Deep Mutational Scanning in Disease-related Genes with Saturation Mutagenesis-Reinforced Functional Assays (SMuRF)

Kaiyue Ma[1,*], Kenneth K. Ng[1], Shushu Huang[1], Nicole J. Lake[1], Jenny Xu[2], Angela Lek[1,3], Lin Ge[1,4], Keryn G. Woodman[1], Katherine E. Koczwara[1], Vincent Ho[1], Christine L. O'Connor[1], Soumya Joseph[5], Melinda A. Brindley[6,7], Kevin P. Campbell[5,7] & Monkol Lek[1,7,8,*].

[1]Department of Genetics, Yale School of Medicine, New Haven, CT, USA
[2]Yale University, New Haven, CT, USA
[3]Muscular Dystrophy Association, Chicago, IL, USA
[4]Department of Neurology, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China
[5]Howard Hughes Medical Institute, Senator Paul D. Wellstone Muscular Dystrophy Specialized Research Center, Department of Molecular Physiology and Biophysics and Department of Neurology, Roy J. and Lucille A. Carver College of Medicine, The University of Iowa, Iowa City, IA, USA
[6]Department of Infectious Diseases, Department of Population Health, University of Georgia, Athens, GA, USA.
[7]Senior Authors
[8]Lead Contact
[*]Correspondence: kaiyue.ma@yale.edu (K.M.), monkol.lek@yale.edu (M.L.)

Abstract

Interpretation of disease-causing genetic variants remains a challenge in human genetics. Current costs and complexity of deep mutational scanning methods hamper crowd-sourcing approaches toward genome-wide resolution of variants in disease-related genes. Our framework, Saturation Mutagenesis-Reinforced Functional assays (SMuRF), addresses these issues by offering simple and cost-effective saturation mutagenesis, as well as streamlining functional assays to enhance the interpretation of unresolved variants. Applying SMuRF to neuromuscular disease genes *FKRP* and *LARGE1*, we generated functional scores for over 99.8% of all possible coding single nucleotide variants and resolved 310 clinically reported variants of uncertain significance with high confidence, enhancing clinical variant interpretation in dystroglycanopathies. SMuRF also demonstrates utility in predicting disease severity, resolving critical structural regions, and providing training datasets for the development of computational predictors. Our approach opens new directions for enabling variant-to-function insights for disease genes in a manner that is broadly useful for crowd-sourcing implementation across standard research laboratories.

Main

The recent advancement in Next-generation sequencing (NGS) and population studies[1–4], and the establishment of large biobanks have improved disease-associated variant detection and interpretation[5–7]. However, for many genetic disease patients, it remains challenging to identify the specific disease-causing variants, which hinders subsequent treatment development and patients' disease management and family planning[8,9]. The emergence of gene therapies has further heightened the importance of finding diagnoses for genetic disease patients as an initial and crucial prerequisite towards clinical trial readiness. The variants discovered in patients that

are difficult to interpret are classified as variants of unknown significance (VUS), which is often a result of insufficient clinical evidence to determine pathogenicity[10]. The rate of VUS observed in disease-associated genes is typically higher in individuals of non-European ancestry due to limited diversity in biomedical databases[11].

To provide an additional line of evidence for the interpretation of variants, deep mutational scanning (DMS) was proposed to unbiasedly generate functional scores for all possible variants[12]. DMS employs pooled model cells that carry large amounts of variants as "patients in flasks" and characterize the variants with appropriate high-throughput functional assays[13], which has been considered a prominent strategy in the creation of an Atlas of Variant Effects[14]. However, despite its promise, the high cost and unpredictable research outcomes discourage many researchers, especially those in the modestly funded rare disease field, from utilizing DMS.

In saturation mutagenesis, well-established methods include the insertion of variant-carrying tiles[15], saturation genome editing (multiplex homology-directed repair)[16],reversibly-terminated inosine mutagenesis[17], and recently emerged saturation prime editing[18]. However, these methods often come with limitations like intensive labor requirements, high expenses, disparate variant representation, and limited spanning regions. Hence, there is a need for a simple and cost-effective saturation mutagenesis method that provides comprehensive and unbiased coverage. Programmed Allelic Series (PALS) was previously developed as a saturation mutagenesis method with relatively simple steps and unbiased coverage[19]. PALS had the potential to reduce costs by using minimized-length oligos instead of long DNA fragments to introduce variants. However, its reliance on special reagents has limited its widespread adoption in DMS studies. A previous optimization of PALS removed the need for special reagents but transformed the method from requiring only a single-tube reaction to necessitating independent reactions for each targeted codon, significantly increasing the labor involved[20].

Besides saturation mutagenesis, there are challenges in the development of robust and accurate functional assays. Typical functional assays used in DMS to enrich variants include growth assays and flow cytometry assays[21]. Growth assays are primarily suitable for genes significantly affecting cell viability or growth rates and may not apply to many genes. In contrast, flow cytometry assays convert gene function-related biological signals into quantifiable fluorescence. However, they are relatively costly due to cell sorter reliance, particularly when saturation mutagenesis efficiency is low, and more cells require sorting. Second, precisely controlling protein expression levels poses another challenge, as overexpression of the targeted gene may rescue variant pathogenic effects[22], diminishing assay sensitivity for variant interpretation. Lastly, barcoding is a common practice in DMS studies to facilitate variant detection using short-read NGS. However, this process typically necessitates an additional round of NGS to assign the barcodes to the variants, incurring extra expenses[23].

Addressing these challenges is crucial for applying DMS to disease genes. Dystroglycanopathies are a set of rare autosomal recessive diseases with clinical heterogeneity ranging from brain malformation in Walker-Warburg syndrome (WWS) to milder muscular symptoms in Limb-Girdle Muscular Dystrophies (LGMDs)[24]. Pathogenic variants in *DAG1*, the gene that encodes alpha-dystroglycan (α-DG), and genes encoding enzymes involved in α-DG glycosylation disrupt the binding between α-DG and extracellular matrix ligands, which compromises the

muscle cell integrity and leads to dystroglycanopathies[25]. The most severe dystroglycanopathy cases can lead to miscarriage and neonatal deaths, highlighting the critical need for a better understanding of the clinical significance of variants in genetic testing[26–28].

In this study, DMS was modularized into two components: saturation mutagenesis and functional assays. Saturation Mutagenesis-Reinforced Functional Assays (SMuRF) was developed to address challenges that were identified for each of these components. We built upon the core concept of PALS but achieved saturation mutagenesis without the need for special reagents, equipment, or labor-intensive steps, resulting in a method named Programmed Allelic Series with Common procedures (PALS-C), which is accessible to most molecular biology laboratories. We demonstrated the versatility of SMuRF by applying it to two key enzyme genes, *FKRP* and *LARGE1*, involved in the glycosylation of α-DG. For this implementation of SMuRF, we employed lentiviral delivery to introduce variants into cells for the flow cytometric functional assay, achieving high variant coverage. In addition, we utilized the weak UbC promoter[29], and streamlined the workflow to control gene expression levels, ensuring precise variant assessment. Lastly, we implemented a block-by-block strategy throughout the saturation mutagenesis and functional assay, eliminating the need for barcoding and significantly reducing costs.

Results

Establishing the functional assay in SMuRF

SMuRF was developed with the aim of creating a universal DMS workflow that is adaptable for various genes by employing appropriate assays. Initially, the functional assay was established to characterize all possible coding single nucleotide variants (SNVs) in genes related to dystroglycanopathies. Most variants of α-DG glycosylation enzymes, including FKRP and LARGE1, lack clinical reports and those reported remain poorly interpreted (Extended Data Fig. 1), with many unique to the families[30]. FKRP adds the second ribitol-5-phosphate (Rbo5P) to the Rbo5P tandem[31] while LARGE1 is responsible for adding the repeated disaccharide units of matriglycan[32] (Fig. 1a). New drugs[33], gene therapies[34,35] and cell therapies[36] are being actively developed for dystroglycanopathies related to FKRP or LARGE1, emphasizing the need for improved variant interpretation to facilitate the patient enrollment in these gene-specific trials.

To develop a functional assay suitable for DMS, we initially identified α-DG hypoglycosylation as the molecular phenotype associated with dystroglycanopathies and the focal point for assessment. The IIH6C4 antibody, extensively used in α-DG-related research and clinical diagnoses, specifically binds to the matriglycan chain of glycosylated α-DG Core M3, enabling the quantification of α-DG glycosylation levels[37–50]. Fibroblasts from patients with dystroglycanopathies have been characterized using the IIH6C4 antibody in a fluorescence flow cytometry (FFC) assay[51]. The human haploid cell line, HAP1, has become a widely utilized platform in α-DG-related research, covering various areas such as dystroglycanopathy gene discovery[52], enzymatic functions[42], and α-DG binding properties and functions[53]. Additionally, previous studies have established the compatibility of HAP1 cells with FFC[54]. We modified the IIH6C4 FFC assay originally designed for fibroblasts to be suitable for HAP1 cells while improving its sensitivity to differentiate between variants (Extended Data Fig. 3a,b). SMuRF was

developed based on this IIH6C4 FFC assay and applied to generate functional scores for all possible coding SNVs in *FKRP* and *LARGE1* (Fig. 1b).

Streamlining the saturation mutagenesis and FACS assay in SMuRF

To establish the engineered cell line platforms for the saturation mutagenesis, the endogenous gene of interest (GOI), specifically *FKRP* or *LARGE1*, was knocked out to make *GOI*-KO HAP1 lines, which lack the endogenous *GOI* function (Extended Data Fig. 4a,b). *DAG1* overexpression was achieved by Lenti-*DAG1* transduction (Extended Data Fig. 4c-e). Subsequent experiments were performed using monoclonal *GOI*-KO Lenti-*DAG1* HAP1 lines. Since most known dystroglycanopathy cases are caused by missense variants[55], we decided to do saturation mutagenesis for all possible coding SNVs. We first constructed the plasmids carrying the wild-type (WT) GOIs. To better control the expression level (Supplementary Fig. 1), we employed the weak promoter UbC for GOI expression[29], creating Lenti-UbC-*FKRP*-EF1α-*BSD* and Lenti-UbC-*LARGE1*-EF1α-*BSD* (Fig. 2a).

Next, we performed PALS-C, a two-way extension cloning method, to introduce all possible coding SNVs using the WT plasmids as templates (Fig. 2b). The variants were introduced initially as 64-nt reverse PCR primers. The short length of the oligos allowed for a low cost in synthesizing the oligo pool. We adopted a block-by-block strategy where we divided the GOI variants into multiple non-overlapping blocks (6 for *FKRP* and 10 for *LARGE1*). For each GOI, PALS-C initiates with one single pool of oligos and eventually generates an isolated lentiviral plasmid pool for each block. PALS-C steps are massively multiplexed: Step1 requires only a single-tube reaction; the following steps can be done in a single-tube reaction for each block (Methods and Supplementary Method 4). All downstream experiments following PALS-C, up to the NGS, were conducted individually for each block. This block-by-block strategy allowed us to employ short-read NGS to examine variant enrichment while avoiding the requirement of an additional NGS to assign barcodes to variants spanning the entire CDS and the expenses associated with it[56].

Variant representation in the lentiviral plasmid pools generated by PALS-C was evaluated using a shallow NGS service, according to which, more than 99.6% of all possible SNVs were represented (Extended Data Fig. 5). The plasmid pools were packaged into lentiviral particles, which were subsequently delivered into the platform cells through transduction. Once transduced cells are sufficiently expanded following drug selection, they can be sorted using fluorescence-activated cell sorting (FACS) to isolate cells with either high or low glycosylation levels. Staining conditions and gating parameters in FACS were optimized with mini-libraries (Extended Data Fig. 6a,b). α-DG glycosylation level was quantified by IIH6C4-FITC signal (Fig. 2c and Extended Data Fig. 6c,d). The FACS events of each group achieved a minimum of ~1000 × coverage.

Genomic DNA from each group of each block was used to build the multiplexed sequencing library using a 3-round PCR strategy (Extended Data Fig. 7a). Raw NGS datasets were analyzed with our customized analytical pipeline “Gargamel-Azrael” to generate SMuRF scores for all variants (Extended Data Fig. 7b). Essentially, SMuRF score is the normalized relative

enrichment of a variant in the FACS groups (Methods). High SMuRF scores indicate normal functionality of variants in α-DG glycosylation while low scores indicate deleterious effects.

SMuRF recapitulated and expanded the knowledge gained from variant databases

SMuRF scores were generated for ~99.9% of all possible coding SNVs of *FKRP* (4450/4455) and 100% of *LARGE1* (6804/6804) (Supplementary Table 4, 5). Two transduction replicates were performed for *FKRP* to confirm the reproducibility of the workflow (Extended Data Fig. 8). SMuRF scores align with the anticipated patterns of different variant types (Fig. 3a,b). The synonymous variants display scores that closely approximate the WT score (SMuRF=0), exhibiting a narrow range of values. *FKRP* synonymous variants have a median of 0.17, with a 95% confidence interval (CI) of 0.14~0.21, while *LARGE1* synonymous variants have a median of 0.20 (95% CI: 0.16~0.24). The nonsense variants consistently exhibit low SMuRF scores, with sparse outliers observed. *FKRP* nonsense variants have a median of -2.27 (95% CI: -2.42~-2.17), while *LARGE1* nonsense variants have a median of -2.02 (95% CI: -2.11~-1.93). Two noteworthy outliers among the nonsense variants are *FKRP* c.1477G>T (p.Gly493Ter) (SMuRF=-0.42) and *LARGE1* c.2257G>T (p.Glu753Ter) (SMuRF=0.04). These two are the nonsense variants positioned closest possible to their respective canonical stop codons. The relatively high SMuRF scores of these two variants suggest that their impact on the enzymatic function is negligible in the context of the CDS constructs. Furthermore, since both variants are in the last exon of their respective transcripts, it is also unlikely for them to be substantially influenced by nonsense-mediated decay (NMD) *in vivo*[57].

