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

Daniel Beltran, Janita Thusberg, Michael Anderson, Yuya Kobayashi, John Garcia, Tom Winder, Matteo Vatta, Scott Topper, Keith Nykamp
Invitae Corporation, San Francisco, CA

## Introduction

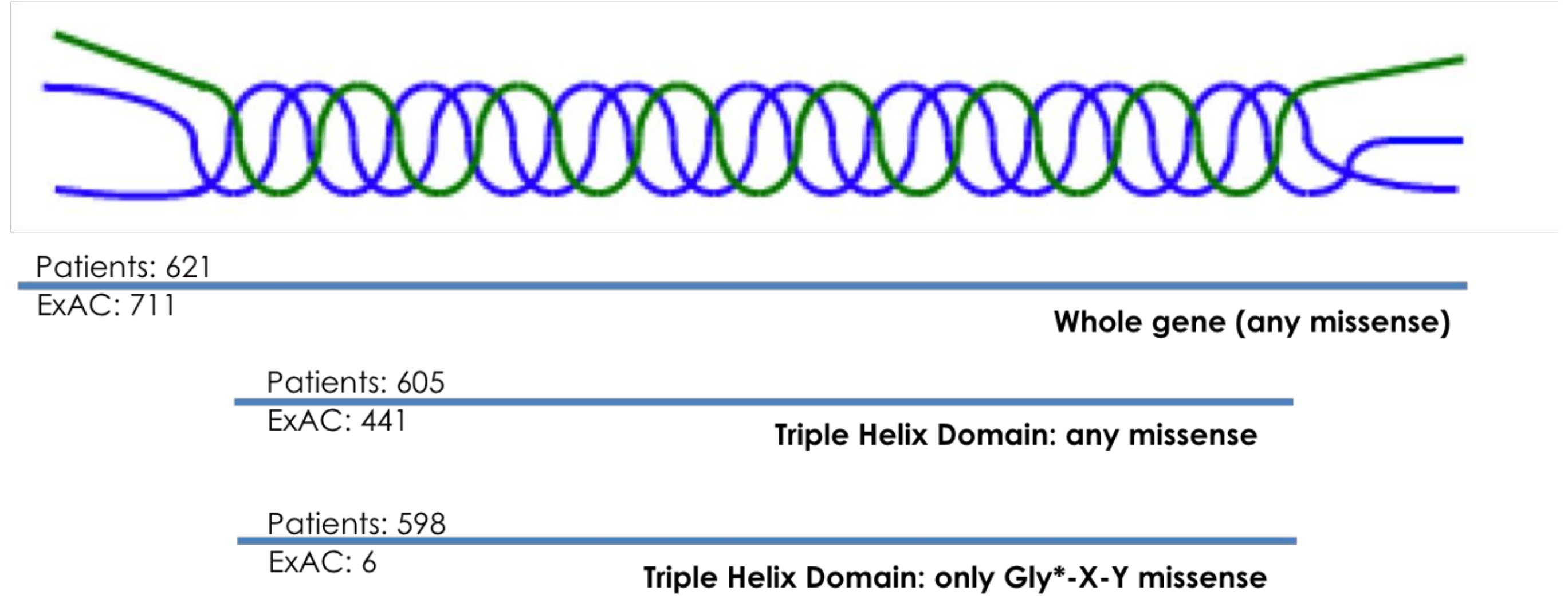
Amino acid residues that are critical for proper protein structure or function are intolerant of variation, and this understanding is key in the accurate clinical variant classification of novel variants. However, clear guidelines for objectively identifying a residue or region as "critical" do not exist. Therefore, we developed an approach to critical residue classification that takes into account background variation, the distribution of clinically observed variants, and knowledge of protein structure. We intend to incorporate this new category of evidence into Sherloc, our evidence-based system for variant interpretation. 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 COL3A1 protein comprises a triple-helix (TH) domain encoded by 343 Gly-X-Y (GXY) repetitions, and Gly residues therein are crucial to 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. We have applied this analysis to all the collagen genes associated with human disease and to missense variants in cysteine residues in known protein domains in fibrillin-1 (FBN1).

