Universidad Autónoma de Madrid

Escuela Politécnica Superior
Máster Universitario en
BioInformática y Biología Computacional

Final Master's Thesis

Identificación de módulos asociados a fenotipos patológicos

Author: Adrián Garcia Moreno
Thesis Director: Florencio Pazos Cabaleiro
Second Advisor: Carlos Aguirre Maeso

February - 2018

Table of contents

Acknowledgements

I want to thank all the attention received in the
National Centre of Biotechnology,
to the Systems Biology Group and especially:
my tutor Dr Pazos Cabaleiro, Florencio,
and his colleague
Dr Aguirre Maeso. Carlos.

Also to Dr Garcia Ranea, Juan Antonio
and his group,
at the Malaga University
for letting me work with their data.

Finally, thanks to all that people who
was "annoyed" and did their bit of help
during the realisation of this Master's Thesis:
Ferrer Salvador, Angel
Zarco Andujar, Virginia
Ortiz Rossini, Pablo
Ortiz Corrales, Andrés
Ruiz Rodríguez, Laura
Soriano Bermúdez, Jesús José

Abstract

Copy Number Variations (CNVs) are genomic structural variations frequently observed in healthy individuals, but can also lead to disease. They are the etiological cause of many rare genomic disorders that affect a large number of people in population, constituting a major public health problem. Unlike other small mutations, deleterious CNVs can reach millions of nucleotides containing several genes and other functional DNA regions. Many of these CNVs have yet unknown relationships to the phenotypes observed in patients. Therefore, the identification of the potentially affected molecular and genetical mechanisms in the CNVs and their relation with certain phenotypes in patients with rare deleterious disorders, nowadays, remains as a big challenge for clinical geneticists. Based on different datasets that links phenotypes, patients and genomic loci, two systemic approaches were used to understand the molecular basis that underlie those CNVs. Firstly, a functional analysis of the genes coded in these regions is carried out to realise which are the biological processes affected by the CNVs mutations thus to the phenotypes. Secondly, a network propagation analysis is done to expand the knowledge of the query genes and its interactome context. The results obtained for a cluster of patients and a number of phenotypes of clinical interest are briefly explained.

Keywords: CNVs, phenotype, ontology, network, pathologies.

1. Introduction

One of the main tasks of bioinformaticians is to develop and/or to combine existing computational methodologies in the pursuit of the understanding of the complex mutations that conform biodiversity. The mutation, as a concept, is poorly understood by society, so it has a negative connotation to those people unfamiliar with biological sciences. The term itself refers to a change that could have equally positive, negative or none effects at all. In any case they are largely determined by natural selection.

In the context of this Master's thesis, the mutations to be studied are DNA changes which seem to have unfavourable impacts in human population. These genomic alterations are of interest to the health system due to their relation with several different diseases. Although there are many types of mutations depending on how you classify them (Saitou, N. 2013)[34], this work is focused on Copy Number Variations. They are a genomic structural variation involving the repetition of DNA segments ranging from one kilobase (kb) to several megabases (Clancy, S., 2008)[10]. CNVs are defined as pathological when the number of repetitions of the pattern segment diverge from a reference genome (Redon, R., et al., 2006)[31]. This variability can be classified into deletions, if the number of copies is lower compared with the reference genome, or duplications, if the number is higher, and like any other mutation CNVs are found in organisms by inheritance or spontaneous occurrence (de novo). Two examples of diseases where CNVs have an important role are fragile X syndrome (Lozano, R. et al., 2014)[19] and Huntington's disease (Vittori, A. et al., 2014)[45].

It has been observed that the distribution of pathological CNVs length is significantly displaced towards higher values in relation to healthy population (Reyes-Palomares A. et al., 2016)[32]. Therefore, deleterious CNVs, unlike small mutations such as Single Nucleotide Polymorphisms (SNPs), can affect a larger number of genes and other functional DNA regions. Because of that, an intensified severity and complexity of the phenotypes is found on pathological CNVs, which makes still a challenge the identification of the actual genetic causes of the pathologies associated to this kind of mutation. It becomes even more difficult to set a reliable link between any specific CNV and a certain phenotype when large CNVs are usually related to a low prevalence in the population.

In order to shed some light on pathological CNVs Dr. Ranea has created an undirected tripartite network that associates three entities: phenotypes, patients and "de novo" CNVs. This network has thousands of phenotype-patient-genotype associations extracted from a dataset of patients with low prevalent genetic disorders included in the DECIPHER resource (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources - version of March 2016: approximately 17,000 patients, 24,000 CNVs divided in 13,000 deletions and 11,000 duplications -) (Firth, H.V. et. al., 2009)[11]. These CNVs genotypes are worldwide collected and identified by microarray comparative analysis and other technologies, which have been demonstrated to be efficient techniques for the discovery of new chromosomal syndromes in many cases (Ballif BC. et al., 2007; Shaw-Smith C. et al., 2006)[5, 37]. Although the majority of rare genetic syndromes have a monogenic cause, phenotypes are associated to imbalances in different functionally related loci distributed along the genome.

In view of the huge variability that all CNVs from DECIPHER might have, each CNV locus is defined as a small overlapping region (SOR) determined by the genomic regions shared among patients in the network. The phenotypes are defined according to the Human Phenotype Ontology (HPO) (Köhler et al., 2017)[35]. Since HPO is a hierarchical vocabulary, every term is expanded with its parents. Patients, who voluntarily provided informed consent, are totally anonymised and the majority is not diagnosed with a specific illness.

A schematic representation of how is created the tripartite network is in Figure 1.

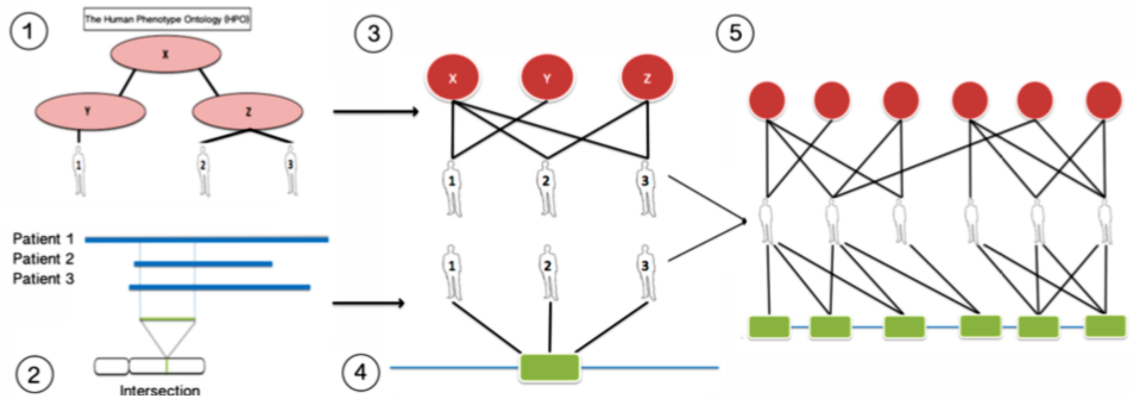


Figure 1. Generation of a tripartite network using DECIPHER patient data. Circles represent phenotypes and rectangles loci. 1) Patients are phenotypically annotated using HPO terms; 2) A locus is generated as the chromosomal region where at least two patients CNVs with the same phenotype overlap; 3) The HPOs-patients subnetwork; 4) The patients-loci subnetwork; 5) The final tripartite network.

The main purpose of the mentioned network is to find where the information converges regarding each type of edge in the network (e.g. CNV-loci to patients or phenotypes to patients). Therefore, if patients with different genomic disorders, pathological CNVs in our case, have the same phenotype we must look for common ground in their connection.

Chagoyen and Pazos showed that clinical signs, which are the phenotypes under study in this project, are reflected as modules at the molecular level to at least the same extent as diseases do (Chagoyen M. and Pazos F. 2016)[9]. The idea is that the biological processes and network modules associated to phenotypes or clusters of patients would provide a biological interpretation for the corresponding clinical signs or diseases, and eventually point to new disease markers or targets.

In this master project we approach the study CNVs-loci relation to phenotypes from a systemic perspective. Firstly, we are going to check the biological processes enriched in the set of genes contained in the affected loci. Secondly, we will look for network modules where these genes tend to map. To this extent, the interpretation, in biological terms, of the relation between the genomic regions and 1) a given phenotype or 2) a cluster of patients with similar phenotypes is not trivial due to limitations of the methods mentioned above.

2. Objectives

1. Genomic annotation of all loci associated to each phenotype and cluster of patients.
   a. Division of loci in separate *bed* format files per phenotype and cluster.
   b. Genomic annotation using *bedtools.*
   c. Genes extraction.
2. Functional annotation.
   a. Programmatic automation of Gene Ontology overrepresentation analysis.
3. Network modules identification.
4. Integration of the methodologies in a pipeline.
5. Assessment of some results based on bibliographic sources.