Interestingly, the start-loss variants exhibit markedly low SMuRF scores, significantly lower than those observed in most nonsense variants (p-value=9.2e-5, *FKRP*; 0.0034, *LARGE1*). *FKRP* start-loss variants have a median of -3.09 (95% CI: -3.47~-2.93), while *LARGE1* start-loss variants have a median of -2.55 (95% CI: -2.93~-2.25). This observation indicates that, at least in the context of the *FKRP* and *LARGE1* SMuRF CDS constructs, there is a lack of effective genetic compensation to counter the start-loss variants, such as functional downstream alternative start codons[58]. The homozygous start-loss variant *FKRP* c.1A>G (SMuRF=-3.31) has been reported to be associated with WWS, the most severe FKRP-related disorder. This variant has been documented in two cases, with one resulting in the unfortunate death of a child at the age of 6 days and the other leading to a terminated pregnancy[28]. Therefore, the start-loss variants may warrant increased attention in genetic testing protocols.

Allele frequency (AF) refers to the relative frequency of a genetic variant at a specific chromosomal locus within a population. When compared to AF data in The Genome Aggregation Database (gnomAD, v4.0.0)[6,59], SMuRF scores aligned with the selection against pathogenic variants: low allele frequency variants (AF<1.5e-05) exhibited a wide range of functional scores, while variants with higher frequency showed functional scores converging towards the WT score as common variants are depleted of pathogenic variants (Fig. 3c,d). As one of the largest population databases, gnomAD v4 currently provides reports for only 1250 *FKRP* coding SNVs (28.1%) and 1374 *LARGE1* coding SNVs (20.2%). The ability of SMuRF scores to recapitulate the patterns observed in gnomAD makes them a significant addition to gnomAD.

The α-DG glycoepitope, as well as the enzymes involved in its glycosylation, are largely conserved within Metazoa[60]. Evolutionary conservation scores, such as PhyloP scores, indicate the degree of conservation of a variant derived from multiple sequence alignments across species[61]. When compared with PhyloP scores calculated from 100 vertebrates, SMuRF demonstrated the evolutionary tolerance of relatively harmless variants and the selection against damaging variants in both *FKRP* and *LARGE1* (Supplementary Fig. 2), with a tendency for missense variants to be more disruptive at the more conserved sites (rho=-0.38, *FKRP*; -0.21, *LARGE1*).

SMuRF is sensitive and accurate in predicting variant pathogenicity and severity

ClinVar is a public archive of reports of human variants[62], where the variants were classified according to clinical supporting evidence into different categories including: Benign(B), Benign/likely benign (B/LB), Likely benign(LB), Likely pathogenic (LP), Pathogenic/likely pathogenic (P/LP), Pathogenic(P), and Variants of Uncertain Significance (VUS). SMuRF scores correlate well with clinical classification in ClinVar (Fig. 4a,b and Extended Data Fig. 9a,b). Furthermore, dystroglycanopathies encompass a spectrum of diseases with varying severity, including severe cases like WWS and muscle-eye-brain disease (MEB), intermediate cases like congenital muscular dystrophies (CMD), and relatively mild cases like LGMDR9 (LGMD2I)[24,28]. We wanted to examine whether SMuRF scores could be used to predict the severity of a variant. We employed a naive additive model where the functional scores of the variants on both alleles were summed to calculate the biallelic functional score. We aggregated data from 8 well-curated cohorts and compared them with SMuRF scores (Supplementary Table 6,7)[28,47,55,63–67]. The functional scores of the variants associated with mild cases were significantly higher compared to those of the intermediate and severe cases, as expected (Fig. 4c). Additionally, SMuRF scores showed a correlation with the reported disease onset age (Fig. 4d), where high-function variants were associated with later onset (rho=0.61; 0.48, male; 0.77, female). These analyses indicate the potential utility of the SMuRF scores for improving variant interpretation, where confidence scores were also introduced to account for technical confounders and enhance credibility (Supplementary Table 1,2 and Supplementary Fig. 9,10).

In addition to directly assisting in variant classification, SMuRF scores can also be used to validate and improve computational predictors. Computational prediction is currently an active area of research with a wealth of methods recently developed, including CADD[68], metaSVM[69], REVEL[70], MVP[71], EVE[72], MutScore[73],PrimateAI-3D[74], ESM1b[75], MAVERICK[76], and AlphaMissense[77,78]. However, these methods often perform differently depending on the genetic context. To compare SMuRF with these computational predictors, we assessed the receiver operating characteristic (ROC) curves for all methods using the P, P/LP, and LP variants in ClinVar as true positives (Fig. 4e,f). A higher Area Under Curve (AUC) value indicates better discriminatory ability in classifying pathogenic variants. SMuRF outperforms all computational methods for *LARGE1* (AUC=0.87). For *FKRP*, two predictors, REVEL (AUC=0.79) and EVE (AUC=0.78) exhibit comparable performance to SMuRF (AUC=0.78).

We checked the correlation between the predictors and SMuRF. Most predictors assigned higher scores to pathogenic variants, hence are negatively correlated with SMuRF scores (except ESM1b). Among all the predictors examined, AlphaMissense has the strongest correlation with

SMuRF (rho=-0.65, *FKRP*; -0.45, *LARGE1*) (Fig. 4g-i and Supplementary Table 3). EVE scores are derived from an evolutionary model trained on sequences from over 140k species[72], which also shows a relatively strong correlation with SMuRF (rho=-0.615, *FKRP*; -0.412, *LARGE1*), but it is slightly weaker when compared to PrimateAI-3D, a more recent predictor trained on common primate variants (rho=-0.625, *FKRP*; -0.418, *LARGE1*) (Extended Data Fig. 9c-f)[74]. REVEL, among the predictors, demonstrates the best performance according to the ROC curves and also exhibits a relatively good correlation with SMuRF (rho=-0.58, *FKRP*; -0.39, *LARGE1*) (Extended Data Fig. 9g,h). Taken together, these findings demonstrate the potential of SMuRF scores to enhance variant interpretation, both as a standalone line of evidence and in combination with computational predictors.

SMuRF highlighted the critical structural regions

The currently known disease-related mutations in *FKRP* and *LARGE1* are distributed throughout their entire sequences (Supplementary Fig. 5), and only limited critical structural sites have been identified and associated with specific disease mechanisms. SMuRF can contribute to highlighting critical structural regions in the enzymes. The protein structures of both FKRP and LARGE1 have been previously studied. FKRP is known to have a stem domain (p.1-288) and a catalytic domain (p.289-495)[79]. SMuRF scores revealed that missense variants in the catalytic domain are generally more disruptive than those in the stem domain (p-values<2.22e-16) (Fig. 5a). Furthermore, it has been reported that a zinc finger loop within the catalytic domain plays a crucial role in FKRP enzymatic function[79]. SMuRF analysis demonstrated that missense variants in the zinc finger loop exhibit greater disruption compared to variants in the remaining regions of the catalytic domain (p-value=0.0016). The observed differences in the domains are only significant in the case of missense variants, and they are not driven by technical artifacts such as block differences and positional effects as there are no significant differences observed among synonymous variants (p-values>0.1) (Fig. 5b).

LARGE1 has two catalytic domains: a xylose transferase (XylT) domain (p.138-413) and a glucuronate transferase (GlcAT) domain (p.414-756)[80]. They are each responsible for adding one unit of the polysaccharide matriglycan chain, which consists of alternating xylose and glucuronate units. SMuRF revealed that the missense variants in both catalytic domains tend to be significantly more disruptive than the variants in the N-terminal domain (p-values<2.22e-16) (Fig. 5c). Interestingly, SMuRF also showed that the missense variants in the XylT domain tend to be more disruptive than those in the GlcAT domain (p-value<2.22e-16). Again, the differences in SMuRF scores between domains were not observed in synonymous variants (p-values>0.05) (Fig. 5d). A previous IIH6C4 western blot experiment revealed a similar observation, demonstrating that mutations deactivating the GlcAT domain, but not the XylT domain, can generate a faint band indicative of glycosylated matriglycan[80]. Moreover, AlphaMissense also presents the same effects, where missense variants in the XylT domain tend to be more disruptive (Supplementary Fig. 16f and Supplementary Discussion 6).

We further mapped the SMuRF scores of SNV-generated single amino acid substitutions (SNV-SAASs) onto the 3D structures of the enzymes (Fig. 5e,f) (FKRP: PDB 6KAM; LARGE1: PDB 7UI7), thereby highlighting the critical regions that are susceptible to missense disruptions. SMuRF confirmed the functional importance of p.Cys318 in FKRP (log2 mean missense =-2.15),

which is required for Zn2+ binding in the zinc finger loop[81]. A p.Cys318Tyr variant (SMuRF=-2.12) has been reported to be associated with WWS[66]. SMuRF also highlighted the functional importance of p.Phe473 in FKRP (log2 mean missense =-1.78), which is located in a small hydrophobic pocket essential for CDP-ribitol substrate binding within the catalytic domain[81].

Three important amino acids in the FKRP stem domain were labeled on the 3D structure as well: p.Tyr88 (log2 mean missense =-2.06) and p.Ser221 (log2 mean missense =-0.59), which are both situated at the subunit-subunit interface involved in FKRP tetramerization *in vivo* (Supplementary Fig. 15), and p.Leu276 (log2 mean missense =0.35), which interacts with the catalytic domain[79]. p.Tyr88Phe is likely associated with disease[82], and has a low SMuRF score (-3.52). p.Ser221Arg was associated with CMD-MR (MR: mental retardation)[83]. All three p.Ser221Arg SNVs have low SMuRF scores (c.661A>C: -2.13; c.663C>A: -2.21; c.663C>G: -2.18). Moreover, c.663C>A was examined in the mini-library screen and presented low function (Extended Data Fig. 6a). p.Leu276Ile is a founder mutation in the European population[84], which is commonly associated with milder symptoms[85]. Interestingly, it has a relatively higher SMuRF score (-0.57) and performed more similarly to the benign variants rather than other pathogenic variants in the mini-library screen (Extended Data Fig. 6a). In addition, SMuRF highlighted the importance of p.Asp242 (log2 mean missense =-2.20) and p.Asp244 (log2 mean missense =-1.96) in LARGE1, which are crucial for XylT activity, as well as p.Asp563 (log2 mean missense =-0.48) and p.Asp565 (log2 mean missense =-0.55), which are required for GlcAT activity[80]. Variants affecting different enzyme domains may require distinct treatment approaches[86]. SMuRF can assist in selecting appropriate treatments for different variants by highlighting critical regions in different domains.

Validations confirmed SMuRF findings in the myogenic context

One caveat of SMuRF is that the HAP1 platform cell line, although widely used in α-DG-related studies, may not fully reflect the clinical relevance of dystroglycanopathies, which primarily affect neuromuscular tissues[87]. To address this issue, we generated myogenic platform cell lines by engineering MB135, a human control myoblast cell line[88]. Endogenous *FKRP* or *LARGE1* were knocked out respectively in the MB135 cell line. Monoclonal homozygous KO lines were established for both genes (Extended Data Fig. 10a). Despite being incompatible with the flow cytometric assay (Extended Data Fig. 10c), the KO MB135 myoblasts were effectively utilized for individual variant validation using an immunofluorescence assay that we developed.

The *FKRP*-KO and *LARGE1*-KO MB135 myoblasts were rescued by different individual variants using lentivirus and differentiated into myotubes for IIH6C4 IF staining (Fig. 6a and Extended Data Fig. 10b). The results were consistent with the SMuRF scores, the mini-library screen (Extended Data Fig. 6a,b) and the ClinVar reports. Again, the *FKRP* founder mutation p.Leu276Ile displayed an intermediate α-DG glycosylation signal, lying between the benign variants and the other pathogenic variants.

To further validate the SMuRF scores, we developed an orthogonal assay that is independent of the IIH6C4 antibody. Proper glycosylation of α-DG is crucial for the viral entry of Lassa fever virus (LASV)[89]. LASV glycoprotein complex (LASV-GPC) has been employed to generate recombinant vesicular stomatitis virus (rVSV-LASV-GPC) as a safer agent for investigating

LASV entry[90,91]. rVSV-LASV-GPC was utilized in a gene-trap screen in HAP1 cells to identify crucial genes involved in α-DG glycosylation, where cells with dysfunctional α-DG glycosylation genes exhibited increased resistance to rVSV-LASV-GPC infection, resulting in their enrichment in the population[52]. We increased the sensitivity of this VSV assay and employed it to validate interesting findings from the flow-cytometry assay (Fig. 6b). Instead of using rVSV-LASV-GPC, whose genome contains the LASV-GPC coding sequence, allowing the virus to re-enter cells, we utilized pseudotyped ppVSV-LASV-GPC. The ppVSV cannot re-enter cells as it is pseudotyped using a LASV-GPC helper plasmid and lacks a viral glycoprotein coding sequence in its genome (Methods). The *GOI*-KO MB135 myoblasts were first transduced with Lenti-*GOI* variants and selected with blasticidin. Subsequently, half of the cells were infected by the ppVSV-LASV-GPC, while the other half was kept as the non-infected control.

We then mixed 10 Lenti-*FKRP* variants with Lenti-WT-*FKRP* to rescue *FKRP*-KO MB135 (Fig. 6c), and mixed 11 Lenti-*LARGE1* variants with Lenti-WT-*LARGE1* to rescue *LARGE1*-KO MB135 (Fig. 6d). *FKRP* c.135C>T (p.Ala45=; SMuRF=0.21; AF= 0.14) and *LARGE1* c.435C>T (p.Ala145=; SMuRF=-0.58; AF=0.48) have the highest AFs in gnomAD v4, which were used as high-function references. *FKRP* c.663C>A (p.Ser221Arg; SMuRF=-2.21) and *LARGE1* c.992C>T (p.Ser331Phe; SMuRF=-1.81) are Pathogenic in ClinVar, which were used as low-function references. Nonsense variants with high SMuRF scores, as well as synonymous outliers with either low or ultra-high scores, were included to be validated (inclusion criteria: Supplementary Method 11).

The ppVSV assay supports the notion that the most 3' nonsense variants, *FKRP* c.1477G>T (p.Gly493Ter; SMuRF=-0.42) and *LARGE1* c.2257G>T (p.Glu753Ter; SMuRF=0.04), can retain the enzymatic functions, while other nonsense variants with relatively high SMuRF scores did not exhibit high functionality in the ppVSV assay. With the successful application of the ppVSV assay in small-scale variant interpretation experiments, we have laid the foundation for optimizing the conditions to make the assay applicable for large-scale, all-possible variant interpretation experiments (Supplementary Discussion 2).

