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

Michael J. Xie[1,2]†, Gareth A. Cromie[1]†, Katherine Owens[1,3], Martin S. Timour[1], Michelle Tang[1], J. Nathan Kutz[3], Ayman W. El-Hattab[4], Richard N. McLaughlin Jr.[1] and Aimée M. Dudley[1,2]*

[1] Pacific Northwest Research Institute, Seattle, Washington, USA

[2] Molecular Engineering Graduate Program, University of Washington, Seattle, Washington, USA

[3] Department of Applied Mathematics, University of Washington, Seattle, Washington, USA

[4] Department of Clinical Sciences, College of Medicine, University of Sharjah, Sharjah, United Arab Emirates

†These authors contributed equally to this work.

Keywords: Yeast, Serine Deficiency, Neu-Laxova Syndrome, PSAT1, Inborn Errors of Metabolism

***Corresponding Author:**

Aimée M. Dudley, Ph.D.

Pacific Northwest Research Institute

720 Broadway

Seattle, WA 98122

aimee.dudley@gmail.com

Abstract

Background: Pathogenic variants in *PHGDH*, *PSAT1*, and *PSPH* cause a set of rare, autosomal recessive diseases known as serine biosynthesis defects. Serine biosynthesis defects present in a broad phenotypic spectrum that includes, at the severe end, Neu–Laxova syndrome, 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. 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.

Methods: Recently, our laboratory established a yeast-based assay for human *PSAT1* function. We have now applied it at scale to assay the functional impact of 1,914 SNV-accessible amino acid substitutions. In addition to assaying the functional impact of individual variants in yeast haploid cells, we can assay pairwise combinations of *PSAT1* alleles that recapitulate human genotypes, including compound heterozygotes, in yeast diploids.

Results: Results of our assays of individual variants (in haploid yeast cells) agree well with clinical interpretations and protein structure-function relationships, supporting the use of our data as functional evidence under the ACMG interpretation guidelines. Results from our diploid assay successfully distinguish patient genotypes from those of healthy carriers and agree well with disease severity. Finally, we present 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.

Conclusions: Taken together, our work provides an example of how large-scale functional assays in model systems can be powerfully applied to the study of a rare disease.

Background

High throughput sequencing technologies have accelerated the discovery of human genetic variation. Data from the Genome Aggregation Database (gnomAD) [1] suggest that the average person harbors approximately 200 rare (allele frequency <0.1%) protein coding variants [2]. However, leveraging genomic information for the diagnosis and treatment of medically actionable diseases is currently limited by the relatively small number of alleles for which clinical interpretation is readily available. Most of the ~550K missense variants in ClinVar [3], even those in well-studied disease genes, are variants of uncertain significance (VUS). The difficulty of variant interpretation is compounded further in rare diseases, and patients often face long diagnostic odysseys or remain undiagnosed.

Inborn errors of metabolism (IEM), genetic alterations that disrupt metabolic homeostasis, are a relatively large class of rare diseases. Over 700 IEMs [4] have been described and new disorders continue to be discovered [5]. Although individually rare, collectively these diseases are common, with estimates suggesting that IEMs may affect 1 in 800-2600 live births [6,7]. 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. Therapeutic strategies often focus on amending these imbalances. 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 [8–10].

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 (**Fig. 1**). 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 [11]. 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. As a result, serine biosynthesis defects negatively impact nervous system development and function, with clinical manifestations that include microcephaly, seizures, intellectual disability, and neuropathies. 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 [12–16]. Prompt diagnosis is crucial, as the impact of therapeutic intervention, including prenatal dietary supplementation, is greatest before patients become symptomatic and irreversible neurological damage occurs.

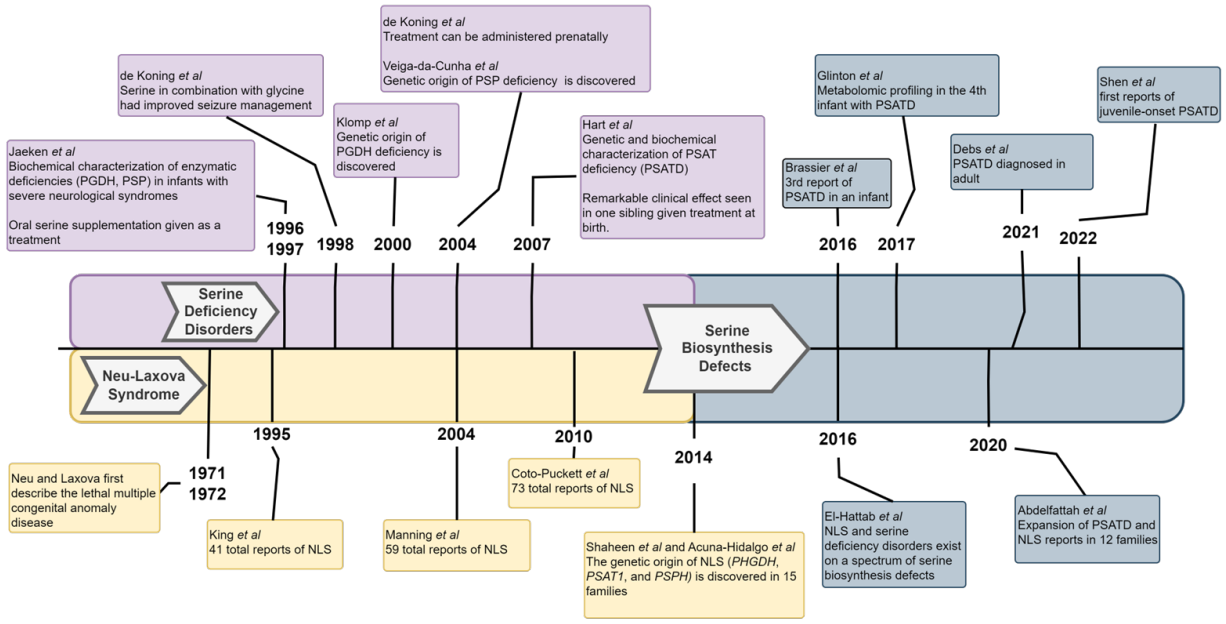


Fig. 1 Key developments in the study of serine deficiency disorders, a set rare but treatable inborn errors of metabolism [11–30].

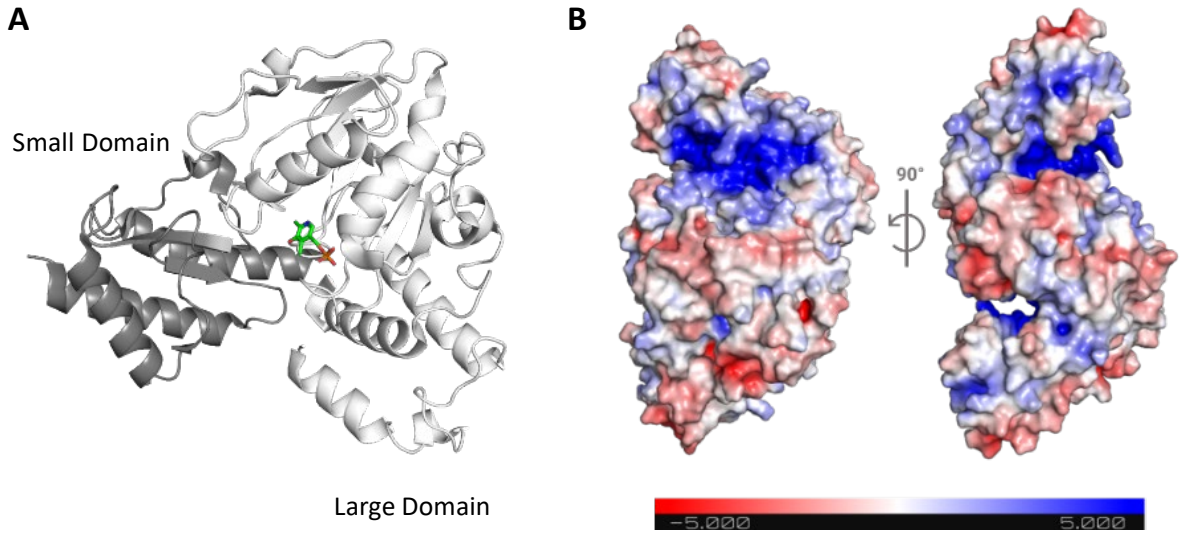
Unfortunately, like many rare diseases, diagnosing serine deficiency disorders is difficult in the absence of a family history of disease. One challenge is that the disorders display a broad phenotypic spectrum (reviewed in [11,31]). At one extreme, Neu-Laxova syndrome (NLS) is a severe, lethal, multiple

