

Cobey JH Donelson^{1,2}, Héctor Martín Merinero¹, Seth J Welsh¹, Santiago Rodríguez de Cordoba³, Richard JH Smith^{1,2}, Yuzhou Zhang¹

¹Molecular Otolaryngology and Renal Research Laboratories, University of Iowa, Iowa City, USA

²Molecular Physiology and Biophysics Graduate Program, University of Iowa, Iowa City, USA

³Centro de Investigaciones Biológicas Margarita Salas, Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

Introduction

Complement Factor H (FH) is a soluble glycoprotein composed of 20 short consensus repeats (SCRs), each approximately 60 amino acids in length. These SCRs are organized into three functional domains: the N-terminal region (SCRs 1–4), which regulates complement activation; the C-terminal region (SCRs 19–20), which mediates recognition of host cell surfaces; and the mid-region (SCRs 5–18), which remains poorly characterized and contains numerous variants of uncertain significance (VUS).

There remains a lack of functional evidence for the mid-region of FH (SCRs 5–18), which contributes to the vast number of VUSs identified. Here, we integrate CADD, REVEL, and structure-based $\Delta\Delta G$ (DDGun3D) to assess variant impact and guide reclassification, with molecular dynamic (MD) simulations for validation. This work identifies likely benign and likely pathogenic variants in the mid-region of FH with emphasis on those driving disease.

Methods

1. Identify missense variants with a minor allele frequency (MAF) <0.1% in SCRs 5-18 of *CFH* in the genome aggregation database (gnomAD).
2. Classify variants using ACMG criteria as benign/likely benign (B/LB), pathogenic/likely pathogenic (P/LP) or VUS.
3. Calculate CADD and REVEL scores for each variant.
4. Calculate the $\Delta\Delta G_{\text{Fold}}$ for each variant using DDGun3D.
5. Use thresholds defined by REVEL, CADD, and $\Delta\Delta G_{\text{Fold}}$ values to reclassify VUSs.
6. Validate results with molecular dynamic (MD) simulations.