## Methods

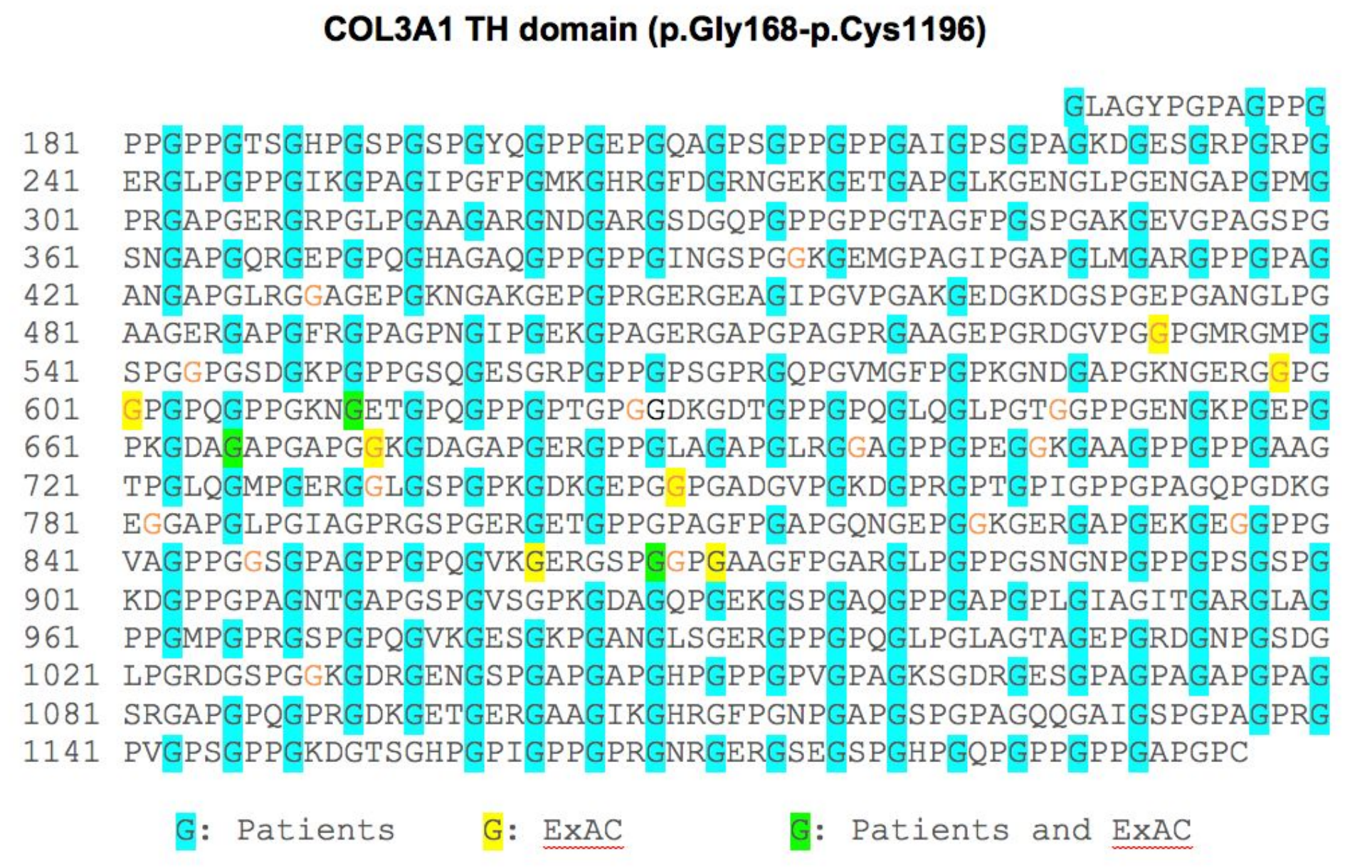
We identified all reported missense variants in the COL3A1 protein and analyzed their location and frequency in both patients and the general population. Patients were collated from the Ehlers Danlos Syndrome Variant Database (https://eds.gene.le.ac.uk), and the ExAC database was used for the general 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 excluded from this analysis. We then calculated the patient/control ratio both for any missense variants and for Gly missense variants (Gly*-X-Y) within the TH domain. Finally, enrichment of patients with vEDS and a TH Gly missense variant was calculated as a ratio of patients/controls with TH Gly missense variant over patients/controls with any *COL3A1* missense variant and denoted as the key residue ratio. Similar analyses were performed for TH Gly variants in other collagen genes associated with human disease and Cys missense variants in *FBN1*. For the determination of the GXY motif, we used data from *COL1A1*, *COL1A2*, *COL2A1*, *COL3A1*, *COL4A1*, *COL6A1*, *COL6A2*, *COL6A3*, and *COL7A1*.

