OXFORD UNIVERSITY PRESS | Bioinformatics

PlanNET: Homology-based predicted interactome for multiple planarian transcriptomes

Journal:	*Bioinformatics*
Manuscript ID	Draft
Category:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Castillo-Lara, Sergio; Universitat de Barcelona, Facultat de Biologia, Av. Diagonal 643, Genètica, Microbiologia i Estadística Abril, Josep; Universitat de Barcelona, Facultat de Biologia, Av. Diagonal 643, Genètica, Microbiologia i Estadística
Keywords:	Annotation, Biological networks, Data integration, Machine learning, Protein-protein interaction, Transcriptomics

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Databases and ontologies

PlanNET: Homology-based predicted interactome for multiple planarian transcriptomes

S. Castillo-Lara [1], J.F. Abril [1,*]

[1] Computational Genomics Lab, Genetics, Microbiology & Statistics Dept.,Universitat de Barcelona; Institut de Biomedicina (IBUB) Barcelona, Catalonia, Spain.

* To whom correspondence should be addressed.

Abstract

Motivation: Planarians are emerging as a model organism to study regeneration in animals. However, the little available data of protein-protein interactions hinders the advances in understanding the mechanisms underlying its regenerating capabilities.
Results: We have developed a protocol to predict protein-protein interactions using sequence homology data and a reference Human interactome. This methodology was applied on ten *Schmidtea mediterranea* transcriptomic sequence datasets. Then, using Neo4j as our database manager, we developed PlanNET, a web application to explore the multiplicity of networks and the associated sequence annotations. By mapping RNA-seq expression experiments onto the predicted networks, and allowing a transcript-centric exploration of the planarian interactome, we provide researchers with a useful tool to analyse possible pathways and to design new experiments, as well as a reproducible methodology to predict, store, and explore protein interaction networks for non-model organisms.
Availability: The web application PlanNET is available at `https://compgen.bio.ub.edu/PlanNET`. The source code used is available at `https://github.com/scastlara/PlanNET`.
Contact: jabril@ub.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The freshwater planarian *Schmidtea mediterranea*, a platyhelminth of the class *Turbellaria*, has become a model for studying regeneration in animals due to its ability to regenerate its whole body even from small parts of it. Planarians only have one cell type able to divide by mitosis, named neoblasts, which are responsible for the extraordinary regeneration capabilities of these organisms (Wagner and Wang, 2011).

In recent years, several studies have been performed in order to unravel the molecular mechanisms of planarian regeneration, as well as its regulation (for instance: Cebrià 2007; Scimone *et al.* 2010; Fernandez-Taboada *et al.* 2010). Additionally, different *high-throughput* RNA-seq experiments have been carried out; up to nine of those transcriptomes are publicly available for *S. mediterranea* alone (Abril *et al.*, 2010; Adamidi *et al.*, 2011; Blythe *et al.*, 2010; Rouhana *et al.*, 2012; Galloni, 2012; Kao *et al.*, 2013; Labbé *et al.*, 2012; Resch *et al.*, 2012; Sandmann *et al.*, 2011; Solana *et al.*, 2012), and more datasets are coming for this and related species (Brandl *et al.*, 2016).

Gene or protein expression analyses take into account significant statistical differences between two or more experimental conditions; however, the large amount of collected data and the fact that this data usually refers to specific proteins or transcripts can lead to key functional elements to remain hidden. Approaches based in systems biology can help to unravel the importance of the different proteins in particular functional processes, as to help to identify similarities between different protein interactions networks. Those techniques will pinpoint missing components of the network (relative to networks from different species like humans) that may reveal driver components of planarian-specific processes such as regeneration. Furthermore, it is possible that those approaches will also suggest homologous functional candidates to test in planarians as an *in vivo* model. Cross-referencing pathways information with genome and transcriptome data may also be useful for researchers, facilitating the link to the functional annotation over the sequences and cis-regulatory elements around the genic *loci*.

Instead of studying and analyzing individual genes or proteins, focusing on the environment of such elements where those components play their roles may reveal interesting insights. Molecular medicine based

on gene and protein networks has been expanding rapidly, and has shown that most disease-causing genes often work together, either forming protein complexes or participating in the same signalling pathways.

Several approaches have been developed in order to infer protein interactions networks from different sources. Sequence homology can be used to predict interactions that have been conserved between species, and the information about these protein interactions can be transferred from one species to another using different approaches (Garcia-Garcia *et al.*, 2012; Murakami and Mizuguchi, 2014; Schuette *et al.*, 2015). In the context of planarians, Lobo and levin developed a method to infer regulatory networks from morphological phenotypes distilled from genetic, surgical and pharmacological experiments (Lobo and Levin, 2015). However, this approach is limited by the currently available phenotypic data on planarians described in the literature; and, although the amount of data collected for this organism is increasing, other approaches based on high-throughput experiment results and large-scale sequence analyses to predict planarian protein-protein interactions will be very useful to the developmental and regeneration research community.

Linking a predicted planarian interactome with a human network may not only provide a useful tool for researchers in order to associate planarian genes with certain cellular functions, but it may also provide a link between planarian regeneration and human molecular pathways.

2 Methods

2.1 Summary of the protocol

A protocol based on sequence homology was developed to infer possible *interolog* relationships between proteins of one arbitrary species and human. In this work, we predicted interactions for ten *Schmidtea mediterranea* transcriptomes (Supplementary Fig.1). The method searched for human homologs to a set of transcripts of the desired species through BLAST searches (Altschul *et al.*, 1990), PFAM domain meta-alignments (Punta *et al.*, 2011), and EggNOG alignments (Huerta-Cepas *et al.*, 2016). Then, a set of features was computed for each possible pair of transcripts, using information from 3did (Mosca *et al.*, 2014), Gene Ontology (GO; Carbon *et al.* 2009), and a human interactome graph. The protocol was first applied to *Drosophila melanogaster*'s transcript sequences; then a random forest classifier was built using this data.

The program TransPipe was implemented in order to automate the whole procedure, taking as input a FASTA file with the *S. mediterranea* transcripts, a hidden Markov model domain database, a FASTA with human sequences and an EggNOG hidden Markov model database. The program also allows to adjust the E-value cutoff for each of the alignment methods independently, as well as providing several plots generated using the R module *ggplot* (Wickham, 2009) to visualize the results. The source code is available from `https://compgen.bio.ub.edu/PlanNET/downloads`, alongside the install information and the required dependencies. The program is distributed under the free software GNU 2 license, but users should register first.

