Phylogenetic Investigation of the Phenotype-Genotype Relationship Between Pathogenic COL5A2 and COL1A1 Variants and Classical Ehlers-Danlos-Syndrome

Question: Do pathogenic variants of COL5A2 and COL101 scatter along the gene randomly, transcending conserved regions?

Introduction

Ehlers-Danlors syndrome is a set of heterogenous disorders characterized by connective tissue pathology (Defendi *et al*., 2015). As of 2017, current classification of the disorder comprises 13 subtypes with a high degree of overlapping between subtypes (Malfait *et al*., 2017). As of 2020, 21 genes are listed as a causal factor for Ehlers-Danlors syndrome (OrphaNet). Thus, Ehlers-Danlors syndrome might be classified as a complex disease rather than a Mendelian disorder, in which variants in a single gene are realized as a completely new phenotype. One subtype, classical Ehlers-Danlors syndrome is the simplest form, characterized by skin hyperextensibility, widened atrophic scars, and generalized joint hypermobility (Pauker *et al.,* 2016). Among causal genes, variants in COL1A1 and COL5A2 are thought to complicate connective tissue integrity, resulting in phenotypes like classical Ehlers-Danlors syndrome. Challenges to interpret genotype-phenotype interplay continue to arise despite the growing body of genomic data. Researchers attempt to combine existing wet lab and dry lab tools and mathematical modelling to tackle the challenges. Miles of success have been recorded; however, many genetic disorders remain etiologically unelucidated, awaiting diagnostic, prognostic and therapeutic tools. This piece will try to employ simple tools like multiple sequence alignment and phylogenetic analysis to provide as much as insights into the issue.

Results

49 reviewed sequences of IPR000885 and 76 reviewed sequences from Homo sapiens out of 386 reviewed sequences of IPR008160 are used for downstream analysis along with one pathogenic variant of COL1A1, namely p.Gly788Ser and one pathogenic variant of COL5A2, namely p.Gly1149Arg . ClustalW performed on MEGAX computed an MSA, referred as MSA I, of length 1982 and a mean conservation score of 0.26 for IPR000885 sequences and an MSA, referred as MSA II, of length 4089 and a mean conservation score of 0.07. Visualizing conservation scores across the MSA I reveals a long stretch of region along the Fibrillar-collagen, C-terminal super family members, characterized by a periodicity of Glycine amino acids (Figure 3). However, visualizing conservation scores across the MSA II reveals a noisier picture of conservation (Figure 4). Conservation score for 788th residue of COL1A1 (1286th position of MSA I, Figure 5), the position of the variant p.Gly788Ser, is approximately 0.55, and conservation score for the 1149th residue of COL5A2 (1614th position of MSA I, Figure 6), the position of the variant p.Gly1149Arg is approximately 0.52. UPGMA tree constructed on MSA I comprises of taxa with the same type of protein sequenced from different types of organisms. Moreover, sequences of COL1A1 and COL1A2 are clustered as two separate taxa connected with a single branch and COL5A2 from Homo sapiens are connected to the tree after this branch (Figure 7). UPGMA tree constructed on MSA II comprises of taxa including various forms of Collagen helix repeat super family members (Figure 8), one specific taxa places COL1A1 and COL5A2 together (Figure 9).

