Using Saturation Mutagenesis-Reinforced Functional Assays (SMuRF) to improve the variant interpretation for alpha-dystroglycan glycosylation enzymes

Kaiyue Ma, Kenneth Ng, Shushu Huang, Nicole Lake, Jenny Xu, Lin Ge, Keryn Woodman, Katherine Koczwara, Angela Lek & Monkol Lek.

Abstract

Our inability to interpret the consequences of rare variants is an unappreciated challenge in the diagnosis of rare diseases. We developed a democratized workflow called Saturation Mutagenesis-Reinforced Functional assays (SMuRF) to inspect the direct impact variants have on enzymatic activity. We employed SMuRF to score all possible coding single nucleotide variants (SNVs) of Dystroglycanopathies-related enzyme-coding genes, *FKRP* and *LARGE1*. The utility of SMuRF scores was enhanced through the assignment of confidence scores and orthogonal assays for validation. SMuRF recapitulated and significantly expanded the knowledge gained from clinical reports and population databases, aiding in alleviating ethnic disparity in biomedical databases and improving variant classification. SMuRF expanded the training datasets of the computational predictors with the potential to improve their variant classification capability. SMuRF highlighted the critical regions in the enzyme structure which shed light on different disease mechanisms. SMuRF is the first high-throughput functional workflow to study dystroglycanopathy variants and opens the door for better variant interpretation underlying other rare diseases.

Main

Rare diseases affect approximately 3.5% – 5.9% of the worldwide population, of which 72% are genetic [1]. The emergence of gene therapies has further heightened the significance of finding diagnoses for rare genetic disease patients as an initial and crucial prerequisite towards clinical trial readiness. The recent advancement of Next-generation sequencing (NGS) and the establishment of large biobanks have improved disease-associated variant detection and interpretation [2–4]. However, for many rare disease patients, it remains challenging to identify the specific disease-causing variants, which hinders subsequent treatment development and patients' disease management and family planning [5,6]. 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 due to ultra-low population frequency [7]. The issue was worsened by the ethnic disparity in biomedical databases [8]. To complement the insufficient clinical evidence, deep mutational scanning (DMS) was proposed to unbiasedly generate functional scores for all possible variants [9]. DMS employs pooled model cells that carry large scales of variants as "patients in flasks" and characterize the variants with appropriate high-throughput functional assays [10].

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) [11]. 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. [12–14]. Most known dystroglycanopathy cases are caused by missense variants [15]. 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 (Fig. 1a and Extended Data Fig. 2) [16]. Among the enzymes, FKRP adds the second ribitol-5-phosphate (Rbo5P) to the Rbo5P tandem [17] while LARGE1 is responsible for adding the repeated disaccharide units of matriglycan [18]. Both enzymes are associated with many rare disease cases, for which novel treatments are being developed vigorously, including new drugs [19], gene therapies [20,21] and cell therapies [22]. This further emphasizes the need for improved variant interpretation to facilitate the enrollment of more patients in these gene-specific trials. However, 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 [23]. Despite the recent advancements in population studies [24–27], the challenges associated with comprehensively understanding recessive variants persist, calling for the implementation of DMS.

Hypoglycosylation is a molecular phenotype underlying dystroglycanopathies. The IIH6C4 antibody is widely used for α-DG-related research and is an effective tool for detecting this hallmark in clinical diagnoses. IIH6C4 specifically binds to the matriglycan chain of glycosylated α-DG Core M3, allowing for quantification of α-DG glycosylation levels [28]. Functions of the enzymes involved in α-DG Core M3 glycosylation have been evaluated with IIH6C4 in previous studies, including GMPPB [29], DPM1/2/3 [30], DOLK [31], POMT1/2 [32], POMK [33], POMGNT2 [34], B3GALNT2 [35], ISPD [36], FKTN [37], FKRP [38], TMEM5 [39], B4GAT1 [40] and LARGE1 [41] (Fig. 1a). Intriguingly, variants of POMGNT1, an enzyme that participates in the glycosylation of Core M1 and M2, can also perturb the IIH6C4 signal through an unknown mechanism [28], while variants of GnT-Vb/IX, an enzyme participates in the glycosylation of Core M2, are unlikely to alter IIH6C4 signal [42] (Extended Data Fig. 2). Fibroblasts from patients with dystroglycanopathies have been characterized using the IIH6C4 antibody in a fluorescence flow cytometry (FFC) assay [43]. 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 [44], enzymatic functions [33], and α-DG binding properties and functions [45]. Additionally, previous studies have established the compatibility of HAP1 cells with FFC [46]. Building upon these previous studies, we adapted the IIH6C4 FFC assay developed for fibroblasts to be applicable to HAP1 cells. Through this adaptation, we also increased the sensitivity of the assay, enabling the discrimination of differences between variants (Extended Data Fig. 3a,b).

The IIH6C4 FFC assay provided a great opportunity for us to improve variant interpretation of the α-DG glycosylation enzymes in a robust and scalable manner and meet both the needs in genetic testing and the diagnostic needs for novel trials. Building upon this assay, we developed a universal DMS workflow called Saturation Mutagenesis-Reinforced Functional Assays

(SMuRF) (Fig. 1b), and we employed SMuRF to generate functional sores for all possible coding SNVs of FKRP and LARGE1.

Results

SMuRF is a universal DMS workflow to characterize SNVs of α-DG glycosylation enzymes

SMuRF was developed with the aim of creating a universal DMS workflow that is adaptable for various genes by employing appropriate assays. Initially, it was established to characterize all possible coding SNVs for FKRP by employing the IIH6C4-based flow cytometric assay, and subsequently, its adaptability to all α-DG core M3 glycosylation enzymes was demonstrated by its application to LARGE1. The established IIH6C4 SMuRF workflow has 4 major steps:

(1) Establishing engineered cell line platforms. The endogenous gene of interest (GOI), specifically *FKRP* or *LARGE1*, was knocked-out to make *GOI*-KO HAP1 lines (Extended Data Fig. 4a,b). *DAG1* overexpression was achieved by Lenti-*DAG1* transduction (Extended Data Fig. 4c,d,e). Subsequent experiments were performed using monoclonal *GOI*-KO Lenti-*DAG1* HAP1 lines.

(2) Creating lentiviral pools of all possible coding SNVs to transduce the platform cells. To accomplish this step, we first constructed the plasmids carrying the wildtype (WT) GOIs. A key consideration here is to employ a weak promoter for the study of enzymatic activity, as over-expression may rescue the pathogenic effects of the low-function variants [47]. We employed a weak promoter UbC for GOI expression [48], creating Lenti-UbC-*FKRP*-EF1α-*BSD* and Lenti-UbC-*LARGE1*-EF1α-*BSD* (Fig. 2a). *BSD* encodes Blasticidin S deaminase (BSD), which confers blasticidin resistance in transduced cells.

