Jing Zhang ORCID iD: 0000-0003-4190-3065

Panagiotis Katsonis ORCID iD: 0000-0002-7172-1644

Nick Grishin ORCID iD: 0000-0003-4108-1153

Assessing predictions on fitness effects of missense variants in calmodulin

Jing Zhang[2], Lisa N. Kinch[1], Qian Cong[2], Panagiotis Katsonis[3], Olivier Lichtarge[3], Castrense Savojardo[4], Giulia Babbi[4], Pier Luigi Martelli[4], Emidio Capriotti[4], Rita Casadio[4], Aditi Garg[5], Debnath Pal[5], Jochen Weile[6,7,8], Song Sun[6,7,8], Marta Verby[6,7,8], Frederick P. Roth[6,7,8,9] and Nick V. Grishin[1,2,#]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA.

[2]Departments of Biophysics and Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390-8816, USA.

[3]Department of Molecular and Human Genetics, Department of Biochemistry & Molecular Biology, Department of Pharmacology, Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX, USA.

[4]Biocomputing Group, FABIT/Giorgio Prodi Interdepartmental Center for Cancer Research, University of Bologna, Via F. Selmi 3, Bologna, 40126, Italy.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.23857.

[5]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, 560 012, India.

[6]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto Ontario M5G 1X5, Canada.

[7]The Donnelly Centre and Departments of [8]Molecular Genetics and [9]Computer Science University of Toronto, Toronto, Ontario M5S 3E1, Canada.

Corresponding author:

[#]Nick V. Grishin

Howard Hughes Medical Institute,

University of Texas Southwestern Medical Center,

5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, USA.

Email: grishin@chop.swmed.edu

Grant numbers:

The CAGI experiment coordination is supported by NIH U41 HG007346 and the CAGI conference by NIH R13 HG006650. The assessment of this challenge is supported by grants (to NVG) from the National Institutes of Health GM127390 and the Welch Foundation I-1505. OL and PK were supported by the NIH-GM079656 and the NIH-GM066099 grants.

Abstract

This paper reports the evaluation of predictions for the "CALM1" challenge in the 5th round of the Critical Assessment of Genome Interpretation held in 2018. In the challenge, the participants were asked to predict effects on yeast growth caused by missense variants of human calmodulin, a highly conserved protein in eukaryotic cells sensing calcium concentration. The performance of predictors implementing different algorithms and methods is similar. Most predictors are able to identify the deleterious or benign variants with modest accuracy, with a baseline predictor based purely on sequence conservation slightly outperforming the submitted predictions. Nevertheless, we think that the accuracy of predictions remains far from satisfactory, and the field awaits substantial improvements. The most poorly predicted variants in this round surround functional CALM1 sites that bind calcium or peptide, which suggests that better incorporation of structural analysis may help improve predictions.

Key words: CAGI, predictors, missense variants, calmodulin, disease

Introduction

The Critical Assessment of Genome Interpretation (CAGI), round 5, is aimed to provide an objective evaluation of computational methods for predicting phenotypic impacts of genomic variations. There are 14 challenges in round 5, and we present here the assessment of challenge called "CALM1". In this challenge, fitness scores were provided by a complementation assay developed in Fritz Roth's Lab. The assay evaluated the ability of human CALM1 missense variants to rescue a temperature sensitive mutation of the yeast ortholog CMD1 (Weile et al., 2017). Conceptually, the fitness score represents

the relative growth rate of yeast with *CALM1* missense variants to that of yeast with wild type *CALM1*. Thus, the deleterious missense variants have fitness scores closer to 0 and benign variants have fitness score closer to 1. In the challenge, participants were asked to predict fitness scores for 1813 missense variants of CALM1. Although exact values of experimental fitness scores were not given, the distribution was provided to help normalize predictions.

Calmodulin (CALM1) is a calcium-sensing protein that modulates the activity of a large number of proteins in the cell. It has dumbbell-shaped structure composed of two globular domains connected by a flexible linker (Babu, Bugg, & Cook, 1988). Each globular domain has two calcium-binding motifs that make up an EF-hand. As a calcium sensor, calmodulin is involved in numerous cellular processes, and is especially important for normal function of neuron and muscle cells. Its variants have been found to be causally associated with two diseases, ventricular tachycardia, catecholaminergic polymorphic, 4 and long QT syndrome 14 (Boczek et al., 2016; Nyegaard et al., 2012; Yu et al., 2016). Choosing calmodulin as a target to assess the current state-of-the-art in computational methods for variance prediction has a couple of advantages. First, calmodulin is ubiquitous in most eukaryotic cells (Stevens, 1983) providing numerous sequence homologs for sequence analysis. Second, numerous structures of calmodulin complexes are available (Drum et al., 2002; Meador, Means, & Quiocho, 1992; Shen, Zhukovskaya, Guo, Florian, & Tang, 2005). These structures aid in understanding the functional relevance of mutations. Third, various studies have been done to decipher how calmodulin is involved in different pathways (Berchtold & Villalobo, 2014; Parry & June, 2003; Sorensen, Sondergaard, & Overgaard, 2013; Stull, 2001). Overall, the

abundance of existing knowledge for calmodulin permits various methods to be applied and thus is a good target for evaluating computational methods.

Current predictors can be divided into three main types according to their features: (1) prediction based on sequence conservation; (2) incorporation of both sequence and structural information; (3) integration of predictions from several predictors. We received 7 predictions from 4 groups, which include all three mentioned types of methods. The predictors included two published methods: Evolutional Action (group 1) (Katsonis & Lichtarge, 2014) and INPS3D (group 3) (Savojardo, Fariselli, Martelli, & Casadio, 2016). Group 2 used average values from PhD-SNP (Capriotti, Calabrese, & Casadio, 2006), PANTHER (Thomas et al., 2003) and SNPs&GO (Calabrese, Capriotti, Fariselli, Martelli, & Casadio, 2009) and group 4 used molecular dynamics. The assessment shows that all predictors except group 4 are capturing qualitative (e.g., deleterious vs. benign) effects of variants on proteins. However, the quantitative agreement between predictions and experimental measures remains modest. Most predictors are able to differentiate deleterious variants and benign variants. However, the accuracy of exact values is waiting for substantial improvements.

MATERIALS AND METHODS

Positive control and the baseline predictor

As in CAGI4, we defined a positive control and a baseline predictor. The positive control consists of fitness scores for each variant randomly drawn from an assumed Gaussian distribution with the given fitness score as mean and the experimental standard error as standard deviation. The baseline predictor was based on the frequency of amino acids

at each position in a CALM1 multiple sequence alignment (MSA). About 1133 ortholog/inparalog sequences of calmodulin were extracted from orthoDB (Kriventseva et al., 2015) at the metazoa level and were aligned using Promals3D (Pei & Grishin, 2014). The original predicted score for each variant was calculated using the following formula:

$$ln\frac{Q_m}{P_m} - ln\frac{Q_w}{P_w}$$

where Q_m and Q_w are the estimated probabilities of mutated and wild-type amino acids at a mutated position in the alignment as defined in, and P_m and P_w are Robinson-Robinson background frequencies of the mutated and wild-type amino acids. The original predicted scores were normalized according to the distribution of experimental fitness scores.

