PROJECT REPORT

Investigating the molecular cause of hypermobile Ehlers-Danlos syndrome

Cay Rahn
6255648

Course: MSB1014 Network Biology
Program: Master Systems Biology
Faculty: FSE
Academic Year: 2023/24

October 27, 2023

1. Introduction

Ehlers-Danlos syndromes are a group of heritable connective tissue disorders that can be classified into multiple subtypes. The current classification describes 14 subtypes [1, 2], with hypermobile EDS (hEDS) being the most common form and the only subtype with an unknown molecular cause, leading to a diagnosis based on clinical presentation. Identifying the molecular cause is crucial to improving the diagnosis process, understanding the disease and finding potential treatment options [3].

Despite intensive research, a clear molecular cause with related to connective tissue has yet to be established [4], and results of extensive gene analysis of many affected individuals are only expected in 2025 [5]. Until then, utilising already collected data is essential to understand more about hEDS. However, a change in diagnosis criteria in 2017 resulted in data collected earlier not being usable anymore [6, 7].

This project aims to investigate the molecular cause of hypermobile EDS by studying the influence of differentially expressed genes in hEDS patients. It mainly tries to find molecular functions and biological processes affected by differentially expressed genes similar to the ones affected by the molecular cause of other EDS types. This analysis eventually aims to find genes differentially expressed in hEDS that are candidates for being the molecular cause of hEDS and to relate the findings to existing research.

2. Methods

The following structured approach will be pursued to answer the research question:

Analysis of Differentially Expressed Genes (DEGs). The used dataset of gene expression profiles from dermal fibroblasts from patients with hEDS and healthy controls is available at the NCBI GEO database with the accession number GSE218012 [3]. The analysis is performed with the R-packages DeSeq2 (v1.40.2) and limma (v.3.56.2) based on the R-Script from GEO2R to identify up-regulated and down-regulated genes [8, 9]. Genes with a log2-fold change $> \pm 0.5$ and a by the Benjamin-Hochberg procedure adjusted p-value < 0.05 are included. The cut-offs were chosen based on similar research [10, 11].

Network Creation. The Protein-protein interaction (PPI) network is created in Cytoscape (v3.10.1) [12] by querying the before-identified DEGs from the STRING database with a medium confidence interaction score (> 0.4) [13]. Since hEDS belongs to the family of Ehlers-Danlos syndromes, its molecular cause is most likely closely related to other EDS types. Therefore, the network is expanded by additionally querying genes related to other EDS types retrieved from Disease ontology (Disease Ontology ID 13359) [14]. The resulting network is annotated with the differential expression data.

Gene Ontology and Clustering. GeneOntology (GO) enrichment is performed using the R-package clusterProfiler (v4.8.3) to gain insight into biological processes and molecular functions affected by DEGs, with results with $p < 0.05$ being considered significant [15, 16, 17]. To attain more detailed insights into specific parts, the created network is clustered using two different algorithms: MCODE and Community Clustering. MCODE is a clustering algorithm designed to find highly connected regions in PPI networks that might represent molecular complexes [18]. On the other hand, GLay, a community clustering algorithm, is designed for a functional interpretation of biological processes and pathways. Clustering is performed in Cytoscape using the clusterMaker2 app (v2.3.4) with the default parameters [12, 19]. Only clusters of more than 15 genes are considered to ensure relevance and keep the analysis feasible. Analysis of the clusters includes investigating whether genes are clustered with genes that cause other EDS types and whether the clusters consist of up-regulated or down-regulated genes or a combination of both. Heat Diffusion is applied on larger clusters to identify genes closely connected to EDS genes not captured by smaller clusters, starting with the EDS nodes using Cytoscape

functionality [20], using a time parameter of $t = 0.3$.

3. Results and Discussion

3.1. Differentially Expressed Genes and Network Creation

Under the chosen thresholds discussed in section 2, 908 genes were found to be differentially expressed (495 up-regulated and 413 down-regulated). STRING could query 828 of them; using other ID types did not change this. After querying the additional EDS-related genes, the resulting network consists of 847 genes and 6129 connections. The position of the known EDS genes in the network is, on average, more central than expected by chance based on degree, clustering coefficient, betweenness centrality and closeness centrality, supporting the close relationship between hEDS and other EDS types.

