Data Science 10593052 Bioinformatics & Network Medicine 2020-2021

Date: 20 Dicember 2020

Network Medicine Project report

|  |
| --- |
| Network Medicine@Data Science A.Y. 2020-2021  Manuscript Title  Author1, Author1 and Author1  1Group no. X Abstract Max 150 words. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. |

Introduction

The aim of this project is to carry out a network medicine analysis about a specific pathological condition, namely **dilated cardiomyopathy** (**DCM**). This is a primary disorder of the hearth muscle characterized by dilated left ventricle with systolic dysfunction of the left or both ventricles. DCM develops along with other pathogenetic mechanisms, such as hemodynamic overload, ventricular remodeling, accelerated apoptosis and genetic mutations.

Causes of DCM include genetics, viral infections, toxic substances such as alcohol, cocaine and other drugs. It has been observed that genetics factors play a role in 20 to 35% of cases and involve genes that encode a heterogeneous group of molecules that participate in force generation, force transmission, sarcomere integrity, cytoskeletal and nuclear architecture, electrolyte homeostasis, mitochondrial function and transcription. Indeed, many pathogenetic pathways have been hypothesized to explain the development of DCM depending the affected genes and the dislodged intracellular structures or pathways, also suggesting to think of DCM as a group of diseases, instead of a single form of cardiomyopathy.

In the following, starting from the existing resources about the genes involved in DCM, we will build the complete disease interactome and use different learning algorithms to identify potential putative disease modules.

Seed genes

The first step of our project was to build the seed list, that is, the list of all human genes involved in our target disease. We have downloaded the dataset about “*Curated gene-disease association*” available on the online open-source database ***DisGeNet*** and we have filtered it so as to keep only the records related to dilated cardiomyopathy; specifically, we just kept the records having diseaseId = C0007193*,* where the eight-digit code is the identifier of DMC on the DisGeNet database. The result of our query was a set of 48 genes identified by their Entrez Gene Id and correlated with many other information, such as their gene symbol, their disease specificity and pleiotropy indexes, their source and its related score.

Before gathering more data about our seed genes, we used the multi-symbol checker on the ***HGNC*** database to check that all of them were updated and approved, obtaining a positive feedback for the whole set. We also discovered that three genes (*FAS, RAC1, RAF1*) match both an approved and an alias symbol.

We then moved to the ***Uniprot*** website to collect some more information about the seed genes, namely the Uniprot Accession Number (*Uniprot AC*), the name of the encoded protein and a brief description of its function. The table below reports a summary of the data we have collected:

|  |  |  |  |
| --- | --- | --- | --- |
| GeneId | GeneName | UniprotAC | ProteinName |
| 58 | ACTA1 | P68133 | Actin, alpha skeletal muscle |
| 70 | ACTC1 | P68032 | Actin, alpha cardiac muscle 1 |
| 153 | ADRB1 | P08588 | Beta-1 adrenergic receptor |
| 154 | ADRB2 | P07550 | Beta-2 adrenergic receptor |
| 355 | FAS | P25445 | Tumor necrosis factor receptor |
| 356 | FASLG | P48023 | Tumor necrosis factor ligand |
| 472 | ATM | Q13315 | Serine-protein kinase ATM |
| 948 | CD36 | P16671 | Platelet glycoprotein 4 |
| 1440 | CSF3 | P09919 | Granulocyte colony-stimulating factor |
| 1482 | NKX2-5 | P52952 | Homeobox protein Nkx-2.5 |
| 1499 | CTNNB1 | P35222 | Catenin beta-1 |
| 1756 | DMD | P11532 | Dystrophin |
| 1956 | EGFR | P00533 | Epidermal growth factor receptor |
| 2194 | FASN | P49327 | Fatty acid synthase |
| 2876 | GPX1 | P07203 | Glutathione peroxidase 1 |
| 3688 | ITGB1 | P05556 | Integrin beta-1 |
| 4000 | LMNA | P02545 | Prelamin-A/C |
| 4306 | NR3C2 | P08235 | Mineralocorticoid receptor |
| 4624 | MYH6 | P13533 | Myosin-6 |
| 4625 | MYH7 | P12883 | Myosin-7 |
| 4878 | NPPA | P01160 | Natriuretic peptides A |
| 4879 | NPPB | P16860 | Natriuretic peptides B |
| 5318 | PKP2 | Q99959 | Plakophilin-2 |
| 5663 | PSEN1 | P49768 | Presenilin-1 |
| 5664 | PSEN2 | P49810 | Presenilin-2 |
| 5879 | RAC1 | P63000 | Ras-related C3 botulinum toxin substrate 1 |
| 5894 | RAF1 | P04049 | RAF proto-oncogene serine/threonine-protein kinase |
| 5973 | RENBP | P51606 | N-acylglucosamine 2-epimerase |
| 6331 | SCN5A | Q14524 | Sodium channel protein type 5 subunit alpha |
| 6389 | SDHA | P31040 | Succinate dehydrogenase |
| 6443 | SGCB | Q16585 | Beta-sarcoglycan |
| 6462 | SHBG | P04278 | Sex hormone-binding globulin |
| 6584 | SLC22A5 | O76082 | Solute carrier family 22 member 5 |
| 6648 | SOD2 | P04179 | Superoxide dismutase |
| 6934 | TCF7L2 | Q9NQB0 | Transcription factor 7-like 2 |
| 7112 | TMPO | P42166 | Lamina-associated polypeptide 2, isoform alpha |
| 7137 | TNNI3 | P19429 | Troponin I, cardiac muscle |
| 7139 | TNNT2 | P45379 | Troponin T, cardiac muscle |
| 7273 | TTN | Q8WZ42 | Titin |
| 7350 | UCP1 | P25874 | Mitochondrial brown fat uncoupling protein 1 |
| 7840 | ALMS1 | Q8TCU4 | Alstrom syndrome protein 1 |
| 8313 | AXIN2 | Q9Y2T1 | Axin-2 |
| 10060 | ABCC9 | O60706 | ATP-binding cassette sub-family C member 9 |
| 55759 | WDR12 | Q9GZL7 | Ribosome biogenesis protein WDR12 |
| 64651 | CSRNP1 | Q96S65 | Cysteine/serine-rich nuclear protein 1 |
| 137735 | ABRA | Q8N0Z2 | Actin-binding Rho-activating protein |
| 150094 | SIK1 | P57059 | Serine/threonine-protein kinase SIK1 |
| 347273 | CAVIN4 | Q5BKX8 | Caveolae-associated protein 4 |

