UNIVERSITÉ LIBRE DE BRUXELLES
Faculty of Sciences

Networks in developmental disorders

Charlotte NACHTEGAEL

Master thesis for the Master in Bioinformatics and Modelling

Promoters: T. Lenaerts, G. Smits and C. Olsen

Academic year 2016 - 2017

Acknowledgements

Many many thanks to all three of my promoters for their relentless research for better science and their enthusiastic involvement all through this master thesis. Thank you for helping me when I was stuck in informatics or biologic dilemmas and thank you for seeing the writing of the thesis through !

Thanks to Youssef, Julie and Yann-Äel who taught me many things and helped me understanding the data. Shout out particularly to Youssef for his work on the DDD dataset so we could all enjoy clean and useful data.

Thanks to my family for their indestructible faith in me and their incredible capacity to bear with me during this thesis. The late nights working were not lonely because of you !

Thanks to all my friends, in biomedical sciences, in bioinformatics and in informatics, for their support this whole year and their numerous offers of fleeing in Hawaii to take a break. We will still go there one day.

Contents

Chapter 1

Introduction

Since the sequencing of the first two human genomes [1, 2], the genomics field evolved in order to be more accessible to the public. While the Human Genome Project costed around $2.7 billion, it is possible nowadays for someone to have their genome sequenced for less than $1000 [3]. This price allows the daily application of genomics in diverse fields such as the clinics and the medico-legal and the exploration of the genetic cause of a plethora of disorders.

However, the huge amount of data and the complexity of the genomic data itself made the use of the *bioinformatics* necessary. Bioinformatics is a transdisciplinary field joining computer science and biology. Computer science methods help to understand and organize the information associated with the genome.

While bioinformatics, with the emergence of next-generation sequencing, unveiled genes linked to monogenic disorders, genome-wide associations studies and linkage analysis are limited by the lack of well defined control cases and insufficient sample size [4]. This approach is also limited when confronted with oligogenic and polygenic disorders with high genetic heterogeneity, underlying the need of other methods for these disorders.

The lack of consensus between patients of their disease-causing genes in more complex diseases promotes to search for a consensus on a higher level of gene annotation, such as the role of these genes in pathways, their temporal or local co-expression levels, as well the study of their protein interactors [5]. The study of biological networks helped to improve the understanding of complex diseases and their treatment [6, 7].

In this work, we will specifically focus on neurodevelopmental disorders. Neurodevelopmental disorders regroup a wide range of diseases such as intellectual disability and autism spectrum disorders [8]. The genetic etiology of neurodevelopmental disorders is largely yet unknown due to their large spectrum of phenotypes, their genetic heterogeneity and the fact that patients present an overlap of both phenotypes and genotypes [8].

To complete this objective, we will use a multidisciplinary approach to face the challenges posed in the study of neurodevelopmental disorders and their genetic factors. To shed some light on the genetic cause of these disorders and the relationship between their genotype and phenotype, we will explore the protein-protein interaction networks built from relevant genes of patients suffering from selected disorders or presenting a particular characteristic.

Because these pathologies are not well understood and because of the massive overlap between their phenotypes and genotypes, we are today unable to pinpoint the genetic cause of their phenotypes and understand the physiopathology of these phenotypes. For this reason, we will study their biological networks in a specific cohort to find convergent biological terms and specific network profiles to map to a specific disorder.

As the possibilities of the biological networks are wide, we limit ourselves to the protein-protein interaction networks as a first foray. We will use the Deciphiring Developmental Disorders (DDD) cohort from the Sanger Institute as primary material. The cohort is composed of exomic information of 1133 children and their parents. Understanding what kind of phenotypes we encounter in this cohort and the genetic information we have access to with them is primordial.

With all of this in mind, the scientific questions we would like to address are:

1. Which kind of phenotypic and genetic architectures can be found in patients suffering from neurodevelopmental disorders ?
2. Can specific disorders be discriminated based on their network and their topologies ?
3. Which pathways and biological terms are linked to the disorders and can they be mapped to specific clusters in the network ?

Chapter 2

State of the art

In this chapter, we will explore the notions needed to apprehend the multidisciplinary subject of this thesis.

- In Section 2.1, we will explain human genetic variation, its different forms, as well as the possible impact of genetic variation on genetic diseases, more particularly developmental disorders.
- In Section 2.2, we will introduce different types networks that can be studied in biology, as well as their applications.
- In Section 2.3, we will describe different existing protein-protein interaction databases.

Once these definitions and concepts are introduced, we will delve into the application of networks to developmental disorders in Section 2.4.

2.1 Biological notions

2.1.1 General genetics

In 2001, the publication of the first two human genome references allowed the comparison and discovery of the differences between individuals based on their genetic makeup, called genetic variation [1, 2]. These differences revealed an insight into the contribution of the genetic variation to the phenotype diversity [9]. The genetic variation, also called *genetic polymorphism*, is defined as the variation in a DNA sequence between distinct individuals (or chromosomes) of a given species (or population).

Genetic variants can be qualified as either common or rare according to the frequency of the minor allele in the human population. The minor allele is the less common allele of a

polymorphism. Common variants have a minor allele frequency (MAF) of at least 1%, while rare ones have a MAF of less than 1%. The frequency of any variation depends on multiple factors such as the DNA stability, the location in a coding region, the probability of repair, the consequent effect of the variation, the reproductive fitness, the human population history, the chromosomal location and recombination rates [10].

Variation is born from the mutation of the DNA. While most of the mutations found in an individual are inherited from its parents, a mutation appearing only in the individual's genome and not its parents is called a *de novo* mutation. The father seems to be instrumental in genetic variation. A child inherits a larger number of mutations from its father than its mother and the age of the father is proportional to the *de novo* mutation rate [11, 12].

Another proposed classification for the genetic variants is according to their scale. Genetic variations vary from the single nucleotide base to the whole chromosome. In this work, we will study in more detail the small and intermediate scales of genetic variation, the single nucleotide polymorphism and the structural variants respectively (Fig 1). *Structural variation* is defined as all genetic changes in a region of DNA larger than one single pair. These variations include short deletion-insertions, inversions and balanced translocations and genomic deletions, commonly referred to as copy number variants [13]. We will briefly discuss each type of variant falling into these two categories.

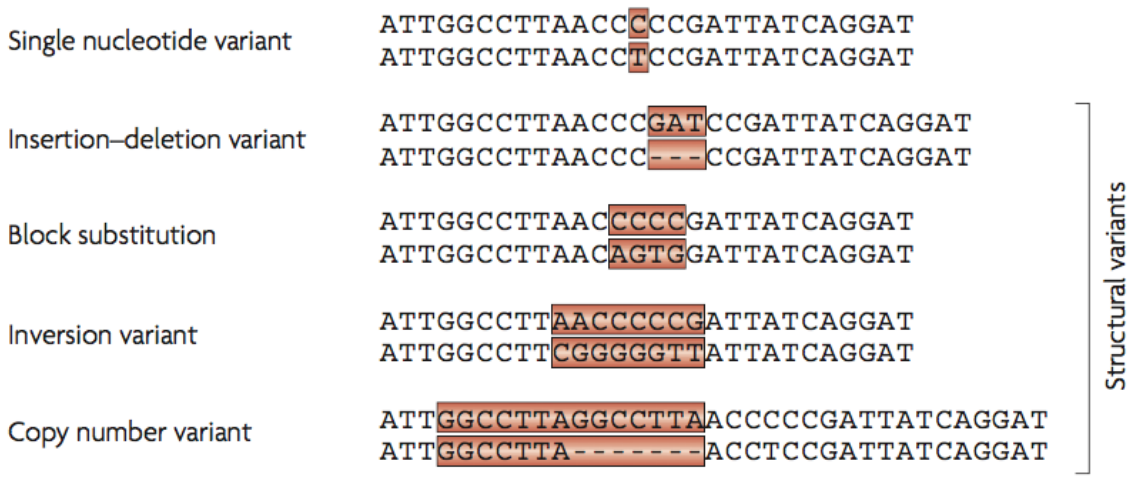


Figure 1: Classes of human genetic variants. Figure from Frazer et al. [14]

The single nucleotide polymorphism (SNP) is the substitution of one single nucleotide by another. Change of one purine base to another purine base is a transition, whereas change of a purine to a pyrimidine base, or the opposite, is a transversion. SNPs are the most common genetic variation. Two individuals' genomes differ from one another by approximately 0.1% of nucleotide sites, found all across the genome. The SNPs are commonly studied for their association to genetic diseases [15, 16, 17].

Block substitution is a string of adjacent SNPs. As with all genomic variation, the majority of this variation is found outside of coding regions. Even so, a substantial amount of block substitutions is found within coding exons and they have a strong potential to be involved in disease pathology. However, while they have a potential higher impact due to the magnitude of the change, the frequency of this type of variant is only around 1% of the SNPs [18].

Short insertions and deletions (INDELs) are the second most abundant form of human genetic variation. An INDEL occurs when one or more bases are present in one individual's genome but not in the genome of another, resulting from the insertion or the deletion of these bases in one of the genomes. Contrarily to the SNP, INDELs are uninformly distributed across the genome [19]. INDELs are investigated with high-throughput sequencing to determine their impact on human genes and their consequent impact on genetic diseases such as cancer [20].

An *inversion* is a continuous nucleotide sequence which is complementary to the original sequence and flipped, but is found at the same position. For example, the sequence AGTTCC would become GGAACT. This is the result of a rearrangement in which an internal segment of a chromosome has been broken twice, flipped 180 degrees, and rejoined [21]. Small sequence inversions are not uncommon, but have no great impact unless found in a coding sequence. Large sequence inversions occur more rarely but result occasionally in a disease. [22]. One of the best example is Hemophilia A, a disorder caused by mutations in the factor VIII gene, where 43% of the patients possess a recurrent inversion [21].

A *copy number variant* (CNV) is defined as the phenomenon where a large DNA sequence of at least 1000 bases is duplicated, the number of copies varying between individuals. The CNVs were linked to genetic disorders such as DiGeorge and Prader-Willi syndromes [23, 24]. While 4.8–9.5% of the genome contributes to the CNVs, the CNVs do not seem to have phenotypic consequences and at least 100 genes were deleted without any effect [25].

When they are situated in the coding sequence of a gene, these mutations have different consequences depending on their location and how they alter the sequence. SNPs have three different possible mutation types: the *silent* mutation which does not change the resulting amino acid compared to the normal, despite a change in the codon, the *missense* type where the change in the codon results in a different amino acid than the original and the *nonsense* type which introduces a premature stop codon. The two latter are the most dangerous types as they could lead to functional changes. CNVs can affect the density of the chromosome and could promote or inhibit the access of transcription factors to gene promoters.

In clinics, variants are characterized based on the genetic variation within the gene and the genetic variation compared to the parents (Fig 2). Individuals have two alleles of the same gene which can possess or not a genetic variation inherited or not from the parents. A variant not inherited from the parents is called *de novo*. When both alleles present the mutation, its is called *homozygous*, when only one of them does, *heterozygous*. Two main categories of variants

can be found in the autosome: autosomal dominant, single heterozygous variant with generally a strong effect, which can be inherited or de novo, and the autosomal recessive, where the combination of variants in the gene results in the effect. The latter category is divided in two classes: *compound heterozygous* variants which are different heterozygous variants found in the same gene in the child and *recessive homozygous* variants which are variants found in their heterozygous form in one or both parents and are found in the homozygous form in the child (Fig 2). The variants can be either de novo, inherited from the parents, or a combination of both.

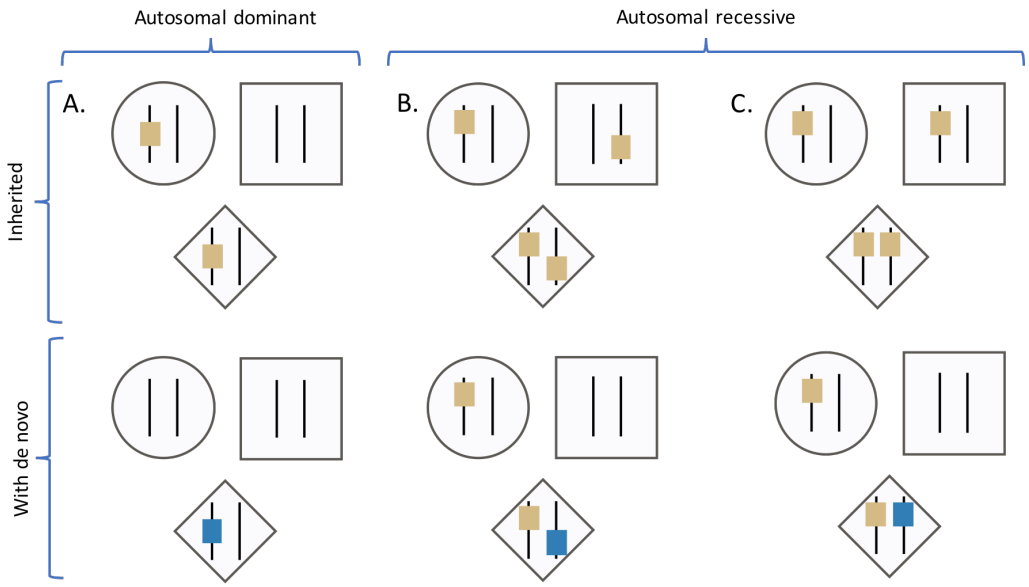


Figure 2: Schema of the clinical variants studied in this work. A, B and C correspond to autosomal dominant, compound heterozygous and recessive homozygous variants respectively. The variants in gold are inherited variants and the variants in blue are de novo variants. De novo variants are variants not found in the parents and inherited variants are variants coming from at least one of the parents.

2.1.2 Genetic diseases

Genetic diseases are caused by mutations in the genome affecting the organism's function. These mutations can be inherited or *de novo* mutations and contribute with varying degrees to the phenotype.

Monogenic to polygenic continuum

Genetic diseases can be categorized according to their genotype. *Monogenic* diseases are caused by one single defective gene. *Mendelian* diseases are a subset of monogenic diseases where the genetic etiology is hereditary. Monogenic diseases present severe phenotypes, but are rare compared to the oligogenic and polygenic diseases [5]. *Polygenic* diseases, also called *complex*

trait diseases, result from a multitude of mutations in a large number of genes. Diseases such as diabetes type 2 and schizophrenia are polygenic. However, genetic modifiers can influence the phenotype of previously considered as monogenic. From these observations, a third category of genetic diseases called *oligogenic* emerged. Oligogenic disorders have a primarily genetic cause, but require the combined action of mutant alleles in a small number of genes. One example of this phenomenon is the Fuchs corneal dystrophy, an autosomal dominant disorder born from the mutation in the TCF4 gene, where mutations in the TCF8 gene cause to severe prognosis and severe manifestation of the disease [26, 27].

Many examples indicate that a continuum exists between Mendelian and complex traits, between monogenic and polygenic diseases. The mapping of a disease on this continuum depends on three main criteria: whether one major gene is the main cause of the phenotype, the number of genes involved and the contribution of environmental factors, such as the alcohol intake or the obesity [5].

With the emergence of high-throughput sequencing and the increase of genomic data, insights into genetic diseases and their cause have evolved. When one gene is causative for multiple phenotypes, the influence of the genomic context should be taken into account[28]. Moreover, the same phenotype can be linked to different genotypes, sometimes due to large genetic heterogeneity (more than 1000 genes) [29]. Another difficulty in the discovery of the genetic causes of complex diseases is the new understanding that genetic causes do not exclusively consist of rare variants for rare diseases and common variants for common diseases, but also a combination of both [30].

Neurodevelopmental disorders

Neurodevelopmental disorders (NDDs) are a group of complex diseases in which the growth and development of the brain and/or of the central nervous system are impaired [8]). Neurodevelopmental disorders englobe intellectual disability (ID), developmental delay, autism spectrum disorder (ASD), attention-deficit hyperactivity disorder, speech and language disorders, specific learning disorders and many more. Multiple phenotypes such as neuropsychiatric problems, impaired motor function, learning and non-verbal communication are observed. These disorders occur also different degrees of severity: they may be mild and easily manageable, or they may be severe and require daily support [8].

Genetic etiology of some neurodevelopmental disorders are known, generally rare, monogenic syndromes. Examples include X fragile syndrome, Down syndrome, Rett syndrome, Prader-Willi syndrome and Angelman syndrome [31, 32, 33, 34]. The syndromes all result from one major genetic event, from the gene scale to the chromosomal one.

We will focus in this thesis on the disorders without a known genetic etiology, either because even though some cases can be explained by a monogenic form, this form did not apply to other

cases, or because non-genetic factors have to be taken into account [8].

In ASD for example, less than 10% of the cases are monogenic and are highly associated with de novo mutations [35, 36, 37]. However, the heritability of ASD has been estimated to be 85-92% based on twin studies [38, 39]. The oligogenic and polygenic genetic diagnosis of ASD is currently absent due to the difficulty to pinpoint disease-causing variants. While the first variants were identified by linkage analysis of large families with multiple affected family members [38], genome-wide associations studies brought to light only common variants [40]. This method however did not identify any loci of genome-wide significance, casting doubts on the effect of these variants, as well as to the common variant/common disease theory [30]. Deletions and duplications variants have also been linked to neurodevelopmental disorders, demonstrating mainly a risk factor effect (i.e. incomplete penetrance depending of the genomic and environmental background) [36]. The variability of severity as well as the incomplete penetrance of variants increases the difficulty to correctly determine the variants causing the disease. ASD is considered highly heterogeneous, with more than 800 genes associated with its etiology, supporting the hypothesis of oligogenic burden along the monogenic-polygenic continuum [41].

2.2 Networks

2.2.1 Network topology and properties

Networks are composed of *nodes* and *edges* connecting the nodes with each other. Nodes represents objects, such as genes or proteins, and edges represents any relationship existing between these objects. Networks can be undirected or directed if the edges have a direction associated to them. Networks can be characterized according to different measures [6]:

- The *degree* of a node is defined as the number of edges linking this node to nodes in the network. In directed networks, we can distinguish different types of degree: incoming degree, number of edges pointing towards the node, and outgoing degree, number of edges outgoing.
- The *degree distribution*, P(k), is defined as the fraction of nodes in the network with degree k. P(k) is calculated by obtaining the number of nodes with k edges and dividing this number by the total number of nodes.
- The *path length* is the distance in a network which can be measured by counting the number of edges that must be traversed to get from one node to another. The shortest path between two nodes is the minimum number of edges connecting them. The mean path length is the average shortest path between all pairs of nodes.

- The *betweenness centrality* is a measure that indicates how central a node is in a network. It is the fraction of shortest paths between all pairs of nodes in a network that go through a given node.

