UNIVERSIDAD DE MÁLAGA

Análisis de relaciones fenotípicas mediante el uso de ontologías biomédicas: Aplicaciones a enfermedades genéticas

TESIS DOCTORAL

Rocío Rodríguez López

Directores:
Francisca Sánchez Jiménez y
Armando Reyes Palomares

UNIVERSIDAD DE MÁLAGA

AUTOR: Rocío Rodríguez López

http://orcid.org/0000-0002-0518-9119

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

Análisis de relaciones fenotípicas mediante el uso de ontologías biomédicas: Aplicaciones a enfermedades genéticas

Rocío Rodríguez López

Noviembre 2017

Tesis Doctoral

Departamento de Biología Molecular y Bioquímica
Universidad de Málaga

Director 1: Francisca Sánchez Jiménez
Director 2: Armando Reyes Palomares

A QUIEN CORRESPONDA,

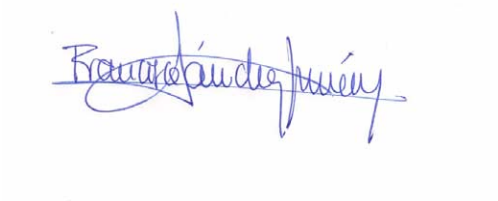
FRANCISCA MARÍA SÁNCHEZ JIMÉNEZ, DNI 45062521R, Catedrática de Bioquímica y Biología Molecular de la Universidad de Málaga e IP del grupo PAIDI BIO-267 y la Unidad 741 de Ciberer (ISCIII) y ARMANDO REYES PALOMARES, DNI 44593340C, Investigador Postdoctoral del *European Molecular Biology Laboratory* (EMBL) en Heidelberg (Alemania)

CERTIFICAN,

Que Dña. Rocío Rodríguez López, Licenciada en Ingeniería Informática por la Universidad de Málaga, ha realizado bajo nuestra dirección conjunta en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga el trabajo de investigación correspondiente a su Tesis Doctoral que lleva por título "***Análisis de relaciones fenotípicas mediante el uso de ontologías biomédicas: Aplicaciones a enfermedades genéticas***".

Este trabajo reúne, a nuestro juicio, contenido científico suficiente y las condiciones necesarias para ser presentado y defendido ante el tribunal correspondiente para optar al grado de Doctor.

Atentamente,



Esta Tesis Doctoral ha sido subvencionada por los proyectos SAF2011-26518, del Ministerio de Economía y Competitividad (MINECO), el proyecto PAIDI P10-CVI6585 (Junta de Andalucía), otros fondos del grupo BIO-267 (Junta de Andalucía), el CIBER de Enfermedades Raras (CIBERER) del Instituto de Salud Carlos III y el proyecto INTERCONECTA-AMER (CDTI, MINECO). Se agradece el apoyo a la Universidad de Málaga y el proyecto Andalucía-Tech. Parte de los resultados y metodologías recogidas en la presente memoria han dado lugar a las siguientes publicaciones:

- Molecular networks: Biomedical implications. Francisca Sánchez Jiménez, Almudena Pino Ángeles, Rocío Rodríguez López, María Morales, José Luis Urdiales. Pharmacological Research (2016), 114, 90-102.

- Histamine and its receptors as a module of the biogenic amine diseasome. Rocío Rodríguez López, María Morales, Francisca Sánchez Jiménez. Histamine Receptors (2016), 28, 173-214.

- Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. Armando Reyes-Palomares, Aníbal Bueno, Rocío Rodríguez López, Miguel Ángel Medina, Francisca Sánchez Jiménez, Manuel Corpas, Juan Antonio García Ranea. BMC Genomics (2016), 17, 232.

- PhenUMA: a tool for integrating the biomedical relationships among genes and diseases. Rocío Rodríguez López, Armando Reyes Palomares, Francisca Sánchez Jiménez, Miguel Ángel Medina. BMC Bioinformatics (2014), 15, 375.

- Global analysis of the human pathophenotypic similarity gene network merges disease module components. Armando Reyes Palomares, Rocío Rodríguez López, Juan Antonio García Ranea, Francisca Sánchez Jiménez, Miguel Ángel Medina. PLOS ONE (2013), 8-2, e56653.

Y a las siguientes comunicaciones:

- Studying drug response variability using network-based approaches (Póster). Rocío Rodríguez López, Christian Arnold, Armando Reyes Palomares, Francisca Sánchez Jiménez, Judith B. Zaugg. International Conference on Systems Biology, Septiembre 2016, Barcelona, España.

- Network approaches and gene expression profiles to study drug response (Póster). Rocío Rodríguez López, Christian Arnold, Armando Reyes Palomares, Francisca Sánchez Jiménez, Judith B. Zaugg. EMBL Partnership Conference Perspectives in Translational Medicine, Junio 2016, Heidelberg, Alemania.

- Visibility of amine-related elements in human diseasomes (Charla). Rocío Rodríguez López, Armando Reyes Palomares, Miguel Ángel Medina, Francisca Sánchez Jiménez. 44th European Histamine Research Society Meeting, Mayo 2015, Torremolinos (Málaga), España.

- PhenUMA: a biomedical tool for the integration and visualization of phenotypic relationships among genes (Póster). Armando Reyes Palomares, Rocío Rodríguez López, Miguel Ángel Medina Torres, Francisca Sánchez Jiménez. 37th FEBS, 22nd IUBMB and 35th SEBBM Congress, Septiembre 2012, Sevilla, España.

- PhenUMA: web tool for integration of phenotypical and functional information (Póster). Rocío Rodríguez López, Armando Reyes Palomares, Miguel Ángel Medina, Francisca Sánchez Jiménez. II European Conference: "Genomics of Complex Diseases: New Challenges", Abril 2014, Málaga, España.

- The human pathophenotypic network (Póster). Armando Reyes Palomares, Rocío Rodríguez López, Miguel Ángel Medina, Francisca Sánchez Jiménez. 37th FEBS, 22nd IUBMB and 35th SEBBM Congress, Septiembre 2012, Sevilla, España.

Agradecimientos

En primer lugar, quiero mostrar mi agradecimiento a la Universidad de Málaga por proporcionar recursos, formación e instalaciones que han hecho posible el desarrollo mi carrera académica. En segundo lugar, le debo un especial agradecimiento a mis directores: Kika y Armando. Kika, sin quien esta tesis no hubiera sido posible, me ayudó a superar cada una de las dificultades e hizo todo lo posible para que pudiera culminar mi carrera académica. Por esto y por todo el trabajo a lo largo de estos años: muchas gracias! Espero que la disfrutes de la mejor forma posible esta nueva etapa que acaba de comenzar. Armando, mi compañero de equipo y director, es protagonista indiscutible de este trabajo. Sus propuestas, proyectos e ideas han sido de gran inspiración en mi desarrollo profesional. Muchas gracias por transmitirme tus conocimientos y por todo lo que me has enseñado. No dudo en que te espera una prometedora trayectoria científica, la cual ya ha comenzado como se merece en unos de los mejores centros del mundo. Por último, una mención especial a Miguel Ángel por su colaboración y apoyo, a mí y a mis directores, a lo largo del desarrollo de esta tesis.

Special thanks to Judith Zaugg who gave me the opportunity to join her group. Thanks for the valuable comments and suggestions and thank you for your time and support. I also want to express my gratitude to the members of Zaugg's group: Christian, Mariana, Ivan, Ignacio, Pooja, and Ece. It was really nice working with you, and I wish all of you the best of luck and every success in the future.

A mis compañeros de laboratorio, principalmente a Bea y Aníbal, con quien he compartido penas, alegrías y largas horas de trabajo. Nuestras experiencias, viajes y largas conversaciones han hecho mucho más fácil esta etapa. Muchas gracias por todo. A Raúl, de quien he aprendido mucho de nuestras discusiones y de la forma que tiene de ver la ciencia. Una mención especial a María, por su colaboración en parte del contenido de esta tesis y por elegirnos para realizar su TFG. Finalmente, gracias a todos aquellos con los que he compartido laboratorio durante estos seis años: Aurelio, Almudena, Ian, James, Mariavi, Juan Antonio, Pedro, David, Luz y Cristóbal. Y mucha suerte a la nueva generación

Elena, Nando y José Joaquín. Por último agradecer este trabajo a mi familia y amigos por su paciencia y su apoyo a lo largo de todo este proceso.

Lista de Abreviaturas

5'-HT Serotonina

ALDH5A1 Aldehído deshidrogenasa 5 miembro A1

AOC1 Diamino oxidasa

AOC2 Amino oxidasa cobre dependiente 2

AOC3 Amino oxidasa cobre dependiente 3

ARG1 Arginasa 1

ARG2 Arginasa 2

AUC *Area under de curve* (Área bajo la curva)

AZIN1 Inhibidor de antizimas 1

AZIN2 Inhibidor de antizimas 2

BCKDH Deshidrogenasa de alfa-cetoácidos de cadena ramificada

BCKDHA Deshidrogenasa de alfa-cetoácidos de cadena ramificada E1, polipéptido alfa

BCKDHB Deshidrogenasa de alfa-cetoácidos de cadena ramificada E1, subunidad beta

BioNER *Biomedical named entity recognition*

BP *Biological process*

CC *Cellular component*

CGH *Comparative genomic hybridization*

CNV *Copy number variation*

DA Dopamina

DBT *Dihydrolipoamide branched chain transacylase E2*

DDC Dopa descarboxilasa

DDC L-Aminoácido aromático descarboxilasa

DECIPHER *Database of Chromosomal Imbalance and Phenotype in Humans using Esembl Resources*

DFMO Eflornitina

DGN *Disease gene network*

DGV *Database of genomic variants*

DLD Dihidrolipoil deshidrogenasa

DRD1 Receptor de dopamina 1

DRD2 Receptor de dopamina 2

DRD3 Receptor de dopamina 3

DRD4 Receptor de dopamina 4

DRD5 Receptor de dopamina 5

EBI *European bioinformatics institute*

EDNRA Receptor de endotelina tipo A

EEG Electroencefalograma

EHR *Electronic health records*

FDR *False discovery rate*

FORGE *Finding of rare disease genes*

FP Falsos positivos

GAD Grafo acíclico dirigido

GHR *Genetics home reference*

GO *Gene ontology*

GOA *Gene ontology annotation*

GPCR Receptor acoplado a proteínas G

GRIN1 Receptor de N-metil aspartator 1

GRIN2A Receptor de N-metil aspartator 2A

GRIN2B Receptor de N-metil aspartator 2B

GWAS *Genome wide association atudies*

HDC Histidina descarboxilasa

HDGN *Human disease gene network*

HDN *Human disease network*

Hia Histamina

HLGT *High level group term*

HLT *High livel term*

HNMT Histamina N-metil transferasa

HPO *Human phenotype ontology*

HRH1 Receptor de histamina 1

HRH2 Receptor de histamina 2

HRH3 Receptor de histamina 3

HRH4 Receptor de histamina 4

HTR1A Receptor de 5-hidroxitriptamina 1A

HTR1B Receptor de 5-hidroxitriptamina 1B

HTR2A Receptor de 5-hidroxitriptamina 2A

HTR2B Receptor de 5-hidroxitriptamina 2B

HTR2C Receptor de 5-hidroxitriptamina 2C

HTR3A Receptor de 5-hidroxitriptamina 3A

HTR3B Receptor de 5-hidroxitriptamina 3B

HTR3C Receptor de 5-hidroxitriptamina 3C

HTR3D Receptor de 5-hidroxitriptamina 3D

HTR3E Receptor de 5-hidroxitriptamina 3E

HTR4 Receptor de 5-hidroxitriptamina 4

HTR5A Receptor de 5-hidroxitriptamina 5A

HTR6 Receptor de 5-hidroxitriptamina 6

HTR7 Receptor de 5-hidroxitriptamina 7

IC *Information content* (Contenido Informativo)

IEA *Inferred from electronic annotation*

IRDiRC *International rare diseases research consortium*

KEGG *Kyoto encyclopedia of genes and genomes*

LLT *Lowest level term*

MAOA Monoamina oxidasa A

MAOB Monoamina oxidasa B

MD *Monogenic disease* (Enfermedad monogénica)

MeSH *Medical subject heading*

MF *Molecular fucntion*

MG *Monotropic gene* (Gen monotrópico)

MGN *Metabolic gene network*

MICA *Most informative common ancestor* (Ancestro común más informativo)

MSUD Enfermedad de orina con olor a jarabe de arce

NCBI *National centre of biotechnology information*

NIH *National institutes for health*

NLM *National library of medicine*

NMDA N-metil aspartato

OAZ1 Antizima 1 de la ornitina-descarboxilasa

OAZ2 Antizima 2 de la ornitina-descarboxilasa

OAZ3 Antizima 3 de la ornitina-descarboxilasa

OBO *Open biomedical ontologies*

ODC Ornitina descarboxilasa

ODC1 Ornitina descarboxilasa

ODGN *Orphan disease gene network*

ODN *Orphan disease network*

OMIM *Online mendelian inheritance in man*

PA Poliaminas

PAOX Poliamina oxidasa (PAOX)

PCCB Propionil-CoA carboxilasa subunidad beta

PD *Polygenic diseases* (Enfermedad poligénica)

PEL *Phenotypically enriched loci* (*Loci* fenotípicamente enriquecido)

PG *Pleiotropic gene* (Gen pleiotrópico)

PIN *Protein interaction network* (Red de interacción entre proteínas)

PINK1 *PTEN induced putative kinase 1*

PSGN *Pathophenotypic similarity gene network*

PT *Preferred term*

Put Putrescina

ROC *Receiver operating characteristic*

SAT1 Espmermidina/espermina acetil transferasa

SCL18A2 Transportador de soluto familia 18 miembro 2

SLC12A8 Transportador de soluto familia 12

SLC18A1 Transportador de soluto familia 18 miembro 1

SLC22A2 Transportador de soluto familia 22 miembro 2

SLC22A3 Transportador de soluto familia 22 miembro 3

SLC3A2 Transportador de soluto familia 3 miembro 2

SLC6A3 Transportador de soluto familia 6 miembro 3

SLC6A4 Transportador de soluto familia 6 miembro 4

SMOX Espermina oxidasa

SMS Espermina sintasa

SNCA Alfa-sinucleína

SNOMED CT *Systematized nomenclature of medicine – clinical terms*

SNP *Single nuclotide polymorphism*

SOC *System organ class*

Spd Espermidina

Spm Espermina

SRM Espermidina sintasa

SSADHD *Succinic semialdehyde dehydrogenase deficency*

SVM *Support vector machine*

TGM1 Transglutaminasa 1

TGM2 Transglutaminasa 2

TH Tirosina hidroxilasa

TNF Factor de necrosis tumoral

TPH1 Triptófano hidroxilasa

TPH2 Triptófano hidroxilasa 2

UMLS *Unified medical language system*

VP Verdaderos positivos

WES *Whole exome sequencing*

WGS *Whole genome sequencing*

Y2H *Yeast-two hybrid*

Summary

Introduction

Recent advances in high-throughput techniques, such as human genome sequencing, transcriptomics and proteomics have provided a massive amount of data. This has propelled the study of the genetic and molecular basis of human diseases. For instance, Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) has greatly contributed to the identification of the genetic causes for a high number of orphan diseases (Bamshad et al., 2011; Boycott et al., 2013; Ng et al., 2009, 2010) Genome-Wide Association Studies (GWAS) have also enabled the identification of thousands of *loci* associated with common diseases (Lee et al., 2014). However, the understanding of the molecular background for many genetic conditions is still a big challenge (MacArthur et al., 2014).

Research in rare diseases is especially complex due to the heterogeneity of etiologies and the small samples size of cohort of patients (Samuels 2010). Even when a 6%-8% of the population is affected by a rare disease, the number of individuals diagnosed with a specific rare disease is, sometimes, extremely low. This makes it difficult to determine the genetic and molecular basis of these diseases. To address these problems, several international platforms have been created to provide repositories of genomic and phenotypic data of rare disease patients. For example, DECIPHER (Database of Chromosomal Imbalance and Phenotype in Humans Using Esembl Resources) (Firth et al., 2009), RD-Connect (Thompson et al., 2014), Non-Diagnostic Diseases Program (Gahl et al., 2012), Canadian Care4Rare (`http://care4rare.ca/`) and FORGE (Finding of Rare Disease Genes) (Beaulieu et al., 2014) work to identify candidate genes related to rare diseases.

Access to data from patients diagnosed with rare diseases could be useful for studying relationships between phenotypes and genetic variants. Due to the phenotypic complexity of these pathologies, the analysis and integration of patient data together to their respective molecular and genetic information could enhance the identification of candidate genes, as well as shed light on the molecular bases involved in the pathology development. For this reason, the combination of clinical information with biomedical knowledge bases

- including relationships between genes and diseases (McKusick, 2007; Weinreich et al., 2008), sequences (Pruitt et al., 2007), interactomes (Joshi-Tope et al., 2004; Kerrien et al., 2012; Szklarczyk et al., 2011; Turner et al., 2010), genetic variations (Landrum et al., 2014), patients (Firth et al., 2009), and so on - allows researchers to study the interactions that take place inside and outside the cells and explore disease-causing genes and/or the involved genetic regions.

These databases provide the accessibility to biomedical data, making it available to the scientific community. Nevertheless, further efforts are still needed in the development of new tools to analyze this information and extract new knowledge (Medina, 2013). Therefore, contributions of computer scientists, mathematicians and statisticians are crucial to speed up integrative strategies, so helping out analyses and understanding of the development of pathological processes (Barabási et al., 2011; Joyce and Palsson, 2006; Medina, 2013; Vidal et al., 2011).

In this respect, graph/network theory has been extensively used for studying complex properties of biological systems. For instance, the interactions among the elements involved in biological processes can be represented as networks, where the nodes could be proteins, genes or metabolites. These networks represent interactomes, that is, the whole set of interactions between macromolecules present in the same system (for instance, a cell). Interactomes provide a simplified perspective of the cellular behaviour, however they allow to unveil knowledge not observable *a priori*(Vidal et al., 2011)).

Network analysis has also been used to explore the relationships among diseases. Two outstanding works are "The human disease network"(Goh et al., 2007) and "The orphan disease network"(Zhang et al., 2011), which study networks built from gene-disease relationships provided by OMIM and Orphanet, respectively. These approaches (or *diseasomes*) were the foundations of the recently field called “Network medicine”, which is based on the integration of biomedical information: relationships among diseases, interactions among molecules inside the cell and social networks (Barabási, 2007).

Diseasomes are useful for looking into the relationships among diseases, where diseases are treated as indivisible entities. However, if we consider the phenotypic diversity of human diseases, they can be broken down into multiple different pathological phenotypes (patho-phenotypes). This makes it possible to distinguish with a higher precision associations between biological processes and disease-specific phenotypes (Barabási et al., 2011). In particular, phenotypic data provides two new levels of information that are omitted in disease networks studies, the phenotype specificity and the phenotypic relationships among phenotypically similar diseases (Hidalgo et al., 2009; Lussier and Liu, 2007; Nakazato et al., 2009). For this reason, the symptomatology of genetic diseases cannot be

omitted from clinical analyses, as it is required to study the problem from systemic perspectives. However, the lack of a standard vocabulary to represent the phenotype information has been an important handicap to progress in this area.

In biomedicine and biology, biomedical ontologies (OBO) are tools extensively used. These resources consist in a standardized vocabulary organized hierarchically. The most consolidated example of biomedical ontology is Gene Ontology (GO), which contains concepts referred to biological functions (GO terms). GO organizes these terms in three domains: biological processes, molecular functions and cellular components. Annotations are genes-GO terms relationships and provide the description of the biological function of these genes. In turn, Human Phenotype Ontology (HPO) hierarchically organizes a vocabulary of terms corresponding to pathological phenotypes associated with genetic diseases (Robinson and Mundlos, 2010). The use of biomedical ontologies (HPO and GO) makes it possible to compare phenotypic and functional profiles, respectively, and to estimate similarity between genes, diseases or any object annotated in the ontology.

Currently, there are numerous tools that use GO and are available in Bioconductor (`https://www.bioconductor.org/`): GOSemSim (Yu et al., 2010), PCAN (Godard and Page, 2016) or topGO (`http://bioconductor.org/packages/release/bioc/html/topGO.html`). However, the number of tools using HPO is still very small. An example is Phenomizer (Köhler et al., 2009), a tool for obtaining diseases or genes phenotypically similar to a set of input phenotypes. Other tools, as PhenomeNET (Hoehndorf et al., 2011) combines HPO with phenotypic information of other species and also makes it possible to explore biomedical relationships among genes or diseases.

Phenotypic information is a key piece for a better understanding of molecular mechanisms that lie behind pathologies. For this reason, the study of genetic diseases requires deep phenotyping and genotyping at the individual level to build a precise map of the relationships among the phenotypic manifestations and the genetic variations along the genome. Several initiatives are focused in establishing relationships between phenotypes and genetic variations. For instance, databases such as ClinVar (Landrum et al., 2014) or DECIPHER (Firth et al., 2009) have datasets of genetic variations (both SNP and structural) and phenotypic information (HPO terms) at the sample level.

In summary, the use of the phenotypic profiles at disease, gene or individual level could be very useful to improve our understanding of the etiology of human diseases. The study of phenotypic relationships between genes within their molecular context will also help to point out to the putative altered biological process(es) in a given pathological condition. Some current initiatives, aimed to develop platforms for sharing molecular, functional and/or phenotypic data, are indeed enhancing the progress of research in genetic

and molecular bases of diseases. Thus, this kind of resources should be especially useful to advance towards a better understanding of many rare diseases.

Hypothesis and Objectives

Taking into account the reasons mentioned above, and with the major aim to help the advance of human disease characterization (especially rare diseases), the present PhD project was planned to contrast the following hypothesis:

The study of phenotypic information on complex diseases and the integration of this information with functional information could help to better understand the molecular mechanisms involved in many pathological processes.

To do that, we established the following objectives:

- **Objetive 1:** Analysis of the phenotypic relationships among genes.
- **Objetive 2:** Development of a tool (PhenUMA) to query and explore phenotypic and functional relationships.
- **Objetive 3:** Extraction of gene-disease relationships not included in public databases by using text mining and ontologies.
- **Objetive 4:** Analysis of genotype-phenotype relationships in a heterogeneous population of patients of genomic syndromes.

Methods

Databases and repositories

OMIM (Online Mendelian Inheritance in Man): OMIM (`www.omim.org`) is a repository of genetic diseases and disease-causing genes. Morbidmap.txt file - downloaded in October 2012 - was used to obtain 4,261 relationships between 3,794 genes and 3,486 diseases.

Orphanet: Orphanet (`www.orpha.net`) is a catalogue of rare diseases. It includes information about the molecular bases of these pathologies, known-related genes and detailed descriptions in several languages. In this study, the file *Diseases with their associated genes* has been used to obtain gene-disease relationships (downloaded in October 2012). This file contains 4,472 relationships between 2,614 genes and 2,555 diseases.

DECIPHER: It is a database that allows physicians around the world to report mutations and clinical features of patients affected by rare genomic disorders. DECIPHER – available from `https://decipher.sanger.ac.uk/` - is mainly focus on structural variations such as Copy Number Variations (CNV). The information used in this study has been downloaded in May 2014. This file includes the associations between the CNVs detected and the symptoms observed in the patients. The resulting data consist of 6,564 patients with 9,186 mutations associated with 1,860 phenotypes (Firth et al., 2009).

DGV (Database of Genomic Variants): It is a collection of structural variations from the control population built from results of multiple genetic association studies. This database has information about mutated chromosomal regions, types of mutations, bibliographic references, and the platform used for the analyses of healthy individual CNVs. The file GRCh37_hg19_variants_2014-10-16.txt was downloaded from the link `http://dgv.tcag.ca/dgv/app/home`

ClinVar: is a database storing relationships among variations and phenotypes (Landrum et al., 2014). The information provided by ClinVar comes from a multitude of groups and laboratories that submit the pathological interpretation of a genetic variant. Currently, ClinVar contains around 158,000 records involving more than 125,000 variants (Landrum et al., 2014).

Gene-disease relationships (*diseasomes*) and its classification

The analysed associations between genes and diseases derived from information about mutations proposed as the cause of a respective disease. These associations were used to build the bipartite networks HDN (Human Disease Network) and ODN (Orphan Disease Network), from OMIM and Orphanet respectively, these types of networks are often called *diseasomes* (Goh et al., 2007; Zhang et al., 2011). To compare both *diseasomes*, it was initially proposed a classification of the gene-disease relationships:

- MD-MG (Monogenic Disease-Monotropic Gene)
- MD-PG (Monogenic Disease-Pleiotropic Gene)
- PD-MG (Polygenic Disease-Monotropic Gene)
- PD-PG (Polygenic Disease-Pleiotropic Gene)

A disease is monogenic when is related with only one gene and polygenic when is related with more than one gene. On the other side, we called monotropic genes to those

related to only one disease and finally, a pleiotropic gene is a gene related to more than one disease.

Gene-gene networks (Unipartite Projections)

HDN and ODN were used for building HDGN (Human Disease-Gene Network) and ODGN (Orphan Disease-Gene Network). These networks were the unipartite projections of HDN and ODN and contain relationships among genes related to the same disease or diseases.

Phenotypic Similarity Gene Network (PSGN)

Phenotypic profiles were defined using Human Phenotype Ontology (HPO), which provides a standardized vocabulary of pathological phenotypes. The phenotypic profiles for each OMIM disease were obtained from *phenotypic_annotation.omim* downloaded from HPO website. The phenotypic profile of each gene was defined from the phenotypic profile/s of those diseases related to the gene in genetic association studies.

Text-mining for integration of biomedical relationships

The semantic similarity method used in this work was based on the Resnik's measure (Resnik, 1995). This approach uses the Information Concept (IC) associated with each term of the ontology. Each of these terms (HPO terms) corresponds to a pathological phenotype. The Resnik's similarity measure is based on the Most Informative Common Ancestor (MICA) between two terms. If two HPO terms are close in the ontology, the MICA value will be greater than the farthest terms. Then, the phenotypic similarity between two genes or diseases is related to the similarity (or closeness) of the phenotypes (HPO terms) taking part of their phenotypic profiles.

A gene phenotypic similarity network was built comparing the phenotypic profiles of each pair of genes and using the measure of semantic similarity proposed by Köhler et al. 2009. This procedure was also used to build phenotype similarity networks included in PhenUMA (Objective 2).

Systematic method to identify phenotypically enriched *loci* for genomic disorders

We developed a network-based approach to study genotype-phenotype relationships in a set of patients from DECIPHER database (`https://decipher.sanger.ac.uk/`). For

this purpose, a patient network was constructed from patients with similar structural variants (CNVs), where the nodes represent patients and edges (interactions) represent an overlap within the same genomic *loci* of patient CNVs.

Python NetworkX package (`https://networkx.github.io/`) is used to obtain the cliques of the network. A clique is a complete subgraph, i.e. all its nodes are connected. The number of the nodes of the cliques goes from k = 3, without limit in the maximum size. For each clique, an enrichment analysis was carried out by applying a hypergeometric test. The result was several clique-phenotype relationships that were filtered according to the following criteria:

- Adjusted p-value <0.05.
- At least three patients are related to the same phenotype.
- At least the 50 % of the patients are related to the phenotype.

Finally, this integrative network approach was used to identify Phenotypically Enriched Loci (PEL) using a genome-wide analysis. We consider as a PEL those genomic regions where the frequency of mutated individuals was significantly higher in patients (DECIPHER) than in control population (DGV).

Results

Analysis of the phenotypic relationships among genes

OMIM and Orphanet provide gene-diseases relationships used for building the *diseasomes*: Human Disease Network (HDN) (Goh et al., 2007) and Orphan Disease Network (ODN) (Zhang et al., 2011)). In this work, the updated versions of the *diseasomes* were explored. HDN contains 2,525 genes related to 3,132 diseases and ODN contains 2,331 genes related to 2,125 diseases.

The gene-disease relationships in HDN and ODN were classified by the number of genes involved in each disease (monogenic diseases or polygenic diseases), as well as the number of diseases related with each gene (monotropic genes or pleiotropic genes), as it was described in methods. In both networks, monotropic genes represent a high percentage of the total set of disease-causing genes: 72 % (1,810 genes) of the OMIM genes and 69 % (1,625 genes) of the Orphanet genes.

Although the gene-disease relationships included in OMIM and Orphanet are similar, the resulting unipartite networks show great differences. Firstly, there is a contrast in

the number of nodes and edges. ODGN (1,492 nodes and 6,380 edges) is quite larger than HDGN (749 nodes and 2,654 edges). Besides, there are some differences in the number of unconnected nodes: 1,776 in HDGN (Human Disease-Gene Network), and 839 in ODGN (Orphan Disease-Gene Network). The number of isolated genes is directly related to the amount of information in the network. The higher number of unconnected nodes the lower the number of gene-gene relationships. Therefore, these results could point out to a lack of information (relationships) for genes included in HDGN.

Taking into account that the information contained in both databases is similar, the differences between the networks could be due to the way in which the information is organized. For example, frequently Orphanum identifiers (used by Orphanet to identify diseases) are general pathologies (e.g. Diabetes) that correspond to several types of diseases in OMIM (e.g. Diabetes type I. Diabetes type II, and others). On the other side, Orphanet diseases are usually related with a greater number of genes than OMIM diseases. These differences reinforce the need for consensus to determine gene-disease relationships, the definition of disease and its classification.

All diseases grouped into a single general pathology have something in common: symptoms. In this thesis, it is considered that the analysis of the symptomatology of the diseases could provide more precise gene-disease relationships. Besides, the previous analysis reinforces the hypothesis about the need for a systematic procedure of phenotypic characterization of genetic diseases.

Novel pathological relationships among disease-causing genes emerge from the comparison of phenotype profiles

Phenotypic variability and complexity are properties frequently observed in genetic diseases (Loscalzo and Barabasi, 2011). Therefore, the use of the phenotypic profiles (defined as the set of symptoms of a disease) could be useful to study phenotypic diversity in order to establish new disease-causing gene relationships.

Phenotypic profiles were used to calculate similarity between each pair of genes using semantic similarity. We obtained more than 3,000,000 pairs of genes with a value of semantic similarity greater than 0. However, we only consider as significant those located at the 98th percentile (2 % of the comparisons with the highest phenotypic similarity value). These were included in the phenotypic similarity gene network (PSGN), which contains 26,197 relationships among 1,705 genes.

The use of hierarchical structures, as HPO, allows us to consider the specificity of the phenotypes to establish relationships among genes (or diseases). In this work, specificity

is defined according to the frequency of each phenotype in the global set of genes. This approach allows us to prioritize phenotypically similar genes related to the most specific phenotypes. That is, a high value of phenotypic similarity between two genes indicates that both genes share similar phenotypes and these ones are very specific. In fact, the relationship set included in PSGN (2,6197 in total) correspond to the gene pairs associated sharing very specific phenotypes, as their similarity values represent the top 2% of the highest phenotypic similarity values. From this approach, the resulting phenotypic relationships are useful to study the underlying biomolecular processes involved in the different pathological conditions.

Integration of phenotypic and functional information

Some interactomes were used to study the underlying functional information of the phenotypic relationships:

- Protein-protein interaction network (PIN), with 74,657 interactions among 9,580 genes.
- Functional similarity relationships (GO) with 496,973 relationships and 9,157 genes.
- Metabolic relationships involving 9,812 relationships among 535 enzymes.

The percentage of phenotypic relationships included in PSGN that were supported by a known functional interaction was calculated for each type of interactome. As a result, 23.7% of the physical interactions, 11.8% of the metabolically associated genes, and 8.1% of the functional relationships obtained from GO (Biological Process) were included in PSGN. Despite the set of functionally and phenotypically related genes is small, they provide a starting point for the study of molecular mechanisms related to genetic diseases.

PhenUMA

The results shown in the previous points highlight the advantages of using phenotypic relationships and their integration with functional information. The access of the scientific community and physicians to this information could help in the formulation of new hypotheses. PhenUMA was developed to provide an easy access to this information. PhenUMA is a platform for the analysis of phenotypic and functional relationships among genes and diseases. It is publicly accessible from the url `www.phenuma.uma.es`.

PhenUMA database integrates several types of relationships that help to discover relationships between genes or diseases that cannot be deduced from the source databases. The relationships are classified into three groups: functional relationships, inferred relationships (unipartite projections of the *disasomes*) and phenotypic relationships. The functional interactions are protein-protein interactions from the STRING database (Szklarczyk et al., 2011), metabolic relationships between genes based on flux-correlations (Veeramani and Bader, 2009) and functional similarity relationships among genes were calculated using GO. Inferred relationships were obtained by using different graph projections of gene-diseases networks, from OMIM and Orphanet. We distinguished two types of inferred relationships: gene-gene relationships, where both genes are related with the same disease (or diseases) and disease-disease relationships, where both diseases are related with the same gene (or genes). Finally, phenotypic similarity relationships among genes or diseases were estimated using HPO and the semantic similarity measure used by Köhler et al. (2009) were incorporated into the database.

PhenUMA has been designed to provide biomolecular and clinical information associated with a set of genes, diseases or phenotypes of interest. To perform a query, it is necessary to provide a set of genes, diseases or phenotypes and a type of relationship (e.g. phenotypic, functional, metabolic, etc.). Using this input data, PhenUMA builds a network that integrates all the relationships included in the databases among the queried genes (diseases or phenotypes). Figure 1 shows how networks are built from each type of input. First, input query is used to obtain a seed network, containing the relationships between the input data and other genes or diseases included in the database. Then, this network is enriched with other types of interactions: protein-protein interactions, metabolic interactions, etc.

When PhenUMA was released, there were some other platforms to explore biological and biomedical relationships between genes. For instance, GeneMANIA integrates physical, metabolic and co-expression relationships (Warde-Farley et al., 2010). However, GeneMANIA does not include pathological relationships. Other tools such as MalaCards (Rappaport et al., 2013), Phenomizer (Köhler et al., 2009) and PhenomeNET (Hoehndorf et al., 2011) use phenotypes associated with genes and diseases, but these tools do not integrate phenotypical and functional information.

Currently, the analysis of phenotypic information has been integrated in several databases such as DECIPHER, OMIM, ClinVar, MalaCards and Orphanet. Some other recent resources such The Monarch Initiative platform (Mungall et al., 2017) integrates several tools to directly analyse phenotype and genetic data from patients. This platform includes interesting tools such as PhenIX (Smedley et al., 2015) for the prediction of genes from exomes

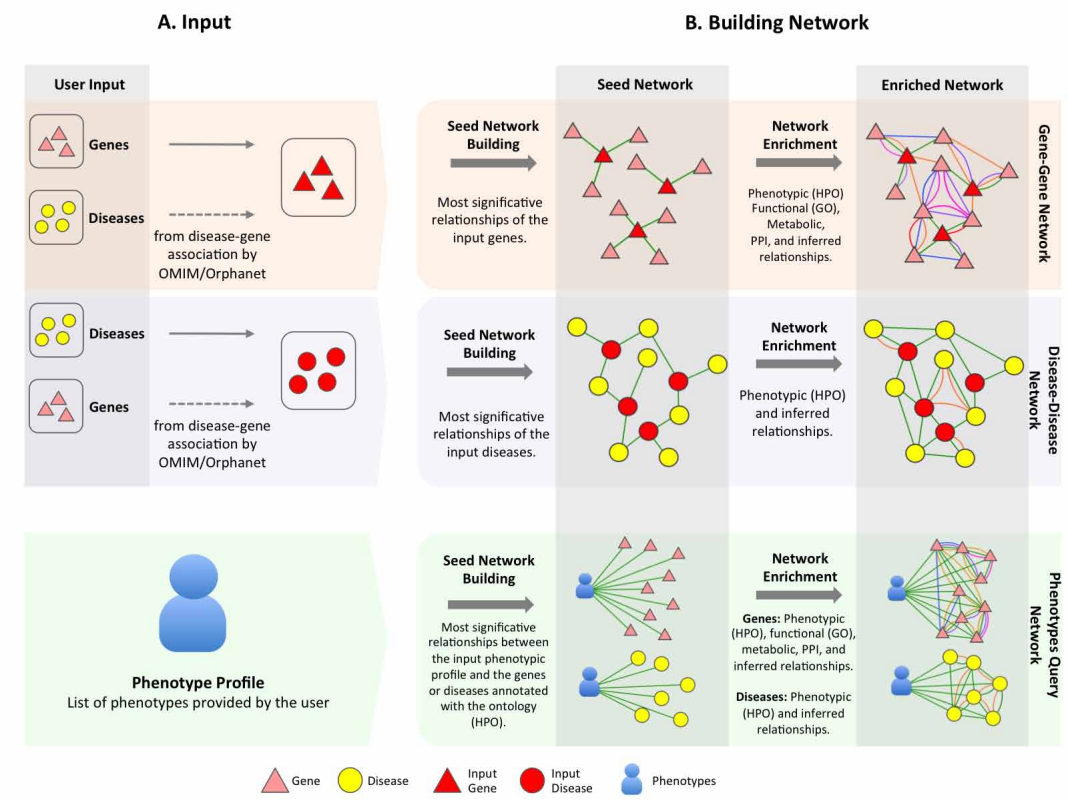


Figura 1: PhenUMA workflow. This figure shows the steps followed by PhenUMA when a query is executed. **A.** Various types of data can be queried: genes, diseases or phenotypes. **B.** The network creation process is slightly different for each type of query. First, a "seed network" is created that includes the relationships between the input data (for example, genes) and the rest of the information included in the database. Subsequently, this network is enriched with the rest of the relationships included in the database. As shown in the figure, the type of relationships included in this process depends on the type of nodes in the network (genes or diseases).

and phenotypic profiles or the Patient Archive (PA, `http://patientarchive.org`) that is a repository oriented to clinicians and researchers with the aim of sharing patient information similar to DECIPHER (Firth et al., 2009), PhenoTips (Girdea et al., 2013) or Gene2MP (Chong et al., 2015).

Integration of phenotypic profiles of biogenic amines-related genes from literature using text-mining tools

The relationships included in PhenUMA are based on the information included in public databases: OMIM, Orphanet and HPO. PhenUMA stores relationships among genes and their available phenotypic information. Nevertheless, there are not phenotypic profiles of genes without any pathogenic variant located so far, a circumstance that can induce a gap of information in PhenUMA. Eventually, the lack of information in the databases contrasts with the pathological information present in literature. This can lead to the fact that PhenUMA can return an empty output for some genes, in spite of their well-documented pathological implications. This issue is observed in genes involved in the metabolism of biogenic amines. This group of genes were taken as a case of use for integrative analyses starting from information retrieved from literature by using text mining techniques. Specifically, we explore the phenotypic information concerning genes related to the metabolism of polyamines (putrescine, spermidine and spermine), histamine, dopamine y serotonin.

The total number of references retrieved from PubMed associated with ”human polyamine” or ”human histamine” is more than 40,000. With respect to dopamine and serotonin the results include around 60,000 references. If we add in the search the word “disease”, the number of references ranges from 6,000 to 20,000, respectively. This supports the existence of studies about the associations between medical conditions and biogenic amines in the biomedical literature. However, this information is not reflected in the repositories of human diseases like OMIM, Orphanet or DECIPHER.

Using text-mining tools, 429 relationships between 53 genes (out of 60 included in the analysis) and 129 diseases were obtained. The 96 % of these relationships comes only from the literature, and 36 relationships (4 %) were also described in OMIM. The low number of diseases included in databases is because these databases catalogue gene-disease relationships that are supported by disease-causing variants. However, only a few mutations or variants of amine-related genes related to the manifestation of some pathology. PhenUMA only provides an output network for some genes involved in dopamine metabolism (DRD2, TH and DDC). In respect to polyamines, there is information only for SAT1,

SMS and ARG1. No result was obtained for genes related with histamine. However, after a deep analysis of the literature is possible to identify biological processes, involving amine-related genes, that are affected and responsible for many conditions.

In order to study the phenotypic information of amine-related genes, the gene-disease associations were used to assign a phenotypic profile to each gene. Fifty-three (out of the initial input of 60 genes) were related with diseases by literature and annotated in the ontology. In total, the phenotypic information associated to these genes is composed by 2,975 annotations between the 53 genes and 928 phenotypes (HPO terms) distributed by several branches of the ontology.

Phenotypic profiles allowed us to integrate these genes into the whole gene-phenotype network and into PhenUMA. Considering the phenotypic profiles from the literature, 7,754 phenotypic similarity relationships (above the 98th percentile) were obtained among 1,149 genes and the 53 amine-related genes.

The phenotypic relationships were integrated with information from STRING and GO. The combination of text-mining tools and the integration of functional and phenotypic information allowed us to establish relationships among amine-related genes and other genes: such as SINCA, TFN or EDNRA. Results indicate that is possible to obtain a list of candidate genes, which could be related to the pathophysiology of amine-related genes.The fact that the strategy allows us to locate well-known information can be considered as a validation proof.

There are multiple non-genetic causes that influence proteins expression or activity. Thus, this strategy that combines text-mining and network analysis techniques can provide a useful strategy to reach a better understanding of the complex biological problems involved in different pathological processes.

Study of genotype-phenotype relationships in DECIPHER patients

In the previous sections, the biomedical information related to genes and diseases were analysed in order to reach a better understanding of the molecular mechanism underlying medical conditions. The incorporation of the phenotypic profiles can be also used to study the consequences of genetic variations. For this reason, the DECIPHER database (Firth et al., 2009) was used to study the relationships among structural variations and pathological processes. DECIPHER is a repository of patients, which provides the copy number variations or CNVs (including deletions and duplications) and a set of symptoms (described as HPO terms) related to each patient. In total, this database includes more

than 45,000 patients (March 2015), of which over 10,000 gave their consent to share their medical data (Firth et al., 2009).

Here we analyse 9,186 CNVs genotyped in 6,564 patients. This dataset is coming from patients suffering of heterogeneous genomic disorders characterised of very diverse clinical features such as developmental delay, intellectual disability, and congenital malformations. Our exploration involved patient genotypes and phenotypes to perform a pipeline that included network analysis, phenotypic enrichment, and genetic association studies. In this way, it was possible to investigate similarities between genetic micro-variations (CNVs) and pathological phenotypes. The combined use of these methods allowed us to point out putative novel syndromes so far uncatalogued in databases as ClinVar or DECIPHER.

First, both the CNVs and phenotypic profiles included in the analysis were explored. With respect to phenotypic profiles, it was observed that patient profiles with a low number of phenotypes (1-5 phenotypes) were the most frequent ones. On the contrary, profiles with a greater number of phenotypes were less frequent. Additionally, the complexity of these profiles is higher for patients with "de novo"variations, compared to those inherited from their progenitors. This effect is similar when CNV lengths were compared between patients and controls. Patients CNVs (DECIPHER) are frequently larger than CNVs observed in healthy individuals.

The objective of this work is to study the CNVs included in DECIPHER and identify regions related to some phenotypic manifestations. For this purpose, relationships were established among patients whose CNVs overlap at least one base pair. The result was a network, where nodes correspond to patients and the relationships indicate an overlap between the mutated regions of two individuals. In total, the network consisted of 6,342 nodes and 89,526 relations.

The patient network was explored in order to identify phenotypically enriched *loci*, also called PELs (Phenotypically Enriched Loci). The objective was to identify patient groups having overlapped mutated regions (at least one base pair - dsDNA sequence - overlapped among them). So, firstly, we identified cliques, grouping at least 3 patients. In graph theory, a clique is a set of nodes forming a complete graph (i.e., all the nodes of the graph are related among them). Then, in this network, cliques are groups of patients, whose CNVs overlap at the same region. The second phase was the phenotypic enrichment of each clique using the phenotypes (HPO terms) associated with each patient. This process allows us to select cliques, which patients share similar phenotypes, and identify the most representative phenotypes of each group.

The results contain 1,042 *locus*-phenotype relationships involving 487 PELs and 195 enriched HPO terms. Control cases were used to identify significant genotype-phenotype relationships, using a similar procedure as Cooper et al. (2011). After this filter, the results contained 387 *locus*-specific phenotype relationships among 336 PELs and 115 different phenotypes.

The next analysis was focused on checking whether the method was able to identify as PEL those CNVs that were previously reported as pathogenic by other genetic studies. For this purpose, we used ClinVar, a public database that includes relationships between variations and genetic diseases. ClinVar incorporates details about their clinical significance so the variation-phenotype can be labelled as pathogenic, likely pathogenic, benign, and likely benign. To check how many PELs are present in ClinVar we selected 2,243 pathologic or likely pathogenic CNVs-OMIM associations and another 75 genomic regions associated with DECIPHER syndromes. Next, we studied the overlapping of these regions with the identified PELs identified by our approach. We compared the regions affected by the same type of mutation (i.e. deletions or duplications). As a result, there were obtained a total of 93 and 15 genomic regions associated with genomic syndromes including ClinVar and DECIPHER, respectively.

ClinVar regions were used for visualization of the patient network, that represents a map of the phenotype associations identified from DECIPHER (Figure 2). In this network, we observed that, in some cliques, not all the patients are affected by a mutation identified as pathogenic in ClinVar. Specifically, 164 (50 %) of the CNVs of the PELs overlap with a region in ClinVar, indicating that PELs are potentially related to known genomic disorders.

Conclusions

The next points summarize the conclusions of this thesis:

- Phenotype information of genetic diseases can be used to build phenotypic similarity gene networks (PSGN). The resulting network allows us the identification of novel relationships among disease-causing genes, which are useful to explore the molecular bases of pathological processes.

- A tool for the analysis of functional and phenotypic information was developed. PhenUMA has demonstrated to be a useful resource to study biomedical relationships among disease-causing genes, diseases and phenotypes. PhenUMA provides novel phenotypic and functional relationships among disease and genes that are not directly accessible in disease databases.

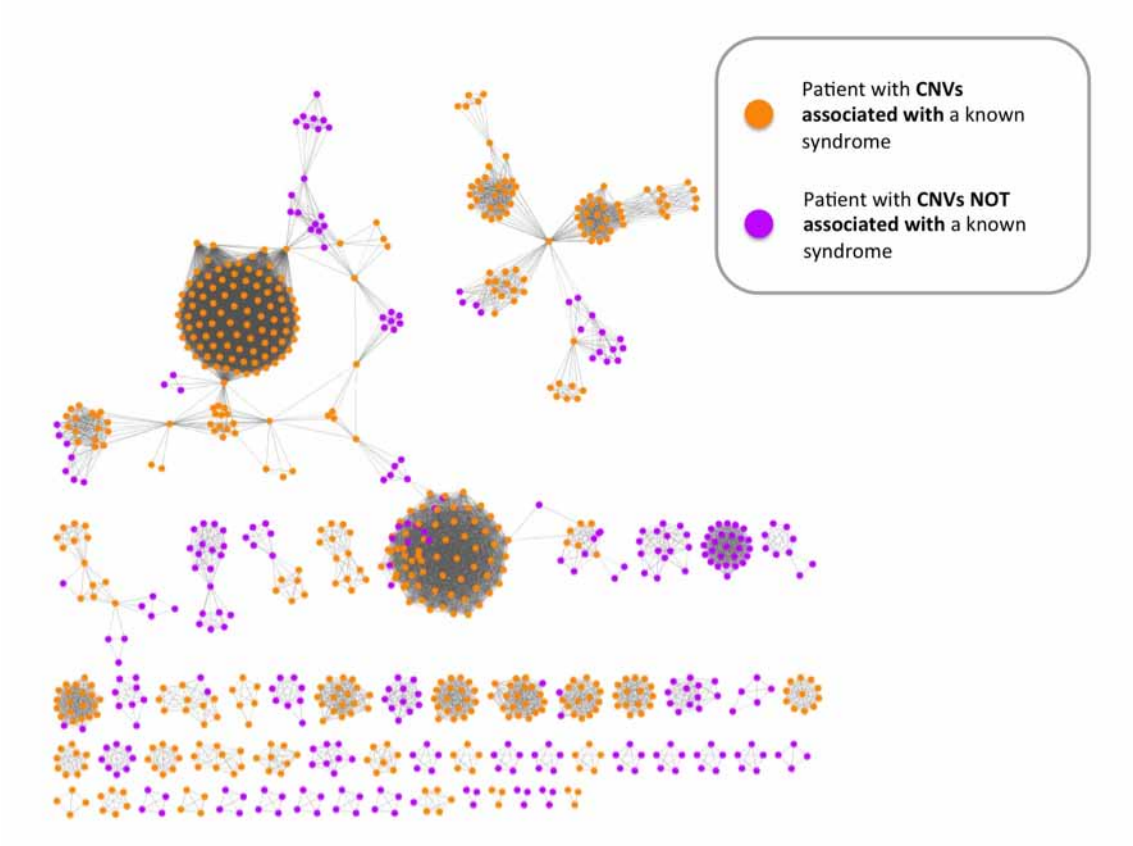


Figura 2: Patient network. This network includes relationships among patients whose CNVs overlap. The minimum overlap is 1bp. ClinVar data were used to highlight the patients with regions overlapping with any known disease-causing variation. Orange nodes indicate patients having any CNV that overlaps with a CNV related to a known syndrome. In purple, nodes corresponding to patients with CNV non-related to syndromes catalogued in disease databases.

- Text-mining techniques were used to define phenotypic profiles of amine-related genes that do not have genetic associations with disease in databases. These techniques could be a complement for PhenUMA to detect literature-based relationships between gene and diseases that are not present in disease databases.
- Individual phenotypic and genotype information of a heterogeneous population of patients from DECIPHER database has been used to build a patient network. Further analyses on the resulting patient network has revealed significant relationships among symptoms and structural variations. In addition, known genomic syndromes can be suspected for some undiagnosed patients and putative novel syndromes could be suggested.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En las últimas décadas, los avances en técnicas experimentales de alto rendimiento, como la genómica, la transcriptómica o la proteómica, han generado una enorme cantidad de datos. Esto ha impulsado el estudio de las bases genéticas y moleculares de un gran número de enfermedades humanas. Prueba de ello son los estudios de asociación del genoma completo o GWAS (*Genome Wide Association Studies*), que han permitido la identificación de miles de *loci* asociados a enfermedades comunes. Sin embargo, conocer las bases moleculares de muchas enfermedades supone aún un reto de largo alcance (Stessman et al., 2014).

Parte de la dificultad radica en la complejidad de algunas enfermedades, que pueden ser consecuencia de alteraciones en varias regiones del genoma (Botstein and Risch, 2003; Vidal et al., 2011), y no de una sola variación. Además, una gran parte de las variaciones genéticas asociadas a enfermedades están localizadas en regiones no codificantes, afectando en mayor medida a regiones reguladoras del genoma (Grubert et al., 2015; Maurano et al., 2012). Por otro lado, la información genética no siempre es suficiente para explicar los mecanismos y procesos implicados en una enfermedad. Al componente genético hay que sumarle la influencia de otros factores, tales como el estilo de vida, el entorno y procesos estocásticos que influyen en la expresión génica, y por tanto, en la manifestación de determinados fenotipos (Hu et al., 2016; Lehner, 2013). A pesar de que los GWAS han sido útiles para identificar cientos de miles de relaciones entre variaciones genéticas y enfermedades, estos esfuerzos son aún insuficientes para entender la etiología y las bases moleculares de las enfermedades más complejas (Visscher et al., 2012).

1.1. Aproximaciones biocomputacionales para el estudio de enfermedades raras

El incremento en la eficiencia de las técnicas de alto rendimiento, sumado al desarrollo en paralelo de métodos de integración y análisis de información, han ayudado a minimizar estas limitaciones. Esto permite estudiar las enfermedades en varios niveles de complejidad (Auffray et al., 2009): interacciones genéticas o reguladoras, relaciones metabólicas, interacciones físicas, etc. El estudio de toda esta información en conjunto ha proporcionado otra visión de enfermedades complejas y multifactoriales como el cáncer, la obesidad o la diabetes (Auffray et al., 2009). Con respecto a las enfermedades raras, el problema es especialmente complejo debido a la heterogeneidad de las etiologías y la baja disponibilidad de muestras de pacientes y familias para el análisis molecular y fenotípico de estas patologías (Boycott et al., 2013).

En Europa, una enfermedad se considera rara si afecta a menos de 1 individuo de cada 2.000, en cambio, en Estados Unidos el criterio es que afecte a menos de 200.000 individuos. Aunque es difícil determinar el número exacto de enfermedades raras, existen aproximadamente entre 6.000 y 8.000 enfermedades catalogadas como raras y se conoce que aproximadamente un 80 % son de origen genético. Con respecto a los genes asociados con enfermedades raras, el número oscila entre 7.000 y 15.000 genes (Boycott et al., 2013).

Se estima que entre un 6 % y un 8 % de la población está afectada por alguna enfermedad rara. A pesar de que estos porcentajes representan una porción considerable de la población, la cantidad de individuos afectados por una enfermedad concreta es en ocasiones muy bajo. Este hecho implica un número de muestras de pacientes insuficiente para estudiar las patologías de menor prevalencia. Por ello, determinar las bases genéticas y moleculares de estas enfermedades supone una gran dificultad. La secuenciación completa de genomas (WGS) y exomas (WES) ha contribuido enormemente a la identificación de las causas genéticas de numerosas enfermedades de baja prevalencia (Boycott et al., 2013; Buske et al., 2015; Smedley et al., 2014; Smedley and Robinson, 2015). Las estrategias para identificar genes o variaciones candidatas usando WGS y WES se basan en identificar las variaciones potencialmente patogénicas y localizar y anotar su posición en el genoma (Boycott et al., 2013). Sin embargo, algunas variaciones son extremadamente raras, que junto con el número reducido de pacientes dificulta aún más la tarea (Firth et al., 2009). A esto hay que sumarle las posibles imprecisiones de las relaciones entre variaciones genéticas y enfermedades raras. Estudios recientes, como el realizado por *Exome Aggregation Consortium* (Lek et al., 2016), han destacado la frecuencia en población sana de variacio-

nes presentes en bases de datos especializadas (como OMIM o ClinVar) y relacionadas con enfermedades raras.

Para afrontar estas dificultades se han puesto en marcha plataformas cuyo objetivo es proporcionar un repositorio común de datos genómicos y fenotípicos de pacientes de enfermedades raras. Por ejemplo, la base de datos DECIPHER (Firth et al., 2009) (*Database of Chromosomal Imbalance and Phenotype in Humans using Esembl Resources*), que se inició en 2004 en Reino Unido, proporciona una interfaz en la que los grupos clínicos pueden subir la información de sus pacientes. Estos datos pueden ser utilizados por la comunidad científica bajo demanda y tras la firma de un acuerdo con DECIPHER. Otra plataforma es RD-Connect (Thompson et al., 2014), fundada por la Unión Europea y por IRDiRC (*International Rare Diseases Research Consortium*) (http://www.irdirc.org/). RD-Connect es una plataforma que permite almacenar datos genómicos y registros de pacientes, y pone a disposición de la comunidad científica herramientas útiles para análisis de datos relacionados con enfermedades raras (Thompson et al., 2014). Otras iniciativas como *Undiagnosed Diseases Program* (Gahl et al., 2012) del *National Institutes for Health* (NIH) o las canadienses Care4Rare (http://care4rare.ca/) y FORGE (*Finding of Rare Disease Genes*) (Beaulieu et al., 2014) trabajan para la identificación de genes candidatos relacionados con enfermedades raras.

La accesibilidad a datos de pacientes diagnosticados de enfermedades raras supone un recurso de gran utilidad para el estudio de las relaciones entre los genes afectados y los fenotipos observados en dichos pacientes. Dada la complejidad fenotípica de estas patologías, un análisis sistémico de sus sintomatologías podría potenciar la identificación de genes candidatos, así como arrojar luz sobre las bases moleculares implicadas en el desarrollo de las mismas. Por otro lado, el análisis de las regiones afectadas, así como su relación con fenotipos patológicos, podrían permitir priorizar aquellas regiones del genoma asociadas con enfermedades genéticas y contribuir a la identificación de nuevos síndromes.

Además de la información contenida en las bases de datos de pacientes, en la actualidad existen un gran número de repositorios consolidados que contienen una extensa variedad de información biomédica: relaciones entre genes y enfermedades (Weinreich et al., 2008; McKusick, 2007), secuencias (Pruitt et al., 2007), interactomas (Joshi-Tope et al., 2004; Kanehisa and Goto, 2000; Kerrien et al., 2012; Szklarczyk et al., 2011), datos de variaciones genéticas (Landrum et al., 2014), entre otros. Estas bases de datos ponen esta información a disposición de la comunidad científica, otorgándole un mayor potencial. La integración de esta información aporta una perspectiva global, más completa, del conjunto de interacciones que acontecen en el interior y exterior de las células (Aittokallio

and Schwikowski, 2006). El análisis de estas interacciones permite abordar la complejidad de los procesos biológicos implicados en enfermedades raras desde un punto de vista holístico (Medina, 2013).

Es un hecho que las ciencias de la computación, las matemáticas y la estadística son cruciales para acelerar la integración, el análisis y la comprensión de los mecanismos involucrados en los procesos biológicos y en patologías (Barabási et al., 2011; Joyce and Palsson, 2006; Medina, 2013; Vidal et al., 2011). En el área de la bioinformática se han desarrollado un gran número de algoritmos, herramientas y *pipelines* para el análisis de datos de secuenciación y el alineamiento de secuencias (McGinnis and Madden, 2004). A esto hay que sumarle la aplicación de técnicas como la minería de datos (Hall et al., 2009; Huang et al., 2008; Rebhan et al., 1998), ontologías como *Gene Ontology* (GO) (Ashburner et al., 2000) - de la que se hablará más adelante - así como, la aplicación de teoría de grafos para el análisis de redes biológicas (Goh et al., 2007; Vidal et al., 2011).

En definitiva, la bioinformática es indispensable para complementar los avances en técnicas experimentales (Kitano, 2002), ya que potencia la extracción de nuevo conocimiento, y el planteamiento de nuevas hipótesis. Esta estrategia holista ha permitido avanzar en el estudio de las causas genéticas de las enfermedades raras y del resto de patologías.

1.2. Teoría de redes en biología

Una de las aproximaciones usadas para el estudio de la complejidad de los sistemas biológicos implica la construcción y análisis de redes. Gran parte de la información obtenida a partir de técnicas experimentales puede expresarse en forma de grafo. La teoría de redes o teoría de grafos es un área de las matemáticas con aplicación en distintas disciplinas (física, sociología, economía, ciencias de la computación, etc.) para representar las interacciones entre los elementos de un sistema. En teoría de grafos se define una red (G) como un conjunto de nodos o vértices (V) y un conjunto de aristas (E) entre estos nodos, formalmente G={V,E}. La tabla 1.1 incluye algunos de los conceptos básicos relacionados con la teoría de redes y que serán usados en varios capítulos de esta tesis.

Las interacciones entre los elementos que participan en los procesos biológicos de las células se pueden expresar en forma de grafos, siendo los nodos proteínas, genes o metabolitos. Entre las redes más relevantes con las que se ha trabajado en las últimas décadas están las redes de interacción entre proteínas, generalmente obtenidas a partir de experimentos *yeast-two hybrid* (Y2H) (Kerrien et al., 2012; Szklarczyk et al., 2011). También se han construido redes metabólicas que representan las reacciones bioquímicas que se

Tabla 1.1: Definición de conceptos relacionados con la teoría de redes.

Concepto	**Definición**
Red bipartita	Red que en los que las aristas (E) relacionan dos tipos de nodos. Por ejemplo, una red entre genes y enfermedades.
Red unipartita	Red que solo contiene un tipo de nodo
Grado	Para un nodo de la red el grado o *degree* es el número de relaciones de ese nodo.
Red no dirigida	Red en la que las relaciones no tienen dirección. Una relación entre un nodo A y un nodo B es igual a la relación entre el nodo B y el nodo A.
Red dirigida	Red en las que las relaciones tienen orientación.
k-*clique* o *clique*	Subred o subgrafo de k nodos en la que todos ellos están conectados entre sí (subgrafo completo).

producen en las células, y redes de regulación de la expresión génica (ej.: interacciones entre elementos cis y trans del transcriptoma) (Joshi-Tope et al., 2004; Kanehisa and Goto, 2000). Estas redes representan interactomas, es decir, el conjunto de interacciones entre las macromoléculas presentes en una misma célula. Los interactomas proporcionan una perspectiva simplificada del comportamiento celular, sin embargo, permiten extraer conocimiento no observable *a priori* (Vidal et al., 2011). Los análisis topológicos de los interactomas permiten deducir el modo en el que se estructuran los elementos de la red y sus propiedades, siendo muy útiles para estudiar los principios fundamentales de organización de los sistemas biológicos.

Estos análisis topológicos han revelado algunas propiedades comunes entre distintos interactomas: i) Pueden considerarse *small-worlds* (Wagner and Fell, 2001), lo que indica que existen pocas conexiones entre cada par de nodos independientemente del tamaño de la red y el nivel de dispersión de sus nodos (Montañez et al., 2010). Esta propiedad podría ser un indicador de la robustez del sistema, ya que una perturbación afectaría a una zona de la red y no se extendería por el resto de los nodos (Hu et al., 2016; Solé and Valverde, 2008). ii) Las redes biológicas tienen una estructura *scale-free* (Montañez et al., 2010; Wagner and Fell, 2001), es decir, el grado (*degree*) de sus nodos es muy heterogéneo, hay nodos con una o dos conexiones y otros que acumulan un gran número que suelen considerarse *hubs*. iii) Las redes biológicas suelen mostrar una estructura modular. Un módulo es un conjunto de nodos con una mayor tendencia a interactuar entre ellos que con nodos del exterior. La modularidad de las redes biológicas ha contribuido al planteamiento de la hipótesis de que los módulos puede considerarse unidades funcionales (Wang and Zhang,

2007) y que es posible estudiar los mecanismos moleculares comunes entre enfermedades genéticas a partir del análisis del solapamiento de los módulos de dichas enfermedades (Barabási et al., 2011; Oti and Brunner, 2006). Estas propiedades son indicadores de la robustez de los sistemas biológicos. Sin embargo, a pesar de esta robustez, una alteración que implique el aumento de la vulnerabilidad del sistema pueden hacerlos más sensibles a perturbaciones que se asocien con la aparición de enfermedades (Kitano, 2004).

El estudio de estos módulos y de las interacciones entre ellos nos permitirá analizar las relaciones entre enfermedades y si existen funciones biológicas compartidas entre ellas (Barabási et al., 2011). Varios autores han trabajado en la hipótesis de que es posible estudiar las enfermedades como perturbaciones que afectan a elementos interconectados en el interior de la célula. En este tipo de aproximaciones, se entienden las enfermedades como consecuencia de la alteración de un conjunto de procesos que conforman grupos de elementos altamente interconectados (Vidal et al., 2011).

Una de las estrategias para relacionar enfermedades consiste en el planteamiento de que dos enfermedades estarán relacionadas si son consecuencia de una mutación en el mismo gen o conjunto de genes. De esta hipótesis surgen varias propuestas de redes de enfermedades (*diseosomes*, véase Figura 1.1), como son *The human disease network* (Goh et al., 2007) y *The orphan disease network* (Zhang et al., 2011) obtenidas a partir de las relaciones gen-enfermedad procedentes de las bases de datos de OMIM y Orphanet, respectivamente. La exploración de las conexiones entre enfermedades permite estudiar cómo diferentes síntomas, en ocasiones, surgen de la alteración de los mismos procesos biológicos (Barabási, 2007). Por esta razón, no solo se requiere el estudio de los interactomas y las relaciones entre enfermedades, sino también la integración de esta información para tener una visión más completa de los elementos implicados en ellas.

Además, en el riesgo de padecer determinadas patologías no solo contribuyen las variaciones genéticas responsables de las perturbaciones que afectan a nivel molecular, sino que también influye el entorno y las interacciones sociales. Por ejemplo, los genes HLA-DQB1 y HLA-DRB1, junto con otros 21 genes, se asocian con diabetes mellitus tipo I; no obstante, está demostrado que existen factores de riesgo de diabetes causados por el entorno (Federoff and Gostin, 2009). Otro ejemplo es la obesidad, para la que una mutación en FTO supone un factor de riesgo, sin embargo, el entorno social de los individuos puede aumentar el riesgo de padecer esta enfermedad incluso en un 40 % (Barabási, 2007). La medicina de redes (*Network Medicine*) consiste en analizar la integración de toda esta información: relaciones entre enfermedades, interacciones entre moléculas en el interior de la célula y las redes sociales (Barabási, 2007), para estudiar de forma sistémica los mecanismos asociados con enfermedades genéticas (Hood and Flores, 2012).

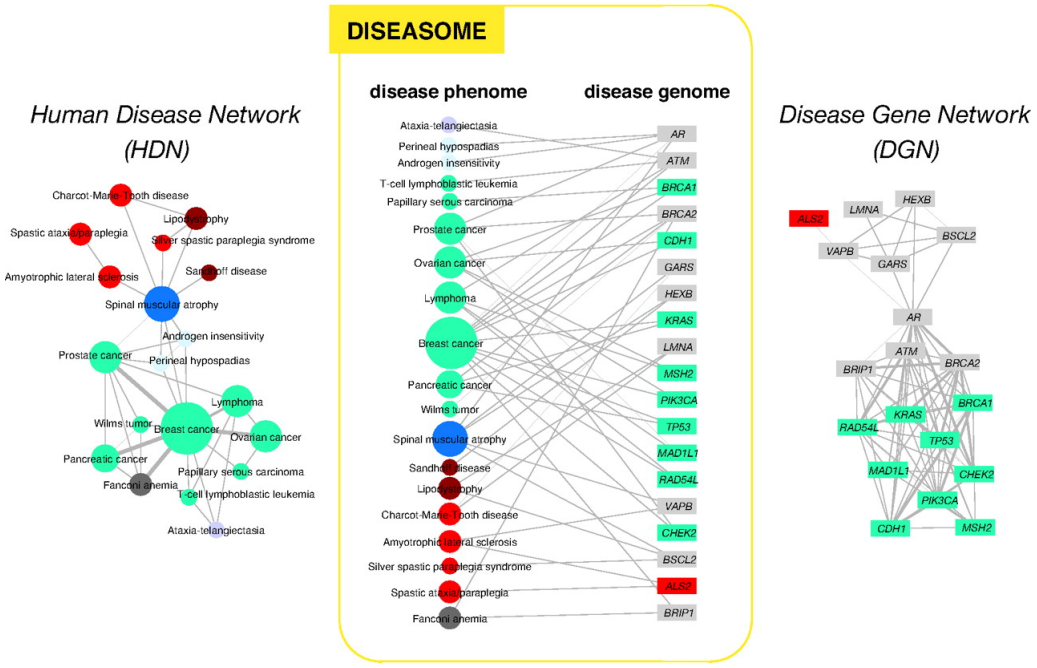


Figura 1.1: Esquema del *Diseasoma*. (Figura tomada de (Goh et al., 2007)). En el centro de la figura se incluye el *diseasoma* que está formado por las relaciones entre genes y enfermedades. A ambos lados se encuentran las proyecciones obtenidas a partir de la red bipartita (gen-enfermedad) central. A la izquierda, la *Human Disease Network* (HDN) formada por relaciones entre enfermedades asociadas con el mismo gen, y a la derecha la *Disease Gene Network* (DGN) formada por las relaciones entre genes asociados con la misma enfermedad.

1.3. Representación e integración de información fenotípica

En biología el término fenotipo tiene varias acepciones (Robinson, 2014), sin embargo, la definición más extendida hace referencia a un rasgo o característica de un individuo por influencia del genotipo y del entorno, que puede ser una manifestación a nivel fisiológico, bioquímico, molecular o de comportamiento (Pierce, 2014). No obstante, tanto en un contexto clínico como en esta tesis, esta definición se utilizará con ciertos matices. Cuando se trata de procesos patológicos, un fenotipo o síntoma representa una anomalía o desviación de la morfología o el comportamiento fisiológico canónico, como consecuencia de una enfermedad (Robinson, 2014). Por otro lado, aunque en la literatura frecuentemente se califique como fenotipo a una patología concreta (e.g. Parkinson, cáncer o diabetes) (Chong et al., 2015), en este trabajo, un fenotipo hará referencia a cada uno de los síntomas de una enfermedad. Ejemplos de fenotipos pueden ser la demencia, cambios de personalidad, depresión o alucinaciones característicos del Parkinson.

Las enfermedades raras manifiestan sintomatologías muy heterogéneas, incluso entre pacientes con la misma enfermedad, pudiendo afectar a múltiples tejidos u órganos que dan lugar a problemas morfológicos, fisiológicos o de comportamiento (EURORDIS, 2005). Los modelos de estudio detallados anteriormente (e.g. *diseasomes*) permiten estudiar nuevas relaciones entre genes y enfermedades, donde las enfermedades son tratadas como entidades indivisibles en las que se omite la complejidad asociada a una detallada descripción de sus síntomas. No obstante, las enfermedades pueden desglosarse en fenotipos y, de esta forma, considerar la diversidad fenotípica presente en enfermedades de baja prevalencia. Esto permite discriminar con mayor precisión los procesos biológicos que podrían estar implicados en algunas enfermedades o en sus fenotipos patológicos (pato-fenotipos) (Barabási et al., 2011).

Varios trabajos han destacado la relevancia de los fenotipos patológicos para establecer el grado de relación entre enfermedades (Hidalgo et al., 2009; Lussier and Liu, 2007; Nakazato et al., 2009). Es por ello que la integración de esta información es indispensable para estudiar de forma sistémica la complejidad de las enfermedades genéticas. El grado de precisión de los datos que se utilicen en dicho proceso de integración va a repercutir directamente en el potencial para generar nuevo conocimiento (Smith et al., 2003). Por este motivo, la sintomatología de las enfermedades genéticas no puede ser omitida en los análisis con perspectivas sistémicas. Sin embargo, la carencia prolongada de un lenguaje estandarizado en la comunidad médica para representar esta información ha supuesto un

obstáculo para avanzar en este terreno. Las descripciones de los síntomas de los pacientes se suelen expresar en texto libre dando lugar a un elevado grado de ambigüedad. Este hecho ha constituido una gran dificultad para avanzar en la comparación y el análisis automatizado de esta información. Si se realiza una consulta, tanto en la literatura científica como las descripciones de enfermedades en bases de datos públicas (OMIM, Orphanet, GeneCards,...), se pueden encontrar fácilmente distintas expresiones que hacen referencia a un mismo fenotipo o términos que pueden tener multitud de significados (Robinson, 2015). Por ejemplo, un síntoma característico del síndrome Laurin-Sandrow (ORPHA: 2378) es la polidactilia (*polydactyly*), que consiste en la existencia de un número de dedos anormal en manos o pies. Al buscar en la literatura encontramos varias formas de hacer referencia a este síntoma *mirror-image polydactyly* o *mirror hand polydactyly* (Innis and Hedera, 2004), *polysyndactyly of hands* (Mariño-Enríquez et al., 2008). La comparación y el agrupamiento de pacientes y enfermedades con esta sintomatología supondría un problema sin un consenso en la nomenclatura. Para resolver estos problemas del lenguaje natural surge la necesidad de definir un vocabulario normalizado que permita la representación de un dominio de conocimiento que es objeto de estudio. Alcanzar este objetivo facilitará la integración de los perfiles fenotípicos y su asociación a los distintos elementos que van a ser explorados en esta tesis tales como enfermedades, variaciones genéticas, genes y/o individuos. Este proceso de estandarización posibilita la integración de información fenotípica con los distintos interactomas, así como su análisis, permitiendo así una comprensión más precisa de los procesos moleculares involucrados en estos fenotipos.

Sin embargo, un conjunto normalizado de términos no es suficiente para representar toda la complejidad intrínseca de las características observadas en enfermedades genéticas. Lo que caracteriza a los sistemas complejos no son solo los conceptos que los integran, sino las relaciones existentes entre ellos. Siguiendo con el ejemplo anterior, otra propiedad de la información fenotípica es que la variabilidad de estos síntomas permite agruparlos en categorías, es decir, en síntomas más generales. Por ejemplo, la polidactilia es un fenotipo relativamente general, pero podría ser clasificado como “alteración de las extremidades” junto con otros síntomas que afecten a extremidades superiores o inferiores. Pero este fenotipo puede llegar a ser más específico, por ejemplo la “polidactilia post axial en manos” o “polidactilia pre axial” en manos son fenotipos aún más específicos pero siguen perteneciendo a la misma categoría. Por lo tanto, el modelo debe considerar el grado de especificidad de las relaciones entre los conceptos que representa. La necesidad de proporcionar una visión abstracta del conocimiento no es exclusiva de la biomedicina, ni de la biología. En multitud de disciplinas se requiere de una perspectiva simplificada de

un universo de discurso con la máxima fidelidad posible; las ontologías son herramientas útiles para cubrir esta necesidad (Hastings, 2017).

El término ontología es relativamente amplio y su definición puede diferir en función de la disciplina. De hecho, aunque este concepto se ha utilizado ampliamente en ciencias de la computación su origen es más antiguo. Etimológicamente la palabra ontología proviene del griego *onto* (lo que existe) y *logía* (estudio) y conforma la rama más relevante de la metafísica que investiga la categorías básicas del ser y cómo se relacionan entre ellas (Gruber, 1995). Las ciencias de la computación han incorporado este término para hacer referencia a la representación abstracta de todo lo que existe en el área de estudio de interés (Gruber, 1995; Hoehndorf et al., 2015). De forma muy resumida, una ontología se compone de clases, relaciones, un dominio o vocabulario y axiomas (Hastings, 2017; Hoehndorf et al., 2015).

Para diseñar una ontología, es necesario identificar cada una de las clases contenidas en el universo que se quiere representar, así como definir las relaciones entre dichas clases. Un ejemplo sencillo podría ser una ontología para definir las interacciones entre proteínas humanas. Para ello sería necesario definir la clase “proteína” que va a agrupar a todas las proteínas humanas y la relación “interacciona con” para definir la interacción física de dos proteínas, es decir, las clases y las relaciones conforman el conjunto de elementos que vamos a necesitar para definir el universo que queremos representar. Además, es necesaria la definición de un vocabulario normalizado para etiquetar cada una de esas clases, por ello, cada clase incorpora descripciones y sinónimos para hacer comprensible su significado a un lector humano. Por último, y más importante, es necesario definir el conjunto de axiomas que definan las reglas lógicas para relacionar las clases.

1.4. Ontologías biomédicas

Dentro de la biomedicina y la biología se ha extendido el uso de estructuras también denominadas ontologías. Las ontologías biomédicas (OBO) constan de un conjunto de conceptos organizados de forma jerárquica. El ejemplo más consolidado de ontología biomédica es la Gene Ontology (GO), que es una herramienta que proporciona un vocabulario estandarizado para describir las funciones biológicas en las que están involucradas proteínas de distintos organismos (Consortium, 2000; Camon et al., 2004; Thomas, 2017).

1.4.1. Gene Ontology: vocabulario organizado de referencia

Gene Ontology es una de las primeras ontologías biomédicas y ha sido un modelo de referencia para otros recursos que se han desarrollado con posterioridad, como por ejemplo *Human Phenotype Ontology* (HPO), que se describirá en secciones posteriores. Una de las motivaciones de esta herramienta fue el estudio de las funciones de genes a través de distintas especies. La definición de un conjunto de términos para describir la función de un gen dentro de un organismo permite compartir, comparar y analizar la información. Esto permite indagar en los procesos biológicos conservados evolutivamente y los que son comunes a organismos separados filogénicamente (Consortium, 2000; Thomas, 2017). Esto explica que las primeras versiones de GO estuvieran focalizadas en describir las funciones de los genes de varios organismos modelos como el ratón (*Mouse Genome Informatics*), levadura (*Saccharomyces Gemone Database*) o la mosca de la fruta (*FlyBase*). Hoy en día existen perfiles funcionales para alrededor de 30 especies en GO.

Anteriormente se mencionaron las ventajas de usar un lenguaje común con el objetivo de integrar información e incrementar el valor del conocimiento de un área. Para conceptualizar el "universo" de conocimiento compuesto por funciones biológicas, se consideró que para describir el papel del producto de un gen dentro de la célula se deben especificar: i) las funciones moleculares en las que está involucrado, ii) la localización celular donde se expresa y iii) el proceso biológico global en el que las funciones moleculares de dicho gen están contenidas (Thomas, 2017). Con este objetivo, GO proporciona un vocabulario normalizado, organiza los elementos de este vocabulario jerárquicamente formando un grafo acíclico dirigido (GAD) y los agrupa en tres dominios principales que podrían considerarse tres ontologías independientes: procesos biológicos, funciones moleculares y componentes celulares (Consortium, 2000; Smith et al., 2003). Los términos aparecen relacionados entre sí usando varios tipos de relaciones, siendo *is_a* las más común y que se usa para indicar que un término es subclase de otro. La Figura 1.2 muestra un resumen de las relaciones más comunes entre términos GO junto con *is_a*. A parte de los tipos de relaciones incluidos en la Figura 1.2 existen otras como *has_part* que es la relación complementaria a *part_of* desde la perspectiva de la clase ascendente.

El proyecto GO surgió con el objetivo de permitir la definición de perfiles funcionales de proteínas mediante el uso de un vocabulario normalizado de términos (denominados comúnmente términos GO). Al conjunto de relaciones entre términos GO y genes o proteínas se les denomina anotaciones, las cuales tienen asociadas un código de evidencia que indica la procedencia de la relación. El grupo de anotaciones mayoritario en GO corresponde a aquellas etiquetadas como IEA (*Inferred from Electronic Annotation*), que son

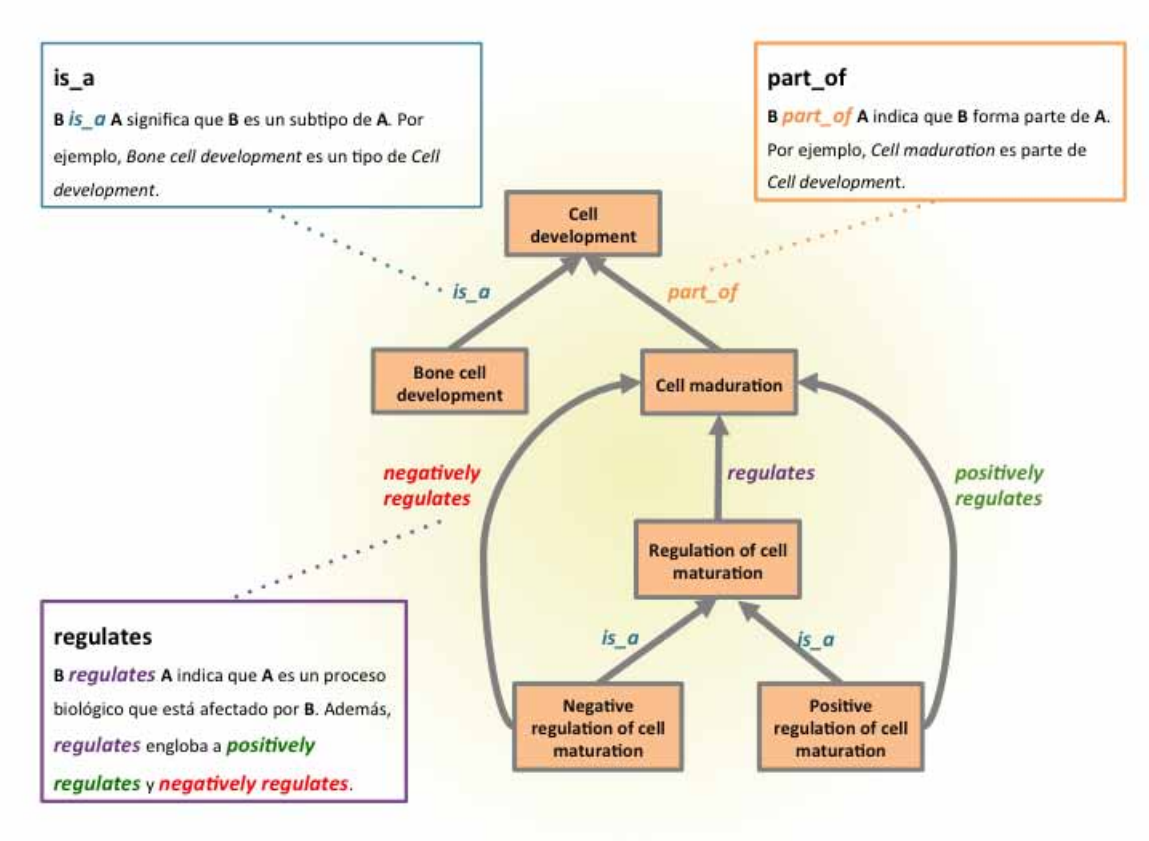


Figura 1.2: Tipos de relaciones entre términos de GO. Entre términos GO pueden existir varios tipos de relaciones. Esta figura muestra las más comunes: *is_a*, *part_of* y *regulates*.

aquellas que no han sido incluidas manualmente. El resto se clasifican en función de su procedencia: experimentos, análisis computacionales, o literatura.

El proyecto *Gene Ontology Annotation* (GOA) desarrollado por el EBI (*European Bioinformatics Institute*), asociado con el proyecto GO, es una colaboración entre GO Consortium y UniProt con el propósito de establecer anotaciones entre términos GO y códigos UniProt (Barrell et al., 2009; Camon et al., 2004). Desde el año 2004 y hasta la fecha, GOA ha proporcionado más de 368 millones de anotaciones para casi 54 millones de proteínas en más de 480.000 grupos taxonómicos. Aunque las asociaciones se han realizado entre proteínas y términos, actualmente existe una iniciativa para anotar ADN no codificante (Huntley et al., 2015).

Tanto GO como otros vocabularios que se construyeron inicialmente han supuesto un referente para creación de otras muchas herramientas con estructura similar. La creación de GO permite la definición del formato para representar estos vocabularios, denominado *Open Biomedical Ontology Format* (OBO Format). En la página web de OBO Foundry (http://www.obofoundry.org/) se puede encontrar el listado actual de todos los vocabularios disponibles, como Chebi, DO (*Disease Ontology*) o SO (*Sequence Ontology*).

1.4.2. Ontologías de fenotipos

Al inicio de esta sección se mencionaron las limitaciones para describir las sintomatologías observadas en enfermedades usando un vocabulario estandarizado. Representar esta información en un formato normalizado implica una revisión de la literatura y un curado manual. En los últimos años, han surgido varias iniciativas para facilitar el análisis e integración de esta información. Una de las aproximaciones se ha basado en el uso de minería de textos para extraer los fenotipos presentes en las descripciones de enfermedades incluidas, por ejemplo, en OMIM (van Driel et al., 2006; Nakazato et al., 2009) o en historiales médicos (PDN) (Hidalgo et al., 2009). Otra iniciativa más extendida y reciente consiste en el análisis de EHR (*Electronic Healh Records*) para obtener información de pacientes (Jensen et al., 2012). Sin embargo, la falta de un vocabulario común para representar la información sigue siendo una limitación. Para solventar esta limitación algunos autores han optado por usar una nomenclatura propia (PhenoDB) (Lussier and Liu, 2007) u otros recursos como MeSH (*Medical Subject Heading*) (Nakazato et al., 2009), ICD-9 (Hidalgo et al., 2009) o UMLS (*Unified Medical Language System*) (Bodenreider, 2004).

UMLS es una de las ontologías más populares y se compone de un conjunto de ficheros y *software* para integrar varios vocabularios. En total, UMLS se compone de alrededor de 2 millones de nombres para unos 900.000 conceptos de interés, más de 60 familias de

vocabularios biomédicos y 12 millones de relaciones entre conceptos. Entre los vocabularios integrados se incluyen: NCBI taxonomy, Gene Ontology, *Medical Subject Headings* (MeSH), OMIM y *Digital Anatomist Symbolic Knowledge Base* (Bodenreider, 2004).

Tanto MeSH como SNOMED CT están integradas en UMLS, sin embargo, pueden usarse independientemente. MeSH es un vocabulario normalizado elaborado y mantenido por la *National Library of Medicine* (NLM) y consiste en un conjunto de términos en una estructura jerárquica que permite realizar búsquedas en varios niveles de especificidad. MeSH consta de 27.883 descriptores con más de 87.000 términos (o sinónimos). Además de usarse en el contexto médico, MeSH es útil para indexar los conceptos médicos de la literatura incluida en MEDLINE (Coletti et al., 2001).

Por otro lado, *Systematized Nomenclature of Medicine – Clinical Terms* o SNOMED CT (Brown et al., 2006) contiene alrededor de medio millón de términos relativos a enfermedades, anatomía, morfología, fármacos, etc. La estructura de esta ontología es similar a otras ontologías biomédicas (jerárquica) y su uso en el campo médico y académico es bajo licencia. En SNOMED CT, la información biomédica está representada a partir de tres tipos de elementos: Conceptos, Descripciones y Relaciones. Los conceptos se agrupan en varias ramas, en función de la clase a la que pertenezcan: patologías (|Clinical finding|), especie (|Organism|) o tratamiento (|Procedure|). Además, dichos conceptos poseen identificadores numérico únicos y se organiza jerárquicamente desde el más general (siendo la raíz global el término |SNOMED CT concept|) al más específico. Los términos se relacionan a partir de asociaciones *is_a* (véase apartado 1.4.1) que permiten relacionar conceptos entre sí para proporcionar definiciones formales y otras propiedades a los conceptos (Brown et al., 2006; Ruch et al., 2008).

MedDRA (http://www.meddra.org/) es una herramienta usada en el ámbito médico, que de forma similar a los recursos anteriores, proporciona un conjunto de términos con información biomédica y clínica. La estructura es similar a los vocabularios vistos hasta ahora, es decir, jerárquica. En el caso de MedDRA los términos se organizan en una estructura que posee cinco niveles de especificidad: *System Organ Class* (SOC), *High Level Group Term* (HLGT), *High Livel Term* (HLT), *Preferred Term* (PT) y *Lowest Level Term* (LLT). El nivel más específico es LLT, que agrupa a más de 70.000 términos. Cada elemento de LLT está asociado con un PT, que es un descriptor que hace referencia a un síntoma, enfermedad diagnosticada, procedimiento quirúrgico o historial familiar. Los niveles más generales HLT y HLGT agrupan los términos de PT basándose en conceptos como anatomía, patología, fisiología, etiología o función. Por último, todos los términos de la ontología se agrupan en términos SOCs que hacen referencia a la etiología (ej: infecciones), anatomía

(ej: enfermedad gastro-intestinal) o propósito (ej: procedimiento quirúrgico) (Brown et al., 1999).

1.4.2.1. Human Phenotype Ontology (HPO)

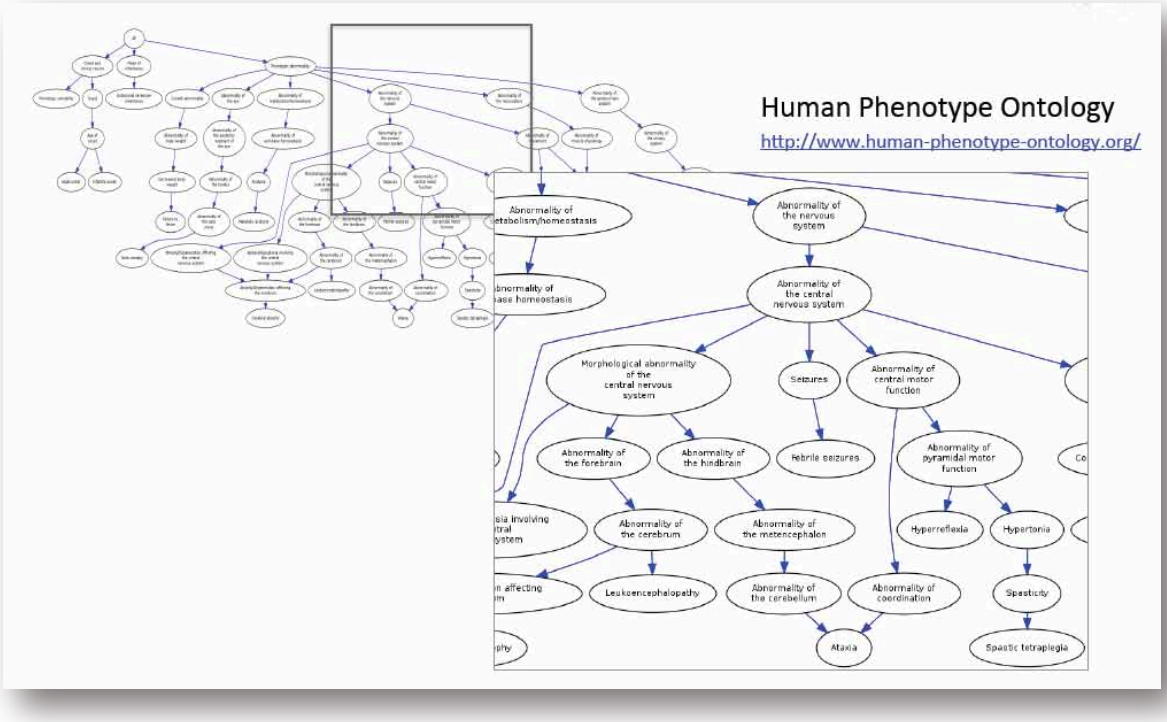


Figura 1.3: ***Human Phenotype Ontology.*** Esta figura fue generada a partir de capturas de las herramienta Phenomizer (Köhler et al., 2009) y muestra como se organizan los fenotipos en HPO.

Human Phenotype Ontology (HPO) organiza jerárquicamente un vocabulario de términos que corresponden a fenotipos patológicos asociados con alguna enfermedad genética (Robinson et al., 2008) siguiendo el formato usado para GO (OBO *Format*) (Figura 1.3). A pesar de que han surgido varias iniciativas para trabajar de forma automatizada con síntomas de enfermedades, la que mayor trascendencia ha tenido ha sido HPO. Quizás gran parte de su éxito se deba a que fue diseñada con una estructura similar a la GO, en la que cada término hace referencia a un fenotipo patológico asociado con alguna enfermedad. Por otra parte, HPO se construyó a partir de los datos de OMIM, haciendo uso de las especificaciones de las enfermedades incluidas en esa base de datos, por lo tanto, la ontología

estaba, ya en sus inicios, ampliamente enriquecida de anotaciones. La ontología consiste en aproximadamente 10.000 fenotipos patológicos que se agrupan en una rama principal denominada *Phenotypic Abnormality* que constan a su vez de 24 sub-ramificaciones más. Los fenotipos se asocian mayoritariamente con relaciones *subclass_of* que es equivalente a *is_a* (Figura 1.2), típica de GO, aun así, ha ido incorporando otros tipos de relaciones como *has_part*, para indicar que un término se compone de varios elementos.

1.4.2.2. Definición de perfiles fenotípicos

HPO proporciona un amplio y variado conjunto de términos para la construcción de perfiles fenotípicos asociados a enfermedades, genes, pacientes o variaciones genéticas. Un perfil fenotípico se compone de aquellos términos HPO que mejor se ajustan a la descripción o sintomatología de una enfermedad, gen, paciente, etc. Gran parte de los perfiles fenotípicos de enfermedades se construyeron a partir de las descripciones de OMIM (Robinson et al., 2008). La Figura 1.4 (paneles A y B) muestra un ejemplo de como, a partir de la especificación de la enfermedad "Kifafa" (*Kifafa seizure disorder*) en OMIM (OMIM 245180), es posible construir su perfil fenotípico seleccionando los síntomas presentes en su descripción, que corresponden a términos HPO.

