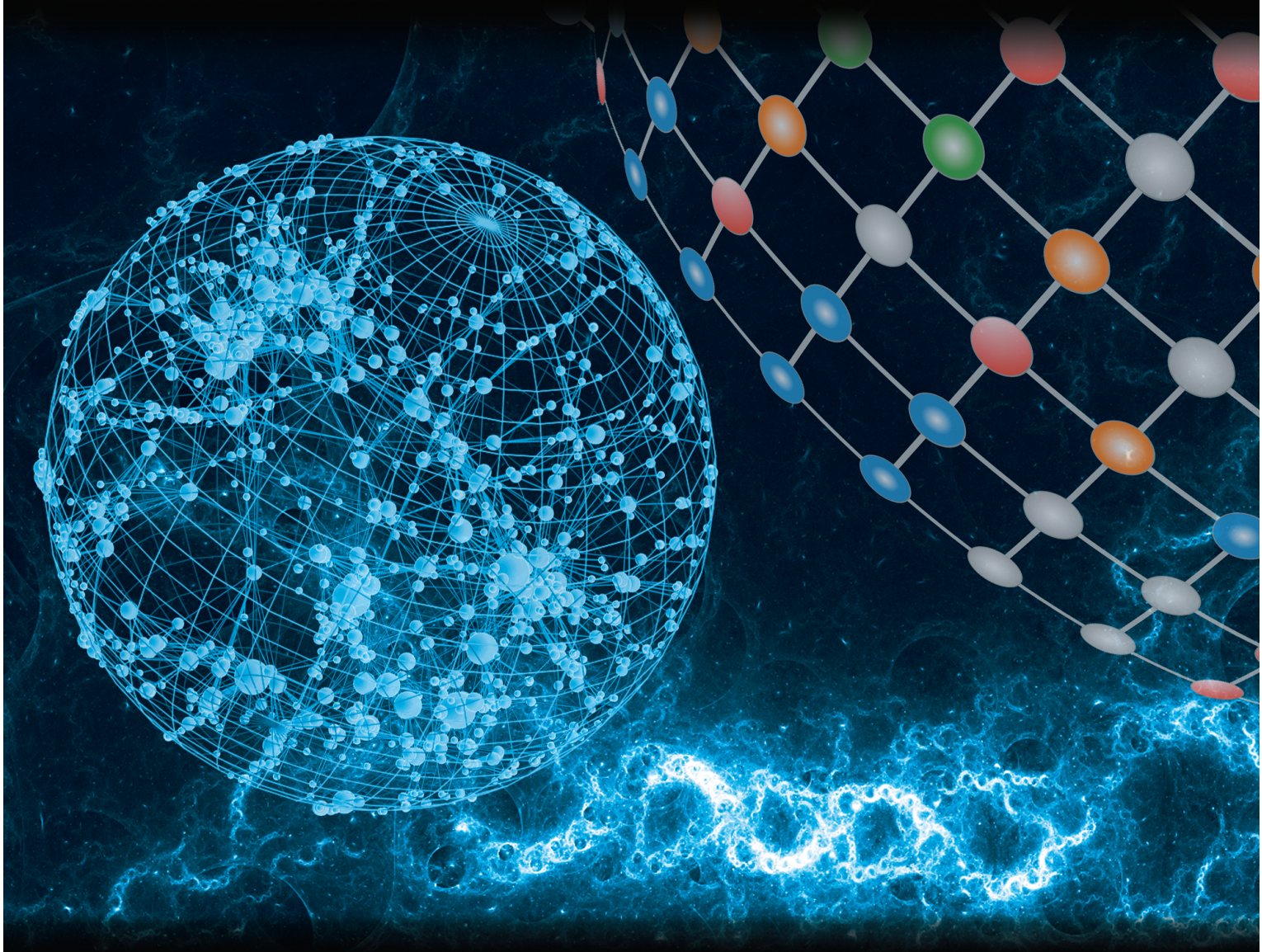
Tesis Doctoral

Systems Biology Approaches to Evaluate Disease Modularity



Armando Reyes Palomares

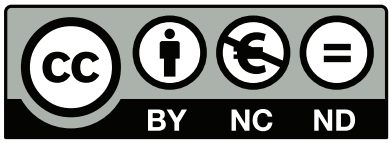
Málaga 2014

UNIVERSIDAD DE MÁLAGA

Publicaciones y
Divulgación Científica

AUTOR: Armando Reyes Palomares

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga



Systems Biology Approaches to Evaluate Disease Modularity

Armando Reyes Palomares

TESIS DOCTORAL

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

UNIVERSIDAD DE MÁLAGA

Málaga, 2014

Systems Biology Approaches to Evaluate Disease Modularity

Memoria presentada para optar al grado de
Doctor por la Universidad de Málaga

Armando Reyes Palomares

DIRECTORES

Miguel Ángel Medina Torres
Catedrático de Bioquímica y
Biología Molecular de la
Universidad de Málaga

Francisca Sánchez Jiménez
Catedrática de Bioquímica y
Biología Molecular de la
Universidad de Málaga

UNIVERSIDAD
DE MÁLAGA

Miguel Ángel Medina Torres, Catedrático de Bioquímica y Biología Molecular de la Universidad Málaga, Francisca Mª Sánchez Jiménez, Catedrática de Bioquímica y Biología Molecular de la Universidad Málaga

CERTIFICAN

Que D. Armando Reyes Palomares, Licenciado en Biología 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 "Systems Biology Approaches to Evaluate Disease Modularity".

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.

Málaga, Marzo de 2014

Miguel Ángel Medina Torres Francisca Sánchez Jiménez

The research in this Doctoral Thesis titled 'Systems Biology Approaches to Evaluate Disease modularity' was carried out at the group PAIDI BIO-267 of the Department of Molecular Biology and Biochemistry of University of Malaga. The author and supervisors also belongs to the unit 741 of Centre for Biomedical Network Research on Rare Diseases (CIBERER).

This research has been funded by the CIBERER and Grants SAF2011-26528 (MEC, Spain), CVI-06585 and CTS-1507 (Junta de Andalucia and FEDER), and AMER Consorptium (FEDER-Innterconecta, CDTI, Spain). This work is one of the activities for the Bioinformatic Platform for Rare Diseases (BIER) of CIBERER, which is an initiative of "Instituto de Salud Caros III".

The author of this Thesis has been recipient of a Training University Lecturers (Competitive FPU Grant from Spanish Ministry of Education).

PUBLICATIONS

List of publications included in this Thesis:

A combined model of hepatic polyamine and sulfur amino acid metabolism to analyze S-adenosyl methionine availability
Reyes-Palomares A, Montañez R, Sánchez-Jiménez F, Medina MA.
Amino Acids. 2012 Feb;42(2-3):597-610.
doi:10.1007/s00726-011-1035-7.

Systems biology metabolic modeling assistant: an ontology-based tool for the integration of metabolic data in kinetic modeling
Reyes-Palomares A*, Montañez R*, Real-Chicharro A, Chniber O, Kerzazi A, Navas-Delgado I, Medina MA, Aldana-Montes JF, Sánchez-Jiménez F.
Bioinformatics. 2009 Mar 15;25(6):834-5. doi:10.1093/bioinformatics/btp061.

Global analysis of the human pathophenotypic similarity gene network merges disease module components
Reyes-Palomares A, Rodríguez-López R, Ranea JA, Sánchez Jiménez F, Medina MA.
PLoS One. 2013;8(2):e56653.
doi: 10.1371/journal.pone.0056653.

PhenUMA: A Tool for Integrating the Biomedical Relationships among Genes and Diseases
Rodríguez-López R*, Reyes-Palomares A*, Sánchez-Jiménez F, and Medina MA.
(Submitted to BMC Genomics)

Network Medicine Approaches for Systematic Identification of Phenotype and Structural Variants Associations
Reyes-Palomares A. *et al.*
(Manuscript in preparation)

* These authors have contributed equally.

OTHER PUBLICATIONS

List of publications not included in this Thesis:

Regulatory cross-talk of mouse liver polyamine and methionine metabolic pathways: a systemic approach to its physiopathological consequences.
Correa-Fiz F, Reyes-Palomares A, Fajardo I, Melgarejo E, Gutiérrez A, García-Ranea JA, Medina MA, Sánchez-Jiménez F.
Amino Acids. 2012 Feb;42(2-3):577-95.
doi: 10.1007/s00726-011-1044-6.

What is known on angiogenesis-related rare diseases? A systematic review of literature.
Rodríguez-Caso L, Reyes-Palomares A, Sánchez-Jiménez F, Quesada AR, Medina MÁ.
J Cell Mol Med. 2012 Dec;16(12):2872-93.
doi: 10.1111/j.1582-4934.2012.01616.x. Review.

Histamine: an undercover agent in multiple rare diseases?
Pino-Ángeles A, Reyes-Palomares A, Melgarejo E, Sánchez-Jiménez F.
J Cell Mol Med. 2012 Sep;16(9):1947-60.
doi: 10.1111/j.1582-4934.2012.01566.x. Review.

First steps in computational systems biology: A practical session in metabolic modeling and simulation.
Reyes-Palomares A, Sánchez-Jiménez F, Medina MÁ.
Biochem Mol Biol Educ. 2009 May;37(3):178-81.
doi: 10.1002/bmb.20281.

AGRADECIMIENTOS

Estimados compañeros, son muchos años con vosotros y la lista de agradecimientos es extensa afortunadamente. Antes de escribir estas palabras no he podido evitar pensar en perspectiva de lo que representa esa lista. Así que mentalmente la he esbozado como una red (deformación profesional), en la que los nodos sois vosotros y las relaciones pues se establecen en función de lo compartido como esfuerzos, afinidades y vivencias tanto personales como profesionales; o también los hay por pura coincidencia, pero que también cuentan porque el roce hace el cariño.

A simple vista, en esa "red" de relaciones humanas y profesionales destacan dos enormes "hubs". Miguel Ángel y Kika, muchas gracias por permitirme disfrutar del uso libre de mi propio razonamiento. Las pruebas las tenéis, podría haber hecho esto algo antes, pero no ha sido así porque estaba muy a gusto con vosotros. Me siento muy afortunado de que seáis mis directores, compañeros y amigos.

Miguel Ángel, tu me fichaste. Mis ambiciones estaban adormecidas y carecía de un reto claro ante un sistema educativo muy limitado para las mentes inquietas y creativas. Sin embargo, con tus asignaturas me presentaste un nuevo horizonte, una forma alternativa de expresar mis conocimientos y de entrenar mis habilidades. Desde que me senté en tu despacho para plantear los primeros proyectos, me lo dejaste tajantemente claro, "ten en cuenta que para este trabajo tienes que ser autodidacta". A priori no entendía muy bien eso de ser autónomo en un colectivo de investigación; pero no tuvo que pasar mucho tiempo para darme cuenta del reto que me planteabas, algo así como: "*sapere aude* chaval" (eso de "chaval", obviamente, es cosecha propia). Muchas gracias por enseñarme unas pautas tan esenciales, dejar que me desahogue con mis tensiones y ayudarme a relativizarlas, así como darme tu opinión, siempre cuidadosamente aséptica para no influenciarme, en la toma de decisiones muy importantes para mi.

Kika, que te voy a decir que no sepas todo lo que te debo, como tu dices conectamos bien y no nos hacen falta muchas palabras para entendernos. Pero, muchas gracias porque sin acabar la carrera apostate por mí para formar parte de un nuevo proyecto para el grupo al que nos presentamos como una unidad completa de biocomputación. En esa época, no era muy consciente de como funcionaban estas cosas, pero ahora con más perspectiva pienso con absoluta franqueza que fuiste valiente. Todo maestro tiene su librillo pero tu debes de tener unos cuantos porque nadie ponía en duda la seguridad que emanaba ese sexteto de becarios (el largo, almu, raúl, ian, ale y un servidor). Puede que fuéramos –y seamos– algo frikis pero con la moral de legionarios. Muchas gracias por esa inyección de moral, ese buen rollo, tus empujones para tomar iniciativas y ser tan ecuánime al tomar decisiones difíciles dentro del colectivo. He sido y soy testigo directo de tu perseverancia y dedicación a la profesión. Así como lo has sido en los asuntos más cercanos a mi también lo has sido con él/la/los que se sienta/n a mi lado (between o among lo mismo da :)).

Muchas gracias al "Computational Biology and Data Mining Research Group" del "Max Delbrück Center for Molecular Medicine" por acogerme durante 3 meses de estancia predoctoral. Fue una experiencia fabulosa y me sentí como en casa, me encantó el grupo y conocí a muy buenos amigos: Miguel, Nancy, Enrique, Jean-Fred, David, Arvind, Martin y Merie.

En cuanto a los Seniors del grupo. Juan Antonio, son muchos años mesa con mesa y compartiendo almuerzos. Gracias por tu valiosos consejos. Para mi es un placer tenerte como amigo y compañero, y opino que gente como tu, con perspectiva y gran dinamismo, es fundamental para agilizar la universidad más aún si cabe en esta época. Nos conocemos bastante bien, tu sabes muchos de mis secretos y yo sé que no te vas sin comprobar tu archivador ;) . José Luis e Ignacio, aunque no lo sepáis en los seminarios aprendí mucho de vuestros comentarios, preguntas y discusiones con Kika o MAM sobre los experimentos de los compañeros con bata, gracias. Ana, directora, gracias por tu amabilidad facilitando esas variopintas consultas o gestiones.

Los demás componentes de esta red modular. El lado oscuro, en el que he encontrado buenos compañeros y amigos, además de compartir una gran cantidad de horas de trabajo. Alejandro y Rocío habéis sido mis dos tendones de Aquiles. Si alguien me dice bioinformático tiene que saber que la parte informática es fundamentalmente gracias a vosotros. He tenido la suerte de conoceros y entenderme perfectamente con ambos, tanto en lo profesional como en lo personal. Rocío, eres una gran "aprendiz". En un período corto de tiempo hemos entablado una buena amistad, obtenido buenos resultados, organizado eventos científicos y nos han agraciado con un premio honorífico. ¡¡no se puede pedir más!! ;) Rubio se te echa de menos, tu personalidad era la chispa del mejor de los ambientes... si, si, ambientes de los que a ti te gustan. Nunca olvidaré ese viaje a Ámsterdam dos chavalines con todo el equipo de desarrolladores de Cytoscape... un momento genuino!! Raúl, gracias compañero por compartir esas largas jornadas de trabajo y combinarlas con buenas aventuras (Tenerife, Barcelona, Valencia, etc.). Eres un trabajador nato y un artista pero de los de verdad. Aure y Almu, dos inseparables que en menos de un año habéis volado los dos. Durante tres añitos fuimos el equipo ciberero todo un escuadrón. Muchas gracias a ambos porque vuestra efectiva organización siempre me ha sido de gran ayuda y referencia. Ian, el matemático descifrador, horas y horas de matlab, con mucha personalidad y excelente persona. Jim, my British mate. Thanks for sharing scientific discussion with me, but also talk to you about whatever topic society, sports, perspectives... I was thinking, it should be fine to resume beer sessions. Bea y Aníbal, la "nueva" remesa del lado oscuro, proyectos de ingenieros bioinformáticos, ánimos que ya mismo os vendrá la mejor parte, y sabéis que podéis contar conmigo. Muchas gracias también al resto del grupo; un pedazo de "wet lab" (pasado y presente, os podría decir mucho a todos y cada uno de vosotros pero sois tantos): Hicham (queda pendiente ese cafelito de moda), Betty :), Gianni, Melissa (me pusiste las pilas con la Tesis, gracias), Flor, María Victoria, Javi, Luiso (muchas gracias por tus reflexivas visitas) y Carmen, Esther, Casimiro, Auxi, Carlos^2, Joaquín.

Por último, quiero agradecérselo a mis familiares y amigos, a los que les debo la vida así como el día a día.

En primer lugar, a mi clon genético mi hermano. Gracias por todo Arturo, vivo con la tranquilidad y la confianza, estando más lejos o más cerca, de tenerte siempre a mi lado. Eres una gran persona, estoy muy orgulloso de ti. Este trabajo –en parte– también es tuyo porque a saber en que momento embrionario los dos dejamos de ser uno.

A mi madre, Sissi la emperatriz de Málaga. Gracias por velar por nuestra felicidad incondicionalmente, nos cuidaste dando el más sólido apoyo, y nos concediste la libertad y la responsabilidad de ser nosotros mismos y con nuestras circunstancias. Siempre nos tendrás a los dos aunque nunca podremos recompensarte por todo lo que nos has dado.

A mi padre, siempre he dicho que uno puede ser feliz descubriendo y trabajando de lo que más conoce. Tu nos enseñaste a deducir y a razonar, para reconocer con criterio la importancia de las cosas y la relatividad de las apariencias. Te agradezco tus lecciones y tu empeño en que las aprendiéramos, siempre nos ayudaste para que aprendiéramos a motivarnos.

A Antonia y Lázaro, mi mezcla peligrosa ;) que nos criaron y aguantaron toda nuestra infancia y adolescencia, con la paciencia y la sabiduría de los ancianos. Nunca olvidaré lo que representáis para nosotros. A Manolo, muchas gracias por confiar en nosotros y querernos como a sobrinos, siempre ayudas a ver lo sencilla que pueden ser las cosas, la vida. José desde pequeño seguimos juntos primos y mejores amigos, a Lulú, a mi abuelo Adolfo y a la Chica, a Lourdes y Almudena, demás familia y amigos, muchas gracias por compartir vuestras vidas conmigo y apoyarme de corazón.

Por último, mi compañera la dueña de mis pasiones, María, que has sido testigo en primera línea de este trabajo. Gracias por aguantarme, mimarme y amarme como desde el primer día.

Gracias

CONTENTS

PREÁMBULO

En el curso 2006-2007, sin haber acabado la carrera fue cuando me inicié en el mundo de la investigación. Pero mi aventura realmente comenzó un curso antes, estudiando bioquímica metabólica que impartía mi director de tesis, Miguel Ángel Medina Torres. En aquel curso académico descubrí un mundo de posibilidades y distinto a lo que habitualmente cualquier alumno aprende en las aulas. A este mundo, en concreto, lo suelo llamar como el "mercado del conocimiento" en biología molecular y bioquímica, como explicaré más adelante. Pero permita que continúe explicando mi experiencia didáctica. Miguel Ángel, nos propuso como única tarea obligatoria que los propios alumnos diseñáramos una práctica lo razonablemente afín a los contenidos de la asignatura. A pesar de que algunos no entienden la finalidad de esta tarea, todo el que la trabaja suele acabar convencido de que es una experiencia enriquecedora. La realidad objetiva de esta tarea es que si quieres aprender algo, lo que tienes que demostrar es que eres capaz de explicarlo. Él exige esta garantía para que el alumno estudie con bastante profundidad un tema como para llegar casi a dominarlo.

De alguna manera, para darle alguna pista de mi argumento, pretendo recalcar que cualquier conocimiento adquirido debe ser formalizado –o estandarizado– para que otros lo puedan entender. Por aquel entonces, ya disponía de ciertas habilidades para la informática pero carecía de conocimientos en programación y computación. Después de explorar y reflexionar, le propuse a Miguel Ángel diseñar una práctica con la que cualquier alumno pudiera hacer un modelo matemático de una ruta metabólica como la glucolisis. Tras su visto, como alumno me planteé ¿qué alumno de 3º de biología puede aprender en una sola práctica todos esos conocimientos en matemáticas y en programación? Fue entonces, en junio de 2005, después de días documentándome e investigando cuando descubrí las puertas del "mercado de conocimiento". Y digo las puertas porque esto solo es el principio del viaje hasta acabar de escribir esta tesis. En

primer lugar, encontré mucho software de programas para diseñar y simular rutas metabólicas (CellDesigner®, CoPaSi o GePaSi, Systems Biology WorkBench, etc.), diversas bases de datos de información metabólica y enzimática (BRENDA, SABIO-RK, EMP, KEGG) y repositorios con modelos metabólicos completos y funcionales (JWS-Online y BioModels). Una de los mayores ventajas de este mercado es que todo es open-source, y cualquiera podía –y aún puede– servirse a su gusto con fines académicos.

