Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Scienze della Vita, della Terra e dell'Ambiente

Ciclo XXXI

Settore Concorsuale: 05/E1 - BIOCHIMICA GENERALE

Settore Scientifico Disciplinare: BIO/10 - BIOCHIMICA

THE BIOLOGICAL COMPLEXITY OF THE GENOTYPE-PHENOTYPE RELATION: FROM GENES AND PROTEINS TO PHENOTYPES AND DISEASES

Presentata da: Giulia Babbi

Coordinatore Dottorato	**Supervisore**
Prof. Giulio Viola	Prof. Pier Luigi Martelli

Esame finale anno 2019

Abstract

To unveil the biological complexity at the basis of the genotype-phenotype relation it is fundamental to integrate knowledge, that is to integrate the different omics describing the levels of biological complexity: genomics, proteomics, transcriptomics, metabolomics, interactomics. The annotation process is time consuming because the various information on genes and proteins are collected in different databases of annotation, and we still need a unified framework collecting all the different levels of knowledge.

The situation gets more complicated when we move the focus to diseases and phenotypes. The identification of molecular mechanisms behind different phenotypes offers a way to understand the processes that lead to disease insurgence and progression. In the context of precision medicine, the challenge is to ascribe the ensemble of phenotypes to a small number of possibly altered biological functions. Another issue in computational biology is the prediction of specific phenotypic effect of gene and protein variants, to test the performance of computational methods towards experiment in vivo and in vitro.

The main aim of this thesis is to study the relations among genes, variations, diseases and phenotypes with the approaches of computational biology, integrating information from different resources to make a step forward in the direction of unveiling the biological complexity. After a general introduction (chapter 1), we present the webservers eDGAR (chapters 2, 3) and PhenPath (chapter 4), collecting and analysing the gene-disease associations and the phenotypes-biological processes associations, respectively.

We then assessed whether disease-related variations induce perturbations of the protein stability. To this aim, we developed a new predictor called INPS-3D (chapter 5). We test our predictors participating in international experiments (chapters 6, 7, 8) on specific study cases. Thanks to the expertise acquired in the field, we also collaborate with the Sant'Orsola Genetic Medical Unit of the Department of Medicine and Surgery of the University of Bologna, building a series of model of protein structure of myosin 1F and its variants related to the thyroid cancer (chapter 9).

Concluding (chapters 10, 11), we tried to depict the biological complexity merging a large-scale approach with the analysis of specific study cases, providing webservers, tools and computation methods to help researchers in directing further experiments.

Index

1 Introduction

1.1 Unravelling the Biological Complexity

The advent of Next Generation Sequencing (NGS) technologies and the possibility to generate genetic, transcriptomic, epigenetic data and other genome-wide data for a relatively small cost opened numerous opportunities for translation into the clinic (Casey G et al, 2013). How could we transfer the knowledge derived from NGS studies to have a real impact on the clinical management of diseases? How to handle the massive amount of data?

Data annotation and data integration are part of the answer to this biological problem.

It is clear that to unveil the biological complexity at the basis of the genotype-phenotype relation it is fundamental to integrate knowledge, which is to integrate the different omics describing the levels of biological complexity: genomics, proteomics, transcriptomics, metabolomics, interactomics. This integration of features increases our understanding of the molecular mechanisms leading from a genetic variation to a specific phenotype. The study of the general genotype-phenotype relation is at the basis of the comprehension of the variant-disease relation, that is a research area involving a series of bioinformatics approaches that may be defined as 'translational bioinformatics'. Nowadays, we are far from having a complete understanding of the intricate network of the molecular processes involved in disorders, and we are still searching for cures for most complex diseases (Kann MG, 2009).

As data on gene-disease relations accumulate, it emerges that an increasing number of diseases are associated with several genes. Such multigenic diseases are defined as heterogeneous or polygenic diseases, on the basis of their association to independent or concomitant alterations in sequence and/or in expression of sets of genes, respectively (McClellan J and King MC, 2010). A crucial goal in the direction of precision medicine is to understand the molecular mechanisms that connect the different genes associated to the same disease. This aspect is very complicated, because the information on genes is stored in different databases of annotation, and it is difficult to collect all the different levels of knowledge in a unifying framework.

The situation gets more complicated when we move the focus from diseases to phenotypes. In fact, many diseases are associated to symptom complexes and co-occurrence of different phenotypes whose diversity complicates the understanding of the underlying molecular mechanisms (Fisch GS, 2017). The identification of molecular mechanisms behind different phenotypes offers a way to understand the processes that lead to disease insurgence and progression. In the context of personalized medicine, the challenge is to ascribe the ensemble of phenotypes to a small number of possibly altered biological functions.

It is clear the need of comprehensive resources to help researchers in directing future efforts, collecting and merging the annotation features from many different source databases, as well as providing clues on possible genes and biological pathways related to diseases and/or phenotypes to be investigated.

This thesis aims to make a step forward in the direction of unveiling the biological complexity of the genotype-phenotype relation. In order to do so, we decided to analyse the genotype-phenotype relations with two different approaches:

i) Large scale studies: to develop methods that integrate knowledge, to study the emerging features of the gene-diseases associations and of the phenotype-molecular pathways associations, to provide databases and tools for the researchers' community.

ii) Specific study cases: to investigate real applications of the general models developed in the large-scale studies, to test the current computational resources comparing them with experimentally validated data, to collaborate with clinicians in developing new strategies in the direction of precision medicine.

1.2 Resources for gene and protein annotation

There are many databases collecting genes and their associated features that can be used for gene annotation. Annotating a gene or a protein means to endow it with specific biological features (e.g. molecular functions, biological processes and pathways, protein 3D structure, disease associations).

Routinely, the databases used for the annotation process may be specific for a particular type of features (e.g. the protein products, the protein variants, the disease associations) or they may regard a subset of specific genes (e.g. organism specific databases).

Among the great number of available resources, we selected a set of annotation databases to retrieve information about genes and proteins. Selection considered different criteria, such as the amount of entries described, the level of data curation by experts, the frequency of updating and releasing, the usage of international standards to report the data. In the following paragraphs it is reported a brief list of the main resources for data annotation used in this thesis project as well as in the associated papers.

1.2.1 UniProt

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data (The UniProt Consortium, 2017). In particular in this thesis we refer to the UniProt Knowledgebase (UniProtKB), that is composed of two databases: SwissProt (release of August 2018, accessed September 12, 2018), containing 558,125 protein sequences, that is reviewed and manually curated with information extracted from literature and curator-evaluated computational analysis, and TrEMBL, that contains 124,797,108 protein sequences automatically annotated.

Another useful resource provided by the UniProt Consortium is Humsavar, a collection of all the missense variants annotated in human SwissProt entries, actually containing 78,049 single amino acid variants (release of August 2018, accessed September 12, 2018). Humsavar classifies its variants in disease variants, polymorphisms and unclassified variants, on the basis of the curated information retrieved by experts working on SwissProt. The current release account for 30,251 disease related variants and 39,963 polymorphisms.

1.2.2 OMIM

The Online Mendelian Inheritance in Man (OMIM) (Amberger JS et al, 2015) is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily. OMIM contains information on Mendelian disorders over 15,000 genes, collecting genotype-phenotype relations. Presently (accessed October 8, 2018) it contains information on 6,259 diseases with known molecular basis, associated with 3,961 genes. The OMIM classification for Mendelian diseases has become a standard in the international community as long as the MIM-numbers (a six-digit code univocally associated to a disease) have been widely used by researchers to identify human disorders.

1.2.3 ClinVar

ClinVar is a database of human variations and their relations with diseases (Landrum MJ et al, 2016). ClinVar maps the variants to reference sequences according to the HGVS standard, and it reports for each variant the patient samples, the clinical significance and other supporting data. The level of confidence in the accuracy of variation calls and assertions of clinical significance depends in large part on the supporting evidence, so this information is very important. A review status is assigned to each assertion, to support communication about its trustworthiness. Presently (accessed October 22, 2018), ClinVar contains data on 458,485 variations with interpretation, among them only 10,529 variations are curated by experts. ClinVar variations span across many genes, for a total of 30,219 genes; among them, only 6,035 protein coding genes are associated with specific variants not overlapping more genes.

1.2.4 DisGeNET

The DisGeNET database collects human gene-disease associations from different resources, merging various curated databases and text-mining derived associations including Mendelian, complex and environmental diseases. The crucial operation of integration is performed via gene and disease vocabulary mapping (Piñero J et al, 2017).

The information in DisGeNET can be accessed via multiple access points, including many different user interfaces that are increasing the usage and the spread of this resource.

The current version of DisGeNET (v5.0) contains 561,119 gene-disease associations, between 17,074 genes and 20,370 disorders and traits, and 135,588 variant-disease associations, between 83,002 single nucleotide polymorphisms and 9,169 diseases.

1.2.5 HPO

The Human Phenotype Ontology (HPO) (Köhler S et al, 2017) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease.

The HPO is being increasingly adopted as a standard for phenotypic abnormalities by diverse groups such as international rare disease organizations, registries, clinical labs, biomedical resources, and clinical software tools and will thereby contribute toward nascent efforts at global data exchange for identifying disease aetiologies (Köhler S et al, 2017).

HPO currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases (accessed October 10, 2018). The terms are arranged in a directed acyclic graph and they are connected by *is-a* (subclass-of) edges, such that a term represents a more specific or limited instance of its parent term(s). Phenotypic abnormality is the main subontology of HPO and it contains descriptions of clinical districts and their phenotypes. Additional subontologies are provided to describe inheritance patterns, onset/clinical course and modifiers of abnormalities.

1.2.6 Gene Ontology

The Gene Ontology (GO; Gene Ontology Consortium, 2017) is a vocabulary of functional terms that is composed of three main categories: i) Cellular Component (CC), ii) Molecular Function (MF), iii) Biological Process (BP). Cellular component terms describe the different cellular localizations, such as organelles, membranes, other anatomical structures (e.g. cytoplasm) or specific gene product complexes (e.g. ribosome). Molecular function terms describe activities that occur at the molecular level (e.g. enzymatic reactions). Biological process terms describe biological events (e.g. negative regulation of apoptotic process) that results from the concerted organization of many molecules with various molecular functions. The last version of the GO resource (Amigo 1.8, Carbon S et al, 2009) annotates more than 90% of human genes with 45,043 biological functions (29,691 GO:BP, 11,150 GO:MF and 4,202 GO:CC, accessed October 26, 2018).

1.2.7 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes, Kanehisa M et al, 2017) is a database for the systematic analysis of gene functions. The database is composed of 15 main manually curated databases which are categorized into systems, genomic, chemical and health information. It is one of the most spread databases used in functional annotation of genes, proteins and small molecules. In particular, KEGG PATHWAY collects manually drawn pathway maps connecting with links the genes/proteins and associating them, resulting in higher level of biological complexity. Pathway maps represent the dual aspect of the metabolism: the genomic network, connecting the genome-encoded enzymes catalysing biochemical reactions, and the chemical network, composed by the compounds that are transformed by means of these enzymes (Kanehisa M, 2013). To increase the organization of information stored in KEGG, they developed KEGG BRITE that provides a functional hierarchy of the KEGG objects.
The last release of KEGG (Release 88.1, accessed October 26, 2018) annotates a total of 7,469 human genes in 330 pathways (considering only the lowest level of the hierarchy).

1.2.8 Reactome

Reactome (Fabregat A et al, 2018) is a manually curated database of pathways and processes. Starting from the physical interactions occurring in cells, Reactome describes the chemical reactions in the framework of biological pathways, providing information about proteins and small molecules and their related pathways. Different connected reactions are grouped into pathways, and then pathways are structured in a hierarchy of biological events. Reactome maps describe canonical biochemical pathways and cellular processes, as well as the molecular pathways involved in diseases. A unique characteristic of Reactome is that it divides genes in specific "modules" that are part of more general biochemical pathways. The last version of Reactome (v.66, accessed October 26, 2018) annotates a total of 10,870 genes in 2,244 pathways describing 12,047 reactions.

1.2.9 Protein Data Bank

The Protein Data Bank (PDB, Berman HM et al, 2002) is an open access experimental data archive, providing access to 3D structure data for large biological molecules (proteins, DNA, and RNA). Knowing the 3D structure of a biological macromolecule is essential for understanding its function and consequently its role in human health and diseases. The current version of PDB (accessed October 9, 2018) contains 144,871 biological macromolecular structures. *Homo sapiens* is one of the most represented organisms with more than 41,427 related entries.

1.2.10 NET-GE

NET-GE is a method for network-based gene enrichment analysis (Di Lena P et al, 2015; Bovo S et al, 2016). NET-GE relies on the STRING Human Interactome (release 10, Szklarczyk D et al, 2015) and the annotation derives from the Gene Ontology resource (The Gene Ontology Consortium, 2017), KEGG (Kanehisa M et al, 2017) and Reactome (Fabregat A et al, 2018) databases.

Starting from data stored in STRING interactome (Szklarczyk D et al, 2015), NET-GE first performs a module building procedure, aimed at extracting connected and compact subgraphs of the STRING interactome (Figure 1).

The resulting modules are then used to address the problem of functional association.

Over-representation analysis is performed by mapping the input gene set of genes or proteins on each module, determining through a Fisher's exact test whether there are significant overlaps among the input set and the modules. NET-GE implements both a standard and a network-based gene enrichment procedure. Entering with a 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 WS, 2009).

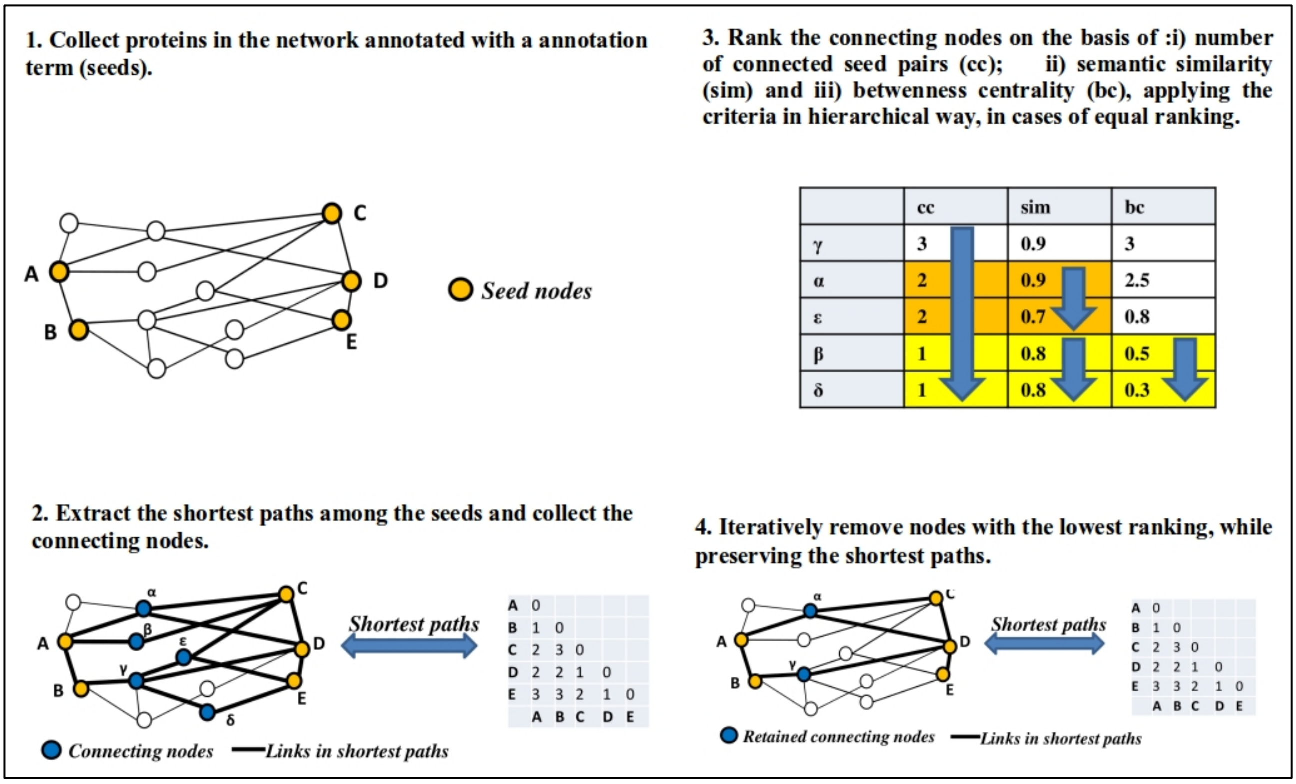

1. Collect proteins in the network annotated with a annotation term (seeds).
A
B
C
D
E
Seed nodes
3. Rank the connecting nodes on the basis of :i) number of connected seed pairs (cc); ii) semantic similarity (sim) and iii) betwenness centrality (bc), applying the criteria in hierarchical way, in cases of equal ranking.
cc
sim
bc
γ 3 0.9 3
α 2 0.9 2.5
ε 2 0.7 0.8
β 1 0.8 0.5
δ 1 0.8 0.3
2. Extract the shortest paths among the seeds and collect the connecting nodes.
α
β
ε
γ
δ
Shortest paths
A 0
B 1 0
C 2 3 0
D 2 2 1 0
E 3 3 2 1 0
A B C D E
Connecting nodes
Links in shortest paths
4. Iteratively remove nodes with the lowest ranking, while preserving the shortest paths.
Shortest paths
Retained connecting nodes
Links in shortest paths

Figure 1. Modules extraction procedure in NET-GE. 1. All the proteins of the network sharing a specific annotation term are collected into a *seed set*. 2. Each *seed set* is expanded into a function-specific module by computing the shortest paths among each pair of seed nodes. 3. Nodes connecting the *seed set* are collected and ranked by using graph theoretic and information theoretic measures. 4. Each module is minimized by filtering out the less informative connecting nodes while preserving the shortest paths (adapted from Di Lena P et al, 2015).

1.3 Workflow

The main aim of this thesis is to study the relations among genes, variations, diseases and phenotypes with the typical approaches of computational biology, integrating information from different resources and merging the various levels determining biological complexity. In details, starting from public databases available online, we analysed the gene-disease associations characterizing human genes, describing the molecular functions, the involved biological processes, the transcription regulation, the protein products and their interactions. We implemented a database of Disease-Gene Associations with annotated Relationships among genes (eDGAR, Babbi G et al, 2017, chapter 2) collecting data on 2,672 diseases associated with 3,658 genes, for a total of 5,729 gene-disease associations. Every gene in eDGAR is well annotated with features derived from different ontologies. Moreover, eDGAR reports the interactions of the protein product of each gene in stable complexes, as well as the known protein-protein interactions. We also report the data derived from NET-GE (see paragraph 1.2.10) to endow the group of genes associated to the same disease with new features, and to identify the biological processes that are specific of each disorder.

We then used the collected information in eDGAR to study the principal characteristic of the diseases associated to more than one gene (612 polygenic and heterogeneous diseases over the 2,672 of eDGAR). We found out that 96% of these diseases are associated to at least one couple of genes sharing the same function; 14% have at least a couple of associated genes that belong to the same macromolecular complex; 56% of the proteins associated to polygenic or heterogeneous diseases are in direct interaction. Moreover, in the 44% of the diseases, there are at least two genes co-regulated by the same transcription factor. These results have been published in the scientific journal BMC Genomics in 2017 (Babbi G et al, 2017, chapter 2). Thanks to a collaboration with the IMIM group at the PRBB centre in Barcelona, and in particular with the group of Prof. Laura I Furlong (Universitat Pompeu Fabra), we integrate our expertise with the one of the curators of DisGeNET (Piñero J et al, 2017), a database of gene-disease associations (see paragraph 1.2.4). We prepared the new version of eDGAR, eDGAR+, that includes an updated data-set of gene-disease associations (12,560 associations, connecting 5,574 diseases to 6,580 genes) and new features of gene annotation, regarding the tissue of expression and the variants known in the literature (chapter 3).

Furthermore, to enlarge the analysis of gene-disease associations, we analysed the associations among phenotypes and diseases, connecting 7,137 phenotypes to 4,292 diseases and 3,446 genes. Starting from these studies, we built a new platform called PhenPath (Babbi G et al, submitted in 2018, chapter 4). We used the NET-GE algorithm to functional enrich the genes associated with 7,137 phenotypes, annotating they biological processes and pathways. Currently, results of this analysis are under review at the journal BMC Genomics.

Beside the study of gene-disease and gene-phenotype associations with a large-scale approach, we also analysed in deep the relations among genetic mutations, protein variants and their associations to diseases and phenotypes. In particular, we built INPS-3D, a tool for the prediction of the effect of protein variants on the protein stability (ΔΔG), based on the 3D-structure of proteins (Martelli PL et al, 2016, chapter 5). Using this powerful predictor, we participated into two editions of the Critical Assessment of Genome Interpretation (CAGI), an international experiment with the aim of testing computational methods for the predictions of phenotypic effects of genetic mutations or protein variants. In particular, using INPS-3D and other strategies, we competed in 16 challenges in the last three years: 6 challenges in the CAGI4 edition (2015-2016) and 10 challenges in the CAGI5 edition (2017-2018). We obtained good results in many challenges presenting our work to an international audience of experts in the field, and we also collaborated in 2 publications in the CAGI4 special issue (Daneshjou R et al, 2017, chapter 7; Xu Q et al, 2017, chapter 8). We are actually collaborating in 3 more publications for the CAGI5 special issue. To better describe the huge work conceiving CAGI challenges and to show the various approaches that we proposed in these years, we collected the data of 5 different challenges predicted with INPS-3D (chapter 6). The various protocols used for the predictions of the phenotypic effects associated the computational approach (predictors data) with the study of the literature and the structural biology knowledge derived from the protein experimental structure, when available.

Thanks to the expertise acquired in the field, we also collaborated with the Sant'Orsola Genetic Medical Unit of the Department of Medicine and Surgery of the University of Bologna, building a series of models of protein structures of myosin 1F and of its variants related to the thyroid cancer (Familial Non-Medullary Thyroid Carcinoma, FNMTC) (Diquigiovanni C et al, 2018, chapter 9).

1.4 ELIXIR

In the bioinformatics era, it is fundamental to integrate data and to share resources. To collaborate with other organizations around the world, it is important to create communities of researches that are interested in the same field. In this direction, enlarging the network of collaborations is fundamental, and being part of the ELIXIR community is important especially for what regards resources integration and interoperability.

ELIXIR is an intergovernmental organization that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers. The goal of ELIXIR is to coordinate these resources, making easier for scientists to find and share data, exchange expertise, and agree on best practices. The Bologna Biocomputing Group is part of the ELIXIR community and we collaborate in the improvement of the quality and adoption of Bioschemas, a set of semantic annotations for tools, data and samples developed by the ELIXIR Interoperability platform.

Tools should be easily accessible by other automated services. This could be done via well-defined Application Program Interfaces (APIs): as an example, eDGAR (Babbi G et al, 2017, chapter 2), a database of disease-gene associations, implements RD-Connect API used by the Rare Disease community.

2 eDGAR

2.1 Contribution to the state of the art

Here we present eDGAR, a database of gene-disease associations, with a specific focus on the annotations of intergenic relations in heterogeneous and polygenic diseases. We merge, without redundancy, data from OMIM (Amberger JS et al, 2015), ClinVar (Landrum MJ et al, 2016), and Humsavar (The UniProt Consortium, 2017).

With the advent of Next Generation Sequencing techniques, lists of genes involved in several diseases have been determined. Although many collections of gene-disease associations already exist (e.g. OMIM, ClinVar, Humsavar, MalaCards (Rappaport N et al, 2017) and DisGeNET (Piñero J et al, 2017)), the need of a resource for the deep investigation on the features shared among genes/proteins co-involved in the same disease is still unfilled. Indeed, the analysis of their relations can help targeting the important biological processes and pathways implicated in the disease and can therefore narrow the search of other possibly involved genes. At present, a database collecting data only on digenic diseases (related to concomitant defects in pairs of genes) is available (DIDA, Gazzo AM et al, 2016) and reports the relations between pairs of genes involved in 54 diseases.

For each gene in eDGAR, the database reports many features like the cytogenetic location, links to the Ensembl (Zerbino DR et al, 2018), SwissProt (The UniProt Consortium, 2017), PDB entries (Berman HM et al, 2002), Gene Ontology (GO, Gene Ontology Consortium, 2017) annotations and links to the KEGG (Kanehisa M et al, 2017) and REACTOME pathways (Fabregat A et al, 2018), when available.

For sets of genes involved in the same disease, the database collects from publicly available databases different types of features: physical interactions, co-occurrence in protein complexes, regulatory interactions, shared functions and pathways, functional terms significantly enriched with NET-GE (Di Lena P et al, 2015; Bovo S et al, 2016) in the set when possible and the co-localization in neighbouring cytogenetic loci. Information is organized in a relational database and an interface allows customized data search and retrieval.

eDGAR offers a new resource to analyse disease-gene associations, especially in multigenic diseases where genes can share physical interactions and/or co-occurrence in the same functional processes.

2.2 General information on the paper

The presented paper can be found in the following publication:

Authors: Giulia Babbi, Pier L. Martelli, Giuseppe Profiti, Samuele Bovo, Castrense Savojardo and Rita Casadio

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

Journal: BMC Genomics

Volume: 18(Suppl 5) **Issue:** 554 **Pages:** 25-34

Year: 2017

Impact Factor: 3.73

Quartile and subject: 1st quartile in Biotechnology and Applied Microbiology

DOI: 10.1186/s12864-017-3911-3.

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

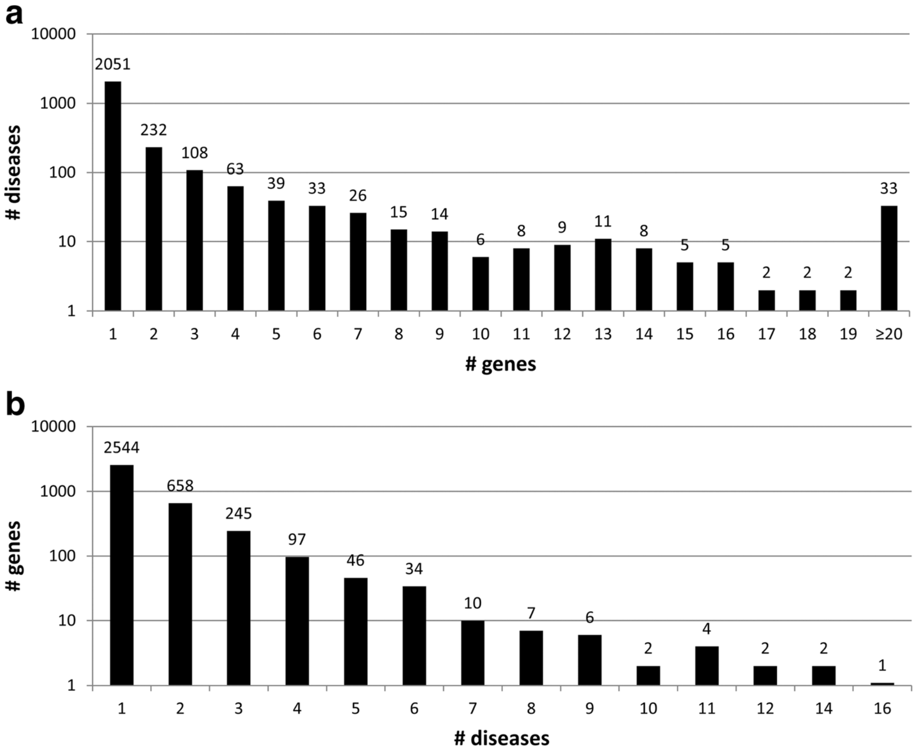


Fig. 1 Distribution of gene-disease associations. The Y-axis scale is logarithmic. **a** Number (#) of genes associated with diseases. 2672 diseases are distributed with respect to the number of associated genes. 2051 diseases are monogenic; 621 diseases are associated with multiple genes (from 2 to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

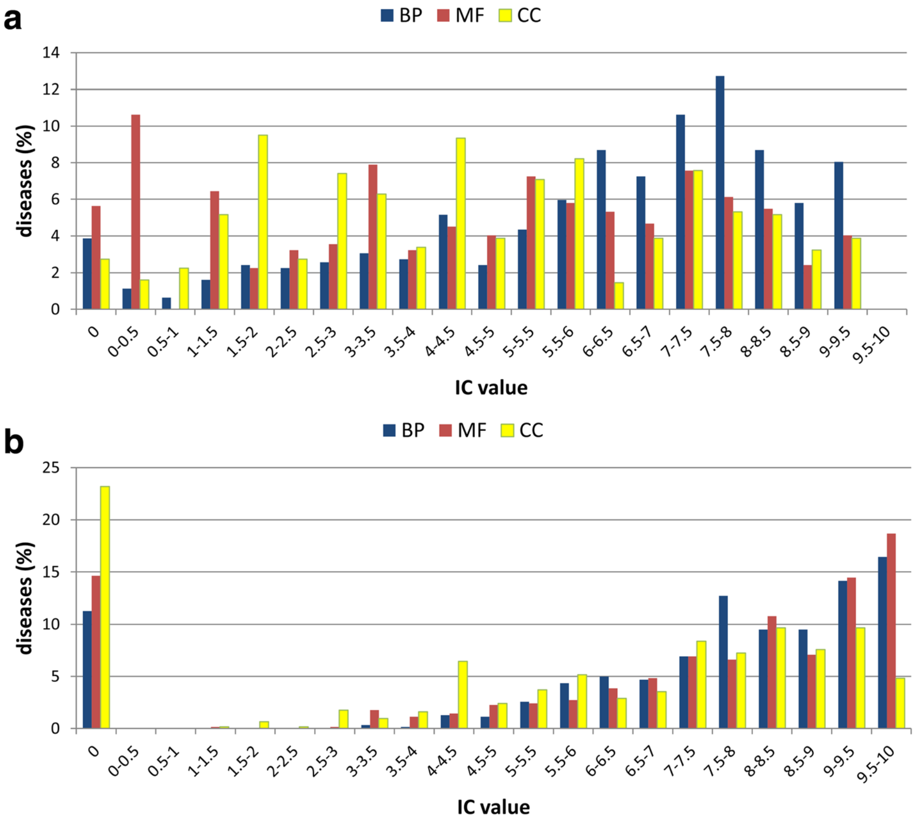


Fig. 2 Distribution of best IC values of GO terms for genes involved in multigenic diseases. **a** GO terms shared by genes; **b** GO terms after enrichment with NET-GE. For each multigenic disease, IC values of gene-associated GO terms (of the three different roots) are evaluated (Eq. 1). In the figure, 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.

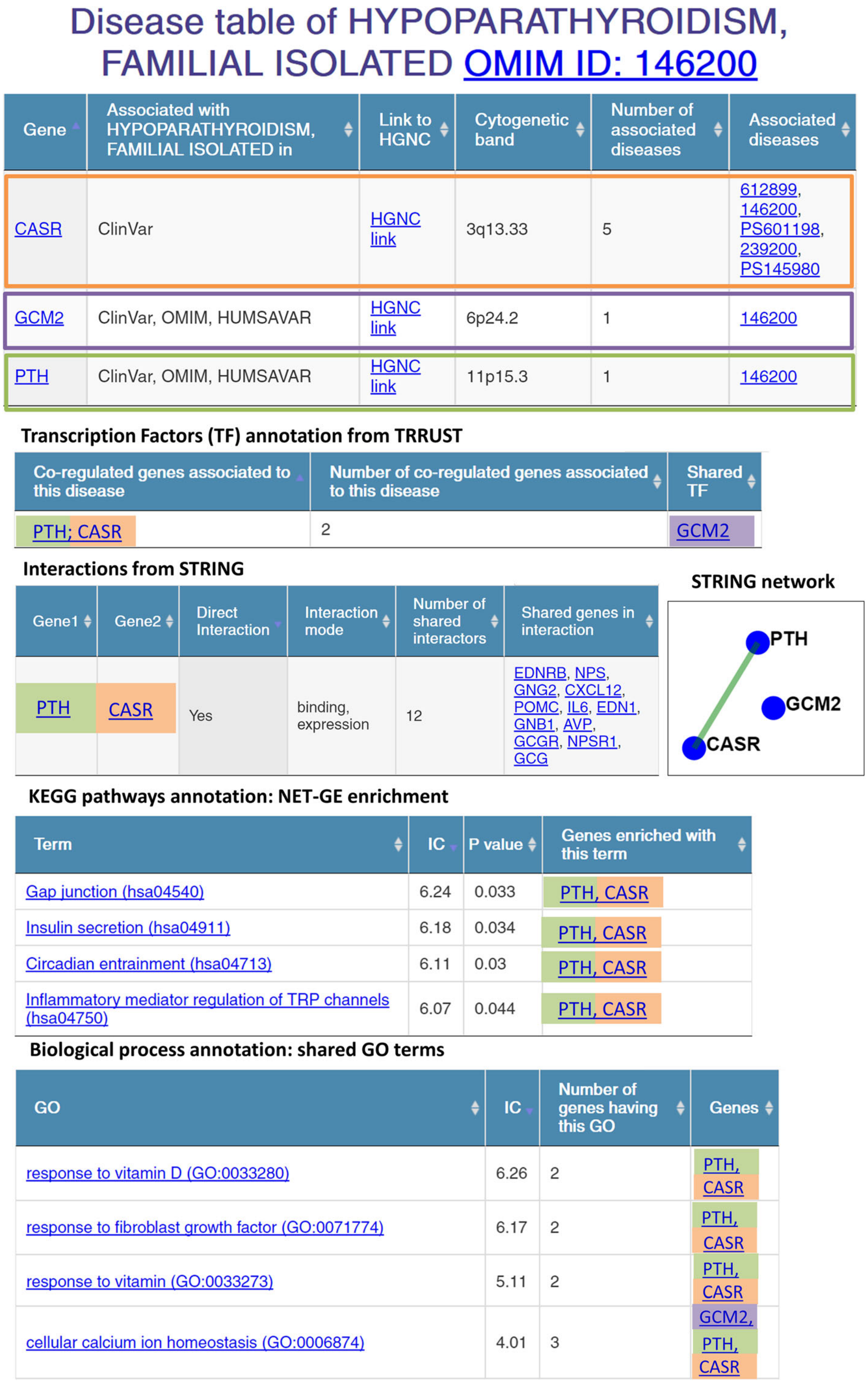
In eDGAR the familial isolated hypoparathyroidism (OMIM 146200) is associated with three different genes: GCM2 and PTH (both reported in OMIM, ClinVar and Humsavar) and CASR (reported only in ClinVar). CASR is an extracellular calcium-sensing receptor whose activity is mediated by G-proteins, PTH is 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



Disease table of HYPOPARATHYROIDISM, FAMILIAL ISOLATED OMIM ID: 146200

Gene	Associated with HYPOPARATHYROIDISM, FAMILIAL ISOLATED in	Link to HGNC	Cytogenetic band	Number of associated diseases	Associated diseases
CASR	ClinVar	HGNC link	3q13.33	5	612899, 146200, PS601198, 239200, PS145980
GCM2	ClinVar, OMIM, HUMSAVAR	HGNC link	6p24.2	1	146200
PTH	ClinVar, OMIM, HUMSAVAR	HGNC link	11p15.3	1	146200

Transcription Factors (TF) annotation from TRRUST

Co-regulated genes associated to this disease	Number of co-regulated genes associated to this disease	Shared TF
PTH; CASR	2	GCM2

Interactions from STRING

Gene1	Gene2	Direct Interaction	Interaction mode	Number of shared interactors	Shared genes in interaction
PTH	CASR	Yes	binding, expression	12	EDNRB, NPS, GNG2, CXCL12, POMC, IL6, EDN1, GNB1, AVP, GCGR, NPSR1, GCG

STRING network

KEGG pathways annotation: NET-GE enrichment

Term	IC	P value	Genes enriched with this term
Gap junction (hsa04540)	6.24	0.033	PTH, CASR
Insulin secretion (hsa04911)	6.18	0.034	PTH, CASR
Circadian entrainment (hsa04713)	6.11	0.03	PTH, CASR
Inflammatory mediator regulation of TRP channels (hsa04750)	6.07	0.044	PTH, CASR

Biological process annotation: shared GO terms

GO	IC	Number of genes having this GO	Genes
response to vitamin D (GO:0033280)	6.26	2	PTH, CASR
response to fibroblast growth factor (GO:0071774)	6.17	2	PTH, CASR
response to vitamin (GO:0033273)	5.11	2	PTH, CASR
cellular calcium ion homeostasis (GO:0006874)	4.01	3	GCM2, PTH, CASR

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.