3. Materials and Methods

3.1. Initial Data

Our starting point is the results obtained by the research group of Dr. Ranea. Seven different datasets were obtained from the study of large-scale associations in the tripartite network described in the introduction (Perkins J., 2014; Reyes Palomares et al., 2013)[29,32]. These datasets are in the form of different collections of loci linked to phenotypes and to clusters of patients whose record of clinical signs in their clinical history has a certain degree of similarity. The datasets are divided in two groups regarding the two mentioned links of loci, four shows CNVs directly related to phenotypes and three are referred to clusters of patients.

It is important to highlight that CNVs are annotated according to their mutation origin, which can be either by deletion or duplication. Some datasets gather the CNVs of a specific origin and their associated phenotypes according to the previous division. The reason for this division is that it is known that each kind of mutation in the same genomic region of the CNV could cause different effects. For example, in the 19p13.3 and 19p13.13 syndromes, microdeletions lead to macrocephaly while microduplications cause microcephaly (Nevado et al., 2015; Dolan et al., 2010)[28, 12], whereas in the region 17p11.2, microdeletions lead to Smith-Magenis Syndrome and microduplications cause the Potocki-Lupski Syndrome. Even though these are just two examples there are many microdeletion-microduplication syndromes (Weise, A., et al., 2012)[46].

The relations between the phenotypes (HPO term) and the genomic regions are measured by a statistic-based index, the hypergeometric test (HyI). This index represents the log-transformed probability of observing an equal or greater number of shared nodes than expected by chance, and consequently measures the significance of the connections between the nodes in a directly proportional manner. Whilst the clusters of patients have been obtained using the Brainerd-Robinson coefficient, a similarity measure that was developed within archaeology, for comparing assemblages in terms of the proportions of types or other categorical data. This coefficient was calculated according to the phenotypical profiles of each patient once that the information content index of each HPO term has been obtained.

In the first group of datasets (relationships HPO-loci) we found 4 different files: 1) where CNVs were "de novo" inherited, 2) CNVs that consist only in "de novo" deletions, 3) only "de novo" duplications, and 4) both "de novo" deletions and duplications. In a general oversight of the previous datasets altogether we find that 8575 CNVs loci are related to 880 different phenotypes. Some CNVs have different HyI values for each linked HPO term, which led to 1158942 associations. The distribution of the HyI scores, ranging from 0 to 18, is represented in Figure 2.

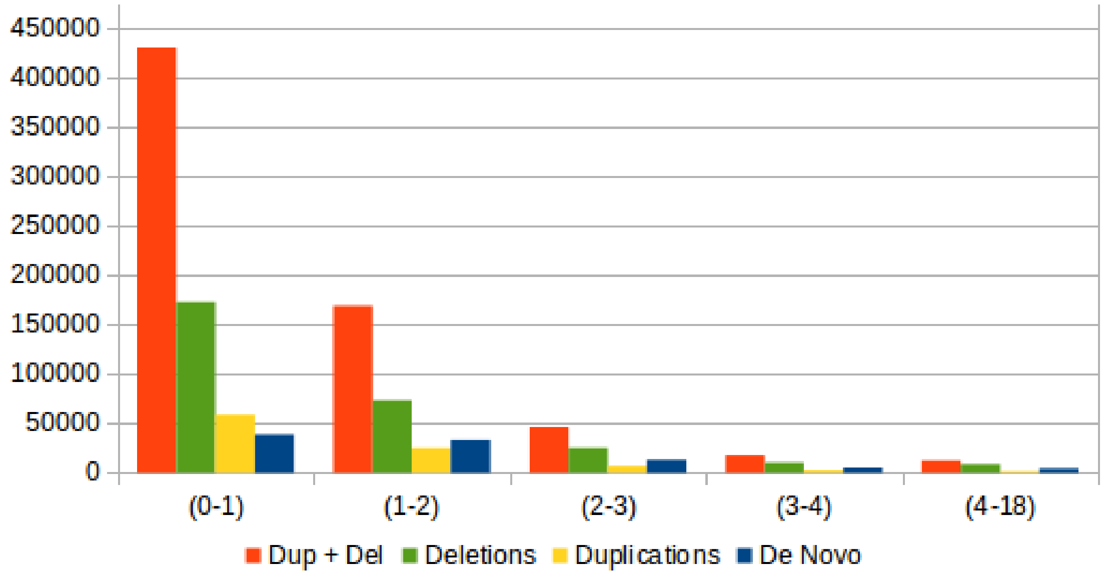


Figure 2. Representation of the number of loci within each HyI score range in every dataset.

In the second group we found 3 datasets and, like in the previous group, they classify the "de novo" CNVs by duplications, deletions and both mixed. The files contain 4230 unique CNV loci distributed in 14227 different clusters of patients with similar clinical signs, which gives back 421496 loci-cluster associations. A cluster could be understood as a syndrome or disease where loci of patients with similar phenotypic profiles are gathered. Despite the previous interpretation, it might be difficult to acknowledge that we have that many syndromes information. That is because these clusters are not independent but form a hierarchical structure, where a given cluster could be entirely or partially enclosed in a larger one.

The CNVs in these last datasets are associated, besides to their mutation origin, to a score that follows the DECIPHER's pathogenicity classification system. Since we are aware of microdeletion/microduplication syndromes, we must consider the unique loci and its duplication and deletion variants, what provide us with 4266 variants. There are 34 of these variants excluded from the following chart, Figure 3, because they have more than one different pathogenicity grade annotated. The motive of this disagreement can be explained due to the fact that researchers do not come to the same conclusions, and its reflected in the notorious percentages of unknown and uncertain consequences.

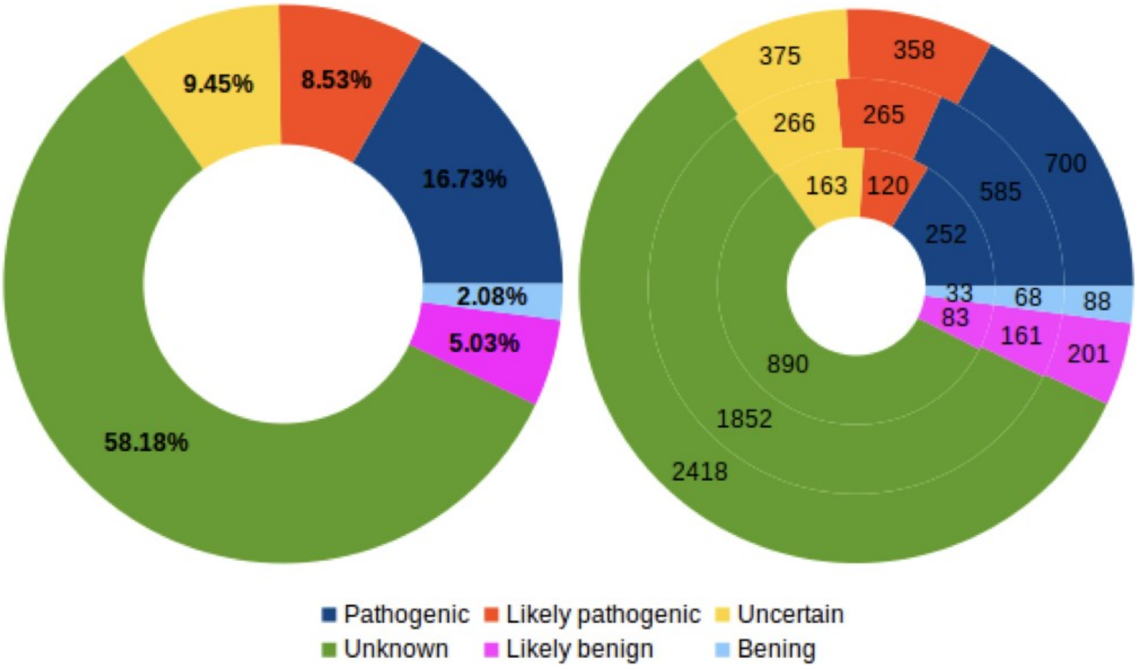


Figure 3. Distribution of CNV variants according to their pathogenicity. Left: Proportional distribution of the 4232 different variants. Around 58% of the CNVs have consequences not yet known and 9% its suspected. Right: Quantitative distribution of variants of each clusters dataset (from inside to outside: first CNVs with duplications, secondly deletions, and thirdly both).

3.2. Genomic Annotation

Each CNV could contain various genomic elements but we are interested in retrieving all the genes that overlap with our genomic regions in at least one base-pair. This can be done using the software of *Bedtools* (Quinlan AR. and Hall IM, 2010)[30] along with a proper annotation file and the datasets described in the previous point once transformed into *bed* format.

The genomic regions retrieved from DECIPHER are given in terms of the Genome Reference Consortium Human Build 37 (GRCh37), also known as hg19. We recognise that this assembly is not the most recent, GRCh38, and although there are tools to transform the loci coordinates into the current assembly, we decided not to transform our initial data to avoid possible confusions. In this scenario we used the latest hg19 annotation file (in GFF3 format) downloaded from the file transfer protocol (ftp) site of the National Centre for Biotechnology Information (NCBI). The main advantage of using the files from these sites is that they are often updated daily and with a small script it would be possible to download and use the latest annotation and the software that needs them in a single command. The downside is that the data in this files might not be enough curated for certain types of studies.