2.2 Datasets

2.2.1 Sequences and hidden Markov models

With the aim to have a sequence assigned to each of the HUGO Gene Nomenclature Comittee (HGNC) symbols (Gray *et al.*, 2015), a list of identifiers and synonyms was downloaded from that project website. One set of human sequences was built using three databases: SwissProt (version 2014/09) (Wasmuth and Lima, 2016), TrEMBL (version 2014/09), and ENSEMBL (gene build 79, GRCh38.p2, Yates *et al.* 2016).

The mapping of HGNC identifiers against human sequences was done sequentially. First, priority was given to Swissprot sequences, followed by ENSEMBL and finally TrEMBL sequences. Each sequence was assigned to a specific HGNC symbol using the aforementioned synonyms table, looking for sequences in the next database only if a symbol remained unassigned. This constitued the H-Prot dataset.

The PFAM domains were downloaded from the PFAM site, version 27.0; and the EggNOGs hidden Markov models, animals meNOG version 4.0, from the database website. The *Drosophila melanogaster* transcript sequences to train the random forest classifier were downloaded from FlyBase release r5.56 (Gramates *et al.*, 2017).

We predicted interactions over 10 planarian transcripts datasets: Adamidi (Adamidi *et al.*, 2011), Blythe (Blythe *et al.*, 2010), Consolidated (Kao *et al.*, 2013), Cthulhu (kindly provided by Kerstin Bartscherer lab), Dresden (Brandl *et al.*, 2016), Graveley (Resch *et al.*, 2012), Illuminaplus (Sandmann *et al.*, 2011), Newmark (Rouhana *et al.*, 2012), Pearson (Labbé *et al.*, 2012) and Smed454 (Abril *et al.*, 2010).

2.2.2 Protein-protein interactions

The human protein-protein interactions dataset was retrieved from BioGRID (version 3.4.133, Stark *et al.* 2006) and STRING (version 10, Von Mering *et al.* 2003). All the nodes were renamed to HGNC symbols when possible, using the HGNC synonyms table, and when no synonym was found, the node remained as an ENSEMBL protein (ENSP) identifier. This whole human gene/protein network included 26,934 nodes and 794,052 edges.

Drosophila melanogaster's protein-protein interactions were downloaded from DroiD FlyBase curated PPI dataset (version 2015_12, Yu *et al.* 2008).

2.3 Homology prediction

The transcript sequences were aligned to the H-Prot dataset using BLASTX and TBLASTN, with an E-value cutoff of 10^{-10} in both cases. From the resulting alignments the *best reciprocal hits* were selected. In order to simplify the whole protocol, we selected the translated longest open reading frame (ORF) for each of all the transcript sequences. These ORF were used for the two following procedures.

The alignment to the EggNOG hidden markov models were performed using *hmmsearch* (Eddy, 1998), with an E-value cutoff of 10^{-10}. We have chosen the subset meNOG (version 4.1), restricting the dataset to only those domains that contained a human protein with an HGNC identifier. The program *hmmsearch* was used in order to annotate the PFAM domains on the transcript sequences, using an E-value cutoff of 10^{-10} and the hidden markov model database of PFAM-A, release 27.0.

The redundancy of the annotation of the domains over the transcripts was reduced by joining several consecutive domains. The conditions used for that merge were the following:

1. Both domains should be equal and consecutive.
2. Both domains annotated over the ORF should represent different regions of the domain. In order to decide if this condition was met, the overlap between both annotations had to be less than 25% of the real length of the PFAM domain.
3. The distance between the domains over the ORF had to be equal or less than the real length of the domain that is not annotated over the transcript, plus a 25% of the total length of the domain.

Once the PFAM domains were annotated, each transcript sequence and each protein of the H-Prot dataset was transformed into a meta-sequence where the annotated domains were concatenated, producing a string of domain symbols suitable for a meta-alignment. Those constructs were then aligned using the *Needleman-Wunsch* algorithm, with a *match* value of +30, a *missmatch* value of -30, and a *gap* value of -5. The *match* score was also adjusted to the percentage of the domain annotated on the transcript sequence. Best reciprocal hits were also selected.

The best homologous human protein was selected for each transcript using the following criteria:

1. If a protein is a unique best reciprocal hit in the EggNOG alignment, set it as the best homolog for that particular transcript.
2. Contrarily, if a unique protein has the largest number of supporting evidences from all the different methods, select it.
3. Otherwise, if a unique sequence is the best hit in the EggNOG alignment (lower E-value), set it as an homolog.
4. Or then, if only a sequence is the best BLAST hit (lower E-value), select it.
5. Else, select the best scoring hit in the PFAM domain meta-alignment.
6. If no condition is met, the contig is discarded.

These decision rules were established because of the EggNOG aligment was set to be more reliable than the others, given that it uses hidden Markov profiles instead of similarity searches, and also given that we had assesed the performance of each method separately.

2.4 Prediction of interactions

A set of 19 features was computed for each possible pair of transcripts with at least one human homolog:

1. **Path length**. The shortest path between the homologous proteins in the human interactome was computed. If no path was found, a value of -1 was assigned. Self-interactions (those pairs with a shortest path of 0) were removed. In order to speed up the prediction, all the shortest paths between all the human proteins were pre-computed using the python module *graph_tool* (Peixoto, 2014).
2. **Domain interaction score**. This score is the number of all the PFAM domain pairs found in the transcripts using hmmsearch ($E\text{-}value \leq 10^{-10}$) that are annotated as interacting in the 3did database.
3. **Gene ontology normalized term overlap** (NTO) between the homologous proteins (Mistry and Pavlidis, 2008). This GO similarity measure was chosen because of its simplicity and the speed to compute it compared to other similarity scores. For each pair of transcripts and each of the GO domains ("molecular function", "cellular component" and "biological process") all the parents in the gene ontology graph for the annotated terms of the two homologous proteins were retrieved. Then, the overlap of these two sets (normalized over the minimum set) was computed. This feature takes values between 0 (no GO term overlap) and 1 (all the annotated GO terms are the same).
4. **Alignment measures**. Several of the alignment measures reported by BLAST, hmmsearch, and the meta-alignment, were used to train the classifier: BLAST and EggNOG E-values, BLAST query coverage and PFAM meta-alignment score. Finally, a boolean variable for each of the alignments and each of the three methods was defined. This variable was set to "True" if the transcript-human sequence pairs were best reciprocal hits, and "False" otherwise.

