in its core may indicate more flexibility. We speculate that, since the flexibility of CYP active sites is correlated with its promiscuity(Nair, McKinnon, and Miners 2016; Skopalík, Anzenbacher, and Otyepka 2008), this may be one component that allows 2C9 to bind more substrates (Wishart et al. 2018).

Variants that impart novel function, like those that alter substrate specificity, often reduce thermodynamic stability (DePristo, Weinreich, and Hartl 2005). To determine if CYP2C9's altered substrate profile and lower thermodynamic stability relative to CYP2C19 (Thomson 2021), constituted such a tradeoff, we analyzed the 43 divergent positions between CYP2C19 and 2C9. We found that positions 241, 288, and 289 are the likely locus of such a tradeoff because these three 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 (**Fig. 2.10B**). Position 288 was the only position in CYP2C19 where installing the 2C9 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 at least 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. Likewise, CYP2C9 K241E introduces the same opposing negative charge adjacent to E288. This pattern of loss of abundance suggests that CYP2C9 may have evolved from CYP2C19. This is because CYP2C9 is the only CYP2C 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 three 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 CYP2C9's reduced thermodynamic stability. CYP2C9's new substrate binding capabilities apparently made this loss of thermodynamic stability evolutionarily tolerable.

Here, we present abundance measurements for thousands of CYP2C19 variants, useful for understanding CYP structure, function and evolution as well as the impact of human variants. However, there are some key limitations to our abundance data. First, we expressed a CYP2C19 cDNA from an inducible promoter, so we cannot detect variants that induce splicing defects or affect transcriptional regulation. Second, variants can affect function without affecting abundance. For example, variants may disrupt a critical substrate binding position or prohibit binding to critical cofactors like cytochrome P450 reductase (CPR) or cytochrome b5. 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 tolerance across multiple CYPs,

ultimately creating a model that can predict variant effects in any human CYP regardless of whether they have been directly assayed. In addition to improving personalized drug dosing, such comprehensive profiling could expand our overall understanding of CYP function.

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 Table S3

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, 100U/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 by detachment with trypsin 0.5%. All cell lines were tested negative for mycoplasma on a monthly basis.

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 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 Table S2 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, 2uM of mixed primers, and 5% DMSO in a 5 µl reaction volume of KAPA HiFi Hotstart 2x 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 µl of phosphorylated product was combined with 1 µl of 10x T4 ligase buffer and 0.5 µl of T4 DNA ligase (NEB), incubated at 16℃ overnight, and cleaned and concentrated. The ligated products were transformed into electrocompetent E. 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 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 AflII at 37℃ for 1h, followed by heat inactivation at 65℃ for 20 minutes. We ordered barcode oligos with 18bp random sequences from IDT, resuspended them at 100 µM, and annealed them by combining 1 µl of each primer with 4 µl of CutSmart buffer and 34ul of ddH2O and running 98℃ for 3 minutes, ramping down to 25℃ at -0.1 ℃/s. The annealed oligos were then Klenow filled by combining 0.8 µl Klenow polymerase (exonuclease negative, NEB) with 1.35 µl of 1mM dNTPs with 40 µl of product to fill in the barcode oligo, using cycling conditions of 25℃ for 15 min, 70℃ for 20 min, ramping down to 37℃ at -0.1℃/s. The resulting products were then

ligated overnight at 16℃. The barcoded library was transformed into electrocompetent E. coli cells (NEB C2989K) and midiprepped (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 libraries with Illumina sequencing. The forward and reverse reads were merged using Pear (J. Zhang et al. 2014), and barcode counts were estimated using Bartender (L. Zhao et al. 2018). Barcodes with fewer than 10 reads were filtered out, resulting in ~200,000 unique barcodes for an average of 21x coverage.

PacBio sequencing for barcode-variant mapping

PacBio sequencing libraries were generated with SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences) according to 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% 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 cell in a Sequel II v2.0 run using a 15 hour movie.

We filtered long reads for a minimum of 3 passes. We then analyzed the circular consensus reads (CCSs) 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 Supplemental Table S3 for details).

FACS-based deep mutational scan (VAMP-seq)

HEK293T with a Bxb1 serine recombinase landing pad with an inducible Caspase 9 cassette (HEK293T-LLP-iCasp9)(Matreyek et al. 2020) that enable expression of one variant per cell were used for all human cell experiments. 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 one tube, 7.1 µg of barcoded library plasmid was mixed with 0.48 µg of Bxb1 plasmid in 710 µL of OptiMEM. In a separate tube, 28.5 µl of Fugene was diluted into 685 µl of OptiMEM. The Fugene and DNA tubes were then combined and incubated at room temperature for 15 minutes. 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 µg/mL. 24 h after doxycycline was added, 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 four 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. Three biological replicates from separate transfections were collected for the FACS-based deep mutational scan. Sorted library amplification and sequencing

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 Kit (QIAGEN) according to the manufacturer's instructions, with the addition of a 30-minute incubation step at 37°C with RNase during the resuspension step. For the first round of PCR, eight 50 µL reactions were set up for each sample, with a final concentration of 50 ng/µL input genomic DNA, 1x Q5 High-Fidelity Master Mix, and 0.25 µM of JS454 and JS1004 primers. The reaction conditions

were 95°C for 30 seconds, 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 3 minutes, repeated 4 additional times, followed by 72°C for 2 minutes and a 4°C hold. The eight reactions were then combined, bound to AMPure XP (Beckman Coulter) at 0.6X bead volume to sample volume, cleaned, and eluted with 38.5 µL water. 15 µL (40%) of the eluted volume was mixed with Q5 High-Fidelity Master Mix, GB001, and one of the indexed reverse primers, JS385 through JS473, added at 0.25 µM each. The PCR reaction was run with SYBR Green I on a Bio-Rad MiniOpticon. The reaction was denatured for 3 minutes at 95°C, cycled 18 times at 95°C for 15 seconds, 67°C for 30 seconds, and 72°C for 45 seconds, with a final 2-minute extension at 72°C.

The indexed amplicons were then run on a Tapestation according to the manufacturer's instructions. For each sample, 1 µL sample was mixed with 3 µL 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% agarose gel with SYBR Safe, and then the gel was extracted using a freeze and 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 greater than 4 x 10^-5 across all bins (**Fig. 2.3**). 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. Missense variant abundance scores ranged from -0.09 to 1.5.

Abundance classes were determined as in previous studies(Amorosi et al. 2021; Matreyek et al. 2018). To discern 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 missense variants. The missense 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 cloned 11 variants using IVA cloning(García-Nafría, Watson, and Greger 2016) site directed mutagenesis into the VAMP-seq recombination vector (attB-CYP2C19-eGFP-IRES-mCherry) via primers listed in **Table S3** (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 transformed chemically competent *E. coli* cells (NEB C2987 or Bioline BIO-85027). Bacterial

clones were prepped with a midiprep kit, validated by Sanger sequencing and whole plasmid nanopore sequencing. We then transfected the preps into and HEK293T-LLP-iCasp9 landing pad cells in 6-well plates with 400,000 cells per well. 2.7 µg of plasmid was 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 minutes. After incubation, the Lipofectamine/DNA mixture was added to cells dropwise and the plates were placed in an incubator at 37°C. After 24 hours, the cells were induced with doxycycline at a final concentration of 2.5 µg/mL, and at least 24 hours later we selected for recombinant cells by adding small molecule AP1903 which causes inducible Caspase 9 in unrecombined landing pad cells to dimerize, activate, 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). Three biological replicates of each individual variant were measured.

multidms **analysis of CYP2C19 and 2C9 deep mutational scans**

The *multidms* approach is only compatible with DMS data where most mutations are observed in multiple unique variants, each with a functional score, as the number of variants with a given mutation is the basis by which shifts are regularized. In the CYP2C19 and CYP2C9(Amorosi et al. 2021) DMS data, each mutation was only present in one variant, but nearly all variants were associated with several unique barcodes. Thus, to use *multidms*, we calculated abundance scores at the level of individual barcodes and used these data as input to *multidms*, where shifts are regularized based on the number of barcodes associated with a given mutation. We then fit

a single *multidms* model to DMS data from both homologs. The model was trained to predict abundance scores in CYP2C9 and shifts in scores in CYP2C19 relative to CYP2C9, minimizing the difference in predicted and experimentally measured scores, as quantified by a Huber loss function. We used L1 regularization to drive shifts to zero unless they were strongly supported by the data. We used a penalty weight of $\lambda = 1e\text{-}5$. At this weight, we detected 1,912 mutations with non-zero shift values.

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 15 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. P-values were then adjusted for repeated hypothesis testing using the Benjamini-Hochberg method(Benjamini and Hochberg 1995) with a false discovery rate of 5%. Positions with p-values less than 0.05 were considered significant.

Part II - Heterogeneous drug response between cells

Chapter 3: Morphological heterogeneity in drug treatment

3.1 Cellular heterogeneity complicates the study of biology

Part I describes challenges introduced by heterogeneity at the organismal level. These same challenges are present at the cellular level, even amongst clonal cell lines. Cellular heterogeneity has classically been acknowledged as a fundamental property of biological systems, meaning that the assumption that bulk biological measurements come from a single, normally distributed population is often incorrect (Elsasser 1984; Rubin 1990). Failing to account for heterogeneity can result in incomplete or incorrect scientific findings (Altschuler and Wu 2010). The consequences of the flawed scientific findings resulting from heterogeneous cell populations depend on the composition of the heterogeneity.

Bulk measurements from a population that contains multiple dominant subpopulations will result in inaccurate data for most or all members of the population, meaning that any resulting findings will be flawed (Altschuler and Wu 2010). When a population contains small subpopulations, a bulk measurement will yield data that still accurately represents the majority of the cells in the population (Altschuler and Wu 2010). However, findings based on these data will not be predictive for members of the small subpopulations. In some cases, lacking predictive power for small subpopulations can have dire consequences. For example, tumor composition is often heterogeneous, containing subsets of cells that respond differently to chemotherapeutic interventions. Applying a treatment targeted towards the majority population may simply purify a resistant population, potentially resulting in a new, treatment-resistant tumor (Fisher, Pusztai, and Swanton 2013; Sun and Yu 2015; Raatz et al. 2021).

Biologically meaningful cellular heterogeneity is now appreciated as the default, rather than the exception, especially in cellular responses to environmental stress and disease. As a result, single-cell technologies that can capture quantitative biological measurements for individual cells have been widely applied, and are capable of characterizing various heterogeneous molecular phenotypes (Chattopadhyay and Roederer 2015; Cohen et al. 2008; Almendro, Marusyk, and Polyak 2013). For example, single-cell RNA sequencing quantifies thousands of transcripts per cell, single-cell ATAC-seq measures chromosome accessibility, and Memory-Seq tracks cellular memory across cell divisions (Shaffer et al. 2020; Cao et al. 2017, 2018; Cusanovich et al. 2015; Cuomo et al. 2023). However, these approaches generally do not capture heterogeneous cellular morphology and behavior.

3.2 Cell morphology is an important biological signal

Cell morphology is a potent indicator of cell state. When treated with drugs, even clonal cell populations can elicit biologically meaningful morphological heterogeneity (Slack et al. 2008). Moreover, the distributions of morphological subpopulations across several markers can be used to successfully group drugs by their mechanisms of action (Slack et al. 2008). High content screening (HCS) leverages the link between morphology and drug action. HCS uses automated workflows to profile thousands of chemical compounds at various doses (Caicedo, Singh, and Carpenter 2016; Rohban et al. 2017). In these workflows, cells are seeded into 96- or 384-well plates, treated with hundreds of drugs across different doses, stained with organelle-specific dyes, imaged, and analyzed to identify drug hits that induced changes in visual phenotype. Since many of these visual phenotypes are too subtle for the human eye, machine learning algorithms are employed (**Fig. 3.1**). By identifying drugs that induce visual phenotypes, HCS is poised to make drug discovery faster, cheaper, and more efficient (Lin et al. 2020).

A major limitation of current HCS methods is the need to pre-select a small number of proteins of interest to visualize. While methods like Cell Painting use dyes that span most of the critical organelles, identifying the overall molecular cell state of morphologically distinct subpopulations remains challenging (Rohban et al. 2017; Cimini et al. 2023). Combining Cell Painting with the L1000 assay, which estimates the gene-expression profile by directly quantifying a reduced representation of the transcriptome, unifies single-cell transcriptional state with morphology (Way et al. 2022). However, like other single-cell technologies, the morphological and spatial information is lost when cells are removed from their imaging plate. Thus, morphological information cannot be directly tied to cell state. Moreover, the transcriptome does not directly measure proteins, which are the functional units of the cells, and are not always accurately quantified by mRNA abundance (Vogel and Marcotte 2012).

3.3 Technologies combining mass spectrometry-based proteomics with microscopy

Mass spectrometry-based proteomics (MS) is the gold standard of proteome-wide protein quantification, and is capable of reliably quantifying proteome-wide protein abundance, changes in translation and degradation, and protein interaction networks (Lobingier et al. 2017; Doherty et al. 2009; Nilsson et al. 2010; Zubarev 2013). However, most MS technologies must be applied to populations of cells, meaning that morphological heterogeneity information is lost. Thus, microscopy-based techniques that measure proteins are important for investigating drugs and other stimuli that provoke morphologically heterogeneous phenotypes in a cell population.

MS imaging (MSI) was first described more than 50 years ago, and remains the current leading technique for unbiased quantification of hundreds of spatially resolved proteins (Dilmetz et al.

2021; Buchberger et al. 2018). The most common method is matrix-assisted laser adsorption/desorption ionization (MALDI) (Niehaus et al. 2019). In MALDI workflows, tissue sections are mounted and coated with a matrix to facilitate analyte ionization. Proteins on the surface of the sample are digested into peptides. The peptides are then ionized with laser pulses that are adsorbed by the matrix, resulting in desorption/ionization of the sample and matrix molecules into a gaseous plume. The mass spectra are collected from the plume and analyzed at each spatial coordinate (Dilmetz et al. 2021). Multi-modal MSI systems pair morphology information with the molecular image by asynchronously capturing images using traditional microscopy techniques and computationally aligning them afterwards (Buchberger et al. 2018). While powerful, MSI methods require specialized equipment and expertise, and must balance trade-offs in spatial resolution, proteomic coverage, quantification, and sample processing.

Many of the challenges inherent to MSI are circumvented by separating sample collection from the MS analysis. Improvements in image analysis and low-input MS have been leveraged to combine laser capture microdissection with MS analysis (Piehowski et al. 2020; Dewez et al. 2020; Mund et al. 2022). For example, Deep Visual Proteomics (DVP) uses custom instrumentation to automatically image tissue slices or fixed cells, microdissect regions or morphologies of interest, capture the samples in a plate for sample preparation with a low-input, single-pot method, then analyze the samples on a newly developed MS (Mund et al. 2022). DVP was used to analyze the proteomes of 80-100 cells with distinct nuclear morphologies, revealing single-cell heterogeneity represented by 515 significant protein groups from 3,653 unique proteins. DVP was also applied to salivary gland carcinoma tissue sections to separate normal-appearing acinar cells from acinic cell carcinoma cells. By comparing healthy tissue with cancerous tissue, they discovered possible new therapeutic targets (Mund et al. 2022).

Deep Visual Proteomics improves on many of the limitations of MSI methods, providing high spatial resolution, deeper proteome coverage, and robust protein quantification. However, DVP suffers from similar barriers to entry as MSI methods, limiting its widespread application. In particular, DVP requires laser capture microdissection for spatially resolved sample collection. An alternative sample collection method called Visual Cell Sorting (VCS) provides similar spatial resolution, but with more approachable techniques, less specialized instrumentation, and yields substantially more material from cultured cells.

3.4 Visual Cell Sorting is a versatile method for separating cells with distinct morphologies

Visual Cell Sorting is a microscope-based method that separates cells based on their visual features. In short, cells express a photoconvertible green fluorescent protein called Dendra2. Automated microscope software captures images, segments cells, classifies cells into subpopulations based on morphologies of interest, then selectively illuminates nuclei of each subpopulation with a pulse of UV light, red-shifting the Dendra2 (Hasle et al. 2020). After all fields of view have been imaged and activated, cells are physically recovered using fluorescence-activated cell sorting (FACS) and analyzed using sequencing or other biochemical or molecular assays (**Fig. 3.2**). At 20x magnification, VCS can capture 6,000 fields of view in a ~13 hour experiment,.

The VCS workflow was previously combined with single cell transcriptomics to study heterogeneous cell response to the chemotherapeutic paclitaxel (PTX) (Hasle et al. 2020). In cells treated with a low dose of PTX, a subset had abnormal appearing nuclei with irregular lobulations. The transcriptomes of cells that were able to maintain normal appearing nuclei in

the presence of PTX had higher levels of known PTX resistance genes, indicating that, despite being a clonal population, some cells may have been in transcriptional state similar to “pre-resistance” (Mitra et al. 2016; Symmons and Raj 2016; Shaffer et al. 2017, 2018). Thus, the morphological heterogeneity in PTX-treated cells was biologically meaningful, and was only discoverable by pairing visually distinct cells with their transcriptional states.

In Chapter 4, I will describe my work combining VCS with MS-based proteomics (VCS-MS), which presents several advantages over methods like DVP that use laser capture microdissection. First, DVP is comparatively low throughput when applied to cultured cells. Compared to DVP, which captures hundreds of cells per phenotype, VCS can isolate tens of thousands. Moreover, DVP is incompatible with live cells since samples have to be pre-treated for laser capture microdissection, whereas VCS can be used with incubation microscope stages. I show that I can reliably separate mouse and human cells using VCS-MS, and quantify >5,000 proteins from ~50,000 cells with high accuracy. By introducing VCS-MS as another spatially resolved MS method, I have introduced an accessible and versatile tool to investigate cellular heterogeneity in cultured cells.

Chapter 4: Proteomics on visually distinct cell populations

Abstract

Visual phenotypes are potent indicators of cell state. However, investigating the link between visual phenotype and cell state requires pairing microscopy with -omic technologies. While mass spectrometry is a powerful tool for unbiased measurement of the proteome, technologies that combine mass spectrometry with microscopy are limited. Visual Cell Sorting (VCS) is a flexible workflow that separates cells based on visual features. Here, combine VCS with mass spectrometry-based proteomics (VCS-MS) to investigate morphologically heterogeneous cell populations. In an experimental evaluation, we apply VCS-MS to mixed populations of mouse and human cells, identifiable by fluorescent reporters. We demonstrate that VCS-MS separates the visually distinct cells with similar fidelity to fluorescence-assisted cell sorting. In low-input samples containing 50,000 cells, we reliably detect ~7,700-9,800 unique proteins represented by ~72,000-97,000 unique peptides with minimal cross-species contamination. Thus, VCS-MS is a powerful new method for interrogating the molecular underpinnings of morphological heterogeneity in cultured cells.

Introduction

Single cell technologies can reveal how and why cells differ, and are fundamental for understanding processes like development and adaptation to environmental stress and disease (Altschuler and Wu 2010; Chattopadhyay and Roederer 2015; Almendro, Marusyk, and Polyak 2013; Cohen et al. 2008). In particular, single-cell sequencing approaches are widely applied, but these approaches generally do not capture heterogeneous cellular morphology which can indicate how cells are responding to stimuli (Rohban et al. 2017; Slack et al. 2008). Moreover, sequencing-based approaches do not directly measure proteins, the functional units of the cell. Thus, microscopy-based techniques that measure proteins are important for investigating drugs and other stimuli that provoke morphologically heterogeneous phenotypes in a cell population. However, most microscopy-based techniques that measure proteins are antibody-based and are therefore limited in throughput and in discovery-based application because they require pre-selected targets of interest.

Mass spectrometry (MS) is capable of reliably quantifying proteome-wide protein abundance (Zubarev 2013; Nilsson et al. 2010), changes in translation and degradation (Doherty et al. 2009), and protein interaction networks (Lobingier et al. 2017). However, most MS technologies require populations of cells, meaning that information about heterogeneous phenotypes, including morphology, is lost. Recently, MS has been combined with microscopy to quantify proteins in spatially resolved samples. For example, laser capture microdissection paired with low-input MS (Piehowski et al. 2020; Dewez et al. 2020; Mund et al. 2022) and MS imaging (Buchberger et al. 2018) allows detection of 2,000-4,500 proteins. However, they require extensive expertise, specialized equipment, and cannot be used on live samples.

Visual Cell Sorting (VCS) is a microscope-based method that identifies individual cells with a morphology of interest and selectively illuminates them at 1 μm resolution with a short pulse of blue light to irreversibly photoconvert a fluorophore (Hasle et al. 2020). These photoconverted cells can be separated with fluorescence-activated cell sorting (FACS) and

analyzed using sequencing or other biochemical or molecular assays. Because VCS employs an automated microscope, 6,000 fields of view at 20x magnification can be captured in a ~13 hour experiment. The VCS workflow was previously combined with single cell transcriptomics to study heterogeneous cell response to the chemotherapeutic paclitaxel (PTX) (Hasle et al. 2020). In cells treated with a low dose of PTX, some cells had abnormal appearing nuclei with irregular lobulations. The transcriptomes of cells that maintained normal appearing nuclei had higher levels of known PTX resistance genes, indicating that, despite being a clonal population, some cells were in a transcriptional state similar to “pre-resistance” (Mitra et al. 2016; Symmons and Raj 2016; Shaffer et al. 2017, 2018). Thus, the morphological heterogeneity in PTX-treated cells uncovered a novel PTX response state, which was only discoverable by pairing visual features with transcriptional state.

VCS is an unbiased, high throughput, inexpensive, and flexible workflow capable of isolating visually distinct cells. Moreover, outfitting an automated microscope for VCS requires only three components: a live cell incubation chamber, a digital micromirror device, and a 405 nm laser totaling ~$50,000, making it particularly accessible.

In this work, we develop a workflow to couple VCS with MS (VCS-MS) and are able to measure >7,700 unique proteins from 50,000 cells with minimal contamination from cell culture serum used during both culturing and FACS. In an experimental evaluation of VCS-MS, we separate mouse cells from human cells by their fluorescent reporters, and demonstrate the method is highly reproducible and performs as well as standard protocols using FACS alone. We demonstrate that VCS-MS is a robust protocol that can be applied to cultured cells with heterogeneous visual phenotypes.

Results

VCS-MS workflow

Visual Cell Sorting (VCS) (Hasle et al. 2020) physically separates visually distinct cells using automated microscopy to image cells, classify them based on visual phenotype, and irreversibly photoconvert a fluorescent reporter in cells of interest. After microscopy, cells are separated using fluorescence-activated cell sorting (FACS). To enable the proteomic characterization of visually distinct cells, we first established a workflow to couple VCS with mass spectrometry (VCS-MS; **Fig. 4.1A**).