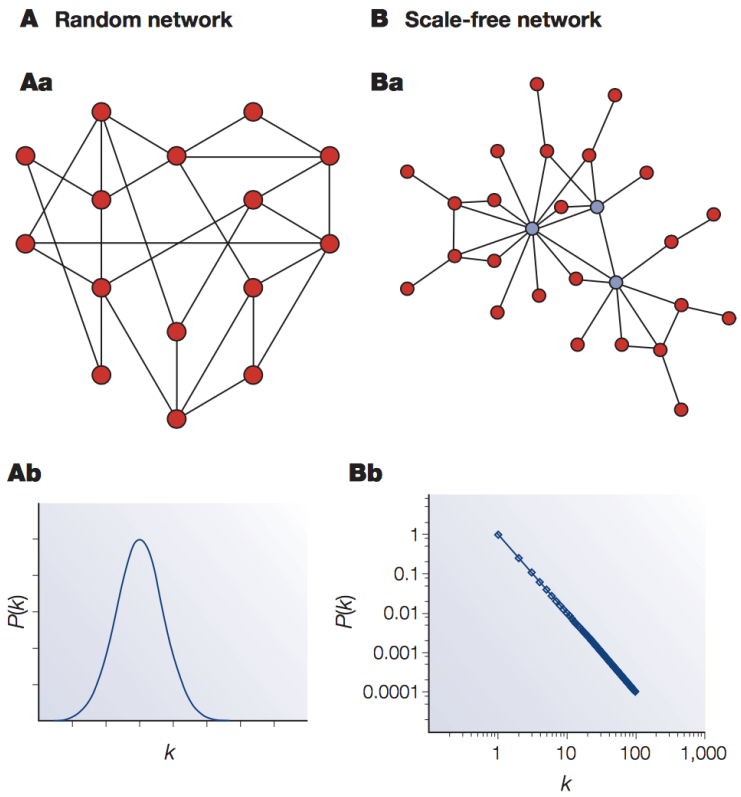


Figure 3: Types of networks and their typical degree distribution. Part of the figure from the article of Barabasi and Oltvai [6].

We will focus on two different types of network [6]:

- *Random* networks, with a degree distribution following a Poisson distribution which indicates that nodes significantly different from the average degree are rare. The mean path length is proportional to the logarithm of the network size (Fig 3 A).
- *Scale-free* networks, with a power-law degree distribution which means that the probability that a node is highly connected is more significant than in a random graph. A small number of highly connected nodes, called *hubs*, are typically found in this type of network (Fig 3 B).

2.2.2 Biological networks

We can represent biological information in the form of networks. Biological networks have been claimed to be scale-free but typically fall more in the broad-scale category [6]. Biological

networks can be divided in two categories: *interaction* networks and *association* networks. Edges of interactions networks represent a direct physical interaction between nodes, whereas the edges of association networks reflect only a link based on a shared and/or common property of the two nodes [7].

A good example of interaction networks is *protein-protein interaction* (PPI) network where the nodes represent the proteins and the edges represents a physical interaction (binding) between them. The first network resulted from two-hybrid studies in yeast [42]. *Transcriptional regulatory* networks represent both transcription factors and DNA sequences as nodes and the edges point from the regulator towards their targets. These networks are often used in diseases where transcriptional changes occur, such as cancer [43].

The *gene co-expression* network is an example of an association network. Nodes represent genes and the weights of the edges linking them represent the similarity of their expression profiles, based on the similarity of the location of their products in the tissues as well as the time they are transcribed [44, 45]. The *genetic interaction* network indicates that the phenotype of a double mutant differs from what is expected from each individual mutant, bringing new insight to oligogenic diseases [46]. Another network type is the *global disease* network which connects diseases that share at least one mutated gene, or in the opposite way it can connect genes together if they both are mutated in the same disease. This network connects genotype and phenotype in a new fashion [47].

We decided to work in this thesis with protein-protein interaction networks. Dynamics networks such as co-expression networks, while bringing relevant information, are changing depending on several factors such as the age of the individuals and the tissue. Protein-protein interaction networks allow to build a more global view of the biological processes of individuals.

2.3 Protein-protein interaction databases

Protein-protein interactions are essential for biological processes and molecular functions. They are commonly used to better understand protein function, protein complexes and provide some explanations about phenotype-genotype relationships [48, 49, 50]. Several methods exists to obtain protein-protein interaction information and can be divided in four categories: binary mapping, capturing, curation and prediction. *Binary mapping* systematically tests all pairs of every biomolecules for interaction. *Capturing* targets the collection of molecular complexes and the identification of their compononents, assumed to be protein partners. *Curation* extracts the information from the literature. Finally, the *prediction* methods base the identification of protein partners on computational methods working with the available data and knowledge about proteins.

Several databases have compiled protein interactions based on a mix of these methods. We will

briefly describe some of the major protein-protein interaction databases (Table 1).

Database	Proteins	Interactions
BioGRID (v 3.4.148)	21,008	280,910
HRPD (release 9)	30,047	41,327
IntAct & Mint (v 4.2.6)	98,329	484,110
InWEB_IM (version 2016_09_12)	17,429	612,996

Table 1: Basic statistics for the protein-protein databases in *Homo sapiens*

BioGRID

BioGrid is the result of manual curation and text-mining of experimentally discovered protein-protein interaction publications. Curators annotate each interaction according to specific criteria such as the type of the experiment. The database also recently expanded to chemical and genetic interactions [51].

HRPD

HPRD is based on experimental evidence from the literature similar to BioGRID. Additionally, it includes protein-protein interactions, as well as information about post-translational modifications, subcellular localization, protein domain architecture, tissue expression and association with human diseases. HRPD is divided in two sub-categories: binary for experiment with only two protein candidates, and complex for more than two proteins [52].

IntAct & Mint

The database includes, in addition to protein-protein interaction, a brief description of the experimental method and the literature citation of the human proteins as well as the proteins derived from several other species. The information is obtained from literature or from direct deposition of the data. The Mint database was merged with the IntAct, adding to the database experimentally verified protein interactions mainly of mammalian species [53, 54].

InWEB_IM

InWEB_IM integrates protein-protein interactions from eight different databases from both human and non-human organisms. Non-human protein-protein interactions are transferred to the humans only if the majority of these databases agree on the phylogenetic relationship between protein pairs in the model organism and humans [55].

2.4 Networks in developmental disorders

Due to the genetic heterogeneity of neurodevelopmental disorders and the not yet entirely understood environmental factors involvement, large-scale genetic studies have only begun to delve into the secrets of their genetic etiology. However, the use of biological networks points towards a convergence of biological functions [56].

While some studies integrate different types of biological networks [57, 58, 59], most of network-based studies in neurodevelopmental diseases tend to focus on only one type of network. Both protein-protein interaction networks [60, 61, 62] and co-expression networks [45, 63, 64, 65] are used for the study of the physiopathology of developmental disorders. However, these studies focus primarily on *de novo* variants at the expense of inherited ones, due to the assumption that they contribute less to the phenotype [35, 37].

The choice of the network type is very important. Protein-protein interactions are abundant, but they tend to be biased towards the most studied proteins [49]. On the other hand, the gene expression in central nervous system is highly heterogenous in location and time all along brain development and maturation, and is also less abundant. These observations lead several groups to integrate these networks in order to grasp the real association and weights to give to the genes [58, 59].

Different strategies were used to analyse the resulting networks: identification of risk-genes [58, 66], of pathways involved in the physiopathology [59, 63] and even the creation of networks to profile neurodevelopmental disorders [60, 67].

Chapter 3

Material & Methods

3.1 Datasets

3.1.1 Deciphering Developmental Disorders

We use mainly the Deciphering Developmental Disorders (DDD) dataset from the Sanger Institute [68]. The study aims to reunite exomic sequencing of children with diverse and undiagnosed developmental disorders, as well as their parents. The current dataset is composed of 1133 children from 1,101 families, with 1,071 unrelated children and 62 siblings.

Using exomic sequencing means that we do not have any information about the non-coding variants, as the exome is only composed of genes transcribed and ultimately translated into proteins. In addition to exomic sequencing, phenotypic information about the children and the parents is supplied. This phenotyping is done through Human Phenotype Ontology (HPO) terms. Each HPO term describes a clinical abnormality which can be general or more specific [69]. These terms can be used to categorize the patients in the dataset. The most frequent HPO terms are found in Table 2.

In this work, we use the variants of the DDD dataset. Variants are obtained from the exome sequences with the help of SAMtools [70], Dindel [71] and GATK [72]. The variants obtained were stored in Variant Call Format (VCF) files.

Variant annotation

These variants were originally annotated by the Sanger group with the most severe consequence predicted by Ensembl Variant Effect Predictor [73], and minor allele frequencies from a combination of the 1000 Genomes project, UK10K, the NHLBI Exome Sequencing Project, Scottish Family Health Study, UK Blood Service and unaffected DDD parents.

HPO term	Number of patients	Percentage of patients
Global developmental delay	541	47.7%
Microcephaly	179	15.8%
Delayed speech and language development	176	15.5%
Seizures	175	15.4%
Intellectual disability	127	11.2%
Specicific learning disability	120	10.6%
Autism	103	9.1%

Table 2: Most frequent HPO terms in the DDD dataset, with the number of patients with this HPO term and the percentage of patients with the HPO term in the DDD dataset. Overlaps between their patients are present, as the counts include all patients who have at least the given term.

The variants are stored into the Highlander database [74]. Highlander further annotates the variants with the help of a number of tools. One of them is dbNSFP (database for nonsynonymous SNPs' functional predictions) [75, 76]. It compiles prediction scores from different algorithms, such as SIFT, Polyphen2, FATHMM, MutationAssessor and MutPred, and other related information including allele frequencies observed in the 1000 Genomes Project phase 3 data, UK10K cohorts data, ExAC consortium data and the NHLBI Exome Sequencing Project ESP6500 data and various gene IDs from different databases. Another tool used for the annotation is SNPeff [77]. SNPeff evaluates which kind of effect the variant will have such as synonymous or non-synonymous amino acid replacement, start and stop codon gain or loss, or frame shifts. According to this classification, the impact of the effect will be scored as high, moderate, low or modifier.

The data was downloaded from the Hadoop cluster housing the Highlander installation with the help of an R script and the R packages *RMySQL* and *RJDBC*, and grouped by individual.

Filtering the data

Several criteria can be used to filter the variants.

Some criteria are based on the quality of the variant itself during the sequencing. Sometimes, multiple alleles are found at a single nucleotide position and are taken out. We also possess the information if the variant passed the filters executed after the sequencing in order to distinguish sequencing errors and use this to only keep the variants that passed all quality filters.

Other criteria select damaging variants. Damaging variants are variants that potentially explain the phenotype. The first criteria is to have some consensus of the prediction tools about the effect of the variant, so we are confident in the prediction. The second criteria is to have an

high or moderate impact of the effect on the coding region, predicted by the SNPeff tool. High or moderate effect means that the variant could be at the origin of functional changes of the coded protein or the transcription of the mRNA.

Finally, much more detailed criteria can be used to obtain specific types of variants of clinical interest. In this work we look only for the de novo autosomal dominant, the inherited autosomal recessive compound heterozygous and the inherited autosomal recessive homozygous (Fig 2). We do not study autosomal dominant inherited or autosomal recessive variants with a de novo component. While they could be of a clinical interest, their detection is difficult, so we decided to limit ourselves to variants for which we can confirm their nature.

Three subsets can consequently be distinguished in the data: the quality subset, regrouping all variants having passed the quality filters of sequencing quality and multiple allelic variants, the damaging subset, with variants predicted as damaging or with enough impact on the coding gene to possibly be damaging and for which the prediction tools reached a consensus, and the clinical subset, composed of variants of clinical interest, meaning the de novo autosomal variants, inherited autosomal recessive compound heterozygous and the inherited autosomal recessive homozygous, of good quality and with possible damaging effect (Table 3).

	Quality	Damaging	Clinical
No multiple alleles	X		X
Sequencing filters	X		X
Consensus of effect in prediction tools		X	X
High or moderate effect		X	X
Types of variant of clinical interest			X

Table 3: Filters used for each subset of the DDD dataset. Quality filters exclude multiallelic variants and variants that failed either the quality filters of the sequencing machine, or the tool used to call the variants. Damaging filters keep variants with a potential impact on the coding sequence and for whoch the prediction tools reached a consensus. The clinical subset is composed of de novo autosomal dominant, the inherited autosomal recessive compound heterozygous and the inherited autosomal recessive homozygous that oassed both the quality and the damaging filters.

3.1.2 Protein-protein interactions database

For the protein-protein interactions database, we chose the InWeb_IM database. Our choice was motivated by the abundance of interactions compared to the number of proteins (Table 1). This database regroups several of the other existing databases (BioGRID, DIP, IntAct, MatrixDB, Reactome, WikiPathways, NetPath, BIND), thusly enriching their content. The

additional metadata of the database allows the use of several other interaction databases and information for the following analysis.

Another reason for our choice is the computation of the score attributed to the interaction. The initial score of the interaction is based on the literature, notably the consensus of the interaction across publications and the type of experiment leading to the discovery of the interaction. For example, large-scale experiments weight less than small-scale. The score is then calibrated according to the network topology, for example interactions between proteins who do not share many partners are punished. Finally the score is calibrated with a gold-standard set of interactions derived from pathways, in order to transform the score into a lower bound of the probability of interaction with that initial score or higher is true [55]. This means that the score evaluates the probability that the score obtained for the protein-protein interaction is true according to the "true" protein-protein interactions found with this score or higher in the gold-standard set.

3.2 Protein-protein interactions networks

3.2.1 Construction

We wrote a R script to build the protein-protein interaction network. The program takes as input the list of variants of interest. From this list, we extract the list of distinct proteins resulting from the coding genes in which the variants are present. At this point, we have two options:

- We can build the network only with the variants. In this case, we only keep interactions between the proteins of our list. This has the advantage of keeping a reasonable number of nodes and edges in the network, however we could lose information as mutation in a variant could impact its partners.
- The other option is to build the network with the proteins, as well as the their direct first interactors. The first part consists in extracting all the interactions involving all the proteins of our list. The second part consist in extracting the interactions between all the proteins we reunited in the first phase, as well as the original proteins. While with this method we can study in a more comprehensive way the impact of the variants on the interactome, this choice also results in a huge number of nodes and edges and in a more complex network

We implemented both methods, but used only the first method in this work.

3.2.2 Visualization

Visualizing the network facilitates the interpretation and highlights interesting patterns present in the network such as highly connected sub-networks, modules or hubs.

For this purpose, we use the open software Cystoscape[1](version 3.5.1). Cytoscape provides a basic set of features for data integration, analysis, and visualization of complex networks. In order to directly download the network to Cytoscape from R, we use the *RCy3* package[2]. This package also allows to annotate the nodes of the network with the variant information. We also customized the network so the shapes represent the types of variants (recessive homozygous, compound heterozygous, de novo, a combination of them) and the colour the origin of the variant (parents in green or none in yellow).

3.2.3 Topology

In order to characterize the different networks, we use the R package *igraph*[3]. With this package, we can compute different network statistics, such as the degrees, the betweenness centrality, the detached components. Analysing these detached components is of interest as they could be representative of specific phenotypes or specific attributes of the variants.

3.2.4 Clustering

Clustering aims at creating groups of elements of a dataset that share similar attributes inside a number of separate sets (clusters). Different clustering methods exist and are separated in different classes: partitional, hierarchical and density-based. *Partitional* technique of clustering iteratively refines a set of clusters by relocating elements from one cluster to another [78]. *Hierarchical* technique result in a tree of cluster, where the root of the tree is the original set of data and the leaf nodes are the elements forming individual clusters. *Density-based* clustering forms clusters by looking at neighbouring high density regions, where the density is based on the number of elements in the neighbourhood.

In this work, we work with density-based clustering. We chose this clustering method to take advantage of the protein-protein interactions and to find potential protein complexes.

Molecular Complex Detection or MCODE is a clustering algorithm based on local density [79]. The first phase consists in computing the weight of each node based on its local network density, by defining the core-clustering of a node using the highest k-core of the node neighborhood. A k-core is a graph of minimal degree k ($\forall$ nodes n in graph G, degree(n) $\geq$ k). The highest k-core of a graph is the central most densely connected subgraph. The core clustering of a node is the

[1] `http://www.cytoscape.org/`

[2] `https://www.rdocumentation.org/packages/RCy3/versions/1.2.0`

[3] `http://igraph.org/r/`

density of the highest k-core of the immediate neighborhood of this nodes (all nodes direclty connected to it) including the node. The weight of a node is calculated by the product of its core clustering times the highest k-core of its immediate neighbourhood. Once all the weights are calculated, clusters are built by taking the node with the highest weight as a seed and recursively building the cluster by including its immediate neighbourhood where the weight is above a chosen threshold. This way, highly connected and dense clusters are built.

3.2.5 Enrichment analysis

Enrichment analysis is a powerful method to analyse large gene sets in order to identify over- or under-represented pathways. Neurodevelopmental diseases typically involve groups of genes. Multiple genes are linked to a single biological pathway, and so it likely is the combination of mutations in these gene sets that leads to the differences in the phenotype.

We used an app of Cytoscape to perform the enrichment analysis, called ClueGO [4]. ClueGO integrates the Gene Ontology (GO) terms and creates an integrative network of them, allowing the calculation of enrichment and/or depletion tests for terms and groups based on the hypergeometric distribution. GO terms annotate gene and gene products across species. The GO terms are divided in three categories: *biological process*, regrouping pathways and molecular events, *molecular function*, which contains the activities of a gene product, and *cellular component*, which describes the location where the gene product can be found. The GO terms are organized hierarchically where parent terms are more general and wide than these of children [80]. Different annotation databases can be added, such as KEGG and Reactome. KEGG is a database that regroups information on the genomic and molecular-level of high-level functions of the biological system. It also contains disease and drug information, such as perturbations to the biological system [81]. Reactome is a curated database of pathways and reactions in human biology. Information is collected by expert researchers and cross-referenced to other resources like NCBI, Ensembl, UniProt, UCSC Genome Browser, HapMap, KEGG, ChEBI, PubMed and GO, as well as inferred from orthologous organisms [82].

ClueGO computes first a binary gene-term matrix with the selected terms and their associated genes from our genes list. Based on this matrix, a term–term similarity matrix is calculated using chance corrected kappa statistics to determine the association strength between the terms. The kappa statistics is equal to $1 - \frac{1-p_o}{1-p_e}$ where p_o is the relative observed agreement among the terms and p_e is the hypothetical probability of chance agreement. The functional groups are created by iterative merging of initially defined groups based on the predefined kappa score threshold [83].

To correct the p-values of the functional groups, we decided to use the Bonferroni correction method. If you test a group of 10,000 genes, 500 might be found to be significant by chance.

[4] `http://www.ici.upmc.fr/cluego/`

Therefore, it is important to correct the p-value of each gene when performing a statistical test on a group of genes. Multiple testing correction adjusts the individual p-value for each gene to keep the overall error rate to less than or equal to the user-specified p-cutoff value. Bonferroni takes the p-value of each gene and multiply it by the number of genes in the gene list. If the corrected p-value is still below the cutoff, the gene will be significant. We retain results with a p-value lower than 0.05 to reduce the terms obtained by chance as much as possible.

ClueGO also allows to choose at which level we are looking for the terms: global, medium or specific terms. We choose to look at each level in order to study broad, general pathways, as well as specific biological processes.

Chapter 4

Results

4.1 Creation of the protein-protein interaction networks

In order to create protein-protein interaction networks, we first need to extract a list of relevant protein-coding genes. These protein-coding genes are obtained from selected variants of the patients in the DDD dataset. Exploring and understanding the architecture of the DDD dataset is important as it impacts directly on the selected proteins and therefore the protein-protein interaction network.

4.1.1 Exploration of the DDD dataset

Considering the size of the DDD dataset, we need to filter the data to extract the potentially interesting variants for each patient. The filters chosen are typically used in the clinical environment for diagnosis and create different subsets within the DDD dataset (Fig 4, Section 3.1.1).

We study more in detail the quality and the damaging subsets (Table 3) and their effects on the dataset (Fig 5). The quality subset contains 78.6% of all the variants. Among the variants eliminated due to the quality filters, 96.8% were eliminated either due to sequencing errors or to some poor results with the filters used to call the variants from the reads by GATK, 2.1% had multiple alleles and were mainly INDELs type of variants and 1.1% failed both filters. More variants are eliminated by the damaging filters, the damaging subset containing only 5.3% of all the variants. Among the eliminated variants, 4% were excluded because they had either a modifier or low impact on the coding region. Modifier variants are usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact. 6.8% of the eliminated variants of the damaging subset were excluded because they do not reach any consensus with the prediction effect tools. The rest of the eliminated variants, 89.2%, failed both the impact and the consensus filters (Fig 5). When we combine both quality

and damaging filters, resulting in 4% of all the variants, the variants are mainly excluded due to the damaging filters than the quality filters (Fig 5). The quality filters mainly eliminated the variants that did not pass the filters of the sequencing machine or of the GATK.