3 eDGAR+

3.1 Introduction

The annotation procedure is fundamental, especially for genes and proteins related to diseases that may share feature of interest to direct the research and find new clues for therapies. To overcome the problem of retrieving the comprehensive annotation for genes associated with diseases we provided eDGAR (Babbi G et al, 2017), an online resource of gene-disease associations with a specific focus of the annotated relations among genes involved in the same disease. We derived the data from different online resources (OMIM (Amberger JS et al, 2015), ClinVar (Landrum MJ et al, 2016), and Humsavar (The UniProt Consortium, 2017)), collecting 5,729 associations for 2,672 diseases and 3,658 genes.

We built a pipeline for the annotation of genes retrieving information form many ontologies and databases (see chapter 2).

We are now preparing the new version of eDGAR, eDGAR+, besides an updated version of the data on annotated gene-disease associations, it contains new source of information, including a subset of 8,811 curated gene-disease associations from DisGeNET.

DisGeNET (Piñero J et al, 2017) is one of the largest collections of gene-disease association studies. The current version of DisGeNET (v5.0) contains 561,119 gene-disease associations, between 17,074 genes and 20,370 diseases, disorders, traits, and clinical or abnormal human phenotypes, and including also a collection of 135,588 variant-disease associations, between 83,002 SNPs and 9,169 diseases and phenotypes.

Moreover, we include in eDGAR+ the information on variants on genes and proteins, with a specific interest on the variants related to diseases. Among the various resources of features for the annotation, we include Human Protein Atlas (HPA, Uhlén M et al, 2015), describing the tissue of expression of genes and proteins.

We believe that eDGAR+ is one of the most comprehensive resources for retrieving the annotation of genes and proteins related to diseases insurgence, helping researchers in directing their analysis highlighting the features shared by variants, genes and proteins associated with the same malady.

3.2 Methods

3.2.1 Data collection

The updating of eDGAR+ is currently running; we provide here some preliminary statistics. We collect the data from 4 different resources: UniProt, ClinVar, OMIM and DisGeNET (Table 1). To standardize data we map every gene to the current version of the HGNC identifiers, controlled with the HGNC multi-symbol checker (Yates B et al, 2017). We also consider the mapping to the relative Ensembl identifiers (Zerbino DR et al, 2018) and the associated proteins on UniProt-SwissProt Accession numbers in order to help the user in recovering the information of interest with an accessibility that consider different international standards.

Database	**Associations**	**Genes**	**Diseases**
UniProt	5,673	3,804	4,938
ClinVar	7,630	4,434	5,018
OMIM	6,145	4,228	5,408
DisGeNET	8,811	4,909	4,253
TOTAL	**12,560**	**6,580**	**5,574**

Table 1: Statistics on the resources of gene-disease associations.

Finally, considering unique gene-disease associations removing redundancy among the databases, we obtain 12,560 associations among 6,580 genes and 5,574 diseases (Table 2).

Entries type	**eDGAR # entries**	**eDGAR+ # entries**	**% increment**
Gene-disease associations	5,729	12,560	+119%
Genes	3,658	6,580	+80%
Diseases	2,672	5,574	+109%

Table 2: Comparison among eDGAR and eDGAR+ data on gene-disease associations.

3.2.2 Annotation procedure

The annotation procedure is central in the eDGAR+ approach to the problem of gene-disease associations.

Each gene page in eDGAR+ contains first the extended gene name and a *Gene-disease associations table* stating the associations with diseases. If the gene is protein coding and annotated with a 3D-structure in PDB, we report a *Structural analysis* of the PDB with the best resolution, highlighting variants and region of interest of the protein.

We then report the *Annotation of the gene* collecting the features retrieved with a pipeline comprising many resources and we provide the different features in separated tables In particular, in the new version eDGAR+, we include tables about variants (dbSNP identifiers (Sherry ST et al, 2001) annotated with the relative amino acid change and the disease association) and expression in normal tissues – HPA (highlighting the tissue of expression, the cell type and the level of expression)

We provide also disease pages, presenting first the disease name and OMIM identifier, then *Gene-disease associations table* stating the associations with genes and variants. Finally, for the diseases related at least to two genes, we analyse the *Relations among genes.*

We use NET-GE (Di Lena P et al, 2015 and Bovo S et al, 2016) a tool for standard and network gene enrichment, for retrieving new annotations for the set of genes associated to each disease, analysing the significantly enriched terms for each gene set considering Gene Ontology, KEGG and REACTOME.

All these features are reported in tables and networks, with the aim of building unified frameworks collecting the current knowledge on genes associated to the same disease.

3.2.3 Variant standardization

Variants in protein coding regions can be defined with specific and unique identifiers (e.g. dbSNP identifiers, Sherry ST et al, 2001), or as the amino acidic change on a specific protein sequence, or only as a mutation in a specific position of the genomes. When we retrieve variants from different resources, they may be defined using different approach: to compare them and to create a coherent database it is important to standardize variant definition. Dealing with variant standardization is a problematic issue: relating the genetic variations to the protein variant in its different isoforms is quite complicated on a large-scale computational approach, especially when we are interested in the keeping correct disease associations. We derive gene variations and protein variants from three different resources: DisGeNET (81,561 variants), UniProt (77,917 variants) and Intact (Ochard S et al, 2013; 11,694 variants).

The annotation of variants with international standards and identifiers is still an issue: among the three different resources, DisGeNET has dbSNP identifiers associated with 82% of the variants, being the highly standardized resources among those examined. UniProt collected dbSNP identifiers for the 74% of the variants; Intact generally does not provide dbSNP for variants. To compare and select the variants for eDGAR+ it is necessary to homogenize these data. To homogenize our data, first we have to answer to the question: what identify univocally a variant? In eDGAR, a variant is a combination of a variation on a gene identified by a dbSNP identifier and the relative amino-acidic substitution.

We then map all the variants with the tool for variant annotation provided by Ensembl (Zerbino DR et al, 2018). Finally, we obtain 107,226 variants in 12,541 genes. Over the 107,226 variants, 61,885 are associated with 4,217 diseases, for a total for 73,882 variant-disease associations over 3,309 genes.

3.3 Results

3.3.1 Diseases and genes annotation

The current version of eDGAR+ contains 12,560 gene-disease associations, involving 6,580 genes, increasing the number of genes associated with disease of 80% respect to the previous version.

We report in Table 3 the general statistics for gene annotation, comparing the precedent version of eDGAR with eDGAR+.

For each category of features, the total number of annotated genes is increased, while the percentage of annotated genes over the total genes in the webserver remains similar in the two versions of eDGAR. The percentage of annotated genes depends also on the updating of the relative annotation primary sources. We are currently revising our annotation pipeline with the most updated version of the annotation database to increase the number of annotated genes. We are also considering other primary resources for annotation to enlarge the number of annotated genes thanks to the advantage of retrieving the annotation from many different sources.

3.3.2 Data visualization

eDGAR+ has improved data visualization, regarding protein structures and networks of shared features. In each gene page, if the gene codes for a protein with known protein structure, the protein structure is shown with all the variants highlighted. Moreover, for each heterogeneous or polygenic disease, we are building networks summing up all the features shared by the associated genes. Each network shows a different feature, like co-expression in the same tissue, expression regulation by shared TFs to protein-protein interactions and co-occurrence in protein complexes. The possibility of analysing many layers of information in a glance just looking at different networks lets the user grasp the relations among genes associated to the same disease in term of physical interactions, biological pathways, transcription regulation etc.

The new release has a new fancy user-friendly style, collecting information along the web-pages in a well-organized way.

Entry type	**Genes** **in eDGAR**	**Genes** **in eDGAR+**
Total number of genes with SwissProt	3,628 (100.0%)	6,085 (100.0%)
with PDB	1,682 (46.4%)	2,693 (44.3%)
being in a tandem repeat	381 (10.5%)	702 (11.7%)
being TF	253 (7.0%)	397 (6.5%)
being regulated by TF	942 (26.0%)	1,288 (21.2%)
being regulated by TF - not TF	783 (21.6%)	1,048 (17.3%)
with GO BP	3,538 (97.5%)	5,797 (95.6%)
with GO MF	3,419 (94.2%)	5,778 (95.3%)
with GO CC	3,576 (98.6%)	5,915 (97.6%)
with KEGG	2,057 (56.7%)	3,014 (49.5%)
with REACTOME	2,278 (62.8%)	3,897 (64.0%)
with CORUM	714 (19.7%)	812 (13.3%)
with CENSUS	696 (19.2%)	504 (8.3%)
with BIOGRID physical	3,307 (91.3%)	5,488 (90.2%)
with BIOGRID genetic	351 (9.7%)	593 (9.7%)
with STRING	2,746 (75.7%)	5,210 (85.6%)
in membrane	2,059 (56.8%)	2,940 (48.3%)

Table 3: A comparison of the level of gene annotation in eDGAR and in eDGAR+.

3.4 Conclusions

eDGAR+ is the new version of the webserver DGAR, collecting new data from different primary resources, new features and sources of gene annotation. It will be soon available online with an improved web interface and a scientific paper describing the new version is under preparation.

Users and researchers in the field may take advantages of the networks of the shared features of genes related to the same disease to direct their experiments, analysing new clues on possible related pathways and significant interactions among gens associated to diseases.

4 PhenPath

4.1 Contribution to the state of the art

The co-occurrence of different phenotypes complicates the understanding of the underlying molecular mechanisms leading to disease insurgence. We propose a resource for the identification of molecular mechanisms underpinning different phenotypes, to ascribe the ensemble of phenotypes to a small number of possibly altered biological functions.

Here we present PhenPath, a webserver of disease-phenotype relations with information at the molecular level, comprising a tool able to retrieve diseases, genes and functional annotations associated to a given set of phenotypes.

We propose our resource for directing scientific efforts, speeding up the diagnosis and retrieving new possible association among biological processes and diseases. We believe that biotechnologists, physicians and medical researchers may find in PhenPath a useful resource of information, especially when studying complex and rare diseases.

4.2 General information on the paper

The presented paper is now under review in the Journal BMC Genomics.

Authors: Giulia Babbi, Pier L. Martelli and Rita Casadio

Title: PhenPath: a tool for characterizing biological functions underlying different phenotypes

Journal: BMC Genomics

Submission year: 2018

Impact Factor: 3.73

Quartile and subject: 1st quartile in Biotechnology and Applied Microbiology

PhenPath: a tool for characterizing biological functions underlying different phenotypes

Paper submitted to BMC Genomics in 2018, under review

Authors

Giulia Babbi[1,2], Pier Luigi Martelli[1,3,]* and Rita Casadio[1,3,4]

* Corresponding author

Affiliations

[1] University of Bologna, FABIT, Via San Donato 15, 40126 Bologna, Italy.

[2] University of Bologna, Department of BIGEA, Piazza di Porta S. Donato, 1, 40126 Bologna, Italy.

[3] University of Bologna, CIG, Interdepartmental Center “Luigi Galvani” for integrated studies of Bioinformatics, Biophysics and Biocomplexity, Via G.Petroni 26, 40126 Bologna, Italy.

[4] CNR, Institute of Biomembrane and Bioenergetics (IBIOM), Via Giovanni Amendola 165/A - 70126 Bari Italy.

E-mail addresses:

giulia.babbi3@unibo.it (GB)

pierluigi.martelli@unibo.it (PLM)

rita.casadio@unibo.it (RC)

Abstract

Background. Many diseases are associated with complex patterns of symptoms and phenotypic manifestations. Parsimonious explanations aim at reconciling the multiplicity of phenotypic traits with the perturbation of one or few biological functions. For this, it is necessary to characterize human phenotypes at the molecular and functional levels, by exploiting gene annotations and known relations among genes, diseases and phenotypes. This characterization makes it possible to implement tools for retrieving functions shared among phenotypes, co-occurring in the same patient and facilitating the formulation of hypotheses about the molecular causes of the disease.

Results. We introduce PhenPath, a new resource consisting of two parts: PhenPathDB and PhenPathTOOL. The former is a database collecting the human genes associated with the phenotypes described in Human Phenotype Ontology (HPO) and OMIM Clinical Synopses. Phenotypes are then associated with biological functions and pathways by means of NET-GE, a network-based method for functional enrichment of sets of genes. The present version considers only phenotypes related to diseases. PhenPathDB collects information for 18 OMIM Clinical synopses and 7,137 HPO phenotypes, related to 4,292 diseases and 3,446 genes. Enrichment of Gene Ontology annotations endows some 87.7%, 86.9% and 73.6% of HPO phenotypes with Biological Process, Molecular Function and Cellular Component terms, respectively. Furthermore, 58.8% and 77.8% of HPO phenotypes are also enriched for KEGG and Reactome pathways, respectively. Based on PhenPathDB, PhenPathTOOL analyses user-defined sets of phenotypes retrieving diseases, genes and functional terms which they share. This information can provide clues for interpreting the co-occurrence of phenotypes in a patient.

Conclusions. The resource allows finding molecular features useful to investigate diseases characterized by multiple phenotypes, and by this, it can help researchers and physicians in identifying molecular mechanisms and biological functions underlying the concomitant manifestation of phenotypes. The resource is freely available at http://edgar.biocomp.unibo.it/phenpath/.

Keywords

Phenotype, diseases, molecular pathway, biological process, enrichment

1 Background

Co-occurrence of different phenotypes often associated with symptom complexes hampers the understanding of the molecular mechanisms which characterize diseases and their insurgence [1]. Furthermore, the analysis of epidemiological data reveals that different phenotypes associated with specific diseases frequently co-occur in the same individuals during their lifespan [2,3]. In both situations, highlighting functional molecular mechanisms underlying disease insurgence and progression offers a way to understand possible associations between phenotypes and diseases. In the context of personalized medicine, this approach can be in principle adopted to analyze phenotypes that are peculiar of every single patient. The challenge is to reconcile the ensemble of phenotypes with a small number of possibly altered biological functions. Along this line, Brodie et al. (2014), [4], reported a large-scale analysis of Genome Wide Studies (GWAS) results demonstrating that phenotypes can be significantly associated to specific pathways, where SNPs cluster, depending on the specific disease.

Several resources are presently available to exploit data for associating phenotypes to diseases. The Phenotype-Genotype Integrator (PheGenI) [5], merges data from genome-wide association study (GWAS) stored at the National Human Genome Research Institute (NHGRI, http://www.genome.gov) with several databases housed at the National Center for Biotechnology Information (NCBI), including Gene, dbGaP, OMIM, eQTL and dbSNP (https://www.ncbi.nlm.nih.gov/gap/phegeni). This phenotype-oriented resource aims at facilitating prioritization of variants, from GWAS studies, for generation of biological hypotheses and it is quite useful for a search based on chromosomal location, gene, SNP, or phenotype. Search results include annotated tables of SNPs, genes and association results, a dynamic genomic sequence viewer, and gene expression data.

For the molecular diagnosis of rare genetic diseases, the recently developed Phenopolis ([6], https://phenopolis.org/about) is an open platform for harmonization and analysis of sequencing and phenotype data. The platform offers per phenotype, a prioritized list of genes, based on known association and gene enrichment analysis.

Other resources provide associations between diseases and phenotypes, including the Human Phenotype Ontology (HPO) [7] and the OMIM Clinical synopses [8]. Exploiting these associations, methods have been developed to cluster different diseases through shared phenotypes. In particular, the Phenotypic Disease Network [9] focuses on phenotypic links among co-occurring diseases to address the comorbidity problem.

The Phenomizer tool [10], provided by the Human Phenotype Ontology consortium, analyzes lists of phenotypes/symptoms with the aim of assisting the clinical workflow and providing diagnoses.

While many resources focus on the relationship among phenotypes, diseases and genes, little is known about the relevance of molecular functions and functional processes underlying the occurrence of phenotypes.

The goal of our research is to supplement disease-phenotype associations with information at the molecular level. To this aim, here we describe a resource (PhenPath) able to retrieve diseases, genes and functional annotations associated with a given set of phenotypes.

Our resource builds on supplementing known disease-phenotype links with the molecular information on the association between genes and diseases. This last knowledge is stored in different databases, including Humsavar [11], ClinVar [12] and OMIM [8], previously integrated by DisGeNet [13] and by eDGAR [14], which exploits also functional annotations.

Phenotype-disease and disease-gene relationships can be represented with a graph and, after collapsing the disease layer, direct associations between genes and phenotypes emerge. Furthermore, efficient enrichment procedures help in associating groups of genes to specific biological processes and/or metabolic pathways, endowing the group with statistically validated functional annotations. Among other procedures, our NET-GE [15] exploits proximity relationships among genes as derived from gene-gene interaction networks [16], and here it is adopted to functionally annotate phenotype-related genes. Considering the relationship among diseases, genes and functions, and the association among diseases and phenotypes, PhenPath allows the association of phenotypes to biological processes and pathways, reconciling their manifestation with molecular events.

2 Results

We implemented a new resource, PhenPath, to help researchers and physicians in studying complex diseases, characterized by one or multiple phenotypes.

PhenPath consists of two parts: a database collecting relationships among genes, diseases, phenotypes and biological functions (PhenPathDB), and a tool allowing to retrieve genes, diseases and biological functions shared by a group of phenotypes, provided by the user (PhenPathTOOL).

2.1 PhenPathDB

PhenPathDB is generated considering the three main steps described in the following: i) a phenotype-disease association procedure; ii) a disease-gene association procedure; and iii) a phenotype functional annotation derived by collapsing the gene layer, after an enrichment procedure of the functional annotation of the different disease-associated genes. Functional annotations consider Gene Ontology [17] terms of the three main roots (Molecular Function, Biological Process and Cellular Component), KEGG [18] and Reactome [19] pathways**.**

2.1.1 Phenotype-disease association

PhenPathDB builds upon the known associations among phenotypes, diseases and genes***.*** PhenPathDB includes information about the following phenotypic terms (Table 1): i) 18 phenotypic general categories from the OMIM Clinical Synopsis [8], which classifies 4,165 OMIM diseases, grouped according to the affected human body districts; ii) 7,173 phenotypic terms from HPO [7], annotating 4,292 OMIM diseases (59% of the 12,111 phenotypic terms of HPO, which are disease-associated). HPO Ontology includes five main sub-ontologies (Phenotypic Abnormalities, Clinical Modifier, Clinical Course, Mode of Inheritance, and Frequency). Specific terms, called leaf terms, are 3,837 and they annotate at the deepest level 4,023 diseases. The most populated sub-ontology is Phenotypic Abnormalities, which includes 78% of the HPO disease-related phenotypes with 24 main categorizations referring to body districts and physiological functions. They expand into 5,661 terms associated with 4,273 diseases, of which 3,802 are leaf terms annotating 3,721 diseases (Table 1).

Table 1. Phenotypic terms included in PhenPathDB

Ontology	**Phenotypic terms (#)**	**Diseases associated to Phenotypic terms (#)**	**Phenotypic leaf terms (#)**	**Diseases associated to Phenotypic leaf terms (#)**
OMIM Clinical Synopsis *	18	4,165	-	-
HPO	7,173	4,292	3,837	4,023
HPO sub-ontology Phenotypic Abnormalities §	5,661	4,273	3,802	3,721

* OMIM Clinical Synopsis is not organized in a graph, and as a consequence, it does not contain distinction among root, intermediate and leaf terms. § HPO Phenotypic Abnormalities are the subset of HPO, organized according to body districts and physiological functions into 24 different main terms.

Most of the OMIM diseases are associated with more than one HPO leaf term (Figure 1). Only 15% of the diseases are associated with one phenotype, and about half of the diseases are associated with 5 or more phenotypes. The extreme case is the Rubinstein-Taybi syndrome that is annotated with 48 HPO leaf terms.

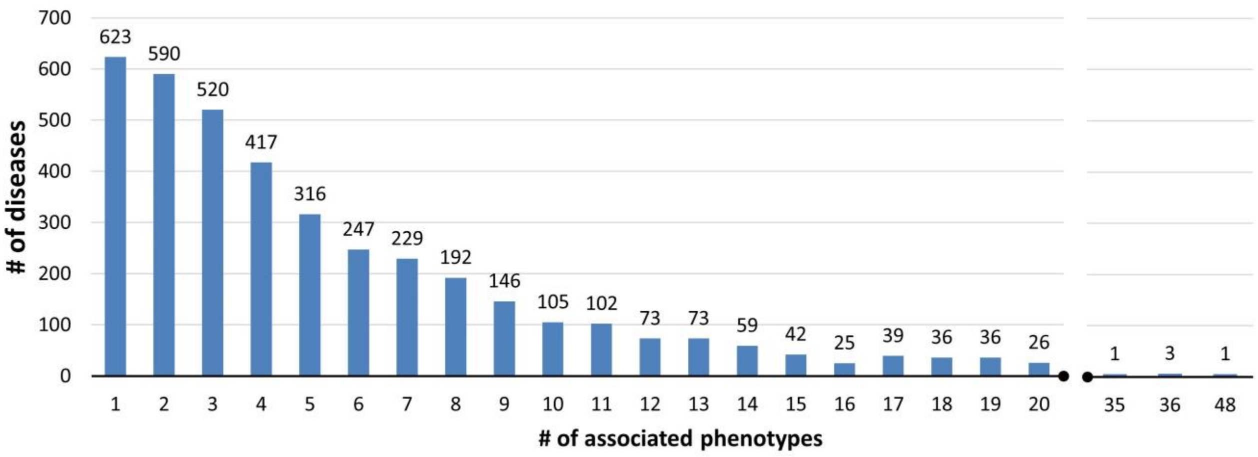


Figure 1: OMIM diseases as a function of associated HPO phenotypes. Data include 3,837 HPO phenotypes (leaves of the HPO ontology) associated with 4,023 OMIM diseases (Table 1, second row). Only 623 diseases (15%) are associated with a single phenotype, while about half of the diseases (47%) are associated with 5 or more phenotypes. Rubinstein-Taybi syndrome has the maximum number of associated HPO phenotypes (48, considering only leaves of the HPO graph).

2.1.2 Disease-gene association

Each phenotype-disease link described in 2.1.1 is supplemented with a set of genes, by exploiting the gene-disease relationships reported in eDGAR [15]. Figures 2 and 3 show the number of diseases (blue bars) and genes (red bars) associated to the 18 terms of the OMIM Clinical Synopsis and to the 24 main categories of the HPO Phenotypic Abnormalities sub-ontology, respectively. With eDGAR, Phenotypic OMIM Clinical Synopsis terms and HPO terms are associated with 3,230 and 3,446 genes, respectively.

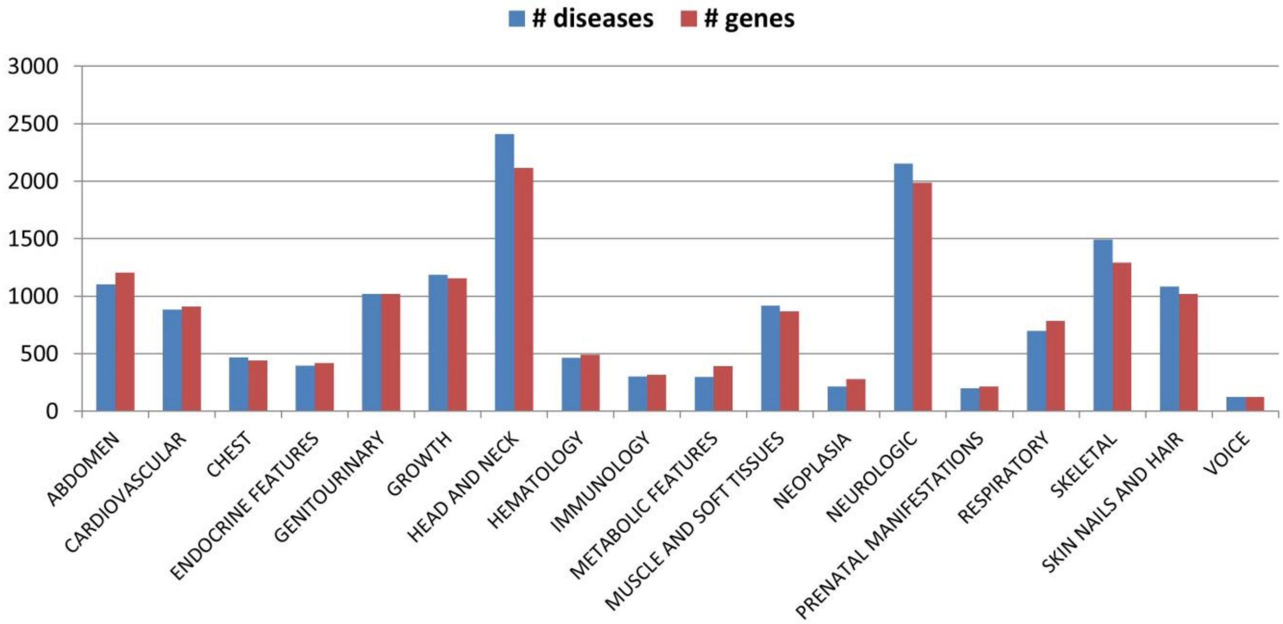


Figure 2: Number of diseases and genes associated with OMIM Clinical Synopsis terms. Blue bars (diseases); red bars (genes).

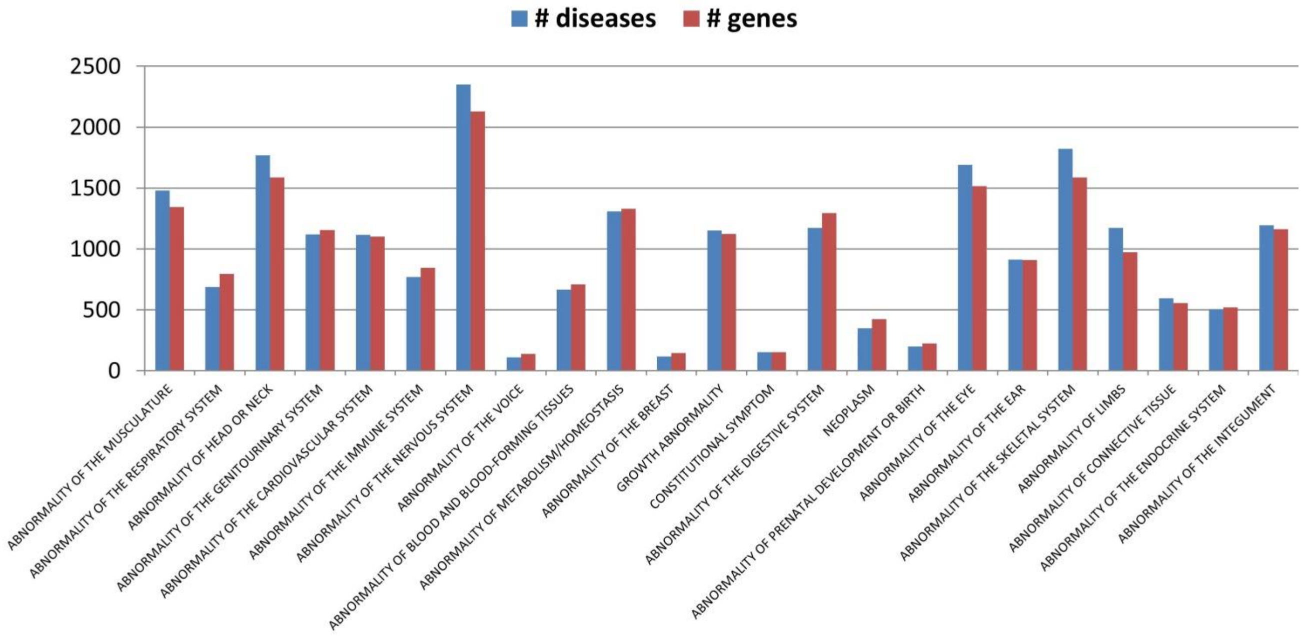


Figure 3: Number of diseases and genes associated with the 24 main categories of HPO Phenotypic Abnormalities sub-ontology. The 24 roots refer to anatomic districts and physiological functions. Blue bars (diseases); red bars (genes).

2.1.3 Functional annotation of phenotypes

According to our procedure, any phenotype links one or more disease/s, which are associated with specific genes. Any set of genes can be functionally characterized by adopting an enrichment procedure. Here, we adopt NET-GE, a tool for the functional enrichment analysis of genes (two or more) [15]. NET-GE considers the relationships among annotated genes as described in the STRING interactome, from which it derives a function-specific gene module to be used as a basis for the overrepresentation analysis. This procedure takes into consideration Gene Ontology terms, KEGG and Reactome pathways.

Following enrichment, most phenotypes included in Table 1, are annotated with Gene Ontology (GO) terms, as shown in Table 2. In particular, 87.7% and 86,9% of HPO terms are enriched with GO terms of Biological Process (BP) and Molecular Function (MF), respectively.

Table 2: Functional annotation of HPO terms.

FUNCTIONAL ANNOTATION	Phenotypes (#)	HPO terms (%)	Non redundant functional terms (#)
with GO BP	6256	87.7%	6838 GO BP
with GO MF	6202	86.9%	2211 GO MF
with GO CC	5254	73.6%	946 GO CC
with KEGG	4198	58.8%	326 KEGG
with REACTOME	5550	77.8%	1369 REACTOME

Statistics refer to 7,137 HPO terms comprised in PhenPath and associated with 4,292 diseases and 3,446 genes. Terms included in PhenPath comprise 59% of the 12,111 terms listed in HPO. BP: Biological Process; MF: Molecular Function; CC: Cellular Component. #: number of.

2.2 PhenPathDB interface

PhenPathDB organizes associations among phenotypes, diseases, genes and functional annotations in two major entering tables: OMIM Clinical Synopsis and HPO Phenotypic Abnormality (http://edgar.biocomp.unibo.it/phenpath/). Each Table contains links to our results grouped into:

i) *general analysis*, which for each phenotypes, lists diseases, associated genes and the functional characterization derived from the enrichment procedure;

ii) *intersection analysis*, which allows to derive features shared between two phenotypes, highlighting the common diseases, genes and functional annotations.

More specifically, *general analysis* reports diseases and genes associated with the phenotype, the annotation obtained with NET-GE, along with the Bonferroni-corrected p-value of the enrichment procedure, and the Information Content (IC) evaluating the specificity of the term (see Methods section for further details). The page lists also the genes accounting for the enrichment of each functional term and the associated diseases and describing the association of specific functional terms with the phenotype under consideration. Diseases and genes are linked to the corresponding OMIM and Human Gene Nomenclature Committee (HGNC) [20] entries

The *intersection analysis* is based on the pre-computed shared features of pairs of phenotypes out of the same ontology (18 categories of OMIM Clinical Synopsis or 24 main categories of HPO Phenotypic

Abnormalities sub-ontology). Furthermore, shared GO terms, KEGG and Reactome pathways, enriched for both groups of associated genes, are listed. For each functional term, the IC value is reported as well as the Bonferroni-corrected p-values of the two enrichment procedures. The phenotype page provides also the list of genes associated with a particular functional term.

It is possible to access the database either by browsing the PhenPathDB page or by searching for specific phenotypes in the Search page. For HPO, the 24 main categories of the Phenotypic Abnormalities are present in the browsing page, and all terms can be retrieved with a search.

2.3 PhenPathTOOL

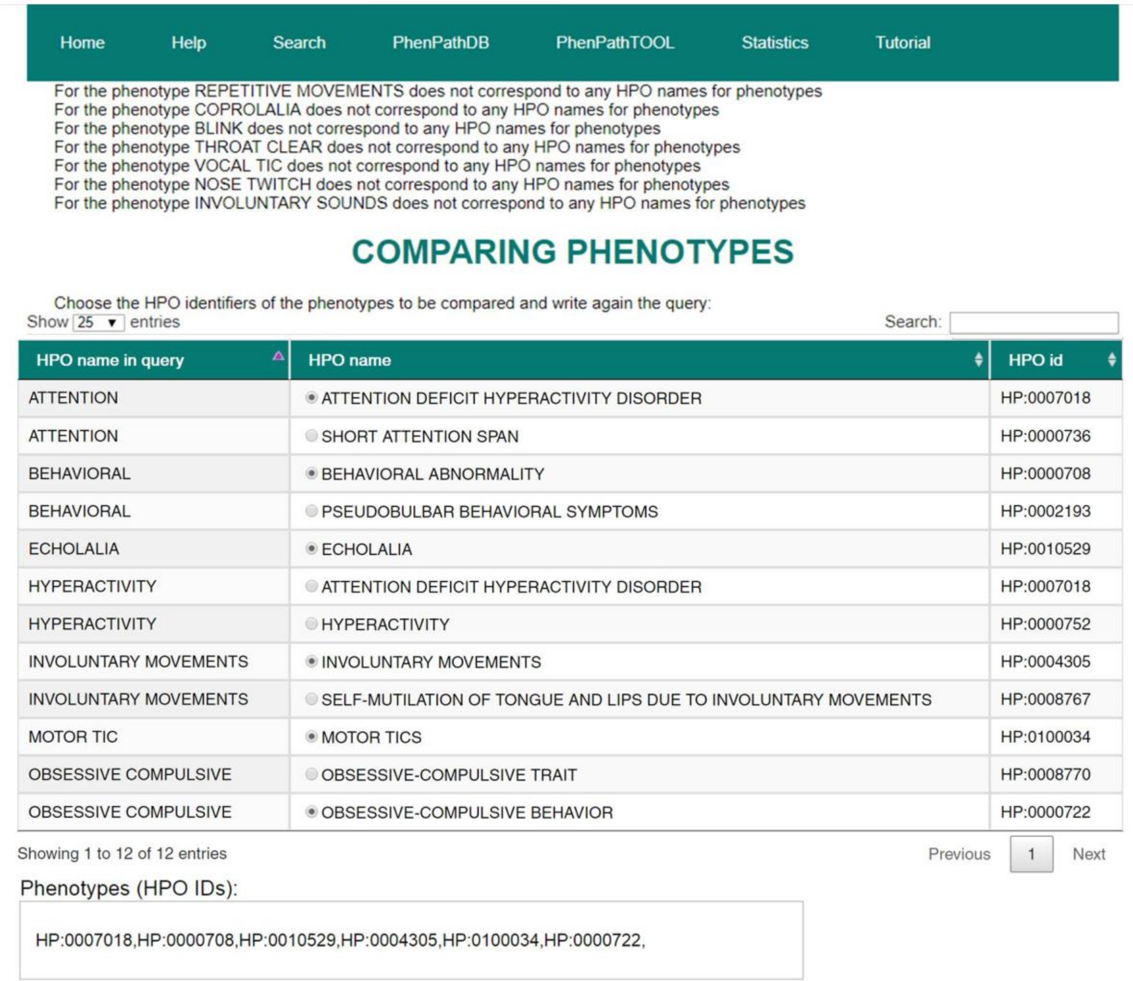
PhenPathTOOL is a web application that, given a set of phenotypes, retrieves the shared diseases, genes and functional terms. PhenPathTOOL is user-friendly, accepting as input HPO IDs as well as names of phenotypes. The intersection is computed in real-time. PhenPathTOOL allows investigating the relationship among groups of phenotypes at different levels. Firstly, it retrieves whether there is an intersection among the lists of diseases associated with the input phenotypes. In this way, it highlights when the phenotype co-occurrence is already known and points towards specific diseases. Occasionally, when input phenotypes do not share common diseases, PhenPathTOOL can retrieve shared genes, possibly related to their concomitant manifestation. Furthermore, even when phenotypes do not share genes, they may share the enriched biological functions (GO terms, KEGG and Reactome pathways), accounting for a common mechanism. The interface reports in different tables the lists of shared GO terms, KEGG and REACTOME pathways, obtained as described above. Each table lists the IC of the term, as well as the Bonferroni-corrected p-value for each association (see Methods for further details).