3.2. Gene Ontology and clustering

GO-enrichment is performed on the DEGs to acquire an overview of over-represented molecular functions, biological processes, and cellular components. In contrast to later results, which include the known EDS genes, this analysis is performed purely on the differentially expressed genes. The over-representation analysis shows terms related to chromatin, nucleosomes, the cell cycle signalling and regulation and the chromosomal region. The exact GO terms are shown in Table 1 and are generally consistent with findings of previous research [21, 7, 6].

GO term	Description
GO:0030527	structural constituent of chromatin
GO:0046982	protein heterodimerization activity
GO:0005201	extracellular matrix structural constituent

(a) Molecular Function

GO term	Description
GO:0000786	nucleosome
GO:0032993	protein-DNA complex
GO:0098687	chromosomal region
GO:0000775	chromosome, centromeric region
GO:0062023	collagen-containing extracellular matrix

(b) Cellular Component

GO term	Description
GO:0006334	nucleosome assembly
GO:0034728	nucleosome organization
GO:0065004	protein-DNA complex assembly
GO:0071824	protein-DNA complex organization
GO:0000075	cell cycle checkpoint signaling
GO:1901988	negative regulation of cell cycle phase transition
GO:0010948	negative regulation of cell cycle process
GO:0045786	negative regulation of cell cycle

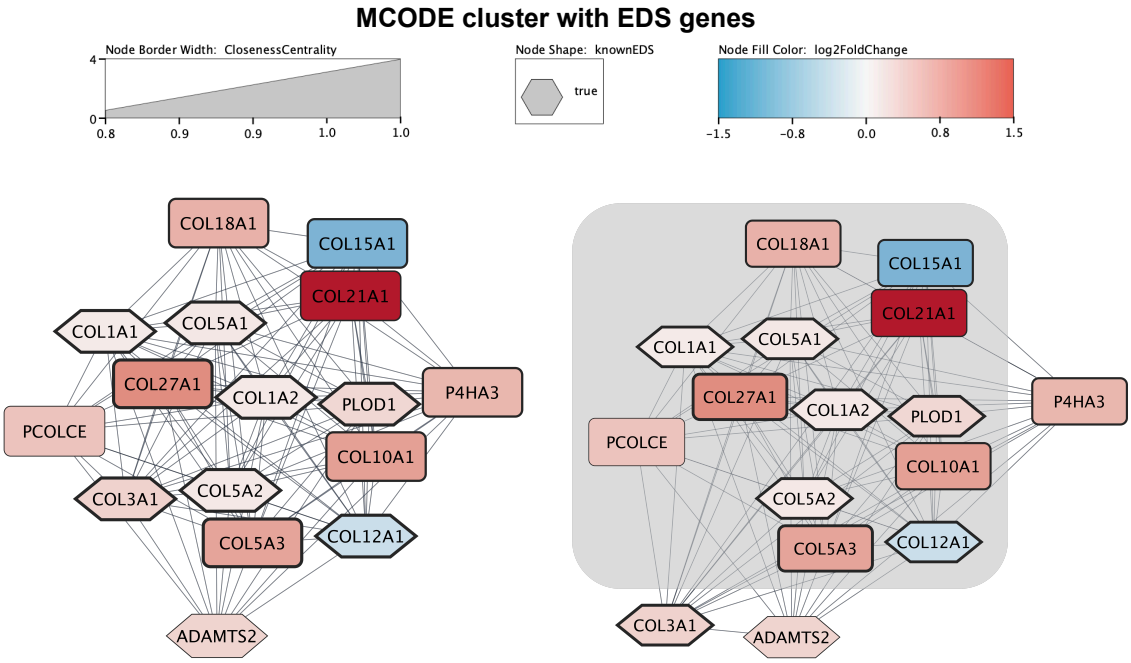
(c) Biological Process

Table 1: GO terms for the differentially expressed genes.

3.2.1. MCODE

Running MCODE on the created networks finds 3 clusters with more than 15 genes, one with 66 genes and 1953 connections, one with 44 genes and 686 connections and one with 16 genes and 114 connections, with the two larger clusters containing up-regulated genes only and no genes known to cause other EDS types.