Figures

Figure 1: MSA I Focused on the 788th Residue of COL1A1

Figure 2: MSA I Focused on the 1149th Residue of COL5A2

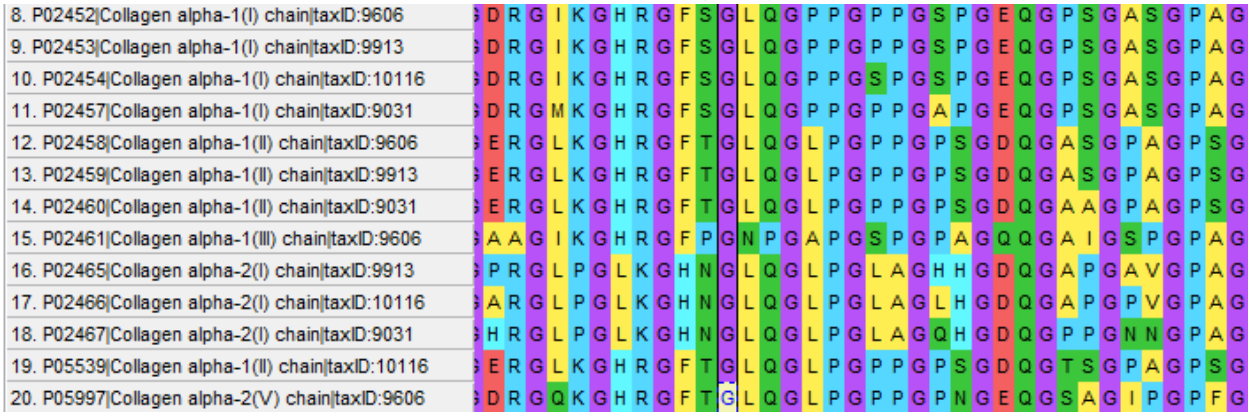


Figure 3: Conservation Scores across the MSA I

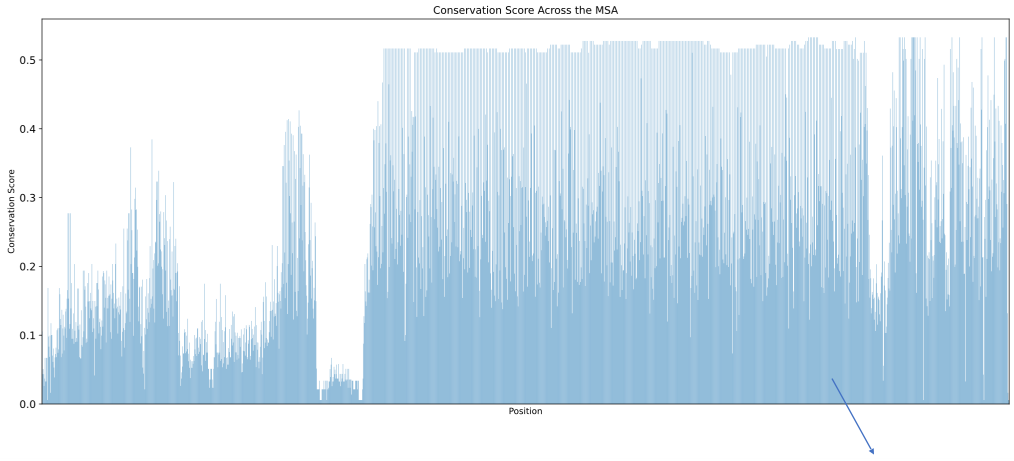


Figure 4: Conservation Scores across the MSA II

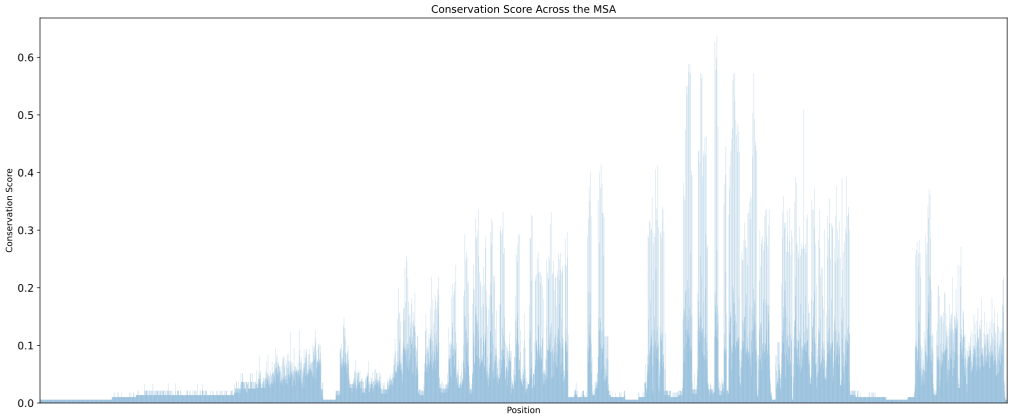


Figure 5: Conservation Scores across the MSA I, Position Interval: [1201, 1299]