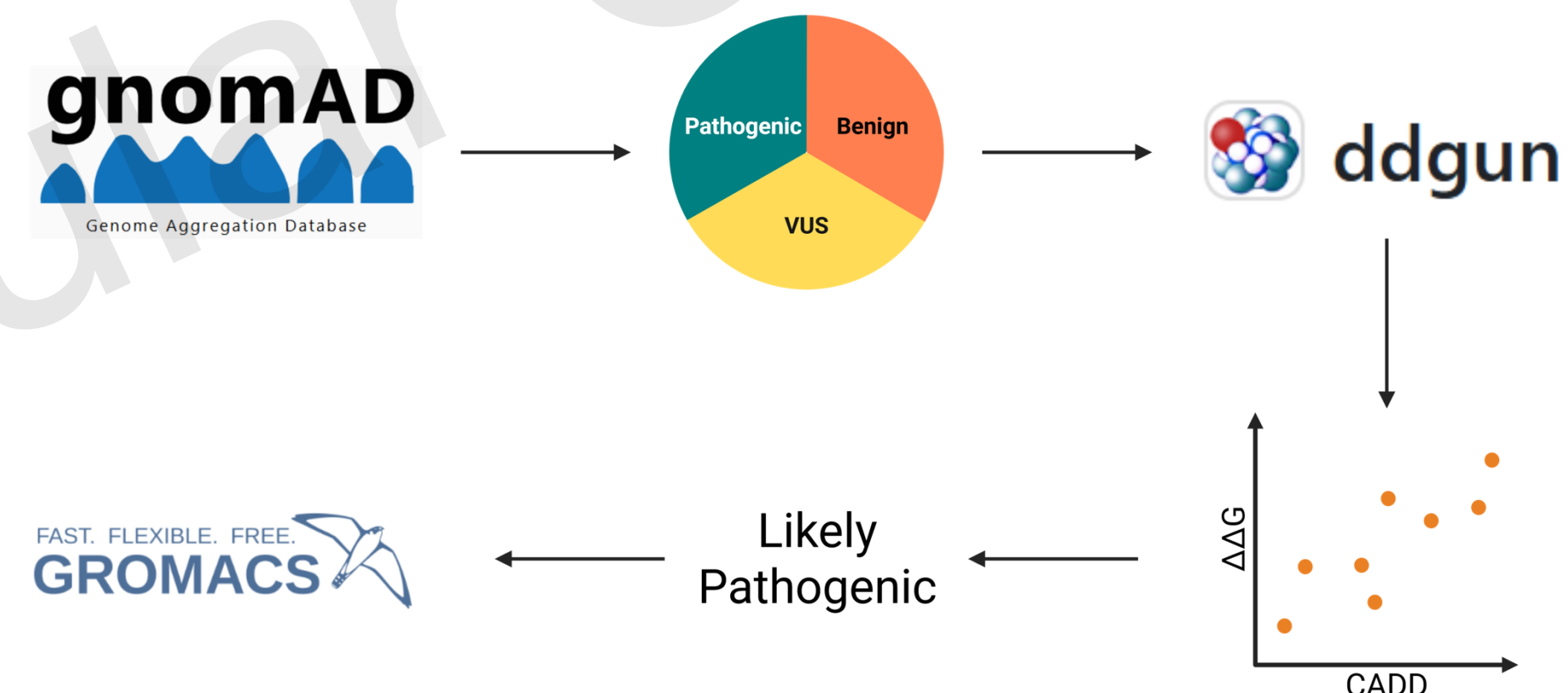


Figure 1. Experimental design. Variants in the *CFH* mid-region (SCRs 5–18) in gnomAD were grouped based on their ACMG classification and MAF. Folding free energy was calculated with DDGun3D. The folding free energy vs CADD and REVEL scores were plotted for each variant. MD simulations were run on predicted P/LP variants.