Quantile transformation of original predictions

Although the distribution of experimental fitness scores was provided, most participants did not calibrate their predictions using it. Thus, normalization of predictions was required to make predictors comparable in their scale, which is especially important for numeric comparison. We performed quantile transformation of the original predictions from participants and of our baseline predictor. Because predictors were not allowed to predict negative values,. all negative competitive growth scores were shifted to 0 before transformation. The variants were ranked by the predicted values, and each variant was assigned the experimental score with the same rank. The assigned experimental scores

for mutants that are predicted to be ties are further averaged to obtain the final transformed predictions.

Measures for prediction assessment

Each predictor was evaluated by their ability (1) to classify variants into categories such as deleterious and non-deleterious variants (classification), (2) to rank variants by their impacts on yeast fitness (ordinal association), and (3) to predict experimental fitness scores (numeric comparison). For the assessment, variants were assigned to the following categories by their experimental fitness score: less than 0.3 for deleterious, between 0.3 and 0.8 for intermediate, from 0.8 to 1.0 for wild type. Table 1 summarizes all scores used for the evaluation.

Evaluation of overall performance and its statistical significance

Four of the measures listed in Table 1 (i.e. the three ordinal associations and the AUC) are purely based on rank and are not sensitive to the distribution of numeric values. Five others depend on the distribution of numeric values and thus were calculated with both original and quantile-transformed predictions. For each measure, we transformed the original scores to Z scores, and positive control and baseline predictor were excluded from the calculation of mean and standard deviation of original scores to avoid their influence on the score distribution. The average Z scores of the rank-based, original-value-based, and transformed-value-based measures were computed and summed up to be the final score to assess the performance on each subset.

To take experimental errors into consideration, we assumed that the fitness score for each variant can be randomly drawn from a Gaussian distribution defined by the reported fitness score and the standard error. We simulated 50 datasets using above method. Then, we performed bootstrap resampling on each simulated dataset 100 times, and thus generated 5000 mock datasets. We obtained the distribution of ranks for each group on 5000 mock datasets.

Results

Most variants have minor or no effects on yeast survival.

The distribution of experimental fitness scores of variants is depicted in Figure 1. Negative fitness scores were shifted to 0, as the challenge requires non-negative predictions. A majority of variants are either detrimental or benign, and thus the distribution is bimodal. About 71% of variants with fitness scores equal or above 0.6 and 56% of variants with fitness scores equal or above 0.8, suggesting variants are biased towards being benign to yeast survival.

Functional suggestions of variants

The Calmodulin gene CALM1 encodes a Ca-binding protein with two tandem EF-hand domains. CALM1 structures adopt various different conformations in response to Ca, and provide selectivity for interacting with cellular targets to drive a wide range of biological processes (Bhattacharya, Bunick, & Chazin, 2004). CALM1 achieves specific recognition of these targets by adopting multiple conformations with and without bound Ca. Figure 2A and B highlight two such alternate conformations. The calmodulin EF-hands bind to the IQ domain of the Ca(v)1.2 Ca2+ channel in a compact parallel

conformation (peptide2 binding mode, Figure 2A), with both domains binding the peptide through hydrophobic surfaces (Fallon, Halling, Hamilton, & Quiocho, 2005). Calmodulin binds the inactivation gate (DIII-IV linker) of the cardiac sodium channel in an alternate extended conformation (peptide1 binding mode, Figure 2B), with the CALM1 C-lobe contacting the bound peptide.

Each of the structures binds four Ca2+ ions, with each site using four key acidic residues. The experimental fitness scores for multiple mutations at these Ca-binding sites are plotted in figure 2C. While a few mutations of key Ca-binding residues are detrimental, most exhibit intermediate and benign fitness. This skewed fitness distribution of functional mutations suggests that Ca binding might be redundant. Indeed, fitness scores mapped to the extended CALM1 structure highlight the extreme difference between minimum scores measuring detrimental mutations (Figure 3A) and mean scores highlighting a broader range of fitness levels (Figure 3B) with respect to the peptide1 binding mode. The distribution of minimum mean and maximum fitness scores for residues contacting peptide in both binding modes, peptide1 binding mode and peptide2 binding mode are plotted in comparison to the same distributions of Ca binding residues (Figure 3C). The distributions suggest peptide1 binding mode might contribute more to fitness than peptide2. The relatively lower fitness of the C-terminal Ca binding residues, which contribute to peptide1 binding, also support this notion.

Predictions and experimental fitness scores have disparate distributions

We also plotted the distribution of predicted fitness scores of each participant in Figure 1. Unfortunately, most participants did not normalize their predictions according to the given distribution of experimental fitness scores. The Kolmogorov–Smirnov test

indicates that only predictions from group 1 replicate the experimental distribution ($p>0.05$), and the distribution of predictions from group 2-1 is most dissimilar to the experimental distribution. Group 2-1 predicted most variants to have mildly deleterious effects on yeast fitness, with few variants predicted to be benign. Considering that different scales of predictions may bias evaluation and conceal the real capacity of predictors to detect the effects of variants, we applied quantile transformation of predicted values of each group to make the results comparable with each other.

Overall performance of predictors is comparable and far behind accuracy

A similar evaluation strategy (Table 1) as CAGI4 is applied to the predictions from this round to assess the ability of methods to (1) classify variants; (2) rank variants by their effects on fitness and (3) numerically predict fitness scores of variants. The performance of the predictors on each measure is shown in Table 2. All participants except group 4 show significantly better than random predictions where the best performing group exhibits a Kendall's tau correlation of 0.17. As in the previous CAGI4 round, a baseline control calculated from amino acid frequency in a multiple sequence alignment has comparable and even slightly better performance with the other predictors in this challenge. However, the baseline control stands out more with respect to quantitative metrics as compared to qualitative measures. Group 1, group 2 and the baseline predictor are on par with each other in their ability to rank variants' effects on yeast fitness. Group 2 marginally outperforms the other two, but the worse original numeric predictions make it rank behind. When comparing group 1 and the baseline predictor, whose prediction distributions resemble the experimental fitness score distribution more closely, the baseline predictor outperformed group 1 in classifying deleterious variants using either

original predicted values or re-scaled predicted values as criteria. These results suggest the baseline predictor has surpassing ability to identify extremely detrimental variants.

To access the significance of our evaluation, we simulated 5000 datasets by assuming a Gaussian distribution of fitness scores of each variant and using the experimental fitness score and standard error as mean and standard deviation for the distribution, respectively. For each simulated dataset, we calculated assessment measures and obtained a Z score for each prediction. The distributions of Z scores of predictors (Figure 4A) do not show clear separation and cover similar range, suggesting comparable performance of several predictors. A striking gap between all predictors and the positive control suggests substantial improvements are needed for accurate predictions. Consistent with Z score results, the distribution of ranks on 5000 simulated datasets exhibits a tie between the baseline predictor and group 1 (Figure 4B). Intriguingly, both predictors (baseline and group 1) normalized their predictions to the distribution of experimentally determined fitness scores.

