Yeast-based assays for studying the functional impact of missense variants in a rare human disease gene at scale

Michael Xie

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Aimée Dudley, Chair

Douglas Fowler

Julian Simon

Program Authorized to Offer Degree:

Molecular Engineering

University of Washington

Abstract

Yeast-based assays for studying the functional impact of missense variants in a rare human disease gene at scale

Michael Xie

Chair of the Supervisory Committee:

Aimée Dudley

Department of Genome Sciences

Advancements in high-throughput sequencing technologies have accelerated the discovery of human genetic variation. However, leveraging genomic information for precision medicine is currently limited by the relatively small number of variants for which there is enough supporting evidence to interpret them clinically. An example of clinically actionable diseases for which large-scale functional data can have an enormous impact on patient health is for serine biosynthesis defects; a group of rare inherited metabolic disorders caused by pathogenic variants in *PHGDH*, *PSAT1*, and *PSPH*. However, because L-serine supplementation, especially if started early, can ameliorate and in some cases even prevent symptoms, knowledge of pathogenic variants is highly actionable. Here, we use a yeast-based complementation assay to measure the functional impact of 1,914 amino acid substitutions in human PSAT, ~88% of all unique SNV-accessible missense variants. Our assay scores agree well with known biological features of the enzyme and existing clinical annotations, supporting its use as functional evidence for variant interpretation. We then extend this approach to assay a subset of pairwise *PSAT1* allele

combinations in yeast diploids. Results from our diploid assay successfully distinguish patient genotypes from those of healthy carriers and agree well with disease severity. Additionally, we developed a linear model that uses individual allele measurements (in haploid yeast cells) to accurately predict the biallelic function (in diploid yeast cells) of ~1.8 million allele combinations corresponding to potential human genotypes. Finally, we present a method that could be used to experimentally measure large numbers of variant combinations in yeast diploids. Taken together, our work provides an example of how large-scale functional assays in model systems can be powerfully applied in the study of rare disease and to inform future diagnostic efforts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many kind, talented, brilliant, and patient people. I would like to first thank my advisor, Aimée Dudley. Her guidance, mentorship, and passion for science has shaped who I am today and who I aspire to be in the future. I was continually met with compassion and enthusiasm to perform impactful research, and guidance to connect with people that shared those values. I am thankful for Maitreya Dunham, Doug Fowler, and Julian Simon for serving on my committee. Everyone's openness and willingness to teach and train myself as well as others, while being leaders in the field, remains inspiring to me. I started in the lab with brief wet lab and data science skills and am grateful for all the mentorship and help from Russell Lo, Gareth Cromie, Amy Sirr, Cathy Ludlow, Lauren Ames, Michelle Tang, Julee Ashmead, and Marty Timour.

I want to thank my family for their support through every endeavor I have ever taken on. I will always be grateful for the encouragement and support from my mother and father to pursue my passions. Thank you to my older brother Tim, for always being a role model for me. Thank you to my younger brother Shaun, for motivating me to be a role model for you. Becoming a first-generation doctoral graduate is something that will be held close to my heart and is proudly a culmination of all the support that I have received from so many wonderful people. Finally, thank you to my partner, Amber Mayle, for the constant unconditional love and always being there for me.

Chapter 1: Introduction

1.1 Challenges in variant interpretation

One of the main goals of human genetics is to understand how genetic variants impact human health. In particular, the identification of sequence variants that influence traits related to the predisposition, onset, and progression of human disease. Advancements in high-throughput sequencing technologies have accelerated the discovery of genetic variation. Since the landmark work to draft the human genome [1], the estimated cost of sequencing a human genome has reduced from 100 million to less than 1,000 USD over the last two decades [2]. As a result, it is possible to generate large amounts of genetic data from many individuals. The Genome Aggregation Database (gnomAD) [3] represents the largest publicly available aggregation of sequence data from over 140,000 people. Data from the gnomAD [3] found ~4.6 million missense variants from these individuals and suggest that the average person harbors approximately 200 rare (allele frequency <0.1%) protein coding variants [4]. Therefore, one major challenge lies in distinguishing disease-causing variants from the enormous number of individual variants that can exist in each human genome.

Analytical approaches, such as family-based linkage and case-control association studies, have driven tremendous growth in the discovery of rare and common disease-associated genetic variation during the same period [5]. While these approaches can associate genes with particular diseases, determining the role of individual variants in pathogenicity is challenging [6]. As genetic testing continues to be widely applied as a diagnostic tool, clinicians are tasked with drawing meaningful conclusions from the variants that arise from the resulting reports.

The American College of Medical Genetics and Genomics (ACMG) developed guidelines to address this need and provide a standardized framework to interpret sequence variants clinically

[7]. A criterion of evidence derived from datasets such as population, computational, functional, and segregation data, contribute to whether a high confidence classification can be made. However, in the absence of enough supporting evidence they are classified as variants of uncertain significance (VUS) and provide no clinical utility. Within ClinVar [8–10], a publicly available archive of variation interpretations, the total number of reports and VUS continues to skyrocket. Single nucleotide variant (SNV) entries encoding missense substitutions has increased nearly 4.5-fold, from ~140,000 to over 630,000, over the last 5 years. Majority (~78%) of these variants, even those in well-studied disease genes, are classified as VUS (**Figure 1-1**).

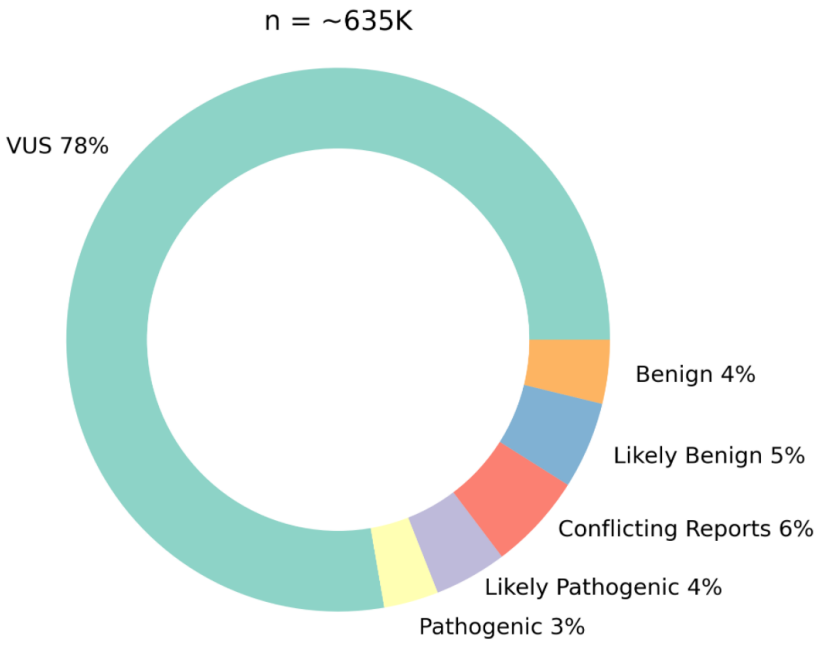

Figure 1-1. Distribution of reported clinical interpretations for ClinVar SNV entries encoding missense substitutions. Interpretations with reports of both benign and likely benign were grouped into the likely benign category in this plot. Similarly, interpretations with both pathogenic and likely pathogenic were grouped into the likely pathogenic category in this plot. The data presented here was taken from the January 5th, 2023, release of ClinVar [10].

It is evident that the discovery of novel sequence variants greatly outpaces the ability to interpret them clinically. This presents a critical roadblock in realizing the potential of precision

medicine to provide improved patient care. To be able to infer pathogenicity for rare variants that will only be present in a small number of people, comprehensive approaches are required to tackle the scale of potential variation. Considering only a single type of variation in protein coding genes, over 60 million missense variants are possible [11]. Furthermore, as humans are diploid organisms, each variant can exist in homozygous and heterozygous states and the possible number of allele combinations increases exponentially.

1.2 Existing Strategies

Computational methods are one approach that can make variant effect predictions at scale. Predictors can be built from leveraging features such as evolutionary conservation, protein structural data, and population data [12]. For example, EVmutation [13] is an unsupervised statistical method that can predict the effect of protein variants from sequence co-variation in multiple sequence alignments. Precomputed effect scores for all possible amino acid substitutions at each position for every human protein can be made readily available. Although computational prediction methods can be applied at scale, their relatively high error rates limit their clinical utility [12].

Functional assays are another approach that experimentally measures the functional consequences of variants in a model system. Advances in high-throughput DNA synthesis and sequencing technologies have facilitated the development of Multiplexed Assays of Variant Effect (MAVEs), in which the functional effects of thousands of variants can be measured simultaneously [14]. Over the last decade, studies applying MAVEs have generated functional effects for over 200,000 variants [15]. MAVEs share a similar overall framework where variants are constructed, implemented into a model experimental system in which genotype is linked to a selectable

phenotype, and library sequencing is used to determine the frequency of each variant before and after selection [16]. MAVEs can be applied to study a variety of functional genetic elements such as UTR, promoters, mRNA, and proteins.

Deep mutational scanning is a high-throughput approach that can experimentally measure the impact of all possible amino acid substitutions at each position along a given protein [17]. Upon application of selection pressure, variants can be distinguished by their effect on phenotype. A variety of selection pressures can be utilized such as fluorescence or growth. Fluorescence based selections can stratify variants based on their levels of fluorescence and the resulting sorted bins can be sequenced for relative variant frequencies. Variant abundance by massively parallel sequencing (VAMP-seq) [18] represents one approach that utilizes protein variant-fluorescent reporter fusions to measure the levels of steady state protein abundance in cultured human cells. Another method, click-seq [19], leverages click-chemistry and fluorescently labeled activity-based probes to measure variant enzyme activity in yeast cells. In cellular growth assays, variants with wild type like function grow normally and variants exhibiting loss of function have their growth impaired [15].

Genetic complementation is one approach that leverages the capability of human protein coding sequences to functionally replace their orthologs in model organisms. Cellular fitness is then linked to the human protein's overall ability to carry out required function. Any relative changes in fitness from human protein variants can then be attributed to changes in overall function. One study showed that a panel of yeast-based complementation assays for 179 variants in 22 human disease-genes could more accurately predict pathogenic and benign variants over computational methods [20]. *Saccharomyces cerevisiae* is a strong model organism to establish complementation assays, as many human-yeast orthologs that encode proteins with conserved

function in many biological processes. A recent large-scale study showed that 200 essential yeast genes could be replaced by their human ortholog and predicts a high success rate (80-90%) for establishing complementation assays for metabolic genes [21]. Coupled with low cost and ease of genetic manipulation for this model organism, yeast-based functional assays provide a strong platform for high-throughput assessment of variant effect in many human protein coding genes.

1.3 Serine biosynthesis defects

Inborn errors of metabolism (IEM), genetic alterations that disrupt metabolic homeostasis, are a relatively large class of rare diseases. Over 700 IEMs [22] have been described and new disorders continue to be discovered [23]. Although individually rare, collectively these diseases are common, with estimates suggesting that IEMs may affect 1 in 800-2600 live births [24,25]. Many are severe and manifest early in life, but some are also medically actionable and timely diagnosis can prevent the onset of irreversible damage. In these diseases, disruptions of metabolic pathways can lead to toxic levels of substrate accumulation or deficiency of an essential product, and therapeutic strategies often focus on amending these imbalances [26]. For example, dietary restriction is prescribed for the treatment of urea cycle disorders, phenylketonuria, and galactosemia, while dietary supplementation is prescribed for homocystinuria and pyridoxine-dependent epilepsy [26–28].

Serine biosynthesis defects are a group of clinically actionable IEMs of varying severity that were first described in the 1970's, characterized biochemically in the 1990's, and mapped genetically in the 2000's (**Figure 1-2**). Impairment of any of the three L-serine biosynthesis pathway enzymes, phosphoglycerate dehydrogenase (PGDH; encoded by *PHGDH*), phosphoserine aminotransferase (PSAT; encoded by *PSAT1*), and phosphoserine phosphatase (PSP; encoded by *PSPH*), results in systemic serine deficiency [29]. Serine metabolism is central

to numerous biological processes, including the synthesis of proteins, nucleotides, and phospholipids, as well as the formation of the neuromodulators D-serine and glycine [29–31]. Serine biosynthesis defects present in a broad phenotypic spectrum that includes, at the severe end, Neu–Laxova syndrome (NLS), a lethal multiple congenital anomaly disease, intermediately in the form of infantile serine biosynthesis defects with severe neurological manifestations and growth deficiency, and at the mild end, as childhood disease with intellectual disability [29,32]. Case studies have demonstrated that oral serine supplementation can reduce and, in some cases, prevent the onset of these severe symptoms that typically manifest very early in life [33–37]. Prompt diagnosis is crucial, as the impact of therapeutic intervention, including prenatal dietary supplementation [33], is greatest before patients become symptomatic and irreversible neurological damage occurs.

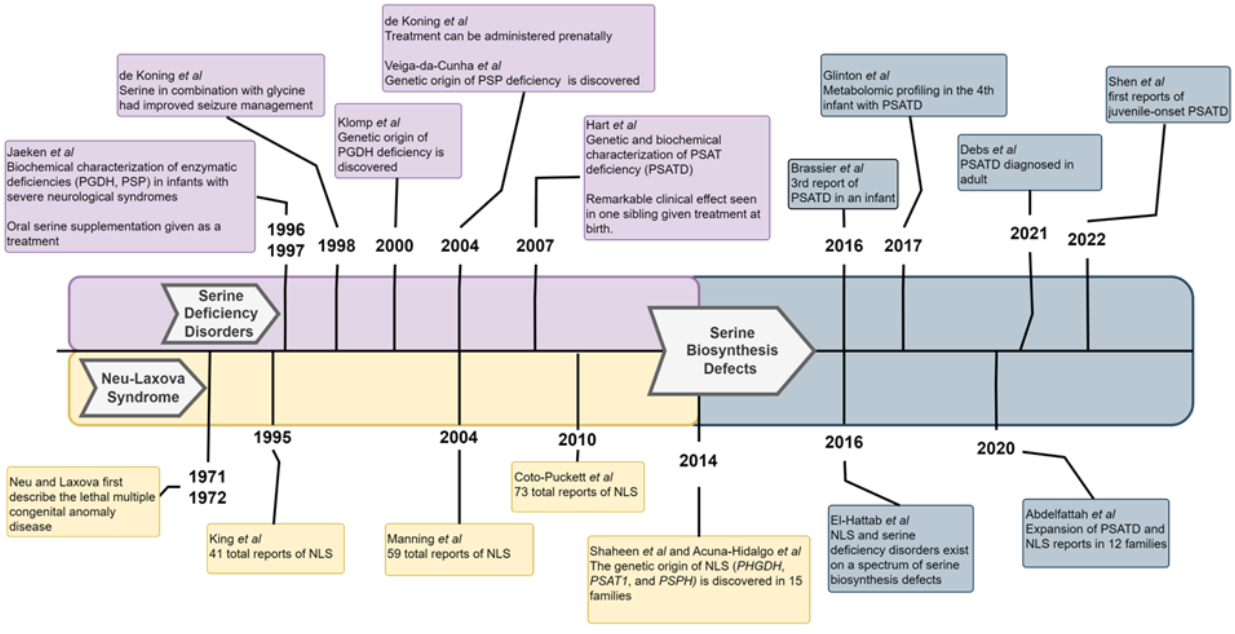

Figure 1-2. Key developments in the study of serine biosynthesis defects. A set rare but treatable inborn errors of metabolism [29,33–51].

Unfortunately, like many rare diseases, diagnosing serine biosynthesis defects is difficult in the absence of a family history of disease. One challenge is that the disorders display a broad phenotypic spectrum, as described above. Another challenge is that biochemical measurement of amino acid levels in patient serum or cerebral spinal fluid (CSF) requires comparison to closely age-matched controls [52]. Although the serum-directed metabolic screen is more widely used, generally considered in children with intellectual disability, serine and glycine plasma concentrations can be normal in the presence of serine biosynthesis defects if the blood sample is not obtained in a fasted state [48,53]. In contrast, CSF serine and glycine concentrations are not affected by meals, however, CSF amino acid analysis is not typically done in the absence of seizures [48,53]. Therefore, the metabolic work up may fail to identify children with serine biosynthesis defects. Furthermore, while significantly lower levels of glycine and serine serum or CSF can indicate suspected cases, the relative levels of impairment do not correlate with disease severity [32].

While *in vitro* biochemical assays that measure the enzymatic activity of fibroblast extracts or purified recombinant proteins are available for all three enzymes [34,35,37,41,42], they have only been applied in a handful of suspected cases to provide further molecular confirmation. For PSAT, there are only two studies that characterized the enzymatic activity of recombinant protein variants [37,47]. These approaches are limited in throughput and present barriers for providing timely diagnosis for current and future patients. Although it has been hypothesized that residual enzyme activity may explain the phenotypic variability [46,49,54] within serine biosynthesis defects, functional studies regarding this subject are sparse and the natural histories of the disorders are largely unknown.

1.4 Overview of thesis

The work presented in this dissertation aims to demonstrate how high-quality large-scale data can inform our understanding of biology and future diagnostic efforts. I begin with an example of how large-scale functional data can be applied in the study of enzyme function and in the context of a rare metabolic disease. In Chapter 2, I apply a high-throughput yeast-based cellular assay to measure the impact of 1,914 amino acid substitutions on human phosphoserine aminotransferase (PSAT; encoded by *PSAT1*) function. I extend this approach to assay a subset of patient *PSAT1* genotypes in a diploid yeast assay and develop a computational model that makes predictions for ~1.8 million potential allele combinations. In Chapter 3, I present an alternative strategy to improve methods for generating double barcoded combinatorial libraries. This approach could be used to experimentally measure large numbers of variant combinations in yeast diploids. In Chapter 4, I summarize the research described in this dissertation and outline future areas of study and the broader context of this work in the field of clinical genomics.

Chapter 2:

Predicting the functional effect of compound heterozygous genotypes from large scale variant effect maps of missense variation

The work described here was performed by Michael J. Xie and collaborators. Aimée M. Dudley conceived the project and supervised all experimental work. Michael J. Xie constructed and assayed the variant library. Michael J. Xie and Martin S. Timour performed the DNA sequencing. Gareth A. Cromie, Michelle Tang, and Katherine Owens developed computational pipelines. J. Nathan Kutz supervised software development. Gareth A. Cromie and Michael J. Xie performed all data analysis with supervision by Aimée M. Dudley. Michael J. Xie performed all protein structure analysis with supervision from Richard N. McLaughlin. Gareth A. Cromie and Michael J. Xie performed all comparisons to published clinical data with oversight by Ayman W. El-Hattab. Michael J. Xie, Gareth A. Cromie, and Aimée M. Dudley wrote the manuscript and incorporated comments from Ayman W. El-Hattab, Richard N. McLaughlin and J. Nathan Kutz. Writing from this chapter is under revision and has been submitted as a manuscript under the same name.

2.1 Background

Identifying the presence of pathogenic sequence variants through whole exome/genome sequencing has several potential advantages for diagnosing inherited metabolic diseases, such as serine biosynthesis defects. However, this approach relies on having clinical interpretations for rare variants, which are generally not available. Instead, within the current disease literature, clinical sequencing has primarily been used as a tool by experts to provide molecular confirmation once a patient has become symptomatic. High-throughput approaches to variant interpretation that can be applied to rare variants are needed to increase the success rate of sequencing-based diagnostics. Functional assays that can quantitatively measure variant effects on protein activity [16] are an approach that can be used as part of the criteria for variant interpretation established by the American College of Medical Genetics (ACMG) [7]. Improvements in DNA synthesis technology have enabled time-and-cost effective methods for building large variant libraries. When combined with high-throughput phenotyping methods, functional assays can

comprehensively assess variants that have been identified in the human population as well as those that may arise in the future [14,15].

Recently, our laboratory established a yeast-based assay for human *PSAT1* function [55]. PSAT catalyzes the reversible conversion of glutamate to α-ketoglutarate and 3-phosphohydroxypyruvate (3-PHP) to phosphoserine [56]. PSAT belongs to a large family of fold type I pyridoxal phosphate (PLP; the active form of vitamin b6) cofactor-dependent enzymes, and more specifically, the class IV group of aminotransferases [57,58]. The crystal structure of human PSAT complexed with PLP (PDB: 3e77; residues L17-L370) shows that the protein adopts an s-shaped homodimer assembly in which two separate PLP-containing active sites surrounded by an overall positive charge distribution reside along the interface (**Figure 2-1.B**). Each subunit of the homodimer consists of two domains; a large domain, which forms the dimer interface and binds PLP, and a small domain (**Figure 2-1.A**). Despite extensive literature on the structure and biochemistry of PSAT [56], there is not a comprehensive understanding of which mutations at which positions impair the ability of PSAT to support normal serine biosynthesis.

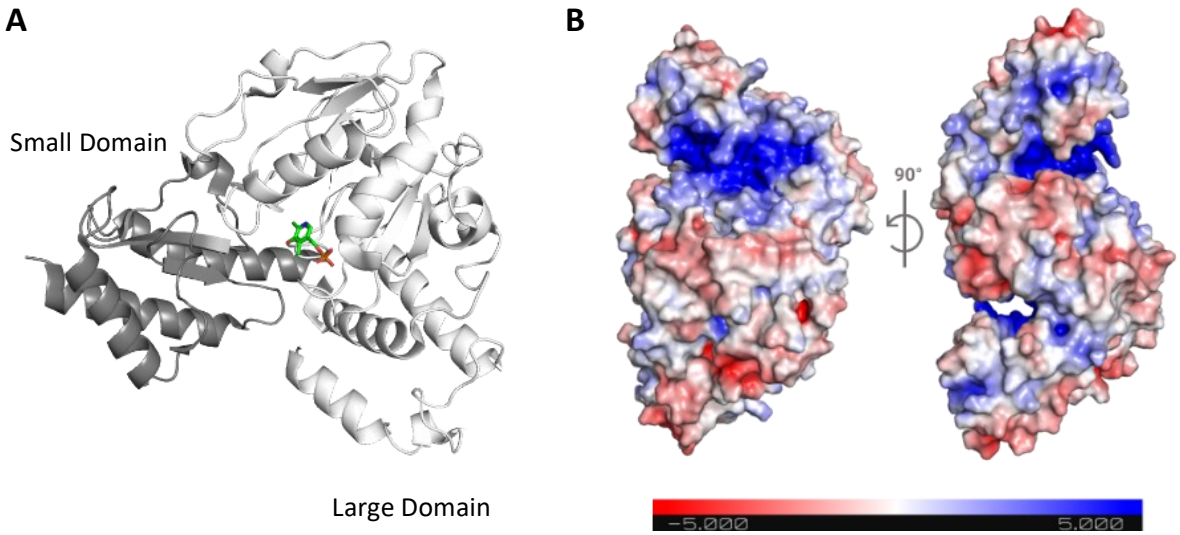

Figure 2-1. Overall subunit organization and homodimer assembly of human PSAT. **A** Ribbon diagram representation of the secondary structure of a subunit of PSAT (PDB 3e77; L17-L370) complexed with PLP (green) in stick representation. The small C-terminal domain is colored as dark grey and large N-terminal domain as white. **B** Local charge distribution on the surface of the PSAT homodimer assembly, on a scale of negative charge (-5 kT; red) to positive charge (+5 kT; blue) as calculated by the Adaptive Poisson–Boltzmann Solver program.