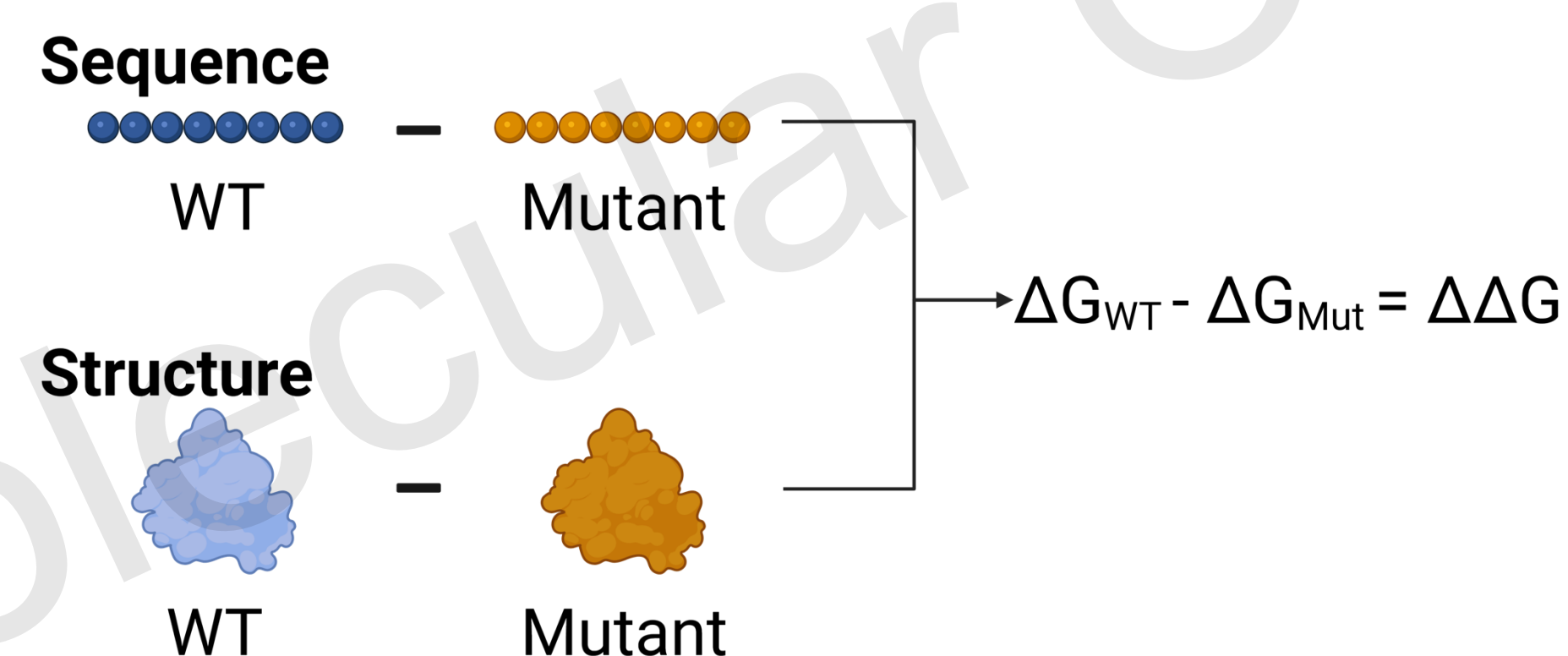


Figure 2. DDGun3D. DDGun3D predicts the folding free energy change ($\Delta\Delta G_{\text{Fold}}$) associated with missense variants. DDGun3D scores are based on sequence information and the protein's 3D structure. Sequence-based inputs include BLOSUM62 substitution scores, local interaction energies, and hydrophobicity differences, while structural features account for interaction energies within the 3D structure and changes in solvent-accessible surface area.

Filtering Cohort

We identified 1,026 variants in the mid-region of FH with a MAF < 0.1% in gnomAD. Of these, 1,011 are classified as VUS, 13 as B/LB, and only 2 as P/LP. Previous studies have proposed CADD score thresholds of <19 for B/LB and ≥ 19 for P/LP specifically in the mid-region of FH (1). Additional reports have suggested REVEL scores ≤ 0.29 and ≥ 0.64 to support B/LB and P/LP classifications, respectively (2). From our training set of 35 B/LB and 15 P/LP variants, we established DDGun3D thresholds of ≤ 0.4 for B/LB and ≥ 0.6 for P/LP.

To sum, we applied the following cutoffs:

- B/LB: CADD<19; REVEL ≤ 0.29 ; $\Delta\Delta G_{\text{Fold}}\leq 0.4$
- P/LP: CADD ≥ 19 ; REVEL ≥ 0.64 ; $\Delta\Delta G_{\text{Fold}}\geq 0.6$

These metrics yielded a positive predictive value (PPV) of 60% and a negative predictive value (NPV) of 60%.

Predicted P/LP variants

We identified 48 P/LP variants: 36 are associated with the loss of Cys; the remaining 12 are shown in Table 1.

Table 1. Non-Disulfide Variants Predicted to be P/LP

SCR	Protein	$\Delta\Delta G_{\text{fold}}$	CADD	REVEL
8	W499R	1.3	25.6	0.823
8	K507I	0.6	24.7	0.662
9	Y534C	0.7	23.6	0.651
11	Y657H	1.3	26.0	0.686
11	G667E	0.6	25.5	0.750
11	P682S	0.7	24.8	0.667
12	P742S	0.6	25.2	0.713
12	W738R	2.0	25.8	0.810
14	W858R	2.4	24.7	0.815
16	W978G	3.1	26.2	0.795
17	W1037R	2.1	24.6	0.807
17	W1037C	1.9	25.3	0.773