Modest performance for predicting deleterious variants and wild type variants

Differentiating deleterious variants and benign variants *in silico* is considered the major challenge for current computational methods. Thus, we specifically evaluated predictors' ability to identify deleterious (fitness score <0.3) and benign (fitness score >=0.8) variants. A receiver operating characteristic (ROC) curve exhibits the diagnostic ability of predictors to classify variants into deleterious or benign. The area under ROC curve indicates the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one. The ROC curves for group 1, 2 and the

baseline predictor are tangled together suggesting equivalent performance in classifying deleterious variants (Figure 5). The higher true positive rate at the beginning of the ROC (low false positive) for the baseline predictor implies the most detrimental variants predicted by baseline predictor are more likely to be truly detrimental compared with other predictors. We also calculated Matthews correlation coefficients (MCCs) to evaluate predictors' performance to classify deleterious or benign variants. MCC for classifying benign variants of group1 and group2 is higher than that for classifying benign variants of two groups, indicating that the predictors are more reliable in detecting benign variants (Table 2).

Inaccurate predictions on calcium-binding sites and peptide binding sites

The average performance of predictors for each position along the primary sequence of CALM1 is shown in a heatmap (Figure 6A), scaled from green (good performance) to red (poor performance). The performance of predictions around calcium-binding sites is below average. Variants for calcium binding site residues were predicted to be detrimental by most predictors, yet most variants did not exhibit obvious effects on yeast growth (Figure 2C). For example, position D21 is one of the sites where most of variants' effects are poorly predicted. D21 coordinates Ca in the first EF-hand Ca-binding motif and is conserved among vertebrate CALM1 orthologs. Given this conservation and contribution to function, the lack of detrimental variants at this position is surprising and might suggest that the first Ca binding site in human CALM1 does not contribute to fitness in the yeast complementation system.

Meanwhile, a number of CALM1 variants were generally predicted as benign, but they exhibited detrimental experimentally determined competitive fitness scores (Figure 6B-

D). One of these poorly predicted variants, Q136M, maps to the C-terminal EF-hand lobe near the Ca binding site (within 4Å, Figure 6B). This residue displays relatively low conservation and does not coordinate the Ca in the extended structure, which likely resulted in the tendency for benign predictions. However, the experimental fitness score for this variant was zero, suggesting that the swap from a polar side chain to a hydrophobic one is detrimental. The backbone of this residue coordinates Ca and perhaps requires a polar side chain interacting with the surrounding solution to adopt the correct orientation.

Two relatively conservative variants of aromatic side chains to hydrophobic ones (F13M and F69M) were also predicted by the community as benign. While they do not bind peptide1 in the extended conformation of calmodulin bound to a peptide from the cardiac sodium channel (Na(V)1.5, Figure 6C), they do interact with peptide in an alternate peptide2 binding mode (Figure 6D). The wild type aromatic sidechains form pi stacking interactions with aromatic residues from the peptide, potentially explaining the detrimental effect of the variants. An additional poorly predicted as benign variant, Q9A also interacts with the peptide2 binding mode (Figure 6E), suggesting that the altered binding surface caused by the variant is detrimental. Finally, the poorly predicted as benign variant Y100T is also in the Ca binding site (Figure 6F).

Discussions

Fitness scores from yeast complementation assay can be double-edged

Datasets for evaluating mutation fitness are one of most important factors contributing to the conclusions of the assessment. Many predictors are trained using public datasets such

as OMIM (Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2015), dbSNP (Sherry et al., 2001) and ClinVar (Landrum et al., 2018) or directly extract variant information from them. Thus, using public datasets for evaluation may have following disadvantages: 1) biased assessment; 2) overly optimistic performance; 3) inability to extend functional effects to new variants; 4) errors in public databases (Coovadia, 2017; Grimm et al., 2015).

To overcome these shortcomings, the CAGI committee provides a new experimentally determined dataset of variant fitness that is not yet available to the public. This dataset will not have significant overlap with training data used by existing predictors and the large number of missense variants will reveal full capacity of predictors to predict functional effects caused by new variants. However, such datasets do not come without risks. Although yeast and human share a striking number of orthologs and biological pathways, there are numerous human proteins without equivalents in yeast. Some interactors with human calmodulin may be absent in yeast. Human is composed of organs that consist of differentiated cells with disparate functions, but yeast is a single-cell organism. While the yeast complementation assay uses growth rate as criteria to judge effects of variants, calmodulin variants in human may affect other phenotypes unique to higher organisms, such as muscle contraction. Human calmodulin only shares about 60% identity with yeast. Thus, the variants which are deleterious or benign for human may not show the same effects in yeast and vice versa. For example, several disease-related variants in human (N54I, N98S and E141G) with yeast-derived fitness score very close to 1. Thus, the experimentally determined fitness scores did not capture some variants contributing to human disease.

Several predictors show comparable performances and are slightly better than others

Although predictors participating in the challenge use different methodologies, not a single group significantly outperformed the others. The baseline predictor and group 1 perform slightly better, with both concentrating on sequence conservation and amino acid frequency. However, they are also the only predictors that normalized predictions according the experimental distribution. Thus, their better performance may be due in part to good normalization. Group 2 incorporated predictions from several published predictors by using their average prediction values. This method shows a marginally higher value in Kendall's tau correlation, Spearman’s rank correlation coefficient and area under ROC curve for detecting deleterious variants. Thus, incorporation of predictions from several methods may provide a strategy for improving performance in the future. However, how to integrate predictions from various sources to obtain a significantly finer prediction is unclear.

Group 3 used a published predictor called INSP3D, which is designed to predict protein stability change upon single point mutation from sequence and structures. Its performance is worse than the baseline predictor, group 1 and 2, as it is possible that many variants on calmodulin affect protein-protein interactions or protein conformational changes instead of protein stability. Therefore, its performance in this challenge may not reflect its real ability to predict protein stability change. Group 4 is the only group with worse than random predictions. It has 20% predictions that are anti-correlating, due to which the overall performance indicators become poor. Group 4 used molecular dynamics to estimate the change in the flexibility profile of a mutant with respect to that

of the wild type structure. They hypothesized this change is proportional to the change in the function of a mutant.

The performance of predictors decreased compared with CAGI4

As assessors for both CAGI4 (Zhang et al., 2017) and CAGI5, we noticed that the performance of predictors did not improve in this round. In fact, performance of predictors was slightly worse than in the previous round. This disappointing trend is possibly due to the small number of participants or the short time for observing improvements of methods since last challenge. The median Kendall's tau correlation coefficient for the CALM1 challenge was 0.15, as compared to 0.26 for CAGI4. However, these comparisons might not accurately reflect predictor performance, as experimental determination of fitness scores and choice of protein contribute to the results. The CALM1 yeast ortholog evolves faster in fungi, and budding yeast cells can survive with all EF-hands ablated although CALM1 is essential for yeast (Geiser, van Tuinen, Brockerhoff, Neff, & Davis, 1991). However, disease-related variants in human predominately surround calcium-binding sites in the C-terminus (Jensen, Brohus, Nyegaard, & Overgaard, 2018). Thus, using budding yeast as organism for testing the functional effects of variants at calcium binding positions could be problematic, although the system seems to work reasonably well for pathogenicity prediction (Weile et al., 2017).