Había topado con la cuna de la biología de sistemas, una corriente de planteamientos innovadores, atractivos y acordes a las necesidades de lo que buscaba para mi tarea de bioquímica metabólica. Toda una casta de científicos con un perfil mayoritariamente tecnológico, tales como Hiroaki Kitano del "Systems Biology Institute" o John C. Doyle y Michael Hucka del "California Institute of Technology", fundaron una plataforma que se denominó "SBML Systems Biology Markup Language" (SBML, http://www.sbml.org/). Esta plataforma, se inició en 2001 y fue evolucionando progresivamente hasta convertir al SBML en el formato estándar para transferir y almacenar la información mínima que se requiere para diseñar un modelo biológico, principalmente orientado a la modelización de reacciones bioquímicas. Este formato es compatible con múltiples programas que utilizan la misma información pero con propósitos muy distintos, un mercado que presenta múltiples ofertas y que cada uno puede servirse libremente. Por ejemplo, algunos utilizan ficheros SBML para visualizar la información utilizando diversas notaciones gráficas, como por ejemplo la "Systems Biology Graphical Notation" (SBGN, http://www.sbgn.org/). Otros permiten hacer un estudio estructural del conjunto de reacciones descritas en los modelos (ficheros SBML) a partir de sus estequiometrias, como por ejemplo el análisis de las sub-redes mínimas que permiten mantener un estado estacionario ("elementary flux modes"). Sin embargo, para mi práctica, utilicé programas para diseñar y analizar modelos metabólicos a partir de la información cinética enzimática conocida para cada reacción. Escogí este tipo de programas porque permiten estudiar a la vez las propiedades dinámicas y estructurales de un conjunto de reacciones bioquímicas en cuestión. Estos programas solo requieren introducir de forma ordenada las ecuaciones y los parámetros cinéticos de cada reacción para hacer las simulaciones y estudiar como cambian las concentraciones de los metabolitos y los flujos metabólicos a lo largo del tiempo.

La trayectoria que he vivido en primera persona del proyecto SBML, me permite afirmar que formalizó unos estándares básicos –y prácticos– para asentar los fundamentos de la modelización de la biología de sistemas actual, tras consolidarse como disciplina académica hace algo más de una década. Gracias al software disponible, a la variedad de algoritmos desarrollados, a los repositorios

de modelos online y a las facilidades que ha aportado este proyecto al resto de miembros de la comunidad científica, es posible una lección práctica sobre la modelización de la glucolisis en una sesión de 3 horas en el aula de informática. Utilizando todos estos recursos fue como me inicié en la investigación, diseñando una actividad con fines docentes que todavía impartimos Miguel Ángel y yo desde entonces.

En el año 2007, me incorporé como asistente de investigación a la unidad 741 del Centro de Investigaciones Biomédicas en Red de Enfermedades Raras (CIBERER) dirigida por Francisca Sánchez Jiménez, mi co-directora. CIBERER era un proyecto de ámbito nacional y gracias al cual siguen contratados algunos miembros de nuestro grupo. En 2010, el Ministerio de Educación y Ciencia me concedió una beca del Programa Nacional de Formación de Profesorado Universitario que he disfrutado hasta la fecha propuesta para la defensa de esta Tesis Doctoral.

ABSTRACT

Biological systems are in a constant process of innovation as an essential precondition to evolve. For this reason, the emergence of phenotypic variation is an inherent property of complex adaptive systems. Even though living systems acquire robustness to internal and external disturbances during the evolutionary process, pathological conditions entail impairments to their functionality. This is the rationale for studying biomedical issues according to the organizational properties of biological systems, with the aim of understanding the mechanisms of diseases.

I first discuss the theoretical background that is suitable for the research included in this Thesis, such as my own interpretation of systems biology, the current theories about the origin of the biological modularity and some evolutionary considerations that concern in the genotype-phenotype relationships. In this section, I also argue the use and the development of integrative systems biology methods that should be addressed to evaluate disease modules: computational models (i.e. mathematical and network-based models) and other standardized efforts (ontologies and different databases with biological and biomedical data). Then, I enunciate the hypothesis and declare the objectives that motivated this research: i) mathematical modelling based on kinetic law formalism for studying the functional modularity of the metabolism; ii) the development of a workflow to integrate metabolic and kinetic data from different databases for metabolic modelling; iii) the evaluation of the functional coherence in phenotypic relationships between disease-causing genes by using network-based analysis; iv) the development of an integrative framework of biomedical information; v) the use of network medicine approaches to study the phenotypic and genotypic relationships in a heterogeneous group of patients with genetic syndromes. Finally, the results derived from the research carried out in this Thesis are included in the form of already published articles and manuscripts (either submitted or in preparation).

CHAPTER 1
INTRODUCTION

The challenge of modern biology is to understand the structural and functional properties of complex adaptive systems. In the case of cellular networks, the notion of module is the abstraction of interconnected biomolecules from whole molecular interactions in cells[1,2]. I argue that this abstraction can be supported by topological and functional criteria, but to understand how biological structures integrate functionality is necessary to consider their dynamical singularities. This is the rationale behind the approaches of systems biology[3–5]. In fact, only on the basis of the fundamental organization of biological systems is possible to evaluate how modularity –or any other systemic property– pervades living things. Systems biology approaches are aimed for a deeper understanding of certain biomedical issues[6,7]. Many integration efforts are made in this direction. Here, I focus only on diseases with a genetic origin, where their molecular bases are known or they show a pattern of inheritance. This means that at least one genotype predisposes to suffer or express any or several of the clinical features classically recognised in these medical conditions. This points to another issue concerning how these pathological conditions depend on the modular architecture of the underlying biology.

In this section, the theoretical background suitable for the research studies included in this Thesis is introduced. First, I describe my perception on the current systemic view of biology and the current theories about the origin of the modular organization in biology. I continue with an overview of the evolutionary processes that involve the emergence of phenotypes and how they can be related to the modular nature of genetic diseases. Finally, I summarize the state of the art of established and emerging 'omic technologies with biomedical applications and some aspects of the integrative systems biology approach used in this Thesis.

NOTA: Los contenidos de la Introducción comprendidos entre las páginas 10 y 30 no se muestran en esta versión de la Tesis por corresponderse con material inédito sujeto a derechos de *copyright* impuestos por la editorial de la revista donde se pretende publicar.

CHAPTER 2
HYPOTHESIS & OBJECTIVES

Living systems are in a constant process of innovation as an essential precondition to evolve. This innovation process is not intended to provide the most appropriate features in the short-term, but simply to create new biological features. Subsequent evolutionary processes are the responsible for shaping structural and functional properties of biological systems over long periods of time. Under these circumstances, it should be assumed that the manifestation of pathological process is inherent in our biology. Despite biological systems acquire robustness to internal and external changes such as genetic variations and environmental fluctuations respectively[29]; the disease state entails an impairment of functionality (loss of phenotypic robustness). This is the rationale for studying human diseases according to the organizational properties of biological adaptive systems, with the aim of understanding how they affect their fitness. The incidence of genetic diseases in the population is the consequence of certain biological constraints and susceptible to be studied by a systemic view. It could be considered as the main hypothesis underlying this thesis. Nonetheless, there are two premises that actually constitute the specific hypotheses of this Thesis.

First hypothesis: *The computational approaches are required to study the modular organization of biological systems.* In particular, I will focus on biochemical reaction systems that conform metabolic modules related to the S-adenosylmethionine.

Second hypothesis: *The phenotypic coherence between disease-causing genes depends on their functional context.* For this reason, the integration of molecular interactions underlying the genotype-phenotype relationship is helpful to elucidate the modularity of genetic diseases.

This Thesis emphasizes the use and the development of integrative systems biology approaches, such as computational models and standard methods, which should be addressed to outline the modularity of diseases. Thus, to study and test these hypotheses we defined the following objectives:

Objective 1. The use of mathematical modelling based on kinetic law formalism to study the functional properties of the modularity in the metabolism. In particular, I will use different metabolic models to evaluate how biochemical reactions determine the availability of S-adenosylmethione, one of the main precursors of polyamines and an essential methyl donor in cells.

Objective 2. The development of a workflow that integrates metabolic and kinetic data from different databases and helpful to design or extend metabolic models. For this objective, the workflow should comply with the current proposals to standardize models of biochemical reactions, concretely those that have been agreed in the various consortia and international groups of systems biology, such as the use of ontologies and formats agreed (i.e. SBML).

Objective 3. The evaluation of functional coherence in phenotypic relationships between disease-causing genes by using network-based analysis. The phenotypic relationships will be established by calculating semantic similarity between genes through the use of an ontology of human abnormalities.

Objective 4. The development of a tool to integrate phenotypic and functional relationships between genes, diseases and phenotypes. This tool will be freely available online and will be built according to the discussions and conclusions derived from the results obtained for the fulfilment of the previous aim.

Objective 5. The use of network medicine approaches to study the phenotypic and genotypic relationships in a heterogeneous group of patients with genetic syndromes.

CHAPTER 3
RESULTS. THESIS PUBLICATIONS.

3.1. A combined model of hepatic polyamine and sulfur amino acid metabolism to analyze S-adenosyl methionine availability

3.2. Systems biology metabolic modeling assistant: an ontology-based tool for the integration of metabolic data in kinetic modeling

3.3. Global analysis of the human pathophenotypic similarity gene network merges disease module components

3.4. PhenUMA: A Tool for Integrating the Biomedical Relationships among Genes and Diseases

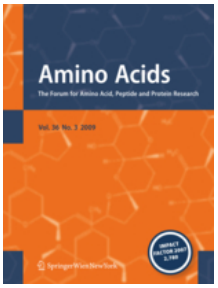
3.5. Network Medicine Approaches for Systematic Identification of Phenotype and Structural Variants Associations

CHAPTER 3. PUBLICATION 1

A Combined Model of Hepatic Polyamine and Sulfur Amino Acid Metabolism to Analyze S-Adenosyl Methionine Availability

Armando Reyes-Palomares, Raúl Montañez, Francisca Sánchez Jiménez and Miguel Angel Medina

Amino Acids 2012 42(2-3): 597-610. Published: February 2012



Print ISSN: 0939-4451

Online ISSN: 1438-2199

Supplementary Material: Yes

Status: Published

DOI: 10.1007/s00726-011-1035-7

CHAPTER 3. PUBLICATION 1

Supplementary Material for:

A Combined Model of Hepatic Polyamine and Sulfur Amino Acid Metabolism to Analyze S-Adenosyl Methionine Availability

Supplementary material INCLUDED in this Thesis:

Online Resource 3.
PDF file containing the whole set of equations included in the two versions of the combined model.

Online Resource 4.
PDF file containing Tables S1 and S2.

Online Resource 6.
PDF file containing the whole set of abbreviations.

This supplementary material is NOT INCLUDED in this Thesis but it is available online:

Online Resource 1.
Combined model (version 1) in SBML format.

Online Resource 2.
Combined model (version 2) in SBML format.

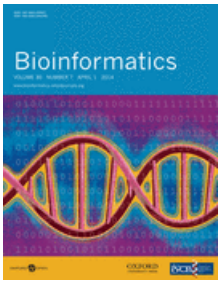
Online Resource 5.
Excel file containing the whole set of data concerning the sensitivity analysis.

CHAPTER 3. PUBLICATION 2

Systems biology Metabolic Modeling Assistant: an Ontology-Based Tool for the Integration of Metabolic Data in Kinetic Modeling

Armando Reyes-Palomares, Raul Montañez, Alejando Real-Chicharro, Othmane Chniber, Amine Kerzazi, Ismael Navas-Delgado, Miguel Ángel Medina, José F. Aldana-Montes, Francisca Sánchez-Jiménez

Bioinformatics 2009 25 (6): 834-835. Published: February 2012



Print ISSN: 1367-4803

Online ISSN: 1460-2059

Supplementary Material: No

Status: Published

DOI: 10.1093/bioinformatics/btp061

CHAPTER 3. PUBLICATION 3

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

Armando Reyes-Palomares, Rocío Rodríguez-López, Juan A. G. Ranea, Francisca Sánchez Jiménez and Miguel Angel Medina

PLoS ONE 2013 8(2): e56653. Published: February 2013

Print ISSN:

Online ISSN: 1932-6203

Supplementary Material: Yes

Status: Published

DOI: 10.1371/journal.pone.0056653

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

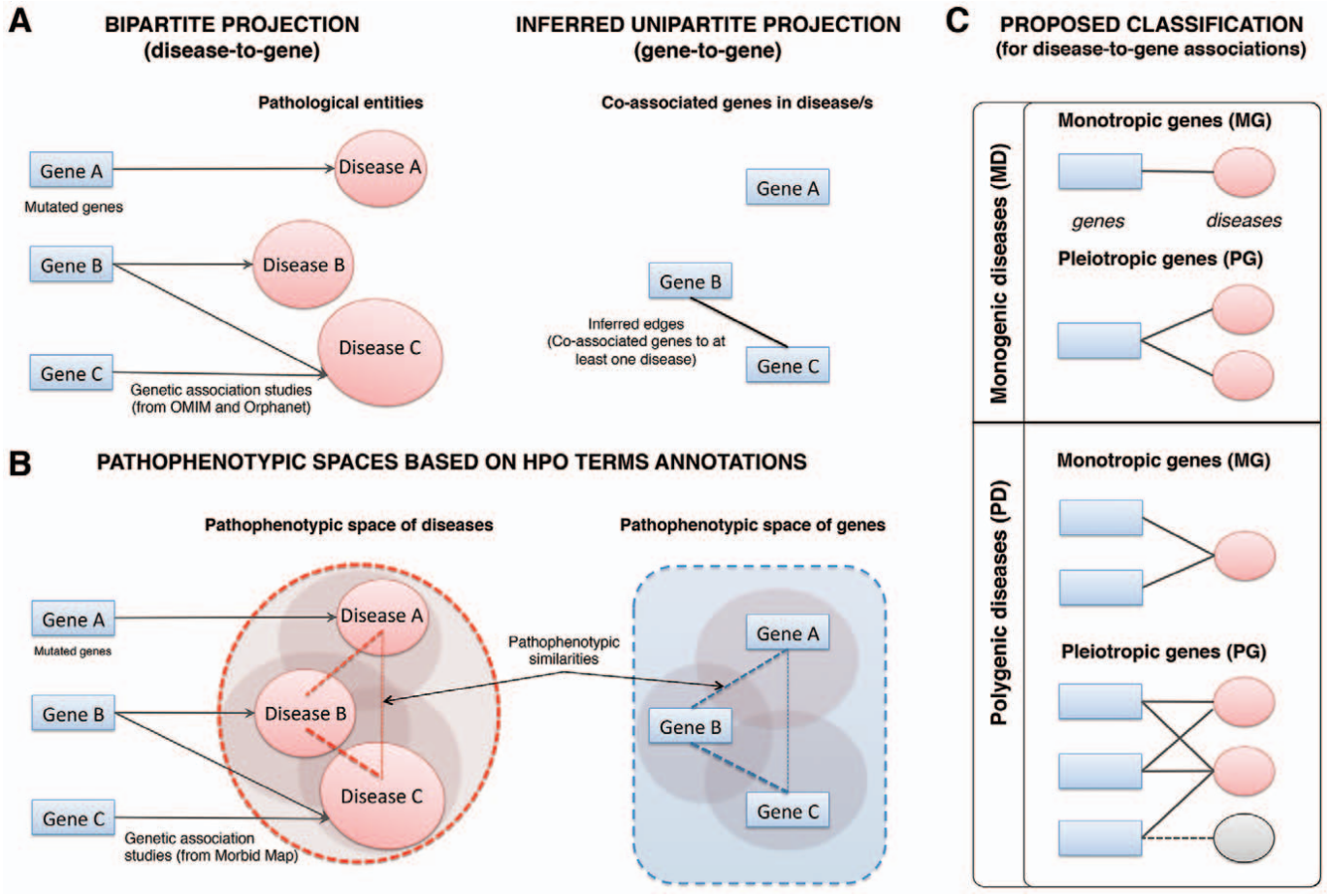


Figure 1. Schematic representation of distinct disease-to-gene relationships. Different disease associations between genes using (A) the data from genetic disease databases or (B) their associated pathological phenotypes. (A) The co-associations of genes in disease/s allow the inference of gene-to-gene 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.

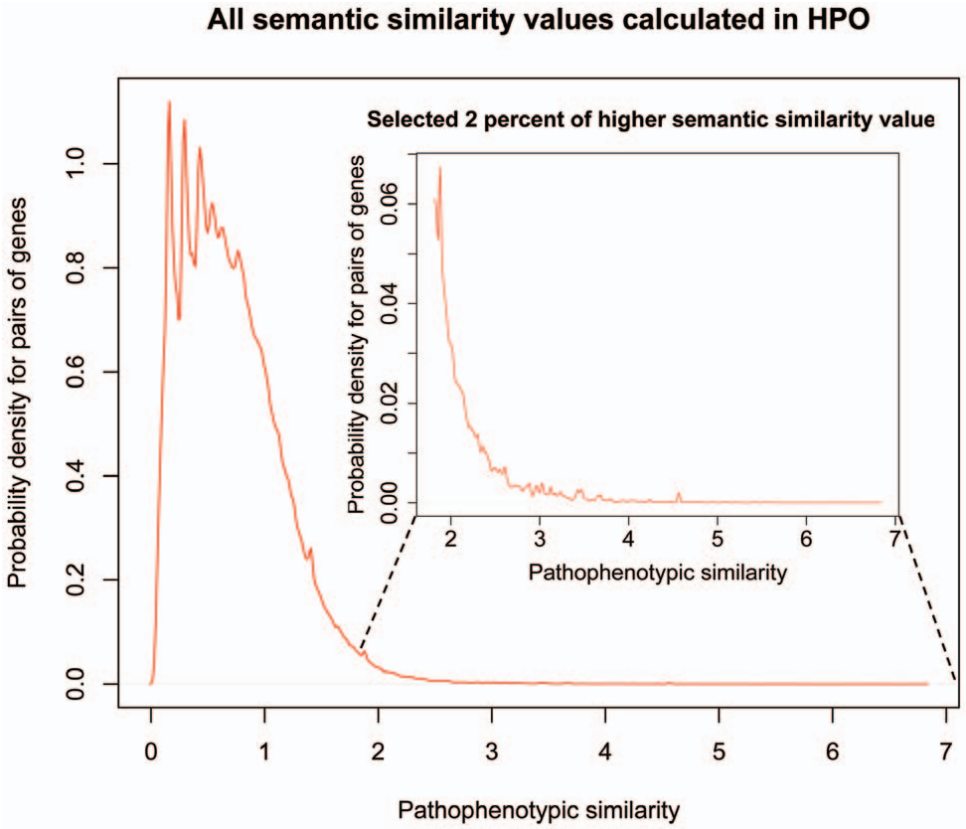


Figure 2. Probability density function for pathophenotypic similarities among pairs of genes in HPO. Densities of the pathophenotypic similarity values for all annotated genes in HPO (outer plot) and for the top 2% of gene pairs with the highest pathophenotypic similarities (inner plot). The bandwidth 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

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

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:

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

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:

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

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:

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

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:

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

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,

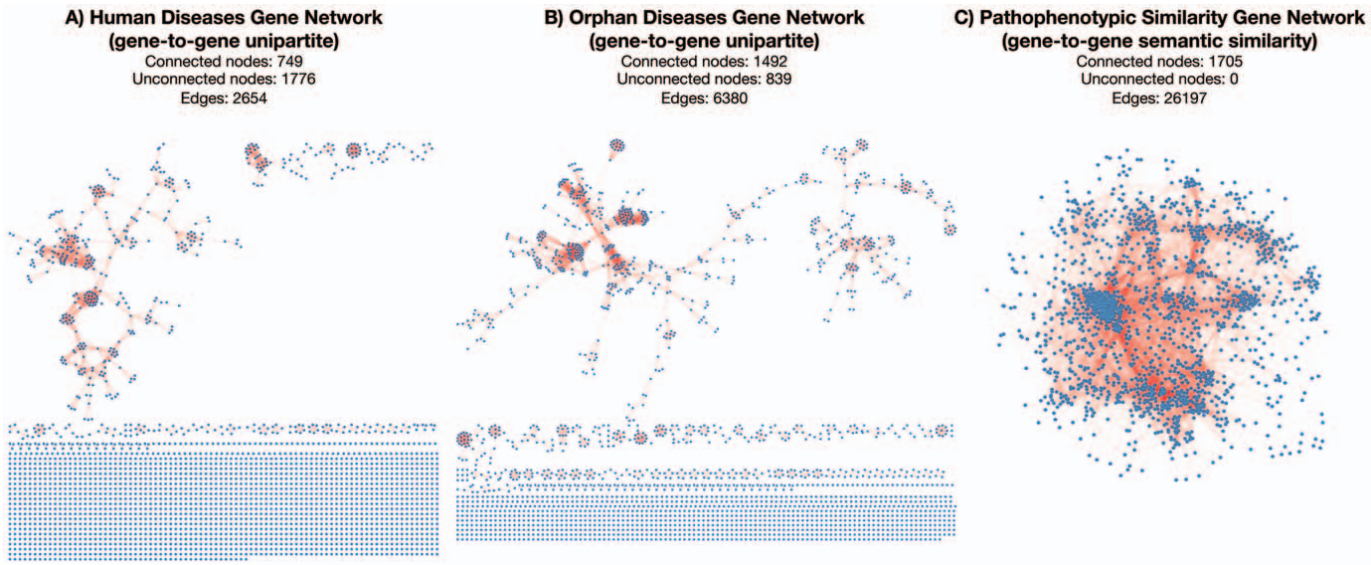


Figure 3. Unipartite gene-to-gene projections of the disease networks and the pathophenotypic similarity gene network. Human diseases genetic network (HDGN in panel A), Orphan diseases genetic network (ODGN in Panel B) and Human pathophenotype similarity gene network (PSGN in panel C). PSGN consists 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.