Next, we performed saturation mutagenesis to introduce all possible SNVs using the WT plasmids as templates. Currently, well-established saturation mutagenesis methods include but are not limited to the insertion of variant-carrying tiles [49], multiplex homology-directed repair [50], and reversibly-terminated inosine mutagenesis [51]. However, these methods are usually subjected to one or more of the limitations, including intensive labor requirements, high expenses, disparate variant representation, and limited spanning regions. To address these issues, we developed a 2-way extension cloning method called Programmed Allelic Series with Common procedures (PALS-C), which was adapted from PALS [52]. We managed to achieve saturation mutagenesis without the special reagents and equipment required for PALS or the laborious steps required in a previous optimization for PALS [53], which makes PALS-C simple and accessible to most molecular biological laboratories (Fig. 2b). A highly detailed protocol encompassed key considerations for oligo design, experimental conditions, and our customized computational tools were provided to ensure easy and successful application of PALS-C by others (Methods and Supplementary Method 4).

We adopted a multi-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. All downstream experiments, up to NGS, were conducted individually for each block. This

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 [54].

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.

(3) Isolating lentiviral-rescued cells with high or low glycosylation levels through IIH6C4-based fluorescence-activated cell sorting (FACS). Once the transduced cells have undergone drug selection and expanded to an adequate quantity, they can be utilized in FACS. To establish the working conditions for the FACS, staining conditions and gating parameters were tested with mini-libraries to achieve optimized separation of variants with different functional levels (Extended Data Fig. 6a,b). α-DG glycosylation level was quantified by IIH6C4-FITC signal (Fig. 2c and Extended Data Fig. 6c,d), based on which cells were sorted to the high-glycosylation group and the low-glycosylation group. The FACS events of each group achieved a minimum of ~1000 × coverage.

(4) Building the NGS library and generating the SMuRF scores. Genome DNA from each group of each block was used to build the sequencing library using a 3-round PCR strategy (Extended Data Fig. 7a). Raw NGS datasets were analyzed with our customized workflow 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 high function of variants to glycosylate α-DG while low scores indicate low function. Two transduction replicates were performed for *FKRP* to confirm the reproducibility of the workflow (Extended Data Fig. 8).

SMuRF recapitulated and expanded the knowledge gained from population databases

SMuRF scores were generated for ~99.9% of all possible coding SNVs of *FKRP* (4450/4455) and 100% of *LARGE1* (6804/6804). The variants missing from the *FKRP* pools were: c.279G>C, c.430A>C, c.432G>C, c.439G>C and c.454A>C.

SMuRF scores align with the anticipated patterns of different variant types (Fig. 3a,b). The SMuRF score of the WT was set to 0. The synonymous variants display scores that closely approximate the WT score, 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) [55]. Notably, 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 [56]. 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 [14]. The multi-exon CDS structure of *LARGE1* may confer genetic compensations for start-loss variants that are undetectable by SMuRF [57]. However, in the case of *FKRP*, where there is only one coding exon, the SMuRF scores effectively indicate that start-loss variants pose a high risk of being highly damaging. Therefore, these variants warrant increased attention in genetic testing protocols.

Allele frequency refers to the relative frequency of a genetic variant at a specific chromosomal locus within a population. When combined with allele frequency data obtained from population databases, SMuRF scores were consistent with the selection against pathogenic variants. The Genome Aggregation Database (gnomAD) is a resource that aggregated and harmonized both exome and genome sequencing data from a wide variety of large-scale sequencing projects [3]. The SMuRF scores of the *FKRP* and *LARGE1* variants reported in the large population database gnomAD v3.1.2 [58] were examined. Low allele frequency variants (Allele count=1 or 2) exhibited a wide range of functional scores, while variants with higher frequency showed functional scores converging towards the WT score due to the selective exclusion of pathogenic variants from the population (Fig. 3c,d). As one of the largest population databases, gnomAD currently provides reports for only 374 *FKRP* coding SNVs (8.4%) and 426 *LARGE1* coding SNVs (6.3%). The ability of SMuRF scores to recapitulate the patterns observed in gnomAD makes them a significant expansion to gnomAD.

The α-DG glycoepitope, as well as the enzymes involved in its glycosylation, are largely conserved within Metazoa. Orthologous sequence similarities of both human *FKRP* and *LARGE1* can be identified in organisms as primitive as choanoflagellates [59]. Evolutionary conservation scores, such as PhyloP scores, indicate the degree of conservation of a variant derived from multiple sequence alignments across species [60]. 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 (Spearman's rho=-0.38, *FKRP*; -0.21, *LARGE1*).

SMuRF improved the scope of clinical interpretation of rare variants and provided good training datasets for computational predictors

ClinVar is a public archive of reports of human variants [61], where the variants were classified according to clinical supporting evidence into different categories including: Benign(B),

Benign/likely benign (B/LB), Likely benign(LB), Pathogenic(P), Pathogenic /likely pathogenic (P/LP), Likely pathogenic (LP), and Variants of Uncertain Significance (VUS). SMuRF scores correlate well with clinical classification in ClinVar, with B, B/LB, and LB variants having scores close to the WT score while P, P/LP, and LP variants having low scores (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) [11,14]. 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 combined through addition to calculate the biallelic functional score. Initially, we conducted the analysis using the disease conditions reported in ClinVar, but no significant pattern was observed (Supplementary Fig. 3). We believe this is likely due to the suboptimal accuracy in the ClinVar reports. We came to realize that well-curated reports are essential for such analysis. Hence, we aggregated data from 8 well-curated cohorts and compared them with SMuRF scores [14,15,38,62–66]. The functional scores of the variants associated with mild cases were significantly higher compared to those of the intermediate and severe cases (Extended Data Fig. 9c). Additionally, SMuRF scores showed a correlation with the reported disease onset age (Extended Data Fig. 9d), where high-function variants were associated with later onset (Spearman's rho=0.61; 0.48, male; 0.77, female).

These analyses above indicate the potential utility of the SMuRF scores for improving variant interpretation. The SMuRF scores can be used to aid in the classification of VUSs and improve the interpretation of LP and LB variants. Additionally, the knowledge gained from well-curated reports can be used to predict the severity of the variants. However, it is important to acknowledge the presence of technical outliers in the SMuRF scores, which primarily arise from inherent limitations of PALS-C and coverage issues in the FACS method. To enhance credibility, we have introduced confidence scores in the reclassification process (Supplementary Table 1,2 and Supplementary Fig. 9,10). Our ultimate goal is to utilize SMuRF reclassification as an additional line of evidence in clinical variant interpretation, thereby aiding in clinical decision-making.