Aunque los perfiles fenotípicos se componen de los términos más específicos asociados a una enfermedad, las anotaciones en HPO (o cualquier ontología con estructura similar) implican relaciones entre estas enfermedades y todas las categorías ascendentes a los términos que componen un perfil. La Figura 1.4C muestra una visión más completa del conjunto de términos de la enfermedad, que incluye los fenotipos más específicos (**anotaciones directas**) y las **anotaciones inferidas**. Aunque en la figura se muestre solo una selección de relaciones inferidas, el perfil fenotípico completo se compone de todos los fenotipos ascendentes, desde las anotaciones directas hasta la raíz (*Phenotypic Abnormality*). Esto va a permitir establecer relaciones teniendo en cuenta todos estos fenotipos, y de esta forma, podrían emerger un gran número de relaciones que estaban pasando desapercibidas. Las relaciones entre genes o enfermedades y términos HPO (o GO) se incluyen en ficheros de anotaciones. La sintaxis de estos ficheros es similar a la usada por el proyecto *Gene Ontology Annotation*, es decir GAF 2.0 (Huntley et al., 2015) y suelen especificar solo las anotaciones directas, puesto a partir de la ontología se puede inferir el resto de anotaciones.

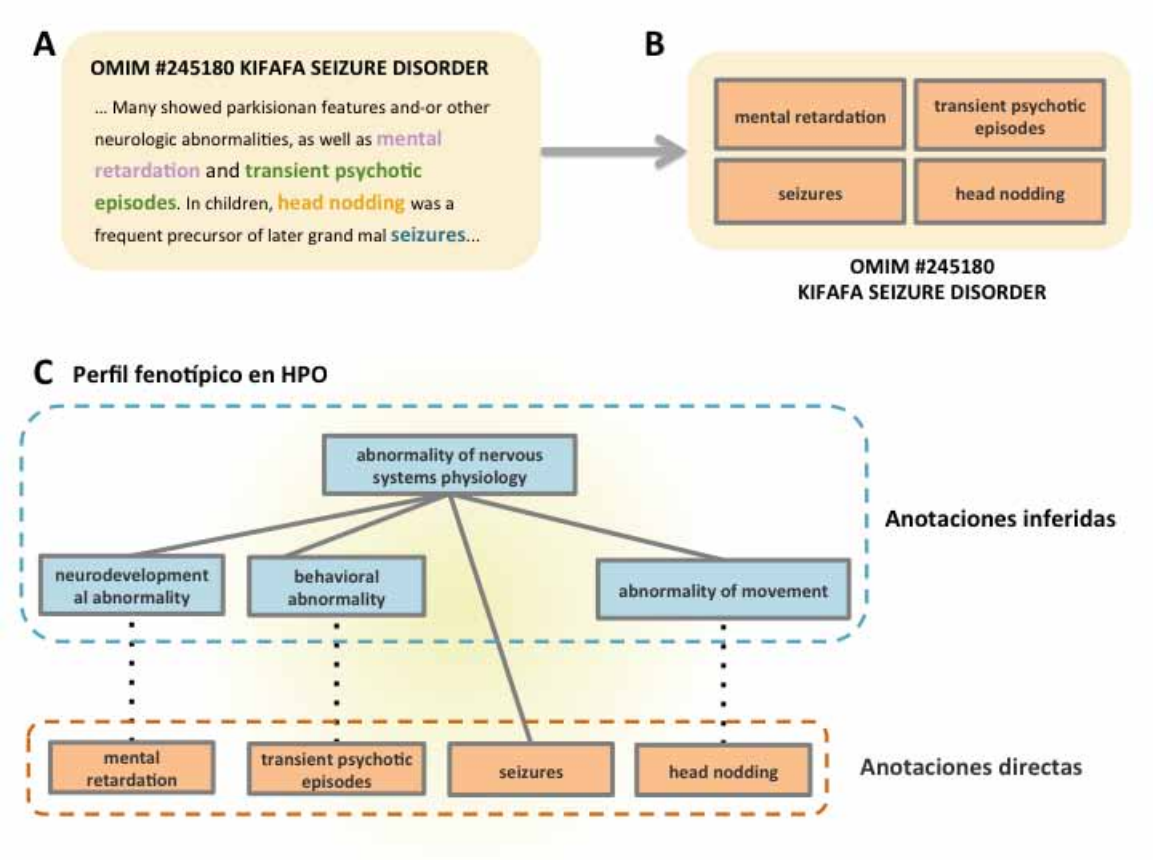


Figura 1.4: Definición de perfil fenotípico. La descripción de las enfermedades incluidas en OMIM permite la definición de perfiles fenotípicos si se dispone de un vocabulario normalizado. **A.** Descripción de la enfermedad Kifafa procedente de OMIM. **B.** Perfil fenotípico seleccionando los términos que describen la enfermedad. **C.** Perfil fenotípico en HPO teniendo en cuenta las anotaciones directas (naranja) y el resto de anotaciones inferidas (azul). Las relaciones inferidas se establecen entre todos los términos HPO desde los más específicos hasta la raíz.

1.4.2.3. Similitud semántica en ontologías biomédicas

El uso de ontologías biomédicas permite que emerjan asociaciones no observadas previamente. Al tratarse de estructuras jerárquicas es posible relacionar fenotipos a partir de un ancestro común, que será un fenotipo más inespecífico o genérico. Por ejemplo, en la Figura 1.5 se muestran dos enfermedades "cefalea en racimos" (OMIM 119915) y "malformación de Chiari" (OMIM 207950) relacionadas con los fenotipos "cefalea en racimos" (Lipton et al., 2004) y "neuralgia occipital" (Stevenson, 2004), respectivamente. El término "Dolor de cabeza" es un ancestro común en la ontología para ambos síntomas. Los dos términos hacen referencia a manifestaciones distintas de dolor de cabeza; la "cefalea en racimos" presenta dolores extremadamente intensos, en episodios (entre 15-180 minutos) y localizada en la zona de la cara y las órbitas (Lipton et al., 2004). Por otro lado, la "neuralgia occipital" implica dolores crónicos en la parte superior del cuello, la espalda y detrás de los ojos, zonas es las que se encuentran el nervio occipital mayor y menor. Las relaciones entre ambas enfermedades emergen gracias al uso de esta estructura. Si se utilizara solo un vocabulario de términos (no relacionados entre sí), establecer la relación entre estas enfermedades (o fenotipos) no sería posible, al menos de forma sistemática.

Contenido informativo: Entre el ejemplo de la Figura 1.5A y el representado en los paneles B y C se observan diferencias a simple vista. La narcolepsia (OMIM 609039) y la enfermedade de Niemann-Pick (OMIM 257220) manifiestan el mismo fenotipo (Cataplexia), sin embargo, no ocurre lo mismo en el ejemplo de la Figura 1.5 (B y C). Por lo tanto, el valor de similitud para el primer ejemplo (Figura 1.5A) debe ser mayor que el valor de similitud de la relación entre "cefalea en racimos" (OMIM 119915) y "malformación de Chiari" (OMIM 207950). Para medir la similitud entre términos es posible utilizar una medida que represente la proximidad entre dos fenotipos o conjuntos de fenotipos en una ontología. En la literatura se pueden encontrar varias propuestas para comparar perfiles usando ontologías biomédicas (Jiang and Conrath, 1997; Pesquita et al., 2009; Resnik, 1995); sin embargo, en este trabajo se han usado medidas basadas en la propuesta por Resnik (Resnik, 1995) que usa el concepto de **contenido informativo** o **information content** (IC) para comparar términos.

IC o contenido informativo es un concepto usado en teoría de la información y proporciona un valor numérico que indica la **especificidad** de un término. En este trabajo, un término (o fenotipo en este caso) se considera específico si su frecuencia es baja dentro de un conjunto de datos. Por ejemplo, un fenotipo que es observado en una sola enfermedad tendrá una alta especificidad. Por el contrario, un término HPO como "anomalías del sistema nervioso" tendrá asociado un gran número de enfermedades, implicando una baja

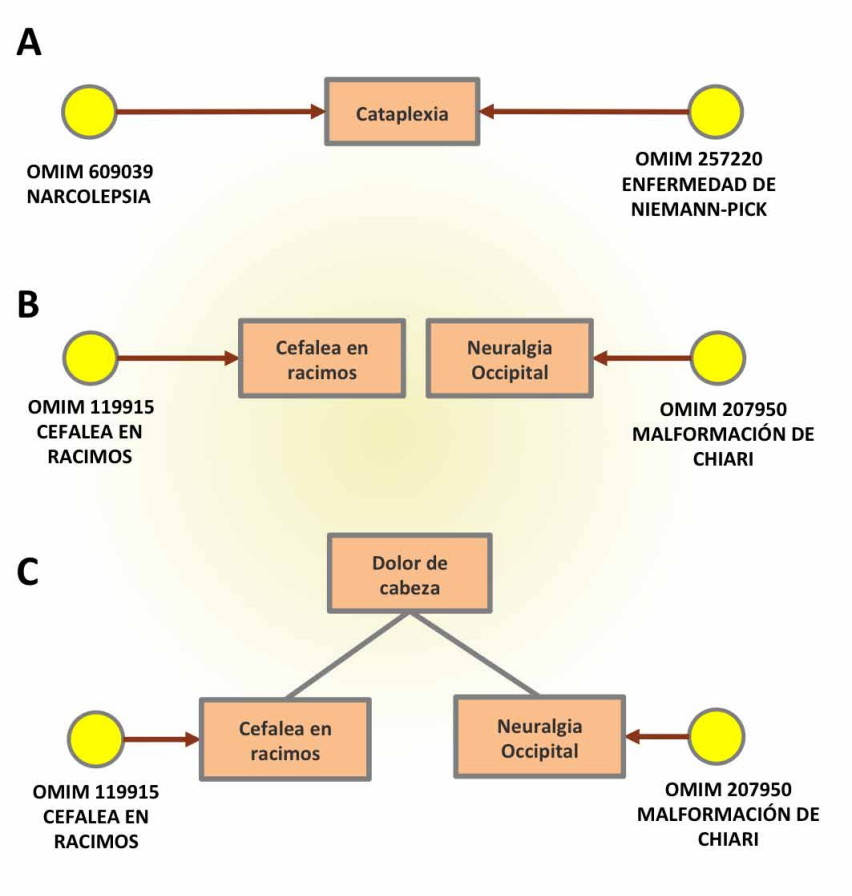


Figura 1.5: Relaciones entre enfermedades a partir de sus fenotipos. A. La narcolepsia y la enfermedad de Niemann-Pick se asocian con la catalepsia por lo tanto pueden relacionarse fenotípicamente. **B.** La cefalea en racimos y la malformación de Chiari por el contrario se asocian con términos HPO distintos. **C.** Por lo tanto, la relación entre esos dos términos será a partir de un término más general (Dolor de cabeza).

especificidad. Un fenotipo muy específico será además más informativo, ya que apuntará a un conjunto de enfermedades muy concretas. Por lo tanto, el IC proporciona un valor de especificidad y de información, como su propio nombre indica, de cada término de la ontología.

En una ontología, el cálculo de IC se realiza como se indica en el ejemplo de la Figura 1.6. Junto a cada término de la figura se muestran tres valores: el número de elementos asociados al término (N), la probabilidad o frecuencia (p) y el contenido informativo (IC).

El número de elementos asociados a un término ($N(t)$), en este caso, es el número de enfermedades anotadas a un término t. En las ontologías biomédicas se sigue una regla denominada *true path rule*, que indica que todos los objetos (enfermedades) anotados a un término, lo estarán también con sus términos ascendentes. Análogamente, el conjunto de anotaciones de cada uno de los términos constará de las asociaciones directas y todas las asociaciones a términos descendientes. Un ejemplo de ello puede verse en la Figura1.6: "dolor de cabeza" tiene asociadas tres enfermedades (N=3), adiposis dolorosa, cefalea en racimos y malformación de Chiari, aunque solo la adiposis dolorosa está relacionada directamente con "dolor de cabeza", las otras dos (cefalea en racimos y malformación de Chiari) están anotadas a términos más específicos. Es por eso que el número de elementos (N) de la raíz "anomalías del sistema nervioso" es 5, aunque no haya ninguna enfermedad directamente asociada a ese término. De hecho, en toda ontología con estructura similar, el término más general o raíz tendrá asociado todos los objetos anotados en la ontología.

A partir del número de elementos asociados a un término (N) es posible obtener la probabilidad de cada término $p(t)$. En concreto, dicha probabilidad se calcula dividiendo el número de elementos asociados a cada término (N) entre el número total de anotaciones (5 en este caso) como muestra la leyenda de la Figura 1.6. El IC es el valor inverso a la probabilidad (p) y matemáticamente se define como $IC(t)$ $log(p(t))$. Siguiendo esta fórmula, la raíz ("anomalías del sistema nervioso") de la ontología tendrá un contenido informativo de 0, por lo tanto, no es nada informativa, puesto que todas las enfermedades se relacionan con dicho fenotipo y carece de especificidad. El contenido informativo aumenta en términos más cercanos a las hojas, y por lo tanto también la especificidad, como puede verse en la Figura 1.6. Por ejemplo, "neuralgia occipital" es un fenotipo específico puesto que solo está relacionado con la malformación de Chiari.

Similitud semántica entre perfiles fenotípicos: El valor de similitud semántica entre dos términos corresponde con el IC del ancestro común más informativo o MICA (*Most Informative Common Ancestor*). Para obtener el MICA entre dos términos, en primer lugar es necesario obtener el conjunto total de ancestros comunes a ambos. Sean t_i y t_j dos

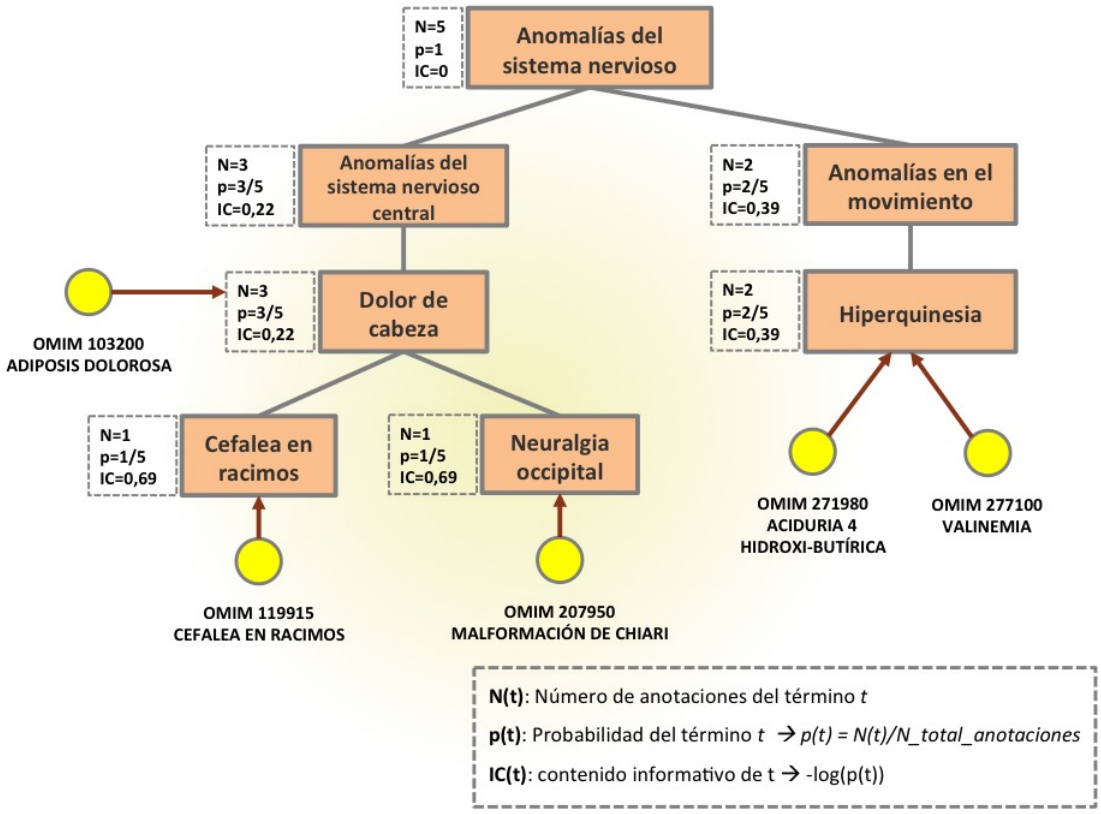


Figura 1.6: Calculo del contenido informativo (IC). El contenido informativo se calcula a partir de la probabilidad de cada término p(t), siendo t un término de la ontología, como indica la leyenda de la figura. A su vez, la probabilidad depende del número de anotaciones de cada término (N). Las anotaciones de cada término (N(t)) son el total de las anotaciones directas y todas las enfermedades relacionadas con los fenotipos descendientes de t.

términos de la ontología, los ancestros comunes son el resultado de la intersección de los ancestros de t_i y t_j. Formalmente:

$$ancestroscomunes(t_i, t_j) \quad ancestros(t_i) \quad ancestros(t_j) \tag{1.1}$$

Por lo tanto, la similitud entre estos dos términos corresponde con el máximo IC del conjunto de ancestros comunes:

$$sim(t_i, t_j) \quad MICA(t_i, t_j) \quad \max_{t \ ancestroscomunes(t_i,t_j)} IC(t) \tag{1.2}$$

Volviendo al ejemplo de la Figura 1.6 el valor de similitud semántica entre "cefalea en racimos" y "neuralgia occipital" será 0.22 que es el IC asociado al ancestro común más informativo que es "dolor de cabeza". Sin embargo, los genes y enfermedades están relacionados con perfiles ontológicos formados por varios términos. Por ello, la similitud entre dos genes (o dos enfermedades) se obtiene a partir de la comparación de dos perfiles. Pero antes de definir el cálculo de la similitud entre perfiles, se define como la similitud entre un término t y un conjunto de términos S:

$$sim(t, S) \quad max_{t_i \ S} sim(t, t_i) \tag{1.3}$$

Para calcular la similitud entre dos perfiles, en esta tesis se han usado dos medidas diferentes para la similitud entre perfiles funcionales (GO) y entre perfiles fenotípicos (HPO). La similitud funcional (GO) entre genes se obtiene a partir de la medida propuesta por Resnik. Este método ha dado buenos resultados en varios trabajos (Lord et al., 2003; Sevilla et al., 2005; Xu et al., 2008)) y se obtiene con la siguiente fórmula:

Resnik:

$$sim_{max}(S_i, S_j) \quad \max_{t_i \ S_i} sim(t_i, S_j) \tag{1.4}$$

La similitud fenotípica (HPO) se obtiene a partir de la medida también usada por los autores de HPO (Couto et al., 2007; Köhler et al., 2009), la cual está basada en la media de las comparaciones de dos grupos de términos. Sea S_i y S_j dos perfiles fenotípicos:

Resnik media:

$$sim_{media}(S_i, S_j) \quad \frac{_{t_i \ S_i} \max_{t_j \ S_j} sim(t_i, t_j)}{S_i} \tag{1.5}$$

La Ecuación 1.5 tiene el inconveniente de que no es simétrica, es decir $sim(S_i, S_j)$! $sim(S_j, S_i)$ por ello se usó una versión simétrica de dicha formula:

Resnik simétrica:

$$sim_{simetrica}(S_i, S_j) \quad \frac{sim_{media}(S_i, S_j) \quad sim_{media}(S_j, S_i)}{2} \tag{1.6}$$

1.5. Desarrollos anteriores al presente trabajo

En trabajos previos se ha observado que los genes que manifiestan fenotipos comunes tienden a estar relacionados funcionalmente, ya sea porque interaccionan físicamente o participan en la misma ruta metabólica (van Driel et al., 2006; Hidalgo et al., 2009). En las enfermedades complejas suelen estar involucrados varios genes, que afectan a varios procesos biológicos a la vez. Comparar las enfermedades teniendo en cuenta la heterogeneidad de sus síntomas y la integración de los espacios fenotípicos con información funcional va a permitir priorizar los elementos que puedan estar involucrados en estas enfermedades (Oti et al., 2008). De hecho, varios autores han destacado la importancia de la definición de fenomas (Baker, 2013) que ayuden a estudiar las manifestaciones de los genotipos y la influencia del entorno (Butte and Kohane, 2006).

La minería de textos ha sido un recurso utilizado en varios trabajos para asociar conceptos biomédicos (fenotipos, funciones, drogas, ...) a genes o enfermedades. Hidalgo et al., 2009 extrae la información de registros de pacientes con el objetivo establecer relaciones entre enfermedades que co-ocurren con frecuencia y estudiar su comorbilidad. Gendoo (Nakazato et al., 2009) es una herramienta para la consulta de términos biomédicos MeSH asociados a genes y enfermedades. A cada par OMIM-MeSH establece un p-valor y le asocia un valor de especificidad. El uso de herramientas de minería de textos (*text-mining*) permite asociar perfiles fenotípicos a enfermedades y construir redes de enfermedades en las que las relaciones representan la similitud entre ellas (van Driel et al., 2006). Este trabajo establece un esquema para extraer información fenotípica y establecer relaciones a partir de dicha información. Los elementos principales de este esquema son: i) descripciones de las patologías que suelen estar incluidas en bases de datos como OMIM, la literatura o informes de pacientes, ii) un vocabulario normalizado de conceptos biomédicos (MeSH, SNOMED-CT, UMLS, ...), iii) herramientas de minería de textos y iv) medidas para comparar los perfiles fenotípicos. van Driel et al., 2006 usan la herramienta MMTx para extraer los fenotipos de las enfermedades incluidas en OMIM y esto les permite definir perfiles fenotípicos para cada una de ellas. Estos perfiles son vectores de términos UMLS, que se emplean para comparar las enfermedades midiendo el solapamiento entre los perfiles, usando el coseno del ángulo entre los vectores. Esta misma

medida es usada por Lage et al., 2007, siguiendo una metodología similar. En dicho trabajo se identifican los perfiles fenotípicos asociados a enfermedades y genes analizando la base de datos de OMIM para predecir interacciones entre proteínas que manifiestan fenotipos similares.

HPO surge en 2008 (Bauer et al., 2008) y su construcción siguió el mismo procedimiento que se detalla en el párrafo anterior: obtención de los perfiles fenotípicos a partir de la descripción de las enfermedades. Además, en este trabajo se proporciona una ontología propia (HPO), permitiendo definir de forma más completa el perfil fenotípico de las enfermedades y, como consecuencia, establecer un mayor número de relaciones entre ellas. El desarrollo de este tipo de recursos propicia el desarrollo de protocolos, herramientas, algoritmos de priorización y análisis estadístico.

En la actualidad, se han desarrollado numerosas herramientas que hacen uso de ontologías, sobre todo de GO. Su uso está muy extendido e incluso está disponible en plataformas como Bioconductor (https://www.bioconductor.org/). Algunos ejemplos son GOSemSim (Yu et al., 2010), topGO (disponible en `http://bioconductor.org/packages/release/bioc/html/topGO.html`), goSTAG (Bennett and Bushel, 2017), HPOSim (Deng et al., 2015) o PCAN (Godard and Page, 2016). También, existen otras muchas herramientas desarrolladas en otros lenguajes de programación desde aplicaciones de escritorio hasta herramientas web para la consulta de información y la aplicación de algoritmos. Estos recursos facilitan un acercamiento de dicha información a miembros de la comunidad científica, no familiarizados con el uso de técnicas más sofisticadas de análisis bioinformático. Por ejemplo, se destaca Phenomizer (Köhler et al., 2009) como recurso que usa información (pato)fenotípica y permite la consulta de genes, enfermedades y perfiles fenotípicos para obtener una lista de enfermedades o genes similares ordenados por valor de similitud. Otras herramientas amplían la información fenotípica con datos de otras especies. Aparte de HPO, existen otras ontologías con fenotipos ortólogos, usadas en PhenomeNET (Hoehndorf et al., 2011), que es una herramienta con la misma finalidad que Phenomizer, que integra información biomédica de varios organismos modelo (mosca, pez cebra, ratón, etc.)

En secciones anteriores se comentó la importancia de la integración de información para analizar las relaciones patológicas y estudiarlas en el contexto molecular. GeneMania (Warde-Farley et al., 2010) permite la integración de relaciones entre un conjunto de productos génicos. Aunque no incorpora información fenotípica, tiene en cuenta diversos interactomas y permite visualizar una serie de relaciones en una misma red a partir de un conjunto de genes. Como se muestra en la Figura 1.7, estas relaciones se estable-

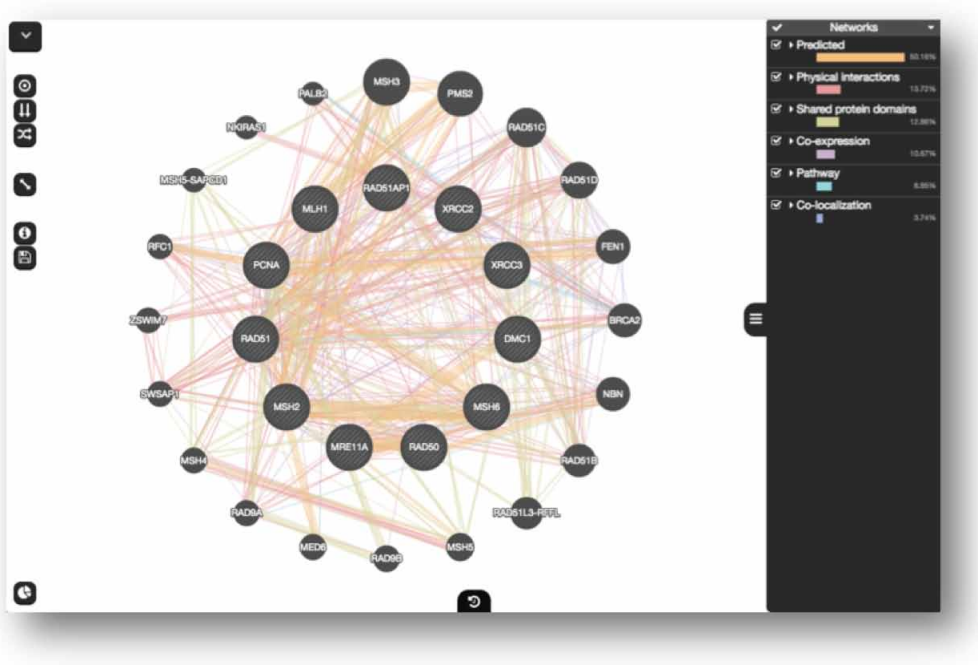


Figura 1.7: Ejemplo de la consulta de un conjunto de genes en GeneMania (Warde-Farley et al., 2010).

cen en función de la co-expresión, las interacciones físicas, las relaciones metabólicas, la co-localización y la similitud de dominios entre los productos génicos.

La información fenotípica es crucial para conocer los mecanismos moleculares que están tras estas patologías. Por lo tanto, también lo es facilitar la integración de información, así como facilitar el acceso de la comunidad científica a esa información. Sin embargo, las relaciones entre enfermedades y fenotipos que encontramos en HPO proceden de OMIM, es decir, solo vamos a encontrar perfiles de enfermedades asociados con genes en los que una mutación se asocia con una enfermedad. Por esta razón, conseguir una visión global de las manifestaciones fenotípicas relacionadas con mutaciones a lo largo del genoma supondría un gran potencial para el estudio de las enfermedades genéticas. Hay varias iniciativas para establecer relaciones entre fenotipos y variaciones genéticas. Bases de datos como ClinVar (Landrum et al., 2014) o DECIPHER (Firth et al., 2009) son repositorios de variaciones genéticas (tanto SNP como estructurales) que están incorporando identificadores de términos HPO, lo cual pone de manifiesto la utilidad de esta ontología y que su uso es cada vez más extendido.

En definitiva, el uso de los perfiles fenotípicos puede ser de gran utilidad para analizar las relaciones entre enfermedades. Las relaciones entre los genes incluidos en el *diseaso-*

ma pueden enriquecerse considerando las enfermedades como grupos de síntomas en lugar de como entidades indivisibles. El estudio de estas relaciones dentro del contexto molecular va a permitir estudiar los procesos biológicos alterados asociados con estos fenotipos. Este puede ser un posible abordaje para entender mejor el contexto de algunas enfermedades raras. Dadas las dificultades de su estudio experimental (debidas fundamentalmente al escaso número de pacientes), la posibilidad de estudiar la similitud entre sus sintomatologías y las de otras enfermedades humanas podrían desvelar similitudes funcionales que apunten a procesos biológicos concretos. Por tanto, queda claro que para seguir avanzando en la caracterización, prevención e intervención de patologías humanas, tanto raras como prevalentes, es necesario facilitar el acceso de la comunidad científica a todos estos datos y también al uso de estos recursos bioinformáticos. Por otra parte, es también indispensable, que los profesionales de la clínica se familiaricen con el uso de una ontología de fenotipos humanos (como la HPO o futuras alternativas consensuadas por científicos y clínicos) para caracterizar de forma sistemática los síntomas observados durante la exploración de los pacientes. Este esfuerzo de estandarización permitiría comparar con más fiabilidad las patologías, los pacientes y las variaciones genéticas potencialmente patológicas identificadas tras la secuenciación de los genomas o exomas de estos pacientes.

Capítulo 2

Hipótesis y Objetivos

La integración de datos masivos generados a partir de técnicas experimentales de alto rendimiento facilita el estudio de las enfermedades genéticas o de baja prevalencia. Trabajos como el desarrollado por (Goh et al., 2007) resaltan la importancia de estudiar la complejidad de las relaciones entre los genes asociados a enfermedades. Este estudio muestra que las enfermedades generalmente manifiestan diversos síntomas y que distintas patologías se pueden asociar a fenotipos patológicos similares; asociados a alteraciones en procesos biológicos relacionados funcionalmente. Si bien uno de los grandes problemas de la integración de información (pato)fenotípica radica en lo sujeta que está al lenguaje natural, herramientas como HPO facilitan esta integración. Por esta razón, para profundizar en las relaciones fenotípicas entre enfermedades, en esta tesis se trabaja con la hipótesis de que:

El estudio de las enfermedades complejas a partir del perfil fenotípico que las describe y la integración de esta información en su contexto biomolecular permite identificar los mecanismos moleculares afectados en múltiples procesos patológicos.

Objetivo 1 – Analizar las relaciones fenotípicas entre genes: Definir los perfiles fenotípicos de cada gen y construir una red de relaciones fenotípicas entre genes. La red fenotípica resultante se comparará con las redes unipartitas entre genes previamente publicadas (Goh et al., 2007; Zhang et al., 2011), y se realizará una estimación global de las implicaciones funcionales de aquellos genes fenotípicamente similares.

Objetivo 2 - Desarrollar una herramienta para la consulta de relaciones fenotípicas: Este trabajo de análisis e integración de información generará una gran cantidad de información emergente, especialmente útil en el ámbito de la investigación biomédica. Por esta razón, el siguiente objetivo será desarrollar una herramienta que permita el acceso a estas relaciones y la información asociada. Las tareas en este punto consistirán en el

diseño de una base de datos que integre toda esta información y de una plataforma que permita la consulta de genes o enfermedades de interés.

Objetivo 3 – Combinar minería de textos con ontologías para establecer relaciones gen-enfermedad no incluidas en bases datos públicas: PhenUMA integra información de varias bases de datos. Sin embargo, mucha de la información biomédica presente en la literatura no lo está en dichos repositorios, como ocurre con bastantes genes asociados con el metabolismo de las aminas biogénicas. En este apartado, se hará uso de herramientas de minería de textos para extraer de la literatura relaciones entre estos genes y enfermedades. A continuación se identificará qué fenotipos pueden estar asociados a dichos genes. Y por último se procederá a establecer relaciones fenotípicas entre estos genes y su integración con información funcional asociada a ellos. El desarrollo de este objetivo puede servir de proyecto piloto para abordar estudios similares con otros módulos del metabolismo y la fisiopatología humana.

Objetivo 4 – Analizar las relaciones genotipo-fenotipo en un conjunto heterogéneo de pacientes con síndromes genómicos: Las enfermedades, en ocasiones, son la consecuencia de mutaciones que abarcan grandes regiones del genoma (duplicaciones o deleciones). DECIPHER es un repositorio que incluye las CNVs (variaciones en número de copias) y los fenotipos de un conjunto de pacientes. Para analizar la información fenotípica asociadas a estas CNVs, en este apartado se analizará la información genómica y fenotípica de los pacientes de la base de datos DECIPHER. Se estudiará la sintomatología común a los pacientes cuyas regiones afectadas son solapantes, para identificar regiones potencialmente patológicas. Este método podría facilitar el diagnóstico de otros pacientes afectados en los mismos fragmentos.

Capítulo 3

Material y Métodos

3.1. Test estadísticos

3.1.1. Test Mann-Whitney U

La prueba o test Mann-Whitney U (o Wilcoxon test) se usó para comparar distribuciones. Consiste en un test no paramétrico cuyo objetivo es validar la hipótesis nula (H_0) de que, teniendo la distribución de dos muestras, existe la misma probabilidad de seleccionar aleatoriamente un valor de la primera muestra que sea mayor que (o menor que) un elemento aleatorio de la segunda muestra. Este test está incluido en la mayoría de los paquetes estadísticos. En esta tesis se ha usado función *wilcox.test*, que implementa esta prueba para comparar dos distribuciones, y está contenido en el módulo de estadística *stats* de R.

3.1.2. Test hipergeométrico

El test hipergeométrico permite calcular la probabilidad de obtener específicamente k éxitos en n extracciones de una población con N elementos. Este test es usado con frecuencia para calcular si una población está muy representada o poco representada en una muestra. La probabilidad se obtiene partir de la función:

$$P \quad Value(X \quad k) \quad \frac{\binom{d}{k}\binom{N}{n}\binom{d}{k}}{\binom{N}{n}} \tag{3.1}$$

donde N es el número de elementos de la población, n es el número de la muestra extraída, d es el número de éxitos y k el tamaño de la muestra.

3.1.3. Corrección de significación en múltiples comparaciones

En experimentos en los que se llevan a cabo múltiples pruebas de comparaciones independientes, el número de comparaciones afecta a la probabilidad de considerar erróneamente un resultado como estadísticamente significativo (error tipo I) por lo tanto, es necesaria la corrección para reducir el número de falsos positivos (Noble, 2009).

El método más frecuente y más sencillo es la corrección de Bonferroni. El objetivo es controlar los falsos positivos (error de tipo I), es decir, las hipótesis erróneamente validadas. Formalmente, H_1, H_2, ..., H_m serán las hipótesis que quieren ser validadas y p_1, p_2, ..., p_m sus correspondiente p-valores. Boferroni rechaza la hipótesis si p_i $\frac{}{m}$ donde m es el número de hipótesis analizadas y es el umbral de confianza, que por convención suele ser 0,05 o 0,01. Todos los p-valores menores a $\frac{}{m}$ se consideran estadísticamente significativos (Noble, 2009). Este método es uno de los más conservadores y esta rigidez es mayor si el número de hipótesis que se quiere comparar aumenta. Cuando el conjunto de datos, y por lo tanto el número de hipótesis, es muy elevado, este test se hace más restrictivo, ya que a medida que aumenta el número de comparaciones (m) desciende el umbral ($\frac{}{m}$) lo que hará menos probable que se rechace la hipótesis nula (Shaffer, 1995).

El ajuste de Benjamini-Hochberg, trata de corregir esto, introduciendo el concepto de FDR (false discovery rate) como la proporción de hipótesis erróneamente rechazadas con respecto al conjunto total de hipótesis rechazadas; siendo $H_1, H_2, \ldots, H_m$ las hipótesis que quieren ser validadas y $p_1, p_2, \ldots, p_m$ sus correspondiente p-valores. Mediante el método Benjamini-Hochberg la hipótesis es rechazada si p_i $\frac{i}{m}$ (Benjamini and Hochberg, 1995).

En esta tesis, se ha usado mayoritariamente el método Benjamini-Hochberg para corregir los p-valores, ya que en algunos análisis el número de comparaciones era alto y se producía una penalización en el número de hipótesis validadas. Sin embargo, la aproximación de Bonferroni también se usó en algunas comparaciones.

3.2. Bases de datos

Para desarrollar los análisis y llevar a cabo los objetivos descritos en el capítulo anterior se ha hecho uso de varios repositorios públicos con información de enfermedades y de las asociaciones gen-enfermedad:

3.2.1. OMIM

OMIM (www.omim.org) es un catálogo de enfermedades genéticas y genes asociados con estas enfermedades. Aunque en sus inicios fue creada por el doctor McKusick-Nathans, ahora es gestionada por el *National Centre of Biotechnology Information* (NCBI) y la universidad Johns Hopkins de medicina (Hamosh et al., 2005). OMIM proporciona ficheros de texto con la información contenida en la base de datos. Concretamente, para este trabajo se ha usado el fichero - descargado en octubre de 2012 - morbidmap.txt para obtener 4.261 relaciones entre 3.794 genes y 3.486 enfermedades.

3.2.2. Orphanet

Orphanet (www.orpha.net) es un catálogo de enfermedades de baja prevalencia que incluye las bases moleculares de este tipo de patologías, genes relacionados y descripciones detalladas en varios idiomas. Orphanet proporciona toda la información disponible en la base de datos separada en varios ficheros en formato XML, accesibles desde www.orphadata.org. En este trabajo se ha usado el fichero *Diseases with their associated genes* para obtener las relaciones gen-enfermedad, concretamente 4.472 relaciones entre 2.614 genes y 2.555 enfermedades. Además, el fichero *Disorders, cross referenced with other nomenclatures* incluye las correspondencias entre identificadores Orphanum y otras bases de datos como OMIM, ID-9 o UMLS.

3.2.3. DECIPHER

DECIPHER (Firth et al. 2009), accesible desde https://decipher.sanger.ac.uk/, es una base de datos que permite registrar las mutaciones de pacientes afectados por enfermedades genéticas a clínicos de todo el mundo. Las variaciones en número de copias o *Copy Number Variations* (CNV) (frecuencia menor al 1 %) de pacientes con enfermedades de baja prevalencia se descargaron de la base de datos de DECIPHER, en mayo de 2014, a través de un acuerdo para el acceso a esos datos. Este fichero consta de las asociaciones entre las CNV detectadas y los síntomas observados en cada uno de los pacientes, para los que se ha consentido la distribución de dicha información. Con respecto a las mutaciones, este fichero proporciona la región cromosómica afectada, el tipo de mutación (deleción o duplicación), el tipo de herencia y los fenotipos, que es una información aportada por los clínicos que han llevado a cabo la exploración de los pacientes. Sin embargo, no todos los pacientes tienen información fenotípica. Los pacientes sin descripción fenotípica asociada no se incluyeron en el estudio. Además, se usaron solo las mutaciones detectadas con

arrays CGH, que conforman la mayoría de las muestras. Los datos resultantes constan de 6.564 pacientes con 9.186 mutaciones asociados a 1.860 fenotipos.

3.2.4. Database of Genomic Variants (DGV)

Para validar los análisis realizados con los datos de pacientes de DECIPHER se usaron las variaciones estructurales observadas en pacientes sanos procedentes de *Database of Genomic Variants* (DGV). DGV incluye la región cromosómica, el tipo de mutación, la referencia bibliográfica y la plataforma usada para el análisis de cada caso. Para este trabajo se usó el fichero GRCh37_hg19_variants_2014-10-16.txt descargado desde `http://dgv.tcag.ca/dgv/app/home`.

3.2.5. ClinVar

ClinVar (`http://www.ncbi.nlm.nih.gov/clinvar/`) es una base de datos que incluye relaciones entre variaciones y fenotipos (Landrum et al., 2014). La información que proporciona ClinVar procede de multitud de grupos y laboratorios quienes suben la interpretación patológica de una variante genética. Además, ClinVar incorpora el sentido clínico (*Clinical Significance*) de cada par variación-enfermedad: *pathogenic*, *likely pathogenic*, *benign*, *likely benign*, etc. Actualmente, en ClinVar se incluyen más de 158.000 registros involucrando a más de 125.000 variantes que afectan a alrededor de 26.000 genes (Landrum et al., 2016).

3.3. Redes e interactomas

A continuación, se describirán las características fundamentales de los repositorios de información utilizados:

3.3.1. Relación gen-enfermedad

Las asociaciones entre enfermedades y genes que se han tenido en cuenta proceden de OMIM u Orphanet. En estos repositorios se proporcionan asociaciones entre un gen y una enfermedad, basándose en la existencia de una mutación en un gen que pueda ser causante de dicha enfermedad. Los pares gen-enfermedad de ambas bases de datos se usaron separadamente, formando dos conjuntos de relaciones bipartitas, correspondiente a los *diseasomas* (véase Figura 1.1) que se han tenido en cuenta en esta tesis (Goh et al., 2007;

Zhang et al., 2011). MIM proporciona esta información en el fichero *MorbidMap* disponible desde su sitio web bajo demanda. De este fichero se extrajeron las relaciones entre identificadores OMIM de enfermedades e identificadores OMIM de genes de los cuales se obtuvo su código Entrez. Las relaciones entre enfermedades Orphanet y genes se obtienen de Orphadata (`http://www.orphadata.org/`) que es gestionado por Orphanet y que proporciona distintos ficheros en formato XML, en los que organiza toda la información de Orphanet: genes, síntomas, prevalencia, etc. En este trabajo se ha hecho uso del fichero *Disorders with their associated genes* que contiene las relaciones entre enfermedades raras y genes. El número de enfermedades OMIM son 3.132 relacionadas con 2.525 genes. Las relaciones derivadas de Orphanet incluyen 2.331 genes relacionados con 2.125 enfermedades.

3.3.2. Clasificación de relaciones gen-enfermedad

Para analizar tanto las relaciones gen-enfermedad así como las proyecciones unipartitas hemos clasificado ambas en función del número de relaciones entre ellos. Para ello introducimos los conceptos de enfermedad monogénica o poligénica si están relacionadas con un solo gen o varios respectivamente. Con respecto a los genes usamos los términos de monotrópico y pleiotrópico para indicar si el gen está asociado con una enfermedad o varias. Teniendo en cuenta esta nomenclatura, toda las relaciones gen-enfermedad se clasifican en cuatro grupos (Figura 3.1):

- MD-MG (*Monogenic Disease-Monotropic Gene*): enfermedad monogénica relacionada con gen monotrópico
- MD-PG (*Monogenic Disease-Pleiotropic Gene*): enfermedad monogénica relacionada con gen pleiotrópico
- PD-MG (*Polygenic Disease-Monotropic Gene*): enfermedad poligénica relacionada con gene monotrópico
- PD-PG (*Polygenic Disease-Pleiotropic Gene*): enfermedad poligénica relacionada con gen pleiotrópico

3.3.3. Redes unipartitas de genes y enfermedades

En teoría de redes una proyección, en este caso de una red bipartita, tiene como objetivo reducir la complejidad de la red dando como resultado otra red en la que existen

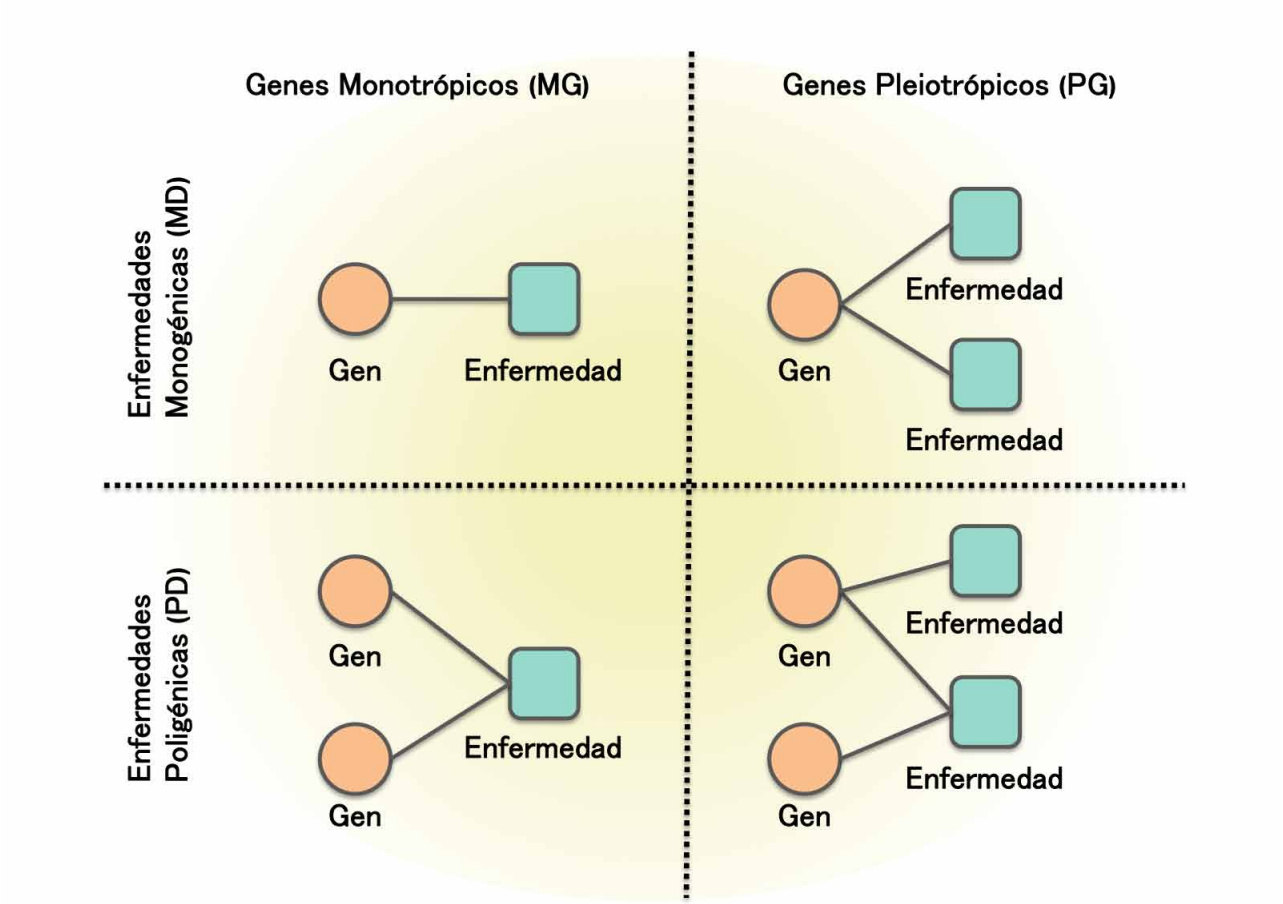


Figura 3.1: Las relaciones se dividieron en cuatro grupos en función del número de genes asociados con la enfermedad y viceversa.

relaciones entre elementos de un solo tipo, o unipartita. Puesto que se parte de una red bipartita (gen-enfermedad o *diseasome*) es posible obtener dos tipos de proyecciones distintas: la proyección de las enfermedades y la proyección de los genes, izquierda y derecha de la Figura 3.2 respectivamente, método usado por (Goh et al., 2007)

En este trabajo se han usado versiones actualizadas de ambas proyecciones: *Human Disease Network* (HDN) y *Human Disease Gene Network* (HDGN)(Goh et al., 2007). La nueva versión de la HDN que consta de 2.885 relaciones entre 1.843 enfermedades y la HDGN que consta de 2.654 relaciones y 749 genes. Siguiendo el mismo procedimiento se ha obtenido también una versión actualizada de la red de enfermedades raras de Orphanet (Zhang et al. 2011). A partir de la red de relaciones entre genes y enfermedades raras (Orphanet) hemos obtenido las redes unipartitas de enfermedades (*Orphan Disease Network* o ODN) y genes (*Orphan Disease Gene Getwork* o ODGN). Las redes resultantes constan de 3.568 relaciones entre 1.655 enfermedades y 4.456 relaciones entre 3.657 genes, respectivamente. En esta tesis se usaron dos versiones de las redes unipartitas gen-gen. Para el análisis y comparación con redes de similitud fenotípica entre genes (Objetivo 1) se usaron solo las relaciones gen-gen; sin embargo, en el desarrollo de la herramienta PhenUMA (Objetivo 2), se incorporó el número de enfermedades compartidas entre cada par de genes. Este número se corresponde con el peso de la relación.

3.3.4. Interactomas

Para el análisis funcional de las relaciones fenotípicas (Objetivo 1) entre genes se usaron las interacciones físicas entre proteínas de *CRG Human Interactome* (Bossi and Lehner, 2009) que incluye 10.299 genes y 80.922 interacciones usando aquellas interacciones validadas por al menos un experimento. Las relaciones metabólicas entre genes se obtuvieron de la red propuesta por Veeramani (Veeramani and Bader, 2009) que se basa en el análisis de balance de flujo de la red metabólica humana Recon 1 (Duarte et al., 2007). En la base de datos de PhenUMA (Objetivo 2) las relaciones entre proteínas proceden de STRING v9.05 (Szklarczyk et al., 2011), que integra distintas relaciones: físicas, metabólicas, co-expresión, literatura, etc. El fichero *protein.links.v9.05.txt.gz* (descargado en septiembre de 2012) contiene todas las relaciones entre proteínas integradas en STRING v9.05. Se seleccionaron aquellas relaciones que involucran a proteínas humanas y cuya puntación están contenidas en el 95º percentil. En total, se incluyeron 96.856 relaciones entre 10.316 genes. La red de relaciones metabólicas (Veeramani and Bader, 2009) también se incluyó en la base de datos de PhenUMA y consta de 9.812 relaciones entre 535 genes.

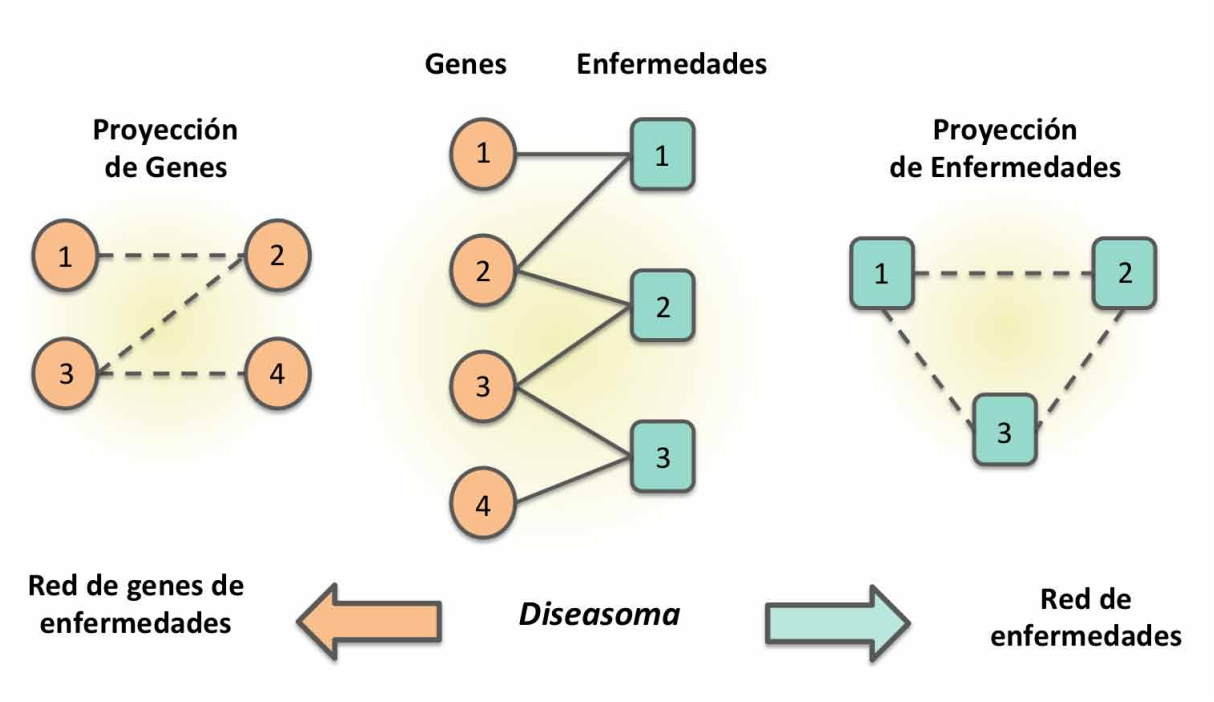


Figura 3.2: Proyecciones de la red gen-enfermedad. Las redes unipartitas (aquellas que consisten en relaciones entre genes o entre enfermedades) se obtienen a partir de la proyección de las relaciones gen-enfermedad del "diseasoma". La red unipartita de la izquierda consiste en las relaciones entre genes asociados con la misma enfermedad. A la derecha se incluye la red unipartita de enfermedades. En esta red dos enfermedades estarán relacionadas si están causadas por el mismo gen (o genes).

3.4. Relaciones de similitud fenotípica

Las relaciones de similitud fenotípica se obtienen a partir de la comparación de los perfiles fenotípicos asociados a cada par de objetos, es decir, al conjunto de términos HPO relacionados con cada objeto. Los objetos con los que trabajamos son genes, enfermedades o enfermedades raras, pacientes y regiones genómicas (DECIPHER)

3.4.1. Perfiles fenotípicos

En esta tesis, se denomina perfil fenotípico a un conjunto términos de HPO que representa la sintomatología de una enfermedad, gen, paciente, etc. A continuación se describe la procedencia de los perfiles fenotípicos de cada uno de los elementos:

3.4.1.1. Perfiles fenotípicos de enfermedades

En esta tesis, se han analizado relaciones fenotípicas entre enfermedades de OMIM y Orphanet. Para obtener estas relaciones es necesario comparar los perfiles fenotípicos de estas enfermedades. Con respecto a OMIM, los perfiles fenotípicos se obtienen del fichero de anotaciones *phenotype_annotation.omim*, descargado desde `http://human-phenotype-ontology.github.io/downloads.html` y creado por los autores de la ontología (HPO). En este fichero se encuentran las relaciones entre cada enfermedad OMIM y los términos HPO que definen su sintomatología, es decir, las anotaciones de las enfermedades a términos de la ontología. El fichero utilizado, descargado en 2011, incluye información para 4.965 enfermedades.

Con respecto a Orphanet, los perfiles fenotípicos se han construido a partir de la información fenotípica de OMIM. Para esta tarea se ha usado el fichero *Disorders, cross referenced with other nomenclatures* (descargado de Orphadata, `http://www.orphadata.org/`), en el que se incluye la correspondencia de los identificadores Orphanum (identificador usado en Orphanet para clasificar enfermedades) con otras bases de datos de enfermedades, entre ellas OMIM. El conjunto de anotaciones resultante contiene información para 3.143 enfermedades de Orphanet. Afortunadamente, este procedimiento no sería necesario actualmente, ya que Orphanet proporciona los perfiles fenotípicos de cada enfermedad. Sin embargo, fue necesario al comienzo de la realización del presente proyecto.

3.4.1.2. Perfiles fenotípicos de genes

El perfil fenotípico asociado a cada gen también se ha obtenido de la web de HPO, concretamente del fichero *gene_to_phenotype.txt*. Las relaciones gen-fenotipo son el resultado de la proyección de las relaciones gen-enfermedad-fenotipo, de manera, que el conjunto de fenotipos asociado a cada gen corresponderá con la unión de los fenotipos de las enfermedades con las que está relacionado dicho gen. El fichero utilizado incluye información para 1.806 genes.

3.4.2. Similitud semántica

Para comparar los perfiles fenotípicos (HPO) y funcionales (GO) se usaron dos medidas diferentes basadas en el cálculo del contenido informativo (IC). Para obtener la similitud fenotípica entre genes y enfermedades se usó la versión simétrica del método propuesto por Resnik (Resnik, 1995) y utilizado por (Köhler et al., 2009), definida en la Introducción (sección 1.4.2.3) y que denominamos Resnik_Simétrica (Ecuación 1.6). La similitud semántica funcional entre genes se calculó a partir de la medida propuesta por Resnik (Ecuación 1.4). Para la implementación de estas medidas se usó el código fuente de la herramienta Ontologizer (Bauer et al., 2008).

3.4.3. Selección del umbral de corte

El número de las relaciones de similitud semántica obtenidas entre genes, enfermedades y enfermedades raras es muy elevado. A estos elementos se les asignan descripciones fenotípicas que pueden contener síntomas muy poco específicos. A modo de recordatorio, decir que la especificidad de un síntoma o fenotipo está inversamente relacionada con su frecuencia en el conjunto de enfermedades o genes. Es decir, un fenotipo relacionado con un gran número de enfermedades será poco específico (y por consiguiente su IC será bajo). Los perfiles fenotípicos poco específicos podrían dar a lugar a relaciones de similitud débiles, con valor de similitud bajo, a lo que se podría denominar “ruido”.

Con el propósito de reducir el nivel de ruido, en el conjunto de relaciones fenotípicas, se planteó establecer un umbral estadístico óptimo, para así poder discriminar entre valores de similitud semántica significativos y no significativos. Para analizar el ruido existente en las comparaciones de los tres grupos de elementos considerados (genes, enfermedades OMIM y enfermedades raras) se siguieron dos estrategias. Por un lado, se comparó la evolución del número de elementos a medida que incrementamos el umbral de corte (Figura 3.3A). En segundo lugar, se usó el índice Jaccard, que representa el por-

centaje de solapamiento entre dos conjuntos. Sean dos conjuntos de elementos A y B, $Jaccard(A,B)$ $\frac{(A \quad B)}{(A \quad B)}$. Usando Jaccard se evaluó el solapamiento de las relaciones de similitud fenotípica y las relaciones inferidas entre genes y enfermedades (obtenidas como se indicó en la Figura 3.2) procedentes de los *diseasomas*. El resultado de esta evaluación se muestra en la Figura 3.3B.

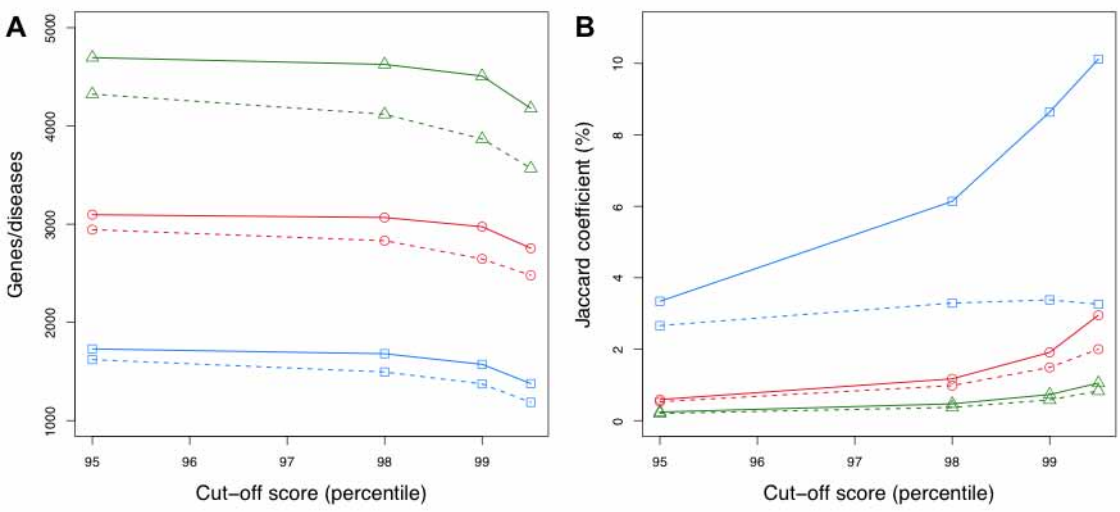


Figura 3.3: Número de nodos y de relaciones a medida que aumentan el umbral de corte de similitud semántica. (tomada de (Reyes-Palomares et al., 2013)). Los cuadros azules hacen referencia a los genes, los círculos rojos a enfermedades OMIM y los triángulos verdes a enfermedades Orphanet. Las líneas corresponden al número de relaciones que se obtiene para los distintos valores de similitud semántica: Resnik (en líneas continuas, Ecuación 1.4, Introducción apartado 1.4.2.3) y la versión simétrica de Resnik usando la media (en líneas punteadas, Ecuación 1.6, Introducción apartado 1.4.2.3). **A.** Número de nodos (genes o enfermedades) en las redes a medida que se aumenta el umbral de corte establecido para los valores de similitud semántica (de menor a mayor) en orden creciente de percentiles, a partir del 95º percentil. **B.** Índice Jaccard resultante de la comparación del conjunto de relaciones de similitud fenotípica y las incluidas en los *diseasomas*.

Las redes, calculadas usando tanto la medida de Resnik (líneas continuas) como la versión simétrica (líneas discontinuas), se han filtrado estableciendo umbrales situados en los percentiles 95º, 98º, 99º y 99.5º. Se observó un punto de inflexión en el percentil 98º: el número de elementos de las redes desciende bruscamente y el solapamiento entre las relaciones de similitud fenotípica y las inferidas (Figura 3.2) aumenta considerablemente a partir de ese valor, sobre todo para el caso de los genes (cuadrados azules). Teniendo en cuenta estos resultados, se consideró el valor de similitud correspondiente al 98º percentil como umbral de corte que minimiza la pérdida de información y maximizar la consisten-

cia de las relaciones. La Tabla 3.1 muestra el número de nodos y relaciones de cada una de las redes de similitud fenotípica.

Tabla 3.1: Número de nodos y relaciones de las redes de similitud semántica usando el corte seleccionado (98º percentil).

Tipo	Número de Nodos	Número de relaciones
OMIM-OMIM	4.627	14.9689
Orphanum-Orphanum	3.068	75.924
Gen-Gen	1.681	24.902

Con respecto a la similitud funcional, el número de relaciones es mucho mayor. En consecuencia, se establece un límite superior (percentil 99.5º) con el objetivo de seleccionar las relaciones más significativas y reducir el posible ruido producido por similitudes no informativas. La Tabla 3.2 muestra el número de genes y relaciones de las redes de similitud funcional para cada una de las tres subontologías de GO.

Tabla 3.2: Número de genes (nodos) y relaciones de similitud funcional calculadas a partir de cada una de las ramas de GO.

Subontologías	Número de Nodos	Número de Relaciones
Procesos Biológicos *Biological Process* (BP)	9.123	486.982
Funciones Moleculares *Molecular Function* (MF)	8.087	565.739
Componente Celular *Cellular Component* (CC)	6.046	565.739

Para los análisis correspondientes al Objetivo 1 se usaron solo las relaciones funcionales entre genes usando la rama de procesos biológicos. Para el resto de trabajos como PhenUMA y la exploración de relaciones funcionales de genes involucrados en el metabolismo de las aminas (Objetivos 2 y 3) se usaron las tres ramas de la ontología.

3.4.4. Validación de las relaciones de similitud fenotípica y cálculo de la curva ROC

Para validar las relaciones de similitud fenotípica se usó un clasificador binario usando como información de referencia la intersección de varias redes con información funcional: red de interacción entre proteínas (Bossi and Lehner, 2009), red de relaciones meta-

bólicas (Veeramani and Bader, 2009) - introducidas en el apartado 3.3.4 de este capítulo - y la red de relaciones funcionales entre genes a partir de GO. Para comprobar la validez de estas relaciones se construyó una curva ROC (*Receiver Operating Characteristic*) que es la representación gráfica de la sensibilidad o la representación del ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la especificidad o razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos).

Para la representación de la curva ROC se consideró como conjunto de verdaderos positivos (VP), o conjunto de éxitos, a los pares de genes con similitud fenotípica alta (por encima del 98º percentil) presente es la redes de referencia. Para determinar el conjunto de falsos positivos se obtuvo un conjunto de relaciones aleatorias a partir de los interactomas. Los falsos positivos (FP) serán aquellos pares de genes con similitud fenotípica (también por encima del 98º percentil) presentes en el conjunto de relaciones aleatorias.

3.4.5. Cálculo de los p-valores para las consultas de fenotipos

Una de las funcionalidades de la herramienta desarrollada en el Objetivo 2 es identificar el conjunto de genes (o enfermedades) con mayor similitud respecto a un conjunto de fenotipos. La herramienta proporciona una lista ordenada con el valor de similitud semántica resultante de la comparación del conjunto de fenotipos de entrada con los genes o enfermedades anotados en la ontología. Sin embargo, además del valor de similitud semántica se proporciona la significación de dicha consulta, que corresponde con la probabilidad de que un conjunto de fenotipos aleatorios muestre un valor de similitud mayor, con un gen o enfermedad, que los fenotipos de entrada.

El tamaño de las consultas puede ser variable, y la similitud semántica asociada a las relaciones entre un conjunto de fenotipos de entrada y un gen o enfermedad puede ser poco significativa, ya que el valor de similitud obtenido puede deberse al azar, por ello, junto el valor de similitud semántica es necesario incluir un p-valor que indique como de significativo es el resultado. Para calcular la significación de los valores de similitud semántica resultantes de la comparación de los fenotipos de entrada con los elementos anotados en la ontología, se usó un procedimiento basado en el método de Monte Carlo similar al usado por Phenomizer (Köhler et al., 2009), que consiste en la generación de muestras aleatorias de fenotipos. Puesto que PhenUMA no establece ningún límite en cuanto al tamaño de las consultas, resulta inviable calcular la significación de todas las consultas posibles, sin embargo, se obtuvieron los p-valores para muestras de tamaño desde 1 hasta 10 fenotipos (para muestras de mayor tamaño se usa los p-valores para consultas de tamaño 10).

El procedimiento para calcular los p-valores fue como se detalla a continuación:

- Para cada enfermedad anotada en la ontología y para cada tamaño de consulta (desde 1 hasta 10 fenotipos) se obtuvieron 100.000 muestras aleatorias compuestas por fenotipos de la ontología. Cada una de estas muestras (m_i) pueden definirse formalmente como:

$$m_i \quad \{p_1, p_2, ..., p_i, ...\} \tag{3.2}$$

  donde p_i son términos anotados en la HPO.

- Posteriormente se calculó la similitud fenotípica entre cada uno de los 100.000 perfiles de fenotipos aleatorios y cada enfermedad anotada en la ontología. Por lo tanto, para cada enfermedad se obtuvo un conjunto de pares formados por un valor de similitud y una muestra:

$$similitud_{enf} \quad \{(m_1, s_1), (m_2, s_2), ..., (m_j, s_j), ...\} \tag{3.3}$$

  donde m_i representa a cada una de las muestras aleatorias de fenotipos y s_i los valores de similitud semántica entre cada muestra y la enfermedad (enf).

- El siguiente paso fue calcular la frecuencia de cada una de estos valores de similitud:

$$frecuencia_{enf} \quad \{(f_1, s_1), (f_2, s_2), ..., (f_k, s_k), ...(f_n, s_n)\} \tag{3.4}$$

- A continuación, ser ordenó el conjunto $frecuencia_enf$ en función de los valores de similitud semántica (s_i), de menor a mayor, y se calculó la frecuencia acumulada para cada valor de similitud:

$$frec.acumulada_{enf} \quad \{(P_1, s_1), (P_2, s_2), ..., (P_i, s_i), ...(P_n, s_n)\} \tag{3.5}$$

  siendo F_i el número total de valores de similitudes semánticas mayores a s_i.

- Por último, a partir de las frecuencias acumuladas es posible obtener la probabilidad de que exista un valor de similitud mayor para cada s_i por azar:

$$p \quad valores_{enf} \quad \{(P_1, s_1), (P_2, s_2), ..., (P_i, s_i), ...(P_n, s_n)\} \tag{3.6}$$

donde P_i $\frac{\sum_{j=i}^{n} F_j}{100{,}000}$, o lo que es lo mismo el cociente entre el número de valores de similitud semántica mayores o iguales a s_i y el número total de muestras aleatorias. Es decir, P_i representa la probabilidad de obtener un valor de similitud semántica mayor a s_i entre un conjunto de fenotipos aleatorios y la enfermedad enf.

Este proceso se repitió para cada enfermedad y para distintos tamaños de muestra (desde 1 hasta 10 fenotipos) y se almacenó en la base de datos de PhenUMA.

3.5. Enriquecimiento fenotípico y funcional

Se denomina enriquecimiento fenotípico o funcional a la obtención de los términos, HPO o GO respectivamente, más representativos asociados a un conjunto de elementos. Estos pueden ser genes, enfermedades, pacientes o cualquier objeto anotado en la ontología. El enriquecimiento se obtiene calculando, para cada término t de la ontología, el solapamiento entre los elementos anotados a t y el conjunto de elementos del estudio (o muestra). Para obtener la significación estadística de dicho solapamiento se obtiene el p-valor a partir de un test hipergeométrico:

$$P \quad Valor(t) \quad \sum_{k \quad n_t}^{\min(m_t,n)} \frac{\binom{m_t}{n} \quad \binom{m \quad m_t}{n \quad k}}{\binom{m}{n}} \tag{3.7}$$

siendo:

- n el número de elementos de la muestra
- m el número total de elementos anotados en la ontología
- n_t el número de elementos de la muestra asociados con el término t
- m_t el número de elemento de la ontología asociados con el término t

El p-valor asociado a cada termino t indica la probabilidad de que exista un conjunto aleatorio de elementos, con tamaño igual al de la muestra (n) y seleccionado del conjunto total de elementos anotados a la ontología, cuyo número de anotaciones al término t es mayor o igual que el de la muestra (n_t). El p-valor se obtiene para cada término de la ontología. Por lo tanto, se trata de una comparación de hipótesis múltiple y es necesaria la corrección de los p-valores obtenidos para cada término, para lo cual, se ha usado el método de Bonferroni (veáse 3.1.3). Un término se considera enriquecido por un conjunto de elementos anotados a la ontología si su p-valor ajustado es menor a 0,05 (Bauer et al., 2008).

3.6. Integración de relaciones biomédicas a partir de minería de textos.

Las relaciones entre genes y enfermedades integradas en PhenUMA proceden de las relaciones gen-enfermedad incluidas en OMIM y Orphanet. Sin embargo, en la literatura se pueden entontrar relaciones entre genes y enfermedades que no están presentes en dichas bases de datos, y por lo tanto, tampoco en PhenUMA. Para extraer esta información se propone un flujo de trabajo (Figura 3.4) cuyo objetivo es extraer estas relaciones mediante el uso herramientas de minería de textos. En este trabajo se ha aplicado esta metodología a genes implicados en el metabolismo de las aminas biogénicas. Este grupo de genes se caracteriza por tener una baja presencia en bases de datos y sin embargo, su relación con varias patologías está muy presente en la literatura.

Para extraer relaciones gen-enfermedad no incluidas en OMIM y Orphanet se usaron las herramientas DISEASES (Pletscher-Frankild et al., 2014) y DisGeNET (Bravo et al., 2014):

- DisGeNET contiene 429.111 relaciones entre 17.181 genes y 14.619 enfermedades. La información procede de varias bases de datos como *The Comparative Toxicogenomics Database* (Davis et al., 2013) y UniProtKB (`http://www.uniprot.org/`). Además se incluye información procedente de la minería de textos de artículos de *Genome-Wide Association Studies* (GWAS) (Becker et al., 2004), y de *Literature-driven Human Gene-Disease Network* (Liu et al., 2013) usando BeFree (Bravo et al., 2014) que es una herramienta para extraer información de textos biomédicos. BeFree se compone de un módulo para la extracción de entidades en la literatura (*Biomedical Named Entity Recognition*, BioNER) y un módulo para identificar relaciones entre dichas entidades (ER) basado en *Support Vector Machine* (SVM), para obtener las asociaciones correctas entre entidades detectadas en la literatura.

- Por otro lado, DISEASES extrae relaciones gen-enfermedad a partir del análisis de los resúmenes de *Genetics Home Reference* (GHR) (`https://ghr.nlm.nih.gov`) UniProtKB (`http://www.uniprot.org/`), GWAS y DistiLD (Palleja et al., 2012) e información de mutaciones procedente de *Catalog of Somatic Mutations in Cancer Catalog of Somatic Mutations in Cancer* (`http://cancer.sanger.ac.uk/cosmic`).

A partir del conjunto de relaciones gen-enfermedad obtenidas a partir de minería de textos, se construyeron los perfiles fenotípicos (conjunto de términos HPO) de los genes relacionados con el metabolismo de las aminas (Figura 3.4B). De esta forma ya es posible

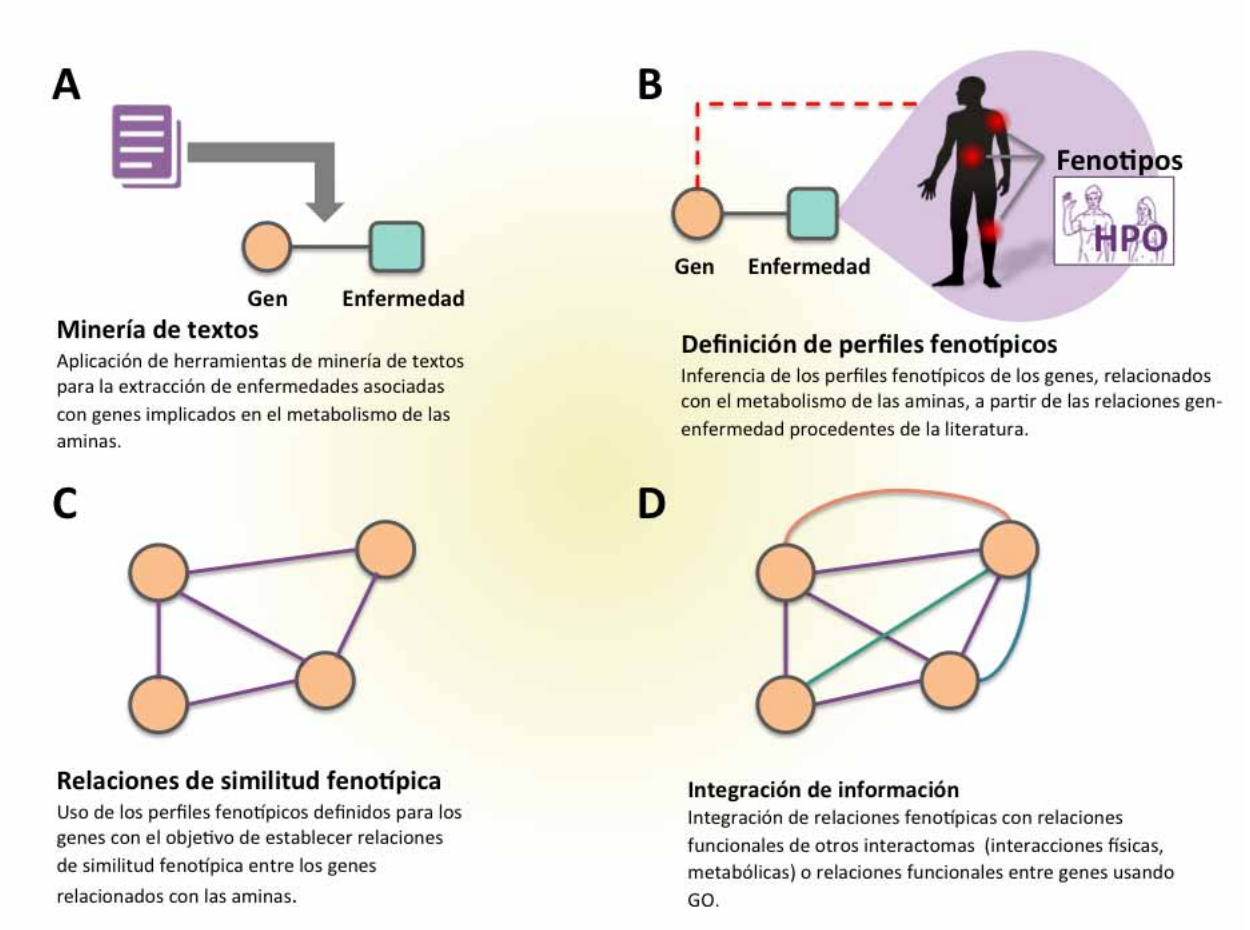


Figura 3.4: Esquema del procedimiento para extraer las relaciones fenotípica entre genes relacionados con aminas biogénicas. A. Uso de herramientas de minería de textos para extraer las relaciones gen-enfermedad. **B.** Definición de perfiles fenotípicos de estos genes a partir de las relaciones gen-enfermedad extraídas y las relaciones entre enfermedades OMIM y fenotipos (apartado 3.4.1.1) **C.** Uso de medidas de similitud semántica para establecer relaciones de similitud fenotípica entre los genes asociados con el metabolismo de las aminas y otros genes anotados en la ontología. **D.** Integración de las relaciones de similitud fenotípica con información funcional para estudiar el contexto funcional de la información fenotípica.

establecer relaciones de similitud fenotípica entre estos genes y otros genes anotados en la ontología. Por último, al igual que en PhenUMA, se integró la información fenotípica con las relaciones funcionales procedentes de GO y STRING (Figura 3.4D).

3.7. Caracterización de los *loci* fenotípicamente enriquecidos (PEL) a partir de los datos de DECIPHER

En esta tesis se han analizado el solapamiento entre CNV de varios pacientes (DECIPHER) y su enriquecimiento fenotípico (Objetivo 4). Para el análisis de las relaciones genotipo-fenotipo en el conjunto de pacientes se define PEL (*Phenotypically Enriched Loci*) o *loci* fenotípicamente enriquecido. La Figura 3.5 muestra un resumen del proceso para la obtención de los *loci* fenotípicamente enriquecidos.

3.7.1. Construcción de la red de pacientes

Usando los datos de la base de datos de DECIPHER (Firth et al., 2009), descrita en el punto 3.2.3 de este capítulo, se construyó una red de pacientes a partir de las regiones afectadas por alguna CNV como indica la Figura 3.5A usando como genoma de referencia GRCh37/hg19. En esta red los nodos son pacientes y existirá una relación entre dos pacientes si existe algún solapamiento entre las regiones genómicas alteradas en ambos. Los solapamientos se calcularon para deleciones y duplicaciones separadamente, siendo un par de bases el solapamiento mínimo considerado entre dos pacientes.

3.7.2. Cálculo de los *cliques* de pacientes

A partir de esta red se obtuvieron la lista de *cliques* (Figura 3.5B). Un *clique* es un conjunto de nodos en el que existe una relación entre cada par de nodos (es decir, un subgrafo completo). El cálculo de los *cliques* se realizó usando en paquete de Python NetworkX (`https://networkx.github.io/`). Siendo k el número de nodos de un *clique*, el tamaño de los *cliques* va desde k=3 sin límite en el tamaño máximo. El algoritmo puede generar *cliques* redundantes compuestos por los mismos pacientes. Para eliminar las redundancias, se eliminaron los *cliques* formados por los mismos pacientes con el objetivo de obtener una lista de *cliques* únicos.

La metodología usada para relacionar los pacientes podría identificar *cliques* en los que no todos los pacientes solapan en la misma región. Esto es debido a que algunos pacientes pueden estar afectados por más de una CNV. Si además, estas variaciones son cer-

canas en el genoma podrían solapar con pacientes, asociados con CNV de gran tamaño. Para evitar esto y para realizar los siguientes análisis se filtraron aquellos *cliques* en los que existe al menos un par de bases solapante entre todas las mutaciones de los pacientes (incluidos en dicho *clique*). El conjunto de *cliques* resultante se usó para realizar el análisis fenotípico, descrito en el siguiente apartado.

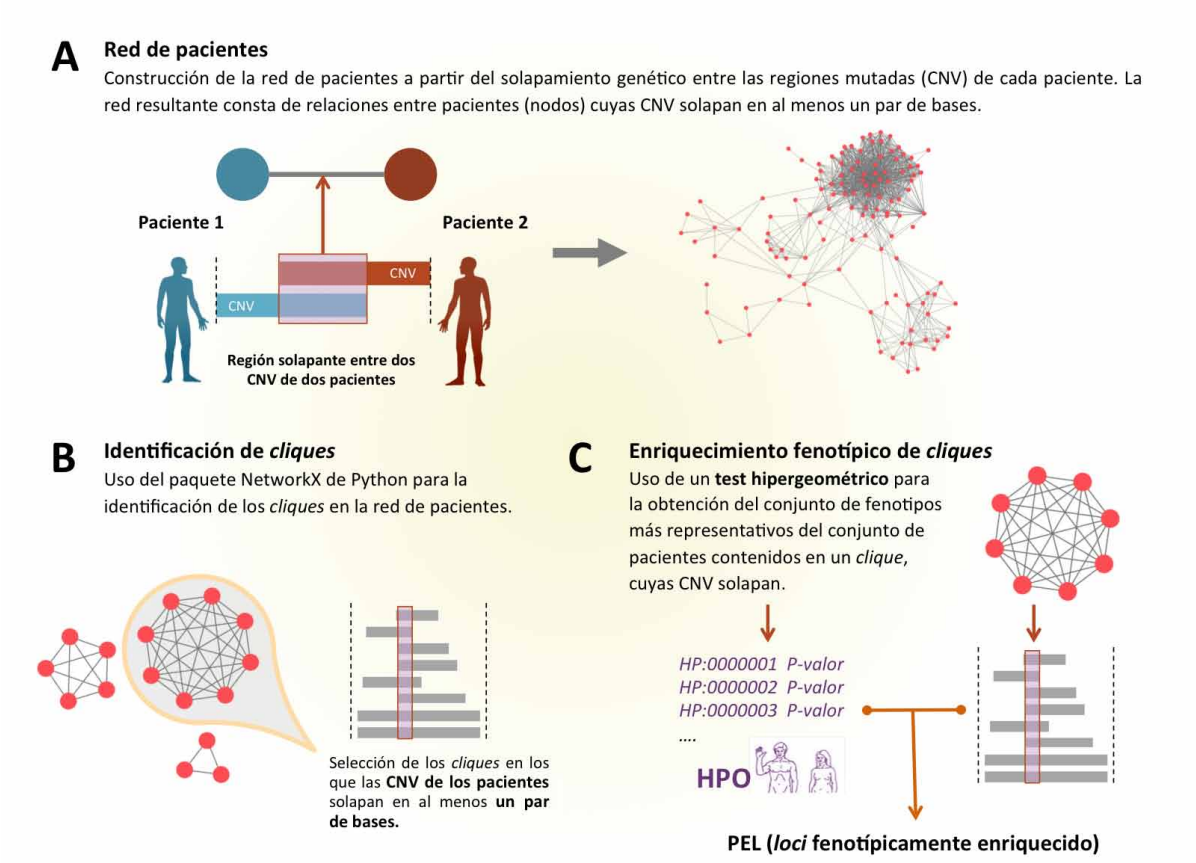


Figura 3.5: Esquema del proceso para la determinación de los PEL a partir de la información fenotípica y genética proporcionada por DECIPHER . A. El primer paso en la construcción de una red de pacientes. Para establecer una relación entre dos pacientes se obtiene el solapamiento entre las CNV observadas en dichos pacientes. **B.** A continuación, se identifican los *cliques* (grafos completos) seleccionado aquellos que representan un solapamiento de al menos un par de bases. **C.** Por último, se calcula el enriquecimiento fenotípico de estos *cliques*, con el objetivo de obtener los términos HPO más representativo del conjunto de pacientes incluidos en uno de ellos y asociados al *loci* solapante, determinando así los PEL.

3.7.3. Enriquecimiento fenotípico de *cliques*

En este trabajo definiremos "*loci* fenotípicamente enriquecido" (Figura 3.5C) como la región solapante entre un conjunto de pacientes enriquecida fenotípicamente, es decir, existe un conjunto de fenotipos para todos o la mayoría de los pacientes con un p-valor significativo. Para cada uno de los *cliques* seleccionados en el apartado anterior se calculó el enriquecimiento fenotípico a partir de un test hipergeométrico, usando un procedimiento similar al expuesto en el apartado 3.5. El resultado fueron varios subconjuntos de relaciones *clique*-fenotipo que fueron filtrados siguiendo los siguientes criterios:

- que el p-valor del test hipergeométrico sea menor que 0,05
- que existan al menos tres pacientes asociados con un mismo fenotipo
- y que al menos el 50 % de los pacientes estén asociados al fenotipo.

Tras realizar este filtro se identificaron *cliques* enriquecidos con términos HPO redundantes, relacionados jerárquicamente (relaciones padre-hijo). Para evitar sesgos ocasionados en los posteriores análisis se eliminaron los fenotipos redundantes, dando prioridad a los más significativos, es decir, aquellos con menor p-valor.

3.7.4. Caracterización de los PEL

Para validar el método se utilizaron las CNV presentes en la población sana (sección 3.2.4). El objetivo era evaluar la frecuencia de los PEL, identificados mediante el método detallado en el punto anterior, en la población control. Para ello, se usó un test de Fisher, de manera que, a partir de cada PEL se construyó una tabla de contingencia con la siguiente información: a) número de pacientes del PEL asociado con un fenotipo enriquecido, comparado con el conjunto total de pacientes asociados con ese fenotipo, y b) el número de individuos sanos (muestras de DGV) con regiones mutadas que solapan (al menos 1 par de bases) con el PEL comparado con el resto de individuos sanos. Tras aplicar el test de Fisher se ajustaron los p-valores usando Benjamini-Hochberg (veáse 3.1.3) y se estableció el corte en p-valores <0.05. El objetivo de este análisis es calcular la significación estadística de las relaciones HPO-*loci* comparándolo con la frecuencia de la CNV en la población de individuos sanos.

3.7.5. Aleatorizaciones para calcular la significación de los PEL potencialmente patológicos

Para determinar la significación de los PEL que permita identificarlos como potencialmente causantes de enfermedad se diseñaron cinco modelos aleatorios:

- **CNV aleatorias (casos) procedentes de DGV:** la información incluida en DGV se usó para estudiar la presencia, en el conjunto de controles, de las regiones de PEL potencialmente patológicas observadas en pacientes. Para validar si la identificación de PEL patológicos se debe al azar, se repitieron los análisis usando un modelo aleatorio basado en la aleatorización de las CNV de los controles.

- **Localización aleatoria de CNV de pacientes:** generación de regiones aleatorias manteniendo la distribución del tamaño de las CNV de los pacientes y su frecuencia en los cromosomas. El objetivo es comprobar si la frecuencia de los PEL detectados usando datos reales, es similar para las CNV aleatoriamente distribuidas en el genoma.

- **Localización aleatoria de CNV de controles:** en este análisis se siguió un procedimiento similar al anterior pero aplicado a las regiones observadas en los controles. Se obtuvo un conjunto de CNV aleatoriamente distribuidas en el genoma, manteniendo el tamaño de las mutaciones observadas en DGV. Con este modelo aleatorio se quiere comprobar si la frecuencia de los PEL obtenidos en el análisis es menor cuando se aleatorizan las CNV del conjunto de controles.

- **Aleatorización de las relaciones paciente-CNV:** este modelo consiste en usar un conjunto aleatorio de relaciones paciente-CNV, para comprobar si usando este conjunto aleatorio se obtiene un resultado similar que usando los datos reales.

- **Aleatorización de las relaciones paciente-fenotipo manteniendo la frecuencia de los fenotipos:** El objetivo de este modelo es identificar si el número de PEL detectados es menor usando perfiles fenotípicos aleatorios de los pacientes. Para cada modelo se realizaron 1.000 aleatorizaciones y se calculó el número de PEL resultantes para cada caso y su significación a partir del enriquecimiento fenotípico.

Capítulo 4

Resultados y discusión

4.1. Análisis de la red de similitud fenotípica entre genes

4.1.1. Comparación de redes de genes asociados con enfermedades genéticas procedentes de OMIM y Orphanet

Las enfermedades complejas en ocasiones son consecuencia de mutaciones de varios genes que afectan a varios procesos biológicos. Además, el entorno y los procesos estocásticos juegan un papel importante en el desarrollo de estas enfermedades. Tradicionalmente, la definición y clasificación de las enfermedades se ha basado en el cuadro sintomatológico de las observaciones en un órgano o sistema concreto. Esta filosofía de la medicina, basada en la evidencia, ha minimizado la importancia del estudio de los procesos bioquímicos involucrados, la información genética o el entorno. Puesto que todos estos elementos participan en la etiología de enfermedades genéticas, es posible que incluso una misma enfermedad se manifieste de forma diferente (con síntomas diferentes) en varios individuos (Loscalzo and Barabasi, 2011). Esta problemática se refleja en la ambigüedad y falta de consenso a la hora de clasificar e identificar algunas de estas patologías. Este hecho se pone de manifiesto, por ejemplo, en las diferencias entre bases de datos actuales, como OMIM y Orphanet. Estas bases de datos proporcionan las relaciones gen-enfermedad que fueron usadas para la construcción de los *diseasomas*, analizados en trabajos anteriores (Zhang et al., 2011; Goh et al., 2007). En este trabajo, se analizaron versiones actualizadas de estos *diseasomas* y se compararon entre sí, con el objetivo de estudiar las diferencias en la definiciones de enfermedades en ambas bases de datos.

La red resultante generada a partir de los datos de OMIM relaciona 2.525 genes con 3.132 enfermedades y, usando los datos de Orphanet, la red resultante contiene 2.331 ge-

nes relacionados con 2.125 enfermedades. Para analizar estas relaciones se han agrupado tal como se indica en el capítulo de Material y Métodos (Figura 3.1), es decir, en función del número de genes implicados en cada enfermedad (enfermedades monogénicas o enfermedades poligénicas), así como, el número de enfermedades en los que están involucrados cada gen (genes monotrópicos o genes pleiotrópicos). La Tabla 4.1 muestra cómo se distribuyen las relaciones gen-enfermedad contenidas en estas redes según la clasificación propuesta. Esta clasificación permite observar algunas diferencias entre ambas redes. Entre ellas destaca el número de relaciones biunívocas (MD-MG), siendo las incluidas en OMIM el doble de las que se pueden encontrar en Orphanet. Se observa, además, que en Orphanet el conjunto de PD-MG es predominante al resto, es decir, enfermedades poligénicas asociadas a genes monotrópicos (Tabla 4.1). En ambos *diseasomas*, los genes monotrópicos representan un porcentaje alto del conjunto total de genes causantes de enfermedades. En concreto, el 72 % de los genes (1.810 genes) de OMIM y el 69 % de los genes (1.625 genes) de Orphanet se relacionan con una sola enfermedad.

Tabla 4.1: Distribución de las relaciones gen-enfermedad.

	Red gen-enfermedad (OMIM)		**Red gen-enfermedad (Orphanet)**	
	Ratio	**Genes(%)**	**Ratio**	**Genes(%)**
MD-MG	1,00	1.431 (56,7)	1,00	717 (30,8)
MD-PG	2,57	639 (25,3)	2,71	435 (18,7)
PD-MG	0,46	379 (15,0)	0,40	908 (39,0)
PD-PG	2,13	371 (14,7)	1,68	584 (25,1)

Las relaciones gen-enfermedad se usaron para generar las redes entre genes siguiendo la metodología usada en los *diseasomas* (véase Figura 3.2 de Material y Métodos). Las redes resultantes se muestran gráficamente en las Figuras 4.1A y 4.1B.

Aunque los dos repositorios de enfermedades contienen una información similar - el 69 % de las enfermedades que está en Orphanet también está en OMIM y con respecto a los genes el solapamiento es del 81 % - las diferencias entre las redes gen-gen son mayores de las esperadas. Las Figuras 4.1A y 4.1B permiten observar visualmente el contraste entre la cantidad de nodos aislados: 1.776 para la HDGN (*Human Disease-Gene Network*) y 839 para la ODGN (*Orphan Disease-Gene Network*). Un mayor número de nodos no conectados implica un número menor de relaciones entre los genes de HDGN, con respecto a los de ODGN. Para entender mejor este fenómeno se puede volver a la clasificación de relaciones gen-enfermedad anterior, concretamente al conjunto de relaciones biunívocas, correspondiente al conjunto MD-MG. Los genes que se incluyen en este conjunto son

precisamente los que quedan excluidos de las redes de genes (HDGN y ODGN). Como se indicó al inicio, en la Tabla 4.1 se puede comprobar que el conjunto de relaciones biunívocas acumula un mayor número de relaciones en ambos casos (56.7 % en OMIM y 30.8 % en Orphanet), siendo en OMIM más de la mitad de todas las relaciones.