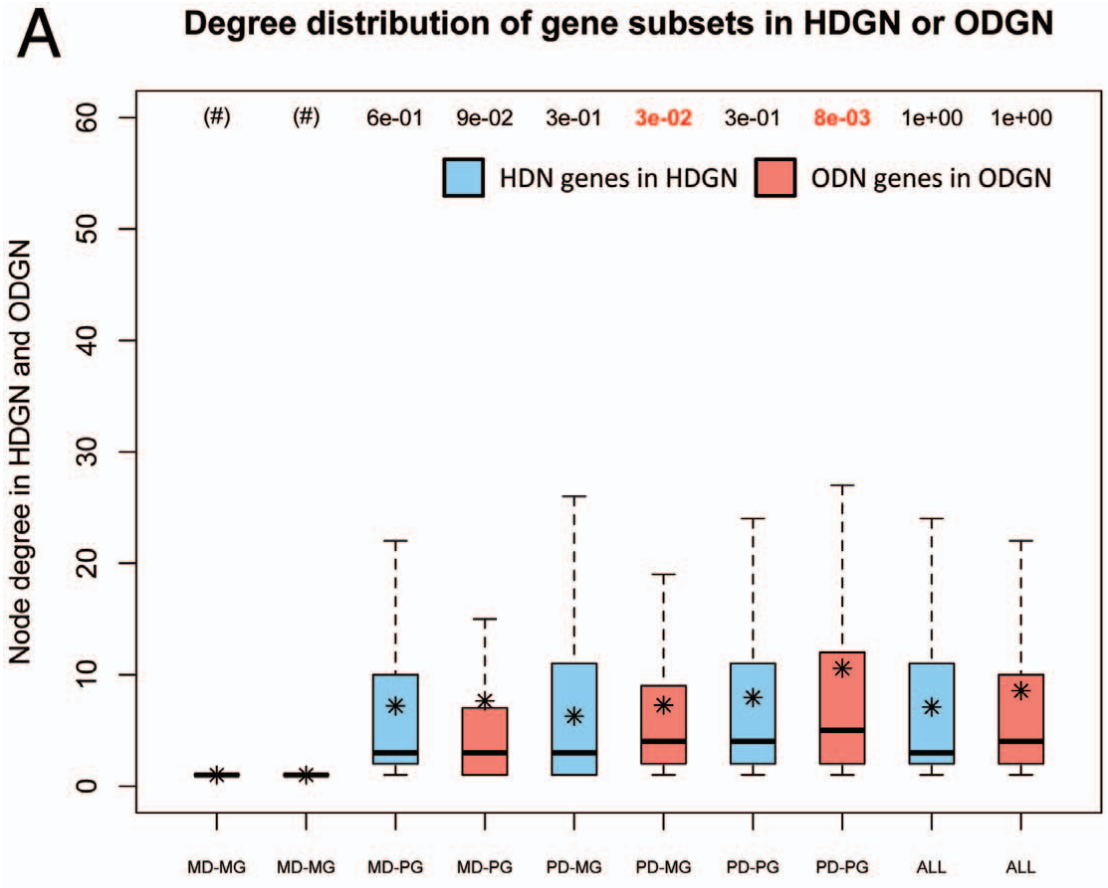
The resulting intersections of PSGN with each unipartite projection proved 528 shared nodes and 1055 shared edges for HDGN and 931 and 1669 for ODGN (Table 4). Therefore, 39% and 26% of inferred pathological relationships intersect with pathophenotypic similarities of PSGN, even improving the intersection 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


A
Degree distribution of gene subsets in HDGN or ODGN
Node degree in HDGN and ODGN
(#) (#) 6e-01 9e-02 3e-01 3e-02 3e-01 8e-03 1e+00 1e+00
HDN genes in HDGN
ODN genes in ODGN
0 10 20 30 40 50 60
MD-MG MD-MG MD-PG MD-PG PD-MG PD-MG PD-PG PD-PG ALL ALL

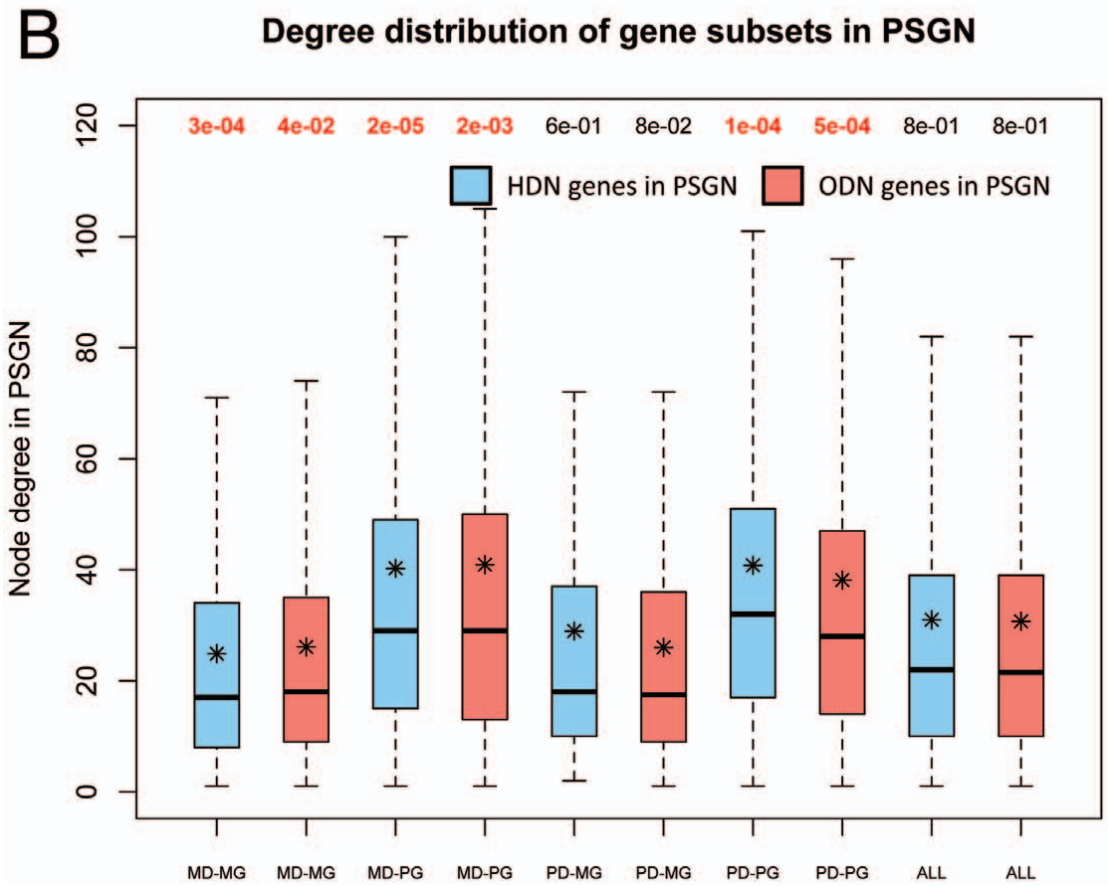

B
Degree distribution of gene subsets in PSGN
Node degree in PSGN
3e-04 4e-02 2e-05 2e-03 6e-01 8e-02 1e-04 5e-04 8e-01 8e-01
HDN genes in PSGN
ODN genes in PSGN
0 20 40 60 80 100 120
MD-MG MD-MG MD-PG MD-PG PD-MG PD-MG PD-PG PD-PG ALL ALL

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.

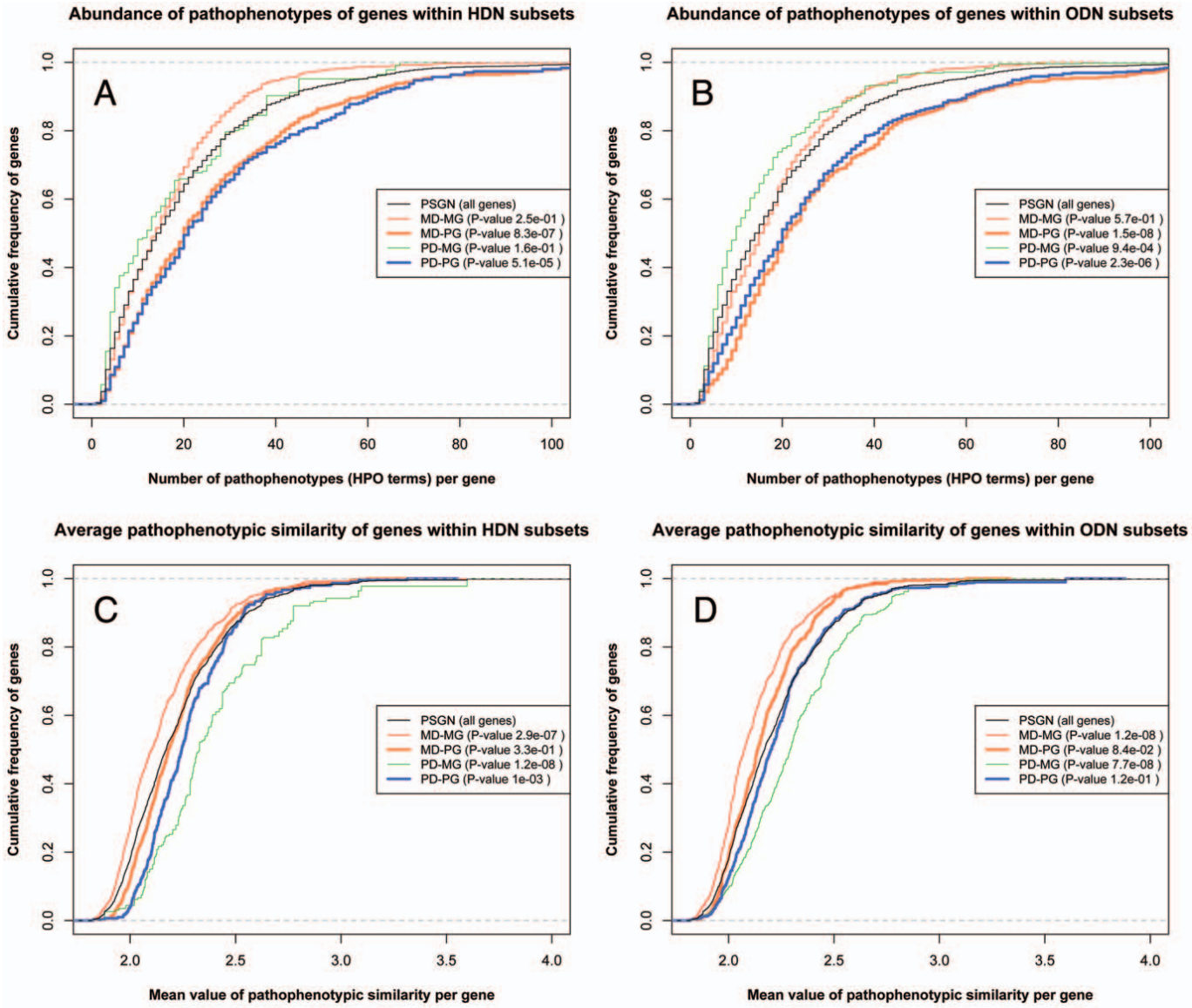


Figure 5. Distributions of the number of pathophenotypes and pathophenotypic similarities in each subset. MD-MG (red line), MD-PG (orange line), PD-MG (green line), PD-PG (blue line) and PSGN (Black line). Upper panels represent the cumulative frequency of the number of specific pathophenotypes annotated 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

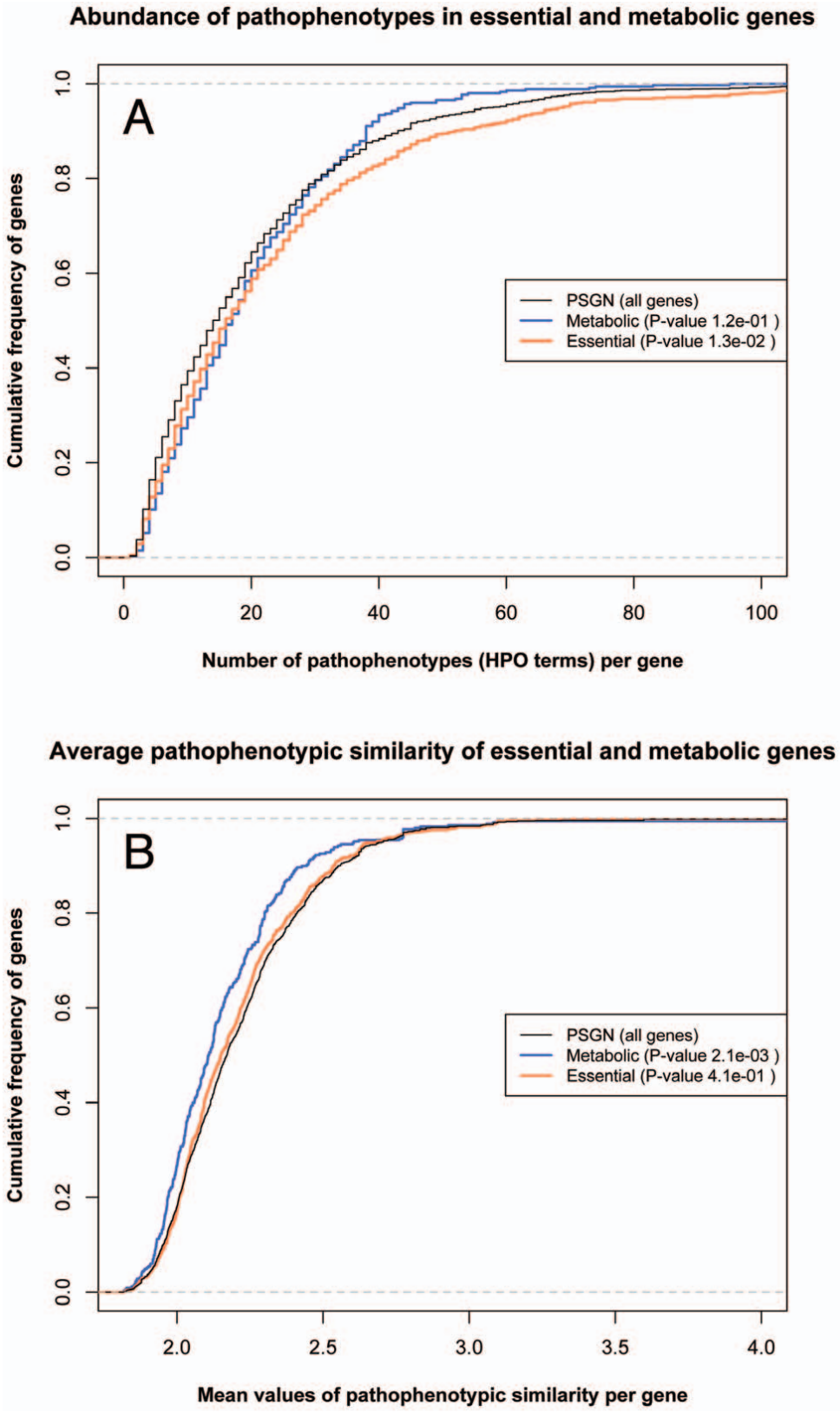


Figure 6. Distributions of the number of pathophenotypes and the pathophenotypic similarities for metabolic and essential genes. Metabolic genes (orange line), essential genes (orange line) and the PSGN (Black line). Upper panel (A) represents the cumulative frequency of the number of specific pathophenotypes 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
ÔR	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(ÔR)	Expected edges in the intersection	E(ÔR) = Q*R/M	36.01	27.96	27.43	24.33	543.57	196.95
ÔR	Observed edges in the intersection	–	422	35	124	17	2473	194
Q (%)	Percentage of query edges	perc_Q = 100*ÔR/Q	23.72	2.54	11.70	1.81	8.16	1.77
R (%)	Percentage of reference edges	perc_R = 100*ÔR/R	2.71	0.23	38.63	5.30	14.35	1.13
Jac_sim	Jaccard coefficient of similarity	Jac_sim = ÔR/(QvR)	0.0250	0.0021	0.0986	0.0137	0.0549	0.0069
P value	P-value of the intersection	Pval = P(X> = ÔR)	**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

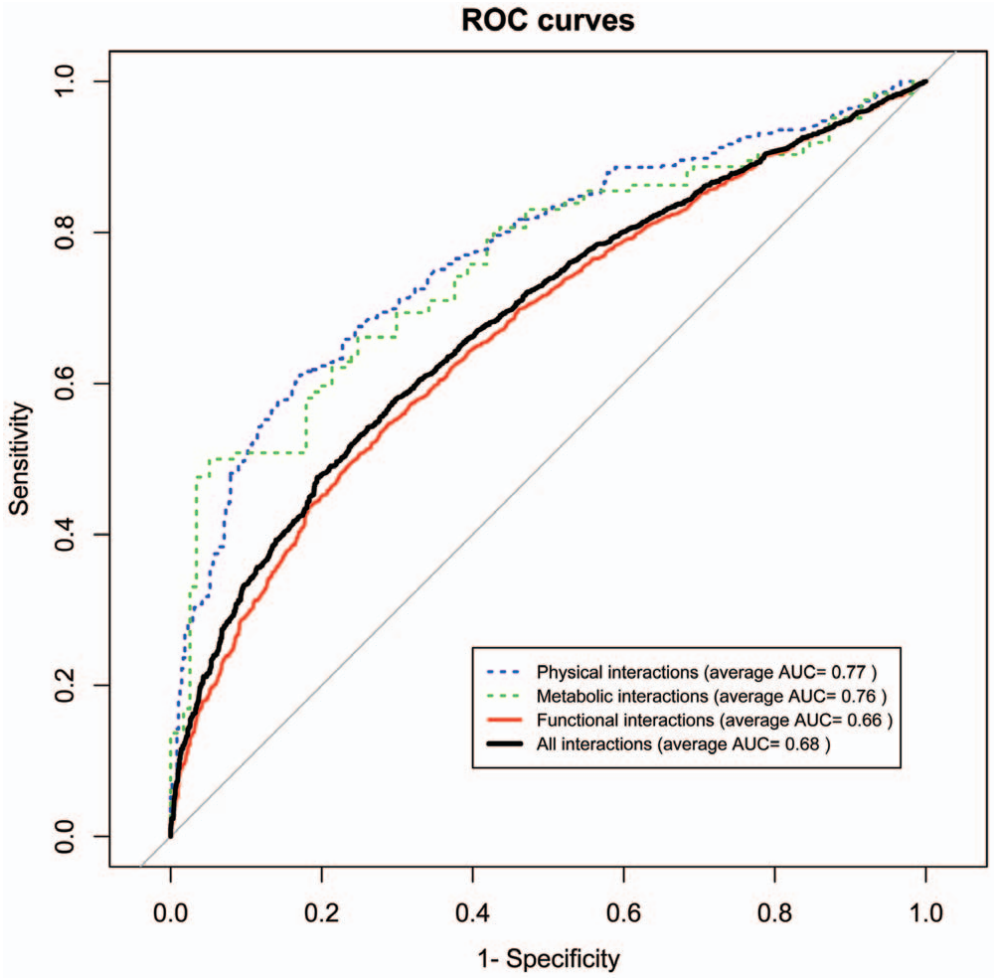


Figure 7. Receiver operative characteristic (ROC) curve performance by biomolecular interactions of pathophenotypic similarities. Physical interactions (dashed blue line), metabolic flux correlations (dashed green line), functional interactions (red continuous line) and an integrated interactome generated 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