To build the random forest classifier, a training set of 11,595 *Drosophila melanogaster* interacting pairs was retrieved from DroiD (Flybase curated dataset), and 853,023 random pairs filtered against the DroiD pairs constituted the non-interacting protein pairs. All the features were manually discretized into fixed ranges specific to each variable. We used the R module `randomForest` (version 4.6-10, Liaw and Wiener 2002), setting the number of trees to 1,000 and downsampling the non-interacting pairs so that for building each tree the ratio between non-interacting and interacting pairs was 5:1.

For all the performance validation measures the out-of-bag votes reported by the module were used. A cutoff of 0.6 votes was set to decide if a pair is interacting. In order to reduce the search space of interologs, the program TransPipe only considers those pairs with a *path length* $<= 2$, and removes all the pairs that are not connected on the human interactome (*path length* $= -1$).

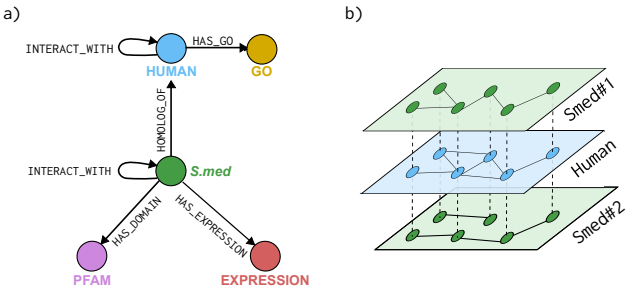


Fig. 1. Neo4j database core schema used to store the predicted interactomes, along with the reference human interactome and the sequence annotations for the planarian datasets. a) Diagram summarizing types of relationships and labels used in the database. b) Example of two planarian interactomes (Smed #1 and Smed #2) connected through HOMOLOG_OF relationships (dotted lines in the figure) to the Human interactome. This database schema allows us to incorporate any number of predicted interactomes in the database, connect them through the Human protein-protein interactions network, and relate similar nodes.

2.5 Neo4j database

All the predicted interactomes, as well as the annotations of the different planarian transcripts were stored in a Neo4j database, version 3.1.1 (Robinson *et al.*, 2013). The choice of a graph database instead of a relational database such as MySQL was driven by the nature of the data itself: an interactome can be easily stored as a series of nodes and connections. Traversing the graph can then be done in a very time-efficient way, and operations such as obtaining the transcripts/proteins connected to a given node through an arbitrary number of intermediate connections is trivial.

In addition, having the interactomes stored as a series of nodes and connection not only allows us to perform queries faster, but gives us the ability to use different types of connections that are associated to different meanings. All the homology relationships between planarian contigs and human sequences were also stored as connections between nodes. This allows us to map the predicted interactomes over the human protein-protein interactions. Thanks to this, we are able to search planarian interactions using human protein symbols, as well as comparing subgraphs and pathways across all the different predicted interactomes. The PFAM alignments were also stored in this graph, as well as the Gene Ontology annotations, giving us the ability to, for example, look for pathways were the genes involved have a particular GO code or a PFAM domain.

Finally, gene expression information from a Digital Gene Expression (DGE) experiment (Rodríguez-esteban *et al.*, 2015) was also stored in the Neo4j database. As can be seen in Fig.1, we used up to five different types of connections, each one with a set of attributes storing the relevant features of that relationship: for example, *HOMOLOG_OF* relationships have attributes such as the BLAST E-value and the sequence alignment coverage.

3 Results

3.1 Performance of the predictor

The performance of the classification of contig pairs as interacting or non-interacting was evaluated using the following measures: precision, sensitivity, specificity, out-of-bag error rate (OOB error), and area under the curve of the receiver operating characteristic (ROC); see Fig.2. The area under the curve calculated using different votes cutoffs was 0.82. In order to improve all the performance measures, but at the same time, to give the user the freedom to choose or focus on more or less confident

Table 1. Results of the prediction of protein-protein interactions for ten S. mediterranea transcriptome datasets. The "Average degree" describes the connectivity of each graph as $interactions/nodes$. The "Percentage of Plen1" corresponds to the fraction of interactions in each network that are also found in the reference human interactome.

Transcriptome	Total contigs	Contigs with homolog	Human homologs	Contigs in interactome	Number of interactions	Average degree	Percentage of Plen1
Adamidi	18,547	9,478	5,187	4,903	32,626	6.657	36.8%
Blythe	24,008	10,930	5,564	5,929	32,892	5.548	34.7%
Consolidated	23,545	12,775	5,809	7,098	53,609	7.553	30.8%
Cthulhu	117,763	10,793	5,741	5,411	36,049	6.662	41.9%
Dresden	40,480	14,626	5,889	7,713	68,805	8.921	30.4%
Graveley	19,503	8,475	4,329	3,796	14,254	3.755	30.6%
Illuminaplus	28,926	10,090	5,182	5,263	29,574	5.619	36.8%
Newmark	53,898	20,665	6,359	11,188	100,138	8.950	39.0%
Pearson	25,889	10,465	5,656	5,176	30,538	5.810	31.0%
Smed454	46,602	14,720	4,711	8,734	112,512	4.061	57.4%
Mean		12,302	5,441	6,521	51,100	6,354	
Std.Dev		3,412	568	2,085	31,047	1,687	

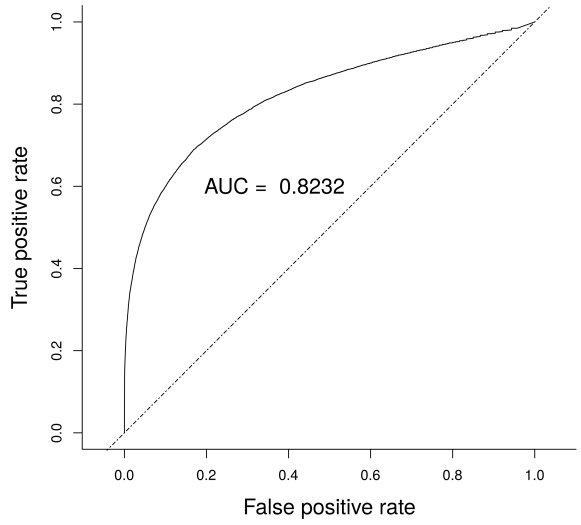


Fig. 2. Receiver operating characteristic (ROC) curve of the random forest classifier. Using Drosophila melanogaster's protein-protein interactions data downloaded from DroiD and a Human interactome as reference, we built a random forest classifier to predict interactions from transcript alignments to human proteins. The ROC curve was built using the Out-of-Bag votes of the random forest for the Drosophila interacting and non-interacting transcript pairs.