Here, we apply our yeast-based assay at scale to quantify the functional impact of 1,914 amino acid substitutions that are accessible via a single nucleotide variant (SNV) in the human *PSAT1* coding sequence. Our results agree well with clinically interpreted alleles and with protein structure-function relationships, supporting the use of our data as functional evidence under the ACMG interpretation guidelines. In addition to assaying the functional impact of single variants in yeast haploids, we construct and assay yeast diploids with pairwise combinations of *PSAT1* alleles that recapitulate human genotypes. The results of this diploid functional assay distinguish patient genotypes from those of healthy carriers and agree with the stratification of patient genotypes by disease severity. Finally, we developed a mathematical model that can accurately predict biallelic function in diploids from pairwise combinations of individual allele activity measured experimentally in haploids. Taken together, our work provides an example of how large-

scale functional assays in model systems can be powerfully applied in the study of rare disease and to inform future diagnostic efforts.

2.2 Methods

Strain library construction

All *Saccharomyces cerevisiae* strains used in this study (**Supplementary Table S1 and S2**) were derived from the isogenic lab strains FY4 (MAT**a**) and FY5 (MATα) [59]. Unless otherwise noted, strains were grown in rich medium (YPD, 1% yeast extract, 2% peptone, and 2% glucose) or minimal medium (SD, without amino acids, 2% glucose) using standard media conditions and methods for yeast genetic manipulation [60]. Methodology concerning the design and construction of the yeast codon-optimized version (*yPSAT1*) of human *PSAT1* isoform 1, the wild type (*yPSAT1*) strain, the deletion (*ser1Δ0)* strain, and individual variant strains were previously described in Sirr *et al* [55].

Here, we describe the design and construction of a variant library ~10 fold larger and modifications of the growth assay for high throughput as follows: The variant library was designed to capture the amino acid substitutions resulting from all SNV-accessible missense mutations across 369 codons, excluding the start and stop, in the human *PSAT1* isoform 1 cDNA sequence (CCDS6660.1; Consensus Coding Sequence database [61]). The 9 possible single nucleotide variants at each codon resulted in 4-7 unique amino acid substitutions. *yPSAT1* variants encoding the complete set of these unique amino acid substitutions (n = 2,182) were synthesized (Twist Biosciences) in *yPSAT1* as an oligonucleotide library in which each well of a 96 well plate contained all amino acid substitutions (n=4-7) at a given amino acid.

Individual wells of these plates were amplified to approximately 500 ng of DNA by 15-cycles of high-fidelity PCR (Phusion High-Fidelity DNA Polymerase; Thermo Scientific) and

transformed into a MAT**a** haploid deletion strain (*ser1Δ0*) using standard methods. Single colonies from the yeast transformations were isolated such that 6,335 individual transformants, each encoding a single amino acid substitution, were arrayed into 96-well plates containing rich medium. Because individual transformants are isolated and maintained as separate stocks (one strain per well in a 96-well plate), each strain is an independently constructed biological isolate of the variant it contains. For downstream phenotype normalization, each library plate also contained replicates of the same control strains: 2 deletion (*ser1Δ0)* and 4 wild type (*yPSAT1*) strains.

Variant library sequence confirmation

Because of the transformation approach described above, for each *yPSAT1* transformant, we know which codon is mutated (target codon), but not which specific variant is present. To determine this, we used a custom MinION (Oxford Nanopore Technologies) sequencing pipeline. Briefly, individual transformants were pooled in groups of 12, so that no target codon was represented more than once in a single pool. Each pool was then sequenced and, at each target codon, the most frequent potential variant codon was identified (candidate variant) as well as the second most frequent variant. Because we know which target codon corresponds to which transformant, this allows us to associate each candidate variant with a single transformant.

For any given DNA sequence, the frequency and pattern of MinIon sequencing errors varies greatly from base to base. These errors can occur at frequencies high enough to generate spurious matches to variant codons. However, the error patterns are also reproducible, allowing us to develop an error model for each variant, describing the frequency with which it is generated by sequencing errors. This frequency can be compared to the frequency observed for each candidate variant in the pooled sequencing, allowing true variants to be distinguished from sequencing noise.

Using this approach, candidate variants underwent quality control. Candidates were rejected if the observed frequency of the variant was less than 3.3x the estimated error frequency for that variant. In addition, candidates were also rejected if the second most frequent variant was both enriched (≥10-fold) relative to its error frequency and was observed at >30% of the frequency of the candidate variant. Finally, candidates were rejected if they were supported by less than 15 reads, or if any missense or nonsense secondary mutations were present in the *yPSAT1* sequence. The end result of this process was that each transformant was assigned either a high confidence call for the variant present in that transformant, or an NA call that resulted in that transformant being removed from analysis.

Validating the MinION Variant Calls Using Illumina Sequencing

To validate our Oxford Nanopore pipeline, we performed amplicon sequencing on the Illumina sequencing platform to confirm the variant in each well of each variant plate. These calls were then compared to the result from the Oxford Nanopore pipeline. For the Illumina sequencing, the *yPSAT1* ORF was divided up into five ~300bp overlapping segments. The specific segment to be amplified and sequenced corresponded to the known codon position of the variant within our library. 35 cycles of PCR amplification were performed using genomic DNA as a template. The initial PCR reaction was then diluted and used for an additional 24 cycles of PCR to add on Illumina index sequences. The final PCR reactions were pooled and gel purified before performing Illumina sequencing on a Nextseq 500 using 300-cycle, dual-indexed, paired-end sequencing. Reads were demultiplexed and aligned to the *yPSAT1* reference sequence as described [55]. For each isolate, the basecalls across the target codon position were used to identify the variant codon sequence.

Comparison of the MinION and Illumina results revealed excellent agreement (95.5%) for the MinION calls passing the quality filters (**Figure 2-2**). Most of the "variants" failing the MinION quality filters are occasions when, as determined by the Illumina sequencing, no change had actually been made at the target codon (i.e. wild type *yPSAT1* sequence). As expected, for these "no-change" isolates, the most frequent potential variant identified by the MinION pipeline occurred at a low frequency similar to the error frequency for that variant (**Figure 2-2.A**). In contrast, when a variant was present, it was observed at higher frequency and was correctly identified with high accuracy.

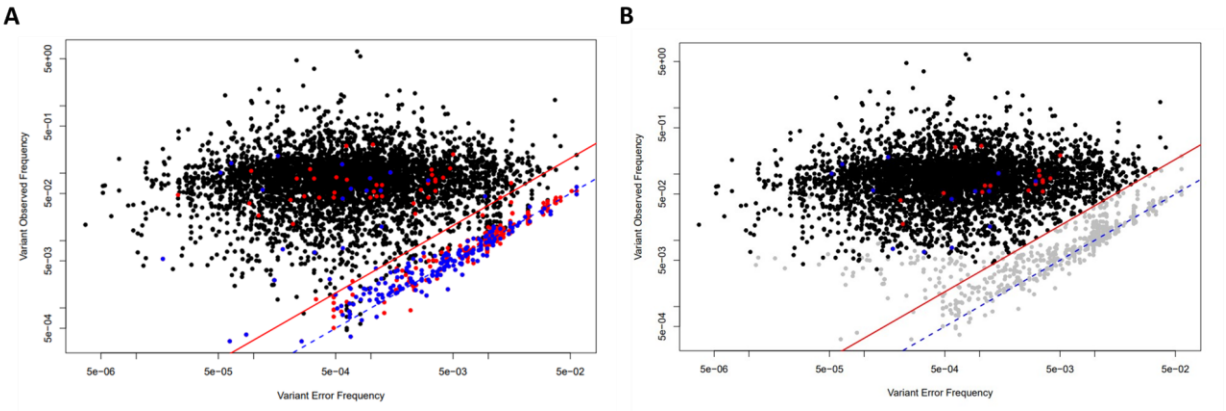

Figure 2-2.Agreement between yPSAT1 variant codons identified by Oxford Nanopore versus Illumina sequencing. For each variant call, the observed ratio of Oxford Nanopore variant to reference reads (expected to be 1/12) is plotted against the empirically calculated error ratio. Blue dotted line indicates 1:1 relationship between observed and error ratios. Quality filters require observed variant frequency to be >3.3 times the error frequency (red line). **A** Variant calls in agreement with Illumina shown in black. Disagreements where Illumina identifies no change at the target codon (i.e. reference sequence) shown in blue. Remaining disagreements shown in red. **B** As in **A** except that variants failing quality control filters are colored in grey.

Generation of Yeast Diploids

Unless noted, all *yPSAT1* variants are in a single mating type (MAT**a**) (Additional File 1: Table S1). To generate diploid strains harboring combinations of *yPSAT1* alleles (missense variant, wild type, or null), variants were first introduced into a strain of the opposite mating type. The

resulting *yPSAT1* MATα strains were then mated in a pairwise manner to the relevant *yPSAT1* MAT**a** strains using standard methods [62]. The resulting diploid strains, harboring combinations of *yPSAT1* alleles (Additional File 1: Table S2), were then arrayed in an alternating checkerboard pattern that minimizes the influence of nutrient competition from neighboring colonies during phenotype measurement.

Growth assays

Before phenotyping, the set of variant strains generated in this study were extended to include the haploid strains carrying missense variants constructed and sequence confirmed by Sirr *et al* [55]. These additional strains were re-arrayed to match the phenotyping layout used in this study, with at least 2 isolates of each variant. Additional control plates included a plate with wild type at every position, and plates consisting entirely of D100A or A99V (with control positions) *yPSAT1* strains, which are ClinVar pathogenic alleles encoding missense variants.

A Biomek i7 robot outfitted with a V&P 96-pin head was utilized to pin strain plate libraries between different culture media. Strains were initially grown to saturation in rich medium and then pinned onto solid medium utilizing glycerol as the central carbon source (YPG,1% yeast extract, 2% peptone, 2% glycerol) to remove any yeast cells lacking mitochondria (petite). Strains were then pinned back into rich medium and grown to saturation. Each plate was then pinned in replicate (n=3-6) onto solid minimal medium, which lacks serine, and grown for 3 days at 30°C. A mounted Canon PowerShot SX10 IS compact digital camera was used to take images (ISO200, f4.5, 1/40s exposure) every 24 hours for three days under consistent lighting, camera to subject distance, and zoom. Each plate was labeled with a custom code39 barcode that was included in the frame of view. Images were acquired as jpg files.

Image-based growth quantification

Each plate image was processed using PyPl8 (https://github.com/lacyk3/PyPl8) to extract features from each strain patch in each barcoded agar plate as follows (**Figure 2-3**): First, the barcode within each image was detected and decoded to rename each file using the corresponding plate name, replicate number, condition, and timepoint. Next, each image was cropped into 96 square tiles, segmented, and each replica pinned patch was identified using Otsu's method or circle detection. Finally, the sum of the gray scale pixel intensities within each strain patch (pixelsum), was extracted and used as the metric for growth estimation.

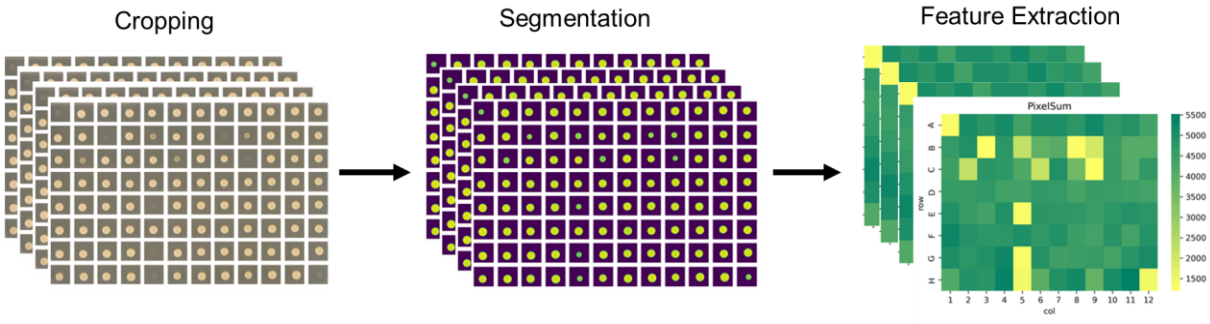

Figure 2-3. Overview of the image-based growth quantification processing pipeline.

Growth data fitting and normalization

For haploid strains, raw phenotypic values were normalized, quality control filters were applied to each isolate, and a final relative growth estimate for each variant was determined, as described in Lo *et al* [63]. Briefly, normalization steps were carried out to account for the effects of plate-to-plate variation, relative growth of neighboring patches, and plate edge effects. Pin effect normalization did not reduce noise and was omitted. Isolates with (nonsynonymous) secondary mutations were removed from the dataset as were all isolates of variants that showed a high degree of variation in isolate-to-isolate growth values. This left a final filtered dataset of 5,164

independent isolates. Finally, a linear model was used to estimate the relative growth of each genotype on a scale with growth of null controls set to 0 and growth of wild type set to 1. The script carrying out the growth normalization steps is provided as Additional File 4.

A similar approach was applied to phenotypic values extracted from the diploid growth assay, although neighbor effects were assumed negligible because of the checkerboard pinning arrangement. For normalization, a linear model was used to simultaneously estimate the effects of genotype, plate edge positioning and plate-to-plate variation on growth. Genotype effects were rescaled to set homozygous null to 0 and homozygous wild type to 1.

Predicting diploid growth

We developed a linear model to predict the growth of diploid strains based on which pair of *yPSAT1* variants is present, using individual growth estimates of each allele in haploids. In this model, for each diploid, the (haploid) growth estimates of the lower and higher growth alleles were labelled as minimum and maximum (j, k), respectively. In cases of homozygous combinations, the minimum and maximum were equal. Strains carrying a single copy of the wild type allele (*yPSAT1*) or null (*ser1*Δ0) had their respective haploid estimates set equal to 1 and 0. To predict diploid growth (d) from the more impaired (x_j) and less impaired (x_k) alleles, we performed an ordinary least squares regression to fit an additive pairwise model ($d = a + bx_j + cx_k$), with resulting coefficients being a=0.05, b=0.28, and c =0.73. We also compared this model to two simpler regression models. In the first model, diploid growth was predicted from the mean of the haploid estimates ($x_j = x_k$). In the second, diploid growth was predicted from the higher growing haploid allele only, i.e. complete dominance ($x_j = 0$). Leave one out cross validation was used to assess model performance in the best model.

Next, we evaluated the performance of experimental versus predicted growth as a binary classifier for identifying genotypes matching those of patients diagnosed with *PSAT1*-related serine biosynthesis defects vs carrier parents. We fit a logistic regression model to both experimentally measured diploid growth values (Log(p/1-p) = e + fy_e) and predicted diploid growth values (Log(p/1-p) = g + hy_p) calculated from the additive pairwise model. The resulting coefficients for the regression models were e=-7.7, f=10.7, g=-20.0, and h=27.2. The logistic regression models indicated a threshold (p=0.5) decision value of 72% and 73% for experimental and predicted diploid growth, respectively.

Protein structure and conservation analysis

Structural features were derived from the crystal structure of human phosphoserine aminotransferase isoform 1 complexed with pyridoxal phosphate cofactor (PDB: 3e77; subunit L17-L370, homodimer biological assembly). Secondary structure features were extracted from this crystal structure using DSSP software [64,65]. AlphaFold (version 2022-11-01) [66,67] was used to predict the structure of the missing 16 N-terminal residues. Molecular visualizations were created using the PyMol Molecular Graphics System Version 2.3.2 (Schrödinger). Tools from the publicly available PyMol script and plugin repository were used to determine interfacial residues and charge distribution. Dimer interface residues were defined as residues at which the solvent-exposed surface area in the monomeric model is greater than the solvent-exposed surface area in the dimer model (cutoff value = 0.5 Å^2). Macromolecular electrostatics were estimated for each residue of PDB 3e77 using the Adaptive Poisson-Boltzmann Solver PyMol plugin. Evolutionary conservation scores and 'grades' for each position of the full-length amino acid sequence of human PSAT (UniProtID [68] Q9Y617-1) were computed using ConSurf [69]. All parameters for the

ConSurf calculation were the same as the methodology outlined in creating the ConSurf-DB repository [70], with the best evolutionary model determined to be WAG. The evolutionary conservation scores and 'grades' represent the calculated positional conservation based on an amino acid alignment of 300 diverse homologs of PSAT. The determined structure and conservation features for each amino acid position is provided in Additional File 1: Table S6.

2.3 Results

2.3.1 Surveying the functional impact of large-scale missense variation in *PSAT1*

Our laboratory previously established a yeast-based complementation assay that leveraged the ability of the human *PSAT1* coding sequence to functionally replace its yeast ortholog, *SER1* [55]. In this assay, growth on minimal medium lacking serine provides a quantitative readout of PSAT activity, allowing the functional impact of protein coding variants to be assessed. Variant impact is expressed on a relatively intuitive scale of activity between the level of yeast growth associated with no activity (that of a complete gene deletion) and that conferred by the wild type human protein coding sequence (*yPSAT1*). Variants causing a reduction in PSAT activity could have their effect via decreases in enzymatic activity, protein stability, or a combination of both.

In this study, we applied the assay at scale to measure the effect of thousands of amino acid substitutions as follows. First, with the exception of the translational initiation and termination codons, we identified all unique amino acid substitutions (n = 2,182) that were accessible via a SNV across the full length of human *PSAT1* isoform 1 cDNA (1,113 bp). Variant codons encoding each amino acid substitution were then introduced into the yeast codon-optimized version of *PSAT1* (*yPSAT1*). The resulting variant library was transformed into a haploid *SER1* deletion strain

(*ser1Δ0*) and integrated in single copy at the *SER1* locus of the yeast genome, under the control of the endogenous *SER1* transcriptional promoter and terminator. Next, transformants were individually arrayed in 96-well plate format, and the identity of the variant codon present in each transformant was determined by Oxford Nanopore MinION sequencing (Methods). Strains that harbored secondary mutations in the protein coding sequence were removed from further consideration. The arrayed strain library was then grown, in triplicate, on solid medium lacking serine and imaged after three days at 30°C. Data from these images were used to measure the growth of each strain, relative to wild type (*yPSAT1*) and null (*ser1Δ0*), using a custom automated image analysis pipeline and normalization software (Methods).

Here, we report the functional impact of 1,914 amino acid substitutions in *PSAT1* (Additional File 1: Table S3), corresponding to ~88% of all unique SNV-accessible amino acid substitutions (n=2,182) in the human *PSAT1* cDNA sequence. A subset of these substitutions (n=196) was also assayed in our previous study [55] which included the full set of *PSAT1* missense variants described in ClinVar, gnomAD, or the clinical literature at that time. Reanalysis in the current study allowed us to assess these amino acid substitutions using an improved data processing pipeline as the 1,718 new substitutions, thereby placing them on a common scale to facilitate direct comparison. Despite the use of slightly different data analysis pipelines, there was excellent agreement (R^2=0.94) between the normalized growth estimates of the 196 substitutions in the two studies.

Among the full set of 1,914 amino acid substitutions, the distribution of growth values was bimodal, with a large peak centered near the value of the wild type *yPSAT1* strain (normalized growth=1) and a smaller peak centered around the value of the null (deletion) strain (normalized growth=0) (**Figure 2-4**). These results are consistent with a large group of protein coding variants

showing little or no functional impairment relative to wild type, and a smaller group behaving as complete loss of function alleles, comparable to the null control. The remaining protein coding variants showed varying levels of functional impairment (**Figure 2-4**). On this basis, we classified amino acid substitutions in our assay as follows. Substitutions with ≤95% normalized growth were considered functionally impaired relative to wild type and the remaining substitutions (>95% normalized growth) were considered unimpaired. Among the impaired class, we further defined any substitutions resulting in ≤5% normalized growth (i.e. comparable to that of the null control) as amorphs, and any substitutions resulting in less severe functional impairment (>5% and ≤95% normalized growth) as hypomorphs.

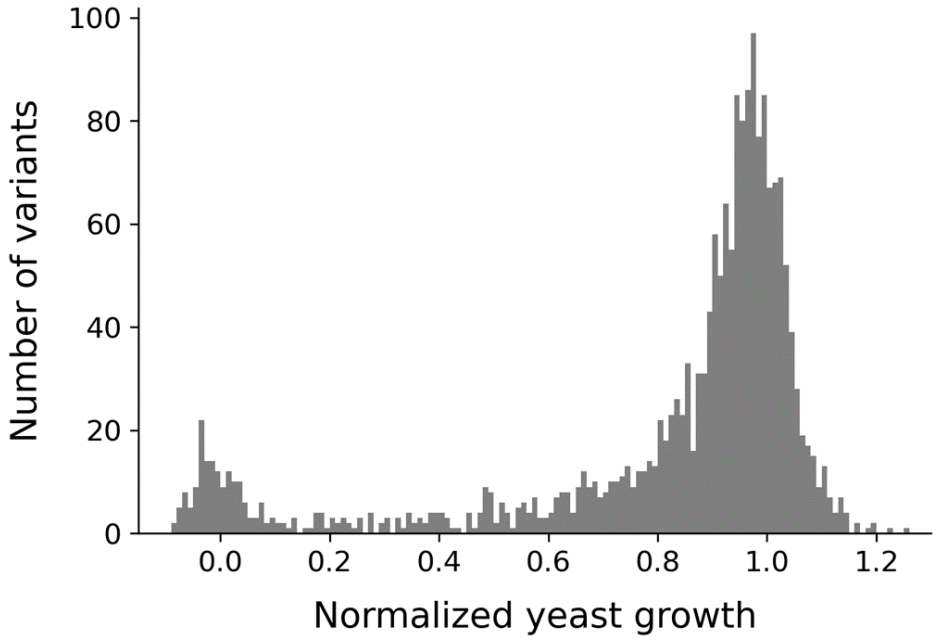

Figure 2-4. Distribution of variant effect for 1,914 PSAT1 unique SNV-accessible missense amino acid substitutions. Frequency (in 5% intervals) of experimentally measured yeast growth values scaled relative to wild type yPSAT1 (normalized growth=1) and null (normalized growth=0).

