



EvoPPI 1.0: a Web Platform for Within- and Between-Species Multiple Interactome Comparisons and Application to Nine PolyQ Proteins Determining Neurodegenerative Diseases

Noé Vázquez^{1,2} · Sara Rocha^{3,4} · Hugo López-Fernández^{1,2,3,4,5}  · André Torres^{3,4} · Rui Camacho⁶ · Florentino Fdez-Riverola^{1,2,5} · Jorge Vieira^{3,4} · Cristina P. Vieira^{3,4} · Miguel Reboiro-Jato^{1,2,5}

Received: 24 August 2018 / Revised: 7 January 2019 / Accepted: 9 January 2019 / Published online: 1 February 2019
© International Association of Scientists in the Interdisciplinary Areas 2019

Abstract

Protein–protein interaction (PPI) data is essential to elucidate the complex molecular relationships in living systems, and thus understand the biological functions at cellular and systems levels. The complete map of PPIs that can occur in a living organism is called the interactome. For animals, PPI data is stored in multiple databases (e.g., BioGRID, CCSB, DroID, FlyBase, HIPPIE, HitPredict, HomoMINT, INstruct, Interactome3D, mentha, MINT, and PINA2) with different formats. This makes PPI comparisons difficult to perform, especially between species, since orthologous proteins may have different names. Moreover, there is only a partial overlap between databases, even when considering a single species. The EvoPPI (<http://evoppi.i3s.up.pt>) web application presented in this paper allows comparison of data from the different databases at the species level, or between species using a BLAST approach. We show its usefulness by performing a comparative study of the interactome of the nine polyglutamine (polyQ) disease proteins, namely androgen receptor (AR), atrophin-1 (ATN1), ataxin 1 (ATXN1), ataxin 2 (ATXN2), ataxin 3 (ATXN3), ataxin 7 (ATXN7), calcium voltage-gated channel subunit alpha 1 A (CACNA1A), Huntingtin (HTT), and TATA-binding protein (TBP). Here we show that none of the human interactors of these proteins is common to all nine interactomes. Only 15 proteins are common to at least 4 of these polyQ disease proteins, and 40% of these are involved in ubiquitin protein ligase-binding function. The results obtained in this study suggest that polyQ disease proteins are involved in different functional networks. Comparisons with *Mus musculus* PPIs are also made for AR and TBP, using EvoPPI BLAST search approach (a unique feature of EvoPPI), with the goal of understanding why there is a significant excess of common interactors for these proteins in humans.

Keywords Protein–protein interactions databases · Inter-specific comparisons · PolyQ disease proteins

1 Introduction

Information on the function and molecular properties of individual proteins is available in major databases such as UniProt [1]. To be functional, most proteins establish physicochemical dynamic connections with other proteins. Finding these interactions provides opportunities to explore their biological functions [2]. The map of the protein–protein

interactions (PPIs) in a particular organism is called the interactome [3]. Aberrant PPIs are detected in multiple aggregation-related diseases, such as polyglutamine diseases, Creutzfeldt–Jakob, Parkinson’s, Alzheimer’s, and cancer [4, 5]. The comparison of PPI networks in patients and controls can elucidate the molecular basis of these diseases and lead to the identification of possible therapeutic targets.

While several computational and experimental methods based on single or high-throughput screens have been implemented for detecting PPIs, all present advantages and disadvantages [6]. Computational methods (e.g., text mining, docking, machine learning, interolog mapping, and so forth [7]), are able to detect thousands of PPIs in much less time and at a lower cost than experimental methods; however, since these methods are based on predictions and not on experimental data, accuracy is always an issue.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12539-019-00317-y>) contains supplementary material, which is available to authorized users.

✉ Hugo López-Fernández
hlfernandez@uvigo.es

Extended author information available on the last page of the article

Nevertheless, computational methods can be very useful to understand which interactions may be missing in the available experimental dataset. In addition, the use of high-throughput screen methods cannot guarantee the capture of all interactions, and the most used experimental high-throughput screen techniques (i.e., mass spectrometry, two-hybrid assays, and tandem affinity purification) can produce rates of false positive interactions up to 50% [2]. Finally, methods that are unlikely to generate false positives, such as X-ray crystallography, are not easy to scale up and thus cannot be used to study large numbers of PPIs.

PPI datasets of several species, obtained with different detection methods, are publicly accessible and can be downloaded from databases such as BioGRID [8, 9], CCSB [10–14], DroID [15], FlyBase [16], HIPPIE [2], HitPredict [17], HomoMINT [18], INstruct [19], Interactome3D [20], Mentha [21], MINT [22], or PINA [23]. Although there is some degree of overlap between databases, every database reports an exclusive set of information, and since interactions can be reported in different formats, the comparison between databases can be demanding (e.g., BioGRID, MINT and CCSB report interactions using gene identifiers, UniProt numbers, and gene names, respectively). Databases can be human-curated (e.g., BioGRID, HIPPIE, and MINT) and can report the source of each PPI (e.g., BioGRID and MINT). Furthermore, functionally equivalent proteins can have distinct names in different species, making the comparison across species difficult to achieve. Since each method and database presents advantages and disadvantages, interactions reported in several independent studies, or in distinct species, are expected to be more reliable than those reported in a single study using high-throughput methods. Furthermore, as stated above, the comparison of interactomes obtained under different conditions (e.g., patients and controls) can be informative.

This paper presents EvoPPI (<http://evoppi.i3s.up.pt>), an open-source web application that aims to effortlessly compare PPI datasets across databases and species. Since proteins can have different names in the species being compared, a BLAST-based approach is used for across-species comparisons, allowing users to specify different criteria to select the proteins that are considered functionally equivalent. It should be noted, however, that EvoPPI is not an application for PPI inference using homology. Four parameters can be adjusted by the user: (1) number of descriptions to report, which controls the number of sequences to be reported in the output; (2) the expect value, which describes the number of hits expected by chance when searching a database of a particular size (lower *E*-value represents more “significant” match); (3) the minimum percentage of identity that the sequence alignment must have to be considered a positive match; and (4) the minimum length of the aligned block, which specifies the size the sequence alignment must

have to be considered a positive match. These features are useful when comparing organisms such as *Homo sapiens* and *Drosophila melanogaster*, where two rounds of whole genome duplication occur in human lineage [24], implying that the majority of *Drosophila* genes have multiple paralogs in humans. In short, EvoPPI presents distinctive features such as the use of a BLAST approach for the identification of orthologous/paralogous genes (where the user can define the number of descriptions, the minimum expect value, the minimum length of alignment blocks, and the minimum identity), and the use of colour codes for an effortless detection of differences between datasets.