In addition to assisting in variant re-classification, SMuRF scores can also be used to validate and improve computational predictors. Computational prediction is currently the most scalable and cost-effective method to interpret and predict the functional impact of all novel variants discovered. It is an active area of research with a wealth of methods recently developed, including CADD [67], metaSVM [68], REVEL [69], MVP [70], EVE [71] and MutScore [72]. 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. 4c,d). 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. The predictors assigned higher scores to pathogenic variants, hence are negatively

correlated with SMuRF scores. EVE scores are derived from an evolutionary model trained with sequences from over 140k species and reported a high concordance with functional assays [71]. Indeed, EVE, among the predictors we tested, has the highest correlation coefficient with SMuRF (Spearman's rho=-0.62, *FKRP*; -0.41, *LARGE1*) (Extended Data Fig. 9e,f and Supplementary Table 3). REVEL, among the predictors, demonstrates the best performance according to the ROC curves and also exhibits a relatively good correlation with SMuRF (Spearman's 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). As there are only limited known pathogenic variants, no clear pattern has been concluded to identify critical structural regions and associate them to 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 and a catalytic domain [73]. 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 [73]. 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 and a glucuronate transferase (GlcAT) domain [74]. 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 variants in the XylT domain tend to be more disruptive than those in the GlcAT domain (p-values<2.22e-16). 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 [74]. Together, these observations suggest that the addition of a single xylose to α-DG is sufficient to be detected by IIH6C4, albeit generating a very weak signal. It is uncertain how much physiological function can be achieved by this single xylose in comparison to the complete matriglycan chain. Again, the differences in SMuRF scores between domains were observed exclusively in missense variants and not in synonymous variants (p-values>0.05) (Fig. 5d).

We further mapped the SMuRF scores of SNV-accessible 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 structural significance of the critical regions. 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 [75]. A p.Cys318Tyr variant (SMuRF=-2.12) has been reported to be associated with WWS [65]. 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 [75]. 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 situated at the subunit-subunit interface involved in FKRP tetramerization *in vivo*, and p.Leu276 (log2 mean missense =0.35), which interacts with the catalytic domain [73]. P.Tyr88Phe is likely associated with disease [76], and has a low SMuRF score (-3.52). p.Ser221Arg was associated with CMD-MR (MR: mental retardation) [77]. 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 [78], which is commonly associated with milder symptoms [79]. 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 [74]. It is important to note that at a specific amino acid site, SNV-SAASs can involve substitutions from one amino acid to another with similar or different biochemical properties. The co-occurrence of these scenarios can moderate the mean SMuRF score of this site. Variants affecting different enzyme domains may require distinct treatment approaches [80]. SMuRF, by highlighting critical regions in different domains, can assist in selecting appropriate treatments for different variants.

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 [81]. To address this issue, we generated a myogenic platform cell line by engineering MB135, a human control myoblast cell line [82]. Endogenous *FKRP* or *LARGE1* were knocked out respectively in the MB135 cell line. Monoclonal 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 founder mutation Leu276Ile displayed an intermediate α-DG glycosylation signal, lying between the benign variants and the other pathogenic variants.

Additionally, we explored an orthogonal assay to further validate SMuRF results. Proper glycosylation of α-DG is crucial for the viral entry of Lassa fever virus (LASV) [83]. 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 [84]. 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 [44].

We adapted this methodology and utilized rVSV-LASV-GPC/ppVSV-LASV-GPC to infect Lenti-*GOI* variant pool-rescued *GOI*-KO MB135 myoblasts (Fig. 6b). The rVSV genome incorporates the LASV-GPC coding sequence, allowing it to re-enter cells. In contrast, the ppVSV lacks this capability as it is pseudotyped using a LASV-GPC plasmid (Methods). Likely due to the re-entering feature of rVSV, the rVSV screen did not yield meaningful results (Supplementary Fig. 6a). The ppVSV screen for both FKRP and LARGE1 showed a tendency where start-loss variants were the most enriched in the infected group, suggesting a higher disruptive effect on α-DG glycosylation. Nonsense variants were the next most enriched, followed by missense variants, while synonymous variants were the least enriched (Supplementary Fig. 6b,c). This tendency aligns with what we observed in the SMuRF FACS assay. However, the ppVSV assay, in general, lacks the sensitivity to distinguish differences among the variants.

Discussion

The lack of definitive diagnoses for rare disease patients is a significant challenge faced in clinical practice. The presence of VUSs in patient genes poses a challenge for clinicians in establishing the disease-causing gene and making informed decisions regarding the suitability of gene-specific treatments. Consequently, such patients are often excluded from receiving appropriate treatments or participating in clinical trials involving novel therapeutics. In this context, the implementation of SMuRF can be beneficial. SMuRF scores can serve as an additional line of evidence to support clinical variant interpretation, aiding in the diagnostic process.

However, in clinical practice, the majority of patients exhibit compound mutations, which raises the need for further investigation on how to apply SMuRF scores in diagnosing such cases. In this study, we employed 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. While this model demonstrated a promising correlation between SMuRF scores and disease severity (Extended Data Fig. 9c,d), further assessments may be needed. Additionally, our results emphasized the significance of well-curated reports in predicting disease severity. Our findings revealed a correlation between the disease onset age and SMuRF scores, while 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 [85].

An important consideration in the clinical implementation of the SMuRF score is that the IIH6C4 assay may only capture one of the multiple functions associated with a gene. For example, FKRP is also known to participate in the glycosylation of fibronectin [86]. Hence, a gene variant may

have a limited impact on function in the context of a specific assay but may have potentially damaging effects in another assay. Furthermore, the clinical implementation of the SMuRF score poses challenges in three additional aspects. Firstly, 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 [59], may act as an effective modifier in *LARGE1*-related diseases [87–89]. Secondly, the SMuRF score for nonsense mutations 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, further eliminating any residual functions that the truncated proteins may possess [90]. Lastly, the IIH6C4 assay used in SMuRF may detect technical signals without physiological significance. This is particularly relevant for LARGE1 as LARGE1 is directly involved in the formation of matriglycan, the target of IIH6C4. Specifically, the difference between *LARGE1* missense variants in the GlcAT domain and the XylT domain may be partly technical (Fig. 5c).

SMuRF effectively recapitulated the selection against pathogenic variants by utilizing AF data from gnomAD v3.1.2. A noteworthy aspect of the population study is that the variants favored by selection may not necessarily correspond to the ones that confer optimal enzymatic activity. 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 [74,91].

The lentiviral expression level is essential in the SMuRF workflow. Its significance lies in the fact that for certain variants that compromise the enzymatic activity, the negative impact can be counteracted through the over-expression of the enzyme [47]. This was demonstrated during the early development stage of SMuRF, where we initially attempted to use the EF-1α core promoter to drive the *FKRP* expression and failed to achieve the expected separation (Supplementary Fig. 1). This aspect is especially relevant in the context of the lentiviral expression system, where the lentiviral expression level can vary from the endogenous level, resulting from the complex effects of both copy numbers and promoter strength. The impact of the Multiplicity of Infection (MOI) of the lentiviral pool on the outcome of the functional characterization is profound (Supplementary Fig. 7). In the SMuRF workflow, the MOI was controlled to ensure that the lentiviral GOI RNA level is comparable to the endogenous level in WT cells. The titer was determined in pre-experiments. Future improvement to the workflow entails identifying and employing the endogenous promoter core element as a replacement for the UbC promoter [92], or employing saturation mutagenesis methods that can introduce variants to the correct genomic context [93].

