Data Science 10593052 Bioinformatics & Network Medicine 2020-2021

Date: 20 Dicember 2020

Network Medicine Project report

|  |
| --- |
| Network Medicine@Data Science A.Y. 2020-2021  Dilated Cardiomyopathy:  A Network Medicine Analysis  Clara Punzi1, Nicolò Petrungaro1 and Vishal Kumar Matta1  1Group no. 4 Abstract Biological systems are made up of a number of components with complex functional interdependencies which can be represented as complex networks. |

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** | **Gene name** | **UniprotAC** | **Protein name** |
| 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 built 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.

* After the first lookup, the mapping tool reported that 4864 out of 4886 genes were successfully mapped to 4959 reviewed UniProtKB IDs, which means that 22 proteins are not mapped in *Uniprot* and that some of the genes encode for many proteins, given the big difference between the number of genes (4864) and that of proteins retrieved (4959).
* To avoid duplicate records mapping the same gene to different proteins, we decided to gather the list of all the proteins encoded by one gene in the same row reducing the total of mapping to 4900 rows.
* Of the 34 missing mappings, we already knew that 22 were not mapped, whereas, after a double check on the retrieved Uniprot gene symbols, we realize that the remaining 12 genes are only mapped to unreviewed proteins, therefore we kept them unmapped.

Finally, we joined the table of the interaction data with that containing the Uniprot gene-protein mapping. As for the seed genes interactome, we obtained at the end a smaller dataset than the original one, since different interactions are reported to happen between the same pair of proteins. Thus, the final disease interactome contains a total of 212556 records, although the total number of PPIs were 271837.

Here are the first records of the disease interactome:

|  |  |  |  |
| --- | --- | --- | --- |
| **Official Symbol**  **Interactor A** | **Official Symbol**  **Interactor B** | **UniprotAC Interactor A** | **UniprotAC Interactor B** |
| ADRB1 | GIPC1 | P08588 | O14908 |
| PSEN2 | CAPN1 | P49810 | P07384 |
| CAPN3 | TTN | P20807 | Q8WZ42 |
| MAGI1 | CTNNB1 | Q6P9H4, Q96QZ7 | P35222 |
| DCN | EGFR | P07585 | P00533 |
| SUMO1 | FAS | P63165 | P25445 |
| FLNA | ITGB1 | P21333 | P05556 |
| … | … | … | … |

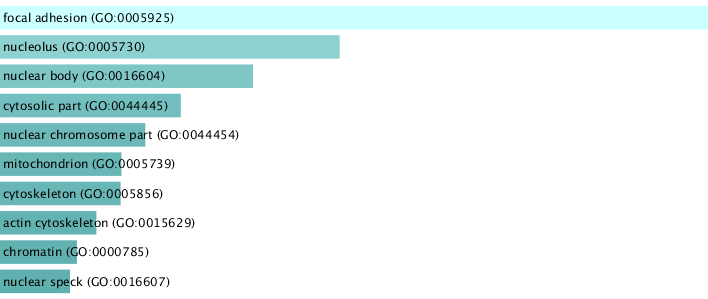
*Disease interactome (BioGRID, Uniprot)*

# Enrichment analysis

We have concluded the data collection part of the project by performing the enrichment analysis using the ***Enrichr*** webservice. Our aim was to identify the overrepresented genes in the disease interactome since they could be associated with our target disease phenotype. As references, we used the three **Gene Ontology** categories (Molecular Function, Cellular Component and Biological Process), as well as **KEGG HUMAN 2019** to detect the overrepresented pathways. The following charts show the results of our analysis, limiting to the first 10 terms for each category:

