Large-scale mutational analysis of transporters in the Solute Carrier Family 22: applications in rare disease and pharmacogenetics

by
Megan Koleske

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Jason Gestwicki
4909848DBB404E5...
Jason Gestwicki
Chair

DocuSigned by:
Kathleen Giacomini
Kathleen Giacomini

DocuSigned by:
Renata Gallagher
D2132452BBFF4E9...
Renata Gallagher

Committee Members

ACKNOWLEDGMENTS AND DEDICATION

There are many people to acknowledge and thank for their part in supporting me throughout graduate school and the completion of this dissertation. Below I'd like to express my most sincere gratitude to each of you.

First, to my mentor and dissertation advisor, Dr. Kathleen Giacomini, for years of guidance, leadership, and support. You challenged me until I discovered I am capable of much more than I dreamed possible, and then continued to challenge me until I truly believed it. Thank you for trusting me and giving me the independence to fly. I am not the same scientist, nor woman, I was before I stepped foot in your lab, and I am better for it.

To my thesis committee members, Drs. Renata Gallagher and Jason Gestwicki, for your scientific expertise and input that greatly shaped this work for the better, but more so for your guidance, mentorship, and kindness throughout the journey.

To the Biohub Mavericks team, for helping to shape the foundation of what would morph into my main thesis project, and specifically to Greg McInnes, for being such a brilliant and giving collaborator.

To my undergraduate research advisor, Dr. Przemysław Radwański, for giving me my first research experience, teaching me the foundations, and giving me the independence to build confidence as a scientific researcher.

To all Giacomini Lab members past and present, specifically Sook Wah Yee and Elizabeth Ennis Green who trained me, and my mentees Mattias Rodin and Sebastian Jakobsen who taught me so much through training you. Most notably, to my fellow Giacomini Lab graduate students, Bianca

Vora, Dina Buitrago, and Xujia Zhou: in addition to countless hours of scientific conversations working through ideas and problem-solving, I am most grateful for the friendship, encouragement, and laughter that provided much respite throughout this journey when long days in lab didn't go so well. I am honored to have shared this experience with you, to have learned and grown together.

To my fellow PSPG classmates, for the community that first made SF feel like home. To the ia's, Tia Tummino and Capria Rinaldi, for being the most caring and understanding support system during this time. To those that came before me, specifically Adam Melgoza, Arielle Shkedi, Serena Tamura, for the wisdom and guidance passed on to me from your experiences.

To friends in San Francisco, Jennifer Stokes, Ankita Dhussa, Kristin Boyd, Sandhya Rangarajan, Emily Bassett, and Neena Joshi, for your patience in the hours of listening to me talk about this journey, for celebrating the big and little wins with me, and for all the laughter-filled moments away from it all.

To old friends in distant places, Tori Smith, Grace Stammen, Samantha Saraceni, Michelle Kryc, and Chantelle Hobbs, for the strongest bonds of friendship that endured the test of both distance and time. I am beyond proud and empowered to be surrounded by such brilliantly successful women who impress, uplift, and inspire me endlessly.

To my love, Samuel, for being my daily sounding board, closest confidant, adventure partner, and best friend. For always knowing how to make me laugh and calm me down. For reminding me it's just a small piece of something greater. I am so lucky to share this life with you.

Most importantly, to my family, for the love that shapes the most integral parts of my being. To my brother, Nick, for being the most kind and patient role model throughout my entire life. When we were little, I followed you everywhere and wanted to do exactly what you did; to this day you are the best person I know. To my parents, for giving me a life more beautiful than I could have dreamed. To my father, Matt, for offering the best advice, always having a solution, and knowing the perfect moment to crack a joke. From you I learned two of the most important skills in life: how to have a strong work ethic and how to make anything fun. Watching you seamlessly weave the two together has taught me so much. You have always been my biggest inspiration. To my mother, Sharon, for being my #1 supporter and answering every single phone call. I have lived my entire life in deep admiration of the woman you are, the kindness and empathy that touches everyone you meet, the selflessness that radiates from you, the joy and energy you bring to a room. You are the greatest friendship of my life.

This dissertation is dedicated to any version of my former self that ever doubted she was smart enough, worked hard enough, or was a good enough scientist to finish this journey. You were, you are, you could, you did. I'm proud of you.

CONTRIBUTIONS

Several chapters in this dissertation contain material that has been published previously or is currently under consideration/review for publication. Chapters as presented here do not necessarily represent the final published form and in some cases have been edited slightly.

Chapter 1 is reproduced in part with permission from the publication: Koleske ML, Liang X, Enogieru OJ, Buitrago D, Giacomini KM (2022). Organic Cation and Zwitterion Transporters. In You G & Morris ME (Eds.), *Drug Transporters: Molecular Characterization and Role in Drug Disposition, 3rd Edition*. Wiley. All authors wrote, reviewed, and approved the manuscript.

Chapter 2 is reproduced in full with permission from the publication: McInnes G*, Sharo AG*, Koleske ML*, Brown JEH*, Norstad M, Adhikari AN, Wang S, Brenner SE, Halpern J, Koenig BA, Magnus DC, Gallagher RC, Giacomini KM, Altman RB. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021 Apr 1;108(4):535-548. G.M., A.G.S., M.L.K., and J.E.H.B. wrote the manuscript, all authors reviewed and approved the manuscript. *These authors contributed equally to this work.

Chapter 3 is reproduced in full from a manuscript currently in review for publication: Koleske ML, McInnes G, Brown JEH, Thomas N, Hutchinson K, Chin MY, Koehl A, Arkin MR, Schlessinger A, Gallagher RC, Song YS, Altman RB, Giacomini KM. Large-scale functional characterization of OCTN2 variants informs protein-specific variant effect predictor for Carnitine Transporter Deficiency. M.L.K. wrote the manuscript with contributions from G.M., J.E.H.B., and N.T., all authors reviewed and approved the manuscript.

Chapter 4 is written by Megan Koleske and contains unpublished experimental results with contributions from Sook Wah Yee and Willow Coyote-Maestas. Willow Coyote-Maestas aided in the design of the OCT1 variant library, Sook Wah Yee generated the library.

Chapter 5 is written by Megan Koleske.

"I HAVE NO SPECIAL TALENT. I AM ONLY PASSIONATELY CURIOUS."

- ALBERT EINSTEIN

Large-scale mutational analysis of transporters in the Solute Carrier Family 22: applications in rare disease and pharmacogenetics

Megan Koleske

ABSTRACT

From disease diagnostics to precision dosing of medication, genome sequencing has the potential to revolutionize healthcare. A key challenge in translating genetic information into clinical action is understanding the phenotypic consequences of genetic variants in clinically important genes. Transporters encoded by genes in the Solute Carrier (SLC) family 22 have strong clinical relevance in rare genetic disease and pharmacogenetics. For example, loss-of-function variants in *SLC22A5*, encoding the carnitine transporter OCTN2, cause the rare metabolic disorder Carnitine Transporter Deficiency (CTD), and variants with functional consequences in *SLC22A1*, encoding the hepatic uptake transporter OCT1, contribute to interindividual differences in exposure and response for many commonly used medications. Experimental studies to uncover phenotypic consequences of coding region variants in transporters and other genes have struggled to match pace with the rate at which variants are identified by next-generation sequencing, slowing translation into clinically actionable information. The goal of this dissertation research is to experimentally and computationally address the key challenge of understanding the phenotypic effects of genetic variants in SLC22 transporters, with a primary focus on OCTN2 and OCT1.

The dissertation begins with an overview of current practices and limitations in the interpretation of variation in genes underlying inborn errors of metabolism and drug response, detailing

experimental approaches to validating a variant as causative or pathogenic and summarizing advances in computational methods aiming to predict variant effect on protein function. We propose a vision for a genomic learning healthcare system (GLHS) that facilitates the translation of a patient's genome into clinically actionable information for diagnostic and therapeutic purposes. After the overview, we present a rich set of experimental and computational approaches, which were developed and used to improve the functional prediction of genetic variants in OCTN2 for diagnosis of CTD. We functionally characterized 150 OCTN2 missense variants and found that 71% of variants had a significant effect on the uptake of carnitine. 25% of variants reduced transporter function to less than 20% of the wild-type OCTN2, a clinically meaningful threshold for CTD. We asked what was causing reduced function, and identified improper subcellular localization to be a major loss-of-function mechanism affecting 62% of variants. These data were then used in machine learning to build a protein-specific variant effect prediction model that accurately classified variants of OCTN2 as functional (>20%) or LOF (<20%) (area under the receiver operating characteristics curve 0.895±0.025). The machine learning models outperformed current models in terms of functional predictions of genetic variants in OCTN2. Limitations, however, included the number of variants experimentally tested to inform the models, which were limited by experimental methodologies. Therefore, we asked how we can increase throughput of SLC transporter variant phenotyping and transfer predictive models to other SLC22 family members. To this end, we developed a platform for deep mutational scanning (DMS) of a homolog of OCTN2 in the SLC22 family, OCT1 (*SLC22A1*) to increase the scale and diversify the phenotypes with which we can investigate functional genomics of transporters. We generated a landing-pad based cell system for expression of OCT1 and validated the system with multiple assays involving diverse substrates and phenotypes:

uptake of the fluorescent substrate ASP^+; uptake of the radiolabeled substrates MPP^+ and metformin, and cytotoxicity of the OCT1 substrates, oxaliplatin and platinum analogs SM73 and SM85. We confirmed that OCT1 variants exhibit substrate-specific functional effects with variants p.R61C, p.P117L, and p.G401S. Then, we constructed a library of all 11,572 possible missense and single amino acid deletion variants to undergo functional and spatial characterization by the established DMS system. Data generated with this platform will be useful in the interpretation of OCT1 variants and clinically actionable for drug dosing purposes in precision medicine.

In summary, this dissertation research led to a top performing model for predicting the functional effects of variants in OCTN2, which may be causal for CTD. Importantly, we addressed limitations in our model and developed experimental methodologies that extended both the scale of the genetic variants under investigation and the functional phenotypes assessed. Collectively, these methods together with artificial intelligence including machine and transfer learning methods should lead to comprehensive models for accurately predicting the function of coding region variants of all genes in the SLC22 family. These studies will pave the way to a new understanding of the effects of genetic variants in SLC transporters that are causal for human disease and diverse pharmacogenetic phenotypes.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

Chapter 1: Organic Cation and Zwitterion Transporters

1.1 Abstract

Transporters in the solute carrier superfamily (SLC) play critical roles in the absorption and disposition of numerous solutes in the human metabolome. The focus of this chapter is on transporters for two major categories of solutes: organic cations and zwitterions. The chapter begins with an overview of organic cation transporters (OCTs) and places a major focus on OCT1 (*SLC22A1*). Zwitterion transporters are described in the second section of the chapter, with the main focus on OCTN2 (*SLC22A5*). For both transporters, information is provided on tissue distribution, ligand selectivity, and transport mechanism. In addition, we include information from genetically engineered mouse models, as well as human genetic and pharmacogenomic studies describing clinical associations between genetic polymorphisms or mutations in the individual transporters and clinical phenotypes. As over half of prescription drugs are basic compounds, polymorphisms in OCTs have been associated with many pharmacogenomic traits. Further, as carnitine, a zwitterion, is a key molecule in fatty acid oxidation, many associations with zwitterion transporters include phenotypes that are ultimately related to disorders in energy production. The chapter ends with a brief discussion of future research that is needed to advance our understanding of organic cation and zwitterion transporters.

1.2 Introduction to the Organic Cation Transporter Family

Within the human SLC22 transporter family, the electrogenic organic cation transporter (OCT) subfamily consists of three members: OCT1 (*SLC22A1*), OCT2 (*SLC22A2*), and OCT3 (*SLC22A3*). These transporters play important physiological and pharmacological roles, as they transport a variety of structurally diverse endogenous compounds and xenobiotics that have a net

positive charge at physiological pHs. Alterations in the expression and function of these transporters can lead to various pathophysiological conditions. As the first member of the family, rat OCT1 (*Slc22a1*) was cloned and characterized in 1994, followed by rat OCT2 (*Slc22a2*) in 1996 (1). The third member, OCT3, was identified and cloned in both rat and human in 1998 (1). These three transporters share similar transport mechanisms and have overlapping ligand specificities; however, they differ in terms of their tissue distribution and the mechanisms involved in the regulation of their expression.

1.2.1 Tissue Distribution

Despite the similarity of ligand specificity and transport function, the tissue distribution of the three OCTs varies greatly in humans and other species (Fig. 1.1). Though expressed in many tissues, human OCT1 is most highly expressed in the liver and localized to the basolateral membrane of hepatocytes (1). In rodents, high expression of OCT1 is also observed on the sinusoidal membrane of hepatocytes (1). However, the expression and location of OCT1 in the kidney differs between human and rodents.

Whereas in rodents, OCT1 is also expressed in the kidney and located on the basolateral membrane of proximal tubules, very low levels of OCT1 mRNA are detected in human kidney (2). In addition to the expression pattern in the kidney and liver of rodents, Oct1 is also expressed in rodent intestine on the basolateral membrane of enterocytes. In contrast, in the trachea and bronchi of human, rat, and mouse, OCT1 appears to be expressed on the luminal membrane of epithelial cells (1). In addition, low expression levels of human OCT1 (hOCT1) have been detected in other tissues including brain, heart, skeletal muscle, peripheral leukocytes, adrenal gland, mammary gland, immune cells, and adipose tissue (3).

1.2.2 Structure–Function Relationship

In human, the genes *SLC22A1, SLC22A2,* and *SLC22A3,* which encode OCT1, OCT2, and OCT3, respectively, are localized within a cluster on chromosome 6q26-27 (1). Each of these genes comprise 11 exons and 10 introns. OCT1–3 contain 554, 555, and 556 amino acids, respectively. hOCT1 and hOCT2 are approximately 70% identical in amino acid sequence, whereas hOCT3 shares 50% sequence identity with hOCT1 and hOCT2 (1). The precise mechanism of binding and transport is not fully uncovered due to the lack of a high-resolution crystal structure (4). The predicted 2 and 3D structures of hOCTs consist of a typical major facilitator superfamily (MFS) fold of 12 α-transmembrane helix domains arranged in a barrel-shaped structure with a large cleft that opens in cytoplasm. The modeled 3D structures of all hOCTs are based on the crystal structure of the human *SLC2A3* (GLUT3) transporter and display an outward-open conformation and putative cyclic C1 protein symmetry (4). The NH_2- and COOH-terminal ends of the OCTs are intracellular (1). All three transporters contain a large (100+ amino acids) extracellular loop between transmembrane domain (TMD) 1 and TMD2 and a relatively large intracellular loop between TMD6 and TMD7 (1, 4). The large extracellular loop contains N-glycosylation sites (Asn-Xaa-Ser/Thr) and cysteine residues, features indicative of putative roles in drug binding and uptake. The large intracellular loop contains several predicted sites for protein kinase C (PKC)-dependent phosphorylation. Phosphorylation of these sites changes substrate selectivity (1, 4). In addition, homology models of inward-facing and outward-facing tertiary structures of OCTs have been generated based on *E. coli* transporters, lactose permease LacY and the glycerol-3-phosphate transporter GlpT (5). The transmembrane domains, and in particular the 4th and 10th transmembrane domains, are thought to be critically involved in substrate recognition by the OCTs, and differences between the three isoforms in

terms of substrate specificity may be related to differences in these regions. Extensive site-directed mutagenesis followed by functional characterization of mutants has indicated that transported organic cations bind to amino acids in the innermost cavity of the outward open binding cleft. The binding sites for different transported organic cations are overlapping but nonidentical so that exchange of one amino acid in this innermost cleft may change affinity for one substrate but not another (5). These results suggest that OCT1, and likely all OCTs, contains multiple overlapping but nonidentical recognition sites for the various structurally diverse substrates. Further mutational analyses in OCT1 and OCT2 support the occurrence of a complex binding pocket in these transporters. The binding pocket might appear in inward- or outward-oriented conformation and these conformations can differ in substrate affinity (6). On the basis of uptake studies for hOCT2, a model has been suggested where two substrates can bind simultaneously to the transporter. Upon binding, the resulting transporter/substrate1/substrate2 complex cannot be translocated (7), suggesting an inhibition mechanism. However, it is important to note that given the broad substrate selectivity of the OCTs, the key domains or residues involved in substrate recognition may differ by substrate, even within the same protein.

1.2.3 Transport Mechanism

OCT1–3 function as uniporters (Fig. 1.2B), facilitating diffusion of substrates across the plasma membrane. Transport can be bidirectional, depending on substrate, and is driven by electrochemical gradients. The OCTs share several common features related to transport mechanism. First, modeling and mutational analysis suggest that OCTs follow an alternating-access transport model. The substrate binds to the outward-open conformation of the transporter, which induces a conformational change. Then the substrate–transporter complex passes a transient occluded state to the inward-open conformation. Lastly, the substrate is released to the

cytoplasm and the transporter returns to the outward open conformation (6) (Fig. 2.2A). The structural changes of OCTs during the transport cycle require a rigid body movement of the six N-terminal TMDs against the six C-terminal TMDs, and a hinge domain in TMD 11 is crucial for this movement (5, 8). Second, the translocation of organic cations by OCTs is electrogenic and independent of sodium and chloride ions (5). The net transport of organic cations is driven by the intracellular negative membrane potential and the concentration gradient. Positively charged cations are taken up into cells according to the electrochemical gradient, and this process is membrane sensitive. Artificially modulating the membrane potential by replacement of extracellular Na^+ with K^+ changes the rate of transport by OCTs (9). Third, the transport direction of OCTs is bidirectional, and as noted, net transmembrane flux is dependent on the electrochemical gradient. In addition to cation influx, OCTs acting as efflux transporters have been demonstrated in multiple studies (1). Fourth, OCT1-3, defined as "poly-specific" OCTs, can transport a variety of substrates with diverse molecular structures. As such, their substrates tend to have higher K_m values than those of the substrates for the more specific transporters, such as the neurotransmitter transporters (SLC6). In addition, OCT1–3 can be inhibited by a large number of compounds that are not transported. Common substrates of all OCTs are relatively low molecular mass (below 500 g/mol) and hydrophilic organic cations such as the prototypical cation tetraethylammonium (TEA), the neurotoxin MPP^+, and the endogenous compound N-methylnicotinamide (NMN). Several clinically important drugs have been shown to interact with all of the OCTs, including the antidiabetic drug metformin. Besides this, endogenous compounds such as the biogenic amine neurotransmitters (i.e., dopamine, epinephrine, nor- epinephrine, histamine, and serotonin) have been shown to interact with one or more OCT transporters (1). Although the OCT family shows broad overlap in substrate specificity, there are examples of

relatively isoform-specific and species-specific substrates and inhibitors. More details will be provided in the next section.

1.3 OCT1

1.3.1 Substrate and Inhibitor Selectivity

Compounds that are commonly transported by hOCT1 include the model cations MPP^+, TEA, tetrapropylammonium (TPrA), tetrabutylammonium (TBuA), N-methylquinine, and N-(4-4-azo-n-pentyl)-21-deoxyajmalinium, the endogenous compounds choline, acetylcholine and agmatine, and the drugs quinidine, quinine, acyclovir, ganciclovir, metformin, sumatriptan, ondansetron, morphine, and several anticancer agents, (e.g., anthracyclines) (Table 1.1) (10). Both human and mouse OCT1 are high-capacity thiamine (vitamin B1) transporters that respond to the uptake of dietary thiamine to liver (11). The series of n-tetraalkylammonium (nTAA) compounds has been shown to have different affinity among human, rabbit, rat, and mouse OCT1. While the larger nTAAs are transported at greater rate in hOCT, the smaller nTAAs are transported at greater rate in rOCT1 or mOCT1 (12). It is suggested that molecular mass or hydrophobicity may affect differences in recognition of OCT substrates across species. In terms of inhibition, some inhibitors of OCTs show differences in potency among the individual subtypes. For example, the inhibition potency of phencyclidine, diphenhydramine, prazosin, citalopram, and atropine is greater for hOCT1 com- pared with hOCT2 and hOCT3. In contrast, corticosterone shows stronger inhibition on hOCT3 than OCT1 (1). Besides influx transport, OCT1 has been demonstrated as an efflux transporter of acylcarnitine from liver to plasma (13).

1.3.2 Regulation

In human, OCT1 is predominantly expressed in hepatocytes. Thus, the *SLC22A1* gene is suggested to be regulated by the liver-enriched transcription factors, such as hepatocyte nuclear

factor 4a (HNF4a), CCAAT/enhancer binding proteins α and β, hepatocyte nuclear factor 1a (HNF1a) and 3c (HNF3c). Two co-operating HNF4a response elements have been identified between nucleotides -1479 and -1441 of the 5′-flanking regions of the *SLC22A1* gene, which are upstream of the transcription initiation sites. In electrophoretic mobility shift assays (EMSA), recombinant HNF4a directly interacts with both sites. Mutation of these sites in *SLC22A1* promoter luciferase reporter constructs abolish transactivation (14). These motifs are conserved in primates, but not rodents, indicating different patterns of *SLC22A1/ Slc22a1* gene regulation in these species (15). In addition, upstream stimulating factors USF1 and USF2 have been identified to regulate basal hepatic expression of OCT1 via a cognate E-box (15). OCT1 expression can be modulated by ligand-dependent nuclear receptors such as pregnane X receptor, farnesoid X receptor, constitutive androstane receptor, glucocorticoid receptor or peroxisome proliferator- activated receptor α and γ. In addition, the transcription expression level of OCT1 can be regulated by epigenetic methylation (16). Moreover, OCT1 can be subject to post-translational modulation. Stimulation of either protein kinase A (PKA) or PKC increases uptake of the fluorescent compound ASP^+ in HEK293 cells stably transfected with rat OCT1 (1). However, this effect may be species-dependent (17). As with rOCT1, hOCT1 appears to be positively regulated by the p56lck tyrosine kinase, as evidenced by reduced hOCT1 activity after treatment with aminogenistein. Human OCT1 has further been shown to be regulated by the Ca^{2+}/calmodulin complex, which appears to affect the affinity of the tested substrates (17), possibly due to phosphorylation of the OCT1 protein.

1.3.3 Animal Models

In vivo studies in mice, in which individual transporters have been removed genetically (knockout mice) provide valuable insights in potential physiologic and biomedical functions of

OCTs. However, species differences between OCTs of humans and rodents impose limitations on the ability to apply conclusions obtained from the mouse experiments to humans. Oct1 knockout mice are viable and fertile, with no obvious physiological abnormalities when compared with their wild-type littermates, suggesting that Oct1 has minimal impact on normal physiology (18). However, at the biochemical level, disruption of Oct1 in mice affects both lipid and glucose metabolism by reducing the hepatic uptake of thiamine (16). Additionally, disruption of Oct1 in mice has significant impact on the disposition of organic cations. For example, when administered the prototypical organic cation, TEA, $Oct1^{-/-}$ mice show significantly reduced uptake of TEA into the liver. In accordance with reduced hepatic uptake of TEA, biliary excretion is lower in $Oct1^{-/-}$ mice (18). Additionally, direct intestinal excretion of TEA is reduced by approximately 50%. In addition to TEA, $Oct1^{-/-}$ mice have similar decreases in hepatic uptake of other OCT1 substrates (i.e., MPP^{+} and meta-iodobenzylguanidine (MIBG)) (18).

In addition to pharmacokinetic effects, knockout of Oct1 in mice can have implications for prescription drugs, exemplified by metformin, an anti-hyperglycemic pre- scription medication used as a first-line treatment for Type 2 diabetes. Despite similar pharmacokinetic profiles between $Oct1^{-/-}$ and wild-type mice, $Oct1^{-/-}$ mice showed greater than 30-fold decrease in metformin uptake into liver, the site of action for metformin, compared with wild-type littermates (19). Further studies investigated the role of Oct1 in the development of metformin-induced lactic acidosis, a leading toxicity from this drug. A significant increase in serum lactic acid concentration was observed after administration of metformin to wild-type mice, but only slight elevations in serum lactate were seen in $Oct1^{-/-}$ mice (20). Taken together, these results suggest that OCT1-mediated metformin transport is a limiting step in metformin uptake into

liver, and that the lactic acidosis induced by metformin is related to the availability of the drug to its target organ. Recent studies have demonstrated large effects of knocking out Oct1 on the hepatic uptake and clearance of sumatriptan and fenoterol, and lesser effects on ondansetron (21).

1.3.4 Human Genetic Studies

Human OCT1 is the most polymorphic of the OCTs in terms of missense variants. In 2002, 7 missense variants and an amino acid deletion, 420del, in OCT1 were identified in samples from Europeans, and in 2003, 14 missense variants and 420del were identified in OCT1 in samples from four major ancestral groups (22, 23). Since then, over 1000 single-nucleotide polymorphisms (SNPs) have been identified (24), 22 of which have been related to treatment outcome for drugs transported by OCT1. Among these variants, 21 are located in the protein-coding region causing amino-acid substitutions, while one results in an amino-acid deletion (p.M420del). Notably, the frequency of missense alleles in OCT1 is ancestry-specific. European, African, and Latin American (Puerto Rican, Colombian and Mexican) populations present with higher variability than Asians and Pacific Islanders (24). Six variants in OCT1 have a global allele frequency > 0.02 (Fig. 1.3).

Many nonsynonymous polymorphisms of OCT1 have been functionally characterized *in vitro*. The uptake of the cation OCT1 substrate ^{3}H-MPP^{+} was reduced in *Xenopus* oocytes expressing variants p.R61C, p.C88R, p.G401S, p.P341L, p.G220V, and p.G465R that were identified in a large sample of ancestrally diverse healthy subjects (22, 23). Of the variants with significant functional differences from the reference OCT1, five (p.S14F, p.R61C, p.P341L, p.G401S, and p.G465R) occur at >1% allele frequency in at least one ancestral group. OCT1 variants p.P283L and p.R287G exhibit no uptake of either ^{14}C-TEA or ^{3}H-MPP^{+}, although the protein level on the

membrane of these variants is comparable to reference OCT1. The results suggest that residues Pro283 and Arg287 have a substantial role in substrate recognition or the transport cycle of OCT1 (16). In another study, 12 OCT1 nonsynonymous variants were stably expressed in HEK cells, and metformin was used to characterize the uptake function of these variants (25). Although the mRNA expression of these variants is comparable to reference allele, 7 OCT1 variants exhibit significantly reduced or lost metformin uptake. The GFP-tagged p.G465R and p.R61C variants display abnormal localization on the plasma membrane (25). Furthermore, the uptake of metformin is significantly reduced in cells expressing variants identified in Chinese and Japanese populations including p.Q97K, p.P117L, and p.R206C relative to the OCT1 reference (26).

Due to a high level of evidence from many *in vitro* uptake studies and clinical pharmacogenomic studies on *SLC22A1*, the importance of OCT1 as an emerging transporter on drug disposition, response, and toxicity has been highlighted and discussed by the International Transporter Consortium (ITC) (27). Many of these genetic associations have focused on the antidiabetic drug, metformin. However, studies of the effects of OCT1 polymorphisms on metformin pharmacokinetics and pharmacodynamics have been inconsistent. For example, while significant associations of missense OCT1 polymorphisms with metformin plasma concentrations were observed in several pharmacogenomics studies, other studies failed to observe such effects in either healthy sub- jects or patients with type 2 diabetes (16). Positron emission tomography (PET)/computed tomography (CT) using ^{11}C-metformin showed that individuals who are carriers of the OCT1 reduced function variants, p.M420del and p.R61C, have decreased concentrations of metformin in the liver without changes in systemic plasma levels compared with individuals with reference OCT1 (16). Additional associations between OCT1 polymorphisms and drug

levels or response to prescription drugs other than metformin have been studied. These studies have demonstrated significant associations between reduced function nonsynonymous variants of SLC22A1 and the antimigraine drug, sumatriptan (16), the antinausea drug, ondansetron (28), and opiate analgesic drugs or their metabolites, including morphine and O-desmethyltramadol (16). More recently, OCT1 variants were also shown to significantly affect physiology and pathology. Human genome-wide association study data indicate a possible correlation between metabolic phenotypes and OCT1 genotypes, which may be related to the disposition of its endogenous substrates, thiamine, and acylcarnitine (11, 13, 29).

1.3.5 Biomarkers and FDA Guidances for Transporter-Mediated DDIs

Polypharmacy commonly exists in older and chronic disease populations. Transporters can interact with a wide range of endogenous and xenobiotic substrates. Significant drug–drug interactions (DDI) can lead to unfavorable efficacy and safety concerns, and therefore, industry, academia, and regulatory agencies have increased the recognition of transporter-mediated drug interactions. In 2010, the ITC proposed seven transporters as sites for DDIs including P-gp, BCRP, OATP1B1 and 1B3, OAT1 and 3, and OCT2 (30), which was updated to include MATEs. More recently, the ITC suggested that OCT1 and OATP2B1 be added (31). The FDA cites manuscripts from the ITC recently published DDI guidance documents, including one focused on *in vitro* DDI assessment and the other focused on clinical DDI evaluation. These guidances describe the conduct of *in vitro* transporter studies and the use of specific criteria to assess the potential for drugs to interact with transporters and either perpetrate DDIs or be subjected to DDIs. More recently, potential endogenous biomarkers for transporters are being explored as an additional approach to assess the DDI liability of drug candidates (32). For OCTs, potential biomarkers, which may lack specificity for individual OCT isoforms, include NMN,

tryptophan, and creatinine in addition to thiamine (32). Full assessment of the bio- synthesis and elimination pathways of these compounds as well as extensive studies validating their specificity and usefulness in predicting clinical DDIs are needed.

1.4 Introduction to the Zwitterion Transporters

Within the human SLC22 transporter family, the zwitterion transporter subfamily is composed of hOCTN1 (*SLC22A4*), hOCTN2 (*SLC22A5*), FLIPT1/SLC22A15 (*SLC22A15*), and CT2 (*SLC22A16*), among others. These transporters play important physiological and pharmacological roles, acting in the influx and efflux of essential endogenous compounds (e.g., carnitine and ergothioneine), drugs, and various xenobiotics. Alterations in the expression and function of these transporters can lead to various pathophysiological conditions.

1.4.1 Tissue Distribution

In humans, OCTN2 is expressed ubiquitously at low levels in most tissues. Highest expression is observed in skeletal muscle, brain, kidney, intestine, cardiac tissue, and reproductive organs. Many of these tissues have high energy demands and rely heavily on fatty acid β-oxidation for ATP production. OCTN2 expression in these tissues ensures carnitine stores are available to conjugate to intra- cellular long-chain fatty acids for translocation into the mitochondrial matrix where β-oxidation occurs. In the kidney, OCTN2 is localized to the apical membrane of the renal proximal tubule where it functions largely in the reabsorption of renally excreted carnitine from urine to maintain systemic levels.

1.4.2 Structure-Function Relationship

Human OCTN (hOCTN) transporters are localized to the plasma membrane of the cell and are involved in the bidirectional transport of cations and zwitterions. The genes encoding hOCTN1

and hOCTN2 are found in relative proximity at the same locus on chromosome 5q31 (16). Each are encoded by 10 exons, with hOCTN1 composed of 551 amino acids and hOCTN2 composed of 557 amino acids. They are homologues—with 78% sequence identity at the mRNA level and 76% sequence identity at the protein level. Additionally, these transporters share about 30% protein identity to OCT1–3. Similar to OCTs, OCTNs have predicted topology with 12 transmembrane domains, cytoplasmic N- and C-termini, a large extracellular loop between TMD1 and TMD2 containing multiple glycosylation sites, and an intracellular nucleotide binding sequence motif (16). Facilitating sodium-dependent transport of some substrates, the OCTN transporters also contain sodium-recognition sites and can function as symporters (Fig. 1.2B).

1.4.3 Transport Mechanism

Similar to the OCTs, the zwitterion transporters are believed to function through the alternating-access transport mechanism. Multiple transport mechanisms have been observed for the OCTN transporters, depending on substrate. The zwitterion transporters can function as uniporters, like OCT1-3, translocating single substrates, or as cotransporters, transporting multiple substrates in the same direction (symport) or opposite directions (antiport) (Fig. 1.2B) (33). OCTN1 transports the zwitterion ergothioneine via a sodium-dependent symport uptake mechanism, but can transport other zwitterions (e.g., gabapentin) independent of sodium. OCTN1 can also act as a pH- dependent proton/cation antiporter (e.g., TEA), or a bidirectional organic cation uniporter (e.g., acetylcholine). OCTN2 acts as a secondary active sodium-dependent cotransporter for carnitine, facilitating symport of sodium and carnitine at a 1:1 ratio (16). OCTN2 transport of some cations, including TEA, is sodium-independent, while other cations are transported via proton/cation antiport (16).

1.5 OCTN2

1.5.1 Substrate and Inhibitor Selectivity

In vitro, hOCTN2 is multi-specific and has been shown to transport a number of endogenous compounds and xenobiotics (16). Primarily, OCTN2 is a sodium-dependent, high-affinity L-carnitine transporter with a K_m of 4 μM. To a lesser extent, OCTN2 transports some short-chain acylcarnitines, including acetyl-L-carnitine and the drug metabolites pivaloylcarnitine and valproylcarnitine. OCTN2 transports the prototypical cation, TEA, in a sodium-independent manner. Other substrates of OCTN2 include drugs ipratropium, mildronate, amisulpride, sulpiride, etoposide, ethambutol, cephaloridine, quinidine, and verapamil (Table 1.1). *In vitro*, many approved drugs act as inhibitors of OCTN2. Transport of L-carnitine is inhibited by β-lactam antibiotics including cefepime, cefoselis, cephaloridine, cefuroxime, cephalexin, and cefazolin with varying IC_{50} values, likely attributed to the presence of a quaternary amine functional group similar to carnitine. Other strong inhibitors span many drug classes, including cardiac drugs including verapamil, quinidine, and amiodarone, proton-pump inhibitors such as omeprazole, and anticancer agents including tamoxifen, gefitinib, and cedirinib, among others.

In vivo, OCTN2 has not been reported as a target for drug–drug interactions to date. However, multiple drugs have been observed to cause carnitine deficiency through various mechanisms. Administration of the anticonvulsant valproic acid and the antibiotic pivalic acid causes reduced plasma carnitine levels and, in some cases, clinically relevant carnitine deficiency. Multiple mechanisms have been proposed. One possibility is that the valproate and pivalate directly inhibit the binding pocket for carnitine in OCTN2. Other studies have suggested that rather than inhibit OCTN2 directly, these drugs form carnitine conjugates and are likely effluxed out of the kidney with poor reabsorption, resulting in carnitine wasting and depletion. Alternatively, the

valproyl- and pivaloyl-carnitine esters could block reabsorption of free carnitine at OCTN2. Regardless of mechanism, these cases of drug-induced carnitine deficiency have been fatal in patients with carnitine transporter deficiency who already have reduced systemic carnitine levels.

In recent years, OCTN2 has become a target of drug delivery optimization strategies. Multiple properties make it an attractive drug target. First, it is theorized to increase oral bioavailability of targeted drugs due to high expression in the small intestine. Second, it has the potential to increase blood–brain barrier (BBB) permeability of substrates due to expression at the BBB. Third, it allows for the targeting of drugs to the kidney. And fourth, it has been hypothesized to increase delivery of asthma therapeutics to the lung (34). Multiple carnitine-conjugated prodrugs have been developed, including butyrate used in treatment for gut inflammation, nepotic acid used to treat seizures, and the chemotherapeutic drug, gemcitabine. Carnitine-conjugated gemcitabine exhibits 5-fold bioavailability over gemcitabine alone. In addition, nanoparticles are being explored for targeted delivery via OCTN2 for other cancer drugs like paclitaxel.

1.5.2 Regulation

Transcriptional regulation of OCTN2 is mediated in part by the peroxisome proliferator-activated receptor α (PPARα). PPARα plays an important role in the regulation of genes involved in lipid metabolism and energy homeo- stasis and is highly expressed in tissues that use fatty acid oxidation as a primary energy source, including heart muscle, skeletal muscle, and kidney (35). Notably, OCTN2 is expressed highly in these tissues as well. Tentative PPAR response elements (PPREs) are found in the promoter and intronic regions of *SLC22A5* in several species. In rats, treatment with the PPARα agonist clofibrate leads to increased transcription of OCTN2 in the liver and small intestine, but not the kidney or muscle tissue. Aligned with the upregulation of OCTN2 in rat liver, hepatic concentrations of carnitine are also increased by

PPARα activation. These findings are supported by the downregulation of OCTN2 and overall reduction in systemic carnitine levels in PPARα -null mice. Upregulation of OCTN2 by PPARα has also been demonstrated in pigs. In addition to fibrates, PPARα -mediated regulation of OCTN2 is affected by cisplatin. Cisplatin is hypothesized to inhibit DNA binding to the PPARα/RXR complex, resulting in an overall down- regulation of OTN2 and an increase in urinary carnitine wasting in mice (36). The PPARγ/RXRα complex also modulates OCTN2 expression in the large intestine. Human colonocytes and mouse colon exhibit altered expression of OCTN2 in response to PPARγ, but not PPARα (37). In a mouse model of IBD, proinflammatory cytokines interact with the PPARγ/RXRα complex to reduce OCTN2 expression, contributing to disease pathology. Treatment with the PPARγ agonist luteolin to rescue OCTN2 expression results in the reduction of colonic inflammation (38).