Discussion

In this study, we introduced SMuRF as a more accessible workflow for DMS studies and demonstrated its utility in improving variant interpretation in genetic diseases. We developed an accessible saturation mutagenesis method and functional assays with the sensitivity and accuracy required by DMS studies. We generated functional scores for over 99.8% of all possible coding SNVs of *FKRP* and *LARGE1*. The SMuRF scores can enable severity prediction, which was validated with well-curated patient reports. SMuRF also resolved critical protein structural regions susceptible to missense disruptions, aiding in treatment decision making.

Previous DMS studies have explored all possible single amino acid substitutions (SAAS). We decided to explore all possible SNVs relative to the human reference sequence, therefore all amino acid substitutions were not represented. However, population sequencing data has demonstrated that multiple SNVs in the same codon (i.e. multi-nucleotide variants) are extremely rare[7], thus justifying that representing all possible SNVs is more practical and cost-effective when data is to be used to interpret variants observed in sequenced individuals. Furthermore, the

majority of VUS are SNVs resulting in missense variants[55]. Lastly, it is worth noting that the PALS-C saturation mutagenesis method in the SMuRF workflow can be extended to analyze small-sized variants beyond SNVs, including SAAS, small insertions, or deletions.

It is important to consider certain factors when utilizing scores from SMuRF or other DMS studies. The majority of dystroglycanopathies are recessive and patients may harbor compound heterozygous variants, which raises the need for further investigation on how to apply DMS functional scores in interpreting such cases. In this study, we explored a naive additive model where the biallelic functional scores were calculated by the simple addition of the SMuRF scores of the variants on both alleles. This model demonstrated a promising correlation between SMuRF scores and disease severity (Fig. 4c,d), however further future development will be needed. Additionally, our results emphasized the significance of well-curated reports in predicting disease severity (Supplementary Fig. 3). Our findings revealed a correlation between the disease onset age and SMuRF scores (Fig. 4d and Supplementary Fig. 11). However, creatine kinase (CK) values did not show a significant correlation with SMuRF scores (Supplementary Fig. 4). This observation is consistent with the knowledge in the field that CK levels can fluctuate with activity and decrease when muscle mass is lost over time[92].[92]

The use of SMuRF scores for interpretation of genetic variants poses challenges in the following aspects. Firstly, the IIH6C4 assay may only capture one of multiple functions associated with FKRP and LARGE1. For example, FKRP is also known to participate in the glycosylation of fibronectin[93]. Second, the presence of a homologous gene may impact the clinical relevance of the SMuRF scores. For instance, *LARGE2*, a paralog of *LARGE1* resulting from a duplication event first observed in Chondrichthyes[60], may act as an effective modifier in *LARGE1*-related diseases[94–96]. Third, SMuRF analysis for nonsense variants has inherent limitations. In the context of the CDS constructs used in SMuRF, a nonsense mutation does not trigger exon-junction complex (EJC)-enhanced NMD. However, in the correct multi-exon genomic context, when a nonsense mutation is located upstream of the last EJC, it undergoes EJC-dependent NMD, eliminating any residual functions that the truncated proteins may possess[97]. Lastly, variants that confer optimal enzymatic activity may not necessarily correspond to the ones favored by selection in a population. α-DG glycosylation plays a crucial role in LASV viral entry. It is possible that certain variants, despite conferring lower enzymatic activity for α-DG glycosylation, are favored by selection in populations where LASV is epidemic[80,98]. These four aspects emphasize the need for orthogonal validation assays. In our study, we created myogenic platform cell lines to validate the findings from the flow cytometry assay. We employed a ppVSV viral-entry-based assay as an orthogonal assay. This marks the first instance of enhancing the sensitivity of the VSV assay from a gene discovery level to a variant interpretation level in α-DG research.

In addition to FKRP and LARGE1, IIH6C4-SMuRF can be readily employed for other enzymes involved in α-DG Core M3 glycosylation (Fig. 1a). Additionally, variants in POMGNT1, an enzyme that mainly participates in the glycosylation of Core M1 and M2, can perturb the IIH6C4 signal[37] (Extended Data Fig. 2), potentially by serving as an “enzymatic chaperone” for FKTN[60,99–101]. However, variants in GnT-Vb/IX, an enzyme that participates in Core M2 glycosylation, are unlikely to alter IIH6C4 signal[102]. Therefore, IIH6C4-SMuRF is readily applicable to at least 12 other dystroglycanopathy enzymes. Furthermore, IIH6C4-SMuRF can

be adapted for other dystroglycanopathy candidate genes, such as *SLC35A1* and *SLC35A4*, which are involved in the transport of CDP-ribitol[103].

The SMuRF workflow is versatile and can be adapted to utilize other gene/pathway-specific functional assays, therefore broadening its utility across disease genes. Notably, the MaveDB database aggregates data from published DMS studies, but so far, only about 50 human genes have been recorded[104]. Efforts to chart the "Atlas of Variant Effects" should be crowdsourced, with experts in their respective fields contributing, which requires the technology used in DMS to be accessible to as many laboratories as possible. We developed SMuRF workflow to reduce DMS cost and complexity such that more researchers can contribute to the crowd-sourcing efforts required for coverage across thousands of disease genes.

Methods

Cell culture

Wildtype HAP1 (C631) and *DAG1*-KO HAP1 (HZGHC000120c016) were ordered from Horizon Discovery. All HAP1 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM) (Gibco, 12440053) with 10% Fetal Bovine Serum (FBS, R&D Systems, S11150) and 1× Antibiotic-Antimycotic (Anti-anti, Gibco, 15240062). The medium was replaced every 2 days, unless otherwise stated. HAP1 cells tend to grow into multi-layers; hence, to keep the cells in optimal status, TrypLE Express Enzyme (Gibco, 12605010) was used to passage the cells to maintain the cells in healthy confluency (30-90%). The HAP1 cells used in SMuRF were immortalized using lentivirus packaged with pLV-hTERT-IRES-hygro (Addgene, 85140), a gift from Tobias Meyer, as previously described[105]. HEK293T cells were cultured in DMEM (Gibco, 11995065) with 10% FBS and 1× Anti-anti. All MB135 cells were cultured in Ham's F-10 Nutrient Mix (Gibco, 11550043) with 20% FBS, 1× Anti-anti, 51 ng/ml dexamethasone (Sigma-Aldrich, D2915) and 10 ng/mL basic fibroblast growth factor (EMD/Millipore, GF003AF-MG). The medium was replaced every 2 days, unless otherwise stated. The MB135 cells were differentiated in Skeletal Muscle Differentiation Medium (PromoCell, C-23061) with 1× Anti-anti. The differentiation medium was replaced every 4 days, unless otherwise stated.

CRISPR RNP nucleofection

Synthetic Single Guide RNA (sgRNA) Kits and SpCas9 2NLS Nuclease were ordered from Synthego. RNP complexes were prepared in SE Cell Line Nucleofector Solution (Lonza, PBC1-00675) and delivered into cells with a Lonza 4D-Nucleofector. The program used for HAP1 was EN-138; the program used for MB135 was CA-137. Single clones were isolated from pooled nucleofected cells and genotyped by targeted Sanger sequencing. sgRNA sequences, RNP complex preparation conditions, and genotyping primers can be found in Supplementary Method 1. *FKRP*-KO HAP1 carries a 1-bp insertion (c.181Adup); *FKRP*-KO MB135 is homozygous for the same mutation. *LARGE1*-KO HAP1 carries a 94-bp deletion (c.121_214del); *LARGE1*-KO MB135 is homozygous for the same mutation.

Plasmid construction

Lenti-*DAG1* plasmid used the backbone of lentiCRISPR v2, which was a gift from Feng Zhang (Addgene, 52961). *DAG1* coding exons were cloned from human genome DNA by PCR. Lenti-*FKRP* plasmids and Lenti-*LARGE1* plasmids used the backbone of lentiCas9-Blast, which was a

gift from Feng Zhang (Addgene, 52962). *FKRP* coding exon was cloned from HAP1 genome DNA. *LARGE1* coding sequence was cloned from HEK293T cDNA. HEK293T carries a *LARGE1* mutation (c.1848G>A) on one allele, which was removed from the Lenti-*LARGE1* plasmids to make the pooled variant library. The removal of this mutation used the same strategy as the introduction of individual variants to the lentiviral plasmids for the mini-libraries: briefly, a short localized region was cut with restriction enzymes from the wildtype plasmid and 2 variant-carrying inserts, each covering 1 of 2 sides of this region were inserted. The UbC promoter was cloned from pAAV-UbC-eGFP-F, which was a gift from Pantelis Tsoulfas (Addgene, 71545). The EF-1α promoter was taken from lentiGuide-Puro, which was a gift from Feng Zhang (Addgene, 52963). BSD-WPRE was from lentiCas9-Blast. The lentiviral plasmids used for the pooled library contain a UbC-driven gene-of-interest CDS and an EF-1α-driven *BSD*. *BSD* encodes Blasticidin S deaminase (BSD), which confers blasticidin resistance in transduced cells. Plasmid assemblies were achieved either with NEBuilder HiFi DNA Assembly Master Mix (NEB, E2621) or T4 DNA Ligase (M0202). Cloning details of plasmid construction and the list of plasmids deposited to Addgene can be found in Supplementary Method 2.

RT-PCR and RT-qPCR

RT-PCR and RT-qPCR were performed following manufacturers' manuals. PrimeScript RT Reagent Kit (Takara, RR037) was used for cDNA synthesis. Phusion High-Fidelity DNA Polymerase (NEB, M0530) was used for PCR reactions. SsoAdvanced Universal SYBR Green Supermix (Bio-Rad, 1725271), Hard-Shell 96-Well PCR Plates (Bio-Rad, HSP9601), Plate Sealing Film (Bio-Rad, MSB1001) and Bio-Rad C1000 Touch Thermal Cycler were used for qPCR experiments. Primers can be found in Supplementary Method 3.

Lentivirus packaging and transduction

Lentivirus was packaged by HEK293T cells. For a 10-cm dish (90% confluency), 1.5 mL Opti-MEM (Gibco, 31985062), 10 µg psPAX2 (Addgene, 12260; or use dR8.91 instead), 2 µg pMD2.G (Addgene, 12259), 9 µg lentiviral plasmid, and 50 µL TransIT-LT1 Transfection Reagent (Mirus, MIR 2300) were mixed at room temperature for 15 mins and then added to the cells. 3.5 mL DMEM was added to the cells. 72 hrs later, the supernatant in the dish was filtered with 0.45 µm PES filter (Thermo Scientific, 165-0045), mixed with 5 mL Lenti-X Concentrator (Takara, 631232) and rocked at 4 °C overnight. The viral particles were then spun down (1800 ×g, 4 °C, 1hr) and resuspended in 200 µL DMEM. Lentivirus was titrated with Lenti-X GoStix Plus (Takara, 631280). For lentiviral transduction, the cells to be transduced were plated in wells of plates. One day after seeding, the medium was replaced and supplemented with polybrene (final conc. 8 µg/mL). Lentivirus was then added to the wells for a spinfection (800 ×g, 30 °C, 1hr). One day post-transduction, the medium was replaced, and drug selection was started if applicable. For constructs with BSD, Blasticidin S HCl (Gibco, A1113903, final conc. 5 µg/mL) was used for drug selection. For constructs with PuroR, Puromycin Dihydrochloride (Gibco, A1113803, final conc. 1 µg/mL) was used. Drug selection was performed for 10-14 days.

PALS-C cloning for saturation mutagenesis

Each variant of all possible CDS SNVs (Extended Data Fig. 5a,b) was included in a 64-bp ssDNA oligo. The oligos were synthesized (one pool per GOI) by Twist Bioscience. PALS-C is an 8-step cloning strategy to clone lentiviral plasmid pools from the oligos. An elaborate protocol can be found in Supplementary Method 4. Briefly, the oligos were used as PCR reverse

primers, which were annealed to the plasmid template and extended towards the 5' end of the gene of interest. The resulting products of each block were isolated using block-specific primers. Then the variant strands were extended towards the 3' end to get the full-length sequences, which were subsequently inserted into the plasmid backbone using NEBuilder (NEB, E2621). The purifications for PALS-C steps were done with NucleoSpin Gel and PCR Clean-Up kit (Takara, 740609). Final assembled products were delivered to Endura Electrocompetent Cells (Lucigen, 60242-1) via electrotransformation (Bio-Rad Gene Pulser II). Transformed bacteria were grown overnight and plasmid pools were extracted using the PureLink Midiprep Kit (Invitrogen, K210014). To check library complexity, colony forming units (CFUs) were calculated and a minimum 18 × coverage was achieved for the plasmid pool of each block of *FKRP* and *LARGE1*. Variants that created new type2S enzyme recognition sites tended to be underrepresented in the pool. These variants are reported in Supplementary Method 5.

Quality control (QC) of plasmid pools and saturation mutagenesis

QC was performed for the plasmid pools using the Amplicon-EZ service provided by GENEWIZ (Extended Data Fig. 5c,d and Supplementary Method 6). 99.6% of the SNVs of both genes were represented in the plasmid pools (Extended Data Fig. 5e,f). Lentivirus of each block was packaged by HEK293T cells in one 10-cm dish. Small-scale pre-experiments were performed to determine the viral dosage for optimal separation. GoStix Value (GV) quantified by the Lenti-X GoStix App (Takara) was used to scale the titer of each block to be the same. GV is subject to viral-packaging batch effects; hence, lentiviral pools of all blocks were packaged at the same time using the reagents and helper plasmids of the same batch. Depending on the specific batch and packaging system, 1e3-1e5 GV×µL of lentivirus was used for each block. Considerations for controlling the Multiplicity of Infection (MOI) of the lentiviral pool were discussed in Supplementary Discussion 1. For each block, 600k HAP1 cells or 200k MB135 cells were plated in a well of a 6-well plate for transduction. The cell number was counted with an Automated Cell Counter (Bio-Rad, TC20). The cell number for each block was expanded to more than 30M for FACS.

Proof-of-concept mini-libraries for the FACS assay

Mini-libraries of variants were employed to examine and optimize the separation of the FACS assay. Lentiviral constructs were cloned and packaged individually for 8 *FKRP* variants and 3 *LARGE1* variants in addition to the wild-type constructs. These lentiviral particles were mixed to make a mix-9 *FKRP* mini-library and a mix-4 *LARGE1* mini-library. Conditions of transduction, staining, sorting and gDNA extraction were optimized using the mini-libraries. Relative enrichment of variants was defined as the ratio of the variants' representation in the high-glycosylation sample to their representation in the low-glycosylation group, which was quantified with either Sanger sequencing or Amplicon-EZ NGS (Supplementary Method 7).