93 congenital anomaly disease characterized by marked intrauterine growth restriction, microcephaly,
94 distinctive facial features, limb and skin defects, and variable brain malformations including
95 lissencephaly, corpus callosum agenesis, and hypoplastic cerebellum and pons. Infantile serine
96 biosynthesis defects represent the intermediate form of the disorder spectrum, typically presenting with
97 intrauterine growth restriction, microcephaly, feeding difficulties, vomiting, irritability, hypertonia
98 evolving into spastic tetraplegia, nystagmus, early onset seizures, and poor psychomotor development.
99 Other findings include congenital cataracts, adducted thumbs, megaloblastic anemia, brain atrophy,
100 hypomyelination, and hypoplastic cerebellum and pons. At the mild end of the disorder spectrum are
101 childhood serine biosynthesis defects characterized by developmental delay, intellectual disability, and
102 behavioral abnormalities. Other findings include epilepsy, ataxia, nystagmus, polyneuropathy, and
103 hypertonia. Although it has been hypothesized that residual enzyme activity may explain the phenotypic
104 variability [25,28,32] within serine deficiency disorders, functional studies regarding this subject are
105 sparse and the natural histories of the disorders are largely unknown. Biochemical diagnosis of serine
106 biosynthesis defects can be challenging as it requires measurement of amino acid levels in patient serum
107 or cerebral spinal fluid (CSF) with comparison to closely age-matched controls. Although the serum-
108 directed metabolic screen is more widely used, considered generally in children with intellectual
109 disability, serine and glycine plasma concentrations can be normal in the presence of serine biosynthesis
110 defects if the blood sample is not obtained in a fasted state. In contrast, CSF serine and glycine
111 concentrations are not affected by meals, however, CSF amino acid analysis is not typically done in the
112 absence of seizures. Therefore, the metabolic work up may fail to identify children with serine
113 biosynthesis defects [27,33].

114 Identifying the presence of pathogenic sequence variants through whole exome/genome sequencing
115 has several potential advantages for diagnosing inherited metabolic diseases, such as serine biosynthesis
116 defects. However, this approach relies on having clinical interpretations for rare variants, which are
117 generally not available. Instead, within the current disease literature, clinical sequencing has primarily
118 been used as a tool by experts to provide molecular confirmation once a patient has become
119 symptomatic. High-throughput approaches to variant interpretation that can be applied to rare variants
120 are needed to increase the success rate of sequencing-based diagnostics. Although computational
121 prediction methods can be applied at scale, their relatively high error rates limit their utility [34].
122 Functional assays that can quantitatively measure variant effects on protein activity [35] are an
123 alternative approach that can be used as part of the criteria for variant interpretation established by the
124 American College of Medical Genetics (ACMG) [36]. Improvements in DNA synthesis technology have
125 enabled time-and-cost effective methods for building large variant libraries. When combined with high-
126 throughput phenotyping methods, functional assays can comprehensively assess variants that have been
127 identified in the human population as well as those that may arise in the future [37,38].

128 Recently, our laboratory established a yeast-based assay for human PSAT function [39]. PSAT catalyzes
129 the reversible conversion of glutamate to α-ketoglutarate and 3-phosphohydroxypyruvate (3-PHP) to
130 phosphoserine [40]. PSAT belongs to a large family of fold type I pyridoxal phosphate (PLP; the active

131 form of vitamin b6) cofactor-dependent enzymes, and more specifically, the class IV group of
132 aminotransferases [41,42]. The crystal structure of human PSAT complexed with PLP (PDB: 3e77;
133 residues L17-L370) shows that the protein adopts an s-shaped homodimer assembly in which two
134 separate PLP-containing active sites surrounded by an overall positive charge distribution reside along
135 the interface (Fig. 2B). Each subunit of the homodimer consists of two domains; a large domain, which
136 forms the dimer interface and binds PLP, and a small domain (Fig. 2A). Despite extensive literature on
137 the structure and biochemistry of PSAT [40], there is not a comprehensive understanding of which
138 mutations at which positions impair the ability of PSAT to support normal serine biosynthesis.



139

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

145

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

Methods

Strain library construction

All *Saccharomyces cerevisiae* strains used in this study (Additional File 1: Tables S1 and S2) were derived from the isogenic lab strains FY4 (MAT**a**) and FY5 (MATα) [43]. 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 [44].

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* [39]. 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 [45]). 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 (described in detail in Additional File 2). 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.

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

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

208 **Validating the MinION Variant Calls Using Illumina Sequencing**

209 To validate our Oxford Nanopore pipeline, we performed amplicon sequencing on the Illumina
210 sequencing platform to confirm the variant in each well of each variant plate. These calls were then
211 compared to the result from the Oxford Nanopore pipeline. For the Illumina sequencing, the *yPSAT1* ORF
212 was divided up into five ~300bp overlapping segments. The specific segment to be amplified and
213 sequenced corresponded to the known codon position of the variant within our library. 35 cycles of PCR
214 amplification were performed using genomic DNA as a template. The initial PCR reaction was then
215 diluted and used for an additional 24 cycles of PCR to add on Illumina index sequences. The final PCR
216 reactions were pooled and gel purified before performing Illumina sequencing on a Nextseq 500 using
217 300-cycle, dual-indexed, paired-end sequencing. Reads were demultiplexed and aligned to the *yPSAT1*
218 reference sequence as described [39]. For each isolate, the basecalls across the target codon position
219 were used to identify the variant codon sequence.

220 Comparison of the MinION and Illumina results revealed excellent agreement (95.5%) for the MinION
221 calls passing the quality filters (Additional File 2: Fig. S1). Most of the "variants" failing the MinION quality
222 filters are occasions when, as determined by the Illumina sequencing, no change had actually been made
223 at the target codon (i.e. wild type *yPSAT1* sequence). As expected, for these "no-change" isolates, the
224 most frequent potential variant identified by the MinION pipeline occurred at a low frequency similar to
225 the error frequency for that variant (Additional File 2: Fig. S1A). In contrast, when a variant was present,
226 it was observed at higher frequency and was correctly identified with high accuracy.

227 The accession number for the Illumina and MinION read sequences generated for this study is (in
228 process, available upon manuscript acceptance).

229 **Generation of Yeast Diploids**

230 Unless noted, all *yPSAT1* variants are in a single mating type (MAT**a**) (Additional File 1: Table S1). To
231 generate diploid strains harboring combinations of *yPSAT1* alleles (missense variant, wild type, or null),
232 variants were first introduced into a strain of the opposite mating type. The resulting *yPSAT1* MATα
233 strains were then mated in a pairwise manner to the relevant *yPSAT1* MAT**a** strains using standard
234 methods [46]. The resulting diploid strains, harboring combinations of *yPSAT1* alleles (Additional File 1:
235 Table S2), were then arrayed in an alternating checkerboard pattern that minimizes the influence of
236 nutrient competition from neighboring colonies during phenotype measurement.

237 **Growth assays**

238 Before phenotyping, the set of variant strains generated in this study were extended to include the
239 haploid strains carrying missense variants constructed and sequence confirmed by Sirr *et al* [39]. These
240 additional strains were re-arrayed to match the phenotyping layout used in this study, with at least 2
241 isolates of each variant. Additional control plates included a plate with wild type at every position, and
242 plates consisting entirely of D100A or A99V (with control positions) *yPSAT1* strains, which are ClinVar
243 pathogenic alleles encoding missense variants.