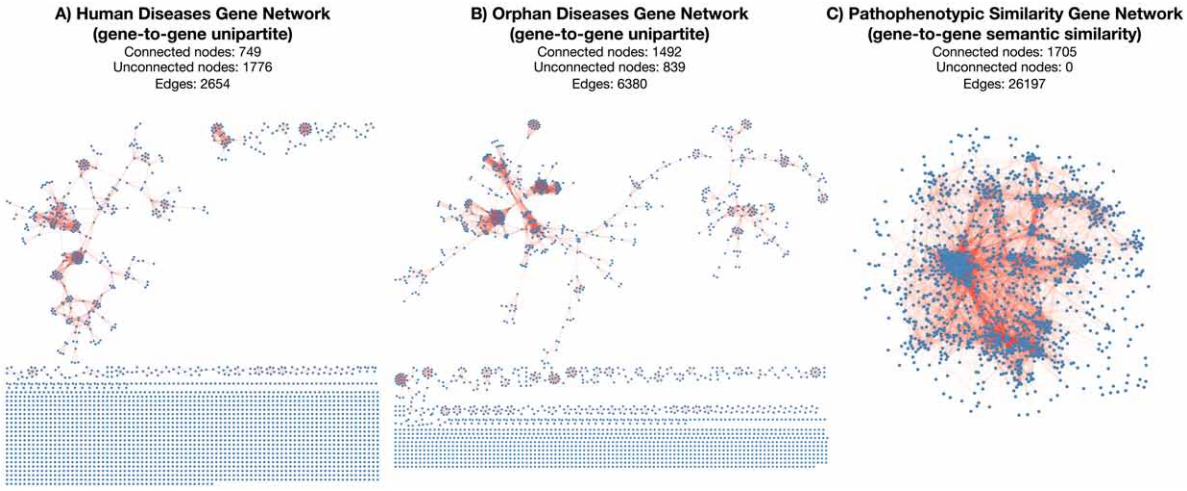


Figura 4.1: Redes entre genes a partir de enfermedades y fenotipos (tomada de Reyes-Palomares et al. 2013). A. *Human Disease Gene Network* (HDGN). **B.** *Orphan Disease Gene Network* (ODGN). **C.** *Pathophenotypic Similarity Gene Network* (PSGN)

Teniendo en cuenta la información contenida en ambas bases de datos, se podía esperar una mayor similitud entre ambas redes. Sin embargo, las diferencias encontradas podrían no ser una consecuencia de la información que representan, sino de cómo está estructurada. La forma en la que son clasificadas las enfermedades puede ser una de las causas de estas diferencias. De hecho, el identificador Orphanum, usado por Orphanet para identificar enfermedades suele agrupar varios términos OMIM. En ocasiones, cada elemento de Orphanet corresponde a una enfermedad (ej. Diabetes), que incluye a varios subtipos de una misma patología (ej. Diabetes Tipo I, Diabetes Tipo II, ...), mientras que OMIM usa un identificador diferente para cada subtipo de esa misma enfermedad. Si se comprueban las relaciones OMIM-Orphanum (obtenidas del fichero *Disorders, cross referenced with other nomenclatures* descargado desde Orphadata, ver Material y Métodos 3.2.2) se obtiene que el número de códigos OMIM por cada Orphanum puede ser muy elevado. Un ejemplo llamativo es el de la Retinitis pigmentosa (ORPHA:791) que se relaciona con 80 identificadores OMIM: Retinitis pigmentosa 1 (OMIM 180100), Retinitis pigmentosa 2 (OMIM 312600), Retinitis pigmentosa 51 (OMIM 613464), etc. El número de genes asociados a esta enfermedad en Orphanet es 71, sin embargo, en OMIM la mayoría de

estas enfermedades son monogénicas y se relacionan con genes monotrópicos, es decir, estas relaciones se incluirán en el conjunto MD-MG. Los genes incluidos este grupo aparecerán aislados en HDGN, construida a partir de los datos de OMIM, pero no si usamos los datos de Orphanet (ODGN).

Además de las diferencias en el número de nodos aislados, la intersección de las relaciones de ambas redes es inesperadamente pequeño. Para realizar esta comparativa se ha usado el índice Jaccard, que indica el porcentaje del total de relaciones que intersectan (que son solapantes) entre dos conjuntos, excluyendo los nodos aislados. En concreto, el índice Jaccard entre ODGN y HDGN es solo el 7.9% de todas las relaciones incluidas en la unión de ambas redes. También es notable la diferencia de tamaño entre ambas redes, siendo HDGN bastante más pequeña que la ODGN, la cual duplica en número de relaciones a la primera. Una de las principales causas de estas diferencias son nuevamente los criterios utilizados en cada base de datos para asignar identificadores a enfermedades. Un identificador de Orphanet se relaciona de media con más genes que un identificador de OMIM (ver Tabla 4.1). Por ejemplo, el grupo de relaciones PD-MG es considerablemente más abundante para la red derivada de Orphanet que la de OMIM, 908 pares frente a 379, respectivamente. Esta diferencia en el número de genes del conjunto PD-MG es la que produce una diferencia sustancial en el número de relaciones inferidas.

Estas diferencias observadas entre ambas redes refuerzan la necesidad de un consenso para determinar las relaciones gen-enfermedad, la definición de enfermedad y su clasificación. Las discrepancias entre ambas bases de datos hacen que sea complicado unificarlas e incluso decidir qué información es la más adecuada. Las relaciones gen-enfermedad en OMIM son más específicas, el porcentaje de relaciones biunívocas es muy grande (56,7%), lo que implica que muchas enfermedades, que son subtipos de otra más general, se consideren como una enfermedad independiente con identificador propio. Este detalle puede ser positivo para el estudio de cada patología de forma independiente, pero es un inconveniente para el análisis del *diseasoma* y sus proyecciones, ya que aumenta considerablemente el número de nodos aislados. Sin embargo, esto no ocurre en Orphanet, donde la información se estructura de forma que permite establecer relaciones entre genes que van a estar conectados y afectan a procesos biológicos comunes, a pesar de que es más difícil determinar los genes asociados a una subclase de enfermedad en particular.

En resumen, cada base de datos tiene ventajas e inconvenientes, pero llegados a este punto surgen varias preguntas, ¿qué tienen en común dos enfermedades para que se consideren subtipos de otra más general? ¿cómo podemos solucionar las ambigüedades derivadas del uso de dos fuentes de datos aparentemente similares? En esta tesis se considera que para solventar estos inconvenientes puede ser útil el uso de la sintomatología

de las enfermedades para compararlas y clasificarlas. Todas las enfermedades agrupadas en una sola patología general tienen algo en común y es el solapamiento en síntomas o fenotipos patológicos, pero algunas de estas diferencias son muy sutiles. Estos resultados refuerzan la hipótesis de la necesidad de utilizar un procedimiento sistemático de caracterización fenotípica de enfermedades genéticas.

4.1.2. Nuevas relaciones entre genes a partir del uso de información (pato)fenotípica

Las enfermedades genéticas manifiestan una gran complejidad y variabilidad fenotípica (Loscalzo and Barabasi, 2011). Por ello, en este trabajo se consideró que el uso de los perfiles fenotípicos (definido como el conjunto de síntomas de una enfermedad) podría permitir tener en cuenta dicha variabilidad para establecer nuevas relaciones entre genes. Además, la definición de los perfiles fenotípicos usa la sintomatología observada en los pacientes, evitando así la ambigüedad derivada del uso de identificadores establecidos por distintos repositorios.

La definición de los perfiles fenotípicos se realizó usando mediante la *Human Phenotype Ontology* (HPO), que proporciona un vocabulario normalizado y permite establecer similitudes patofenotípicas entre objetos anotados a la misma. Para realizar estas comparaciones, se construyó el perfil fenotípico de cada gen como se describe en Material y Métodos (apartado 3.4.1) utilizando las relaciones gen-enfermedad. La construcción de la red de similitud fenotípica se llevó a cabo a partir de la comparación de los perfiles fenotípicos de cada par de genes, utilizando la medida de similitud semántica (Resnik Simétrica, véase Introducción 1.4.2.3) propuesta por Köhler et al. (2009). Como resultado se obtienen más de 3.000.000 de relaciones entre genes, sin embargo, solo consideramos como significativas aquellas situadas en el percentil 98º, o lo que es lo mismo, el 2 % de las comparaciones con mayor valor de similitud fenotípica (Figura 3.3 en Material y Métodos). La red de similitud fenotípica entre genes (PSGN) consta de 26.197 relaciones entre 1.705 genes. Antes de comentar con mayor profundidad las relaciones de PSGN, y comparar esta red con las anteriores (HDGN y ODGN), en la Figura 4.1 se observa a simple vista que posee una estructura diferente, sin nodos aislados y con una mayor densidad en el número de relaciones.

Tanto las redes de relaciones inferidas entre genes, usando las relaciones incluidas en HDGN y ODGN, como la red de similitud fenotípica entre genes (PSGN) constan de relaciones establecidas a partir de información patológica. En HDGN y ODGN se usan las enfermedades compartidas por cada par de genes para establecer las relaciones y en PSGN

se consideran los fenotipos asociados a esas mismas enfermedades. Por esta razón, era esperado que un porcentaje alto de las relaciones de HDGN y ODGN estén incluidas también en PSGN. Se obtuvo el conjunto de genes y relaciones comunes (intersección) entre PSGN y cada una de las redes HGDN y ODGN.

El resultado de esta comparación muestra que PSGN contiene el 39 % (HDGN) y el 26 % (ODGN) de las relaciones entre genes incluidas en las proyecciones unipartitas. El número de genes comunes es bajo, si tenemos en cuenta que los perfiles fenotípicos y la HDGN tiene originariamente la misma procedencia (relaciones gen-enfermedad en OMIM). Sin embargo, hay que recordar que solo se están comparando las relaciones de similitud fenotípica más significativas (el 2 %), y la mayoría de las relaciones de similitud (mayores a 0) no superan ese umbral.

El uso de la ontología y las medidas de similitud influyen en el valor asignado a cada par de genes. En la introducción de esta tesis se explicó cómo afecta el uso de la estructura jerárquica para comparar términos (apartado 1.4.2.3). Si comparamos dos fenotipos muy inespecíficos (muy frecuentemente asociados con genes) el valor de similitud semántica entre ellos va a ser muy bajo, independientemente de su cercanía dentro de la ontología. Por lo tanto, para que dos genes tengan un valor de similitud alto, no basta con que los dos genes estén asociados a los mismos fenotipos, sino que además estos perfiles deben tener un alto valor informativo, o lo que es lo mismo, deben de estar compuestos por fenotipos específicos (alejados de la raíz de la ontología).

La exactitud en la definición de los perfiles fenotípicos de cada gen, así como la especificidad de estos fenotipos, depende en gran medida de las descripciones de las enfermedades asociadas a dicho gen en las bases de datos (OMIM). Esto es debido a que las asociaciones gen-fenotipo (o término HPO) se establecen a partir de dichas descripciones (véase Material y Métodos 3.4.1). El uso de los fenotipos, su organización en la estructura jerárquica y las medidas basadas en el contenido informativo van a priorizar relaciones entre perfiles fenotípicos más específicos. Esto influye en las diferencias observadas entre la red PSGN y las dos proyecciones, puesto que en HDGN y ODGN solo se tiene en cuenta si entre dos genes hay una enfermedad común o no. En estas redes (HDGN y ODGN) no se tiene en cuenta la especificidad de las patologías comunes a dos genes. En PSGN, se establecen las relaciones usando HPO, por tanto sí se considera este factor.

El uso de esta metodología podría resultar poco intuitiva, dado que es posible que la comparación de dos perfiles muy similares dé como resultado un valor pequeño de similitud, si se componen de fenotipos poco específicos. Sin embargo, este método permite discriminar los genes fenotípicamente similares y asociados a fenotipos muy específicos, de los pares de genes asociados a fenotipos muy generales. De hecho, el conjunto de rela-

ciones incluidas en PSGN (26.197 en total) consta de pares de genes asociados a fenotipos poco frecuentes entre los genes anotados en la ontología. Estas relaciones podrían implicar a genes involucrados en procesos biomoleculares comunes.

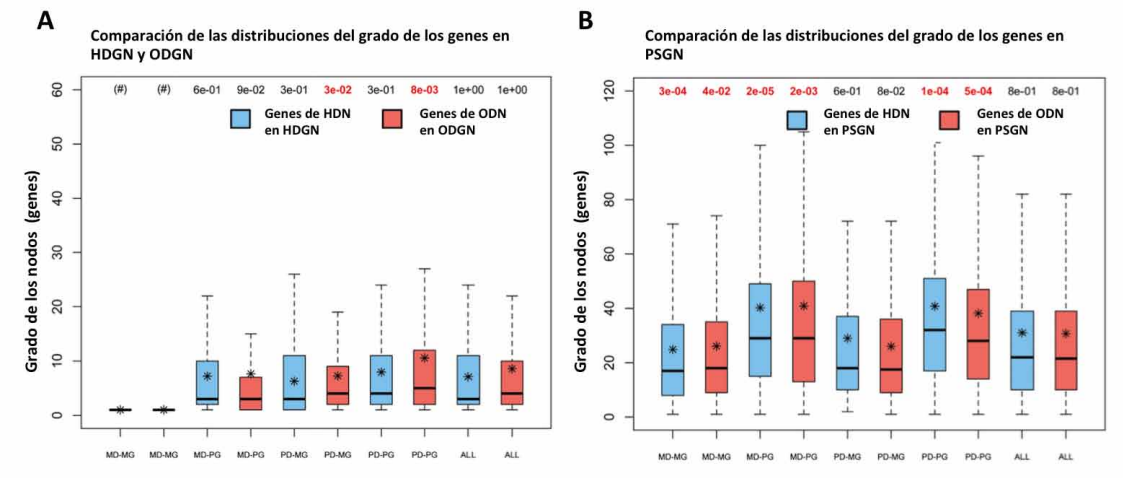


Figura 4.2: Distribución del número de relaciones entre los distintos subgrupos de relaciones genes-enfermedad. Esta figura muestra la comparación de la distribución del grado en cada conjunto. **A.** Relaciones obtenidas a partir de la proyección de las relaciones gen-enfermedad. **B.** Relaciones de similitud fenotípica entre genes. En la parte superior de ambas gráficas se incluye el p-valor, que muestra la significación de comparar cada conjunto con la red completa. Los p-valores significativos aparecen en color rojo. La media en el grado de cada nodo se indica en cada caja con un asterisco.

4.1.3. Análisis de las relaciones fenotípicas entre genes

En términos cuantitativos, la principal diferencia entre PSGN y las proyecciones HDGN y ODGN está en el número de relaciones, siendo mucho mayor para PSGN, a pesar de que el número de nodos en PSGN no dista tanto del de las otras redes. En consecuencia, es previsible que el número de relaciones o grado (*degree*) de cada nodo (gen) en PSGN sea también mayor. Para evaluar estas diferencias, se ha comparado la distribución del grado de cada una de las redes (HDGN, ODGN y PSGN) y para cada conjunto de genes (MD-MG, MD-PG, PD-MG y PD-PG). El test Mann-Whitney U se usó evaluar si las diferencias entre la distribución del grado para cada conjunto y para las redes completas eran significativas. La Figura 4.2 muestra las gráficas de las distribuciones del número de conexiones para los genes de cada conjunto, teniendo en cuenta los genes incluidos tanto en HDGN como en PSGN. El contraste más evidente se observa en el número de relaciones biunívocas (MD-

MG), que se corresponde con genes aislados o no conectados. En la Figura 4.2B, donde se muestra la distribución del grado para los genes en HDGN y ODGN, vemos que los genes incluidos en este grupo aparecen relacionados con una media de 25 genes, a diferencia del resultado obtenido para la red PSGN (Figura 4.2A).

En general, en la comparación de la distribución del grado en ambas redes, se puede ver que el uso de los perfiles fenotípicos para establecer relaciones entre genes aumenta el número de relaciones en todos los grupos. No obstante, en la gráfica (Figura 4.2B) se observa que los genes pleiotrópicos acumulan un mayor número de relaciones. Esto puede deberse a que el número de relaciones de cada gen en PSGN puede verse afectado por el número de fenotipos que tiene asociado. Este hecho es una consecuencia del método usado para asignar el perfil fenotípico de los genes, a partir de las relaciones gen-enfermedad (véase Material y Métodos 3.4.1). Esto implica que para los genes pleiotrópicos (MD-PG, PD-PG) se espere un conjunto de fenotipos mayor que para los monotrópicos (MD-MG, PD-MG).

Puesto que el conjunto de genes pleiotrópicos va a estar asociado a un mayor número de enfermedades, se comprobó si este conjunto de genes estaba también relacionado con perfiles fenotípicos más numerosos. Para ello se compararon las distribuciones del número de fenotipos por gen para cada subconjunto de genes (Figura 4.3). Los genes pleiotrópicos (líneas azul y naranja) presentan un mayor número de fenotipos con respecto a la red completa (línea negra), mientras que los monotrópicos (líneas verde y roja) presentan un menor número de fenotipos con respecto a la red completa. Este resultado evidencia el enriquecimiento de fenotipos en aquellos genes asociados a varias enfermedades, indicando a su vez que puedan participar en un mayor número de relaciones (como muestra la Figura 4.2). Sin embargo, un gen con un perfil fenotípico que abarque un amplio número de síntomas, puede generar relaciones poco específicas y valores de similitud fenotípica bajos.

4.1.4. Integración de información fenotípica y funcional

En los análisis anteriores se han evaluado las diferencias en el número de relaciones emergentes gracias al uso de los perfiles fenotípicos. Aunque el número de relaciones fenotípicas inferidas es mucho mayor que en los diseasomas, muchas de estas relaciones entre genes no necesariamente estarán asociadas a los mismos procesos biomoleculares que causan la patología o el fenotipo en particular. No obstante, puesto que las perturbaciones en las interacciones a nivel celular pueden relacionarse con la progresión de enfermedades, el estudio de todas estas relaciones fenotípicas en su conjunto e integradas en

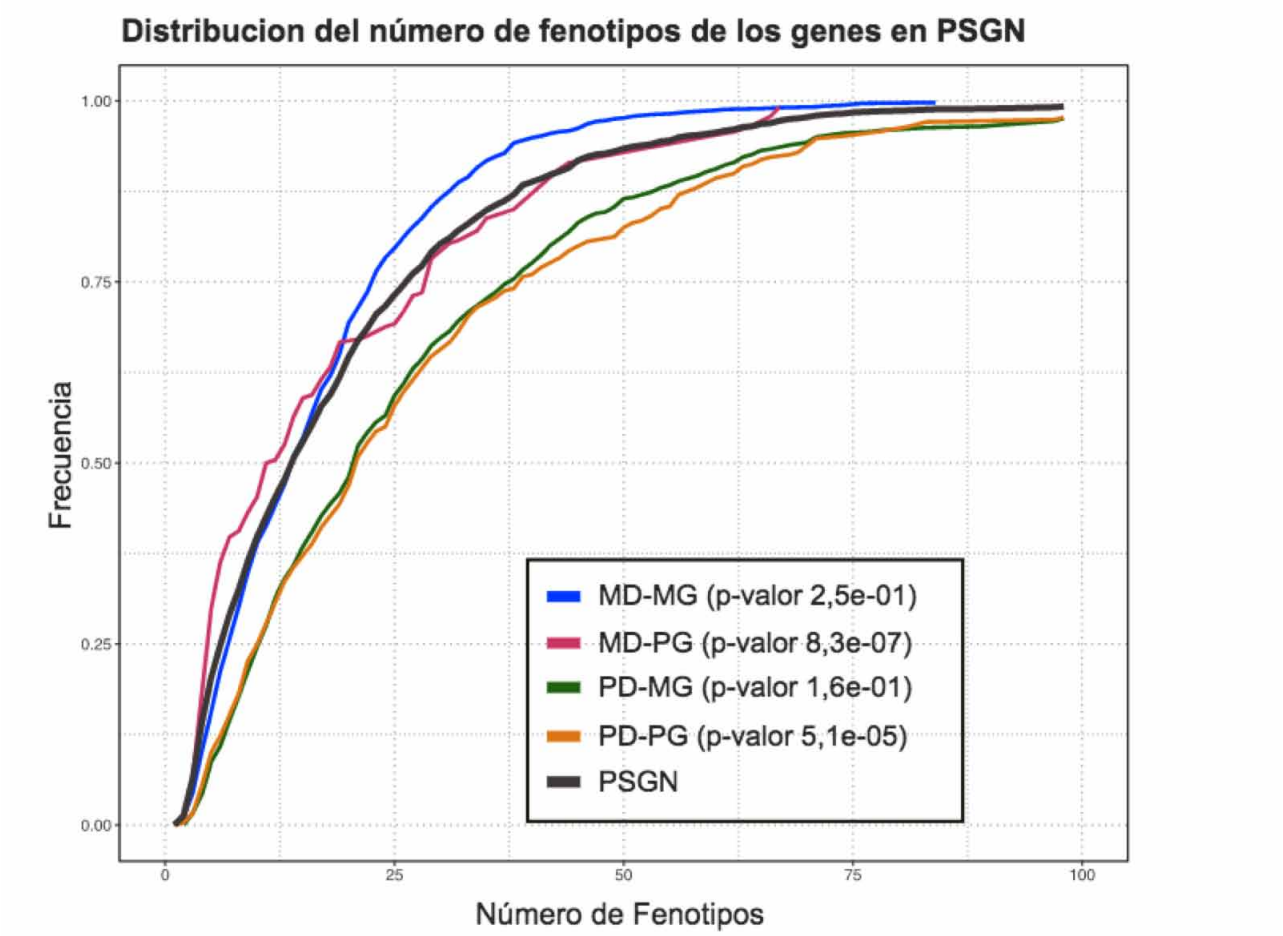


Figura 4.3: Distribución del número de fenotipos en los distintos subgrupos de relaciones genes-enfermedad. En esta gráfica se muestra la frecuencia acumulada respecto al número de fenotipos asociados a los genes. Cada una de las distribuciones se comparó comparó con la de la red global (PSGN).

su contexto funcional pueden permitir identificar relaciones potencialmente asociadas a enfermedades.

Con el propósito de estudiar la información funcional subyacente de las relaciones fenotípicas, se usaron varios interactomas:

- Red de interacciones entre de proteínas (PIN, interactoma humano), con 74.657 interacciones entre 9.580 genes.
- Relaciones de similitud funcional (FSGN) a partir de GO con 496.973 relaciones y 9.157 genes.
- Relaciones metabólicas (MGN) que implican 9.812 relaciones entre 535 enzimas.

Se calculó la intersección con PSGN de cada uno de los interactomas (Tabla 4.2). Dado que el objetivo es explorar las interacciones funcionales de las relaciones emergentes gen-gen a partir de las relaciones fenotípicas, se seleccionaron solo las relaciones de los interactomas que involucraban a genes incluidos en PSGN. La Tabla 4.2 muestra el número de nodos y relaciones resultantes de calcular la intersección de la red de relaciones fenotípicas y los interactomas. El índice Jaccard – que mide el porcentaje de solapamiento entre dos conjuntos – indica el bajo solapamiento entre los interactomas y PSGN. A pesar de ello, podemos destacar a las relaciones metabólicas con casi un 10 % de solapamiento.

Tabla 4.2: Intersección entre PSGN y relaciones funcionales.

		Nodos	**Relaciones**	**Jaccard(relaciones)**
$PSGN$	PIN	896	422	2,5
$PSGN$	MGN	127	124	9,8
$PSGN$	$FSGN$	1.370	2.473	5,4

Si en lugar del solapamiento calculamos el porcentaje de relaciones en PSGN con información funcional obtenemos que el 23.7 % de las relaciones en PSGN se corresponden con interacciones físicas, un 11,8 % con genes asociados metabólicamente y el 8.1 % son relaciones funcionales obtenidas a partir GO (*Biological Process*). Al integrar relaciones funcionales y fenotípicas, el conjunto de relaciones es reducido, por lo tanto suponen un buen punto de partida para el estudio de los mecanismos moleculares relacionados con enfermedades genéticas. Las mejoras futuras en la definición de los perfiles fenotípicos de genes, enfermedades y variaciones genéticas, así como los avances en las estructuras de las ontologías y en el diseño de medidas de similitud, pueden permitir refinar las relaciones de similitud fenotípica. Por otro lado, la integración de información relativa al entorno

podría enriquecer estas relaciones y permitir estudiar con mayor exactitud las causas de los procesos patológicos, ya que en muchas de las enfermedades el factor ambiental juega un papel importante.

A pesar del bajo solapamiento, entre las relaciones de similitud fenotípica y las relaciones funcionales, se analizó si las relaciones funcionales entre genes se corresponden a pares de genes cuyos rangos de similitud fenotípica son altos. Para ello, se construyó una curva ROC (*Receiver Operating Characteristic*, o Característica Operativa del Receptor) usando como referencia la información funcional (PIN, MSG y FSGN). El conjunto de verdaderos positivos (*true positives* o TP) está formado por las relaciones incluidas en las intersecciones de PSGN y cada interactoma. Los falsos positivos son las relaciones incluidas en la intersección de PSGN con aleatorizaciones de cada interactoma, manteniendo la distribución de la conectividad. Las curvas ROC resultantes muestran un área bajo la curva (*area under curve* o AUC) de 0,77 para interacciones físicas (PIN), 0,76 para relaciones metabólicas (MGN) y 0,66 para relaciones funcionales (FSGN). El área bajo la curva de una señal aleatoria es 0,5, por esta razón, tanto las relaciones físicas como las metabólicas, asociadas más directamente a reacciones biomoleculares, muestran una fuerte señal alejada del azar. Este resultado refuerza la idea de que es posible estudiar las enfermedades a partir de la integración de información (pato)fenotípica y funcional, y por otro lado, hay que tener en cuenta el ruido determinado por la baja especificidad de muchas relaciones en este tipo de análisis.

4.1.5. Genes candidatos a partir de la integración de información funcional y fenotípica: Síndrome de la orina con olor a jarabe de arce

A modo ilustrativo y como ejemplo se analizaron los genes asociados con el síndrome de la orina con olor a sirope de arce (MSUD, MIM #248600). Este ejemplo muestra las relaciones fenotípicas entre genes asociados con MSUD y se profundiza en el contexto funcional de las mismas. El objetivo es mostrar que a partir de la información fenotípica es posible revelar nuevas asociaciones de los genes previamente asociados a MSUD con otros genes que pudieran estar también involucrados en el desarrollo de dicha enfermedad.

MSUD es una enfermedad genética agrupada en las aminoacidurias y causada por una disminución de la actividad del complejo deshidrogenasa de alfa-cetoácidos de cadena ramificada (*branched-chain alpha-ketoacid dehydrogenase* BCKDH), el cual cataliza el primer paso para la degradación aminoácidos ramificados (valina, leucina y isoleucina). El complejo enzimático tiene tres subunidades (E1, E2 y E3) codificadas por cuatro ge-

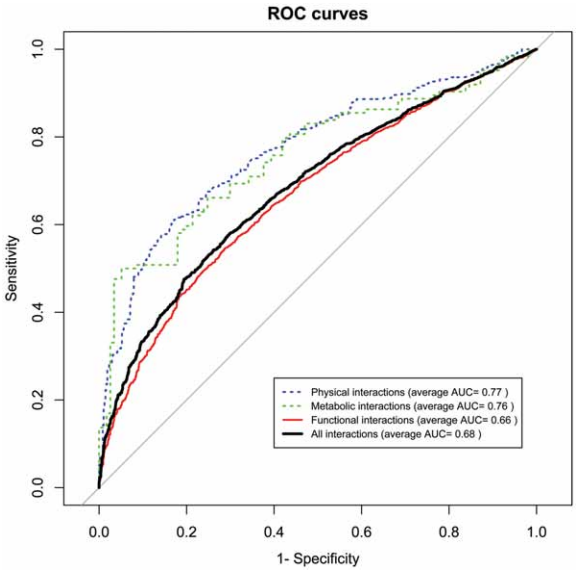


Figura 4.4: Curva ROC de la similitud fenotípica entre genes (tomada de Reyes-Palomares et al. 2013). Esta gráfica muestra las curvas ROC resultantes de evaluar la señal de las relaciones fenotípicas para los tres interactomas (relaciones físicas (azul), metabólicas (verde) y funcionales (rojo)). El área bajo la curva muestra una fuerte señal en los tres casos, sobre todo para relaciones físicas y metabólicas.

nes BCKDHA-E1A (EntrezID 539), BCKDHA-E1B (EntrezID 594), DBT-E2 (EntrezID 1629) y DBT-E3 (EntrezID 1738). Esta enfermedad está bien caracterizada fenotípicamente; los síntomas frecuentes son: cerumen con olor a jarabe de arce, niveles altos de aminoácidos ramificados, cetonuria, encefalopatía, coma y fallo respiratorio. Se seleccionaron todas las relaciones entre los genes relacionados con MSUD y otros genes en PSGN. Algunos de los genes relacionados en esta selección también presentan relaciones metabólicas de correlación de flujo con BCKDHA, BCKDHB, DBT y DLD (Veeramani and Bader, 2009) y la mayoría de ellos participan en diferentes reacciones incluidas en la ruta de la degradación de la valina, leucina e isoleucina.

Se obtuvieron las relaciones fenotípicas entre estos genes y el resto de genes incluidos en PSGN, cuyo resultado se muestra en la Figura 4.5A. En dicha figura se destacan, adicionalmente, las relaciones metabólicas en las que están involucrados estos genes. Además, como muestra la Figura 4.5C es posible identificar los fenotipos asociados a cada relación entre genes. Por ejemplo, PCCA y PCCB están asociados con alteraciones bioquímicas similares y correlacionados con MSUD, como por ejemplo altos niveles de ácido láctico y cuerpos cetónicos. Por otro lado, otros fenotipos comunes se manifiestan en un nivel patofisiológico: edema cerebral, pancreatitis o coma. A partir de este esquema se puede des-

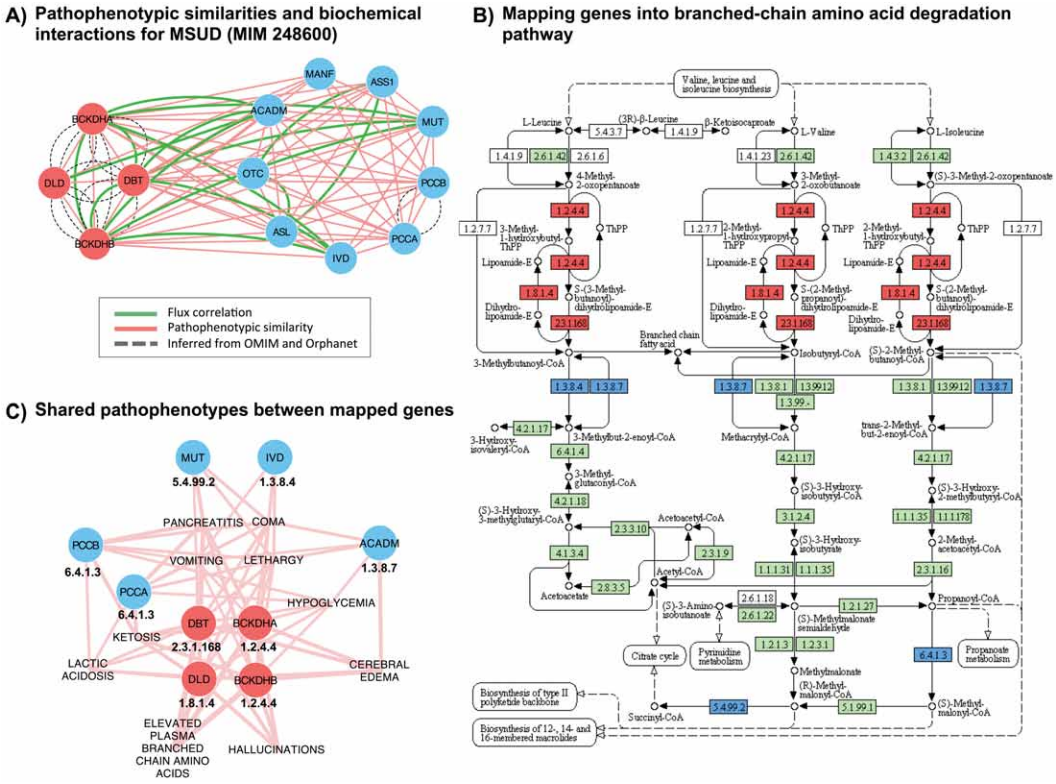


Figura 4.5: Relaciones fenotípicas y metabólicas de los genes implicados en síndrome de la orina con olor a jarabe de arce (MSUD) (MSUD) (tomada de Reyes-Palomares et al. 2013). A. Integración de relaciones fenotípicas y metabólicas entre genes asociados con MSUD (en rojo) y otros genes incluidos en PSGN y MGN. **B.** Esquema de la ruta metabólica encargada de la degradación de aminoácidos ramificados. Este esquema se obtuvo de *Kyoto Encyclopedia of Genes and Genomes* (KEGG). Las enzimas codificadas por genes humanos aparecen en verde, en rojo las enzimas codificadas por genes relacionados con MSUD, y en azul los genes fenotípicamente similares a estos genes. **C.** Fenotipos asociados a cada par de genes involucrados en la ruta metabólica.

tacar el potencial de PSGN y la integración de esta información en el contexto funcional. A partir de estos análisis es posible identificar nuevos genes candidatos e indagar en el contexto funcional de muchas enfermedades genéticas.

Es decir, este ejemplo muestra que es posible extraer nuevos genes candidatos a estar afectados en pacientes que padecen estos síntomas, como PCCA o PCCB. Por ello, la accesibilidad a esta información, para profesionales del ámbito clínico y de investigación biomédica, podría ser de gran utilidad para identificar genes y relaciones candidatas implicadas en los procesos patológicos que son objeto de estudio.

4.2. PhenUMA

4.2.1. Descripción de la herramienta

Los resultados mostrados en los puntos anteriores destacan las ventajas del uso de las relaciones fenotípicas y su integración con información funcional. El caso expuesto en la Figura 4.5 pone de manifiesto además que estudiar las relaciones fenotípicas en un contexto funcional nos permite generar nuevas hipótesis. Con este objetivo se presenta PhenUMARodríguez-López et al. (2014), una herramienta (accesible desde la url: `www.phenuma.uma.es`) que facilita el análisis simultáneo de información fenotípica y funcional, así como descubrir relaciones entre genes o enfermedades que no pueden deducirse a partir de bases de datos como OMIM y Orphanet. PhenUMA también integra relaciones de interactomas como STRING (Szklarczyk et al., 2011) y correlaciones de flujos metabólicos (Veeramani and Bader, 2009). PhenUMA permite la consulta de genes, enfermedades o fenotipos para obtener subredes que integren información fenotípica y funcional. Para ello, la herramienta proporciona una representación visual e interactiva de los resultados en forma de redes. De esta forma, se facilita la exploración de la red, la consulta de información asociada a cada nodo (gen o enfermedad) o relación y, además, la descarga de los resultados. Adicionalmente, PhenUMA permite consultar el enriquecimiento funcional y fenotípico de grupos de genes o enfermedades, es decir, es posible obtener una lista ordenada de los términos GO o HPO más representativos de elementos seleccionados. El método implementado para calcular los enriquecimientos se basa en un test hipergeométrico (descrito en Material y Métodos, apartado 3.5).

4.2.1.1. Base de datos de PhenUMA

PhenUMA es una herramienta web, desarrollada en Java. El visualizador está basado en el *plugin* de Cytoscape para web: CytoscapeWeb (Lopes et al., 2011). Todas las relaciones entre genes y enfermedades se integran en una base de datos implementada en MySQL. La Figura 4.6 muestra un resumen de las relaciones incluidas en la base de datos, que integra tanto relaciones unipartitas entre genes y enfermedades, como bipartitas (gen-enfermedad). La información incluida puede agruparse en tres grupos: relaciones funcionales, relaciones inferidas (proyecciones unipartitas de los disasomas) y relaciones de similitud fenotípica. Como se ha comentado en el párrafo anterior, la información funcional procede de varios interactomas (STRING y relaciones metabólicas) y de relaciones de similitud semántica a partir de GO. La base de datos incluye las relaciones gen-enfermedad procedentes de OMIM y sus proyecciones unipartitas: relaciones entre genes que comparten una o varias enfermedades, y relaciones entre enfermedades asociadas con el mismo gen (o genes). Por último, se incorporaron en la base de datos las relaciones de similitud fenotípica entre genes, enfermedades y enfermedades raras usando HPO y las medidas de similitud semántica descritas en la introducción de esta tesis (apartado 1.4.2.3). La Tabla 4.3 incluye el número de relaciones y nodos para cada tipo de relación de la base de datos.

4.2.1.2. Construcción de redes en PhenUMA

PhenUMA está diseñada para proporcionar el acceso a la información asociada a un conjunto de genes, enfermedades o fenotipos de interés. Para ejecutar una consulta en PhenUMA basta con introducir los datos de entrada. El procedimiento que sigue la herramienta internamente se incluye en la Figura 4.7, que muestra el proceso de construcción de redes para cada tipo de entrada.

Cuando se ejecuta una consulta se seleccionan el tipo de entrada y el tipo de relaciones de interés que se quiere consultar (relaciones fenotípicas, funcionales, físicas, etc.). Como puede verse en la figura los datos de entrada se usan para obtener una red inicial que consta de todas las relaciones entre esos datos y otros genes o enfermedades incluidos en la base de datos. El siguiente paso es enriquecer esta información incluyendo todas las relaciones entre cada par de genes de la red resultante. Con el objetivo de proporcionar flexibilidad para seleccionar la especificidad de las relaciones resultantes, PhenUMA permite al usuario seleccionar entre tres niveles de confianza (*low*, *medium* y *high*), correspondiente a los percentiles 98º, 99º y 99.5º para la similitud fenotípica en HPO y 99.5º, 99.8º y 99.9º para la similitud funcional en GO. Estos umbrales se aplican para las rela-

Tabla 4.3: Número de nodos y relaciones para cada tipo de interacción en la base de datos de PhenUMA.

Tipo de Red	Tipo de relación	Número de nodos	Número de relaciones
Relaciones fenotípicas			
OMIM-OMIM	Inferidas a partir de Genes (OMIM)	1.843	2.885
OMIM-OMIM	Similitud fenotípica (HPO)	4.627	149.689
Orpha-Orpha	Inferidas a partir de Genes (Orphanet)	1.655	3.568
Orpha-Orpha	Similitud fenotípica (HPO)	3.068	75.924
Gen-Gen	Inferidas a partir de OMIM	784	3.217
Gen-Gen	Inferidas a partir de Orphanet	1.641	8.292
Gen-Gen	Similitud fenotípica (HPO)	1.681	24.902
Relaciones funcionales			
Gen-Gen	Similitud funcional. Procesos biológicos (GO)	9.123	486.982
Gen-Gen	Similitud funcional. Componentes celulares (GO)	6.046	565.739
Gen-Gen	Similitud funcional. Funciones moleculares(GO)	8.087	397.683
Gen-Gen	Interacciones entre proteínas (STRING)	10.316	96.856
Gen-Gen	Relaciones metabólicas (Veeramani and Bader 2009)	535	9.812

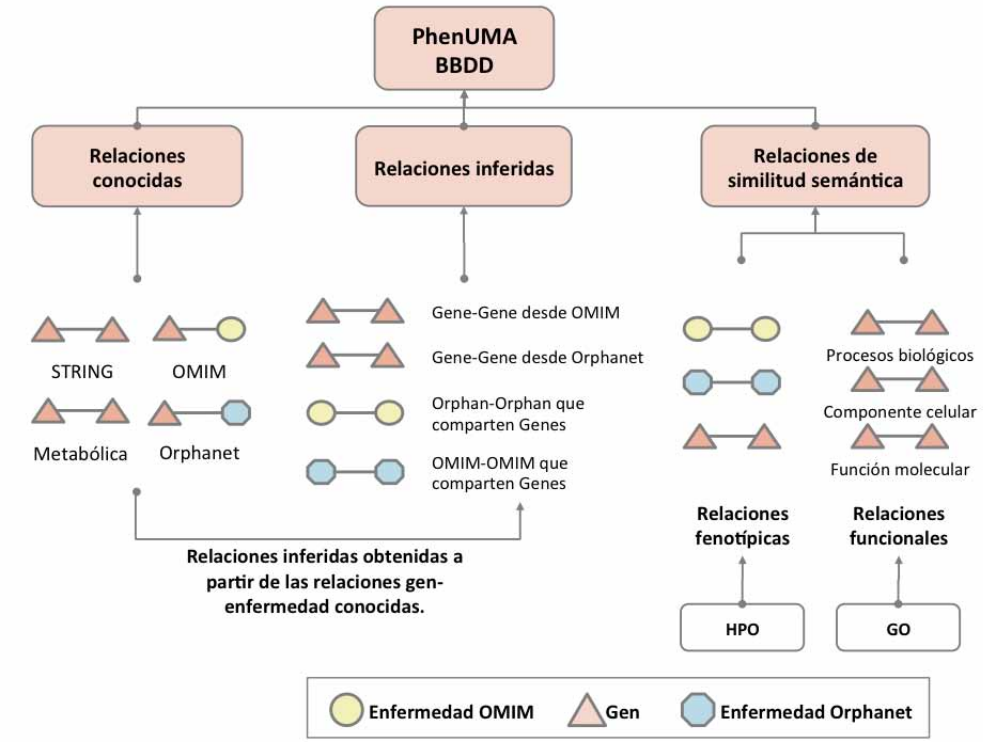


Figura 4.6: Base de datos de PhenUMA. Esta figura muestra un esquema de la base de datos de PhenUMA, que consiste en relaciones conocidas procedentes de algunos interactomas, relaciones inferidas a partir de las relaciones genenfermedad y relaciones de similitud semántica a partir de GO y HPO.

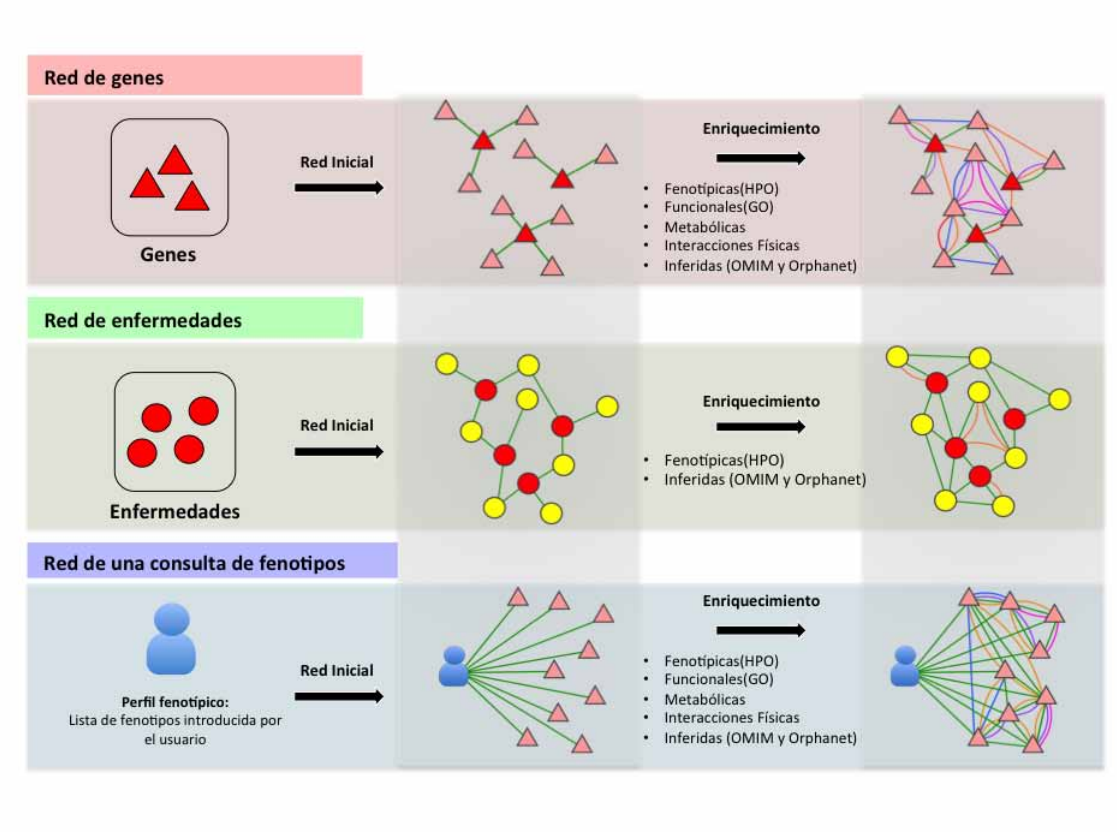


Figura 4.7: Proceso de construcción de las redes en PhenUMA. Esquema del proceso de construcción de las redes de genes o enfermedades a partir de varios tipos de entradas: genes, enfermedades o fenotipos. El mecanismo es similar en los tres casos. En primer lugar, se obtienen las relaciones del conjunto de elementos de entrada y, posteriormente, se incorpora la información funcional incluida en la base de datos.

ciones de similitud semántica, tanto funcionales (GO) como fenotípicas (HPO). Para las consultas de relaciones procedentes de STRING o Veeramani and Bader (2009) no es posible aplicar diferentes niveles de confianza. Esto se debe a la densidad variable presente en las redes de similitud semántica. El número de relaciones de un gen o enfermedad va a depender de la especificidad de su perfil fenotípico o funcional. En ocasiones, la red resultante puede contener un número demasiado extenso de relaciones, que harán difícil su estudio. Por esta razón, se permite elegir entre diferentes umbrales y, de esta forma, aumentar o reducir la cantidad de información contenida en la red.

Aunque el procedimiento desarrollado para construir las redes es similar para cada tipo de entrada, en consultas de fenotipos el proceso difiere ligeramente. En ese caso, el conjunto de fenotipos de entrada se considera como un perfil fenotípico y se compara con todos los genes o enfermedades anotadas en la ontología. La red resultante consta de un nodo, que representa el conjunto de fenotipos de entrada, y de las relaciones entre dicho perfil fenotípico y genes o enfermedades fenotípicamente similares. La red final incorpora las relaciones resultantes consideradas estadísticamente significativas, con p-valor ajustado menor a 0,05 (el cálculo del p-valor para las consultas de fenotipos se explicó en la sección de Material y Métodos, apartado 3.4.5).

4.2.2. Phenuma permite explorar las relaciones fenotípicas entre enfermedades

Succinic semialdehyde dehydrogenase deficency (SSADHD. OMIM 171980), en castellano conocida como síndrome de deficiencia de succínico semialdehido deshidrogenasa, es una enfermedad metabólica rara relacionada con mutaciones en el gen ALDH5A1 (ALDH5A1, OMIM 610045). Usando PhenUMA se identificaron las enfermedades fenotípicamente similares a SSADHD aplicando los distintos niveles de confianza: *low*, *medium* y *high* (Figura 4.8). Este resultado muestra cómo se forman diferentes grupos de enfermedades a medida que se aumenta el umbral de corte. Por ejemplo, aplicando el nivel de confianza más bajo (98º percentil) se obtiene una red con nodos interconectados que agrupa enfermedades asociadas con la epilepsia, convulsiones, procesos neurodegenerativos, anomalías neurofisiológicas y de comportamiento. La deficiencia de SSADH se relaciona con muchas enfermedades asociadas a convulsiones, epilepsia y cambios del comportamiento. Estas relaciones son aún más evidentes si subimos el nivel de confianza hasta observarse grupos más definidos, Figura 4.8B.

Aunque aumente el umbral de corte y el número de relaciones sea menor, las relaciones entre las enfermedades de cada grupo permanecen al aplicar el nivel de corte mayor;

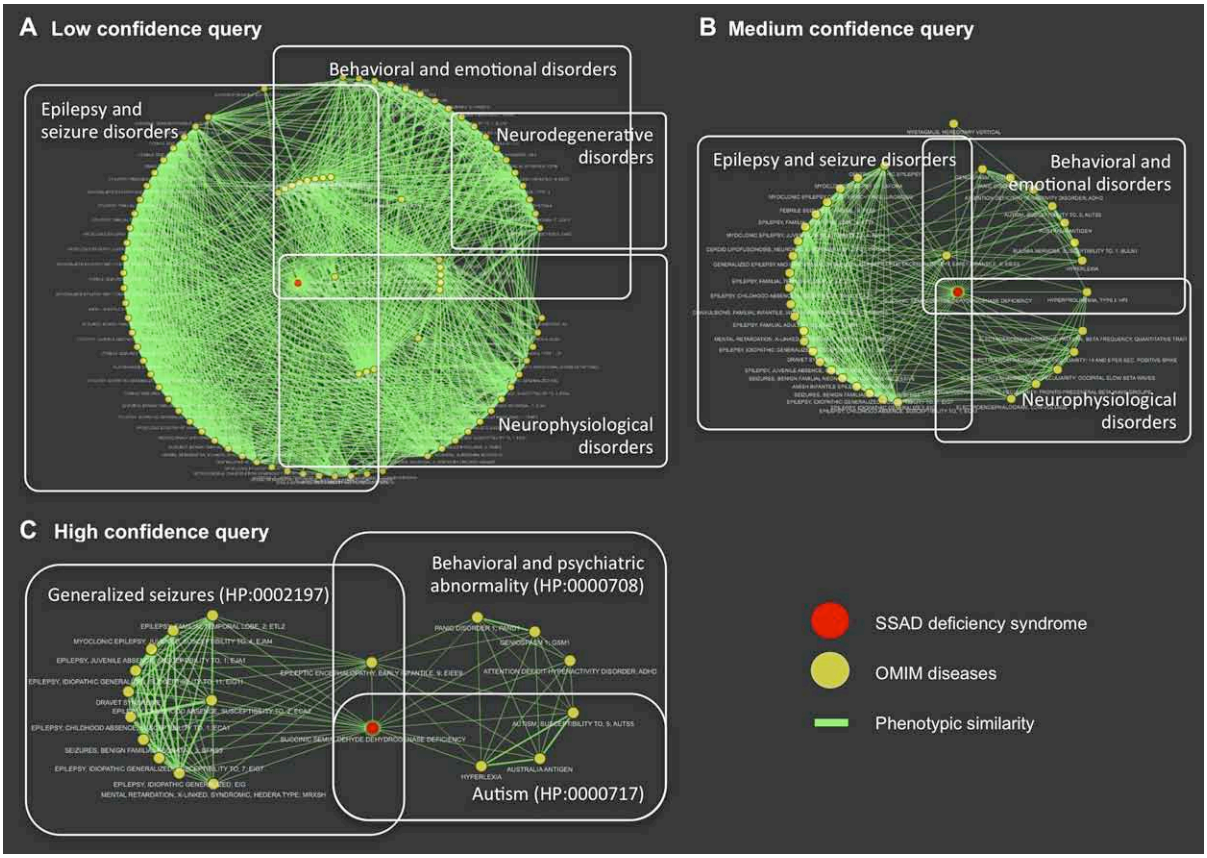


Figura 4.8: Redes de similitud fenotípica de enfermedades relacionadas con el síndrome SSADHD (tomada de RodríguezLópez et al. 2014). La figura muestra el resultado en PhenUMA de la consulta de las relaciones fenotípicas entre SSADHD y el resto de enfermedades aplicando varios niveles de confianza.

así en la Figura 4.8C se muestran grupos de enfermedades aún más definidos y enriquecidas por fenotipos algo más específicos, como puede ser el autismo. Con este ejemplo, se pone de manifiesto la utilidad de usar varios niveles de confianza, las facilidades que proporciona la herramienta para explorar las conexiones y la posibilidad estudiar los fenotipos representativos de los grupos de enfermedades y genes que contenidos en las redes resultantes.

4.2.3. Comparación con otras herramientas

En el momento de concluir el desarrollo de PhenUMA existían otras herramientas que permitían al usuario consultar información biomédica y se consideró adecuado realizar una comparativa para destacar las ventajas de cada una. Las principales características que se tuvieron en cuenta para comparar PhenUMA con el resto de herramientas fueron la capacidad para proporcionar acceso a relaciones fenotípicas y su integración con información funcional. Además, se consideró como características destacables la visualización de las relaciones y la capacidad para explorar los nodos y las relaciones, puesto que permiten el estudio de las relaciones con más profundidad/claridad a través de otros recursos. La Tabla 4.4 muestra un resumen de la comparación entre PhenUMA y otras herramientas (accedidas en el año 2014) para la consulta de información fenotípica.

Tabla 4.4: Comparativa de PhenUMA con otras herramientas similares.

Herramienta	Información Fenotípica	Medida de Similitud	Integración de Información	Descarga de Resultados	Visualizador de Redes
PhenUMA	Sí	IC	Sí	Sí	Sí
Phenomizer	Sí	IC	No	Sí	No
GeneMania	No	-	Sí	Sí	Sí
PhenomeNET	Sí	Jaccard	Sí	Sí	No
Malacards	No	MCRDS	Sí	No	Sí

Uno de los objetivos de PhenUMA es reunir información procedente de varias bases de datos, ontologías e interactomas. GeneMania (Warde-Farley et al., 2010) se desarrolló con un propósito similar. La interfaz de GeneMania permite la visualización de redes de genes incluyendo distintos tipos de información tales como: interacciones físicas entre proteínas, relaciones metabólicas, relaciones de co-expresión, co-localización y relaciones a partir de la similitud entre dominios de proteínas. Aunque GeneMania proporciona una visión muy completa de las interacciones que ocurren en el interior de la célula a

partir de unos genes de entrada, el objetivo de PhenUMA es incluir además información patológica. GeneMania no facilita de forma directa el estudio de las relaciones fenotípicas en el contexto biomolecular de un conjunto de genes. Sin embargo, desde PhenUMA esta información se visualiza de forma conjunta proporcionando relaciones funcionales asociadas a relaciones fenotípicas.

Otras herramientas, como MalaCards (Rappaport et al., 2013), se presenta como un completo repositorio de información, que integra información patológica y funcional, con la particularidad de que MalaCards incluye información fenotípica asociada a enfermedades de ratón y no en humanos. A parte de esto, la principal diferencia entre MalaCards y PhenUMA está en las diferentes filosofías para mostrar la información. MalaCards proporciona un resumen para cada gen o enfermedad que incluye diferentes puntos, mientras que PhenUMA, incorpora toda la información en una red, permitiendo ver con mayor facilidad toda la información asociada a una enfermedad o gen.

PhenomeNET (Hoehndorf et al., 2011) y Phenomizer (Köhler et al., 2009) son dos herramientas que usan la sintomatología de las enfermedades para establecer relaciones entre ellas, pero a diferencia de PhenUMA, estas relaciones no son proporcionadas dentro de un contexto funcional. Phenomizer demuestra los potenciales beneficios de los métodos de similitud semántica y del uso de ontologías para establecer relaciones entre información biomédica y su aplicación en el diagnóstico de enfermedades. Estas características han sido también incluidas en PhenUMA. Una propiedad adicional de PhenomeNET es que hace uso de varias ontologías para comparar genes y enfermedades de OMIM u Orphanet y fenotipos. Las ontologías integradas en PhenomeNET permiten ampliar las relaciones entre enfermedades mediante HPO con información fenotípica de otras especies.

Puesto que PhenomeNET y Phenomizer son las dos herramientas más comparables con PhenUMA, se han contrastado con mayor profundidad los resultados devueltos por ambas herramientas y PhenUMA. Para realizar esta comparación, se ha usado la lista de enfermedades fenotípicamente similares a la deficiencia de succínico semialdehído deshidrogenasa (OMIM 271980) y se ha realizado una comparación directa de los resultados obtenidos con Phenuma, Phenomizer y PhenomeNET. En primer lugar, se ha ordenado las lista de enfermedades obtenida por su valor de similitud y se han seleccionado las 50 primeras. A continuación, se ha realizado un enriquecimiento fenotípico usando un test hipergeométrico y usando Bonferroni (Material y Métodos, apartado 3.4.5) para corregir los p-valores resultantes. En la Tabla 4.5 se incluye el resultado de dicho enriquecimiento mostrando solo los fenotipos directamente asociados con SSADHD, junto con el valor asociado al contenido informativo de cada uno, lo que da una idea de su nivel especificidad. Por ejemplo, *status epilepticus* que es el fenotipo con mayor IC asociado, indica que es el

más específico relacionado con la SSADHD. Como también se observa en la Tabla 4.5, el resultado proporcionado por PhenUMA incluye un conjunto de enfermedades (similares a SSADHD) enriquecidas fenotípicamente por el fenotipo *status epilepticus*, sin embargo, no se ha encontrado un enriquecimiento de estas características para PhenomeNET o Phenomizer. Esto significa que, *status epilepticus* es un fenotipo representativo de las enfermedades que muestran una sintomatología común a SSADHD en PhenUMA.

Tabla 4.5: Enriquecimiento fenotípico de las enfermedades OMIM similares a deficiencia de SSAHD.

Fenotipos	**IC**	**PhenUMA**	**Phenomizer**	**PhenomeNET**
Status epilepticus	0,71	$1,49e^{5}$	0,58	0,14
Crisis de Ausencia	0,68	$6,75e^{11}$	$1,91e^{11}$	0,29
Convulsiones mioclónicas generalizadas	0,60	$2,3e^{3}$	$6,52e^{4}$	1
Ansiedad	0.58	$6,47e^{3}$	0,48	
Psicosis	0.56	$6,61e^{4}$	1	1
Convulsiones tónico-clónicas generalizadas	0,56	$3,53e^{29}$	$2,2e^{25}$	$1,26e^{14}$
Comportamiento agresivo	0.54	$6,5e^{9}$	1	1
Anomalía en el EEG	0.49	$4,31e^{14}$	$5,47e^{5}$	1
Hiperactividad	0.48	$3,51e^{3}$	1	1
Ataxia	0.317	1	1	$8,26e^{21}$
Movimiento de ojos anormal	0,307	0,875		$1,09e^{19}$
Hipotonía muscular	0,28	1		$2,73e^{7}$
Discapacidad intelectual	0,214	1	1	$1,72e^{3}$

Los algoritmos para calcular las medidas de similitud usados por PhenUMA y Phenomizer son similares, sin embargo, PhenomeNET usa Jaccard para calcular la similitud entre las enfermedades. Los valores de similitud entre cada par de enfermedades se proporcionan a través de la web de PhenomeNET, en el fichero "borderflow0.1". Este fichero incluye la similitud entre perfiles fenotípicos de varias especies: humanos, gusano, ratón, pez cebra, etc. Se usaron los valores de similitud para todas las enfermedades en PhenomeNET para compararlos con las similitudes entre enfermedades proporcionadas por PhenUMA. Para los resultados de ambas herramientas se construyó la curva ROC y se mostró gráficamente la FDR (*False Discovery Rate*), usando como datos de referencia las relaciones inferidas entre enfermedades. En la Figura 4.9 se incluye el resultado de ambas comparaciones.

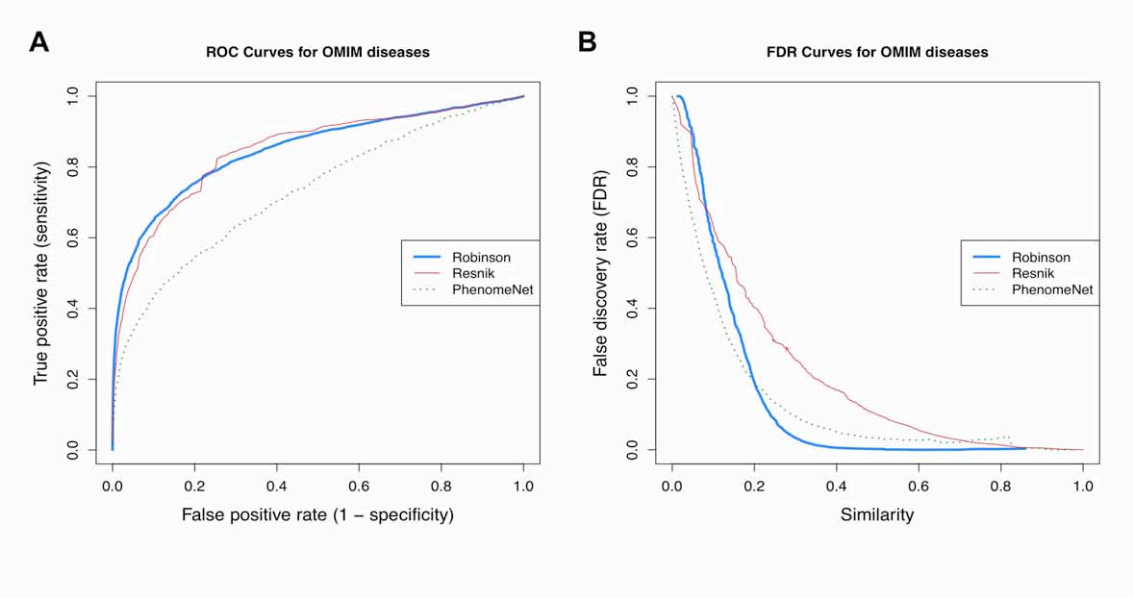


Figura 4.9: Curva ROC y FDR (*False Discovery Rate*) de la similitud fenotípica calculada con PhenUMA y PhenomeNET (tomada de Rodríguez-López et al. 2014). A. Curva ROC para la similitud fenotípica entre enfermedades OMIM. Para todos los casos se usaron las relaciones entre enfermedades inferidas, es decir, enfermedades relacionas porque comparten el mismo gen o genes. **B.** FDR para los valores crecientes de similitud fenotípica.

Aunque la medida de similitud que se usó en PhenUMA es la usada por Robinson en HPO (Resnik_Simétrica, Ecuación 1.6), comparamos también los resultados de PhenomeNET con los valores de similitud usando Resnik (ver Introducción, Ecuación 1.4). La curva ROC resultante muestra mejores resultados para las medidas de similitud (Resnik y Resnik_Simétrica) que para los resultados descargados de PhenomeNET. Con respecto a la FDR (Figura 4.9B), los resultados muestran una menor FDR, para la medida de similitud propuesta por Robinson que la de PhenomeNET, que a su vez muestra mejores resultados que Resnik. En conclusión, PhenUMA y las medidas utilizadas para establecer relaciones entre enfermedades proporcionan mejores resultados con respecto al conjunto de relaciones de referencia.

Actualmente el análisis de la información fenotípica se ha extendido considerablemente. Esto ha sido propiciado por la integración de HPO en varias bases de datos como DECIPHER , OMIM, ClinVar, MalaCards, Orphanet, etc. (Koehler et al., 2017). Herramientas más recientes como Phenolyzer (Yang et al., 2015) proporcionan una lista priorizada de genes candidatos a partir de una lista de fenotipos o enfermedades, similar a PhenUMA, e integran además de HPO, información patológica de varias bases de datos para construir esta lista priorizada: OMIM, Disease Ontology, ClinVar, GWAS. No obstante, la tendencia en los últimos años ha sido el uso de información fenotípica para analizar datos genéticos de pacientes. La plataforma Monarch Iniciative (Mungall et al., 2017) integra varias de estas herramientas como PhenIX (Smedley and Robinson, 2015), para la predicción de genes candidatos a partir de la exomas y cuadros fenotípicos, y Patient Archive (PA, `http://patientarchive.org`) que es un repositorio orientado a clínicos e investigadores con el objetivo de compartir información de pacientes. Con respecto a esta última, existen otros recursos con el mismo objetivo como DECIPHER (Firth et al., 2009), PhenoTips (Girdea et al., 2013) y Gene2MP (Chong et al., 2015).

4.3. Minería de textos para el estudio de relaciones fenotípicas entre genes implicados en el metabolismo de las aminas biogénicas

La información incluida en PhenUMA se obtiene a partir de varias bases de datos públicas: OMIM, Orphanet y HPO. La base de datos de PhenUMA integra relaciones fenotípicas y funcionales entre genes. Sin embargo, puesto que las relaciones entre genes y enfermedades proceden de estudios genéticos, no será posible asignar información fenotípica a genes de los que no se ha identificado ninguna variación causante de enfermedad. En algunas ocasiones, la falta de información en estas bases de datos no se corresponde con la disponible en la literatura. Esto produce un vacío de resultados en PhenUMA para algunos genes, incluso a pesar de que sus implicaciones patológicas estén bien documentadas.

En este trabajo, se han usado los genes implicados en el metabolismo de las aminas como caso de uso (Tabla 4.6) (Rodríguez-López et al., 2016). A pesar de la enorme cantidad de información fenomenológica y bioquímica sobre aminas biogénicas y su implicación en procesos patológicos, las dos bases de datos más completas sobre enfermedades humanas (OMIM o Orphanet) no reflejan este hecho. En concreto, se exploraron los genes relacionados con el metabolismo de las aminas biogénicas derivadas de la descarboxilación de los aminoácidos catiónicos ornitina e histidina, es decir, putrescina, espermidina y espermina (Put, Spd, Spm, respectivamente), y de aminoácidos aromáticos como serotonina y dopamina (5'-HT y DA) - que son ambos productos de la actividad dopa descarboxilasa (DDC), también nombrada como descarboxilasa de L-aminoácidos aromáticos - y las macromoléculas relacionadas con todos estos elementos. Estos genes se usaron como modelo para extraer relaciones gen-enfermedad de la literatura y de esta forma estudiar las relaciones fenotípicas y funcionales de los mismos.

Recientemente se han publicado una serie de revisiones que proporcionan información estructural, funcional actualizada y complementaria a la que aparece en el siguiente apartado sobre los elementos del metabolismo de aminas (Beaulieu et al., 2014; Hoyer et al., 1994; Sánchez-Jiménez et al., 2013; Panula et al., 2015).

4.3.1. Introducción a las aminas biogénicas

Las aminas biogénicas juegan un papel importante en los procesos fisiológicos más importantes, desde la proliferación y diferenciación celular hasta la nutrición, la respuesta

inmune, la neurobiología y la reproducción. Estos efectos se propagan a través de una amplia variedad de receptores específicos, señalización y vías metabólicas.

Entre las aminas biogénicas solo la diamina derivada de ornitina (Put) y las poliaminas (PA) derivadas de ornitina y S-adenosilmetionina descarboxilada (Spd y Spm), son sintetizadas por casi todas las células humanas con capacidad proliferativa. La razón es que las poliaminas (Spd y Spm) son esenciales para la síntesis y conformación de los ácidos nucleicos y las proteínas. La ornitina descarboxilasa (ODC), enzima responsable de la síntesis de la Put, es una diana de tratamientos terapéuticos anti-cáncer, por ejemplo DFMO, que es un inhibidor de ODC que se utiliza en terapias de cáncer de colon y neuroblastoma (Raj et al., 2013; Saulnier Sholler et al., 2015).

Por otro lado, la histamina (Hia) es la diamina producto de la reacción de la histidina descarboxilasa (HDC). Hia es el ligando de al menos cuatro tipos de receptores de membrana específicos, denominados H1R-H4R (codificados por los genes HRH1-4), miembros de la familia de receptores acoplados a proteínas G (GPCR, G Protein-Coupled Receptors) (Thurmond, 2015). La Hia es un ligando de receptores de N-metil aspartato (NMDA, codificados por los genes GRIN); sin embargo, usa un sitio de unión distinto a PA, quienes también se unen a este receptor (Burban et al., 2010). Por un lado, está claro que la Hia es capaz de modular la proliferación celular (Glatzer et al., 2013; Martinel Lamas et al., 2015), a veces mostrando una pauta de síntesis y efectos antagónicos a PA (Fajardo et al., 2001; García-Faroldi et al., 2009), pero además, es un neurotransmisor que está funcionalmente conectado con otras aminas neurotransmisoras (Ellenbroek and Ghiabi, 2014; Medina et al., 2008; Panula et al., 2014; Panula and Nuutinen, 2013). De hecho, se ha probado que HRH1 es una diana terapéutica para varias enfermedades neurológicas; además, se está investigando el uso farmacológico de antagonistas y antagonistas inversos de HRH3 (Passani and Blandina, 2011). La Hia es también un conocido mediador inmune que juega un papel central en alergias y otras patologías inmunes (Thurmond et al., 2008; Zampeli and Tiligada, 2009). Además, la Hia está implicada en la fisiología gástrica y es responsable de la secreción de ácido gástrico. Por consiguiente, las funciones fisiológicas más importantes y complejas están moduladas por aminas biogénicas (neurología, inmunología, nutrición, reproducción, proliferación y diferenciación) (Sánchez-Jiménez et al., 2013).

Los productos de la DDC, fundamentalmente serotonina (5'-HT) y dopamina (DA), son neurotransmisores y compuestos neuroendocrinos que transmiten su señal a través de una serie de miembros de la familia GPCR (Beaulieu et al., 2015). Las perturbaciones en la síntesis, transporte, degradación y señalización son la base de varias enfermedades neurológicas (por ejemplo, esquizofrenia, Parkinson, ansiedad, depresión, déficit de atención, etc.), enfermedades del sistema inmunológico y circulatorio (hipertensión, alergias

o psoriasis) y enfermedades raras (deficiencia de L-aminoácido aromático descarboxilasa, síndrome de Lech-Nyhan, síndrome de Prader-Willy, entre otros) (Arreola et al., 2015; Choi, 2015; Gratwicke et al., 2015; Johnston and Skene, 2015).

Se puede consultar más información sobre el metabolismo de aminas biogénicas, y de las proteínas implicadas en su metabolismo y sistemas de señalización de sus efectos biológicos, en las siguientes revisiones (Beaulieu et al., 2014; Hoyer et al., 1994; Sánchez-Jiménez et al., 2013; Panula et al., 2015).

Las aminas biogénicas derivadas de aminoácidos catiónicos y aromáticos están involucradas en multitud de interacciones con otros elementos moleculares (genes, proteínas y metabolitos) (Medina et al., 2005; Sánchez-Jiménez et al., 2007). Además, estas interacciones y sus efectos difieren entre distintos tipos celulares y tejidos. Sin embargo, la coordinación de todas estas interacciones es necesaria para la salud y estabilidad del organismo. Por ello, la comprensión de todas estas interacciones es absolutamente necesaria para entender numerosas patologías y síntomas. Debido a la complejidad de estas relaciones, es necesaria la integración de información bioquímica, molecular y fenotípica relacionada con aminas biogénicas, para proporcionar una visión global del problema. Es, por lo tanto, necesario el desarrollo de herramientas y análisis sistémicos para organizar, priorizar y curar la información clínica y molecular relacionada con los genes relacionados con el metabolismo de las aminas biogénicas.

4.3.2. Contraste de la presencia de genes relacionados con aminas biogénicas en bases de datos y en la literatura

La complejidad de los procesos biológicos en los que están involucrados las aminas biogénicas podría indicar su implicación en multitud de procesos patológicos, que pueden ser consecuencia de la perturbación de la actividad de alguno(s) de estos genes. Sin embargo, en bases de datos como OMIM, la presencia de los genes incluidos en la Tabla 4.6 es sorprendentemente escasa. Aunque en OMIM si se incluyen varias enfermedades asociadas a genes relacionados con la dopamina y serotonina, existe una falta de información en el caso de Hia y PA casi absoluta (Tabla 4.7). Por ejemplo, entre los elementos relacionados con PA, solo existe información patológica para 3 de los 13 genes que se han incluido en este análisis. Concretamente, solo existe información sobre el gen de la espermidina sintasa (SMS), que aparece relacionado el síndrome de Snyder-Robinson (OMIM 309583) (Pegg, 2014); las variaciones genéticas de la ornitina decarboxilasa (ODC), que aparecen relacionadas con un menor riesgo de padecer cáncer de colon (OMIM 114500); y por último, ARG1 relacionado con arginemia (OMIM 207800). En el caso de la histamina,

Tabla 4.6: Genes (nombre y *GeneSymbol*) relacionados con el metabolismo de las aminas.

Poliaminas	**Histamina**
Arginasa 1 (ARG1)	Histidina descarboxilasa (HDC)
Arginasa 2 (ARG2)	Histamina N-metil transferasa (HNMT)
Ornitina descarboxilasa (ODC1)	Receptor de histamina 1 (HRH1)
Antizima 1 de la ornitina-descarboxilasa (OAZ1)	Receptor de histamina 2 (HRH2)
Antizima 2 de la ornitina-descarboxilasa (OAZ2)	Receptor de histamina 3 (HRH3)
Antizima 3 de la ornitina-descarboxilasa (OAZ3)	Receptor de histamina 4 (HRH4)
Inhibidor de antizimas 1 (AZIN1)	
Inhibidor de antizimas 2 (AZIN2)	
Espermidina sintasa (SRM)	
Espermina sintasa (SMS)	
Espmermidina/espermina acetil transferasa (SAT1)	
Poliamina oxidasa (PAOX)	
Espermina oxidasa (SMOX)	
Dopamina/Serotonina	**Elementos compartidos por el metabolismo de varias aminas**
Tirosina hidroxilasa (TH)	Diamino oxidasa (AOC1)
Triptófano hidroxilasa (TPH1)	Amino oxidasa cobre dependiente 2 (AOC2)
Triptófano hidroxilasa 2 (TPH2)	Amino oxidasa cobre dependiente 3 (AOC3)
L-Aminoácido aromático descarboxilasa (DDC)	Monoamina oxidasa A (MAOA)
Receptor de dopamina 1 (DRD1)	Monoamina oxidasa B (MAOB)
Receptor de dopamina 2 (DRD2)	Transportador de soluto familia 22 miembro 2 (SLC22A2)
Receptor de dopamina 3 (DRD3)	Transportador de soluto familia 22 miembro 3 (SLC22A3)
Receptor de dopamina 4 (DRD4)	Transportador de soluto familia 3 miembro 2 (SLC3A2)
Receptor de dopamina 5 (DRD5)	Transportador de soluto familia 6 (SLC6A3, SLC6A4)
Receptor de 5-hidroxitriptamina 1 (HTR1A, HTR1B)	Transportador de soluto familia 12 (SLC12A8)
Receptor de 5-hidroxitriptamina 2 (HTR2A, HTR2B, HTR2C)	Transportador de soluto familia 18 (SLC18A1)
Receptor de 5-hidroxitriptamina 3 (HTR3A, HTR3B, HTR3C, HT3D, HTR3E)	Transportador de soluto familia 18 (SLC18A2)
Receptor de 5-hidroxitriptamina 4 (HTR4)	Transglutaminasa 1 (TGM1)
Receptor de 5-hidroxitriptamina 5 (HTR5A)	Transglutaminasa 2 (TGM2)
Receptor de 5-hidroxitriptamina 6 (HTR6)	Receptor de N-metil aspartator 1 (GRIN1)
Receptor de 5-hidroxitriptamina 7 (HTR7)	Receptor de N-metil aspartator 2 (GRIN2A, GRIN2B)

en OMIM encontramos la relación entre las forma truncadas de HDC con susceptibilidad a padecer síndrome de Gilles de la Tourette (OMIM 137580) (Castellan Baldan et al., 2014). Adicionalmente, la falta de Histamina N-Metil transferasa (HNMT), una enzima que participa en la degradación Hia, se asocia con la susceptibilidad de padecer asma (OMIM 600807) y con retraso mental (OMIM 616937).

La información extraída de OMIM y expuesta en la Tabla 4.7, contrasta con la información presente en la literatura biomédica relacionada con aminas biogénicas. De hecho, al realizar una exploración preliminar en la base de datos PubMed (`http://www.ncbi.nlm.nih.gov/pubmed/`) para analizar la presencia de estos genes en la literatura, se encontró que el número de referencias bibliográficas asociadas a "human polyamines" o "human histamine" supera las 40.000 citas; de igual modo, si se consulta "human dopamine" o "human serotonin" el número de citas es alrededor de 60.000. Incluso siendo más restrictivo, e incluyendo la palabra "disease", para consultar información patológica de algunas de las aminas, se obtienen entre 6.000 y 20.000 referencias. No obstante, a pesar de la presencia en la literatura biomédica de los genes relacionados con aminas, dicho conocimiento no está reflejado en repositorios de enfermedades humanas como OMIM u Orphanet.

Estos hechos ponen de manifiesto un "hueco" en la información fisiopatológica de enfermedades relacionadas con aminas biogénicas en los repositorios actuales de información, especialmente para los casos de las diaminas y las poliaminas. Por lo tanto, es necesario aplicar otras estrategias para estudiar las patologías relacionadas con estas enfermedades. Una primera aproximación, realizada por nuestro grupo de investigación, consistió en la construcción de una red de enfermedades relacionadas con Hia, que contenía alrededor de 20 enfermedades (Pino-Ángeles et al., 2012), concluyendo que el uso de herramientas de minería de textos permite que surjan nuevas relaciones asociadas con el metabolismo de las aminas biogénicas.

Por esta razón, en este trabajo se presenta un flujo de trabajo (Material y Métodos, apartado 3.6) cuyo objetivo fue extraer relaciones gen-enfermedad de la literatura y, de esta forma, construir los perfiles fenotípicos de estos genes. Este flujo de trabajo, combina el uso herramientas de minería de textos con ontologías biomédicas. Esto permitió el estudio de información funcional y fenotípica entre genes relacionados con el metabolismo de las aminas biogénicas.

Tabla 4.7: Relaciones gen-enfermedad presentes en OMIM. En las columnas de esta tabla se muestra el gen (geneSymbol), el código OMIM de la enfermedad y el nombre de cada enfermedad.

Gen	OMIM	Nombre
Dopamina/Serotonina		
TH	605407	Síndrome de Segawa
TPH2	631003	Susceptibilidad a déficit de atención
	608516	Susceptibilidad a depresión unipolar
DDC	608643	Deficiencia de L-Aminoácido aromático descarboxilasa
DRD3	190300	Temblores (susceptibilidad)
	181500	Susceptibilidad a la esquizofrenia
DRD4	143465	Déficit de atención/Hiperactividad
DRD5	143465	Déficit de atención/Hiperactividad (susceptibilidad)
	606798	Blefaroespasmo
HTR1A	614674	Fiebre periódica, asociada al ciclo menstrual
HTR2A	103750	Susceptibilidad a dependencia del alcohol
	606788	Susceptibilidad a anorexia nerviosa
	608516	Susceptibilidad a depresión unipolar
	164230	Susceptibilidad a trastorno obsesivo-compulsivo
	181500	Susceptibilidad a la esquizofrenia
	608516	Susceptibilidad a trastorno afectivo
Poliaminas		
ARG1	207800	Arginemia
ODC1	114500	Menor riesgo a padecer cáncer de colon
SMS	309538	Síndrome de Synder-Robinson
Histamina		
HDC	137580	Susceptibilidad a Gilles de la Tourette
HNMT	619739	Retraso mental
	600807	Susceptibilidad al asma
Elementos compartidos por el metabolismo de varias aminas		
MAOA	300615	Síndrome de Brunner
SLC6A3	613135	Distonía-parkinsonismo infantil
	188890	Protección contra la dependencia a la nicotina
TGM1	242300	Ictiosis
GRIN1	614254	Retraso mental
GRIN2A	245570	Epilepsia con trastorno del habla
GRIN2B	616139	Encefalopatía epiléptica
	613970	Retraso mental (dominante autosómico 6)

4.3.3. La incorporación de herramientas de minería de textos permite enriquecer las relaciones gen-enfermedad presentes en las bases de datos

Como se detalló en el capítulo Material y Métodos (sección 3.6), se usaron herramientas de minería de textos (DISEASES y DisGeNet) para extraer de la literatura relaciones entre enfermedades y genes asociados con el metabolismo de las aminas biogénicas. En total, se obtuvieron 429 relaciones entre 53 genes (de los 60 incluidos en el análisis) y 129 enfermedades. El 96 % de las relaciones proceden solo de la literatura, siendo solo 36 (el 4 %) las incluidas además en OMIM. La baja presencia de estas relaciones en OMIM puede deberse a la procedencia de la información contenida en las mismas. En dichas bases de datos se incluyen variaciones en genes candidatos a ser causantes de alguna enfermedad. Sin embargo, para la gran mayoría de los casos no se han detectado -o estudiado- mutaciones/variantes (por ejemplo, SNP) asociadas a genes del metabolismo de aminas que permitan anotarlos directamente a determinadas patologías. No obstante, si se analiza la literatura con mayor profundidad es posible identificar genes o procesos biológicos que podrían estar afectados y ser responsables de estas enfermedades, como se expuso anteriormente. El caso de estudio seleccionado es una muestra de que la complejidad de la etiología de algunas enfermedades, en ocasiones no se debe a un solo gen, sino que implican interacciones más complejas. En dichas interacciones, probablemente están incluidos elementos reguladores de la expresión de los genes que se muestran en la Tabla 4.6 que aún desconocemos parcial o totalmente.

En la Tabla 4.8 se muestra la información completa de dichas relaciones, y en la Figura 4.10 se incluyen las relaciones que involucran a enfermedades relacionadas con más de una amina. El conjunto de enfermedades más extenso es el que comparten los genes relacionados con la dopamina/serotonina y los elementos compartidos por los metabolismos de varias aminas, que mayoritariamente se componen de enfermedades neurológicas. Por el contrario, los genes asociados con la histamina se asocian con un número menor de enfermedades, algunas de ellas también neurodegenerativas o relacionadas con alteraciones de sistema inmunológico como asma, rinitis alérgica o dermatitis. Varias enfermedades se asocian con genes de varias aminas (35 de las 129); sin embargo, solo están asociadas a los cuatro grupos el síndrome de Gilles de Tourette y la obesidad.

Más de la mitad de las enfermedades (56 %) encontradas en la literatura se relacionan con el metabolismo de la dopamina y serotonina cuyo conjunto de genes es el más extenso. El resultado no solo es el esperado por la cantidad de patologías relacionadas, sino también por el tipo de enfermedades, mayoritariamente neurológicas y neurodege-

nerativas. Como se ha comentado anteriormente, la serotonina (5'-HT) y dopamina (DA) son neurotransmisores. Anomalías en la síntesis, transporte, degradación o recepción son parte de la etiología de varias enfermedades neurológicas, del sistema inmunológico y circulatorio, así como también de algunas enfermedades raras. Además, en este conjunto de enfermedades se incluyen patologías asociadas a adicciones, como al tabaco, el juego y el alcohol. Estos resultados ponen de manifiesto la complejidad de ciertas enfermedades, sobre todo las neurológicas, que suelen ser poligénicas: la esquizofrenia (31 genes), autismo (23 genes), síndrome de Gilles de la Tourette (21 genes), déficit de atención (21 genes) o la dependencia al alcohol (20 genes).

Tanto en la Figura 4.10 como en la Tabla 4.8 se incluyen algunos transportadores de membrana como SLC6A3 y SLC6A4 (transportadores de serotonina) relacionados con varias enfermedades neurológicas, que también se relacionan con la dopamina; por ejemplo, algunas adicciones, autismo (Sutcliffe et al., 2005), dislexia, demencia con cuerpos de Lewy (O'Brien et al., 2004), síndrome de Gilles de la Tourette (Gunther et al., 2012), pánico (Strug et al., 2010), fobias (Furmark et al., 2004) o susceptibilidad al síndrome de Asperger (Wendland et al., 2008). Además, estos neurotransmisores se relacionan con patologías como la esquizofrenia (Sáiz et al., 2010), susceptibilidad a la migraña (Liu et al., 2011), déficit de atención (Tong et al., 2015), y temblores, que también aparecen vinculados a HNMT.

Tabla 4.8: Relaciones gen-enfermedad a partir de las herramientas de minería de textos. En las columnas de esta tabla se muestra el nombre de las enfermedades, su código OMIM y el conjunto de genes con el que están relacionadas (geneSymbol), respectivamente.

Enfermedad	**OMIM**	**Genes**
Aciduria arginosuccínica	207900	ARG1
Aciduria D-2-hidroxiglutárica	600721	GRIN2A; GRIN2B
Adenoma pituitario	600634	DRD2
Alzheimer	104300	AOC2; AOC3; GRIN2A; GRIN2B; HRH3; HTR1A;MAOA; MAOB; SLC6A4; TH
Amiloidosis	105150	AOC2; AOC3
Apnea central del sueño	107640	MAOA
Argininemia	207800	ARG1
Autismo	209850	DRD1; DRD2; DRD3; DRD4; DRD5; GRIN2A; GRIN2B; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR3A; HTR3C; HTR5A; HTR7; MAOA; MAOB; SLC18A1; SLC6A3; SLC6A4; TPH1; TPH2

Enfermedad	OMIM	Genes
Blefaroespasmo	606798	DRD5
Cáncer colorrectal	114500	ODC1; PAOX
Cáncer de mama	114480	HNMT; HTR3A; ODC1
Cáncer de paratiroides	608266	SMS
Cáncer de próstata	176807	ODC1
Cáncer de pulmón	211980	DDC; HTR3A; ODC1; SAT1
Cáncer de testículos	273300	DRD1
Cáncer de uretra	191600	TPH1
Cáncer folicular de tiroides	188470	ARG2
Carcinoma adrenocortical	202300	HTR2B
Carcinoma hepatocelular	114550	ARG1; AZIN1; OAZ3; ODC1; TH
Cinetosis	158280	HRH1; HTR1A; HTR3A
Cistinuria	220100	SLC3A2
Cutis laxa	123700	SLC3A2
Daltonismo	216900	HRH2
Deficiencia de L-aminoácido aromático descarboxilasa	608643	DDC
Deficiencia sistémica primaria de carnitina	212140	SLC22A2
Deficit de atención-hiperactividad	143465	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; GRIN2A; HRH3; HTR1B; HTR2A; HTR2C; HTR4; HTR7; MAOA; MAOB; SLC6A3; SLC6A4; TH; TPH1; TPH2
Demencia con cuerpos de Lewy	127750	MAOA; MAOB; SLC6A3; SLC6A4; TH
Dependencia del alcohol	103780	DRD1; DRD2; DRD3; DRD4; GRIN1; GRIN2A; GRIN2B; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR3B; HTR7; MAOA; MAOB; SLC6A3; SLC6A4; TPH1; TPH2
Depresión	608520	DRD3; DRD4; HTR2C; MAOA
Depresión unipolar	608516	HTR1A; HTR2C; HTR3A; MAOA
Dermatitis	603165	HRH1; HRH4
Dermografismo	125635	HDC; HRH1
Diabetes mellitus tipo 1	125853	AOC3; HTR2C

Enfermedad	OMIM	Genes
Disautonomía familiar	223900	MAOA
Disostosis cleidocraneal	119600	OAZ1; OAZ2
Distonía-parkinsonismo infantil	613135	SLC6A3
Distrofia muscular facioescapulohumeral	158900	SLC22A3
Enfermedad de Addison	103230	DDC; TPH1
Enfermedad de Darier-White	124200	SAT1
Enfermedad de Huntington	143100	DRD1; DRD2; DRD5; GRIN2A; GRIN2B; MAOB; TH
Enfermedad de Machado-Joseph	109150	HTR1A
Enfermedad de Wilson	277900	DRD2
Enfermedad del espectro PBD-Zellweger	214100	TPH2
Enfermedad venooclusiva pulmonar	265450	SLC6A4
Esclerosis lateral amiotrófica	105400	MAOB; TH
Esclerosis tuberosa 1	191100	HTR6
Esquizofrenia	181500	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; GRIN1; GRIN2A; GRIN2B; HRH1; HRH3; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR3B; HTR3D; HTR4; HTR5A; HTR6; HTR7; MAOA; MAOB; SLC18A1; SLC6A3; SLC6A4; TH; TPH1; TPH2; SLC3A2
Fenilcetonuria	261600	TH; TPH1
Feocromocitoma	171300	DDC; TH
Fobia	608251	MAOA; SLC6A4
Hernia de hiato	142400	OAZ1
Hernia diafragmática congénita	142340	PAOX; SMOX
Heterotropía	300049	GRIN2A
Hiperprolactinemia	615555	DRD2
Hipertensión intracraneal idiopática	243200	HTR2C
Hipertensión pulmonar	178600	SLC6A4

Enfermedad	**OMIM**	**Genes**
Holoprosencefalia	142945	HTR5A
Ictus isquémico	601367	AOC3; GRIN2A; GRIN2B; HTR1A; MAOB; SLC6A4; SMOX; TH
Insensibilidad congénita al dolor	162400	MAOA
Insomnio familiar fatal	600072	TPH1
Ludopatía	606349	DRD1; DRD2; DRD3; DRD4; DRD5; GRIN1; HTR1B; HTR2A; MAOA; MAOB; SLC6A3; SLC6A4
Lupus eritematoso sistémico	152700	GRIN2A
Mastocytosis/urticaria pigmentosa	154800	HDC; TPH1
Melanoma cutáneo (Maligno)	155600	ODC1
Migraña hemipléjica familar	141500	DRD2
Miopia	160700	TH
Miositis	160750	HTR3A; SLC6A4
Muerte súbita del lactante	272120	HTR1A; MAOA; SLC6A4; TH; TPH1;
Necrosis estriatal bilateral infantil	271930	DDC; TH; TPH1
Nefropatía epidémica	124100	AOC3
Obesidad	601665	AOC3; DRD2; HRH3; HTR1B; HTR2A; HTR2C; SAT1
Obesidad/Síndrome metabólico	605552	HTR2C
Oncocitoma	553000	SLC3A2
Osteoporosis	166710	SMS
Pánico	167870	DRD1; DRD2; DRD4; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; MAOA; SLC6A4; TPH1; TPH2
Papiloma del plexo coroideo	260500	HTR2C
Parálisis periódica cardiodisrítmica de Andersen	170390	SLC18A1
Parálisis supranuclear progresiva	601104	DRD2; TH
Parkinson	168600	DDC; DRD1; DRD2; DRD3; DRD4; GRIN2A; GRIN2B; HNMT; HRH3; HTR1A; HTR3A; MAOA; MAOB; SLC6A3; SLC6A4; TH; TPH1

Enfermedad	OMIM	Genes
Polidactilia postaxial tipo A1	174200	HTR6
Poliposis adenomatosa familiar	175100	ODC1
Psicosis maniaco-depresiva	125480	SLC6A3
Raquitismo hipofosfatímico	193100	SMS
Reflujo gastroesofágico	109350	HRH2
Retraso mental	300046	MAOA
Rinitis alérgica	607154	HDC; HNMT; HRH1; HRH2; HRH4
Síndrome de Bloom	210900	HTR5A
Síndrome de Brunner	300615	MAOA
Síndrome de disquinesia ciliar	244400	OAZ3
Síndrome de Fanconi primario renotubular	134600	SLC22A2
Síndrome de Gilles de la Tourette	137580	DRD1; DRD2; DRD3; DRD4; DRD5; HDC; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR3B; HTR5A; HTR7; MAOA; OAZ3; SLC22A3; SLC6A3; SLC6A4; TPH2
Síndrome de Landau-Kleffner	245570	GRIN2A
Síndrome de Lennox-Gastaut	606369	GRIN2A
Síndrome de Lesh-Nyhan	300322	DRD5
Síndrome de Prader-Willi	176270	HTR2B, HTR2C
Síndrome de Rett	312750	HTR7; TH
Síndrome de Seckel	210600	HTR5A
Síndrome de Segawa	605407	TH
Síndrome de Snyder-Robinson	309583	SMS
Síndrome de Sotos	117550	DRD1
Síndrome de Weaver	277590	TH
Síndrome de Wernicke-Korsakoff	277730	TPH1
Síndrome de Wolf-Hirshhorn	194190	DRD5
Síndrome de Wolfram	222300	DRD5
Síndrome del adelanto de fase	604348	TPH2
Síndrome nefrótico congénito	256300	SLC3A2
Suceptibilidad a atrofia multisitémica	146500	DDC; DRD2; TH; TPH1

Enfermedad	OMIM	Genes
Suceptibilidad a espondiloartropatía	106300	ARG2
Susceptibilidad a adicción al tabaco	188890	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; HTR1A; HTR2A; HTR3B; HTR7; MAOA; SLC6A3; SLC6A4; TH; TPH1
Susceptibilidad a glioma	137800	ODC1
Susceptibilidad a infarto de miocardio	608446	TH
Susceptibilidad a la dislexia	127700	DRD3; DRD4; DRD5; SLC6A3
Susceptibilidad a la enfermedad de Hirshsprung	142623	TH
Susceptibilidad a leishmaniasis	608207	ARG1; ODC1
Susceptibilidad a malaria	609148	HRH1; OAZ1; ODC1; SRM
Susceptibilidad a migraña con aura (tipo 7)	609179	DRD2; DRD4; HTR2B; HTR2C; MAOA; SLC6A4
Susceptibilidad a migraña con o sin aura	157300	DRD1; DRD2; DRD3; DRD4; DRD5; HRH3; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR7; MAOA; SLC6A4; TPH1
Susceptibilidad a migraña sin aura	607501	DRD1; DRD2; DRD3; DRD4; DRD5; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR6; MAOA; SLC6A4
Susceptibilidad a neuroblastoma	256700	DDC; HTR3A; MAOA; MAOB; OAZ2; ODC1; TH
Susceptibilidad a psoriasis	177900	ODC1
Susceptibilidad al asma	600807	ARG1; ARG2; HDC; HNMT; HRH1; HRH2; HRH4
Susceptibilidad al cáncer de pulmón	612571	SAT1
Susceptibilidad al síndrome de Asperger	300494	HTR2A; SLC6A4
Temblores	190300	DRD3; HNMT;
Tricotilomanía	613229	MAOA
Vitiligo	193200	DDC

Enfermedad	OMIM	Genes
Vitreorretinopatía exudativa familiar	133780	MAOA

También se localizaron, además de las enfermedades neurológicas, algunos tipos de cáncer asociados a alteraciones en el metabolismo de las aminas (principalmente poliaminas): cáncer de pulmón, de hígado, mama, próstata, colorrectal, tiroides y piel; y otras como la histamina se asocian al cáncer de mama (He and Zhang, 2006; Martinel Lamas et al., 2015; Medina et al., 2008).