A second major difference that could lead to decreasing performance is that UBE2I and calmodulin have different interaction behaviors. The interaction between calmodulin and various targets involves a large interface, a buried surface ranging from 2400–3000 Å2 for calmodulin/peptide and 5900 Å2 for EF/calmodulin complex (Hoeflich & Ikura,

2002). A large interface may lead to difficulty to predict the effects of missense variants on interactions. A variant on interface may decrease the affinity but the difference may not result in any detectable functional effects. It will be difficult to infer quantitative relationship between reduction in affinity of interactions and functional effects and thus results in poor predictions.

Figure 1 The distribution of experimental fitness scores and predictions. The 3D plot depicts the ratio (Y-axis) of fitness scores (X-axis) from experiment (exp) and all participants (depth axis). All negative fitness scores and predictions are shifted to 0, as the challenge requires non-negative predictions.

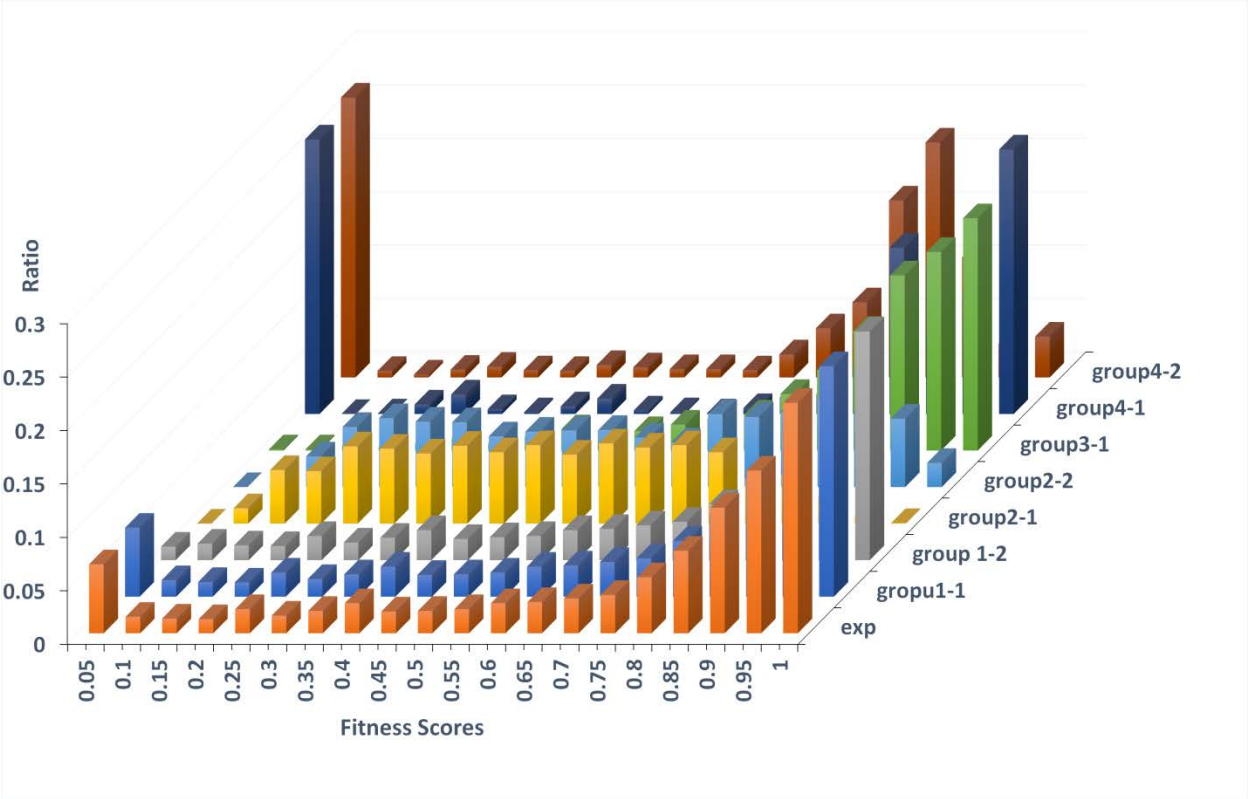


Figure 2 CALM1 Ca-binding Functional Residue Mutations Exhibit Redundancy. Calcium binding sites are labeled numerically (Ca1-Ca4) and colored according to primary sequence order. Calmodulin structures are colored in rainbow from the N-terminus (blue) to the C-terminus (red). **(A)** Compact calmodulin structure conformation depicts Ca-dependent binding to the hydrophobic IQ Domain (pink cartoon, peptide2 mode) of the Cardiac Ca(v)1.2 Calcium Channel [PDB:2f3y]. **(B)** Extended calmodulin structure conformation depicts Ca-dependent binding to the inactivation gate DIII-IV linker (magenta cartoon, peptide1 mode) of the cardiac sodium channel (Na(V)1.5) [PDB: 4djc]. **(C)** Experimental competitive fitness scores (unscaled) for Ca-binding site mutations diverge from wildtype (green section) through intermediate (yellow section) to detrimental (red section).