## Results

*Glycine variants in the COL3A1 TH are enriched in vEDS patients.*

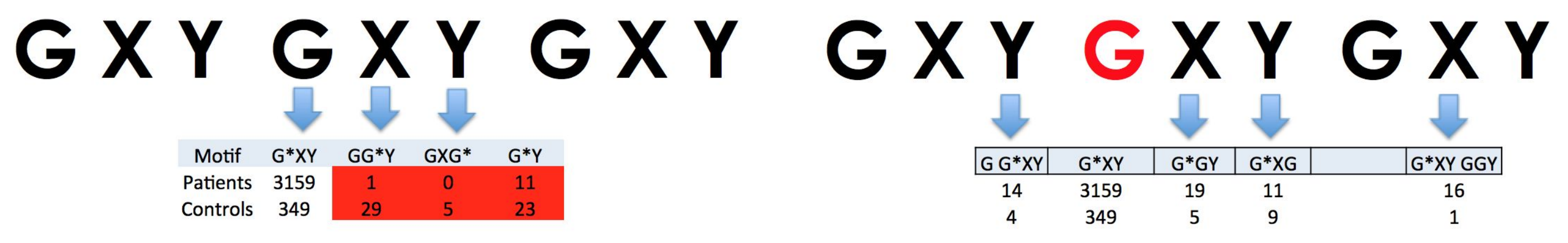

Patients: 621
ExAC: 711
Whole gene (any missense)
Patients: 605
ExAC: 441
Triple Helix Domain: any missense
Patients: 598
ExAC: 6
Triple Helix Domain: only Gly*-X-Y missense