Prior to the annotation, each dataset is divided into different subsets, each containing all the loci that correspond to a specific HPO term or cluster of patients so that we can study all the phenotypes or clusters individually. Additionally, the loci associated directly to HPOs were split with different HyI scores cut-offs, ranging from 2 to 5 in steps of 0.5. We ended up with four datasets broken down in seven restrictive HyI scores and three clusters datasets.

Finally, *Bedtools* will help us to annotate everything included in the genomic regions and with the design of a small script we are able to extract the genes included in the loci of each phenotype and cluster.

3.3. Functional Annotation

There are two types of analysis when a functional annotation is performed that should be differentiated. Sometimes the query genes might have quantitative associated values, for example to indicate differential expression or just a rank used to prioritise them. In this case we will be carrying out an enrichment analysis. However, if no gene-wise scores are provided, this analysis should be called a functional overrepresentation and it is the one pertinent to our data. The overrepresentation analysis reports the biological processes where our query genes have greater presence, not being possible to differentiate the effect of genes in those processes unlike the enrichment analysis.

In this project, the Gene Ontology (GO) database is chosen as the source of terms that reflects functional units (Ashburner et al., 2000; The Gene Ontology Consortium. 2017)[4,40]. Even though, there are three categories of Gene Ontology, we used the following two: cellular component and biological process. The molecular function ontology was ruled out because it does not reflect the functional information related to biological systems we are interested in.

We need to find out which are these processes and cellular compartments for each gene list, that there are as many as different clusters and phenotypes we have annotated in the previous point. It also could be found that the same phenotype appears in different datasets, as it is shown in Figure 4, which is a reason for the huge number of genes lists.

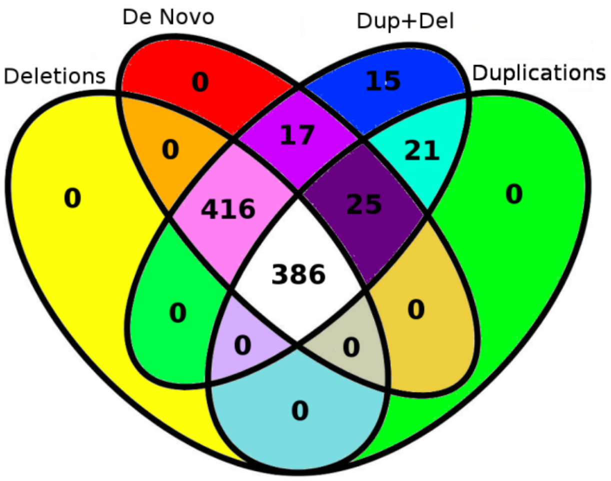


Figure 4. Venn diagram showing the phenotypes distribution across the initial datasets.

Once we take into consideration the amount of data that is being handled, the use of a non-programmatic tool to do the analysis is discarded. It was difficult to find a proper tool that is not either a web service or one based in a user interface. Nevertheless, after a comparative search It was decided to employ the topGO R package because it is constantly maintained and widely used by scientific community (Alexa, A. and Rahnenfuhrer, J. 2016)[2].

A GO overrepresentation analysis is based on gene counts. Tests like Fisher's exact test, hypergeometric test and binomial test can be used for that (Draghici, S. et al., 2006)[13]. TopGO offers the possibility to choose between these and other kind of predefined statistical tests while could also be user-defined. In our case we opted for the Fisher's exact test.

TopGO needs a gene to GO identifier map file, which can be found on some annotation databases of R packages but is preferable to create a new one. To achieve this, I used the gene to GO annotation file found in the gene data folder at the NCBI ftp site. Like the GFF3 file commented earlier, this one is re-calculated daily and, in a short R script, its download, transformation to a mapping file and subsequent use, could be included, along with the overrepresentation analysis workflow.

3.4. Network Modules

In systems biology there is a technique, widely used in other professional sectors such as law enforcement and marketing, that is growing in importance called network propagation. It is a link analysis algorithm that explores the relations of the nodes in a network, which in bioinformatics it normally represents an interactome of genes and/or proteins, but could also represent any other molecular phenomena in terms of entities and relationships. This algorithm could be implemented in two ways. The first implementation is known by several names: PageRank, the random surfer model or random walk with restart. The second is called heat diffusion and it is the one used in our last approach in the study, because this algorithm has shown computational advantages, especially for bigger networks and queries, in comparison with the others (Al-Mohy et al., 2011)[1].

Like its own name indicates the heat diffusion algorithm considers the network as a closed conductive medium in which an input heat, determined by the sum of that assigned in the query nodes, diffuses through it. The diffusion happens until a certain time is reached, if time would tend to infinite then heat will be uniformly distributed in the network. At that time, each node will have a specific value of heat associated. Those with highest values are the ones in closest relationship to the query set.

The heat diffusion algorithm that is going to be used is implemented by Carlin et al (Carlin, DE. et al., 2017)[7] and all its details can be found in their publication. Interestingly, they showed that the Spearman correlation, between adjacent time steps for various network sizes and the same for a fixed network of 10.000 nodes with various proportions of the graph included in the query, is stronger with time close to 0.1. This means that heat distribution is similar regardless the number of nodes in the network and the query set when this time is used, so it is the default value for their algorithm although it could be personalised.

This algorithm has been implemented in Cytoscape 3.6 and could be installed for previous versions searching in the apps by its name, Diffusion. But its usage is not dependent on Cytoscape, which means, its developers freed the code on Github so that it can be used programmatically outside Cytoscape. Diffusion is callable via a REST-base Application Programming interface (API) in order to offer the possibility of being used in several programming languages and via cURL. However, it is recommended to rely on Python because the network, with an initial heat amount assigned to the query nodes, must reach the service in CX json format and the main library to handle this format, Networkx, is in that language. This powerful library also offers the opportunity to use the service massively and retrieve the results according to our own criteria.

The network used for this analysis is the latest version of the human interactome in Biogrid (Stark, C. et al., 2006)[38] merged with the network generated by Menche combining different resources (Menche, J. et al., 2015)[24]. This final network has 17253 genes and 335635 edges.

3.5. Pipeline

Once all methodologies are in place, they were integrated into a single pipeline. The actual implementation let the user choose a certain threshold score (HyI in our case) to filter the loci of the initial bed file. After that, loci are grouped together in individual bed files regarding each relation of them with the associated feature (phenotype/patients cluster in the context of our data). The next step is to retrieve the genomic elements annotation using *bedtools* systematically. As we are interested on coding elements, from that annotation we take the genes in individual lists attending to the previous division. Optionally, the genes could be associated with the score of the genomic loci. These genes lists could continue in two different ways. The first one, already implemented, is the GO overrepresentation analysis whereas the second choice, which is still not integrated in the pipeline, is the network propagation analysis with heat diffusion algorithm.

The GO overrepresentation analysis uses a default p-value cut-off of 0.05 but any other value could be provided. It also uses the three ontologies (Cellular Compartment, Molecular Function and Biological Process) as default but the user could choose those of interest.

The results are then merged into a single summary file which shows the genes below their original genomic region and the different results of the functional overrepresentation in individual tables of the chosen GO categories. A second p-value threshold could be chosen to filter again the results. Finally, this file is formatted into a html (Hypertext Markup Language) to facilitate its inspection as well as including links to additional data and websites: HPO pages, NCBI genes pages, the loci visualisation in the UCSC (University of California, Santa Cruz) genome browser in the hg19 assembly, files with the significant genes, generated during the overrepresentation analysis, and the graphic representation of the significant biological processes in the GO network by TopGO and AmiGO.

The different steps of the pipeline can be started at any point and it allows the user to give as an input a folder with the initial query files or a single file. All the scripts that conforms the workflow and the pipeline itself are handled with command-line arguments when a principal script is invoked.

The whole pipeline was made freely available at GitHub, as python and R scripts: *https://github.com/EidrianGM/FromBed2NetPipeLine*

A schematic representation of the pipeline, in the context of this Master's thesis is given in Figure 5.

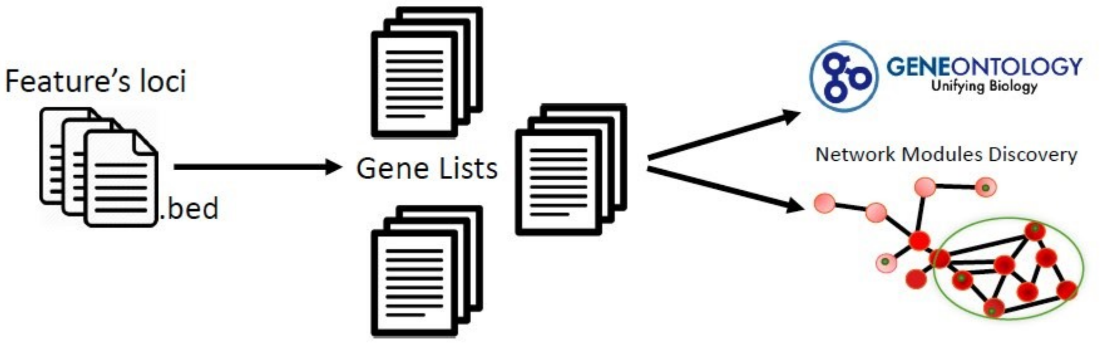


Figure 5. From any bed file that associates loci to different features, its genes are extracted individually for each one. Then a functional annotation and network propagation analysis is carried out with the genes of every feature.