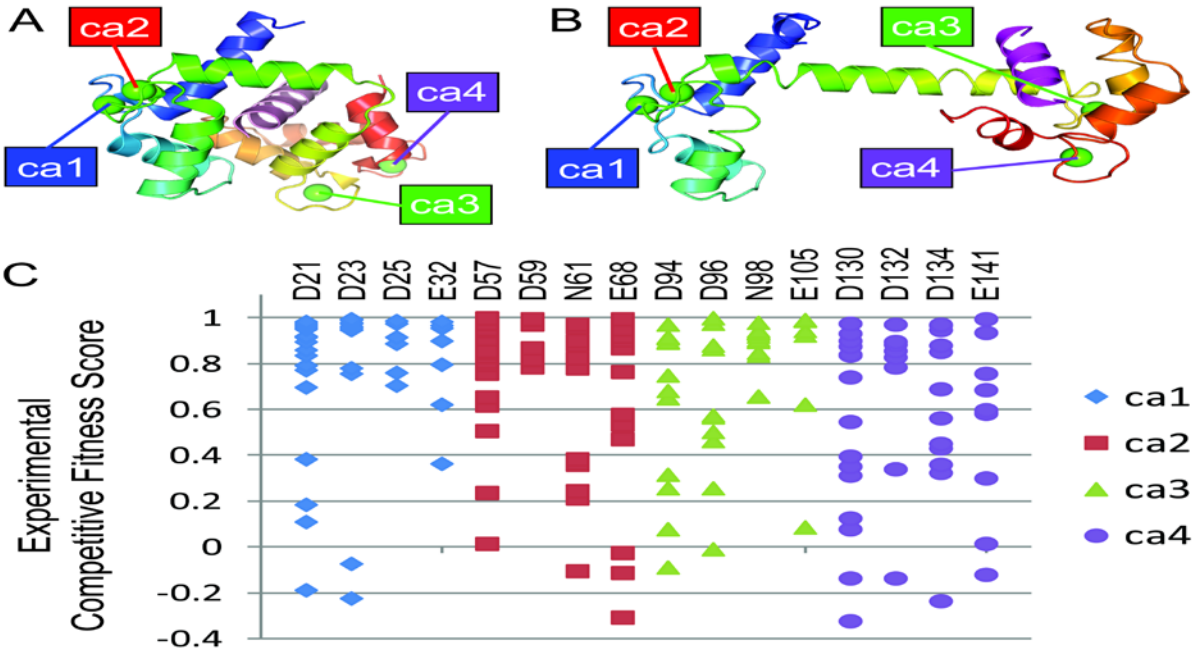


Figure 3 CALM1 Peptide binding Site Mutations Exhibit Diverse Fitness Consequences. Extended CALM1 conformation displayed in surface representation bound to peptide 1 (yellow cartoon). CALM1 residues are colored by scale from damaging (red) to benign (blue) competitive fitness score. **(A)** CALM1 colored by the minimum competitive fitness score per site. **(B)** CALM1 colored by the mean competitive fitness score per site. **(C)** Experimental competitive fitness scores for residues interacting with both peptide binding modes (red), peptide1 binding mode (blue) and peptide2 binding mode (green) on the left are compared to the Ca-binding residues on the right (grey background, same coloring as in Figure 2C). Maximum fitness scores (square symbols), mean fitness scores (triangle symbols), and minimum fitness scores (diamond symbols) per residue position are indicated by a solid line for the respective average over each category of binding mode residues.

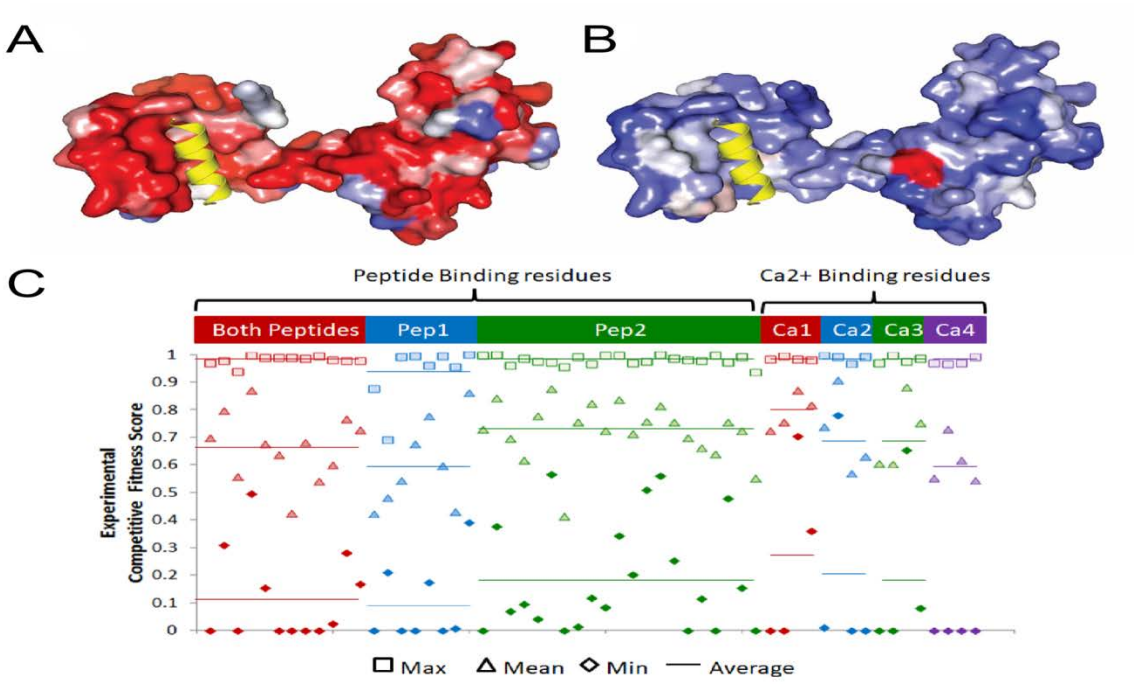


Figure 4 Statistical robustness of Z-scores and rank of predictors. The boxplots illustrate the confidence interval of (A) Z scores and (B) rank of predictor performance. The red lines indicate the median of Z scores/rank, the boxes extend from first quantile to the third quantile and whiskers show the 95% confidence interval range. positive, positive control; baseline, baseline predictor.

