Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine learning-based analysis of genotype–phenotype correlations in OFD1-associated phenotypes

Brunella Franco

franco@tigem.it

Federico II University of Naples https://orcid.org/0000-0001-5588-4569

Francesco Cifarelli

Telethon Institute of Genetics and medicine https://orcid.org/0009-0005-1449-8692

Nunziana Pezzella

Telethon Institute of Genetics and medicine

Ange-Line Bruel

FHU-TRANSLAD, CHU DIJON https://orcid.org/0000-0002-0526-465X

Christel Thauvin-Robinet

Hôpital d'Enfants https://orcid.org/0000-0002-4155-139X

Domenico Iervolino

Pasquale Dolce

Article

Keywords: OFD1, Ciliopathies, Machine Learning, SuperLearner, Genotype-phenotype correlations

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11400982/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no duality of interest

Machine learning-based analysis of genotype–phenotype correlations in *OFD1*-associated phenotypes

Francesco Cifarelli[1], Nunziana Pezzella[1], Ange-Line Bruel[3], Christel Thauvin-Robinet[3], Domenico Iervolino[4], the OFD1 collaborative group, Pasquale Dolce[5,*], Brunella Franco[1,2,6,*]

1. Telethon Institute of Genetics and Medicine (TIGEM), Via Campi Flegrei, 34, 80078, Pozzuoli, Naples, Italy

2. Genomic and Experimental Medicine (GEM) program, Scuola Superiore Meridionale, Naples, Italy

3. Université Bourgogne Europe, CHU Dijon Bourgogne, Inserm CTM UMR1231 GAD team, FHU TRANSLAD, Laboratoire de Génomique Médicale, Dijon, France

4. PhD Program in Translation Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinic Epidemiology", University of Padua

5. Department of Translational Medical Sciences, University of Naples Federico II, Via S Pansini 5, 80131, Naples, Italy

6. Medical Genetics, Department of Translational Medical Sciences, University of Naples Federico II, Via S Pansini 5, 80131, Naples, Italy

*co-corresponding authors

Abstract

Establishing genotype–phenotype correlations in pleiotropic genes remains a major clinical challenge. We collected clinical and genomic data from 356 individuals (61 newly described) carrying *OFD1* variants and classified phenotypes into cystic, neurocognitive, orocraniofacial, and skeletal groups. We combined positional and co-occurrence analyses with a SuperLearner ensemble model to predict phenotypic outcomes of missense variants, interpreted using SHAP. The framework was benchmarked on *LMNA,* another highly pleiotropic gene. Positional analysis revealed that C-terminal domain depletion is specifically associated with motile cilia impairment. Furthermore, co-occurrence analysis showed that while orocraniofacial, skeletal, and neurocognitive phenotypes co-segregate, cystic disease did not co-segregate with this triad, suggesting a partly distinct pathogenic trajectory. The SuperLearner model outperformed single algorithms with higher accuracy, identifying evolutionary conservation and structural domains as primary pathogenic drivers. Our machine learning framework provides a valuable, interpretable tool for predicting organ-specific risks and interpreting newly discovered variants, demonstrating high potential for extension to other pleiotropic genes characterized by complex clinical heterogeneity.

Keywords: OFD1; Ciliopathies; Machine Learning; SuperLearner; Genotype-phenotype correlations.

Introduction

The *OFD1* gene codes for a pleiotropic protein associated with different X-linked conditions, namely, Oro-facio-digital (OFD) type 1, primary ciliary dyskinesia (PCD), Joubert syndrome type 10 (JBS10), retinitis pigmentosa 23 (RP23), all associated with cilia dysfunction [1]. OFD type 1 is a multisystemic disorder due to dysfunction of primary cilia, classically thought to be inherited as a male lethal dominant trait and characterized by orocraniofacial abnormalities, skeletal defects, brain developmental abnormalities, and/or isolated cognitive defects and visceral involvement (e.g., cystic disease affecting kidneys, pancreas, and ovaries) [2]. JBS10 is a neurodevelopmental disorder also typically associated with primary cilia impairment, inherited as an X-linked recessive condition and defined by the presence of a specific cerebellar and brainstem malformation, the molar tooth sign, and by progressive involvement of retina, kidneys, and liver [3]. Conversely, PCD, also inherited as an X-linked recessive trait, is due to motile cilia dysfunction and is accordingly associated with respiratory problems and infertility [4, 5]. The last condition, RP23, is a non-syndromic form of retinitis pigmentosa described, to date, only in four individuals and transmitted as a recessive trait [6].

While next-generation sequencing has significantly improved variant prediction, the clinical trajectory associated with pleiotropic genes such as *OFD1* remains one of the most challenging tasks in clinical genomics. Recently, machine learning strategies such as AlphaMissense and PrimateAI-3D have been developed to predict whether a variant is pathogenic or benign. However, despite their exceptional potential, they are unable to define the phenotypic trajectories of a pathogenic missense variant [7, 8] . This limitation often complicates genetic counseling and the interpretation of variants of uncertain significance (VUS), particularly in patients with atypical clinical manifestations. To bridge this gap, we implemented a machine learning framework that applies a SuperLearner ensemble model to

predict the phenotypic manifestations of missense variants in *OFD1*. We also benchmarked our framework on another pleiotropic, but non-ciliary, gene, *LMNA*, and demonstrated its potential as a standard tool to predict organ-specific risks and interpret VUS across complex genetic syndromes, beyond the simple duality of pathogenic and benign.

Materials and methods

Study design and patient population

The study was approved by the appropriate institutional review board (IRB), and informed consent was obtained from participants enrolled as new cases. A comprehensive retrospective study was conducted by integrating clinical and genomic data from the OFD1 international collaborative group with data obtained through a systematic review of the literature. Inclusion criteria were the presence of molecularly confirmed non-synonymous variants in individuals with a phenotype compatible with one of the above-described *OFD1*-associated conditions. Novel variants were identified using sequencing-based approaches. PubMed was queried up to January 31, 2026, using the search terms: "OFD1", "Oral-Facial-Digital Syndrome type 1", “Orocraniofacialdigital Syndrome type 1”, "Joubert Syndrome type 10", "JBS10", "Primary Ciliary Dyskinesia", "PCD", "Retinitis Pigmentosa 23", and "RP23". Duplicate cases were identified by cross-referencing demographic data and variant details and were excluded to prevent statistical bias. A dataset of 39 variants classified as benign was retrieved from the ClinVar database (Table S1) to define a baseline of structural tolerance. Synonymous variants were systematically filtered out and only benign non-synonymous substitutions were retained to ensure the control group matched the molecular nature of the pathogenic missense cohort. These 39 variants constituted the strict phenotype-absent class (0) in the training matrix.

To ensure standardized and accurate reporting, all variants were systematically verified using the VariantValidator tool [9] and annotated according to HGVS nomenclature guidelines based on the *OFD1* reference transcript NM_003611.3 (GRCh38/hg38). Splice-site variants occurring outside canonical sites were evaluated using SpliceAI to predict their impact on mRNA processing and reading frame conservation. Given the retrospective nature of literature-derived data, incomplete phenotypic reporting may have influenced the classification of certain traits as absent.

Phenotypic classification

To ensure consistency across heterogeneous clinical sources, phenotypic features were mapped to standard Human Phenotype Ontology (HPO) terms. Clinical signs were dichotomized (Present/Absent) based on the following definitions: the cystic phenotype was defined as the presence of multiple cysts detected by ultrasound, computed tomography, or magnetic resonance imaging; motile cilia impairment was established by the presence of a clinical history of recurrent lower respiratory tract infections (pneumonia, bronchitis), bronchiectasis, and/or *situs inversus*, confirmed, when available, by low nasal nitric oxide levels or ultrastructural defects detected via transmission electron microscopy; neurocognitive involvement was defined by the presence of intellectual disability (IQ<70), global developmental delay according to the DSM-5-TR, and specific CNS malformations (e.g., agenesis of the corpus callosum, molar tooth sign); an orocraniofacial phenotype was designated when at least one of the following clinical signs was observed: abnormal hair/alopecia, facial milia, dysmorphism, cleft lip, tongue anomalies, teeth abnormalities, cleft or high arched palate, aberrant oral frenula; and skeletal anomalies included polydactyly, syndactyly, brachydactyly, clinodactyly, and short stature.

Collection of nucleotide variant features

The biological and structural features of missense variants were integrated to characterize their impact. The following predictive properties were extracted: conservation score, Relative Solvent Accessibility (RSA), secondary structure (q3), Grantham score, polarity shift, and functional domains. The conservation score was calculated using the ConSurf server [10] (consurf.tau.ac.il) based on multiple sequence alignments. RSA was calculated using NetSurfP-3.0 [11] (services.healthtech.dtu.dk/services/NetSurfP-3.0/). Residues were assigned to α-helices, β-sheets, or random coils based on NetSurfP-3.0 predictions. The Grantham score and polarity shifts were calculated to quantify the biophysical disparity between wild-type and mutant amino acids. Mapping data for OFD1 functional domains were retrieved from UniProt (ID: O75665) and/or from the literature.