**GO Biological Process**

Immagine che contiene testo

Descrizione generata automaticamente

**GO Cellular Component**

Immagine che contiene freccia

Descrizione generata automaticamente

**GO Molecular Function**

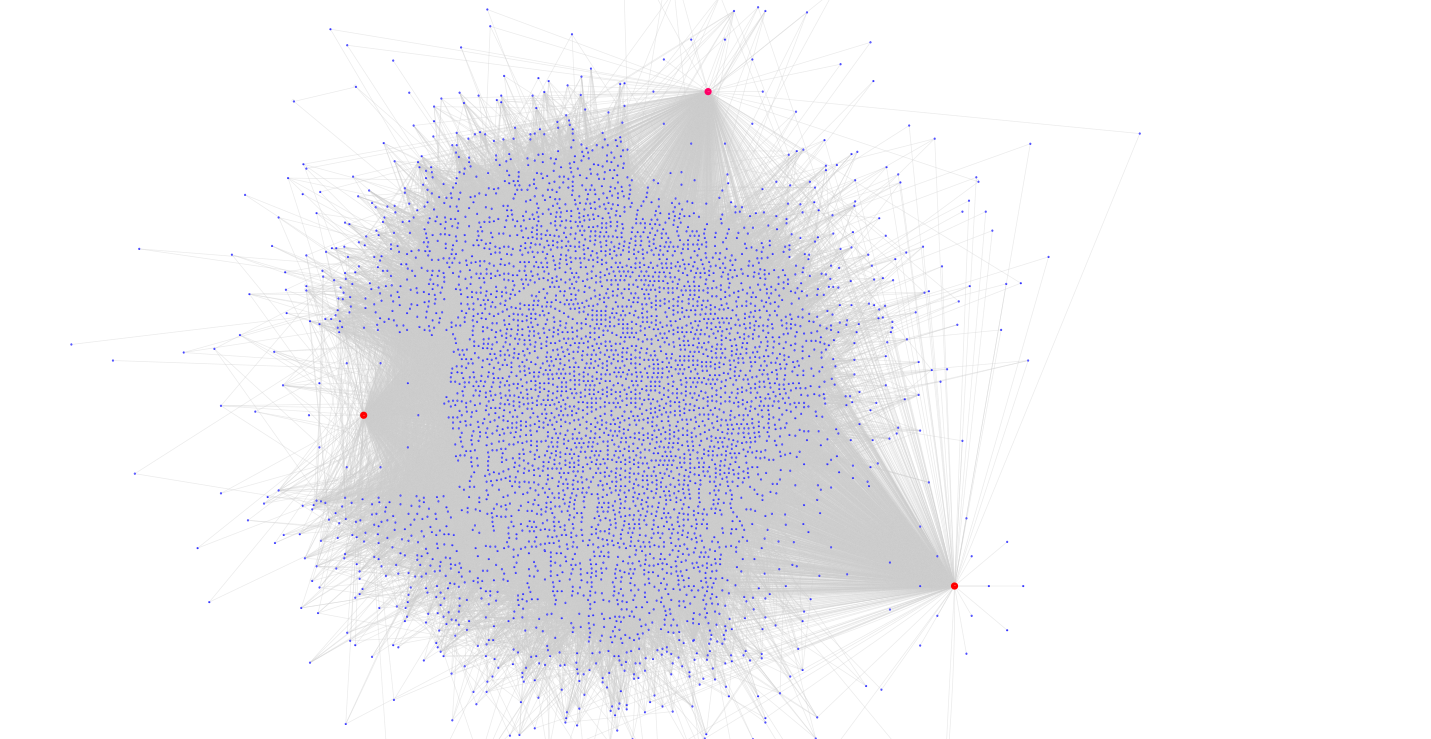
# Immagine che contiene testo, screenshot Descrizione generata automaticamente

**KEGG HUMAN 2019**

The first thing one can notice looking at the charts is that viral factors seem to play an important role in the development of DCM. Indeed, the first overrepresented pathway for our disease interactome is the viral carcinogenesis (p-value = 3.062108e-42) , which indicates a similar behavior of DCM with that of viral cancers; in the latter case the effects of the tumor viruses are, for example, the disruption of pathways that are necessary for the maintenance of the integrity of the host cellular genome and the promotion of aberrant cell-proliferation via the modulation of cell-signaling pathways to escape from cellular defense systems such as blocking apoptosis, the disruption of pathways that are necessary for the maintenance of the integrity of the host cellular genome. These effects are also coherent with the overrepresented GO Biological Processes, whose first three terms are exactly viral process (p-value = 1.032337e-46), regulation of the apoptosis (p-value = 4.125531e-43) and positive regulation of gene expression (p-value = 9.265969e-43). This also explains why, at molecular level, we found that the most overrepresents term is RNA binding (p-value = 1.417229e-118). Finally, with regards to the GO Cellular Component, the term with the highest rank is focal adhesion (p-value = 1.978364e-68), that we also found among the first positions of KEGG ranking, followed by nucleus and nuclear body. Again, this suggests us the involvement of our genes in the encoding of proteins having important regulatory functions.

