Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE BIOTECNOLOGICHE E FARMACEUTICHE

Ciclo XXX

Settore Concorsuale: 05/E1 - BIOCHIMICA GENERALE E BIOCHIMICA CLINICA

Settore Scientifico Disciplinare: BIO/10 - BIOCHIMICA

DEVELOPMENT OF COMPUTATIONAL METHODS FOR BIOLOGICAL COMPLEXITY

Presentata da: SAMUELE BOVO

Coordinatore Dottorato — **Prof. Santi Mario Spampinato**

Supervisore — **Prof.ssa Rita Casadio**

Esame finale anno 2018

(This page has been left blank intentionally)

In the appendix are included the printed versions of following journal articles:

1. **Bovo S**, *et al.* (2016) **NET-GE: a web-server for NETwork-based Gene Enrichment of sets of human genes related to the same phenotype.** *Bioinformatics*, 32(22):3489-3491 PMID:27485441
2. **Bovo S**, *et al.* (2017) **From Protein Variations to Biological Processes and Pathways with NET-GE.** *Genomics and Computational Biology*, 3(3); e45;
3. Babbi G, *et al.* (2017) **eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes.** *BMC Genomics*, 18(Suppl5):554. PMID:28812536.
4. Xu Q, *et al.* (2017) **Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4.** *Human Mutation*, 38(9):1123-1131. PMID:28370845.
5. Daneshjou R, *et al.* (2017) **Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges.** *Hum Mutation*, 38(9):1182-1192. PMID 28634997.

(This page has been left blank intentionally)

Abstract (English)

The cell is a complex system. In this system, the different layers of biological information establish complex links converging in the space of functions; processes and pathways talk each other defining cell types and organs. In the space of biological functions, this lead to a higher order of "emergence", greater than the sum of the single parts, defining a biological entity a complex system. The introduction of omic techniques has made possible to investigate – in a single shot – the complexity of each biological layer. With the different technologies we can have a near complete readout of the different biomolecules. However, it is only through data integration that we can let emerge and understand biological complexity. Given the complexity of the problem, we are far from having fully understood and developed exhaustive computational methods. Thus, this make urgent the exploration of biological complexity through the implementation of more powerful tools relying on new data and hypotheses. To this aim, Bioinformatics and Computational Biology play determinant roles. The present thesis describes computational methods aimed at deciphering biological complexity starting from genomic, interactomic, metabolomic and functional data.

The first part (chapters 1 – 5) describes NET-GE, a network-based gene enrichment tool aimed at extracting biological functions and processes of a set of gene/proteins related to a phenotype. NET-GE exploits the information stored in biological networks, to better define the biological events occurring at gene/protein level. When tested against set of genes related to OMIM diseases (#244 gene set), NET-GE retrieves enriched terms not detectable by standard method, in a number of diseases ranging from 19% to 40% of the whole OMIM set.

The first part (chapters 6 – 7) describes also eDGAR, a database collecting and organizing gene-disease associations data as retrieved from OMIM, Humsavar and ClinVar. The database is aimed at providing a comprehensive knowledge of the molecular signatures at the basis of 2,672 diseases (621 are polygenic, i.e. associated to multiple genes). Thanks to its ability in detect new functional terms, NET-GE is used in eDGAR to enhance the understanding of the biological function involved in the development of diseases.

The second part (chapter 8 – 9) deals with metabolomics. I describe a new way to perform metabolite enrichment analysis. Given the strict relationship among genes, proteins and metabolites, I explore the metabolome by exploiting the features of an interactome. To do that, I developed NET-GE^{M}, a version of NET-GE rewritten to enrich functions and pathways starting from a list of metabolites. The NET-GE^{M} analysis of a set of 41 metabolites related to Parkinson's disease (PD) detected pathways not highlighted by a canonical metabolite enrichment tool. A fraction of these pathways is retrievable also when performing a NET-GE analysis with a gene set of genes related to PD. This highlights the notion that complementary information is stored in the gene and metabolite layers of biological complexity.

The third part (chapter 10 – 11) describes the methods and results obtained in the CAGI experiment, a community experiment aimed at assessing computational methods used to predict the impact of genomic variation on a phenotype. Different challenges were proposed in the CAGI4 edition, and here I describe the methods related to three of them: two challenges involving the prediction of the healthy status based on exome data, and one challenge related to the prediction of the effects of mutation on the activity and allosteric regulation of the human pyruvate kinase.

Overall, the developed methods aim at efficiently integrate different data and boost the way to decipher and understand biological complexity.

Abstract (Italian)

La cellula è un sistema complesso. In questo sistema, i differenti *layer* biologici stabiliscono relazioni che danno origine a funzioni e *pathway* molecolari. Un sistema è solitamente caratterizzato da proprietà emergenti, dove il risultato finale e maggiore rispetto alla somma delle singole pasrti. Nel contesto biologico, le interazioni tra *layer* fanno emergere informazioni come la specificità cellulare e di organo.

L'introduzione delle diverse tecnologie omiche ha reso possibile investigare la complessità di ogni *layer*. Con una singola analisi, le diverse tecnologie riescono a restituire un profilo biomolecolare quasi completo. Tuttavia, per avere una chiara comprensione del problema biologico, i diversi profili devono essere integrati. Attualmente siamo distanti dall'avere sviluppato metodi per l'analisi della complessità in modo esauriente; questo rende urgente l'implementazione di metodi sempre potenti, bastati su nuovi dati e nuovi approcci. A tale scopo, la Bioinformatica e la Biologia computazionale giocano un ruolo fondamentale.

Questa tesi descrive nuovi metodi computazioni atti alla comprensione della complessità biologica a partire da dati genomici, di interazione, di metaboliti e di funzione biologica.

Nella prima parte (capitoli 1 – 5) introduco NET-GE, un *tool* che si propone di estrarre funzioni biologiche condivise tra geni, tramite l'utilizzo di reti di interazioni proteica (STRING). Quest'ultime permettono infatti di meglio definire le funzioni stesse. Testato con *set* di geni coinvolti in malattie (#244 malattie ricavate da OMIM), in un numero di malattie che varia dal 19% al 40% (a seconda del metodo e delle funzioni analizzate), NET-GE ha arricchito per nuove funzioni. Nella prima parte (capitoli 6 – 7) descrivo anche eDGAR, un *database* che raccoglie associazioni gene-malattia da diverse risorse (OMIM, ClinVar, Humsavar). Per un totale di 2,672 malattie (621 poligeniche; legate a più geni), eDGAR si propone di dare una completa caratterizzazione molecolare del fenotipo finale. Grazie all'abilità di arricchire per nuove funzioni, NET-GE è usato in eDGAR per migliorare la caratterizzazione funzionale delle diverse malattie.

La seconda parte (capitoli 8 – 9) tratta di metabolomica. Descrivo un nuovo metodo per l'analisi funzionale di dati metabolomici basato sullo stretto rapporto tra geni, proteine e metaboliti. Ho sviluppato NET-GE^{M}, una versione di NET-GE riscritta per lavorare con dati metabolomici. In un caso di studio relativo alla malattia di Parkinson (PD), l'analisi di 41 metaboliti ha portato ad ottenere risultati interessanti: NET-GE^{M} arricchisce per *pathway* non individuabili con metodi standard (ma riscontrati in letteratura). Inoltre, parte di questi risultati si è ottenuta tramite l'analisi – basata su NET-GE – di un set di geni relativo al PD. Questo ci permette di rimarcare quanto genomica e metabolomica siano tecniche complementari, con parte dell'informazione condivisa dai diversi *layer* biologici.

La terza ed ultima parte (capitoli 10 – 11) tratta del CAGI, un esperimento internazionale atto a valutare metodi computazionali utilizzati nella predizione dell'impatto di varianti genomiche a livello fenotipico. Diverse "sfide" sono state rilasciate nella quarta edizione. Qui descrivo i metodi sviluppati per affrontare tre di loro. Due sfide riguardavano la predizione dello stato di salute (sano/malato) partendo da dati di esomica. La terza sfida ha riguardato la predizione dell'effetto di mutazioni sull'attività e regolazione allosterica della piruvato kinasi umana.

Complessivamente, questi metodi si propongono di integrare in modo efficiente diversi dati biologici, migliorando il modo di comprendere e di analizzare la complessità biologica.

Acknowledgments

It goes without saying that many people have contributed in a scientific way to this thesis. I thank them all for the continuous support, guidance, insights and discussions given to me during the last three years.

Personally speaking, I would like to thank:

- my advisor, Prof. Rita Casadio, for teaching me what determination is and how to put it in every day of my life;
- all the members of the Bologna Biocomputing Group: Prof. Pier Luigi Martelli, Dr. Giuseppe Profiti, Dr. Giulia Babbi, Dr. Castrense Savojardo, Dr. Francesco Aggazio, Dr. Pietro Di Lena and Prof. Piero Fariselli (names randomly ordered);
- Prof. Luca Fontanesi, for providing me the opportunity to learn and participate in many different projects since the first time I met him;
- all the members of the Fontanesi's Lab: Dr. Giuseppina Schiavo, Dr. Anisa Ribani, Dr. Valerio Joe Utzeri and Dr. Claudia Geraci;
- Dr. Gianluca Mazzoni, a special person always there for me; there are no words to express my gratitude to him;
- the Ph.D. program coordinator, Prof. Santi Mario Spampinato, for its passion and availability;
- the people I met in Bologna and around the world, that made the Ph.D. experience the best one I ever had.

A special thanks to my family, for having trusted me and encouraged any my decisions.

(This page has been left blank intentionally)

Contents

ABSTRACT (ENGLISH) III

ABSTRACT (ITALIAN) IV

ACKNOWLEDGMENTS V

Part I - Network-based gene enrichment analysis with NET-GE

1 INTRODUCTION 5

1.1 DECIPHERING BIOLOGICAL COMPLEXITY 5
1.2 PROTEIN-PROTEIN INTERACTION: WHY IS IT SO IMPORTANT? 7
1.3 STRING AS COMPREHENSIVE PPI NETWORK 8
1.4 PROTEIN-PROTEIN INTERACTIONS NETWORKS AS GRAPHS 9
1.5 GENE/PROTEIN ANNOTATION AND RELATED DATABASES 11
1.5.1 KEGG 12
1.5.2 REACTOME 12
1.5.3 GENE ONTOLOGY 13
1.5.4 HIERARCHICAL REPRESENTATION OF FUNCTIONAL ANNOTATIONS 13
1.6 QUANTIFYING THE FUNCTION SPECIFICITY AND SIMILARITY. 14
1.7 GENE ENRICHMENT ANALYSIS 15
1.7.1 GENE ENRICHMENT ANALYSIS: HOW DOES IT WORKS? 15
1.7.2 CLASSIFICATION OF GENE ENRICHMENT METHODS 16
1.7.3 FROM STANDARD TO NETWORK-BASED GENE ENRICHMENT 17

2 NET-GE 19

2.1 DATABASES 19
2.2 MODULE EXTRACTION 20
2.3 ENRICHMENT PROCEDURE 23

3 NET-GE WEB SERVER 24

3.1 NET-GE WEB SERVER: INPUT 25
3.2 NET-GE WEB SERVER: OUTPUT 25
3.3 GRAPHICAL REPRESENTATION OF THE ENRICHED TERMS 27

4 BENCHMARKING THE METHOD: A QUANTITATIVE APPROACH. 27

5 STUDY CASES: A QUALITATIVE EVALUATION OF NET-GE. 29

5.1 THE ADHD STUDY CASE 30
5.2 THE OCD STUDY CASE 32

6 NET-GE AND EDGAR: MOLECULAR SIGNATURES OF DISEASES 34

6.1 FUNCTIONAL RELATIONSHIPS OF DISEASE-RELATED GENES 34
6.2 HYPOPARATHYROIDISM AS CASE STUDY 37

7 CONCLUSIONS 37

Rethinking metabolite enrichment analysis: NET-GE^{M}

8 INTRODUCTION: WHY METABOLOMICS? 41

8.1 HOW TO FUNCTIONALLY INTERPRET METABOLOMIC DATA 42
8.1.1 METABOLITE-RELATED DATABASES: A BRIEF OVERVIEW. 42
8.2 NET-GE^{M}: RETHINKING METABOLITE FUNCTIONAL ASSOCIATION. 43
8.2.1 NET-GE^{M}: HOW DOES IT WORK? 43
8.2.2 IMPLEMENTING NET-GE^{M}: DEFINING A METABOLITE-GENE MAPPING TABLE 44
8.2.3 WHY GENES INSTEAD OF METABOLITES? A CRITICAL VIEW 44
8.2.4 THE STATISTICS AT THE BASIS OF NET-GE^{M} AND MBROLE2.0 46
8.2.5 TESTING THE METHOD: THE PARKINSON'S DISEASE CASE STUDY 46

9 CONCLUSIONS 50

Part III
The Critical Assessment of Genome Interpretation (CAGI) experiment

10 INTRODUCTION 52

10.1 THE PYRUVATE KINASE CHALLENGE 53
10.2 DATASETS 55
10.3 METHOD 56
10.4 PERFORMANCE ASSESSMENT 57
10.5 RESULTS 59
10.5.1 PREDICTION OF THE L-PYK ENZYME ACTIVITY 59
10.5.2 PREDICTION OF $Q_{AX\text{-}ALA}$ 61

10.5.3 PREDICTION OF $Q_{AX\text{-}F\text{-}1,6\text{-}BP}$ 63
10.6 CONCLUSION 63

11 CAGI4 EXOME CHALLENGES 64

11.1 CROHN'S DISEASE 64
11.2 BIPOLAR DISORDER 65
11.3 DATASETS 65
11.3.1 CROHN'S DISEASE 65
11.3.2 BIPOLAR DISORDER 66
11.4 METHODS 66
11.4.1 DATA QUALITY ASSESSMENT 66
11.4.2 VARIANT/GENE ANNOTATION 67
11.4.3 GENE SELECTION 67
11.4.4 ASSESSORS EVALUATION 68
11.5 RESULTS 69
11.5.1 CROHN'S DISEASE EXOME CHALLENGE. 69
11.5.2 BIPOLAR DISORDER EXOME CHALLENGE. 71
11.6 CONCLUSION 73

12 GENERAL CONCLUSIONS 73

13 REFERENCES 75

14 SUPPLEMENTARY MATERIAL 87

15 LIST OF PUBLICATIONS 90

15.1 PEER-REVIEWED PUBLICATIONS RELATED TO THIS THESIS 90
15.2 OTHER PEER-REVIEWED PUBLICATIONS 90
15.3 ABSTRACTS AND POSTERS 91

16 APPENDIX 95

Part I

Network-based gene enrichment analysis with NET-GE

1 Introduction

1.1 Deciphering biological complexity

Omics techniques have changed the way to investigate the complexity of a biological system (Gullapalli *et al.*, 2012; Robinson, 2014). But, what is a biological system? And why is it defined as complex?

A system can be defined as an integration of parts or elements, connected in some form of interaction or interdependence to form a (complex) unitary whole (Misra, 2008). Usually, systems share common characteristics such as: (i) structure, defined by components and their parts, (ii) behaviour, which involves input, processing and output of mass, energy, information and data, (iii) interconnectivity, meaning that the different parts of the system have functional and structural relationships to each other (Schreuder, 2014). In the biological context, proteins are systems, so are cells and organs. Moreover, all these entities linked together define an organism as a system too.

Although there is no clear definition of "complex system", there is an understanding that when the single parts of a systems interact, the resulting global system displays emergent collective properties, culminating in a higher order of organization and functions, whose behaviour cannot be reduced to the sum of its parts (Casti *et al.*, 2003). The interactions among the single parts are organized in non-random and non-completely ordered way. Moreover, the system units establish non-linear interactions (meaning that a small perturbation may cause a large effect, a proportional effect, or no effect at all) making the systems chaotic. When dealing with biological systems, we have also to consider that they cannot be treated as isolated entities: noise and stochastic behaviour have to be taken into account.

Omics techniques allow to explore the different single levels of biological complexity. However, it is only through multi-omic data integration that the understanding of each single biological layer can be boosted. Genomics can reveal the involvement of some genes or variations in the development of a disease or phenotype. However, how do these genes lead to the final phenotype? Beyond the investigation at the genome layer, genes/proteins must be analysed in the context of their interactions and in relation to biological processes/pathways, shedding light on the properties of the systems that cannot be derived from the analysis of the isolated elements. Thus, alongside the development/use of experimental techniques, we need to develop computational methods for data interpretation and integration in order to achieve a higher-level biological knowledge.

In the last twenty years of genomic research, a massively amount of data has been produced to decipher, understand and characterize phenotypes. As consequence, NGS based experiments and big data analysis have provided lists of "interesting" variants/genes/proteins that, in the context of functional genomics (Pevsner, 2015), need annotations for been reconciled with known and putatively common biological processes/pathways describing the phenotype. To highlight this "genes/proteins – biological processes/pathways – phenotype" relation, functional interpretation of gene/protein sets is usually done by applying gene enrichment analysis, i.e. a statistical procedure that consist of testing whether a gene/protein set is enriched with certain biological functions (Huang *et al.*, 2009a). Gene enrichment procedures can be classified in two macro-classes: standard and network-based. The latter class take advantage from biological networks, by modelling the complexity of processes occurring in the cell through algorithms exploiting graph features (Junker *et al.*, 2008).

The work I describe in this first part is a novel network-based method for gene enrichment analysis. The method, called NET-GE (network-based gene enrichment analysis; Di Lena *et al.*, 2015; Bovo *et al.*, 2016; Bovo *et al.*, 2017), uses graph- and information- theoretic measures to mine the STRING interactome and build functional modules starting from genes annotated with a shared functional term. The resulting modules are then used to address the problem of the functional association. The web server implementation (Bovo *et al.*, 2016) provides annotations based on the Gene Ontology resource (Gene Ontology Consortium, 2015), the KEGG (Kanehisa *et al.*, 2016) and Reactome (Fabregat *et al.*, 2016) pathways. Moreover, it implements both a standard and a network-based gene enrichment analysis. One peculiarity of NET-GE is the possibility to enrich terms that are not present in the annotations of the starting gene set (and thus not detectable through a standard gene enrichment method). When tested on a OMIM-derived benchmark sets (disease related gene sets), NET-GE was able to enrich for biologically meaningful terms neglected by other methods (Di Lena *et al.*, 2015, Bovo *et al.*, 2017). To prove the ability of NET-GE in dissecting biological complexity, two study cases were investigated (Bovo *et al.*, 2017). Moreover, NET-GE have been recently used in the development of eDGAR, a database collecting and organizing gene/disease associations (Babbi *et al.*, 2017).

Given this brief introduction about NET-GE, in the first chapter I introduce some notions necessary to understand how NET-GE has been built and how it works, such as the concept of protein-protein interactions, their representation as network/graph and some concepts of graph theory. Then, I introduce the concept of gene annotation by reviewing some of the main

databases used in the field. Lastly, the concept of gene enrichment analysis and the main methodologies at the basis of it are presented. In the second chapter I introduce the methods at the basis of NET-GE, the third chapter describes the implementation of NET-GE as web-service, the fourth chapter discusses the benchmarking procedure and the obtained results, the fifth chapter presents some study cases and applications exploiting NET-GE, while in the sixth chapter I introduce eDGAR.

1.2 Protein-Protein Interaction: why is it so important?

Proteins are biological objects that rarely act alone. In fact, a protein is often modulated – in terms of its function and activity – by other proteins it interacts with (Phizicky *et al.,* 1995). Commonly, protein-protein interactions (PPIs) are described as physical contacts among two (or more) proteins: intermolecular forces and steric complementarity among surface patches determine precise patterns of relationships. However, in some cases the term "protein interaction" encompasses also functional and logical interaction events, such as the ones resulting from genetic interactions (De Las *et al.,* 2010). Examples of functional interaction are the presence of proteins in the same pathway (not necessarily involving the physical contact) and the transcriptional relationship, by which a protein (transcription factor) influences the expression of other genes. Different techniques have been devised to catch this type of interaction. Among them: (i) neighbourhood relationships, i.e. functionally related proteins organized very closely in the genomes and likely inherited together during evolution process), (ii) gene fusion events, i.e. single-domain containing proteins in a given genome are joined together in a multidomain protein in another genome), (iii) gene co-expression, i.e. similar pattern of expression between genes and (iv) phylogenetic profiles, i.e. the analysis of the patterns of co-occurrence of different groups of genes in different genomes.

To understand the role of each protein in the cell, PPI studies have become of fundamental importance (Safari-Alighiarloo *et al.,* 2014) and the huge amount of interaction data collected over the years has made it possible to construct several "interactomes", also named protein-protein interaction networks (PPINs). Moreover, thanks to the advent of HTS techniques (such as purification-mass spectrometry, cross-linking MS analysis, MS-based protein correlation profiling and yeast two-hybrid screens; a review of these techniques is presented in Mehta *et al.,* 2016) interactomes are more and more complete and complex. In fact, HTS techniques can determine in a single shot a huge number of interaction events (both at the gene and at the protein level) leading to an increment of the amount of detectable physical and

functional links. However, not all the interactions events discovered by HTS techniques are truly physiological interactions, and their incorporation in an interactome often add a certain degree of noise to the PPIN (von Mering *et al.*, 2002).

Despite the problem of having or not a complete and noiseless PPIN, interactomes proved a discrete success in solving biological problems. PPINs have been used to solve different tasks such as the prediction of protein function and the identification of functional modules (Nabieva *et al.*, 2005, Sharan *et al.*, 2007; Chen *et al.*, 2009, Tripathi *et al.*, 2016). Moreover, PPINs are playing a more and more fundamental role in systems biology and systems medicine in order to elucidate the biological events at the basis of the different phenotypes/diseases.

1.3 STRING as comprehensive PPI network

The Search Tool for the Retrieval of Interacting Genes, formerly named STRING, is a biological database of known and predicted PPIs (Szklarczyk *et al.*, 2010). Established in the year 2000 (Snel *et al.*, 2000), STRING aims at providing a critical assessment and integration of PPIs from different resources by including physical and functional associations.

STRING sources of PPI data can be subdivided in seven channels (type of evidence) (Szklarczyk *et al.*, 2017):

1) the *experiments* channel, collecting interactions experimentally observed in laboratories (bio-chemical/physical and genetic experiments). Sources of PPI are primary databases organized in the IMEx consortium (Orchad *et al.*, 2012) plus BioGRID (Breitkreutz *et al.*, 2008);
2) the *database* channel, where PPI evidences are imported from manually curated pathways databases;
3) the *text-mining* channel, where an association score is given to each pair of proteins frequently mentioned together in the same paper, abstract or even sentence;
4) the *co-expression* channel, where an association score is given to each pair of proteins consistently similar in their expression patterns. Co-expression data are retrieved from gene expression experiments carried out by using microarray and/or RNAseq approaches;
5) the *neighbourhood* channel, where an association score is given to each pair of proteins when they are consistently observed in each other's genome neighbourhood;

6) the *fusion* channel, where an association score is given to each pair of proteins when there is at least one organism in which their respective orthologs have fused into a single, protein-coding gene;
7) the *co-occurrence* channel: where an association score is given to each pair of proteins when their orthologs tend to be observed as "present" or "absent" in the same subsets of phylogenetically related organisms.

The last version of STRING (v.10.5; May 14, 2017) counts 9,643,763 proteins and 1,380,838,440 different interactions, for a total of 2,031 organisms. For each interaction, by integrating the probabilities of the seven channels, STRING provides a combined confidence score (scaled between 0 and 1) representing the estimated likelihood that a given interaction is biologically meaningful, specific and reproducible, given the supporting evidences (Szklarczyk *et al.*, 2017). Confidence limits are given by STRING as following: low confidence, 0.15; medium confidence, 0.4; high confidence, 0.7; highest confidence, 0.9. Usually, the combined score is used to (i) draw PPINs at different confidence levels and (ii) to implement a weighted graph where scores are used to operate on the network. This second usage of the combined score will be briefly discussed in the next sub-chapter.

1.4 Protein-Protein Interactions networks as graphs

A PPIN is usually represented by means of a graph, a mathematical structure used to model different problems. In this structure, nodes represent genes/proteins and edges their interactions.

A graph is usually defined as $G = (V, E)$, where V is the set of vertices representing the nodes $\{V_1, V_2, V_3, \ldots, V_n\}$ and E is a set of edges representing the links among the nodes. An edge is defined as $E = \{(u, v) \mid u, v \in V\}$, where u and v are the connected vertices.

Graphs can be directed or undirected depending on whether an edge direction is provided or not, respectively. In a directed graph, an edge $E = (u, v)$ is directed from u to v and it indicates that it is possible traverse the graph only from u to v and not vice versa (unless also the edge $E = (v, u)$ is present).

Moreover, when edges (or nodes) are labelled with a score, the originated graph is defined "weighted". Generally, a PPIN is an undirected and unweighted graph.

When considering all the different information stored in it, STRING can be treated as a weighted and directed graph. In fact, by incorporating functional events (e.g. activation,

repression) the graph acquires directionality, while the assignment of a confidence score to edges makes it weighted.

To better understand the next paragraphs, I introduce some definitions and proprieties of a graph. A *path* from a source $u \in V$ to a target $v \in V$, or (u,v)-path for short, is an alternating sequence of vertices and edges $u,(u, V_1),V_1,(V_1,V_2),V_1,\ldots,(V_k, v),v$ starting with u and ending with v, such that the vertices before and after an edge are its tail and head, respectively (Brandes, 2008). A path never passes two times over the same edge or node. A graph is *connected* if it is possible to find a path between the all pairs of vertices. The *path length* of an (u, v)-path is the number of edges it contains, and the distance (Ulrik 2008). If the graph is connected, two vertices could be connected by different paths of different lengths.

The *shortest path* between two vertices (**Figure 1**) is defined as the path which minimize the sum of the weights of its constituent edge (in an unweighted graph, the shortest path is the one with the lowest number of edges).

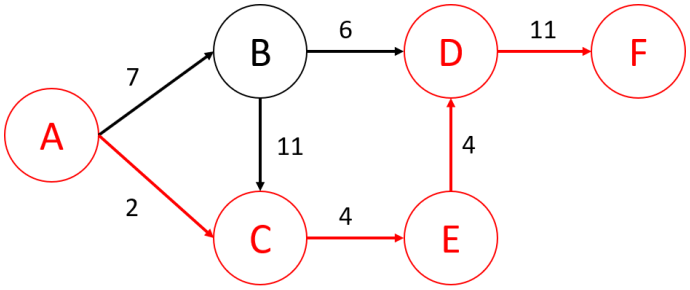


Figure 1. Shortest path between vertices A and F. Considering the proposed weighted directed graph, the shortest path minimizing the sum on the weights w is the ACEDF, with $w = 22$.

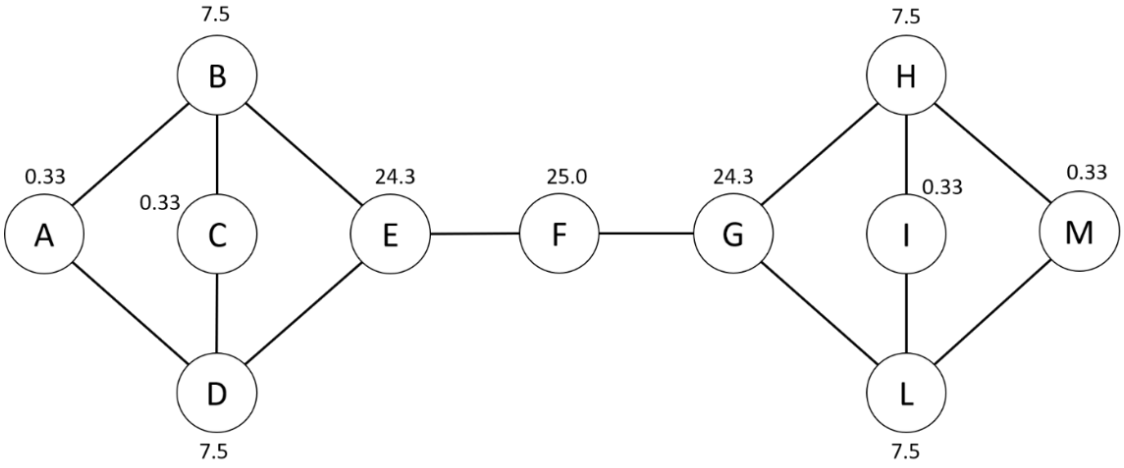


Figure 2. Betweenness centrality (BC) scores. Considering the proposed graph, the vertex *F* has the highest BC since it lies on all the shortest paths among the pairs of vertices.

To estimate the importance of a node for the connectivity (or the information flow) of a graph, different measures of node centrality have been developed. Among them, the betweenness centrality index measures the centrality of a node by considering the shortest paths (Figure 2). Given a node *v* of a graph *G*, its betweenness centrality is defined as:

$$BC(v) = \sum_{(i,j)\in V} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \qquad (1)$$

where, for nodes *i*, *j*, $v \in V$, with $i \neq j \neq v$, σ_{ij} is the number of shortest paths from node *i* to node *j*, and $\sigma_{ij}(v)$ is the number of shortest paths that pass through *v*. The measure is usually interpreted as the degree to which a vertex has control over pairwise connections between other vertices, based on the assumption that the importance of connections is equally divided among all shortest paths for each pair (Ulrik, 2008).

Many other graph-theoretic measures can be computed/extracted when studying a PPIN. Even though we discussed only those that are used in the development of NET-GE, it is clear that the study of a PPIN from a graph-theoretic point of view is necessary to elucidate the biological complexity at the basis of the phenomena under investigation.

1.5 Gene/protein annotation and related databases

Gene annotation databases are commonly used to evaluate the functional properties of genes. Annotating genes/proteins means endow sequences with specific biological features (i.e. the different protein domains, the different functions, biological processes and pathways in which the genes/proteins are involved). However, because experimental investigation is costly and time-consuming, only a small part of protein functions has been experimentally determined. To overcome this problem, many computational methods aimed at predicting protein function have been developed. These methods rely essentially on the expansion of the relatively small number of experimentally determined functions to large collections of proteins (Valencia, 2005) by means of some measure of similarity.

Several databases have been developed to classify genes according their roles in the cell. Among them, KEGG (Kanehisa *et al.*, 2016) and Reactome (Fabregat *et al.*, 2016) collect genes/proteins as well as metabolites, in maps representing the different biochemical pathways. The Gene Ontology (Gene Ontology Consortium, 2015) is a database mainly used to annotate genes with biological processes, functions and the cellular locations. However, differently from the KEGG and the Reactome databases – in which annotation is manually

curated – GO terms can be electronically assigned to a gene/protein. Moreover, alongside the difference in content, these databases vary in size.

In the following subchapter, I briefly introduce the KEGG, REACTOME and GO resources.

1.5.1 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database for the systematic analysis of gene functions. The database consists of 15 main manually curated databases which are categorized into systems, genomic, chemical and health information (Kanehisa *et al.*, 2015, Kanehisa *et al.*, 2016). Due to its comprehensiveness, it makes one of the most attractive databases used in the annotation of gene/protein functions as well as small molecules. Among them, KEGG PATHWAY collects manually drawn pathway maps representing the links among genomic information and higher order functional information. Pathway maps are drawn to represent the dual aspect of the metabolism: the genomic network of how genome-encoded enzymes are connected to catalyse biochemical reactions and the chemical network of how compounds are transformed by means of those enzymes (Kanehisa, 2013). Alongside the canonical biochemical pathways representing the metabolism, KEGG provides maps depicting the processing of genetic and environmental information, cellular processes, and the molecular pathways involved in human diseases. Moreover, KEGG BRITE provides then a functional hierarchy of the KEGG objects.

The last release of KEGG (v.84.0, Dec. 2017) annotate a total of 7,314 human genes in 323 pathways (considering only the lowest level of the hierarchy). It is evident that about the 2/3 of the human genome lacks a KEGG functional annotation.

1.5.2 Reactome

Reactome (Fabregat *et al.*, 2016) is a manually curated database of pathways and processes. It describes biological pathways as chemical reactions that closely mirror the physical interactions occurring in the cell. Reactome provides information about proteins and small molecules and how they participate in pathways to coordinate cellular events. Reactions are grouped into pathways, which in turn are assembled into a hierarchy of biological events. (Croft *et al.*, 2011; Milacic *et al.*, 2012). Like KEGG maps, the REACTOME ones describe canonical biochemical pathways, cellular processes, and the molecular pathways involved in diseases. However, REACTOME differs in the number of maps and in hierarchical structure:

REACTOME confines genes in specific “modules” of more general biochemical pathways. The last version of REACTOME (v.62, Sept. 2017) annotate a total of 10,712 genes in 2,176 pathways. By using REACTOME, it is possible to annotate about half of the human genome.

1.5.3 Gene Ontology

The Gene Ontology (GO; Gene Ontology Consortium 2015) is a controlled vocabulary of functional terms subdivided into three main categories: i) molecular function (MF); ii) cellular component (CC) and biological process (BP). Biological process terms describe biological events (e.g. negative regulation of apoptotic process) accomplished by one or more organized assemblies of molecular functions. Molecular function terms describe activities that occur at the molecular level (e.g. protein kinase binding) and the cellular component terms describe a component of a cell such as an anatomical structure (e.g. cytoplasm) or a gene product group (e.g. ribosome). The last version of the GO resource (Amigo 1.8) annotates more than 90% of human genes with 54,440 biological functions (17,788 GO:BP, 17,734 GO:MF and 18,918 GO:CC). It is evident that this resource functionally annotates the greatest part of the human genome. However, it is worth to note that out of the 442,065 associations, only about the 40% of associations has been experimentally derived.

1.5.4 Hierarchical representation of functional annotations

The three databases (GO, KEGG, Reactome) are structured in a hierarchic way, where parent–child relationships are defined. A parent terms represents a more general concept than its children terms. In the case of KEGG, term-term relationships are represented as a simple tree, while the Reactome and GO databases originate a directed acyclic graph (DAG).

In a DAG, a term can have multiple parents and it is not possible to have a path in graph that starting from a term points back to the same term. An example of a simple tree and DAG is given in **Figure 3**.

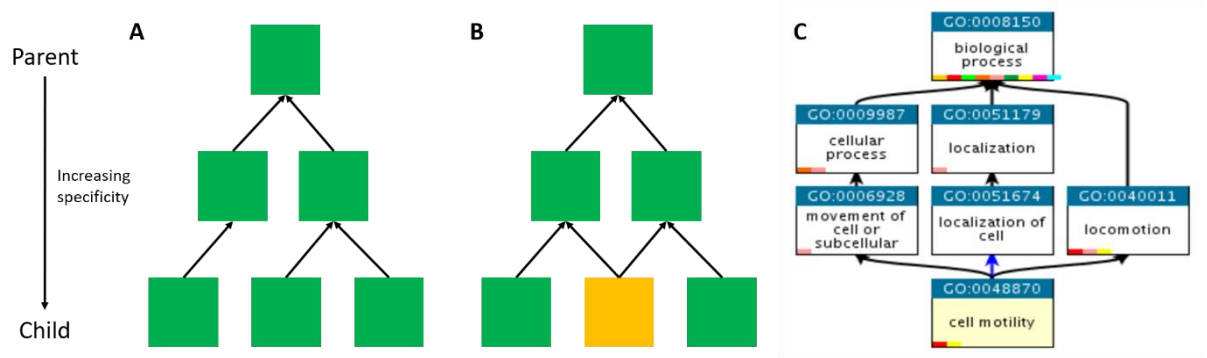


Figure 3. Simple trees and directed acyclic graphs. Both structures are directed graphs in which boxes represent nodes (functional terms) and arrows represent edges. A) An example of a simple tree, in which each child has only one parent. Based on the BRITE hierarchy, KEGG pathways are structured as a simple tree. B) A directed acyclic graph (DAG) in which each child can have one or more parents. The node with multiple parents is highlighted in yellow. Like a simple tree, a DAG has directed edges and does not have cycles (a path cannot start and end at the same node). Reactome and GO terms are structured as a DAG. C) An example of a functional term of the GO:BP branch and its parent-child relationships.

1.6 Quantifying the function specificity and similarity.

How is it possible to understand if a protein function is quite specific or general? When a gene is annotated with a functional term, the associations between the gene and the terms' parents are implicitly inferred (Rhee *et al.*, 2008). Based on these term-term relationships, a way to describe the function specificity of a given term as function of the annotation hierarchy, is to use the number of nodes in the shortest path connecting the term to its root. However, even if the level in the hierarchy is often assumed to be indicative of the terms specificity, this is not completely true. To overcome this problem other measures have been introduced. Among them, the Information Content (IC) score provides an alternate measure of functional specificity (Louie *et al.*, 2010). The IC of a given term *v* is computed as follows:

$$IC(v) = -log_2(p(v)) \qquad (2)$$

where *p(v)* is the relative frequency of occurrence of the term *v* in the gene/protein dataset under consideration. Terms that occur less frequently have higher IC and are assumed to be re specific.

Alongside the function specificity concept, there is the functional similarity one. A gene function can be annotated with different terms, but are these terms similar or different? An

accurate measure of semantic similarity of terms is critical for an accurate measurement of gene functional similarities (Song *et al.,* 2014). Several approaches have been proposed in the last decades for measuring functional similarity. The most common measures used are the Resnik's (Resnik, 1995) and Lin's (Lin, 1998), which are measures relying on the IC score. Resnik's measures similarity between two terms as simply the IC of their most informative common ancestor (MICA), while Lin's considers how distant the terms are from their common ancestor. In this way, the Lin's relate the IC of the MICA to the IC of the terms being compared as following:

$$sim_{Lin}(v_i, v_j) = \frac{2 \times IC\,(v_{MICA})}{IC(v_i) + IC(v_j)} \qquad (3)$$

where $IC(v_i)$, $IC(v_j)$ and $IC(v_{MICA})$ are the information contents of the two terms and their common ancestor, respectively.

1.7 Gene Enrichment Analysis

Omics techniques identify lists of "interesting" genes/proteins characterizing the investigated phenotype. However, given this data, how can we understand the biology at the basis of the phenotype? How is it possible to reconcile genes/proteins with known and putatively shared functional information? Dealing with the interpretation of these gene/protein sets, gene enrichment analysis (also called pathway analysis) has become the most widely used approaches for functional associations, thanks to its ability to provide valuable insights into the biological events underlying a gene/protein set (Tipney *et al.,* 2010).

1.7.1 Gene enrichment analysis: how does it works?

The principal foundation of enrichment analysis is that if a biological event is abnormal in a given study, the co-functioning genes/protein should have a higher (enriched) potential of being selected as a relevant group by the omic technologies (Huang *et al.,* 2009a).

Enrichment analysis is generally performed by mapping genes/proteins to their associated biological annotations (e.g. processes, functions, pathways), comparing the distribution of the target gene set (investigated gen set) against the background distribution of these terms (i.e. the annotated genes) (Tipney *et al.,* 2010). In doing this, enrichment is quantitatively measured by applying different statistical methods, such as the hypergeometric distribution or

its analogous one-sided Fisher's exact test (Huang *et al.*, 2009a). Given a number of genes *N*, with *F* of these genes associated with a particular functional term and *T* of these genes in the target set, then the probability that *b* or more genes from the target set are associated with the functional term is given by the hypergeometric tail:

$$Prob(X \geq b) = \sum_{i=b}^{\min(T,F)} \frac{\binom{T}{i}\binom{N-T}{F-i}}{\binom{N}{F}} \qquad (4)$$

To each term, the Bonferroni or Benjamini-Hotchberg (FDR) procedure is then used to counteract the problem of multiple comparisons (Noble, 2009).

1.7.2 Classification of gene enrichment methods

Given a list of genes, which tool for functional association should be used? Are all the methods comparable? Dealing with the problem of functional association, several gene enrichment analysis tools have been developed over the years. Based on the algorithms on which they rely, they can be divided into three main classes (Huang *et al.*, 2009a): (i) singular enrichment analysis (SEA), (ii) gene set enrichment analysis (GSEA) and (iii) modular enrichment analysis (MEA).

Singular enrichment analysis – Commonly referred as "gene enrichment analisys" or as "Over-Representation Analysis" (ORA), SEA is the most traditional strategy for gene enrichment analysis. Examples of tools implementing the ORA strategy are DAVID (Huang *et al.*, 2009a; Huang *et al.*, 2009b) and GoRILLA (Eden *et al.*, 2009). They are based on statistics like the hypergeometric distribution or its analogous one-sided Fisher's exact test. SEA methods simply assess the over-representation of biological annotations in a pre-selected "candidate" gene list, resulting very efficient in extracting meaningful biological features. However, as drawback SEA methods tend to generate a very large list of enriched terms. Moreover, results are highly impacted by the procedures/methods used to extract the candidate gene (Huang *et al.*, 2009a; Irizarry *et al.*, 2011).

Gene set enrichment analysis - GSEA methods were introduced having in mind the following: is it possible to overcome the limits of using pre-selected "candidate" genes? This powerful approach allows to weight the importance of genes ranking them by some statistics (e.g. the

fold change of differentially expressed genes), by extending the analysis also to those genes not included in the "candidate" set, but marginally contributing to the biology of the phenotype. Thus, GSEA methods result very suitable in the analysis of microarray or RNAseq data, by allowing gene that cannot pass a hypothetical selection threshold (e.g. differential expressed genes with p-value > 0.05) to contribute to the enrichment analysis (Huang *et al.*, 2009a). Differently from ORA methods, GSEA approaches rely on different statistics.

Among the GSEA strategies, GSEA (Subramanian *et al.,* 2005) was the first method to be developed. It is based on a Kolmogorov–Smirnov-like statistic. GSEA computes an enrichment score (ES) that reflects the degree to which a set of genes sharing an annotation term is overrepresented at the extremes (top or bottom) of the whole ranked list of genes. Subsequently, to estimate the statistical significance of the ES, a permutation test is applied. The underlying assumption is that the genes ranked in higher positions, and driving the enrichment procedure, are likely the most contributors to the biology of the phenomena. However, this is not always true since in real biology a small change of some signal transduction genes can result in a larger downstream biological impact.

Modular enrichment analysis - MEA is the only class that takes advantage from the relationships existing between annotation terms when performing enrichment calculations, i.e. using composite annotation terms (join terms). The use of composite annotation terms may therefore be able to provide biological insight lacking in analyses that treat single terms as independent objects (Huang *et al.,* 2009a). In this class we have tool like Ontologizer (Grossmann *et al.,* 2007) and topGO (Alexa *et al.,* 2006).

1.7.3 From standard to network-based gene enrichment

Is it possible to improve methods for functional association? If yes, how? One of the major improvements made in the field of functional association has been the exploiting of information contained in biological networks. Standard methods treat genes/proteins as isolated objects, completely neglecting the functional/physical links among them. However, the analysis of gene sets in the context of their interactions could provide new valuable biological insights (Di Lena *et al.,* 2015). Taking the advantage from this kind of information, a new class of tools, denoted as network-based enrichment analysis tools, has emerged. Based on the strategies and algorithms used to perform enrichment, this new class can be broadly classified into two sub-classes (Di Lena *et al.*, 2015):

A) methods that use the topology of a PPIN to infer how much similar distinct sets of gene/proteins are (e.g. SPIA (Tarca *et al.,* 2009), SEPEA (Thomas *et al.,* 2009), PWEA (Hung *et al.,* 2010), TopoGSA (Glaab et sl., 2010), EnrichNET (Glaab *et al.,* 2012), SANTA (Cornish *et al.,* 2014), JEPETTO (Winterhalter *et al.,* 2014) and TPEA (Yang *et al.,* 2017));

B) methods that identify functionally-related modules in a PPIN and then infer protein/gene biological roles from such modules (e.g. PINA (Wu *et al.,* 2009; Cowley *et al.,* 2012), FunMod (Natale *et al.,* 2014), MetaCORE (Bessarabova *et al.,* 2012)).

In both classes, graph-theoretic measures and graph properties (such as shortest paths, degree, etc.) are used to extract meaningful information from an interactome. Among the publicly available tools, EnrichNet and PINA are the two most cited methods, representative of the A and B classes above, respectively. To give an idea of this new category, we briefly discuss the strategies at the basis of PINA and EnrichNet.

PINA (Protein Interaction Network Analysis) is a web resource developed by the integration of PPIs data from six different databases (IntAct (Kerrien *et al.,* 2006), MINT (Chatr-aryamontri *et al.,* 2007), BioGRID (Breitkreutz *et al.,* 2008), DIP (Salwinski *et al.,* 2004), HPRD (Peri *et al.,* 2003) and MIPS MPact (Güldener *et al.,* 2006)). The core of PINA consists in the identification of functional modules (clusters of densely interconnected nodes) which are likely to represent sets of functionally related proteins. After the module construction, modules are annotated by looking for enriched term coming from four databases (KEGG (Kanehisa *et al.,* 2016), GO (Gene Ontology Consortium, 2015), PFAM (Finn *et al.,* 2016a) and MSigDB (Liberzon *et al.,* 2011)). Given an input set of genes/proteins, it is mapped on the pre-computed modules and the over-represented modules are identified by means of a hypergeometric test. As result, the input gene set is characterized by the significantly enriched annotations of the over-represented modules (Wu *et al.,* 2009; Cowley *et al.,* 2012).

EnrichNet is a web application based on PPIN integrating different information: molecular interactions (STRING, Szklarczyk *et al.,* 2017), cellular pathways (KEGG (Kanehisa *et al.,* 2016), BioCarta (http://www.biocarta.com), WikiPathways (Kutmon *et al.,* 2016), REACTOME (Fabregat *et al.,* 2016), PID (Schaefer *et al.,* 2009)), biological annotations (GO (Gene Ontology Consortium, 2015), InterPro (Finn *et al.,* 2016b)) and tissue-specific gene expression data. The enrichment procedure at the basis of EnrichNET consists of two steps (Glaab *et al.,* 2012): 1) the target genes are mapped on reference datasets in the network meanwhile scoring their distance using a random walk with restart procedure, and 2) the

significance of the distance scores is assessed by using a background model. Significantly annotations are then retained.

2 NET-GE

NET-GE is a novel method for network-based gene enrichment analysis (Di Lena *et al.*, 2015; Bovo *et al.*, 2016; Bovo *et al.*, 2017). Considering the different methods described in the previous chapter, like the methods of the class B (e.g. PINA), NET-GE is based on a pre-processing phase aimed at extracting modules from a PPIN. However, differently from all the other methods, the modules built by NET-GE are function-specific by construction, since each one is construct starting with genes/proteins sharing a specific biological annotation (seed set).

By using graph- and information- theoretic measures, NET-GE builds functional modules by expanding each seed set into a compact and connected subgraph of a PPIN (Di Lena *et al.*, 2015; Bovo *et al.*, 2017). The resulting modules are then used to address the problem of the functional association.

Over-representation analysis is performed by mapping the input gene/protein set on each module, determining through a Fisher's exact test whether there are significant overlaps among the input gene/protein set and the modules. To facilitate the use of NET-GE, a web-application have been released at http://net-ge.biocomp.unibo.it/enrich (Bovo *et al.*, 2017).

The following chapters will introduce the databases, the algorithm and the datasets used to build and test NET-GE. **The full articles Bovo *et al.*, 2016; Bovo *et al.*, 2017; and Babbi *et al.*, 2017 are reported in the appendix of the thesis.**

2.1 Databases

NET-GE relies on the STRING Human Interactome (release 10; http://version10.string-db.org/). After its download, STRING was processed by retaining all the links (with the exclusion of self-loops) with a documented action, irrespectively of the STRING combined score and of the supporting evidence. The resulting network comprised 15,632 nodes and 307,413 links. Modules were also built by using a filtered version of STRING, here named STRING0.9, in which only the links with a STRING combined score ≥ 0.9 were retained. The filtered version comprised 9,422 nodes and 80,112 links. Annotation sets were retrieved from

the Gene Ontology resource (UniProt-GOA human 145 web resource; https://www.ebi.ac.uk/GOA), the KEGG (release 77; http://www.kegg.jp/) and Reactome (release 53; https://reactome.org/) databases.

Genes were associated to each annotation term by means of the UniProtKB accession numbers. Given the hierarchical structure of the annotation features, associations were propagated till the root. The resulting seed sets were then expanded as will be described in the next paragraphs. Statistics about annotation sets are reported in **Table 1**.