Statistical analysis and machine learning framework

Descriptive statistical analyses were used to summarize cohort characteristics. Categorical variables were compared using the chi-square test or Fisher's exact test, as appropriate. Quantitative variables were compared using Student's t-test or Mann–Whitney U test, depending on their distribution. Statistical significance was set at $\alpha=0.05$ for all tests. Analyses were performed in R (version 4.3.3) [12] and GraphPad Prism (version 8.0). To analyze phenotypic clustering, a pairwise co-occurrence matrix was constructed, and the log odds ratio (LogOR) was calculated for each phenotype pair to quantify the strength of these associations. LogOR was used to quantify correlations between phenotypes. This metric provides a normalized scale in which positive values indicate co-occurrence and negative values indicate mutual exclusivity. For predictive modeling, a SuperLearner ensemble framework was implemented using the SuperLearner R package (version 2.0-40) [13]. To guarantee computational reproducibility, a fixed pseudo-random number seed (set.seed(123))

was initialized prior to all data partitioning, sampling, and training workflows. The ensemble included four base algorithms: Random Forest (randomForest, configured with ntree = 200) [14], Extreme Gradient Boosting (xgboost) [15], Support Vector Machines with a radial basis function kernel (caret) [16], and Generalized Linear Models (GLM) with Elastic Net regularization (glmnet) [17]. For the GLM component, an adaptive Elastic Net framework was used and both the mixing parameter (α) and the regularization penalty (λ) were optimized via grid search. This allowed the model to dynamically select the optimal penalization strategy, ranging from pure LASSO ($\alpha=1$) for sparse feature selection to Elastic Net ($0 < \alpha < 1$) for correlated predictors, with optimization performed independently for each phenotype. The SuperLearner meta-algorithm combined out-of-fold predictions from base classifiers using non-negative least squares (method.NNLS) to optimize convex combination weights and minimize cross-validated risk. All models were trained using three-fold cross-validation with five repeats to minimize overfitting. Class imbalance was addressed by implementing an upsampling strategy within the training loop, while stratified sampling was used to preserve the original class distribution in the test sets, thereby ensuring an unbiased evaluation of model performance. The risk of data leakage is minimal because repeated missense variants are rare in our dataset ($n \leq 5$ pathogenic). Moreover, individuals sharing the same variant often displayed different clinical signs. This phenotypic heterogeneity prevents the algorithm from artificially inflating its accuracy through mere variant memorization, ensuring true generalizability. Due to severe class imbalance between females and males in our cohort, patient sex was excluded as a predictive feature, focusing the model purely on the variant's structural and biophysical properties. Before model training, correlations between predictive variables were evaluated to assess potential redundancy. None of the variables showed complete collinearity, supporting their inclusion as complementary predictors. Optimal probability thresholds for the classification models were determined by maximizing the

Youden index to optimize the balance between sensitivity and specificity. Performance was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, balanced accuracy, F1 score, and log loss. Model calibration was assessed using the Brier score. Model interpretability was achieved through calculation of SHapley Additive exPlanations (SHAP) values [18]. The global significance of the different predictors was then visualized using bee swarm plots [19], which rank the top 10 variables based on their collective impact on model performance. Primary model evaluation was conducted using standard cross-validation. To confirm that performance estimates were not inflated by hyperparameter optimization or data leakage, we additionally performed a complete nested cross-validation, separating hyperparameter tuning (inner loop) from unbiased evaluation (outer loop).

Results

Definition of OFD1 domains

Standard databases often lack updated, integrated maps of gene structure that incorporate newly described functional or interaction domains. We updated the OFD1 map through an extensive review of the literature and public databases to provide the model with a structural framework incorporating spatial coordinates for evaluating the functional effects of pathogenic variants. We collected the localization of the LisH domain, regions of homo-oligomerization, and basic and acidic residues from UniProt (ID:O75665). We confirmed the specific positioning of the six coiled-coil domains, the seven low-complexity regions, and the four disordered regions, using bioinformatic tools [20]. In addition, we added the PKA-p motif [21], the LIR motifs [22] and the domains of interaction with eIF4E, RuvBl1, PAXILLIN, PCM1, SDCCAG8, and the ARP2/3 complex, all experimentally confirmed through pull-

187 down assays and described in the literature [23, 24] (Table S2). The results are schematically
188 illustrated in Figure 1A.
189
190 ## Patient cohort and variant distribution
191 We collected a cohort of 356 individuals displaying 228 different variants in *OFD1*,
192 including 61 newly described cases identified through cooperation within the OFD1
193 international collaborative effort (Table 1 and Table S3). Variants in the *OFD1* transcript are
194 associated with four distinct clinical entities (Figure 1A) distributed as follows: 296 cases of
195 OFD type 1 (including nine OFD1-XLR individuals), 35 JBS10, 21 PCD, and 4 RP23 (Figure
196 1B). The most common variant type was frameshift (51.9%), followed by splicing (13.7%),
197 missense (12.9%), nonsense (12.6%), and in-frame deletion/insertion variants (3.3%) (Figure
198 1C). Nineteen deletions (5.3%) encompassing different portions of the *OFD1* transcript were
199 also included in the dataset. Mapping these variants revealed distinct phenotype-specific
200 distribution patterns across the gene. Variants associated with classic OFD type 1 syndrome
201 in females are distributed across the first 18 exons. Interestingly, JBS10-associated variants
202 exhibit a bimodal distribution: non-truncating variants localize to exons 3-10, whereas
203 truncating variants localize to exons 17-21. PCD variants are exclusively truncating and
204 located at the C-terminus (exons 16-21). The four variants associated with RP23 involve
205 exons 4, 16, and 22, with one additional deep intronic variant located in intron 9. OFD type 1
206 has been traditionally considered an X-linked dominant male lethal disorder; however, our
207 review of the scientific literature highlighted a viable sub-cohort of nine males with an OFD
208 type 1 phenotype. Consequently, we defined a sub-category (OFD1-XLR) to highlight these
209 rare viable male patients, whose survival is likely linked to specific hypomorphic allelic
210 effects or somatic mosaicism (Table S3).
211

Truncating variants induce cystic phenotypes and motile cilia impairment in a position-dependent manner

We stratified our cohort into (i) truncating and (ii) non-truncating variants and observed that among the phenotypes evaluated, the cystic phenotype (CYS) was significantly more frequent in the truncating variant group (p-value<0.05) while signs of motile cilia impairment (MCI) were identified exclusively in individuals carrying truncating variants (Figure 2A). Exon-by-exon phenotypic mapping unveiled a specific distribution of the two phenotypes associated with truncating variants: CYS-associated variants were located within the first 16 exons, whereas truncating variants associated with MCI were specifically located in exons 16, 18, 20, and 21, underscoring a clear functional compartmentalisation within the protein (Figure 2B).

The cystic phenotype emerges as an independent clinical entity

We applied a pairwise phenotypic co-occurrence matrix approach to identify clinical signs that significantly co-segregate, revealing clusters that reflect shared pathogenic mechanisms. Our analysis showed an interconnected syndromic core including orocraniofacial (OCF), skeletal (SK), and neurocognitive (NC) defects. The significant co-segregation was strongly supported by elevated LogOR values (OCF vs. SK: LogOR = 3.71; OCF vs. NC: LogOR = 2.20; SK vs. NC: LogOR = 1.58). Notably, the cystic phenotype showed no significant co-occurrence with any of the other three phenotypes (Figure 2C), suggesting that cystic involvement follows a distinct clinical trajectory, independent of the course and development of the clinical triad OCF-SK-NC.

Genotype-phenotype correlation of missense variants

Unlike truncating variants, the pathogenicity of missense variants cannot be predicted solely from their position in the coding sequence, as they can have a variable effect on protein structure and function. To further investigate the effects of missense variants, we analyzed five distinct parameters: evolutionary conservation, Relative Solvent Accessibility (RSA), Grantham score, polarity shift, and secondary structure. Evolutionary conservation emerged as a global determinant, showing significant correlations across the entire clinical spectrum, including the CYS phenotype (Figure 3A). Conversely, RSA correlated exclusively with the classical syndromic triad (OCF, SK, NC; $P < 0.05$), showing no significant association with the cystic phenotype (CYS; $P > 0.05$) (Figure 3B). Other physico-chemical variables were insufficient to stratify the cohort, the Grantham score was associated only with the NC phenotype (Figure 3C), whereas polarity shift showed no significant association (Figure 3D). Finally, secondary structure (preferential localization within α-helices) acted similarly to evolutionary conservation, correlating across the entire clinical spectrum (Figure 3E), prompting the application of a multivariate machine-learning approach.

A SuperLearner ensemble model demonstrates the best performance in predicting *OFD1*-associated phenotypes

To overcome the limitations of traditional univariate statistical analysis, we first trained four distinct algorithms (Random Forest, XGBoost, SVM, and GLMNet) with the five variables described above together with the functional domains to predict the effect of *OFD1* missense variants. Because no single algorithm consistently outperformed the others across all four phenotypes (Table S4), we applied a SuperLearner ensemble framework to integrate the strengths of multiple models into a single robust predictor. The SuperLearner ensemble

model achieved higher performance than individual algorithms, reaching an Area Under the ROC Curve (AUC) ranging from 0.82 to 0.969 and balanced accuracy between 0.76 and 0.91 (Figure 4A, Table S4). Notably, the SuperLearner exhibited high specificity across the evaluated traits, predicting the absence of clinical phenotypes as effectively as their presence. Model calibration was assessed using the Brier score, indicating good agreement between predicted probabilities and observed outcome frequencies (Table S4). To confirm the robustness of our standard cross-validation results, all models were re-evaluated using a complete Nested Cross-Validation. This stringent evaluation confirmed the stability of the ensemble model, maintaining high predictive metrics across all phenotypic categories (Table S5).

Identifying the structural and domain-specific drivers of *OFD1* pathogenicity

To interpret predictions generated by the SuperLearner model, we used the SHapley Additive exPlanations (SHAP) framework, which represents the contribution of each variable to the model's prediction both at the single-variant level and globally across the entire cohort. The conservation score emerged as the most influential predictor across all clinical phenotypes. Secondary structure (q3) represented a top predictor for CYS and SK manifestations, whereas RSA was a top predictor for NC and OCF phenotypes. The Grantham score ranked among the top ten predictors but outside the top five most influential features, suggesting that physicochemical differences between wild-type and mutant residues contribute only partially to phenotype determination. Variants affecting the LisH domain were consistently associated with a higher predicted probability of phenotypic manifestation, identifying this region as a critical element. Conversely, variants located within interaction domains with Paxillin, RuvBL1, or SDCCAG8 were associated with a lower predicted probability of clinical

phenotypes. Overall, the integration of the SuperLearner framework with SHAP analysis demonstrates that each clinical phenotype is characterized by a distinct "risk signature," in which evolutionary conservation, structural context, and domain localization interact in complex, non-linear ways to determine the specific pathological outcome (Figure 4B).