Figure 4: Different subsets of the DDD dataset based on the filters used in Table 3.

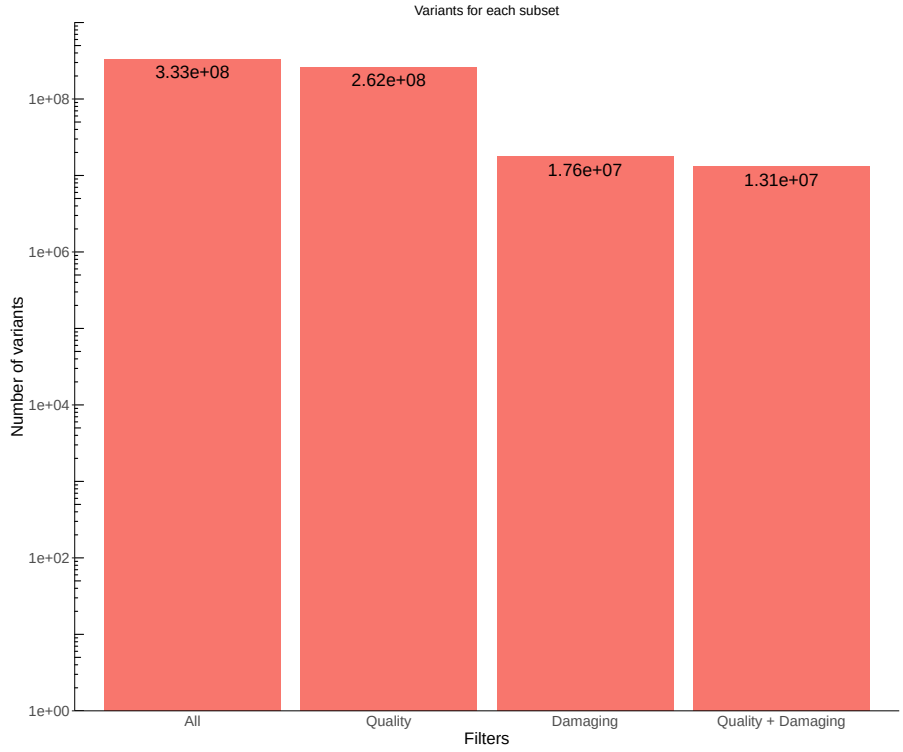


Figure 5: Number of variants of each different subset. The filters of each subset are found in Table 3. The quality subset regroups 78.6% of all the variants, the damaging subset 5.3% and the combination of both type of filters 4%. The quality filters eliminated 71,272,125 variants. The damaging filters exclude many more variants than the quality filters, namely 315,315,174 variants. The combination of both filters excludes 319,885,485 variants.

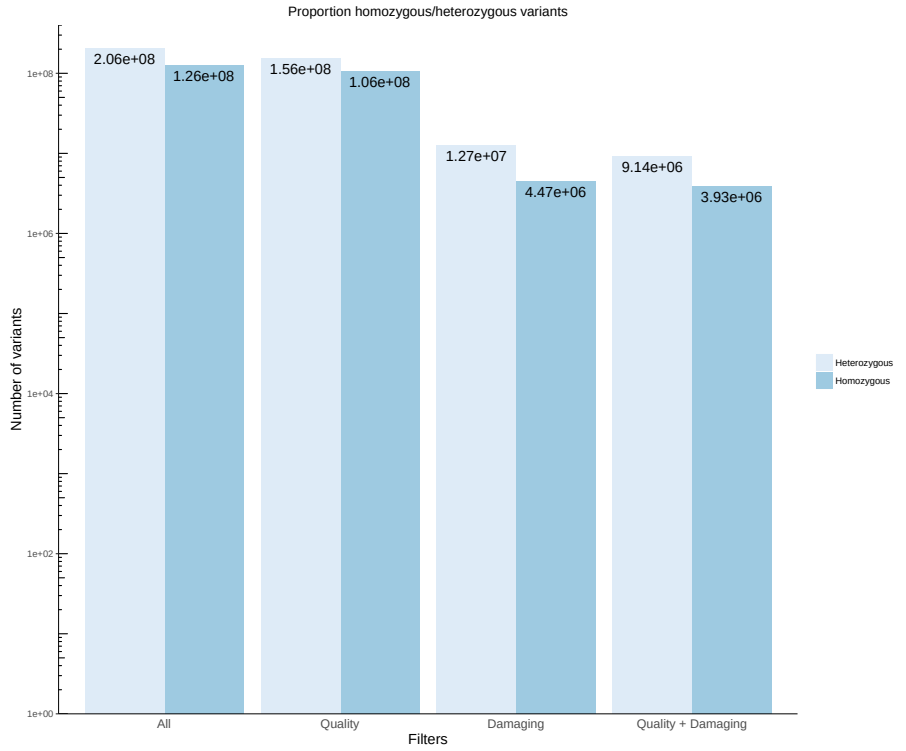


Figure 6: Number of homozygous and heterozygous variants of the different subsets of the DDD dataset. The percentage of homozygous/heterozygous are : 37.9%/62.1%, 40.5%/59.5%, 30.1%/69.9% and 26%/74% for all the variants, the quality subset, the damaging subset and the combination of both quality and damaging filters respectively.

A second information about the dataset is the proportion of homozygous and heterozygous variants (Fig 6). While we observe the conventional 2:1 ratio in favor of the heterozygous in the overall DDD dataset and the quality subset, the heterozygous percentage rises to 74% in the damaging subset and 69.9% in the subset comprising the variants passing the quality and the damaging filters. This is because there is a bias towards heterozygous damaging effect compared to the homozygous. Dominant heterozygous variants are the most damaging variant that can be found in the genome . Because of this, prediction tools for damaging effect and disease-causing effect are biased towards the heterozygous variants [84, 85].

Finally, we study the dataset according to the proportion of autosomal variants and variants present on the sex chromosomes (Fig 7). We detect no bias in the proportion of the autosomal and sex-chromosomes variants as they follow the same distribution as in the 1000 Genomes Project of around 98% of autosomal variants and 2% on the X and Y chromosomes [10].

The clinical subset of the DDD dataset that we use as a basis for the construction of the protein-protein interactions networks is based on the combination of the quality and the damaging filters. We also limit ourselves to the autosomal variants, as we do not want to bias our studies towards the variants in the XY chromosomes. The variants in XY chromosome would

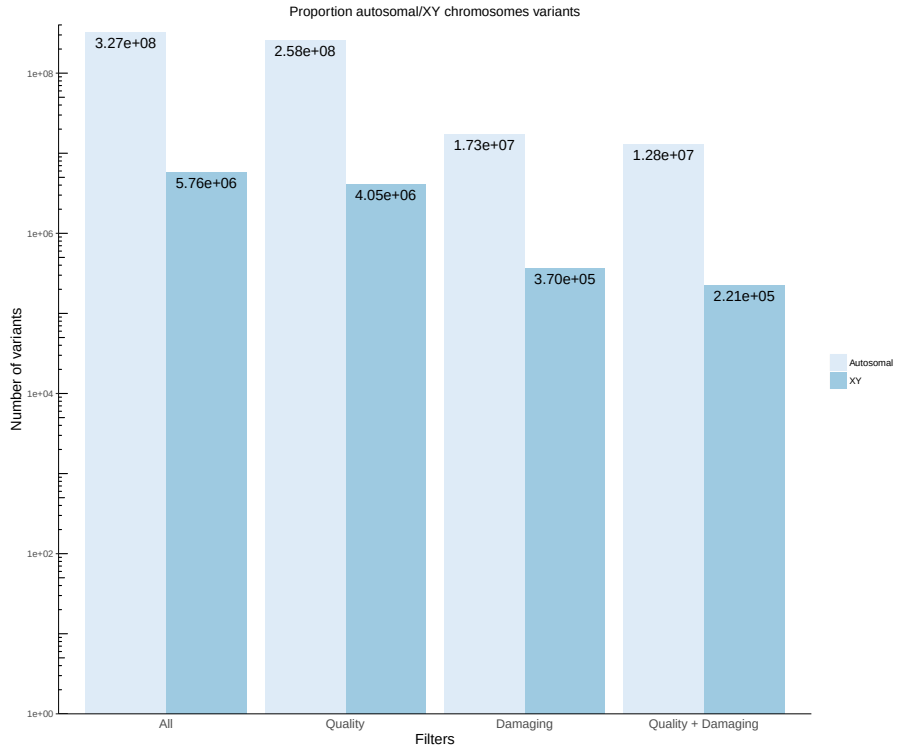


Figure 7: Number of autosomal and sex-chromosome variants in the different subsets of the DDD dataset. The proportion of autosomal and sex-chromosome variants are 97.9%/2.1%, 98.4%/1.6%, 98.3%/1.7% and 98.3%/1.7% for the all the variants, the quality subset, the damaging subset and the combination of both quality and damaging filters respectively.

heavily bias towards the male patients, as clinically relevant variants, such as the x-linked variants, can only be found in males. The explored clinical variant types are also restricted and divided in two categories: the *autosomal dominant de novo* variants and the *autosomal recessive inherited* variants. The latter type groups recessive homozygous (RH) variants and compound heterozygous (CH) variants. Only inherited variants are considered, even though by a combination between inherited and de novo variants, we could also obtain the same types of variants. By putting stringent criteria for the selection of the variants, we hope to obtain clinically pertinent genes that could bring light to mechanism of developmental diseases.

4.1.2 Cohort protein-protein interaction networks

After filtering the variants and obtaining the clinical subset of the DDD dataset, we can build the protein-protein interaction networks. We decided to build the protein-protein network for the whole cohort, meaning we use all the variants found for all the patients to obtain the list of protein for the network (See Section 3.2). However, the current clinical variant subset includes common and rare variants. We know that common variants are involved in the neurodevelopmental diseases, but their real implication is still subject to debate [40, 41], while the impact of rare variants has been proved in general more consequent [30]. Additionally, rare

variants are nowadays the variants that can be detected in clinics, making them a primary source of information for diagnosis. In light of this, we chose to work with rare variants and to exclude the most common ones. However, we need to choose at which minor allele frequency we will filter out the variants.

To decide, we create different lists of genes based on different threshold values for the minor allele frequency in order to choose the most appropriate list and protein-protein interaction network. We filter the variants based on their minor allele frequency lower than or equal to 0.0001, 0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09 and 0.1, then we apply the regular clinical filters. From these lists of variants, we extract the gene lists and build the protein-protein interaction network only with the proteins derived from the gene lists. However, we observe that with the regular filters we obtain a great number of de novo variants. A great number of these presumed de novo variants are actually present in the parents' variants. These variants were excluded from the initial list of variants due to the quality filters. In clinics, with such a situation, we would look directly into the sequence alignment map files to check the presence or the absence of the variant in the parents and determine the cause of the rejection of the variant. As these files were not made public by DDD, we can only filter out these potential false positive de novo variants based on the unfiltered parents' list of variants. This results in a list of highly constrained variants which we can use with more confidence than the list obtained with the regular filters (Fig 8).

We can note that while the number of variants grows almost linearly when we allow the frequent variants, the number of genes, as well as the number of protein-protein interactions, seems to decrease their growing rate after reaching 0.01 for the minor allele frequency. This observation is linked to the number of variants, genes and protein-protein interactions of the de novo variants which seems to reach a plateau after 0.01. This plateau is due to the fact that it is difficult to find de novo variant with a high frequency in the population, as it contradicts the nature of the de novo variant itself. The RH and CH variants seem to be the reason of the growing rate. Inherited variants are not limited by the allele frequency as the de novo variants are, as we inherit our genome from our parents so we will always find inherited variants.

With these results, we decided to choose to study 0.01 as the threshold value for the minor allele frequency filter with the highly constrained variants. While assuring that we are studying relevant clinical variants and lowering the false positive rate of the de novo variants, the threshold value of 0.01 is also commonly used in clinics to discriminate rare variants. This decision also has the advantage of reducing the size of the subset of proteins studied to a more manageable one for the subsequent analyses.

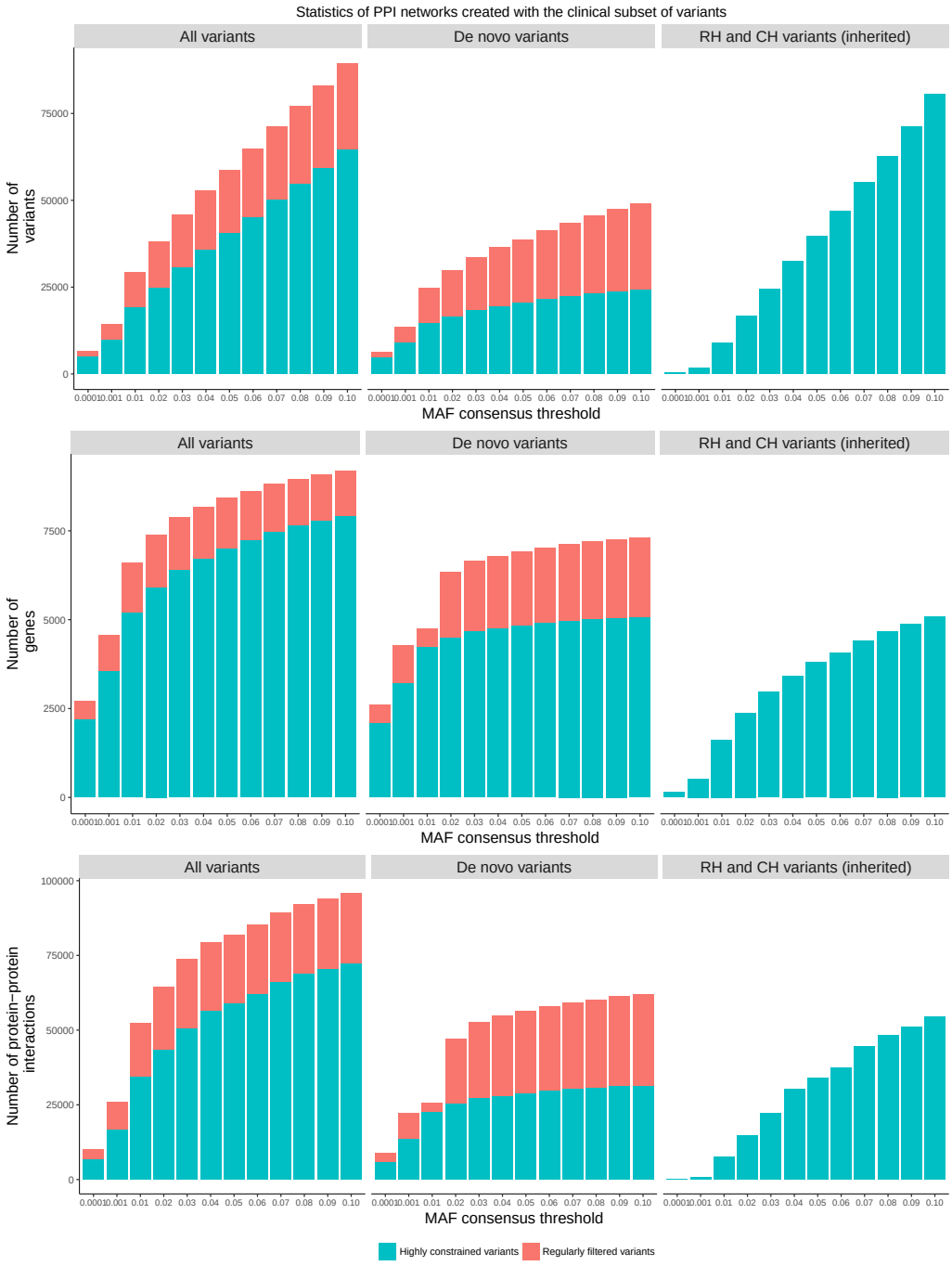


Figure 8: Number of variants, genes and protein-protein interactions for different threshold values of minor allele frequency for the filtering of the DDD dataset. De novo variants corresponds to autosomal dominant de novo variants and RH and CH correspond to autosomal inherited recessive homozygous and autosomal inherited compound heterozygous respectively. The red represent the false positive de novo variants that were not filtered out when we apply the filters to both the parents and the children. The blue are the variants obtained when we filter the de novo variants with the unfiltered parents' lists of variants. The number of false positive de novo variants increases when we admit higher minor allele frequency. Using more constrained filters reduces the number of de novo variants greatly.

4.1.3 Subclasses of the DDD dataset

The DDD dataset regroups 1133 patients suffering from different symptoms and exhibiting different phenotypes. In consequence, due to this diversity, finding strong links between the data and the diseases is very difficult. We subset the data three ways: the patients suffering from Intellectual Disability (ID) vs the other patients, the patients suffering from Autism Spectrum Disorder (ASD) vs the other patients and the monogenic patients vs the others (Fig 9). The ID and ASD patients are determined according to their HPO terms (Table 4). The monogenic patients are based on a list of patients diagnosed by the Sanger group [68]. We excluded from this list the patients with Copy Number variation and uniparental disomy as we do not have neither of these type of variation in our data.

HPO terms for ID	**HPO terms for ASD**
Intellectual disability, mild	Autism
Intellectual disability	Autism spectrum disorder
Intellectual disability, profound	Autistic behavior
Intellectual disability, moderate	Autism with high cognitive abilities
Intellectual disability, progressive	Impaired social interactions
Intellectual disability, borderline	Stereotypic behavior
Intellectual disability, severe	Repetitive compulsive behavior
Specific learning disability	Recurrent hand flapping
Global developmental delay	Inappropriate behavior
Developmental delay	
Developmental regression	

Table 4: HPO terms used to filter the ID or ASD patients in the DDD dataset. These are the terms most commonly used in clinics, but the list is non exhaustive, as the attribution of the HPO terms is done subjectively by the clinician.

We observe that a majority of the patients in the DDD dataset suffers from ID, which was also observed when we computed the frequencies of the HPO terms of the dataset (Table 2, Section 3.1.1). Only a small part of the patients present a autistic phenotype. It means that the sample size is considerably reduced and we need to proceed cautiously when comparing the subclass with others of bigger size. The monogenic patients regroup one third of the DDD dataset. We split according to the subclass of the patients the highly constrained clinical variants with a minor allele frequency below or equal to 0.01 which correspond to 10,967 distinct variants (Fig 10). Obviously, the number of patients greatly influences the number of variants for each subclass. The patients however, despite being in two different subclasses, share common variants. These variants, and their subsequent genes, will be used to discriminate subclass-specific results.