2.3.2 Mapping functional effects to protein structure and conservation

Our dataset of 1,914 PSAT variants assesses the functional impact of the 4-7 SNV-accessible amino acid substitutions at each position across the length of the human PSAT protein (370 aa). At some positions, the majority of SNV-accessible variants exhibited some degree of functional impairment, while other positions were mutationally tolerant to all sampled variants (**Figure 2-5.A**). To provide a better understanding of the impact of missense substitutions on protein function, we examined the results of our assay in the context of evolutionary and structural features of the protein [40].

The overall structural organization and active site architecture of PSAT are conserved with other members of the class IV family of aminotransferases [58], and phosphoserine aminotransferases from eukaryotic and prokaryotic organisms [71]. To examine our functional scores in the context of evolutionary conservation, we used ConSurf [69] scores (Methods), which are derived from amino acid sequence alignments of homologs. As illustrated in **Figure 2-5**, more conserved (negative ConSurf score) regions localize near the interfacial active sites of the PSAT protein. We expected that highly conserved residues in PSAT would be sensitive to amino acid substitutions. Consistent with this hypothesis, there was a highly significant correlation (Spearman rank correlation=0.47, $p < 7.2 \times 10^{-22}$, **Figure 2-6.A**) between the ConSurf score for a residue and the median yeast assay score for substitutions at that residue. As expected, more conserved residues displayed lower median growth. Examining the relationship between the ConSurf scores at a residue and the growth of individual substitutions at that residue further resolved this relationship (**Figure 2-6.B**). Amino acid substitutions that impaired PSAT activity were concentrated at highly conserved residues within PSAT (**Figure 2-5.C**). However, while few

amino acid substitutions at weakly conserved residues exhibited impairment of PSAT function, many substitutions at conserved positions did not (**Figure 2-5** and **Figure 2-6.B**).

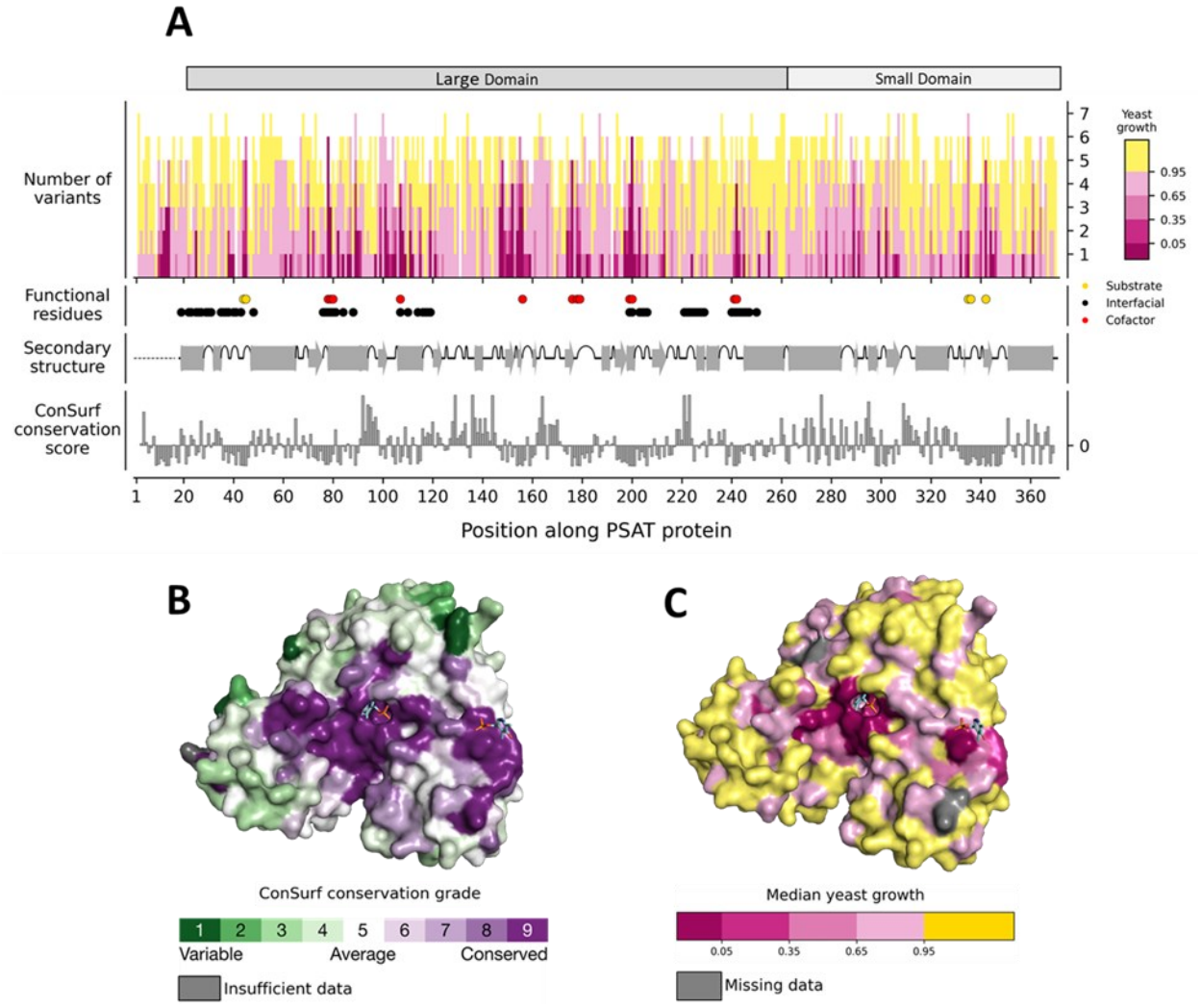

Figure 2-5 Missense variant effect map across the length of human PSAT compared to structural features and conservation. **A** The top plot depicts the growth of the possible 4-7 substitutions introduced at each amino acid position, ordered from highest (top) to lowest (bottom) growth. Overlayed below are the functional residues of PSAT, secondary structure, and ConSurf evolutionary conservation scores. Subunit domains are indicated above. Residues implicated in substrate binding, formation of the dimer interface, or cofactor binding are shown as circles (righthand legend). Helices and beta-sheets are depicted as gray cylinders and arrows, respectively. Turns shown as upward half-coils. The black dotted line represents secondary structures that were unavailable in the solved crystal structure (PDB: 3e77; L17-L370). More negative ConSurf scores indicate more conserved positions. **B** Subunit surface colored by the ConSurf evolutionary conservation grade and **C** subunit surface colored by median yeast growth estimate per position. Two PLP cofactors are shown in each surface representation (**B,C**) to indicate the other interfacial active site bound to the opposite subunit (not depicted).

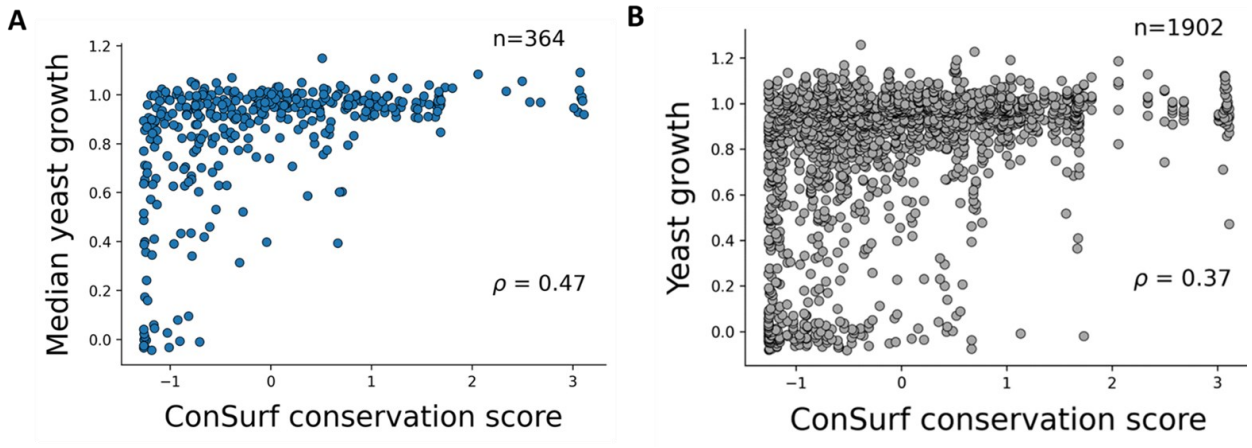

Figure 2-6. Comparison of haploid yeast growth scores and ConSurf conservation scores for each corresponding PSAT amino acid position. More negative conservation scores indicate more conserved scores. **A** Scatter plot of the median yeast substitution score and ConSurf score per amino acid position. **B** Scatter plot of haploid yeast growth scores for each substitution and the ConSurf score for their amino acid position. The corresponding Spearman rank correlation (ρ) is labeled on each plot ($p<7.2 \times 10\text{-}22$ in **A**, $p<1.5 \times 10\text{-}62$ in **B**).

We next considered results of our assay in the context of protein structure. We analyzed a subset of highly conserved functional positions around the active sites, where we see a concentration of residues intolerant to substitutions (**Figure 2-5**). This subset consisted of 12 cofactor binding residues and 5 substrate binding residues. Consistent with the expectation that these sites would be sensitive to amino acid substitutions, we observed that the median growth of variants at these positions was only 7.3%, significantly, and very substantially, lower than the global median of 93.6% ($p< 1x10^{-5}$, one sided permutation test).

We next examined the functional impact of variants at each of these positions individually. In the crystal structure of PSAT complexed with PLP (PDB: 3e77), 12 residues directly interact with the functional groups of the cofactor (**Figure 2-7.A** and **Figure 2-7.B**) either in pyridoxal ring binding/ coordination or in phosphate group binding. We expected that amino acid substitutions at residues involved in pyridoxal ring binding or coordination would be sensitive to amino acid substitutions (**Figure 2-7**.**A**), as this functional group is directly involved in catalysis [58,72,73]. In fact, all tested substitutions at these six residues (K200, D176, T156, W107, S178,

and S179) were functionally impaired. The majority of these, (19/29), including all SNV-accessible missense variants at the catalytic lysine (K200) were amorphic (**Figure 2-7.C**). Next, we examined the six residues that participate in phosphate group binding, which may not participate directly in catalysis and whose functional role is less well understood (**Figure 2-7.B**). The phosphate group may act as an anchor point to the protein [74], or even directly interact with nearby substrates [73]. At three of these positions (G78, G79, and Q199) all variants were functionally impaired and had amorphic median growth scores (**Figure 2-7.C**). Functionally impaired variants were also observed at C80, N241*, and T242*(asterisk indicates residues from opposite subunit), although some substitutions were tolerated (**Figure 2-7**.**C**). Interestingly, most substitutions at the C80 residue are not impaired relative to wild type with the only exception being C80R (growth score=80%). It has been hypothesized that C80 contributes to the higher phosphoserine substrate affinity (Km = 5 μM) observed in human PSAT relative to other phosphoserine aminotransferases [71], and possibly the positive charge introduction represented by C80R negatively impacts this binding.

Because the human PSAT crystal structure was solved only without bound substrate, we considered 5 conserved substrate binding residues that were identified in *Escherichia coli* [75], *Bacillus alcalophilus* [76], and *Arabidopsis thaliana* [77] phosphoserine aminotransferases that had crystal structures of phosphoserine or α-methyl-L-glutamate (analog) bound states. Amino acid substitutions at all but one of these sites, corresponding to H335, R336, R342, H44* and R45* (asterisk indicates residues from opposite subunit) in human PSAT, show strong loss of function (**Figure 2-7.D** and **Figure 2-7.E**). The exception is R336, which displayed activity similar to wild type for all tested substitutions. Thus, R336 may be a conserved residue that does not directly participate in substrate stabilization in PSAT, but does in other orthologs.

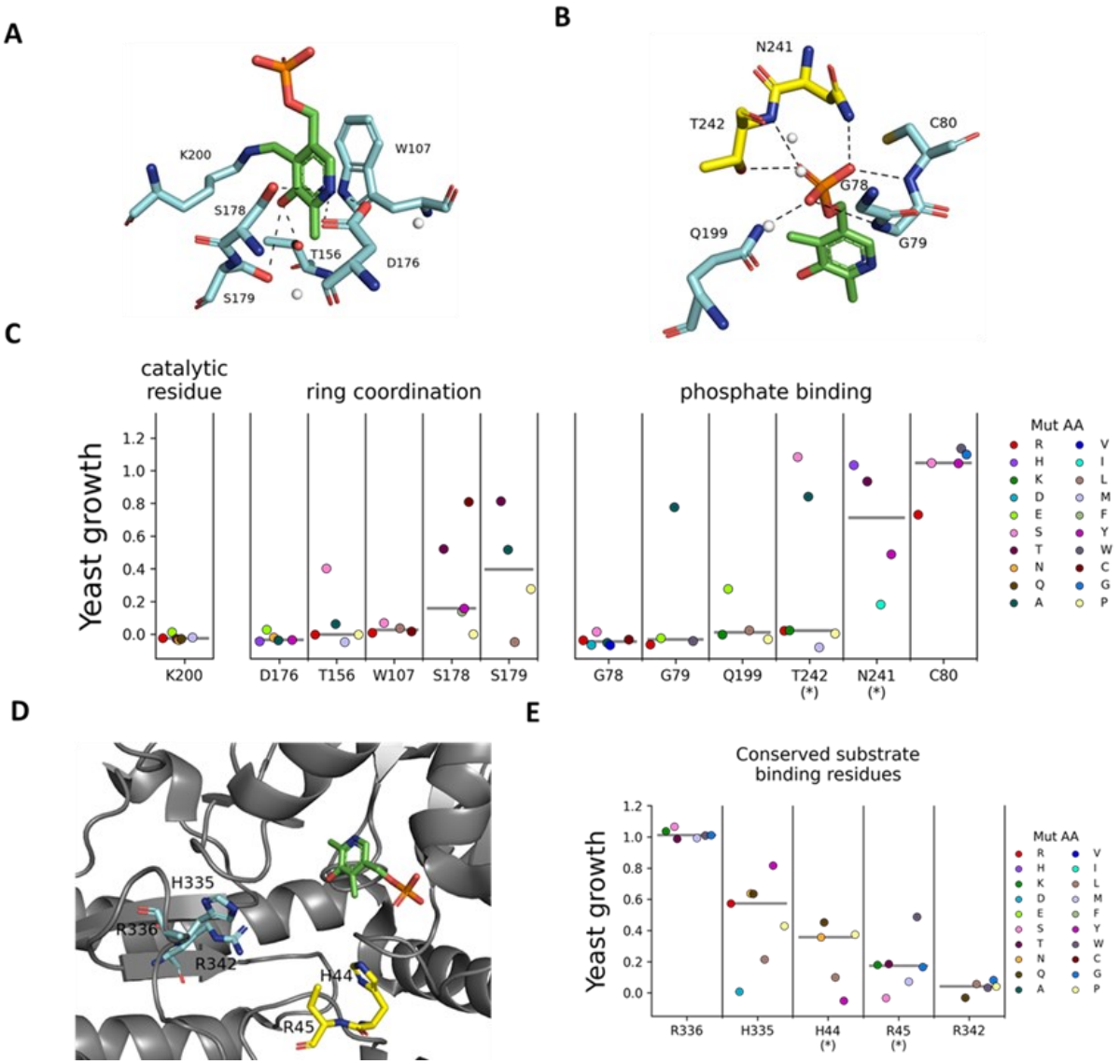

Figure 2-7. Functional impact of amino acid substitutions at active site residues. Stick representations for residues in one subunit are colored in cyan, residues from the opposite subunit in the homodimer assembly are in yellow, and the PLP cofactor in green in all panels where the crystal structure of human PSAT complexed with PLP (PDB: 3e77) is visualized. **A-B** Organization of PLP cofactor binding residues. Hydrogen bonding is depicted by dashed black lines and nearby water molecules are represented by white spheres. **A** Residues that bind or coordinate PLP's heterocyclic ring, with the catalytic lysine shown covalently bound to PLP. **B** Residues that bind to the phosphate group of PLP. **C** The distribution of variant growth scores for cofactor binding residues, colored according to the amino acid that it is substituted as shown (Mut AA). Asterisks here denote amino acids from the other subunit. The median growth score for each position is shown as a horizontal grey bar. **D** Relative position of highly conserved substrate binding residues to PLP, in the unbound (no substrate) state. **E** The distribution of functional

effects for these conserved substrate binding residues, with the same coloring scheme and annotation as in **C**.

The solved human PSAT crystal structure also lacks the first 16 N-terminal amino acids (PDB: 3e77), in which we observe a cluster of sites intolerant to amino acid substitution (residues 11-14) (**Figure 2-7.A**). AlphaFold [66,67] predicts that these residues are located in the small domain, proximal to the active site (**Figure 2-8**). Interestingly, a deletion of 4 N-terminal residues in *Entamoeba histolytica* phosphoserine aminotransferase, corresponding to P12-A15 in human PSAT, yielded a mutant enzyme that had reduced substrate (phosphoserine) affinity, as well as a 10-fold reduction in activity [78]. Additionally, a serine residue involved in binding the analog substrate has been identified in *E. coli* [75], corresponding to P12 in human PSAT. Thus, residues within the first 16 N-terminal residues of human PSAT protein may be involved in substrate binding.

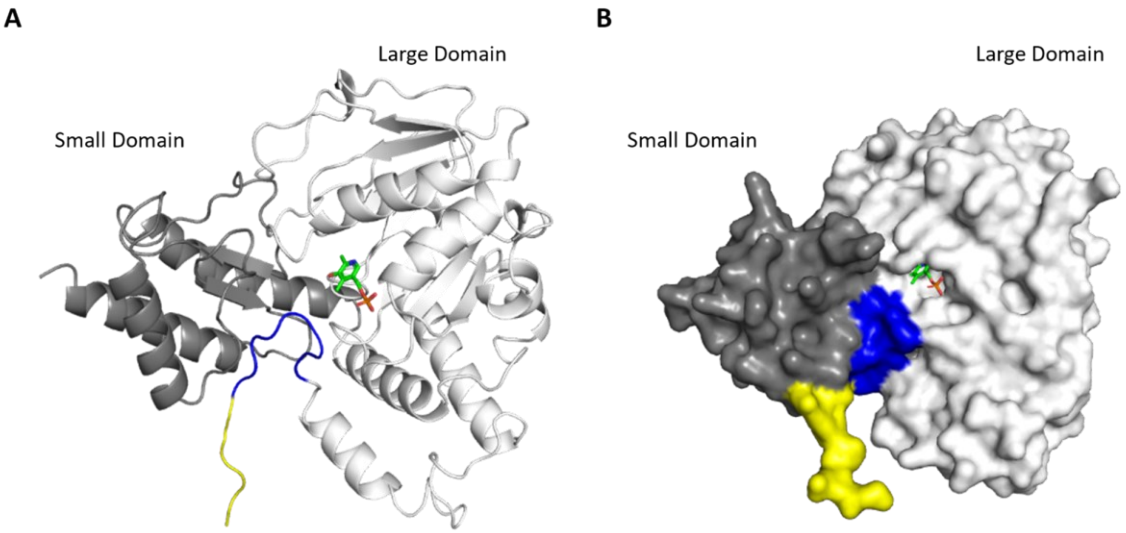

Figure 2-8. Predicted subunit structure for human PSAT. **A** Cartoon and **B** surface representation of the subunit structure for human PSAT (UniProtID=Q9Y617-1, version 2002-01-23) [68], as predicted by AlphaFold (version 2022-11-01) [66,67]. Small and large domains are labeled on both. The PLP cofactor (green) is represented as a stick and was transplanted into the predicted structure using AlphaFill [79], to better visual the location of one of the active sites. Residues with AlphaFold per-residue confidence scores (pLDDT) of very confident ratings (>90) are colored in blue (V8-K16). Residues below this confidence rating of very confident (<90), are shown in yellow (M1-V7).

2.3.3 Mapping functional effects to clinical interpretations

We next compared our results to clinical classifications for all corresponding missense variants in ClinVar. We expected that substitutions derived from known pathogenic variants would produce strong reductions in PSAT activity in our assay, while those derived from known benign variants would have activities comparable to wild type or display only weak reductions in activity. Of the 90 amino acid substitutions derived from variants currently in ClinVar (November 3, 2022, release), three (S179L, A99V, and D100A) are from variants annotated as pathogenic/likely pathogenic, two (I123V and P295R) from variants annotated as benign/likely-benign, and three (V149M, A234S, and R306C) from variants with conflicting interpretations [80]. Since our previous study [55], while the small number of missense alleles with definitive clinical significance calls remained largely unchanged, the number of missense VUS increased nearly 12-fold (from 7 to 82). Consistent with expectation and our previous results [55], the pathogenic/likely pathogenic substitutions all resulted in substantial functional impairment in our assay, with normalized growth values of 82% or below. Similarly, the benign/likely-benign substitutions demonstrated little impairment, with normalized growth values above 91% (**Figure 2-9**).

Data from validated functional assays are potentially valuable as supporting evidence for variant annotation according to the ACMG guidelines [7]. A challenge for rare diseases, such as serine deficiency disorders, is that the limited number of well characterized pathogenic and benign variants precludes the use of current methods, such as odds of pathogenicity [81], for deriving formal measures of classification confidence. As an alternative approach for providing guidance for the use of our data in clinical interpretation, we used a simple thresholding approach, similar that that used in our previous study [55]. Guided by the assay scores of amino acid substitutions with definitive clinical significance calls (**Figure 2-9**), we classified as deleterious any

substitutions with ≤82% growth in our assay, which is less than or equal to the scores of substitutions derived from known pathogenic variants (**Figure 2-9**). We also defined a non-deleterious range for scores ≥91%, which is greater than or equal to the scores of substitutions derived from known benign variants (**Figure 2-9**). Because increases in enzyme activity above wild type have not been associated with serine deficiency disorders, we included values greater than wild type in the non-deleterious range. Finally, because there are no clinically annotated variants associated with substitutions having scores between 82% and 91%, this range of the assay is of unknown clinical significance, and we have labeled this range uncertain. Within these ranges of our assay, 534 (28%) of all amino acid substitutions are deleterious, 1,100 (57%) are non-deleterious, and 280 (15%) are uncertain (**Figure 2-9**). Of the 80 ClinVar missense variants currently annotated as VUS that were tested in our assay, 22 result in substitutions that fall in our deleterious range, 41 in our non-deleterious range and 17 in the uncertain range. Thus, our dataset provides functional information for a large number of variants that lack clinical interpretation, including 77% of the missense VUS currently listed in ClinVar.