OCTN2 is upregulated by the estrogen receptor (ER) in breast cancer cells and tumor tissue, an effect attributed to the identification of a novel estrogen-responsive element (ERE) in an intronic region of *SLC22A5* (16).

OCTN2 is further regulated by PDZ domain-containing proteins (39). PDZK1 colocalizes with OCTN2 at the apical membrane of renal tubule cells. PDZK1 stimulates carnitine uptake via OCTN2 by increasing the V_{max} of the transporter, although cell-surface expression of OCTN2 is unchanged suggesting PDZK1 stimulates translocation of carnitine. The four terminal amino acids at the carboxyl end of OCTN2 serve as a PDZ binding motif, and deletion or substitution of these residues eliminates PDZK1 stimulation of OCTN2. PDZK2 also increases the transport capacity of OCTN2, but through a different mechanism, increasing localization of OCTN2 to the plasma membrane (39).

Lastly, OCTN2 expression is regulated by heat shock transcription factor 1 (HSF1) (16). The promoter variant −207G>C disrupts a consensus sequence for an HSF binding element. Cells with the −207G wild-type promoter containing the intact HSF1 binding site have higher expression of OCTN2 after heat-shock compared to cells with the −207C variant, which results in disrupted HSF1 binding.

1.5.3 Animal Models

Octn2$^{-/-}$ knockout mice have been characterized as a model of carnitine deficiency, resulting from a point mutation that causes a change from the amino acid leucine to arginine at residue 352. This substitution causes complete OCTN2 loss-of-function *in vitro* and *in vivo*. The mice, deemed juvenile visceral steatosis (jvs) mice, present with growth retardation and enlarged abdomen due to hepatic steatosis, as well as hyperammonia and hypoglycemia (40). Pharmacokinetics in jvs mice reveal drastically altered carnitine parameters, including reduced bioavailability, decreased volume of distribution, decreased tissue-to-plasma concentration ratios, and increased clearance of carnitine compared with wild-type mice (41). Furthermore, jvs mice display spontaneous intestinal apoptotic phenotypes including ulceration and gut perforation, and an immune response involving macrophage and lymphocyte infiltration (42). Inflammation and intestinal apoptosis are reduced when mice are treated with carnitine supplementation.

1.5.4 Human Genetic Studies

Biallelic loss-of-function mutations in OCTN2 result in a Mendelian disease known as carnitine transporter deficiency (CTD, also referred to as primary carnitine deficiency, systemic carnitine deficiency, or carnitine uptake defect; OMIM #212140 (43)). Almost 200 OCTN2 mutations have been identified among patients, the majority of which are missense, followed in frequency

by nonsense, frameshift, and noncoding mutations that affect splice sites or regulatory regions (44). Systemically, patients display extremely low plasma carnitine levels caused by reduced dietary carnitine absorption and excessive carnitine wasting in the urine due to loss of reabsorption by OCTN2 (45). The disorder varies in time to onset and disease severity and/or presentation, with common symptoms including cardiomyopathy, cardiac arrhythmias, hepatic encephalopathy, and hypoglycemia. Missed diagnosis can be fatal in infants, thus many developed countries screen infants for CTD, among other disorders, at birth. Diagnosis in individuals with less severe disease can be delayed well into adulthood, exemplified by the identification of maternal carnitine deficiency from low carnitine levels in the newborn during screening. CTD is treated with supplemental carnitine at high doses, up to 200 mg/kg multiple times per day, with decent outcomes. Lack of adherence to treatment has resulted in sudden death in at least one report. Incidence of CTD varies globally, affecting 1:120,000 individuals in Australia, 1:75,000 in the United States, 1:40,000 in Japan, 1:27,000 in China, and up to 1:300 individuals in the Faroe Islands (45). In at least one case, CTD has manifested as intellectual disability and autism spectrum disorder (46).

Extensive functional genomic studies have been conducted for OCTN2. The common promoter variant −207G>C is well characterized, known to decrease transcription of OCTN2 due to disruption of HSF1 transcription factor binding. Another variant in the 5′-UTR of the gene (−149G>A) creates an early ATG translation start site, reducing the translation of wild-type OCTN2 and decreasing carnitine transport (47). This variant has been found repeatedly in CTD patients for whom no or only one known deleterious variant is detected. Many studies have functionally characterized OCTN2 variants associated with CTD. Loss-of-function results from

multiple mechanisms, including altered kinetic parameters decreasing carnitine affinity or capacity for transport, decreased affinity for sodium, and reduced plasma membrane localization.

In GWAS, genetic polymorphisms in the OCTN2 locus have strong associations with blood metabolite levels (carnitine, acetylcarnitine, LDL cholesterol, and fatty acids), fat-free mass, and autoimmune disease (e.g., Crohn's disease). Forming a haplotype with OCTN1 at the IBD5 locus, OCTN2 is associated with Type 1 diabetes and Crohn's disease (16).

1.6 Conclusion

To understand physiologic and pharmacologic systems, it is critical to identify all of the components or proteins involved in those systems. The last few decades have ushered in a new understanding of the physiologic and pharmacologic roles of important zwitterions and organic cations, as the transporters involved in their absorption and disposition have been identified. It is now clear that transporters in the SLC22 family, along with a few other transporters, play key roles as determinants of systemic and tissue levels of cationic and zwitterionic drugs. At all levels from molecular to physiologic and pathophysiologic, there are major gaps in our knowledge. First and foremost, transporters for organic cations and zwitterions need to be discovered. Many transporters in the SLC superfamily, and in particular within the SLC22 family, remain orphans and need to be deorphaned. Further, no transporter in the SLC22 family has been crystallized; therefore, the precise molecular transport mechanism is not known. Moreover, though many associations have been reproducibly observed between genetic variants in organic cation and zwitterion transporters and various clinical phenotypes, the mechanisms by which the transporter contributes to the phenotypes remain poorly understood. Rare variants in the transporters such as in *SLC22A5* (OCTN2) are associated with fatal diseases, yet the function

of these variants remain unknown, and therapies remain poor at best. Finally, the physiologic, pharmacologic, and pathophysiologic systems that include these transporters need to be fully understood in order to obtain a full understanding of human biology and pharmacology.

1.7 Figures

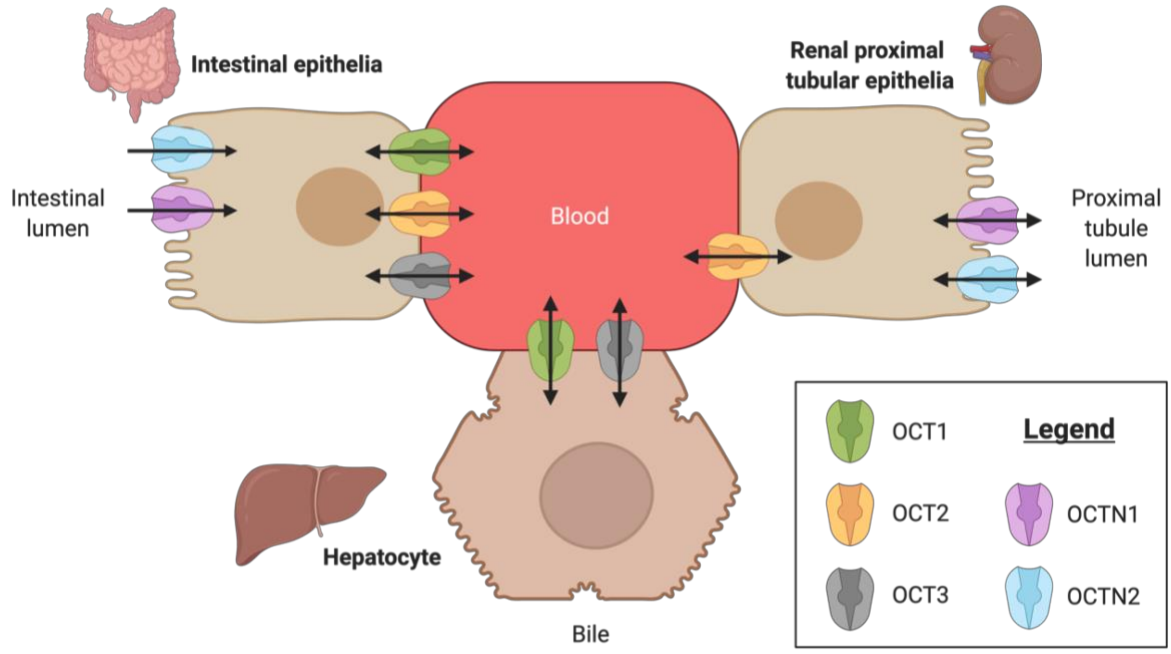


Figure 1.1. Tissue distribution and membrane localization of organic cation and zwitterion transporters. Created with BioRender.com.

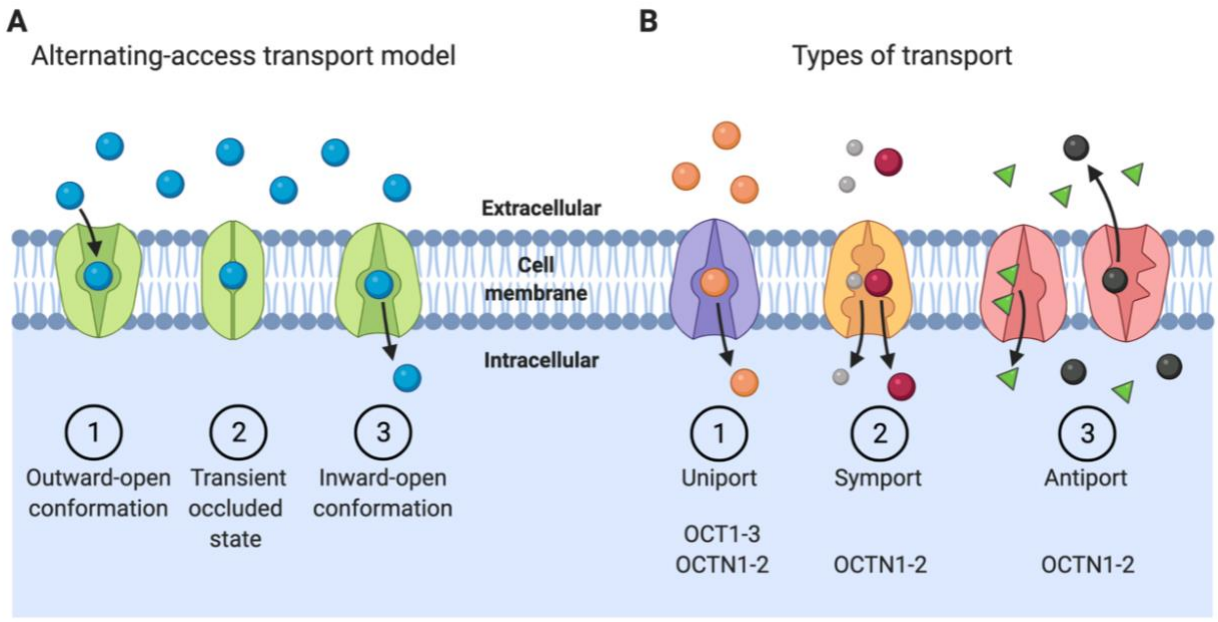


Figure 1.2. Transport of substrates by organic cation and zwitterion transporters. (A) Alternating-access transport model for translocation of substrates. **(B)** Types of transport mechanisms and transporters demonstrated to function by each mechanism. Created with BioRender.com.

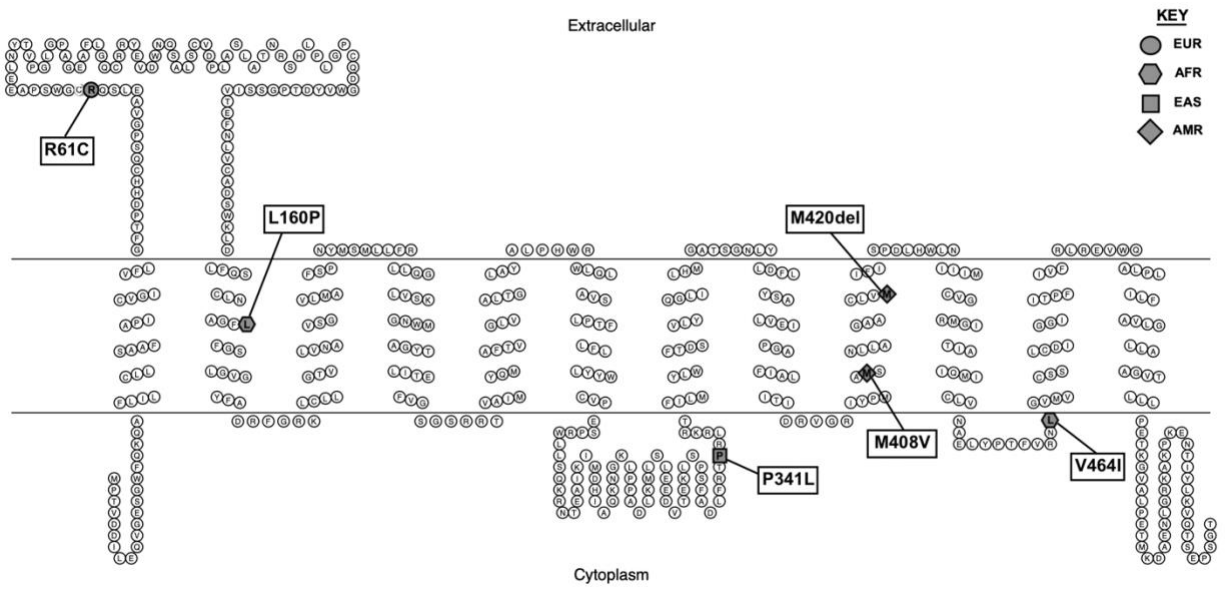


Figure 1.3. Predicted secondary structure of OCT1 with most common missense variants highlighted. Common defined as Global Allele Frequency (GAF) > 0.02. The group with the highest allele frequency for each variant is shown. Created with TOPO2.

1.8 Tables

Table 1.1. Selected substrates and inhibitors of the major organic cation and zwitterion transporters, OCT1-3 and OCTN1-2.

Transporter	Model Substrates	Substrates	Model Inhibitors	Inhibitors
OCT1	MPP^+, TEA, ASP^+, metformin	Endogenous: serotonin, acylcarnitines, choline, acetylcholine, creatinine, agmatine Exogenous: acyclovir, quinidine, quinine, thiamine, sumatriptain, ondansetron, morphine	Quinidine, verpamil	Exogenous: Atropine, abacavir, zidovudine, tenofovir, spironolactone, ondansetron, quinine, midazolam
OCT2	TEA, MPP^+, ASP^+, NBD-MTMA, metformin	Endogenous: creatinine, choline, serotonin, dopamine, histamine Exogenous: amphetamine, cisplatin, cimetidine, phenformin	quinidine, cimetidine	Endogenous: testosterone Exogenous: doxepin, zolpidem, ritonavir, imipramine, tramadol, tacrine, olanzapine
OCT3	MPP^+, ASP^+, metformin	Endogenous: Creatinine, agmatine, dopamine, progesterone, testosterone Exogenous: atropine, prazosin, cimetidine, verapamil, nicotine	corticosterone	Endogenous: progesterone, B-estradiol, corticosterone Exogenous: verapamil, carvedilol, imipramine, cimetidine, metformin
OCTN1	L-ergothioneine, TEA	Endogenous: L-ergothioneine, L-carnitine, acetylcholine Exogenous: cytarabine, amisulpride, ethambutol, ipratropium, gapapentin	TEA	Endogenous: L-carnitine, acetylcarnitine, choline, acetylcholine, gamma-butyrobetaine Exogenous: carvedilol, flecainide, lidocaine, verapamil, mitoxantrone, dipyridamole, doxorubicin
OCTN2	L-carnitine	Endogenous: L-carnitine, acetyl-L-carnitine, choline Exogenous: D-carnitine, mildronate, ipratropium, etoposide, amisulpride	TEA, verapamil	Endogenous: L-carnitine, acetylcarnitine Exogenous: clozapine, emetine, vinblastine, omeprazole, verapamil, β-lactam antibiotics

Abbreviations:
MPP^+: N-methyl-4-phenylpyridinium.
TEA: tetraethylammonium.
ASP^+: 4-(4-(diethylamino)styryl)-*N*-methylpyridinium.
NBD-MTMA: N,N,N-trimethyl-2-[methyl(7-nitrobenzo[c][l,2,5]oxadiazol-4-yl)amino]ethanaminium.

1.9 References

1. Koepsell H, Lips K, Volk C. Polyspecific organic cation transporters: structure, function, physiological roles, and biopharmaceutical implications. Pharm Res. 2007;24(7):1227-51.
2. Motohashi H, Sakurai Y, Saito H, Masuda S, Urakami Y, Goto M, et al. Gene expression levels and immunolocalization of organic ion transporters in the human kidney. J Am Soc Nephrol. 2002;13(4):866-74.
3. Wagner DJ, Hu T, Wang J. Polyspecific organic cation transporters and their impact on drug intracellular levels and pharmacodynamics. Pharmacol Res. 2016;111:237-46.
4. Dakal TC, Kumar R, Ramotar D. Structural modeling of human organic cation transporters. Comput Biol Chem. 2017;68:153-63.
5. Koepsell H. Substrate recognition and translocation by polyspecific organic cation transporters. Biol Chem. 2011;392(1-2):95-101.
6. Volk C. OCTs, OATs, and OCTNs: structure and function of the polyspecific organic ion transporters of the SLC22 family. Wiley Interdisciplinary Reviews: Membrane Transport and Signaling. 2014;3(1):1-13.
7. Harper JN, Wright SH. Multiple mechanisms of ligand interaction with the human organic cation transporter, OCT2. Am J Physiol Renal Physiol. 2013;304(1):F56-67.
8. Egenberger B, Gorboulev V, Keller T, Gorbunov D, Gottlieb N, Geiger D, et al. A substrate binding hinge domain is critical for transport-related structural changes of organic cation transporter 1. J Biol Chem. 2012;287(37):31561-73.
9. Chien HC, Zur AA, Maurer TS, Yee SW, Tolsma J, Jasper P, et al. Rapid Method To Determine Intracellular Drug Concentrations in Cellular Uptake Assays: Application to

Metformin in Organic Cation Transporter 1-Transfected Human Embryonic Kidney 293 Cells. Drug Metab Dispos. 2016;44(3):356-64.

10. Morrissey KM, Wen CC, Johns SJ, Zhang L, Huang SM, Giacomini KM. The UCSF-FDA TransPortal: a public drug transporter database. Clin Pharmacol Ther. 2012;92(5):545-6.

11. Chen L, Shu Y, Liang X, Chen EC, Yee SW, Zur AA, et al. OCT1 is a high-capacity thiamine transporter that regulates hepatic steatosis and is a target of metformin. Proc Natl Acad Sci U S A. 2014;111(27):9983-8.

12. Dresser MJ, Gray AT, Giacomini KM. Kinetic and selectivity differences between rodent, rabbit, and human organic cation transporters (OCT1). J Pharmacol Exp Ther. 2000;292(3):1146-52.

13. Kim HI, Raffler J, Lu W, Lee JJ, Abbey D, Saleheen D, et al. Fine Mapping and Functional Analysis Reveal a Role of SLC22A1 in Acylcarnitine Transport. Am J Hum Genet. 2017;101(4):489-502.

14. Hyrsova L, Smutny T, Carazo A, Moravcik S, Mandikova J, Trejtnar F, et al. The pregnane X receptor down-regulates organic cation transporter 1 (SLC22A1) in human hepatocytes by competing for ("squelching") SRC-1 coactivator. Br J Pharmacol. 2016;173(10):1703-15.

15. Hyrsova L, Smutny T, Trejtnar F, Pavek P. Expression of organic cation transporter 1 (OCT1): unique patterns of indirect regulation by nuclear receptors and hepatospecific gene regulation. Drug Metab Rev. 2016;48(2):139-58.

16. Koepsell H. Organic Cation Transporters in Health and Disease. Pharmacol Rev. 2020;72(1):253-319.

17. Ciarimboli G, Struwe K, Arndt P, Gorboulev V, Koepsell H, Schlatter E, et al. Regulation of the human organic cation transporter hOCT1. J Cell Physiol. 2004;201(3):420-8.

18. Jonker JW, Wagenaar E, Mol CA, Buitelaar M, Koepsell H, Smit JW, et al. Reduced hepatic uptake and intestinal excretion of organic cations in mice with a targeted disruption of the organic cation transporter 1 (Oct1 [Slc22a1]) gene. Mol Cell Biol. 2001;21(16):5471-7.

19. Wang DS, Jonker JW, Kato Y, Kusuhara H, Schinkel AH, Sugiyama Y. Involvement of organic cation transporter 1 in hepatic and intestinal distribution of metformin. J Pharmacol Exp Ther. 2002;302(2):510-5.

20. Wang DS, Kusuhara H, Kato Y, Jonker JW, Schinkel AH, Sugiyama Y. Involvement of organic cation transporter 1 in the lactic acidosis caused by metformin. Mol Pharmacol. 2003;63(4):844-8.

21. Morse B, Kolur A, Hudson L, Hogan A, Chen L, Brackman R, et al. Pharmacokinetics of Organic Cation Transporter 1 (OCT1) Substrates in Oct1/2 Knockout Mice and Species Difference in Hepatic OCT1-Mediated Uptake. Drug Metab Dispos. 2020;48(2):93-105.

22. Kerb R, Brinkmann U, Chatskaia N, Gorbunov D, Gorboulev V, Mornhinweg E, et al. Identification of genetic variations of the human organic cation transporter hOCT1 and their functional consequences. Pharmacogenetics. 2002;12(8):591-5.

23. Shu Y, Leabman MK, Feng B, Mangravite LM, Huang CC, Stryke D, et al. Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1. Proc Natl Acad Sci U S A. 2003;100(10):5902-7.

24. Arimany-Nardi C, Koepsell H, Pastor-Anglada M. Role of SLC22A1 polymorphic variants in drug disposition, therapeutic responses, and drug-drug interactions. Pharmacogenomics J. 2015;15(6):473-87.

25. Shu Y, Brown C, Castro RA, Shi RJ, Lin ET, Owen RP, et al. Effect of genetic variation in the organic cation transporter 1, OCT1, on metformin pharmacokinetics. Clin Pharmacol Ther. 2008;83(2):273-80.

26. Chen L, Takizawa M, Chen E, Schlessinger A, Segenthelar J, Choi JH, et al. Genetic polymorphisms in organic cation transporter 1 (OCT1) in Chinese and Japanese populations exhibit altered function. J Pharmacol Exp Ther. 2010;335(1):42-50.

27. Yee SW, Brackman DJ, Ennis EA, Sugiyama Y, Kamdem LK, Blanchard R, et al. Influence of Transporter Polymorphisms on Drug Disposition and Response: A Perspective From the International Transporter Consortium. Clin Pharmacol Ther. 2018;104(5):803-17.

28. Tzvetkov MV, Saadatmand AR, Bokelmann K, Meineke I, Kaiser R, Brockmöller J. Effects of OCT1 polymorphisms on the cellular uptake, plasma concentrations and efficacy of the 5-HT(3) antagonists tropisetron and ondansetron. Pharmacogenomics J. 2012;12(1):22-9.

29. Liang X, Yee SW, Chien HC, Chen EC, Luo Q, Zou L, et al. Organic cation transporter 1 (OCT1) modulates multiple cardiometabolic traits through effects on hepatic thiamine content. PLoS Biol. 2018;16(4):e2002907.

30. Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, Chu X, et al. Membrane transporters in drug development. Nat Rev Drug Discov. 2010;9(3):215-36.

31. Giacomini KM, Galetin A, Huang SM. The International Transporter Consortium: Summarizing Advances in the Role of Transporters in Drug Development. Clin Pharmacol Ther. 2018;104(5):766-71.

32. Chu X, Chan GH, Evers R. Identification of Endogenous Biomarkers to Predict the Propensity of Drug Candidates to Cause Hepatic or Renal Transporter-Mediated Drug-Drug Interactions. J Pharm Sci. 2017;106(9):2357-67.

33. Koepsell H. The SLC22 family with transporters of organic cations, anions and zwitterions. Mol Aspects Med. 2013;34(2-3):413-35.

34. Kou L, Sun R, Ganapathy V, Yao Q, Chen R. Recent advances in drug delivery via the organic cation/carnitine transporter 2 (OCTN2/SLC22A5). Expert Opin Ther Targets. 2018;22(8):715-26.

35. Eder K, Ringseis R. The role of peroxisome proliferator-activated receptor alpha in transcriptional regulation of novel organic cation transporters. Eur J Pharmacol. 2010;628(1-3):1-5.

36. Lancaster CS, Hu C, Franke RM, Filipski KK, Orwick SJ, Chen Z, et al. Cisplatin-induced downregulation of OCTN2 affects carnitine wasting. Clin Cancer Res. 2010;16(19):4789-99.

37. D'Argenio G, Petillo O, Margarucci S, Torpedine A, Calarco A, Koverech A, et al. Colon OCTN2 gene expression is up-regulated by peroxisome proliferator-activated receptor gamma in humans and mice and contributes to local and systemic carnitine homeostasis. J Biol Chem. 2010;285(35):27078-87.

38. Li P, Wang Y, Luo J, Zeng Q, Wang M, Bai M, et al. Downregulation of OCTN2 by cytokines plays an important role in the progression of inflammatory bowel disease. Biochem Pharmacol. 2020;178:114115.

39. Kato Y, Sai Y, Yoshida K, Watanabe C, Hirata T, Tsuji A. PDZK1 directly regulates the function of organic cation/carnitine transporter OCTN2. Mol Pharmacol. 2005;67(3):734-43.

40. Koizumi T, Nikaido H, Hayakawa J, Nonomura A, Yoneda T. Infantile disease with microvesicular fatty infiltration of viscera spontaneously occurring in the C3H-H-2(0) strain of mouse with similarities to Reye's syndrome. Lab Anim. 1988;22(1):83-7.

41. Yokogawa K, Higashi Y, Tamai I, Nomura M, Hashimoto N, Nikaido H, et al. Decreased tissue distribution of L-carnitine in juvenile visceral steatosis mice. J Pharmacol Exp Ther. 1999;289(1):224-30.

42. Shekhawat PS, Srinivas SR, Matern D, Bennett MJ, Boriack R, George V, et al. Spontaneous development of intestinal and colonic atrophy and inflammation in the carnitine-deficient jvs (OCTN2(-/-)) mice. Mol Genet Metab. 2007;92(4):315-24.

43. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7.

44. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

45. Longo N. Primary Carnitine Deficiency and Newborn Screening for Disorders of the Carnitine Cycle. Ann Nutr Metab. 2016;68 Suppl 3:5-9.

46. Guevara-Campos J, González-Guevara L, Guevara-González J, Cauli O. First Case Report of Primary Carnitine Deficiency Manifested as Intellectual Disability and Autism Spectrum Disorder. Brain Sci. 2019;9(6).

47. Ferdinandusse S, Te Brinke H, Ruiter JPN, Haasjes J, Oostheim W, van Lenthe H, et al. A mutation creating an upstream translation initiation codon in SLC22A5 5'UTR is a frequent cause of primary carnitine deficiency. Hum Mutat. 2019;40(10):1899-904.

Chapter 2: Opportunities and challenges for the computational interpretation of rare variation in clinically important genes

2.1 Abstract

Genome sequencing is enabling precision medicine—tailoring treatment to the unique constellation of variants in an individual's genome. The impact of recurrent pathogenic variants is often understood, however there is a long tail of rare genetic variants that are uncharacterized. The problem of uncharacterized rare variation is especially acute when it occurs in genes of known clinical importance with functionally consequential variants and associated mechanisms. Variants of uncertain significance (VUSs) in these genes are discovered at a rate that outpaces current ability to classify them with databases of previous cases, experimental evaluation, and computational predictors. Clinicians are thus left without guidance about the significance of variants that may have actionable consequences. Computational prediction of the impact of rare genetic variation is increasingly becoming an important capability. In this paper, we review the technical and ethical challenges of interpreting the function of rare variants in two settings: inborn errors of metabolism in newborns and pharmacogenomics. We propose a framework for a genomic learning healthcare system with an initial focus on early-onset treatable disease in newborns and actionable pharmacogenomics. We argue that (1) a genomic learning healthcare system must allow for continuous collection and assessment of rare variants, (2) emerging machine learning methods will enable algorithms to predict the clinical impact of rare variants on protein function, and (3) ethical considerations must inform the construction and deployment of all rare-variation triage strategies, particularly with respect to health disparities arising from unbalanced ancestry representation.

2.2 Introduction

We are approaching an era in which genome sequencing at birth may become a widespread practice with the potential to revolutionize healthcare. Interpretation of the genetic variants identified by sequencing, however, remains a significant challenge and limits the use of DNA sequencing as a primary diagnostic screen (1). Current algorithms used to interpret the significance of genetic mutations are not reliable enough to be used without additional clinical data (2). Yet, accumulating biomedical data enables machine learning algorithms to predict the consequence of genetic variants with increasing accuracy. The pairing of modern algorithms and widespread genome sequencing is beginning to deliver precision medicine in limited settings (3) but the broad interpretation of rare genetic variation requires both algorithmic advances and improved access to data. The identification of rare variation responsible for unusual clinical phenotypes is a particularly difficult challenge because both the responsible gene and the associated variation must be identified. A slightly more tractable problem is the identification of clinically important variants in genes that are already known to be clinically significant and have known mechanisms for influencing phenotype.

This paper focuses on two clinical domains that have known clinically important genes and in the near term should benefit greatly from improved rare variant interpretation: pharmacogenomics (PGx) and inborn errors of metabolism (IEMs). IEMs and PGx are examples of genetic practice characterized by monogenic phenotypes for which therapeutic action can be taken in response to clinically important variants in known genes. Both fields have been revolutionized by low-cost sequencing and the curation of large databases cataloging the effects of specific genetic variants. Furthermore, both fields struggle with interpretation of the phenotypic effects of rare variants that have not been clinically evaluated.

As an interdisciplinary team supported by the Chan Zuckerberg Biohub, we approach these two challenges by addressing both computational and ethical issues in order to develop a framework for genome-informed medical care that benefits all. Here, we review the current practices and limitations of variant interpretation in PGx and IEMs and highlight recent computational advances that will allow researchers to improve precision medicine. Ethical considerations include health disparities because existing genetic and genomic databases are not inclusive of individuals of diverse ancestries. As the recent strategic vision from the US National Human Genomic Research Institute (NHGRI) attests, there are significant societal implications of a genomic learning healthcare system that we cannot afford to oversimplify (4). Our focus on genes of known consequence should generalize ultimately to the more difficult cases where the gene, function, and mechanism are not well understood.

2.3 PGx and IEMs in current clinical practice

For both PGx and IEMs, our detailed understanding of the biological processes at play (the genes that are critical and how they interact) has reached a point at which routine genetic screens can inform clinical decision-making. In the United States, PGx testing is mandated by the Food and Drug Administration for a number of drugs because of safety concerns and is recommended for many others. Testing for IEMs is routine practice for nearly all newborns in the United States, but the role of genetic testing is largely limited to second-tier screens and carrier testing. These two clinical domains are linked in more ways than it may superficially appear. The clinical implications for most known PGx- and IEM-driven phenotypes are often caused by variants in a single gene. As monogenic traits, there is not only a critical importance in understanding the impact of variants in the underlying genes but also in narrowing the problem space for a tractable

solution. Additionally, the mechanisms of disease and treatment response are generally understood.

PGx describes how an individual's response to medication is influenced by genetic variation in pharmacogenes: genes encoding proteins involved in the pharmacokinetics and pharmacodynamics of a drug (5). Many pharmacogenes have common genetic variants with known clinical significance. These variants can affect the metabolism, transport, and action of drugs throughout the body and may influence efficacy or lead to adverse events. Studies have shown as many as 99.8% of individuals carry at least one genetic variant that could lead to adverse outcomes for at least one drug (6-8). In the past, clinical practice overlooked the influence of genetics on drug response and—except for several extreme case (9)—used a standardized dose of any particular drug for most patients, with some trial-and-adjustment to determine the ideal drug and dosage. This error-prone process can lead to decreased efficacy and increased incidence of adverse events that could be otherwise avoided (10). Clinical practice may be moving toward genetic testing prior to drug dosing, although at present, current practice is still limited to physician-guided treatment: genotyping or sequencing is ordered by a physician and carried out clinically (Fig. 2.1A). To date, there are 60 drugs with clinical dosing guidelines published by the Clinical Pharmacogenomics Implementation Consortium (CPIC) and 94 drugs with guidelines from the Dutch Pharmacogenomics Working Group (DPWG) (10). As the inexpensive interrogation of genetic information gains a foothold in clinical medicine, pharmacogenetic information will increasingly become standard care. Importantly, when genetic information is used to guide dosing, the current focus is on common polymorphisms in individuals of European ancestry. Common polymorphisms in other ancestral groups and rare variants are generally not included in current clinical dosing guidelines. This can lead to health

disparities based on a patient's ancestry and is problematic for all individuals because rare variants are estimated to contribute to as much as 50% of interindividual variation in drug response (11).

IEMs encompass more than 1,000 genetic disorders, including organic acidemias, urea cycle defects, lysosomal storage disorders, and disorders of amino acid metabolism (12). IEMs are characterized by monogenic mutations that can affect protein function and result in altered metabolite levels. The majority are autosomal recessive disorders. Many IEMs are severe, early-onset conditions amenable to therapeutic intervention, and early treatment can lead to significantly improved clinical outcomes. Because the consequences of unrecognized IEMs in pre-symptomatic newborns can be catastrophic, detection before symptom manifestation is essential. Newborn screening (NBS), a near-universal public health practice, detects over 40 of the most common, treatable IEMs via biochemical tests performed in blood samples taken shortly after birth. IEMs occur in ~1 in 2,000 births worldwide and are present in all ancestral groups (13). Comparing incidence across ancestry is difficult because of differences in screening between countries and the fact that ancestry is not consistently categorized within countries (14). One study of ancestrally diverse California newborns suggested that newborns with Middle Eastern ancestry had the highest incidence of IEMs (>1 in 1,000) and newborns with Japanese or Pacific Island ancestry had the lowest incidence of IEMs (<1 in 5,000) (15).

Presently, NBS detects IEMs by identifying elevated metabolites in blood, which is performed with tandem mass spectrometry (MS/MS), an inexpensive and rapid test. However, disorders may be missed, some analytes are non-specific, and follow-up testing may be time consuming and complex (1, 16). DNA sequencing has the potential to more accurately identify disorders for

which MS/MS detection is not optimal and also identify disorders for which there is no appropriate metabolite screen.

Carrier testing provides an opportunity to detect rare variants in IEMs and other disease-associated genes (17) before conception. However, interpretation of genetic screening results still faces significant challenges (18), especially in cases identifying variants of uncertain significance (VUSs) where risk for inherited disease cannot be definitively assessed and actionability is questionable. The falling cost of next-generation sequencing will continue to expand the identification of genomic variants that may cause IEMs or alter drug response. Although many genetic variants have established associations with disease phenotypes or drug response, the majority are of unknown clinical consequence. Generating experimental data to validate the pathogenicity of individual variants is tedious and expensive, although recent advances have facilitated more large-scale generation of data (19). Several databases attempt to catalog variants in disease-causing genes, but there is no central catalog for associated functional data. Thus, alternative methods for determining or predicting functional effects of genetic variants are urgently needed.