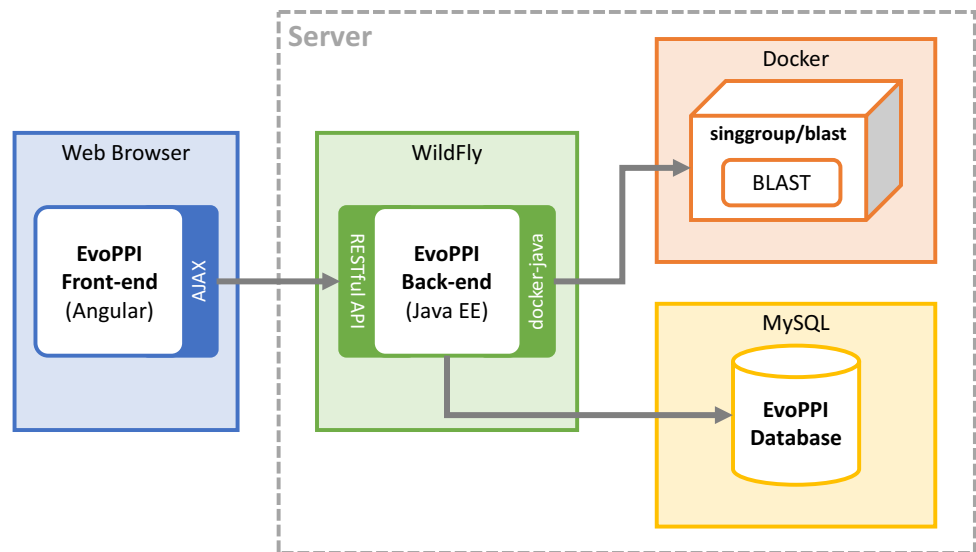
To demonstrate the usefulness of EvoPPI, we will compare the human interactomes for ATXN1, ATXN2, CACNA1A, ATXN7, TBP, ATXN3, HTT, ATN1, and AR, the nine polyglutamine (polyQ) proteins that are associated to degenerative disorders due to an expansion of the polyQ tract. These proteins are responsible for six spinocerebellar ataxias (SCA) types 1, 2, 6, 7, 17, Machado–Joseph disease (MJD or SCA3), Huntington’s disease (HT), dentatorubral pallidolusian atrophy (DRPL), and spinal and bulbar muscular atrophy X-linked 1 (SBMA), respectively [25]. We will begin by demonstrating that there is no protein in common to all the polyQ disease interactomes. We will then show that when considering those proteins shared between the interactomes of at least four of the polyQ disease proteins, six are found to belong to the ubiquitin pathway. Comparisons with *Mus musculus* PPIs are also made for AR and TBP, using the EvoPPI BLAST search approach for distinct species comparisons, to explore why there is a significant excess of common interactors for these proteins in humans.

2 Materials and Methods

2.1 Data

EvoPPI relies on two main types of data to perform the analyses: reference genomes of the species (FASTA files) and interactomes (TSV files with the interactions). The current version of EvoPPI includes the reference genomes of ten animal species: *Bos taurus*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus*, *Rattus norvegicus*, and *Xenopus laevis* (see Supplementary Table 1 for more details). For each of these 10 species, more than 100 PPIs are available in at least 1 interactome database. The reference genomes were downloaded from NCBI in GenBank Flat File Format (GBFF) and parsed to extract the CoDing sequences (CDSs) and create the FASTA files required by EvoPPI. In the same operation, dictionaries of gene synonyms for each species were also created. These dictionaries

Fig. 1 Architecture of EvoPPI, showing its main components and their interactions



are used by EvoPPI to allow users to look for different gene names.

The current version of EvoPPI also includes 52 interactomes for the 10 species (see Supplementary Table 1 for additional details). We downloaded all the available interactomes from the following databases: BioGRID, CCSB, DroID, FlyBase, HIPPIE, HitPredict, HomoMINT, Instruct, Interactome3D, mentha, MINT, and PINA. We then parsed each database file to convert them into a unified format that EvoPPI can handle, that is, we converted each file into a simple TSV file with two columns that represent the Gene-ID identifiers of the genes involved in each reported interaction. It is important to note that this process requires converting the gene identifiers from their source formats (UniProtKB-ID, Gene name, or FlyBase; Supplementary Table 1) into Gene-ID, which sometimes requires a two-step conversion: first converting them to UniProtKB-ID and then to Gene-ID. To perform this step, we used the mapping API offered by UniProtKB.¹ However, some interactions were lost, because they could not be converted (see Supplementary Table 1 for additional details).

All the information managed by EvoPPI, including interactomes, species, and gene data, is stored in a relational database. This allows fast information retrieval, reducing the time required to process user queries. The current version of EvoPPI also includes support for user registration, allowing users to keep and manage their query results.

2.2 EvoPPI architecture

EvoPPI is composed of two different applications that act as the front-end and the back-end components, respectively.

The front-end application is a web application that was implemented using the Angular v6 framework² in combination with the Angular Material v6 library³ and the Material Dashboard Angular 5 template,⁴ for a richer user interface. The back-end application was implemented using the Java EE 7 platform.⁵ This application provides a RESTful API [26] with resources to access data and to request PPIs calculation. Communication between front-end and back-end applications is done using Asynchronous JavaScript and XML (AJAX) and JavaScript Object Notation (JSON) for data encoding.

EvoPPI relies on BLAST to perform sequence alignment between the gene sequences of distinct species, to identify orthologous/paralogous genes. As explained before, this identification is needed to enable a comparison of interactomes belonging to distinct species. To avoid installation and configuration issues, a Docker⁶ container was created with a BLAST v2.6.0 installation. This container is invoked from the back-end application using the docker-java v3.0.13 library.⁷

Figure 1 represents the general architecture and deployment of EvoPPI, including the components described above. EvoPPI is currently running in a WildFly v10.1.0 application server⁸ and uses a MySQL v5.7 database management system⁹ to store the information.

² <https://angular.io/>.

³ <https://material.angular.io/>.

⁴ <https://www.creative-tim.com/product/material-dashboard-angular2/>.

⁵ <https://www.oracle.com/es/java/technologies/java-ee.html>.

⁶ <https://www.docker.com/>.

⁷ <https://github.com/docker-java/docker-java>.

⁸ <http://wildfly.org/>.

⁹ <https://www.mysql.com/>.

¹ https://www.uniprot.org/help/api_idmapping.

EvoPPI 1.0 is publicly accessible at <http://evoppi.i3s.up.pt>. It is an open-source software distributed under a GPLv3 license. The source code of the front-end application is publicly available at <https://github.com/sing-group/evoppi-front-end>, while the source code of the back-end application is available at <https://github.com/sing-group/evoppi-backend>. Finally, the Docker container with the BLAST installation can be found at <https://hub.docker.com/r/singgroup/evoppi-blast/>.

2.3 Interactome Comparison Algorithms

EvoPPI allows users to compare the interactions of a gene (i.e., the *query gene*) in two or more interactomes, which may belong to the same or distinct species. Depending on this aspect, the algorithm used to perform the calculations is different.