Since VCS experiments produce samples with 25,000 - 75,000 cells, we prepared cells after sorting using a low-input mass spectrometry sample preparation protocol called Single-Pot Solid-Phase-enhance Sample Preparation (SP3) (Hughes, Sorensen, and Morin 2019) and measured samples using LC/MS-MS with data-dependent acquisition (DDA). Although the workflow is similar to another study that used SP3 to measure 25,000 FACS sorted macrophages (Sielaff et al. 2017), our samples contained bovine serum albumin (BSA) derived peptides that made up >80% of the total signal intensity, accounting for ~13% of all detected peptides (**Fig. 4.3A-B**). Such a substantial amount of BSA suppresses signal from peptides arising from the proteomes of the cells of interest, limiting depth of coverage.

Previous studies have sorted cells into receptacles with filter material that captures cells but allows aqueous solutions to pass through, enabling multiple PBS washes without losing cells to repeated rounds of pelleting and aspiration. After washing, cells can be lysed in the filter chamber and the lysate centrifuged into fresh tubes for downstream sample preparation(Myers et al. 2019). As an approach to remove BSA from our samples, we tested three PBS washes across four types of tubes with different membranes: BioInert™, wwPTFE, hydrophilic PVDF, and regenerated cellulose (**Fig. 4.3C-D**). The NanosepMF tubes with a BioInert™ membrane from Pall Biosciences resulted in the lowest BSA signal and fewest BSA peptides relative to

proteome peptides (**Fig. 4.3C-D**). However, BSA still accounted for >40% of the total signal detected (**Fig. 4.3C-D**). We hypothesized that after the first wash, cells were clumping on the membrane, forming a hydrophobic shell preventing subsequent washes from removing BSA from the interstitial space between cells. To test this hypothesis, we vortexed the cells to resuspend them between washes. However, we detected only modest improvement (**Fig. 4.3E-F**).

As an alternative approach to remove BSA, we turned to the Laminar Wash™ MINI System (Curiox) which can wash cells in suspension without centrifugation or cell loss. This is critical for VCS due to the limited number of cells generated in each experiment. However, BSA peptides still made up ~50% of the total signal and ~5% of all detected peptides giving only modest improvement over the filter-based wash method (**Fig. 4.3F-G**).

As a third approach, we replaced the BSA resuspension solution with protease-resistant BSA (rBSA) prior to sorting. To replace BSA with rBSA, we used an altered protocol to lift adherent cells wherein we washed the adhered cells on the plate, incubated them in a dissociative reagent that does not require media to quench, and added a small amount of rBSA to the lifted cells to facilitate pelleting (see Methods). Cells were then sorted by FACS directly into a concentrated lysis buffer. This approach nearly eliminated signal arising from BSA peptides and produced highly correlated protein and peptide quantifications (**Fig. 4.1B-D, Fig. 4.4A-B**). Using this optimized protocol, we detected hundreds more unique proteins and thousands more peptides from low-input, FACS sorted cells, representing a ~10% increase in proteome depth (**Fig. 4.1E-G**). Therefore, this approach was used for all subsequent VCS-MS experiments.

VCS-MS can measure proteomes in visually distinct cells

To determine the ability of VCS-MS to differentiate between proteomes of visually distinct cells, we co-cultured human RPE1 cells expressing a histone tagged with a near infrared

fluorescent protein (H2B-miRFP) and mouse 3T3 cells with no fluorescent marker. Both cell types also express the photoconvertible fluorescent protein Dendra2 localized to the nucleus. Since only the human cells express H2B-miRFP, we used miRFP fluorescence to mimic a visual phenotype (**Fig. 4.2A**).

We applied VCS-MS to an equal mixture of co-cultured mouse and human cells (**Fig. 4.2A**). Using custom CellProfiler scripts, we segmented mouse and human cell nuclei based on Dendra2 fluorescence. We then activated Dendra2+/miRFP+ nuclei with 300 ms and Dendra2+/miRFP- nuclei 1,200 ms laser pulses to photoconvert different amounts to Dendra2 in mouse versus human cells which enabled subsequent FACS to separate each population. To assess whether activation bin affected the resulting sample purity, we also completed the experiment with the inverse activation schema, activating Dendra2+/miRFP+ nuclei with 1,200 ms and Dendra2+/miRFP- nuclei 300 ms laser pulses (**Fig. 4.2B-C**).

After isolating the visually distinct mouse and human cells, we analyzed their proteomes using a data-independent acquisition (DIA) MS method. Across all samples, we detected ~72,000-97,000 peptides and ~7,700-9,800 unique proteins from 50,000 cells per sample (**Fig. 4.5A-B**). The method is highly reproducible with good correlation (Pearson's $r \sim 0.9$) observed between biological replicates (**Fig. 4.5C**). Previously, we found that VCS did not substantially alter cell transcriptomes (Hasle et al. 2020). Likewise, the peptide abundances of high and low activation samples of the same cell type were highly correlated and clustered together demonstrating that VCS does not appreciably affect the proteome either (**Fig. 4.5C**). We filtered the resulting proteomes for human- and mouse-specific peptides and assessed how accurately VCS-MS separated cell subpopulations. We detected ~7-20% of the total signal from human peptides in mouse samples and ~7-12% from mouse peptides in human samples (**Fig. 4.2B-C**). FACS-based separation based on miRFP fluorescence resulted in ~7% total signal from human peptides in mouse samples, and ~2% from mouse peptides in human samples. Thus, compared to FACS-based separation, VCS-MS performed similarly (**Fig. 4.2B-C**). Low activation samples

had higher amounts of signal from the opposite species (**Fig. 4.2B-C**), likely due to Dendra2 turnover after ~13 hours of imaging.

Discussion

Visual phenotypes are powerful readouts for drug screens to identify new candidate compounds (Rohban et al. 2017). Moreover, high content imaging, transcriptomics, and proteomics provide complementary information about cell state (Way et al. 2022). However, visuospatially resolved proteomics technologies are limited, largely due to specialized and expensive instrumentation. In response, we developed VCS-MS, a flexible and inexpensive workflow that combines visual cell sorting with unbiased proteomics to quantify thousands of proteins in visually distinct subpopulations of cells. We demonstrated that VCS-MS can physically isolate visually distinct cells with high fidelity and can quantify thousands of proteins, laying a foundation for future applications.

A notable strength of VCS-MS is its compatibility with live cells, avoiding several challenges introduced by other microscope-based MS workflows that rely on laser capture microdissection (Mund et al. 2022; Piehowski et al. 2020; Dewez et al. 2020). Since VCS-MS uses fluorescent labeling to separate cells, it does not require the extensive sample processing needed for laser capture microdissection. Moreover, the high powered laser used to excise regions of interest could damage cells and proteins, especially when used to collect individual cells and nuclei. VCS-MS's use of fluorescent activation with a ~1 second pulse of UV light avoids these circumvents these issues. Finally, VCS-MS is exceptionally well suited for *in vitro* perturbations and workflows requiring live cells. For example, VCS-MS could be readily paired with Stable Isotope Labeling with Amino acids in Cell culture (SILAC) to measure protein degradation and synthesis (Doherty et al. 2009). Changes in proteostasis networks can be missed if changes in synthesis and degradation are balanced. By pairing VCS-MS with SILAC, additional

dimensions of drug-induced morphological heterogeneity can be explored. Moreover, our workflow would provide unprecedented control of SILAC culture conditions since sample comparisons can be made between cells sorted from the same well.

Visuospatially resolved proteomics methods like VCS-MS are critical to elucidate the link between cell morphology and cell state. For example, Image2Omics is a deep learning model that predicts proteomic cell state directly from images. The algorithm was trained on images and bulk proteomics datasets from M1 and M2 stimulated macrophages and can predict proteomic cell state at ~38-41% accuracy from images alone (Mehrizi et al. 2023). Inaccuracy in Image2Omics prediction may be driven by cell heterogeneity as even clonal cell populations often have heterogeneous morphological responses to stimuli (Slack et al. 2008). Thus, bulk proteomic measurements may not accurately represent a heterogeneous population (Altschuler and Wu 2010). Since VCS-MS can select for visual phenotypes amongst a heterogeneous population, cell morphologies and proteomes can be more accurately linked, resulting in more robust training data than bulk measurements for tools like Image2Omics.

In future work, the VCS-MS workflow can be expanded to interrogate many distinct cellular morphologies within a single experiment. As Dendra2 is tunable to up to 4 bins using different activation times (Hasle et al. 2020), multiple distinct cell morphologies can be isolated for proteomic analysis. However, as Dendra2 activation occurs while cells are alive, turnover of Dendra2 leads to dilution of the ratio of activated to unactivated Dendra2 within each cell over time. This leads to cells in the high activation bin dropping into lower bins or cells in the low activation bin being lost from the assay. Exploring methods that fix cells during imaging could overcome this issue. However, dissociating intact, fixed cells from the imaging dish is challenging and could affect proteomic depth. However, these improvements to VCS-MS would

enable its use for more time-sensitive phenotypes, longer imaging times, and additional phenotypic bins.

Finally, VCS-MS's flexibility means that it will continue to improve alongside new MS instrumentation and technologies. For example, samples collected with VCS could be enriched for phosphorylated peptides (Leutert et al. 2019; Villén and Gygi 2008) or other post-translational modifications to investigate changes in protein signaling in morphologically distinct subpopulations. Furthermore, VCS-MS could be used to measure other protein biophysical properties, like thermodynamic stability using thermal proteome profiling (Mateus et al. 2022). VCS-MS's compatibility with other methods of interrogating proteomic cell state positions it as an especially versatile technology.

Overall, VCS-MS is an inexpensive, accessible, reproducible and powerful proteomics method capable of quantifying >7,700 proteins from visually distinct cell subpopulations. VCS-MS can elucidate the link between visual phenotypes and proteomic cell state, improving high content screening and enabling the study of heterogeneous cell response to chemical perturbation.

Methods

Cell lines

hTERT immortalized RPE-1 cells expressing NLS-Dendra2x3/H2B-miRFP/NES-mBeRFP and 3T3 cells expressing NLS-Dendra2x3 were generated as previously described (Hasle et al. 2020). In short, lentiviral vectors for each fluorescent construct were transduced into their respective parental lines (ATCC, RPE-1: CRL-4000, 3T3: CRL-1658), and single-cell sorted. Cells were sorted for high green fluorescence prior to mixing experiments.

Visual Cell Sorting imaging and activation

A Leica DMi8 inverted microscope was outfitted for Visual Cell Sorting as previously described (Hasle et al. 2020). The microscope has live cell imaging capabilities and a digital micromirror device with a 405 nm laser capable of activating at ~1 μm resolution. The automated image acquisition was controlled with the Metamorph Advanced Image Acquisition Software package (v7.10.1.161; Molecular Devices, San Jose, CA). Prior to imaging, a 6-well glass-bottom, black-walled plate was seeded with an equal mixture of RPE-1 and 3T3 cells at 400,000 total cells per well in media without phenol red. After 24 hours, the plate was imaged with custom journals to segment nuclei, classify images based on visual phenotype, and activate classified nuclei for 300 ms or 1200 ms (available on GitHub at https://github.com/FowlerLab/VCS-MS).

Visual Cell Sorting fluorescence-activated cell sorting (FACS)

After imaging and activation for ~13 hours, the 6-well plate was removed from the microscope, and each well was washed three times with DPBS to remove proteins from cell culture media. Cells were dissociated by incubating in 500 μl ACCUTASE™ (STEMCELL Technologies, cat. no. 07920) at 37℃ for 5 minutes. The ACCUTASE™ was deactivated by dilution using 2.5 ml DPBS per well, and 30 μl of 0.5% acetylated BSA (Sigma-Aldrich, B8894; rBSA) was added to each well to facilitate pelleting. Cells from each well were spun for 5 minutes at 300 x *g*, the supernatant was aspirated, and pellets were resuspended in 0.5% rBSA before FACS.

Cells were sorted using a FACS Aria III (BD Biosciences). A FSC-A vs. SSC-A plot was used to gate for live cells, and FSC-W vs. FSC-A was used to gate for single cells. Dendra2-positive cells were gated using a FITC-A histogram plot. “High activation” (1,200 ms), “low activation” (300 ms) and “no activation” (0 ms) cell populations were then selected using a PE-YG-A vs. FITC-A plot. Activated cells were sorted into 2 ml Protein LoBind® tubes (Eppendorf, cat. no. 022431102) containing 20 μl 10x lysis buffer (20% SDS, 150 mM NaCl, 0.5 M Tris·Cl, pH 8.0). Using an 85 μm nozzle, 50,000 sorted cells resulted in a final volume of ~220 μl, diluting the

lysis buffer to a 1x working concentration. Samples were snap frozen using liquid nitrogen and stored at -80℃ until sample preparation.

Denaturing cell lysis, protein reduction, and alkylation

Cells sorted into the concentrated lysis buffer were sonicated in a water bath for 20 minutes. Proteins were reduced with 3 mM dithiothreitol (DTT) for 30 minutes at 55℃ and alkylated with 9 mM iodoacetamide for 30 minutes at room temperature in the dark. The alkylation reaction was quenched with an additional 3 mM DTT.

Whole proteome sample preparation

Denatured lysates were digested using R2-P1, an automated, bead-based sample preparation method based on Single-Pot, Solid-Phase Sample Preparation (Hughes et al. 2019; Leutert et al. 2019). Briefly, samples were diluted to 75% ethanol in a KingFisher 96 microtiter deep-well plate containing 0.5 mg/ml carboxylated magnetic beads and processed using a magnetic particle processing robot (KingFisher™ Flex). The protein-bound magnetic beads were washed in four plates containing 80% ethanol. The beads were then deposited into 50 mM Tris·Cl, pH 8.9 and digested using 0.01 ug/ul LysC (Wako Chemicals) for 4 hours at 37℃. The beads were incubated in a final water wash and the water wash was combined with the digestion solution. Digestions were acidified to pH 2 with trifluoroacetic acid and desalted over Empore C18 stage tips (Rappsilber, Mann, and Ishihama 2007).

Mass spectrometry data acquisition

Peptide samples were resuspended in a 4% formic acid, 3% acetonitrile solution and analyzed by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) on a hybrid quadrupole orbitrap mass spectrometer (QExactive or Orbitrap Exploris 480, Thermo Fisher Scientific). Samples were loaded 100 µm ID x 3 cm precolumn packed with Reprosil C18 3 µm

beads (Dr. Maisch GmbH) and separated by reverse phase chromatography on a 100 µm ID x 30 cm analytical column packed with Reprosil C18 1.9 µm beads (Dr. Maisch GmbH) housed in a column heater set to 50℃. Peptides were separated using a gradient of 7-50% acetonitrile in 0.125% formic acid delivered at 450 nl/min over 95 minutes with a total 120 minute acquisition time.

For all experiments, the mass spectrometer was operated in data-dependent acquisition mode. For each cycle, one full mass spectrometry scan was acquired from 300 to 1500 m/z at 70,000 resolution with a fill target of 3E6 ions and automated calculation of injection time. The 20 most abundant ions from the full MS scan were selected for fragmentation using 2 m/z precursor isolation window and beam-type collisional-activation dissociation (HCD) with 30% normalized collision energy. MS/MS spectra were acquired at 30,000 resolution by setting the AGC target to standard and injection time to automated mode. Fragmented precursors were dynamically excluded from selection for 40 seconds.

For experiments comparing VCS on mixed human and mouse cells, samples were also measured in data-independent acquisition (DIA) mode. The DIA acquisition scheme was set up based on Pino *et al. (Pino et al. 2020)*. For DIA measurements, 30 MS/MS DIA spectra were acquired in 24-m/z precursor isolation windows covering 363-1095 m/z using a staggered window pattern at 30,000 resolution with an AGC target of 1E6, automated injection time and a normalized collision energy of 27. Precursor spectra were acquired after every 30 MS/MS at 60,000 resolution with standard AGC target and automated injection time.

Mass spectrometry data analysis

The human and mouse reference proteomes FASTA databases were downloaded from Uniprot (human: UP000005640, downloaded in 2023; mouse: UP000000589, downloaded in 2023).

DDA data was searched using Comet (release 2019.01.2; (Eng, Jahan, and Hoopmann 2013)). The precursor mass tolerance was set to 20 ppm. Constant modification of cysteine carbamidomethylation (57.0215 Da) and variable modifications of methionine oxidation (15.9949 Da), N-terminal acetylation (42.0106 Da), and lysine acetylation (42.0106 Da) were used for all searches. A maximum of 2 of each variable modification was allowed per peptide. Search results were filtered to a 1% false discovery rate at the peptide spectrum match level using Percolator (Käll et al. 2007).

To analyze DIA data, a spectral library was first generated using both DDA and DIA using MSFragger-DIA in the FragPipe computational platform (Kong et al. 2017; Yu et al. 2022). Standard search parameters were used including fixed modification of cysteine carbamidomethylation and variable modification of methionine oxidation. A PSM and peptide FDR cutoff of < 0.01 were chosen. DIA-NN (v1.8.1)(Demichev et al. 2020) was then used with the generated spectral library to search and quantify the DIA raw files using the “Robust LC (high precision)” setting.

Bioinformatics

If not specified otherwise, R version 4.3.1 (https://www.r-project.org/) with the “tidyverse” package collection (https://www.tidyverse.org) and “protti” package collection (https://github.com/jpquast/protti) was used for all analyses. Code is available on GitHub (https://github.com/FowlerLab/VCS-MS)

Conclusion

The presence of meaningful heterogeneity introduces prominent challenges to the study of biological systems. My work addresses some of these challenges by employing and developing tools to study heterogeneity across organisms and in clonal cells. At both the organismal and cellular level, detecting and understanding subpopulations is critical to delivering effective treatments and avoiding harmful ones.

As discussed in Part I, heterogeneity amongst populations of humans can obscure the application of evidence-based medicine. As a response, the field of pharmacogenomics seeks to tailor drug dosing to a patient's genetic makeup. Although identifying variants in patients is becoming easier, knowing the appropriate clinical adjustment remains challenging since the consequences of most variants are unknown. To remedy this, I completed the first comprehensive analysis of variant effects in two CYP homologs, 2C19 and 2C9, whose function is critical for elimination and bioactivation of hundreds of drugs. In Chapter 2, I described how the results of this study revealed regions where CYP2C19 and 2C9's variant abundance differed, possibly explaining functional differences between the two highly similar homologs.

The joint analysis method I used, *multidms*, is capable of analyzing more than two DMSs, making it a powerful tool to continue exploring the mutational landscape of CYP homologs. Incorporating DMSs from other CYPs that are important for pharmacogenomics, like CYP2D6 and 3A4, would provide substantial insight into variant effect commonalities and differences across CYPs. For example, deleterious positions with shift values near zero would indicate positions that are indispensable for abundance. Conversely, positions with large shift values may highlight important differences in protein folding between CYPs, and could warrant followup studies to discern the mechanistic underpinnings.

One key limitation to the VAMP-seq assay is that it can only identify variants whose loss of function is due to decreased abundance. Therefore, variants that have disrupted enzymatic function but normal abundance may be missed. For example, in Chapter 2, I discovered six positions in CYP2C19 that were surprisingly tolerant to mutation despite being highly conserved and located in the hydrophobic core where most variants are deleterious. These positions are critical for enzymatic function, cofactor binding, and substrate selectivity in the closely related CYP2C9, meaning they likely serve a similar purpose in CYP2C19. Thus, abundance measurements alone can miss variants that may impact enzymatic function.

A possible modification to VAMP-seq to provide a readout of functionality is to integrate substrate binding into the abundance measurements. Ligand binding stabilizes proteins, altering their thermodynamic stability (Mateus et al. 2022). Thus, variants bound to a given substrate may have shifted thermodynamic stability, and therefore steady-state abundance, relative to their unbound counterpart. By using VAMP-seq to measure variant abundance in the presence and absence of a substrate, this method could be used to identify variants and positions that are important for substrate specificity at scale. For example, cells expressing a CYP2C19 library could be treated with diclonefac, which is a substrate for CYP2C9 but not 2C19. Variants in CYP2C19 that enable diclonefac binding may have higher abundance in the diclonefac condition than in the vehicle condition. If feasible, this method would be a massive increase in throughput compared to current approaches that rely on individual measurements of substrate metabolites by purified CYP protein.

While VAMP-seq, like other DMS methods, can readily measure all single amino acid variants, limitations in assay throughput and library construction prevent adequate coverage of all possible CYP double mutants. Single amino acid deviations from WT can substantially alter

variant effect (Faber et al. 2019), indicating that even the addition of a second mutation could reveal many previously unknown epistatic interactions. Since creating variant libraries is costly and labor intensive, an alternative is to modify an existing barcoded variant library by inserting a new mutation to replace the WT background with a new background. However, this approach would require high efficiency to avoid introducing noise to the subsequent assay. For example, if only a portion of barcode-variant library molecules successfully receive the new background, a given barcode may represent a variant on both the WT and the new background. Thus, an epistatic effect driven by the new background may be obscured. Recent innovations in site directed mutagenesis have enabled >93% efficiency in plasmids up to 13.2 kb (K. Zhang et al. 2021), which is compatible with the ~6.8 kb VAMP-seq vector employed in Chapter 2. Followup VAMP-seq studies on CYP2C19 double mutants containing the variants with highest population frequency in gnomAD, like I331V (*1b) and E92D (not reported in ClinVar), could elucidate non-additive variant effects which are notoriously difficult to predict computationally.

A final limitation to VAMP-seq assays is that our model is not a perfect representation of the native expression environment. Although it uses mammalian cells, the CYP enzyme is encoded by a cDNA that lacks intronic regions, it is driven by an ectopic promoter, and the fused GFP reporter may alter the stability of some variants. Thus, the abundance readout in VAMP-seq may not fully recapitulate CYP expression in its native environment.

In Part II, I described VCS-MS, a new technology that can be used to separate visually distinct subpopulations in heterogeneous cell mixtures, a common feature in drug treatment of clonal cells (Hasle et al. 2020; Slack et al. 2008; Raatz et al. 2021). VCS-MS extends Visual Cell Sorting (Hasle et al. 2020) for use with proteomics. As a benchmark for the method, I separated an equal mixture of mouse and human cells, and quantified species-specific proteins. VCS-MS

separated the visually distinct mouse and human cells with high fidelity, demonstrating its potential for use in other applications.