Residue Hotspots

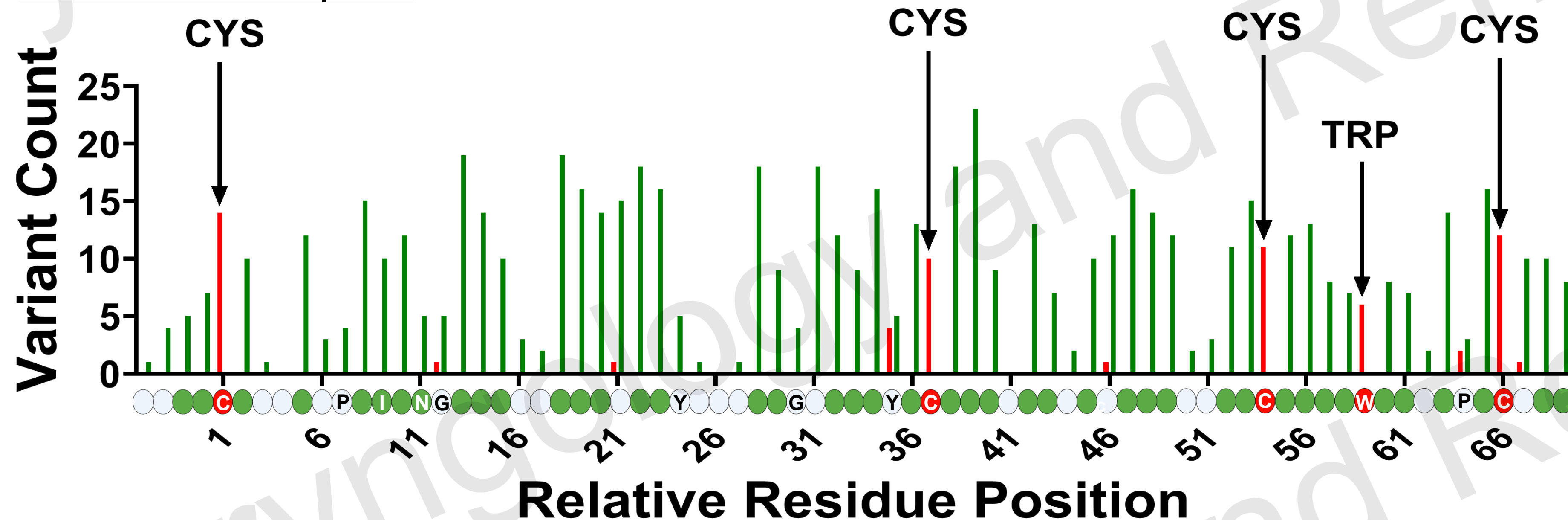


Figure 5. Predicted Classification of Standardized Residues in SCRs 5–18. The sequences from SCRs 5–18 were overlaid to identify conserved residues. The conserved residues were used as “landmarks” to map the P/LP and B/LB variants onto a standardized SCR. The linear, standardized FH protein is displayed just below the x-axis. Variants previously classified as P/LP, those newly classified as P/LP were grouped together (red), as were previously and newly classified B/LB variants (green). VUSs are grey. Highly conserved Cys and Trp residues are highlighted in the bar graph.