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

253 **Image-based growth quantification**

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

261 **Growth data fitting and normalization**

262 For haploid strains, raw phenotypic values were normalized, quality control filters were applied to each
263 isolate, and a final relative growth estimate for each variant was determined, as described in Lo *et al*
264 [47]. Briefly, normalization steps were carried out to account for the effects of plate-to-plate variation,
265 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 [48,49]. AlphaFold (version 2022-11-01) [50,51] was used to predict the structure of the missing

302 16 N-terminal residues. Molecular visualizations were created using the PyMol Molecular Graphics
303 System Version 2.3.2 (Schrödinger). Tools from the publicly available PyMol script and plugin repository
304 were used to determine interfacial residues and charge distribution. Dimer interface residues were
305 defined as residues at which the solvent-exposed surface area in the monomeric model is greater than
306 the solvent-exposed surface area in the dimer model (cutoff value = 0.5 $Å^2$). Macromolecular
307 electrostatics were estimated for each residue of PDB 3e77 using the Adaptive Poisson-Boltzmann Solver
308 PyMol plugin. Evolutionary conservation scores and 'grades' for each position of the full-length amino
309 acid sequence of human PSAT (UniProtID [52]: Q9Y617-1) were computed using ConSurf [53]. All
310 parameters for the ConSurf calculation were the same as the methodology outlined in creating the
311 ConSurf-DB repository [54], with the best evolutionary model determined to be WAG. The evolutionary
312 conservation scores and 'grades' represent the calculated positional conservation based on an amino
313 acid alignment of 300 diverse homologs of PSAT. The determined structure and conservation features
314 for each amino acid position is provided in Additional File 1: Table S6.

315 **Results**

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

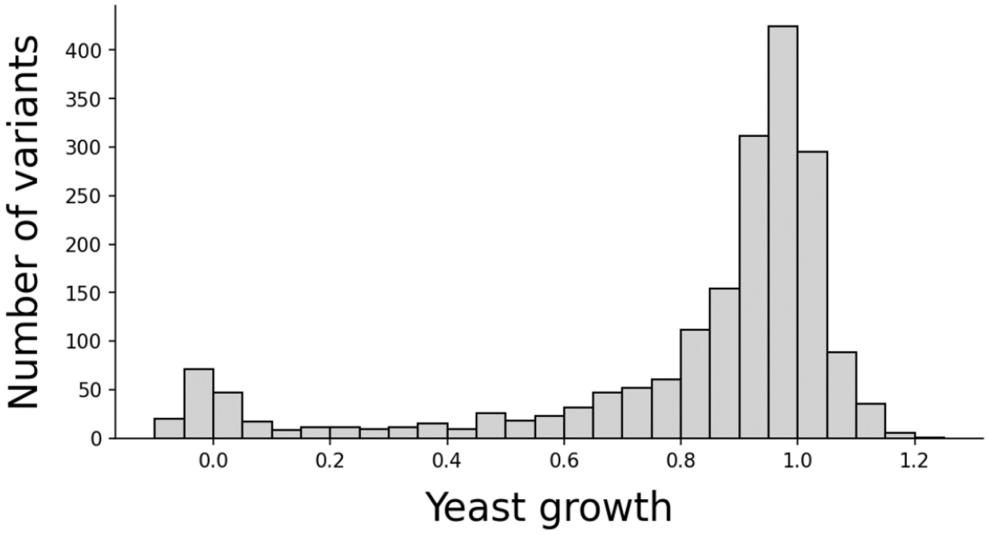
317 Our laboratory had previously established a yeast-based complementation assay that leveraged the
318 ability of the human *PSAT1* coding sequence to functionally replace its yeast ortholog, *SER1* [39]. In this
319 assay, growth on minimal medium lacking serine provides a quantitative readout of PSAT activity,
320 allowing the functional impact of protein coding variants to be assessed. Variant impact is expressed on
321 a relatively intuitive scale of activity between the level of yeast growth associated with no activity (that
322 of a complete gene deletion) and that conferred by the wild type human protein coding sequence
323 (*yPSAT1*). Variants causing a reduction in PSAT activity could have their effect via decreases in enzymatic
324 activity, protein stability, or a combination of both.

325 In this study, we applied the assay at scale to measure the effect of thousands of amino acid substitutions
326 as follows. First, with the exception of the translational initiation and termination codons, we identified
327 all unique amino acid substitutions (n = 2,182) that were accessible via a SNV across the full length of
328 human *PSAT1* isoform 1 cDNA (1,113 bp). Variant codons encoding each amino acid substitution were
329 then introduced into the yeast codon-optimized version of *PSAT1* (*yPSAT1*). The resulting variant library
330 was transformed into a haploid *SER1* deletion strain (*ser1Δ0*) and integrated in single copy at the *SER1*
331 locus of the yeast genome, under the control of the endogenous *SER1* transcriptional promoter and
332 terminator. Next, transformants were individually arrayed in 96-well plate format, and the identity of
333 the variant codon present in each transformant was determined by Oxford Nanopore MinION
334 sequencing (Methods). Strains that harbored secondary mutations in the protein coding sequence were
335 removed from further consideration. The arrayed strain library was then grown, in triplicate, on solid
336 medium lacking serine and imaged after three days at 30°C. Data from these images were used to
337 measure the growth of each strain, relative to wild type (*yPSAT1*) and null (*ser1Δ0*), using a custom
338 automated image analysis pipeline and normalization software (Methods).

339 Here, we report the functional impact of 1,914 amino acid substitutions in *PSAT1* (Additional File 1: Table
340 S3), corresponding to ~88% of all unique SNV-accessible amino acid substitutions (n=2,182) in the human
341 *PSAT1* cDNA sequence. A subset of these substitutions (n=196) was also assayed in our previous study
342 [39] which included the full set of *PSAT1* missense variants described in ClinVar, gnomAD, or the clinical
343 literature at that time. Reanalysis in the current study allowed us to assess these amino acid substitutions
344 using an improved data processing pipeline as the 1,718 new substitutions, thereby placing them on a
345 common scale to facilitate direct comparison. Despite the use of slightly different data analysis pipelines,
346 there was excellent agreement (R^2=0.94) between the normalized growth estimates of the 196
347 substitutions in the two studies.

348 Among the full set of 1,914 amino acid substitutions, the distribution of growth values was bimodal, with
349 a large peak centered near the value of the wild type *yPSAT1* strain (normalized growth=1) and a smaller
350 peak centered around the value of the null (deletion) strain (normalized growth=0) (Fig. 3). These results
351 are consistent with a large group of protein coding variants showing little or no functional impairment
352 relative to wild type, and a smaller group behaving as complete loss of function alleles, comparable to
353 the null control. The remaining protein coding variants showed varying levels of functional impairment
354 (Fig. 3). On this basis, we classified amino acid substitutions in our assay as follows. Substitutions with
355 ≤95% normalized growth were considered functionally impaired relative to wild type and the remaining
356 substitutions (>95% normalized growth) were considered unimpaired. Among the impaired class, we
357 further defined any substitutions resulting in ≤5% normalized growth (i.e. comparable to that of the null
358 control) as amorphs, and any substitutions resulting in less severe functional impairment (>5% and ≤95%
359 normalized growth) as hypomorphs.

360



361

362 **Fig. 3** Distribution of variant effect for 1,914 PSAT1 unique SNV-accessible missense amino acid substitutions. Frequency (in
363 5% intervals) of experimentally measured yeast growth values scaled relative to wild type yPSAT1 (normalized growth=1) and
364 null (normalized growth=0).