*Gly variants are distributed throughout the TH in the Gly*-X-Y motif.*



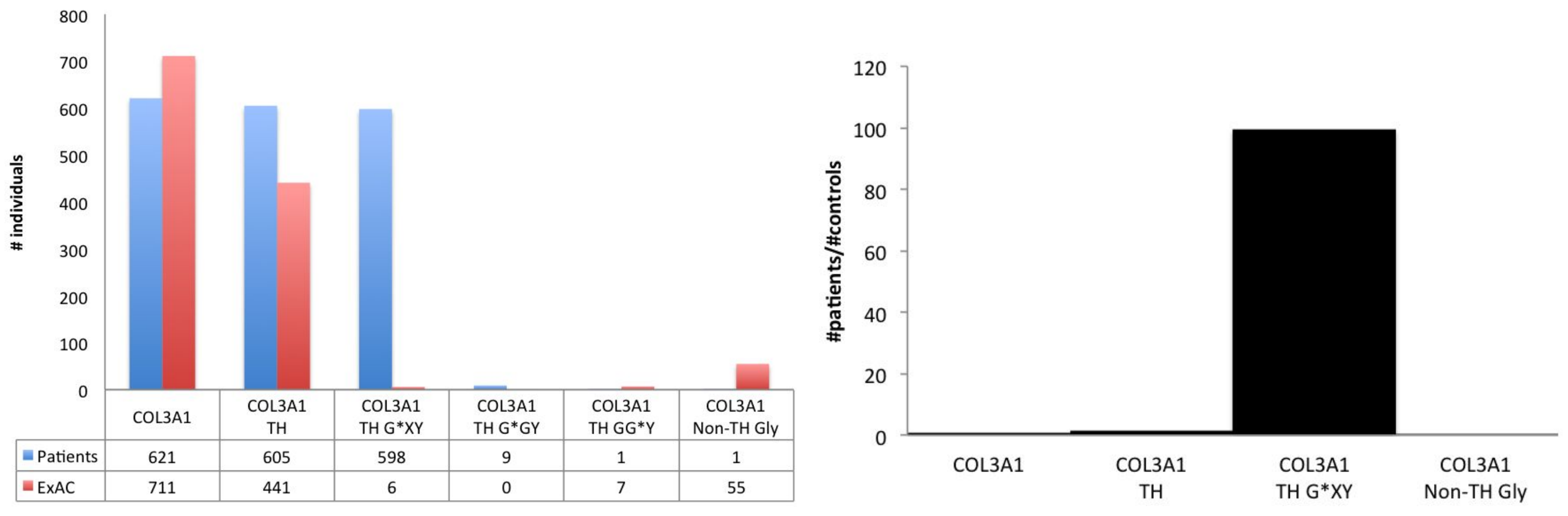
| Motif | Variant | ExAC counts |
|---|---|---|
| GG*Y | p.Gly532Ser | 1 |
| GG*Y | p.Gly598Ala | 1 |
| GG*Y | p.Gly601Arg | 1 |
| G*XY | p.Gly612Ala | 1 |
| G*XY | p.Gly666Ser | 2 |
| GG*Y | p.Gly673Ser | 1 |
| GG*Y | p.Gly748Ser | 3 |
| G*XY | p.Gly861Ser | 1 |
| G*GY | p.Gly867Cys | 1 |
| G*XY | p.Gly870Val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

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



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

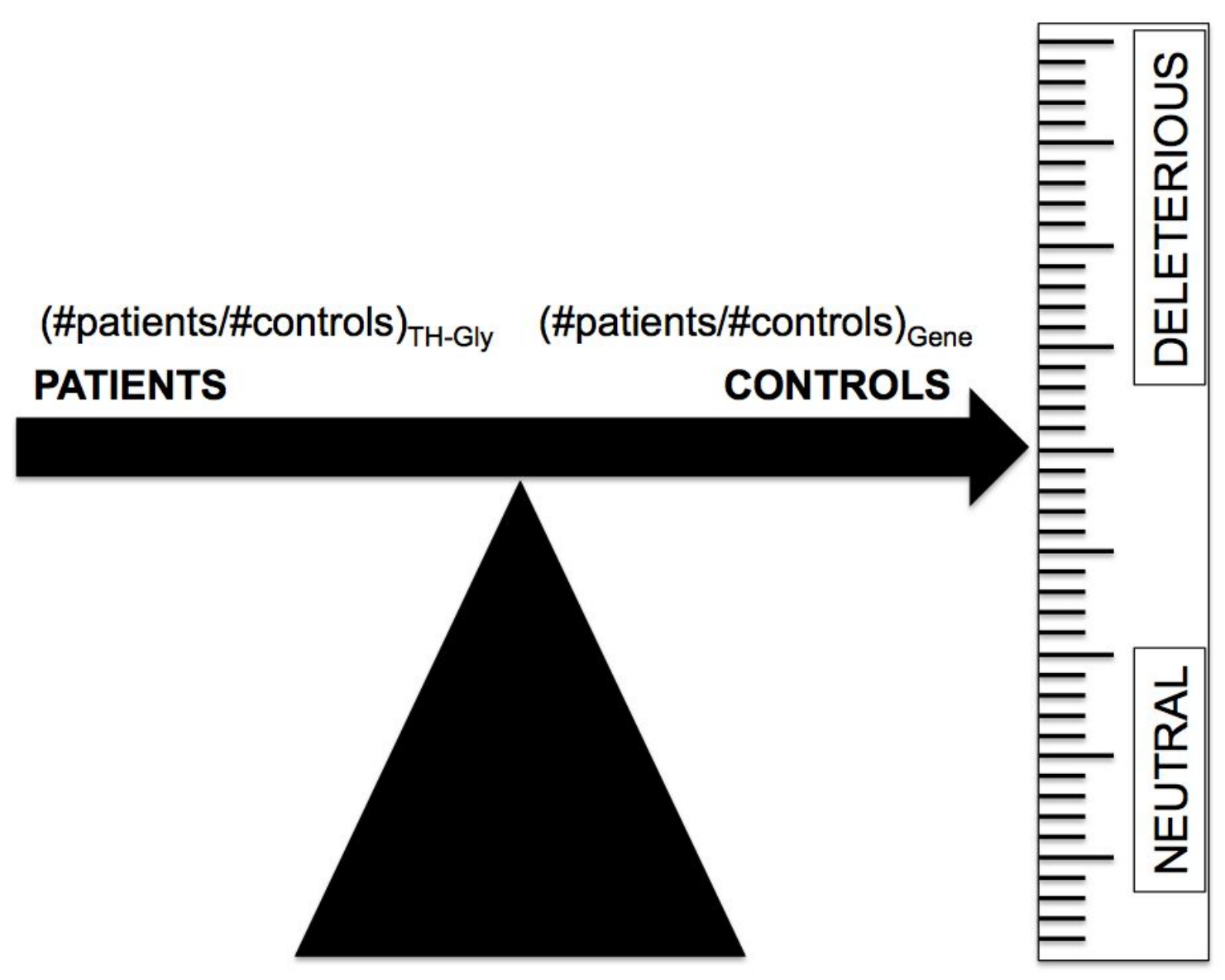
| G G*XY | G*XY | G*GY | G*XG | | G*XY GGY |
|---|---|---|---|---|---|
| 14 | 3159 | 19 | 11 | | 16 |
| 4 | 349 | 5 | 9 | | 1 |