Generalizability of the framework across pleiotropic genes: application to *LMNA*

To verify whether our SuperLearner framework is transferable beyond ciliopathies, we benchmarked it on missense variants in the pleiotropic *LMNA* gene, which encodes Lamin A/C. Starting from the cohort curated by Lin et al. [25], we classified *LMNA*-associated phenotypes in four clinical groups (Cardiomyopathy/Arrhythmia, Metabolic/Lipodystrophy, Muscular/Neuromuscular, and Progeroid/Aging). The SuperLearner achieved an overall mean AUC of 0.923 (Figure 4C) and a mean balanced accuracy of 0.86 (Table S6). Notably, predictive performance was exceptionally high for Metabolic/Lipodystrophy (AUC = 0.976), while remaining robust across severe tissue-specific and progeroid presentations (Figure 4C, Table S6). Consistent with our primary *OFD1* findings, the SuperLearner model provided superior probabilistic calibration compared to individual candidate algorithms, yielding the lowest mean Brier score (0.081) and mean log loss (0.263) across all phenotypes (Table S6) and confirming its scalability as a standard tool to dissect complex pleiotropic genetic architectures.

Discussion

The interpretation of genotype–phenotype relationships in pleiotropic genes remains one of the most challenging aspects of clinical genomics. Despite the widespread adoption of next-generation sequencing, predicting the clinical trajectory associated with specific variants

remains challenging. Current tools are largely limited to the dichotomic classification of variants as benign or pathogenic [7, 8]; this gap represents a major limitation for genetic counseling and patient management. It also complicates the implementation of tailored surveillance strategies, particularly for missense substitutions, whose pathogenic mechanisms are often poorly captured by standard interpretative frameworks.

Our work proposes a model that could provide a starting point for the development of future comprehensive and clinically actionable diagnostic tools to guide organ-specific diagnostic and therapeutic strategies.

We tested our framework on a cohort of 356 individuals carrying 228 different variants in the pleiotropic gene *OFD1*, including 61 newly identified cases. The increased cohort size allowed us to move beyond descriptive reporting toward structured genotype–phenotype modelling integrating genomic position, affected functional domain and protein-structural features. Our mapping analysis showed that, particularly for truncating variants, the simple loss of a portion of the protein is not sufficient to explain the phenotype; rather, the precise position of the variant influences the clinical outcome. Indeed, while truncating variants in the first 18 exons [1] may result in a complete loss of function, those located at the C-terminus may selectively affect regions required for motile cilia function. This spatial segregation could explain the presence of respiratory and reproductive defects in PCD [4, 5, 26] and in a few patients with JBS10 [3] and is further supported by the cases described by Budny et al.[4], where individuals are affected both by OFD1 syndrome and motile cilia impairment. The case reported by Budny et al. marks a key biological transition. The frameshift variant found in exon 16 (c.2122-2125dupAAGA) leads to the production of a stably truncated protein that lacks its C-terminus but retains the full N-terminal and central regions, enabling the recruitment of the minimal machinery needed for primary ciliogenesis. This remaining function prevents hemizygous male embryos from undergoing prenatal lethality that is

normally linked to early null alleles. Exon 16 therefore appears to represent a critical transition point, revealing the bifunctional role of the OFD1 gene both during embryonic development and in adult differentiated tissues. Although we cannot exclude that rare amino acid substitutions at the C-terminus might also disrupt motile cilia function, no such variants were identified in our dataset, precluding their inclusion in the missense-based modeling.

Beyond the spatial segregation of truncating variants, our phenotypic co-occurrence analysis suggested a distinct clinical trajectory for cystic involvement. The cystic phenotype (CYS) did not show significant co-occurrence with the orocraniofacial, skeletal, and neurocognitive triad (OCF-SK-NC), which were instead closely inter-correlated. This supported a partially independent pathogenic trajectory for CYS. Indeed, the OCF-SK-NC triad has been linked to early alterations in embryonic developmental morphogens, such as the Wnt and Sonic Hedgehog pathways [27, 28, 29, 30, 31]. In contrast, previous studies [32, 33, 34] suggest that cystogenesis could involve later, interconnected networks that depend on the loss of ciliary mechanosensation, which in turn disrupts planar cell polarity (PCP) and alters intracellular calcium influx. This calcium deficiency may cause a disturbance in cAMP homeostasis, resulting in adenylate cyclase activation and subsequent fluid secretion and epithelial proliferation via the PKA/ERK pathway [32, 33]. At the same time, ciliary dysfunction could lead to hyperactivation of the mTOR pathway, causing abnormal cell growth [34]. Taken together, these complex molecular interactions could provide a reasonable mechanistic explanation for the rare instances of isolated polycystic kidney disease linked to *OFD1* [35]. To date, the number of patients with CYS may be underestimated due to a lack of long-term follow-up. These findings support targeted organ surveillance (kidney ultrasound and nephrological follow-up), regardless of how mild or absent the remaining syndromic presentation may be.

A central aspect of our study is the analysis of missense variants, which, unlike truncating variants, require residue-level analysis of structural and functional perturbations. We incorporated benign variants as controls to establish a baseline of structural tolerance and implemented a SuperLearner ensemble framework integrating multiple machine-learning algorithms. The SuperLearner framework decoded a non-linear system better than the individual classifiers and demonstrated satisfactory calibration across phenotypes. To interpret the SuperLearner predictions, SHapley Additive exPlanations (SHAP) analysis was applied to identify the features most relevant to the model's prediction [36]. Across all clinical phenotypes, the main predictive variables were conservation, RSA and secondary structure (q3). Importantly, these features collectively define the global topological and thermodynamic impact of a variant on the protein architecture and folding, which proved to be a stronger disease determinant than mere local chemical alterations, such as polarity shifts. Our structural analysis highlighted distinct roles for different biophysical parameters in shaping the clinical outcome. In univariate analysis, altered RSA correlated exclusively with the OCF-SK-NC triad, while secondary structure perturbations (q3) affected the entire clinical spectrum, including the CYS phenotype. When evaluated through the multivariate SHAP analysis to determine the most relevant features for the model, q3 emerged as a top predictor for CYS and SK, whereas RSA was a top predictor for NC and OCF. We hypothesize that CYS and SK manifestations might be highly sensitive to perturbations in local secondary structure elements, which could destabilize specific functional domains required for downstream pathways like mTOR signaling. In contrast, NC and OCF phenotypes appear more strictly dependent on surface-exposed residues. Variants altering these accessible interfaces might directly disrupt critical early-embryonic protein-protein interactions without triggering a global collapse of the molecular scaffold.

The LisH domain is the domain most strongly associated with an increased predicted probability of severe phenotypes, consistent with its established role in ciliary proteins dimerization. Conversely, variants located within the binding domains for Paxillin, RuvBL1, or SDCCAG8 were associated with a lower predicted probability of severe manifestations, suggesting that, unlike the disruption of the LisH domain, they do not cause the complete collapse of the entire molecular scaffold at the basal body [37, 38, 39].

To enhable our model to better handle rare phenotypes, such as the CYS phenotype, we addressed class imbalance through upsampling and assessed model performance using Nested Cross-Validation. Although external validation was not possible because of the limited availability of independent *OFD1* cohorts with detailed phenotypic annotation, prospective validation in larger international cohorts will be needed to confirm the generalizability of our findings.

To verify that our model could be applied beyond ciliary biology, we benchmarked the framework on missense variants of *LMNA*, a pleiotropic gene responsible for a spectrum of cardiac, metabolic, neuromuscular, and premature aging disorders [40]. The SuperLearner model achieved a good predictive accuracy across all four *LMNA* clinical groups. Retraining the same pipeline on *LMNA* yielded comparable discrimination, indicating that the approach is transferable to other pleiotropic genes. This does not constitute external validation of the OFD1 models, which remain to be tested in independent cohorts. Several caveats require consideration. The retrospective design and reliance on literature-derived cases introduce potential ascertainment and reporting bias. Certain phenotypes were excluded from the predictive modeling analysis for specific biological and statistical reasons. Retinitis pigmentosa (RP23) was not included due to only four reported cases, which precluded meaningful statistical analysis. Similarly, motile cilia impairment (MCI) was omitted from the missense-based modeling because all variants associated with this phenotype in our

cohort were truncating, making them unsuitable for an analysis specifically focused on missense substitutions. This restricts the pathogenic mechanism of MCI to C-terminal loss rather than localized amino acid substitutions. Anothre important consideration is that in females, X-chromosome inactivation (XCI) drives phenotypic variability [41]. However, these epigenetic factors could not be modeled because XCI data were not available for all patients in our cohort. This is a broader field challenge; indeed, even advanced computational tools like AlphaMissense struggle to interpret X-linked variants, as purely structural algorithms remain blind to patient-specific epigenetic dynamics [42]. Despite these limitations, our approach offers significant translational value. In a clinical genomics context, "black-box" predictions are insufficient for variant interpretation. By employing SHAP analysis, our framework provides transparent, phenotype-specific predictions where the contribution of each structural variable is clearly quantified. Within the ACMG/AMP variant interpretation guidelines, computational predictors contribute as a supporting line of evidence. Therefore, our model could provide structured, interpretable data that could actively assist clinicians in prioritizing organ-specific risks and reclassifying variants of uncertain significance (VUS), translating computational insights into actionable diagnostic support.

In conclusion, this study expands the mutational landscape of *OFD1* and introduces a robust structural and machine-learning framework to support the decoding of clinical heterogeneity in *OFD1*-related disorders. It may ultimately contribute to more precise genotype–phenotype prediction in ciliopathies and beyond.