Methods that directly pair MS measurements with visuospatial coordinates, like MSI, are currently the predominant approaches employed to study visually resolved phenotypes. However, separating phenotype selection and sample isolation from the MS analysis has drastically improved proteome depth and coverage. Technologies like Deep Visual Proteomics use laser capture microdissection to excise regions or cells of interest, deposit them in a 96- or 384-well plate, then process the samples with low-input MS (Mund et al. 2022). As an alternative, VCS-MS marks cells with visual phenotypes of interest using a fluorescent tag and isolates them using FACS. While similar in concept, VCS-MS and DVP have distinct use cases.

Since DVP relies on laser capture microdissection, it is capable of isolating hundreds of cells of interest (Mund et al. 2022) whereas VCS-MS can isolate tens of thousands. The higher throughput of VCS-MS means more biological material is collected, making it more amenable to analyses that require high amounts of protein lysate, like phosphoproteomics (Leutert et al. 2019). Moreover, since VCS-MS is compatible with live cells, it can be more readily used to measure protein turnover with metabolic labeling, like with SILAC (Doherty et al. 2009). Thus, VCS-MS is better suited for cell culture applications than DVP.

On the other hand, DVP has been applied to patient tissue samples, separately analyzing cancerous tissue from nearby normal tissue (Mund et al. 2022). Currently, VCS-MS relies on an engineered cell line expressing the photoconvertible green fluorescent protein, dendra2. Thus, it cannot yet be applied to patient samples. However, VCS-MS could be applied to tissue slices with an alternative photoconversion system. For example, an antibody could be conjugated to a photoconvertible fluorescent reporter. Using an antibody reporter system would also enable

selection based on other visual phenotypes beyond cell or nuclear morphology like protein or organelle distribution.

While a strength in some applications, VCS-MS's restricted use to live cultured cells also has limitations. During automated imaging and activation, the cells are still growing, dividing, and turning over their proteome. Therefore, photoconverted dendra2 is consistently being degraded and replaced with new, green dendra2, aberrantly reducing the ratio of green:red dendra2 that is used to separate VCS-activated cells. As a result, wells that are imaged later in the experiment have better separation than those imaged earlier. One solution is to create a Tet-Off dendra2 cell line to halt dendra2 production at the start of imaging upon introduction of doxycycline. This solution would maintain the post-activation green:red fluorescence value since the cells would be degrading all dendra2, activated and non-activated, at an equal rate, and not replacing it with green dendra2. Another solution is to modify the VCS-MS workflow for use with fixed cells, halting all dendra2 turnover. However, cell fixation introduces additional challenges. To measure whole cell proteomes, the cells need to be dissociated from the imaging dish intact prior to FACS, potentially requiring the use of enzymatic cleavage and fixation reversal that could complicate downstream MS analysis. If optimized, VCS-MS on fixed cells would enable new applications, like unbiased proteomic analyses of cell cycle stages based on nuclear morphology without chemicals that stall cells in specific stages of cell division.

Future Directions

In the near-term, future directions in the fields of pharmacogenomics can be broadly categorized into either extensions of breadth or depth. Until recently, pharmacogenomics research has been occupied with questions of "depth" in low throughput, thorough characterization of small

numbers of CYP genes. Currently, however, the biggest contributions to progress in pharmacogenomics will be those that expand breadth.

First, pharmacogenomics and other genomics-based personalized medicine initiatives have recently been criticized for the need for the lack of breadth demonstrated by their underrepresentation of non-White groups in genomics datasets. Major initiatives to include underrepresented groups are already underway. For example, All of Us is a multi-institution program designed to sequence genomes from >1,000,000 Americans to speed up health research discoveries. An explicit goal of All of Us is to include participants from groups that have been underrepresented in health research in the past. In addition to discovering hundreds or thousands of new CYP alleles, research programs like All of Us will accurately measure allele frequencies within ethnic groups, which vary widely (Dai et al. 2015) and can have implications for drug dosing in patients with non-White ethnicities.

Second, variant discovery is vastly outpacing variant interpretation (Fayer et al. 2021). Indeed, prior to my work, nearly all of the >400 missense mutations in the gnomAD database had unknown consequences. Here, I showed that over half likely have decreased abundance (**Fig. 2.11**). Thus, multiplexed assessment of variant effects and *in silico* predictors, like AlphaMissense (Cheng et al. 2023), will be needed to close the gap.

At the cellular level, multi-omics promises to capture the immense complexity of biology (Bray et al. 2016; Subramanian et al. 2020; Hasin, Seldin, and Lusis 2017; Mehrizi et al. 2023). Pairing visual phenotypes with –omics data is still a relatively new field of research. Since most morphology-based proteomics or transcriptomics methods are cumbersome and expensive, there has been a recent focus on developing methods with lower barriers to entry, like Visual Cell Sorting (Hasle et al. 2020). Thus, research focus has been more on designing new

techniques rather than applying them to biological questions. In the next stages, this field will benefit most from close collaborations with clinicians and specialized biologists who focus on disease areas with known spatial and morphologic features, especially in cancer and neurology.

Collectively, this dissertation demonstrates the broad importance of studying heterogeneity, both between people and amongst clonal cell populations. As humanity's knowledge of our own biology expands, we continue to uncover new nuance and ever-expanding complexity. Thus, continuing to develop and employ technologies capable of investigating subpopulations is critical to a holistic understanding of human health and disease.

Tables

Table 2.1. CYP2C19 library fluorescence activated cell sorting.

Four-way sort of the HEK293T abundance CYP2C19 library. Cells were binned into equal 25% bins based on their GFP:mCherry fluorescence. The number of events sorted for each experiment, grown out for 24-48 hours prior to gDNA harvest.

CYP2C19 library fluorescence activated cell sorting				
Experiment number	**Cells sorted in Bin1**	**Cells sorted in Bin2**	**Cells sorted in Bin3**	**Cells sorted in Bin4**
1	7,514,524	6,522,331	7,325,607	6,920,966
2	7,518,266	7,296,360	7,435,898	7,671,996
3	6,938,146	6,860,475	6,933,813	7,144,143

Table 2.2. Library statistics from barcode-variant mapping.

Barcoded CYP2C19 library sequenced on a Sequel II v2.0. CCS reads (circular consensus reads) generated with ccs2. During subassembly, barcodes with more than 3 CCS reads were retained.

SMRT cells	1
CCS reads with 3 or more passes	1,910,631
Unique barcodes (coverage)	199,053 (9.6x)
Barcodes with one consensus read	12,360
Barcodes with two consensus reads	10,430
Barcodes with three or more concensus reads	176,372
Barcodes associated with WT CYP2C19 sequence	4,830
Barcodes associated with indel	38,712
Barcodes associated with two or more amino acid mutations	35,435
Barcodes associated with single amino acid mutation (mean, median barcodes per single amino acid mutation)	106,432 (11.87, 7)

# single amino acid mutations (percent possible)	9,055 (88.2%)
# unique nucleotide sequences	12,559

Figures

Figure 1.1

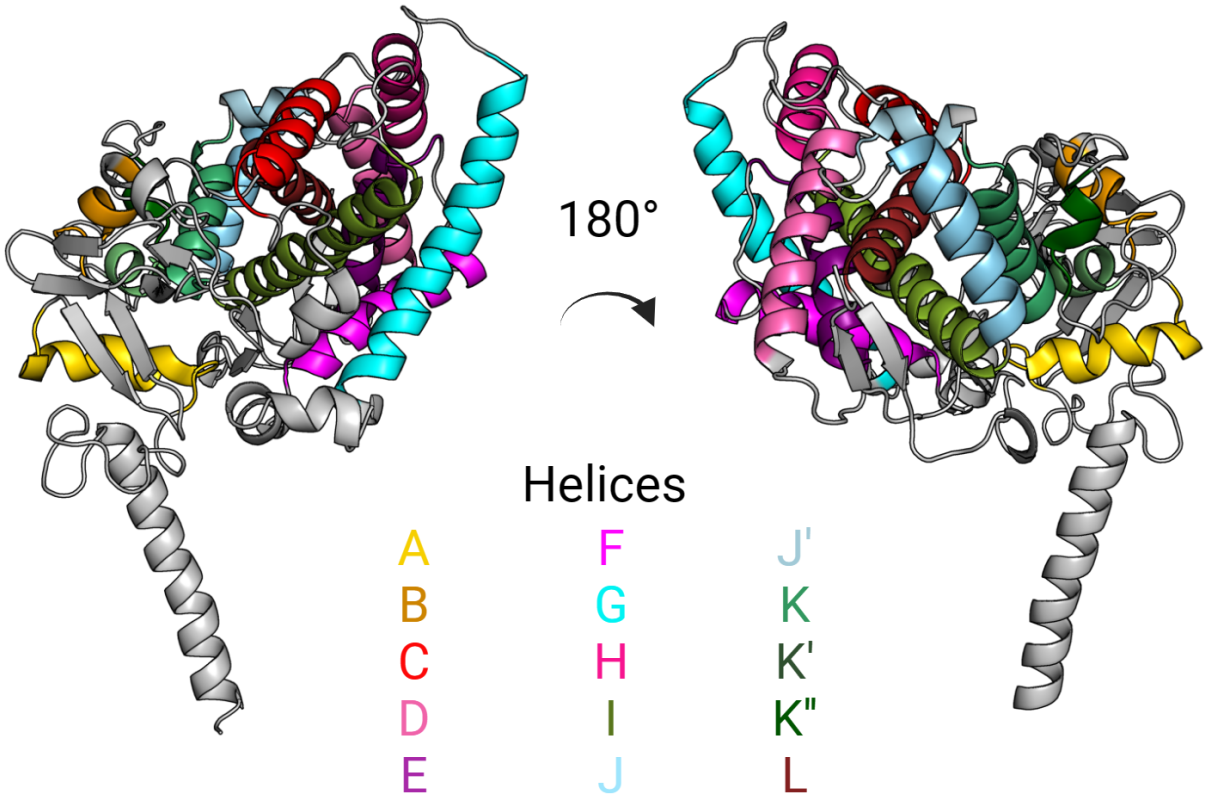


Figure 1.1. CYP2C19 structure colored by α-helix. Structure from molecular dynamics simulation (Mustafa et al. 2020) is depicted as a cartoon. Helix A is show in yellow, helix B in mustard, helix C in red, helix D in light pink, helix E in purple, helix F in magenta, helix G in cyan, helix H in pink, I in olive, J and J` in light blue, K, K` and K`` in shades of forest green, and L in brown.

Figure 1.2

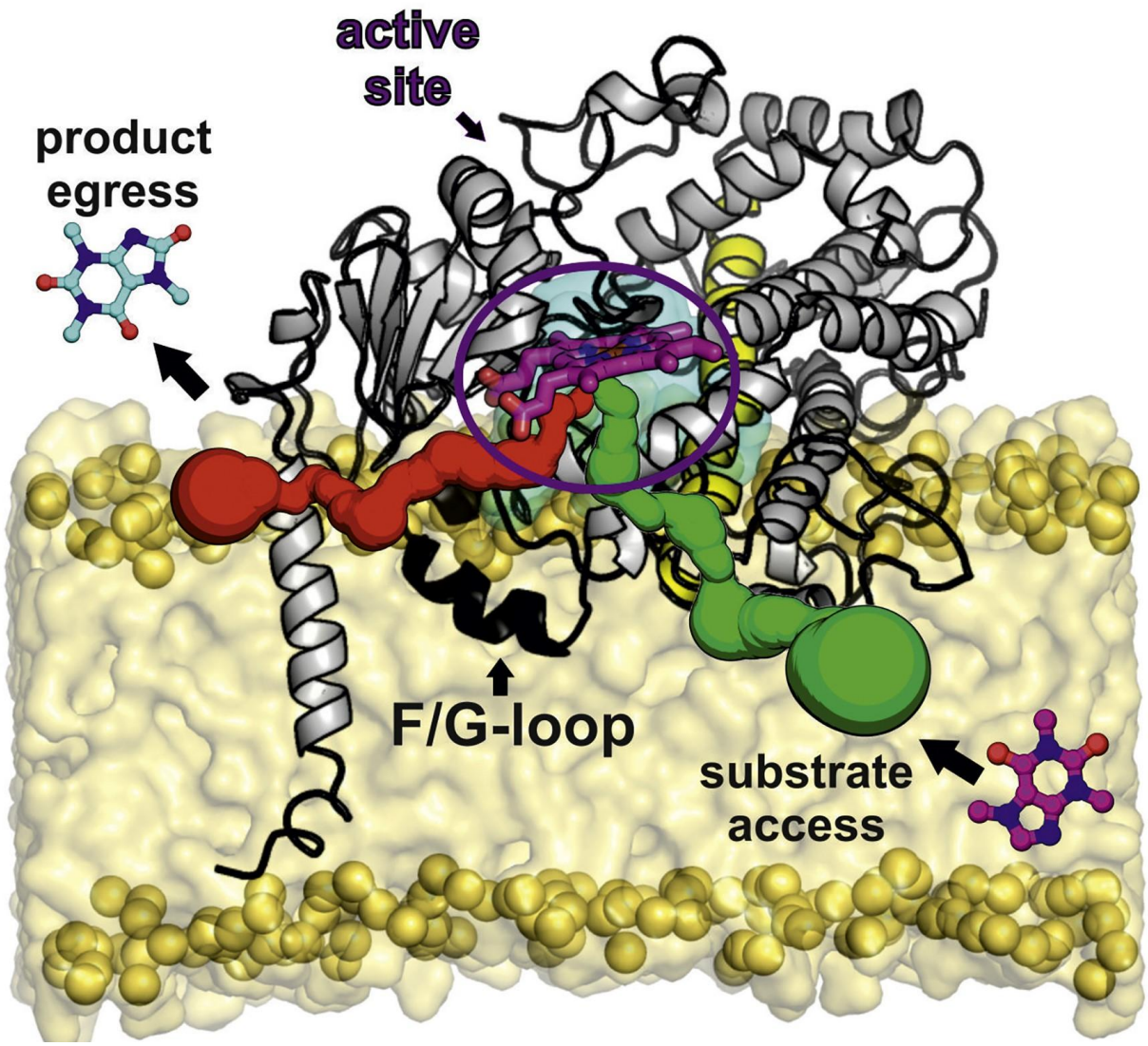


Figure 1.2. Localization of substrate access and product egress channels in the structure of CYP3A4. The F/G loop is highlighted in black, and the arrow shows the F`/G` loop that compose its tip. Amphiphilic compounds enter the CYP active site from the membrane interior via a family 2 channel shown in green. In the active site housing the heme cofactor (violet), they are transformed and the resulting products leave via the solvent channel shown in red. Adapted from “Membrane-attached mammalian cytochromes P450: An overview of the membrane's

effects on structure, drug binding, and interactions with redox partners" by Martin Šrejber, Veronika Navrátilová, Markéta Paloncýová, Václav Bazgier, Karel Berka, Pavel Anzenbacher, and Michal Otyepka, 2018, *Journal of Inorganic Biochemistry*, 183, 117-136; https://doi.org/10.1016/j.jinorgbio.2018.03.002.

Figure 1.3

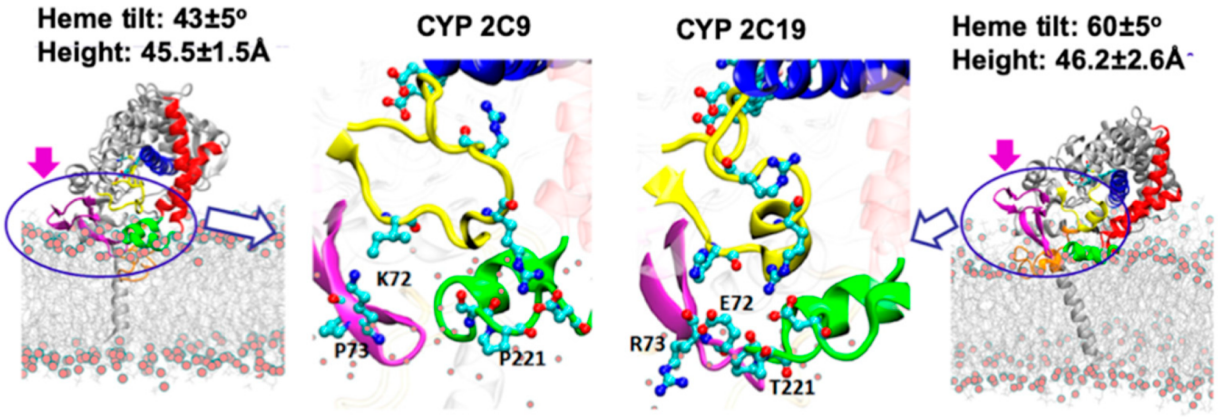
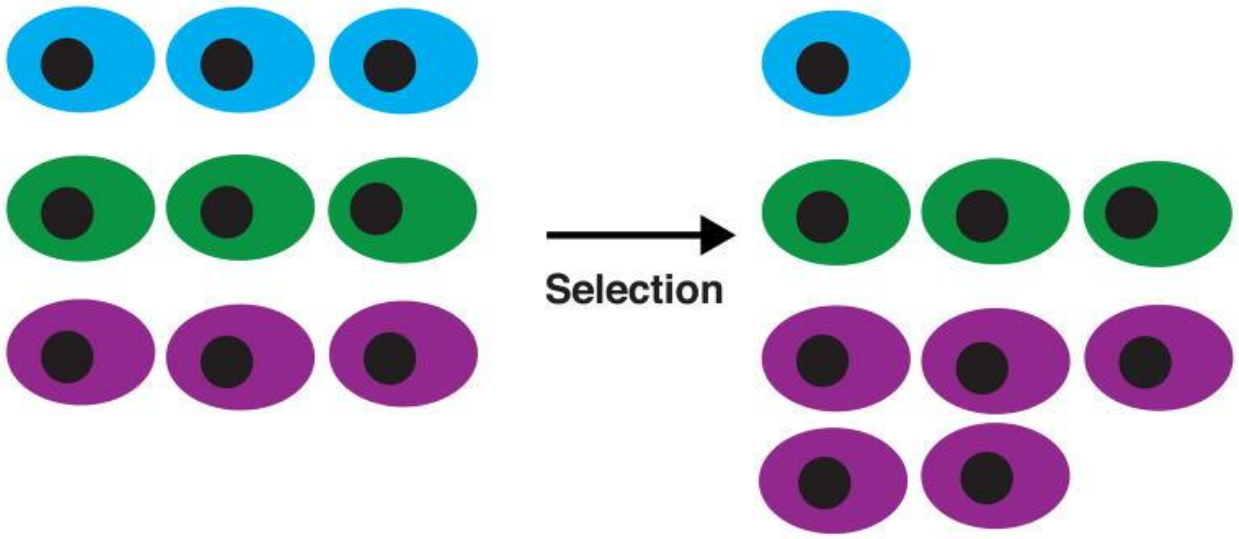


Figure 1.3 Differences in the arrangements of the CYP 2C9 and CYP 2C19 residues at the membrane interface. The last frames of from molecular dynamics simulations of CYP2C9 (left) and CYP2C19 (right) with full system and membrane interface region. Side chains are shown in ball-and-stick representation colored by atom type with cyan carbons. β-sheets 1 and 2 are shown in magenta, the B–C loop in yellow, the F and G helices in red, the F` and G` helices in green, and the central I-helix in blue. Adapted from "Differing Membrane Interactions of Two Highly Similar Drug-Metabolizing Cytochrome P450 Isoforms: CYP 2C9 and CYP 2C19" by Ghulam Mustafa, Prajwal P. Nandekar, Neil J. Bruce, and Rebecca C. Wade, 2019, *International Journal of Molecular Sciences, 20(18), 4328;* *https://doi.org/10.3390/ijms20184328*.

Figure 1.4



Variant	Mutation	Counts input	Counts selected	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67

Figure 1.4. Deep mutational scanning generates large-scale mutational data. Cells are transfected with a variant library containing all possible single amino acid variants. When placed under a selective pressure less functional variants drop out and variants with higher functionality expand. A functional score is calculated for each variant by comparing the number of cells with each variant in the input and after selection. Adapted from “Deep mutational scanning: a new style of protein science” by Douglas M Fowler and Stanley Fields, 2014, *Nature Methods,* 11(8):801-7. doi: 10.1038/nmeth.3027. Copyright © 2014.

Figure 2.1

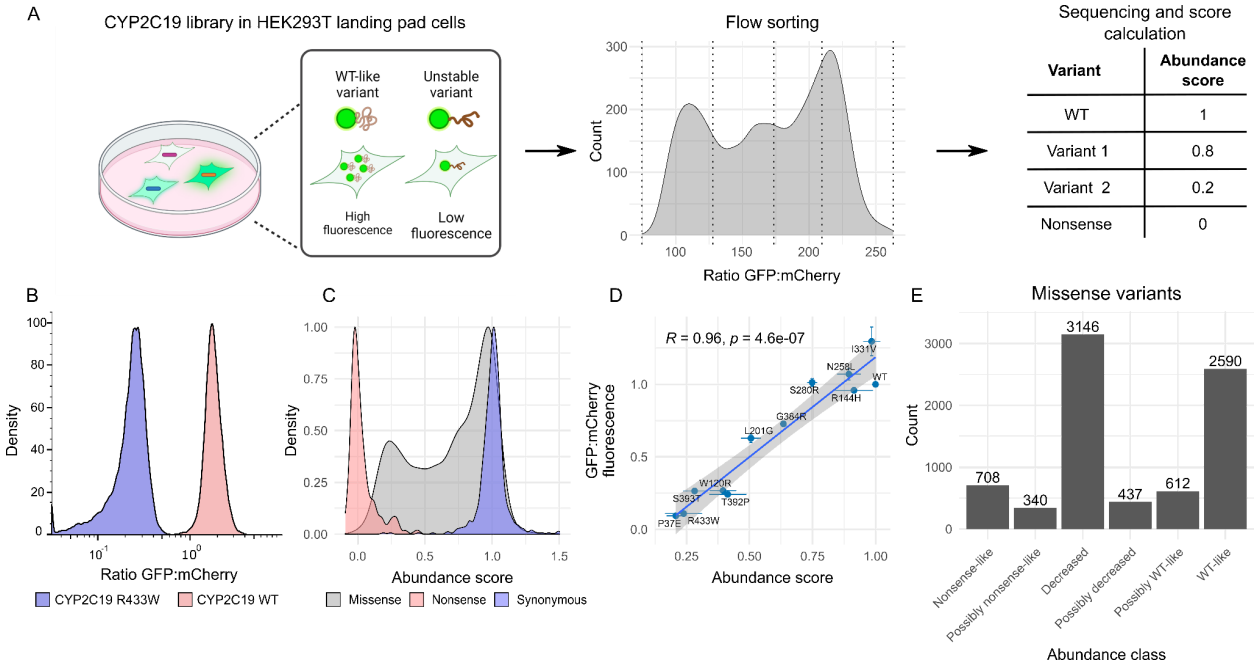


Figure 2.1. Multiplexed measurement of CYP2C19 abundance. Variant assessment by massively parallel sequencing (VAMP-seq) measures variant abundance at scale. A) In VAMP-seq, a barcoded library fused to GFP is recombined into a genomically integrated landing pad in HEK293T cells. mCherry is expressed co-transcriptionally via an IRES. Unstable variants are degraded by the proteostasis machinery of the cell, resulting in lower GFP signal compared to wild type (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. B) GFP:mCherry ratio for cells expressing either CYP2C19 WT (red) or the R433W destabilizing variant (blue) (n ~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 standard deviation of abundance scores (x-axis) or mean fluorescence (y-axis). E) Number of missense variants in each abundance class.