Network measures

Starting from the disease interactome, we have built its representation as mathematical network in order to analyze and discover its structural and topological characteristics.



*Graph representation of the LCC (largest connected component).*

*The red nodes correspond to the genes FASN, LMNA and EFGR. We have highlighted them because their very high degree let them look like hubs for our network; indeed, all the three of them are the known to encode for proteins having relevant biological functions.*

1. First, we have calculated the number of connected components of our graph. Since we obtained that we had only one, we concluded that the largest connected component (LCC) exactly overlapped with the original network. After that, we have evaluated the main global measures of the network, which are summarize in the table below:

|  |  |
| --- | --- |
| **Global Measure** | Value |
| Nr. Nodes | 4934 |
| Nr. Links | 196283 |
| Average path length | 2.34 |
| Average degree | 79.56 |
| Average clustering coefficient | 0.19 |
| Network diameter | 5 |
| Network radius | 3 |
| Centralization | 0.33 |

*Global measures of the LCC*

1. Second, we have also calculated some local measures for each node of the network. In particular we focused on the centrality measures that will be later crucial for the definition of the putative disease modules.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gene** | **Node degree** | **Betweenness centrality** | **Eigenvector centrality** | **Closeness centrality** | **Betweenness/**  **Degree Ratio** |
| EGFR | 1319 | 0.063853 | 0.063318 | 0.570157 | 4.841e-05 |
| LMNA | 887 | 0.045209 | 0.055337 | 0.543881 | 5.096e-05 |
| PLEKHA4 | 1719 | 0.038615 | 0.116295 | 0.59881 | 2.246e-05 |
| FASN | 884 | 0.037205 | 0.050656 | 0.539716 | 4.209e-05 |
| KIAA1429 | 1406 | 0.036433 | 0.092424 | 0.580421 | 2.591e-05 |
| TRIM25 | 1084 | 0.023724 | 0.0745 | 0.552593 | 2.188e-05 |
| ESR2 | 1225 | 0.023001 | 0.075539 | 0.56474 | 1.877e-05 |
| APP | 811 | 0.020837 | 0.044435 | 0.535904 | 2.569e-05 |
| ELAVL1 | 875 | 0.019188 | 0.060298 | 0.541018 | 2.193e-05 |
| ESR1 | 1224 | 0.019108 | 0.090085 | 0.562037 | 1.561e-05 |
| NTRK1 | 1293 | 0.01906 | 0.094833 | 0.571081 | 1.474e-05 |
| MYC | 1285 | 0.018293 | 0.101819 | 0.568122 | 1.423e-05 |
| HNRNPL | 680 | 0.015494 | 0.047479 | 0.529803 | 2.278e-05 |
| KRAS | 794 | 0.013958 | 0.049679 | 0.531745 | 1.758e-05 |
| CTNNB1 | 708 | 0.012508 | 0.061617 | 0.532376 | 1.767e-05 |
| KIF14 | 1024 | 0.010911 | 0.086977 | 0.552655 | 1.065e-05 |
| HNRNPH1 | 615 | 0.009792 | 0.051401 | 0.526243 | 1.592e-05 |
| TP53 | 767 | 0.009747 | 0.064001 | 0.533297 | 1.271e-05 |
| HIST1H4A | 967 | 0.009546 | 0.085427 | 0.542625 | 9.872e-06 |
| RNF4 | 707 | 0.009513 | 0.048048 | 0.527368 | 1.345e-05 |

*20 highest ranking genes for betweenness*

# Putative disease modules

The last part of our analysis is focused on the identification of the putative disease modules, that is, the groups of nodes (i.e., cellular components) whose perturbation can be linked to dilated cardiomyopathy phenotype, this assuming the fact that all nodes belonging to the same module have a high likelihood of being involved in the same disease.