Data availability

The novel variants identified in this study have been submitted to the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) under accession numbers indicated in Table 1 (held under embargo until publication). All other datasets generated and analyzed during the current study are available from the corresponding author upon request.

Code availability

The computational analysis was performed using standard open-source packages in R version 4.3.3, as detailed in the Materials and Methods section. The custom R scripts and pipelines generated during the current study to implement the SuperLearner ensemble model and SHAP interpretation are available from the corresponding author upon reasonable request.

Acknowledgements

We thank Drs. Sandro Banfi and Roberta Russo for helpful discussions and critical reading of the manuscript and the Italian Telethon Foundation and Scuola Superiore Meridionale for continuous support.

Author contributions

F.C., N.P., and B.F. conceptualized, critically reviewed, and edited the manuscript. F.C. performed all computational analyses and drafted the manuscript. P.D. and D.I. supported the bioinformatics design and data interpretation. A.-L.B. and C.T.R. and the OFD1 C.G. provided clinical data. B.F. supervised the research. All authors approved the final version.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics declarations

This study included both newly recruited patients and previously published cases. The study was approved by the ethical commette of the Centre Hospitalier Universitaire Dijon and is authorized under the protocol NCT01962129 (https://clinicaltrials.gov/study/NCT01962129?viewType=Card&locStr=Dijon,%20France&country=FR&state=Bourgogne-Franche-Comt%C3%A9&city=Dijon&cond=Orofaciodigital%20Syndromes&rank=1**)**. For newly collected patient data, informed consent was obtained from participants or their legal guardians. Previously published cases were retrieved from the scientific literature and publicly available resources.

Competing interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Gemini to assist with English language editing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

The OFD1 collaborative group

Note for the publisher: The following individuals are members of the OFD1 collaborative group and served as non-author contributors (collaborators). Please index them in PubMed/MEDLINE as collaborators.

Albert Ong, Sheffield Children's Hospital, Sheffield, UK

O. W. Quarrel, Sheffield Children's Hospital, Sheffield, UK

André Travessa, Medical Genetics Department, Hospital de Santa Maria, Lisbon, Portugal

Juliette Dupont, Medical Genetics Department, Hospital de Santa Maria, Lisbon, Portugal

Patrícia Dias, Medical Genetics Department, Hospital de Santa Maria, Lisbon, Portugal

C. F. Sirri, Celal Bayar University, Manisa, Turkey

Cindy Skinner, Greenwood Genetic Center, Greenwood, South Carolina, USA

Sara Cathey, Greenwood Genetic Center, Greenwood, South Carolina, USA

Enza Maria Valente, University of Pavia, Pavia, Italy

Valentina Serpieri, University of Pavia, Pavia, Italy

Florence Petit, University of Lille, Lille, CHU Lille, ULR7364 RADEME, FHU G4 Génomique, F-59000 Lille, France

Fortunato Lonardo, P.O. Gaetano Rummo, Azienda Ospedaliera San Pio, Benevento, Italy

Paolo Fontana, P.O. Gaetano Rummo, Azienda Ospedaliera San Pio, Benevento, Italy

Francesca Mari, University of Siena, Siena, Italy

Alessandra Renieri, University of Siena, Siena, Italy

Francesca Santamaria, Department of Translational Medical Sciences, Federico II University, Naples, Italy

Francesco Brancati, University of L'Aquila, L'Aquila, Italy

Fuad Chowdhury, Alberta Children's Hospital, University of Calgary, Calgary, Canada

Ashish Marwaha, Alberta Children's Hospital, University of Calgary, Calgary, Canada

Gabriella Voso, Saint Peter’s University Hospital, New Brunswick, New Jersey, USA

Giulia Carignani, Ospedale Careggi, Florence, Italy

Gretchen Oswald, Johns Hopkins University, Baltimore, Maryland, USA

Harry Dietz, Johns Hopkins University, Baltimore, Maryland, USA

Katia Bessaci, American Memorial Hospital CHU de Reims, Centre de Ressources et de Compétence de la Mucoviscidose, F-51000 Reims, France

Marie Legendre, APHP, Sorbonne Université, Inserm, Hôpital Armand Trousseau, Paris, France

Maimoona Zariwala, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Timothy John Vece, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Stephanie Davis, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Marco Seri, Medical Genetics, S. Orsola-Malpighi Hospital, Bologna, Italy

Mercè Artigas López, Servicio de Genética Médica, Hospital Virgen del Camino, Pamplona, Spain

Pia Castelluccio, Ospedale Cardarelli, Naples, Italy

Regaan Klatt, The Hospital for Sick Children, Toronto, Canada

Ahmad Teebi, The Hospital for Sick Children, Toronto, Canada

Serdar Ceylaner, INTERGEN Genetics and Rare Diseases Diagnosis Research & Application Center, Ankara, Turkey

Nesrin Ceylan, INTERGEN Genetics and Rare Diseases Diagnosis Research & Application Center, Ankara, Turkey

Tadeusz Kałużewski, GENOS Sp. z o.o., Łódź, Poland

References

1. Pezzella N, Bove G, Tammaro R, Franco B. OFD1: One gene, several disorders. Am J Med Genet C Semin Med Genet 2022; 190: 57–71. https://doi.org/10.1002/ajmg.c.31962
2. Ferrante MI, Giorgio G, Feather SA, Bulfone A, Wright V, Ghiani M, et al. Identification of the gene for oral-facial-digital type 1 syndrome. Am J Hum Genet 2001; 68: 569–76. https://doi.org/10.1086/318802
3. Coene KLM, Roepman R, Doherty D, Afroze B, Kroes HY, Letteboer SJF, et al. OFD1 Is Mutated in X-Linked Joubert Syndrome and Interacts with LCA5-Encoded Lebercilin. Am J Hum Genet 2009; 85: 465–81. https://doi.org/10.1016/j.ajhg.2009.09.002
4. Budny B, Chen W, Omran H, Fliegauf M, Tzschach A, Wisniewska M, et al. A novel X-linked recessive mental retardation syndrome comprising macrocephaly and ciliary dysfunction is allelic to oral–facial–digital type I syndrome. Hum Genet 2006; 120: 171–8. https://doi.org/10.1007/s00439-006-0210-5
5. Wallmeier J, Nielsen KG, Kuehni CE, Lucas JS, Leigh MW, Zariwala MA, et al. Motile ciliopathies. Nat Rev Dis Primers 2020; 6: 77. https://doi.org/10.1038/s41572-020-0209-6
6. Webb TR, Parfitt DA, Gardner JC, Martinez A, Bevilacqua D, Davidson AE, et al. Deep intronic mutation in OFD1, identified by targeted genomic next-generation sequencing, causes a severe form of X-linked retinitis pigmentosa (RP23). Hum Mol Genet 2012; 21: 3647–54. https://doi.org/10.1093/hmg/dds194
7. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 2023; 381: eadg7492. https://doi.org/10.1126/science.adg7492

8. Gao H, Hamp T, Ede J, Schrider CB, McRae JF, Chandonia JM, et al. The landscape of tolerated genetic variation in humans and primates. Science 2023; 380: eabn8197. https://doi.org/10.1126/science.abn8197
9. Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping and formatting of sequence variation descriptions. Hum Mutat 2018; 39: 31-38. https://doi.org/10.1002/humu.23348
10. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 2016; 44: W344. https://doi.org/10.1093/nar/gkw408
11. Høie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, et al. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. Nucleic Acids Res 2022; 50: W510–5. https://doi.org/10.1093/nar/gkac439
12. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2026. https://www.R-project.org/
13. Polley E, LeDell E, Kennedy C, van der Laan M. SuperLearner: Super Learner Prediction. 2025. https://CRAN.R-project.org/package=SuperLearner
14. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002; 2: 18–22. https://CRAN.R-project.org/doc/Rnews/
15. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. xgboost: Extreme Gradient Boosting. 2025. https://CRAN.R-project.org/package=xgboost
16. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw 2008; 28: 1–26. https://doi.org/10.18637/jss.v028.i05

17. Tay JK, Narasimhan B, Hastie T. Elastic Net Regularization Paths for All Generalized Linear Models. J Stat Softw 2023; 106: 1–31. https://doi.org/10.18637/jss.v106.i01
18. Greenwell B. fastshap: Fast Approximate Shapley Values. 2024. https://CRAN.R-project.org/package=fastshap
19. Mayer M. shapviz: SHAP Visualizations. 2025. https://CRAN.R-project.org/package=shapviz
20. Jagodzik P, Zietkiewicz E, Bukowy-Bieryllo Z. Conservation of OFD1 Protein Motifs: Implications for Discovery of Novel Interactors and the OFD1 Function. Int J Mol Sci 2025; 26. https://doi.org/10.3390/ijms26031167
21. Senatore E, Chiuso F, Rinaldi L, Intartaglia D, Delle Donne R, Pedone E, et al. The TBC1D31/praja2 complex controls primary ciliogenesis through PKA-directed OFD1 ubiquitylation. EMBO J 2021; 40: e106503. https://doi.org/10.15252/embj.2020106503
22. Morleo M, Brillante S, Formisano U, Ferrante L, Carbone F, Iaconis D, et al. Regulation of autophagosome biogenesis by OFD1-mediated selective autophagy. EMBO J 2021; 40: e105120. https://doi.org/10.15252/embj.2020105120
23. Iaconis D, Monti M, Renda M, Van Koppen A, Tammaro R, Chiaravalli M, et al. The centrosomal OFD1 protein interacts with the translation machinery and regulates the synthesis of specific targets. Sci Rep 2017; 7: 1156. https://doi.org/10.1038/s41598-017-01156-x
24. Cao M, Zou X, Li C, Lin Z, Wang N, Zou Z, et al. An actin filament branching surveillance system regulates cell cycle progression, cytokinesis and primary ciliogenesis. Nat Commun 2023; 14: 3734. https://doi.org/10.1038/s41467-023-37340-z