Staining for FFC and FACS

Reagent volumes were determined based on sample size. Below, the staining for samples of one gene block is described as an example. The cells were washed twice with DPBS (Gibco, 14190144), digested with Versene (Gibco, 15040066), and counted. 30M cells were used for staining, which was performed in a 15 mL tube. The cells were spun down (700 ×g, 4 °C, 15 mins) and resuspended in 3mL DPBS supplemented with 30 µL Viobility 405/452 Fixable Dye (Miltenyi Biotec, 130-130-420). All the following steps were done in the dark. The sample was

gently rocked at room temperature for 30 mins, and then 7 mL PEB buffer (1 volume of MACS BSA Stock Solution, Miltenyi Biotec, 130-091-376 ;19 volumes of autoMACS Rinsing Solution, Miltenyi Biotec, 130-091-222) was added to the tube. The cells were spun down (700 ×g, 4 °C, 15 mins) and resuspended in 3mL DPBS supplemented with 30 µL Human BD Fc Block (BD Pharmingen, 564220). The sample was gently rocked at room temperature for 30 mins, and then 7 mL DPBS was added. The cells were spun down (700 ×g, 4 °C, 10 mins) and resuspended in 3mL MAGIC buffer (5% FBS; 0.1% NaAz w/v; 10% 10× DPBS, Gibco, 14200166; water, Invitrogen, 10977015) supplemented with 15 µL IIH6C4 antibody (Sigma-Aldrich, 05-593, discontinued; or antibody made in Dr. Kevin Campbell's lab). The sample was gently rocked at 4 °C for 20 hrs. 7 mL MAGIC buffer was added before the cells were spun down (700 ×g, 4 °C, 10 mins) and resuspended in 3 mL MAGIC buffer supplemented with 60 µL Rabbit anti-Mouse IgM FITC Secondary Antibody (Invitrogen, 31557). The sample was gently rocked at 4 °C for 20 hrs. 7 mL DPBS was added to the sample before the cells were spun down (700 ×g, 4 °C, 10 mins), resuspended with 4 mL DPBS and filtered with a 40 µm Cell Strainer (Falcon, 352340). Important: DO NOT use IIH6C4 antibody from Santa Cruz, sc-73586.

FFC and FACS gating parameters

FFC experiments were performed with a BD LSR II Flow Cytometer; FACS experiments were performed with a BD FACSAria Flow Cytometer. Forward scatter (FSC) and side scatter (SSC) were used to exclude cell debris and multiplets. Singlets were isolated for downstream analysis. Pacific Blue (450/50 BP) or an equivalent channel was used to detect the Viobility 405/452 Fixable Dye and isolate the live cells for analysis. FITC (530/30 BP), GFP (510/20 BP) or an equivalent channel was used to detect the FITC secondary antibody signal. 20k events were recorded for each block to decide the gating parameters. For FACS, the top ~20% of the cells were isolated as the high-glycosylation group and the bottom ~40% of the cells were isolated as the low-glycosylation group. The .fcs files, the FlowJo .wsp files and the software interface reports of the sorter were made available on FlowRepository. A minimum ~1000 × coverage (*e.g.*, 750k cells harvested for a block with 750 variants) was achieved for both groups of each block.

NGS library construction

The cells were spun down (800 ×g, 4 °C, 10 mins), and gDNA was harvested from each sample with PureLink Genomic DNA Mini Kit (Invitrogen, K182002). A 3-step PCR library construction was performed to build the sequencing library. Step1: lentiviral sequence isolation. A pair of primers specific to the lentiviral backbone was used to amplify the lentiviral CDS sequences of each sample. Step2: block isolation. Each primer contained a 20-bp flanking sequence of the specific block and a partial Illumina adaptor sequence. F primers contained the barcodes to distinguish the high-glycosylation group and the low-glycosylation group. Step3: adaptor addition. Step2 products were multiplexed and the rest of the Illumina adaptor was added to the amplicons. An elaborate protocol can be found in Supplementary Method 8. The NGS libraries were sequenced using Psomagen's HiSeq X service. ~400M reads were acquired per library.

SMuRF score generation

Enrichment of a variant (E_var) in a FACS group is calculated as a ratio of the count of the variant (c_var) to the total count (c_total) at the variant site:

$$E_var = c_var / c_total$$

Enrichment of the WT (E_WT) is calculated separately for each block. E_WT is calculated as a ratio of the number of the reads without variant (r_WT) to the number of the reads with one or no variant (r_clean).

$$E_WT = r_WT / r_clean$$

Relative enrichment (rE) is a ratio of the enrichment in the high-glycosylation group to the enrichment in the low-glycosylation group:

$$rE_var = E_var_high / E_var_low$$
$$rE_WT = E_WT_high / E_WT_low$$

The functional score of a variant is calculated as the ratio of its relative enrichment to that of the WT sequence in the corresponding block, and the SMuRF score is calculated as the log2 value of the functional score:

$$Functional_score = rE_var / rE_WT$$
$$SMuRF = log2(Functional_score)$$

Count of variants and reads were generated from raw sequencing data using the analytical pipeline deposited in the GitHub repository Gargamel. SMuRF scores were calculated using the scripts deposited in the GitHub repository Azrael.

The variants missing from the *FKRP* pools were: c.279G>C, c.430A>C, c.432G>C, c.439G>C and c.454A>C.

Confidence score generation and classification

In order to account for technical confounders, we have developed a confidence scoring system to assess the reliability of functional scores assigned to each variant. Our approach assumes that synonymous variants, which are not expected to have a functional effect, can serve as a null model (Supplementary Discussion 3). Hence, we expect the functional score for synonymous variants to be 1 (SMuRF=0). We have identified two key technical confounders: 1) the position of the variant in the block for the functional assay, and 2) the coverage (defined as the sum of "high" and "low" reads).

First, we developed a confidence score for each of these confounders. For the position in the block, we binned the variants into groups, with each bin representing 10 base pairs. Within each bin, we took the mean of the functional score of all synonymous variants, and the confidence score per bin was then derived as follows:

$$Confidence\ (block\ position) = \frac{1}{mean(synoymous\ functional\ score)}$$

A confidence score that is closer to 1 indicates higher confidence/reliability. We employed a similar approach for the coverage-based confidence score; We binned the synonymous variants into groups representing every 5th percentile of the coverage distribution, and the per-bin confidence score was derived with the formula above. Subsequently, we assigned both confidence scores to all variants in the testing set based on their position in the block and

coverage. To integrate the effects of both confounding variables, the mean of both confidence scores for each variant was calculated. After this, a percentile-based confidence rank score was calculated based on the difference between the combined confidence score and 1 (which is taken to be the null, and further deviations from 1 indicate lower reliability). The final confidence rank score ranges from 0 to 1, with values closer to 1 indicating variants with higher confidence within the set. After which, the variants were assigned "LOW" confidence if their rank score was < 0.05, "MEDIUM" confidence if their rank score was ≥ 0.05 and < 0.50, and "HIGH" confidence if their rank score was above 0.50.

Visualization of data on protein 3D structures

Protein 3D structures from Protein Data Bank (PDB) were visualized using UCSF ChimeraX v1.3[106]. The crystal structure of human FKRP (PDB:6KAM)[79] and the electron microscopy structure of LARGE1 (PDB:7UI7)[80] were used. Figures displaying domain locations or the log2 mean missense score per residue on 3D structures were generated using custom ChimeraX command files. Domain coordinates displayed are per Ortiz-Cordero *et al.* for FKRP[81], and Joseph *et al.* for LARGE1[80]. The log2 of the mean missense score per residue was calculated using a custom Python script.

Immunofluorescence

15 mm round Thermanox Coverslips (Thermo Scientific, 174969) were placed in the wells of 24-well plates. To coat the coverslips, 0.1% gelatin (Sigma-Aldrich, G9391) was added to the wells and immediately removed. After the coverslips were air-dried, 250k MB135 cells were resuspended in 0.5 mL growth medium and seeded into each well. One day after plating the cells, the medium was changed to the differentiation medium, and cells were differentiated for 3-7 days until myotubes were formed. The cells were washed with DPBS and fixed with 4% PFA (Sigma-Aldrich, 158127) for 10 mins at room temperature. The cells were blocked with 2% Bovine Serum Albumin (BSA, Sigma-Aldrich, A9647) at room temperature for 1 hr before undergoing incubation with the IIH6C4 antibody (1:200 in 2% BSA, Sigma-Aldrich, 05-593, discontinued) at 4 °C for 20 hrs. The cells were then washed with DPBS before undergoing incubation with the secondary antibody (1:100 in 2% BSA, Invitrogen, 31557) at room temperature for 2 hrs in the dark. Antifade Mounting Medium with DAPI (Vector Laboratories, H1500) was dropped onto Microscope Slides (Fisher Scientific, 22-037-246). The coverslips were washed again with DPBS and put onto the drops on the slides facedown and kept at room temperature for 30 mins in the dark. Pictures were taken with a Revolve ECHO microscope. (DO NOT use IIH6C4 antibody from Santa Cruz, sc-73586. An alternative IIH6C4 antibody may be acquired from Dr. Kevin Campbell.)

Packaging and infection of rVSV / ppVSV

rVSV-LASV-GPC viral particles, ppVSVΔG-VSV-G viral particles, and the LASV-GPC plasmid were obtained from Dr. Melinda Brindley's group. To package ppVSV-LASV-GPC viral particles, HEK293T cells were transfected with the LASV-GPC plasmid and then transduced with ppVSVΔG-VSV-G viral particles. The resulting particles were referred to as ppVSV-LASV-GPC-Generation1. A new batch of LASV-GPC transfected HEK293T cells were subsequently transduced with ppVSV-LASV-GPC-Generation1 to produce ppVSV-LASV-GPC-Generation2, reducing residual VSV-G in the pseudotyped particles. Similarly, later generations can be packaged. The experiments in this study utilized ppVSV-LASV-GPC-Generation2 and 3.

The 50% tissue culture infectious dose (TCID50) of the VSV was determined using the Spearman-Karber method[107,108]. Lentiviral transduction and blasticidin drug selection were performed in the same manner as those in the FACS assay. Afterwards, cells were divided into two groups (~1M cells each): a no-infection group and an infection group. rVSV infection was conducted at an approximate MOI of 0.5. NH4Cl (Sigma-Aldrich, A9434, final conc. 5mM) was added during the infection and subsequent recovery. After 60 hours of infection, the medium was replaced, and the cells were allowed to recover for 12 hours before harvesting. ppVSV infection was performed at an approximate MOI of 1~3, and the infected cells were recovered to ~1M prior to harvesting. Detailed ppVSV packaging protocol and considerations for application can be found in Supplementary Method 10.

Mini-libraries for the validation using the ppVSV assay

10 Lenti-*FKRP* variants were mixed with Lenti-WT-*FKRP* to rescue *FKRP*-KO MB135. 11 Lenti-*LARGE1* variants were mixed with Lenti-WT-*LARGE1* to rescue *LARGE1*-KO MB135. The lentiviral plasmids were individually constructed and validated using the whole-plasmid sequencing service provided by Plasmidsaurus. The individual plasmids were mixed to package the *FKRP* mix11 and *LARGE* mix12 lentiviral pools. Lentiviral transduction, drug selection, and ppVSV infection were performed similarly as stated above. After harvesting the cells, the lentiviral GOI sequence was amplified with the PCR1 protocol in Supplementary Method 8 and sequenced with nanopore sequencing provided by Plasmidsaurus. The functional score was quantified by the ratio of a variant's enrichment in the non-infected group to its enrichment in the ppVSV-infected group. The list of the variants included in the mini-libraries and other details can be found in Supplementary Method 11.

Statistics

Two-sided Wilcoxon tests were performed with the "ggsignif" R package. Spearman's rank correlation coefficients were calculated with the "cor.test" function in R.

Declaration of AI-assisted technologies in the writing process

We used ChatGPT 3.5 to improve the readability and language in this manuscript. The manuscript was first drafted by us and polished with ChatGPT sentence-by-sentence where we deemed necessary. We then reviewed and finalized the text. We take full responsibility for the contents of this manuscript.

Data availability

Scores generated in this manuscript can be found in Supplementary Table 4 (*FKRP*) and Supplementary Table 5 (*LARGE1*). Plasmids were made available on Addgene: Lenti-*DAG1* (205149), Lenti-UbC-*FKRP*-EF1α-BSD (205150), and Lenti-UbC-*LARGE1*-EF1α-BSD (205151). FFC and FACS datasets were made available on FlowRepository (FR-FCM-Z6LL, FR-FCM-Z6LM, FR-FCM-Z6LN, FR-FCM-Z6LP, FR-FCM-Z6LQ, FR-FCM-Z6LR, FR-FCM-Z6LX). NGS raw data were deposited to the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) (accession: PRJNA993285). Detailed experimental protocols are available on protocols.io. All scripts used in this manuscript can be found on GitHub (https://github.com/leklab). Scripts in the Balthazar repository were used for oligo design and other pre-SMuRF experiments. Analytical pipeline in the Gargamel repository

was used for processing the raw NGS data. Scripts in the Azrael repository were used for generating SMuRF scores and downstream analyses.

References

1. Fridman, H. *et al.* The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. *Am. J. Hum. Genet.* **108**, 608–619 (2021).
2. Balick, D. J., Jordan, D. M., Sunyaev, S. & Do, R. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. *Am. J. Hum. Genet.* **109**, 33–49 (2022).
3. Barton, A. R., Hujoel, M. L. A., Mukamel, R. E., Sherman, M. A. & Loh, P.-R. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. *Am. J. Hum. Genet.* **109**, 1298–1307 (2022).
4. Schmenger, T., Diwan, G. D., Singh, G., Apic, G. & Russell, R. B. Never-homozygous genetic variants in healthy populations are potential recessive disease candidates. *NPJ Genom. Med.* **7**, 54 (2022).
5. Boycott, K. M. *et al.* A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell* **177**, 32–37 (2019).
6. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
7. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
8. Blöß, S. *et al.* Diagnostic needs for rare diseases and shared prediagnostic phenomena: Results of a German-wide expert Delphi survey. *PLoS ONE* **12**, e0172532 (2017).