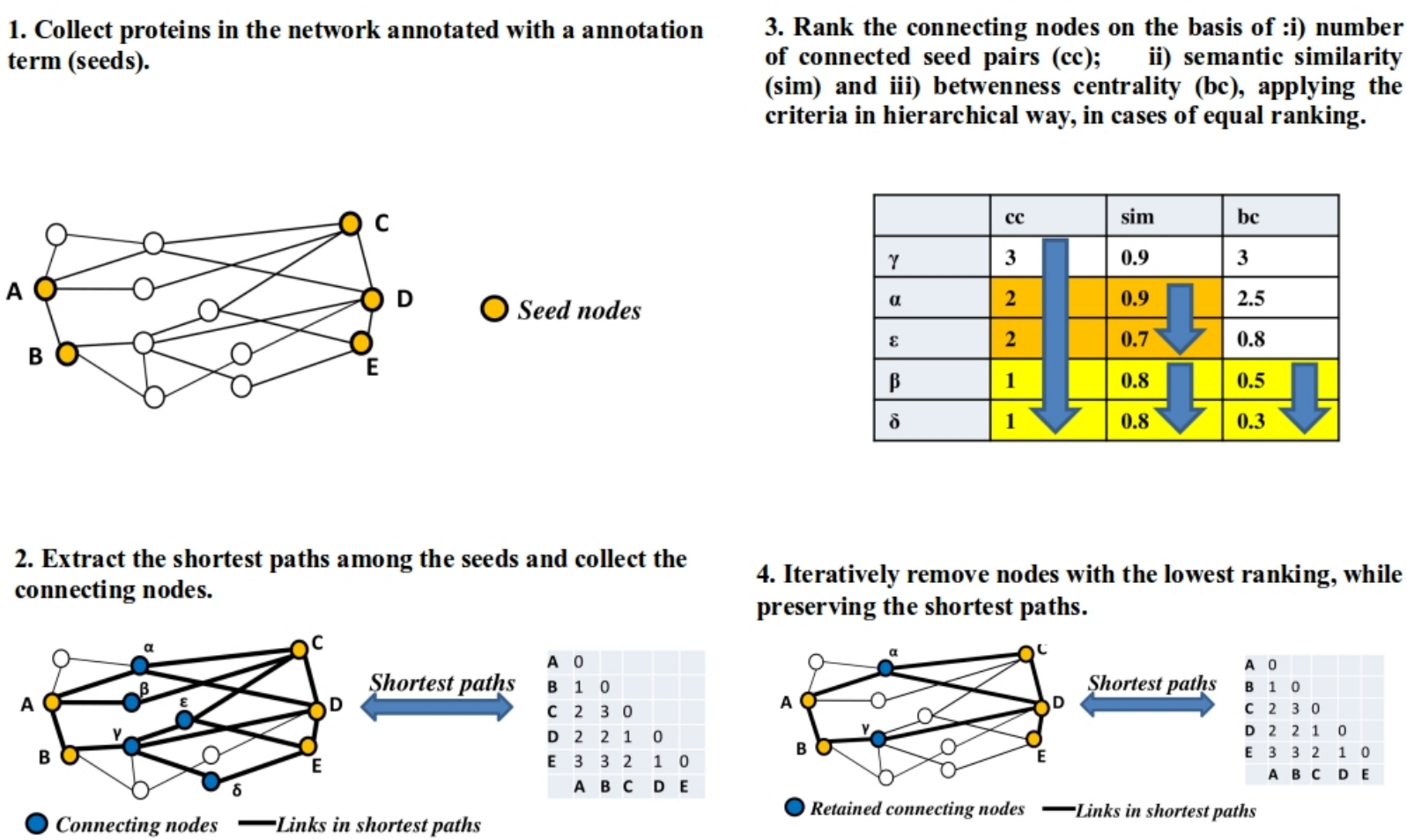
Table 1. NET-GE statistics. Number of annotations and genes are presented.

		GO:BP	GO:MF	GO:CC	KEGG	Reactome
Standard Enrichment						
	Terms	12,783	4,076	1,461	340	1,731
	Genes	18,626	18,254	19,150	6,972	8,093
Network-based (STRING)						
	Genes	17,390	17,499	16,523	9,769	10,103
Network-based (STRING0.9)						
	Genes	17,958	17,282	18,854	7,833	8,708

2.2 Module extraction

The module extraction procedure, aimed at extracting connected and compact subgraphs of the STRING interactome, consisted of four major steps (**Figure 4**). Briefly:

1) all the proteins of the network sharing a specific annotation term were collected into a seed set;
2) each seed set was expanded into a function-specific module by computing the shortest paths among each pair of seed nodes;
3) nodes connecting the seed set were collected and ranked by using graph-theoretic and information-theoretic measures;
4) the module was minimized by filtering out the less informative connecting nodes while preserving the shortest paths.



5)

6)

Figure 4. Outline of the network module generation of NET-GE. The four different steps are highlighted. Adapted from Di Lena *et al.*, 2015.

The ranking step, and the subsequent minimization, are steps aimed at simplifying the sub-network topology by highlighting its main structure. In fact, through the generation of the complete set of shortest paths (step 2), a very large module can be produced, leading to the risk of including nodes that are not representative of the annotation term under expansion.

To overcome this problem, a graph filtering procedure was applied to generate a minimal connecting network (MCN). The graph filtering procedure here applied (step 4) relied on some metrics (step 3) ensuring that the minimized network, when compared against the original one, preserves the number of shortest paths connecting the seed pairs. Details of the steps 2-4 are given as following.

Shortest-path extraction – Given a seed set and the STRING interactome, the all-pairs shortest path algorithm was applied by treating STRING as an undirected and unweighted graph. Self-loops were discarded from the graph and seed proteins not appearing into STRING were kept as isolated nodes in the minimal network.

Ranking procedure – Nodes connecting the seed proteins were ranked on the basis of three measures: 1) seed centrality, 2) semantic similarity with the reference annotation term and 3) betweenness centrality.

These measures were used as primary, secondary and tertiary sort key, respectively. The seed centrality score was adopted as measure of importance of a node. For each node it was computed by counting the number of distinct seed pairs connected by it. In this way, a higher score reflects the probability that such node appears in a subgraph. The semantic similarity score, here defined as the Lin's measure, was used to measure the degree of relationship between the annotation terms and the reference annotation term of each connecting node. We defined the maximum semantic similarity of a connecting node with respect to the reference GO term as the highest Lin's score between the GO terms associated to the connecting node and the reference GO term. The betweenness centrality score measured the importance (centrality) of a node in the networks. Here, this measure was computed by considering the sub-network under analysis. Betweenness centrality was here used to assess the ranking of those connecting nodes presenting the same ranking with respect to the other two scores.

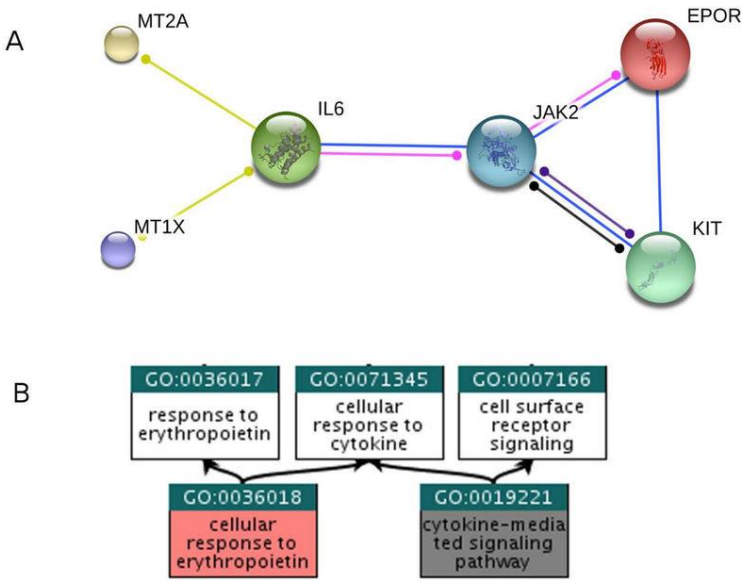


Figure 5. Minimal connecting network for GO:0036018. A) Minimal connecting network extracted from STRING 9.2 (http://www.string-db.org) build for the Biological Process term GO:003601 (cellular response to erythropoietin). The seed genes, directly annotated with GO:0036018, are HGNC:MT2A, HGNC:KIT, HGNC:EPOR and HGNC:MT1X. The connecting genes HGNC:JAK2 and HGNC:IL6, recovered by the minimization procedure, are associated to GO:0019221 (cytokine-mediated signalling pathway). B) Relationships between the reference GO term (GO:0036018) and the GO associated to the connecting genes (GO:0019221). Figure extracted from Di Lena *et al.*, 2015.

Module minimization – Modules were minimized by using the measures above reported. For each module, connecting nodes were ranked from the most to the less informative, and, starting from the last one, they were iteratively removed while preserving the shortest path. One example of minimal connecting network is provided in **Figure 5.**

2.3 Enrichment procedure

NET-GE implements both a standard and a network-based gene enrichment procedure. Entering with a gene/protein set, each gene/protein is mapped into the modules of a selected annotation database. Over-representation is tested through the Fisher's exact test. However, while the standard gene enrichment includes only annotations of the seed nodes, the network-based one includes, for each module, the seeds and their connecting nodes. Multiple testing correction is then applied by using either the Bonferroni or the Benjamini-Hochberg (FDR) procedure (Noble, 2009).

3 NET-GE web server

NET-GE have been released as web-application. The front-end of the web server follows the Model-View-Controller (MVC) paradigm thanks to the web2py framework (http://www.web2py.com/). After submitting the query, the server displays a book-markable page reporting the status of the job. This page is periodically updated and, at the end of the gene enrichment procedure, a link to the results is provided.

The final visualization of the results exploits the Graphviz library (http://www.graphviz.org/) for laying out the acyclic directed graphs for both Gene Ontology, KEGG and REACTOME. Enriched terms are then highlighted. In addition, the web server shows dynamic network renderings, based on the JavaScript library d3.js (http://d3js.org/), for the visualization of the underlying interaction networks involving a specific term. The user can also provide an e-mail address used to e-mail he/she as soon as results are ready.

For multiple submissions, each request is queued and runs as soon as there is available computing power. Running time depends on size of the input set and on the number of functionally related terms.

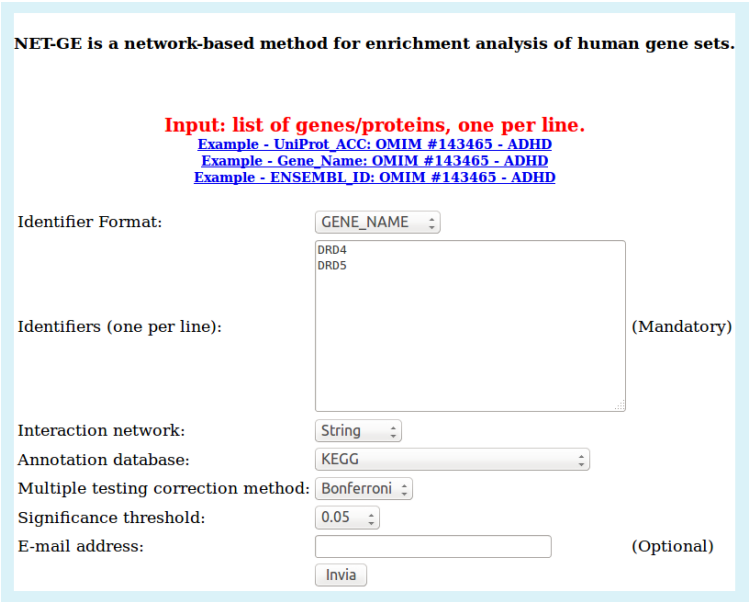


Figure 6. NET-GE web interface. To use the NET-GE web interface, the user is required to perform the following steps: i) choose an identifier format; ii) copy and paste a list of gene/protein identifiers; iii) choose a STRING network; iv) chose the annotation database and v) choose the multiple testing correction procedure and a cut-off. Optionally, the user can enter the e-mail address in order to be notified as soon as results are ready.

3.1 NET-GE web server: input

NET-GE web interface (**Figure 6**) accepts UniProtKB Accession Numbers, Ensembl genes and HGNC gene names. The end user can select: 1) annotation modules based on STRING or STRING0.9; 2) the annotation database (GO terms, KEGG, Reactome); 3) the multiple testing correction method (either the Bonferroni or the FDR based correction) and 4) the significance threshold.

3.2 NET-GE web server: output

The output page reports two tables, one for the standard procedure and another for the network-based one. Both tables list the significantly enriched terms ranked by their adjusted p-value. Each table comprises seven fields: i) the procedure used to enrich the term i.e. standard or network-based enrichment (S and N, respectively), ii) the identifier of the term, linked to the corresponding database, iii) the number and the list of input genes associated to the term, iv) the number and the list of the genes (seeds and/or connecting) annotated with the term, v) the p-value of the association (Bonferroni- or Benjamini-Hochberg- corrected), vi) the name of the term and vii) the term-specific network (only for the network-based enrichment). In the network-based enrichment mode, new terms are highlighted with the symbol N** (**Figure 7**).

By default, the two tables display five significant enriched terms; it is however possible to expand the list of results and visualize all the enriched terms by clicking on the 'Show/Hide all results' button. Results can be downloaded as tab delimited plain text by clicking on the "Download results in tab-delimited format" button.

Moreover, the result page reports a graphical representation of the enriched terms in the context of their relationships (DAG; **Figure 7**).

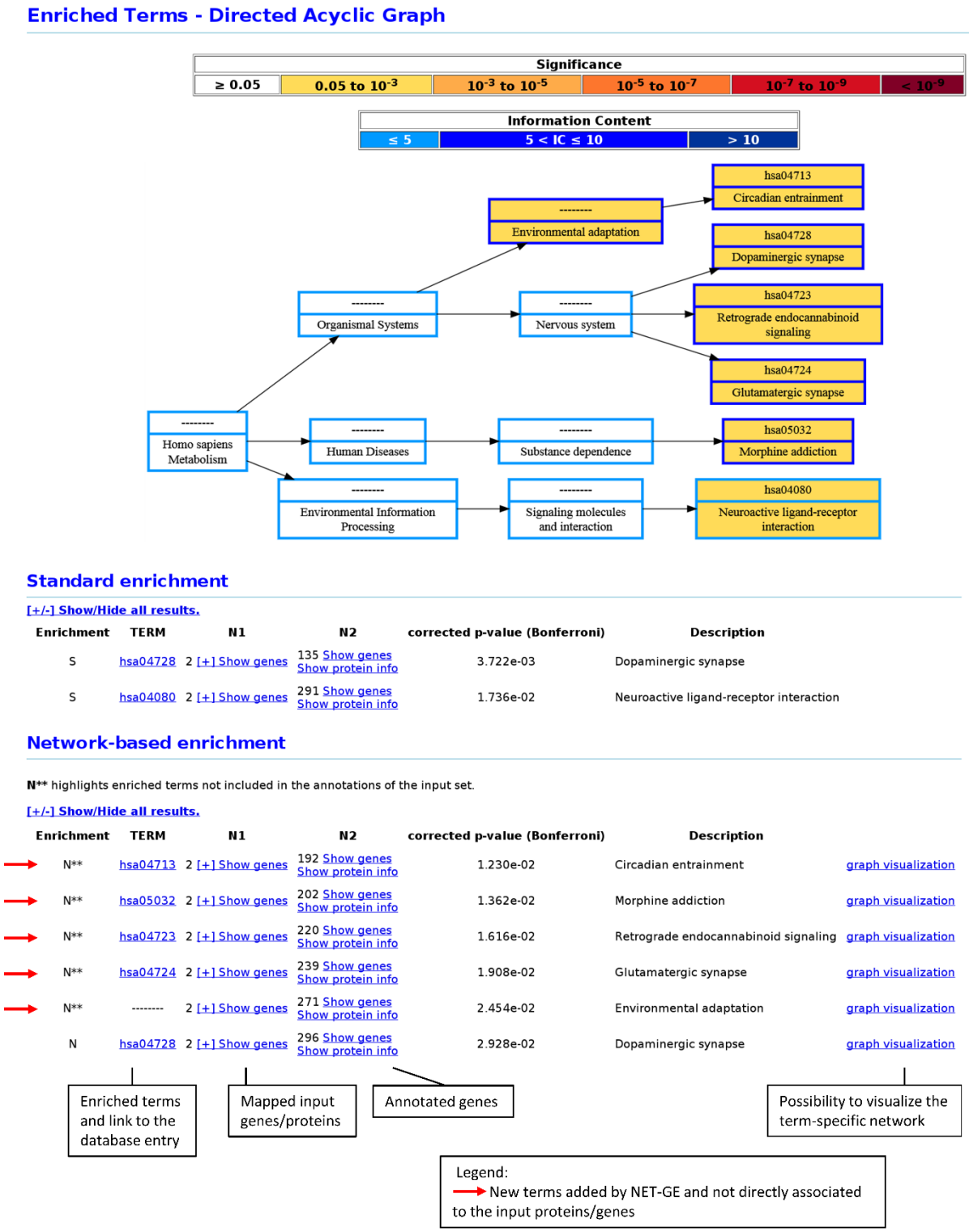


Figure 7. Example of NET-GE output. Results of the ADHD study case (see chapter 5) are here presented as example of the NET-GE analysis over the KEGG database. Results of the standard gene enrichment and of the network-based one are reported in two separated tables. Enriched terms are graphically presented using the DAG structure. The colour code of the box reflects both the degree of enrichment and the information content of the term. Figure extracted from Bovo *et al.*, 2017.

3.3 Graphical representation of the enriched terms

The nodes in the DAG are color-coded according to the significance of the enrichment and to the information content. The graph can be downloaded as image by clicking the "Save image (*.svg)" button. Whenever a term is enriched by the network-based procedure, the term-specific network can be explored by clicking on the "graph visualization" button. Nodes (proteins) in the network are color-coded highlighting seed, connecting and input proteins (**Figure 8**). The term-specific network can be also downloaded as plain text files, where nodes and arcs information are provided.

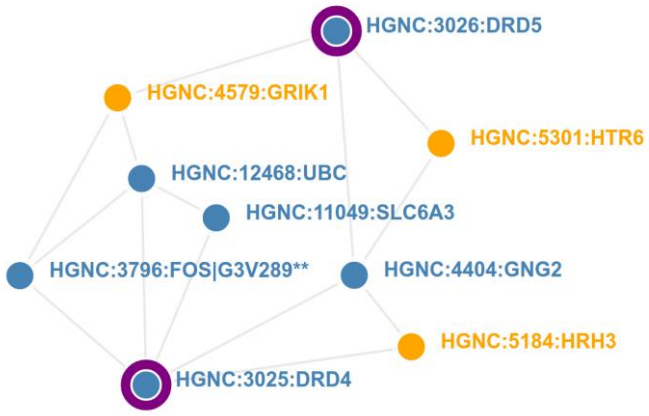


Figure 8. NET-GE module exploration. The graph of the two input proteins (ADHD study case; see chapter 5) and their first neighbours for the new enriched module GO:0014052 (regulation of secretion). Seed and connecting genes are highlighted in orange and blue, respectively. Nodes with a purple border identify the submitted IDs. Being a new enriched term, the two genes enter the module as connecting nodes.

4 Benchmarking the method: a quantitative approach.

NET-GE was benchmarked by using sets of genes involved in mendelian diseases as retrieved from the Online Mendelian Inheritance in Man (OMIM; Hamosh *et al.*, 2005) resource. The dataset comprised 244 OMIM-related gene sets, with a number of protein associated to each disease ranging from 2 to 29, and an average number equal to 4. This dataset was analysed by using both the standard (S) and the network-based (N) methods. Annotation terms with a p-value < 0.05, Bonferroni corrected, were considered over-represented. Both the STRING and STRING0.9 implementations were tested for each

annotation database (GO terms, KEGG and REACTOME pathways), separately. For each OMIM disease we counted:

a) the number of terms enriched by the standard method;

b) the number of terms enriched by the network (new terms excluded);

c) the number of new terms (terms enriched only by the network -based approach and not present in the annotation sets of the input genes/proteins).

While evaluating the network-based approach, we focused only on terms that were not enriched by the standard method, filtering out also all the terms that were ancestors of terms enriched by the standard method.

Considering the three categories above reported, out of the 244 OMIM sets we computed:

a) the number if diseases for which neither the network-based nor the standard method retrieved significantly overrepresented terms (set “n.a.”);

b) the number of diseases for which the standard method retrieved more terms compared to the network-based method (set “S>N”);

c) the number of diseases for which both the methods retrieved an equal number of terms (S=N);

d) the number of diseases for which the network-based procedure enriched more terms compared to the standard method (set “N>S”);

e) the number of diseases for which the network-based procedure added term not included in the annotations of the input set (set “New Terms”).

For each annotation database, results are presented in **Figure 9**, panels A and B for STRING and STRING0.9, respectively. STRING based analyses were characterized by more than 50% of OMIM diseases presenting over-represented terms, considering also the new terms and the terms present equally or in a greater number than the one enriched only by the standard method. By using STRING0.9 similar results were retrieved.

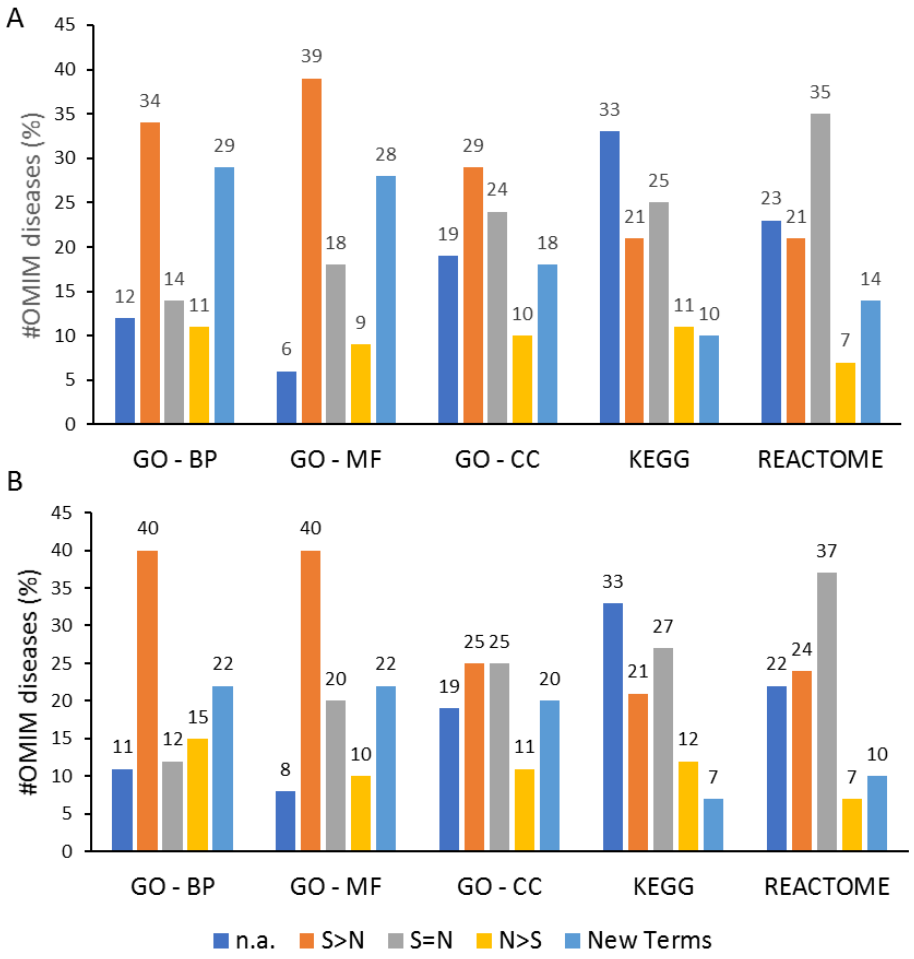


Figure 9. NET-GE benchmark results: A) STRING based and B) STRING0.9 based. The bar plot presents, for each annotation database, the percentage of diseases (out of the 244 tested) for which: neither the network-based nor the standard method retrieved significantly overrepresented terms ("n.a.", blue), the standard method retrieved more terms compared to the network-based one ("S>N"; orange), both the methods retrieved the same number of terms ("S=N"; grey), the network-based approach enriched more terms compared to the standard method ("N>S"; yellow) and finally, the network-based approach added term not included in the annotations of the input set ("New Terms"; light blue).

5 Study cases: a qualitative evaluation of NET-GE.

NET-GE was evaluated by testing two gene sets related to 1) the Attention Deficit Hyperactivity Disorder (ADHD) and 2) the Obsessive-Compulsive Disorder (OCD). While the ADHD represented a hypothetical study case – because genes were retrieved from the OMIM database – the OCD study case represented a real study case because the investigated gene set resulted from the analyses of sequencing data.

5.1 The ADHD study case

ADHD is the most common neuro-behavioural problems afflicting children between 6 and 17 years of age (Sharma *et al.,* 2014). ADHD is considered a chronic neurobehavioral disorder characterized by developmentally inappropriate levels of hyperactivity, impulsivity and inattention (Sharma *et al.,* 2014; Tarver *et al.,* 2014). In the last decade, several studies have supported a strong genetic contribution for ADHD. A review about the genetic of ADHD have been recently published by Bonvicini *et al.,* 2016.

We tested NET-GE considering the ADHD-related genes reported in OMIM (OMIM #143465): the dopamine receptors DRD4 (UniProtKB AC: P21917) and DRD5 (UniProtKB AC: P21918). Enrichment analyses were carried out setting the significance threshold at 0.05 on the Bonferroni corrected p-values. For sake of clarity, we report only the most informative terms, neglecting all the parent terms of the hierarchy. Analyses ran over the GO:BP and KEGG databases.

Focusing on the GO:BP branch, terms enriched by NET-GE are listed in **Table 2**. The standard enrichment statistically over-represented processes strictly related to the disease, such us the involvement of psychiatric functions (cognition, learning), the response to some chemical compounds (amphetamine, cocaine, alkaloids, ammonium ion) and the involvement of the dopaminergic pathway (including the intracellular signal transduction/second messenger pathways of cAMP). As reported in **Table 2**, the network-based approach added terms related to behavioural characters (such as the response to fear) or to the response to other chemical compounds (histamine). Interestingly, the network-based procedure added new terms (not associated to the input protein), unexpectedly involved in ADHD, such as the GABAergic pathway, process experimentally involved in the development of the disorder (Edden *et al.,* 2012).

The standard enrichment over the KEGG database highlighted the involvement of the dopaminergic synapses and the neuroactive ligand-receptor interactions (**Table 3**). The network-based procedure added new terms related to the response to morphine (addiction), the involvement of the glutamatergic synapses, the pathway of retrograde endocannabinoid, and the circadian rhythm system (**Table 3**).

All these processes, although non-characterizing the input proteins, have been previously described in literature as being ADHD-related (Maltezos *et al.,* 2014; Centonze *et al.,* 2009; Gamble *et al.,* 2013; Zhu *et al.,* 2008).

When compared with NET-GE, PINA and EnrichNET did not retrieve any significantly over-represented term/module. However, for reasons of statistical reliability, it is worth to note the EnrichNET authors recommend the analysis of gene sets with at least 10 genes/proteins.

Table 2. ADHD study case: over-represented biological processes.

Enrichment[1]	Term[2]	N1[3]	N2[4]	p-value[5]	Description[6]
S	GO:0001963	2	12	1.38E-04	synaptic transmission, dopaminergic
S	GO:0007212	2	26	6.80E-04	dopamine receptor signaling pathway
S	GO:0001975	2	28	7.91E-04	response to amphetamine
S	GO:0042220	2	34	1.17E-03	response to cocaine
S	GO:0045761	2	61	3.83E-03	regulation of adenylate cyclase activity
S	GO:0007188	2	122	1.54E-02	adenylate cyclase-modulating G-protein coupled receptor signaling pathway
N**	GO:0014052	2	13	2.64E-04	regulation of gamma-aminobutyric acid secretion
N	GO:0034776	2	19	5.79E-04	response to histamine
N	GO:0051954	2	52	4.49E-03	positive regulation of amine transport
N	GO:0001662	2	62	6.40E-03	behavioral fear response
N**	GO:0032228	2	85	1.21E-02	regulation of synaptic transmission, GABAergic
N	GO:0050805	2	96	1.54E-02	negative regulation of synaptic transmission

[1]Enrichment: Standard (S) and Network-based (N). N** indicates new enriched terms not associated to the input proteins; [2]Term: identifier: [3]N1: Input proteins belonging to the term; [4]N2: proteins characterizing the term; [5]p-value: Bonferroni corrected p-value; [5]Description: brief description of the term.

Table 3. ADHD study case: over-represented KEGG pathways.

Enrichment	Term	N1	N2	p-value	Description
S	hsa04728	2	135	3.72E-03	Dopaminergic synapse
S	hsa04080	2	291	1.74E-02	Neuroactive ligand-receptor interaction
N**	hsa04713	2	192	1.23E-02	Circadian entrainment
N**	hsa05032	2	202	1.36E-02	Morphine addiction
N**	hsa04723	2	220	1.62E-02	Retrograde endocannabinoid signaling
N**	hsa04724	2	239	1.91E-02	Glutamatergic synapse
N	hsa04728	2	296	2.93E-02	Dopaminergic synapse

Columns descriptors are given as in Table 2.

5.2 The OCD study case

Obsessive-compulsive disorder (OCD) is a severe neuropsychiatric disorder that can have disabling effects on both adults and children. This disease is characterized by obsessions (intrusive unwanted thoughts and/or images) and/or compulsions (ritualized repetitive behaviours) (Pauls 2010). Several twin-based and family-based studies have provided evidences about the involvement of genetic factors for the expression of OCD. A review about the genetics of OCD have been recently published by Browne *et al.,* 2014.

A more recent study has investigated OCD by using whole-exome sequencing (WES) (Cappi *et al.,* 2016). The authors sequenced twenty OCD cases and their unaffected parents (parent–child trios) looking for *de novo* mutations (i.e. mutation present only in the affected individual). At the end of their analyses, the authors identified 27 genes carry mutations (**Table S1**), that they analysed in the context of their interactions by using the IPA (www.qiagen.com/ingenuity) and DADA (Erten *et al.,* 2011) toolboxes: the former performing gene enrichment analyses, the latter gene prioritization. Among the 27 OCD-related genes, network topology analyses identified WWP1 and SMAD4 as brokers' genes, while the gene enrichment analysis highlighted three over-represented pathways (TGF-β signalling, p-value = 1.3E-25; BMP signalling pathway, p-value = 2.5E-15; glucocorticoid receptor signalling, p-value = 7.1E-12) and the DADA analysis prioritized SMAD4 and WWP1 genes. However, focusing our attention on the results of the IPA pathway enrichment, among the 27 OCD-related genes, only SMAD4 (related to one patient) was present in the three enriched pathways. Thus, we investigated the provided gene set comparing their results with the ones obtained by NET-GE.

Over-representation analysis was initially carried out by NET-GE over the GO:BP branch. The significance threshold was set equal to 0.05, considering Bonferroni corrected results.

Also in this case, we report only the most informative terms, neglecting all the ancestors in the hierarchy. The standard procedure did not find any shared biological process in the input set. The network-based procedure enriched for processes involving the purinergic nucleotides/nucleosides (**Table 4**). Interestingly, the metabolism of purine has been linked to several neurological disorders (Micheli *et al.,* 2011; Moretti *et al.,* 2003; Hines *et al.,* 2014). Moreover, these processes characterized 10 of the 27 initial genes, reconciling common biological features to 9 out of the 14 screened individuals.

Table 4. OCD study case: over-represented biological processes.

Enrichment	Term	N1	N2	p-value	Description
N	GO:0009207	10	1506	1.51E-02	purine ribonucleoside triphosphate catabolic process
N	GO:0046130	10	1559	2.05E-02	purine ribonucleoside catabolic process
N	GO:0042454	10	1585	2.37E-02	ribonucleoside catabolic process

Columns descriptors are given as in Table 2.

Table 5. OCD study case: over-represented molecular functions.

Enrichment	Term	N1	N2	p-value	Description
S	GO:0016887	5	392	1.36E-02	ATPase activity
N**	GO:0004800	2	8	8.50E-03	thyroxine 5'-deiodinase activity
N	GO:0019904	9	1988	2.69E-02	protein domain specific binding

Columns descriptors are given as in Table 2.

To broadly understand the molecular functions representing the gene set, we moved over the GO:MF branch. The standard procedure statistically over-represented the "ATPase activity" while the network-based procedure enriched for the term "thyroxine 5'-deiodinase activity" (**Table 5**). Proteins annotated with this activity (i.e. deiodinases, also named as DIOs) are involved in the regulation of the thyroid hormone activity. As reviewed by Mermi *et al.,* 2016, several studies investigated the role of thyroid glands in OCD, even though contradictory results have been pointed out due to the investigation of patients affected by other psychiatric symptoms (comorbidity). However, the authors investigated a group of patients who had not any comorbid condition, pointing out that thyroid hormone alterations may be associated with occurrence or maintenance of OCD. By using NET-GE, the SMAD4 and WWP1 genes mapped on the "thyroxine 5'-deiodinase activity". Among their first neighbours we had the DIO1, DIO3 and UBC genes (all seed nodes, except UBC). Interestingly, via SMADs, DIO3 is transcribed upon TGF-β stimulation (Huang *et al.,* 2005). Moreover, the first step of thyroid hormone action is the activation of thyroxine by the outer ring deiodination, mechanism promoted by ubiquitin (UBC) (Egri P *et al.,* 2014). More interestingly, WWP1 is involved in the direct transfer of the ubiquitin to targeted substrates. Thus, the identification of this pathway by NET-GE seems to complement the finding about the TGF-β signalling pathway reported in Cappi *et al.,* 2016.

Overall, by considering the GO:BP and GO:MF over-represented terms, our results seem elucidate a more global vision of the disorder.

6 NET-GE and eDGAR: molecular signatures of diseases

The complex nature of the association between genes and diseases is one of the major challenges of Precision Medicine programs. The molecular mechanisms at the basis of the pathogenesis are often uncharacterized. Thus, the investigation of functional relationships among genes involved in the same disease may give fundamental indications about the disease development. eDGAR (Babbi *et al.*, 2017) is a database collecting and organizing gene-disease associations data as retrieved from OMIM (Hamosh *et al.*, 2005), Humsavar (UniProt Consortium) and ClinVar (Landrum *et al.*, 2016).

The database is aimed at providing a comprehensive knowledge of the molecular signatures at the basis of diseases. eDGAR lists 2,672 diseases related to 3,658 different genes, for a total of 5,729 gene-disease associations. A total of 2,051 diseases are monogenic (associated to just one gene) while 621 are polygenic (associated to multiple genes). For each gene set, eDGAR provides information about: (i) interactions as retrieved from PDB (Berman *et al.*, 2000), BIOGRID and STRING; (ii) co-occurrence in stable and functional structural complexes; (iii) shared GO terms, KEGG and REACTOME pathways; (iv) enriched functional annotations, (v) regulatory interactions as derived from TRRUST (Han *et al.*, 2017) and vi) localization on chromosomes and/or co-localisation in neighbouring loci.

NET-GE functional enrichment was used to enhance the understanding of the biological processes and pathways playing a role in the development of the different pathologies, by exploiting its ability in detecting functional terms not present in the list of annotations of a target gene/protein set.

6.1 Functional relationships of disease-related genes

Given a set of disease-related genes, do they share some biological feature explaining the finial phenotype? In eDGAR, the large majority of diseases (from 94.4% to 97.3%, depending on the sub-ontology) is associated with at least one pair of genes sharing GO terms. More than 90% of all the possible gene pairs involved in the same disease have common GO:BP and GO:CC terms; the percentage is somehow smaller (76%) for the GO:MF branch. The total number of GO annotations shared by pairs of genes for GO:BP, GO:MF and GO:CC is 72,787 (unique terms: 4,582), 13,113 (unique terms: 915) and 16,298 (unique terms: 656),

respectively. These data confirm the notion that genes associated with the same disease share some level of functional similarity (Oti *et al.*, 2007).

However, being GO terms organized in a DAG, the information conveyed by the shared annotations (measured by means of IC) can be very different (going from very general to very specific terms). Our dataset is characterized by a IC value ranging from 0 (the root terms) to 10 (the most specific term). The average IC values for MF, BP and CC shared terms are 5.8 ± 1.7, 5.9 ± 1.7, and 5.8 ± 1.9, respectively. For each disease, the specificity of the annotation is evaluated by extracting the best IC values among the GO terms shared by pairs of co-associated genes (**Figure 10**, panel A). For all the sub-ontologies, the best IC values are very spread, and it is evident that on average the most specific terms (highest IC values) belong to the BP sub-ontology: genes pairs sharing GO:BP, GO:MF and GO:CC terms with IC ≥ 5 are present in 72%, 49% and 46% of the diseases, respectively (**Figure 10**, panel A).

By using NET-GE, for the majority of diseases it is possible to statistically enrich GO terms and pathways. The total number of GO annotations enriched for polygenic diseases is 17,029, 4,851 and 3,910 with average IC values 6.1 ± 1.8, 7.1 ± 2, and 6.4 ± 2 for GO:BP, GO:MF and GO:CC respectively (**Figure 10**, panel B). What emerge is that the statistically validated terms further support the notion that genes shard by a disease have some level of functional similarity. Moreover, what NET-GE highlight is the involvement of informative functional terms (higher IC).

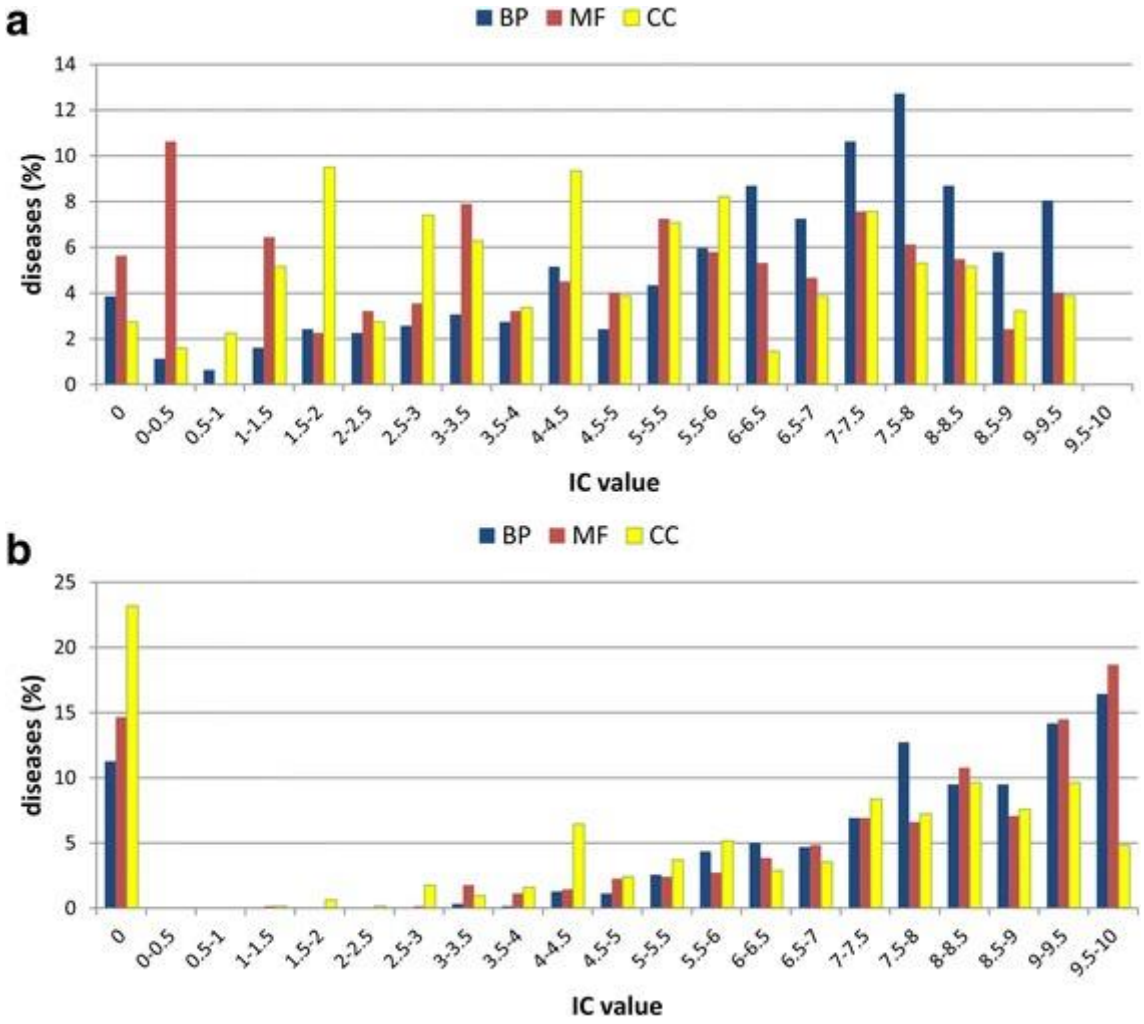


Figure 10. Distribution of best IC values of GO terms for genes involved in multigenic diseases. **a**) GO terms shared by genes; **b**) GO terms after enrichment with NET-GE. For each multigenic disease, IC values of gene-associated GO terms (of the three different roots) are evaluated (Eq. 2). In the figure, the highest IC for each disease is shown. The frequency is computed with respect to the total number of multigenic diseases (621). When IC = 0, genes associated with multigenic disease do not share or enrich GO terms (panel **a** and **b** respectively). Figure taken from Babbi et al., 2017.

6.2 Hypoparathyroidism as case study

The comprehensiveness of eDGAR can be highlighted by examining the following disease: Hypoparathyroidism. This disease is an endocrine deficiency characterized by low serum calcium levels, elevated serum phosphorus levels, and absent/low levels of parathyroid hormone (PTH) in blood (Bilezikian *et al.*, 2011). In eDGAR, the familial isolated hypoparathyroidism (OMIM #146200) is associated with three different genes: GCM2 and PTH (a probable transcriptional regulator and the parathyroid hormone, respectively; both reported in OMIM, ClinVar and Humsavar) and CASR (an extracellular calcium-sensing receptor; reported only in ClinVar).

Gene enrichment analysis over the GO:BP branch highlights two new terms: “regulation of amino acid transport” and “negative regulation of muscle contraction”. Interestingly, these new annotations are related to the severe symptoms of hypothyroidisms, namely tetany and seizure (Shoback 2008).

The enrichment analysis over the KEGG database highlights unexpected terms such as “Circadian entrainment”, “Inflammatory mediator regulation of TRP channels”, “Gap junction” and “Insulin secretion”. Impairments of both the circadian rhythms and the insulin secretion process have been reported in patients affected by hypoparathyroidism (Bauer *et al.*, 1992; Yang *et al.*, 2015).

Taken together, these results highlight: (i) the usefulness of a database integrating data from different sources and (ii) the superiority of the network-based gene enrichment analysis, that coupled together enhance the understanding of biological complexity.

7 Conclusions

Here, I presented NET-GE (**Figure 11**), a tool developed for tackling the problem of the human biological complexity. Specifically, NET-GE is a tool for associating biological processes, functions and pathways to sets human genes/proteins of interest. NET-GE allows both standard and network-based gene enrichment analysis, considering genes as isolated object in the standard mode, and exploiting the STRING interactome to better provide valuable biological insights in the network-based mode

In the last years, several different network-based tools have been implemented. However, differently from other methods, NET-GE explores the STRING interactome by building

modules, function-specific by construction, derived from genes annotated with a specific biological feature.

The method has been benchmarked by performing qualitative and quantitative analyses, considering 244 disease-related gene sets coming from OMIM. NET-GE was able to enrich for terms neglected by the standard methods and not originally present in the annotation of the starting gene/protein set.

Given the possibility to work with small input sets and the ability to over-represent new terms with a valuable biological insight, NET-GE results suitable for deciphering the biological complexity and helping the formulation of new hypotheses on the biological events underlying a phenotype. Given this ability, it has been used in the development of eDGAR, a database collecting and organizing gene-disease associations data as retrieved from OMIM, Humsavar and ClinVar. The use of NET-GE in the analysis of multigenic diseases extended the comprehension of diseases from a biological (functional) point of view.

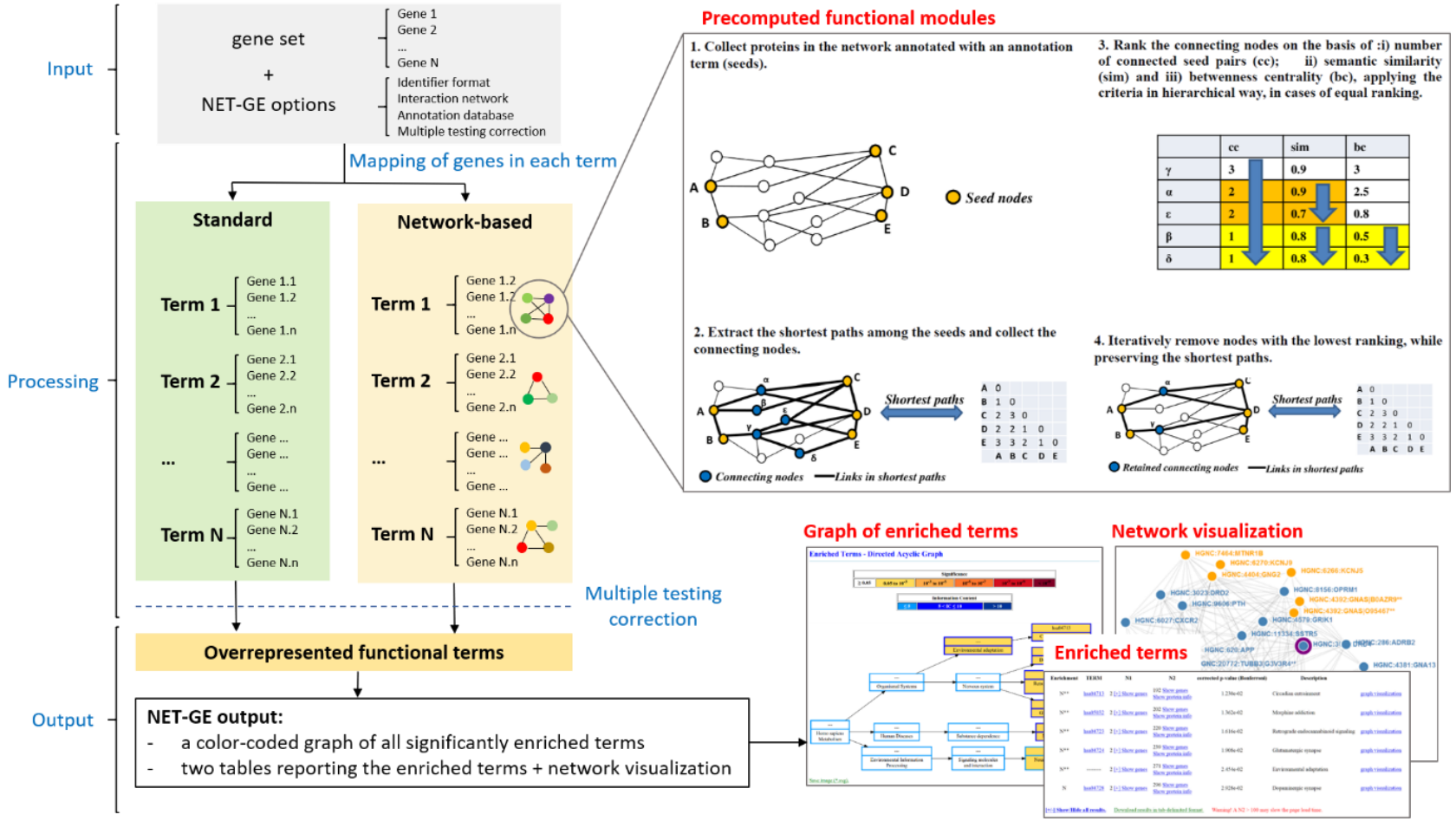


Figure 11. **Schematic representation of the NET-GE analysis procedure.** After the submission of a gene set and the selection of the analysis parameters, both a standard and a network-based analysis is performed (the network-based is based on the precomputed expanded functional modules). As results NET-GE returns two lists of enrichment terms (one for each enrichment mode), their representation as DAG and the possibility to explore the functional modules. Figure adapted from Bovo *et al.*, 2016.

Part II

Rethinking metabolite enrichment analysis: NET-GE^{M}

8 Introduction: why metabolomics?

The last years have seen the development of Metabolomics, thanks to different 'omic technologies aimed at investigating the repertory of metabolites (the chemical compounds that are transformed during metabolism) of a biological sample (Hollywood *et al.,* 2006). By exploring cell complexity at the metabolite level, Metabolomics aims at providing a functional readout of the changes determined by genetic blueprint, regulation, protein abundance/modification and environmental modification (Altmaier *et al.*, 2008).

In the series of the 'omics techniques, genomics, proteomics and transcriptomics result highly valuable, but merely indicate the potential cause for phenotypical response. Instead, metabolomics provides details on potentially effected pathways, because metabolite concentrations differences provide a closest link to the phenotypical response (Wishart et al., 2009). Being the metabolites the downstream products of genomic, transcriptomic and proteomic processes, they represent an "intermediate phenotype" more closely related the final phenotype (i.e. clinical endpoint/disease) than genes and proteins (Gieger et al., 2011). As a consequence, metabolomics is getting more and more adopted as potential source of biomarkers (Vinayavekhin *et al.,* 2010).

Typically, a metabolomics experiment (case/control experiment) ends up with a list of differentially abundant metabolites. Over the years, different experimental approaches and technologies have been adopted in metabolomics (e.g. targeted or non-targeted, Mass Spectrometry (MS) based or Nuclear Magnetic Resonance (NMR) based; for a recent review about metabolomics procedures and technique see Jacob *et al.,* 2017). However, like the other omics-based experiments, a post-processing of data is necessary: metabolites need annotations for reconciling them with known and putatively shared biological pathways describing the phenotype. Routinely, the metabolite complexity is investigated by exploring the link to the functional space. However, as discussed in the first part of this thesis, to gain a global overview of the different processes shaping the final phenotype/endpoint, biological data should be combined. In this perspective, we developed NET-GE^{M}, a version of NET-GE rewritten to work with metabolites.

In this second part of the thesis, I briefly discuss the implementation and results of NET-GE^{M}. **The work has not been published yet and it is still under development. Only preliminary results are available.** Here I give only a brief overview of the findings when considering over-representation of the KEGG pathways.

8.1 How to functionally interpret metabolomic data

Given a list of small molecules, how is it possible to reconcile them with shared biological features? Several tools for the functional interpretation of metabolomics experiments have been developed. They can be broadly divided in two classes: (A) tools that allow only a mapping and representation of metabolites over metabolic pathways (without apply any statistics) and (B) tools performing metabolite enrichment analysis. The principle at the basis of metabolite enrichment analysis is the same of the gene enrichment analysis. Class B tools are generally preferred since a statistical validation is provided.

Moreover, methods for metabolite enrichment can be subdivided in two groups: (i) methods that perform a classic Over-Representation Analysis (ORA) (e.g. MetPa (Xia *et al.*, 2010a), MBROLE2.0 (López-Ibáñez *et al.,* 2016)) and (ii) methods that incorporate also a numeric value associated to each metabolite, known as Metabolite Set Enrichment Analysis (MSEA) or Quantitative Enrichment Analysis (QEA) methods (e.g. MSEA (Xia *et al.*, 2010b), MPEA (Kankainen *et al.,* 2011), MetPa (Xia *et al.,* 2010), IMPaLa (Kamburov *et al.*, 2011)). In particular, tools like MetPa and IMPaLa carry additional features. MetPA provides a "pathway impact score" for each pathway, by performing a pathway topological analysis based on centrality measures (betweenness centrality and out degree centrality) of a metabolite in a given metabolic network (Xia *et al.,* 2010). IMPaLA integrates pathway analysis of metabolomics data alongside with gene expression or protein abundance data (Kamburov *et al.*, 2011).