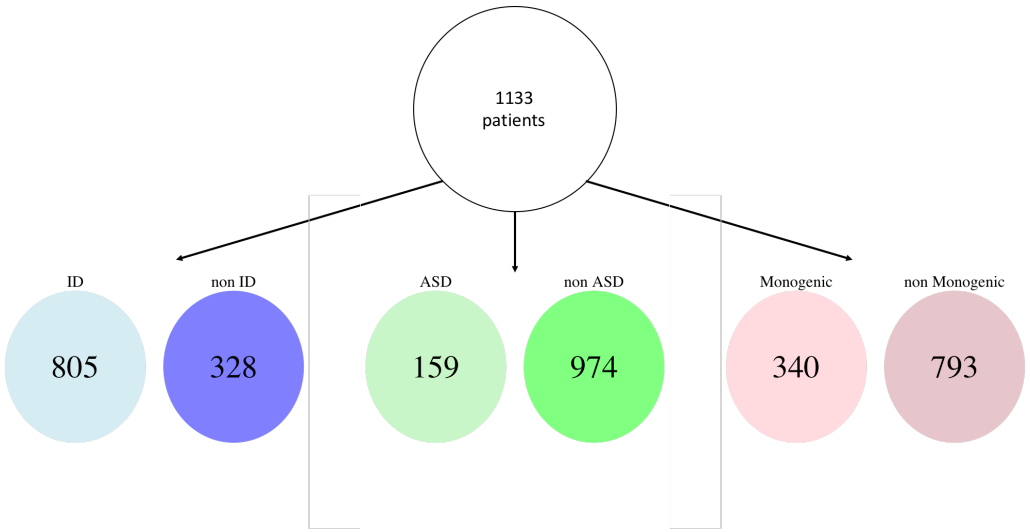


Figure 9: Distribution of the patients according to the three classes : ID vs non ID, ASD vs non ASD and Monogenic vs non Monogenic patients. The discrimination between subclasses was done with the HPO terms for ID and ASD, and with the list of patients diagnosed by the DDD for the monigenic patients. Patients suffering from ID constitutes around 71% of the dataset, while only 14% of the patients present a phenotype of autistic nature. The monogenic patients regroup 30% of the DDD dataset.

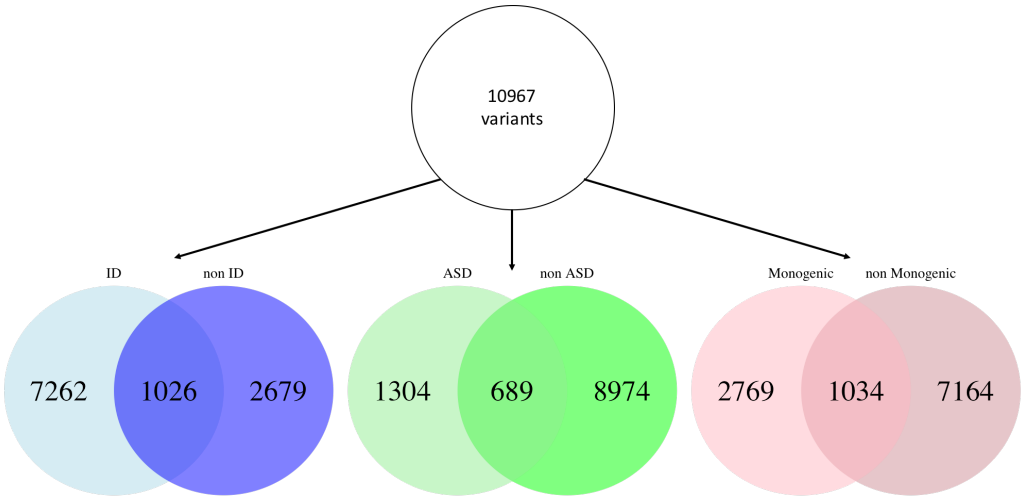


Figure 10: Distribution of the variants according to the three classes. The list of variants correspond to the clinical subset of highly constrained variants with a minor allele frequency lower than or equal to 0.01. The number of variants seems to be proportional to the number of patients.

As we want to discriminate the subclasses, we characterize the variants according to their genotype quality, their minor allele frequency and the predicted impact of the variants to see if we find differences between the subclasses. We note that the variants shared between

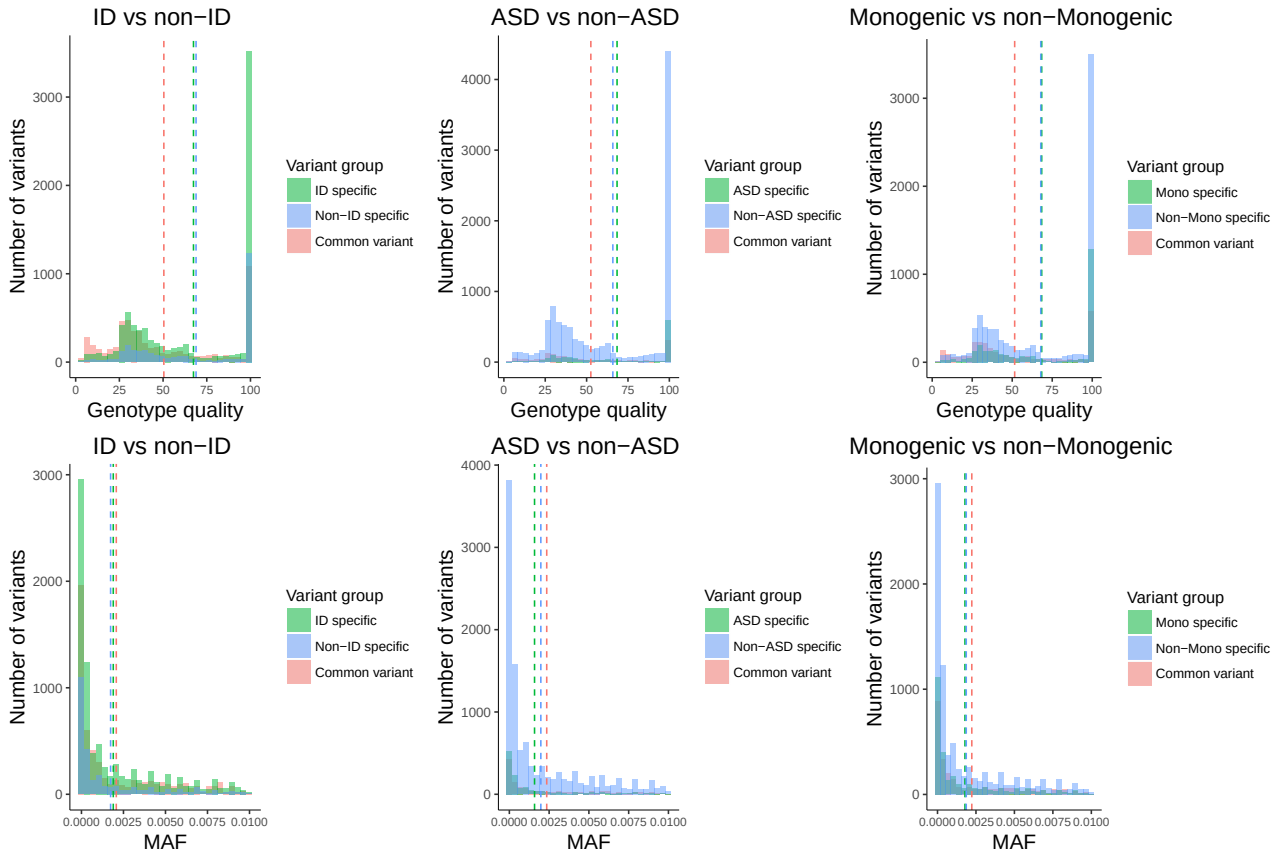


Figure 11: Characterization of the quality and the frequency of the variants according to their subclasses. We distinguish three groups: in green, the variants specific to the subclass of the phenotype studied, in blue, the variants specific to the subclass of patients not presenting the phenotype studied, and in red the variants shared by both subclasses. The dotted lines represent the average for each group of variants. No difference between the variants specific to subclasses can be found in the genotype quality or the minor allele frequency, however common variants seems to have a lower mean gentoype quality and are more frequent in average.

the subclasses are in average of a worse genotype quality, around 50%, than the average of the variants specific to the subclasses. The variants common to the subclasses are also more frequent than the specific (Fig 11). However, we do not distinguish any marking difference in the quality or the frequency between the variants specific to opposite subclasses.

The study of the proportion of the different effects of the variants of the different subclasses revealed that, while non-ASD and non-ID specific variants were mainly non-synonymous coding variants, the variants specific to the ID subclass were also non-synonymous coding variants (Fig 12). These variants could have an impact on the protein itself, but they could have also other effects, such as a change in the chromosome density and perhaps modifying the access to the gene or neighbouring ones.

From the list of variants of each subclass we obtain the list of coding genes which we use to build

the protein-protein interaction networks (Fig 13). It is important to note that the number of genes is directly related to the number of patients, so any direct analysis of the absolute number of genes has to take into account this bias. We observe that despite the fact that we had more variants specific to the subclasses, the number of genes does not reflect that fact. It can be explained by the fact that, while a variant might be specific to one subclass or the other, this variant could be on the same gene as another variant of the opposite subclass. This is why we observe this great number of common genes. This observation also reflects the fact that neurodevelopmental disorders overlap each other both by their genotype and their phenotype.

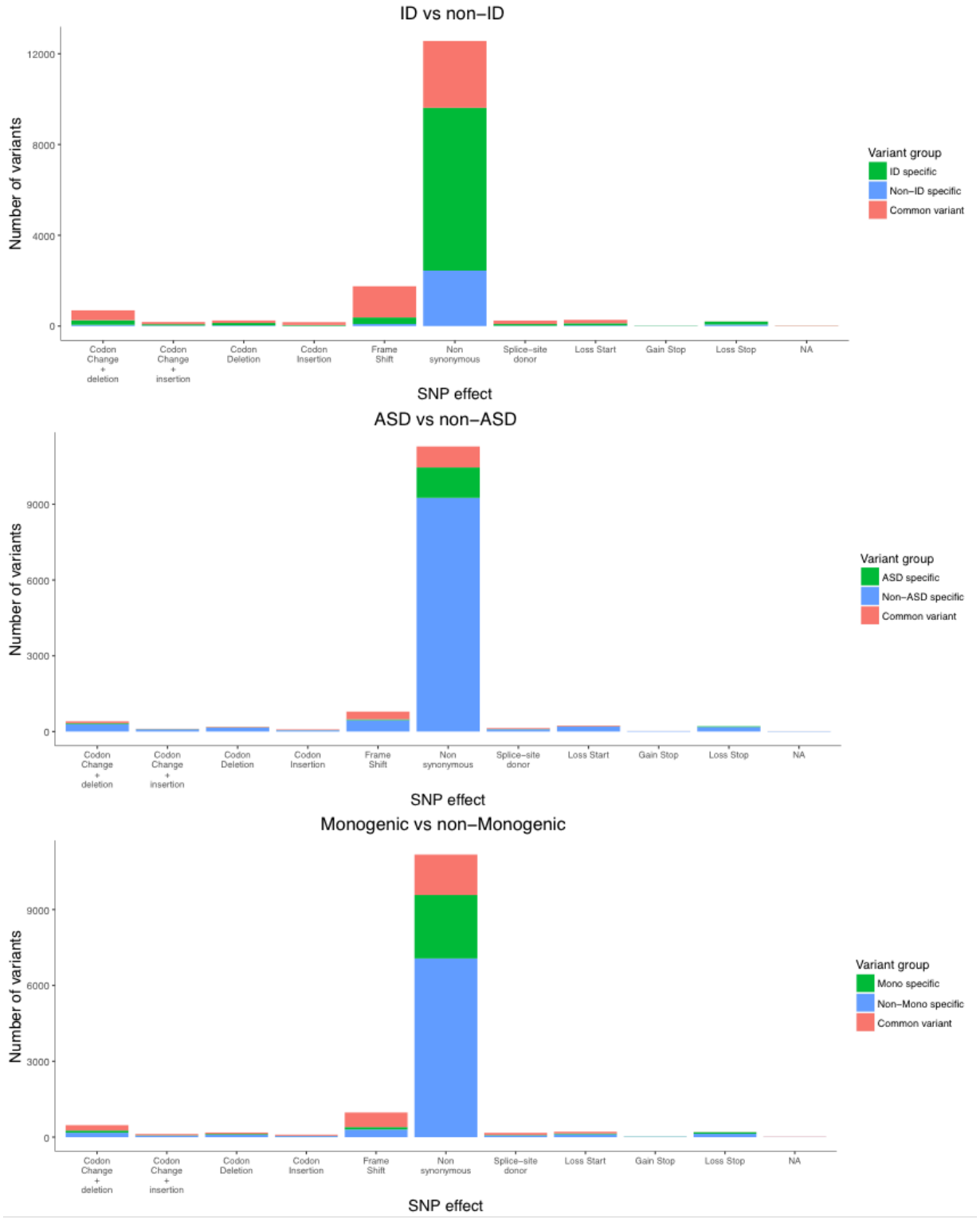


Figure 12: Characterization of the impact of the variants in the coding sequence according to their subclasses. We distinguish three groups of variants: in green, the variants specific to the subclass of the phenotype studied, in blue, the variants specific to the subclass of patients not presenting the phenotype studied, and in red the variants shared by both subclasses. The SNP effects are divided in ten categories plus one of unlabelled effect. The most represented category is the non-synonymous coding mutation.

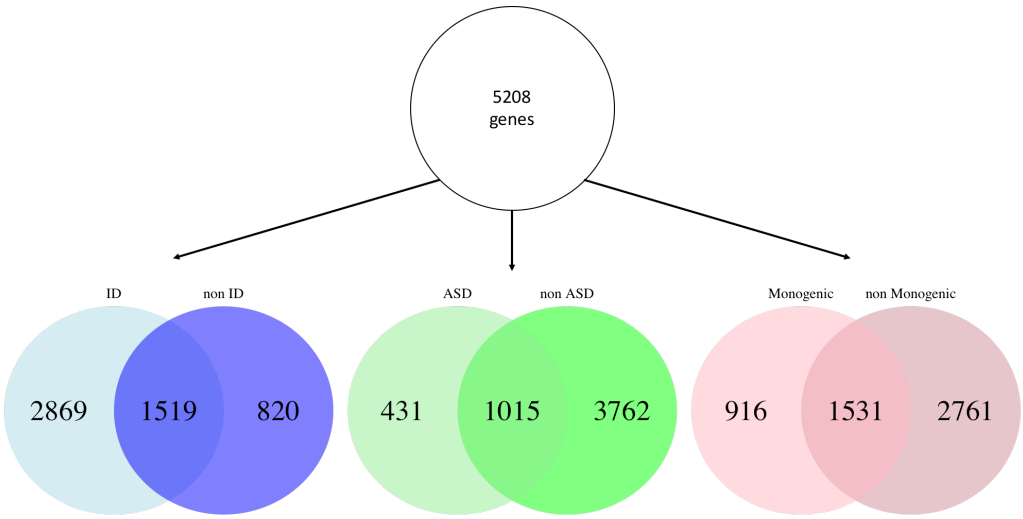


Figure 13: Distribution of the unique genes according to the three classes. The number of common genes represents around 20% of the total number of genes. The proportion of genes specific to each subclass follow the proportion of patients for each subclass.

4.2 Analysis of the protein-protein interactions networks

4.2.1 Topology of the networks

The study of the protein-protein interactions networks of the different subclasses reveals a similar pattern across all of them. The networks are divided in two parts: one big cluster regrouping nodes highly connected and their neighbourhood which will be called the *mainland*, and a variable number of isolated nodes or nodes with few edges called the *archipelago* (Fig 14).

We characterized the networks according to the network statistics of the overall network and the statistics of the mainland and the archipelago (Table 5.a). As expected, the mainland constitutes the bulk of the nodes and edges of the overall network. We hypothesize that the reason for the high number of nodes without any edges in the archipelago is because the archipelago is composed of proteins with few information about their protein partners, transcription factors or even because they are less known proteins. This unfortunately means that this part of the network needs additional information to be of real biological interest. We could try to group the protein according to their cellular location, as spatial information could be a clue of their function.

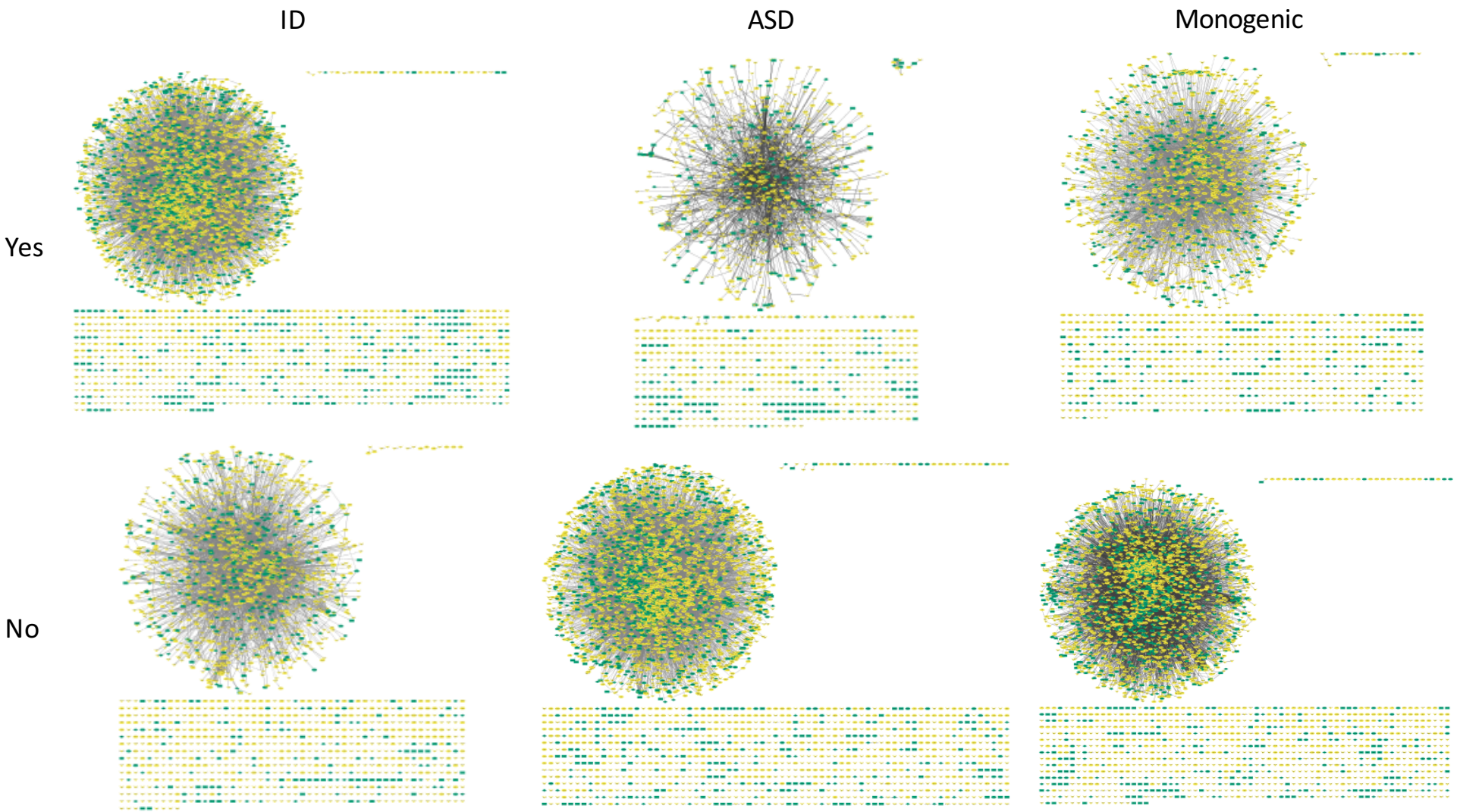


Figure 14: Protein-protein interaction networks of each subclass. The color of the node represent the origin of the variants : yellow for none (de novo) and green for the parents (inherited). All networks are divided in two parts. The highly connected one is called the mainland and all the disconnected islands around constitute the archipelago.

In order to compare the different groups of phenotype between themselves, we normalize the values of the Table 5a according to the proportion of the patients in the subclass (Table 5b). Unfortunately, this normalization still brings a lot of bias due to the size of the subclass. We compute the ratio of the normalized value of the subclass of the studied phenotype on the normalized value of the subclass with the rest of the patients (Table 5c) to compare the values of the subclass displaying the phenotype and the subclass not displaying it.

