Glycine missense variants in the COL3A1 triple helix domain: assessing functional domain data during clinical variant interpretation

Daniel Beltrán, Janita Thusberg, Michael Anderson, Yuya Kobayashi, John Garcia, Tom Winder, Scott Topper, Keith Nykamp

Invitae Corporation, San Francisco, CA

Disclosure statement: All authors are employees and stockholders of Invitae Corporation.

Introduction

Amino acid residues that are critical for proper protein structure or function are intolerant of variation, and this understanding is key in the accurate clinical variant classification of novel variants. However, clear guidelines for objectively identifying a residue or region as “critical” do not exist, and overestimation of the importance of such regions is a common source of erroneous and over-confident clinical classifications. Therefore, we developed an approach to critical residues classification that takes into account background variation, the distribution of clinically observed variants, and knowledge of the three-dimensional structure of proteins. We intend to incorporate this new category of evidence into Sherloc, our evidence-based system for variant interpretation system. We selected *COL3A1* as the paradigm for protein domain evaluation. *COL3A1* is a well-characterized gene associated with vascular Ehlers-Danlos syndrome (vEDS). Most of the protein comprises a triple-helix (TH) domain encoded by 343 Gly-X-Y (GXY) repetitions, and Gly residues therein are important for protein structure and macromolecular assembly. Missense genetic variation at these residues is significantly lower than background variation affecting other residues and regions of COL3A1, and vEDS cohorts show an enrichment for variants at TH Gly residues. Together, these observations provide a method for identifying critical residues, which we applied to the other collagen genes as well. Notably, by this measure, variants affecting many Glycine residues in COL5A1 should NOT be classified as pathogenic without additional supporting information.

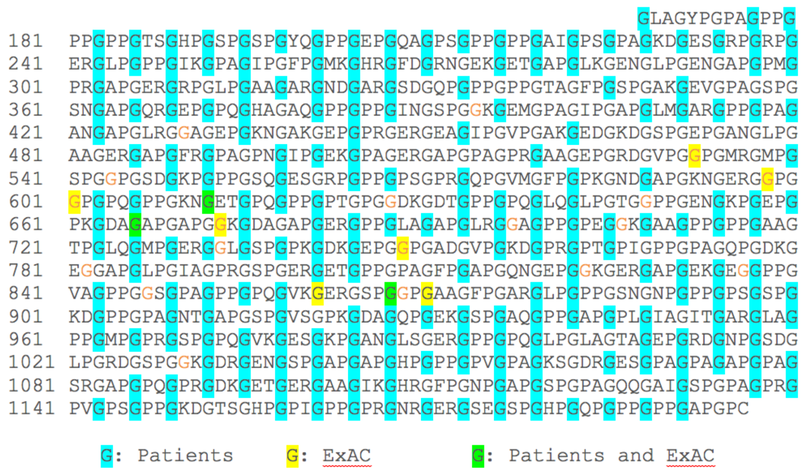
Methods

We identified all reported missense variants in the COL3A1 protein, and analyzed their location and frequency in both patients and the general population. Patients were collated from the Ehlers Danlos Syndrome Variant Database (EDSdb, https://eds.gene.le.ac.uk) and the ExAC database was used for the general population. We first calculated the ratio of patients with any COL3A1 missense variant to the number of individuals (“controls”) in ExAC with any COL3A1 missense variant (i.e. Whole gene: patients/controls). Variants with a frequency greater than 0.1% in ExAC were not used for this analysis. We then calculated the patient/control ratio for any missense in the Triple Helix (TH) domain and specifically for Glycine missense variants (Gly*-X-Y) within the TH domain. Finally, enrichment of patients with vEDS and a TH Glycine missense variant was calculated as a ratio of (patients/controls with TH Gly missense) over (patients/controls with any *COL3A1* missense) and denoted as “key residue ratio”. A similar analysis was performed for Gly TH variants in *COL5A1*, *COL10A1* and Cys missense variants in *FBN1*. For the determination of the Gly-X-Y motif, we used data from *COL1A1*, *COL1A2*, *COL2A1*, *COL3A1*, *COL4A1*, *COL6A1*, *COL6A2*, *COL6A3* and *COL7A1*.