8.1.1 Metabolite-related databases: a brief overview.

Tools for the functional interpretation of metabolites sets rely on databases broadly classified into two main classes: (A) databases providing a chemical classification of molecules, such as the Human Metabolome Database (HMDB; Wishart *et al.,* 2013), CheEBI (Hastings *et al.,* 2013) and PubChem (Kim *et al.,* 2016), and (B) databases providing a functional annotation such as KEGG, Reactome or BioCyc (Caspi *et al.*, 2016).

Among them, the HMDB resource is the most comprehensive database of human metabolites. Aside chemical information, HMDB provides also clinical and molecular biology/biochemistry data. HMDB users can easily retrieve information about the link between a metabolite and the associated diseases (e.g. via an external link to OMIM), the related biological pathways/processes (e.g. via an external link to KEGG) and the

enzymes/proteins controlling/linked to the metabolite (e.g. via an external link to UniProtKB). HMDB provides also links to other metabolites databases, such as KEGG COMPOUND. This branch of the KEGG resource collects and catalogues small molecules, biopolymers, and other chemical compounds relevant to biological systems.

8.2 NET-GE^{M}: rethinking metabolite functional association.

Databases like KEGG PATHWAYS usually treat metabolites as objects confined in some specific pathways. However, since metabolites are strictly connected to proteins, they can influence other biological pathways due to the interactions occurring between proteins. This means that relations among metabolites can be derived from genes/proteins interactomes and this procedure establishes new links among pathways. Given the strict relation among genes, proteins and metabolites, we explored the metabolome by exploiting the features of an interactome. In order to do that, we developed NET-GE^{M}, a version of NET-GE rewritten to work with metabolites.

8.2.1 NET-GE^{M}: how does it work?

Instead of performing a standard metabolite enrichment analysis, NET-GE^{M} performs a gene-based metabolite enrichment analysis. From the HMDB resource we retrieved the mapping table "metabolite-gene" (HMDB ID to UniProtKB; the mapping table is further discussed).

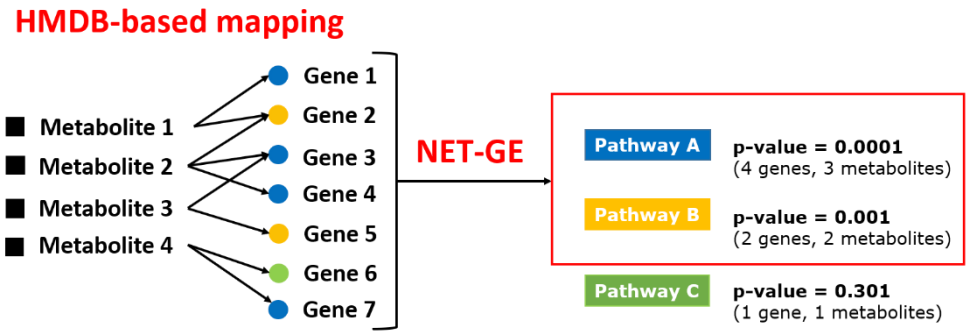


Figure 12. Schematic representation of the NET-GE^{M} enrichment procedure. By using the mapping table provided by the HMDB resource, metabolites are linked to genes. Then, enrichment analysis is performed using the metabolite-related genes. Over-represented term characterized by at least 2 different genes linked to at least two different input metabolites (HMDB IDs) are retained.

Entering NET-GE^{M} with a set of metabolites, compounds are mapped on the genes acting on them (via the mapping table) and ORA is successively performed considering genes instead metabolites (**Figure 12**).

8.2.2 Implementing NET-GE^{M}: defining a metabolite-gene mapping table

NET-GE^{M} relies on a metabolite-protein and the gene-process associations as derived from HMDB and NET-GE, respectively. To make the results of NET-GE^{M} comparable with the results of a standard metabolite enrichment tools, like MBROLE2.0, the metabolite-protein mapping table was pruned by removing the metabolites not available in MBROLE2.0. The metabolite-protein mapping table initially accounted for a total of 22,133 HMDB IDs (primary IDs) linked to 5,631 UniProtKB accession numbers (ACs). Out of the 5,631 UniProtKB ACs, 5,568 (99,9%) are annotated with NET-GE functional sets. The final mapping table links 22,130 metabolites (HMDB IDs) to 5,358 genes.

8.2.3 Why genes instead of metabolites? A critical view

Given the metabolite-gene mapping table, we analysed both the number of HMDB IDs related to each gene and the number of genes associated to each HMDB ID. Distributions are shown in **Figure 13** panel A and B, respectively. The distribution of the number of genes associated to each HMDB ID highlights a pick characterized by 16,852 HMDB. This pick is composed of a set of 13,631 HMDB IDs (over 22,133 HMDBs IDs; ~ 62% of the dataset) associated to the same group comprising 34 genes.

An over-representation analysis (by using NET-GE) of these 34 genes over the KEGG pathways database highlighted the involvement of different lipid metabolisms (data not shown). The 13,631 HMDB IDs are mainly lipids of different classes sharing a common backbone and characterized by a different rearrangement of the C=C double bonds. When mapped over the KEGG compounds database, these molecules were grouped under shared identifiers.

An example is the “phosphatidic acid”, identified in KEGG as the compound C00416, which is linked in HMDB to 17 compounds. By considering this simple example, we have to think at the impact that different “nomenclatures” can have on the development of a tool for metabolite enrichment analysis.

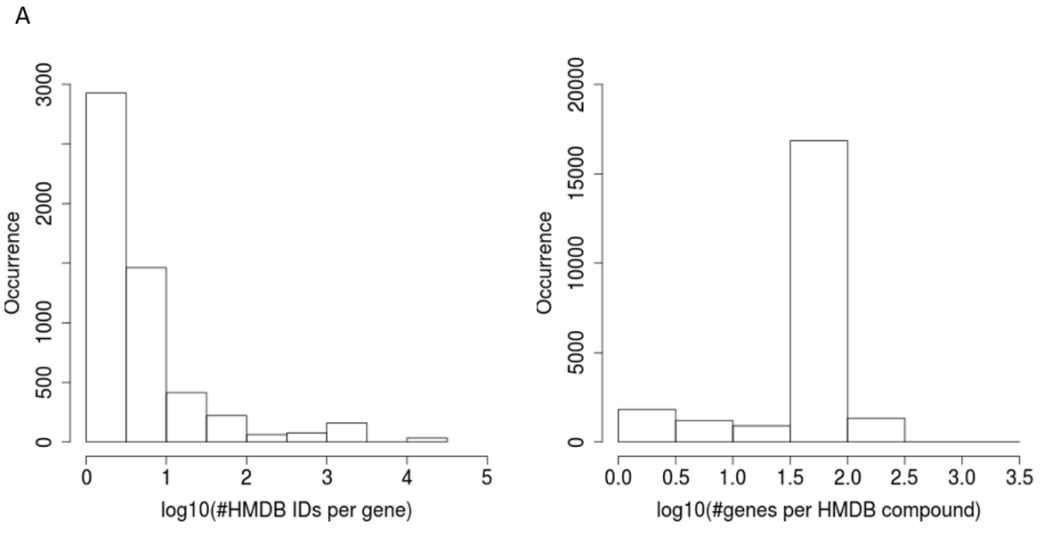


Figure 13. Distribution of A) number of HMDB IDs per gene, and B) number of genes per HMDB ID. The second graph (**B**) highlights a pick mainly characterized by a set of 34 genes associated with 13,631 HMDB IDs.

For example, the KEGG pathway hsa00561 (Glycerolipid metabolism) counts 29 KEGG linked to 154 HMDB IDS. The proportion of compounds included in this pathway with respect to the total number of compound used as background (3586 KEGG compounds or the 22,133 HMDB IDS), lead to different p-values when performing the Fisher's exact test and determine different sets of statistically enriched functional annotations. This example illustrates the problem of the best metabolite representation to be used to build a tool for ORA in metabolomics. In other words, representing metabolites with compound ontologies or chemical classes, as in standard tools, can be problematic on the statistical analysis point of view. NET-GEM tries to overcome this issue by referring to a more stable underlying layer, that is by representing each compound in terms of the genes involved in its metabolisms and their interactions. Another reason for that choice is the fact that the metabolome is still under discovery and a stable tool based on compound ontologies cannot be easily developed, as discovery of new metabolites can lead to an additional expansion and imbalance of annotations. With NET-GE^{M} we counteract the problem of stability by using genes instead metabolites, since the last 20 years of genetics and genomics have provided a stable number of genes having a curated annotation. Moreover, since gene annotations seem stable enough, a tool relying only on metabolite-gene annotations can be "easily" updated

8.2.4 The statistics at the basis of NET-GE^{M} and MBROLE2.0

The Fisher's exact test was used to carry out enrichment analysis over the KEGG pathways database, both in MBROLE2.0 and in NET-GE^{M}. Analyses were carried out entering a list of HMDB IDs in both the systems. The no. of metabolites in MBROLE2.0 and the no. of genes in NET-GE, were used as background. The Benjamini-Hochberg procedure was adopted to counteract the problem of multiple testing. Results of the standard and network-based procedures were merged together. Only KEGG pathways with a FDR < 0.001 are considered significantly over-represented. When performing ORA with NET-GE^{M}, only over-represented terms characterized by at least 2 different genes linked to at least two different input metabolites are retained.

8.2.5 Testing the method: the Parkinson's disease case study

To evaluate NET-GE^{M} we used a dataset of 54 metabolites related to Parkinson's disease (PD), a neurodegenerative disorder characterised mainly by tremors and other motor symptoms caused by the formation of Lewy bodies and a loss of dopaminergic neurons in the substantia nigra (Kalia *et al.,* 2015). The 54 metabolites were collected as KEGG compounds by looking for metabolite-disease associations published in PubMed in the 2006-2016 decade as described in Kori *et al.,* 2016 (**Table S2**).

A total of 41 KEGG compounds (corresponding to 41 HMDB IDs) are available and linked to 475 genes in our mapping table. As we highlighted above, some compounds are missing since we did not use the updated version of the database, while other have not a link to a gene for biological reasons (like the compound “creatinine” which results from a spontaneous dehydration of “creatine”, not involving any enzyme).

The 41 HMDB IDs were used as input elements both in NET-GE^{M} and in MBROLE2.0. The results of NET-GE^{M} were compared against the results of MBROLE2.0. Moreover, to further validate the over-represented terms not retrievable by MBROLE2.0, we looked in PubMed for experimental evidences.

To understand how metabolomics is complementary to genomics, we also analysed via NET-GE a gene set of 98 PD-related genes (genes supported by at least 5 publications) as retrieved from Phenopedia (Yu *et al.,* 2010).

Over the KEGG PATHWAYS database, MBROLE2.0 detects only 3 over-represented pathways. NET-GE^{M} enriches for a total of 67 pathways, while NET-GE detects 33 pathways. **Figure 14** presents the Venn diagram produced by the three different methods.

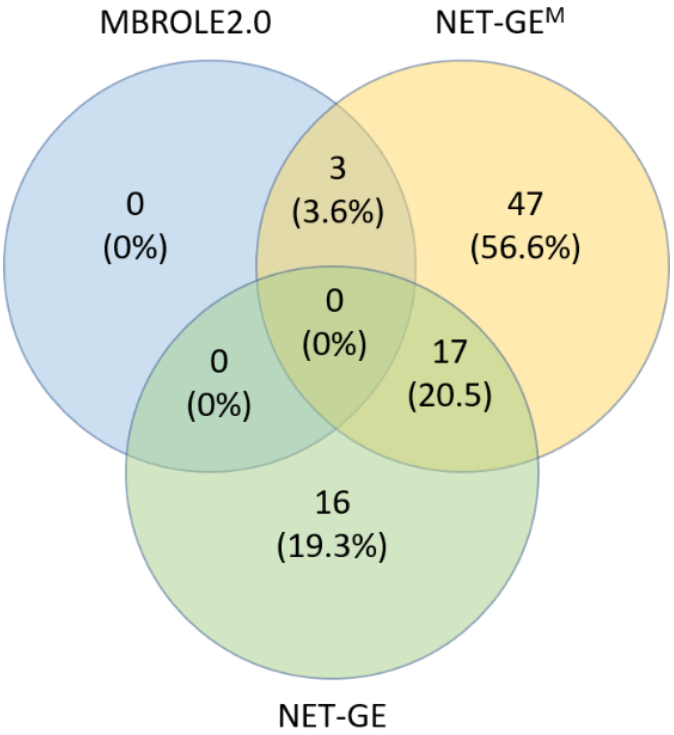


Figure 14. Venn diagram of the pathways over-represented by the three different methods.

The 3 pathways detected by MBROLE2.0 (the "Alanine, aspartate and glutamate metabolism", the "Citrate cycle" and the "Glyoxylate and dicarboxylate metabolism") are comprised in the set of enriched terms detected by NET-GE^M.

Compared to MBROLE2.0, NET-GE^M adds 64 pathways: 6 enriched only by the standard method (S), 56 enriched both by the standard and network-based method (N/S), and 2 added only by using the network -based (N).

The terms exclusively enriched by NET-GE^M (47 leaf terms, **Table 6**) point to pathways of great interest. A first pathway involves the GABAergic and Glutamatergic synapses. In fact, dopamine can modulate glutamatergic transmission by the convergence effect onto medium spiny neurons, by acting on D2-R located pre-synaptically on glutamatergic inputs or by modulating excitatory inputs onto GABAergic and cholinergic interneurons (Gardoni *et al.*, 2015). Another interesting metabolism is the "Taurine and hypotaurine" one. In animal models of PD, protective properties of cystamine have been evidenced. Interestingly, the metabolism of cystamine generates several intermediates including hypotaurine and taurine (Bousquet *et al.*, 2010). Different lipids metabolism have been over-represented. Interestingly, in a lipidomics study of PD substantial changes in sphingolipid and glycerophospholipid biosynthetic pathways have been reported (Cheng *et al.*, 2011).

Table 6. Parkinson's disease: KEGG pathways over-represented only by NET-GE^{M}.

	Standard			Network-based			
ID	**G**	**M**	**p-value**	**G**	**M**	**p-value**	**Description**
hsa01200	48	16	1.09E-30	57	18	3.93E-29	Carbon metabolism
hsa00053	25	5	2.47E-29	25	5	1.05E-28	Ascorbate and aldarate metabolism
hsa00330	35	13	3.19E-27	41	14	6.68E-27	Arginine and proline metabolism
hsa00040	25	8	1.45E-23	26	8	1.04E-23	Pentose and glucuronate interconversions
hsa00260	30	12	1.87E-28	31	12	4.99E-23	Glycine, serine, threonine metabolism
hsa05033	29	3	5.47E-26	32	5	9.00E-23	Nicotine addiction
hsa04727	42	7	2.36E-27	46	9	5.34E-22	GABAergic synapse
hsa00860	25	5	1.48E-20	28	7	6.90E-21	Porphyrin and chlorophyll metabolism
hsa00620	26	7	2.16E-22	27	10	1.78E-19	Pyruvate metabolism
hsa00010	31	7	6.56E-20	34	11	1.84E-17	Glycolysis / Gluconeogenesis
hsa04723	36	6	2.29E-18	44	8	2.82E-17	Retrograde endocannabinoid signaling
hsa01210	16	6	1.19E-18	16	6	6.13E-17	2-Oxocarboxylic acid metabolism
hsa01230	29	10	3.76E-17	33	11	1.43E-16	Biosynthesis of amino acids
hsa00340	18	11	5.29E-18	20	13	1.61E-15	Histidine metabolism
hsa00350	25	13	4.35E-23	26	13	2.47E-15	Tyrosine metabolism
hsa05032	32	4	5.87E-16	39	7	7.05E-15	Morphine addiction
hsa04724	37	9	3.54E-17	42	11	1.67E-14	Glutamatergic synapse
hsa00500	24	3	9.18E-15	26	4	2.77E-14	Starch and sucrose metabolism
hsa00983	21	4	8.59E-14	24	7	5.01E-14	Drug metabolism - other enzymes
hsa00410	17	10	1.42E-12	22	12	1.47E-13	beta-Alanine metabolism
hsa00360	15	11	7.78E-16	15	11	8.45E-12	Phenylalanine metabolism
hsa00380	20	11	1.39E-13	24	12	3.03E-11	Tryptophan metabolism
hsa00270	18	7	1.02E-12	19	7	4.74E-10	Cysteine and methionine metabolism
hsa00430	7	4	9.24E-07	9	6	9.49E-08	Taurine and hypotaurine metabolism
hsa00280	16	10	5.13E-08	19	11	2.76E-07	Valine, leucine and isoleucine degradation
hsa00071	16	5	2.72E-08	17	5	3.93E-07	Fatty acid degradation
hsa04976	16	7	1.30E-05	23	9	2.15E-06	Bile secretion
hsa05031	15	6	3.42E-05	22	9	4.30E-06	Amphetamine addiction
hsa00290	4	4	4.54E-05	5	6	4.63E-06	Valine, leucine and isoleucine biosynthesis
hsa00564	21	5	4.52E-07	25	7	1.13E-05	Glycerophospholipid metabolism
hsa04713	17	4	1.85E-04	23	6	3.56E-05	Circadian entrainment
hsa00460	6	3	1.14E-05	6	3	7.09E-05	Cyanoamino acid metabolism
hsa05230	14	7	7.33E-05	16	8	7.16E-05	Central carbon metabolism in cancer
hsa00640	10	6	2.18E-05	10	6	1.40E-04	Propanoate metabolism
hsa00970	-	-	-	12	8	1.75E-04	Aminoacyl-tRNA biosynthesis
hsa00650	-	-	-	9	8	3.32E-04	Butanoate metabolism
hsa00471	4	3	4.46E-05	4	3	5.20E-04	D-Glutamine and D-glutamate metabolism
hsa00310	13	7	3.11E-05	14	7	8.15E-04	Lysine degradation
hsa00400	4	4	1.99E-04	4	4	9.71E-04	Phenylalanine, tyrosine and tryptophan biosynthesis
hsa04964	7	6	7.58E-04	-	-	-	Proximal tubule bicarbonate reclamation
hsa00561	13	8	6.70E-05	-	-	-	Glycerolipid metabolism
hsa04024	26	8	5.47E-04	-	-	-	cAMP signaling pathway
hsa05231	16	4	6.98E-04	-	-	-	Choline metabolism in cancer
hsa01220	3	3	7.05E-04	-	-	-	Degradation of aromatic compounds

^Functional terms enriched only by NET-GE^{M} and MBROLE2.0. All the other terms are enriched by NET-GE^{M} and NET-GE; $^{\#}$Number of input metabolite-related genes associated to the term; *Number of input metabolites associated to the term via genes; §FDR corrected p-value.

Lastly, NET-GE^{M} highlights the involvement of circadian rhythms. This is plausible since dopamine plays a pivotal role in the regulation of sleep and circadian homeostasis (Videnovic *et al.*, 2013).

Moreover, only via the network-based approach we could highlight the link between PD and the aminoacyl-tRNA metabolism. This metabolism is very interesting since the aminoacyl-tRNA synthetase-interacting factor AIMP2 as shown to be the substrate of Parkin, a protein promoting ubiquitination and proteasome degradation. The control of expression of Parkin substrates through ubiquitination and degradation is critical for dopaminergic cell survival. (Park *et al.*, 2008).

Other pathways, although more generic, are highlighted by NET-GE^{M}: the different metabolisms of amino-acids, the "Pyruvate metabolism" and the "Glycolysis/Gluconeogenesis" (Liu *et al.*, 2016).

Table 7. Parkinson's disease: KEGG pathways over-represented by NET-GE^{M} and confirmed by either NET-GE or MBROLE 2.0. NET-GE^{M} enrichments are presented both for the standard and the network-based method.

	Standard			Network-based			
ID	**G$^{\#}$**	**M***	**p-value§**	**G$^{\#}$**	**M***	**p-value§**	**Description**
hsa01100	232	32	1.49E-77	251	32	4.80E-44	Metabolic pathways
hsa00982	44	11	9.85E-36	46	13	5.71E-30	Drug metabolism - cytochrome P450
^hsa00250	27	10	2.50E-26	31	10	3.25E-25	Alanine, aspartate and glutamate metabolism
^hsa00020	25	6	5.49E-26	25	6	1.79E-24	Citrate cycle (TCA cycle)
hsa00980	35	8	1.22E-22	38	11	1.07E-20	Metabolism of xenobiotics by cytochrome P450
hsa05204	37	13	2.55E-23	39	15	2.51E-20	Chemical carcinogenesis
hsa00830	29	7	3.13E-19	32	10	1.39E-15	Retinol metabolism
hsa00140	22	5	7.58E-13	27	8	8.31E-12	Steroid hormone biosynthesis
^hsa00630	15	11	4.01E-12	15	11	3.83E-11	Glyoxylate and dicarboxylate metabolism
hsa00480	18	8	4.00E-09	19	10	4.24E-09	Glutathione metabolism
hsa04080	47	5	1.77E-10	54	11	1.83E-08	Neuroactive ligand-receptor interaction
hsa05030	15	6	5.08E-07	21	7	1.27E-07	Cocaine addiction

Columns descriptors are given as **in Table 6**.

At the functional level, our comparison of genomics and metabolomics data shows a reasonable level of concordance. NET-GE and NET-GE^M have 17 shared pathways (**Figure 14**). Considering only the leaf terms (no. 9), enriched terms highlights pathways related to PD (**Table 7**). Among them, we have the "retinol metabolism" (Maden 2007), the "glutathione metabolism" (Smeyne et al., 2013), the involvement of xenobiotics and the of P450 system (Steventon *et al.,* 2001; Miksys *et al.,* 2002) and the "neuroactive ligand-receptor interaction".

However, we had to admit that in the fraction of terms enriched by NET-GE^M we had terms detected also by MBROLE, but with a p-value$_{FDR}$ < 0.001. This suggests that the use of genes instead of metabolites, with the addition of an interactome at the basis of the analysis, can counteract problems due to statistics, letting emerge the pathways describing (correctly) the cell complexity at the metabolite level.

9 Conclusions

NET-GE^M provides a new alternative way to analyse metabolomic data. By exploiting the functional modules at the basis of NET-GE to perform a gene-based metabolite enrichment analysis, with NET-GE^M we analyse metabolomics data in the context of the different layers of biological complexity. We qualitatively tested NET-GE^M with a metabolite set linked to Parkinson's disease. The preliminary results obtained over the KEGG pathways database are satisfactory. NET-GE^M enriches more terms than MBROLE2.0, a canonical tool used in the functional analysis of metabolite set. While MBROLE2.0 detects only 3 pathways, NET-GE^M detects other 64 pathways. These pathways were qualitatively evaluated looking in PubMed for experiments supporting these evidences and many of them can be reconducted to the PD. Moreover, to understand the level of complementarity between genomics and metabolomics – in terms of retrievable functional features – we analysed via NET-GE a gene set of 98 PD-related genes. We observed a partial overlap (50% of terms enriched by NET-GE were retrieved also by NET-GE^M; 27% of the terms enriched by NET-GE^M were retrieved also by NET-GE) highlighting that shared biological information is stored in the different layers of biological complexity.

Although preliminary, the results obtained from this "hybrid" method seem promising. However, a lot of work is still necessary to test the performance of NET-GE^M.

Part III

The Critical Assessment of Genome Interpretation (CAGI) experiment

10 Introduction

The Critical Assessment of Genome Interpretation (CAGI, \ˈkā-jē\, https://www.genomeinterpretation.org/) is a community experiment aimed at evaluating computational methods for determining the phenotypic impacts of genomic variants. The CAGI goals are: (i) to evaluate the capability of state-of-the-art methods to make useful predictions of molecular, cellular, or organismal phenotypes from genomic data, (ii) to identify bottlenecks in genome interpretation that suggest critical areas of the future research; (iii) to standardize the field by suggesting appropriate assessment methods and defining what is required for an accurate prediction, (iv) to engage and connect researchers from the diverse disciplines whose expertise is essential to methods for genome interpretation and (v) to highlight innovation.

Participants taking part in CAGI experiments (challenges) are provided genetic variants for which blind predictions of the resulting phenotypes are made. Usually, a CAGI experiment is conducted over a period of one year, that starts with the identification/development of suitable challenges (release of unpublish data and formulation of related questions) followed by a period during which participants are invited to analyse data and submit predictions. After the closure of challenges, independent assessors evaluate predictions against gold-standard experimental or clinical data. CAGI experiments ends with a meeting to discuss the outcomes. Started in 2010, four CAGI experiments have been conducted to date. Moreover, participants (data providers, predictors and evaluators) are encouraged to publish their finding. This year, a special issue of Human Mutation has bene completely dedicated to the CAGI experiments (see Hoskins *et al.*, 2017).

During the four editions of CAGI, the same challenge (with new blind datasets) and new challenges have been proposed. These challenges spanned a wide range of relationships between genetic variants and phenotypes such us challenges about the effect of single-base variants on RNA expression levels and protein activity or the interpretation of exome and genome sequencing data for assigning complex traits phenotypes. Other challenges regarded the ability to predict the effect of mutations in cancer driver genes on cell growth and challenges in which participants were asked to identify causative variants for rare diseases in a given gene panel.

In the following chapter, I will introduce the methods proposed in facing three CAGI challenges: the prediction of the effect of variants on the Liver Pyruvate kinase activity and

allosteric regulation, and the interpretation of Crohn's and Bipolar exomes for discriminating healthy and sick individuals. **The full version of the journal articles describing the methods, Daneshjou *et al.*, 2017; and Xu *et al.*, 2017 are reported in the appendix.**

10.1 The Pyruvate Kinase challenge

Pyruvate kinase (PYK) is an enzyme (EC 2.7.1.40) which regulates the last step of the glycolytic pathway (Gupta *et al.*, 2010). PYK catalyses the transfer of phosphate from phosphoenolpyruvate (PEP) to ADP, to generate ATP as following:

$$\text{Phosphoenolpyruvate (PEP)} + \text{ADP} \rightleftarrows \text{Pyruvate} + \text{ATP}$$

Depending on the different tissue requirements, mammals express four isozymic forms: 1) R-PYK, restricted to erythrocytes, 2) L-PYK, found predominantly in liver and kidney, 3) M1-PYK, expressed in muscle and brain, and 4) M2-PYK, found in fetal tissues and in proliferating cells (Morgan *et al.*, 2013).

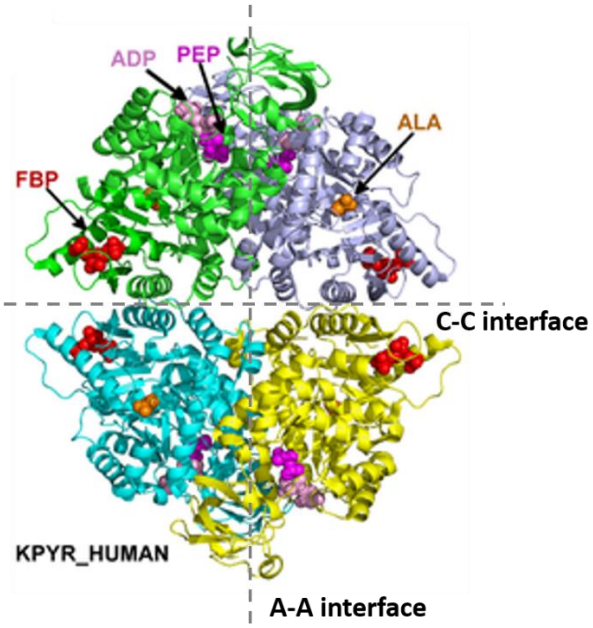


Figure 15. Structure of tetrameric human pyruvate kinase. The structure was assembled by superimposing monomers from several structures of homologues of L-PYK with PEP, ADP, and alanine bound onto a tetrameric structure of human L-PYK with fructose-1,6-bisphosphate (FBP) bound (PDB: 4IP7). PEP, ADP, ALA, and FBP are shown in spheres, coloured in magenta, pink, orange, and red, respectively. Figure adapted from Xu *et al.*, 2017.

As reviewed by Dombrauckas *et al.,* 2005, PYK is a tetrameric protein of identical subunits which are arranged in a dimer-of-dimers configuration (**Figure 15**) with approximate D_2 symmetry. Each subunit contains one active site and is composed of four domains: the A-, B-, C- and N-terminal domains. The A-domain, is the core of the monomer and has an α_8/β_8 barrel tertiary structure motif. The active site is localized at one end of the barrel, in a cleft between the A-and B-domains. The B domain is mobile domain that moves toward the A-domain closing on the active site upon binding of the Mg^{2+}-ADP substrate complex. The C-domain, found on the opposite side of the A domain, consists of both α and β structural elements. The allosteric FBP-binding pocket is located entirely within the C-domain. The tetramer is held together by reciprocal hydrogen bonds across the C-C (small) interface, between the neighbouring C-domains, and the A-A (large) interface, between the neighbouring A-domains.

Regulation of the M2, L, and R isoforms is accomplished by (i) the phosphorylation of the N-terminus at S12 and (ii) allosteric regulation by fructose-1,6-bisphosphate (Fru-1,6-BP), alanine, and ATP (Prasannan *et al.,* 2012). Fructose 1,6 bisphosphate acts as allosteric activator while alanine as allosteric inhibitor (Fenton *et al.,* 2009a).

Regulation of L-PYK plays a pivotal role in the maintenance of glucose homeostasis, preventing hyperglycaemia and hypoglycaemia. Several non-synonymous variants of R/L-PYK in the PYK deficiency patients have been observed falling in or near the allosteric effector binding sites, and modifications in allostery seem sufficient to cause disease (Xu *et al.,* 2017).

Given the complex regulation of L-PYK, the understanding of the allosteric regulation is of fundamental importance in order to develop allosteric drugs acting on the L-PYK. In the case of L-PYK, allosteric regulation can be defined as how one ligand *A* binds to a protein *E* in the presence vs. absence of a second ligand *X* (Prasannan *et al.,* 2012). This definition describes allostery by a thermodynamic energy cycle composed by four enzymatic states (**Figure 16**). Moreover, it also defines an allosteric coupling constant (Q_{ax}) as following:

$$Q_{ax} = \frac{K_{ia}}{K_{ia/x}} \qquad (5)$$

where K_{ia} and $K_{ia/x}$ are the dissociation constants for binding the substrate *A* in the absence or presence, respectively, of the allosteric effector *X*, as defined in **Figure 16**. A value of $Q_{ax} = 1$ indicates that the system is not allosteric, while a $Q_{ax} \neq 1$ indicates allosteric coupling between

the binding of *A* to a protein and the binding of *X* to the same protein. In particular a Q_{ax} >1 indicates a positive allosteric coupling, while Q_{ax}<1 denotes a negative or inhibitory coupling between *X* and *A* sites.

Given two sets of mutations in the L-PYK, retrieved from two different experiments, participants were asked to submit predictions on the effect of the mutations on L-PYK enzyme activity and allosteric regulation. Experimental results on enzyme activity are categorize into two classes: we were asked to enter the probability that the mutant enzyme retains activity (0 = no activity detected, 1 = activity detected). Allosteric coupling is measured as a continuous assay result; we were asked to enter the predicted numeric value for Q_{ax}. For experiment #1, we had to predict the results of the assay to measure coupling of the allosteric inhibitor alanine. For experiment #2, we had to predict the results of the assay to measure coupling of the allosteric inhibitor alanine plus the allosteric activator F-1,6-BP.

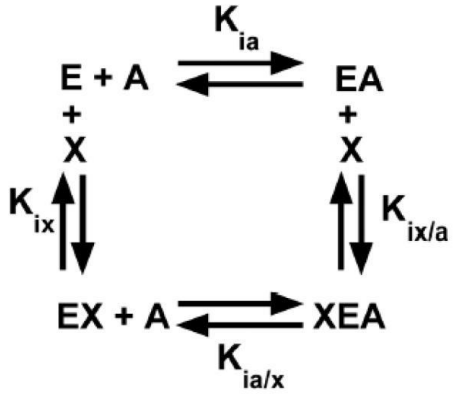


Figure 16. Allosteric energy cycle. In the reaction scheme, the enzyme *E* can bind one substrate *A* and one allosteric effector *X*. K_{ia} is the binding of the substrate *A* to the enzyme *E* in the absence of effector *X*. $K_{ia/x}$ is the binding of the substrate *A* to the enzyme *E* in the presence of saturating concentrations of effector *X*. K_{ix} is the binding of effector *X* to the enzyme *E* when substrate *A* is absent. $K_{ix/a}$ is the binding of effector *X* to the enzyme *E* in the presence of saturating concentrations of substrate *A*. Figure from Prasannan *et al.*, 2012.

10.2Datasets

Predictors were provided with two datasets representing the results of two site directed mutation studies of the human L-PYK protein (expressed in E. coli). Briefly, for each mutant the Q_{ax} value was determined by measuring the affinity of the enzyme for PEP (via titration of

activity over a concentration of PEP) over a concentration range of effector. Details of the assays and data evaluations are reported in several papers (Fenton *et al.,* 2009a; Fenton *et al.,* 2009b, Ishwar *et al.,* 2015).

The first experiment generated a dataset of 113 mutations at 9 residues that contact bound alanine or are very near the bound alanine: Arg55, Ser56, Asn82, Arg118, His476, Val481, Pro483, and Phe514. Only one mutation per mutant protein was introduced by using degenerate codons/random substitutions. The resulting proteins were evaluated for presence/absence of enzyme activity. A total of 23 mutant proteins were completely inactive, while the remaining 90 conserved activity and have therefore been used to evaluate the coupling constant $Q_{ax\text{-}Ala}$.

In the second experiment, the alanine scanning mutagenesis approach was used to evaluate which non-alanine/non-glycine residues in the L-PYK contribute to allosteric functions. A total of 430 residues were mutated into alanine. Coupling constants $Q_{ax\text{-}Ala}$ and $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$ were separately evaluated. Allosteric coupling of alanine and F-1,6-BP could not be measured for 37 and 19 mutant proteins, respectively.

For sake of clarity, when the challenge opened, only information about the residue position and the substitution type were available and participants were asked to submit prediction for all the mutants. Details about the number of mutants where it has been possible to measure enzymatic activity and allosteric coupling were released only during the assessment phase.

10.3Method

To predict the effect of the different residue substitutions on the enzyme activity (task 1), we initially remapped the alanine binding site on the crystal structure of human L-PYK (PDBs: 4IP7, 4IMA, Holoyak *et al.,* 2013) considering the human M2 pyruvate kinase (PDB:4FXJ, Morgan *et al.,* 2013) locked into the T state by phenylalanine. Then, we measured the distance of each mutated residues from: 1) the Citrate/Mn/ATP/Fru-1,6-BP molecules in the respective binding sites and 2) the residues at the C-C (small) and A-A (large) interfaces of the tetramer. To determine the residues at the interfaces we made use of the DSSP program (Kabsch and Sander, 1983) to compute the solvent-accessible surface area (SASA) of each residue. By denoting with $SASA_T$ and $SASA_M$ the solvent-accessible surface

area of the residues in the T and M forms, respectively, we considered a residue at the interface when the parameter Δ_{SASA} = $SASA_T$ - $SASA_M$ was ≥ 10 Å^2.

We predicted a mutation as neutral for the protein activity when it is located in region far apart (>0.5 nm) from the active site, the effector sites and the interface domains. As results, we provided a binary classification of enzyme activity: 1 for active mutant and 0 for inactive mutant.

In the prediction of the allosteric coupling constant $Q_{ax\text{-}Ala}$ (task 2), we considered: 1) the distance (DR) between the effector binding site and the mutated residue, and 2) the substitution weight (w) of the mutation as derived from the BLOSUM62 scoring matrix. In our prediction we did not provide a continuous value as requested. Instead, we adopted a binary classification, assigning a value of 0.1 to a wild type coupling (DR > 0.5nm or w ≥ 0), and a value of 1 to mutations abolishing the allosteric coupling (DR ≤ 0.5 nm and w < 0). Our hypothesis was that the higher the conservation, the lower the effect.

The prediction of the allosteric coupling constant $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$ was analogous to the $Q_{ax\text{-}Ala}$, except for considering the distance from the F-1,6-BP, instead of the alanine.

10.4Performance assessment

The performance of the method was evaluated by an external evaluator. To score the prediction performance of the first task, the evaluator made use of the following scoring indexes:

$$TPR = \frac{TP}{TP+FN} \quad (6)$$

$$TNR = \frac{TN}{TN+FP} \quad (7)$$

$$PPV = \frac{TP}{TP+FP} \quad (8)$$

$$NPV = \frac{TN}{TN+FN} \quad (9)$$

To assess the overall accuracy, the evaluator calculated also the total accuracy (ACC), the balanced accuracy (BACC), the Matthews correlation coefficient (MCC) and the F1 score.

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \quad (10)$$

$$BACC = \frac{1}{2}(TPR + TNR) \quad (11)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (12)$$

$$F1 = 2\frac{TPR \times PPV}{TPR+PPV} \quad (13)$$

In the evaluation of the $Q_{ax\text{-}Ala}$ and $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$, the Spearman's rho (ρ) and Kendall's tau (τ) correlation indexes were used. The Spearman's rho (ρ) index, also defined as Spearman's rank correlation coefficient, measures the monotonic correlation between prediction and experimental data. Give the dataset (*p*, *e*), where p_i and e_i are i^{th} predicted and experimental data points, respectively, prediction data points are converted into ranks Rp_i and experimental data points are converted into ranks Re_i. Then, ρ is calculated as following:

$$\rho = \frac{cov(Rp,Re)}{\sigma_{Rp}\sigma_{Re}}, -1 \leq \rho \leq 1 \quad (14)$$

where *cov(Rp,Re)* is the covariance and σ_{Rp} , σ_{Re} are the standard deviations of the ranked variables.

The Kendall's tau (τ) index, also define Kendall rank correlation coefficient, is another measure the ranks correlation between two variables. Given a dataset (*p*, *e*), any pair of (p_i, e_i) and (p_j, ej), where $i \neq j$, are said to be concordant if both $p_i > p_j$ and $e_i > e_j$, or if both $p_i < p_j$ and $e_i < e_j$. They are discordant, if both $p_i < p_j$ and $e_i < e_j$, or if $p_i > p_j$ and $e_i > e_j$. If $p_i = p_j$ and $e_i = e_j$, the pair is neither concordant nor discordant. Then, τ is calculated as following:

$$\tau = \frac{(n.of\ concordant\ pairs) - (n.of\ discordant\ pairs)}{n(n-1)/2}, -1 \leq \tau \leq 1 \quad 15)$$

Considering the two indexes, a value equal to 1 indicates that the predicted and the experimental data points have identical rankings, a value equal to -1 indicates that one ranking is the reverse of the other and a value equal to 0 indicates that the sets of data are independent.

10.5 Results

The L-PYK challenge was accessed by more than 30 researchers. However, only four groups accepted the challenge. Researchers approached the challenge considering different features, such as evolutionary information, the location of mutations and molecular docking. Details about the groups and the different approaches are reported in Xu *et al.*, 2017.

In the following subchapters I will discuss the methods and results obtained by our approach.

10.5.1 Prediction of the L-PYK enzyme activity

The first task of the L-PYK challenge regarded the prediction of the effect of single mutations on the enzyme activity, in terms of retained activity or not (and not the level of activity). Given a set of 113 mutations, predictors were invited to submit a prediction in a binary form: 0 means inactive, 1 active.

We hypothesized that a mutation is neutral to the protein activity provided that it is located in region far apart (>0.5 nm) from the active site, the effector sites and the interface domains. Based on this hypothesis, in the first dataset we predicted all the mutations as neutral to the protein activity since they were located outside the main functional sites. The performance indexes of our method are reported in **Table 8**, while performance indexes of all the different participants groups are presented in **Figure 17**.

Overall the method performed well with an ACC = 0.796. The hypothesis we made was quite reasonable since all the mutations were in or near the alanine binding site which is distant from the active site. However, a group performed better than us, reaching an ACC = 0.867. Given the imbalance of the dataset, a measure like the F1 score was suited for the performance evaluation since it only includes positive predictions and experimental phenotypes, and omits the negative ones. Since this dataset consisted of majority of active enzymes (80%), having predicted a larger fraction of the enzymes to be active, we achieved good results (F1 = 0.887). We tried to figure out what did we miss. Among the 23 misprediction, we missed the correct prediction of the V481 (6/13 mutations) and P483 (10/11 mutation). Based on our definition of interface, they were classified as residues not at the interface. However, they are part of the Cβ1 that is integrated into the β-sheet whose Cβ5 is part of the C-C interface.

Table 8. L-PYK enzyme activity task: binary prediction results.

	TPR	TNR	PPV	NPV	ACC	BACC	MCC	F1
EXP1*	1	0	0.796	0	0.796	0.5	0	0.887
EXP2#	0.838	0.205	0.901	0.127	0.772	0.521	0.034	0.868

*EXP1: 113 mutations at 9 residues that contacts bound alanine; #EXP2: alanine scanning.

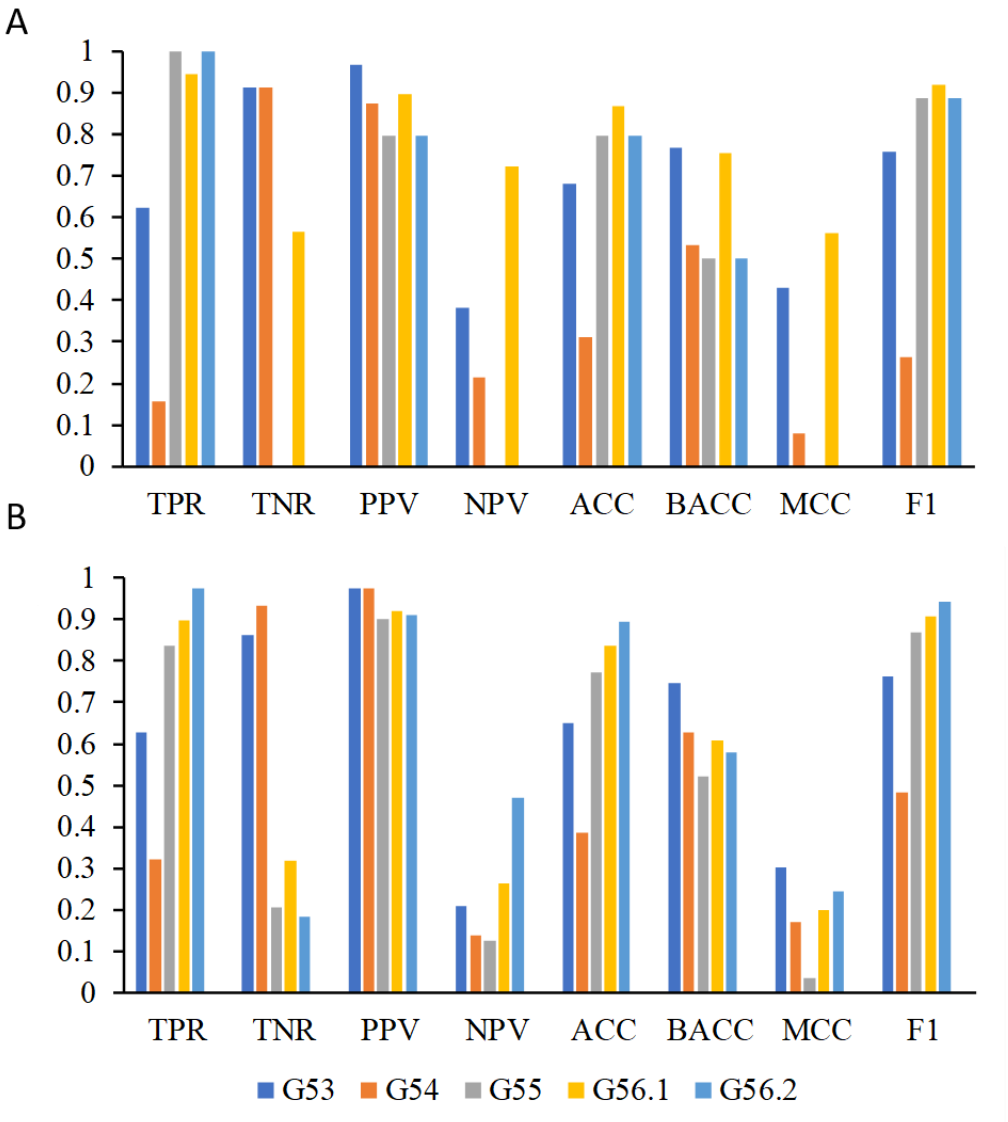


Figure 17. L-PYK enzyme activity task: binary prediction results of the four participant groups. Our group the G55 one (gravy colour). Data are taken from (Xu *et al.*, 2017) A) Datasets of 113 mutations at 9 residues that contacts bound alanine or is very near the bound alanine. B) Alanine scanning datasets.

The same hypothesis was tested in the analysis of the alanine scanning derived dataset (mutations across the entire enzyme). Also in this case, we reached an ACC = 0.772, and a F1 equal to 0.868.

However, in the assessment of binary phenotype predictions with imbalanced dataset a measure like the BACC index is recommended (see eq. 11) (Wei *et al.*, 2013). BACC is the average of the rate of correctly predicting the experimentally active mutants (TPR) and the rate of correctly predicting the experimentally inactive mutants (TNR). A value of BACC = 0.50 correspond to a random prediction, since if one predicts all of the phenotypes in one class – like in our case – the BACC is automatically equal to 0.50. By using this index, given the score around 0.5 for both the dataset, it emerges that the hypothesis we made probably is too naïve for the task.

10.5.2 Prediction of $Q_{ax\text{-}Ala}$

The second task regarded the prediction of the effect of a mutation on the allosteric regulation. Predictors were asked to estimate the inhibitory allosteric effect of alanine on binding of the substrate PEP, by submitting a value in the range of $0 < Q_{ax\text{-}Ala} \leq 1$.

We hypothesized that the $Q_{ax\text{-}Ala}$ predicted values could depend on the distance of the mutated residue from: 1) the active site, 2) the allosteric inhibitor alanine or 3) the residues at the interfaces. Moreover, we scored the mutations based on the weight derived from BLOSUM 62, assuming that the higher the conservation, the lower the effect. We adopted a binary classification, where $Q_{ax\text{-}Ala}$ = 0.1 indicates wild type coupling while $Q_{ax\text{-}Ala}$ = 1 indicates abolition of the allosteric coupling.

In the first experiment, 23 mutants out of 90 did not have a measurable allosteric coupling ($Q_{ax\text{-}Ala}$ = 1) index, since mutants were inactive, while the remaining ones had a $Q_{ax\text{-}Ala}$ ranging from 0.014 to 0.590. Our group was the best performing one, achieving favourable correlations with ρ = 0.351 and τ = 0.299 (p-values = 0.002 for both), while the other predictors had p-values in the range of 0.17 – 0.88 (no correlation). The challenge was assessed using also the TPR, TNR, etc. scores while considering the experimental $Q_{ax\text{-}Ala}$ values as binary data. We obtained: TPR = 17/23 = 0.739, TNR = 39/55 = 0.709, BACC = 0.724. This was better than random and explained the positive correlation coefficients.

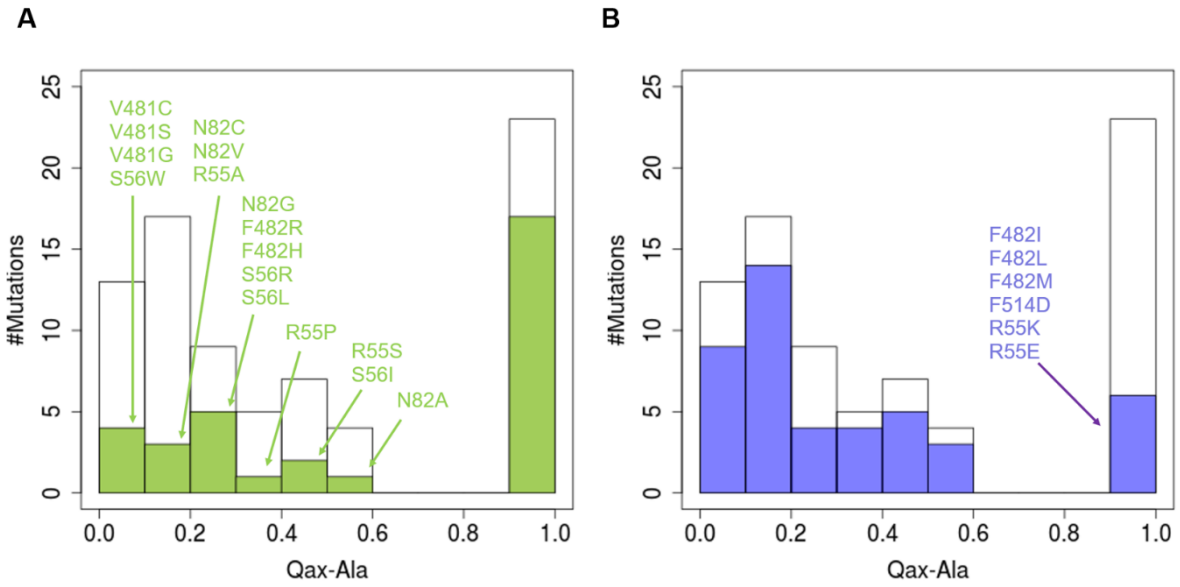


Figure 18. Prediction of the effect of the mutations near alanine on the L-PYK allosteric regulation (Experiment #1) The two bar plots report the experimental measures of $Q_{ax\text{-}Ala}$. Mutants abolishing allosteric coupling have $Q_{ax\text{-}Ala}$= 1, while mutants preserving the allosteric coupling to some extent have $Q_{ax\text{-}Ala}$ ranging between 0.014 to 0.590 ($Q_{ax\text{-}Ala}$ of the wild type protein is 0.08). The two panels highlight on the experimental distribution the fraction of residues predicted as A) abolishing (predicted $Q_{ax\text{-}Ala}$ = 1) and B) preserving the allosteric coupling to some extent (predicted $Q_{ax\text{-}Ala}$ = 0.1). Misclassified residues are reported.