The main differences between the groups is caused by the number of edges in the networks and the statistics of the archipelago. While the ID subclass has a more connected mainland than the subclass with the patients with no ID, the mainland of the ASD subclass is less connected than the non ASD, despite having more nodes. However, while being the most connected, the ID subclass has comparatively less nodes than its opposite compared to the ASD or monogenic (in red in the table). On the other hand, the archipelago of ID is less connected and possess less nodes than the non ID one, whereas the archipelagos of ASD and monogenic patients is more connected with more nodes than their opposite (in grey in the table). The difference observed is based on a very small number of edges, so this could not be significant. We could explain this difference by the hypothesis that ID is caused by variants in protein complexes, represented here by the nodes in the mainland, and ASD is much more a result of a precise stochiometry and combination of several different variants. This hypothesis could also explain why the de novo trio approach to find the genetic causality of ID has been much more efficient than in ASD patients. However, we must also remember than a large part of ASD patients also displays ID phenotypes. This means that the ASD subclass overlaps greatly the ID subclass. Indeed, on the 1446 genes of the ASD subclass, 1354 (93.6%) are also present in ID. From the 92 ASD-only genes, 64% of these are found in the mainland and 36% in the archipelago, meaning that it is difficult to really infer from the topological differences observed between the networks a biological cause without relying on a circular argument. However, because of the bias introduce by the size of the network due to the difference in the number of patients in the different subclasses, we cannot compare the subclasses between themselves without being sure of how we should normalize the values in order to reduce the bias as much as possible.

To further explore the differences between the networks, we decide to analyse independently the mainland and the archipelago.

Table 5: Networks statistics for each subclass, their normalized values according the the percentage of patients in the subclass and the ratio of the phenotypic subclass on the value of the non-phenotypic subclass.

		ID	non ID	ASD	non ASD	Mono	non Mono
	% patients	71	29	14	86	30	70
Overall network characteristics	# nodes	4388	2339	1446	4777	2447	4292
	# edges	24937	6706	2640	29108	7822	23521
Mainland characteristics	# nodes	3334	1564	848	3658	1710	3237
	# edges	24933	6698	2605	29104	7815	23520
	Average degree	15.0	8.5	6.1	15.9	9.1	14.5
Archipelago characteristics	# nodes	1074	775	598	1119	737	1055
	# edges	4	8	35	4	7	1

(a) Networks statistics for each subclass.

		ID	non ID	ASD	non ASD	Mono	non Mono
Overall network characteristics	# nodes	61.8	80.7	103.3	55.5	81.6	61.3
	# edges	351.2	231.2	188.6	338.5	260.7	336.0
Mainland characteristics	# nodes	47.0	53.9	60.6	42.5	57.0	46.2
	# edges	351.2	231.0	186.1	338.4	260.5	336.0
	Average degree	0.2	0.3	0.4	0.2	0.3	0.2
Archipelago characteristics	# nodes	0.1	0.2	0.3	0.1	0.2	0.1
	# edges	15.1	26.7	42.7	13.0	24.6	15.1

(b) Normalized values of the network statistics by dividing the values by the percentage of patients in the subclass over all the patients.

		Ratio ID/nonID	Ratio ASD/nonASD	Ratio Mono/nonMono
Overall network characteristics	# nodes	0.8	1.9	1.3
	# edges	1.5	0.6	0.8
Mainland characteristics	# nodes	0.9	1.4	1.2
	# edges	1.5	0.5	0.8
	Average degree	0.7	2.4	1.5
Archipelago characteristics	# nodes	0.5	3.6	2.1
	# edges	0.6	3.3	1.6

(c) Ratio of the normalized values of the subclass of the phenotype divided by the subclass comprising the patients not displaying the phenotype. The ID subclass possess comparatively more edges than the non-ID subclass, compared to the other two subclass (in blue). The mainland of the ID subclass has less nodes but more edges than the non-ID subclass, while it is the opposite in the ASD and monogenic subclasses (in red). On the other hand, the archipelago in the ID has less edges and less nodes than the non ID, whereas it is the complete opposite again in the archipelago of the ASD and the monogenic subclasses (in grey).

4.2.2 Mainland analysis

In an effort to create a map of the different phenotypes according to their network, we decide to divide the mainland into clusters according to their local density (See Section 3.2.4) that we can afterwards characterize with the help of an enrichment analysis (See Section 3.2.5). To obtain the clusters specific to the subclass, we compute the percentage of overlap of the genes of each cluster with the list of the common genes between the subclasses and kept the clusters with an overlap with the common genes equal to or less than 75% (See the overlaping part in the Venn diagrams of Fig 13). These clusters are then independently characterized with an enrichment analysis regrouping GO BP, KEGG and Reactome terms. We keep the biological terms found only in the subclass displaying the phenotype in comparison to the subclass not displaying the phenotype in order to obtain at the end the specific biological terms of the subclass displaying the phenotype. The summary of the results obtained for the different subclasses are found in Table 6. The different biological terms found were reunited into categories for the analysis.

	ID	**non ID**	**ASD**	**non ASD**	**Mono**	**non Mono**
# clusters	39	31	20	41	33	42
# filtered clusters	37	21	15	40	26	39
# biological terms	633	228	97	764	222	540
# specific biological terms to studied phenotype	426		40		105	

Table 6: Mainlands analysis results. The number of biological terms for the phenotypic subclasses are based on the biological terms of the filtered clusters, while for their opposite subclasses the biological terms are for all of the clusters. The filtered biological terms are consequently filtered with the biological terms obtained with this method. The number of clusters, filtered or not, does not depend on the number of genes. The number of terms found are proportional to the number of genes of each subclass.

ID vs non ID

On 39 clusters, 37 have a overlap percentage with the common genes with a value equal to or less than 75% (Table 7). We take a special interest in the cluster completely specific to the ID subclass, in other words cluster with an overlap of 0% with the common genes. Only two clusters of three genes fill this criterion.

The first clusters is composed of UGT2B17, UGT1A4 and UGT2B28. They all are uridine diphosphate (UDP) glucuronosyltransferase, involved in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogenous compounds. Each protein in the cluster targets eugenol and testosterone for UGT2B17,bilirubin for UGT1A4 and steroid substrates

Threshold values of % overlap		100	75	50	25	0
Number of clusters		39	37	35	6	2
Max cluster size	**# nodes**	142	142	142	16	3
	# edges	2009	2009	2009	120	3
Min cluster size	**# nodes**	3	3	3	3	3
	# edges	3	3	3	3	3
Median cluster size	**# nodes**	6	8	9	8	3
	# edges	9	11	12	10	3

Table 7: Number of clusters of the mainland of ID subclass with different threshold values of the percentage overlap with the common genes of the subclasses. The number of clusters decrease with the threshold values. The biggest cluster has a overlap with common genes of between 50 and 25%, explaining why the median number of nodes and edges increases until 50% and decreases afterwards.

for UGT2B28.

The biological terms attached to this clusters are:

- Pentose and glucuronate interconversions
- Ascorbate and aldarate metabolism
- Steroid hormone biosynthesis
- Glucuronidation
- Retinol metabolism
- Porphyrin and chlorophyll metabolism
- Metabolism of xenobiotics by cytochrome P450
- Drug metabolism

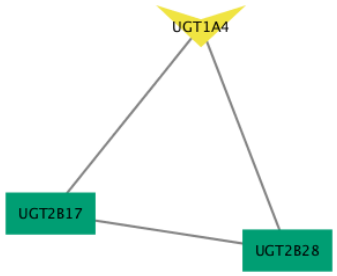
This means that some ID cases are directly related to problems in the detoxification of xenobiotics. Some forms of ID are hypothesized as being caused by inborn metabolism defects and are believed to be curable if we correct the metabolism errors [86]. These forms could be similar to the one we detected with this specific cluster.

The second cluster is composed of ORAI3, ORAI1 and TRPC3. They all are cation channels, ORAI1 and ORAI3 being calcium-specific. Mutations of ORAI1 and ORAI3 result in immunodificiency and myopathy [87]. Mutations in TRPC channels are linked to a myriad of

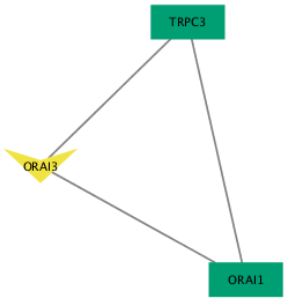
neurological disorders [88]. Mutations in cation channels have a potential impact on the neurodevelopmental disorders as they control the depolarization of neurons and thusly a myriad of the brain functions.

However, no biological terms could be attached to this particular cluster.

Both clusters possess the same characteristics: they are composed of three genes, connected in a triangular shape by three edges and each gene is the result of one unique variant in the cohort. They are both composed of one gene with a de novo variant and two genes with recessive homozygous variants (Fig 15). This could support the theory that while the inherited genetic context of the child is important and obviously plays a role in the apparition of some symptoms, de novo variants could be what is tipping the balance towards the disease.



(a) First cluster specific to ID, regrouping UGT2B17, UGT1A4 and UGT2B28.



(b) Second cluster specific to ID, regrouping ORAI3, ORAI1 and TRPC3.

Figure 15: Networks of the specific clusters of ID. Yellow and V shape color represents de novo origin, while the green represents that the variant is inherited and the rectangular shape that the variant is recessive homozygous. Both clusters are composed of one de novo variant and two recessive homozygous variants.

From the 426 biological terms specific to the ID subclass, we can distinguish several categories (Appendix A):

- Metabolism (carbohydrates mainly)
- RNA transport and metabolism
- Transport and cytoskeleton organization
- Hormone secretion and transport
- Phocytose and endocytose
- Virus life cycle
- DNA repair and metabolism
- Drug metabolism and detoxification
- Signaling
- Phagocytose and endocytose
- Synapse function and neurotransmitter receptor
- Axon guidance and projection

The biological terms were found across 16 clusters (Fig 18). We could relate the hormone secretion and transport category of terms to a cluster of 16 genes, which is a fairly specific cluster of ID with an overlap of only 25% with the common genes with the non-ID subclass.

ASD vs non ASD

On 20 clusters, 16 have a overlap percentage with the common genes with a value equal to or less than 75% (Table 8). We find only one cluster of three genes with no overlap with the genes shared between ASD and non ASD subclasses, which also correspond to the smallest cluster.

The cluster is composed of HRH1, LTB4R2 and EDN2. HRH1 is a histamine receptor expressed in the central nervous system. Histamine is a neurotransmitter involved in numerous functions such as the neuroinflammation, cognition, sleep, attention, sensory function and motor function [89]. LTB4R2 is a leukotriene receptor gene produced in the neuronal system. Leukotriene is an inflammatory mediator that is typically expressed at the same time as histamine. Leukotriene was detected as elevated in the blood of autistic children and is promoted as a early biomarker for austism [90, 91]. EDN2 is coding for endothelin 2, a vasoconstrictive peptide. EDN2 was linked indirectly to autistic behaviour in a CAPDS2 knockout mouse. The KO mouse

Threshold values of % overlap		100	75	50	25	0
Number of clusters		20	16	3	1	1
Max cluster size	**# nodes**	39	39	15	3	3
	# edges	202	202	36	3	3
Min cluster size	**# nodes**	3	3	3	3	3
	# edges	3	3	3	3	3
Median cluster size	**# nodes**	6	6.5	5	3	3
	# edges	12	12.5	10	3	3

Table 8: Number of clusters of the mainland of ASD subclass with different threshold values of the percentage overlap with the common genes of the subclasses. The number of clusters decreases with the threshold values.

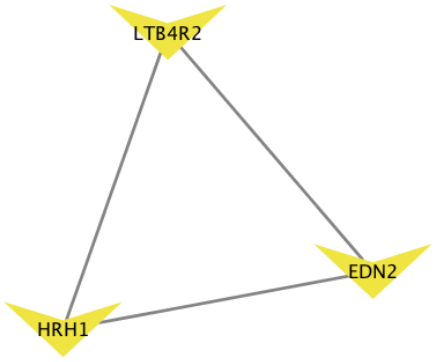


Figure 16: Network of the specific cluster of ASD. The yellow and V shape represent the de novo variant. The cluster is composed of three de novo variants

exhibited autistic-like behaviour and impaired cerebellar development in addition. EDN2 was differentially expressed [92]. EDN2 was also detected as a gene for interest in a genome-wide search for candidate genes for autism [93].

No biological term could be linked to this cluster.

The topology of this specific cluster is similar to the specific clusters of ID, where the three genes are connected in a triangular shape with three edges. However, the nature of the variants in ASD of this specific clusters is different, as we have only three de novo variants (Fig 16). This could mean that this phenotype of neural inflammatory response is specific to ASD.

The 40 biological terms specific to the ASD subclass, found across 2 clusters, are divided in several categories (Appendix B):

- RNA metabolism and transport

- DNA repair and metabolism
- Virus life cycle and infection
- Transport and cytoskeleton organization
- Signaling
- Immunology and autoimmune diseases
- Metabolism
- Phagocytose and endocytose

The first cluster regroups the biological terms of RNA transport, transport and organization of the cytoskeleton and DNA metabolism whereas the second cluster englobes the signaling and the immunology terms (Fig 18). The first more general and metabolic cluster has an overlap of 71.8% with the common genes, probably explaining the nature of its biological terms.

It is surprising to not find any biological term linked to neural function or the behaviour. However, this difference observed between ASD terms and ID terms also points towards the monogenic character of ID causality. If we can find biological terms of neuronal function in ID, this means that the amount of genes linked to this category of terms is greater and implies a direct link to the neurodevelopmental process. In ASD, the genetic cause is hypothesized to be oligogenic, which would explain why we obtain biological terms not directly involved in neurodevelopment. However, this difference could also be caused by the difference in the number of patients in ID compared to the ASD. The size difference means that we have a larger number of genes to related to enriched biological terms.

Monogenic vs non Monogenic

Out of 33 clusters, 26 have a overlap percentage with the common genes with a value equal to or less than 75% (Table 9). Again, we only find one cluster with 0% overlap with shared genes, which also correspond to the smallest cluster.

The cluster is composed of ARID2, SS18 and DPF3. All three proteins are involved in chromatin-remodeling process. ARID2 interacts with BAF chromatin-remodeling complex, which facilitates ligand-dependent transcriptional activation by nuclear receptors. Mutations in ARID2 was also linked to delay in the development and intellectual disability [94]. SS18 is a subunit protein also involved in the chromatin remodeling complex BAF. BAF complex plays a required role in development and interacts with CDH7 in neural development, the latter being linked to ASD as a rare syndromic variant also linked to CHARGE syndrome [95, 96]. Mutations and interference with BAF complex are linked to several neurological disorders ranging

Threshold values of % overlap		100	75	50	25	0
Number of clusters		33	26	11	4	1
Max cluster size	**# nodes**	74	74	8	4	3
	# edges	342	342	23	6	3
Min cluster size	**# nodes**	3	3	3	3	3
	# edges	3	3	3	3	3
Median cluster size	**# nodes**	5	5	4	4	3
	# edges	8	7.5	5	4.5	3

Table 9: Number of clusters of the mainland of the monogenic subclass with different threshold values of the percentage overlap with the common genes of the subclasses, as well as their maximum, minimum and median number of nodes and edges of the clusters for each threshold value. The size and the number of the clusters decrease with the threshold value of percentage of overlap.

from intellectual disability to psychiatric conditions such as schizophrenia, and neurodegenerative diseases [97]. DPF3 is a subunit of the nBAF (neural BAF) chromatin-remodeling complex, which plays a important role in the switch of neural stem cells towards the neural differentiation, as well as a role into the dendrite growth.

It is interesting to note that none of the genes of the variants used for the diagnoses of the monogenic patients were found in the cluster.

Similar to the previous clusters found in the other subclasses, the cluster is composed of three genes in a triangular shape connected by three edges. However, the composition of the nature of the variants at the origin of the three genes is again different. One gene has a de novo variant for origin, while the two others have two compound heterozygous variants for origin (Fig 17). While the nature of the variants is different from the one in ID, we can still observe the ratio two inherited genes with one de novo.

Six different biological terms are linked to the cluster:

- Neurophilin interactions with VEGF and VEGFR
- Semaphorin interactions
- Sema3A PAK dependent Axon repulsion
- SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion
- CRMPs in Sema3A signaling
- Signal transduction by L1

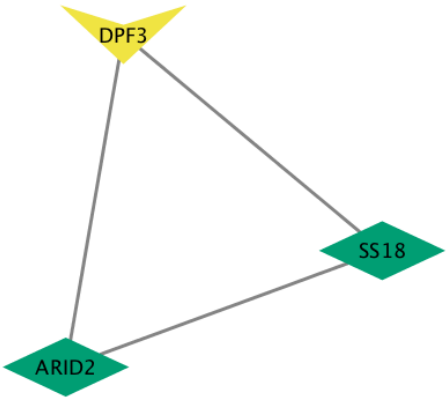


Figure 17: Network of the cluster specific to the monogenic subclass. The yellow color and V shape represent a de novo variant, the green color represent an inherited variant and the diamond shape is for the compound heterozygous variant. The cluster is composed of one de novo variant and two compound heterozygous variants.

All these terms are linked to neuronal functions, meaning that the specific cluster brings evidences of mutations in genes involved in brain function. Neurophilin (or neuropilin) is a receptor in neurons responsible for axon guidance, particularly in combination with the semaphorin proteins (Sema3A). Both the neuropilin and the semaphorins were linked to neurological disorders [98, 99]. The L1 signal transduction provides cues to the neuron for axon guidance. L1 interacts also with the neuropilin and Sema3A receptor to mediate Sema3A induced growth cone collapse and axon repulsion.

The specificity of these genes to the monogenic patients and their biological terms indicates that axon guidance could be bias the genetic causality more towards the monogenic. The same biological terms could be found in ID in which strong monogenic cases can easily be found.

The categories found for the 105 biological terms characterizing the clusters of monogenic patients (Appendix C):

- Hormone secretion and transport
- Virus Life cycle and infection
- Metabolism
- Transport and cytoskeleton organization
- Immunology response and autoimmune diseases
- RNA transport and metabolism
- DNA repair and metabolism

- Signaling

The biological terms were found across 7 clusters (Fig 18). The hormone and secretion transport could be linked to a specific cluster of 52 genes, the second biggest cluster with a overlap value of 61.5%. The transport and cytoskeleton organization, as well as the RNA transport could be linked to a cluster of 74 genes, the biggest cluster of the mainland (Table 9), which is also linked to the metabolism and virus life cycle and infection categories, with an overlap value of 60.8%. The immunology could be specifically linked to a cluster of 5 genes which was also linked to virus life cycle and infection, as well as signalling. The other categories were found across different clusters.

Distribution of the biological terms in subclasses

The categories of biological terms are generally distributed across several clusters in the subclasses (Fig 18). We can assign specific categories to clusters only in ASD and in the monogenic subclass, which could be due to the fact that we have less genes in these subclasses, reducing in consequence the number of clusters with an overlap equal to or less than 75% (Table 8 and Table 9), subsequently reducing also the number of biological terms specific to the subclass (Table 6). However, we observe that ASD seems to possess similar categories of biological terms to the ID, except for the immunology category (Fig 18). This is due to the fact that ASD share a lot of genes with ID (See Section 4.2.1). This could also explain the fact that a lot of ASD patients also suffer from ID (123 children on 159).