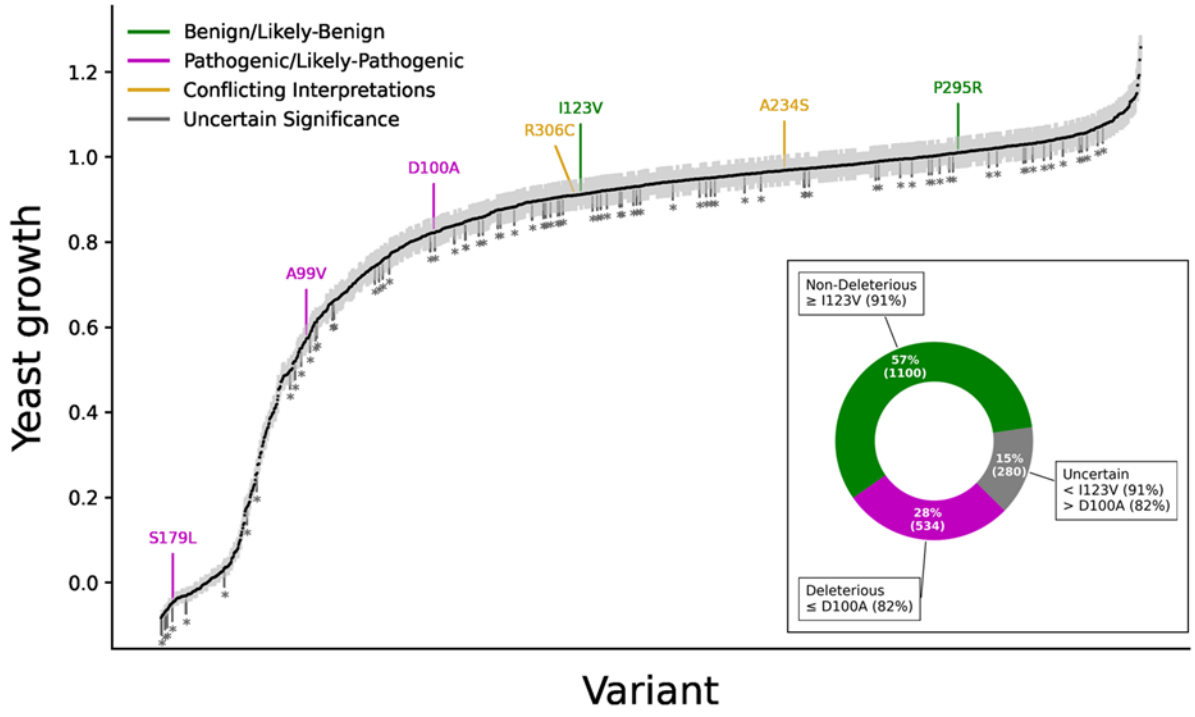

Figure 2-9. Mutational scan of 1,914 PSAT1 unique SNV-accessible missense amino acid substitutions. Rank ordered (lowest to highest) normalized variant growth. Black dots represent the mean normalized growth estimate for each substitution, with light grey bars indicating standard errors. All current ClinVar annotations associated with tested substitutions are labelled above their growth estimate as benign/likely-benign (green), pathogenic/likely-pathogenic (magenta), conflicting interpretations (yellow), or as an asterisk below their growth estimate if they are of uncertain significance (dark grey). The boxed panel depicts the stratification of normalized growth estimates, based on thresholds derived from ClinVar clinical annotation, into functionally unimpaired (green), indeterminate (gray), or functionally impaired (magenta) ranges.

2.3.4 Experimental models of homozygous and compound heterozygous genotypes

Pathogenic variants in *PSAT1* cause disease that ranges from severe (Neu-Laxova syndrome 2, NLS2) to milder forms (PSAT deficiency, PSATD). These are collectively referred to as NLS2/PSATD, or individually when clinical severity is discussed specifically. Because NLS2/PSATD is an autosomal recessive disease, clinical manifestation depends on the enzymatic function conferred by the combination of *PSAT1* alleles in the patient's genome. The ability to generate stable diploid yeast cells by mating haploids allows us to model diploid human *PSAT1*

genotypes in yeast and quantitively assess the function of allele pairs. As in our previous study [55], we used our growth assay to functionally assess pairwise combinations of protein-coding variants across all reported unique patient (and carrier parent) genotypes available at the time (

Table 1). Since our previous study [55], six additional reports [49–51,82–84] have added descriptions of 20 new NLS2/PSATD patients and twelve unique genotypes to the disease literature (

Table 1).

Table 1. Disease literature review of patients diagnosed with *PSAT1*-related serine biosynthesis defects.

Reference (PMID)	Patient	Parent Allele 1	Parent Allele 2	Phenotype
36061210	Shen_Patient_1	c.43G>C (p.Ala15Pro)	c.43G>C (p.Ala15Pro)	PSATD
36061210	Shen_Patient_2	c.43G>C (p.Ala15Pro)	c.43G>C (p.Ala15Pro)	PSATD
34089226	Debs_Patient_1	c.43G>C (p.Ala15Pro)	c.467C>T (p.Thr156Met)	PSATD
26610677	Brassier_Patient_2	c.129T>G (p.Ser43Arg)	c.129T>G (p.Ser43Arg)	PSATD
32579715	Abdelfattah_Patient_6a	c.129T>G (p.Ser43Arg)	c.129T>G (p.Ser43Arg)	Mild NLS / PSATD
17436247	Hart_Patient_1	c.299A>C (p.Asp100Ala)	c.delG107 (p.Gly36Ala_fs*5)	PSATD
17436247	Hart_Patient_2	c.299A>C (p.Asp100Ala)	c.delG107 (p.Gly36Ala_fs*5)	PSATD
29269105	Glinton_Patient_4	c.44C>T (p.Ala15Val)	c.432delA (p.Asp145Met_fs*49)	PSATD
30122079	Zaltsberg_Patient_1	c.44C>T (p.Ala15Val)	c.432delA (p.Asp145Met_fs*49)	PSATD
27161889	El-Hattab_Patient_2	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2 / PSATD
25152457	Acuna-Hidalgo_Patient_1	c.del1023_1027delinsAGACCT (p.Arg342Asp_fs*6)	c.del1023_1027delinsAGACCT (p.Arg342Asp_fs*6)	NLS2
25152457	Acuna-Hidalgo_Patient_2	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2
25152457	Acuna-Hidalgo_Patient_3	c.536C>T (p.Ser179Leu)	c.536C>T (p.Ser179Leu)	NLS2
25152457	Acuna-Hidalgo_Patient_4	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2
25152457	Acuna-Hidalgo_Patient_5	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2
25152457	Acuna-Hidalgo_Patient_6	c.296C>T (p.Ala99Val)	c.536C>T (p.Ser179Leu)	NLS2

32579715	Abdelfattah_Patient_5a	c.1A->G (p.?)	c.1A->G (p.?)	NLS2
32579715	Abdelfattah_Patient_5b	c.1A->G (p.?)	c.1A->G (p.?)	NLS2
32579715	Abdelfattah_Patient_7a	c.181C>T (p.Arg61Trp)	c.296C>T (p.Ala99Val)	NLS2
32579715	Abdelfattah_Patient_7b	c.181C>T (p.Arg61Trp)	c.296C>T (p.Ala99Val)	NLS2
32579715	Abdelfattah_Patient_8	c.235G>T (p.Gly79Trp)	c.235G>T (p.Gly79Trp)	NLS2
32579715	Abdelfattah_Patient_9	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2
32579715	Abdelfattah_Patient_10	c.296C>T (p.Ala99Val)	c.870?1G>T (splicing)	NLS2
32579715	Abdelfattah_Patient_12	c.733T>C (p.Cys245Arg)	c.733T>C (p.Cys245Arg)	NLS2
32579715	Abdelfattah_Patient_13	c.463G>C (p.Glu155Gln)	c.870?1G>T (splicing)	NLS2
32579715	Abdelfattah_Patient_14	c.870?1G>T (splicing)	c.870?1G>T (splicing)	NLS2
32579715	Abdelfattah_Patient_15a	c.955delA (p.Arg319Asp_fs*14)	c.955delA (p.Arg319Asp_fs*14)	NLS2
32579715	Abdelfattah_Patient_15b	c.955delA (p.Arg319Asp_fs*14)	c.955delA (p.Arg319Asp_fs*14)	NLS2
28600779	Monies_16W-0250/16N-0116	c.233G>C (p.Gly78Ala)	c.233G>C (p.Gly78Ala)	NLS2
31903955	Ni_Patient_1	c.208T>A (p.Tyr70Asn)	c.1024C>T (p.Arg342Trp)	NLS2
35885441	Olave_Patient_1	c.296C>T (p.Ala99Val)	c.296C>T (p.Ala99Val)	NLS2

To experimentally measure the functional impact of these allele combinations in our yeast assay, we constructed diploid strains harboring *yPSAT1* allele combinations encoding the same pair of PSAT1 amino acid sequences as the human genotypes. We then assayed the growth of these strains relative to homozygous wild type (*yPSAT1*) and null (*ser1*Δ0) on minimal medium lacking serine. Previously constructued *yPSAT1* diploid strains encoding homozygous A99V or S43R, and the compound heterzygote A99V / S179L [55], were also included for comparision. Although the growth of strains modelling some patient genotypes was close to that of strains modelling some carrier genotypes, all modelled patient genotypes (Additional File 1: Table S5) displayed reduced growth relative to their respective carrier parents and the homozygous wildtype (**Figure 2-10.B**).

Consistent with our previous study [55], we also observed good agreement between the degree of *PSAT1* functional impairment in our diploid assay and disease severity (**Figure 2-10.B**). Diploids corresponding to genotypes from NLS2 patients had normalized growth from 0% to 77%, and diploids corresponding to genotypes from PSATD patients had a higher normalized growth range of 70% to 88%.

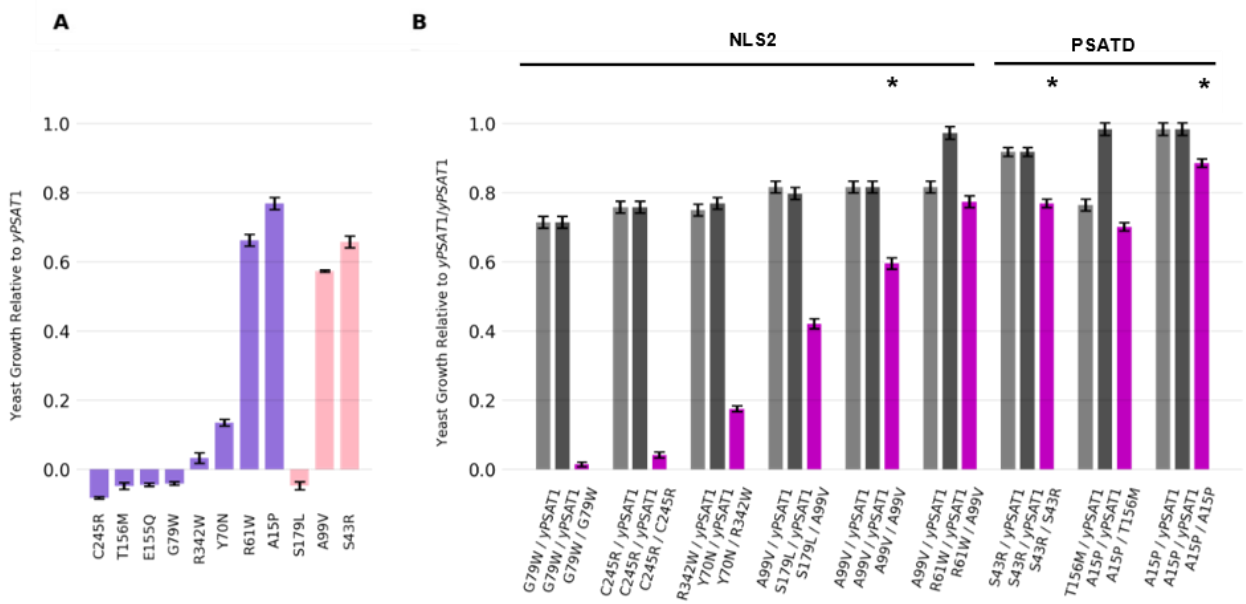

Figure 2-10 Modeling PSAT activity in patient genotypes. Both bar plots show the mean normalized growth estimate and standard errors. **A** Haploid growth estimates for novel (purple) [28–30,64] and established [25,26] (pink) missense alleles in patient amino acid substitutions. **B** Diploid yeast models of patient trios. Carrier parents are shown in light and dark grey, and the offspring in magenta. Recurrent carrier genotypes are repeated for ease of reference relative to the offspring. The clinical severity (NLS2 or PSATD) associated with patient genotypes is indicated. Asterisks denote offspring genotypes that have been reported in multiple unrelated families.

Our assay results are also consistent with clinical stratification within the NLS2 and PSATD patient groups. For PSATD, the A15P / A15P homozygous genotype is the least impaired in our assay (88% growth) and is also the mildest form of PSATD reported to date. A recent report [51] described two unrelated patients that, other than congenital or childhood ichthyosis, were developmentally normal and displayed neuropathy onset at ages 16 and 17. For NLS2, the two

genotypes with the lowest growth in our assay (G79W / G79W and C245R / C245R) correspond to patients displaying severe NLS2 [49]. Three of the remaining NLS2 genotypes modelled here (Y70N / R342W, A99V / S179L, A99V / R61W) display higher growth in our assay and correspond to patients described as having moderate NLS expression [46,49,82]. Among these, Y70N / R342W was the most impaired in our assay (18% growth), and the patient had a postnatal survival of 8 weeks [82]. In contrast, the individual harboring R61W / A99V (77% growth, least impaired) had the longest NLS2-associated postnatal survival described to date (4 months) [49]. Patients homozygous for the A99V genotype (60% growth) exhibit variable postnatal survival (ranging from 1 day to 9 weeks) [46,49,54]. Together, these results highlight the potential value of our model organism assay for variant interpretation in the context of diploid *PSAT1* genotypes.

2.3.5 Predicting biallelic effects from individual allele measurements

Experimentally assessing the level of PSAT activity associated with all allele-pairs would require construction and assays of 1.83 million diploid yeast strains. As a more labor and cost-effective alternative, we evaluated whether a relationship exists between individual haploid estimates and their resulting diploid growth in combination. For homozygous diploids, a linear relationship was seen between haploid and diploid growth estimates (**Figure 2-11**, R^2=0.98).

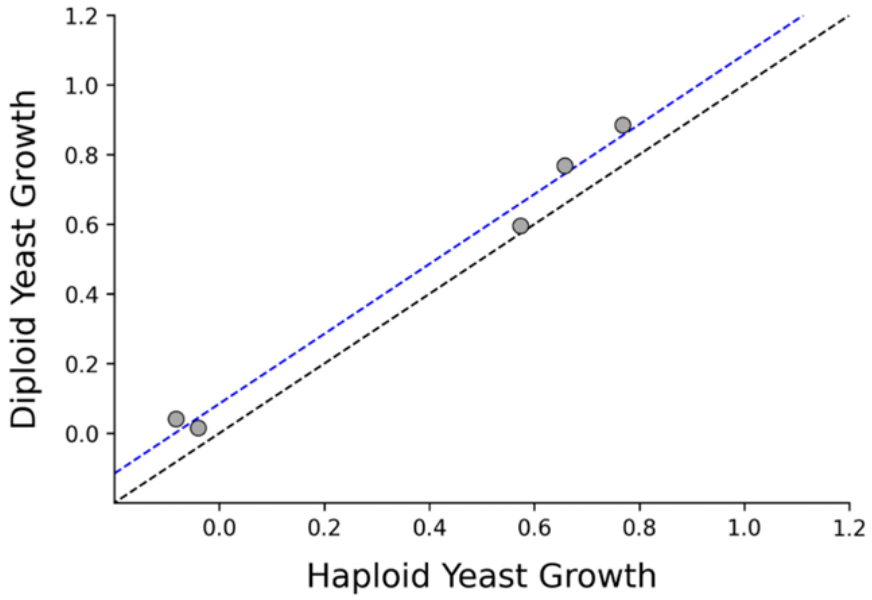

Figure 2-11. Comparison of experimentally measured haploid and diploid growth scores for PSAT variants that are in homozygous patient (NLS2 or PSATD) genotypes. Scatter plot of normalized haploid and diploid growth scores. Haploid scores were scaled relative to wild type *yPSAT1* (normalized growth=1) and null (normalized growth=0). Diploid scores were scaled relative to homozygous wild type *yPSAT1 / yPSAT1* (normalized growth=1) and *null / null* (normalized growth=0). The black dotted line represents a 1:1 correspondence between the haploid and diploid scores. The blue dotted line indicated the observed correlation (R^2=0.98) with slope of 1.0 and intercept of 0.09.

To extend this analysis to heterozygous genotypes, we fit a model to our experimental diploid dataset of both patient and carrier genotypes (n = 23), to predict diploid growth as a linear combination of the allele with the higher and the allele with the lower haploid growth values (**Figure 2-12.A**). This model assumed a uniform degree of dominance of the higher growing allele over the lower growing allele and explained 97% of variance in the growth of the diploids. To compare how this model performed relative to alternatives with simpler assumptions, we also generated models that assumed either complete dominance of the higher-growing allele or an equal contribution of both alleles (**Figure 2-12.B** and **Figure 2-12.C**). Our full model performed significantly better than either of the alternative models (Anova, Df=1; $p=<5.4x10^{-6}$ and $p<2.3x10^{-}$

^{6}, respectively). Leave one out cross validation of the pairwise model displayed an excellent ability to predict diploid growth from haploid measurements (RMSE = 0.0698 and MAE=0.0566).

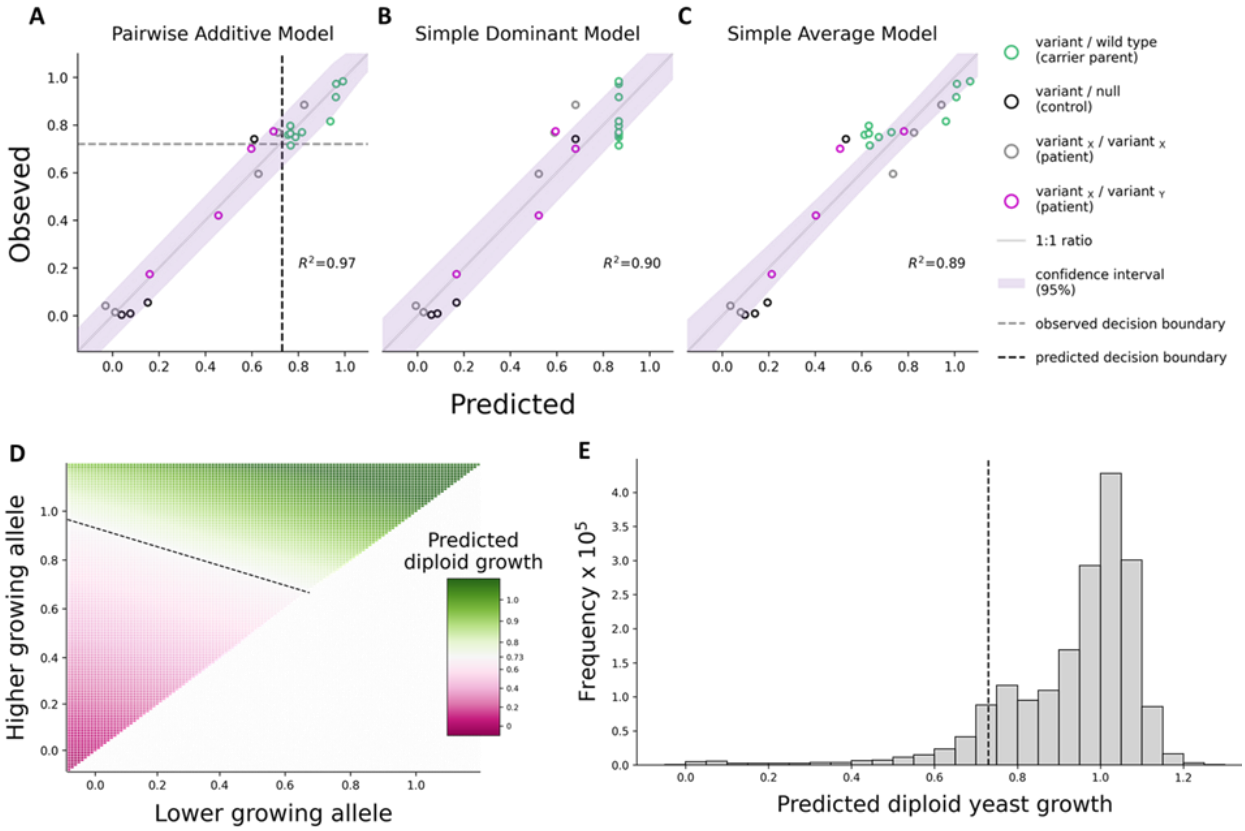

Figure 2-12. Experimental and predicted biallelic growth values as clinical classifiers. Observed versus fitted values shown for three linear regression models (pairwise additive, **A**; simple dominant, **B**; and simple average, **C**) predicting diploid growth (allele pairs) as a function of haploid growth (single alleles). Each circle represents a unique biallelic combination. Patient, carrier, and control strains are labeled as indicated. A diagonal line of perfect correspondence (1:1 ratio) and the coefficient of determination (R^2) for each model are included for ease of comparison. Decision boundaries for observed (72%) versus predicted (73%) diploid growth as a binary classifier for identifying patient genotypes modeled in our diploid assay are shown as horizontal and vertical dotted grey and black lines in **A**, respectively. Confidence intervals (95%) for each model are shaded in purple. **D** Heatmap of the predicted diploid growth value for all possible pairwise combinations of haploid allele measurements. E Frequency of predicted diploid growth values (in 5% bins) for all ~1.8M combinations of alleles. The decision boundary of 73% is shown as a black dotted line in **D** and **E**.

We next compared the results of applying logistic regression to experimental and predicted diploid growth values to produce binary classifiers capable of distinguishing the low growing

genotypes of NLS2/PSATD patients (n=9) from the high growing carrier parent genotypes in our trio models (n=10). The threshold growth values at the decision boundary (p=0.5) were similar for both approaches: 72% for experimental and 73% for predicted diploid growth. Interestingly, the decision boundary using predicted values performed better than that of the experimental values, misclassifying only one patient genotype (A15P / A15P) versus four for the model using experimental values (**Figure 2-12.A**). It is likely the haploid estimates used to make predictions on diploid growth were more accurate than the experimental diploid estimates as they were generated from a larger number of experimental replicates. These results suggest that diploid growth values predicted from haploid measurements perform well in clinical classification of human *PSAT1* genotypes.

On this basis, we computed the predicted diploid growth from all unique pairwise combinations (~1.8 million) of haploid estimates from 1,914 *yPSAT1* missense alleles, wild type (*yPSAT1*), and null (*ser1*Δ0) using the pairwise-additive model **Figure 2-12.A**). We then classified whether the resulting values fell above or below the disease classification boundary of 73% growth. A large majority (89%) of genotypes fell above the decision boundary threshold, suggesting these genotypes would not lead to NLS2/PSATD **Figure 2-12.D** and **Figure 2-12.E**). However, over 200,000 *yPSAT1* biallelic combinations displayed predicted diploid growth below the classification threshold, consistent with these genotypes resulting in disease (**Figure 2-12.E**).