3 Discussion

3.1 Study case: Tourette syndrome

The first example describes the use of PhenPathTOOL for retrieving a characterized disease starting from a list of phenotypes and the possibility to enrich the annotation of involved biological functions. Tourette syndrome is a neurobehavioral disorder that causes motor and vocal tics associated with behavioral abnormalities, like attention-deficit–hyperactivity disorder and obsessive-compulsive disorder [21]. Possible symptoms include involuntary or semi-voluntary movements or sounds, repetitive movements, blinking, nose twitching, throat clearing to echolalia or coprolalia.

We searched with PhenPathTOOL the typical phenotypic traits of the Tourette syndrome, using a plain list of phenotype names ("motor tic, vocal tic, behavioral, attention, hyperactivity, obsessive-compulsive, involuntary movements, involuntary sounds, repetitive movements, blink, nose twitch, throat clear, echolalia, coprolalia"). The interface presents a selectable list of HPO terms whose names contain the input terms (Figure 4).



For the phenotype REPETITIVE MOVEMENTS does not correspond to any HPO names for phenotypes
For the phenotype COPROLALIA does not correspond to any HPO names for phenotypes
For the phenotype BLINK does not correspond to any HPO names for phenotypes
For the phenotype THROAT CLEAR does not correspond to any HPO names for phenotypes
For the phenotype VOCAL TIC does not correspond to any HPO names for phenotypes
For the phenotype NOSE TWITCH does not correspond to any HPO names for phenotypes
For the phenotype INVOLUNTARY SOUNDS does not correspond to any HPO names for phenotypes

COMPARING PHENOTYPES

Choose the HPO identifiers of the phenotypes to be compared and write again the query:

Show 25 ▾ entries Search:

HPO name in query	HPO name	HPO id
ATTENTION	● ATTENTION DEFICIT HYPERACTIVITY DISORDER	HP:0007018
ATTENTION	○ SHORT ATTENTION SPAN	HP:0000736
BEHAVIORAL	● BEHAVIORAL ABNORMALITY	HP:0000708
BEHAVIORAL	○ PSEUDOBULBAR BEHAVIORAL SYMPTOMS	HP:0002193
ECHOLALIA	● ECHOLALIA	HP:0010529
HYPERACTIVITY	○ ATTENTION DEFICIT HYPERACTIVITY DISORDER	HP:0007018
HYPERACTIVITY	○ HYPERACTIVITY	HP:0000752
INVOLUNTARY MOVEMENTS	● INVOLUNTARY MOVEMENTS	HP:0004305
INVOLUNTARY MOVEMENTS	○ SELF-MUTILATION OF TONGUE AND LIPS DUE TO INVOLUNTARY MOVEMENTS	HP:0008767
MOTOR TIC	● MOTOR TICS	HP:0100034
OBSESSIVE COMPULSIVE	○ OBSESSIVE-COMPULSIVE TRAIT	HP:0008770
OBSESSIVE COMPULSIVE	● OBSESSIVE-COMPULSIVE BEHAVIOR	HP:0000722

Showing 1 to 12 of 12 entries

Previous 1 Next

Phenotypes (HPO IDs):

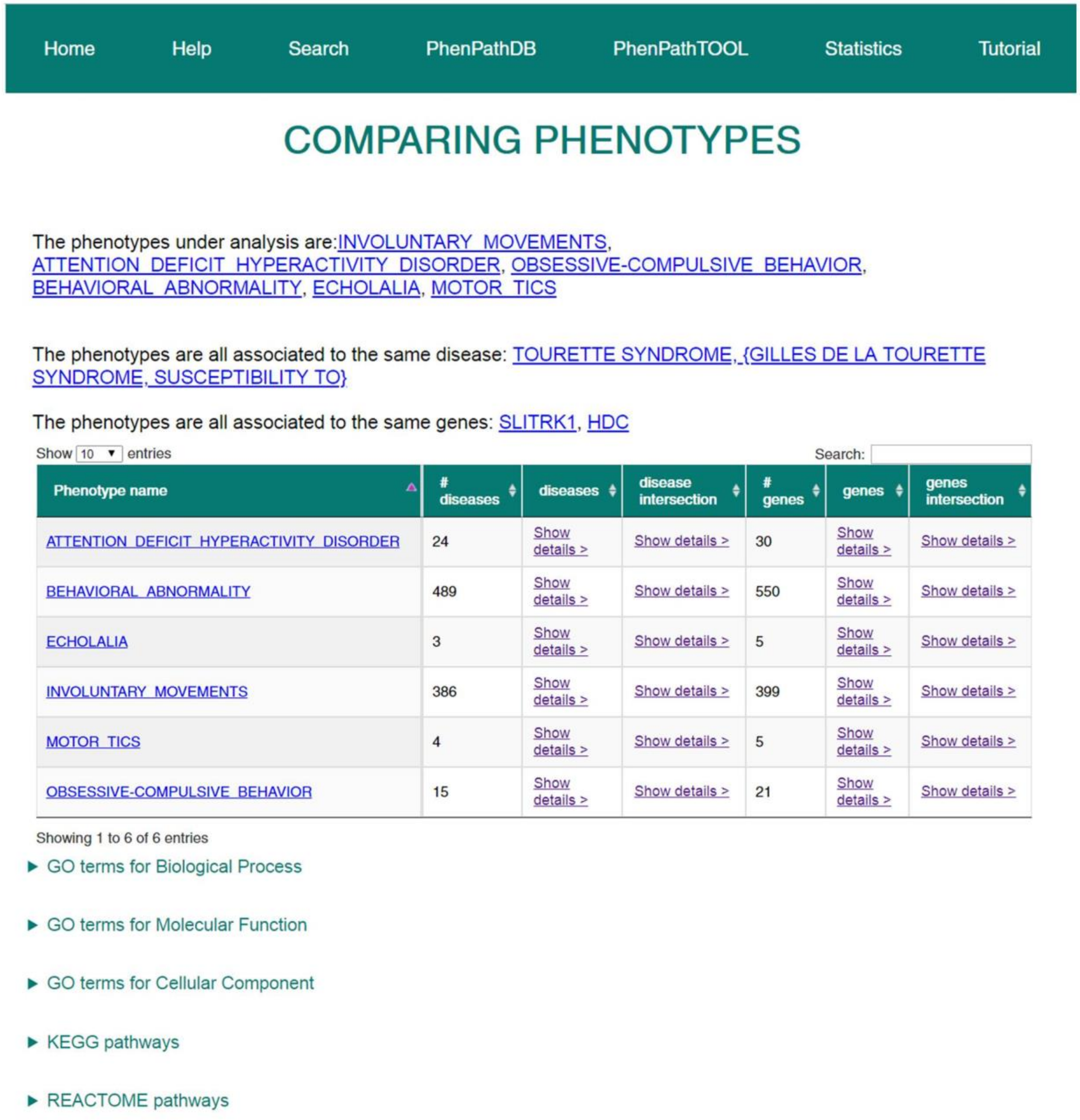
HP:0007018,HP:0000708,HP:0010529,HP:0004305,HP:0100034,HP:0000722,

Figure 4: Selection of phenotypes in PhenPathTOOL. After searching with a list of different names, the web interface shows all the names that do not correspond to any HPO identifier and then a table with all the HPO terms matching the input. The user may then select the most appropriate phenotypes to be analyzed.

In this particular study case, we selected, among the proposed HPO terms, the 6 that better describe the phenotypes of Tourette syndrome to perform further analyses: "attention deficit hyperactivity disorder

(HP:0007018), behavioral abnormality (HP:0000708), echolalia (HP:0010529), involuntary movements (HP:00043059), motor tics (HP:0100034), obsessive-compulsive behavior (HP:0000722)".

PhenPathTOOL returns the diseases and genes shared among the phenotypes, as long as the shared enriched pathways (GO terms, KEGG and REACTOME, Figure 5).



COMPARING PHENOTYPES

The phenotypes under analysis are:INVOLUNTARY_MOVEMENTS, ATTENTION_DEFICIT_HYPERACTIVITY_DISORDER, OBSESSIVE-COMPULSIVE_BEHAVIOR, BEHAVIORAL_ABNORMALITY, ECHOLALIA, MOTOR_TICS

The phenotypes are all associated to the same disease: TOURETTE SYNDROME, {GILLES DE LA TOURETTE SYNDROME, SUSCEPTIBILITY TO}

The phenotypes are all associated to the same genes: SLITRK1, HDC

Phenotype name	# diseases	diseases	disease intersection	# genes	genes	genes intersection
ATTENTION_DEFICIT_HYPERACTIVITY_DISORDER	24	Show details >	Show details >	30	Show details >	Show details >
BEHAVIORAL_ABNORMALITY	489	Show details >	Show details >	550	Show details >	Show details >
ECHOLALIA	3	Show details >	Show details >	5	Show details >	Show details >
INVOLUNTARY_MOVEMENTS	386	Show details >	Show details >	399	Show details >	Show details >
MOTOR_TICS	4	Show details >	Show details >	5	Show details >	Show details >
OBSESSIVE-COMPULSIVE_BEHAVIOR	15	Show details >	Show details >	21	Show details >	Show details >

Showing 1 to 6 of 6 entries

- ▸ GO terms for Biological Process
- ▸ GO terms for Molecular Function
- ▸ GO terms for Cellular Component
- ▸ KEGG pathways
- ▸ REACTOME pathways

Figure 5: PhenPathTOOL results. The figure shows the webpage of PhenPathTOOL after the analysis of 6 different HPO phenotypes. First, a list of the shared diseases and genes is reported. Then, a general table collects data on diseases and genes associated with each phenotype, allowing direct intersection. The last section reports the links to the analysis of GO terms, KEGG and Reactome pathways.

PhenPathTOOL correctly recognizes that the concomitance of phenotypes points to the Tourette syndrome, and to two genes (SLITRK1, HDC) that are associated with the disease [21]. Interestingly enough, only the intersection of functional terms shared by different phenotypes is able to retrieve relevant common shared annotations. 30 terms are shared by at least 4 phenotypes. Among them, besides the general annotations like “behavior”, “cognition” or “learning or memory”, there are interesting clues on more specific pathways such as “catecholamine metabolic process”. Interestingly, symptomatic therapies for the Tourette syndrome involve the control of neurotransmission from dopamine and adrenaline, which are members of the catecholamine family [22]. Although the pathogenesis of the disorder remains obscure, the catecholamine metabolic process pathway has already been studied in relation to the Tourette syndrome [23].

3.2 Study case: Obesity, Diabetes and Ovarian Cyst

Here PhenPathTOOL compares three phenotypes that, although not being related to a common disease, are often co-occurring: obesity, diabetes and ovarian cysts. Epidemiological studies report that women affected by polycystic ovarian syndrome, for which ovarian cysts is the main phenotypes, are often showing also obesity and diabetes phenotypes [24]. In particular, increasing evidence point to an increase of type 2 diabetes in women affected by polycystic ovarian syndrome [25].

We analyzed with PhenPathTOOLS the three co-occurring phenotypes: Obesity (HP:0001513), Diabetes Mellitus type II (HP:0005978) and Ovarian Cyst (HP:0000138). Routinely, the three terms refer to specific diseases: however, in HPO they indicate phenotypes associated to different disorders.

As expected, no disease is common to all the input phenotypes. Diabetes and obesity share 3 diseases: Prader-Willi Syndrome, Morbid obesity and spermatogenic failure, and Microcephalic osteodysplastic primordial dwarfism, type II. No disease links ovarian cysts to either obesity or diabetes.

The analysis at the gene level retrieves only one gene shared among the three phenotypes: PPARG, the Peroxisome proliferator-activated receptor gamma, a nuclear receptor involved in lipid uptake and adipogenesis. More genes are shared between pairs of phenotypes: NPP1, AKT2 between diabetes and obesity, HNF1A, INSR and PPP1R3A between ovarian cysts and diabetes, and PTEN between ovarian cysts and obesity.

A better characterization of the common ground of the three phenotypes comes from the analysis of shared functional annotations. 7 GO terms for molecular function are shared, being *hormone receptor binding* (GO:0051427) the most specific one (IC=6.84). Moreover, 58 GO terms for biological process are shared, 16 of which with IC values greater than 5. These include *generation of precursor metabolites and energy* (GO:0006091), *energy derivation by oxidation of organic compounds* (GO:0015980), *cellular response to peptide hormone stimulus* (GO:0071375), *developmental process involved in reproduction* (GO:0003006), *response to peptide hormone* (GO:0043434), *cellular response to hormone stimulus* (GO:0032870), *response to hormone* (GO:0009725), *response to insulin* (GO:0032868), *regulation of growth* (GO:0040008*).* Each term is associated with the three phenotypes by means of many genes, including PPARG. On the overall, the annotation points towards phenomena associated with the response to hormones, in particular insulin. Specifically, the *response to insulin* is associated with each phenotype with a corrected p-value of 1E-9, 0.04 and 0.005, respectively for Diabetes, Obesity and Ovarian Cyst.

The novelty with PhenPathTOOL is that the co-occurrence of the three phenotypes is ascribed to defects of the response to insulin. Interestingly, recent literature confirms that insulin resistance is a common background for both obesity and diabetes mellitus type 2 [26] and that insulin is a key factor also in the uptake of glucose by ovarian tissues during the menstrual cycle of some rodent, primate and ruminant species [27]. In particular, the link between metabolic disorders and cystic ovarian disease has been studied in animal models [28], specifically for the insulin resistance as a pathogenic factor. Our analysis is also supported by the finding that the activity of PPARG, the only gene shared among the three phenotypes under investigation, is sufficient for whole-body insulin sensitization [29].

3.3 Study case: Rett syndrome

PhenPathTOOL can be adopted to endow a disease (described with a set of phenotypes) with novel links to genes and functional terms, retrieved by intersecting the sets of genes and functional terms associated with the single phenotypes in PhenPathDB. As a study case, we here apply PhenPathTOOL to the detection of new associations between genes and Rett syndrome (RTT). RTT is a neurodevelopmental disorder corresponding to two OMIM entries (#312750 and #613454) linked to genes MECP2 (encoding methyl CpG binding protein 2) and FOXG1 (encoding the forkhead box protein G1), respectively [30,31]. RTT primarily affects females and it is characterized by loss of language and communication skills, microcephaly, learning impairment, coordination, and other brain functions. Affected girls may lose the use of their hands and begin making repeated hand-wringing, washing, or clapping motions. Atypical forms of RTT, not reported in OMIM, have been described in patients not carrying mutations on FOXG1 nor MECP2 and manifesting additional phenotypes such as breathing abnormalities, spitting or drooling, unusual eye movements, cold hands and feet, irritability, sleep disturbances, seizures and scoliosis [32, https://ghr.nlm.nih.gov/condition/rett-syndrome]. Recently, literature reported new genes associated with RTT, including cyclin-dependent kinase-like 5 (CDKL5), myocyte-specific enhancer factor 2C (MEF2C), and transcription factor 4 (TCF4) [33-35]. These associations are not yet reported in major databases and, consequently, they are not included in PhenPathDB. We tested the ability of PhenPathTOOL to recover these associations starting from the phenotype description. We entered 9 HPO terms, characterizing the classical and atypical RTT, namely *breathing dysregulation* (HP:0005957), *abnormality of coordination* (HP:0011443), *drooling* (HP:0002307), *irritability* (HP:0000737), *severe expressive language delay* (HP:0006863), *specific learning disability* (HP:0001328), *microcephaly* (HP:0000252), *scoliosis* (HP:0002650), and *sleep disturbance* (HP:0002360).

As a first step, PhenPathTOOL intersects the gene sets associated with the phenotypes. Although no gene is common to the nine phenotypes, 5 genes (MECP2, CDKL5, UBE3A, SLC2A1, SLC16A2) are shared by 5 phenotypes. MECP2 and CDKL5 have been previously reported [32, 35]. Interestingly, our analysis highlights the association with CDKL5, which is not present in PhenPathDB.

PhentPathTOOL then retrieves the intersection of GO terms, KEGG and Reactome pathways enriched for the different phenotypes. Focusing on GO BP, 440 terms are shared among two or more phenotypes. In particular, when restricting to terms with medium/high specificity (IC > 4.5), 12 enriched terms are common to 5 or more phenotypes. Among them, the seven terms listed in Table 4 describe biological processes that involve the two genes known to be related with RTT (MECP2 and FOXG1), as well as TCF4, that has been only recently associated with RTT (Table 3).

These findings illustrate the efficacy of PhenPathTOOL in linking a set of phenotypes to genes and functional annotations, which can be adopted for planning further experimental analysis.

Table 3: A selection of GO BP terms shared by the phenotypes in input after enrichment procedure.

GO BP term	IC value	# of associated phenotypes	Associated phenotypes	Related genes associated with RTT
cellular component morphogenesis	5.4	6	*microcephaly, sleep disturbance, scoliosis, breathing dysregulation, abnormality of coordination, specific learning disability*	**TCF4**
Behavior	5.23	6	*microcephaly, specific learning disability, sleep disturbance, scoliosis, abnormality of coordination, drooling*	MECP2
cell projection organization	4.95	6	*microcephaly, specific learning disability, scoliosis, breathing dysregulation, abnormality of coordination, sleep disturbance*	MECP2
neurological system process	4.65	6	*microcephaly, specific learning disability, scoliosis, abnormality of coordination, sleep disturbance, drooling*	FOXG1, MECP2
system development	4.75	5	*microcephaly, sleep disturbance, scoliosis, abnormality of coordination, severe expressive language delay*	**TCF4**, FOXG1, MECP2
anatomical structure formation involved in morphogenesis	4.69	5	*microcephaly, specific learning disability, scoliosis, breathing dysregulation, abnormality of coordination*	**TCF4**, FOXG1
single-organism behavior	5.67	5	*microcephaly, sleep disturbance, scoliosis, abnormality of coordination, drooling*	**TCF4**, MECP2

The table reports some of the most interesting biological processes associated with the phenotypes given as input to PhenPathTOOL. For each term, the IC value is shown with the specific phenotype associations. Noticeably, TCF4 has been only recently associated with RTT [35].

3.4 Study case: Associating genes to uncharacterized diseases

We propose PhenPath as a resource for formulating hypotheses on the molecular mechanisms underlying the manifestation of concomitant phenotypes, in particular in case of non-well characterized diseases. Here we estimate the performance of PhenPathTOOL in retrieving relevant associations between groups of co-occurring phenotypes and possible causative genes, collecting from Orphanet [36] a blind set consisting of 87 diseases, not included in OMIM nor, consequently, used to build PhenPathDB. Orphanet associate these diseases with both HPO phenotypic terms and sets of possibly causative genes (see Methods section 5.3 for further details on the dataset).

We evaluate the efficiency of PhenPathTOOL in retrieving genes starting from the phenotypic characterization of diseases. For each disease in the blind set, we entered in PhenPathTOOL the Orphanet-associated HPO terms and we retrieved the corresponding lists of shared genes. We then compared the genes retrieved with PhenPathTOOL with the genes proposed by Orphanet

For 61 diseases out of 87 (70%), PhenPathTOOL retrieves at least one of the genes associated by Orphanet. Overall, out of the 100 genes associated by Orphanet, 58 are recovered with PhenPathTOOL (58%). In particular for 2 diseases, “Pituitary stalk interruption syndrome “and “Hypothyroidism due to deficient transcription factors involved in pituitary development or function “, PhenPath retrieves 5 out of 7 and 5 out 5 Orphanet-associated genes, respectively.

A summary of all the results obtained for the external dataset is provided as supplementary material (https://drive.google.com/file/d/1PkH48TMpxA33RxXRYNGj6RQnD4qtdEI2/).

4 Conclusions

PhenPath offers a new approach for investigating the molecular mechanisms leading to the correlated manifestation of different phenotypes. PhenPath may be used to explore the possible connections among different phenotypes co-occurring in a patient, offering new clues on the biological mechanisms that may explain its clinical conditions.

Four case studies show the potential use of PhenPath for retrieving diseases starting from a set of phenotypes, if existing, and/or for better characterize the functions and pathways possibly involved in the manifestation of different symptoms. We propose our resource for directing scientific efforts, helping the diagnosis and retrieving new possible associations among biological processes and diseases. We believe that biotechnologists, physicians and medical researchers may find PhenPath a useful resource of information, especially when studying complex and rare diseases.

5 Methods

5.1 Associations among phenotypes, diseases and genes

PhenPathDB stands on the merging of disease-phenotype and disease-gene relationships. In PhenPath, a phenotype is defined as an actual physical characteristic, and we follow the phenotype characterization provided by HPO and OMIM Clinical Synopsis. We define a disease as a medical condition associated with specific phenotypes, and we classify diseases according to OMIM identifiers.

In detail, two lists of phenotype terms have been considered: the OMIM Clinical Synopsis (March 2017 release) and the HPO Phenotypic Abnormalities categories (May 2017 release). OMIM Clinical Synopsys groups OMIM diseases within 22 phenotypic categories, 18 referring to systems of the human body (e.g.: respiratory system, musculature, etc.) and 4 referring to further level of characterization (inheritance, laboratory abnormalities, molecular basis, and miscellaneous). In PhenPath, we retained the former and discharged the latter, ending up with the phenotypic characterization for 3,230 diseases.

The HPO consists of 12,111 different phenotypes organized into a direct acyclic graph (DAG) including 3,837 leaf phenotypes. A leaf in a graph is a node without sub-nodes (children), and by consequence, a leaf phenotypic term provides the most detailed level of annotation. When a phenotype is associated with a disease by HPO, the annotation is extended to all the parent phenotypes in the HPO DAG. On the overall, 4,292 OMIM diseases are associated with 7,137 HPO phenotypes, which represent the 59% of all the HPO phenotypes. In particular, 4,023 diseases are associated with 3,837 leaf phenotypes. Of particular interest are the phenotypes originating from 24 main categories, referring to human body districts and physiological functions (musculature, respiratory system, head or neck, genitourinary system, cardiovascular system, immune system, nervous system, voice, blood and blood-forming tissues, metabolism/homeostasis, breast, growth, constitutional symptoms, digestive system, neoplasm, thoracic cavity, prenatal development or birth, eye, ear, skeletal system, limbs, connective tissue, endocrine system, integument). These categories are grouped into the Phenotypic Abnormalities sub-ontology. It comprises 5661 phenotypes, among which 3802 are leaves.

Gene-disease associations are extracted from our curated database, eDGAR [15] (August 2017 release), which collects information from OMIM, Humsavar and ClinVar.

5.2 Enrichment analysis

For each group of genes associated to the same phenotype, the functional characterization is performed with NET-GE [16], an algorithm for standard and network-based gene enrichment analysis that includes the annotations derived from GO, KEGG and Reactome pathways. Briefly, it relies on the STRING Human Interactome [17], to build function-specific modules of interacting genes, starting from genes/proteins annotated with a given term. Then, given a list of genes/proteins, the over-represented modules are retrieved and scored with a p-value computed with an exact Fisher test and corrected with the Bonferroni procedure. A significance threshold of 0.05 has been considered.

The Information Content (IC) is computed for each GO term, KEGG and REACTOME pathways, adopting the following equation:

$$IC_{term} = -log_2\left(\frac{N_{term}}{N_{root}}\right) \qquad (1)$$

where N_{term} is the number of human genes endowed with the particular GO, KEGG or REACTOME term and N_{root} is the number of human genes annotated in the ontology. IC lower limit is zero; high IC values indicate that a small number of genes are annotated with a particular term in the human genome and therefore the annotation is highly informative.

For every phenotype in PhenPath, we perform the enrichment procedure via NETGE algorithm and we report the results in the PhenPathDB webpages. Using PhenPathTOOL, the users may compare different phenotypes retrieving the enriched biological pathways shared over the phenotypes in input. For each term describing a pathway, we report the Pvalue of the significant associations to every phenotype in input.

5.3 Blind dataset for the performance evaluation

For the evaluation of the performance of PhenPathTOOL we collected a dataset of phenotype-disease-gene associations from Orphanet, a resource for rare diseases with high-quality information [36]. In Orphanet (release Dec 2018), 3765 diseases are associated both with HPO phenotype terms and genes. We filtered out all diseases mapped to OMIM and therefore used for the implementation of PhenPathDB, retaining 550 Orphanet diseases. We then collected diseases associated with 2 or more HPO phenotypes, ending up with 87 diseases, which form a blind set for testing PhenPathTOOL. For each disease, we entered in PhenPathTOOL the associated HPO phenotypic terms and we retrieved the list of genes they. We compare these proposed genes with the genes reported by Orphanet for the disease. The evaluation dataset is provided as supplementary material (https://drive.google.com/file/d/1PkH48TMpxA33RxXRYNGj6RQnD4qtdEI2/)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analyzed during the current study are available at:

http://edgar.biocomp.unibo.it/phenpath/

Competing interests

The authors declare that they have no competing interests

Funding

University of Bologna (R.F.O. 2017 to R.C and P.L.M.).

Authors' contributions

RC, PLM, and GB conceived and designed the work and wrote the paper. GB collected and curated data. PLM, GB and RC analyzed and interpreted data. All authors critically revised and approved the manuscript.

Acknowledgements

Not applicable.

References

1. Fisch GS. Whither the genotype-phenotype relationship? An historical and methodological appraisal. Am J Med Genet C Semin Med Genet 2017; 175:343-353.
2. Hu JX, et al. Network biology concepts in complex disease comorbidities. Nat Rev Genet. 2016; 17:615-629.
3. Zielinski A, Halling A. Association between age, gender and multimorbidity level and receiving home health care: a population-based Swedish study. BMC Research Notes 2015; 8:714.
4. Brodie A, et al. Large-Scale Analysis of Phenotype-Pathway Relationships Based on GWAS Results. PLoS ONE 2014;9:e100887.
5. Ramos EM, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet 2014; 22:144-147.

6. Pontikos N, et al. Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. Bioinformatics 2017; 33:2421-2423.
7. Köhler S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res 2017; 45:865–876.
8. Amberger JS, et al. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res 2015; 43: D789-98.
9. Hidalgo CA, et al. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Computational Biology 2009; 5:e1000353.
10. Köhler S, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies, Am J Human Gen 2009; 85:457-464.
11. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017; 45: D158–D169.
12. Landrum MJ, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016; 44:D862-868.
13. Piñero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017; 45: D833–D839.
14. Babbi G, et al. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. BMC Genomics 2017; 5:554.
15. Bovo S, et al. NET-GE: a web-server for NETwork-based human gene enrichment. Bioinformatics. 2016; 32:3489-3491.
16. Szklarczyk D, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: D447-452.
17. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res 2017; 45: D331-D338.
18. Kanehisa M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017; 45:D353-D361.
19. Fabregat A, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res 2018; 46:D649-D655.
20. Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res 2017; 45:D619-625.
21. Jankovic J. Tourette's syndrome. New Eng J Med 2001; 345:1184-1192.
22. Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/tourette-syndrome/diagnosis-treatment/drc-20350470 (Accessed 7 August 2017)
23. Hawksley J, et al. The role of the autonomic nervous system in Tourette Syndrome. Front Neurosci 2015;9:117
24. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol 2018; 14:270-284.

25. Rubin KH, et al. Development and Risk Factors of Type 2 Diabetes in a Nationwide Population of Women with Polycystic Ovary Syndrome. J Clin Endocr Metabolism 2017; 102:3848–3857.
26. Al-Goblan AS, et al. Mechanism linking diabetes mellitus and obesity. Diabetes Metab Syndr Obes. 2014;7: 587–591.
27. Dupont J, Scaramuzzi RJ. Insulin signalling and glucose transport in the ovary and ovarian function during the ovarian cycle. Biochem. J 2016; 473:1483–1501.
28. Opsomer G, et al. Insulin resistance: the link between metabolic disorders and cystic ovarian disease in high yielding dairy cows? Anim Reprod Sci 1999; 56:211-222.
29. Shigeki S, et al. PPARγ activation in adipocytes is sufficient for systemic insulin sensitization. Proc Natl Acad Sci USA 2009;106:22504-22509.
30. Amir RE, et al. Rett syndrome is caused by mutations in Xlinked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet 1999; 23:185–188.
31. Philippe C, et al. Phenotypic variability in Rett syndrome associated with FOXG1 mutations in females J Med Gen 2010; 47:59– 65.
32. Tarquinio DC, et al. The Changing Face of Survival in Rett Syndrome and MECP2-Related Disorders. Pediatr Neurol 2015; 53:402– 411.
33. Dolce A, et al. Rett syndrome and epilepsy: an update for child neurologists Pediatr Neurol 2013;48:337– 345
34. Armani R, et al. Transcription factor 4 and myocyte enhancer factor 2C mutations are not common causes of Rett syndrome Am J Med Genet A 2012; 158A:713– 719.
35. Evans JC, et al. Early onset seizures and Rett-like features associated with mutations in CDKL5. Eur J Hum Gen 2005;13:1113– 1120
36. Pavan S, et al. Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS One. 2017; 12:e0170365

5 INPS-3D

5.1 Contribution to the state of the art

We describe INPS-3D, a predictor based on protein structure for computing the effect of single residue variations on protein stability (ΔΔG), scoring at the state-of-the-art.

Change of protein stability upon variation appears to assume a particular relevance in annotating whether a single residue substitution can or cannot be associated to a given disease. Thermodynamic properties of human proteins and of their disease related variants are still lacking. In the present work, we take advantage of the available three-dimensional structure of human proteins for predicting the role of disease related variations on the perturbation of protein stability.

We then filter 368 OMIM disease related proteins known with atomic resolution with 4,717 disease related single residue variations and 685 polymorphisms without clinical consequence. Our analysis indicates that OMIM disease related variations in proteins promote a much larger effect on protein stability than polymorphisms non-associated to diseases. Disease related variations with a slight effect on protein stability frequently occur at the protein accessible surface suggesting that they are located in protein-protein interactions patches in putative human biological functional networks. The hypothesis is corroborated by proving that proteins with many disease related variations that slightly perturb protein stability are on average more connected in the human physical interactome (IntAct, Ochard S et al, 2013) than proteins with variations predicted larger effect on protein stability.

5.2 General information on the paper

The presented paper can be found in the following publication:

Authors: Pier L. Martelli, Piero Fariselli, Castrense Savojardo, Giulia Babbi, Francesco Aggazio and Rita Casadio

Title: Large scale analysis of protein stability in OMIM disease related human protein variants

Journal: BMC Genomics

Volume: 17 (Suppl 2) **Issue:** 397 **Pages:** 239-247

Year: 2016

Impact Factor: 3.73

Quartile and subject: 1st quartile in Biotechnology and Applied Microbiology

DOI: 10.1186/s12864-016-2726-y.

BMC Genomics

RESEARCH ARTICLE Open Access

Large scale analysis of protein stability in OMIM disease related human protein variants

Pier Luigi Martelli[1,2*†], Piero Fariselli[1,3†], Castrense Savojardo[1,2†], Giulia Babbi[1,2], Francesco Aggazio[1,2] and Rita Casadio[1,2]

From VarI-SIG at ISMB 2015
Dublin, Ireland. 11/07/2015

Abstract

Background: Modern genomic techniques allow to associate several Mendelian human diseases to single residue variations in different proteins. Molecular mechanisms explaining the relationship among genotype and phenotype are still under debate. Change of protein stability upon variation appears to assume a particular relevance in annotating whether a single residue substitution can or cannot be associated to a given disease. Thermodynamic properties of human proteins and of their disease related variants are lacking. In the present work, we take advantage of the available three dimensional structure of human proteins for predicting the role of disease related variations on the perturbation of protein stability.

Results: We develop INPS3D, a new predictor based on protein structure for computing the effect of single residue variations on protein stability (ΔΔG), scoring at the state-of-the-art (Pearson's correlation value of the regression is equal to 0.72 with mean standard error of 1.15 kcal/mol on a blind test set comprising 351 variations in 60 proteins). We then filter 368 OMIM disease related proteins known with atomic resolution (where the three dimensional structure covers at least 70 % of the sequence) with 4717 disease related single residue variations and 685 polymorphisms without clinical consequence. We find that the effect on protein stability of disease related variations is larger than the effect of polymorphisms: in particular, by setting to |1 kcal/mol| the threshold between perturbing and not perturbing variations of the protein stability, about 44 % of disease related variations and 20 % of polymorphisms are predicted with |ΔΔG| > 1 kcal/mol, respectively. A consistent fraction of OMIM disease related variations is however predicted to promote |ΔΔG| ≤ 1 kcal/mol and we focus here on detecting features that can be associated to the thermodynamic property of the protein variant. Our analysis reveals that some 47 % of disease related variations promoting |ΔΔG| ≤ 1 are located in solvent exposed sites of the protein structure. We also find that the increase of the fraction of variations that in proteins are predicted with |ΔΔG| ≤ 1 kcal/mol, partially relates with the increasing number of the protein interacting partners, corroborating the notion that disease related, non-perturbing variations are likely to impair protein-protein interaction (70 % of the disease causing variations, with high accessible surface are indeed predicted in interacting sites). The set of OMIM surface accessible variations with |ΔΔG| ≤ 1 kcal/mol and located in interaction sites are 23 % of the total in 161 proteins. Among these, 43 proteins with some 327 disease causing variations are involved in signalling, structural biological processes, development and differentiation.

(Continued on next page)

* Correspondence: gigi@biocomp.unibo.it
†Equal contributors
[1]Biocomputing Group, University of Bologna, Via San Giacomo 9/2, 40126 Bologna, Italy
[2]Department BiGeA, University of Bologna, Via Selmi 3, 40126 Bologna, Italy
Full list of author information is available at the end of the article

(Continued from previous page)

Conclusions: We compute the effect of disease causing variations on protein stability with INPS3D, a new state-of-the-art tool for predicting the change in ΔΔG value associated to single residue substitution in protein structures. The analysis indicates that OMIM disease related variations in proteins promote a much larger effect on protein stability than polymorphisms non-associated to diseases. Disease related variations with a slight effect on protein stability (|ΔΔG| < 1 kcal/mol) frequently occur at the protein accessible surface suggesting that they are located in protein-protein interactions patches in putative human biological functional networks. The hypothesis is corroborated by proving that proteins with many disease related variations that slightly perturb protein stability are on average more connected in the human physical interactome (IntAct) than proteins with variations predicted with |ΔΔG| > 1 kcal/mol.

Keywords: Protein stability, Disease related-variations, Residue solvent accessibility, Interactomics networks

Background

One of the key goals in the postgenomic era is the elucidation of the mechanisms at the basis of the relationship between genotype and phenotype. In particular, understanding how human genetic variations are associated to diseases is still an open problem and its solution is a crucial issue for exploiting the possibilities offered by the modern sequencing techniques in the framework of precision medicine [1, 2].