Results

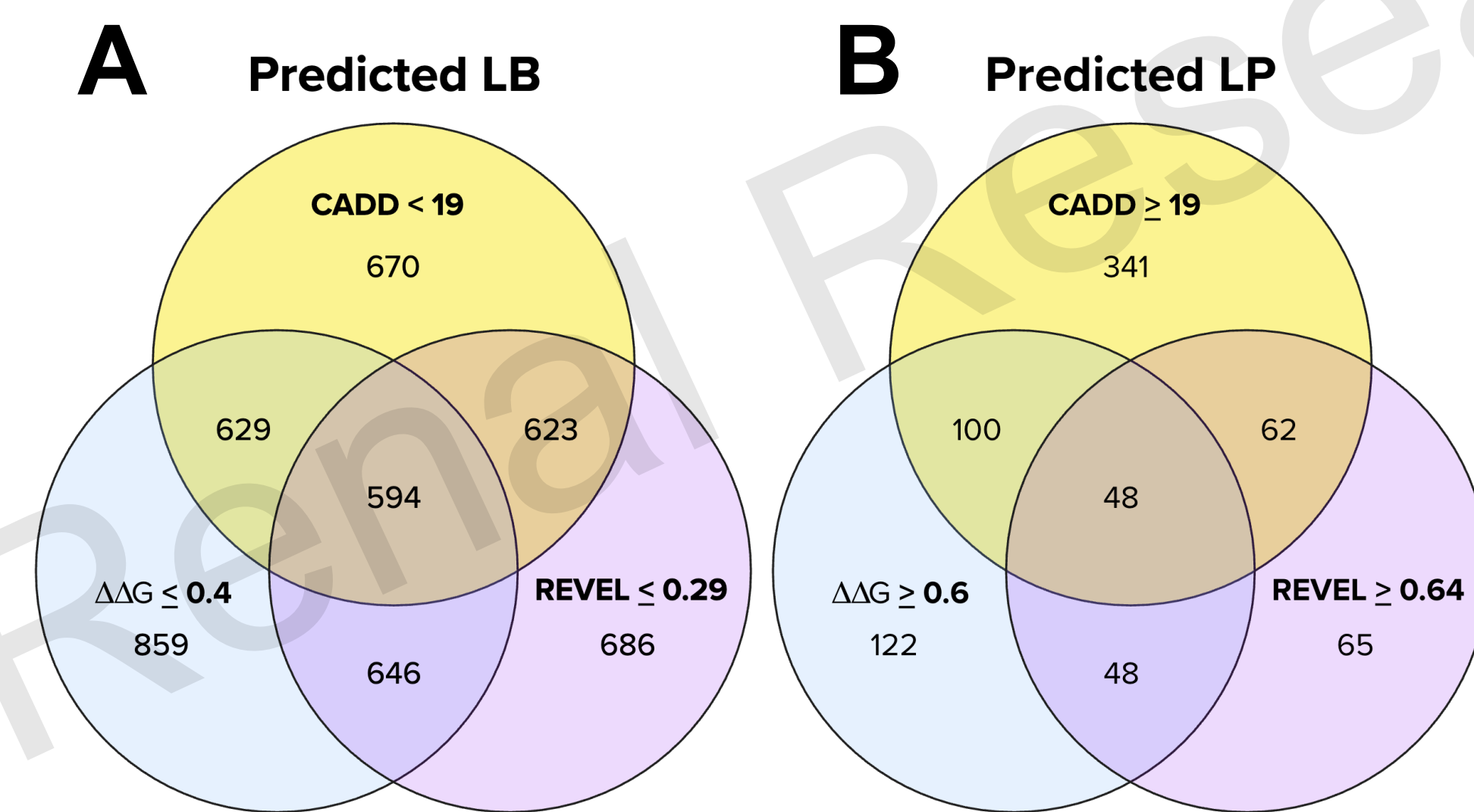


Figure 3. Workflow for Prioritizing Variants. Panel A illustrates the filtering process for variants with evidence supporting a B/LB reclassification. 594 of 1,011 VUSs (59%) are now predicted B/LB. Panel B shows the filtering of variants with evidence supporting a P/LP reclassification. 48 of 1,011 VUSs (5%) are now predicted P/LP.