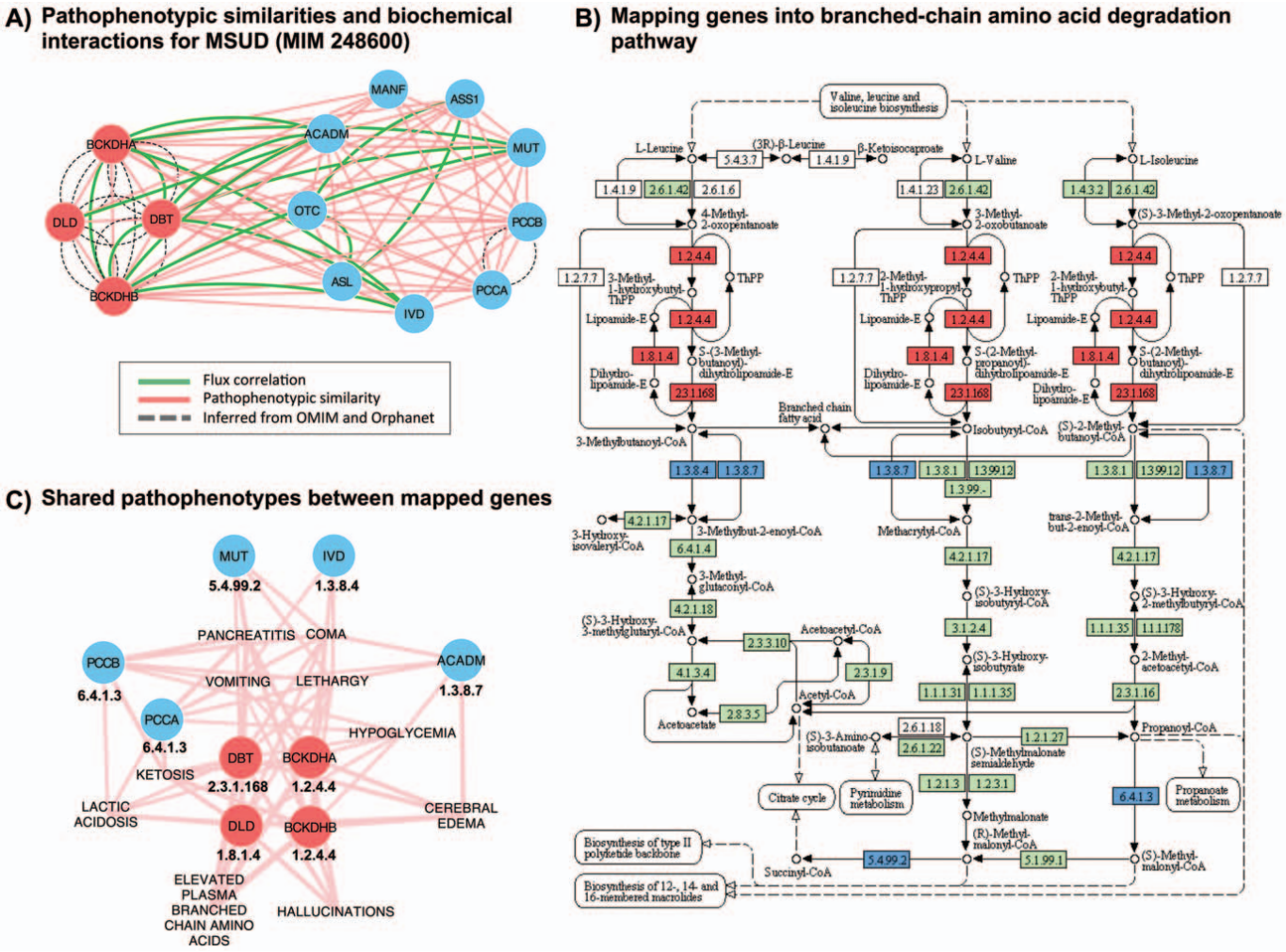


Figure 8. Maple syrup urine disease pathological and metabolic interactions. In red genes associated with MSUD and in blue pathophenotypic similar genes. (A) Pathophenotypic similarity gene sub-network for MSUD causing genes. It can be noteworthy that there are no inferred relationships between MSUD 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.

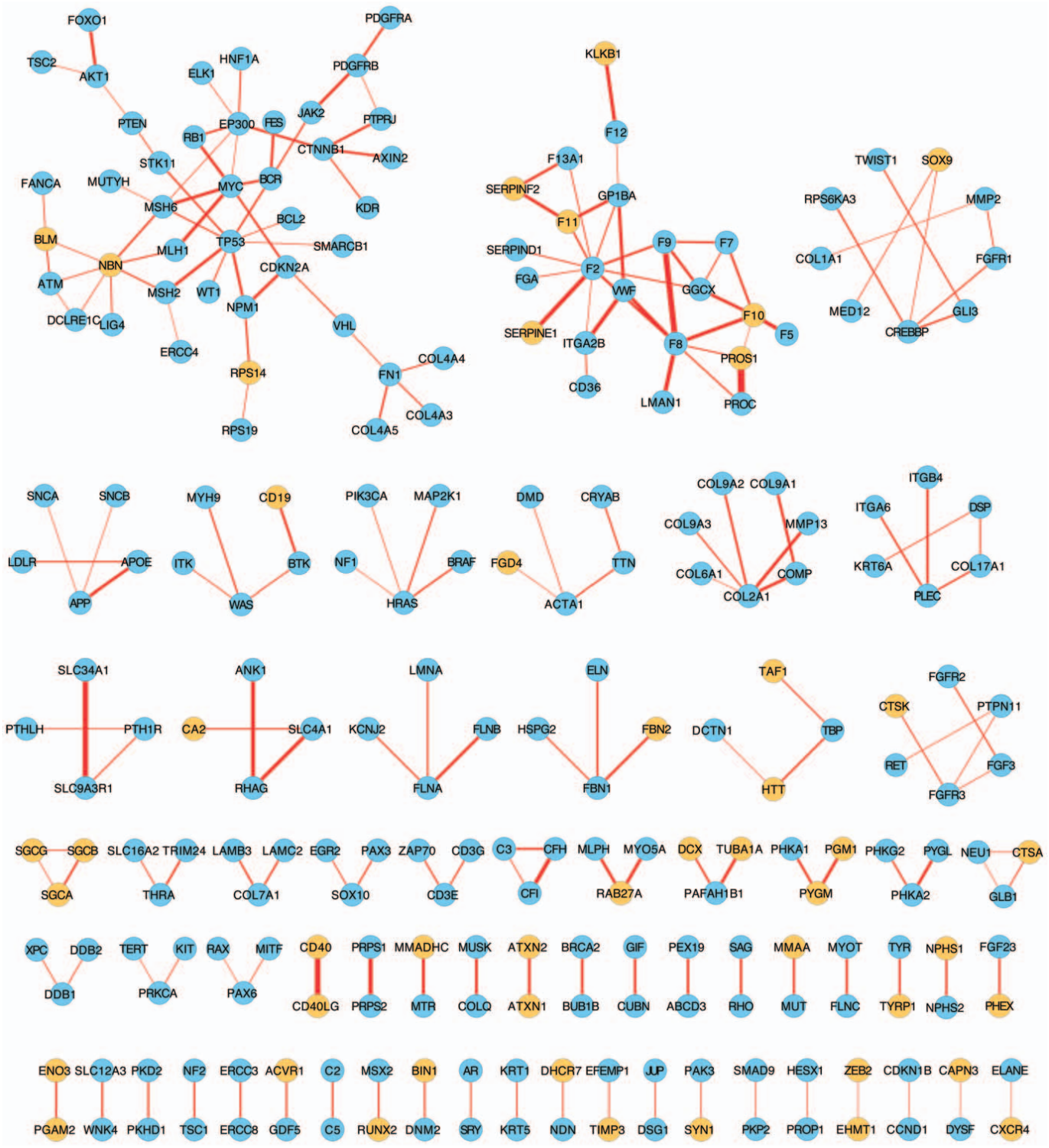


Figure 9. Physical interactions between genes with similar phenotypic lost in the current networks of diseases. This figure is the result of the difference of the resulting intersection between PSGN and PIN after removing those interactions present in HDGN and ODGN. Those genes that are MD-MG in HDN 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.

CHAPTER 3. PUBLICATION 3

Supplementary Material for:

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

Supplementary material INCLUDED in this Thesis:

Methods S1.
doi:10.1371/journal.pone.0056653.s014

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.
doi:10.1371/journal.pone.0056653.s001

Figure S2.
Spearman correlation between the number of pathophenotypes per gene and the average pathophenotypic similarity per gene for PSGN genes.
doi:10.1371/journal.pone.0056653.s002

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.
doi:10.1371/journal.pone.0056653.s003

Table S8.
Spearman correlations between gene degrees in PSGN and HDGN/ODGN.
doi:10.1371/journal.pone.0056653.s011

Table S9.
Spearman correlation between gene degrees in PSGN and biomolecular interactomes.
doi:10.1371/journal.pone.0056653.s012

This supplementary material is NOT INCLUDED in this Thesis but it is available online:

Table S1.
Bipartite and unipartite projections of the updated version of the human diseases network.
doi:10.1371/journal.pone.0056653.s004

Table S2.
Bipartite and unipartite projections of the updated version of the orphan disease network.
doi:10.1371/journal.pone.0056653.s005

Table S3.
Different gene subsets in the human diseases network following proposed classification.
doi:10.1371/journal.pone.0056653.s006

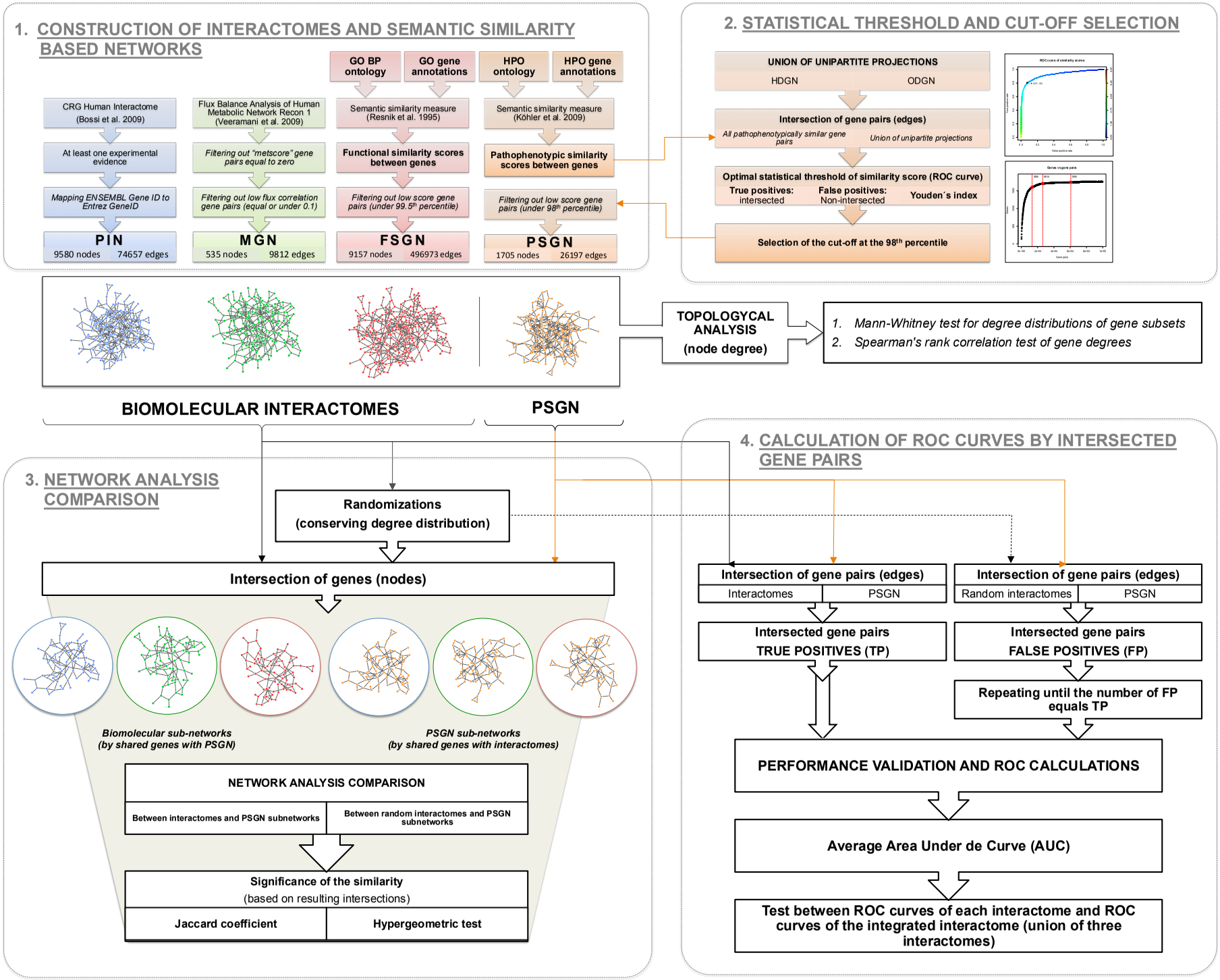
Table S4.
Different gene subsets in the orphan diseases network following proposed classification.
doi:10.1371/journal.pone.0056653.s007

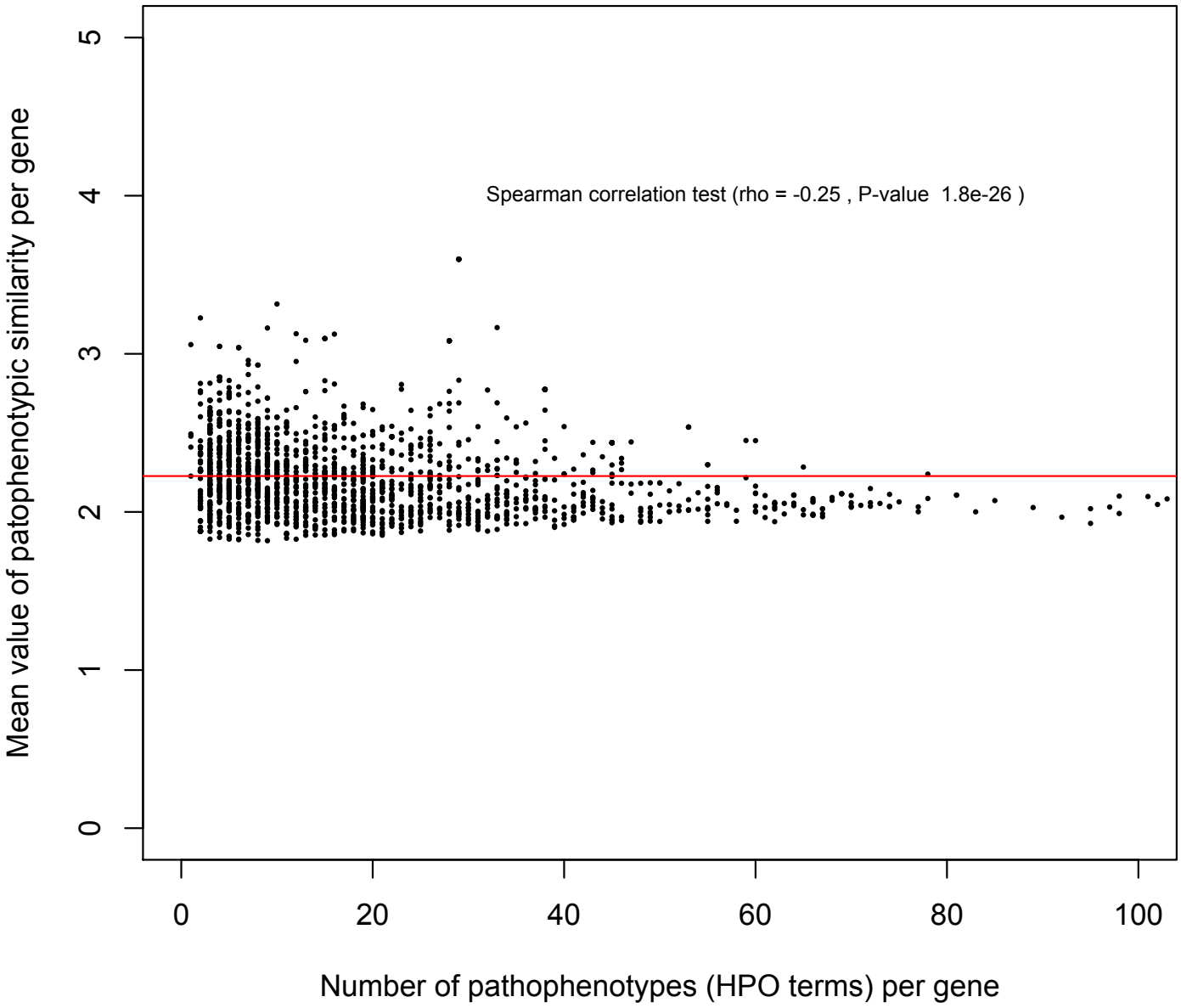
Table S5.
Different biomolecular interactomes based on physical, metabolic and functional interactions.
doi:10.1371/journal.pone.0056653.s008

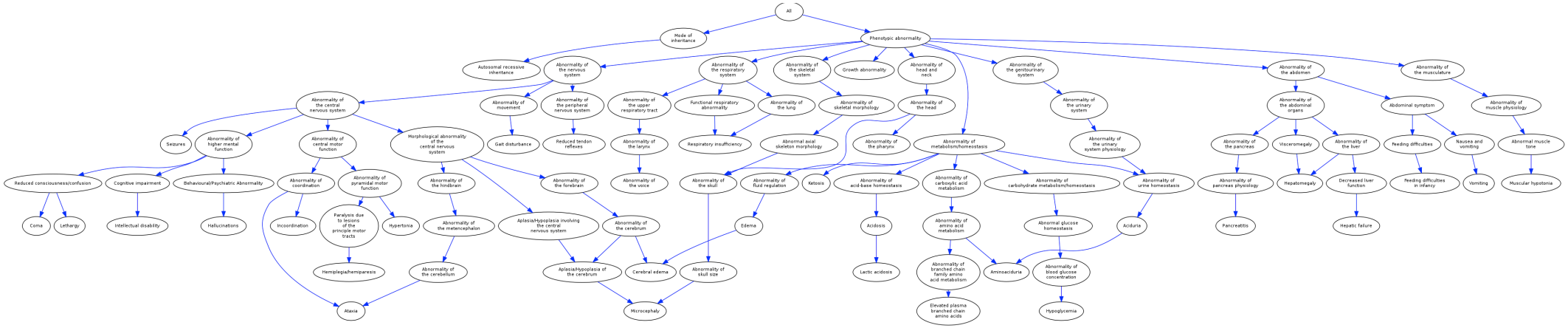
Table S6.
Biomolecular interactome and PSGN sub-networks after nodal intersections.
doi:10.1371/journal.pone.0056653.s009

Table S7.
Pathophenotypic similarity gene network.
doi:10.1371/journal.pone.0056653.s010

Table S10.
Network intersection between PSGN and PIN removing inferred gene-to-gene associations.
doi:10.1371/journal.pone.0056653.s013


1. CONSTRUCTION OF INTERACTOMES AND SEMANTIC SIMILARITY BASED NETWORKS
GO BP ontology
GO gene annotations
HPO ontology
HPO gene annotations
CRG Human Interactome (Bossi et al. 2009)
Flux Balance Analysis of Human Metabolic Network Recon 1 (Veeramani et al. 2009)
Semantic similarity measure (Resnik et al. 1995)
Semantic similarity measure (Köhler et al. 2009)
At least one experimental evidence
Filtering out "metscore" gene pairs equal to zero
Functional similarity scores between genes
Pathophenotypic similarity scores between genes
Mapping ENSEMBL Gene ID to Entrez GeneID
Filtering out low flux correlation gene pairs (equal or under 0.1)
Filtering out low score gene pairs (under 99.5th percentile)
Filtering out low score gene pairs (under 98th percentile)
PIN
9580 nodes 74657 edges
MGN
535 nodes 9812 edges
FSGN
9157 nodes 496973 edges
PSGN
1705 nodes 26197 edges
2. STATISTICAL THRESHOLD AND CUT-OFF SELECTION
UNION OF UNIPARTITE PROJECTIONS
HDGN
ODGN
Intersection of gene pairs (edges)
All pathophenotypically similar gene pairs
Union of unipartite projections
Optimal statistical threshold of similarity score (ROC curve)
True positives: intersected
False positives: Non-intersected
Youden´s index
Selection of the cut-off at the 98th percentile
TOPOLOGYCAL ANALYSIS (node degree)
1. Mann-Whitney test for degree distributions of gene subsets
2. Spearman's rank correlation test of gene degrees
BIOMOLECULAR INTERACTOMES
PSGN
3. NETWORK ANALYSIS COMPARISON
Randomizations (conserving degree distribution)
Intersection of genes (nodes)
Biomolecular sub-networks (by shared genes with PSGN)
PSGN sub-networks (by shared genes with interactomes)
NETWORK ANALYSIS COMPARISON
Between interactomes and PSGN subnetworks
Between random interactomes and PSGN subnetworks
Significance of the similarity (based on resulting intersections)
Jaccard coefficient
Hypergeometric test
4. CALCULATION OF ROC CURVES BY INTERSECTED GENE PAIRS
Intersection of gene pairs (edges)
Interactomes
PSGN
Intersection of gene pairs (edges)
Random interactomes
PSGN
Intersected gene pairs TRUE POSITIVES (TP)
Intersected gene pairs FALSE POSITIVES (FP)
Repeating until the number of FP equals TP
PERFORMANCE VALIDATION AND ROC CALCULATIONS
Average Area Under de Curve (AUC)
Test between ROC curves of each interactome and ROC curves of the integrated interactome (union of three interactomes)