Figure 2.2

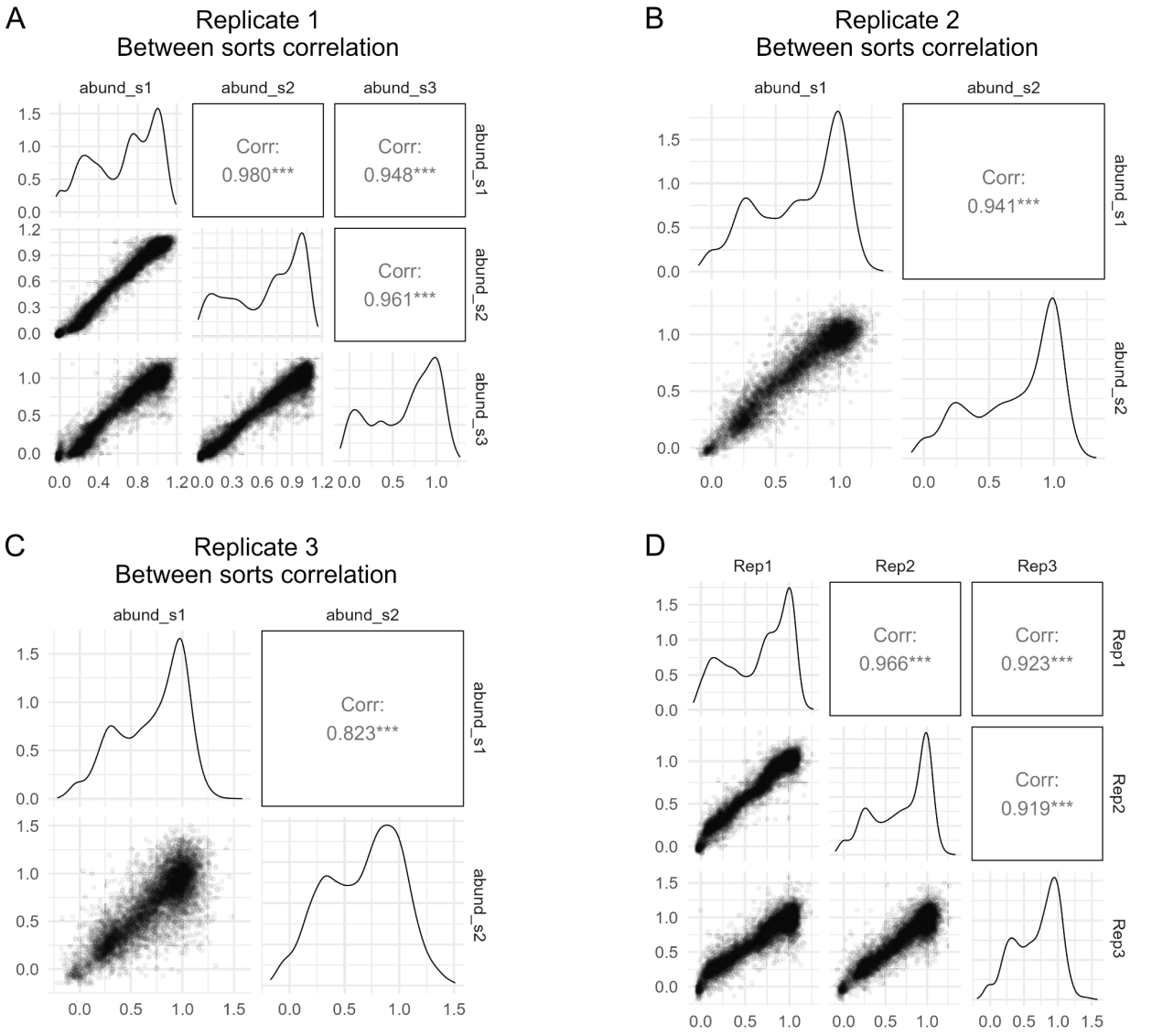


Figure 2.2. Abundance score correlation matrices

A-C) Scatter plots showing correlation between variant abundance scores between each sort for three replicates. D) Scatter plot showing correlation of variant abundance score across all replicates. Sorts were combined by averaging variant abundance scores.

Figure 2.3

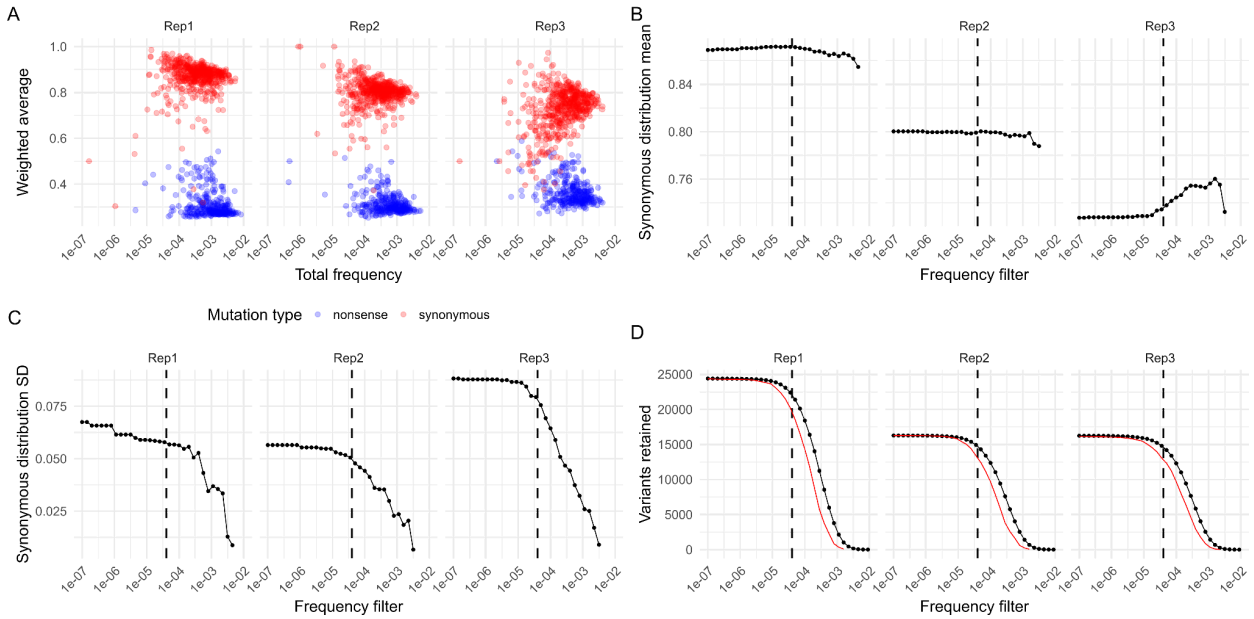


Figure 2.3. Determining variant frequency filters

A) Scatter plots of variant weighted averages and frequencies for synonymous (red) and nonsense (blue) variants across three replicates. B-D) Frequency filters for abundance scores. Dot plots show the synonymous distribution means (B), standard deviations (C), and number of variants (D; black) or 14 times the number of synonymous variants (red) for each replicate across frequency cutoffs. For plots E-H, the frequency filter of $4x10^{-5}$ used is shown as a dotted line.

Figure 2.4

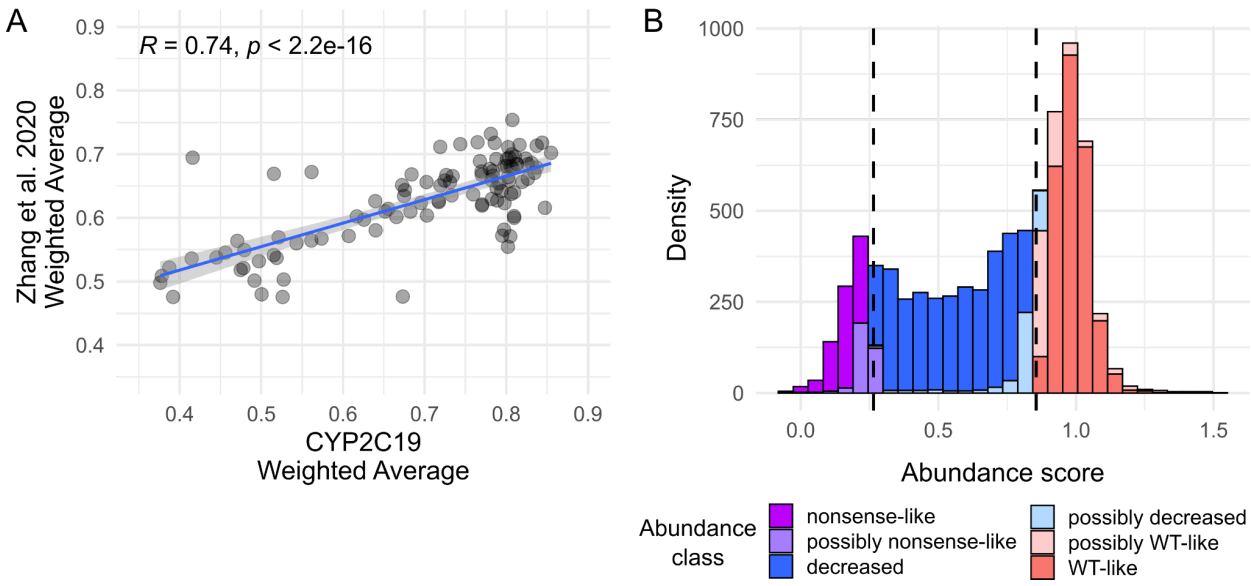


Figure 2.4. Categorization of variant abundance scores into classes

A) Scatterplot of CYP2C19 abundance weighted averages compared between our assay and another VAMP-seq assay completed by Zhang et al, 2020(L. Zhang et al. 2020). B) Missense variant abundance score histogram colored by abundance class. In dotted lines, scores lower than the 5th percentile of the synonymous distribution (right) and 95th percentile of the nonsense distribution (left) were used for classification. Variant abundance class was determined by whether their scores and confidence intervals fell in the synonymous and nonsense distribution thresholds, as described in the Methods.

Figure 2.5

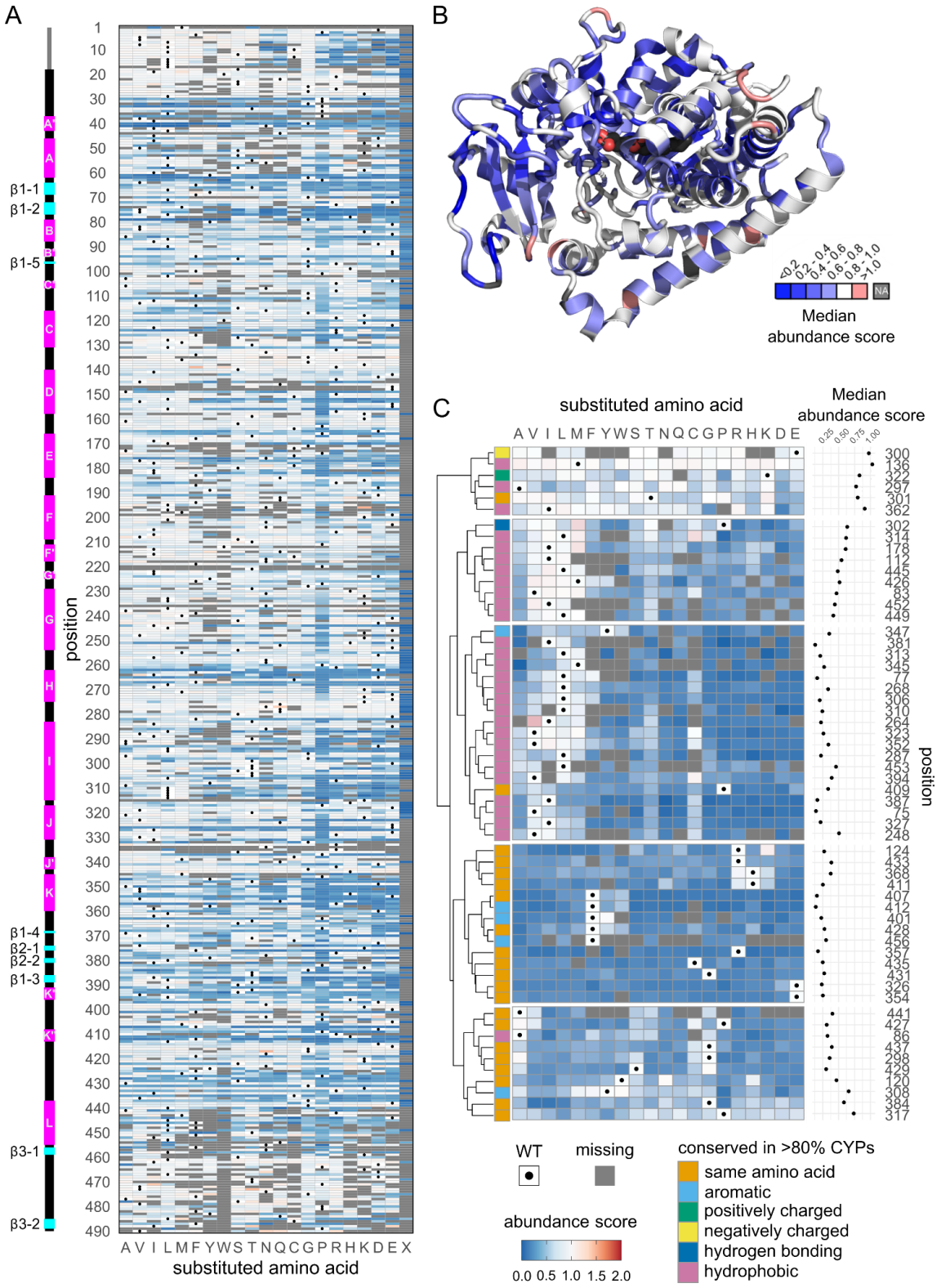

A
B
C
position
substituted amino acid
A V I L M F Y W S T N Q C G P R H K D E X
β1-1
β1-2
β1-5
β1-4
β2-1
β2-2
β1-3
β3-1
β3-2
<0.2
0.2 - 0.4
0.4 - 0.6
0.6 - 0.8
0.8 - 1.0
>1.0
NA
Median abundance score
Median abundance score
0.25
0.50
0.75
1.00
WT
missing
abundance score
0.0 0.5 1.0 1.5 2.0
conserved in >80% CYPs
same amino acid
aromatic
positively charged
negatively charged
hydrogen bonding
hydrophobic

Figure 2.5. 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. 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 binned median score, with missing scores represented in gray. Heme is colored by element (carbon: black, nitrogen: blue, oxygen: red, iron: yellow). C) Hierarchical clustering of CYP2C19 abundance scores 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, Vogel, and Pleiss 2014).

Figure 2.6

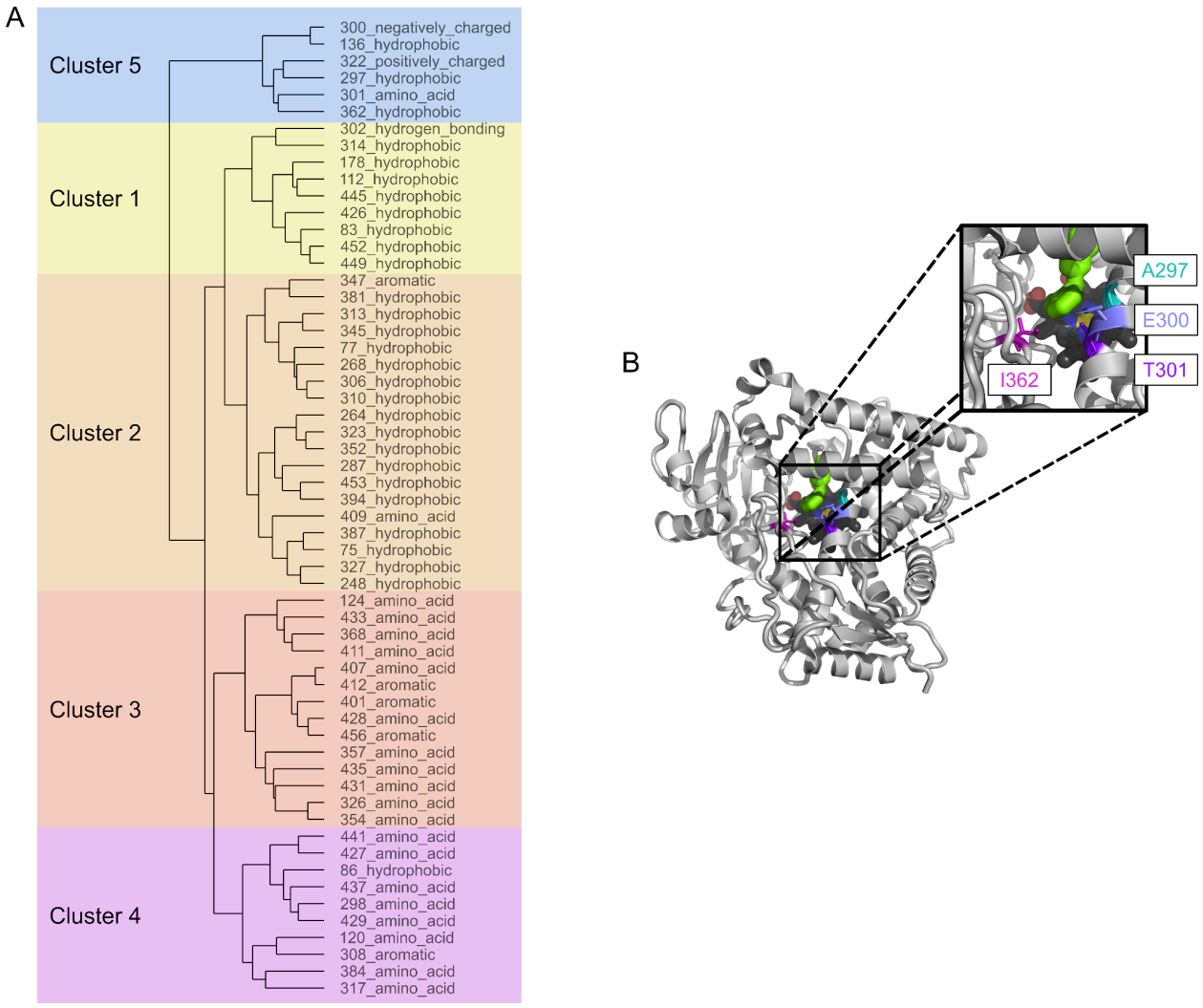


Figure 2.6. Mutationally tolerant conserved sites in the hydrophobic core.

A) Dendrogram showing hierarchical clustering of sites conserved in >80% eukaryotic CYPs (Gricman, Vogel, and Pleiss 2014). B) CYP2C19 structure (PDB: 4GQS). Heme is colored by element (carbon: black, nitrogen: blue, oxygen: red, iron: yellow), and PDB chemical 0XV in green. Positions A297 (cyan), E300 (lavender), T301 (purple), and I362 (magenta) are located in the hydrophobic core near heme.

Figure 2.7

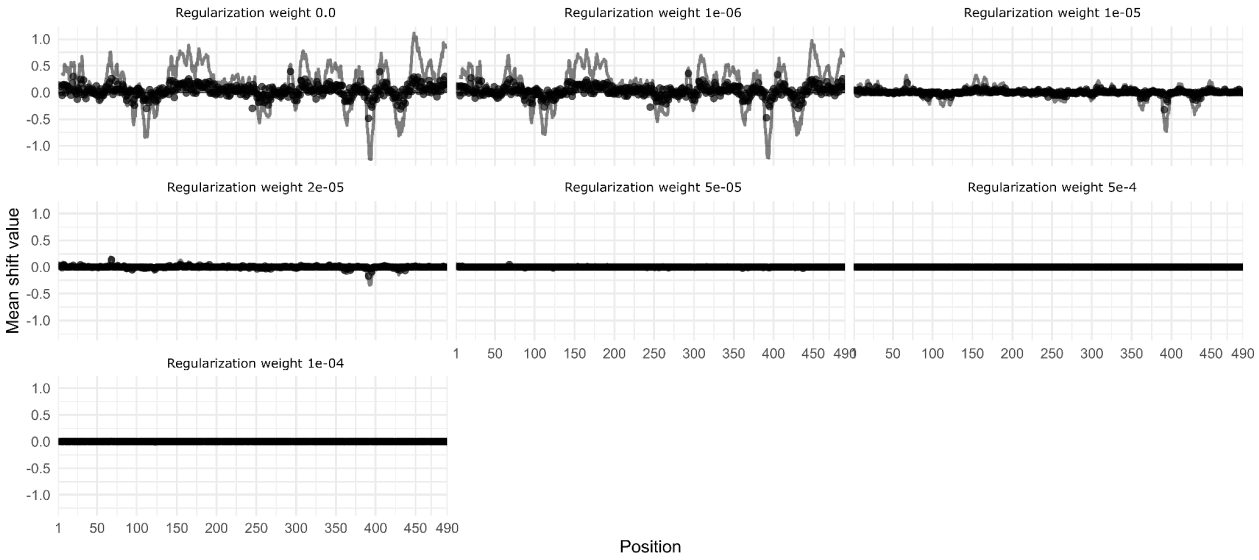


Figure 2.7. CYP2C19 shift values across regularization weights. Scatter plot depicting mean CYP2C19 shift value per position. Gray line is the rolling sum of the mean shift values with k=5. Shift values were calculated using regularization weights ranging from 0.0 to 1e-4. A weight of 1e-5 was selected as the optimal value.

