Genome-wide inference of the *Camponotus floridanus* protein-protein interaction network using homologous mapping and interacting domain profile pairs

Shishir K. Gupta1,2,4, Mugdha Srivastava1,4, Özge Osmanoglu1 & Thomas Dandekar1,3*

Apart from some model organisms, the interactome of most organisms is largely unidentified. High-throughput experimental techniques to determine protein-protein interactions (PPIs) are resource intensive and highly susceptible to noise. Computational methods of PPI determination can accelerate biological discovery by identifying the most promising interacting pairs of proteins and by assessing the reliability of identified PPIs. Here we present a first in-depth study describing a global view of the ant *Camponotus floridanus* interactome. Although several ant genomes have been sequenced in the last eight years, studies exploring and investigating PPIs in ants are lacking. Our study attempts to fill this gap and the presented interactome will also serve as a template for determining PPIs in other ants in future. Our *C. floridanus* interactome covers 51,866 non-redundant PPIs among 6,274 proteins, including 20,544 interactions supported by domain-domain interactions (DDIs), 13,640 interactions supported by DDIs and subcellular localization, and 10,834 high confidence interactions mediated by 3,289 proteins. These interactions involve and cover 30.6% of the entire *C. floridanus* proteome.

In terms of biodiversity and biomass, insects are the most successful animals on earth. They provide major beneficial impacts such as pollination, food source and soil improvement. On the contrary, some insects damage crops and spread deadly diseases as vectors. One of the harmful pests is the ant *Camponotus floridanus* which is widely distributed throughout Florida and the neighboring states1. They hollow the wood softened by moisture and damage the structural integrity of houses by affecting the wood work with their strong mandibles. Besides this, this ant species serves as a good model system to understand host-endosymbiont relationships regarding its bacterial endosymbiont *Blochmannia*2.

The complete genome sequences of *C. floridanus* has revealed the composition of proteins, based mainly on theoretical predictions utilizing their corresponding DNA sequence. We analyzed the transcriptome level evidence of protein existence and re-annotated the *C. floridanus* gene models and proteins3. How these proteins interact is not yet explored, in part due to limited genetic studies in this organism, high cost, time-consuming and labor-intensive nature of experimental methods. Protein interactions are at the core of nearly all biological processes, and knowledge about protein-protein interactions (PPIs) is vital for understanding biological systems. Despite advances in high-throughput experimental methods for detecting PPIs, the interaction networks for even the well-studied experimental model organisms are far from complete4. Nevertheless, high throughput assays typically include false positives PPIs5 which stipulate an enduring need for efficient computational methods to complement existing experimental approaches. In this context, combining the interolog method6 with adding domain information7, gene ontology (GO) annotation8 and cellular localization9,10 yields a graphical representation of the interaction networks, a robust and well-established approach to provide an intuitive vision and useful insights to help and analyze complex relations therein, as indicated by several previous studies in the reconstruction and
understanding of PPIs in various organisms. Here we used domain information, subcellular localization and isoform information to filter the preliminary global PPI network of C. floridanus reconstructed on stringent interolog based criteria. We focus on interactions predicted with high confidence to reduce noise. This conservative approach rejects 79.1% of the preliminary predicted interactions. We then explored the topologically important and evolutionary conserved proteins by analyzing the reconstructed interactome regarding cellular functions.

Results and Discussion
Generating the interactome of ant C. floridanus. PPIs are typically mediated by interactions between domains that are often evolutionary conserved across species and form stable interactions. PPI (protein-protein interaction) maps from experiments on D. melanogaster were collected and augmented by PPI data from the DIP database (Database of Interacting Proteins). This provided a basis for interaction predictions according to interologs from C. floridanus: conserved proteins compared to Drosophila should also be conserved in their interactions (see Materials and methods for details).

Optimally, for such predictions several methods are combined (Fig. 1). We combined the orthology prediction methods InParanoid and OrthoMCL. This did yield a first estimate of the C. floridanus interactome with 6274 nodes and 51866 edges. However, the preliminary ant PPI network could have several false positive interactions acquired from the interologs of template data as shown previously in similar studies, including transfer to curated databases. To reduce false predictions, we counter-checked all our data by domain-domain interactions (DDI). DDI are often used as an approach independent from sequence homology-based methods to predict protein-protein interaction networks and thus strongly reduce the number of false positives. Generally, some of the PPIs are achieved via interactions between short motifs that are often transient interactions. On the other hand, conserved interactions are mediated by conserved interaction domains across species. Moreover, many signals and processes in the cell rely on conserved interacting protein domains. There were 51866 conserved proteins (interologs) and 20544 ant protein-protein interactions that also were associated with DDI pairs, yielding a curated C. floridanus interactome with 4589 nodes and 20544 edges. For final curation of the interactome we used the subcellular localization of ant proteins: interacting proteins have to share the same subcellular localization (summarized in Table 1), predicted interactions between proteins not in the same location were removed. This led to a consolidated ant interactome consisting of 3914 nodes and 13640 edges. The highest proportion of interactions were identified in the cytoplasm following by nucleus and plasma membrane respectively. A closer inspection of the interactions that were enriched across subcellular compartments (such as Golgi apparatus-cytoplasm) showed that in numerous cases at least one of the interacting proteins was alternatively localized to a compartment other than its major site of localization and thus the interacting proteins did indeed share a common compartment. For instance, in 482 interaction pairs (Table 1) at least one protein showed both the Golgi apparatus localization and cytoplasmic localization. It should be noted that these interaction partners are multiple localized proteins and may also appear in other cellular compartments. This is not an uncommon situation, as > 50% of proteins of our final interactome network annotated with predicted subcellular localization information are, in fact, localized at two or more compartments.