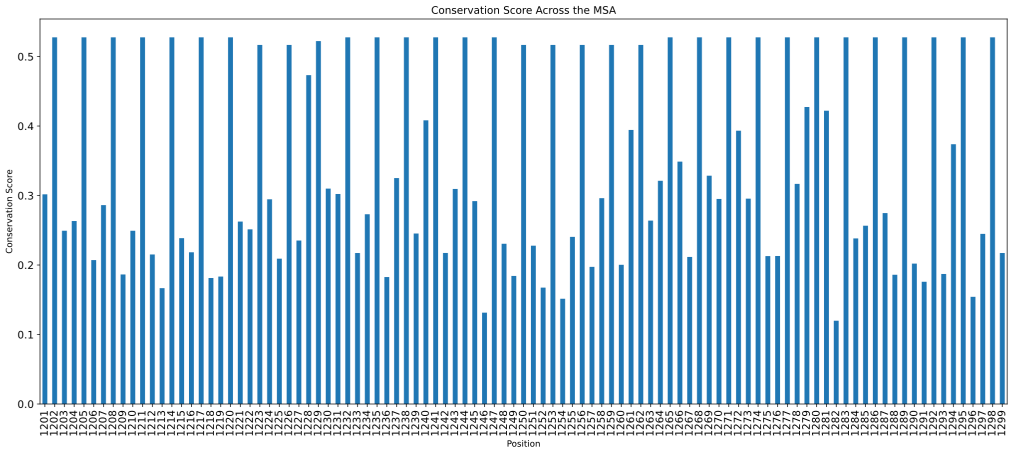


Figure 6: Conservation Scores across the MSA 1, Position Interval: [1601, 1699]