Figure 2.8

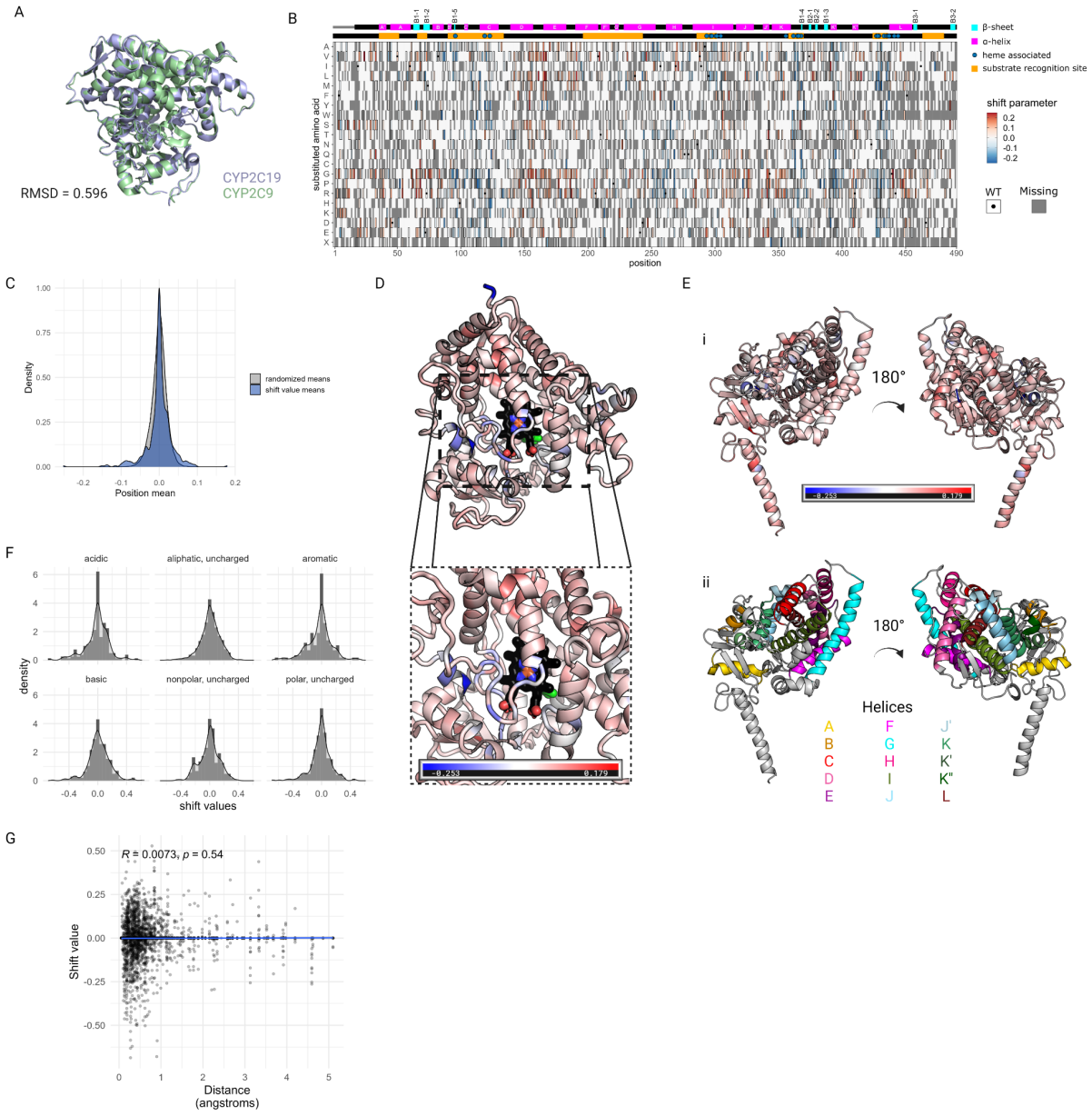


Figure 2.8. CYP2C19 and CYP2C9 comparison. A) PyMOL alignment of CYP2C19 (PDB: 4GQS) and CYP2C9 (PDB: 1OG2) crystal structures with RMSD = 0.596 Å. B) Heatmap of CYP2C19 shift values across all positions and substituted amino acids. CYP2C19 WT is shown in white with a black dot, and missing data is gray. Substrate recognition regions are shown above the heatmap in orange, and sites that interact with heme are shown with blue diamonds.

Secondary structure of CYP2C19 above the heatmap is represented with α-helices shown in magenta and β-sheets shown in cyan. C) Density distribution of mean shift values in blue and randomized mean shift values in gray, as described in Methods. D) CYP2C19 structure (PDB: 4GQS) colored by position mean shift value. Heme is colored by element (carbon: black, nitrogen: blue, oxygen: red, iron: yellow), and PDB chemical 0XV in green. E) CYP2C19 structure from molecular dynamics simulation(Mustafa et al. 2020) colored by i) position mean shift value or ii) α-helix. G) Scatterplot showing position mean shift value versus the distance between Cα of aligned CYP2C19 (PDB: 4GQS) and CYP2C9 (PDB: 1OG2) structures. Pearson's R = 0.04.

Figure 2.9

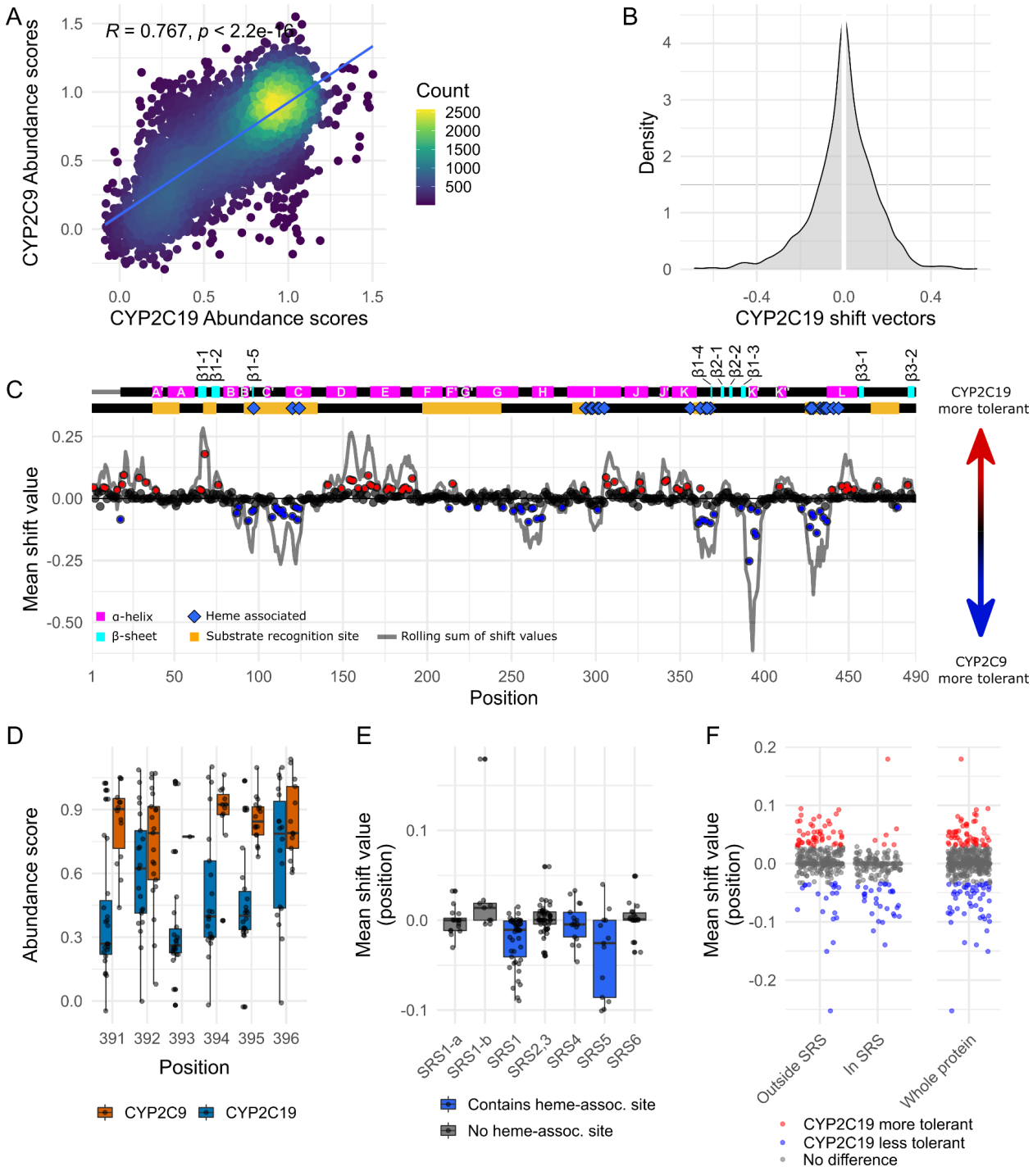


Figure 2.9. Comparison of VAMP-seq mutational tolerance of CYP2C19 and CYP2C9. A) Scatterplot of 5,979 abundance scores present in CYP2C19 and our previous CYP2C9 VAMP-seq experiment (Amorosi et al. 2021). B) Distribution of non-zero shift values calculated

using multiDMS (Haddox et al. 2023). Shift values are shown only for variants present in both datasets. C) Scatterplot of mean shift values for each position. Filled dots represent positions significantly more tolerant in CYP2C19 (red) or more tolerant in CYP2C9 (blue) with FDR-controlled p-values < 0.05 using a randomization test. The rolling sum of the mean shift values is depicted by the gray line. Substrate recognition regions are shown in orange, and sites that interact with heme are shown with blue diamonds. Secondary structure of CYP2C19 represented with α-helices shown in magenta and β-sheets shown in cyan. D) Boxplot of variant abundance scores for CYP2C19 (blue) and CYP2C9 (orange) across positions in the K' helix. Dots represent variant abundance scores. E) Boxplot of shift values in substrate recognition sites (SRSs) with or without a heme-associated site. Dots represent mean shift values at each position within the SRS. F) Dot plot of mean shift values for each position separated by whether 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.

Figure 2.10

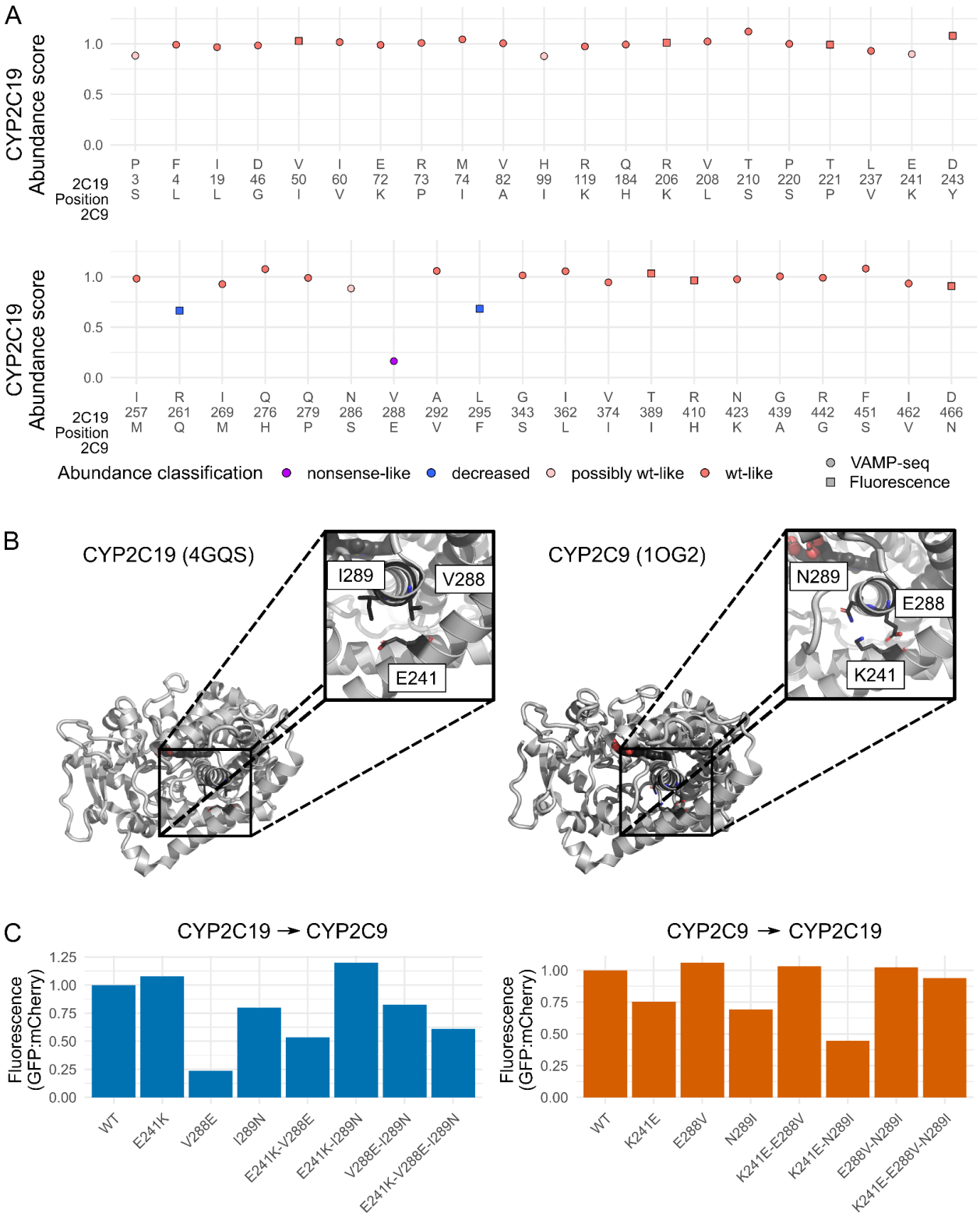

A
CYP2C19 Abundance score
1.0
0.5
0.0
2C19 Position 2C9
P 3 S
F 4 L
I 19 L
D 46 G
V 50 I
I 60 V
E 72 K
R 73 P
M 74 I
V 82 A
H 99 I
R 119 K
Q 184 H
R 206 K
V 208 L
T 210 S
P 220 S
T 221 P
L 237 V
E 241 K
D 243 Y
CYP2C19 Abundance score
1.0
0.5
0.0
2C19 Position 2C9
I 257 M
R 261 Q
I 269 M
Q 276 H
Q 279 P
N 286 S
V 288 E
A 292 V
L 295 F
G 343 S
I 362 L
V 374 I
T 389 I
R 410 H
N 423 K
G 439 A
R 442 G
F 451 S
I 462 V
D 466 N
Abundance classification
nonsense-like
decreased
possibly wt-like
wt-like
VAMP-seq
Fluorescence
B
CYP2C19 (4GQS)
I289
V288
E241
CYP2C9 (1OG2)
N289
E288
K241
C
CYP2C19 → CYP2C9
Fluorescence (GFP:mCherry)
1.25
1.00
0.75
0.50
0.25
0.00
WT
E241K
V288E
I289N
E241K-V288E
E241K-I289N
V288E-I289N
E241K-V288E-I289N
CYP2C9 → CYP2C19
Fluorescence (GFP:mCherry)
1.00
0.75
0.50
0.25
0.00
WT
K241E
E288V
N289I
K241E-E288V
K241E-N289I
E288V-N289I
K241E-E288V-N289I

Figure 2.10. 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 2C9 WT amino acid installed. The WT amino acids for CYP2C19 and 2C9 are shown above and below the position. Dots are colored by 2C19 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. 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 (n = 50,000 cells). Fluorescence of each variant is normalized to its respective WT CYP.

Figure 2.11

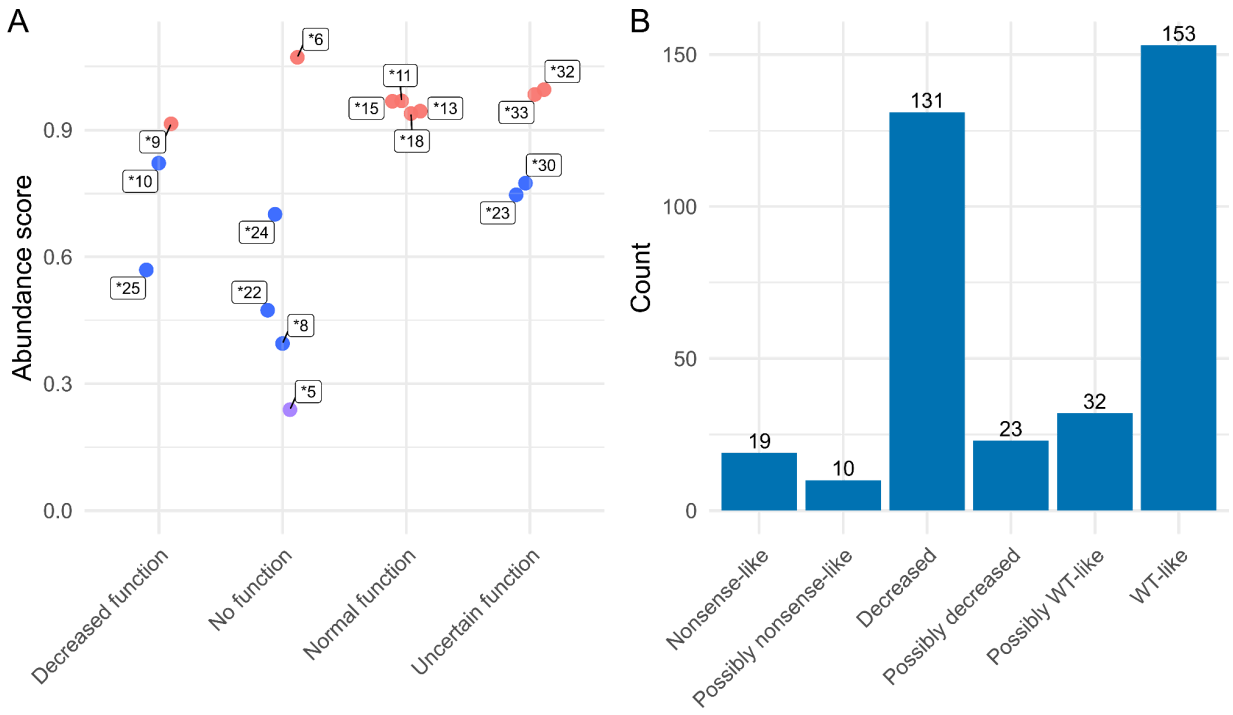


Figure 2.11. CYP2C19 abundance scores for variants found in humans. A) Scatter plot of CYP2C19 abundance scores of star (*) alleles with clinical functional status according to CPIC database (accessed May 18, 2022). Dots are colored by abundance score classification and labeled by their star allele designation. B) Bar plot representing abundance score classification of variants in gnomAD (accessed May 18, 2022).

Figure 2.12

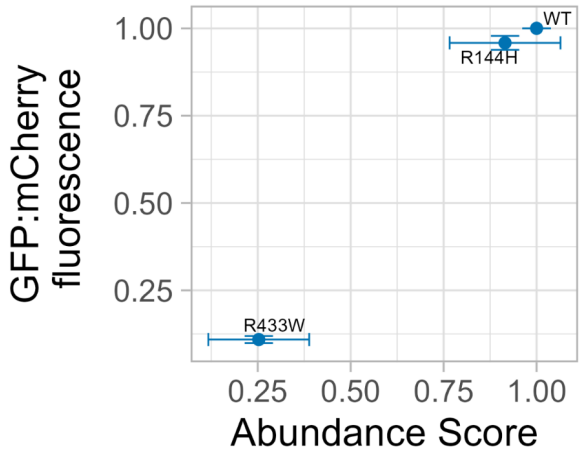


Figure 2.12. Individual variant validation of *9 (R144H). Geometric mean of the fluorescence signal ratio of GFP:mCherry vs abundance score for CYP2C19 WT, *5 (R433W) with known destabilizing properties, and *9 (R144H). Error bars show standard deviation of fluorescence distributions (y-axis) or standard deviation of abundance scores (x-axis).