602 25. Lin EW, Brady GF, Kwan R, Nesvizhskii AI, Omary MB. Genotype-phenotype
603 analysis of *LMNA*-related diseases predicts phenotype-selective alterations in lamin
604 phosphorylation. FASEB J 2020; 34: 9051. https://doi.org/10.1096/fj.202000500r
605 26. Hannah WB, DeBrosse S, Kinghorn BA, Strausbaugh S, Aitken ML, Rosenfeld M, et
606 al. The expanding phenotype of OFD1-related disorders: Hemizygous loss-of-
607 function variants in three patients with primary ciliary dyskinesia. Mol Genet
608 Genomic Med 2019; 7: e911. https://doi.org/10.1002/mgg3.911
609 27. Kurosaka H, Iulianella A, Williams T, Trainor PA. Disrupting hedgehog and WNT
610 signaling interactions promotes cleft lip pathogenesis. J Clin Invest 2014; 124: 1660–
611 71. https://doi.org/10.1172/jci72688
612 28. Macca M, Franco B. The molecular basis of oral-facial-digital syndrome, type 1. Am
613 J Med Genet C Semin Med Genet 2009; 151C: 318–25.
614 https://doi.org/10.1002/ajmg.c.30224
615 29. Bimonte S, De Angelis A, Quagliata L, Giusti F, Tammaro R, Dallai R, et al. Ofd1 is
616 required in limb bud patterning and endochondral bone development. Dev Biol 2011;
617 349: 179–91. https://doi.org/10.1016/j.ydbio.2010.09.020
618 30. D'Angelo A, de Angelis A, Avallone B, Piscopo I, Tammaro R, Studer M, et al. Ofd1
619 controls dorso-ventral patterning and axoneme elongation during embryonic brain
620 development. PLoS One 2012; 7: e52937.
621 https://doi.org/10.1371/journal.pone.0052937
622 31. Ferrante MI, Zullo A, Barra A, Bimonte S, Messaddeq N, Studer M, et al. Oral-facial-
623 digital type I protein is required for primary cilia formation and left-right axis
624 specification. Nat Genet 2006; 38: 112–7. https://doi.org/10.1038/ng1684
625 32. Zullo A., Iaconis D., Barra A., Cantone A., Messaddeq N., Capasso G., et al.. Kidney-
626 specific inactivation of Ofd1 leads to renal cystic disease associated with upregulation

627 of the mTOR pathway. Hum Mol Genet. 2010 Jul 15;19(14):2792-803.
628 https://doi.org/10.1093/hmg/ddq180
629 33. Buqaileh R., Alshriem L. A., AbouAlaiwi W. Ciliary G-Protein Coupled Receptor
630 Signaling in Polycystic Kidney Disease. Int J Mol Sci 2025; 26: 4971.
631 https://doi.org/10.3390/ijms26114971
632 34. Al-Orjani Q., Alshriem L. A., Gallagher G., Buqaileh R., Azizi N., AbouAlaiwi W.
633 Mechanistic Insights into the Pathogenesis of Polycystic Kidney Disease. Cells 2025;
634 14: 1203. https://doi.org/10.3390/cells14151203
635 35. Zhang HW, Su BG, Yao Y. OFD1 mutation induced renal failure and polycystic
636 kidney disease in a pair of childhood male twins in China. World J Clin Cases 2020;
637 8: 331–6. https://doi.org/10.12998/wjcc.v8.i2.331
638 36. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and
639 epigenomics. Brief Bioinform 2021; 22: bbaa177. https://doi.org/10.1093/bib/bbaa177
640 37. Van De Weghe JC, Rusterholz TDS, Latour B, Grout ME, Aldinger KA, Shaheen R,
641 et al.. Mutations in ARMC9, which Encodes a Basal Body Protein, Cause Joubert
642 Syndrome in Humans and Ciliopathy Phenotypes in Zebrafish. Am J Hum Genet.
643 2017 Jul 6;101(1):23-36. https://doi.org/10.1016/j.ajhg.2017.05.010
644 38. DeSpenza, T., Jr, Singh, A., Allington, G., Zhao, S., Lee, J., Kiziltug, E., et al.
645 Pathogenic variants in autism gene *KATNAL2*cause hydrocephalus and disrupt
646 neuronal connectivity by impairing ciliary microtubule dynamics. Proc Natl Acad Sci
647 U S A. 2024 Jul 2;121(27):e2314702121. https://doi.org/10.1073/pnas.2314702121
648 39. Huang Y, Jiang Z, Gao X, Luo P, Jiang X. ARMC Subfamily: Structures, Functions,
649 Evolutions, Interactions, and Diseases. Front Mol Biosci. 2021 Nov 29;8:791597.
650 https://doi.org/10.3389/fmolb.2021.791597

40. Worman HJ, Bonne G. “Laminopathies”: A wide spectrum of human diseases. Exp Cell Res 2007; 313: 2121–2133. https://doi.org/10.1016/j.yexcr.2007.03.028
41. Tukiainen T, Villani AC, Yen A, Rivas MA, Marshall JL, Satija R, et al. Landscape of X chromosome inactivation across human tissues. Nature 2017; 550: 244-248. https://doi.org/10.1038/nature24265
42. Krenn M, Schmidt A, Wagner M, Ernst M, Graf E, Zulehner G, et al. AlphaMissense prediction for the evaluation of missense variants in the diagnostic setting of neuromuscular disorders. Eur J Neurol 2025; 32: e16544. https://doi.org/10.1177/22143602251370957.

Figures legend

Figure 1. Distribution of *OFD1*-associated variants. A) *OFD1*-associated disorders: Oral-Facial-Digital syndrome type 1 (OFD1), Joubert syndrome type 10 (JBS10), Primary Ciliary Dyskinesia (PCD), and Retinitis Pigmentosa 23 (RP23). The relative clinical signatures are indicated. OFD1 and JBS10 exhibit convergent features (skeletal abnormalities, neurocognitive impairment, and cystic organ disease), PCD is characterized by defects in motile cilia-bearing tissues, and RP23 presents as isolated retinal dystrophy. The exon organization of the *OFD1* gene and the identified functional domains are also shown. Different colors indicate the domains identified. The upper track maps the distribution of disease-specific variants across the gene structure (visualized via ProteinPaint). **B)** Study cohort featuring 61 newly identified cases (OFD1, n=52; JBS10, n=3; PCD, n=5; RP23, n=1). **C)** Distribution of variant types (Frameshift, n=185; Splicing, n=49; Missense, n=46; Nonsense, n=45; In frame, n=12; Deletions, n=19).

Figure 2. Clinical and genetic composition of the *OFD1* cohort. A) Bar plot comparing the frequency of clinical manifestations between individuals harboring truncating and non-truncating variants. The cystic phenotype (CYS) is significantly more frequent in the truncating variant cohort. **B)** Distribution of truncating variants across exons. Variants associated with the cystic phenotype (CYS) are localized within exons 1–16 (blue), whereas variants linked to motile cilia impairment (MCI) are confined to exons 16–21 (orange). **C)** The co-occurrence matrix shows the phenotypic overlap between clinical features. Color intensity reflects the Log Odds Ratio (LogOR). A robust clinical cluster is observed for OCF, SK, and NC, while the CYS phenotype remains statistically independent.

Statistical significance was determined using Fisher's exact test (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$).

Figure 3. Analysis of missense variants. A-C) Violin plots show the results of the Mann-Whitney U test performed on individuals with (+) and without (-) each phenotypes (CYS, cystic; NC, neurocognitive; OCF, orocraniofacial; SK, skeletal) for the five variables selected; **A)**, Conservation scores, **B)** Relative Solvent Accessibility, and **C**) Grantham scores. **D-E)** Bar plots. Fisher's exact test shows the association of polarity shift (D) and secondary structure (E) with each phenotype.

(* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$).

Figure 4. Machine learning prediction of OFD1-associated phenotypes and cross-gene validation on *LMNA*. A) Receiver Operating Characteristic (ROC) curves for the prediction of the four clinical phenotypes. Each curve shows the performance of the five algorithms used: RF, Random Forest; XGB, Extreme Gradient Boosting; SVM, Support Vector Machine; GLMNET, Generalized Linear Models with Elastic Net regularization, and SuperLearner. The

SuperLearner, which represents the ensemble of the other four algorithms, has the best performance. **B)** SHapley Additive exPlanations (SHAP) beeswarm plots reveal the relative contribution of each variable to the model's predictions. Factors such as conservation scores, RSA, and q3 emerge as key predictors for phenotypic outcomes. **C)** ROC curves comparing the predictive performance of the SuperLearner ensemble against four base algorithms (Random Forest, XGBoost, SVM, and GLMNet) across four distinct *LMNA*-associated clinical groups: Cardiomyopathy/Arrhythmia, Metabolic/Lipodystrophy, Muscular/Neuromuscular, and Progeroid/Aging.

Abbreviations: AUC, Area Under the Curve; CYS, cystic phenotype; NC, neurocognitive phenotype; OCF, orocraniofacial phenotype; SK, skeletal phenotype; RSA, Relative Solvent Accessibility; q3, secondary structure. Domain abbreviations are provided in Figure 1A.