Synonymous variants are typically considered to have limited effects on gene function [94]. In our study, we adopted the assumption that synonymous variants have no effects when calculating our confidence scores. However, it is possible that some synonymous variants can affect RNA motifs, structure, or splicing, leading to potential changes in gene function [95–97]. While evaluating the effects of variants on splicing is challenging within the SMuRF framework, it is possible that the effects of synonymous variants on RNA motifs can be captured by SMuRF. Indeed, we identified two synonymous variants in *LARGE1*, c.639T>C (SMuRF=1.18) and c.642A>G (SMuRF

score=1.22), which may have gain-of-function effects by removing a poly(A) motif from the coding sequence (Supplementary Fig. 8). The poly(A) signal is associated with transcription termination, and its removal may increase the level of functional transcripts [98]. However, despite this poly(A) motif is not separated by exon junctions *in vivo*, it is still important to note that this observation may not fully apply to the endogenous *LARGE1* due to regulatory mechanisms unique to the correct genomic context.

One of the motivations to develop SMuRF was to democratize DMS so that modestly funded laboratories that are experts in particular disease genes can contribute to the community effort to achieve an "Atlas of Variant Effects". SMuRF achieved this democratization with inexpensive reagents, commonly used laboratory equipment, and open-source software for data analysis (Supplementary Methods). We have provided comprehensive protocols for implementing SMuRF on *FKRP* and *LARGE1*, with the aim of supporting other researchers in applying this approach to investigate additional genes. SMuRF can be readily adapted for studying other enzymes involved in α-DG glycosylation. In this manuscript, we employed SMuRF to analyze all possible coding SNVs of *FKRP* and *LARGE1*, as this type of variant is the most common causal variant observed in dystroglycanopathies. However, it is worth noting that the PALS-C saturation mutagenesis method developed for SMuRF can be applied to small-sized variants beyond SNVs, including SAAS and small insertions or deletions. Moreover, integrating PALS-C with other functional assays specific to other genes or regulatory elements can further enhance the versatility of SMuRF.

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%). 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. 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), 2 µg pMD2.G (Addgene, 12259) and 9 µg lentiviral plasmid 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 suppled 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 dosage 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, 1e3-1e4 GV×µL of lentivirus was used for each block. For each block, 600k HAP1 cells or 200k MB135 cells were plated in a well of a 6-well plate. 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

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 wildtype 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 for 30 mins before 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 for 30 mins before 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 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.

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. 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 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 is < 0.05, "MEDIUM" confidence if their rank score is ≥ 0.05 and < 0.50, and "HIGH" confidence if their rank score is above 0.50.

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 infected 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 infected with ppVSV-LASV-GPC-Generation1 to produce ppVSV-LASV-GPC-Generation2, reducing residual VSV-G in the pseudotyped particles. The experiments utilized ppVSV-LASV-GPC-Generation2. The MOI of the ppVSV was determined as described previously [99]. 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 a concentration of 2e5 TCID50/mL. 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 3, and the infected cells were recovered to ~1M prior to harvesting.

Statistics

Wilcoxon tests were performed with the “ggsignif” R package. Spearman’s rank correlation coefficients were calculated with the “cor.test” function in R.

Data availability

Scores generated in this script can be found in Supplementary Table 4 (*FKRP*) and Supplementary Table 5 (*LARGE1*). Plasmids were made available on Addgene. FFC and FACS datasets were made available on FlowRepository. NGS raw data were deposited to 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. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
2. Boycott, K. M. *et al.* A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell* **177**, 32–37 (2019).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
5. 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).
6. Libell, E. M. *et al.* The outcomes and experience of pregnancy in limb girdle muscular dystrophy type R9. *Muscle Nerve* **63**, 812–817 (2021).
7. Starita, L. M. *et al.* Variant interpretation: functional assays to the rescue. *Am. J. Hum. Genet.* **101**, 315–325 (2017).

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

9. 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).

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

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

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

13. 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).

14. 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).

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

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

17. 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).

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

19. 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).

20. 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).

21. 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).

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

23. 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).

24. 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).

25. 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).

26. 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).

27. 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).

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

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

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

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

32. 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).

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

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

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

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

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

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

39. 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).

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

41. 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).

42. 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).

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

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

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

46. 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).

47. 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).

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

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

50. 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).

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

52. 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).

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

54. 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).

55. 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).

56. 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).

57. Kimura, K. *et al.* Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).

58. 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.

59. 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).

60. 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).

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

62. 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).

63. Unnikrishnan, G. *et al.* Phenotype Genotype Characterization of FKRP-related Muscular Dystrophy among Indian Patients. *J. Neuromuscul. Dis.* (2023) doi:10.3233/JND-221618.

64. 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).

65. 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).

66. 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).

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

68. 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).

69. 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).

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

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

72. 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).

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

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

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

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

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

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

79. 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).

80. 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).

81. 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).

82. 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.

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

84. 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).

85. 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).

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

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

88. 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).

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

90. 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).

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

92. Pontiller, J., Gross, S., Thaisuchat, H., Hesse, F. & Ernst, W. Identification of CHO endogenous promoter elements based on a genomic library approach. *Mol. Biotechnol.* **39**, 135–139 (2008).

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

94. Dhindsa, R. S. *et al.* A minimal role for synonymous variation in human disease. *Am. J. Hum. Genet.* **109**, 2105–2109 (2022).

95. Mueller, W. F., Larsen, L. S. Z., Garibaldi, A., Hatfield, G. W. & Hertel, K. J. The silent sway of splicing by synonymous substitutions. *J. Biol. Chem.* **290**, 27700–27711 (2015).

96. Gaither, J. B. S. *et al.* Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population. *Gigascience* **10**, (2021).

97. Zeng, Z., Aptekmann, A. A. & Bromberg, Y. Decoding the effects of synonymous variants. *Nucleic Acids Res.* **49**, 12673–12691 (2021).

98. Zhang, Y. *et al.* Alternative polyadenylation: methods, mechanism, function, and role in cancer. *J. Exp. Clin. Cancer Res.* **40**, 51 (2021).

99. Willard, K. A., Alston, J. T., Acciani, M. & Brindley, M. A. Identification of residues in lassa virus glycoprotein subunit 2 that are critical for protein function. *Pathogens* **8**, (2018).

Acknowledgements

We thank the members of the Lek laboratory for their critical feedback on the manuscript. We thank S. Pajusalu and H. Best for their efforts during the early stage of this project. We thank V. Ho and C. O'Connor for their support in lab logistics. We thank K. Campbell and the Campbell laboratory for providing the IIH6C4 antibody and many suggestions for the project. We thank M. Brindley and the Brindley laboratory for providing the VSV materials and many suggestions for the project. 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". 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 Ng, Shushu Huang, Nicole Lake, Lin Ge, Keryn Woodman, Katherine Koczwara, 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

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.W. and K.K. participated in the early development of this study. K.M., K.N., N.L. and M.L. performed computational

analyses, with the help of J.X. and L.G.. K.M., K.N., N.L., J.X. and M.L. wrote this manuscript with the help of other authors.

Corresponding author

Correspondence to Monkol Lek.