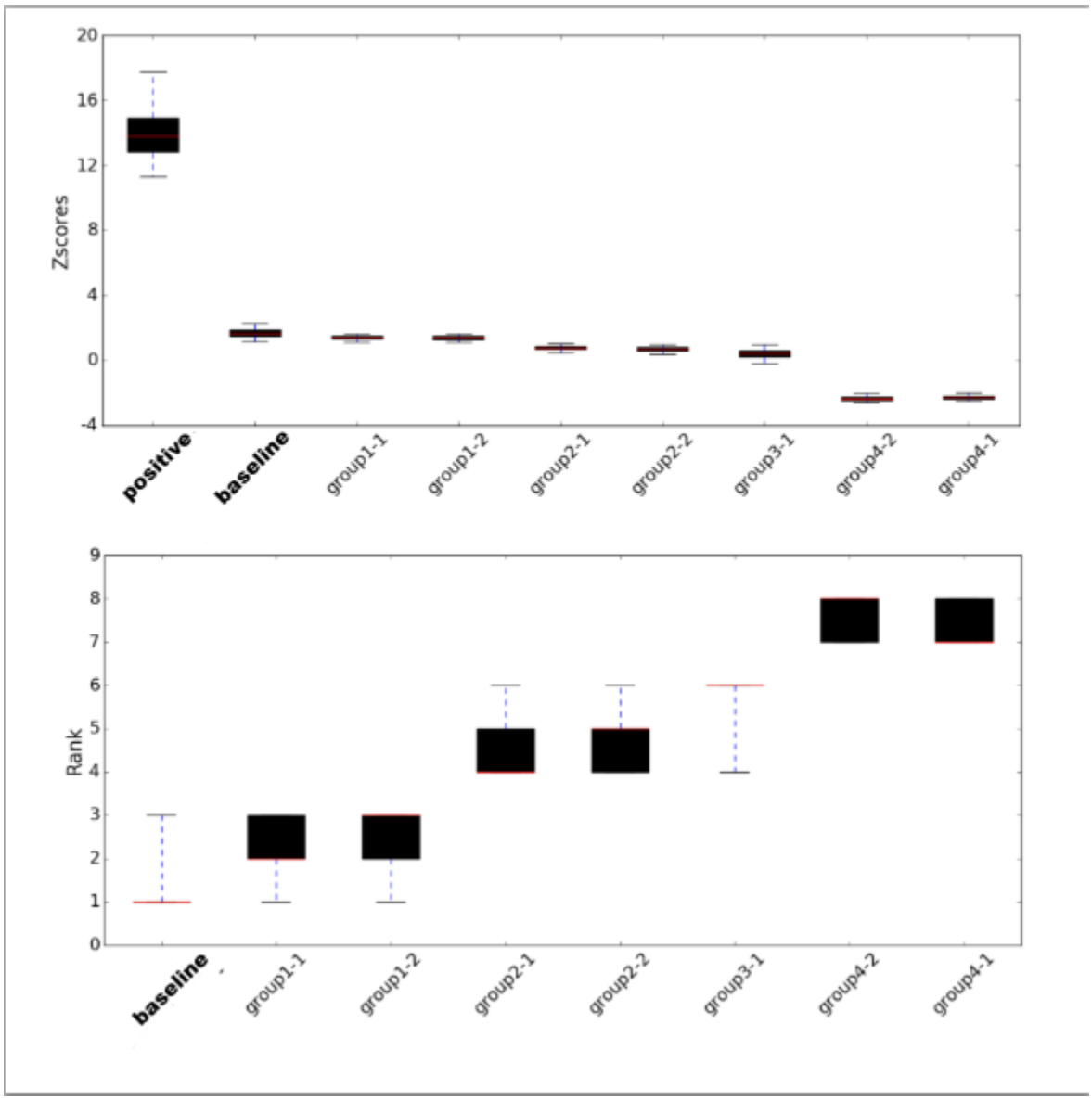


Figure 5 ROC showing performance of predictors for predicting deleterious variants. baseline, baseline predictor; positive, positive control.

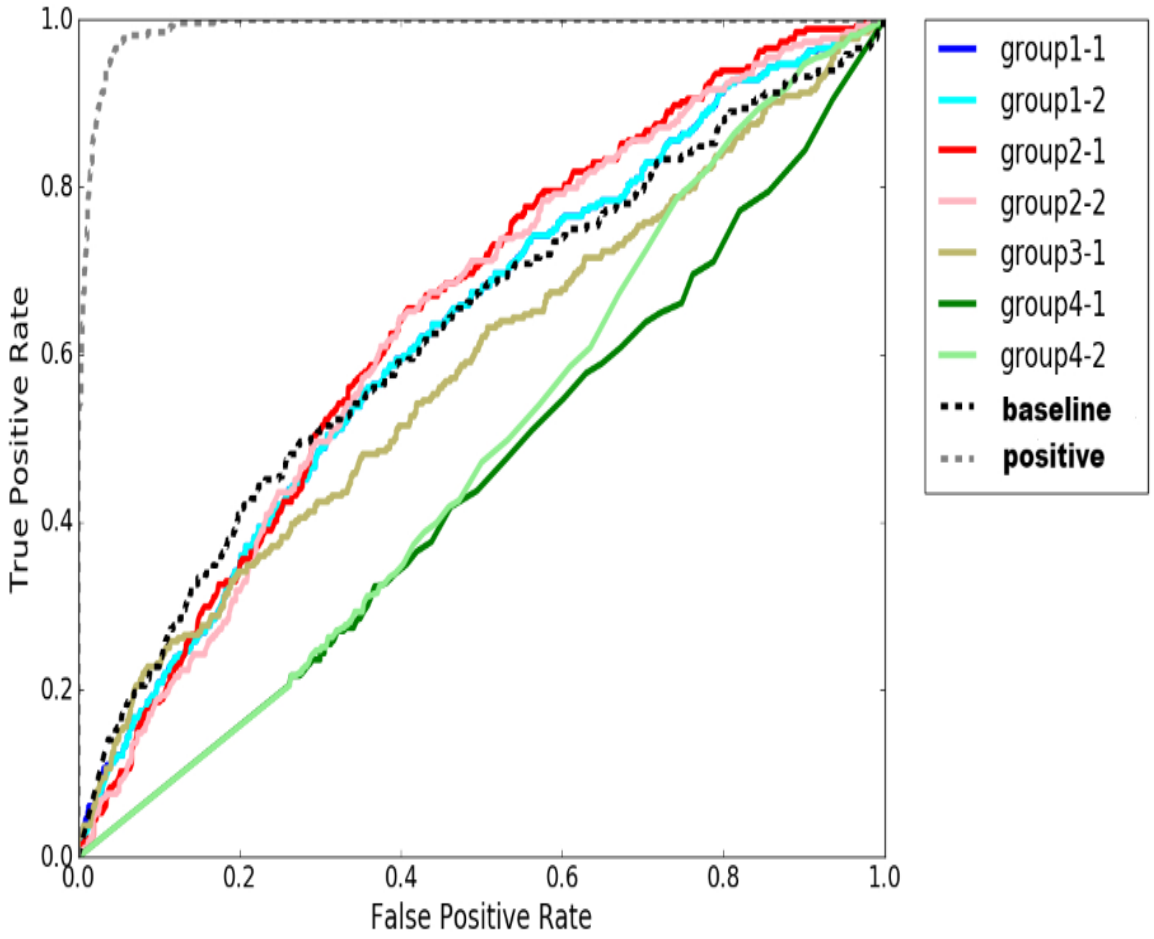
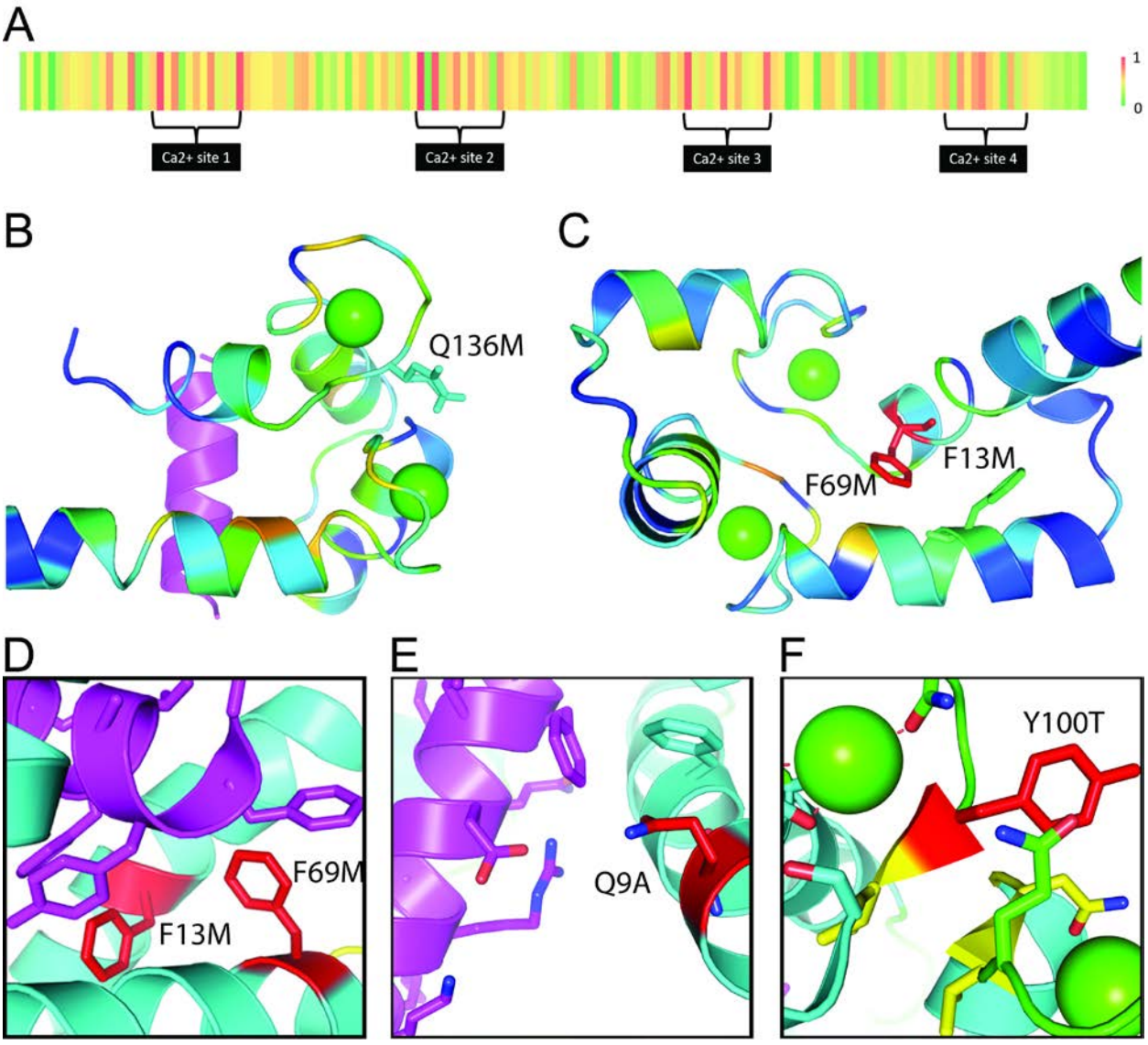