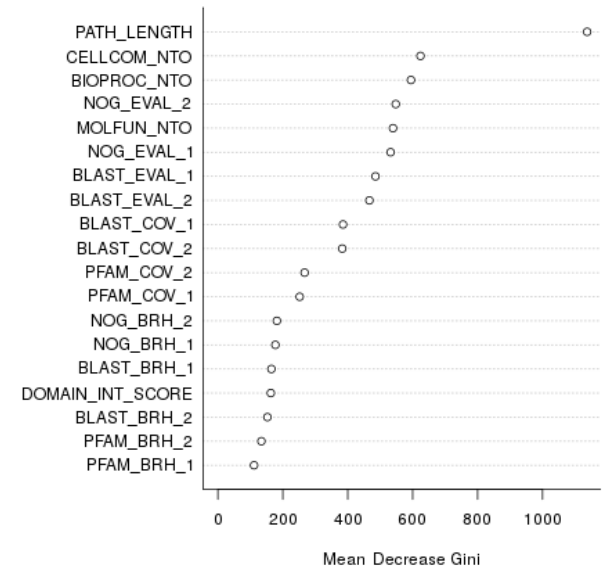


Fig. 3. Variable importance of the features used by the random forest classifier to predict interacting protein pairs.

predictions, we decided to use a votes cut-off of 0.6. By using this cut-off, we obtained a precision of 0.35, a sensitivity of 0.34, a specificity of 0.99 and an OOB error of 2.67%.

We analyzed the relative importance of the 19 features used for the classification of each contig pair using the "Gini importance" index provided by the R `randomForest` package, as it has been shown to be useful for feature selection in classification problems (Menze *et al.*, 2009). The most useful feature to predict protein-protein interactions resulted to be the distance between the homologous proteins in the human interactome graph, defined as "PATH_LENGTH" in Fig.3. Although other features, such as the Gene Ontology NTOs, are also relevant for the classification, there is a clear separation in terms of importance between path length and all the other features.

3.2 Prediction of planarian interactions

The contigs of the 10 planarian transcriptomes were aligned to the H-Prot dataset using BLAST searches, HMMER alignments to EggNOG models, and PFAM meta-alignments. We selected the best hit for each contig and we computed the 19 features for each possible pair of contigs required for the random forest classifier. Although each planarian contig had only one selected homolog, several human proteins had more than one homologous planarian contig in the selected pairs used for the prediction. However, among the selected contig pairs, most human proteins had between one and two homologous contigs for all the datasets (Supplementary Fig.2). Two human proteins (ACTB and ACTG1) had 6,117 and 4,222 homologous contigs in the Smed454 dataset, possibly due to the fact that the corresponding RNA libraries were unnormalized. Because of limitations of computing power when predicting the interactions, all except

one contig for each of these two human proteins were removed. We selected the contig with the lowest E-value in the EggNOG alignment for each of these cases.

The classifier was used to predict interactions in 10 planarian transcriptomes. As it can be seen in Table 1, the number of contigs with an homolog varies from 9,478 to 20,665, while the number of Human homologs for each dataset shows a way lower variation. The final number of predicted interactions is also highly variable. However, the number of interactions strongly correlates with the initial number of contigs with an homolog in each dataset ($Spearman's\ rho = 0.939$, *p-value* $< 10^{-10}$). The number of contigs in each predicted interactome is also dependant on the initial contig count ($Spearman's\ rho = 0.988$, *p-value* $< 10^{-10}$).

In order to compare the confidence of each prediction, we plotted the distribution of votes of the classifier for each dataset (including the OOB votes for the testing dataset). As can be seen in Supplementary Fig.3, all the planarian interactomes have a very similar distribution of votes, with the votes for the testing dataset being slightly higher. Most predictions fall between 60% of votes and 70% of votes, but there is a big number of predicted interacting pairs with a high percentage of votes in all the datasets. For all the datasets, the proportion of contig pairs with interacting homologs in human (*path length* $= 1$) was less than 1%, while this proportion increased significantly when considering only the available predicted interactions (Table 1). Additional information about both the predictions and the sequence alignments for each dataset is available at the protocol summary page (`https://compgen.bio.ub.edu/PlanNET/datasets`).

3.3 PlanNET web application

In order to explore the predicted interactomes and the sequence annotations of the planarian sequences, we implemented a web application called PlanNET using the python web development framework `Django` and the javascript plugin `cytoscape.js` (Franz *et al.*, 2015). The starting form is divided in four sections that serve as different entry points to the Neo4j database.

GeneSearch provides a text-based search by keywords, thanks to it, the user can look for all the annotated features of the planarian contigs using either planarian contig identifiers or human protein symbols. The latter will retrieve all the *S. mediterranea* contigs of a particular dataset that are homologous to the specified human protein.

We also provide a way to explore the predicted interaction networks utilizing `cytoscape.js` in *NetExplorer*, where the user can search for nodes across the different planarian protein-protein interaction networks, either using contig identifiers, human protein symbols (wildcards allowed), PFAM identifiers, or GO codes.

Thanks to the graph-based database manager Neo4j, traversing the networks to retrieve any subset of them does not have a huge performance impact; we took advantage of this capability to implement *PathwayFinder*. This application looks for all the possible paths between two protein/contigs in the specified interactome, rating all these paths depending on their overall confidence. This score was defined as the mean of the random forest votes for each of its predicted interactions. Just like in *NetExplorer*, users can search by human protein symbols, PFAM identifiers, contig identifiers and GO codes.

For the sake of completeness, we also implemented a BLAST web form to look for contigs on the graphs using sequence homology searches.

4 Discussion

In this work we introduce a tool to predict protein-protein interactions from transcript sequences using sequence alignments and a reference Human interactome. This tool was then used to predict ten different protein interactions networks from ten *S. mediterranea* transcript datasets. As a result, we provide PlanNET, a web application that allows researchers to explore these networks in different ways, as well as to access to all sequence annotations performed in order to predict the interactions.

Given the out-of-bag performance evaluation of the predictor, we conclude that this random forest classifier is useful for inferring interologies between two species, for instance planarian and human. The area under the ROC curve of 0.82 strongly indicates the significant improvement from a random predictor of our tool. The low precision and sensitivity (0.35 and 0.34, respectively) can be attributed to the fact that from all the possible pairs of proteins of a given organism, only a tiny subset of them really interact. It has been described that the protein interaction network of any given species is always very sparse, as the degree distribution of most of them follow the power-law (Barabási and Oltvai, 2004). This fact alone makes it harsh for any predictor to retrieve a large amount of interactions out of those pairs, without retrieving many false positives. However, the developed predictor can be further improved in different ways; for example, introducing new features such as the confidence of the annotated interactions in the reference human network, or adding new reference interactomes from other species.