711 **Table 1: Novel *OFD1* variants and associated clinical phenotypes.**

Patient ID	ClinVar ID	Gene	Transcript	cDNA Variant (HGVS)	Protein Variant (HGVS)	Variant type	Location	Diagnosis	CYS	SK	NC	OCF
PED0254.1	SCV008086502	*OFD1*	NM_003611.3	c.86_90del	p.Asp29Glyfs17	frameshift	Exon 2	OFD1	+	+		+
PED4588.1	SCV008086503	*OFD1*	NM_003611.3	c.112-2del	p.?	splicing	Intron 2	OFD1				+
PZ. 132	SCV008086504	*OFD1*	NM_003611.3	c.112A>C	p.Thr38Pro	missense	Exon 3	OFD1		+		+
PED0249.1	SCV008086505	*OFD1*	NM_003611.3	c.219_220insCT	p.Ser74Leufs*5	frameshift	Exon 3	OFD1				
PED1417.4	SCV008086506	*OFD1*	NM_003611.3	c.221_222dupCT	p.Asn75Leufs*4	frameshift	Exon 3	OFD1		+		+
PED1551.1	SCV008086507	*OFD1*	NM_003611.3	c.221delinsTAAAAGAGCTGCC	p.Ser74delinsLeuLysGluLeuPro	in-frame	Exon 3	OFD1		+		+
PZ.183	SCV008086508	*OFD1*	NM_003611.3	c.257G>A	p.Gly86Asp	missense	Exon 3	OFD1	+	+	+	+
PED0268.1	SCV008086509	*OFD1*	NM_003611.3	c.283_286del	p.Phe95Glnfs*49	frameshift	Exon 3	OFD1	+	+		+
PED2031.1	SCV008086510	*OFD1*	NM_003611.3	c.312+1del	p.?	splicing	Intron 3	OFD1	+			+
PED0261.1	SCV008086511	*OFD1*	NM_003611.3	c.313-2A>G	p.?	splicing	Intron 3	OFD1		+	+	+
PED3378.1	SCV008086512	*OFD1*	NM_003611.3	c.344T>A	p.Ile115Asn	missense	Exon 4	OFD1				+
PED1172.1	SCV008086513	*OFD1*	NM_003611.3	c.381+1G>A	p.?	splicing	Intron 4	OFD1		+	+	+
PED5081.1	SCV008086514	*OFD1*	NM_003611.3	c.400_403del	p.Glu134Ilefs*10	frameshift	Exon 5	OFD1	+	+		+
PED0247.1	SCV008086515	*OFD1*	NM_003611.3	c.412+1G>C	p.?	splicing	Intron 5	OFD1		+	+	+
PED3171.1	SCV008086516	*OFD1*	NM_003611.3	c.412+1G>T	p.?	splicing	Intron 5	OFD1			+	+
PED7766.1	SCV008086517	*OFD1*	NM_003611.3	c.431dup	p.Leu144Phefs*11	frameshift	Exon 6	OFD1				
PZ. 300	SCV008086518	*OFD1*	NM_003611.3	c.587T>C	p.Leu196Ser	missense	Exon 7	OFD1	-	+	+	+
PED0278.1	SCV008086519	*OFD1*	NM_003611.3	c.613_617delAGAGAinsTAT	p.Arg205Tyrfs*18	frameshift	Exon 7	OFD1				+
PED5137.1	SCV008086519	*OFD1*	NM_003611.3	c.613_617delinsTAT	p.Arg205Tyrfs*18	frameshift	Exon 7	OFD1			+	+

PED0234.1	SCV008086520	*OFD1*	NM_003611.3	c.616_620del	p.Glu206Argfs*16	frameshift	Exon 7	OFD1				+
Case 3	SCV008086521	*OFD1*	NM_003611.3	c.628C>T	p.Gln210*	nonsense	Exon 7	OFD1	+	+	+	+
PED1559.1	SCV008086522	*OFD1*	NM_003611.3	c.655-8A>G	p.?	splicing	Intron 7	OFD1		+		+
PED4587.1	SCV008086523	*OFD1*	NM_003611.3	c.710dupA	p.Tyr238Valfs*2	frameshift	Exon 8	OFD1		+		+
ID72	SCV008086523	*OFD1*	NM_003611.3	c.710dupA	p.Tyr238Valfs*2	frameshift	Exon 8	OFD1	+	+		+
PZ. 180	SCV008086523	*OFD1*	NM_003611.3	c.710dupA	p.Tyr238Valfs*2	frameshift	Exon 8	OFD1		+		+
PED0251.1	SCV008086523	*OFD1*	NM_003611.3	c.710dupA	p.Tyr238Valfs*2	frameshift	Exon 8	OFD1	+	+	+	+
PED0241.1	SCV008086524	*OFD1*	NM_003611.3	c.712del	p.Tyr238Metfs*5	frameshift	Exon 8	OFD1		+		+
PZ. 152	SCV008086525	*OFD1*	NM_003611.3	c.790del	p.Glu264Lysfs*27	frameshift	Exon 8	OFD1	+	+	+	+
PED0243.1	SCV008086526	*OFD1*	NM_003611.3	c.813_814insA	p.His272Thrfs*7	frameshift	Exon 8	OFD1		+		+
Case 1	SCV008086527	*OFD1*	NM_003611.3	c.877_878del	p.Met293Glyfs*15	frameshift	Exon 9	OFD1	-	+	+	+
Case 2	SCV008086527	*OFD1*	NM_003611.3	c.877_878del	p.Met293Glyfs*15	frameshift	Exon 9	OFD1	+	-	+	+
PED0240.1	SCV008086528	*OFD1*	NM_003611.3	c.898_899dup	p.Ala301Lysfs*4	frameshift	Exon 9	OFD1	+	+	+	+
PED0257.1	SCV008086529	*OFD1*	NM_003611.3	c.913_917del	p.Gln305Serfs*2	frameshift	Exon 9	OFD1		+	+	+
PED0244.1	SCV008086530	*OFD1*	NM_003611.3	c.1056-2A>G	p.?	splicing	Intron 10	OFD1		+	+	+
PED0270.1	SCV008086531	*OFD1*	NM_003611.3	c.1056-1G>A	p.?	splicing	Intron 10	OFD1	+			+
PED2161.1	SCV008086532	*OFD1*	NM_003611.3	c.1061_1065del	p.Gln354Argfs*5	frameshift	Exon 11	OFD1	+	+	+	+
PED4775.1	SCV008086533	*OFD1*	NM_003611.3	c.1100delinsAA	p.Arg367Glnfs*4	frameshift	Exon 11	OFD1		+		+
PED0266.1	SCV008086534	*OFD1*	NM_003611.3	c.1103_1106del	p.Ile369Lysfs*18	frameshift	Exon 11	OFD1	+	+	+	+
PZ.227	SCV008086534	*OFD1*	NM_003611.3	c.1106_1109del	p.Ile369Lysfs*18	frameshift	Exon 11	OFD1	+		+	+
PZ. 181	SCV008086535	*OFD1*	NM_003611.3	c.1117_1121del	p.Arg373Glufs*2	frameshift	Exon 11	OFD1		+	+	+
PED1547.1	SCV008086536	*OFD1*	NM_003611.3	c.1205G>A	p.Arg402His	missense	Exon 12	OFD1				
POL	SCV008086537	*OFD1*	NM_003611.3	c.1222-2A>C	p.?	splicing	Intron 12	OFD1			+	
PZ. 189	SCV008086538	*OFD1*	NM_003611.3	c.1268_1272del	p.Gln423Profs*5	frameshift	Exon 13	OFD1		+		+
PED0201.1	SCV008086539	*OFD1*	NM_003611.3	c.1405C>A	p.Leu469Ile	missense	Exon 13	OFD1	+	+		+
PED0253.1	SCV008086540	*OFD1*	NM_003611.3	c.1604dup	p.Leu535Phefs*22	frameshift	Exon 15	OFD1		+		+
PED0262.1	SCV008086541	*OFD1*	NM_003611.3	c.1613_1620del insTTTACA	p.Gln538Leufs*18	frameshift	Exon 15	OFD1		+	+	+

PZ. 301	SCV008086542	*OFD1*	NM_003611.3	c.1639A>G	p.Lys547Glu	missense	Exon 15	OFD1	-		+	
PED1911.1	SCV008086543	*OFD1*	NM_003611.3	c.1878del	p.Glu626Aspfs*3	frameshift	Exon 16	OFD1		+		+
PZ. 303	SCV008086544	*OFD1*	NM_003611.3	c.1960C>T	p.Arg654Trp	missense	Exon 16	RP23	-	-	-	-
#3588	SCV008086545	*OFD1*	NM_003611.3	c.2273C>G	p.Ser758*	nonsense	Exon 17	PCD	-	-	+	-
#3845	SCV008086545	*OFD1*	NM_003611.3	c.2273C>G	p.Ser758*	nonsense	Exon 17	PCD		+	+	-
PED0200.1	SCV008086546	*OFD1*	NM_003611.3	c.2387+79_2387+80insT	p.?	splicing	Intron 18	OFD1		+	+	+
DCP1170	SCV008086547	*OFD1*	NM_003611.3	c.2638G>T	p.Glu880*	nonsense	Exon 20	JBS10		-	+	-
PZ. 228	SCV008086548	*OFD1*	NM_003611.3	c.2725C>T	p.Arg909*	nonsense	Exon 20	PCD		-	-	-
PZ. 229	SCV008086548	*OFD1*	NM_003611.3	c.2725C>T	p.Arg909*	nonsense	Exon 20	JBS10			+	
PZ. 304	SCV008086548	*OFD1*	NM_003611.3	c.2725C>T	p.Arg909*	nonsense	Exon 20	PCD	-	+		+
8822GM000025	SCV008086549	*OFD1*	NM_003611.3	c.2725C>T	p.Arg909*	nonsense	Exon 20	PCD	-	-	+	-
17GM02557	SCV008086550	*OFD1*	NM_003611.3	c.2806del	p.Met936Trpfs*5	frameshift	Exon 21	JBS10			+	
PZ. 174	-	*OFD1*	Genomic	c.(?-*1)*(2806+1_?)del	p.?	Large deletion	Xp22.2	OFD1		+		+
PZ. 137	-	*OFD1*	Genomic	c.1056-?_1221+?del	p.?	Large deletion	Exons 10–11	OFD1	+	+		+
PZ. 175	-	*OFD1*	Genomic	c.1_917+?del	p.?	Large deletion	Exons 1–9	OFD1		+		+

712

Table 1. Molecular and clinical features of the novel *OFD1* variants identified in our patient cohort. Small variants are described in compliance with the Human Genome Variation Society (HGVS) guidelines using the canonical *OFD1* RefSeq transcript NM_003611.3 as the reference sequence. Nucleotide numbering reflects cDNA positions.