4. Results and Discussion

To begin with, we must count the number of clusters and phenotypes whose most significant GO biological process is associated at a certain p-value, it should be noted that TopGO returns a maximum of 10E^{-30}. This count is done for each dataset and it is shown in Figure 6 and 7 which can be used to select a p-value cut-off separating the most interesting cluster-GO and HPO-GO associations for further study.

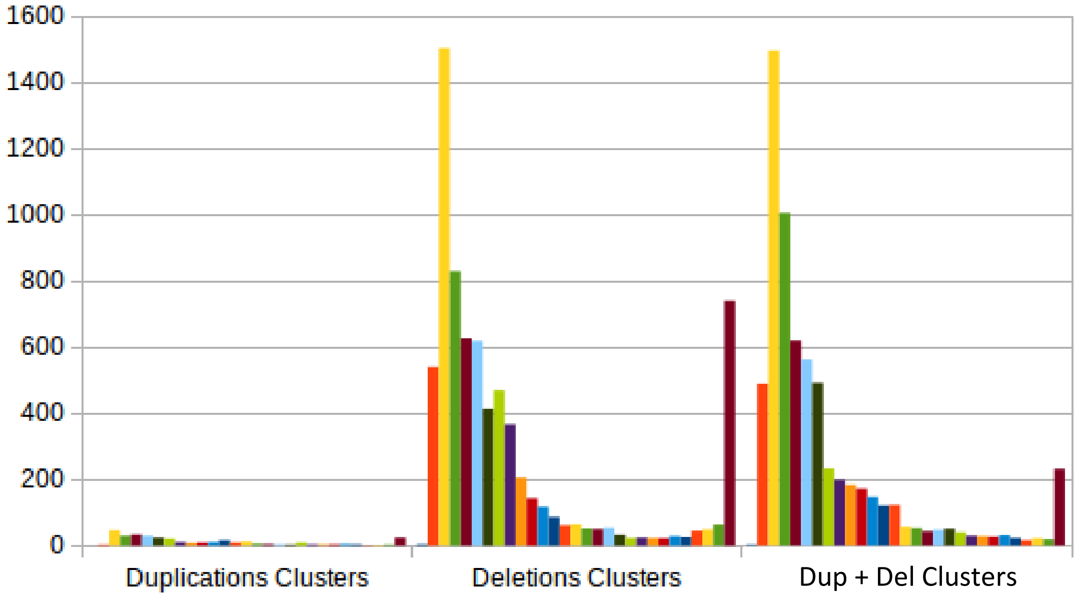


Figure 6. Number of clusters whose most significant biological process falls into a certain level of significance (*-log(p-value)*) from 2 to 30, left to right, in each dataset. Surprisingly and in contraposition with the phenotypes results we find more clusters than expected for the highest value of GO terms significance.

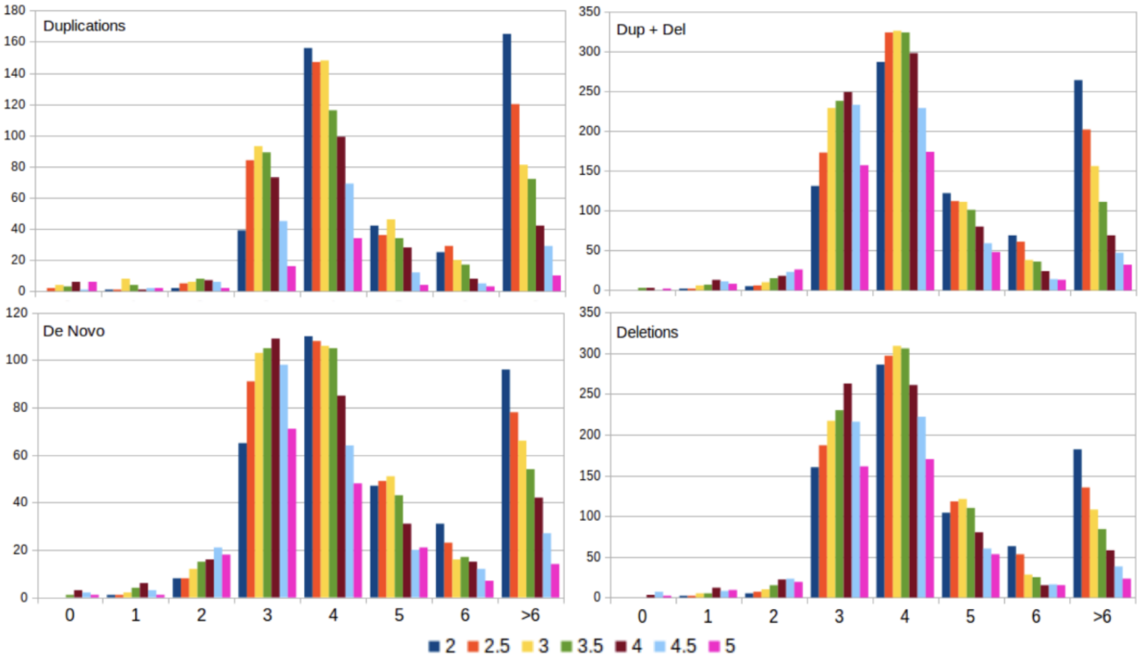
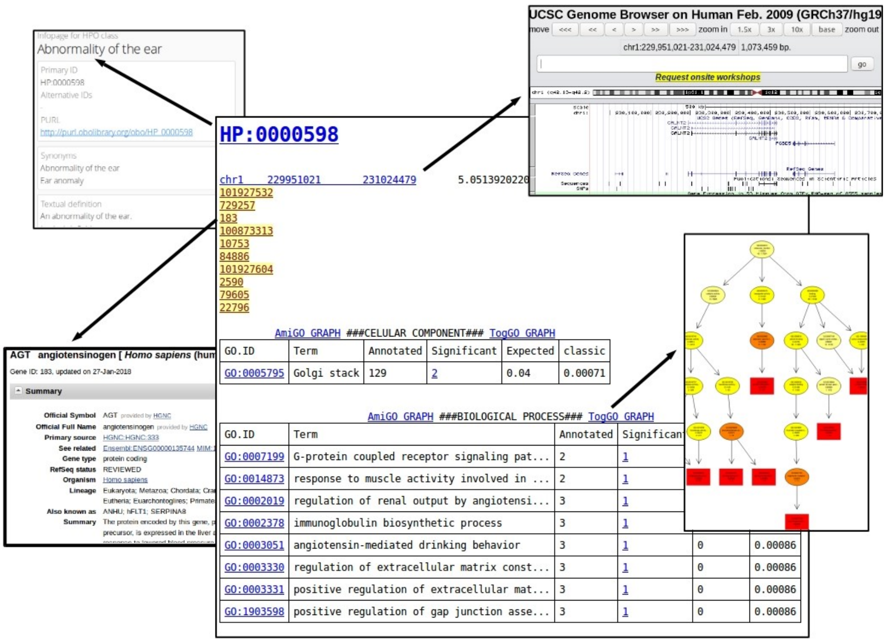


Figure 7. Number of phenotypes whose most significant biological process falls into a certain level of significance (-*log(p-value)*) from 0 to above 6 in each dataset. In the legend it is displayed the different HyI cut-offs in our initial datasets. As it is expectable the number of phenotypes decrease with higher HyI and GO terms p-values.

It seems reasonable to start studying the results from the loci with an HyI score above 4. For both, clusters and phenotypes, the minimum p-value of their most enriched biological process should be 0.001.

As it shown in Figure 8, thanks to the html summary report of the results is easier to expand the information associated with a certain phenotype thus making it easier to understand its molecular implications.


Infopage for HPO class
Abnormality of the ear
Primary ID
HP:0000598
Alternative IDs
-
PURL
http://purl.obolibrary.org/obo/HP_0000598
Synonyms
Abnormality of the ear
Ear anomaly
Textual definition
An abnormality of the ear.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19
chr1:229,951,021-231,024,479 1,073,459 bp.
Request onsite workshops

HP:0000598

chr1 229951021 231024479 5.0513920220
101927532
729257
183
100873313
10753
84886
101927604
2590
79605
22796

AmiGO GRAPH ###CELULAR COMPONENT### TogGO GRAPH

GO.ID	Term	Annotated	Significant	Expected	classic
GO:0005795	Golgi stack	129	2	0.04	0.00071

AmiGO GRAPH ###BIOLOGICAL PROCESS### TogGO GRAPH

GO.ID	Term	Annotated	Significan		
GO:0007199	G-protein coupled receptor signaling pat...	2	1		
GO:0014873	response to muscle activity involved in ...	2	1		
GO:0002019	regulation of renal output by angiotensi...	3	1		
GO:0002378	immunoglobulin biosynthetic process	3	1		
GO:0003051	angiotensin-mediated drinking behavior	3	1	0	0.00086
GO:0003330	regulation of extracellular matrix const...	3	1	0	0.00086
GO:0003331	positive regulation of extracellular mat...	3	1	0	0.00086
GO:1903598	positive regulation of gap junction asse...	3	1	0	0.00086

AGT angiotensinogen [*Homo sapiens* (hum
Gene ID: 183, updated on 27-Jan-2018
Summary

Official Symbol	AGT provided by HGNC
Official Full Name	angiotensinogen provided by HGNC
Primary source	HGNC:HGNC:333
See related	Ensembl:ENSG00000135744 MIM:1
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Cra Eutheria; Euarchontoglires; Primate
Also known as	ANHU; hFLT1; SERPINA8
Summary	The protein encoded by this gene, p precursor, is expressed in the liver a

Figure 8. Screenshot of summary of results in html format.