As a final step of network reduction, isoforms of proteins are shown as a single node. These steps of successive filtering ultimately reduce the complexity of the network and increase the confidence of the C. floridanus interactome. Figure 1 summarizes the C. floridanus protein-protein interaction databases, our workflow, pruning steps and resulting ant network. It consists of 3289 nodes and 10834 edges (more details in (23)). The complete four networks are provided in the Datasheets 1-4 in Supplementary Material. We also identified several novel interactions predicted to be present in C. floridanus. For instance, an interaction was observed between S-phase kinase-associated protein 1 (SkpA, Cflo_N_g10272) and immune receptor peptidoglycan-recognition protein LC (PGRP-LC, Cflo_N_g10272). As an important component of ubiquitin-proteasome pathway SkpA is involved in Immune Deficiency (IMD) pathway regulation in D. melanogaster. Since PGRP-LC is also a regulator of the ant IMD pathway, the interaction we identified suggests that SkpA can modulate the IMD pathway by the interaction with PGRP-LC. Not only the interaction between protein complexes such as laminin subunit beta-1 (Cflo_N_g14102) and laminin subunit gamma-1 (Cflo_N_g9869) but also the interaction between Cflo_N_g14102 and C-type lectin precursor (Cflo_N_g765) was resolved (see Datasheets 3 in Supplementary Material for all the interactions).

To further supplement the proposed ant interactome, we performed a topology-based scoring of the network. The method CAPPIC used the intrinsic modularity of PPI network for assessing the confidence of individual interactions. 88.5% of the total interactions are high confidence (Fig. 2) while 9.65% were assigned to medium confidence and 1.8% to low confidence.

We applied the Mann-Whitney test to compare the average confidence scores of all four PPI networks and observed significant increase of confidence score for the first three steps from the preliminary network through DDI mediated filtering and localization-based filtering (Supplementary Fig. 1). The mean confidence score of the final interactome, after the isoform merging, did not change much. This is because the merging of this last step also eliminated some high confidence PPIs mediated by the isoforms. Nevertheless, the comparison of the proportions of high-confidence PPIs in the preliminary interactome and the final ant interactome indicates that it has a significantly increased number of high confidence interactions (in the preliminary network these are 78%, in the final 89%; Fisher’s exact test p-value < 2.2e-16). Note that the applied filtering steps also eliminated most of the low confidence PPIs (see low confidence zone in Supplementary Fig. 1). To further confirm the elimination after successive filtering steps, we compared the low confidence PPIs proportions in all four interactomes in a pairwise way with Fisher’s exact test and show a significant decrease in the number of the low confidence PPIs between the preliminary, DDI-filtered and localization-filtered interactomes (in preliminary 4.5%, in DDI-filtered 2.2%, in localization-filtered 1.6% with maximum p-value < 3.4e-05). These analyses clearly demonstrate the improvement of network quality after filtering steps.
Network analysis of *C. floridanus* interactome and accuracy assessment. The resulting PPI summarizes the whole network and reveals central connecting nodes. The final high confidence ant interactome showed a clustering coefficient of 0.094 with a mean shortest path length of 4.359, network diameter 14 and an average degree of 6.970. As a typical biological network it shows small-world connectivity and scale-free topology.

We further tested whether the proposed interactome aligns with the properties of a real biological network. To assess this, we derived three independent datasets and compared their topological properties with the proposed network. The average z-statistic value (Datasheet 5 in Supplementary Material) clearly indicates comparatively less variation of the ant interactome from the 'Barabási-Albert scale free model' (z-statistic = 23.06, −5.28) in terms of clustering coefficient and mean shortest path. However, the differences were high while comparing that of with 'Watts–Strogatz small world graph model' (z-statistic = −30.95, −58.49) and 'Erdős-Rényi flat-random' network model (z-statistic = 171.03, −52.72). Scale-free networks have been often observed in biological systems such as PPI and gene regulatory networks, therefore the bias towards such a network is an indicator of the equalization of the reconstructed ant interactome. To test another factor, the degree distribution of the ant interactome was...
much closer to Watts–Strogatz model ($z$-statistic $= 0.49$), although the differences with Barabási–Albert model was not too high ($z$-statistic $= 2.45$). The nodes in the network obey a power-law distribution indicating a typical, biological small-world and scale-free network.

