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].

Although many studies investigated several genes, no clear molecular cause with a connection to connective tissue has been established yet [4]. While an ongoing reseach aims to find the genetic cause of hEDS by analysing the genes of many affected individuals, results 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 and limma 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 [12] by querying the before-identified DEGs from the STRING database with an interaction score > 0.4, which reflects medium confidence [13]. Since hEDS belongs to the family of Ehlers-Danlos syndromes, its molecular cause is most likely closely related to other EDS types. The PPI network of the DEGs is therefore 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 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, resulting in different cluster structures: MCODE to analyse the molecular function and Community clustering to investigate biological processes. Only clusters of more than 15 genes are considered to ensure relevance and keep the analysis feasible. Further analysis of the resulting 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.

MCODE MCODE is a clustering algorithm designed to find highly connected regions in PPI networks that might represent molecular complexes [18]. MCODE was applied in Cytoscape with the clusterMaker2 app using the default parameters [12, 19].

Community Clustering GLay, a community clustering algorithm, was designed to be used for a func-

tional interpretation of clusters in networks and used for this purpose here [20]. Analogous to the MCODE, clustering was performed with clusterMaker2 and Cytoscape using the default parameters [12, 19].

Heat Diffusion is applied on larger clusters to identify genes closely connected to EDS genes that are not captured by smaller clusters, starting with the EDS nodes using Cytoscape functionality [21], 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. Around half (495) are up-regulated, and the remaining 413 are 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. Enrichment analysis 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.

Biological processes over-represented in the DEGs include cell-cycle regulation, signalling and transition, nucleosome assembly and organisation and protein-DNA assembly and organisation. Cellular components affected are the nucleosome (GO:0000786), the chromosomal region (GO:0098687 and GO:0000775), the protein-DNA complex (GO:0032993) and the collagen-containing extracellular matrix (ECM) (GO:0062023). Over-representation analysis for the molecular function is less meaningfull on a large network. Still, three terms are related to a significantly higher number of genes than the rest: The structural constituent of chromatin (GO:0030527), protein heterodimerization activity (GO:0046982) and extracellular matrix structural constituent (GO:0005201).

[TODO: Do i want to look at Pathways? lupus and alcoholism maybe not that interesting, but Cell cycle?]

[TODO: Analysis]

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 but also two down-regulated genes. Some do not show a high differential expression but are genes known to cause other EDS types. The cluster consists of eight EDS genes with a $|\text{log2FoldChange}| < 0.5$ and nine differentially expressed genes. One of the EDS genes is also one of the two down-regulated genes. All known EDS genes besides ADAMTS2 have a high Closeness Centrality, consistent with the findings of EDS genes being more central in the complete network. COL21A1 shows the highest differential expression ($\text{log2FoldChange} > 2$), more than twice as high as the other genes while being less central in the cluster.

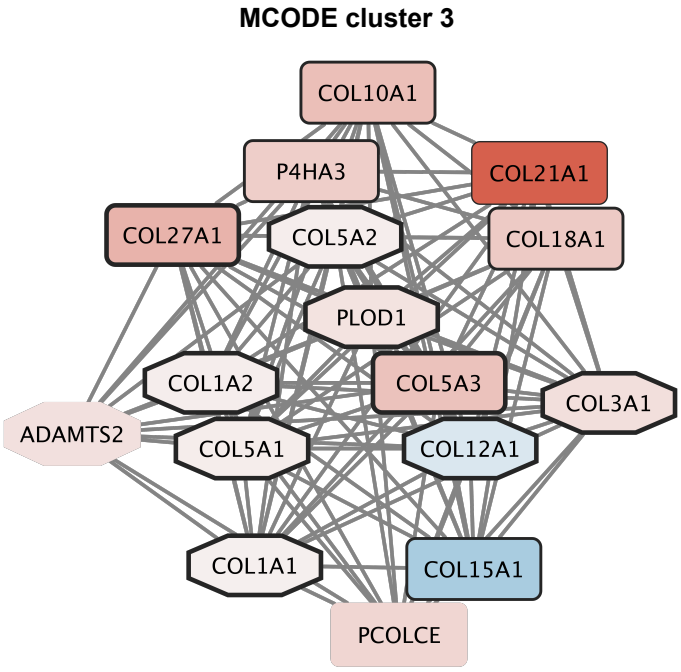


Figure 1: The MCODE cluster contains many genes known to cause other types of EDS. [TODO: add legend for shape, border and colour]