9. Libell, E. M. *et al.* The outcomes and experience of pregnancy in limb girdle muscular dystrophy type R9. *Muscle Nerve* **63**, 812–817 (2021).

10. Starita, L. M. *et al.* Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

11. Caswell-Jin, J. L. *et al.* Racial/ethnic differences in multiple-gene sequencing results for hereditary cancer risk. *Genet. Med.* **20**, 234–239 (2018).

12. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).

13. Wei, H. & Li, X. Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Front. Genet.* **14**, 1087267 (2023).

14. Fowler, D. M. *et al.* An Atlas of Variant Effects to understand the genome at nucleotide resolution. *Genome Biol.* **24**, 147 (2023).

15. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48**, 1570–1575 (2016).

16. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* **513**, 120–123 (2014).

17. Haller, G. *et al.* Massively parallel single-nucleotide mutagenesis using reversibly terminated inosine. *Nat. Methods* **13**, 923–924 (2016).

18. Erwood, S. *et al.* Saturation variant interpretation using CRISPR prime editing. *Nat. Biotechnol.* **40**, 885–895 (2022).

19. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–6, 4 p following 206 (2015).

20. Jia, X. *et al.* Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk. *Am. J. Hum. Genet.* **108**, 163–175 (2021).

21. Findlay, G. M. Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. *Hum. Mol. Genet.* **30**, R187–R197 (2021).

22. Tucker, J. D., Lu, P. J., Xiao, X. & Lu, Q. L. Overexpression of mutant FKRP restores functional glycosylation and improves dystrophic phenotype in FKRP mutant mice. *Mol. Ther. Nucleic Acids* **11**, 216–227 (2018).

23. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol.* **18**, 150 (2017).

24. Meilleur, K. G. *et al.* Clinical, pathologic, and mutational spectrum of dystroglycanopathy caused by LARGE mutations. *J. Neuropathol. Exp. Neurol.* **73**, 425–441 (2014).

25. Kanagawa, M. Dystroglycanopathy: from elucidation of molecular and pathological mechanisms to development of treatment methods. *Int. J. Mol. Sci.* **22**, (2021).

26. Borisovna, K. O. *et al.* Compound heterozygous POMGNT1 mutations leading to muscular dystrophy-dystroglycanopathy type A3: a case report. *BMC Pediatr.* **19**, 98 (2019).

27. Geis, T. *et al.* Clinical long-time course, novel mutations and genotype-phenotype correlation in a cohort of 27 families with POMT1-related disorders. *Orphanet J. Rare Dis.* **14**, 179 (2019).

28. Van Reeuwijk, J. *et al.* A homozygous FKRP start codon mutation is associated with Walker-Warburg syndrome, the severe end of the clinical spectrum. *Clin. Genet.* **78**, 275–281 (2010).

29. Qin, J. Y. *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS ONE* **5**, e10611 (2010).

30. Johnson, K. *et al.* Detection of variants in dystroglycanopathy-associated genes through the application of targeted whole-exome sequencing analysis to a large cohort of patients with unexplained limb-girdle muscle weakness. *Skelet. Muscle* **8**, 23 (2018).

31. Kanagawa, M. *et al.* Identification of a Post-translational Modification with Ribitol-Phosphate and Its Defect in Muscular Dystrophy. *Cell Rep.* **14**, 2209–2223 (2016).

32. Goddeeris, M. M. *et al.* LARGE glycans on dystroglycan function as a tunable matrix scaffold to prevent dystrophy. *Nature* **503**, 136–140 (2013).

33. Wu, B. *et al.* Ribitol dose-dependently enhances matriglycan expression and improves muscle function with prolonged life span in limb girdle muscular dystrophy 2I mouse model. *PLoS ONE* **17**, e0278482 (2022).

34. Vannoy, C. H., Leroy, V. & Lu, Q. L. Dose-Dependent Effects of FKRP Gene-Replacement Therapy on Functional Rescue and Longevity in Dystrophic Mice. *Mol. Ther. Methods Clin. Dev.* **11**, 106–120 (2018).

35. Yonekawa, T. *et al.* Large1 gene transfer in older myd mice with severe muscular dystrophy restores muscle function and greatly improves survival. *Sci. Adv.* **8**, eabn0379 (2022).

36. Dhoke, N. R. *et al.* A universal gene correction approach for FKRP-associated dystroglycanopathies to enable autologous cell therapy. *Cell Rep.* **36**, 109360 (2021).

37. Yoshida-Moriguchi, T. & Campbell, K. P. Matriglycan: a novel polysaccharide that links dystroglycan to the basement membrane. *Glycobiology* **25**, 702–713 (2015).

38. Panicucci, C. *et al.* Mutations in GMPPB Presenting with Pseudometabolic Myopathy. *JIMD Rep.* **38**, 23–31 (2018).

39. Barone, R. *et al.* DPM2-CDG: a muscular dystrophy-dystroglycanopathy syndrome with severe epilepsy. *Ann. Neurol.* **72**, 550–558 (2012).

40. Lefeber, D. J. *et al.* Autosomal recessive dilated cardiomyopathy due to DOLK mutations results from abnormal dystroglycan O-mannosylation. *PLoS Genet.* **7**, e1002427 (2011).

41. Hu, H. *et al.* Conditional knockout of protein O-mannosyltransferase 2 reveals tissue-specific roles of O-mannosyl glycosylation in brain development. *J. Comp. Neurol.* **519**, 1320–1337 (2011).

42. Walimbe, A. S. *et al.* POMK regulates dystroglycan function via LARGE1-mediated elongation of matriglycan. *eLife* **9**, (2020).

43. Endo, Y. *et al.* Milder forms of muscular dystrophy associated with POMGNT2 mutations. *Neurol. Genet.* **1**, e33 (2015).

44. Stevens, E. *et al.* Mutations in B3GALNT2 cause congenital muscular dystrophy and hypoglycosylation of α-dystroglycan. *Am. J. Hum. Genet.* **92**, 354–365 (2013).

45. Willer, T. *et al.* ISPD loss-of-function mutations disrupt dystroglycan O-mannosylation and cause Walker-Warburg syndrome. *Nat. Genet.* **44**, 575–580 (2012).

46. Ujihara, Y. *et al.* Elimination of fukutin reveals cellular and molecular pathomechanisms in muscular dystrophy-associated heart failure. *Nat. Commun.* **10**, 5754 (2019).

47. Lee, A. J. *et al.* Clinical, genetic, and pathologic characterization of FKRP Mexican founder mutation c.1387A>G. *Neurol. Genet.* **5**, e315 (2019).

48. Manya, H. *et al.* The Muscular Dystrophy Gene TMEM5 Encodes a Ribitol β1,4-Xylosyltransferase Required for the Functional Glycosylation of Dystroglycan. *J. Biol. Chem.* **291**, 24618–24627 (2016).

49. Willer, T. *et al.* The glucuronyltransferase B4GAT1 is required for initiation of LARGE-mediated α-dystroglycan functional glycosylation. *eLife* **3**, (2014).

50. Longman, C. *et al.* Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha-dystroglycan. *Hum. Mol. Genet.* **12**, 2853–2861 (2003).

51. Stevens, E. *et al.* Flow cytometry for the analysis of α-dystroglycan glycosylation in fibroblasts from patients with dystroglycanopathies. *PLoS ONE* **8**, e68958 (2013).

52. Jae, L. T. *et al.* Deciphering the glycosylome of dystroglycanopathies using haploid screens for lassa virus entry. *Science* **340**, 479–483 (2013).

53. Sheikh, M. O. *et al.* Cell surface glycan engineering reveals that matriglycan alone can recapitulate dystroglycan binding and function. *Nat. Commun.* **13**, 3617 (2022).

54. Beigl, T. B., Kjosås, I., Seljeseth, E., Glomnes, N. & Aksnes, H. Efficient and crucial quality control of HAP1 cell ploidy status. *Biol. Open* **9**, (2020).

55. Song, D. *et al.* Genetic variations and clinical spectrum of dystroglycanopathy in a large cohort of Chinese patients. *Clin. Genet.* **99**, 384–395 (2021).

56. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).

57. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).

58. Benitez-Cantos, M. S. *et al*. Translation initiation downstream from annotated start codons in human mRNAs coevolves with the Kozak context. *Genome Res.* **30**, 974–984 (2020).

59. Chen, S. *et al.* A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *BioRxiv* (2022) doi:10.1101/2022.03.20.485034.

60. Bigotti, M. G. & Brancaccio, A. High degree of conservation of the enzymes synthesizing the laminin-binding glycoepitope of α-dystroglycan. *Open Biol.* **11**, 210104 (2021).

61. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

62. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-5 (2014).

63. Awano, H. *et al*. FKRP mutations cause congenital muscular dystrophy 1C and limb-girdle muscular dystrophy 2I in Asian patients. *J. Clin. Neurosci.* **92**, 215–221 (2021).

64. Unnikrishnan, G. *et al.* Phenotype Genotype Characterization of FKRP-related Muscular Dystrophy among Indian Patients. *J. Neuromuscul. Dis.* **10**, 615–626 (2023).

65. Brown, S. C., Fernandez-Fuente, M., Muntoni, F. & Vissing, J. Phenotypic Spectrum of α-Dystroglycanopathies Associated With the c.919T>a Variant in the FKRP Gene in Humans and Mice. *J. Neuropathol. Exp. Neurol.* **79**, 1257–1264 (2020).

66. Beltran-Valero de Bernabé, D. *et al*. Mutations in the FKRP gene can cause muscle-eye-brain disease and Walker-Warburg syndrome. *J. Med. Genet.* **41**, e61 (2004).

67. Louhichi, N. *et al*. New FKRP mutations causing congenital muscular dystrophy associated with mental retardation and central nervous system abnormalities. Identification of a founder mutation in Tunisian families. *Neurogenetics* **5**, 27–34 (2004).

68. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

69. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).

70. Ioannidis, N. M. *et al*. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

71. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).

72. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

73. Quinodoz, M. *et al.* Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am. J. Hum. Genet.* **109**, 457–470 (2022).

74. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *BioRxiv* (2023) doi:10.1101/2023.05.01.538953.

75. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

76. Danzi, M. C. *et al.* Deep structured learning for variant prioritization in Mendelian diseases. *Nat. Commun.* **14**, 4167 (2023).

77. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

78. Ljungdahl, A. *et al.* AlphaMissense is better correlated with functional assays of missense impact than earlier prediction algorithms. *BioRxiv* (2023) doi:10.1101/2023.10.24.562294.

79. Kuwabara, N. *et al.* Crystal structures of fukutin-related protein (FKRP), a ribitol-phosphate transferase related to muscular dystrophy. *Nat. Commun.* **11**, 303 (2020).

80. Joseph, S. *et al.* Structure and mechanism of LARGE1 matriglycan polymerase. *BioRxiv* (2022) doi:10.1101/2022.05.12.491222.

81. Ortiz-Cordero, C. *et al.* NAD+ enhances ribitol and ribose rescue of α-dystroglycan functional glycosylation in human FKRP-mutant myotubes. *eLife* **10**, (2021).

82. Liang, W.-C. *et al.* Limb-girdle muscular dystrophy type 2I is not rare in Taiwan. *Neuromuscul. Disord.* **23**, 675–681 (2013).

83. Topaloglu, H. *et al.* FKRP gene mutations cause congenital muscular dystrophy, mental retardation, and cerebellar cysts. *Neurology* **60**, 988–992 (2003).

84. Poppe, M. *et al.* The phenotype of limb-girdle muscular dystrophy type 2I. *Neurology* **60**, 1246–1251 (2003).

85. Krag, T. O. & Vissing, J. A New Mouse Model of Limb-Girdle Muscular Dystrophy Type 2I Homozygous for the Common L276I Mutation Mimicking the Mild Phenotype in Humans. *J. Neuropathol. Exp. Neurol.* **74**, 1137–1146 (2015).

86. Cataldi, M. P., Lu, P., Blaeser, A. & Lu, Q. L. Ribitol restores functionally glycosylated α-dystroglycan and improves muscle function in dystrophic FKRP-mutant mice. *Nat. Commun.* **9**, 3448 (2018).

87. Taniguchi-Ikeda, M., Morioka, I., Iijima, K. & Toda, T. Mechanistic aspects of the formation of α-dystroglycan and therapeutic research for the treatment of α-dystroglycanopathy: A review. *Mol. Aspects Med.* **51**, 115–124 (2016).

88. Jagannathan, S. *et al.* Model systems of DUX4 expression recapitulate the transcriptional profile of FSHD cells. *Hum. Mol. Genet.* (2016) doi:10.1093/hmg/ddw271.

89. Joseph, S. & Campbell, K. P. Lassa Fever Virus Binds Matriglycan-A Polymer of Alternating Xylose and Glucuronate-On α-Dystroglycan. *Viruses* **13**, (2021).

90. Acciani, M. *et al.* Mutational Analysis of Lassa Virus Glycoprotein Highlights Regions Required for Alpha-Dystroglycan Utilization. *J. Virol.* **91**, (2017).

91. Lay Mendoza, M. F., Acciani, M. D., Levit, C. N., Santa Maria, C. & Brindley, M. A. Monitoring Viral Entry in Real-Time Using a Luciferase Recombinant Vesicular Stomatitis Virus Producing SARS-CoV-2, EBOV, LASV, CHIKV, and VSV Glycoproteins. *Viruses* **12**, (2020).

92. Baird, M. F., Graham, S. M., Baker, J. S. & Bickerstaff, G. F. Creatine-kinase- and exercise-related muscle damage implications for muscle performance and recovery. *J. Nutr. Metab.* **2012**, 960363 (2012).

93. Wood, A. J. *et al.* FKRP-dependent glycosylation of fibronectin regulates muscle pathology in muscular dystrophy. *Nat. Commun.* **12**, 2951 (2021).

94. Fujimura, K. *et al.* LARGE2 facilitates the maturation of alpha-dystroglycan more effectively than LARGE. *Biochem. Biophys. Res. Commun.* **329**, 1162–1171 (2005).

95. Moore, C. J., Goh, H. T. & Hewitt, J. E. Genes required for functional glycosylation of dystroglycan are conserved in zebrafish. *Genomics* **92**, 159–167 (2008).

96. Esser, A. K. *et al.* Loss of LARGE2 disrupts functional glycosylation of α-dystroglycan in prostate cancer. *J. Biol. Chem.* **288**, 2132–2142 (2013).