**Gene ontology (GO) enrichment analysis.** The molecular function GO term over-representation analysis indicates enriched protein functions in the ant networks ($FDR < 0.05$; Table 2 and Datasheet 6 in Supplementary Material). Over-represented functional categories include the term ‘binding’ as to be expected from the PPI construction and a validation criterion. Out of 2804 proteins annotated as GO term GO:0005488 ‘binding’ in $C. floridanus$ proteome, 46.11% proteins are present in the final interactome. The number given here in parentheses is the number of proteins that do not have subcellular localization in any of the other compartment in the table.

**Figure 2.** Density plot of the confidence scores for interactions in the $C. floridanus$ interactome. How sure are the predicted protein-protein interactions? The distribution of the different confidence levels were computed with CAPPIC$^{31}$. Score distributions were separated into low, medium and high confidence category and the density for each category was plotted. In the three subsets scores range between 0 and 0.3 for subset 1 (green, low confidence), 0.3 and 0.7 for subset 2 (blue, medium confidence) and 0.7 and 1 for subset 3 (red, high confidence). Stippled coloured lines indicate median values for the different categories.

| Subcellular Localization | Number of PPIs | Number of Proteins |
|-------------------------|----------------|-------------------|
| Cytoplasm (2176)        | 6871           |                   |
| Golgi apparatus (243)   | 482            | 720               |
| Peroxisome (21)         | 11             | 3                 |
| Other (1)               | 1              | 0                 |
| Nucleus (1866)          | 4686           | 227               |
| Cytoskeleton (427)      | 1580           | 134               |
| Lysosome (50)           | 106            | 17                |
| Mitochondrion (319)     | 443            | 39                |
| ER (270)                | 360            | 59                |
| Extracellular (123)     | 120            | 16                |
| Plasma membrane (1352)  | 2365           | 383               |

| Subcellular Localization | Number of PPIs | Number of Proteins |
|-------------------------|----------------|-------------------|
| Cytoplasm (2176)        | 6871           |                   |
| Golgi apparatus (243)   | 482            | 720               |
| Peroxisome (21)         | 11             | 3                 |
| Other (1)               | 1              | 0                 |
| Nucleus (1866)          | 4686           | 227               |
| Cytoskeleton (427)      | 1580           | 134               |
| Lysosome (50)           | 106            | 17                |
| Mitochondrion (319)     | 443            | 39                |
| ER (270)                | 360            | 59                |
| Extracellular (123)     | 120            | 16                |
| Plasma membrane (1352)  | 2365           | 383               |

**Table 1.** Numbers of PPIs, by subcellular localization of interacting proteins in localization supported network. The numbers in the parentheses indicate the total number of proteins in the interactome that are present in each subcellular compartment (based on localization prediction, including multiple localized proteins). ER, endoplasmic reticulum. The number given here in parentheses is the number of proteins that do not have subcellular localization in any of the other compartment in the table.
random networks each with 100 random interactions among the proteins that were assigned to level-4 molecular function GO annotations using a custom-made Perl script which can be accessed from the GitHub repository (https://github.com/ShishirGupta-Wu/ant_ppi). We made sure the random networks did not contain any proteins pairs apparent in the preliminary interactome. Using the GOGO algorithm semantic similarity scores were also assigned to the random networks (non-PPIs) and these scores were further compared with the interacting proteins in a pairwise way using the Mann-Whitney U test. We observed that the interacting protein set had not only the highest average score of 0.47, this was also well separated and significantly higher than the average score in all the 30 non-PPI sets (Fig. 3). This comparison demonstrates the interactions in our calculated ant interactome are functionally relevant and clearly different from random networks.

**C. floridanus** interactome protein conservation compared with seven organisms. Proteins that perform essential functions are expected to be evolutionary conserved. We further investigated the evolutionary conservation of ant interactome proteins. Higher degree proteins are generally evolutionary better conserved, some caveats are discussed in [38,40]. To analyze this, node degree and the fraction of proteins present in the ant interactome that are conserved in different model organisms were compared. It turns out that in general the interactions are conserved and supported by most species tested and not just by one (Fig. 4). For exact quantification we did not check the possible restricted conservation of the binary ant PPIs, but more general the conservation of proteins that are present in the ant interactome and have orthologs in seven other species. For instance, in the ant interactome there are 535 proteins of degree 2. Out of these 535 proteins 451 have an ortholog in Anopheles, 209 have an ortholog in Arabidopsis, 298 have an ortholog in *C. elegans*, 404 have an ortholog in mouse, 82 have an ortholog in plasmodium, 151 have an ortholog in yeast, and 402 have a human ortholog. There was a positive correlation between degree and conservation in the evolutionary closest analyzed species *A. gambiense* (Spearman’s rank $r = 0.62$, p-value $= 3.8e-09$). Similar correlations are observed between ant and human ($r = 0.60$), and mouse ($r = 0.51$). Between ant and worm the correlation was weak ($r = 0.33$), while no significant correlation is observed between ant and *A. thaliana*, *P. falciparum*, and yeast. An ortholog table is provided in Datasheet 7 in Supplementary Material.