Abbreviations:

- CYS: Cystic kidney disease / Polycystic kidney disease
- SK: Skeletal abnormalities
- NC: Central Nervous System involvement
- OCF: Oro-facio-digital features
- OFD1: Orofaciodigital syndrome 1
- PCD: Primary Ciliary Dyskinesia
- JBS10: Joubert syndrome 10
- RP23: Retinitis Pigmentosa 23
- (+) Clinical feature present; (-) Clinical feature absent; (blank spaces indicate unavailable or non-applicable clinical data).

Note on protein nomenclature: In accordance with standard HGVS notation, the protein consequence for variants mapping within canonical splice sites (intronic positions +1, +2, -1, -2) or deep intronic regions is designated as p.?, as the exact impact on the mRNA transcript has not been experimentally validated via functional assays (e.g., RT-PCR/RNA-Seq).

Note on Structural Variants: For patients PZ.174, PZ.137, and PZ.175, precise genomic breakpoints at the single-nucleotide resolution are unknown. Consequently, cDNA descriptions for these entries employ HGVS uncertainty question markers (?) to indicate uncertain boundaries, and single-entry ClinVar Accession IDs are not applicable.

Supplementary tables legends:

Table S1. ClinVar benign non-synonymous *OFD1* variants. List of 39 missense variants classified as benign retrieved from the ClinVar database, used as a baseline of structural tolerance.

Table S2. OFD1 protein functional domains. Localization of structural motifs, interaction domains, and specific binding sites within the OFD1 protein sequence based on the literature.

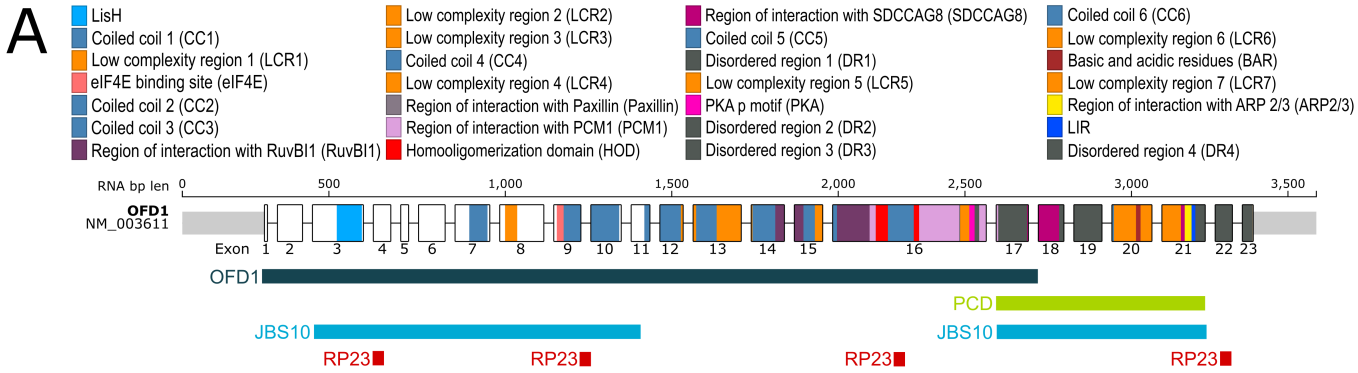
Table S3. Comprehensive cohort of *OFD1*-associated variants. Detailed clinical and genetic information for the 356 individuals carrying molecularly confirmed non-synonymous variants.

Table S4. Predictive performance of individual algorithms vs SuperLearner. Summary of standard cross-validation metrics (including AUC, balanced accuracy, and Brier score) across *OFD1* phenotypes.

Table S5. Nested Cross-Validation performance evaluation. Confirmation of the ensemble model's stability across phenotypic categories evaluated through rigorous Nested Cross-Validation to exclude data leakage.

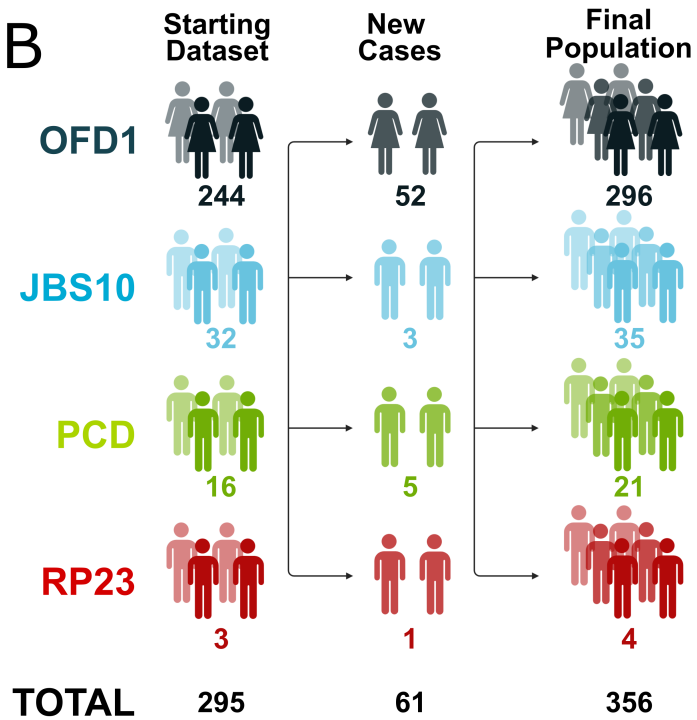
Table S6. Machine learning prediction of *LMNA*-associated phenotypes. Cross-gene validation metrics demonstrating the transferability and performance of the SuperLearner framework on the pleiotropic *LMNA* gene.

A

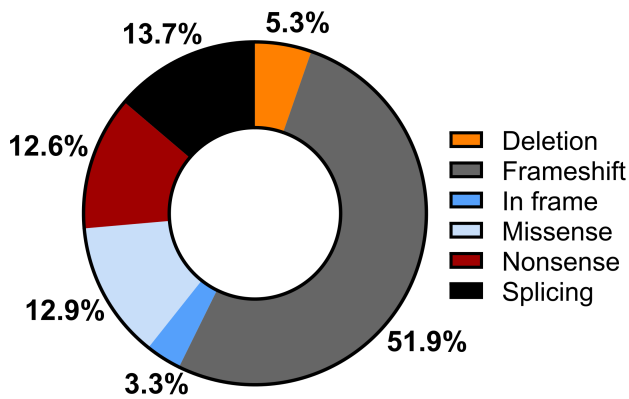


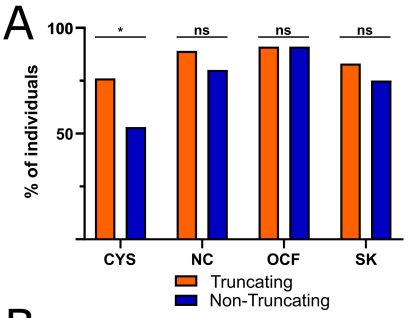
Oro-facio-digital syndrome type 1 (OFD1)	Joubert Syndrome Type 10 (JBS10)	Primary Ciliary Dyskinesia (PCD)	Retinitis pigmentosa type 23 (RP23)
•Craniofacial abnormalities •Cystic disease •Skeletal defects •CNS malformations •Cognitive impairment	•Molar tooth sign •Craniofacial abnormalities •Cystic disease •Skeletal defects •Retinal dysfunction •Cognitive impairment •Motile cilia dysfunction	•Recurrent respiratory infections •Decreased fertility	•Retinal dysfunction •Decreased visual fields •Abnormal color vision •Retinal pigment epithelium atrophy