The task of identifying disease modules can be accomplished through different learning algorithms, both cluster and non-cluster based. For the purpose of our project, we have the employed two different methods: the python implementation of the Markov clustering algorithm (**MCL**) and the **DIAMOnD** tool.

1. **Markov Clustering**: this algorithm is based on the simulation of stochastic flows in the graph and rely on the assumption that the significance of disease proteins is linked to their density.

We applied the MCL algorithm on our disease interactome using the default values for the inflation hyperparameter (we tried to optimize it but the evaluation of the modularity measure was computationally too expensive for our machines). Once we have found the clusters, we have filtered them in order to keep only those being composed of at least 10 nodes, thus reducing the number of clusters from 186 to only 26. Next, we restricted our focus even more by running the hypergeometric test to keep only those modules in which the seed genes were statistically overrepresented, i.e., having p-value <0.05. The final resulting 3 sets represent the putative disease modules for DCM.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Module** | **Nr. of seed genes** | **Total nr. of genes** | **Ratio nr. seed genes/ tot nr. of genes** | **p-value** |
| 1 | 2 | 11 | 0.181818 | 0.005019 |
| 2 | 2 | 12 | 0.166666 | 0.005985 |
| 3 | 2 | 21 | 0.095238 | 0.017991 |

*Summary table of the putative disease modules found, ranked by their statistical significance*

1. **DIAMOnD**: this algorithm is not cluster-based and relies on the assumption, different from MCL, that the topological modules (i.e. pure mathematical clusters) don’t capture disease modules; instead, it is more meaningful to evaluate the significance of disease protein connections. Tu run the algorithm we first had to prepare two input files: one with the full list of all human protein-protein interactions as previously retrieved from *Biogrid* and one with our seed gene list. Fed with these two files, the algorithm determines the connectivity p-value of all genes connected to those in the seed list, rank them and add the one with highest significance to the seed gene list. At the end of 200 iterations, we have defined our new putative disease module as being composed of the resulting 200 proteins encoded by the genes selected by the *DIAMOnD* algorithm. These are the first 30 genes identified:

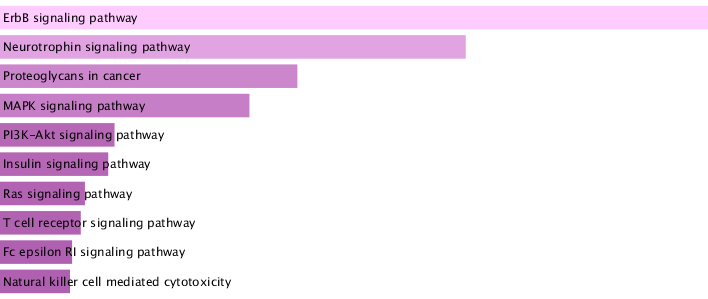
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#rank** | **Gene** |  | **#rank** | **Gene** |  | **#rank** | **Gene** |
| 1 | HSPB2 |  | 11 | YWHAZ |  | 21 | ACTB |
| 2 | SRC |  | 12 | SNCA |  | 22 | CBL |
| 3 | PIK3R1 |  | 13 | AKT1 |  | 23 | IRS1 |
| 4 | ABL1 |  | 14 | GAPDH |  | 24 | IGF1R |
| 5 | YWHAQ |  | 15 | STUB1 |  | 25 | ERBB2 |
| 6 | CFL1 |  | 16 | MAPT |  | 26 | ERRFI1 |
| 7 | ARRB1 |  | 17 | YWHAE |  | 27 | GRB2 |
| 8 | ARRB2 |  | 18 | CSNK1A1 |  | 28 | PTK2 |
| 9 | NEDD4 |  | 19 | PARK2 |  | 29 | JAK2 |
| 10 | MAP3K5 |  | 20 | TRAF6 |  | 30 | FYN |