**Overall conservation and infection induced hubs and bottlenecks in the ant interactome.** We also evaluated the overall conservation of all the ant proteins with the other seven model organisms and compared the relatedness of the ant interactome proteins using the chi-square test. The analysis indicated the relatedness of corresponding proportions with p-value $< 0.05$ in each case. The differences in the number of orthologs can be clearly visualized (Fig. 5a) in case of ant comparison with protozoan parasite, yeast and plant. Due to the large phylogenetic distance to these three organisms there are less orthologs but these are well conserved (chi-square test).

**Table 2.** Top 20 over-represented GO molecular function terms in the ant-interactome. See Datasheet 6 in Supplementary material for the complete list of over-represented GO molecular functions terms in ant-interactome.

| GO ID         | GO molecular function term                                           | FDR     |
|---------------|---------------------------------------------------------------------|---------|
| GO:0005488    | Binding                                                             | 3.27E-130 |
| GO:0003824    | Catalytic activity                                                  | 8.17E-92 |
| GO:0097159    | Organic cyclic compound binding                                     | 9.44E-88 |
| GO:1901363    | Heterocyclic compound binding                                       | 9.44E-88 |
| GO:0036094    | Small molecule binding                                              | 3.59E-80 |
| GO:1901265    | Nucleoside phosphate binding                                        | 7.83E-78 |
| GO:0000166    | Nucleotide binding                                                  | 7.83E-78 |
| GO:0043168    | Anion binding                                                       | 2.45E-66 |
| GO:0017076    | Purine nucleotide binding                                           | 5.15E-59 |
| GO:0032553    | Ribonucleotide binding                                              | 7.09E-59 |
| GO:0032555    | Purine ribonucleotide binding                                       | 1.76E-58 |
| GO:0035639    | Purine ribonucleoside triphosphate binding                          | 2.85E-58 |
| GO:0097367    | Carbohydrate derivative binding                                     | 2.68E-55 |
| GO:0043167    | Ion binding                                                         | 3.31E-52 |
| GO:0016817    | Hydrolase activity, acting on acid anhydrides                       | 2.18E-39 |
| GO:0005515    | Protein binding                                                     | 2.18E-39 |
| GO:0016818    | Hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides | 9.29E-39 |
| GO:0016462    | Pyrophosphatase activity                                            | 6.95E-38 |
| GO:0017111    | Nucleoside-triphosphatase activity                                  | 3.90E-37 |
| GO:0030554    | Adenyl nucleotide binding                                           | 3.90E-37 |
The remaining set of the other four organisms including insect, human, mouse and worm together consists of orthologs in comparison to the ant proteins (Fig. 5b and Datasheet 7 in Supplementary Material). 187 proteins of the ant interactome are ant-specific in this comparison: they do not have orthologs in any of the analysed organisms (Fig. 5b). The analysis of central topological properties of a PPI network helps to identify key multifunctional components of the network41. Infection induced proteins of *C. floridanus* are conserved in related organisms including key interactions. The degree of the node42 and the betweenness centrality43...
represent the most important properties in the PPI network because of their role in maintaining the functional integrity and connectivity of the network. Proteins with high degree are termed as hubs while the proteins with high betweenness centrality are termed bottlenecks.