Some of the results are really interesting because they inform about biological processes that are already known to be related with certain phenotypes, thus allowing us to suppose that these CNVs might be a cause. This can be seen in the following phenotypes, *microcephaly* (HP:0000252), *small nail* (HP:0001792), *cerebellar hypoplasia* (HP:0001321), *general myoclonic seizures* (HP:0002123), *proportionate short stature* (HP:0003508), *hyperactivity* (HP:0000752). The first three mentioned clinical signs, have CNVs loci with HyI score above 5, conversely, the HyI score of the other phenotypes is lower than 5 but greater than 4. In the following paragraphs some aspects of these results are addressed.

4.1. Microcephaly

Microcephaly is the condition where the head is significantly smaller than the average. In the duplications CNVs, results showed that only GO processes related with box H/ACA snoRNA are reported as significant thanks to the presence of DKC1. This gene is related to dyskeratosis congenita, a multisystem disorder caused by defective telomere maintenance. Some of the patients suffering this disorder, described in Marrone et al. had microcephaly associated (Marrone, A, et al., 2007)[22]. In the deletions CNVs dataset, while lithium ion related processes where issued, in the dataset with loci with both deletions and duplications, mainly iron and copper transportation processes are reported. However, the loci associated to microcephaly, in both mentioned datasets, enclosed genes that result in enriched processes related chorion and pancreas development. CDK6, one of the genes involved in pancreas development, is within a CNV locus associated to the phenotype in question (Muhammad S. et al, 2013)[26].

4.2. Small Nail

The nail is a plate-like, keratinous, translucent structure that consists of highly specialised epithelial cells, called keratinocytes, that grows from a deep groove in the dermis of the skin. Above the layer of keratinocytes, we find the cornified envelope, composed by several layers of terminally differentiated, dead keratinocytes (Candi et al., 2016)[6]. As it could be expected, the three most significantly biological processes, reported by the loci linked to *small nail* phenotype, are keratinisation, keratinocytes differentiation, and cornification. On the one hand, several keratins of type II (KRTs 1, 2, 3, 4, 8, 73, 77, 78, 79) and type I (KRT18) among other genes are coded in the corresponding loci. On the other hand, many homeobox genes (HOXC 4, 5, 6, 9, 11, 12, 13) are also encoded in this loci and HOXC13 is known to have a direct impact in nails and hair development (Lin, Z. et al., 2012; Li, X. et al., 2017)[18,16].

4.3. Cerebellar Hypoplasia

Cerebellar hypoplasia is the term used to illustrate the underdevelopment of the cerebellum. The CNVs that are related to this phenotype have a specific locus that includes two monoamine oxidases, MAOA and MAOB. Alternatively, a larger number of loci contain the gene CASK, a calcium/calmodulin dependent serine protein kinase. The presence of these genes reported as significant biological processes those associated with neurotransmitters catabolism in general, and catecholamines in particular, specially dopamine. The association between MAOA and the phenotype has been reported by Alzghoul et al. in mice (Alzghoul, L. et al., 2012)[3]. On the other hand, all patients under study by Naijm et al. have mutations of CASK and showed cerebellar hypoplasia (Naijm et al., 2008)[27]. Moreover, we find the Norrie disease pseudoglioma gene (NDP) which was extensively studied in its relation with cerebellum development (Tokarew, N, 2017)[43].

Only the previous mentioned genes are included in our network, thus they will be the initial heat diffusion nodes. After doing the network propagation analysis we find three of the previous genes in two separated modules with the top heated nodes applying a cut-off at 0.01. The heat distribution is reflected in Figure 9.

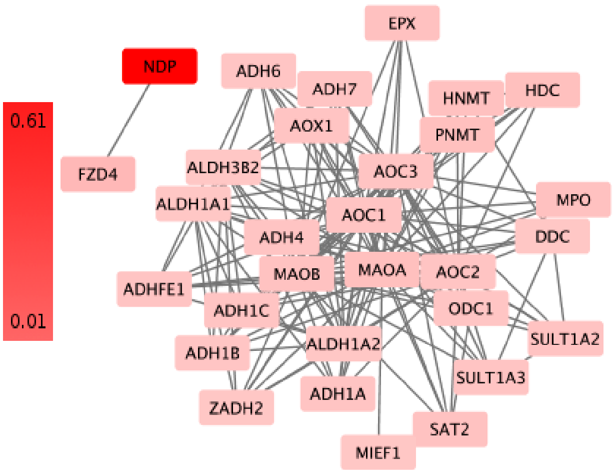


Figure 9. Heat distribution results after the diffusion from the genes within the loci associated to Cerebellar hypoplasia. Heat amount is represented as a decreasing gradient from red to light red.

NDP is the most heated node and its paired with FZD4, one of the two receptors of the product of NDP, norrin protein. Norrin/Frizzled4 signaling activates the canonical Wnt pathway to control retinal vascular development. Using genetically engineered mice, it was found that in the adult retina and cerebellum, gain or loss of Norrin/Fz4 signaling results in a cell-autonomous gain or loss, respectively, of blood retina barrier (BRB) and blood brain barrier (BBB) function, indicating an ongoing requirement for Frizzled signaling in barrier maintenance and substantial plasticity in mature CNS vascular structure (Ye, X. et al., 2010)[47].

The monoamine oxidases are in a bigger module with several alcohol and aldehyde dehydrogenases (ADHs and ALDHs), decarboxylases of ornithine, dopa and histidine, and other amine oxidases. In general, and according to KEGG functional annotation these genes are related with tyrosine, retinol, histidine, phenylalanine and cytochrome P450 drug metabolism.

4.4. General Myoclonic Seizures

General myoclonic seizures are brief, shock-like jerks of a muscle or a group of muscles that usually do not last more than a second or two but could happen in sequence within a short time. To some degree, it may occur occasionally to otherwise healthy people, for instance, hiccups may be considered a type of myoclonus. Nonetheless, this clinical sign is mainly found in epilepsy, the fourth most common neurological disorder in human population, but also in multiple sclerosis, Parkinson's disease, Alzheimer's disease, or Creutzfeldt-Jakob disease. As a result of studying the loci associated to the phenotype in question, there are three genes frequently associated in literature with epilepsy, SCN1A, SCN2A, SCN9A (Meisler et al.,2005; Catterall 2012; Zhou et al., 2018)[23, 8, 48]. The product of these genes are 3 subunits of voltage-gated sodium channels whose function is the generation and propagation of action potentials in neurones and muscle.

The only genes in the loci associated to *general myoclonic seizures* that are present in our network are the genes SCN1A, SCN2A, TTC21B and GALNT3. The modules found when applying the previous mention cut-off are shown in Figure 10.

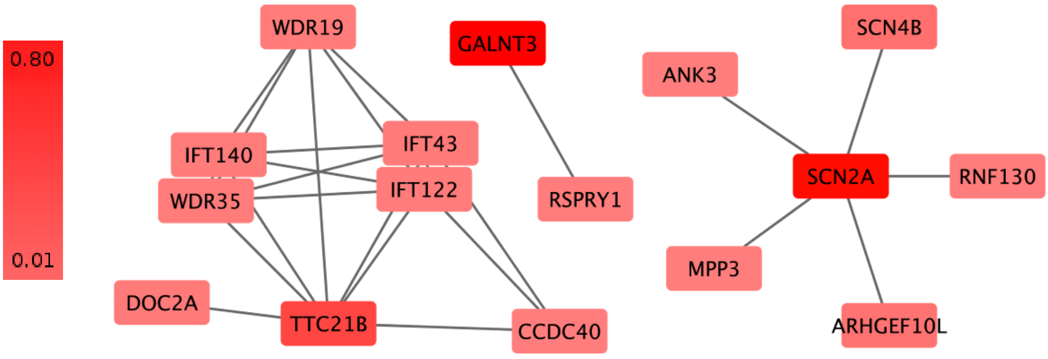


Figure 10. Network modules associated to *general myoclonic seizures.* Heat amount is represented as a decreasing gradient from red to white.