At present, validation of genetic variants as causal for IEMs or important for PGx is complex, involving consideration of layers of information at the genetic, phenotypic, clinical, and familial levels (20). Variants in genes underlying IEMs frequently require functional characterization to be validated as causal. Functional validation can be carried out with a myriad of model systems, including patient-derived cells or blood, immortalized cell lines, and animal models (21). Robust functional assays suitable for the validation of variants as causal are not always available because they require a biological or biochemical measurement directly related to the function of the gene of interest. Common experimental methods to validate pathogenicity include overexpression

models to assess function of the variant allele, genetic rescue whereby introduction of the wild-type allele rescues phenotype, and transgenic expression for phenotyping in model organisms such as *E. coli*, yeast, *Drosophila*, *C. elegans*, zebrafish, and mice (21). CRISPR-Cas9 technology allows for high-throughput functional characterization in many systems. Assays investigating mRNA and protein expression (i.e., RNA sequencing [RNA-seq] and immunoblot) can reveal variant consequences on splicing and allele expression or differential protein expression, respectively (21). The validation of clinically important variants relating to PGx is also complex. Targeted functional assays evaluating variant effects on gene function can be carried out *in vitro* when feasible via similar methods and models as for IEMs. Examples include enzyme activity assays (22) and transporter uptake assays (23). Pharmacogenetic variation can further be validated as clinically important in pharmacokinetic/pharmacodynamic studies, whereby individuals with a particular genotype exhibit significantly different drug response compared with individuals with a different genotype for the variant in question.

2.4 Ethical considerations in rare variant interpretation

Genome-informed precision medicine must include analysis of ethical, legal, and social implications (ELSIs) in order to improve upon rather than exacerbate existing health disparities (4). We have identified six chief concerns with enhancing computational predictors for the phenotypic effects of rare variation at the scale proposed here. First, the uncertainty of results and, second, the return of clinical results can either improve or compromise clinical care. Although enhanced computational predictors for IEMs and PGx can minimize harm from the trial and error of current clinical practices, consistency in clinical education and approaches to ambiguous and incidental findings will be critical to determining societal benefit. Third, research and clinical stakeholder perspectives in approaching the classifications of VUSs can differ.

Fourth, the underrepresentation of minority groups in current datasets and the underlying research that informs them needs particular attention in order to create a larger and more diverse reference genome so that biases can be reduced. Fifth, an effective genomic learning healthcare system must account for data security and privacy risks. Sixth, there needs to be transparent data sharing expectations across all levels of participation in the learning system. Building on previous ethical frameworks (24, 25) and the need for a nuanced approach (26), we suggest that trade-offs between ensuring individual control over data and the social obligations of individuals have yet to be resolved at the level of ethical governance provisions. Discussion of these concerns is guided by three central ethical questions, summarized in Table 2.1 and elaborated within the ethics spotlight sections.

2.4.1 Ethics spotlight 1: Can genome sequencing improve the uncertainty of results and return of clinical results?

For the use of predictive algorithms as the primary methods of analysis for IEMs and PGx to be ethically justified, these methods must provide equal or greater certainty than current methods. Improving screening and predictive analysis for IEMs and PGx at the testing level is contingent upon the accuracy of results, the provisions around returning results, and the impact on clinical care. Even pathogenic results can have variable penetrance and/or VUSs and, given the possibility of reclassification over time, can cause significant consternation on both the part of the clinician and patient (27). Perhaps most thoroughly documented in cancer genetics (28), the clinical return of genetic results is rarely straightforward. The prohibition against the return of uncertain results, outlined by the American College of Medical Genetics and Genomics (ACMG), is such that even if there is a suspicion that an uncertain variant is pathogenic, it

should conservatively be classified as a VUS because this information is used in medical decisions (2).

The follow-up of uncertain results is complicated by clinician/researcher and patient expectations and understandings of actionability. Genomic literacy across different healthcare professional roles is limited (14, 29). The disclosing of sequencing results should be contingent upon what has been previously explained to the patient/parent about incidental findings and potential treatments (30). As healthcare delivery is already biased with regard to decisions about referrals or withdrawals of care, including decisions made through racial discrimination, it will be challenging for algorithms to correct for existing biases in the handling of results (31). Uncertain and incidental (or secondary) results in clinical care should be considered in the context of existing slippages of fiduciary obligations—such as clinician biases and/or patient mistrust—that emerging tests may or may not be able to compensate for (32). The NHGRI has called for greater diversity among the genomic scientist workforce (4).

In order to contain immediate risks around uncertainty of results and focus resources, is there a case for tiered approaches? For example, beginning with targeted sequencing and, upon accuracy improvements, expanding programs to include non-targeted sequencing, or at the individual level, only sequencing specific genes as a second-tier option if a positive test result arises in genome sequencing? Certainly, implementing genome sequencing at the routine screening level requires greater computational accuracy, accessibility, and more nuanced ethical safeguards (4, 26). In the US healthcare context, it is difficult to resolve the issue of healthcare insurance coverage. Can financial disparity in the follow-up of results be partially alleviated with temporary coverage through risk-sharing agreements between payers and manufacturers of tests (33)? Can ethical priorities of the clinician and patient transaction be made compatible with the

needs of the genomic learning healthcare system—which must maximize scarce resources—such that genomic sequencing improves healthcare across all of society?

2.5 Evaluating variants of uncertain significance

Variants in functionally important genes are often suspected to lead to clinical consequences. For IEMs and PGx, there are hundreds of genes in which nonsense and missense variants are associated with clinical outcomes. Although additional genetic, epigenetic, and environmental factors alter disease risk and drug response, the gene sequence is the primary determinant of phenotype for these genes. Thousands of pathogenic rare variants in these genes have been characterized with clinical consequences often well understood and cataloged. Yet exome and genome sequencing continue to identify novel variants in these genes at a rapid pace. The ACMG has developed guidelines to interpret these variants, but by design, conclusive evidence is required to assert a variant is pathogenic, even in known disease genes (2). For example, defects in *PAH* (OMIM: 612349) cause phenylketonuria (PKU [OMIM: 261600]), an IEM that can lead to severe intellectual disability and seizures when untreated. In gnomAD (34), a population database of variants seen in more than 100,000 individuals, 57% of observed protein-altering variants in *PAH* have unknown pathogenicity. Individuals who are homozygous for these variants at birth will have an unknown risk of developing PKU, and carriers of these variants cannot be advised of their risk of having a child with PKU. Thus, predicting the functional consequence of rare variants in IEMs and PGx is an important challenge.

To begin to address this issue, numerous publicly available databases actively catalog genetic variants and associated disease and drug response phenotypes. These databases are typically human curated and bring together information that would otherwise be dispersed across the

literature, allowing researchers and clinicians to quickly access existing knowledge. Several databases focus on the pathogenicity of variants genome wide, including thousands of variants in IEM and PGx genes. These include ClinVar, ClinGen, the Human Gene Mutation Database (HGMD), and Online Mendelian Inheritance in Man (OMIM) (35-37). Such platforms have a shared goal of linking genes with disease, although they take different approaches. ClinVar allows submissions from clinical laboratories, research groups, and specialized databases, presenting all submitted data through an online interface. Most submissions are not manually vetted and are presented as submitted. ClinGen and OMIM attempt to provide authoritative curation of known variants and their relationship to disease. Curators review literature and experimental data to determine pathogenicity of genetic variants. ClinVar and ClinGen share and collaboratively curate data. In addition to being used for standardizing the set of variants with known consequences, these databases are also used by researchers and clinicians to evaluate the evidence that an uncatalogued VUS causes disease based on its similarity to cataloged variants (e.g., if a VUS results in the same amino acid change as a cataloged pathogenic variant, this VUS now has strong evidence for being pathogenic) (2). Similarly, efforts have been made to catalog the relationship between genetic variation and drug response, exemplified by databases including PharmVar and PharmGKB (38-40). Like ClinVar, PharmVar relies on user submissions of discovered haplotypes in genes related to pharmacogenomics.

These variant databases encapsulate the combined expertise of thousands of clinical researchers across the world but also reveal a large amount of uncertainty. The majority of possible missense variants in IEM and PGx genes are classified as VUSs or are altogether missing from databases. ClinVar alone contains more than 6,000 variants classified as VUSs in IEM genes and more than 10,000 VUSs in PGx genes (Fig. 2.2A-B). Variants in ClinVar change classification as

researchers submit new evidence, but very few VUSs are resolved as fully pathogenic or benign (Fig. 2.2C-D). Instead, many variants are subject to conflicting classifications. Indeed, 41% of IEM and PGx variants in ClinVar are of uncertain significance or have conflicting interpretations of clinical importance. For novel variants, it is often challenging to establish pathogenic certainty until they are observed by multiple clinicians who submit consistent classifications to a variant database. For VUSs without further clinical or experimental evidence, computational methods offer a possible resolution.

Most computational approaches predict the functional impact of single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs) by using predictive machine learning models. The popular tool CADD uses a logistic regression model and more than 60 genomic features to learn the features that distinguish randomly generated variants from recently fixed variants in humans (41). The resulting predictor has been used to predict the pathogenicity of clinical variants and is currently used in clinical analysis pipelines (42). REVEL, a meta-predictor, uses the ensemble of scores from several prediction algorithms like CADD, each with different strengths and weaknesses, and is trained to differentiate rare unlabeled variants from HGMD pathogenic variants (43). Both CADD and REVEL are capable of predicting the effects of variants in any gene, which is typical of predictors used in clinical research. However, predictors that are gene-, gene family-, or locus-specific generally perform better for both IEMs and PGx in comparison to predictors that rely on data from the entire genome (44-51). Despite their promise, such bespoke methods are constrained by the limited data available for most genes, such as the number of known pathogenic variants and associated functional data. Because these methods are designed to predict the functional impact of a variant, their predictions can be some layers removed from the clinical consequence. Additionally,

pharmacogenes are not under the same evolutionary constraint as genes involved in disease, limiting the effectiveness of most predictive algorithms (46, 52).

To combine the best features of variant databases and computational predictors, automated systems that use both in tandem are already being tested to predict the pathogenicity of rare variants. Consider one recent study evaluating IEM detection by sequencing dried blood spots (DBSs) obtained from newborns (1). This study compared the performance of MS/MS to exome sequencing as a primary screen for IEMs on a set of 805 newborns with confirmed IEMs. Variants identified by sequencing were automatically assessed on rarity, protein consequence, and predicted pathogenicity (including CADD) and matched with cataloged pathogenic variants in ClinVar and HGMD to predict disease status. Overall, this combination was neither sufficiently sensitive nor specific compared to MS/MS, and exome sequencing notably missed a number of cases in which a pair of rare, protein-altering variants were absent from the causal gene. However, performance varied among IEMs and, in some cases, provided more specific diagnoses than conventional MS/MS analyte testing. 32% of pathogenic variants were absent from HGMD and ClinVar. Critically, sequencing led to several false positives in which an individual harbored a pair of rare, protein-altering variants in an IEM gene but did not have the associated disorder. These false positives significantly limit the ability to use DNA sequencing for screening and could be mitigated by more accurate computational methods that distinguish pathogenic from benign protein-altering variants.

2.5.1 Ethics spotlight 2: Can we view the classification of VUSs as a social justice opportunity?

Whether the classification of VUSs and IEMs can offer a fairer distribution of the benefits of sequencing technologies across all population groups is a significant question. Most large

datasets in the US contain homogeneous ancestry that is unrepresentative of the whole population (53, 54). In addition to the need to improve predictive methods for IEMs, screened individuals need to be considered as part of a social group in relationship to a wider and unequal social system. The moral obligations embedded within the ethics of clinical research and practice need to be better integrated (24). For individuals seeking healthcare, polygenic risk scores are more accurate for patients of European ancestry because the data from which algorithms are trained are derived largely from individuals of European ancestry (55, 56). Similarly, variant impact predictors tend to be derived from cataloged variants in databases, which are not representative of all ancestries. For example, ClinVar was recently found to be missing a large number of hearing impairment variants that primarily affect individuals of African ancestry (57), most likely indicative of a broader pattern. For variant predictors, this bias will lead to greater reliance on European ancestry variants and European genetic context, producing less accurate classification of IEM and PGx variants in other ancestral populations (e.g., African), which would only compound existing injustice in healthcare access for underrepresented populations (58, 59). Disparity in ancestry representation is especially stark in data sources for genome-wide association studies, where European ancestry disproportionately represents 81% of the dataset population (53).

Can we alleviate healthcare disparity by closing current ancestry gaps in genetics research? Given evidence that polygenic risk scores can be improved upon by incorporating datasets from a broader range of genetic ancestries (60), it is imperative that the genetics field strives for fairer training data. As the field matures to consider the role of genetic modifiers (61), as well as social and environmental interactions (62), genotypes of diverse individuals are needed to consider the effects of genetic modifiers and the environment on variants. Newborn screening programs, with

their mandatory collection and the near universal application of testing, provide a diverse and truly representative set of individuals (15). That said, racial discrimination in healthcare and healthcare research is not simply resolvable through technical fixes. Redressing data underrepresentation and health equity in machine learning precision medicine must be viewed in the context of governance and broader social change, which we discuss in "Ethics Spotlight 3," regarding questions of social obligation.

2.6 Opportunities in rare variant evaluation

In predicting the effect of a variant on gene function, we can predict its effects on the system, such as a metabolic pathway, and then on the physiology and/or pathophysiology. Cataloging observed likely clinically impactful variants in databases such as ClinVar and PharmVar (37) can be effective for determining the pathogenicity of more frequent rare variants (allele frequency between 0.01% and 1%). These variants are common enough that they have been identified in multiple individuals, and therefore, the effect on phenotype can be verified. However, ultra-rare variants, defined as having an allele frequency less than 0.01%, are responsible for a large portion of rare genetic disorders. Publicly available databases of PKU patients indicate that 60% of cases involve at least one ultra-rare SNV, and in 28% of cases, the affected individual carries an ultra-rare variant on both copies of *PAH*. Some of these ultra-rare variants may be *de novo* mutations, and the individual may be the only person known to harbor that exact variant (63). The vast majority of ultra-rare variants are absent from clinical databases, indicating that the current approach of cataloguing observed genetic variants fails when allele frequencies are especially low. For *PAH*, which is one of the most studied metabolic genes, only 9% of possible SNVs have functional impact classified in ClinVar.

Emerging computational algorithms may serve as a means for evaluating the impact of rare variants in IEM and PGx genes. As noted above, existing algorithms have limited ability to accurately predict the impact of variants in these genes, especially among rare variants. Methods have been developed to specifically evaluate variants in pharmacogenes, but these are largely based on existing methods and may have some of the same inherent biases (46). Machine learning has revolutionized computer vision and natural language processing by effectively analyzing spatial and sequential data (64-66). Machine learning is a type of artificial intelligence in which algorithms are taught, or "trained," to make predictions based on existing data. Machine learning forms the basis of existing variant effect prediction algorithms, where an algorithm is trained to predict whether a genetic variant will be deleterious or not on the basis of a training dataset of known deleterious and benign genetic variants. In recent years as computational power and the amount of available data has increased, a type of machine learning that uses deep neural networks, known as deep learning, has become widespread. With the rapid growth in the availability of biological data, deep learning has also been extensively used in bioinformatics (67-74), including transcription factor binding site prediction (75), genome functional annotation (76), and assessment of variant function (77, 78). Several methods have been developed specifically for the evaluation of alleles in pharmacogenes, namely *CYP2D6* (MIM: 124030) (79, 80). These purpose-built models outperform existing methods and are capable of assessing the impact of any combination of variants observed in a haplotype rather than single variants. One major drawback of deep learning is that it requires an immense amount of data in order to estimate the large number of parameters required for good performance (81, 82).

Transfer learning offers an opportunity to leverage the power of deep learning in situations where data are limited. It is difficult to obtain sufficient data to develop phenotype prediction

algorithms from genomic data via deep learning, especially when we only have tens or hundreds of individuals with both genome sequencing data and well-characterized clinical or molecular phenotypes. Transfer learning is an emerging approach for overcoming the challenge of limited data. The idea is to build models that perform a task (X) that is similar to the goal task (Y) but for which there are large amounts of relevant real or simulated data. Once the model for solving task X is performing well, it can be refined with data relevant to task Y. In the case of predicting variants, we might build a model using data from a well-studied gene (X) and then refine the model with data from a poorly studied gene (Y). The resulting model may perform very well on Y because the "lessons" learned in modeling X transfer well to Y (83-88). There are several flavors of transfer learning that have been applied to applications in genetics and proteomics. Convolutional neural network (CNN)-based approaches pre-train weights of convolutional layers on large datasets that can be finetuned on smaller datasets (79). Transformer-based approaches, frequently used in natural language processing, have been applied to functional predictions of variants in proteins (89, 90). Graph-CNNs have been used to make drug-binding predictions with protein structure data after being pre-trained with an unsupervised learning step (91). These transfer learning methods could in theory be used to create structure-based predictions of the effect of amino acid changes on drug binding. These methods combined with *in silico* representations of drug molecules could be used to create substrate-specific predictions of drug-protein interactions and how genetic variants may influence that behavior.

The underlying homology between gene orthologs and paralogs may allow for an increased ability to perform transfer learning. We may be able to use knowledge learned in some domains to inform others. Not surprisingly, some rare diseases have received more attention than others, often because of the incidence of the disease, serendipitous factors, and scientific opportunities.

These well-studied diseases typically have significantly more variant impact data available than others. PKU has an incidence of 1 in 10,000 newborns, and there are hundreds of disease-associated cataloged variants. In comparison, tyrosine hydroxylase deficiency (THD [MIM: 605407]) affects fewer than 1 in 100,000 newborns and has been associated with fewer than 20 variants in TH (MIM: 191290). Sequencing benefits individuals with THD less simply because the disease is rarer and few known pathogenic variants exist. The chemical similarity of phenylalanine and tyrosine leads to a high degree of homology between PAH and TH, which presents an opportunity to transfer knowledge about PKU variants to better understand THD—for example, in understanding which parts of the protein may be more or less tolerant of non-synonymous mutations. However, although transfer learning may offer some advantages in the assessment of rare variation, this approach relies on the existence of genes that are similar enough to the gene of interest with sufficient data. Transfer learning may be valuable for some domains, but there is still a need to generate large amounts of high-quality data. Ideally, for knowledge to be truly transferable, data collection would be ongoing and from whole-population datasets rather than being limited to existing datasets.

Ultimately, the goal of any variant interpretation method is to improve clinical care. Integration of genetics into the clinic is already quite challenging, and integration of computational methods for predicting variant function is rife with further challenges. Learning health systems have long been proposed as models for improving healthcare (92-94), but integration of genetic data into such a system would allow for the accumulation of data to train more sophisticated predictive models as well as an opportunity to iteratively improve upon such algorithms.

A genomic learning healthcare system would allow for rapid collection and phenotyping of rare variants. Learning health systems have been proposed in healthcare since 2007, but few have

fully integrated genetics to inform patient treatment (94). In existing systems, the algorithms are constantly improving on the basis of a feedback loop of data that are collected over the course of patient treatment. A genomic learning healthcare system (GLHS) would operate in much the same way, but with the addition that clinical decision support is provided on the basis of genetic data as well as clinical data (33). In this proposed system, collection, sequencing, and analysis of patient data would be required as a first step and would need to be available as part of the patient's clinical record in the electronic health system. This would enable clinical decision support for IEM- and PGx-related conditions, providing doctors with diagnosis and treatment guidance. The algorithms underlying the clinical decision support can be evaluated regularly and updated on the basis of newly available patient data. In addition to evaluating the algorithms, sequencing and analyzing important genes for every individual treated will allow for more rapid collection and phenotyping of ultra-rare variants—if ancestrally diverse datasets are available.

The ultimate goal of a GLHS is to improve treatment for all patients by leveraging their genetic data. This includes determining the pathogenicity of rare variants that may be previously unseen in patients and potentially making clinical decisions based on their predicted impact. As a conservative first step, a genomic learning healthcare system could implement existing clinical guidance models for IEMs and PGx, such as the pharmacogenomics dosing recommendations from CPIC. Once genetic data are collected for each patient, predictive models for rare variants can be developed and implemented in clinical practice at such a time when there is sufficient confidence in the predictions of the model. Careful analysis will be needed in selecting and evaluating predictive models for both IEMs and PGx, and it is likely that gene-specific models will be needed. The specific clinical action based on a predicted phenotype will then depend on the application area and the onset and severity of the condition. Severe IEMs may require

immediate intervention (such as PKU), whereas for others, a preventative approach is deployed. Some IEMs respond to pharmaceutical interventions, and genotype may predict likelihood of response to a specific medication (95, 96). For late-onset IEMs, genotype may predict age of onset, which can inform appropriate patient monitoring (97). Similarly for PGx, if the potential consequences of prescribing a drug are life threatening, the clinician may select an alternate therapy. Likewise, if the consequences are mild, they may proceed with caution. We illustrate this framework in Figure 2.3 before turning to the ethical questions to be taken into account (2).

2.6.1 Ethics spotlight 3: How can genomic learning healthcare systems ensure adequate genomic input and data governance?

Data governance and consent for secondary data use will significantly shape whether or not genomic learning healthcare systems can improve accuracy and reduce biases. Learning health care systems present unique ethical challenges that traditional clinical and research ethics—focusing on individual harms and a sharp research/clinical care divide—will find difficult to address (24). Data collection and input (step 1 and step 2 of Fig. 2.3) differs between clinical and public health repositories in terms of provisions around secondary use. One option to improve data privacy and security is through the use of federated learning. This approach involves a centrally pooled dataset with non-co-located data only; data are not shared directly, and model parameters could be protected by research collaboration agreements, advances in data encryption, and a trusted third-party to oversee data access (98).

The use of artificial intelligence in healthcare systems is also complicated by issues arising from the possible encoding and routinization of human bias, even with the use of seemingly neutral data sources (54). Artificial intelligence has been described as “the collective medical mind” (31). More than simply doing no harm, a GLHS should actively support greater health equity (4,

99). Of central importance is whether clinical data, or model parameters if deploying a federated learning approach, could be viewed and secured as a public good insofar as all stakeholders, both healthcare and private industry, hold a moral obligation to use and share clinical data in ways that benefit society over and above individual or commercial interests (25). If viewing clinical data as a public good, determining how to deal with computational predictors and healthcare outcomes that accurately capture differences not so much resulting from human input biases but rather serving unfair social conditions would be of greatest difficulty.

For public health data use, it is important to identify and address social and political inconsistencies in the ethical oversight from institutional review boards and government bodies, particularly in regard to informed consent and anonymization of data (100). This requires careful consideration for how questions of beneficence regarding collections and distribution of quality care across populations can vary and ultimately widen health disparities (101). Taking the case of newborn DBSs, the current justification for the mandatory nature of newborn screening rests on the potential harms to the child were they not screened for these treatable conditions (see (26) for a full historical justification). Safeguards are needed to protect the storage and research use of genetic data, which could become more identifiable (102). With such protections, could the practice of informed consent with individuals be seen as less important than another process to ensure respect for autonomy at a group level in order to meet social obligations to contribute to both greater knowledge and efforts to reduce social inequity in health? (24) Because biobanks of newborn DBSs provide a rich and unique dataset for research and improving newborn screening (and other genetic testing)—with enormous potential for contribution to a GLHS—the loss of such potential, if secondary use of newborn DBSs is only permissible on an individual consent basis, needs to be carefully weighed up against ethical concerns about respect for individual

control. How do we ensure respect for individuals in a GLHS that relies on the collective contributions of entire populations in order for everyone to potentially benefit? Those implementing machine learning research in a GLHS must engage these questions directly.

2.7 Conclusion

The defining problem of the genomic age is the interpretation of human genetic variation. In reviewing computational advancements and ethical concerns, we look to develop gene-specific variant interpretation algorithms with a genomic learning healthcare system that builds from a focus on early-onset treatable disease in newborns and actionable pharmacogenomics recommendations. We seek diagnosis of IEMs and treatment for PGx that is tailored to each individual and treatment outcomes that are shared to improve treatment for future patients across all of society. The existing system is the first step toward this goal, as evidenced by confirmatory sequencing of patients and variant cataloging in databases such as ClinVar. Yet the existing system falls short because it is reactive rather than predictive and accurate treatment depends on whether the variant has been previously seen and cataloged. Importantly, it remains to be determined whether computational methods can alleviate health inequity that is reinforced by these limited variant databases. Pervasive sequencing may indeed present a social justice opportunity: to actively promote a more fair and consistent distribution of treatment across all population groups. Yet, there are many barriers blocking the way, including unrepresentative sequencing databases, secondary data use permissions, barriers to healthcare access, and existing biases at the human interface of research and caregiving.

There are technical challenges, including accurate variant classification, data limitations, and growing numbers of variants of uncertain significance. A combination of a GLHS and transfer

learning can overcome existing data limitations in order to improve the computational prediction of variants. An increased understanding of each patient's variants will enable more precise diagnosis and treatment. Most importantly, as more patients provide information into the system, lessons learned from one patient may inform the care of all patients. A dynamic and fair genomic learning healthcare system will create the greatest patient benefit from the captured genomic and phenotypic information, but this will fundamentally depend on careful consideration of societal implications.

2.9 Figures

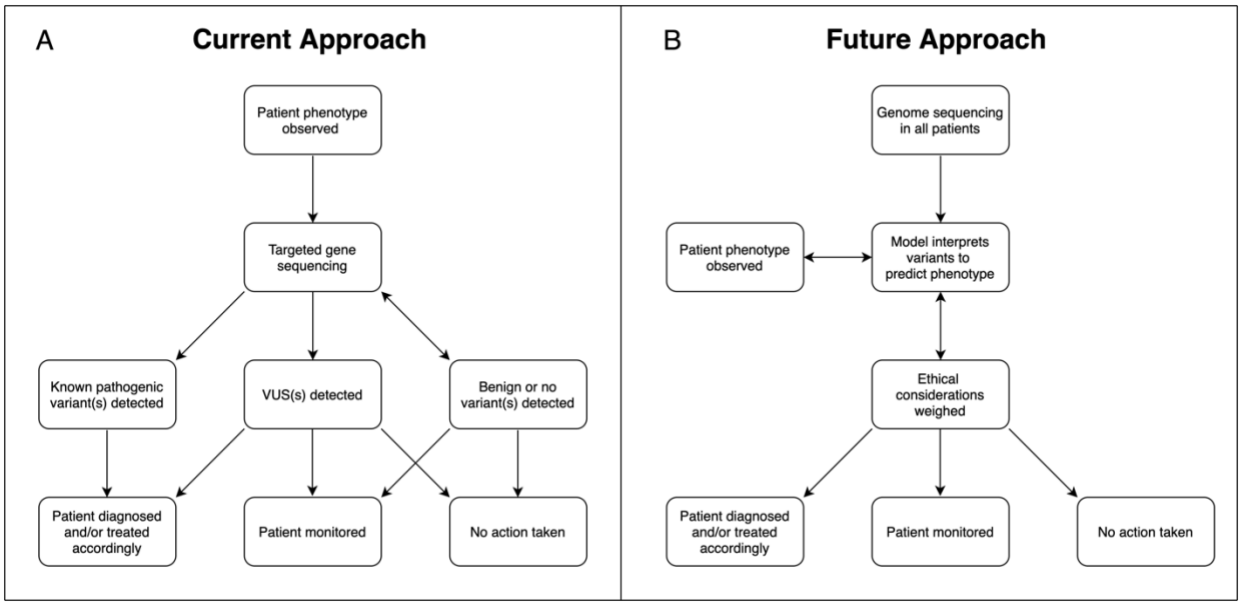


Figure 2.1. Diagram of current treatment workflow and proposed workflow that integrates genomics. Simplified overview of the identification and treatment of patients with IEMs or PGx. We contrast the current approach with our proposed framework, which incorporates early sequencing and analysis of rare variants with machine learning and ethical considerations. **(A)** Current practice for IEMs and PGx begins with an observable phenotype. In IEMs, this may be an altered metabolite detected by newborn screening; in PGx, perhaps an adverse event. Phenotype can also include physical examination, medical history, family history, and relevant labs or studies. Genetic sequencing is then performed, which could include targeted single-gene sequencing with copy number variant detection, gene panel, whole-exome sequencing, or in some cases, family trio sequencing to assess phasing and identify *de novo* variants. If annotated pathogenic variants are identified in the target gene, a patient may be diagnosed with a disease and offered preventative services (as is the case with IEMs) or given a different drug or dose adjustment (as with PGx). Identification of VUSs may result in a diagnosis, depending on the other variants identified. In this simplified figure, VUS refers to a variant in the targeted gene of interest as opposed to an incidental finding not relevant to diagnosis. Patient diagnosis can occur without DNA sequencing, as is the case with some IEMs. **(B)** Hypothetical future approach to patient care in the fields of PGx and IEMs. All individuals undergo whole-genome sequencing at birth. Machine learning models use detected variants to predict phenotype (disease risk or differential drug response). Ethical considerations are addressed, and clinical action is taken accordingly.

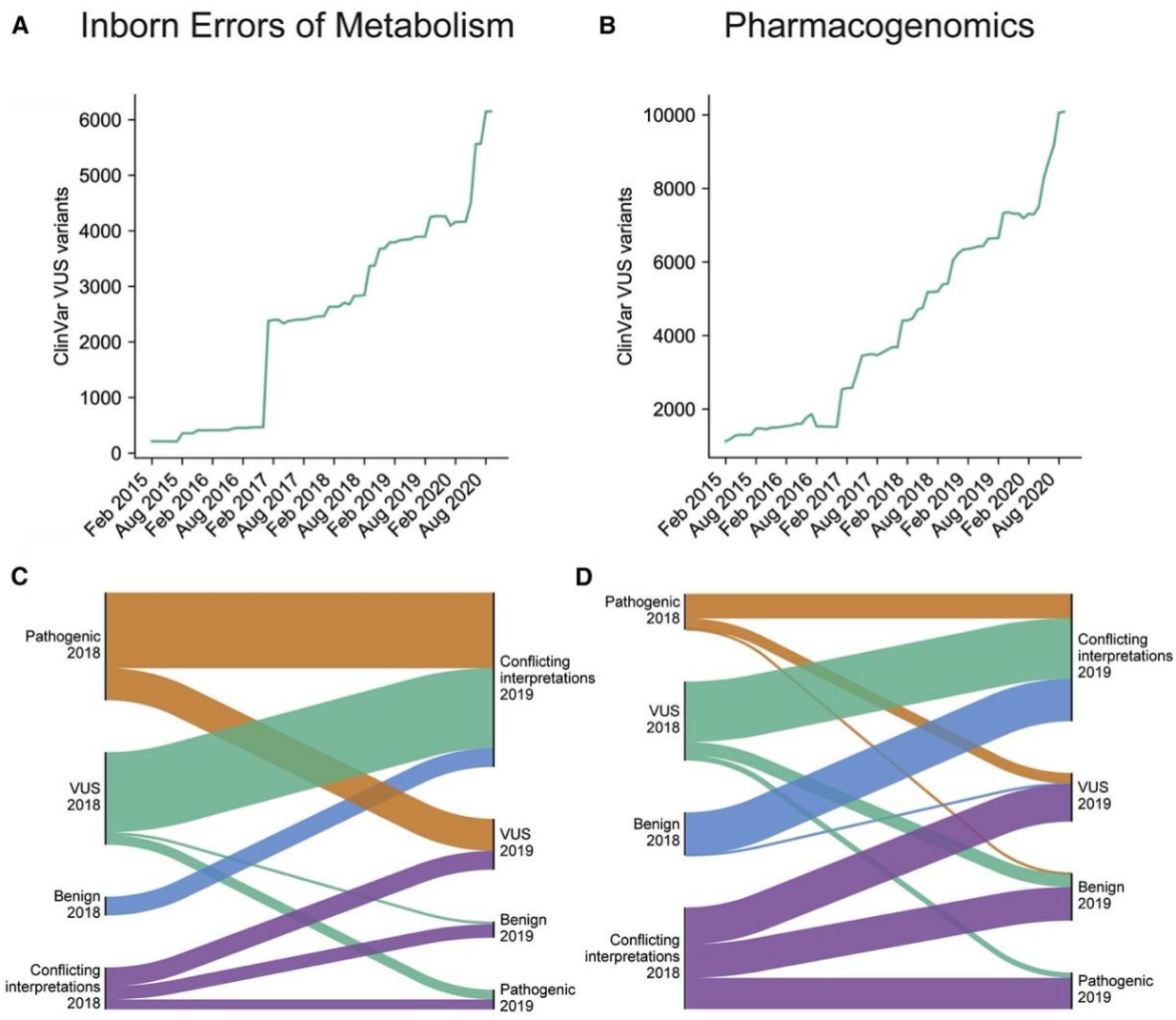


Figure 2.2. ClinVar variants of uncertain significance in genes related to IEMs and PGx. The number of VUSs in ClinVar between 2015 and 2020 in **(A)** IEM and **(B)** PGx genes, respectively. All ClinVar variants in (**C)** IEM and **(D)** PGx genes that were reclassified between February 2018 and February 2019. Height of bars is proportional to number of variants reclassified. A total of 293 variant reclassifications is shown in **(C)** and 434 variant reclassifications are shown in **(D).**

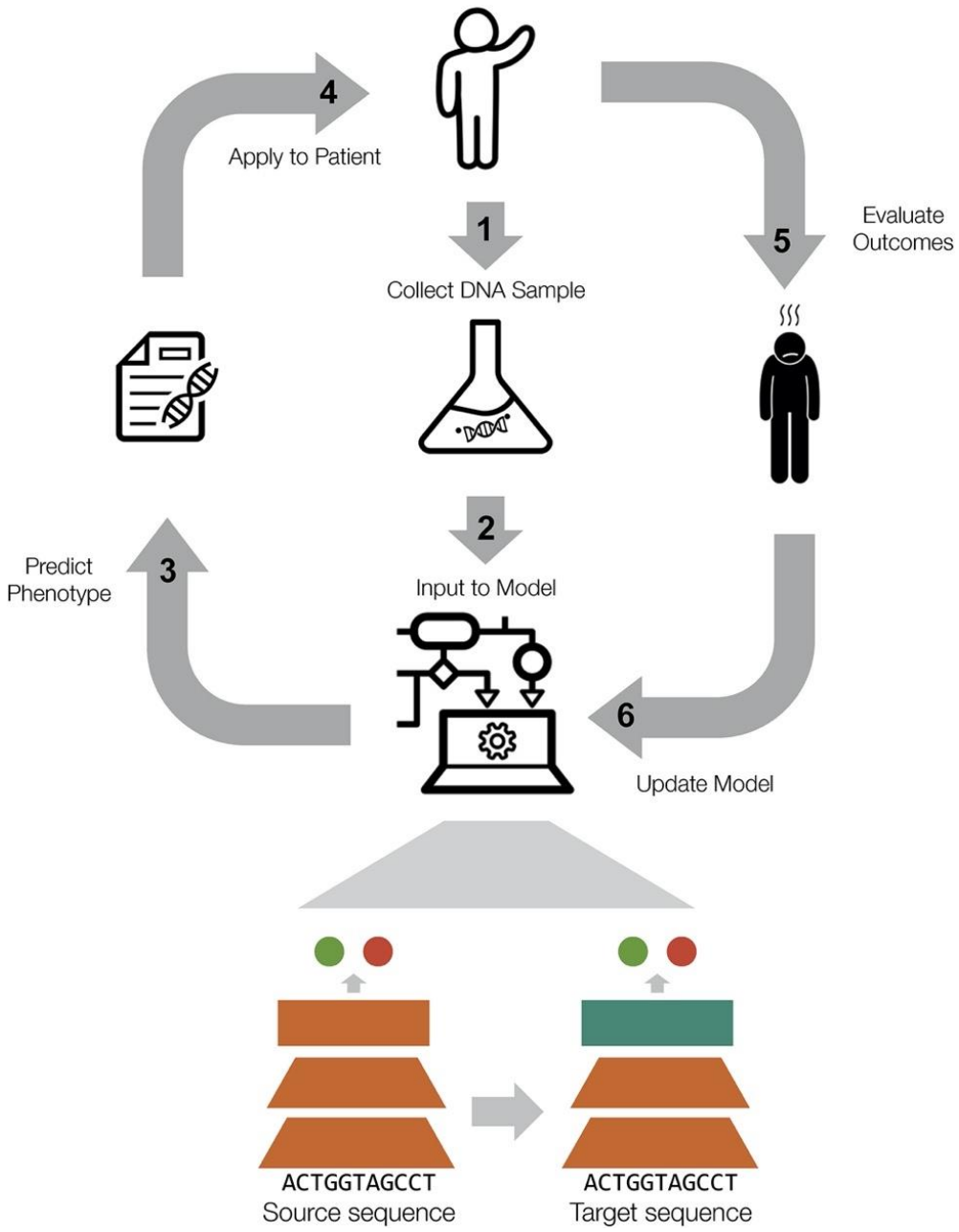


Figure 2.3 Proposed workflow for a genomic learning healthcare system. Patients' DNA samples are collected and sequenced with genomic data input to computational models. The model outputs a predicted phenotype for the patient; results are reviewed by clinicians and applied to the patient. Outcomes are evaluated and the model continues to learn from a feedback loop to improve outcomes for future patients. Icons are from The Noun Project (103-106).