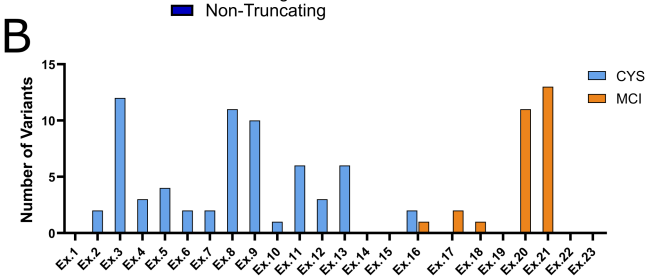
B

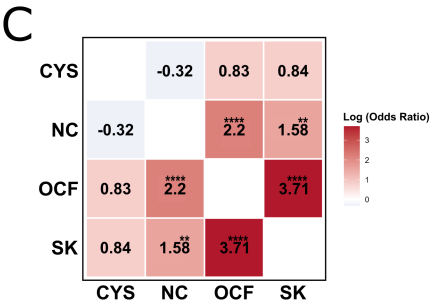


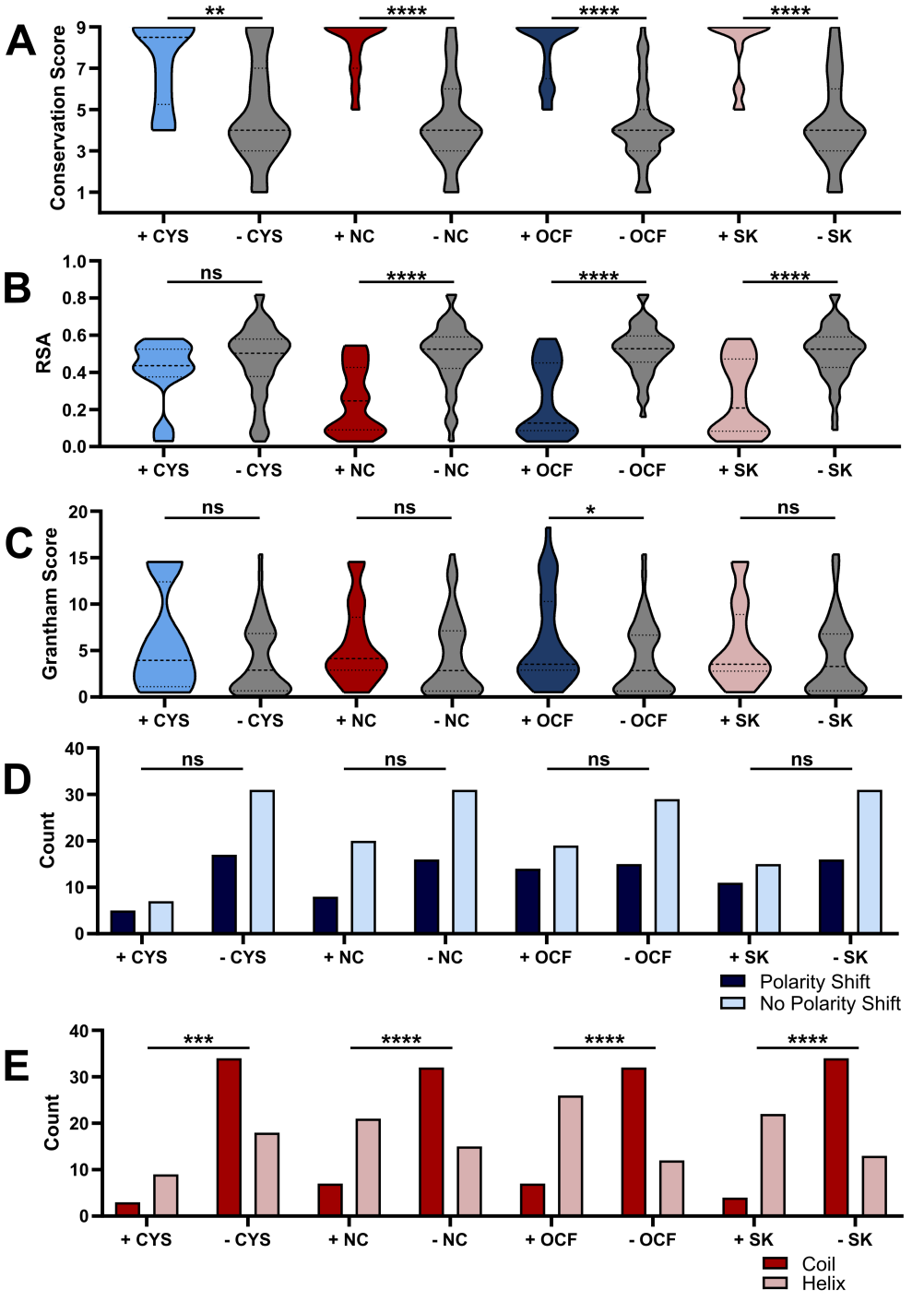
C

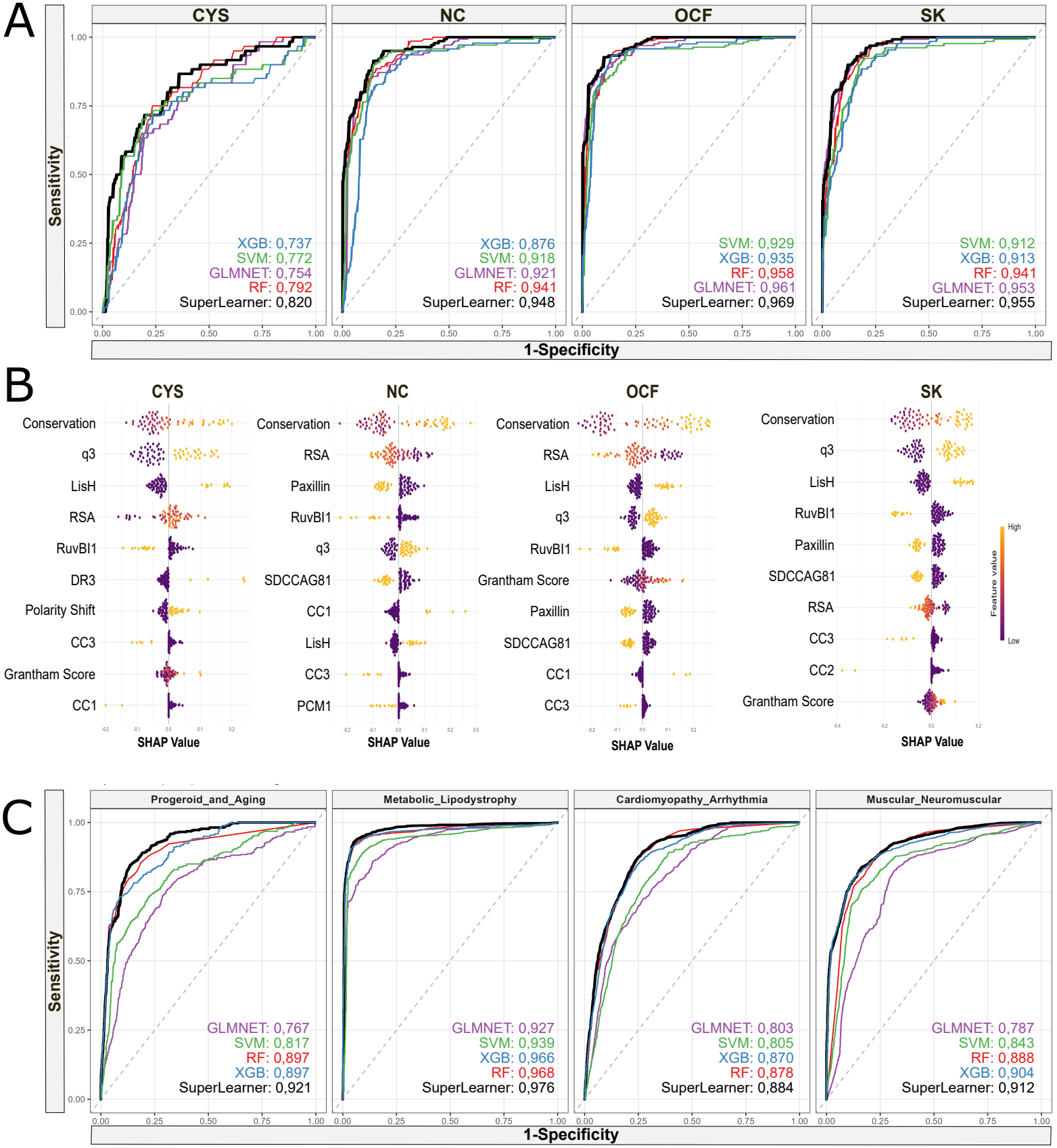



A
100
50
% of individuals
*
ns
ns
ns
CYS
NC
OCF
SK
Truncating
Non-Truncating


B
15
10
5
0
Number of Variants
CYS
MCI
Ex.1
Ex.2
Ex.3
Ex.4
Ex.5
Ex.6
Ex.7
Ex.8
Ex.9
Ex.10
Ex.11
Ex.12
Ex.13
Ex.14
Ex.15
Ex.16
Ex.17
Ex.18
Ex.19
Ex.20
Ex.21
Ex.22
Ex.23


C
CYS
NC
OCF
SK
-0.32
0.83
0.84
-0.32
****
2.2
**
1.58
0.83
****
2.2
****
3.71
0.84
**
1.58
****
3.71
CYS
NC
OCF
SK
Log (Odds Ratio)
3
2
1
0


A
Conservation Score
9
7
5
3
1
**
****
****
****
+ CYS
- CYS
+ NC
- NC
+ OCF
- OCF
+ SK
- SK
B
RSA
1.0
0.8
0.6
0.4
0.2
0.0
ns
****
****
****
C
Grantham Score
20
15
10
5
0
ns
ns
*
ns
D
Count
40
30
20
10
0
ns
ns
ns
ns
Polarity Shift
No Polarity Shift
E
Count
40
30
20
10
0
***
****
****
****
Coil
Helix


A
CYS
NC
OCF
SK
Sensitivity
1-Specificity
XGB: 0,737
SVM: 0,772
GLMNET: 0,754
RF: 0,792
SuperLearner: 0,820
XGB: 0,876
SVM: 0,918
GLMNET: 0,921
RF: 0,941
SuperLearner: 0,948
SVM: 0,929
XGB: 0,935
RF: 0,958
GLMNET: 0,961
SuperLearner: 0,969
SVM: 0,912
XGB: 0,913
RF: 0,941
GLMNET: 0,953
SuperLearner: 0,955
B
CYS
NC
OCF
SK
Conservation
q3
LisH
RSA
RuvBI1
DR3
Polarity Shift
CC3
Grantham Score
CC1
Conservation
RSA
Paxillin
RuvBI1
q3
SDCCAG81
CC1
LisH
CC3
PCM1
Conservation
RSA
LisH
q3
RuvBI1
Grantham Score
Paxillin
SDCCAG81
CC1
CC3
Conservation
q3
LisH
RuvBI1
Paxillin
SDCCAG81
RSA
CC3
CC2
Grantham Score
SHAP Value
Feature value
High
Low
C
Progeroid_and_Aging
Metabolic_Lipodystrophy
Cardiomyopathy_Arrhythmia
Muscular_Neuromuscular
Sensitivity
1-Specificity
GLMNET: 0,767
SVM: 0,817
RF: 0,897
XGB: 0,897
SuperLearner: 0,921
GLMNET: 0,927
SVM: 0,939
XGB: 0,966
RF: 0,968
SuperLearner: 0,976
GLMNET: 0,803
SVM: 0,805
XGB: 0,870
RF: 0,878
SuperLearner: 0,884
GLMNET: 0,787
SVM: 0,843
RF: 0,888
XGB: 0,904
SuperLearner: 0,912

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS3.docx
- TableS4.xlsx
- TableS6.xlsx
- TableS1.xlsx
- TableS2.docx
- TableS5.xlsx