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 21, 2023

1 Introduction

Ehlers-Danlos syndromes are a group of heritable connective tissue disorders that can be classified into multiple subtypes. The current classification proposed in 2017 describes 13 subtypes [1] and was later extended by one additional type [2]. The most common type is hypermobile EDS (hEDS) which is also the only subtype with unknown molecular cause, leading to a diagnosis based on clinical presentation. Identifying the molecular cause is crucial to improve the diagnosis process, the understanding of the disease and to find 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 study aims to find the genetic cause of hEDS by analysing the genes of around 1000 affected individuals, results are not expected before 2025 [5]. Until then, it is essential to utilise the already collected data to understand more about hEDS. One problem with this is that the diagnosis criteria changed with the new subtype classification in 2017, causing individuals diagnosed with hEDS before this change to not fulfilling the criteria anymore and being diagnosed with Hypermobility Spectrum Disorder (HSD) [6]. However, the clinical representation of those two diagnoses overlaps, and the terms are often used interchangeably. Both diagnoses are also often grouped as hEDS/HSD because it is currently not clear whether they are different to each other or not [6, 7].

- we know the clinical representation and how other eds types work (kind of)
- TODO: include information of what is affected, extracellular matrix, collagen, connective tissue

This project aims to investigate the molecular cause of hypermobile EDS by studying the influence of differentially expressed genes in hEDS patients.

- which biological processes and pathways are affected
- similarities to pathways/processes/function affected by other EDS genes
- TODO: what do we want to do in the big picture: find candidate genes? is this to ambitious?

2 Methods

2.1 Analysis of Differentially Expressed Genes and Network Creation

A dataset of gene expression profiles from dermal fibroblasts from patients with hEDS and healthy controls is used. It is available at the NCBI GEO database with the accession number GSE218012 [3]. The analysis is performed with the R-package DeSeq2 based on the analysis script exported from GEO2R to identify up-regulated and down-regulated genes [8]. Genes with a log2-fold change $> \pm 0.5$ and a p-value < 0.05 after adjustment based on False-Discovery-Rate using the Benjamin-Hochberg procedure were considered. The cut-offs were chosen based on similar research [9, 10].

The Protein-protein interaction (PPI) network is then created in Cytoscape [11] by querying the before identified differentially expressed genes from the STRING database with an interaction score > 0.4 which reflects a medium confidence [12]. Since hEDS belongs to the family of Ehlers-Danlos syndroms its molecular cause is most likely closely related to other EDS types. The PPI network of the differentially expressed genes is therefore expanded by additionally querying genes related to other EDS types. The genes are retrieved from Disease Ontolgy with Disease Ontology ID 13359 [13] and

similarly queried from STRING with a medium confidence threshold as well. The resulting network is annotated with the data obtained by the data from the analysis of the differential expression of the genes.

2.2 Enrichment analysis and clustering

- exploratory enrichment analysis to gain insights about affected pathways, biological processes, molecular function and cellular components
- use R-package clusterProfiler [14] for GO-enrichment [15, 16]
- large network, thus cluster before to get better insights for specific parts
- two different cluster methods used with different resulting cluster structure
- create subnetworks for clusters with more than 15 nodes [TODO: why this threshold? to focus on more relevant because many in MCODE between 10 and 15 nodes, for community clustering smaller ones probably found by MCODE]

2.2.1 MCODE

MCODE is a clustering algorithm designed to find highly-connected regions in larger PPI networks [17]. Such regions might represent molecular complexes which makes MCODE a suitable algorithm to analyes the molecular function of the resulting clusters that are mostly small because of MCODEs search for dense regions. MCODE was applied in Cytoscape with the clusterMaker2 app using the default parameters [18].

On the resulting clusters, GO-enrichment is performed to find over-represented molecular functions that might be involved in causing hEDS. Addititonally, it is investigated whether genes are clustered together with genes known to cause other EDS types and whether the resulting clusters consist of upregulated or downregulated genes or a combination of both.