MCODE cluster with EDS genes. The third, smaller cluster, shown in Figure 1, contains mainly up-regulated genes, including eight EDS-related genes with no significant differential expression and nine differentially expressed genes. One of the EDS genes is also one of the two down-regulated genes. Notably, all EDS genes besides ADAMTS2 exhibit high Closeness Centrality, consistent with EDS genes being more central in the complete network.



(a) Without visualisation of the ECM GO-term

(b) With visualisation of the ECM GO-term

Figure 1: The MCODE cluster contains many genes known to cause other types of EDS.

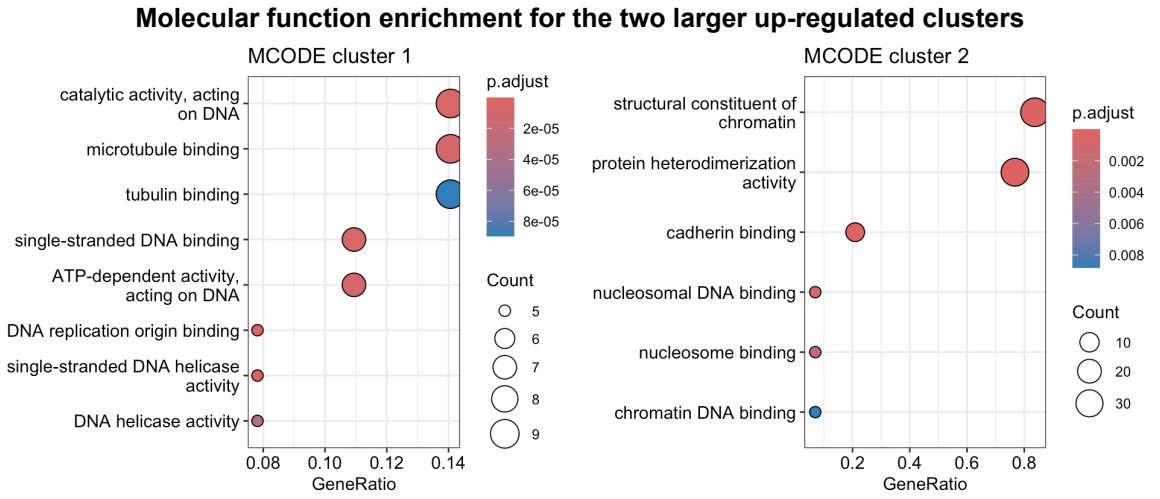
GO-Enrichment testing over-representation of molecular functions of the cluster returns two GO terms related to the extracellular matrix, GO:0005201 and GO:0030020, with the second one being a subterm of the first. Both GO terms contain the same EDS genes and seven, respectively, six other genes, with PCOLE being the only one not present in the subterm. The collagen-related genes COL27A1 and COL5A3 have a central position in the cluster and relatively strong differential expression. The gene COL21A1 is very strongly differentially expressed (log2FoldChange > 2), more than twice as high as the other genes, while being less central. The collagen encoded by this gene maintains the integrity of ECM and is a paralog to COL5A1, a known EDS gene [22]. The deficiency of the collagen type encoded by COL15A1, the only significantly down-regulated gene, is associated with muscle and microvessel deterioration in mice. This gene is a paralog of COL5A1 as well.

Finding connections to ECM in the enrichment analysis is consistent with similar research [7, 6, 21]. Although the connection of the ECM is already known, the visualisation provided in Figure 1 allows us to assess the centrality together with the differential expression, letting the before-mentioned genes stand out.

Up-regulated MCODE Clusters. The two larger, up-regulated MCODE clusters show no relation to ECM terms, as is shown in Figure 2. It is noticeable that only a few genes of the first cluster are part of the enriched terms, while the gene ratio is much higher for the second cluster, with up to around 80 % of the genes being involved in the first two terms. Therefore, the enrichment of the first cluster provides little insight into molecular function in hEDS patients. On the other hand, the up-regulation seen in the second cluster for the GO-term GO:0030527, the structural constituent of chromatin, is interesting because earlier research found down-regulated genes involved in processes related to chromatin in vEDS [21].