Results

Glycines are distributed throughout the Triple Helix as a Gly-X-Y motif*

COL3A1 Triple Helix domain (p.Gly168-p.Cys1196)


GLAGYPGPAGPPG
181 PPGPPGTSGHPGSPGSPGYQGPPGEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPG
241 ERGLPGPPGIKGPAGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMG
301 PRGAPGERGRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGSPG
361 SNGAPGQRGEPGPQGHAGAQGPPGPPGINGSPGGKGEMGPAGIPGAPGLMGARGPPGPAG
421 ANGAPGLRGGAGEPGKNGAKGEPGPRGERGEAGIPGVPGAKGEDGKDGSPGEPGANGLPG
481 AAGERGAPGFRGPAGPNGIPGEKGPAGERGAPGPAGPRGAAGEPGRDGVPGGPGMRGMPG
541 SPGGPGSDGKPGPPGSQGESGRPGPPGPSGPRGQPGVMGFPGPKGNDGAPGKNGERGGPG
601 GPGPQGPPGKNGETGPQGPPGPTGPGGDKGDTGPPGPQGLQGLPGTGGPPGENGKPGEPG
661 PKGDAGAPGAPGGKGDAGAPGERGPPGLAGAPGLRGGAGPPGPEGGKGAAGPPGPPGAAG
721 TPGLQGMPGERGGLGSPGPKGDKGEPGGPGADGVPGKDGPRGPTGPIGPPGPAGQPGDKG
781 EGGAPGLPGIAGPRGSPGERGETGPPGPAGFPGAPGQNGEPGGKGERGAPGEKGEGGPPG
841 VAGPPGGSGPAGPPGPQGVKGERGSPGGPGAAGFPGARGLPGPPGSNGNPGPPGPSGSPG
901 KDGPPGPAGNTGAPGSPGVSGPKGDAGQPGEKGSPGAQGPPGAPGPLGIAGITGARGLAG
961 PPGMPGPRGSPGPQGVKGESGKPGANGLSGERGPPGPQGLPGLAGTAGEPGRDGNPGSDG
1021 LPGRDGSPGGKGDRGENGSPGAPGAPGHPGPPGPVGPAGKSGDRGESGPAGPAGAPGPAG
1081 SRGAPGPQGPRGDKGETGERGAAGIKGHRGFPGNPGAPGSPGPAGQQGAIGSPGPAGPRG
1141 PVGPSGPPGKDGTSGHPGPIGPPGPRGNRGERGSEGSPGHPGQPGPPGPPGAPGPC

G: Patients G: ExAC G: Patients and ExAC

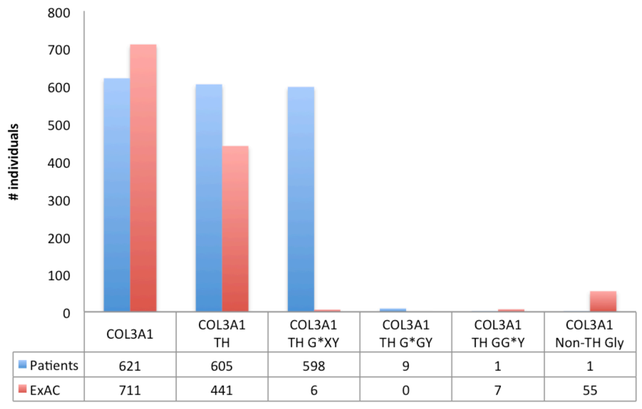
Motif	Variant	ExAC counts
GG*Y	p.Gly532Ser	1
GG*Y	p.Gly598Ala	1
GG*Y	p.Gly601Arg	1
G*XY	p.Gly612Ala	1
G*XY	p.Gly666Ser	2
GG*Y	p.Gly673Ser	1
GG*Y	p.Gly748Ser	3
G*XY	p.Gly861Ser	1
G*GY	p.Gly867Cys	1
G*XY	p.Gly870Val	1

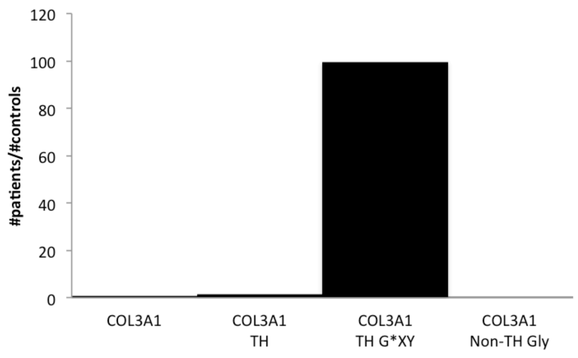
Patient enrichment is confined to the first position of the Gly-X-Y motif*

