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]. 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 [2].

Although many studies investigated several genes, no clear molecular cause with a connection to connective tissue has been established yet

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 [2]. 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 [3]. 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 where considered. The cut-offs where chosen based on similar research [4, 5].

- use Cytoscape to create network [6]
- query differentially expressed genes from string db [7] (confidence cut off 0.4)
- query eds genes related to other eds types additionally: 21 genes retrieved from Disease Ontology with Disease Ontology ID 13359 [8], queried from string, again with a medium confidence cut off of 0.4
- load data about differential expression into network, for EDS genes and differentially expressed genes

2.2 Enrichment analysis and clustering

- exploratory enrichment analysis on whole network
- use R-package clusterProfiler [9] for enrichment
- 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 finds ”densely connected regions in large protein-protein interaction networks that may represent molecular complexes” [10]
- results in smaller clusters
- suited to analyse molecular function
- use default parameters
- analysis of molecular function with clusterProfiler with GO over-representation analysis
- do we see clusters with other molecular functions than expected
- are genes clustered together with other eds genes
- are clusters generelly mostly upregulated or downregulated

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” [11]

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

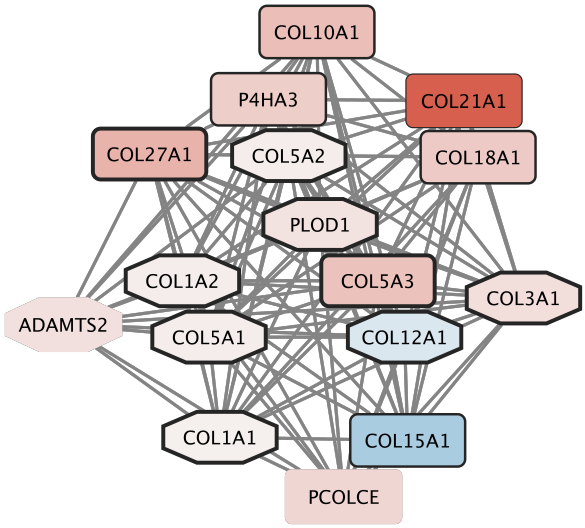

Figure 1: MCODE cluster 3 [TODO: add legend for shape and colour]

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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