Figure 6 Poorly predicted variants. (A) heatmap of average performance of predictors on each position. The averages of absolute difference between predictions and experimental fitness scores were colored from low (green) to high (red) for each position **(B)** CALM1 extended conformation [PDB 4djc] C-terminal EF-hand lobe is colored in rainbow by residue conservation from blue (variable) to red (conserved). Ca (green sphere) and Peptide1 (magenta cartoon) binding site are near intermediate conserved residues. Q136M (sticks) was generally predicted as benign, yet resulted in a competitive fitness score of zero. **(C)** CALM1 extended conformation N-terminal EF-hand lobe depicted as in A shows position of intermediate F13M (stick) and conserved F69M (stick) in site lacking peptide. **(D)** Zoom of F13/F69 site (red stick) in CALM1 with peptide2 [PDB:2f3y]. Both mutations were predicted as benign but had detrimental fitness (zero). **(E)** Zoom of Q9A (red stick) near peptide2 (magenta cartoon) and **(F)** zoom of Y100T (red stick) near Ca highlight additional benign predictions that were experimentally detrimental.



References

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res, 43*(Database issue), D789-798. doi:10.1093/nar/gku1205

Babu, Y. S., Bugg, C. E., & Cook, W. J. (1988). Structure of calmodulin refined at 2.2 A resolution. *J Mol Biol, 204*(1), 191-204.

Berchtold, M. W., & Villalobo, A. (2014). The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer. *Biochim Biophys Acta, 1843*(2), 398-435. doi:10.1016/j.bbamcr.2013.10.021

Bhattacharya, S., Bunick, C. G., & Chazin, W. J. (2004). Target selectivity in EF-hand calcium binding proteins. *Biochim Biophys Acta, 1742*(1-3), 69-79. doi:10.1016/j.bbamcr.2004.09.002

Boczek, N. J., Gomez-Hurtado, N., Ye, D., Calvert, M. L., Tester, D. J., Kryshtal, D., . . . Ackerman, M. J. (2016). Spectrum and Prevalence of CALM1-, CALM2-, and CALM3-Encoded Calmodulin Variants in Long QT Syndrome and Functional Characterization of a Novel Long QT Syndrome-Associated Calmodulin Missense Variant, E141G. *Circ Cardiovasc Genet, 9*(2), 136-146. doi:10.1161/CIRCGENETICS.115.001323

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat, 30*(8), 1237-1244. doi:10.1002/humu.21047

Capriotti, E., Calabrese, R., & Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector

machines and evolutionary information. *Bioinformatics, 22*(22), 2729-2734. doi:10.1093/bioinformatics/btl423

Coovadia, A. (2017). Lost in Interpretation: Evidence of Sequence Variant Database Errors. *J Assoc Genet Technol, 43*(1), 23-28.

Drum, C. L., Yan, S. Z., Bard, J., Shen, Y. Q., Lu, D., Soelaiman, S., . . . Tang, W. J. (2002). Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin. *Nature, 415*(6870), 396-402. doi:10.1038/415396a

Fallon, J. L., Halling, D. B., Hamilton, S. L., & Quiocho, F. A. (2005). Structure of calmodulin bound to the hydrophobic IQ domain of the cardiac Ca(v)1.2 calcium channel. *Structure, 13*(12), 1881-1886. doi:10.1016/j.str.2005.09.021

Geiser, J. R., van Tuinen, D., Brockerhoff, S. E., Neff, M. M., & Davis, T. N. (1991). Can calmodulin function without binding calcium? *Cell, 65*(6), 949-959.

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., . . . Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat, 36*(5), 513-523. doi:10.1002/humu.22768

Hoeflich, K. P., & Ikura, M. (2002). Calmodulin in action: diversity in target recognition and activation mechanisms. *Cell, 108*(6), 739-742.

Jensen, H. H., Brohus, M., Nyegaard, M., & Overgaard, M. T. (2018). Human Calmodulin Mutations. *11*(396). doi:10.3389/fnmol.2018.00396

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res, 24*(12), 2050-2058. doi:10.1101/gr.176214.114

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simao, F. A., Pozdnyakov, I. A., . . . Zdobnov, E. M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res, 43*(Database issue), D250-256. doi:10.1093/nar/gku1220

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., . . . Maglott, D. R. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res, 46*(D1), D1062-D1067. doi:10.1093/nar/gkx1153

Meador, W. E., Means, A. R., & Quiocho, F. A. (1992). Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. *Science, 257*(5074), 1251-1255.