| | COL3A1 | COL3A1 TH | COL3A1 TH G*XY | COL3A1 TH G*GY | COL3A1 TH GG*Y | COL3A1 Non-TH Gly |
|---|---|---|---|---|---|---|
| ■ Patients | 621 | 605 | 598 | 9 | 1 | 1 |
| ■ ExAC | 711 | 441 | 6 | 0 | 7 | 55 |

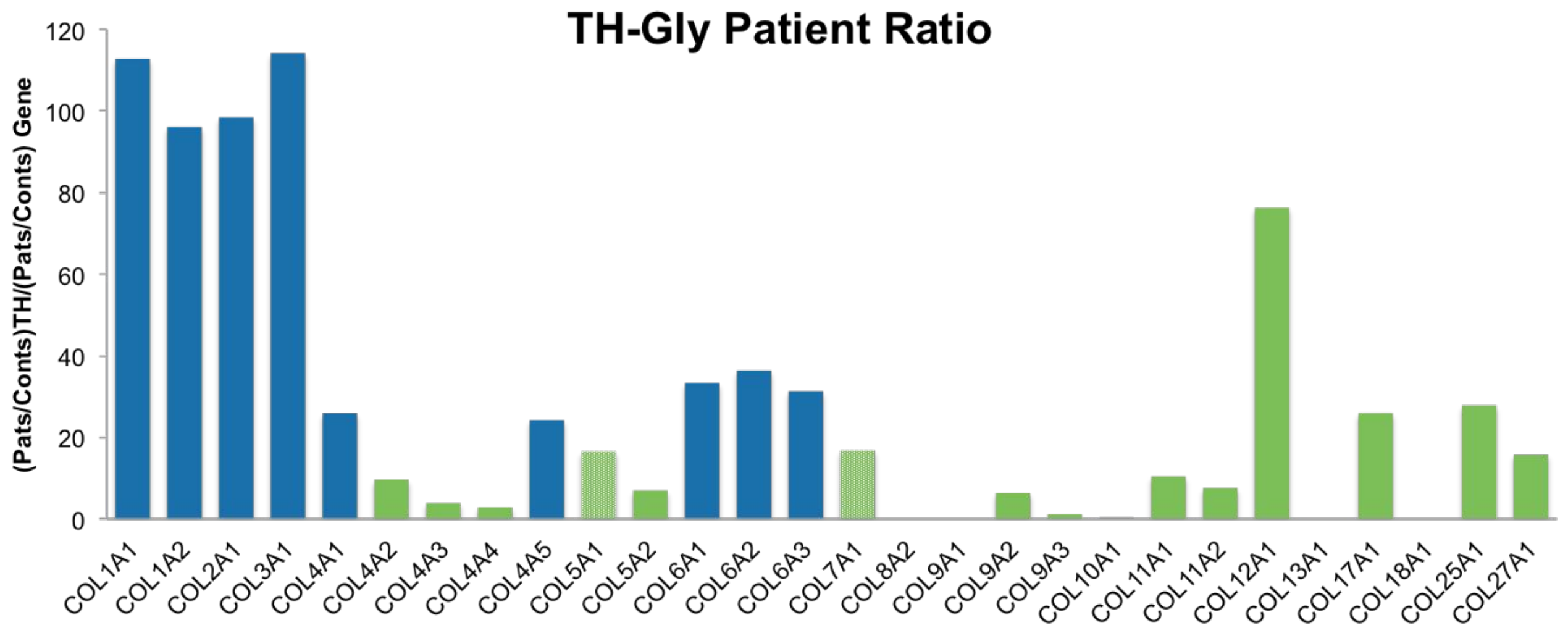
*Establishing a metric to estimate the likelihood of a COL3A1 TH Gly missense variant to be pathogenic and comparing it with other collagens*