Figure 3.1

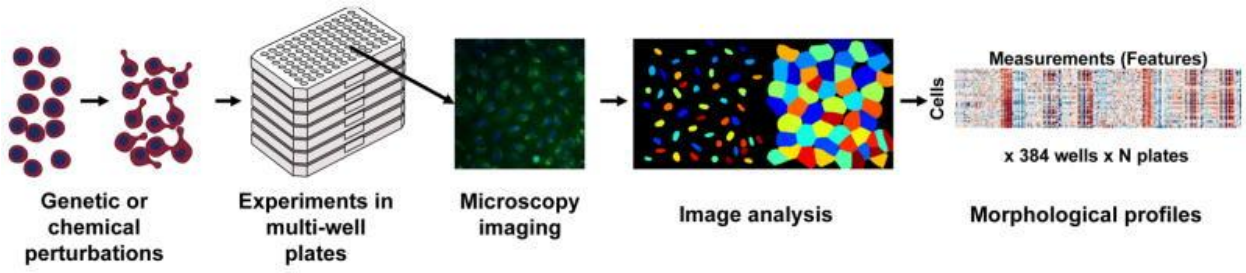


Figure 3.1. Overview of the strategy of morphological profiling using an image-based assay. Cells are seeded into 96- or 384-well plates, treated with chemical compounds across various doses, across different doses, stained with organelle-specific dyes, imaged, and analyzed to identify drug hits that induced changes in visual phenotype. Morphological profiles are analyzed and drug hits are identified for further validation. Adapted from Bray MA, Singh S, Han H, Davis CT, Borgeson B, Hartland C, Kost-Alimova M, Gustafsdottir SM, Gibson CC, Carpenter AE. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016 Sep;11(9):1757-74. doi: 10.1038/nprot.2016.105. Epub 2016 Aug 25. PMID: 27560178; PMCID: PMC5223290. Copyright © 2016

Figure 3.2

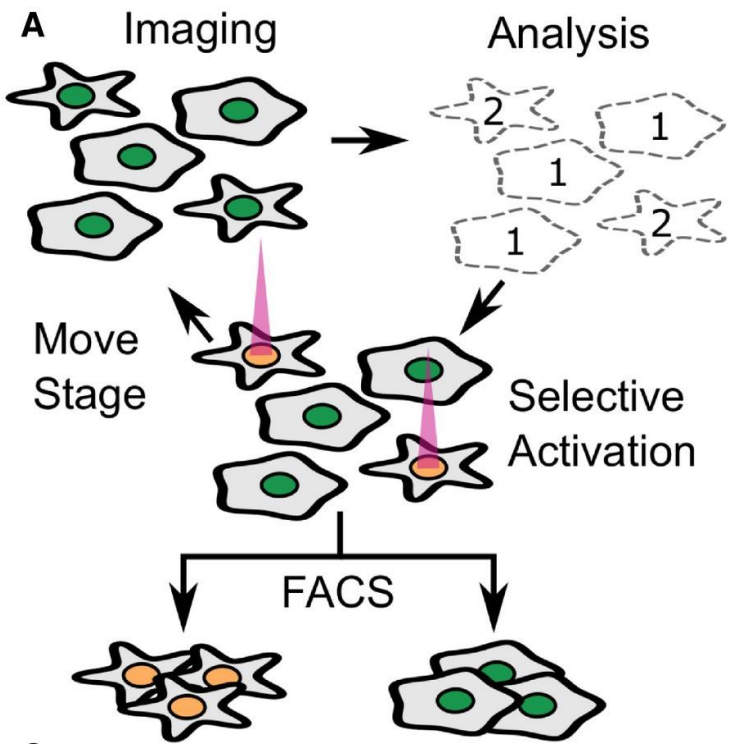


Figure 3.2. Visual Cell Sorting overview. Cells express nuclear-localized Dendra2 protein, which is photoactivated when pulsed with ultraviolet light. Cells are treated with chemical perturbations, then automated microscopy is used to collect images of heterogeneous cell populations. Images are then analyzed by segmenting cells, classifying cells based on visual features, and their nuclei selectively activated using a digital micromirror device directing a pulse of UV light for 200 ms to 1200 ms. After all fields of view are imaged and activated, cells are separated using fluorescence-activated cell sorting based on the ratio of activated to unactivated Dendra2. Adapted from “High-throughput, microscope-based sorting to dissect cellular heterogeneity” by Nicholas Hasle, Anthony Cooke, Sanjay Srivatsan, Heather Huang, Jason J Stephany, Zachary Krieger, Dana Jackson, Weiliang Tang, Sriram Pendyala, Raymond J Monnat Jr., Cole Trapnell, Emily M Hatch, and Douglas M Fowler, 2020. Molecular Systems Biology (2020)16:e9442. Copyright © 2020

Figure 4.1

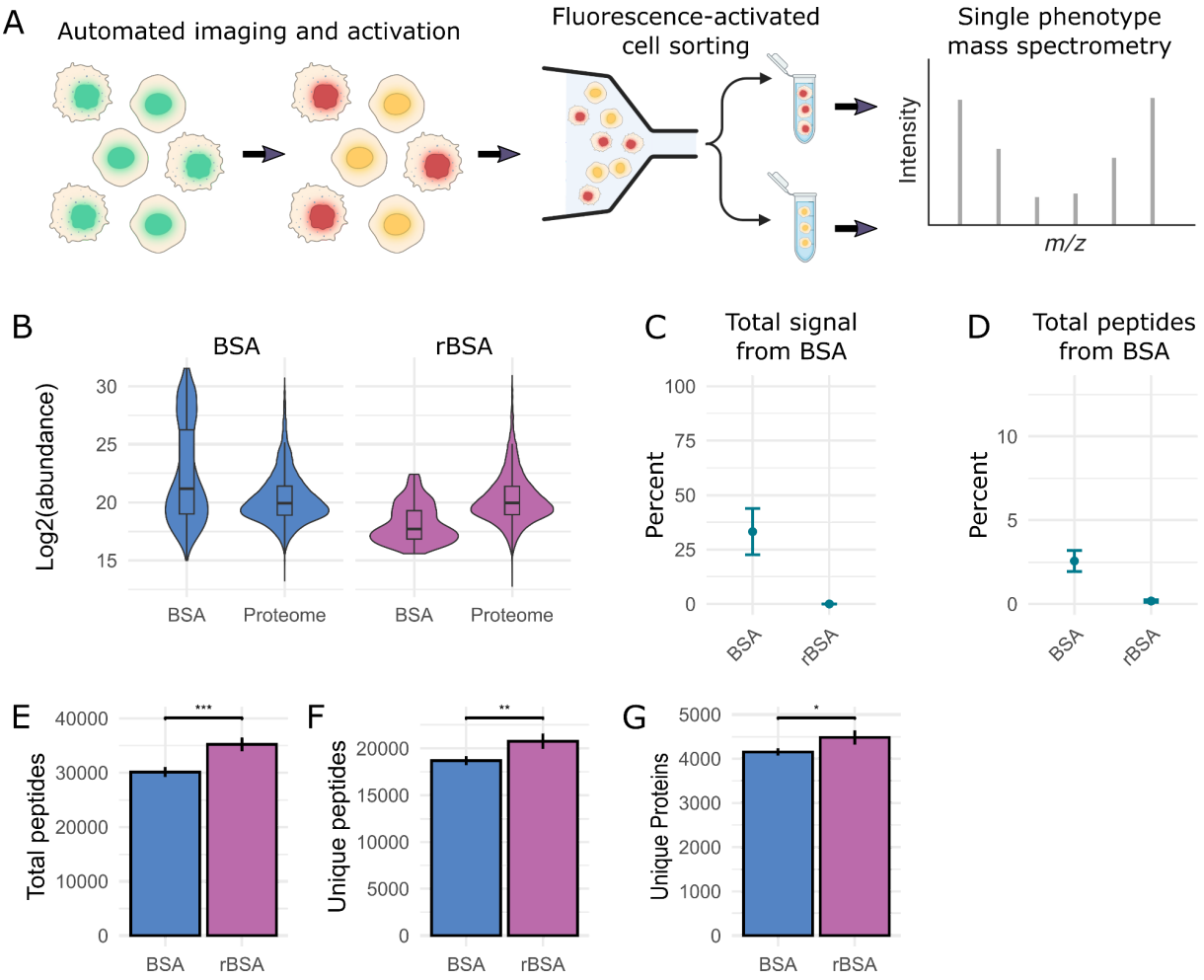


Figure 4.1. Visual Cell Sorting coupled to Mass Spectrometry (VCS-MS).

A) VCS-MS workflow overview. Cells expressing dendra2 are automatically imaged, classified based on visual features, and activated using different duration pulses of 405 nm light. Activated cells are then separated using fluorescence-activated cell sorting, and analyzed using MS. B) Violin plots showing the relative abundance of BSA- or Proteome-associated peptides in samples prepared with BSA or cleavage-resistant rBSA. C) Mean percent total signal from BSA-associated peptides. D) Mean percent of all peptides identified that were derived from BSA in samples prepared with BSA or rBSA. E) Mean total number and F) number of unique, non-BSA peptides, and G) number of unique, non-BSA proteins in samples prepared with BSA

or rBSA. Statistical significance is shown with asterisks. Two-sided Student *t*-test *P*-values indicate statistical significance (* $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$). All experiments were completed with n = 4 replicates, with error bars representing standard deviation.

Figure 4.2

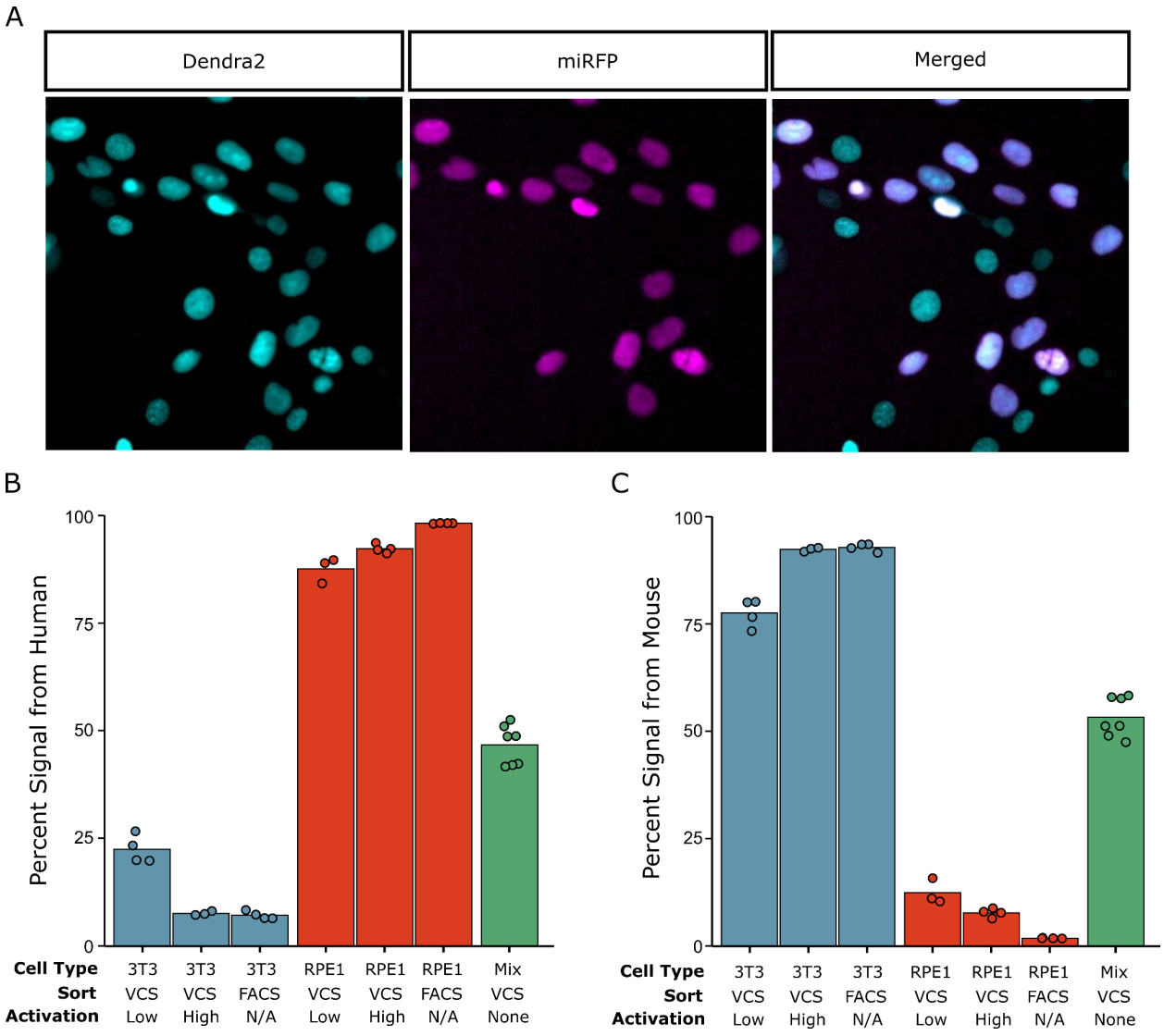


Figure 4.2. Measuring the proteomes of visually distinct cell populations.

Co-cultured human (RPE1) and mouse (3T3) cells were separated using VCS or FACS. A) Representative images showing Dendra2 fluorescence in cyan, miRFP fluorescence in magenta, and merged channels. All cells express Dendra2, and only RPE1 cells express miRFP. B-C) Percent total signal from human- and mouse-specific peptides, respectively. Cell type represents the target population for separation. Cell types were separated using either VCS or FACS where 3T3 cells are Dendra2+/miRFP- and RPE1 cells are Dendra2+/miRFP+.

For samples separated with VCS, low activation samples were pulsed with 300 ms of 405 nm light and high activation samples were pulsed with 1200 ms. “Mix” cells were collected from imaging wells that were not activated.

Figure 4.3

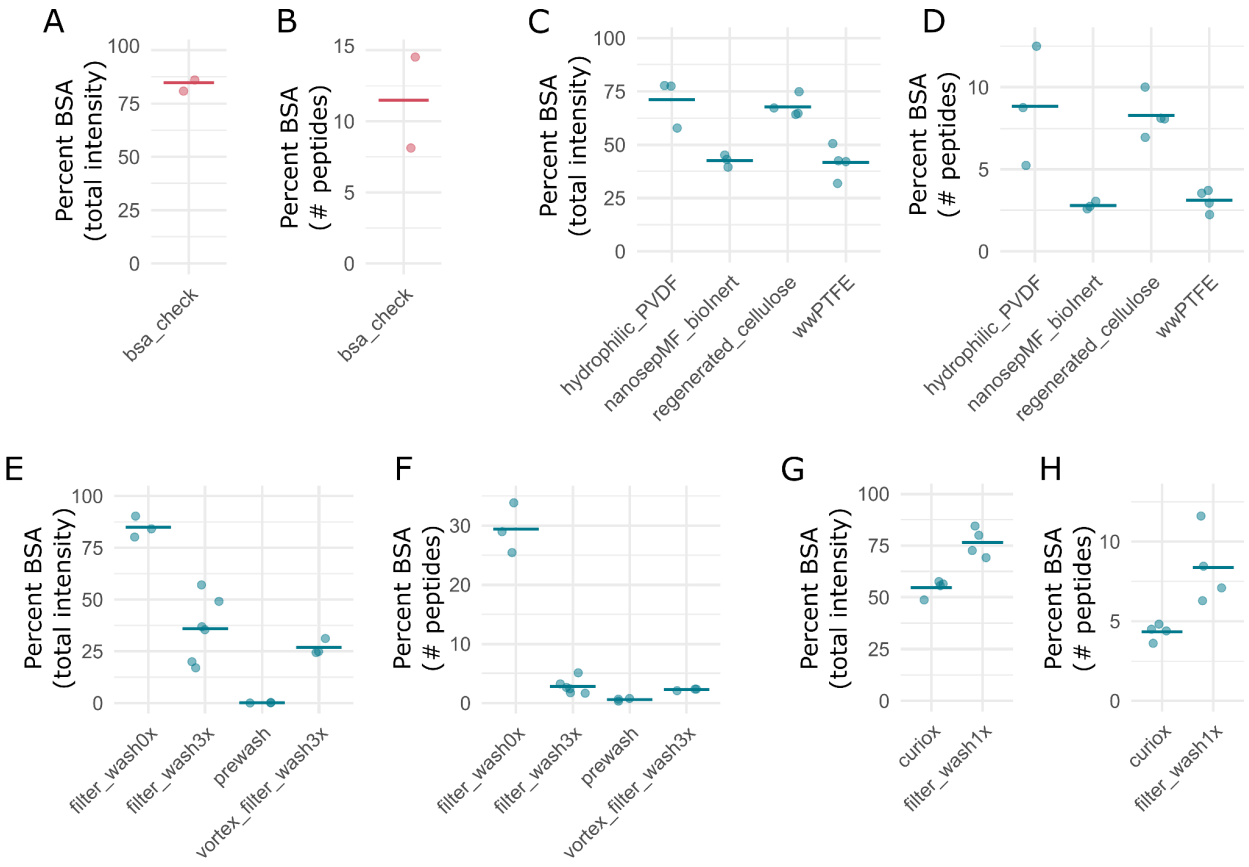


Figure 4.3. BSA content across wash conditions. Each scatter plot shows either percent total intensity from BSA-associated peptides or the percent of the total number of peptide identifications associated with BSA. Each dot represents a sample, and horizontal lines show the mean of the samples. Red dots are from samples collected using FACS. Blue dots are from samples where cells were counted by hemocytometer and collected by pipetting. A-B) Samples were washed once with PBS by pelleting and aspirating, then FACS was used to sort replicates (n=2) containing ~40,000 cells into 0.5% BSA. C-D) Cells were diluted to 50,000 cells per 500 µl in 0.5% BSA, and 500 µl was added to the filter chambers of tubes with hydrophilic PVDF (Millipore, UFC30GV0S), BioInert (Pall, ODM02C34), regenerated cellulose (Thomas Scientific, 1194R90), and wwPTFE (Pall, ODPTFE02C34) membranes. Each sample was then centrifuged at 300-500 x *g* for ~1 minute to flow the supernatant through, leaving ~50 µl. Then, 500 µl PBS

was added, and the procedure was repeated for three total washes. E-F) Samples labeled “prewash” were collected from stock cells washed 3x with PBS by pelleting and aspirating the supernatant. Then, cells were counted by hemocytometer and 50,000 cells were aliquoted to each sample. The remaining samples were collected as in C-D into BioInert membrane tubes and washed as follows: “filter_wash0x” were centrifuged to remove the 0.5% BSA resuspension solution, then lysed and collected; “filter_wash3x” were washed 3 additional times with PBS, leaving ~50 µl supernatant behind; “vortex_filter_wash3x” were washed 3 times with PBS as above, briefly vortexing between washes to resuspend cells. G-H) Cells were counted using a hemocytometer and 25,000 cells were collected in 50 µl aliquots. “Curiox” samples were processed using Laminar Wash™ MINI System (Curiox) with 10 full buffer exchanges, following manufacturer’s instructions. “Filter_wash1x” samples were processed as described in E-F.

Figure 4.4

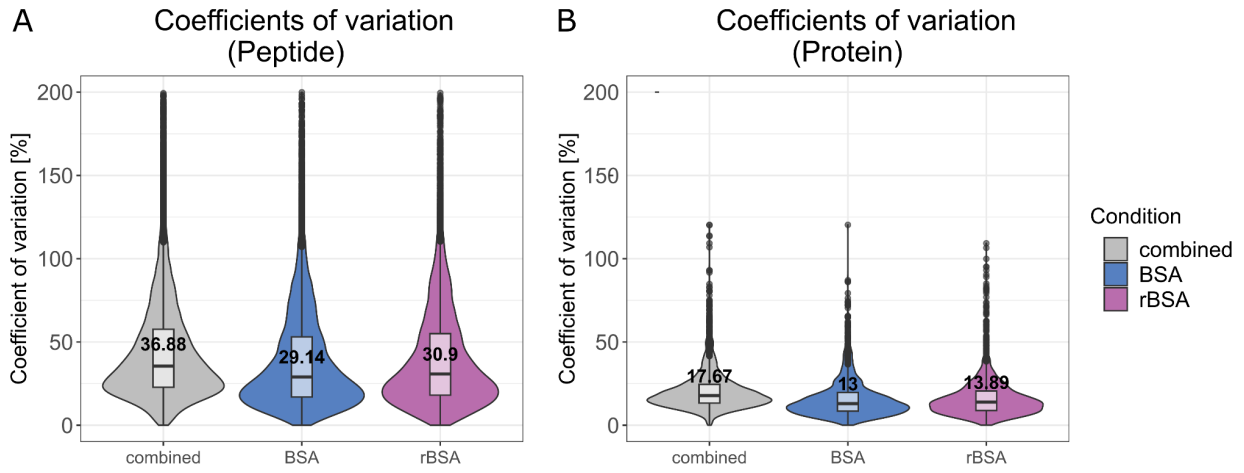


Figure 4.4. Violin plots showing coefficients of variation (CVs) for samples prepared with BSA or cleavage resistant rBSA. Numbers show median CV. A) CVs for peptide-level quantification. B) CVs for protein-level quantification.

Figure 4.5

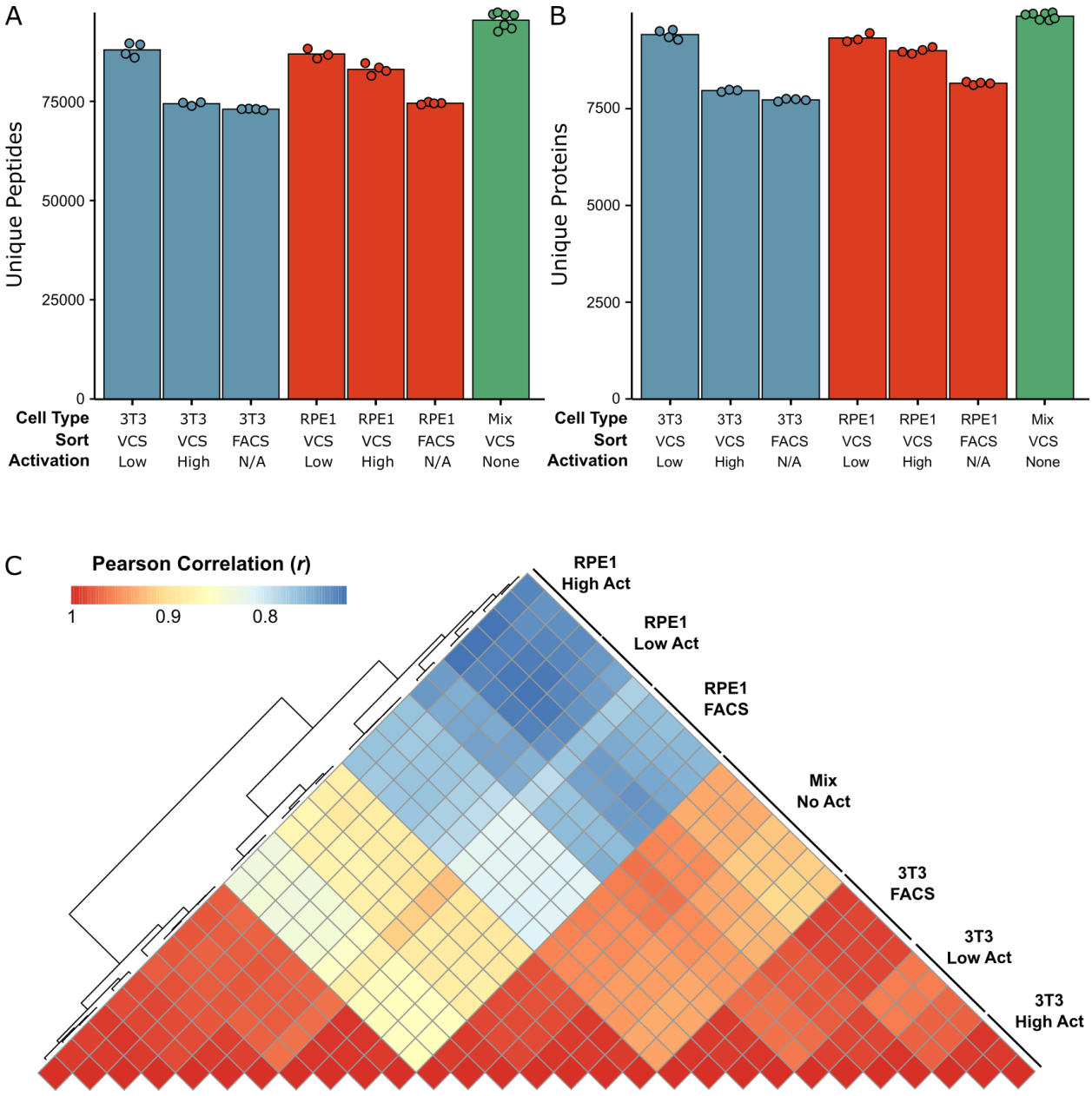


Figure 4.5. Co-cultured human (RPE1) and mouse (3T3) cells separated using VCS or FACS. A) Number of unique peptides and B) unique proteins from human- and mouse-specific peptides. Proteins represented by fewer than 3 unique peptides were removed from the analysis. Cell Type represents the target population for separation. Cell Types were selected using either VCS or FACS where 3T3 cells are dendra2+/miRFP- and RPE1 cells are

dendra2+/miRFP+. For samples separated with VCS, Low Activation samples were pulsed with 300 ms of 405 nm light and High Activation samples were pulsed with 1200 ms. “Mix” cells were collected from imaging wells that were not activated. C) Heatmap showing pairwise peptide abundance Pearson correlations for RPE1 High Act (VCS, n=4), RPE1 Low Act (VCS, n=3), RPE1 (FACS, n=4), Mix No Act (VCS, n=7), 3T3 Low Act (VCS, n=4), 3T3 High Act (VCS, n=3), and 3T3 (FACS, n=4) samples. Sample positions were determined using hierarchical clustering.