Figure 18 reports the distribution of the $Q_{ax\text{-}Ala}$ values, with highlighted the fraction of residues misclassified. Considering our predictions, we had a perfect prediction for three mutated residues: R118 (15 mutations), H476 (5 mutations), P483 (1 mutation).

For the other 6 mutated residues we had a percentage of wrong prediction in the range 8-50%: R55 (10 mutations, 50% misclassified), S56 (8 mutations, 38% misclassified), N82 (8 mutations, 50% misclassified), V481 (7 mutations, 43% misclassified), F482 (13 mutations, 38% misclassified), F514 (12 mutations, 8% misclassified). Considering these results, the method used to address this task seems to work quite well. Given these results, it emerges that the use of the residues position in the predictions seems quite reasonable. However, what we have to better define seems the way to score the residue substitution. Here, we used the BLOSUM62 matrix. However, we should try other BLOSUM (or PAM) matrices, or completely different matrices such as the McLachlan chemical similarity matrix (McLachlan, 1972).

In the second experiment (the alanine scanning) we tested the same hypothesis. With the exclusion of 37 mutants for which $Q_{ax\text{-}Ala}$ could not be measured, the remaining ones had a

$Q_{ax\text{-}Ala}$ ranging from 0.01 to 0.49. We achieved a very weak positive correlations (all the other groups had negative correlation), with p-values ranging from 0.38 to 0.88. However, we have to admit that we erroneously classified some mutations as abolishing the allosteric coupling, and with the exclusion of our accidental mistake, the hypothesis partially works.

10.5.3 Prediction of $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$

In the second experiment predictors were asked to predict the allosteric effect of Fru-1,6-BP binding to L-PYK. $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$ values ranged from 0.5 to 320, with the vast majority of mutants having a $Q_{ax\text{-}F\text{-}1,6\text{-}BP}$ values between 0 and 60. In this task, only our group achieved a positive correlation, even if it was very marginal (both ρ and $\tau \sim 0.05$, with p-value = 0.2). All other predictors had negative correlations.

10.6 Conclusion

In the context of cellular complexity, biological regulation by means of allostery plays a role in the definition and control of cellular pathways. Allostery take place when the binding of an effector molecule induces an effect on the main functional site of a protein. Whole protein site-directed mutagenesis experiments are often used to probe allosteric mechanisms (Carlson et al., 2016). Moreover, the development of computational methods for the prediction of allosteric sites and the effect of mutations on the allosteric mechanisms is of great interest, especially in allosteric drug design.

The L-PYK experiment aimed at evaluating different strategies used to predict the effect of mutations on allosteric regulation while understanding their bottlenecks. For the prediction of allosteric effects of alanine and fructose, we had positive correlations for the $Q_{ax\text{-}Ala}$ challenge in first experiment. Moreover, only our predictions were statistically significant. In the alanine scanning experiment, no one group had a statistically significant positive correlation for their predictions of the $Q_{ax\text{-}Ala}$ and $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$ values. Our predictions of the allosteric effect of alanine and fructose considered the distance of the mutated residues from functionally important sites, while the severity of the mutation from wild type considered the scores of a substitution matrix. It is likely that we predicted many of the mutations that abrogated Ala binding altogether ($Q_{ax\text{-}Ala} = 1$), rather than quantitatively predicting the effect of the mutations on the diverse values of $Q_{ax\text{-}Ala}$ of the remaining mutations ($Q_{ax\text{-}Ala} < 1$).

Given the performance of the same procedure over the alanine scanning dataset (mutations to alanine at 430 sites throughout the protein), it is not likely that our distance-based method would extend readily to the general problem of predicting allosteric effects, especially for residues not in or near the binding site (Xu *et al.,* 2017).

However, considering the results of our methods and the ones of the other predictors, it emerges that additional approaches are needed to decipher the biological complexity of allosteric regulation.

11 CAGI4 exome challenges

With the advent of Next Generation Sequencing technologies, the way to understand the human genome in health and disease drastically changed, moving toward the concept of genomic/personalized medicine: the use of an individual patient's genotypic information in his or her clinical care (Manolio *et al.,* 2013). Whole genome or whole exome sequencing approaches have now become routine in the identification of causative variants (Iglesias *et al.,* 2014). However, the interpretation of genetic data is still one of the major difficulties in the implementation of precision medicine (Fernald et al., 2011).

To evaluate the methods aimed at interpreting genomic variants, three challenges involved making predictions using exome sequence data at CAGI4: the Crohn's disease challenge, the bipolar disorder challenge and the warfarin dosing challenge (Daneshjou *et al.*, 2017).

In the next paragraphs I introduce the methods and the results we obtained when addressed the Crohn's disease and bipolar disorder challenges.

11.1 Crohn’s disease

Crohn’s disease (CD; OMIM #266600) is a chronic idiopathic inflammatory bowel disease (IBD) condition characterized by skip lesions and transmural inflammation that can affect the entire gastrointestinal tract, from the mouth to the anus (Feuerstain *et al.,* 2017). Several twin-based, family-based, candidate gene and genome wide association studies have provided evidences about the involvement of genetic factors for the expression CD (Liu *et al.,* 2014). Moreover, accumulating evidences suggest that the immune tolerance to the normal intestinal bacteria is disturbed in the genetically susceptible individuals (Cho *et al.,* 2008).

Even if the exact causes of CD are unknown, what emerged till now is a complex interplay between environmental factors, including the intestinal microbiota, and the host immune mechanisms in genetically susceptible individuals (Franke *et al.*, 2010).

11.2 Bipolar disorder

Bipolar disorder (BD) is a severe brain disorder that causes unusual shifts in mood characterized by episodes of mood elevation (mania) and depression, interspersed with periods of euthymia. Moreover, the disease is associated with a significant mortality, with high rates of suicides and medical comorbidities (Sagar *et al.*, 2007). Despite a complex etiology, family-based and twin based genetic studies have provided evidences about the role of genetic factors for the expression of BD (Smoller *et al.*, 2003). Moreover, genomic-based and genome wide association studies have identified several genes robustly associated with BD (Goes *et al.*, 2016).

11.3 Datasets

11.3.1 Crohn's disease

Predictors were provided a dataset of 111 unrelated German ancestry exomes (64 cases, 47 controls) sequenced by using an Illumina HiSeq2000 instrument. Data providers processed the data as following: i) sequenced reads were mapped to the human genome build hg19 and 2) variants were called for all the 111 exomes together using the Genome Analysis Toolkit (GATK v. 3.3-0, McKenna *et al.*, 2010) Haplotype Caller. Variant calls were restricted to the TruSeq exome target. GATK was also used in the variant quality score recalibration steps. Only high-quality variants passing the filters were retained, for a total of 247,537 variants.

In addressing this challenge, predictors were invited to use the CD datasets released in previous CAGI editions for calibrating their methods (CAGI2 and CAGI3). In training our method, we made use of the CAGI3 dataset. It comprised 66 German ancestry exomes: 51 cases and 15 controls. The TruSeq exome bed file was used for combined variant calling for all the 66 exomes using the GATK program. GATK was also used in the variant quality score recalibration steps. A total 202,691 of genomic variants were called. In this dataset part of the samples were related. In fact, some cases were selected from pedigrees of families with multiple occurrences of CD. Controls samples were mostly unrelated healthy individuals,

except for the unaffected parents of three cases and the unaffected twin of one case (Daneshjou *et al.*, 2017).

11.3.2 Bipolar disorder

Predictors were provided a processed dataset of 1,000 exomes of unrelated bipolar disorder cases and age/ancestry-matched controls of Northern European ancestry. Samples were sequenced using an Illumina HiSeq2000 machine. The NimbleGen SeqCap EZ v2.0 Exome arrays with ~3.4 Mb additional custom target for promoter, UTR, and intronic information of 1,422 synaptic genes and 60 genes previously associated with BD were used for target capture (Daneshjou *et al.*, 2017). Data provider processed the data as following: 1) sequenced reads were mapped to the human genome build hg19 and 2) variants were called for all the 1,000 samples together using the GATK Unified Genotyper. Only high-quality pass variants were retained. The data providers discarded variants with more than 10% of un-called genotypes or in Hardy-Weinberg disequilibrium at p-value < 1×10^{-6}, as well as specific genotype calls with read depth < 10 or genotype quality < 20. A total of 501,253 genomic variants were provided. The organizers divided the dataset into halves: 500 exomes for training, and 500 exomes for the prediction challenge.

11.4 Methods

The method we implemented for both challenges is based on the assumption that individuals affected by a disease carry harmful variants in specific candidate genes. We performed an ab-initio search for candidate genes associated to Crohn (CD) and Bipolar disorders (BD), separately, by using the two provided training dataset. The candidate gene extraction procedure, consisted of three major steps: 1) data processing, 2) variant annotation and 3) gene selection. Details are given in the next subsections.

11.4.1 Data quality assessment

In each dataset, genomic variants were retained provided that the corresponding genomic position was covered at least in the 90% of the samples and with a genotype quality ≥ 20. In the case of CD, we retained only genomic variations fulfilling the same requirements in the CAGI4 dataset comprising the genomes to be classified.

11.4.2 Variant/gene annotation

The effect of genomic variants on gene products were annotated with the Variant Effect Predictor (VEP; McLaren *et al.*, 2016). We retained only variants promoting frameshifts, stop-gains or losses and residue changes (missense variants).

For each sample, we marked a gene as disease-related if it carried at least one variant in homozygous form (recessive variants) for the alternative genotype. We built a matrix in which, for each individual, genes were labelled either as 0 (indicating a wild type gene) or 1 (presence of a potentially harmful mutation).

In order to better score the potential harmfulness of missense variants we adopted SNPs&GO (Calabrese *et al.*, 2009). Briefly, SNPs&GO is a support vector machine based method that uses different pieces of information, derived from protein sequence, protein sequence profile, and protein function (GO annotations) to predict if a given mutation can be classified as disease-related or not. For each prediction, the probability of being disease-related is used to compute the reliability score. In our analyses we retained all the missense variations predicted as disease-related, independently of the reliability score.

We built a second binary matrix not considering missense variations predicted as neutral.

11.4.3 Gene selection

Candidate genes were identified on the basis of the percentage of healthy and sick persons carrying potentially harmful variants, considering the two different matrices defined in the previous section. In both cases, two thresholds have been set: the minimum percentage of sick (%S) and the maximum percentage of healthy (%H) variant-carriers.

In order to fix the thresholds, we carried out a k-fold (k=3 for CD and k=5 for BD) cross-validation (k-CV) procedure. Briefly, the training dataset was divided in k equal size subsets: a single subset was used as testing set, while the remaining k-1 subsets were used as training data. This process was repeated k times: each one of the k subset was used exactly once as testing set.

In each cross-validation round, a grid search on the parameters %S and %H was performed. A total of 80 points were sampled with %S ranging from 1 to 80 and %H ranging between 0 and 80. Each pair of values allowed to select a set of genes fulfilling the requirements in the

training set. Genomes in the training set were then sorted according to the number of selected genes carrying mutations. The discrimination power of each selected set of genes was assessed with a Receiver Operating Characteristic (ROC) Curve Analysis and the best pair of thresholds (%S and %H) was accordingly retained. Testing sets were adopted only to evaluate the discriminative performance (as reported in **Table 9 and Table 10**, row 4)

Briefly, the ROC curve analysis shows the trade-off between the true positive rate (TPR) and false positive rate (FPR) as one changes the criterion for positivity, in this case the number of selected genes carrying mutations in the genome (Hajian-Tilaki *et al.*, 2013). The Area Under the Curve (AUC) was used as evaluation. A perfect predictor gives an AUC score equal to 1 while a random predictor is characterized by an AUC of 0.5.

Each cross-validation fold retrieves different threshold pairs and, consequently, different gene sets. When building the final classificatory, two different merging criteria were adopted: 1) using as candidate genes the union of the *k* sets (herein called "union set") or 2) using as candidate genes only the genes shared among the *k* sets (herein called "intersection set").

Predictions on the challenge set, for both the disorders, were computed by counting the number of mutated interesting genes. A total of four different submissions were done since two training datasets (SNPs&GO based or not) and two ways to select genes (union or intersection sets) were adopted. For each individual and for each prediction method, a probability value for the disease was predicted by evaluating the ratio of mutated genes among the candidate set.

11.4.4 Assessors evaluation

Assessors evaluated the predictions by using the ROC analysis, by testing the robustness of the prediction accuracy when making predictions on different subsamples of exomes and by assessing the confidence intervals reported by the participants. To capture confidence intervals on the predictions, Monte Carlo sampling was adopted: each prediction was then modified by adding a random value drawn from a normal distribution with a mean of zero and a standard deviation equivalent to that reported by the predictors. If no confidence interval was reported for the original prediction, the standard deviation was taken to be zero. If a prediction for a particular exome was missing, the prediction score for that sample was set to the mean reported prediction value in that submission. In order to compare submissions by a single figure of merit, the average area under the ROC curves from the bootstrap sampling

was used, accompanied by the bootstrapped confidence interval around that area under the curve, to estimate the robustness of differences between prediction performances (Daneshjou *et al.*, 2017).

11.5 Results

14 and 9 groups submitted predictions for the CD and BD exomes challenges, respectively. In addressing these challenges, participants made use of many different methods, and in general the same method was applied in solving both the challenges. Groups used specific disease related variant/gene sets (as retrieved from genome-wide association studies or from genomic catalogues) or predicted variant/gene sets. A range of machine learning approaches were used to build the classifiers: naïve Bayes, logistic regression, neural nets, and random forests. The detailed explanation of the different approaches is reported in Daneshjou *et al.*, 2017.

The method we developed was aimed at identifying a disease related gene set able to discriminate between sick and healthy individuals. One of the two procedures we devised relied on SNPs&GO, a pathogenicity prediction tool based on different information derived from protein sequence, protein sequence profile, and protein function. In the following sub-chapters, I will discuss the results obtained by our methods.

11.5.1 Crohn's disease exome challenge.

Given a set of 111 exomes (247,537 genomic variants), predictors were asked to identify which individuals had Crohn's disease and which ones were healthy.

To address this challenge, we developed a method that used the CAGI3 datasets as training dataset. In this dataset, case samples (n. 51) were collected from German families with a particularly high burden of Crohn's disease (two or more affected family members), including a pair of twins discordant for the disease, and another pair of concordant twins. Additional healthy controls were drawn from the unaffected German general population (Daneshjou *et al.*, 2017). During the evaluation of the CAGI3 Crohn's disease challenge, assessors evidenced a substantial difference in clustering between cases and controls, thus we decided to operate with an unbiased dataset by excluding the 8 control samples clearly clustered together apart. In fact, when implementing our prediction procedure with the whole

set of 66 exomes (data not shown) we were always able to clearly discriminate that 8 controls, while the remaining 7 controls were more challenging (data not shown). The samples we used in training were 58: 51 cases and 7 controls. Moreover, before implementing the discrimination procedure we restricted our analyses to only those genomic variants shared between the CAGI3 and CAGI4 datasets. The two datasets had in common a total of 22,985 genomic variants

The VEP based annotation procedure ended up with a dataset of 58 samples and 22,587 harmful variants (0 frameshift, 219 stop-gain, 41 stop-loss and 22,327 missense variants). SNPs&GO annotations were available for 16,733 variants, of which a small fraction of 824 variants were predicted as disease related.

In each sample, we labelled a gene as disease related if it carried at least one harmful variants in the homozygous form. By indicating with 0 a non-mutated gene, and with 1 a gene carrying the homozygous alternative form of harmful variations, we ended up with a 58 samples × 9,029 genes binary matrix. In one of the two methods we developed, missense variants were considered harmful only if predicted as disease related by SNPs&GO A second binary matrix was built by considering the information of SNPs&GO. In this case we had a binary matrix of size 58 × 7,077 genes.

The two matrices were used separately in the gene selection procedure. In this step, we adopted a 3-CV procedure. The three subsets were composed each one by 17 case and 2-3 control individuals, randomly selected.

Table 9. Crohn's disease. Parameters and performance of the methods**.**

	With SNPs&GO		**Without SNPs&GO**	
	Union	**Intersection**	**Union**	**Intersection**
% of the healthy samples	<25%	<25%	<25%	<25%
% of the sick samples	>25%	>25%	>25%	>25%
n. of selected genes	122	21	667	201
AUC testing	0.95	0.98	0.99	0.99
AUC 111 exomes	0.41	0.44	0.45	0.44
AUC* 111 exomes (LCL, UCL)	0.47 (0.46, 0.48)	0.46 (0.45, 0.47)	0.50 (0.49, 0.51)	0.46 (0.45, 0.47)

*The average area under the ROC curves from the bootstrap sampling is presented (LCL: lower confidence limit, UCL: upper confidence limit). Confidence level was set at 0.95.

The selection procedure aimed at identifying a gene set maximizing the AUC score while considering the percentage of the sick and healthy sub-populations carrying the mutated gene under evaluation. By adopting a 3-CV procedure, 3 gene sets were returned. We tested the ability to discriminate of both the union and the intersection of the gene sets (the combination and the intersection of the 3 gene lists, respectively). For each implementation, **Table 9** lists the parameters and the performance obtained during the training phase.

The four interesting gene sets were then used to predict the health status of the 111 samples. Evaluations are presented in **Table 9**. Despite the good AUC score obtained in training (AUC 0.95 – 0.99), on the CAGI4 dataset all the four models did not performed as expected (AUC 0.41 – 0.44). Possible causes of a near random prediction could be attributed to the low number of controls used during the training phase (and the imbalance of the training dataset itself).

Among the 14 participating groups, the top performing group reached an average area under the ROC curve of 0.65 by using a Naïve Bayes model incorporating information from a set of GWAS loci.

11.5.2 Bipolar disorder exome challenge.

Given a set of 500 exomes (501,203 genomic variants), predictors were asked to identify Bipolar and healthy persons. A dataset of 500 exomes with known healthy status (249 cases and 251 controls) was provided as training dataset.

The quality check of all the 1,000 exomes did not discard any sample. The VEP based annotation procedure ended up with a dataset of 1,000 samples and 182,166 harmful variants (0 frameshift, 3,975 stop-gain, 168 stop-loss and 178,023 missense variants). SNPs&GO annotations were available for 19,019 missense variants, of which a small fraction of 14,955 were predicted as disease related.

In each sample, we labelled a gene as disease related if it carried at least one harmful variants in the homozygous form. By indicating with 0 a non-mutated gene, and with 1 a gene carrying the homozygous alternative form of harmful variations, we ended up with a 500 samples × 17,940 genes binary matrix. In one of the two methods we developed, missense variants were considered harmful only if predicted as disease related by SNPs&GO A second

binary matrix was built by considering the information of SNPs&GO. In this case we had a binary matrix of size 500 samples × 7,746 genes.

The two matrices were used separately in the gene selection procedure. In this step, we adopted a 5-CV procedure. It is worth to note that a substantial difference between the CD challenge and the BD challenge relied on the amount of training data provided: for the bipolar disorder challenge 500 of the 1,000 exomes were randomly selected and provided as training data. Thus, in this challenge we built five more informative and well-balanced subsets each one composed by 49-50 cases and 50-51 controls individuals, randomly selected.

By adopting a 5-CV procedure, 5 gene sets were returned. We tested the ability to discriminate of both the union and the intersection of the gene sets (the combination and the intersection of the 5 gene lists, respectively). For each implementation, **Table 10** lists the parameters and the performance obtained during the training phase.

All the models reached a quite good AUC score (0.71 – 0.82) in training. However, on the blind CAGI4 dataset, all them did not performed as expected (AUC 0.56 – 0.58). The best performance, with an AUC ~ 0.6, was reached by the models incorporating the genomic variants predicted as disease-related by SNPs&GO. However, when evaluated with average area under the ROC curve, a random predictor emerged.

Table 10. Bipolar disorder. Parameters and performance of the methods**.**

	With SNPs&GO		**Without SNPs&GO**	
	Union	**Intersection**	**Union**	**Intersection**
% of the healthy samples	<1%	<1%	<1%	<1%
% of the sick samples	>1%	>1%	>1%	>1%
n. of selected genes	1,047	393	1,006	506
AUC testing	0.79	0.82	0.71	0.74
AUC 500 exomes	0.58	0.58	0.56	0.56
AUC* 500 exomes (LCL, UCL)	0.51 (0.50, 0.52)	0.51 (0.50, 0.52)	0.52 (0.51, 0.52)	0.52 (0.51, 0.52)

*The average area under the ROC curves from the bootstrap sampling is presented (lower confidence limit, upper confidence limit). Confidence level was set at 0.95.

Among the nine participating groups, the top performing one reached an average area under the ROC curve of 0.60by implementing DeepBipolar, a Deep Convolutional Neural Network (DCNN) based tool (Laksshman *et al.*, 2017).

11.6 Conclusion

The use of patient's genotypic information is rapidly changing the way in which molecular knowledge is translated into health care, allowing the practice of genomic/personalized medicine. In this context, the exome challenges proposed in the CAGI4 experiment were aimed at identifying sick and healthy individuals based on exome sequencing data.

Several predictors joined the challenges, taking advantage from different data and computational methods. In addressing the Crohn's and Bipolar challenges, we build a very simple method that exploited the frequency of harmful mutations in the populations under investigation. Moreover, our method took advanced of SNPs&GO to classify a gene as disease-related (or not). However, despite the different variations of the method, the classifiers we built had a very poor performance.

Generally speaking, the methods that predictors applied in the CAGI4 edition were far apart from of being perfect classifiers. Even if some of those reached an AUC > 0.7 (Daneshjou *et al.*, 2017), the obtained results highlighted the fact that a lot of work is still necessary in the field of genomic/personalized medicine.

12 General conclusions

The functional interpretation of biological datasets (genes, proteins, metabolites, …) is not a trivial task. The cell is a complex entity: the different layers of information (the gene space, the protein space, the metabolite space, …) constantly "interact" and determine each other, in a non-unidirectional way. Thus, only by taking in consideration these interactions we can understand what is going on at the functional level. Practically speaking, we have to integrate data.

In this thesis, I described different methods for dissecting cell complexity: NET-GE, NET-GE^{M} and eDGAR. All the tools integrate data to better disclose functional processes at the basis of a phenotype. NET-GE relies on the STRING interactome to enhance the

understanding of the functional features of a set of genes as derived from the GO, KEGG and REACTOME resources; NET-GE^{M} aims at interpreting metabolomics data by using the strict relationships among metabolites, genes and proteins as derived from the HMDB and STRING resources; eDGAR collects human gene-disease associations as derived from different sources: OMIM, ClinVar, and Humsavar. To functionally characterize the diseases, eDGAR relies on interaction data (as derived from BIOGRID, STRING and TRRUST) and on NET-GE. Because of the unavailability of appropriate benchmark datasets, a rigorous evaluation of enrichment methods is not possible. However, the functional enrichments recovered with NET-GE have disclosed in different case studies relevant and non-trivial associations between phenotypes and molecular functions/pathways. In some case we were also able to mine PubMed for experimental evidences described in the available scientific literature. The functional terms retrieved by NET-GE and NET-GE^{M} resulted coherent with the investigated phenotypes, highlighting the efficacy of the methods in dissecting biological complexity by means of integrated data.

This thesis also describes my activities in the context of the CAGI experiment: a community experiment aimed at assessing computational methods for predicting the phenotypic impacts of genomic variants. Different challenges were proposed in the CAGI4 edition. I joined three of them: two challenges regarded the prediction of the health status of individuals form their exome variants (the Crohn's and Bipolar challenges), and one challenge related to the evaluation of the functional effect of mutations on the activity and allosteric regulation of the human pyruvate kinase. In addressing the challenges, ad hoc computational methods were built. The exome challenges took advantage of SNPs&GO to identify genes potentially discriminating sick from healthy persons. The L-PYK challenge relied on the BLOSUM62 matrix to score the effect of mutations on the L-PYK allosteric regulation. We did not make use of other data except the provided ones. We obtained remarkable results in the L-PYK challenge. In the case of exome challenges the results of the community experiment are generally poor (in particular for Bipolar disease) suggesting that the analysis of the genotype is still not sufficient to achieve a complete understanding of complex traits. This consideration strengthens the hypothesis that an effective analysis of complex phenotypes requires the integration of data describing the different components of the biological hierarchies. To this aim, Bioinformatics and Computational Biology play determinant roles, for the development of tools able to connect different levels of information, as NET-GE, NET-GE^{M} and eDGAR.

13 References

Alexa, A., Rahnenfuhrer, J., Lengauer, T. (2006) **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure**. *Bioinformatics*, 22:1600–1607.

Altmaier, E., Ramsay, S.L., Graber, A., Mewes, H.W., Weinberger, K.M., Suhre, K. (2008) **Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication**. *Endocrinology*, 149(7):3478-89

Babbi, G., Martelli, P.L., Profiti, G., Bovo, S., Savojardo, C., Casadio, R. (2017**) eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes**. *BMC Genomics*, 18(Suppl 5):554.

Bauer, M.S., Soloway, A., Dratman, M.B., Kreider, M. (1992) **Effects of hypothyroidism on rat circadian activity and temperature rhythms and their response to light**. *Biological Psychiatry*, 32(5):411–25.

Bender E. (2015) **Big data in biomedicine: 4 big questions.** *Nature*, 527(7576):S19.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000) **The Protein Data Bank.** *Nucleic Acids Research*, 28: 235-242.

Bessarabova, M., Ishkin, A., JeBailey, L., Nikolskaya, T., Nikolsky, Y. (2012) **Knowledge-based analysis of proteomics data.** *BMC Bioinformatics*, 13 Suppl. 16:S13.

Bilezikian, J.P., Khan, A., Potts, J.T.Jr., Brandi, M.L., Clarke, B.L., Shoback, D., Jüppner, H., D'Amour, P., Fox, J., Rejnmark, L., Mosekilde, L., Rubin, M.R., Dempster, D., Gafni, R., Collin,s M.T., Sliney, J., Sanders, J. (2011) **Hypoparathyroidism in the adult: epidemiology, diagnosis, pathophysiology, target-organ involvement, treatment, and challenges for future research.** *Journal of Bone and Mineral Research*, 26(10):2317-37

Bonvicini, C., Faraone, S.V., Scassellati, C. (2016) **Attention-deficit hyperactivity disorder in adults: A systematic review and meta-analysis of genetic, pharmacogenetic and biochemical studies.** *Molecular Psychiatry*, 21(7):872-84.

Bousquet, M., Gibrat, C., Ouellet, M., Rouillard, C., Calon, F., Cicchetti, F. (2010) **Cystamine metabolism and brain transport properties: clinical implications for neurodegenerative diseases.** *J Neurochem.*, 114(6):1651-8.

Bovo, S., Di Lena, P., Martelli, P., Fariselli, P., Casadio, R. (2017). **From Protein Variations to Biological Processes and Pathways with NET-GE.** *Genomics and Computational Biology*, 3(3), e45.

Bovo, S., Di Lena, P., Martelli, P.L., Fariselli, P., Casadio, R. (2016) **NET-GE: a web-server for NETwork-based human gene enrichment.** *Bioinformatics*, 32(22):3489-3491.

Brandes, U. (2008) **On variants of shortest-path betweenness centrality and their generic computation.** *Social Networks*, 30(2):136–145.

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., Dolinski, K., Tyers M. (2008) **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Research*, 36(Database issue):D637-40.

Browne, H.A., Gair, S.L., Scharf, J.M., Grice, D.E. (2014) **Genetics of obsessive-compulsive disorder and related disorders.** *Psychiatric Clinics of North America*, 37(3):319-35.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., Casadio, R. (2009). **Functional annotations improve the predictive score of human disease-related mutations in proteins.** *Human Mutation*, 30(8):1237-1244.

Cappi, C., Brentani, H., Lima, L., Sanders, S.J., Zai, G., Diniz, B.J., Reis, V.N., Hounie, A.G., Conceição do Rosário, M., Mariani, D., Requena, G.L., Puga, R., Souza-Duran, F.L., Shavitt, R.G., Pauls, D.L., Miguel, E.C., Fernandez, T.V. (2016) **Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways.** *Translational Psychiatry*, 6:e764.

Carlson, G.M., Fenton, A.W. (2016) **What Mutagenesis Can and Cannot Reveal About Allostery.** *Biophysical Journal*, 110(9):1912-23

Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothar,i A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D. (2016) **The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases**. *Nucleic Acids Research* ,44(D1):D471-80.

Casti, J., Karlqvist, A. (2003) **Art and Complexity**. JAI, Amsterdam. p.86.

Centonze, D., Bari, M., Di, Michele. B,, Rossi, S., Gasperi, V., Pasini, A., Battista, N., Bernardi, G., Curatolo, P., Maccarrone, M. (2009) **Altered anandamide degradation in attention-deficit/hyperactivity disorder.** *Neurology*, 2(17):1526-7.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., Cesareni, G. (2007) **MINT: the Molecular INTeraction database.** *Nucleic Acids Research*, 35(Database issue):D572-4.

Chen, L., Wang, R. S., & Zhang, X. S. (2009) **Network-Based Prediction of Protein Function.** *Biomolecular Networks: Methods and Applications in Systems Biology*, 231-278.

Cheng, D., Jenner, A.M., Shui, G., Cheong, W.F., Mitchell, T.W., Nealon, J.R., Kim, W.S., McCann, H., Wenk, M.R., Halliday, G.M., Garner, B. (2011) **Lipid pathway alterations in Parkinson's disease primary visual cortex**. *PLoS One*, 6(2):e17299.

Cho, J. H. (2008). **The genetics and immunopathogenesis of inflammatory bowel disease.** *Nature Reviews Immunology*, 8(6):458-466.

Cornish, A.J., Markowetz, F. (2014) **SANTA: Quantifying the Functional Content of Molecular Networks.** *PLOS Computational Biology*, 10:e1003808.

Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., Wu, J. (2012) **PINA v2.0: mining interactome modules.** *Nucleic Acids Research*, 40: D862-865.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L. (2011) **Reactome: a database of reactions, pathways and biological processes.** *Nucleic Acids Research*, 39(Database issue):D691-7.

Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, Lena PD, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat RR, Li X, Pal LR, Kundu K, Yin Y, Moult J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofran Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen BS, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman RB, Klein TE, Hoskins RA, Repo S, Brenner SE, Morgan AA. (2017) **Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges.** *Human Mutation*, 38(9):1182-1192.

De Las Rivas, J., Fontanillo, C. (2010**) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.** *PLoS Computational Biology*, 6(6):e1000807.

Di Lena, P., Martelli, P.L., Fariselli, P., Casadio, R. (2015) **NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases.** *BMC Genomics*, 16 Suppl 8:S6.

Dombrauckas, J. D., Santarsiero, B. D., Mesecar, A. D. (2005). **Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis.** *Biochemistry*, 44(27):9417-9429.

Edden, R.A., Crocetti, D., Zhu, H., Gilbert, D.L., Mostofsky, S.H. (2012) **Reduced GABA concentration in attention-deficit/hyperactivity disorder.** *Archives of General Psychiatry*, 69(7):750-3.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z. (2009). **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.** *BMC Bioinformatics*, 10:48.

Egri, P., Gereben, B. (2014) **Minimal requirements for ubiquitination-mediated regulation of thyroid hormone activation.** *Journal of Molecular Endocrinology*, 53(2):217-26.

Erten, S., Bebek, G., Ewing, R.M., Koyutürk, M. (2011) DADA: **Degree-Aware Algorithms for Network-Based Disease Gene Prioritization.** *BioData Mining*, 4:19.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P. (2016) **The Reactome pathway Knowledgebase.** *Nucleic Acids Research*, 44(D1):D481-7.

Fenton, A. W., & Hutchinson, M. (2009a). **The pH dependence of the allosteric response of human liver pyruvate kinase to fructose-1, 6-bisphosphate, ATP, and alanine.** *Archives of biochemistry and biophysics*, 484(1):16-23.

Fenton, A. W., Alontaga, A. Y. (2009b). **The impact of ions on allosteric functions in human liver pyruvate kinase.** *Methods in enzymology*,466:83-107.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., Altman, R. B. (2011). **Bioinformatics challenges for personalized medicine.** *Bioinformatics*,27(13):1741-1748.

Feuerstein, J.D., Cheifetz, A.S. (2017) **Crohn Disease: Epidemiology, Diagnosis, and Management.** *Mayo Clinic Proceedings*, 92(7):1088-1103.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., ... & Gough, J. (2016b). **InterPro in 2017—beyond protein family and domain annotations.** *Nucleic Acids Research*, 45(D1): D190-D199.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. (2016a) **The Pfam protein families database: towards a more sustainable future.** *Nucleic Acids Research*, 44(D1):D279-85.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., ... & Anderson, C. A. (2010). **Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.** *Nature genetics*, 42(12):1118-1125.

Gamble, K.L., May, R.S., Besing, R.C., Tankersly, A.P., Fargason, R.E. (20139 **Delayed sleep timing and symptoms in adults with attention-deficit/hyperactivity disorder: a controlled actigraphy study.** *Chronobiology International*, 30(4):598-606.

Gardoni, F., Bellone, C. (2015) **Modulation of the glutamatergic transmission by Dopamine: a focus on Parkinson, Huntington and Addiction diseases.** *Front Cell Neurosci.*, 9:25.

Gene Ontology Consortium. (2015) **Gene Ontology Consortium: going forward.** *Nucleic Acids Research*, 43: D1049-1056.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E., Weinberger, K.M., Adamski, J., Illig, T., Suhre, K. (2008) **Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum.** *PLoS Genet*, 4(11):e1000282.

Glaab E, Baudot A, Krasnogor N, Valencia A. (2010) **TopoGSA: network topological gene set analysis.** *Bioinformatics*, 26(9):1271-2.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., Valencia, A. (2012) **EnrichNet: network-based gene set enrichment analysis.** *Bioinformatics*, 28(18):i451-i457.

Goes, F.S. (2015) **Genetics of Bipolar Disorder: Recent Update and Future Directions.** *Psychiatric Clinics of North America*, 39(1):139-55.

Grossmann, S., Bauer, S., Robinson, P.N., Vingron, M. (2007) **Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis.** *Bioinformatics*, 23(22):3024-31.

Güldener, U., Münsterkötter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W., Stümpflen, V. (2006) **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Research*, 34(Database issue):D436-41.

Gullapalli, R. R., Desai, K. V., Santana-Santos, L., Kant, J. A., & Becich, M. J. (2012). **Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics.** *Journal of pathology informatics*, 3:40.

Gupta, V., Bamezai, R. N. (2010). **Human pyruvate kinase M2: a multifunctional protein.** *Protein science*,19(11):2031-2044.

Hajian-Tilaki, K. (2013). **Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation.** *Caspian journal of internal medicine*, 4(2): 627.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A. (2005**) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Research*, 33:D514-7.

Han, H., Cho, J.W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.N., Jung, H., Nam, S., Chung, M., Kim, J.H., Lee, I. (2017) **TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions**. *Nucleic Acids Res.* 2017 Oct 26. doi:10.1093/nar/gkx1013.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., Steinbeck, C. (2013) **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013**. *Nucleic Acids Research*, 41(Database issue):D456-63.

Hines, D. J., Haydon, P. G. (2014). **Astrocytic adenosine: from synapses to psychiatric disorders.** *Philosophical Transactions of the Royal Society B*, 369(1654), 20130594.

Hollywood, K., Brison, D.R., Goodacre, R. (2006) **Metabolomics: current technologies and future trends.** *Proteomics*, 17:4716-23

Hoskins, R.A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., Brenner, S.E. (2017) **Reports from CAGI: The critical assessment of genome interpretation.** *Human Mutation*, 38:1039–1041.

Huang, D.W., Sherman, B.T., & Lempicki, R.A. (2009a) **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Research*, 37:1-13.

Huang, D.W., Sherman, B.T., Lempicki, R.A. (2009b) **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature Protocols*, 4(1):44-57.

Huang, S.A., Mulcahey, M.A., Crescenzi, A., Chung, M., Kim, B.W., Barnes, C., Kuijt, W., Turano, H., Harney, J., Larsen, P.R. (2005) **Transforming growth factor-beta promotes**

inactivation of extracellular thyroid hormones via transcriptional stimulation of type 3 iodothyronine deiodinase. *Molecular Endocrinology*, 19(12):3126-36.

Hung, J.H., Whitfield, T.W., Yang, T.H., Hu, Z., Weng, Z., DeLisi, C. (2010) **Identification of functional modules that correlate with phenotypic difference: the influence of network topology.** *Genome Biology*, 11(2):R23.

Iglesias, A., Anyane-Yeboa, K., Wynn, J., Wilson, A., Cho, M. T., Guzman, Sisson, R., Egan, C., Chung, W.K. (2014). **The usefulness of whole-exome sequencing in routine clinical practice.** *Genetics in Medicine*, 16(12):922-931.

Irizarry, R.A., Wang, C., Zhou, Y., Spee,d T.P. (2011) **Gene set enrichment analysis made simple.** *Statistical Methods in Medical Research*, 18(6):565-75.

Ishwar, A., Tang, Q., Fenton, A. W. (2015**). Distinguishing the interactions in the fructose 1, 6-bisphosphate binding site of human liver pyruvate kinase that contribute to allostery.** *Biochemistry*, 54(7):1516-1524.

Jacob, M., Lopata, A. L., Dasouki, M., & Abdel Rahman, A. M. (2017). **Metabolomics toward personalized medicine**. *Mass Spectrometry Reviews*, doi: 10.1002/mas.21548.

Junker, B.H., Schreiber, F. (2008) **Analysis of Biological Networks.** Wiley, Hoboken (NJ), USA.

Kalia, L.V., Lang, A.E. (2015) **Parkinson's disease.** *Lancet*, 386(9996):896-912.

Kamburov, A., Cavill, R., Ebbels, T.M., Herwig, R., Keun, H.C. (2011) **Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.** *Bioinformatics*, 27(20):2917-8.

Kanehisa M. (2013) **Chemical and genomic evolution of enzyme-catalyzed reaction networks**. *FEBS Lett.* 587(17):2731-2737.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2014**) Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Research*, 42(Database issue):D199-205

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. (2016) **KEGG as a reference resource for gene and protein annotation.** *Nucleic Acids Research*, 44(D1):D457-62.

Kankainen, M., Gopalacharyulu, P., Holm, L., Oresic, M. (2011) **MPEA--metabolite pathway enrichment analysis**. *Bioinformatics*, 27(13):1878-9.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., Hermjakob, H. (2006) **IntAct-open source resource for molecular interaction data.** *Nucleic Acids Research*, 35(Database issue):D561-5.

Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B.A., Wang, J., Yu, B., Zhang, J., Bryant, S.H. (2016**) PubChem Substance and Compound databases.** *Nucleic Acids Research*, 4;44(D1):D1202-13.

Kori, M., Aydın, B., Unal, S., Arga, K.Y., Kazan, D. (2016) **Metabolic Biomarkers and Neurodegeneration: A Pathway Enrichment Analysis of Alzheimer's Disease, Parkinson's Disease, and Amyotrophic Lateral Sclerosis**. *OMICS*, 20(11):645-661.

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E.L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S.R., Miller, R., Coort, S.L., Cirillo, E., Smeets, B., Evelo, C.T., Pico, A.R. (2016) **WikiPathways: capturing the full diversity of pathway knowledge**. *Nucleic Acids Research*, 44(D1):D488-94.

Laksshman, S., Bhat, R.R., Viswanath, V., Li, X. (2017) **DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning.** *Hum Mutation*, 38(9):1217-1224.

Landrum, M,J., Lee J,M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., Maglott, D.R. (2016) **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Research*, 44(D1):D862–D868.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P. (2011) **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics*, 27(12):1739-40.

Lin, Dekang. (1998). **An Information-Theoretic Definition of Similarity.** *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, 296-304.

Liu, J. Z., Anderson, C. A. (2014). **Genetic studies of Crohn's disease: past, present and future.** *Best Practice & Research Clinical Gastroenterology*, 28(3):373-386.

Liu, K., Li, F., Han, H., Chen, Y., Mao, Z., Luo, J., Zhao, Y., Zheng, B., Gu, W., Zhao, W. (2016) **Parkin Regulates the Activity of Pyruvate Kinase M2.** *J Biol Chem.*, 291(19):10307-17.

López-Ibáñez, J., Pazos, F., Chagoyen, M. (2016) **MBROLE 2.0 -functional enrichment of chemical compounds.** *Nucleic Acids Research*, 44(W1):W201-4.

Louie, B., Bergen, S., Higdon, R., Kolker, E. (2010) **Quantifying protein function specificity in the gene ontology.** *Standards in Genomic Sciences*, 2(2):238-44.

Maden, M. (2007) **Retinoic acid in the development, regeneration and maintenance of the nervous system.** *Nat Rev Neurosci*, 8(10):755-65.

Maltezos, S., Horder, J., Coghlan, S., Skirrow, C., O'Gorman, R., Lavender, T.J., Mendez, M.A., Mehta, M., Daly, E., Xenitidis, K., Paliokosta, E., Spain, D., Pitts, M., Asherson, P., Lythgoe, D.J., Barker, G.J., Murphy, D.G. (2014) **Glutamate/glutamine and neuronal integrity in adults with ADHD: a proton MRS study.** *Translational Psychiatry*, 4:e373.

Manolio, T. A., Chisholm, R. L., Ozenberger, B., Roden, D. M., Williams, M. S., Wilson, R., Bick, D., Bottinger, E.P., Brilliant, M.H., Eng, C., Frazer, K.A., Korf, B., Ledbetter, D.H., Lupski, J.R., Marsh, C., Mrazek, D., Murray, M.F., O'Donnell, P.H., Rader, D.J., Relling, M.V., Shuldiner, A.R., Valle, D., Weinshilboum, R., Green, E.D., Ginsburg, G.S. (2013). **Implementing genomic medicine in the clinic: the future is here.** *Genetics in Medicine*, 15(4):258-267.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A. (2010) **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Research*, 20(9):1297-303.

McLachlan, A.D. (1972) **Repeating sequences and gene duplication in proteins**. *J. Mol. Biol.* 64:417–437.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., Cunningham, F. (2016) **The Ensembl Variant Effect Predictor.** *Genome Biology*, 17(1):122.

Mehta, V., Trinkle-Mulcahy, L. (2016) **Recent advances in large-scale protein interactome mapping.** *F1000Research*, 5. pii: F1000 Faculty Rev-782.

Mermi, O., Atmaca, M. (2016). **Thyroid gland functions are affected in obsessive compulsive disorder.** *Anadolu psikiyatri dergisi-anatolian journal of psychiatry*, 17(2):99-103.

Micheli, V., Camici, M., Tozzi, M.G., Ipata, P.L., Sestini, S., Bertelli, M., Pompucci G. (2011) **Neurological disorders of purine and pyrimidine metabolism.** *Current Topics in Medicinal Chemistry*, 11(8):923-47.

Miksys SL, Tyndale RF. (2002) **Drug-metabolizing cytochrome P450s in the brain.** *J Psychiatry Neurosci*. 27(6):406-15.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., Stein, L. (2012) **Annotating cancer variants and anti-cancer therapeutics in reactome.** *Cancers*, 4(4):1180-211.

Misra K.B. (2008) **Engineering Design: A Systems Approach**. *Handbook of Performability Engineering. Springer, London.* pp 13-24.

Moretti, A., Gorini, A., Villa, R. F. (2003**). Affective disorders, antidepressant drugs and brain metabolism.** *Molecular psychiatry*, 8(9):773-785.

Morgan, H. P., O'Reilly, F. J., Wear, M. A., O'Neill, J. R., Fothergill-Gilmore, L. A., Hupp, T., Walkinshaw, M. D. (2013). **M2 pyruvate kinase provides a mechanism for nutrient sensing and regulation of cell proliferation.** *Proceedings of the National Academy of Sciences*, 110(15):5881-5886.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M. (2005) **Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.** *Bioinformatics*, 21 Suppl 1:i302-10.

Natale, M., Benso, A., Di Carlo, S., Ficarra, E. (2014) **FunMod: a Cytoscape plugin for identifying functional modules in undirected protein-protein networks.** *Genomics Proteomics Bioinformatics*, 12(4):178-86.

Noble, W.S. (2009) **How does multiple testing correction work?** Nat Biotechnol. 27(12):1135-7.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., … Hermjakob, H. (2012). **Protein Interaction Data Curation - The International Molecular Exchange Consortium (IMEx).** *Nature Methods*, *9*(4), 345–350

Oti, M., Brunner H.G. (2007). **The modular nature of genetic diseases**. *Clin Genet*. 71(1):1-11.

Park, S.G., Schimmel, P., Kim, Sunghoon. (2008) **Aminoacyl tRNA synthetases and their connections to disease.** *Proc Natl Acad Sci U S A*., 105(32): 11043–11049

Pauls, D.L. (2010) **The genetics of obsessive-compulsive disorder: a review.** *Dialogues in Clinical Neuroscience*, 12(2):149-63.

Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. (2003) **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Research*. 2003 13(10):2363-71.

Pevsner, J. (2015) **Bioinformatics and Functional Genomics, 3rd edition**. Wiley-Blackwell, Hoboken (NJ), USA.

Phizicky, E.M., Fields, S. (1995) **Protein-protein interactions: methods for detection and analysis.** *Microbiololical Reviews*, 59(1):94-123.

Prasannan, C. B., Tang, Q., Fenton, A. W. (2012). **Allosteric regulation of human liver pyruvate kinase by peptides that mimic the phosphorylated/dephosphorylated N-terminus.** *Allostery: Methods and Protocols*, 335-349.

Resnik, P. (1995) **Using information content to evaluate semantic similarity in a taxonomy.** *Proceedings of the 14th international joint conference on Artificial Intelligence*, pp. 448–453.

Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S. (2008) **Use and misuse of the gene ontology annotations.** *Nature Reviews Genetics*, 9(7):509-15.

Robinson P.N. (2014) **Genomic data sharing for translational research and diagnostics.** *Genome Medicine*, 6: 78.

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., Peyvandi, AA. (2014) **Protein-protein interaction networks (PPI) and complex diseases.** *Gastroenterology and Hepatology from Bed to Bench*, (1):17-31.

Sagar, R., Pattanayak, R. D. (2017). **Potential biomarkers for bipolar disorder: Where do we stand?** *The Indian Journal of Medical Research*, 145(1):7.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D. (2004) **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Research*, 32(Database issue):D449-51.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H. (2009) **PID: the Pathway Interaction Database.** *Nucleic Acids Research*, 37:D674-9.

Schreuder, D.A. (2014) **Vision and Visual Perception: The Conscious Base of Seeing**. Archway Publishing, Bloomington. p324.

Sharan, R., Ulitsky, I., & Shamir, R. (2007). **Network-based prediction of protein function.** *Molecular systems biology*, 3(1): 88.

Sharma, A., Couture, J. (2013) **A review of the pathophysiology, etiology, and treatment of attention-deficit hyperactivity disorder (ADHD).** *Annals of Pharmacotherapy*, 48(2):209-25.

Shoback, D. (2008). **Hypoparathyroidism**. *New England Journal of Medicine*, 359(4):391-403.

Smeyne M, Smeyne RJ. (2013) **Glutathione metabolism and Parkinson's disease.** *Free Radic Biol Med*, 262:13-25.

Smoller J, W., Finn, C.T. (2003) **Family, twin, and adoption studies of bipolar disorder.** *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 123C(1):48-58.

Snel, B., Lehmann, G., Bork, P., Huynen, M.A. (2000) **STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene**. *Nucleic Acids Research*, 28(18):3442-4.

Song, X., Li, L., Srimani, P.K., Yu, P.S., Wang, J.Z. (2014) **Measure the Semantic Similarity of GO Terms Using Aggregate Information Content.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(3):468-76.

Steventon GB, Sturman S, Waring RH, Williams AC. (2001) **A review of xenobiotic metabolism enzymes in Parkinson's disease and motor neuron disease.** *Drug Metabol Drug Interact.* 18(2):79-98.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. (2005) **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences*,102(43):15545-50.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., von Mering, C. (2015**) STRING v10: protein-protein interaction networks, integrated over the tree of life.** *Nucleic Acids Research*, 43:D447-52.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C. (2017) **The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.** *Nucleic Acids Research*, 45(D1):D362-D368.

Taft, R. J., & Mattick, J. S. (2003). **Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences.** *Genome Biology*, 5(1), P1.

Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P., Romero, R. (2009) **A novel signaling pathway impact analysis.** *Bioinformatics*, 25(1):75-82.

Tarver, J., Daley, D., Sayal, K. (2014) **Attention-deficit hyperactivity disorder (ADHD): an updated review of the essential facts.** *Child: Care, Health and Development*, 40(6):762-74.

Thomas, R., Gohlke, J.M., Stopper, G.F., Parham, F.M., Portier, C.J. (2009**) Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure.** *Genome Biology*, 10:R44.

Tipney, H., Hunter, L. (2010) **An introduction to effective use of enrichment analysis software.** *Human Genomics*, 4(3):202-6.

Tripathi, S., Moutari, S., Dehmer, M., & Emmert-Streib, F. (2016**). Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules.** *BMC bioinformatics*, 17(1): 129.

UniProt Consortium, (20015) **UniProt: a hub for protein information**. *Nucleic Acids Research*, 43:D204–D212.

Valencia A. (2005) **Automatic annotation of protein function**. Curr Opin Struct Biol 15(3):267-74.

Videnovic, A., Golombek, D. (2013) **Circadian and sleep disorders in Parkinson's disease.** *Exp Neurol.*, 243:45-56.

Vinayavekhin, N., Homan, E.A., Saghatelian, A. (2010) **Exploring disease through metabolomics.** *ACS Chem Biol*, 5(1):91-103.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P. (2002) **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature*, 417(6887):399-403.