La histamina es una amina implicada directa- e indirectamente en múltiples procesos patológicos distintos, por lo que se puede considerar su carácter pleiotrópico. En este sentido, los resultados obtenidos son coherentes con el conocimiento actual, al indicar que elementos involucrados en su metabolismo se asocian tanto con cáncer, como con enfermedades neurológicas o inmunológicas, como la dermatitis y la rinitis alérgica. Este resultado era previsible, dado que la histamina es un mediador inmunológico esencial en alergias y en múltiples patologías inmunológicas, tales como el asma (como puede verse en la Figura 4.10) (Medina et al., 2008; Pino-Ángeles et al., 2012; Zampeli and Tiligada, 2009).

Las interacciones entre diferentes aminas biogénicas podría también tener repercusión en algunos procesos patológicos que incluyen simultáneamente inflamación y proliferación (o regeneración tisular), por ejemplo, leucemias o mastocitosis graves, crecimiento de tumores endocrinos y problemas hepáticos. Este es también el caso de infecciones parasitarias, como la malaria. Del análisis de la literatura se obtuvieron varios genes de poliaminas (SRM, OA1, ODC) asociados con la susceptibilidad a la malaria junto con HRH1. Los parásitos requieren de PA para proliferar y las zonas inflamadas constituyen una fuente importante de aminas biogénicas y ROS, cuyas consecuencias aún no han sido evaluadas (Mecheri, 2012; Sánchez-Jiménez et al., 2013).

4.3.4. Las relaciones emergentes de la literatura permiten ahondar en la información sintomatológica asociadas a aminas biogénicas

A partir de las relaciones gen-enfermedad, mencionadas en el apartado anterior, fue posible asignar un perfil fenotípico a los genes analizados. El procedimiento que se siguió fue el descrito en Material y Métodos (apartado 3.4.1). De los 60 genes estudiados, 53 de ellos aparecen relacionados en la literatura con enfermedades anotadas en la ontología. En total, la información fenotípica asociada a estos genes está compuesta de 2.975 anota-

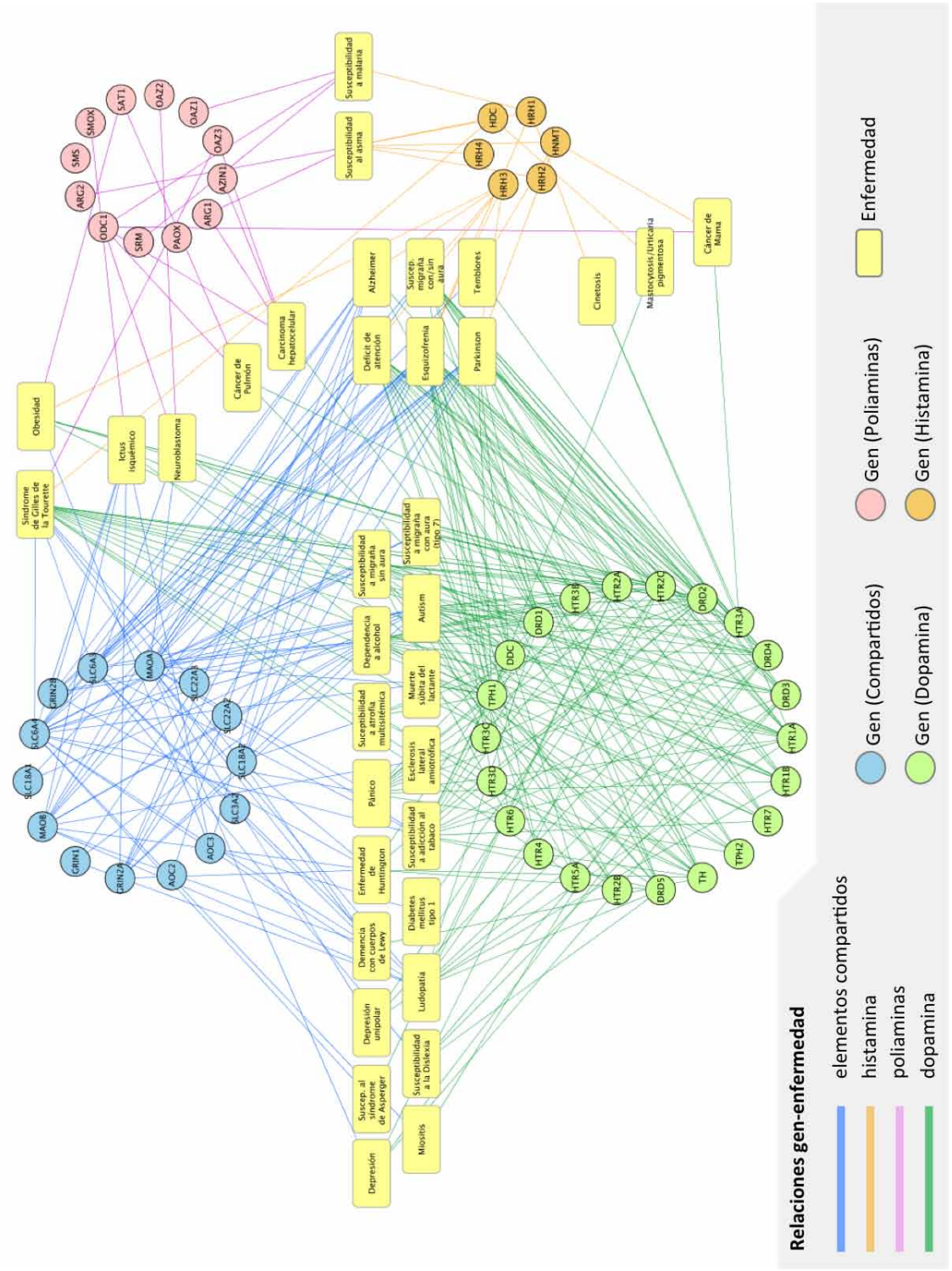


Figura 4.10: Relaciones enfermedades y genes asociados con el metabolismo de las aminas a partir de la literatura. Usando las herramientas DISEASES y DisGeNET se han obtenido nuevas relaciones entre genes y enfermedades que no están presentes en OMIM u Orphanet. En esta figura se han incluido solo las enfermedades compartidas por más de un grupo de genes. El resto de relaciones se encuentran en la Tabla 4.6.

ciones entre los 53 genes y los 928 fenotipos (términos HPO) distribuidos por varias ramas de la ontología. A modo de recordatorio, indicar que la HPO clasifica los fenotipos patológicos (alrededor de 10.000 términos) organizados en distintas ramas, concretamente 23. Dichas ramas hacen referencia a un dominio anatómico afectado por alguna anomalía, (anomalías de las extremidades o anomalías oculares) o a anomalías de un sistema completo (anomalías del sistema endocrino).

Para estudiar qué fenotipos o dominios están más relacionados con las alteraciones del metabolismo de las aminas, se analizaron las anotaciones de los genes para valorar la representatividad de cada rama de la ontología. Para ello, se obtuvieron los genes asociados a los fenotipos de cada rama para todas las aminas y para cada tipo de amina por separado. Posteriormente, se utilizó el test de Fisher para estimar que ramas estaban significativamente enriquecidas. A partir de este test se compararon los genes relacionados con aminas anotados en cada rama frente al conjunto total de genes anotados en dicha rama. Puesto que se trata de un test con múltiples comparaciones, el p-valor obtenido para cada rama se ajustó usando el método Benjamini-Hochberg (Material y Métodos 3.1.3). La Figura 4.11 muestra las ramas enriquecidas (p-valor ajustado <0,05) por estos genes. Se obtuvo una relación significativa para 17 de las 23 ramas, los recuadros en blanco indican que no existen en esa rama fenotipos asociados para ese conjunto de genes, o que la relación no es significativa.

El análisis se realizó para el conjunto total de genes y para cada conjunto por separado. En la primera línea de la tabla de la Figura 4.11 se muestran las ramas enriquecidas por el total de los genes asociados con aminas. Los fenotipos asociados con anomalías del sistema nervioso son los más representativos para este conjunto de genes, de hecho, 48 de los 53 (90 %) genes están asociados con al menos un fenotipo que afecta al sistema nervioso. Además, al menos 36 genes (67 %) se relacionan con fenotipos que afectan a la cara o el cuello. En ocasiones, este tipo de síntomas se relacionan con patologías neurológicas, como es el caso de "*Mask-like facies*" o cara de máscara, que es el fenotipo más frecuente en el conjunto de genes del estudio y que es característico de enfermedades como el Parkinson o el síndrome de Segawa (también conocido como distonía sensible a la dopamina). Otros fenotipos asociados a los genes analizados se corresponden con síntomas característicos de enfermedades raras, por ejemplo, el hipertelorismo ocular, que consiste en una distancia anormal entre los ojos, habitual en pacientes diagnosticados con el síndrome de Sotos, el síndrome de Weaver o el síndrome de Snyder-Robinson. Otro ejemplo es la micrognatia (mandíbula anormalmente pequeña), que corresponden a malformaciones observadas en pacientes de algunas enfermedades raras como Síndrome de Wolf-Hirshhorn. Otras ramas enriquecidas de forma significativa son, por un lado, las anomalías en el ab-

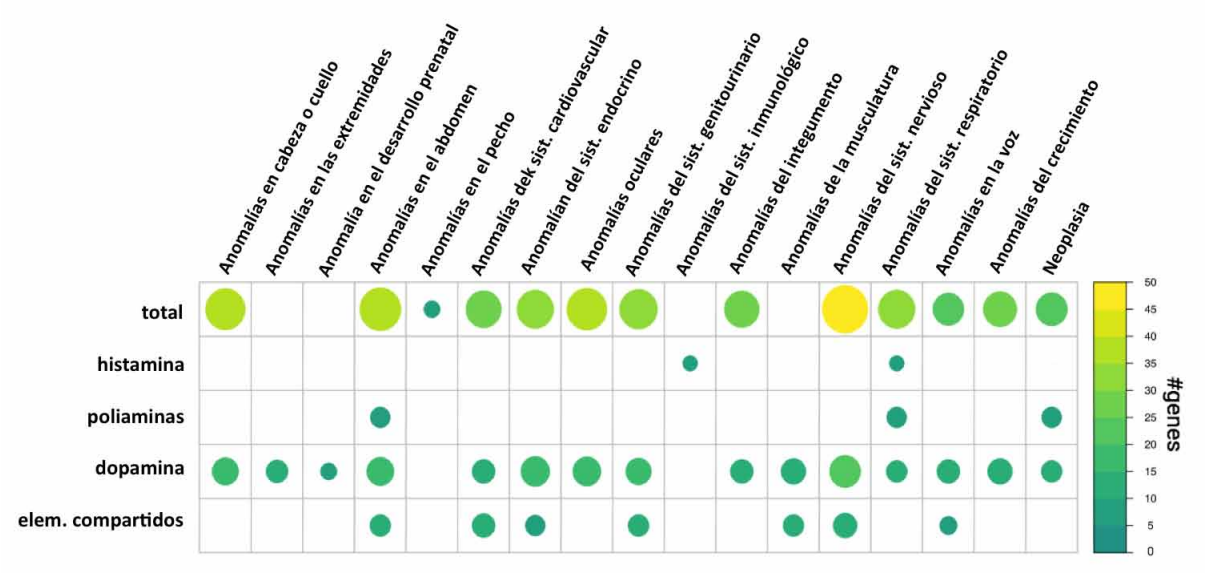


Figura 4.11: Enriquecimiento de cada una de las ramas de la ontología por los distintos grupos de genes asociados las aminas La tabla muestra el número de genes de cada uno de los grupos (Tabla 4.6) asociados con los fenotipos de cada rama de la ontología. Para cada rama y grupo de genes se ha obtenido la significación estadística usando un test de Fisher. En la tabla solo se han incluido los solapamientos estadísticamente significativos (p-valor <0,05).

domen, destacando la frecuencia de algunos síntomas como nauseas, diarrea, disfagia, vómitos o estreñimiento. También predominan los problemas respiratorios y endocrinos. Con respecto al sistema endocrino, se observó que todos los genes relacionados con el metabolismo de la serotonina estaban asociados a un aumento en los niveles de serotonina sérica ("*increased serum serotonin*"), que se asocia con el autismo; sin embargo, no encontramos relación entre estos genes y el autismo en las bases de datos públicas.

Además del análisis global de todos los genes, también se analizó cada conjunto por separado. El enriquecimiento de las ramas va a depender del tamaño del conjunto estudiado. Por ejemplo, los genes relacionados con el metabolismo de la histamina son únicamente 6. Esto explica que solo se encuentre enriquecimiento significativo de dos ramas, lo cual contrasta con la variabilidad fenotípica presente en los genes asociados al metabolismo de la serotonina/dopamina. Los genes asociados a la histamina se asocian a fenotipos del sistema respiratorio e inmunológico, concretamente al asma, como ya se indicaba en las relaciones mostradas en la Figura 4.10.

El análisis anterior tenía como objetivo proporcionar una visión general de los fenotipos asociados con genes implicados en el metabolismo de aminas. No obstante, se consideró oportuno profundizar algo más en la representatividad de los fenotipos más específicos en cada uno de los conjuntos. Para ello, se realizó un enriquecimiento fenotípico de cada conjunto de genes usando un test hipergeométrico (ver Material y Métodos, punto 3.1). La Tabla 4.9 muestra los fenotipos más enriquecidos para cada conjunto de genes, excepto para los genes relacionados con el metabolismo de la histamina, para los cuales no se encontró ningún fenotipo enriquecido de forma significativa.

La Tabla 4.9 muestra los fenotipos más específicos asociados a cada conjunto de genes. Destacan los genes relacionados con la serotonina y dopamina, que están enriquecidos con términos HPO relacionados con trastornos del comportamiento y del movimiento. Estos resultados, junto con los mostrados en la Figura 4.11, ponen de manifiesto la relación entre las poliaminas y varios tipos de cáncer, así como la implicación de la dopamina y la serotonina en patologías neurológicas y neurodegenerativas. Aunque estos resultados eran previsibles, pueden ser un indicativo de la validez del procedimiento propuesto en este trabajo. El esquema propuesto, demuestra que es posible extraer información de forma sistemática sobre los efectos patológicos en los que están implicados las aminas biogénicas. El conocimiento extraído representa adecuadamente el conocimiento actual relativo a esos genes, por tanto, puede constituir un punto de partida para futuras aproximaciones más precisas.

Tabla 4.9: Fenotipos más enriquecidos por los genes relacionados con el metabolismo de las distintas aminas estudiadas en este trabajo.

HPO Id	Nombre	P-Valor
Serotonina/Dopamina (23 genes)		
HP:0000708	Alteración del comportamiento	$4,755e^{08}$
HP:0002360	Trastornos del sueño	$9,996e^{04}$
HP:0001300	Parkinsonismo	$1,753e^{03}$
HP:0002071	Alteración de la función motora extrapiramidal	$1,753e^{03}$
HP:0100022	Alteración del movimiento	$1,753e^{03}$
Poliaminas (11 genes)		
HP:0003218	Oroticaciduria	0,0235
HP:0003765	Psoriasis	0,0235
HP:0030060	Neoplasias en tejido nervioso	0,0235
HP:0030061	Neoplasias neuroectodérmicas	0,0235
HP:0030063	Neoplasias neuroepiteliales	0,0235
Elementos compartidos (18 genes)		
HP:0000708	Alteración del comportamiento	$2,278e^{04}$
HP:0002071	Alteración de la función motora extrapiramidal	0,0349

4.3.5. La integración funcional y fenotípica fomenta la identificación de genes concurrentes con los relacionados con aminas en procesos patológicos

La información fenotípica inferida a partir de las relaciones gen-enfermedad y las anotaciones de enfermedades a HPO permiten integrar estos genes en la red global de similitud fenotípica gen-gen. En PhenUMA, solo se obtienen relaciones para algunos genes implicados en el metabolismo de la dopamina (DRD2, TH y DDC), en el caso de las poliaminas, solo hay información para los genes SAT1, SMS y ARG1 y ninguna relación para los genes asociados con la histamina, como indica la Tabla 4.10. No obstante, a partir de las relaciones obtenidas de la literatura y de los perfiles fenotípicos asignados a dichos genes se obtuvo un número significativo de relaciones no observadas anteriormente. En total, se obtuvieron 7.754 relaciones de similitud fenotípica (por encima del 98º percentil) entre 1.149 genes en total, considerando los 53 genes relacionados con el metabolismo de las aminas y otros genes anotados en HPO para los que se obtuvo alguna relación.

Tabla 4.10: Comparativa del número de relaciones obtenidas para cada grupo de aminas usando los datos de la literatura y las relaciones incluidas en PhenUMA (a partir de las relaciones en OMIM).

	Nuevas relaciones		**PhenUMA**	
	# Relaciones	**# Genes**	**# Relaciones**	**# Genes**
Poliaminas	1.752	554	68	68
Histamina	592	165	0	0
Dopamina	4.072	631	48	35
Compartidos	1.792	409	28	18

Esta aproximación muestra que usando la información presente en la literatura es posible obtener un número mayor de genes al obtenido usando únicamente las relaciones incluidas en las bases de datos de enfermedades. No obstante, estas relaciones deberían ser debidamente analizadas, cribadas y validadas por investigadores y/o clínicos.

A partir de la información fenotípica es posible obtener un conjunto relativamente grande de genes asociados a los procesos patológicos en los que están implicadas las aminas biogénicas. Por ello, para profundizar algo más en estas relaciones, se realizó una exploración de la información funcional subyacente a estas relaciones fenotípicas, correspondiente al último paso del procedimiento propuesto (Material y Métodos, sección 3.6). Para integrar la información fenotípica con información funcional, se buscaron aquellos

pares de genes relacionados fenotípicamente en interactomas como STRING y en relaciones de similitud funcional a partir de GO. El resultado de esta integración se muestra en la Figura 4.12, que incluye las relaciones funcionales entre genes que también se relacionan fenotípicamente. Con respecto a las relaciones de similitud funcional (GO), se destaca la importancia de la rama de procesos biológicos, representadas en color azul.

La Figura 4.12 muestra las relaciones funcionales que existen entre las distintas aminas, que son coherentes con el solapamiento fenotípico entre ellas que se muestra en la Figura 4.10. Un ejemplo de este hecho es la densidad de relaciones que se observa entre los diferentes grupos de genes. Sobre todo entre la dopamina/serotonina y el conjunto genes implicados con el metabolismo varias aminas (los elementos compartidos mencionados en la Tabla 4.8), así como, algunos genes relacionados con la histamina.

En la Figura 4.12 también se observa la aparición de genes, que no estaban inicialmente introducidos en el estudio, pero que comparten información fenotípica y funcional con los genes asociados a distintas aminas. Dichos genes pueden considerarse genes candidatos a estar asociados con las mismas patologías que los genes del metabolismo de aminas y quizá cooperar con ellos en los mismos escenarios fisiopatológicos. Sobre todo, hay que destacar algunos genes asociados a la histamina y a la dopamina que comparten relación patológica y funcional. Un ejemplo es el gen EDNRA, que es un receptor de endotelina y en el que algunas variantes se han relacionado con migraña (Joshi et al., 2011). El factor de necrosis tumoral (TFN) aparece también relacionado con genes asociados con la dopamina e histamina, ya que está directamente relacionado con algunas patologías neurológicas como la demencia, Alzheimer (McCusker et al., 2001), la migraña (Asuni et al., 2009), y también se asocia con la susceptibilidad al asma (Berry et al., 2006), características clínicas relacionadas con ambas aminas (véase Figura 4.10 o Tabla 4.8). Con respecto a las relaciones funcionales, cabe mencionar que TNF se asocia a la regulación negativa de aminas (GO:0051953), y también con procesos biológicos relacionados con HTR1B, HTR2A, DRD4, DRD3 y el receptor de histamina HRH3, lo que sugiere que podría ser un elemento regulador importante en la coordinación de las vías metabólicas y funciones fisiopatológicas de dopamina, serotonina e histamina.

Otro de los genes asociados con varias aminas es SNCA, que codifica la proteína alpha-sinucleina, expresada sobre todo en las neuronas, especialmente en terminales presinápticos. Este gen se asocia con enfermedades neurodegenerativas como el Parkinson o demencia con cuerpos de Lewy (Chartier-Harlin et al., 2001; Golbe and Mouradian, 2004; Ibáñez et al., 1998), de ahí su relación con genes asociados con la dopamina (DRD2) e histamina (HRH3). SNCA también está relacionado con procesos biológicos como la regulación negativa del transporte de aminas (GO:0015874) y el transporte de serotonina

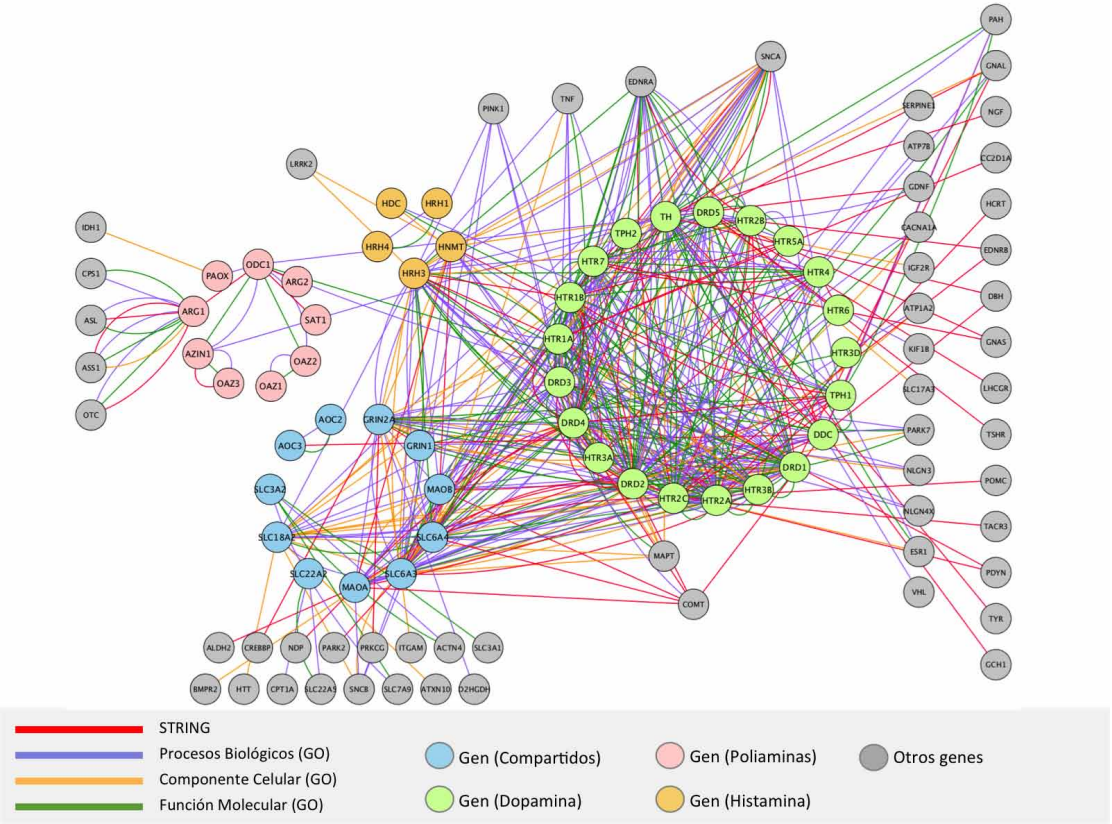


Figura 4.12: Integración funcional de las relaciones fenotípicas entre los genes relacionados con aminas y el resto de genes. Esta figura incluye las relaciones funcionales entre cada par de genes, en los que además existe una relación de similitud fenotípica, es decir, solo se han incluido las relaciones funcionales entre genes que además se asocian con un conjunto de síntomas similar (a partir de la información procedente de la literatura). La información funcional procede de STRING (enlaces rojos), y de las relaciones proporcionadas por GO teniendo cuenta las diferentes ramas: procesos biológicos (púrpura), componente celular (morado) y función molecular (verde). Como puede verse en la leyenda, el color de los nodos indica el grupo al que pertenecen.

(GO:0000683), lo que explica su relación funcional con genes relacionados con la histamina y la dopamina. SNCA presenta una relación fuerte con TH, tanto en GO (las tres sub-ontologías) como en STRING, esto es debido a que SNCA también está relacionado con la síntesis de la dopamina en GO; además, TH y SNCA se expresan en los axones de las neuronas y existen evidencias experimentales de que interaccionan físicamente (Perez et al., 2002).

Por último, se destaca la aparición en la red de PINK1 que se relaciona principalmente con el Parkinson (Pickrell and Youle, 2015), enfermedad relacionada con varios genes de nuestro estudio, sobre todo con genes asociados la dopamina e histamina. Funcionalmente PINK1 está relacionado con los receptores de dopamina, ya que participa en la regulación de la transmisión sináptica (GO:0032225), la secreción de dopamina (GO:0014046) y la regulación positiva del transporte de aminas (GO:00 51954).

En definitiva, en este trabajo se presentó un proyecto piloto, cuyo objetivo era obtener información biomédica asociada con aminas biogénicas, a partir del flujo de trabajo que combina el análisis de la literatura con herramientas de minería de textos y la integración de información fenotípica y funcional. Algunos de los resultados obtenidos en este trabajo son esperados y forman parte del conocimiento actual sobre las implicaciones patológicas de las aminas biogénicas, lo cual valida la aproximación propuesta.

Además, el principal objetivo de este análisis era demostrar precisamente que relaciones presentes en la literatura entre genes y enfermedades quedan fuera de las bases de datos, posiblemente debido a la exclusión expresa de relaciones que no tengan base genética conocida. Mediante la aproximación propuesta es posible el afloramiento de nuevas relaciones que permiten avanzar en la caracterización de la complejidad de las enfermedades humanas. Es necesario tener en cuenta que las causas de la pérdida de homeostasis, que supone la aparición o evolución de una enfermedad, en general, es la consecuencia de la alteración de una o varias funciones biológicas. En ocasiones, hay múltiples causas no genéticas que influyen en la expresión o los niveles de actividad de una proteína. Por tanto, esta información funcional entre productos génicos puede suponer una aproximación para explicar mejor los problemas biológicos complejos implicados en diferentes procesos patológicos.

4.4. Estudio de las relaciones genotipo-fenotipo entre pacientes

En los puntos anteriores se analizaron las relaciones entre genes y enfermedades. El uso de perfiles fenotípicos hace posible que emerjan nuevas relaciones entre genes. Además, su integración con información funcional permite estudiar el contexto molecular de estos procesos patológicos. Sin embargo, para aumentar la precisión de estas relaciones, en este apartado, se propone el uso de perfiles fenotípicos detallados a nivel de individuos, de los cuales se tiene acceso a su información genética. De esta forma, es posible indagar en qué variaciones genéticas podrían estar asociadas a ciertas características clínicas. Para ello, se analizó la información genotípica y fenotípica de miles de pacientes. Los pacientes estudiados en este análisis proceden de la base de datos DECIPHER (Material y Métodos 3.2.3). Este repositorio proporciona variaciones en el número de copias o CNV (deleciones y duplicaciones) y el conjunto de síntomas asociados a cada paciente (utilizando términos HPO) (Firth et al., 2009). En total, los datos almacenados en DECIPHER suman más de 45.000 pacientes (marzo de 2015), de los cuales más de 10.000 han dado su consentimiento para compartir sus datos médicos (Firth et al., 2009).

En este estudio (Reyes-Palomares et al., 2016), se analizó un subconjunto de los pacientes registrados en DECIPHER, que constaba de 6.564 individuos en los que se han detectado un total de 9.186 CNV. Estos pacientes presentan características clínicas y patologías muy heterogéneas, en lo cuales, se observan síntomas relacionados con retrasos en el desarrollo, discapacidad intelectual y malformaciones congénitas.

Para explorar sistemáticamente las relaciones entre las regiones genómicas alteradas y los fenotipos más representativos de los pacientes, se analizaron las relaciones genotípicas y fenotípicas entre ellos. Además, se aplicaron enriquecimientos fenotípicos y estudios de asociación genética. Gracias a este análisis, fue posible identificar grupos de pacientes asociados a trastornos genómicos conocidos, así como destacar potenciales nuevos síndromes no catalogados aún en bases de datos como ClinVar o DECIPHER.

4.4.1. Resumen de los perfiles fenotípicos y del tamaño de las regiones procedentes de DECIPHER

En las primeras fases del análisis se realizó una exploración general de los datos de las CNV y de los perfiles fenotípicos (Figura 4.13). En esta tesis, se usaron CNV asociadas a pacientes (DECIPHER) e individuos aparentemente sanos, que se usaron como grupo

control, y que fueron obtenidos del repositorio DGV (Material y Métodos apartado 3.2.4). La Tabla 4.11 muestra un resumen de las CNV que se incluyeron en este trabajo.

Tabla 4.11: Resumen de los datos de las mutaciones analizadas en este trabajo. Las CNV proceden de DECIPHER (pacientes) y DGV (controles).

	Todos	**Pacientes**	**Controles**
Número de muestras	10.324	6.564	5.072
CNV	14.226	9.186	495.916
Tipo de CNV:			
Deleciones	7.554	5.101	343.489
Duplicaciones	6.672	4.085	152.427
Longitud CNV (Kb)	3.336	3.014	31
Tipo de herencia			
De novo	14.501	2.454	
Heredadas de padres sin fenotipo	9.345	1.945	
Heredadas de padres con fenotipo similar	1.345	240	
Desconocido	21.946	3.638	

En relación a las características clínicas (Figura 4.13A), se observó que perfiles con un menor número de fenotipos, oscilando entre 1 y 5 fenotipos, eran notablemente más frecuentes que pacientes con un perfil fenotípico más variado o heterogéneo. Esto podría deberse parcialmente a sesgos que se producen debido al interés específico del personal clínico e investigador en los estudios de patologías concretas y caracterizadas por un conjunto de síntomas.

Como se muestra en la Tabla 4.11, se consideraron CNV con distintos tipos de herencia. Para comprobar cómo afecta el tipo de herencia al tamaño de los perfiles fenotípicos, se compararon los perfiles asociados con pacientes que presentan mutaciones *de novo* y mutaciones heredadas de sus progenitores. En la Figura 4.13B se aprecia que las mutaciones *de novo* se relacionan con perfiles fenotípicos de mayor tamaño. Aplicando el test de Mann-Whitney U se calculó la significación estadística de la diferencia entre ambas distribuciones, obteniendo un p-valor de $9,63e^{-52}$, por lo tanto, la diferencia entre las distribuciones es claramente significativa.

La longitud de las regiones mutadas también difiere en función del tipo de herencia. La Figura 4.13D muestra que las mutaciones *de novo* tienen una longitud media mayor que aquellas mutaciones que son heredadas de padres que no muestran síntomas de enfermedad. Estos resultados, junto con los observados en la Figura 4.13B, manifiestan cierta relación entre el tamaño de la región mutada y el número de fenotipos observados.

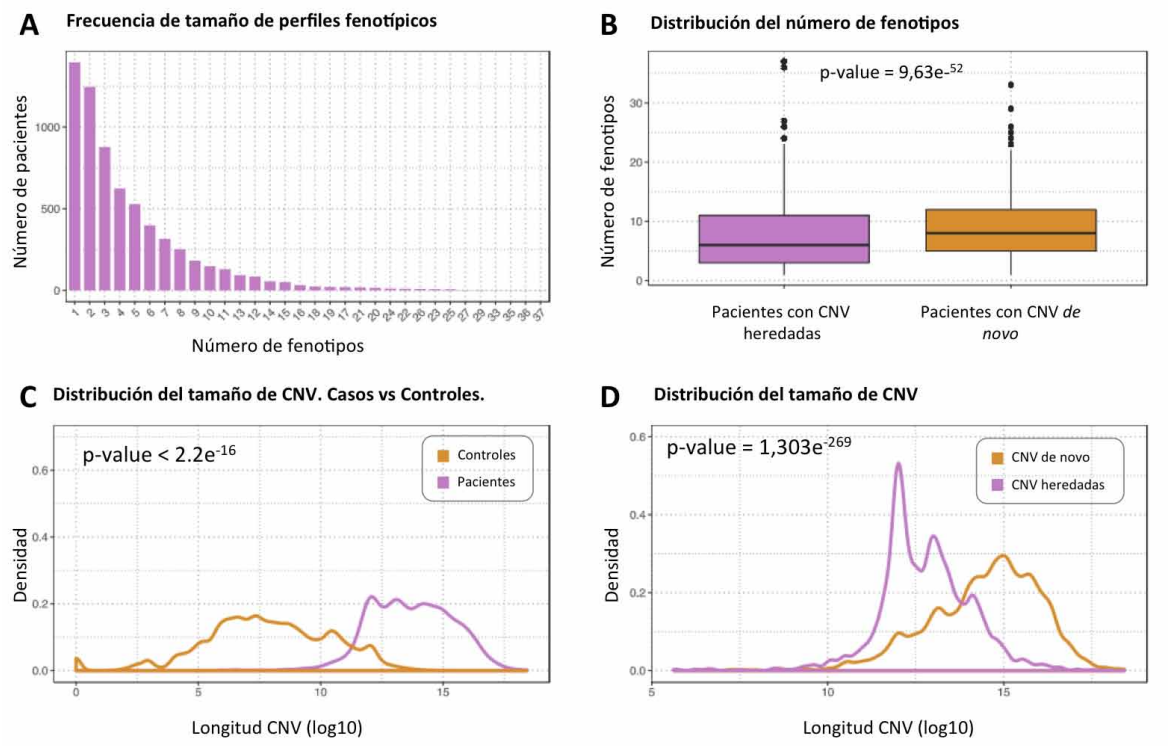


Figura 4.13: Gráficas de la información fenotípica y genética de DECIPHER. A. Frecuencia del número de fenotipos observados en pacientes de DECIPHER. **B.** Comparación de la distribución del tamaño de los perfiles fenotípicos de los pacientes con CNV heredadas o *de novo*. **C.** Comparación de las distribuciones del tamaño de las CNV de pacientes con las CNV observadas en los controles. **D.** Comparación de la longitud de las CNV *de novo* y heredadas.

Para profundizar más en dicha observación, se compararon ambas variables (el número de fenotipos y la longitud de mutaciones) para cada paciente. Puesto que un paciente puede tener más de una CNV, en estos casos, se calculó la longitud en función de la suma de la longitud de todas las CNV detectadas. Por lo tanto, esta aproximación permite considerar el tamaño total de la región genómica mutada en cada paciente. La Figura 4.14 muestra la distribución del número de características clínicas observadas en los pacientes en función de la longitud total de sus regiones mutadas y se puede apreciar un correlación positiva y significativa (coeficiente de correlación de Pearson es de 0,48 y un p-valor $8{,}46e^{-11}$). Dicha relación podría indicar una relación entre la complejidad clínica de los pacientes, determinada por el número de síntomas, y la extensión de las regiones mutadas.

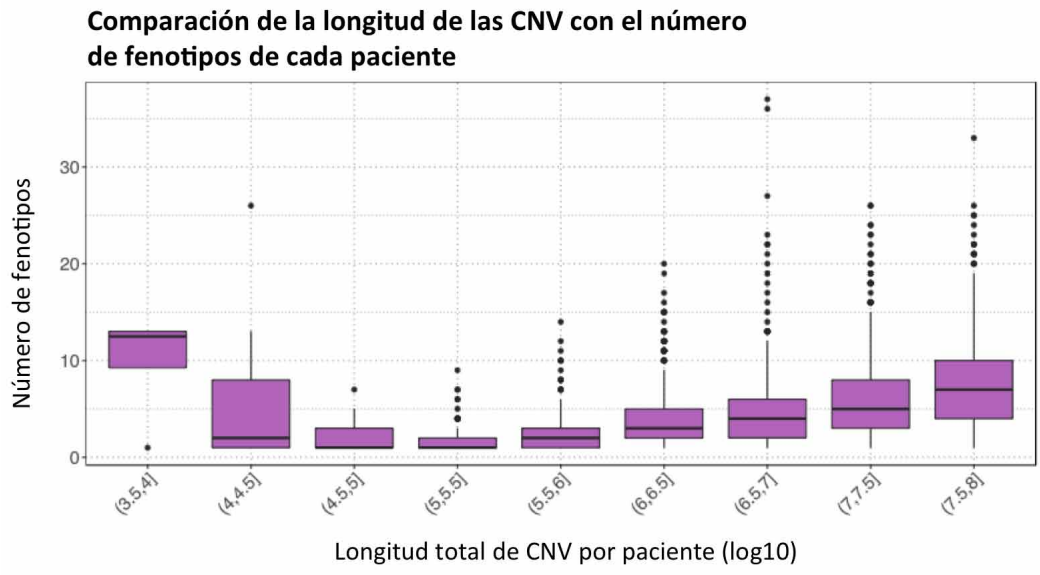


Figura 4.14: Comparación de número de fenotipos y el tamaño de las regiones de cada paciente. En el eje X se incluye el logaritmo en base 10 de la longitud de la región total mutada en cada paciente. La longitud total de las CNV es la suma de las longitudes de las CNV detectadas en un paciente. El eje Y representa al número de fenotipos asociados a cada paciente. El resultado muestra una tendencia a un mayor número de fenotipos en pacientes con regiones mutadas más extensas.

Las diferencias en el tamaño de las CVN también son notables al comparar las CNV de los pacientes con las de los individuos sanos (controles). Al comparar la distribución (Figura 4.14) de la longitud de las CNV de ambos grupos de individuos, se observó que los pacientes presentan CNV de mayor tamaño que los controles. Este resultado no es

Tabla 4.12: Parámetros topológicos de la red de pacientes.

Propiedad topológica	Valor
Número de nodos	6.304
Número de relaciones	89.526
Coeficiente de agrupamiento	0,801
Componentes conectadas	5
Diámetro de la red	10
Número de caminos mínimos	39.482.458
Camino mínimo promedio	3,706
Grado medio (degree)	28,403
Densidad de la red	0,005

sorprendente, ya que las mutaciones que se saben que son patogénicas suelen afectar a regiones cromosómicas de gran tamaño (Nygaard et al., 2016).

4.4.2. Relaciones entre pacientes a partir de solapamiento entre CNV

Para analizar grupos de pacientes cuyas CNV solapan en una misma región se construyó una red de pacientes. En esta red, dos pacientes estarán relacionados si sus CNV solapan en al menos un par de bases (ver Material y Métodos, apartado 3.7.1). En total, la red se componía de 6.342 nodos y 89.526 relaciones. La Tabla 4.12 muestra el resumen de algunos parámetros topológicos que corresponden con los característicos de las *large real-world networks* y se alejan de las propiedades observadas en redes aleatorias.

Una de las propiedades destacables es el coeficiente de agrupamiento (C), que como su propio nombre indica, proporciona un valor numérico que muestra cómo están agrupados los nodos en la red. El cálculo de C es el resultado del cociente entre el número de caminos cerrados de longitud 2 y el número total de caminos de longitud 2 (cerrados y no cerrados). Una red completamente conectada tendrá un coeficiente de agrupamiento máximo, es decir, C=1. El mínimo (C=0) es propio de una red en la que sus nodos están muy dispersos (Newman, 2010).

En la red de pacientes este valor es de 0,8; es decir, el 80 % de los caminos que involucran a tres pacientes son cerrados. Por lo tanto, el coeficiente de agrupamiento es considerablemente alto en esta red e indica que la red se compone de grupos de pacientes altamente interconectados. Además, puede inferirse un alto grado de solapamiento entre las CNV de los distintos pacientes que componen la red.

4.4.3. Obtención de los *loci* fenotípicamente enriquecidos (PEL) a partir de la red de pacientes

Puesto que la red se compone de nodos muy interconectados (alto nivel de agrupamiento), se procedió a la identificación de los grupos de pacientes cuyas regiones mutadas solapan en un mismo *loci*. Para asegurar que todos los pacientes, de los grupos identificados, solapen entre sí se usó el concepto de *clique*. Un *clique* es una red (o subred) en la que todos sus nodos están conectados entre sí, es decir, conforman un grafo completo. Como se especifica en el capítulo de Material y Métodos (sección 3.6), se identificaron los *cliques*, usando el paquete NetworkX, seleccionado aquellos con un mínimo de 3 pacientes y, además, filtrando solo aquellos en los que todos los pacientes se asocian con CNV solapantes en una misma región.

Cada uno de los *cliques* representa una región afectada por varios pacientes. A continuación se usó la información fenotípica de estos pacientes para calcular el enriquecimiento fenotípico de cada *clique*, o lo que es lo mismo, el enriquecimiento fenotípico de la región afectada o *loci*. De esta forma es posible obtener los fenotipos más representativos de los pacientes con CNV solapantes. Como se explicó en Material y Métodos (sección 3.5), se usó un test hipergeométrico para calcular el enriquecimiento de los *loci*, seleccionando aquellos cuyo p-valor es menor 0,05. Además, los fenotipos redundantes y poco informativos fueron eliminados atendiendo a la estructura jerárquica de la ontología. A cada uno de los *loci*, para los que se obtuvo un conjunto representativo de fenotipos, se le denominó *loci* fenotípicamente enriquecido o *phenotypically enriched loci* (PEL).

El resultado consta de 1.042 relaciones *loci*-fenotipo entre 487 PEL y 195 términos HPO. Para valorar si los PEL son regiones mutadas con mayor frecuencia en la población de pacientes respecto a la población control se utilizó un procedimiento similar al descrito en Cooper et al. (2011). A partir de este método se identificaron 387 relaciones específicas *loci*-fenotipo entre 336 PEL y 115 fenotipos diferentes. Casi un 70 % de los PEL (336 de 487) representan mutaciones observadas con más frecuencia en el grupo de pacientes que en los controles (p-valor <0,05, tet de Fisher), por lo tanto, se consideraron PEL potencialmente patológicos.

4.4.4. Validación de los PEL potencialmente patogénicos a partir de modelos aleatorios

Aunque el porcentaje de regiones potencialmente patogénicas es alto, este resultado es esperable si tenemos en cuenta la finalidad de DECIPHER. El objetivo de DECIPHER es

recopilar datos de pacientes y sus CNV candidatas a ser patogénicas. Por lo tanto, era razonable detectar un porcentaje alto de *loci* candidatos a estar asociados a enfermedad. Sin embargo, para determinar con mayor fiabilidad la patogenicidad de las regiones identificadas, y descartar que se deba al azar, se diseñaron varios modelos aleatorios para comparar los resultados reales con distintos tipos de aleatorizaciones. La descripción de los modelos aleatorios se incluyó en el capítulo Material y Métodos (sección 3.7.5), no obstante, a modo de recordatorio, indicar que dicho análisis consiste en la identificación de los PEL:

- Usando, como casos, muestras de CNV aleatorias presentes en los controles (DGV).
- Usando, como casos, regiones aleatorias del genoma, manteniendo la longitud de las CNV y la frecuencias por cromosoma de las mutaciones de los pacientes.
- Usando regiones aleatorias para los controles, manteniendo la longitud de las CNV y la frecuencia por cromosoma.
- Usando relaciones paciente-CNV aleatorias.
- A partir de relaciones paciente-fenotipo aleatorias, manteniendo la frecuencia de los fenotipos.

Para todas las aleatorizaciones, se identificó un número de PEL menor que a partir de datos reales. Además, la significación estadística (p-valor <0,05, test de Fisher), derivada de los estudios de asociación genética, fue mayor para los datos reales que para los datos aleatorios (Figura 4.15). Estos resultados ponen de manifiesto la existencia de una fracción de PEL consistentemente patogénicos.

4.4.5. Identificación de regiones conocidas asociadas a enfermedades a partir de los datos de DECIPHER

Para profundizar en la robustez del método, se estudió su capacidad de identificar regiones patológicas conocidas. Con este objetivo, se comprobó cuántas de las regiones identificadas, incluidas en algún PEL, aparecen relacionadas con alguna patología en ClinVar (Landrum et al., 2014) o DECIPHER. Como se explicó en el capítulo Material y Métodos (apartado 3.2.5), ClinVar es una base de datos pública que incluye relaciones entre variaciones y enfermedades genéticas. Además, ClinVar incorpora el sentido clínico (*Clinical Significance*) de cada par variación-enfermedad: *pathogenic*, *likely pathogenic*, *benign*, *likely benign*, etc. Para comprobar cuantos PEL están presentes en ClinVar se seleccionaron

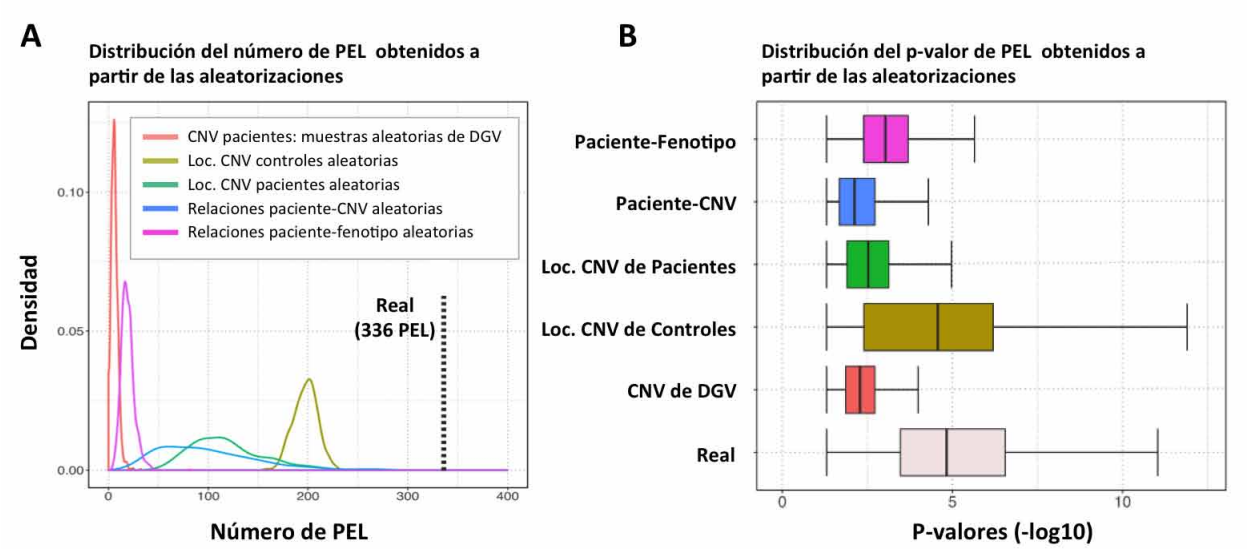


Figura 4.15: Número de PEL identificados a partir de datos aleatorios. A. Distribución del número de PEL obtenidos a partir de los datos aleatorios siguiendo las diferentes estrategias, cada una representada con un color. **B.** Diagrama de cajas que representa las distribuciones de los p-valores, asociados a cada uno de los PEL, obtenidos a partir de los datos reales y las aleatorizaciones.

las 2.243 CNV patológicas (*pathogenic*) o probablemente patológicas (*likely pathogenic*) asociadas con enfermedades OMIM y otras 75 regiones genómicas asociadas con síndromes de DECIPHER. El resultado fue que los PEL identificados solapaban con un total de 93 y 15 regiones clasificadas como patogénicas y asociadas a síndromes descritos en ClinVar y DECIPHER, respectivamente. En la Tabla 4.13 se incluye una lista de PEL cuyas regiones solapan con alguna enfermedad.

Para demostrar la validez de estos resultados, se comprobó si la identificación de regiones patogénicas conocidas era similar para los datos aleatorios. El resultado indica que somos capaces de recapitular un mayor número de regiones asociadas a síndromes a partir de los datos reales, en comparación con los obtenidos a partir de las aleatorizaciones, con la excepción del modelo aleatorio basado en la aleatorización de las CNV en los controles (Figura 4.16).

DECIPHER y ClinVar son bases de datos extensas y contienen información sobre un gran número de patologías. Por ello, se esperaba obtener un mayor número de síndromes con regiones solapantes con los PEL identificados. Sin embargo, los pacientes de DECIPHER, en ocasiones, están previamente diagnosticados con enfermedades raras y poco caracterizadas que podrían no estar aún presentes en bases de datos como ClinVar.

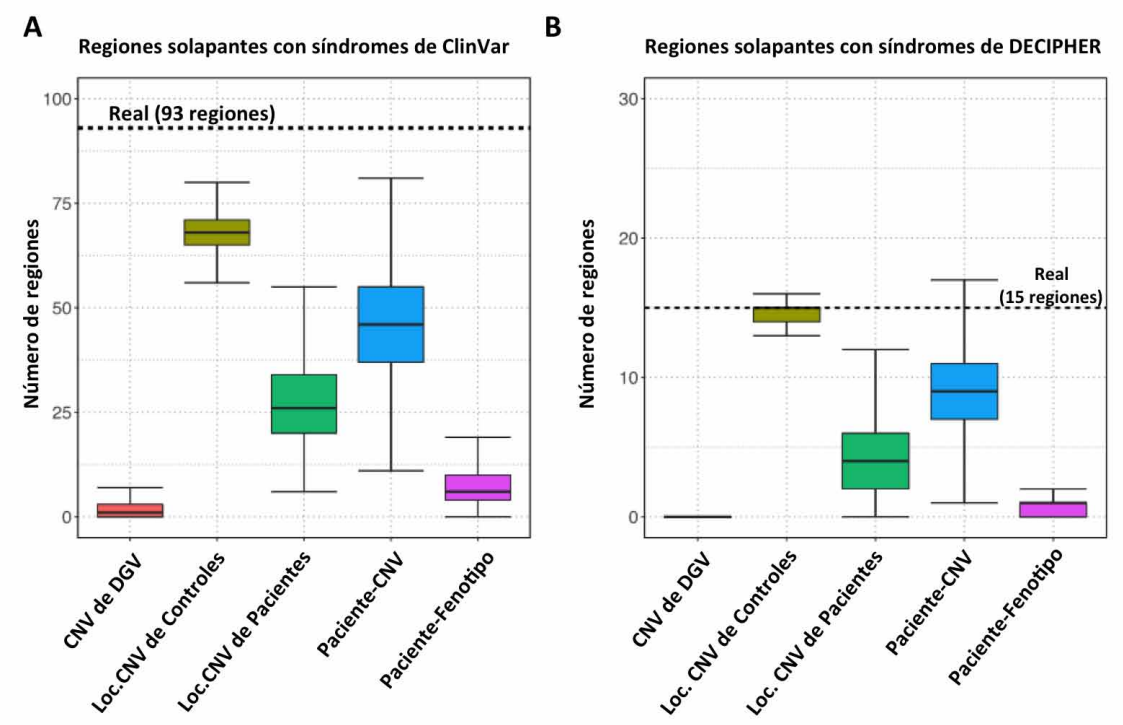


Figura 4.16: Comparación del número de regiones genómicas asociadas a síndromes en ClinVar y DECIPHER. Los diagramas de cajas de la figura representan la distribución del número de síndromes (en ClinVar y DECIPHER, respectivamente) obtenidos por los datos reales y usando cada uno de los modelos aleatorios.

Tabla 4.13: *Loci* fenotípicamente enriquecidos (PEL) cuyas regiones, todas ellas deleciones, coinciden con variaciones genéticas asociadas a alguna enfermedad conocida (MIM).

PEL ID	Chr	Inicio	Long(Kb)	Fenotipos	Casos/ Portadores	p-valor	OMIM
240	1	243981716	12.547	Anomalía en el cráneo	13/18 (0)	$4,50e^{08}$	217990
193	1	243786018	126.15	Anomalía en el cráneo	14/19 (0)	$1,30e^{08}$	217990
68	1	243981716	12.547	Microcefalia	12/18 (0)	$7,40e^{11}$	217990
49	1	243786018	126.15	Microcefalia	13/19 (0)	$9,20e^{12}$	217990
25	1	243981716	12.547	Aplasia/hipoplasia del cerebro	15/18 (0)	$1,80e^{12}$	217990
5	1	243786018	126.15	Aplasia/hipoplasia del cerebro	16/19 (0)	$3,80e^{13}$	217990
70	11	31802605	23.093	Aplasia/hipoplasia ocular	5/8 (0)	$1,00e^{08}$	106210
317	14	55242483	200.932	Anomalías del ojo	6/6 (0)	$1,80e^{04}$	248000
295	4	82082415	31.542	Crecimiento anormal	9/11 (0)	$4,20e^{06}$	601665
484	6	407031	170.484	Anomalías en la región ocular	10/16 (-1)	$4,10e^{06}$	145400
484	6	407031	170.484	Anomalías en la región ocular	10/16 (-1)	$4,10e^{06}$	187350
347	6	1612710	15.026	Anomalías en la región ocular	11/17 (0)	$1,60e^{07}$	145400
347	6	1612710	15.026	Anomalías en la región ocular	11/17 (0)	$1,60e^{07}$	187350
56	6	407031	170.484	Anomalías de la ubicación del globo ocular	9/16 (-1)	$7,90e^{08}$	145400
100	6	2371534	63.584	Hipertelorismo	8/13 (-1)	$2,80e^{08}$	145400
88	6	1612710	357.639	Hipertelorismo	9/16 (-1)	$3,00e^{09}$	145400
58	6	1612710	22.698	Hipertelorismo	10/17 (0)	$3,40e^{11}$	145400
22	8	11610366	83.076	Malformación del corazón y grandes vasos	15/21 (0)	$1,00e^{13}$	265500
6	8	11610366	83.076	Anomalías en el sistema cardiovascular	18/21 (0)	$8,00e^{15}$	265500
7	8	11610366	83.076	Anomalías de morfología cardíaca	17/21 (0)	$6,40e^{15}$	265500
452	X	102585912	9.472	Anomalías de la mano	6/8 (0)	$3,50e^{05}$	108110

El solapamiento entre las CNV de los pacientes y las regiones asociadas con síndromes en ClinVar y DECIPHER, se usaron para visualizar la red de pacientes. De esta forma, se proporciona una representación de las relaciones genéticas entre pacientes y su relación con síndromes en ambas bases de datos (Figura 4.17). Esta red muestra algunos de los *cliques* identificados (como se especificó en Material y Métodos 3.7.2) por el método propuesto. Al analizar la red se identificaron *cliques* de pacientes relacionados entre sí, pero no todos ellos tenían afectada una región asociada con alguna patología en ClinVar o DECIPHER. Esto puede observarse a claramente en la Figura 4.17A. Algunos de los grupos de nodos están formados por nodos de dos colores (morado y naranja). Concretamente, 164 (50 %) de las CNV de los PEL identificados solapan con alguna región catalogada como patológica en ClinVar, lo que indica que dichos PEL están potencialmente relacionados con trastornos genómicos conocidos.

Un ejemplo es el PEL 22, formado por pacientes cuyas regiones delecionadas solapan en 8p23.1. Esta región, coincide con la región relacionada con estenosis pulmonar (OMIM 265500) en ClinVar. En este caso particular, 15 de los 21 pacientes con deleciones en ese *locus* están anotados con "malformación del corazón y grandes vasos"(HP:0002564, p-valor $8,3e$ 10), que es la primera causa de estenosis pulmonar. Además, no hay ningún

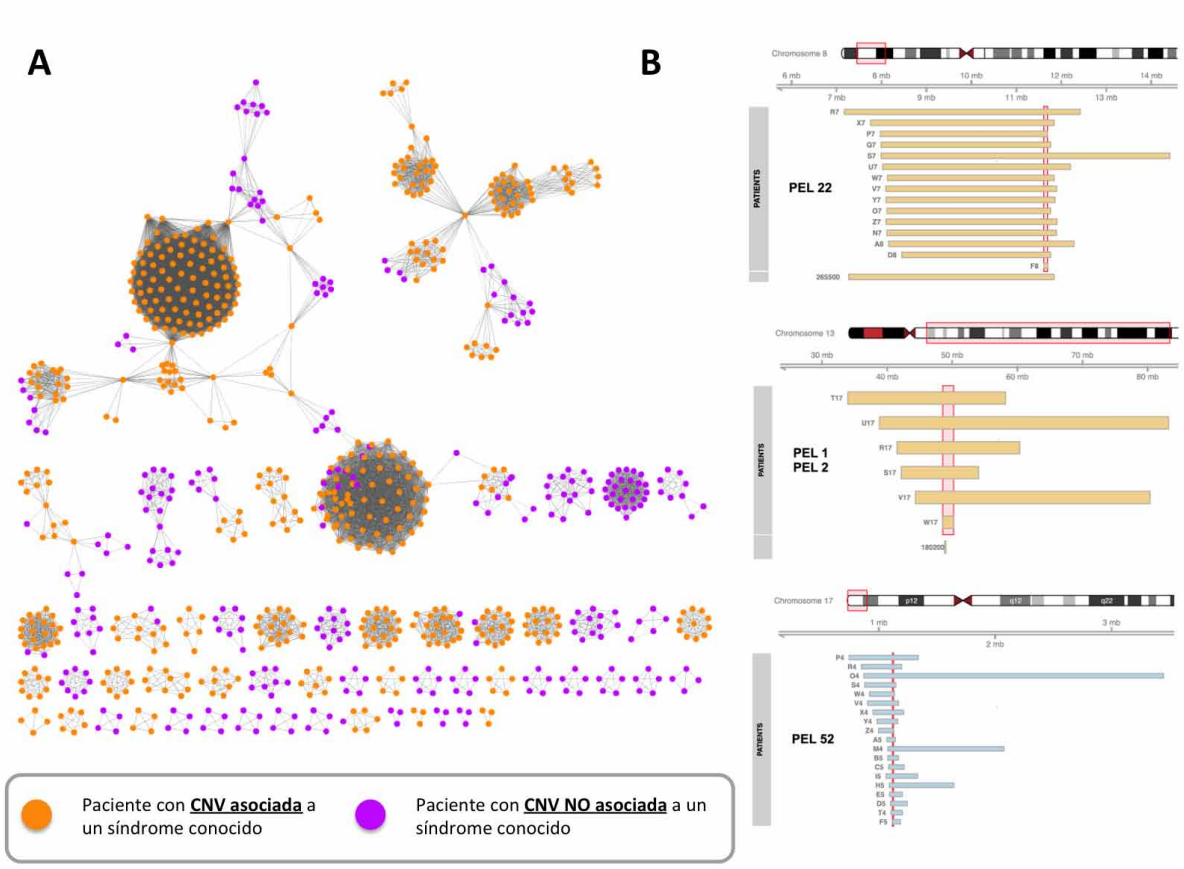


Figura 4.17: A. Red de pacientes relacionados genéticamente. Esta red se compone solo de los pacientes cuya región alterada está incluida en algún PEL, es decir, que está fenotípicamente enriquecida. En naranja se muestran los pacientes cuya CNV solapa con algún síndrome conocido en ClinVar o DECIPHER. En morado, se representan los pacientes que no solapan con ningún síndrome. **B.** Ejemplos de tres PEL. Los pacientes del PEL 22 solapan en regiones asociadas con estenosis pulmonar (OMIM 265500). Los PEL 1 y 2 se componen de pacientes cuyas regiones solapan con el síndrome de deleción en 13q14. Por último, el PEL 52 no se asocia con ningún síndrome conocido y se compone de pacientes que manifiestan mano hendida (HP:0001171) y duplicaciones en 17p13.3.

individuo en el conjunto de los controles que muestre una deleción en ese *locus*, lo que sugiere una alta penetrancia del fenotipo asociado a esa región (Tabla 4.12).

Otro ejemplo es el retinoblastoma (HP:0009919, p-valor $6{,}7e^{-16}$ y $3{,}7e^{-15}$ para PEL1 y 2) asociado a 7 de los pacientes analizados. De estos 7 pacientes, 6 pertenecen al mismo PEL, lo que indica que las CNV de estos pacientes son solapantes. Estas CNV son deleciones en 13q14.2 (chr13:48.544.437-50.206.474, véase la figura 4.17B). Las variaciones estructurales en esa región se asocian con el síndrome de deleción en 13q14, cuya característica clínica más representativa es el retinoblastoma (OMIM 180200) (Friend et al., 1986; Sparkes et al., 1980). Sin embargo, las deleciones en este *locus* son frecuentes en la población control, lo que sugiere una penetrancia reducida para este fenotipo (Mitter et al., 2011). Estos resultados indican que nuestro método es capaz de identificar y priorizar variantes estructurales que están fuertemente asociadas con fenotipos patológicos.

Tabla 4.14: Regiones fenotípicamente enriquecidas no asociadas con ningún síndrome conocido.

PEL ID	Tipo	Chr	Inicio	Long(Kb)	Fenotipos	Casos/ Portadores	p-valor
PEL 3	d	3	181296306	175.931	Anoftalmia	6/9 (0)	$1{,}80e^{-15}$
PEL 5	d	7	95693340	89.973	Ectrodactilia	7/10 (0)	$1{,}70e^{-14}$
PEL 4	d	3	181296306	175.931	Anomalía del tamaño del globo ocular	8/9 (0)	$1{,}90e^{-14}$
PEL 4	d	3	181296306	175.931	Aplasia / hipoplasia ocular	8/9 (0)	$1{,}90e^{-13}$
PEL 71	d	2	200208169	38.268	Anomalía del paladar	11/19 (0)	$2{,}10e^{-11}$
PEL 31	d	3	181166306	576	Anomalía del tamaño del globo ocular	6/9 (0)	$4{,}10e^{-11}$
PEL 105	d	2	200208169	38.268	Anomalía de la cavidad oral	12/19 (0)	$2{,}50e^{-10}$
PEL 84	d	15	100019051	189.992	Retraso en el crecimiento	13/18 (0)	$2{,}70e^{-10}$
PEL 31	d	3	181166306	576	Aplasia/hipoplasia ocular	6/9 (0)	$2{,}70e^{-10}$
PEL 128	d	2	200208169	38.268	Anomalía de la boca	14/19 (0)	$1{,}90e^{-09}$
PEL 131	d	15	100019051	189.992	Anomalía del crecimiento	14/18 (0)	$6{,}40e^{-09}$
PEL 129	d	15	99057570	65.959	Retraso en el crecimiento	11/16 (0)	$8{,}10e^{-09}$
PEL 69	d	11	31735689	39.768	Aplasia/hipoplasia ocular	5/8 (0)	$1{,}00e^{-08}$
PEL 126	d	2	166091754	49.616	Convulsiones	10/15 (0)	$1{,}00e^{-08}$
PEL 175	d	7	112349829	160.71	Retraso del lenguaje y la comunicación	12/18 (0)	$1{,}00e^{-08}$
PEL 78	d	14	29904720	411.94	Aplasia/hipoplasia del cerebro	10/12 (0)	$1{,}50e^{-08}$
PEL 82	d	14	29904720	411.94	Aplasia/hipoplasia del cerebro	10/12 (0)	$1{,}50e^{-08}$
PEL 141	d	7	114297499	533.997	Retraso del lenguaje y la comunicación	11/15 (0)	$4{,}50e^{-08}$
PEL 166	d	2	166244769	311.476	Convulsiones	9/14 (0)	$6{,}00e^{-08}$
PEL 202	d	15	99057570	65.959	Anomalía del crecimiento	12/16 (0)	$8{,}80e^{-08}$
PEL 152	d	14	29904720	411.94	Microcefalia	8/12 (0)	$1{,}60e^{-07}$
PEL 159	d	14	29904720	411.94	Microcefalia	8/12 (0)	$1{,}60e^{-07}$
PEL 216	d	2	200208169	38.268	Anomalía del paladar	7/13 (0)	$1{,}60e^{-07}$
PEL 385	d	2	200246437	0	Anomalía de la cara	16/19 (0)	$1{,}60e^{-07}$
PEL 137	d	6	76509712	359.49	Laxitud articular	5/9 (0)	$1{,}60e^{-07}$
PEL 390	d	7	112349829	160.71	Retraso en el desarrollo neurológico	13/18 (0)	$1{,}60e^{-07}$
PEL 412	d	13	92065689	29.285	Retraso en el crecimiento	9/17 (0)	$2{,}00e^{-07}$
PEL 419	d	13	92065689	29.285	Retraso en el crecimiento	9/17 (0)	$2{,}00e^{-07}$
PEL 436	D	16	3831263	32.469	Anomalía de la cara	15/18 (0)	$4{,}20e^{-07}$
PEL 222	d	2	201936560	57.623	Anomalía de la boca	10/13 (0)	$4{,}80e^{-07}$

PEL ID	Tipo	Chr	Inicio	Long(Kb)	Fenotipos	Casos/ Portadores	p-valor
PEL 222	d	2	200208169	38.268	Anomalía de la boca	10/13 (0)	$4,80e^{07}$
PEL 242	d	7	114297499	533.997	Retraso en el desarrollo neurológico	12/15 (0)	$5,20e^{07}$
PEL 114	d	13	48557360	146.432	Anomalía del tamaño del globo ocular	8/9 (0)	$5,90e^{07}$
PEL 250	d	7	119973023	238.728	Retraso del lenguaje y la comunicación	9/13 (0)	$8,80e^{07}$
PEL 123	d	12	66224830	22.517	Talla baja	7/8 (0)	$1,10e^{06}$
PEL 184	d	1	28743173	21.263	Ojos hundidos	4/7 (0)	$1,90e^{06}$
PEL 462	d	1	11270844	47.828	Anomalías del cráneo	10/14 (0)	$1,90e^{06}$
PEL 417	d	2	201936560	57.623	Anomalía de la boca	9/13 (0)	$2,00e^{06}$
PEL 330	d	14	29781404	230.359	Convulsiones	7/12 (0)	$2,10e^{06}$
PEL 227	d	9	77206264	34.573	Convulsiones	7/10 (0)	$2,10e^{06}$
PEL 133	D	2	219965169	9.153	Sindactilia cutánea de dedo de la mano	3/5 (0)	$2,30e^{06}$
PEL 116	d	3	181296306	6.681	Anomalía de la región ocular	9/9 (0)	$2,50e^{06}$
PEL 173	D	2	59105866	181.965	Hipoplasia de la cara media	3/5 (0)	$4,00e^{06}$
PEL 120	D	2	59105866	181.965	Estrabismo	5/5 (0)	$4,40e^{06}$
PEL 276	d	7	94174003	41.631	Disminución del peso corporal	4/7 (0)	$7,20e^{06}$
PEL 329	d	7	95693340	89.973	Anomalía de la morfología ósea de las extremidades	8/10 (0)	$1,30e^{05}$
PEL 304	d	X	133530468	102.522	Retraso del desarrollo global	6/8 (0)	$1,50e^{05}$
PEL 388	d	10	28842276	86.821	Anomalía de los párpados	6/8 (0)	$1,70e^{05}$
PEL 210	D	2	59105866	181.965	Dificultades para alimentarse durante la lactancia	4/5 (0)	$1,70e^{05}$
PEL 455	d	10	28842276	86.821	Anomalía de las fisuras palpebrales	5/8 (0)	$2,00e^{05}$
PEL 251	d	13	41726952	39.763	Morfología anormal del ojo	6/7 (0)	$2,00e^{05}$
PEL 470	d	10	28842276	86.821	Anormalidad del cabello	5/8 (0)	$2,20e^{05}$
PEL 358	d	1	28743173	21.263	Anomalía de la ubicación del globo ocular	5/7 (0)	$2,90e^{05}$
PEL 460	d	3	181648378	93.928	Anomalía de la región ocular	7/9 (0)	$3,60e^{05}$
PEL 338	d	5	170676605	370.857	Anomalía de los tabiques cardíacos	4/6 (0)	$3,90e^{05}$
PEL 133	D	2	219965169	9.153	Sindactilia de dedos del pie	3/5 (0)	$4,00e^{05}$
PEL 454	d	1	177800358	362.172	Talla baja	5/7 (0)	$4,40e^{05}$
PEL 369	d	14	57423809	185.438	Anomalías en ojos	7/8 (0)	$4,50e^{05}$
PEL 449	d	7	94174003	41.631	Anomalía del pie	5/7 (0)	$4,90e^{05}$
PEL 294	d	1	157149743	12.346	Cantidad anormal de cabello	3/4 (0)	$5,70e^{05}$
PEL 213	d	1	157149743	12.346	Anomalía del labio	4/4 (0)	$5,70e^{05}$
PEL 327	d	19	10640379	140.937	Anormal morfología del sistema genital	4/5 (0)	$7,00e^{05}$
PEL 448	d	14	58205713	144.654	Anomalía del cráneo	7/8 (0)	$8,00e^{05}$
PEL 203	d	13	33963658	138.576	Anomalía del cuello	3/3 (0)	$8,20e^{05}$
PEL 423	d	3	181692255	50.051	Anomalía de la cara	9/9 (0)	$1,10^{e04}$
PEL 324	d	7	94953990	5.573	Crecimiento anormal	6/6 (0)	$2,20e^{04}$
PEL 456	D	2	219965169	9.153	Anomalía de la extremidad inferior	4/5 (0)	$5,20e^{04}$
PEL 361	D	7	106664270	182.398	Estrabismo	3/3 (0)	$5,40e^{04}$
PEL 404	D	7	107527586	136.426	Anomalía del movimiento del ojo	3/3 (0)	$8,20e^{04}$
PEL 407	D	1	113036203	122.933	Anomalía del paladar	3/3 (0)	$1,00e^{3}$

No obstante, más del 50 % (172 de 336) de los PEL patológicos no solapan con ningún síndrome. A pesar de ello, se identificaron varios PEL que, aunque no se asociaron a síndromes conocidos, sí que están fenotípicamente enriquecidos por fenotipos muy específicos tales como ectrodactilia, malformaciones en el corazón, defectos en el septo auricular y anoftalmía (Tabla 4.14). Estos PEL podrían representar a *loci* asociados con nuevos síndromes aún no catalogados en ClinVar. Por ejemplo, se detectó un grupo de pacientes que

muestran una condición médica grave que se conoce como mano hendida (HP: 0001171) con duplicaciones en la región 17p13.3. El PEL asociado con este grupo (PEL 52, p-valor de $1,1e^{-13}$) se compone de pacientes relacionados a una gran variedad de síntomas asociados con esta patología. La característica clínica denominada "anomalías de la mano" (HP: 0001155) fue el término HPO más enriquecido (p-valor = $2,7e^{-07}$ para PEL 52) asociado con este PEL. *A priori* este grupo de pacientes con solapamiento fenotípico y genético podrían estar relacionados con un nuevo síndrome. De hecho, al consultar la literatura se encuentran asociaciones de micro-duplicaciones en este *locus* con un fenotipo similar (Klopocki et al., 2012). Los pacientes de este PEL manifiestan otros síntomas pertenecientes a varias categorías: anomalía de la región ocular, anomalía de la morfología del hueso de la extremidad, anomalía del cráneo, anomalía de la cara, anomalía del cerebro, anomalía del sistema cardiovascular y retraso del crecimiento.

Estos resultados apoyan nuestra hipótesis de que este tipo de estrategias biocomputacionales ayudan a la caracterización y el estudio de las relaciones fenotipo-genotipo y de manera sistemática a partir de datos de individuos.

Capítulo 5

Conclusions

The next points summarize the conclusions of this thesis:

- Phenotype information of genetic diseases can be used to build phenotypic similarity gene networks (PSGN). The resulting network allows us the identification of novel relationships between disease-causing genes, which are useful to explore the molecular basis of pathological processes.
- A tool for the analysis of functional and phenotypic information was developed. PhenUMA has demonstrated to be a useful resource to study biomedical relationships among disease-causing genes, diseases and phenotypes. PhenUMA provides novel phenotypic and functional relationships among disease and genes that is not directly accessible in disease databases.
- Text-mining techniques were used to define phenotypic profiles of amine-related genes that do not have genetic associations with disease in databases. These techniques could be a complement for PhenUMA to detect literature-based relationships between gene and diseases that are not present in disease databases.
- Individual phenotypic and genotype information of a heterogeneous population of patients from DECIPHER database has been used to build a patient network. Further analyses on the resulting patient network provided significant relationships among symptoms and structural variations were unveiled. In addition, known genomic syndromes and putative novel syndromes were detected.

Capítulo 6

Conclusiones

Los siguientes puntos resumen las conclusiones de esta tesis:

- La información fenotípica asociada a enfermedades genéticas permitió la construcción de una red de similitud fenotípica entre genes (PSGN). Esta red contiene nuevas relaciones entre genes causantes de la enfermedad, las cuales fueron útiles para explorar las bases moleculares de los procesos patológicos.
- Se desarrolló una herramienta para el análisis de información funcional y fenotípica. PhenUMA demostró ser un recurso útil para estudiar las relaciones biomédicas entre genes y enfermedades.
- Herramientas de minería de textos fueron usadas para definir los perfiles fenotípicos de genes relacionados con aminas biogénicas. Estas técnicas podrían ser un complemento a PhenUMA para detectar las relaciones procedentes de la literatura, que no están presentes en bases de datos públicas.
- Se usó información fenotípica y genotípicas de individuos de una población heterogénea de pacientes, procedente de DECIPHER. Esta información se usó para construir una red de pacientes. El análisis de esta red permitió identificar relaciones significativas características clínicas y variaciones estructurales. Además,se detectaron potenciales nuevos síndromes.

Bibliografía

Aittokallio, T. y Schwikowski, B. (2006). Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3):243–255.

Arreola, R., Becerril-Villanueva, E., Cruz-Fuentes, C., Velasco-Velázquez, M. A., Garcés-Alvarez, M. E., Hurtado-Alvarado, G., Quintero-Fabian, S., y Pavón, L. (2015). Immunomodulatory Effects Mediated by Serotonin. *Journal of Immunology Research*, 2015:1–21.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., y Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Asuni, C., Stochino, M. E., Cherchi, A., Manchia, M., Congiu, D., Manconi, F., Squassina, A., Piccardi, M. P., y Zompo, M. (2009). Migraine and tumour necrosis factor gene polymorphism. *Journal of Neurology*, 256(2):194–197.

Auffray, C., Chen, Z., y Hood, L. (2009). Systems medicine: the future of medical genomics and healthcare. *Genome Medicine*, 1(1):2.

Baker, M. (2013). Big biology: The 'omes puzzle. *Nature*, 494(7438):416–419.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., y Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755.

Barabási, A.-L. (2007). Network Medicine — From Obesity to the "Diseasome". *New England Journal of Medicine*, 357(4):404–407.

Barabási, A.-L., Gulbahce, N., y Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., y Apweiler, R. (2009). The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(Database issue):D396–D403.

Bauer, S., Grossmann, S., Vingron, M., y Robinson, P. N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651.

Beaulieu, C., Majewski, J., Schwartzentruber, J., Samuels, M., Fernandez, B., Bernier, F., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., Adam, S., Bulman, D., Jones, S., Avard, D., Nguyen, M., Rousseau, F., Marshall, C., Wintle, R., Shen, Y., Scherer, S., Friedman, J., Michaud, J., Boycott, K., y Boycott, K. M. (2014). FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project. *The American Journal of Human Genetics*, 94(6):809–817.

Beaulieu, J.-M., Espinoza, S., y Gainetdinov, R. R. (2015). Dopamine receptors - IUPHAR Review 13. *British Journal of Pharmacology*, 172(1):1–23.

Becker, K. G., Barnes, K. C., Bright, T. J., y Wang, S. A. (2004). The Genetic Association Database. *Nature Genetics*, 36(5):431–432.

Benjamini, Y. y Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bennett, B. D. y Bushel, P. R. (2017). goSTAG: gene ontology subtrees to tag and annotate genes within a set. *Source Code for Biology and Medicine*, 12:6.

Berry, M. A., Hargadon, B., Shelley, M., Parker, D., Shaw, D. E., Green, R. H., Bradding, P., Brightling, C. E., Wardlaw, A. J., y Pavord, I. D. (2006). Evidence of a Role of Tumor Necrosis Factor in Refractory Asthma. *New England Journal of Medicine*, 354(7):697–708.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Bossi, A. y Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5:260.

Botstein, D. y Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33(3s):228–237.

Boycott, K. M., Vanstone, M. R., Bulman, D. E., y MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691.

Bravo, À., Cases, M., Queralt-Rosinach, N., Sanz, F., y Furlong, L. I. (2014). A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed Research International*, 2014:253128.

Brown, E. G., Wood, L., y Wood, S. (1999). The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*, 20(2):109–117.

Brown, S. H., Elkin, P. L., Bauer, B. A., Wahner-Roedler, D., Husser, C. S., Temesgen, Z., Hardenbrook, S. P., Fielstein, E. M., y Rosenbloom, S. T. (2006). SNOMED CT: utility for a general medical evaluation template. *AMIA Annu Symp Proc.*, 2006:101–105.

Burban, A., Faucard, R., Armand, V., Bayard, C., Vorobjev, V., y Arrang, J.-M. (2010). Histamine potentiates N-methyl-D-aspartate receptors by interacting with an allosteric site distinct from the polyamine binding site. *The Journal of Pharmacology and Experimental Therapeutics*, 332(3):912–921.

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W. P., Links, A. E., Washington, N. L., Haendel, M. A., Robinson, P. N., Boerkoel, C. F., Adams, D., Gahl, W. A., Boycott, K. M., y Brudno, M. (2015). PhenomeCentral: A Portal for Phenotypic and Genotypic Matchmaking of Patients with Rare Genetic Diseases. *Human Mutation*, 36(10):931–940.

Butte, A. J. y Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nature Biotechnology*, 24(1):55–62.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., y Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266.

Castellan Baldan, L., Williams, K. A., Gallezot, J.-D., Pogorelov, V., Rapanelli, M., Crowley, M., Anderson, G. M., Loring, E., Gorczyca, R., Billingslea, E., Wasylink, S., Panza, K. E., Ercan-Sencicek, A. G., Krusong, K., Leventhal, B. L., Ohtsu, H., Bloch, M. H., Hughes, Z. A., Krystal, J. H., Mayes, L., de Araujo, I., Ding, Y.-S., State, M. W., y Pittenger, C. (2014). Histidine decarboxylase deficiency causes tourette syndrome: parallel findings in humans and mice. *Neuron*, 81(1):77–90.

Chartier-Harlin, M.-C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M., Waucquier, N., Defebvre, L., Amouyel, P., Farrer, M., y Destée, A. (2001). Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet*, 364(9440):1167–1169.

Choi, M. R. (2015). Renal dopaminergic system: Pathophysiological implications and clinical perspectives. *World Journal of Nephrology*, 4(2):196.

Chong, J., Buckingham, K., Jhangiani, S., Boehm, C., Sobreira, N., Smith, J., Harrell, T., McMillin, M., Wiszniewski, W., Gambin, T., Coban Akdemir, Z., Doheny, K., Scott, A., Avramopoulos, D., Chakravarti, A., Hoover-Fong, J., Mathews, D., Witmer, P., Ling, H., Hetrick, K., Watkins, L., Patterson, K., Reinier, F., Blue, E., Muzny, D., Kircher, M., Bilguvar, K., López-Giráldez, F., Sutton, V., Tabor, H., Leal, S., Gunel, M., Mane, S., Gibbs, R., Boerwinkle, E., Hamosh, A., Shendure, J., Lupski, J., Lifton, R., Valle, D., Nickerson, D., y Bamshad, M. (2015). The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics*, 97(2):199–215.

Coletti, M. H., Bleich, H. L., y AT, M. (2001). Medical Subject Headings Used to Search the Biomedical Literature. *Journal of the American Medical Informatics Association*, 8(4):317–323.

Consortium, T. G. O. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., Kussmann, J., Shashi, V., Johnson, K., Rehder, C., Ballif, B. C., Shaffer, L. G., y Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9):838–846.

Couto, F. M., Silva, M. J., y Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 61(1):137–152.

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., Rosenstein, M. C., Wiegers, T. C., y Mattingly, C. J. (2013). The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Research*, 41(Database issue):D1104–D1114.

Deng, Y., Gao, L., Wang, B., y Guo, X. (2015). HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLos One*, 10(2):e0115692.

Dessimoz, C. y Škunca, N., editors (2017). *The Gene Ontology Handbook*, volume 1446 of *Methods in Molecular Biology*. Springer New York, New York.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., y Palsson, B. Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1777–1782.

Ellenbroek, B. A. y Ghiabi, B. (2014). The other side of the histamine H3 receptor. *Trends in Neurosciences*, 37(4):191–199.

EURORDIS (2005). Rare Diseases: Understanding this Public Health Priority. http://www.eurordis.org/IMG/pdf/princeps_document-EN.pdf.

Fajardo, I., Urdiales, J. L., Medina, M. A., y Sanchez-Jimenez, F. (2001). Effects of phorbol ester and dexamethasone treatment on histidine decarboxylase and ornithine decarboxylase in basophilic cells. *Biochemical Pharmacology*, 61(9):1101–1106.

Federoff, H. J. y Gostin, L. O. (2009). Evolving From Reductionism to Holism. *JAMA*, 302(9):994–996.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. V., Moreau, Y., Pettett, R. M., y Carter, N. P. (2009). REPORT DECIPHER : Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*, 84(4):524–533.

Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., y Dryja, T. P. (1986). A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–646.

Furmark, T., Tillfors, M., Garpenstrand, H., Marteinsdottir, I., Langström, B., Oreland, L., y Fredrikson, M. (2004). Serotonin transporter polymorphism related to amygdala excitability and symptom severity in patients with social phobia. *Neuroscience Letters*, 362(3):189–92.

Gahl, W. A., Markello, T. C., Toro, C., Fajardo, K. F., Sincan, M., Gill, F., Carlson-Donohoe, H., Gropman, A., Pierson, T. M., Golas, G., Wolfe, L., Groden, C., Godfrey, R., Nehrebecky, M., Wahl, C., Landis, D. M., Yang, S., Madeo, A., Mullikin, J. C., Boerkoel, C. F., Tifft, C. J., y Adams, D. (2012). The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genetics in Medicine*, 14(1):51–59.

García-Faroldi, G., Correa-Fiz, F., Abrighach, H., Berdasco, M., Fraga, M. F., Esteller, M., Urdiales, J. L., Sánchez-Jiménez, F., y Fajardo, I. (2009). Polyamines affect histamine synthesis during early stages of IL-3-induced bone marrow cell differentiation. *Journal of Cellular Biochemistry*, 108(1):261–271.

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M. S., Ray, P. N., So, J., Stavropoulos, D. J., y Brudno, M. (2013). PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Human Mutation*, 34(8):1057–1065.

Glatzer, F., Gschwandtner, M., Ehling, S., Rossbach, K., Janik, K., Klos, A., Bäumer, W., Kietzmann, M., Werfel, T., y Gutzmer, R. (2013). Histamine induces proliferation in keratinocytes from patients with atopic dermatitis through the histamine 4 receptor. *The Journal of Allergy and Clinical Immunology*, 132(6):1358–1367.

Godard, P. y Page, M. (2016). PCAN: phenotype consensus analysis to support disease-gene association. *BMC bioinformatics*, 17(1):518.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., y Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690.

Golbe, L. I. y Mouradian, M. M. (2004). Alpha-synuclein in Parkinson's disease: Light from two new angles. *Annals of Neurology*, 55(2):153–156.

Gratwicke, J., Jahanshahi, M., y Foltynie, T. (2015). Parkinson's disease dementia: a neural networks perspective. *Brain: A Journal of Neurology*, 138(Pt 6):1454–1476.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5-6):907–928.

Grubert, F., Zaugg, J., Kasowski, M., Ursu, O., Spacek, D., Martin, A., Greenside, P., Srivas, R., Phanstiel, D., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J., Bustamante, C., Steinmetz, L., Kundaje, A., y Snyder, M. (2015). Genetic Control of Chro-

matin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1051–1065.

Gunther, J., Tian, Y., Stamova, B., Lit, L., Corbett, B., Ander, B., Zhan, X., Jickling, G., Bos-Veneman, N., Liu, D., Hoekstra, P., y Sharp, F. (2012). Catecholamine-related gene expression in blood correlates with tic severity in tourette syndrome. *Psychiatry Research*, 200(2-3):593–601.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., y McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517.

Hastings, J. (2017). *The Gene Ontology Handbook*, chapter Primer in Ontologies. Volume 1446 of Dessimoz and Škunca (2017).

He, X. y Zhang, J. (2006). Toward a molecular understanding of pleiotropy. *Genetics*, 173(4):1885–1891.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., y Christakis, N. A. (2009). A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Computational Biology*, 5(4):e1000353.

Hoehndorf, R., Schofield, P. N., y Gkoutos, G. V. (2011). PhenomeNET: A whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18):e119.

Hoehndorf, R., Schofield, P. N., y Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080.

Hood, L. y Flores, M. (2012). A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnology*, 29(6):613–624.

Hoyer, D., Clarke, D. E., Fozard, J. R., Hartig, P. R., Martin, G. R., Mylecharane, E. J., Saxena, P. R., y Humphrey, P. P. (1994). International Union of Pharmacology classification of receptors for 5-hydroxytryptamine (Serotonin). *Pharmacological Reviews*, 46(2):157–203.

Hu, J. X., Thomas, C. E., y Brunak, S. (2016). Network biology concepts in complex disease comorbidities. *Nature Reviews Genetics*, 17(10):615–629.

Huang, D. W., Sherman, B. T., y Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J., y O'Donovan, C. (2015). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063.

Ibáñez, P., Bonnet, A.-M., Débarges, B., Lohmann, E., Tison, F., Pollak, P., Agid, Y., Dürr, A., y Brice, A. (1998). Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet*, 364(9440):1169–1171.

Innis, J. W. y Hedera, P. (2004). Two patients with monomelic ulnar duplication with mirror hand polydactyly: Segmental Laurin-Sandrow syndrome. *American Journal of Medical Genetics*, 131A(1):77–81.

Jensen, P. B., Jensen, L. J., y Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

Jiang, J. J. y Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *In the Proceedings of ROCLING X*, pages 19–33.

Johnston, J. D. y Skene, D. J. (2015). 60 YEARS OF NEUROENDOCRINOLOGY: Regulation of mammalian neuroendocrine physiology and rhythms by melatonin. *The Journal of Endocrinology*, 226(2):T187–T198.

Joshi, G., Pradhan, S., y Mittal, B. (2011). Vascular Gene Polymorphisms (EDNRA -231 G>A and APOE HhaI) and Risk for Migraine. *DNA and Cell Biology*, 30(8):577–584.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E., y Stein, L. (2004). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue):D428–D432.

Joyce, A. R. y Palsson, B. Ø. (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210.

Kanehisa, M. y Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R. C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S., y Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40(Database issue):D841–D846.

Kitano, H. (2002). Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837.

Klopocki, E., Lohan, S., Doelken, S. C., Stricker, S., Ockeloen, C. W., Soares Thiele de Aguiar, R., Lezirovitz, K., Mingroni Netto, R. C., Jamsheer, A., Shah, H., Kurth, I., Habenicht, R., Warman, M., Devriendt, K., Kordaß, U., Hempel, M., Rajab, A., Mäkitie, O., Naveed, M., Radhakrishna, U., Antonarakis, S. E., Horn, D., y Mundlos, S. (2012). Duplications of BHLHA9 are associated with ectrodactyly and tibia hemimelia inherited in non-Mendelian fashion. *Journal of Medical Genetics*, 49(2):119–125.

Koehler, S., Robinson, P. N., y Mungall, C. J. (2017). "Opposite-o"'-information improves similarity calculations in phenotype ontologies. *bioRxiv*.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., y Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, 85(4):457–464.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., y Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(Database issue):D862–D868.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., y Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue):D980–D985.

Lee, S., Abecasis, G. R., Boehnke, M., y Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics*, 95(1):5–23.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3):168–178.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O?Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., y Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

Lipton, R. B., Bigal, M. E., Steiner, T. J., Silberstein, S. D., y Olesen, J. (2004). Classification of primary headaches. *Neurology*, 63(3):427–435.

Liu, H., Hunter, L., Kešelj, V., y Verspoor, K. (2013). Approximate subgraph matching-based literature mining for biomedical events and relations. *PLos One*, 8(4):e60954.

Liu, H., Liu, M., Wang, Y., Wang, X.-M., Qiu, Y., Long, J.-F., y Zhang, S.-P. (2011). Association of 5-HTT gene polymorphisms with migraine: a systematic review and meta-analysis. *Journal of the Neurological Sciences*, 305(1-2):57–66.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., Bader, G. D., y Dopazo, J. (2011). Cytoscape Web: An interactive web-based network browser. *Bioinformatics*, 27(13):2347–2348.

Lord, P. W., Stevens, R. D., Brass, a., y Goble, C. a. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.

Loscalzo, J. y Barabasi, A.-L. (2011). Systems biology and the future of medicine. *Wiley interdisciplinary reviews. Systems Biology and Medicine*, 3(6):619–627.

Lussier, Y. A. y Liu, Y. (2007). Computational Approaches to Phenotyping: High-Throughput Phenomics. *Proceedings of the American Thoracic Society*, 4(1):18–25.

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A., Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W., y Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469–476.

Mariño-Enríquez, A., Lapunzina, P., Omeñaca, F., Morales, C., y Rodríguez, J. I. (2008). Laurin-Sandrow syndrome: Review and redefinition. *American Journal of Medical Genetics Part A*, 146A(19):2557–2565.

Martinel Lamas, D. J., Rivera, E. S., y Medina, V. A. (2015). Histamine H4 receptor: insights into a potential therapeutic target in breast cancer. *Frontiers in Bioscience (Scholar edition)*, 7:1–9.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R., y Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195.

McCusker, S. M., Curran, M. D., Dynan, K. B., McCullagh, C. D., Urquhart, D. D., Middleton, D., Patterson, C. C., McIlroy, S. P., y Passmore, A. P. (2001). Association between polymorphism in regulatory region of gene encoding tumour necrosis factor alpha and risk of Alzheimer's disease and vascular dementia: a case-control study. *Lancet*, 357(9254):436–439.

McGinnis, S. y Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server):W20–W25.

McKusick, V. A. (2007). Mendelian Inheritance in Man and its online version, OMIM. *American Journal of Human Genetics*, 80(4):588–604.

Mecheri, S. (2012). Contribution of allergic inflammatory response to the pathogenesis of malaria disease. *Biochimica et biophysica acta*, 1822(1):49–56.

Medina, M. Á. (2013). Systems biology for molecular life sciences and its impact in biomedicine. *Cellular and Molecular Life Sciences*, 70(6):1035–1053.

Medina, M. A., Correa-Fiz, F., Rodríguez-Caso, C., y Sánchez-Jiménez, F. (2005). A comprehensive view of polyamine and histamine metabolism to the light of new technologies. *Journal of Cellular and Molecular Medicine*, 9(4):854–864.