3.2.2. Community Clustering

Community Clustering finds six clusters of more than 15 genes. Three are very small and loosely connected clusters containing 18 to 29 genes. These clusters are omitted from the analysis due to



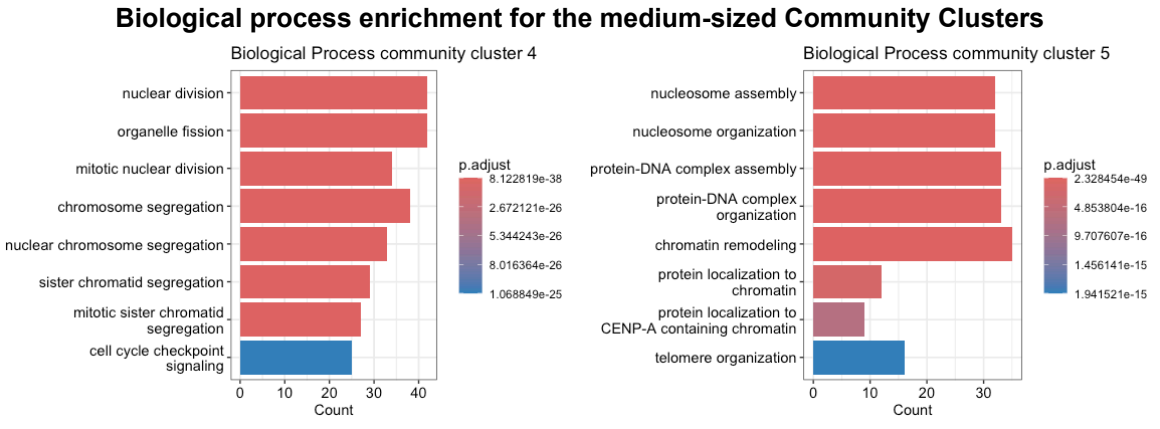
(a) First MCODE cluster

(b) Second MCODE cluster

Figure 2: The results of the GO-enrichment for molecular function of the two up-regulated, larger MCODE clusters.

their small size and loose connectivity. Additionally, there are two medium-sized highly connected clusters with 76 and 105 genes and 1003 and 2330 interactions, respectively, and one huge cluster with 363 genes and 1661 interactions.

Medium-Sized Community Clusters. The medium-sized clusters are highly interconnected and contain mainly up-regulated genes. Genes related to nucleosome and protein-DNA complex are part of one of the clusters, while the genes related to the chromosomal region are clustered in the other one. This distribution is mirrored in the enriched biological processes shown in Figure 3. In the cluster with nucleosome-related genes, five genes are notably highly up-regulated: H3C2, H3C7, H2BC10, ASF1B and WDR37. The first four are nucleosome-related, while WDR37 is related to cell cycle progression, signal transduction and gene regulation. The other cluster has no genes that are significantly more up-regulated than others.



(a) First medium-sized Community cluster

(b) Second medium-sized Community cluster

Figure 3: The results of the GO-enrichment for biological processes of the two medium-sized Community Clusters.

Largest Community Cluster. The largest cluster contains a mix of up-regulated and down-regulated genes and all 21 EDS-related genes. Furthermore, it contains all genes being part of the GO-term

for the ECM found in the over-representation of molecular functions of the MCODE cluster containing the EDS genes. The molecular cluster showing enrichment towards the chromatin part is not a part of this community cluster. Heat Diffusion starting at EDS genes finds 39 DEGs with a heat > 0.1 beside the starting nodes. These hot genes intersect with the MCODE cluster containing the EDS genes, as expected due to their close connection reflected in the clustering. Besides those, there are 31 hot genes, showing that the EDS genes have a central position close to ECM-related genes. GO enrichment of the hot genes not intersecting with the EDS-related MCODE cluster shows an ECM relation as well.

3.3. Conclusion

The analysis is consistent with previous research in finding DEGs that influence the extracellular matrix, chromatin, nucleosome and the cell cycle [21, 7]. Furthermore, clustering revealed which differentially expressed genes are closely related to the ones causing other EDS types. By providing a visualisation of the involvement in the extracellular matrix by simultaneously providing information about the differential expression, genes of particular interest can be identified. These results are complemented by using Heat Diffusion in the largest Community Cluster containing EDS-related genes.