365

366 **Mapping functional effects to protein structure and conservation**

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

373 The overall structural organization and active site architecture of PSAT are conserved with other
374 members of the class IV family of aminotransferases [42], and phosphoserine aminotransferases from
375 eukaryotic and prokaryotic organisms [55]. To examine our functional scores in the context of
376 evolutionary conservation, we used ConSurf [53] scores (Methods), which are derived from amino acid
377 sequence alignments of homologs. As illustrated in Fig. 4, more conserved (negative ConSurf score)
378 regions localize near the interfacial active sites of the PSAT protein. We expected that highly conserved
379 residues in PSAT would be sensitive to amino acid substitutions. Consistent with this hypothesis, there
380 was a highly significant correlation (Spearman rank correlation=0.47, $p < 7.2 \times 10^{-22}$, Additional File 2:
381 Fig. S2A) between the ConSurf score for a residue and the median yeast assay score for substitutions at
382 that residue. As expected, more conserved residues displayed lower median growth. Examining the
383 relationship between the ConSurf scores at a residue and the growth of individual substitutions at that
384 residue further resolved this relationship (Additional File 2: Fig. S2B). Amino acid substitutions that
385 impaired PSAT activity were concentrated at highly conserved residues within PSAT (Fig. 4). However,
386 while few amino acid substitutions at weakly conserved residues exhibited impairment of PSAT function,
387 many substitutions at conserved positions did not (Fig. 4 and Additional File 2: Fig. S2B).

388

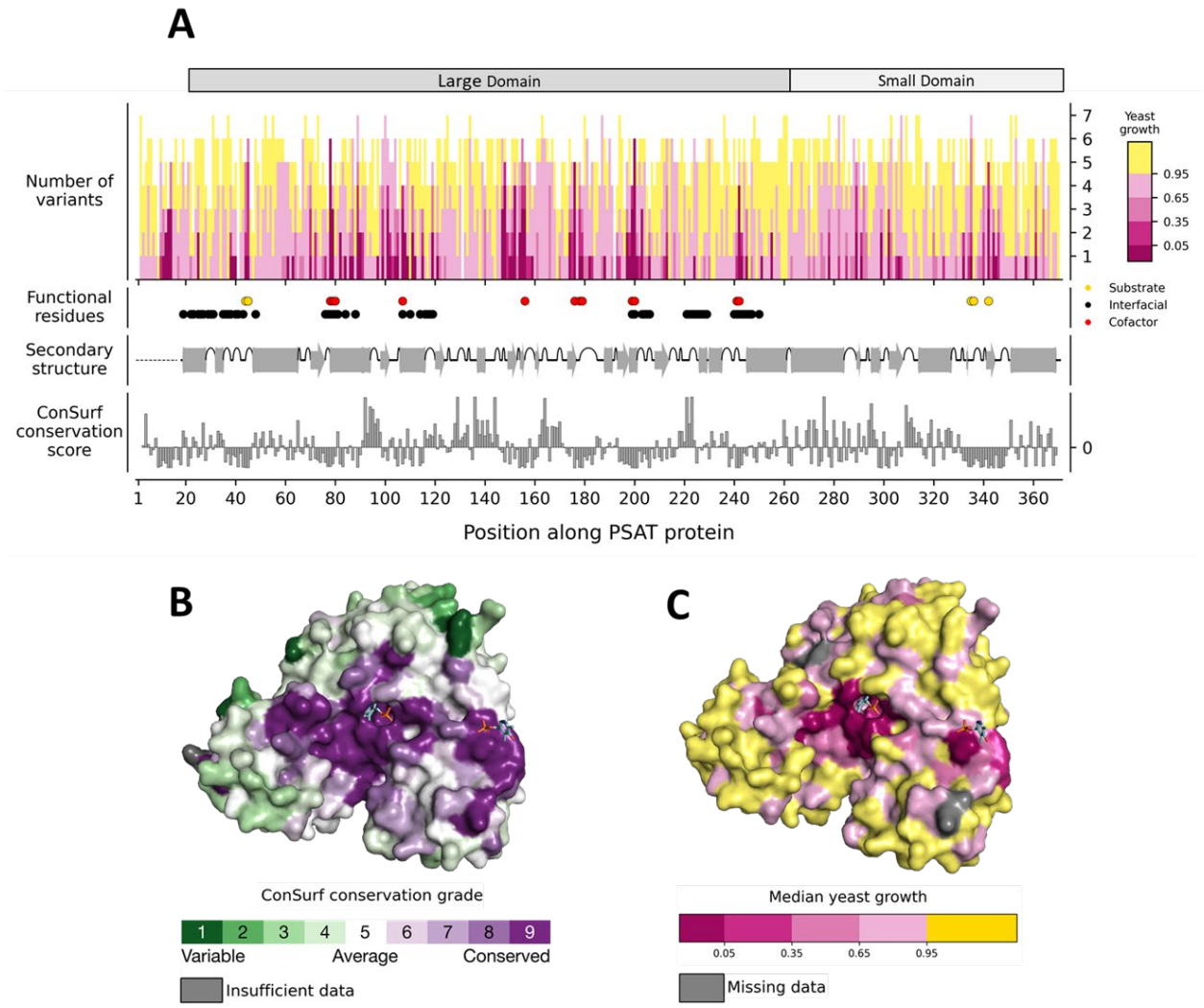


Fig. 4 Missense variant effect map across the length of human PSAT compared to structural features and conservation. **A** The top plot depicts the growth of the possible 4-7 substitutions introduced at each amino acid position, ordered from highest (top) to lowest (bottom) growth. Overlayed below are 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).

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 (Fig. 4). 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%,

406 significantly, and very substantially, lower than the global median of 93.6% ($p < 1x10^{-5}$, one sided
407 permutation test).

408 We next examined the functional impact of variants at each of these positions individually. In the crystal
409 structure of PSAT complexed with PLP (PDB: 3e77), 12 residues directly interact with the functional
410 groups of the cofactor (Fig. 5A and Fig. 5B) either in pyridoxal ring binding/ coordination or in phosphate
411 group binding. We expected that amino acid substitutions at residues involved in pyridoxal ring binding
412 or coordination would be sensitive to amino acid substitutions (Fig. 5A), as this functional group is
413 directly involved in catalysis [42,56,57]. In fact, all tested substitutions at these six residues (K200, D176,
414 T156, W107, S178, and S179) were functionally impaired. The majority of these, (19/29), including all
415 SNV-accessible missense variants at the catalytic lysine (K200) were amorphic (Fig. 5C). Next, we
416 examined the six residues that participate in phosphate group binding, which may not participate
417 directly in catalysis and whose functional role is less well understood (Fig. 5B). The phosphate group may
418 act as an anchor point to the protein [58], or even directly interact with nearby substrates [57]. At three
419 of these positions (G78, G79, and Q199) all variants were functionally impaired and had amorphic
420 median growth scores (Fig. 5C). Functionally impaired variants were also observed at C80, N241*, and
421 T242*(asterisk indicates residues from opposite subunit), although some substitutions were tolerated
422 (Fig. 5C). Interestingly, most substitutions at the C80 residue are not impaired relative to wild type with
423 the only exception being C80R (growth score=80%). It has been hypothesized that C80 contributes to
424 the higher phosphoserine substrate affinity (Km = 5 μM) observed in human PSAT relative to other
425 phosphoserine aminotransferases [55], and possibly the positive charge introduction represented by
426 C80R negatively impacts this binding.

427 Because the human PSAT crystal structure was solved only without bound substrate, we considered 5
428 conserved substrate binding residues that were identified in *Escherichia coli* [59], *Bacillus alcalophilus*
429 [60], and *Arabidopsis thaliana* [61] phosphoserine aminotransferases that had crystal structures of
430 phosphoserine or α-methyl-L-glutamate (analog) bound states. Amino acid substitutions at all but one
431 of these sites, corresponding to H335, R336, R342, H44* and R45* (asterisk indicates residues from
432 opposite subunit) in human PSAT, show strong loss of function (Fig. 5D and Fig. 5E). The exception is
433 R336, which displayed activity similar to wild type for all tested substitutions. Thus, R336 may be a
434 conserved residue that does not directly participate in substrate stabilization in PSAT, but does in other
435 orthologs.