2.4 Discussion

Here, we present a comprehensive functional analysis of amino acid substitutions associated with SNVs in human PSAT. Leveraging such data for clinical interpretation requires an understanding of the relationship between enzyme function and clinical presentation. Our dataset shows clear stratification of existing clinical variants, with benign variants exhibiting little

to no functional impairment, and pathogenic variants exhibiting substantial loss of activity. On this basis, we were able to use the small number of variants with existing clinical interpretation to provide evidence for the likely clinical impact of the large number of variants in similar ranges of the assay. As additional clinical information becomes available through identification of new patients and additional research into the natural history of the disease, the ability to make clinical inferences using our dataset will improve. In the three years since our original study [55], additional patients were described in the disease literature. Not only was there good agreement between the results of our assay and disease severity in these patients, but this additional clinical information also allowed us to extend the range of our assay associated with clinical outcomes.

Given that the therapeutic potential for this medically actionable disorder is highest before patients become symptomatic, clinical sequencing coupled to informative variant interpretation can be a powerful diagnostic tool. Sequencing based diagnostics are becoming more widely applied in newborns, prenatal care [85], and as a part of clinical research efforts such as the Undiagnosed Disease Program [86,87] or Deciphering Developmental Disorders [88]. The most recent AMCG guidelines [89] recommend exome or genome sequencing as a first or second tier option for patients displaying early neonatal (<1 year) congenital abnormalities and for patients exhibiting developmental delay or intellectual disability before the age of 18. In contrast to frameshift and stop-gain mutations, when novel missense variations are identified, even in disease-causing genes, predicting the likely effect on human health is challenging. Relating genotype to phenotype is further complicated by the fact that autosomal recessive diseases are a function of the allele combination in the context of a diploid genome. We can experimentally assay allele combinations in our diploid assay and successfully stratify clinical genotypes. Furthermore, we were able to predict the effects of allele combinations using a computational model that utilizes

combinations of haploid measurements. As a result, we were able to leverage 1,914 yeast measurements to make predictions on 1.8 million potential human genotypes. Given ongoing genome sequencing efforts like the All of US research program and the UK biobank [90], studies that provide information both about individual alleles and the functional impact of allele pairs will play an increasingly important role in realizing the promise of precision medicine.

Chapter 3: Double barcoding combinatorial libraries of yeast diploids with Bxb1 recombinase.

3.1 Background

Advancements in high-throughput technologies have facilitated the development of functional assays that can measure the impact of thousands of variants on protein function [14,15,17]. However, human disease can be influenced by whether a genetic variant is in a monoallelic or biallelic state. A recent study [91] identified variants in known Mendelian disease genes that had an effect in the homozygous and heterozygous states within the Finnish population (FinnGen project [92]; >176,000 individuals). Thus, modeling the impact of variants in allele combinations can inform our understanding of human disease genotype-phenotype relationships in the context of a diploid genome. A major challenge in developing comprehensive approaches for variant effect combinatorics is the scale necessary for experimentation. Thousands of single amino acids substitutions exist for a given protein (N) and measuring all unique pairwise combinations ($N \times N = N^2 / 2$) would increase the size of the screen by several orders of magnitude. Accordingly, large-scale combinatorics require methods to: 1) generate large libraries in a practical way; 2) distinguish combination identities; and 3) have a quantifiable readout that can be multiplexed for high throughput.

Recently, methods for assaying combinatorial libraries have been developed for studying protein-protein and genetic interactions using *Saccharomyces cerevisiae* as a model organism [93–98]. These approaches apply a similar overall strategy that leverages yeast mating for library construction and unique DNA barcodes to identify interacting pairs during sequencing. Haploid yeast libraries (MAT**a** and MATα) encoding a gene or protein of interest are barcoded and mated in bulk to quickly generate all possible pairwise combinations as yeast diploids. Each barcode is

then joined together (double barcode) on a single contiguous DNA strand via *in vivo* recombination to maintain pairwise identity in pooled screening. Selection pressure is then applied to the diploid strain libraries and scored based on the change of double barcode frequency during pooled sequencing. Thus, these methods allow large combinatorial libraries to be constructed efficiently, screened as pools, and are amendable to be tested on multiple conditions.

Regarding *in vivo* recombination, the Cre-lox recombinase system has typically been employed to translocate DNA barcodes onto the same plasmid or chromosome in these yeast-based screens [93–98]. Cre belongs to a family of tyrosine recombinases that rejoins strands of DNA after forming single stranded breaks at lox recognition sites [99,100]. Reactions catalyzed by these enzymes are reversible, resulting in the re-excision of recombined products [101]. The reaction can be biased towards recombined product formation by using orthogonal lox variant sites (e.g lox2272 and lox5171 [102]). Despite the use of lox variant sites in several approaches, low rates of recombination (1-27%) are still reported in yeast combinatorial screens [93,94,97,98]. Therefore, recombination efficiency represents a bottleneck in library construction that can result in missing interaction pairs and more non-uniform distributions in the starting pool. A recent study [103] has shown that the distribution of fitness and frequency of strains in the initial pool of genotypes can impact the accuracy in conventional fold enrichment analysis.

Recombination efficiency can potentially be improved by exploring alternative site-specific recombinases that have different reaction mechanisms. In contrast to tyrosine recombinases like Cre, serine recombinases catalyze unidirectional reactions that re-ligate strands of DNA after forming double stranded breaks at att recognition sites [100,101,104]. These reactions are only reversible in the presence of a recombination directionality factor. Bxb1 is one

type of serine recombinase that has shown high rates of recombination in yeast, mammalian cells, and *in vitro* [100,105,106].

Here we describe an inducible system that utilizes Bxb1-att to recombine barcodes between homologous chromosomes in the genome of diploid yeast. We designed orthogonal donor and recipient integrating cassettes that mark target recombination events with the reassembly of sequence encoding functional green fluorescent protein split with an artificial intron containing the double barcode. Our design allows for downstream library stratification by fluorescent-activated cell sorting (FACS) and does not require the use of auxotrophic markers. We show that complex barcode libraries can be efficiently inserted at barcode landing pad sites in plasmids carrying the integrating cassettes by Gibson assembly [107]. We demonstrate that Bxb1-att improved the estimated lower bounds of recombination efficiency by ~10 fold when compared to a similar design that used Cre-lox [94]. Taken together, this preliminary work highlights the potential utility of alternate recombinases for improving the high-fidelity construction of barcoded combinatorial libraries in yeast diploids.

3.2 Methods

Construction of unbarcoded plasmids

The overall methodology to construct donor and recipient plasmids containing integration cassettes is as follows: generation of DNA fragments with overlapping end homologies, assembling them into a plasmid via Gibson assembly [107], and sequence confirming the constructs with Sanger sequencing (Genewiz). Fragments were ordered as gBlocks (Integrated DNA Technologies; IDT) or were PCR amplified with primers containing 3' overhangs from existing plasmids. All fragments were amplified using a 35-cycle high-fidelity PCR (Phusion High-Fidelity DNA Polymerase; Thermo Scientific) and were either column purified (Zymo-Spin

I; Zymo Research) or gel extracted (Zymoclean Gel DNA recovery; Zymo Research). Amplicons derived from existing plasmids were also digested with DpnI (New England Biolabs; NEB) under the recommended conditions to remove any trace levels of template. Fragments were assembled into plasmids using the NEBuilder HiFi DNA Assembly Kit and protocol (NEB) and transformed into DH5α chemical competent *Escherichia coli* cells (Thermo Scientific). Plasmids were miniprepped (Qigen) from transformants and sequence confirmed by Sanger sequencing (Genewiz).

The vector backbone containing the origin of replication and ampicillin resistance cassette flanked by ~300 bp of upstream and downstream homology to the *HO* locus in *S. cerevisiae* was PCR amplified from pAS34 [108]. Hygromycin (HygMX) and kanamycin (KanMX) resistance cassettes were amplified from pFA6a-hphMX6 [109] and pRS41K [110], respectively. The Bxb1 expression cassette consists of the Bxb1 ORF was amplified from pCMV (pCMV-Bx was a gift from Michele Calos (Addgene plasmid #51552; http://n2t.net/addgene:51552; RRID: Addgene_51552), flanked by an upstream galactose inducible promoter tagged with a N-terminal SV40 nuclear localization signal, and a downstream CYC1 terminator, that were both amplified from pAG415 [111]. The GFP expression cassette, comprised of the GPD yeast promoter, ORF of the yeast codon optimized version of enhanced GFP (yEGFP) [112], and a short synthetic terminator (Tsynth7) [113], were ordered as gBlocks (IDT). An artificial yeast intron (AI) sequence [114] was inserted near the middle of the yeGFP ORF, under recommended spacing (Sasha Levy; personal communication) and was split into two fragments (5’ and 3’).

The 5’ yEGFP-AI fragment and KanMX cassette were placed upstream of the recipient barcode landing pad (NheI restriction site) and Bxb1 attP recognition sequence [105]. All of these fragments were assembled with the vector backbone to generate the pMX-Bxb1-Recipient plasmid

(sequence available as AB724 in Dudley lab LIMS). The 3' yEGFP-AI fragment, HygMX cassette, and Gal-Bxb1 expression cassette were placed downstream of the donor barcode landing pad (AflII restriction site) and attB Bxb1 recognition sequence [105]. All of these fragments were assembled with the vector backbone to generate the pMX-Bxb1-Donor plasmid (sequence available as AB724 in Dudley lab LIMS).

Generating barcoded plasmid libraries

Approximately 5 μg of pMX-Bxb1-Donor and pMX-Bxb1-Recipient plasmids were digested with their respective barcode landing pad associated restriction enzymes, AflII (NEB) and NheI (NEB), to linearize the vector. Linearized vectors were then treated with Antarctic phosphatase (NEB) to reduce the re-ligation of sticky ends before barcode insertion. Fragments were digested with DpnI (NEB), and gel extracted (Zymo Research) to remove trace amounts of residual template plasmid. Single stranded DNA oligonucleotide pools (100-mers) containing a 32 bp barcode containing 20 total random base positions (possible complexity = $4^{20} = 10^{12}$), and 34 bp arms of barcode landing pad homology was ordered (IDT). Stretches of random bases were limited to 2-5 base pairs to prevent the regeneration of undesired restriction sites. 100 ng of linearized vector was assembled with 0.2 μM of the barcode 100-mer (~35x molar excess) using NEBuilder HiFi DNA Assembly Kit in four 20 μL reactions. Assembly reactions were pooled, and column purified (Zymo Research). 5 ng of assembled plasmids were transformed into TOP10 electro competent cells (Thermo Scientific) under their recommended protocol and using these settings on a Gene Pulser Xcell (Bio-Rad): capacitance=25 μF, resistance=200 Ω, and voltage=2.5 kV. Four transformation reactions were pooled into a total volume of 3 mL (cells + SOC medium) and recovered for one hour at 37°C. An aliquot of 100 uL was serially diluted on LB medium

containing ampicillin, to estimate transformation efficiency. The recovered cells (~3 mL) were inoculated into 300 mL of LB medium containing ampicillin, grown for 4-5 doublings at 37°C, and miniprepped (Qigen) to recover ~ 4 ug of barcoded recipient and donor plasmids.

Yeast strain construction

All *S. cerevisiae* strains used in this work were derived from the isogenic lab strains FY4 (MAT**a**) and FY5 (MATα) [59]. Unless otherwise noted, strains were grown in rich medium (YPD, 1% yeast extract, 2% peptone, and 2% glucose), raffinose medium (YPRaff, 1% yeast extract, 2% peptone, and 2% raffinose), or galactose medium (YPGal, 1% yeast extract, 2% peptone, and 2% galactose), using standard media conditions and methods for yeast genetic manipulation [60]. Approximately 500 µg of *HO* integrating cassettes were excised from pMX-Bxb1-Recipient and pMX-Bxb1-Donor plasmids using a NotI (NEB) restriction digest and transformed [115] into haploid FY4 and FY5 yeast strains, respectively. Haploid MAT**a** strains carrying the recipient cassette were selected on YPD containing 400 µg/mL of G418 (Thermo Scientific). Haploid MATα strains carrying the donor cassette were selected on YPD containing 500 µg/mL of Hygromycin B (Thermo Scientific). Confirmation of cassette integration at *HO* was performed using PCR. Diploid strains were constructed by mating the haploid strains and selecting isolates on YPD containing both antibiotics.

Estimating recombination efficiency

To compare the performance of our system to other methods with similar design that utilize Cre-lox, we estimated the recombination efficiency as described previously ([94,98]). This metric is derived from a linear model fit to the number of numbers of recombined diploid recovered post induction and chemical selection, as their system reconstituted an auxotrophic marker. As

recombined versus unrecombined diploids in our Bxb1-att system are not under chemical selection, we utilized a Cannon EOS 5 Mark II camera outfitted with a 520/28-nm band pass filter 488-nm laser to identify yEGFP fluorescent colonies. Three unrecombined diploid isolates carrying both the donor and recipient cassettes were randomly picked. Each isolate was grown to saturation in 5 mL YPD at 30°C. Cells were washed twice with 1 mL of YPRaff, and the pellets were briefly examined with our camera to verify non-fluorescent status. ~1 $x10^8$ cells were resuspended in 5 mL of YPGal to induce recombination and were grown for 24 hours at 30°C. Cells were counted pre and post induction with a hemocytometer to estimate the increase in cells during this period (~2.4 fold growth). Diploid cells were then plated on rich medium at densities of (100, 200, 400 cells/plate) and grown for 48 hours at 30°C. Endpoint plate images were taken with our GFP camera in the dark. Fluorescent colonies were counted from the images, divided by the fold growth, and fit to linear model to estimate the frequency of recombination events (slope).

Flow cytometry

Diploid yeast strains carrying recipient and donor cassettes integrated at the barcode locus (*HO*) were grown, in triplicate, to saturation in YPD medium 30°C. ~1.5 $x10^6$ of cells were back diluted into 5 mL of YPRaff and grown to log phase (OD 0.4-0.6) before induction with galactose (2% final concentration) or without galactose (negative control). A diploid yeast strain carrying a single copy of the yeGFP-AI expression cassette at *HO* was included as a positive control. Cultures were sampled every 24 hours for three days by flow cytometry to quantify the relative proportion of fluorescent cell populations. For cultures growing longer than 24 hours, 1 mL (~1 $x10^8$ cells) were passaged an additional time into 5 mL of YPGal. All flow cytometry experiments were performed on a LE-SH800S Cell Sorter (Sony) using a 100 μm microfluidic sorting chip. Cells

were washed twice with PBS and resuspended to a density of ~1x10^7 cells/mL. yEGFP fluorescence was excited with a 488-nm laser and detected with a 525/50-nm band pass filter. FSC-H and FSC-A gates for identifying singlets were applied for all events. yEGFP fluorescent populations were gated based on the fluorescence intensities of the positive and negative controls. Sub-sampling of fluorescent and non-fluorescent gated events in induced cell cultures were sorted on the ultrapure setting of the instrument directly onto solid medium. The proportion of fluorescent cell populations for at least 10,000 events for each sample was quantified using FloJo software (version 10.5.2; BD biosciences).

3.3 Results

3.3.1 Design of the Bxb1 recombination scheme

We designed a Bxb1-based recombination system that would generate double barcodes at a neutral locus in diploid yeast genome under selectable conditions. To accomplish this, we incorporated three features in our design: 1) an inducible mechanism to control expression of BxB1 and minimize potential off-target effects; 2) a recombination scheme that identifies recombined cells by fluorescence; and 3) places short DNA barcodes adjacent to one another. These features were split into orthogonal donor and recipient cassettes that can be integrated into MAT**a** and MATα genomes, and only recombine in the diploid state after Bxb1 expression is induced.

The donor construct contains the Bxb1 attB site flanked by downstream components in the following order: a barcode landing pad (NheI restriction site), 3' yEGFP-AI fragment, HygMX cassette, and a galactose inducible Bxb1 expression cassette. The recipient construct contains the Bxb1 attP site flanked by upstream components in the following order, a barcode landing pad (AflII restriction site), 3' yEGFP-AI fragment, and KanMX cassette. Both integrating constructs

have ~300bp of upstream and downstream homology to the *HO* locus on each end. By placing the attB and attP recognition sites in parallel orientation at the barcode locus, the resulting products recombine to form attL and attR sites in a crossover like manner. Thus, the desired site-specific recombination reconstitutes the functional yeGFP-AI expression cassette that contains the double barcode within an artificial intron as single copy in the diploid yeast genome (**Figure 3-1**).

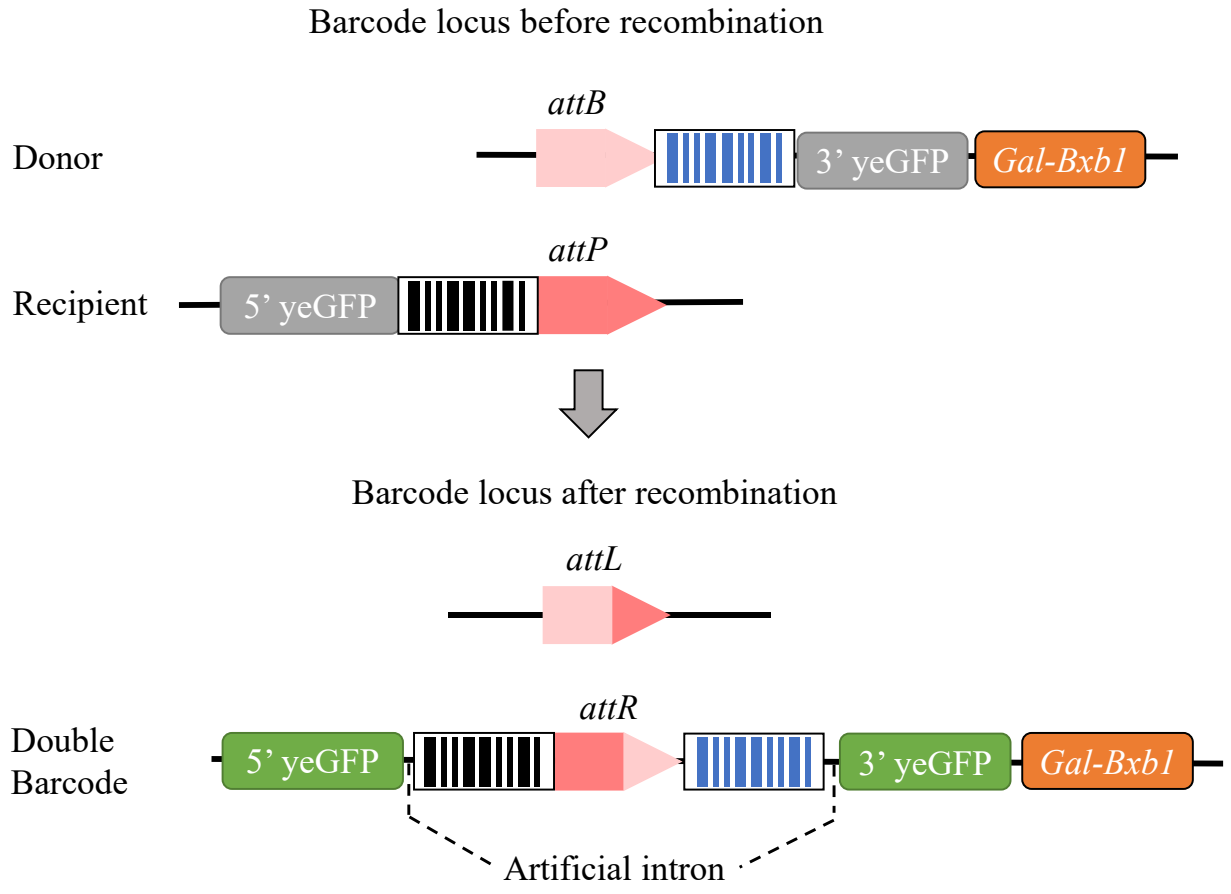

Figure 3-1. Schematic of Bxb1-mediated site-specific barcode recombination.

To determine whether we could generate complex plasmid barcode libraries, we inserted short DNA barcodes into pMX-Bxb1-Donor and pMX-Bxb1-Recipient plasmids. Plasmids were linearized at their unique barcode landing pad site and Gibson assembled [107] with a pool of single stranded barcoded DNA oligonucleotides (100 bp) (**Figure 3-2.A**). Each oligonucleotide is

comprised of barcode with 20 random nucleotide positions and homology to the barcode landing pad at each end. Stretches of random bases were limited to 2-5 base pairs to prevent the regeneration of undesired restriction sites. We expect that the frequency of transformants carrying the same barcode to be rare given the complexity of potential barcodes ($\sim 4^{20} = \sim 10^{12}$). Under this assumption, the number of transformants represents the number of unique barcoded plasmids. We estimated a transformation efficiency of $\sim 2 \times 10^{8}$ CFU/µg for each plasmid. Sanger sequencing a small subset (n=20) barcoded plasmids indicated a unique barcode in each transformant with no errors introduced at overlapping assembly junctions (**Figure 3-2.B**).

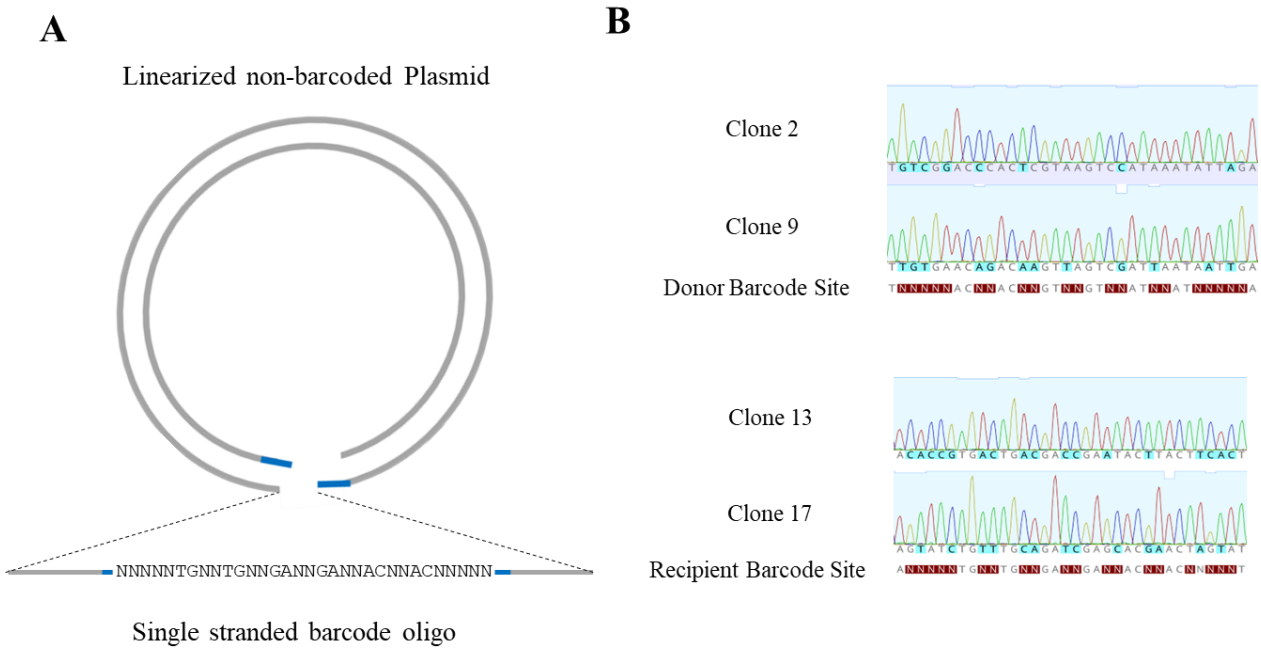

Figure 3-2. Strategy for barcoding plasmid libraries. **A** Recipient or donor plasmids are linearized at their unique barcode landing pad restriction site and Gibson assembled with a barcode oligonucleotide. **B** Sanger sequencing traces for plasmid barcodes isolated from 4 randomly selected transformant clones.