97. Metze, S., Herzog, V. A., Ruepp, M.-D. & Mühlemann, O. Comparison of EJC-enhanced and EJC-independent NMD in human cells reveals two partially redundant degradation pathways. *RNA* **19**, 1432–1448 (2013).

98. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

99. Manya, H., Kuwabara, N., Kato, R. & Endo, T. FAM3B/PANDER-Like Carbohydrate-Binding Domain in a Glycosyltransferase, POMGNT1. *Methods Mol. Biol.* **2132**, 609–619 (2020).

100. Xiong, H. *et al.* Molecular interaction between fukutin and POMGnT1 in the glycosylation pathway of alpha-dystroglycan. *Biochem. Biophys. Res. Commun.* **350**, 935–941 (2006).

101. Kuwabara, N. *et al.* Carbohydrate-binding domain of the POMGnT1 stem region modulates O-mannosylation sites of α-dystroglycan. *Proc Natl Acad Sci USA* **113**, 9280–9285 (2016).

102. Lee, J. K. *et al.* Developmental expression of the neuron-specific N-acetylglucosaminyltransferase Vb (GnT-Vb/IX) and identification of its in vivo glycan products in comparison with those of its paralog, GnT-V. *J. Biol. Chem.* **287**, 28526–28536 (2012).

103. Ury, B., Potelle, S., Caligiore, F., Whorton, M. R. & Bommer, G. T. The promiscuous binding pocket of SLC35A1 ensures redundant transport of CDP-ribitol to the Golgi. *J. Biol. Chem.* **296**, 100789 (2021).

104. Esposito, D. *et al*. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).

105. Hayer, A. *et al.* Engulfed cadherin fingers are polarized junctional structures between collectively migrating endothelial cells. *Nat. Cell Biol.* **18**, 1311–1323 (2016).

106. Pettersen, E. F. *et al.* UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci*. **30**, 70–82 (2021).

107. Tesla, B. *et al*. Temperate conditions limit zika virus genome replication. *J. Virol.* **96**, e0016522 (2022).

108. Ramakrishnan, M. A. Determination of 50% endpoint titer using a simple formula. *World J. Virol.* **5**, 85–86 (2016).

Acknowledgements

We thank J. Cohen for his critical feedback on the manuscript. We thank S. Pajusalu and H. Best for their efforts which helped lead to this project. We thank N. Johnson and the GRASP LGMD consortium for granting us access to their deidentified patient data, which was funded by ML Bio Solutions. This work is supported by a grant to M.L. from the Muscular Dystrophy Association (629095) for “improved clinical interpretation of rare variants in muscle diseases”. K.P.C. is an

investigator of the Howard Hughes Medical Institute. This work was supported in part by a Paul D. Wellstone Muscular Dystrophy Specialized Research Center grant (1U54NS053672 to K.P.C.). M.A.B was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R01AI139238 (MAB). We thank Yale Flow Cytometry for their assistance with FFC and FACS services. The Core is supported in part by an NCI Cancer Center Support Grant # NIH P30 CA016359.

Author information

Authors and Affiliations

Department of Genetics, Yale School of Medicine, New Haven, CT, USA

Kaiyue Ma, Kenneth K. Ng, Shushu Huang, Nicole J. Lake, Lin Ge, Keryn G. Woodman, Katherine E. Koczwara, Vincent Ho, Christine L. O'Connor, Angela Lek & Monkol Lek.

Yale University, New Haven, CT, USA

Jenny Xu

Muscular Dystrophy Association, Chicago, IL, USA

Angela Lek

Department of Neurology, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China

Lin Ge

Howard Hughes Medical Institute, Senator Paul D. Wellstone Muscular Dystrophy Specialized Research Center, Department of Molecular Physiology and Biophysics and Department of Neurology, Roy J. and Lucille A. Carver College of Medicine, The University of Iowa, Iowa City, IA, USA

Soumya Joseph & Kevin P. Campbell

Department of Infectious Diseases, Department of Population Health, University of Georgia, Athens, GA, USA.

Melinda A. Brindley

Contributions

K.M. and M.L. conceived and designed this study. M.L. supervised the experiments and analyses of this study. K.M. designed and performed the experiments for the establishment, application, and validation of SMuRF, with the help of other authors. K.M. and S.H. performed the IF experiments. K.M. and A.L. created the lentiviral plasmid pools. K.G.W. and K.E.K. participated in the early development of this study. V.H. and C.L.O'C assisted in performing the experiments. S.J. and K.P.C. participated in the optimization of the IIH6C4-related experiments. M.A.B participated in designing and interpreting the VSV experiments. K.M., K.K.N., N.J.L. and M.L. performed computational analyses, with the help of J.X. and L.G.. K.M., K.K.N., N.J.L., J.X. and M.L. wrote this manuscript with the help of other authors.

Corresponding author

Correspondence to Kaiyue Ma and Monkol Lek.

Ethics declarations

Competing interests

The authors declare no competing interests.

Figure:

Fig. 1: Overview of variant enzymatic function characterization using SMuRF.

a, Functions of most α-DG glycosylation enzymes can be evaluated by the IIH6C4 antibody. Blue texts mark the enzymes involved in the glycosylation of α-DG Core M3 and its extension. Bold arrows link enzymes to the modifications or glycan additions they catalyze; for instance, POMK catalyzes mannose phosphorylation. Dol-P-Man is dolichol phosphate mannose; GlcNAc, N-acetylglucosamine; GalNAc, N-acetylgalactosamine; Rbo5P, ribitol-5-phosphate; Xyl, xylose; GlcA, glucuronic acid. **b,** A universal workflow of SMuRF. SMuRF accompanies saturation mutagenesis with functional assays. Here, saturation mutagenesis is achieved by delivering variant lentiviral particles to the engineered HAP1 platform where the endogenous gene of interest (GOI) was knocked out and stable *DAG1* overexpression was established through lentiviral integration. A fluorescence-activated cell sorting (FACS) assay was employed to separate the high-function population and the low-function population.

Fig. 2: SMuRF is a universal workflow to characterize SNVs of α-DG glycosylation enzymes.

a, Lenti-*GOI* constructs used for the saturation mutagenesis. The *GOI* CDS expression is driven by a weak promoter UbC. **b,** PALS-C is simple and accessible to most molecular biological laboratories. To accommodate the requirements of downstream short-read NGS, the *GOI* variants were separated into multiple blocks (6 blocks for *FKRP* and 10 blocks for *LARGE1*). The PALS-C 2-way extension cloning generates block-specific lentiviral plasmid pools from 1 oligo pool per GOI. The steps are massively multiplexed: Step1 requires only a single-tube reaction; the following steps can be done in a single-tube reaction for each block. Step8 (not shown) uses electrotransformation to deliver the assembled plasmid pools to bacteria for amplification. **c,** A representative example shows the gating strategy; 20k flow cytometry events of *FKRP* block1 were recorded and reanalyzed with FlowJo.

Fig. 3: SMuRF recapitulated and expanded the knowledge gained from population databases.

a,b, SMuRF scores align with variant types (**a,** *FKRP*; **b,** *LARGE1*): synonymous variants resemble wildtype, nonsense variants consistently have low scores, and start-loss variants exhibit markedly lower scores than nonsense variants. Noteworthy outliers include high-scoring nonsense variants at the end of coding sequences. The box boundaries represent the 25th/75th percentiles, with a horizontal line indicating the median and a vertical line marking an additional 1.5 times interquartile range (IQR) above and below the box boundaries. p-values were calculated using the two-sided Wilcoxon test. Counts of variants were labeled below the boxes. **c,d,** SMuRF revealed functional constraints based on variants reported in gnomAD v4.0.0 (**c,** *FKRP*; **d,** *LARGE1*): Low allele frequency variants had diverse functional scores, while high

allele frequency variants converged towards wildtype due to selection pressures (Gray box: AF<1.5e-05). **a-d**: dashed lines represent WT functional score.

Fig. 4: SMuRF improved the scope of clinical interpretation of genetic variants.

SMuRF scores correlate well with clinical classification in ClinVar (**a,** *FKRP*; **b,** *LARGE1*). (B/LB: Benign, Benign/Likely benign or Likely benign in ClinVar; VUS: Uncertain significance in ClinVar; P/LP: Pathogenic, Pathogenic/Likely pathogenic or Likely pathogenic in ClinVar.) Counts of variants were labeled below the boxes. Real patient data from eight well-curated cohorts demonstrated that SMuRF scores have the potential to predict disease severity. **c,** The SMuRF scores of variants associated with mild cases were significantly higher than those of intermediate and severe cases. Counts of cases were labeled below the boxes. **a-c**: Box plots depict the 25th/75th percentiles (box boundaries), median (horizontal line), and an additional 1.5 times IQR (vertical line) above and below the box boundaries. p-values were calculated using the Wilcoxon test. **d,** The SMuRF scores are correlated with the disease onset age. Dashed trendlines represent linear regression. Spearman's rank correlation rho: 0.61 (all data), 0.48 (male), 0.77 (female). **c,d**: FS1, the functional score of the variant on Allele1; FS2, the functional score of the variant on Allele2. **a-d**: blue dashed lines represent (homozygous) WT functional score. SMuRF demonstrated high performance in accurately classifying pathogenic variants. **e,f,** Receiver operating characteristic (ROC) curves of SMuRF and computational predictors: taking Pathogenic, Pathogenic/Likely pathogenic and Likely pathogenic variants in ClinVar as true positives (**e,** *FKRP*; **f,** *LARGE1*). AUC: Area Under Curve. Higher AUC indicates better performance in classifying pathogenic variants. **g,** The correlation coefficient was calculated between SMuRF scores and scores generated by computational predictors. **h,i,** Among all the predictors examined, AlphaMissense has the strongest correlation with SMuRF(rho=-0.65, FKRP; -0.45, LARGE1). Density was calculated with contour_var = "count" in R. White dashed lines represent linear regression.

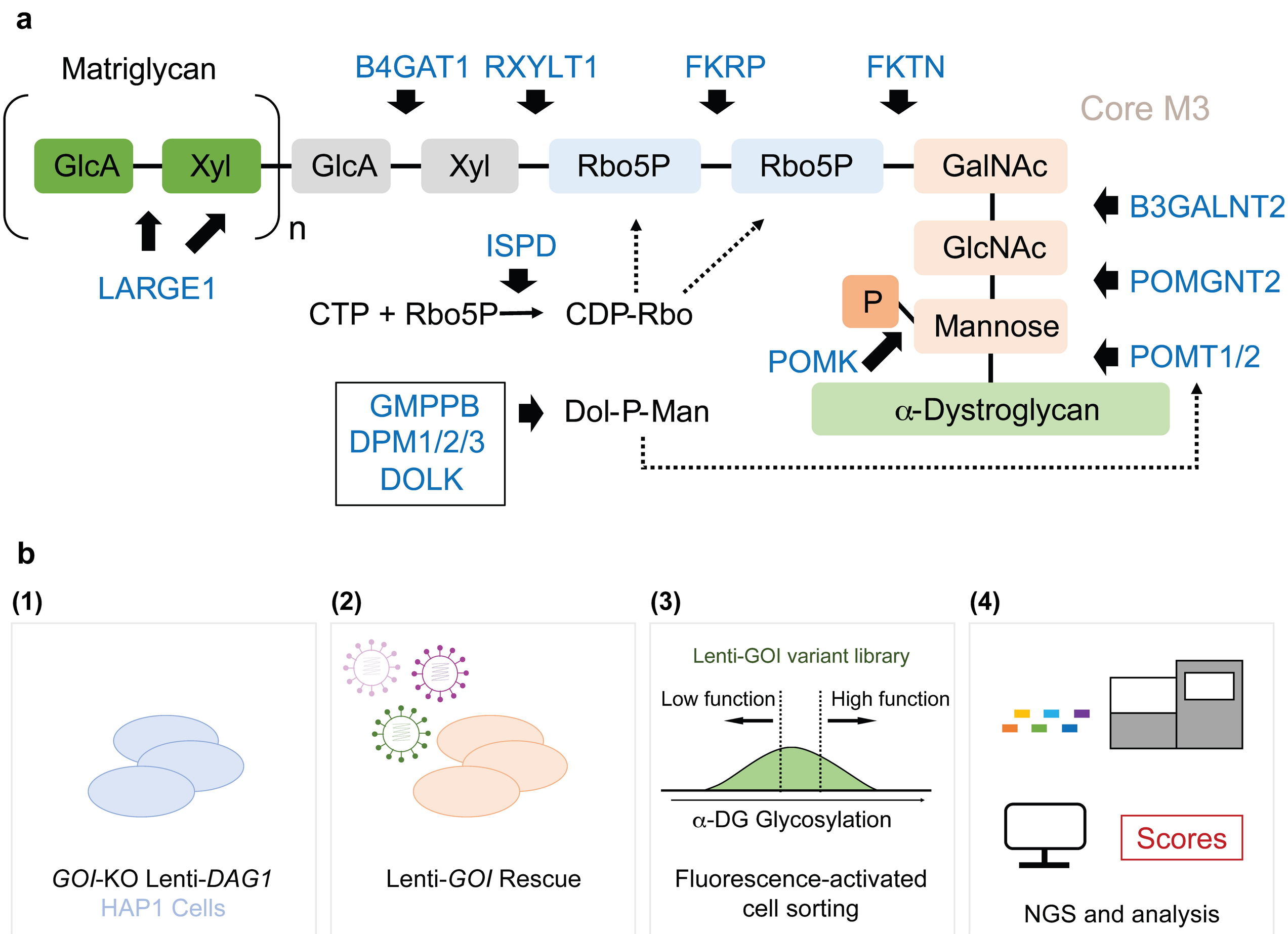
Fig. 5: SMuRF highlighted the critical structural regions.