| | ExAC EXCEPTIONS | | ALL GENE MISSENSE | | | ALL TH MISSENSE | | | TH G*XY | | | RATIO |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TH | 3-7 | >7 | Patients | Controls | Pats/Conts | Patients | ExAC | Pats/Conts | Patients | Controls | Pats/Conts | |
| p.Gly168-p.Cys1196 | 0 | 0 | 621 | 711 | 0.873 | 605 | 441 | 1.37 | 598 | 6 | 99.7 | 114.1 |


$(\#patients/\#controls)_{TH\text{-}Gly}$
$(\#patients/\#controls)_{Gene}$
PATIENTS
CONTROLS
DELETERIOUS
NEUTRAL

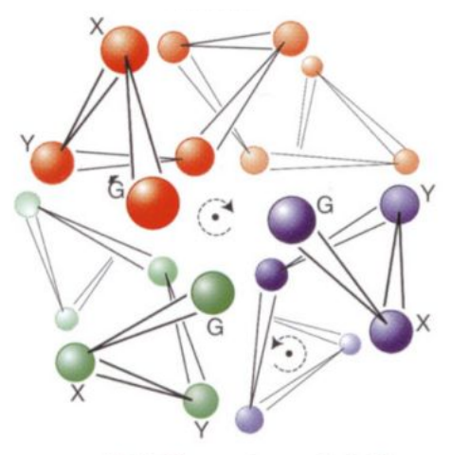
*This analysis can be extended to other collagens.*

| Protein | TH | ExAC EXCEPTIONS 3-7 | ExAC EXCEPTIONS >7 | ALL GENE MISSENSE Patients | ALL GENE MISSENSE Controls | ALL GENE MISSENSE Pats/Conts | TH G*XY Patients | TH G*XY Controls | TH G*XY Pats/Conts | RATIO |
|---|---|---|---|---|---|---|---|---|---|---|
| *COL1A1* | p.Gly179-p.Pro1192 | 0 | 0 | 892 | 913 | 0.978 | 771 | 7 | 110.1 | 112.7 |
| *COL3A1* | p.Gly168-p.Cys1196 | 0 | 0 | 621 | 711 | 0.873 | 598 | 6 | 99.7 | 114.1 |
| *COL5A2* | p.Gly203-p.Leu1229 | 2 | 2 | 14 | 1234 | 0.011 | 7 | 88 | 0.08 | 7.01 |
| *COL10A1* | p.Gly269-p.Ala756 | 2 | 2 | 22 | 696 | 0.032 | 1 | 119 | 0.008 | 0.27 |