436 The solved human PSAT crystal structure also lacks the first 16 N-terminal amino acids (PDB: 3e77), in
437 which we observe a cluster of sites intolerant to amino acid substitution (residues 11-14) (Fig. 4A).
438 AlphaFold [50,51] predicts that these residues are located in the small domain, proximal to the active
439 site (Additional File 2: Fig. S3). Interestingly, a deletion of 4 N-terminal residues in *Entamoeba histolytica*
440 phosphoserine aminotransferase, corresponding to P12-A15 in human PSAT, yielded a mutant enzyme
441 that had reduced substrate (phosphoserine) affinity, as well as a 10-fold reduction in activity [62].
442 Additionally, a serine residue involved in analog substrate has been identified in *E. coli* [59],

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.

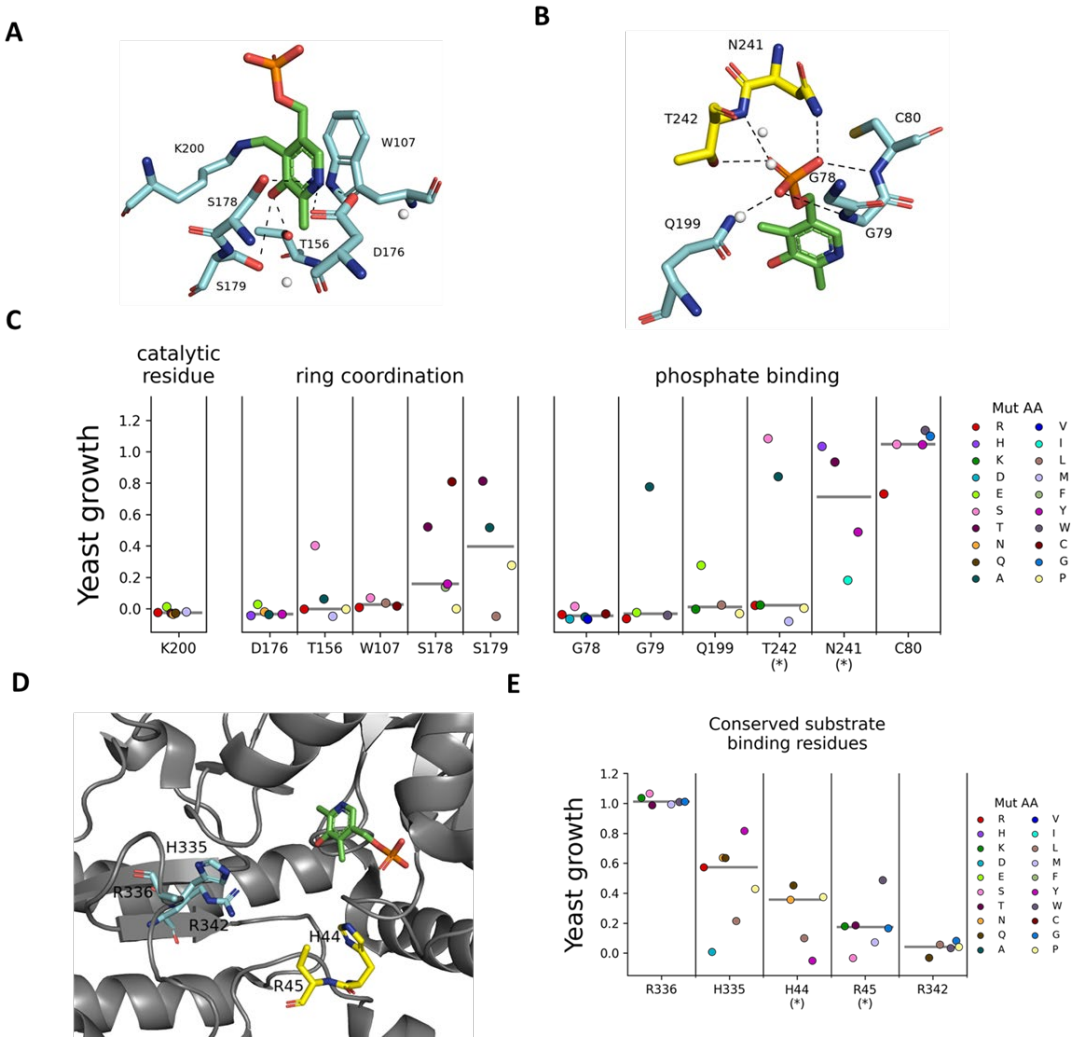


Fig. 5 Functional impact of amino acid substitutions at active site residues. Stick representations for residues in one subunit are colored in cyan, residues from the opposite subunit in the homodimer assembly are in yellow, and the PLP cofactor in green in all panels where the crystal structure of human 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. Each circle is 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) crystal structure of human PSAT. **E** The distribution of functional effects for these conserved substrate binding residues, with the same coloring scheme and annotation as in **C**.

460 **Mapping functional effects to clinical interpretations**

461 We next compared our results to clinical classifications for all corresponding missense variants in ClinVar.
462 We expected that substitutions derived from known pathogenic variants would produce strong
463 reductions in PSAT activity in our assay, while those derived from known benign variants would have
464 activities comparable to wild type or display only weak reductions in activity. Of the 90 amino acid
465 substitutions derived from variants currently in ClinVar (November 3, 2022, release), three (S179L, A99V,
466 and D100A) are from variants annotated as pathogenic/likely pathogenic, two (I123V and P295R) from
467 variants annotated as benign/likely-benign, and three (V149M, A234S, and R306C) from variants with
468 conflicting interpretations [3]. Since our previous study [39], while the small number of missense alleles
469 with definitive clinical significance calls remained largely unchanged, the number of missense VUS
470 increased nearly 12-fold (from 7 to 82). Consistent with expectation and our previous results [39], the
471 pathogenic/likely pathogenic substitutions all resulted in substantial functional impairment in our assay,
472 with normalized growth values of 82% or below. Similarly, the benign/likely-benign substitutions
473 demonstrated little impairment, with normalized growth values above 91% (Fig. 6).

474 Data from validated functional assays are potentially valuable as supporting evidence for variant
475 annotation according to the ACMG guidelines [36]. A challenge for rare diseases, such as serine
476 deficiency disorders, is that the limited number of well characterized pathogenic and benign variants
477 precludes the use of current methods, such as odds of pathogenicity [63], for deriving formal measures
478 of classification confidence. As an alternative approach for providing guidance for the use of our data in
479 clinical interpretation, we used a simple thresholding approach, similar that that used in our previous
480 study [39]. Guided by the assay scores of amino acid substitutions with definitive clinical significance
481 calls (Fig. 6), we classified as deleterious any substitutions with ≤82% growth in our assay, which is less
482 than or equal to the scores of substitutions derived from known pathogenic variants (Fig. 6). We also
483 defined a non-deleterious range for scores ≥91%, which is greater than or equal to the scores of
484 substitutions derived from known benign variants (Fig. 6). Because increases in enzyme activity above
485 wild type have not been associated with serine deficiency disorders, we included values greater than
486 wild type in the non-deleterious range. Finally, because there are no clinically annotated variants
487 associated with substitutions having scores between 82% and 91%, this range of the assay is of unknown
488 clinical significance, and we have labeled this range uncertain. Within these ranges of our assay, 534
489 (28%) of all amino acid substitutions are deleterious, 1,100 (57%) are non-deleterious, and 280 (15%)
490 are uncertain (Fig. 6). Of the 80 ClinVar missense variants currently annotated as VUS that were tested
491 in our assay, 22 result in substitutions that fall in our deleterious range, 41 in our non-deleterious range
492 and 17 in the uncertain range. Thus, our dataset provides functional information for a large number of
493 variants that lack clinical interpretation, including 77% of the missense VUS currently listed in ClinVar.