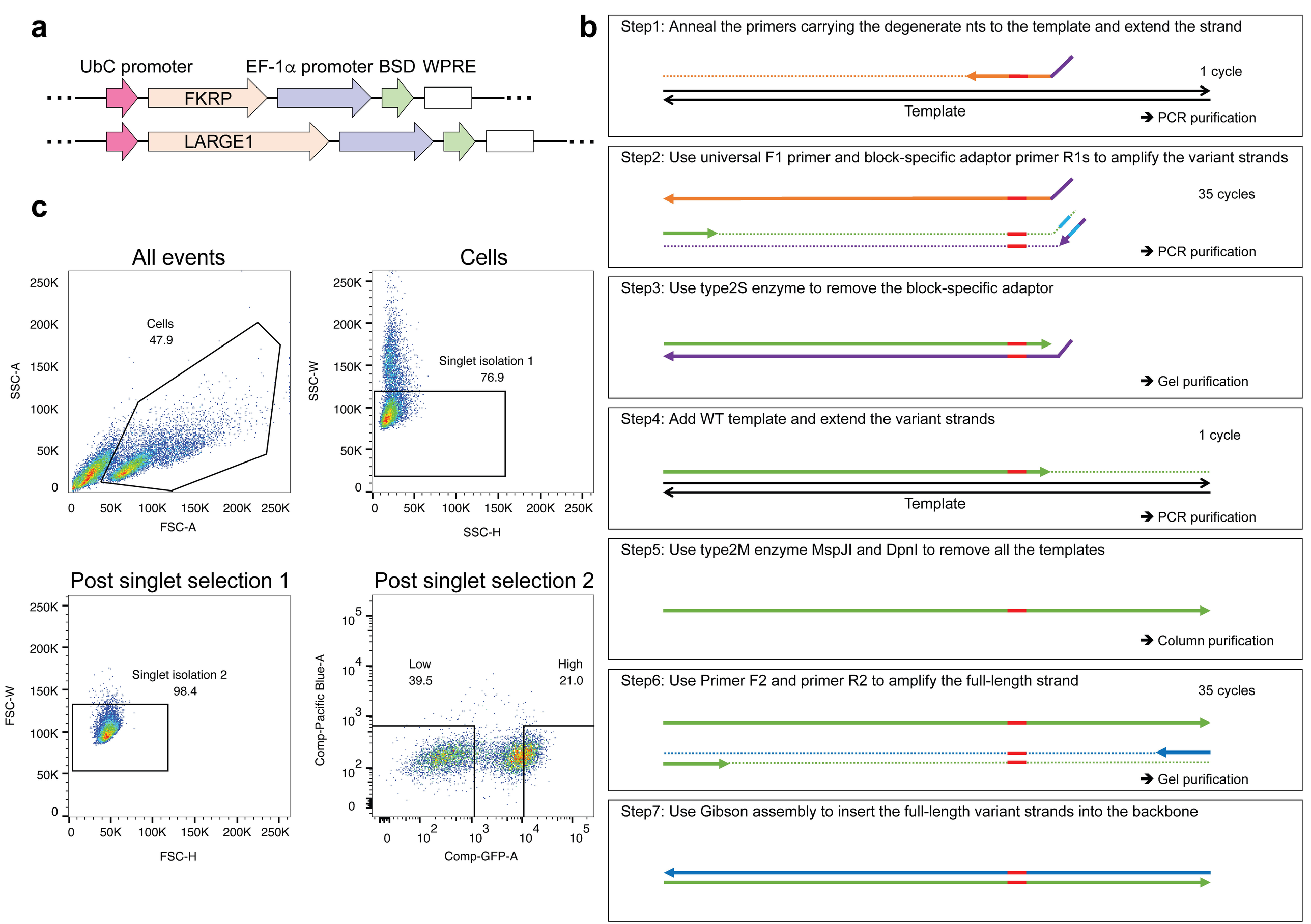
a, SMuRF scores showed higher functional disruption by missense variants in the catalytic domain of *FKRP* compared to the stem domain. The zinc finger loop within the catalytic domain exhibited greater disruption by missense variants. **c,** Missense variants in the catalytic domains of *LARGE1* showed higher disruption compared to the N-terminal domain. Missense variants in the XylT domain were more disruptive than those in the GlcAT domain. The observed domain differences were significant only for missense variants, not synonymous variants (**b,** *FKRP*; **d,** *LARGE1*). Box plots depict the 25th/75th percentiles (box boundaries), median (horizontal line), and an additional 1.5 times IQR (vertical line) above and below the box boundaries. p-values were calculated using the two-sided Wilcoxon test. Counts of variants were labeled below the boxes. Dashed lines represent WT functional score. SMuRF scores were utilized to map SNV-

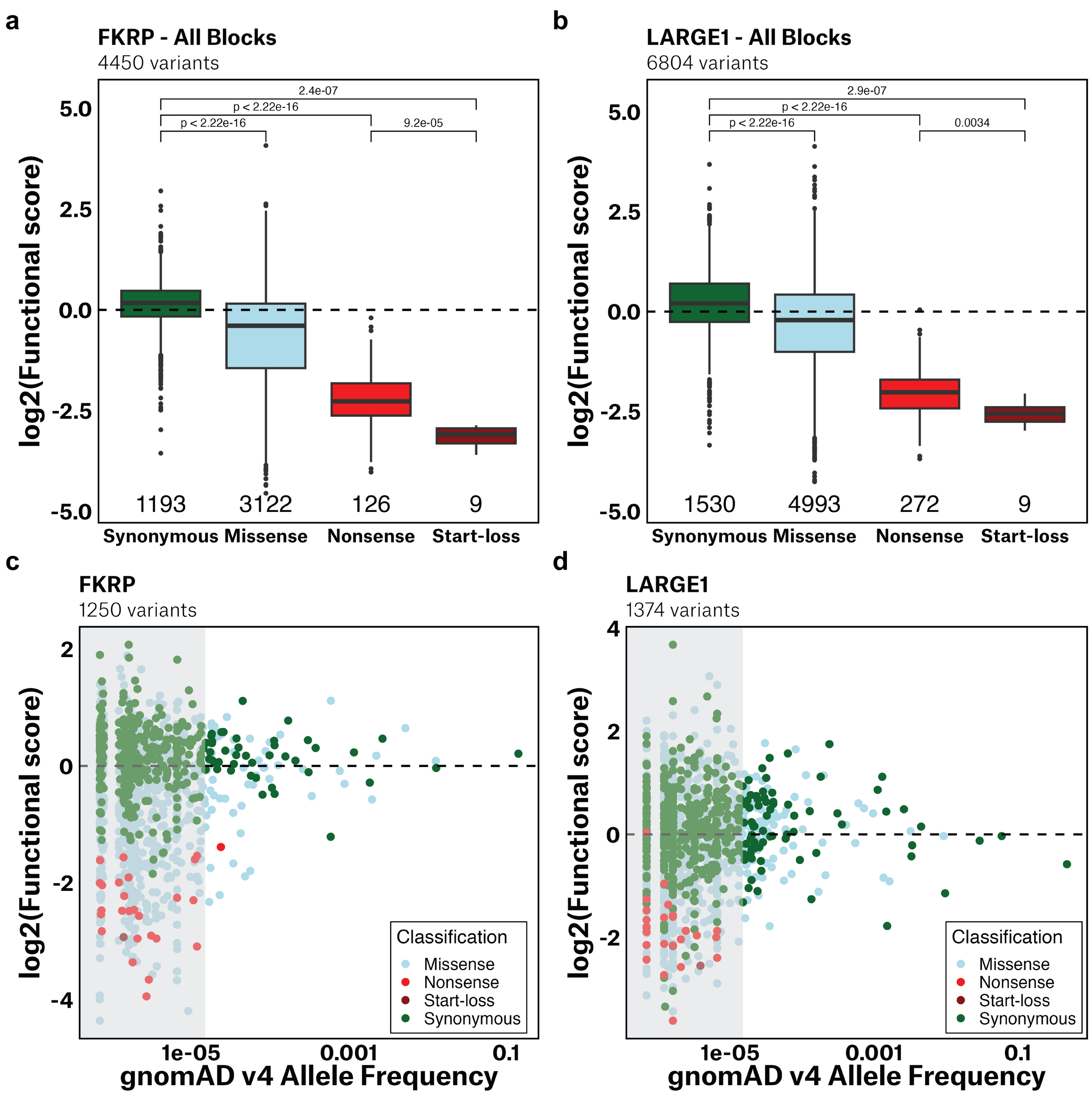
generated single amino acid substitutions (SNV-SAASs) onto the 3D structures of the enzymes (**e,** *FKRP*; **f,** *LARGE1*). The mean SMuRF score per amino acid residue was calculated and visualized using a color scale, where red indicates positions sensitive to substitutions and green is tolerated. The crystal structure of human FKRP (PDB:6KAM, codon: 45-495) and the electron microscopy structure of LARGE1 (PDB:7UI7, codon: 34-756) were used.

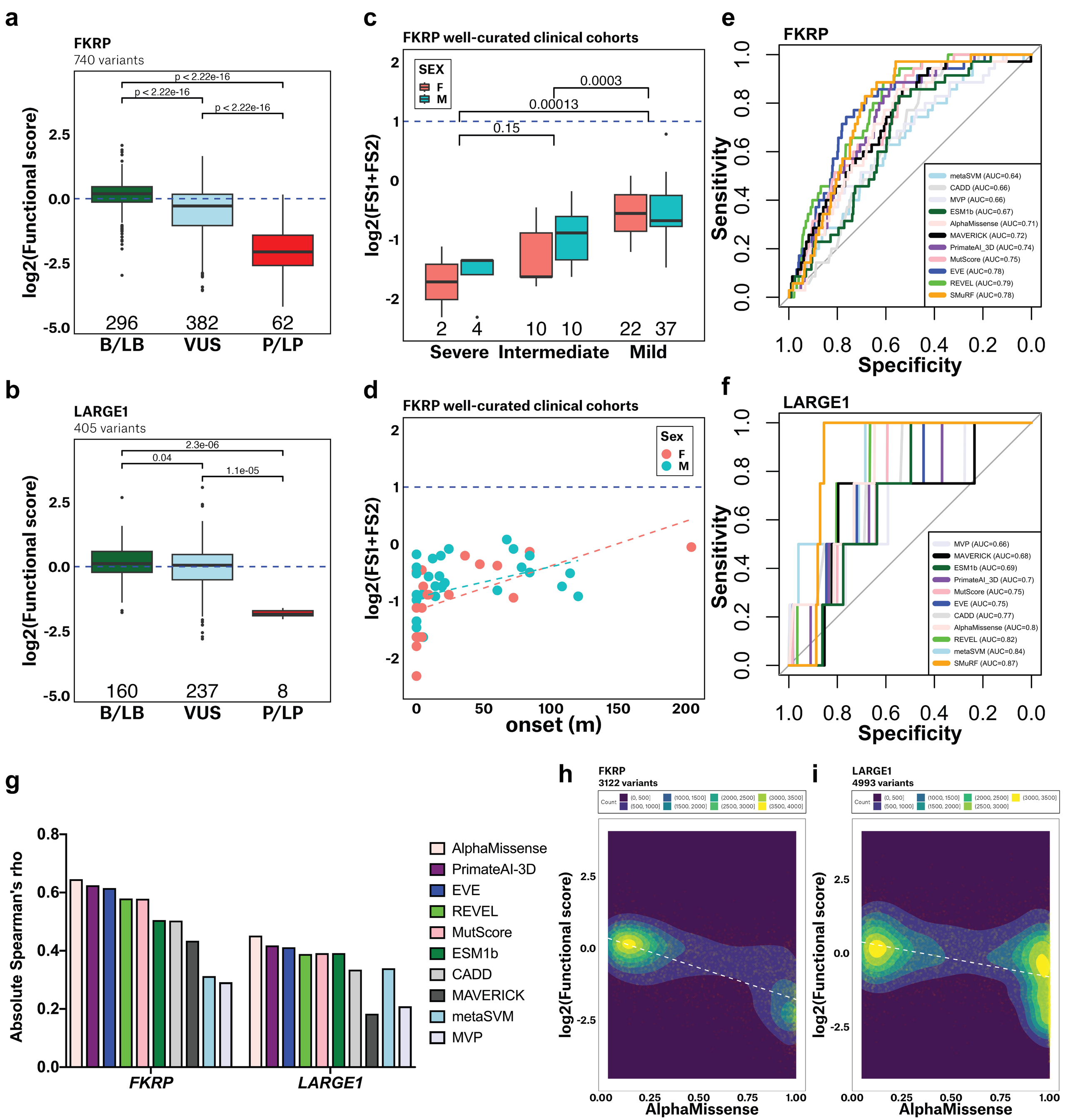
Fig. 6: Validations confirmed SMuRF findings in the myogenic context.