Moreover, clustering revealed that mainly up-regulated genes are closely connected; partially overlapping clusters were found by MCODE and Community Clustering. Some of the related GO terms are known to be involved in other EDS types and to involve down-regulated genes there are up-regulated in the given data. Generally, hEDS has a broad spectrum of clinical representations that might be caused by multiple components being involved as part of the molecular cause. This spectrum is reflected in the found clusters representing different processes involved.

Clear candidate genes cannot be provided although affected biological processes are identified and genes of interest described. Future work should include results from pathway analysis and analyse the created network with the found clusters in more detail.

A. Supplementary Material

All related scripts, data and Cytoscape sessions can be found in the following GitHub repository: `https://github.com/Zianor/MSB1014-NetworkBiology`. Information about versions of the used software and packages is included there as well.

References

[1] Malfait F, Francomano C, Byers P, Belmont J, Berglund B, Black J, et al. The 2017 international classification of the Ehlers–Danlos syndromes. American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 2017 3;175:8-26.

[2] Malfait F, Castori M, Francomano CA, Giunta C, Kosho T, Byers PH. The Ehlers–Danlos syndromes. Nature Reviews Disease Primers 2020 6:1. 2020 7;6:1-25.

[3] Ritelli M, Colombi M. Molecular Genetics and Pathogenesis of Ehlers-Danlos Syndrome and Related Connective Tissue Disorders. Genes. 2020 5;11.

[4] Caliogna L, Guerrieri V, Annunziata S, Bina V, Brancato AM, Castelli A, et al. Biomarkers for Ehlers-Danlos Syndromes: There Is a Role? International Journal of Molecular Sciences 2021, Vol 22, Page 10149. 2021 9;22:10149.

[5] HEDGE (Hypermobile Ehlers-Danlos Genetic Evaluation) Study - The Ehlers Danlos Society;. Available from: `https://www.ehlers-danlos.com/hedge/`.

[6] Gensemer C, Burks R, Kautz S, Judge DP, Lavallee M, Norris RA. Hypermobile Ehlers-Danlos syndromes: Complex phenotypes, challenging diagnoses, and poorly understood causes. De-

velopmental dynamics : an official publication of the American Association of Anatomists. 2021 3;250:318.

[7] Ritelli M, Chiarelli N, Cinquina V, Zoppi N, Bertini V, Venturini M, et al. RNA-Seq of Dermal Fibroblasts from Patients with Hypermobile Ehlers-Danlos Syndrome and Hypermobility Spectrum Disorders Supports Their Categorization as a Single Entity with Involvement of Extracellular Matrix Degrading and Proinflammatory Pathomechanisms. Cells. 2022 12;11.

[8] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 12;15:1-21.

[9] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.

[10] Karimizadeh E, Sharifi-Zarchi A, Nikaein H, Salehi S, Salamatian B, Elmi N, et al. Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis. BMC Medical Genomics. 2019 12;12:1-12.

[11] Lim PJ, Lindert U, Opitz L, Hausser I, Rohrbach M, Giunta C. Transcriptome Profiling of Primary Skin Fibroblasts Reveal Distinct Molecular Features Between PLOD1- and FKBP14-Kyphoscoliotic Ehlers–Danlos Syndrome. Genes 2019, Vol 10, Page 517. 2019 7;10:517.

[12] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003 11;13:2498-504.

[13] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic acids research. 2023 1;51:D638-46.

[14] Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: Classification, content and workflow expansion. Nucleic Acids Research. 2019 1;47:D955-62.

[15] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics 2000 25:1. 2000 5;25:25-9.

[16] Consortium TGO, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 5;224.

[17] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation. 2021 8;2.

[18] Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003 1;4:1-27.

[19] Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. BMC bioinformatics. 2011 11;12.

[20] Carlin DE, Demchak B, Pratt D, Sage E, Ideker T. Network propagation in the cytoscape cyberinfrastructure. PLOS Computational Biology. 2017 10;13:e1005598.

[21] Chiarelli N, Carini G, Zoppi N, Ritelli M, Colombi M. Transcriptome analysis of skin fibroblasts with dominant negative COL3A1 mutations provides molecular insights into the etiopathology of vascular Ehlers-Danlos syndrome. PLOS ONE. 2018 1;13:e0191220.

[22] COL21A1 collagen type XXI alpha 1 chain [Homo sapiens (human)] - Gene - NCBI;. Available from: https://www.ncbi.nlm.nih.gov/gene/81578/#summary.