2.10 Tables

Table 2.1. Ethical considerations for the adoption of novel genomic technologies into learning health system practice.

Areas of IEMs and PGx and ethical issues	Key question
Whole-genome sequencing for newborns: (1) uncertainty of results and (2) return of clinical results, including results from late-onset disorders	Can genome sequencing improve the uncertainty of results and return of clinical results?
Interpreting VUSs: (3) research and clinical divide and (4) social/ racial inequity	Can we view the classification of VUSs as a social justice opportunity to close social and genetic ancestry gaps?
Genomic learning healthcare systems: (5) privacy risks and (6) data sharing	How can genomic learning healthcare systems ensure adequate genomic input and data governance?

VUSs, variants of uncertain significance.

2.11 References

1. Adhikari AN, Gallagher RC, Wang Y, Currier RJ, Amatuni G, Bassaganyas L, et al. The role of exome sequencing in newborn screening for inborn errors of metabolism. Nat Med. 2020;26(9):1392-7.
2. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.
3. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054-6.
4. Green ED, Gunter C, Biesecker LG, Di Francesco V, Easter CL, Feingold EA, et al. Strategic vision for improving human health at The Forefront of Genomics. Nature. 2020;586(7831):683-92.
5. Lavertu A, McInnes G, Daneshjou R, Whirl-Carrillo M, Klein TE, Altman RB. Pharmacogenomics and big genomic data: from lab to clinic and back again. Hum Mol Genet. 2018;27(R1):R72-R8.
6. Van Driest SL, Shi Y, Bowton EA, Schildcrout JS, Peterson JF, Pulley J, et al. Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. Clin Pharmacol Ther. 2014;95(4):423-31.
7. Reisberg S, Krebs K, Lepamets M, Kals M, Mägi R, Metsalu K, et al. Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. Genet Med. 2019;21(6):1345-54.

8. McInnes G, Lavertu A, Sangkuhl K, Klein TE, Whirl-Carrillo M, Altman RB. Pharmacogenetics at Scale: An Analysis of the UK Biobank. Clin Pharmacol Ther. 2021;109(6):1528-37.

9. Martin MA, Klein TE, Dong BJ, Pirmohamed M, Haas DW, Kroetz DL, et al. Clinical pharmacogenetics implementation consortium guidelines for HLA-B genotype and abacavir dosing. Clin Pharmacol Ther. 2012;91(4):734-8.

10. Bank PCD, Caudle KE, Swen JJ, Gammal RS, Whirl-Carrillo M, Klein TE, et al. Comparison of the Guidelines of the Clinical Pharmacogenetics Implementation Consortium and the Dutch Pharmacogenetics Working Group. Clin Pharmacol Ther. 2018;103(4):599-618.

11. Ingelman-Sundberg M, Mkrtchian S, Zhou Y, Lauschke VM. Integrating rare genetic variants into pharmacogenetic drug response predictions. Hum Genomics. 2018;12(1):26.

12. Ferreira CR, van Karnebeek CDM, Vockley J, Blau N. A proposed nosology of inborn errors of metabolism. Genet Med. 2019;21(1):102-6.

13. Waters D, Adeloye D, Woolham D, Wastnedge E, Patel S, Rudan I. Global birth prevalence and mortality from inborn errors of metabolism: a systematic analysis of the evidence. J Glob Health. 2018;8(2):021102.

14. Popejoy AB, Crooks KR, Fullerton SM, Hindorff LA, Hooker GW, Koenig BA, et al. Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am J Hum Genet. 2020;107(1):72-82.

15. Feuchtbaum L, Carter J, Dowray S, Currier RJ, Lorey F. Birth prevalence of disorders detectable through newborn screening by race/ethnicity. Genet Med. 2012;14(11):937-45.

16. Azzopardi PJ, Upshur REG, Luca S, Venkataramanan V, Potter BK, Chakraborty PK, et al. Health-care providers' perspectives on uncertainty generated by variant forms of newborn screening targets. Genet Med. 2020;22(3):566-73.

17. Azimi M, Schmaus K, Greger V, Neitzel D, Rochelle R, Dinh T. Carrier screening by next-generation sequencing: health benefits and cost effectiveness. Mol Genet Genomic Med. 2016;4(3):292-302.

18. Kraft SA, Duenas D, Wilfond BS, Goddard KAB. The evolving landscape of expanded carrier screening: challenges and opportunities. Genet Med. 2019;21(4):790-7.

19. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11(8):801-7.

20. Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, Pugh TJ, et al. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet. 2013;84(5):453-63.

21. Rodenburg RJ. The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis. 2018;41(3):297-307.

22. Suiter CC, Moriyama T, Matreyek KA, Yang W, Scaletti ER, Nishii R, et al. Massively parallel variant characterization identifies NUDT15 alleles associated with thiopurine toxicity. Proc Natl Acad Sci U S A. 2020;117(10):5394-401.

23. Oshiro C, Mangravite L, Klein T, Altman R. PharmGKB very important pharmacogene: SLCO1B1. Pharmacogenet Genomics. 2010;20(3):211-6.

24. Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Cent Rep. 2013;Spec No:S16-27.

25. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP. Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. Radiology. 2020;295(3):675-82.

26. Johnston J, Lantos JD, Goldenberg A, Chen F, Parens E, Koenig BA, et al. Sequencing Newborns: A Call for Nuanced Use of Genomic Technologies. Hastings Cent Rep. 2018;48 Suppl 2:S2-S6.

27. Couzin-Frankel J. Unknown significance. Science. 2014;346(6214):1167-70.

28. Vineis P. Ethical issues in genetic screening for cancer. Ann Oncol. 1997;8(10):945-9.

29. Hippman C, Nislow C. Pharmacogenomic Testing: Clinical Evidence and Implementation Challenges. J Pers Med. 2019;9(3).

30. McCullough LB, Brothers KB, Chung WK, Joffe S, Koenig BA, Wilfond B, et al. Professionally Responsible Disclosure of Genomic Sequencing Results in Pediatric Practice. Pediatrics. 2015;136(4):e974-82.

31. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med. 2018;378(11):981-3.

32. Martinez-Martin N, Magnus D. Privacy and ethical challenges in next-generation sequencing. Expert Rev Precis Med Drug Dev. 2019;4(2):95-104.

33. Lu CY, Williams MS, Ginsburg GS, Toh S, Brown JS, Khoury MJ. A proposed approach to accelerate evidence generation for genomic-based technologies in the context of a learning health system. Genet Med. 2018;20(4):390-6.

34. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

35. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7.

36. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015;372(23):2235-42.

37. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D7.

38. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE, et al. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. Clin Pharmacol Ther. 2018;103(3):399-401.

39. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Twist GP, Klein TE, Miller NA, et al. The Evolution of PharmVar. Clin Pharmacol Ther. 2019;105(1):29-32.

40. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414-7.

41. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.

42. van der Velde KJ, Kuiper J, Thompson BA, Plazzer JP, van Valkenhoef G, de Haan M, et al. Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. Hum Mutat. 2015;36(7):712-9.

43. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85.

44. Li Q, Liu X, Gibbs RA, Boerwinkle E, Polychronakos C, Qu HQ. Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PLoS One. 2014;9(8):e104452.

45. Hamasaki-Katagiri N, Salari R, Wu A, Qi Y, Schiller T, Filiberto AC, et al. A gene-specific method for predicting hemophilia-causing point mutations. J Mol Biol. 2013;425(21):4023-33.

46. Zhou Y, Mkrtchian S, Kumondai M, Hiratsuka M, Lauschke VM. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. Pharmacogenomics J. 2019;19(2):115-26.

47. Adhikari AN. Gene-specific features enhance interpretation of mutational impact on acid alpha-glucosidase enzyme activity. Hum Mutat. 2019;40(9):1507-18.

48. Lal D, May P, Perez-Palma E, Samocha KE, Kosmicki JA, Robinson EB, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. Genome Med. 2020;12(1):28.

49. Heyne HO, Baez-Nieto D, Iqbal S, Palmer DS, Brunklaus A, May P, et al. Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. Sci Transl Med. 2020;12(556).

50. Clerx M, Heijman J, Collins P, Volders PGA. Predicting changes to INa from missense mutations in human SCN5A. Sci Rep. 2018;8(1):12797.

51. Li B, Mendenhall JL, Kroncke BM, Taylor KC, Huang H, Smith DK, et al. Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance. Circ Cardiovasc Genet. 2017;10(5).

52. Jorge LF, Eichelbaum M, Griese EU, Inaba T, Arias TD. Comparative evolutionary pharmacogenetics of CYP2D6 in Ngawbe and Embera Amerindians of Panama and Colombia: role of selection versus drift in world populations. Pharmacogenetics. 1999;9(2):217-28.

53. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.

54. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.

55. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.

56. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100(4):635-49.

57. Chakchouk I, Zhang D, Zhang Z, Francioli LC, Santos-Cortez RLP, Schrauwen I, et al. Disparities in discovery of pathogenic variants for autosomal recessive non-syndromic hearing impairment by ancestry. Eur J Hum Genet. 2019;27(9):1456-65.

58. Perera MA, Cavallari LH, Johnson JA. Warfarin pharmacogenetics: an illustration of the importance of studies in minority populations. Clin Pharmacol Ther. 2014;95(3):242-4.

59. Amendola CP, Silva-Jr JM, Carvalho T, Sanches LC, Silva U, Almeida R, et al. Goal-directed therapy in patients with early acute kidney injury: a multicenter randomized controlled trial. Clinics (Sao Paulo). 2018;73:e327.

60. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.

61. Rahit K, Tarailo-Graovac M. Genetic Modifiers and Rare Mendelian Disease. Genes (Basel). 2020;11(3).

62. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet. 2005;6(4):287-98.

63. Blau N, Shen N, Carducci C. Molecular genetics and diagnosis of phenylketonuria: state of the art. Expert Rev Mol Diagn. 2014;14(6):655-71.

64. Hinton G, Deng L, Yu D, Dahl G, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process Mag 2012;29:82-97.

65. Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, et al. Skip-Thought Vectors. 28 ed: Curran Associates, Inc.; 2015.

66. Collobert R, Weston J. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. New York, NY, USA: ACM2008.

67. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141).

68. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851-69.

69. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12-8.

70. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321-32.

71. Yue T, Wang H. Deep Learning for Genomics: A Concise Overview. arXiv. 2018:1802.00810.

72. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):878.

73. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, et al. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods. 2018;15(4):290-8.

74. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26(7):990-9.

75. Shen Z, Bao W, Huang DS. Recurrent Neural Network for Predicting Transcription Factor Binding Sites. Sci Rep. 2018;8(1):15270.

76. Khodabandelou G, Routhier E, Mozziconacci J. Genome annotation across species using deep convolutional neural networks. PeerJ Comput Sci. 2020;6:e278.

77. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-3.

78. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931-4.

79. McInnes G, Dalton R, Sangkuhl K, Whirl-Carrillo M, Lee SB, Tsao PS, et al. Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol. 2020;16(11):e1008399.

80. van der Lee M, Allard WG, Rolf HA, Baak-Pablo RF, Menafra R, Birgit AL, et al. A unifying model to predict variable drug response for personalised medicine. bioRxiv. 2020.

81. Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S. Why Does Unsupervised Pre-training Help Deep Learning? . J Mach Learn Res. 2010;11:625-60.

82. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.

83. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng. 2010;22:1345–59.

84. Shao L, Zhu F, Li X. Transfer learning for visual categorization: a survey. IEEE Trans Neural Netw Learn Syst 2015;26:1019-34.

85. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016;3:9.

86. Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, Savarese S. Taskonomy: Disentangling task transfer learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3712–22.

87. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks?: Curran Associates, Inc.; 2014.

88. Taroni JN, Grayson PC, Hu Q, Eddy S, Kretzler M, Merkel PA, et al. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Syst. 2019;8(5):380-94 e4.

89. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating Protein Transfer Learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689-701.

90. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. arXiv. 2020.

91. Torng W, Altman RB. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. J Chem Inf Model. 2019;59(10):4131-49.

92. Guise JM, Savitz LA, Friedman CP. Mind the Gap: Putting Evidence into Practice in the Era of Learning Health Systems. J Gen Intern Med. 2018;33(12):2237-9.

93. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med. 2012;157(3):207-10.

94. Etheredge LM. A rapid-learning health system. Health Aff (Millwood). 2007;26(2):w107-18.

95. Leuders S, Wolfgart E, Ott T, du Moulin M, van Teeffelen-Heithoff A, Vogelpohl L, et al. Influence of PAH Genotype on Sapropterin Response in PKU: Results of a Single-Center Cohort Study. JIMD Rep. 2014;13:101-9.

96. Yildiz Y, Talim B, Haliloglu G, Topaloglu H, Akcoren Z, Dursun A, et al. Determinants of Riboflavin Responsiveness in Multiple Acyl-CoA Dehydrogenase Deficiency. Pediatr Neurol. 2019;99:69-75.

97. Grunert SC. Clinical and genetical heterogeneity of late-onset multiple acyl-coenzyme A dehydrogenase deficiency. Orphanet J Rare Dis. 2014;9:117.

98. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3:119.

99. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med. 2018;169(12):866-72.

100. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections. BMC Med Ethics. 2016;17(1):54.

101. Green AR, Tan-McGrory A, Cervantes MC, Betancourt JR. Leveraging quality improvement to achieve equity in health care. Jt Comm J Qual Patient Saf. 2010;36(10):435-42.

102. Shringarpure SS, Bustamante CD. Privacy Risks from Genomic Data-Sharing Beacons. Am J Hum Genet. 2015;97(5):631-46.

103. Lamm V. People Drawn Thin Collection, Person. thenounproject.com.

104. Alberto Gongora H. Big Data Collection, Machine Learning. thenounproject.com.

105. Wray A. Gene Testing. thenounproject.com.

106. ProSymbols. DNA, U.S. STEM Elements Line Icons Collections. thenounproject.com.

Chapter 3: Large-scale functional characterization of OCTN2 variants informs protein-specific variant effect predictor for Carnitine Transporter Deficiency

3.1 Abstract

Genetic variants in *SLC22A5*, encoding the membrane carnitine transporter OCTN2, cause the rare metabolic disorder Carnitine Transporter Deficiency (CTD). CTD is potentially lethal but actionable if detected early, with confirmatory diagnosis involving sequencing of *SLC22A5*. Interpretation of missense variants of uncertain significance (VUS) is a major challenge. In this study, we sought to characterize the largest set to date (n=150) of OCTN2 variants identified in diverse ancestral populations, with the goals of furthering our understanding of the mechanisms leading to OCTN2 loss-of-function (LOF) and creating a protein-specific variant effect prediction model for OCTN2 function. Uptake assays with ^{14}C-carnitine revealed that 105 (70%) of variants significantly reduced transport of carnitine compared to wild-type OCTN2, and 37 variants (25%) severely reduced function to less than 20%. All ancestral populations harbored LOF variants. 62% of GFP-tagged variants impaired OCTN2 localization to the plasma membrane of HEK293T cells and subcellular localization significantly associated with function, revealing a major LOF mechanism of interest for CTD. With these data, we trained a model to classify variants as functional (> 20% function) or LOF (< 20% function). Our model outperformed existing state-of-the-art methods as evaluated by multiple performance metrics, with mean area under the receiver operating characteristic curve (AUC) of 0.895±0.025. In summary, in this study we generated a rich dataset of OCTN2 variant function and localization, revealed important disease-causing mechanisms, and improved upon machine-learning based prediction of OCTN2 variant function to aid in variant interpretation in the diagnosis and treatment of CTD.

3.2 Introduction

Loss-of-function (LOF) variants in transporters in the Solute Carrier Superfamily (SLC) are responsible for over 100 rare genetic diseases (1, 2). Carnitine transporter deficiency (CTD; OMIM #212140 (3); also known as primary carnitine deficiency or carnitine uptake defect) is a rare metabolic disorder caused by biallelic LOF variants in *SLC22A5*, the gene that encodes the plasma membrane carnitine transporter OCTN2. Without timely detection, CTD can be fatal (4-6), but clinical outcomes are relatively successful when diagnosed early and treated with supplemental L-carnitine (7), highlighting the need for sensitive diagnostic practices (8).

As an actionable monogenic disorder, CTD is included in newborn screening (NBS) programs throughout the US. Tandem mass spectrometry (MS/MS) using samples from newborn dried blood spots is the primary screen to flag newborns with abnormally low plasma carnitine levels for further testing. However, use of biochemical assays in NBS for CTD encounters several limitations, and confirmatory diagnosis can be arduous. Biochemical assays in newborn dried blood spots result in many false positive cases, in part because newborn carnitine levels are influenced by maternal carnitine levels, which can be low due to undiagnosed maternal carnitine deficiency or pregnancy-associated reduction in total carnitine (9). Furthermore, the plasma carnitine cutoff value for prompting further workup is not standardized: thresholds that are too high burden NBS programs with many false positives, whereas thresholds that are too low result in false negatives with potentially fatal consequences (9). The poor performance of biochemical-based NBS for CTD in New Zealand resulted in the discontinuation of CTD screening (10), a consideration also underway in Germany where many cases are reportedly missed by NBS (11). Confirmatory testing for CTD following abnormal biochemical results includes sequencing of

the *SLC22A5* gene. Transporter functional assay may also be performed, though this is burdensome and not timely as it requires fibroblasts cultured from a skin biopsy.

Although DNA sequencing may result in early and definitive diagnosis through the identification of variants with already known disease association, the identification of variants of uncertain significance (VUS) and rare or novel variants with unknown clinical consequence can make diagnosis via sequencing difficult. For example, the ClinVar database (12) cataloging clinical significance of genetic variants has entries for 252 unique missense OCTN2 variants. Less than a quarter of these variants have clinical interpretations (7 variants assigned benign or likely benign, 50 variants assigned pathogenic or likely pathogenic), with the remaining 77.4% of variants classified as either conflicting interpretation (n=30) or VUS (n=165).

The interpretation of variation in clinically important genes represents a key challenge in genomic medicine (13). While computational predictions are an important tool that can be used to aid in variant interpretation, there are several barriers to accuracy. Because the majority of prediction models are trained on large datasets of cataloged variants derived from individuals with European ancestry (14-16), an unfair bias is incorporated into the prediction methods with decreased accuracy for variants in individuals of non-European ancestries (17). Further, highly cited gene-agnostic prediction models perform worse for membrane proteins (like OCTN2) compared to soluble proteins (18). Recently, a number of protein-specific variant effect predictors (VEPs) have been successful in outperforming gene-agnostic models (19-23), though none have yet to resolve issues of genomic inclusion.

Training computational models to perform variant effect prediction requires a large amount of data that represents the population served. Thus, the most obvious candidate proteins for such models are those that are linked to highly penetrant monogenic diseases and easily assayable. In this study, we sought to characterize the function and localization of 150 genetic variants in OCTN2 from ancestrally diverse populations, with the ultimate goals of (1) informing inclusive diagnostics and therapeutic strategies for LOF CTD variants and (2) using machine learning to build a protein-specific model to predict functional impact of novel OCTN2 genetic variants.

3.3 Methods

3.3.1 Variant selection and annotation

Variants included in this study were carefully selected to ensure equal representation from diverse ancestral populations available in the Broad Institute's Genome Aggregation Database (gnomAD) (24). 150 *SLC22A5* missense variants were selected for characterization (for simplicity referred to as OCTN2 variants to signify the change at the protein level). Detailed workflow for the selection of OCTN2 variants characterized in this study can be found in Supplementary Figure 3.1. Predicted membrane topology of OCTN2 was modeled from UniProtKB (#O76082). Clinical association of all variants was annotated based on literature and database review and identification of that variant in an individual clinically diagnosed with CTD or suspected of possible CTD due to low carnitine levels and presence of at least one variant in OCTN2. Search terms for literature review included "SLC22A5 mutation", "OCTN2 mutation", "primary carnitine deficiency case study", and "primary carnitine deficiency newborn screening". We compiled a list of previously characterized variants published in the literature to use as an additional dataset for performance evaluation of our machine learning models. Variants included in the dataset met the following criteria: (1) the variant was expressed and assayed in a

mammalian system, and (2) the experiment measured the carnitine transport of a single missense OCTN2 variant. The dataset is referred to as "Literature Variants" (Supplementary Dataset 3.1).

3.3.2 Cell culture

HEK293T cells were cultured in Dubecco's modified Eagle medium (DMEM) (Life Technologies, Carlsbad, CA) supplemented with 10% fetal bovine serum (GE Healthcare Life Sciences, South Logan, UT) and penicillin/streptomycin (100 U/mL) (Life Technologies, Carlsbad, CA) and grown in a humidified incubator at 37°C with 5.0% CO_2.

3.3.3 Construct generation

A custom wild-type OCTN2 plasmid was generated by golden gate cloning as previously described (25). Full length *SLC22A5* cDNA (NM_003060.4) was synthesized (Twist Bioscience, San Francisco, CA) with the start codon removed and adapter sequences synthesized on either end to facilitate golden gate cloning reactions. The linear *SLC22A5* cDNA was domesticated into the MTK0_027 entry vector by golden gate cloning with BsmBI. A second golden gate cloning reaction was performed with BsaI to assemble the transcription unit (TU) plasmid, which included parts MTK1_001, MTK2_023, MTK3a_030, the MTK0_027-SLC22A5 as part 3b, MTK4a_015, MTK4b_001, MTK5_006, and MTK678_001. Description of parts is provided in Supplementary Table 3.1. A third golden gate cloning reaction was performed with BsmBI to insert the *SLC22A5* TU into the MTK0_017 destination vector. MTK plasmids detailed in the Construct Generation section below were a generous gift from the Hana El-Samad Lab (UCSF, San Francisco, CA). Assembled constructs were sequenced (MCLAB, South San Francisco, CA) to validate successful cloning and ensure absence of mutations. All OCTN2 variants selected for

functional characterization were synthesized by site-directed mutagenesis (Genscript, Piscataway, NJ) from the final assembled *SLC22A5* construct.

3.3.4 Transient transfection of plasmids containing OCTN2 variants

Human embryonic kidney cells (HEK293T) containing a landing pad at the hAAVS1 locus were used to create transient or stable cell lines. Transient transfection of constructs encoding the wild-type OCTN2 and OCTN2 variants was achieved by reverse transfection using Lipofectamine LTX transfection reagent (Thermo Fisher Scientific) according to manufacturer's protocol. The MTK0_017 destination vector construct was used as the empty vector. Constructs were mixed with Lipofectamine LTX in OptiMEM media (Life Technologies, Carlsbad, CA), vortexed for 10 seconds, allowed to stand at room temperature for 15 mins, then added to poly-D-lysine coated 96-well plates. Each well received 100 ng of DNA and 0.2 µL Lipofectamine LTX. HEK293T cells were counted and seeded into wells at a density of 35,000 cells/well. After transient transfection, cells were cultured for an additional 48 hours before subsequent experiments were performed (uptake assays or confocal imaging).

3.3.5 In vitro uptake assays

48 hours after reverse transient transfection of OCTN2 variants in poly-D-lysine coated 96-well plates (Fisher Scientific #356461), culture medium was removed and cells were washed three times with Hank's buffered salt solution (HBSS) (Life Technologies, Carlsbad, CA) at 37°C and pre-incubated with the third wash of HBSS for 10 min at 37°C. 80 µL of 1 µM ^{14}C-L-carnitine hydrochloride (Moravek Biochemicals #MC1147, Brea, CA) in HBSS (reaction mix) was added to each well and incubated at 37°C for 10 min, a time point within the linear uptake phase of OCTN2 (26). After 10 min, the reaction mix was aspirated and the cells were washed 3x with

ice-cold HBSS. 280 µL lysis buffer (0.1N NaOH, 0.1% v/v SDS) was added to each well and cells were lysed on an orbital shaker for 1 hour. Then, 230 µL cell lysate was removed from each well and added to liquid scintillation tubes with 2.5 mL Ecolite Liquid Scintillation Cocktail (MP Biomedicals #0188247501, Santa Ana, CA). Tubes were vortexed and the radioactivity in each sample was measured on a Beckman LS6500 liquid scintillation counter (Beckman Coulter, Brea, CA). 25 µL lysate from each well was reserved for protein concentration determination by Pierce BCA Assay. Function of each variant was normalized to wild-type (WT) OCTN2 and expressed as a percentage after background carnitine uptake measured in the empty vector (EV) was subtracted from both, calculated as follows: (Variant – EV)/(WT – EV)*100. Each variant was assayed in triplicate on a 96-well plate and measured in three biological replicates.

3.3.6 Confocal imaging and localization classification

Plasmids encoding OCTN2 variants with a C-terminal monomeric superfolder green fluorescent protein (msfGFP) tag were transiently transfected into HEK293T cells seeded at 20,000 cells/well in black wall poly-D-lysine coated 96 well plates (Greiner Bio-One #655946) as detailed above. One variant was transfected per well. After 48 hours, the plasma membrane was stained for 5 min with Wheat Germ Agglutinin Alexa Fluor 647 Conjugate (Thermo Fisher Scientific) diluted 1:500 in HBSS and cells fixed with 3.7% formaldehyde in HBSS for 20 min. Nuclei were stained with 10 µM Hoechst 33342 dye in HBSS (Thermo Scientific #62249) for 20 min at room temperature. Plates were imaged with the IN Cell Analyzer 6500 confocal high-content imaging system (General Electric Life Sciences/Cytiva, Marlborough, MA), using a 488 nm excitation laser. Nine images were taken per well with all samples imaged on the same day using the same image acquisition settings, and results were then replicated on two independent days. Three independent researchers reviewed images for localization and qualitatively classified

variants into either membrane, intracellular, or mixed categories. Reviewers were given representative images for each of the localization categories to inform their baseline understanding of categorization. All images were displayed using the same brightness and contrast values before given to the reviewers to ensure assessment consistency. Researchers were blinded to variant name or function. Concordance between classification by image reviewer is strong (Supplementary Figure 3.2).

3.3.7 Machine learning

Methods for generation of features used in machine learning are described in the Supplementary Text in Section 3.8. Classification models were trained to predict severe LOF variants with function less than 20% of wild-type OCTN2 function with respect to carnitine transport. For the classification model, three types of models and four feature sets were evaluated. The three types of machine learning models were LASSO penalized logistic regression, random forest, and gradient boosting machines. Four sets of features were generated: 1) sequence-based features describing the resulting amino acid change and position in OCTN2 protein sequence, 2) structure-based features extracted from the AlphaFold-2 structural model (default model download from the AlphaFold Protein Structure Database (27, 28), 3) prediction-based features derived from unsupervised variant effect prediction models, including variational autoencoders (29), Potts models (30), and protein language models (31), and 4) all features combined. Features in each set are provided in Supplementary Table 3.2. A final classification model was trained using a subset of features that were available for every possible amino acid change. Every combination of model type and feature sets was evaluated by training through 100 iterations of random subsampling of the 150 characterized OCTN2 variants using an 80/20 train/test split. In addition to the test set, we evaluated the predictive performance of each model on a set of 82

characterized OCTN2 variants derived from literature that were not characterized in our study (Supplementary Dataset 3.1).

We trained and evaluated the OCTN2 function classifier by training a LASSO penalized logistic regression model using repeated random sampling using all features. We used random splits of the characterized variants with 105 variants used for training and the remaining 45 used for testing in each fold. The models were trained in R using the package Caret with repeated cross validation for hyperparameter tuning. We defined a binarizing cutoff for our model output by maximizing the sum of the sensitivity and specificity over all possible cutoffs. We then used this cutoff to calculate other metrics (e.g., accuracy) in the test set as well as the literature derived variants. The trained model was then used to predict the function of all possible missense variants in OCTN2, including the variants in the test set and those in the literature derived set. We evaluated the relative importance of the input features using coefficients output by the LASSO model.

We additionally trained a regression model to quantitatively predict the measured function of OCTN2 variants, with model selection performed in the same way described in the classification section, however instead of models tuned for classification we trained models for regression. The final model was trained with 110 variants and tested with the remaining 40. LASSO coefficients were again used to evaluate feature importance.

3.3.8 Data analysis

Statistical analysis was performed in R version 3.6.3. (R Core Team, 2020). Plots were generated using R package ggplot2 version 3.3.5. Additional figures were generated with Biorender.

For carnitine uptake assays, data are expressed as mean ± standard error of the mean (SEM) with significance determined by Student's t-test. A Bonferroni correction was used to adjust the significance level to $\alpha = 0.05/150 = 3.3*10^{-4}$. ANOVA was used to determine significance of difference in mean function by variant group. Despite a significant p-value of 0.0234 in the ANOVA, Tukey's post-hoc test revealed there was no significant difference between any of the groups (lowest p-value was 0.059 between Shared and Clinical groups). Difference in mean function by subcellular localization was determined by Welch's ANOVA with Games-Howell post-hoc test (rstatix package in R) due to significant differences in variance.

Multiple metrics were calculated to evaluate the performance of our classification models as well as other published models, including sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), precision or positive predictive value (PPV), negative predictive value (NPV), accuracy, area under the receiver operating characteristic curve (AUC), and Matthew's correlation coefficient (MCC), all of which have been defined previously (19, 22).

3.3.9 Data availability

Functional data were submitted to ARUP OCTN2 mutation database (32) and ClinVar, and are available in Supplementary Dataset 3.2.

3.4 Results

3.4.1 Carnitine uptake studies reveal a continuous spectrum of function of OCTN2 variants

We selected a total of 150 missense variants in OCTN2 for multi-parametric characterization in this study. OCTN2 contains 12 transmembrane domains, 6 extracellular loops, 5 intracellular loops, and intracellular N- and C- termini for a total of 25 domains (Fig. 3.1). To ensure variants selected for characterization had good coverage of the entire protein, we assessed the position of

each variant in the membrane topology of OCTN2. Selected variants spanned the entire predicted secondary structure of the transporter and were present in each intracellular, extracellular, and transmembrane domain of the protein with the exceptions of extracellular loops 2 and 5, which contain 4 and 3 residues, respectively (Fig. 3.1A). The density of variants characterized per transporter domain, calculated as the number of variants assayed in each domain divided by the total number of residues in that domain, ranged from 0.00-0.60, in comparison with the overall variant characterization density of 0.27 (Fig. 3.1B). We characterized variants both associated and unassociated with CTD. Variants associated with CTD were present in most transporter domains (Fig. 3.1B) and accounted for 25.3% of assayed variants (38/150, Supplementary Dataset 3.2).

As the first level of characterization, we performed uptake studies with radiolabeled ^{14}C-carnitine to determine the effect of each variant on OCTN2 function. Interestingly, we observed a continuous functional spectrum, with variant function ranging from -0.25 to 116% of the carnitine transport of the wild-type OCTN2 (Fig. 3.1C). Forty-three OCTN2 variants had no significant impact on transporter function compared to the reference OCTN2, while 107 variants had a statistically significant reduction in carnitine transport (Supplementary Dataset 3.2). Importantly, nearly one quarter of variants assayed (37 variants) exhibited carnitine transport reduced to less than 20% of wild-type function, a threshold previously demonstrated to indicate susceptibility for CTD in patient fibroblasts (33). The majority (26/37) of LOF variants are located in transmembrane domains (Fig. 3.1A).

3.4.2 All ancestral groups harbor variants that exhibit a range of function

OCTN2 variants characterized in this study were carefully selected to ensure equal representation from diverse ancestral populations (Supplementary Figure 3.1). We included variants shared by two or more ancestral populations ("Shared"), variants exclusively found in individuals with African, East Asian, European, Latino, and South Asian ancestries, variants selected from gnomAD at random (blinded to ancestry), and variants with known clinical associations to CTD ("Clinical"). Of note, each group harbored variants spanning a complete range of function (Fig. 3.2A). Median function of OCTN2 variants in the Shared, Random, African, Latino, European, East Asian, South Asian, and Clinical groups was 73.6, 64.7, 62.7, 57.0, 45.5, 44.4, 37.9, and 14.1% of wild-type function, respectively (Fig. 3.2B). Mean function of variants between each group was insignificant ($p \geq 0.059$). We next examined the number of low-functioning variants per group, defined as variants with function less than 20% wild-type function. Interestingly, all groups harbored LOF variants, and Clinical variants had the largest fraction of low-functioning variants (7/10), more than double the fraction of low-functioning variants in any other group (Fig 3.2C-D), as expected.

3.4.3 Confocal imaging reveals variant membrane localization significantly associates with function

To better understand the mechanisms contributing to OCTN2 loss-of-function, we determined the subcellular localization of all 150 OCTN2 variants conjugated to monomeric superfolder green fluorescent protein (msfGFP) in HEK293T cells (34). Confocal imaging revealed that subcellular localization could be classified into three major localization patterns (Fig. 3.3A, Supplementary Dataset 3.2); membrane localization notes the OCTN2 variant localizes primarily to the plasma membrane of the cell similar to the wild-type transporter, intracellular localization

indicates the variant is largely retained in the cytoplasm with minimal or no presence on the plasma membrane, and mixed localization indicates the variant displays a combination of the former patterns, with partial membrane localization in combination with increased intracellular GFP intensity compared to wild-type (Fig. 3.3A*;* inset showing representative cell for each phenotype). Fifty-seven variants displayed membrane localization, 36 variants exhibited intracellular retention, and 57 variants had mixed localization (Fig. 3.3B). Subcellular localization was associated with degree of function: variants on the membrane had the highest median function (72.4% of wild-type OCTN2 function), variants with mixed subcellular localization had a median function of 54.5%, whereas variants retained intracellularly had the lowest median function at 19.0% (Fig. 3.3C). A subset of variants (p.V216L, p.V235G, p.Y243S, p.S470F, and p.R471C) exhibited complete loss-of-function despite proper membrane localization, suggesting that additional mechanisms for loss-of-function may occur.

3.4.4 OCTN2-specific variant effect prediction models outperform existing methods

Most OCTN2 variants identified in CTD patients exhibit severe LOF, with the least functional variants associating with more severe disease presentation (35). Thus, we built a classification model to predict whether OCTN2 missense variants would be LOF, defined as 20% or less than that of wildtype, a clinically meaningful cutoff (33). During model selection we evaluated every combination of three types of machine learning models and four feature sets for each of the 150 OCTN2 variants (see Machine Learning Methods section). We find that a LASSO penalized logistic regression classifier achieves the best performance on the test data and literature derived variants with mean area under the curve (AUC) beneath the receiver operating characteristic (ROC) curve of 0.90 and 0.94 for test data and literature data, respectively (Supplementary Dataset 3.1). We compared the performance of our model to ten other variant prediction models:

REVEL (36), primateAI (37), PolyPhen-2 (38), Rhapsody (39), CADD (40), Dynamut2 (41), ESM-1v (31), MSA Transformer (42), Deep Sequence (29), and EVE (43). We find that our model achieves the best AUC among models tested, indicating that it outperforms existing models in differentiating between functional and LOF OCTN2 variants (Fig. 3.4A, Table 3.1). Additionally, our model achieves an AUC of 0.95 on the functionally characterized OCTN2 variants curated from literature. We evaluated the relative importance of the input features using coefficients from the LASSO model. We find that the most important features for functional prediction were from recent state-of-the-art protein language models (e.g., EVE, ESM-1v, Deep Sequence) in addition to OCTN2-specific descriptors (e.g., intracellular loop, residue number) (Fig. 3.4B).

In addition to the classification model that predicted binary function of variants, we trained a regression model to quantitatively predict the function of OCTN2 variants. We first performed model selection in the same way described in the classification section, finding that the LASSO penalized linear regression model performed best (Supplementary Figure 3.3). We evaluate regression model performance with an R^2 metric, defined as the proportion of variance in measured function that is explained by predicted function. Our model has an R^2 of 0.55 (Supplementary Figure 3.4), considerably higher than any of the other models evaluated in this study. For comparison, the next best performing model is ESM-1v with an R^2 of 0.45.

We used our classification model to generate predictions for the function of all possible missense variants in OCTN2 (n=10,583, Fig. 3.5). We found that 2,097 variants are predicted to cause severe LOF (<20% function), representing 19.8% of all possible variants. From these functional prediction models, we find that charged residue substitutions in transmembrane domains are

predicted to be very damaging to function, whereas hydrophobic substitutions in the TMDs are predicted to have minimal impact on function. Extracellular loop 1, intracellular loop 3 and the intracellular C-terminus are predicted to be most tolerable to substitutions.