Nyegaard, M., Overgaard, M. T., Sondergaard, M. T., Vranas, M., Behr, E. R., Hildebrandt, L. L., . . . Borglum, A. D. (2012). Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *Am J Hum Genet, 91*(4), 703-712. doi:10.1016/j.ajhg.2012.08.015

Parry, R. V., & June, C. H. (2003). Calcium-independent calcineurin regulation. *Nat Immunol, 4*(9), 821-823. doi:10.1038/ni0903-821

Pei, J., & Grishin, N. V. (2014). PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol, 1079*, 263-271. doi:10.1007/978-1-62703-646-7_17

Savojardo, C., Fariselli, P., Martelli, P. L., & Casadio, R. (2016). INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics, 32*(16), 2542-2544. doi:10.1093/bioinformatics/btw192

Shen, Y., Zhukovskaya, N. L., Guo, Q., Florian, J., & Tang, W. J. (2005). Calcium-independent calmodulin binding and two-metal-ion catalytic mechanism of anthrax edema factor. *EMBO J, 24*(5), 929-941. doi:10.1038/sj.emboj.7600574

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311.

Sorensen, A. B., Sondergaard, M. T., & Overgaard, M. T. (2013). Calmodulin in a heartbeat. *FEBS J, 280*(21), 5511-5532. doi:10.1111/febs.12337

Stevens, F. C. (1983). Calmodulin: an introduction. *Can J Biochem Cell Biol, 61*(8), 906-910.

Stull, J. T. (2001). Ca2+-dependent cell signaling through calmodulin-activated protein phosphatase and protein kinases minireview series. *J Biol Chem, 276*(4), 2311-2312. doi:10.1074/jbc.R000030200

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., . . . Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res, 13*(9), 2129-2141. doi:10.1101/gr.772403

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., . . . Roth, F. P. (2017). A framework for exhaustively mapping functional missense variants. *Mol Syst Biol, 13*(12), 957. doi:10.15252/msb.20177908

Yu, C. C., Ko, J. S., Ai, T., Tsai, W. C., Chen, Z., Rubart, M., . . . Chen, P. S. (2016). Arrhythmogenic calmodulin mutations impede activation of small-conductance calcium-activated potassium current. *Heart Rhythm, 13*(8), 1716-1723. doi:10.1016/j.hrthm.2016.05.009

Zhang, J., Kinch, L. N., Cong, Q., Weile, J., Sun, S., Cote, A. G., . . . Grishin, N. V. (2017). Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. *Hum Mutat, 38*(9), 1051-1063. doi:10.1002/humu.23293

Table 1 Summary of Measurements in Assessments.

Classification	
Area Under ROC	$P(X_1 - X_0)$ X_1, predicted score for positive instance; X_0, predicted score for negative instance. The area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one
MCC	$(TP_i \times TN_i - FP_i \times FN_i)/\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}$, $i \in$ (deleterious, intermediate, benign); TP: true positive; TN: true negative; FP: false positive; FN: false negative.
F1	$(2 \cdot precision \cdot recall)/(precision + recall)$, $precision = TP/(TP + FP); recall = TP/(TP + FN)$ TP: true positive; TN: true negative; FP: false positive; FN: false negative.
Ordinal association	
Kendall's tau-b rank correlation	$(n_c - n_d)/\sqrt{(n_0 - n_1)(n_0 - n_2)}$, $n_0 = n(n-1)/2$; $n_1 = \sum_k t_k(t_k - 1)/2$; $n_2 = \sum_j u_j(u_j - 1)/2$; n_c, the number of concordant pairs; n_d, the number of discordant pairs; n, the total number of pairs; t_k, number of values in the k^{th} group of ties by predictions; u_j, number of values in the j^{th} group of ties by experimental scores.
Spearman's rank correlation	$cov(R_{pred}, R_{exp})/\sigma_{R_{pred}}\sigma_{R_{exp}}$ $cov(R_{pred}, R_{exp})$, covariance between predicted and experimental ranks of mutants; $\sigma_{R_{pred}}$ and $\sigma_{R_{exp}}$, standard deviations of predicted and experimental ranks, respectively. Ties were randomly assigned distinct ranks first and then the average of these ranks were assigned to each of them.
Numeric comparison	
Pearson's correlation	$cov(pred, exp)/\sigma_{pred}\sigma_{exp}$ $cov(pred, exp)$, covariance between predictions and experimental scores; σ_{pred}, standard deviation of predictions; σ_{exp}, standard deviation of experimental scores

RMSD	$\sqrt{\frac{1}{N}\sum_{j=1}^{N}(pred_j - exp_j)}$ *N*, the size of a dataset; $pred_j$, j^{th} predictions; exp_j, j^{th} experimental scores
Value agreement test	$\sum Ci$ is the percentage of mutants with the difference between the predicted and experimental growth scores below a certain cutoff *i*. The cutoffs are taken from 0 to 1 with an incremental of 0.01. The area under curve was used as measurement.

group id	rank based scores				original value based						rescaled value based					
	Tau	spearman	del e roc	wil d roc	rms d	pears on	val ue diff	mcc_d ele	mcc_w ild	f1	rms d	pears on	val ue diff	mcc_d ele	mcc_w ild	f1
baseline	0.15	0.23	0.63	0.61	0.37	0.24	0.73	0.16	0.17	0.5	0.37	0.24	0.73	0.16	0.17	0.5
positive	0.85	0.98	0.99	0.98	0.07	0.97	0.95	0.85	0.92	0.92	0.07	0.97	0.96	0.85	0.92	0.92
1-1	0.15	0.22	0.63	0.61	0.38	0.22	0.73	0.11	0.16	0.48	0.38	0.22	0.73	0.11	0.16	0.48
1-2	0.15	0.22	0.63	0.6	0.38	0.23	0.73	0.11	0.16	0.48	0.38	0.23	0.73	0.11	0.16	0.48

2-1	0.17	0.25	0.65	0.61	0.4	0.24	0.66	0.13	0.1	0.29	0.38	0.24	0.73	0.11	0.17	0.48
2-2	0.15	0.23	0.64	0.6	0.37	0.22	0.69	0.1	0.09	0.34	0.38	0.22	0.72	0.09	0.17	0.47
3-1	0.07	0.108	0.58	0.57	0.35	0.17	0.75	0.09	0.1	0.51	0.4	0.17	0.71	0.14	0.08	0.46
4-1	-0.02	-0.03	0.45	0.48	0.51	-0.04	0.61	-0.05	-0.02	0.4	0.43	-0.04	0.67	-0.05	-0.03	0.4
4-2	-0.01	-0.02	0.49	0.49	0.5	-0.03	0.61	-0.03	-0.01	0.34	0.43	-0.03	0.67	-0.05	-0.03	0.4

Table 2 Scores for assessment of predictions. Tau, Kendall's tau coefficient (Tau-b); spearman, Spearman's rank correlation coefficient; dele roc, area under ROC for detecting deleterious variants; wild roc, area under ROC for detecting benign variants; rmsd, root-mean-square deviation; pearson, Pearson correlation coefficient; value diff, area under curve of percentage of variants against the absolute difference between experimental score and predicted score of variants; mcc_dele, matthew correlation coefficient for deleterious variants; mcc_wild, matthew correlation coefficient for benign variants; f1, F-score for three classes (deleterious, mildly deleterious, benign) of variants. baseline: baseline predictor; positive: positive control.