2.3.1 Same Species Comparison

To retrieve the interactions for a given query gene in two or more interactomes belonging to the same species, the following algorithm is applied:

1. *Interactions calculation step*: for each *query interactome*, retrieve from the database the interactions where the query gene is present. EvoPPI allows specifying the *interaction level*, which is the degree of distance (up to a maximum of three) to retrieve transitive interactions. Therefore, if the degree is greater than one, after retrieving the genes that interact with the *query gene*, the process is repeated and the genes that interact with these degree 1 genes are also retrieved. This process is repeated as many times as the degree specified by the user and it results in a set of interactions, each one containing the interacting genes, the degree and the associated source *query interactome*.
2. *Interactions completion step*: iterate over all the interactions resulting from the previous step in order to check if they are present in the other *query interactomes* but were not discovered in the previous step. If so, add them with an unknown *interaction level* (i.e., -1).

For example, as Fig. 2 illustrates, the following situation may occur: using an interaction level of 3 in the first step of the algorithm, the query gene A gives interactions $A \rightarrow B$, $B \rightarrow C$, and $C \rightarrow D$ in Interactome 1, but only the interaction $A \rightarrow B$ is present in Interactome 2. Although the interaction $C \rightarrow D$ is present in Interactome 2, it cannot be discovered because $B \rightarrow C$ does not exist. This completion step adds this kind of interaction with an interaction level of -1 , to indicate that the interaction is present in the interactome but the degree is unknown.

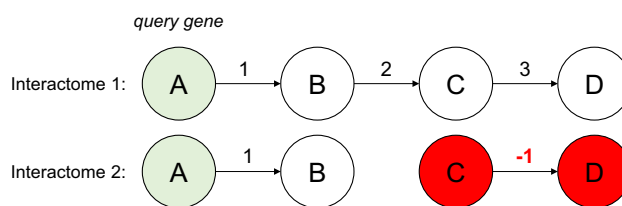


Fig. 2 Exemplification of the Interactions completion step. The interactions set is completed by adding $C \rightarrow D$ (in red), discovered in the Interactome 2 by this second step

2.3.2 Distinct Species Comparison

Queries where interactomes belong to two distinct species follow a more complex process. In this case, the name reference interactomes is given to the interactomes of the species (i.e., reference species) to which the query gene belongs, and the name target interactomes is given to the interactomes of the second, distinct species (i.e., target species). To retrieve the interactions for a given query gene in two or more interactomes belonging to two distinct species, the following algorithm is applied:

1. *Interactions calculation step in reference interactomes*: apply the same procedure described in the interaction calculation step for same species comparisons in all *reference interactomes*.
2. *Interactions completion step in reference interactomes*: apply the same procedure described in the interactions completion step for same species for all the interactions obtained in the previous step.
3. *Query gene BLAST*: perform a BLAST query of the *query gene* and all genes involved in the set of interactions obtained in the previous step against the target genome to find their orthologous/paralogous genes.
4. *Interactions calculation step in target interactomes*: if the *query gene* has any orthologous/paralogous genes in the *target species*, apply the same procedure described in the interaction calculation step for same species comparisons to all its orthologous/paralogous genes in all *target interactomes*. As a restriction, interactions that do not have an orthologous/paralogous gene among the genes retrieved in step 1 are discarded.
5. *Interactions completion step in reference interactomes*: apply the same procedure described in the interactions completion step for same species for all the interactions obtained in the previous step. In this case, the interactions used as reference are those obtained in step 1 (i.e., *reference interactions*), instead of those obtained in step 4 (i.e., *target interactions*). BLAST results obtained in step 3 are used to determine the

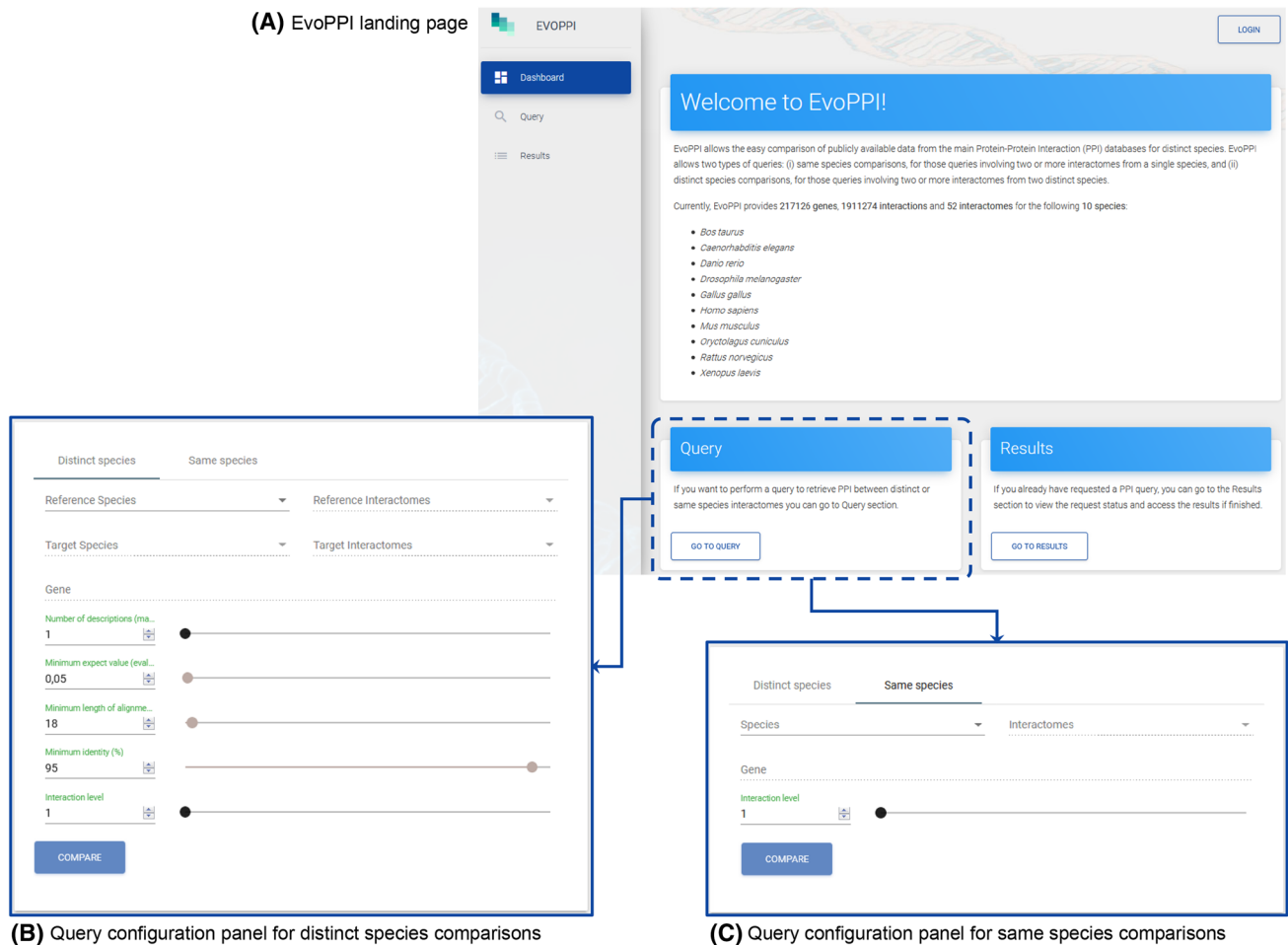


Fig. 3 Screenshots of EvoPPI: **a** the EvoPPI landing page, which gives access to the main functionalities (queries, results management and user login), **b** the query configuration panel for distinct species

comparison, including the BLAST parameters, and **c** the query configuration panel for same species comparisons

orthologous/paralogous relationships between *reference* and *target interactions*.