The role of missense mutations inducing single residue variations (SRVs) in proteins has been widely investigated: several databases collect data about the relationship between SRVs and diseases [3] and several predictive tools have been implemented in order to exploit the available knowledge to predict whether new variants are related to diseases ([4–6]; and others listed in [7]) or are affecting protein function [8].

Biophysical studies allowed to measure the thermodynamic effect that protein variations induce on protein stability [9]. However the number of human proteins whose folding thermodynamics is known in the native and mutated form is still limited due to the time consuming and costly procedure at the basis of experimental investigations. To fill the gap, predictive tools have been trained on the available thermodynamic data to compute the free energy change value upon variation ([10–13], and others listed in [14]). Recently, we introduced INPS [15], a sequence based predictor that well compares with tools taking as input protein structure. When dealing with disease related variations in human protein variants, very little is known about their thermodynamics and it is unclear in annotation processes whether a variation perturbing the protein stability is or not disease related. Extensive comparative analyses of the two classes of datasets (phenotypically vs thermodynamically characterized variations) prove that, on average, variation types most involved in disease are also associated to a large effect on protein stability [16–18]. However, the strength of this association, although recently improved (compare results in [16] with [19]), is not sufficient to consider protein destabilization as the only mechanistic cause explaining the insurgence of diseases. Indeed many variations with |ΔΔG| ≤ 1 kcal/mol are disease-related [12, 13, 15, 16, 19]. In this paper, as a follow up to the problem, we specifically deal with OMIM disease related protein variants whose native structure is known and predict the extent of perturbation that the variation may cause on the native protein stability. To this aim, we develop INPS3D, a new tool for computationally estimating the effect of single residue variations on protein stability based on information extracted from protein three dimensional structure, and compare its performance to state-of-the-art predictors on the blind test set of the OMIM related proteins endowed with well resolved structures. By this, we identify a subset of disease-related variations with |ΔΔG| ≤ 1 kcal/mol and prove that these variations often occurs in sites exposed on the protein accessible surface, with a likelihood to be in interaction sites. Integrating these results with human physical interactomic data, we find that on average, proteins endowed with many interaction partners have disease related variations that are solvent exposed and are characterized by low free energy change values. Our results support the hypothesis that, besides protein stability perturbation, impairment of protein-protein interaction can be also a major mechanism explaining the relation between variations and diseases.

Methods

Data set

We downloaded from the Humsavar dataset (release 2015_10 of 14 Oct 2015) a collection of 27,185 variations related to 3082 OMIM diseases, on 2367 different human proteins and retained only proteins endowed with a PDB structure (3D) covering at least 70 % of the protein sequence. The PDBSWS resource [20] (August 2015 update) was adopted to map the UniProt sequences onto the PDB structures. We ended up with a dataset of 4717 variations related to 484 OMIM diseases on 368 proteins endowed with PDB structures with resolution lower than 3.0 Å (OMIM set). On the same proteins, we also collected 685 polymorphism lacking evidence of association to disease (POLY set).

To train/test (by adopting a cross validation procedure) the predictors, we used S2648, a dataset that was originally derived from the ProTherm database [9] and corrected by the authors of the PoPMuSiC algorithm [11]. It comprises 2648 variations out of 132 different proteins endowed with a 3D structure. We also evaluated the predictor performances on a blind test of 351 variations in 60 proteins, and on 42 variations of the P53 protein not included in the training set and previously described in [12].

INPS3D: a structure based method for the prediction of free energy changes upon protein variations

Here we introduce INPS3D that exploits both sequence and structural information to predict the protein stability changes upon single point mutation. INPS3D takes advantage of the recently released INPS [15] that, starting only from protein sequence, performs similarly to the state-of the-art methods based on protein structure. INPS3D is based on nine input features based on protein sequence and structure. The features extracted from protein sequence are, [15]: 1) substitution score derived from the Blosum62 matrix; 2-3) Kyte-Doolittle hydrophobicity scores of native and mutated residues; 4) mutability index of the native residue; 5-6) molecular weights of native and mutated residues; 7) the difference in the alignment score between the native and mutated sequences and an HMM encoding evolutionary information of the target sequence. Two additional real-valued features derived from the protein structures are: 8) the solvent accessibility of the mutated residue, 9) the energy difference between native and mutated proteins. The solvent accessibility is computed with the DSSP method [21] and normalized as previously described [22]. The energy difference is evaluated by using the residue-based contact potential described in [23]. We consider that two residues are in contact if the minimal distance between all the atoms (not including hydrogen atoms) of two residues is ≤ 5 Å. We used the coordinates of the native protein to compute the contact energy and the energy difference as:

$$\sum_r P(r,w)-P(r,m) \tag{1}$$

where P is the contact potential, w is the wild-type residue, m is the mutated residue, and the r-index runs over the list of w-neighbouring residues. We tested several other potentials, but the performances were similar or lower than those here reported. INPS3D is based on a Support Vector Regression model (SVR) trained on the same dataset adopted for INPS (see data set section). The adopted conventions on the sign are such as when predicting the ΔΔG associated to a variation, positive values refer to the protein stabilization and negative values to protein destabilization.

Analysis of protein surfaces

The solvent accessible surface area of residues in wild-type proteins has been evaluated with the DSSP program [21]. In order to obtain the Relative Solvent Accessibility (RSA), solvent accessibility areas were normalized to the residue-specific maximum solvent accessible area, as previously reported [22]. Residues with RSA ≥ 0.2 are classified as accessible, residues with RSA < 0.2 are classified as buried. RSA has been measured on both the protein isolated chain and the protein complex, as downloaded from the repository of "biological assemblies" of the Protein Data Base [http://www.rcsb.org/pdb/download/download.do#Structures]. To define the interaction interface of the complex, we collected the set of residues that are solvent accessible in the isolated chain and are buried in the complex.

Interactomics analysis

Interacting partners of each protein were retrieved from the IntAct database [24] as downloaded from the IntAct FTP site as to November 2015. The search in the IntAct file was performed using the UniProtKB code and excluding the negative interaction data. The statistical analysis was performed considering only the proteins present in the dataset, at least in one entry.

Table 1 Performance of INPS3D and other state-of-the-art predictors

Method	Cross-validation (2648 variations on 132 proteins)	Blind test set (351 variations on 60 proteins)	Blind test set (42 variations on P53 protein)
INPS[b]	0.53/1.29[a]	0.68/1.26[a]	0.71/1.49[a]
INPS3D	0.58/1.20[a]	0.72/1.15[a]	0.76/1.35[a]
MAESTRO[c]	0.63/1.17[a]	0.71/1.16[a]	0.44/1.71[a,e]
mCSM[d]	0.51/1.26[a]	0.67/1.19[a]	0.68/1.40[a]

[a]Pearson's correlation coefficient/standard error (kcal/mol)
Data are from [b][15]; [c][13]; [d][12], [e]this work, respectively

Results and discussion

INPS3D at work

INPS3D is a new tool for predicting the change of protein folding free energy induced by single residue variations. The performance of the structure based predictor along with that of the sequence based one [15] are shown in Table 1. We report statistical scores obtained benchmarking the predictors with a more stringent per-protein cross-validation procedure [15] on the S2648 set previously described [11], and on a blind test set including some 351 variations in 60 proteins, and a P53 data set (both not included in the training set). Results, reported in Table 1, indicate that INPS3D outperforms INPS, exploiting structure based features not present in the INPS input encoding. INPS3D well compares with the performances obtained with structure-based state-of-the-art methods, mCSM [12], and MAESTRO, recently made available as web server [13].

Predicting the effect of disease related, single residue variations on the stability of OMIM linked proteins

We applied INPS (sequence based), INPS3D (structure based) and MAESTRO (structure based) to the OMIM variation set for estimating the change in protein folding free energy induced by the disease-related variations. For sake of comparison we also ran the tools on the POLY set, containing variations not related to diseases, on the same OMIM proteins. We used polymorphisms from the very same proteins that have also variations related to diseases, in order to constrain the ΔG value of the folded form and avoid possible biases due to the inclusion of other proteins. The results (Fig. 1) confirm that disease related variations tend to produce a larger effect on protein stability than polymorphisms, which, on the other hand, appear to promote free energy perturbations mostly distributed within +/-1 kcal/mol. The result is confirmed by all the predictors. INPS3D predicts that 80 % of polymorphisms and 56 % of disease causing variations promote a |ΔΔG| ≤ 1 kcal/mol with respect to the corresponding native protein.

The results are similar with INPS; with Maestro, the fraction of disease-related variations predicted with low |ΔΔG| values increases to 74 % of the total. Our results, obtained with three independent predictors, corroborate the notion that protein stability perturbation (as detected from the predicted |ΔΔG| > 1 kcal/mol) is associated to disease-related variations. However, at least half of the OMIM set is predicted to promote only a slight change in protein stability (within a range of about 1 kcal/mol in absolute value). The observation poses the question as to whether the thermodynamic property of the protein variant (albeit predicted) can be linked to some structural/ functional feature of the variation, specifically when it is disease causing. Many investigations addressed the issue of which structural features could be associated to disease related variations ([25–29] and references therein). Conclusions are that genetic variations can have dramatic effects on protein stability, hydrogen bonding networks, conformational dynamics, protein activity and protein interaction networks, particularly at the level of functional assemblies [28]. More recently the correlation between the probability of perturbing the protein stability and that of being disease causing was improved [19] with respect to previous data [16]. However, here our analysis addresses the issue from a different perspective: considering that we have predictors of protein stability, the problem is to which extent they label the overall protein in/stability in relation to the corresponding disease related mutation. We find that a high fraction of the protein variants carrying disease-related mutations are predicted with a low |ΔΔG| value, rather independently of the method (compare the INPS3D to MAESTRO results).

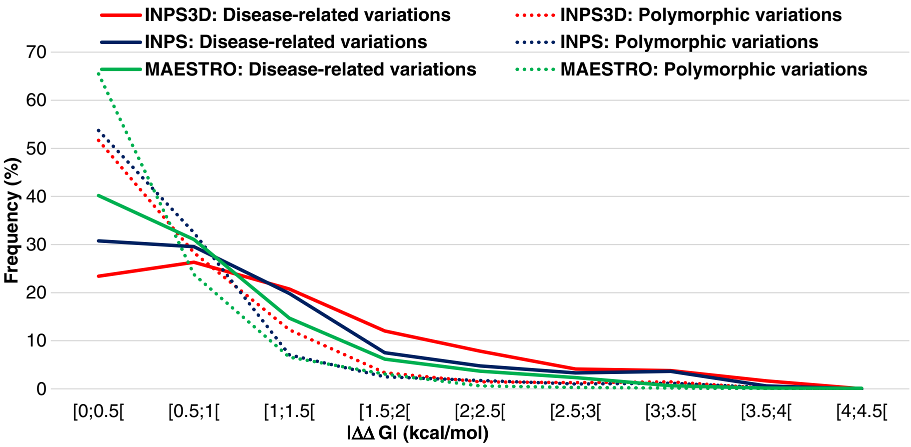


Fig. 1 Distribution of the absolute value of the ΔΔG predicted with INPS3D, MAESTRO and INPS. The set includes 4717 disease related variations and 687 polymorphisms in 368 OMIM proteins

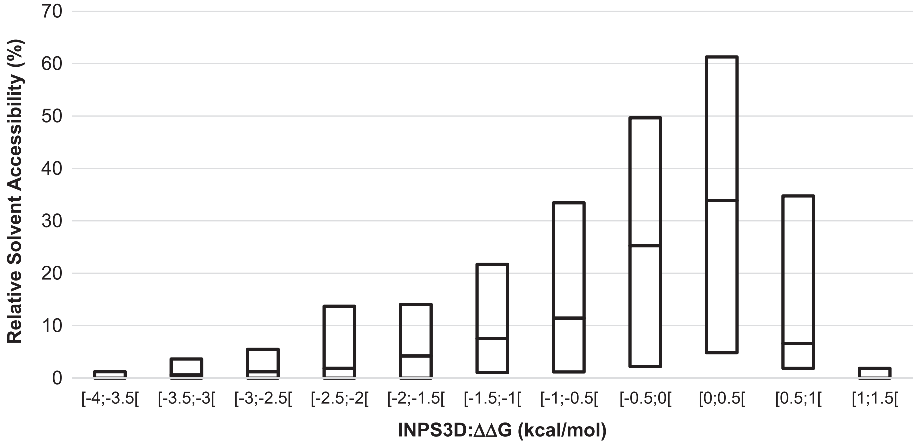


Fig. 2 Relative Solvent Accessibility of the variations as a function of ΔΔG predicted for the variants of the OMIM set. The box-plot reports the median and the lower and upper quartiles of the distribution of relative solvent accessibility for each interval of ΔΔG

Protein |ΔΔG| values and structural/functional properties of the variations

In the following we will consider how some structural properties can be clustered considering perturbing and non- perturbing predicted |ΔΔG| values. The analysis focuses on the Relative Surface Accessibility (RSA), on the propensity of the variation to be or not in an interaction patch, and finally on the relation of the protein variant to be in physical interaction with other proteins, considering ΔΔG values predicted with INPS3D.

We analyse the distribution of the relative solvent accessibility (RSA) of the disease related mutations as a function of the free energy change predicted for the corresponding protein variant. Boxplots in Fig. 2 show that the median and the upper quartile values of RSA are higher in the intervals with ΔΔG values close to zero. This indicates that disease related variations with low ΔΔG values have a more spread out distribution of RSA, and then a larger probability to be solvent accessible.

In Fig. 3, the distribution of the fraction of solvent accessible variations is plotted as a function of the |ΔΔG| values for disease related and polymorphic protein variants. Low |ΔΔG| values are apparently common both to disease causing and polymorphic variations, when they are located in accessible protein sites.

A detailed grouping of the different behaviour of the structural properties of the OMIM related variations is shown in Tables 2 and 3, as a function of the thermodynamic property of the protein variant. Here we focus also on the difference among monomers and assemblies (as documented in the Protein Data Bank, http://www.rcsb.org/pdb/download/download.do#Structures), in order to highlight the role of protein-protein interactions, when present, in the biological functional unit. As an additional feature, we also included the likelihood of each variations to be or not in an interaction patch (computed with our PRED-PPI, [30]). It appears that disease related mutations in proteins variants with low |ΔΔG| values, when solvent

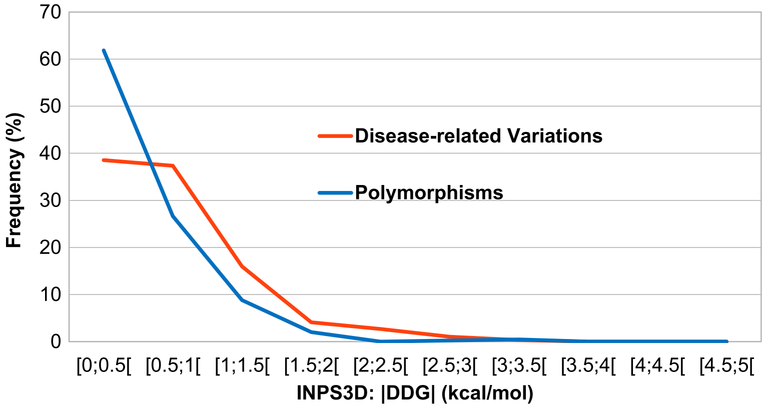


Fig. 3 Frequency of the solvent accessible variations as a function of ΔΔG predicted for the protein variants of the OMIM set

Table 2 Relation between thermodynamic properties and structural properties in proteins with biologically functional monomeric assembly

Disease-related variant	RSA ≥ 0.20	RSA < 0.20
|ΔΔG| ≤ 1	562 (23.4 %)[a]*398*	756 (31.4 %)[a]*39*
|ΔΔG| > 1	176 (7.3 %)[a]*120*	907 (37.8 %)[a]*36*
Polymorphic variant		
|ΔΔG| ≤ 1	194 (59.0 %)[a]*110*	72 (21.9 %)[a]*3*
|ΔΔG| > 1	22 (6.7 %)[a]*10*	41 (12.5 %)[a]*0*

[a]Number of residue predicted to be part of a protein-protein interaction patch (for details on the prediction method, see [30]). Predicted set: 2401 disease related variations and 329 polymorphic variations in 177 proteins

exposed (RSA ≥ 0.20), have also a tendency to be in interaction sites. The property is shared, as expected, with variations that highly perturb protein stability and with polymorphic ones. The low accessibility, in all cases, well agrees with a propensity of being in interaction sites ranging from 0 to 5 %. The value can be considered indicative of the possible range of the false positive rate of the predictor, trained and tested on accessible interaction sites and for which the OMIM set of disease related and polymorphic variations is a blind test set.

Distinguishing functional monomeric from multimeric biological assemblies highlights the relevance of the variations when they are located at the interface of protein complexes [28]. In Table 3, the same grouping of Table 2 is therefore shown for proteins with a biologically functional assembly, as documented in the PDB. Here, it appears that only a small fractions of the total number of disease related mutations in the set occurs at the monomer interface (compare Monomer and Complex at RSA ≥ 0.20) and concomitantly also the number of interaction sites predicted on the complex interface is very low.

From the data reported in Tables 2 and 3, it can be computed that about 70 % of the disease causing variations with high accessible surface in monomers are predicted to be part of an interaction patch. The result is particularly significant considering that the fraction of all accessible residues predicted in interaction patches on the same 368 proteins is 55 %.

Summing up, we show that disease related variations in proteins can promote a low |ΔΔG| value, particularly when they are located in accessible sites that are also interacting sites.

As a follow up, one may consider to which extent protein variants with disease-related mutations located in solvent exposed sites and slightly perturbing the stability, are or not involved in interaction networks of physical interaction, as available in IntAct [24]. We collected from IntAct the number of interacting partners for each protein and analysed it as a function of the fraction of solvent accessible, non-perturbing variations (Fig. 4). The upper quartile and the mean values of the number of interacting partners per protein increase as the fraction of disease related variations predicted as non-perturbing increases. When all the solvent exposed disease related mutations (RSA ≥ 20 %) per protein are related to the number of the corresponding protein interacting partners (Fig. 5), the trend is different from that observed in Fig. 4. This observation highlights the role of predicted ΔΔG values for determining the relation among protein variants with disease-related mutations located in solvent exposed sites and slightly perturbing the stability, and the number of interacting partners in a protein-protein interaction network.

The proteins endowed with a large amount of non-perturbing and solvent exposed disease related variations seem to play a central role in the human protein-protein interaction network. Likely, a variation on the protein surface can affect the interaction affinity, affecting important biological pathways and leading to an altered phenotype, as recently described [31]. Out of the 43

Table 3 Relation between thermodynamic properties and structural properties in proteins with biologically functional multimeric assembly

Disease-related variations	RSA ≥ 0.20	RSA < 0.20
|ΔΔG| ≤ 1	660 (28.5 %) Monomer[a]*465*	650 (28.0 %) Monomer[a]*24*
	550 (25.0 %) Complex[a]*421*	760 (31.5 %) Complex[a]*68*
|ΔΔG| > 1	213 (9.2 %) Monomer[a]*152*	793 (34.2 %) Monomer[a]*24*
	196 (8.5 %) Complex[a]*140*	810 (35.0 %) Complex[a]*36*
Polymorphic variations		
|ΔΔG| ≤ 1	198 (55.6 %) Monomer[a]*131*	84 (23.6 %) Monomer[a]*5*
	186 (52.2 %) Complex[a]*119*	96 (27.0 %) Complex[a]*17*
|ΔΔG| > 1	29 (8.1 %) Monomer[a]*21*	45 (12.6 %) Monomer[a]*9*
	29 (8.1 %) Complex[a]*21*	45 (12.6 %) Complex[a]*9*

[a]Number of residue predicted to be part of a protein-protein interaction patch. 2316 disease related variations and 356 polymorphic variations in 191 proteins. Predictions of INPS-3D and PRED-PPI are independent of the assembly state. RSA values were independently estimated on the monomeric and the complex structures

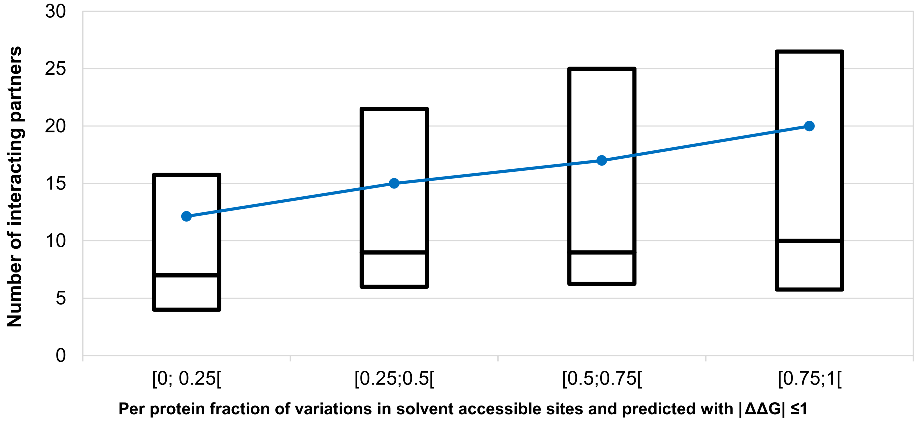


Fig. 4 Relation between the per-protein fraction of non-perturbing, solvent accessible variations and the corresponding number of the wild-type partners of interactions in the human interactome. The box-plot reports the median and the lower and upper quartiles of the number of interactions present in IntAct as a function of the fraction of solvent accessible, non-perturbing variations. The dashed blue line connects the average values. Non perturbing variations are those predicted to promote a |ΔΔG| ≤ 1 kcal/mol with INPS3D and found in protein sites that are solvent accessible. Data refers to 170 proteins with 4037 variations of our data set. Proteins with less than 5 disease-related variations or without interactomic data reported in IntAct are excluded

proteins for which at least 50 % of disease related variations are solvent exposed and predicted with |ΔΔG| ≤ 1, 42 % are involved in differentiation and development processes (including insulin, calmodulin, noggin, angiogenin), 40 % are involved in signalling processes (including the GTPases KRAS, HRAS and NRAS, the serine/threonine kinases PIK3CA and CHEK2), 23 % are structural and adhesion proteins (e.g., actins ACTA1, ACTG2, tubulin TUBA1A and integrin β2).

Conclusions

We address the problem of the perturbations of the protein stability by disease causing variations on a set of OMIM related proteins whose native structure is well solved. To this aim we implemented INPS3D, a tool for computationally estimating the change in ΔΔG value associated to single residue variations, taking as input protein structure. Our strategy is to adopt a predictor that scores at the state-of-the-art and we compare its performance to other state-of-the-art predictors. INPS3D exploits information extracted from protein structures and outperforms the recently released INPS, based only on sequence information. Moreover INPS3D outperforms state-of-the-art structure-based methods that perform similarly to INPS and well compares with MAESTRO, which recently became available as a web

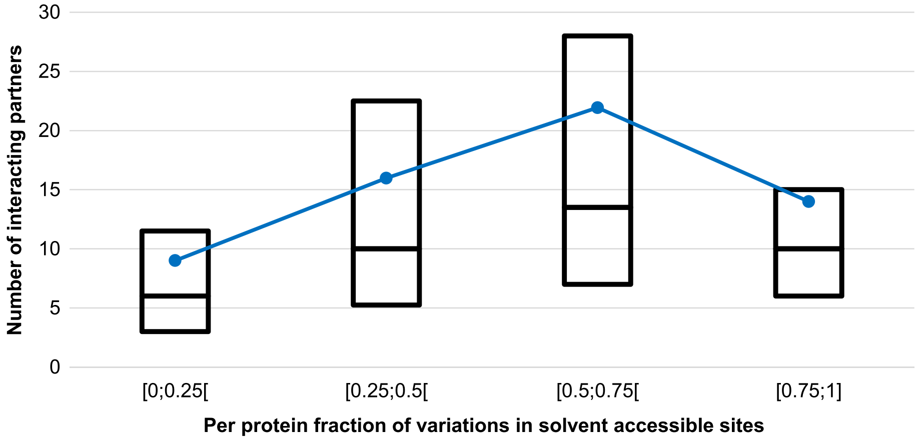


Fig. 5 Relation between the per-protein fraction of solvent accessible variations and the corresponding number of the wild-type partners of interactions in the human interactome. The box-plot reports the median and the lower and upper quartiles of the number of interactions present in IntAct as a function of the fraction of solvent accessible variations. The dashed blue line connects the average values. Data refers to 170 proteins with 4037 variations of our data set. Proteins with less than 5 disease-related variations or without interactomic data reported in IntAct are excluded

server [13]. Both predictors agree up to 90 % even in regions of |ΔΔG| values that can be considered below the error limit of the predictors. We found that OMIM disease-related variations in proteins generally promote a much larger effect on protein stability than polymorphisms non-associated to diseases on the same proteins, confirming that stability perturbation plays a crucial role in impairing protein function (recently confirmed also in [31]). Nevertheless, a significant fraction of disease related variations is predicted to have a small perturbation effect on protein stability: about 50 % of variations promote a |ΔΔG| <1 kcal/mol. The structural analysis of the corresponding proteins reveals that disease-related variations with a slight effect on protein stability often occur on the protein surface suggesting that they can affect the interaction of the proteins within biological functional networks. The analysis of protein-protein interaction networks corroborates the hypothesis that proteins with many non-perturbing disease-related variations are more connected in the human physical interactome (IntAct) than proteins with variations predicted with |ΔΔG| > 1 kcal/mol. The results are however indicative. The error associated to the computed |ΔΔG| value by our predictors (Table 1) is competing with the range of small changes in protein stability and this could increase the number of variations actually destabilising protein stability. It should also be mentioned that for each protein other features that are not exploited in this analysis (e.g., solubility, post-translational modifications, subcellular location, level of expression, etc.) may be considered when labelling a variations as disease causing.

Abbreviations

RSA: relative solvent accessibility.

Funding

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 material

The method is available at http://inpsmd.biocomp.unibo.it/inpsSuite/default/index3D. Data are available upon request.

Authors' contributions

PLM, PF, CS and RC conceived and designed the work and wrote the paper. PF and CS implemented and tested INPS3D. PLM and RC analysed and interpreted data on disease related variations. GB and FA curated the datasets and collaborated in data analysis. All authors critically revised and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Declarations

The publication charges for this article were funded by PRIN 2010-2011 project (Italian Ministry of Education, University and Research, cod 20108XYHJS). This article has been published as part of BMC Genomics Volume 17 Supplement 2, 2016: Proceedings of VarI-SIG 2015: Identification and annotation of genetic variants in the context of structure, function, and disease. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-17-supplement-2.

Author details

[1]Biocomputing Group, University of Bologna, Via San Giacomo 9/2, 40126 Bologna, Italy. [2]Department BiGeA, University of Bologna, Via Selmi 3, 40126 Bologna, Italy. [3]Department BCA, University of Padova, Viale Università 16, 35020 Legnaro (PD), Italy.

Published: 23 June 2016

References

1. Lu YF, Goldstein DB, Angrist M, Cavalleri G. Personalized medicine and human genetic diversity. Cold Spring Harb Perspect Med. 2014;4:a008581.
2. Ashley EA. The precision medicine initiative: a new national effort. JAMA. 2015;313:2119–20.
3. Brookes AJ, Robinson PN. Human genotype-phenotype databases: aims, challenges and opportunities. Nat Rev Genet. 2015;16:702–15.
4. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25:2744–50.
5. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009;30:1237–44.
6. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
7. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011; 32:358–68.
8. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35:3823–35.
9. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. Protherm and Pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res. 2006;34:D204–6.
10. Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics. 2008;9 Suppl 2:S6.
11. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics. 2011;12:151.
12. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30:335–42.
13. Laimer J, Hiebl-Flach J, Lengauer D, Lackner P. MAESTROweb: a web server for structure based protein stability prediction. Bioinformatics. 2016. [Epub ahead of print].
14. Khan S, Vihinen M. Performance of protein stability predictors. Hum Mutat. 2010;31:675–84.
15. Fariselli P, Martelli PL, Savojardo C, Casadio R. INPS: predicting the impact of non-synonymous variations on protein stability from sequence. Bioinformatics. 2015;31:2816–21.
16. Casadio R, Vassura M, Tiwari S, Fariselli P, Martelli PL. Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. Hum Mutat. 2011;32:1161–70.
17. Petukh M, Kucukkal TG, Alexov E. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum Mutat. 2015;36:524–34.
18. Pal LR, Moult J. Genetic basis of common human disease: insight into the role of missense SNPs from genome-wide association studies. J Mol Biol. 2015;427:2271–89.
19. Peng Y, Alexov E. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. Proteins. 2016;84:232–9.

20. Martin AC. Mapping PDB, chains to UniProtKB entries. Bioinformatics. 2005; 21:4297–301.
21. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577–637.
22. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins. 1994;20:216–26.
23. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins. 2001;44:79–96.
24. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012;40(Database issue): D841–6.
25. Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol. 2007;369:1318–32.
26. Gong S, Blundell TL. Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One. 2010;5:e9186.
27. David A, Razali R, Wass MN, Sternberg MJ. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. Hum Mutat. 2012;33:359–63.
28. Wei Q, Xu Q, Dunbrack Jr RL. Prediction of phenotypes of missense mutations in human proteins from biological assemblies. Proteins. 2013;81:199–213.
29. Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E. Molecular mechanisms of disease-causing missense mutations. J Mol Biol. 2013;425:3919–36.
30. Bartoli L, Martelli PL, Rossi I, Fariselli P, Casadio R. The prediction of protein-protein interacting sites in genome-wide protein interaction networks: the test case of the human cell cycle. Curr Protein Pept Sci. 2010;11:601–8.
31. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol. 2012;30:159–64.

6 Solving CAGI challenges with INPS-3D

6.1 Introduction

The Critical Assessment of Genome Interpretation (CAGI, \'kā-jē\, https://www.genomeinterpretation.org/) is an international experiment with the goal of evaluating computational methods for determining the phenotypic impacts of genomic variants. In particular, the aim of the experiment is to evaluate the capability of state-of-the-art methods to make useful predictions of molecular, cellular, or organismal phenotypes from genomic data. Evaluating the state-of-the-art methods helps in standardizing the predictions by suggesting appropriate assessment methods and defining what is required for an accurate prediction, and also to define bottlenecks in genome interpretation that may suggest opportunities for further researches. The internationality of the challenge is an important feature that helps in engaging researchers from around the world, connecting diverse research areas whose expertise is essential to develop and improve methods for genome interpretation and also to highlight and spread innovations.

Usually, a CAGI experiment is conducted over a period of one or two years, that starts with the identification/development of suitable challenges (release of unpublished data and formulation of related questions) followed by a period during which participants are invited to analyse data and submit predictions. Each CAGI edition is structured in many different experiments (challenges), having a similar workflow that we can generalize as follows: data providers plan and complete some real experiment to evaluate the phenotypic effect of some variants of interest; participants (competitors) are provided with genetic variants for which they compute predictions of the resulting phenotypes, without knowing the results of the real experiment; after the closure of challenges, independent assessors evaluate predictions against the results of real experimental or clinical data made by the data providers. CAGI experiments end with a conference to discuss the outcomes. Finally, participants (data providers, predictors and assessors) are encouraged to publish their finding. Since 2010, five CAGI experiments have been conducted to date. Last year, a special issue of Human Mutation has bene completely dedicated to the CAGI experiments (see Hoskins RA et al., 2017) and new scientific papers dedicated to CAGI 5 edition are now under writing process.

CAGI challenges investigate a wide range of relations among genetic variants and phenotypes: i) challenges on the effect of single-base variants on RNA expression levels and protein activity, ii) challenges on the interpretation of exome and genome sequencing data for assigning complex traits phenotypes or clinical panels, iii) challenges regarding the ability to predict the effect of mutations in cancer driver genes on cell growth, iv) challenges in which participants were asked to identify causative variants for rare diseases in a given gene panel.

In the last two CAGI editions we participated into 16 different challenges, here we show the results of 5 challenges in which we use INPS and INPS-3D as principal predictor, introducing the methods proposed in facing these challenges and comparing the results considering the diverse datasets.

6.2 Material and Methods

6.2.1 INPS and INPS-3D

A typical challenge in CAGI experiment consists in determining the stability of a protein variant compared to the stability of the wild type protein. This difference is called ΔΔG value and it corresponds to the difference in unfolding free energy between the variant and wild-type proteins for each variant. To predict the ΔΔG we based our approaches on two main types of analysis: the prediction of stability changes in mutated proteins using INPS (Fariselli P et al, 2015), its variant INPS-3D (Savojardo C et al, 2016) and, when possible, the structural analysis of the protein variations using 3D models of protein structure.

INPS (Impact of Non-synonymous mutations on Protein Stability) is a predictor for the impact of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability. INPS is based on a Support Vector Regression (SVR) approach, trained on seven features extracted from the protein primary sequence, including: BLOSUM substitution score, hydrophobicity (wild type and variants), Dayhoff mutability index of wild type, molecular weights of wild type and variant and evolutionary information derived from multiple sequence alignments.

INPS-3D (Impact of Non-synonymous mutations on Protein Stability - 3D) is a method for predicting the impact of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability, starting from 3D structure (see chapter 5).

6.2.2 The challenges

We used the predictions of INPS and INPS-3D in 5 different CAGI challenges (called by the CAGI commission NAGLU, NPM-ALK, FRATAXIN, CALM, TPMT-PTEN). Among these challenges, only in the FRATAXIN experiment the actual goal was to predict exactly the ΔΔG comparing it to the ΔΔG experimentally measured. In fact, in the other challenges we used INPS and INPS-3D predictions also to estimate other protein characteristic that are partially related to protein stability: i) protein relative abundance in the cell (as a measure of protein stability (TPMT-PTEN challenge) ii) protein variants activity as a ratio over the activity of the wild type (NAGLU challenge) iii) protein variant activity as a result of a competitive growth assay (CALM1 and NPM-ALK challenges).

Predicting the impact of a non-synonymous variant on protein functionality is a very complex task, first of all because the definition of the "functionality" of a protein is complex. Though proteins have a vast range of structures and functions, most proteins share a key requirement: they must be stable enough to perform their role in the cell. Mutations that interfere with thermodynamic stability or folding often cause accelerated turnover and lowered steady-state abundance in cells. Consequently, stability-related reduced protein abundance is a major cause of loss-of-function in monogenic disease (Yue P et al, 2005).

6.3 RESULTS

6.3.1 FRATAXIN

Frataxin is a highly conserved protein found in prokaryotes and eukaryotes that is required for efficient regulation of cellular iron homeostasis. Reduced expression of frataxin is the cause of Friedreich's Ataxia (FRDA), a lethal neurodegenerative disease. 8 single amino acids variants of frataxin have been associated to FRDA (Corey DR, 2016). On the other side, the role of frataxin in cancer is still ambiguous: studies have shown that frataxin protects tumour cells against oxidative stress and apoptosis, but also acts as a tumour suppressor. (Schulz TJ et al, 2006; Guccini I et al, 2011)

The 8 single amino acid variants included in the Frataxin challenge were selected from the COSMIC (Catalog of Somatic Mutations in Cancer) database. These are somatic variants associated with neoplastic diseases and/or detected in cancer tissues. For each variant, participants were asked to predict the ΔΔG value, which is the difference in unfolding free energy between the mutant and wild-type proteins, in kcal/mol.

For this challenge, we used as protein 3D structure the PDB (Berman HM et al, 2002) entry 1EKG, chain A.

First we aligned the protein sequence against Uniref90 (The UniProt Consortium, 2017) to obtain a Multiple sequence alignment (MSA). We made the predictions using INPS-3D, and we assigned to each prediction value a standard deviation of 0.5 to highlight that we trust our predictor.

After the release of experimental values of ΔΔG for each of the 8 variants, we compute some statistics to assess our methods. We obtain a Pearson correlation of 0.71 and a Spearman of 0.62, which are enough good to still trust our approach.