From all the features used by the classifier, the most important one is the distance of the homologous human genes in the Human protein-protein network (PATH_LENGTH). After that, the most crucial features for the correct classification of the protein pairs were the GO similarities and the EggNOG alignment E-value. Those features paired together, ensure that both proteins have a high sequence similarity to their respective human homologs, that those homologs are known to be in a similar cellular location (*cellular component*) and that they share GO *biological process* and *molecular function* annotations. Thus, our tool not only predicts interactions between putative proteins translated from transcripts, but it essentially clusters these contigs according to their functional similarities (instead of, for example, their location in the genome). The relative importance of the EggNOG E-values may be due to two reasons: firstly, the performed protocol favors EggNOG alignments when selecting best hits for each contig, and secondly, hidden markov models are known to detect more distant homologies than sequence searches such as BLAST (Park *et al.*, 1998).

The different number of interactions predicted for each dataset can be attributed to the different number of homologs found for each one, as the strong correlation between the "*Contigs with an homolog*" and "*Number of interactions*" suggests. Therefore, the different level of fragmentation of the contigs from the transcriptomes could greatly affect the final result of the prediction. This reinforces the importance of building a proper full-length mRNA reference set, especially for *S. mediterranea*, as it has been shown for model organisms like human (Cho *et al.*, 2014).

Using Neo4j as our database backend, we developed a web application, called PlanNET, to explore not only the predicted networks, but the sequence annotations as well. Apart from BLAST searches and simple text searches (provided by the applications *BLAST* and *GeneSearch* at PlanNET, respectively), we have also implemented two additional ways to explore the protein networks. *NetExplorer* (Fig.4), will allow researchers to look for predicted interaction networks using contig IDs, GO codes or PFAM domains as baits. The asynchronous `javascript` searches will allow the users to dinamically compare the networks annotated for the different transcriptomes with respect to the reference interactome, human in this case. The application *PathwayFinder* provides a simple way to look for protein interaction pathways in the predicted networks; specifying a *starting protein*, an *ending protein*, and the length of the desired pathways.

We can also integrate gene expression data from different sources, which has been tested by showing results from a DGE experiment performed by Rodriguez-Esteban *et al*. This data can be projected over the graphs, coloring the nodes according to the expression levels; right

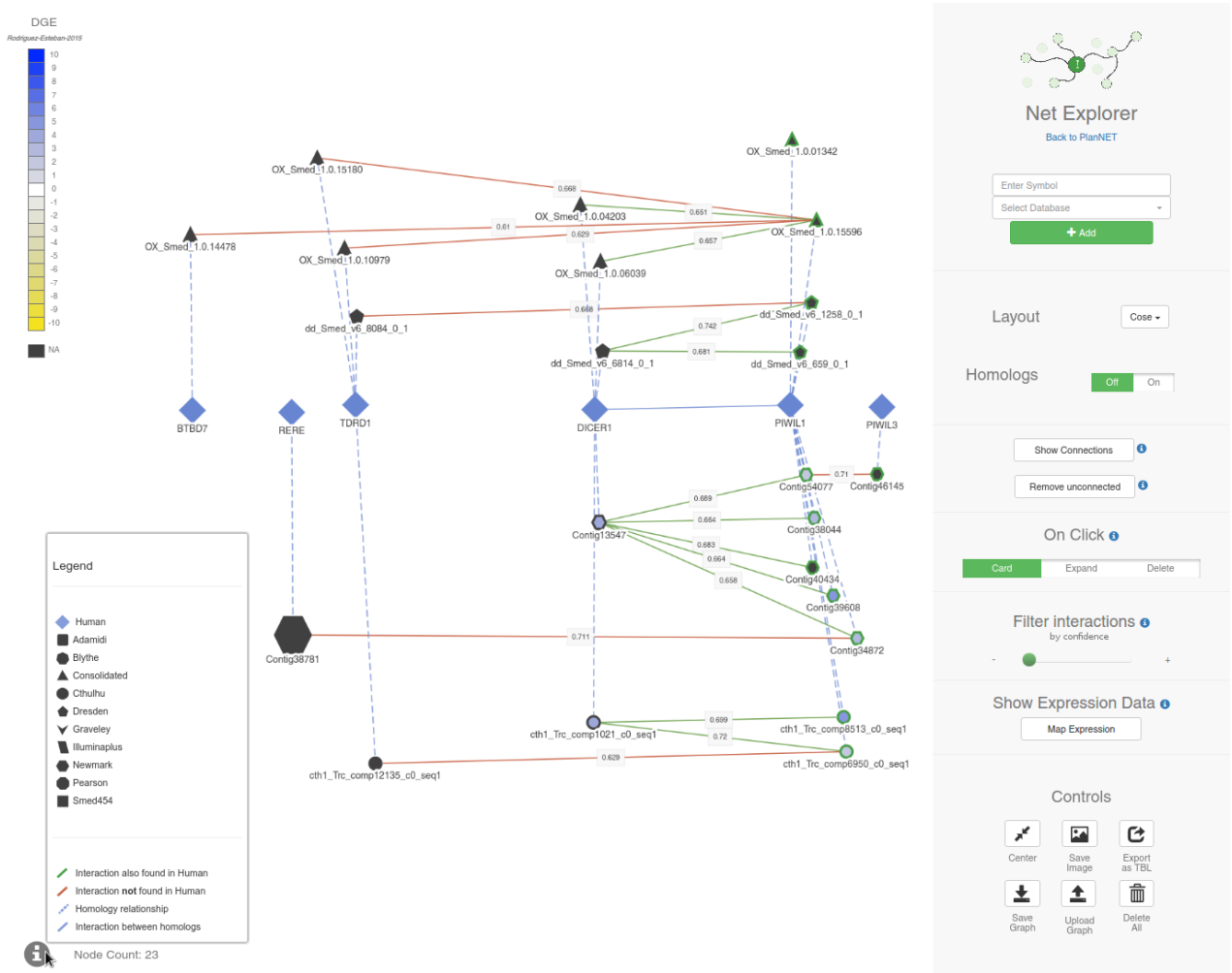


Fig. 4. Screenshot of the PlanNET NetExplorer browser. Each node in the graph represents a protein/contig, the shape of the nodes determines the dataset to which they belong. The size of the nodes depends on the node degree (total number of interactions). The edges color varies depending on the type of relationship (see "Legend" on the lower left corner of this figure). DGE data comparing two samples from the experiment described in Rodríguez-esteban et al. 2015 was projected over this visualization; the color of nodes is based on the expression fold change (scale shown on the upper left corner of this figure). The controls on the right panel allow users to explore the graph further by clicking on nodes, as well as getting information for each contig/protein. Numbers on the edges correspond to the proportion of votes of the random forest classifier, as a measure of the confidence of any given interaction. Users can filter these interactions by confidence value with the slider on the right (under "Filter interactions"). Finally, the interface allows researchers to save and load graphs in `JSON` format.