References

Almendro, Vanessa, Andriy Marusyk, and Kornelia Polyak. 2013. “Cellular Heterogeneity and Molecular Evolution in Cancer.” *Annual Review of Pathology* 8 (January): 277–302.

Altarsha, Muhannad, Tobias Benighaus, Devesh Kumar, and Walter Thiel. 2009. “How Is the Reactivity of Cytochrome P450cam Affected by Thr252X Mutation? A QM/MM Study for X = Serine, Valine, Alanine, Glycine.” *Journal of the American Chemical Society* 131 (13): 4755–63.

Altschuler, Steven J., and Lani F. Wu. 2010. “Cellular Heterogeneity: Do Differences Make a Difference?” *Cell*. https://doi.org/10.1016/j.cell.2010.04.033.

Amorosi, Clara J., Melissa A. Chiasson, Matthew G. McDonald, Lai Hong Wong, Katherine A. Sitko, Gabriel Boyle, John P. Kowalski, Allan E. Rettie, Douglas M. Fowler, and Maitreya J. Dunham. 2021. “Massively Parallel Characterization of CYP2C9 Variant Enzyme Activity and Abundance.” *American Journal of Human Genetics* 108 (9): 1735–51.

Araya, Carlos L., and Douglas M. Fowler. 2011. “Deep Mutational Scanning: Assessing Protein Function on a Massive Scale.” *Trends in Biotechnology* 29 (9): 435–42.

Attia, Tamer Zekry, Taku Yamashita, Mohamed Abdelkhalek Hammad, Akinori Hayasaki, Takumi Sato, Masayoshi Miyamoto, Yuki Yasuhara, et al. 2014. “Effect of Cytochrome P450 2C19 and 2C9 Amino Acid Residues 72 and 241 on Metabolism of Tricyclic Antidepressant Drugs.” *Chemical & Pharmaceutical Bulletin* 62 (2): 176–81.

Baylon, Javier L., Ivan L. Lenov, Stephen G. Sligar, and Emad Tajkhorshid. 2013. “Characterizing the Membrane-Bound State of Cytochrome P450 3A4: Structure, Depth of Insertion, and Orientation.” *Journal of the American Chemical Society* 135 (23): 8542–51.

Beitelshees, Amber L., Cameron D. Thomas, Philip E. Empey, George A. Stouffer, Dominick J. Angiolillo, Francesco Franchi, Sony Tuteja, et al. 2022. “CYP2C19 Genotype-Guided Antiplatelet Therapy After Percutaneous Coronary Intervention in Diverse Clinical Settings.”

Journal of the American Heart Association 11 (4): e024159.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Berka, Karel, Tereza Hendrychová, Pavel Anzenbacher, and Michal Otyepka. 2011. “Membrane Position of Ibuprofen Agrees with Suggested Access Path Entrance to Cytochrome P450 2C9 Active Site.” *The Journal of Physical Chemistry. A* 115 (41): 11248–55.

Bernhardt, Rita. 2006. “Cytochromes P450 as Versatile Biocatalysts” 124: 128–45.

Blaisdell, Joyce, Harvey Mohrenweiser, Jonathan Jackson, Stephen Ferguson, Sherry Coulter, Brian Chanas, Tina Xi, Burhan Ghanayem, and Joyce A. Goldstein. 2002. “Identification and Functional Characterization of New Potentially Defective Alleles of Human CYP2C19.” *Pharmacogenetics* 12 (9): 703–11.

Bray, Mark-Anthony, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. 2016. “Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes.” *Nature Protocols* 11 (9): 1757–74.

Buchberger, Amanda Rae, Kellen DeLaney, Jillian Johnson, and Lingjun Li. 2018. “Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights.” *Analytical Chemistry* 90 (1): 240–65.

Cagiada, Matteo, Kristoffer E. Johansson, Audrone Valanciute, Sofie V. Nielsen, Rasmus Hartmann-Petersen, Jun J. Yang, Douglas M. Fowler, Amelie Stein, and Kresten Lindorff-Larsen. 2021. “Understanding the Origins of Loss of Protein Function by Analyzing the Effects of Thousands of Variants on Activity and Abundance.” *Molecular Biology and Evolution*, March. https://doi.org/10.1093/molbev/msab095.

Caicedo, Juan C., Shantanu Singh, and Anne E. Carpenter. 2016. “Applications in Image-Based Profiling of Perturbations.” *Current Opinion in Biotechnology* 39 (June): 134–42.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. “Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells.” *Science* 361 (6409): 1380–85.

Cao, Junyue, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, et al. 2017. “Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism.” *Science* 357 (6352): 661–67.

Carlsten, Chris, Michael Brauer, Fiona Brinkman, Jeffrey Brook, Denise Daley, Kelly McNagny, Mandy Pui, Diana Royce, Tim Takaro, and Judah Denburg. 2014. “Genes, the Environment and Personalized Medicine: We Need to Harness Both Environmental and Genetic Data to Maximize Personal and Population Health.” *EMBO Reports* 15 (7): 736–39.

Chattopadhyay, P. K., and M. Roederer. 2015. “A Mine Is a Terrible Thing to Waste: High Content, Single Cell Technologies for Comprehensive Immune Analysis.” *American Journal of Transplantation*. Blackwell Publishing Ltd. https://doi.org/10.1111/ajt.13193.

Cheng, Jun, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, et al. 2023. “Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense.” *Science* 381 (6664): eadg7492.

Chiasson, Melissa, Maitreya J. Dunham, Allan E. Rettie, and Douglas M. Fowler. 2019. “Applying Multiplex Assays to Understand Variation in Pharmacogenes” 0 (0): 1–5.

Cimini, Beth A., Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, et al. 2023. “Optimizing the Cell Painting Assay for Image-Based Profiling.” *Nature Protocols*, June. https://doi.org/10.1038/s41596-023-00840-9.

Cohen, A. A., N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, et al. 2008. “Dynamic Proteomics of Individual Cancer Cells in Response to a Drug.” Vol. 322.

Cojocaru, Vlad, Peter J. Winn, and Rebecca C. Wade. 2007. “The Ins and Outs of Cytochrome

P450s.” *Biochimica et Biophysica Acta* 1770 (3): 390–401.

Collins, Francis S., and Harold Varmus. 2015. “A New Initiative on Precision Medicine.” *The New England Journal of Medicine* 372 (9): 793–95.

Coon, Minor J. 2005. “Cytochrome P450: Nature’s Most Versatile Biological Catalyst.” *Annual Review of Pharmacology and Toxicology* 45: 1–25.

Cuomo, Anna S. E., Aparna Nathan, Soumya Raychaudhuri, Daniel G. MacArthur, and Joseph E. Powell. 2023. “Single-Cell Genomics Meets Human Genetics.” *Nature Reviews. Genetics*, April. https://doi.org/10.1038/s41576-023-00599-5.

Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. “Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing.” *Science* 348 (6237): 910–14.

Dai, Da-Peng, Li-Ming Hu, Pei-Wu Geng, Shuang-Hu Wang, Jie Cai, Guo-Xin Hu, and Jian-Ping Cai. 2015. “In Vitro Functional Analysis of 24 Novel CYP2C19 Variants Recently Found in the Chinese Han Population.” *Xenobiotica; the Fate of Foreign Compounds in Biological Systems* 45 (11): 1030–35.

Davidoff, F., K. Case, and P. W. Fried. 1995. “Evidence-Based Medicine: Why All the Fuss?” *Annals of Internal Medicine* 122 (9): 727.

Dean, Laura, and Megan Kane. 2022. *Clopidogrel Therapy and CYP2C19 Genotype*. National Center for Biotechnology Information (US).

Demichev, Vadim, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and Markus Ralser. 2020. “DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput.” *Nature Methods* 17 (1): 41–44.

Denisov, Ilia G., Thomas M. Makris, Stephen G. Sligar, and Ilme Schlichting. 2005. “Structure and Chemistry of Cytochrome P450.” *Chemical Reviews* 105 (6): 2253–77.

DePristo, Mark A., Daniel M. Weinreich, and Daniel L. Hartl. 2005. “Missense Meanderings in

Sequence Space: A Biophysical View of Protein Evolution." *Nature Reviews. Genetics* 6 (9): 678–87.

Derayea, Sayed M., Hirofumi Tsujino, Yukiko Oyama, Yoshinobu Ishikawa, Taku Yamashita, and Tadayuki Uno. 2020. "Impact of Single Nucleotide Polymorphisms (R132Q and W120R) on the Binding Affinity and Metabolic Activity of CYP2C19 toward Some Therapeutically Important Substrates." *Xenobiotica; the Fate of Foreign Compounds in Biological Systems* 50 (12): 1510–19.

Dewez, Frédéric, Janina Oejten, Corinna Henkel, Romano Hebeler, Heiko Neuweger, Edwin De Pauw, Ron M. A. Heeren, and Benjamin Balluff. 2020. "MS Imaging-Guided Microproteomics for Spatial Omics on a Single Instrument." *Proteomics*, August, e1900369.

Dijk, Erwin L. van, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. "Ten Years of next-Generation Sequencing Technology." *Trends in Genetics: TIG* 30 (9): 418–26.

Dilmetz, Brooke A., Yea-Rin Lee, Mark R. Condina, Matthew Briggs, Clifford Young, Christopher T. Desire, Manuela Klingler-Hoffmann, and Peter Hoffmann. 2021. "Novel Technical Developments in Mass Spectrometry Imaging in 2020: A Mini Review." *Analytical Science Advances* 2 (3-4): 225–37.

Doherty, Mary K., Dean E. Hammond, Michael J. Clague, Simon J. Gaskell, and Robert J. Beynon. 2009. "Turnover of the Human Proteome: Determination of Protein Intracellular Stability by Dynamic SILAC." *Journal of Proteome Research* 8 (1): 104–12.

Dorner, Mariah E., Ryan D. McMunn, Thomas G. Bartholow, Brecken E. Calhoon, Michelle R. Conlon, Jessica M. Dulli, Samuel C. Fehling, et al. 2015. "Comparison of Intrinsic Dynamics of Cytochrome p450 Proteins Using Normal Mode Analysis." *Protein Science: A Publication of the Protein Society* 24 (9): 1495–1507.

Doud, Michael B., Orr Ashenberg, and Jesse D. Bloom. 2015. "Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs." *Molecular Biology and Evolution* 32 (11): 2944–60.

Elsasser, W. M. 1984. “Outline of a Theory of Cellular Heterogeneity.” *Proceedings of the National Academy of Sciences* 81 (16): 5126–29.

Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. 2013. “Comet: An Open-Source MS/MS Sequence Database Search Tool.” *Proteomics* 13 (1): 22–24.

Esteves, Francisco, José Rueff, and Michel Kranendonk. 2021. “The Central Role of Cytochrome P450 in Xenobiotic Metabolism-A Brief Review on a Fascinating Enzyme Family.” *Journal of Xenobiotics* 11 (3): 94–114.

Evidence-Based Medicine Working Group. 1992. “Evidence-Based Medicine. A New Approach to Teaching the Practice of Medicine.” *JAMA: The Journal of the American Medical Association* 268 (17): 2420–25.

Faber, Matthew S., Emily E. Wrenbeck, Laura R. Azouz, Paul J. Steiner, and Timothy A. Whitehead. 2019. “Impact of In Vivo Protein Folding Probability on Local Fitness Landscapes.” *Molecular Biology and Evolution* 36 (12): 2764–77.

Fayer, Shawn, Carrie Horton, Jennifer N. Dines, Alan F. Rubin, Marcy E. Richardson, Kelly McGoldrick, Felicia Hernandez, et al. 2021. “Closing the Gap: Systematic Integration of Multiplexed Functional Data Resolves Variants of Uncertain Significance in BRCA1, TP53, and PTEN.” *American Journal of Human Genetics* 108 (12): 2248–58.

Fischer, Markus, Michael Knoll, Demet Sirim, Florian Wagner, Sonja Funke, and Juergen Pleiss. 2007. “The Cytochrome P450 Engineering Database: A Navigation and Prediction Tool for the Cytochrome P450 Protein Family.” *Bioinformatics* 23 (15): 2015–17.

Fisher, R., L. Pusztai, and C. Swanton. 2013. “Cancer Heterogeneity: Implications for Targeted Therapeutics.” *British Journal of Cancer* 108 (3): 479–85.

Foti, Robert S., Dan A. Rock, Xiaogang Han, Robert A. Flowers, Larry C. Wienkers, and Jan L. Wahlstrom. 2012. “Ligand-Based Design of a Potent and Selective Inhibitor of Cytochrome P450 2C19.” *Journal of Medicinal Chemistry* 55 (3): 1205–14.

Fowler, Douglas M., and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of

Protein Science." *Nature Methods* 11 (8): 801–7.

Galli, Mattia, Stefano Benenati, Davide Capodanno, Francesco Franchi, Fabiana Rollini, Domenico D'Amario, Italo Porto, and Dominick J. Angiolillo. 2021. "Guided versus Standard Antiplatelet Therapy in Patients Undergoing Percutaneous Coronary Intervention: A Systematic Review and Meta-Analysis." *The Lancet* 397 (10283): 1470–83.

García-Nafría, Javier, Jake F. Watson, and Ingo H. Greger. 2016. "IVA Cloning: A Single-Tube Universal Cloning System Exploiting Bacterial In Vivo Assembly." *Scientific Reports* 6 (June): 27459.

Geck, Renee C., Gabriel Boyle, Clara J. Amorosi, Douglas M. Fowler, and Maitreya J. Dunham. 2022. "Measuring Pharmacogene Variant Function at Scale Using Multiplexed Assays." *Annual Review of Pharmacology and Toxicology* 62 (January): 531–50.

Goetz, Laura H., and Nicholas J. Schork. 2018. "Personalized Medicine: Motivation, Challenges, and Progress." *Fertility and Sterility* 109 (6): 952–63.

Goldstein, J. A., and S. M. de Morais. 1994. "Biochemistry and Molecular Biology of the Human CYP2C Subfamily." *Pharmacogenetics* 4 (6): 285–99.

Gotoh, O. 1992. "Substrate Recognition Sites in Cytochrome P450 Family 2 (CYP2) Proteins Inferred from Comparative Analyses of Amino Acid and Coding Nucleotide Sequences." *The Journal of Biological Chemistry* 267 (1): 83–90.

Goulding, Rebecca, Diana Dawes, Morgan Price, Sabrina Wilkie, and Martin Dawes. 2015. "Genotype-Guided Drug Prescribing: A Systematic Review and Meta-Analysis of Randomized Control Trials." *British Journal of Clinical Pharmacology* 80 (4): 868–77.

Gricman, Łukasz, Constantin Vogel, and Jürgen Pleiss. 2014. "Conservation Analysis of Class-Specific Positions in Cytochrome P450 Monooxygenases: Functional and Structural Relevance." *Proteins* 82 (3): 491–504.

———. 2015. "Identification of Universal Selectivity-Determining Positions in Cytochrome P450 Monooxygenases by Systematic Sequence-Based Literature Mining." *Proteins* 83 (9):

1593–1603.

Gumulya, Yosephin, Jong-Min Baek, Shun-Jie Wun, Raine E. S. Thomson, Kurt L. Harris, Dominic J. B. Hunter, James B. Y. H. Behrendorff, et al. 2018. “Engineering Highly Functional Thermostable Proteins Using Ancestral Sequence Reconstruction.” *Nature Catalysis* 1 (11): 878–88.

Haddox, Hugh K., Adam S. Dingens, Sarah K. Hilton, Julie Overbaugh, and Jesse D. Bloom. 2018. “Mapping Mutational Effects along the Evolutionary Landscape of HIV Envelope.” *eLife* 7 (March). https://doi.org/10.7554/eLife.34420.

Haddox, Hugh K., Jared G. Galloway, Bernadeta Dadonaite, Jesse D. Bloom, Frederick A. Matsen, and William S. DeWitt. 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.

Haines, D. C., D. R. Tomchick, M. Machius, and J. A. Peterson. 2001. “Pivotal Role of Water in the Mechanism of P450BM-3.” *Biochemistry* 40 (45): 13456–65.

Hasemann, C. A., R. G. Kurumbail, S. S. Boddupalli, J. A. Peterson, and J. Deisenhofer. 1995. “Structure and Function of Cytochromes P450: A Comparative Analysis of Three Crystal Structures.” *Structure* 3 (1): 41–62.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. 2017. “Multi-Omics Approaches to Disease.” *Genome Biology* 18 (1): 83.

Hasle, Nicholas, Anthony Cooke, Sanjay Srivatsan, Heather Huang, Jason J. Stephany, Zachary Krieger, Dana Jackson, et al. 2020. “High-Throughput, Microscope-Based Sorting to Dissect Cellular Heterogeneity.” *Molecular Systems Biology* 16 (6): e9442.

Huang, Weiliang, Wayne A. Johnston, Martin A. Hayes, James J. De Voss, and Elizabeth M. J. Gillam. 2007. “A Shuffled CYP2C Library with a High Degree of Structural Integrity and Functional Versatility.” *Archives of Biochemistry and Biophysics* 467 (2): 193–205.

Hughes, Christopher S., Sophie Moggridge, Torsten Müller, Poul H. Sorensen, Gregg B. Morin,

and Jeroen Krijgsveld. 2019. “Single-Pot, Solid-Phase-Enhanced Sample Preparation for Proteomics Experiments.” *Nature Protocols* 14 (1): 68–85.

Hughes, Christopher S., Poul H. Sorensen, and Gregg B. Morin. 2019. “A Standardized and Reproducible Proteomics Protocol for Bottom-Up Quantitative Analysis of Protein Samples Using SP3 and Mass Spectrometry.” *Methods in Molecular Biology* 1959: 65–87.

Ibeanu, G. C., B. I. Ghanayem, P. Linko, L. Li, L. G. Pederson, and J. A. Goldstein. 1996. “Identification of Residues 99, 220, and 221 of Human Cytochrome P450 2C19 as Key Determinants of Omeprazole Activity.” *The Journal of Biological Chemistry* 271 (21): 12496–501.

Ibeanu, G. C., J. A. Goldstein, U. Meyer, S. Benhamou, C. Bouchardy, P. Dayer, B. I. Ghanayem, and J. Blaisdell. 1998. “Identification of New Human CYP2C19 Alleles (CYP2C19*6 and CYP2C19*2B) in a Caucasian Poor Metabolizer of Mephenytoin.” *The Journal of Pharmacology and Experimental Therapeutics* 286 (3): 1490–95.

Ionescu, Corina, and Mino R. Caira, eds. 2005. “Pathways of Biotransformation — Phase I Reactions.” In *Drug Metabolism: Current Concepts*, 41–128. Dordrecht: Springer Netherlands.

Ionova, Yelena, James Ashenhurst, Jianan Zhan, Hoang Nhan, Cindy Kosinski, Bani Tamraz, and Alison Chubb. 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.” *Clinical and Translational Science* 13 (6): 1298–1306.

Jain, Pankaj C., and Raghavan Varadarajan. 2014. “A Rapid, Efficient, and Economical Inverse Polymerase Chain Reaction-Based Method for Generating a Site Saturation Mutant Library.” *Analytical Biochemistry* 449 (March): 90–98.

Jones, Eric M., Nathan B. Lubock, A. J. Venkatakrishnan, Jeffrey Wang, Alex M. Tseng, Joseph M. Paggi, Naomi R. Latorraca, et al. 2020. “Structural and Functional Characterization of G

Protein-Coupled Receptors with Deep Mutational Scanning." *eLife* 9 (October). https://doi.org/10.7554/eLife.54895.

Jung, F., K. J. Griffin, W. Song, T. H. Richardson, M. Yang, and E. F. Johnson. 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): 16270–79.

Käll, Lukas, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. 2007. "Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets." *Nature Methods* 4 (11): 923–25.

Kircher, Martin, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert J. A. Bell, Joseph F. Costello, Jay Shendure, and Nadav Ahituv. 2019. "Saturation Mutagenesis of Twenty Disease-Associated Regulatory Elements at Single Base-Pair Resolution." *Nature Communications* 10 (1): 3583.

Klein, Melissa D., Craig R. Lee, and George A. Stouffer. 2018. "Clinical Outcomes of CYP2C19 Genotype-Guided Antiplatelet Therapy: Existing Evidence and Future Directions." *Pharmacogenomics* 19 (13): 1039–46.

Klose, T. S., G. C. Ibeanu, B. I. Ghanayem, L. G. Pedersen, L. Li, S. D. Hall, and J. A. Goldstein. 1998. "Identification of Residues 286 and 289 as Critical for Conferring Substrate Specificity of Human CYP2C9 for Diclofenac and Ibuprofen." *Archives of Biochemistry and Biophysics* 357 (2): 240–48.

Kong, Andy T., Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. 2017. "MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics." *Nature Methods* 14 (5): 513–20.

Kravitz, Richard L., Naihua Duan, and Joel Braslow. 2004. "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages." *The Milbank Quarterly* 82 (4): 661–87.

Krenc, Dawid, and Kesara Na-Bangchang. 2022. “Spectroscopic Observations of β-Eudesmol Binding to Human Cytochrome P450 Isoforms 3A4 and 1A2, but Not to Isoforms 2C9, 2C19, and 2D6.” *Xenobiotica; the Fate of Foreign Compounds in Biological Systems* 52 (2): 199–208.

Lazarou, J., B. H. Pomeranz, and P. N. Corey. 1998. “Incidence of Adverse Drug Reactions in Hospitalized Patients: A Meta-Analysis of Prospective Studies.” *JAMA: The Journal of the American Medical Association* 279 (15): 1200–1205.

Lee, Craig R., Jasmine A. Luzum, Katrin Sangkuhl, Roseann S. Gammal, Marc S. Sabatine, Charles Michael Stein, David F. Kisor, et al. 2022. “Clinical Pharmacogenetics Implementation Consortium Guideline for CYP2C19 Genotype and Clopidogrel Therapy: 2022 Update.” *Clinical Pharmacology and Therapeutics*, January. https://doi.org/10.1002/cpt.2526.