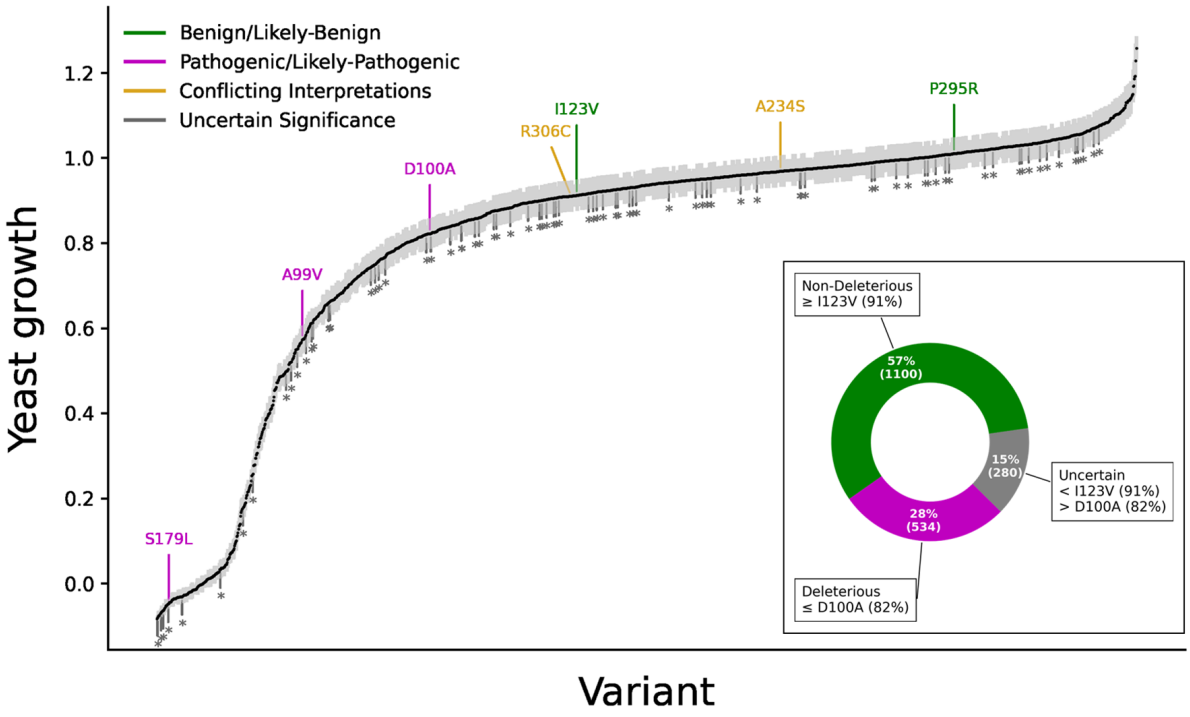
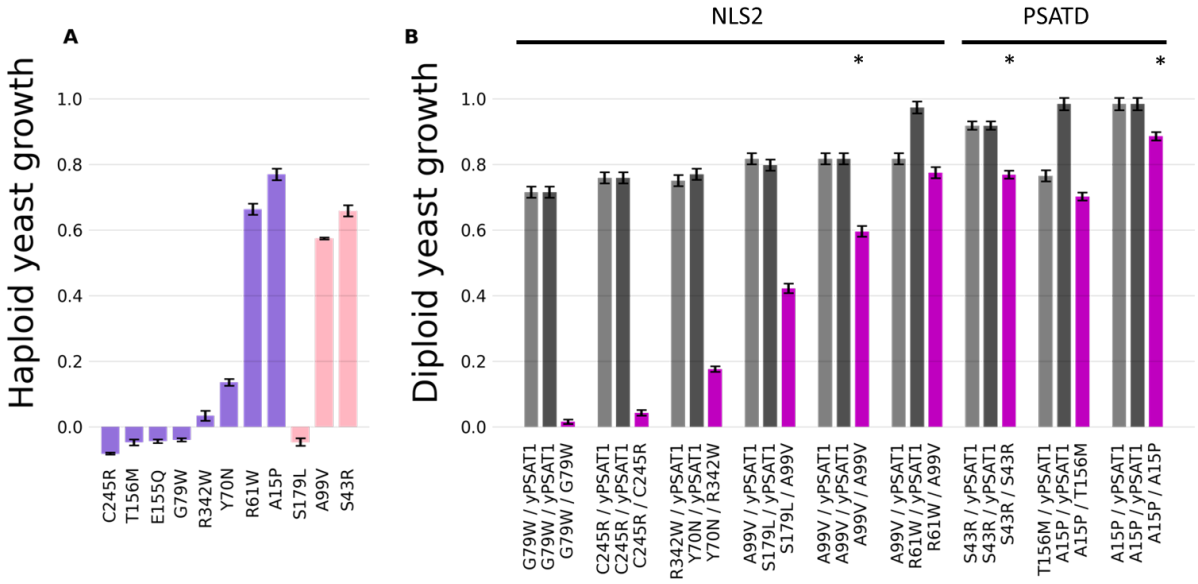


Fig. 6 Mutational scan of 1,914 PSAT1 unique SNV-accessible missense amino acid substitutions. Rank ordered (lowest to highest) normalized variant growth. Black dots represent the mean normalized growth estimate for each substitution, with light grey bars indicating standard errors. All current ClinVar 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.

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 [39], 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 (Additional File 1: Table S4). Since our previous study [39], six additional reports [28–30,64–66] have added descriptions of 20 new NLS2/PSATD patients and twelve unique genotypes to the disease literature (Additional File 1: Table S4). 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

518 the compound heterzygote A99V / S179L [39], were also included for comparision. Although the growth
519 of strains modelling some patient genotypes was close to that of strains modelling some carrier
520 genotypes, all modelled patient genotypes (Additional File 1: Table S5) displayed reduced growth relative
521 to their respective carrier parents and the homozygous wildtype (Fig. 7B). Consistent with our previous
522 study [39], we also observed good agreement between the degree of *PSAT1* functional impairment in
523 our diploid assay and disease severity (Fig. 7B). Diploids corresponding to genotypes from NLS2 patients
524 had normalized growth from 0% to 77%, and diploids corresponding to genotypes from PSATD patients
525 had a higher normalized growth range of 70% to 88%.



526

527 **Fig. 7** Modeling PSAT activity in patient genotypes. Both bar plots show the mean normalized growth estimate and standard
528 errors. **A** Haploid growth estimates for novel (purple) [28–30,64] and established [25,26] (pink) missense alleles in patient
529 amino acid substitutions. **B** Diploid yeast models of patient trios. Carrier parents are shown in light and dark grey, and the
530 offspring in magenta. Recurrent carrier genotypes are repeated for ease of reference relative to the offspring. The clinical
531 severity (NLS2 or PSATD) associated with patient genotypes is indicated. Asterisks denote offspring genotypes that have been
532 reported in multiple unrelated families.

533 Our assay results are also consistent with clinical stratification within the NLS2 and PSATD patient groups.
534 For PSATD, the A15P / A15P homozygous genotype is the least impaired in our assay (88% growth) and
535 is also the mildest form of PSATD reported to date. A recent report [30] described two unrelated patients
536 that, other than congenital or childhood ichthyosis, were developmentally normal and displayed
537 neuropathy onset at ages 16 and 17. For NLS2, the two genotypes with the lowest growth in our assay
538 (G79W / G79W and C245R / C245R) correspond to patients displaying severe NLS2 [28]. Three of the
539 remaining NLS2 genotypes modelled here (Y70N / R342W, A99V / S179L, A99V / R61W) display higher
540 growth in our assay and correspond to patients described as having moderate NLS expression [25,28,64].
541 Among these, Y70N / R342W was the most impaired in our assay (18% growth), and the patient had a
542 postnatal survival of 8 weeks [64]. In contrast, the individual harboring R61W / A99V (77% growth, least
543 impaired) had the longest NLS2-associated postnatal survival described to date (4 months) [28]. Patients
544 homozygous for the A99V genotype (60% growth) exhibit variable postnatal survival (ranging from 1 day

545 to 9 weeks) [25,28,32]. Together, these results highlight the potential value of our model organism assay
546 for variant interpretation in the context of diploid *PSAT1* genotypes.