It is important to note that this algorithm will only retrieve interactions for those genes (or their corresponding orthologous/paralogous) discovered in the first step.

2.4 EvoPPI User Interface




EvoPPI provides an easy-to-use user interface specially designed for users without advanced bioinformatics skills. The landing page of EvoPPI (Fig. 3a) allows users to access the query interface, supporting the two types of analysis: one panel to perform same species comparisons (Fig. 3b) and another for distinct species comparisons (Fig. 3c). In both cases, users start by selecting the species, interactomes, and the query gene to perform the search. Despite EvoPPI using Gene-ID identifiers as the main identifier for the genes, it also keeps other alternative




names. When a user starts to write a gene name, EvoPPI looks for that text in the gene identifier and for alternative names, to show a list of genes from which the user can select the query gene. In addition to these parameters, the interaction level parameter can be used in both query types to select the maximum degree of distance of the retrieved interactions.

The distinct species query form (Fig. 3b) also includes four parameters to configure the BLAST execution and filter the results. These parameters are: (1) the number of descriptions (BLAST *max_target_seqs* parameter); (2) the expect value (BLAST *eval* parameter); (3) the minimum length of alignment blocks; and (4) the minimum identity, expressed as a percentage.

Although queries for the same species are completed in a few seconds, queries across species can take minutes or even hours, due to the BLAST sequence alignment step. Keeping this in mind, EvoPPI was designed to perform the queries asynchronously, so that users can launch a query, leave the

(A) Results list

Distinct species							
Analysis done with interactomes coming from two distinct species							
Scheduled On	Query Gene	Max. Degree	Reference Species	Target Species	#Interactomes	Status	Actions
5/8/2018 9:54:50	ADH1A [124]	1	Homo sapiens	Drosophila melanogaster	1 / 1	COMPLETED	  

Same species							
Analysis done with interactomes coming from the same species							
Scheduled On	Query Gene	Max. Degree	Species	#Interactomes	Status	Actions	
5/8/2018 9:51:26	Fdh [41311]	1	Drosophila melanogaster	3	COMPLETED	  	

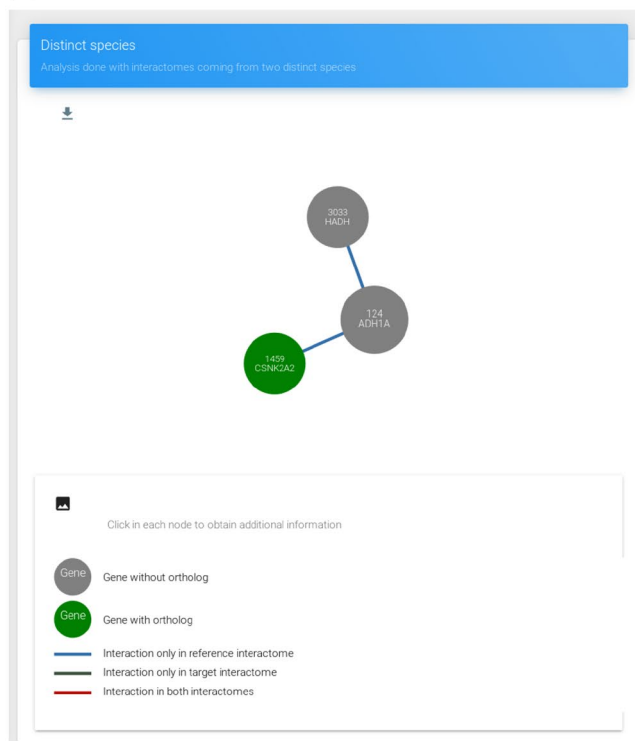
(B) Tabular results view

Distinct species							
Analysis done with interactomes coming from two distinct species							
Gene A	Name A	Gene B	Name B	Homo sapiens: BioGRID	Drosophila melanogaster: BioGRID		
124	ADH1A	1459	CSNK2A2	Degree 1			
124	ADH1A	3033	HADH	Degree 1			

Items per page: 10 1-2 of 2

Click in each Gene name to obtain additional information

Gene without ortholog (grey)
Gene with ortholog (green)

(C) Interaction graph view**(D)** Additional gene information

Gene: CSNK2A1, CK2alpha, CSNK2A2

GeneBank: 1459

Sequences

MFGPAGASRRARYAEVNSLSREYBYEAHVPFSGNQSDQVLRKLRGKYSSEVFEANTNERNVYVLLKQLYDITDFRMYVELLAKLYDCHSKGMHRDVKPHNVMEDHQKLLDLWGLAE MFGPAGASRRARYAEVNSLSREYBYEAHVPFSGNQSDQVLRKLRGKYSSEVFEANTNERNVYVLLKPKKKIKRLEKLEENRSGTNRKLDITVDPYPSKTPALYEVYNTDPKQLYDITDFR MFEELKALDICHSGMBRHSWPHNVMEDHQKLELDWGLAEFFHPAGEYVWVNSRPFKDFELLDYQWYVLSLDMVSLGMLASMRREFFRFGSDHWYQLVRSARVLDTEELDYLYVHLEL MSQNNFVHQLVVEYYSARKNLDSEQKQVYTGKCALDYCHSKGMHRDVKPHNVMEDHQKLLDLWGLAEFFHPAGEYVWVNSRPFKDFELLDYQWYVLSLDMVSLGMLASMRREFFR

BLAST Results: 3

Close Download sequences

Fig. 4 Screenshots of the results management interfaces: **a** the EvoPPI results list, separated in distinct species and same species; **b** tabular results view of a distinct species query involving the gene *ADH1A*; **c** graph view for the same gene; and **d** additional gene information