Medina, V., Croci, M., Crescenti, E., Mohamad, N., Sanchez-Jiménez, F., Massari, N., Nuñez, M., Cricco, G., Martin, G., Bergoc, R., y Rivera, E. (2008). The role of histamine in human mammary carcinogenesis: H3 and H4 receptors as potential therapeutic targets for breast cancer treatment. *Cancer Biology & Therapy*, 7(1):28–35.

Mitter, D., Ullmann, R., Muradyan, A., Klein-Hitpaß, L., Kanber, D., Õunap, K., Kaulisch, M., y Lohmann, D. (2011). Genotype–phenotype correlations in patients with retinoblastoma and interstitial 13q deletions. *European Journal of Human Genetics*, 19(9):947–958.

Montañez, R., Medina, M. A., Solé, R. V., y Rodríguez-Caso, C. (2010). When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays*, 32(3):246–256.

Mungall, C. J., McMurry, J. A., Köhler, S., Balhoff, J. P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., Foster, E., Gourdine, J. P., Jacobsen, J. O. B., Keith, D., Laraway, B., Lewis, S. E., NguyenXuan, J., Shefchek, K., Vasilevsky, N., Yuan, Z., Washington, N., Hochheiser, H., Groza, T., Smedley, D., Robinson, P. N., y Haendel, M. A. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(Database issue):D712–D722.

Nakazato, T., Bono, H., Matsuda, H., y Takagi, T. (2009). Gendoo: Functional profiling of gene and disease features using MeSH vocabulary. *Nucleic Acids Research*, 37(Web Server):W166–W169.

Newman, M. (2010). *Networks An Introduction*. Oxford University Press, New York.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. a., Shendure, J., y Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., y Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276.

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*, 27(12):1135–1137.

Nygaard, M., Debrabant, B., Tan, Q., Deelen, J., Andersen-Ranberg, K., de Craen, A. J. M., Beekman, M., Jeune, B., Slagboom, P. E., Christensen, K., y Christiansen, L. (2016). Copy number variation associates with mortality in long-lived individuals: a genome-wide assessment. *Aging cell*, 15(1):49–55.

O'Brien, J. T., Colloby, S., Fenwick, J., Williams, E. D., Firbank, M., Burn, D., Aarsland, D., y McKeith, I. G. (2004). Dopamine transporter loss visualized with FP-CIT SPECT in the differential diagnosis of dementia with Lewy bodies. *Archives of Neurology*, 61(6):919–925.

Oti, M. y Brunner, H. (2006). The modular nature of genetic diseases. *Clinical Genetics*, 71(1):1–11.

Oti, M., Huynen, M. A., y Brunner, H. G. (2008). Phenome connections. *Trends in Genetics*, 24(3):103–106.

Palleja, A., Horn, H., Eliasson, S., y Jensen, L. J. (2012). DistiLD Database: diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Research*, 40(Database issue):D1036–D1040.

Panula, P., Chazot, P. L., Cowart, M., Gutzmer, R., Leurs, R., Liu, W. L. S., Stark, H., Thurmond, R. L., y Haas, H. L. (2015). International Union of Basic and Clinical Pharmacology. XCVIII. Histamine Receptors. *Pharmacological Reviews*, 67(3):601–655.

Panula, P. y Nuutinen, S. (2013). The histaminergic network in the brain: basic organization and role in disease. *Nature Reviews. Neuroscience*, 14(7):472–87.

Panula, P., Sundvik, M., y Karlstedt, K. (2014). Developmental roles of brain histamine. *Trends in Neurosciences*, 37(3):159–168.

Passani, M. B. y Blandina, P. (2011). Histamine receptors in the CNS as targets for therapeutic intervention. *Trends in Pharmacological Sciences*, 32(4):242–249.

Pegg, A. E. (2014). The function of spermine. *IUBMB Life*, 66(1):8–18.

Perez, R. G., Waymire, J. C., Lin, E., Liu, J. J., Guo, F., y Zigmond, M. J. (2002). A role for alpha-synuclein in the regulation of dopamine biosynthesis. *The Journal of Neuroscience : The Official Journal of The Society for Neuroscience*, 22(8):3090–3099.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., y Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLoS Computational Biology*, 5(7):e1000443.

Pickrell, A. M. y Youle, R. J. (2015). The roles of PINK1, parkin, and mitochondrial fidelity in Parkinson's disease. *Neuron*, 85(2):257–273.

Pierce, B. A. (2014). *Genética. Un enfoque conceptual*. Editorial Médica Panamericana S.A., New York, quinta edición edition.

Pino-Ángeles, A., Reyes-Palomares, A., Melgarejo, E., y Sánchez-Jiménez, F. (2012). Histamine: an undercover agent in multiple rare diseases? *Journal of Cellular and Molecular Medicine*, 16(9):1947–1960.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X., y Jensen, L. J. (2014). DISEASES: Text mining and data integration of disease-gene associations. *Methods*, 74:83–89.

Pruitt, K. D., Tatusova, T., y Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database):D61–D65.

Raj, K. P., Zell, J. A., Rock, C. L., McLaren, C. E., Zoumas-Morse, C., Gerner, E. W., y Meyskens, F. L. (2013). Role of dietary polyamines in a phase III clinical trial of difluoromethylornithine (DFMO) and sulindac for prevention of sporadic colorectal adenomas. *British Journal of Cancer*, 108(3):512–518.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C. P., Safran, M., y Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database: The Journal of Biological Databases and Curation*, 2013:bat018.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D., Kriegel, H.-P., Hoth, D., Oates, J., Peck, C., Schooley, R., Spilker, B., Woodcock, J., y Zeger, S. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *IJCAI*, pages 448–453.

Reyes-Palomares, A., Bueno, A., Rodríguez-López, R., Medina, M. Á., Sánchez-Jiménez, F., Corpas, M., y Ranea, J. A. G. (2016). Systematic identification of phenotypically enriched loci using a patient network of genomic disorders. *BMC Genomics*, 17(1):232.

Reyes-Palomares, A., Rodríguez-López, R., Ranea, J. A. G., Sánchez Jiménez, F., y Medina, M. A. (2013). Global analysis of the human pathophenotypic similarity gene network merges disease module components. *PLos One*, 8(2):e56653.

Robinson, P. y Mundlos, S. (2010). The human phenotype ontology. *Clinical Genetics*, 77(6):525–534.

Robinson, P. N. (2014). Computation Phenotype Analysis in Human Medicine. In Hancock, J. M., editor, *Phenomics*, chapter 2, page 286. CRC Press, Cambridge.

Robinson, P. N. (2015). *Phenomics*, chapter Computational Phenotype Analysis in Human Medicine. Springer International Publishing, Cham.

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., y Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, 83(5):610–615.

Rodríguez-López, R., Morales, M., y Sánchez-Jiménez, F. (2016). Histamine and Its Receptors as a Module of the Biogenic Amine Diseasome. In *Histamine Receptors*, pages 173–214. Springer International Publishing, Cham.

Rodríguez-López, R., Reyes-Palomares, A., Sánchez-Jiménez, F., y Medina, M. Á. (2014). PhenUMA: a tool for integrating the biomedical relationships among genes and diseases. *BMC bioinformatics*, 15(1):375.

Ruch, P., Gobeill, J., Lovis, C., y Geissbühler, A. (2008). Automatic medical encoding with SNOMED categories. *BMC Medical Informatics and Decision Making*, 8 Suppl 1(Suppl 1):S6.

Sáiz, P. A., García-Portilla, M. P., Arango, C., Morales, B., Arias, B., Corcoran, P., Fernández, J. M., Alvarez, V., Coto, E., Bascarán, M.-T., Bousoño, M., Fañanas, L., y Bobes, J. (2010). Genetic polymorphisms in the dopamine-2 receptor (DRD2), dopamine-3 receptor (DRD3), and dopamine transporter (SLC6A3) genes in schizophrenia: Data from an association study. *Progress in Neuro-psychopharmacology & Biological Psychiatry*, 34(1):26–31.

Sánchez-Jiménez, F., Montañez, R., Correa-Fiz, F., Chaves, P., Rodríguez-Caso, C., Urdiales, J. L., Aldana, J. F., y Medina, M. A. (2007). The usefulness of post-genomics tools for characterization of the amine cross-talk in mammalian cells. *Biochemical Society Transactions*, 35(Pt 2):381–385.

Sánchez-Jiménez, F., Ruiz-Pérez, M. V., Urdiales, J. L., y Medina, M. A. (2013). Pharmacological potential of biogenic amine-polyamine interactions beyond neurotransmission. *British Journal of Pharmacology*, 170(1):4–16.

Saulnier Sholler, G. L., Gerner, E. W., Bergendahl, G., MacArthur, R. B., VanderWerff, A., Ashikaga, T., Bond, J. P., Ferguson, W., Roberts, W., Wada, R. K., Eslin, D., Kraveka, J. M., Kaplan, J., Mitchell, D., Parikh, N. S., Neville, K., Sender, L., Higgins, T., Kawakita, M., Hiramatsu, K., Moriya, S.-S., y Bachmann, A. S. (2015). A Phase I Trial of DFMO Targeting Polyamine Addiction in Patients with Relapsed/Refractory Neuroblastoma. *PLos One*, 10(5):e0127246.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martinez-Cruz, L. A., Corrales, F. J., y Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. *IEEEACM Transactions on Computational Biology and Bioinformatics*, 2(4):330–338.

Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, 46(1):561–584.

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., y Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, 10(12):2004–2015.

Smedley, D., Kohler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., Veldboer, J., Zemojtel, T., y Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*, 30(22):3215–3222.

Smedley, D. y Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*, 7(1):81.

Smith, B., Williams, J., y Schulze-Kremer, S. (2003). The ontology of the gene ontology. *AMIA Annu Symp Proc*, 2003:609–613.

Solé, R. V. y Valverde, S. (2008). Spontaneous emergence of modularity in cellular networks. *Journal of The Royal Society Interface*, 5(18):129–133.

Sparkes, R. S., Sparkes, M. C., Wilson, M. G., Towner, J. W., Benedict, W., Murphree, A. L., y Yunis, J. J. (1980). Regional assignment of genes for human esterase D and retinoblastoma to chromosome band 13q14. *Science (New York, N.Y.)*, 208(4447):1042–1044.

Stessman, H. A., Bernier, R., y Eichler, E. E. (2014). A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell*, 156(5):872–877.

Stevenson, K. L. (2004). Chiari Type II malformation: past, present, and future. *Neurosurgical Focus*, 16(2):E5.

Strug, L. J., Suresh, R., Fyer, A. J., Talati, A., Adams, P. B., Li, W., Hodge, S. E., Gilliam, T. C., y Weissman, M. M. (2010). Panic disorder is associated with the serotonin transporter gene (SLC6A4) but not the promoter region (5-HTTLPR). *Molecular Psychiatry*, 15(2):166–176.

Sutcliffe, J. S., Delahanty, R. J., Prasad, H. C., McCauley, J. L., Han, Q., Jiang, L., Li, C., Folstein, S. E., y Blakely, R. D. (2005). Allelic heterogeneity at the serotonin transporter locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors. *American Journal of Human Genetics*, 77(2):265–279.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., Jensen, L. J., y Von Mering, C. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(SUPPL. 1):561–568.

Thomas, P. D. (2017). *The Gene Ontology Handbook*, chapter The Gene Ontology and the Meaning of Biological Function. Volume 1446 of Dessimoz and Škunca (2017).

Thompson, R., Johnston, L., Taruscio, D., Monaco, L., Béroud, C., Gut, I. G., Hansson, M. G., 't Hoen, P.-B. A., Patrinos, G. P., Dawkins, H., Ensini, M., Zatloukal, K., Koubi, D., Heslop, E., Paschall, J. E., Posada, M., Robinson, P. N., Bushby, K., y Lochmüller, H. (2014). RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research. *Journal of General Internal Medicine*, 29(S3):780–787.

Thurmond, R. L. (2015). The histamine H4 receptor: from orphan to the clinic. *Frontiers in Pharmacology*, 6:65.

Thurmond, R. L., Gelfand, E. W., y Dunford, P. J. (2008). The role of histamine H1 and H4 receptors in allergic inflammation: the search for new antihistamines. *Nature Reviews. Drug discovery*, 7(1):41–53.

Tong, J. H., Cummins, T. D., Johnson, B. P., McKinley, L.-A., Pickering, H. E., Fanning, P., Stefanac, N. R., Newman, D. P., Hawi, Z., y Bellgrove, M. A. (2015). An association between

a dopamine transporter gene (slc6a3) haplotype and adhd symptom measures in non-clinical adults. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(2):89–96.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., Morrison, K., Donaldson, I. M., y Wodak, S. J. (2010). irefweb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, 2010:baq023.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., y Leunissen, J. A. M. (2006). A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5):535–542.

Veeramani, B. y Bader, J. S. (2009). Metabolic Flux Correlations, Genetic Interactions, and Disease. *Journal of Computational Biology*, 16(2):291–302.

Vidal, M., Cusick, M., y Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell*, 144(6):986–998.

Visscher, P., Brown, M., McCarthy, M., y Yang, J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Wagner, A. y Fell, D. A. (2001). The small world inside large metabolic networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1478).

Wang, Z. y Zhang, J. (2007). In Search of the Biological Significance of Modular Structures in Protein Networks. *PLoS Computational Biology*, 3(6):e107.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., y Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web Server issue):W214–W220.

Weinreich, S. S., Mangon, R., Sikkens, J. J., en Teeuw, M. E., y Cornel, M. C. (2008). Orphanet: a European database for rare diseases. *Nederlands Tijdschrift Voor Geneeskunde*, 152(9):518–519.

Wendland, J. R., DeGuzman, T. B., McMahon, F., Rudnick, G., Detera-Wadleigh, S. D., y Murphy, D. L. (2008). SERT Ileu425Val in autism, Asperger syndrome and obsessive-compulsive disorder. *Psychiatric Genetics*, 18(1):31–9.

Xu, T., Du, L., y Zhou, Y. (2008). Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9(1):472.

Yang, H., Robinson, P. N., y Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, 12(9):841–843.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., y Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

Zampeli, E. y Tiligada, E. (2009). The role of histamine H4 receptor in immune and inflammatory disorders. *British Journal of Pharmacology*, 157(1):24–33.

Zhang, M., Zhu, C., Jacomy, A., Lu, L. J., y Jegga, A. G. (2011). The orphan disease networks. *American Journal of Human Genetics*, 88(6):755–766.

Anexo

En este anexo se incluyen los artículos publicados a lo largo de esta tesis.

OPEN ACCESS Freely available online

PLOS ONE

Global Analysis of the Human Pathophenotypic Similarity Gene Network Merges Disease Module Components

Armando Reyes-Palomares[1,2], Rocío Rodríguez-López[1,2], Juan A. G. Ranea[1,2], Francisca Sánchez Jiménez[1,2], Miguel Angel Medina[1,2]*

1 Department of Molecular Biology and Biochemistry, Faculty of Sciences, University of Málaga, Málaga, Spain, **2** CIBER de Enfermedades Raras (CIBERER), Málaga, Spain

Abstract

The molecular complexity of genetic diseases requires novel approaches to break it down into coherent biological modules. For this purpose, many disease network models have been created and analyzed. We highlight two of them, "the human diseases networks" (HDN) and "the orphan disease networks" (ODN). However, in these models, each single node represents one disease or an ambiguous group of diseases. In these cases, the notion of diseases as unique entities reduces the usefulness of network-based methods. We hypothesize that using the clinical features (pathophenotypes) to define pathophenotypic connections between disease-causing genes improve our understanding of the molecular events originated by genetic disturbances. For this, we have built a pathophenotypic similarity gene network (PSGN) and compared it with the unipartite projections (based on gene-to-gene edges) similar to those used in previous network models (HDN and ODN). Unlike these disease network models, the PSGN uses semantic similarities. This pathophenotypic similarity has been calculated by comparing pathophenotypic annotations of genes (human abnormalities of HPO terms) in the "Human Phenotype Ontology". The resulting network contains 1075 genes (nodes) and 26197 significant pathophenotypic similarities (edges). A global analysis of this network reveals: unnoticed pairs of genes showing significant pathophenotypic similarity, a biological meaningful re-arrangement of the pathological relationships between genes, correlations of biochemical interactions with higher similarity scores and functional biases in metabolic and essential genes toward the pathophenotypic specificity and the pleiotropy, respectively. Additionally, pathophenotypic similarities and metabolic interactions of genes associated with maple syrup urine disease (MSUD) have been used to merge into a coherent pathological module. Our results indicate that pathophenotypes contribute to identify underlying co-dependencies among disease-causing genes that are useful to describe disease modularity.

Citation: Reyes-Palomares A, Rodríguez-López R, Ranea JAG, Jiménez FS, Medina MA (2013) Global Analysis of the Human Pathophenotypic Similarity Gene Network Merges Disease Module Components. PLoS ONE 8(2): e56653. doi:10.1371/journal.pone.0056653

Editor: Steve Horvath, University of California Los Angeles, United States of America

Received August 29, 2012; **Accepted** January 12, 2013; **Published** February 21, 2013

Funding: The authors' experimental work is supported by grants SAF2011/26518, SAF2009/09839, PI12/01096 and PS09/02216 (Spanish Ministry of Economy and Competitiveness and FEDER), and PIE P08-CTS-3759, CVI-6585 and funds from group BIO-267 (Andalusian Government and FEDER). JR acknowledges grants SAF2009-09839 and SAF2012-33110 and FSJ acknowledges funds from an INTERCONNECTA-AMER grant (Spanish Ministry of Economy and Competitiveness and FEDER). The "CIBER de Enfermedades Raras" is an initiative from the ISCIII (Spain). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: MAM is a PLOS ONE Editorial board member. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

* E-mail: medina@uma.es

Introduction

Phenotypes are the result of the expression of specific genetic backgrounds submitted to the influence of changing environmental conditions [1]. Thus, both the development and resulting symptoms of a given pathology are conditioned by interacting elements at multiple interconnected levels (from molecular to social levels) [2]. These complex interactions can be represented as networks to be analyzed using the principles of Network Theory [3–6]. In this sense, Network Medicine emerged as a new field to study the relationships among diseases and disease-causing genes [7]. Generally, data from genetic association studies establish the basic information for these analyses. Most of these data are available from different public repositories, for instance, Online Mendelian Inheritance in Man (OMIM) [8] and Orphanet [9]. This information can be projected onto networks also known as diseasomes (i.e. "the human disease network" and "the orphan disease networks") [10,11]. These diseasomes open the possibility to work on different types of network projections, treating networks as graphs, which can be used to detect emergent information. For instance, disease-to-gene associations represent bipartite edges (two different types of nodes in every edge) and conform a bipartite graph (as shown in the schematic representation in Figure 1A). On the other hand, projections of gene-to-gene edges and disease-to-disease edges can be inferred from the initial bipartite graph as two different "unipartite" graphs (each with only one type of node). Hence, edges in both inferred unipartite graphs represent either genes associated by a same disease (Figure 1A) or diseases associated through a same gene (these edges were not considered in this study), respectively. The first type of projections (gene-to-gene) are disease-causing gene

networks and the second ones (disease-to-disease edges) are generally known as disease networks [10,11]. Network-based methods enable us to find disease modules that may be understood as all molecular relationships involving disease-causing genes and other genes related to the same pathological processes [7]. In fact, several different biomolecular interactomes based on physical, metabolic or functional interactions have been used to capture some frames of the biological complexity associated with pathologies [12–17]. In this case, one of the most direct applications of network medicine approaches lies in the systematic exploration of the molecular mechanism shared by "apparently" distinct diseases [7]. The emergence of relationships among genes and diseases contribute to obtain more holistic views of the disease origin and environment, to predict new disease-causing genes [17], and possibly to locate new targets for disease diagnosis and/or intervention. All these challenges take part in a wider emergent discipline known as Systems Medicine [18].

However, current pathognomonic classifications are influenced by the traditional clinical procedures used during the 19th century following Oslers principles [19]. These traditional procedures often tend to overvalue the most evident manifested abnormalities (pathophenotypes), causing a direct impact on how pathophenotypic profiles of patients are registered in the clinic [19]. Although it could help the diagnosis, many others pathophenotypes will go unnoticed. As a consequence, most genetic diseases are described as conceptual entities, pathologies, with certain specific clinical features. The disregard of pathophenotypes implies a considerable technical problem for network medicine based methods, since they can be primary consequences of the genetic disturbances. At present, to solve this problem standard phenotypic platforms are required to explore the underlying molecular and cellular mechanisms related to genetic predisposition in developing diseases [20]. Nevertheless, some previous works have claimed that the systematic phenotyping procedure requires ontologies to improve biomedical insights on functional gene communities [21–23]. In this case, the use of ontologies can be an interesting advance in the biomedical integration of this information. The Human Phenotype Ontology (HPO) represents a formalization of the semantic relationships [21,24] among different clinical features described in OMIM (abbreviations used throughout the manuscript are reported in Table 1). Although HPO was initially developed to study the phenotypic associations in order to achieve a potential diagnostic use [25], this standardized biomedical knowledge on human abnormalities allows the identification of

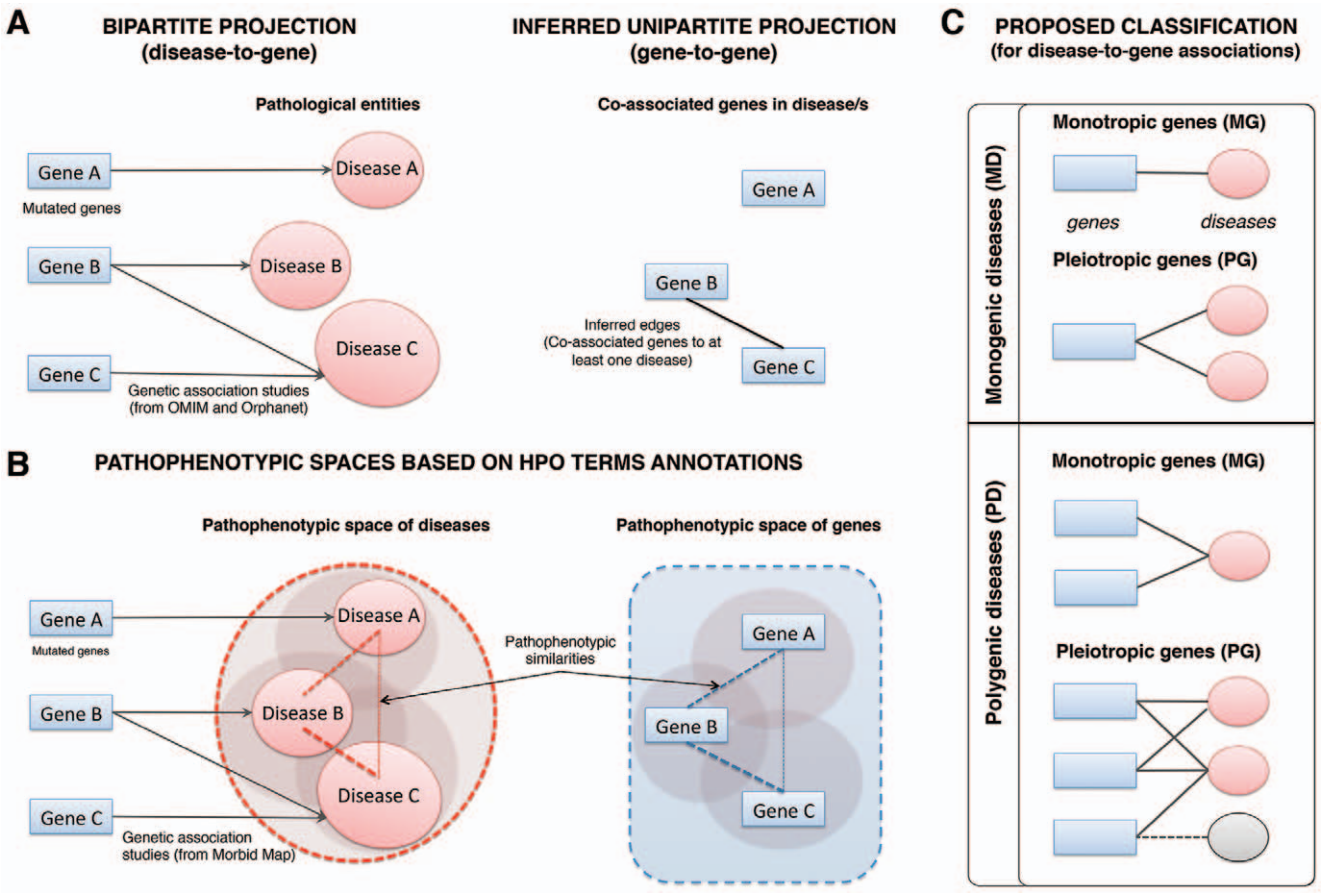


Figure 1. Schematic representation of distinct disease-to-gene relationships. Different disease associations between genes using (A) the data from genetic disease databases or (B) their associated pathological phenotypes. (A) The co-associations of genes in disease/s allow the inference of gene-to-gene projection (unipartite) from the disease-to-gene projection (bipartite). In this case Gene B and Gene C are co-associated with Disease C. (B) The HPO annotations of genetic diseases allow the description of pathophenotypic space for genes and calculation of the semantic similarity (pathophenotypic similarity) between them. In this case, novel relationships emerge as occur between Gene A and Gene B or Gene A and Gene C. (C) The proposed classification in this work: monogenic disease and monotropic genes (MD-MG), monogenic disease and pleiotropic genes (MD-PG), polygenic disease and monotropic gene (PD-MG), polygenic disease and pleiotropic gene (PD-PG). It is noteworthy that genes present in the MD-PG subset can also be present in the PD-PG subset (dashed line linked to monogenic disease in grey).
doi:10.1371/journal.pone.0056653.g001

functional gene-to-gene relationships involved in similar pathological processes [26]. Recent studies conclude that the phenotypic similarity measurement proposed by Robinson and co-workers [25] has a significant contribution to the biological coherence compared to text-mining methods [27]. Therefore, on the one hand, the study of the similarity among pathologies requires representing them as a set of pathophenotypes instead of a pathological entity. On the other side, pathophenotypic information can be used to reinterpret the relationships among diseases identifying a new pathological phenotypic space that makes it possible the study of novel gene-to-gene associations (as can be seen in the schematic representation in Figure 1B). Zhang et al. [11] have recently stressed some limitations of network-based methods suggesting that the relationships between rare diseases cannot be fully captured by gene-to-gene projections alone. Therefore, the efforts to characterize the genetic and functional environment of given diseases (disease modules) can contribute to enrich the usefulness of disease network analyses.

In this work, network medicine approaches have been used to study the pathological relationships among genes using semantic similarities (that in this case are pathophenotypic similarities) instead of inferred unipartite edges (gene-to-gene) from bipartite edges (disease-to-gene associations). For instance, a classification of four distinct disease-to-gene associations is proposed (Figure 1C) to illustrate possible limitations of the current disease-to-gene network models [10,11]. These classes provide four different subsets of genes in agreement with the number of genes associated with a disease (monogenic or polygenic) and the number of diseases associated with a gene (monotropic and pleiotropic). We have also built a pathophenotypic similarity gene network (PSGN) using semantic similarity [25] between genes that are annotated in HPO. The topological features of gene subsets obtained from inferred pathological networks have been analyzed and compared in PSGN. Additionally, the representation of PSGN in three different human biomolecular interactomes based on physical interactions, metabolic flux coupling and functional interactions were also evaluated. For this, a network comparison analysis [28] and a subsequent performance validation have been used to study the degree of contribution of each biomolecular interactome to the biological consistency of gene-to-gene pathophenotypic similarities. In addition, this biological coherence can be used to incorporate novel components in disease-causing gene modules, as we demonstrate for maple syrup urine disease (MSUD), an inborn error of the metabolism of branched-chain amino acids.

Summarizing, this work provides evidence that a standard phenotypic profiling expands the genetic disease associations using a specific ontology for human abnormalities. These pathologic relationships among genes were not obvious and, consequently, disregarded in previous disease network analyses.

Table 1. List of abbreviations used throughout the paper.

Abbreviation	Description
HPO	Human Phenotype Ontology
OMIM	Online Mendelian Inheritance in Man
HDN	Human Disease Network (bipartite projection)
ODN	Orphan Disease Network (bipartite projection)
HDGN	Human Disease Gene Network (unipartite projection)
ODGN	Orphan Disease Gene Network (unipartite projection)
MD-MG	Monogenic Disease and Monotropic Genes
MD-PG	Monogenic Disease and Pleiotropic Genes
PD-MG	Polygenic Disease and Monotropic Genes
PD-PG	Polygenic Disease and Pleiotropic Genes
PSGN	Pathophenotypic Similarity Gene Network
PIN	Physical Interaction Network
MGN	Metabolic Gene Network
FSGN	Functional Similarity Gene Network

doi:10.1371/journal.pone.0056653.t001

Methods

Unipartite Projections of Current Diseasomes

Human disease causing gene network. In the present study, we worked on an updated version of the "Human Diseases Network" (HDN) [10] using Morbid Map from OMIM (http://www.omim.org/). HDN represents a bipartite projection of edges with two types of nodes, genes (MIM genes) and diseases (MIM phenotypes and genes/phenotypes) as described in OMIM. We followed a similar methodology to the one described by Goh et al. [10]. We retrieved all disease-to-gene associations where molecular bases are known and we discarded those phenotypes without MIM numbers. However, unlike previous works [10] we have not grouped diseases according to the similarity between their names. Here, each MIM phenotype or MIM gene/phenotype was considered as a pathological entity and each MIM gene was transformed to its respective Entrez Gene ID. This new version of the HDN consists of 2525 genes (Entrez Gene IDs) associated with 3132 OMIM entries (MIM numbers) generating a network of 5657 nodes and 3862 edges (HDN in Table S1). Hence, we built the respective unipartite projections based on inferred gene-to-gene relationships, named as human disease causing gene network (HDGN). This inference provides emergent gene-to-gene edges if genes are sharing at least one disease.

Orphan disease causing gene network. An updated version of the "Orphan Disease Networks" (ODN) [11] was built using Orphanet data. We used Orphanet because it is focused on genetic and low prevalent diseases; this database is actively updated and continuously reviewed by clinical experts. ODN is the bipartite projection of edges with two types of nodes, genes (Orpha numbers for genes) and orphan diseases (also in Orpha numbers for diseases). All those genes identified with Orpha numbers were transformed to Entrez Gene IDs. This new version of ODN consists of 2331 genes (Entrez Gene IDs) associated with 2125 genetic orphan diseases (ORPHA numbers) generating a network of 4456 nodes and 3657 edges (ODN in Table S2). In a similar procedure to that used for HDN (mentioned above), we built the unipartite projections based on gene-to-gene inferred relationships for ODN, named orphan disease-causing genes network (ODGN).

Classification of Disease-to-gene Associations in Diseasomes

Both HDN and ODN were decomposed into four subclasses, based on the classification of the different types of disease-to-gene associations (Figure 1C): monogenic diseases associated with monotropic genes (MD-MG), monogenic diseases associated with pleiotropic genes (MD-PG), polygenic diseases associated with monotropic genes (PD-MG) and polygenic diseases associated with pleiotropic genes (PD-PG). In the context of the present study, we use the expression "monotropic genes" to refer to genes that have been previously related to only one disease and the expression "pleiotropic genes" to refer to genes that have been previously

related to two or more diseases. Each subclass contains a subset of genes (Tables S3 and S4 Supplementary material).

Pathophenotypic Similarity Gene Network (PSGN)

The pathophenotype gene network was built using pre-calculated values of semantic similarities between genes through the Human Phenotype Ontology (HPO). Previously, we had to describe the pathophenotypic space for genes as the set of clinical features (HPO terms) associated with each gene. Altogether 4669 diseases and 258 genes have direct annotations of their clinical features in HPO, so these diseases and genes have a list of HPO terms describing their phenotypic space. However, the lack of specific HPO terms regarding phenotypic abnormalities for many disease-causing genes hinders the explanation of their semantic relationships in the ontology. Many genes are annotated in the ontology with the sum of all HPO terms that describe their associated diseases in Morbid Map. In these cases, we used the file "gene_to_phenotype.txt" (available on HPO website) to link HPO terms and genes. This file was generated using Morbid Map associations between genes and diseases. Therefore, clinical features described in OMIM were translated in a standardized vocabulary of HPO terms (phenotypic abnormalities) that have been used to define a pathophenotypic space. As mentioned above, this pathophenotypic space for a gene can be directly annotated in HPO or indirectly annotated by the diseases associated with the gene in Morbid Map. We used the phenotypic space of genes to calculate their pathophenotypic semantic similarities with other genes. Only HPO terms with maximal information were used in agreement with the ontology properties and distribution of terms (see semantic similarity calculations section below). We discarded those branches of the ontology without an explicit description of phenotypic abnormalities such as "mode of inheritance" and "onset and clinical course". We obtained a large pathophenotype gene network based on all semantic similarities between genes sharing HPO terms annotated in the phenotypic abnormality branch of the HPO. Despite an extensive literature review we could not detect a systematic methodology to calculate a cut-off score distinguishing between relevant or non-specific semantic similarities. Previous works used the semantic similarity to validate predictions or to evaluate shared biological features between highly specific subset of genes. However, in this case, we needed an optimal statistical threshold from which the signals, pathophenotypic similarities, should be out of the background noise. The cut-off will predetermine the topology of the network, so it could affect arguments and discussion about the "expansion" of pathophenotypic relationships respect to current unipartite projections (HDGN and ODGN). If we select a low similarity score we will introduce exponentially nonspecific relationships. In contrast, a very high score will constraint the model to already known pathological relationships. Therefore, we used the subset of known pathophenotypic similarities (gene pairs) in a binary classification system to estimate the optimal statistical threshold (see supplemental methods and discussion in Methods S1). Finally, the number of unspecific similarities was reduced by selecting the cut-off at the 98th percentile that corresponds to the top 2% of significant gene pairs with higher semantic similarity values. To assess this clustering process of PSGN in the top 2% of phenotypic similarity, we plotted a kernel density distribution of probability of the pathophenotypic similarity for gene pairs (Figure 2).

Biomolecular Interactomes

Physical interaction network (PIN). We used the CRG Human Interactome as the reference for physical interaction network (PIN). This network of protein-to-protein physical interactions contains 10299 genes (Ensembl gene IDs) and 80922 interactions supported by evidence from at least one experiment [29]. The topological analysis of the largest connected component of the CRG Human Interactome was carried out under a similar procedure to that described in previous published works [30,31]. However, all Ensembl gene IDs were transformed to Entrez Gene IDs to enable a node degree correlation and network comparison analysis with PSGN.

Metabolic gene network (MGN) based on metabolic flux correlations. Metabolic networks are usually based on different metabolic coupling approaches such as metabolite sharing (for instance, shared metabolites between enzymes) [15,32,33] and metabolic flux correlations (for instance, correlated metabolic enzymes by flux balance analysis) [34]. In this work, we used the flux-coupling metabolic network built by Veeramani et al. [34]. This network is based on the results of a flux balance analysis [34] of an updated version of the Human Metabolic network Recon 1 [35]. We built MGN using only these gene-to-gene interactions exceeding a metabolic flux correlation value of 0.1 and a "metscore" of 0 from the original network (Table S5, supplementary material).

Functional similarity gene network (FSGN) based on biological processes. The FSGN was built by using the measurement of the semantic similarity between genes described in the branch of biological processes of the Gene Ontology (GO). The functional space of a gene is represented by the set of GO annotations about the biological context where the gene is involved. Thanks to these annotations, genes are directly linked to biological processes describing all the functional features direct or indirectly related to genes. Classical semantic similarity measurements were used to calculate functional similarities between genes according to their functional space. In a similar procedure used for PSGN we removed unspecific functional associations in FSGN generated by irrelevant semantic relationships. However, there are great differences in the number of annotations between HPO and the branch of biological processes of GO. In this case, the main concern is that it resulted in huge size of this dataset. Therefore, we preferred to be quite more restrictive for this threshold, by taking as cut-off the 99.5th percentile instead of the 98th. Thus we selected the top 0.5% of gene pairs with higher functional similarities (Figure S1).

Semantic Similarity Score Calculations (Gene-to-gene)

The way to assign terms to objects is to add annotations. In the present case, the objects represent genes and terms corresponding to phenotypes (HPO terms) or biological processes (GO terms). The specificity of the terms associated with genes allows us to calculate the most significant relationships between them, which use to be related to its proximity to the root. The method we have chosen to calculate the semantic similarity between objects annotated is mainly based on the classical Resnik's measurement [36]. This approach uses the information content (IC) concept that is a way to estimate the specificity of a term [25] and can be defined as the negative natural logarithm of the probability of a term

$$IC(t) = -logp(t) \quad (1)$$

where p(t) is defined on the basis of its frequency (number of term annotated) and the total of terms annotated in the ontology.

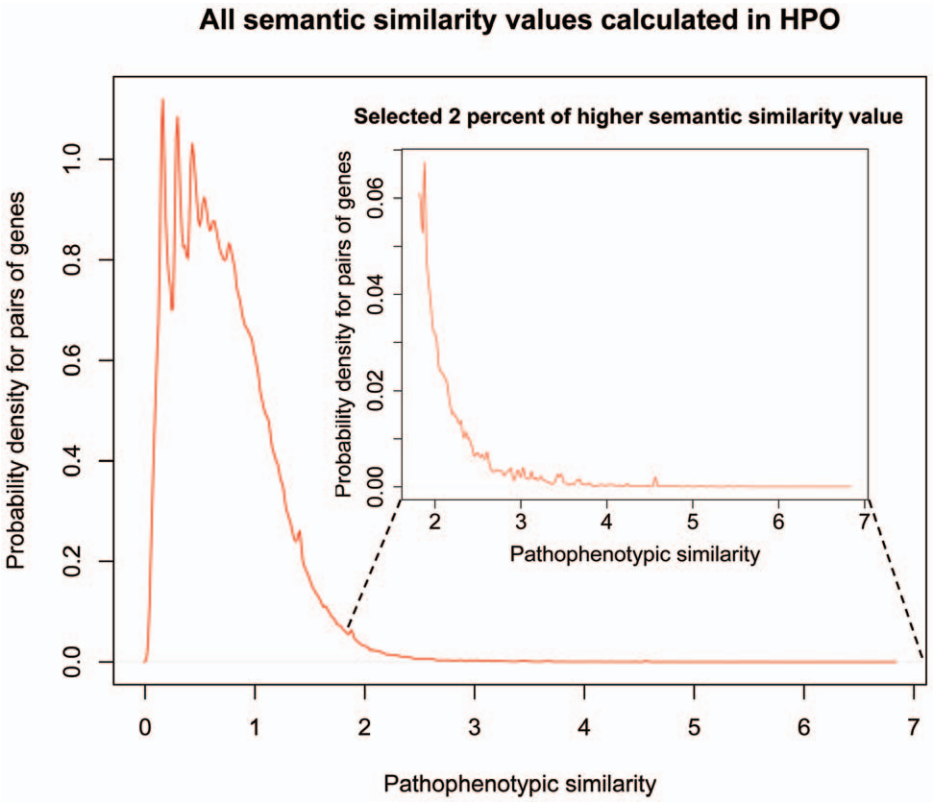


Figure 2. Probability density function for pathophenotypic similarities among pairs of genes in HPO. Densities of the pathophenotypic similarity values for all annotated genes in HPO (outer plot) and for the top 2% of gene pairs with the highest pathophenotypic similarities (inner plot). The bandwidth used was 0.01 and the pathophenotypic similarity value for the cut-off at the top 2% was 1.8179.
doi:10.1371/journal.pone.0056653.g002

$$p(t) = \frac{annotations(t)}{total\ annotations} \quad (2)$$

If the probability decreases then the information content increases and consequently the specificity and the informativeness increase too. Thus, the IC tends to increase as we move away from the root to more specific terms.

For t1 and t2 terms in the ontology, the semantic similarity proposed by Resnik is defined as:

$$sim(t1,t2) = \max_{p \in S(t1,\ t2)} IC(p) \quad (3)$$

where S(t1,t2) is the set of the shared parents of t1 and t2. In other words, the semantic similarity between two terms corresponds with the information content of the most informative common ancestor (MICA) [36].

Functional Semantic Similarity. Many studies so far have made a comparison between semantic similarity measurements using the Gene Ontology, but it seems that there is not a gold standard for semantic similarity measures between set of GO terms. In this work we use:

$$sim(g1,g2) = \max_{ti \in g1\, tj \in g2} sim(ti,tj) \quad (4)$$

a measurement that has been successfully used in some previously published works [37,38]. In (4) g1 and g2 represent genes, where each one is related with a set of ontological terms. The semantic similarity value between sets of terms is calculated by comparing each pair of terms (3), one term of each set, and determined from the maximum value of all pair comparisons.

Pathophenotypic Semantic Similarity. Human Phenotype Ontology is still a novel tool and there are not many works related to the calculation of semantic similarity for this data structure. We have chosen the method proposed by the HPO creators for the comparisons between phenotypic profiles [25]. For g1 and g2 two genes; their semantic similarity is defined as:

$$sim(g1,g2) = \frac{1}{|g1|}\left[\sum_{ti \in g1} \max_{tj \in g2} sim(ti,tj)\right] \quad (5)$$

where firstly is calculated the maximum value of IC, using the equation (3), between each term of g1 and the terms of g2. Finally, a set of values |g1| are used to work out their average.

The previous equation does not provide a symmetric matrix, since the calculated semantic similarity between g1 and g2 will not be the same as semantic similarity between g2 and g1, so Robinson and co-workers [25] suggest a symmetric version:

$$sim_{symmetric}(g1,g2) = \frac{1}{2}sim(g1,g2) + \frac{1}{2}sim(g2,g1) \quad (6)$$

Statistical Computing and Network Based Methods

All statistical computing, data management and graphics were performed in R, a free software environment. Network visualizations and their metadata analyses were performed in Cytoscape [39] and iGraph software, an R package (http://igraph.sourceforge.net/). Due to the large number of subsequent analysis of all built network, we provided a schematic workflow of all the essential steps followed for this study (Figure S1).

Network comparison analysis. Once all networks were built, we carried out a network analysis comparison to compute the nodal and edge intersection between PSGN and the rest of the built networks (HDGN, ODGN, PIN, MGN and FSGN). In the case of disease-causing gene networks (HDGN and ODGN unipartite projections of diseasomes), the intersection could provide a broad view of the similarity of these networks and the PSGN. Previously, we also calculated the intersection of edges between HDGN and ODGN to assess their mutual similarity. For biomolecular interactomes (PIN, MGN and FSGN) the nodal and edge intersection can be useful to explore the underlying molecular events of pathophenotypic similarities. However, biomolecular interactomes require two steps before the intersection analysis. First, we filter networks to ensure that both compared networks have only intersected nodes to minimize their differences in sizes (see schematic diagram of the process in Figure S1). All biomolecular interactomes were filtered to have genes with pathophenotypic data. Hence, we generated three biomolecular sub-networks that contain uniquely genes (nodes) participating in PSGN (Figure S1 and Table S6). This first step was essential for a more accurate value of the significance in the mutual coverage and to reduce the noise in the intersected edges. Moreover, this problem is bidirectional, so we used three different filters for PSGN (one for each cellular network). It will merge in three PSGN sub-networks (Figure S1 and Table S6). To evaluate the significance of the network comparison, we compared PSGN sub-networks with their respective randomized biomolecular interactome, treated and filtered exactly as the original networks. These randomizations were carried out preserving the node connectivity distribution in the respective cellular networks. Subsequently, we used NeAT [28] to compare networks treated as undirected ones. We used different metrics to identify the significance of the intersection: Maximal number of edges in the union, Jaccard coefficient and hypergeometric probability (p-value) [28,40].

Network topological analysis. All gene (node) degrees were calculated for each pathological network and biomolecular interactome, using the iGraph software. Subsequently, a non-parametric test was used to study in each subset of genes the distributions of the node (gene) degree, the number of associated pathophenotypes per gene and the mean value of pathophenotypic similarity per gene. More precisely, a Mann-Whitney test was used to assess the significance of these distributions for gene subsets with the distributions of all genes in PSGN and their respective disease-causing gene network. This non-parametric test was run 1000 times for every subset of genes using a different random sample in each test. These random samples conserved the same size (number of genes) as their respective subset in the correspondent network. Subsequently, we calculated the mean p-value of all runs for every subset. Additionally, a Spearman's rank correlation test ($\alpha = 0.05$) was used to analyze the degree of genes in HDGN, ODGN, PIN, MGN and FSGN with respect to the number of pathophenotypic relationships in PSGN.

Performance validation and ROC calculations. A binary classification system was used to analyze the performance of intersected interactions between different cellular networks (PIN, MGN, FSGN) and phenotypic interactions in PSGN. This binary classification is based on signal detection theory, using a receiver operating characteristic (ROC) analysis [41]. We compared biomolecular interactomes and their respective randomized versions (similar to those ones used in the network comparison analysis) with the PSGN using phenotypic similarities as the value of the signal (Figure S1). ROC curves were obtained considering the intersected interactions of PSGN with biomolecular interactomes as True Positives and those of PSGN with random biomolecular interactions as False Positives (Figure S1). We used randomizations to generate a dataset of False Positives proportional to the number of obtained True Positives for each biomolecular interactome. This procedure was useful to increase the confidence of the ROC analysis. In addition, we calculated the average area under the curve (AUC) for each interactome, calculating about 20 ROC curves following this same procedure.

Results and Discussion

Comprehensive Classification of Disease-to-gene Associations Contained in Currently Available Diseasomes

The projection in networks of the genetic associations data, available in OMIM and Orphanet, shows different patterns of connectivity among diseases and mutated genes (Figure 1A). Thus, we proceeded to build updated versions of existing models of disease networks, the "human disease network" (HDN) [10] and the "orphan disease network" (ODN) [11]. Subsequently, we classified all disease-gene associations of HDN and ODN in order to get an insight regarding their global distribution. For this purpose, we retrieved a total of 2525 and 2331 genes from HDN and ODN, respectively. Each gene dataset was subdivided in four different classes (Tables S3 and S4 for HDN and ODN respectively) according to our proposed criteria (Figure 1C): two monotropic classes (MD-MG and PD-MG) and two pleiotropic classes (MD-PG and PD-PG). Monotropic subsets are exclusive because their relationship with the disease is unique so genes take part in only one subset and they represent 72% and 69% of the total genes in HDN and ODN, respectively. In contrast, pleiotropic genes can be related to monogenic as well as to polygenic diseases so they can be present in both pleiotropic subsets.

The abundance of genes in each subset indicates how genetic association studies tend to distribute genes with different degrees of specificity for pathologies. In both networks, monotropic genes are found to be the most abundant ones, irrespective of the actual number of genes involved in the diseases (Table 2). For instance, "biunivocal" genes (MD-MG subset genes) represent over 56% and 30% of HDN and ODN genes respectively (Table 2). Even more, genes included in the PD-MG class are the most abundant ones in orphan disease network reaching 39% of the total genes. Many PD-MG associations could involve highly co-regulated genes (i.e. coding genes for different subunits of multi-protein complexes), so these genes can be considered a whole functional unit. In this case, we suspect that biunivocal relationships might be underestimated.

The ratios of diseases per gene agree with a pathological convergence (exclusive associations) and divergence (non exclusive associations) for monotropic and pleiotropic genes respectively (Table 2). These results are obvious taking into account our classification criteria. However, they provide a panoramic view of how a set of clinical features (pathophenotypes) observed in patients reach consensus and are attributed to a disease. These

Table 2. Distribution of disease-to-gene associations on proposed classification.

	Human Diseases Network		Orphan Disease Networks	
Subset	**Diseases per gene**	**Genes (%)**	**Diseases per gene**	**Genes (%)**
MD-MG	1.00	1431 (56.7)	1.00	717 (30.8)
MD-PG	2.57	639 (25.3)	2.71	435 (18.7)
PD-MG	0.46	379 (15.0)	0.40	908 (39.0)
PD-PG[a]	2.13	371 (14.7)	1.68	584 (25.1)
All genes[b]	1.24	2525 (100)	0.91	2331 (100)

[a]Pleiotropic genes associated with at least one polygenic diseases.
[b]All genes in HDN and ODN respectively.
doi:10.1371/journal.pone.0056653.t002

results seem to show a human annotation bias that can affect the current disease classifications.

Features of Disease Causing Gene Networks (Unipartite Projections)

From the bipartite projections (disease-to-gene) of HDN and ODN, we built their corresponding unipartite projections (gene-to-gene) (as can be seen in Figure 1A), named as "human disease causing gene network" (HDGN) and "orphan disease causing gene network" (ODGN) respectively (Figures 3A and 3B). Both unipartite projections are based on the emergence of gene-to-gene relationships (edges) inferred from pair of genes co-associated with at least one disease (Figure 1A). Accordingly, all genes in the MD-MG subsets and those uniquely associated with monogenic diseases in MD-PG will appear as unconnected genes in unipartite projections (HDGN and ODGN).

HDGN include 749 genes (nodes) and 2654 inferred gene-gene relationships (edges) among them (Figure 3A and Table S1). However, ODGN is twice as larger as HDGN with 1492 genes and 6380 inferred gene-gene relationships (Figure 3B and Table S2). At first glance, the topological structures of unipartite networks (HDGN and ODGN) are quite similar (Figures 3A and 3B) although an enrichment of unconnected nodes in HDGN is clear when compared to ODGN (1776 and 839 for HDGN and ODGN respectively). This enrichment is mainly due to the higher number of biunivocal relationships (MD-MG) in HDGN (Table 2). Therefore, this is the reason why HDGN shows fewer inferred relationships (2654) than ODGN (6380).

We carried out an analysis of the intersection between both unipartite networks (HDGN and ODGN) to assess an estimation of their similarity. But first we removed all unconnected nodes because they were not considered structural components of these networks. The resulting intersection was 481 genes (intersected nodes) and 662 inferred gene-gene relationships (intersected edges) corresponding to 24% and 10% of edges in HDGN and ODGN respectively (Table 3). Both networks show a Jaccard coefficient of similarity (number of edges in the intersection divided by the number of edges in the union) of 7.9% (Table 3). Surprisingly, the similarity is lower than expected *a priori* which indicates strong differences between the two data sources (OMIM and Orphanet).

These results reinforce the hypothesis that the absence of a systematic procedure in the phenotypically characterization of genetic diseases will affect the utility of network medicine methods. In particular, it leads to the isolation of genes and diseases from their real pathological processes, making it practically impossible to identify groups or subgroups of related pathologies. This observed tendency to the exclusiveness (that is to say, the abundance of monotropic gene-disease relationships) considerably increases the disease-gene association specificity that may be of interest for genetic testing.

Features of Pathophenotypic Similarity Gene Network (PSGN)

The exclusiveness mentioned above could affect pathological processes with many disease variants. In the case of these diseases,

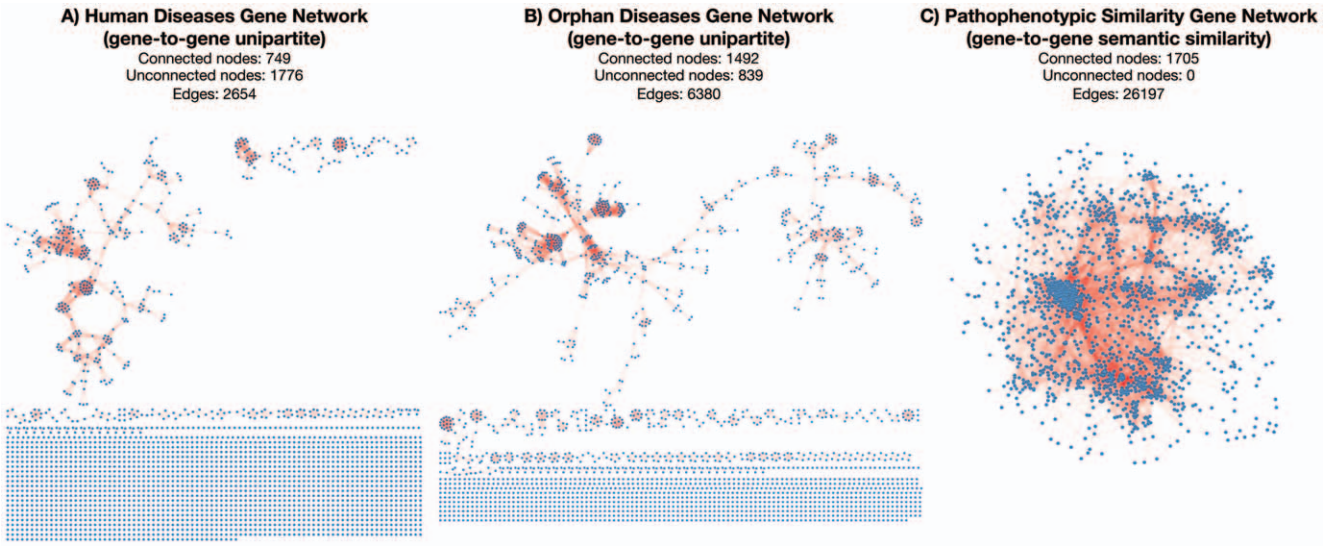


Figure 3. Unipartite gene-to-gene projections of the disease networks and the pathophenotypic similarity gene network. Human diseases genetic network (HDGN in panel A), Orphan diseases genetic network (ODGN in Panel B) and Human pathophenotype similarity gene network (PSGN in panel C). PSGN consists of one connected component (with a few unconnected genes), in contrast to HDGN and ODGN that show a great variety of isolated patterns of association. All unconnected genes (nodes) correspond to those uniquely associated with monogenic diseases, all of them were excluded in unipartite projections.
doi:10.1371/journal.pone.0056653.g003

Table 3. Network intersection analysis between HDGN and ODGN.

Network features	Values
Number of nodes in HDGN	749
Number of nodes in ODGN	1492
Number of edges in HDGN	2654
Number of edges in ODGN	6380
Observed nodes in the intersection	481
Observed edges in the intersection	662
Percentage of edges in HDGN	24.94
Percentage of edges in ODGN	10.38
Jaccard coefficient of similarity	0.079[a]

[a]Fraction of edges in the intersection respect to the total edges in the union.
doi:10.1371/journal.pone.0056653.t003

some genes play a primary role in the progression of the pathology but others modulate the phenotypic variability.

To tackle this problem, HPO offers possibilities for a formal study of the pathophenotypic relationships among genes on the bases of their semantic similarities (pathophenotypic similarities). Therefore, we defined the pathophenotypic space of each gene, consisting of the set of HPO terms associated with the gene (as shown in Figure 1B). These spaces were described using only specific HPO terms, those farthest terms from the root of the ontology, to calculate the semantic similarity value between every two given genes (see methods, Table S7). Higher values of semantic similarity indicate greater specificity in the common pathophenotypic space between a pair of genes. It is known that ontology-based phenotypic similarity methods can also contribute to improve disease-causing gene networks based on phenotypic information built with text-mining analysis [42] or random-walk trajectories between genes considering the ontology as a simple graph [43].

From all calculated pathophenotypic similarities greater than zero, we selected the top 2% of more significant pairs of genes. This selection provides the pathophenotypic similarity gene network (PSGN) with 1075 genes and 26197 gene-to-gene pathophenotypic similarities (Figure 3C and Table S7). Disease-causing gene networks (HDGN and ODGN) exhibit explicit structural differences when they are compared to PSGN (Figure 3); for instance, PSGN consists of only one giant connected component (Figure 3C), which is not the case for HDGN and ODGN.

Almost all the pathophenotypic gene annotations used in HPO originally come from OMIM and they represent the sum of all clinical features of diseases associated with a gene. Accordingly, the pathophenotypic similarity for a gene is somehow dependent on the number of diseases associated with this gene (see methods section). Hence, we proceed with a comprehensive study to assess whether the pathophenotypic similarity can be used to reinterpret the pathological relationships between genes (see supplementary methods and discussion in Methods S1).

Pathophenotypic Similarity Reveals a New Understanding of Pathological Relationships

The survey of the mutual coverage between PSGN and each unipartite projection (HDGN and ODGN) was carried out with an analysis of their intersections.

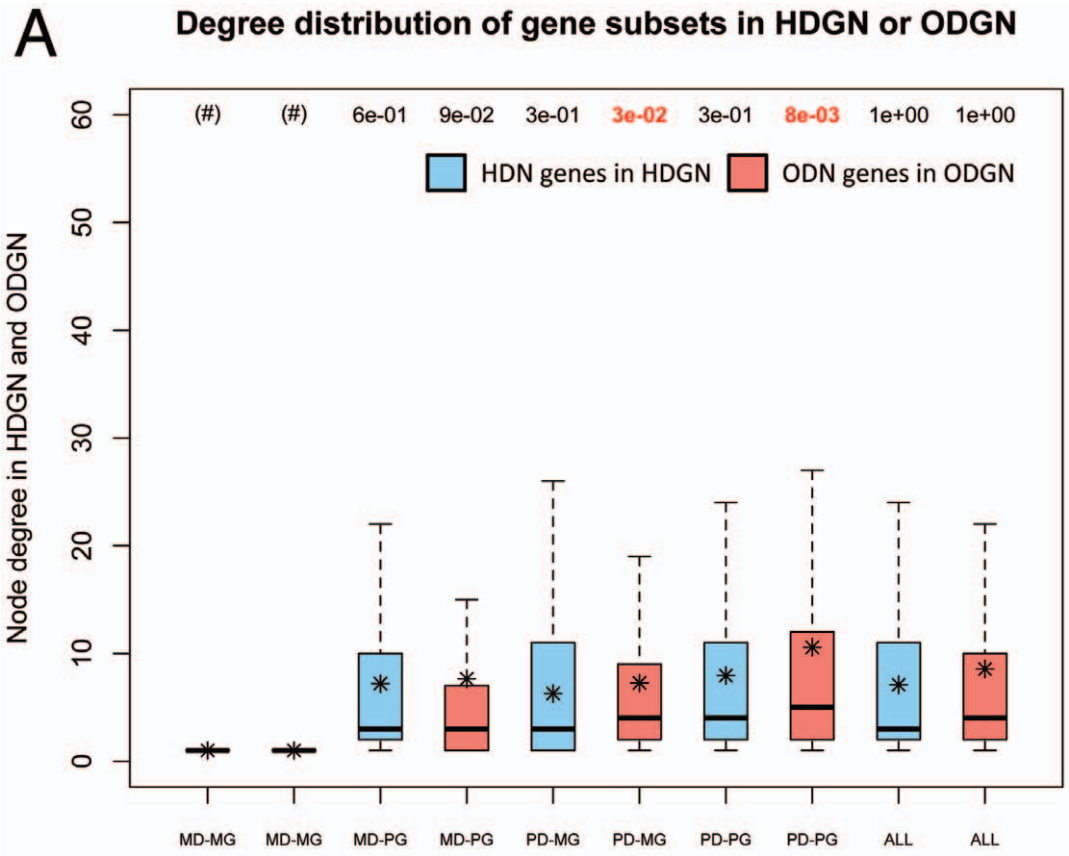
The resulting intersections of PSGN with each unipartite projection proved 528 shared nodes and 1055 shared edges for HDGN and 931 and 1669 for ODGN (Table 4). Therefore, 39% and 26% of inferred pathological relationships intersect with pathophenotypic similarities of PSGN, even improving the intersection between disease causing gene networks (mentioned above). The Jaccard coefficient of similarity of the intersection of PSGN with each pathological network was 3.8% and 5.4% for HDGN and ODGN respectively (Table 4). This can be considered an interesting performance value if we take into account the dependence on the Jaccard coefficient on the different sizes of compared networks (the number of edges in the union are 27796 for HDGN and 30908 for ODGN). Furthermore, there are about 25000 new pathophenotypic similarities, excluding inferred pathological relationships, to be used for the discovery of new underlying pathological relationships among genes.

Topological analysis exhibits the emergence of unnoticed pathological relationships. We have also studied how genes in PSGN are distributed in comparison to HDGN and ODGN. Subsequently, we analyzed the degree distribution of genes for each network (HDGN, ODGN and PSGN), as well as for their respective gene subsets (MD-MG, MD-PG, PD-MG and PD-PG of HDN and ODN). We carried out a Mann-Whitney test to assess the significance of the difference of the degree distribution of each subset in their respective disease-causing gene network and in PSGN (Figure 4, a boxplot was used in all the cases). In agreement with our classification criteria, MD-MG genes (bi-univocal) have null connectivity in their respective disease-causing gene networks (Figure 4A). By contrast, MD-MG genes are phenotypically linked to a mean of 25 genes in PSGN indicating an expansion of pathophenotypic relationships between disease-causing genes in PSGN (Figure 4B). In pathological networks, degree distributions are significantly different for ODGN subsets (PD-MG and PD-PG) but not for HDGN subsets (see their correspondent p-values in Figure 4A). On the other hand, degree distributions in PSGN are quite similar when compared to the equivalent subsets of HDGN and ODGN, where higher node degree for pleiotropic genes and lower for monotropic genes can be appreciated (Figure 4B). In addition, Spearman's rank correlation test was used to explore degree correlations between the pathophenotypic similarity (PSGN) and disease-causing gene networks (HDGN and ODGN) (Table S8). Weak (but statistically significant) positive correlations were found between gene pathological and pathophenotypical relationships (Table S8). These results, as shown in Figure 4 and

Table 4. Network intersection analysis between PSGN and HDGN or ODGN.

Network features	HDGN values	ODGN values
Number of nodes in PSGN	1705	1705
Number of nodes in pahtological network	749	1492
Number of edges in PSGN	26197	26197
Number of edges in pahtological network	2654	6380
Observed nodes in the intersection	528	931
Observed edges in the intersection	1055	1669
Percentage of edges in PSGN	4.03	6.37
Percentage of edges in pahtological network	39.75	26.16
Jaccard coefficient of similarity	0.038[a]	0.054[a]

[a]Fraction of edges in the intersection respect to the total edges in the union.
doi:10.1371/journal.pone.0056653.t004



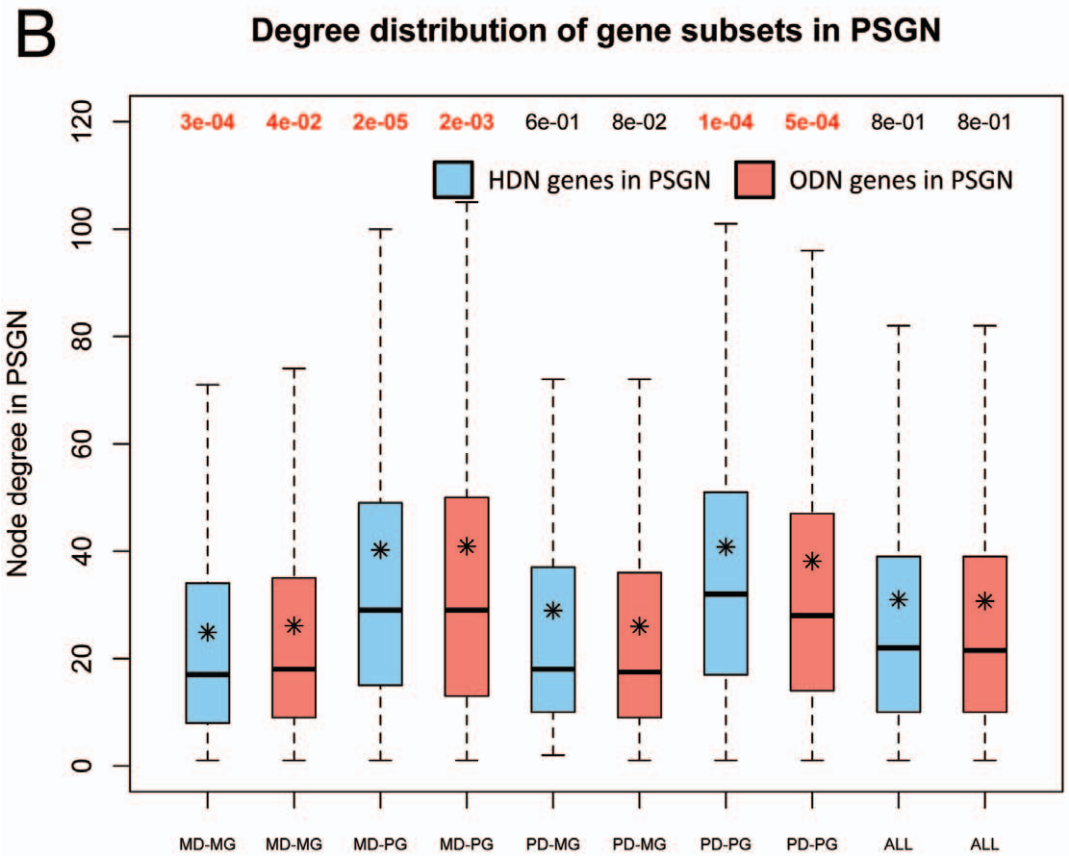


Figure 4. Degree distribution of subset genes in pathological and pathophenotypic gene-to-gene networks. Box plots of the degree of subset genes in HDGN (blue) and ODGN (red). Box plots of the degree of subset genes in PSGN for ODN subsets (blue) and for ODN subsets (red). In bold and red, significant p-values. (*) Mean values. (#) Subsets of completely unconnected genes.
doi:10.1371/journal.pone.0056653.g004

Table S8, clearly show that gene degrees in pathological networks differ from those calculated using pathophenotypic similarities.

The (apparently) most striking observation is that genes uniquely associated with monogenic diseases (genes in MD-MG and many of MD-PG) are present in PSGN. The vast majority of these genes appeared as unconnected genes in the unipartite projections of HDN and ODN (as shown in Figure 3). This means that pathophenotypic similarities lead to the emergence of novel relationships that remained hidden in the gene-to-gene projections of current diseasomes.

Specific contribution of gene subsets to gene-to-gene pathophenotypic similarities. In light of the result discussed above, we consider it necessary to prove the contribution of each type of gene subset to the gene-to-gene similarities of PSGN. This could help to unveil the relationship between the pathological convergences or divergences and the pathophenotypic similarities [30]. Therefore, we analyzed the abundance of pathological phenotypes and the average pathophenotypic similarity per gene.

Figure 5 (panels A and B) represents the distribution of the abundance of pathophenotypes (HPO terms) in genes for HDN and ODN subsets. Pleiotropic genes show distributions significantly different to the distribution of all genes included in PSGN using a Mann-Whitney test (see their correspondent low p-values for MD-PG and PD-PG, Figure 5 panels A and B). On the other hand, monotropic genes seem to be well represented in the pathophenome (whole genes of PSGN) showing only slight differences in the distribution of PD-MG subset for ODN (see the p-value for PD-MG in Figure 5B). Consequently, we can be confident that the phenotypic descriptions used for monotropic genes are not underestimated and they are enough to calculate their pathophenotypic similarities to other genes. By contrast, as expected, pleiotropic genes tend to be annotated in the ontology with more clinical features compared to the whole gene annotations. For an overall estimation of how each subset contributes to the pathophenotypic co-dependence between genes, we calculated the average of pathophenotypic similarity values associated with each gene in the PSGN in order to compare their distributions in different subsets (Figures 5C and 5D). The monotropic subsets contain genes with the highest specific relationships to diseases. Nevertheless, monotropic subsets show very different behavior compared to all genes of the PSGN in the distribution of the average pathophenotypic similarities related to genes within HDN and ODN subsets (see the low p-values for MD-MG and PD-MG in Figures 5C and 5D). MD-MG subsets show lower average pathophenotypic similarity values (Figures 5C and 5D). As a result, these distributions also reveal pathophenotypic relationships among genes that remained lost in the gene-to-gene unipartite projections of HDN and ODN. The distributions of PD-MG subsets show higher average phenotypic similarities between genes (observe that the green curves in Figures 5C and 5D are displaced to the right when compared to the respective red curves, as well as to the rest of curves). This observation could be mainly due to the fact that they are sharing similar sets of annotations, and in many cases they are functional units or strongly co-regulated molecular complexes. With regards to pleiotropic subsets, they seem to be slightly affected by the number of genes involved in the disease (monogenic and polygenic). Nonetheless, their abundance of pathophenotypes could increase the number of non-specific relationships between genes. In this case, non-specific relationships will tend to show low values of similarities decreasing the average value associated with genes. In fact, this agrees with the higher connectivity observed for pleiotropic subsets in both HDN and ODN (Figure 4). For this reason, we analyzed the degree of association between the abundance of pathophenotype per gene and the average similarity value per gene. A weak Spearman correlation was obtained (p-value 1.8E−26 and $r_s = -0.25$, Figure S2) so we can ensure no clear dependence between both parameters. However, there is a tendency to decrease the mean value of pathophenotypic similarity for genes with abundant HPO terms annotations.

Apparently, the use of semantic similarity measurements produces a rearrangement in the pathophenotypic co-dependence between genes overcoming the bias that can be introduced from the original source of data, the Morbid Map. However, the gene pleiotropy dampens their average pathophenotypic similarity values indicating a rise of unspecific relationships with other genes compared to monotropic genes. This observation reinforces our suggestion that the representation of diseasomes as unipartite projections is insufficient to study other underlying (and not necessarily obvious) pathophenotypic relationships.

Overview of the Relationship between Metabolic or Essential Genes and Pathophenotypic Similarity

Taking into account that metabolic and essential disease genes represent about 18% and 34% respectively of the total disease-causing genes, we also studied how they are represented in each subset of genes in our classification (Table 5). The subsequent study of cumulative frequencies per gene of the associated pathopenotypes (Figure 6A) and the average pathophenotypic similarity values (Figure 6B) suggest that gene subsets tend to be associated with different biological properties.

Enrichment of metabolic genes in the MD-MG subclass. Biunivocal classes (MD-MG) are markedly enriched in metabolic coding genes with respect to the other classes; on the contrary, PD-MG is underrepresented by metabolic enzymes. On the other hand, the pathophenotypes corresponding to metabolic genes do not differ from those of the whole pathophenome (see non-significant p value in Figure 6A). However, the mean value of phenotypic similarity is lower for metabolic genes than for the whole pathophenome (Figure 6B). For instance, metabolic genes tend to be involved in more specific pathological processes and exclusively related to pathophenotypes recognized as genetic diseases. It seems relevant that metabolic genes are mainly enriched in the MD-MG subset: 67% and 49% of the whole set of genes in MGN are MD-MG for HDGN and ODGN, respectively. In addition, metabolic genes show a lower distribution of the mean values of pathophenotypes compared to the whole pathophenome (Figure 6). Therefore, dysfunctions in metabolic genes prove a functional bias in disease and gene association studies toward the pathophenotypic specificity (Figure 6). At least two factors could contribute to explain this observation: first, the molecular basis of metabolic dysfunctions can be more precisely identified in these diseases; second, these diseases exhibit pathophenotypes with highly distinguishable features. In any case, both factors can be influenced by the application of routine biochemical analysis in the clinical setup, which allows an easier detection of abnormal concentrations of metabolites in blood or urine.

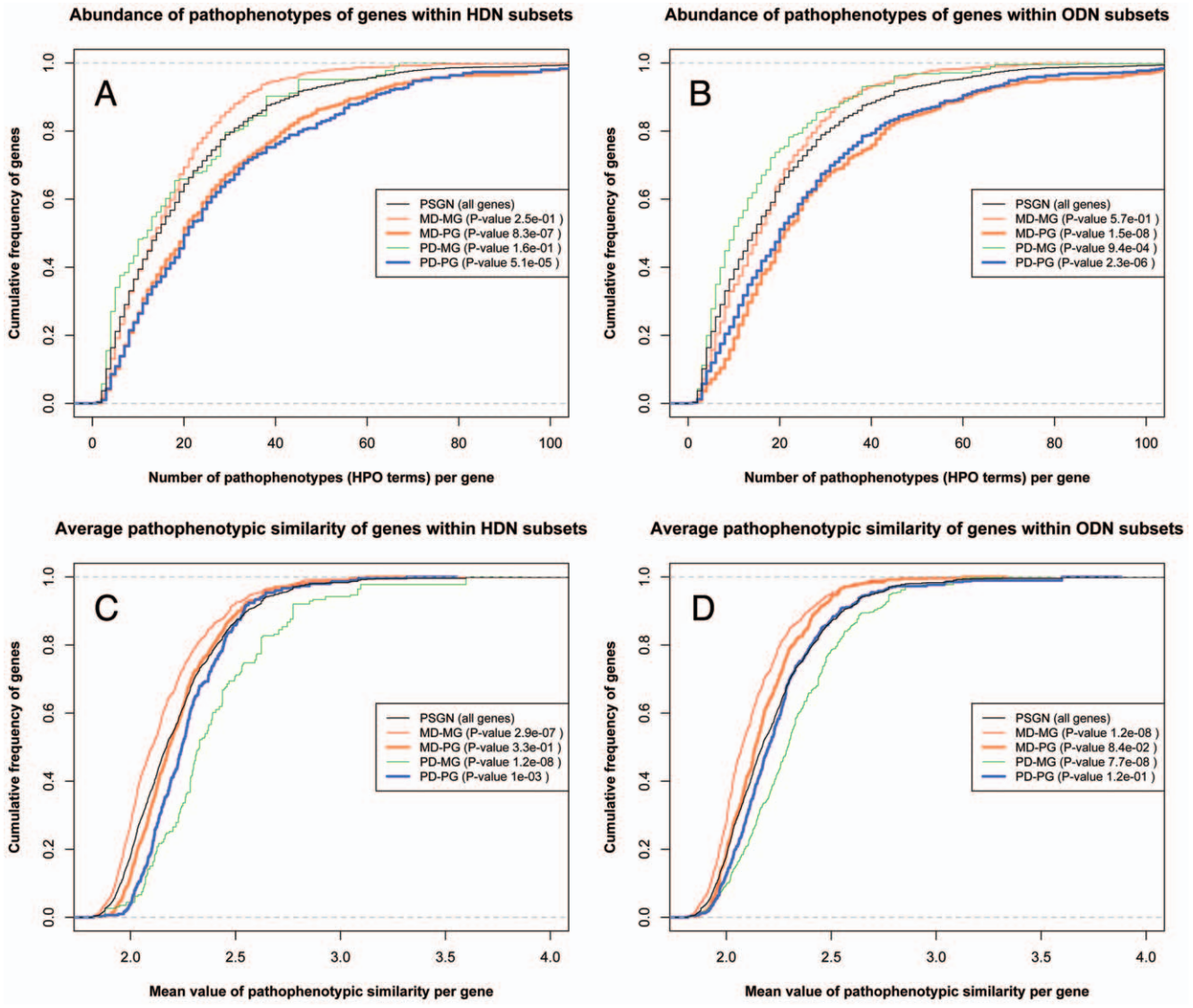


Figure 5. Distributions of the number of pathophenotypes and pathophenotypic similarities in each subset. MD-MG (red line), MD-PG (orange line), PD-MG (green line), PD-PG (blue line) and PSGN (Black line). Upper panels represent the cumulative frequency of the number of specific pathophenotypes annotated for genes in HDN (C) and ODN (D) subsets, the whole set of genes in HPO (PSGN) was used as the reference distribution. Lower panels represent the cumulative frequency of the average pathophenotypic similarity associated with genes in HDN (C) and ODN (D) subsets, the whole set of genes in HPO (PSGN) was used as the reference distribution. The p-values, included in each legend, represent the mean of the resulting p-values after 1000 non-parametric tests (Mann-Whitney test) where each subset was compared, each time, with a random sample of the pathophenome of the same size of the subset (see methods).
doi:10.1371/journal.pone.0056653.g005

Enrichment of essential genes in the pleiotropic subsets. Zhang et al. [11] have reported an enrichment of essential genes in ODN with respect to HDN but our results suggest that both networks show a similar proportion of essential genes (Table 5). In particular, the results shown in Table 5 also indicate that an enrichment of essential genes is produced in pleiotropic gene subclasses. The number of pathophenotypes associated with essential genes is significantly higher than that obtained when using all genes in the PSGN (Figure 6A). But their distribution of mean values of phenotypic similarities is statistically indistinguishable from that of the whole pathophenome (Figure 6B). Some previous network medicine works have discussed how essential genes are represented in different diseaseomes [10,11,30]. Barabasi and co-workers concluded that disease-causing genes are not essential genes because their associated lethality could have severe consequences [10]. Chavali et al. [30] proposed two different topological features for phenotypically divergent genes and essential disease genes, inter-modular and intra-modular hubs respectively. Zhang et al. [11] in their analysis of the orphan disease network found that ODs are enriched in essential genes as compared with the whole set of diseases. In contrast, when we compared the same essential gene dataset used by these authors in the updated versions of HDN and ODN, no detectable differences were found (Table 5). Our observation differs from that of Zhang et al. [11], maybe due to the use an updated version of both disease-causing gene networks and the same dataset of essential genes. In any case, our results do not support the idea that there could be a negative correlation between gene essentiality and disease prevalence. Nonetheless, it seems that there is a certain enrichment of essential genes in the

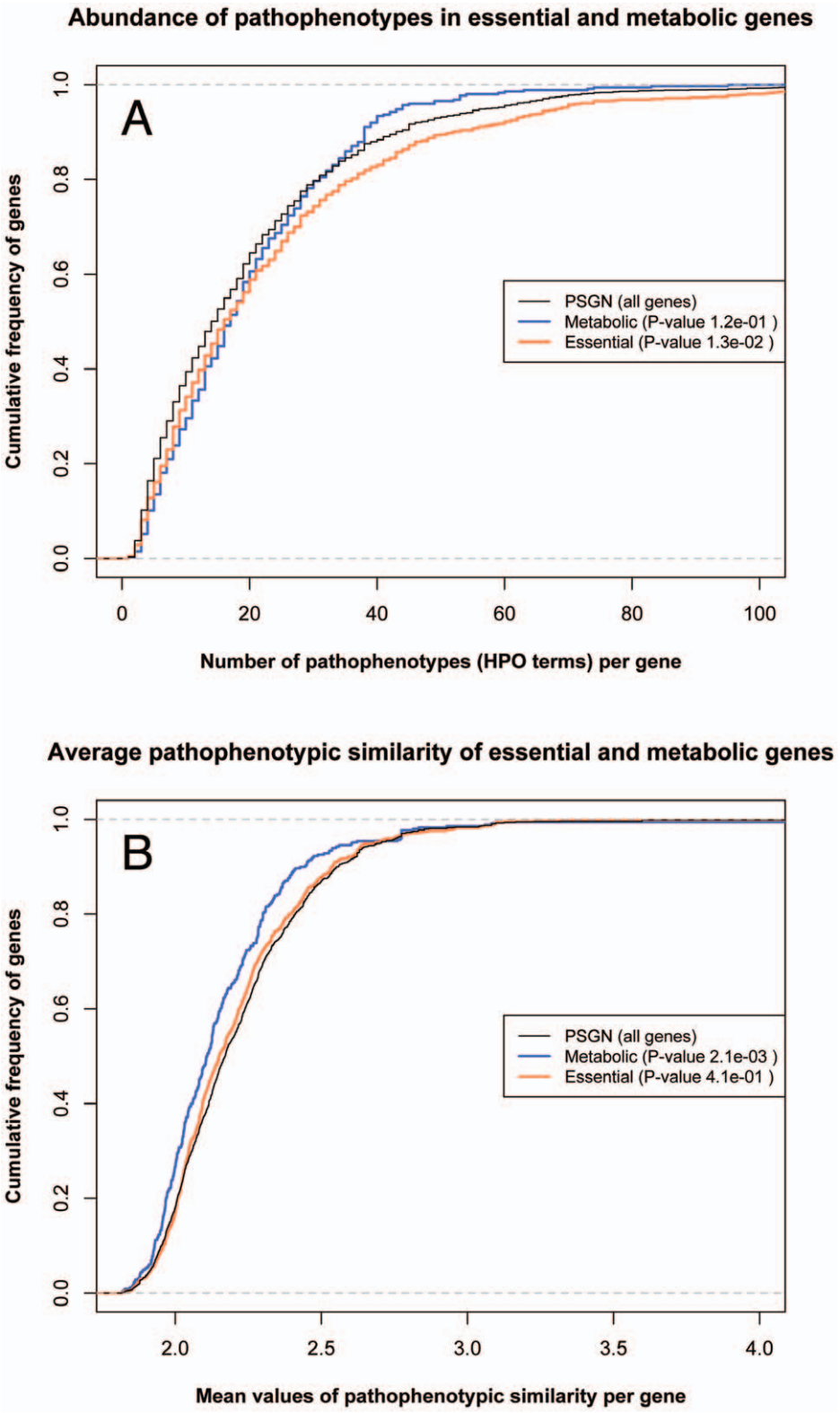


Figure 6. Distributions of the number of pathophenotypes and the pathophenotypic similarities for metabolic and essential genes. Metabolic genes (orange line), essential genes (orange line) and the PSGN (Black line). Upper panel (A) represents the cumulative frequency of the number of specific pathophenotypes annotated for genes, the whole set of genes in HPO (PSGN) was used as the reference distribution. Lower panel (B) represents the cumulative frequency of the average pathophenotypic similarity associated with genes, the whole set of genes in HPO (PSGN) was used as the reference distribution. The p-values, included in each legend, represent the mean of the resulting p-values after 1000 non-parametric tests (Mann-Whitney test) where every set of metabolic and essential genes was compared, each time, with a random sample of genes in PSGN of the same size of their respective set (see methods).
doi:10.1371/journal.pone.0056653.g006

Table 5. Distribution of essential and metabolic genes in current diseases network.

	HDN		ODN	
	Essential	Metabolic	Essential	Metabolic
Subset	genes (% in class)	genes (% in class)	genes (% in class)	genes (% in class)
MD-MG	409 (28.6)	308 (21.5)	219 (30.5)	202 (28.2)
MD-PG	315 (49.2)	79 (12.4)	228 (52.4)	64 (14.7)
PD-MG	106 (28.0)	65 (17.2)	245 (27.0)	105 (11.6)
PD-PG[a]	189 (50.9)	34 (9.2)	286 (49.0)	73 (12.5)
All genes[b]	856 (33.9[c])	458 (18.1)	802 (34.4[c])	409 (17.6)

We determined for each class the percentage of genes considered as essentials and metabolic coding genes included in the built metabolic network (MGN).
[a]Pleiotropic genes associated with at least one polygenic diseases.
[b]All genes in HDN and ODN respectively.
[c]Minimal changes are seen compared to Zhang et al.(2011) [11], these differences are due to updating of data Orphanet.
doi:10.1371/journal.pone.0056653.t005

subsets of ''pleoiotropic'' genes, that is, those associated with more than one disease (Table 5). This result agrees with observed by Chavali et al. in the dataset of shared genes by diseases [30]. The dataset of essential genes used in these works [10,11,30] are human orthologous of lethal mouse genes catalogued in the Mouse Genome database [44].

From our point of view, the enrichment of essential genes in pleiotropic disease-causing genes leads to interesting evolutionary questions on how mutations in these genes are related to their lethality for other mammals and might be involved in the limits of human evolvability [45,46].

Integrative Analysis of PSGN

Built biomolecular interactomes (PIN, MGN and FSGN). The heterogeneity of the cellular interactions among genes affects (either directly or indirectly) the progression of the diseases [13]. Thus, the disturbances caused by genetic mutations can be transmitted in biological systems in several distinct ways. Three different biomolecular interactomes were built to study the association between the pathophenotypic similarity and each type of biological interaction (physical, metabolic and functional interactions). PIN results in 9580 genes connected through 74657 physical interactions (Table S5). MGN contains 535 enzyme-coding genes interconnected by 9812 flux correlations (Table S5). The top 0.5% of functional similarities in the branch of biological processes in the Gene Ontology corresponds to FSGN. FSGN results in 9157 genes and 496973 significant functional similarities (Table S5). For each biomolecular interactome, we evaluated their coverage in PSGN and the contribution of each type of biological interaction to the score of pathophenotypic similarity.

Network comparison analysis between biomolecular interactomes and PSGN. 'A network intersection analysis was carried out using the PSGN as reference and the biomolecular interactomes (PIN, MGN or FSGN) as queries. Nevertheless, the observed differences in size and density of the studied networks could be the cause that the direct network comparison analysis would provide no useful significance values. Therefore, we decided to standardize the contents of the networks by using the intersection of nodes (see methods section) to minimize differences between the reference (PSGN) and the rest of the networks (PIN, MGN or FSGN). This step (Figure S1) provoked a strong structural decomposition from all the original networks that resulted in sub-networks (Table S6). Although we reduced the size differences between the intersected networks, other features are still preserved like the density of edges, which are inherent to the nature of each network (Table 6).

The network comparison results show statistically significant intersections of edges for all biomolecular interactome sub-networks compared to their respective PSGN sub-network (Table 7). This was not the case for randomized networks used as negative controls. The hypergeometric test shows a lower significance of the pathophenotypic similarities resulting in the intersection between PSGN and MGN when compared to PIN and FSGN (Table 7). Nevertheless, the Jaccard coefficient of similarity between biomolecular interactomes and their respective PSGN sub-network was higher for MGN and FSGN (9.8% and 5.4% respectively) than for PIN (2.5%). In this sense, both the percentage of edges remaining in the reference sub-network and the Jaccard coefficient of similarity seem to be good indicators of the size of the phenotypic space covered by the intersection (Table 7). The 23.7% of physical interactions between diseases-causing genes match with pathophenotypic similarities, 11.7% and 8.1% for metabolic flux correlation and functional interactions respectively. FSGN showed the largest and most significant coverage in PSGN (Table 7), which means that the functional relationships of genes based on biological processes define the broadest context of the molecular mechanisms associated with disease-causing genes. Concerning biochemical interactomes (PIN and MGN), PIN exhibits a greater coverage of genes at the intersection than MGN, although the latter presents the highest Jaccard coefficient of similarity (Table 7).

Specific contribution of biomolecular interactions to pathophenotypic similarities. Most of the published network biology studies have made use of the degree of a node (number of connections with other nodes) to assess its relevance in a network. In fact, node degree has been extensively used in physical interaction networks [10,11,30,31] but also in metabolic networks [15,32]. In this work, a topological analysis was carried out in different biomolecular interactomes to calculate the degree of genes (based on gene-to-gene interactions).

To estimate whether the abundance of biological interactions for genes is correlated with the number of phatophenotypic similarities in PSGN, we carried out a Spearman's rank correlation test of gene degrees. This test showed weak, but statistically significant, positive correlations between gene degrees for the whole set of genes (p-value = 2.0E−07, r = 0.15 for HDN; p-value = 3.2E−08, r = 0.16 for ODN) when PIN was compared to PSGN. No significant correlations were found when either MGN or FSGN were compared to PSGN (Table S9). The values

Table 6. Counts of nodes and edges in the comparison of PSGN and biomolecular interactomes.

		PIN		MGN		FSGN	
Symbol	**Description**	**Nodes**	**Edges**	**Nodes**	**Edges**	**Nodes**	**Edges**
R	Reference (PSGN)	1233	15550	131	321	1381	17233
Q	Query (biomolecular interactome)	903	1779	154	1060	1376	30318
QvR	Union	1240	16907	158	1257	1387	45078
$\hat{Q}R$	Intersection	896	422	127	124	1370	2473
Q!R	Query not reference	7	1357	27	936	6	27845
R!Q	Reference not query	337	15128	4	197	11	14760

All calculations were performed using NeAT [28]. The query is PSGN and used reference corresponds to each biomolecular interactomes.
doi:10.1371/journal.pone.0056653.t006

for the different subsets obtained in this analysis clearly show that only physical interactions bear some relation with the abundance of pathophenotypic similarities in pleiotropic genes associated with monogenic diseases (MD-PG). Accordingly, mutations in MD-PG genes seem to "diverge" disturbances more efficiently by protein-protein interactions that determine a pathophenotypic and functional relationship between genes. This result suggests that these genes co-participate in different variants of a given disease and there are functional co-dependencies among them. Thus, we proceeded to assess whether the specificity of the pathophenotypic similarity between genes depends on their type of biological interaction. For that reason, we performed a validation analysis through receiver operating characteristic (ROC) curves to prove the signal in pathophenotypic similarities produced by each biomolecular interactome in PSGN (Figure 7). PIN and MGN showed higher average areas under the ROC curves (AUC values of 0.77 and 0.76, respectively) than functional interactions with an average AUC of 0.66 (Figure 7). Both biochemical interactomes have a strong signal, as depicted by ROC far from the straight line representing randomness (Figure 7). This observation reinforces the idea that strong synergies occur between genes involved in biochemical interactions. The functional network (Figure 7) also shows a signal clearly departed from the straight line representing randomness that is consistent with previous works [27]. However, one should be aware that there is always some degree of nonspecific relationships that can introduce noise in this kind of analysis.

Merging modular components of MSUD using pathophenotypic similarity. We analyzed a metabolic disorder named as maple syrup urine disease (MSUD, MIM 248600). MSUD is a genetic disease grouped into aminoacidurias and caused by a decreased activity of the branched-chain alpha-ketoacid dehydrogenase (BCKD) complex. It catalyzes the first steps for the degradation of branched-chain amino acids (valine, leucine and isoleucine). This enzymatic complex has three subunits (E1, E2, and E3) encoded by four different genes BCKDHA-E1A (Entrez GeneID 593), BCKDHB-E1B (Entrez GeneID 594), DBT-E2 (Entrez GeneID 1629), and DLD-E3 (Entrez GeneID 1738). This inborn error of metabolism is genetically and phenotypically well characterized [47]. The classical clinical features associated with MSDU are: maple syrup odor in cerumen (hours after birth), increased levels of branched -chain amino-acids (valine, leucine and isoleucine), ketonuria, signs of deepening encephalopathy, coma and central respiratory failure. We retrieved a map of all pathophenotypes annotated for MSUD-causing genes (Figure S3). From PSGN, we retrieved all gene pairs including at least one of the MSUD causing genes, but before we removed a dense cluster linked to DLD due to Leigh syndrome (Figure 8 A). Some of the resulting genes also present direct or non-direct metabolic flux correlations with BCKDHA, BCKDHB, DBT or DLD (Figure 8 A) and most of them take part in different reactions of the valine, leucine and isoleucine degradation pathway

Table 7. Significance of the number of edges at the resulting intersection in the network analysis comparison.