547 **Predicting biallelic effects from individual allele measurements**

548 Experimentally assessing the level of PSAT activity associated with all allele-pairs would require
549 construction and assays of 1.83 million diploid yeast strains. As a more labor and cost-effective
550 alternative, we evaluated whether a relationship exists between individual haploid estimates and their
551 resulting diploid growth in combination. For homozygous diploids, a linear relationship was seen
552 between haploid and diploid growth estimates (Fig. 7A and Fig. 7B and Additional File 2: Fig. S4; R^2=0.98).
553 To extend this analysis to heterozygous genotypes, we fit a model to our experimental diploid dataset
554 of both patient and carrier genotypes (n = 23), to predict diploid growth as a linear combination of the
555 allele with the higher and the allele with the lower haploid growth values (Fig. 8A). This model assumed
556 a uniform degree of dominance of the higher growing allele over the lower growing allele and explained
557 97% of variance in the growth of the diploids. To compare how this model performed relative to
558 alternatives with simpler assumptions, we also generated models that assumed either complete
559 dominance of the higher-growing allele or an equal contribution of both alleles (Fig. 8B and Fig. 8C). Our
560 full model performed significantly better than either of the alternative models (Anova, Df=1; $p=<5.4x10^{-}$
561 ^{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).

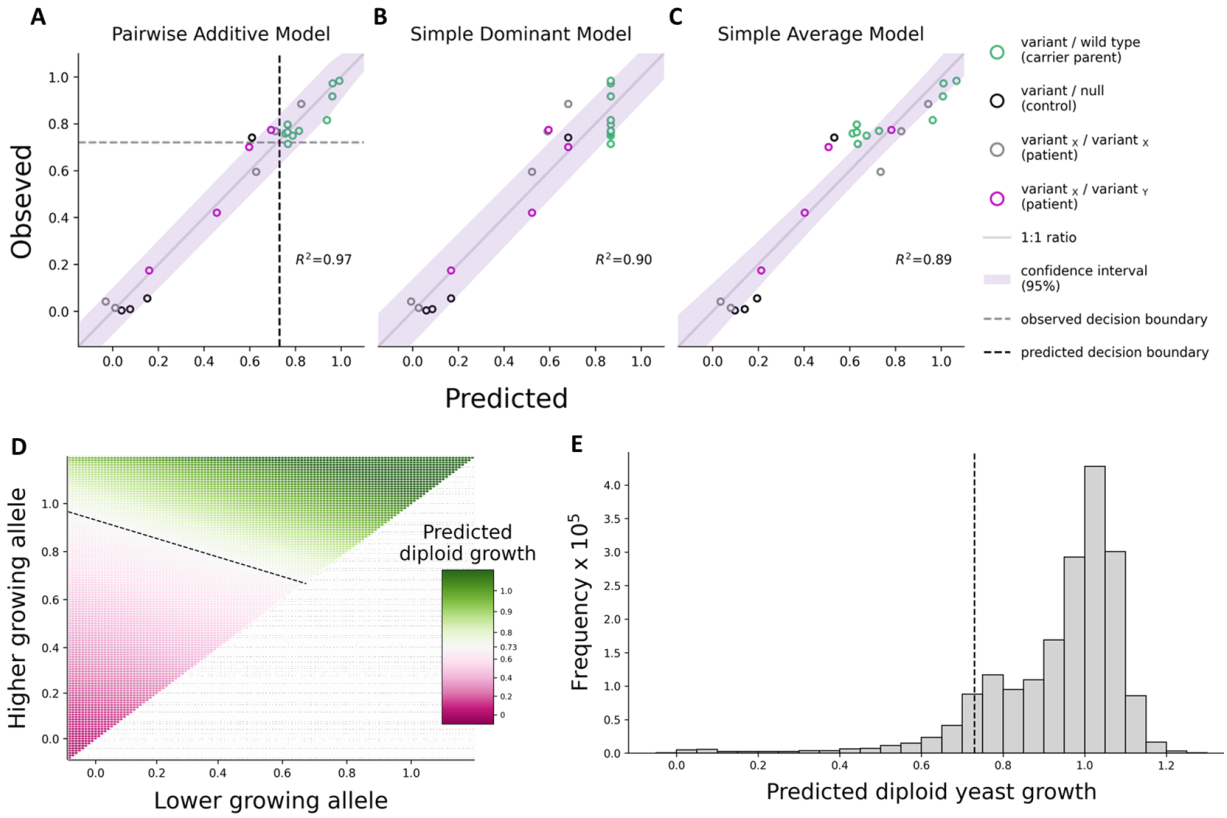


Fig. 8 Experimental and predicted biallelic growth values as clinical classifiers. Observed versus fitted values shown for three linear regression models (pairwise additive, **A**; simple dominant, **B;** and simple average, **C**) predicting diploid growth (allele pairs) as a function of haploid growth (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 (Fig. 8A). It is likely the haploid

584 estimates used to make predictions on diploid growth were more accurate than the experimental diploid
585 estimates as they were generated from a larger number of experimental replicates. These results suggest
586 that diploid growth values predicted from haploid measurements perform well in clinical classification
587 of human *PSAT1* genotypes.

588 On this basis, we computed the predicted diploid growth from all unique pairwise combinations (~1.8
589 million) of haploid estimates from 1,914 *yPSAT1* missense alleles, wild type (*yPSAT1*), and null (*ser1*Δ0)
590 using the pairwise-additive model (Fig. 8A). We then classified whether the resulting values fell above or
591 below the disease classification boundary of 73% growth. A large majority (89%) of genotypes fell above
592 the decision boundary threshold, suggesting these genotypes would not lead to NLS2/PSATD (Fig. 8D
593 and Fig. 8E). However, over 200,000 *yPSAT1* biallelic combinations displayed predicted diploid growth
594 below the classification threshold, consistent with these genotypes resulting in disease (Fig. 8E).

595 **Discussion**

596 Here, we present a comprehensive functional analysis of amino acid substitutions associated with SNVs
597 in human PSAT. Leveraging such data for clinical interpretation requires an understanding of the
598 relationship between enzyme function and clinical presentation. Our dataset shows clear stratification
599 of existing clinical variants, with benign variants exhibiting little to no functional impairment, and
600 pathogenic variants exhibiting substantial loss of activity. On this basis, we were able to use the small
601 number of variants with existing clinical interpretation to provide evidence for the likely clinical impact
602 of the large number of variants in similar ranges of the assay. As additional clinical information becomes
603 available through identification of new patients and additional research into the natural history of the
604 disease, the ability to make clinical inferences using our dataset will improve. In the three years since
605 our original study [39], additional patients were described in the disease literature. Not only was there
606 good agreement between the results of our assay and disease severity in these patients, but this
607 additional clinical information also allowed us to extend the range of our assay associated with clinical
608 outcomes.

609 Given that the therapeutic potential for this medically actionable disorder is highest before patients
610 become symptomatic, clinical sequencing coupled to informative variant interpretation can be a
611 powerful diagnostic tool. Sequencing based diagnostics are becoming more widely applied in newborns,
612 prenatal care [67], and as a part of clinical research efforts such as the Undiagnosed Disease Program
613 [68,69] or Deciphering Developmental Disorders [70]. The most recent AMCG guidelines [71]
614 recommend exome or genome sequencing as a first or second tier option for patients displaying early
615 neonatal (<1 year) congenital abnormalities and for patients exhibiting developmental delay or
616 intellectual disability before the age of 18. In contrast to frameshift and stop-gain mutations, when novel
617 missense variations are identified, even in disease-causing genes, predicting the likely effect on human
618 health is challenging. Relating genotype to phenotype is further complicated by the fact that autosomal
619 recessive diseases are a function of the allele combination in the context of a diploid genome. We can
620 experimentally assay allele combinations in our diploid assay and successfully stratify clinical genotypes.

621 Furthermore, we were able to predict the effects of allele combinations using a computational model
622 that utilizes combinations of haploid measurements. As a result, we were able to leverage 1,914 yeast
623 measurements to make predictions on 1.8 million potential human genotypes. Given ongoing genome
624 sequencing efforts like the All of US research program and the UK biobank [72], studies that provide
625 information both about individual alleles and the functional impact of allele pairs will play an increasingly
626 important role in realizing the promise of precision medicine.