Wei, Q., Dunbrack Jr, R. L. (2013). **The role of balanced training and testing data sets for binary classifiers in bioinformatics.** *PloS one*, 8(7): e67863.

Winterhalter, C., Widera, P., Krasnogor, N. (2014) **JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks.** *Bioinformatics*, 30:1029–1030.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R. (2013) **HMDB 3.0 — The Human Metabolome Database in 2013.** *Nucleic Acids Research*, 41(D1):D801-7.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T.P., Hautaniemi, S. (2009) **Integrated network analysis platform for protein-protein interactions.** *Nature methods*, 6:75-77.

Xia, J., Wishart, D.S. (2010a) **MetPA: a web-based metabolomics tool for pathway analysis and visualization.** *Bioinformatics*, 26(18):2342-4.

Xia, J., Wishart, D.S. (2010b) **MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data**. *Nucleic Acids Research*, 38(WebServer issue)W71-7.

Xu, Q., Tang, Q., Katsonis, P., Lichtarge, O., Jones, D., Bovo, S., Babbi, G., Martelli, P.L., Casadio, R., Lee, G.R., Seok, C., Fenton, A.W., Dunbrack, R.L. Jr. (2017). **Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4.** *Human Mutation*, 38(9):1123-1131.

Yang, N., Yao, Z., Miao, L., Liu, J., Gao, X., Fan, H., Hu, Y., Zhang, H., Xu, Y., Qu, A., Wang, G. (2015) **Novel clinical evidence of an association between Homocysteine and insulin resistance in patients with hypothyroidism or subclinical hypothyroidism.** *PLoS One*, 10(5):e0125922.

Yang, Q., Wang, S., Dai, E., Zhou, S., Liu, D., Liu, H., Meng, Q., Jiang, B., Jiang W. (2017) **Pathway enrichment analysis approach based on topological structure and updated annotation of pathway.** *Briefings in Bioinformatics*, doi: 10.1093/bib/bbx091.

Yu, W., Clyne, M., Khoury, M.J., Gwinn, M. (2010) **Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations.** *Bioinformatics*, 26(1):145-6.

Zhu, J., Reith, M.E. (2008) **Role of the dopamine transporter in the action of psychostimulants, nicotine, and other drugs of abuse.** *CNS & Neurological Disorders-Drug Targets*, 7(5):393-409.

14 Supplementary Material

Table S1. *de novo* mutation identified as characterizing the OCD cohort.

Sample	Gene	Variant effect	Substitution	Detection*
OCD016301	SNUPN	Missense	D255E	Both
OCD018901	ATP2B2	Missense	I1084T	Both
OCD018901	DNAJC19	Silent	R83R	SAMtools
OCD032201	CYP27A1	Silent	P61P	SAMtools
OCD032201	WWP1	Missense	I310M	Both
OCD032201	ERCC6	Nonsense	K1210X	Both
OCD032201	GANAB	Silent	L311L	SAMtools
OCD129101	CCDC108	Missense	E484K	Both
OCD129101	BANK1	Missense	K633M	Both
OCD129101	MYO10	Missense	E199K	Both
OCD139801	GBP4	Silent	T363T	SAMtools
OCD139801	FAM5B	Missense	NA	Both
OCD139801	CR1	Missense	S1748R	Both
OCD139801	SGPP2	Silent	P303P	SAMtools
OCD144601	ACCN4	Silent	X667X	SAMtools
OCD144601	VCX2	Missense	A70G	Both
OCD175901	CIITA	Silent	D1058D	SAMtools
OCD175901	AP1G1	Missense	K155E	Both
OCD181401	BAMBI	Missense	V260I	Both
OCD176501	NDE1	Missense	A986C	GATK
OCD018901	SMAD4	Missense	W302R	GATK
OCD020001	MUC5B	Missense	A4261E	GATK
OCD003301	ARHGAP6	Missense	S134F	GATK
OCD043301	CHD8	Missense	E1327K	GATK
OCD048501	ABCE1	Missense	P243A	GATK
OCD048501	SLC35G5	Missense	S114N	GATK
OCD048501	SAA2	Missense	K102E	GATK

*To ensure the discovery of the maximum number of SNPs, authors used two different bioinformatic pipelines to call SNPs. In the first pipeline, BWA v.0.7.10 (option "aln") was used to align the reads against the human reference genome built 37. Aligned reads were trimmed to the exome target, PCR duplicates were removed using SAMtools and SNPs were called by using the SAMtool option "pileup" (default parameters). In the second pipeline, alignment and variant calling followed the GATK v3 guidelines. Reads were aligned using BWA (option "MEM") and PCR duplicates were marked using Picard v.1.118. GATK v.3.2.3 was used to realign INDELS, recalibrate the quality scores and for SNP calling. To be called, at least 8 unique reads supporting each SNP were required. Data are taken from Cappi *et al.,* 2016.

Table S2. Metabolic set involved in Parkinson's disease. Data are from Kori *et al.*, 2016.

Metabolite	KEGG ID	Status	Tissue	Technique
Acetate	C00033	Decreased	Plasma	1H-NMR
N-acetyl-L-aspartate	C01042	-	S. Nigra	MRS
Alanine	C01401	Decreased	CSF	1H-NMR
γ-aminobutyric acid	C00334	Increased	S. Nigra	3D-MRI
Arginine	C02385	Decreased	Serum	1H-NMR
Ascorbate	C00072	Decreased	Plasma	1H-NMR
Asparagine	C16438	Decreased increased	Serum/CSF/blood	CE-MS/UPLC-ToF-MS
Aspartate	C16433	Increased	CSF	1H-NMR
Choline	C00114	Decreased	S. Nigra	DIES-MS
Citrate	C00158	Decreased	Plasma	1H-NMR
Creatine	C00300	Decreased	S. Nigra	3D-MRI
Creatinine	C00791	Decreased	CSF	GC-ToFMS
Cystine	C01420	Decreased	Plasma	LC-QToF-MS
Dopamine	C03758	Decreased	S. Nigra	3D-MRI
Ethanolamine	C00189	Decreased	Plasma	1H-NMR
Fatty acids	C00162	Increased	Serum	DIES-MS
Galacticol	C01697	Decreased	Plasma	1H-NMR
Gluconate	C00257	Decreased	Plasma	1H-NMR
Glutamic acid	C00302	Decreased	Plasma	LC-QToF-MS
Glutamate	C00025	Decreased Increased	CSF/plasma	1H-NMR
L-glutamine	C00064	Decreased	CSF	1H-NMR
Glutarate	C00489	Decreased	Plasma	1H-NMR
Glutathione	C00051	Decreased	S. Nigra	3D-MRI
Glycine	C00037	Increased	CSF	1H-NMR
Glycolate	C00160	Decreased	Plasma	1H-NMR
Glycerol	C00116	Decreased	Plasma	1H-NMR
24S-hydroxycholesterol (24S-OH)	C13550	Decreased	Plasma	GC-MS/GC-ToFMS
3-hydroxyisovaleric acid	C01013	Decreased	CSF	GC-ToFMS
Homovanillic acid	C05582	Increased	S. Nigra	3D-MRI
Isoleucıne	C16434	Increased	Serum	1H-NMR
Isocitrate	C00311	Decreased	Plasma	1H-NMR
Lipid hydroperoxides	C01025	Increased	Plasma	Chemiluminescence
Lysine	C16440	Decreased	CSF	1H-NMR
Malate	C00149 C00711	Decreased	Plasma	1H-NMR
Malondialdehyde (malonaldehyde)	C19440	Increased	Plasma	HPLC
Methionine	C01733	Decreased	Serum/PMV-CSF	1H-NMR/LCECA
Methylamine	C00218	Decreased	Plasma	1H-NMR
Methylmalonate	C02170	Decreased	Plasma	1H-NMR
Myoinositol	C00137	Decreased Increased	S. Nigra/plasma	3D-MRI/1H-NMR
2-oxoisocaproate (ketoleucine) (4-methyl-2-oxopentanoat)	C00233	Increased	Plasma	GC-ToFMS
Proline	C16435	Increased	Plasma	LC-QToF-MS
Propylene glycol	C00583	Increased	Plasma	1H-NMR
Pyroglutamate	C01879	-	Plasma	GC-ToFMS

Pyruvate	C00022	Increased	Plasma	1H-NMR/ UHPLC-MS-MS
Sorbitol (glucitol)	C00794	Increased	Plasma	1H-NMR
Suberate	C08278	Decreased	Plasma	1H-NMR
Succinate (unmedicated)	C00042	Decreased	Plasma	1H-NMR
Superoxide dismutase	K04564	Increased	Plasma	HPLC
Threonate	C01620	Decreased	Plasma	1H-NMR
Trimethylamine	C00565	Decreased	Plasma	1H-NMR
Tryptophan	C00078	Decreased	CSF	GC-ToFMS
Uric acid	C00366	Decreased	Plasma	Uricase
Urinary biopyrrin	C00486	Increased	Urine	ELISA
Valine	C16436	Increased	Serum	1H-NMR

15 List of Publications

15.1 Peer-reviewed publications related to this thesis

1. Babbi G, Martelli PL, Profiti G, **Bovo S**, Savojardo C and Casadio R. (2017) **eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes.** *BMC Genomics*, 11;18(Suppl5):554.

2. **Bovo S**, Di Lena P; Martelli PL, Fariselli P, Casadio R. (2017) **From Protein Variations to Biological Processes and Pathways with NET-GE.** *Genomics and Computational Biology*, 3(3); e45

3. Xu Q, Tang Q, Katsonis, Lichtarge O, Jones DT, **Bovo S**, Babbi G, Martelli PL, Casadio R, Lee GR, Seok C, Fenton AW and Dunbrack RL. (2017) **Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4.** *Hum Mutation*, 38(9):1123-1131.

4. Daneshjou R, Wang Y, Bromberg Y, **Bovo S**, Martelli PL, Babbi G, Di Lena P, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat R, Li X, Pal L R, Kundu K, Yin Y, Moult J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofran Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen BS, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman R, Klein TE, Hoskins R, Repo S, Brenner SE, Morgan AA. (2017) **Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges.** *Hum Mutation*, 38(9):1182-1192.

5. **Bovo S***, Di Lena P*, Martelli PL, Fariselli P and Casadio R. (2016) **NET-GE: a web-server for NETwork-based Gene Enrichment of sets of human genes related to the same phenotype.** *Bioinformatics*, 32(22):3489-3491. [*First authorship shared]

15.2 Other peer-reviewed publications

1. Utzeri VJ, Ribani A, Schiavo G, Bertolini F, **Bovo S**, Fontanesi L(2018) **Application of next generation semiconductor based sequencing to detect the botanical composition of monofloral, polyfloral and honeydew honey**. *Food Control*. 86: 342-349.

2. Ribani A, Schiavo G, Utzeri VJ, Bertolini F, Geraci C, **Bovo S**, Fontanesi L (2017) **Application of next generation semiconductor based sequencing for species identification in dairy products.** *Food Chemistry*, Available online 3 November 2017, https://doi.org/10.1016/j.foodchem.2017.11.006.

3. **Bovo S**, Mazzoni G, Ribani A, Utzeri JV, Bertolini F. Schiavo G, Fontanesi L. (2017) **A viral metagenomic approach on a non-metagenomic experiment: mining next generation sequencing datasets from pig DNA identified several porcine parvoviruses for a retrospective evaluation of viral infections**. *PLoS One*. 12(6):e0179462.

4. Schiavo G, Hoffmann OI, Ribani A, Utzeri VJ, Ghionda MC, Bertolini F, Geraci C, **Bovo S**, Fontanesi L. (2017) **A genomic landscape of mitochondrial DNA insertions in the pig nuclear genome provides evolutionary signatures of interspecies admixture.** *DNA Research* 2017 dsx019. doi:10.1093/dnares/dsx019 .

5. Fontanesi L, Schiavo G, Galimberti G, **Bovo S**, Russo V, Gallo M, Buttazzoni L. (2017) **A genome wide association study for a proxy of intermuscular fat level in the Italian Large White breed identifies genomic regions affecting an important quality parameter for dry-cured hams.** *Anim Genet*. 48(4):459-465;

6. Fontanesi L, Schiavo G, Gallo M, Baiocco C, Galimberti G, **Bovo S**, Russo V, Buttazzoni L. (2017) **Genome-wide association study for ham weight loss at first salting in Italian Large White pigs: towards the genetic dissection of a key trait for dry-cured ham production.** *Anim Genet*, 48(1):103-107.

7. **Bovo S***, Schiavo G*, Mazzoni G*, Dall'Olio S, Galimberti G, Calò DG, Scotti E, Bertolini F, Buttazzoni L, Samorè AB, Fontanesi L. (2016**) Genome wide association study for the level of serum electrolytes in Italian Large White pigs.** *Anim Genet*, 47(5):597-602. [*First authorship shared]

8. **Bovo S***, Mazzoni G*, Galimberti G, Calò DG, Fanelli F, Mezzullo M, Schiavo G, Manisi A, Trevisi P, Bosi P, Dall'Olio S, Pagotto U, Fontanesi L. (2016) **Metabolomics evidences plasma and serum biomarkers differentiating two heavy pig breeds.** *Animal*. 8:1-8. [*First authorship shared]

9. **Bovo S***, Mazzoni G*, Calò DG, Galimberti G, Fanelli F, Mezzullo M, Schiavo G, Scotti E, Manisi A, Samorè AB, Bertolini F, Trevisi P, Bosi P, Dall'Olio S, Pagotto U and Fontanesi L. (2015) **Deconstructing the pig sex metabolome: Targeted metabolomics in heavy pigs revealed sexual dimorphisms in plasma biomarkers and metabolic pathways.** *J Anim Sci*. [*First authorship shared]

10. **Bovo S**, Bertolini F, Schiavo G, Mazzoni G, Dall'Olio S, Fontanesi L. (2015) **Reduced Representation Libraries from DNA Pools Analysed with Next Generation Semiconductor Based-Sequencing to Identify SNPs in Extreme and Divergent Pigs for Back Fat Thickness.** *Int J Genomics*. 2015:950737.

15.3 Abstracts and Posters

1. Babbi G, Martelli PL, Profiti G, **Bovo S**, Savojardo C and Rita Casadio (2017) **Analysing the relations among genes and polygenic diseases with eDGAR.** F1000Research, doi:10.7490/f1000research.1114767.1

2. Fontanesi L, Bertolini F, Bovo S, Rothschild M.F. (2017) **Characterization of tandem repeat regions in the european sea bass *dicentrarchus labrax* genome** Poster presented at Aquaculture Europe 2017 – Dubrovnik, Croatia – October 17-20, 2017

3. **Bovo S**, Mazzoni G, Ribani A, Utzeri VJ, Bertolini F, Schiavo G, Fontanesi L. (2017) **A metagenomics study on a non-metagenomics experiment: mining next generation sequencing datasets from porcine DNA identified unexpected viral sequences.** Proceedings of the 36th International Conference on Animal Genetics (ISAG); pp. 39-40 & pp. 142-142 - Dublin, Ireland - July 16-21, 2017

4. **Bovo S**, Mazzoni G, Schiavo G, Bertolini F, Galimberti G, Samorè AB, Dall'Olio S, Fontanesi L. (2017) **Genome wide association studies for haematological traits in Italian Large White pigs.** Proceedings of the 36th International Conference on Animal Genetics (ISAG); pp. 136-136 - Dublin, Ireland - July 16-21, 2017

5. Bertolini F, **Bovo S**, Rothschild MF, Fontanesi L. (2017) **Mining the European Sea Bass (Dicentrarchus labrax) genome for the characterization of tandem repeat variability.** Proceedings of the 36th International Conference on Animal Genetics (ISAG); pp. 51-52 & pp. 192-192 - Dublin, Ireland - July 16-21, 2017

6. Schiavo G, Strillacci MG, **Bovo S**, Ribani A, Roman-Ponce SI, Cerolini S, Bertolini F, Bagnato A, Fontanesi L. (2017) **Low number of mitochondrial DNA sequences inserted into the turkey (Meleagris gallopavo) nuclear genome: implications for evolutionary inferences.** Proceedings of the 36th International Conference on Animal Genetics (ISAG); pp. 13-13 & pp. 73-73 - Dublin, Ireland - July 16-21, 2017

7. Babbi G, Martelli PL, Profiti G, **Bovo S**, Savojardo C, Casadio R. (2017) **Analysing the relations among genes and polygenic diseases with eDGAR.** 14th Annual Meeting of the Bioinformatics Italian Society - July 5-7, 2017, Cagliari, Italy

8. **Bovo S**, Mazzoni G, Fanelli F, Mezzullo M, Schiavo, Galimberti G, Bertolini F, Dall'Olio F, Pagotto U, Fontanesi L. (2017) **Comparative analysis of metabolomics profiles of plasma and serum in pigs.** Proceedings of the ASPA 22th Congress, June 13-16 2017, Perugia (Italy). Italian Journal of Animal Science, Vol. 16 Supp. 1, 155

9. **Bovo S**, Mazzoni G, Schiavo G, Bertolini F, Galimberti G, Samorè AB, Dall'Olio S, Fontanesi L. (2017) **Genome wide association studies for haematological and clinical-biochemical parameters in Italian Large White pigs.** Proceedings of the ASPA 22th Congress, June 13-16 2017, Perugia (Italy). Italian Journal of Animal Science, Vol. 16 Supplement 1, 118

10. Babbi G, Martelli PL, Profiti G, **Bovo S**, Savojardo C, Casadio R. (2017) **eDGAR: a webserver for analysing the relationship among genes and polygenic diseases.** Conference Book "NGS 2017: structural variation and population genomics", pp. 46 – 47 - Barcellona, Spain, April 3-5, 2017

11. Utzeri VJ, Ribani A, Schiavo G, Bertolini F, Geraci C, **Bovo S**, Fontanesi L (2017) **Food metagenomics against frauds: applications of next generation semiconductor based sequencing on meat and dairy products and honey.** FoodInnova 2017 - Cesena, Italy, 31st January - 3rd February, 2017

12. **Bovo S**, Martelli PL, Fariselli P, Di Lena P, Profiti G, Savojardo C, Aggazio F, Babbi G and Casadio R. (2016) **AGROFOOD BIOINFORMATICS at the Bologna Biocomputing Group.** Poster presented at the "ITALY - CHINA, Science, technology & Innovation Week". October 25-27, 2016, Bologna, Italy

13. Fontanesi L; Schiavo G; Bertolini F; Galimberti G; **Bovo S**; Gallo M; Russo V; Luca Buttazzoni (2016) **Genome wide association and candidate gene analyses for dry-cured ham quality traits: towards the dissection of important parameters in heavy pig selection programs.** Proceeding of the 9th International Symposium on Mediterranean Pig – Portalegre (Portugal)

14. Schiavo G, Galimberti G, **Bovo S**, Bertolini F, Gallo M, Russo V, Buttazzoni L, Fontanesi L. (2016) **Genome wide association studies for dry-cured ham quality traits in two Italian heavy pig breeds**. Proceeding of the 67th Annual Meeting of the European Federation of Animal Science - Belfast, UK.

15. Fontanesi L, **Bovo S**, Schiavo G, Mazzoni G, Ribani A, Utzeri VJ, Dall'Olio S, Bertolini F, Fanelli F, Mezzullo M, Glimberti G, Calò DG, Martelli PL, Casadio R, Trevisi P, Bosi P, Pagotto U. (2016) **The added value of molecular phenotypes: towards the identification of animal welfare proxies.** Proceeding of the 67th Annual Meeting of the European Federation of Animal Science - Belfast, UK.

16. Ribani A, Utzeri VJ, Schiavo G, **Bovo S**, Geraci C, Fontanesi L. (2016) **Food genomics: application of innovative DNA analysis technologies for authentication of food products.** TRADEIT Entrepreneurship Summer Academy – Postdam (Germany) 6-10 June 2016.

17. Fontanesi L, Schiavo G, Galimberti G, **Bovo S**, Bertolini F, Gallo M, Russo V, Buttazzoni L. (2016) **Genome wide association studies for dry-cured ham quality traits in Italian Large White and Italian Duroc pigs.** Proceedings of the 35th International Conference on Animal Genetics. - J. Anim. Sci. 94(Suppl4):124.

18. Fontanesi L, **Bovo S**, Schiavo G, Mazzoni G, Ribani A, Utzeri VJ, Dall'Olio S, Bertolini F, Fanelli F, Mazzullo M, Galimberti G, Calò DG, Trevisi P, Pagotto U, Bosi P. (2016) **Deconstructing the pig genome-metabolome functional interactions.** Proceedings of the 35th International Conference on Animal Genetics. - J. Anim. Sci. 94(Suppl4):57-58.

19. Schiavo G, Hoffmann OI, Ribani A, Utzeri VJ, Ghionda MC, **Bovo S**, Fontanesi L. (2016) **A genomic landscape of mitochondrial DNA insertions in the nuclear pig genome.** Proceedings of the 35th International Conference on Animal Genetics - J. Anim. Sci. 94(Suppl4):181-182.

20. **Bovo S**, Di Lena P; Martelli PL, Fariselli P, Casadio R. (2016) **From protein variations to biological processes and pathways with NET-GE**. Proceedings of IWWBIO 2016 – 4th International Work-Conference on Bioinformatics and Biomedical Engineering – Granada, Spain, 20-22 April 2016

21. **Bovo S**, Di Lena P; Martelli PL, Fariselli P, Casadio R. (2016) **NET-GE: a web-server for linking protein variations to biological processes and pathways.** Conference Book "NGS 2016: Genome Annotation", pp. 55 – 56 - Barcelona, Spain, April 4-6, 2016

22. **Bovo S**, Babbi G, Martelli PL, Casadio R. (2016) **Liver pyruvate kinase (L-PYK): predict the effects of missense mutations on kinase activity and allosteric regulation.** CAGI4 - Book of abstract, 25-27 March 2016, San Francisco, CA, USA.

23. Babbi G, Martelli PL, **Bovo S**, Casadio R. (2016) **Predicting the effect of variations on protein functional activity and interaction.** CAGI4 - Book of abstract, 25-27 March 2016, San Francisco, CA, USA.

24. **Bovo S**, Martelli PL, Babbi G, Casadio R. (2016) **Discriminating sick from healthy individuals by annotating exome variations.** CAGI4 - Book of abstract, 25-27 March 2016, San Francisco, CA, USA.

25. Babbi G, Martelli PL, **Bovo S**, Casadio R. (2016) **Predicting the effect of variations on NPM-ALK functional activity and binding affinity with Hsp90.** CAGI4 - Book of abstract, 25-27 March 2016, San Francisco, CA, USA.

26. Fontanesi L, Schiavo G, **Bovo S**, Mazzoni G, Fanelli F, Ribani A, Utzeri VJ, Luise D, Samoré AB, Galimberti G, Calò DG, Manisi A, BertolinI F, Mezzullo M, Pagotto U, Dall'Olio S, Trevisi P, Bosi P. (2015) **Dissecting complex traits in pigs: metabotypes illuminate**

genomics for practical applications. Book of Abstracts of the 66th Annual Meeting of the European Federation of Animal Science

27. **Bovo S**, Di Lena P, Martelli PL, Fariselli P and Casadio R. (2015) **From protein variants to biological processes and pathways with NET-GE.** Book of Abstracts of the 58th National Meeting ofthe Italian Society of Biochemistry and Molecular Biology, Urbino (Italy).

28. **Bovo S**, Mazzoni G, Calò DG, Galimberti G, Fanelli F, Mezzullo M, Schiavo G, Scotti E, Manisi A, Samorè AB, Bertolini F, Trevisi P, Bosi P, Dall'Olio S, Pagotto U and Fontanesi L. (2015) **Deconstructing the pig sexome: metabolomics in heavy pigs discovers sex related dimorphisms in metabolic biomarkers and pathways.** Proceedings of the ISAFG 6th Congress, Piacenza (Italy).

29. Luise D, Schiavo G, Galimberti G, Calò DG, **Bovo S**, Mazzoni G, Fanelli F, Scotti E, Mezzullo M, Dall'Olio S, Pagotto U, Bosi P, Fontanesi L, Trevisi P. (2015) **Metabolomics information provides potential predictive internal phenotypes for production and performance traits in Italian Duroc pigs.** Proceedings of the ASPA 21th Congress, Milan (Italy). Italian Journal of Animal Science, Vol. 14 Supp. 1, 25.

30. Fontanesi L, **Bovo S**, Mazzoni G, Schiavo G, Samorè AB, Fanelli F, Bertolini F, Scotti E, Galimberti G, Calò DG, Mezzullo M, Martelli PL, Casadio R, Dall'Olio S, Pagotto U. (2015) **Genomics Meets Metabolomics: Developing a Systems Biology Approach to Understand the Genetic Mechanisms Affecting Complex Traits in Pigs.** Book of abstracts - PAG XXIII, January 10-14, 2015, San Diego, California, USA

16 Appendix

In the following pages are included the printed versions of these papers:

1. **Bovo S**, Di Lena P, Martelli PL, Fariselli P and Casadio R. (2016) **NET-GE: a web-server for NETwork-based Gene Enrichment of sets of human genes related to the same phenotype.** *Bioinformatics*, 32(22):3489-3491 PMID:27485441
2. **Bovo S**, Di Lena P; Martelli PL, Fariselli P, Casadio R. (2017) **From Protein Variations to Biological Processes and Pathways with NET-GE.** *Genomics and Computational Biology*, 3(3); e45;
3. Babbi G, Martelli PL, Profiti G, **Bovo S**, Savojardo C and Casadio R. (2017) **eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes.** *BMC Genomics*, 18(Suppl5):554. PMID:28812536.
4. Xu Q, Tang Q, Katsonis, Lichtarge O, Jones DT, **Bovo S**, Babbi G, Martelli PL, Casadio R, Lee GR, Seok C, Fenton AW and Dunbrack RL. (2017) **Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4.** *Human Mutation*, 38(9):1123-1131. PMID:28370845.
5. Daneshjou R, Wang Y, Bromberg Y, **Bovo S**, Martelli PL, Babbi G, Di Lena P, Casadio R, Edwards M, Gifford D, Jones DT, Sundaram L, Bhat R, Li X, Pal L R, Kundu K, Yin Y, Moult J, Jiang Y, Pejaver V, Pagel KA, Li B, Mooney SD, Radivojac P, Shah S, Carraro M, Gasparini A, Leonardi E, Giollo M, Ferrari C, Tosatto SCE, Bachar E, Azaria JR, Ofran Y, Unger R, Niroula A, Vihinen M, Chang B, Wang MH, Franke A, Petersen BS, Pirooznia M, Zandi P, McCombie R, Potash JB, Altman R, Klein TE, Hoskins R, Repo S, Brenner SE, Morgan AA. (2017) **Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges.** *Hum Mutation*, 38(9):1182-1192. PMID 28634997.

Bioinformatics, 32(22), 2016, 3489–3491
doi: 10.1093/bioinformatics/btw508
Advance Access Publication Date: 2 August 2016
Applications Note

OXFORD

Genome analysis

NET-GE: a web-server for NETwork-based human gene enrichment

Samuele Bovo[1,†], Pietro Di Lena[2,†], Pier Luigi Martelli[1,*], Piero Fariselli[3] and Rita Casadio[1,4]

[1]Biocomputing Group, CIG, Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna, Italy, [2]DISI, University of Bologna, Bologna, Italy, [3]BCA, University of Padova, Padova, Italy and [4]Interdepartmental Center «Giorgio Prodi» for Cancer Research, University of Bologna, Bologna, Italy

*To whom correspondence should be addressed.
[†]These authors contributed equally to this work.
Associate Editor: John Hancock

Received on June 8, 2016; revised on July 18, 2016; accepted on July 20, 2016

Abstract

Motivation: Gene enrichment is a requisite for the interpretation of biological complexity related to specific molecular pathways and biological processes. Furthermore, when interpreting NGS data and human variations, including those related to pathologies, gene enrichment allows the inclusion of other genes that in the human interactome space may also play important key roles in the emergency of the phenotype. Here, we describe NET-GE, a web server for associating biological processes and pathways to sets of human proteins involved in the same phenotype
Results: NET-GE is based on protein–protein interaction networks, following the notion that for a set of proteins, the context of their specific interactions can better define their function and the processes they can be related to in the biological complexity of the cell. Our method is suited to extract statistically validated enriched terms from Gene Ontology, KEGG and REACTOME annotation databases. Furthermore, NET-GE is effective even when the number of input proteins is small.
Availability and Implementation: NET-GE web server is publicly available and accessible at http://net-ge.biocomp.unibo.it/enrich.
Contact: gigi@biocomp.unibo.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Big Data production in biomedicine is rapidly changing the way in which molecular knowledge is translated into health care (Bender, 2015). The spread and establishment of High Throughput Sequencing (HTS) technologies allows retrieving lists of interesting variations characterizing the investigated phenotype. In the context of functional genomics, each phenotype needs annotations for reconciling variations with known and putatively common biological processes and pathways, such as Gene Ontology (GO Consortium, 2015), KEGG (Kanehisa *et al.*, 2016), REACTOME (Fabregat *et al.*, 2016). At this level of biological complexity, a set of genes and their variations can acquire biological meaning and feature annotation only with an enrichment procedure (Laukens *et al.*, 2015). Enrichment helps in identifying within a set of genes some statistically significant and over-represented annotation features. Standard enrichment methods rely on the statistical over representation of the annotations that characterize the genes in the input set. Alternatively, network-based approaches extract graph properties from different interaction networks and pathways for modelling the complexity of the processes occurring in the cell and exploit this information for accomplishing the annotation enrichment in the context of protein functional interaction. Lists of web sites are available (Huang *et al.*, 2009; Laukens *et al.*, 2015; Mooney and Wilmot, 2015). Here, we introduce NET-GE, a web server that implements our method (Di Lena *et al.*,

2015), based on the extraction of subnetworks connecting proteins that share the same functional terms from a protein–protein interaction (PPI) network (Szklarczyk *et al.*, 2015). Differently from other methods also based on networks, our approach extracts modules that are function-specific by constructions and include all the seeds (proteins annotated with the same term) that in the PPI network are related to a specific functional annotation. One peculiarity of NET-GE is the possibility to enrich terms that are not present in the annotation of the starting protein set (and thus not detectable through a standard enrichment). When tested on the OMIM-derived benchmark sets, NET-GE is able to enrich sets of genes related to the same disease with biologically meaningful terms neglected by other methods (Di Lena *et al.*, 2015). The server, in addition, allows annotation based on KEGG and REACTOME pathways and a comparison between standard and network based enrichment.

2 NET-GE

NET-GE includes precomputed subsets of proteins associated to each functional terms of interest (Di Lena *et al.*, 2015). Subnetwork construction is based on the human interactome map downloaded from STRING (release 10.0, http://string-db.org/), or from a filtered version that retains only links with a score ≥ 0.9. Presently STRING includes 15,632 nodes (mapping 18 721 HGNC gene names, http://www.genenames.org/ and 89 085 UniProtKB identifiers) and 307 413 links (in the high quality STRING 0.9 version nodes and links are 9422 and 80 112, respectively).

In the present implementation of NET-GE, annotations are however available for all the 104 569 UniProtKB identifiers (release 2016_01), corresponding to 22 390 genes. The databases for annotating features are GENE ONTOLOGY (from UniProt-GOA human 145 resource, http://www.ebi.ac.uk/GOA); KEGG PATHWAY (release 77.0, http://www.genome.jp/kegg/pathway.html); REACTOME PATHWAY (release 53, http://www.reactome.org/). Redundancy among terms is not taken into consideration.

When generating the annotating subnetworks, for each annotation term we collect the seeds and evaluate the quality of the connecting nodes among seeds (for more details, see Supplementary Fig. 1S). After constraining seed distance, we determine the subset associated to a specific annotation term by retaining the minimal connecting subnetwork (Di Lena *et al.*, 2015). Considering STRING, NET-GE presently includes 20 391 annotation subsets (see http://net-ge.biocomp.unibo.it/enrich/statistics), 14 845 of which contain from two to 10 700 genes. The number of genes per subset is inversely proportional to the information content and the most informative terms correspond to small networks (Fig. 1).

The server implements both a standard and a network-based gene enrichment. Given a gene/protein list, each gene/protein is located in the different subsets of the annotation database. With a Fisher's exact test, the method estimates the overrepresentation significance of input genes/proteins in each precomputed subset for the corresponding annotation term. Standard enrichment includes only annotations of seed nodes; network-based enrichment includes seeds and their connecting nodes. For multiple testing correction, we use both the Bonferroni and the Benjamini-Hochberg (False Discovery Rate, FDR) procedures and evaluate a corrected p-value (Noble, 2009). Updating of the system, including human interactome and annotation databases is planned once a year, following the major releases.

2.1 Web server

NET-GE Web interface accepts UniProtKB Accession Numbers, Ensembl and HGNC gene names. The end user can select: (i) annotation modules based on STRING or STRING 0.9; (ii) the annotation (GO terms, KEGG, REACTOME); (iii) the multiple testing correction methods (Bonferroni or the Benjamini-Hochberg correction); (iv) the significance threshold.

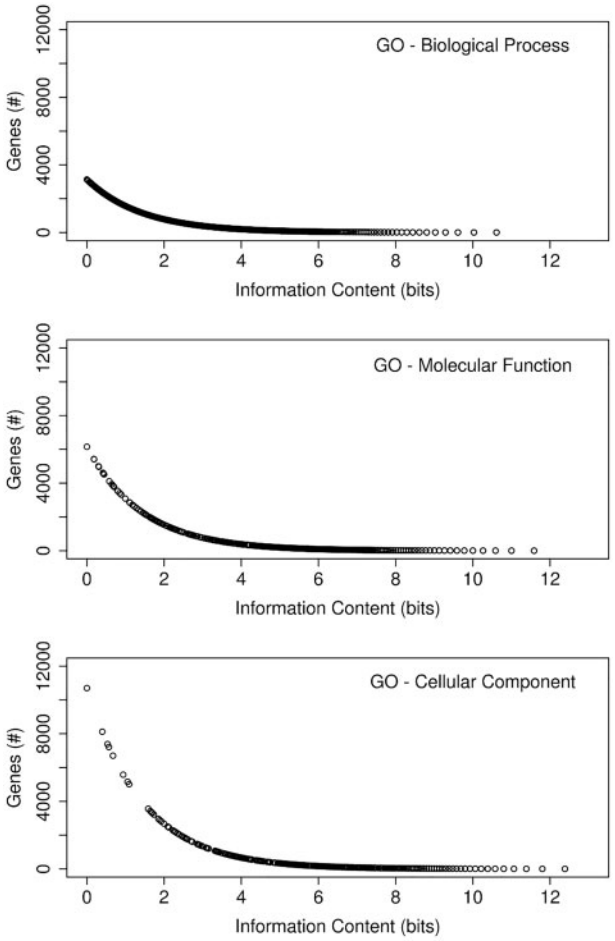


Fig.1. Dimension of subsets (number of genes) as a function of the information content for the Gene Ontology terms of the three main roots. The information content (in bits) is computed adopting standard methods (Shannon, 1948)

The output lists two enrichment tables: one for the standard and one for the network-based method (see the online tutorial for more details). Each table contains the annotation term identifier, linked to the corresponding database; the number and the list of input genes/proteins associated to the term; the *P*-value of the association; the description of the term and for the network based enrichment a visualization of the subnetwork. Enriched terms not included in the annotations of the input gene/protein are highlighted with a double star (see on line tutorial). It is also possible to access the complete set of annotations (for both the enrichment modes) of the submitted genes/proteins through the link provided at the bottom of the page. The front-end for the Web server follows the Model-View-Controller (MVC) paradigm, thanks to the web2py framework (http://www.web2py.com/), and it is optimized to work with all common web browsers. The analysis runs asynchronously: after submitting the query, the server displays a bookmarkable page reporting the status of the job. This page is periodically updated. A link to the results, accessible as soon as the job is completed, is given to the user. The final visualization of the results exploits the Graphviz library (http://www.graphviz.org/) and the JavaScript

library d3.js (http://d3js.org/). The user can also provide an e-mail address used to alert her/him as soon as results are ready. Running time depends on size of the input set (from two up to 200 genes) and ranges about 1–5 min.

Funding

RC thanks COST Action BM1405 (European Union RTD Framework Program) and FARB UNIBO 2012.

Conflict of Interest: none declared.

References

Di Lena,P. *et al*. (2015) NET-GE: a novel NETwork-based gene enrichment for detecting biological processes associated to Mendelian diseases. *BMC Genomics*, **16**, S6.

Fabregat,A. *et al*. (2016) The reactome pathway knowledgebase. *Nucleic Acids Res*., **44**, D481–D487.

Szklarczyk,D. *et al*. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., **44**, D447–D452.

Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res*., **43**, D1049–D1056.,

Huang,D.W. *et al*. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*., **37**, 1–13.

Kanehisa,M. *et al*. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*., **44**, D457–D462.

Laukens,K. *et al*. (2015) Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis. *Proteomics*, **15**, 981–996.

Mooney,M.A. and Wilmot,B. (2015) Gene set analysis: A step-by-step guide. *Am. J. Med. Genet. B Neuropsychiatr. Genet*., **168**, 517–527.

Noble,W.S. (2009) How does multiple testing correction work? *Nat. Biotechnol*., **27**, 1135–1137.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Techn. J*., **27**, 379–423.

GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 3 No. 3 (2017): e45

Research Article

From Protein Variations to Biological Processes and Pathways with NET-GE

Samuele Bovo[1], Pietro Di Lena[2], Pier L. Martelli[1,*], Piero Fariselli[3], Rita Casadio[1,4]

[1]Biocomputing Group, CIG, Interdepartmental Center "Luigi Galvani" for Integrated Studies of Bioinformatics, Bio-physics and Biocomplexity, University of Bologna, Bologna, Italy
[2]DISI, University of Bologna, Bologna, Italy
[3]BCA, University of Padova, Padova, Italy
[4]Interdepartmental Center "Giorgio Prodi" for Cancer Research, University of Bologna, Bologna, Italy

*Correspondence: Email: gigi@biocomp.unibo.it

Received 2016-10-07; Accepted 2017-04-19

ABSTRACT

Gene enrichment analysis is a common technique for highlighting molecular pathways and biological processes of a phenotype. Such technique has recently evolved exploiting the information contained in biological networks. We developed NET-GE, a web server for network-based gene enrichment analyses. NET-GE defines functional associations between a list of genes/proteins and biological processes or pathways by identifying function-specific modules in a molecular interaction network. The peculiarity of NET-GE is the possibility to enrich terms not detectable by standard enrichment procedure. Here, we highlight with two specific applications the performances of NET-GE by computing which functional phenotypes can be associated with two different sets of genes related to Attention Deficit Hyperactivity Disorder and to an Obsessive-compulsive disorder, respectively.

KEYWORDS

Gene enrichment analysis; network-based gene enrichment analysis; functional association

INTRODUCTION

Technologies capable of investigating the organism complexity at different levels of resolution have been revolutionizing healthcare practice [1]. Genomic data are generated more and more to better define, at molecular levels, the origin of the different phenotypes. From a precision/genomic medicine prospective, such phenotypes need annotations in order to reconcile specific variations with common biological processes and pathways, such as GENE ONTOLOGY [2], KEGG [3] and REACTOME [4] pathways. For this purpose, functional association is routinely performed by applying gene enrichment analysis, a technique that assesses the statistically over-represented biological processes and pathways of a given gene/protein set [5].

Presently, enrichment analysis methods mainly group into two classes, standard and network-based. While standard methods rely only on the annotations characterizing the genes/proteins included in the input set, network-based methods consider them in the context of their interaction network. Thus, such methods exploit information derived from functional biological networks, modelling the complexity of the processes occurring in the cell, and implement algorithms that exploit graph properties (such as shortest paths and node degrees).

In the last year, several approaches exploiting the interaction networks for functional association analysis have emerged (see [6–8] for a comprehensive list of available tools). They may be classified into two main categories: A) methods that exploit the topology of the network to infer how similar are sets of genes/proteins, and B) methods that identify functionally related modules, inferring biological features from them. Among the available tools that perform network-based enrichment analysis, EnrichNet [9] and PINA v2.0 [10] are two of the most cited methods, representative of the A and B categories, respectively.

We recently developed NET-GE, a network-based gene enrichment analysis tool [11, 12]. NET-GE falls within the class B and it is based on a pre-processing phase aimed at identifying interconnected and compact modules in a molecular interaction network. However, differently from all the other approaches in class B, the modules found by our method are function-specific by construction, since they are built starting from seed sets collecting all the proteins related to a specific biological annotation

One of the main features of NET-GE is the possibility to enrich terms that are not originally present in the annotation of the starting gene/protein set (and thus not detectable through a standard enrichment). When tested on benchmark sets retrieved from the Online Mendelian Inheritance in Man (OMIM) resource (https://www.omim.org), NET-GE was able to enrich sets of genes related to the same disease, also highlighting new terms (i.e. terms not included in the annotations of the input set) [11].

Here, we present two study cases, demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder (ADHD) and the Obsessive-compulsive disorder (OCD).

METHODS

NET-GE background

The network-based enrichment makes use of precomputed annotation terms, as previously described [11]. Briefly, the human molecular-interaction network was downloaded from STRING v.10 (http://string-db.org). A second version of STRING, named STRING0.9, was obtained by retaining only the links with the STRING combined score ≥ 0.9. The database for annotating features were: GENE ONTOLOGY (as retrieved from the UniProt-GOA human 145 web resource: http://www.ebi.ac.uk/GOA); KEGG PATHWAY v77 and REACTOME PATHWAY v53. For each annotating feature, proteins sharing the same annotation term were collected in a seed set and then extended into a compact and connected module of the molecular-interaction network. Thus, the module was determined by computing all the shortest paths among the seeds genes/proteins and then by reducing the resulting sub-network into the minimal connecting network that preserves the distances among seeds. The minimal connecting network adds to the seeds a set of connecting nodes that are more reliably related to the reference annotation. Details about annotations and module extraction can be found in [11] and [12], respectively.

Over-representation analysis is performed by mapping the input set on each module and determining, through a Fisher's exact test, whether there are significant overlaps between the input set and the modules (seed sets in the case of standard enrichment). Multiple testing correction is then applied using the Bonferroni or the Benjamini-Hochberg (False Discovery Rate, FDR) procedure [13].

When we consider the standard enrichment, the background set is totally disconnected. On the contrary, with the network-based procedure we rely on the human interactome to precompute the annotation modules. Enrichment is computed over a changed reference set that includes also all the nodes connecting seeds with the same annotation. This may change the p-value.

NET-GE web server

A web server, implementing both a standard and a network-based gene enrichment was implemented as described in [12]. Briefly, NET-GE Web interface takes as input a list of genes/proteins (allowed identifiers are: UniProtKB AC, Ensembl and HGNC gene names). The enrichment can be performed considering the annotation modules based on STRING or STRING0.9. The enriched terms can derive from the GENE ONOLOGY (all the sub-ontologies), or from the KEGG or the REACTOME PATHWAYS. The user can select between two kinds of multiple testing correction methods (Bonferroni or the Benjamini-Hochberg correction), and the significance threshold. As output NET-GE reports: 1) two enrichment tables (one for the standard enrichment and one for network-based one), 2) a graph visualizing how the enriched terms are linked, and 3) the complete set of annotations (for both the enrichment modes). Terms not included in the annotations of the input proteins are highlighted with a double star.

RESULTS

To test the performance of NET-GE we used sets of proteins involved in Mendelian diseases [11]. We tested 244 different genetic disorders, each one associated to two or more proteins. Our method was able to detect functional associations not detectable by the standard enrichment. Moreover, the newly enriched terms that were absent in the original annotations of the input genes are likely to provide new knowledge on the phenotype under examination [11].

Here, we present two cases of study demonstrating how NET-GE can help the interpretation and prioritization of variations in sets of genes associated with two complex disorders: the Attention Deficit Hyperactivity Disorder and the Obsessive-compulsive disorder.

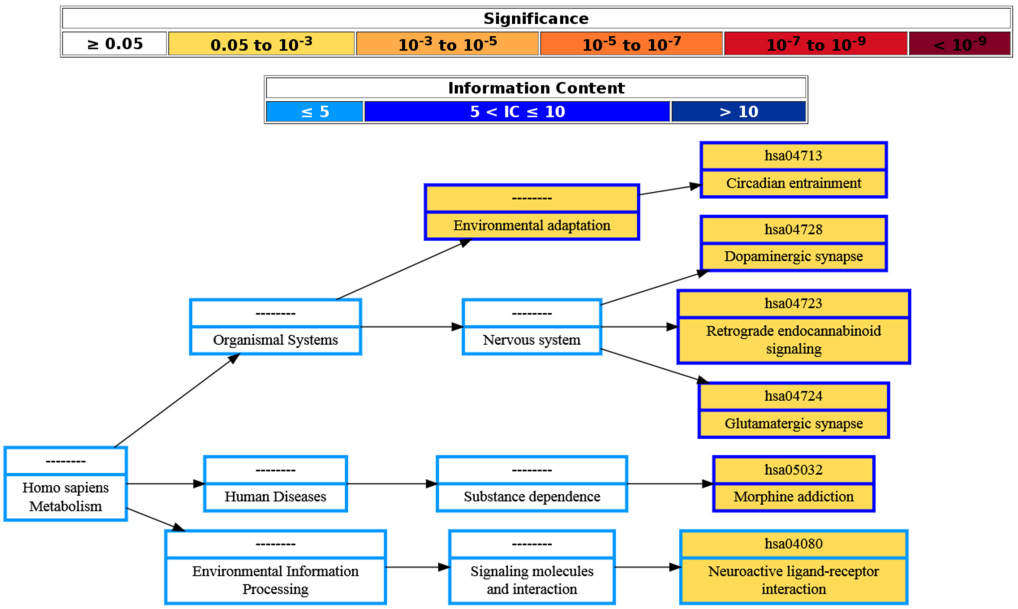
Attention Deficit Hyperactivity Disorder

In the following, we deal with a specific test set (http://net-ge.biocomp.unibo.it/enrich/tutorial) that includes two input proteins related to Attention Deficit Hyperactivity Disorder (ADHD; OMIM #143465), a neurodevelopmental disease of childhood affecting the cognitive and behavioral functions. The genetic disease is associated to variations in the dopamine receptors DRD4 (UniProtKB AC: P21917) and DRD5 (UniProtKB AC: P21918). Using as input the *DRD4* and *DRD5* genes, we carried out enrichment analyses by setting the significance threshold at 0.05 on the Bonferroni corrected p-values. Standard and network-based enrichments ran over the KEGG database. Terms enriched by NET-GE are shown in Figure 1. The standard enrichment on KEGG highlights neuroactive ligand-receptor interaction and dopaminergic synapse as the most significant pathways. The network-based procedure adds new terms, not associated to the input proteins, and involved in ADHD, considering the statistically significant subnetworks. The pathways sorted by significance are: circadian entrainment, morphine addiction, retrograde endocannabinoid signaling and glutamatergic synapse.

Interestingly enough, the enriched pathways had been previously described in literature as being diseases-related. Different experiments have described different pathways [14–17] and the network-based enrichment method retrieved them all from the inclusion of the connecting nodes in the annotation modules.

In Figure 1 the difference in annotation between the standard enrichment procedure and the network-based is shown. As explained in the Methods section, standard enrichment is computed over a totally disconnected reference set. The network-based procedure relies on the precomputed annotation modules and the reference set includes all the nodes that connect seeds with the same annotation. This may increase the p-value as in

Enriched Terms - Directed Acyclic Graph



Standard enrichment

[+/-] Show/Hide all results.

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
S	hsa04728	2 [+] Show genes	135 Show genes Show protein info	3.722e-03	Dopaminergic synapse
S	hsa04080	2 [+] Show genes	291 Show genes Show protein info	1.736e-02	Neuroactive ligand-receptor interaction

Network-based enrichment

N** highlights enriched terms not included in the annotations of the input set.

[+/-] Show/Hide all results.

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description	
N**	hsa04713	2 [+] Show genes	192 Show genes Show protein info	1.230e-02	Circadian entrainment	graph visualization
N**	hsa05032	2 [+] Show genes	202 Show genes Show protein info	1.362e-02	Morphine addiction	graph visualization
N**	hsa04723	2 [+] Show genes	220 Show genes Show protein info	1.616e-02	Retrograde endocannabinoid signaling	graph visualization
N**	hsa04724	2 [+] Show genes	239 Show genes Show protein info	1.908e-02	Glutamatergic synapse	graph visualization
N**	--------	2 [+] Show genes	271 Show genes Show protein info	2.454e-02	Environmental adaptation	graph visualization
N	hsa04728	2 [+] Show genes	296 Show genes Show protein info	2.928e-02	Dopaminergic synapse	graph visualization

Enriched terms and link to the database entry

Mapped input genes/proteins

Annotated genes

Possibility to visualize the term-specific network

Legend:
→ New terms added by NET-GE and not directly associated to the input proteins/genes

Figure 1: Output of NET-GE for the enrichment of KEGG pathways in the ADHD study case. Enrichment analysis was carried out using as input the *DRD4* and *DRD5* genes. The upper panel shows the graph of the enriched terms and their relations. Box filling color represents the corrected p-value associated to the enriched term, while contour color represents its information content (see [11] and [12] for details). The lower panel presents the enriched terms in a tabular format. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.

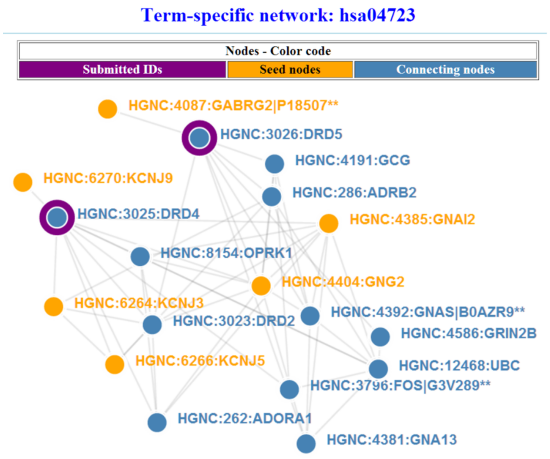


Figure 2: Graph of the first protein neighbours in the ADHD study case. The graph of the KEGG term hsa04723 (Retrograde endocannabinoid signaling) shows the two input proteins (in purple) and the first protein neighbours highlighted as seeds (in yellow) and new connecting genes (in blue). The connecting genes are added to the graph with the network-based enrichment procedure.

the case of the neuroactive ligand receptor interaction that is no longer listed among the terms obtained with the network-based procedure.