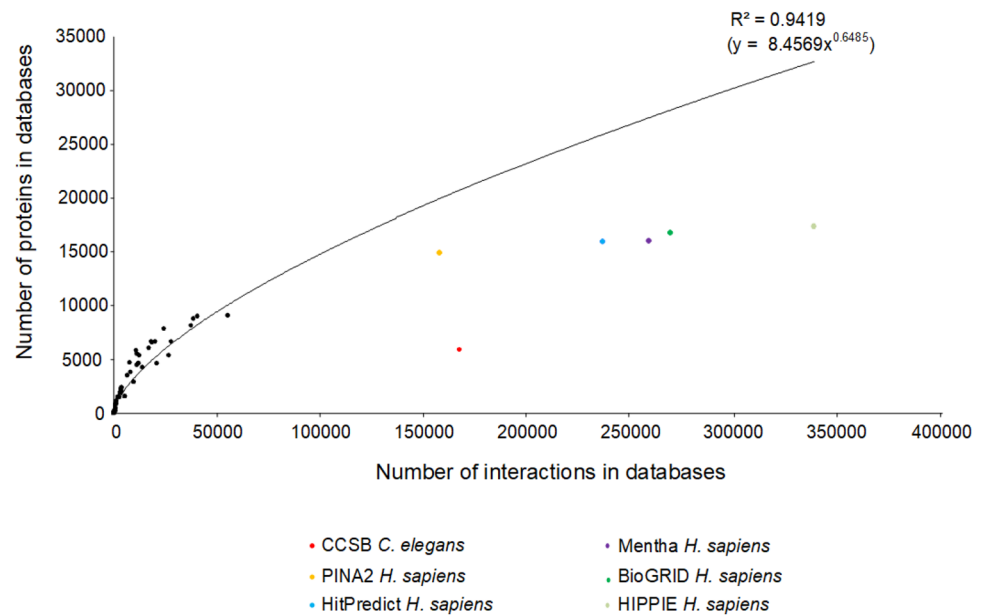
application, and return later to check the execution status. To do so, EvoPPI implements three mechanisms: (1) it generates a unique URL for each query that users can use to return at any time to the query result; (2) it stores the results in the users' browser storage so that they can be reopened later; and (3) it stores the queries in the EvoPPI database when users are logged in the web application.

The query results are listed in the results management interface (Fig. 4a). Each query result is presented in tabular (Fig. 4b) and graph formats (Fig. 4c). The tabular view lists the interactions, including the gene identifiers and names, and the interaction degree in each interactome, while the graph view represents the results as an undirected graph,

where nodes are genes and edges are interactions. Different colours are used to represent the presence or absence of genes and interactions in the interactomes, while node sizes are used to represent the number of interactions for each gene. In both views, genes can be clicked to view detailed information (Fig. 4d), including the gene identifier, alternative names, the protein sequence and, for distinct species queries, the related results of the BLAST alignment.

Finally, the results can be exported in several formats, including a comma-separated values (CSV) file with the retrieved interactions, FASTA files with the protein sequences of the interactomes, and the interactions graph in different image formats.

Fig. 5 The relationship between the number of unique interactions and the number of unique proteins for the 52 datasets used (Supplementary Table 1)



3 Results and Discussion

3.1 Databases

Currently, EvoPPI incorporates data from 12 databases for 10 species, for which a minimum of 100 unique interactions have been reported, totalling 52 PPI datasets (Supplementary Table 1). EvoPPI uses Gene-ID as the main feature by which the corresponding proteins are identified in the database. Since not all publicly available databases use Gene-ID to identify the interacting proteins, data from those databases had to be converted using the UniProtKB ID mapping API, as described in the Material and Methods section. The average rate of conversion is high (87.9%), although it never reaches 100% (Supplementary Table 1).

The relationship between the number of unique interactions and the number of unique proteins for the 52 datasets is presented in Fig. 5. A power function with a coefficient of 0.6485 fits the data well ($R^2 = 0.94$). The six largest datasets, namely *Caenorhabditis elegans* CCSB, and *H. sapiens* PINA2, mentha, HitPredict, BioGRID, and HIPPIE, have a much larger number of unique interactions than what was expected based on the number of unique proteins. This could suggest that: (1) smaller datasets are biased towards proteins showing many interactions; (2) the largest datasets include an important fraction of false positive interactions; or (3) although all data was downloaded from the main PPI databases, the largest datasets may include non-PPI interactions, such as gene interactions. The latter is likely the case for the CCSB *C. elegans* interactome, which integrated the WI8 interactome with evidence for functional relationships based

on mRNA co-expression data available in WormBase12, RNAi phenotypes from RNAiDB24, genetic interactions curated in WormBase12, interolog interactions, and protein–protein interactions from the literature curated dataset [10].

The number of PPI in the datasets overlaps only partially, as shown in Fig. 6, for the nine polyQ disease proteins and for the five largest datasets (PINA2, Mentha, HitPredict, HIPPIE, and BioGRID 3.4). Furthermore, the database with the highest number of PPIs differs from protein to protein. Therefore, the integration of all databases is needed to obtain all PPI available for a particular protein. This can be performed easily and quickly using the EvoPPI “Compare same species” operation.

3.2 Case Study

PolyQ-containing proteins are enriched in protein complexes [27, 29]. Moreover, polyQ regions are usually located close to coiled-coil regions suggesting that they play a role in protein interaction regulation [27–29]. Therefore, it is not surprising that polyQ proteins have more PPI partners than non-polyQ proteins, and that they have a higher tendency to interact with other polyQ proteins than non-polyQ proteins [27]. In humans, 60 polyQ proteins have been described in the complete proteome [30]. Of those, nine of them, when expanded, cause neurodegenerative polyQ diseases [25], namely AR, ATN1, ATXN1, ATXN2, ATXN3, ATXN7, CACNA1A, HTT, and TBP. These nine proteins could be functionally related, despite not having any sequence homology. Therefore, our case study compares the interactors of these nine proteins using the EvoPPI web application.