know, the application allows to color the nodes using expression data from one sample (binning colors by percentils of absolute expression levels) or to compare two samples at the same time from the same experiment (assigning colors in function of $log2(FoldChange)$). We are planning to add further RNA-seq expression data to this tool, which will allow more complex queries to be performed, such as retrieving subnetworks having correlated expression levels across different experiments and samples. Neo4j manager and the Cypher query language make those complex queries simple and fast (Yoon *et al.*, 2017), and they allow to perform them in real time as opposed to pre-computing them.

When designing experiments with the aim to unravel the underlying mechanisms of planarian regeneration, the biological context of any given candidate gene is just as important as its annotations. Thus, a predicted protein-protein interaction network for many planarian transcriptomes will be useful in determining that context from a transcript-centric point of view. Our applications allow researchers to compare the human homologs found in all the transcriptomes, to look for possible interacting proteins, to retrieve sub-networks using Gene Ontology codes and PFAM domains, and to compare expression levels across the transcriptomes. Our approach focuses on the planarian contigs instead of the annotated genes on a reference genome; giving researchers the flexibility to work with any contig as a proper separate entity with its own annotations. When new refined transcriptomes will be made available in the future they will be easily incorporated to PlanNET using our current pipeline, without interfering with all the previous datasets and maintaining the relevance of the application.

At last, in this work we provide one of the firsts practical applications of the database manager Neo4j to store and analyze multiple protein-protein interaction networks. Sequence homologies predicted for the planarian contigs allow us to link the *S. mediterranea* and the human interactome in the database, making quick queries to compare and traverse the networks easy. Our current database design will facilitate the inclusion of genetic interactions, as well as the extension with new interfaces to explore them in the future. In conclusion, this database engine provides a very adaptable framework for storing, modelling and visualizing many networks projected over a reference interactome. This has been crucial to implement a responsive interactive interface, PlanNET, over the extensible interologs network that projects planarian transcriptomes towards human

sequences and vice versa. The analysis pipeline can be applied to any species transcriptomic datasets to map it over a model organism reference interactome, which makes our protocol extensible to a broader range of similar research problems.

Acknowledgements

We are grateful to Emili Saló for his continuous and generous encouragement, as well as for his insights on the planarian molecular biology.

Funding

This work was supported by research grants from Spanish Ministry of Economy (BFU2014-56055P), and from Generalitat de Catalunya (2014SGR687). Sergio Castillo-Lara is fellow of the Catalan Government "AGAUR" (FI-FDR, 2017FI_B_00191).

References

Abril, J. F., Cebrià, F., Rodríguez-esteban, G., Horn, T., Fraguas, S., Calvo, B., Bartscherer, K., and Saló, E. (2010). Smed454 dataset: unravelling the transcriptome of *Schmidtea mediterranea*. *BMC Genomics*, **11**, 731.

Adamidi, C., Wang, Y., Gruen, D., Mastrobuoni, G., You, X., Tolle, D., Dodt, M., Mackowiak, S., Gogol-Doering, A., Oenal, P., Rybak, A., Ross, E., S., A. A., Kempa, S., Dieterich, C., Rajewsky, N., and Chen, W. (2011). De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res.*, **21**(21), 1193–1200.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, **5**(2), 101–113.

Blythe, M. J., Kao, D., Malla, S., Roswell, J., Wilson, R., Evans, D., Jowett, J., Hall, A., Lemay, V., Lam, S., and Aziz Aboobaker, A. (2010). A dual platform approach to transcript discovery for the planarian *Schmidtea mediterranea* to establish RNAseq for stem cell and regeneration biology. *PLOS ONE*.

Brandl, H., Moon, H. K., Vila-Farr??, M., Liu, S. Y., Henry, I., and Rink, J. C. (2016). PlanMine - A mineable resource of planarian biology and biodiversity. *Nucleic Acids Research*, **44**(D1), D764–D773.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Lomax, J., Mungall, C., Hitz, B., Balakrishnan, R., Dolan, M., Wood, V., Hong, E., and Gaudet, P. (2009). AmiGO: Online access to ontology and annotation data. *Bioinformatics*, **25**(2), 288–289.

Cebrià, F. (2007). Regenerating the central nervous system: how easy for planarians!. *Dev. Genes Evol.*, **217**(217), 733–748.

Cho, H., Davis, J., Li, X., Smith, K. S., Battle, A., and Montgomery, S. B. (2014). High-resolution transcriptome analysis with long-read RNA sequencing. *PLoS ONE*, **9**(9).

Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, **14**(14), 755–763.

Fernandez-Taboada, E., Moritz, S., Zeuschner, D., Stehling, M., Schö ler, H. R., Saló, E., and Gentile, L. (2010). *Smed-SmB*, a member of the LSm protein superfamily, is essential for chromatoid body organization and planarian stem cell proliferation. *Development*, **137**(137), 1055–1065.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, **32**(2), 309–311.

Galloni, M. (2012). Global irradiation effects, stem cell genes and rare transcripts in the planarian transcriptome. *Int. J. Dev. Biol.*, **56**, 103–116.

Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., and Oliva, B. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Research*, **40**, 147–151.

Gramates, L. S., Marygold, S. J., Santos, G. d., Urbano, J.-M., Antonazzo, G., Matthews, B. B., Rey, A. J., Tabone, C. J., Crosby, M. A., Emmert, D. B., Falls, K., Goodman, J. L., Hu, Y., Ponting, L., Schroeder, A. J., Strelets, V. B., Thurmond, J., and Zhou, P. (2017). Flybase at 25: looking to the future. *Nucleic Acids Research*, **45**(D1), D663–D671.