6.3.2 TPMT and PTEN

Thiopurine S-methyl transferase (TPMT) is a single domain enzyme involved in the metabolism of thiopurine drugs (Coelho T et al, 2016). Its product is the 6-mercaptopurine, which inhibits de novo purine synthesis leading to cell death. 6-mercaptopurine has been used as a chemotherapeutic agent for Acute-Lymphoblastic Leukemia (ALL) for decades and azathioprine which is converted to 6-mercaptopurine is used to treat autoimmune diseases and to prevent organ rejection after transplant. Overdose with thiopurines leads to treatment interruptions that cause poorer health outcomes and in some cases a life-threatening myelosuppression and hepatotoxicity (Relling MV et al, 2006).

PTEN (Phosphatase and TEnsin Homolog) dephosphorylates phosphatidylinositol (3,4,5)-triphosphate (PIP3), an important secondary messenger molecule promoting cell growth and survival through signalling cascades including those controlled by AKT and mTOR (Song MS et al, 2012). Its important regulatory roles in pro-oncogenic processes results in high rates of PTEN missense mutation in diverse cancers including glioma, endometrial cancer, and melanoma. Germline variation in PTEN results in a collection of developmental abnormalities grouped as PTEN Hamartoma Tumor Syndromes (PHTS) (Eng C, 2003), and is also associated with autism (Butler MG et al, 2005).

For this challenge the data provider is the Fowler laboratory (Fowler DM, Fields S, 2014), that decided to measure the stability of the variant protein as the abundance of the fusion protein and thus the EGFP level of the cell, a protein property that has the advantages of being both informative of variant effect and generalizable to many proteins. To do so, a library of thousands of PTEN and TPMT mutations was assessed to measure the stability of the variant protein using a multiplexed variant stability profiling (VSP) assay, which detects the presence of EGFP fused to the mutated PTEN and TPMT protein respectively.

The dataset is composed by 3,736 PTEN and 2,924 TPMT missense/stop-gain variants.

So practically, the aim of the challenge is to predict the effect of each variant on TPMT and PTEN on protein stability via prediction of the abundances of the fusion protein.

We used protein sequences derived from UniProtKB **(**P51580 for TPMT and P60484 for PTEN) and the protein structure from PDB 2BZG, chain A (TPMT) and 1D5R, chain A (PTEN).

Our approach differs for missense and stop-gain variants: for missense variants we use the prediction of INPS-3D if variant in 3D structure and of INPS for the remaining variants; for stop-gain variants we predict the stability as the ratio of the variant length over the WT length.

For the calibration procedure of the final scores, we used 11 known protein variants of TPMT reported in a functional characterization study (Salavaggione OE et al, 2005) and 27 PTEN variants collected from UniProt and from a functional characterization study (Lee JO et al, 1999). INPS and INPS-3D outcomes on these variants are used to fit a linear model for each protein to remap raw stability change predictions onto the requested range (1=wild type, 0=totally destabilizing, >1 stabilizing).

After the releasing of the experimental results, we assess our prediction computing the Pearson correlation coefficient. Considering TPMT, the Pearson coefficient is 0.40; for the PTEN dataset is 0.51, while considering the two protein dataset together we obtain a Pearson coefficient of 0.44. If we consider only the missense variant, the Pearson coefficient over the two datasets increases to 0.46.

One of the major limits of this challenge is that we used a predictor of ΔΔG to score the abundance of protein variants, but stability and abundance are two different concepts. In fact, many other mechanisms have an effect on protein abundances.

6.3.3 NAGLU

NAGLU is a lysosomal glycohydrolyase that hydrolyzes N-acetyl D-glucosamine from the non-reducing end of heparan sulfate (HS). In humans, deficiency of NAGLU may lead to a rare disorder called Mucopolysaccharidosis IIIB or Sanfilippo B disease (O'Brien JS, 1972; von Figura K, Kresse H, 1972; Valstar MJ et al, 2008) an autosomal recessive disorder affecting lysosomal storage. Specifically, lysosomal HS accumulation causes a neurodegenerative disease whose clinical presentation is associated with many symptoms: from intellectual disability to dementia, including behavioural disturbances. The clinical panel is very negative, because NAGLU deficiency may also lead to death in the second or third decade.

BioMarin Pharmaceutical functionally assessed the enzymatic activity of each of the 165 novel missense mutations in the ExAC dataset.

The challenge consists in predicting the ratio of the activity of the mutated protein over the wild type enzyme; the prediction is a numeric value ranging from 0 (no activity) to 1 (wild-type level of activity) or greater than 1 if the predicted activity is greater than wild-type activity (e.g. 0.5 means 50% of wild-type and 1.5 means 150% of wild-type activity). The predictions are assessed against the numeric values actually measured for each mutation in the enzyme assay.

Recently the structure of NAGLU became available (Ficko-Blean E et al, 2008) with a good resolution of 2.9 Å.

Our approach consists in predicting the protein stability with INPS-3D and then calibrate these predictions using a dataset of 308 NAGLU variants derived from UniProt, ClinVar (Landrum MJ et al, 2016), HGMD (Stenson PD et al, 2012) and dbSNP (Sherry ST et al, 2001), in which 87 protein variants over 308 are already associated with the Sanfilippo B disease.

We used the distribution of their scores computed with INPS-3D to define the threshold that discriminates functional protein variants from not stable and not functional protein variants.

We also consider an expert-based structural analysis taking into consideration the relative solvent accessibility area of each mutated residue and the proximity towards important residues of the active site that are known to impact the protein function

We assess our method towards experimental results: we obtained a low Pearson correlation coefficient (0.24) and a good Spearman coefficient (0.65).

6.3.4 CALM1

Calmodulin is a calcium-sensing protein encoded by the human genes Calmodulin1 (CALM1), Calmodulin2 (CALM2), and Calmodulin3 (CALM3), each encoding exactly the same calmodulin protein sequence. Calmodulin is involved in many different cellular processes, and is especially important for neuron and muscle cell function. Calmodulin has high clinical relevance, as variants of the protein are causally associated with two cardiac arrhythmias: catecholaminergic ventricular tachycardia (Nyegaard M et al, 2012) and long QT syndrome (Crotti L et al, 2013).

A team in Fritz Roth’s Lab at the Donnelly Centre (U. Toronto) and Lunenfeld Tanenbaum Research Institute (Sinai Health Systems), has assessed a large library of calmodulin variants using a high-throughput yeast complementation assay. This assay reveals the overall impact of each variant on the ability of the protein to function in the cell. The functionality of the protein variant is measured with a complementation assay based on ability of calmodulin variants to rescue a yeast strain carrying a temperature-sensitive allele of the yeast calmodulin orthologue CMD1 (Sun S et al, 2016). In fact, CMD1 is an essential gene, and at the restrictive temperature, CMD1 temperature-sensitive mutants do not grow.

The yeast-based functional assays were established and validated in a previous study (Sun S et al, 2016), and they were also validated for the ability to separate pathogenic from non-pathogenic variants (Weile J et al. 2017). The functionality is quantified as a "fitness score, that is a log ratio scaled such that 1 represents full function and 0 represents complete loss of function.

The final provided dataset contains 1,813 variants of human Calmodulin.

The challenge consists in predicting the fitness score (0 complete loss of function, 1 wild-type). We use the protein structure 1CLL, chain A, stored in PDB.

We considered stabilizing ΔΔG (ΔΔG>0) as negative predictions of the fitness scores, and we changed their sign, taking the opposites (ΔΔG value v transformed as -|v|). We directly normalized the predictions in the range 0-1, as requested by the challenge, where 0 means no growth at the restrictive temperature and 1 a wildtype-like growth fitness.

After the releasing of the experimental data, we compute some statistics to assess our method. The Pearson correlation coefficient result to be very low (0.17).

The main critical issues were: i) the experiment has been performed in yeast system, not using human cells; ii) the experimental scores are not a direct measure of a percentage of wild type activity, but the results of a competitive growth assay. We tried to score the result of a competitive growth assay using a predictor of protein stability, and in this specific case the outcome suggests that the approach should be revise because our predictions are not so generalizable.

6.3.5 NPM- ALK

NPM-ALK is a fusion gene originally described in positive anaplastic large cell lymphoma (ALCL). In this tumour, the presence of an NPM oligomerization domain promotes ligand-independent NPM-ALK dimerization, leading to ligand-independent activation of ALK, resulting in constitutive kinase activity, self-phosphorylation and continuous signalling. Although the physiological function and regulation of full-length kinase ALK remains poorly characterized, aberrant expression of constitutively activated NPM-ALK has been clearly established as the leading cause of ALK-positive ALCL.

However, recent studies suggest that inhibitor efficacy may be hampered by several resistance mechanisms including point mutations in ALK (Lovisa F et al, 2015; Lu L et al, 2009). In this context, the inhibition of the molecular chaperone Hsp90 represents an alternative approach to overcoming resistance to kinase inhibitors, since NPM-ALK, like many other kinases, is strictly dependent on molecular chaperones for its maturation and activity (Bonvini P et al, 2002). Conformational stability of ALK is known to be maintained by Hsp90, but the principles of this interaction, the specific domains or motifs recognized, and the impact of mutations on chaperone activity remain obscure.

The Bonvini laboratory has examined the kinase activity and Hsp90 binding affinity of a series of NPM-ALK constructs harbouring single amino acid mutations, multiple amino acid mutations, or deletions in the ALK catalytic domain to define the manner by which nascent NPM-ALK kinase is recognized by Hsp90, and how Hsp90 helps to facilitate NPM-ALK folding, activity, and/or stability. Structural motifs and specific residues in or immediately adjacent to the NPM-ALK catalytic domain were analysed (Bonvini P et al, 2004; Tartari CJ et al, 2008) to identify the determinants of Hsp90 interaction based on the tendency of NPM-ALK to fold.

Participants were asked to submit predictions of both the kinase activity and the Hsp90 binding affinity of each mutant protein relative to the reference.

Our approach starts with the structural analysis of the protein, taking into consideration the relative solvent accessibility area of each variation and the lateral side chain distance to the ATP-binding site of the domain. Also the proximity towards important motifs, like the well-known Y-x-x-x-Y-Y motif called A-loop, was taken into account to define the impact of each variation on protein function and its influence in the affinity to Hsp90. We use the use the ΔΔG values as predicted by INPS-3D to determine the effect of mutations of the different amino acids. For multiple mutations, we consider the mutation with the more severe effect.

Finally, the confidence of each prediction was determined considering the information about specific protein variations described in literature, when available.

Thanks to this approach that is highly manually curated and very protein specific, we obtained a Pearson Coefficient of 0.85 in the task of determining the protein activity.

The prediction of the binding affinity towards Hsp90 on the contrary performs badly, and we may speculate that we lack the structural analysis of the interface between the NPM oligomerization domain/s and the ALK catalytic domain/s as well as the oligomerization surface of the NPM-ALK fusion protein.

6.4 Conclusion

The interest in the genotype-phenotype relation is increasing, in particular for the possibility of predicting with a computational approach the results of in vitro experiment. Testing and scoring the available predictors with specific datasets and results of experiments is important for understanding the state of the art, to define our confidence in computational methods and to highlights bottlenecks and issues that need further studies.

Using the predictions of INPS and INPS-3D in different challenges, we confirm that the tools perform well when used specifically for the task for which they have been trained (the prediction of protein stability). When we try to correlate protein stability to protein abundance or protein activity, the performance is not so good, but we can improve our computational approach when we perform a structural analysis of the protein. If the computational approach is enriched which protein structure analysis and manual curation, including the available knowledge find in literature, the performance may get better also when we use protein stability to predict protein activity (e.g. NPM-ALK challenge if compared with the results of CALM1 challenge). In conclusions, we are in the right direction to fill the gap between computational predictions and in vitro experiment when we train predictors for very specific task and we use them for the same kind of experiments (e.g. protein stability experiment and ΔΔG predictor like INPS-3D). To enlarge the analysis to related experiments, we need to include manual curation and protein structural analysis.

7 Predicting phenotypes from exomes

7.1 Contribution to the state of the art

We present the assessment of three different challenges of the Critical Assessment Genome Interpretation 4 edition (CAGI4, see paragraph 6.1) involving exome-sequencing data: Crohn's disease, bipolar disorder, and warfarin dosing. 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 prediction evaluation: the lessons learned from the exome challenges can be applied to both research and clinical efforts to improve phenotype prediction from genotype. This is a step forward in the direction of precision medicine, aiming to predict a patient's disease risk and best therapeutic options by using that individual's genetic sequencing data.

7.2 General information on the paper

The presented paper can be found in the following publication:

Authors: Roxana Daneshjou, YanranWang, Yana Bromberg, Samuele Bovo, Pier L. Martelli, Giulia Babbi, Pietro Di Lena, Rita Casadio, Matthew Edwards, David Gifford, David T. Jones, Laksshman Sundaram, Rajendra Rana Bhat, Xiaolin Li, Lipika R. Pal, Kunal Kundu, Yizhou Yin, John Moult, Yuxiang Jiang, Vikas Pejaver, Kymberleigh A. Pagel, Biao Li, Sean D. Mooney, Predrag Radivojac, Sohela Shah, Marco Carraro, Alessandra Gasparini, Emanuela Leonardi, Manuel Giollo, Carlo Ferrari, Silvio C. E. Tosatto, Eran Bachar, Johnathan R. Azaria, Yanay Ofran, Ron Unger, Abhishek Niroula, Mauno Vihinen, Billy Chang, Maggie H. Wang, Andre Franke, Britt-Sabina Petersen, Mehdi Pirooznia, Peter Zandi, Richard McCombie, James B. Potash, Russ B. Altman, Teri E. Klein, Roger A. Hoskins, Susanna Repo, Steven E. Brenner and Alexander A. Morgan

Title: Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges

Journal: Human Mutation

Volume: 38 **Issue:** 9 **Pages:** 1182–1192

Year: 2017

Impact Factor: 5.359

Quartile and subject: 1st quartile in Genetics and Heredity

DOI: 10.1002/humu.23280

Received: 5 April 2017 | Revised: 12 June 2017 | Accepted: 12 June 2017

DOI: 10.1002/humu.23280

SPECIAL ARTICLE

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 (≤35 mg or ≥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*.

$$\frac{\sum_{i=1}^{n}\left|\frac{\text{dose\_actual}_i-\text{dose\_predicted}_i}{\text{SD\_predicted}_i}\right|}{n} \tag{1}$$

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).

$$\frac{\sum_{i=1}^{n}\frac{\text{SD\_predicted}_i}{\text{dose\_predicted}_i}}{n} \tag{2}$$

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.

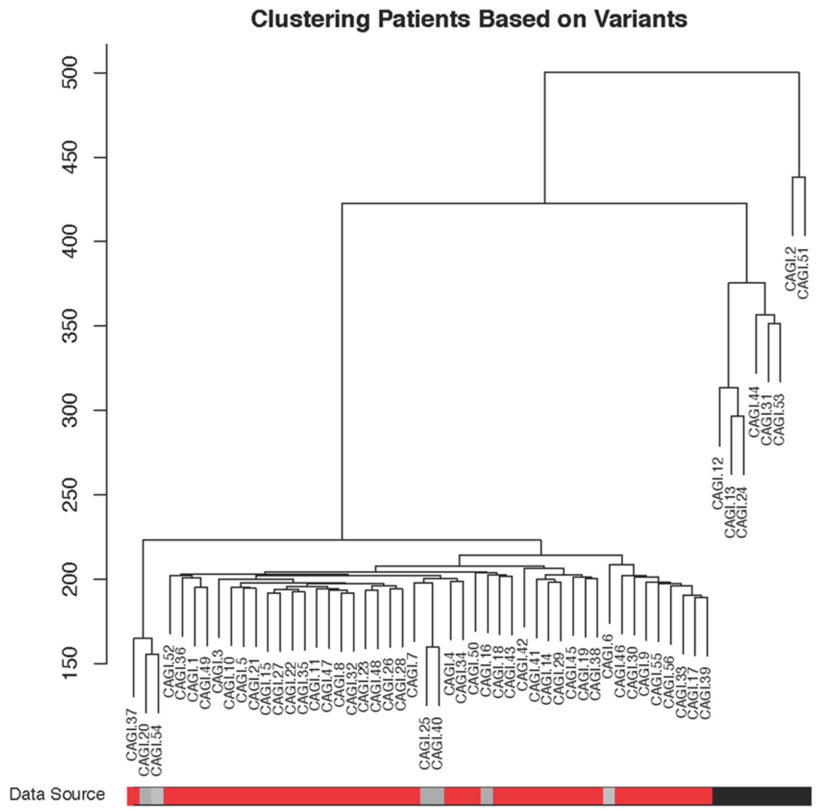


FIGURE 1 Clustering of patients from the CAGI 2 Crohn's disease challenge. The black and gray bars at the bottom represent the controls; the red represents the cases. Many of the controls cluster together, likely due to batch effects. For instance, the controls represented in black were sequenced separately 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

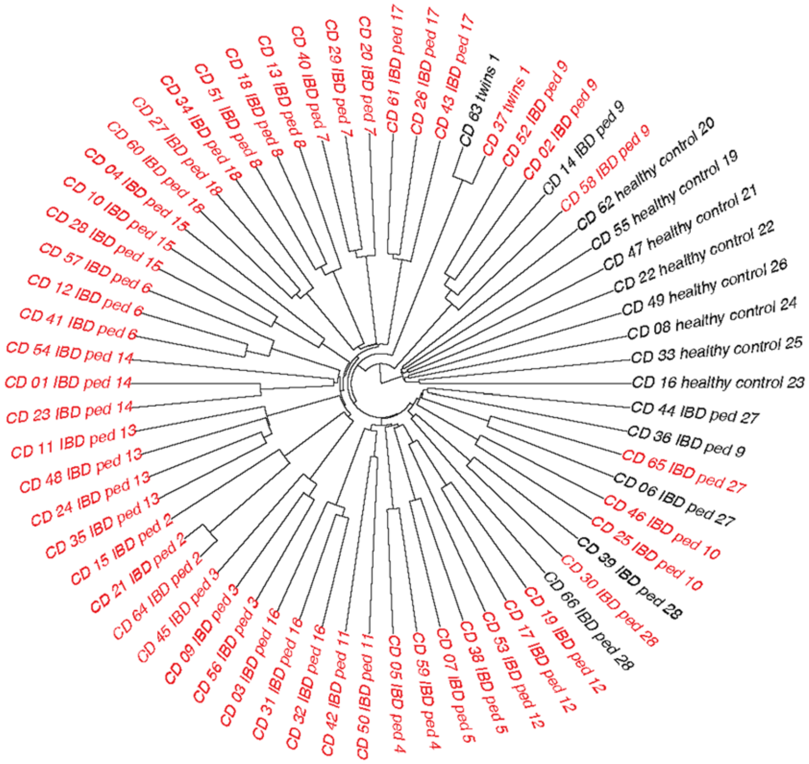


FIGURE 2 Clustering of samples for CAGI 3 Crohn's disease challenge. Black represents controls, whereas red represents cases. This dataset included healthy family members of cases as well as random controls. Samples with a "ped" designation in the sample name came from a pedigree; samples that share 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

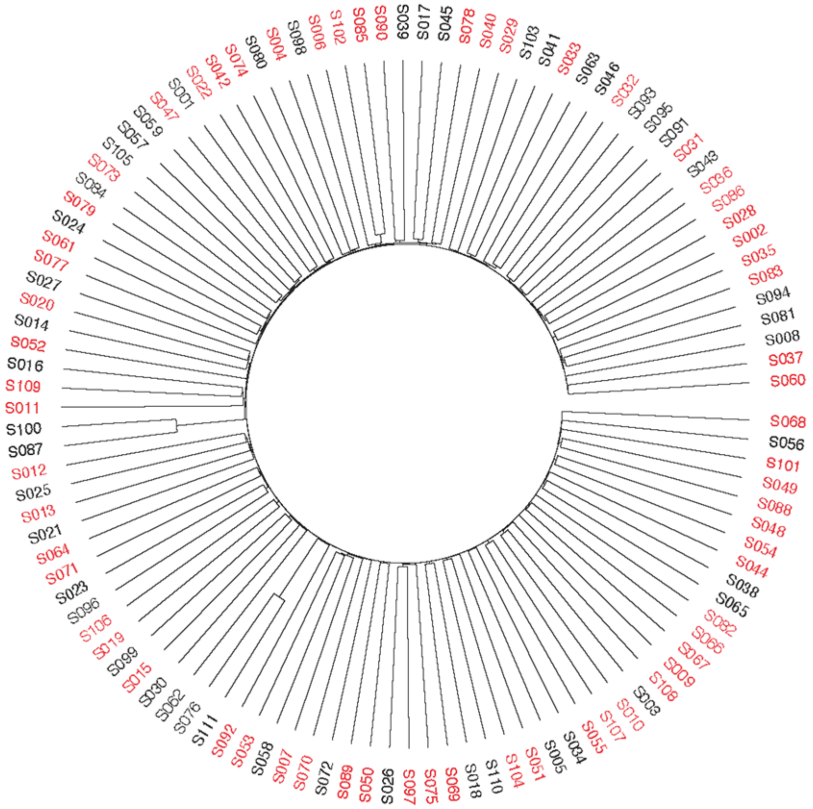


FIGURE 3 Clustering of samples for CAGI 4 Crohn's disease challenge. Black represents controls, and red represents cases

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 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

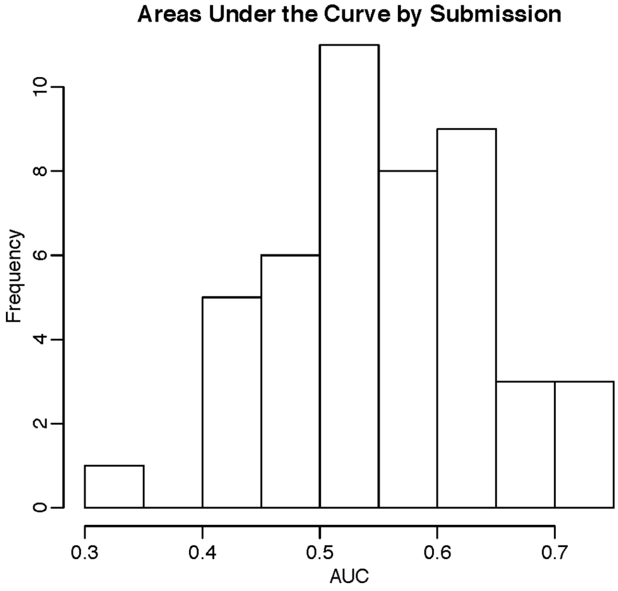


FIGURE 4 CAGI 4 Crohn's disease challenge distribution of AUCs across all methods

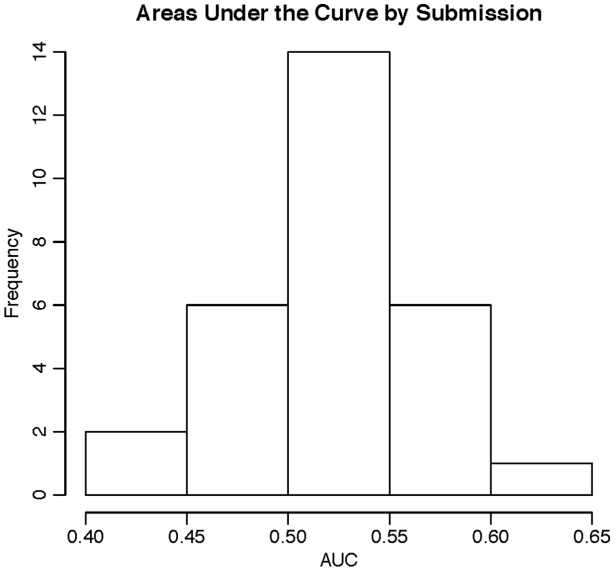


FIGURE 5 CAGI 4 bipolar disorder challenge distribution of AUCs across all methods

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

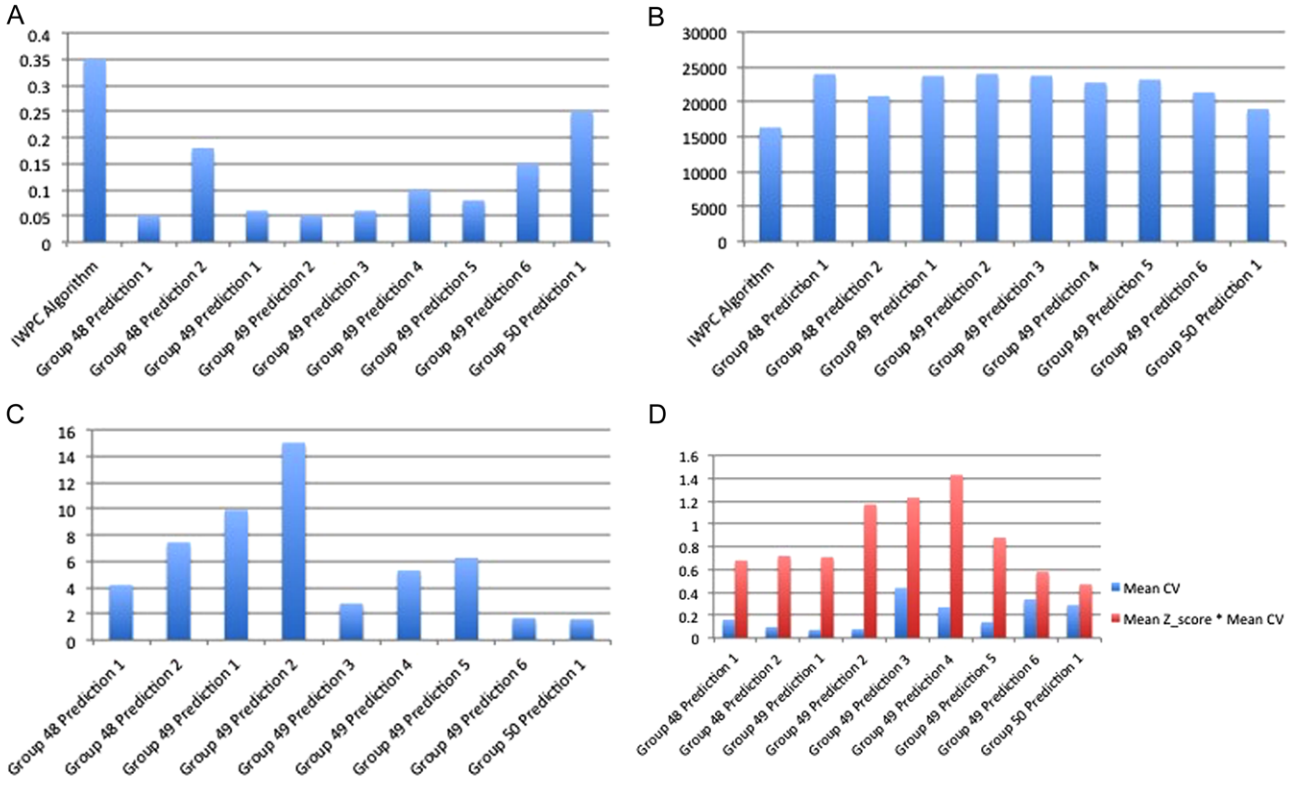


FIGURE 6 **A**: R^2 between predicted doses and actual doses for each group's prediction method as well as the IWPC algorithm. **B**: Sum of squared errors for each group's prediction method and the IWPC algorithm. **C**: Mean *z* scores calculated from each group's predicted doses with predicted standard 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

8 Prediction of allosteric effect in liver pyruvate kinase

8.1 Contribution to the state of the art

Here we present the results of the computational prediction of the allosteric effect of liver pyruvate kinase variants, presented by four different groups at the Critical Assessment Genome Interpretation 4 edition (CAGI4, see paragraph 6.1). 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 highlights a specialized need for new computational tools and utilization of benchmarks that focus on allosteric regulation.

8.2 General information on the paper

The presented paper can be found in the following publication:

Authors: Qifang Xu, Qingling Tang, Panagiotis Katsonis, Olivier Lichtarge, David Jones, Samuele Bovo, Giulia Babbi, Pier L. Martelli, Rita Casadio, Gyu Rie Lee, Chaok Seok, Aron W. Fenton and Roland L. Dunbrack Jr

Title: Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4

Journal: Human Mutation

Volume: 38 **Issue:** 9 **Pages:** 1123–1131.

Year: 2017

Impact Factor: 5.359

Quartile and subject: 1st quartile in Genetics and Heredity

DOI: 10.1002/humu.23222

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

DOI: 10.1002/humu.23222

SPECIAL ARTICLE

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 &

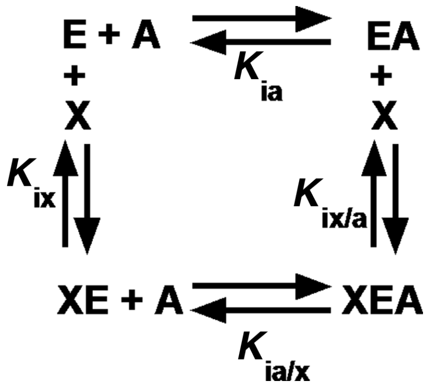


FIGURE 1 Reaction scheme for an allosteric energy cycle in which an enzyme (E) can bind one substrate (A) and one allosteric effector (X). K_{ia} is the equilibrium dissociation constant of the substrate binding to the enzyme in the absence of effector. $K_{ia/x}$ is the equilibrium dissociation constant 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 $Q_{\text{ax-Ala}}$ and $Q_{\text{ax-Fru-1,6-BP}}$

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 ≠ 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 ($n \times (n\text{–}1) / 2$). 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).

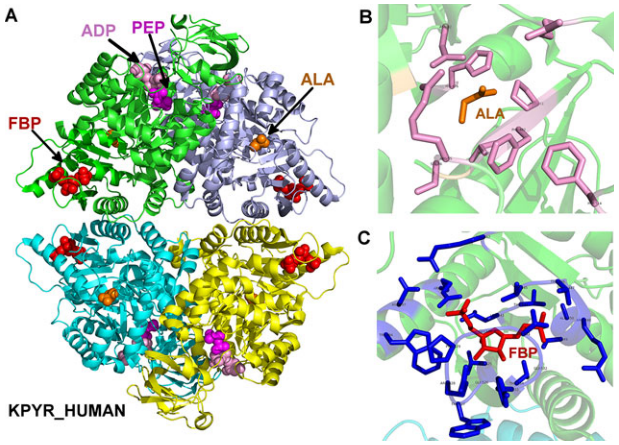


FIGURE 2 Structure of human pyruvate kinase, as well as the binding sites of inhibitor alanine and activator fructose-1,6-bisphosphate. A: A modeled structure of L-PYK tetramer with substrates PEP and ADP, allosteric inhibitor alanine, and allosteric activator. PEP, ADP, alanine (labeled ALA), and fructose-1,6-bisphosphate (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 ≥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

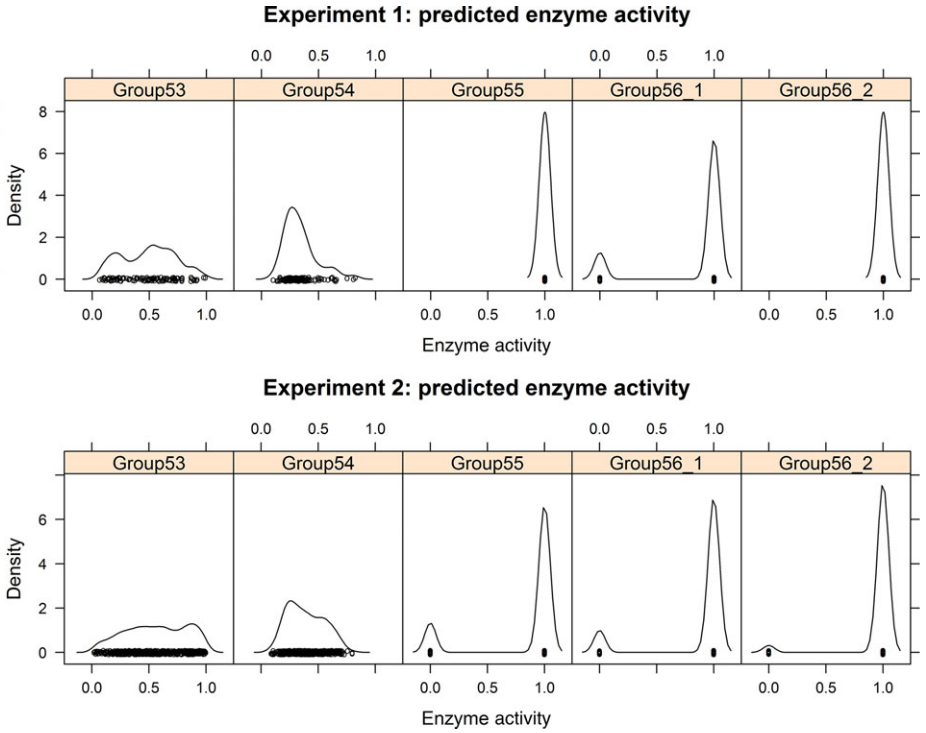


FIGURE 3 Kernel density estimates of five sets of predicted L-PYK enzyme activities

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 ($Q_{\text{ax-Ala}}$)

The second challenge was to estimate the inhibitory allosteric effect of binding alanine, $Q_{\text{ax-Ala}}$ on binding of the substrate PEP. The density estimates of experimental $Q_{\text{ax-Ala}}$ values of two experiments are shown in Figure 4. The wild-type enzyme had a $Q_{\text{ax-Ala}}$ 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 $Q_{\text{ax}} = 1$ (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 $Q_{\text{ax}} = 1$ 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 $Q_{\text{ax-Ala}}$ 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 $Q_{\text{ax-Ala}}$ values is shown in Figure 5. Group 55 provided only binary prediction for $Q_{\text{ax-Ala}}$. 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 $Q_{\text{ax-Ala}}$.