*First 30 genes identified by the DIAMOnD tool*

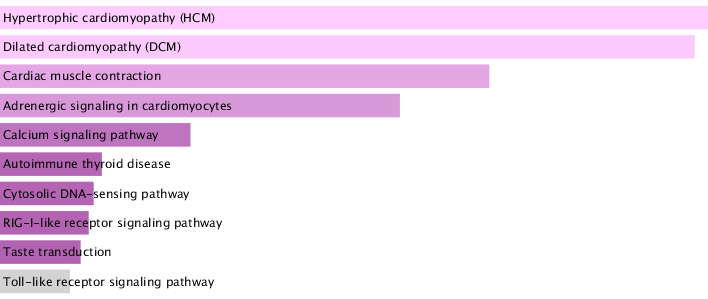
# Enrichment analysis

Finally, we performed again the enrichment analysis on all different putative disease modules in order to draw some conclusions. In the following, for each of three GO categories and KEGG pathway we report the charts corresponding to the three MCL and the one DIAMOnD putative disease modules:

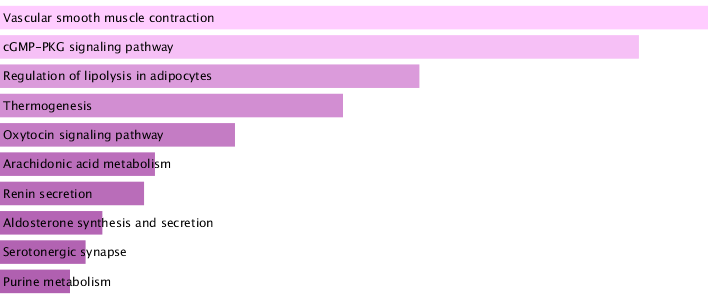
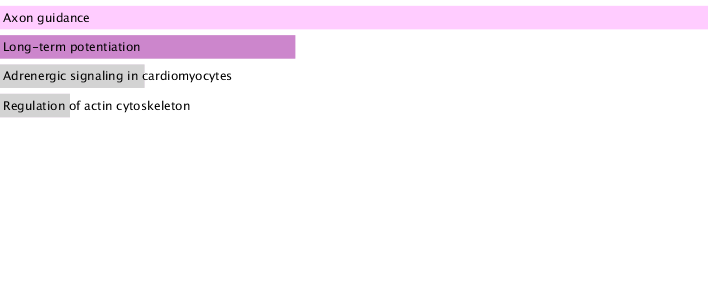
**MCL module 2**