3.4.5 Machine learning enables prediction of variant localization

In addition to predicting function, we trained a model to predict the effect of protein variants on subcellular localization. Informed by the subcellular localization data for all 150 OCTN2 variants obtained from confocal imaging, we aimed to predict whether proteins would be properly localized to the membrane or retained intracellularly. We trained two models: one to predict full membrane localization and a second to predict full intracellular retention. This was done because many variants presented with "mixed localization", i.e., partial but incomplete localization to the membrane. These models attempt to predict whether a protein will have complete localization to the membrane or complete retention within the cell. A logistic regression model is able to differentiate missense variants that make it to the membrane from those that are retained intracellular or have mixed localization with good performance (AUC: 0.74, Accuracy: 0.70, Supplementary Figure 3.5). Similarly, our model is able to differentiate variants that cause intracellular retention from those that have membrane or mixed localization (AUC: 0.74, Accuracy: 0.78).

3.5 Discussion

Interpretation of novel genetic variants in a clinical setting for diagnosis and treatment of genetic disorders or pharmacogenomics remains a major challenge. In this study, we functionally characterized and determined the subcellular localization of 150 missense variants in the plasma membrane carnitine transporter, OCTN2. Our work has important implications toward improving

the diagnosis and treatment of CTD. With this study of 150 missense variants, we substantially increased the number of characterized OCTN2 variants in the literature, expanding our knowledge of low-functioning at-risk variants. Importantly, we identified mislocalization as a common cause of loss-of-function. As variant-specific treatment options become increasingly available for genetic disorders involving membrane proteins (e.g., the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)), this information has the potential to be leveraged in future therapy for individuals harboring particular variants. Finally, this wealth of data was used to build machine learning models to predict function and localization of all possible missense variants in OCTN2. The models greatly improved upon performance of current prediction algorithms for OCTN2 and provided robust predictions for all potential variants in the transporter. Below we describe our major findings in the context of the literature.

Functional assays provide powerful tools for interpretation of genetic variants identified clinically. For example, recent studies demonstrate that *in vitro* functional characterization could aid in the reclassification of the majority of missense VUS (44). Interestingly, our functional characterization of OCTN2 missense variants from the gnomAD database revealed there to be a continuous distribution of OCTN2 function (Fig. 3.1C). This is consistent with previous findings; carnitine transport assayed in fibroblasts from 358 individuals investigated for potential CTD revealed a similar functional spectrum (33). About 25% of the gnomAD variants that we screened were LOF and reduced carnitine transport to 20% or less than control, and in theory have the potential to cause CTD in either individuals homozygous for the variants or in compound heterozygotes. These variants are ultra-rare and have not been observed in homozygous individuals, though we cannot exclude their presence in compound heterozygotes.

To continue to expand the number of characterized OCTN2 variants with known clinical association, we enriched the variants selected for this study with 10 additional variants found in confirmed or suspected cases of CTD (Fig. 3.2A, Supplementary Dataset 3.4). Seven of the clinically associated variants, including all four from confirmed cases were LOF, retaining less than 20% wild-type carnitine transport. Three variants classified by ARUP (32) as VUS were LOF, and three variants retained partial or complete function: p.N91S, p.L202P, and p.D139N functioned at 42.3, 55.9, and 115.5% of wild-type, respectively, suggesting an uncertain role as determinants of low plasma carnitine levels.

Our effort to make an ethical selection of variants for study with equal representation from major ancestral groups in the gnomAD database allowed us to analyze trends in OCTN2 function across diverse populations. We found that loss-of-function OCTN2 variants are identified in all major ancestral populations from gnomAD (Fig. 3.2A). While CTD is rare, prevalence varies globally with estimated incidences of 1:300 in the Faroe Islands (45), as high as 1:8,200 in some regions of China (46), 1:40,000 in Japan (45, 47), and up to 1:75,000 in California (9). We found no significant difference in mean variant function between groups (Fig. 3.2B). Indeed, regions with higher incidence of CTD tend to have common founder variants affecting many individuals (e.g., p.N32S in the Faroe Islands (5) and p.R254X in China (8)), rather than increased number of unique pathogenic variants. Reported incidence rates of diagnosed CTD are substantially lower than expected based on the population-specific allele frequencies of pathogenic variants, suggesting that NBS misses cases at an alarming rate (9-11, 33). Incidence rates of CTD have not been reported in countries with primarily African, Latino, or South Asian ancestries to our knowledge.

In tissue, OCTN2 localizes to the apical membrane of enterocytes in the gut and renal proximal tubular cells in the kidney, where its major function is to absorb and reabsorb carnitine into systemic circulation, respectively. Here for the first time to our knowledge, we identify mislocalization of the carnitine transporter (Fig. 3.3A) to be a common loss-of-function mechanism (Fig. 3.3C), with 62% of variants in our study exhibiting partial or complete intracellular retention (Fig. 3.3B). This new knowledge has the potential to be leveraged in novel therapeutic approaches for CTD, where we suggest that a pharmacochaperone designed to stabilize OCTN2 protein folding could rescue membrane localization and restore a degree of function to missense transporter variants. Even minimal increases in membrane localization and function could be sufficient to maintain systemic carnitine levels (48). Such therapeutic approaches have been successful for CFTR in cystic fibrosis (49), clinically tested for enzyme deficiencies (50), and explored for norepinephrine, dopamine, and serotonin transporters NET/*SLC6A2*, DAT/*SLC6A3*, and SERT/*SLC6A4* (51). In the event that a pharmacochaperone is developed for OCTN2, we envision subcellular localization data to be informative in personalized medicine to identify patients harboring mislocalized variants that may benefit from such treatment.

In addition to variants for which mislocalization appears to be the primary cause of LOF, we identified a population of variants that localized properly to the plasma membrane of the cell yet had greatly impaired function (Fig. 3.3C). Though not directly investigated in our study, we hypothesize that LOF in variants localizing to the plasma membrane is due to an alternative mechanism, such as disrupted transporter kinetics. Carnitine is a zwitterion, and OCTN2 is thought to have distinct carnitine and cation binding sites (52). Variants affecting these binding sites as well as those that create steric hindrance in the binding pocket or alter sodium

recognition in this sodium-dependent transporter could reduce or fully prevent the binding or translocation of carnitine, increasing the K_m of carnitine. Notably, 5/6 variants (p.V216L, p.V235G, p.Y243S, p.S470F, and p.R471C, but not p.N367D) that have less than 20% function and proper membrane localization project into the translocation pore of OCTN2, based on the AlphaFold2 predicted structure. Such variants present on the membrane yet nonfunctional could in theory benefit from rescue by allosteric modulators.

Here we present protein-specific variant effect prediction models trained with ethically selected variants from diverse ancestral populations. Using classification models for both function (Fig. 3.4-5) and localization (Supplementary Figure 3.5), we can predict whether any possible variant in OCTN2 will be functional or have impaired membrane localization. Despite limited experimental capacity to characterize just under 1.5% of all possible missense variants in OCTN2, our models outperform existing methods (Fig. 3.4A). Additionally, our models were trained with data equally representing diverse ancestral groups, aiming to reduce model bias and ensure comparable accuracy in prediction of variants identified across ancestries. We identified 2,097 missense variants predicted to cause severe LOF, and 578 variants predicted to be retained intracellularly, with potential to have function rescued by pharmacochaperone-based therapies. We additionally identified 1,697 variants predicted as LOF despite proper membrane localization. It should be noted that localization and function are highly correlated, as protein that does not localize to the membrane is not functional. However, we do see some differences between function and localization predictions across all 10,583 missense variants (Fig. 3.5, Supplementary Figure 3.5). Investigation of features most important in predicting variant function revealed that protein language models trained on millions of protein sequences were most useful to the model (Fig. 3.4B). The importance of these features suggests that evolutionary

conservation is predictive of OCTN2 function. We make these predictions for all possible missense variants available (Supplementary Dataset 3.3) for use by interested individuals, researchers, or clinicians.

The major limitation of this study is the size of the dataset, in which we functionally and spatially characterized 150 OCTN2 missense variants. While we have greatly increased the number of variants published in the literature with functional characterization, the size of the dataset limits the power to train the machine learning predictive models. Our determination of OCTN2 variant function by radioligand uptake assays allowed for sensitive detection of carnitine transport, a readout directly relevant to CTD etiology. However, the use of radioactivity is currently incompatible with deep mutational scanning (DMS) platforms that have recently increased the scale at which variants can be functionally characterized. Many DMS studies rely on assays that are scalable yet lack direct relevance to a disease phenotype. For example, the use of fluorescence-activated cell sorting (FACS) to detect changes in abundance of OATP1B1-GFP failed to identify variants that are expressed in the cell yet nonfunctional (53). With functional assays that directly measure carnitine transport function, we were able to identify a larger proportion of nonfunctional variants than through fluorescence-based assays alone. An additional limitation is found in the imaging of OCTN2-tagged variants, where we were unable to quantify colocalization between GFP (OCTN2) and the cell membrane stain due to software limitations, and thus provide a qualitative classification of cellular membrane localization.

Clinical interpretation of functional genomic studies can be limited by complex genotype-phenotype relationships. For CTD, the function of a single missense variant assayed *in vitro* must be considered cautiously in a disease context. Conflicting reports have been published on the

absence (54-56)) or presence (35, 57) of a genotype-phenotype correlation. Early reports indicated that patients, and in some instances siblings, with the same OCTN2 variants had variability in symptoms, severity, and age of onset (54-56). In contrast, another study found that symptomatic patients had more nonfunctional variants, namely nonsense and frameshift (35). Finally, a study in patients from the Faroe Islands revealed a significant correlation between residual OCTN2 transporter function and plasma carnitine levels (57). Thus, interpretation an individual's set of variants in a clinical setting must cautiously be made by healthcare professionals in accordance with guidelines by the American College of Medical Genetics (ACMG) (13).

At a patient and community utility level, we recognize the importance of meaningful and respectful translation of sequencing results. For families navigating CTD in young children, there may be divergent views about return of results, result actionability, informed consent and the sufficiency of parental assent for deposition and research use of variant information (58-61). It is desirable to establish patient stakeholder support in the management of and democratization of data sharing (62, 63). Nonetheless, the community benefits of characterizing variants from publicly shared databases, as this study reveals, arguably outweigh risks of not being able to achieve full informed consent for (deidentified) data use.

Ultimately, the purpose of OCTN2-specific variant effect prediction models is to inform clinical diagnostics and decision making, including therapeutic decisions. OCTN2 serves as a unique link between two major fields for which computational interpretation of genetic variants is increasingly needed: inborn errors of metabolism (IEM) and pharmacogenomics (PGx). The solute carrier (SLC) superfamily consists of over 400 transporters, more than 100 of which are

linked to IEM and other Mendelian disorders (1, 2). Encoded by *SLC22A5*, OCTN2 also shares homology with several pharmacogenes in the SLC22 family, namely *SLC22A1* (OCT1), *SLC22A2* (OCT2), *SLC22A6* (OAT1), and *SLC22A8* (OAT3). Functional data is limited for most of these transporters, impeding the interpretation of genetic variation in IEM and PGx. Transfer learning offers a solution whereby algorithms optimized with substantial data from one protein (e.g., OCTN2) could be refined with minimal data for related proteins implicated in IEM or PGx, producing better variant interpretation predictions than would be possible alone (16).

In conclusion, here we present the first comprehensive functional annotation of the largest set known to date (n = 150) of missense variants in OCTN2 from the gnomAD database. For the first time to our knowledge, we show that loss-of-function for many OCTN2 variants can be attributed to failure to traffic to the plasma membrane, revealing a disease-causing mechanism with the potential to be leveraged in therapeutic strategies for the treatment of CTD. Further studies are ongoing to determine the mechanisms for improper sorting of variants to the plasma membrane, which may be leveraged for future therapies. The results of our protein-specific variant effect prediction model for OCTN2, with which we predict the function and localization of OCTN2 variant, substantially outperforms existing methods. We provide these functional and spatial predictions for all possible missense variants in OCTN2 (N=10,583), which may be useful in clinical interpretation of novel variants of uncertain significance in accordance with ACMG guidelines.

3.6 Figures

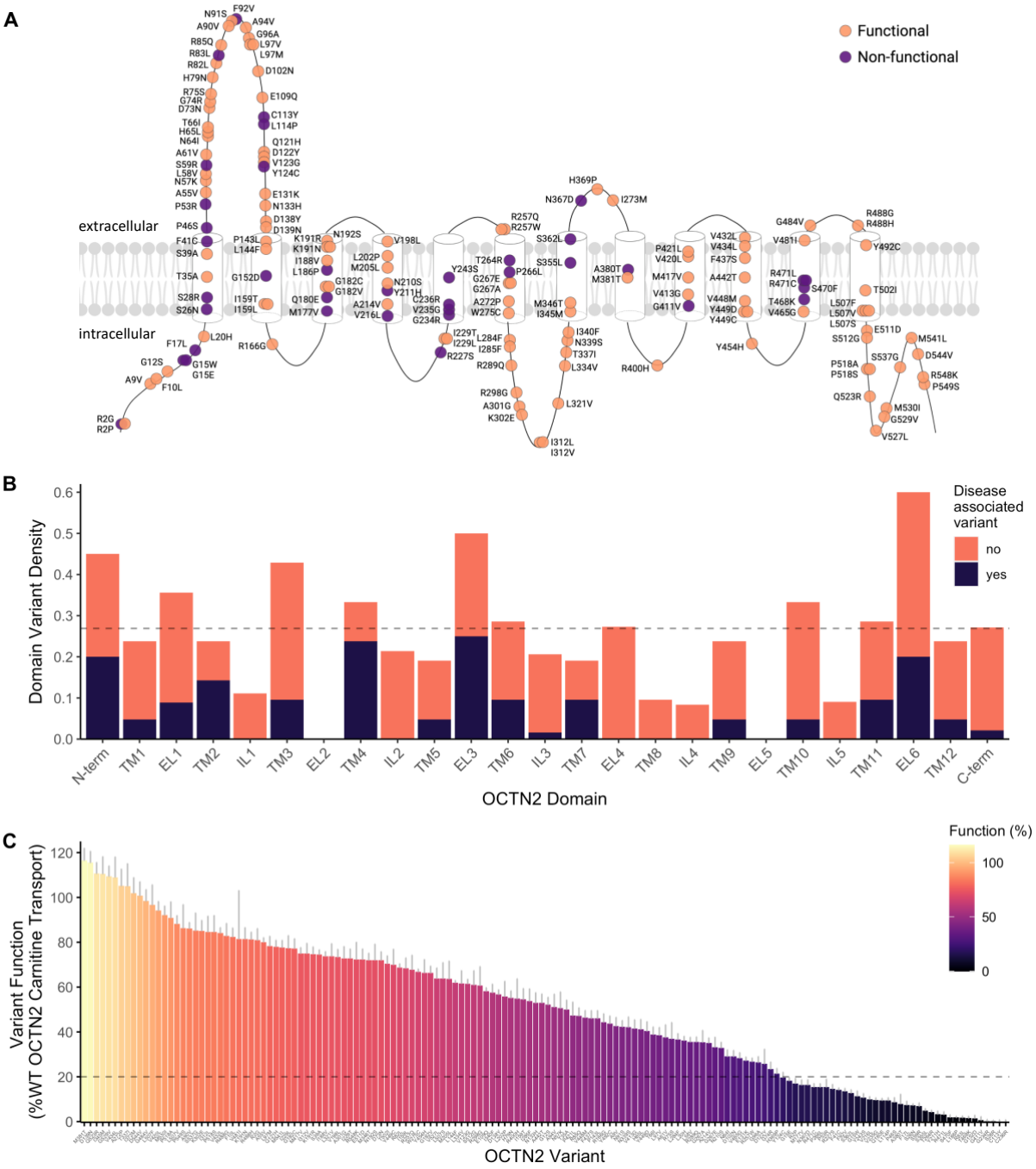


Figure 3.1. Functionally characterized OCTN2 variants. (A) Two-dimensional location of variants selected for characterization along the predicted secondary structure of OCTN2. Variants are colored by functional status, variants in orange are functional (>20% wild-type OCTN2 carnitine transport) and variants in purple are non-functional (<20% transport). **(B)** Density of variants characterized in each domain of OCTN2. Variants that have been clinically associated with CTD are shown in orange and variants that have no known clinical association are shown in purple. The dotted line represents the average density of assayed variants across all

domains. N-term=N-terminus, TM=transmembrane domain, EL=extracellular loop, IL=intracellular loop, C-term=C-terminus. **(C)** Functional characterization of 150 OCTN2 variants with respect to C-14 carnitine uptake in OCTN2-expressing HEK293T cells. Each bar represents the function of an individual OCTN2 variant represented as percentage of wild-type OCTN2 carnitine transport. Data represent mean ± SEM of three individual biological replicates all performed in triplicate.

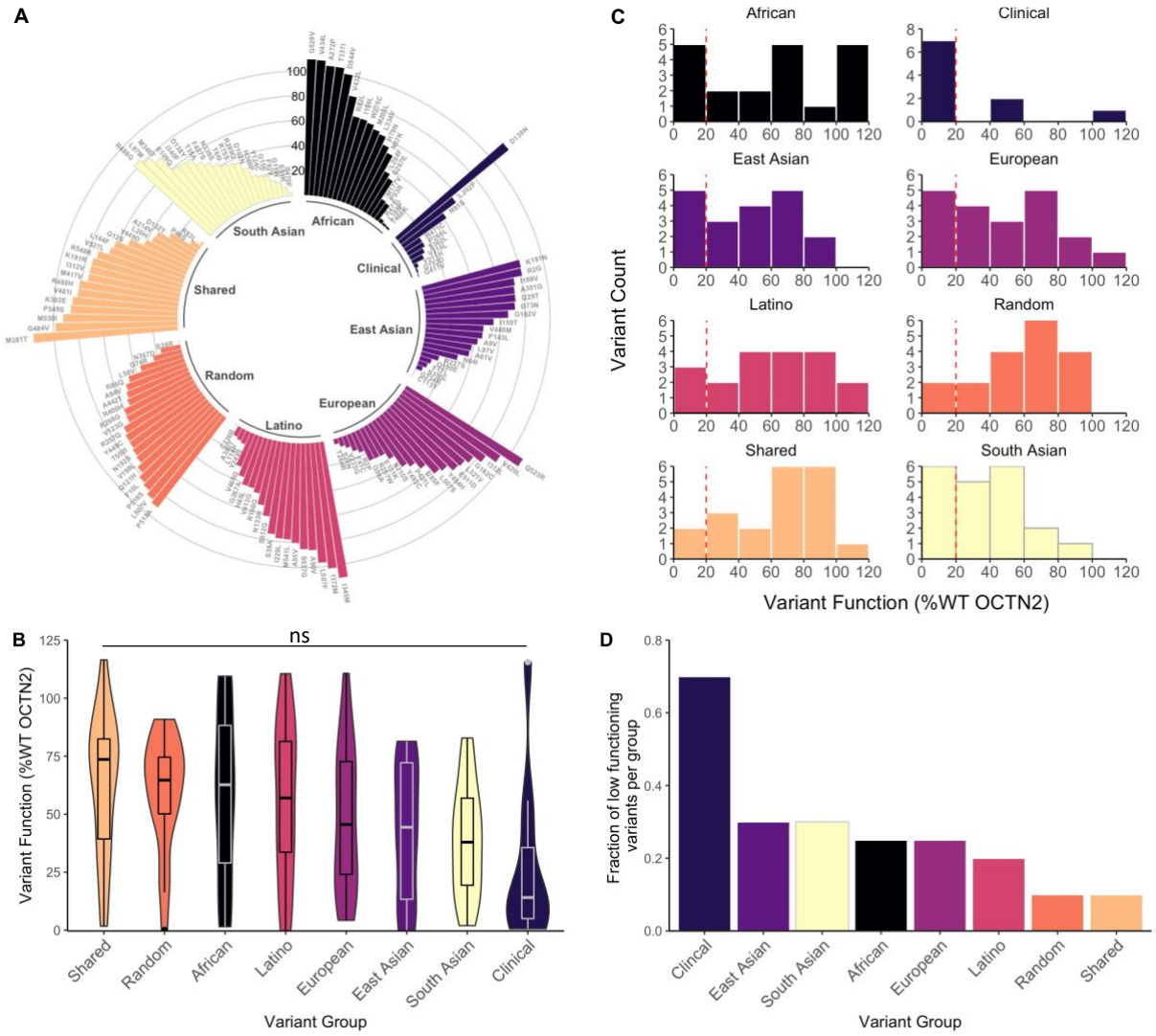


Figure 3.2. Functional distribution of variants by ancestral group. (A) Function of individual OCTN2 variants in each ancestral group. Height of bars (radial y-axis) represents variant function (%WT OCTN2 carnitine transport). **(B)** Violin plots summarize the function of variants in each group. Each embedded boxplot summarizes median, interquartile range, and whiskers in the style of Tukey. ns = not significant as determined by ANOVA with Tukey post-hoc test. **(C)** Histogram of the distribution of variants from each group into functional bins. Vertical red dotted line illustrates the cutoff of 20% function, below which variants have increased risk for CTD. **(D)** Fraction of low functioning variants (<20% WT) assayed in each variant group.

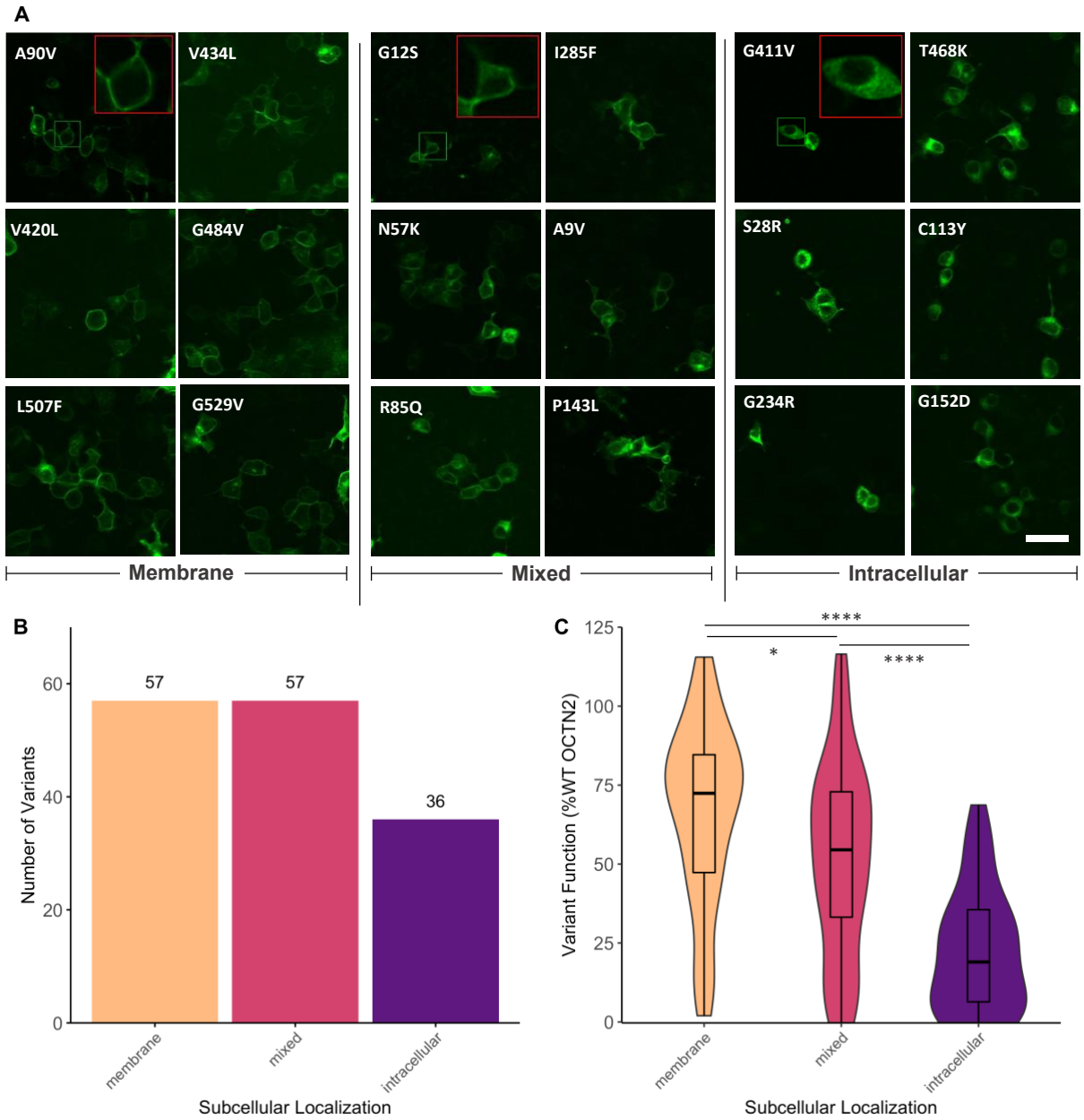


Figure 3.3. Subcellular localization of OCTN2 variants. (A) Representative images of OCTN2 variants conjugated to msfGFP in HEK293T cells. Three distinct patterns are observed: membrane localization (left panel), intracellular localization (middle pane), and mixed localization (right panel). Scale bar in lower right panel represents 50 µm and is consistent for all images. One inset is shown for each localization pattern with original area outlined in green and 3x zoom in upper right corner outlined in red. **(B)** Distribution of variants with each subcellular localization pattern. **(C)** Box plot embedded violin plots show distribution of variant function with respect to carnitine transport based on variant subcellular localization. * indicates p-value < 0.05, **** indicates p-value < 0.0001, Welch's ANOVA between means with Games-Howell post-hoc test.

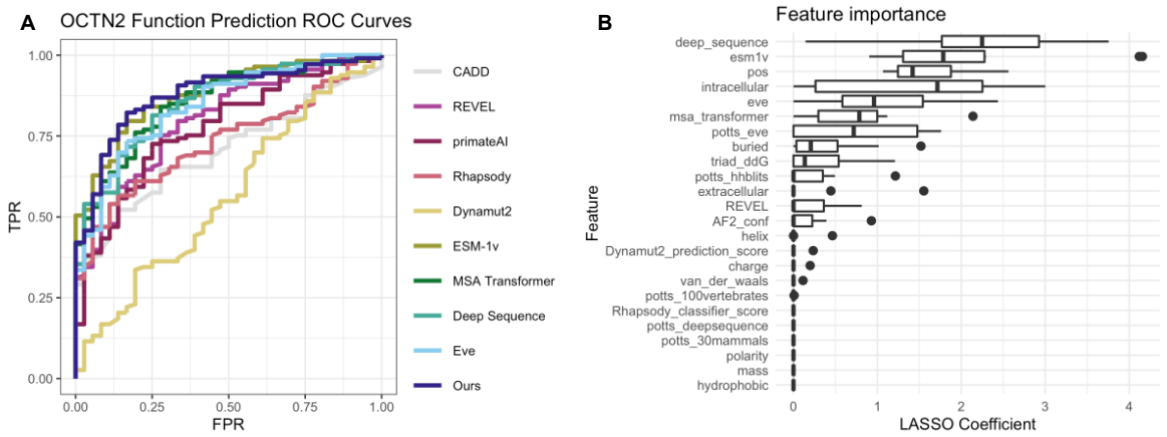


Figure 3.4. Performance of OCTN2 functional classification model. (A) Receiver operator characteristics (ROC) curve for our model compared to other variant effect prediction models. **(B)** Importance of features in performance of our model. Features are described in detail in Section 3.8 Supplementary Text.

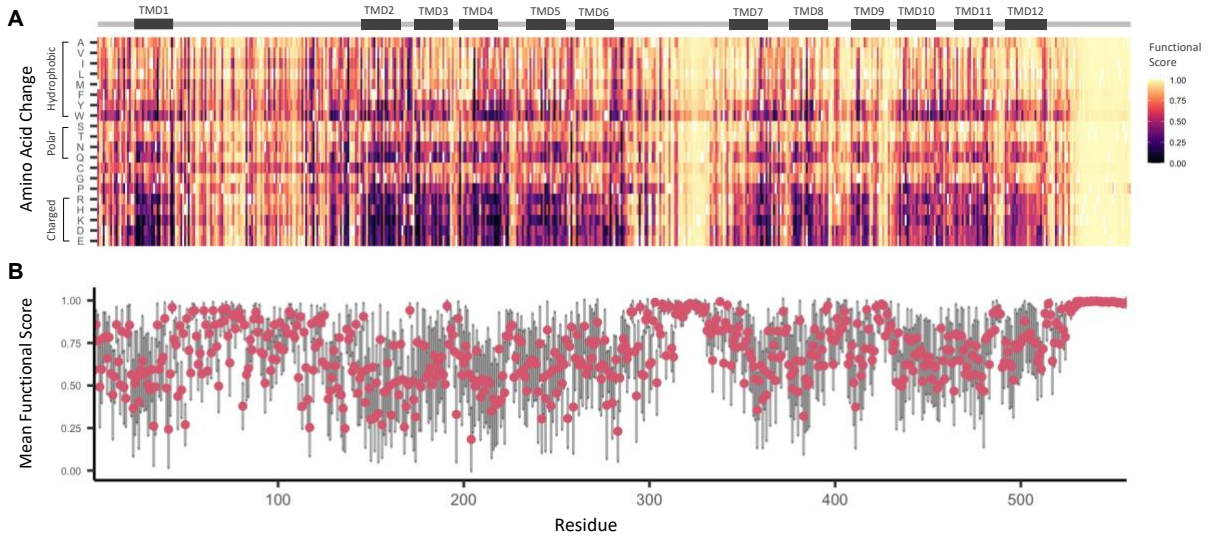
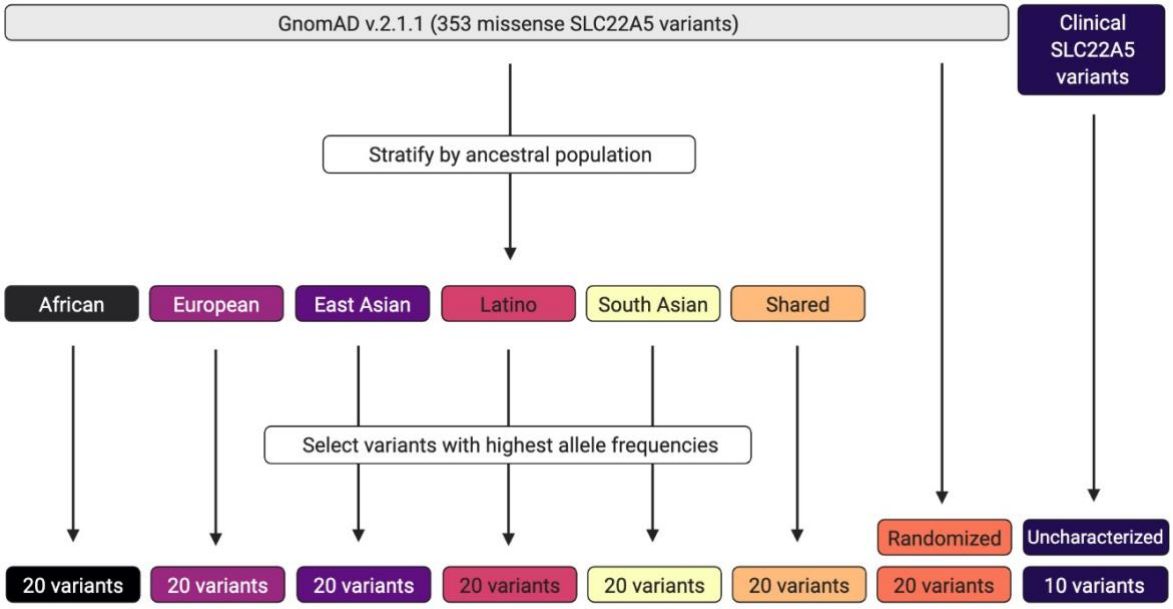
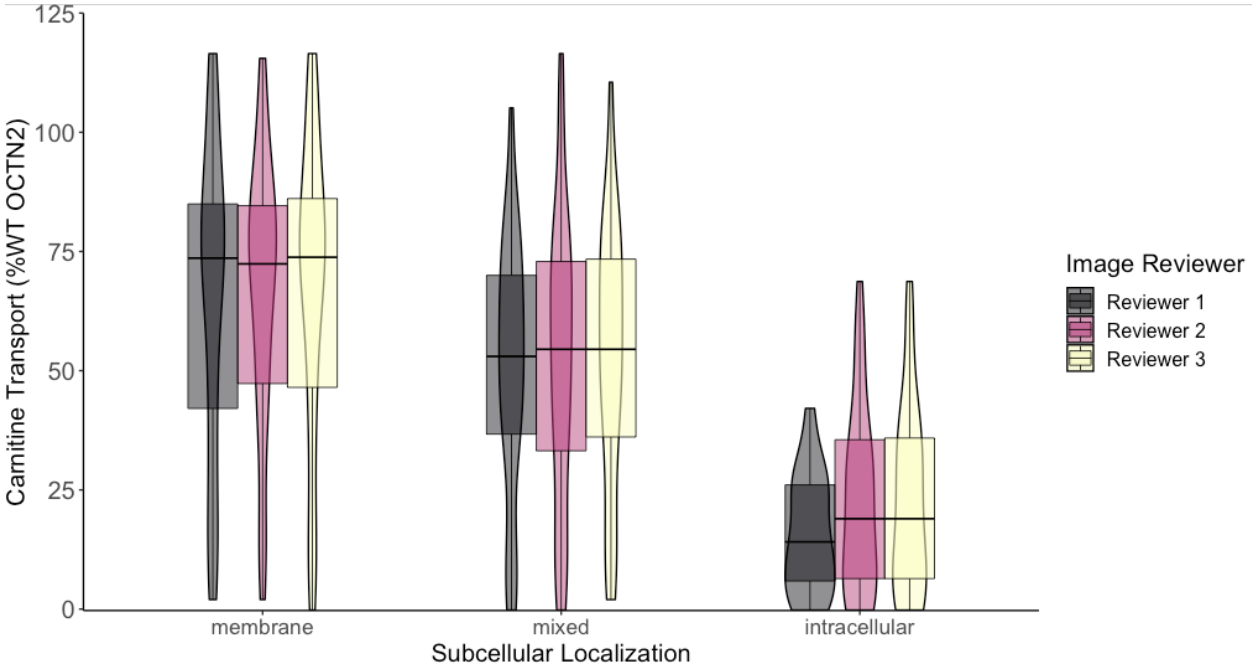


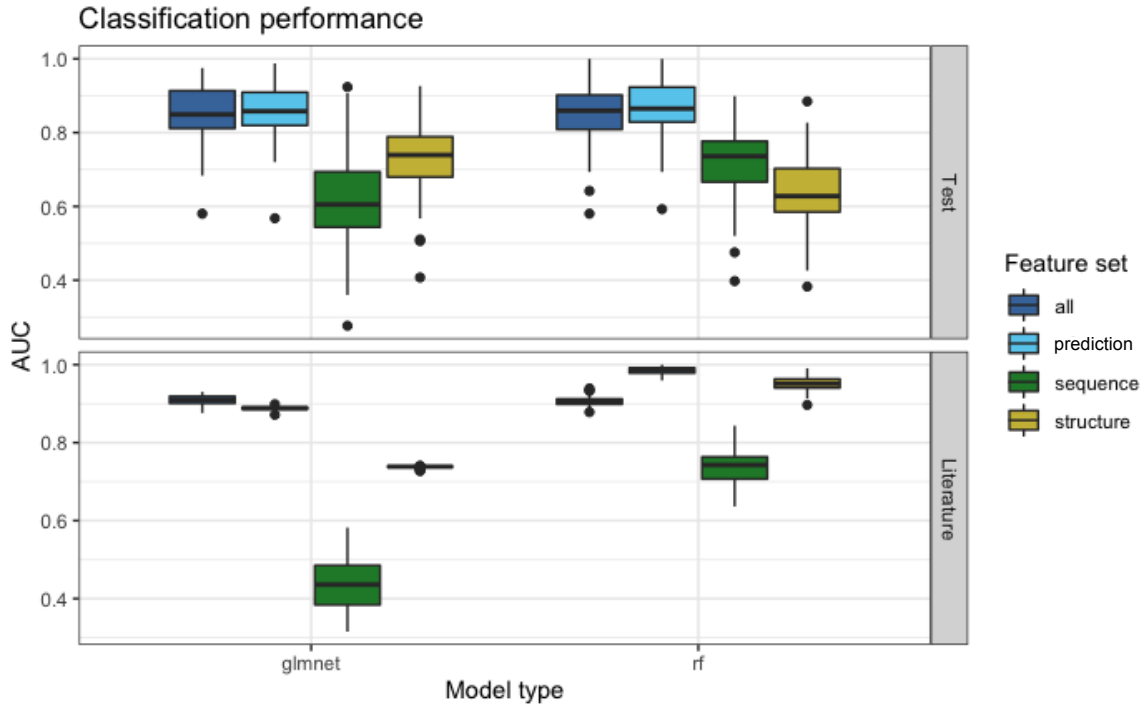
Figure 3.5. Predicted function of all possible missense variants in OCTN2. (A) Normalized functional score for all possible substitutions at every residue. Functional scores greater than 0.5 indicate function greater than 20% of wild-type OCTN2 function, with scores closer to 1 indicating increased confidence in prediction; functional scores less than 0.5 indicate function less than 20% of wild-type OCTN2, with scores closer to 0 indicating increased confidence in prediction. Reference residues are colored in white. Cartoon of OCTN2 secondary structure is aligned above heatmap, TMD = transmembrane domain. **(B)** Mean functional score for each residue position. Dots and bars represent mean and standard deviation of functional score for all residues at that position, respectively.