The module of SCN2A gathers another sodium channel gene, SCN4B. ANK3 is found at the axonal initial segment and nodes of Ranvier of neurons in the central and peripheral nervous systems and play a key role in activities such as cell motility, activation, proliferation, contact, and the maintenance of specialised membrane domains. ARHGEF10L belongs to the RhoGEF subfamily of RhoGTPases which are activated by specific guanine nucleotide exchange factors (GEFs) and are involved in signal transduction. The expression of RNF130 in the mouse counterpart was found to be upregulated in myeloblastic cells following IL3 deprivation, suggesting that this gene may regulate growth factor withdrawal-induced apoptosis of myeloid precursor cells. MPP3 gene product is a member of a family of membrane-associated proteins termed MAGUKs (membrane-associated guanylate kinase homologs). MAGUKs interact with the cytoskeleton and regulate cell proliferation, signaling pathways, and intracellular junctions. This protein contains a conserved sequence, called the SH3 (src homology 3) motif, found in several other proteins that associate with the cytoskeleton and are suspected to play important roles in signal transduction.

The other two most heated nodes are GALNT3 and TTC21B distributed in two separated modules. These genes were found mutated in people suffering from Dravet syndrome, an epileptic encephalopathy using array comparative genomic hybridisation (Marini et al.,2009)[21] where it was confirmed also deletion mutations in SCN1A. However these genes are relatively closed to it and its deletions has been demonstrated that sometimes affects to as many as 20 additional genes (Madia et al., 2006; Suls et al., 2006)[20,39].

4.5. Proportionate Short Stature

The next phenotype, *proportionate short stature*, has associated loci whose genes produce an enrichment in biological processes related with the embryo development. A visual summary of the results is displayed in Figure 11, with a word-cloud plot.

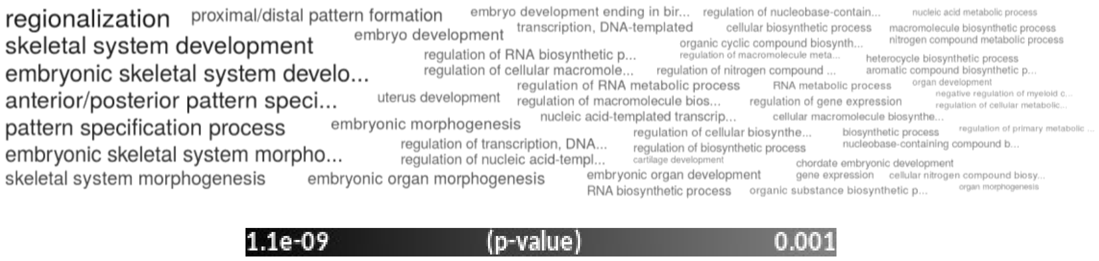


Figure 11. The significance of the biological processes associated with *proportionate short stature*.

The explanation for this functional annotation results is that one of the CNVs loci (chr7:26613851:28749226) covers the homeobox cluster A genes 1, 2, 3, 5, 6, 7, 9, 10 and 11. This homeobox genes encode DNA-binding transcription factors that may regulate gene expression, morphogenesis, and differentiation during embryonic development. Other genes that spanned in the CNVs loci of the short stature phenotype are CHSY1 and ADAMTS17, and they have been directly associated to it in several articles (Khan et al., 2012; Li et al., 2010; Morales et al., 2009; Van Duyvenvoorde et al., 2014)[15,17, 25 ,44].

4.6. Hyperactivity

Hyperactivity is normally attributed to children, but in MedLinePlus, the hyperactive person is defined as that "with increased movement, impulsive actions, and being easily distracted". When it is about children, hyperactivity is often considered more of a problem for schools and parents than it is for themselves. Notwithstanding, many hyperactive children are unhappy, or even depressed because of their different social rhythm. The loci that we found directly associated with to this phenotype contain genes of interleukin receptors and sodium channels genes. The mentioned receptors are interleukin 1 type 1 (IL1R1), 2 (IL1R2) and like 2 (IL1RL2) along with receptor 1 and its accessory protein gene for interleukin 18 (ILR18R1 and IL18RAP). For example, Segman et al. (2002) informs about a possible association between IL1R1 and Attention deficit hyperactivity disorder (Segman et al. 2002)[36.] Links between sodium (e.g. in food additives) and hyperactivity have also been reported (Kemp, A. 2008)[14].

4.7. Cluster Of Patients

Lastly, an example of a cluster of patients in the duplication dataset, whose most significant biological processes are very different between them. As the reader could see in the Figure 12, processes associated to glutathione metabolism, leukotriene biosynthesis, regulation of sprouting angiogenesis, behavior and sulfur metabolism are among the most significant in this cluster.

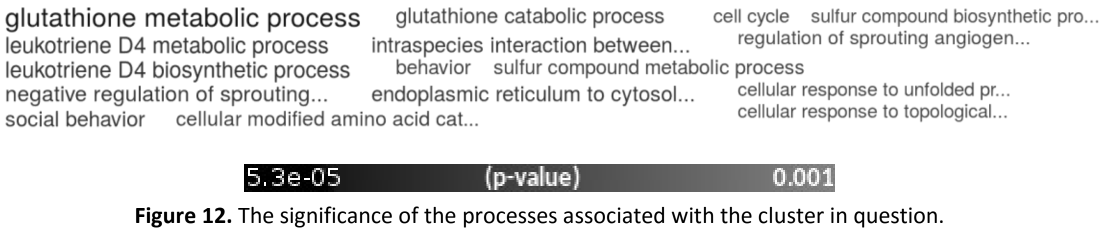


Figure 12. The significance of the processes associated with the cluster in question.

After a bibliographic search of this processes a possible syndrome related with these results is the Ethylmalonic Encephalopathy. This disorder affects several body systems, particularly the nervous system. Neurological signs and symptoms include delay and regression in development, weak muscle tone (hypotonia), seizures, and abnormal movements. The body network of blood vessels (the vascular system) is also affected. It is a mitochondrial disorder caused by genetic abnormalities of sulfide metabolism whose main trigger mutation is on gene ETHE1 (Tiranti et al., 2006; Tiranti et al., 2013)[41, 42]. ETHE1 is a sulfur dioxygenase that plays an essential role in hydrogen sulfide catabolism in the mitochondrial matrix. However, this gene is not reported in the sulfur related GO terms. Despite the similarity of the disorder signs and the biological processes enriched in this cluster it is very difficult to explain what is behind a cluster without further analysis.

4.8. Actual Knowledge

The Human Phenotype Ontology offers a curated annotation file where each phenotype is linked to specific genes. The links are generated using the information about the phenotypes of a particular syndrome and the corresponding genes that are known to cause this syndrome when mutated. The syndromes information comes from OMIM and ORPHANET. The links between genes and phenotypes are offered in two different annotation files, all genes associated or only frequently associated. To compare the gene lists derived from our phenotype-related CNVs and those annotated we used the file that gathers all the genes with known association. In the following table we find the number of phenotype gene lists that have only known genes, with some known and without known for every gene list obtained and for each CNVs-loci HyI score cut-off.

Dataset	HyI Score Cut-off	Without Known Genes	With Some Known Genes	With Only Known Genes
Deletions	**2**	225	577	0
Duplications		121	309	0
Dup + Del		250	630	0
"De Novo"		144	209	5
Deletions	**2.5**	260	538	1
Duplications		173	250	1
Dup + Del		288	590	2
"De Novo"		164	189	5
Deletions	**3**	303	494	1
Duplications		213	191	2
Dup + Del		338	537	1
"De Novo"		185	166	6

Deletions	**3.5**	333	441	1
Duplications		214	131	2
Dup + Del		389	447	5
"De Novo"		202	138	6
Deletions	**4**	352	364	6
Duplications		190	77	1
Dup + Del		402	351	6
"De Novo"		192	110	8
Deletions	**4.5**	296	297	6
Duplications		129	46	1
Dup + Del		335	278	8
"De Novo"		159	82	9
Deletions	**5**	217	238	6
Duplications		58	21	0
Dup + Del		241	224	4
"De Novo"		111	66	8

Table 1. It shows the number of phenotypes that when compared its genes, extracted from the loci of each dataset and for every HyI cut-off, with those from the actual knowledge, have all, some or none genes in common. As it can be seen the number of phenotypes with all its genes in common grows for more restrictive values of HyI until the 4.5 value, what implies that the study of the initial network weight properly the loci.

5. Conclusions

The approach used in this project is an instance of an integrating study. Firstly, because we are able to find already known genetic causes of the phenotypes in our CNVs and secondly because most of the patients under study were not diagnosed with a specific syndrome and now, in both cases, we are able to propose new candidate genes in addition to classify patients with different syndromes.

On the one hand, the functional annotation of all the genes of a certain phenotype or cluster of patients allows us to identify the biological process or processes mutated which are a probable cause of the disorder. After a bibliography search we can corroborate, with empirical evidence, that some of the genes annotated in a certain enriched process are, or are suspected to be, linked with the phenotype. The fact that we find the same associations is a suggestion that our methodology is in the good way.

On the other hand, the network propagation analysis expands the context of the genes to the molecular interactome. It is expected that the query genes of a significant process are gathered in different modules where each one is among the top heated in the diffused network. In the resulting modules, new genes with strong relation to our query could be identified whose inclusion opens a new path of further study because there is little or no direct evidence of involvement with the human phenotype. However, this methodology depends on the network used and the query genes itself. In our case, we found that the network does not always contained all the genes associated to each phenotype or cluster, or the number of query genes was too high to find evident modules.