G X Y G X Y G X Y G X Y G X Y G X Y

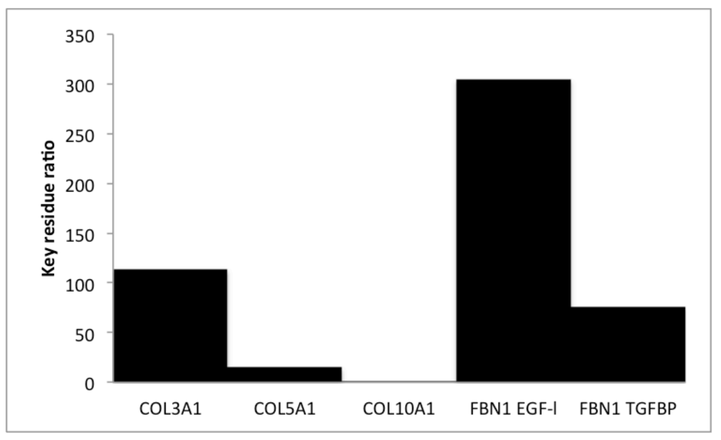
Motif	G*XY	GG*Y	GXG*	G*Y
Patients	3159	1	0	11
Controls	349	29	5	23

G G*XY	G*XY	G*GY	G*XG		G*XY GGY
14	3159	19	11		16
4	349	5	9		1





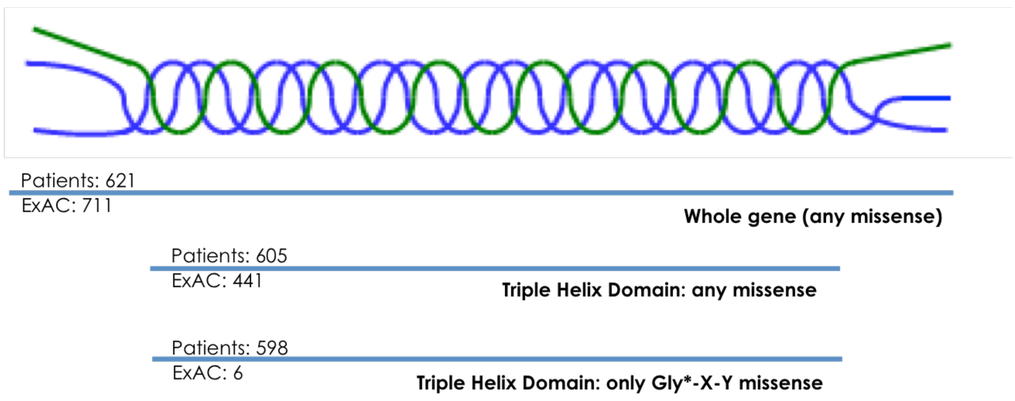
This analysis can be extended to other proteins and key residues

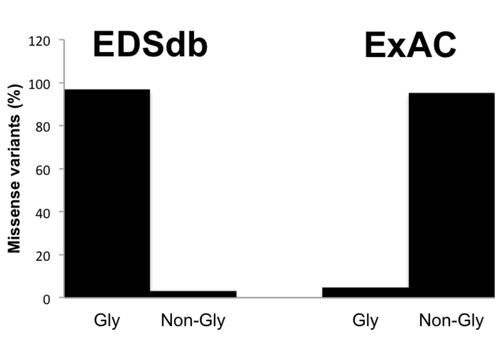


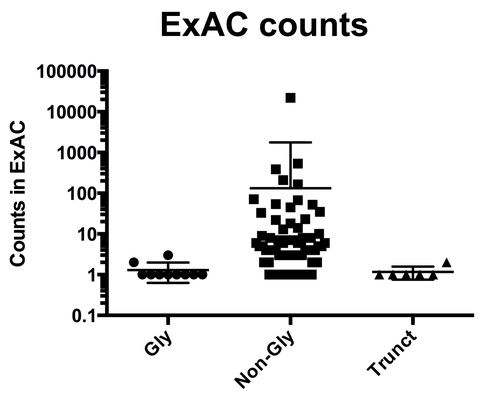
Gene	Patients	ExAC	>3 ExAC	>8 ExAC	Category
COL3A1	598	6	0	0	1
COL5A1	19	59	1	1	2
COL10A1	0	102	2	2	3
FBN1	805	3	0	0	1

Results

Glycine variants in the COL3A1 Triple Helix are enriched in vEDS patients







Conclusions

We used the COL3A1 Triple Helix (TH) domain as a paradigm for evaluating the clinical significance of key amino acid residues in essential protein domains. Missense variants involving the glycine residues of the TH are highly enriched in patients with vEDS. By comparing the frequency of patients and controls with missense variants in COL3A1 we have been able to establish the Gly*-X-Y motif as essential domain and the Gly* as a key residue in this domain. Importantly, a missense change of the glycine in this motif is very likely associated with disease, while a missense variant of the X-Y position is not associated with disease. We also developed an objective metric (#patients/controls with missense variants in a specific domain vs. # patients with missense variants throughout the entire protein) for comparing the likelihood that missense variants in key residues are associated with disease across paralogs and related proteins.