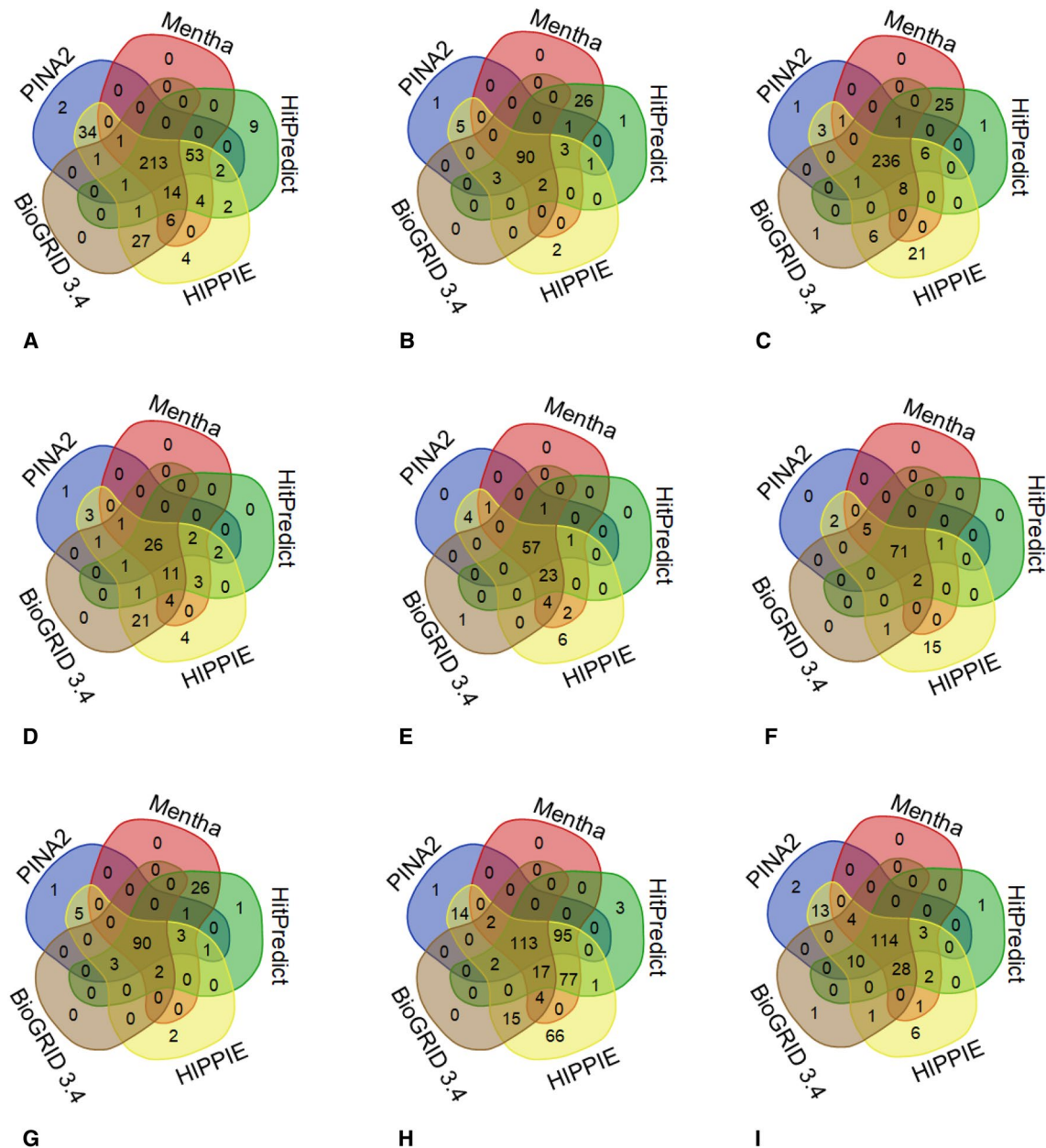


Fig. 6 The number of PPI in the PINA2 (in blue), Mentha (pink), HitPredict (green), HIPPIE (yellow), and BioGRID 3.4 (brown) datasets for: **a** androgen receptor (AR); **b** atrophin-1 (ATN1); **c** ataxin

1 (ATXN1); **d** ataxin 2 (ATXN2); **e** ataxin 3 (ATXN3); **f** ataxin 7 (ATXN7); **g** calcium voltage-gated channel subunit alpha1 A (CACNA1A); **h** Huntingtin (HTT); and **i** TATA-binding protein (TBP)

Four interactions are found among polyQ disease proteins, namely HTT/TBP, AR/TBP, ATXN1/ATXN2 and ATXN7/TBP, which suggests that different proteins participate in the same biological pathways. Nevertheless, none of the proteins reported in the databases interacts with all nine wild-type polyQ disease proteins (Table 1). Indeed, the majority of the interactors bind to a single polyQ disease protein. This is an interesting observation given that some of the datasets considered in this study may be reporting interactions that were detected with proteins having an extended

pathological polyQ only, possibly biasing the results towards an enrichment of common interactors. This suggests that: (1) these proteins participate in different biological processes, and/or (2) there are not many proteins that bind non-specifically to polyQ disease proteins with the help of the polyQ region. Only Polyubiquitin-C (UBC), a protease and RNA-binding protein, is reported to bind to eight out of the nine polyQ disease proteins (Supplementary Table 2). Nevertheless, based on the data available at the 14 *H. sapiens* datasets, UBC likely interacts with more than 50% of all human

Table 1 Distribution of the proteins of the nine polyQ disease interactomes according to presence in one or multiple polyQ disease proteins

Number of polyQ disease proteins	Number of interacting proteins
1	1213
2	196
3	50
4	10
5	3
6	1
7	0
8	1

proteins. There are 14 other proteins that bind to at least 4 of the 9 polyQ disease proteins, namely SUMO1, SUMO2, VCP, PIAS1, CREBBP, EP300, GAPDH, EFEMP2, TP53, TBP, UBE2I, CASP1, CASP3, and NCOR1 (Supplementary Table 2). The list of the 15 interactors that interact with at least 4 of 9 polyQ disease proteins is enriched, according to PANTHER¹⁰ in the molecular function “ubiquitin protein ligase binding” (6 proteins: SUMO1, SUMO2, UBC, PIAS1, TP53 and VCP; fold enrichment = 17.60; FDR = 9.07E−04). At least one of the six proteins that belong to the “ubiquitin protein ligase binding” GO term interacts with each of the polyQ disease proteins in this study (6, 2, 5, 3, 4, 4, 1, 5 and 3 interacts with AR, ATN1, ATXN1, ATXN2, ATXN3, ATXN7, CACNA1A, HTT and TBP, respectively; Supplementary Table 2). This distribution may reflect the number of large-scale studies dedicated to each of the proteins. Therefore, this observation suggests that ubiquitination is an important factor in the regulation of these nine polyQ disease proteins. Indeed, regulation of the ubiquitination machinery has been indicated as a potential therapeutic target in polyglutamine diseases [31–34]. These enzymes target proteins for degradation both by the proteasome and by autophagy [31–34].

Of the 261 proteins that interact with at least two polyQ disease proteins, as many as 42.1% have at least 1 paralogous protein, with an average of 2.44 (Supplementary Table 2). Members of the histone H3 family have as many as ten paralogous in this list. It should be noted that there is a clear co-occurrence of paralogous proteins. For instance, for the AR, only 20% of proteins with a paralogous do not show a presence/absence agreement with all other paralogous proteins. When comparing the interactors of the polyQ disease proteins, in ten cases the number of common interactors is lower than expected by chance, again suggesting that most

polyQ disease proteins are involved in different functional networks (Table 2). For transcription factors AR/TBP, the number of common interactors is larger than expected by chance (Table 2), suggesting that either they are involved in the same biological pathway, or that many proteins are binding due to the presence of a polyQ region, which facilitates the interaction with other proteins. To address this issue, we used the EvoPPI BLAST search approach to identify orthologous/paralogous proteins in *M. musculus* (number of descriptions of 1; minimum expect value of 0.05; minimum length of alignment block of 40; minimum identity of 40%; interaction level 1). We have observed that, since the latter species shows shorter polyQ tracts at the N-terminal region of both AR and TBP proteins, as expected (Supplementary Fig. 1A, B), there are no common interactors when comparing TBP and AR (Fig. 7). 86% of the proteins that interact with TBP in *M. musculus* interact, and were identified as orthologous/paralogous of proteins interacting with TBP in humans, interact only with TBP in humans as well. Moreover, 67% of the proteins that interact with AR in *M. musculus*, and were identified as orthologous/paralogous of proteins interacting with AR in humans, interact only with AR in humans as well (Fig. 7). These observations suggest that, in humans, the sample size for both TBP and AR is already large enough to identify the majority of the proteins that interact with both of them. When we compare the number of common interactions and unique interactions in humans (48 vs. 479) and house mice (0 vs. 30) the proportion is non-significant ($p = 0.10$). The non-significant proportion in the number of common interactors and unique interactors in humans (where both genes encode proteins with polyQ) and in house mice (where there is either no polyQ or the size of the polyQ is shorter than in humans; Supplementary Fig. 1A, B), suggests that the larger number of common interactors between the two proteins could be attributed mainly to the involvement of these proteins in common biological pathways. Nevertheless, despite the large number of reported interactors for both proteins in humans, there are only 12.8% and 24.1% of common interactors for AR and TBP, respectively. Moreover, given the much smaller number of interactors reported for both *M. musculus* AR and TBP than for *H. sapiens*, this test may lack statistical power. This possibility must be considered since, for both humans and yeast, it has been reported that polyQ proteins have more PPI partners than non-polyQ proteins [27]. This tendency is observed in all species analysed in our case study having more than 10,000 interactions reported in at least 1 interactome database (*H. sapiens*, *M. musculus*, *R. norvegicus*, *B. taurus*, *D. melanogaster*, and *C. elegans*; Supplementary Fig. 2A–F, respectively), although only for *H. sapiens*, *M. musculus*, and *D. melanogaster*, polyQ proteins have a significantly larger number of interactors than non-polyQ proteins (Supplementary Fig. 2). Nevertheless, polyQ proteins tend to