Ethics declarations

Competing interests

The authors declare no competing interests.

Figure Captions:

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, glucronic 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. **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 expended the knowledge gained from population databases.

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 Wilcoxon test. SMuRF revealed functional constraints based on variants reported in gnomAD v3.1.2 (**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: Allele Count=1 or 2).

Fig. 4: SMuRF improved the scope of clinical interpretation of rare 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.) 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. Receiver operating characteristic (ROC) curves of SMuRF and computational predictors: taking Pathogenic, Pathogenic/likely pathogenic and Likely pathogenic variants in ClinVar as true positives (**c,** *FKRP*; **d,** *LARGE1*). AUC: Area Under Curve. Higher AUC indicates better performance in classifying pathogenic variants.

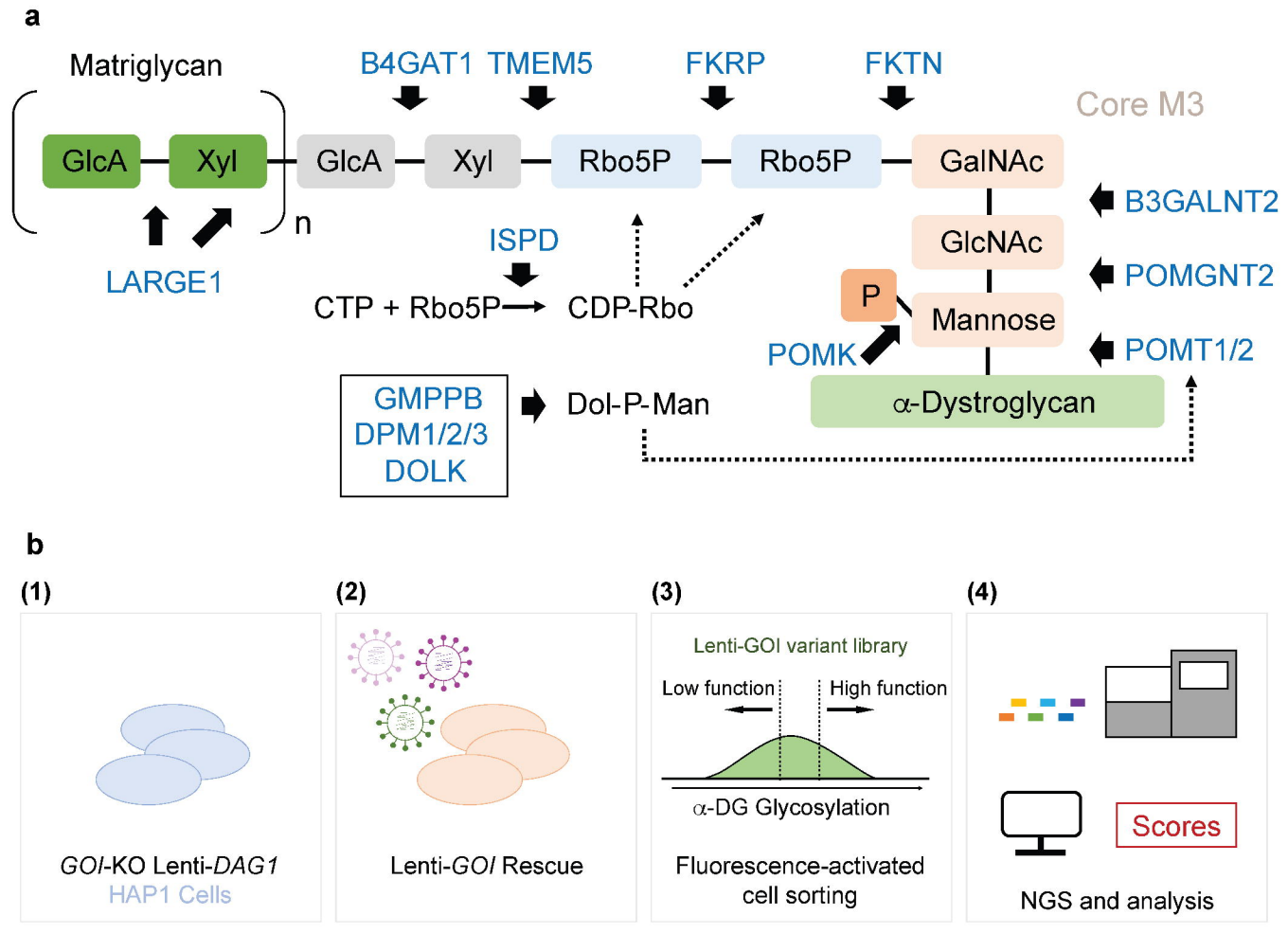
Fig. 5: SMuRF highlighted the critical structural regions.

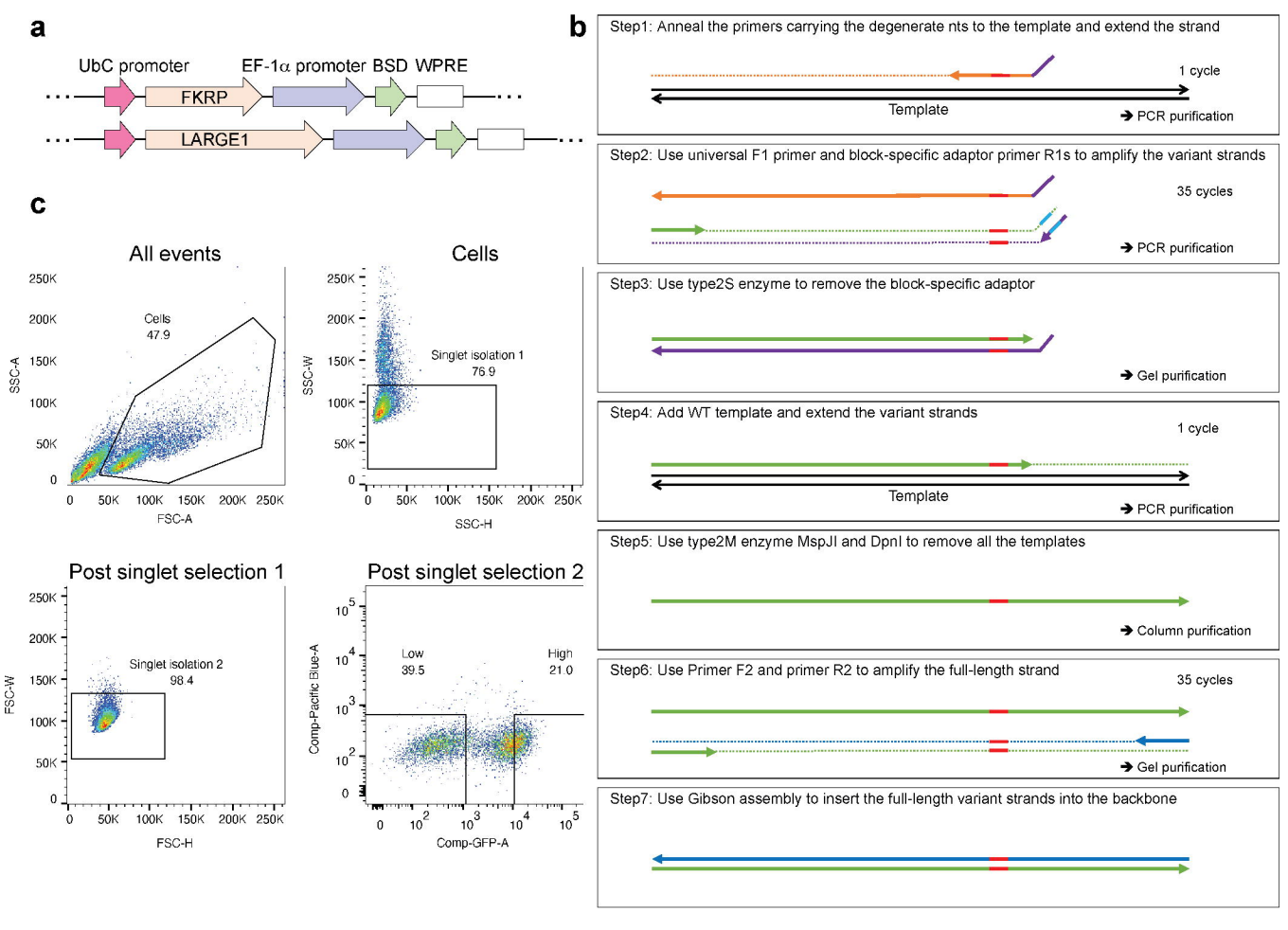
a, SMuRF scores showed higher functional disruption by missense variants in the catalytic domain of *FKRP* compared to the stem domain. The zinc finger loop within the catalytic domain exhibited greater disruption by missense variants. **c,** Missense variants in the catalytic domains of *LARGE1* showed 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 Wilcoxon test. SMuRF scores were utilized to map SNV-accessible 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 are tolerated.