For comparison, we also tried PINA and EnrichNET. Considering as significant p-values<0.05 Benjamini-Hochberg corrected, PINA (tool "Identify enriched Interactome modules") did not retrieve any significantly over-represented module. EnrichNET authors recommend to analyse sets with at least 10 genes/proteins for reasons of statistical reliability. As a consequence, EnrichNET did not retrieve any significant term.

As evaluate the robustness of the method for small input sets composed of two to ten proteins, we computed the effect on the final stability of the enrichment when doubling (with random additions) the sizes of the input sets. We obtain that under these extreme conditions of noise, the stability of the enrichment ranges from 37 to 52%, depending on the annotation term and the network type (see Figure S1).

In Figure 2, the two input proteins are shown in the graph (purple circles) of the first protein neighbors, after network-based enrichment, detailing protein seeds of the Retrograde endocannabinoid signaling KEGG path (hsa04723, *Homo sapiens*) in yellow and the connecting nodes in blues (proteins that are retained after NET-GE based enrichment). The whole annotation network is downloadable (all seeds, nodes and arcs) and it is available for display.

Obsessive-compulsive Disorder

Obsessive-compulsive disorder (OCD) is a severe neuropsychiatric disorder characterized by the presence of obsessions and compulsions [18]. This disorder has been recently investigated in [18] by using whole-exome sequencing (WES).

Twenty OCD cases and their unaffected parents (parent–child trios) were screened for *de novo* missense mutations (i.e. mutation present only in the affected individual), identifying 27 OCD-related genes. Based on Ingenuity software (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis), three signaling pathways were identified as disease-related [18] sharing only one patient/one gene. In fact, among the 27 genes, only SMAD4 (gene mutated in only one patient) was present in the three enriched pathways.

With NET-GE we highlight four biological processes as the most significant ones (Figure 3, panel A), all related to the purine metabolism that has been proven to be associated with several neurological disorders [19–21]. However, and interestingly enough, 10 of the 27 initial genes have common annotations. Testing Molecular Functions, the standard enrichment procedure highlighted ATPase activity and the network-based procedure enriched thyroxine 5'-deiodinase activity (Figure 3, panel B), a new term not associated to the input proteins and involved in OCD [22].

Our results highlight the involvement of processes common to the gene panel and corroborates the notion that network-based enrichment consistently derives information from the connected annotation modules, including genes corresponding to 9 of the 14 patients analyzed in [18].

CONCLUSION

In this article, we presented the NET-GE web server [12], developed for tackling the problem of the human biological complexity. Specifically, NET-GE is a tool for associating biological processes and pathways with sets of human genes/proteins involved in the same phenotype. It performs standard and network-based enrichment analysis. The network-based procedure extracts from the STRING human interactome sub-networks of connecting proteins that share the same annotation [11]. We benchmarked NET-GE on two specific test cases, with a phenotype and its biological functions already described in literature. On this benchmark, the network-based procedure, considering genes/proteins in the context of their functional interaction network, enriched functional annotations that are experimentally validated. This version of NET-GE is preliminary to the inclusion of some additional features that can eventually add to the relevance of detecting emerging functional characteristics from a set of genes, such as the inclusion of ranking scores (e.g. fold of differentially expressed genes) or the usage of tissue-specific interactomes.

ACKNOWLEDGEMENTS

RC thanks COST Action BM1405 (European Union RTD Framework Program) and FARB UNIBO.

A) Biological Process

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
N	GO:0009203	10 [+] Show genes	1506 Show genes Show protein info	1.505e-02	ribonucleoside triphosphate catabolic process
N	GO:0009207	10 [+] Show genes	1506 Show genes Show protein info	1.505e-02	purine ribonucleoside triphosphate catabolic process
N	GO:0009146	10 [+] Show genes	1510 Show genes Show protein info	1.541e-02	purine nucleoside triphosphate catabolic process
N	GO:0009143	10 [+] Show genes	1518 Show genes Show protein info	1.616e-02	nucleoside triphosphate catabolic process

B) Molecular Function

Enrichment	TERM	N1	N2	corrected p-value (Bonferroni)	Description
N**	GO:0004800	2 [+] Show genes	8 Show genes Show protein info	8.505e-03	thyroxine 5'-deiodinase activity
N	GO:0017111	9 [+] Show genes	1786 Show genes Show protein info	1.146e-02	nucleoside-triphosphatase activity
N	GO:0016462	9 [+] Show genes	1919 Show genes Show protein info	2.033e-02	pyrophosphatase activity
N	GO:0016818	9 [+] Show genes	1924 Show genes Show protein info	2.075e-02	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
N	GO:0016817	9 [+] Show genes	1931 Show genes Show protein info	2.136e-02	hydrolase activity, acting on acid anhydrides

Figure 3: Output of NET-GE for Obsessive Compulsive Disorder for Biological Processes (panel A) and Molecular Function (panel B). Genes are derived from [18]. Terms highlighted with a double star are new annotations, not associated to the input proteins and enriched with the network-based procedure. p-values are corrected with the Bonferroni procedure.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files, together with supplementary items listed below, are available at Genomics and Computational Biology online.

Supplementary Figure S1. **Testing the robustness of the network-based enrichment methods**. For small input sets comprising from two to ten proteins (derived from OMIM), we computed the effect of doubling (with random additions) the size on the final stability of the enrichment. This was done for all the annotation terms and the two different version of STRING (see Methods). Errors bars indicated standard deviations over a reference of 123 gene sets.

REFERENCES

1. Sander C. **Genomic Medicine and the Future of Health Care**. Science. 2000;287(5460):1977–1978. doi:10.1126/science.287.5460.1977.
2. The Gene Ontology Consortium. **Gene Ontology Consortium: going forward**. Nucleic Acids Research. 2015;43(D1):D1049–D1056. doi:10.1093/nar/gku1179.
3. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. **KEGG as a reference resource for gene and protein annotation.** Nucleic Acids Research. 2016;44(D1):D457–462. doi:10.1093/nar/gkv1070.
4. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. **The Reactome pathway Knowledgebase.** Nucleic Acids Research. 2016;44(D1):D481–487. doi:10.1093/nar/gkv1351.
5. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. **Impact of outdated gene annotations on pathway enrichment analysis.** Nature Methods. 2016;13(9):705–706. doi:10.1038/nmeth.3963.
6. Laukens K, Naulaerts S, Berghe WV. **Bioinformatics approaches for the functional interpretation of protein lists: from ontology term enrichment to network analysis.** Proteomics. 2015;15(5-6):981–996. doi:10.1002/pmic.201400296.
7. Mooney MA, Wilmot B. **Gene set analysis: A step-by-step guide.** American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics. 2015;168(7):517–527. doi:10.1002/ajmg.b.32328.
8. Huang DW, Sherman BT, Lempicki RA. **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** Nucleic Acids Research. 2009;37(1):1–13. doi:10.1093/nar/gkn923.
9. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. **EnrichNet: network-based gene set enrichment analysis.** Bioinformatics. 2012;28(18):i451–i457. doi:10.1093/bioinformatics/bts389.
10. Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, et al. **PINA v2.0: mining interactome modules.** Nucleic Acids Research. 2012;40(D1):D862–865. doi:10.1093/nar/gkr967.
11. Di Lena P, Martelli PL, Fariselli P, Casadio R. **NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases.** BMC Genomics. 2015;16(Suppl 8):S6. doi:10.1186/1471-2164-16-S8-S6.
12. Bovo S, Di Lena P, Martelli PL, Fariselli P, Casadio R. **NET-GE: a web-server for NETwork-based human gene enrichment.** Bioinformatics. 2016;32(22):3489–3491. doi:10.1093/bioinformatics/btw508.
13. Noble WS. **How does multiple testing correction work?** Nature Biotechnology. 2009;27(12):1135–1137. doi:10.1038/nbt1209-1135.
14. Maltezos S, Horder J, Coghlan S, Skirrow C, O'Gorman R, Lavender TJ, et al. **Glutamate/glutamine and neuronal integrity in adults with ADHD: a proton MRS study.** Translational Psychiatry. 2014;4:e373. doi:10.1038/tp.2014.11.
15. Centonze D, Bari M, Di Michele B, Rossi S, Gasperi V, Pasini A, et al. **Altered anandamide degradation in attention-deficit/hyperactivity disorder.** Neurology. 2009;72(17):1526–1527. doi:10.1212/WNL.0b013e3181a2e8f6.
16. Gamble KL, May RS, Besing RC, Tankersly AP, Fargason RE. **Delayed sleep timing and symptoms in adults with attention-deficit/hyperactivity disorder: a controlled actigraphy study.** Chronobiology International. 2013;30(4):598–606. doi:10.3109/07420528.2012.754454.
17. Zhu J, Reith M. **Role of the Dopamine Transporter in the Action of Psychostimulants, Nicotine, and Other Drugs of Abuse**. CNS & Neurological Disorders - Drug Targets. 2008;7(5):393–409. doi:10.2174/187152708786927877.
18. Cappi C, Brentani H, Lima L, Sanders SJ, Zai G, Diniz BJ, et al. **Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways.** Translational Psychiatry. 2016;6:e764. doi:10.1038/tp.2016.30.
19. Micheli V, Camici M, Tozzi MG, Ipata PL, Sestini S, Bertelli M, et al. **Neurological Disorders of Purine and Pyrimidine Metabolism**. Current Topics in Medicinal Chemistry. 2011;11(8):923–947. doi:10.2174/156802611795347645.
20. Moretti A, Gorini A, Villa RF. **Affective disorders, antidepressant drugs and brain metabolism.** Molecular Psychiatry. 2003;8(9):773–785. doi:10.1038/sj.mp.4001353.
21. Hines DJ, Haydon PG. **Astrocytic adenosine: from synapses to psychiatric disorders.** Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1654):20130594. doi:10.1098/rstb.2013.0594.
22. Mermi O, Atmaca M. **Thyroid gland functions are affected in obsessive-compulsive disorder**. Anatolian Journal of Psychiatry. 2016;17(2):99–103. doi:10.5455/apd.178087.

The Author(s) *BMC Genomics* 2017, **18**(Suppl 5):554
DOI 10.1186/s12864-017-3911-3

BMC Genomics

RESEARCH Open Access

eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes

Giulia Babbi[1], Pier Luigi Martelli[1*], Giuseppe Profiti[1], Samuele Bovo[1], Castrense Savojardo[1] and Rita Casadio[1,2]

From VarI-SIG 2016: identification and annotation of genetic variants in the context of structure, function, and disease
Orlando, Florida, USA. 09 July 2016

Abstract

Background: Genetic investigations, boosted by modern sequencing techniques, allow dissecting the genetic component of different phenotypic traits. These efforts result in the compilation of lists of genes related to diseases and show that an increasing number of diseases is associated with multiple genes. Investigating functional relations among genes associated with the same disease contributes to highlighting molecular mechanisms of the pathogenesis.

Results: We present eDGAR, a database collecting and organizing the data on gene/disease associations as derived from OMIM, Humsavar and ClinVar. For each disease-associated gene, eDGAR collects information on its annotation. Specifically, for lists of genes, eDGAR provides information on: i) interactions retrieved from PDB, BIOGRID and STRING; ii) co-occurrence in stable and functional structural complexes; iii) shared Gene Ontology annotations; iv) shared KEGG and REACTOME pathways; v) enriched functional annotations computed with NET-GE; vi) regulatory interactions derived from TRRUST; vii) localization on chromosomes and/or co-localisation in neighboring loci. The present release of eDGAR includes 2672 diseases, related to 3658 different genes, for a total number of 5729 gene-disease associations. 71% of the genes are linked to 621 multigenic diseases and eDGAR highlights their common GO terms, KEGG/REACTOME pathways, physical and regulatory interactions. eDGAR includes a network based enrichment method for detecting statistically significant functional terms associated to groups of genes.

Conclusions: eDGAR offers a resource to analyze disease-gene associations. In multigenic diseases genes can share physical interactions and/or co-occurrence in the same functional processes. eDGAR is freely available at: edgar.biocomp.unibo.it

Keywords: Gene/disease relationship, Protein-protein interaction, Protein functional annotation, Functional enrichment

Background

The advent of fast and relatively costless techniques for genome screening boosts the research of genetic determinants of human phenotypes, with a specific focus on diseases [1]. By this, lists of genes involved in several diseases/phenotypes are available. One of the most comprehensive database of curated associations between human Mendelian disorders and genes is OMIM [2], collecting 4510 phenotypes with known molecular basis (release of May 2016). Updated resources of associations between variations and diseases are stored in the NCBI-curated ClinVar [3], the UniProt curated Humsavar list [4], and the commercial version of HGMD [5]. Integrative datasets, such as DisGeNet [6] and MalaCards [7] collect lists of gene-disease associations from different sources. MalaCards includes text mining of the scientific literature, gene annotations in terms of shared GO terms and associated pathways. DisGeNet integrates data of disease-associated genes and their variants. Furthermore, a database collecting data on digenic diseases (related to concomitant defects in pairs of genes)

* Correspondence: pierluigi.martelli@unibo.it
[1]Biocomputing Group, BiGeA, University of Bologna, Bologna, Italy
Full list of author information is available at the end of the article

is available (DIDA, [8]) and reports the relationships between pairs of genes involved in 44 diseases.

As data accumulate, it emerges that an increasing number of diseases is associated with several genes. Independent or concomitant alterations in sequence or in expression of sets of genes are associated with the insurgence of genetically heterogeneous and polygenic diseases, respectively [9, 10]. The scenario is even more complicated when different environmental and life-style related factors have strong influence on the insurgence and severity of the pathology [11]. The complex nature of the association between genes and diseases is one of the major challenges of Precision Medicine programs [12].

Dissecting the molecular mechanisms at the basis of the association between genotype and phenotype requires a deep investigation of the features shared among genes (or proteins) co-involved in the same disease. Indeed, by analyzing molecular features and functional interactions, important biological processes and pathways implicated in the disease can emerge and other genes possibly involved in interaction networks can be discovered [13, 14].

This work describes eDGAR, a database of gene-disease associations, supplemented with the annotations of intergenic relationships in heterogeneous and polygenic diseases. We merged, without redundancy, data from OMIM [2], ClinVar [3], and Humsavar [4]. Disease nomenclature derives from OMIM. OMIM phenotype entries are classified according to the OMIM Phenotypic Series, which cluster different entries related to identical or highly similar diseases associated with different genes. As compared to the above mentioned databases, our focus is on specific structural and functional annotations of the genes. For each gene, the database reports the cytogenetic location, links to the Ensembl [15], SwissProt [4] and PDB entries [16], Gene Ontology (GO) [17] annotations and to the KEGG and REACTOME pathways, when available. For sets of genes involved in the same disease, the database collects from publicly available databases different types of relationships: physical interactions, co-occurrence in protein complexes, regulatory interactions, shared functions and pathways, and co-localization in neighboring cytogenetic loci. A network - based approach (NET-GE [18, 19]) provides statistical enrichment to functional terms. Information is organized in a relational database and an interface allows customized data search and retrieval.

The database is freely available at edgar.biocomp.unibo.it.

Construction and content

Data sources of associations between genes and diseases

In order to collect a comprehensive resource of associations among genes and diseases we integrated data from OMIM (May 2016 release) [2], ClinVar (May 2016 release) [3] and Humsavar (June 2016 release) [4]. The primary accessions for genes are HGNC codes [20], while OMIM identifiers are adopted to identify phenotypes. 2839 OMIM phenotype codes corresponding to identical or similar diseases, characterized by genetic heterogeneity, have been clustered into 357 phenotypic series, as defined by OMIM. Synonymic or alternative gene names were reduced to the HGNC gene primary codes, as reported in HGNC (June 2016 release).

On the overall, 5337, 4358 and 3365 gene-disease associations were collected from OMIM, ClinVar and Humsavar, respectively, by retaining only associations with unambiguous identification codes for both genes and diseases. After removing redundancy, the final dataset contains 5729 gene-disease associations, involving 3658 genes associated with 2672 diseases. These 2672 disease IDs correspond to 2315 OMIM IDs for phenotypes and 357 phenotypic series, or to 5154 when the 357 phenotypic series are brought back in 2839 OMIM IDs for phenotypes.

Gene annotation

All genes have been associated with the corresponding Ensembl codes (June 2016 version) [15] with BioMart [21]. Cytogenetic locations on the GrCh38 version of the human genome were therefrom derived. Out of 3658, 30 genes encode for microRNAs and tRNAs. For the 3628 protein coding genes, links to the SwissProt and PDB databases were also retrieved: all genes are linked to at least one SwissProt entry (for a total of 3718 entries) and 1682 genes are linked to at least one PDB entry (for a total of 14,578 PDB entries).

Functional annotation based on Gene Ontology (GO) terms was retrieved from GOOSE, the Online SQL Environment for GO terms implemented in the AmiGO2 portal [22]. All three GO sub-ontologies (Molecular Function: MF; Biological Process: BP; Cellular Component: CC) were considered. Given a GO term, the ancestor terms in the directed acyclic graph of GO (version 2.4) were retrieved by considering the relations "is a subtype of" and "part of". The information content (IC) was computed for each GO term, adopting standard methods [23], with the following equation:

$$IC = -log_2\left(\frac{N_{GO}}{N_{root}}\right) \quad (1)$$

where N_{GO} is the number of human genes endowed with the particular GO term and N_{root} is the number of human genes annotated with all the terms of the considered subontology, as derived from GOOSE [22]. IC lower limit is zero; high IC values indicate that a small number of genes is annotated with a particular GO term in the human genome and therefore the annotation is highly informative.

Associations with KEGG (version 77.0) [24] and REACTOME (version 53) [25] pathways were extracted from SwissProt.

Relationships among genes involved in the same disease

eDGAR integrates several information in order to annotate the possible relationships among protein coding genes related to the same polygenic or heterogeneous disease. The following features are considered:

- Protein-protein interactions, as derived from the multimeric structures deposited at the PDB (February 2016 release) [16], from STRING (version 10.0) [26] and from the experimental data available in BIOGRID (version 3.4) [27]. From the human STRING network, we retained only high confidence links (score ≥ 0.7) with annotated "action". Physical and genetic interactions of BIOGRID are reported separately. For all the considered human interactomes, eDGAR reports both direct and indirect interactions involving one intermediate gene. In addition, we supplemented data on interactions with selected annotations from manually curated features from SwissProt, including links to the PDB and the literature.
- Interactions in stable and functional complexes reported in the following resources: CORUM, listing 2837 mammalian complexes involving 3198 protein chains (16% of the human protein-coding genes) [28], the soluble complex census, listing 622 complexes involving 3006 protein chains [29]. This last resource is referred in the following as CENSUS.
- Functional GO terms and KEGG/REACTOME pathways shared by at least two genes.
- Functional GO terms and KEGG/REACTOME pathways retrieved with NET-GE [18, 19], a network based tool that performs the statistically-validated enrichment analysis of sets of human genes by exploiting the human STRING interactome; a significance of 5% was considered when retrieving statistically enriched terms on the basis of the Bonferroni-corrected *p*-values computed with NET-GE;
- Regulatory interactions derived from TRRUST [30], a curated database of interactions among 748 human transcription factors (TF) and 1975 non-TF targets. Given a set of genes associated with the same disease, eDGAR reports the presence of TF/target pairs and of groups of genes co-regulated by the same TF (belonging or not to the set);
- Co-localization in neighboring loci on the same chromosome: we highlighted genes located in the same cytogenetic band or in the tandem repeat regions listed in the DGD database [31]. DGD collects 945 groups consisting of 3543 genes in humans, likely deriving from duplications of ancestor genes.

Database structure and visualization

The database is implemented with PostgreSQL [32], an open source relational database system. Data stored in the database are retrieved using custom Python programs, while the output of the analysis is visualized in HTML pages using modern technologies like JavaScript. In particular, networks are encoded in JSON format and visualized using the JavaScript library D3.js [33]. We adopted a well known plug-in for jQuery called DataTables [34] for table visualizations, allowing the user to sort tables by columns and text-search inside each table.

Results and discussion

Statistics of the database content

The present release of eDGAR collects 5729 associations between 2672 diseases and 3658 different genes. Figure 1a plots the distribution of the number of genes associated with the same disease, which ranges from one (in 2051 monogenic diseases) to 69 (in the case of the "Retinitis pigmentosa" phenotypic series, OMIM: PS268000). The 621 diseases associated with multiple genes comprise both heterogeneous and polygenic diseases. On the overall, they account for 3678 associations with 2600 genes, 2576 of which code for proteins.

The database also shows a high level of pleiotropy (association of a single gene to several diseases) as shown in Fig. 1b. The most pleiotropic gene is FGFR3 that codes for the fibroblast growth factor receptor 3 and is associated with 16 different diseases.

Statistics of gene annotation

Table 1 lists major annotations of the 3658 genes related to diseases. All but 30 genes are coding for proteins reported in SwissProt; for 46.4% of them, structural information is available in PDB. Membrane proteins, transcription factors and enzymes account for 52%, 7% and 31%, respectively. Almost all the protein-coding genes are functionally annotated: the fraction of genes endowed with GO terms ranges from 94.2% to 98.6%, depending on the sub-ontology (Molecular Function (MF), Biological Process (BP) and Cellular Component (CC)). A smaller percentage of genes are associated with KEGG and REACTOME pathways (56.7% and 62.8%, respectively).

When considering human interactomes, 91.3% and 9.7% of the genes are present in BIOGRID with physical and genetic interactions, respectively; for 82.5% of the genes, STRING reports high confidence interactions (score ≥ 0.7). Some 20% of the genes encode for protein chains involved in functional complexes, as described in the CORUM and CENSUS collections. TRRUST lists some 1036 genes as part of the human regulatory

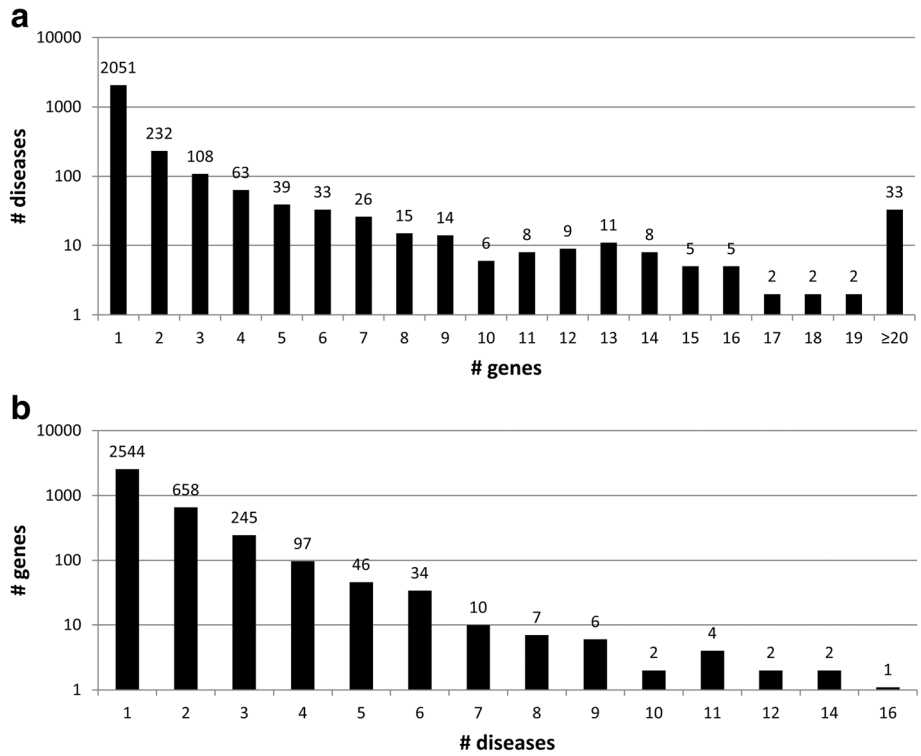


Fig. 1 Distribution of gene-disease associations. The Y-axis scale is logarithmic. **a** Number (#) of genes associated with diseases. 2672 diseases are distributed with respect to the number of associated genes. 2051 diseases are monogenic; 621 diseases are associated with multiple genes (from 2 to 69). **b** Number (#) of diseases associated to genes. 3658 genes are distributed with respect to the number of associated diseases. 2544 genes are associated with a single disease; 1114 genes are associated with multiple diseases (from 2 to 16)

network, of which 253 code for TFs and 783 are non-TF targets.

The level of annotation of the 2576 protein coding genes involved in heterogeneous or polygenic diseases is similar to that of all the genes collected in eDGAR.

Relations among genes associated with the same disease

eDGAR lists the relations among different genes associated with the same multigenic disease (statistics is in Table 2). 21.9% of diseases involve at least one pair of genes located in the same cytogenetic band and in 8.2% of the cases, genes are tandem repeats originated by duplications. These genes are likely to undergo the same regulation mechanisms and to be coexpressed [33].

Many diseases involve at least one pair of genes directly linked in interactomes: 40.3% and 46.9%, considering BIOGRID or STRING networks, respectively. The rates increase to 66.1% and 65.4% when considering also indirect interactions involving one intermediate gene not associated with the disease. 6.3% of diseases involve pairs of genes in a Transcription Factor (TF)/target relationship and 44% involve genes co-regulated by the same TF (considering also TFs not directly associated with the disease). The large majority of diseases (from 94.4% to 97.3%, depending on the sub-ontology) is associated with at least one pair of genes sharing GO terms. More than 90% of all the possible pairs of genes involved in the same disease have common BP and CC terms; the percentage is somehow smaller (76%) for MF sub-ontology. The total number of GO annotations shared by pairs of genes for BP, MF and CC is 72,787 (unique terms: 4582), 13,113 (unique terms: 915) and 16,298 (unique terms: 656), respectively. Overall, these data confirm the notion that genes associated with the same disease share some level of functional similarity, a view previously suggested for a small number of multigenic diseases [14]. However, being GO terms organized in a directed acyclic graph for each root, the information conveyed by the shared annotations can be very different, going from very general to very specific terms. The information content (IC, see Eq. 1) is routinely associated with GO terms in order to evaluate their specificity with respect of the available annotation of all human genes. The IC values of our dataset range from 0 (corresponding to the root GO term) to 10 (corresponding to the most specific terms). The average IC values for MF, BP and CC shared terms are 5.8 ± 1.7, 5.9 ± 1.7, and 5.8 ± 1.9, respectively. For each disease, the specificity of the annotation is evaluated by extracting the best IC values among the GO terms shared by pairs of co-associated genes (Fig. 2a).

Table 1 Gene annotation in eDGAR

	All diseases		Diseases associated with multiple genes	
	# genes[a]	# associated diseases[b]	# genes[a]	# associated diseases[b]
Total number	3658	2672	2600	621
Protein coding genes	3628 (100%)	2655 (100%)	2576 (100%)	619 (100%)
with PDB entry	1682 (46.4%)	1625 (61.2%)	1176 (45.7%)	512 (82.7%)
Membrane proteins	1891 (52.1%)	1644 (61.9%)	1364 (53.0%)	517 (83.5%)
Enzymes (with E.C number)	1112 (30.7%)	1045 (39.4%)	688 (26.7%)	363 (58.6%)
Reported in TRRUST (as TF)	253 (7.0%)	358 (13.5%)	179 (6.9%)	157 (25.4%)
Reported in TRRUST (as target)	783 (21.6%)	969 (36.5%)	570 (22.1%)	405 (65.4%)
Annotated with GO MF	3419 (94.2%)	2575 (97.0%)	2419 (93.9%)	617 (99.7%)
Annotated with GO BP	3538 (97.5%)	2619 (98.6%)	2514 (97.6%)	618 (99.8%)
Annotated with GO CC	3576 (98.6%)	2644 (99.6%)	2533 (98.3%)	618 (99.8%)
Associated with KEGG pathways	2057 (56.7%)	1868 (70.4%)	1430 (55.5%)	549 (88.7%)
Associated with REACTOME	2278 (62.8%)	2007 (75.6%)	1595 (61.9%)	563 (91.0%)
With physical BIOGRID interactions	3307 (91.3%)	2502 (94.2%)	2346 (91.2%)	609 (98.4%)
With genetic BIOGRID interactions	351 (9.7%)	472 (17.8%)	259 (10.1%)	247 (39.9%)
With STRING interactions	2992 (82.5%)	2341 (88.2%)	2146 (83.3%)	609 (98.4%)
Part of CORUM complexes	714 (19.7%)	706 (26.6%)	558 (21.7%)	340 (54.9%)
Part of CENSUS complexes	696 (19.2%)	689 (26.0%)	501 (19.4%)	296 (47.8%)
In tandem repeats	381 (10.5%)	448 (16.9%)	280 (10.9%)	234 (37.8%)

[a]Percentages are computed with respect to the number of protein coding genes
[b]Percentages are computed with respect to the number of diseases associated with protein coding genes

Table 2 Features shared by genes involved in the same heterogeneous or polygenic diseases

	# diseases	# pairwise relations	# protein coding genes
Total number	621	25,100	2576
With pairs of genes:			
In same cytogenetic band	136 (21.9%)	326 (1.3%)	335 (13.0%)
In tandem repeat	51 (8.2%)	58 (0.2%)	92 (3.6%)
In TF/target pairs	39 (6.3%)	81 (0.3%)	94 (3.6%)
Co-regulated by the same TF (not involved in the disease)	273 (44.0%)	2308 (9.2%)	626 (24.3%)
Sharing MF GO	586 (94.4%)	19,075 (76.0%)	2369 (92.0%)
Sharing BP GO	597 (96.1%)	22,948 (91.4%)	2502 (97.1%)
Sharing CC GO	604 (97.3%)	23,645 (94.2%)	2519 (97.8%)
Sharing KEGG pathway	349 (56.2%)	3129 (12.5%)	1074 (41.7%)
Sharing REACTOME pathway	474 (76.3%)	9806 (39.1%)	1554 (60.3%)
Interacting in PDB	96 (15.5%)	207 (0.8%)	199 (7.7%)
In the same CORUM complex	86 (13.8%)	469 (1.9%)	225 (8.7%)
In the same CENSUS complex	45 (7.2%)	166 (0.7%)	119 (4.6%)
Directly linked in STRING	291 (46.9%)	1535 (6.1%)	932 (36.2%)
Indirectly linked in STRING	115 (18.5%)	4355 (17.4%)	1346 (52.3%)
Directly linked in BIOGRID (physical interaction)	250 (40.3%)	944 (3.8%)	799 (31.0%)
Indirectly linked in BIOGRID (physical interaction)	160 (25.8%)	5228 (20.8%)	1607 (62.4%)
Directly linked in BIOGRID (genetic interaction)	9 (1.4%)	13 (0.1%)	19 (0.7%)
Indirectly linked in BIOGRID (genetic interaction)	25 (4.0%)	45 (0.2%)	62 (2.4%)

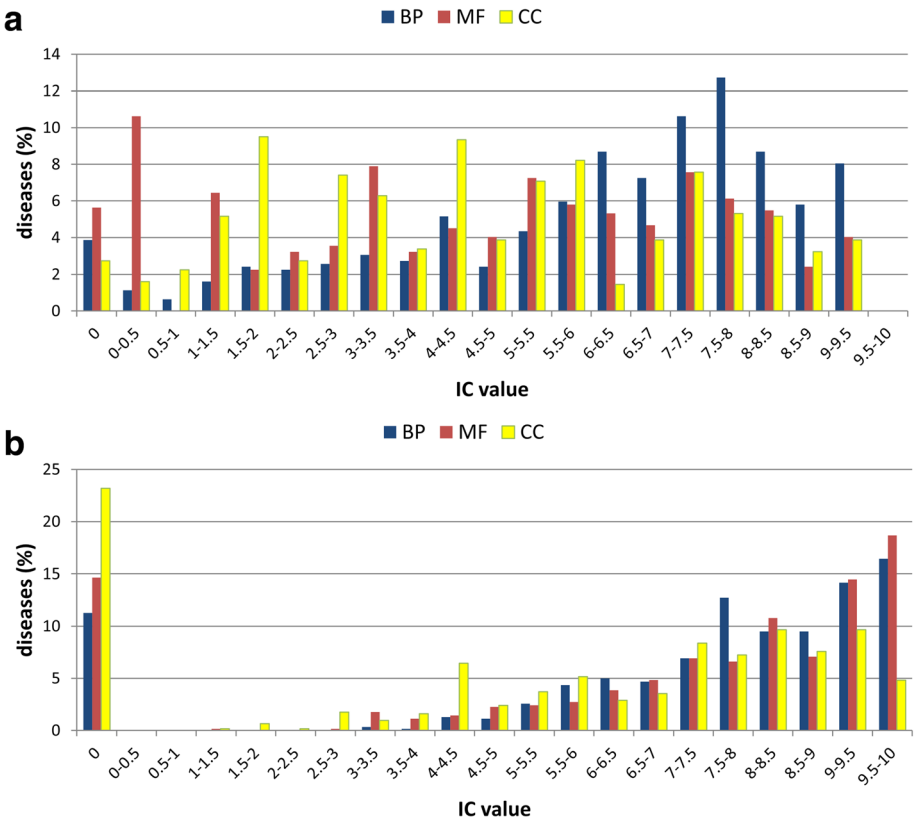


Fig. 2 Distribution of best IC values of GO terms for genes involved in multigenic diseases. **a** GO terms shared by genes; **b** GO terms after enrichment with NET-GE. For each multigenic disease, IC values of gene-associated GO terms (of the three different roots) are evaluated (Eq. 1). In the figure, the highest IC for each disease is shown. The frequency is computed with respect to the total number of multigenic diseases (621). When IC = 0, genes associated with multigenic disease do not share or enrich GO terms (panel **a** and **b** respectively)

For all the sub-ontologies, the best IC values are very spread, and it is evident that on average the most specific terms (highest IC values) belong to the BP sub-ontology: genes pairs sharing BP, MF and CC terms with IC ≥ 5 are present in 72%, 49% and 46% of the diseases, respectively (see Fig. 2a). When a different distribution based on a median is adopted, the pattern is very similar (Additional file 1: Fig. S1A). Genes involved in the same disease share also KEGG and REACTOME pathways (56.2% and 76.3%, respectively (Table 2)).

NET-GE enrichment

In order to better highlight functions shared by groups of genes associated with the same disease, we adopt NET-GE [18, 19], our recently developed network based tool for functional enrichment. For each functional sets of GO terms and/or KEGG or REACTOME pathways, NET-GE builds a network containing all the human genes annotated with the terms (seeds) and including all the connecting genes (the reference human interactome is derived from STRING). Input genes are mapped into the pre-computed NET-GE networks and enrichment analysis is performed. Outputs are Bonferroni-corrected *p*-values, measuring the overrepresentation of each term in the input set. Due to its network-based nature, NET-GE can enrich terms not present in the list of annotations of the input set. Table 3 lists the results of NET-GE on the groups of genes associated with the same disease, considering a 5% significance. For the majority of diseases, NET-GE enriches GO terms of the three sub-ontologies and pathways of KEGG and REACTOME. BP is the sub-ontology type most frequently enriched. The total number of GO annotations enriched for heterogeneous and polygenic diseases is 17,029, 4851 and 3910 (Table 3, rightmost column), with average IC values 6.1 ± 1.8, 7.1 ± 2, and 6.4 ± 2 for BP, MF and CC

Table 3 NET-GE functional enrichment of groups of genes involved in the same disease

	# diseases	# annotations
KEGG pathways	412 (66.3%)	2753
REACTOME pathways	488 (78.6%)	4130
GO MF terms	530 (85.3%)	4851
GO BP terms	551 (88.7%)	17,029
GO CC terms	477 (76.8%)	3910

respectively (Fig. 2b, reporting the distribution of the best IC values among the terms enriched for each disease; for a different distribution based on IC median values, see Additional file 1: Figure S1B).

The user interface

eDGAR is publicly available as a web server at edgar.biocomp.unibo.it with browsing and search options. Browsing is performed with the "Main Table" page that contains all the collected associations between genes and diseases, along with the indication of source databases.

The Search engine allows to access the database with different identifiers: HGNC symbols and Ensembl identifiers for genes, UniProt accession for proteins, OMIM identifiers or disease names for phenotypes and phenotypic series. The user may also search with a set of genes and retrieve shared annotation features.

Two types of pages can be visualized: i) gene specific pages, reporting the associations to diseases and the available gene annotations; ii) disease specific pages, reporting the associations with genes and, in case of heterogeneous and polygenic diseases, the list of relationships linking the different genes, organized into different tables. Interactions from STRING, PDB, BIOGRID, CORUM, CENSUS can also be visualized by means of graphs, reporting direct and indirect interactions. The graphs show the gene associated with the disease as blue nodes and other genes in interactions as pale blue nodes; the direct interactions are visualized as green edges and the indirect interactions as thin black edges (see Fig. 3). Clicking on a node, the user is redirected to the correspondent gene page.

A case study: Hypoparathyroidism

Hypoparathyroidism (OMIM 146200) is an endocrine deficiency disease characterized by low serum calcium levels, elevated serum phosphorus levels and absent or low levels of parathyroid hormone (PTH) in blood [35]. The metabolism of the patient may be altered: the vitamin D supply is inadequate and the magnesium metabolism is irregular. In some clinical panel, hypocalcemia can lead to dramatic effects such as tetany, seizures, altered mental status, refractory congestive heart failure, or stridor.

In eDGAR the familial isolated hypoparathyroidism (OMIM 146200) is associated with three different genes: GCM2 and PTH (both reported in OMIM, ClinVar and Humsavar) and CASR (reported only in ClinVar). CASR is an extracellular calcium-sensing receptor whose activity is mediated by G-proteins, PTH is the parathyroid hormone, whose function is to increase calcium level both by promoting the solution of bone salts and by preventing their renal excretion, and GCM2 (Glial cell missing homolog 2) is a probable transcriptional regulator, considering the SwissProt annotation. The "Transcription Factor (TF) annotation from TRRUST" table in eDGAR reports that GMC2 is a TF that regulates the expression of both PTH and CASR. Moreover, when considering "Interactions from STRING" table, PTH and CASR are in direct interaction, labelled as "binding" and "expression". The shared BP GO terms with the highest IC values are "response to vitamin D" and "response to fibroblast growth factor", both involving CASR and PTH. The response to vitamin D, whose metabolism is often altered in hypoparathyroidism, and a strict interplay between fibroblast growth factors and parathyroid hormone have been previously reported [36–38]. PTH and CASR are also involved in the same REACTOME pathways related to GPCR ligand binding and signaling. No shared KEGG term is found.

NET-GE enrichment for BP for the three genes include new terms endowed with high IC values, like "regulation of amino acid transport", "negative regulation of muscle contraction". Some of these new annotations are related to the severe symptoms of hypothyroidisms, namely tetany and seizure. NET-GE allows retrieving enriched KEGG pathways, such as "Circadian entrainment (hsa04713)", "Inflammatory mediator regulation of TRP channels (hsa04750)", "Gap junction (hsa04540)" and "Insulin secretion (hsa04911)". None of the three genes is directly involved in the four pathways; PTH and CASR are part of the networks defined by NET-GE exploiting the STRING network. Interestingly, these new annotations highlight previously reported impairments of both circadian rhythms impairment and insulin secretion associated with hypoparathyroidism [39, 40].

Figure 3 reports a summary of the information provided by eDGAR for hypothyroidism (OMIM 146200), showing how it allows to collect the different types of relations among the involved genes in a unique page integrating data from many resources.

Conclusions

eDGAR is a resource for the study of the associations between genes and diseases. It collects 2672 diseases, associated with 3658 different genes, for a total number of 5729 gene-disease associations. The novelty of eDGAR is the integration of different sources of gene annotation and in particular, for the 621 heterogeneous/polygenic diseases, eDGAR offers the possibility of analyzing functional and structural relations among co-involved genes. We provide direct interactions between pairs of genes (reported in STRING or BIOGRID) for 291 diseases and indirect interactions for some other 250 diseases. For 273 diseases, at least

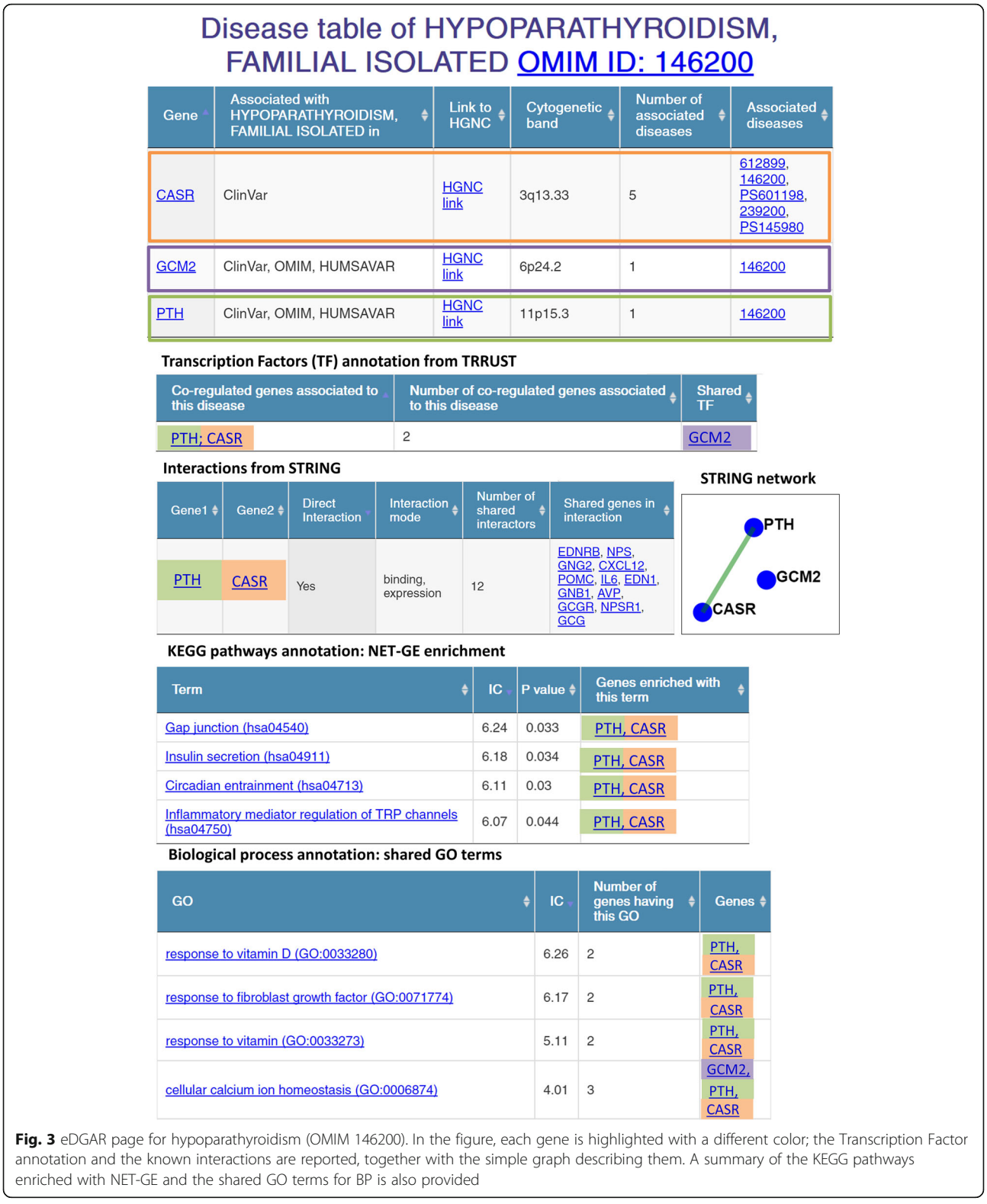


Fig. 3 eDGAR page for hypoparathyroidism (OMIM 146200). In the figure, each gene is highlighted with a different color; the Transcription Factor annotation and the known interactions are reported, together with the simple graph describing them. A summary of the KEGG pathways enriched with NET-GE and the shared GO terms for BP is also provided

one pair of genes is under regulatory interaction of the same TF, while 39 disease are associated with genes being a TF/target couple. For 612 diseases, at least one pair of genes share GO terms and/or KEGG/REACTOME pathways. In particular, genes involved in the same disease most frequently share terms of the BP sub-ontology. This is confirmed also when analyzing the statistically significant functional

terms enriched with NET-GE for 606 diseases. The relations among genes involved in the same disease are often complex and different pairs of genes are linked in different ways. eDGAR is a resource for better tackling the complexity of gene interactions at the basis of multigenic diseases. The database will be updated following the major releases of the different underlying data resources at least once a year.

Additional file

Additional file 1: Figure S1. Distribution of median IC values of GO terms for genes involved in multigenic diseases. A: GO terms shared by genes; B: GO terms enriched with NET-GE. For each multigenic disease, IC value of gene-associated GO terms (of the three different roots) are evaluated (Eq. 1). In the figure the median IC for each disease is shown. The frequency is computed with respect to the total number of multigenic diseases (621). When IC = 0, genes associated with multigenic disease do not share or enrich GO terms (panel A and B respectively). (PNG 393 kb)

Acknowledgements
Not applicable.

Funding
Publication costs for this article were provided by PRIN 2010-2011 project 20108XYHJS (to P.L.M.) (Italian MIUR); COST BMBS Action TD1101 and Action BM1405 (European Union RTD Framework Program, to R.C); PON projects PON01_02249 and PAN Lab PONa3_00166 (Italian Miur to R.C. and P.L.M.); FARB UNIBO 2012 (to R.C.).

Availability of data and materials
The dataset generated during the current study is available and downloadable at edgar.biocomp.unibo.it.

Authors' contributions
RC, PLM, and GB conceived and designed the work and wrote the paper. GB collected and curated data. SB ran the NET-GE predictions. GB, GP, and CS implemented the web server. PLM, GB and RC analysed and interpreted data on disease related variations. All authors critically revised and approved the manuscript.

Ethics approval and consent to participate
The authors declare that they used only public data.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Biocomputing Group, BiGeA, University of Bologna, Bologna, Italy. [2]Interdepartmental Center «Giorgio Prodi» for Cancer Research, University of Bologna, Bologna, Italy.

Published: 10 August 2017

References

1. Kann MG. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Brief Bioinform. 2010;11(1):96–110.
2. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.Org: online Mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue): D789–98.
3. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
4. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–12.
5. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics. 2012;39:1.13:1.13.1–1.13.20.
6. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucl Acids Res. 2016;45(D1):D833–9.
7. Rappaport N, Twik M, Plaschkes I, Nudel R, Stein TI, Levitt J, Gershoni M, Morrey CP, Safran M. Lancet D; MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucl Acids Res. 2016;45(D1):D877–87.
8. Gazzo AM, Daneels D, Cilia E, Bonduelle M, Abramowicz M, Van Dooren S, Smits G, Lenaerts T. DIDA: a curated and annotated digenic diseases database. Nucleic Acids Res. 2016;44(D1):D900–7.
9. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210–7.
10. Weeks DE, Lathrop GM. Polygenic disease: methods for mapping complex disease traits. Trends Genet. 1995;11(12):513–9.
11. Fu W, O'Connor TD, Akey JM. Genetic architecture of quantitative traits and complex diseases. Curr Opin Genet Dev. 2013;23(6):678–83.
12. Cardon LR, Harris T. Precision medicine, genomics and drug discovery. Hum Mol Genet. 2016;25(R2):R166–72.
13. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A. 2007;104(21):8685–90.
14. Oti M, Brunner H. The modular nature of genetic diseases. Clin Genet. 2007;71:1–11.
15. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe K, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P, Searle SM. The Ensembl gene annotation system. Database (Oxford). 2016; pii: baw093.
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN. Bourne PE the protein data bank. Nucleic Acids Res. 2000;28:235–42.
17. The Gene Ontology Consortium.. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2016. pii: gkw1108.
18. Di Lena P, Martelli PL, Fariselli P, Casadio R. NET-GE: a novel NETwork-based Gene Enrichment for detecting biological processes associated to Mendelian diseases. BMC Genomics. 2015;16(Suppl 8):S6.
19. Bovo S, Di Lena P, Martelli PL, Fariselli P, Casadio R. NET-GE: a web-server for NETwork-based human gene enrichment. Bioinformatics. 2016;32(22):3489–91.
20. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2016. pii: gkw1033.
21. Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford). 2011:bar049.
22. Munoz-Torres M, Carbon S. Get GO! Retrieving GO data using AmiGO, QuickGO, API, files, and tools. Methods Mol Biol. 2017;1446:149–60.
23. Shannon CE. A mathematical theory of communication. Bell Syst Techn J. 1948;27:379–423.
24. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457–62.
25. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2016;44(D1):D481–7.
26. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447–52.

27. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43(Database issue):D470–8.
28. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. 2010;38(Database issue):D497–501.
29. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, Babu M, Craig SA, Hu P, Wan C, Vlasblom J, Dar VU, Bezginov A, Clark GW, Wu GC, Wodak SJ, Tillier ER, Paccanaro A, Marcotte EM, Emili A. A census of human soluble protein complexes. Cell. 2012;150(5):1068–81.
30. Han H, Shim H, Shin D, Shim JE, Ko Y, Shin J, Kim H, Cho A, Kim E, Lee T, Kim H, Kim K, Yang S, Bae D, Yun A, Kim S, Kim CY, Cho HJ, Kang B, Shin S, Lee I. TRRUST: a reference database of human transcriptional regulatory interactions. Sci Rep. 2015;5:11432.
31. Ouedraogo M, Bettembourg C, Bretaudeau A, Sallou O, Diot C, Demeure O, Lecerf F. The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. PLoS One. 2012;7(11):e50653.
32. PostgreSQL. https://www.postgresql.org/. Accessed 1 December 2016.
33. Data-Driven. Documents. https://d3js.org/. Accessed 1 December 2016.
34. DataTables. https://datatables.net/. Accessed 1 December 2016.
35. Bilezikian J, Khan A, Potts J, et al. Hypoparathyroidism in the adult: epidemiology, diagnosis, pathophysiology, target organ involvement, treatment, and challenges for future research. J Bone Miner Res. 2011;26(10):2317–37.
36. Lai Y, Wang H, Xia X, Wang Z, Fan C, Wang H, Zhang H, Ding S, Teng W, Shan Z. Serum fibroblast growth factor 19 is decreased in patients with overt hypothyroidism and subclinical hypothyroidism. Medicine (Baltimore). 2016;95(39):e5001.
37. Domouzoglou EM, Fisher FM, Astapova I, Fox EC, Kharitonenkov A, Flier JS, Hollenberg AN, Maratos-Flier E. Fibroblast growth factor 21 and thyroid hormone show mutual regulatory dependency but have independent actions in vivo. Endocrinology. 2014;155(5):2031–40.
38. Lee Y, Park YJ, Ahn HY, Lim JA, Park KU, Choi SH, Park DJ, Oh BC, Jang HC, Yi KH. Plasma FGF21 levels are increased in patients with hypothyroidism independently of lipid profile. Endocr J. 2013;60(8):977–83.
39. Bauer MS, Soloway A, Dratman MB, Kreider M. Effects of hypothyroidism on rat circadian activity and temperature rhythms and their response to light. Biol Psychiatry. 1992;32(5):411–25.
40. Yang N, Yao Z, Miao L, Liu J, Gao X, Fan H, Hu Y, Zhang H, Xu Y, Qu A, Wang G. Novel clinical evidence of an association between Homocysteine and insulin resistance in patients with hypothyroidism or subclinical hypothyroidism. PLoS One. 2015;10(5):e0125922.

Received: 17 January 2017 | Revised: 16 March 2017 | Accepted: 24 March 2017

DOI: 10.1002/humu.23222

SPECIAL ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4

Qifang Xu[1] | Qingling Tang[2] | Panagiotis Katsonis[3] | Olivier Lichtarge[3] | David Jones[4] | Samuele Bovo[5] | Giulia Babbi[5] | Pier L. Martelli[5] | Rita Casadio[5] | Gyu Rie Lee[6] | Chaok Seok[6] | Aron W. Fenton[2] | Roland L. Dunbrack Jr[1]

[1]Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia, Pennsylvania

[2]Department of Biochemistry and Molecular Biology, The University of Kansas Medical Center, Kansas City, Kansas

[3]Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, Texas

[4]Department of Computer Science, University College London, London, United Kingdom

[5]Biocomputing Group, CIG/Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna, Italy

[6]Department of Chemistry, Seoul National University, Seoul, Republic of Korea

Correspondence
Aron W. Fenton, The University of Kansas Medical Center, Biochemistry and Molecular Biology, MS 3030, 3901 Rainbow Boulevard, Kansas City, Kansas 66160.
Email: afenton@kumc.edu
Roland L. Dunbrack, Jr. Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Ave., Philadelphia, PA 19111.
Email: roland.dunbrack@fccc.edu

Contract grant sponsors:NIH (R01 GM084453, R13 HG006650, U41 HG007346, R13 HG006650).

For the CAGI Special Issue

Abstract

The Critical Assessment of Genome Interpretation (CAGI) is a global community experiment to objectively assess computational methods for predicting phenotypic impacts of genomic variation. One of the 2015–2016 competitions focused on predicting the influence of mutations on the allosteric regulation of human liver pyruvate kinase. More than 30 different researchers accessed the challenge data. However, only four groups accepted the challenge. Features used for predictions ranged from evolutionary constraints, mutant site locations relative to active and effector binding sites, and computational docking outputs. Despite the range of expertise and strategies used by predictors, the best predictions were marginally greater than random for modified allostery resulting from mutations. In contrast, several groups successfully predicted which mutations severely reduced enzymatic activity. Nonetheless, poor predictions of allostery stands in stark contrast to the impression left by more than 700 PubMed entries identified using the identifiers "computational + allosteric." This contrast highlights a specialized need for new computational tools and utilization of benchmarks that focus on allosteric regulation.

KEYWORDS

allosteric effect, CAGI experiment, liver pyruvate kinase, missense mutation

1 | INTRODUCTION

Blind challenge experiments, such as CASP (Moult et al., 2016) and CAPRI (Lensink et al., 2017), have provided independent assessment of computational prediction methods in structural biology. They have spurred the development of new methods and the integration of multiple methods in prediction pipelines. The Critical Assessment of Genome Interpretation (CAGI) experiment seeks to achieve the same goals by providing prediction challenges in a number of different areas. In this report, we describe a challenge involving the effect of mutations on the allosteric coupling of effectors and substrate binding to human liver pyruvate kinase (L-PYK). The focus of this competition was to predict the influence of mutations on the allosteric regulation of L-PYK by a negative regulator, alanine, and a positive effector, fructose-1,6-bisphosphate (Fru-1,6-BP). Numerous methods for predicting the effect of mutations on allosteric effector binding have been published in recent years (Collier & Ortiz, 2013; Feher et al., 2014).

The definition of allostery applicable to studies of L-PYK is the affinity of the enzyme for its substrate, phosphoenolpyruvate (PEP), in the absence versus presence of an allosteric effector, recognizing that the effector binds to a site distinct from the active site (Carlson & Fenton, 2016; Fenton, 2008, 2012; Fenton & Alontaga, 2009; Fenton &

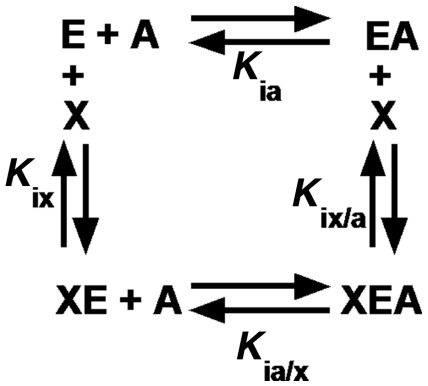


FIGURE 1 Reaction scheme for an allosteric energy cycle in which an enzyme (E) can bind one substrate (A) and one allosteric effector (X). K_{ia} is the equilibrium dissociation constant of the substrate binding to the enzyme in the absence of effector. $K_{ia/x}$ is the equilibrium dissociation constant of the substrate binding to the enzyme in the presence of saturating concentrations of effector. K_{ix} is the equilibrium dissociation constant of the effector when substrate is absent, whereas $K_{ix/a}$ is the equilibrium dissociation constant of effector in the presence of saturating concentrations of substrate

Hutchinson, 2009; Fenton et al., 2010; Ishwar et al., 2015). This definition describes allostery by four enzyme forms that constitute the corners of a thermodynamic energy cycle (Fig. 1), and it provides a mechanism to quantify allosteric function in the form of the allosteric coupling constant (Q_{ax}) (Fenton, 2012; Reinhart, 1983, 1988, 2004; Weber, 1972):

$$Q_{ax} = \frac{K_{ia}}{K_{ia/x}} = \frac{K_{ix}}{K_{ix/a}}$$

K_{ia} and $K_{ia/x}$ are equilibrium dissociation constants for binding the substrate (A) in the absence or presence, respectively, of an allosteric effector, X, as defined in Figure 1. $Q_{ax} = 1$ indicates that the system is not allosteric. When $Q_{ax} > 1$, there is positive allosteric coupling between the binding of X to a protein and the binding of A to the same protein at distinct sites. When $Q_{ax} < 1$, there is a negative or inhibitory coupling between the X and A sites.

The predictors were provided two sets of mutations for predictions of enzyme activity and allosteric effects in L-PYK. Q_{ax} was determined for each active mutant protein by determining PEP affinity (via titrations of activity over a concentration range of PEP) over a concentration range of effector. Experiment 1 consisted of 113 mutations at nine sites in or near to the binding of the negative allosteric regulator, alanine. Participants were asked to provide a probability that each mutant enzyme was active (i.e., not the level of activity) and the value of Q_{ax} for alanine for each mutant. Experiment 2 consisted of mutations to alanine at 430 sites throughout the protein. Participants were then asked to predict the enzyme activity and Q_{ax} values for the effectors alanine and Fru-1,6-BP. Since alanine is a negative regulator, all values of $Q_{ax\text{-}Ala}$ are between 0 and 1, whereas the value of Q_{ax} for Fru-1,6-BP is unbounded. Predictors were provided with the maximum value ($Q_{ax\text{-}Fru\text{-}1,6\text{-}BP} = 320$) found in the alanine-scanning experiment.

2 | METHODS AND MATERIALS

2.1 | Experimental data generation

Wild-type and mutant human L-PYK were expressed in the *E. coli* FF50 strain, which lacks endogenous *pyk* genes, and partially purified using ammonium sulfate fractionation followed by dialysis, as previously described (Fenton & Alontaga, 2009; Ishwar et al., 2015). L-PYK catalyzes the following reaction:

$$\text{Phosphoenolpyruvate} + \text{ADP} \rightarrow \text{Pyruvate} + \text{ATP}$$

Activity measurements were performed at 30°C using a lactate dehydrogenase assay to detect the production of pyruvate by L-PYK. Lactate dehydrogenase catalyzes the following reversible reaction:

$$\text{Pyruvate} + \text{NADH} \rightleftharpoons \text{NAD}^{*} + \text{Lactate}$$

As the L-PYK reaction proceeds, producing pyruvate, the concentration of NADH decreases, which can be detected by monitoring absorbance at 340 nm (A_{340}). Reaction conditions contained 50 mM HEPES or bicine, 10 mM $MgCl_2$, 2 mM (K)ADP, 0.1 mM EDTA, 0.18 mM NADH, and 19.6 U/ml lactate dehydrogenase. PEP and effector concentrations were varied. The rate of the decrease in A_{340} due to NADH utilization was recorded at each concentration of PEP and these initial velocity rates as a function of PEP concentration were used to evaluate the apparent affinity for PEP ($K_{app\text{-}PEP}$) at any one effector concentration. K_{ix} and Q_{ax} for each mutant and the wild type were obtained by fitting the observed $K_{app\text{-}PEP}$ to the equation:

$$K_{app-PEP} = K_a \left(\frac{K_{ix} + [X]}{K_{ix} + Q_{ax}[X]} \right)$$

where $K_a = K_{app\text{-}PEP}$ when the concentration of effector $[X] = 0$.

The dataset represents two experiments, which are characterizations of mutant human L-PYK proteins expressed in *E. coli*, named experiment 1 and experiment 2. Experiment 1 consisted of site-directed mutations at residue positions with a side chain contacting with alanine or very near the bound alanine. A total of 113 substitutions were introduced at nine different sites, of which 23 mutant proteins were completely inactive (no measurable enzyme activity). $Q_{ax\text{-}Ala}$ was determined for the 90 mutant proteins with activity. In experiment 2, 430 residues were mutated into alanine across the entire protein, of which 44 did not have detectable enzyme activity. Allosteric coupling Q_{ax} for inhibition by alanine and activation by Fru-1,6-BP were separately determined.

2.2 | Performance assessment of L-PYK enzyme activity

From the binary experimental enzyme activity data (1 = positive = active; 0 = negative = inactive), we calculated the number of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) for all participating groups in experiment 1 and experiment 2. From these, we calculated the true-positive rate (TPR),

TABLE 1 Groups participating in L-PYK enzyme activity and allostery prediction challenges

Group number	Affiliation	Authors
53	Department of Human and Molecular Genetics, Baylor College of Medicine, Houston, TX	Panagiotis Katsonis, Olivier Lichtarge
54	Department of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom	David Jones
55	Biocomputing Group, CIG/Interdepartmental Center «Luigi Galvani» for Integrated Studies of Bioinformatics, Biophysics and Biocomplexity, University of Bologna, Bologna, Italy	Samuele Bovo, Giulia Babbi, Pier Luigi Martelli, Rita Casadio
56	Department of Chemistry, Seoul National University, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea	Gyu Rie Lee, Chaok Seok

true-negative rate (TNR), positive predictive value (PPV), and negative predictive value (NPV):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$

We also calculated four measures that assess overall accuracy: total accuracy (ACC), balanced accuracy (BACC), Matthews correlation coefficient (MCC) (Matthews, 1975), and F1 score. F1 score is the harmonic mean of precision (PPV) and sensitivity (TPR).

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

$$\mathrm{BACC} = \frac{1}{2}(\mathrm{TPR}+\mathrm{TNR})$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

$$\mathrm{F1} = 2\frac{\mathrm{TPR}\times\mathrm{PPV}}{\mathrm{TPR}+\mathrm{PPV}}$$

Since some predictors provided real values (between 0 and 1), these were converted into binary predictions as described below in the *Results* section.

2.3 | Evaluation of predictions of $Q_{ax\text{-}Ala}$ and $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$

Spearman's rho (ρ), or Spearman's rank correlation coefficient, measures the monotonic correlation between prediction and experimental data. $\rho = 1$ means the predictions and experimental data points have identical rankings. For data set (p_i, e_i), prediction data points are converted into ranks Rp_i, and experimental data points are converted into ranks Re_i. Then, ρ is calculated from the formula:

$$\rho = \frac{\mathrm{cov}(\mathrm{Rp},\mathrm{Re})}{\sigma_{\mathrm{Rp}}\sigma_{\mathrm{Re}}}, \quad -1 \le \rho \le 1$$

Kendall's tau (τ), or Kendall rank correlation coefficient, like Spearman's rho, measures the rank correlation between two variables. For data set (p, e), any pair of (p_i, e_i) and (p_j, e_j), where $i \neq j$, are said to be concordant if both $p_i > p_j$ and $e_i > e_j$, or if both $p_i < p_j$ and $e_i < e_j$. They are discordant, if both $p_i > p_j$ and $e_i < e_j$, or if $p_i < p_j$ and $e_i > e_j$. If $p_i = p_j$ or $e_i = e_j$, the pair is neither concordant nor discordant. We use C for the set of concordant pairs, and D for the set of discordant pairs. τ is defined as the difference between the number of concordant pairs (|C|) and the number of discordant pairs (|D|), divided by the total number of pair combinations ($n \times (n-1)/2$). The formula is given as following:

$$\tau = \frac{|C| - |D|}{n(n-1)/2}$$

All statistical calculations and kernel density estimates of the data were performed in R (R Core Team, 2015).

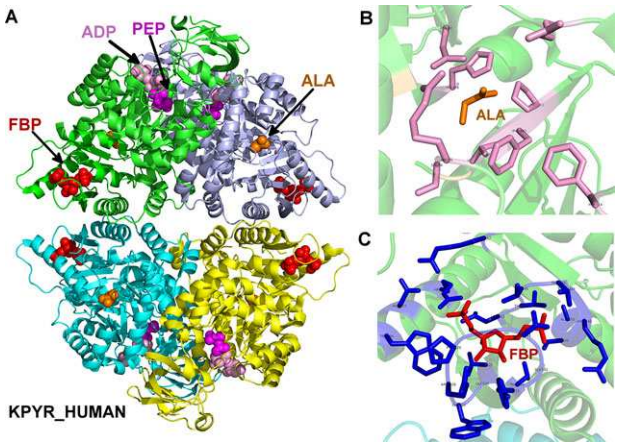


FIGURE 2 Structure of human pyruvate kinase, as well as the binding sites of inhibitor alanine and activator fructose-1,6-bisphosphate. A: A modeled structure of L-PYK tetramer with substrates PEP and ADP, allosteric inhibitor alanine, and allosteric activator. PEP, ADP, alanine (labeled ALA), and fructose-1,6-bisphosphate (labeled FBP) are shown in spheres, colored in magenta, pink, orange, and red, respectively. The structure was assembled by superposing monomers from several structures of homologues of L-PYK with PEP, ADP, and alanine bound onto a tetrameric structure of human L-PYK with fructose-1,6-bisphosphate bound (PDB: 4IP7). B: The allosteric binding site of alanine. Alanine is shown in sticks and colored in orange. Residues that were mutated in experiment 1 are shown in sticks, and colored in pink. C: The binding site of fructose-1,6-bisphosphate (FBP). FBP is shown in sticks and colored in red. Interacting residues are shown in sticks and colored in blue

3 | RESULTS

In this assessment, four groups (53, 54, 55, and 56; Table 1) submitted a total of five prediction sets, of which two were from group 56, labeled 56_1 and 56_2. The methods utilized by each group are provided in the Supp. Materials as are the instructions and information provided to predictors at the time of the experiment.

Human L-PYK is a tetrameric enzyme with distinct binding sites for its reactants, pyruvate, and ADP, and its allosteric effectors, alanine, and Fru-1,6-BP. The structure of the tetramer is shown in Figure 2A, where molecules at the three sites are represented as spheres in each monomer. This composite structure was created by superposing monomers from structures containing alanine (PDB: 2G50, a structure of rabbit L-PYK) (Williams et al., 2006), PEP (PDB: 4HYV, *Trypanosoma brucei* pyruvate kinase) (Zhong et al., 2013), and ADP (PDB: 3GR4, human pyruvate kinase M2) (Hong et al., unpublished, DOI: 10.2210/pdb3gr4/pdb) onto each member of the tetrameric biological assembly of human L-PYK (PDB: 4IP7) (Holyoak et al., 2013). Experiment 1 consisted of 113 mutations spread across nine amino acid positions in or near the alanine-binding site (Fig. 2B): Arg55, Ser56, Asn82, Arg118, His476, Val481, Pro483, and Phe514. Experiment 2 consisted of alanine-scanning mutations across the entire protein, except wild-type positions that are Gly or Ala. The Fru-1,6-BP site is shown in Figure 2C.

3.2 | Prediction of L-PYK enzyme activity

The first challenge was to provide a probability that each enzyme was active. This was a binary outcome, not the level of activity. Even weakly active enzymes were considered active in the experiment. In both experiments, some mutants had no detectable activity, and these were labeled 0; the rest were labeled 1. The active mutants included some enzymes with very low but detectable activity. In experiment 1, 79.6% of mutants were active and 20.4% were inactive. In experiment 2, 88.8% of the mutants were active and 10.2% were inactive. Two of the groups (53 and 54) submitted real values between 0 and 1, instead of binary indicators. For these groups, we labeled all predictions with values $\geq$0.5 as active and the rest as inactive. Figure 3 shows the density functions of predicted enzyme activities. For experiment 1, two groups (55 and 56_2) predicted all mutants to be active (a value of 1) (Fig. 3, top row). This is not unreasonable since all of the mutations were in or near the alanine effector-binding site, which is distant from the active site.

Table 2 provides an assessment of the predictions of enzyme activity for each group for both experiments. We also included values obtained from the PolyPhen-2 server, which is commonly used to predict phenotypes of missense mutations (Adzhubei et al., 2010). Group 56 achieved the highest ACC in both experiments (ACC of 0.867 for group 56_1 in experiment 1; ACC of 0.894 for group 56_2 in experiment 2). Since the goal was to predict whether enzymes were active or inactive, rather than the level of activity, this is a successful result. In the case of experiment 1, predicting all mutants as active would result in an accuracy of 0.796, whereas in experiment 2, a value of 0.888 would be obtained. At least for experiment 1, group 56 achieved better predictions than the simple prediction that all mutants were active.

In most binary phenotype prediction assessments (Wei & Dunbrack, 2013), it is important to balance the success of positive predictions and/or experimental outcomes with negative predictions and/or experimental outcomes. One such measure is the BACC, which is the average of the rate of correctly predicting the experimentally active mutants (TPR) and the rate of correctly predicting the experimentally inactive mutants (TNR). For experiment 1, only groups 53 and 56_1 achieved BACC values above 0.5, with BACC = 0.768 and 0.755, respectively. A BACC of 0.50 is trivial to achieve, since if one predicts all of the phenotypes in one class, the BACC is automatically 0.50 (e.g., groups 55 and 56_2 for experiment 1). Groups 53 and 56_1 achieved their results in contrasting manners: group 53 has low TPR and high TNR, and group 56_1 has high TPR and low TNR. For experiment 2, which contained mutations across the entire protein and is therefore a more real-world prediction task, only group 53 has TPR and TNR > 0.5, resulting in a BACC of 0.745.

Similarly, the MCC and F1 values also balance positive and negative predictions and experimental values but in different ways than BACC (see *Materials and Methods*). F1, in particular, only includes positive predictions and experimental phenotypes and omits negative predictions and phenotypes. Since both data sets consisted of majority of active enzymes (80% and 88% for experiments 1 and 2, respectively), groups that predicted a larger fraction of the enzymes to be active did better in F1 (groups 55, 56_1, and 56_2) than the other groups. Group 54 predicted a majority of the mutants to be inactive in both experiments and thus achieved much lower values for F1 than the other groups.

We compared the results of CAGI groups with that of PolyPhen-2, a server that is commonly used to predict the phenotypes of missense mutations in proteins. PolyPhen-2, like other servers, predicts phenotypes to be deleterious or neutral, or "damaging" versus "benign." This is not necessarily directly associated with enzyme activity, since a deleterious mutation might affect protein expression or the ability to regulate the protein by allosteric mechanisms. Also, the inactive enzymes were only those with no activity, and not those with significant reduction in activity. In experiment 1, PolyPhen-2 predicted most mutants to be inactive, probably because the alanine-binding site is very highly conserved in L-PYK enzymes in order to retain the negative effector capability of alanine. This resulted in a BACC of 0.539. In experiment 2, mutations were spread across the protein and PolyPhen-2 does better, with a BACC of 0.674. Nevertheless, group 53 was able to achieve better results on all four measures of overall success in experiment 2.

As mentioned above, groups 53 and 54 provide real values (not binary values) for the enzyme activity. We speculated that a cutoff of 0.5 might not be ideal to turn their real values into binary predictions. We calculated BACC as function of the cutoff and found that for group 53, a value of 0.5 was still the best for both experiments. But for group 54, values of 0.3 for experiment 1 and 0.35 for experiment 2 provide better results. The values of BACC are 0.724 and 0.696, respectively, which are much better than the 0.5 cutoff (0.534 and 0.627, respectively). But this is only possible with reference to the experimental data, which would not be available in real-world situations. Since the density for predictions for group 54 were

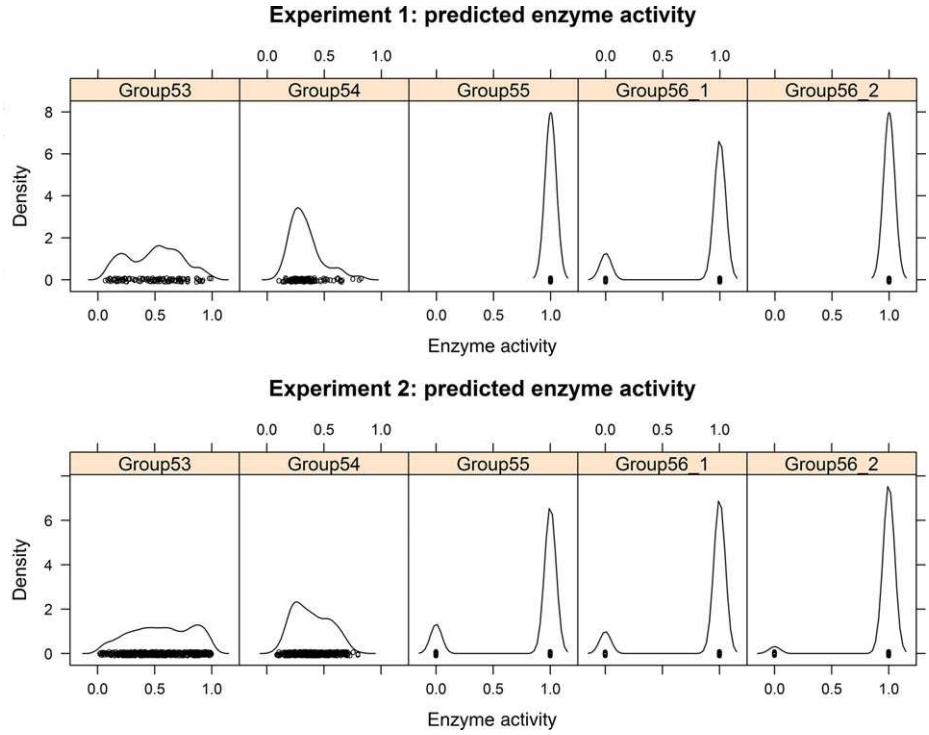


FIGURE 3 Kernel density estimates of five sets of predicted L-PYK enzyme activities

TABLE 2 Binary prediction results of L-PYK enzyme activity

	Experiment 1						Experiment 2					
Method	Group 53	Group 54	Group 55	Group 56_1	Group 56_2	PPH2	Group 53	Group 54	Group 55	Group 56_1	Group 56_2	PPH2
TPR	0.622	0.156	1	0.944	1	0.122	0.626	0.322	0.838	0.898	0.976	0.392
TNR	0.913	0.913	0	0.565	0	0.957	0.864	0.932	0.205	0.318	0.182	0.953
PPV	0.966	0.875	0.796	0.895	0.796	0.917	0.976	0.976	0.901	0.920	0.912	0.987
NPV	0.382	0.216	0	0.722	0	0.218	0.210	0.137	0.127	0.264	0.471	0.150
ACC	0.681	0.310	0.796	**0.867**	0.796	0.292	0.650	0.385	0.772	0.838	**0.894**	0.449
BACC	**0.768**	0.534	0.5	0.755	0.5	0.539	**0.745**	0.627	0.521	0.608	0.579	0.673
MCC	0.431	0.079	0	**0.561**	0	0.103	**0.301**	0.169	0.034	0.199	0.246	0.218
F1	0.757	0.264	0.887	**0.919**	0.887	0.217	0.762	0.484	0.868	0.907	**0.943**	0.562

Notes:
The highest score in each row for the four global measures is in bold and underlined.
0, inactive; 1, active.
TPR, true-positive rate; FPR, false-positive rate; TNR, true-negative rate; PPV, positive predictive value; NPV, negative predictive value; ACC, accuracy; BACC, balanced accuracy; MCC, Matthews correlation coefficient; F1, F1 score.

unimodal (Fig. 3), it was not possible to define a cutoff based on a minimum of density between a low-activity and a high-activity mode in the data.

3.3 | Prediction of allosteric inhibition of alanine ($Q_{ax\text{-}Ala}$)

The second challenge was to estimate the inhibitory allosteric effect of binding alanine, $Q_{ax\text{-}Ala}$ on binding of the substrate PEP. The density estimates of experimental $Q_{ax\text{-}Ala}$ values of two experiments are shown in Figure 4. The wild-type enzyme had a $Q_{ax\text{-}Ala}$ value of ~0.08 in both experiments. In experiment 1, 23 out of 90 mutants did not have measurable allosteric coupling, shown in a peak at $Q_{ax} = 1$ (Fig. 4, left). One possiblity is that alanine continues to bind to these mutant proteins, but that binding does not alter PEP affinity. In other cases, the $Q_{ax} = 1$ outcome is likely because the mutation eliminated binding of Ala to L-PYK altogether (at least to the maximum concentration tested in the experiments). In experiment 2, after excluding 37 mutants for which the allosteric coupling effect could not be measured, the $Q_{ax\text{-}Ala}$ values of 325 (83%) mutants were between 0 and 0.2, relatively similar to the wild-type enzyme.

A comparison by scatter plot of the experimental and the predicted $Q_{ax\text{-}Ala}$ values is shown in Figure 5. Group 55 provided only binary prediction for $Q_{ax\text{-}Ala}$. Group 56_1 and 56_2 provided identical values for both experiments. The scatter plots do not show any obvious correlations between the predicted and experimental $Q_{ax\text{-}Ala}$.

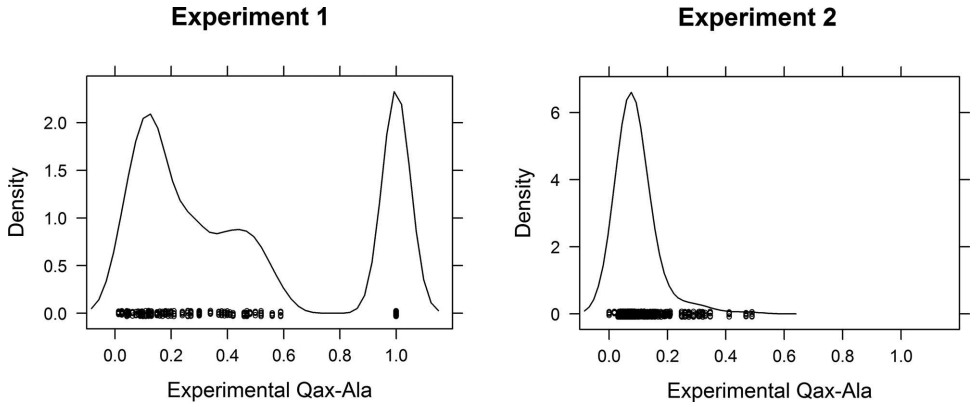


FIGURE 4 Kernel density estimates of experimental $Q_{ax\text{-}Ala}$ values of experiments 1 and 2

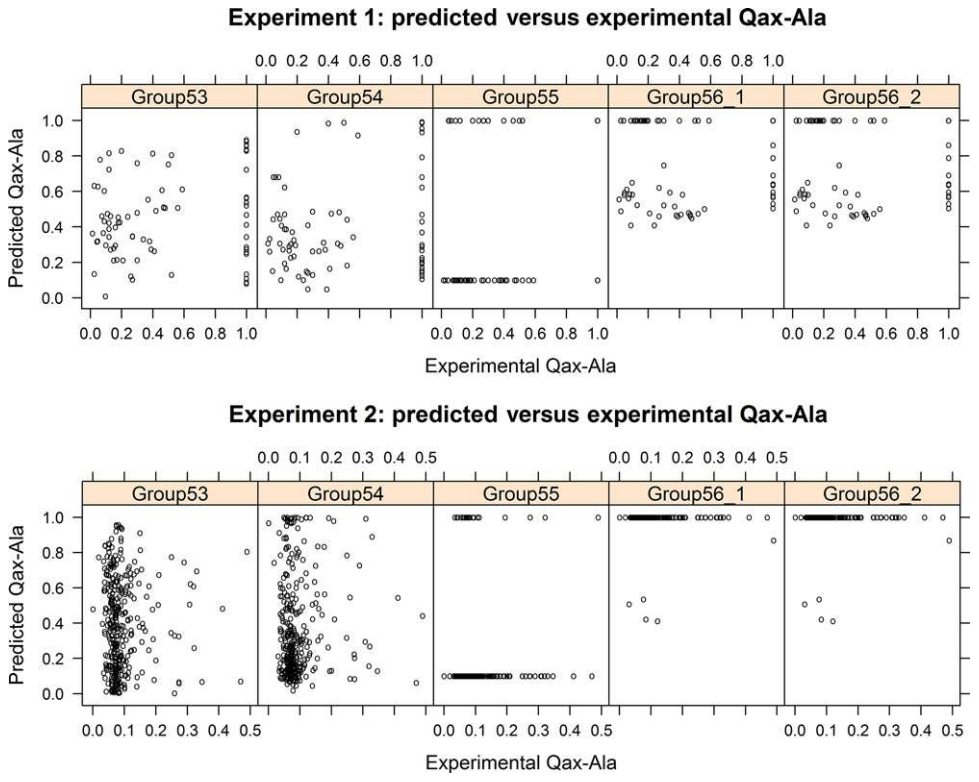


FIGURE 5 Scatter plot of the experimental $Q_{ax\text{-}Ala}$ versus the predicted $Q_{ax\text{-}Ala}$ values

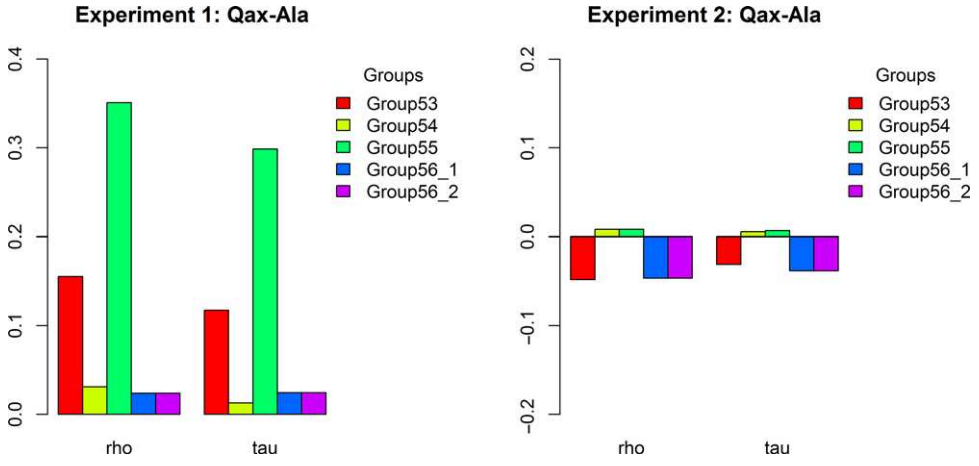


FIGURE 6 Correlations represented by Spearman's ρ and Kendall's τ between the predicted and experimental $Q_{ax\text{-}Ala}$ values of two experiments

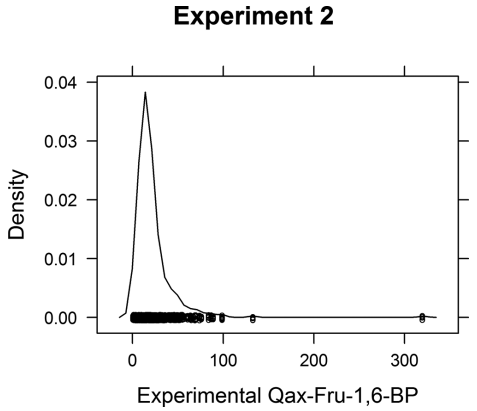


FIGURE 7 Kernel density estimate of experimental $Q_{\text{ax-Fru-1,6-BP}}$ from experiment 2

We calculated Spearman's ρ and Kendall's τ coefficients as nonparametric tests of the correlation of the predictions with the experiments, since the data and predicted values are not unimodal or normally distributed. Only group 55 in experiment 1 achieves a favorable correlation, with $\rho = 0.351$ and $\tau = 0.299$ with P values of 0.002 for both (Fig. 6). All of the other P values are in the range of 0.17–0.88, which implies there is no correlation between the predicted and experimental $Q_{\text{ax-Ala}}$ values. If we treat the experimental $Q_{\text{ax-Ala}}$ values as binary for experiment 1 (Fig. 4, left), we can calculate binary assessment measures such as TPR, TNR, and so on. We did this for group 55, which provided binary prediction values (0.1 and 1.0) with the following results (where positive indicates $Q_{\text{ax-Ala}} = 1$): TPR = 17/23 = 0.739; TNR = 39/55 = 0.709; BACC = 0.724. This is better than random and explains the positive correlation coefficients.

The results for experiment 2 are negatively correlated for three of the groups, and only very weak positive correlations were achieved by groups 54 and 55 (Fig. 6, right). The P values are in the range of 0.38–0.88.

3.4 | Prediction of allosteric activation of Fru-1,6-BP ($Q_{\text{ax-Fru-1,6-BP}}$)

Participants were asked to predict the allosteric effect of Fru-1,6-BP binding to L-PYK for the mutants created in experiment 2 and were told that the maximum value in the experiments was 320. The wild-type protein has a $Q_{\text{ax-Fru-1,6-BP}}$ value of 14.2. The density estimate of experimental $Q_{\text{ax-Fru-1,6-BP}}$ values is shown in Figure 7, showing that the vast majority of mutants had values between 0 and 60. The scatter plots of the predicted $Q_{\text{ax-Fru-1,6-BP}}$ versus experimental $Q_{\text{ax-Fru-1,6-BP}}$ show that groups 53 and 54 provided real values over the full range of the experimental values and group 55 provided discrete values (1, 50, 250, and 320), whereas group 56 provided an approximate wild-type value of 15.3 for most of the mutants and other values for 18 mutants in the range from 1 to 28.3 (Fig. 8).

We calculated Spearman's ρ and Kendall's τ to evaluate the correlations between predicted and experimental $Q_{\text{ax-Fru-1,6-BP}}$ values (Fig. 9). Only group 55 has positive correlations, both very marginal (both ρ and $\tau \sim 0.05$, with P value of 0.2). All others have negative correlations, especially for group 53 and 54. The P values of group 53 are 7.5E-05 for ρ and 8.98E-05 for τ, and the P values of group 54 are 0.0003 for both ρ and τ.

4 | DISCUSSION

We may summarize the results of the CAGI experiment on L-PYK as follows. Groups 53 and 56 had good predictions of the L-PYK enzyme activity in experiments 1 and 2 as measured by BACC (group 53) and ACC (group 56). In these cases, the results were better than that achieved by PolyPhen-2. Group 54 had good predictions only if we set a new cutoff for binary enzyme activity from their real-valued results in both experiments 1 and 2.

For the prediction of allosteric effects of alanine and fructose, groups 55 and 53 had positive correlations for the $Q_{\text{ax-Ala}}$ challenge in experiment 1, but only group 55 had a statistically significant positive correlation. No group had statistically significant, positive correlations for their predictions of $Q_{\text{ax-Ala}}$ or $Q_{\text{ax-Fru-1,6-BP}}$ in experiment 2.

At the conclusion of this experiment, we are left to contemplate why the overall success of predicting allosteric effects was underwhelming. This consideration is particularly valuable given the indications of success of computational approaches reported in the literature. As noted, the only statistically significant result for predicting allosteric data was for group 55 on the $Q_{\text{ax-Ala}}$ challenge in experiment 1. This group used a very simple model that considered the distance each wild-type residue was from bound Ala (as modeled from the structure of human pyruvate kinase M2) and the severity of the mutation from wild type (as determined by scores from a substitution matrix). It is likely that they correctly predicted many of the mutations that abrogated Ala binding altogether ($Q_{\text{ax-Ala}} = 1$), rather than quantitatively predicting the effect of the mutations on the diverse values of $Q_{\text{ax-Ala}}$ of the remaining mutations ($Q_{\text{ax-Ala}} < 1$). It is not likely that their distance-based method would extend readily to the general problem of predicting allosteric effects, especially for residues not in or near the binding site. The results for experiment 2, where mutations were made throughout the protein, confirm this.

It is also clear from the experiment that methods that predominantly used evolutionary considerations (groups 53 and 54) were not able to predict the effects of mutation on allosteric behavior. Group 53 used the evolutionary action of each mutation, a number that can be calculated from phylogenetic sequence analysis (Katsonis & Lichtarge, 2014). Group 54 used covariation of amino acids in pairs of positions within a multiple sequence alignment of homologues of L-PYK (Jones et al., 2015).

Group 56 calculated the binding affinity of each effector to each mutant with docking calculations (Shin et al., 2013), and made the assumption that Q_{ax} was directly proportional to these values. In fact, $Q_{\text{ax}} = K_{\text{ix}}/K_{\text{ix/a}}$ where K_{ix} is the equilibrium dissociation constant of the effector X and $K_{\text{ix/a}}$ is the equilibrium dissociation constant of the effector X when the substrate A is bound. The approximation is not unreasonable given the experimental data from experiment 2: the Pearson and Kendall correlation coefficients between the experimental values of Q_{ax} and K_{ix} for alanine are 0.73 and 0.59, respectively, and for Fru-1,6-BP they are 0.80 and 0.64, respectively (all P values $< 1.0 \times 10^{-15}$).

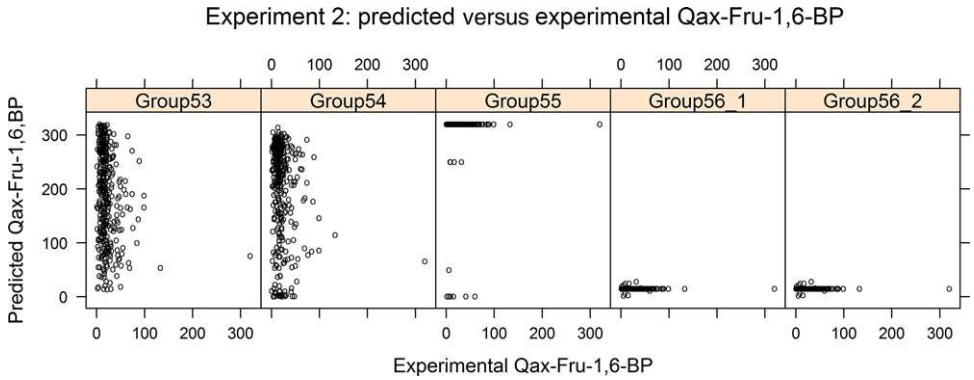


FIGURE 8 Scatter plot of the predicted versus experimental $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$ values from experiment 2

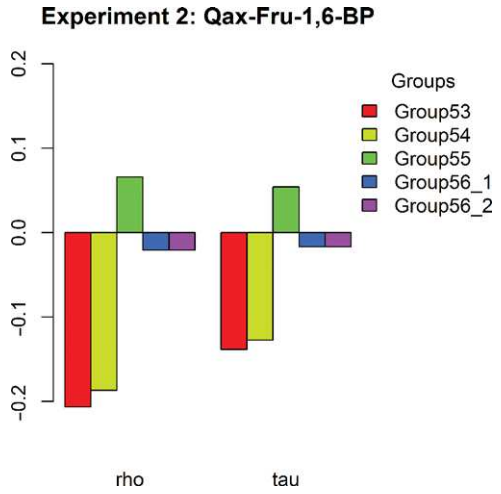


FIGURE 9 Correlations represented by Spearman's ρ and Kendall's τ between the predicted and experimental $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$ values in experiment 2

Group 56 only performed docking calculations to mutations in the binding sites of alanine and Fru-1,6-BP, and submitted values for all other positions of 1.0 for $Q_{ax\text{-}Ala}$ (no inhibition of PEP-binding by Ala) and 15.3 for $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$ (the experimental value). This resulted in only eight mutations with $Q_{ax\text{-}Ala}$ not equal to 1.0, only five of which had experimental values available. If we restrict the calculation of correlation coefficients to these five values, the *P* values for the Spearman and Kendall correlation coefficients are greater than 0.8, and the values of rho and tau are 0.1 and 0, respectively. For $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$, group 56 produced values for 17 mutations adjacent to the Fru-1,6-BP site, only 11 of which had enough enzyme activity to measure $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$. The correlation coefficients with $Q_{ax\text{-}Fru\text{-}1,6\text{-}BP}$ were both ~0.2 with *P* values of ~0.5. Unless docking calculations are able to discern changes in binding affinity of the effector (in the presence or absence of the substrate) for sites far from their binding sites, it is not possible to determine whether such calculations provide valuable information on allosteric behavior.

It is clear from the quality of predictions in this study that additional approaches are needed. Many of the methods reported in the literature involve molecular dynamics simulations that are very computationally intensive (Blacklock & Verkhivker, 2014; Hertig et al., 2016; Weinkam et al., 2012). Several simulations of other forms of pyruvate kinase (Naithani et al., 2015) and mutants thereof have been performed (Kalaiarasan et al., 2015). However, whether such methods could be used in a predictive fashion has yet to be determined. The current data set could be used to benchmark such methods, if a sufficient number of mutants can be simulated.

Allosteric regulation is sometimes presented as a Rube Goldberg-type mechanism initiated by the effector associating with the enzyme/protein (binding causes change A; change A causes change B; change B causes change C, etc.). However, the definition for allostery based on an energy cycle (Fig. 1) implies that allostery is an equilibrium mechanism (Carlson & Fenton, 2016). As such, the allosteric mechanism would be a comparison of changes in the fully equilibrated enzyme forms represented in Figure 1 and not a Rube Goldberg mechanism that would be associated with a kinetics mechanism. Calculations of this sort remain a challenge for computational approaches to predicting the effects of mutations on allosteric regulation.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249.

Blacklock, K., & Verkhivker, G. M. (2014). Computational modeling of allosteric regulation in the hsp90 chaperones: A statistical ensemble analysis of protein structure networks and allosteric communications. *PLoS Computational Biology*, *10*(6), e1003679.

Carlson, G. M., & Fenton, A. W. (2016). What mutagenesis can and cannot reveal about allostery. *Biophysical Journal*, *110*(9), 1912–1923.

Collier, G., & Ortiz, V. (2013). Emerging computational approaches for the study of protein allostery. *Archives of Biochemistry and Biophysics*, *538*(1), 6–15.

Feher, V. A., Durrant, J. D., Van Wart, A. T., & Amaro, R. E. (2014). Computational approaches to mapping allosteric pathways. *Current Opinion in Structural Biology*, *25*, 98–103.

Fenton, A. W. (2008). Allostery: An illustrated definition for the 'second secret of life'. *Trends in Biochemical Sciences*, *33*(9), 420–425.

Fenton, A. W. (Ed.). (2012). *Allostery: Methods and protocols*). New York: Humana Press: Springer Science.

Fenton, A. W., & Alontaga, A. Y. (2009). The impact of ions on allosteric functions in human liver pyruvate kinase. *Methods in Enzymology*, *466*, 83–107.

Fenton, A. W., & Hutchinson, M. (2009). The pH dependence of the allosteric response of human liver pyruvate kinase to fructose-1,6-bisphosphate, ATP, and alanine. *Archives of Biochemistry and Biophysics*, *484*, 16–23.

Fenton, A. W., Johnson, T. A., & Holyoak, T. (2010). The pyruvate kinase model system, a cautionary tale for the use of osmolyte perturbations to support conformational equilibria in allostery. *Protein Science*, *19*, 1796–1800.

Hertig, S., Latorraca, N. R., & Dror, R. O. (2016). Revealing atomic-level mechanisms of protein allostery with molecular dynamics simulations. *PLOS Computational Biology*, *12*(6), e1004746.

Holyoak, T., Zhang, B., Deng, J., Tang, Q., Prasannan, C. B., & Fenton, A. W. (2013). Energetic coupling between an oxidizable cysteine and the phosphorylatable N-terminus of human liver pyruvate kinase. *Biochemistry*, *52*(3), 466–476.

Ishwar, A., Tang, Q., & Fenton, A. W. (2015). Distinguishing the interactions in the fructose 1,6-bisphosphate binding site of human liver pyruvate kinase that contribute to allostery. *Biochemistry*, *54*(7), 1516–1524.

Jones, D. T., Singh, T., Kosciolek, T., & Tetchner, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, *31*(7), 999–1006.

Kalaiarasan, P., Kumar, B., Chopra, R., Gupta, V., Subbarao, N., & Bamezai, R. N. (2015). In silico screening, genotyping, molecular dynamics simulation and activity studies of SNPs in pyruvate kinase M2. *PLOS ONE*, *10*(3), e0120469.

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Research*, *24*(12), 2050–2058.

Lensink, M. F., Velankar, S., & Wodak, S. J. (2017). Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition. *Proteins*, *85*, 359–377.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, *405*(2), 442–451.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins*, *84* (Suppl 1), 4–14.

Naithani, A., Taylor, P., Erman, B., & Walkinshaw, M. D. (2015). A molecular dynamics study of allosteric transitions in Leishmania mexicana pyruvate kinase. *Biophysical Journal*, *109*(6), 1149–1156.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reinhart, G. D. (1983). The determination of thermodynamic allosteric parameters of an enzyme undergoing steady-state turnover. *Archives of Biochemistry and Biophysics*, *224*(1), 389–401.

Reinhart, G. D. (1988). Linked-function origins of cooperativity in a symmetrical dimer. *Biophysical Chemistry*, *30*(2), 159–172.

Reinhart, G. D. (2004). Quantitative analysis and interpretation of allosteric behavior. *Methods in Enzymology*, *380*, 187–203.

Shin, W. H., Kim, J. K., Kim, D. S., & Seok, C. (2013). GalaxyDock2: Protein-ligand docking using beta-complex and global optimization. *Journal of Computational Chemistry*, *34*(30), 2647–2656.

Weber, G. (1972). Ligand binding and internal equilibria in proteins. *Biochemistry*, *11*(5), 864–878.

Wei, Q., & Dunbrack, R. L. Jr. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLOS ONE*, *8*(7), e67863.

Weinkam, P., Pons, J., & Sali, A. (2012). Structure-based model of allostery predicts coupling between distant sites. *Proceedings of the National Academy of Sciences*, *109*(13), 4875–4880.

Williams, R., Holyoak, T., McDonald, G., Gui, C., & Fenton, A. W. (2006). Differentiating a ligand's chemical requirements for allosteric interactions from those for protein binding. Phenylalanine inhibition of pyruvate kinase. *Biochemistry*, *45*(17), 5421–5429.