Figure 5. Conservation of C. floridanus proteins. (a) Network showing the similarity of C. floridanus proteins (central node) with other organisms proteins (peripheral nodes) quantitively by color codes. Color gradients represent the number of proteins in each species that have orthologs in either C. floridanus interactome (itor, inner circle) or in the whole C. floridanus proteome (tor, outer circle). A similar color to the ant node shows close orthology relation while nodes of more distant species have different colors. A correlation between the colors of two circles is expected to show that the interactome successfully represent the orthology relationships between ant and other species. For instance, if both circles of a species show a similar color to ant, it depicts that the close orthology to ant can be also observed in the interactome. Pearson's Chi-squared test with Yates' continuity correction showed relationship (p-value < 2.2e-16) between the total ant proteins orthologs and interactome proteins orthologs for each corresponding organisms. (b) Conservation of ant interactome proteins in set of any analyzed organisms. Number of ant interactome proteins that are unique to ant and that are orthologous to proteins in different numbers of analyzed species is shown. (c) Differentially expressed top hubs and bottlenecks (represented by black dots) and presence/absence of their orthologs in other seven organisms. Differential expression values (log2FC) of hubs and bottlenecks in bacterial infection of C. floridanus are represented by a color gradient.
We applied Fisher’s exact test to compare the proportion of multi-localized proteins in hubs and bottlenecks to non-hubs and non-bottlenecks, respectively. Supplementary Fig. 2 shows differences between the localization of bottlenecks and hubs of the ant interactome. For bottleneck proteins, 70% were found to be multi-localized (versus 56% for non-bottleneck proteins; significant difference, \( p = 9.6 \times 10^{-19} \)). On the other hand, 62% of the hub proteins had multiple localizations (versus 56% for non-hub proteins; significant difference, \( p = 0.001575 \)).

Integration of the RNASeq data with the ant interactome revealed differentially expressed infection-induced hubs and bottlenecks during the bacterial infection of *C. floridanus* (Fig. 5c). These include also well-known key proteins involved in *C. floridanus* immune response such as nuclear factor NF-kappa-B p110 (Relish, Cflo_N_g6682), acidic mammalian chitinase (*Cflo_N_g2277*), as well as stress-related protein cytochromes P450 6A1 (*Cflo_N_g11706*). Given the high importance of hubs and bottlenecks in PPI networks and their differential expression during bacterial infection, all the identified proteins are expected to participate in the defense against bacterial pathogen, and hence can also be examined for decoding immune mechanisms. The insect peritrophic membrane (PM) imposes protective physical barriers over the midgut epithelium. The PM related proteins have shown their potential as targets for pest control. Therefore, the important ant peritrophic membrane protein 1 (*Cflo_N_g4555*) (Fig. 5c) with no human homology could be further tested as a potential pest target. However, differential expression does not guarantee a protein to be the best target and therefore, other topologically important proteins in the network without human homology (Datasheet 7 in Supplementary Material) should also be considered as potential pest targets in future.

**Conclusions**

Our curated ant interactome is the first large-scale PPI network of an ant. It allows besides numerous analysis of network biology to study how different cellular processes connect to each other including hub proteins and different types of crosstalk, for instance in immunity.

Similarly, the PPI maps of other sequenced ants can be reliably predicted using the interologs of the reconstructed high-confidence *C. floridanus* interactome. Moreover, detailed cross-validation, comparison with random networks, GO annotation, and conservation analysis support the high quality of the resulting ant interactome and its construction steps. The network analysis including evolutionary conserved network proteins further suggest that topologically important proteins could also be exploited as future pest targets. For instance, cytochrome P450 6A1 (*Cflo_N_g11706*), peritrophic membrane protein 1 (*Cflo_N_g4555*), flexible cuticle protein 12 (*Cflo_N_g6859*), endocuticle structural glycoprotein SgAbd-1 (*Cflo_N_g7775*) were identified as topologically important differentially expressed proteins with no human orthologs. Nevertheless, specific interactions highlighted from our global analysis will need individual follow up by detailed investigations.

**Materials and Methods**

**Reconstructing protein-protein interaction map of *C. floridanus***. We compiled the list of experimentally verified high-confidence PPIs available in Database of interacting proteins (DIP) from DroID database which includes data from different studies including interactions from high throughput Gal4 proteome-wide yeast two-hybrid (Y2H) screens, LexA Y2H system screens, PPIs from fly protein interaction map, interactions determined in large-scale co-affinity purification (co-AP)/MS screens, interactions from BIND, BioGRID, MINT, IntAct, and databases available in DroID v2014_10.

The *C. floridanus* interologs of the entire template PPIs were determined using orthology predictions from the software InParanoid and OrthoMCL. These were further customized using own perl and bash scripts. For DIP interactors we used the default parameters of InParanoid. For the fly data orthology was determined using the stricter Blast e-value of 1e-05 was used and the MCL inflation index set to 1.5. InParanoid distinguished seed orthologs with co-orthologs and left fewer possibilities of mixing outparalogs in orthologous clusters. Consensus predictions of InParanoid and OrthoMCL were added to InParanoid seed orthologs to create a set of interologs.

**Pruning PPIs with domain-domain interactions.** The amino acid sequences of non-redundant preliminary PPIs were extracted and domains were assigned to them using Pfam version 27.0. The list of non-redundant domain-domain interactions was prepared from the meta-databases Domine, DIMA 3.0, and IDDI database. These use complexes available in the Protein Data Bank (PDB) to identify by interacting domains the Pfam families containing these domains. These Pfam families are then predicted to be interacting. This list was used to parse the template PPIs. All interactions were categorized whether they are supported (good interactions, used for further filtering steps) or not by domain-domain interactions (DDIs).