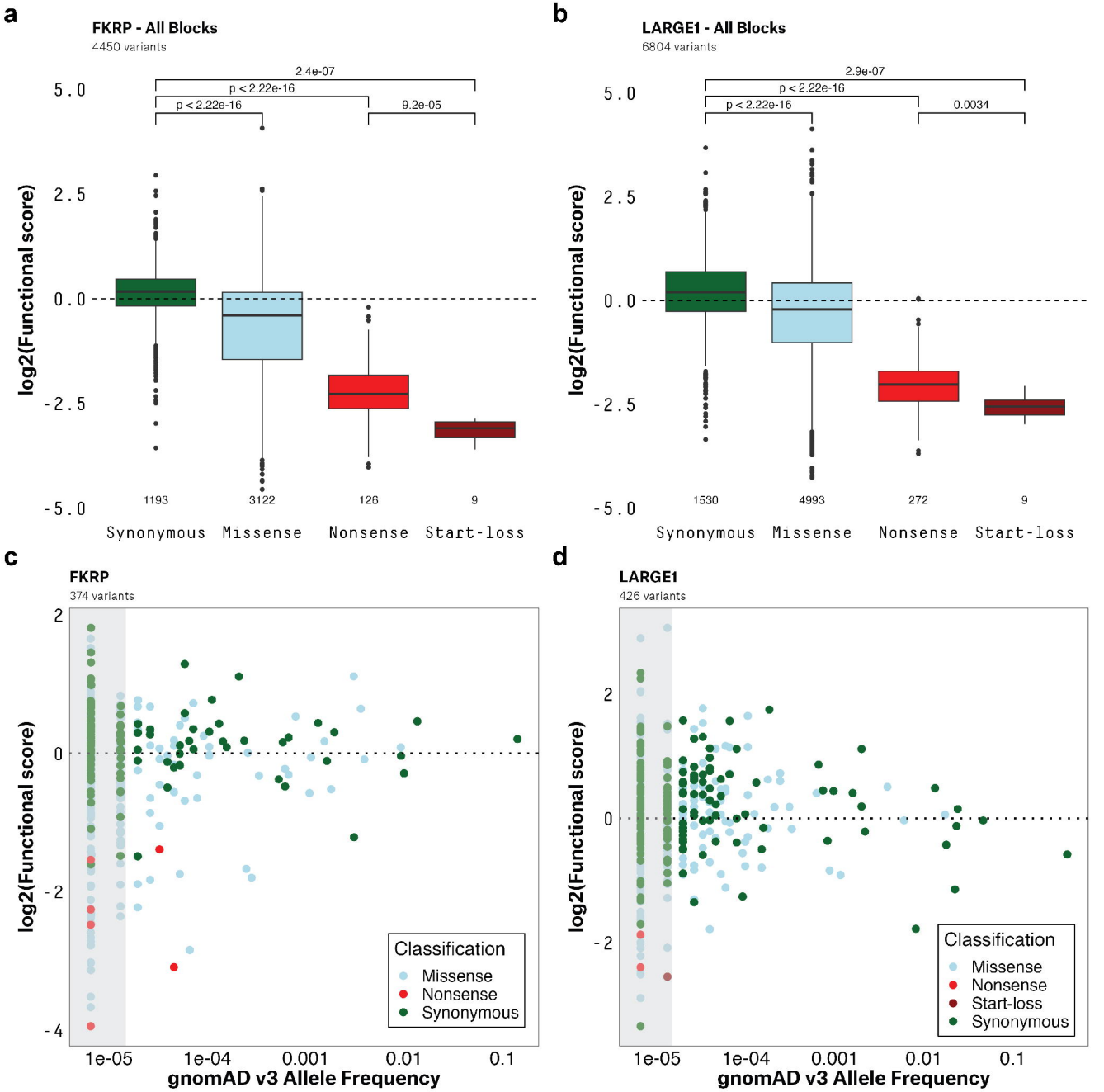
Fig. 6: Validations confirmed SMuRF findings in the myogenic context.