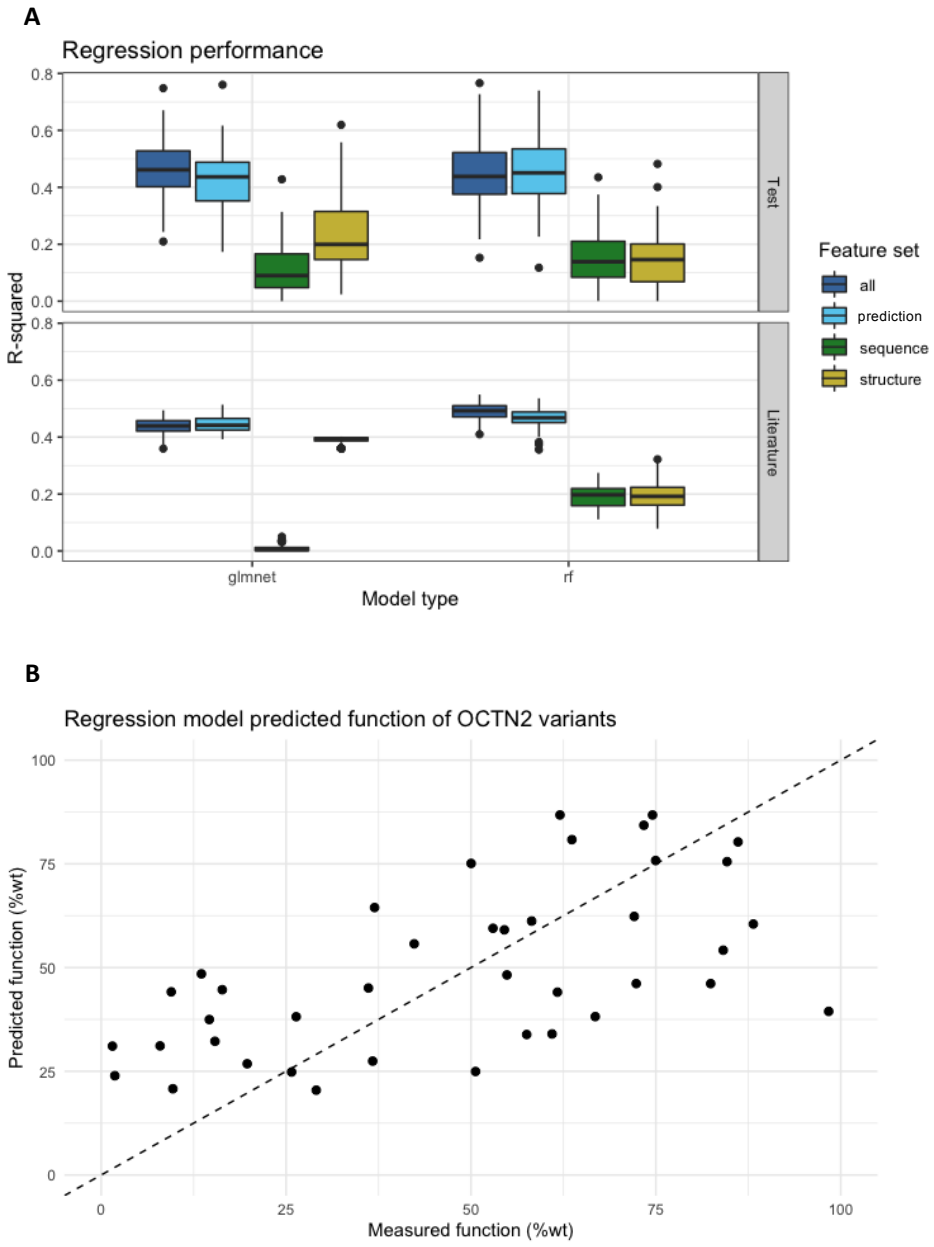
Supplementary Figure 3.1. Workflow for selection of OCTN2 variants characterized in this study. First, gnomAD variants were stratified by ancestral population in which they were identified/classified. The top 20 population-specific variants by allele frequency were selected from each of the African, Latino, East Asian, European, and South Asian populations and were exclusive to that ancestral population (i.e., not found in any other population). Twenty additional variants were selected from the "Shared" group, defined as found in at least two gnomAD populations listed above. In addition, 20 variants were selected at random from the remaining gnomAD OCTN2 missense variants, irrespective of ancestry. Finally, 10 uncharacterized variants clinically associated with diagnosed or suspected CTD were included for study.



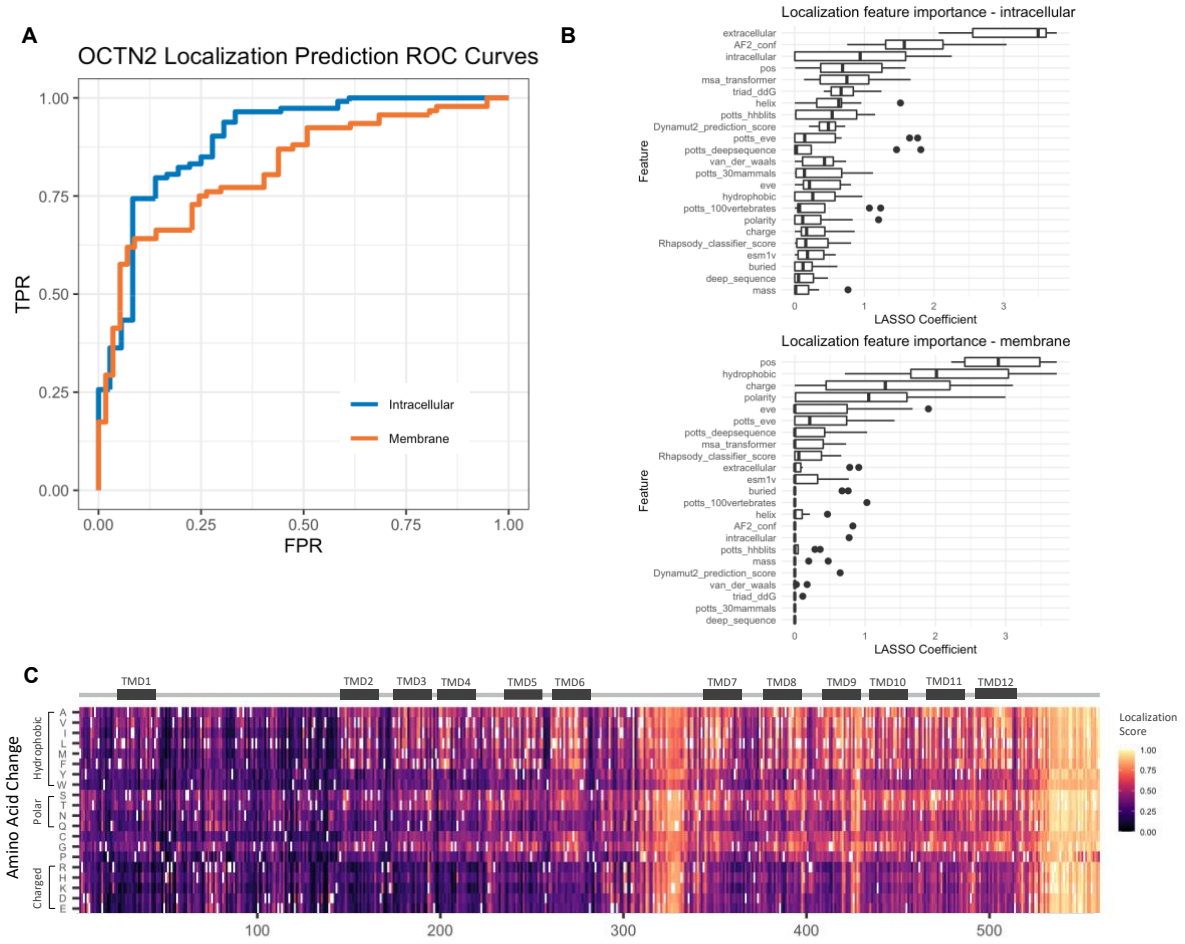
Supplementary Figure 3.2. Concordance of subcellular localization of GFP-tagged OCTN2 variants classified by three independent image reviewers blinded to the variant name or function.



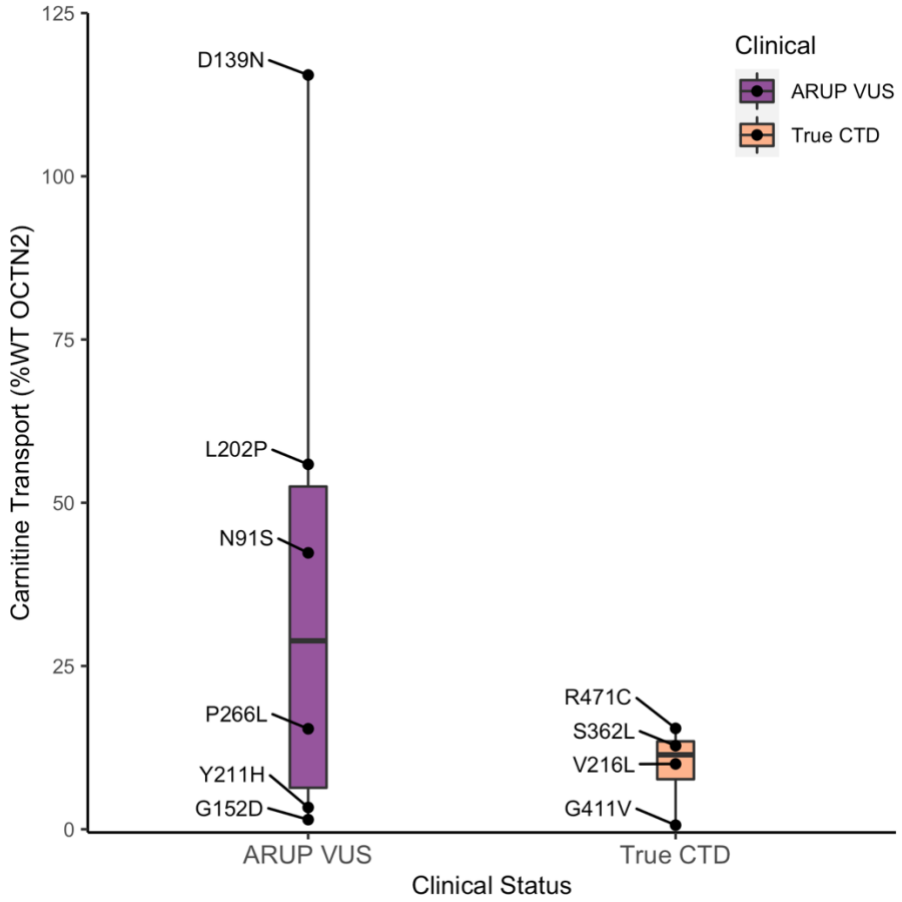
Supplementary Figure 3.3. Performance of classification machine learning models to predict OCTN2 function evaluated during model selection. Model selection was made by comparing the performance of LASSO penalized logistic regression (glmnet) and random forest (rf) models on four different feature sets: prediction derived features, sequence derived features, structure derived features, a feature set of all features combined. Every combination of model type and feature sets was evaluated by training predictive models through 100 iterations of random subsampling of the 150 characterized OCTN2 variants using 80% of the variants for training and 20% for test in each subsample. In addition, model performance was evaluated on variants from the literature. Model performance is evaluated by AUC under the ROC curve.



Supplementary Figure 3.4. Performance of regression machine learning models to predict OCTN2 function. (A) Model selection was made by comparing the performance of LASSO penalized logistic regression (glmnet) and random forest (rf) models on four different feature sets: prediction derived features, sequence derived features, structure derived features, a feature set of all features combined. Every combination of model type and feature sets was evaluated by training predictive models through 100 iterations of random subsampling of the 150 characterized OCTN2 variants using 80% of the variants for training and 20% for test in each subsample. Model performance is evaluated by the R-squared metric. **(B)** Performance of the final regression model (random forest with prediction derived features) trained on 110 variants and tested with the remaining 40 variants. Experimentally measured function is compared to model predicted function.



Supplementary Figure 3.5. Performance of machine learning models to predict OCTN2 localization. (A) Receiver operator characteristics (ROC) curve for our models. Intracellular classification notes variants predicted to have intracellular localization vs membrane or mixed localization; membrane classification notes variants predicted to have membrane localization vs intracellular or mixed localization. **(B)** Importance of features in performance of the intracellular (top panel) and membrane (bottom panel) classification models. **(C)** Normalized localization score for all possible substitutions at every residue. Localization scores greater than 0.5 indicate predicted membrane localization, with scores closer to 1 indicating increased confidence in prediction; functional scores less than 0.5 indicate predicted intracellular localization, with scores closer to 0 indicating increased confidence in prediction. Reference residues are colored in white.



Supplementary Figure 3.6. Function of the "clinical" variants added to the study in addition to the 140 variants selected from gnomAD. Variants identified in true confirmed cases of CTD are in orange, and variants in suspected cases in the ARUP database currently classified as variants of unknown significance (VUS) are in purple.

3.7 Tables

Table 3.1. Performance of the models in classification of OCTN2 variants as loss-of-function (<20%) or functional (>20%). Highest score for each metric is highlighted in bold font.

Predictor	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	MCC
Ours	**0.895 (±0.025)**	**0.856 (±0.034)**	**0.890 (±0.053)**	0.845 (±0.052)	0.653 (±0.079)	**0.964 (±0.017)**	**0.673 (±0.061)**
ESM-1v	0.879	0.805	0.796	0.833	**0.938**	0.566	0.563
MSA Transformer	0.854	0.772	0.761	0.806	0.925	0.518	0.501
Deep Sequence	0.854	0.799	0.814	0.750	0.911	0.563	0.517
EVE	0.845	0.758	0.833	0.735	0.50	0.933	0.496
REVEL	0.797	0.745	0.722	0.752	0.481	0.895	0.422
primateAI	0.774	0.732	0.750	0.726	0.466	0.901	0.418
Polyphen	0.753	0.779	0.556	**0.850**	0.541	0.857	0.401
Rhapsody	0.734	0.624	0.889	0.540	0.381	0.938	0.370
CADD	0.702	0.604	0.861	0.522	0.365	0.922	0.331
Dynamut2	0.563	0.450	0.336	0.806	0.844	0.279	0.132

Supplementary Table 3.1. Constructs from the Mammalian Toolkit used in the generation of SLC22A5 constructs in this study (25).

MTK part type	Description	DNA topology	Resistance
MTK1_001	Encodes ConS connector	Circular	Chloramphenicol
MTK2_023	Encodes pCMV-H promoter	Circular	Chloramphenicol
MTK3a_030	Encodes start codon and 6xHIS 3xFLAG	Circular	Chloramphenicol
MTK0_027-SLC22A5 (part 3b)	Encodes SLC22A5	Circular	Chloramphenicol
MTK4a_015	Encodes msfGFP	Circular	Chloramphenicol
MTK4b_001	Encodes BghPA	Circular	Chloramphenicol
MTK5_006	Encodes ConRE connector	Circular	Chloramphenicol
MTK678_001	Encodes ColE1-AmpR backbone vector	Circular	Ampicillin
MTK0_017	Encodes BxB1 attB KanR destination vector	Circular	Kanamycin
MTK0_027	Encodes Part Entry Vector	Circular	Chloramphenicol
JPF0335	Encodes BxB1 recombinase (pCAG_NLS_HA_Bxb1_Addgene51271)	Circular	Ampicillin
SLC22A5 transcriptional unit	SLC22A5 CDS with adaptor sequences for Golden Gate Cloning	Linear	NA

Supplementary Table 3.2. Machine learning features.

Feature set	Feature
Sequence-based	Change in polarity
	Change in charge
	Change in mass
	Change in hydrophobicity
	Buried accessible surface area
	Van del Waals forces
	Helix potential
	Residue position
	2D-structural domain (intracellular loop, extracellular loop, transmembrane domain)
Structure-based	Solvent access area
	Solvent access relative score
	Network centrality degree
	Network centrality cluster coefficient
	Network centrality closeness
	Network centrality betweenness
	Network centrality eigenvector centrality
	Network centrality average neighbor degree
	AlphaHelix/turn/coil
Prediction-based	Triad ddG
	AF2 confidence
	ESM-1v
	MSA Transformer
	DeepSequence
	EVE
	Triad DDG
	AF2 confidence
	REVEL
	CADD
	Rhapsody score
	DynaMut2 score
	Potts_EVE
	Potts_hhblits
	Potts_100vertebrates
	Potts_deepsequence
	Potts_30mammals

3.8 Supplementary Text

Methods

Feature Generation

Allele frequency was calculated from gnomAD (24). Pathogenicity scores for all assayed OCTN2 variants were obtained from PolyPhen-2 (38), CADD (40), REVEL (36), PrimateAI (37). As no crystal structure has been solved for OCTN2, we used the AlphaFold2 predicted structure (Entry O76082) (27) for generation of structural features. Predicted stability change ($\Delta\Delta G^{stability}$) was generated with DynaMut2 (41). Solvent accessibility was generated using GETAREA (64). Prediction of variant effect on protein dynamics was generated using Rhapsody (39). Network centrality analysis (centrality degree, centrality cluster coefficient, centrality closeness, centrality betweenness, eigenvector centrality, average neighbor degree) was generated with Network Analysis of Protein Structure (NAPS) (65). Generation of other features is described below.

pLDDT and triad_ddG scores

The AlphaFold 2 predicted structure of human SLC22A5 was obtained from the AlphaFold (27) protein structure database. pLDDT scores that represent model confidence for each position in the structure were extracted as a set of features. pLDDT is continuously valued between 0 and 100, and is a prediction of the lDDT-Ca score (66) that is used to compare two models by reporting on distances between their Ca atoms at equivalent positions. The plDDT score reported by AlphaFold is a learned prediction that was calibrated using distance from the ground-truth structure during model training.

The AlphaFold 2-predicted model was then directly used as input for ddG calculations using Triad Protabit design software (https://triad.protabit.com). The structure was first standardized using the Standardize Structure App within Triad, and then single mutant stability scores (ddG) were calculated with default parameters (float distance = 7Å and backbone only scoring = off) using the Rosetta scoring function. These parameters ensure that local interactions around the mutation site are repacked using rosetta prior to delta G calculation.

Modeling variant effects with models of evolutionary data

In order to predict the effects of mutations, we utilized conservation information derived from evolutionary homologs of the transporter SLC22A5. The state-of-the-art unsupervised variant effect predictors (that is, predictors that have not been trained with any variant screening data) fit a statistical model on a set of related, functional sequences. These models can take the form of protein language models (31), variational autoencoders (VAEs) (29) and Potts models (30). The likelihood of a sequence under the derived statistical models has been shown to correlate well with the probability that the sequence is functional.

The features of our ensemble model include previously published methods for variant effect prediction and additional features, some of which we designed ourselves. The pre-existing methods for variant effect prediction that we used are DeepSequence (29), EVE (43), ESM-1v (31), and MSA Transformer (42). DeepSequence, EVE, and MSA Transformer require an MSA to make predictions, while ESM-1v requires only a sequence. DeepSequence and MSA Transformer features were constructed using the "DeepSequence" from "Constructing Alignments." EVE features use EVE alignments.

Constructing alignments

To combine the advantages of alignments that represent different evolutionary timescales, we used 5 different alignments for downstream training. We combined deep, diverged alignments (Alignments 1, 2, 3), which we hypothesized would include coarse-grained fold information, with more alignments that sample more closely-related organisms (Alignments 4, 5).

1. HHblits: To produce this alignment, we reimplemented the alignment generation procedure described in (67). Our implementation can be run using the `mogwai-align` command in https://github.com/nickbhat/mogwai. We used UniprotKB O76082 as the query sequence**.** HHblits performed 1 iteration searching against Uniclust30 with an e-value 1e-80 to produce an alignment of 7201 sequences.
2. EVE: The EVE alignment was downloaded directly from https://evemodel.org/download/protein/S22A5_HUMAN.
3. Deepsequence: We follow the approach in (30, 31) to form a deep alignment. The query sequence was searched against the UniRef100 database using the profile HMM homology search tool jackhmmer (68). Non-redundant sequences were kept with a 0.8 sequence similarity threshold. This resulted in an alignment with 7205 sequences. The alignment script can be found at https://github.com/rmrao/DeepSequence/blob/master/align.py
4. 100 vertebrates: FASTA alignments for coding regions of the UCSC Known Genes corresponding to the human reference genome (hg38/GRCh38, Feb. 2009) aligned to 100 vertebrate genome assemblies as described in http://genomewiki.ucsc.edu/index.php/Hg38_100-way_conservation_alignment were downloaded from

http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/alignments/knownCanonical.multiz100way.protAA.fa.gz and subset to those fasta records with the header prefix corresponding to the ENSEMBL transcript name for SLC22A5, "ENST00000245407."

5. 30 mammals: FASTA alignments for coding regions of the UCSC Known Genes corresponding to the human reference genome (hg38/GRCh38, Feb. 2009) aligned to 30 mammalian genome assemblies as described in http://genomewiki.ucsc.edu/index.php/Hg38_30-way_conservation_alignment were downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz30way/alignments/knownCanonical.multiz30way.protAA.fa.gz and subset to those fasta records with the header prefix corresponding to the ENSEMBL transcript name for SLC22A5, "ENST00000245407."

All of these alignments can be found at https://github.com/songlab-cal/slc22a5.

Modeling variant effects with Potts models

A Potts model is an undirected Markov Random Field model which has been shown to capture information about protein structure (69, 70) as well as protein function (30, 71). Adding the unsupervised likelihood as a feature has been shown to improve the performance of regression models used to predict protein function (72, 73).

In the Potts model, which models marginal effects and pairwise interactions, the likelihood of a sequence x is given by:

$$E(x) = \sum_i h_i(x(i)) + \sum_{i<j} J_{ij}(x(i), x(j))$$

$$\mathcal{L}(x) = \frac{1}{Z} \exp\left(-E(x)\right)$$

Where Z is the partition function.

To use a Potts model to predict variant effects, we compute the energy difference between the variant and the wildtype reference sequence, in this case the human reference gene SLC22A5.

$$\Delta E(x) = E(x) - E(x_0)$$

Note that this variant effect can be computed without computing the partition function Z.

Separate Potts models were fit on each of the 5 alignments described in "Constructing Alignments." Each model was used separately to predict variant effects by computing the energy difference between the variant and the wildtype reference sequence.

Fitting the Potts model

Due to the combinatorial complexity of computing the partition function Z, we cannot maximize the true likelihood of the sequences. Instead, we estimate the coupling parameters J and the marginal effects h to maximize the pseudolikelihood, which follows the established approach in (30, 71, 74). For the optimization, we use a modified version of Adam (75) which ties together all squared updates. The Potts model implementation we used can be found in our open source MRF library https://github.com/nickbhat/mogwai. All Potts models were trained on a single NVIDIA RTX 2080 Ti GPU for 5000 gradient update steps, with a batch size of 4096 sequences and a learning rate of 0.5.

3.9 Supplementary Files

Four additional datasets are available in a separate file.

Supplementary Dataset 3.1 (separate file). OCTN2 variants with function reported in the literature not assayed in our study. These literature variants were used in evaluation of machine learning models as an additional test/validation set. 82 unique variants had reported function. 12 variants had multiple measurements reported by different publications, for a total of 94 entries. Function was averaged for variants with multiple measurements.

Supplementary Dataset 3.2 (separate file). OCTN2 variants characterized in the study, including function, localization, statistical significance, associated features.

Supplementary Dataset 3.3 (separate file). Functional predictions for all 10,583 missense variants in OCTN2. "mean_pred" is a score that represents the probability that the variant is functional (>20% WT OCTN2 carnitine transport). The "deleterious" column binarizes the mean_pred score based on a cutoff that maximizes specificity and sensitivity. A deleterious score of 0 indicates variant is functional; a score of 1 indicates the variant is prediction to be loss-of-function (<20% WT OCTN2 carnitine transport).

Supplementary Dataset 3.4 (separate file). Clinical variants assayed in this study identified in individuals with confirmed or suspected CTD.

3.10 References

1. Lin L, Yee SW, Kim RB, Giacomini KM. SLC transporters as therapeutic targets: emerging opportunities. Nat Rev Drug Discov. 2015;14(8):543-60.
2. Ferreira CR, van Karnebeek CDM, Vockley J, Blau N. A proposed nosology of inborn errors of metabolism. Genet Med. 2019;21(1):102-6.
3. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7.
4. Rinaldo P, Stanley CA, Hsu BY, Sanchez LA, Stern HJ. Sudden neonatal death in carnitine transporter deficiency. J Pediatr. 1997;131(2):304-5.
5. Rasmussen J, Duno M, Lund AM, Steuerwald U, Hansen SH, Joensen HD, et al. Increased risk of sudden death in untreated primary carnitine deficiency. J Inherit Metab Dis. 2020;43(2):290-6.
6. Gelinas R, Leach E, Horvath G, Laksman Z. Molecular Autopsy Implicates Primary Carnitine Deficiency in Sudden Unexplained Death and Reversible Short QT Syndrome. Can J Cardiol. 2019;35(9):1256 e1- e2.
7. Cederbaum SD, Koo-McCoy S, Tein I, Hsu BY, Ganguly A, Vilain E, et al. Carnitine membrane transporter deficiency: a long-term follow up and OCTN2 mutation in the first documented case of primary carnitine deficiency. Mol Genet Metab. 2002;77(3):195-201.
8. Lin Y, Zhang W, Huang C, Lin C, Lin W, Peng W, et al. Increased detection of primary carnitine deficiency through second-tier newborn genetic screening. Orphanet J Rare Dis. 2021;16(1):149.

9. Gallant NM, Leydiker K, Wilnai Y, Lee C, Lorey F, Feuchtbaum L, et al. Biochemical characteristics of newborns with carnitine transporter defect identified by newborn screening in California. Mol Genet Metab. 2017;122(3):76-84.

10. Wilson C, Knoll D, de Hora M, Kyle C, Glamuzina E, Webster D. The decision to discontinue screening for carnitine uptake disorder in New Zealand. J Inherit Metab Dis. 2019;42(1):86-92.

11. Schiergens KA, Weiss KJ, Roschinger W, Lotz-Havla AS, Schmitt J, Dalla Pozza R, et al. Newborn screening for carnitine transporter defect in Bavaria and the long-term follow-up of the identified newborns and mothers: Assessing the benefit and possible harm based on 19 (1/2) years of experience. Mol Genet Metab Rep. 2021;28:100776.

12. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D7.

13. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.

14. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.

15. Cohn EG, Hamilton N, Larson EL, Williams JK. Self-reported race and ethnicity of US biobank participants compared to the US Census. J Community Genet. 2017;8(3):229-38.

16. McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021;108(4):535-48.

17. Genevieve LD, Martani A, Shaw D, Elger BS, Wangmo T. Structural racism in precision medicine: leaving no one behind. BMC Med Ethics. 2020;21(1):17.

18. Kroncke BM, Duran AM, Mendenhall JL, Meiler J, Blume JD, Sanders CR. Documentation of an Imperative To Improve Methods for Predicting Membrane Protein Stability. Biochemistry. 2016;55(36):5002-9.

19. Li B, Mendenhall JL, Kroncke BM, Taylor KC, Huang H, Smith DK, et al. Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance. Circ Cardiovasc Genet. 2017;10(5).

20. Clerx M, Heijman J, Collins P, Volders PGA. Predicting changes to INa from missense mutations in human SCN5A. Sci Rep. 2018;8(1):12797.

21. Hart SN, Polley EC, Shimelis H, Yadav S, Couch FJ. Prediction of the functional impact of missense variants in BRCA1 and BRCA2 with BRCA-ML. NPJ Breast Cancer. 2020;6:13.

22. Adhikari AN. Gene-specific features enhance interpretation of mutational impact on acid alpha-glucosidase enzyme activity. Hum Mutat. 2019;40(9):1507-18.

23. Del Angel G, Reynders J, Negron C, Steinbrecher T, Mornet E. Large-scale in vitro functional testing and novel variant scoring via protein modeling provide insights into alkaline phosphatase activity in hypophosphatasia. Hum Mutat. 2020;41(7):1250-62.

24. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

25. Fonseca JP, Bonny AR, Kumar GR, Ng AH, Town J, Wu QC, et al. A Toolkit for Rapid Modular Construction of Biological Circuits in Mammalian Cells. ACS Synth Biol. 2019;8(11):2593-606.

26. Wang Y, Meadows TA, Longo N. Abnormal sodium stimulation of carnitine transport in primary carnitine deficiency. J Biol Chem. 2000;275(27):20782-6.

27. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

28. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50(D1):D439-D44.

29. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018;15(10):816-22.

30. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128-35.

31. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv. 2021:2021.07.09.450648.

32. SLC22A5 Database [Internet]. 2022. Available from: http://arup.utah.edu/database/OCTN2/OCTN2_welcome.php.

33. Frigeni M, Balakrishnan B, Yin X, Calderon FRO, Mao R, Pasquali M, et al. Functional and molecular studies in primary carnitine deficiency. Hum Mutat. 2017;38(12):1684-99.

34. Urban TJ, Gallagher RC, Brown C, Castro RA, Lagpacan LL, Brett CM, et al. Functional genetic diversity in the high-affinity carnitine transporter OCTN2 (SLC22A5). Mol Pharmacol. 2006;70(5):1602-11.

35. Rose EC, di San Filippo CA, Ndukwe Erlingsson UC, Ardon O, Pasquali M, Longo N. Genotype-phenotype correlation in primary carnitine deficiency. Hum Mutat. 2012;33(1):118-23.

36. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85.

37. Sundaram L, Gao H, Padigepati S, McRae J, Li Y, Kosmicki J, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-70.

38. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7 20.

39. Ponzoni L, Penaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. 2020;36(10):3084-92.

40. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.

41. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Sci. 2021;30(1):60-9.

42. Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, et al. MSA Transformer. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research. 139: PMLR; 2021.

43. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91-5.

44. Brnich SE, Rivera-Munoz EA, Berg JS. Quantifying the potential of functional evidence to reclassify variants of uncertain significance in the categorical and Bayesian interpretation frameworks. Hum Mutat. 2018;39(11):1531-41.

45. Rasmussen J, Kober L, Lund AM, Nielsen OW. Primary Carnitine deficiency in the Faroe Islands: health and cardiac status in 76 adult patients diagnosed by screening. J Inherit Metab Dis. 2014;37(2):223-30.

46. Yang X, Li Q, Wang F, Yan L, Zhuang D, Qiu H, et al. Newborn Screening and Genetic Analysis Identify Six Novel Genetic Variants for Primary Carnitine Deficiency in Ningbo Area, China. Front Genet. 2021;12:686137.

47. Koizumi A, Nozaki J, Ohura T, Kayo T, Wada Y, Nezu J, et al. Genetic epidemiology of the carnitine transporter OCTN2 gene in a Japanese population and phenotypic characterization in Japanese pedigrees with primary systemic carnitine deficiency. Hum Mol Genet. 1999;8(12):2247-54.

48. Amat di San Filippo C, Pasquali M, Longo N. Pharmacological rescue of carnitine transport in primary carnitine deficiency. Hum Mutat. 2006;27(6):513-23.

49. Baatallah N, Elbahnsi A, Mornon JP, Chevalier B, Pranke I, Servel N, et al. Pharmacological chaperones improve intra-domain stability and inter-domain assembly via distinct binding sites to rescue misfolded CFTR. Cell Mol Life Sci. 2021;78(23):7813-29.

50. Pampalone G, Grottelli S, Gatticchi L, Lombardi EM, Bellezza I, Cellini B. Role of misfolding in rare enzymatic deficits and use of pharmacological chaperones as therapeutic approach. Front Biosci (Landmark Ed). 2021;26(12):1627-42.

51. Bhat S, Newman AH, Freissmuth M. How to rescue misfolded SERT, DAT and NET: targeting conformational intermediates with atypical inhibitors and partial releasers. Biochem Soc Trans. 2019;47(3):861-74.

52. Ohashi R, Tamai I, Inano A, Katsura M, Sai Y, Nezu J, et al. Studies on functional sites of organic cation/carnitine transporter OCTN2 (SLC22A5) using a Ser467Cys mutant protein. J Pharmacol Exp Ther. 2002;302(3):1286-94.

53. Zhang L, Sarangi V, Ho MF, Moon I, Kalari KR, Wang L, et al. SLCO1B1: Application and Limitations of Deep Mutational Scanning for Genomic Missense Variant Function. Drug Metab Dispos. 2021;49(5):395-404.

54. Wang Y, Taroni F, Garavaglia B, Longo N. Functional analysis of mutations in the OCTN2 transporter causing primary carnitine deficiency: lack of genotype-phenotype correlation. Hum Mutat. 2000;16(5):401-7.

55. Wang Y, Korman SH, Ye J, Gargus JJ, Gutman A, Taroni F, et al. Phenotype and genotype variation in primary carnitine deficiency. Genet Med. 2001;3(6):387-92.

56. Lamhonwah AM, Olpin SE, Pollitt RJ, Vianey-Saban C, Divry P, Guffon N, et al. Novel OCTN2 mutations: no genotype-phenotype correlations: early carnitine therapy prevents cardiomyopathy. Am J Med Genet. 2002;111(3):271-84.

57. Rasmussen J, Lund AM, Risom L, Wibrand F, Gislason H, Nielsen OW, et al. Residual OCTN2 transporter activity, carnitine levels and symptoms correlate in patients with primary carnitine deficiency. Mol Genet Metab Rep. 2014;1:241-8.

58. Wolf SM, Burke W, Koenig BA. Mapping the Ethics of Translational Genomics: Situating Return of Results and Navigating the Research-Clinical Divide. J Law Med Ethics. 2015;43(3):486-501.

59. Halley MC, Young JL, Fernandez L, Kohler JN, Undiagnosed Diseases N, Bernstein JA, et al. Perceived utility and disutility of genomic sequencing for pediatric patients: Perspectives from parents with diverse sociodemographic characteristics. Am J Med Genet A. 2022;188(4):1088-101.

60. Shabani M, Dyke SOM, Marelli L, Borry P. Variant data sharing by clinical laboratories through public databases: consent, privacy and further contact for research policies. Genet Med. 2019;21(5):1031-7.

61. Rush A, Battisti R, Barton B, Catchpoole D. Opinions of Young Adults on Re-Consenting for Biobanking. J Pediatr. 2015;167(4):925-30.

62. Acmg Board Of D. Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet Med. 2017;19(7):721-2.

63. Tabor HK, Goldenberg A. What Precision Medicine Can Learn from Rare Genetic Disease Research and Translation. AMA J Ethics. 2018;20(9):E834-40.

64. Fraczkiewicz R. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comput Chem. 1998;19(3):319-33.

65. Chakrabarty B, Parekh N. NAPS: Network Analysis of Protein Structures. Nucleic acids research. 2016;44(W1):W375-82.

66. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29(21):2722-8.

67. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A. 2020;117(3):1496-503.

68. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7(10):e1002195.

69. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Elife. 2014;3:e02030.

70. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One. 2011;6(12):e28766.

71. Figliuzzi M, Barrat-Charlaix P, Weigt M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? Mol Biol Evol. 2018;35(7):1821.

72. Hsu C, Nisonoff H, Fannjiang C, Listgarten J. Learning protein fitness models from evolutionary and assay-labeled data. Nat Biotechnol. 2022.

73. Russ WP, Figliuzzi M, Stocker C, Barrat-Charlaix P, Socolich M, Kast P, et al. An evolution-based model for designing chorismate mutase enzymes. Science. 2020;369(6502):440-5.

74. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A. 2013;110(39):15674-9.

75. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv. 2014.

Chapter 4: Development of a functional screening platform for deep mutational scanning of the Organic Cation Transporter 1 (OCT1, *SLC22A1*)

4.1 Abstract

The organic cation transporter 1 (OCT1), encoded by *SLC22A1*, is a poly-specific membrane transporter with important pharmacogenetic implications. Current biological understanding of the mechanisms of OCT1 transport, poly-specificity, expression, and localization is limited, in part due to the lack of a crystal structure for OCT1 or any substantially homologous proteins. Polymorphisms in OCT1 can have substrate-specific effects and are known to influence exposure, efficacy, and adverse events for many FDA-approved drugs from diverse classes. Functional data is lacking, especially for less-common variants, though rare variation is sure to contribute to interindividual differences in pharmacokinetics and pharmacodynamics of drugs that are substrates of OCT1. Deep mutational scanning (DMS) is an emerging approach that utilizes next-generation sequencing (NGS) techniques to match thousands of variants to their respective function as determined by a number of functional assays. In this study, we established a screening platform for DMS of OCT1 and generated a library of 11,572 OCT1 variants comprising all possible missense variants and single amino acid deletions. We validated the function of our HEK293T landing-pad based system expressing wild-type OCT1 and variants p.R61C, p.P117L, and p.G401S and demonstrated sensitivity to detect significant differences in uptake of fluorescent (ASP+) and radiolabeled (MPP+ and metformin) substrates in cells expressing reduced-function variants. We established a cytotoxicity-based assay using the OCT1 substrate and anti-cancer platinum analog SM85 that can be used to select for loss-of-function variants identifiable by NGS. We propose that DMS of OCT1 will be critical in gaining understanding of OCT1 transporter biology and identifying variants significantly affecting the function of OCT1 that have the potential to be clinically actionable in the implementation of pharmacogenetic data in dosing of drugs that are OCT1 substrates.

4.2 Introduction

Genetic variants in more than 20 genes are known to influence exposure and/or response to over 80 medications (1). Gene-drug relationships are relevant in three major arms of the pharmaceutical world: 1) in drug target discovery whereby variation in genes/proteins associates with a disease phenotype and suggests potential drug targets for disease treatment, 2) in clinical pharmacogenetic studies conducted by the drug development sector to evaluate investigational new drug safety and efficacy in individuals with different genotypes, and 3) in the post-market stage whereby pharmacogenetic data can be used to inform dosing of approved drugs in clinical practice. In the first example, recent studies suggest that drugs designed for genetically-validated targets are more successful in Phase II and Phase III clinical trials (2). In the second instance, the FDA and other regulatory agencies provide guidance documents (3) for the pharmaceutical industry for the conduct and evaluation of clinical pharmacogenomics studies. In the third case, the Clinical Pharmacogenetics Implementation Consortium (CPIC) provides guidelines to interpret genetic variation in pharmacogenes and translate this information into actionable prescribing recommendations for relevant drugs (4).

Pharmacogenes encode proteins, typically enzymes and transporters, which, in general, influence the metabolism or disposition of drugs. Numerous transporters in the ATP-binding cassette (ABC) and Solute Carrier (SLC) transporter superfamilies are encoded by pharmacogenes. Clinical pharmacogenetic studies are recommended to investigate drug-variant interactions for new drugs that are substrates of transporters with well-established polymorphisms, such as OATP1B1 (*SLCO1B1*)(5). Additionally, nine transporters are known to be mediators of clinically-relevant drug-drug-interactions (DDIs): P-glycoprotein (P-gp, *ABCB1*), breast cancer resistance protein (BCRP, *ABCG2*), organic anion transporting polypeptide 1B1 and 1B3

(OATP1B3, *SLCO1B3*), organic anion transporters 1 (OAT1/*SLC22A6*) and 3 (OAT3/*SLC22A8*), multidrug and toxin exclusion proteins (MATE1/*SLC47A1* and MATE2K/*SLC47A2*), and organic cation transporter 2 (OCT2/*SLC22A2*) (6). Emerging evidence suggests additional transporters with DDI and pharmacogenetic relevance, including OCT1 (7).

SLC22A1, encoding the organic cation transporter 1 (OCT1), is a pharmacogene with polymorphisms known to associate with exposure, response, and/or side effects for a number of drugs with varied structures across diverse classes (7, 8) (Fig 4.1). Reduced function OCT1 variants have been reported to have conflicting effects on hepatic exposure and pharmacodynamics of the antidiabetic drug metformin (7), increase systemic exposure of the antimigraine drug sumatriptan (9), increase maximal plasma concentration (C_{max}) and efficacy of antiemetic drugs ondansetron and tropisetron (10), decrease clearance (11) and increase side effects of the analgesic drug morphine (12), increase surrogate markers of efficacy of the analgesic drug tramadol (13) and increase C_{max} of its active metabolite O-desmethyltramadol (14), and increase exposure and side effects of the anti-asthmatic fenoterol (15).

To date, at least 16 major OCT1 alleles have been identified in more than 1,000 individuals from 53 diverse populations. Extreme global variability exists amongst the percentage of diverse populations with functional OCT1 alleles: individuals carrying homozygous or compound heterozygous loss-of-function alleles make up 80% of some South American populations and 9% of Caucasian populations, yet just less than 2% of East Asian and Oceanic populations (16). While some variants in pharmacogenes are common (i.e., polymorphisms) and thus able to be statistically associated with phenotype (e.g., drug exposure or response), interpretation of rare variants that are novel or sparsely observed is a major challenge. Comprehensive functional

characterization of variants in OCT1 has the potential to aid in pharmacogenetic variant interpretation and precision medicine/dosing.

To understand the function of genetic variants in all genes, including pharmacogenes, it is essential to conduct experimental studies (17). Functional genomic studies of SLC transporters are most commonly conducted *in vitro* in mammalian cell-based expression systems. Site-directed mutagenesis can be used to introduce single nucleotide variants or small insertions and deletions into a DNA construct. After transient or stable overexpression of the encoded transporter variant of interest, function can be measured by substrate uptake assays involving radioligands, fluorescent substrates, genetically encoded biosensors, or unlabeled substrates detected by mass spectrometry (18). In addition, transporters that are electrogenic can be functionally assessed through electrophysiology experiments (18). These methods, although the gold-standard in the field, are relatively low throughput, making the functional characterization of many genetic variants time consuming and costly.

Advancing technologies have greatly increased the scale at which functional studies are possible. Deep mutational scanning (DMS) is an emerging approach used to functionally characterize thousands of protein variants in parallel (19). Identification and development of a relevant assay for generating high-quality large-scale functional data can be challenging. Here, we present an experimental platform compatible with DMS of 11,572 variants in the pharmacogene OCT1, including all possible single missense variants, synonymous variants, and single amino acid deletions. We validate the integration and function of OCT1 and representative OCT1 variants in landing pad-based stable cell lines with fluorescent uptake assays with the substrate ASP^+ and radioligand uptake assays with substrates MPP^+ and metformin, and demonstrate detectable

differences in OCT1-mediated cytotoxicity upon exposure to platinum compounds SM73 and SM85 (Fig. 4.1).

4.3 Methods:

4.3.1 OCT1 wild-type construct assembly

The reference *SLC22A1* coding sequence (NM_003057.3) was cloned into the hAAVS1 landing pad vector by Gibson assembly (20). Briefly, template and backbone vector were PCR amplified with custom Gibson primers using PrimeSTAR® Max DNA Polymerase (Takara Bio, Kusatsu, Shiga, Japan). PCR products were run on a 1% agarose gel and extracted with the Zymo DNA Clean & Concentrator-5 kit (Zymo #11-302C, Zymo Research, Irvine, CA). Gibson assembly was performed with NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, Ipswich, MA) with a 2:1 mass ratio of backbone vector to insert for 60 min at 20°C. Then, assembled construct was transformed into NEB® 5-alpha competent cells (New England Biolabs, Ipswich, MA) and plated on LB-Ampicillin plates which were incubated at 37°C overnight. Colonies were picked and grown in 5 mL LB-Ampicillin broth for 16 hr at 37°C in a shaker/incubator. DNA was isolated with the Zyppy™ Plasmid Miniprep Kit (Zymo Research, Irvine, CA). Sequencing was performed to ensure correct assembly and absence of PCR errors (MCLAB, South San Francisco, CA).

4.3.2 Site-directed mutagenesis

Site-directed mutagenesis (SDM) was used to generate three OCT1 missense variants for experimental optimization and platform development. OCT1 variants p.R61C, p.P117L, and p.G401S were introduced into the OCT1 construct generated by Gibson assembly (section 4.3.1 above) with Q5® Site-Directed Mutagenesis Kit (New England Biolabs, Ipswich, MA)

following manufacturer's protocol. Sequencing was performed to ensure correct assembly and absence of PCR errors (MCLAB, South San Francisco, CA).

4.3.3 Cell culture

Cells were cultured in Dubecco's modified Eagle medium (DMEM) (Life Technologies, Carlsbad, CA) supplemented with 10% fetal bovine serum (GE Healthcare Life Sciences, South Logan, UT) and penicillin/streptomycin (100 U/mL) (Life Technologies, Carlsbad, CA) and grown in a humidified incubator at 37°C with 5.0% CO_2.

4.3.4 Stable cell line generation

Stable cell lines expressing OCT1 wild-type, p.R61C, p.P117L, and p.G401S were created by transfection into TetBxB1BFP-iCasp-Blast Clone 12 HEK293T cells harboring a landing pad for single-site specific integration (PMID: 31612958, referred to herein as HEK293T-landing pad cells). Cells were seeded at a density of $6.5x10^5$ cells/well in a 6-well plate. One day after seeding, cells were co-transfected with either OCT1 reference or variant destination vectors and the BxB1 expression vector (pCAG-NLS-BxB1) at a 1:1 mass ratio with Lipofectamine LTX transfection reagent (Thermo Fisher Scientific) according to manufacturer's protocol. Two days after transfection, cells were split into T25 flasks and doxycycline (2 μg/mL, Sigma-Aldrich) was added to induce OCT1 expression. One day after doxycycline was added, AP1903 (10 nM, MedChemExpress) was added for negative selection of cells with unsuccessful recombination. After one week selection with AP1903, expression of OCT1 was verified with *in vitro* uptake assays as described in sections 4.3.5 and 4.3.6 before further experiments were performed.

4.3.5 In vitro ASP^+ uptake assays

The activity of each OCT1 variant was measured by uptake of 4-Di-1-ASP (4-(4-(Dimethylamino)styryl)-N-Methylpyridinium Iodide (ASP^+) (21) in stable cell lines described above. Briefly, cells were seeded at 50,000 cells/well in poly-D-lysine coated black, clear bottom 96-well plates (Greiner Bio-One, Monroe, NC). One day later when cells were at approximately 95% confluency, the culture media was removed and replaced with 100 µL of 1 µM ASP^+ (Invitrogen™ #D288) in HBSS warmed to 37°C in the absence or presence of OCT1 inhibitors carvedilol or ketoconazole at 20, 50, or 100 µM (reaction mixture). After 10 min uptake, the reaction mixture was removed and cells were washed three times with ice-cold HBSS. The final wash was removed and ASP^+ fluorescence was measured with the GloMax® Explorer microplate reader (Promega, Madison, WI) with excitation and emission filters tuned to 475 nm and 580-640 nm wavelengths, respectively.

4.3.6 In vitro radioligand uptake assays

One day after seeding of OCT1 wild-type and variant stable cell lines in poly-D-lysine coated 96-well white bottom plates (Perkin Elmer #6005070) culture medium was removed and cells were washed three times with Hank's buffered salt solution (HBSS) (Life Technologies, Carlsbad, CA) at 37°C and pre-incubated with the third wash of HBSS for 10 min at 37°C. 0.5 µM metformin [biguanido-14C] hydrochloride (#ARC1738, American Radiolabeled Chemicals, Inc., St. Louis, MO) or 6.25 nM ^{3}H-methyl-4-phenylpyridinium acetate ("MPP^+", #ART0849, American Radiolabeled Chemicals, Inc., St. Louis, MO) in HBSS (reaction mix) was added to the cells and incubated at 37°C for 5 min, a time point within the linear uptake phase of OCT1 (22). After 5 min, the reaction mix was aspirated and the cells were washed three times with ice-cold HBSS. 200 µL MicroScint-20 (Perkin Elmer) was added to each well and cells were lysed

on an orbital shaker for 1 hr before plates were sealed with an adhesive plastic cover and read in a MicroBeta2® Microplate Counter (Perkin Elmer, Waltham, MA). Function of each variant was normalized to wild-type (WT) OCT1 and expressed as a percentage after background uptake measured in the empty vector (EV) was subtracted from both, calculated as follows: (Variant – EV)/(WT – EV)*100. Each cell line was assayed in triplicate on a 24-well plate and measured in three biological replicates.

4.3.7 Cytotoxicity assays with platinum compounds

The cytotoxicity of the platinum compounds was measured with the CellTiter-Glo® assay in 96-well plates. Briefly, stable cell lines were seeded in clear poly-D-lysine coated 96 well plates in culture medium with doxycycline (2 µg/mL) for 24, 48, or 72 hr treatment with platin compounds. Cells were seeded at a density of 25,000 cells/well for 24 hr treatment, 10,000 cells/well for 48 hr treatment, and 4,000 cells/well for 72 hr. 24 hr after seeding, culture medium was replaced with fresh medium containing doxycycline (2 µg/mL) and serial dilutions were performed to treat cells with SM73 or SM85 (US Patent No. 9217007B2 (23)) in concentrations ranging from 0-200 µM. After the treatment duration (24, 48, or 72 hr) had elapsed, media was replaced with 45 µL DMEM + 10% FBS per well. 45 µL CellTiter-Glo® reagent was added to each well and plates were placed on a shaker/incubator for 10 minutes. Then, cell viability was measured with a GloMax® Explorer microplate reader (Promega, Madison, WI). Cell viability at each concentration and time point was calculated as a percentage of maximum cell viability for each cell line. The maximum cell viability was the CellTiter-Glo® readout in the wells not treated with platin compounds for each cell line.

4.3.8 OCT1 DMS library generation

At the end of the *SLC22A1* sequence the stop codon was removed and replaced with a 3x Gly-Ser linker followed by mNeonGreen_11 and a downstream expression marker miRFP670 co-expressed via a self-cleaving P2A sequence. The DMS library was designed with SPINE (24) and library construction has been described previously (25). Briefly, oligos were designed and synthesized (Agilent Technologies, Inc., Santa Clara, CA) to contain each possible amino acid substitution at every position. Due to high error rates in oligo synthesis that limit maximal oligo length to 230 base pairs, the SLC22A1 coding region sequence (NM_003057.3) was divided into 11 blocks for oligo synthesis (Supplementary Table 4.1). Corresponding to each SLC22A1 block, 11 primer sets were designed by SPINE to amplify each oligo block and the respective section of the SLC22A1 backbone vector into which the oligo block is cloned for the generation of seamless circular constructs containing each variant in the complete gene. The 11 backbones were amplified by 25 PCR cycles with PrimeSTAR GXL Polymerase (Takara Bio, Kusatsu, Shiga, Japan) and 1 ng backbone DNA was used as template. The oligo library was amplified in 11 separate PCR reactions with respective primers for the amplification of each oligo block by 25 PCR cycles with PrimeSTAR GXL Polymerase and 1 µL of the oligo library (resuspended in 1mL TE buffer) used as template. After PCR, amplified backbones and inserts were run on a 1% agarose gel and extracted with the Zymo DNA Clean & Concentrator-5 kit (Zymo #11-302C, Zymo Research, Irvine, CA). To assemble the backbones with respective inserts, BsaI Golden Gate cloning reactions were set up in 20 µL reactions containing 100 ng of amplified backbone DNA, 20 ng of amplified oligo DNA, 0.2 BsaI-HFv2 (New England Biolabs, Ipswich, MA), 0.4 µL T4 DNA ligase (New England Biolabs, Ipswich, MA), 2 µL T4 DNA ligase buffer, and 2 µL 10mg/mL BSA. The Golden Gate cloning reactions were placed in a thermocycler overnight

with the following protocol: (5 min at 42°C, 10 min at 16°C)*40 cycles, then 20 min at 42°C, then 10 min at 80°C, then 4°C hold. The 11 reactions containing assembled constructs were cleaned with the Zymo DNA Clean & Concentrator-5 kit at eluted in 6 µL elution buffer. The constructs were then transformed into E. Cloni 10G electrocompetent cells (Lucigen Corp., Middleton, WI) by electroporation according to manufacturer's instructions. After transformation, cells were grown at 30°C for 6-10 hr in 30 mL LB broth with kanamycin (40 µg/mL) until optimal optical density (OD600) was achieved. A small amount of transformed cells was plated at multiple dilutions to evaluate transformation efficiency and validate successful assembly and presence of single missense mutations by sequencing (MCLAB, South San Francisco, CA) and from the remainder of transformed cells DNA was purified with Zyppy™ Plasmid Miniprep Kit. Each of the 11 sublibraries was combined at equimolar ratio to make the complete mutational library. Finally, the complete library was cloned into a destination vector plasmid with BxB1-compatible attB recombination sites for stable integration into the HEK293T-landing pad cells. Briefly, existing constructs were amplified by inverse PCR with primers that add complementary BsmBI cut sites, then Golden Gate cloning was performed using BsmBI and T4 ligase (New England Biolabs, Ipswich, MA) with the same protocol as described above for BsaI to generate the final deep mutational scanning library of attB-SLC22A1-mNeonGreen-P2A-puromycinR constructs.

4.3.9 Data analysis

Statistical analysis for uptake assays was performed in R version 3.6.3 (R Core Team, 2020) and plots were generated using R package ggplot2 version 3.3.5. Student's t-test was used to determine significance between groups, with $p<0.05$ used as the significance threshold. For IC_{50} curves from cytotoxicity assays, statistical analysis, curve fitting, and figures were generated

with GraphPad Prism version 8 software (La Jolla, CA). Additional figures were generated with Biorender and chemical structures were drawn with Marvin Sketch 19.9.

4.4 Results

4.4.1 OCT1 expression and function in landing pad cell lines

The expression and function of human OCT1 and selected variants p.R61C, p.P117L, and p.G401S after stable integration into the HEK293T landing pad cells were confirmed by measuring the uptake of OCT1 substrates ASP^+, MPP^+, and metformin as described below.

4.4.2 Effect of OCT1 variants on uptake of fluorescent substrate ASP^+

Here we validated the function of HEK239T-OCT1 stable cells by assessing uptake of the fluorescent substrate ASP^+ and sought to establish optimal inhibition conditions. We selected some of the most potent inhibitors of OCT1 identified previously (26) and tested them at multiple concentrations. Despite low IC_{50} values reported previously, (2.6 μM for ketoconazole and 1.6 μM for carvedilol), we found that 20 μM of either inhibitor reduced ASP^+ uptake by just 67-73% (p-value<0.001, Student's t-test, Fig. 4.2A). At 100 μM, ketoconazole resulted in near-complete inhibition of ASP^+ uptake, and carvedilol completely abolished ASP^+ uptake (p-value<0.001). We then tested the effect of OCT1 variants on the uptake of ASP^+, proceeding with carvedilol as an inhibitor and evaluating inhibitability at 20, 50, and 100 μM. Stable cell lines expressing OCT1 p.R61C and p.P117L did not exhibit any significant change in ASP^+ uptake from OCT1 WT, yet uptake was significantly reduced in cells expressing the reduced function OCT1 p.G401S variant (p-value<0.001), which had a lower baseline uptake of ASP^+ than the other OCT1 variants (Fig. 4.2B). OCT1 p.R61C was more inhibitable than the WT or p.P117L variant, with greatest reduction in uptake of ASP^+ at all three concentrations of carvedilol tested.

4.4.3 Effect of OCT1 variants on uptake of radiolabeled substrates MPP^+ and metformin

Uptake of the prototypical cation MPP^+ and metformin was determined by radioligand uptake assays. The uptake of MPP^+ and metformin was significantly higher in landing pad cells stably transfected with OCT1 constructs compared to untransfected landing pad cells ($p < 0.001$, Fig. 4.3A-B). The uptake of MPP^+ and metformin was significantly reduced by the OCT1 inhibitor carvedilol at 100 μM ($p < 0.001$). ^{3}H-MPP^+ uptake in stable cell lines expressing OCT1 variants was assessed and was found to be significantly increased in p.R61C, unchanged in p.P117L, and significantly decreased in p.G401S compared to OCT1 WT. ^{14}C-metformin uptake was significantly decreased in p.R61C and p.G401S and unchanged in p.P117L. Carvedilol (100 μM) completely inhibited uptake of ^{14}C-metformin and almost completely inhibited uptake of ^{3}H-MPP^+ by OCT1 WT and all variants tested.

4.4.4 Time and concentration dependence of OCT1-mediated cytotoxicity by platinum compounds oxaliplatin, SM73, and SM85

OCT1 has previously been shown to mediate the influx of the platinum-based anticancer drug oxaliplatin in human colon cancer cell lines (27). In HEK293T cells stably expressing hOCT1, we measured cytotoxicity of oxaliplatin up to concentrations of 200 μM. Even with 72 hr exposure to 200 μM oxaliplatin, only partial cytotoxicity was observed in OCT1 expressing cells (Supp. Fig. 4.1). Therefore, we sought to evaluate the cytotoxicity of two synthesized platinum compounds, SM73 and SM85. HEK293T-landing pad and HEK293T-OCT1 WT cells were treated with SM73 and SM85 at 12 concentrations ranging from 0-50 μM (2-fold dilutions) for 24, 48, and 72 hr before cell viability was measured. OCT1 expressing cells exhibited robust platinum-mediated cytotoxicity compared to landing pad cells (Fig 4.5). Cell viability was completely (48 hr and 72 hr time points) or near completely (24 hr time point) abolished in both

cell lines after treatment with 50 μM SM73 or SM85, and cell viability was unaffected by concentrations as low as 50 nM at all time points. IC_{50} values for SM73 and SM85 were calculated for both cell lines at all three time points (Table 4.1). Resistance factors, defined as the IC_{50} in the HEK293T-landing pad cells divided by the IC_{50} in the HEK293T-OCT1 cells, ranged from 40.3 to 167.4 μM for SM73 and 86.2 to 2934.1 μM for SM85 (Table 4.1), indicating that cells overexpressing OCT1 are much more sensitive to platinum-induced cytotoxicity than cells without OCT1 overexpression. Resistance factors were highest at the 48 hr time point for both SM73 and SM85. IC_{50} values decreased as treatment time increased for both platinum compounds, indicating sensitivity increases with increased drug exposure.

4.4.5 Effect of OCT1 variants on the cytotoxicity of platinum compound SM85

SM85 was found to be more potent and have greater resistance factors than SM73 in platinum-induced cytotoxicity assays (see section 4.3.4 above), thus we moved forward with SM85 in assessing viability of cells expressing OCT1 variants. We assessed cell viability after exposure to 12 concentrations of SM85 (0-50 μM, 2-fold dilutions, Fig. 4.5) for 48 hr, the time point that resulted in the greatest resistance factor in previous experiments. We found that cells expressing OCT1 p.P117L were most sensitive to SM85-induced cytotoxicity (RF = 11.0, Table 4.2), consistent with a mild gain-of-function phenotype observed for this variant with other substrates. Cells expressing OCT1 p.R61C were slightly less sensitive (RF = 7.2) to SM85 than those expressing OCT1 WT (RF = 11.0), and p.G401S exhibited strong resistance to SM85-induced cytotoxicity (RF = 0.4, Table 4.2).

4.4.6 Substrate specificity of common OCT1 variants

Here we summarize the substrate specificity of OCT1 variants. We directly measured uptake of MPP^+, metformin, and ASP^+. We found that OCT1 p.P117L in general functions very similar to

OCT1, with minimal [significant or insignificant] reductions in transport of MPP^+ and metformin. OCT1 p.G401S is largely a reduced function variant, with severe reduction in uptake for both metformin and ASP^+ yet moderate function for MPP^+ uptake. In contrast, OCT1 p.R61C is a variant with variable degree of function based on substrate. p.R61C is gain-of-function for MPP^+ uptake, loss-of-function for metformin uptake, and normal function for ASP^+ uptake. These results are summarized in Figure 4.6.

4.4.7 Generation of OCT1 deep mutational scanning variant library

The OCT1 variant library for deep mutational scanning was constructed as described in Methods section 4.2.8. The library contains up to 11,572 total sequences. The human OCT1 protein contains 554 amino acids. The library contains sequences encoding all 19 possible single missense variants as well as synonymous variants and single amino acid deletions at every residue position in OCT1 with the exception of methionine at position 1. Thus, the library contains 19 missense variants * 553 residues =10,507 variants, 1 synonymous variant * 522 residues = 522 synonymous variants (no synonymous variants exist for Met or Trp residues, of which OCT1 contains 19 and 13, respectively), and 1 deletion * 543 amino acids = 11,572 total OCT1 variants present in the DMS library. The constructs encode OCT1 with the split fluorescent protein fragment $mNeonGreen2_{11}$ attached to the C-terminus by a 3x Gly-Ser linker, followed by miRFP670 co-expressed via a P2A self-cleaving peptide. $mNeonGreen2_{11}$ is a 16-amino acid peptide fragment of the mNeonGreen2 fluorescent protein that is non-fluorescent by itself but can be co-expressed with the $mNeonGreen2_{1-10}$ fragment to self-complement into a fluorescent protein. $mNeonGreen2_{1-10/11}$ has multiple benefits: 1) the small size of the $mNeonGreen2_{11}$ fragment compared to a complete fluorescent protein is less likely to interfere with the folding, stability, or function of OCT1, and 2) it exhibits increased brightness upon self-

complementation and reduced background fluorescence of non-complemented fragments compared to the split fluorescent protein $GFP_{1-10/11}$. Inclusion of miRFP670 is used to control for OCT1 expression.

4.5 Discussion

Here we present a screening platform for the high-throughput assessment of functional impact of 11,572 variants in the pharmacogene SLC22A1, encoding the organic cation transporter 1, OCT1. The platform utilized a landing-pad based stable cell system, which was experimentally validated to function and exhibit sensitivity to detect changes in OCT1 function caused by missense variants or single amino acid deletions. We demonstrate expression of OCT1 by measuring significant increase in uptake of the radiolabeled OCT1 substrates MPP^+ and metformin that is reduced in the presence of the OCT1 inhibitor carvedilol. We establish two diverse assays that can be used to determine the effect of missense variants on the function of OCT1: (1) a fluorescence-based assay that measures uptake of the OCT1 substrate, ASP^+ into cells and (2) a cytotoxicity-based assay that selects for loss-of-function or reduced function variants, identifiable by next-generation sequencing. Our studies confirm previous findings that missense variants of OCT1 exhibit different effects on transport function depending on the substrate. Finally, we describe the generation of an OCT1 deep mutational scanning library containing 11,572 OCT1 variants that can both be functionally characterized with the aforementioned assays as well as spatially characterized to determine subcellular localization utilizing a conjugated fragment of the split fluorescent protein $mNeonGreen2_{1-10/11}$.

We compared our results to existing reports in the literature and identified a number of interesting differences. With ASP^+ uptake assays in HEK293T-landing pad cells expressing

OCT1, we assessed inhibition by two drugs reported to be potent inhibitors of OCT1: ketoconazole and carvedilol. Reported IC_{50} values were 2.6 μM for ketoconazole and 1.6 μM for carvedilol. Here, we found that 20 μM of either inhibitor only partially reduced uptake of ASP^+. IC_{50} values can be influenced by many factors, including concentrations of substrates and inhibitors used and incubation time. Compared to the previous study, we used a lower concentration ASP^+ (1 μM vs 2 μM) and had a longer assay duration (10 min vs 2 min), factors that might be contributing to differences observed here. Another interesting finding was that ASP^+ uptake was not statistically different in cells expressing OCT1 p.R61C or p.P117L compared to wild-type OCT1, yet uptake was significantly reduced by p.G401S. Interestingly, despite similar ASP^+ uptake to wild-type OCT1 and p.P117L, p.R61C exhibited a greater degree of inhibition with all three concentrations of carvedilol tested. Though speculative, it is possible that p.R61C does not affect affinity for or capacity to transport ASP^+, but has increased affinity for carvedilol.

The pooled OCT1 DMS library will be transfected and stably integrated into the HEK293T-landing pad cells as described in the methods section, with library diversity verified by next-generation sequencing (NGS). In the fluorescence-based ASP^+ uptake assay, OCT1 variant function can be determined by fluorescence-activated cell sorting (FACS). After incubation with ASP^+ in suspension, cells can be sorted into bins of fluorescent intensity indicative of variant function. We hypothesize that loss-of-function variants will not uptake ASP^+ and thus have low fluorescence intensity, whereas functional variants will have high fluorescence intensity, with moderately functioning variants in between. After sorting into bins, cells can be lysed and genomic DNA harvested for sequencing by NGS to determine the identity of variants in each bin. Because OCT1 genetic variants exhibit distinct effects on different substrates, follow-up

studies will be needed to understand the substrate specificity of individual OCT1 variants. In the cytotoxicity assay, cells stably expressing OCT1 variants will be exposed to platinum compound SM85 to determine sensitivity. We hypothesize that functional OCT1 variants will uptake SM85 resulting in toxicity and cell death, whereas LOF variants will be resistant to SM85-mediated cytotoxicity similar to the HEK293T-landing pad cells. There may be a reduction in abundance in cells expressing variants with moderate function compared to resistant cells as exposure time to SM85 increases. After treatment with SM85, cells will be washed to remove dead cells and then lysed and genomic DNA harvested and sequenced by NGS to determine abundance of OCT1 variants. Variants enriched in sequencing reads primarily from live cells are resistant to SM85-induced cytotoxicity and hypothesized to be LOF.

The presence of split fluorescent protein fragment mNeonGreen2$_{11}$ at the C-terminus of OCT1 allows for detection upon co-expression and complementation with mNeonGreen2$_{1-10}$. This can be used to quantify protein expression by FACS and identify variants that may have reduced expression due to thermodynamic instability and/or increased protein degradation (28). In addition, membrane localization of OCT1 variants can be quantified by FACS upon binding to anti-OCT1 antibody and a fluorescent secondary antibody. Previous studies revealed that 38% (6/16) of major OCT1 alleles display improper subcellular localization leading to loss-of-function for all substrates (16). Upon comparison with functional data from uptake and cytotoxicity assays, we hypothesize that OCT1 variants absent from the plasma membrane or with extreme reduction in overall expression will be LOF, thus enriched in sequencing reads after SM85 treatment and falling into low fluorescence bins after ASP$^+$ uptake.

Deep mutational scanning of OCT1 has the potential to reveal important binding and translocation sites within OCT1, as well as continue to inform our understanding of the

mechanisms by which OCT1 transports its substrates. OCT1 is a polyspecific transporter, with more than 150 substrates identified (29). Multiple binding sites have been proposed within OCT1, some of which are perhaps overlapping, and multiple proposed pharmacophore models have substantial differences. No crystal structures exist for OCT1 or any related proteins in the SLC22 family, and homology models built to fill the gap are based on solved structures sharing less than 20% amino acid identity with OCT1 (30). DMS has previously been used to determine protein structure (31) and is sure to enrich our understanding of structure-function relationships within OCT1.

One limitation of the proposed deep mutational scan of OCT1 is the substrate-specific functional heterogeneity caused by genetic variation. For example, the most common variant p.M420del greatly reduces transport of some substrates (e.g., metformin) but has no or minimal effect on the uptake of other substrates (e.g., morphine) (9, 29). This greatly limits the generalizability of an activity score determined for one substrate. However, strong pairwise correlations exist for functional effects between some substrates. The effect of OCT1 polymorphisms on the transport of MPP^+ was highly correlated with the effect on transport of O-desmethyltramadol, ASP^+ was highly correlated with morphine, and metformin was highly correlated with sumatriptan (correlation coefficients of 0.869, 0.838, and 0.776, respectively) (29). This limitation motivates our drive to carry out deep mutational scanning with multiple diverse assays using more than one OCT1 substrate, as well as to perform detailed follow-up studies to understand the specificity of individual variants of OCT1. Specificity differences among OCT1 missense variants suggest different binding sites and translocation pathways that may be utilized by different substrates. Such differences are also consistent with a protein that tolerates a diverse array of substrates.

The rich layers of data revealed by DMS can be used to train machine learning models to predict impact of a variant on expression, function, and localization of OCT1 and related transporters. A model built with mutagenesis data from OCT1 can be fine-tuned to predict these parameters for related transporters with far less data available in a process known as transfer learning (32). OCT1 is a member of the SLC22 family which contains multiple transporters with clinical pharmaceutical interest (OCT2/*SLC22A2*, OAT1/*SLC22A6*, OAT3/*SLC22A8*), and genetic disease implications (OCTN2/*SLC22A5*, URAT/*SLC22A12*). Protein-specific models for interpreting functional impact of variants in these transporters could have great clinical relevance in precision dosing of certain medications that are substrates of these transporters and in diagnostics of associated diseases, though ethical and technical challenges need attention before this vision becomes a reality (32).

In summary, here we describe a deep mutational scanning platform for the multi-parametric characterization of all possible single amino acid variants in OCT1. We establish and validate several functional assays as well as comment on a split fluorescence protein fragment that can be used to quantify protein expression levels of OCT1 variants. DMS of OCT1 can ultimately be used to broadly enhance our biological understanding of the mechanisms of OCT1 transport, polyspecificity, expression, and localization. Variants significantly affecting the function of OCT1 have the potential to be clinically actionable in the implementation of pharmacogenetic data in dosing of drugs that are OCT1 substrates. Finally, as the first SLC transporter to be fully characterized by DMS to our knowledge, lessons learned from OCT1 can be insightful for related transporters with known clinical importance.

4.6 Figures

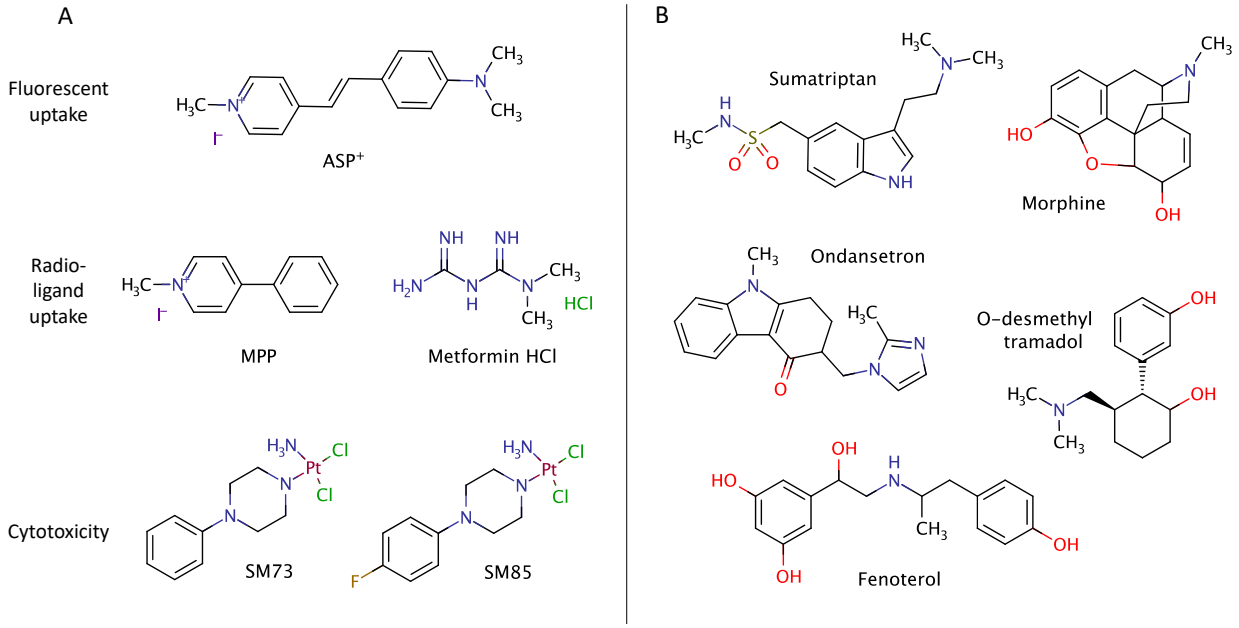


Figure 4.1. OCT1 substrates. (A) Substrates used in the validation of the DMS screening platform and **(B)** selected substrates with documented clinical evidence of differential exposure, efficacy, and/or side effects in individuals with reduced function OCT1 variants.