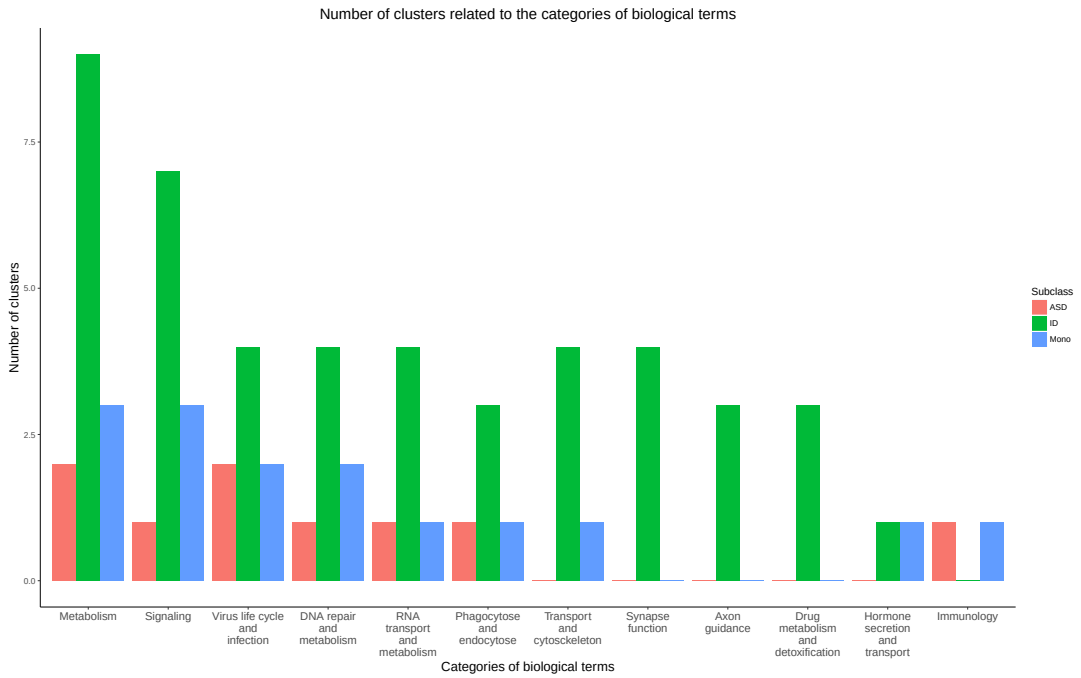


Figure 18: Number of clusters related to each category of specific biological terms found in the different subclasses.

4.2.3 Archipelago analysis

To complete the analysis of the network, we need to look at the archipelago. As the nodes in the archipelago are mainly isolated, we do not cluster this part of the network and use directly the list of genes for an enrichment analysis to characterize the archipelago in a biological way.

After obtaining the biological terms for each subclass, we filter the biological terms in order to have only the terms specific to the studied phenotype subclass (Table 10).

	ID	**non ID**	**ASD**	**non ASD**	**Mono**	**non Mono**
# biological terms	2	8	8	5	3	14
# specific biological terms to the studied phenotype	1		7		3	

Table 10: Archipelago enrichment analysis results. The number of terms obtained is very low.

Despite the fact that more genes can be found in the ID archipelago compared to the non ID (and vice-verso for the ASD and non-ASD), the number of biological terms found with the enrichment analysis does not reflect this statistics. This is due to two reasons. The first one is that the genes in the archipelago are by definition the less known genes. Consequently, less information about their function and their involvement in the human physiology can be found and use to annotate them. The second reason is the fact that the statistical value of a biological term in the enrichment analysis is based on the number of genes found for this term compared to the expected number that could be found by chance for this term according to the size of the gene list. This means that to be statistically significant some terms needs a lot more genes than others, which could be impaired if the size of the gene list increases. This means that when you have a bigger list of genes, you need more genes to overcome the expected number of genes found by chance for some terms.

ID vs non ID

The only biological term found is cilium or flagellum-dependent cell motility. This biological term is covered by nine genes of the archipelago and is covered by 30% by these genes. However, we cannot explain yet the biological relevance of such a term in the context of intellectual disability. In order to understand the interest of this term, we explore the genes of the archipelago composing it.

CCSAP is a protein involved in the microtubule stabilization, especially at the spindle poles during mitosis. CFAP46 is part of the cilium axoneme, involved in cell projection. TEKT5 is related to the sperm flagellum. The six other genes are all heavy chains of the dynein axonemal, which is a microtubule-associated motor protein for transport. No connection between these

genes and neurodevelopmental disorders could be found.

ASD vs non ASD

We find 7 biological terms specifically linked to the ASD subclass.

- Homophilic cell adhesion via plasma membrane adhesion molecules
- Skeletal muscle fiber differentiation
- Mesodermal cell migration
- Neurotransmitter Release Cycle
- Glutamate Neurotransmitter Release Cycle
- Transport of inorganic cations/anions and amino acids/oligopeptides
- HCN channels

While we did not find any biological term linked to neural function in the mainland of ASD, the four last terms of the archipelago are involved in brain function. Two subcategories can be distinguished: the first two terms are linked to the synaptic function, while the two others are related to the polarization and action potential of the neuron.

16 genes from the archipelago are at the origin of these terms (SLC12A7, SLC15A3, SLC17A7, SLC1A6, SLC1A7, SLC24A3, SLC24A4, SLC26A1, SLC26A9, SLC3A1, SLC4A9, SLC7A5, SLC9A5, TSPOAP1, HCN2, HCN3). 13 of these genes code for excitatory amino acid transporter, directly involved in the synaptic function for clearing neurotransmitter such as the glutamate. TSPOAP1 is a peripheral-type benzodiazepine receptor-associated protein, a modulatary allosteric site on the GABAA receptors. GABA is the major inhibitory neurotransmitter of the central nervous system. Mutations inhibitory subunits of the GABA receptor are associated with epilepsy and seizures [100]. The last two genes code for cation channels and participate in the hyperpolarization of the neuron. Mutations in cation channels can impact the potential action of the neurons and thusly the brain function.

Monogenic vs non Monogenic

We find three biological terms specific to the monogenic archipelago.

- Cilium or flagellum-dependent cell motility
- Cilium movement
- Detection of light stimulus involved in visual perception

The first two terms regroups the same genes as with the ID subclass of the same term, along with three new genes: CCDC40, a protein necessary for motile cilia function,, and CFAP100 and CFAP61, both of the same family as CFAP46, which are protein associated to the cilium and flagellum and are involved in cell projection.

The third term is due to 5 genes. ATP8A2 is involved in the transport of aminophospholipids in brain cells, retinal photoreceptors and testis. A missense mutation in this gene has been strongly linked to the cerebellar ataxia, mental retardation and dysequilibrium syndrome [101]. BEST1 is coding for a protein of the bestrophin family. Bestrophins form or regulate ion-channel that can be found in retinal cells. They are linked to various retinopathias. EYS can be found in the photoreceptor layer of the retine and is at the origin of the autosomal recessive retinitis pigmentosa. GJA10 is a connexin medtiating the gap junction trafficking in the retinal cells. Finally, SEMA5B, the most interesting gene, codes for a semaphorin, involved in the axon growth during development of the nervous system. SEMA5B is considered as a gene of interest for developmental disorders such as ASD, due to his involvement in synapse connexion and density [102].

4.2.4 Union of the mainland and archipelago analyses

No association in the biological terms could be found between the mainland and the archipelago. This is due to the fact that the archipelago is composed of less known genes and thusly the number of enriched terms found is smaller than for the mainland. In average, 5.6% of the genes of the mainland do not have any attached biological term, whereas this is the case for 23% of the genes in the archipelago.

One interesting thing to note is that in ID and monogenic subclasses, the brain function related terms can be found in the mainland. For the ASD, and partially for the monogenic, these terms are found in the archipelago. This reflects what we found in the basic characteristics of the topology of the network (Table 5). The interesting information of the ASD subclass is in the archipelago, just as the archipelago is uncharacteristically enriched compared to the archipelago of the ID subclass.

Chapter 5

Conclusion

With the study of the specific disorders networks, we could observe a difference in the topology of intellectual disability and autism spectrum disorder, the former being in general more often monogenic than the latter. This difference was also found with the search for biological terms, where the terms with a neural connotation were found in different parts of the network, where the topological difference was also found. This difference could be at the origin of the monogenic character of the intellectual disability compared to autism, especially when similar findings were found with the monogenic subclass. However, we fear that this difference could also be due to the difference of the sample size which could introduce a bias. We should find for the future a better way to normalize our values to compare class of phenotypes between themselves.

Mapping categories of biological terms to clusters seems to be difficult because of the variety of the biological terms associated with the clusters. No exact category could be attributed to a specific set of clusters, meaning that the clustering of the networks by density clustering do not necessarily give functionally connected clusters.

It would be interesting to add information on the different proteins, like the burden for example. Burden can be described as the weight of the damage that a protein brings to the physiology because of the number of variants attached to this protein, their type of mutation or the scores of these mutations calculated with the different prediction tools. A burden score could also be used as a way to cluster proteins in the networks that are possibly more damaging than the others. We would need to design ways to refine such a score, like machine learning techniques.

While we implemented the study of networks from the patient scale to the family scale and the whole cohort scale, we decided that by looking at the cohort network we use the power of the size of the cohort to search for convergent biological terms. However, it would be interesting in the future to return to the smaller scale and use what we discovered with the whole cohort and determine if the same conclusions can be drawn from the study of the individual networks. Another limitation is the fact that we use only the genes from relevant clinical variants to build

our protein-protein interactions networks and not also their first interactors. The latter choice could unearth connections between proteins of the same or connecting pathways or protein complexes. However, we would then also need to handle much larger and complex networks. This would need the use of the big data methods and to adapt the framework so it can work in a parallelized way.

Another way of expanding our research would be to explore other types of networks and integrate their information together to enrich the basic protein-protein interactions networks.

Despite the fact that we made interesting observations, they can only be considered preliminary results to this domain of research. Future work will be built up on the already existing framework in order to extend the study to the exploration of different types of networks in neurodevelopmental disorders, as well as on several different scales in order to create maps of these disorders.

Appendix A

Biological terms of ID mainland

Biological term	**Cluster ID**	**Category**
Citrate cycle (TCA cycle)	2	Metabolism
Pyrimidine metabolism	2	Metabolism
Lysine degradation	2	Metabolism
Starch and sucrose metabolism	2	Metabolism
Membrane Trafficking	2	Metabolism
Pyruvate metabolism	2	Metabolism
Glycogen storage diseases	2	Metabolism
Glycogen synthesis	2	Metabolism
Myoclonic epilepsy of Lafora	2	Synapse function
Glycogen storage disease type 0 (liver GYS2)	2	Metabolism
Glycogen storage disease type IV (GBE1)	2	Metabolism
Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide, LLO) and transfer to a nascent protein	2	Metabolism
Asparagine N-linked glycosylation	2	Metabolism
Synthesis of substrates in N-glycan biosythesis	2	Metabolism
Lysosomal glycogen catabolism	2	Metabolism
Glycogen storage disease type II (GAA)	2	Metabolism
Vesicle-mediated transport	2	Phagocytose and endocytose
Diseases of carbohydrate metabolism	2	Metabolism
Diseases of metabolism	2	Metabolism
GDP-fucose biosynthesis	2	Metabolism

COPI-dependent Golgi-to-ER retrograde traffic	2	Transport and cytoskeleton
Retrograde transport at the Trans-Golgi-Network	2	Transport and cytoskeleton
Intra-Golgi and retrograde Golgi-to-ER traffic	2	Transport and cytoskeleton
Glycogen breakdown (glycogenolysis)	2	Metabolism
Gluconeogenesis	2	Metabolism
Glucose metabolism	2	Metabolism
Golgi-to-ER retrograde transport	2	Transport and cytoskeleton
SNARE interactions in vesicular transport	2	Phagocytose and endocytose
Synaptic vesicle cycle	2	Synapse function
Shigellosis	2	Virus life cycle and infection
Central carbon metabolism in cancer	2	Metabolism
receptor-mediated endocytosis	2	Phagocytose and endocytose
actin filament organization	2	Transport and cytoskeleton
endomembrane system organization	2	Transport and cytoskeleton
cellular component disassembly	2	Transport and cytoskeleton
actin filament depolymerization	2	Transport and cytoskeleton
regulation of actin filament depolymerization	2	Transport and cytoskeleton
negative regulation of actin filament depolymerization	2	Transport and cytoskeleton
negative regulation of actin filament polymerization	2	Transport and cytoskeleton
negative regulation of protein complex assembly	2	Transport and cytoskeleton
negative regulation of protein polymerization	2	Transport and cytoskeleton

regulation of actin filament-based process	2	Transport and cytoskeleton
negative regulation of protein complex disassembly	2	Transport and cytoskeleton
regulation of protein complex disassembly	2	Transport and cytoskeleton
cellular protein complex assembly	2	Transport and cytoskeleton
ATP metabolic process	2	Metabolism
protein depolymerization	2	Transport and cytoskeleton
negative regulation of cytoskeleton organization	2	Transport and cytoskeleton
actin filament capping	2	Transport and cytoskeleton
clathrin-dependent endocytosis	2	Phagocytose and endocytose
supramolecular fiber organization	2	Transport and cytoskeleton
regulation of protein depolymerization	2	Transport and cytoskeleton
negative regulation of protein depolymerization	2	Transport and cytoskeleton
regulation of supramolecular fiber organization	2	Transport and cytoskeleton
negative regulation of supramolecular fiber organization	2	Transport and cytoskeleton
Free fatty acids regulate insulin secretion	3	Metabolism
Synthesis, secretion, and inactivation of Glucose-dependent Insulinotropic Polypeptide (GIP)	3	Metabolism
Fatty Acids bound to GPR40 (FFAR1) regulate insulin secretion	3	Metabolism
Renin secretion	4	Metabolism
hormone transport	4	Metabolism
peptide hormone secretion	4	Metabolism
hormone secretion	4	Metabolism
Transport of the SLBP independent Mature mRNA	5	RNA metabolism
Transport of the SLBP Dependant Mature mRNA	5	RNA metabolism

Transport of Mature mRNA Derived from an Intronless Transcript	5	RNA metabolism
Transport of Mature mRNAs Derived from Intronless Transcripts	5	RNA metabolism
Influenza Life Cycle	5	Virus life cycle and infection
Transport of Ribonucleoproteins into the Host Nucleus	5	Transport and cytoskeleton
Influenza Viral RNA Transcription and Replication	5	Virus life cycle and infection
Viral Messenger RNA Synthesis	5	Virus life cycle and infection
Regulation of Glucokinase by Glucokinase Regulatory Protein	5	Metabolism
Interactions of Vpr with host cellular proteins	5	Transport and cytoskeleton
Nuclear import of Rev protein	5	Transport and cytoskeleton
Vpr-mediated nuclear import of PICs	5	Virus life cycle and infection
vRNP Assembly	5	Virus life cycle and infection
Gene Silencing by RNA	5	RNA metabolism
Nuclear Envelope Breakdown	5	Transport and cytoskeleton
Nuclear Pore Complex (NPC) Disassembly	5	Transport and cytoskeleton
Regulation of HSF1-mediated heat shock response	5	Transport and cytoskeleton
Cellular response to heat stress	5	Transport and cytoskeleton
Transcriptional regulation by small RNAs	5	RNA metabolism
tRNA processing in the nucleus	5	RNA metabolism
rRNA modification in the nucleus and cytosol	5	RNA metabolism
Major pathway of rRNA processing in the nucleolus and cytosol	5	RNA metabolism
Mitotic Prophase	5	DNA repair and metabolism

Glucose transport	5	Metabolism
Gluconeogenesis	5	Metabolism
Glucose metabolism	5	Metabolism
rRNA processing	5	RNA metabolism
rRNA processing in the nucleus and cytosol	5	RNA metabolism
Ribosome biogenesis in eukaryotes	5	Metabolism
ribosome biogenesis	5	DNA repair and metabolism
Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template	7	DNA repair and metabolism
Recognition of DNA damage by PCNA-containing replication complex	7	DNA repair and metabolism
Translesion Synthesis by POLH	7	DNA repair and metabolism
Recognition and association of DNA glycosylase with site containing an affected pyrimidine	7	DNA repair and metabolism
Cleavage of the damaged pyrimidine	7	DNA repair and metabolism
Recognition and association of DNA glycosylase with site containing an affected purine	7	DNA repair and metabolism
Cleavage of the damaged purine	7	DNA repair and metabolism
Displacement of DNA glycosylase by APEX1	7	DNA repair and metabolism
POLB-Dependent Long Patch Base Excision Repair	7	DNA repair and metabolism
Resolution of AP sites via the multiple-nucleotide patch replacement pathway	7	DNA repair and metabolism
Resolution of AP sites via the single-nucleotide replacement pathway	7	DNA repair and metabolism
Downregulation of ERBB4 signaling	7	Signaling
HIV Transcription Initiation	7	Virus life cycle and infection
RNA Polymerase II HIV Promoter Escape	7	Virus life cycle and infection
Polymerase switching on the C-strand of the telomere	7	DNA repair and metabolism

Processive synthesis on the C-strand of the telomere	7	DNA repair and metabolism
Telomere C-strand (Lagging Strand) Synthesis	7	DNA repair and metabolism
Telomere C-strand synthesis initiation	7	DNA repair and metabolism
Removal of the Flap Intermediate from the C-strand	7	DNA repair and metabolism
Activation of ATR in response to replication stress	7	DNA repair and metabolism
Unwinding of DNA	7	DNA repair and metabolism
PPARA activates gene expression	7	Drug metabolism and detoxification
Biological oxidations	7	Drug metabolism and detoxification
Phase 1 - Functionalization of compounds	7	Drug metabolism and detoxification
CYP2E1 reactions	7	Drug metabolism and detoxification
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	7	Drug metabolism and detoxification
Metabolism of xenobiotics by cytochrome P450	7	Drug metabolism and detoxification
Drug metabolism	7	Drug metabolism and detoxification
Fatty acid, triacylglycerol, and ketone body metabolism	7	Metabolism
Mismatch Repair	7	DNA repair and metabolism
Mismatch repair (MMR) directed by MSH2:MSH6 (MutSalpha)	7	DNA repair and metabolism
Mismatch repair (MMR) directed by MSH2:MSH3 (MutSbeta)	7	DNA repair and metabolism
Mitochondrial translation initiation	7	DNA repair and metabolism
Mitochondrial translation	7	DNA repair and metabolism

Mitochondrial translation elongation	7	DNA repair and metabolism
Mitochondrial translation termination	7	DNA repair and metabolism
Regulation of TP53 Activity	7	DNA repair and metabolism
APEX1-Independent Resolution of AP Sites via the Single Nucleotide Replacement Pathway	7	DNA repair and metabolism
PCNA-Dependent Long Patch Base Excision Repair	7	DNA repair and metabolism
Termination of translesion DNA synthesis	7	DNA repair and metabolism
HDR through Single Strand Annealing (SSA)	7	DNA repair and metabolism
HDR through Homologous Recombination (HRR)	7	DNA repair and metabolism
DNA Double-Strand Break Repair	7	DNA repair and metabolism
Resolution of D-Loop Structures	7	DNA repair and metabolism
Homology Directed Repair	7	DNA repair and metabolism
Resolution of D-loop Structures through Synthesis-Dependent Strand Annealing (SDSA)	7	DNA repair and metabolism
HDR through Homologous Recombination (HR) or Single Strand Annealing (SSA)	7	DNA repair and metabolism
Resolution of D-loop Structures through Holliday Junction Intermediates	7	DNA repair and metabolism
Homologous DNA Pairing and Strand Exchange	7	DNA repair and metabolism
Processing of DNA double-strand break ends	7	DNA repair and metabolism
Presynaptic phase of homologous DNA pairing and strand exchange	7	Synapse function
Gap-filling DNA repair synthesis and ligation in GG-NER	7	DNA repair and metabolism
Fanconi Anemia Pathway	7	Metabolism