In both cases, detecting the GO biological process or the network module associated to a given phenotype allows to 1) put that phenotype into a molecular context, and 2) find other genes (not in the original loci) that could serve as new markers or even drug targets for that phenotype. It should be taken into account that these phenotypes are associated to patients who are undiagnosed.

With regards to perform this kind of analysis in in future research works, the pipeline will be continuously improved to face any issue and add new features. However, a difficulty is derived from the network propagation analysis given that it is difficult to set an arbitrary cut-off to select a certain number of top heated nodes. A study that tries to find the relation between the number of the query nodes, its topological features and those of the network could bring in new and original results.

It has been also proposed to do a clustering analysis of the phenotypes using its related genes as categorical features. In a first attempt we used *Graphext*, a private tool that offers personalised clustering services. The gene profile of each phenotype are joined in a matrix which, due to the high variability that our datasets have, must be transformed using a Principal Component Analysis (PCA). The resulting matrix is then uploaded to *Graphext* whose algorithm calculates the distance between each pair of similar arrays, the gene profiles of each phenotype in our case. Then a network is drawn with those pair of vectors in profuse force layout. Finally, the clusters are identified as communities using Louvain Methodology. The results are shown in Figure 13.

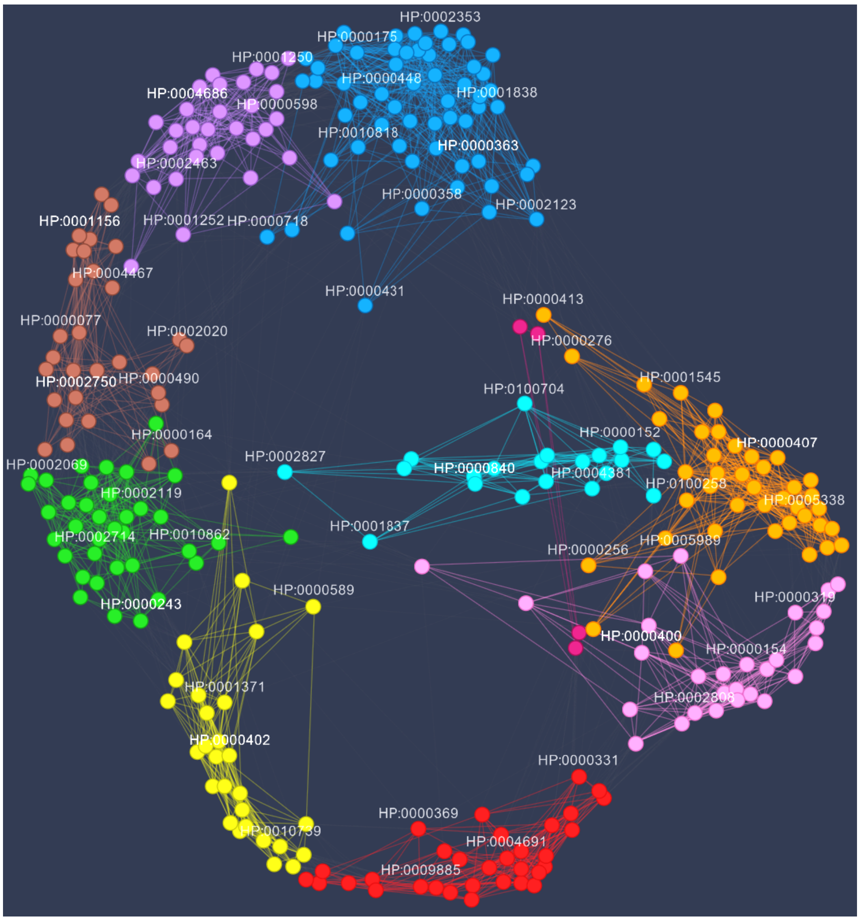


Figure 13. Clustering of genotypic profiles of each phenotype for the dataset of "de novo" with loci of HyI index above 4 associated.

Exploring this clustering analysis could offer interesting conclusions about the comorbidity of the phenotypes, because if they share a significant number of mutated genes it could be expected that the phenotypes appears in a patient altogether.

Another future prospect is try to devise a way of automatically analyse all the GO-HPO and GO-cluster associations that have been found (thousands). For this master project, only a qualitative discussion of some interesting cases was feasible. Such large scale analysis could involve text-mining of the literature.

Finally, a more ambitious future sequel of this study would involve analysing other genomic elements present in the loci, besides the coding genes. Such analysis, involving, among other things, epigenetic information, would point to other possible molecular mechanisms not directly reflected in the gene repertories.

6. Bibliography

1. Al-Mohy AH, Higham NJ. Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators. SIAM J Sci Comput. 2011; 33: 488–511.
2. Alexa A and Rahnenfuhrer J (2016). topGO: Enrichment Analysis for Gene Ontology. R package version 2.30.0.
3. Alzghoul, L., Bortolato, M., Delis, F., Thanos, P. K., Darling, R. D., Godar, S. C., … Shih, J. C. (2012). Altered Cerebellar Organization and Function in Monoamine Oxidase A Hypomorphic Mice. Neuropharmacology, 63(7), 1208–1217. http://doi.org/10.1016/j.neuropharm.2012.08.003
4. Ashburner et al. Gene ontology: tool for the unification of biology (2000) Nat Genet 25(1):25-9.
5. Ballif BC, Hornor SA, Jenkins E, Madan-Khetarpal S, Surti U, Jackson KE et al. (2007) Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. Nat.Genet; 39, 1071–1073.
6. Candi, Eleonora, Knight, Richard A, Panatta, Emanuele, Smirnov, Artem, and Melino, Gerry (2016) Cornification of the Skin: A Non-apoptotic Cell Death Mechanism. In: eLS. John Wiley & Sons Ltd, Chichester.
7. Carlin DE, Demchak B, Pratt D, Sage E, Ideker T (2017) Network propagation in the cytoscape cyberinfrastructure. PLoS. Comput. Biol. 13(10):e1005598. https://doi.org/10.1371/journal.pcbi.1005598
8. Catterall WA (2012). Sodium Channel Mutations and Epilepsy. In: Noebels JL, Avoli M, Rogawski MA, et al., editors. Jasper's Basic Mechanisms of the Epilepsies. 4th edition. Bethesda (MD): National Center for Biotechnology Information (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK98185/
9. Chagoyen M, Pazos F. (2016). Characterization of clinical signs in the human interactome. Bioinformatics. 2016 Feb 9. pii: btw054.
10. Clancy, S. (2008) Copy number variation. Nature Education 1(1):95.
11. Firth, H.V. et al (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources. Am. J. Hum.Genet 84, 524-533. DOI: dx.doi.org/10/1016/j.ajhg.2009.03.010
12. Dolan, M., Mendelsohn, N. J., Pierpont, M. E., Schimmenti, L. A., Berry, S. A., Hirsch, B. (2010) A novel microdeletion/microduplication syndrome of 19p13.13 Journal of Genetics In Medicine. 12, 503 – 511 http://dx.doi.org/10.1097/GIM.0b013e3181e59291
13. Draghici, S., Sellamuthu, S., and Khatri, P. (2006). Babel's tower revisited: a universal resource for cross-referencing across annotation databases. Bioinformatics (Oxford, England), 22:btl372v1–2939. 10.1093/bioinformatics/btl372 .
14. Kemp, A. (2008). Food additives and hyperactivity. BMJ : British Medical Journal, 336(7654), 1144. http://doi.org/10.1136/bmj.39582.375336.BE
15. Khan AO, Aldahmesh MA, Al-Ghadeer H, Mohamed JY, Alkuraya FS. (2012) Familial spherophakia with short stature caused by a novel homozygous ADAMTS17 mutation. Ophthalmic Genet., 33(4), 235-9. doi: 10.3109/13816810.2012.666708
16. Li, X., Orseth, M. L., Smith, J. M., Brehm, M. A., Agim, N. G. and Glass, D. A. (2017), A Novel Homozygous Missense Mutation in HOXC13 Leads to Autosomal Recessive Pure Hair and Nail Ectodermal Dysplasia. Pediatr Dermatol, 34: 172–175. doi:10.1111/pde.13074
17. Li, Y., Laue, K., Temtamy, S., Aglan, M., Kotan, L. D., Yigit, G., Wollnik, B. (2010). Temtamy Preaxial

Brachydactyly Syndrome Is Caused by Loss-of-Function Mutations in Chondroitin Synthase 1, a Potential Target of BMP Signaling. American Journal of Human Genetics, 87(6), 757–767. http://doi.org/10.1016/j.ajhg.2010.10.003