3.3.2 Generation of recombined diploids

To test whether our design yielded yEGFP fluorescent (yEGFP+) cells that had the desired recombination product, we induced expression of Bxb1 in diploid strains harboring the unbarcoded

recipient and donor constructs at the *HO* locus. First, the recipient integrating construct was transformed in haploid MAT**a** (FY4) strains and selected on YPD+Kan. The donor integrating construct was then transformed in haploid MATα (FY5) strains and selected on YPD+Hyg. Haploid strains were then mated and diploids were selected on medium containing both antibiotics. Consistent with the expectation that splitting the yEGFP-AI across two chromosomes would yield non-functional proteins, the resulting haploid and diploid strains were non-fluorescent (yEGFP-) (**Figure 3-3.A**). A positive control diploid strain carrying a single copy of yeGFP-AI at *HO* indicated that the artificial yeast intron could be successfully spliced, and functional yEGFP translated (**Figure 3-3.A**).

Next, we induced diploids and non-selectively sampled cells after 24 hours of growth to see if our method could generate the desired site-specific recombination. Our results show a large proportion of cells became fluorescent (**Figure 3-3.B**) after galactose induction. We sought to sequence the att sites in a small subset of yEGFP+ and yEGFP- diploids (n=20) to characterize the recombination sites. Inspection of the att site in yEGFP+ diploids indicated that the attP recognition site was correctly recombined to form the product attR site flanked by intact barcode landing pads (**Figure 3-3.C**). yEGFP- diploids revealed that their attP and attB recognition sites were also intact. These results suggest that Bxb1 could successfully introduce and recombine dsDNA breaks in high-fidelity. It is possible that unintended rearrangements can be caused by the endogenous homologous DNA repair system in *S. cerevisiae*. Based on our initial survey, we expect these events to occur at a relatively rare frequency.

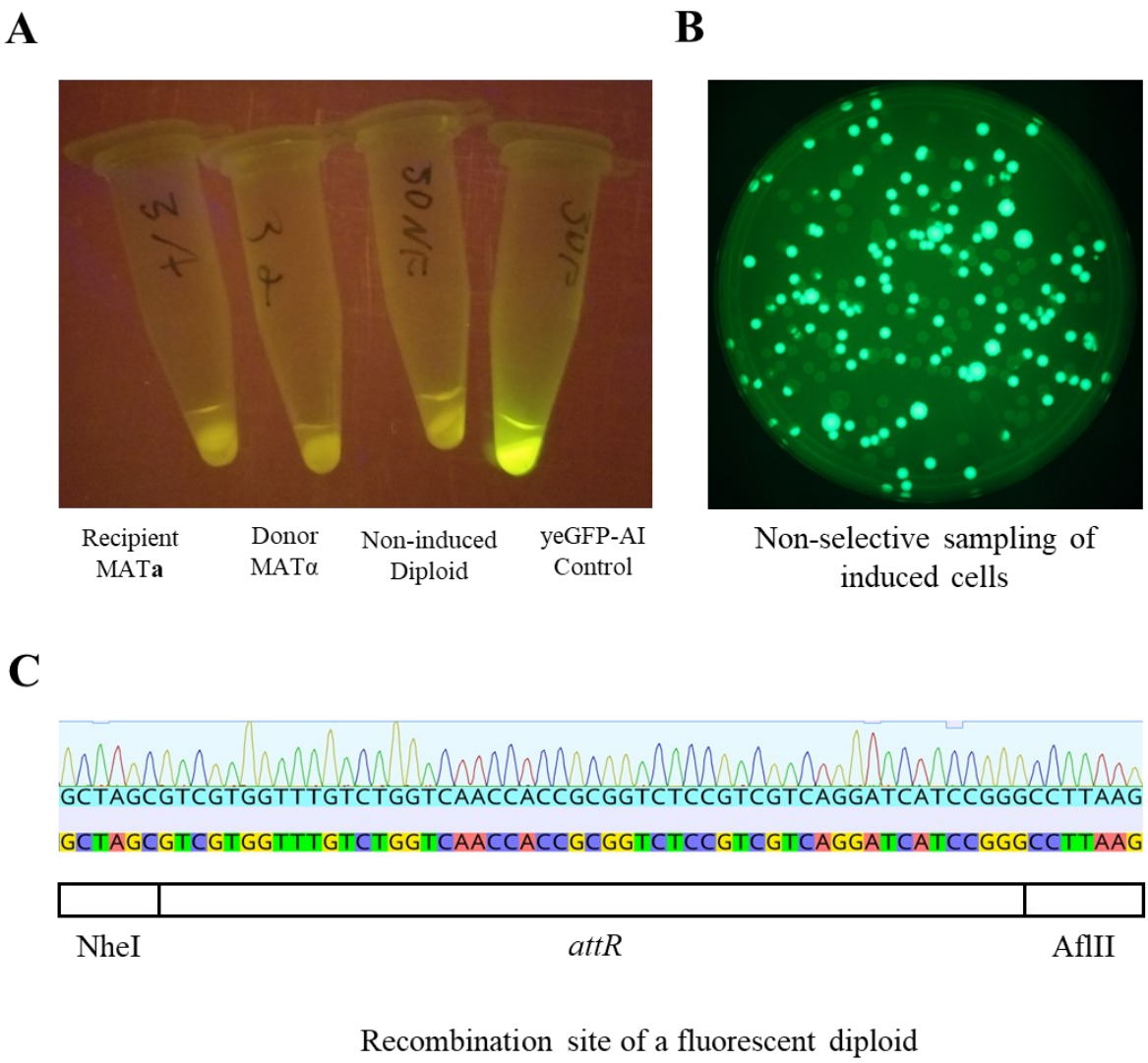

Figure 3-3. Bxb1 site-specific recombination in yeast diploids. **A** Fluorescence of yeast cell pellets in Eppendorf tubes. **B** Fluorescence of colonies sampled from induced cells and plated on YPD. **C** Sanger sequencing trace for the recombined attR site in a yeGFP+ diploid post galactose induction.

3.3.3 Estimating rates of recombination

To compare our results to a similar design that utilized Cre-lox and reconstitution of a split auxotrophic marker (*URA3*-AI), we estimated rates of recombination as previously described [94]. Briefly, the lower bounds of recombination efficiency represent the slope of a linear model fit to

the number of recombined diploids recovered from plating various cell densities, adjusted for fold growth during induction. We estimate that the lower bounds of recombination efficiency for our system to be ~21%, approximately 10-fold higher than previous approaches that used Cre-lox [94,98].

One difference in our approach is that recombined diploids are identified by yEGFP-AI fluorescence and are not under chemical selection. The fluorescence of each recombined cell is a function of the total number of functional yEGFP-AIs. In determining the lower bounds of recombination efficiency, fluorescent colonies that grew from a single recombined diploid cell (**Figure 3-3.B**) represent the accumulation of many copies of yEGFP-AIs that can be distinguished from the background autofluorescence of yeast cells. To be able to stratify induced diploids into libraries of recombined diploids by FACS, we evaluated the fluorescence of single cells by flow cytometry. We expect that the galactose promoter to be fully induced shortly after the addition of galactose to strongly express Bxb1. However, the time scale and factors that compete or inhibit Bxb1-mediated recombination between att sites in the context of homologous yeast chromosomes is not understood. Thus, the rate at which recombined sites are formed impact when cells start expressing functional yEGFP and if they can be selected in downstream FACS.

We characterized diploid populations at three timepoints of varying growth periods in the presence of galactose to characterize the proportion of yeGFP+ cells. We observed a clear difference in the fluorescence distributions between control strains (**Figure 3-4**) and used these differences to set thresholds for populations in induced diploid cultures. Cells that had induction growth periods longer than 24 hours displayed higher proportions of yeGFP+ populations (~60%). Sampling yEGFP+ populations at all time points (96 events/plate) indicated that approximately all cells were viable and fluorescent. However, sampling yEGFP- gated populations at the 24-hour

time point showed that ~40% of sorted events (n=768) became fluorescent after 2 days of growth on solid media containing glucose. This behavior was also observed in yEGFP- gated populations at the 48-hour timepoint but were overall less frequent (~9%; n=768). We expect that it is unlikely that sampled cells are recombining after sorting onto plates with media conditions that repress Bxb1 expression. Suggesting that the threshold for gated yEGFP- populations may be too broad currently, or that considerable lag in the rate of recombination exists in diploid populations of the same initial genotype.

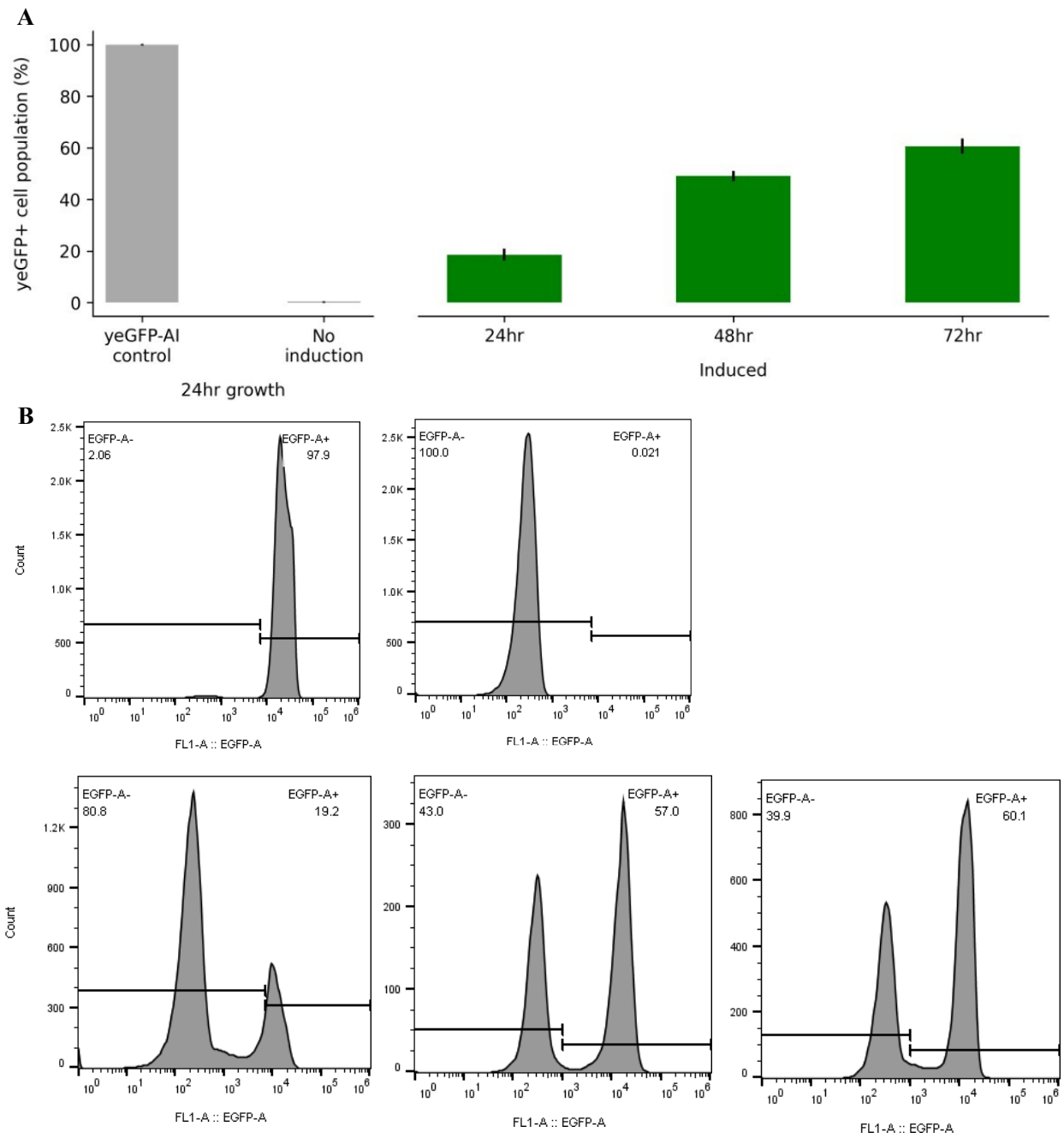

Figure 3-4. Determining the proportion of fluorescent diploid populations by flow cytometry. **A** Proportion of yEGFP fluorescent diploid yeast cells at several timepoints of varying growth periods in induction conditions. Bar chart of the average number of fluorescent singlets over 10,000 events, over 3 replicates. Error bars represent the standard deviation. Grey bars represent control strains that were grown for 24 hours. **B** Sample histogram of the distribution of yEGFP intensities for one replicate shown.

3.4 Discussion

Here we present preliminary work that demonstrates how our Bxb1-mediated design can generate recombined diploid strains that can be identified by their fluorescence. Resulting double barcodes are maintained as a single copy in the diploid yeast genome within an artificial yeast intron, that can be amplified as a short DNA fragment (<300 bp) and be sequenced by existing methods (e.g Illumina, Oxford Nanopore). We show that high rates of Bxb1-mediated recombination are possible in our initial tests of recombining unbarcoded diploid cells carrying recipient and donor constructs. The lower bounds of recombination efficiency of our design (~21%) compared to a similar platform [94], which utilized Cre-lox and reconstitution of auxotrophic marker, was ~ 10-fold higher. However, a direct comparison requires generating additional constructs in our scheme that swap the recombinase ORF and recognitions sites to Cre-lox. Measuring and sampling yEGFP+ and yEGFP- populations of the same initial genotype in our design also suggest that recombined populations can reach ~60%. Compared to a previous design that attempted to use homology directed repair (**Appendix A**), unidirectional site-specific recombination had much higher rates of desired product formation. Rates of recombination may be more accurately estimated by sequencing recombined double barcoded diploid strains to determine the distribution of frequency, complexity, and effects of growth during library construction.

Further improvements can be made to our design by implementing a yeast-codon optimized version of Bxb1 and the use of att recognition site variants. A recent study demonstrated that an attB and attP variants (Bxb1-GA) had improved rates of site-specific integration in mammalian cells [116]. Our strategy is also amendable to plasmids instead of chromosomal DNA, although they are maintained in variable copy numbers per cell. Future design considerations may include

moving the Gal-Bxb1 expression cassette from the integrating donor cassette to a plasmid that can be cured post recombination to further reduce risk of undesired DNA damage. Alternative strategies include utilizing an additional marker that can be counter selected (example in [93]) or coupling an inducible promoter to disrupt plasmid segregation (example in [117]). Furthermore, this design could be expanded to use split drug resistance markers for chemical selection instead of fluorescence (example in [98]) to reduce time and costs associated with FACS.

In summary, we present an alternative method that displays high rates of recombination for double barcoding yeast diploids. Improvements in recombination efficiency can potentially result in more comprehensive construction of combinatorial libraries and less variation in the starting frequency distribution.

Chapter 4: Conclusions and outlook

The work presented in this dissertation outlines two high-throughput methods for generating functional datasets that can be used to inform genotype-phenotype relationships in biology and human health. As DNA sequencing becomes more cost-effective and widely applied, the rate at which human genetic variants are discovered will continue to skyrocket. Modeling, measuring, and understanding the impact of genetic variants remain a central challenge in precision medicine. High-throughput functional assays are an important approach in overcoming this challenge.

In Chapter 2, we demonstrate how large-scale functional assays in model systems can be powerfully applied in the study of rare disease and to inform future diagnostic efforts. The strong agreement between experimentally measured results from our assay and known features of the protein structure together with (albeit limited) clinical data support its use as a well-validated assay of human PSAT function. The corresponding dataset provides meaningful functional information for a large proportion (88%) of all unique SNV-accessible missense substitutions across the length of the protein. Furthermore, our computational model extends the use of the dataset to predict the functional impact of ~1.8 million allele combinations corresponding to potential human genotypes. As such, this approach leverages a relatively small amount of clinical data to classify large numbers of variants making it especially powerful for rare diseases. In Chapter 3, we explore an alternative site-specific recombinase to improve methods for barcoding combinatorial libraries in yeast diploids, which could be used to experimentally measure the functional effect of pairwise allele combinations at scale. In this chapter, I outline future areas of study and the broader context of this work in the field of clinical genomics.

4.1.1 Future work in PSAT protein science and potential applications

Mutational scanning experiments can inform our current understanding of protein science and outline areas of study that can be further explored. For enzymes like human PSAT, aminotransferases, and PLP-dependent enzymes, the mechanistic role of the cofactor's pyridoxal ring group in catalysis has been extensively characterized [58,73,118]. Our yeast assay was able to recapitulate this functional importance, with all pyridoxal ring interacting residues being mutationally intolerant. In contrast, the role of PLP's phosphate group in catalysis is overall less understood for these enzymes and may act as an anchor point to the protein [74] or directly interacts with substrates [73]. Our assay identified some substitutions that were tolerated at PLP phosphate group binding residues, and points to the need for further crystallographic studies of human PSAT in substrate bound states to determine if cofactor-substrate interactions exist.

Along these lines, a deeper understanding of the factors that govern substrate specificity can further inform future areas of study such as the development of therapeutics or industrial strain engineering. For example, a recent study demonstrated that introducing two missense mutations at two substrate binding sites in *E. coli* phosphoserine aminotransferase could switch substrate specificity to the non-canonical L-homoserine and drive production 1,3-propanediol [119]. Regarding therapeutics, targeting human PSAT is of particular interest as it is overexpressed and or upregulated in many types of cancers such as breast, colorectal, and lung cancers [30,120,121]. Several preclinical tumor models have shown improved anti-cancer effects when the serine biosynthesis pathway is disrupted genetically (e.g gene knockouts) [122]. However, few PSAT specific inhibitors have been described or developed [123]. One example is aminooxyacetate, although it functions as a general inhibitor of aminotransferases and PLP-dependent enzymes [124]. Thus, large-scale functional data in combination with more detailed mechanisms of substrate binding presents an opportunity to aid the development of classic small molecule

inhibitors or even molecular glues (targeted protein degradation) [125] for future PSAT cancer therapeutics.

The overall importance of the cofactor in PLP-dependent enzymes also extends beyond catalysis and may impact protein folding and stability [126]. However, the mechanistic role of PLP as a chaperone for most enzymes, including human PSAT, has not been explored in detail. Interestingly, a study showed that serine synthesis is impaired in neuronal cells grown in vitamin b6 (precursor of PLP) deficient conditions [127]. One advantage of our yeast assay is the control of extracellular environment and conditions of low and high vitamin b6 can be experimentally evaluated (like in [128]). Future studies may also include implementation of a reporter for abundance in our assay, such as PSAT fusions to a fluorescent protein or enzyme reporter like luciferase, to further characterize the effects of PLP on stability. While serine supplementation is the recommended treatment for serine biosynthesis defects and has seen remarkable clinical effects (examples in [33,37]), there may be a potential therapeutic benefit in combinatorial supplementation with vitamin b6 as with other disorders of vitamin b6 metabolism and deficiencies of PLP-dependent enzymes [126,129].

4.1.2 Potential of yeast-based functional assays for related enzymes

Our yeast-based complementation assay provides a strong example of how orthologs encoding proteins with conserved enzymatic function can be studied in a heterologous system. Suggesting the suitability of *S. cerevisiae* to detect loss of function variants in an active site architecture that is shared within each sub-family of aminotransferases. Tens of highly related (by multiple amino acid sequence alignment) yeast and human orthologs have been identified for other aminotransferases [58], and may be good candidates for establishing future complementation assays. For example, pathogenic variants in human *TAT* (encoding tyrosine aminotransferase)

result in a deficiency to metabolize tyrosine and cause a set of rare IEMs termed tyrosinemia type II [130]. Its ortholog in *S. cerevisiae* (*ARO8*) is also unable to grow on minimal medium supplemented with tyrosine [131]. Thus, searching within the family of aminotransferases may be a strong starting point for functionally probing other metabolic enzymes that potentially can have their selectable phenotypes linked to variant growth in yeast.

Beyond aminotransferases, *S. cerevisiae* may also be a strong model organism to functionally assay other PLP-dependent enzymes. In another large-scale study [128] that applied a yeast-based complementation assay for human cystathionine beta-synthase (CBS), the results from their screen also showed that positions in proximity to PLP were intolerant to variation. That study [128] also identified over 40 human genes that encode PLP-dependent enzymes, which could be utilized as another starting point for exploring future yeast complementation assays.

4.1.3 Future pooled screens of PSAT function in yeast

Our high-throughput approach in Chapter 2 leveraged synthesized DNA libraries, arrayed strain libraries, nanopore sequencing, and automated image analysis to independently measure growth for thousands of variants with multiple biological and technical replicates. Ongoing work in the Dudley laboratory has also successfully applied this methodology to assay other genes (e.g *OTC, ASL, ASS1*) associated with urea cycle disorders. After establishing a yeast complementation assay, a single researcher can construct, sequence, and test a variant library in 4-7 weeks. Towards comprehensively assaying the effect of pairwise variant combinations in yeast diploids, this strategy is not time or cost effective and would likely require batched phenotyping.

Adaptation of this approach to pooled format would facilitate the ability to perform large-scale screens of variant combinations in a single experiment. One strategy that to accomplish this would be to barcode variant strains and determine the frequency of reads before and after selection

to estimate relative fitness in a pooled competitive growth assay. In Chapter 3 I presented a generalizable method that could be used to barcode existing haploid libraries and generate double barcoded diploid strains in the resulting bulk cross.