			PIN		MGN		FSGN	
Symbol	**Description**	**Formula**	**Network**	**Random**	**Network**	**Random**	**Network**	**Random**
N	Nodes in the union	–	1240	1238	158	158	1387	1387
M	Max number of edges in the union	$M = N*(N-1)/2$	768180	765703	12403	12403	961191	961191
$E(\hat{Q}R)$	Expected edges in the intersection	$E(\hat{Q}R) = Q*R/M$	36.01	27.96	27.43	24.33	543.57	196.95
$\hat{Q}R$	Observed edges in the intersection	–	422	35	124	17	2473	194
Q (%)	Percentage of query edges	$perc\_Q = 100*\hat{Q}R/Q$	23.72	2.54	11.70	1.81	8.16	1.77
R (%)	Percentage of reference edges	$perc\_R = 100*\hat{Q}R/R$	2.71	0.23	38.63	5.30	14.35	1.13
Jac_sim	Jaccard coefficient of similarity	$Jac\_sim = \hat{Q}R/(QvR)$	0.0250	0.0021	0.0986	0.0137	0.0549	0.0069
P value	P-value of the intersection	$Pval = P(X >= \hat{Q}R)$	**4.0E−308**	1.1E−01	**2.7E−51**	9.6E−01	**1E−321**[a]	5.9E-01

All calculations were performed using NeAT [28]. The query is PSGN and used reference corresponds to each biomolecular interactomes. In bold, those significant p-values.
[a]The limit of precision for the hypergeometric test.
doi:10.1371/journal.pone.0056653.t007

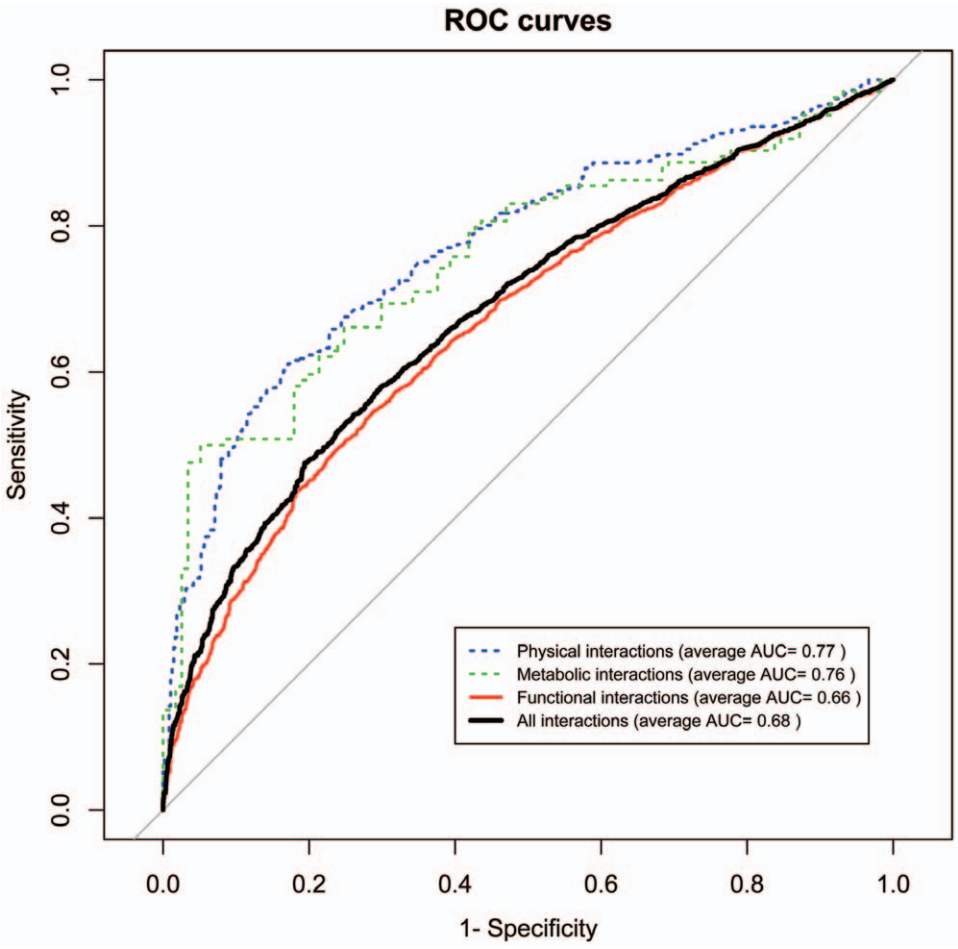


Figure 7. Receiver operative characteristic (ROC) curve performance by biomolecular interactions of pathophenotypic similarities. Physical interactions (dashed blue line), metabolic flux correlations (dashed green line), functional interactions (red continuous line) and an integrated interactome generated by the sum of all other interactomes (black continuous line). ROC curves were computed to assess the signal of pathophenotypic similarities for biological interactions. True positives (TP) were those interactions that where found in the intersection between PSGN and each biomolecular interactome (PIN, MGN and FSGN). The dataset of false positives (FP) was calculated from intersected gene pairs between PSGN and randomizations of each biomolecular interactome. We obtained severals different FP datasets to calculate the average area under the curve (AUC), it was 0.77 for PIN, 0.76 for MGN, 0.66 for FSGN and 0.68 for the integrated interactome. Only biochemical interactomes show significantly different AUCs to that of the integrated interactome (average p-values of 2.2E−6 and 4.1E−2 for PIN and MGN respectively).
doi:10.1371/journal.pone.0056653.g007

(Figure 8 B). This evidence that integrating functional co-dependencies and pathophenotypic similarities merge apparently non-related genes into a module of the molecular pathobiology. Furthermore, we can breakdown the module relationships to map shared pathophenotypes between genes (Figure 8 C). For instance, IVD and ACADM are genes included in MD-MG subsets for both HDN and ODN. However, in this sub-network (Figure 8 A) we detect that they are sharing pathophenotypes with MSUD genes (Figure 8 C). It is possible to identify the set of the most specific pathophenotypes for MSUD, elevated plasma branched chain aminoacids or hallucinations. In addition, PCCA and PCCB appear with similar clinical biochemistry parameters highly correlated with MSUD, such as high levels of lactic acid and ketone bodies (Figure 8 C). In contrast, other pathophenotypes point to disorders at a systemic or pathophysiological level, such as cerebral edema, pancreatitis, lethargy and coma (Figure 8 C). Nevertheless, these genes are grouped in the same biological context (Figure 8 B) and, it is important to remark, that all of them are in the mitochondrial matrix.

This metabolic syndrome illustrates the potentials of PSGN. This network provides novel pathological similarities between genes and outlines the pathobiology and functional context of disease-causing genes using metabolic interactions.

Overlapped physical and pathophenotypic interactions disregarded in unipartite projections. Finally, given the relevance of the physical interactions, we carried out a manual exploration of the intersection between PIN and PSGN. This is to remove all those gene-to-gene edges in both HDGN and ODGN from the resulting intersection. This resulted in the selection of all the disregarded relationships between genes in unipartite projections of diseasomes that are phenotypically and physically related (Figure 9 and Table S10). Therefore, tuning the balance between the "noise" and the confidence of interactions may improve the predictive power of new disease-related genes using network medicine approaches based on pathophenotypic term.

Conclusions

Current studies in medical genetics are mainly centered in establishing associations among diseases and genetic variations for

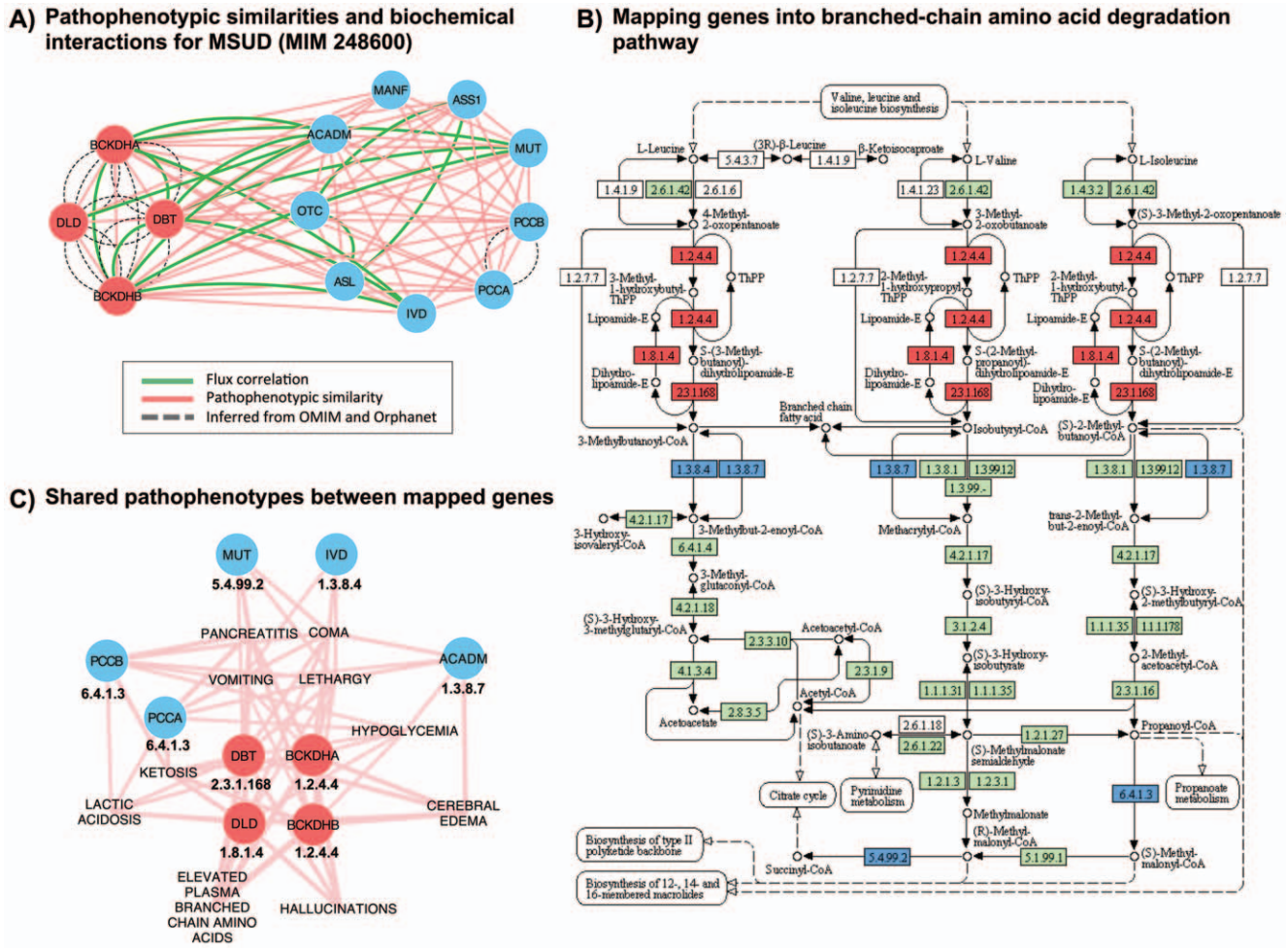


Figure 8. Maple syrup urine disease pathological and metabolic interactions. In red genes associated with MSUD and in blue pathophenotypic similar genes. (A) Pathophenotypic similarity gene sub-network for MSUD causing genes. It can be noteworthy that there are no inferred relationships between MSUD genes and the rest. (B) Map of branched-chain amino acid degradation pathway from. This map has been extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG, hsa:00280) developed by Kanehisa Laboratories. Enzymes encoded by human genes are in green. (C) Pathophenotypes shared between genes in the same metabolic module.
doi:10.1371/journal.pone.0056653.g008

personalized medicine. Many of these genetic variations are located in intragenic regions of DNA and they constitute the basic data to build disease-causing gene networks [10,11]. These networks are useful to find new genetic interactions between diseases, as well as to predict the influence of gene functions in existing pathologies [48–50]. In the present work, we have classified the different patterns of gene-disease associations in four subsets according to two different criteria (MD-MG, MD-PG, PD-MG, PD-PG, as depicted in Figure 1C). This is in contrast to previously published works in which only one criterion was used, either specific and shared genes by diseases [30] or monogenic or polygenic disease-causing genes [31,51]. Our findings indicate that the inferred associations are insufficient to describe properly both interactions among diseases and among genes. This effect can be easily observed when analyzing bipartite graphs composed of gene-to-disease edges. In these networks, more than 30% of the genes participate in "bi-univocal" relationships (that is, genes associated exclusively with a single disease). This specificity can be useful for diagnostics, but it makes it more difficult to establish groups or to identify interactions among diseases. On the other hand, our results have also uncovered an enrichment of metabolic genes in bi-univocal subsets, as well as an enrichment of essential genes in pleiotropic subsets. The lack of cellular and molecular phenotyping platforms constrains the possibility to detect shared features among pathologies. Consequently, this reduces the possibilities of generating new knowledge on the molecular bases of the pathophenotypic profiles, to distinguish classes and subclasses of a given disease more precisely [7,11,26]. However, medical semantics remains the standard tool to establish the sets of observed clinical features associated with pathologies. In the case of diseases with predominantly genetic origins, pathophenotypes are usually very conserved among patients. We have shown that pathophenotypic similarity gene networks can be a great resource to uncover the molecular mechanisms involved in the responses of organisms to genetic disturbances. For instance, it shows to be useful to merge biomolecular components involved in a same pathological process like MSUD.

In the future, network integration and standardization of molecular and cellular phenotypes could improve the understanding of the evolutionary mechanisms involved in pathological processes. Further experimental and analytical efforts in this direction are warranted.

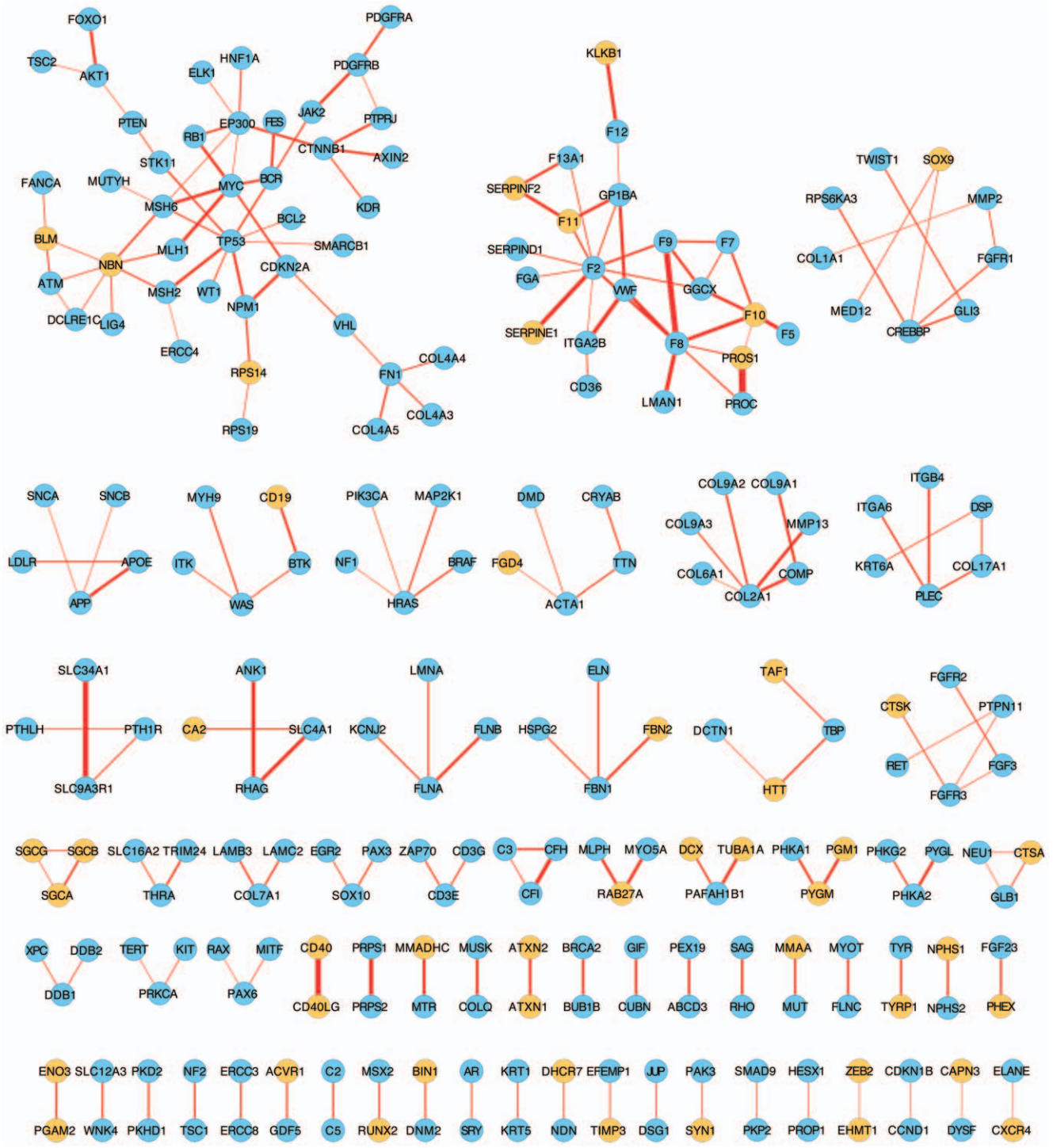


Figure 9. Physical interactions between genes with similar phenotypic lost in the current networks of diseases. This figure is the result of the difference of the resulting intersection between PSGN and PIN after removing those interactions present in HDGN and ODGN. Those genes that are MD-MG in HDN and ODN have been coloured in orangeThese genes indicate that they present underlying pathophenotypical relationships with other genes that had been disregarded by the inference of shared disease genes.
doi:10.1371/journal.pone.0056653.g009

Supporting Information

Figure S1 Schematic representation of the workflow of essential steps followed in this study: building network processes, optimal statistical threshold selection, network comparisons, topological analysis and ROC curve construction.
(PDF)

Figure S2 Spearman correlation between the number of pathophenotypes per gene and the average pathophenotypic similarity per gene for PSGN genes.
(PDF)

Figure S3 Graph of the pathophenotypes annotated to maple syrup urine syndrome. Parental nodes are close to the root in the human phenotype ontology and, therefore, with lower specificity. In contrast, child nodes are the most informative and specific pathological phenotypes.
(PDF)

Table S1 Bipartite and unipartite projections of the updated version of the human diseases network.
(XLS)

Table S2 Bipartite and unipartite projections of the updated version of the orphan disease network.
(XLS)

Table S3 Different gene subsets in the human diseases network following proposed classification.
(XLS)

Table S4 Different gene subsets in the orphan diseases network following proposed classification.
(XLS)

Table S5 Different biomolecular interactomes based on physical, metabolic and functional interactions.
(XLS)

Table S6 Biomolecular interactome and PSGN subnetworks after nodal intersections.
(XLS)

Table S7 Pathophenotypic similarity gene network.
(XLS)

Table S8 Spearman correlations between gene degrees in PSGN and HDGN/ODGN.
(PDF)

Table S9 Spearman correlation between gene degrees in PSGN and biomolecular interactomes.
(PDF)

Table S10 Network intersection between PSGN and PIN removing inferred gene-to-gene associations.
(XLS)

Methods S1
(PDF)

Acknowledgments

The authors thank J.R. Perkins and I. Morilla for useful comments and suggestions.

Author Contributions

Conceived and designed the experiments: ARP RRL. Performed the experiments: ARP RRL. Analyzed the data: ARP RRL JAGR FSJ MAM. Contributed reagents/materials/analysis tools: ARP RRL JAGR FSJ MAM. Wrote the paper: ARP RRL FSJ MAM.

References

1. Benfey PN, Mitchell-Olds T (2008) From Genotype to Phenotype: Systems Biology Meets Natural Variation. Science 320 : 495–497.
2. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA (2009) A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol 5: e1000353.
3. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
4. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74: 47–97.
5. Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21: 1010–1024.
6. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118: 4947–4957.
7. Barabási A-L, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12: 56–68.
8. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Research 37 : D793–D796.
9. Aymé S (2003) Orphanet, an information site on rare diseases. Soins: 46–47.
10. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104 : 8685–8690.
11. Zhang M, Zhu C, Jacomy A, Lu LJ, Jegga AG (2011) The orphan disease networks. Am J Hum Genet 88: 755–766.
12. Vidal M, Cusick ME, Barabási A-L (2011) Interactome Networks and Human Disease. Cell 144: 986–998.
13. Park J, Lee D-S, Christakis NA, Barabasi A-L (2009) The impact of cellular networks on disease comorbidity. Mol Syst Biol 5.
14. Ideker T, Sharan R (2008) Protein networks in disease. Genome Research 18 : 644–652.
15. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, et al. (2008) The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A 105 : 9880–9885.
16. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, et al. (2008) A Genomewide Functional Network for the Laboratory Mouse. PLoS Comput Biol 4: e1000165.
17. Linghu B, Snitkin E, Hu Z, Xia Y, DeLisi C (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome Biol 10: R91.
18. Auffray C, Chen Z, Hood L (2009) Systems medicine: the future of medical genomics and healthcare. Genome Med 1: 2.
19. Loscalzo J, Barabasi A-L (2011) Systems biology and the future of medicine. Wiley Interdiscip Rev Syst Biol Med 3: 619–627.
20. Robinson PN (2012) Deep phenotyping for precision medicine. Hum Mutat 33: 777–780.
21. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, et al. (2008) The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. Am J Hum Genet 83: 610–615.
22. Osborne J, Flatow J, Holko M, Lin S, Kibbe W, et al. (2009) Annotating the human genome with Disease Ontology. BMC Genomics 10: S6.
23. Espinosa O, Hancock JM (2011) A Gene-Phenotype Network for the Laboratory Mouse and Its Implications for Systematic Phenotyping. PLoS ONE 6: e19693.
24. Robinson PN, Mundlos S (2010) The Human Phenotype Ontology. Clin Genet 77: 525–534.
25. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, et al. (2009) Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. Am J Hum Genet 85: 457–464.
26. Oti M, Huynen MA, Brunner HG (2009) The Biological Coherence of Human Phenome Databases. Am J Hum Genet 85: 801–808.
27. Zhang S, Chang Z, Li Z, DuanMu H, Li Z, et al. (2012) Calculating phenotypic similarity between genes using hierarchical structure data based on semantic similarity. Gene 497: 58–65.
28. Brohee S, Faust K, Lima-Mendez G, Vanderstocken G, Van Helden J (2008) Network Analysis Tools: from biological networks to clusters and pathways. Nat Protocols 3: 1616–1629.
29. Bossi A, Lehner B (2009) Tissue specificity and the human protein interaction network. Mol Syst Biol 5: 260.
30. Chavali S, Barrenas F, Kanduri K, Benson M (2010) Network properties of human disease genes with pleiotropic effects. BMC Syst Biol 4: 78.
31. Cai JJ, Borenstein E, Petrov DA (2010) Broker Genes in Human Disease. Genome Biol Evol 2 : 815–825.
32. Lee D-S (2010) Interconnectivity of human cellular metabolism and disease prevalence. J Stat Mech 12015: P12015.
33. Montañez R, Medina MA, Solé R V, Rodríguez-Caso C (2010) When metabolism meets topology: Reconciling metabolite and reaction networks. Bioessays 32: 246–256.
34. Veeramani B, Bader JS (2009) Metabolic Flux Correlations, Genetic Interactions, and Disease. J Comput Biol 16: 291–302.
35. Rolfsson O, Palsson B, Thiele I (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. BMC Syst Biol 5: 155.
36. Resnik P (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI. 448–453.
37. Mistry M, Pavlidis P (2008) Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 9: 327.
38. Xu T, Du L, Zhou Y (2008) Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics 9: 472.

39. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27 : 431–432.
40. Brohée S (2012) Using the NeAT Toolbox to Compare Networks to Networks, Clusters to Clusters, and Network to Clusters. Methods in molecular biology (Clifton, N.J.). Springer New York, Vol. 804. 327–342.
41. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27: 861–874.
42. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM (2006) A text-mining analysis of the human phenome. Eur J Hum Genet 14: 535–542.
43. Xie M, Hwang T, Kuang R (2012) Reconstructing Disease Phenome-genome Association by Bi-Random Walk. Bioinformatics 1: 1–8.
44. Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res 36: D724–8.
45. Wagner GP, Zhang J (2011) The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat Rev Genet 12: 204–213.
46. Hill WG, Zhang X-S (2012) On the Pleiotropic Structure of the Genotype–phenotype Map and the Evolvability of Complex Organisms. Genetics.
47. Nellis MM, Danner DJ (2001) Gene preference in maple syrup urine disease. Am J Hum Genet 68: 232–237.
48. Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, et al. (2009) Systems biology approaches and pathway tools for investigating cardiovascular disease. Mol Biosyst 5: 588–602.
49. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18: 2078–2090.
50. Cerami E, Demir E, Schultz N, Taylor BS, Sander C (2010) Automated Network Analysis Identifies Core Pathways in Glioblastoma. PLoS ONE 5: e8918.
51. Feldman I, Rzhetsky A, Vitkup D (2008) Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A 105 : 4323–4328.

BMC Bioinformatics

SOFTWARE Open Access

PhenUMA: a tool for integrating the biomedical relationships among genes and diseases

Rocío Rodríguez-López[1,2†], Armando Reyes-Palomares[1,2†], Francisca Sánchez-Jiménez[1,2] and Miguel Ángel Medina[1,2*]

Abstract

Background: Several types of genetic interactions in humans can be directly or indirectly associated with the causal effects of mutations. These interactions are usually based on their co-associations to biological processes, coexistence in cellular locations, coexpression in cell lines, physical interactions and so on. In addition, pathological processes can present similar phenotypes that have mutations either in the same genomic location or in different genomic regions. Therefore, integrative resources for all of these complex interactions can help us prioritize the relationships between genes and diseases that are most deserving to be studied by researchers and physicians.

Results: PhenUMA is a web application that displays biological networks using information from biomedical and biomolecular data repositories. One of its most innovative features is to combine the benefits of semantic similarity methods with the information taken from databases of genetic diseases and biological interactions. More specifically, this tool is useful in studying novel pathological relationships between functionally related genes, merging diseases into clusters that share specific phenotypes or finding diseases related to reported phenotypes.

Conclusions: This framework builds, analyzes and visualizes networks based on both functional and phenotypic relationships. The integration of this information helps in the discovery of alternative pathological roles of genes, biological functions and diseases. PhenUMA represents an advancement toward the use of new technologies for genomics and personalized medicine.

Keywords: Functional relationships, Phenotypic relationships, Gene-disease relationships, Systems biology, Network medicine, Network biology

Background

Integration of clinical and biomolecular data is a key step in the advancement of current biomedical research and development. One of the greatest limitations of this process is the absence of standard platforms to merge clinical and research studies [1]. Some recent initiatives have focused on data sharing to provide precise phenotypic descriptions of patients in combination with genetic variation [2,3]. An effective integration of clinical features with their molecular context, including genetic, physical and metabolic interactions, is expected to produce new insights for biomedical research [4]. In fact, the phenome and the interactome were recently listed among the five most up-and-coming 'omes' that may offer new insights in science [5]. Therefore, new integrative data tools are required to establish these functional and phenotypic links for genome-scale analyses.

Although inherited disorder databases such as OMIM [6] and Orphanet [7], provide extremely valuable details about the molecular nature of pathological conditions, these databases lack direct procedures for integrating biomolecular information. Biomedical ontologies are promising standard resources to address a systematic integration of phenotypes into the molecular background of mutated genomic regions [1,8,9]. For instance, the Human Phenotype Ontology (HPO) currently contains over 10,000 terms that represent each one an individual phenotype [10]. An intuitive approach for determining similarities between sets of ontological terms (HPO terms), that could represent the

* Correspondence: medina@uma.es
†Equal contributors
[1]Departamento de Biología Molecular y Bioquímica, Universidad de Málaga, Andalucía Tech, Facultad de Ciencias, and IBIMA (Biomedical Research Institute of Málaga), Málaga, Spain
[2]CIBER de Enfermedades Raras (CIBERER), E-29071 Málaga, Spain

phenotypic spaces of disorders or even genes, is to estimate their proximity in the ontology.

On the other hand, the Gene Ontology (GO) is an organized vocabulary of terms that can be subdivided into three sub-ontologies: biological processes, cellular components and molecular functions. Genes are associated with consistent annotations that conform sets of GO terms that are useful to describe the cellular and molecular events involving genes [11]. Furthermore, biomolecular interactomes, such as protein-protein interactions and metabolic and gene regulatory networks, should also be used to obtain a systemic view of the molecular and biochemical reactions related to disease-causing genes [12].

In particular, because ontologies have been beneficial in understanding diseases as a set of phenotypes rather than conceptual entities, studying correlations among distinct biological conditions affected by genetic variations would be very useful [13].

The main purpose of this application is to provide a friendly platform that facilitates the analysis of phenotypic and functional information and the discovery of emergent or unnoticed relationships between pairs of genes or genetic diseases. PhenUMA also complies useful biological information from different interactomes, including protein-protein interactions from STRING [14] and metabolic flux correlations [15]. Altogether, PhenUMA may be useful for discovering interesting new insights on or features shared by human diseases, increasing the potential for diagnosis and pharmacological intervention.

Implementation

Knowledge base: data processing and storage

The initial stages of the development of PhenUMA were focused on building a consistent knowledge base, and subsequent efforts were dedicated to design a user-friendly web application. The knowledge base contains all of the information necessary to create the output networks, and the source data were retrieved from consolidated databases or from inferred relationships determined using different data processing methods (Figure 1A, schematic representation of the knowledge base). The web interface was implemented to make the query execution easier and to allow the visualization of outcome networks according to the Cytoscape Web 1.0.3 utility [16]. The tool was developed in Java, and the database was built using MySQL 5.0.45. PhenUMA and other resources such as tutorials and downloadable processed data are available on the web (http://www.phenuma.uma.es/). An illustrative example of all of the types of gene-gene relationships is shown in Figure 1B.

Known relationships

The Gene Map file provided by OMIM was used to extract 4,261 relationships between 2,794 OMIM genes and 3,486 OMIM phenotypes; OMIM genes were mapped to their GeneID. The PhenUMA knowledge base also contains the associations between Orphanet diseases and genes. This information was extracted from the file “Diseases with their associated genes”, included at Orphadata [17], and was used to develop 4,472 connections between 2,614 GeneIDs and 2,555 orphan diseases. We also included the diverse interactomes of human protein-protein interactions (96856 relationships) that were found with STRING [14] and 9812 gene pairs that had positive flux correlations in the metabolic network [15].

Inferred relationships

The inferred relationships between genes or diseases and orphan diseases are due to binary relationships, resulting in four different types of networks. For instance, an inference between two genes will be considered if at least one or more OMIM/Orphan diseases are associated with both genes. A stronger interaction between two genes will be considered when they share more than one disease. Overall, the scores that indicate the intensity of the relationship is the number of disorders involved in the relationship. The same criterion was applied to establish the inferred relationships between OMIM and Orphan disorders. In this case, the number of genes shared by the disorders is considered the score.

Semantic similarity relationships

HPO and GO were used to calculate the phenotypic similarities between genes or diseases and the functional similarities between genes, respectively. We used Ontologizer 2.0, an open-source tool, to determine the functional similarities, and it was also adapted to compute phenotypic similarities [18]. Each gene or disease is represented by a set of terms that defines its functional or phenotypic profile. Only the most specific terms are included in the annotation files because the “true path rule” is met. This rule implies that each object related to a term also relates to all of the ancestors of this term to the root. For instance, the OMIM (MIM# 200500) disorder “Acheiropody” is associated with both “Humeral hypoplasia” (HP:0005792) and all of its ancestors, such as “Aplasia/Hypoplasia of the humerus” (HP:0006507).

Two different semantic similarity measures that are based on Resnik’s approach were used to calculate the functional similarity among genes and the phenotypic similarity among phenotypic profiles. Both measures are based on the concept of information content (IC), which is calculated using the logarithm of the probability of each term (the ratio of the number of annotations of a term to the total number of annotations). If the probability decreases, then the IC increases, and consequently, the specificity and the informativeness also increase. The semantic similarity between two terms of a given ontology, as proposed by

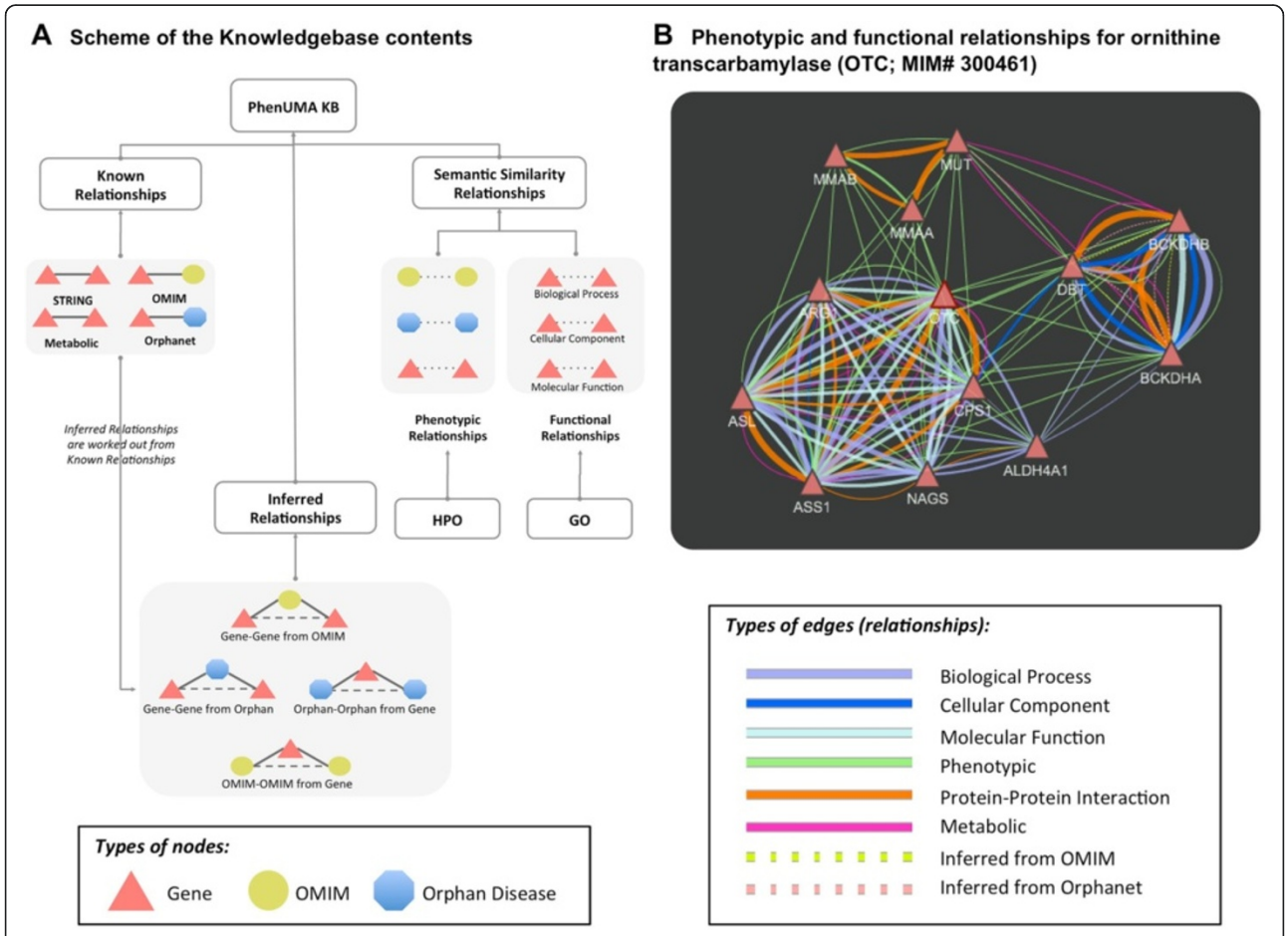


Figure 1 PhenUMA knowledge base. A: Schematic representation of the PhenUMA knowledge base contents. Three types of relationships are included in the knowledge base: i) "known relationships" (solid lines) were extracted from the databases OMIM, Orphanet and STRING and also include the metabolic interactions from Veeramani and Bader [15]; ii) "inferred relationships" (dashed lines) were taken from the OMIM and Orphanet known relationships; and iii) "semantic similarity relationships" (dotted lines). For the semantic similarity relationships, scores were calculated using the HPO and GO. Genes (red triangles), OMIM diseases (yellow circles) and Orphanet diseases (blue octagons) are the components of these relationships. This schematic describes how inferred relationships were determined from known relationships; that is, how the dashed lines were deduced from the solid lines. **B**: Illustrative example of integration between phenotypic and functional gene-gene relationships as retrieved in PhenUMA for ornithine transcarbamylase (*OTC*; MIM# 300461) at a medium confidence level.

Resnik [19], is determined by the IC of the most informative common ancestor (MICA). The similarity score between groups of terms was obtained by selecting the maximum MICA from all possible pairs of terms. This algorithm has produced suitable results for calculating functional similarity among genes on several occasions [20-22] and is based on the most specific GO terms. This allows relating genes considering the closest molecular mechanisms between them. Regarding to the phenotypic similarity, we have used the complete set of symptoms (HPO terms), associated with a disease or gene, because is more adequate to compare phenotypic profiles. For this reason, we used the method applied by Robinson and co-workers [23], based on Resnik combined with the best-match average. Briefly, if p1 and p2 are two different phenotypic profiles, the semantic similarity of this pair of HPO terms is defined as:

$$sim(p1,\ p2) = \frac{\sum_{ti\in\ p1} max_{tj\in p2} sim(ti,\ tj)}{|p1|} \quad (1)$$

where t_i and t_j represent each HPO term that is included in the profiles p1 and p2. This equation is not symmetric. Robinson and co-workers use a symmetric version for HPO [23]:

$$\mathrm{sim}_{symmetric}(p1,\ p2) = \frac{\mathrm{sim}(p1,\ p2)}{2} + \frac{\mathrm{sim}(p2,\ p1)}{2} \quad (2)$$

The annotation files that include the relationships between genes or diseases and their ontological profiles were required to calculate semantic similarity. We downloaded

the annotation file "gene_annotations.goa_human", which relates GO terms to human genes, from the GO website. Two additional files, named "phenotype_annotation.tab" for OMIM and orphan diseases and "gene2phenotype.txt" for gene annotations, were downloaded from the HPO website. In this case, only the annotations of the descendent terms from the "Phenotypic Abnormality (HP:0000118)" term were used for the calculations. This process compiled the associations of 4,965 OMIM diseases plus 3,143 orphan diseases with sets of HPO terms and relationships between 1,806 genes and HPO terms. Table 1 summarizes the different types of semantic similarities processed by PhenUMA.

Optimal threshold selection of semantic similarities

Each type of semantic similarity calculation requires the establishment of an optimal statistical threshold to differentiate between significant and non-significant similarity scores. Therefore, a minimal meaningful threshold was estimated for each class of phenotypic and functional similarity listed in Table 1. Four different reference datasets were generated from the information in the PhenUMA knowledge base: one for each phenotypic similarity (OMIM-OMIM, Orphan Disease-Orphan Disease and Gene-Gene) and another for all different types of functional similarity. In particular, we compared each dataset of disease pairs, which was inferred from the gene-disease association studies found in OMIM and Orphanet, to the phenotypic similarities between the diseases. The dataset for phenotypic similarities between genes was generated from the union of all inferred pairs obtained from OMIM and Orphanet. The fourth reference dataset resulted from the combination of interactomes from both metabolic and protein-protein interactions; the same dataset was used for all of the functional similarities.

Initially, we built a binary classifier system that compares all of the computed scores between semantically similar genes or disease pairs with their respective reference datasets. However, the estimated thresholds in each ROC curve were meaningful (Additional file 1), but they are impractical as optimal cutoffs because of the large size of the resulting networks. Therefore, we analyzed cutoff variations in the phenotypic similarity datasets using a similar approach as in one of our recent studies [13]. First, we removed all pairs of genes or diseases that had a similarity score below the 95th percentile. Next, we studied both the influence of cutoff variations on the number of gene or disease entries and the resulting Jaccard's similarity coefficients when comparing the semantic similarity networks to their respective reference datasets network (Figure 2). More specifically, the Jaccard's similarity coefficient represents the number of intersected pairs of gene or disease entries divided by the number of pairs of entries in the union.

As shown in Figure 2A, the number of genes and diseases began to decrease at the 98th percentile of all phenotypic similarities. Robinson's measurement clearly conserved more genes and diseases at the same cutoff points than Resnik's did measurement (solid lines above dashed lines, Figure 2A). The phenotypic similarity networks that result in different cutoffs are more similar to the reference dataset networks as we increase the similarity score cutoffs (solid lines above dashed lines, Figure 2B). This trend is especially notable for the evolution of Jaccard's similarity coefficient for the phenotypic similarity gene networks at the 98th percentile, where Resnik's measurement has a maximum similarity of approximately 3% and Robinson's one increases up to 10%. Indeed, this coefficient even decreased in Resnik's measurement at the 99th percentile

Table 1 Summary of main relationships in the knowledge base

Type of network	Type of interaction (source)	Nodes	Relationships
Phenotypic relationships			
OMIM-OMIM	Inferred by Genes (OMIM)	1843	2885
OMIM-OMIM	Phenotypic Similarity (HPO)	4627	149689[a]
Orphan Disease-Orphan Disease	Inferred by Genes (Orphanet)	1655	3568
Orphan Disease-Orphan Disease	Phenotypic Similarity (HPO)	3068	75924[a]
Gene-Gene	Inferred by OMIM (OMIM)	784	3217
Gene-Gene	Inferred by Orphan Disease (Orphanet)	1641	8292
Gene-Gene	Phenotypic Similarity (HPO)	1681	24902[a]
Functional relationships			
Gene-Gene	Functional Similarity (GO Biological Process)	9123	486982[a]
Gene-Gene	Functional Similarity (GO Cellular Component)	6046	565739[a]
Gene-Gene	Functional Similarity (GO Molecular Function)	8087	397683[a]
Gene-Gene	Protein-protein interactions (STRING)	10316	96856
Gene-Gene	Metabolic interactions [Veeramani and Bader[15]]	535	9812

[a]Resulting relationships to apply the respective cutoff for low confidence level.

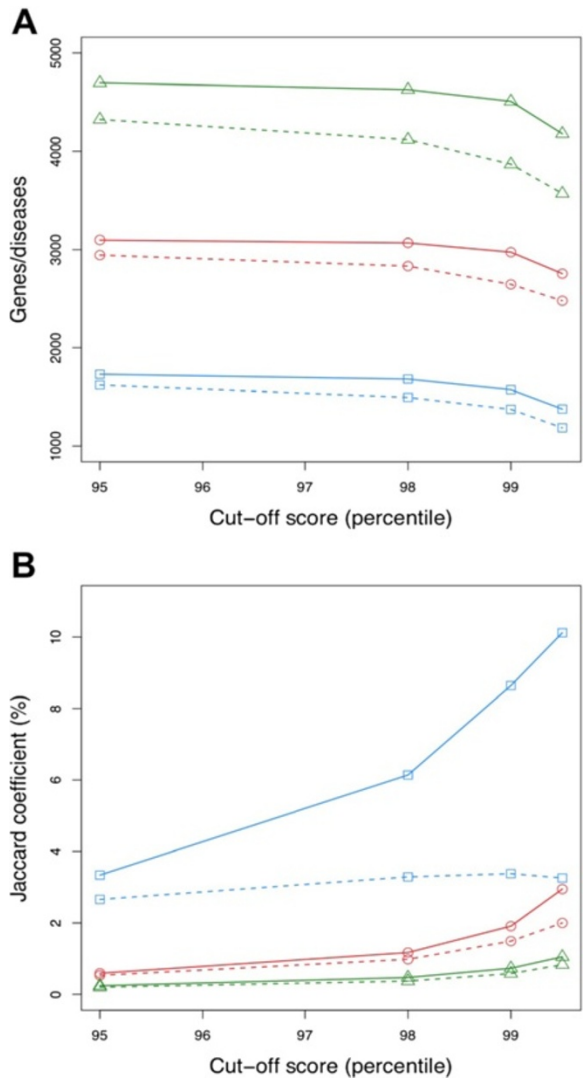


Figure 2 Effects of phenotypic similarity cutoff variations on the number of elements and Jaccard coefficients. Computed phenotypic similarities for gene pairs (blue squares), OMIM disease pairs (red circles) and Orphanet disease pairs (green triangles) were filtered at the 95th percentile, and different cutoff scores corresponding to the 95th, 98th, 99th and 99.5th percentiles were used. The Resnik and Robinson measurements are shown as solid and dashed lines, respectively. **A**: Variations in the number of genes and diseases that are involved in phenotypic similarities at increasing values of the similarity score. **B**: Variations of the Jaccard's similarity coefficients calculated from the resulting intersection between the phenotypic similarity-based networks and their respective inferred networks is represented as the distinct similarity scores.

(blue squares and dashed line, Figure 2B). The phenotypic similarity disease networks also had slightly higher Jaccard's similarity coefficients for Robinson's measurement from the 95th percentile to the top similarity score (red circles and a solid line for OMIM diseases and a green line, Figure 2B).

As it was foreseeable, the semantic similarity measurement applied by Robinson produced better performance for phenotypic similarities than Resnik's method (see Additional file 1). This analysis revealed the 98th percentile as a suitable threshold that provided a balanced tradeoff between a gain in specificity for phenotypic similarities and a loss of information for disease and gene pairs (Figure 2). For this reason, we selected the 98th percentile of Robinson's measurement as the lowest similarity value and the minimal appropriate cutoff to build phenotypic similarity based networks.

On the other hand, functional similarities are strongly dependent on large ontological domains that cluster genes with similar scores. Consequently, we set the lower cutoff at the 99.5th percentile, which considerably increases the similarity's significance and reduces noise from non-informative similarities. Therefore, phenotypic- and functional similarity-based networks were stored in the knowledge base using the 98th and 99.5th percentile as the minimal levels of confidence, respectively (Table 1). All of the scores were normalized following a min-max normalization method, and therefore the scores take values between 0 and 1, where 0 corresponds to the minimal score greater than the cutoff, and 1 represents the highest score for semantic similarity. This method results in confident semantic similarity relationships and a manageable size of networks to be processed by PhenUMA.

Results

Network building process

PhenUMA allows the retrieval of information related with a set of genes, diseases or phenotypes of interest. Figure 3 shows the building network stages for each type of input and output. When a query is executed, firstly a seed network is created from the input reported by the user; subsequently, this network is populated with the relationships included in the database for the type of data related (Figure 3B). For example, if a phenotypic similarity network is requested for one gene or one list of genes, the resulting network is populated with the functional, protein-protein interaction, metabolic and inferred relationships (see an example for ornithine transcarbamylase in Figure 1B). PhenUMA allows users to select among three different levels of confidence, termed low, medium and high, for both phenotypic similarities (the 98th, 99th and 99.5th percentiles, respectively) and functional similarities (the 99.5th, 99.8th and 99.9th percentiles, respectively).

The process of network building is quite different if a set of phenotypes is used as input. In this case, the set of phenotypes is considered as a new phenotypic profile. The similarity between this set and the phenotypic space of other genes or diseases is calculated using Robinson's semantic similarity measure. In the outcome network, the

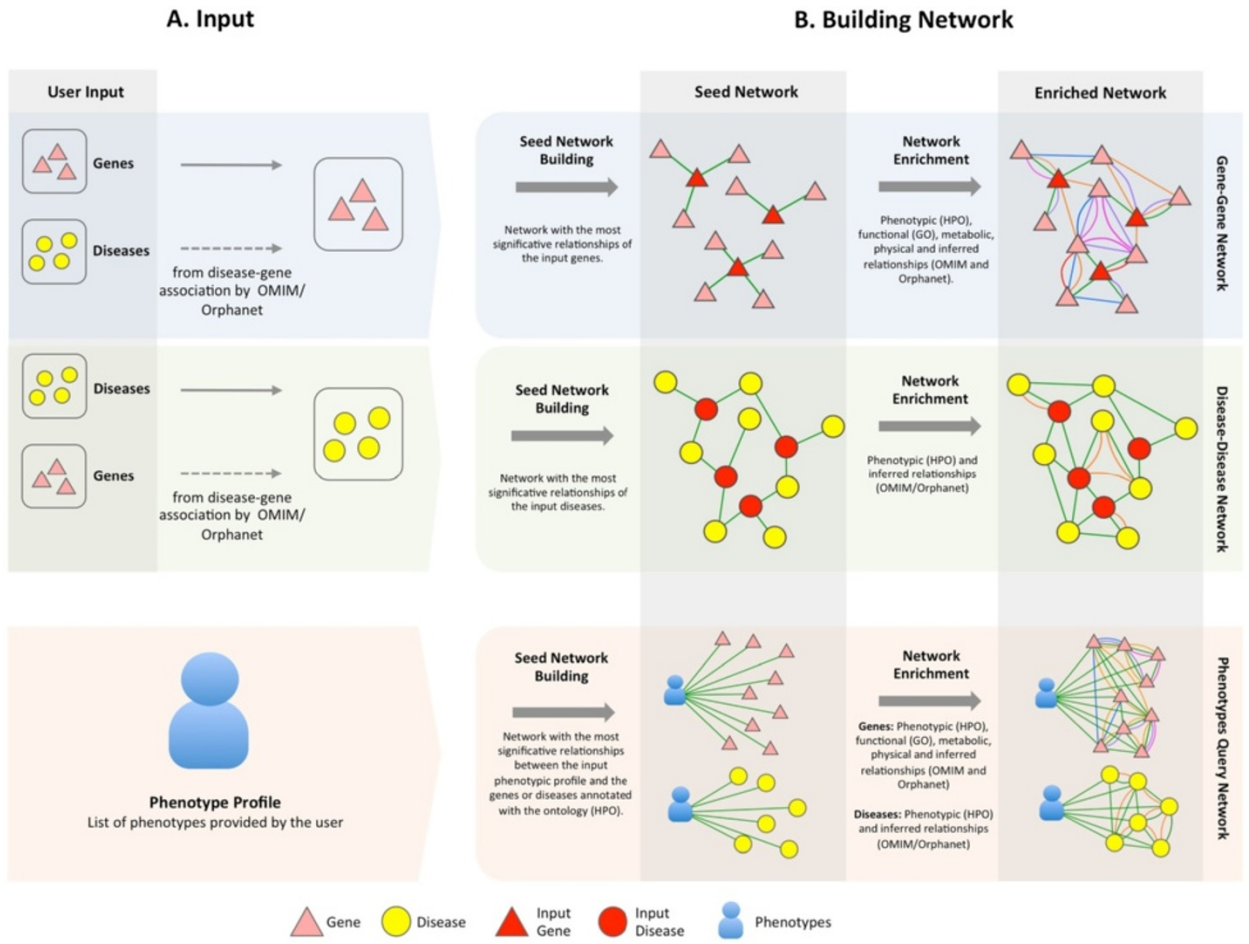


Figure 3 Building network process. A: Input provided by the user of PhenUMA. *Gene-Gene network* building allows a set of genes or diseases as input (OMIM or Orphan diseases). In case of providing a disease list, genes associated with each disease (OMIM or Orphanet associations) are use to create the gene-gene network in the building network stage. *Disease-Disease network* can relate OMIM diseases or Orphan diseases and in both cases the input type are similar: a list of diseases or a set of genes. *Phenotype query network* building require of a set of phenotypes (HPO) as input, which is taken as a phenotype profile. **B**: Building network stage is divided in two parts: the seed network building that contains the relationships between de input set (genes, diseases or phenotypes) and the rest of elements included in the database and the network enrichment that consist in the addition of the rest of relationships included in the knowledge base (see Figure 1) between the elements related in each network.

set of phenotypes is represented as a node, and only the significant relationships (*P-value* <0.05) among the genes or diseases are included. *P-values* are the probability of obtaining a greater score, between the input query and each gene or disease annotated to the ontology, in the comparison with a random set of phenotypes with same size as the input set. The calculation of *P-values* was performed using the Monte Carlo method based on the generation of random samples (1000000 of samples for each size of query from 1 to 10) of phenotypes to calculate a estimation of the probability of a greater score, similar to those used in Phenomizer [24]. For example, if the P-value associated to the score of the relationships between a query of five phenotypes and a disease is $5 \cdot 10^{-6}$ means that only 5 of 1000000 random combinations of five phenotypes provides a greater score that the input set in the comparison with a specific disease.

Novel pathological relationships between genes

The gene-gene network obtained using semantic similarity methods and the gene-gene inference network from known interactions (both OMIM and Orphanet) were compared to study their mutual coverage. Three distinct subsets were distinguished (Figure 4): inferred pairs of genes that are not included in phenotypic similarity gene network (Inferred OUT), inferred pairs of genes that are in the phenotypic similarity gene network (Inferred IN) and novel pairs of genes that are exclusively in the phenotypic similarity gene network. These latter genes represent more than 90% of all computed phenotypic similarities

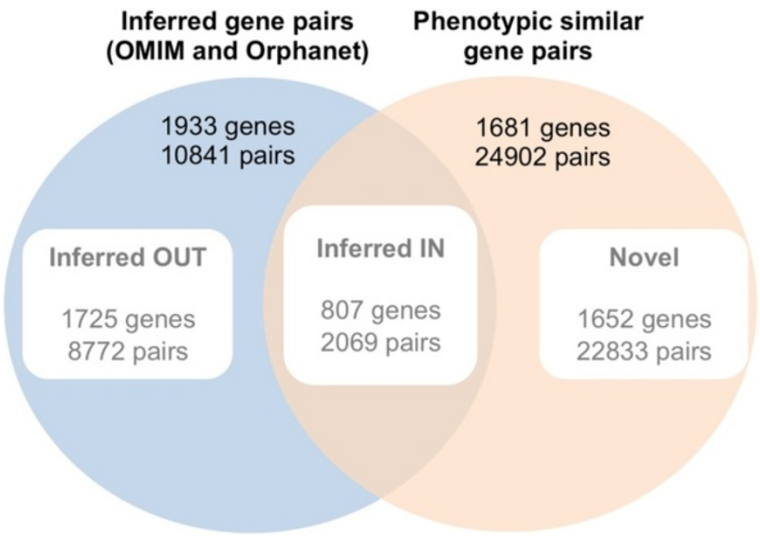


Figure 4 Subsets of inferred and phenotypically similar gene pairs. Venn diagram showing the distribution of gene pairs between a dataset of inferred relationships (from the union of OMIM and Orphanet) and the phenotypic similarity gene network at a low level of confidence corresponding to the 98th percentile.

(22,833 of 24,902 gene pairs). They are considered novel because the involved genes are not co-associated with the same genetic disease based on the current information in OMIM and Orphanet. Notably, 1606 genes in OMIM and 792 genes Orphanet are associated with only one monogenic disease so they would appear as unconnected in inferred networks. Nevertheless, more than 49% and 61% of these genes, respectively, are linked to other genes with phenotypic similarity in PhenUMA.

PhenUMA can detect whether genes are directly or indirectly involved in similar pathological events via the semantic similarity of their phenotypic profiles. For instance, some mutations in *carbonic anhydrase II* (*CA2*; MIM# 611492) are uniquely related to a monogenic disease named osteopetrosis with renal tubular acidosis (MIM# 259730 or ORPHA 2785). When using as output network of gene-gene semantic similarities from HPO with low confidence in PhenUMA, *CA2* shows phenotypic similarities to *TNFSF11* (MIM# 602642), *TBCE* (MIM# 604934) and *SLC4A1* (MIM# 109270). *CA2* also has a physical interaction with *SLC4A1* and a functional similarity for a biological process with *TNFSF11*. In agreement with the whole set of HPO annotations for *CA2*, the most specific clinical features for this gene include: distal renal tubular acidosis (HP:0008341), extramedullary hematopoiesis (HP:0001978), periodic hypokalemic paresis (HP:0008153), optic nerve compression (HP:0007807), elevated serum acid phosphatase (HP:0003148) and diaphyseal sclerosis (HP:0003034). *TNFSF11* presents phenotypic similarities with *CA2* for extramedullary hematopoiesis (HP:0001978), cranial nerve compression (HP:0001293), diaphyseal sclerosis (HP:0003034), hepatosplenomegaly (HP:0001433) and cranial hyperostosis (HP:0004437). Indeed, *TNFSF11* and *CA2* are positive regulators in bone remodeling (GO:0046852) and reabsorption (GO:0045780). *SLC4A1* shares phenotypes with *CA2*, including periodic paralysis (HP:0003768), renal tubular acidosis (HP:0001947) and hypokalemia (HP:0002900) and is also biochemically related to *CA2* by physical interactions. *TBCE* and *CA2* are not functionally associated, but both genes are associated phenotypically with renal tubular dysfunction (HP:0000124) and increased bone mineral density (HP:0011001). This example illustrates the novel phenotypic similarities for *CA2* that are integrated with other functional relationships and additional information processed by PhenUMA. All of these results can be retrieved from PhenUMA combining network visualization, informative panels and other features such as phenotypic and functional enrichment analysis of selected nodes in resulting networks.

Clustering diseases by phenotypic similarity

PhenUMA allows users to obtain coherent disease and gene clusters related to a particular disease, gene or set of phenotypes for research purposes. As an example, we will examine succinic semialdehyde dehydrogenase deficiency (SSADHD; MIM# 271980), also known as 4-Hydroxy butyric aciduria, a rare inborn error of metabolism associated with mutations in Locus *ALDH5A1* (*ALDH5A1*; MIM# 610045). We used PhenUMA to search for all of the phenotypic similarities to SSADH deficiency at each of the confidence levels of low, medium and high. These results show how different clusters of diseases are generated and belong to distinguishable groups according to their

phenotypic similarity score (Figure 5). For instance, a low cutoff for phenotypic similarity gives four large overlapped and densely interconnected clusters of disorders associated with epilepsy, seizures, neurodegenerative processes, neurophysiological abnormalities and behavioral problems (Figure 5A). SSADH deficiency has a higher frequency of connections to the disorders that involve convulsions, epilepsy or changes in behavior, and the connection becomes more evident when we increase the similarity score to the medium level of significance (Figure 5B). In this case, the established clusters have a more clearly defined structure and relationships to SSADH deficiency. Indeed, three non-overlapped clusters are apparent in Figure 5B. However, although the phenotypic coherence increased, the interconnections between clusters (OMIM diseases) remained abundant in the resulting network (Figure 5B). Therefore, we constrained the query to the most significant phenotypic similarities for SSADH deficiency by selecting the "high confidence" option in PhenUMA.

At least three types of specific phenotypes including behavioral or psychiatric abnormalities (HP:0000708), autism (HP:0000717) and generalized seizures (HP:0002197) involve a succinic semialdehyde dehydrogenase deficiency (Figure 5C). Interestingly, the clusters of disorders associated with behavioral and seizure abnormalities are interconnected by two monogenic diseases: succinic semialdehyde dehydrogenase deficiency (SSADHD, MIM# 271980) and early infantile epileptic encephalopathy-9 (EIEE9, MIM# 300088). Table 2 shows the results of a phenotypic enrichment for the 19 OMIM disorders shown in Figure 5C using the hypergeometric test provided by PhenUMA. These observations demonstrate how phenotypic similarity and network-based methods are useful in studying the pathobiology of human diseases. In particular, this method also provides an alternative procedure to understanding groups of diseases that share similar clinical features.

Comparison with other resources

A comparison between PhenUMA and related web-based tools was performed to analyze several criteria, including the integration of information, the phenotypic information used to relate genes and diseases, the visualization of

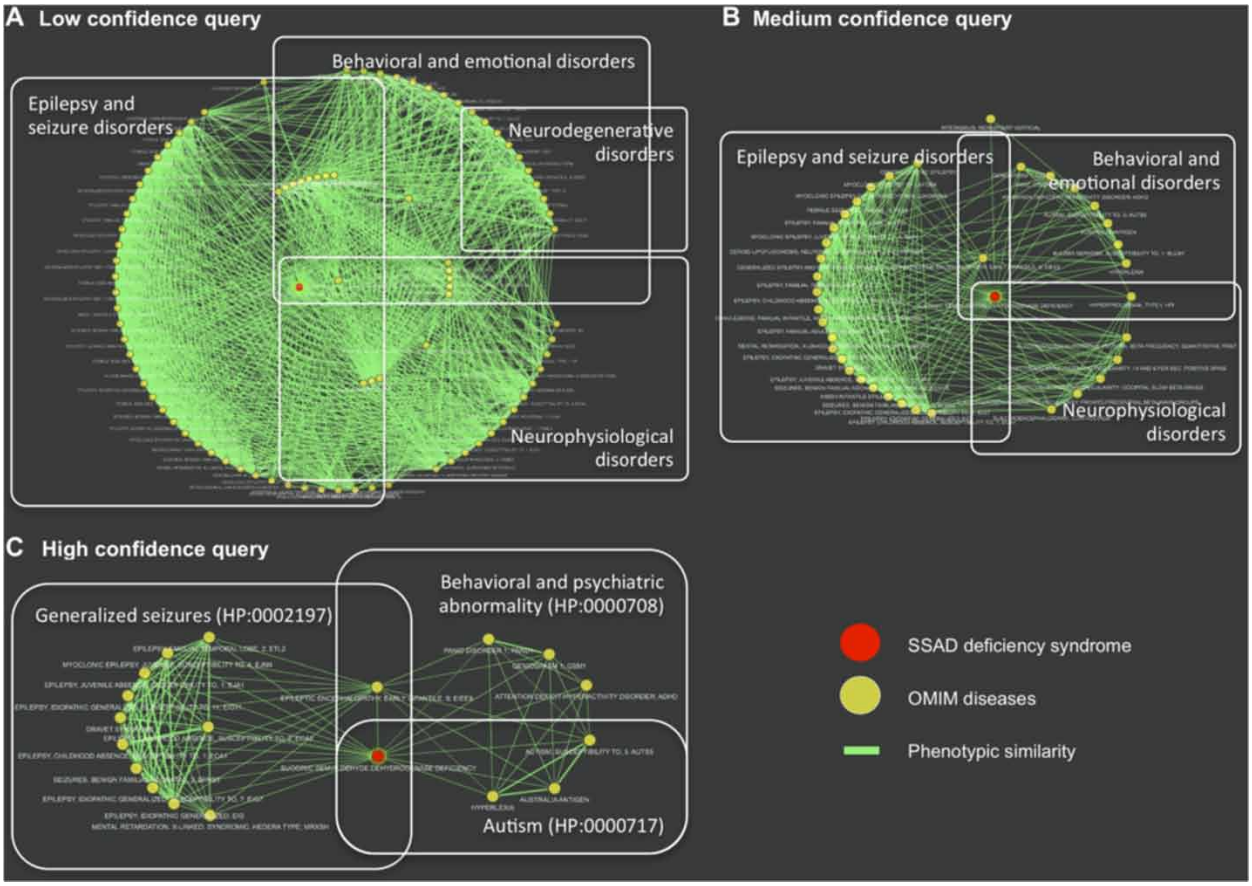


Figure 5 Phenotypically similar disorders associated with SSADH deficiency at different confidence levels. PhenUMA results of the query for SSADH deficiency (MIM# 271980) at different levels of confidence **A**: Low, **B**: Medium and **C**: High. All panels are screenshots of the PhenUMA results that were edited to highlight the main clinical features associated with each OMIM disease cluster.

Table 2 Phenotypic enrichment of SSADHD and high confidence similar disorders

HPO term	Name	Annotated diseases	Study	*P-value*	MIM diseases
HP:0002197	Generalized seizures	70	13	4.87E-19	(607628, 607681, 611364, 600669, 608096, 607631, 607208, 300423, 608217, 600131, 271980, 604827, 300088)
HP:0002123	Generalized myoclonic seizures	27	6	2.62E-08	(611364, 600669, 607631, 607208, 271980, 604827)
HP:0002133	Status epilepticus	11	4	3.53E-06	(608096, 607208, 271980, 300088)
HP:0002392	EEG with polyspike wave complexes	4	3	1.35E-05	(607681, 600669, 600131)
HP:0000717	Autism	35	4	5.29E-04	(606053, 238350, 209800, 271980)
HP: 0000708	Behavioural/Psychiatric Abnormality	406	8	4.47E-03	(143465, 606053, 238350, 167870, 209800, 271980, 300088, 190100)
HP:0001311	Neurophysiological abnormality	83	4	1.65E-02	(607681, 600669, 600131, 271980)
HP:0000739	Anxiety	33	3	1.71E-02	(167870, 271980, 190100)

information and the availability of the datasets. Table 3 summarizes all of the features considered when comparing PhenUMA with other, similar tools.

PhenUMA aims to integrate information using network-based methods, and GeneMANIA is a useful example of the integration of biomolecular data [25]. This web interface generates gene networks based on many different types of relationships such as protein and genetic interactions, pathways, coexpression, colocalization and protein domain similarities. However, in addition to functional interactions, PhenUMA also includes the pathological and phenotypic relationships between genes as shown in Table 3. Other tools, such as MalaCards, integrate the pathological and functional information related to human diseases by supplying an extensive repository of different information, where mouse phenotypes are used instead of human phenotypes [26]. Two notable tools that integrate phenotypic information are Phenomizer and PhenomeNET, but both tools are not specifically designed to integrate this information with biomolecular data, which is required for an extensive systemic analysis. Phenomizer demonstrates the potential benefits of semantic- and ontology-based methods when they are applied for the systematic diagnosis of diseases [24]; these features were also included in PhenUMA. PhenomeNET is another tool that allows users to retrieve the semantic similarities between a single OMIM/Orphan disease, gene or phenotype and other genes or diseases, including cross-species information [27] and uses a Jaccard's index to calculate phenotypic similarity. Conversely, the similarity score between the diseases as calculated by MalaCards, named the Malacards Composite Related Diseases Score (MCRDS), combines an enrichment analysis of disease descriptors with different search engine ranks [26]. The resulting ranked scores in MalacCards are also used to build disease networks based on their shared disease descriptors, but it uses murine phenotypes instead of human phenotypes.

PhenomeNET and Phenomizer are the most comparable to PhenUMA. Therefore, a more systematic comparison was performed between the results of PhenUMA and PhenomeNET. To do so, we downloaded the file "borderflow-0.1", which contains relationships and similarity scores between the phenotypes of several species, such as worm, fly, rat, mouse, zebra fish and human, from the PhenomeNET website. Given this cross-species phenotype network, we selected only OMIM disease pairs. A ROC curve was built using the same reference set of inferred relationships between OMIM diseases that share one or several genes. The resulting ROC curves from Resnik's and Robinson's measures give better results than those provided by PhenomeNET (Figure 6A). We analyzed the fraction of expected false discoveries by calculating the false discovery

Table 3 Comparison of PhenUMA with other tools

Tool	Phenotypic relationships	Phenotypic similarity method	Gene querying	Phenotype querying	Information integration	Download results	Network display
PhenUMA	Yes	IC-based	Yes	Yes	Yes	Yes	Yes
Phenomizer	Yes	IC-based	No	Yes	No	Yes	No
GeneMania	No	-	Yes	No	Yes	Yes	Yes
PhenomeNET	Yes	Jaccard's Index	Yes	Yes	Yes	Yes	No
MalaCards	No*	MCRDS	Yes	Yes	Yes	No	Yes

*Mouse Phenotypes (from Mammalian Phenotype Ontology) are related with the disease queried but not Human Phenotypes.

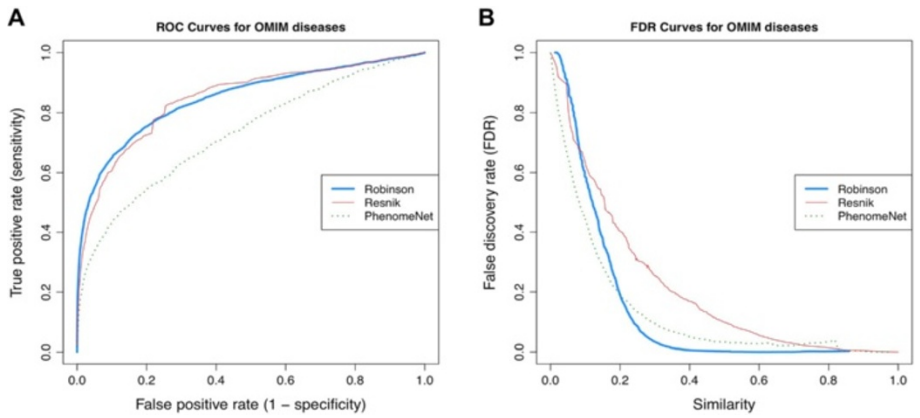


Figure 6 ROC curve and false discovery rates (FDR) for phenotypic similarities between diseases provided by PhenUMA and PhenomeNET. A: ROC curves for phenotypic similarities between OMIM diseases. For all the cases we used the same reference dataset. This dataset are all inferred OMIM disease pairs that are those diseases associated with the same gene/s. It is noteworthy that the results from Robinson and Resnik are equivalent to those in Additional file 1: Figure S1A and S1B, respectively, **B**: FDR for increasing values of phenotypic similarity scores.

rate for each system (Figure 6B). In this case, we observed a lower false discovery rate for PhenUMA, which uses the Robinson's measure, compared to the similarity score computed using PhenomeNET (Figure 6B). However, PhenomeNET gives a lower fraction of expected false positives than the classical Resnik's measure.

Finally, using the lists of diseases that are phenotypically similar to SSADH deficiency (OMIM #271980), we made a direct comparison of the results obtained from PhenUMA, Phenomizer and PhenomeNET. First, these lists were ranked by their phenotypic similarity to SSADH deficiency, and the top 10 and 50 of the OMIM diseases were selected. Then, we performed a phenotypic enrichment of each top list using a hypergeometric test and its corresponding Bonferroni correction. In the Table 4, we summarized the results of the phenotype enrichments by comparing them both to the list of phenotypes that are related to SSADH deficiency and to their respective IC values that indicate their level of specificity. For instance, status epilepticus showed the highest IC value, which indicates that it is the most specific phenotype associated with SSADH deficiency (Table 4).

PhenUMA gives a significant enrichment of status epilepticus in the top 10 and 50 of ranked diseases, while no significant enrichment was found for Phenomizer and PhenomeNET. Consequently, the diseases more phenotypically similar to SSADH deficiency are also associated with status epilepticus in PhenUMA. In addition, from the 22 phenotypes annotated for SSADH deficiency, we can count 9 significant phenotypes in the top 50 of the similar diseases retrieved by our system (Table 4). However, Phenomizer and PhenomeNET have only 4 and 5 phenotypes with a *P-value* below 0.05, respectively. Interestingly, there is a gradual enrichment of specific phenotypes in PhenUMA and Phenomizer as we constrain the conditions from the top 50 to the top 10 (Table 4). In contrast, the enrichment of phenotypes in PhenomeNET gives phenotypes with low IC values.

Discussion

PhenUMA provides an integrative framework for biomedical and biomolecular relationships among genes and genetic diseases by combining network methods and semantic similarity calculations. This integration process uses pathological and functional information from different databases, inferences of already known relationships and computed semantic similarities using biomedical ontologies (HPO and GO), as shown in Table 1. To achieve this goal, PhenUMA uses several biocomputational technologies to unify in the same platform information that apparently is unconnected. One of the primary applications of this platform is to explore how disease-associated genes are phenotypically and functional associated. PhenUMA was shown to be useful for discovering novel pathological relationships between genes and as a new way to study groups of diseases based on the similarity of their phenotypic profiles. These phenotypic similarity relationships are strongly dependent on the ontology structure and the threshold selection. The Human Phenotype Ontology is a standardized platform with recognized clinical value [24], but the selection of an optimal threshold requires reference datasets to assess the precise significance of the similarity score. In PhenUMA, we set a score for semantic similarity that is suitable to detect implicit relationships in databases. The reference datasets used here were built from the inferred relationships (the union of the sets Inferred IN and Inferred OUT of Figure 4) of disease or gene pairs from OMIM or Orphanet that share at least one disease or one gene, respectively. Each

Table 4 Phenotypic enrichment of OMIM diseases similar to SSADH Deficiency (OMIM 271980)

		Bonferroni corrected P-values					
		PhenUMA		Phenomizer		PhenomeNET	
Phenotypes	IC	Top 10	Top 50	Top 10	Top 50	Top 10	Top 50
Status epilepticus	0,709	**7,36E-03**	**1,49E-05**	6,81E-01	5,83E-01	1	1,39E-01
Absence seizures	0,681	**1,21E-02**	**6,75E-11**	**1,02E-02**	**1,91E-11**		2,99E-01
Hyperkinesis	0,658		1			1	1
Hallucinations	0,613		7,55E-01		1		1
Generalized myoclonic seizures	0,604	**6,21E-04**	**2,30E-03**	**5,20E-04**	**6,52E-04**	1	1
Anxiety	0,581	6,90E-02	**6,47E-03**	5,78E-02	4,79E-01		
Autism	0,574	7,76E-02	1,51E-01		5,67E-01		1
Psychosis	0,565	1	**6,61E-04**	1	1		1
Generalized tonic-clonic seizures	0,562	**1,19E-09**	**3,54E-29**	**2,05E-05**	**2,20E-25**	**1,44E-13**	**1,26E-14**
Delayed speech and language development	0,543		1		1	1	1
Aggressive behavior	0,540	1	**6,50E-09**	1	1		1
Hypokinesia	0,491		1				1
EEG abnormality	0,489	1	**4,31E-14**	1	**5,47E-05**	1	1
Increased body weight	0,486		1				
Hyperactivity	0,484		**3,50E-03**		1	1	1
Hyporeflexia	0,437		1			1	1,14E-01
Motor delay	0,420		1		1		1
Ataxia	0,317		1		1	1	**8,27E-21**
Abnormality of eye movement	0,307		8,75E-01			1,54E-01	**1,10E-19**
Muscular hypotonia	0,281		1			1	**2,74E-07**
Intellectual disability	0,214		1		1	8,65E-01	**1,72E-03**
Abnormality of metabolism/homeostasis	0,123	1	1	1	1	1	1

In bold, Bonferroni corrected P-values ≤0.05, hypergeometric tests.

type of inference has a different biomedical meaning. For example, an inferred relationship between two disorders, where both present genetic variations associated with the same gene, might indicate a potential functional dependence between these pathologies and the molecular mechanisms involving this gene. If these disorders are phenotypically similar, it supports the hypothesis that perturbations in this gene will produce similar clinical features. Therefore, the resulting thresholds for phenotypically similar diseases are the minimal scores that distinguish disease pairs that are potentially related to the same molecular background. On the other hand, an inferred relationship between genes suggests that both genes could be part of close functional modules. Therefore, mutations in these genes may be canalizing perturbations effects to cause the same clinical features. The resulting optimal threshold is useful for determining the minimal similarity score for two genes that may be involved in the same pathological processes.

Our analysis provides evidence that Robinson's measurement, which uses the entire phenotypic profile of disorders to calculate similarities between genes and diseases, performs better than the classical Resnik's measurement (Additional file 1: Figure S1). As the similarity score increases, it implies a higher phenotypic specificity between gene and disease pairs. Robinson's measure conserves more information (Figure 2A) and the resulting networks are more similar to the used reference datasets (Figure 2B). In addition, PhenUMA provides more confident phenotypic similarities between OMIM diseases than do other similar systems, such as PhenomeNET (Figure 6A and B). To compute similarity scores, both systems use the entire phenotypic profile of OMIM diseases instead of the most specific phenotype in the relationship. It means that the entire phenotypic profile of a disease will be more informative than the most specific phenotype, reinforcing the need for deep phenotyping [1]. Our system also has a lower false positive rate than PhenomeNET (Figure 6B). A possible explanation for these differences is that PhenomeNET uses cross-species information, so it may be influencing the similarity scores.

Furthermore, we also used a case of study of SSADH deficiency to show how phenotypic similarity generates comprehensive clusters of diseases in PhenUMA (Figure 5).

The resulting phenotypic enrichments of ranked OMIM diseases by their similarity to SSADH deficiency are quite different for PhenUMA and Phenomizer compared to PhenomeNET. For instance, PhenUMA and Phenomizer, which use the same similarity measures, are more significantly enriched with the clinical features associated with SSADH deficiency than those of PhenomeNET (Table 4). Our results suggest that clusters of phenotypically similar diseases are more coherent in PhenUMA compared to other current similar systems.

Our assessment of the integration of functional and phenotypic relationships was based in a network comparison and correlation analysis of distinct subsets of pairs of genes. In general, phenotypic similarity clusters genes that interact in close molecular and cellular biological conditions. While it remains difficult to systematically distinguish between meaningful relationships and background noise, phenotypic similarity gene network is significantly enriched with functional interactions. For instance, the resulting network of gene pairs from the "Novel subset" is coherent and abundant in functional interactions, especially for protein-protein interactions and functional similarities in biological process (see Additional file 1). In general, protein-protein interactions and pairs of genes with similar cellular localizations likely give more direct evidence for the inferred pathological relationships [28], as observed for the "Inferred IN" and "Inferred OUT" subsets (see Additional file 1). Notably, these results may be influenced by a biomedical research bias, especially for genes that are associated with the same genetic disease [29,30]. Nevertheless, PhenUMA includes the option to filter results with the highest semantic similarity by offering a range of specificity of interactions between genes or diseases. Future improvements on this feature will be needed to extend the validity and the variety of biological interactions.

Conclusions

In conclusion, the information produced by PhenUMA integrates clinical and biomolecular information to supply wider insights on the phenotypic and molecular characteristics of pathological processes. This tool is useful to help clinical and basic researchers to reinterpret their results and to redesign experiments by considering apparently non-related elements a priori. PhenUMA users can download detailed tutorials and stored networks from the knowledge base on the website. Returns, including comments and criticisms, from final users will be considered for future improvements of this tool.

Availability and requirements

Project Name: PhenUMA
Project home page: www.phenuma.uma.es
Operating system(s): platform independent
Programming language: Java

Additional file

Additional file 1: Evaluation of methods and integration of information. Evaluation of the measures purposed by Resnik and the approach used by Robinson in the semantic similarity calculation and evaluation of the integration of phenotypic and functional relationships. **Figure S1.** ROC curves for functional and phenotypic relationships. **Figure S2.** Similarity and significance of the intersection between subsets and interactomes. **Figure S3.** Distribution of functional similarity scores in the subsets of inferred and phenotypically similar gene pairs.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

RRL and ARP conceived this project. RRL and ARP wrote this paper. ARP performed the data analysis and approaches evaluation. RRL and ARP designed the database and the tool. RRL implemented the database and the tool. FSJ and MAM supervised this work. All authors read and approved the manuscript.

Acknowledgements

The authors thank PN Robinson, S. Köhler and S. Bauer for clarifying and providing details on how to associate phenotypes with genes and OMIM diseases. The authors also thank AR Palomares, JR Perkins and JAG Ranea for useful comments and suggestions.
This work is one of the activities for the Platform "Bioinformática para Enfermedades Raras" of CIBERER, which is an initiative of ISCIII.

Funding

This work was funded by CIBERER, contract AMER (CDTI, MINECO, Spain), and Grants SAF2011-26528 (MEC, Spain), CVI-06585 (Junta de Andalucia and FEDER) and PS09/02216 (MEC, ISCIII and FEDER).

Received: 21 April 2014 Accepted: 4 November 2014
Published online: 25 November 2014

References

1. Robinson PN: **Deep phenotyping for precision medicine.** *Hum Mutat* 2012, **33:**777–780.
2. Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chénier S, Chitayat D, Faghfoury H, Meyn MS, Ray PN, So J, Stavropoulos DJ, Brudno M: **PhenoTips: patient phenotyping software for clinical and research use.** *Hum Mutat* 2013, **34:**1057–1065.
3. Hamosh A, Sobreira N, Hoover-Fong J, Sutton VR, Boehm C, Schiettecatte F, Valle D: **PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features.** *Hum Mutat* 2013, **34:**566–571.
4. Schofield PN, Hancock JM: **Integration of global resources for human genetic variation and disease.** *Hum Mutat* 2012, **33:**813–816.
5. Baker M: **Big biology: the'omes puzzle.** *Nature* 2013, **494:**416–419.
6. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) [http://www.omim.org]
7. Orphanet: an online rare disease and orphan drug data base. © INSERM 1997 [http://www.orpha.net]
8. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S: **Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users.** *Hum Mutat* 2012, **33:**803–808.
9. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP: **REPORT DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources.** *Am J Hum Genet* 2009, **84:**524–533.
10. Robinson PN, Mundlos S: **The human phenotype ontology.** *Clin Genet* 2010, **77:**525–534.
11. Mistry M, Pavlidis P: **Gene ontology term overlap as a measure of gene functional similarity.** *BMC Bioinformatics* 2008, **9:**327.
12. Vidal M, Cusick ME, Barabási A-L: **Interactome networks and human disease.** *Cell* 2011, **144:**986–998.
13. Reyes-Palomares A, Rodríguez-López R, Ranea JAG, Sánchez Jiménez F, Medina MA: **Global analysis of the human pathophenotypic similarity**

gene network merges disease module components. *PLoS One* 2013, **8:**e56653.
14. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C: **The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.** *Nucleic Acids Res* 2011, **39**(Database issue):D561–D568.
15. Veeramani B, Bader JS: **Metabolic flux correlations, genetic interactions, and disease.** *J Comput Biol* 2009, **16:**291–302.
16. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser.** *Bioinformatics* 2010, **26:**2347–2348.
17. Orphadata: Free access data from Orphanet. © INSERM 1997 [http://www.orphadata.org]
18. Bauer S, Grossmann S, Vingron M, Robinson PN: **Ontologizer 2. 0 — a multifunctional tool for GO term enrichment analysis and data exploration.** *Bioinformatics* 2008, **24:**1650–1651.
19. Resnik P: **Using information content to evaluate semantic similarity in a taxonomy.** *IJCAI* 1995, **1:**448–453.
20. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19:**1275–1283.
21. Xu T, Du L, Zhou Y: **Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data.** *BMC Bioinformatics* 2008, **9:**472.
22. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A: **Correlation between gene expression and GO semantic similarity.** *IEEEACM Trans Comput Biol Bioinforma* 2005, **2:**330–338.
23. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S: **The human phenotype ontology: a tool for annotating and analyzing human hereditary disease.** *Am J Hum Genet* 2008, **83:**610–615.
24. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN: **Clinical diagnostics in human genetics with semantic similarity searches in ontologies.** *Am J Hum Genet* 2009, **85:**457–464.
25. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q: **The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function.** *Nucleic Acids Res* 2010, **38**(Web Server issue):W214–W220.
26. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Iny Stein T, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D: **MalaCards: an integrated compendium for diseases and their annotation.** *Database (Oxford)* 2013, **2013:**bat018.
27. Hoehndorf R, Schofield PN, Gkoutos GV: **PhenomeNET: a whole-phenome approach to disease gene discovery.** *Nucleic Acids Res* 2011, **39:**e119.
28. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104:**8685–8690.
29. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics* 2006, **22:**773–774.
30. Wang J, Zhou X, Zhu J, Zhou C, Guo Z: **Revealing and avoiding bias in semantic similarity scores for protein pairs.** *BMC Bioinformatics* 2010, **11:**290.

doi:10.1186/s12859-014-0375-1
Cite this article as: Rodríguez-López *et al.*: PhenUMA: a tool for integrating the biomedical relationships among genes and diseases. *BMC Bioinformatics* 2014 **15**:375.

Chapter7
Histamine and Its Receptors as a Module of the Biogenic Amine Diseasome

Rocío Rodríguez-López, María Morales,and Francisca Sánchez-Jiménez

Abstract Biogenic amines play important roles in most important physiological processes, from cell proliferation and differentiation to nutrition, immune response, and neurobiology and reproduction. These effects are spread through a wide variety of cell-specific receptors, cell-specific signaling, and metabolic pathways. However, the biochemical events underlying these effects conform very complex networks of interactions that are far from being completely understood in most cases. In addition, two or more biogenic amines can coexist in the same physiological scenarios keeping cross talk events with influence in their respective physiological functions. In this respect, histamine seems to be the most pleiotropic biogenic amine keeping biochemical and functional interactions with both growth-related polyamines and neurotransmitters in different cell models and tissues. As diseases are the consequence of a biochemical imbalance in one or more tissues, the physiological importance of these compounds and their multiple relationships must have a reflection in the human *diseasome*, the scope of which is not yet known. This fact impedes development of new solutions for diagnosis, prognosis, and treatment of the multiple diseases involving the action of biogenic amines. This work is a further effort of our group to integrate genetic, functional, and clinical information about biogenic amine-related diseases assisted by text mining and network theory-based tools with the aim of helping to advance in personalized biomedical strategies.

R. Rodríguez-López
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus Andalucía-Tech, Campusd eT eatinos, Malaga 29071, Spain

Institutod eB iomedicinaI BIMA, Campusd eT eatinos, Malaga 29071, Spain
e-mail: rorodriguez@uma.es; kika@uma.es

M. Morales
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus Andalucía-Tech, Campusd eT eatinos, Malaga 29071, Spain
e-mail: mariamoralesmar@hotmail.com

F. Sánchez-Jiménez, Ph.D.(✉)
Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus Andalucía-Tech, Campusd eT eatinos, Malaga 29071, Spain

Unidad7 41C IBERd eE nfermedadesR aras, Campusd eT eatinos, Malaga 29071, Spain

Institutod eB iomedicinaI BIMA, Campusd eT eatinos, Malaga 29071, Spain

P. Blandina, M.B. Passani (eds.), *Histamine Receptors*, The Receptors 28,
DOI 10.1007/978-3-319-40308-3_7

Keywords Histamine • G-protein-coupled receptors • Dopamine • Serotonin • Polyamines • Amine oxidases • Cancer • Inflammation • Neurotransmission • Rare diseases• Systemsme dicine

Abbreviations

5′-HT	5′-Hydroxitriptamineors erotonin
Ac	Acetylm oiety
ADHD	Attention-deficith yperactivityd isorder
DA	Dopamine
DFMO	Difluoromethylornithine
GABA	Gamma-aminobutyrica cid
GO	GeneO ntology
GWAS	Genome-widea ssociations tudies
H_1R	Histaminere ceptortype 1(prote in)
H_2R	Histaminere ceptortype 2(prote in)
H_3R	Histaminere ceptortype 3(prote in)
H_4R	Histaminere ceptortype 4(prote in)
Hia	Histamine
HPO	Humanp henotypeontology
ODC	Ornithine decarboxylase(prote in)
OMIM	OnlineM endelianinh eritanceinma n
PA	Polyamines
PLP	Pyridoxal5 ′-phosphate
Put	Putrescine
ROS	Reactiveoxyge ns pecies
Spd	Spermidine
Spm	Spermine

Genes and their encoded proteins are abbreviated by their official symbol as recommended in the NCBI gene database and listed in Table 6.1.

7.1 Introduction

Metabolism of amino acid derivatives has almost been neglected for years in general biochemistry books as it was considered part of the secondary metabolism. However, the progressive knowledge integration of different biomedical areas and the more systemic view of the biological processes are progressively revealing the great importance of these derivatives in human physiopathology. For instance, decarboxylation products of amino acids, commonly known as biogenic amines, are

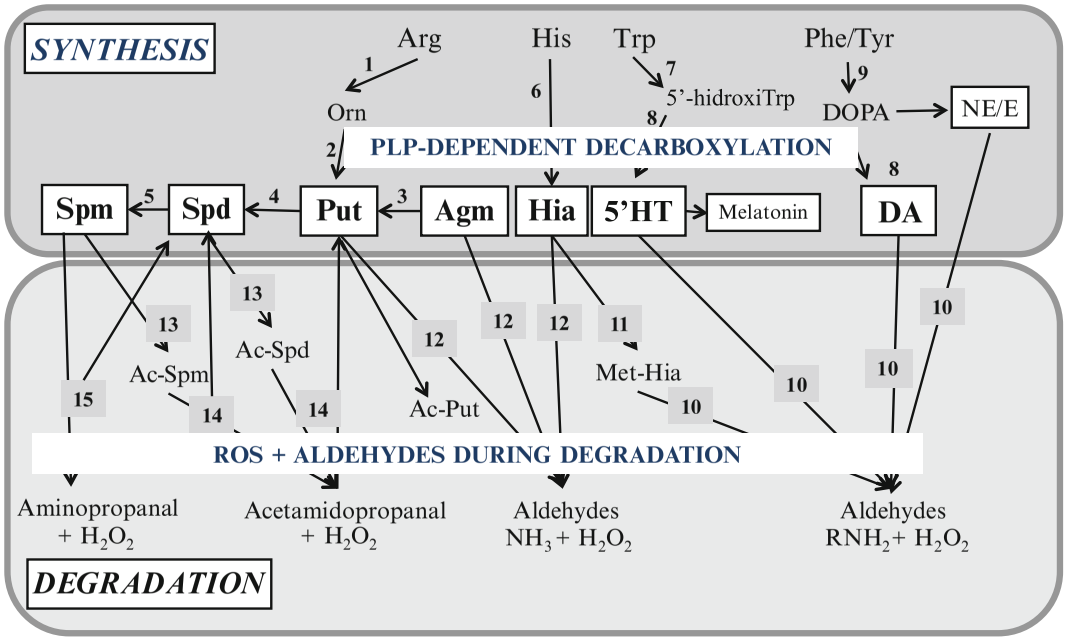


Fig. 7.1 Scheme of synthesis and degradation pathways for biogenic amines derived from cationic and aromatic L-amino acids. *White* cases crossing the pathways label common steps among them, and *bold* lettering points out the biogenic amines considered in the frame of this chapter. The enzymes considered in this work are located by numbers in the figure, with the correspondence to the official symbols of the respective genes (see Table 7.1) as follows: 1, ARG; 2, ODC; 3, AGMAT; 4, SMS; 5, SRM; 6, HDC; 7, TPH; 8, DDC; 9, TH; 10, MAO; 11, HNMT; 12; DAO; 13, SAT; 14, PAOX; 15, SAOX

essential biomolecules for all the most important physiological functions of a human being. A synthetic scheme of the synthesis and degradation pathways of biogenic amines is depicted in Fig. 7.1, showing some of the common features among them. Biogenic amine synthesis includes a decarboxylation step that, in some cases, is the only reaction required for their synthesis; i.e., putrescine (Put), histamine (Hia), and gamma-aminobutyric acid (GABA). In mammals, L-amino acid decarboxylases are pyridoxal 5′-phosphate (PLP)-dependent enzymes [1]. Degradation of biogenic amines includes the action of different amine oxidases, which can lead to the overproduction of ROS and toxic aldehydes [2]. Moreover, Hia and Put share the possibility of being degraded by diamine oxidase, the AOC1 encoding protein (Table 7.1) [3]. Further information on the enzymes involved in these pathways can be obtained from previous references [4–8].

Among biogenic amines, only polyamines that are putrescine, spermidine, and spermine (Put, Spd, and Spm, respectively) are synthetized in nearly all human cell types, at least in those with the capability to proliferate. The so-called higher polyamines (Spd and Spm) are essential for nucleic acid synthesis and conformation, Spd being the precursor of hypusine, an essential moiety for activity of the translation factor eIF-5A [37]. In fact, the enzyme responsible for Put synthesis, ornithine decarboxylase (ODC), is being used as an anticancer drug target. The ornithine analog difluoromethylornithine (DFMO) or eflornithine is a suicide ODC inhibitor that is being used in clinical works against several types of cancer (i.e., as chemopreventive in colon cancer and in therapies against neuroblastoma) [38, 39] and has antiparasitic properties [40]. In addition, it has been proposed as potentially useful for many other neoplasias. Evidence is also claiming for a role of polyamines as neuroactive compounds [41] as well as required for the correct maturation of some immune cell types (i.e., mast cells) [42]. No structure of any specific mammalian polyamine transport system has been characterized, in spite of valuable efforts by several research groups [43, 44]. Nevertheless, it is accepted that the simultaneous inhibition of PA synthesis and transport could provide a good strategy against pathologies involving undesirable cell proliferation [29]. Although, several key questions on polyamine biochemistry, molecular biology, and physiology are still open in spite of the demonstrated relevance of the compounds in cell life/death equilibrium [45].