2.2.2 Community Clustering

- small cluster not helpful for biological processes and pathways
- therefore use second clustering method
- community clustering with GLay, ”more suitable for functional interpretation” [19]

3 Results

3.1 Differentially Expressed Genes and Network creation

- 908 differentially expressed genes with chosen thresholds section 2.1
- around half are upregulated, around half downregulated (Upregulated: 495, Downregulated: 413)
- String was able to query 828 of them, using alternative ID/Query methods did not change this
- after querying additional genes that are known to be related to other eds types: resulting network with 847 nodes and 6129 edges
- position of known eds genes in network is on average more central (checked degree, clustering coefficient, betweenness centrality and closeness centrality)

3.2 Enrichment analysis and clustering

3.2.1 MCODE

- 3 clusters with more than 15 nodes (first with 66 nodes & 1953 connections, second with 44 nodes & 686 connections, third with 16 nodes & 114 connections)
- first and largest cluster upregulated only
- second one as well
- 3rd MCODE cluster, shown in figure 1, is mostly upgregulated with 2 downregulated nodes, some not relevantly differentially expressed, also only 9/16 are not known eds genes, 8 are
- the eds genes are all below the threshold of $|\text{log2FoldChange}| > 0.5$, 1 of them is 1 of the two downregulated genes
- quite interesting to see genes closely related to other eds genes, COL21A1 is also strongly upregulated ($\text{log2FoldChange} > 2$)
- all known EDS genes with ADAMTS2 as exception have a high Closeness Centrality
- regarding the enrichment of this cluster: not surprising to see extracellular matrix in two GO-Terms when testing for overrepresentation of molecular functions
- first go term: GO:0005201 - The action of a molecule that contributes to the structural integrity of the extracellular matrix. overall 13 of the 16 genes in this GO-Term, 6 are known eds genes, the other 6 differentially expressed (COL10A1, COL15A1, PCOLCE, COL5A3, COL18A1, COL21A1, COL27A1)
- second go term: GO:0030020 - A constituent of the extracellular matrix that enables the matrix to resist longitudinal stress, subtype of the first. Same EDS genes and same genes as in first term except PCOLCE, PCOLE is also the only downregulated gene [TODO: interprete]

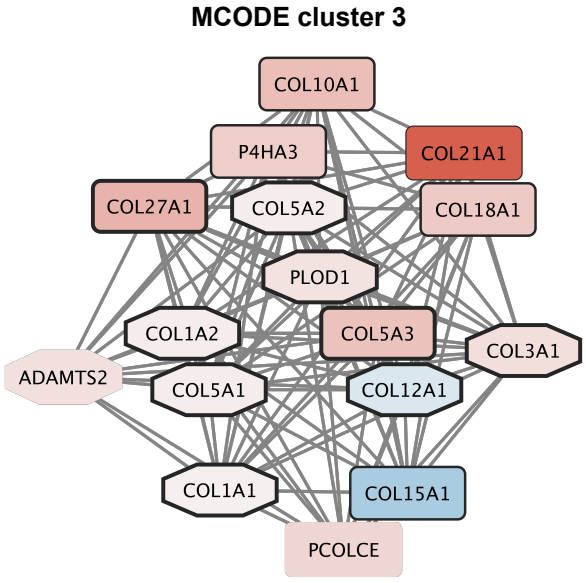

Figure 1: The MCODE cluster containing many genes known to cause other types of EDS. [TODO: add legend for shape and colour]

3.2.2 Community Clustering

- 6 clusters with more than 15 nodes

- 3 very small loosely connected clusters (18 nodes & 17 connections, 29 nodes & 32 connections, 29 nodes & 29 connections), two medium sized highly interconnected (76 nodes & 1000 conections and 105 nodes & 2330 connections) and one very large cluster (363 nodes & 1661 connections)
- especially medium sized clusters highly interconnected
- biggest one mix of up-regulated and down-regulated genes, contains all 21 genes known to cause other EDS types
- both medium sized clusters mostly upregulated $\rightarrow$ interesting!
- smallest cluster with 18 nodes to small for meaningful enrichment analysis regarding processes

4 Discussion

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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[18] 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.

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