**DIAMOnD module**



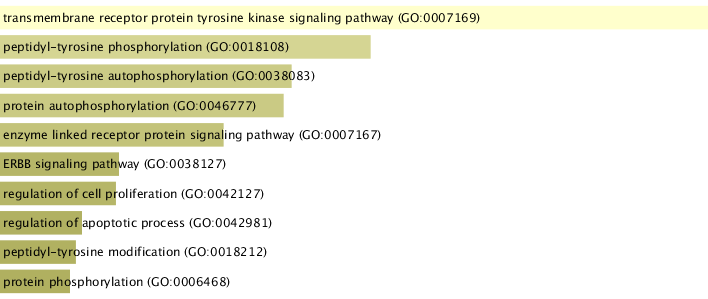
**MCL module 1**



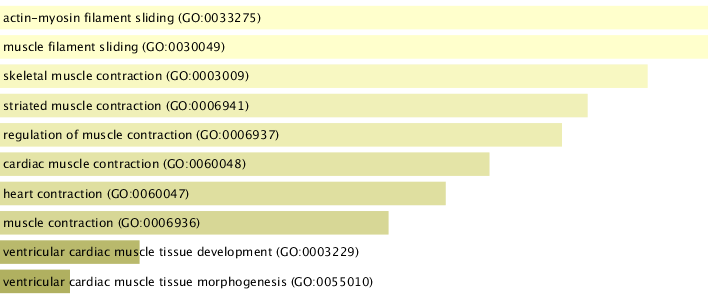
**MCL module 3**

**KEGG HUMAN 2019**

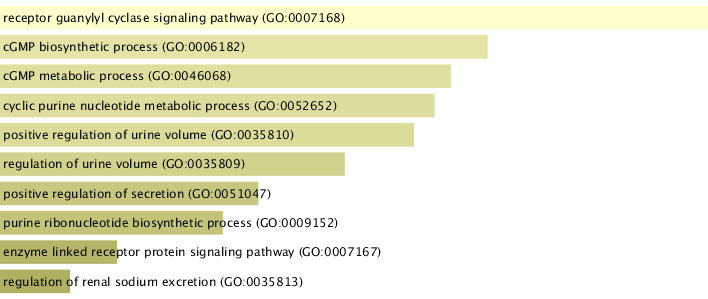
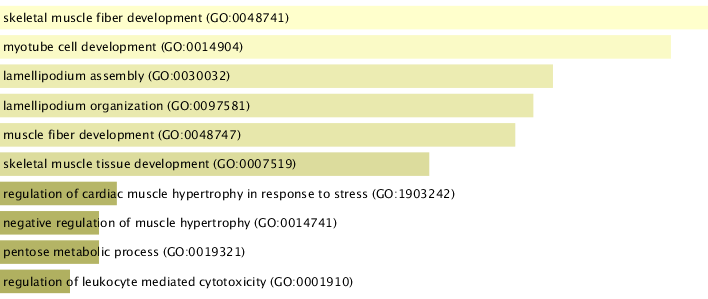
**GO BIOLOGICAL PROCESS**



**DIAMOnD module**

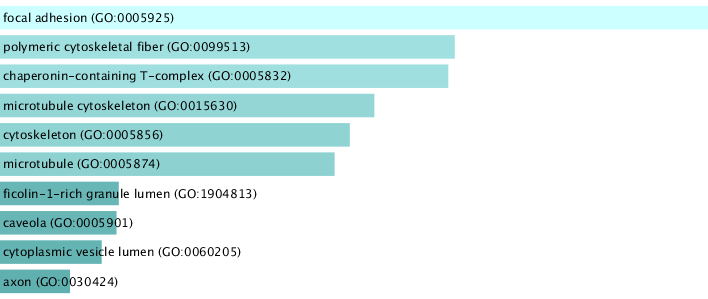


**MCL module 1**

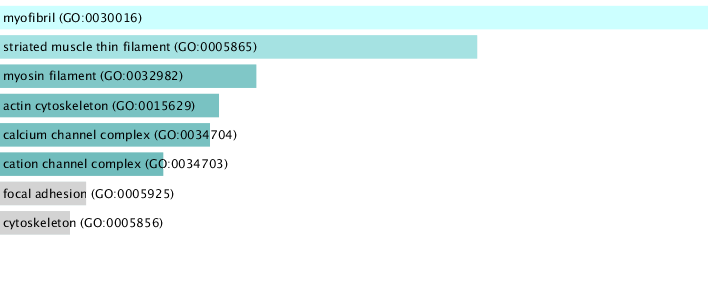


**MCL module 3**

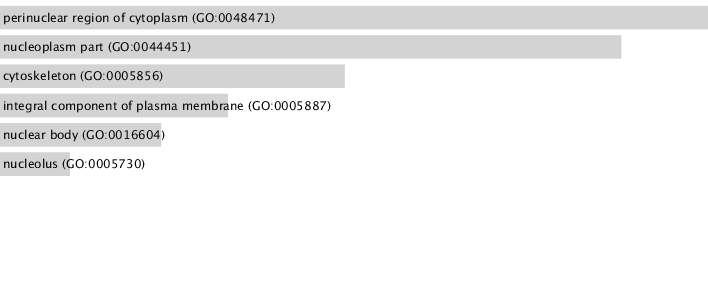
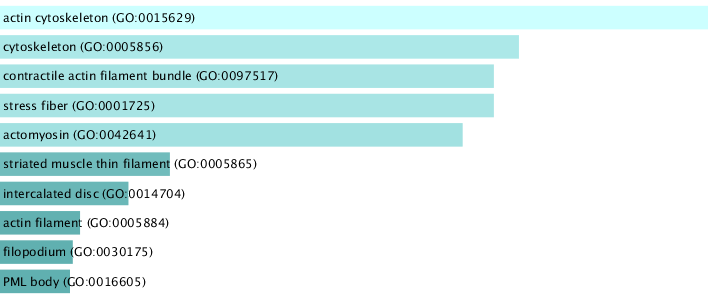
**MCL module 2**