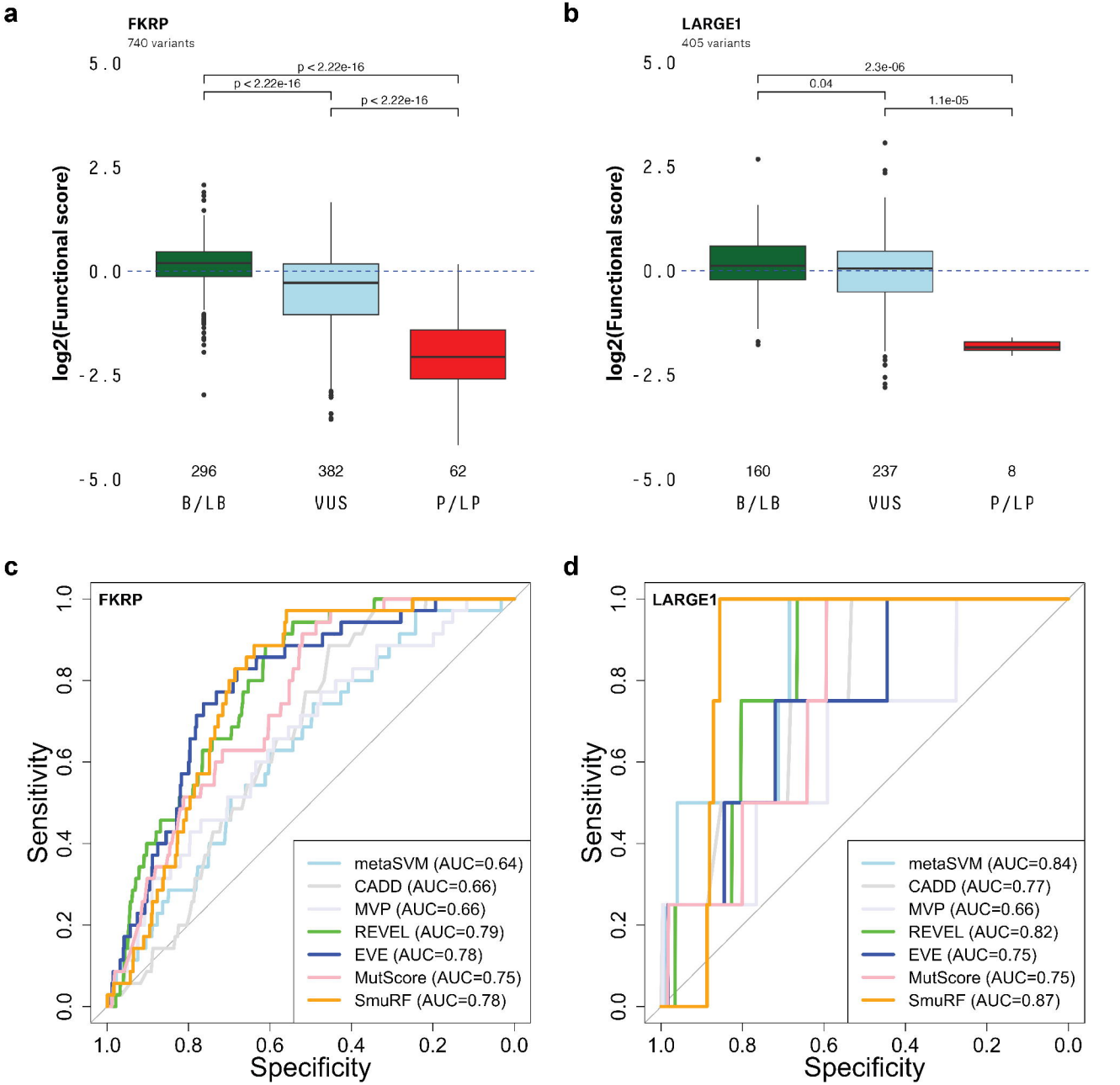
a, Validation of individual *FKRP* variants using an IIH6C7 IF assay. The myoblasts underwent transduction and drug selection, followed by differentiation into myotubes, which were subsequently used for IF. ".r" denotes lentiviral transduction of an individual variant. Blue: DAPI. Green: IIH6C4, α-DG 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

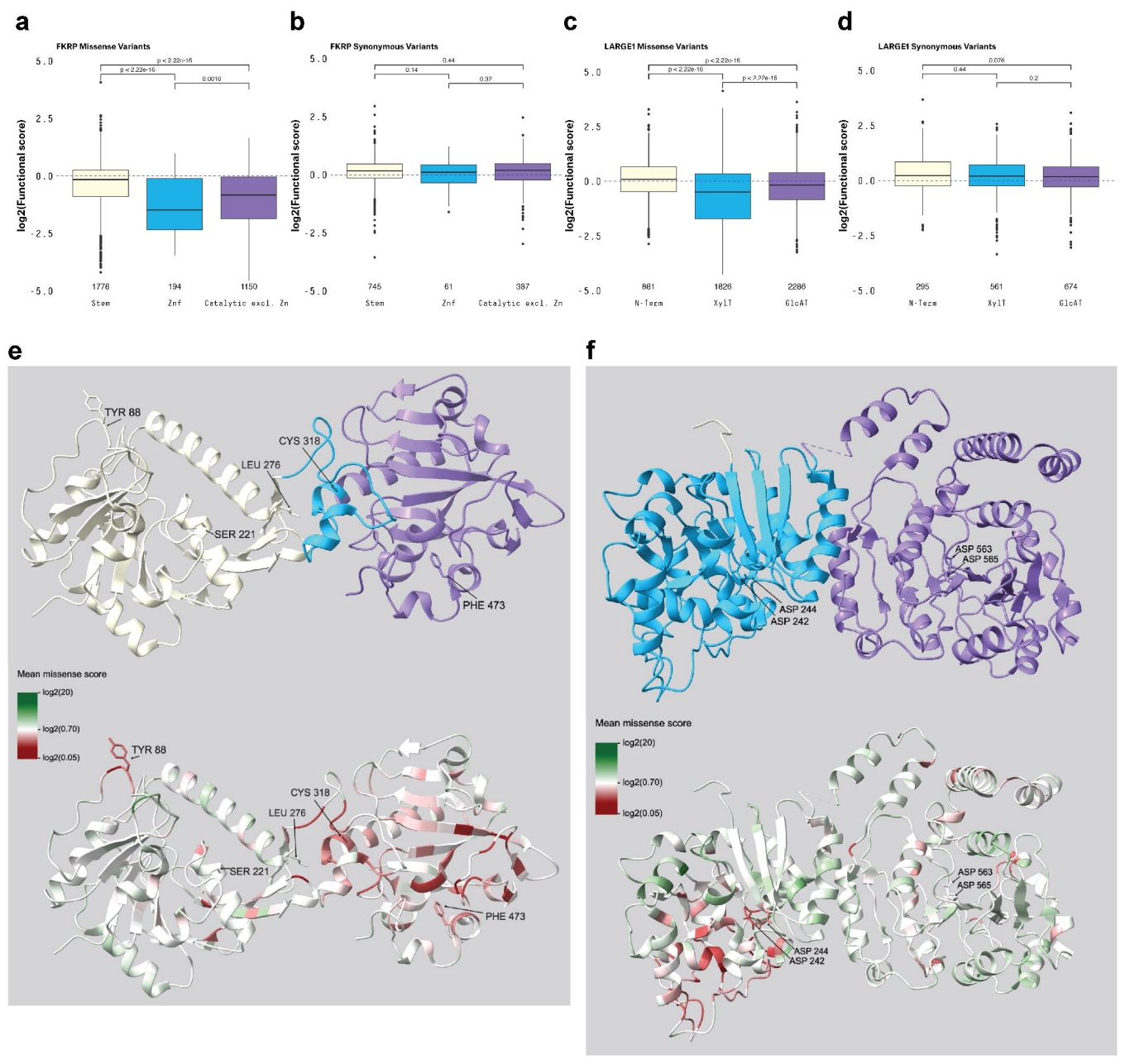
on α-DG-dependent viral entry. Vesicular stomatitis virus (VSV) with Lassa fever virus glycoprotein complex (LASV-GPC) can infect cells in an α-DG-dependent manner. Variant enrichment before/after VSV infection can be used to quantify their performances regarding α-DG glycosylation.