Histamine (Hia) is the product of the histamine decarboxylase reaction. Hia is the ligand of at least four specific membrane receptor types that are members of the G-protein-coupled receptor (GPCR) family and named as H_1R–H_4R (encoded by HRH1–4 genes, Table 7.1) [17, 46]. Histamine, along to polyamines, is a ligand of the N-methyl aspartate receptors (encoded by GRIN genes, Table 7.1), but uses a different binding site in the target [47]. Histamine can be considered the most pleiotropic biogenic amine. On the one hand, it is clear that histamine is able to modulate cell proliferation [48, 49], sometimes showing antagonic synthesis with respect to polyamines [50, 51]. In vitro, it is demonstrated that Hia is able to bind DNA changing its structure [52], and it has been detected in human breast cancer cell nuclei [53]. On the other hand, histamine is a neurotransmitter that is functionally connected to other biogenic amine neurotransmitters [54–57]. In fact, it has been proven that H_3R is a therapeutic target for several emergent neurological diseases, and there are advanced clinical trials checking the pharmacological usefulness of

Table 7.1 Biogenic amine-related genes (considered in this work) identified by the encoded protein/polypeptide names, the gene official symbols, and the respective Entrez Gene ID

Proteinna me	Official gene symbol	Gene ID	Reference
Polyamines			
Arginase1	ARG1	383	[9]
Arginase 2	ARG2	384	[10]
Ornithinede carboxylase	ODC1	4953	[11]
Ornithinede carboxylasea ntizyme1	OAZ1	4946	[12]
Ornithinede carboxylasea ntizyme2	OAZ2	4947	[12]
Ornithinede carboxylasea ntizyme3	OAZ3	51686	[12]
Antizymei nhibitor1	AZIN1	51582	[12]
Antizymei nhibitor2	AZIN2	113451	[13]
Spermidines ynthase	SRM	6723	[4]
Spermines ynthase	SMS	6611	[14]
Spermidine/sperminea cetylt ransferase	SAT1	6303	[7]
Polyamineoxi dase	PAOX	196743	[7]
Spermineoxi dase	SMOX	54498	[7]
Histamine			
Histidinede carboxylase	HDC	3067	[15]
HistamineN -methylt ransferase	HNMT	3176	[16]
Histaminer eceptor1	HRH1	3269	[17]
Histaminer eceptor2	HRH2	3274	[17]
Histaminer eceptor3	HRH3	11255	[17]
Histaminer eceptor4	HRH4	59340	[17]
Dopamine/serotonin			
Tyrosineh ydroxylase	TH	7054	[18]
Tryptophanh ydroxylase1	TPH1	7166	[19]
Tryptophanh ydroxylase2	TPH2	121278	[19]
AromaticL- aminoa cidde carboxylase	DDC	1644	[20]
Dopaminer eceptor1	DRD1	1812	[21]
Dopaminer eceptor2	DRD2	1813	[21]
Dopaminer eceptor3	DRD3	1814	[21]
Dopaminer eceptor4	DRD4	1815	[21]
Dopaminer eceptor5	DRD5	1816	[21]
5-Hydroxytryptaminer eceptor 1	HTR1A	3350	[22]
	HTR1B	3351	
5-Hydroxytryptaminer eceptor2	HTR2A	3356	[23, 24]
	HTR2B	3357	
	HTR2C	3358	

(continued)

Table 7.1 (continued)

Proteinna me	Official gene symbol	Gene ID	Reference
5-Hydroxytryptaminer eceptor3	HTR3A	3359	[23]
	HTR3B	9177	
	HTR3C	170572	
	HTR3D	200909	
	HTR3E	285242	
5-Hydroxytryptaminer eceptor4	HTR4	3360	[23]
5-Hydroxytryptaminer eceptor5	HTR5A	336	[23]
5-Hydroxytryptaminer eceptor 6	HTR6	3362	[23]
5-Hydroxytryptaminer eceptor7	HTR7	3363	[23]
Shared elements between biogenic amines (*)			
Diamineoxi dase	AOC1	26	[3, 25]
Retinaa mineoxi dase	AOC2	314	[26]
Semicarbazides ensitivea mineoxi dase	AOC3	8639	[27]
Monoamineoxi daseA	MAOA	4128	[28]
Monoamineoxi daseB	MAOB	4129	[28]
Organicc ationt ransporter2	SLC22A2	6582	[29]
Organicc ationt ransporter3	SLC22A3	6581	[29]
Solutec arrierf amily3	SLC3A2	6520	[29]
Solutec arrierf amily6	SLC6A3	6531	[29]
	SLC6A4	6532	
Solutec arrierf amily 8	SLC12A8A	84561	[29]
Vesiculara minetra nsporter1	SLC18A1	6570	[30]
Vesiculara minetra nsporter2	SLC18A2	6571	[31]
Transglutaminase1	TMG1	7051	[32]
Transglutaminase2	TMG2	7052	[33, 34]
N-Methyla spartater eceptor1	GRIN1	2902	[35]
N-Methyla spartater eceptor 2	GRIN2A	2903	[36]
	GRIN2B	14811	

* Elements involved in metabolism of more than one biogenic amine subset

H_3R antagonist and inverse agonists against them [58]. Hia is also a well-known immune mediator with a major role in allergies among other immune pathologies [6, 59–61], which in turn could take part in the inflammation-carcinogenesis interplay [62]. In addition, Hia plays an important role in gastric physiology and is responsible for gastric acid secretion [63, 64]. A role in Leydig cell functions has also been suggested for Hia [65]. Thus, the most important and complex human physiological functions are modulated by this biogenic amine (neurology, immunology, nutrition, reproduction, proliferation, and differentiation).

The products of aromatic L-amino acid decarboxylase or DOPA decarboxylase (DDC), mainly serotonin (5′-HT) and dopamine (DA), are also neurotransmitters and neuroendocrine compounds also transmitting their signals through a series of members of the GPCR family [21]. It is known that disturbances in their synthesis, transport, degradation, or reception are in the bases of many emergent neurological

disorders (i.e., schizophrenia, Parkinson's, anxiety and depression, attention-deficit hyperactive disorder, bipolar disorder, etc.), circulatory and immunological problems (i.e., hypertension, allergies, psoriasis), as well as rare diseases (i.e., aromatic L-amino acid decarboxylase deficiency, Lesch-Nyhan syndrome, Prader-Willi syndrome, among many others) [20, 30, 66–68].

There are other biogenic amines derived from L-aromatic amino acids playing very key roles in our neurophysiology, i.e., melatonin, epinephrine, and norepinephrine, which play very important roles as modulators of our circadian cycle and coordination of physical activity, alert/relax shift, etc. [68, 69]. In addition, glutamate decarboxylase produces gamma-aminobutyric acid (GABA), the most important neurotransmitter that reduces neuronal excitability and controls muscle tone. Alterations of GABA-related molecular elements are related to many human diseases (i.e., fragile X syndrome, Rett syndrome, Down syndrome, schizophrenia, Tourette's syndrome, neurofibromatosis, tremor, epilepsy, etc.) [70, 71]. Many pharmaceutical investments are currently devoted to drug development against both GABA- and catecholamine-related diseases. This fact has contributed to a higher degree of information about gene-disease-drug relationships with respect to PA and Hia, as we will see later.

In this chapter, we will focus our attention on the biogenic amines derived from cationic amino acids (Put, Spd, Spm, and Hia) and the DDC product (5′-HT and DA) and their related macromolecules (Table 7.1); our expertise is mainly with biological problems related to these amines and their related elements [72, 73]. Nevertheless, our opinion is that the biogenic amine physiopathology needs a full integration of the information concerning the entire family of biogenic amine-related elements. Thus, this chapter should be considered as the starting point for a more ambitious integrative project on the molecular and biomedical information of all biogenic amines.

When the physiological responses associated with the different amines are observed, Hia appears as a structural and functional connector among them. Hia is the product of a cationic amino acid and is able to modulate cell growth as well as share neurological functions with biogenic amines derived from L-aromatic amino acids. Our group has obtained multiple evidences about the cross talk between PA and Hia summarized in several previous reviews [73, 74]. This experience led us to the following perspective: The biomedical universe of biogenic amines derived from cationic and aromatic amino acids consists of multiple subnetworks of interactions among biomolecular elements (genes, proteins, and metabolites), each one involving hundreds of molecular elements synthetized in a cell-type-specific manner. In addition, these subnetworks also keep cross interactions among them through different events in different tissues/organs. This competition for the same ligands or targets has metabolic and physiological consequences that are not well characterized so far. Nevertheless, all of these interactions must be coordinated to keep a healthy state of a human organism. Thus, further characterization of these complex and intertwined biogenic amine-related physiological scenarios is essential to fully understand a long list of pathological symptoms and diseases and requires an effort to integrate all the biochemical, molecular, and phenotypic data around the elements related to biogenic amine metabolism and signaling [72, 73]. This is to say that we need to advance toward a more holistic view of the problem. These efforts should help for future and more efficient intervention strategies. Taking this into account,

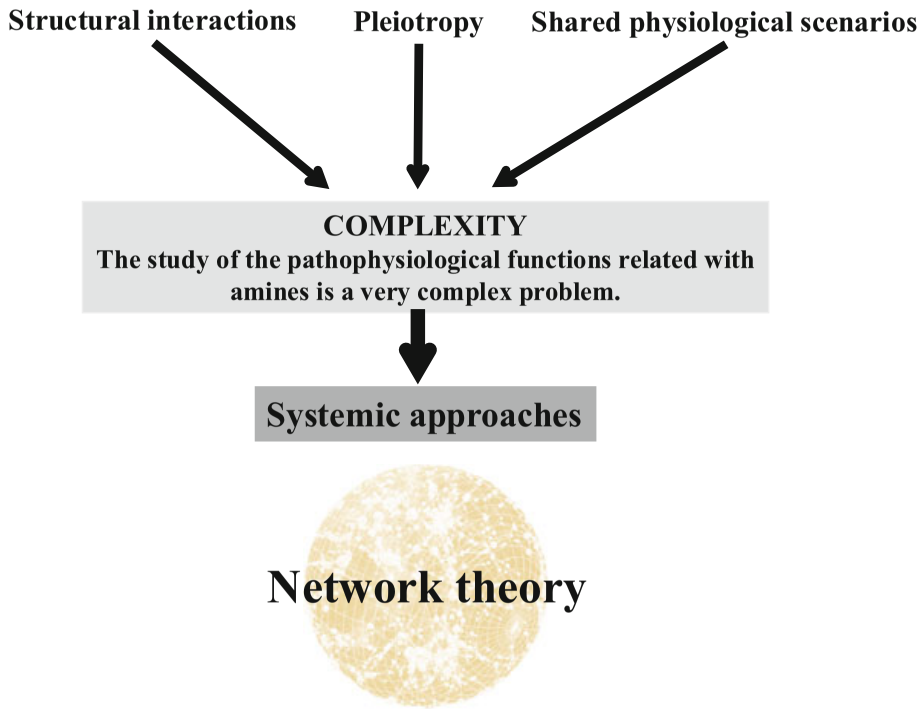


Fig. 7.2 A scheme of the hypothesis and strategy of the present work

we need to develop biocomputational support (databases, analytical tools) to organize, prioritize, and curate the molecular and clinical information. Figure 7.2 is a scheme of the hypothesis and strategy of the present work.

Thus, herein we locate and present a first set of integrative information on relationships of biogenic amine-related elements in the context of human pathologies. This information comes from computer-assisted searches and statistical calculations, in addition to our own experience. Nevertheless, we encourage the organization of a multinational open platform to be progressively enriched and curated by the "aminer community." This platform could include structural and functional cross talk events among all biogenic amine-related elements, with the aim of understanding the network topology better, as well as genetic and pharmacological data. It would consequently help the characterization and intervention of still obscure biomedical problems as important as behavior abnormalities, psychosomatic problems, brain-gut axis abnormalities, roles of immune cells in neurodegenerative diseases, additive/synergistic effects among genetic variants, and so on.

7.2 Histamine: ASys temicC ontrollerSynthe tizedbyJus t a Few Selected Cell Types

Histamine is able to scatter intercellular communication signals to a wide variety of cell types of a human body by using different tissue-specific receptor targets (see other chapters of this book), but synthetized and stored by a very reduced set of cells, known as histamine-producing cells: histaminergic neurons, enterochromaffin-like cells, and mast cells [75–78]. Other immune cells and some tumor types can synthetize but not store histamine into specific endosomes [79].

It is also known that histamine metabolism-related elements share functions and associated pathologies with elements of other biogenic amine subnetworks (Fig. 7.3). With respect to PA and cell growth, we have observed that non-producing cells (i.e., HEK-293, derived from human embryo kidney), transfected to overproduce Hia, reduce their ODC activity, PA levels, as well as cell viability and cell cycle progression [80], with concomitant induction of caspases 3/7 and alpha-synuclein [81]. In fact, it is hard to get stable transfected cells overexpressing human histidine decarboxylase (HDC), which is coherent with the lack of experimental models in literature overexpressing HDC. Maybe it is related to the fact that HDC activity is sorted and secluded in lumen of endoplasmic reticulum before maturation/activation in histamine-producing cells [82]. It is worth mentioning that malignant forms of human mastocytosis express high levels of HDC; curiously this type of neoplasia does not exhibit a high rate of cell proliferation [83, 84].

It has also been observed that elevation in histamine levels reduces the levels of PA (and/or ODC activity/expression) in different mouse cultured mast cells and during mast cell differentiation in vivo [42, 50, 51, 85, 86]. In turn, human myeloid leukemia cell differentiation to macrophage is negatively regulated by Spm [87]. At a physiological level, they also establish a cross talk in other different scenarios, for instance, progression of several human cancer types and gastrointestinal and neurological functions

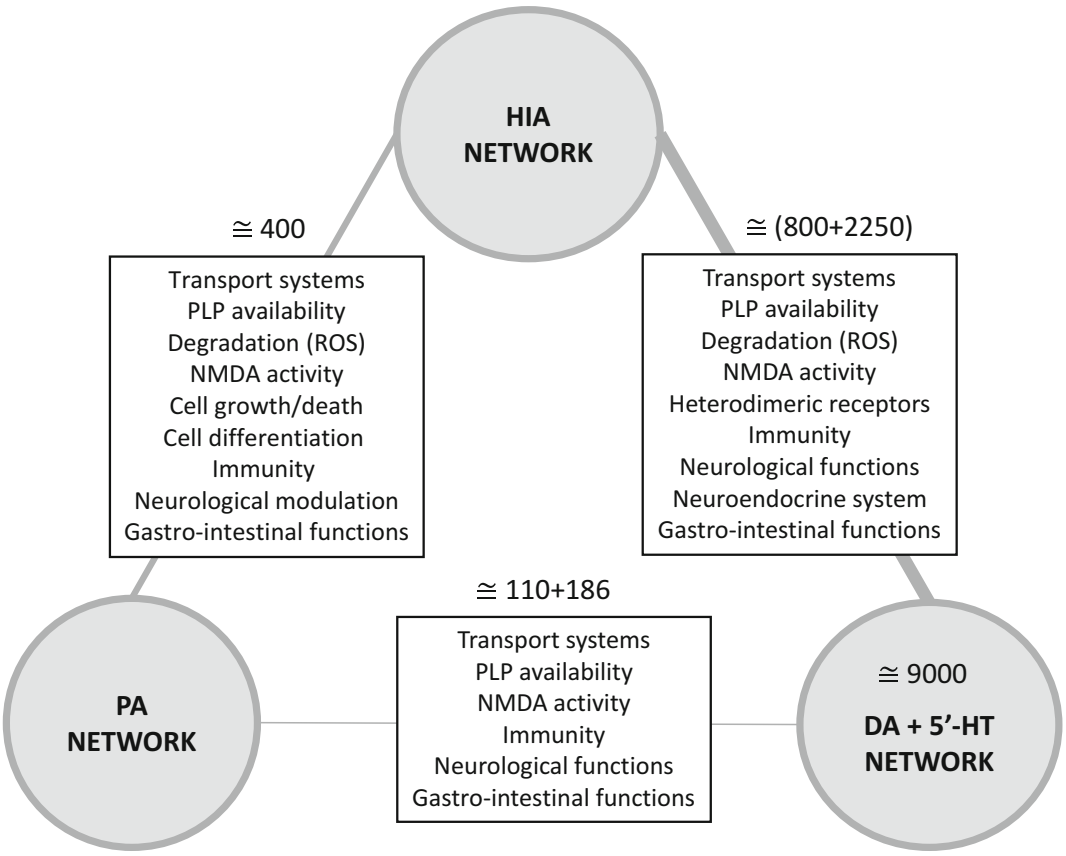


Fig. 7.3 Biochemical and physiological processes shared among biogenic amine subnetworks. Numbers on the edges or inside the DA/5′-HT subnetwork indicate the number of PubMed database publications retrieved under the order "human plus the name of two of them"

[61, 74, 88]. At metabolic levels, Hia and PA coincide in several metabolic points: receptors (i.e., NMDA) [47] and transport systems (i.e., several organic cation transporters and vesicular membrane amine transporters) [29], amino oxidases [3, 89], transglutaminase activity [90], and detox elements (i.e., Cytochrome P450) [91].

These facts should be taken into account in order to characterize the role of Hia-PA interplay in different cancer models, as well as in genesis and/or evolution of Parkinson's and other neurological and neurodegenerative diseases [92]. Several international groups are producing very interesting results on Hia implication in these pathologies [93–97].

Hia and 5′-HT are two immune mediators stored by mast cell granules [66]. In fact, an equilibrium is kept inside the granule between proteoglycans (anionic biomolecules) and Hia or 5′-HT (organic cations) [98]. This suggests that both amines could exclude each other as intravesicular components of mast cells. Both amines also play important roles in appetite and digestion. Serotonin is considered to be a member of the gut-brain axis linked to orexigenic signals [99]. Hia is also related to the orexin/hypocretin system [100]. Hia is the key inducer of gastric acid secretion during digestion and is included in the ghrelin-gastrin-hia-HCl axis [101, 102]. In turn, ghrelin increases the concentration of dopamine in the substantia nigra [103].

Respecting both 5′-HT and DA, it is well known that Hia shares roles with these amines in similar scenarios inside and outside the brain [61]. Both amines are related to many neurological and neuroendocrine disorders, for instance, schizophrenia; Parkinson's; Alzheimer's; affective disorders; hyperactive, addictive, and aggressive behaviors; ADHD; and appetite disorders [78, 104–108]. Physical interaction has been reported between D1 or D2 receptors and H_3R in striatal postsynaptic membranes [109, 110], which have also been proposed as being important for these neurologic disorders [56]. In addition, as mentioned above, recent results indicate that an excess of newly nascent Hia in cytosol induces the synthesis of α-synuclein in an HDC transfected model (human embryo kidney cells-297) [81]. Increased levels of intracranial PA have also been detected in Alzheimer's patients [111].

Other molecular and/or functional tripartite cross interactions have been suggested among PA, Hia, and/or DA/5′-HT. Human HDC and DDC share more than 50 % of the protein sequence; in fact, they share some ligands (i.e., PLP, histamine, and EGCG) [112, 113]. In spite of ODC and HDC/DDC apoenzymes not being homologous proteins, they share PLP as the cofactor, PLP acting as a chaperone of their respective apoenzyme native conformations [114]. Therefore under conditions of vitamin B6 deficiency, PLP availability could affect many steps of amino acid metabolism including synthesis of all biogenic amines mentioned in this chapter. Degradation of all biogenic amines produces ROS through activity of different polyamine, diamine, and monoamine oxidases (Fig. 7.1) with deleterious effects in different tissues, as mentioned above [89, 115].

Gastrointestinal microbiota (including pathogen organisms) produces biogenic amines, which can be an important source of these compounds for human beings. PA, Hia, DA, and 5′-HT play important roles in gastric and intestinal functions. PA is important for gastrointestinal epithelial proliferation [116]; Hia is needed for gastric acid secretion but deleterious in the case of inflammatory bowel diseases [117]. A competence between Hia and Put incorporation into rat enterocytes involving transglutaminase activity has been reported [118]. As well as DDC products, Hia plays a role in the brain-gut axis, as mentioned before [119]. Food and microbiota seem to be the sources of agmatine for human beings. This biogenic amine is the product of arginine decarboxylase activity (apparently absent in human cells) and a precursor of putrescine [120]. Beneficial effects have been assigned to agmatine in human health, and its therapeutic use has been proposed for a wide spectrum of pathologies, i.e., diabetes mellitus, neurodegenerative diseases, opioid addiction, mood disorders, cognitive disorders, and cancer [121–123].

The effects of dietary biogenic amines have been the subjects of two European COST Actions: COST 917 [124] and 922 [125]. There were several clear conclusions from these communication forums, one of them being that further efforts are required to clarify the absorption rates and transport systems determining the real concentrations and consequently the influence of microbiota-derived biogenic amines versus endogenous synthesis in human physiopathology. In addition, to evaluate the neurological effects of dietary amines, we should get more data on blood-brain barrier permeability to amines and amine derivatives [126].

7.3 Needand Str ategyfor aFir st Biocomputational InformationI ntegrationE ffortinthe B iogenic Amine-Related Human Pathology Field

Figure 7.3 is a scheme of the biochemical and physiological processes shared by Hia subnetwork with the other two biogenic amine subnetworks considered in this chapter, as synthetized in the previous section. Each one of the mentioned processes involves hundreds or even thousands of molecular, metabolic, genetic, and cellular elements, that is to say, too much information to be managed just for a human brain. Something similar occurs for references. For instance, in PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) more than 40,000 references are retrieved under the order "human polyamines" or "human histamine" and around 60,000 asking for "human dopamine" or "human serotonin." Even when further restricted by including the word "diseases" in the order, we can get from 6000 to 20,000 references depending on the amine. The number of papers containing "human and the name of two of these amines" and retrieved by PubMed database is also shown in Fig. 7.3. It seems to be very low considering the similarity among the biochemical and functional list of processes, suggesting that there must be an important quantity of information on physiopathological relationships among biogenic amines still undisclosed. It is therefore clear that the full characterization of the biogenic amine "universe", in general, needs the help of integrative bioinformatics assisted resources. The effort requires not only the development of the biocomputational tools but also a cooperative frame among international experts; we have been claiming this for years [72, 73].

As a paradox, in spite of the huge quantity of phenomenological and biochemical information on the roles of biogenic amines in human disease models and samples, the most complete databases on human diseases (OMIM, Orphanet, Decipher) provide incomplete information on relationships between diseases and biogenic amine-related elements as they usually consider genetic information only. This fact is specially marked in the cases of Hia and PA.

Among polyamine-related elements, inactivating mutations of the spermidine synthase gene codification causes Snyder-Robinson syndrome [14]. Ornithine decarboxylase genetic variations are related to APC-dependent colon cancer risk. In pediatric neuroblastoma, PA metabolism also plays a key role in the metabolic remodeling, which is essential for tumor survival and proliferation [127]. In fact, there are ongoing clinical trials on both types of neoplasias based on these findings [128, 129]. Nevertheless, there is an impressive quantity of information on polyamine and other types of cancer for which the current databases are almost blind.

In the case of histamine, OMIM (http://omim.org/) relates expression of truncated forms of human HDC to Gilles de la Tourette's syndrome [130]. The lack of histamine N-methyl transferase (HNMT), an enzyme that participates in histamine degradation, is related to susceptibility to asthma [131]. The role of H_1R (HRH1) in susceptibility to encephalomyelitis/multiple sclerosis is still controversial (Table 7.2) [132, 133]. Again these facts indicate important gaps of physiopathological information in the current repositories on multigenic and complex diseases as those involving cationic biogenic amines. This delay in systematic integration of biogenic

Table 7.2 Diseases associated with histamine-related elements at NCBI gene and/or OMIM databases

Gene symbols	Related disease names	Disease OMIM
HDC	Tourette'ss yndrome	137580
HNMT	Susceptibilityt oa sthma	600807
HRH1	Susceptibilityt om ultiples clerosis	126200
AOC1	Cysticfi brosis	219700[a]
MAOB	Parkinson'sdi sease	168600[a]

[a]Not fully validated by the databases

Table 7.3 Functional Hia element-disease relationships revealed previously by text mining tools [61]

Diseasena me	DiseaseO MIM
Neurological diseases	
Hereditarye ssentialt remor	190300
Myoclonicdys tonia	159900
Narcolepsya ndc ataplexy	161400
NeuroinØammatory diseases	
Hereditary sensory and autonomic neuropathies IV and V	256800 and 608654
Multiples clerosis	126200
Immune/inØammatory diseases	
Crohndis eases/ulcerativec olitis	266600
Familial cold autoinflammatorys yndrome	120100
Idiopathica plastica nemia	609135
Infantile neurologic cutaneous articular syndrome	607115
Muckle-Wellss yndrome	191900
Psoriatica rthritis	607507
Systemic juvenile psoriatic arthritis	604302
Rare diseases	
Acutemye loidle ukemia	252270
Brugadas yndrome	601144
Congenitala drenalh yperplasia	145295
Familiar long QT syndrome	152427
Mastocytosis	154800
Vitamin D-dependent rickets type 2A	277440
VonW illebranddi sease	193400
Zollinger-Ellisons yndrome	131100

amine information is blocking the advance of biomedical knowledge in the field and consequently the development of new intervention strategies.

On the one hand, in a first attempt to reduce the "dark matter" of the histamine network, our group located around 20 diseases for which clear evidence exists in the involvement of histamine-related elements; this work was assisted by text mining tools [61] (Table 7.3). It is proof of concept that further biocomputational integrative efforts will give rise to emergent information on biogenic amine physiopathology.

On the other hand, in the last few years, our group has developed the tool PhenUMA [134]. This tool takes the advantages provided by biomedical ontologies, Gene Ontology (GO, http://geneontology.org) and Human Phenotype Ontology (HPO, http://human-phenotype-ontology.github.io/) [135]. These structures are standardized vocabularies organized in a hierarchical structure. Each one of the elements of these ontologies (called terms) is ordered from the most general (terms placed close to the root) to the more specific ones (terms placed close to the leaves). The use of standardized vocabularies allows the definition of functional profiles (GO) or phenotypic profiles (HPO) for both genes and diseases. A profile is built selecting the terms of the ontology that provided the best description of the functional processes or the phenotypic manifestation for a gene or a disease. Several approaches can be used over these profiles to establish similarities between them, and these approaches are called semantic similarity measures [134]. The objective of these measures is to score the similarity between two genes or diseases using terms of the ontology related to them.

PhenUMA uses HPO to establish phenotypic relationships among genes and diseases and integrates these relationships with functional and physical information. Additionally, PhenUMA allows the query of a set of genes, diseases, or phenotypes to retrieve networks that integrate different kinds of relationships with respect to the input data. Thus, in addition to GO and HPO, the tool works with data from international open source ontologies and databases, for instance, OMIM (http://omim.org/), Orphanet (ww.orpha.net), and STRING (http://string-db.org/). Currently PhenUMA is also open to the web. Figure 7.4 provides a scheme of PhenUMA database. Thus, it integrates known relationships among genes from several interactomes and public resources and, in addition, similarities among genes and diseases combining biomedical ontologies and semantic similarity measures [136, 137].

By using PhenUMA, the abovementioned information gap with respect to involvement of both PA- and Hia-related elements in human pathologies is inherited by our tool, making impossible to establish any relationships among these genes. As an example, Fig. 7.5 (panel A) shows the results obtained from the tool when asking for all DA receptor-related genes, compared to those obtained when asking for specific Hia-related elements listed in Table 7.1 (Fig. 7.5, panel B).

On the bases of this previous experience, we decided to combine both text mining and biomedical ontologies. Briefly the working plan was as follows. Firstly, text mining resources were used to retrieve diseases associated with amine-related genes from literature, and, secondly, new gene-gene relationships were predicted using the phenotypic profile associated with these diseases. Figure 7.6 is a scheme of the whole procedure used in the present work, including biocomputational workflows, steps of manual curation, and tests. The molecular elements (genes) considered as the seed of the search are shown in Table 7.1.

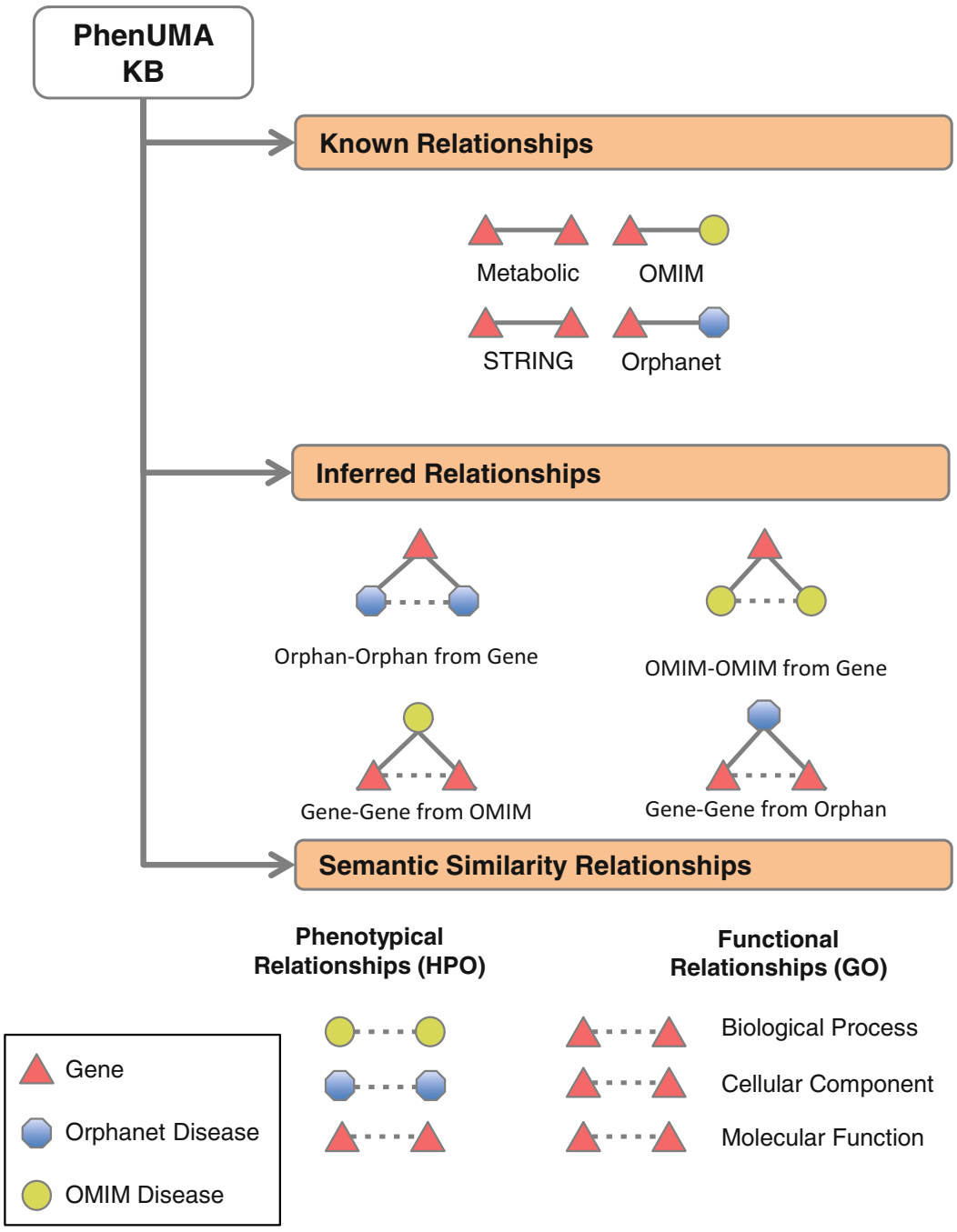


Fig. 7.4 PhenUMA working scheme. The tool integrates: known relationships between pairs of genes (STRING and metabolic relationships) and gene-disease relationships (OMIM and Orphanet), inferred relationships (*dashed lines*) between genes and/or diseases, and semantic similarity relationships among genes or diseases by using Gene Ontology (functional relationships) and Human Phenotype Ontology (phenotypic relationships)

Fig. 7.5 Representation of the network retrieved from PhenUMA initially asking for gene-gene pathologic relationships of the five dopamine receptors DRD1-5 (*panel A*) or for gene-gene pathological relationships of the four histamine receptors (HRH1-4 and HNMT) (*panel B*)

7.3.1 TextM iningPr ocedures

The first step of the workflow (Fig. 7.6) was the usage of tools that use text mining to retrieve gene-disease associations. Those genes considered as the seed of the search are shown in Table 7.1. Two tools were used in this stage: DisGeNET [138] and DISEASES [139].

DisGeNET contains 429,111 curated relationships among 17,181 genes and 14,619 diseases with a continuous updating system. The information came from

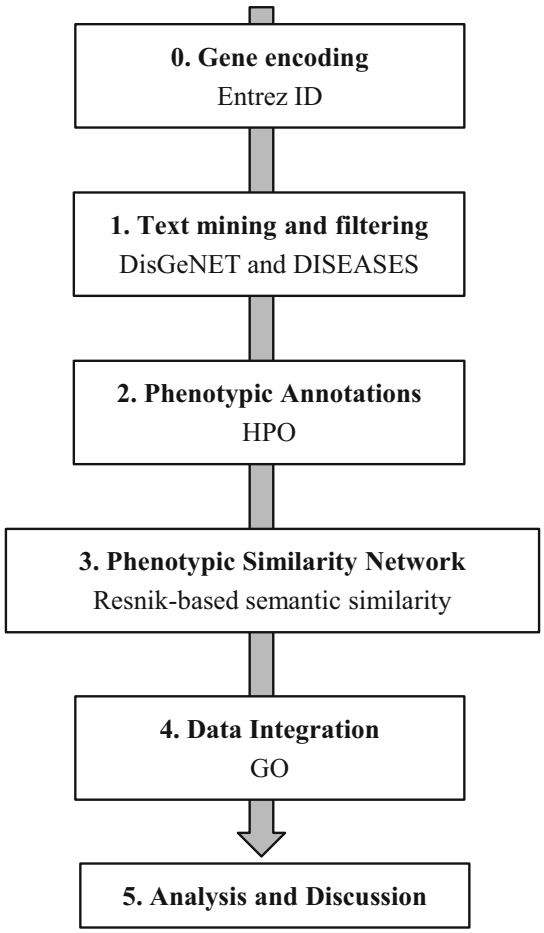


Fig.7.6 Workflowus edt o detect relationships establishedbya mine-related genes. The biocomputational resources used in each step are mentioned

other resources such as Comparative Toxicogenomics Database (CTD), UniProt, Rat Genome Database (RGD), and Mouse Genome Database (MGD) and from previous text mining initiatives like revised articles of genome-wide association studies (GWAS) and Genetic Association Database (GAD), Literature-derived Human Gene-DiseaseN etwork(LH GDN),a nd BeFree[140].

DISEASES provides an approach for retrieving gene-disease relationships from abstracts coming from Genetics Home Reference (GHR), UniProtKB, results of genome-wide association studies (GWAS) and DistiLD (http://distild.jensenlab.org), and mutation data from the Catalog of Somatic Mutations in Cancer (COSMIC).

Score values provided by both tools were normalized. Then a filter was applied; only those relationships within the highest 10 % of the normalized score ranking were considered from here on.

7.3.2 Phenotypic Annotation

At this stage of the workflow, the next objective is to define the phenotypic profiles associated with the genes related to the diseases gathered in the previous section. To do that, we used two types of relationships: gene-disease relationships

(obtained from the previous step) and disease-phenotype relationships. The latter were downloaded from HPO website, which provided the phenotypic profile associated with each OMIM disease. So, using the gene-disease associations, we are able to assign a set of phenotypes (HPO terms) for each amine-related gene. Some of these genes are associated with more than one disease; in these cases, the phenotypic profile is determined by the union of the profiles of the diseases associated with the gene.

7.3.3 SemanticSimilar ityR elationships

The next step was the calculation of the semantic similarity among amine-related genes and the rest of the genes annotated to the ontology. This measure allowed assigning a score to the overlapping between the phenotypic profiles of two genes; i.e., the similarity between the symptoms of two genes or diseases. In this case, the measure used [137] is based on the information content (IC) concept. IC is defined by $-log(probability(t))$, where t is a term of the ontology, and gives us an idea of the specificity of each phenotype, as explained previously [134]. The similarity between two genes is determined through the comparison of all the elements included in the phenotypic profile of both genes. All the genes with pathological information are compared, and the most significant values (in our case, those over the 98th percentile) are taken into account.

7.3.4 DataI ntegration

All these phenotypic similarity relationships among genes obtained in step 3 (Fig. 7.6) are integrated with functional information coming from the resources mentioned above. The objective of this part of the workflow is to highlight the relationships among genes that are involved in both the same functional processes and the same phenotypic characteristics. For this purpose, the functional semantic similarity among these genes was calculated using the Gene Ontology (GO) and its three sub-ontologies (biological process, cellular component, and molecular function).

7.3.5 DataPr ocessing

The retrieved information was manually analyzed. The resulting data and networks are briefly described and discussed in the next sections.

7.4 RetrievedI nformation: TheStar tingPoint for AMINETWORKING 1.0

Figure 7.7 is a representation of the network obtained following the procedure described in point 7.3.1 (text mining, normalization, and filtering). Elements of the different gene subgroups of Table 7.1 are differentially colored (see Fig. 7.7 caption). As the resulting edges between gene-disease pairs are too many to be properly observed as a network, they are also listed in Table 7.4. Nevertheless, the network topology is also informative as discussed below.

Results include many of the previously known relationships mentioned in Tables 7.2 and 7.3; they are represented as red edges. It is an internal validation of our strategy, as it is indeed able to automatically locate validated information from bibliography. Many other new relationships are inferred from our strategy when compared with the information present in the most common databanks of gene-

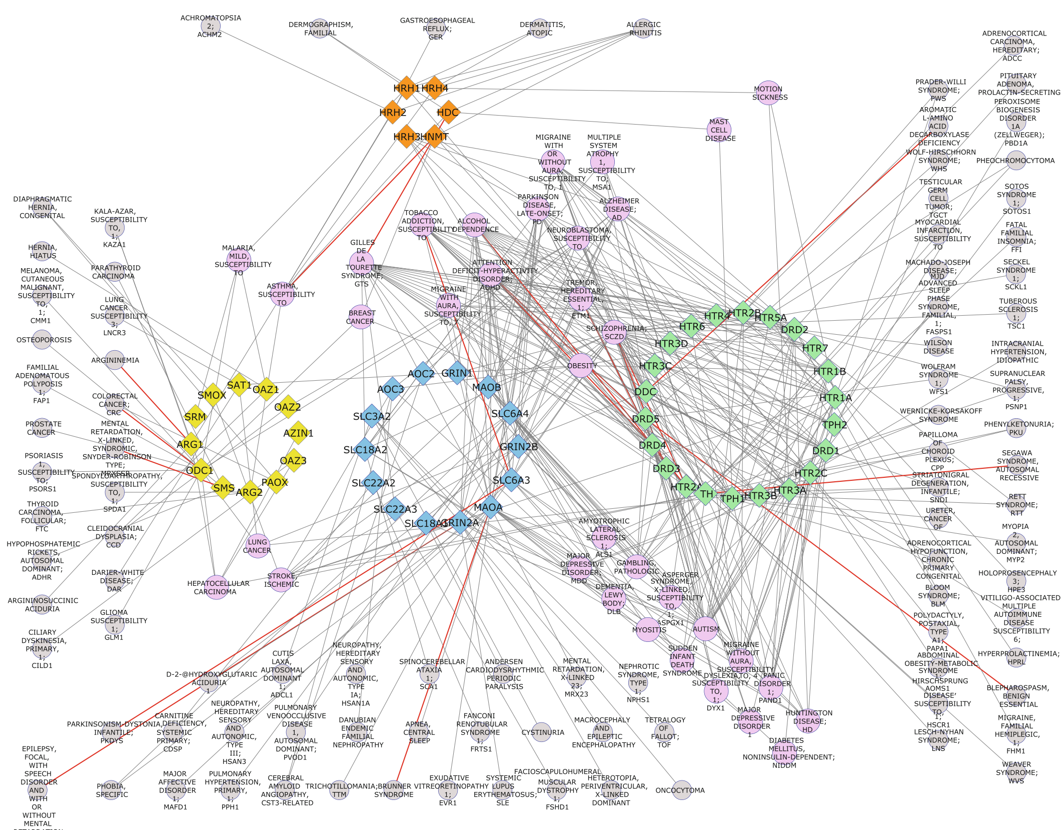


Fig. 7.7 Network obtained from our workflow (Fig. 7.6) by asking for relationships between any of the amine-related genes of Table I and human diseases. Elements of the different gene subgroups of Table 7.1 are differentially colored: *yellow*, PA-related elements; *orange*, Hia-related elements; *green*, DA/5′-HT; *blue*, shared-elements. Diseases related to more than one group of amine-related genes listed in Table 7.1 are colored in magenta. *Red edges* represent gene-disease relationships previously located and included in Tables 7.2 or 7.3

Table 7.4 Biogenic amine-related genes associated with human diseases deduced from the text mining procedure described in Sect. 7.3.1

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Abdominalobe sity-metabolic syndrome 1	605552	3358	HTR2C
Achromatopsia2	216900	3274	HRH2
Adrenocorticalc arcinoma, hereditary	202300	3357	HTR2B
Adrenocorticalh ypofunction, chronic primary congenital	103230	1644;7166	DDC;T PH1
Advanced sleep phase syndrome, familial, 1	604348	121278	TPH2
Alcoholde pendence	103780	1812;1813; 1814; 1815; 2902; 2903; 14811; 3350; 3351; 3356; 3358; 3359; 9177; 3363; 4128; 4129; 6531; 6532; 7166; 121278	DRD1; DRD2; DRD3; DRD4; GRIN1; GRIN2A; GRIN2B; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR3B; HTR7; MAOA; MAOB; SLC6A3; SLC6A4; TPH1; TPH2
Allergicrhi nitis	607154	3067;3176; 3 269; 3274; 59340	HDC; HNMT; HRH1; HRH2; HRH4
Alzheimer'sdis ease	104300	314;8639; 2903; 14811; 11255; 3350; 4128; 4129; 6532; 7054	AOC2; AOC3; GRIN2A; GRIN2B; HRH3; HTR1A; MAOA; MAOB; SLC6A4; TH
Amyotrophic lateral sclerosis 1	105400	4129;7054	MAOB;T H
Andersen cardiodysrhythmic periodic paralysis	170390	6570	SLC18A1
Apnea,c entrals leep	107640	4128	MAOA
Argininemia	207800	383	ARG1
Argininosuccinica ciduria	207900	383	ARG1
Aromatic L-amino acid decarboxylase deficiency	608643	1644	DDC
Asperger'ss yndrome, x-linked, susceptibility to, 1	300494	3356;6532	HTR2A;S LC6A4
Asthma, susceptibility to	600807	383; 384; 3067; 3176; 3269; 3274; 59340	ARG1; ARG2; HDC; HNMT; HRH1; HRH2; HRH4
Attention-deficit hyperactivity disorder	143465	1644;1812; 1813; 1814; 1815; 1816; 2903; 11255; 3351; 3356; 3358; 3360; 3363; 4128; 4129; 6531; 6532; 7054; 7166; 121278	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; GRIN2A; HRH3; HTR1B; HTR2A; HTR2C; HTR4; HTR7; MAOA; MAOB; SLC6A3; SLC6A4; TH; TPH1; TPH2

(continued)

Table 7.4 (continued)

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Autism	209850	1812;1813; 1814; 1815; 1816; 2903; 14811; 3350; 3351; 3356; 3357; 3358; 3359; 170572; 3361; 3363; 4128; 4129; 6570; 6531; 6532; 7166; 121278	DRD1; DRD2; DRD3; DRD4; DRD5; GRIN2A; GRIN2B; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR3A; HTR3C; HTR5A; HTR7; MAOA; MAOB; SLC18A1; SLC6A3; SLC6A4; TPH1; TPH2
Blepharospasm,be nign essential	606798	1816	DRD5
Blooms yndrome	210900	3361	HTR5A
Breast cancer	114480	3176; 3359; 4953	HNMT; HTR3A; ODC1
Brunners yndrome	300615	4128	MAOA
Carnitinede ficiency, systemic primary	212140	6582	SLC22A2
Cerebral amyloid angiopathy, cst3 related	105150	314;8639	AOC2;A OC3
Ciliarydys kinesia,pri mary,1	244400	51686	OAZ3
Cleidocranialdys plasia	119600	4946;4947	OAZ1;O AZ2
Colorectalc ancer	114500	4953;196743	ODC1;P AOX
Cutis laxa, autosomal dominant 1	123700	6520	SLC3A2
Cystinuria	220100	6520	SLC3A2
D-2-Hydroxyglutarica ciduria 1	600721	2903;14811	GRIN2A;G RIN2B
Danubian endemic familial nephropathy	124100	8639	AOC3
Darier-Whitedis ease	124200	6303	SAT1
Dementia, Lewy body	127750	4128; 4129; 6531; 6532; 7054	MAOA; MAOB; SLC6A3; SLC6A4; TH
Dermatitis,a topic	603165	3269;59340	HRH1;H RH4
Dermographism,f amilial	125635	3067;3269	HDC;H RH1
Diabetesme llitus, noninsulin-dependent	125853	8639;3358	AOC3;H TR2C
Diaphragmatiche rnia, congenital	142340	196743;54498	PAOX;S MOX
Dyslexia, susceptibility to, 1	127700	1814; 1815; 1816; 6531	DRD3; DRD4; DRD5; SLC6A3
Epilepsy, focal, with speech disorder and with or without mental retardation	245570	2903	GRIN2A
Exudativev itreoretinopathy1	133780	4128	MAOA
Facioscapulohumeral muscular dystrophy 1	158900	6581	SLC22A3

(continued)

Table 7.4 (continued)

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Familiala denomatous polyposis 1	175100	4953	ODC1
Fanconi renotubular syndrome 1	134600	6582	SLC22A2
Fatalf amilial insomnia	600072	7166	TPH1
Gambling,pa thologic	606349	1812;1813; 1814; 1815; 1816; 2902; 3351; 3356; 4128; 4129; 6531; 6532	DRD1; DRD2; DRD3; DRD4; DRD5; GRIN1; HTR1B; HTR2A; MAOA; MAOB; SLC6A3; SLC6A4
Gastroesophagealr eflux	109350	3274	HRH2
Gilles de la Tourette's syndrome	137580	1812;1813; 1814; 1815; 1816; 3067; 3350; 3351; 3356; 3357; 3358; 9177; 3361; 3363; 4128; 51686; 6581; 6531; 6532; 121278	DRD1; DRD2; DRD3; DRD4; DRD5; HDC; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR3B; HTR5A; HTR7; MAOA; OAZ3; SLC22A3; SLC6A3; SLC6A4; TPH2
Gliomas usceptibility1	137800	4953	ODC1
Hepatocellularc arcinoma	114550	383;51582; 51686; 4953; 7054	ARG1; AZIN1; OAZ3; ODC1; TH
Hernia, hiatus	142400	4946	OAZ1
Heterotopia,pe riventricular, x-linked dominant	300049	2903	GRIN2A
Hirschsprungdi sease, susceptibility to, 1	142623	7054	TH
Holoprosencephaly3	142945	3361	HTR5A
Huntingtondi sease	143100	1812;1813; 1816; 2903; 14811; 4129; 7054	DRD1; DRD2; DRD5; GRIN2A; GRIN2B; MAOB; TH
Hyperprolactinemia	615555	1813	DRD2
Hypophosphatemicric kets, autosomal dominant; ADHR	193100	6611	SMS
Intracranialh ypertension, idiopathic	243200	3358	HTR2C
Kala-azar, susceptibility to, 1	608207	383; 4953	ARG1; ODC1
Lesch-Nyhans yndrome	300322	1816	DRD5
Lungc ancer	211980	1644;3359; 4953; 6303	DDC; HTR3A; ODC1; SAT1
Lungc ancers usceptibility3	612571	6303	SAT1
Machado-Josephdi sease	109150	3350	HTR1A
Macrocephalya nd epileptic encephalopathy	606369	2903	GRIN2A
Majora ffectivedis order1	125480	6531	SLC6A3
Major depressive disorder	608516	3350; 3358; 3359; 4128	HTR1A; HTR2C; HTR3A; MAOA

(continued)

Table 7.4 (continued)

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Major depressive disorder 1	608520	1814; 1815; 3358; 4128	DRD3; DRD4; HTR2C; MAOA
Malaria, mild, susceptibility to	609148	3269; 4946; 4953; 6723	HRH1; OAZ1; ODC1; SRM
Mastc elldis ease	154800	3067;7166	HDC;T PH1
Melanoma, cutaneous malignant, susceptibility to, 1	155600	4953	ODC1
Mental retardation, x-linked 23	300046	4128	MAOA
Mental retardation, x-linked, syndromic, Snyder-Robinson type	309583	6611	SMS
Migraine with aura, susceptibility to, 7	609179	1813;1815; 3357; 3358; 4128; 6532	DRD2; DRD4; HTR2B; HTR2C; MAOA; SLC6A4
Migraine with or without aura, susceptibility to, 1	157300	1812;1813; 1814; 1815; 1816; 11255; 3350; 3351; 3356; 3358; 3359; 3363; 4128; 6532; 7166	DRD1; DRD2; DRD3; DRD4; DRD5; HRH3; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR7; MAOA; SLC6A4; TPH1
Migraine without aura, susceptibility to, 4	607501	1812;1813; 1814; 1815; 1816; 3350; 3351; 3356; 3357; 3358; 3362; 4128; 6532	DRD1; DRD2; DRD3; DRD4; DRD5; HTR1A; HTR1B; HTR2A; HTR2B; HTR2C; HTR6; MAOA; SLC6A4
Migraine, familial hemiplegic, 1	141500	1813	DRD2
Motion sickness	158280	3269;3350; 3359	HRH1;H TR1A;H TR3A
Multiple system atrophy 1, susceptibility to	146500	1644;1813; 7054; 7166	DDC; DRD2; TH; TPH1
Myocardiali nfarction, susceptibility to	608446	7054	TH
Myopia 2, autosomal dominant	160700	7054	TH
Myositis	160750	3359;6532	HTR3A;S LC6A4
Nephrotics yndrome,t ype1	256300	6520	SLC3A2
Neuroblastoma, susceptibility to	256700	1644; 3359; 4128; 4129; 4947; 4953; 7054	DDC; HTR3A; MAOA; MAOB; OAZ2; ODC1; TH
Neuropathy,h ereditary sensory and autonomic, type 1A	162400	4128	MAOA
Neuropathy, hereditary sensory and autonomic, type 3	223900	4128	MAOA

(continued)

Table 7.4 (continued)

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Obesity	601665	8639;1813; 11255; 3351; 3356; 3358; 6303	AOC3; DRD2; HRH3; HTR1B; HTR2A; HTR2C; SAT1
Oncocytoma	553000	6520	SLC3A2
Osteoporosis	166710	6611	SMS
Panicdis order 1	167870	1812;1813; 1815; 3350; 3351; 3356; 3358; 3359; 4128; 6532; 7166; 121278	DRD1; DRD2; DRD4; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; MAOA; SLC6A4; TPH1; TPH2
Papillomaof c horoidpl exus	260500	3358	HTR2C
Parathyroidc arcinoma	608266	6611	SMS
Parkinson's disease, late onset	168600	1644; 1812; 1813; 1814; 1815; 2903; 14811; 3176; 11255; 3350; 3359; 4128; 4129; 6531; 6532; 7054; 7166	DDC; DRD1; DRD2; DRD3; DRD4; GRIN2A; GRIN2B; HNMT; HRH3; HTR1A; HTR3A; MAOA; MAOB; SLC6A3; SLC6A4; TH; TPH1
Parkinsonism-dystonia, infantile	613135	6531	SLC6A3
Peroxisomebi ogenesis disorder 1a (Zellweger)	214100	121278	TPH2
Phenylketonuria	261600	7054;7166	TH;T PH1
Pheochromocytoma	171300	1644;7054	DDC;T H
Phobia,s pecific	608251	4128;6532	MAOA;S LC6A4
Pituitary adenoma, prolactin secreting	600634	1813	DRD2
Polydactyly, postaxial, type A1	174200	3362	HTR6
Prader-Willis yndrome	176270	3357	HTR2B
Prader-Willis yndrome	176270	3358	HTR2C
Prostate cancer	176807	4953	ODC1
Psoriasis1,s usceptibilityt o	177900	4953	ODC1
Pulmonaryh ypertension, primary, 1	178600	6532	SLC6A4
Pulmonaryv eno-occlusive disease 1, autosomal dominant	265450	6532	SLC6A4
Retts yndrome	312750	3363;7054	HTR7;T H
Schizophrenia	181500	1644;1812; 1 813; 1814; 1815; 1816; 2902; 2903; 14811; 3269; 11255; 3350; 3351; 3356; 3358; 3359; 9177; 200909; 3360; 3361; 3362; 3363; 4128; 4129; 6570; 6531; 6532; 7054; 7166; 121278; 6520	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; GRIN1; GRIN2A; GRIN2B; HRH1; HRH3; HTR1A; HTR1B; HTR2A; HTR2C; HTR3A; HTR3B; HTR3D; HTR4; HTR5A; HTR6; HTR7; MAOA; MAOB; SLC18A1; SLC6A3; SLC6A4; TH; TPH1; TPH2; SLC3A2

(continued)

Table 7.4 (continued)

Diseasena mes	OMIM	Entrezge ne	Genes ymbols
Seckels yndrome1	210600	3361	HTR5A
Segawa syndrome, autosomal recessive	605407	7054	TH
Sotoss yndrome 1	117550	1812	DRD1
Spondyloarthropathy, susceptibility to, 1	106300	384	ARG2
Striatonigralde generation, infantile	271930	1644; 7054; 7166	DDC; TH; TPH1
Stroke,is chemic	601367	8639;2903; 14811; 3350; 4129; 6532; 54498; 7054	AOC3; GRIN2A; GRIN2B; HTR1A; MAOB; SLC6A4; SMOX; TH
Sudden infant death syndrome	272120	3350; 4128; 6532; 7054; 7166;	HTR1A; MAOA; SLC6A4; TH; TPH1;
Supranuclearpa lsy, progressive, 1	601104	1813;7054	DRD2;T H
Systemiclupus e rythematosus	152700	2903	GRIN2A
Testicularge rmc ell tumor	273300	1812	DRD1
Thyroidc arcinoma,f ollicular	188470	384	ARG2
Tobaccoa ddiction, susceptibility to	188890	1644;1812; 1813; 1814; 1814; 1815; 1816; 3350; 3356; 9177; 3363; 4128; 6531; 6532; 7054; 7166	DDC; DRD1; DRD2; DRD3; DRD4; DRD5; HTR1A; HTR2A; HTR3B; HTR7; MAOA; SLC6A3; SLC6A4; TH; TPH1
Tremor, hereditary essential, 1	190300	1814; 3176; 6531	DRD3; HNMT;
Trichotillomania	613229	4128	MAOA
Tuberouss clerosis1	191100	3362	HTR6
Ureter,c ancerof	191600	7166	TPH1
Vitiligo-associated multiple autoimmune disease susceptibility 6	193200	1644	DDC
Weavers yndrome	277590	7054	TH
Wernicke-Korsakoffs yndrome	277730	7166	TPH1
Wilson disease	277900	1813	DRD2
Wolf-Hirschhorns yndrome	194190	1816	DRD5
Wolframs yndrome 1	222300	1816	DRD5

diseases relationships, as far as we know. Diseases represented as magenta circles are at least shared by two biogenic amine elements. After this text mining effort, it is clear that Hia and PA are not unplugged modules of the human *diseasome*.

Among these results there is an impressive quantity of inferred information that, of course, needs to be analyzed and curated by the "aminer community". We invite all our colleagues interested in biogenic amines to cooperate in the task. We are open to receive proof of concepts (references, personal communication, etc.), suggestions, and comments through the email account aminetworking@uma.es.

When focusing on the subsets of diseases related to the PA-related elements (Fig. 7.7 and Table 7.4), we can observe that, as expected, this module is mainly related to an important list of cancer types. The module of genes related to DDC-derived biogenic amines, as expected, is linked to a huge catalog of neurological and neuroendocrine disorders. Hia-related elements also establish multiple relationships with neurological disorders and the subnetwork of DDC-derived elements, but they seem to present a wider spectrum of affected organs/tissues. In the next paragraph, we will mainly focus our discussion on diseases involving elements from at least two different modules.

In fact, results remark the molecular complexity of several neurological diseases involving ten (or more) biogenic amine-related genes from different subsets. This is the case of (in alphabetic order) alcohol dependence, Alzheimer's diseases, attention-deficit hyperactivity disorder, autism, gambling (pathological), Gilles de la Tourette's syndrome, migraine (susceptibility with or without aura), panic disorder, Parkinson's disease, schizophrenia, and tobacco addiction (susceptibility) (Table 7.4). The concurrence of Hia-related genes and DDC product-related genes associated with different addictive behaviors is remarkable. It is a very interesting field that, for sure, involves both functional and even physical interactions between the gene products that are not fully characterized yet [141].

Taking into account the density of concurrent relationships between Hia-related genes and DDC-product-related genes under different neurological circumstances, it is clear that discussion of genetic/biochemical data of a single element in the context of a patient (or experimental model) of one of these diseases (or susceptibilities) should not ignore the probable contribution of the concurrent elements of its own module or the others. This strategy would help a personalized location of molecular contributors to a given patient and disease, as well as the efficient advance in the molecular characterization, prevention, and intervention of many other potential patients. This hypothesis encourages us to claim for the convenience of an integrative multinational project that incorporates information on patients of prevalent and emergent neurological abnormalities or susceptibilities to them.

Several interesting concurrence events on the same disease (or susceptibility to a given disease) are worth mentioning since it enriches our information of common physiopathological scenarios and could contribute to getting new insights for discussion of high-throughput results and personalized medicine initiatives. Briefly, we will mention them in the paragraphs below.

Our text mining search gave a few but very interesting diseases described as involving both PA and Hia-related elements (Table 7.4). Breast cancer has been

related to elements of both modules [53, 142, 143] and different elements of their metabolic pathways have been proposed as markers and/or therapeutic targets. DA/5′-HT-related elements have also been related to breast cancer progression, and DRD1 ligands have been proposed for breast cancer chemotherapy [144]. It is clear that the degree of knowledge on PA roles in cancer growth and progression is much higher for PA than for the other biogenic amines, in spite of the suggested roles and mechanisms for different Hia-related elements and tumor types from over 30 years ago [145]. This topic is being helped by recent findings of correlation between polymorphisms of Hia-related genes (HDC, HNMT, and HRH3) and breast cancer [48, 143]. As several metabolic points of mutual PA-Hia interference have been detected (as mentioned in introduction), we think that the PA-Hia-DA/5′-HT metabolic interplay in the context of human breast cancer deserves more attention from the "aminer community". This interest could be extended to other cancer types where PA and other biogenic amine elements could share the same environment (for instance, mastocytosis, myeloid leukemias, melanomas, and brain, lung, and gastrointestinal cancer types). The pattern of Hia receptors expressed in different human cancer types will be determinant for the Hia effects induced on each type. They could even be antagonic depending on both the receptor expression pattern and also on the specific proteome of each tumor. Fortunately, multiple ongoing initiatives are trying to integrate and classify cell/tissue-specific "omic" information that could be used to clarify the pleiotropic effects of Hia not only in cancer but in other biomedical scenarios [146–148].

A common phenotype to all biogenic amine subgroups is "susceptibility to asthma". Up to 5 Hia-related elements were associated. They include elements taking part in synthesis, degradation, and signaling [149–153], as well as arginase [154], which is the first and essential enzyme for both NO and PA synthesis in mammalian tissues [155]. Interestingly, it is well known that amino acids, arginine, and histidine, are considered essential amino acids during the first years of human life, just when the immune system is being conformed. It is tempting to speculate that there is a putative link between children's diets (in terms of Arg/His content) and immune characteristics of human beings. In fact, it is proposed that PA is important for the ideal children's immune system development [156]. Unfortunately, it indeed is not a well-explored scenario, in spite of the recent increasing evidence of the essential role of PA in nonmalignant myeloid cell differentiation [42].

Literature describes the PA-related element named antizyme inhibitor 2 (AZIN2) as a regulator of the intracellular vesicle trafficking involved in secretory processes (immune cells and others) [86, 157]. AZIN2 is an inactive ODC paralog/pseudogene, capable of binding ODC antizymes (AOZ). AZIN2 is also expressed in the brain (specific neurons of the hippocampus and cerebellum) [157]; in fact it is one of the few human organs expressing AZIN2 [158]. It is reported that AZIN2 accumulates during Alzheimer's disease progression [159]. Several membrane amine transporters such as SLC6A3, SLC6A4 (known as serotonin transporters), and SLC22A3 (or OCT3), which are also able to accept PA as ligands [29, 160], appear as related to multiple neurological diseases; for instance, addictions, ADHD [161], autism [162], dyslexia, dementia associated to Lewy body [163], Gilles de la

Tourette's syndrome [164], major affective disorders [165], panic [166], Parkinson's disease (late onset) [167], phobias [168], schizophrenia [169], susceptibility to Asperger's syndrome [170], susceptibility to migraine [171], and tremor (Table 7.4) [164]. This concurrence of PA and neurotransmitters at the level of transport systems suggests the possibility of still undisclosed aspects of PA influence on neurological problems and neurodegenerative diseases, as well as in other physiological scenarios where they also coincide (for instance, intestine).

We can also find interesting intermodule relationships outside the brain. From Table 7.4 and even Fig. 7.7, it can be deduced that all gene subgroups have elements related to strokes. These are solute carrier 6A4, the NMDA-polypeptides GRIN2A and GRIN2B, and several amine oxidases, MAOB (monoamine oxidase), semicarbazide-sensitive amine oxidase (AOC3), and SMOX (a specific spermine oxidase); all of them are ROS sources covering a wide spectrum of amine substrates (Fig. 7.1). In fact, Igarashi and Kashigagi propose the use of polyamine metabolites as markers for stroke and renal failure [172]. Ictus is also a pathological event involving arginine/NO metabolism-, immune-, and stress response-related genes. Consequently, a cross talk among different biogenic amines can be hypothesized [173].

Concurrence of different biogenic amine degradation systems could also play importance in other pathological processes involving simultaneously inflammation and proliferation, for instance, inflammation-associated carcinogenesis, tumor growth progression, hepatic injury, and so on, as discussed elsewhere [73]. This is also the case of parasitic infections (i.e., malaria, Table 7.4). Our search found the phenotype "susceptibility to malaria" as related to several PA elements, spermidine synthase (SRM), antizyme 1 (OAZ1), and ODC as well as histamine receptor type 1 (HRH1). Parasite requires PA to proliferate, and the surrounding inflammatory cells constitute an important source of other biogenic amines and ROS, with consequences that are still not well evaluated [72, 174].

Regarding the different malignancies mentioned in Table 7.4 associated with several biogenic amines, some of them had been reviewed in previous works, for instance, mast cell and rare gastric malignancies [61]. Up to seven biogenic amine-related elements coming from all four gene subsets (Tables 7.1 and 7.4) were related to "susceptibility to neuroblastoma." As mentioned above, it is one of the diseases where PA synthesis inhibition is getting promising success in clinical work [39], which is in agreement with the recently demonstrated tight coordination between energy metabolism, protein synthesis, and PA synthesis in this pediatric cancer [127]. At present, we are interested in locating metabolic/genetic features contributing to the differential sensitivity to PA synthesis inhibitors observed among patients [39]. In the data analyses, integration of molecular and functional information of the other biogenic amine gene subsets is considered very convenient.

In addition to neuroblastoma, Table 7.4 includes many other low-prevalence and rare diseases. Approaches capable of saving biological samples and experiments are especially valuable for these diseases. Many of them had been located previously [61, 175], but the present work adds several to the list. As members of "Centro de Investigación Biomédica En Red en Enfermedades Raras," the Spanish institute for research in rare diseases, we hope this information can be beneficial to

many biomedical research and clinical groups working on these pathologies that are currently worldwide considered a health priority.

As mentioned in the previous section, the computer-assisted procedure described in Sect. 7.3 had as its objective the location of gene-gene interactions involving both functional and phenotypic concurrence between biogenic amine-related and other human genes, as they must reflect biochemical, molecular, and cellular interactions that could be involved in human diseases. This step could provide useful insights for genomic result analyses and initiatives of personalized medicine.

Figure 7.8 shows the crude results of the phenotypic similarity network of all gene-gene relationships inferred from the procedures described in Sects. 7.3.3 and 7.3.4. Nodes are human genes sharing phenotypic similarities with any of the genes listed in Table 7.4 and scored over the 98th percentile. In bright blue, the biogenic amine-related genes are shown. The other human genes phenotypically related to biogenic amine genes are depicted in gray. Among the results there is an impressive quantity of inferred information. The full analysis and validation of these sets of results is out of the frame of the present work.

The phenotypic relationships were therefore enriched with functional information as described in Sect. 7.3.4. Thus, Fig. 7.9 represents gene-gene relationships among genes sharing both phenotypic and annotated functional characteristics. The quality

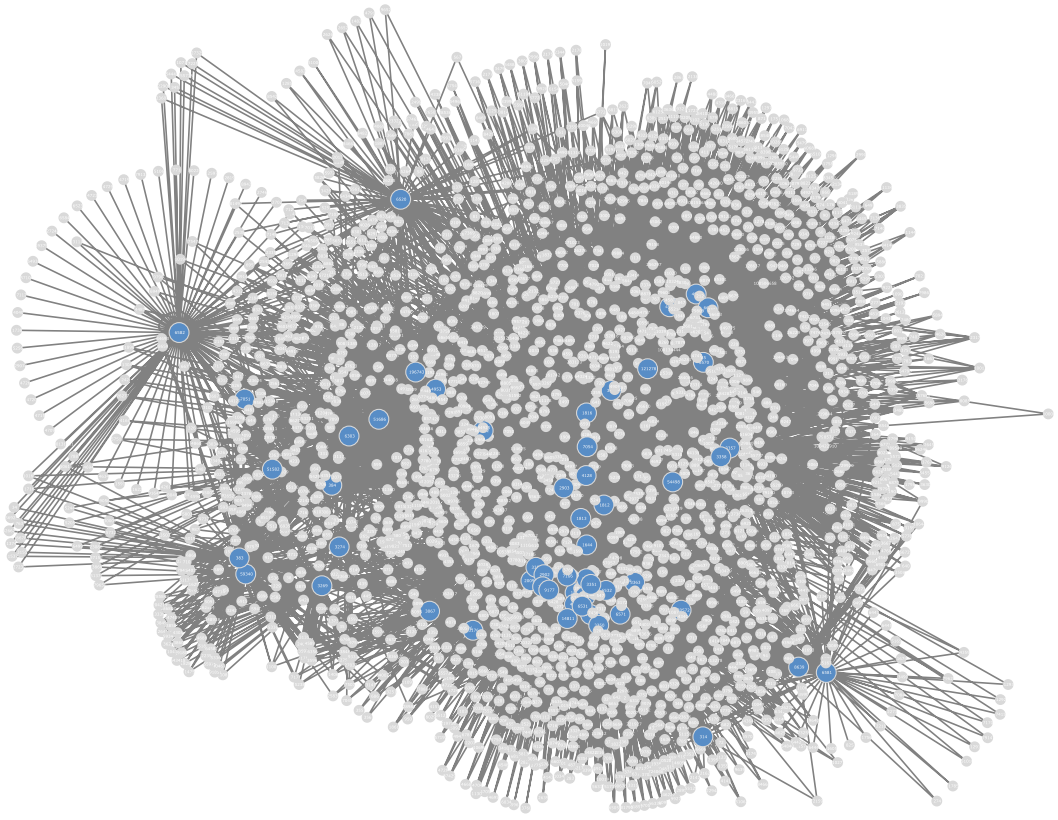


Fig. 7.8 Network of phenotypic relationships between biogenic amine genes and any other human gene. Biogenic amine genes included in Table 7.1 are colored as *blue spheres*

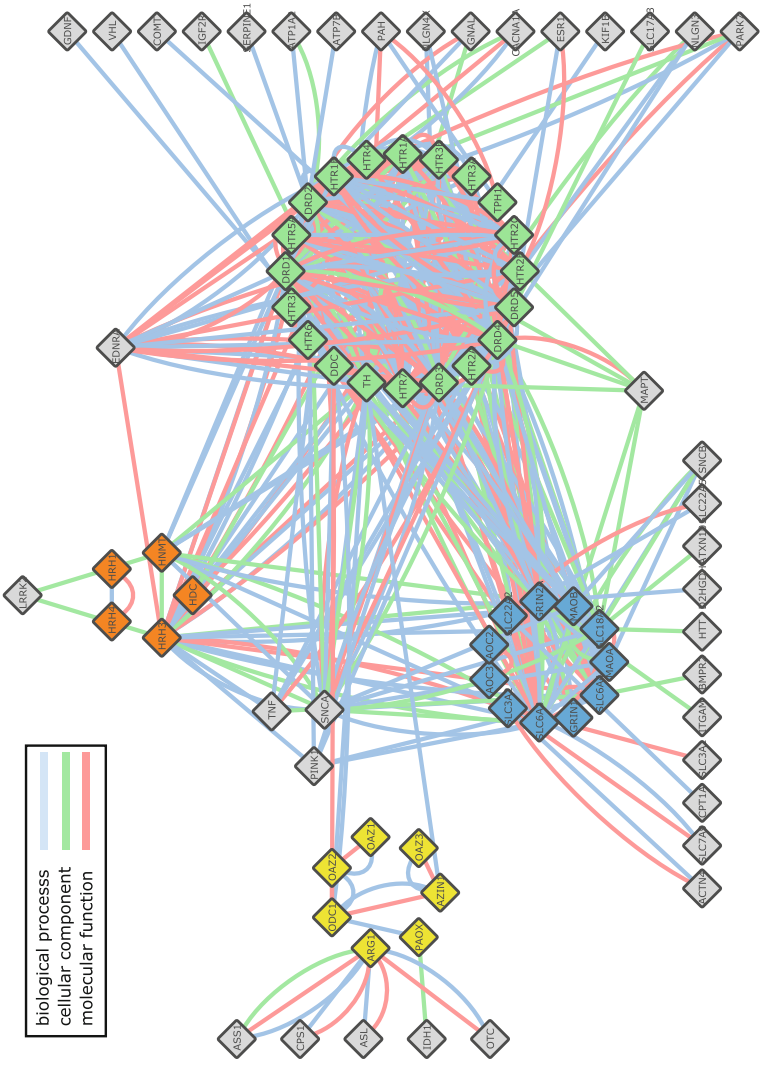


Fig. 7.9 Network of biogenic amine-related genes related to other human genes by sharing both phenotypic and functional characteristics. Elements of the different gene subgroups (Table 7.1) are differentially colored: *yellow*, PA-related elements; *orange*, Hia-related elements; *green*, DA/5′-HT; *blue*, shared-elements. The functional characteristics (*edges*), named as in the Gene Ontology, are also distinguished with different colors

of the functional interactions (edges in the figure) is distinguished by colors, as well as the different amine modules (as specified in the lettering included in the figure and its footnote). Many shared (GO-defined) biological processes (blue edges) can be observed, as well as many (GO-defined) molecular functions (pink edges) that are shared specially among neurotransmitter-related elements. Interactions with other human genes (gray diamonds) are also detected.

The full set of information will be included in the first version of an integrative platformonbiogenicamine-relatedbiomedicalinformationnamedAMINETWORKING 1.0, which is an ongoing project at present. These sets of information should be curated and validated by the "aminer community". We invite all our colleagues interested in biogenic amines to cooperate in the task. We are open to receive proof of concepts (references, personal communication, etc.), suggestions, and comments at present through the email account aminetworking@uma.es.

7.5 ConcludingR emarks

Brain function and associated diseases, immunology and immune diseases, and cancer and rare diseases are all currently biomedical international priorities. Biogenic amines are involved in all of them, and sometimes the cross talk events among the different amine subnetworks are proposed to play an important role in the cross talk among the processes, for instance, in the case of the brain-gut axis and several immune, neurological, and neuroendocrine diseases and in the immune system-cancer/parasite interaction events during cancer/infection progression.

The present work underlies our claim to go toward the construction of a database specialized on the whole biogenic amine network. In the future, it should include information on genes, their variants and associated RNA species, tissue-specific expression data, information of the encoded protein structures and their kinetic data, protein-protein interaction data, intracellular location, as well as gene associations with human phenotypes/diseases, and gene (protein target)-drug associations. This work is just a first step that could really grow if it were supported by the collaboration of international groups with expertise in different biogenic amine modules and/or diseases. We could contribute with our previous experience in the development of biocomputational predictive models and tools for location and integration of metabolic and enzymatic data [176, 177], gene-phenotype/disease and protein/ligand association tools and analyses, [134, 178–180], and development of social curation tools [181]. As mentioned, the data sets behind the figures shown in the present work could be named as AMINETWORKING 1.0 and will be available to any researcher or clinician interested in the relationships mentioned/depicted in Table 7.4 and Figs. 7.7, 7.8, and 7.9. We have opened the email address aminetworking@uma.es to encourage the interchange of information and comments on this initiative with other research groups.

As a consequence of the exponentially growing *omics* initiatives, this kind of biocomputational support is becoming essential to get an efficient yield of the analytic

investments. On the one hand, data on human genomic variants and transcript variability on the elements included in Table 7.1 is rapidly increasing, and the integration of this information could provide valuable support for personalized (predictive) medicine in a wide catalog of diseases for which more than a single biogenic amine module is involved. In fact, it could help to combine therapies or more accurate disease susceptibility predictions. On the other hand, the integration of drug-target information in a single repository could give light to new drug discovery initiatives and combined therapies, due to the similarity among structures, reaction mechanisms, interactions, and requirements of elements of different subnetworks, as experienced by our group on different occasions [113, 182]. In this sense, taking into account the intracellular location and the tissue-specific molecular and functional characteristic of the different targets is a very important issue to progress toward the right direction as also claimed in a previous work by our group [183] Unfortunately many of these translational possibilities are being currently delayed without the support of an integrative platform as the one we propose here.

Acknowledgments This work was supported by Grants SAF2011-26518 (MINECO, Spain) and PAIDI Grant P10-CVI6585 (Andalusian Government). We also thank the support (RRHH) from CIBERER, as well as from the University of Málaga and Andalucia Tech facilities. CIBERER is an initiative of Instituto de Salud Carlos III.

References

1. Amadasi A, Bertoldi M, Contestabile R, Bettati S, Cellini B, di Salvo ML, et al. Pyridoxal 5′-phosphate enzymes as targets for therapeutic agents. Curr Med Chem. 2007;14:1291–324. doi:10.2174/092986707780597899.
2. Agostinelli E, Seiler N. Non-irradiation-derived reactive oxygen species (ROS) and cancer: therapeutic implications. Amino Acids. 2006;31:341–55. doi:10.1007/s00726-005-0271-8.
3. Kirschner KM, Braun JFW, Jacobi CL, Rudigier LJ, Persson AB, Scholz H. Amine oxidase copper-containing 1 (AOC1) is a downstream target gene of the Wilms tumor protein, WT1, during kidney development. J Biol Chem. 2014;289:24452–62. doi:10.1074/jbc.m114.564336.
4. Pegg AE. Mammalian polyamine metabolism and function. IUBMB Life. 2009;61:880–94. doi:10.1002/iub.230.
5. Casero RA, Pegg AE. Polyamine catabolism and disease. Biochem J. 2009;421:323–38. doi:10.1042/bj20090598.
6. Stegaev V, Nies AT, Porola P, Mieliauskaite D, Sánchez-Jiménez F, Urdiales JL, et al. Histamine transport and metabolism are deranged in salivary glands in Sjogren's syndrome. Rheumatology (Oxford). 2013;52:1599–608. doi:10.1093/rheumatology/ket188.
7. Battaglia V, DeStefano SC, Murray-Stewart T, Casero RA. Polyamine catabolism in carcinogenesis: potential targets for chemotherapy and chemoprevention. Amino Acids. 2014;46:511–9. doi:10.1007/s00726-013-1529-6.
8. Bertoldi M. Mammalian Dopa decarboxylase: structure, catalytic activity and inhibition. Arch Biochem Biophys. 2014;546:1–7. doi:10.1016/j.abb.2013.12.020.
9. Cederbaum SD, Yu H, Grody WW, Kern RM, Yoo P, Iyer RK. Arginases I and II: do their functions overlap? Mol Genet Metab. 2004;81 Suppl 1:S38–44. doi:10.1016/j.ymgme.2003.10.012.

10. Satriano J. Arginine pathways and the inflammatory response: interregulation of nitric oxide and polyamines: review article. Amino Acids. 2004;26:321–9. doi:10.1007/s00726-004-0078-4.
11. Perez-Leal O, Merali S. Regulation of polyamine metabolism by translational control. Amino Acids. 2012;42:611–7. doi:10.1007/s00726-011-1036-6.
12. Kahana C. Regulation of cellular polyamine levels and cellular proliferation by antizyme and antizyme inhibitor. Essays Biochem. 2009;46:47–61. doi:10.1042/bse0460004.
13. López-Contreras AJ, Ramos-Molina B, Cremades A, Peñafiel R. Antizyme inhibitor 2: molecular, cellular and physiological aspects. Amino Acids. 2010;38:603–11. doi:10.1007/s00726-009-0419-4.
14. Pegg AE. The function of spermine. IUBMB Life. 2014;66:8–18. doi:10.1002/iub.1237.
15. Komori H, Nitta Y, Ueno H, Higuchi Y. Structural study reveals that Ser-354 determines substrate specificity on human histidine decarboxylase. J Biol Chem. 2012;287:29175–83. doi:10.1074/jbc.m112.381897.
16. Heidari A, Tongsook C, Najafipour R, Musante L, Vasli N, Garshasbi M, et al. Mutations in the histamine N-methyltransferase gene, HNMT, are associated with nonsyndromic autosomal recessive intellectual disability. Hum Mol Genet. 2015;24:5697–710. doi:10.1093/hmg/ddv286.
17. Panula P, Chazot PL, Cowart M, Gutzmer R, Leurs R, Liu WLS, et al. International union of basic and clinical pharmacology. XCVIII. Histamine receptors. Pharmacol Rev. 2015;67:601–55. doi:10.1124/pr.114.010249.
18. Tekin I, Roskoski R, Carkaci-Salli N, Vrana KE. Complex molecular regulation of tyrosine hydroxylase. J Neural Transm. 2014;121:1451–81. doi:10.1007/s00702-014-1238-7.
19. Bennett PJ, McMahon WM, Watabe J, Achilles J, Bacon M, Coon H, et al. Tryptophan hydroxylase polymorphisms in suicide victims. Psychiatr Genet. 2000;10:13–7.
20. Montioli R, Dindo M, Giorgetti A, Piccoli S, Cellini B, Voltattorni CB. A comprehensive picture of the mutations associated with aromatic amino acid decarboxylase deficiency: from molecular mechanisms to therapy implications. Hum Mol Genet. 2014;23:5429–40. doi:10.1093/hmg/ddu266.
21. Beaulieu J-M, Espinoza S, Gainetdinov RR. Dopamine receptors—IUPHAR review 13. Br J Pharmacol. 2015;172:1–23. doi:10.1111/bph.12906.
22. Hartig PR, Hoyer D, Humphrey PP, Martin GR. Alignment of receptor nomenclature with the human genome: classification of 5-HT1B and 5-HT1D receptor subtypes. Trends Pharmacol Sci. 1996;17:103–5. doi:10.1016/0165-6147(96)30002-3.
23. Hoyer D, Clarke DE, Fozard JR, Hartig PR, Martin GR, Mylecharane EJ, et al. International Union of Pharmacology classification of receptors for 5-hydroxytryptamine (Serotonin). Pharmacol Rev. 1994;46:157–203.
24. Tohda M. Serotonin 2C receptor as a superhero: diversities and talents in the RNA universe for editing, variant, small RNA and other expected functional RNAs. J Pharmacol Sci. 2014;126:321–8. doi:10.1254/jphs.14r06cr.
25. García-Martín E, Martínez C, Serrador M, Alonso-Navarro H, Ayuso P, Navacerrada F, et al. Diamine oxidase ꞩ10156191 and rs2052129 variants are associated with the risk for migraine. Headache. 2015;55:276–86. doi:10.1111/head.12493.
26. Kaitaniemi S, Elovaara H, Grön K, Kidron H, Liukkonen J, Salminen T, et al. The unique substrate specificity of human AOC2, a semicarbazide-sensitive amine oxidase. Cell Mol Life Sci. 2009;66:2743–57. doi:10.1007/s00018-009-0076-5.
27. Buffoni F. Semicarbazide-sensitive amine oxidases: some biochemical properties and general considerations. Prog Brain Res. 1995;106:323–31.
28. Bortolato M, Chen K, Shih JC. Monoamine oxidase inactivation: from pathophysiology to therapeutics. Adv Drug Deliv Rev. 2008;60:1527–33. doi:10.1016/j.addr.2008.06.002.
29. Abdulhussein AA, Wallace HM. Polyamines and membrane transporters. Amino Acids. 2014;46:655–60. doi:10.1007/s00726-013-1553-6.
30. German CL, Baladi MG, McFadden LM, Hanson GR, Fleckenstein AE. Regulation of the dopamine and vesicular monoamine transporters: pharmacological targets and implications for disease. Pharmacol Rev. 2015;67:1005–24. doi:10.1124/pr.114.010397.

31. Lazarov NE, Reindl S, Fischer F, Gratzl M. Histaminergic and dopaminergic traits in the human carotid body. Respir Physiol Neurobiol. 2009;165:131–6. doi:10.1016/j.resp.2008.10.016.
32. Johnson KB, Petersen-Jones H, Thompson JM, Hitomi K, Itoh M, Bakker ENTP, et al. Vena cava and aortic smooth muscle cells express transglutaminases 1 and 4 in addition to transglutaminase 2. Am J Physiol Heart Circ Physiol. 2012;302:H1355–66. doi:10.1152/ajpheart.00918.2011.
33. Ientile R, Currò M, Caccamo D. Transglutaminase 2 and neuroinflammation. Amino Acids. 2015;47:19–26. doi:10.1007/s00726-014-1864-2.
34. Qiao S-W, Piper J, Haraldsen G, Oynebråten I, Fleckenstein B, Molberg O, et al. Tissue transglutaminase-mediated formation and cleavage of histamine-gliadin complexes: biological effects and implications for celiac disease. J Immunol. 2005;174:1657–63.
35. Williams K. Extracellular Modulation of NMDA Receptors. In: Van Dongen AM, editor. Biology of the NMDA receptor. Boca Raton: CRC Press/Taylor & Francis; 2009. http://www.ncbi.nlm.nih.gov/books/NBK5272/.
36. Hirose T, Saiki R, Yoshizawa Y, Imamura M, Higashi K, Ishii I, et al. Spermidine and Ca(2+), but not Na(+), can permeate NMDA receptors consisting of GluN1 and GluN2A or GluN2B in the presence of Mg(2+). Biochem Biophys Res Commun. 2015;463:1190–5. doi:10.1016/j.bbrc.2015.06.081.
37. Landau G, Bercovich Z, Park MH, Kahana C. The role of polyamines in supporting growth of mammalian cells is mediated through their requirement for translation initiation and elongation. J Biol Chem. 2010;285:12474–81. doi:10.1074/jbc.m110.106419.
38. Raj KP, Zell JA, Rock CL, McLaren CE, Zoumas-Morse C, Gerner EW, et al. Role of dietary polyamines in a phase III clinical trial of difluoromethylornithine (DFMO) and sulindac for prevention of sporadic colorectal adenomas. Br J Cancer. 2013;108:512–8. doi:10.1038/bjc.2013.15.
39. Saulnier Sholler GL, Gerner EW, Bergendahl G, MacArthur RB, VanderWerff A, Ashikaga T, et al. A phase I trial of DFMO targeting polyamine addiction in patients with relapsed/refractory neuroblastoma. PLoS One. 2015;10, e0127246. doi:10.1371/journal.pone.0127246.
40. Birkholtz L-M, Williams M, Niemand J, Louw AI, Persson L, Heby O. Polyamine homoeostasis as a drug target in pathogenic protozoa: peculiarities and possibilities. Biochem J. 2011;438:229–44. doi:10.1042/bj20110362.
41. Fiori LM, Turecki G. Implication of the polyamine system in mental disorders. J Psychiatry Neurosci. 2008;33:102–10.
42. García-Faroldi G, Rodríguez CE, Urdiales JL, Pérez-Pomares JM, Dávila JC, Pejler G, et al. Polyamines are present in mast cell secretory granules and are important for granule homeostasis. PLoS One. 2010;5:e15071. doi:10.1371/journal.pone.0015071.
43. Masuko T, Kusama-Eguchi K, Sakata K, Kusama T, Chaki S, Okuyama S, et al. Polyamine transport, accumulation, and release in brain. J Neurochem. 2003;84:610–7. doi:10.1046/j.1471-4159.2003.01558.x.
44. Poulin R, Casero RA, Soulet D. Recent advances in the molecular biology of metazoan polyamine transport. Amino Acids. 2012;42:711–23. doi:10.1007/s00726-011-0987-y.
45. Miller-Fleming L, Olin-Sandoval V, Campbell K, Ralser M. Remaining mysteries of molecular biology: the role of polyamines in the cell. J Mol Biol. 2015;427:3389–406. doi:10.1016/j.jmb.2015.06.020.
46. Thurmond RL. The histamine H4 receptor: from orphan to the clinic. Front Pharmacol. 2015;6:65. doi:10.3389/fphar.2015.00065.
47. Burban A, Faucard R, Armand V, Bayard C, Vorobjev V, Arrang J-M. Histamine potentiates N-methyl-D-aspartate receptors by interacting with an allosteric site distinct from the polyamine binding site. J Pharmacol Exp Ther. 2010;332:912–21. doi:10.1124/jpet.109.158543.
48. Martinel Lamas DJ, Rivera ES, Medina VA. Histamine H4 receptor: insights into a potential therapeutic target in breast cancer. Front Biosci (Schol Ed). 2015;7:1–9. doi:10.2741/420.
49. Glatzer F, Gschwandtner M, Ehling S, Rossbach K, Janik K, Klos A, et al. Histamine induces proliferation in keratinocytes from patients with atopic dermatitis through the histamine 4 receptor. J Allergy Clin Immunol. 2013;132:1358–67. doi:10.1016/j.jaci.2013.06.023.

50. Fajardo I, Urdiales JL, Medina MA, Sanchez-Jimenez F. Effects of phorbol ester and dexamethasone treatment on histidine decarboxylase and ornithine decarboxylase in basophilic cells. Biochem Pharmacol. 2001;61:1101–6. doi:10.1016/s0006-2952(01)00567-6.
51. García-Faroldi G, Correa-Fiz F, Abrighach H, Berdasco M, Fraga MF, Esteller M, et al. Polyamines affect histamine synthesis during early stages of IL-3-induced bone marrow cell differentiation. J Cell Biochem. 2009;108:261–71. doi:10.1002/jcb.22246.
52. Ruiz-Chica AJ, Soriano A, Tuñón I, Sánchez-Jiménez FM, Silla E, Ramírez FJ. FT-Raman and QM/MM study of the interaction between histamine and DNA. Chem Phys. 2006;324:579–90. doi:10.1016/j.chemphys.2005.11.022.
53. Medina V, Croci M, Crescenti E, Mohamad N, Sanchez-Jiménez F, Massari N, et al. The role of histamine in human mammary carcinogenesis: H3 and H4 receptors as potential therapeutic targets for breast cancer treatment. Cancer Biol Ther. 2008;7:28–35. doi:10.4161/cbt.7.1.5123.
54. Panula P, Nuutinen S. The histaminergic network in the brain: basic organization and role in disease. Nat Rev Neurosci. 2013;14:472–87. doi:10.1038/nrn3526.
55. Panula P, Sundvik M, Karlstedt K. Developmental roles of brain histamine. Trends Neurosci. 2014;37:159–68. doi:10.1016/j.tins.2014.01.001.
56. Ellenbroek BA, Ghiabi B. The other side of the histamine H3 receptor. Trends Neurosci. 2014;37:191–9. doi:10.1016/j.tins.2014.02.007.
57. Flik G, Folgering JHA, Cremers TIHF, Westerink BHC, Dremencov E. Interaction between brain histamine and serotonin, norepinephrine, and dopamine systems: in vivo microdialysis and electrophysiology study. J Mol Neurosci. 2015;56:320–8. doi:10.1007/s12031-015-0536-3.
58. Passani MB, Blandina P. Histamine receptors in the CNS as targets for therapeutic intervention. Trends Pharmacol Sci. 2011;32:242–9. doi:10.1016/j.tips.2011.01.003.
59. Thurmond RL, Gelfand EW, Dunford PJ. The role of histamine H1 and H4 receptors in allergic inflammation: the search for new antihistamines. Nat Rev Drug Discov. 2008;7:41–53. doi:10.1038/nrd2465.
60. Zampeli E, Tiligada E. The role of histamine H4 receptor in immune and inflammatory disorders. Br J Pharmacol. 2009;157:24–33. doi:10.1111/j.1476-5381.2009.00151.x.
61. Pino-Ángeles A, Reyes-Palomares A, Melgarejo E, Sánchez-Jiménez F. Histamine: an undercover agent in multiple rare diseases? J Cell Mol Med. 2012;16:1947–60. doi:10.1111/j.1582-4934.2012.01566.x.
62. Yang XD, Ai W, Asfaha S, Bhagat G, Friedman RA, Jin G, et al. Histamine deficiency promotes inflammation-associated carcinogenesis through reduced myeloid maturation and accumulation of CD11b+ Ly6G+ immature myeloid cells. Nat Med. 2011;17:87–95. doi:10.1038/nm.2278.
63. Chen D, Aihara T, Zhao C-M, Håkanson R, Okabe S. Differentiation of the gastric mucosa. I. Role of histamine in control of function and integrity of oxyntic mucosa: understanding gastric physiology through disruption of targeted genes. Am J Physiol Gastrointest Liver Physiol. 2006;291:G539–44. doi:10.1152/ajpgi.00178.2006.
64. Nozaki K, Weis V, Wang TC, Falus A, Goldenring JR. Altered gastric chief cell lineage differentiation in histamine-deficient mice. Am J Physiol Gastrointest Liver Physiol. 2009;296:G1211–20. doi:10.1152/ajpgi.90643.2008.
65. Pagotto RM, Monzón C, Moreno MB, Pignataro OP, Mondillo C. Proliferative effect of histamine on MA-10 Leydig tumor cells mediated through HRH2 activation, transient elevation in cAMP production, and increased extracellular signal-regulated kinase phosphorylation levels. Biol Reprod. 2012;87:150. doi:10.1095/biolreprod.112.102905.
66. Arreola R, Becerril-Villanueva E, Cruz-Fuentes C, Velasco-Velázquez MA, Garcés-Alvarez ME, Hurtado-Alvarado G, et al. Immunomodulatory effects mediated by serotonin. J Immunol Res. 2015;2015:1–21. doi:10.1155/2015/354957.
67. Choi MR. Renal dopaminergic system: Pathophysiological implications and clinical perspectives. World J Nephrol. 2015;4:196–212. doi:10.5527/wjn.v4.i2.196.
68. Gratwicke J, Jahanshahi M, Foltynie T. Parkinson's disease dementia: a neural networks perspective. Brain. 2015;138:1454–76. doi:10.1093/brain/awv104.