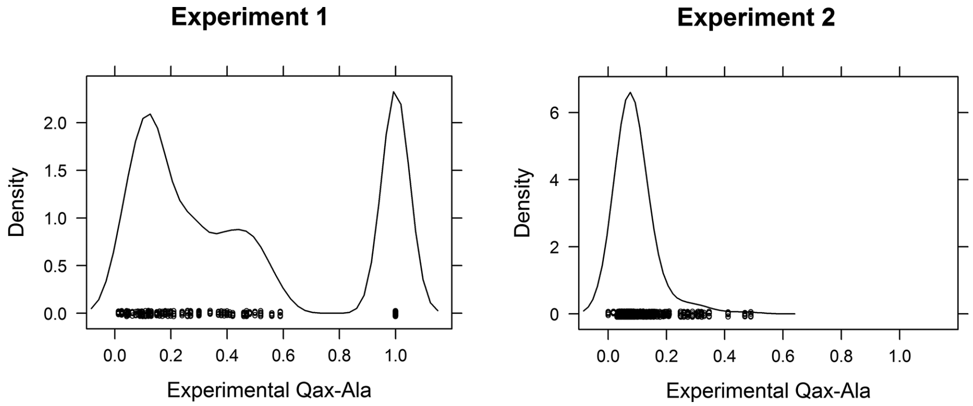


FIGURE 4 Kernel density estimates of experimental $Q_{\text{ax-Ala}}$ values of experiments 1 and 2

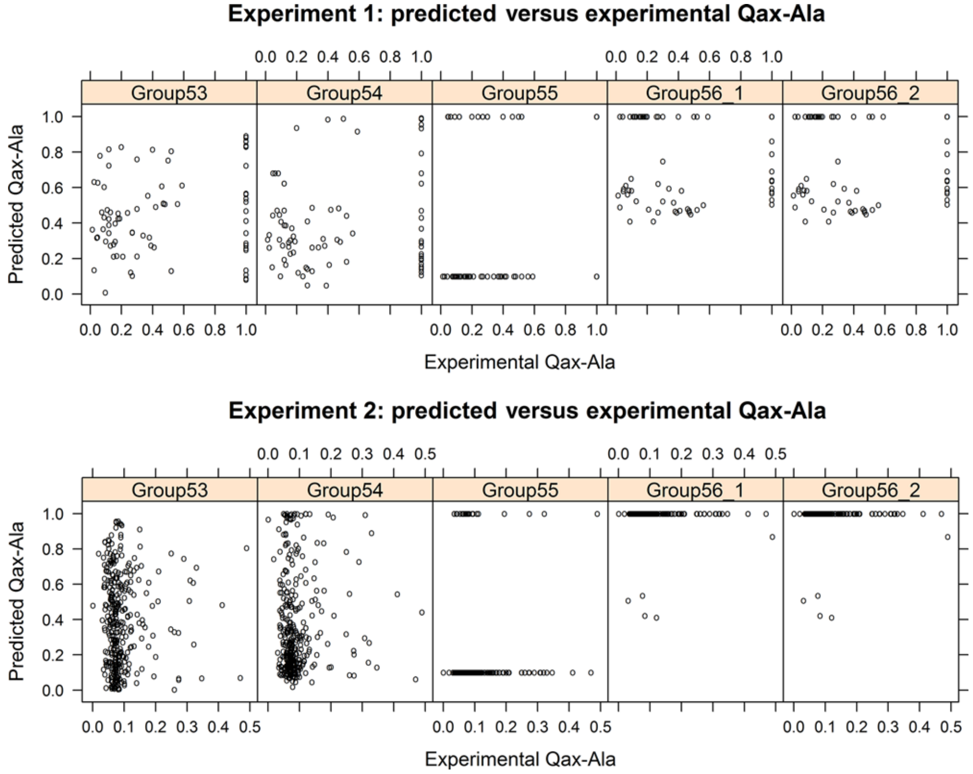


FIGURE 5 Scatter plot of the experimental $Q_{\text{ax-Ala}}$ versus the predicted $Q_{\text{ax-Ala}}$ values

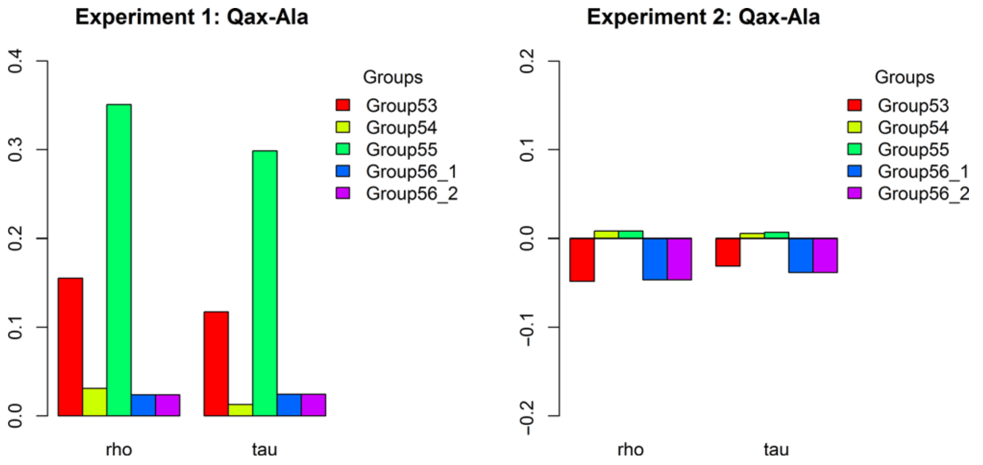


FIGURE 6 Correlations represented by Spearman's ρ and Kendall's τ between the predicted and experimental $Q_{\text{ax-Ala}}$ values of two experiments

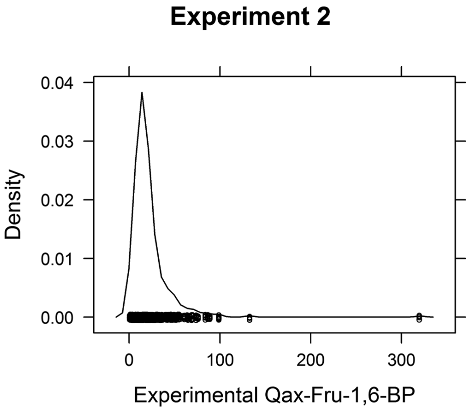


FIGURE 7 Kernel density estimate of experimental $Q_{\text{ax-Fru-1,6-BP}}$ from experiment 2

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 $< 1.0 \times 10^{-15}$).

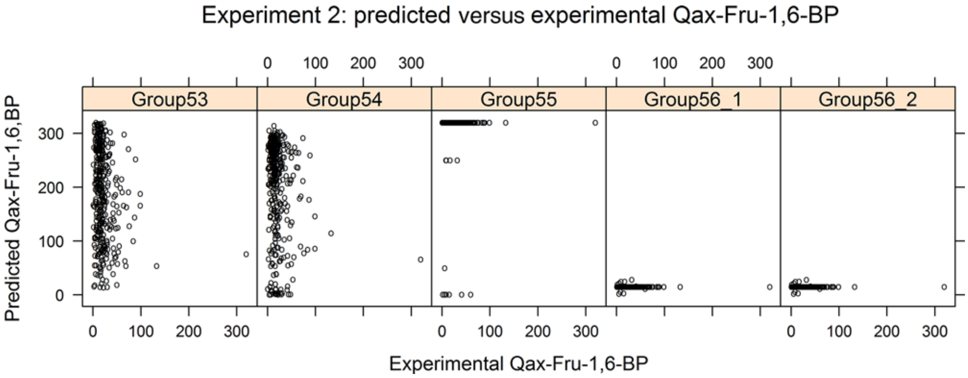


FIGURE 8 Scatter plot of the predicted versus experimental $Q_{\text{ax-Fru-1,6-BP}}$ values from experiment 2

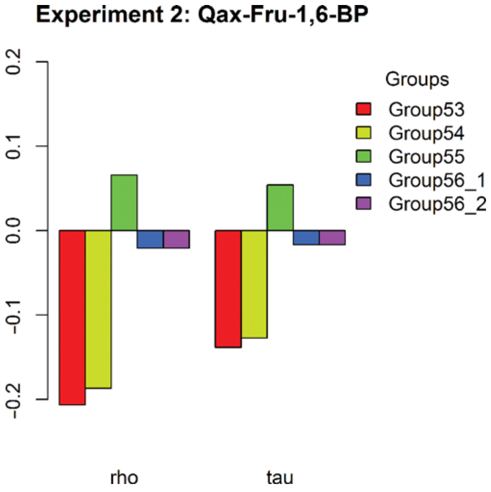


FIGURE 9 Correlations represented by Spearman's ρ and Kendall's τ between the predicted and experimental $Q_{\text{ax-Fru-1,6-BP}}$ values in experiment 2

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 $Q_{\text{ax-Ala}}$ (no inhibition of PEP-binding by Ala) and 15.3 for $Q_{\text{ax-Fru-1,6-BP}}$ (the experimental value). This resulted in only eight mutations with $Q_{\text{ax-Ala}}$ 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 $Q_{\text{ax-Fru-1,6-BP}}$, 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 $Q_{\text{ax-Fru-1,6-BP}}$. The correlation coefficients with $Q_{\text{ax-Fru-1,6-BP}}$ 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

9 Myosin 1F variants in relation with thyroid cancer

9.1 Contribution to the state of the art

Here we present a study that identifies for the first time a role for myosin 1F (MYO1F) in Familial Non-Medullary Thyroid Cancer (FNMTC). Whole exome sequencing analysis in the family affected by FNMTC with oncolytic feature reveals a novel heterozygous mutation (c.400G > A, NM_012335; p.Gly134Ser) in exon 5 of MYO1F. We experimentally observed an altered mitochondrial network in cell model expressing the mutant MYO1F p.Gly134Ser protein, leading to increased mitochondrial mass and a significant increase in both intracellular and extracellular reactive oxygen species, compared to cells expressing the wild-type (wt). These phenotypic effects conferred a significant advantage in colony formation, invasion and anchorage-independent growth. These results were corroborated by in vivo studies in zebrafish. Thanks to an additional screening of 192 FNMTC families we identify another variant in MYO1F exon 7, which leads to exon skipping, and we computationally build the model of the protein variant and predict the alteration of the ATP-binding domain in MYO1F.

9.2 General information on the paper

The presented paper can be found in the following publication:

Authors: Chiara Diquigiovanni, Christian Bergamini, Cecilia Evangelisti, Federica Isidori, Andrea Vettori, Natascia Tiso, Francesco Argenton, Anna Costanzini, Luisa Iommarini, Hima Anbunathan, Uberto Pagotto, Andrea Repaci, Giulia Babbi, Rita Casadio, Giorgio Lenaz, Kerry J. Rhoden, Anna M. Porcelli, Romana Fato, Anne Bowcock, Marco Seri, Giovanni Romeo and Elena Bonora

Title: Mutant MYO1F alters the mitochondrial network and induces
tumor proliferation in thyroid cancer

Journal: International Journal of Cancer

Volume: 143 **Pages:** 1706–1719

Year: 2018

Impact Factor: 7.360

Quartile and subject: 1st quartile in Oncology

DOI: 10.1002/ijc.31548.

IJC
International Journal of Cancer

Mutant MYO1F alters the mitochondrial network and induces tumor proliferation in thyroid cancer

Chiara Diquigiovanni[1], Christian Bergamini[2], Cecilia Evangelisti[3], Federica Isidori[1], Andrea Vettori[4], Natascia Tiso[4], Francesco Argenton[4], Anna Costanzini[1,2], Luisa Iommarini[2], Hima Anbunathan[5], Uberto Pagotto[1], Andrea Repaci[6], Giulia Babbi[2], Rita Casadio[2], Giorgio Lenaz[3], Kerry J. Rhoden[1], Anna Maria Porcelli[2], Romana Fato[2], Anne Bowcock[5], Marco Seri[1], Giovanni Romeo[1] and Elena Bonora[1]

[1] Department of Medical and Surgical Sciences, DIMEC, St. Orsola-Malpighi Hospital, University of Bologna, Bologna, Italy
[2] Department of Pharmacy and Biotechnology, FABIT, University of Bologna, Bologna, Italy
[3] Department of Biomedical and Neuromotor Sciences, DIBINEM, University of Bologna, Bologna, Italy
[4] Department of Biology, University of Padova, Padova, Italy
[5] National Heart and Lung Institute, Imperial College, London, United Kingdom
[6] Endocrinology Unit, St. Orsola-Malpighi Hospital, Bologna, Italy

Familial aggregation is a significant risk factor for the development of thyroid cancer and familial non-medullary thyroid cancer (FNMTC) accounts for 5–7% of all NMTC. Whole exome sequencing analysis in the family affected by FNMTC with oncocytic features where our group previously identified a predisposing locus on chromosome 19p13.2, revealed a novel heterozygous mutation (c.400G>A, NM_012335; p.Gly134Ser) in exon 5 of *MYO1F*, mapping to the linkage locus. In the thyroid FRTL-5 cell model stably expressing the mutant MYO1F p.Gly134Ser protein, we observed an altered mitochondrial network, with increased mitochondrial mass and a significant increase in both intracellular and extracellular reactive oxygen species, compared to cells expressing the wild-type (wt) protein or carrying the empty vector. The mutation conferred a significant advantage in colony formation, invasion and anchorage-independent growth. These data were corroborated by *in vivo* studies in zebrafish, since we demonstrated that the mutant MYO1F p.Gly134Ser, when overexpressed, can induce proliferation in whole vertebrate embryos, compared to the wt one. *MYO1F* screening in additional 192 FNMTC families identified another variant in exon 7, which leads to exon skipping, and is predicted to alter the ATP-binding domain in MYO1F. Our study identified for the first time a role for *MYO1F* in NMTC.

Introduction

Familial aggregation is a significant risk factor for the development of thyroid cancer derived from follicular epithelial cells (non-medullary thyroid carcinoma, NMTC). When the primary cancer site is considered, the thyroid gland shows the highest estimate of familial relative risk among all organs (5- to 10-fold compared to 1.8 and 2.7 for breast and colon cancer, respectively).[1] Familial NMTC (FNMTC) accounts for 5–7% of all NMTC and may occur as a part of familial cancer syndromes (familial adenomatous polyposis, Gardner's syndrome, Cowden's disease, Carney's complex type 1, Werner's syndrome and papillary renal neoplasia) or as a primary feature (familial NMTC) (for a review see Ref. 2,3). FNMTC has become a well-recognized, unique clinical entity. Although still debated, there are some epidemiologic and clinical kindred studies that have shown an association between FNMTC and more aggressive behavior than sporadic cases, with higher rates of multicentric tumors, lymph node metastasis and extrathyroidal invasion, a younger age of onset and shorter disease-free survival.[2–4] A search for

Key words: non-medullary thyroid carcinoma, TCO locus, whole exome sequencing, MYO1F, mitochondrial network
Additional Supporting Information may be found in the online version of this article.
Conflict of interest: The authors declare that they have no potential conflicts of interest.
Grant sponsor: Associazione Italiana per la Ricerca sul Cancro (AIRC); **Grant number:** IG2015-17069; **Grant sponsor:** Italian Ministry of Health; **Grant numbers:** GR-2012 "DIANE" and RF-2011-02350857; **Grant sponsor:** EU FP7 Marie Curie project; **Grant number:** MEET-317433
DOI: 10.1002/ijc.31548
History: Received 2 Oct 2017; Accepted 11 Apr 2018; Online 19 Apr 2018

Correspondence to: Marco Seri, Unit of Medical Genetics, Department of Medical and Surgical Sciences, St. Orsola-Malpighi Hospital, University of Bologna, 40138 Bologna, Italy, Tel: +39-0512-088-421, Fax: +39-0512-088-416, E-mail: marco.seri@unibo.it; or Elena Bonora, Unit of Medical Genetics, Department of Medical and Surgical Sciences, St. Orsola-Malpighi Hospital, University of Bologna, 40138 Bologna, Italy, Tel: +39-0512-088-434, Fax: +39-0512-088-416, email: elena.bonora6@unibo.it

What's new?

Evidence suggests that familial non-medullary thyroid carcinoma (FNMTC) is highly heterogeneous, complicating the identification of underlying mutations in family pedigrees. Here, investigation of chromosome 19p13.2, which contains a known thyroid cancer-predisposing locus, led to the identification of a novel mutation in the gene *MYO1F*. Relative to wild-type controls, thyroid cell models carrying mutant *MYO1F* exhibited a significant increase in colony formation and greater potential for invasion and anchorage-independent growth. Mutated cells further showed an altered mitochondrial phenotype, similar to the one observed in human thyroid tumors. The findings suggest that *MYO1F* has a role in thyroid cancer predisposition.

susceptibility genes, undertaken using linkage-based approaches, led to the identification of several predisposing loci: MNG1 (14q32), TCO (19p13.2), fPTC/PRN (1q21), NMTC1 (2q21), FTEN (8p23.1-p22) and the telomere–telomerase complex.[2,3] Mutations were identified at the 14q31 locus in the *DICER1* gene, which encodes for an enzyme required for miRNA maturation.[5] Recent data have also shown that dysregulation of miRNA expression is a hallmark of thyroid cancer[6] and an altered splicing regulation has been reported in FNMTC patients carrying a germline mutation in the *SRRM2* gene, encoding a splicing machinery subunit.[7] Additional studies have identified predisposing risk-variants in non-coding genes, including miRNAs[8] and a long non-coding gene *PTCSC2*.[9] Mutations in genes encoding regulators of the RAS pathway such as *RASAL1*[10] and *SRGAP1*[11] were also identified in FNMTC cases. Taken together, these data indicate that the genetic predisposition to FNMTC is characterized by a high degree of heterogeneity, hampering the identification of the underlying mutations in the corresponding pedigrees.

We previously mapped a predisposing locus for FNMTC on chromosome 19p13.2 in a multigenerational family with multiple individuals affected by thyroid carcinoma with oncocytic features (oxyphilia; TCO), with autosomal dominant inheritance.[12] In our study, we report whole exome sequencing (WES) data and functional studies providing evidence that mutant *MYO1F*, mapping to the TCO locus on chromosome 19p13.2, lead to NMTC.

Materials and Methods

The study was approved by the committee for protection of persons in biomedical research of Lyon (CCPRB A-96.18) and by the IARC Ethical Review Board (Project 95–050, amendment 01–013). Informed consents were obtained by clinicians, in each collaborating center.

Subjects

The TCO family has been previously reported[12] and the main clinical characteristics are reported in the Supporting Information. Papillary thyroid carcinomas (PTCs) were diagnosed in individuals II-5, III-3 and III-7 at the ages of 41, 27 and 11 years, respectively. In total, 192 FNMTC patients included in the mutation screening came from the families collected between 1996 and 2012 through the International Consortium for the Genetics of Non-Medullary Thyroid Carcinoma; 149 female patients and 43 males were included (age of onset: 11–84 years, mean age = 42 years), thyroid cancer diagnosis is reported in Supporting Information Table S1.

WES analysis

WES was performed on three individuals from the TCO family, two affected by thyroid carcinoma with oncocytic features (individuals II-3; III-7, Fig. 1*a*) and one affected by adenoma (II-4), according to the pipeline reported in the Supporting Information. Variants were confirmed by polymerase chain reaction (PCR) and direct sequencing.

Cell lines

The FRTL-5 cell line is a stable thyroid cell line derived from normal thyroid glands from 5 to 6-week-old Fisher rats.[13] All cells were cultured in 6H5 medium consisting of Coon's modified Ham's F12 medium (Sigma-Aldrich, St. Louis, MO) supplemented with 5% newborn calf serum (NCS) (Sigma-Aldrich), 1 μg/ml insulin, 10 nM hydrocortisone, 5 μg/ml apo-transferrin, 10 ng/ml gly-his-lys, 10 ng/ml somatostatin, 1 mU/ml TSH (Sigma-Aldrich, St. Louis, MO) and penicillin/streptomycin (EuroClone, Milan, Italy). Cells were propagated in a fully humidified atmosphere of 5% CO_2 at 37°C.

COS7 cells derived from monkey kidney tissue were grown in Dulbecco's modified Eagle's medium, 10% fetal bovine serum, 2 mM L-glutamine, 100 U/ ml penicillin and 100 μg/ml streptomycin, in a humidified incubator at 37°C with 5% CO_2.

pCMV6-MYO1F p.Gly134Ser plasmid generation *via* site-directed mutagenesis

The construct pCMV6 encoding wild-type (wt) *MYO1F* (RC207069) was purchased from OriGene OriGene Technologies, Rockville, MD) in frame with the tag (polypeptide chain containing the aminoacid sequence Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys, or DYKDDDDK (DDK) and containing neomycin resistance (G418) for stable selection. The mutation c.400G > A was inserted using the Q5 Site-direct Mutagenesis kit, according to the manufacturer's instruction (New England Biolabs, Ipswich, MA) using the oligonucleotides forward 5′-AGGTGTCTGGCGGAAGCGAGAAGGTCCAG-3′ and reverse 5′-TGGAGATGTAGCCCATGATTATTTGGCT-3′. The site-directed mutagenesis was verified by plasmid direct sequencing.

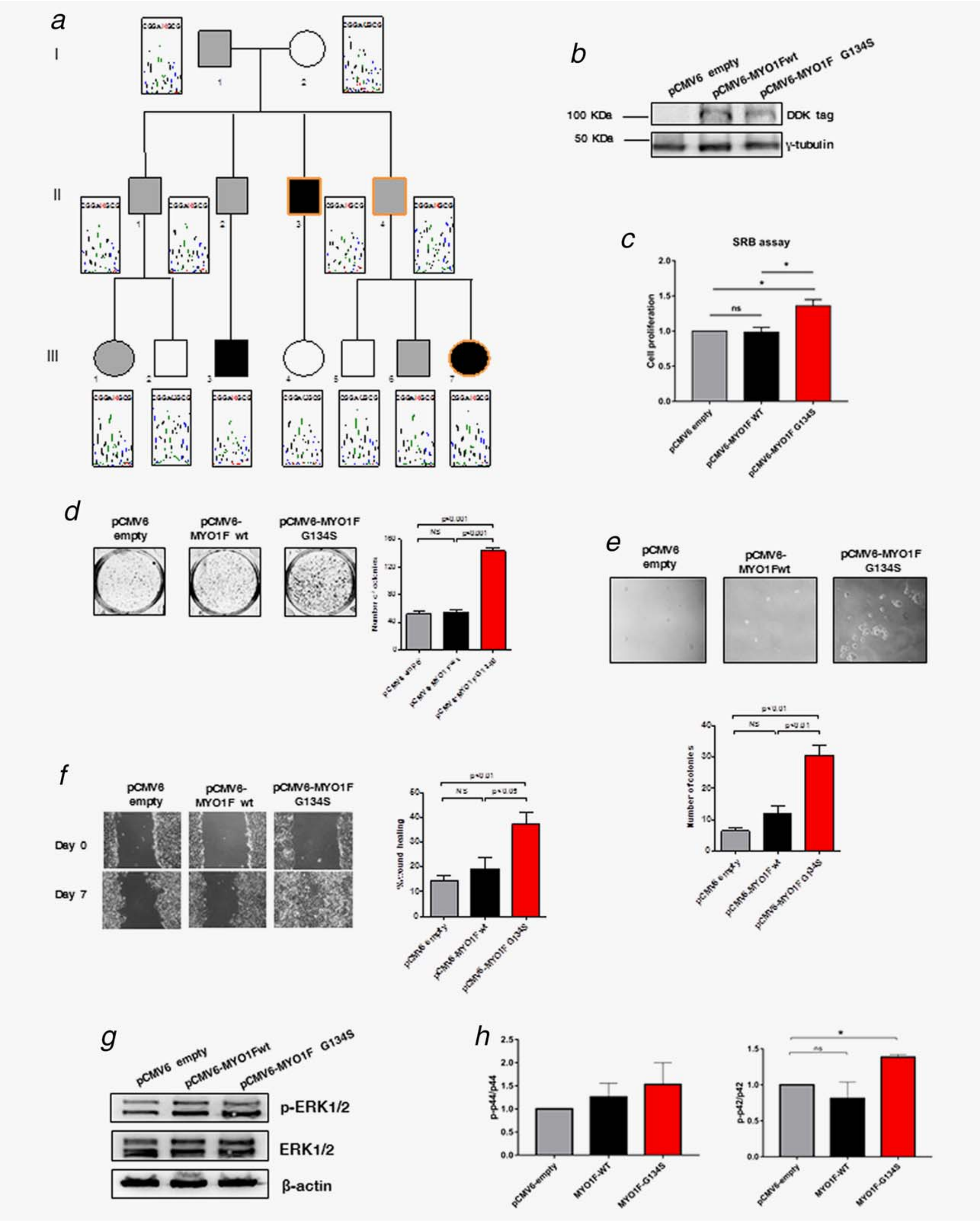


Figure 1. Study of MYO1F p.Gly134Ser variant. (*a*) Pedigree of the TCO family: electropherograms of the sequences of available family members, showing the co-segregation of the change (in red) with the oncocytic carcinoma (black)/adenoma (grey) phenotype. (*b–h*) Functional analysis of the MYO1F p.Gly134Ser variant. All experiments were repeated at least three times. Scale bars indicate standard error, stars indicate $p < 0.05$. (*b*) Western blot analysis showing the recombinant MYO1F protein in stably expressing FRTL-5 cells, using a specific anti-DDK antibody. Cell stably transfected with the empty vector are indicated as pCMV6-empty. Cells stably expressing the wt protein are indicated as pCMV6-MYOF wt, cells stably expressing the mutant protein are indicated as pCMV6-MYOF G134S. (*c*) SRB assay showed a significant increase in cell growth and proliferation for the pCMV6-MYOF G134S cells. (*d*) Plate colony formation potential using SRB assay showing an increased number of colonies formed by FRTL-5 expressing the mutant MYO1F protein, compared to cells expressing either the empty vector or the wt protein. (*e*) Growth in soft agar: FRTL-5 cells expressing the MYO1F mutant protein p.Gly134Ser significantly generated more colonies, compared to the empty and the cells expressing the wt protein. (*f*) Wound healing assay: FRTL-5 cells expressing the MYO1F mutant protein p.Gly134Ser filled the gap significantly faster compare to the other two cell lines. (*g, h*) Western blotting analysis of ERK1/2 phosphorylation in the three cell lines and densitometric quantification. [Color figure can be viewed at wileyonlinelibrary.com]

Generation of FRTL-5-stably transfected cell lines

A 7.5-µg of pCMV6 empty, pCMV6-MYO1F-wt and pCMV6-MYO1F-G134S plasmids were transfected using liposomes according to the manufacturer's instructions (Lipofectamine 2000, ThermoFisher Scientific, Grand Island, NY). Forty eight hours after transfection, selection was obtained by supplementing complete medium with 500 µg/ml G418 (ThermoFisher Scientific) for 2 weeks. Isolated clones were grown with 200 µg/ml G418.

Western blot

A detailed protocol is reported in the Supporting Information, including the list of primary antibodies used.

Sulforhodamine B (SRB) assay to investigate cell proliferation and plate colony formation

For cell growth and proliferation assays, 2.5×10^5 cells were seeded in duplicate and incubated 96 hours at 37°C. For plate colony formation, 2.5×10^4 cells were seeded in duplicate and incubated for 20 days at 37°C. Cells were washed in phosphate-buffered saline (PBS) and fixed with cold trichloroacetic acid (TCA) 50% at 4°C for 1 hr, then TCA was eliminated and cells were dried at room temperature for 16 hrs. Cells were stained with SRB 0.4% in 1% acetic acid for 30 min, washed with 1% acetic acid for four times. For the proliferation assay, cells were solubilized in TrisHCl 10 mM pH 10.5, mixing for 10 min on a rotatory plate. Absorbance was read at $\lambda = 564$ nm using a Beckman Coulter DU-530 spectrophotometer. For plate colony assay, cells were photographed with ChemiDoc XRS+ (Biorad). Area and number of colonies were quantified with the *ImageJ* software (National Institute of Health, Bethesda, MD) discarding colonies <1 pixel.

Soft agar colony assay

Stable cell lines were seeded in triplicate in a 0.48% top agar in growth medium over a layer of 0.8% agar in a six-well plate at a density of 1×10^5 cells/ml. Plates were incubated at 37°C and 5% CO_2 for 12 days, monitoring for colony formation. Medium was replaced every 5 days. After 12 days, colonies were photographed and analyzed with *ImageJ* software.

Wound healing assay

Stable cell lines were plated onto six-well plates and allowed to form a confluent monolayer. The cell monolayer was then scratched in a straight line to make a "scratch wound" with a 10-µl tip and the cell debris was removed by washing the cells with PBS. 5H5 (6H5 medium without Thyroid-Stimulating Hormone (TSH)) medium supplemented with 10% NCS and 200 µg/ml of neomycin was added with or without 1 mM *N*-acetyl-L-cysteine (NAC), and images of the closure of the scratch were captured at 0 and 7 days. Images were analyzed with the *TScratch* software.[14]

Iodide transport

Iodide uptake by FRTL-5 cells was measured by live cell imaging with the fluorescent halide biosensor yellow fluorescent protein (YFP)-H148Q/I15L, as described previously.[15,16]

Mitochondrial morphology and mass assessment *via* live cell imaging

Mitochondrial morphology was assessed by live imaging with or without 1 mM NAC, using a Nikon Eclipse 80 microscope (Nikon, Tokio, Japan) according to Ref. 17. Circularity measurements were collected using *ImageJ* standard tools.

Mitochondrial mass measurements

In 96-well culture plates, 1×10^4 FRTL-5-stable cell lines were seeded in quadruplicates. The next day, cells were loaded with 50 nM MitoTracker Green (MTG) for 30 min at 37°C in complete medium. After washing twice with medium, MTG fluorescence was recorded in a plate reader (EnSpire, PerkinElmer). MTG fluorescence values were expressed as relative fluorescence unit (RFU)/viable cells. Cell viability was assessed with a resazurin-based method.

Mitochondrial potential measurement *via* JC-1

The fluorescent probe JC-1 (5, 5′,6, 6′-tetrachloro-1, 1′, 3, 3′-tetraethylbenzimidazol carbocyanine iodide) was used to measure the mitochondrial membrane potential ($\Delta\phi$), as described in the Supporting Information.

Cellular respiration

Oxygen consumption in intact cells. Approximately 1.5×10^6 FRL5-stable cell lines were harvested at 70–80% confluence, washed in PBS, resuspended in complete medium and assayed for oxygen consumption at 30°C using a thermostatically controlled oxygraph chamber (Instech Mod. 203, Plymouth Meeting, PA). Basal respiration was measured in their respective media and compared with the one obtained after injection of oligomycin (1 µM) and Carbonyl cyanide-p-trifluoromethoxyphenylhydrazone (FCCP) (1–6 µM). Antimycin A (5 µM) was added at the end of experiments to completely block the mitochondrial respiration. Data were normalized to protein content determined by the Lowry method.

ATP/ADP synthesis ratio determination. Nucleotides were extracted and detected using a Kinetex C18 column (250 × 4.6 mm, 100 Å, 5 µm; Phenomenex, CA), with a two pump Agilent 1100 series system. Absorbance (260 nm) was monitored with a photodiode array detector (Agilent 1100 series system). Nucleotide peaks were identified by comparison and coelution with standards and quantification by peak area measurement compared with standard curves.[18]

ROS quantification

Intracellular ROS. FRTL-5-stable cell lines were seeded at 5×10^4 cells per well and incubated 16 hrs. Cells were treated with 10 µM 2',7'-dichlorodihydrofluorescein diacetate

(DCFDA) dissolved in medium for 1 hr. Then, cells were washed twice with PBS and incubated for 12 hrs in complete medium. Finally, cells were washed with PBS and the fluorescence emission from each well was measured (λexc = 485 nm; λem = 535 nm) with a multi-plate reader (Enspire, Perkin Elmer). Data are reported as the mean ± SD of at least three independent experiments.

Extracellular ROS. FRTL-5-stable cell lines were seeded at 5 × 10^4 cells per well and incubated 16 hrs. Cells were treated with 10 μM Amplex red (N-acetyl-3,7-dihydroxyphenoxazine), 0.025 U/ml horseradish peroxidase dissolved in complete medium for 16 hrs. The medium was collected and measured (λexc 530, λem 590) with a multiplate reader (Enspire, Perkin Elmer). Data were normalized for cell number using resazuring assay. Data are reported as the mean ± SD of at least three independent experiments.

In vivo study of mutant MYO1F

Zebrafish embryos and adults were maintained and mated according to standard procedures. Mutant and wt capped *MYO1F* mRNAs were synthesized with the SP6 mMESSAGE mMACHINE kit (Ambion, ThermoFisher Scientific) using as template the PCS2 + MYO1F-G134S and PCS2 + MYO1F-wt plasmids, respectively. wt zebrafish embryos were injected at one-cell stage with 150 pg of MYO1F-wt or MYO1F-G134S mRNA and then fixed at 48 hrs post fertilizations (hpf). To determine the cell proliferation patterns, a whole-mount immunostaining with the anti-phospho-Histone H3 (pH3) antibody (Millipore, Darmstadt, Germany) was performed. We counted the mitotic cells along the trunk of each fish (from the yolk extension to the tip of the tail) and calculated the average number of pH3-positive cells per embryo to compare the difference among groups. Statistical analysis was performed using Student's unpaired *t* test. Differences were considered significant for $p < 0.05$.

MYO1F mutation screening in FNMTC pedigrees

PCR primers for human *MYO1F* (NM_012335) were designed with Primer3 v4.0 (http://primer3.ut.ee) and are available on request. Genomic DNA extracted from peripheral blood was amplified according to standard PCR conditions, and PCR products were analyzed by direct sequencing, as reported in the Supporting Information.

P1 pAltermax MYO1F exon 7-minigene generation

PCR of *MYO1F* genomic region encompassing exons 7 and 8 was performed using primer forward 5′ GGGGAATTCAGAAGGGAAGAGAGGCAAGG-3′, inserting an *EcoRI* restriction site, and primer reverse 5′-CCCTCTAGAAACTCAGGAGGGTTTCTGGG-3′, inserting an *XbaI* restriction site from a heterozygous carrier. We generated the minigene reporter as described previously.[19] The PCR products were cloned into the digested P1 pAltermax and plasmids sequenced to identify the plasmids with the wt or the variant alleles. The splicing alteration analysis was performed as reported in Ref. 19 and in Supporting Information.

Structural modeling

Modeling of the protein structure was performed adopting a building obtained by comparison procedures based on MODELLER (https://salilab.org/modeller/). The template was MYO1C_HUMAN (PDB code: 4BYF_A), and the final structural superimposition indicated a 45% sequence identity among the computed and experimental structures. Given the coverage of the template to the target, modeling was possible in the protein region spanning amino acids 16–714. From structural superimposition, it was also possible to locate the ATP-binding domain.

Statistical analysis

Statistical analyses were performed using the one-way analysis of variance (ANOVA) with Tukey's multiple comparison test. All tests were completed using Prism (GraphPad, San Diego, CA). A $p < 0.05$ was considered statistically significant. All experiments were carried out at least in triplicates.