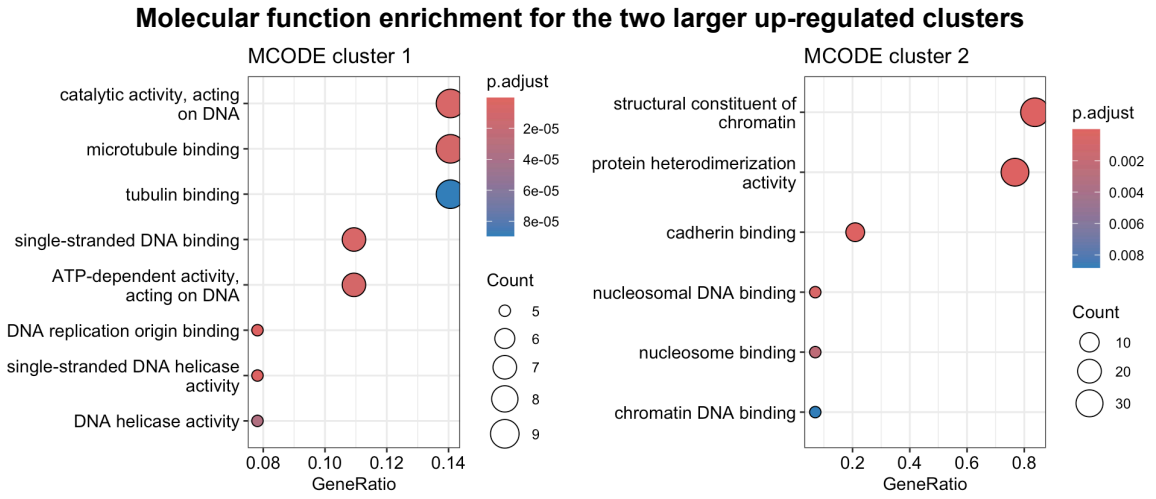
GO-Enrichment testing overrepresentation of molecular functions of the cluster returns extracellular matrix in two terms, GO:0005201 and GO:0030020, with the second one being a subterm of the first. The first describes the action of a molecule that contributes to the structural integrity of the extracellular matrix; the second is a constituent of the extracellular matrix that enables the matrix to resist longitudinal stress. Both GO terms contain the same EDS genes and seven, respectively, six other genes, with PCOLE being the only gene not present in the GO subterm. These genes are investigated in more detail regarding their centrality, differential expression and what is known about them. COL27A1, a fibrillar collagen gene, has a central position in the cluster and relatively strong differential expression. The same applies to COL5A3, another gene related to collagen. The gene COL21A1 is very strongly differentially expressed, as mentioned before. It encodes the alpha chain of XXI collagen, which maintains the integrity of ECM and is a paralog to COL5A1, a known EDS gene [22].

To find connections to ECM in the enrichment analysis is consistent with findings of similar research [7]. [Todo: there was other research, find] The affection of ECM with particular disorganisation of collagen and fibronecting was found in hEDS and two other EDS types [23]. [TODO: point out what my new findings are, COL21A1, etc.]

Up-regulated MCODE cluster

The two larger, up-regulated MCODE clusters show no over-representation in ECM terms, as is shown in Figure 2. It is noticeable that the enrichment of the first cluster shows that only a few genes are part of the enriched terms. For the second cluster, the gene ratio is much higher, with up to around 80 % of the genes being involved in the second 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



(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

involved in processes related to chromatin in vEDS [23].

3.2.2. Community Clustering

Community Clustering results in six clusters with more than 15 genes. Three are very small and loosely connected clusters, containing 18 to 29 genes and approximately the same amount of connections as genes. Since clusters of less than 20 genes are not large enough for analysis of biological processes, smaller clusters are omitted from the analysis. Additionally, there are two medium-sized highly connected clusters with 76 genes and 100 connections, and 105 genes and 2330 connections, respectively, and one huge cluster with 363 genes and 1661 connections. The medium-sized clusters are highly interconnected and contain mainly up-regulated genes.

Largest Community Cluster

The largest cluster contains a mix of up-regulated and down-regulated genes, and all 21 genes related to other EDS types. 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 DEG with a heat > 0.1 beside the starting nodes. These hot genes intersect with the MCODE cluster containing the EDS genes as it is expected due to their close connection reflected in the clustering. Beside those, there are 31 hot genes.

[TODO: check whether they were mentioned in other research and if we see them somewhere else]

Medium Sized Community Clusters

- Nucleosome and protein-DNA complex in one of the two medium sized clusters while the genes related to the chromosomal region are clustered in the other one.
- first one: 5 highly differentielly expressed genes noticeable: H3C2, H3C7, H2BC10, ASF1B, WDR37

3.3. Discussion and Conclusion

- ECM/collagen in genes closely related to genes of other EDS types, seen in community cluster and MCODE cluster
- TODO: link to research, meaning

- chromatin related terms in a MCODE cluster and a community cluster found, in community cluster together with nucleosome
- chromatin interesting because down-regulated in other EDS type
- TODO: link to research, meaning
- heat diffusion in largest community cluster showed other genes additionally to ECM genes that are probably interesting
- project showed genes fullfilling similar roles than genes in other EDS types, candidates for causing hEDS
- hEDS has wide spectrum of representations, probably multiple components represented in the found clusters

A. Supplementary Material

All related scripts, data and Cytoscape sessions can be found in the following GitHub repository: `https://github.com/Zianor/MSB1014-NetworkBiology`.

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. Developmental 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] Su G, Kuchinsky A, Morris JH, States DJ, Meng F. GLay: community structure analysis of biological networks. Bioinformatics. 2010 12;26:3135-7.

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

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

[23] 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.