One important consideration in developing pooled growth assays for metabolic enzymes like PSAT, is determining if strains release the autotrophically required nutrient to other cells in the population [132,133]. The potential introduction of cross-feeding between higher and lower fitness strains could obscure variant effect estimations. To test for this phenomenon, a subset of PSAT variant haploid strains covering a broad range of effect (as determined in Chapter 2) could be barcoded and used to compare individual growth rates to their relative barcode frequency in the pooled competitive growth assay at multiple timepoints. In the case that cross-feeding does exist, future strain engineering may include introducing a conditional knockout of yeast genes encoding transporters that uptake exogenous serine (Gnp1 and Agp1) [134]. Furthermore, experiments that vary the initial distribution of fitness can also inform statistical methods to model the impact of mean population fitness in fold enrichment analysis.

4.1.4 Outlook for serine biosynthesis defects

Attempts to establish yeast complementation assays for the other two serine biosynthesis pathway genes (*PHGDH* and *PSPH*) in a similar method to *PSAT1* were not successfully on initial attempts. Although there are promising steps for future assay engineering that may achieve this goal (**Appendix B**). Humanizing an entire metabolic pathway in yeast has the potential to assay polygenic combinations of single variants, as well to study flux and regulation in a genetically tractable organism.

Unfortunately, like many other rare diseases, there is no dedicated repository for the aggregation and comparison of genotypic, phenotypic, functional, or therapeutic data for serine

biosynthesis defects. Often reports of novel variants rely on literature review and manual curation (examples in [49–51]) to compare biomedical traits. Regarding case reports of individuals with genetic characterization, there are 45 for *PHGDH* (reviewed in [135]), 31 for *PSAT1* (reviewed in [49], and recent reports in [50,51,82–84]), and several for *PSPH* [42,136,137]. This approach is time consuming and case reports can be missed. For a rare disease that affects a small number of people, each case report holds substantial value. Efforts from programs like the International Rare Diseases Research Consortium (IRDiRC) will hopefully improve data sharing in the future.

The functional datasets generated in Chapter 2 can serve as a valuable resource in study and diagnosis of *PSAT1*-related serine biosynthesis defects (NLS2 and PSATD). A simple thresholding approach based on haploid growth scores corresponding to available clinical annotations allowed us to classify thousands of variants into deleterious, uncertain, and non-deleterious categories. As more clinical interpretations are established, calculating formal measures of classification for our assay, such as odds of pathogenicity [81,138], will also become possible. Given that the therapeutic potential for these medically actionable disorders is highest before patients become symptomatic, clinical sequencing coupled to informative variant interpretation can be a powerful diagnostic tool.

As clinical sequencing continues to be more widely applied, there will be more opportunities to diagnose all forms serine biosynthesis defects if knowledge of pathogenic variants is available. Prenatal exome sequencing in two recent large-scale studies [85,139] demonstrated the ability to increase diagnostic yield in fetuses with detected congenital abnormalities. For severe cases of NLS, a lethal congenital anomaly disease, this approach may allow for diagnosis and potentially treatment during pregnancy. The most recent AMCG guidelines [89] recommended exome or genome sequencing as a first or second tier option for pediatric patients with neonatal

congenital abnormalities, developmental delay, or intellectual disability. Milder forms of serine-biosynthesis defects in infants and children have symptoms that fall into this recommendation and become more likely to be sequenced. There are also three [50,136,140] cases of serine-biosynthesis defects that were diagnosed in adulthood. Notably one patient was enrolled under the Undiagnosed Disease Program [86,87], and the others in tertiary neurological care. Thus, future sequencing efforts in prenatal, pediatric, and specialized care that are targeted towards phenotypes that overlap with serine biosynthesis defects may greatly benefit this rare, but actionable set of IEMs.

4.1.5 Outlook for variant effect combinatorics

While variant interpretation is a bottleneck that can hamper the progress of clinical genomics, diagnosis requires an understanding of the level of gene function conferred by the combination of alleles in the patient's genome. The aggregation of genotype and longitudinal phenotype data in biobanks are allowing researchers to investigate the effect of allele combinations at larger scales. Approaches like phenome wide association studies (PheWAS) and Phenotype Risk Scores (PheRS) are identifying variants in Mendelian disease genes with presumed recessive inheritance patterns that have significant heterozygous disease effects [91,141]. These studies further support the idea that Mendelian and complex diseases may exist on a spectrum, and gaining insight on this can benefit future clinical interpretations.

APPENDIX

Appendix A:

Double barcoding combinatorial libraries of yeast diploids via homology directed repair

Background

Eukaryotes repair potentially lethal genomic DNA damage through homologous recombination (HR), or non-homologous end joining (NHEJ) [142,143]. This biology has been exploited to enhance rates of recombination for integrating cassettes flanked with homology at genomic sites with an induced double stranded DNA (dsDNA) break [144]. The robust nature of endogenous DNA repair to translocate DNA flanked by enough homology, suggests a high-fidelity platform for barcode fusion.

The double barcoding design via homology directed repair is as follows: integrating recipient and donor cassettes at *HO* locus in *S. cerevisiae*, inducing a dsDNA break in the recipient, and engineering tracts of homology adjacent to the break to direct template (donor) mediated repair (**Figure A1**). The homing mega-endonuclease I-OnuI was for chosen for break induction, due to its high specificity, activity, and ability to be expressed transiently by galactose induction [145,146]. In this experimental design, barcode fusion events are selectively marked by the reconstitution of full length GFP that contains an artificial intron [147]. Homologous recombination is restricted to gap-repair by inverting the recipient cassette's orientation (relative to donor), where repair by crossover or break induced replication (BIR) generate lethal dicentric chromosomes. This orientation to selects against break induced replication, an error prone process that can introduce genomic instability [148].

Results and Discussion

Upon induction and growth for 17 hours at 30°C, only a minority of the population became fluorescent (**Figure A1**). Sampling of the GFP negative population revealed that the majority (>99%) had been produced by a gap-repair process that drove the loss of the entire recipient cassette (**Figure A1**). In which homology outside the cassette on chromosome IV is reached through long range re-sectioning, and the non-homologous tail (recipient cassette) is clipped to allow for gap-repair DNA synthesis [149]. Thus, this method is limited by the competition from existing extensive endogenous homology outside of the recipient cassette, which disfavors gap-repair directed by the short amount of flanking homology (50 bp) engineered into the cassette. An attempt to limit long-range re-sectioning through inhibition with caffeine (1μM-10mM), which depletes two key long-range endonucleases (Dna2 and Sae2) [150], yielded no significant shifts in GFP^+ populations.

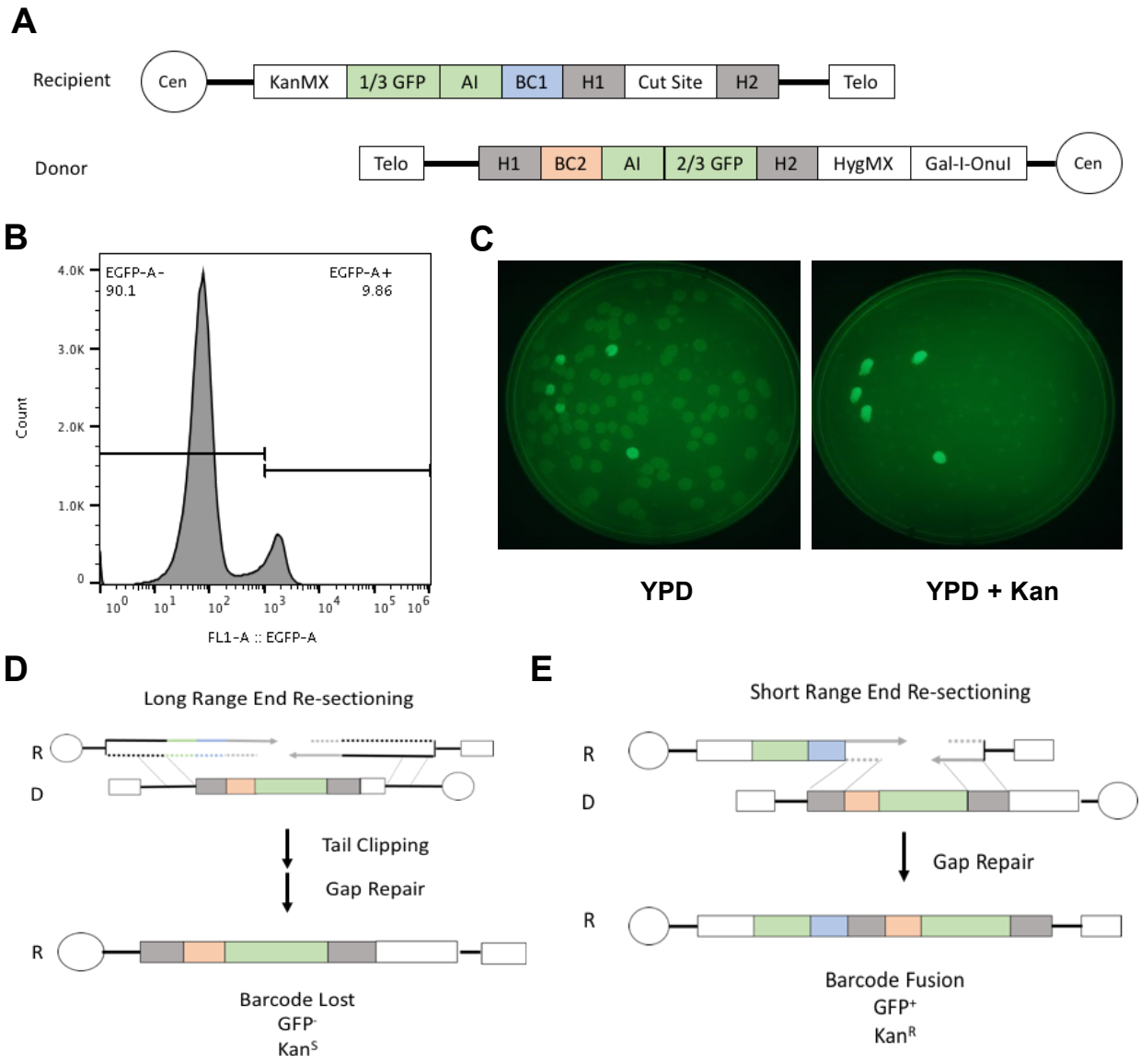

Figure A1. Chromosomal barcode translocation via directed gap repair. **A** Diagram of the recombination scheme at the barcode locus. **B** Proportions of yEGFP fluorescent cells for a single sample after 17 hours of growth in YPGal. **C** Replica plating of ~100 cells, non-selectively sampled from an induced culture, on rich medium (YPD) and rich medium containing G418 (Kan). **D** Preposed mechanism for the large population of yEGFP^{-} Kan^{S} diploids. **E** Preposed mechanism for the small population of yEGFP^{+} Kan^{-} diploids.

Appendix B:

Preliminary work towards establishing yeast complementation assays for *PHGDH* and *PSPH*

Background

Impairment of any of the three L-serine biosynthesis pathway enzymes, phosphoglycerate dehydrogenase (PGDH; encoded by *PHGDH*), phosphoserine aminotransferase (PSAT; encoded by *PSAT1*), and phosphoserine phosphatase (PSP; encoded by *PSPH*), result in serine deficiency disorders and can present with identical clinical phenotypes [29]. To extend characterization to the entire pathway, we applied a similar strategy in our PSAT assay (**Chapter 2 Methods**) to test whether the human coding sequences for *PHGDH* and *PSPH* could functionally replace their yeast orthologs. In *S. cerevisiae*, two genes (*SER3* and *SER33*) encode the ortholog for human *PHGDH*, and a single gene (*SER2*) encodes the ortholog for human *PSPH*.

Methods

Deletion strains (*ser3Δ0 ser33*Δ0, *ser2Δ0*) were generated by replacing orthologous yeast genes with a selectable drug marker at their native locus. Complementation strains were generated by integrating a single copy of the yeast codon optimized version of the human coding sequence (*yPHGDH*, *yPSPH*) at a neutral locus (*HO*) under the control of their respective orthologous yeast promoter and terminator. In the case of *yPHGDH*, the *SER33* promoter and terminator was chosen, as strains harboring single deletions (*ser3Δ0* or *ser33*Δ0) had growth approximately equal to wild type on minimal media lacking serine. To test whether integration at a neutral locus affected wild type function, control strains were constructed with yeast *SER33* or *SER2* under the control of their endogenous promoter and terminator integrated at *HO*. The wild type FY4 strain and previously constructed (**Chapter 2 Methods**) *yPSAT1* and *ser1Δ0* strains were also included for comparison.

Strains were grown to saturation in YPD and pinned onto YPGlycerol to remove petites. Strains were to then transferred to YPD, grown to saturation, and pinned onto four different conditions: minimal medium containing either glucose (1% final concentration), glycerol (120mM), ethanol (180 mM), or acetate (180 mM). Plates were imaged every 24 hours for 4 days and raw phenotypic values were extracted using PyPl8 software.

Results and discussion

All deletion strains (*ser3Δ0 ser33*Δ0, *ser2Δ0*) were unable to grow on minimal medium lacking serine, however the same was observed for *yPHGDH* and *yPSPH* complementation strains (**Figure A2**). Rescue strains indicated that integration at *HO* did not impair function relative to wild type FY4. A recent study [151] demonstrated that human PGDH could only restore growth in yeast strains with media utilizing non-fermentable carbon sources. Screening similar conditions with glycerol, ethanol, and acetate as the carbon source as in [151], however, strains harboring *yPHGDH* were still unable to grow (**Figure A2**). One difference in design is that the human *PGHDH* coding sequence was under the control of a strong constitutive yeast promoter (GPD) [151]. Therefore, expression levels higher than the promoter sequences for *SER3* or *SER33* may be required for human *PHGDH* to restore function in yeast. Another consideration for why *yPSPH* failed to complement, is the levels of calcium in the standard composition of minimal medium may have inhibited human PSP function [152]. Thus, there are several assay engineering steps that could be explored in the future to establish functional assays for the remaining genes in the human serine biosynthesis pathway.

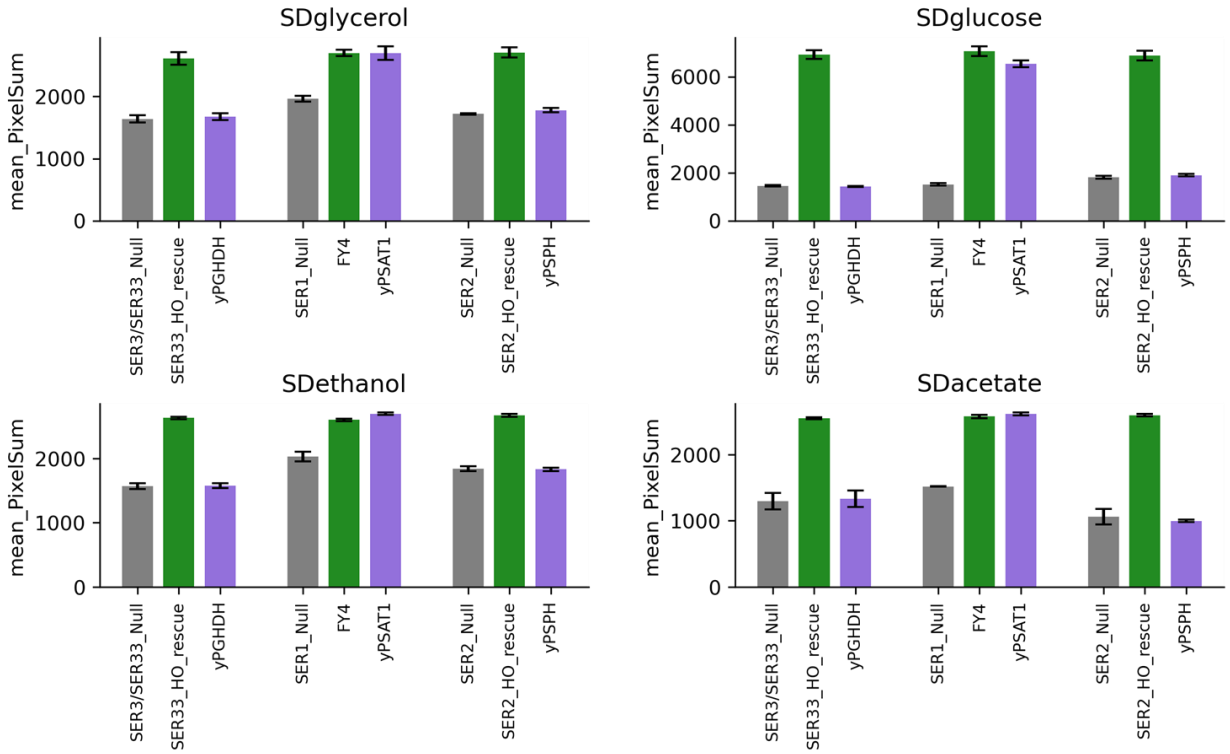

Figure A2. Yeast complementation assays for human serine biosynthesis pathway genes on different media conditions (listed above). Bar chart of the mean pixelsum (sum of the gray scale pixel intensities within each strain patch) after four days of growth for three replicates. Error bars represent the standard deviation.

BIBLIOGRAPHY

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

2. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed [October 29th, 2019].

3. 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. Nature Research; 2020;581:434–43.

4. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat [Internet]. 2021; Available from: https://onlinelibrary.wiley.com/doi/10.1002/humu.24309

5. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577:179–89.

6. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508:469–76.

7. 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. Genetics in Medicine. Nature Publishing Group; 2015;17:405–24.

8. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–5.

9. Landrum MJ, Kattman BL. ClinVar at five years: Delivering on the promise. Hum Mutat. 2018;39:1623–30.

10. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020;48:D835–44.

11. Zou J, Valiant G, Valiant P, Karczewski K, Chan SO, Samocha K, et al. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. Nat Commun. 2016;7:13293.

12. Liu Y, Yeung WSB, Chiu PCN, Cao D. Computational approaches for predicting variant impact: An overview from resources, principles to applications. Front Genet. Frontiers Media S.A.; 2022.

13. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. Nature Publishing Group; 2017;35:128–35.
14. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. Nat Protoc. Nature Publishing Group; 2016. p. 1782–7.

15. Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. Hum Genet. Springer Verlag; 2018. p. 665–78.

16. Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, et al. Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet. Cell Press; 2017. p. 315–25.

17. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11:801–7.

18. 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;50:874–82.

19. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. The American Journal of Human Genetics. 2021;108:1735–51.

20. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 2016;26:670–80.

21. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science (1979). 2015;348:921–5.

22. Illsinger S, Das AM. Impact of selected inborn errors of metabolism on prenatal and neonatal development. IUBMB Life. 2010;62:403–13.

23. Ferreira CR, Rahman S, Keller M, Zschocke J, Abdenur J, Ali H, et al. An international classification of inherited metabolic disorders (ICIMD). J Inherit Metab Dis. John Wiley and Sons Inc; 2021;44:164–77.

24. Sanderson S, Green A, Preece MA, Burton H. The incidence of inherited metabolic disorders in the West Midlands, UK. Arch Dis Child. 2006;91:896–9.

25. Dionisi-Vici C, Rizzo C, Burlina AB, Caruso U, Sabetta G, Uziel G, et al. Inborn errors of metabolism in the Italian pediatric population: a national retrospective survey. J Pediatr. 2002;140:321–7.

26. Gambello MJ, Li H. Current strategies for the treatment of inborn errors of metabolism. Journal of Genetics and Genomics. Institute of Genetics and Developmental Biology; 2018. p. 61–70.

27. Morris AAM, Kožich V, Santra S, Andria G, Ben-Omran TIM, Chakrapani AB, et al. Guidelines for the diagnosis and management of cystathionine beta-synthase deficiency. J Inherit Metab Dis. Springer Netherlands; 2017. p. 49–74.

28. Stockler S, Plecko B, Gospe SM, Coulter-Mackie M, Connolly M, van Karnebeek C, et al. Pyridoxine dependent epilepsy and antiquitin deficiency. Mol Genet Metab. 2011;104:48–60.

29. El-Hattab AW. Serine biosynthesis and transport defects. Mol Genet Metab. Academic Press Inc.; 2016. p. 153–9.

30. Pan S, Fan M, Liu Z, Li X, Wang H. Serine, glycine and one-carbon metabolism in cancer (Review). Int J Oncol. NLM (Medline); 2021. p. 158–70.

31. Pernot P, Maucler C, Tholance Y, Vasylieva N, Debilly G, Pollegioni L, et al. D-serine diffusion through the blood-brain barrier: Effect on D-serine compartmentalization and storage. Neurochem Int. 2012;60:837–45.

32. de Koning TJ. Amino acid synthesis deficiencies. J Inherit Metab Dis. Springer Netherlands; 2017. p. 609–20.

33. de Koning TJ, Klomp LWJ, Oppen ACCV, Beemer PFA, Dorland L, Berg IETV den, et al. Prenatal and early postnatal treatment in 3-phosphoglycerate-dehydrogenase deficiency. Lancet. Elsevier B.V.; 2004;364:2221–2.

34. Jaeken J, Detheux M, van Maldergem L, Foulon M, Carchon H, van Schaftingen E, et al. 3-Phosphoglycerate dehydrogenase deficiency: an inborn error of serine biosynthesis. Archives ofDisease in Childhood. 1996.

35. Jaeken J, Detheux M, Fryns JP, Collet JF, Alliet P, van Schaftingen E. Phosphoserine phosphatase deficiency in a patient with Williams syndrome. J Med Genet. BMJ Publishing Group; 1997;34:594–6.

36. de Koning TJ, Duran M, van Maldergem L, Pineda M, Dorland L, Gooskens R, et al. Congenital microcephaly and seizures due to 3-phosphoglycerate dehydrogenase de¢ciency: Outcome of treatment with amino acids. J. Inherit. Metab. Dis. 2002.

37. Hart CE, Race V, Achouri Y, Wiame E, Sharrard M, Olpin SE, et al. Phosphoserine aminotransferase deficiency: A novel disorder of the serine biosynthesis pathway. Am J Hum Genet. University of Chicago Press; 2007;80:931–7.

38. Neu RL, Kajii T, Gardner LI, Nagyfy SF. A lethal syndrome of microcephaly with multiple congenital anomalies in three siblings. Pediatrics. 1971;47:610–2.

39. Laxova R, Ohara PT, Timothy JA. A further example of a lethal autosomal recessive condition in sibs. J Ment Defic Res. 1972;16:139–43.

40. King JAC, Gardner V, Chen H, Blackburn W. Neu-laxova syndrome: Pathological evaluation of a fetus and review of the literature. Fetal Pediatr Pathol. Informa Healthcare; 1995;15:57–79.

41. Klomp LW, de Koning TJ, Malingré HE, van Beurden EA, Brink M, Opdam FL, et al. Molecular characterization of 3-phosphoglycerate dehydrogenase deficiency--a neurometabolic disorder associated with reduced L-serine biosynthesis. Am J Hum Genet. 2000;67:1389–99.