627 **Conclusion**

628 The strong agreement between experimentally measured results from our assay and known features of
629 the protein structure together with (albeit limited) clinical data support its use as a well-validated assay
630 of human PSAT function. The corresponding dataset provides meaningful functional information for a
631 large proportion (88%) of all unique SNV-accessible missense substitutions across the length of the
632 protein. Furthermore, our computational model extends the use of the dataset to predict the functional
633 impact of ~1.8 million allele combinations corresponding to potential human genotypes. As such, this
634 approach leverages a relatively small amount of clinical data to classify large numbers of variants making
635 it especially powerful for rare diseases.

636 **References**

637 1. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint
638 spectrum quantified from variation in 141,456 humans. Nature. Nature Research; 2020;581:434–43.

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

642 3. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: Improving access to
643 variant interpretations and supporting evidence. Nucleic Acids Res. Oxford University Press;
644 2018;46:D1062–7.

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

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

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

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

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

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

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

660 11. El-Hattab AW. Serine biosynthesis and transport defects. Mol Genet Metab. Academic Press Inc.;
661 2016. p. 153–9.

662 12. de Koning TJ, Klomp LWJ, Oppen ACCV, Beemer PFA, Dorland L, Berg IETV den, et al. Prenatal and
663 early postnatal treatment in 3-phosphoglycerate-dehydrogenase deficiency. Lancet. Elsevier B.V.;
664 2004;364:2221–2.

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

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

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

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

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

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

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

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

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

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

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

693 24. Shaheen R, Rahbeeni Z, Alhashem A, Faqeih E, Zhao Q, Xiong Y, et al. Neu-laxova syndrome, an inborn
694 error of serine metabolism, is caused by mutations in PHGDH. Am J Hum Genet. Cell Press; 2014;94:898–
695 904.

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

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

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

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

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

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

713 31. de Koning TJ. Amino acid synthesis deficiencies. J Inherit Metab Dis. Springer Netherlands; 2017. p.
714 609–20.

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

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

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

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

724 36. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the
725 interpretation of sequence variants: A joint consensus recommendation of the American College of
726 Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine.
727 Nature Publishing Group; 2015;17:405–24.

728 37. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants.
729 Nat Protoc. Nature Publishing Group; 2016. p. 1782–7.

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

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

735 40. Youn BAEK J, Youn JUN D, Taub D, Ho KIM Y. Characterization of human phosphoserine
736 aminotransferase involved in the phosphorylated pathway of L-serine biosynthesis. Biochem. J. 2003.

737 41. Murtas G, Marcone GL, Sacchi S, Pollegioni L. L-serine synthesis via the phosphorylated pathway in
738 humans. Cellular and Molecular Life Sciences. Springer Science and Business Media Deutschland GmbH;
739 2020. p. 5131–48.

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

743 43. Winston F, Dollard C, Ricupero-Hovasse SL. Construction of a set of convenient saccharomyces
744 cerevisiae strains that are isogenic to S288C. Yeast. Chichester, UK: John Wiley & Sons, Ltd; 1995;11:53–
745 5.

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

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

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

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

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

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

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

760 51. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein
761 Structure Database: massively expanding the structural coverage of protein-sequence space with high-
762 accuracy models. Nucleic Acids Res. 2022;50:D439–44.

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

765 53. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved
766 methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res.
767 2016;44:W344-50.

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

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

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

775 57. Liang J, Han Q, Tan Y, Ding H, Li J. Current advances on structure-function relationships of pyridoxal
776 5'-phosphate-dependent enzymes. Front Mol Biosci. Frontiers Media S.A.; 2019.

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

779 59. Hester G, Stark W, Moser M, È rg Kallen Zora Markovic Â-Housley J, Jansonius JN. Crystal Structure
780 of Phosphoserine Aminotransferase from Escherichia coli at 2.3 A Ê Resolution: Comparison of the
781 Unligated Enzyme and a Complex with a a a-Methyl-L-Glutamate. J. Mol. Biol. 1999.

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

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

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

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

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

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

66. 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.

67. 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.

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

69. 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

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

71. 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.

72. 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.

Supplementary information

Additional File 1: Table S1. *S. cerevisiae* strains used in the haploid assay. **Table S2.** *S. cerevisiae* strains used in the diploid assay. **Table S3.** Normalized *yPSAT1* haploid growth scores and standard error estimates. **Table S4.** Disease literature review of patients diagnosed with PSAT1-related serine biosynthesis defects. **Table S5.** Normalized *yPSAT1* diploid growth scores and standard error estimates. **Table S6.** Human PSAT structure, conservation, and median haploid yeast growth data per amino acid position.

Additional File 2: Further description of the MinIon sequencing pipeline. **Fig S1.** Agreement between yPSAT1 variant codons identified by Oxford Nanopore versus Illumina sequencing. **Fig S2.** Comparison of haploid yeast growth scores and ConSurf conservation scores for each corresponding PSAT amino acid position **Fig S3.** Predicted subunit structure for human PSAT. **Fig. S4.** Comparison of experimentally measured haploid and diploid growth scores for PSAT variants that are in homozygous patient (NLS2 or PSATD) genotypes**.**

Additional File 3: Growth normalization script**.**

Additional File 4: MinIon variant calling script**.**

835 **Abbreviations**

836 3-PHP: 3-phosphohydroxypyruvate; ACMG: American College of Medical Genetics; cDNA: Protein coding
837 sequence; Df: Degree of freedom; DNA: deoxyribonucleic acid; gnomAD: Genome Aggregation Database;
838 IEM: Inborn errors of metabolism; MAE: Mean absolute error; MATa: Yeast mating type a; MATα: Yeast
839 mating type alpha; NA: Not applicable; NLS: Neu-Laxova syndrome; NLS2: Neu-Laxova syndrome 2; ORF:
840 Open reading frame; PCR: Polymerase chain reaction; PGDH: Phosphoglycerate dehydrogenase; PLP:
841 Pyridoxal Phosphate; PSAT: Phosphoserine aminotransferase; PSATD: Phosphoserine aminotransferase
842 deficiency ; PSP: Phosphoserine phosphatase; R^2: Coefficient of determination ; RMSE: Root mean
843 squared error; SD: Minimal medium; SNV: Single nucleotide variant; VUS: Variants of uncertain
844 significance ; YPD: Rich medium; yPSAT1: yeast codon-optimized version of the human phosphoserine
845 aminotransferase isoform 1 cDNA sequence

846 ## Declarations

847 **Ethics approval and consent to participate**

848 Not applicable. No human or animal subjects were used in this work.

849 **Availability of data and materials**

850 Provided upon manuscript acceptance.

851 **Competing interests**

852 AMD is a scientific advisor to and has a financial interest in Fenologica Biosciences, Inc., a company that
853 develops instrumentation and analytical tools for the analysis of microbial colonies. The remaining
854 authors declare that they have no competing interests.

855 **Funding**

856 This work was funded in part by NIH/ NIGMS award R01 GM134274 to AMD and R35 GM142773 to RNM.
857 KO and JNK acknowledge support from the National Science Foundation AI Institute in Dynamic Systems
858 (grant number 2112085).

859 **Authors' contributions**

860 AMD conceived the project. MJX constructed and assayed the variant library. MJX and MST performed
861 the DNA sequencing. AMD supervised all experimental work. GAC, MT, and KO developed computational
862 pipelines. JNK supervised software development. GAC and MJX performed all data analysis with
863 supervision by AMD. MJX performed all protein structure analysis with supervision from RNM. GAC and
864 MJX performed all comparisons to published clinical data with oversight by AWE-H. MJX, GAC, and AMD
865 wrote the manuscript and incorporated comments from AWE-H, RNM and JNK. All authors approved the
866 manuscript.

867 **Acknowledgements**

868 We thank Douglas Fowler, Maitreya Dunham, Russell Lo, and Fowzan Alkuraya for helpful discussions.

869