**Subcellular localization filtering.** The subcellular localization of *C. floridanus* proteins was determined with orthology to Swiss-Prot proteins and the extended version of KnowPredsite available at UniLoc server, a knowledge-based classifier for protein subcellular localization. If in a binary interaction, both proteins do not share the same localization or at least one compartment in multiple localized proteins, the interaction was ruled out as probable not occurring.

**Isoform filtering.** The information on *C. floridanus* protein isoforms and their function was extracted from our previous publication of *C. floridanus* re-annotation and transcriptome sequencing. To reduce network complexity and noise, isoforms of any specific protein present in the network were represented as a single node. Although, the data files for all the networks are provided in the Supplementary Tables 1–3 which allow interested readers to analyze the network of their choice further if they wish.
Assigning the confidence score. In fact, the preliminary network is filtered successively as mentioned above to reconstruct the final network, in this way the final network is already of high-confidence as many network biologists working on PPI networks have used DDIIs and subcellular localization either to increase confidence or validate the interacting pairs. Here additionally we used topology-based method CAPPIC (cluster-based assessment of protein-protein interaction confidence) to assign the interaction confidence score \( r_{PPI} \) in the filtered network. In brief, CAPPIC calculations are based on the assumption that the proteins existing in the same network module are expected to have a higher number of common neighbours (neighbourhood interconnectedness) and a short path length between them. For scoring the confidence level, CAPPIC first performs the clustering of the network using a robust clustering algorithm, Markov Cluster (MCL) and then scores the interactions according to their level of compliance with the basic assumptions of topology-based methods. For the clustering we used an MCL inflation value of 1.5. Scores were classified to three subsets; low confidence score between 0 to 0.3, medium confidence score between 0.3 to 0.7, and high confidence score between 0.7 to 1.

Network analysis and visualization. The C. floridanus interactome was subjected to topological analysis using Network Analyzer plugin version 2.7 of Cytoscape version 2.8.1. The node degree distribution, mean path length, network diameter and betweenness centrality (BC) were determined with graph theoretic analysis implemented with CentiScaPe. For the network \( G(V,E) \), the BC of node \( n \) is defined as follows

\[
BC(n) = \sum_{s \neq n \neq t} \left( \frac{\sigma_{st}(n)}{\sigma_{st}} \right)
\]

here \( s \) and \( t \) are network nodes different from node \( n \), \( \sigma_{st} \) is the number of shortest paths from \( s \) to \( t \), and \( \sigma_{st}(n) \) gives the number of shortest paths from \( s \) to \( t \) that go through node \( n \).

Hubs and bottlenecks in the network were identified with cytoHubba. Hubs were defined as proteins connecting with \( \geq 5 \) proteins. Moreover, top 20% of bottlenecks and hubs were considered for mapping of the RNASeq expression data which was collected from our previous publication.

Random networks. We generated random networks following the Erdős-Rényi Model, Barabási-Albert Model and randomized the proposed (final) ant interactome while preserving the total number of interactome nodes using the Network Randomizer plugin of Cytoscape. A total of 1000 random simulation were employed to generate the undirected random graphs. For all three network sets we computed topological parameters, mean shortest path, degree distribution and clustering coefficient and compared their differences to the native ant interactome using the statistical Z-test.

Functional annotation. Blast2GO was used to annotate the Gene Ontology (GO) terms of proteins involved in the reconstructed interactome. Over-representation analyses of GO terms was performed using the Gossip package of the Blast2GO suite. A two-tailed Fisher's exact test followed by false discovery rate (FDR) correction for multiple testing was applied to see the functional difference of ant interactome proteins annotations (foreground set) and full C. floridanus proteome annotations (background set). Only differences having an adjusted p-value < 0.05 were considered significant.

Orthology analysis. InParanoid was used to identify the orthologs of topologically important nodes in seven model organisms: Anopheles gambiense, Arabidopsis thaliana, Caenorhabditis elegans, Homo sapiens, Mus musculus, Plasmodium falciparum, and Saccharomyces cerevisiae. Only the ortholog with 100% bootstrap support was considered as true ortholog. As a note of caution, the conservation was calculated rather conservatively demanding double orthology relations. Hence, the absence of an ortholog (Suppl. Datasheet 7) only indicates that the highly restrictive threshold was not met. Generally, a sequence related protein may still be found by less restrictive algorithms (e.g. BLAST).

For exact quantification of the degree of conservation of ant PPIs we did not check the possible restricted conservation of the binary ant PPIs, but more general the conservation of proteins that are present in the ant interactome and have orthologs in seven other species. After calculation of the orthology relationships between ant and other organisms we identified for every degree the occurrence value of the ant interactome and how many orthologs are present in other species. For each organism the fraction of proteins at a particular ant degree is considered as the number of ant protein orthologs at that particular degree and greater divided by the number of proteins in the set.