¹⁰ <http://pantherdb.org>.

Table 2 Comparisons of the interactors of the polyQ disease proteins

Proteins	Status*	ATN1	ATXN1	ATXN2	ATXN3	ATXN7	CACNA1A	HTT	TBP (68)
AR (112)	C	10	28	9	15	8	7	40	48
	OO	102/45	84/66	103/18	97/28	104/49	105/39	72/69	64/20
	N	104	83	131	121	100	110	80	129
ATN1 (56)	C		23	7	7	14	17	11	1
	OO		32/71	48/20	48/36	41/43	38/29	44/98	54/67
	N		135	186	170	163	177	108	139
ATXN1 (94)	C			10	7	21	6	39	10
	OO			84/17	87/36	73/36	88/40	55/70	84/58
	N			150	131	131	127	97	109
ATXN2 (29)	C				5	4	4	5	4
	OO				22/38	23/53	24/43	22/104	23/64
	N				196	181	190	130	170
ATXN3 (43)	C					8	4	26	6
	OO					35/49	39/42	17/83	37/62
	N					169	176	135	156
ATXN7 (57)	C						18	20	9
	OO						39/28	37/89	48/59
	N						176	115	145
CACNA1A (47)	C							8	3
	OO							38/101	43/65
	N							114	148
HTT (109)	C								18
	OO								91/50
	N								102

Significant comparisons are highlighted in bold (Fisher's exact test; $p < 0.05$). Significant comparisons where the number of common interactors is lower than expected by chance are in italics, and the significant comparison where the number of common interactors is higher than expected are in bold. The number of proteins that interact with at least two reference proteins is indicated in parentheses

*Status: *C* common, *OO* only in one, *N* in none

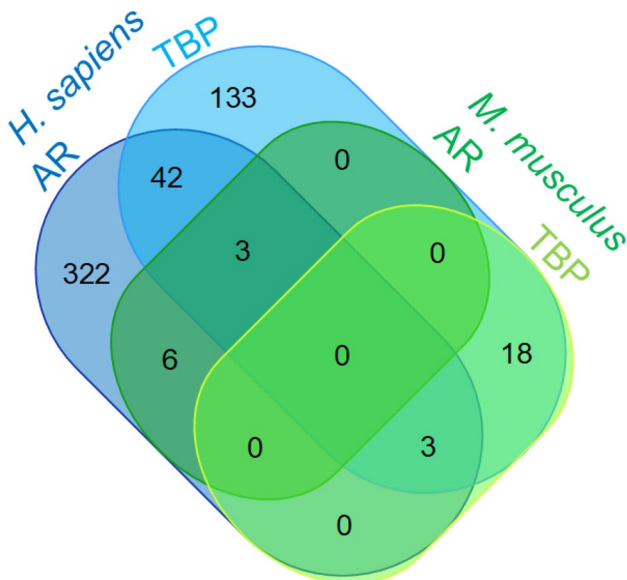


Fig. 7 *H. sapiens* (in blue) and *M. musculus* (in green) AR and TBP interactors

be transcription factors [27, 29] and these could have more interactors than the remaining protein categories. Therefore, we also addressed whether the proteins that function as transcription factors have more interactions than those that (1) are not transcription factors and (2) have, or not, a polyQ region. When this is taken into account, a clear effect of the polyQ on the number of interactors is observed in humans only. Therefore, we cannot exclude the effect of the polyQ region on the large number of common interactors between AR and TBP in humans, despite the surprising result obtained for the other species.

4 Conclusions

This paper has presented EvoPPI, an open-source web application tool that enables users to compare the interactions of a protein across interactomes from the same or different species. To compare interactomes from different species, EvoPPI uses a versatile BLAST search approach, which, we

believe, is a distinctive feature of EvoPPI, since comparable tools only identify orthologs with the same name across species. The current version of EvoPPI also includes support user registration, allowing users to retain their query results for future management.

We have also shown the use of EvoPPI to compare the interactomes of the 9 human polyQ disease genes (those proteins that, when the polyQ tract is expanded, cause neurodegenerative disorders) in 14 datasets. Although polyQ genes show a large number of protein interactions, we found only a small set (15) that are common to at least four of these polyQ disease genes. Of these 15 proteins, 40% are involved in ubiquitin protein ligase-binding function. Ubiquitin/proteasome system dysfunction has been suggested in a range of polyglutamine neurodegenerative diseases [31–34]. Using the unique EvoPPI feature *Compare different species*, the comparisons of the human and mouse AR and TBP interactomes revealed a significant excess of common proteins. In humans, and for AR and TBP only, we cannot confidently discard the polyQ region as the cause of the observed excess of common interactions. For the other seven polyQ disease proteins, no excess of common interactors was observed.

The current development of EvoPPI includes, but it is not limited to: (1) the creation of a management interface to enable users to include new interactomes and species in the EvoPPI database, (2) the addition of new data visualization and analysis options, and (3) an improvement of the distinct species comparison algorithm to make it more complete and efficient.

Acknowledgements This article is a result of the project Norte-01-0145-FEDER-000008—Porto Neurosciences and Neurologic Disease Research Initiative at I3S, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). Sara Rocha is also supported by this project. H. López-Fernández is supported by a post-doctoral fellowship from Xunta de Galicia (ED481B 2016/068-0). SING group thanks Centro de Investigación, Transferencia e Innovación (CITI) from University of Vigo for hosting its IT infrastructure. Financial support from the Xunta de Galicia (Centro singular de investigación de Galicia accreditation 2016–2019) and the European Union (European Regional Development Fund—ERDF), is gratefully acknowledged.