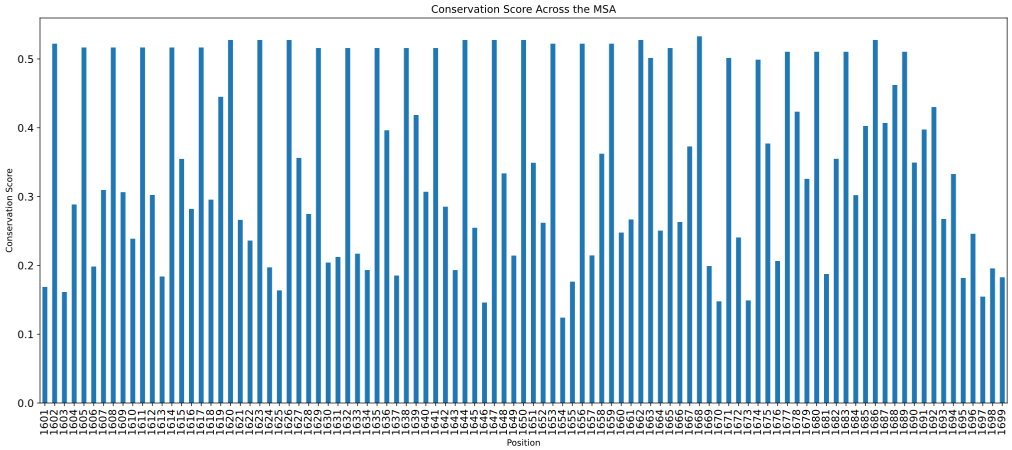


Figure 7: UPGMA Tree Constructed on MSA I

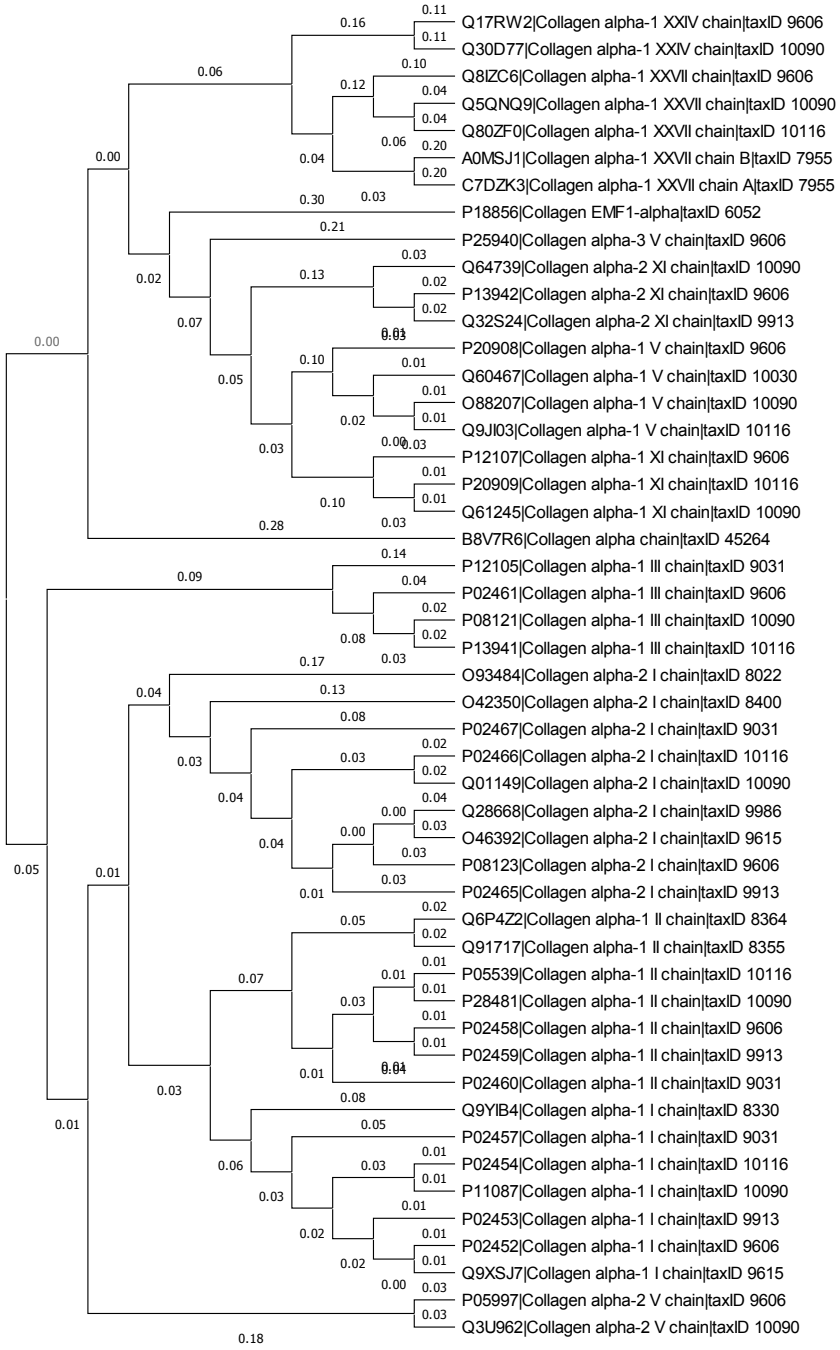


Figure 8: UPGMA Tree Constructed on MSA II

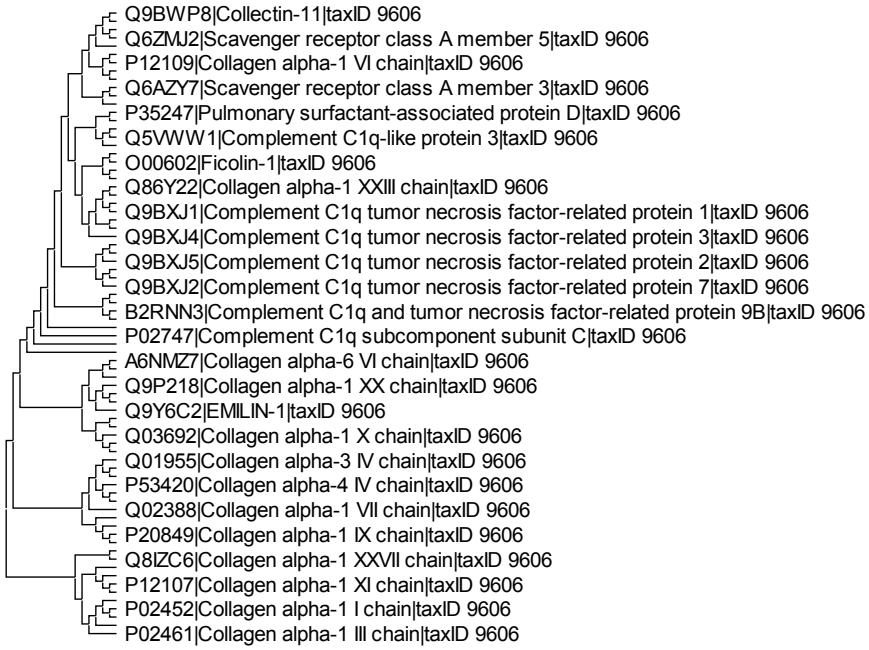
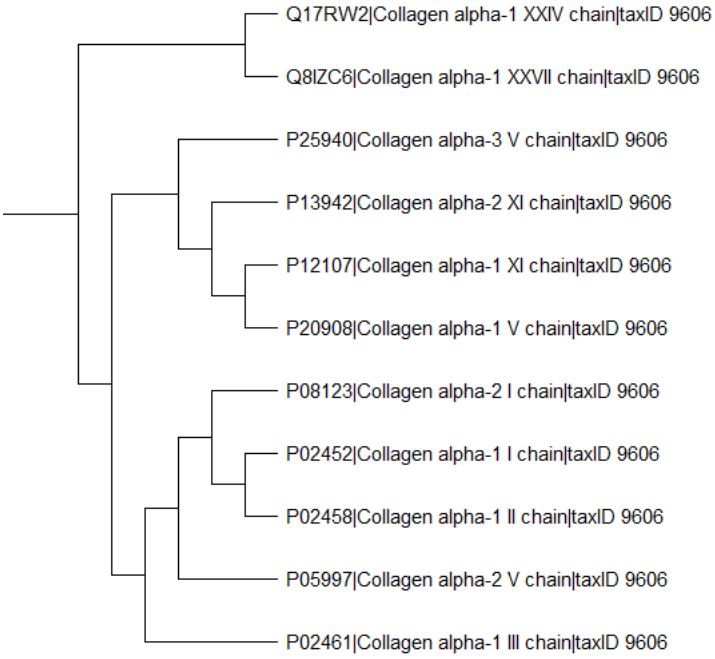