Results

Identification of a novel missense mutation in MYO1F conferring tumor-like properties to thyroid cells

WES was performed in three members of the original TCO family where the linkage locus was identified[12] (II-3, II-4 and III-7; Fig. 1*a*), in two individuals affected by thyroid carcinoma and one affected by thyroid adenoma, all with oncocytic features. All variants were queried with ANNOVAR and filtered based on Single Nucleotide Polymorphisms Database (dbSNP) database annotation. Potentially deleterious mutations were selected according to their functional class, and prioritization was given to those lying in the chr19p13.2 linkage region and present in all three cases. A unique novel heterozygous variant in the linkage interval shared by all three individuals fulfilled the criteria for pathogenicity: the mutation c.400G > A in *MYO1F* cDNA (NM_012335), leading to a missense p.Gly134Ser substitution, predicted to be damaging by PolyPhen-2 and Provean (Supporting Information Table S2), not present in the NHLBI Exome Sequencing Project (ESP), in the Exome Aggregation (ExAc) and Genome Aggregation (gnomAD) databases and absent from 1000 in-house control chromosomes. The variant cosegregated with the carcinoma/adenoma phenotype in the family and appeared to be a likely candidate for the NMTC gene residing at 19p13.2 (Fig. 1*a*). *MYO1F* consists of 28 exons encoding a 1098-amino-acid protein of the class of unconventional myosins.[20] The p.Gly134Ser amino acid change resides in a very well conserved position in the ATP-binding domain of the protein. Since thyroid tumor tissue from patients was not available for additional studies, we generated cell models stably expressing the wt or mutant MYO1F (mut) after transfection with the corresponding

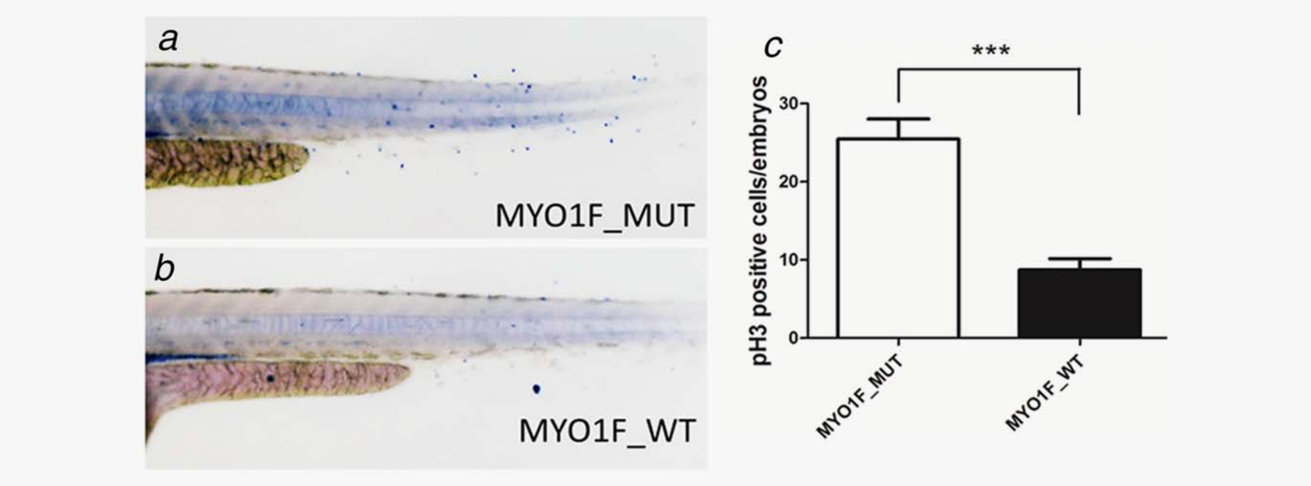


Figure 2. Proliferation analysis in zebrafish overexpressing either wt or mutant MYO1F p.Gly134Ser: (*a,b*) Immunostaining of phospho-histone H3 (pH3) performed in 48 hpf zebrafish larvae. An increase in cell proliferation can be observed in embryos injected with mutant MYO1F mRNA compared with embryos injected with the wt transcript of MYO1F. (*c*) Quantification of pH3-positive cells in injected embryos (48 hpf) was performed with manual counting of mitotic cells (blue nuclei) along the left side of the embryonic trunk, between the yolk extension and the tip of the tail. For each group, 22 embryos were analyzed (MYO1F_MUT: 25.45 ± 2.584; MYO1F_WT: 8.727 ± 1.445). ***, $p < 0.001$, Student's unpaired *t* test. [Color figure can be viewed at wileyonlinelibrary.com]

episomal plasmids, and a control cell line stably expressing the corresponding empty vector, pCMV6, *via* G418 selection. We used highly differentiated and functional FRTL-5 rat thyroid cell line[13] to reveal the effects of the MYO1F variant. The p.Gly134Ser mutation was inserted by site-directed mutagenesis in the construct encoding wt *MYO1F* in frame with the DDK tag. Western blotting with anti-DDK antibody in stably transfected cells showed that both wt and mut proteins were expressed in similar amounts (Fig. 1*b*). Stable cell lines expressing either the wt or the mut MYO1F protein were tested for their proliferative and tumorigenic potential in comparison with cells transfected with the empty vector. A significant increase in the proliferation was observed in mut cells, compared to cells expressing either the empty vector or the wt recombinant protein (one-way ANOVA with Tukey's multiple comparisons test $p = 0.0044$; pCMV6 empty *vs*. pCMV6-MYO1F G134S, $p = 0.0092$; pCMV6-MYO1F WT *vs*. pCMV6-MYO1F G134S $p = 0.0072$; pCMV6 empty *vs*. pCMV6-MYO1F WT $p = 0.9853$, Fig. 1*c*). A significant increase in the number of colonies in anchorage-dependent and independent growth was also observed in mut cells, compared to cells expressing either the empty vector or the wt recombinant protein (one-way ANOVA $p < 0.0001$, Fig. 1*d*). Anchorage-independent growth was monitored as colony formation in soft agar. Mutant MYO1F-expressing cells showed a significant increase in colony formation in soft agar, compared to cells stably transfected with the wt protein or the empty vector (ordinary one-way ANOVA $p = 0.0005$; Fig. 1*d*, lower panel).

The wound-healing assay, showed that mutant cells had a significantly greater invasive potential after 7 days in culture, compared to cells stably transfected with the empty vector or the wt protein, as quantified with *TScratch* software[14] (ordinary one-way ANOVA $p = 0.0024$; Fig. 1*e*). The wound healing assay was performed in a medium lacking TSH. Since proliferation of thyroid cells is totally TSH-dependent, we could discriminate between proliferation and invasiveness. Therefore, our data do indeed indicate that the mutant cells have a greater invasive potential.

To relate the observed changes in growth to the activation of specific cellular pathways, we investigated different kinases with key roles in cell proliferation and migration, including Akt and ERK1/2. We found a specific increase in the phosphorylation of ERK1/2 kinases in cells expressing the mutant protein, in particular for the p42 isoforms (Figs. 1*f* and [1]*g*; $p = 0.0042$, empty *vs*. pCMV6 MYO1F G134S). Taken together, these findings support a role for the MYO1F mutation in the modulation of tumorigenic potential *in vitro* (*i.e*., in the modulation of proliferation and invasivity).

Mutant MYO1F p.Gly134Ser stimulates proliferation in zebrafish embryos

To analyze the pro-proliferative function of MYO1F *in vivo*, we evaluated the effects of the human p.Gly134Ser MYO1F protein in zebrafish (*Danio rerio*) embryos. The zebrafish genome encodes a single *myo1f* orthologue (GenBank ref seq. NM_001256671.2; NP_001243600.1), with 85% similarity and 76% identity at amino acidic level to human MYO1F. Notably, the position corresponding to human Glycine 134 is conserved in the zebrafish Myo1f protein, indicating a putative functional role of this aminoacidic residue (Supporting Information Fig. S1).

To test whether the mutant MYO1F variant can induce cell proliferation *in vivo*, one-cell stage embryos were injected with either wt or p.Gly134Ser MYO1F mutated mRNA. At 48 hpf, the injected embryos were fixed and stained with antibodies against phospho-histone H3 (pH3), a widely used marker to reveal cell mitosis in zebrafish.[21–23] Embryos injected with the mutant mRNA showed a significant increase in the number of pH3-positive cells, compared to their siblings injected with the MYO1F wt allele (Figs. 2*a* and 2*b*). In particular, we observed an increased number of mitotic cells, especially in the caudal region ($p < 0.0001$, Fig. 2*c*) indicating that, when ubiquitously expressed, the MYO1F mutant protein can induce proliferation also in zebrafish embryos.

Iodide influx is not altered by the mutation MYO1F p.Gly134Ser

FRTL-5 cells are highly differentiated thyroid cells and a suitable model to measure iodide transport *in vitro*. We measured iodide uptake by live cell imaging after transient transfection with a vector encoding YFP-H148Q/I152L, a modified YFP whose fluorescence is quenched by I^- in a concentration-dependent manner.[15,16] We did not detect any differences in I^- uptake between the different cell lines (one-way ANOVA $p = 0.4816$; Supporting Information Figs. S2*a* and S2*b*).

Mutation MYO1F p.Gly134Ser alters the mitochondrial network

Since the oncocytic phenotype is characterized by mitochondrial hyperplasia in the tumors of affected individuals of the TCO family,[12] we analyzed the mitochondrial network of stably transfected FRTL-5 cells by live-cell microscopy using the MitoTracker Green probe. Mitochondria in the mutant cell lines appeared more fragmented compared to mitochondria in wt and empty cell lines (Fig. 3*a*), as shown by the significant increase in circularity value of mutant cells mitochondria when compared to wt and empty cell mitochondria (Fig. 3*b*).

The total mitochondrial mass was significantly greater in mutant cell lines, as determined by MitoTracker fluorescence quantification, normalized for cell viability using a resazurin-based assay (ordinary one-way ANOVA $p < 0.0001$; Fig. 3*c*). The increase in mitochondrial mass in the mutant cells was confirmed *via* Western blotting for voltage-dependent anion-selective channel (VDAC) (ordinary one-way ANOVA $p = 0.0136$, Fig. 3*d*).

Since an impaired mitochondrial network may alter mitochondrial function, we evaluated the levels of proteins and their phosphorylated forms (phospho-DRP1), involved in mitochondrial fission/fusion, that is, DRP1 and MFN1, but we did not detect any significant difference between the various cell lines (Figs. 3*e* and ^{3}*f*; Supporting Information Figs. S3*a* and S3*b*, respectively).

We measured the mitochondrial membrane potential and oxidative phosphorylation (OXPHOS) activity of the different cell lines. The mitochondrial membrane potential was measured with the probe JC-1,[24,25] and normalized for cell viability using a resazurin-based assay. No differences were found between empty vector-expressing cells, wt and mutant cells (one-way ANOVA $p = 0.0720$; Supporting Information Fig. S3*c*). The addition of oligomycin A did not alter the fluorescence ratio of JC-1, indicating that ATP hydrolysis by ATPase was not involved in maintaining the mitochondrial potential (Supporting Information Fig. S3*c*).

We measured the ATP/ADP ratio in the different cell lines, showing that the cells expressing mutant MYO1F exhibit a significant lower ratio in comparison to wt cells, due to the concomitant decrease in ATP and increase in ADP levels ($p = 0.0289$, one-way ANOVA, Fig. 3*g*). However, there were no differences in respiratory activity between the different cell lines under basal conditions (one-way ANOVA $p = 0.5014$, Supporting Information Fig. S3*d*) in the ratio of FCCP/oligomycin-treated cells (one-way ANOVA $p = 0.3900$; Supporting Information Fig. S3*e*. Extracellular lactate measurement also showed no changes between the different cell lines (ordinary one-way ANOVA $p = 0.4069$; Supporting Information Fig. S3*f*).

ROS are elevated in FRTL-5 cells expressing MYO1F p.Gly134Ser

Since differentiated thyroid cells produce a great amount of hydrogen peroxide (H_2O_2) necessary for thyroid hormone synthesis,[26] we investigated whether reactive oxygen species (ROS) production in transfected FRTL-5 cell lines was deranged by the MYO1F mutation.

Intracellular ROS levels, measured with the fluorescent probe DCF-DA, were significantly increased in the mutant cells (one-way ANOVA $p = 0.0015$, Fig. 4*a*). To understand whether this phenomenon was due to alterations/decreases of intracellular ROS detoxifying enzymes, we performed Western blotting analysis of catalase, mitochondrial manganese superoxide dismutase (SOD2) and peredoxin-3 (Prx3), using GAPDH as endogenous reference. The steady state levels of the analyzed proteins were not significantly different between all cell lines (Fig. 4*b*; Supporting Information Figs. S4*a*–S4*c*; one-way ANOVA $p = 0.1328$ for catalase, $p = 0.8592$ for SOD2, $p = 0.6837$ for Prx3).

Interestingly, treatment for 24 hrs with the antioxidant compound NAC partially recovered the defects in the mitochondrial network in cells expressing mutant MYO1F, confirming the role of ROS in mitochondrial fragmentation ($p < 0.0001$; Fig. 4*c*). In concordance, we observed a decrease in cell invasion between the FRTL-5 cell lines treated with NAC, compared to the untreated ones, as measured by the wound healing assay (Student's *t* test, untreated *vs.* treated $p = 0.0236$ pCMV6 empty, $p = 0.0338$ pCMV6-MYO1F wt, $p = 0.0488$ pCMV6-MYO1F-G134S; Fig. 4*d*). This effect was observed in all cell lines, not only for the mutant MYO1F cells.

To measure extracellular ROS, we used the fluorescent probe Amplex Red, which is unable to cross the plasma membrane. We observed a significantly higher amount of extracellular ROS in mutant cell lines, compared to the empty vector-transfected cells and the wt ones. Moreover, we detected, a significant decrease in extracellular ROS in the cells expressing MYO1F wt, when compared to the empty vector (one-way ANOVA $p = 0.0004$; pCMV6-empty *vs.*

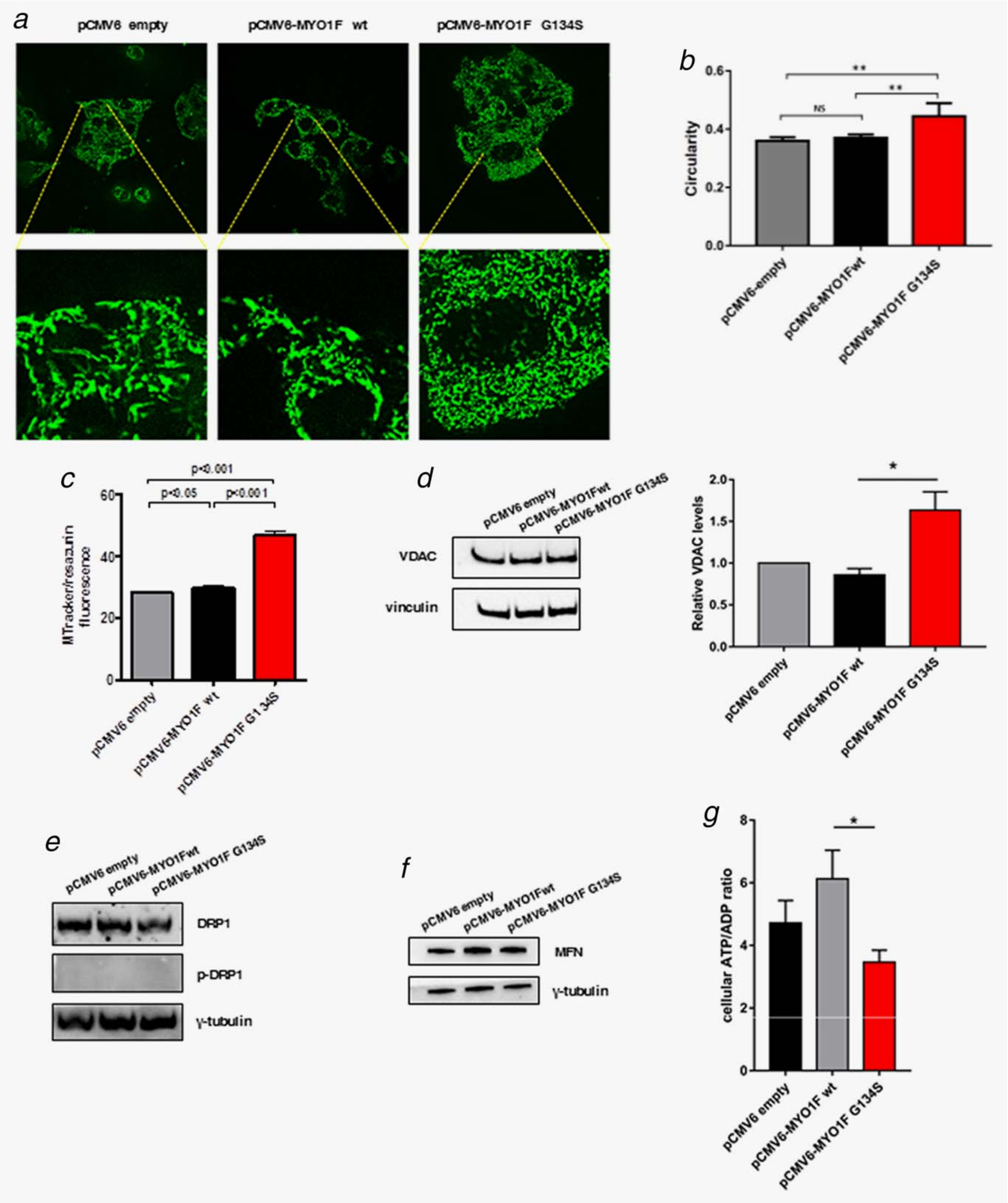


Figure 3. Mitochondrial defects in FRTL-5- MYO1F p.Gly134Ser cells. (*a*) Representative fluorescence images of pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134S stained with Mitotracker Green to evaluate mitochondrial network. The cells expressing the mutant protein show more circular (*b*) and more abundant (*c*) mitochondria and more fragmented mitochondrial network in comparison with wt and cells bearing empty vector. MitoTracker signal quantification was normalized on viable cell number assessed by resazurin-based assay. (*d*) Representative image of Western blotting analysis for VDAC in pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134S cells and relative quantification, compared to reference protein (vinculin). Scale bars indicate standard errors. Stars indicate significant *p* values. (*e*–*f*) Representative images of Western blotting analysis for DRP1-phospho-DRP1 (*e*) and MFN1 (*f*) in pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134S cells. (*g*) ATP/ADP ratio in cellular extracts from pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134S cells, showing a decreased ATP/ADP ratio in the mutant FTRL5 cells. Scale bars indicate standard errors. Stars indicate significant *p* values. [Color figure can be viewed at wileyonlinelibrary.com]

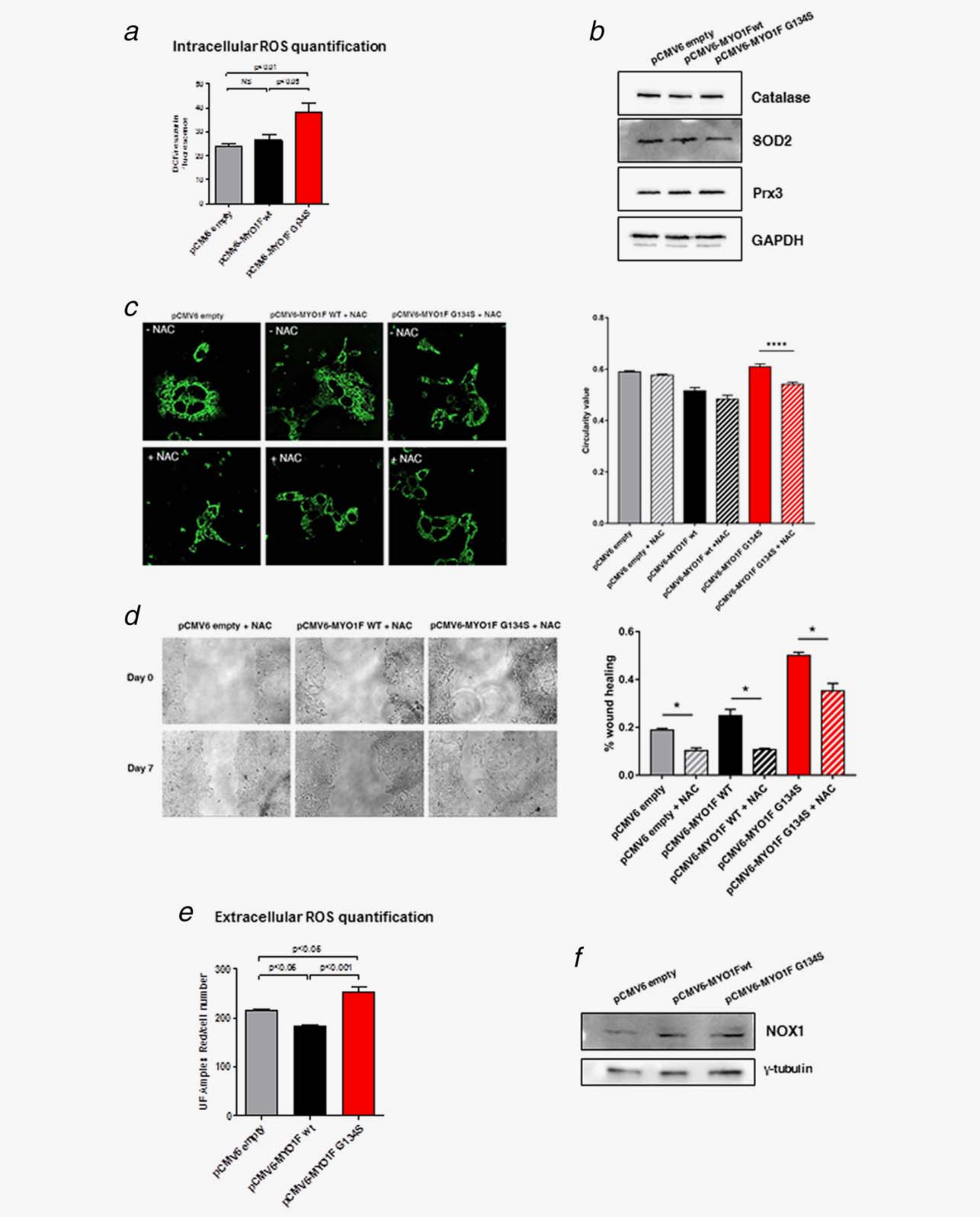


Figure 4. ROS production. (*a*) Intracellular ROS production measured by DCFDA fluorescent probe. Data show a significant ROS production increase in the FRTL-5 cells expressing MYO1F p.Gly134Ser in comparison to wt and cells bearing the empty pCMV6 vector. Data are expressed as arbitrary fluorescence units ± SD, normalized on viable cell number. (*b*) Representative Western blot analysis showing the expression of detoxifying enzymes (catalase, SOD2, and Prx3) in the three cell lines. GAPDH was used as endogenous loading control. (*c*) Representative fluorescence images of pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134-S cells stained for 24 hrs with 1 mM *N*-acetyl-L-cysteine (NAC) or vehicle. Live cells were stained with 40 nM Mitotracker Green to evaluate mitochondrial network. Circularity value analysis was performed using *ImageJ* software standard tool. Data indicate a significant recovery for the NAC-treated mutant cells *versus* the untreated mutant ones. (*d*) Wound healing assay in presence of NAC and relative quantification, showing a significant decrease in invasive potential in the cell lines. (*e*) Extracellular ROS production measured by Amplex red fluorescent probe. Data show that FRTL-5 cells expressing MYO1F p.Gly134Ser presented the highest levels of extracellular ROS, whereas the cells expressing the wt protein presented a reduced amount of extracellular ROS. Data are expressed as arbitrary fluorescence units ± SD normalized on viable cell number. Cell viability was assessed by resazurin-based method. (*f*) Representative image of Western blotting analysis for NOX1 in pCMV6-empty, pCMV6-MYOF wt, and pCMV6-MYOF G134S cells. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Rare coding variants identified in *MYO1F*-targeted mutation screening

Chr19 genomic position (hg19)	Amino acid change (NP_036467)	MAF in famNMTC ($N = 192$)	MAF in gnomAD (European only)
g.8616995 C>T rs184748543	p.Lys186[1]	0.0026	0.004224
g.8615552C>T rs201962739	p.Pro266	0.0026	0.001478
g.8615513C>G	p.Gly368[2]	0.0026	0
g.8610599G>T	p.Ile430	0.0026	0
g.8587411C>T rs201982814	p.Val1024Met[3]	0.0026	0.0007455[4]

[1] SNV not changing the corresponding amino acid, but with an altered ESE profile compared to wt cDNA, and removing SR-binding domains. The SNV co-segregated with the NMTC phenotype in the available members of the corresponding family.
[2] SNV not segregating with the NMTC phenotype in the corresponding families.
[3] Missense variant predicted to be "benign" (PolyPhen-2) and "tolerated" (SIFT).
[4] One homozygous individual present in European population.

pCMV6-MYO1F wt $p < 0.05$; pCMV6-MYO1F wt *vs*. pCMV6-MYO1F G134S $p < 0.001$; Fig. 4*e*). We evaluated NOX1 protein levels, but we did not detect any significant variation between the different cell lines (Fig. 4*f*; Supporting Information Fig. S4*d*; one-way ANOVA $p = 0.5900$).

Mutation screening of human MYO1F in FNMTC patients

To identify additional patients carrying predisposing germline mutations in *MYO1F*, we performed a mutation screening *via* Sanger sequencing of genomic DNA from peripheral blood of 192 independent FNMTC cases. These patients represented a heterogeneous group of cases affected by PTC/FTC, and the presence of oncocytic features was not always investigated. These data were available only for a small subgroup of patients (Supporting Information Table S1). We identified several rare/novel coding variants in *MYO1F* (Table 1), including a rare silent change in exon 7, present in both the affected individuals of the corresponding family, from whom DNA was available (Supporting Information Fig. S5*a*). This change potentially removed an exonic sequence enhancer (ESE) in exon 7, as predicted by the ESE Finder v3.0 program (Supporting Information Fig. S5*b*). The change, corresponding to the genomic coordinates chr19:g.8616995C > T (rs184748543), was present with a minor allele frequency (MAF) of 0.003168 in the general population and a MAF of 0.004224 in individuals of European ancestry-only (gnomAD; Table 1). The variant frequency was not significantly different between the NMTC cases and general population controls; moreover, one individual in the gnomAD database was homozygous for the variant allele, suggesting that it might have no severe functional consequences.

Nevertheless, to study whether it could hamper the inclusion of exon 7 in the final *MYO1F* transcript, since no fresh RNA was available from the affected patients carrying the rs184748543 variant allele, we generated a minigene plasmid carrying either the wt or mutant sequence, and transfected simian COS7 cells to study transcription (Figs. 5*a* and 5*b*). RT-PCR with minigene-specific synthetic primers and direct sequencing revealed that the wt exon was correctly spliced, whereas the mutant transcript lacked exon 7 (Fig. 5*c*). This altered transcript is predicted to produce a shorter MYO1F protein, with an in-frame deletion of 43 amino acids (G169-Q212) in the motor domain of MYO1F, that may alter the structure of the ATP-binding domain in the molecular motor of MYO1F (residues 110–117 and 162–166; Fig. 5*d*).

Discussion

The etiology of differentiated thyroid cancer is still poorly understood, but this type of cancer is influenced by both genetic and environmental factors. Large genome-wide case–control association studies have identified genetic variants conferring NMTC susceptibility in the general population.[27–29] A number of common single nucleotide polymorphisms (SNPs) have been reported to be associated with NMTC risk, but few studies have been conducted in high-risk NMTC families to examine the transmission of the risk allele to the affected members.[30]

In our study, we report the identification of MYO1F as the gene mutated at the TCO locus. We provide functional evidence that the MYO1F p.Gly134Ser mutation leads to an increased oncogenic potential *in vitro*, in terms of cell growth and invasion. FRTL-5 cells, a cell model resembling a functional thyrocyte,[13] stably transfected with the plasmid encoding mutant MYO1F p.Gly134Ser showed increased proliferation, generated significantly more colonies in soft agar and showed a significantly greater invasive potential compared to cells stably transfected with the empty vector or with wt *MYO1F*.

These *in vitro* data were supported by *in vivo* findings in zebrafish, showing that the mutant MYO1F p.Gly134Ser, when overexpressed, can induce proliferation in whole vertebrate embryos, supporting the idea that the novel missense change identified in exon 5 of *MYO1F* is the causative mutation at the TCO locus.

The TCO locus in the original pedigree was associated with an oncocytic phenotype, that is, enriched in mitochondria.[12] Previous work by our group uncovered a tight correlation between the co-occurrence of mitochondrial DNA (mtDNA) alterations in oncocytic thyroid cancer, and a marked dysfunction of OXPHOS complexes, in particular complex I.[31–33] Since thyroid follicular cells generate H_2O_2 by membrane-bound dual oxidases for the synthesis of thyroid hormones, these cells are at increased risk of oxidative stress

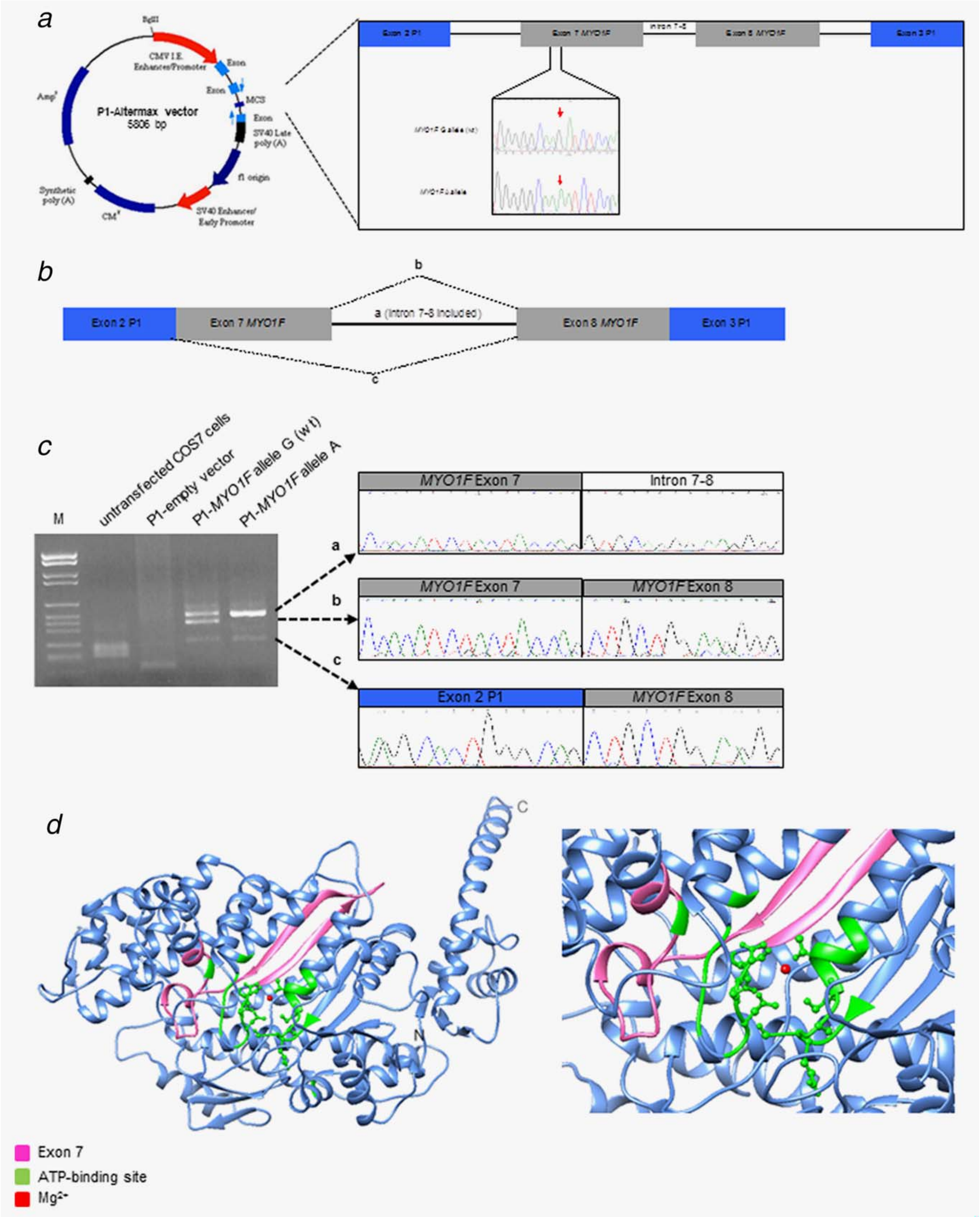


Figure 5. MYO1F rs184748543. (*a*) Map of the minigene plasmid, showing the genomic insert of the wt and mutant alleles (red arrows). (Blue arrows = position of the primers used for the specific RT-PCR). (*b*) RT-PCR of COS7-transfected with the *MYO1F* allele-specific minigenes. *Upper panel*: predicted final transcripts generated by the correct splicing of mini-gene-specific exons (blue) and *MYO1F*-specific exons (grey). *Lower panel*: 2% agarose gel image (left) of the RT-PCR products, showing the different sizes of the transcripts and corresponding electropherograms (right): the wt *MYO1F* allele promoted the inclusion of the exon 7 in the final transcript, whereas the mutant allele induced an exon skipping in the final transcript, as predicted by the removal of the ESE in the exon 7. (*d*) Structure prediction of the MYO1F molecular motor region, with the ATP-binding region highlighted in green. Pink indicates the residues corresponding to exon 7 and red is the ion of magnesium. [Color figure can be viewed at wileyonlinelibrary.com]

and ROS-mediated DNA damage. Indeed, an imbalance between pro- and anti-oxidative factors has been suggested as an important mechanism in thyroid tumorigenesis.[34,35] Oxidative stress generated by mitochondrial dysfunction can also promote migration and stimulate MAPK-mediated cell death. We therefore sought to evaluate: (*i*) the functionality of the mitochondrial respiratory chain as a whole; (*ii*) the response to oxidative stress of FRTL-5 cells stably expressing the wt or mutant recombinant MYO1F protein, compared to cells expressing the empty-vector. We found that the mitochondrial membrane potential and OXPHOS activities were similar in all cell lines, suggesting that mitochondria were still functional. However, analysis of the mitochondrial network by live-cell visualization revealed that in the mutant cell lines, mitochondria appeared as separated rod-shaped organelles.

The mitochondrial features of mutant MYO1F cells were therefore reminiscent of the oncocytic features described previously in the tumor tissues of the patients carrying the p.Gly134Ser change.[12]

In our experimental setting, we found that cells with the MYO1F p.Gly134Ser mutation, in addition to having an altered mitochondrial network and an increased mitochondrial mass, produced significantly more intracellular and extracellular ROS.

It has been reported that the establishment and maintenance of a transformed state is related to the presence of extracellular ROS, in particular, superoxide anion generated by a specific membrane-associated NADPH-oxidase, NOX1.[35] In fact, oncogenic activation of proliferative/mitogenic pathways has been associated with increased ROS production due to activation of the membrane-bound NADPH oxidases.[36] We did not detect differences among NOX1 protein levels, but this finding did not exclude an increased activation of this enzyme in the cells carrying the mutant MYO1F protein. Extensive analysis of tumor cell lines derived from different tissues, including thyroid carcinomas, has shown that they were all characterized by extracellular ROS generation, not found in cells derived from normal tissues.[34] This is paralleled by our findings, since extracellular ROS production was increased only in FRTL-5 cells expressing the mutant MYO1F p.Gly134Ser protein, suggesting that the mutation is sufficient to generate a transformed phenotype.

Interestingly, when the cells with the MYO1F p.Gly134Ser mutation were treated with an antioxidant (NAC), we observed a partial but significant rescue of the mitochondrial fragmentation, confirming the role of ROS in this phenomenon.[37] In agreement with this result, treatment with NAC also decreased the invasiveness of all cell lines including mutant MYO1F cells, as indicated by the wound-healing assay. These pilot data on phenotype rescue suggest that the treatment with antioxidants may be effective for this type of tumors.

Since the "mitochondria-rich" phenotype may be underreported by histologic analysis,[32] we screened additional FNMTC patients to identify other *MYO1F* germline variants that could predispose to thyroid tumor development. However, the available samples represented a heterogeneous group of familial cases affected by NMTC, and the high genetic heterogeneity of thyroid cancer might have hampered the discovery of a number of additional predisposing variants in *MYO1F*. We identified a rare variant in two affected sibs in exon 7, which promoted the skipping of the exon from mature mRNA *in vitro*. No data regarding the presence of an oncocytic phenotype were available for the two affected sibs. In addition, the unavailability of fresh RNA from tissues of these patients prevented us from confirming that this exon skipping event actually occurs *in vivo*. Moreover, the allele frequency of the exon 7 variant allele in our FNMTC cases was not significantly different from the one present in the control individuals in public databases; therefore, its contribution to NMTC predisposition remains elusive. Our results regarding MYO1F mutation screening in FNMTC cases stress once again the high genetic heterogeneity underlying familial thyroid cancer. Nevertheless, our study shows that a defective MYO1F protein promotes the development of an oncocytic phenotype, that is, mitochondrial proliferation, indicating that this cellular characteristic can develop not only from mitochondrial DNA defects[31–33] but also from nuclear defects in specific genes, that is, *MYO1F*. Mitochondrial dysfunction and stress has been widely related to cancer, in particular, in thyroid cancer predisposition.[31,32] More broadly, an altered mitochondrial function is a hallmark of many cancers, although the nature of functional modification depends on the type of cancer.[33]

It is interesting to note that F-actin is one of the few known interactors of MYO1F[20] and has been recently implicated in mitochondrial fission control.[38] Blockade of F-actin polymerization/depolymerization altered the mitochondrial network.[38] Similarly to what has been observed in other autosomal dominant disorders due to mutations in myosin genes, such as MYH9,[39,40] we can hypothesize that the modified conformation of MYO1F may block actin filament recycling; therefore, concurrently altering the mitochondrial network organization.