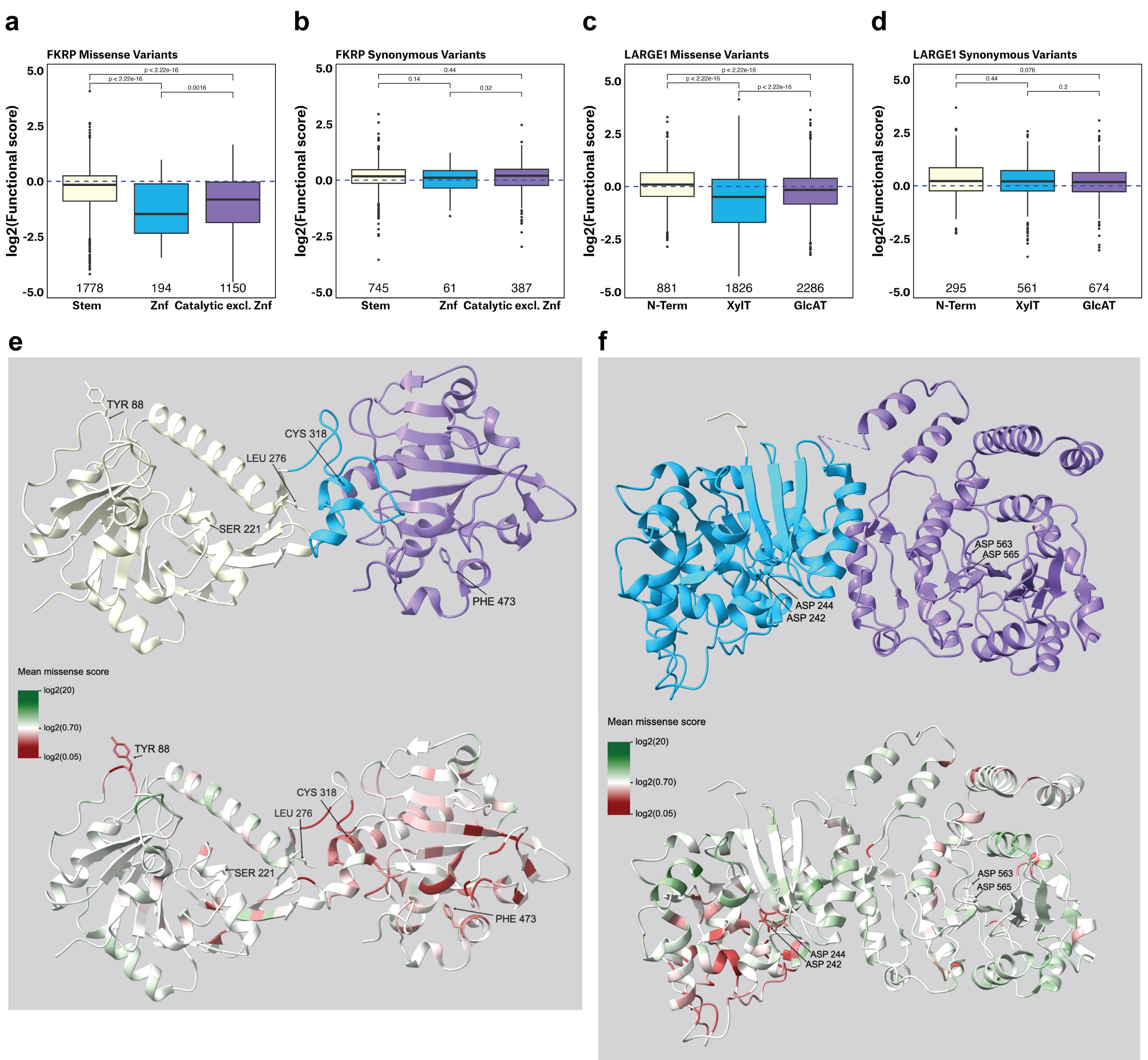
a, Validation of individual *FKRP* variants using an IIH6C4 IF assay. The myoblasts underwent transduction and drug selection, followed by differentiation into myotubes, which were subsequently used for IF. ".r" denotes lentiviral transduction of an individual variant. Blue: DAPI. Green: IIH6C4, α-DG the glycosylation level. Nine individual transductions were performed, including WT and 8 variants (Supplementary Method 7). The brightness and contrast of the photos were adjusted in Adobe Photoshop with the same settings. **b,** An orthogonal assay based on α-DG-dependent viral entry. Vesicular stomatitis virus (VSV) with Lassa fever virus glycoprotein complex (LASV-GPC) can infect cells in an α-DG-dependent manner. Variant enrichment before/after VSV infection can be used to quantify their performances regarding α-DG glycosylation. **c,d,** The ppVSV assay can be employed to validate the findings from the flow cytometry assay (c, *FKRP*; d, *LARGE1*). 10 Lenti-*FKRP* variants were mixed with Lenti-WT-*FKRP* to rescue *FKRP*-KO MB135. 11 Lenti-*LARGE1* variants were mixed with Lenti-WT-*LARGE1* to rescue *LARGE1*-KO MB135. The functional score was quantified by the ratio of a variant's enrichment in the non-infected group to its enrichment in the ppVSV-infected group. Higher functional score indicates better performance in α-DG glycosylation. *FKRP* c.135C>T (p.Ala45=) and *LARGE1* c.435C>T (p.Ala145=) have the highest AFs in gnomAD v4. *FKRP* c.663C>A (p.Ser221Arg) and *LARGE1* c.992C>T (p.Ser331Phe) are Pathogenic in ClinVar. Biological replicate N=3: Lentiviral transduction and ppVSV infection were both performed independently. Figures display mean values with SEM. Additional discussions for the ppVSV results are included in Supplementary Discussion 2-5.

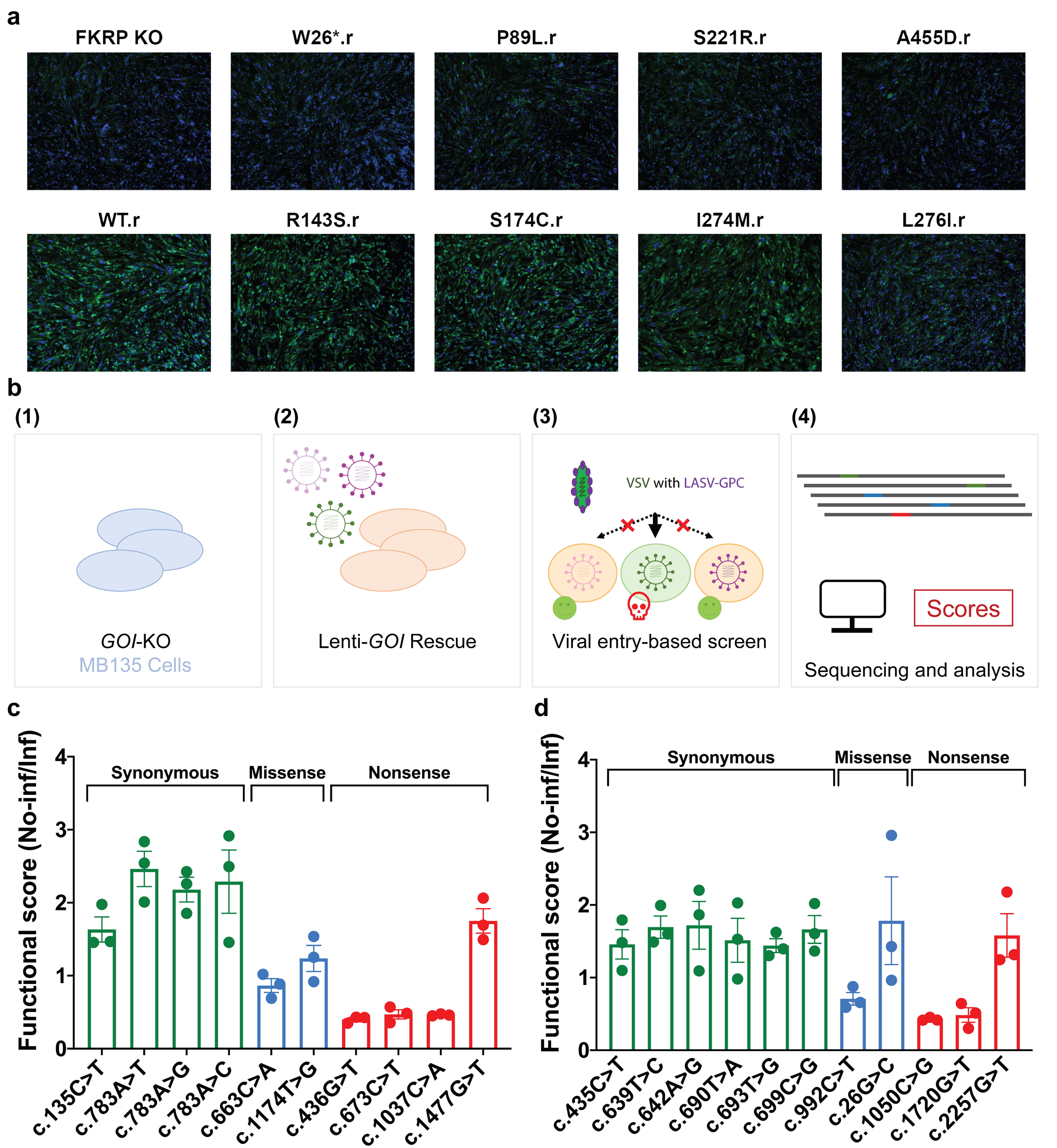

a
Matriglycan
GlcA
Xyl
n
LARGE1
B4GAT1
RXYLT1
FKRP
FKTN
GlcA
Xyl
Rbo5P
Rbo5P
Core M3
GalNAc
GlcNAc
Mannose
P
B3GALNT2
POMGNT2
POMT1/2
POMK
ISPD
CTP + Rbo5P
CDP-Rbo
GMPPB
DPM1/2/3
DOLK
Dol-P-Man
α-Dystroglycan
b
(1)
GOI-KO Lenti-DAG1
HAP1 Cells
(2)
Lenti-GOI Rescue
(3)
Lenti-GOI variant library
Low function
High function
α-DG Glycosylation
Fluorescence-activated
cell sorting
(4)
Scores
NGS and analysis


a
UbC promoter
EF-1α promoter
BSD
WPRE
FKRP
LARGE1
b
Step1: Anneal the primers carrying the degenerate nts to the template and extend the strand
1 cycle
Template
➔ PCR purification
Step2: Use universal F1 primer and block-specific adaptor primer R1s to amplify the variant strands
35 cycles
➔ PCR purification
Step3: Use type2S enzyme to remove the block-specific adaptor
➔ Gel purification
Step4: Add WT template and extend the variant strands
1 cycle
Template
➔ PCR purification
Step5: Use type2M enzyme MspJI and DpnI to remove all the templates
➔ Column purification
Step6: Use Primer F2 and primer R2 to amplify the full-length strand
35 cycles
➔ Gel purification
Step7: Use Gibson assembly to insert the full-length variant strands into the backbone
c
All events
Cells
47.9
SSC-A
FSC-A
Cells
Singlet isolation 1
76.9
SSC-W
SSC-H
Post singlet selection 1
Singlet isolation 2
98.4
FSC-W
FSC-H
Post singlet selection 2
Low
39.5
High
21.0
Comp-Pacific Blue-A
Comp-GFP-A


a
FKRP - All Blocks
4450 variants
2.4e-07
p < 2.22e-16
p < 2.22e-16
9.2e-05
log2(Functional score)
5.0
2.5
0.0
-2.5
-5.0
1193
3122
126
9
Synonymous
Missense
Nonsense
Start-loss
b
LARGE1 - All Blocks
6804 variants
2.9e-07
p < 2.22e-16
p < 2.22e-16
0.0034
log2(Functional score)
5.0
2.5
0.0
-2.5
-5.0
1530
4993
272
9
Synonymous
Missense
Nonsense
Start-loss
c
FKRP
1250 variants
log2(Functional score)
2
0
-2
-4
Classification
Missense
Nonsense
Start-loss
Synonymous
1e-05
0.001
0.1
gnomAD v4 Allele Frequency
d
LARGE1
1374 variants
log2(Functional score)
4
2
0
-2
Classification
Missense
Nonsense
Start-loss
Synonymous
1e-05
0.001
0.1
gnomAD v4 Allele Frequency


a
FKRP
740 variants
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
log2(Functional score)
2.5
0.0
-2.5
-5.0
296
382
62
B/LB
VUS
P/LP
b
LARGE1
405 variants
2.3e-06
0.04
1.1e-05
log2(Functional score)
2.5
0.0
-2.5
-5.0
160
237
8
B/LB
VUS
P/LP
c
FKRP well-curated clinical cohorts
SEX
F
M
0.0003
0.00013
0.15
log2(FS1+FS2)
2
1
0
-1
-2
2
4
10
10
22
37
Severe
Intermediate
Mild
d
FKRP well-curated clinical cohorts
Sex
F
M
log2(FS1+FS2)
2
1
0
-1
-2
0
50
100
150
200
onset (m)
e
FKRP
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
metaSVM (AUC=0.64)
CADD (AUC=0.66)
MVP (AUC=0.66)
ESM1b (AUC=0.67)
AlphaMissense (AUC=0.71)
MAVERICK (AUC=0.72)
PrimateAI_3D (AUC=0.74)
MutScore (AUC=0.75)
EVE (AUC=0.78)
REVEL (AUC=0.79)
SMuRF (AUC=0.78)
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
f
LARGE1
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
MVP (AUC=0.66)
MAVERICK (AUC=0.68)
ESM1b (AUC=0.69)
PrimateAI_3D (AUC=0.7)
MutScore (AUC=0.75)
EVE (AUC=0.75)
CADD (AUC=0.77)
AlphaMissense (AUC=0.8)
REVEL (AUC=0.82)
metaSVM (AUC=0.84)
SMuRF (AUC=0.87)
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
g
Absolute Spearman's rho
0.8
0.6
0.4
0.2
0.0
FKRP
LARGE1
AlphaMissense
PrimateAI-3D
EVE
REVEL
MutScore
ESM1b
CADD
MAVERICK
metaSVM
MVP
h
FKRP
3122 variants
Count
(0, 500]
(500, 1000]
(1000, 1500]
(1500, 2000]
(2000, 2500]
(2500, 3000]
(3000, 3500]
(3500, 4000]
log2(Functional score)
2.5
0.0
-2.5
0.00
0.25
0.50
0.75
1.00
AlphaMissense
i
LARGE1
4993 variants
Count
(0, 500]
(500, 1000]
(1000, 1500]
(1500, 2000]
(2000, 2500]
(2500, 3000]
(3000, 3500]
log2(Functional score)
2.5
0.0
-2.5
0.00
0.25
0.50
0.75
1.00
AlphaMissense


a
FKRP Missense Variants
log2(Functional score)
p < 2.22e-16
p < 2.22e-16
0.0016
1778
194
1150
Stem
Znf
Catalytic excl. Znf
b
FKRP Synonymous Variants
0.44
0.14
0.32
745
61
387
c
LARGE1 Missense Variants
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
881
1826
2286
N-Term
XylT
GlcAT
d
LARGE1 Synonymous Variants
0.076
0.44
0.2
295
561
674
e
TYR 88
CYS 318
LEU 276
SER 221
PHE 473
Mean missense score
log2(20)
log2(0.70)
log2(0.05)
f
ASP 563
ASP 565
ASP 244
ASP 242


a
FKRP KO
W26*.r
P89L.r
S221R.r
A455D.r
WT.r
R143S.r
S174C.r
I274M.r
L276I.r
b
(1)
(2)
(3)
(4)
GOI-KO
MB135 Cells
Lenti-GOI Rescue
VSV with LASV-GPC
Viral entry-based screen
Scores
Sequencing and analysis
c
Functional score (No-inf/Inf)
Synonymous
Missense
Nonsense
c.135C>T
c.783A>T
c.783A>G
c.783A>C
c.663C>A
c.1174T>G
c.436G>T
c.673C>T
c.1037C>A
c.1477G>T
d
Functional score (No-inf/Inf)
Synonymous
Missense
Nonsense
c.435C>T
c.639T>C
c.642A>G
c.690T>A
c.693T>G
c.699C>G
c.992C>T
c.26G>C
c.1050C>G
c.1720G>T
c.2257G>T