Gray, K. a., Daugherty, L. C., Gordon, S. M., Seal, R. L., Wright, M. W., and Bruford, E. a. (2015). Genenames.org: The HGNC resources in 2015. *Nucleic Acids Research*, **43**(D1079-D1085).

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., Von Mering, C., and Bork, P. (2016). EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, **44**(D1), D286–D293.

Kao, D., Felix, D., and Aboobaker, A. (2013). The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. *BMC Genomics*, **14**(14), 797.

Labbé, R. M., Irimia, M., Currie, K. W., Lin, A., Zhu, S. J., Brown, D. D. R., Ross, E. J., Voisin, V., Bader, G. D., Blencowe, B. J., and Pearson, B. J. (2012). A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells*, **30**(30), 1734–1745.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, **2**(3), 18–22.

Lobo, D. and Levin, M. (2015). Inferring Regulatory Networks from Experimental Morphological Phenotypes: A Computational Method Reverse-Engineers Planarian Regeneration. *PLOS Computational Biology*, **11**(6), e1004295.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, **10**(1), 213.

Mistry, M. and Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC bioinformatics*, **9**(1), 327.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, **42**(D1), 374–379.

Murakami, Y. and Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. *BMC bioinformatics*, **15**(1), 213.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of molecular biology*, **284**(4), 1201–1210.

Peixoto, T. P. (2014). The graph-tool python library. *figshare*.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, a., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, a., and Finn, R. D. (2011). The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Resch, A. M., Palakodeti, D., Lu, Y. C., Horowitz, M., and Graveley, B. R. (2012). Transcriptome analysis reveals strain-specific and conserved

stemness genes in *Schmidtea mediterranea. PLOS ONE*, **7**(7).

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O'Reilly Media, Inc.

Rodríguez-esteban, G., González-sastre, A., Rojo-laguna, J. I., and Saló, E. (2015). Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in *Schmidtea mediterranea*. *BMC genomics*, **16**, 361.

Rouhana, L., Vieira, a. P., Roberts-Galbraith, R. H., and Newmark, P. a. (2012). PRMT5 and the role of symmetrical dimethylarginine in chromatoid bodies of planarian stem cells. *Development*, **139**(6), 1083–1094.

Sandmann, T., Vogg, M. C., Owlarn, S., Boutros, M., and Bartscherer, K. (2011). The head-regeneration transcriptome of the planarian *Schmidtea mediterranea. Genome Biol.*, **12**, R76.

Schuette, S., Piatkowski, B., Corley, A., Lang, D., and Geisler, M. (2015). Predicted protein-protein interactions in the moss *Physcomitrella patens*: a new bioinformatic resource. *BMC Bioinformatics*, **16**(1), 89.

Scimone, M. L., Meisel, J., and Reddien, P. W. (2010). The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian *Schmidtea mediterranea. Development*, **137**(137), 1231–1241.

Solana, J., Kao, D., Mihaylova, Y., Jaber-Hijazi, F., Malla, S., Wilson, R., and Aboobajer, A. (2012). Defining the molecular profile of planarian pluripotent stem cells using a combinatorial RNA-seq, RNAi and irradiation approach. *Genome Biol.*, **13**(3), R19.

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.

Wagner, D. E. and Wang, I. E. (2011). Clonogenic neoblasts are pluripotents adult stem cells that underlie planarian regeneration. *Science*, **332**(332), 811–816.

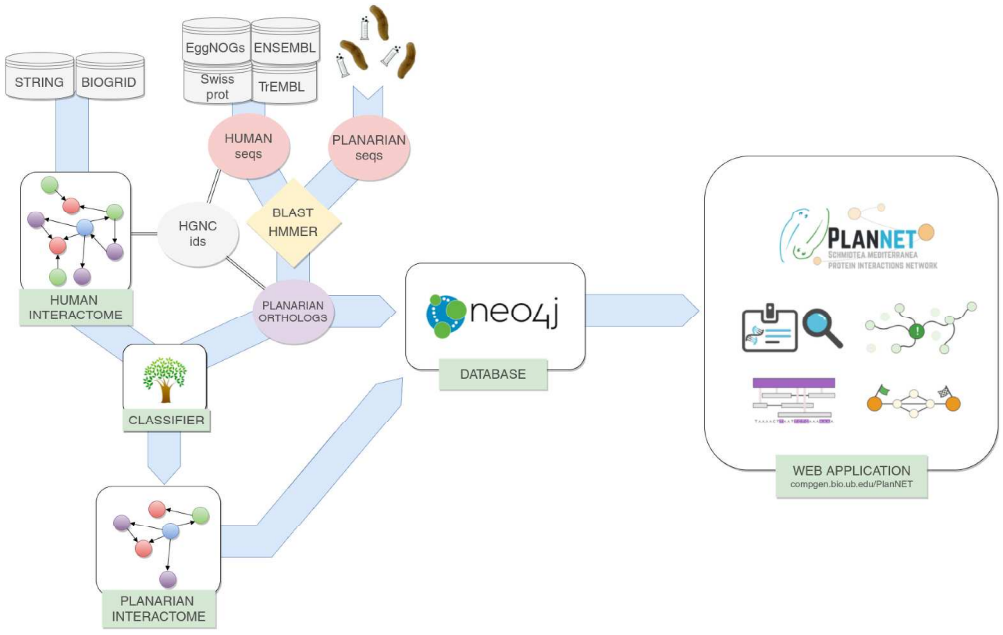
Wasmuth, E. V. and Lima, C. D. (2016). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **45**, 1–12.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., Gir??n, C. G., Gordon, L., Hourlier, T., Hunt, S. E., Janacek, S. H., Johnson, N., Juettemann, T., Keenan, S., Lavidas, I., Martin, F. J., Maurel, T., McLaren, W., Murphy, D. N., Nag, R., Nuhn, M., Parker, A., Patricio, M., Pignatelli, M., Rahtz, M., Riat, H. S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S. P., Zadissa, A., Birney, E., Harrow, J., Muffato, M., Perry, E., Ruffier, M., Spudich, G., Trevanion, S. J., Cunningham, F., Aken, B. L., Zerbino, D. R., and Flicek, P. (2016). Ensembl 2016. *Nucleic Acids Research*, **44**(D1), D710–D716.