DNA replication initiation	7	DNA repair and metabolism
Activation of the pre-replicative complex	7	DNA repair and metabolism
Polymerase switching	7	DNA repair and metabolism
Leading Strand Synthesis	7	DNA repair and metabolism
Removal of the Flap Intermediate	7	DNA repair and metabolism
Processive synthesis on the lagging strand	7	DNA repair and metabolism
Lagging Strand Synthesis	7	DNA repair and metabolism
DNA strand elongation	7	DNA repair and metabolism
G2/M DNA damage checkpoint	7	DNA repair and metabolism
RNA Polymerase II Promoter Escape	7	RNA metabolism
RNA Polymerase II Transcription Pre-Initiation And Promoter Opening	7	RNA metabolism
Base Excision Repair	7	DNA repair and metabolism
DNA Damage Bypass	7	DNA repair and metabolism
DNA Repair	7	DNA repair and metabolism
Depurination	7	DNA repair and metabolism
Depyrimidination	7	DNA repair and metabolism
Base-Excision Repair, AP Site Formation	7	DNA repair and metabolism
Abasic sugar-phosphate removal via the single-nucleotide replacement pathway	7	DNA repair and metabolism
Resolution of Abasic Sites (AP sites)	7	DNA repair and metabolism

Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex	7	DNA repair and metabolism
RNA Polymerase II Transcription Initiation	7	RNA metabolism
RNA Polymerase II Transcription Initiation And Promoter Clearance	7	RNA metabolism
Aryl hydrocarbon receptor Signaling	7	Signaling
Negative regulators of RIG-I/MDA5 signaling	7	Signaling
TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling	7	Signaling
DNA replication	7	DNA repair and metabolism
Base excision repair	7	DNA repair and metabolism
Nucleotide excision repair	7	DNA repair and metabolism
Mismatch repair	7	DNA repair and metabolism
Homologous recombination	7	DNA repair and metabolism
DNA replication	7	DNA repair and metabolism
translation	7	DNA repair and metabolism
snRNA metabolic process	9	RNA metabolism
ncRNA processing	9	RNA metabolism
Neurotransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	10	Synapse function
Transmission across Chemical Synapses	10	Synapse function
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	12	Metabolism
GRB2 events in ERBB2 signaling	12	Signaling
PI3K events in ERBB2 signaling	12	Signaling
AKT-mediated inactivation of FOXO1A	12	Signaling
ATF4 activates genes	12	Signaling
Ephrin signaling	12	Synapse function
mRNA decay by 3' to 5' exoribonuclease	12	RNA metabolism

Neurofascin interactions	12	Synapse function
Butyrate Response Factor 1 (BRF1) binds and destabilizes mRNA	12	RNA metabolism
Tristetraprolin (TTP, ZFP36) binds and destabilizes mRNA	12	RNA metabolism
KSRP (KHSRP) binds and destabilizes mRNA	12	RNA metabolism
Respiratory electron transport	12	Metabolism
ERBB2 Regulates Cell Motility	12	Signaling
ERBB2 Activates PTK6 Signaling	12	Signaling
Signaling by PTK6	12	Signaling
Downregulation of ERBB2 signaling	12	Signaling
Axon guidance	12	Axon guidance and projection
Non-alcoholic fatty liver disease (NAFLD)	12	Metabolism
Alzheimer's disease	12	Synapse function
Parkinson's disease	12	Synapse function
regulation of GTPase activity	12	Metabolism
positive regulation of GTPase activity	12	Metabolism
regulation of Ras protein signal transduction	12	Signaling
ephrin receptor signaling pathway	12	Signaling
regulation of small GTPase mediated signal transduction	12	Signaling
negative regulation of intracellular signal transduction	12	Signaling
MAPK3 (ERK1) activation	13	Signaling
PKA-mediated phosphorylation of CREB	13	Signaling
Calmodulin induced events	13	Synapse function
Ca-dependent events	13	Synapse function
CaM pathway	13	Synapse function
RAF-independent MAPK1/3 activation	13	Signaling
MAPK1 (ERK2) activation	13	Signaling
Prolactin receptor signaling	13	Signaling
DAG and IP3 signaling	13	Signaling
Glucagon signaling in metabolic regulation	13	Metabolism
PKA activation	13	Signaling
Integration of energy metabolism	13	Metabolism
PKA activation in glucagon Signaling	13	Signaling
Activation of AKT2	13	Signaling
PLC-gamma1 Signaling	13	Signaling

Adenylate cyclase activating pathway	13	Signaling
Neurophilin interactions with VEGF and VEGFR	13	Synapse function
EGFR interacts with phospholipase C-gamma	13	Signaling
SHC-related events triggered by IGF1R	13	Signaling
Integrin alphaIIb beta3 signaling	13	Signaling
GRB2:SOS provides linkage to MAPK signaling for Integrins	13	Signaling
p130Cas linkage to MAPK signaling for integrins	13	Signaling
Semaphorin interactions	13	Axon guidance and projection
CTLA4 inhibitory signaling	13	Signaling
Sema3A PAK dependent Axon repulsion	13	Axon guidance and projection
SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion	13	Axon guidance and projection
CRMPs in Sema3A signaling	13	Axon guidance and projection
Other semaphorin interactions	13	Axon guidance and projection
Regulation of insulin secretion	13	Metabolism
Transport of glycerol from adipocytes to the liver by Aquaporins	13	Metabolism
Passive transport by Aquaporins	13	Phagocytose and endocytose
RSK activation	13	Signaling
Signal transduction by L1	13	Axon guidance and projection
Aquaporin-mediated transport	13	Phagocytose and endocytose
VEGFR2 mediated vascular permeability	13	Phagocytose and endocytose
MAP2K and MAPK activation	13	Signaling
Negative regulation of MAPK pathway	13	Signaling
IL-6-type cytokine receptor ligand interactions	13	Signaling
Signaling by moderate kinase activity BRAF mutants	13	Signaling
Signaling by high-kinase activity BRAF mutants	13	Signaling

RAS signaling downstream of NF1 loss-of-function variants	13	Signaling
Paradoxical activation of RAF signaling by kinase inactive BRAF	13	Signaling
MET activates RAS signaling	13	Signaling
Signaling by FGFR3 fusions in cancer	13	Signaling
Rap1 signaling pathway	13	Signaling
Phospholipase D signaling pathway	13	Signaling
Oocyte meiosis	13	DNA repair and metabolism
Longevity regulating pathway	13	DNA repair and metabolism
Longevity regulating pathway	13	DNA repair and metabolism
Adrenergic signaling in cardiomyocytes	13	Signaling
Apelin signaling pathway	13	Synapse function
Retrograde endocannabinoid signaling	13	Synapse function
Glutamatergic synapse	13	Synapse function
Cholinergic synapse	13	Synapse function
GABAergic synapse	13	Synapse function
Inflammatory mediator regulation of TRP channels	13	Synapse function
Insulin secretion	13	Metabolism
GnRH signaling pathway	13	Signaling
Progesterone-mediated oocyte maturation	13	Signaling
Estrogen signaling pathway	13	Signaling
Thyroid hormone synthesis	13	Metabolism
Regulation of lipolysis in adipocytes	13	Metabolism
Aldosterone synthesis and secretion	13	Metabolism
Endocrine and other factor-regulated calcium reabsorption	13	Metabolism
Vasopressin-regulated water reabsorption	13	Phagocytose and endocytose
Salivary secretion	13	Metabolism
Gastric acid secretion	13	Metabolism
Pancreatic secretion	13	Metabolism
Bile secretion	13	Metabolism
Morphine addiction	13	Synapse function

Dilated cardiomyopathy	13	Metabolism
purine nucleotide biosynthetic process	13	DNA repair and metabolism
purine ribonucleotide biosynthetic process	13	DNA repair and metabolism
nucleotide biosynthetic process	13	DNA repair and metabolism
cyclic nucleotide metabolic process	13	DNA repair and metabolism
cyclic nucleotide biosynthetic process	13	DNA repair and metabolism
hormone-mediated signaling pathway	13	Signaling
intracellular receptor signaling pathway	13	Signaling
steroid hormone mediated signaling pathway	13	Signaling
cAMP metabolic process	13	Metabolism
ribose phosphate biosynthetic process	13	DNA repair and metabolism
cyclic purine nucleotide metabolic process	13	DNA repair and metabolism
purine-containing compound biosynthetic process	13	DNA repair and metabolism
nucleoside phosphate biosynthetic process	13	DNA repair and metabolism
Interleukin-7 signaling	15	Signaling
Role of phospholipids in phagocytosis	15	Phagocytose and endocytose
Tie2 Signaling	15	Signaling
Signaling by MET	15	Signaling
Negative regulation of MET activity	15	Signaling
InlB-mediated entry of Listeria monocytogenes into host cell	15	Virus life cycle and infection
Listeria monocytogenes entry into host cells	15	Virus life cycle and infection
Shigellosis	15	Virus life cycle and infection
Toxoplasmosis	15	Virus life cycle and infection

Amoebiasis	15	Virus life cycle and infection
Pancreatic cancer	15	Metabolism
Prostate cancer	15	Metabolism
Chronic myeloid leukemia	15	Metabolism
Central carbon metabolism in cancer	15	Metabolism
response to peptide	15	Metabolism
cellular response to peptide	15	Metabolism
Insulin processing	19	Metabolism
VxPx cargo-targeting to cilium	19	Transport and cytoskeleton
Cargo trafficking to the periciliary membrane	19	Transport and cytoskeleton
Trafficking of myristoylated proteins to the cilium	19	Transport and cytoskeleton
TRAF3-dependent IRF activation pathway	19	Signaling
TRAF6 mediated IRF7 activation	19	Signaling
RIG-I-like receptor signaling pathway	19	Signaling
Macroautophagy	21	Phagocytose and endocytose
Pentose and glucuronate interconversions	23	Drug metabolism and detoxification
Ascorbate and aldarate metabolism	23	Drug metabolism and detoxification
Glucuronidation	23	Drug metabolism and detoxification
Retinol metabolism	23	Drug metabolism and detoxification
Porphyrin and chlorophyll metabolism	23	Drug metabolism and detoxification
Metabolism of xenobiotics by cytochrome P450	23	Drug metabolism and detoxification
Drug metabolism	23	Drug metabolism and detoxification
Drug metabolism	23	Drug metabolism and detoxification
Activation of TRKA receptors	32	Signaling

TRKA activation by NGF	32	Signaling
Axonal growth inhibition (RHOA activation)	32	Axon guidance and projection
Regulated proteolysis of p75NTR	32	Axon guidance and projection
p75NTR regulates axonogenesis	32	Axon guidance and projection
Signaling to STAT3	32	Axon guidance and projection
NFG and proNGF binds to p75NTR	32	Axon guidance and projection
Axonal growth stimulation	32	Axon guidance and projection
Glutathione conjugation	35	Drug metabolism and detoxification
Glutathione synthesis and recycling	35	Drug metabolism and detoxification
Glutathione metabolism	35	Drug metabolism and detoxification
Metabolism of xenobiotics by cytochrome P450	35	Drug metabolism and detoxification
Drug metabolism	35	Drug metabolism and detoxification
Chemical carcinogenesis	35	Metabolism
Type I hemidesmosome assembly	37	Transport and cytoskeleton

Table 11: Specific biological terms of the ID mainland, the cluster in which they are found and the category attributed.

Appendix B

Biological terms of ASD mainland

Biological terms	Cluster ID	Category
Transport of the SLBP independent Mature mRNA	1	RNA transport and metabolism
Transport of the SLBP Dependant Mature mRNA	1	RNA transport and metabolism
Transport of Mature mRNA Derived from an Intronless Transcript	1	RNA transport and metabolism
Transport of Mature mRNAs Derived from Intronless Transcripts	1	RNA transport and metabolism
Transport of Ribonucleoproteins into the Host Nucleus	1	Virus cycle life and infection
Viral Messenger RNA Synthesis	1	Virus cycle life and infection
Regulation of Glucokinase by Glucokinase Regulatory Protein	1	Metabolism
Interactions of Vpr with host cellular proteins	1	Virus cycle life and infection
Nuclear import of Rev protein	1	Virus cycle life and infection
Vpr-mediated nuclear import of PICs	1	Virus cycle life and infection
Nuclear Envelope Breakdown	1	Transport and cytoskeleton
Nuclear Pore Complex (NPC) Disassembly	1	Transport and cytoskeleton

tRNA processing in the nucleus	1	RNA transport and metabolism
Mitotic Prophase	1	DNA repair and metabolism
Glucose transport	1	Metabolism
Interleukin-7 signaling	6	Signalling
Generation of second messenger molecules	6	Signalling
Role of phospholipids in phagocytosis	6	Phagocytose and endocytose
Tie2 Signaling	6	Signalling
Costimulation by the CD28 family	6	Immunology
CD28 co-stimulation	6	Immunology
CD28 dependent PI3K/Akt signaling	6	Immunology
Interferon gamma signaling	6	Immunology
Growth hormone receptor signaling	6	Signalling
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	6	Immunology
Th1 and Th2 cell differentiation	6	Immunology
Th17 cell differentiation	6	Immunology
Intestinal immune network for IgA production	6	Immunology
Adipocytokine signaling pathway	6	Signalling
Type II diabetes mellitus	6	Metabolism
Type I diabetes mellitus	6	Metabolism
Leishmaniasis	6	Virus cycle life and infection
Autoimmune thyroid disease	6	Immunology
Allograft rejection	6	Immunology
Graft-versus-host disease	6	Immunology
Viral myocarditis	6	Virus cycle life and infection
antigen processing and presentation	6	Immunology

Table 12: Specific biological terms of the ASD mainland, the cluster in which they are found and the category attributed.

Appendix C

Biological terms of the monogenic mainland

Biological terms	**Cluster ID**	**Category**
Lysosphingolipid and LPA receptors	2	Metabolism
hormone transport	2	Hormone secretion and transport
hormone secretion	2	Hormone secretion and transport
regulation of hormone secretion	2	Hormone secretion and transport
Transport of the SLBP independent Mature mRNA	3	RNA transport and metabolism
Transport of the SLBP Dependant Mature mRNA	3	RNA transport and metabolism
Transport of Mature mRNA Derived from an Intronless Transcript	3	RNA transport and metabolism
Transport of Mature mRNAs Derived from Intronless Transcripts	3	RNA transport and metabolism
Transport of Ribonucleoproteins into the Host Nucleus	3	Virus life cycle and infection
Viral Messenger RNA Synthesis	3	Virus life cycle and infection
Regulation of Glucokinase by Glucokinase Regulatory Protein	3	Metabolism
Interactions of Vpr with host cellular proteins	3	Virus life cycle and infection

Nuclear import of Rev protein	3	Virus life cycle and infection
Vpr-mediated nuclear import of PICs	3	Nuclear transport
Starch and sucrose metabolism	3	Metabolism
Nuclear Envelope Breakdown	3	Transport and cytoskeleton
Nuclear Pore Complex (NPC) Disassembly	3	Transport and cytoskeleton
tRNA processing in the nucleus	3	RNA transport and metabolism
Mitotic Prophase	3	DNA repair and metabolism
Glucose transport	3	Metabolism
Alpha-defensins	11	Immunology
Binding and entry of HIV virion	11	Virus life cycle and infection
Phosphorylation of CD3 and TCR zeta chains	11	Immunology
Generation of second messenger molecules	11	Signaling
PD-1 signaling	11	Immunology
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	11	Immunology
Th1 and Th2 cell differentiation	11	Immunology
Intestinal immune network for IgA production	11	Immunology
Type I diabetes mellitus	11	Immunology
Leishmaniasis	11	Virus life cycle and infection
Staphylococcus aureus infection	11	Virus life cycle and infection
Autoimmune thyroid disease	11	Immunology
Allograft rejection	11	Immunology
Graft-versus-host disease	11	Immunology
Viral myocarditis	11	Virus life cycle and infection
antigen processing and presentation	11	Immunology
Translesion synthesis by Y family DNA polymerases bypasses lesions on DNA template	13	DNA repair and metabolism
Signaling by ERBB2	13	Signaling

Signaling by ERBB4	13	Signaling
SHC1 events in ERBB2 signaling	13	Signaling
PI3K events in ERBB4 signaling	13	Signaling
SHC1 events in ERBB4 signaling	13	Signaling
PLCG1 events in ERBB2 signaling	13	Signaling
Nuclear signaling by ERBB4	13	Signaling
GRB7 events in ERBB2 signaling	13	Signaling
Downregulation of ERBB2:ERBB3 signaling	13	Signaling
Signaling by EGFR in Cancer	13	Signaling
Activation of AKT2	13	Signaling
GRB2 events in ERBB2 signaling	13	Signaling
PI3K events in ERBB2 signaling	13	Signaling
SHC-related events triggered by IGF1R	13	Signaling
Integrin alphaIIb beta3 signaling	13	Signaling
CTLA4 inhibitory signaling	13	Signaling
Constitutive Signaling by EGFRvIII	13	Signaling
Signaling by EGFRvIII in Cancer	13	Signaling
Signaling by Ligand-Responsive EGFR Variants in Cancer	13	Signaling
Signaling by Overexpressed Wild-Type EGFR in Cancer	13	Signaling
Inhibition of Signaling by Overexpressed EGFR	13	Signaling
Signaling by FGFR4 in disease	13	Signaling
Termination of translesion DNA synthesis	13	Signaling
MAP2K and MAPK activation	13	Signaling
Signaling by high-kinase activity BRAF mutants	13	Signaling
RAS signaling downstream of NF1 loss-of-function variants	13	Signaling
DNA Damage Bypass	13	DNA repair and metabolism
Platelet Aggregation (Plug Formation)	13	Metabolism
ERBB2 Activates PTK6 Signaling	13	Signaling
Signaling by PTK6	13	Signaling
MET activates RAS signaling	13	Signaling
Signaling by FGFR3 fusions in cancer	13	Signaling
PTK6 promotes HIF1A stabilization	13	Signaling
Downregulation of ERBB2 signaling	13	Signaling
Negative regulators of RIG-I/MDA5 signaling	13	Signaling

Nucleotide excision repair	13	DNA repair and metabolism
regulation of body fluid levels	13	Metabolism
PKB-mediated events	23	Signaling
PI3K Cascade	23	Signaling
Signaling by FGFR in disease	23	Signaling
mTOR signalling	23	Signaling
mTORC1-mediated signalling	23	Signaling
Signaling by cytosolic FGFR1 fusion mutants	23	Signaling
FGFR1 mutant receptor activation	23	Signaling
Signaling by FGFR1 in disease	23	Signaling
Macroautophagy	26	Phagocytose and endocytose

Table 13: Specific biological terms of the monogenic mainland, the cluster in which they are found and the category attributed.

Bibliography

[1] Consortium, I.H.G.S. Initial sequencing and analysis of the human genome. *Nature*, 409 (6822):860–921, Feb 2001.

[2] Project, H.G. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL `http://science.sciencemag.org/content/291/5507/1304`.

[3] Institute, N.H.G.R. The cost of sequencing a human genome, 07 2016. URL `https://www.genome.gov/sequencingcosts/`.

[4] Londin, E. et al. *Use of Linkage Analysis, Genome-Wide Association Studies, and Next-Generation Sequencing in the Identification of Disease-Causing Mutations*, pages 127–146. Humana Press, Totowa, NJ, 2013. ISBN 978-1-62703-435-7. doi: 10.1007/978-1-62703-435-7{_}8. URL `http://dx.doi.org/10.1007/978-1-62703-435-7_8`.