Data availability
All data generated and analysed during this study are included in this published article (and its supplementary Information files). The dataset and codes for random network generation are also available at https://github.com/ShishirGupta-Wu/ant_ppi.

Received: 9 May 2019; Accepted: 22 January 2020;
Published online: 11 February 2020

References
1. Gadou, J., Heinz, J., Holldobler, R. & Schmid, M. Population and colony structure of the carpenter ant Camponotus floridanus. Mol. Ecol. 5, 785–792 (1996).
2. Zientz, E., Beyaert, I., Gross, R. & Feldhaar, H. Relevance of the endosymbiosis of Blochmannia floridanus and carpenter ants at different stages of the life cycle of the host. Appl. Env. Microbiol. 72, 6027–6033, https://doi.org/10.1128/AEM.00933-06 (2006).
25. Cusick, M. E.

24. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research?

23. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high-throughput interaction data.

22. Gupta, S. K. Re-annotation of Camponotus floridanus Genome and Characterization of Innate Immunity Transcriptome Responses to Bacterial Infections PhD thesis, Universität Würzburg, (2016).

21. Ganguly, S. & Saha, A. Computational prediction of host-pathogen protein-protein interactions.

20. Gehrke, K., Knoth, O., Heil, F. & Baumeister, W. H. Protein-protein interaction networks reveal essential biology.

19. Pereira, C., Denise, A. & Lespinet, O. A meta-approach for improving the prediction and the functional annotation of orthologous protein-protein interactions.

18. Frasier, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein-protein interactions.

17. Schuster-Bockler, B. & Bateman, A. Reuse of structural domain-domain interactions in protein networks.

16. Wojcik, J. & Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs.

15. Itzhaki, Z., Akiva, E., Altuvia, Y. & Margalit, H. Evolutionary conservation of domain-domain interactions.

14. Wang, Y. C.

13. Wang, Y. C.

12. Remmelle, C. W.

11. Dyer, M. D., Murali, T. M. & Sobral, B. W. Computational prediction of host-pathogen protein-protein interactions.

10. Akhoon, B. A.

9. Yu, H.

8. Mahdavi, M. A. & Lin, Y. H. False positive reduction in protein-protein interaction predictions using gene ontology annotations.

7. Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein-protein interactions.

6. Yu, H.

5. von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nat. 417, 399–403, https://doi.org/10.1038/nature00750 (2002).

4. Ben-Hur, A. & Noble, W. S. Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinforma. 7(Suppl 1), S2, https://doi.org/10.1186/1471-2105-7-S1-S2 (2006).

3. Gupta, S. K. et al. Literature-curated protein interaction datasets. Nat. methods 6, 39–46, https://doi.org/10.1038/nmeth.1284 (2009).

2. Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein-protein interactions.

1. Yu, H. et al. Annotation transfer between genomes: protein–protein interologs and protein–DNA regologs. Genome Res. 14, 1107–1118, https://doi.org/10.1101/gr.177404 (2004).


56. Guruharsha, K. G.

58. Stark, C.

54. Formstecher, E.

52. Stanyon, C. A.

49. Salwinski, L.

45. Sajjadian, M. & Hosseininaveh, V.

46. Zhang, X. & Guo, W.

53. Zhong, J., Zhang, H., Stanyon, C. A., Tromp, G. & Finley, R. L. Jr. A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. Genomes Res. 13, 2691–2699, https://doi.org/10.1101/gr.1134603 (2003).

50. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052, https://doi.org/10.1016/j.jmb.2001.09.081 (2001).

51. Finn, R. D. et al. Pfam: the protein families database. Nucleic acids Res. 24, 306–310, https://doi.org/10.1093/nr/naf057 (2006).

52. Croft, D. B., Mundo, N. F., Ouzounis, C. A. & Marchler-Bauer, A. Pfam: the protein families database. Nucleic acids Res. 39, D292–298, https://doi.org/10.1093/nar/gkr848 (2011).

53. Ouzounis, C. A. et al. Pfam: the protein families database. Nucleic acids Res. 39, D292–298, https://doi.org/10.1093/nar/gkr848 (2011).

54. Ewing, B., Hillier, L. V., Wendl, M. C. & Green, P. O. A novel approach for sequencing human genomic DNA. Science 263, 1–14, https://doi.org/10.1126/science.2513851 (1993).

55. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921, https://doi.org/10.1038/nature01097 (2001).

56. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921, https://doi.org/10.1038/nature01097 (2001).

57. Rubin, E. M., Venter, J. C., Remington, K., Heitz, F. & Gocayne, J. D. The DNA sequence of Drosophila melanogaster. Science 289, 2469–2473, https://doi.org/10.1126/science.289.5483.2469 (2000).