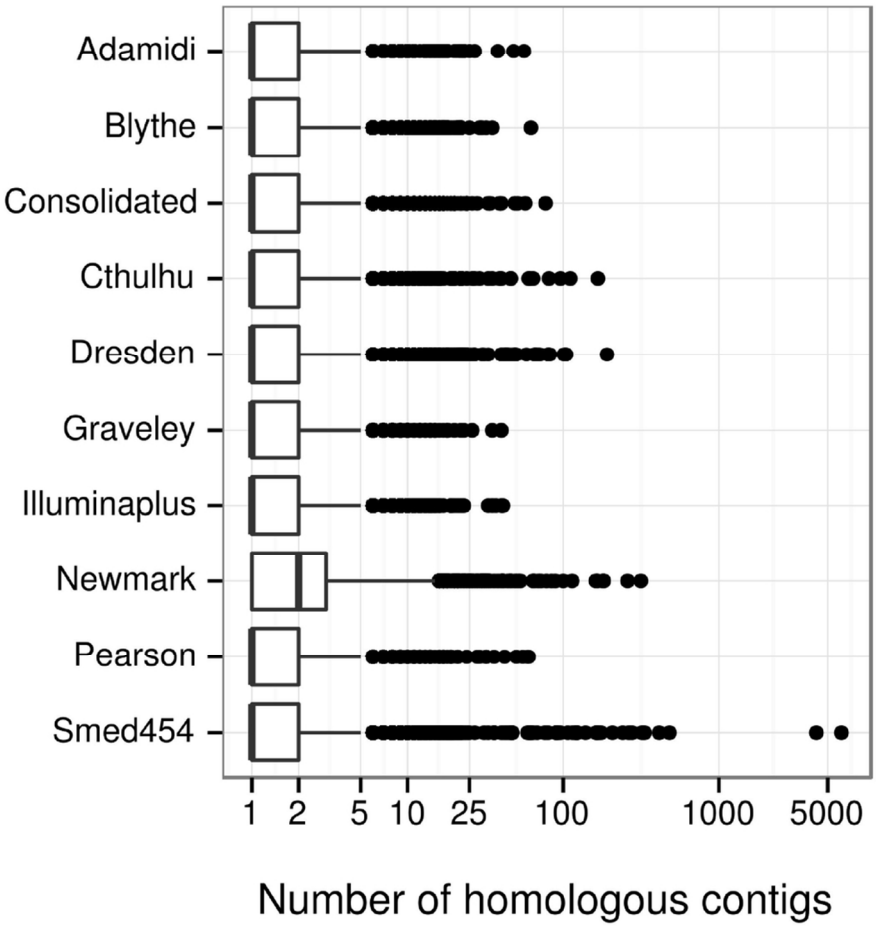
Yoon, B.-H., Kim, S.-K., Kim, S.-Y., Jensen, L., Bork, P., and Felix, V. (2017). Use of Graph Database for the Integration of Heterogeneous Biological Data. *Genomics & Informatics*, **15**(1), 19.

Yu, J., Pacifico, S., Liu, G., and Jr, R. L. F. (2008). DroID: the *Drosophila* Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC genomics*, **9**, 1–9.



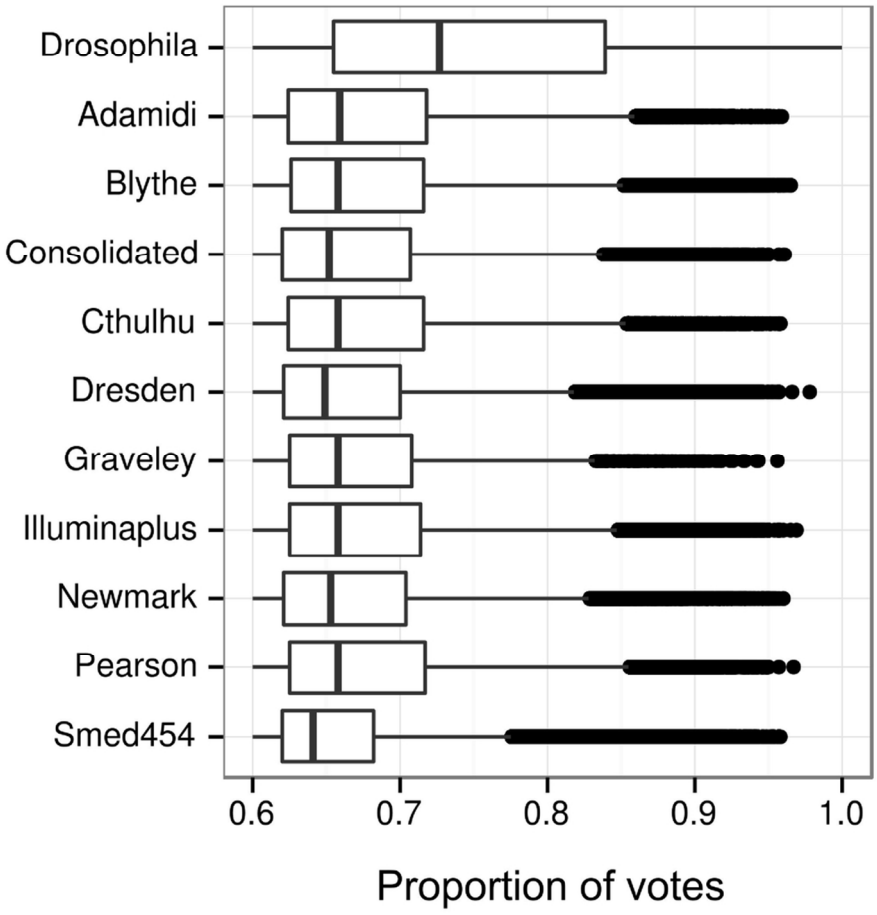
General overview of the protocol used to predict protein-protein interactions on Schmidtea mediterranea transcriptomes. Using a human protein-protein interactions network retrieved from BioGrid and String, we predicted protein interactions for planarian transcripts. First, we searched for homology relationships between human proteins and planarian transcripts using BLAST and HMMER, and then, using a random forest classifier trained with Drosophila melanogaster sequences, we predicted a different interactome over each transcript set. All the information gathered during this process was uploaded to a Neo4j database, and we built a web interface called PlanNET to navigate through those networks and explore the connectivity and the nodes content such as sequence and domain information or projected expression levels.

650x408mm (600 x 600 DPI)



Number of homologous contigs for each human protein with at least one homolog in the analyzed transcriptomes.

101x101mm (300 x 300 DPI)



Proportion of positive votes of random forest classification of protein-protein interactions per dataset. Note that the votes displayed for the training set correspond to the out of bag votes.

101x101mm (300 x 300 DPI)