[5] Badano, J.L. and Katsanis, N. Beyond mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet*, 3(10):779–789, 10 2002. URL `http://dx.doi.org/10.1038/nrg910`.

[6] Barabasi, A.L. and Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 02 2004. URL `http://dx.doi.org/10.1038/nrg1272`.

[7] Jinawath, N. et al. Bridging the gap between clinicians and systems biologists: from network biology to translational biomedical research. *Journal of Translational Medicine*, 14(1):324, 2016. doi: 10.1186/s12967-016-1078-3. URL `http://dx.doi.org/10.1186/s12967-016-1078-3`.

[8] Hu, W.F., Chahrour, M.H. and Walsh, C.A. The diverse genetic landscape of neurodevelopmental disorders. *Annual Review of Genomics and Human Genetics*, 15(1):195–213, 2017/05/14 2014. doi: 10.1146/annurev-genom-090413-025600. URL `http://dx.doi.org/10.1146/annurev-genom-090413-025600`.

[9] Collins, F.S. et al. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 04 2003.

[10] Consortium, T..G.P. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 11 2012. URL http://dx.doi.org/10.1038/nature11632.

[11] Kong, A. et al. Rate of de novo mutations, father's age, and disease risk. *Nature*, 488 (7412):471–475, 08 2012. doi: 10.1038/nature11396. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3548427/.

[12] Campbell, C.D. and Eichler, E.E. Properties and rates of germline mutations in humans. *Trends in Genetics*, 29(10):575–584, 2017/05/13 2013. doi: 10.1016/j.tig.2013.04.005. URL http://dx.doi.org/10.1016/j.tig.2013.04.005.

[13] Feuk, L., Carson, A.R. and Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97, 02 2006. URL http://dx.doi.org/10.1038/nrg1767.

[14] Frazer, K.A. et al. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–251, 04 2009. URL http://dx.doi.org/10.1038/nrg2554.

[15] Manolio, T.A., Brooks, L.D. and Collins, F.S. A hapmap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605, 05 2008. doi: 10.1172/JCI34772. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2336881/.

[16] Consortium, I.H. The international hapmap project. *Nature*, 426(6968):789–796, 12 2003.

[17] Mayeux, R. Mapping the new frontier: complex genetic disorders. *Journal of Clinical Investigation*, 115(6):1404–1407, 06 2005. doi: 10.1172/JCI25421. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1137013/.

[18] Rosenfeld, J.A., Malhotra, A.K. and Lencz, T. Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. *Nucleic Acids Research*, 38(18):6102–6111, 10 2010. doi: 10.1093/nar/gkq408. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2952858/.

[19] Montgomery, S.B. et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Research*, 23(5):749–761, 05 2013. doi: 10.1101/gr.148718.112. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3638132/.

[20] Wajnberg, G. and Passetti, F. Using high-throughput sequencing transcriptome data for indel detection: challenges for cancer drug discovery. *Expert Opinion on Drug Discovery*, 11(3):257–268, 2016. doi: 10.1517/17460441.2016.1143813. URL http://dx.doi.org/10.1517/17460441.2016.1143813. PMID: 26787005.

[21] Feuk, L. Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine*, 2(2):11–11, 2010. doi: 10.1186/gm132. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2847702/.

[22] Sanders, A.D. et al. Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Research*, 26(11):1575–1587, 11 2016. doi: 10.1101/gr.201160.115. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088599/.

[23] Inoue, K. and Lupski, J.R. Molecular mechanisms for genomic disorders. *Annual Review of Genomics and Human Genetics*, 3(1):199–242, 2002. doi: 10.1146/annurev.genom.3.032802.120023. URL http://dx.doi.org/10.1146/annurev.genom.3.032802.120023. PMID: 12142364.

[24] McCarroll, S.A. and Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat Genet*, 06 2007.

[25] Zarrei, M. et al. A copy number variation map of the human genome. *Nat Rev Genet*, 16 (3):172–183, 03 2015. URL http://dx.doi.org/10.1038/nrg3871.

[26] Riazuddin, S.A. et al. Missense mutations in tcf8 cause late-onset fuchs corneal dystrophy and interact with fcd4 on chromosome 9p. *The American Journal of Human Genetics*, 86(1):45–53, 2017/05/14 2009. doi: 10.1016/j.ajhg.2009.12.001. URL http://dx.doi.org/10.1016/j.ajhg.2009.12.001.

[27] Agarwal, S. and Moorchung, N. Modifier genes and oligogenic disease. *Journal of Nippon Medical School*, 72(6):326–334, 2005. doi: 10.1272/jnms.72.326.

[28] Zhu, X. et al. One gene, many neuropsychiatric disorders: lessons from mendelian diseases. *Nat Neurosci*, 17(6):773–781, 06 2014. URL http://dx.doi.org/10.1038/nn.3713.

[29] Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat Genet*, 46(8):881–885, 08 2014. URL http://dx.doi.org/10.1038/ng.3039.

[30] Schork, N.J. et al. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, 19(3):212–219, 6 2009. doi: https://doi.org/10.1016/j.gde.2009.04.010. URL http://www.sciencedirect.com/science/article/pii/S0959437X09000884.

[31] Cornish, K. et al. Annotation: Deconstructing the attention deficit in fragile x syndrome: a developmental neuropsychological approach. *Journal of Child Psychology and Psychiatry*, 45(6):1042–1053, 2004. ISSN 1469-7610. doi: 10.1111/j.1469-7610.2004.t01-1-00297.x. URL http://dx.doi.org/10.1111/j.1469-7610.2004.t01-1-00297.x.

[32] Weijerman, M.E. and de Winter, J.P. Clinical practice. *European Journal of Pediatrics*, 169(12):1445–1452, 2010. ISSN 1432-1076. doi: 10.1007/s00431-010-1253-0. URL http://dx.doi.org/10.1007/s00431-010-1253-0.

[33] Amir, R.E. et al. Rett syndrome is caused by mutations in x-linked mecp2, encoding methyl-cpg-binding protein 2. *Nat Genet*, 23(2):185–188, 10 1999. URL http://dx.doi.org/10.1038/13810.

[34] Buiting, K. Prader–willi syndrome and angelman syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 154C(3):365–376, 2010. ISSN 1552-4876. doi: 10.1002/ajmg.c.30273. URL http://dx.doi.org/10.1002/ajmg.c.30273.

[35] Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823):445–449, 04 2007. doi: 10.1126/science.1138659. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2993504/.

[36] Marshall, C.R. et al. Structural variation of chromosomes in autism spectrum disorder. *American Journal of Human Genetics*, 82(2):477–488, 02 2008. doi: 10.1016/j.ajhg.2007.12.009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2426913/.

[37] Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299, 04 2012. doi: 10.1016/j.neuron.2012.04.009. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3619976/.

[38] Le Couteur, A. et al. A broader phenotype of autism: The clinical spectrum in twins. *Journal of Child Psychology and Psychiatry*, 37(7):785–801, 1996. ISSN 1469-7610. doi: 10.1111/j.1469-7610.1996.tb01475.x. URL http://dx.doi.org/10.1111/j.1469-7610.1996.tb01475.x.

[39] Folstein, S. and Rutter, M. Infantile autism: A genetic study of 21 twin pairs. *Journal of Child Psychology and Psychiatry*, 18(4):297–321, 1977. ISSN 1469-7610. doi: 10.1111/j.1469-7610.1977.tb00443.x. URL http://dx.doi.org/10.1111/j.1469-7610.1977.tb00443.x.

[40] Anney, R. et al. A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, 19(20):4072–4082, 10 2010. doi: 10.1093/hmg/ddq307. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2947401/.

[41] Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat Genet*, 46(8):881–885, 08 2014. URL http://dx.doi.org/10.1038/ng.3039.

[42] Uetz, P. et al. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 02 2000. URL http://dx.doi.org/10.1038/35001009.

[43] Scharer, C.D. et al. Genome-wide promoter analysis of the sox4 transcriptional network in prostate cancer cells. *Cancer research*, 69(2):709–717, 01 2009. doi: 10.1158/0008-5472.CAN-08-3415. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2629396/.

[44] Lu, P. et al. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotech*, 25(1):117–124, 01 2007. URL http://dx.doi.org/10.1038/nbt1270.

[45] Ruan, J., Dean, A.K. and Zhang, W. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Systems Biology*, 4:8–8, 2010. doi: 10.1186/1752-0509-4-8. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2829495/.

[46] Boucher, B. and Jenna, S. Genetic interaction networks: better understand to better predict. *Frontiers in Genetics*, 4:290, 2013. doi: 10.3389/fgene.2013.00290. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3865423/.

[47] Goh, K.I. et al. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690, 05 2007. doi: 10.1073/pnas.0701361104. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1885563/.

[48] Ma, X. and Gao, L. Biological network analysis: insights into structure and functions. *Briefings in Functional Genomics*, 11(6):434, 2012. doi: 10.1093/bfgp/els045. URL +http://dx.doi.org/10.1093/bfgp/els045.

[49] Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 11 2014. doi: 10.1016/j.cell.2014.10.050. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4266588/.

[50] Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 10 2005. URL http://dx.doi.org/10.1038/nature04209.

[51] Chatr-aryamontri, A. et al. The biogrid interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, 01 2017. URL http://dx.doi.org/10.1093/nar/gkw1102.

[52] Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 10 2003. doi: 10.1101/gr.1680803. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC403728/.

[53] Zanzoni, A. et al. Mint: a molecular interaction database. *FEBS Letters*, 513(1):135–140, 2002. ISSN 1873-3468. doi: 10.1016/S0014-5793(01)03293-8. URL http://dx.doi.org/10.1016/S0014-5793(01)03293-8.

[54] Orchard, S. et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363, 01 2014. URL http://dx.doi.org/10.1093/nar/gkt1115.

[55] Li, T. et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Meth*, 14(1):61–64, 01 2017. URL http://dx.doi.org/10.1038/nmeth.4083.

[56] Berg, J.M. and Geschwind, D.H. Autism genetics: searching for specificity and convergence. *Genome Biology*, 13(7):247–247, 2012. doi: 10.1186/gb-2012-13-7-247. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491377/.

[57] Lin, C.C. et al. Dynamic functional modules in co-expressed protein interaction networks of dilated cardiomyopathy. *BMC Systems Biology*, 4:138–138, 2010. doi: 10.1186/1752-0509-4-138. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2978157/.

[58] Li, J. et al. Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Molecular Systems Biology*, 10(12):774, 12 2014. doi: 10.15252/msb.20145487. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300495/.

[59] Hormozdiari, F. et al. The discovery of integrated gene networks for autism and related disorders. *Genome Research*, 11 2014. URL http://genome.cshlp.org/content/early/2014/11/05/gr.178855.114.

[60] Gilman, S.R. et al. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, 70 (5):898–907, 2017/05/15 2011. doi: 10.1016/j.neuron.2011.05.021. URL http://dx.doi.org/10.1016/j.neuron.2011.05.021.

[61] Ben-David, E. and Shifman, S. Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Mol Psychiatry*, 18(10):1054–1056, 10 2013. URL http://dx.doi.org/10.1038/mp.2012.148.

[62] O'Roak, B.J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397):246–250, 05 2012. doi: 10.1038/nature10989. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3350576/.

[63] Gulsuner, S. et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, 154(3):518–529, 08 2013. doi: 10.1016/j.cell.2013.06.049. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894107/.

[64] Parikshak, N.N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, 155(5):1008–1021, 11 2013. doi: 10.1016/j.cell.2013.10.031. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934107/.

[65] Willsey, A.J. et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, 155(5):997–1007, 11 2013. doi: 10.1016/j.cell.2013.10.020. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3995413/.

[66] Corominas, R. et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature Communications*, 5:3650, 04 2014. doi: 10.1038/ncomms4650. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996537/.

[67] Cristino, A.S. et al. Neurodevelopmental and neuropsychiatric disorders represent an interconnected molecular system. *Mol Psychiatry*, 19(3):294–301, 03 2014. URL http://dx.doi.org/10.1038/mp.2013.16.

[68] Study, T.D.D.D. Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, 519(7542):223–228, 03 2015. URL http://dx.doi.org/10.1038/nature14135.

[69] Köhler, S. et al. The human phenotype ontology in 2017. *Nucleic Acids Research*, 45 (D1):D865–D876, 01 2017. URL http://dx.doi.org/10.1093/nar/gkw1039.

[70] Li, H. et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16): 2078–2079, 08 2009. doi: 10.1093/bioinformatics/btp352. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/.

[71] Albers, C.A. et al. Dindel: Accurate indel calls from short-read data. *Genome Research*, 21(6):961–973, 06 2011. doi: 10.1101/gr.112326.110. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3106329/.

[72] McKenna, A. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, 09 2010. doi: 10.1101/gr.107524.110. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/.

[73] McLaren, W. et al. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics*, 26(16):2069–2070, 08 2010. doi: 10.1093/bioinformatics/btq330. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2916720/.

[74] Helaers, R. Highlander: variant filtering made easy. URL http://sites.uclouvain.be/highlander/index.html.

[75] Liu, X., Jian, X. and Boerwinkle, E. dbnsfp: A lightweight database of human nonsynonymous snps and their functional predictions. *Human Mutation*, 32(8):894–899, 2011. ISSN 1098-1004. doi: 10.1002/humu.21517. URL http://dx.doi.org/10.1002/humu.21517.

[76] Liu, X., Jian, X. and Boerwinkle, E. dbnsfp v2.0: A database of human non-synonymous snvs and their functional predictions and annotations. *Human Mutation*, 34(9):E2393–E2402, 2013. ISSN 1098-1004. doi: 10.1002/humu.22376. URL http://dx.doi.org/10.1002/humu.22376.

[77] Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly*, 6(2):80–92, 04 2012. doi: 10.4161/fly.19695. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679285/.

[78] Fraley, C. and Raftery, A.E. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

[79] Bader, G.D. and Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003. doi: 10.1186/1471-2105-4-2. URL http://dx.doi.org/10.1186/1471-2105-4-2.

[80] Consortium, T.G.O. et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 05 2000. doi: 10.1038/75556. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/.

[81] Kanehisa, M. et al. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(Database issue):D353–D361, 01 2017. doi: 10.1093/nar/gkw1092. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210567/.

[82] Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl_1):D428–D432, 01 2005. URL http://dx.doi.org/10.1093/nar/gki072.

[83] Bindea, G. et al. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, 04 2009. doi: 10.1093/bioinformatics/btp101. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2666812/.

[84] Ng, P.C. et al. Genetic variation in an individual human exome. *PLOS Genetics*, 4(8):1–15, 08 2008. doi: 10.1371/journal.pgen.1000160. URL https://doi.org/10.1371/journal.pgen.1000160.

[85] Xue, Y. et al. Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing. *American Journal of Human Genetics*, 91(6):1022–1032, 12 2012. doi: 10.1016/j.ajhg.2012.10.015. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516590/.

[86] van Karnebeek, C.D.M. and Stockler, S. Treatable inborn errors of metabolism causing intellectual disability: A systematic literature review. *Molecular Genetics and Metabolism*, 105(3):368–381, 03 2012. doi: 10.1016/j.ymgme.2011.11.191. URL http://dx.doi.org/10.1016/j.ymgme.2011.11.191.

[87] Lacruz, R.S. and Feske, S. Diseases caused by mutations in orai1 and stim1. *Annals of the New York Academy of Sciences*, 1356(1):45–79, 11 2015. doi: 10.1111/nyas.12938. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4692058/.

[88] Griesi-Oliveira, K., Suzuki, A.M. and Muotri, A.R. *TRPC Channels and Mental Disorders*, pages 137–148. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-1088-4. doi: 10.1007/978-94-024-1088-4{_}12. URL http://dx.doi.org/10.1007/978-94-024-1088-4_12.

[89] Haas, H.L., Sergeeva, O.A. and Selbach, O. Histamine in the nervous system. *Physiological Reviews*, 88(3):1183, 07 2008. URL http://physrev.physiology.org/content/88/3/1183.abstract.

[90] El-Ansary, A. and Al-Ayadhi, L. Lipid mediators in plasma of autism spectrum disorders. *Lipids in Health and Disease*, 11(1):160, 2012. doi: 10.1186/1476-511X-11-160. URL http://dx.doi.org/10.1186/1476-511X-11-160.

[91] Qasem, H., Al-Ayadhi, L. and El-Ansary, A. Cysteinyl leukotriene correlated with 8-isoprostane levels as predictive biomarkers for sensory dysfunction in autism. *Lipids in Health and Disease*, 15:130, 2016. doi: 10.1186/s12944-016-0298-0. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988023/.

[92] Sadakata, T. et al. Analysis of gene expression in ca2+-dependent activator protein for secretion 2 (cadps2) knockout cerebellum using genechip and kegg pathways. *Neuroscience Letters*, 639:88–93, 2 2017. doi: https://doi.org/10.1016/j.neulet.2016.12.068. URL http://www.sciencedirect.com/science/article/pii/S0304394016310163.

[93] Wall, D.P. et al. Comparative analysis of neurological disorders focuses genome-wide search for autism genes. *Genomics*, 93(2):120–129, 2 2009. doi: https://doi.org/10.1016/j.ygeno.2008.09.015. URL http://www.sciencedirect.com/science/article/pii/S0888754308002292.

[94] Shang, L. et al. Mutations in arid2 are associated with intellectual disabilities. *neurogenetics*, 16(4):307–314, 2015. doi: 10.1007/s10048-015-0454-0. URL http://dx.doi.org/10.1007/s10048-015-0454-0.

[95] Bajpai, R. et al. Chd7 cooperates with pbaf to control multipotent neural crest formation. *Nature*, 463(7283):958–962, 02 2010. doi: 10.1038/nature08733. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2890258/.

[96] Gamsiz, E.D. et al. Discovery of rare mutations in autism: Elucidating neurodevelopmental mechanisms. *Neurotherapeutics*, 12(3):553–571, 07 2015. doi: 10.1007/s13311-015-0363-9. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489950/.

[97] Son, E.Y. and Crabtree, G.R. The role of baf (mswi/snf) complexes in mammalian neural development. *American journal of medical genetics. Part C, Seminars in medical genetics*, 0(3):333–349, 09 2014. doi: 10.1002/ajmg.c.31416. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4405377/.

[98] Shiflett, M.W., Gavin, M. and Tran, T.S. Altered hippocampal-dependent memory and motor function in neuropilin 2–deficient mice. *Translational Psychiatry*, 5(3):e521–, 03 2015. doi: 10.1038/tp.2015.17. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4354347/.

[99] Hosseinpour, M. et al. Neuropilin-2 rs849563 gene variations and susceptibility to autism in iranian population: A case-control study. *Metabolic Brain Disease*, pages 1–4, 2017. doi: 10.1007/s11011-017-0024-2. URL http://dx.doi.org/10.1007/s11011-017-0024-2.

[100] Macdonald, R.L., Kang, J.Q. and Gallagher, M.J. Mutations in gaba(a) receptor subunits associated with genetic epilepsies. *The Journal of Physiology*, 588(Pt 11):1861–1869, 06 2010. doi: 10.1113/jphysiol.2010.186999. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2901974/.

[101] Emre Onat, O. et al. Missense mutation in the atpase, aminophospholipid transporter protein atp8a2 is associated with cerebellar atrophy and quadrupedal locomotion. *European Journal of Human Genetics*, 21(3):281–285, 03 2013. doi: 10.1038/ejhg.2012.170. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3573203/.

[102] Duan, Y. et al. Semaphorin 5a inhibits synaptogenesis in early postnatal- and adult-born hippocampal dentate granule cells. *eLife*, 3:e04390, 2014. doi: 10.7554/eLife.04390. URL https://doi.org/10.7554/eLife.04390.