18. Lin, Z., Chen, Q., Shi, L., Lee, M., Giehl, K. A., Tang, Z., … Yang, Y. (2012). Loss-of-Function Mutations in HOXC13 Cause Pure Hair and Nail Ectodermal Dysplasia. American Journal of Human Genetics, 91(5), 906–911. http://doi.org/10.1016/j.ajhg.2012.08.029
19. Lozano, R., Hagerman, R. J., Duyzend, M., Budimirovic, D. B., Eichler, E. E., & Tassone, F. (2014). Genomic studies in fragile X premutation carriers. Journal of Neurodevelopmental Disorders, 6(1), 27. http://doi.org/10.1186/1866-1955-6-27
20. Madia F, Striano P, Gennaro E, Malacarne M, Paravidino R, Biancheri R, Budetta M, Cilio MR, Gaggero R, Pierluigi M, Minetti C, Zara F. (2006) Cryptic chromosome deletions involving SCN1A in severe myoclonic epilepsy of infancy. Neurology 67(7):1230–1235.
21. Marini, C., Scheffer, I. E., Nabbout, R., Mei, D., Cox, K., Dibbens, L. M., McMahon, J. M., Iona, X., Carpintero, R. S., Elia, M., Cilio, M. R., Specchio, N., Giordano, L., Striano, P., Gennaro, E., Cross, J. H., Kivity, S., Neufeld, M. Y., Afawi, Z., Andermann, E., Keene, D., Dulac, O., Zara, F., Berkovic, S. F., Guerrini, R. and Mulley, J. C. (2009), SCN1A duplications and deletions detected in Dravet syndrome: Implications for molecular diagnosis. Epilepsia, 50: 1670–1678. doi:10.1111/j.1528-1167.2009.02013.x
22. Marrone, A., Walne, A., Tamary, H., Masunari, Y., Kirwan, M., Beswick, R., Vulliamy, T., Dokal, I. (2007) Telomerase reverse-transcriptase homozygous mutations in autosomal recessive dyskeratosis congenita and Hoyeraal-Hreidarsson syndrome. Blood 110: 4198-4205.
23. Meisler, M. H., & Kearney, J. A. (2005). Sodium channel mutations in epilepsy and other neurological disorders. Journal of Clinical Investigation, 115(8), 2010–2017. http://doi.org/10.1172/JCI25466
24. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S., Vidal, M., Loscalzo, J., & Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete human interactome. Science (New York, N.Y.), 347(6224), 1257601. http://doi.org/10.1126/science.1257601
25. Morales, J., Al-Sharif, L., Khalil, D. S., Shinwari, J. M. A., Bavi, P., Al-Mahrouqi, R. A., Al Tassan, N. (2009). Homozygous Mutations in ADAMTS10 and ADAMTS17 Cause Lenticular Myopia, Ectopia Lentis, Glaucoma, Spherophakia, and Short Stature. American Journal of Human Genetics, 85(5), 558–568. http://doi.org/10.1016/j.ajhg.2009.09.011
26. Muhammad S. Hussain, Shahid M. Baig, Sascha Neumann, Vivek S. Peche, Sandra Szczepanski, Gudrun Nürnberg, Muhammad Tariq, Muhammad Jameel, Tahir N. Khan, Ambrin Fatima, Naveed A. Malik, Ilyas Ahmad, Janine Altmüller, Peter Frommolt, Holger Thiele, Wolfgang Höhne, Gökhan Yigit, Bernd Wollnik, Bernd A. Neubauer, Peter Nürnberg, Angelika A. Noegel; CDK6 associates with the centrosome during mitosis and is mutated in a large Pakistani family with primary microcephaly, Human Molecular Genetics, Volume 22, Issue 25, 20 December 2013, Pages 5199–5214, https://doi.org/10.1093/hmg/ddt374
27. Najm J, et al. Mutations of CASK cause an X-linked brain malformation phenotype with microcephaly and hypoplasia of the brainstem and cerebellum. Nat. Genet. 2008; 40:1065–1067.
28. Nevado, J., Rosenfeld, J. A., Mena, R., Palomares-Bralo, M., Vallespín, E., Ángeles Mori, M., … Lapunzina, P. (2015). PIAS4 is associated with macro/microcephaly in the novel interstitial 19p13.3 microdeletion/microduplication syndrome. European Journal of Human Genetics, 23(12), 1615–1626. http://doi.org/10.1038/ejhg.2015.51
29. Perkins J, Ayuso P, Cornejo-García JA, Ranea JA. The study of severe cutaneous drug hypersensitivity reactions from a systems biology perspective. Curr Opin Allergy Clin. Immunol. Aug;14(4):301-6. (2014).

30. Quinlan AR and Hall IM, (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.
31. Redon, R., et al. Global variation in copy number in the human genome. Nature 444, 444–454 (2006) doi:10.1038/nature053
32. Reyes-Palomares A, Bueno A, Rodríguez-López R, Medina MÁ, Sánchez-Jiménez F, Corpas M, Ranea JA (2016). Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. BMC Genomics. 16;16(1):608.
33. Reyes-Palomares A, Rodríguez-López R, Ranea JA, Sánchez-Jiménez F, Medina MA. (2013) Global analysis of the human pathophenotypic similarity gene network merges disease module components. PLoS One. 8(2):e56653.
34. Saitou, N. (2013) Chapter: Mutation. Introduction to evolutionary genomics., XXIII, 461 p. 227 illus., Hardcover ISBN: 978-1-4471-5303-0. http://www.springer.com/978-1-4471-5303-0
35. Köhler Sebastian, Vasilevsky Nicole, Engelstad Mark, Foster Erin, et al. (2017) The Human Phenotype Ontology in 2017 Nucl. Acids Res. doi: 10.1093/nar/gkw1039
36. Segman RH, Meltzer A, Gross-Tsur V, Kosov A, Frisch A, Inbar E, Darvasi A, Levy S, Goltser T, Weizman A, Galili-Weisstub E. (2002) Preferential transmission of interleukin-1 receptor antagonist alleles in attention deficit hyperactivity disorder. Mol Psychiatry, 7: 72-74.
37. Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S et al. (2006Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet; 38, 1032 – 1037.
38. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Res. Jan 1, 2006; 34:D535-9.
39. Suls A, Claeys KG, Goossens D, Harding B, Van Luijk R, Scheers S, Deprez L, Audenaert D, Van Dyck T, Beeckmans S, Smouts I, Ceulemans B, Lagae L, Buyse G, Barisic N, Misson JP, Wauters J, Del-Favero J, De Jonghe P, Claes LR. (2006) Microdeletions involving the SCN1A gene may be common in SCN1A-mutation-negative SMEI patients. Hum Mutat 27(9):914–920.
40. The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, 45(Database issue), D331–D338. http://doi.org/10.1093/nar/gkw1108
41. Tiranti V, Briem E, Lamantea E, Mineri R, Papaleo E, De Gioia L, Forlani F, Rinaldo P, Dickson P, Abu-Libdeh B, Cindro-Heberle L, Owaidha M, Jack RM, Christensen E, Burlina A, Zeviani M. ETHE1 mutations are specific to ethylmalonic encephalopathy. J Med Genet. 2006 Apr;43(4):340-6.
42. Tiranti V, Zeviani M. Altered sulfide (H_2S) metabolism in ethylmalonic encephalopathy. Cold Spring Harb Perspect Biol. 2013 Jan 1;5(1):a011437. doi: 10.1101/cshperspect.a011437.
43. Tokarew, Nicholas. (2017) The Role of Norrie Disease Pseudoglioma (Ndp) in Cerebellar Development/Tumorigenesis and its Relationship with the Sonic Hedgehog Pathway. PhD degree in Biochemistry Department of Biochemistry, Microbiology and Immunology. Faculty of Medicine. University of Ottawa
44. Van Duyvenvoorde, H. A., Lui, J. C., Kant, S. G., Oostdijk, W., Gijsbers, A. C., Hoffer, M. J., … Wit, J. M. (2014). Copy number variants in patients with short stature. European Journal of Human Genetics, 22(5), 602–609. http://doi.org/10.1038/ejhg.2013.203
45. Vittori, A., Breda, C., Repici, M., Orth, M., Roos, R. A. C., Outeiro, T. F., … the REGISTRY investigators of the European Huntington's Disease Network. (2014). Copy-number variation of the neuronal glucose transporter gene SLC2A3 and age of onset in Huntington's disease. Human Molecular Genetics, 23(12), 3129–3137. http://doi.org/10.1093/hmg/ddu022
46. Weise, A., Mrasek, K., Klein, E., Mulatinho, M., Llerena, J. C., Hardekopf, D., … Liehr, T. (2012). Microdeletion and Microduplication Syndromes. Journal of Histochemistry and Cytochemistry,

60(5), 346–358. http://doi.org/10.1369/0022155412440001

47. Ye, X., Wang, Y., & Nathans, J. (2010). The Norrin/Frizzled4 signaling pathway in retinal vascular development and disease. Trends in Molecular Medicine, 16(9), 417–425. http://doi.org/10.1016/j.molmed.2010.07.003
48. Zhou, P., He, N., Zhang, J.-W., Lin, Z.-J., Wang, J., Yan, L.-M., Meng, H., Tang, B., Li, B.-M., Liu, X.-R., Shi, Y.-W., Zhai, Q.-X., Yi, Y.-H. and Liao, W.-P. (2018), Novel mutations and phenotypes of epilepsy-associated genes in epileptic encephalopathies. Genes, Brain and Behavior, e12456. Accepted Author Manuscript. doi: 10.1111/gbb.12456