Zhong, W., Morgan, H. P., McNae, I. W., Michels, P. A., Fothergill-Gilmore, L. A., & Walkinshaw, M. D. (2013). 'In crystallo' substrate binding triggers major domain movements and reveals magnesium as a co-activator of Trypanosoma brucei pyruvate kinase. *Acta Crystallographica Section D: Biological Crystallography*, *69*(Pt 9), 1768–1779.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Xu, Q, Tang, Q, Katsonis, P, Lichtarge, O, et al. Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. *Human Mutation*. 2017;38:1123–1131. https://doi.org/10.1002/humu.23222

Received: 5 April 2017 | Revised: 12 June 2017 | Accepted: 12 June 2017

DOI: 10.1002/humu.23280

SPECIAL ARTICLE

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges

Roxana Daneshjou[1] | Yanran Wang[2] | Yana Bromberg[2] | Samuele Bovo[3] | Pier L Martelli[3] | Giulia Babbi[3] | Pietro Di Lena[4] | Rita Casadio[3,5] | Matthew Edwards[6] | David Gifford[6] | David T Jones[7] | Laksshman Sundaram[8] | Rajendra Rana Bhat[8] | Xiaolin Li[8] | Lipika R. Pal[9] | Kunal Kundu[9,10] | Yizhou Yin[9,10] | John Moult[9,11] | Yuxiang Jiang[12] | Vikas Pejaver[12,13] | Kymberleigh A. Pagel[12] | Biao Li[14] | Sean D. Mooney[13] | Predrag Radivojac[12] | Sohela Shah[15] | Marco Carraro[16] | Alessandra Gasparini[16,17] | Emanuela Leonardi[17] | Manuel Giollo[16,18] | Carlo Ferrari[18] | Silvio C E Tosatto[16,19] | Eran Bachar[20] | Johnathan R. Azaria[20] | Yanay Ofran[20] | Ron Unger[20] | Abhishek Niroula[21] | Mauno Vihinen[21] | Billy Chang[22] | Maggie H Wang[22,23] | Andre Franke[24] | Britt-Sabina Petersen[24] | Mehdi Pirooznia[25] | Peter Zandi[26] | Richard McCombie[27] | James B. Potash[28] | Russ B. Altman[1] | Teri E. Klein[1] | Roger A. Hoskins[29] | Susanna Repo[29] | Steven E. Brenner[29] | Alexander A. Morgan[30]

[1]Department of Genetics, Stanford School of Medicine, Stanford, California

[2]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, New Jersey

[3]Biocomputing Group, BiGeA/CIG, "Luigi Galvani" Interdepartmental Center for Integrated Studies of Bioinformatics, Biophysics, and Biocomplexity, University of Bologna, Bologna, Italy

[4]Biocomputing Group/Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

[5]"Giorgio Prodi" Interdepartmental Center for Cancer Research, University of Bologna, Bologna, Italy

[6]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts

[7]Bioinformatics Group, Department of Computer Science, University College London, London, United Kingdom

[8]Large-scale Intelligent Systems Laboratory, NSF Center for Big Learning, University of Florida, Gainesville, Florida

[9]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland

[10]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, Maryland

[11]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland

[12]Department of Computer Science and Informatics, Indiana University, Bloomington, Indiana

[13]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, Washington

[14]Gilead Sciences, Foster City, California

[15]Qiagen Bioinformatics, Redwood City, California

[16]Department of Biomedical Science, University of Padova, Padova, Italy

[17]Department of Woman and Child Health, University of Padova, Padova, Italy

[18]Department of Information Engineering, University of Padova, Padova, Italy

[19]CNR Institute of Neuroscience, Padova, Italy

[20]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, Israel

[21]Protein Structure and Bioinformatics Group, Department of Experimental Medical Science, Lund University, Lund, Sweden

[22]Division of Biostatistics and Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, Chinese University of Hong Kong, Shatin, N.T., Hong Kong

[23]CUHK Shenzhen Research Institute, Shenzhen, China

[24]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Kiel, Germany

[25]Department of Psychiatry, The Johns Hopkins University School of Medicine, Baltimore, Maryland

[26]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

[27]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

[28]Department of Psychiatry, University of Iowa, Iowa City, Iowa

[29]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, California

[30]Stanford School of Medicine, Stanford, California

Correspondence
Roxana Daneshjou, Department of Genetics, Stanford School of Medicine, Stanford, California.
Email: roxanad@stanford.edu

Contract grant sponsors: NIH (U41 HG007446, R13 HG006650, U01 GM115486, U24 MH068457); Informatics Research Starter grant from the PhRMA Foundation; National Science Foundation of China (81473035, 31401124); Israeli Science Foundation (772/13).

For the CAGI Special Issue

Abstract

Precision medicine aims to predict a patient's disease risk and best therapeutic options by using that individual's genetic sequencing data. The Critical Assessment of Genome Interpretation (CAGI) is a community experiment consisting of genotype–phenotype prediction challenges; participants build models, undergo assessment, and share key findings. For CAGI 4, three challenges involved using exome-sequencing data: Crohn's disease, bipolar disorder, and warfarin dosing. Previous CAGI challenges included prior versions of the Crohn's disease challenge. Here, we discuss the range of techniques used for phenotype prediction as well as the methods used for assessing predictive models. Additionally, we outline some of the difficulties associated with making predictions and evaluating them. The lessons learned from the exome challenges can be applied to both research and clinical efforts to improve phenotype prediction from genotype. In addition, these challenges serve as a vehicle for sharing clinical and research exome data in a secure manner with scientists who have a broad range of expertise, contributing to a collaborative effort to advance our understanding of genotype–phenotype relationships.

KEYWORDS

bipolar disorder, Crohn's disease, exomes, machine learning, phenotype prediction, warfarin

1 | INTRODUCTION

Precision medicine aims to use a patient's genomic and clinical data to make predictions about medically relevant phenotypes such as disease risk or drug efficacy (Ashley, 2015; Ashley et al., 2010).

The Critical Assessment of Genome Interpretation (CAGI) is a community experiment, which aims to advance methods for phenotype prediction from genotypes through a series of "challenges" with real data (CAGI, 2011). Exome-sequencing data, which captures exons and nearby flanking regulatory regions, is already being used clinically to solve medical mysteries with well-defined symptoms (Brown & Meloche, 2016). However, in order to advance precision medicine, clinicians and scientists will need to be able to make inferences about disease risk or drug efficacy from genetic data. Interpretation of genetic data is one of the major difficulties in the implementation of precision medicine (Fernald, Capriotti, Daneshjou, Karczewski, & Altman, 2011).

CAGI is an example of the Common Task Framework, a phrase coined by Mark Liberman to describe the approach of using shared training and testing datasets and evaluation metrics to advance machine learning (Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine, & Schwalbe, 2016; Donoho, 2015). The Common Task Framework has been called the "secret sauce" behind the recent successes in machine learning (Donoho, 2015). Starting with common task challenges in the 1980s for machine translation, this approach has led to significant gains in speech recognition and dialog systems, protein structure prediction, biomedical natural language processing, autonomous vehicles, and collaborative filtering for consumer preferences (Bell & Koren, 2007; Morgan et al., 2008; Moult, Fidelis, Kryshtafovych, Schwede, & Tramontano, 2014; Thrun et al., 2006; Walker et al., 2001). Through this same approach, CAGI aims to push forward the field of precision medicine.

At CAGI 4 held in 2016, three challenges involved making predictions using exome sequence data: a Crohn's disease challenge, a bipolar disorder challenge, and a warfarin dosing challenge. These challenges represent the spectrum of phenotypes seen in clinical practice. Bipolar disorder and Crohn's disease are discrete phenotypes, with the former being a clinical diagnosis (based on meeting clinical criteria) and the latter a pathological diagnosis (based on biopsies). Therapeutic warfarin dose, on the other hand, is a continuous phenotype.

The Crohn's disease challenge has been a part of previous CAGI iterations, whereas the bipolar disorder and warfarin dosing challenges debuted during CAGI 4. We will describe the nature of each challenge in greater detail. The number of groups participating in each challenge can be found in Table 1.

TABLE 1 The number of predictors and predictions for each CAGI challenge

Challenge	Number of predictors	Number of predictions
Crohn's disease exomes challenge	CAGI 2 – 10 groups	CAGI 2 – 33 predictions
	CAGI 3 – 14 groups	CAGI 3 – 58 (+3 late) predictions
	CAGI 4 – 14 groups	CAGI 4 – 46 predictions
Bipolar exomes challenge	CAGI 4 – 9 groups	CAGI 4 – 29 predictions
Warfarin exomes challenge	CAGI 4 – 3 groups	CAGI 4– 9 predictions

1.1 | Crohn's disease challenge

Crohn's disease is a chronic inflammatory bowel disease marked by transmural inflammation of the gastrointestinal tract that can occur anywhere from the mouth to the rectum (Cho, 2008). Symptoms include pain and debilitating diarrhea, which can lead to malnutrition (Cho, 2008). Monozygotic twin studies have shown a concordance of 40%–50%, and genome-wide association studies have identified genetic risk loci (Cho, 2008; Halfvarson, Bodin, Tysk, Lindberg, & Jarnerot, 2003). Age of onset is typically between 20 and 40 years old, but early age of onset, such as in early childhood, is associated with more severe disease features (Uhlig et al., 2014).

The 2011 (CAGI 2) dataset has 56 exomes (42 cases, 14 controls), all of German ancestry (Ellinghaus et al., 2013). The 2013 (CAGI 3) dataset has 66 exomes (51 cases, 15 controls). Though these samples were also of German ancestry, cases were selected from pedigrees of German families with multiple occurrences of Crohn's disease. As such, some of these cases were related. For the most part, the samples sequenced as controls were unrelated healthy individuals; the exceptions to this were the unaffected parents of three cases and the unaffected twin of one case. The most recent challenge, CAGI 4 in 2016, was to identify cases from controls in 111 unrelated German ancestry exomes (64 cases, 47 controls). For CAGI 4, submitting groups were allowed to use the data from the Crohn's disease CAGI challenges of 2011 and 2013. In all iterations of the challenge, groups were asked to report a probability of Crohn's disease (between 0 and 1) for each individual and a standard deviation representing their confidence in that prediction. For the most recent Crohn's disease evaluation, teams were also asked to predict whether age of onset was greater or less than 10 years of age; an age cutoff selected by CAGI based on the literature (Uhlig et al., 2014). Additional details of the challenges can be found in Supp. Exhibit 1.

1.2 | Bipolar disorder challenge

Bipolar disorder is a mood disorder marked by elevated mood (mania or hypomania) and depressed mood that disrupts an individual's ability to function (Craddock & Sklar, 2013). In the general population, the lifetime risk of bipolar disorder is 0.5%–1% (Craddock & Jones, 1999). However, bipolar disorder has a high component of heritability, with studies demonstrating a 40%–70% monozygotic twin concordance (Craddock & Jones, 1999). In this CAGI 4 challenge, 1,000 exomes of unrelated bipolar disorder cases and age/ancestry-matched controls of Northern European ancestry were provided. Five-hundred exomes were used as the training set and 500 exomes were used for the prediction set (Monson et al., 2017). Groups were asked to report a probability of bipolar disorder (between 0 and 1) for each individual and a standard deviation representing their confidence in that prediction. Additional information on the challenge can be found in Supp. Exhibit 2.

1.3 | Warfarin dosing challenge

Warfarin is an anticoagulant with over 30 million prescriptions written in 2011 (IMS Institute of Healthcare Informatics, 2012). Warfarin remains a clinical staple despite the introduction of novel oral anticoagulants because of multiple factors—warfarin's lower cost, longer half-life, and clinical indications for which novel oral anticoagulants have not yet been approved (Bauer, 2011). However, warfarin is responsible for one-third of hospitalizations due to adverse drug events because of its narrow therapeutic index and high interindividual dose variability (Budnitz, Lovegrove, Shehab, & Richards, 2011). Both clinical and genetic factors affect the therapeutic dose of warfarin (Klein et al., 2009). For this challenge, participants were provided with exomes of African Americans on tail ends of the warfarin dose distribution ($\leq$35 mg or $\geq$49 mg) (Daneshjou et al., 2014). Clinical covariates were provided for all exomes. The training set consisted of 50 exomes, and participants submitted dose predictions with standard deviations on 53 test set exomes. Additional details of the challenge can be found in Supp. Exhibit 3.

2 | METHODS

2.1 | Data distribution

Data were distributed to the participants who consented to the CAGI data use agreement. Data providers worked with their home institution to ensure adherence with local privacy regulations and predicting groups agreed not to share the anonymized data. Data were provided as described above, with genetic variant data shared in the VCF file format.

2.2 | Predicting phenotypes

Participants required to return a simple text file with appropriate predicted values (such as disease status and confidence in prediction) for each sample. They were also provided with a validation script to check their output formatting. Participants were asked to submit a methods description for each submission. The prediction results from selected groups that submitted predictions and methods descriptions were presented at the CAGI meeting. Additionally, the ground truth data and scoring scripts used to perform the evaluation were shared with participants.

2.3 | Data quality

For the Crohn's disease and bipolar disorder exome challenges, biases in the data were assessed using principal component analysis and clustering after pruning for linkage disequilibrium using plink (Purcell et al., 2007).

For the warfarin challenge, data had previously undergone QC using ancestry informative markers to confirm self-reported ancestry and identity by state (IBS) analysis in order to ensure that samples were not related, as previously described (Daneshjou et al., 2014).

2.4 | Assessing discrete phenotypes (Crohn's disease and bipolar disorder)

A simple accuracy of prediction per sample score, such as derivable from setting a threshold for prediction (such as 0.5), although tantalizing in its simplicity neither supports the goals of CAGI nor is it representative of a likely clinically relevant scenario for prediction. Because the genetic datasets from CAGI are drawn from case-control studies, as well as pedigree studies in families with a strong burden of disease, it does not represent a random sampling of the population. Requiring a fixed threshold for evaluation and reporting a basic accuracy score of prediction in such a dataset would obscure interpretation. Also, using this as a figure of merit for ranking encourages participants to optimize their system predictions for the anticipated case/control distribution instead of focusing on features that selectively prioritize and rank disease likelihood in the absence of that calibration. The use of receiver operator characteristics (ROC) curves for genomic test evaluation has been previously investigated by Wray, Yang, Goddard, and Visscher (2010).

The ROC offers many advantages for evaluating a test, and is often used to characterize clinical tests. The shape of a ROC curve can help differentiate between highly sensitive tests, which could rule in a possible diagnosis, and highly specific tests that could rule out a diagnosis. The prediction of Crohn's disease status from sequencing data might be used in either of those situations depending on clinical presentation, risk factors, or stage of patient evaluation. Additionally, ROC curves allow easy selection of a classification threshold (based on selecting a position on the curve). Based on the selected threshold, a positive or negative likelihood ratio can be derived and applied in standard evidence-based techniques of patient diagnosis, which rely on a Bayesian framework that takes into account the pretest probabilities and the characteristics of a given test depending on the threshold chosen for prediction (Fagan, 1975).

We evaluated the robustness of the prediction accuracy when making predictions on different subsamples of exomes and assessed the confidence intervals reported by the participants.

To capture confidence intervals on the predictions, multiple samples with replacement were drawn. Each prediction was then modified by adding a random amount drawn from a normal distribution with a mean of zero and a standard deviation equivalent to the standard deviation reported for the original prediction. If no confidence interval was reported for the original prediction, the standard deviation was taken to be zero. If a prediction for a particular exome was missing, the prediction score for that sample was set to the mean reported prediction value in that submission. In order to compare submissions by a single figure of merit, the average area under the ROC curves from the bootstrap sampling was used, accompanied by the bootstrapped confidence interval around that area under the curve, to estimate the robustness of differences between prediction performances. The evaluation scripts were provided to all participants.

A cross-validated logistic regression-based metaclassifier using lasso regularization was also trained on the submissions as features for CAGI 4 Crohn's disease and CAGI 4 bipolar disorder. This step allowed us to assess whether combining the features selected across the different groups would improve prediction over a single method. If a metaclassifier could perform better than any single method, then a combination of methods might lead to meaningfully better performance.

2.5 | Assessing continuous phenotypes (therapeutic warfarin dose)

For the warfarin exomes challenge, several metrics of assessment were used. Each participant provided a predicted therapeutic dose of warfarin for each individual as well as a standard deviation for that prediction.

To look at the amount of variation in dose explained by the predicted doses, we used linear regression with the linear model function (lm) in the R statistical package (v 2.15.3). We evaluated each method using the R^2 and the sum of squared errors. Additionally, we compared each prediction against one of the best performing warfarin-predictive algorithms, the International Warfarin Pharmacogenetic Consortium (IWPC) algorithm (Klein et al., 2009).

To assess, on average, how many participant-provided standard deviations the predicted dose was from the actual dose, we used a mean of the absolute value of the *z* score for each prediction, as seen in Equation (1). Here, dose_actual is the known therapeutic dose of warfarin for each individual i, whereas dose_predicted is the therapeutic dose predicted by that group for that individual. SD_predicted is the standard deviation for each individual's predicted dose, as provided by the participant's prediction method. The number of individuals is *n*.

$$\frac{\sum_{i=1}^{n}\left|\frac{\text{dose\_actual}_i - \text{dose\_predicted}_i}{\text{SD\_predicted}_i}\right|}{n} \quad (1)$$

To assess the range of the each prediction's standard deviation compared with the predicted dose, we calculated the mean of the coefficient of variation, which was the mean of the standard deviation for each prediction divided by the predicted dose, as seen in Equation (2).

$$\frac{\sum_{i=1}^{n}\frac{\text{SD\_predicted}_i}{\text{dose\_predicted}_i}}{n} \quad (2)$$

We also evaluated the mean absolute value of the *z* score multiplied by the mean coefficient of variation for each method. This value allowed us to assess the mean *z* scores with a penalization for mean *z* scores whose values were closer to 0 because of larger standard deviations.

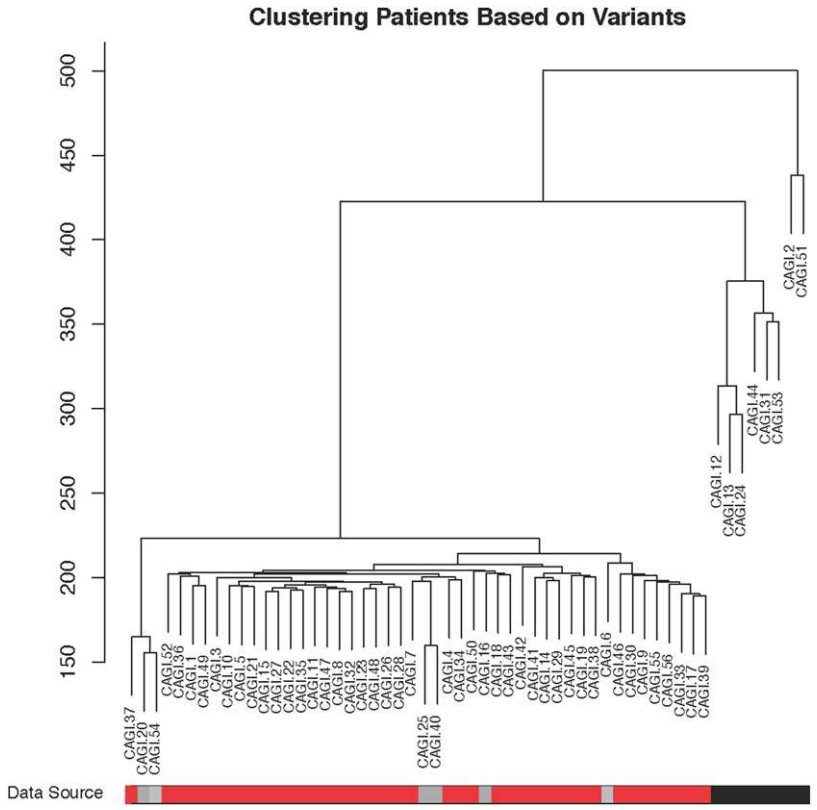


FIGURE 1 Clustering of patients from the CAGI 2 Crohn's disease challenge. The black and gray bars at the bottom represent the controls; the red represents the cases. Many of the controls cluster together, likely due to batch effects. For instance, the controls represented in black were sequenced separately from the gray controls and the cases

We calculated rho and *P* values using the spearman rank correlation between (1) each group's predicted warfarin doses and the actual therapeutic doses across individuals and (2) each group's predicted warfarin doses and the IWPC-predicted doses across individuals. These calculations were made with the spearmanr command from the stat package in scipy (python v 2.7.5).

3 | RESULTS

With each year, CAGI has expanded the number of challenges and participants. Table 1 displays the number of participants and predictions for each CAGI challenge.

3.1 | Crohn's disease exomes challenge (CAGI 2–4)

For the 2011 Crohn's disease (CAGI 2) challenge, during the assessment phase, a substantial batch effect was discovered in the data as a side effect of sample preparation and sequencing (Fig. 1). Overall, the control samples that clustered separately due to this batch effect had fewer variants reported that did not match the reference genome. The participants were not aware of this batch effect; their methods were not designed to exploit it. However, this raises the possibility that techniques that used a very large list of genes were more likely to correctly identify case samples as coming from individuals with Crohn's disease. Indeed, many different methods did better than random based on AUC, with a maximum AUC of 0.94, and in general approaches that favored a large list of potentially Crohn's disease-related genes and gave more weight to rarer variants did the best. A full description of all methods used by the participants can be found in Supp. Exhibit 1:CAGI 2. Supp. File 1 shows comparative results of the CAGI 2 Crohn's disease challenge predictive methods. It is certainly biologically plausible that increased burden of variation in a large number of Crohn's disease-related genes leads to increased likelihood of disease; however, it is also possible that there was systematic over-reporting of variation as a batch effect. Therefore, it was important to re-evaluate with more data.

In the 2013 CAGI 3, a much greater effort was made to carefully collect and prepare samples in a completely consistent way. In this instance, case samples were collected from German families with a particularly high burden of Crohn's disease (two or more affected family members), including a pair of twins discordant for the disease, and another pair of twins concordant with the disease. Additional healthy controls were drawn from the unaffected German general population. During the 2013 CAGI 3, there was once again a substantial difference in clustering between cases and controls, but in this dataset there was substantially more homogeneity in the cases. Individuals from different case families clustered much more closely with each other than with unrelated controls (Fig. 2). This prompted two possible hypotheses. The first is that there might be a hidden founder effect, and these families with a high burden of disease may all actually be closely related. The second is that reduced heterogeneity and perhaps

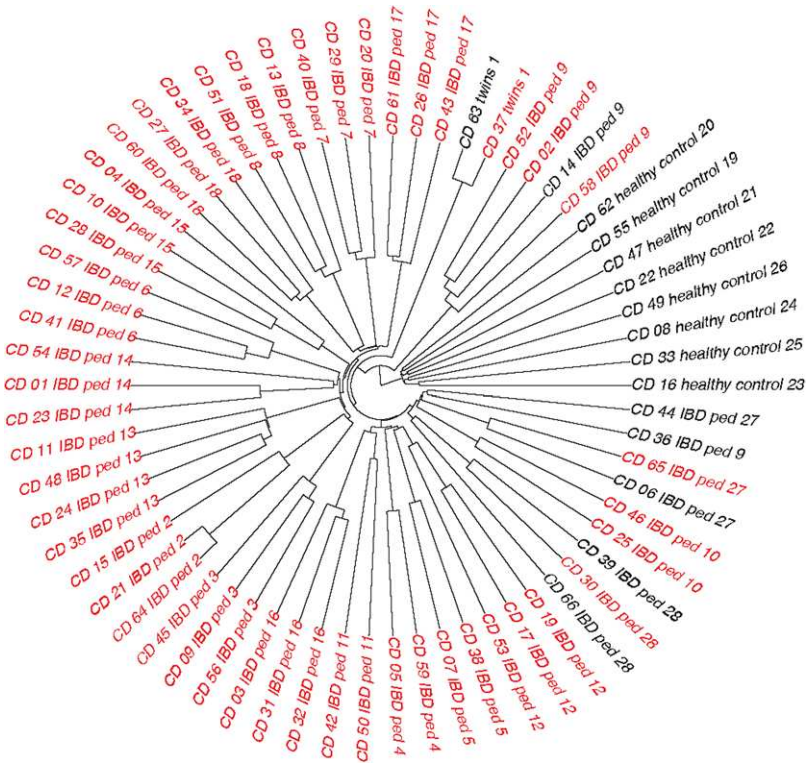


FIGURE 2 Clustering of samples for CAGI 3 Crohn's disease challenge. Black represents controls, whereas red represents cases. This dataset included healthy family members of cases as well as random controls. Samples with a "ped" designation in the sample name came from a pedigree; samples that share the same "ped" number came from the same pedigree

increased ancestor consanguinity may contribute to increased risk of Crohn's disease in these families with a high burden. Either one alone or a mixture of both possibilities is biologically plausible. In this instantiation of CAGI, groups that simply did some version of partitioning the test datasets based on hierarchical clustering did quite well, and the top performing methods had an AUC of 0.87. Once again, all of these methods were implemented without awareness of the bias in the data. A full description of all methods used by the participants can be found in Supp. Exhibit 1:CAGI 3. Supp. File 2 shows comparative results of the CAGI 3 Crohn's disease challenge.

In CAGI 4, 111 exomes were derived from a mix of 64 Crohn's disease patients, with a skew toward early onset of disease, and 47 healthy controls, all taken from individuals of German descent. With this data, the simple separation of cases and controls based on genetic variants was not present (Fig. 3), suggesting the problems with batch effects and sampling bias were no longer present; the only noticeable structure indicated the possibility of a few related samples, as seen in the PCA and IBD plots shown in Supp. Figures S1 and S2. Correspondingly, the peak performance dropped from previous CAGI iterations down to an AUC of 0.72. However, given the elimination of biases in the data, this incarnation of the Crohn's disease challenge is likely the best reflection of how the prediction methods perform. A metaclassifier created by the assessment team using all submitted methods for this challenge, as shown in Supp. Figure S3, had an AUC of 0.78, a small improvement over the top method. The distribution of AUCs across methods is shown in Figure 4. A full description of all methods used by the participants can be found in Supp. Exhibit 1:CAGI 4. Supp. File 3 shows comparative results of the CAGI 4 Crohn's disease challenge.

The top approach in CAGI 4 used a compiled list of genes and genomic regions associated with Crohn's disease from prior studies, used imputation to evaluate risk contribution from known regions associated with Crohn's disease but not covered by exome sequencing, and used the Welcome Trust Case Control Consortium (WTCCC) Crohn's disease genotyping array data to train a disease classifier to score relative risk for each sample.

Across participants, numerous methods were used for selecting the covariates, highlighting the many different approaches to building a Crohn's disease classifier. Similar to the top approach, many groups used variants previously found to be associated in genome-wide association studies; the NHGRI catalog was a popular choice to identify these associated variants (Welter et al., 2014). Other approaches relied on gene lists of associated and "predicted" Crohn's disease genes to select variants of interest. To create the "predicted" list of Crohn's disease genes, groups used a variety of methods. Examples include using (1) existing tools such as Phenolyzer, which associates disease terms with genes based on prior research, expands the gene list by using gene–gene relationships, and then creates a ranked list of candidate genes; (2) creating gene lists based on GO pathways enriched with Crohn's disease-associated variants; and (3) using natural language processing to identify genes of interest from PubMed abstracts (Ashburner et al., 2000; Yang, Robinson, & Wang, 2015). From a gene level, different groups would then devise different strategies to select variants of interest. For some approaches, population level frequency

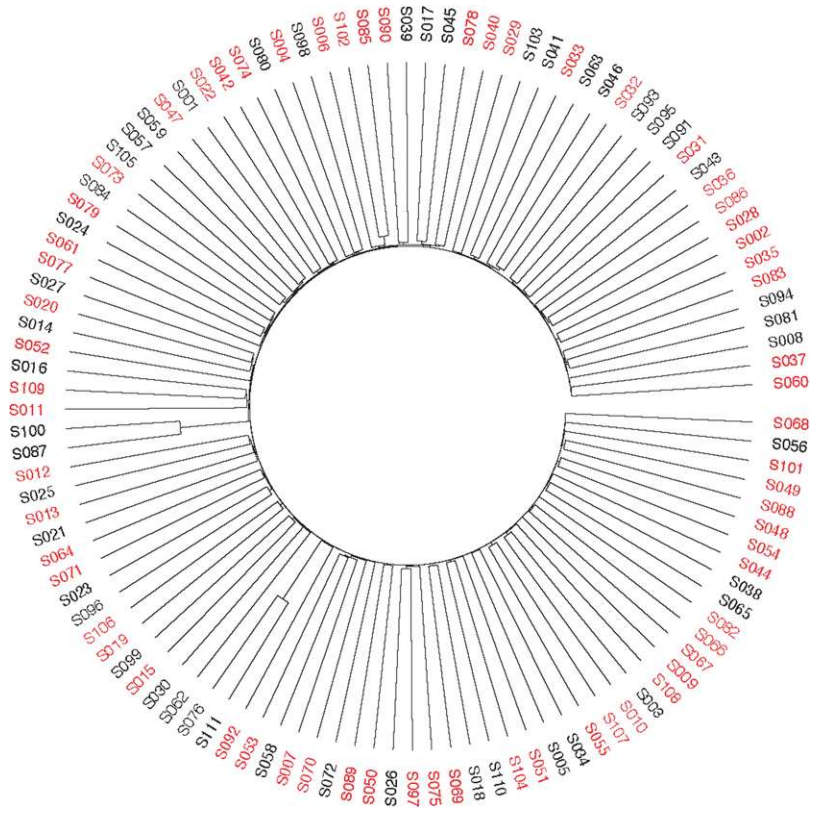


FIGURE 3 Clustering of samples for CAGI 4 Crohn's disease challenge. Black represents controls, and red represents cases

data was used to help distinguish variants more likely to be pathogenic. Other methods relied on pathogenicity prediction tools such as SNAP, PON-P2, SNPs&GO, and Variant Effect Predictor to inform variant selection and weighting (Bromberg & Rost, 2007; Calabrese, Capriotti, Fariselli, Martelli, & Casadio, 2009; McLaren et al., 2010; Niroula, Urolagin, & Vihinen, 2015).

A range of machine learning approaches were used to actually build the classifiers: naïve Bayes, logistic regression, neural nets, and random forests. Additionally, some groups improved on prior iterations by creating metaclassifiers based on combinations of prior methods.

3.2 | Bipolar disorder exomes challenge (CAGI 4)

As noted, a substantial difference between the Crohn's disease phenotypic prediction challenge and the bipolar disorder challenge was that a substantial amount of training data was provided for the bipolar disorder challenge, with 500 of the 1,000 exomes randomly selected and provided as training data for the challenge. These samples were unrelated, and analysis steps assessing the relationships between samples can be found in Supp. Figs. S4–S6. The top performing group had a method with an AUC of 0.64. The distribution of AUCs across methods is shown in Figure 5. Although many groups used approaches similar to those used for the Crohn's disease challenge, the top performing group (which did not apply this method to Crohn's disease data) treated the genotype data as linear features and trained a neural network with three hidden layers, with the middle layers looking at local features in the linear space of the ordered SNPs of the

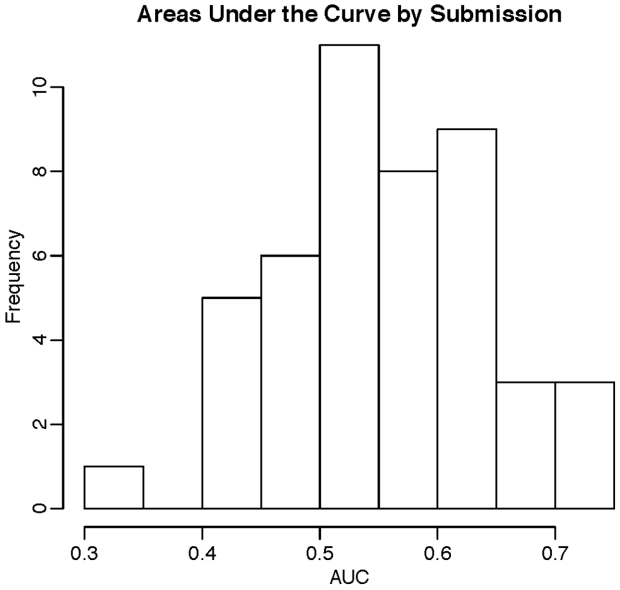


FIGURE 4 CAGI 4 Crohn's disease challenge distribution of AUCs across all methods

VCF file, tuning for performance using cross-validation on the test data. Importantly, this approach used essentially no prior knowledge of genetics or the results of prior studies on disease–gene relationships. Supp. File 4 shows comparative results of the CAGI 4 bipolar disorder challenge. Overall descriptions of prediction methods are available under Supp. Exhibit 2: CAGI 4. A metaclassifier created by the assessment team using all submitted methods for this challenge, as shown in

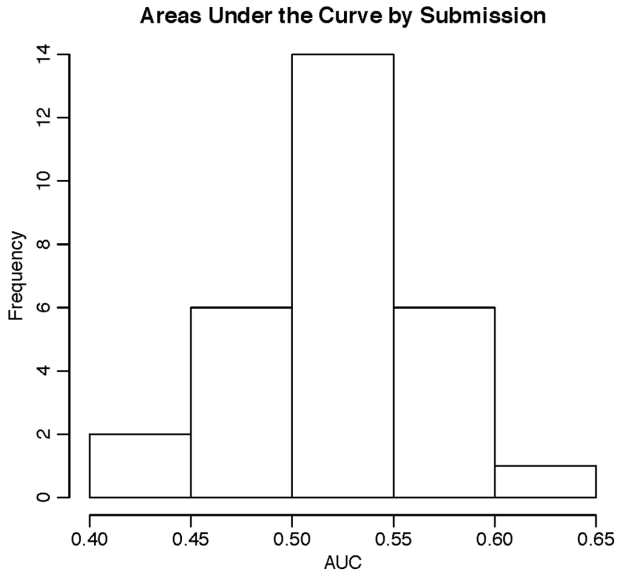


FIGURE 5 CAGI 4 bipolar disorder challenge distribution of AUCs across all methods

Supp. Figure S7, had an AUC of 0.64, which was not notably different from the top method.

3.3 | Warfarin exomes challenge (CAGI 4)

With the warfarin exomes challenge, similar to the Crohn's disease challenge, many groups utilized a priori data to create a list of covariates to use for their models. This included known pharmacokinetic and pharmacodynamic warfarin genes, genes mentioned in the literature, and also using tools to find functional neighbors of the known gene set.

One prediction method (Group 50, Prediction 1) was ahead of the others when looking across multiple performance metrics described in the methods section—R^2, mean absolute value of z score, and mean absolute value of z score multiplied by the coefficient of variation (Fig. 6A–D; Supp. Table S1). The R^2 of the top prediction method was 0.25, compared with 0.35 for the IWPC prediction method, one of the best performing published predictive algorithms. A visualization of the predictions compared with the actual dose can be seen in Supp. Figures S8 and S9. Details of all methods can be found in Supp. Exhibit 3:CAGI 4.

The methods submitted for this challenge had several similar features. Every method submitted took advantage of the fact that the range of the actual doses were published in the paper from which the data came. Thus, these methods either fit rankings to the dose range or set predicted doses above or below the known range to the lower or upper limits. Additionally, most methods used prior information from the literature to help set the initial clinical and genetic covariates to consider in their models.

4 | DISCUSSION

The CAGI exomes challenges revealed lessons specific to each particular challenge as well as generalizable principles for future genotype–phenotype prediction challenges.

4.1 | Crohn's disease

Overall, there were substantial challenges with bias and population stratification in the datasets that made the evaluation and comparison of techniques for identifying Crohn's disease status from exome data difficult. In the latest crop of prediction systems, it may be that techniques such as using imputation to infer variants in regions not covered by the exome sequencing and using large external microarray SNP chip datasets for classifier training were key factors in superior performance. The top AUC varied across the three evaluations, demonstrating the substantial differences in the data sets. Groups who created metaclassifiers based on combining previous methods from previous CAGI challenges demonstrated the value of applying the Common Task Framework to genetic problems—through iteratively improving their methods based on prior learning. Importantly, across the three CAGI evaluations, the average system performance performed better than random, including in the most recent, CAGI 4, implying that there is some level of useful information in predicting the likelihood of Crohn's disease from exome data in the population, something previously not demonstrated.

4.2 | Bipolar disorder

Surprisingly, the group that created the best performing prediction in the bipolar disorder challenge acknowledged having little background in biomedicine or genetics. This group approached the problem as purely a data classification challenge. On the one hand, this may be hailed as another example of the unreasonable effectiveness of data and the success of machine learning over human expertise; the quotation "Every time I fire a linguist, the performance of our speech recognition system goes up," has been attributed to Fred Jelinek in the 1980s, and something similar may be afoot in genomics, promising an exciting future as datasets expand and machine learning techniques improve. However, one of the major challenges is that prediction accuracy with case-control data does not really reflect most applications we can envision for a phenotypic prediction system. Moreover, while not detected by any of our quality control methods, it is still possible that the top performing method picked up on hidden population stratification/biases in the data. Although we were unable to find evidence of this, a sophisticated machine learning system may be identifying features that partition the cases and controls but that are not related to biological drivers of disease risk. Unfortunately, the tools to dissect the deep neural net architecture in the context of genomic features are currently too primitive to help us deepen our biological understanding using these results. There has been recent work into advanced techniques to understand the decisions made by previous black box systems in areas like image processing and natural language processing; however, similar tools for understanding genomic prediction systems are less developed (Ribeiro, Singh, & Guestrin, 2016)).

4.3 | Warfarin

Predicting warfarin dose using clinical information and genetics is a difficult problem; one of the best performing algorithms (IWPC) has an R^2 of 0.35 on this data set. Existing algorithms have poorer performance

on diverse populations since most algorithms are trained on European descent populations (Daneshjou et al., 2014; Klein et al., 2009). For this challenge, the winning method had an R^2 of 0.25.

The warfarin exomes challenge had several limitations. The sample size was limited, with only 50 samples for training and 53 for testing. Data were generated at a time when exome sequencing was more expensive; falling costs may allow an expansion of available exome data. Additionally, all groups used the known dose range of the cohort when assigning their predicted doses. Because of the use of this known range, some of these methods may be tailored particularly to this challenge and not be generalizable to the wider population.

4.4 | Overall lessons from CAGI exomes challenges

An advantage of the common task structure is the ability to iterate quickly and learn from the setbacks of the groups analyzing the data. The exomes challenges allowed us to glean several important lessons that will inform future iterations of CAGI.

The importance of population stratification, batch effects, and hidden biases became evident early on with the CAGI 2 Crohn's disease challenge (Fig. 1). In that particular instance, either population stratification or batch effects created a discernable difference between cases and controls that was unlikely related to actual disease status. Based on that finding in CAGI 2, every subsequent CAGI challenge included a preanalysis of the whole-exome data trying to identify whether there were samples that clustered together inappropriately based on case-control status. Population stratification has long been an issue in genetic studies. The most obvious issue arises when cases and controls come from distinctly different ancestral populations, such as comparing Northern European cases against Chinese controls. However, less obvious stratification can also be an issue, such as differences in admixture/population substructure or cryptic relatedness (Price, Zaitlen, Reich, & Patterson, 2010). Batch effects can occur at many different steps in the pipeline, for example, if samples from the cases and controls have differences in sample preparation, DNA quality, sequencing coverage, or genotype calling. Any of the above can result in prediction methods that perform well due to systemic biases between cases and controls rather than true features that define case-control status.

How these challenge datasets emulate the real world was another important consideration and was a topic of discussion among the CAGI 4 community.

A majority of the challenges used samples of Northern European ancestry, only the warfarin dose prediction challenge used samples of African American ancestry. In order for the methods to be generalizable to real-world populations, representation of human diversity is necessary, particularly since disease risk and pharmacogenetic variants can be population-specific (Rosenberg et al., 2010). Moreover, the CAGI exome datasets all came from research studies, which are often designed to maximize the possibility of picking up a significant signal. One way to achieve this is through selecting for extreme phenotypes—a strategy employed by both the Crohn's disease exome

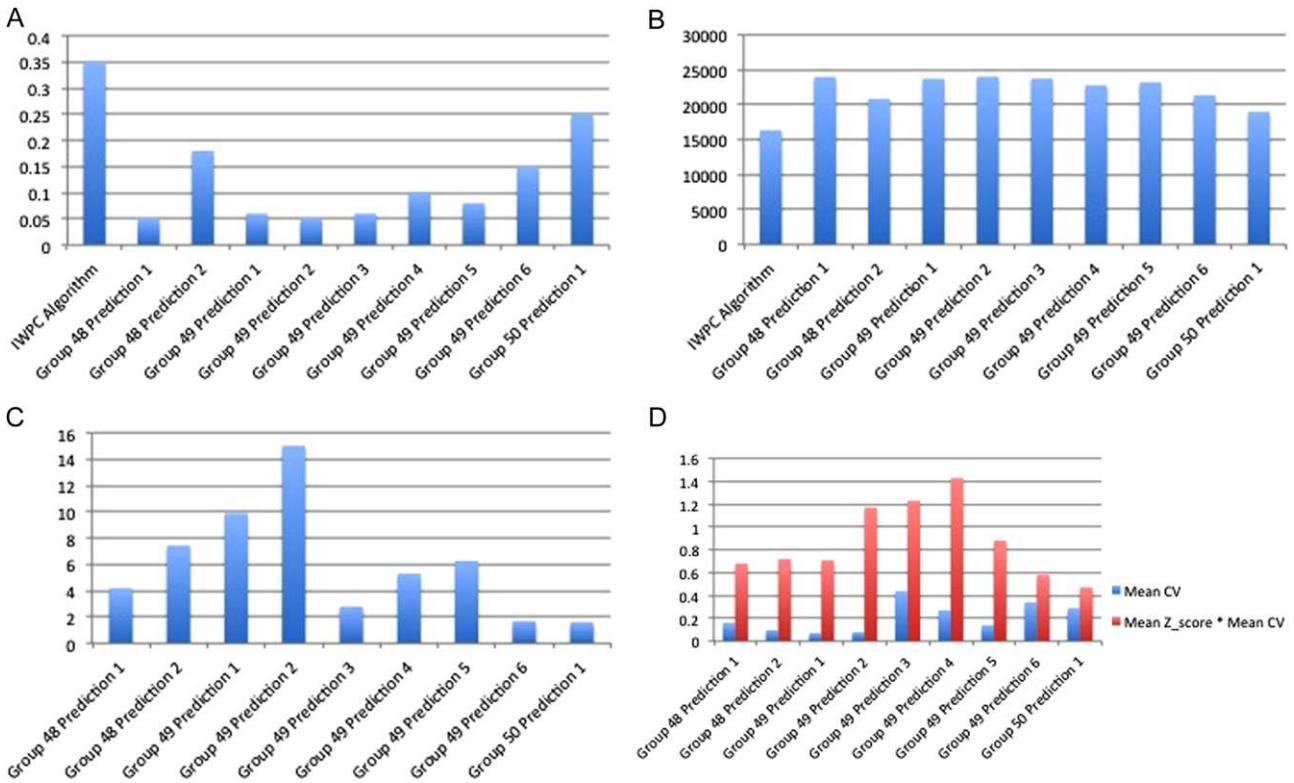


FIGURE 6 **A**: R^2 between predicted doses and actual doses for each group's prediction method as well as the IWPC algorithm. **B**: Sum of squared errors for each group's prediction method and the IWPC algorithm. **C**: Mean *z* scores calculated from each group's predicted doses with predicted standard deviations and actual doses. **D**: Mean coefficient of variation (CV) and mean CV multiplied by mean *z* score for each group's prediction method

dataset (which selected a subset of cases who had early-onset Crohn's disease) and the warfarin prediction exome dataset (selected from individuals requiring "low" and "high" doses to achieve the therapeutic effect) (Manolio et al., 2009). However, while this strategy works well for increasing signal strength in research, using such data for building a classifier may lead to a biased predictor that has difficulty differentiating between the more subtle variations seen in the real world. Having larger datasets and using data generated for clinical use may help remedy some of these issues in the future.

Finally, one of the most promising lessons from CAGI was on the effectiveness of data. As mentioned before, for complex tasks, the common task framework has provided a way to have many people work on a problem and iterate quickly. After each challenge ended, the evaluation scripts and the challenge answers were shared so that participants could analyze when their prediction methods succeeded or failed. This process allowed groups to have information for future improvement. Additionally, large datasets, even if imperfect, have also been shown to be a critical part of developing algorithms to tackle a complicated task (Pereira, Norvig, & Halevy, 2009). Critical to accumulating large enough datasets is data sharing, and the open data movement aims to encourage increased biomedical data sharing (McNutt, 2016). However, one of the difficulties with genetic data that includes protected health information is sharing data in a secure manner. CAGI, which includes data encryption and verifies the groups participating, can provide a platform to facilitate sharing such data. As a result of the data accumulated thus far, CAGI has demonstrated how data can, in certain cases, surmount prior biological knowledge. For CAGI 4, the bipolar disease challenge was the best example; individuals with no biological background, but a strong background in data science, had the best performance. In particular, this should inspire a more multidisciplinary approach to genotype–phenotype prediction and a greater effort to engage those whose backgrounds are more data driven rather than biologically driven.

Overall, the CAGI exomes challenges provided an opportunity to begin building the classifiers required to implement precision medicine. While there is still a long road ahead for genotype–phenotype prediction, the accumulation of larger datasets and the participation of more groups with every subsequent CAGI holds promise for continued improvement.

ACKNOWLEDGMENTS

We would like to thank and acknowledge the CAGI planning committee, as well as all data providers and participants.

DISCLOSURE STATEMENT

R.M. has participated in Illumina-sponsored meetings over the last 4 years and received travel reimbursement and an honorarium for presenting at these events. Illumina had no role in decisions relating to the study/work to be published, data collection, and analysis of data and the decision to publish.

R.M. has participated in Pacific Biosciences-sponsored meetings over the last 3 years and received travel reimbursement for presenting at these events.

R.M. is a founder and shared holder of Orion Genomics, which focuses on plant genomics and cancer genetics.

R.M. is a SAB member for RainDance Technologies, Inc.

All the other authors have no conflict of interest to declare.

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29.

Ashley, E. A. (2015). The precision medicine initiative: A new national effort. *JAMA*, *313*(21), 2119–2120.

Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., … Altman, R. B. (2010). Clinical assessment incorporating a personal genome. *Lancet*, *375*(9725), 1525–1535.

Bauer, K. A. (2011). Recent progress in anticoagulant therapy: Oral direct inhibitors of thrombin and factor Xa. *Journal of Thrombosis and Haemostasis*, *9*(Suppl 1), 12–19.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *SIGKDD Explorations Newsletter*, *9*(2), 75–79.

Bromberg, Y., & Rost, B. (2007). SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Research*, *35*(11), 3823–3835.

Brown, T. L., & Meloche, T. M. (2016). Exome sequencing a review of new strategies for rare genomic disease research. *Genomics*, *108*(3–4), 109–114.

Budnitz, D. S., Lovegrove, M. C., Shehab, N., & Richards, C. L. (2011). Emergency hospitalizations for adverse drug events in older Americans. *The New England Journal of Medicine*, *365*(21), 2002–2012.

CAGI. (2011). Critical Assessment of Genome Interpretation. Retrieved from https://genomeinterpretation.org/.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutation*, *30*(8), 1237–1244.

Cho, J. H. (2008). The genetics and immunopathogenesis of inflammatory bowel disease. *Nature Reviews Immunology*, *8*(6), 458–466.

Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Academies of Sciences, Engineering, and Medicine, & Schwalbe, M. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington, D. C.: National Academies Press.

Craddock, N., & Jones, I. (1999). Genetics of bipolar disorder. *Journal of Medical Genetics*, *36*(8), 585–594.

Craddock, N., & Sklar, P. (2013). Genetics of bipolar disorder. *Lancet*, *381*(9878), 1654–1662.

Daneshjou, R., Gamazon, E. R., Burkley, B., Cavallari, L. H., Johnson, J. A., Klein, T. E., … Perera, M. A. (2014). Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood*, *124*(14), 2298–2305.

Donoho, D. (2015). 50 years of data science. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

Ellinghaus, D., Zhang, H., Zeissig, S., Lipinski, S., Till, A., Jiang, T., … Franke, A. (2013). Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*, *145*(2), 339–347.

Fagan, T. J. (1975). Letter: Nomogram for Bayes theorem. *The New England Journal of Medicine*, *293*(5), 257.

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, *27*(13), 1741–1748.

Halfvarson, J., Bodin, L., Tysk, C., Lindberg, E., & Jarnerot, G. (2003). Inflammatory bowel disease in a Swedish twin cohort: A long-term follow-up of concordance and clinical characteristics. *Gastroenterology*, *124*(7), 1767–1773.

IMS Institute of Healthcare Informatics. (2012). The use of medicines in the United States: Review of 2011. Retrieved from https://www.imshealth.com/files/web/IMSH%20Institute/Reports/The%20Use%20of%20Medicines%20in%20the%20United%20States%202011/IHII_Medicines_in_U.S_Report_2011.pdf.

Klein, T. E., Altman, R. B., Eriksson, N., Gage, B. F., Kimmel, S. E., Lee, M. T., … Johnson, J. A. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *The New England Journal of Medicine*, *360*(8), 753–764.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, *26*(16), 2069–2070.

McNutt, M. (2016). #IAmAResearchParasite. *Science*, *351*(6277), 1005–1005.

Monson, E. T., Pirooznia, M., Parla, J., Kramer, M., Goes, F. S., Gaine, M. E., … Willour, V. L. (2017). Assessment of whole-exome sequence data in attempted suicide within a bipolar disorder cohort. *Molecular Neuropsychiatry*, *3*, 1–11.

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., … Hirschman, L. (2008). Overview of BioCreative II gene normalization. *Genome Biology*, *9*(Suppl 2), S3.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins 82* (Suppl 2), 1–6.

Niroula, A., Urolagin, S., & Vihinen, M. (2015). PON-P2: Prediction method for fast and reliable identification of harmful variants. *PLoS One*, *10*(2), e0117380.

Pereira, F., Norvig, P., & Halevy, A. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*, 8–12.

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Review Genetics*, *11*(7), 459–463.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Retrieved from http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, *11*(5), 356–366.

Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., … Mahoney, P. (2006). Stanley: The robot that won the DARPA Grand Challenge. *Journal of Field Robotics*, *23*(9), 661–692.

Uhlig, H. H., Schwerd, T., Koletzko, S., Shah, N., Kammermeier, J., Elkadri, A., … COLORS in IBD Study Group and NEOPICS. (2014). The diagnostic approach to monogenic very early onset inflammatory bowel disease. *Gastroenterology*, *147*(5), 990–1007.e3.

Walker, M. A., Passonneau, R., & Boland, J. E. (2001). Quantitative and qualitative evaluation of Darpa Communicator spoken dialogue systems. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Toulouse, France: Association for Computational Linguistics, 515–522.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–D1006.

Wray, N. R., Yang, J., Goddard, M. E., & Visscher, P. M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genetics*, *6*(2), e1000864.

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

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Daneshjou R, Wang Y, Bromberg Y, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*. 2017;38:1182–1192. https://doi.org/10.1002/humu.23280