Spearman correlation test (rho = -0.25 , P-value 1.8e-26)
Mean value of patophenotypic similarity per gene
Number of pathophenotypes (HPO terms) per gene
0
1
2
3
4
5
0
20
40
60
80
100


All
Mode of inheritance
Phenotypic abnormality
Autosomal recessive inheritance
Abnormality of the nervous system
Abnormality of the respiratory system
Abnormality of the skeletal system
Growth abnormality
Abnormality of head and neck
Abnormality of the genitourinary system
Abnormality of the musculature
Abnormality of the abdomen
Abnormality of the central nervous system
Abnormality of movement
Abnormality of the peripheral nervous system
Abnormality of the upper respiratory tract
Functional respiratory abnormality
Abnormality of the lung
Abnormality of skeletal morphology
Abnormality of the head
Abnormality of the urinary system
Abnormality of the abdominal organs
Abdominal symptom
Abnormality of muscle physiology
Seizures
Abnormality of higher mental function
Abnormality of central motor function
Morphological abnormality of the central nervous system
Gait disturbance
Reduced tendon reflexes
Abnormality of the larynx
Respiratory insufficiency
Abnormal axial skeleton morphology
Abnormality of the pharynx
Abnormality of metabolism/homeostasis
Abnormality of the urinary system physiology
Abnormality of the pancreas
Visceromegaly
Abnormality of the liver
Feeding difficulties
Nausea and vomiting
Abnormal muscle tone
Reduced consciousness/confusion
Cognitive impairment
Behavioural/Psychiatric Abnormality
Abnormality of coordination
Abnormality of pyramidal motor function
Abnormality of the hindbrain
Abnormality of the forebrain
Abnormality of the voice
Abnormality of the skull
Abnormality of fluid regulation
Ketosis
Abnormality of acid-base homeostasis
Abnormality of carboxylic acid metabolism
Abnormality of carbohydrate metabolism/homeostasis
Abnormality of urine homeostasis
Abnormality of pancreas physiology
Hepatomegaly
Decreased liver function
Feeding difficulties in infancy
Vomiting
Muscular hypotonia
Coma
Lethargy
Intellectual disability
Hallucinations
Incoordination
Paralysis due to lesions of the principle motor tracts
Hypertonia
Abnormality of the metencephalon
Aplasia/Hypoplasia involving the central nervous system
Abnormality of the cerebrum
Edema
Acidosis
Abnormality of amino acid metabolism
Abnormal glucose homeostasis
Aciduria
Pancreatitis
Hepatic failure
Hemiplegia/hemiparesis
Abnormality of the cerebellum
Aplasia/Hypoplasia of the cerebrum
Cerebral edema
Abnormality of skull size
Lactic acidosis
Abnormality of branched chain family amino acid metabolism
Aminoaciduria
Abnormality of blood glucose concentration
Ataxia
Microcephaly
Elevated plasma branched chain amino acids
Hypoglycemia

Supplementary methods and discussion section for

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

Armando Reyes-Palomares[1,2], Rocio Rodríguez-López[1,2], Juan AG 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, E-29071 Málaga, Spain

[2]CIBER de Enfermedades Raras (CIBERER), E-29071 Málaga, Spain

[*]To whom correspondence should be addressed.

Email: medina@uma.es

Procedures to select the similarity score cutoff

1. Introduction

A systematic methodology to identify significant semantic similarities (similarity scores) has not been yet established. We therefore propose a few systematic steps to set aside what can be considered as significant similarity scores, in agreement to the current genetic association and ontological structure knowledge. The semantic similarity proposed by Resnik computes as informative is each term [1]. The information content (also known by IC) of an ontological term depends on its relative frequency, the number of annotations of a term respect to the whole set of annotations (corpus). It means that a high of informativeness indicates more specificity and fewer annotations. Hence, the similarity between terms will be assessed calculating the IC of the most informative common ancestor (MICA) derived from the ontological structure.

We have calculated all similarity scores (pathophenotypic similarities) between annotated genes in the Human Phenotype Ontology (HPO) [2]. For this, we used an IC-based measure [3] resulting in 1309578 similarity scores between 1812 genes, excluding zero scores. In this study, genes are annotated with a set of HPO terms describing the pathological processes related to them by genetic association studies. Accordingly, the computed scores are themselves probabilistic estimations of the pathophenotypic similarity between genes respect to the total number of pathological phenotypes (HPO terms) comprehended in the study.

In this case, the percentile can be considered as a suitable parameter to decide a level of significance for all ranked scores. However, it is assumed that most of the computed pathophenotypic relationships between genes are non-informative although they are statistically recognizable by their similarity score. Furthermore, HPO is a controlled vocabulary that has been manually revised and structured as a direct acyclic graph (DAG) [4]. Consequently, the similarity scores depends on the current knowledge represented in HPO, where some relationships can be missing and general domains cluster many terms. It means that noise could be present in any computed similarity score using biomedical ontologies [5], so the identification of an optimal statistical threshold remains as the greatest difficulty. This is to set the limit of similarity score

from which gene-gene relationships are significantly informative. Therefore, we have two main purposes to identify this optimal statistical threshold and, once it has been assessed, to select the most appropriate cutoff of similarity score.

2. Procedure

We started by analyzing the properties of all the computed pathophenotypic similarities. The similarity score distribution reveals that lower values are clearly more frequent (more probable) than higher ones (Figure 1). Indeed, it is in line with the information content computed by the semantic similarity measurement applied for this work [6], which also uses the relative frequency.

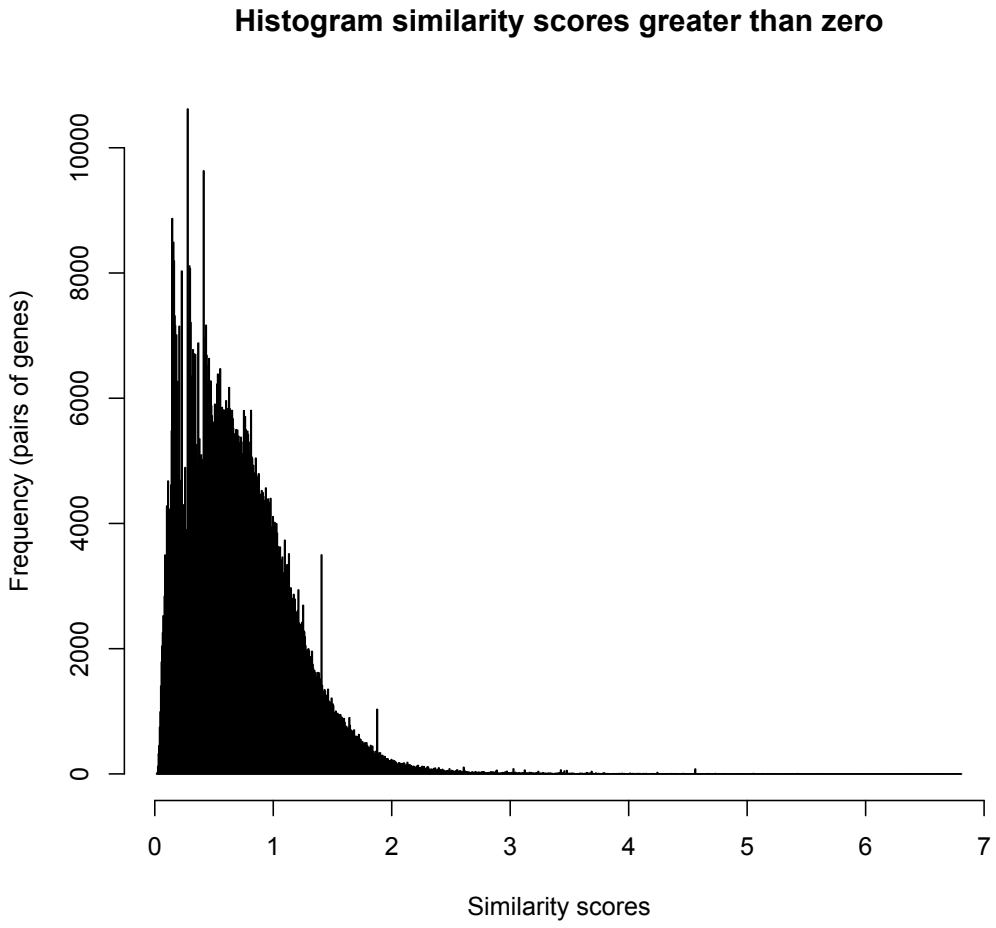


Figure 1. Histogram of gene pairs per similarity score. The number of gene pairs for each similarity score after to remove zero scores. It can be note that a high frequency for bars located the region of lower similarity scores.

Figure 1 represents the number of gene pairs counted for each score. In some cases, similarity scores reach more than 10000 pairs of genes and they are located along the region of smaller values. This may be due to the fact that low similarity scores have been computed by HPO terms close to the root of the ontology conforming general domains (clusters) of non-specific pathophenotypic similarities between genes. Therefore, these domains lack of significant information to become constituent elements of the pathophenotypic gene similarity network (PSGN). These clustering effects can be observed by analyzing the number of resulting genes at different cutoffs among the whole range of similarity scores. Thus we used 1000 different cutoffs to study in details variations on this distribution. For instance, it could be appreciated that the number of participating genes decreases as the similarity score cut-off increases (Figure 2); what indicates that genes are not taken into account if they are not participating in any relationship over the threshold.

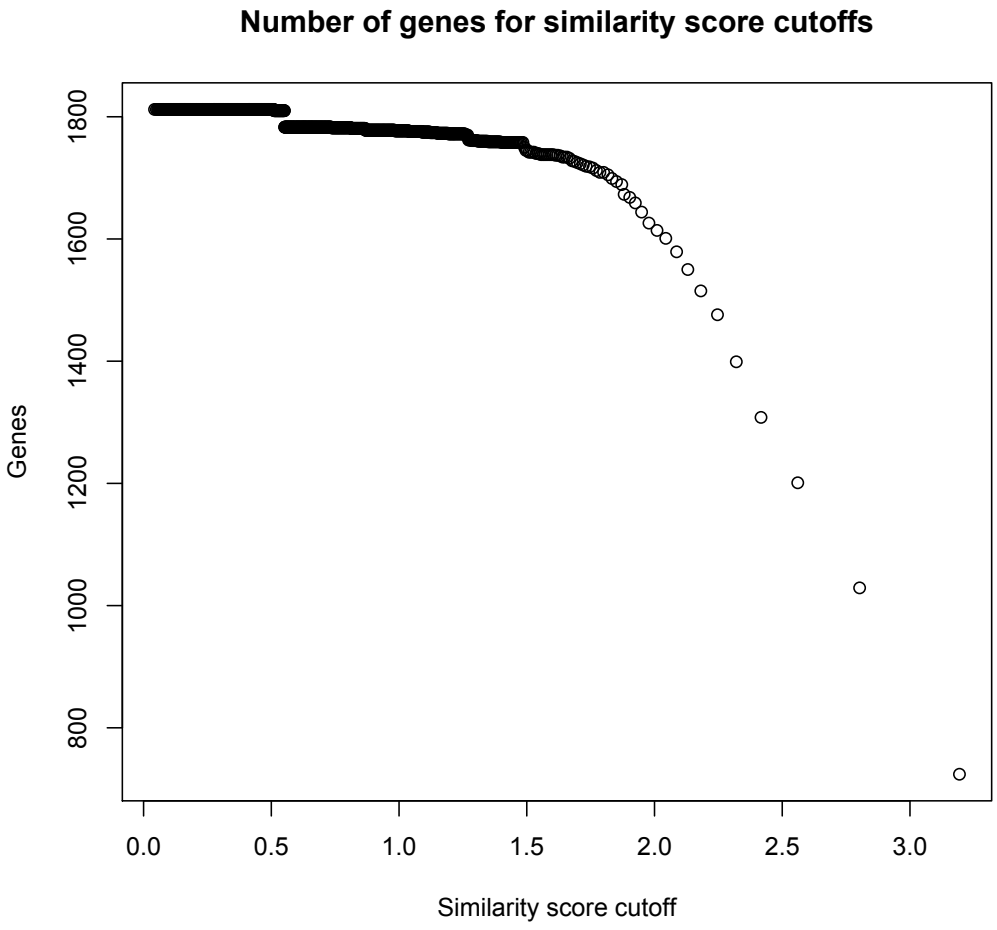


Figure 2. Number of genes counted at each similarity score cutoffs. One thousand equidistant cutoffs were established along the whole range of similarity scores. Subsequently, gene pairs with a score equal or greater than the used cutoff were selected and resulting genes were counted.

As can be noted, the clustering effects are more evident for higher similarity scores where little variations lead to introduce strong differences in the number of genes. In contrast, the number of genes is well conserved when using low similarity scores as cutoffs. However, along this flat area, we observe abrupt shifts in the number of genes, whose can be due to discard scores associated with HPO terms grouping many genes (group of genes sharing HPO annotations). The number of genes is stable but the gene pairs decreased exponentially until it reaches scores higher than approximately 1.5 (Figure 3).

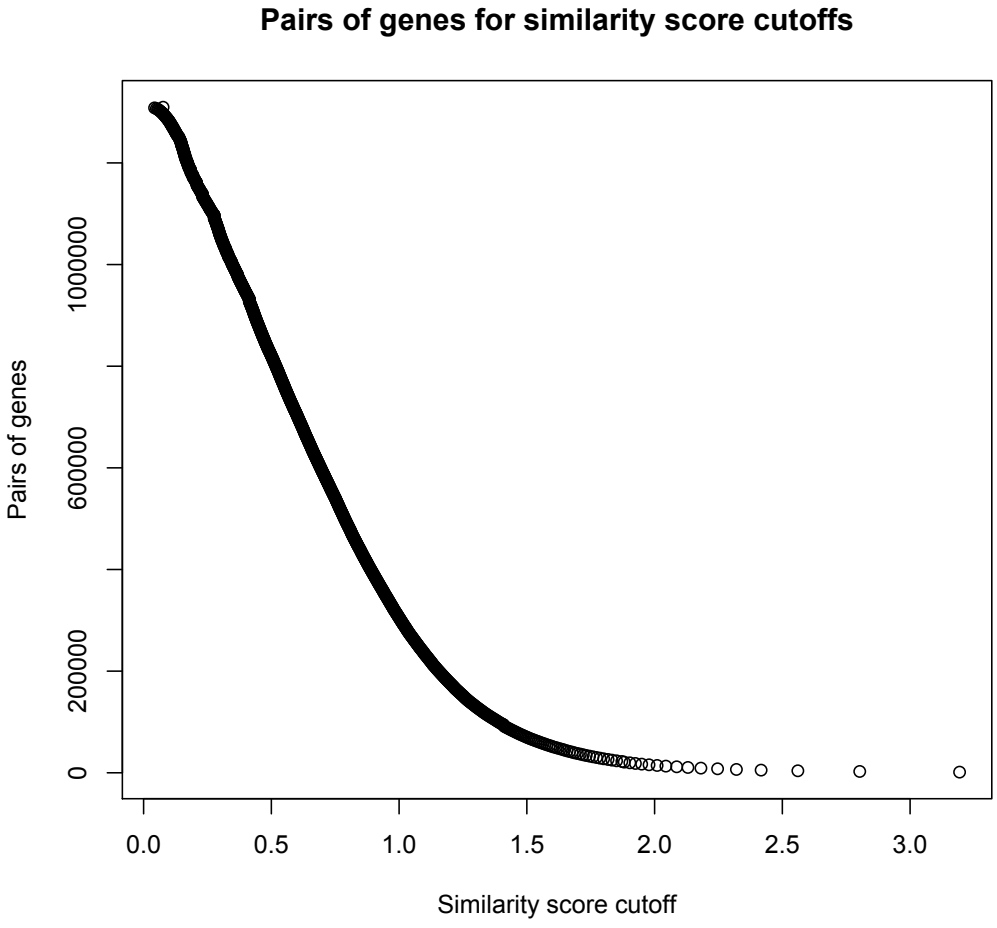


Figure 3. Number of gene pairs for similarity scores cutoff. One thousand equidistant cutoffs were established along the whole range of similarity scores. All gene pairs with a score equal or greater than the used cutoff were selected. Three types of areas can be approximately defined: vertical asymptote (approx. from 0 to 1.0 cutoffs), turning point of the curvature (approx. from 1.0 to 2.0 cutoffs) and horizontal asymptote (approx. from 2.0 to 1.0 cutoffs).

Figure 3 represents a curve with three different zones: a vertical asymptote, a curvature and a horizontal asymptote. The vertical asymptote represents a clear lineal dependency between the used cutoff and the number of gene pairs. Then, many gene-gene relationships are affected by minor variations in the similarity score during the analysis. This result can be partially explained by hierarchical dependencies between HPO terms in the ontology and suggests the exponential growth of nonspecific similarities as the score drops. On the other hand, the horizontal asymptote seems to indicate that the number of gene pairs is conserved when the cutoff has reached a certain score. But what really happens is that the specificity increases considerably with higher values of semantic similarity, what means that genes are sharply clustered. Hence, the curvature represents the threshold to distinguish from low to high signal-to-noise ratio, in the vertical and horizontal asymptote, respectively (Figure 3).

3. Results for threshold and cutoff selection

Different approaches were used to assess the biological relevance of this threshold in order to maximize noise-reduction and specificity. Subsequently, this optimal threshold should be the reference value from which to set the most appropriated cut-off to build the PSGN.

First off all, we performance a binary classifier system and illustrate it in a receiver operating characteristic (ROC) curve. This binary system was built from gene pairs showing pathophenotypic similarity (1309578 gene pairs) that were present or absent in the union of both unipartite projections (HDGN and ODGN). We joined all inferred interactions from HDGN and ODGN considering as unique the redundant ones. It should be pointed, that both unipartite projections (HDGN or ODGN) are based on genes that are associated with at least one same disease in OMIM and Orphanet, respectively. The resulting network, in turn, was compared with the dataset of pairs of genes obtained after to compute similarity scores between genes. It resulted in a cross-tabulation of overlapped and non-overlapped gene pairs to be considered respectively as true positives and false positives (Table 1).

Table 1. Binary classification based on the intersection of gene pairs

	Genes	Gene pairs
Pathophenotypic similarities	1812	1309578
Unipartite projections (HDGN + ODGN)	1760	8372
Overlapped or True Positives (TP)	1064	3271
Non-overlapped or False Positives (FP)	748	1306307

This classification model has been used to predict the optimal threshold of pathophenotypic similarity score involving a pair of genes on the same pathological process (Figure 4).