*Table of seed genes involved in dilated cardiomyopathy and related proteins (DisGeNet, Uniprot)*

Summary on interaction data

The next step was to collect all binary protein interaction data. To get them we referred to the current release of the biomedical interaction repository ***BioGRID***.

* Since we were only interested in human interactions, we first filtered out all records in which one or both the interacting organism were not human, i.e. with (Organism ID Interactor A = 9606) or (Organism ID Interactor B = 9606).
* Next we moved to limit our dataset to the interactions involving at least one seed gene. The first thing we noticed was that the protein encoded by the gene *CAVIN4* seemed to participate to no interaction, but since when we have instead interrogated the database according to its Entrez Gene Id (347273) we got many results, we have realized that the problem was that on to the *BioGRID* database the official symbol for the gene 347273 is MURC, whereas CAVIN4 is just one of its aliases. Thus, we have modified our seed list to meet this convention. After interrogating the BioGRID database on (Entrez Gene Interactor A.isin(seed\_list)) or (Entrez Gene Interactor B.isin(seed\_list)) we obtained our first set of 12081 PPIs, of which 180 between seed genes only. We found 4886 more genes interacting with our seed genes.
* Finally, we have also retrieved the interactions among the not seed genes, this time limiting our query to those records in which both interactors were of interest. We thus obtained 259756 more PPIs.
* We gathered the interaction data in a single table of 271837 rows.

The following table summarizes the results of this section:

|  |  |
| --- | --- |
|  | Total number |
| Seed genes (DisGeNet) | 48 |
| Seed genes (BioGRID) | 48 |
| Interacting proteins | 4934 |
| Interactions | 271837 |

*Summary of the interaction data (BioGRID)*

# Interactomes data

Rearranging the interaction data previously retrieved, we have build the following two tables having as columns only the official gene symbol and the protein identifier of both interactors:

1. **Seed genes interactome**, composed of interactions involving seed genes only.

Although the total number of PPIs retrieved is 180, we discovered that the gene pairs are only 59, since many different interactions can actually happen between the same pair of proteins or between the protein and itself. As an example, 43 different interactions are reported in BioGRID to happen between P00533 (epidermal growth factor receptor ) and itself.

We only report the first few lines of the seed genes interactome:

|  |  |  |  |
| --- | --- | --- | --- |
| Official Symbol  Interactor A | Official Symbol  Interactor B | UniprotAC Interactor A | UniprotAC Interactor B |
| FASN | FASN | P49327 | P49327 |
| PSEN1 | CTNNB1 | P49768 | P35222 |
| FAS | FASLG | P25445 | P48023 |
| EGFR | CTNNB1 | P00533 | P35222 |
| ATM | ATM | Q13315 | Q13315 |
| EGFR | EGFR | P00533 | P00533 |
| FASLG | FAS | P48023 | P25445 |
| … | … | … | … |

*Seed gene interactome (BioGRID, Uniprot)*

1. **Disease interactome**, composed of all interactions involving at least one seed gene

We first need to retrieve the *UniprotAC* identifier of all the non-seed genes involved in the selected interactions. We can accomplish this using the *Retrieve Id/Mapping* tool on the Uniprot website as already done for the seed genes. The mapping tool reported that 4864 out of 4886 genes were successfully mapped to 4959 UniProtKB IDs, which means that 22 proteins are not mapped in *Uniprot*.

# Enrichment analysis

Explain briefly the methods you followed to carry out the enrichment analysis and add the related tables/charts.

# Notes and comments

References (if any, this is the format to be used)

Suboc,T. (2019) Dilated Cardiomyopathy, MD, *MSD Manual*

De Paris, V., Biondi, F., Stolfo, D. et al. (2019) Dilated Cardiomyopathy: From Genetics to Clinical Management (Chapter 3). Springer

Dormand,J.R. and Prince,P.J. (1980) A family of embedded Runge–Kutta formulae. *J. Comp. Appl. Math.*, **6**, 19–26.