Recent data have shown the contribution of mitochondrial dynamics toward tumor initiation and progression, although the exact mechanism is not known. Excessive fission and reduced fusion is a feature of many tumors.[41–43] For example, in human pancreatic cancer, expression of oncogenic Ras/activation of MAPK pathway induces ERK2-mediated Drp1 phosphorylation leading to increased mitochondrial fragmentation and the inhibition of this phosphorylation in xenografts is sufficient to block tumor growth.[44] Interestingly, recent data indicated that ERK2 also phosphorylated MFN1 to control mitochondrial morphology and apoptosis.[45] We did not find difference in Drp1 levels and phosphorylation in our cell models, and it will be of interest to evaluate also this pathway in the framework of the observed altered mitochondrial network present in the mutant MYO1F cells.

It is becoming increasingly clear that mitochondrial fission and fusion play a critical role in quality control and mitochondrial damage/repair in cancer. Therefore, our data showing a fragmented mitochondrial network due to MYO1F p.Gly134Ser mutation highlight a potential novel pathway that may be deranged in thyroid cancer, that is, an altered myosin/F-actin regulated interaction.[20]

To date, no other mutations have been reported in myosin-encoding genes in thyroid cancer; however, it is interesting that MYH9, a non-muscle myosin involved in sensorineural deafness and thrombocytopenia,[39,40] has recently been found to regulate the ncRNA genes *PTCSC2* and *FOXE1* at the 9q22 thyroid cancer susceptibility locus.[46] In the TCGA database, somatic mutations in *MYO1F* are reported in 352 cases from various cancer types (Supporting Information Fig. S6*a*). The mutation identified at the TCO locus in MYO1F was not reported. In the COSMIC database several mutations are present in MYO1F in different types of cancer (Supporting Information Fig. S6*b*), but only a somatic variant is reported in thyroid carcinoma (COSM4132813). However, *MYO1F* overexpression was reported in 24 of 513 (4.68%) cases (Supporting Information Fig. S6*c*). These and our data suggest that MYO1F dysregulation may predispose to cancer in a subgroup of cases. Indeed, the oncocytic phenotype, observed in the family with the MYO1F p.Gly134Ser mutation, represents a specific, although rare, group of thyroid neoplasms, in which *MYO1F* mutation screening may be more relevant than in other FNMTC cases. The identification of the molecular cause(s) of specific thyroid cancer subtypes will help tailor patients' treatment for a more personalized therapy.

URL

Catalogue of Somatic Mutations in Cancer (COSMIC): http://cancer.sanger.ac.uk/ ESEfinder 3.0: rulai.cshl.edu/tools/ESE/ Exome Aggregation database (ExAc): http://http://exac.broadinstitute.org/Genome Aggregation database (gnomAD): http://gnomad.broadinstitute.org/ MODELLER: https://salilab.org/modeller/ PolyPhen-2: genetics.bwh.harvard.edu/pph2 PROVEAN (including SIFT): provean.jcvi.org/ Primer 3: primer3.ut.ee The Cancer Genome Atlas (TCGA): https://tcga-data.nci.nih.gov/

Acknowledgements

We thank all patients and families that participated in the study, Mr. F. Baccianti, Dr. F. Bianco and Dr. G. Zuccheri for technical support in cell imaging, Dr. I. Kurelac for helpful suggestions and discussion and Dr. L. Pivotti and Dr. M. Milanetto for their help in managing Padua Zebrafish Facility. The work was supported by Associazione Italiana per la Ricerca sul Cancro (AIRC) grant IG2015–17069 (to M.S.), Italian Ministry of Health grants GR-2012 "DIANE" (to E.B.), RF-2011–02350857 (to K.J.R.), and EU FP7 Marie Curie project MEET-317433 (to A.M.P.).

References

1. Malchoff CD, Malchoff DM. Familial nonmedullary thyroid carcinoma. *Cancer Control* 2006;13: 106–10.
2. Guilmette J, Nosè V. Hereditary and familial thyroid tumours. *Histopathology* 2018;72:70–81.
3. Navas-Carrillo D, Ríos A, Rodríguez JM, et al. Familial nonmedullary thyroid cancer: screening, clinical, molecular andgenetic findings. *Biochim Biophys Acta* 2014;1846:468–76.
4. Bonora E, Tallini G, Romeo G. Genetic predisposition to familial nonmedullary thyroid cancer: an update of molecular findings and state-of-the-art studies. *J Oncol* 2010;2010:1
5. Rio Frio T, Bahubeshi A, Kanellopoulou C, et al. DICER1 mutations in familial multinodular goiter with and without ovarian Sertoli-Leydig cell tumors. *JAMA* 2011;305:68–77.
6. Dettmer M, Perren A, Moch H, et al. Comprehensive microRNA expression profiling identifies novel markers in follicular variant of papillary thyroid carcinoma. *Thyroid* 2013;23:1383–9.
7. Tomsic J, He H, Akagi K, et al. A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Sci Rep* 2015;5:10566
8. Swierniak M, Wojcicka A, Czetwertynska M, et al. In-depth characterization of the microRNA transcriptome in normal thyroid and papillary thyroid carcinoma. *J Clin Endocrinol Metab* 2013; 98:E1401–9.
9. He H, Li W, Liyanarachchi S, et al. Genetic predisposition to papillary thyroid carcinoma: involvement of FOXE1, TSHR, and a novel lincRNA gene, PTCSC2. *J Clin Endocrinol Metab* 2015;100:E164–72.
10. Liu D, Yang C, Bojdani E, et al. Identification of RASAL1 as a major tumor suppressor gene in thyroid cancer. *J Natl Cancer Inst* 2013;105: 1617–27.
11. He H, Bronisz A, Liyanarachchi S, et al. SRGAP1 is a candidate gene for papillary thyroid carcinoma susceptibility. *J Clin Endocrinol Metab* 2013;98:E973–80.
12. Canzian F, Amati P, Harach HR, et al. A gene predisposing to familial thyroid tumors with cell oxyphilia maps to chromosome 19p13.2. *Am J Hum Genet* 1998;63:1743–8.
13. Meli A, Perrella G, Curcio F, et al. In vitro cultured cells as probes for space radiation effects on biological systems. *Mutat Res* 1999;430: 229–34.
14. Gebäck T, Schulz MM, Koumoutsakos, et al. TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays. *Biotechniques* 2009;46:265–74.
15. Rhoden KJ, Cianchetta S, Stivani V, et al. Cell-based imaging of sodium iodide symporter activity with the yellow fluorescent protein variant YFP-H148Q/I152L. *Am J Physiol Cell Physiol* 2007;292:C814–23.
16. Rhoden KJ, Cianchetta S, Duchi S, et al. Fluorescence quantitation of thyrocyte iodide accumulation with the yellow fluorescent protein variant YFP-H148Q/I152L. *Anal Biochem* 2008;373: 239–46.
17. Chazotte B. Labeling mitochondria with MitoTracker dyes. *Cold Spring Harb Protoc* 2011;2011: pdb.prot5648–92.
18. Bergamini C, Moruzzi N, Volta F, et al. Role of mitochondrial complex I and protective effect of CoQ10 supplementation in propofol induced cytotoxicity. *J Bioenerg Biomembr* 2016;48: 413–23.
19. Bonora E, Evangelisti C, Bonichon F, et al. Novel germline variants identified in the inner mitochondrial membrane transporter TIMM44 and their role in predisposition to oncocytic thyroid carcinomas. *Br J Cancer* 2006;95:1529–36.
20. Kim SV, Mehal WZ, Dong X, et al. Modulation of cell adhesion and motility in the immune system by Myo1f. *Science* 2006;314:136–9.
21. Verduzco D, Amatruda JF. Analysis of cell proliferation, senescence, and cell death in zebrafish embryos. *Methods Cell Biol* 2011;101:19–38.
22. Mendieta-Serrano MA, Schnabel D, Lomelí H, et al. Cell proliferation patterns in early zebrafish development. *Anat Rec (Hoboken)* 2013;296: 759–73.
23. Luo N, Li H, Xiang B, et al. Syndecan-4 modulates the proliferation of neural cells and the formation of CaP axons during zebrafish embryonic neurogenesis. *Sci Rep* 2016;6:25300
24. Smiley ST, Reers M, Mottola-Hartshorn C, et al. Intracellular heterogeneity in mitochondrial membrane potentials revealed by a J-aggregate-forming lipophilic cation JC-1. *Proc Natl Acad Sci USA* 1991;88:3671–5.
25. Chazotte B. Labeling mitochondria with JC-1. *Cold Spring Harb Protoc* 2011;2011: pdb.prot065490
26. Yoshihara A, Hara T, Kawashima A, et al. Regulation of dual oxidase expression and H2O2 production by thyroglobulin. *Thyroid* 2012;22:1054–62.
27. Cavaco BM, Batista PF, Sobrinho LG, et al. Mapping a new familial thyroid epithelial neoplasia susceptibility locus to chromosome 8p23.1-p22 by high-density single-nucleotide polymorphism genome-wide linkage analysis. *J Clin Endocrinol Metab* 2008;93:4426–30.
28. Jazdzewski K, Murray EL, Franssila K, et al. Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. *Proc Natl Acad Sci USA* 2008;105:7269–74.
29. Gudmundsson J, Sulem P, Gudbjartsson DF, et al. Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat Genet* 2009;41:460–4.

30. Bonora E, Rizzato C, Diquigiovanni C, et al. The FOXE1 locus is a major genetic determinant for familial nonmedullary thyroid carcinoma. *Int J Cancer* 2014;134:2098–107.
31. Bonora E, Porcelli AM, Gasparre G, et al. Defective oxidative phosphorylation in thyroid oncocytic carcinoma is associated with pathogenic mitochondrial DNA mutations affecting complexes I and III. *Cancer Res* 2006;66:6087–96.
32. Gasparre G, Porcelli AM, Bonora E, et al. Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. *Proc Natl Acad Sci USA* 2007; 104:9001–6.
33. Porcelli AM, Ghelli A, Ceccarelli C, et al. The genetic and metabolic signature of oncocytic transformation implicates HIF1alpha destabilization. *Hum Mol Genet* 2010;19:1019–32.
34. Bauer G. Targeting extracellular ROS signaling of tumor cells. *Anticancer Res* 2014;34:1467–82.
35. Laurent E, McCoy JW, Macina RA, et al. Nox1 is over-expressed in human colon cancers and correlates with activating mutations in K-Ras. *Int J Cancer* 2008;123:100–7.
36. Carvalho DP, Dupuy C. Role of the NADPH oxidases DUOX and NOX4 in thyroid oxidative stress. *Eur Thyroid J* 2013;2:160–7.
37. Iqbal S, Hood DA. Oxidative stress-induced mitochondrial fragmentation and movement in skeletal muscle myoblasts. *Am J Physiol Cell Physiol* 2014;306:C1176–83.
38. Li S, Xu S, Roelofs BA, et al. Transient assembly of F-actin on the outer mitochondrial membrane contributes to mitochondrial fission. *J Cell Biol* 2015;208:109–23.
39. Seri M, Cusano R, Gangarossa S, et al. Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. The May-Heggllin/Fechtner Syndrome Consortium. *Nat Genet* 2000;26:103–5.
40. Seri M, Pecci A, Di Bari F, et al. MYH9-related disease: May-Hegglin anomaly, Sebastian syndrome, Fechtner syndrome, and Epstein syndrome are not distinct entities but represent a variable expression of a single illness. *Medicine (Baltimore)* 2003;82:203–15.
41. Inoue-Yamauchi A, Oda H. Depletion of mitochondrial fission factor DRP1 causes increased apoptosis in human colon cancer cells. *Biochem Biophys Res Commun* 2012;421:81–5.
42. Zhao J, Zhang J, Yu M, et al. Mitochondrial dynamics regulates migration and invasion of breast cancer cells. *Oncogene* 2013;32:4814–24.
43. Ferreira-da-Silva A, Valacca C, Rios E, et al. Mitochondrial dynamics protein Drp1 is overexpressed in oncocytic thyroid tumors and regulates cancer cell migration. *PLoS One* 2015;10: e0122308
44. Kashatus JA, Nascimento A, Myers LJ, et al. Erk2 phosphorylation of Drp1 promotes mitochondrial fission and MAPK driven tumor growth. *Mol Cell* 2015;57:537–51.
45. Pyakurel A, Savoia C, Hess D, et al. Extracellular regulated kinase phosphorylates mitofusin 1 to control mitochondrial morphology and apoptosis. *Mol Cell* 2015;58:244–54.
46. Wang Y, He H, Li W, et al. MYH9 binds to lncRNA gene PTCSC2 and regulates FOXE1 in the 9q22 thyroid cancer risk locus. *Proc Natl Acad Sci USA* 2017;114: 474–9.

10 Discussion

The goal of this thesis is to make a step forward in the direction of unveiling the biological complexity of the genotype-phenotype relation. We believe that studying the molecular mechanisms and the biological functions at the basis of phenotypes manifestation and disease insurgence is useful to understand the processes characterising the phenotypes/diseases and to retrieve new possible related genes and variants to be tested in further researches. To analyse the shared features of phenotype and disease related genes is fundamental to integrate knowledge that means to integrate the various omics collecting different level of annotation.

Our approach focused on this problem via large-scale studies whose significance was proved with specific study cases. In order to do so, we built resources like eDGAR (Babbi G et al, 2017; chapter 2) with the aim of merging features describing gene-disease associations. The idea beyond eDGAR is that curating and collecting the information on gene-disease associations is crucial to help researchers and physicians studying complex diseases. These concepts are at the basis of many other integrative datasets, such as DisGeNet (Piñero J et al, 2017) and MalaCards (Rappaport N et al, 2017) that 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, while DisGeNet integrates data of disease-associated genes and their variants. Although these resources are already endowed with much information, we strongly believed that eDGAR may contribute in the field thanks to the huge variety of resources used for gene annotation, including information on regulatory interactions, co-localization in neighbouring loci, protein-protein interactions and co-occurrence in protein complexes. This variety of resources allows a deep investigation of the features shared among genes (or proteins) co-involved in the same disease, letting emerge biological processes and pathways implicated in the disease. In eDGAR dataset, 621 diseases are associated with multiple genes (23% of the dataset). Investigating the features of polygenic diseases is crucial to better understand the nature of these maladies; in fact, studying the shared features (e.g.: biological pathways, molecular functions, stable interactions in complexes) we may find important processes characterizing the disease insurgence. Our results confirmed this hypothesis: we were able to endow the greatest majority of the polygenic diseases in the dataset with functional relations. For example, considering Gene Ontology (Gene Ontology Consortium, 2017) terms for biological process, almost all the polygenic diseases have at least a pair of genes in the same biological pathway, while considering other resources like REACTOME (Fabregat A et al, 2018) and KEGG

(Kanehisa M et al, 2017) this percentage remains in any case above the 50% (chapter 2). Being part of the same biological pathways could be index of protein-protein interactions. We confirmed this idea analysing protein-protein interactions in stable complexes, physical interactions and indirect interactions through a intermediator. Our results showed that 14% of the protein in polygenic diseases are part of the same protein complex, and the percentage increases (40-45%) considering less stable interactions derived from STRING (Szklarczyk D et al, 2015) and in BIOGRID (Chatr-Aryamontri A et al, 2015). Considering indirect protein-protein interactions, we retrieved 25% more of polygenic diseases having related proteins in interactions.

Therefore, diseases related to multiple genes shared biological processes and moreover they are characterized by protein-protein interactions of the protein products of the associated genes. We studied these protein-protein interactions having in mind the concepts that proteins in the same complex or in interaction should be co-expressed. Accordingly, we decided to analyse the regulations of the genes in polygenic maladies, taking into consideration transcription factors not directly linked to the disease that regulate the expression of genes associated to the same malady. Consequently, we analyse this property and we found out that half of the diseases of the dataset (44%) are associated with at least a couple of co-regulated genes. Concluding, we want to reinforce the notion that genes associated to the same disease shared functional features and thus it is important to compare the annotation of the genes related to a malady to discover new possible biological pathways, protein complexes or transcription factors that may be analysed in further research to understand the molecular mechanisms at the basis of disease insurgence and progression.

The novelty of eDGAR is that it allows a comprehensive analysis of the shared features of genes related to the same disease, and we believe that this resource may give a contribute in the direction of precision medicine to understand the molecular mechanisms that connect the different genes associated to the same disease.

eDGAR has already been successfully used as a resource to retrieve well annotated genes associated with amyotrophic lateral sclerosis, with the aim of studying the expression of these genes in relation with developmental neurogenesis pathways (Swindell WR et al, 2018).

In the same perspective of eDGAR, we built PhenPath (Babbi et al, under revision; chapter 4) to study specifically the biological processes underneath the appearance of several phenotypes. Many other resources for studying the molecular mechanisms leading to different phenotypes have been recently developed, like Phenopolis (Pontikos N et al, 2017), an open platform for harmonization and analysis of sequencing and phenotype data, offering a prioritized list of genes per phenotype, based on known association and gene enrichment analysis.

Other resources provide associations between diseases and phenotypes, including the Human Phenotype Ontology (HPO; Köhler S et al, 2017) and the OMIM Clinical synopses (Amberger JS et al, 2015). In particular, the Phenomizer tool (Köhler S et al, 2009), provided by the Human Phenotype Ontology consortium, analyses lists of phenotypes/symptoms with the aim of assisting the clinical workflow and suggesting diagnoses.

While many resources focus on the relationship among phenotypes, diseases and genes, little is known about the relevance of molecular functions and functional processes underlying the co-occurrence of phenotypes. Our hypothesis is that phenotypes co-occurrence may derive from an alteration of a limited number of biological processes underneath the phenotypes in exam, and thus retrieving the shared biological pathways and functions is very useful to study the phenotypes insurgence, especially when the number of genes associated to phenotypes is restricted. Accordingly with this idea, in PhenPath we focused on supplement gene-disease-phenotype associations with functional annotations associated with a given set of phenotypes. PhenPath offers a new approach for investigating the molecular mechanisms leading to the correlated manifestation of different phenotypes. PhenPath may be used to explore the possible connections among different phenotypes co-occurring in a patient, offering new clues on the biological mechanisms that may explain its clinical conditions.

Although the paper describing PhenPath is still under revision, we already have proofed the efficacy of this resource with study cases. In particular when we analysed Rett syndrome associated phenotypes with PhenPathTOOL, via comparison of the biological pathways shared by the phenotypes in input, we recovered genes that have been only recently associated with Rett syndrome and not previously reported in the gene-disease datasets used to build PhenPath. These findings illustrate the efficacy of PhenPathTOOL in linking a set of phenotypes to genes and functional annotations, retrieving new genes involved the disease insurgence to be studied in further experiments.

Beside the study of gene-disease and gene-phenotype associations with a large-scale approach, we also analysed in deep the relations among genetic mutations, protein variants and their associations to diseases and phenotypes. In particular, with INPS-3D (Martelli PL et al, 2016, chapter 5), we participated into two editions of the Critical Assessment of Genome Interpretation (CAGI), an international experiment with the aim of testing computational methods for the predictions of phenotypic effects of genetic mutations or protein variants. Over the 16 challenges in which we competed in the last three years, we obtained good results presenting our work to an international audience of experts in the field, and we also collaborate in 2 publications in the CAGI4 special issue (Daneshjou R et al, 2017, chapter 7; Xu Q et al, 2017, chapter 8). Other 3 publications have already been submitted to the Human Mutation CAGI5 special issue. We recently compared our results in CAGI5 edition with the ones of other predictors summarizing the lessons learnt in a paper under review (Savojardo et al, 2019, Human Mutation, submitted).

Here, we highlight the evaluation of the performance of our INPS-3D predictor, which has been used to generate predictions submitted to CAGI5 for the challenge of Frataxin and TPMT-PTEN. In particular for Frataxin challenge, evaluation was carried out using the same procedure applied during the official assessment of the challenge (performed by Emidio Capriotti, University of Bologna, Italy). According to the official CAGI5 assessment, INPS-3D is among the top-performing methods participating to this challenge.

We can say that the good performance achieved by INPS-3D in this experiment reflects the fact that the challenge required to predict the ΔΔG value upon variation, which is exactly the same experimental evidence used to train our predictor.

In the TPMT-PTEN challenge our approaches show performances differentiated between the two proteins, with correlations that are lower for TPTM and higher for PTEN. This behaviour is in line to what observed for all participants to the challenge, as pointed-out during the official assessment (performed by Yana Bromberg, Rutgers University, NJ, USA). Overall, our submissions are in the top 50% among challenge participants as highlighted in the assessment.

Comparing results of Frataxin and TPMT-PTEN challenges, it is worth noting that, using essentially the same prediction approach, we achieved very different levels of performance. It is clear that, as soon as the prediction task deviates from the original scope of the predictor, performances progressively decrease.

Thanks to the expertise acquired in the field, we also collaborate with the Sant'Orsola Genetic Medical Unit of the Department of Medicine and Surgery of the University of Bologna, building a series of models of protein structure of myosin 1F and its variants related to the thyroid cancer (Familial Non-Medullary Thyroid Carcinoma, FNMTC) (Diquigiovanni C et al, 2018, chapter 9). Here we want to highlight that our approach merged basic and applied research, keeping focused on real problems like the annotation of specific protein variants in relation with disease, with direct application in medicine.

11 Conclusions

The annotation of genes, proteins and their variants is still an issue in computational biology. In a large-scale perspective, a great effort is continuously made by the scientific community with the aim of creating resources for the annotation, maintaining and updating current databases and curating the stored information. The problem of standardization of the nomenclature that we use to define genes as well as proteins and variants and their features is not solved; we need to map our data to many different classifications to be user friendly, waiting for a definitive homogenization of international standards. In this direction, enlarging the network of collaborations is fundamental, and being part of the ELIXIR community (see paragraph 1.4) is important especially for what regards resources integration and interoperability.

In this thesis, we proposed webservers and tools available online to help researchers in directing their experiment and speed up the annotation procedures. With eDGAR (Babbi G et al, 2017) and its new version eDGAR+ we provide a database of very well annotated gene-disease associations, with the possibility of comparing and analysing in deep the relations among genes associated with the same disease. To understand the biological process that leads to the appearance of different phenotypes, we provide PhenPath (Babbi G et al, submitted in 2018), comprising a database of precomputed analysis and a tool for the online comparison of set of phenotypes, retrieving shared genes, diseases and shared molecular mechanisms.

We proposed a great variety of approaches for the prediction of the phenotypic effects of genetic variants, participating in two editions of the CAGI experiment. We test our predictors (e.g. INPS-3D, Savojardo et al, 2016) and compared our outcomes with the one obtained by other researchers in the field. We report some of the best results to describe the most effective methods and we already published two scientific papers on the Special Issue of CAGI 4 edition on Human Mutations (Hoskins RA et al, 2017) and other papers are now under writing process. Thanks to the expertise acquired in defining the phenotypic effect of variants, we collaborate directly with the Sant'Orsola Genetic Medical Unit to compute protein models of myosin 1F variants related to Thyroid Cancer, helping in directing their research with our computational approach.

In conclusions, we tried to depict the biological complexity merging a large-scale approach with the analysis of specific study cases. Although we are still far being able to predict the whole phenotypic appearance and disease state of a human being based only on the genetic information, we are now able to predict some simple phenotypic effects of gene variants and to relate gene-disease associations understanding the molecular mechanisms shared by genes involved in the same disease. With the study cases, we demonstrate that our computational methods have great results in predicting the outcome of simple experiments. Altogether, these findings help researchers and scientist in directing further efforts and in planning their experiments, and we believe that building networks of web servers, predictors and tools is a fundamental step for understanding the biological complexity.

12 References

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gku1205

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*. http://doi.org/10.1186/s12864-017-3911-3

Babbi, G., Martelli, P. L., & Casadio, R. (2018) PhenPath: a tool for characterizing biological functions underlying different phenotypes*. BMC Genomics* (submitted in 2018, under revision)

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., … Zardecki, C. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*. http://doi.org/10.1107/S0907444902003451

Bonvini, P., Gastaldi, T., Falini, B., & Rosolen, A. (2002). Nucleophosmin-anaplastic lymphoma kinase (NPM-ALK), a novel Hsp90-client tyrosine kinase: Down-regulation of NPM-ALK expression and tyrosine phosphorylation in ALK+CD30+lymphoma cells by the Hsp90 antagonist 17-allylamino,17-demethoxygeldanamycin. *Cancer Research*.

Bonvini, P., Rosa, H. D., Vignes, N., & Rosolen, A. (2004). Ubiquitination and Proteasomal Degradation of Nucleophosmin-Anaplastic Lymphoma Kinase Induced by 17-Allylamino-Demethoxygeldanamycin: Role of the Co-Chaperone Carboxyl Heat Shock Protein 70-Interacting Protein. *Cancer Research*. http://doi.org/10.1158/0008-5472.CAN-03-3531

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*. http://doi.org/10.1093/bioinformatics/btw508

Butler, M. G., Dazouki, M. J., Zhou, X. P., Talebizadeh, Z., Brown, M., Takahashi, T. N., … Eng, C. (2005). Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *Journal of Medical Genetics*. http://doi.org/10.1136/jmg.2004.024646

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., … Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btn615

Casey, G., Conti, D., Haile, R., & Duggan, D. (2013). Next generation sequencing and a new era of medicine. *Gut*. http://doi.org/10.1136/gutjnl-2011-301935

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., … Tyers, M. (2015). The BioGRID interaction database: 2015 update. *Nucleic acids research*. http://doi.org/10.1093/nar/gku1204

Coelho, T., Andreoletti, G., Ashton, J. J., Batra, A., Afzal, N. A., Gao, Y., … Ennis, S. (2016). Genes implicated in thiopurine-induced toxicity: Comparing TPMT enzyme activity with clinical phenotype and exome data in a paediatric IBD cohort. *Scientific Reports*. http://doi.org/10.1038/srep34658

Corey, D. R. (2016). Synthetic nucleic acids and treatment of neurological diseases. *JAMA Neurology*. http://doi.org/10.1001/jamaneurol.2016.2089

Crotti, L., Johnson, C. N., Graf, E., De Ferrari, G. M., Cuneo, B. F., Ovadia, M., … George, A. L. (2013). Calmodulin mutations associated with recurrent cardiac arrest in infants. *Circulation*. http://doi.org/10.1161/CIRCULATIONAHA.112.001216

Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P. L., Babbi, G., … Morgan, A. A. (2017). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Human Mutation*. http://doi.org/10.1002/humu.23280

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*. http://doi.org/10.1186/1471-2164-16-S8-S6

Diquigiovanni, C., Bergamini, C., Evangelisti, C., Isidori, F., Vettori, A., Tiso, N., … Bonora, E. (2018). Mutant MYO1F alters the mitochondrial network and induces tumor proliferation in thyroid cancer. *International Journal of Cancer*. http://doi.org/10.1002/ijc.31548

Eng, C. (2003). PTEN: One gene, Many syndromes. *Human Mutation*. http://doi.org/10.1002/humu.10257

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkx1132

Fariselli, P., Martelli, P. L., Savojardo, C., & Casadio, R. (2015). INPS: Predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btv291

Ficko-Blean, E., Stubbs, K. A., Nemirovsky, O., Vocadlo, D. J., & Boraston, A. B. (2008). Structural and mechanistic insight into the basis of mucopolysaccharidosis IIIB. *Proceedings of the National Academy of Sciences*. http://doi.org/10.1073/pnas.0711491105

Fisch, G. S. (2017). Whither the genotype-phenotype relationship? An historical and methodological appraisal. *American Journal of Medical Genetics, Part C: Seminars in Medical Genetics*. http://doi.org/10.1002/ajmg.c.31571

Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: A new style of protein science. *Nature Methods*. http://doi.org/10.1038/nmeth.3027

Gazzo, A. M., Daneels, D., Cilia, E., Bonduelle, M., Abramowicz, M., Van Dooren, S., … Lenaerts, T. (2016). DIDA: A curated and annotated digenic diseases database. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkv1068

Guccini, I., Serio, D., Condò, I., Rufini, A., Tomassini, B., Mangiola, A., … Malisan, F. (2011). Frataxin participates to the hypoxia-induced response in tumors. *Cell Death and Disease*. http://doi.org/10.1038/cddis.2011.5

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*. http://doi.org/10.1002/humu.23290

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1092

Kanehisa, M. (2013). Chemical and genomic evolution of enzyme-catalyzed reaction networks. *FEBS Letters*. http://doi.org/10.1016/j.febslet.2013.06.026

Kann, M. G. (2009). Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*. http://doi.org/10.1093/bib/bbp048

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., … Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*. http://doi.org/10.1016/j.ajhg.2009.09.003

Köhler, S., Vasilevsky, N. A., Engelstad, M., Foster, E., McMurry, J., Aymé, S., … Robinson, P. N. (2017). The human phenotype ontology in 2017. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1039

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkv1222

Lee, J. O., Yang, H., Georgescu, M. M., Cristofano, A. Di, Maehama, T., Shi, Y., … Pavletich, N. P. (1999). Crystal structure of the PTEN tumor suppressor: Implications for its phosphoinositide phosphatase activity and membrane association. *Cell*. http://doi.org/10.1016/S0092-8674(00)81663-3

Lovisa, F., Cozza, G., Cristiani, A., Cuzzolin, A., Albiero, A., Mussolin, L., … Bonvini, P. (2015). ALK kinase domain mutations in primary anaplastic large cell lymphoma: Consequences on NPM-ALK activity and sensitivity to tyrosine kinase inhibitors. *PLoS ONE*. http://doi.org/10.1371/journal.pone.0121378

Lu, L., Ghose, A. K., Quail, M. R., Albom, M. S., Durkin, J. T., Holskin, B. P., … Cheng, M. (2009). ALK mutants in the kinase domain exhibit altered kinase activity and differential sensitivity to small molecule ALK inhibitors. *Biochemistry*. http://doi.org/10.1021/bi8020923

Martelli, P. L., Fariselli, P., Savojardo, C., Babbi, G., Aggazio, F., & Casadio, R. (2016). Large scale analysis of protein stability in OMIM disease related human protein variants. *BMC Genomics*. http://doi.org/10.1186/s12864-016-2726-y

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell*. http://doi.org/10.1016/j.cell.2010.03.032

Noble, W. S. (2009). How does multiple testing correction work? *Nature Biotechnology*. http://doi.org/10.1038/nbt1209-1135

Nyegaard, M., Overgaard, M. T., Sondergaard, M. T., Vranas, M., Behr, E. R., Hildebrandt, L. L., … Borglum, A. D. (2012). Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *American Journal of Human Genetics*. http://doi.org/10.1016/j.ajhg.2012.08.015

O 'Brien, J. S. (1972). Sanfiippo Syndrome: Profound Deficiency of Alpha-Acetylglucosaminidase Activity in Organs and Skin Fibroblasts from Type-B Patients. *Proc. Nat. Acad. Sci. USA*. http://doi.org/10.1073/pnas.69.7.1720

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., … Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkt1115

Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., … Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw943

Pontikos, N., Yu, J., Moghul, I., Withington, L., Blanco-Kelly, F., Vulliamy, T., … Plagnol, V. (2017). Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btx147

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Stein, T. I., Levitt, J., … Lancet, D. (2017). MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1012

Relling, M. V., Pui, C. H., Cheng, C., & Evans, W. E. (2006). Thiopurine methyltransferase in acute lymphoblastic leukemia [4]. *Blood*. http://doi.org/10.1182/blood-2005-08-3379

Salavaggione, O. E., Wang, L., Wiepert, M., Yee, V. C., & Weinshilboum, R. M. (2005). Thiopurine S-methyltransferase pharmacogenetics: Variant allele functional and comparative genomics. *Pharmacogenetics and Genomics*. http://doi.org/10.1097/01.fpc.0000174788.69991.6b

Savojardo, C., Fariselli, P., Martelli, P. L., & Casadio, R. (2016). INPS-MD: A web server to predict stability of protein variants from sequence and structure. *Bioinformatics*. http://doi.org/10.1093/bioinformatics/btw192

Savojardo, C., Babbi, G., Bovo, S., Capriotti, E., Martelli, P. L., & Casadio, R. (2019). Are machine learning based methods suited to address complex biological problems? Lessons from CAGI-5 challenges*. Human Mutation*. (submitted in 2019, under revision)

Schulz, T. J., Thierbach, R., Voigt, A., Drewes, G., Mietzner, B., Steinberg, P., … Ristow, M. (2006). Induction of oxidative metabolism by mitochondrial frataxin inhibits cancer growth: Otto Warburg revisited. *Journal of Biological Chemistry*. http://doi.org/10.1074/jbc.M511064200

Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. http://doi.org/10.1093/nar/29.1.308

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shaw, K., & Cooper, D. N. (2012). The human gene mutation database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics*. http://doi.org/10.1002/0471250953.bi0113s39

Song, M. S., Salmena, L., & Pandolfi, P. P. (2012). The functions and regulation of the PTEN tumour suppressor. *Nature Reviews Molecular Cell Biology*. http://doi.org/10.1038/nrm3330

Sun, S., Yang, F., Tan, G., Costanzo, M., Oughtred, R., Hirschman, J., … Roth, F. P. (2016). An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*. http://doi.org/10.1101/gr.192526.115

Swindell, W. R., Bojanowski, K., Kindy, M. S., Chau, R., & Ko, D. (2018). GM604 regulates developmental neurogenesis pathways and the expression of genes associated with amyotrophic lateral sclerosis. *Translational neurodegeneration*. . http://doi.org/10.1186/s40035-018-0135-7

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., … Von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gku1003

Tartari, C. J., Gunby, R. H., Coluccia, A. M. L., Sottocornola, R., Cimbro, B., Scapozza, L., … Gambacorti-Passerini, C. (2008). Characterization of some molecular mechanisms governing autoactivation of the catalytic domain of the anaplastic lymphoma kinase. *Journal of Biological Chemistry*. http://doi.org/10.1074/jbc.M706067200

The gene ontology consortium. (2017). Expansion of the gene ontology knowledgebase and resources.. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1108

The UniProt Consortium. (2017). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1099

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*. http://doi.org/10.1126/science.1260419

Valstar, M. J., Ruijter, G. J. G., van Diggelen, O. P., Poorthuis, B. J., & Wijburg, F. A. (2008). Sanfilippo syndrome: A mini-review. *Journal of Inherited Metabolic Disease*. http://doi.org/10.1007/s10545-008-0838-5 von Figura, K., & Kresse, H. (1972). The Sanfilippo B corrective factor: A N-acetyl-α-D-glucosaminidase. *Biochemical and Biophysical Research Communications*. http://doi.org/10.1016/S0006-291X(72)80044-5

Weile, J., Sun, S., Cote, A. G., Knapp, J., Verby, M., Mellor, J. C., … Roth, F. P. (2017). Expanding the Atlas of Functional Missense Variation for Human Genes. *BioRxiv*. http://doi.org/10.1101/166595

Yates, B., Braschi, B., Gray, K. A., Seal, R. L., Tweedie, S., & Bruford, E. A. (2017). Gene-names.org: The HGNC and VGNC resources in 2017. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkw1033

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*. http://doi.org/10.1016/j.jmb.2005.08.020

Xu, Q., Tang, Q., Katsonis, P., Lichtarge, O., Jones, D., Bovo, S., … Dunbrack, R. L. (2017). Bench-marking predictions of allostery in liver pyruvate kinase in CAGI4. *Human Mutation*. http://doi.org/10.1002/humu.23222

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkx1098