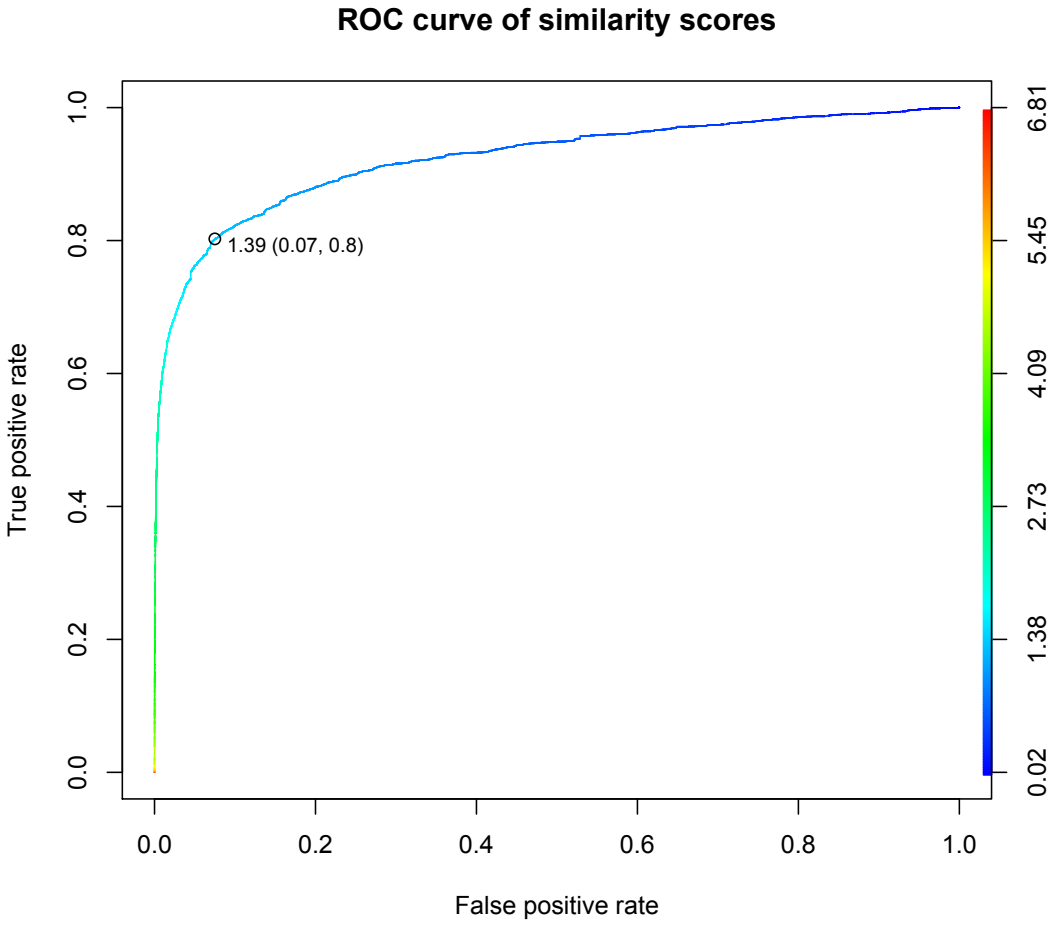


Figure 4. ROC curve and optimal threshold of pathophenotypic similarity. In this performance analysis, those pairs of genes showing pathophenotpic similarity and co-associated with at least one diseases are considered as true positives and the rest of gene pairs showing pathophenotypic similarity are considered as false positive. Black circle indicates the coordinates (0.07 and 0.8 for false and true positive rates, respectively) for the Youden's index (1.39) that determines the optimal threshold. The color palette maps for each threshold the corresponding similarity scores. ROCR [7] and pROC [8] packages were used to represent and calculate optimal threshold.

As can be appreciated from the ROC curve the pathophenotypic similarity (similarity score) is a good indicator to assess the underlying relationships of genes in a particular pathological process. The optimal threshold of similarity score that maximizes the trade-off between the rates of true and false positives is 1.39 (which corresponds to the 92th percentile, Table 2) for the Youden's index [9,10] (Figure 4). We also calculated the inflection point at the curvature in Figure 3, which can be worth as an alternative approach to locate the turning point. We fit the curve of computed similarity scores to an exponential function and the inflection point results in 1.55 corresponding to the 95th percentile (Table 2). This curve-fitting analysis was carried out in MATLAB.

Table 2. Results using top score percentiles as cutoffs

Percentile	Score	Genes	Gene pairs
92th	1.39[a]	1759	99777
95th	1.55[b]	1742	60742
98th	**1.82**	**1705**	**26197**
99th	2.04	1601	13098

[a] Patho-phenotypically meaningful threshold by ROC curve analysis.

[b] Turning point fitting similarity scores to an exponential function.

Therefore, we have two different thresholds to estimate the appropriate cutoff to be used to build the pathophenotypic similarity gene network. However, the 92th and 95th percentiles are located in regions under the influence of large clusters of non-specific similarities. For instance, we observe that similarities with scores below 1.5 show abrupt changes in the number of genes (as can be seen in Figure 2). Thus, we carried out a more detailed analysis for these regions by filtering all similarity scores below 1.39, which could be considered as the patho-phenotypically meaningful threshold (Figure 4).

In particular, the number of genes begins to drop around the value of 20.000 computed similarities scores (Figure 5). In this case, the score in the 98th percentile (1.8179) represents a reference value from which to ensure high specificity for pathophenotypic similarities without losing information. Therefore, 1.8179 corresponds to the lowest similarity value at the top 2% of all ranked scores and the most appropriate cutoff to build PSGN, for the reasons discussed above.

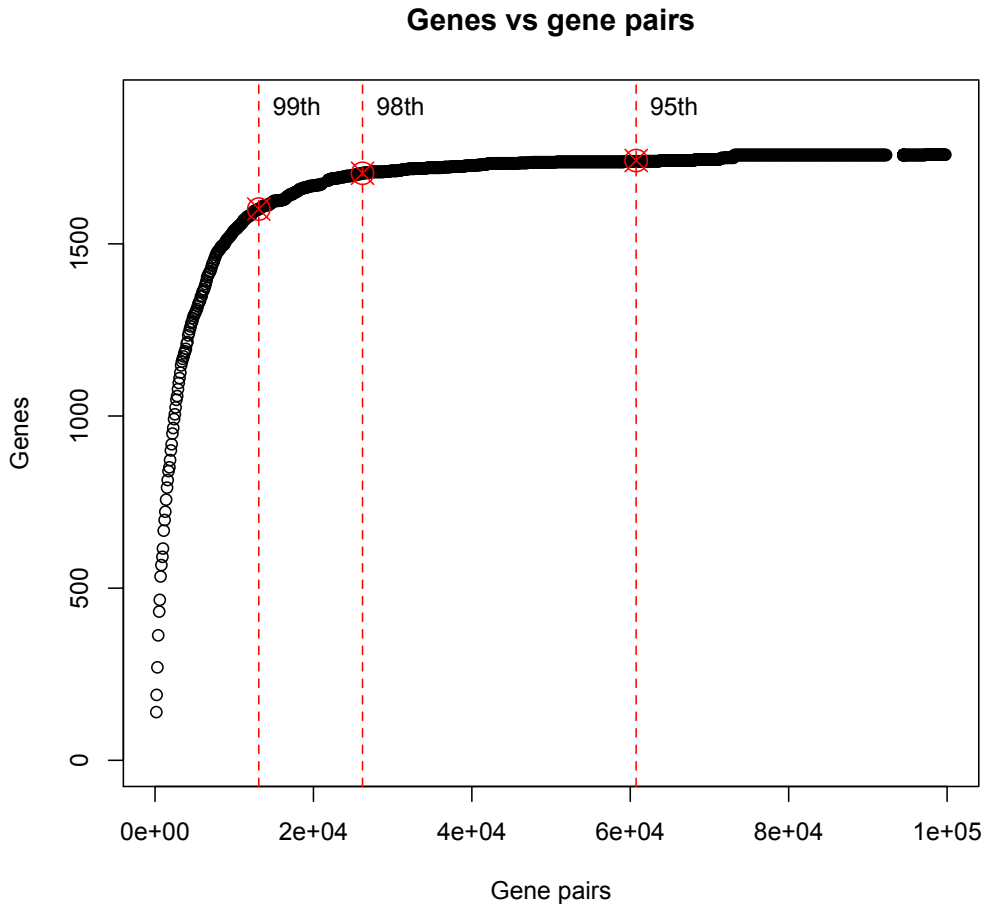


Figure 5. The number of genes and gene pairs for each similarity score cutoffs. One thousand equidistant cutoffs were established along the range of similarity scores from the 92th percentile, it means removing scores below 1.39. All gene pairs with a score equal or greater than the cutoff were selected and resulting genes were counted. Red circles and vertical lines indicate exact location for 95th, 98th and 99th percentiles.

4. References

1. Resnik P (1995) Using Information Content to Evaluate Semantic Similarity in a Taxonomy. IJCAI. pp. 448–453.

2. Robinson PN, Mundlos S (2010) The Human Phenotype Ontology. Clin Genet 77: 525–534.

3. Köhler S, Doelken SC, Rath A, Aymé S, Robinson PN (2012) Ontological phenotype standards for neurogenetics. Hum Mutat 33:1333-1339

4. Robinson PN (2012) Deep phenotyping for precision medicine. Hum Mutat 33: 777–780.

5. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5: 12.

6. 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.

7. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940–3941.

8. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77.

9. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3: 32–35.

10. Perkins NJ, Schisterman EF (2006) The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 163: 670–675.

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

	HDGN			ODGN		
Class	**n (genes)**	**r_s**	**P-value**	**n (genes)**	**r_s**	**P-value**
MD-MG[a]	-	-	-	-	-	-
MD-PG	247	0.06	3.4E-01	280	0.11	6.6E-02
PD-MG	226	0.26	7.3E-05	446	0.25	7.4E-08
PD-PG[b]	303	0.06	3.2E-01	485	0.13	4.2E-03
All genes[c]	528	0.17	9.2E-05	931	0.22	1.0E-11

[a] Biunivocal genes are not present in diseases causing gene networks.
[b] Pleiotropic genes associated with at least one polygenic diseases.
[c] All intersected genes between PSGN and HDGN or ODGN respectively.

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

		PIN			MGN			FGN		
Network	**Class**	**n (genes)**	r_s	**P-value**	**n (genes)**	r_s	**P-value**	**n (genes)**	r_s	**P-value**
HDN	MD-MG	555	0.13	2.9E-03	87	-0.01	9.4E-01	634	0.01	7.7E-01
HDN	MD-PG	420	0.18	2.6E-04	31	-0.02	9.3E-01	456	-0.04	4.0E-01
HDN	PD-MG	158	0.06	4.5E-01	29	0.37	5.0E-02	175	-0.06	4.3E-01
HDN	PD-PG[b]	247	0.10	1.0E-01	14	-0.25	3.9E-01	267	-0.12	6.0E-02
HDN	All genes in HDN[c]	1175	0.15	2.0E-07	150	0.10	2.4E-01	1313	0.01	6.5E-01
ODN	MD-MG	364	0.11	3.0E-02	68	0.02	8.7E-01	411	0.01	9.3E-01
ODN	MD-PG	312	0.21	2.2E-04	26	-0.23	2.7E-01	340	-0.10	6.3E-02
ODN	PD-MG	322	0.18	1.3E-03	40	0.16	3.4E-01	354	-0.04	4.3E-01
ODN	PD-PG[b]	384	0.13	1.4E-02	28	0.01	9.5E-01	423	-0.01	9.3E-01
ODN	All genes in ODN[c]	1151	0.16	3.2E-08	149	0.07	3.9E-01	1279	0.01	5.9E-01

[a] Network based on functional similarities from biological processes branch of Gene Ontology.
[b] Pleiotropic genes associated with at least one polygenic diseases.
[c] All intersected genes between the respective biomolecular interactome and PSGN.

CHAPTER 3. PUBLICATION 4

PhenUMA: a Tool for Integrating the Biomedical Relationships among Genes and Diseases

Rocío Rodríguez-López, **Armando Reyes-Palomares**, Francisca Sánchez Jiménez and Miguel Angel Medina

BMC Genomics

Print ISSN:

Online ISSN: 1471-2164

Supplementary Material: Yes

Status: Submitted

DOI:

CHAPTER 3. PUBLICATION 4

Supplementary Material for:

PhenUMA: a Tool for Integrating the Biomedical Relationships among Genes and Diseases

Supplementary material INCLUDED in this Thesis:

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.

CHAPTER 3. PUBLICATION 5

Network Medicine Approaches for Systematic Identification of Phenotype and Structural Variants Associations

Armando Reyes-Palomares, *et al.*

Print ISSN:

Online ISSN:

Supplementary Material:

Status: **Manuscript in preparation**

DOI:

CHAPTER 4
DISCUSSION

Functional modularity in biological systems

The modularity is an essential property to study the architecture of the functionality in biological systems, at whatever of their scale. This is one of the main arguments to evidence that genetic information is not the unique level of causality[86], but there are many other biological scales determining the function. In this regard, systems biology approaches aim to take into account the dynamical singularities to understand how biological systems function[4,5].

In the research group where this Thesis has been carried out, I have studied the modular and functional behaviour of the metabolism using a particular case of the sulfur amino acids. Our previous work modelling polyamine metabolism in mammals[87] suggested an unexpected relevant role of S-adenosyl methionine (SAM) in the control of polyamines levels. To evaluate it further, we decided that the first task within this Doctoral Thesis work would be the design of a metabolic model integrating those metabolic modules that are linked by SAM, such as the polyamine[87] and methionine metabolism[88]. In addition, we also included the folate[89,90] and glutathione metabolism[91] for an extended view of the regulatory processes. Time-course simulations of our model suggest a relevant role of SAM in polyamines homeostasis. In proliferative conditions, MAT-II is expressed. This enzyme is inhibited by its own product (SAM); which cellular levels are decreased. SAM and ornithine are the immediate precursors of polyamines. Since polyamines are necessary for proliferation, their levels should be maintained or even increased under proliferative conditions. Accordingly, we proposed alternative regulatory mechanisms in polyamine metabolism that depend on SAM availability. *In silico* experiments with our model under proliferative conditions indicate that the

metabolic flux redistribution to balance polyamine levels when levels of SAM are diminished could be explained by an increased activity of ornithine decarboxylase (ODC). These predictions have been experimentally validated as can be deduced from the conclusions of the recent work published by Lu, Mato and co-workers[92]. This computational work illustrates how the metabolic modularity, in the case of amine metabolism, is an operative feature of the structure of the metabolic network. Additionally, this model is also useful to raise novel hypotheses that will require experimental validation to amplify the relative information about complex biological systems.

Evaluating disease modularity

All the current hypotheses agree that the origin of biological modularity and the subsequent performance require genetic variations[19,22,23]. In this case, the genetic level is the unique level of reference to track the underlying evolutionary mechanism[18]. In the present section, I discuss the different perspectives about the relationships between genotypic and phenotypic variability. Both are measurable features that can show a different degree of "resistance" (robustness) to the change[26,29]. In contrast, living beings (as complex adaptive systems) require plasticity to face the constant fluctuations that occur in nature. Therefore, they need to be constantly changing their features to ensure their survival or –even– to improve their fitness (evolvability).

The complexity of the genotype-phenotype relationships is multidimensional because of the presence of multiples genotypes associated with the same phenotype, as well as the existence of different phenotypic traits associated with the same phenotype. For this reason, the one gene-one enzyme hypothesis should be considered more as an exception than the rule in biology. In this line, there is a very interesting debate about the pleiotropy[45,47,93], defined as the multiple traits that can be associated with a genotype or a gene. This debate discusses actively if there is a universal or restricted pleiotropy, what means that every genotype has an effect on all the possible measurable traits[43] or there is a modular pleiotropy where the effects of a genotype are restricted to a reduced number of traits[45]. This debate implicates to dig into the biological meaningful role of genes (or genotypes) in fitness but also for the expression of their related characters. For this purpose, the objective is to search for novel measurements of pleiotropy to evaluate traits more directly related to the biological function of genes (or a genotype). These phenotypes would be likely the reflection of the molecular functions that gene products may carry out into cells. In this sense, the relevance of the functions carried out by genes is the more appropriate approach to assess their role the different modules they take part in.

As mentioned in the Introduction, the phenome is considered one of the 'omes that may provide new insights into science. But I want to be even more explicit: the deep phenotyping should be referred to high-throughput screening of cellular and molecular phenotypes in multivariate cellular states to get better correlations between traits and genetic variations. For instance, Andreas Wagner showed in a recent publication the potential benefits to study direct molecular phenotypes, such as transcription binding affinity, together with DNA variability[38].

As far as possible with the current biomedical information available, further studies are necessary to trace which cellular and molecular mechanisms are underlying pathological phenotypes are necessary. For this purpose, our first issue was to define diseases in suitable terms to approach them as far as possible to their molecular and biological context. For that, we built and analysed the *human pathophenome gene network* [94] and this network was compared to the previously published disease networks called HDN[95] and ODN[96]. Unlike these previous disease networks, the pathophenome uses semantic similarities. This pathophenotypic similarity was calculated for pairs of genes using their phenotypic annotations in the "Human Phenotype Ontology" and, subsequently, comparing their phenotypic spaces. The resulting human pathophenome network contains 1706 genes (nodes) and 26192 significant pathophenotypic similarities (edges). This network reveals a strong re-arrangement of the pathological relationships among genes and, moreover, they are measurable by phenotypic similarities. Many novel pathophenotypic interactions between genes have been uncovered. 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 might contribute to discover pre-clinical stages and to identify underlying co-dependencies among disease-causing genes that are useful to describe disease modularity.

Standardization efforts

An integrative framework for metabolic modelling

We developed a user-friendly application, named as Systems Biology Metabolic Modelling Assistant (SBMM Assistant, http://www.sbmm.uma.es/) able to integrate kinetic and metabolic information of any organism. SBMM assistant, works on the bases of an ontology-based mediator developed to integrate data

from KEGG, CHEBI, BRENDA and SABIORK. SBMM assistant is characterized by the following features: It is an SBML-compatible and friendly tool able to guide to the novel or experienced user to capture, enrich, generate and visualize biological networks, to make basic queries on enzymatic kinetics and regulation, and to annotate this information following the MIRIAM specifications.

In this sense SBMM assistant is an example of such a specific application, having the aim to act as a complementary tool (assistant) for metabolic modelling analysis programs. Thus, it is not only able to capture, enrich and store information in models, but also helps the cross-talk among the different resources and tools in a friendly way.

An integrative framework for biomedical information

Within this Doctoral Thesis, we present PhenUMA (http://www.phenuma.uma.es/), a framework for the integrative analysis of biomedical information that can help with the discovery of alternative pathological roles of genes, biological processes and phenotypes. PhenUMA knowledge base includes pairwise relationships that result from: (i) repositories of genetic association studies (OMIM and Orphanet), protein-protein interactions (STRING) and metabolic interactions; (ii) network inferences in known relationships; and (iii) semantic similarity measurements using biomedical ontologies (Gene Ontology and Human Phenotype Ontology) for functional and phenotypic similarity, respectively. The workflow begins by building "seed networks", which are used as backbones and extended with phenotypic and functional associations, and can be later analysed to look for phenotypic or functional enrichment. We also report a systematic method to set the optimal threshold of phenotypic similarity that is suitable to detect meaningful relationships using association indices (i.e. jaccard index or hypergeometric test) and performance validation[94].

The main advantage of PhenUMA over other systems is to unify network, semantic similarity and enrichment analysis in the same platform. This will allow users to manage vast amount of a priori unconnected (or at least unreachable in an easy way) biomedical data. This framework is useful to evaluate phenotypic similarities between functionally related genes and clusters of genes or medical conditions sharing specific clinical features. Furthermore, PhenUMA aids the evaluation of the enrichment of phenotypes or biological attributes in reported results. Clusters of phenotypically related diseases are more coherent in PhenUMA compared to other similar resources.

PhenUMA represents an advance towards the use of new technologies for genomics and personalized medicine. In a near future update, we will include phenotype-genotype association data and more repositories of curated molecular interactions. Since PhenUMA is under continuous development, user comments and suggestions are welcome.

Network medicine approaches using patient data

The combined use of network analysis methods together genomic association studies for biomedical applications emerged as a new field, named as Network Medicine[49,97]. This is a promising approach, specially now, that the scientific community is assessing the main standards necessary for the expected increasing genome sequencing projects for the next years[72,98]. In this Thesis, I present a preliminary study comprising 4627 unbalanced CNVs in 3315 patients (cases) from a heterogeneous group of disorders showing developmental delay, intellectual disability and congenital malformations from DECIPHER Database[78,99]. Our aim has been to use individual clinical features to compare genotypic and phenotypic relationships among these patients. In particular, we present a combined analysis of network-based approaches with genetic association studies. To this end, we first have built a network where vertices symbolize patients and edges represent that both patients show both genotypic and phenotypic relationship. Subsequently, we use phenotypic enrichment analysis and clustering methods (using clique percolation method) to identify locus-phenotypic enrichment and novel phenotype-genotype associations, for this latter using a control data set of reference. This workflow allows us to identify more than 500 different genotype-phenotype associations. Finally, we use our results to build a high-resolution genomic map of phenotypes associations with overlapping CNVs between patients. This study illustrates how network medicine approaches are very helpful to characterize very low prevalent disorders as those included analysed in this case data set. The use of integrative analysis of large data sets provides clinicians and researchers both depth and wider interpretation of patient profiles. This works evidences the need to advance in consolidated standards and public repositories of genomic and health data for the advancement of genomic medicine.

CHAPTER 5
CONCLUSIONS

Conclusions from Objective 1

An integrated model of the metabolism of polyamines and sulfur amino acids has been build from previously published mathematical models of its constituent modules. The model allows us to evaluate and predict the S-adenosyl methionine availability in different physiological conditions. The predictions of this model have been subsequently confirmed by independent experimental studies.