Figure 9: A Single Taxa from the UPGMA Tree Constructed on MSA II



Discussion

Even though, understanding of the etiology of rare diseases are relatively easier than understanding the etiology of complex diseases such as Alzheimer's and Parkinson's disease, much more remains to be elucidated in terms of understanding genotype-phenotype interplay for rare diseases. One approach taken by geneticists to tackle the pathogenicity of variants across genome involves phylogenetic analysis. Here, it's showed that a simple formulation of a conservation score may provide insights into this interplay. Formulated conservation score revealed the high-degree conservation of long stretches along Fibrillar collagen, C-terminal superfamily, revealing its periodic nature, where Glycine amino acids are interspaced by two amino acids which are as well conserved to a relative degree than the Glycine amino acids. The pathogenic variants selected from ClinVar showed that the conservation score provides insights into the pathogenicity of a variant. These variants also reveal that motif disturbing missense situations may cause pathological results. Phylogenetic analyses reveal that the interspecies variation among Collagen members is less than the intraspecies variation. The analyses also reveal that COL1A1 and COL5A2 sequences are closely related, providing more evidence that the knowledge on COL1A1 and COL5A2 variants as pathogenic for classical Ehlers-Danlors syndrome. More analyses required to conclude that the variants of COL1A1 and COL5A2 scatter along the gene in a biased manner rather than a random manner, however our findings suggest that variants within conserved regions are more susceptible to realize a pathogenic phenotype.

Materials and Methods

Ehlers-Danlors syndrome has been selected through searching OrphaNet database. The list of causal genes is exhausted and COL1A1 along with COL5A2 are selected for downstream analysis. UniProtKB search on P02452 (COL1A1) and P05997 (COL5A2) revealed that these proteins are a member of Collagen triple helix repeat (IPR008160) and Fibrillar collagen, C-terminal (IPR000885) super families. Reviewed sequences from both families are retrieved from Interpro database as fasta files. Sequences of IPR008160 are portioned by a Python script, to subset the sequences that belong to Homo sapiens. Subset IPR008160 and whole IPR00885 sequences are then aligned on MEGAX software by ClustalW, then UPGMA trees are built by MEGAX. In order to assess the conservation along the multiple sequence alignment (MSA) a conservation score for a given MSA column is formulated as follows:

$$C = 1 - \frac{H}{\max(H)}$$

$$H = -\sum_{i \epsilon AA} p_i * log_2(p_i)$$

$$AA = the\ set\ of\ amino\ acid\ symbols,\ exluding\ the\ symbol\ for\ a\ gap$$

A custom Python script is used to compute and visualize conservation scores along the MSA. Lastly, pathogenic variants of COL1A1 and COL5A2 are searched on ClinVar database, two variants are selected for visualization of the residue within the MSA and for visualization of the conservation for the residue containing MSA column.

References

1. Defendi GL. Genetics of Ehlers-Danlos Syndrome. *Medscape Reference*. August, 2015; http://emedicine.medscape.com/article/943567-overview.
2. Malfait, F., Francomano, C., Byers, P., Belmont, J., Berglund, B., Black, J., Bloom, L., Bowen, J. M., Brady, A. F., Burrows, N. P., Castori, M., Cohen, H., Colombi, M., Demirdas, S., De Backer, J., De Paepe, A., Fournel-Gigleux, S., Frank, M., Ghali, N., … Tinkle, B. (2017). The 2017 international classification of the Ehlers–Danlos syndromes. *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*, *175*(1), 8–26. https://doi.org/10.1002/ajmg.c.31552
3. Orphanet. [cited 2021 Jan 4]. Available from: https://www.orpha.net/
4. Pauker SP & Stoler J. Clinical manifestations and diagnosis of Ehlers-Danlos syndromes. *UpToDate*. February 22, 2016; http://www.uptodate.com/contents/clinical-manifestations-and-diagnosis-of-ehlers-danlos-syndromes.