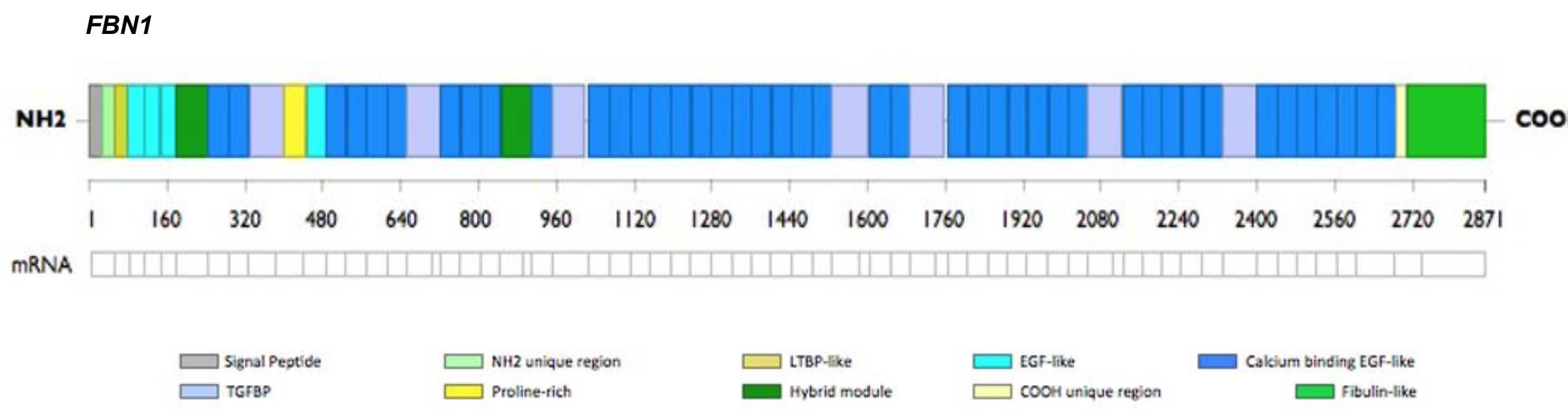


*Can this analysis be extended to infer severity?*

| | N-FLANKING Glycines | | | C-FLANKING Gs | | |
|---|---|---|---|---|---|---|
| | GXG G*XY | GGY G*XY | GY G*XY | G*GY | G*GG | G*XG |
| Patients | 14 | 25 | 1 | 17 | 4 | 11 |
| Controls | 4 | 2 | 1 | 3 | 0 | 4 |
| Pats/Conts | 3.5 | 12.5 | 1 | 5. 7 | NA | 2.75 |
| Ratio | 0.2 | 0.73 | 0.067 | 0.33 | NA | 0.16 |



*This analysis can be extended to other proteins and key residues.*



From Collod-Beroud *et al.*, ~~(2003)~~ Update of the UMD-FBN1 mutation database and creation of an FBN1 polymorphism database. ***Hum Mutat.* 2003; 22:199-208.**

| | | Cys forming | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Category 1 | All missense | FBN1 | Domain | Non-domain | EGF-L | TGFBP | Hybrid | 4-Cys | Cys disrupting |
| | 1773 | 243 | 241 | 2 | 140 | 69 | 14 | 18 | 805 |
| Patients | 1568 | 34 | 13 | 21 | 10 | 3 | 0 | 0 | 3 |
| Controls | 1.13 | 7.06 | 18.5 | 0.1 | 14 | 23 | NA | NA | 268 |
| Pats/Conts | 1 | 6.2 | 16.2 | 0.09 | NA | 20.4 | NA | NA | 237 |

## Conclusions

We used the *COL3A1* as a paradigm for evaluating the clinical significance of key amino acid residues in essential protein domains. Missense variants involving the Gly residues of the TH are highly enriched in patients with vEDS. By comparing the frequency of missense variants in *COL3A1* in patients and controls, we were able to establish the Gly*-X-Y motif as an essential domain and Gly* as a key residue in this domain. Importantly, a missense change of the Gly in this motif is very likely associated with disease, whereas 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.