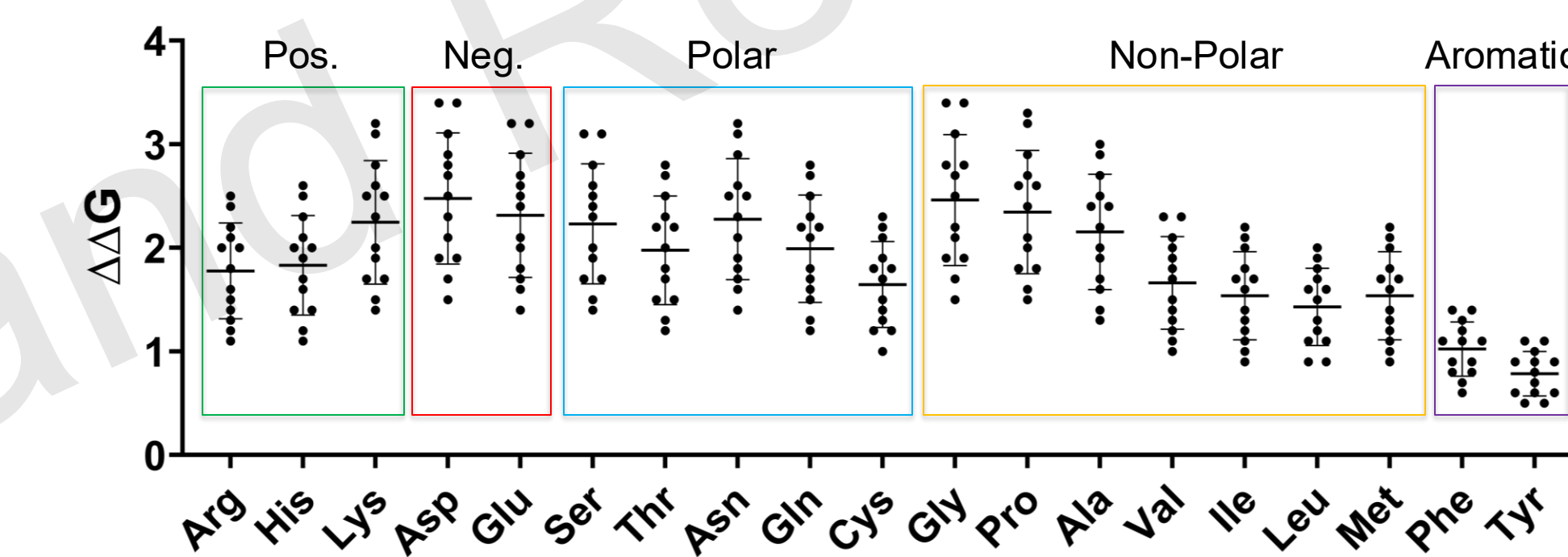


Figure 4. Tryptophan Substitutions. $\Delta\Delta G_{\text{Fold}}$ values were calculated at each tryptophan site by substituting all 19 other amino acids. Aromatic substitutions exhibited the most benign effects on protein stability.