References

1. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169
2. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH (2017) HIP-PIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res* 45:D408–D414
3. Cusick ME, Klitgord N, Vidal M, Hill DE (2005) Interactome: gateway into systems biology. *Hum Mol Genet* 14:R171–R181
4. Chiti F, Dobson CM (2017) Protein misfolding, amyloid formation, and human disease: a summary of progress over the last decade. *Annu Rev Biochem* 86:27–68
5. Cescatti M, Saverioni D, Capellari S, Tagliavini F, Kitamoto T, Ironside J, Giese A, Parchi P (2016) Analysis of conformational stability of abnormal prion protein aggregates across the spectrum of Creutzfeldt–Jakob disease prions. *J Virol* 90:6244–6254
6. Peng X, Wang J, Peng W, Wu F-X, Pan Y (2017) Protein–protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 18:798–819
7. Folador E, de Oliveira Junior A, Tiwari S, Jamal S, Ferreira R, Barh D, Ghosh P, Silva A, Azevedo V (2015) In silico protein–protein interactions: avoiding data and method biases over sensitivity and specificity. *Curr Protein Pept Sci* 16:689–700
8. Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz B-J, Dolinski K, Tyers M (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res* 45:D369–D379
9. Stark C (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34:D535–D539
10. Simonis N, Rual J-F, Carvunis A-R, Tasan M, Lemmens I, Hirozane-Kishikawa T, Hao T, Sahalie JM, Venkatesan K, Gebreab F, Cevik S, Klitgord N, Fan C, Braun P, Li N, Ayivi-Guedehoussou N, Dann E, Bertin N, Szeto D, Dricot A, Yildirim MA, Lin C, de Smet A-S, Kao H-L, Simon C, Smolyar A, Ahn JS, Tewari M, Boxem M, Milstein S, Yu H, Dreze M, Vandenhoute J, Gunsalus KC, Cusick ME, Hill DE, Tavernier J, Roth FP, Vidal M (2009) Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. *Nat Methods* 6:47–54
11. Rolland T, Taşan M, Charloreaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis A-R, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruysinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejada AO, Trigg SA, Twizere J-C, Vega K, Walsh J, Cusick ME, Xia Y, Barabási A-L, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M (2014) A proteome-scale map of the human interactome network. *Cell* 159:1212–1226
12. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, Svrikapa N, Hirozane-Kishikawa T, Rietman E, Yang X, Sahalie J, Salehi-Ashtiani K, Hao T, Cusick ME, Hill DE, Roth FP, Braun P, Vidal M (2011) Next-generation sequencing to generate interactome datasets. *Nat Methods* 8:478–480
13. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamasos E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhoute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437:1173–1178
14. Venkatesan K, Rual J-F, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh K-I, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet A-S, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabási A-L, Vidal M (2009) An empirical framework for binary interactome mapping. *Nat Methods* 6:83–90
15. Murali T, Pacifico S, Yu J, Guest S, Roberts GG, Finley RL (2011) DroID 2011: a comprehensive, integrated resource for protein,

- transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res* 39:D736–D743
16. Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ (2016) FlyBase Consortium: FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Res* 44:D786–D792
 17. López Y, Nakai K, Patil A (2015) HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. *Database* 2015:bav117
 18. Persico M, Ceol A, Gavrilu C, Hoffmann R, Florio A, Cesareni G (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinform* 6:S21
 19. Meyer MJ, Das J, Wang X, Yu H (2013) INstruct: a database of high-quality 3D structurally resolved protein interactome networks. *Bioinformatics* 29:1577–1579
 20. Mosca R, Céol A, Aloy P (2013) Interactome3D: adding structural details to protein networks. *Nat Methods* 10:47–53
 21. Calderone A, Castagnoli L, Cesareni G (2013) mentha: a resource for browsing integrated protein–interaction networks. *Nat Methods* 10:690–691
 22. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonico E, Castagnoli L, Cesareni G (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40:D857–D861
 23. Cowley MJ, Pinese M, Kassahn KS, Waddell N, Pearson JV, Grimmond SM, Biankin AV, Hautaniemi S, Wu J (2012) PINA v2.0: mining interactome modules. *Nucleic Acids Res* 40:D862–D865
 24. Mendivil Ramos O, Ferrier DEK (2012) Mechanisms of gene duplication and translocation and progress towards understanding their relative contributions to animal genome evolution. *Int J Evol Biol* 2012:1–10
 25. Fan H-C, Ho L-I, Chi C-S, Chen S-J, Peng G-S, Chan T-M, Lin S-Z, Harn H-J (2014) Polyglutamine (PolyQ) diseases: genetics to treatments. *Cell Transpl* 23:441–458
 26. Fielding RT (2000) Architectural styles and the design of network-based software architectures. University of California, Irvine
 27. Schaefer MH, Wanker EE, Andrade-Navarro MA (2012) Evolution and function of CAG/polyglutamine repeats in protein–protein interaction networks. *Nucleic Acids Res* 40:4273–4287
 28. Petrakis S, Schaefer MH, Wanker EE, Andrade-Navarro MA (2013) Aggregation of polyQ-extended proteins is promoted by interaction with their natural coiled-coil partners. *Insights Perspect BioEssays* 35:503–507
 29. Fiumara F, Fioriti L, Kandel ER, Hendrickson WA (2010) Essential role of coiled coils for aggregation and activity of Q/N-rich prions and polyQ proteins. *Cell* 143:1121–1135
 30. Butland SL, Devon RS, Huang Y, Mead C-L, Meynert AM, Neal SJ, Lee S, Wilkinson A, Yang GS, Yuen MM, Hayden MR, Holt RA, Leavitt BR, Ouellette BF (2007) CAG-encoded polyglutamine length polymorphism in the human genome. *BMC Genom* 8:126
 31. Nath SR, Lieberman AP (2017) The ubiquitination, disaggregation and proteasomal degradation machineries in polyglutamine disease. *Front Mol Neurosci* 10:78
 32. Pratt WB, Gestwicki JE, Osawa Y, Lieberman AP (2015) Targeting Hsp90/Hsp70-based protein quality control for treatment of adult onset neurodegenerative diseases. *Annu Rev Pharmacol Toxicol* 55:353–371
 33. Rusmini P, Crippa V, Cristofani R, Rinaldi C, Cicardi ME, Galbiati M, Carra S, Malik B, Greensmith L, Poletti A (2016) The role of the protein quality control system in SBMA. *J Mol Neurosci* 58:348–364
 34. Ciechanover A, Brundin P (2003) The ubiquitin proteasome system in neurodegenerative diseases. *Neuron* 40:427–446

Affiliations

Noé Vázquez^{1,2} · Sara Rocha^{3,4} · Hugo López-Fernández^{1,2,3,4,5}  · André Torres^{3,4} · Rui Camacho⁶ · Florentino Fdez-Riverola^{1,2,5} · Jorge Vieira^{3,4} · Cristina P. Vieira^{3,4} · Miguel Reboiro-Jato^{1,2,5}

¹ ESEI-Escuela Superior de Ingeniería Informática, Universidad de Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain

² Centro de Investigaciones Biomédicas (Centro Singular de Investigación de Galicia), Vigo, Spain

³ Instituto de Investigação e Inovação em Saúde (I3S), Universidade do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

⁴ Instituto de Biologia Molecular e Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal

⁵ SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain

⁶ LIAAD and DEI and Faculdade de Engenharia, Universidade do Porto, Porto, Portugal