**DIAMOnD module**



**MCL module 1**

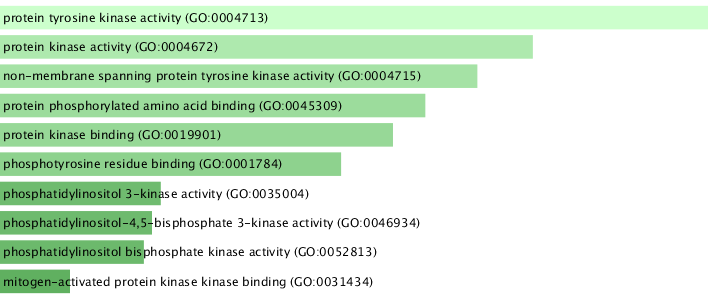


**MCL module 3**

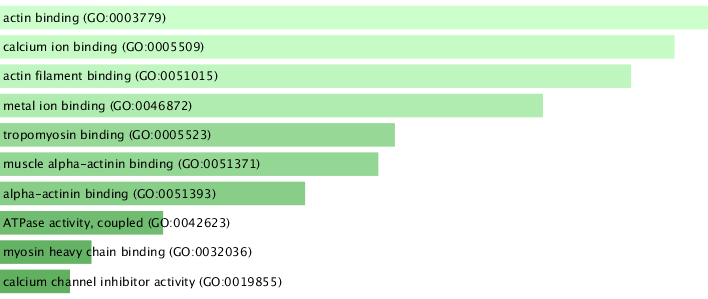
**MCL module 2**

**GO CELLULAR COMPONENT**

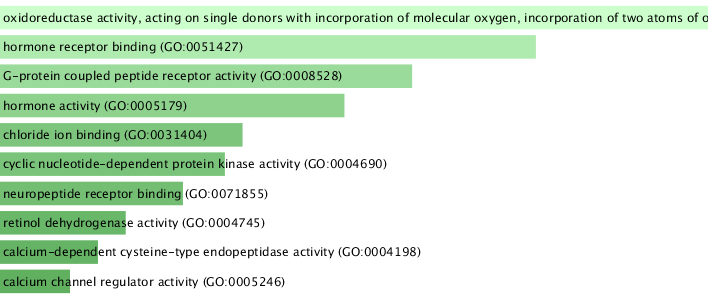
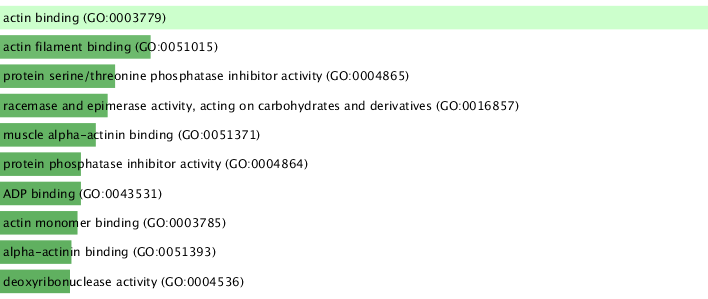
**GO MOLECULAR FUNCTION**



**DIAMOnD module**

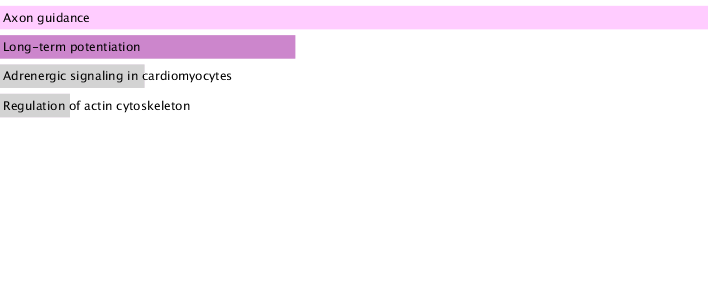


**MCL module 1**



**MCL module 3**

**MCL module 2**



# 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