69. Johnston JD, Skene DJ. 60 YEARS OF NEUROENDOCRINOLOGY: Regulation of mammalian neuroendocrine physiology and rhythms by melatonin. J Endocrinol. 2015;226:T187–98. doi:10.1530/joe-15-0119.
70. Deidda G, Bozarth IF, Cancedda L. Modulation of GABAergic transmission in development and neurodevelopmental disorders: investigating physiology and pathology to gain therapeutic perspectives. Front Cell Neurosci. 2014;8:119. doi:10.3389/fncel.2014.00119.
71. Yuan H, Low C-M, Moody OA, Jenkins A, Traynelis SF. Ionotropic GABA and glutamate receptor mutations and human neurologic diseases. Mol Pharmacol. 2015;88:203–17. doi:10.1124/mol.115.097998.
72. Sánchez-Jiménez F, Ruiz-Pérez MV, Urdiales JL, Medina MA. Pharmacological potential of biogenic amine-polyamine interactions beyond neurotransmission. Br J Pharmacol. 2013;170:4–16. doi:10.1111/bph.12109.
73. Sánchez-Jiménez F, Montañez R, Correa-Fiz F, Chaves P, Rodríguez-Caso C, Urdiales JL, et al. The usefulness of post-genomics tools for characterization of the amine cross-talk in mammalian cells. Biochem Soc Trans. 2007;35:381–5. doi:10.1042/bst0350381.
74. Medina MA, Correa-Fiz F, Rodríguez-Caso C, Sánchez-Jiménez F. A comprehensive view of polyamine and histamine metabolism to the light of new technologies. J Cell Mol Med. 2005;9:854–64. doi:10.1111/j.1582-4934.2005.tb00384.x.
75. Seifert R, Strasser A, Schneider EH, Neumann D, Dove S, Buschauer A. Molecular and cellular analysis of human histamine receptor subtypes. Trends Pharmacol Sci. 2013;34:33–58. doi:10.1016/j.tips.2012.11.001.
76. Micallef S, Stark H, Sasse A. Polymorphisms and genetic linkage of histamine receptors. Life Sci. 2013;93:487–94. doi:10.1016/j.lfs.2013.08.012.
77. Schneider EH, Neumann D, Seifert R. Modulation of behavior by the histaminergic system: Lessons from HDC-, H3R- and H4R-deficient mice. Neurosci Biobehav Rev. 2014;47:101–21. doi:10.1016/j.neubiorev.2014.07.020.
78. Schneider EH, Neumann D, Seifert R. Modulation of behavior by the histaminergic system: lessons from H(1)R-and H(2)R-deficient mice. Neurosci Biobehav Rev. 2014;42:252–66. doi:10.1016/j.neubiorev.2014.03.009.
79. Schneider E, Leite-de-moraes M, Dy M. Histamine, immune cells and autoimmunity. Adv Exp Med Biol. 2010;709:81–94. doi:10.1007/978-1-4419-8056-4_9.
80. Abrighach H, Fajardo I, Sánchez-Jiménez F, Urdiales JL. Exploring polyamine regulation by nascent histamine in a human-transfected cell model. Amino Acids. 2010;38:561–73. doi:10.1007/s00726-009-0417-6.
81. Caro-Astorga J, Fajardo I, Ruiz-Pérez MV, Sánchez-Jiménez F, Urdiales JL. Nascent histamine induces α-synuclein and caspase-3 on human cells. Biochem Biophys Res Commun. 2014;451:580–6. doi:10.1016/j.bbrc.2014.08.022.
82. Furuta K, Nakayama K, Sugimoto Y, Ichikawa A, Tanaka S. Activation of histidine decarboxylase through post-translational cleavage by caspase-9 in a mouse mastocytoma P-815. J Biol Chem. 2007;282:13438–46. doi:10.1074/jbc.m609943200.
83. Krauth M-T, Agis H, Aichberger KJ, Simonitsch-Klupp I, Müllauer L, Mayerhofer M, et al. Immunohistochemical detection of histidine decarboxylase in neoplastic mast cells in patients with systemic mastocytosis. Hum Pathol. 2006;37:439–47. doi:10.1016/j.humpath.2005.11.015.
84. Garcia-Montero AC, Jara-Acevedo M, Teodosio C, Sanchez ML, Nunez R, Prados A, et al. KIT mutation in mast cells and other bone marrow hematopoietic cell lineages in systemic mast cell disorders: a prospective study of the Spanish Network on Mastocytosis (REMA) in a series of 113 patients. Blood. 2006;108:2366–72. doi:10.1182/blood-2006-04-015545.
85. Fajardo I, Urdiales JL, Paz JC, Chavarría T, Sánchez-Jiménez F, Medina MA. Histamine prevents polyamine accumulation in mouse C57.1 mast cell cultures. Eur J Biochem. 2001;268:768–73. doi:10.1046/j.1432-1327.2001.01930.x.
86. Kanerva K, Lappalainen J, Mäkitie LT, Virolainen S, Kovanen PT, Andersson LC. Expression of antizyme inhibitor 2 in mast cells and role of polyamines as selective regulators of serotonin secretion. PLoS One. 2009;4, e6858. doi:10.1371/journal.pone.0006858.

87. Gavin IM, Glesne D, Zhao Y, Kubera C, Huberman E. Spermine acts as a negative regulator of macrophage differentiation in human myeloid leukemia cells. Cancer Res. 2004;64:7432–8. doi:10.1158/0008-5472.can-04-0051.
88. Ding XQ, Chen D, Rosengren E, Persson L, Hakanson R. Comparison between activation of ornithine decarboxylase and histidine decarboxylase in rat stomach. Am J Physiol. 1996;270:G476–86.
89. Seiler N. Catabolism of polyamines. Amino Acids. 2004;26:217–33. doi:10.1007/s00726-004-0070-z.
90. Ballas SK, Mohandas N, Clark MR, Embury SH, Smith ED, Marton LJ, et al. Reduced transglutaminase-catalyzed cross-linking of exogenous amines to membrane proteins in sickle erythrocytes. Biochim Biophys Acta. 1985;812:234–42. doi:10.1016/0005-2736(85)90543-7.
91. LaBella FS, Brandes LJ. Interaction of histamine and other bioamines with cytochromes P450: implications for cell growth modulation and chemopotentiation by drugs. Semin Cancer Biol. 2000;10:47–53. doi:10.1006/scbi.2000.0307.
92. Kallweit U, Aritake K, Bassetti CL, Blumenthal S, Hayaishi O, Linnebank M, et al. Elevated CSF histamine levels in multiple sclerosis patients. Fluids Barriers CNS. 2013;10:19. doi:10.1186/2045-8118-10-19.
93. Rinne JO, Anichtchik OV, Eriksson KS, Kaslin J, Tuomisto L, Kalimo H, et al. Increased brain histamine levels in Parkinson's disease but not in multiple system atrophy. J Neurochem. 2002;81:954–60. doi:10.1046/j.1471-4159.2002.00871.x.
94. Ballerini C, Aldinucci A, Luccarini I, Galante A, Manuelli C, Blandina P, et al. Antagonism of histamine H4 receptors exacerbates clinical and pathological signs of experimental autoimmune encephalomyelitis. Br J Pharmacol. 2013;170:67–77. doi:10.1111/bph.12263.
95. Büttner S, Broeskamp F, Sommer C, Markaki M, Habernig L, Alavian-Ghavanini A, et al. Spermidine protects against α-synuclein neurotoxicity. Cell Cycle. 2014;13:3903–8. doi:10.4161/15384101.2014.973309.
96. Benetti F, Furini CRG, de Carvalho MJ, Provensi G, Passani MB, Baldi E, et al. Histamine in the basolateral amygdala promotes inhibitory avoidance learning independently of hippocampus. Proc Natl Acad Sci. 2015;112:E2536–42. doi:10.1073/pnas.1506109112.
97. Shan L, Dauvilliers Y, Siegel JM. Interactions of the histamine and hypocretin systems in CNS disorders. Nat Rev Neurol. 2015;11:401–13. doi:10.1038/nrneurol.2015.99.
98. Ringvall M, Rönnberg E, Wernersson S, Duelli A, Henningsson F, Abrink M, et al. Serotonin and histamine storage in mast cell secretory granules is dependent on serglycin proteoglycan. J Allergy Clin Immunol. 2008;121:1020–6. doi:10.1016/j.jaci.2007.11.031.
99. Smitka K, Papezova H, Vondra K, Hill M, Hainer V, Nedvidkova J. The role of "mixed" orexigenic and anorexigenic signals and autoantibodies reacting with appetite-regulating neuropeptides and peptides of the adipose tissue-gut-brain axis: relevance to food intake and nutritional status in patients with anorexia nervosa. Int J Endocrinol. 2013;2013:483145. doi:10.1155/2013/483145.
100. Sundvik M, Panula P. Interactions of the orexin/hypocretin neurones and the histaminergic system. Acta Physiol (Oxf). 2015;213:321–33. doi:10.1111/apha.12432.
101. Ai W, Liu Y, Langlois M, Wang TC. Kruppel-like factor 4 (KLF4) represses histidine decarboxylase gene expression through an upstream Sp1 site and downstream gastrin responsive elements. J Biol Chem. 2004;279:8684–93. doi:10.1074/jbc.m308278200.
102. Sakurada T, Ro S, Onouchi T, Ohno S, Aoyama T, Chinen K, et al. Comparison of the actions of acylated and desacylated ghrelin on acid secretion in the rat stomach. J Gastroenterol. 2010;45:1111–20. doi:10.1007/s00535-010-0269-6.
103. Andrews ZB, Erion D, Beiler R, Liu Z-W, Abizaid A, Zigman J, et al. Ghrelin promotes and protects nigrostriatal dopamine function via a UCP2-dependent mitochondrial mechanism. J Neurosci. 2009;29:14057–65. doi:10.1523/jneurosci.3890-09.2009.
104. Lee M, Ryu YH, Cho WG, Kang YW, Lee SJ, Jeon TJ, et al. Relationship between dopamine deficit and the expression of depressive behavior resulted from alteration of serotonin system. Synapse. 2015;69:453–60. doi:10.1002/syn.21834.

105. Seyedabadi M, Fakhfouri G, Ramezani V, Mehr SE, Rahimian R. The role of serotonin in memory: interactions with neurotransmitters and downstream signaling. Exp Brain Res. 2014;232:723–38. doi:10.1007/s00221-013-3818-4.
106. Naoi M, Riederer P, Maruyama W. Modulation of monoamine oxidase (MAO) expression in neuropsychiatric disorders: genetic and environmental factors involved in type A MAO expression. J Neural Transm (Vienna). 2015;123(2):91–106. doi:10.1007/s00702-014-1362-4.
107. Nelson DL, Gehlert DR. Central nervous system biogenic amine targets for control of appetite and energy expenditure. Endocrine. 2006;29:49–60. doi:10.1385/endo:29:1:149.
108. Idova GV, Alperina EL, Cheido MA. Contribution of brain dopamine, serotonin and opioid receptors in the mechanisms of neuroimmunomodulation: evidence from pharmacological analysis. Int Immunopharmacol. 2012;12:618–25. doi:10.1016/j.intimp.2012.02.010.
109. Ferrada C, Ferré S, Casadó V, Cortés A, Justinova Z, Barnes C, et al. Interactions between histamine H3 and dopamine D2 receptors and the implications for striatal function. Neuropharmacology. 2008;55:190–7. doi:10.1016/j.neuropharm.2008.05.008.
110. Moreno E, Hoffmann H, Gonzalez-Sepúlveda M, Navarro G, Casadó V, Cortés A, et al. Dopamine D 1 -histamine H 3 receptor heteromers provide a selective link to MAPK signaling in GABAergic neurons of the direct striatal pathway. J Biol Chem. 2011;286:5846–54. doi:10.1074/jbc.m110.161489.
111. Inoue K, Tsutsui H, Akatsu H, Hashizume Y, Matsukawa N, Yamamoto T, et al. Metabolic profiling of Alzheimer's disease brains. Sci Rep. 2013;3:2364. doi:10.1038/srep02364.
112. Moya-Garcia AA, Medina MA, Sánchez-Jiménez F. Mammalian histidine decarboxylase: from structure to function. Bioessays. 2005;27:57–63. doi:10.1002/bies.20174.
113. Ruiz-Pérez MV, Pino-Ángeles A, Medina MA, Sánchez-Jiménez F, Moya-García AA. Structural perspective on the direct inhibition mechanism of EGCG on mammalian histidine decarboxylase and DOPA decarboxylase. J Chem Inf Model. 2012;52:113–9. doi:10.1021/ci200221z.
114. Cellini B, Montioli R, Oppici E, Astegno A, Voltattorni CB. The chaperone role of the pyridoxal 5′-phosphate and its implications for rare diseases involving B6-dependent enzymes. Clin Biochem. 2014;47:158–65. doi:10.1016/j.clinbiochem.2013.11.021.
115. Meiser J, Weindl D, Hiller K. Complexity of dopamine metabolism. Cell Commun Signal. 2013;11:34. doi:10.1186/1478-811x-11-34.
116. Timmons J, Chang ET, Wang J-Y, Rao JN. Polyamines and gut mucosal homeostasis. J Gastrointest Dig Syst. 2012;S7:001.
117. Kotlyar DS, Shum M, Hsieh J, Blonski W, Greenwald DA. Non-pulmonary allergic diseases and inflammatory bowel disease: a qualitative review. World J Gastroenterol. 2014;20:11023–32. doi:10.3748/wjg.v20.i32.11023.
118. Guihot G, Blachier F. Histidine and histamine metabolism in rat enterocytes. Mol Cell Biochem. 1997;175:143–8.
119. Bertrand PP, Bertrand RL. Serotonin release and uptake in the gastrointestinal tract. Auton Neurosci. 2010;153:47–57. doi:10.1016/j.autneu.2009.08.002.
120. Kanerva K, Mäkitie LT, Pelander A, Heiskala M, Andersson LC. Human ornithine decarboxylase paralogue (ODCp) is an antizyme inhibitor but not an arginine decarboxylase. Biochem J. 2008;409:187–92. doi:10.1042/bj20071004.
121. Piletz JE, Aricioglu F, Cheng J-T, Fairbanks CA, Gilad VH, Haenisch B, et al. Agmatine: clinical applications after 100 years in translation. Drug Discov Today. 2013;18:880–93. doi:10.1016/j.drudis.2013.05.017.
122. Arndt MA, Battaglia V, Parisi E, Lortie MJ, Isome M, Baskerville C, et al. The arginine metabolite agmatine protects mitochondrial function and confers resistance to cellular apoptosis. Am J Physiol Cell Physiol. 2009;296:C1411–9. doi:10.1152/ajpcell.00529.2008.
123. Moretti M, Matheus FC, de Oliveira PA, Neis VB, Ben J, Walz R, et al. Role of agmatine in neurodegenerative diseases and epilepsy. Front Biosci (Elite Ed). 2014;6:341–59. doi:10.2741/710.

124. Morgan DML, Bauer F, White A, editors. COST Action 917: biogenically active amines in food. Vol. VII. Luxembourg: Office for official publications of the European Communities, Luxembourg; 2005.
125. Wallace HY, Hughes A, editors. COST Action 922: health implications of dietary amines. Review of current status. Luxembourg: Official publications of the European Commission; 2004.
126. Fogel WA, Lewinski A, Jochem J. Histamine in food: is there anything to worry about? Biochem Soc Trans. 2007;35:349–52. doi:10.1042/bst0350349.
127. Ruiz-Pérez MV, Medina MÁ, Urdiales JL, Keinänen TA, Sánchez-Jiménez F. Polyamine metabolism is sensitive to glycolysis inhibition in human neuroblastoma cells. J Biol Chem. 2015;290:6106–19. doi:10.1074/jbc.m114.619197.
128. Rial NS, Meyskens FL, Gerner EW. Polyamines as mediators of APC-dependent intestinal carcinogenesis and cancer chemoprevention. Essays Biochem. 2009;46:111–24. doi:10.1042/bse0460008.
129. Lozier AM, Rich ME, Grawe AP, Peck AS, Zhao P, Chang AT, et al. Targeting ornithine decarboxylase reverses the LIN28/Let-7 axis and inhibits glycolytic metabolism in neuroblastoma. Oncotarget. 2015;6:196–206. doi:10.18632/oncotarget.2768.
130. Castellan Baldan L, Williams KA, Gallezot J-D, Pogorelov V, Rapanelli M, Crowley M, et al. Histidine decarboxylase deficiency causes tourette syndrome: parallel findings in humans and mice. Neuron. 2014;81:77–90. doi:10.1016/j.neuron.2013.10.052.
131. Yamauchi K. Regulation of gene expression of L-histidine decarboxylase and histamine N-methyl-transferase, and its relevance to the pathogenesis of bronchial asthma. Nihon Rinsho. 1996;54:377–88.
132. Saligrama N, Case LK, del Rio R, Noubade R, Teuscher C. Systemic lack of canonical histamine receptor signaling results in increased resistance to autoimmune encephalomyelitis. J Immunol. 2013;191:614–22. doi:10.4049/jimmunol.1203137.
133. Lu C, Diehl SA, Noubade R, Ledoux J, Nelson MT, Spach K, et al. Endothelial histamine H1 receptor signaling reduces blood-brain barrier permeability and susceptibility to autoimmune encephalomyelitis. Proc Natl Acad Sci U S A. 2010;107:18967–72. doi:10.1073/pnas.1008816107.
134. Rodríguez-López R, Reyes-Palomares A, Sánchez-Jiménez F, Medina M. PhenUMA: a tool for integrating the biomedical relationships among genes and diseases. BMC Bioinformatics. 2014;15:375. doi:10.1186/s12859-014-0375-1.
135. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83:610–5. doi:10.1016/j.ajhg.2008.09.017.
136. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85:457–64. doi:10.1016/j.ajhg.2009.09.003.
137. Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J Artif Intel Res. 1999;11:95–130. doi:10.1613/jair.514.
138. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015:bav028. doi:10.1093/database/bav028.
139. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods. 2014;74:83–9. doi:10.1016/j.ymeth.2014.11.020.
140. Bravo À, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. A knowledge-driven approach to extract disease-related biomarkers from the literature. Biomed Res Int. 2014;2014:253128. doi:10.1155/2014/253128.
141. Vanhanen J, Nuutinen S, Lintunen M, Mäki T, Rämö J, Karlstedt K, et al. Histamine is required for H_3 receptor-mediated alcohol reward inhibition, but not for alcohol consumption or stimulation. Br J Pharmacol. 2013;170:177–87. doi:10.1111/bph.12170.

142. Xu T, Du L, Zhou Y. Evaluation of GO-based functional similarity measures using S cerevisiae protein interaction and expression profile data. BMC Bioinformatics. 2008;9:472. doi:10.1186/1471-2105-9-472.
143. He G-H, Lin J-J, Cai W-K, Xu W-M, Yu Z-P, Yin S-J, et al. Associations of polymorphisms in histidine decarboxylase, histamine N-methyltransferase and histamine receptor H3 genes with breast cancer. PLoS One. 2014;9, e97728. doi:10.1371/journal.pone.0097728.
144. Borcherding DC, Tong W, Hugo ER, Barnard DF, Fox S, LaSance K, et al. Expression and therapeutic targeting of dopamine receptor-1 (D1R) in breast cancer. Oncogene. 2015. doi:10.1038/onc.2015.369.
145. Medina VA, Rivera ES. Histamine receptors and cancer pharmacology. Br J Pharmacol. 2010;161:755–67. doi:10.1111/j.1476-5381.2010.00961.x.
146. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015;348:660–5. doi:10.1126/science.aaa0355.
147. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509:575–81. doi:10.1038/nature13302.
148. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509:582–7. doi:10.1038/nature13319.
149. Gervasini G, Agúndez JAG, García-Menaya J, Martínez C, Cordobés C, Ayuso P, et al. Variability of the L-Histidine decarboxylase gene in allergic rhinitis. Allergy. 2010;65:1576–84. doi:10.1111/j.1398-9995.2010.02425.x.
150. North ML, Grasemann H, Khanna N, Inman MD, Gauvreau GM, Scott JA. Increased ornithine-derived polyamines cause airway hyperresponsiveness in a mouse model of asthma. Am J Respir Cell Mol Biol. 2013;48:694–702. doi:10.1165/rcmb.2012-0323oc.
151. Szczepankiewicz A, Bręborowicz A, Sobkowiak P, Popiel A. Polymorphisms of two histamine-metabolizing enzymes genes and childhood allergic asthma: a case control study. Clin Mol Allergy. 2010;8:14. doi:10.1186/1476-7961-8-14.
152. Welter D, Gray WA, Kille P, Ontology TG. Investigation of semantic similarity as a tool for comparative genomics. In: Wagner R, Revell N, Pernul G, editors. Database and expert systems applications. Berlin: Springer; 2007. p. 772–9.
153. Weltman JK. Histamine as a regulator of allergic and asthmatic inflammation. Allergy Asthma Proc. 2003;24:227–9.
154. Zhang R, Kubo M, Murakami I, Setiawan H, Takemoto K, Inoue K, et al. l-Arginine administration attenuates airway inflammation by altering l-arginine metabolism in an NC/Nga mouse model of asthma. J Clin Biochem Nutr. 2015;56:201–7. doi:10.3164/jcbn.14-140.
155. Montañez R, Sánchez-Jiménez F, Aldana-Montes JF, Medina MA. Polyamines: metabolism to systems biology and beyond. Amino Acids. 2007;33:283–9. doi:10.1007/s00726-007-0521-4.
156. Dandrifosse G, Peulen O, El Khefif N, Deloyer P, Dandrifosse AC, Grandfils C. Are milk polyamines preventive agents against food allergy? Proc Nutr Soc. 2000;59:81–6. doi:10.1017/S0029665100000100.
157. López-Garcia C, Ramos-Molina B, Lambertos A, López-Contreras AJ, Cremades A, Peñafiel R. Antizyme inhibitor 2 hypomorphic mice. New patterns of expression in pancreas and adrenal glands suggest a role in secretory processes. PLoS One. 2013;8, e69188. doi:10.1371/journal.pone.0069188.
158. Ramos-Molina B, López-Contreras AJ, Cremades A, Peñafiel R. Differential expression of ornithine decarboxylase antizyme inhibitors and antizymes in rodent tissues and human cell lines. Amino Acids. 2012;42:539–47. doi:10.1007/s00726-011-1031-y.
159. Mäkitie LT, Kanerva K, Polvikoski T, Paetau A, Andersson LC. Brain neurons express ornithine decarboxylase-activating antizyme inhibitor 2 with accumulation in Alzheimer's disease. Brain Pathol. 2010;20:571–80. doi:10.1111/j.1750-3639.2009.00334.x.
160. Li DC, Nichols CG, Sala-Rabanal M. Role of a hydrophobic pocket in polyamine interactions with the polyspecific organic cation transporter OCT3. J Biol Chem. 2015;290:27633–43. doi:10.1074/jbc.M115.668913.

161. Tong JHS, Cummins TDR, Johnson BP, McKinley L-A, Pickering HE, Fanning P, et al. An association between a dopamine transporter gene (SLC6A3) haplotype and ADHD symptom measures in nonclinical adults. Am J Med Genet B Neuropsychiatr Genet. 2015;168B:89–96. doi:10.1002/ajmg.b.32283.
162. Sutcliffe JS, Delahanty RJ, Prasad HC, McCauley JL, Han Q, Jiang L, et al. Allelic heterogeneity at the serotonin transporter locus (SLC6A4) confers susceptibility to autism and rigid-compulsive behaviors. Am J Hum Genet. 2005;77:265–79. doi:10.1086/432648.
163. O'Brien JT, Colloby S, Fenwick J, Williams ED, Firbank M, Burn D, et al. Dopamine transporter loss visualized with FP-CIT SPECT in the differential diagnosis of dementia with Lewy bodies. Arch Neurol. 2004;61:919–25. doi:10.1001/archneur.61.6.919.
164. Gunther J, Tian Y, Stamova B, Lit L, Corbett B, Ander B, et al. Catecholamine-related gene expression in blood correlates with tic severity in tourette syndrome. Psychiatry Res. 2012;200:593–601. doi:10.1016/j.psychres.2012.04.034.
165. Horschitz S, Hummerich R, Lau T, Rietschel M, Schloss P. A dopamine transporter mutation associated with bipolar affective disorder causes inhibition of transporter cell surface expression. Mol Psychiatry. 2005;10:1104–9. doi:10.1038/sj.mp.4001730.
166. Strug LJ, Suresh R, Fyer AJ, Talati A, Adams PB, Li W, et al. Panic disorder is associated with the serotonin transporter gene (SLC6A4) but not the promoter region (5-HTTLPR). Mol Psychiatry. 2010;15:166–76. doi:10.1038/mp.2008.79.
167. Kelada SNP, Checkoway H, Kardia SLR, Carlson CS, Costa-Mallen P, Eaton DL, et al. 5′ and 3′ region variability in the dopamine transporter gene (SLC6A3), pesticide exposure and Parkinson's disease risk: a hypothesis-generating study. Hum Mol Genet. 2006;15:3055–62. doi:10.1093/hmg/ddl247.
168. Furmark T, Tillfors M, Garpenstrand H, Marteinsdottir I, Långström B, Oreland L, et al. Serotonin transporter polymorphism related to amygdala excitability and symptom severity in patients with social phobia. Neurosci Lett. 2004;362:189–92. doi:10.1016/j.neulet.2004.02.070.
169. Sáiz PA, García-Portilla MP, Arango C, Morales B, Arias B, Corcoran P, et al. Genetic polymorphisms in the dopamine-2 receptor (DRD2), dopamine-3 receptor (DRD3), and dopamine transporter (SLC6A3) genes in schizophrenia: data from an association study. Prog Neuropsychopharmacol Biol Psychiatry. 2010;34:26–31. doi:10.1016/j.pnpbp.2009.09.008.
170. Wendland JR, DeGuzman TB, McMahon F, Rudnick G, Detera-Wadleigh SD, Murphy DL. SERT Ileu425Val in autism, Asperger syndrome and obsessive-compulsive disorder. Psychiatr Genet. 2008;18:31–9. doi:10.1097/ypg.0b013e3282f08a06.
171. Liu H, Liu M, Wang Y, Wang X-M, Qiu Y, Long J-F, et al. Association of 5-HTT gene polymorphisms with migraine: a systematic review and meta-analysis. J Neurol Sci. 2011;305:57–66. doi:10.1016/j.jns.2011.03.016.
172. Igarashi K, Kashiwagi K. Use of polyamine metabolites as markers for stroke and renal failure. Methods Mol Biol. 2011;720:395–408.
173. Ortega SB, Noorbhai I, Poinsatte K, Kong X, Anderson A, Monson NL, et al. Stroke induces a rapid adaptive autoimmune response to novel neuronal antigens. Discov Med. 2015;19:381–92.
174. Mecheri S. Contribution of allergic inflammatory response to the pathogenesis of malaria disease. Biochim Biophys Acta. 1822;2012:49–56. doi:10.1016/j.bbadis.2011.02.005.
175. Correa-Fiz F, Reyes-Palomares A, Medina MA, Sanchez-Jiménez F. Roles of biogenic amines in emergent and rare diseases. In: Dandrifosse G, editor. Biological aspects of biogenic amines and polyamine conjugates. Kerala: Research Signpost; 2009. p. 339–419.
176. Reyes-Palomares A, Montañez R, Real-Chicharro A, Chniber O, Kerzazi A, Navas-Delgado I, et al. Systems biology metabolic modeling assistant: an ontology-based tool for the integration of metabolic data in kinetic modeling. Bioinformatics. 2009;25:834–5. doi:10.1093/bioinformatics/btp061.
177. Schlüter A, Real-Chicharro A, Gabaldón T, Sánchez-Jiménez F, Pujol A. PeroxisomeDB 2.0: an integrative view of the global peroxisomal metabolome. Nucleic Acids Res. 2010;38:D800–5. doi:10.1093/nar/gkp935.

178. Real-Chicharro A, Ruiz-Mostazo I, Navas-Delgado I, Kerzazi A, Chniber O, Sánchez-Jiménez F, et al. Protopia: a protein-protein interaction tool. BMC Bioinformatics. 2009;10 Suppl 1:S17. doi:10.1186/1471-2105-10-s12-s17.
179. Moya-García AA, Ranea JAG. Insights into polypharmacology from drug-domain associations. Bioinformatics. 2013;29:1934–7. doi:10.1093/bioinformatics/btt321.
180. Reyes-Palomares A, Rodríguez-López R, Ranea JAG, Sánchez Jiménez F, Medina MA. Global analysis of the human pathophenotypic similarity gene network merges disease module components. PLoS One. 2013;8, e56653. doi:10.1371/journal.pone.0056653.
181. Navas-Delgado I, Real-Chicharro A, Medina MÁ, Sánchez-Jiménez F, Aldana-Montes JF. Social pathway annotation: extensions of the systems biology metabolic modelling assistant. Brief Bioinform. 2011;12:576–87. doi:10.1093/bib/bbq061.
182. Castro-Oropeza R, Pino-Ángeles A, Khomutov MA, Urdiales JL, Moya-García AA, Vepsäläinen J, et al. Aminooxy analog of histamine is an efficient inhibitor of mammalian L-histidine decarboxylase: combined in silico and experimental evidence. Amino Acids. 2014;46:621–31. doi:10.1007/s00726-013-1589-7.
183. Sánchez-Jiménez F, Reyes-Palomares A, Moya-García AA, Ranea JAG, Medina MÁ. Biocomputational resources useful for drug discovery against compartmentalized targets. Curr Pharm Des. 2014;20:293–300. doi:10.2174/13816128113199990030.

Reyes-Palomares et al. BMC Genomics (2016) 17:232
DOI 10.1186/s12864-016-2569-6

BMC Genomics

RESEARCH ARTICLE Open Access

Systematic identification of phenotypically enriched loci using a patient network of genomic disorders

Armando Reyes-Palomares[1,2,4*], Aníbal Bueno[1], Rocío Rodríguez-López[1,2], Miguel Ángel Medina[1,2], Francisca Sánchez-Jiménez[1,2], Manuel Corpas[3] and Juan A. G. Ranea[1,2*]

Abstract

Background: Network medicine is a promising new discipline that combines systems biology approaches and network science to understand the complexity of pathological phenotypes. Given the growing availability of personalized genomic and phenotypic profiles, network models offer a robust integrative framework for the analysis of "omics" data, allowing the characterization of the molecular aetiology of pathological processes underpinning genetic diseases.

Methods: Here we make use of patient genomic data to exploit different network-based analyses to study genetic and phenotypic relationships between individuals. For this method, we analyzed a dataset of structural variants and phenotypes for 6,564 patients from the DECIPHER database, which encompasses one of the most comprehensive collections of pathogenic Copy Number Variations (CNVs) and their associated ontology-controlled phenotypes. We developed a computational strategy that identifies clusters of patients in a synthetic patient network according to their genetic overlap and phenotype enrichments.

Results: Many of these clusters of patients represent new genotype-phenotype associations, suggesting the identification of newly discovered phenotypically enriched *loci* (indicative of potential novel syndromes) that are currently absent from reference genomic disorder databases such as ClinVar, OMIM or DECIPHER itself.

Conclusions: We provide a high-resolution map of pathogenic phenotypes associated with their respective significant genomic regions and a new powerful tool for diagnosis of currently uncharacterized mutations leading to deleterious phenotypes and syndromes.

Background

Genomic Structural Variations are one of the main sources of human genome variation. Copy Number Variations (CNVs) naturally occur in the genome of healthy individuals [1, 2], some of them leading to disease [3]. CNVs consist of thousands to millions of bp deletions, duplications, insertions or inversions, recurrent in the population either by inheritance or spontaneous occurrence (*de novo*) [4]. Although the discovery of CNVs was relatively recent, a plethora of genetic association studies have been carried out to understand their evolutionary [5], functional [6] and phenotypic effects [4]. It has been estimated that two genomes can differ approximately about 0.4 % due to CNVs [7] and that these variations have a considerable impact on human health. Several known chromosome imbalances causing complex genomic disorders have been characterized by different medical conditions such as developmental [8, 9], neuropsychiatric [10–12], cancer [13], autoimmune diseases [14] and idiopathic learning disability [15]. However, recent genome wide association studies suggest that the lack of data for individual's medical records is an important limitation to fully understand the genetic basis for many genomic disorders [16, 17]. Initiatives such as the Personal Genomes Project (PGP) [18], Genomics England (http://www.genomicsengland.co.uk/) and the Precision Medicine program [19] aim to provide descriptive records

* Correspondence: armando.reyes@embl.de; ranea@uma.es
[1]Universidad de Málaga, Andalucía Tech, Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, and IBIMA (Biomedical Research Institute of Málaga), E-29071 Málaga, Spain
Full list of author information is available at the end of the article

and associated genomic data accessible for research. These datasets, however, are still unavailable or pose different challenges when looking into genetic association studies: e.g., lack of sizable data (e.g., PGP) or too restrictive access (e.g., Genomics England). These shortcomings may encourage genetic association studies to oversimplify complex phenotypic profiles of individuals, focusing on the most representative clinical features [20]. This makes it more difficult to characterize pathophysiological associations between clinical features observed in studied individuals [20]. New systematic and standardized methods are thus required that make use of limited accessible clinical genotype and phenotype profiling datasets to enhance our understanding of the genetic impact of CNVs on human health [21]. The present work uses individual clinical and genetic information stored in the DECIPHER Database [21], a database of sub-microscopic chromosome abnormalities (deletions and duplications) observed in clinic with a potential pathogenic association. Data currently stored by DECIPHER add up to more than 45,000 patients (march 2015), of which more than 10,000 have given consent to share their medical data [22] under an ethically regulated data access protocol. We focus our study on a subset of these data of 9,186 unbalanced CNVs from 6,564 patients that included a heterogeneous set of pathophenotypes, including developmental delay, intellectual disability and congenital malformations. Network analyses has been used in previous studies to characterize affected pathways by CVNs in cancer [23]. Here we applied network medicine approaches, phenotypic enrichment analyses and genetic association studies to build patient networks to explore the similarities between reported genetic microvariations (CNVs) and pathological phenotypes. We represented patients as nodes connected with edges to other patients whose CNVs overlap. Our resulting networks allowed the systematic identification of genetically related clusters of patients by finding cliques [24, 25]. A phenotypic enrichment analysis of patient clusters was performed to identify overrepresented phenotypes for each cluster. We named *Phenotypically Enriched Locus* (PEL) an affected genomic location showing significant associations with phenotypes. Significant genotype-phenotype associations were retrieved through the comparison of patients (cases) and healthy (controls) datasets, using a case–control association analysis. The combined use of these methods allowed us to build a high-resolution genotype-phenotype map that identifies a) already known, b) potentially novel genomic disorders and c) the additive phenotypic effects found in some proximal structural variations.

Methods

Case and control datasets

Cases

Rare CNVs (frequency of <1 %) from patients with low prevalent genomic disorders were downloaded from DECIPHER database (08/05/2014; http://decipher.sanger.ac.uk/) through its Data Access Agreement. This dataset contains genotype-phenotype annotations of consented DECIPHER patients, including chromosome locations, type of structural variant (gain or loss), mode of inheritance (de novo, inherited from unaffected parent, inherited from affected parent and unknown) and clinical phenotypes observed by expert physicians. When available, patients in DECIPHER are assigned phenotypes from the Human Phenotype Ontology (HPO), a standard controlled vocabulary of pathological terms [26]. Patients not annotated with HPO phenotypes were removed from our study. To reduce heterogeneity among collected patient data from DECIPHER, we only selected CNVs originated from array CGH technology, which corresponds to the majority of the database's genotypic data. A final dataset of 6,564 patients with 9,186 CNVs presenting 1,860 non-redundant HPO terms was chosen for this study (Additional file 1: Table S1). Access to DECIPHER genomic coordinates of chromosomal microdeletions, microduplications and associated phenotypes were obtained through a Data Access Agreement. All data shared by the DECIPHER database have signed a consent form obtained by the submitting clinician. Those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it by the Recipient or its Registered Users.

Controls

CNVs from healthy individuals were retrieved from the Database of Genomic Variants (DGV, http://dgv.tcag.ca/) [27], which provides a curated collection of human structural variations in control data from multiples studies. DGV offers information about CNVs of individual samples such as chromosome locations, type of structural variation (gain or loss) and reference (PubMed ID) of the study and the platform used in the analysis. The control structural variants dataset ("*GRCh37_hg19_variants_2014-10-16.txt*") was downloaded from DGV. This dataset combines CNV data from diverse studies. Using DGV as the control dataset has the caveat that it does not distinguish unrelated from related samples (i.e., the same patient CNV retrieved from different studies). Although in practice this overrepresentation of the same patient may seldom happen, it may still overestimate the number of so-called independent CNVs, affecting our final results. This overestimation of the frequency of CNVs in controls drove us to make a stricter assessment of the statistical

significance of our predicted pathogenic CNVs. The types of effects this inflation of non-pathogenic CNVs may cause include an increase of the number of false negatives (i.e. true pathogenic CNVs that overlap with an over-estimated number of control CNVs) and a reduction of the number of false positives (i.e. false pathogenic CNVs overlapping with an over-estimated number of control CNVs). Therefore, we have considered CNVs from DGV only as a quantitative control for preventing misclassifications of CNVs as pathogenic.

Building the genotype-based patient network

We designed a workflow to systematically identify all the existing genotype-phenotype associations in the case dataset (Fig. 1). First, the overlap between patient CNVs belonging to the same class (either gains or loses) was computed using the GRCh37/hg19 reference genome. For the purposes of this study, we assumed that two patient CNVs overlap if at least they share one common base pair. The resulting genetic relationships were used to build the network, where nodes are patients and edges represent the overlap between patient CNVs (Fig. 1).

Clustering of patients using cliques

Finding all the k-cliques associated with each patient provides all complete graphs from the resulting genotype-based patient network. These cliques correspond to sets of variable numbers of nodes where all are connected to all by edges [24]. To find all the cliques associated with each node from the patient network, we used the algorithm of the function "cliques_containing_node" available in the Python package named NetworkX. The minimum size of cliques was limited to three patients (k nodes ≥ 3) but no limitation was applied to maximum size clique detection. We then merged into one clique all those containing identical sets of patients with the aim of getting a unique list of cliques resulting from the patient network. This list of unique cliques is of high interest for our approach because it allows the systematic identification of the whole set of patients sharing similar genotypes by mining directly the clusters of the network. Taking into account that CNV lengths can be very variable across the case population, a large patient CNV can overlap with other patient CNVs at different genomic regions. These complex interactions in the patient network imply that some cliques might not necessarily represent a cluster of patients where all their CNV overlap. Thus, we selected only those cliques that were fully represented by patients with mutations on the same genomic region. The resulting cliques were used as the list of clusters of patients to be used for downstream analyses, i.e., phenotype enrichment analysis.

Phenotype enrichment analysis

The Human Phenotype Ontology (HPO) was used as a relational graph to identify common phenotypes among all the clique patients. The hierarchical organization of HPO terms (phenotypes) by parent–child relationships allows the detection of phenotype enrichments when their annotations co-occur at the same ontological level. We used this relational graph to detect the common phenotypes in a given cluster –or clique– of patients. To systematically assess the phenotype significance in each

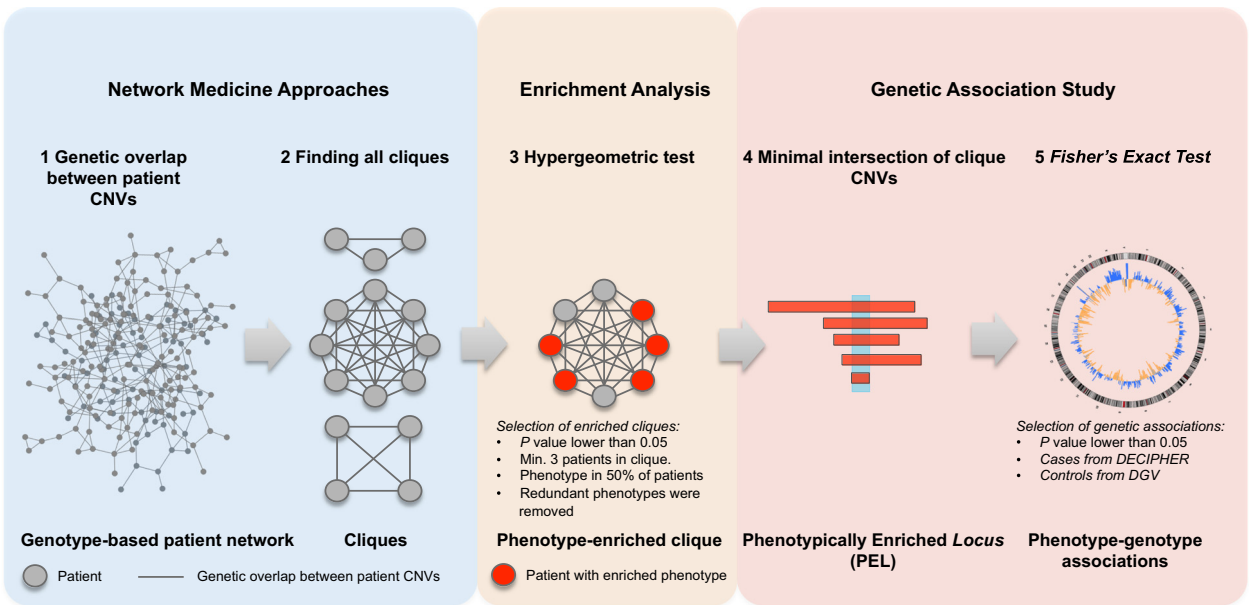


Fig. 1 Workflow used to carry out the present study. More details can be found out in Methods

clique, we used a hypergeometric test and adjusted the *P-values* using Bonferroni. This test compares the frequency of every HPO term in each clique (number of observed cases in the sample) against their frequency in the whole dataset of annotated patients (observed cases in the population). To carry out this test, we used the number of individuals per clique as the sample size, the number of patients in the samples presenting a phenotype as the observed cases, and the total number of patients in DECIPHER database presenting the phenotype as the population size. We selected clique-phenotype enrichment associations by applying three different thresholds: 1) $P < 0.05$ from hypergeometric test, 2) counting at least three patients annotated with the enriched phenotype, and 3) if at least 50 % of the patients in the clique are annotated with the enriched phenotype. Once this selection process ended, we found that many of these cliques were enriched with HPO terms that are closely related in the ontology (i.e. parent–child relationship), producing some redundancy that does not add information. In those cases, redundancies were removed by selecting the most significant (lowest *P-values*) HPO terms as the representative ones.

Characterizing phenotypically enriched *loci* (PELs)

We defined a phenotypically Enriched Locus (PEL) as the minimal common intersection among all the CNVs of patients in every clique that is significantly enriched with phenotypes (Fig. 1). We studied PELs' incidence in patients (cases) by comparing them to a healthy population (control). Their statistical significance was assessed using a Fisher's exact test from a contingency table. This table consisted of a) the number of patients in a PEL associated with an enriched phenotype versus the total number of observed cases with that particular phenotype, and b) the number of healthy individuals –or samples from DGV dataset– with structural variants overlapping to this PEL versus the rest of observed controls (i.e., healthy population). We checked overlaps between PELs and individual control CNVs that overlapped at least 1 bp. After applying the Fisher's exact test, the *P-values* were adjusted using Benjamini & Hochberg and only those PEL sites with $P < 0.05$ were considered. This procedure allowed us to calculate the statistical significance of associations between enriched phenotypes (HPO-term) and a PEL compared to frequency of CNVs from the healthy population on the same locus. Finally, the penetrance of enriched phenotypes for each *locus* was calculated as the proportion of individuals showing the enriched phenotype –cases- over the healthy population –control-, by using a similar approach to the one recently published by Cooper et al. [8, 28].

Randomization analysis on case and control datasets

Five randomization analyses were designed to test different null hypotheses: (i) Arbitrarily selected CNVs from the control dataset without replacement and it was used to test if the frequency of detected PELs is lower than from a case population (DECIPHER) when using CNVs from a healthy population (DGV). This randomization analysis was named "random patient CNVs from DGV". (ii) The second type of randomized case dataset was generated from arbitrary genomic regions while keeping the CNV length distribution and chromosome frequencies from the case dataset and it was named "random patient CNV location". This randomized dataset was used to test if the frequency of detected PELs is lower when individual case CNVs are randomly distributed across the genome compared to real patient CNVs from DECIPHER. (iii) A similar approach as mentioned above was used to generate the third type of randomized dataset but using the control dataset (DGV) instead of the case dataset. This randomization analysis, named "random control CNV location", was used to test if the frequency of PELs is lower when individual control CNVs are randomly distributed across the genome compared to real CNVs from DGV. (iv) The fourth type of randomization analysis was carried out by randomly shuffling the patient-CNVs relations (named as "rewiring patient-CNV") to test if the frequency of PELs is lower when using arbitrary phenotype-genotype relationships. (v) Finally, randomized case datasets were built using arbitrary phenotype descriptions of patients while keeping the phenotype frequency, to ensure that the representativeness of phenotypes from the real data is preserved. This randomization analysis was used to test that the frequency of detected PELs is lower using arbitrary phenotype descriptions for patients. We carried out one thousand randomization experiments for each randomized dataset and counted the number of PELs as well as the significances derived from the phenotypic enrichment analysis (*P-values* < 0.05, hypergeometric test) and genetic association study (*P-values* < 0.05, Fisher's exact test).

Results and discussion

Phenotypic and genotypic features of patient population

The subset of 6,564 patients from the DECIPHER database used in this study includes the CNVs and clinical features (i.e., HPO phenotypic terms) observed by expert physicians in these patients. Table 1 summarizes the data analyzed for case (patients) and control (healthy population) datasets. The distribution of different phenotypes (HPO terms) associated with patients (Fig. 2a) showed that almost half of patients were annotated with just one HPO term, while the remaining cases showed more complex phenotypic profiles with two or more associated terms. The distributions of *de novo* and inherited

Table 1 Population dataset descriptions

	All patients	Cases	Control
Samples	10,324	6,564	5,072[b]
Identified CNVs	14,226[a]	9,186	495,916
Type of CNVs:			
Loss	7,554	5,101	343,489
Gain	6,672	4,085	152,427
Average CNV length (Kb)	3,336	3,014	31
Type of inheritance:			
De novo constitutive	14,501	2,454	
Inherited from normal parent	9,345	1,945	
Inherited from parent with similar phenotype to child	1,345	240	
Unknown	21,946	3,638	

The table shows genotyped patients in DECIPHER database (*All*), the genotyped and phenotyped patients from DECIPHER used in this work (*Cases*) and the healthy individuals from the DGV repository (Control). The first column indicates the distribution of data based on number of individuals, number and type of CNVs and their type of inheritance. [a] This is a pre-selection of CNVs from DECIPHER that are potentially pathogenic. [b] This number does not correspond to individual samples

patients were explored based on the complexity of their phenotypic profiles (Fig. 2b). It is observed that the *de novo* CNVs show a significant ($P < 2.2E\text{-}16$, Mann–Whitney *U* test) bias toward more complex –or diverse– phenotype profiles than the inherited group (Fig. 2b). The distribution of CNV lengths in patients is biased toward higher lengths as compared with those of control CNVs, something that should be expected if clinicians remove the non-pathological CNVs (Fig. 2c). Within the observed patient dataset, those including *de novo* CNVs showed the highest average length compared to the inherited set (Fig. 2d). These results suggest a positive relationship between CNV length and the complexity of annotated patient phenotypes. This is not a surprising observation, since larger CNVs affect more genes in the genome, producing an additive effect to observed clinical features.

Analysis of phenotypically enriched *loci* (PELs)

We built a patient network, consisting of 6,324 nodes (patients) connected by 89,526 interactions based on the genetic overlapping between patient CNVs, and we calculated some topological parameters (Table 2). The resulting network showed low density, which means that the portion of potential interactions is low compared to

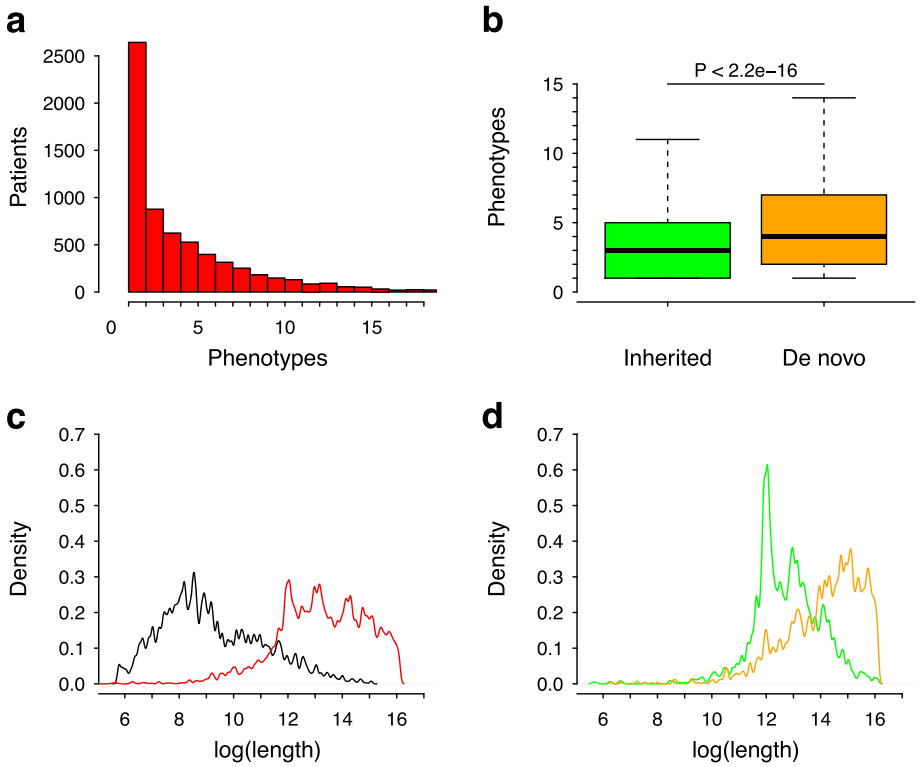


Fig. 2 CNV length vs. phenotype relationships. **a** Histogram for the number of phenotypes observed in DECIPHER patients. **b** Boxplots of the number of phenotypes observed in patients showing inherited or *de novo* CNVs (because this CNV was absent in parents). For this plot, we only took into account those patients for whom only one CNV was detected. **c** Length CNV distributions for control (black line) and case (red line) populations. **d** Length CNV distributions in cases for *de novo* CNVs (orange line) and inherited CNVs by parents that do not manifest any pathogenic phenotype (green line)

Table 2 Topological parameters and properties of patient network

Network parameter	Value
Nodes	6,304
Number of interactions	89,526
Clustering coefficients	0.801
Connected components	5
Network diameter	10
Shortest paths	39,482,458
Average shortest path length	3.706
Average degree	28.403
Network density	0.005

the actual interactions in the network, and a high average clustering coefficient, which measures how nodes (patients) tend to cluster together. In addition, we also observed other properties such as a heterogeneous degree distribution, a small average shortest path length, and a high average clustering coefficient of network nodes, available in Additional file 2: Figure S1. These network properties suggest that the patient network appeared to show general features of most large real-world networks in contrast to random networks.

From the patient network, we proceed to study PELs; i.e., significantly enriched genomic *loci* with phenotypes in patient clusters. We designed network-based and enrichment analyses to find genetically and phenotypically related clusters of patients (cliques; see Methods and Fig. 1). In total, 1,042 *locus*-phenotype associations between 487 PELs and 195 enriched phenotypes (HPO terms) were generated. We performed a genome-wide study of CNVs, using as control a dataset of healthy population, to evaluate the significance of genotype-phenotype associations in PELs. A Fisher's exact test (see Methods) related to previous works was applied [8]. However, our experiment defined genetic associations to exploit patient network relationships, evaluating each *locus* independently instead of using sliding windows as previous works. In addition, redundant and uninformative phenotypes were also removed according to their parent–child relationships (see Methods). Using this systematic approach, we reported 387 specific *locus*-phenotype associations between 336 PELs and 115 different phenotypes (HPO terms; Additional file 3: Table S2). Almost 70 % (336 of 487) PELs were

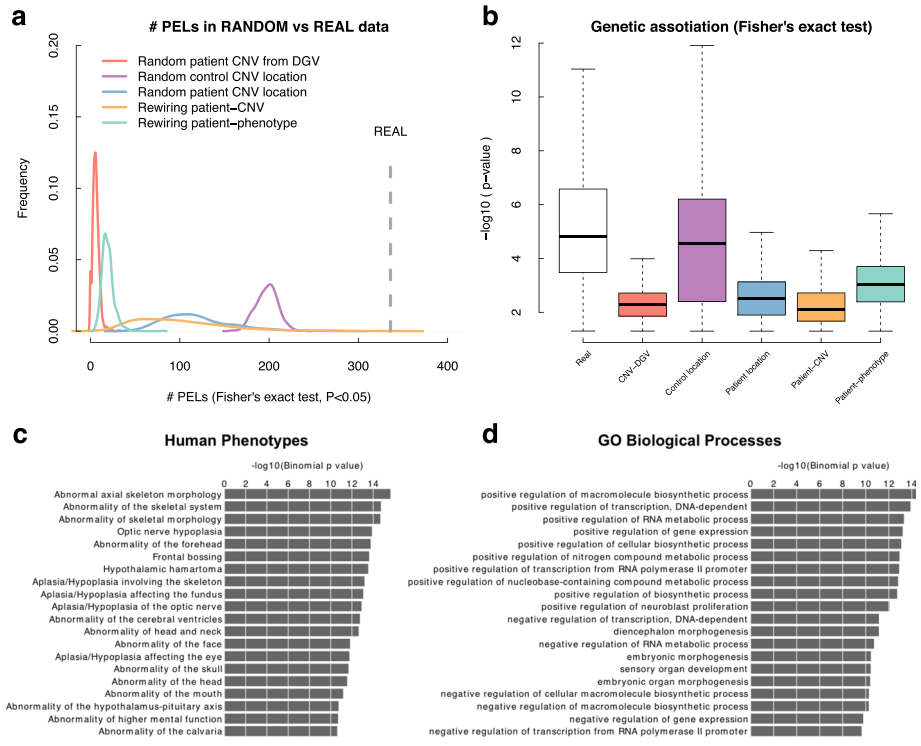


Fig. 3 Functional analysis of pathogenic phenotypically enriched *loci*. **a** Distribution of the resulting number of PELs after 1000 randomization experiments, each type of randomization experiment is represented by a colored line tat is described in the legend. The dashed grey line corresponds the number of PELs obtained using the real data. **b** Boxplot of the distribution of *P-values* ($P < 0.05$, Fisher's exact test) that results from case control analysis. **c** Enrichment of human phenotype in genomic regions related to pathogenic PELS, we used a binomial test from GREAT. **d** Enrichment of biological process in genomic regions related to pathogenic PELS, we used a binomial test from GREAT

significantly more frequently mutated in patients compared to healthy individuals ($P < 0.05$, Fisher's exact test). We denoted these as pathogenic PELs. Given the nature of collecting pathogenic CNVs in DECIPHER, it is not surprising that we obtained this high percentage (70 %) of potentially pathogenic PELs.

To assess whether these *loci* are potentially pathogenic and that our results are not due to chance, we did several randomization analyses with the aim of comparing real and random results. Five different types of randomization analyses were designed using randomized case and control datasets to test if the frequency of detected PELs is lower than real cases (Fig. 3a): (i) we generated random datasets of mutations in patients from random sets of CNVs that were selected from the control dataset (DGV), we used random locations for (ii) patient CNVs and (iii) control CNVs by selecting random genomic regions while keeping CNV length distributions and chromosome frequencies, (iv) the rewiring of the patient-CNV relations, and, finally, (v) the rewiring of phenotype descriptions of patients conserving the phenotype frequencies (see Methods).

We found that the number of PELs identified by using the real data (336) was substantially higher compared to that resulted from the different randomization experiments (Fig. 3a). In addition, the significances (*P-values* < 0.05, Fisher's exact test) derived from the genetic association study are also higher in real than in randomized datasets (Fig. 3b). The small differences with respect the control dataset with random CNV locations suggest that there is a portion of CNVs in the control population (DGV) that are randomly distributed across the genome, something that might be expected in natural genetics populations (Fig. 3b). Overall these results reveal the existence of a fraction of PELs in DECIPHER that are consistently pathogenic, where both the number of resulting PELs and the median significance of Fisher's exact test are higher when using real data compared to random datasets (Fig. 3a and b, respectively).

We then studied which annotations from diverse biomedical ontologies are associated with these *loci* using GREAT [29]. It was found that these regions are significantly enriched for human phenotypes (Fig. 3c), reinforcing the probable clinical implication of mutations affecting these genomic regions. In addition, we also found that these PELs are enriched for cis-regulatory domains involved in biosynthetic processes, regulatory elements and embryonic morphogenesis (Fig. 3d). The experimental and functional characterization of these genomic regions might improve our current understanding of the molecular basis of these genomic disorders.

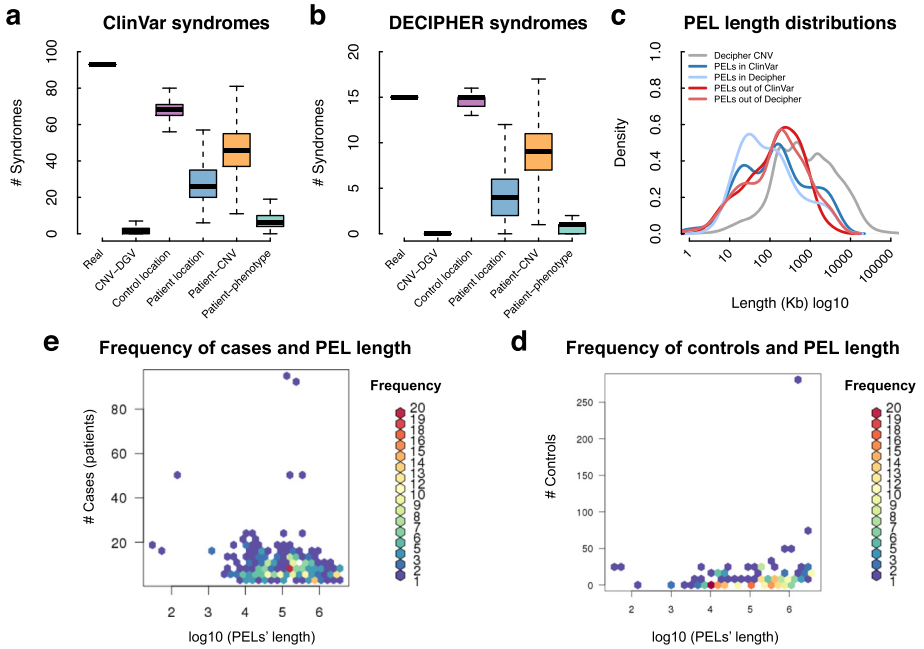


Fig. 4 Pathogenicity of phenotypically enriched *loci* (PELs). **a** Boxplot of the distribution of the number of ClinVar syndromes overlapping PELs. Every boxplot represents the resulting number of PELs in real and permuted data; there is only one value for real data. **b** Boxplot of the distribution of the number of DECIPHER syndromes overlapping PELs. **c** Distribution of the PEL lengths overlapping with ClinVar and DECIPHER syndromes, the grey line represent the distribution of the length of the raw patient CNVs from DECIPHER. **d** Relationship between the PEL length and their number of cases (patients), the colors represent the frequency of the observations. **e** Relationship between the PEL length and their number of overlapped controls (DGV), the colors represent the frequency of the observations

Pathogenicity of PELs

With the aim to validate the resulting phenotype-genotype associations, we searched how many pathogenic PELs match with known genomic disorders in ClinVar [30]. For this we selected 2,243 pathogenic or likely pathogenic CNVs associated with any OMIM phenotype and other 75 genomic regions described as DECIPHER syndromes. We then studied if our method retrieves genomic syndromes from ClinVar or DECIPHER. First, we looked for those PELs overlapping known syndrome from both databases (Additional file 4: Table S3 and Additional file 5: Table S4 for ClinVar and DECIPHER respectively) and having the same type of mutation as the described for syndromes (i.e. deletions or duplications). The number of syndromes was determined and real results were compared versus random results (Fig. 4a and b, for ClinVar and DECIPHER respectively). From the real datasets, we counted a total of 93 and 15 syndromes overlapping PELs from ClinVar and DECIPHER respectively. These numbers are higher than the ones obtained from the randomization experiments (Fig. 4a and B), with the exception of those using control CNVs with random locations across the genome. The distributions of the randomizations were similar in ClinVar and DECIPHER but with considerable differences in the number of syndromes (Fig. 4a and b). Although a higher number of known syndromes could be expected, it should be taken into account that DECIPHER includes several cohorts of patients with rare genomic disorders that have not been well characterized. This means that some cohorts of patients that have been already diagnosed for well-characterized syndromes have probably not been sent to the DECIPHER database. To study how the length of PELs could be affecting our approach, we compared their length distributions across the different subset of PELs (Fig. 4c). The average length of PELs overlapping known syndromes is slightly shorter than those classified as potential novel syndromes, and the length of raw CNV from DECIPHER are considerably longer (Fig. 4c). Subsequently, we compared the length of PELs and the number of patient CNVs and control CNVs overlapping these PELs (Fig. 4d and e, for

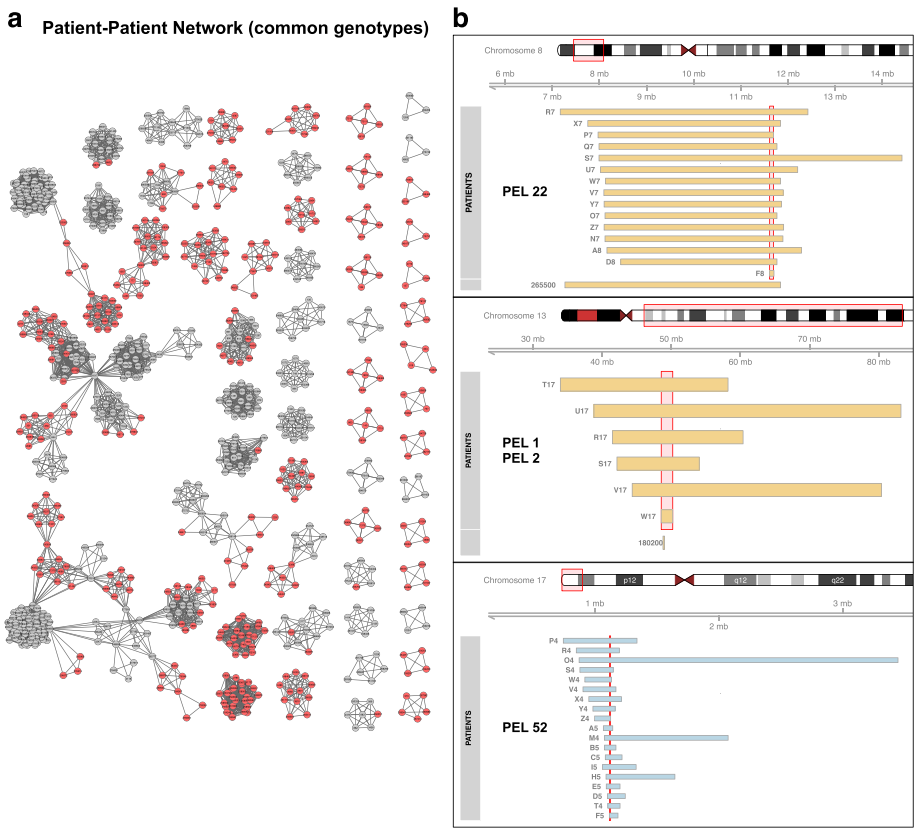


Fig. 5 Genetic and phenotypic relationships between patients. **a** Network of the patients associated with the 336 pathogenic PELs. It includes 830 patients (nodes) and 9606 pairwise relationships supported by genotype-phenotype associations (edges). Grey nodes indicate that patient PELs are associated with at least one known syndrome and red nodes indicate that patient PELs do not overlap with any known genomic syndrome. **b** Examples of known and novel PELs. Patients of PEL 52 that coincide with deletions associated with pulmonic stenosis (MIM 265500). PEL 1 and 2 are patients showing coincidences with the 13q14 deletion syndrome in which the most representative clinical feature is retinoblastoma (MIM 180200). PEL 52 is not associated with any known syndrome and it has patients showing split hand (HP:0001171) and duplications in 17p13.3

patients and controls, respectively). We observed that the frequency of patients overlapping a PEL is independent to their length (Fig. 4a). This effect could be also explained by the specific cohorts of patient CNVs that are collected in DECIPHER. However, it is observable that the frequency of controls that overlap PELs, despite being very low, increases with PEL length (Fig. 4b). This observation agrees with the random distribution of control CNVs across the genome. Overall, these results evidence that our approach is robust at finding phenotypically enriched *loci* (PELs) from a heterogeneous population of patients of different genomic disorders.

We also built a patient network from the genotype and phenotype data of individuals related to pathogenic PELs, revealing clusters of patients that correspond to cliques or sets of them. The resulting network represents a map of the most relevant genotype-phenotype associations that we found in the DECIPHER dataset (Fig. 5a). From ClinVar information, we identified patient CNVs with or without an overlap to known genomic disorders (grey and red nodes in Fig. 5a, respectively). A detailed exploration of these clusters of patients revealed that 164 (~50 %) of the pathogenic PELs (see previous section) overlapped pathogenic CNVs in ClinVar, indicating that PELs are potentially related to known genomic disorders (Table 3 and Additional file 5: Table S4). For instance, in Fig. 5b, the PEL associated with the 8p23.1 deletion coincides with the same genomic location as the genetic variants related to pulmonic stenosis (MIM 265500) in ClinVar. In this particular case, 15 out of 21 patients with deletions in this *locus* (Fig. 5b and PEL 22 from Table 3) were annotated with "Malformation of the heart and great vessels" (HP:0002564, *P-value* of the enrichment 8.3E-10), which is the primary cause of pulmonic stenosis. In addition, there was no healthy individual from the control dataset showing a deletion in this *locus*, suggesting a high penetrance of this phenotype associated to this locus (Table 3).

Another example is retinoblastoma (HP:0009919, *P-value* of the enrichment 6.7E-16 and 3.7E-15 for PEL 1 and 2 respectively; Additional file 3: Table S2) where 6 out of the 7 cases from the patient dataset belong to the same PEL, consisting on deletions in 13q14.2 (chr13:48,544,437-50,206,474, see Fig. 5b). It has been documented that structural variations in this *locus* are associated with the 13q14 deletion syndrome in which the most representative clinical feature is retinoblastoma (MIM 180200) [31, 32]. However, deletions in this *locus*

Table 3 Phenotypically enriched locus overlapping with phenotypically similar known genomic syndromes

PEL ID	Type*	Chr	Start	Length (Kb)	Phenotype	Cases/Carrier (DGV)	*P* value[a]	P [b]	MIM[c]
PEL 240	d	1	243981716	12.547	Abnormality of the skull	13/18 (0)	4.50E-08	100	217990
PEL 193	d	1	243786018	126.15	Abnormality of the skull	14/19 (0)	1.30E-08	100	217990
PEL 68	d	1	243981716	12.547	Microcephaly	12/18 (0)	7.40E-11	100	217990
PEL 49	d	1	243786018	126.15	Microcephaly	13/19 (0)	9.20E-12	100	217990
PEL 25	d	1	243981716	12.547	Aplasia/Hypoplasia of the cerebrum	15/18 (0)	1.80E-12	100	217990
PEL 15	d	1	243786018	126.15	Aplasia/Hypoplasia of the cerebrum	16/19 (0)	3.80E-13	100	217990
PEL 70	d	11	31802605	23.093	Aplasia/Hypoplasia affecting the eye	5/8 (0)	1.00E-08	100	106210
PEL 317	d	14	55242483	200.932	Abnormality of the eye	6/6 (0)	1.80E-04	100	248000
PEL 295	d	4	82082415	31.542	Growth abnormality	9/11 (0)	4.20E-06	100	601665
PEL 484	d	6	407031	170.484	Abnormality of the ocular region	10/16 (1)	4.10E-06	30.2	145400
PEL 484	d	6	407031	170.484	Abnormality of the ocular region	10/16 (1)	4.10E-06	30.2	187350
PEL 347	d	6	1612710	15.026	Abnormality of the ocular region	11/17 (0)	1.60E-07	100	145400
PEL 347	d	6	1612710	15.026	Abnormality of the ocular region	11/17 (0)	1.60E-07	100	187350
PEL 156	d	6	407031	170.484	Abnormality of globe location	9/16 (1)	7.90E-08	28	145400
PEL 100	d	6	2371534	63.584	Hypertelorism	8/13 (1)	2.80E-08	25.7	145400
PEL 88	d	6	1612710	357.639	Hypertelorism	9/16 (1)	3.00E-09	28	145400
PEL 58	d	6	1612710	22.698	Hypertelorism	10/17 (0)	3.40E-11	100	145400
PEL 22	d	8	11610366	83.076	Malformation of the heart and great vessels	15/21 (0)	1.00E-13	100	265500
PEL 6	d	8	11610366	83.076	Abnormality of the cardiovascular system	18/21 (0)	8.00E-15	100	265500
PEL 7	d	8	11610366	83.076	Abnormality of cardiac morphology	17/21 (0)	6.40E-15	100	265500
PEL 452	d	X	102585912	9.472	Abnormality of digit	6/8 (0)	3.50E-05	100	108110

* Duplication (D) and deletion (d). [a] Adjusted P-values from the Fisher's Exact test of the case–control analysis. [b] P is the penetrance, this table show only those PELs with a penetrance higher than 25 %. The penetrance was calculated as described by Cooper et al. [8, 28]. [c] OMIM genomic disorders from ClinVar showing phenotypes that were similar to those found in the respective PEL

Table 4 The novel pathogenic phenotypically enriched locus

PEL ID	Type*	Chr	Start	Length (Kb)	Phenotype	Cases/Carrier (DGV)	*P* value[a]	P [b]
PEL 3	d	3	181296306	175.931	Anophthalmia	6/9 (0)	1.80E-15	100
PEL 5	d	7	95693340	89.973	Ectrodactyly	7/10 (0)	1.70E-14	100
PEL 4	d	3	181296306	175.931	Abnormality of globe size	8/9 (0)	1.90E-14	100
PEL 4	d	3	181296306	175.931	Aplasia/Hypoplasia affecting the eye	8/9 (0)	1.90E-13	100
PEL 71	d	2	200208169	38.268	Abnormality of the palate	11/19 (0)	2.10E-11	100
PEL 31	d	3	181166306	576	Abnormality of globe size	6/9 (0)	4.10E-11	100
PEL 105	d	2	200208169	38.268	Abnormality of the oral cavity	12/19 (0)	2.50E-10	100
PEL 84	d	15	100019051	189.992	Growth delay	13/18 (0)	2.70E-10	100
PEL 31	d	3	181166306	576	Aplasia/Hypoplasia affecting the eye	6/9 (0)	2.70E-10	100
PEL 128	d	2	200208169	38.268	Abnormality of the mouth	14/19 (0)	1.90E-09	100
PEL 131	d	15	100019051	189.992	Growth abnormality	14/18 (0)	6.40E-09	100
PEL 129	d	15	99057570	65.959	Growth delay	11/16 (0)	8.10E-09	100
PEL 69	d	11	31735689	39.768	Aplasia/Hypoplasia affecting the eye	5/8 (0)	1.00E-08	100
PEL 126	d	2	166091754	49.616	Seizures	10/15 (0)	1.00E-08	100
PEL 175	d	7	112349829	160.71	Delayed speech and language development	12/18 (0)	1.00E-08	100
PEL 78	d	14	29904720	411.94	Aplasia/Hypoplasia of the cerebrum	10/12 (0)	1.50E-08	100
PEL 82	d	14	29904720	411.94	Aplasia/Hypoplasia of the cerebrum	10/12 (0)	1.50E-08	100
PEL 141	d	7	114297499	533.997	Delayed speech and language development	11/15 (0)	4.50E-08	100
PEL 166	d	2	166244769	311.476	Seizures	9/14 (0)	6.00E-08	100
PEL 202	d	15	99057570	65.959	Growth abnormality	12/16 (0)	8.80E-08	100
PEL 152	d	14	29904720	411.94	Microcephaly	8/12 (0)	1.60E-07	100
PEL 159	d	14	29904720	411.94	Microcephaly	8/12 (0)	1.60E-07	100
PEL 216	d	2	200208169	38.268	Abnormality of the palate	7/13 (0)	1.60E-07	100
PEL 385	d	2	200246437	0	Abnormality of the face	16/19 (0)	1.60E-07	100
PEL 137	d	6	76509712	359.49	Joint laxity	5/9 (0)	1.60E-07	100
PEL 390	d	7	112349829	160.71	Neurodevelopmental delay	13/18 (0)	1.60E-07	100
PEL 412	d	13	92065689	29.285	Growth delay	9/17 (0)	2.00E-07	100
PEL 419	d	13	92065689	29.285	Growth delay	9/17 (0)	2.00E-07	100
PEL 436	D	16	3831263	32.469	Abnormality of the face	15/18 (0)	4.20E-07	100
PEL 222	d	2	201936560	57.623	Abnormality of the mouth	10/13 (0)	4.80E-07	100
PEL 222	d	2	200208169	38.268	Abnormality of the mouth	10/13 (0)	4.80E-07	100
PEL 242	d	7	114297499	533.997	Neurodevelopmental delay	12/15 (0)	5.20E-07	100
PEL 114	d	13	48557360	146.432	Abnormality of the globe	8/9 (0)	5.90E-07	100
PEL 250	d	7	119973023	238.728	Delayed speech and language development	9/13 (0)	8.80E-07	100
PEL 123	d	12	66224830	22.517	Short stature	7/8 (0)	1.10E-06	100
PEL 184	d	1	28743173	21.263	Deeply set eye	4/7 (0)	1.90E-06	100
PEL 462	d	1	11270844	47.828	Abnormality of the skull	10/14 (0)	1.90E-06	100
PEL 417	d	2	201936560	57.623	Abnormality of the mouth	9/13 (0)	2.00E-06	100
PEL 330	d	14	29781404	230.359	Seizures	7/12 (0)	2.10E-06	100
PEL 227	d	9	77206264	34.573	Seizures	7/10 (0)	2.10E-06	100
PEL 133	D	2	219965169	9.153	Cutaneous finger syndactyly	3/5 (0)	2.30E-06	100
PEL 116	d	3	181296306	6.681	Abnormality of the ocular region	9/9 (0)	2.50E-06	100
PEL 173	D	2	59105866	181.965	Midface retrusion	3/5 (0)	4.00E-06	100
PEL 120	D	2	59105866	181.965	Strabismus	5/5 (0)	4.40E-06	100

Table 4 The novel pathogenic phenotypically enriched locus *(Continued)*

PEL 276	d	7	94174003	41.631	Decreased body weight	4/7 (0)	7.20E-06	100
PEL 329	d	7	95693340	89.973	Abnormality of limb bone morphology	8/10 (0)	1.30E-05	100
PEL 304	d	X	133530468	102.522	Global developmental delay	6/8 (0)	1.50E-05	100
PEL 388	d	10	28842276	86.821	Abnormality of the eyelid	6/8 (0)	1.70E-05	100
PEL 210	D	2	59105866	181.965	Feeding difficulties in infancy	4/5 (0)	1.70E-05	100
PEL 455	d	10	28842276	86.821	Abnormality of the palpebral fissures	5/8 (0)	2.00E-05	100
PEL 251	d	13	41726952	39.763	Abnormal eye morphology	6/7 (0)	2.00E-05	100
PEL 470	d	10	28842276	86.821	Abnormality of the hair	5/8 (0)	2.20E-05	100
PEL 358	d	1	28743173	21.263	Abnormality of globe location	5/7 (0)	2.90E-05	100
PEL 460	d	3	181648378	93.928	Abnormality of the ocular region	7/9 (0)	3.60E-05	100
PEL 338	d	5	170676605	370.857	Abnormality of the cardiac septa	4/6 (0)	3.90E-05	100
PEL 133	D	2	219965169	9.153	Toe syndactyly	3/5 (0)	4.00E-05	100
PEL 454	d	1	177800358	362.172	Short stature	5/7 (0)	4.40E-05	100
PEL 369	d	14	57423809	185.438	Abnormality of the eye	7/8 (0)	4.50E-05	100
PEL 449	d	7	94174003	41.631	Abnormality of the foot	5/7 (0)	4.90E-05	100
PEL 294	d	1	157149743	12.346	Abnormal hair quantity	3/4 (0)	5.70E-05	100
PEL 213	d	1	157149743	12.346	Abnormality of the lip	4/4 (0)	5.70E-05	100
PEL 327	d	19	10640379	140.937	Abnormal genital system morphology	4/5 (0)	7.00E-05	100
PEL 448	d	14	58205713	144.654	Abnormality of the skull	7/8 (0)	8.00E-05	100
PEL 203	d	13	33963658	138.576	Abnormality of the neck	3/3 (0)	8.20E-05	100
PEL 423	d	3	181692255	50.051	Abnormality of the face	9/9 (0)	1.10E-04	100
PEL 324	d	7	94953990	5.573	Growth abnormality	6/6 (0)	2.20E-04	100
PEL 456	D	2	219965169	9.153	Abnormality of the lower limb	4/5 (0)	5.20E-04	100
PEL 361	D	7	106664270	182.398	Strabismus	3/3 (0)	5.40E-04	100
PEL 404	D	7	107527586	136.426	Abnormality of eye movement	3/3 (0)	8.20E-04	100
PEL 407	D	1	113036203	122.933	Abnormality of the palate	3/3 (0)	1.00E-03	100

*Duplication (D) and deletion (d). [a] Adjusted *P-values* from the Fisher's Exact test of the case–control analysis. [b] This table show only those PELs with a penetrance higher than 100 %. The penetrance was calculated as described by Cooper et al. [8, 28]

are frequent in control population (286 samples, Additional file 3: Table S2), suggesting a reduced penetrance for the retinoblastoma phenotype [33] where other factors might be influencing this medical condition. These results indicate that our method is able to identify and prioritize structural variants that are strongly associated with pathological phenotypes.

In addition, several clusters of patients associated with pathogenic PELs that were found not to be apparently associated with known genomic syndromes but significantly enriched for highly specific clinical features such as ectrodactyly, malformations in the heart, defects in atrial septum, and anophthalmia (Table 4). More than 50 % (172 out of 336) of the pathogenic PELs do not overlap with any known genomic disorder in ClinVar so they can be candidates for novel syndromic *loci*. For instance, we detected a cluster of patients showing a severe medical condition that is known as split hand (HP:0001171) with duplications in 17p13.3 (Fig. 5b). The PEL associated with this cluster (PEL 52, *P-value* of 1.1E-13 for Fisher's exact test in Additional file 3: Table S2) shows a very high penetrance for this phenotype, but its patients display a broad spectrum of specific clinical outcomes that are associated with this medical condition. The phenotype "abnormality of the hand" (HP:0001155) was the most enriched HPO term (*P-value* of the enrichment 2.7E-07 for PEL 52 in Additional file 3: Table S2) associated with this PEL (Table 4). *A priori* this cluster of genetically and phenotypically related patients could be considered a novel genomic disorder. Indeed, after reviewing the available clinical literature we found evidence of syndromic presence in micro-duplications spanning this *locus*, related to a previous familiar study with a similar phenotype [34]. We distinguished seven broad domains of phenotypic abnormalities through the examination of the phenotypic relationships between patients from PELs (Additional file 3: Table S2): abnormality of the ocular region, abnormality of the limb bone morphology, abnormality of the skull, abnormality of the face, abnormality of the cerebrum, abnormality of the cardiovascular system

and growth delay. Our results show that this approach provides a new tool for the characterization and the study of phenotype-genotype relationships in a systematic genome-wide manner. For instance, it is possible to characterize the pleotropic effects of pathogenic CNVs or to study mutations on different mutated genomic regions that are related to similar phenotypes.

Additive phenotypic effects of pathogenic CNVs

We observed that the length of CNVs is correlated to complex phenotypic profiles of DECIPHER patients, as shown in Fig. 2a. This complexity is here defined as the number of distinct clinical features that have been observed by a physician in a patient. Thus, it was explored if the length of significant PELs is associated with complex pathogenicity or adds more phenotypes according to the number of different genomic regions that are affected. To illustrate this effect, we analyzed the phenotypic relationships between significant PELs that are in close genomic regions. For instance, deletions in 10q25.13 (PEL 149) and 10q26.13 (PEL 239) are related to different phenotypes such as abnormalities of the cardiovascular system and the genitourinary system respectively (Fig. 6a). Most cases with deletions in

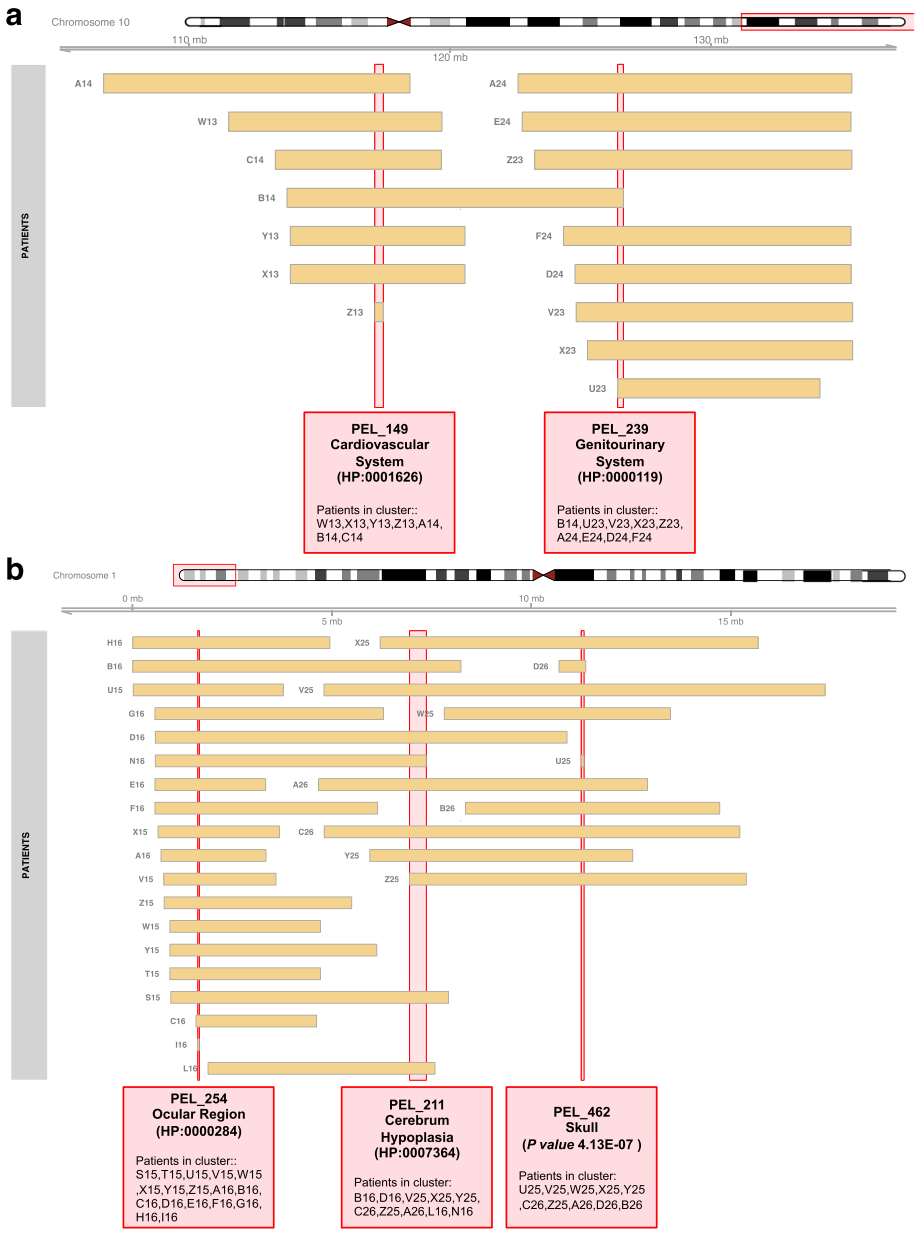


Fig. 6 Illustrative examples of additive phenotypic effects of PELs

10q25.13 (5 of 7 cases) are associated with malformations of the heart and great vessels, denoting a very specific clinical feature. In addition, cases with deletions in 10q26.13 are related to defects in the genitourinary system (PEL 239 in Fig. 6a). The patient B14 (Fig. 6a) shows both phenotypes and has a deletion that overlaps both genomic *loci* (PEL 149 and PEL 239, Fig. 6a). This example illustrates an additive effect, accumulating specific clinical features according to the extension of structural variants with respect to the genome of reference. This effect is also noticeable for more complex genetic relationships among *loci* of patient CNVs associated with significant PELs as those represented in Fig. 6b. In this case, three different clusters (cliques) of highly interconnected patients were detected, indicating that some individuals are included in more than one cluster or PEL. These different PELs were found to be associated with abnormalities of the ocular region, aplasia/hypoplasia of the cerebrum and abnormalities of the skull (PEL 254, 211 and 462, respectively, Fig. 6b). All patients overlapping these regions from significant PELs show the phenotype if they have the structural variation, except for patient S15 who apparently does not have signs of hypoplasia of the cerebrum. Different PELs associated with the same phenotype (HPO terms) were found located in contiguous or even the same genomic region. In some other cases, distinct PELs were essentially the same clusters of patients except with variations in one or two individuals (they should be considered one PEL). Thus, despite the precise identification of genomic coordinates of individual CNVs being a technological limitation, the wide adoption of next generation sequencing methods by clinical studies may solve the current shortcomings in the array-based CNV data used for this analysis.

Conclusions

This work presents a combined analysis of network-based approaches, phenotype enrichment and genetic association studies for patient CNVs in the DECIPHER database. A set of methods was developed to identify clusters of patients that are genetically and phenotypically related. The newly developed methods used here have potential usefulness for a wide range of applications, such as prediction of unknown syndromes, characterization of candidate pathogenic structural variants and the identification likely associated phenotypes with a specific *locus*. This procedure could be improved using more specific clinical features of the patients, so physicians should be encouraged to submit detailed phenotype data. This work evidences the need for advancement in consolidated standards and public repositories for genomic and medical records in genomic and personalized medicine.

Additional files

Additional file 1: Table S1. Collection of DECIPHER patients CNVs, mode of inheritance and phenotypes that have been analyzed from DECIPHER. (XLSX 2440 kb)

Additional file 2: Figure S1. Distribution of topological parameters calculated from the patient network based on the overlapping between individual DECIPHER CNVs. (PDF 84 kb)

Additional file 3: Table S2. Phenotypically enriched loci (PELs) after the enrichment analysis (hypergeometric test, *P*-values <0.05.) and case–control analysis (Fisher's exact test, *P*-values <0.05). (XLSX 269 kb)

Additional file 4: Table S3. Relationship between Phenotypically enriched loci (PELs) and genomic disorders used from ClinVar, OMIM phenotypes caused by likely pathogenic and pathogenic CNVs from ClinVar. (XLSX 64 kb)

Additional file 5: Table S4. Relationship between Phenotypically enriched loci (PELs) and genomic disorders used from DECIPHER. (XLSX 59 kb)

Abbreviations

CNV/s: Copy number variation/s; PEL/s: Phenotypically enriched locus/loci, HPO, human phenotype ontology; DGV: Database of genomic variants; OMIM: Online Mendelian Inheritance in Man; MIM: OMIM identification number; DECIPHER: DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ARP and JAGR developed the methods. ARP, FSJ, MAM, MC and JAGR designed the study and supervised all the analysis. ARP, AB, RRL and JAGR implemented and optimized the workflow. ARP, AB, RRL and JAGR performed statistical analyses and analyzed DECIPHER data. ARP, MC and JAGR wrote the manuscript with contributions from all other authors. All authors read and approved the final manuscript.

Acknowledgements

This work was funded by CIBERER (U741), EU-FP7-Systems Microscopy NoE (Grant Agreement 258068), and grants SAF2011-26518, SAF2012-33110 (MEC, Spain), BIO2014-56092-R (MINECO and FEDER, Spain), and CTS-486, CTS-1507 and CVI-06585 Excellence Grants (Junta de Andalucía, Spain), and BIO-267 (fondos PAIDI, Junta de Andalucía, Spain). MC is grateful to UK's BBSRC for core funding. The "CIBER de Enfermedades Raras" is an initiative from the ISCIII (Spain). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. ARP is recipient of a postdoctoral fellowship granted by Fundación Ramón Areces. This study makes use of data generated by the DECIPHER community. A full list of centres who contributed to the generation of the data is available from http://decipher.sanger.ac.uk and via email from decipher@sanger.ac.uk. Funding for the project was provided by the Wellcome Trust.

Author details

[1]Universidad de Málaga, Andalucía Tech, Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, and IBIMA (Biomedical Research Institute of Málaga), E-29071 Málaga, Spain. [2]CIBER de Enfermedades Raras (CIBERER), E-29071 Málaga, Spain. [3]The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK. [4]Present address: The European Molecular Biology Laboratory Heidelberg, 69117 Heidelberg, Germany.

Received: 22 September 2015 Accepted: 7 March 2016
Published online: 15 March 2016

References

1. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36:949–51.

2. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. Science. 2004;305:525–8.
3. Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet. 2015;16:172–83.
4. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14:125–38.
5. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
6. Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang N, Hurles ME, Haendel M, Washington NL, Smedley D, Mungall CJ, Lewis SE, Ott C-E, Bauer S, Schofield PN, Mundlos S, Spielmann M, Robinson PN. Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol. 2014;15:423.
7. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008;453:56–64.
8. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43:838–46.
9. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, Schuurs-Hoeijmakers JH, Hoischen A, Pfundt R, Krumm N, Carvill GL, Li D, Amaral D, Brown N, Lockhart PJ, Scheffer IE, Alberti A, Shaw M, Pettinato R, Tervo R, de Leeuw N, Reijnders MRF, Torchia BS, Peeters H, Thompson E, O'Roak BJ, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46:1063–71.
10. Cook EH, Scherer SW. Copy-number variations associated with neuropsychiatric conditions. Nature. 2008;455:919–23.
11. Malhotra D, Sebat J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell. 2012;148:1223–41.
12. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Almeida J, Bacchelli E, Bader GD, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bölte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Bryson SE, Carson AR, Casallo G, Casey J, Chung BHY, Cochrane L, Corsello C, et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature. 2010;466:368–72.
13. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho Y-J, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463:899–905.
14. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, Mangion J, Roberton-Lowe C, Marshall AJ, Petretto E, Hodges MD, Bhangal G, Patel SG, Sheehan-Rooney K, Duda M, Cook PR, Evans DJ, Domin J, Flint J, Boyle JJ, Pusey CD, Cook HT. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006;439:851–5.
15. Knight SJ, Regan R, Nicod A, Horsley SW, Kearney L, Homfray T, Winter RM, Bolton P, Flint J. Subtle chromosomal rearrangements in children with unexplained mental retardation. Lancet. 1999;354:1676–81.
16. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, et al.. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31:1102–10.
17. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13:395–405.
18. Church GM. The personal genome project. Mol Syst Biol. 2005;1:2005.0030.
19. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 2015;372:793–5.
20. Loscalzo J, Barabasi A-L. Systems biology and the future of medicine. Wiley Interdiscip Rev Syst Biol Med. 2011;3:619–27.
21. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T, Bevan AP, Bragin E, Chatzimichali EA, Gribble S, Jones P, Krishnappa N, Mason LE, Miller R, Morley KI, Parthiban V, Prigmore E, Rajan D, Sifrim A, Swaminathan GJ, Tivey AR, Middleton A, Parker M, Carter NP, Barrett JC, Hurles ME, et al.. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2014;385:1305–14.
22. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ. DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res. 2014;42(Database issue):D993–D1000.
23. Hwang TH, Atluri G, Kuang R, Kumar V, Starr T, Silverstein KA, Haverty PM, Zhang Z, Liu J. Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers. BMC Genomics. 2013;14:440.
24. Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 2005;435:814–8.
25. Palla G, Barabási A-L, Vicsek T. Quantifying social group evolution. Nature. 2007;446:664–7.
26. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83:610–5.
27. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42(Database issue):D986–92.
28. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C, Stalker H, Hamid R, Hannig V, Abdel-Hamid H, Bader P, McCracken E, Niyazov D, Leppig K, Thiese H, Hummel M, Alexander N, Gorski J, Kussmann J, Shashi V, Johnson K, Rehder C, Ballif BC, Shaffer LG, Eichler EE. Corrigendum: A copy number variation morbidity map of developmental delay. Nat Genet. 2014;46:1040.
29. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28:495–501.
30. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5.
31. Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, Dryja TP. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature. 1986;323:643–6.
32. Sparkes RS, Sparkes MC, Wilson MG, Towner JW, Benedict W, Murphree AL, Yunis JJ. Regional assignment of genes for human esterase D and retinoblastoma to chromosome band 13q14. Science. 1980;208:1042–4.
33. Mitter D, Ullmann R, Muradyan A, Klein-Hitpass L, Kanber D, Ounap K, Kaulisch M, Lohmann D. Genotype-phenotype correlations in patients with retinoblastoma and interstitial 13q deletions. Eur J Hum Genet. 2011;19:947–58.
34. Klopocki E, Lohan S, Doelken SC, Stricker S, Ockeloen CW, Soares Thiele de Aguiar R, Lezirovitz K, Mingroni Netto RC, Jamsheer A, Shah H, Kurth I, Habenicht R, Warman M, Devriendt K, Kordass U, Hempel M, Rajab A, Mäkitie O, Naveed M, Radhakrishna U, Antonarakis SE, Horn D, Mundlos S. Duplications of BHLHA9 are associated with ectrodactyly and tibia hemimelia inherited in non-Mendelian fashion. J Med Genet. 2012;49:119–25.