58. Venter, J. C. et al. The sequence and organization of the human genome. Science 291, 1304–1351, https://doi.org/10.1126/science.1059916 (2001).

59. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. Science 287, 2185–2195, https://doi.org/10.1126/science.1008782 (1999).

60. Stanyon, C. A. et al. Drosophila protein-interaction map centered on cell-cycle regulators. Cell 147, 690–703, https://doi.org/10.1016/j.cell.2011.08.047 (2011).

61. Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314, 1041–1052, https://doi.org/10.1016/j.jmb.2001.09.081 (2001).

62. Finn, R. D. et al. Pfam: the protein families database. Nucleic acids Res. 24, 306–310, https://doi.org/10.1093/nr/naf057 (2006).

63. Ouzounis, C. A. et al. Pfam: the protein families database. Nucleic acids Res. 39, D292–298, https://doi.org/10.1093/nar/gkr848 (2011).

64. Luo, Q., Pagel, P., Vilne, B. & Frishman, D. DIMA 3.0: Domain Interaction Map. Nucleic acids research 41, (2013).

65. Vlasblom, J. & Wodak, S. J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. Cell.2011.08.047 (2011).

66. Berman, H. M. et al. The Protein Data Bank. Nucleic acids Res. 28, 235–242 (2000).

67. Lin, H. N., Chen, C. T., Sung, T. Y., Ho, S. Y. & Hsu, W. L. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. BMC Bioinformatics. 10(Suppl 1), S9, https://doi.org/10.1186/1471-2105-10-S1-S9 (2012).

68. Berman, H. M. et al. The Protein Data Bank. Nucleic acids Res. 28, 235–242 (2000).

69. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: a web tool for confidence scoring of biological interactions. Nucleic acids Res. 39, D730–735, https://doi.org/10.1093/nar/gkj229 (2011).

70. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world. Proc. Natl Acad. Sci. U S Am. 100, 15966–15971, https://doi.org/10.1073/pnas.1520426100 (2003).

71. Kuchaiev, O., Rasajski, M., Higham, D. J. & Pizurki, N. Geometric de-noising of protein-protein interaction networks. PLoS computational biology 5, e1000454, https://doi.org/10.1371/journal.pcbi.1000454 (2009).

72. Vladislam, J. & Wodak, S. J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics. 10, 99, https://doi.org/10.1186/1471-2105-10-99 (2009).

73. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504, https://doi.org/10.1101/gr.123903 (2003).

74. Scardoni, G., Petterlini, M. & Laudanna, C. Analyzing biological network parameters with CentiScaPe. Bioinformatics. 25, 2857–2859, https://doi.org/10.1093/bioinformatics/btp517 (2009).

75. Chin, C. H. et al. cytoHubba: identifying hub objects and sub-networks from complex networks. BMC Syst. Biol. 8(Suppl 4), S11, https://doi.org/10.1186/1752-0509-8-S4-S11 (2014).

76. Erdos, P. & Rényi, A. On Random Graphs I. Publicationes Mathematicae 6, 290–297 (1959).

77. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. Science 286, 509–512 (1999).

78. Tosadori, G., Bestvina, I., Spoto, F., Laudanna, C. & Scardoni, G. Creating, generating and comparing random network models with NetworkRandomizer. F1000Research 5, 2524, https://doi.org/10.12688/f1000research.9203.3 (2016).

79. Kreyzig, E. Applied Mathematics, fourth ed., (Hoboken, NJ: John Wiley & Sons, (1979).

80. Bluhmgen, N. et al. Biological profiling of gene groups utilizing Gene Ontology. Genome Inf. 16, 106–115 (2005).

81. Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289–300 (1995).

Acknowledgements

We gratefully acknowledge funding of this work by the German Research Foundation (DFG-GR1243/8-1) and T.D. by project number 210879364 – TRR 124/B-1). S.K.G. gratefully acknowledges that the present work is a partial output of his PhD thesis and fruitful discussions with GRK 2157 3D tissue infect (project number 270563345). M.S. would like to thank for the financial support from Frauenbeauftragten Büro, University of Wuerzburg, Germany. S.K.G. and Ö.O. would like to thank Dr. Atanas Kamburov (Harvard Medical School,
Massachusetts, USA) for help and useful discussions on the CAPPIC calculations. This publication was funded by the German Research Foundation (DFG) and the University of Wuerzburg in the funding programme Open Access Publishing.

**Author contributions**
S.K.G. and T.D. conceived and designed the study. S.K.G., Ö.O., and M.S. performed the procedure and analyzed the data. All authors contributed to biological interpretation of the results. M.S. and T.D. wrote the manuscript. All authors read and agreed to the final version of the manuscript.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-59344-1.

**Correspondence** and requests for materials should be addressed to T.D.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

![Open Access](https://creativecommons.org/licenses/by/4.0/)

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020