Conclusions from Objective 2

We have developed a workflow for integrating metabolic and kinetic data in a user-friendly tool to provide access to this information in a unique platform. This tool is helpful in the development of similar projects related to objective 1.

Conclusions from Objective 3

We have shown how semantic similarity based on pathological phenotypes is a useful resource to construct gene networks that allow us to discover the molecular mechanism underlying the development of genetic diseases. Phenotypic similarity enables to identify and evaluate the distinct components from the same metabolic module, as in the case of phenotypically similar genes to those associated with Maple Syrup Urine Disease.

Conclusions from Objective 4

We have built an integrative framework for biomedical and biological information that it is useful to study phenotypic and functional relationships between genes, as well as to cluster disease that shared similar phenotypes.

Conclusions from Objective 5

The combination of network-based analysis with genetic association studies improves the systematic identification of significant relationships between genomic regions associated with pathological phenotypes from a heterogeneous group of patients with low prevalence diseases.

CHAPTER 5
CONCLUSIONES

Conclusiones derivadas del objetivo 1

Se ha construido un modelo integrado del metabolismo de las poliaminas y de los aminoácidos azufrados a partir de modelos matemáticos de sus módulos constituyentes previamente publicados. El modelo permite estudiar y predecir la disponibilidad de S-adenosil metionina ante distintas condiciones fisiológicas. Las predicciones de este modelo han sido posteriormente confirmadas en estudios experimentales independientes.

Conclusiones derivadas del objetivo 2.

Se ha desarrollado un "workflow" para integrar información metabólica y cinética en una herramienta de fácil manejo para que los usuarios puedan acceder a esa información desde una única plataforma. Esta herramienta ayuda al desarrollo de proyectos similares al relacionado con el objetivo 1.

Conclusiones derivadas del objetivo 3

Hemos mostrado cómo la similitud semántica basada en los fenotipos patológicos es un recurso útil para construir redes de genes que permitan descubrir los mecanismos moleculares implicados en el desarrollo de las enfermedades genéticas. La similitud fenotípica posibilita identificar y estudiar los distintos componentes de un mismo módulo metabólico, como es el caso de los genes fenotípicamente similares a los relacionados con la enfermedad del jarabe de arce.

Conclusiones derivadas del objetivo 4

Se ha construido un "framework" para integrar información biomédica y biológica que es útil para estudiar relaciones fenotípicas y funcionales entre genes, así como para agrupar enfermedades que comparten fenotipos similares.

Conclusiones derivadas del objetivo 5

La combinación del análisis de redes y estudios de asociación genética ha permitido identificar de forma sistemática relaciones significativas entre regiones genómicas asociadas a fenotipos específicos en un grupo heterogéneo de pacientes con enfermedades de baja prevalencia.

REFERENCES

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).

2. Lauffenburger, D. A. Cell signaling pathways as control modules: Complexity for simplicity? *Proc Natl Acad Sci* **97,** 5031–5033 (2000).

3. Kitano, H. Perspectives on systems biology. *New Gener Comput* **18,** 199–216 (2000).

4. Kitano, H. Systems biology: a brief overview. *Science (80-)* **295,** 1662–1664 (2002).

5. Kitano, H. Computational systems biology. *Nature* **420,** 206–10 (2002).

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

7. Wolkenhauer, O., Auffray, C., Jaster, R., Steinhoff, G. & Dammann, O. The road from systems biology to systems medicine. *Pediatr Res* **73,** 502–7 (2013).

8. Von Bertalanffy, L. *General System Theory: Foundations, Development, Applications (Revised Edition)*. (George Braziller Inc., 1968).

9. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* **117,** 500–544 (1952).

10. Noble, D. A modification of the Hodgkin--Huxley equations applicable to Purkinje fibre action and pacemaker potentials. *J Physiol* **160,** 317–352 (1962).

11. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2,** 343–372 (2001).

12. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu Rev Cell Dev Biol* **26,** 721–44 (2010).

13. Cassman, M. Barriers to progress in systems biology. *Nature* **438,** 1079 (2005).

14. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *Trends Microbiol* **15,** 45–50 (2007).

15. Noble, D. *The Music of Life: Biology Beyond Genes [Paperback]*. 176 (Oxford University Press, 2006).

16. Noble, D. Claude Bernard, the first systems biologist, and the future of physiology. *Exp Physiol* **93,** 16–26 (2008).

17. Medina, M. Á. Systems biology for molecular life sciences and its impact in biomedicine. *Cell Mol Life Sci* 1–19 (2012). doi:10.1007/s00018-012-1109-z

18. Wagner, G. P., Pavlicev, M. & Cheverud, J. M. The road to modularity. *Nat Rev Genet* **8,** 921–31 (2007).

19. Sole, R. V. & Fernandez, P. Modularity "for free" in genome architecture? (2003).

20. Solé, R. V & Valverde, S. Spontaneous emergence of modularity in cellular networks. *J R Soc Interface* **5,** 129–33 (2008).

21. Kashtan, N., Parter, M., Dekel, E., Mayo, A. E. & Alon, U. Extinctions in heterogeneous environments and the evolution of modularity. *Evolution* **63,** 1964–75 (2009).

22. Kashtan, N. & Alon, U. Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A* **102,** 13773–8 (2005).

23. Clune, J., Mouret, J.-B. & Lipson, H. The evolutionary origins of modularity. *Proc Biol Sci* **280,** 20122863 (2013).

24. Guimerà, R., Arenas, A. & Díaz-Guilera, A. Communication and optimal hierarchical networks. *Phys A Stat Mech its Appl* **299,** 247–252 (2001).

25. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297,** 1551–5 (2002).

26. Stelling, J., Sauer, U., Szallasi, Z., Doyle, F. J. & Doyle, J. Robustness of cellular functions. *Cell* **118,** 675–85 (2004).

27. Corominas-Murtra, B., Goñi, J., Solé, R. V & Rodríguez-Caso, C. On the origins of hierarchy in complex networks. *Proc Natl Acad Sci U S A* **110,** 13316–21 (2013).

28. Wagner, A. Robustness and evolvability: a paradox resolved. *Proc Biol Sci* **275,** 91–100 (2008).

29. Kitano, H. Biological robustness. *Nat Rev Genet* **5,** 826–37 (2004).

30. Schaefer, C., Schlessinger, A. & Rost, B. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. *Bioinformatics* **26,** 625–31 (2010).

31. Rorick, M. M. & Wagner, G. P. Protein structural modularity and robustness are associated with evolvability. *Genome Biol Evol* **3,** 456–75 (2011).

32. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5,** 101–13 (2004).

33. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2,** 2006.0008 (2006).

34. Draghi, J. A., Parsons, T. L., Wagner, G. P. & Plotkin, J. B. Mutational robustness can facilitate adaptation. *Nature* **463,** 353–5 (2010).

35. Kirschner, M. & Gerhart, J. Evolvability. *Proc Natl Acad Sci* **95,** 8420–8427 (1998).

36. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* **9,** 965–974 (2008).

37. Wagner, A. The role of robustness in phenotypic adaptation and innovation. *Proc Biol Sci* **279,** 1249–58 (2012).

38. Payne, J. L. & Wagner, A. The Robustness and Evolvability of Transcription Factor Binding Sites. *Science (80-)* **343,** 875–877 (2014).

39. Ciliberti, S., Martin, O. C. & Wagner, A. Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A* **104,** 13591–6 (2007).

40. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett* **579,** 1772–8 (2005).

41. Stearns, F. W. One hundred years of pleiotropy: a retrospective. *Genetics* **186,** 767–73 (2010).

42. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends Genet* **29,** 66–73 (2013).

43. Zhang, J. & Wagner, G. P. On the definition and measurement of pleiotropy. *Trends Genet* **29,** 383–4 (2013).

44. He, X. & Zhang, J. Toward a molecular understanding of pleiotropy. *Genetics* **173,** 1885–91 (2006).

45. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12,** 204–213 (2011).

46. Hill, W. G. & Zhang, X.-S. Assessing pleiotropy and its evolutionary consequences: pleiotropy is not necessarily limited, nor need it hinder the evolution of complexity. *Nat Rev Genet* **13,** 296; author reply 296 (2012).

47. Wagner, G. P. & Zhang, J. Universal pleiotropy is not a valid null hypothesis: reply to Hill and Zhang. *Nat Rev Genet* **13,** 296 (2012).

48. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat Rev Genet* **14,** 168–78 (2013).

49. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12,** 56–68 (2011).

50. Hidalgo, C. A., Blumm, N., Barabási, A.-L. & Christakis, N. A. A Dynamic Network Approach for the Study of Human Phenotypes. *PLoS Comput Biol* **5,** e1000353 (2009).

51. Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity . *Proc Natl Acad Sci* **105 ,** 9880–9885 (2008).

52. Ibáñez, K., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A. & Valencia, A. Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers Detected by Transcriptomic Meta-analyses. *PLoS Genet* **10,** e1004173 (2014).

53. Park, J., Lee, D.-S., Christakis, N. A. & Barabasi, A.-L. The impact of cellular networks on disease comorbidity. *Mol Syst Biol* **5,** (2009).

54. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31,** 1102–10 (2013).

55. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization . *Bioinforma* **27 ,** 431–432 (2011).

56. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat Methods* **9,** 1069–76 (2012).

57. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435,** 814–8 (2005).

58. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41,** W77–83 (2013).

59. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4,** 44–57 (2009).

60. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25,** 1091–3 (2009).

61. Bindea, G., Galon, J. & Mlecnik, B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics* **29,** 661–3 (2013).

62. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21,** 3448–9 (2005).

63. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19,** 524–531 (2003).

64. Hoops, S. *et al.* COPASI--a COmplex PAthway SImulator. *Bioinformatics* **22,** 3067–74 (2006).

65. Sauro, H. M. *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* **7,** 355–72 (2003).

66. Le Novère, N. *et al.* Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* **23,** 1509–15 (2005).

67. Robinson, P. N. *et al.* The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *Am J Hum Genet* **83,** 610–615 (2008).

68. Jiang, R., Gan, M. & He, P. Constructing a gene semantic similarity network for the inference of disease genes. *BMC Syst Biol* **5,** S2 (2011).

69. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *interjournal.org* (2006).

70. Albert, R. Scale-free networks in cell biology. *J Cell Sci* **118,** 4947–57 (2005).

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

72. Sheridan, C. Illumina claims $1,000 genome win. *Nat Biotechnol* **32,** 115–115 (2014).

73. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

74. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).

75. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43,** 585–9 (2011).

76. Yim, H.-S. *et al.* Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet* **46,** 88–92 (2014).

77. Sollie, A. *et al.* A New Coding System for Metabolic Disorders Demonstrates Gaps in the International Disease Classifications ICD-10 and SNOMED-CT which can be Barriers to Genotype-Phenotype data Sharing. *Hum Mutat* (2013). doi:10.1002/humu.22316

78. Firth, H. V *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84,** 524–533 (2009).

79. Rath, A. *et al.* Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat* **33,** 803–8 (2012).

80. Robinson, P. N. Deep phenotyping for precision medicine. *Hum Mutat* **33,** 777–780 (2012).

81. Vinayagam, A. *et al.* Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nat Methods* **11,** 94–9 (2014).

82. Girdea, M. *et al.* PhenoTips: Patient Phenotyping Software for Clinical and Research Use. *Hum Mutat* (2013). doi:10.1002/humu.22347

83. Hamosh, A. *et al.* PhenoDB: A New Web-Based Tool for the Collection, Storage, and Analysis of Phenotypic Features. *Hum Mutat* **34,** 566–71 (2013).

84. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144,** 986–998 (2011).

85. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148,** 1293–307 (2012).

86. Noble, D. A theory of biological relativity: no privileged level of causation. *Interface Focus* **2,** 55–64 (2012).

87. Rodríguez-Caso, C., Montañez, R., Cascante, M., Sánchez-Jiménez, F. & Medina, M. A. Mathematical modeling of polyamine metabolism in mammals. *J Biol Chem* **281,** 21799–21812 (2006).

88. Reed, M. C., Nijhout, H. F., Sparks, R. & Ulrich, C. M. A mathematical model of the methionine cycle. *J Theor Biol* **226,** 33–43 (2004).

89. Reed, M. C. *et al.* A mathematical model gives insights into nutritional and genetic aspects of folate-mediated one-carbon metabolism. *J Nutr* **136,** 2653–61 (2006).

90. Ulrich, C. M. *et al.* Mathematical modeling of folate metabolism: predicted effects of genetic polymorphisms on mechanisms and biomarkers relevant to carcinogenesis. *Cancer Epidemiol Biomarkers Prev* **17,** 1822–31 (2008).

91. Reed, M. C. *et al.* A mathematical model of glutathione metabolism. *Theor Biol Med Model* **5,** 8 (2008).

92. Tomasi, M. L. *et al.* Polyamine and methionine adenosyltransferase 2A crosstalk in human colon and liver cancer. *Exp Cell Res* **319,** 1902–11 (2013).

93. Hill, W. G. & Zhang, X.-S. On the Pleiotropic Structure of the Genotype–phenotype Map and the Evolvability of Complex Organisms . *Genet* (2012).

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

95. Goh, K.-I. *et al.* The human disease network . *Proc Natl Acad Sci* **104 ,** 8685–8690 (2007).

96. Zhang, M., Zhu, C., Jacomy, A., Lu, L. J. & Jegga, A. G. The orphan disease networks. *Am J Hum Genet* **88,** 755–766 (2011).

97. Zanzoni, A., Soler-López, M. & Aloy, P. A network medicine approach to human disease. *FEBS Lett* **583,** 1759–1765 (2009).

98. Jones, B. Genomics: personal genome project. *Nat Rev Genet* **13,** 599 (2012).

99. Bragin, E. *et al.* DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. *Nucleic Acids Res* **42,** D993–D1000 (2014).

GLOBAL SUMMARY OF RESULTS

METABOLIC MODELLING

Cellular metabolism depends on enzymes and transporters that are differentially expressed in mammal cells. Most of these enzymes or transporters are poorly characterized, although it is well known that their dysfunctions are associated with inherited metabolic diseases. The metabolism is a complex biological system and requires systemic approaches, such as the dynamic analysis of biochemical reaction networks to understand how metabolic processes are regulated. Indeed, the kinetic modelling is among the reference computational techniques –together with the structural analysis of metabolic networks– in this resurgence of the systemic view of the biology. According to the notion of the metabolism as a hierarchical and modular structure, we evaluate how the metabolism conserves functionality by integrating discrete metabolic modules into a single model.

Introduction to the metabolic modelling and the management of metabolic information from databases

Advances in systems biology have boosted the development of computational tools and resource facilities for modelling the pre-existing biological knowledge from multiple perspectives. Through international cooperation and support of many research groups, there is a wide catalogue of tools amongst which I would like to highlight all those that are compatible with the Systems Biology Markup Language (SBML). Most of the software compatible with SBML is also accessible to users coming from different disciplines and who are not familiar with programming language. During the first phase of my pre-doctoral training I set up an innovative educational methodology by designing a practical session for metabolic modelling.

This practical session consists of two distinct parts, the management of metabolic data and the design of a simple metabolic model of the glycolysis. For the first part, we select many different resources, databases and software, to acquire metabolic data such as metabolic pathways (KEGG, Reactome or Panther), biochemical reaction kinetics data (BRENDA or Sabio-RK) and repositories of published models in SBML (JWS Online Cellular Systems Modelling and BioModels). In the second part, students design and simulate a simple model, using CellDesigner and CoPaSi, to illustrate how glycolysis regulation exhibits an oscillatory behaviour. We conclude that this practical session is very useful and allow students to understand how biochemical reactions take place into cells, by using time-course simulations to explore changes in metabolite concentrations and metabolic fluxes.

Functional and Modular Integration of Amine Metabolism.

An integrated metabolic model of the metabolism of polyamines and sulfur amino acids in hepatocyte.

S-Adenosyl methionine (SAM) is the main donor of methyl groups and a metabolic hub that interconnects the polyamine, histamine, sulfur-containing amino acid and folate metabolisms. Previous experimental studies have proposed SAM as a physiological biomarker of cell stage in hepatocytes, as well as a key regulatory metabolite. SAM levels decrease after a switch in the genetic expression of two methionine adenosyltransferase (MAT) genes, from MAT1A to MAT2A. This switch is induced under proliferating conditions in hepatic cells (i.e. hepatocellular carcinoma). Our previous work in the modelling of the polyamine metabolism in mammals suggested an unexpected relevant role of SAM in the control of polyamines levels. Previous computational works suggested the role of polyamines in this change of activity by distinct MATs. Here we studied how decreased levels of SAM, the main biosynthetic precursor of polyamines, affect polyamine metabolism. We also propose a mathematical model to explain which mechanisms are involved to restore –or even to increase– spermine and spermidine levels under proliferative conditions.

For this purpose, we designed a metabolic model integrating those metabolic modules that are linked by SAM, such as polyamine and methionine metabolism. In addition, we also included the folate and glutathione metabolism for an extended view of the regulatory processes. The methodological approach consisted in to analysis of a mathematical model of ordinary differential equations (ODEs). First, we analysed individually those models previously published and they were contrasted to the experimental results. Later, we combined these models in a single model, in a SBML file, and we created a second version of the model considering the proliferative conditions. Previous models and the two combined versions of the models were submitted to BioModels Database (BIOMD0000000450 and MODEL1305060000).

The simulations of our model suggest a relevant role of SAM in polyamines homeostasis. In proliferative conditions, MAT-II is expressed which is inhibited by its own product, S-adenosylmethione. Therefore, cellular levels of SAM are decreased. SAM and ornithine are the main precursors of polyamines. Since polyamines are necessary for proliferation, their levels should be maintained or even increased. Accordingly, we proposed alternative regulatory mechanisms in the polyamine metabolism that depends on SAM availability. *In silico* experiments with our model under proliferative conditions, indicate that the metabolic flux redistribution to balance polyamines levels by lowered levels of SAM could be explained by an increased activity of ornithine decarboxylase (ODC). These predictions can be considered as experimentally validated as can be deduced from the conclusions of the recent work published by Mato and co-workers.

This computational work illustrates how the metabolic modularity, in the case of amine metabolism, is an operative feature of the structure of metabolic networks. Additionally, this model is also useful to raise novel hypothesis that will require experimental validation to amplify the relative information about complex biological systems.

An ontology-based tool to integrate metabolic data for kinetic modelling

Metabolic modelling requires data about biochemical reactions, enzymes, metabolites and modulators (activators or inhibitors), as well as to recognize those classical functional modules know as metabolic pathways. One of the most recognised problems in the field of kinetic modelling is the lack of kinetic data about the activity of enzymes.

We developed a user-friendly application, named as Systems Biology Metabolic Modelling Assistant (SBMM Assistant, http://www.sbmm.uma.es/)

able to integrate kinetic and metabolic information of any organism. SBMM assistant, works on the bases of an ontology-based mediator developed to integrate data from KEGG, CHEBI, BRENDA and SABIORK. SBMM assistant is characterized by the following features: It is an SBML-compatible and friendly tool able to guide to the novel or experienced user to capture, enrich, generate and visualize biological networks, to make basic queries on enzymatic kinetics and regulation, and to annotate this information following the MIRIAM specifications. Semantic-web technologies have been claimed to be applied specifically to solve the present shortcomings in the workflow of metabolic modelling, in this sense SBMM assistant is an example of such a specific application, having the aim to act as a complementary tool (assistant) for metabolic modelling analysis programs. Thus, it is not only able to capture, enrich and store information in models, but also helps the cross-talk among the different resources and tools in a friendly way.

This resource has been used to enrich of data other databases about peroxisomal disorders (PeroxisomeDB 2.0) and to generate a pilot knowledge base (Amine Knowledge Base: asp.uma.es/amineKB). A subsequent version of SBMM assistant, names as SBMM Assistant: Social Pathway Annotation, was oriented to the automatic curating process of databases. These efforts are, specially, interesting in the case of inborn errors of the metabolism to share information due to limited biomedical information about these rare genetic diseases.

NETWORK MEDICINE APPROACHES

A systemic view of rare diseases, perhaps not so rare

Systems biology has been proposed in numerous studies to increase understanding of complex diseases like cancer or neurodegenerative diseases. Complex diseases are usually polygenic and each genetic variant point to a risk value, but they also show and strong multifactorial and environmental influences. On the other hand, rare diseases use to be genetically characterized and there is variability in terms of the number of genes that may be associated with disease. In addition, their low prevalence (less than 0.05% in the population) and their typically conserved clinical features make them very interesting from a biological point of view. Our hypothesis is that the incidence of these diseases in the population is the consequence of certain biological constraints and susceptible to be studied by a systemic view.