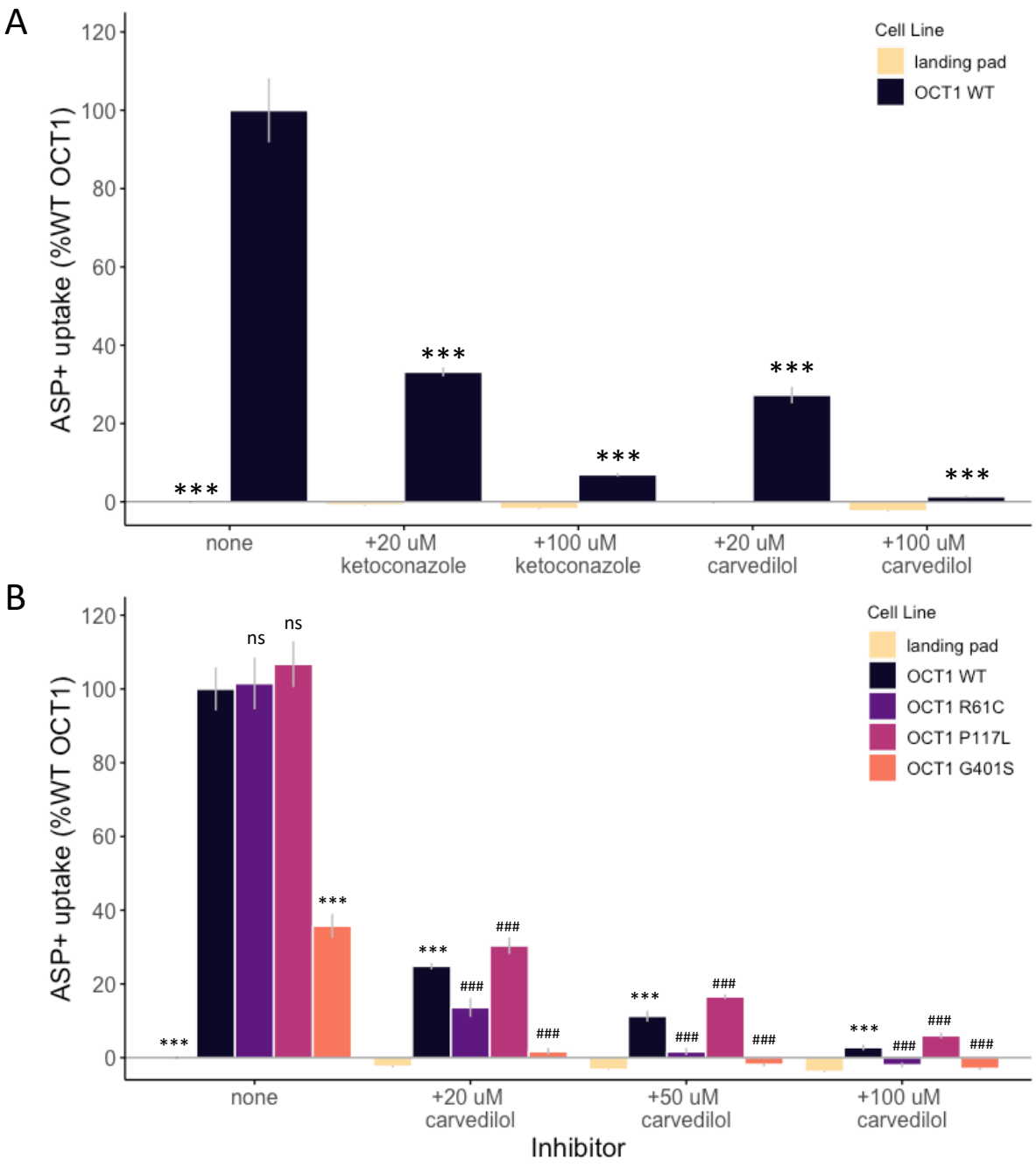


Figure 4.2. Effect of OCT1 and OCT1 variants on uptake and inhibition of ASP$^+$. (A) Uptake of ASP$^+$ in HEK293T-landing pad and HEK293T-OCT1 expressing cells in absence or presence of OCT1 inhibitors ketoconazole and carvedilol at 20 µM or 100 µM. **(B)** Effect of OCT1 variants R61C, P117L, and G401S on ASP$^+$ uptake and inhibition by carvedilol (20, 50, or 100 µM). Data are expressed as mean ± SEM from two biological replicates (n = 4 per condition for each replicate). *** indicates significant difference ($p<0.001$) from uptake in OCT1 WT cells with no inhibitor, ### indicates significant difference ($p<0.001$) from uptake in respective variant cell line with no inhibitor, ns = not significant from OCT1 WT.

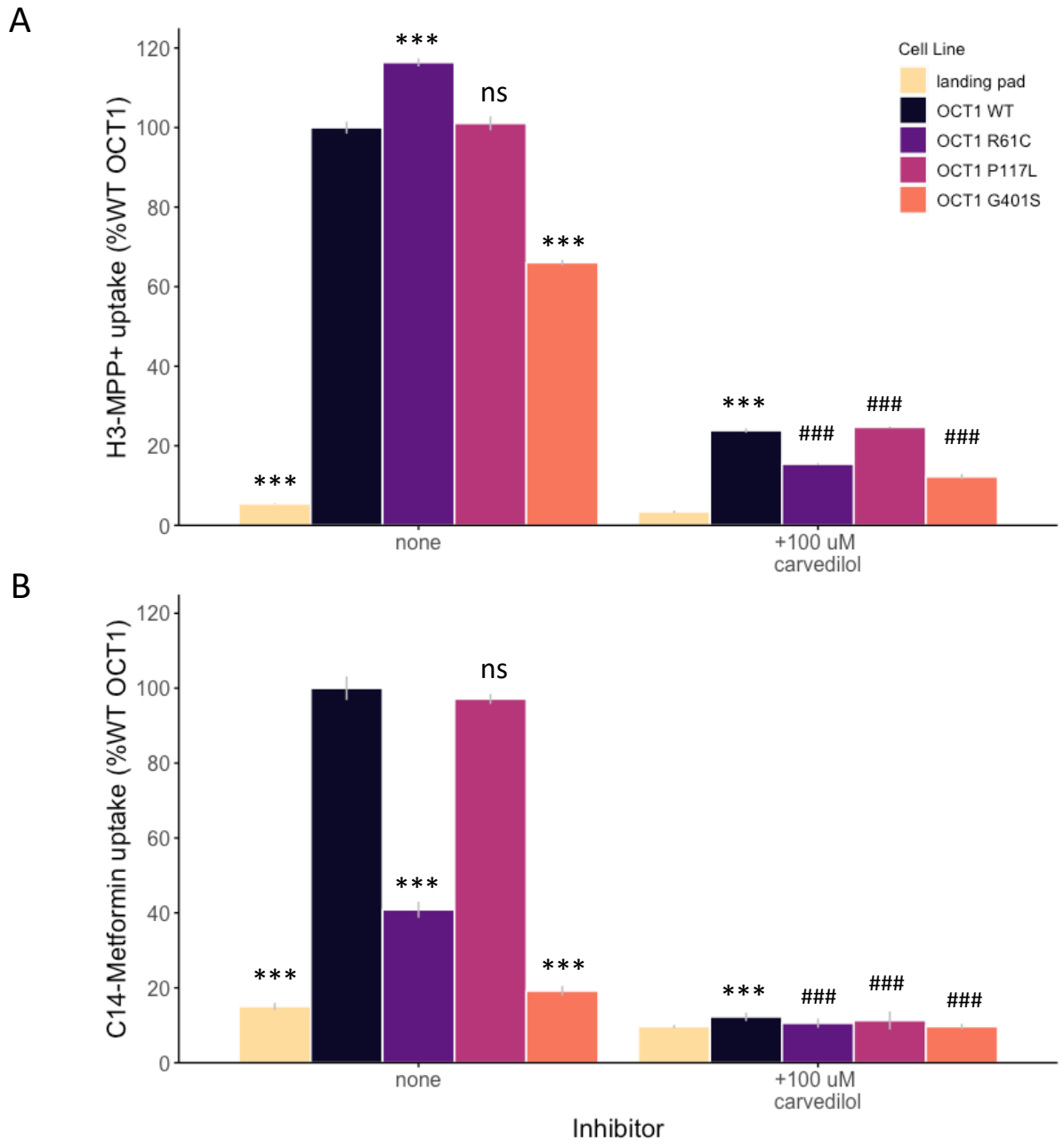


Figure 4.3. Effect of OCT1 variants on uptake and inhibition of radiolabeled substrates. Function of stable cell lines HEK293T-landing pad, and HEK293T-landing pad cells expressing OCT1 WT, OCT1 R61S, OCT1 P117L, and OCT1 G401S with respect to uptake of **(A)** ^{3}H-MPP^{+} and **(B)** ^{14}C-metformin**.** Uptake of substrates was determined in the absence or presence of 100 µM carvedilol to inhibit OCT1. Data are expressed as mean ± SEM from three biological replicates (n = 4 per condition for each replicate). *** indicates significant difference ($p<0.001$) from uptake in OCT1 WT cells with no inhibitor, ### indicates significant difference ($p<0.001$) from uptake in respective variant cell line with no inhibitor, ns = not significant from OCT1 WT.

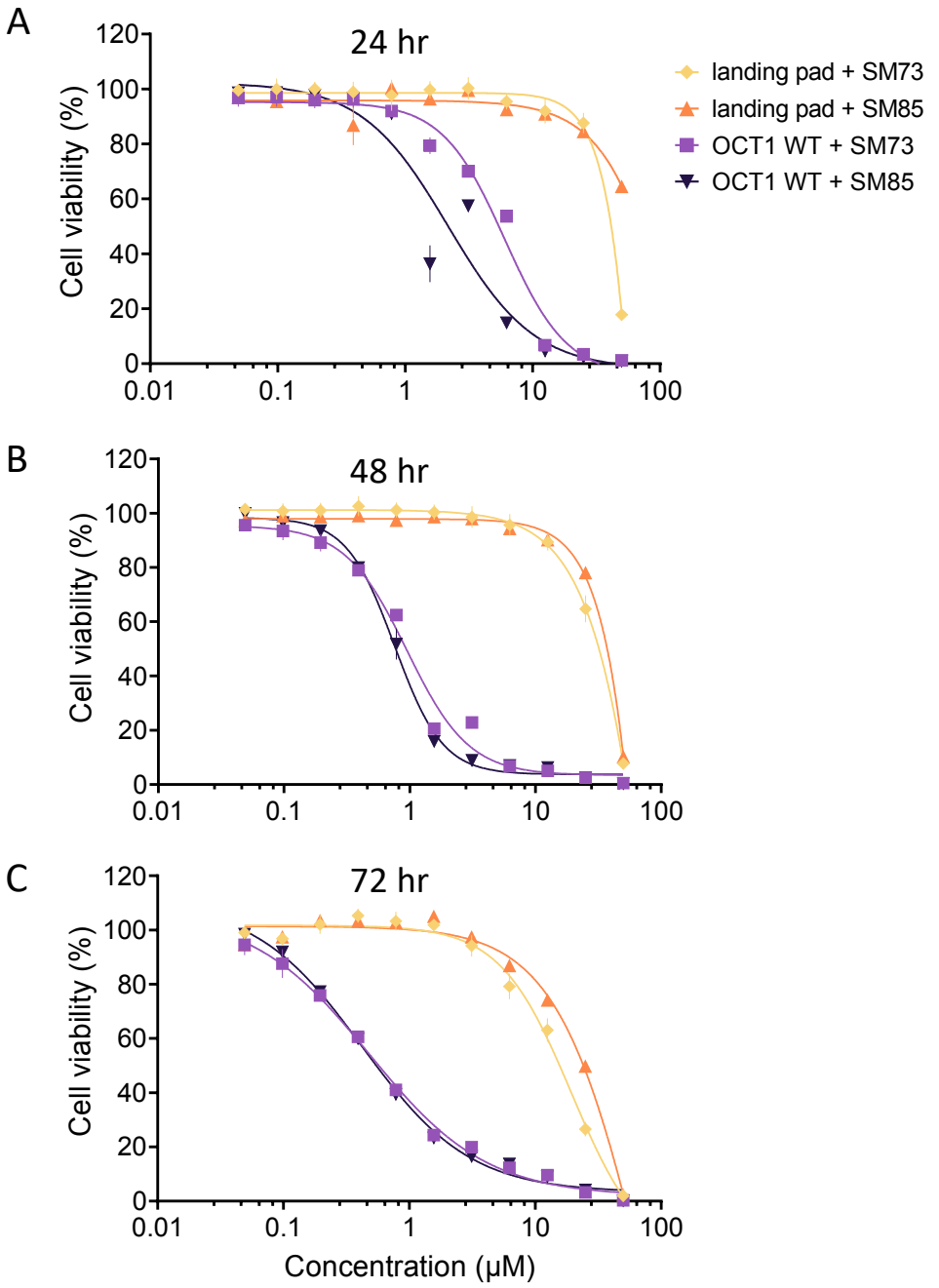


Figure 4.4. Time- and concentration-dependent cytotoxicity of platinum compounds SM73 and SM85 in cells expressing OCT1 WT and HEK293T-landing pad control cells. HEK293T-landing pad control cells and HEK293T-landing pad cells expressing OCT1 were exposed to SM73 or SM85 at concentrations ranging from 0-50 μM for **(A)** 24 hr, **(B)** 48 hr, or **(C)** 72 hr. Data are expressed as mean ± SEM and are representative of two biological replicates (n = 3 per condition for each replicate).

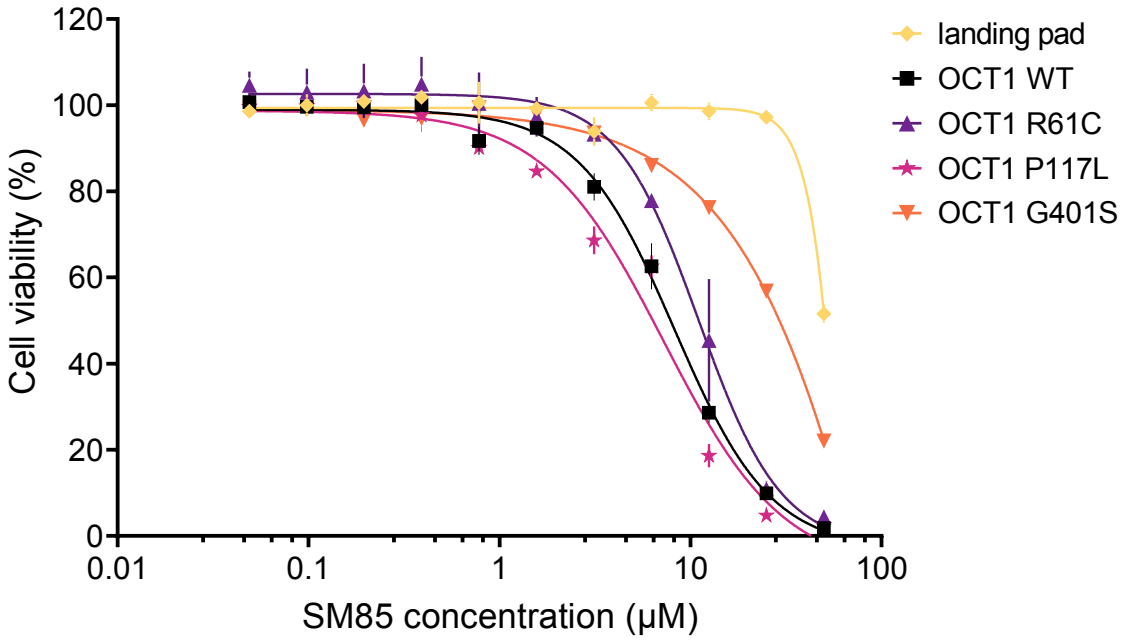


Figure 4.5. Effect of OCT1 variants on cytotoxicity of platinum compound SM85 in stable cell lines. HEK293T-landing pad cells and HEK293T-landing pad cells expressing OCT1 WT and missense variants of OCT1 were exposed to concentrations of SM85 ranging from 0-50 μM for 48 hr. Data are expressed as mean ± SEM for each concentration (n = 3 per condition for each replicate).

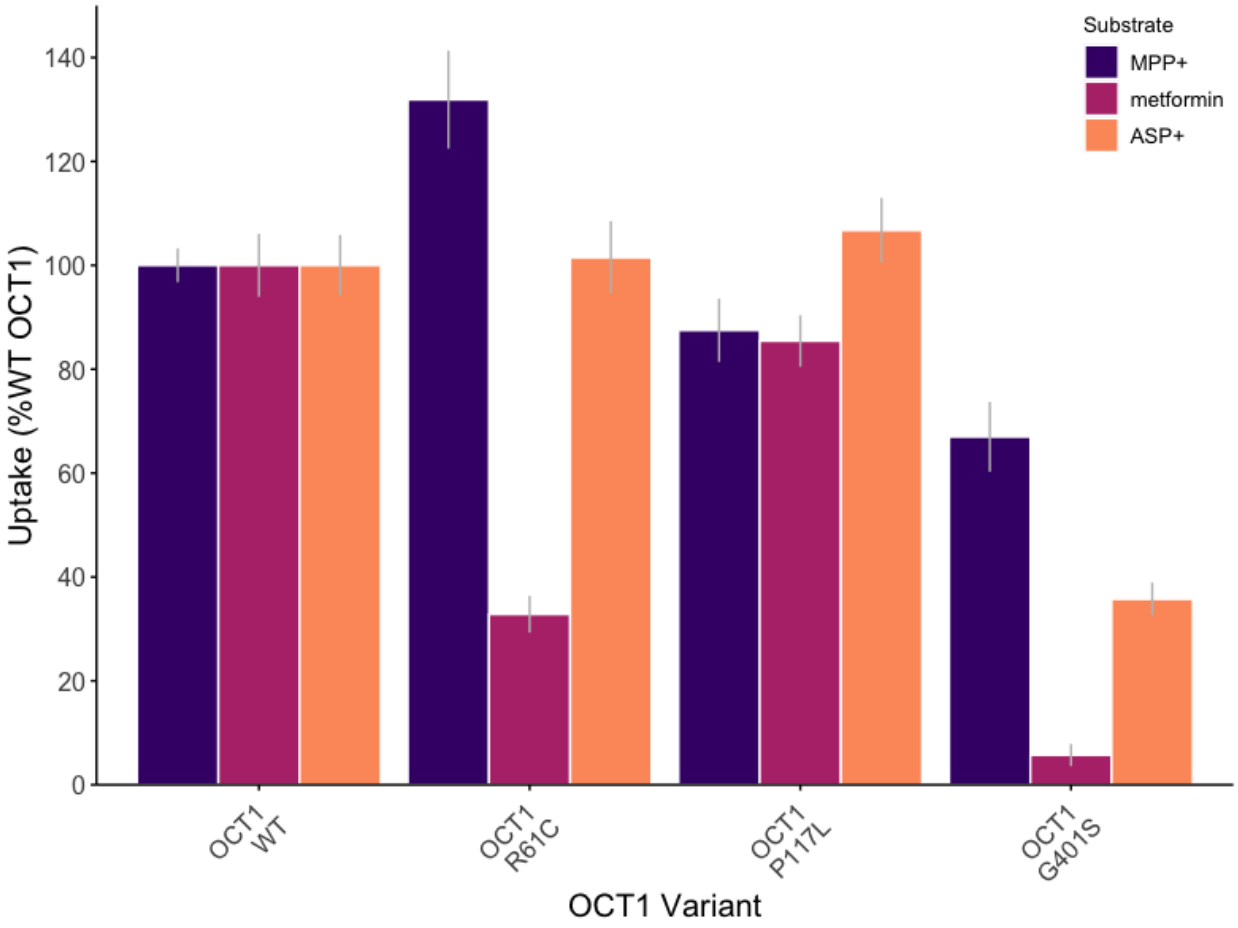
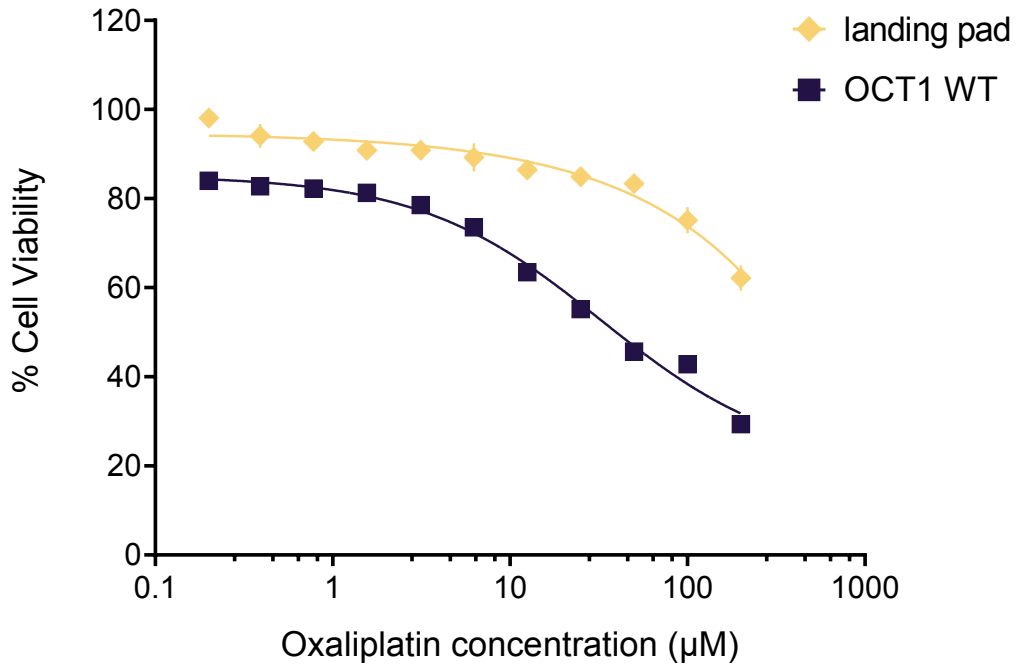


Figure 4.6. Summary of the effect of OCT1 variants on uptake function of MPP^+, metformin, and ASP^+. Data are expressed as mean ± SEM. For MPP^+ and metformin, data from three replicate experiments are shown (n = 4 for each variant per replicate). For ASP^+, data from two replicate experiments are shown (n = 4 for each variant per replicate).



Supplementary Figure 4.1. Cytotoxicity of oxaliplatin in cells expressing OCT1 WT. HEK293T-landing pad cells and OCT1 cells were exposed to concentrations ranging from 0-200 μM for 72 hr. Data are expressed as mean ± SEM and are representative of three biological replicates (n = 4 per condition for each replicate).

4.7 Table

Table 4.1. Drug sensitivity of oxaliplatin and platinum compounds SM73 and SM85 in HEK293T-landing pad and HEK293T-OCT1 cell lines. IC_{50}s were determined by cell viability as measured by the CellTiterGlo assay as described in Methods. Data are expressed as mean of one to two experiments with each done in triplicate. The resistance factor was defined as the ratio of the mean IC_{50} values in HEK293T-landing pad over HEK293T-OCT1 cells. ND: not determined.

		IC_{50} (µM)		
Compound	**Exposure (hr)**	**HEK293T-landing pad**	**HEK293T-OCT1**	**Resistance Factor (RF)**
Oxaliplatin	72	ND	32.55	ND
SM73	24	584	6.03	96.8
SM73	48	159	0.95	167.4
SM73	72	19.4	0.481	40.3
SM85	24	187	2.17	86.2
SM85	48	2227	0.759	2934.1
SM85	72	79.1	0.423	187.0

Table 4.2. Drug sensitivity of SM85 platinum analogue in OCT1-transfected cells. Data are expressed as mean of triplicate. The resistance factor was defined as the ratio of the mean IC_{50} values in HEK293T-landing pad over HEK293T-OCT1 WT or variant expressing cells.

Cell Line	IC_{50} (µM)	Resistance Factor (RF)
HEK293T landing pad	79.9	1.0
HEK293T-OCT1 WT	8.3	9.7
HEK293T-OCT1 R61C	11.0	7.2
HEK293T-OCT1 P117L	7.2	11.0
HEK293T-OCT1 G401S	187.8	0.4

Supplementary Table 4.1. OCT1 oligo block design. OCT1 oligo blocks synthesized for assembly by golden gate cloning in the construction of the DMS variant library.

OCT1 oligo block	Residues encoded
1	1-50
2	51-102
3	103-153
4	154-204
5	205-254
6	255-304
7	305-354
8	355-404
9	405-454
10	455-504
11	505-554

Supplementary Table 4.2. Simplified overview of the protocol for assembly of the OCT1 deep mutational scanning library.

Step	Description
Step 1	Variant library with SPINE (24)
Step 2	OCT1 oligo block synthesis (11 blocks)
Step 3	PCR to amplify each block and complementary backbone
Step 4	Gel purification of PCR products
Step 5	Golden gate cloning with BsaI to assemble constructs
Step 6	DNA purification of assembled constructs
Step 7	Transformation into E. coli by electroporation
Step 8	Growth of bacteria in LB broth
Step 9	Miniprep of DNA constructs
Step 10	Pooling of 11 sublibraries into single complete variant library
Step 11	PCR to add BsmBI cut sites
Step 12	Golden gate cloning with BsmBI to assemble OCT1 variants into destination vector
Step 13	DNA purification of assembled constructs
Step 14	Transfection into HEK293T landing pad cells
Step 15	Selection of stably integrated cells

4.8 References

1. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature. 2015;526(7573):343-50.
2. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. 2019;15(12):e1008489.
3. US Deparment of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH). Clinical Pharmacogenomics: Premarket Evaluation in Early-Phase Clinical Studies and Recommendations for Labeling. Guidance for Industry. . Silver Springs, MD, USA; 2013.
4. Caudle KE, Klein TE, Hoffman JM, Muller DJ, Whirl-Carrillo M, Gong L, et al. Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. Curr Drug Metab. 2014;15(2):209-17.
5. Lee HH, Ho RH. Interindividual and interethnic variability in drug disposition: polymorphisms in organic anion transporting polypeptide 1B1 (OATP1B1; SLCO1B1). Br J Clin Pharmacol. 2017;83(6):1176-84.
6. US Deparment of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER). In Vitro Metabolism- and Transporter-Mediated Drug-Drug Interaction Studies. Guidance for Industry. . Silver Springs, MD, USA; 2017.

7. Zamek-Gliszczynski MJ, Giacomini KM, Zhang L. Emerging Clinical Importance of Hepatic Organic Cation Transporter 1 (OCT1) in Drug Pharmacokinetics, Dynamics, Pharmacogenetic Variability, and Drug Interactions. Clin Pharmacol Ther. 2018;103(5):758-60.

8. Yee SW, Brackman DJ, Ennis EA, Sugiyama Y, Kamdem LK, Blanchard R, et al. Influence of Transporter Polymorphisms on Drug Disposition and Response: A Perspective From the International Transporter Consortium. Clin Pharmacol Ther. 2018;104(5):803-17.

9. Matthaei J, Kuron D, Faltraco F, Knoch T, Dos Santos Pereira JN, Abu Abed M, et al. OCT1 mediates hepatic uptake of sumatriptan and loss-of-function OCT1 polymorphisms affect sumatriptan pharmacokinetics. Clin Pharmacol Ther. 2016;99(6):633-41.

10. Tzvetkov MV, Saadatmand AR, Bokelmann K, Meineke I, Kaiser R, Brockmöller J. Effects of OCT1 polymorphisms on the cellular uptake, plasma concentrations and efficacy of the 5-HT(3) antagonists tropisetron and ondansetron. Pharmacogenomics J. 2012;12(1):22-9.

11. Fukuda T, Chidambaran V, Mizuno T, Venkatasubramanian R, Ngamprasertwong P, Olbrecht V, et al. OCT1 genetic variants influence the pharmacokinetics of morphine in children. Pharmacogenomics. 2013;14(10):1141-51.

12. Balyan R, Zhang X, Chidambaran V, Martin LJ, Mizuno T, Fukuda T, et al. OCT1 genetic variants are associated with postoperative morphine-related adverse effects in children. Pharmacogenomics. 2017;18(7):621-9.

13. Stamer UM, Musshoff F, Stuber F, Brockmoller J, Steffens M, Tzvetkov MV. Loss-of-function polymorphisms in the organic cation transporter OCT1 are associated with reduced postoperative tramadol consumption. Pain. 2016;157(11):2467-75.

14. Tzvetkov MV, Saadatmand AR, Lötsch J, Tegeder I, Stingl JC, Brockmöller J. Genetically polymorphic OCT1: another piece in the puzzle of the variable pharmacokinetics and pharmacodynamics of the opioidergic drug tramadol. Clin Pharmacol Ther. 2011;90(1):143-50.

15. Tzvetkov MV, Matthaei J, Pojar S, Faltraco F, Vogler S, Prukop T, et al. Increased Systemic Exposure and Stronger Cardiovascular and Metabolic Adverse Reactions to Fenoterol in Individuals with Heritable OCT1 Deficiency. Clin Pharmacol Ther. 2018;103(5):868-78.

16. Seitz T, Stalmann R, Dalila N, Chen J, Pojar S, Dos Santos Pereira JN, et al. Global genetic analyses reveal strong inter-ethnic variability in the loss of activity of the organic cation transporter OCT1. Genome Med. 2015;7(1):56.

17. Rodenburg RJ. The functional genomics laboratory: functional validation of genetic variants. J Inherit Metab Dis. 2018;41(3):297-307.

18. Dvorak V, Wiedmer T, Ingles-Prieto A, Altermatt P, Batoulis H, Barenz F, et al. An Overview of Cell-Based Assay Platforms for the Solute Carrier Family of Transporters. Front Pharmacol. 2021;12:722889.

19. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11(8):801-7.

20. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6(5):343-5.

21. Ahlin G, Karlsson J, Pedersen JM, Gustavsson L, Larsson R, Matsson P, et al. Structural requirements for drug inhibition of the liver specific human organic cation transport protein 1. J Med Chem. 2008;51(19):5932-42.

22. Nies AT, Hofmann U, Resch C, Schaeffeler E, Rius M, Schwab M. Proton pump inhibitors inhibit metformin uptake by organic cation transporters (OCTs). PLoS One. 2011;6(7):e22163.

23. Giacomini KM, More, S., inventor; The Regents of the University of California, assignee. Platinum Anticancer Agents. USA2015.

24. Coyote-Maestas W, Nedrud D, Okorafor S, He Y, Schmidt D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. Nucleic Acids Res. 2020;48(2):e11.

25. Coyote-Maestas W, Nedrud D, He Y, Schmidt D. Determinants of trafficking, conduction, and disease within a K+ channel revealed through multiparametric deep mutational scanning. bioRxiv. 2022.

26. Chen EC, Khuri N, Liang X, Stecula A, Chien HC, Yee SW, et al. Discovery of Competitive and Noncompetitive Ligands of the Organic Cation Transporter 1 (OCT1; SLC22A1). J Med Chem. 2017;60(7):2685-96.

27. Zhang S, Lovejoy KS, Shima JE, Lagpacan LL, Shu Y, Lapuk A, et al. Organic cation transporters are determinants of oxaliplatin cytotoxicity. Cancer Res. 2006;66(17):8847-57.

28. Matreyek KA, Starita LM, Stephany JJ, Martin B, Chiasson MA, Gray VE, et al. Multiplex assessment of protein variant abundance by massively parallel sequencing. Nat Genet. 2018;50(6):874-82.

29. Meyer MJ, Tzvetkov MV. OCT1 Polyspecificity-Friend or Foe? Front Pharmacol. 2021;12:698153.

30. Koepsell H. Organic Cation Transporters in Health and Disease. Pharmacol Rev. 2020;72(1):253-319.

31. Schmiedel JM, Lehner B. Determining protein structures using deep mutagenesis. Nat Genet. 2019;51(7):1177-86.

32. McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. Am J Hum Genet. 2021;108(4):535-48.

Chapter 5: Conclusions and Perspectives

In this dissertation, we sought to address the challenge of translating genetic data into clinically actionable information by increasing understanding of phenotypic consequences and improving experimental and computational approaches to interpreting coding region variants in clinically important genes. We performed functional genomic studies of Solute Carrier (SLC) transporters, namely OCTN2 and OCT1, to identify genetic variants with significant effects on function, relevant in the pathology of rare disease and interindividual differences in drug response, respectively. We outlined how these data can be used to inform machine learning models to improve upon the computational interpretation of variants of uncertain significance (VUS), both within these transporters and the other members of the SLC22 family, particularly to inform diagnostics and selection/dosing of therapeutics.

Chapter 1 provided foundational information on the biology and clinical relevance of organic cation (OCT1) and zwitterion (OCTN2) transporters in the SLC22 family from pharmacologic, physiologic, and pathophysiologic perspectives. Though many substrates and inhibitors are known to interact with these transporters, specific information on structure-function relationships, transport mechanisms, and functional genomics remains to be discovered. In the absence of crystal structures for any transporters in the SLC22 family or closely related families, our understanding of the precise mechanisms by which these transporters function remains poor at best. Much more work needs to be done on SLC transporters, including the elucidation of crystal structures, to expand our understanding of their impactful role in human biology and physiology.

Chapter 2 addressed approaches and challenges to interpreting genetic variation in clinically important genes. In the era of advanced genome sequencing, we can envision a future in which the genome of every individual is sequenced at birth. That genomic data has the power to

revolutionize healthcare practices from diagnostic to therapeutics – once we know what to do with it. At present, accurate interpretation of the effects of genetic variants is far behind identification of said variants. With each individual harboring millions of variants compared to a reference genome, detailed review of every variant for each individual is outside the scope of human capability. Advances in machine learning and artificial intelligence will be likely to bridge the gap, with much improvement in performance needing to be made first. Emerging experimental techniques, including deep mutational scanning, are increasing the scale at which we understand the functional or phenotypic effects of variants. With more data, we can build better models to interpret genetic variation, eventually enabling the possibility of a genomic learning healthcare system to process genomic data at birth into clinically actionable recommendations for lifelong improvement of healthcare.

Chapter 3 presented the characterization of a large set of OCTN2 variants and the development of protein-specific machine learning models for predicting variant effect on function with relevance for the rare disease Carnitine Transporter Deficiency (CTD). While a few hundred OCTN2 variants have been identified in patients with CTD, the disease-causing potential of the thousands of thus undetected variants is unknown. Knowledge of the function of an individual's set of OCTN2 variants can be informative in clinical decision making. Here, we were limited by experimental methodologies with a radioactive substrate in the number of variants we were able to characterize. Regardless, we identified mechanistic trends amongst loss-of-function variants and built a protein-specific variant effect model that outperforms existing models in predicting variant function. We can only imagine the power that larger datasets would supply to machine learning for even more substantial improvements to variant interpretation. In addition, we emphasize the importance of designing functional genomic studies to include variants from

diverse ancestries, and the benefit that inclusive machine learning models will have for health equity. The identification of improper membrane localization as a major loss-of-function mechanism opens the door to innovative therapeutic approaches for CTD, with future efforts needed to improve treatment options and clinical outcomes for those affected by CTD.

Next, we aimed to address the limited throughput of transporter variant phenotyping faced in Chapter 3. In Chapter 4, we described the development of a deep mutational scanning (DMS) platform to scale up functional characterization of the pharmacogene OCT1. Comprehensive DMS of all possible variants has not been done for any SLC transporters to our knowledge. OCT1 is known to transport more than 150 cationic drugs and endogenous molecules, and is highly polymorphic with variants having documented substrate-specific functional effects. We propose that DMS of OCT1 will be groundbreaking at both the molecular and physiologic level. Experimental approaches characterizing multiple diverse phenotypes will aid in the identification of variants with significant functional and substrate-specific effects. Data from DMS can even be used to resolve protein structures, which would be revolutionary for the SLC22 transporter family. Notably, these data will be important in clinical implementation of pharmacogenetic dosing for OCT1 substrates, whereby an individual may need to be prescribed another dose or a different medication altogether to best tailor treatment to their genetic code. Additionally, rich layers of data from DMS of OCT1 may be beneficial for improving the interpretation of genetic variation in other related transporters in the SLC22 family and beyond with the help of artificial intelligence.

In conclusion, this dissertation aims to address and tackle challenges with translating genetic data into the clinic. With this work, we demonstrate the impact of understanding phenotypic consequences of variants in genes and proteins with clinical relevance, both at the molecular and

physiologic levels. Though much work remains to be done, we show that through the bridging of experimental and computational approaches, significant improvements to variant interpretation are possible and will one day enable better translation of genomic information in healthcare.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Megan Kolleske
DEC6E616F18F45E...
Author Signature

5/27/2022
Date