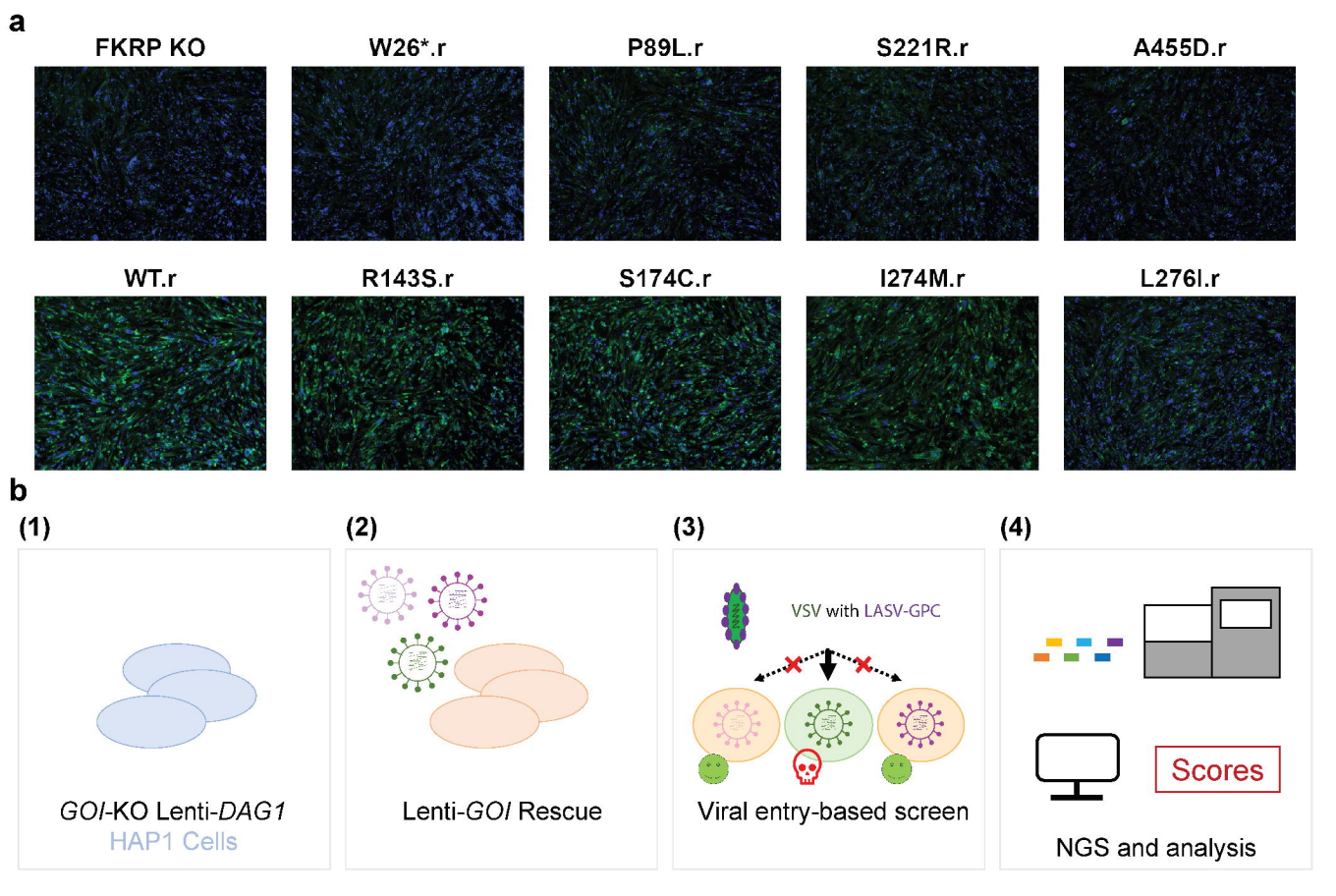

a
Matriglycan
GlcA
Xyl
n
LARGE1
B4GAT1
TMEM5
FKRP
FKTN
GlcA
Xyl
Rbo5P
Rbo5P
GalNAc
Core M3
B3GALNT2
GlcNAc
POMGNT2
ISPD
CTP + Rbo5P
CDP-Rbo
P
Mannose
POMK
POMT1/2
GMPPB
DPM1/2/3
DOLK
Dol-P-Man
α-Dystroglycan
b
(1)
(2)
(3)
(4)
GOI-KO Lenti-DAG1
HAP1 Cells
Lenti-GOI Rescue
Lenti-GOI variant library
Low function
High function
α-DG Glycosylation
Fluorescence-activated
cell sorting
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
374 variants
log2(Functional score)
2
0
-2
-4
1e-05
1e-04
0.001
0.01
0.1
gnomAD v3 Allele Frequency
Classification
Missense
Nonsense
Synonymous
d
LARGE1
426 variants
log2(Functional score)
2
0
-2
1e-05
1e-04
0.001
0.01
0.1
gnomAD v3 Allele Frequency
Classification
Missense
Nonsense
Start-loss
Synonymous


a
FKRP
740 variants
p < 2.22e-16
p < 2.22e-16
p < 2.22e-16
log2(Functional score)
5.0
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)
5.0
2.5
0.0
-2.5
-5.0
160
237
8
B/LB
VUS
P/LP
c
FKRP
Sensitivity
Specificity
metaSVM (AUC=0.64)
CADD (AUC=0.66)
MVP (AUC=0.66)
REVEL (AUC=0.79)
EVE (AUC=0.78)
MutScore (AUC=0.75)
SmuRF (AUC=0.78)
d
LARGE1
Sensitivity
Specificity
metaSVM (AUC=0.84)
CADD (AUC=0.77)
MVP (AUC=0.66)
REVEL (AUC=0.82)
EVE (AUC=0.75)
MutScore (AUC=0.75)
SmuRF (AUC=0.87)


a
FKRP Missense Variants
p < 2.22e-16
p < 2.22e-16
0.0016
log2(Functional score)
5.0
2.5
0.0
-2.5
-5.0
1778
194
1150
Stem
Znf
Catalytic excl. Zn
b
FKRP Synonymous Variants
0.44
0.14
0.37
745
61
387
Stem
Znf
Catalytic excl. Zn
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
N-Term
XylT
GlcAT
e
TYR 88
CYS 318
LEU 276
SER 221
PHE 473
Mean missense score
log2(20)
log2(0.70)
log2(0.05)
TYR 88
CYS 318
LEU 276
SER 221
PHE 473
f
ASP 563
ASP 565
ASP 244
ASP 242
Mean missense score
log2(20)
log2(0.70)
log2(0.05)
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)
GOI-KO Lenti-DAG1
HAP1 Cells
(2)
Lenti-GOI Rescue
(3)
VSV with LASV-GPC
Viral entry-based screen
(4)
Scores
NGS and analysis