MD Validation/Future Directions

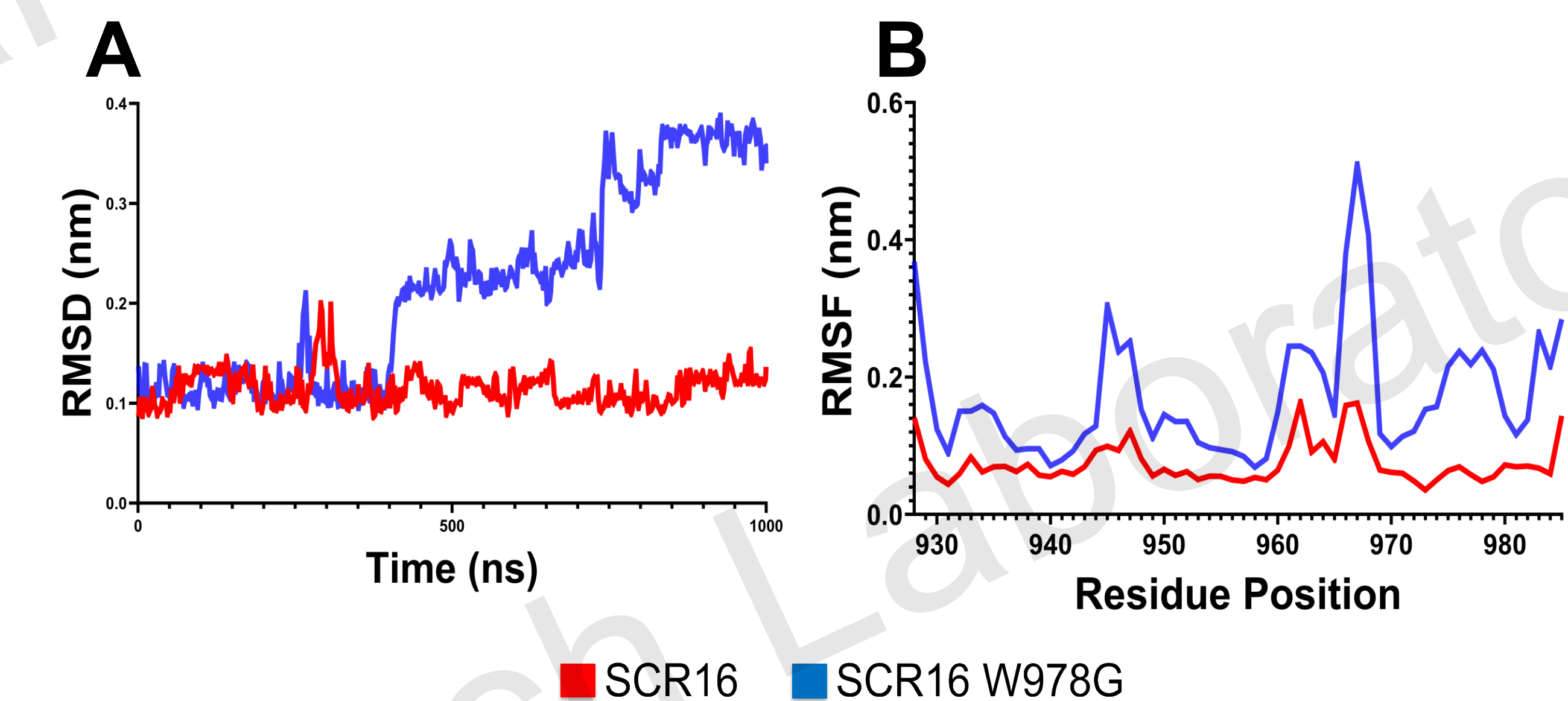


Figure 6. Validation with Molecular Dynamics Simulations. Examples of SCR16 wild-type (red) and W978G (blue) were simulated in explicit solvent using the AMBER force field for 1 microsecond. Panel A shows the Root Mean Square Deviation (RMSD) over time for WT and W978G. Panel B shows the Root Mean Square Fluctuation (RMSF) for WT and W978G throughout the simulation.

Conclusions

From gnomAD, we identified 1,011 rare (MAF < 0.1%) missense VUSs in the FH protein and provided evidence supporting the reclassification of 594 (59%) as B/LB, and 48 (5%) as P/LP. These P/LP variants consistently disrupt disulfide bonds or alter conserved tryptophan residues. We propose two hypotheses to explain why the loss of conserved cysteine residues results in P/LP classification: (1) the loss of a disulfide bond may lead to unfolding of the SCR domain and subsequent degradation of FH prior to secretion; (2) disruption of disulfide bonds generates a free cysteine residue which—in FH's oxidizing extracellular environment—is likely to form aberrant pairings that distort the protein's structure. Regarding tryptophan substitutions, we found that conserved tryptophans contribute to hydrogen bonding within SCR domains, and we hypothesize that their substitution destabilizes domain architecture, promoting unfolding and degradation. Notably, we identified an individual with FH deficiency carrying the W978L variant (3), which has been classified as P/LP; this tryptophan substitution supports the predictive accuracy of our variant classification algorithm. In summary, we present an alternative approach to characterize variants in the mid-region of FH, with the ultimate goal of improving the identification of P/LP variants.

Acknowledgements

National Institutes of Health R01 DK110023
Molecular Otolaryngology and Renal Research Laboratories
Images were created using BioRender.com

References

1. Martín Merinero H, Zhang Y, Arjona E, et al. Functional characterization of 105 factor H variants associated with aHUS: lessons for variant classification. *Blood*. 2021;138(22):2185-2201. doi:10.1182/blood.2021012037
2. Pejaver V, Byrne AB, Feng BJ, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am J Hum Genet*. 2022;109(12):2163-2177. doi:10.1016/j.ajhg.2022.10.013
3. Tortajada A, Gutiérrez E, Goicoechea de Jorge E, et al. Elevated factor H-related protein 1 and factor H pathogenic variants decrease complement regulation in IgA nephropathy. *Kidney Int*. 2017;92(4):953-963. doi:10.1016/j.kint.2017.03.041