Network biology: a direct approach to study rare diseases genetic relationships

A dataset of 2125 RDs associated with 2331 genes has been modelled and analysed by network biology methods. These analyses were carried out attending to the relationships patterns between genes and RDs, for this reason we subset genes according to the number of genes associated with diseases (evaluating pathological gene co-associations) and the number of diseases associated with a gene (pleiotropy). Functional characteristics and topological properties of rare disease-causing genes have been analysed in metabolic and protein-protein interactions (PPI) networks from published interactomes. We have used different tools like BioMart, ClueGO, and Cytoscape plugins to analyse this dataset with relevant biological knowledge. Different R packages and own scripts were used for data management and statistical analysis.

Our results indicate that the associations between genes and rare diseases depend on complex interactions, but we observe shared and distinguishable features in each subset of genes. Genes associated with monogenic rare disease are functionally different from those genes that are associated with more than one RDs. RDs related genes exhibit a functional bias according to the biological network in which they are mainly implicated, either metabolic or protein-protein interactions networks. In particular, there is likely a dependence on the functional context in which genetic variants cause dysfunctions. Therefore, network analyses are useful to get a deeper understanding about the molecular aetiology and intervention capacity on low-prevalence diseases.

The human pathophenotype-causing genes network

Diseases Networks are useful to study the molecular complexity of genetic diseases. Two main disease networks, "the human diseases networks" (HDN, Goh *et al.* 2007) and "the orphan disease networks" (ODN, Zhang *et al.* 2007), have been published to date among others. However, in these networks, each single node is a disease, characterized as a set of clinical features descriptions (pathophenotypes) represented as pathological entities. Most of these diseases were described using evidence-based medicine methods allowing physicians systematically to differentiate types and sub-types of diseases. Therefore, the representation of diseases as entities, without relationships to other phenotypically similar diseases, affects to network medicine methods. We hypothesize that the pathophenotypic relationships among diseases can help to find out interrelations in molecular events originated by mutations.

In this work, we built and analysed the human pathophenome network to be compared to HDN and ODN. Unlike these previous networks, the pathophenome uses semantic similarities. The pathophenotypic similarities were calculated between pair of genes annotating phenotypic abnormalities in the "Human Phenotype Ontology" and, subsequently, comparing gene phenotypic spaces. The resulting human pathophenome network contains 1706 genes (nodes) and 26192 significant pathophenotypic similarities (edges). This network reveals a strong re-arrangement of the pathological relationships among genes and, moreover, they are measurable by phenotypic similarities. Many novel pathophenotypic interactions between genes have been uncovered. Our results indicate that pathophenotypes might contribute to discover pre-clinical stages and co-dependencies among disease- causing genes.

PhenUMA: an integrative tool of biomedical relationships among genes and diseases.

Several types of regulatory genetic interactions can be directly or indirectly associated with human mutations. These gene-gene relationships are usually based on their co-associations to biological processes, co-existence in cellular locations, co-expression in cell lines, physical interactions, etc. In addition, pathological processes can share similar phenotypic features and mutations in the same genomic location, or even to exhibit genetic variations in different genomic regions. Thus, integrative analyses of all these complex interactions can help us prioritize those relationships between genes and diseases most deserving to be studied by researchers and physicians.

PhenUMA (www.phenuma.uma.es) is a web application to build, analyse and visualize networks based on both functional and phenotypic relationships. This novel tool uses semantic similarity methods to study interconnected genes using their functional and phenotypic features. Furthermore, phenotypic similarities can be useful to analyse clusters of diseases sharing specific phenotypes or to find diseases related to reported phenotypes. This tool enables the inference of new links relative to genes, biological functions and diseases.

Conclusions: This framework provides networks built using integrated information from biomedical and biomolecular data repositories. PhenUMA represents a further step towards the advance of new technologies for genomic and personalized medicine.

Network medicine approaches to study phenotypic similarities based networks

Phenotypic variance is a feature of all biological systems under the influence of complex molecular process and environmental changes. Network medicine is a promising field based on systems biology approaches to study biomedical issues through networks. These network models provide an integrative framework for the analysis 'omics data to characterize the molecular aetiology of pathological processes. Previous network models considered diseases as conceptual entities. Here, we use patients as the nodes in combination with their genetic information. In this work, we will explore the benefits of using individual phenotypic profile of patients to explorer their similarities. Here we also present a systematic method to study patients sharing phenotypes and presenting similar copy number variations, structural variants. Finally, we report a high-resolution map of pathogenic phenotypes associated with their respective significant genomic locations.

RESUMEN GLOBAL DE LOS RESULTADOS

A lo largo de los trabajos de investigación desarrollados recogidos en esta tesis he participado en más de 20 congresos nacionales e internacionales. Por imperativo legal, se exige a los doctorandos que presenten un resumen de más de 5000 palabras que se ajusta muy poco a lo que desde mi punto de vista es un resumen. A pesar de que la medida a priori no me pareció adecuada a la estructura que yo he decidido para mi Tesis Doctoral por compendio de artículos, he reconsiderado junto con mi directores que una forma interesante para rastrear y hacer un seguimiento de la evolución de mi trabajo es precisamente ver lo que he presentado en cada uno de estos congresos, hasta la fecha.

MODELADO METABÓLICO.

El metabolismo celular depende de enzimas y transportadores que se expresan de forma diferencial en las células de mamíferos. La gran mayoría de enzimas o transportadores están pobremente caracterizados, pero sí se sabe que sus disfunciones son el origen de muchas enfermedades, enfermedades metabólicas hereditarias. El metabolismo es un sistema biológico complejo que requiere de abordajes sistémicos, tales cómo el análisis dinámico de redes de reacciones bioquímicas para entender como se regulan los procesos metabólicos. De hecho, el modelado cinético está entre las técnicas computacionales de referencia –junto con el análisis estructural de redes metabólicas– en este resurgimiento del enfoque sistémico de la biología.

En este sentido, partiendo de la concepción del metabolismo celular cómo una estructura jerárquica y modular, nosotros estudiamos cómo el metabolismo conserva su funcionalidad integrando módulos metabólicos discretos en un mismo modelo.

Introducción al modelado metabólico y a la gestión de información metabólica en bases de datos

Los avances en la biología de sistemas, han propulsado el desarrollo de herramientas y servicios de recursos computacionales para modelar el conocimiento biológico preexistente desde múltiples perspectivas. Gracias a la cooperación y a la competitividad entre grupos se han desarrollado un amplio catálogo de herramientas, entre las que me gustaría destacar aquellas orientadas al análisis de reacciones bioquímicas y que son compatibles con el formato SBML (*Systems Biology Markup Language*). Muchos de estos programas suelen ser accesibles a un público interdisciplinar que no está familiarizado al uso de lenguajes de programación. En la experiencia de poner en marcha algunas de estas herramientas encontramos -entre muchas otras aplicaciones- una interesante y sencilla metodología docente. Durante la primera fase de mi formación pre-doctoral diseñé una sesión práctica seleccionando recursos bioinformáticos, como programas y bases de datos, para la modelización de rutas metabólicas. La primera parte de la sesión práctica se centró en capturar información puntual de bases de datos de rutas metabólicas (REACTOME y Panther), de carácter enzimológico (BRENDA) y otras de modelos publicados en SBML (JWS Online Cellular Systems Modelling y BioModels). La segunda parte de la sesión, consistió en desarrollar un sencillo modelo para estudiar de forma ilustrativa la dinámica oscilatoria de la glucólisis usando programas como CellDesigner y COPASI. Podemos concluir que con esta práctica los alumnos no solo se han introducido en la gestión de información metabólica de origen experimental en distintas bases de datos, sino que mediante la monitorización de un tutorial han realizado simulaciones y el análisis del estado estacionario de un modelo oscilatorio de la glucólisis.

Integración modular y funcional del metabolismo.

Un modelo integrado del metabolismo de poliaminas y aminoácidos azufrados en hepatocitos.

La S-adenosilmetionina (SAM) es el principal donador de grupos metilo y un "hub" metabólico que interconecta el metabolismo de poliaminas, histamina, aminoácidos azufrados y folatos. Estudios previos proponen que SAM es un marcador fisiológico del hepatocito y un elemento clave de regulación metabólica. Los niveles de SAM disminuyen cuando se produce un cambio de expresión génica entre las distintas metionina adenosiltransferasas (MAT) de MAT-I/III a MAT-II, en condiciones de proliferación celular (como es el caso de los hepatocarcinomas). Nuestro trabajo previo en el modelado del metabolismo de poliaminas en mamíferos sugiere un papel más relevante de SAM en el metabolismo de poliaminas que el hasta entonces reconocido. Algunos trabajos biocomputacionales han estudiado el papel de la activación de las poliaminas ante el cambio de actividad por las distintas MATs. Sin embargo, este estudio previo ha considerado poco significativo el flujo de SAM como precursor de espermina y espermidina. Nuestro objetivo era demostrar que en condiciones de proliferación celular, en las que los niveles de SAM disminuyen, aumentan las tasas de biosíntesis de poliaminas para mantener el pool de las mismas.

Para ello se diseñó un modelo metabólico que integra aquellos módulos metabólicos que interconecta SAM, tales como el metabolismo de las poliaminas, los ciclos de los folatos y de los metilos activados. El enfoque metodológico ha consistido en desarrollar un modelo matemático en un sistema de ecuaciones diferenciales ordinarias. En un primer lugar, se analizaron de forma individual los modelos previamente publicados y se contrastaron con datos experimentales. Posteriormente, se realizó la integración de toda la información en ficheros SBML generado dos versiones distintas del modelo, una para condiciones

normales y otra para condiciones proliferativas, que están disponibles online en BioModels (BIOMD0000000450 and MODEL1305060000).

Las simulaciones de nuestro modelo sugieren un papel importante de SAM y acetil-CoA en la homeostasis de poliaminas, algo que no se había considerado en estudios experimentales previos. En estado proliferativo actúa MAT-II que es inhibida por su propio producto, SAM, bajando sus niveles celulares. SAM junto con la ornitina son los precursores inmediatos de poliaminas. En estado proliferativo las poliaminas son relevantes y sus niveles celulares se mantienen e incluso aumentan. En este sentido, se proponen mecanismos de regulación alternativos en el metabolismo de las poliaminas que dependan directamente de los niveles de SAM. Nuestro modelo integrado, propone que la redistribución de flujos metabólicos per se, compensa la perdida del precursor, SAM, fundamentalmente por el aumento de actividad de ODC. Nosotros observamos que, en función del estado fisiológico celular, las poliaminas pueden ser un importante destino metabólico de SAM.

Las conclusiones derivadas de este modelo se pueden considerar demostradas experimentalmente, tal y como se puede desprender de las conclusiones de los trabajos publicados por el grupo del Dr. Mato y colaboradores [92], un trabajo experimental que se publicó posteriormente a nuestro modelo. Este trabajo de la biología computacional ejemplifica cómo la modularidad metabólica en el contexto del metabolismo de aminas en mamíferos es completamente operativa. Además, este modelo ayuda a su vez a plantear nuevas hipótesis que precisan de validación experimental para amplificar la información relativa a los sistemas biológicos complejos.

Una herramienta basada en ontologías para la integración de información metabólica para modelado cinético

El modelado metabólico precisa de todo el conocimiento relacionado con las reacciones, las enzimas, los metabolitos, los modificadores (activadores e

inhibidores), así como de la estructura de las rutas metabólicas. Sin embargo, en el terreno del modelado cinético uno de los más reconocidos problemas es la laguna de conocimiento de la cinética de las enzimas.

En el terreno de la integración, hemos desarrollado una herramienta, denominada Systems Biology Metabolic Modeling Assistant (SBMM Assistant: www.sbmm.uma.es) capaz de integrar información cinética y metabólica sobre cualquier enzima de cualquier organismo. SBMM Assistant es una aplicación online compatible con SBML que sirve de guía al usuario para hacer consultas básicas de metabolismo, crear redes metabólicas, consultar parámetros cinéticos y anotar la información siguiendo las especificaciones de MIRIAM. Por el momento, dicha aplicación integra, mediante un sistema mediador basado en ontologías, una selección de la información metabólica disponible en KEGG, CHEBI, BRENDA y SABIORK. Esta información integrada permite generar un entorno rico en datos metabólicos del conjunto de reacciones bioquímicas involucradas en nuestro estudio.

Este recurso ya se está aplicando en el terreno de las enfermedades raras para estudiar algunas patologías relacionadas con el metabolismo de aminas y aminoácidos y para el desarrollo y enriquecimiento de una base de datos sobre proteínas peroxisomales (PeroxisomeDB). También se ha utilizado esta herramienta para desarrollar un piloto de generación facilitada de una base de datos (Amine Knowledge Base: asp.uma.es/amineKB) y una subsiguiente versión denominada "SBMM Assistant: Social Pathway Annotation" (http://www.sbmm.uma.es/spa/) que se diseñó para el curado automático de bases de datos. Estos esfuerzos son, especialmente, interesantes en el caso de las enfermedades raras para compartir información y debido a la escasez de datos biológicos.

APROXIMACIONES BASADAS EN REDES DE LA MEDICINA.

La aproximaciones de la biología de sistema se están usando en numerosos estudios para comprender mejor las enfermedades complejas como el cáncer o las neurodegenerativas. Las enfermedades complejas suelen ser poligénicas y cada variante genética que esté asociada presenta a un valor de riesgo, pero ellas también muestran un fuerte influencia multifactorial y por el ambiente. Por otro lado, las enfermedades raras suelen estar claramente caracterizadas genéticamente y presentan una gran variabilidad en el número de genes involucrados. Además, su baja prevalencia (menos a un 0.05 % de la población) y el cuadro fenotípico típicamente conservado, las hace muy interesantes desde un punto de vista biológico.

Nuestra hipótesis es que la incidencia de esas enfermedades en la población es la consecuencia de determinadas restricciones biológicas que son susceptibles de estudiarse bajo una perspectiva sistémica.

Biología de redes: una aproximación directa para estudiar las relaciones genética entre enfermedades raras.

Un conjunto de datos de 2150 enfermedades raras (ERs) y 2331 genes causantes de enfermedad se han usado para modelar y analizar utilizando métodos de la biología de redes. Estos análisis se han realizado teniendo en cuenta los patrones de relaciones que se producen entre esos genes y las ERs. Por eso hemos agrupado los genes en función del número de genes asociados a las enfermedades (genes que presentan una elevada co-asociación patológica) y el número de enfermedades asociadas al gen (pleiotropía). Las características topológicas y funcionales de los genes causantes de enfermedades raras fueron analizadas para una red metabólica y otra de interacciones proteína-proteína (PPI), interactomas publicados. Para ello he utilizado distintas herramientas tales como BioMart, ClueGO y plugins de Cytoscape para analizar este dataset con información biológica relevante. También he utilizado diversas librerías de R y mis propios scripts para la gestión, representación y el análisis estadístico de los datos.

Nuestros resultados indican que las asociaciones entre genes y enfermedades raras dependen de complejas interacciones, pero nosotros observamos características compartidas y diferenciables entre los distintos sets de genes. Por ejemplo, los genes asociados a una sola enfermedad rara monogénica son funcionalmente diferentes a aquellos genes que están asociados a más de una enfermedad. Los genes asociados a dichas enfermedades muestran un sesgo funcional según la red biomolecular metabólica o de interacciones entre proteínas a la que estén principalmente asociados. En concreto, en muchos casos hay una especie de dependencia entre las relaciones de genes causantes de enfermedades parecidas al contexto funcional en el que las variantes genéticas causan disfunciones. Por lo tanto, los análisis basados en redes son muy útiles para tener una comprensión más profunda de la etiología molecular así como de las posibilidades de intervención en las enfermedades de baja prevalencia.

La red de genes que causan fenotipos patológicos

Las redes de enfermedades son útiles para estudiar la complejidad de las enfermedades genéticas. Dos "diseasomas" o "enfermedomas" se han construido a partir de la información de las bases de datos más importantes de estudios de asociación genética, "the human diseases networks" generada a partir de los datos contenidos en OMIM (HDN, Goh et al. 2007) y "the orphan disease networks" generada a partir de los datos contenidos en Orphanet (ODN, Zhang et al. 2011). En estas redes, cada nodo representa una enfermedad caracterizada por un conjunto de características clínicas y síntomas (fenotipos patológicos) que se consideran como conceptos unitarios. La mayoría de esas enfermedades se han descrito siguiendo los métodos de la medicina basada en la evidencia lo que ha permitido a los médicos diferenciar de forma sistemática entre tipos y subtipos de enfermedades para el diagnóstico. Pero la representación de las enfermedades como entidades, sin establecer las relaciones fenotípicas que presenta con otras enfermedades similares, afecta a las aproximaciones de la medicina en red. Nosotros postulamos que las relaciones basadas en los fenotipos patológicos entre las enfermedades pueden encontrar interrelaciones en los eventos moleculares que se desencadenan a partir de las variaciones genéticas.

En este trabajo, nosotros hemos construido y analizado una red del pato-fenoma humano y lo hemos comparado con los "diseasomas" ya conocidos HDN y ODN. A diferencia de estas redes previas, el pato-fenoma utiliza similitud semántica. Estas similitudes pato-fenotípicas se calcularon entre pares de genes a partir de los pato-fenotípicos a los que se asocian en la "Human Phenotype Ontology" y, posteriormente, se compararon sus espacios pato-fenotípicos. La red del pato-fenoma humano resultante contiene 1706 genes (nodos) y 26192 similitudes pato-fenotípicas entre pares de genes significativas (aristas). Esta red revela un fuerte re-ordenamiento de las relaciones pato-fenotípicas entre los genes y, más aún, se pueden medir a partir de su valor de similitud. Además, se han descubierto muchas nuevas relaciones pato-fenotípicas entre los genes. Estos

resultados indican que los pato-fenotipos pueden contribuir a descubrir fases pre-clínicas y co-dependencias entre los genes causantes de enfermedad.

PhenUMA: una herramienta para integrar relaciones biomédicas entre los genes y las enfermedades

Muchos tipos de interacciones genéticas reguladores puede estar directa o indirectamente asociadas al efecto de las mutaciones humanas. Esas relaciones gen-gen se basan usualmente en su co-asociación a procesos biológicos, su co-existencia en localizaciones celulares, su co-expresión en líneas celulares, interacciones físicas entre proteínas, etc. Además, los procesos patológicos pueden compartir características fenotípicas similares y mutaciones en la misma localización genómica, o incluso mostrar variaciones genéticas en regiones genómicas diferentes. Por eso, análisis integrales que consideren todas esas relaciones complejas puede ayudarnos a priorizar aquellas relaciones entre genes y enfermedades que requieren ser estudiadas por investigadores y médicos.

PhenUMA (http://www.phenuma.uma.es/) es una aplicación web para construir, analizar y visualizar redes basadas en relaciones funcionales y fenotípicas. Esta nueva herramienta utiliza diversas medidas de similitud semántica para estudiar genes interconectados a partir de sus anotaciones funcionales y fenotípicas. Más aún, las similitudes fenotípicas se pueden utilizar para generar clústeres de enfermedades que comparten fenotipos específicos o encontrar enfermedades que estén asociadas a un conjunto de fenotipos. Esta herramienta permite la inferencia de nuevas relaciones en procesos patológicos de genes, procesos biológicos y enfermedades. Este marco de trabajo provee redes construidas a partir de información integrada de distintos repositorios de información biológica y biomédica. PhenUMA representa un paso más hacia el avance de las nuevas tecnologías para la medicina genómica y personalizada.

Aproximaciones de la medicina de redes para estudiar las redes basadas en similitud fenotípica

La variabilidad fenotípica es una característica de los sistemas biológicos bajo la influencia de procesos moleculares complejos y cambios ambientales. La medicina de red usa los modelos de redes para proveer un marco de trabajo integral para el análisis de esas complejas interacciones junto con datos procedentes de las técnicas ómicas. Estos datos permiten estudiar la etiología molecular y los procesos patológicos. La mayoría de esos modelos consideran a las enfermedades como entidades conceptuales y esta noción puede reducir de forma drástica la utilidad de los análisis basados en redes. En este trabajo, he explorado los beneficios de utilizar la los fenotipos para construir redes de enfermedades y de genes en base sus similitudes fenotípicas. Además, también presentamos un método sistemático para analizar los perfiles fenotípicos de pacientes que presentan mutaciones estructurales, variaciones de número de copias, potencialmente patogénicas. Finalmente, nosotros hemos aportado un mapa de alta resolución de los fenotipos patológicos y su grado de asociación a regiones genómicas específicas.