Lertkiatmongkol, Panida, Anunchai Assawamakin, George White, Gaurav Chopra, Pornpimol Rongnoparut, Ram Samudrala, and Sissades Tongsima. 2013. “Distal Effect of Amino Acid Substitutions in CYP2C9 Polymorphic Variants Causes Differences in Interatomic Interactions against (S)-Warfarin.” *PloS One* 8 (9): e74053.

Leutert, Mario, Ricard A. Rodríguez-Mias, Noelle K. Fukuda, and Judit Villén. 2019. “R2-P2 Rapid-Robotic Phosphoproteomics Enables Multidimensional Cell Signaling Studies.” *Molecular Systems Biology* 15 (12): e9021.

Lewis, D. F., M. Dickins, R. J. Weaver, P. J. Eddershaw, P. S. Goldfarb, and M. H. Tarbit. 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; the Fate of Foreign Compounds in Biological Systems* 28 (3): 235–68.

Lin, Sean, Kenji Schorpp, Ina Rothenaigner, and Kamyar Hadian. 2020. “Image-Based High-Content Screening in Drug Discovery.” *Drug Discovery Today* 25 (8): 1348–61.

Lobingier, Braden T., Ruth Hüttenhain, Kelsie Eichel, Kenneth B. Miller, Alice Y. Ting, Mark von Zastrow, and Nevan J. Krogan. 2017. “An Approach to Spatiotemporally Resolve Protein Interaction Networks in Living Cells.” *Cell* 169 (2): 350–60.e12.

Longford, N. T. 1999. “Selection Bias and Treatment Heterogeneity in Clinical Trials.” *Statistics in Medicine* 18 (12): 1467–74.

Mateus, André, Nils Kurzawa, Jessica Perrin, Giovanna Bergamini, and Mikhail M. Savitski. 2022. “Drug Target Identification in Tissues by Thermal Proteome Profiling.” *Annual Review of Pharmacology and Toxicology* 62 (January): 465–82.

Matreyek, Kenneth A., Lea M. Starita, Jason J. Stephany, Beth Martin, Melissa A. Chiasson, Vanessa E. Gray, Martin Kircher, et al. 2018. “Multiplex Assessment of Protein Variant Abundance by Massively Parallel Sequencing.” *Nature Genetics* 50 (6): 874–82.

Matreyek, Kenneth A., Jason J. Stephany, Melissa A. Chiasson, Nicholas Hasle, and Douglas M. Fowler. 2020. “An Improved Platform for Functional Assessment of Large Protein Libraries in Mammalian Cells.” *Nucleic Acids Research* 48 (1): e1.

Matreyek, Kenneth A., Jason J. Stephany, and Douglas M. Fowler. 2017. “A Platform for Functional Assessment of Large Variant Libraries in Mammalian Cells.” *Nucleic Acids Research* 45 (11): e102.

Mega, Jessica L., Willibald Hochholzer, Andrew L. Frelinger 3rd, Michael J. Kluk, Dominick J. Angiolillo, Dean J. Kereiakes, Steven Isserman, et al. 2011. “Dosing Clopidogrel Based on CYP2C19 Genotype and the Effect on Platelet Reactivity in Patients with Stable Cardiovascular Disease.” *JAMA: The Journal of the American Medical Association* 306 (20): 2221–28.

Mehrizi, Rahil, Arash Mehrjou, Maryana Alegro, Yi Zhao, Benedetta Carbone, Carl Fishwick, Johanna Vappiani, et al. 2023. “Multi-Omics Prediction from High-Content Cellular Imaging with Deep Learning.” *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/2306.09391.

Mestres, Jordi. 2005. “Structure Conservation in Cytochromes P450.” *Proteins* 58 (3): 596–609.

Mitra, A. K., U. K. Mukherjee, T. Harding, J. S. Jang, H. Stessman, Y. Li, A. Abyzov, et al. 2016. “Single-Cell Analysis of Targeted Transcriptome Predicts Drug Sensitivity of Single Cells within Human Myeloma Tumors.” *Leukemia* 30 (5): 1094–1102.

Mund, Andreas, Fabian Coscia, András Kriston, Réka Hollandi, Ferenc Kovács, Andreas-David Brunner, Ede Migh, et al. 2022. “Deep Visual Proteomics Defines Single-Cell Identity and Heterogeneity.” *Nature Biotechnology* 40 (8): 1231–40.

Munro, Andrew W., Hazel M. Girvan, Amy E. Mason, Adrian J. Dunford, and Kirsty J. McLean. 2013. “What Makes a P450 Tick?” *Trends in Biochemical Sciences* 38 (3): 140–50.

Mustafa, Ghulam, Prajwal P. Nandekar, Neil J. Bruce, and Rebecca C. Wade. 2019. “Differing Membrane Interactions of Two Highly Similar Drug-Metabolizing Cytochrome P450 Isoforms: CYP 2C9 and CYP 2C19.” *International Journal of Molecular Sciences* 20 (18). https://doi.org/10.3390/ijms20184328.

Mustafa, Ghulam, Prajwal P. Nandekar, Goutam Mukherjee, Neil J. Bruce, and Rebecca C. Wade. 2020. “The Effect of Force-Field Parameters on Cytochrome P450-Membrane Interactions: Structure and Dynamics.” *Scientific Reports* 10 (1): 7284.

Myers, Samuel A., Andrew Rhoads, Alexandra R. Cocco, Ryan Peckner, Adam L. Haber, Lawrence D. Schweitzer, Karsten Krug, et al. 2019. “Streamlined Protocol for Deep Proteomic Profiling of FAC-Sorted Cells and Its Application to Freshly Isolated Murine Immune Cells.” *Molecular & Cellular Proteomics: MCP* 18 (5): 995–1009.

Nair, Pramod C., Ross A. McKinnon, and John O. Miners. 2016. “Cytochrome P450 Structure-Function: Insights from Molecular Dynamics Simulations.” *Drug Metabolism Reviews* 48 (3): 434–52.

Nebert, Daniel W., Kjell Wikvall, and Walter L. Miller. 2013. “Human Cytochromes P450 in Health and Disease.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368 (1612): 20120431.

Nelson, David R. 2011. “Progress in Tracing the Evolutionary Paths of Cytochrome P450.”

Biochimica et Biophysica Acta 1814 (1): 14–18.

Niehaus, M., J. Soltwisch, M. E. Belov, and K. Dreisewerd. 2019. “Transmission-Mode MALDI-2 Mass Spectrometry Imaging of Cells and Tissues at Subcellular Resolution.” *Nature Methods* 16 (9): 925–31.

Nilsson, Tommy, Matthias Mann, Ruedi Aebersold, John R. Yates 3rd, Amos Bairoch, and John J. M. Bergeron. 2010. “Mass Spectrometry in High-Throughput Proteomics: Ready for the Big Time.” *Nature Methods* 7 (9): 681–85.

Niwa, Toshiro, Akira Kageyama, Kae Kishimoto, Yoshiyasu Yabusaki, Fumihide Ishibashi, and Masanao Katagiri. 2002. “Amino Acid Residues Affecting the Activities of Human Cytochrome P450 2C9 and 2C19.” *Drug Metabolism and Disposition: The Biological Fate of Chemicals* 30 (8): 931–36.

Niwa, Toshiro, and Hiroshi Yamazaki. 2012. “Comparison of Cytochrome P450 2C Subfamily Members in Terms of Drug Oxidation Rates and Substrate Inhibition.” *Current Drug Metabolism* 13 (8): 1145–59.

Oguri, K., H. Yamada, and H. Yoshimura. 1994. “Regiochemistry of Cytochrome P450 Isozymes.” *Annual Review of Pharmacology and Toxicology* 34: 251–79.

Otyepka, Michal, Josef Skopalík, Eva Anzenbacherová, and Pavel Anzenbacher. 2007. “What Common Structural Features and Variations of Mammalian P450s Are Known to Date?” *Biochimica et Biophysica Acta* 1770 (3): 376–89.

Paloncýová, Markéta, Veronika Navrátilová, Karel Berka, Alessandro Laio, and Michal Otyepka. 2016. “Role of Enzyme Flexibility in Ligand Access and Egress to Active Site: Bias-Exchange Metadynamics Study of 1,3,7-Trimethyluric Acid in Cytochrome P450 3A4.” *Journal of Chemical Theory and Computation* 12 (4): 2101–9.

Peng, Chi-Chi, Jonathan L. Cape, Tom Rushmore, Gregory J. Crouch, and Jeffrey P. Jones. 2008. “Cytochrome P450 2C9 Type II Binding Studies on Quinoline-4-Carboxamide Analogues.” *Journal of Medicinal Chemistry* 51 (24): 8000–8011.

Pereira, Naveen L., Charanjit Rihal, Ryan Lennon, Gil Marcus, Sanskriti Shrivastava, Malcolm R. Bell, Derek So, et al. 2021. “Effect of CYP2C19 Genotype on Ischemic Outcomes During Oral P2Y12 Inhibitor Therapy: A Meta-Analysis.” *JACC. Cardiovascular Interventions* 14 (7): 739–50.

Piehowski, Paul D., Ying Zhu, Lisa M. Bramer, Kelly G. Stratton, Rui Zhao, Daniel J. Orton, Ronald J. Moore, et al. 2020. “Automated Mass Spectrometry Imaging of over 2000 Proteins from Tissue Sections at 100-μm Spatial Resolution.” *Nature Communications* 11 (1): 8.

Pino, Lindsay K., Seth C. Just, Michael J. MacCoss, and Brian C. Searle. 2020. “Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries.” *Molecular & Cellular Proteomics: MCP* 19 (7): 1088–1103.

Polgár, Tímea, Dóra K. Menyhárd, and György M. Keseru. 2007. “Effective Virtual Screening Protocol for CYP2C9 Ligands Using a Screening Site Constructed from Flurbiprofen and S-Warfarin Pockets.” *Journal of Computer-Aided Molecular Design* 21 (9): 539–48.

“Precision Health: Improving Health for Each of Us and All of Us.” 2022. September 13, 2022. https://www.cdc.gov/genomics/about/precision_med.htm.

Raatz, Michael, Saumil Shah, Guranda Chitadze, Monika Brüggemann, and Arne Traulsen. 2021. “The Impact of Phenotypic Heterogeneity of Tumour Cells on Treatment and Relapse Dynamics.” *PLoS Computational Biology* 17 (2): e1008702.

Rappsilber, Juri, Matthias Mann, and Yasushi Ishihama. 2007. “Protocol for Micro-Purification, Enrichment, Pre-Fractionation and Storage of Peptides for Proteomics Using StageTips.” *Nature Protocols* 2 (8): 1896–1906.

Relling, M. V., and T. E. Klein. 2011. “CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network.” *Clinical Pharmacology and Therapeutics* 89 (3): 464–67.

Rendic, Slobodan, and F. Peter Guengerich. 2012. “Contributions of Human Enzymes in

Carcinogen Metabolism." *Chemical Research in Toxicology* 25 (7): 1316–83.

Reynald, R. Leila, Stefaan Sansen, C. David Stout, and Eric F. Johnson. 2012. "Structural Characterization of Human Cytochrome P450 2C19: Active Site Differences between P450s 2C8, 2C9, and 2C19." *The Journal of Biological Chemistry* 287 (53): 44581–91.

Roden, Dan M., Howard L. McLeod, Mary V. Relling, Marc S. Williams, George A. Mensah, Josh F. Peterson, and Sara L. Van Driest. 2019. "Pharmacogenomics." *The Lancet* 394 (10197): 521–32.

Rohban, Mohammad Hossein, Shantanu Singh, Xiaoyun Wu, Julia B. Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S. Boehm, and Anne E. Carpenter. 2017. "Systematic Morphological Profiling of Human Gene and Allele Function via Cell Painting." *eLife* 6 (March). https://doi.org/10.7554/eLife.24060.

Roses, A. D. 2000. "Pharmacogenetics and Future Drug Development and Delivery." *The Lancet* 355 (9212): 1358–61.

Rubin, H. 1990. "The Significance of Biological Heterogeneity." *Cancer Metastasis Reviews* 9 (1): 1–20.

Saydam, Faruk, İrfan Değirmenci, Alparslan Birdane, Mahmut Özdemir, Taner Ulus, Cansu Özbayer, Ertuğrul Çolak, Necmi Ata, and Hasan Veysi Güneş. 2017. "The CYP2C19*2 and CYP2C19*17 Polymorphisms Play a Vital Role in Clopidogrel Responsiveness after Percutaneous Coronary Intervention: A Pharmacogenomics Study." *Basic & Clinical Pharmacology & Toxicology* 121 (1): 29–36.

Schmiedl, S., M. Rottenkolber, J. Szymanski, B. Drewelow, W. Siegmund, M. Hippius, K. Farker, 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 Opinion on Drug Safety* 17 (2): 125–37.

Scott, Emily E., C. Roland Wolf, Michal Otyepka, Sara C. Humphreys, James R. Reed, Colin J. Henderson, Lesley A. McLaughlin, et al. 2016. "The Role of Protein-Protein and

Protein-Membrane Interactions on P450 Function." *Drug Metabolism and Disposition: The Biological Fate of Chemicals* 44 (4): 576–90.

Seifert, Alexander, Stephan Tatzel, Rolf D. Schmid, and Jürgen Pleiss. 2006. "Multiple Molecular Dynamics Simulations of Human p450 Monooxygenase CYP2C9: The Molecular Basis of Substrate Binding and Regioselectivity toward Warfarin." *Proteins* 64 (1): 147–55.

Shaffer, Sydney M., Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, et al. 2017. "Rare Cell Variability and Drug-Induced Reprogramming as a Mode of Cancer Drug Resistance." *Nature* 546 (7658): 431–35.

Shaffer, Sydney M., Benjamin L. Emert, Raúl A. Reyes Hueros, Christopher Cote, Guillaume Harmange, Dylan L. Schaff, Ann E. Sizemore, et al. 2020. "Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors." *Cell* 182 (4): 947–59.e17.

Shaffer, Sydney M., Benjamin L. Emert, Ann E. Sizemore, Rohit Gupte, Eduardo Torre, Danielle S. Bassett, and Arjun Raj. 2018. "Memory Sequencing Reveals Heritable Single Cell Gene Expression Programs Associated with Distinct Cellular Behaviors." https://doi.org/10.1101/379016.

Sielaff, Malte, Jörg Kuharev, Toszka Bohn, Jennifer Hahlbrock, Tobias Bopp, Stefan Tenzer, and Ute Distler. 2017. "Evaluation of FASP, SP3, and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range." *Journal of Proteome Research* 16 (11): 4060–72.

Sim, Sarah C., and Magnus Ingelman-Sundberg. 2010. "The Human Cytochrome P450 (CYP) Allele Nomenclature Website: A Peer-Reviewed Database of CYP Variants and Their Associated Effects." *Human Genomics* 4 (4): 278–81.

Sirim, Demet, Michael Widmann, Florian Wagner, and Jürgen Pleiss. 2010. "Prediction and Analysis of the Modular Structure of Cytochrome P450 Monooxygenases." *BMC Structural Biology* 10 (October): 34.

Skopalík, Josef, Pavel Anzenbacher, and Michal Otyepka. 2008. “Flexibility of Human Cytochromes P450: Molecular Dynamics Reveals Differences between CYPs 3A4, 2C9, and 2A6, Which Correlate with Their Substrate Preferences.” *The Journal of Physical Chemistry. B* 112 (27): 8165–73.

Slack, M. D., E. D. Martinez, L. F. Wu, and S. J. Altschuler. 2008. “Characterizing Heterogeneous Cellular Responses to Perturbations.” *Proceedings of the National Academy of Sciences* 105 (49): 19306–11.

Šrejber, Martin, Veronika Navrátilová, Markéta Paloncýová, Václav Bazgier, Karel Berka, Pavel Anzenbacher, and Michal Otyepka. 2018. “Membrane-Attached Mammalian Cytochromes P450: An Overview of the Membrane’s Effects on Structure, Drug Binding, and Interactions with Redox Partners.” *Journal of Inorganic Biochemistry* 183 (June): 117–36.

Subramanian, Indhupriya, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. 2020. “Multi-Omics Data Integration, Interpretation, and Its Application.” *Bioinformatics and Biology Insights* 14 (January): 1177932219899051.

Suiter, Chase C., Takaya Moriyama, Kenneth A. Matreyek, Wentao Yang, Emma Rose Scaletti, Rina Nishii, Wenjian Yang, et al. 2020. “Massively Parallel Variant Characterization Identifies NUDT15 Alleles Associated with Thiopurine Toxicity.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (10): 5394–5401.

Sultana, Janet, Paola Cutroneo, and Gianluca Trifirò. 2013. “Clinical and Economic Burden of Adverse Drug Reactions.” *Journal of Pharmacology & Pharmacotherapeutics* 4 (Suppl 1): S73–77.

Sun, Xiao-Xiao, and Qiang Yu. 2015. “Intra-Tumor Heterogeneity of Cancer Cells and Its Implications for Cancer Treatment.” *Acta Pharmacologica Sinica* 36 (10): 1219–27.

Symmons, Orsolya, and Arjun Raj. 2016. “What’s Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism.” *Molecular Cell* 62 (5): 788–802.

Thomson, R. 2021. “Structural and Functional Characterisation of Ancestral Cytochromes P450

from Family 2 in Tetrapods." espace.library.uq.edu.au. https://espace.library.uq.edu.au/view/UQ:a159633.

Tsao, C. C., M. R. Wester, B. Ghanayem, S. J. Coulter, B. Chanas, E. F. Johnson, and J. A. Goldstein. 2001. "Identification of Human CYP2C19 Residues That Confer S-Mephenytoin 4'-Hydroxylation Activity to CYP2C9." *Biochemistry* 40 (7): 1937–44.

Villén, Judit, and Steven P. Gygi. 2008. "The SCX/IMAC Enrichment Approach for Global Phosphorylation Analysis by Mass Spectrometry." *Nature Protocols* 3 (10): 1630–38.

Vogel, Christine, and Edward M. Marcotte. 2012. "Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses." *Nature Reviews. Genetics* 13 (4): 227–32.

Vries, E. N. de, M. A. Ramrattan, S. M. Smorenburg, D. J. Gouma, and M. A. Boermeester. 2008. "The Incidence and Nature of in-Hospital Adverse Events: A Systematic Review." *Quality & Safety in Health Care* 17 (3): 216–23.

Wada, Yasunobu, Maori Mitsuda, Yasuhiro Ishihara, Masatomo Watanabe, Masahiko Iwasaki, and Satoru Asahi. 2008. "Important Amino Acid Residues That Confer CYP2C19 Selective Activity to CYP2C9." *Journal of Biochemistry* 144 (3): 323–33.

Way, Gregory P., Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C. Caicedo, et al. 2022. "Morphology and Gene Expression Profiling Provide Complementary Information for Mapping Cell State." *Cell Systems* 13 (11): 911–23.e9.

Wei, Huijin, and Xianghua Li. 2023. "Deep Mutational Scanning: A Versatile Tool in Systematically Mapping Genotypes to Phenotypes." *Frontiers in Genetics* 14 (January): 1087267.

Werck-Reichhart, D., and R. Feyereisen. 2000. "Cytochromes P450: A Success Story." *Genome Biology* 1 (6): REVIEWS3003.

Wishart, David S., Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, et al. 2018. "DrugBank 5.0: A Major Update to the DrugBank Database for

2018.” *Nucleic Acids Research* 46 (D1): D1074–82.

Yang, Yue, Sergio E. Wong, and Felice C. Lightstone. 2014. “Understanding a Substrate’s Product Regioselectivity in a Family of Enzymes: A Case Study of Acetaminophen Binding in Cytochrome P450s.” *PloS One* 9 (2): e87058.

Yeh, Chiann-Ling C., Clara J. Amorosi, Soyeon Showman, and Maitreya J. Dunham. 2022. “PacRAT: A Program to Improve Barcode-Variant Mapping from PacBio Long Reads Using Multiple Sequence Alignment.” *Bioinformatics* 38 (10): 2927–29.

Yu, Fengchao, Guo Ci Teo, Andy T. Kong, Ginny Xiaohe Li, Vadim Demichev, and Alexey I. Nesvizhskii. 2022. “One-Stop Analysis of DIA Proteomics Data Using MSFragger-DIA and FragPipe Computational Platform.” *bioRxiv*. https://doi.org/10.1101/2022.10.28.514272.

Zanger, Ulrich M., and Matthias Schwab. 2013. “Cytochrome P450 Enzymes in Drug Metabolism: Regulation of Gene Expression, Enzyme Activities, and Impact of Genetic Variation.” *Pharmacology & Therapeutics* 138 (1): 103–41.

Zhang, Jiajie, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. 2014. “PEAR: A Fast and Accurate Illumina Paired-End reAd mergeR.” *Bioinformatics* 30 (5): 614–20.

Zhang, Kewei, Xiaomei Yin, Kaituo Shi, Shihua Zhang, Juan Wang, Shasha Zhao, Huan Deng, et al. 2021. “A High-Efficiency Method for Site-Directed Mutagenesis of Large Plasmids Based on Large DNA Fragment Amplification and Recombinational Ligation.” *Scientific Reports* 11 (1): 10454.

Zhang, Lingxin, Vivekananda Sarangi, Ming-Fen Ho, Irene Moon, Krishna R. Kalari, Liewei Wang, and Richard M. Weinshilboum. 2021. “SLCO1B1: Application and Limitations of Deep Mutational Scanning for Genomic Missense Variant Function.” *Drug Metabolism and Disposition: The Biological Fate of Chemicals* 49 (5): 395–404.

Zhang, Lingxin, Vivekananda Sarangi, Irene Moon, Jia Yu, Duan Liu, Sandhya Devarajan, Joel M. Reid, Krishna R. Kalari, Liewei Wang, and Richard Weinshilboum. 2020. “CYP2C9 and CYP2C19: Deep Mutational Scanning and Functional Characterization of Genomic

Missense Variants.” *Clinical and Translational Science* 13 (4): 727–42.

Zhao, Lu, Zhimin Liu, Sasha F. Levy, and Song Wu. 2018. “Bartender: A Fast and Accurate Clustering Algorithm to Count Barcode Reads.” *Bioinformatics* 34 (5): 739–47.

Zhao, Mingzhe, Jingsong Ma, Mo Li, Yingtian Zhang, Bixuan Jiang, Xianglong Zhao, Cong Huai, et al. 2021. “Cytochrome P450 Enzymes and Drug Metabolism in Humans.” *International Journal of Molecular Sciences* 22 (23). https://doi.org/10.3390/ijms222312808.

Zubarev, Roman A. 2013. “The Challenge of the Proteome Dynamic Range and Its Implications for in-Depth Proteomics.” *Proteomics* 13 (5): 723–26.