42. Veiga-da-Cunha M, Collet J-F, Prieur B, Jaeken J, Peeraer Y, Rabbijns A, et al. Mutations responsible for 3-phosphoserine phosphatase deficiency. Eur J Hum Genet. 2004;12:163–6.

43. Manning MA, Cunniff CM, Colby CE, El-Sayed YY, Hoyme HE. Neu-Laxova Syndrome: Detailed Prenatal Diagnostic and Post-Mortem Findings and Literature Review. Am J Med Genet. Wiley-Liss Inc.; 2004;125 A:240–9.

44. Coto-Puckett WL, Gilbert-Barness E, Steelman CK, Stuart T, Robinson HB, Shehata BM. A spectrum of phenotypical expression of Neu-laxova syndrome: Three case reports and a review of the literature. Fetal Pediatr Pathol. 2010;29:108–19.

45. Shaheen R, Rahbeeni Z, Alhashem A, Faqeih E, Zhao Q, Xiong Y, et al. Neu-laxova syndrome, an inborn error of serine metabolism, is caused by mutations in PHGDH. Am J Hum Genet. Cell Press; 2014;94:898–904.

46. Acuna-Hidalgo R, Schanze D, Kariminejad A, Nordgren A, Kariminejad MH, Conner P, et al. Neu-laxova syndrome is a heterogeneous metabolic disorder caused by defects in enzymes of the l-serine biosynthesis pathway. Am J Hum Genet. Cell Press; 2014;95:285–93.

47. Brassier A, Valayannopoulos V, Bahi-Buisson N, Wiame E, Hubert L, Boddaert N, et al. Two new cases of serine deficiency disorders treated with l-serine. European Journal of Paediatric Neurology. W.B. Saunders Ltd; 2016;20:53–60.

48. Glinton KE, Benke PJ, Lines MA, Geraghty MT, Chakraborty P, Al-Dirbashi OY, et al. Disturbed phospholipid metabolism in serine biosynthesis defects revealed by metabolomic profiling. Mol Genet Metab. Academic Press Inc.; 2018;123:309–16.

49. Abdelfattah F, Kariminejad A, Kahlert AK, Morrison PJ, Gumus E, Mathews KD, et al. Expanding the genotypic and phenotypic spectrum of severe serine biosynthesis disorders. Hum Mutat. John Wiley and Sons Inc.; 2020;41:1615–28.

50. Debs S, Ferreira CR, Groden C, Kim HJ, King KA, King MC, et al. Adult diagnosis of congenital serine biosynthesis defect: A treatable cause of progressive neuropathy. Am J Med Genet A. John Wiley and Sons Inc; 2021;185:2102–7.

51. Shen Y, Peng Y, Huang P, Zheng Y, Li S, Jiang K, et al. Juvenile-onset PSAT1-related neuropathy: A milder phenotype of serine deficiency disorder. Front Genet. Frontiers Media SA; 2022;13.

52. Moat S, Carling R, Nix A, Henderson M, Briddon A, Prunty H, et al. Multicentre age-related reference intervals for cerebrospinal fluid serine concentrations: Implications for the diagnosis and follow-up of serine biosynthesis disorders. Mol Genet Metab. 2010;101:149–52.

53. Benke PJ, Hidalgo RJ, Braffman BH, Jans J, Gassen KLI van, Sunbul R, et al. Infantile Serine Biosynthesis Defect Due to Phosphoglycerate Dehydrogenase Deficiency: Variability in Phenotype and Treatment Response, Novel Mutations, and Diagnostic Challenges. J Child Neurol. 2017;32:543–9.

54. El-Hattab AW, Shaheen R, Hertecant J, Galadari HI, Albaqawi BS, Nabil A, et al. On the phenotypic spectrum of serine biosynthesis defects. J Inherit Metab Dis. Springer Netherlands; 2016;39:373–81.

55. Sirr A, Lo RS, Cromie GA, Scott AC, Ashmead J, Heyesus M, et al. A yeast-based complementation assay elucidates the functional impact of 200 missense variants in human PSAT1. J Inherit Metab Dis. John Wiley and Sons Inc.; 2020;43:758–69.

56. Youn Baek J, Youn Jun D, Taub D, Ho Kim Y. Characterization of human phosphoserine aminotransferase involved in the phosphorylated pathway of L-serine biosynthesis. Biochem. J. 2003.

57. Murtas G, Marcone GL, Sacchi S, Pollegioni L. L-serine synthesis via the phosphorylated pathway in humans. Cellular and Molecular Life Sciences. Springer Science and Business Media Deutschland GmbH; 2020. p. 5131–48.

58. Koper K, Han SW, Pastor DC, Yoshikuni Y, Maeda HA. Evolutionary origin and functional diversification of aminotransferases. Journal of Biological Chemistry. American Society for Biochemistry and Molecular Biology Inc.; 2022.

59. Winston F, Dollard C, Ricupero-Hovasse SL. Construction of a set of convenient saccharomyces cerevisiae strains that are isogenic to S288C. Yeast. Chichester, UK: John Wiley & Sons, Ltd; 1995;11:53–5.

60. Rose M, Winston F, and Hieter P. Methods in Yeast Genetics: A Laboratory Course Manual. Cold Spring Harbor Laboratory Press; 1990.

61. Farrell CM, O'Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, et al. Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. 2014;42:D865–72.

62. Illuxley C, Green ED, Dunbam I. Rapid assessment of S. cerevisiae mating type by PCR. Trends in Genetics. 1990;6:236.

63. Lo RS, Cromie GA, Tang M, Teng K, Owens K, Sirr A, et al. The functional impact of 1,570 SNP-accessible missense variants in human OTC. Available from: https://doi.org/10.1101/2022.10.26.513893.

64. Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, et al. A series of PDB-related databanks for everyday needs. Nucleic Acids Res. 2015;43:D364-8.

65. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.

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

67. 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:D439–44.

68. Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523–31.

69. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016;44:W344-50.

70. Goldenberg O, Erez E, Nimrod G, Ben-Tal N. The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res. 2009;37:D323-7.

71. Singh RK, Kumar D, Gourinath S. Phosphoserine Aminotransferase has Conserved Active Site from Microbes to Higher Eukaryotes with Minor Deviations. Protein Pept Lett. Bentham Science Publishers Ltd.; 2021;28:996–1008.

72. Eliot AC, Kirsch JF. Pyridoxal phosphate enzymes: Mechanistic, structural, and evolutionary considerations. Annu Rev Biochem. 2004. p. 383–415.

73. Liang J, Han Q, Tan Y, Ding H, Li J. Current advances on structure-function relationships of pyridoxal 5'-phosphate-dependent enzymes. Front Mol Biosci. Frontiers Media S.A.; 2019.

74. Denesyuk AI, Denessiouk KA, Korpela T, Johnson MS. Functional attributes of the phosphate group binding cup of pyridoxal phosphate-dependent enzymes. J Mol Biol. Academic Press; 2002;316:155–72.

75. Hester G, Stark W, Moser M, È rg Kallen Zora Markovic Â-Housley J, Jansonius JN. Crystal Structure of Phosphoserine Aminotransferase from Escherichia coli at 2.3 A Ê Resolution:

Comparison of the Unligated Enzyme and a Complex with a a a-Methyl-L-Glutamate. J. Mol. Biol. 1999.

76. Battula P, Dubnovitsky AP, Papageorgiou AC. Structural basis of l-phosphoserine binding to Bacillus alcalophilus phosphoserine aminotransferase. Acta Crystallogr D Biol Crystallogr. 2013;69:804–11.

77. Sekula B, Ruszkowski M, Dauter Z. Structural analysis of phosphoserine aminotransferase (isoform 1) from arabidopsis Thaliana– the enzyme involved in the phosphorylated pathway of serine biosynthesis. Front Plant Sci. Frontiers Media S.A.; 2018;9.

78. Singh RK, Tomar P, Dharavath S, Kumar S, Gourinath S. N-terminal residues are crucial for quaternary structure and active site conformation for the phosphoserine aminotransferase from enteric human parasite E. histolytica. Int J Biol Macromol. Elsevier B.V.; 2019;132:1012–23.

79. Hekkelman ML, de Vries I, Joosten RP, Perrakis A. AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Methods. 2022;

80. 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. Oxford University Press; 2018;46:D1062–7.

81. Brnich SE, Abou Tayoun AN, Couch FJ, Cutting GR, Greenblatt MS, Heinen CD, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. BioMed Central Ltd.; 2020.

82. Ni C, Cheng RH, Zhang J, Liang JY, Wei RQ, Li M, et al. Novel and recurrent PHGDH and PSAT1 mutations in Chinese patients with Neu-Laxova syndrome. European Journal of Dermatology. John Libbey Eurotext; 2019;29:641–6.

83. Shapira Zaltsberg G, McMillan HJ, Miller E. Phosphoserine aminotransferase deficiency: imaging findings in a child with congenital microcephaly. Journal of Maternal-Fetal and Neonatal Medicine. Taylor and Francis Ltd; 2020;33:1033–5.

84. Serrano Olave A, Padín López A, Martín Cruz M, Monís Rodríguez S, Narbona Arias I, Jiménez López JS. Prenatal Diagnosis of Neu–Laxova Syndrome. Diagnostics. MDPI AG; 2022;12:1535.

85. Fu F, Li R, Yu Q, Wang D, Deng Q, Li L, et al. Application of exome sequencing for prenatal diagnosis of fetal structural anomalies: clinical experience and lessons learned from a cohort of 1618 fetuses. Genome Med. BioMed Central Ltd; 2022;14.

86. Gahl WA, Wise AL, Ashley EA. The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. JAMA. 2015;314:1797–8.

87. Macnamara EF, D'Souza P, Tifft CJ. The undiagnosed diseases program: Approach to diagnosis. Ferreira C, editor. Transl Sci Rare Dis [Internet]. 2019;4:179–88. Available from: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/TRD-190045

88. Firth H v, Wright CF, DDD Study. The Deciphering Developmental Disorders (DDD) study. Dev Med Child Neurol. 2011;53:702–3.

89. Manickam K, McClain MR, Demmer LA, Biswas S, Kearney HM, Malinowski J, et al. Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine. Springer Nature; 2021;23:2029–37.

90. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.

91. Heyne HO, Karjalainen J, Karczewski KJ, Lemmelä SM, Zhou W, FinnGen, et al. Mono- and biallelic variant effects on disease at biobank scale. Nature. 2023;613:519–25.

92. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613:508–18.

93. Yachie N, Petsalaki E, Mellor JC, Weile J, Jacob Y, Verby M, et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. Mol Syst Biol. EMBO; 2016;12:863.

94. Schlecht U, Liu Z, Blundell JR, St Onge RP, Levy SF. A scalable double-barcode sequencing platform for characterization of dynamic protein-protein interactions. Nat Commun. Nature Publishing Group; 2017;8.

95. Díaz-Mejía JJ, Celaj A, Mellor JC, Coté A, Balint A, Ho B, et al. Mapping DNA damage-dependent genetic interactions in yeast via party mating and barcode fusion genetics . Mol Syst Biol. EMBO; 2018;14.

96. Younger D, Berger S, Baker D, Klavins E. High-Throughput Characterization of Protein-Protein Interactions by Reprogramming Yeast Mating. High-throughput characterization of protein–protein interactions by reprogramming yeast mating. 2017;122143.

97. Jaffe M, Sherlock G, Levy SF. iSeq: A new double-barcode method for detecting dynamic genetic interactions in yeast. G3: Genes, Genomes, Genetics. Genetics Society of America; 2017;7:143–53.

98. Liu X, Liu Z, Dziulko AK, Li F, Miller D, Morabito RD, et al. iSeq 2.0: A Modular and Interchangeable Toolkit for Interaction Screening in Yeast. Cell Syst. Cell Press; 2019;8:338-344.e8.

99. Durrant MG, Fanton A, Tycko J, Hinks M, Chandrasekaran SS, Perry NT, et al. Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. Nat Biotechnol. 2022;

100. Merrick CA, Zhao J, Rosser SJ. Serine Integrases: Advancing Synthetic Biology. ACS Synth Biol. American Chemical Society; 2018;7:299–310.

101. Grindley NDF, Whiteson KL, Rice PA. Mechanisms of site-specific recombination. Annu Rev Biochem. 2006. p. 567–605.

102. Lee G, Saito I. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. Gene. 1998;216:55–65.

103. Li F, Salit ML, Levy SF. Unbiased Fitness Estimation of Pooled Barcode or Amplicon Sequencing Studies. Cell Syst. Cell Press; 2018;7:521-525.e4.

104. Stark WM. Making serine integrases work for us. Curr Opin Microbiol. 2017;38:130–6.

105. Xu Z, Brown WRA. Comparison and optimization of ten phage encoded serine integrases for genome engineering in Saccharomyces cerevisiae. BMC Biotechnol. BioMed Central Ltd.; 2016;16.

106. Matreyek KA, Stephany JJ, Chiasson MA, Hasle N, Fowler DM. An improved platform for functional assessment of large protein libraries in mammalian cells. Nucleic Acids Res. Oxford University Press; 2020;48.

107. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6:343–5.

108. Maddirevula S, Shamseldin HE, Sirr A, AlAbdi L, Lo RS, Ewida N, et al. Exploiting the Autozygome to Support Previously Published Mendelian Gene-Disease Associations: An Update. Front Genet. 2020;11.

109. Hentges P, van Driessche B, Tafforeau L, Vandenhaute J, Carr AM. Three novel antibiotic marker cassettes for gene disruption and marker switching in Schizosaccharomyces pombe. Yeast. 2005;22:1013–9.

110. Taxis C, Knop M. System of centromeric, episomal, and integrative vectors based on drug resistance markers for Saccharomyces cerevisiae. Biotechniques. 2006;40:73–8.

111. Cuperus JT, Lo RS, Shumaker L, Proctor J, Fields S. A tetO Toolkit to Alter Expression of Genes in Saccharomyces cerevisiae. ACS Synth Biol. American Chemical Society; 2015;4:842–52.

112. Kaishima M, Ishii J, Matsuno T, Fukuda N, Kondo A. Expression of varied GFPs in Saccharomyces cerevisiae: Codon optimization yields stronger than expected expression and fluorescence intensity. Sci Rep. Nature Publishing Group; 2016;6.

113. Curran KA, Morse NJ, Markham KA, Wagman AM, Gupta A, Alper HS. Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. ACS Synth Biol. American Chemical Society; 2015;4:824–32.

114. Yoshimatsu T, Nagawa F. Control of Gene Expression by Artificial Introns in Saccharomyces Cerevisiae. Science (1979) [Internet]. Washington: The American Association for the Advancement of Science; 1989;244:1346. Available from: https://www.proquest.com/scholarly-journals/control-gene-expression-artificial-introns/docview/213535414/se-2?accountid=14784.

115. Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc. 2007;2:31–4.

116. Jusiak B, Jagtap K, Gaidukov L, Duportet X, Bandara K, Chu J, et al. Comparison of Integrases Identifies Bxb1-GA Mutant as the Most Efficient Site-Specific Integrase System in Mammalian Cells. ACS Synth Biol. American Chemical Society; 2019;8:16–24.

117. Haase MAB, Truong DM, Boeke JD. Superloser: A Plasmid Shuffling Vector for Saccharomyces cerevisiae with Exceedingly Low Background. G3 (Bethesda). 2019;9:2699–707.
118. John RA. Pyridoxal phosphate-dependent enzymes. Biochim Biophys Acta. 1995;1248:81–96.

119. Zhang Y, Ma C, Dischert W, Soucaille P, Zeng A. Engineering of Phosphoserine Aminotransferase Increases the Conversion of l-Homoserine to 4-Hydroxy-2-ketobutyrate in a Glycerol-Independent Pathway of 1,3-Propanediol Production from Glucose. Biotechnol J. 2019;14:1900003.

120. Feng M, Cui H, Tu W, Li L, Gao Y, Chen L, et al. An integrated pan-cancer analysis of PSAT1: A potential biomarker for survival and immunotherapy. Front Genet. 2022;13.
121. Yang M, Vousden KH. Serine and one-carbon metabolism in cancer. Nat Rev Cancer. Nature Publishing Group; 2016. p. 650–62.

122. Buqué A, Galluzzi L, Montrose DC. Targeting Serine in Cancer: Is Two Better Than One? Trends Cancer. Cell Press; 2021. p. 668–70.

123. Ravez S, Spillier Q, Marteau R, Feron O, Frédérick R. Challenges and Opportunities in the Development of Serine Synthetic Pathway Inhibitors for Cancer Therapy. J Med Chem. 2017;60:1227–37.

124. Korangath P, Teo WW, Sadik H, Han L, Mori N, Huijts CM, et al. Targeting Glutamine Metabolism in Breast Cancer with Aminooxyacetate. Clinical Cancer Research. 2015;21:3263–73.

125. Dong G, Ding Y, He S, Sheng C. Molecular Glues for Targeted Protein Degradation: From Serendipity to Rational Discovery. J Med Chem. 2021;64:10606–20.

126. Cellini B, Montioli R, Oppici E, Astegno A, Borri Voltattorni C. The chaperone role of the pyridoxal 5'-phosphate and its implications for rare diseases involving B6-dependent enzymes. Clin Biochem. 2014. p. 158–65.

127. Ramos RJ, Pras-Raves ML, Gerrits J, van der Ham M, Willemsen M, Prinsen H, et al. Vitamin B6 is essential for serine de novo biosynthesis. J Inherit Metab Dis. 2017;40:883–91.

128. Sun S, Weile J, Verby M, Wu Y, Wang Y, Cote AG, et al. A proactive genotype-to-patient-phenotype map for cystathionine beta-synthase. Genome Med. BioMed Central Ltd.; 2020;12.
129. Wilson MP, Plecko B, Mills PB, Clayton PT. Disorders affecting vitamin B6 metabolism. J Inherit Metab Dis. John Wiley and Sons Inc.; 2019. p. 629–46.

130. Beyzaei Z, Nabavizadeh S, Karimzadeh S, Geramizadeh B. The mutation spectrum and ethnic distribution of non-hepatorenal tyrosinemia (types II, III). Orphanet J Rare Dis. BioMed Central Ltd; 2022.

131. Urrestarazu A, Vissers S, Iraqui I, Grenson M. Phenylalanine- and tyrosine-auxotrophic mutants of Saccharomyces cerevisiae impaired in transamination. Mol Gen Genet. 1998;257:230–7.
132. Pronk JT. Auxotrophic Yeast Strains in Fundamental and Applied Research. Appl Environ Microbiol. 2002;68:2095–100.

133. Shou W, Ram S, Vilar JMG. Synthetic cooperation in engineered yeast populations. Proceedings of the National Academy of Sciences. 2007;104:1877–82.

134. Esch BM, Limar S, Bogdanowski A, Gournas C, More T, Sundag C, et al. Uptake of exogenous serine is important to maintain sphingolipid homeostasis in Saccharomyces cerevisiae. PLoS Genet. 2020;16:e1008745.

135. Fu J, Chen L, Su T, Xu S, Liu Y. Mild phenotypes of phosphoglycerate dehydrogenase deficiency by a novel mutation of PHGDH gene: Case report and literature review. International Journal of Developmental Neuroscience. 2023;83:44–52.

136. Byers HM, Bennett RL, Malouf EA, Weiss MD, Feng J, Scott CR, et al. Novel report of phosphoserine phosphatase deficiency in an adult with myeloneuropathy and limb contractures. JIMD Rep. Springer; 2016. p. 103–8.

137. Vincent JB, Jamil T, Rafiq MA, Anwar Z, Ayaz M, Hameed A, et al. Phosphoserine phosphatase (PSPH) gene mutation in an intellectual disability family from Pakistan. Clin Genet. 2015;87:296–8.

138. Tavtigian S v., Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genetics in Medicine. Nature Publishing Group; 2018;20:1054–60.

139. Lefebvre M, Bruel A-L, Tisserant E, Bourgon N, Duffourd Y, Collardeau-Frachon S, et al. Genotype-first in a cohort of 95 fetuses with multiple congenital abnormalities: when exome sequencing reveals unexpected fetal phenotype-genotype correlations. J Med Genet. 2021;58:400–13.

140. Méneret A, Wiame E, Marelli C, Lenglet T, van Schaftingen E, Sedel F. A serine synthesis defect presenting with a Charcot-Marie-Tooth-like polyneuropathy. Arch Neurol. 2012;69:908–11.

141. Bastarache L, Hughey JJ, Hebbring S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized mendelian disease patterns. Science (1979). American Association for the Advancement of Science; 2018;359:1233–9.

142. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res. 2008;18:99–113.

143. Chapman JR, Taylor MRG, Boulton SJ. Playing the End Game: DNA Double-Strand Break Repair Pathway Choice. Mol Cell. 2012. p. 497–510.

144. Storici F, Durham CL, Gordenin DA, Resnick MA. Chromosomal site-specific double-strand breaks are efficiently targeted for repair by oligonucleotides in yeast. Proc Natl Acad Sci U S A. 2003;100:14994–9.

145. Cuperus JT, Lo RS, Shumaker L, Proctor J, Fields S. A tetO Toolkit to Alter Expression of Genes in Saccharomyces cerevisiae. ACS Synth Biol. 2015;4:842–52.

146. Lambert AR, Hallinan JP, Shen BW, Chik JK, Bolduc JM, Kulshina N, et al. Indirect DNA Sequence Recognition and Its Impact on Nuclease Cleavage Activity. Structure. 2016;24:862–73.

147. Yoshimatsu T, Nagawa F. Control of gene expression by artificial introns in Saccharomyces cerevisiae. Science (1979). 1989;244:1346–8.

148. Kramara J, Osia B, Malkova A. Break-Induced Replication : The Where , The Why , and The How. Trends in Genetics. Elsevier Ltd; 2018;34:518–31.

149. Hum YF, Jinks-Robertson S. Mismatch recognition and subsequent processing have distinct effects on mitotic recombination intermediates and outcomes in yeast. Nucleic Acids Res. 2019;47:4554–68.

150. Tsabar M, Eapen V V., Mason JM, Memisoglu G, Waterman DP, Long MJ, et al. Caffeine impairs resection during DNA break repair by reducing the levels of nucleases Sae2 and Dna2. Nucleic Acids Res. 2015;43:6889–901.

151. Paczia N, Becker-Kettern J, Conrotte JF, Cifuente JO, Guerin ME, Linster CL. 3-Phosphoglycerate Transhydrogenation Instead of Dehydrogenation Alleviates the Redox State Dependency of Yeast de Novo l -Serine Synthesis. Biochemistry. American Chemical Society; 2019;58:259–75.

152. Peeraer Y, Rabijns A, Collet J-F, van Schaftingen E, de Ranter C. How calcium inhibits the magnesium-dependent enzyme human phosphoserine phosphatase. Eur J Biochem. 2004;271:3421–7.