Evolution of resilience in protein interactomes across the tree of life

Marinka Zitnik[a], Rok Sosič[a], Marcus W. Feldman[b,1], and Jure Leskovec[a,c,1]

[a]Department of Computer Science, Stanford University, Stanford, CA 94305; [b]Department of Biology, Stanford University, Stanford, CA 94305; and [c]Chan Zuckerberg Biohub, San Francisco, CA 94158

Contributed by Marcus W. Feldman, December 18, 2018 (sent for review October 19, 2018; reviewed by Edoardo Airoldi and Aviv Bergman)

Phenotype robustness to environmental fluctuations is a common biological phenomenon. Although most phenotypes involve multiple proteins that interact with each other, the basic principles of how such interactome networks respond to environmental unpredictability and change during evolution are largely unknown. Here we study interactomes of 1,840 species across the tree of life involving a total of 8,762,166 protein–protein interactions. Our study focuses on the resilience of interactomes to network failures and finds that interactomes become more resilient during evolution, meaning that interactomes become more robust to network failures over time. In bacteria, we find that a more resilient interactome is in turn associated with the greater ability of the organism to survive in a more complex, variable, and competitive environment. We find that at the protein family level proteins exhibit a coordinated rewiring of interactions over time and that a resilient interactome arises through gradual change of the network topology. Our findings have implications for understanding molecular network structure in the context of both evolution and environment.

protein–protein interaction networks | molecular evolution | ecology | network resilience | network rewiring

The enormous diversity of life shows a fundamental ability of organisms to adapt their phenotypes to changing environments (1). Most phenotypes are the result of an interplay of many molecular components that interact with each other and the environment (2–5). The study of life's diversity has a long history and extensive phylogenetic studies have demonstrated evolution at the DNA sequence level (6–8). While studies based on sequence data alone have demonstrated evolution of genomes, mechanistic insights into how evolution shapes interactions between proteins in an organism remain elusive (9, 10).

DNA sequence information has been used to associate genes with their functions (11), determine properties of ancestral life (12, 13), and understand how the environment affects genomes (14). Despite these advances in understanding DNA sequence evolution, little is known about basic principles that govern the evolution of interactions between proteins. In particular, evolution of DNA and amino acid sequences could lead to pervasive rewiring of protein–protein interactions and create or destroy the ability of the interactions to perform their biological functions.

The importance of protein–protein interactions has spurred experimental efforts to map all interactions between proteins in a particular organism, its interactome, namely the complex network of protein–protein interactions in that organism. A large number of high-throughput experiments have reported high-quality interactomes in a number of organisms (15–19). Because interactomes underlie all living organisms, it is critical to understand how these networks change during evolution (20, 21) and elucidate key principles of their structure.

Here, we use protein interactions measured by these large-scale interactome mapping experiments and study the evolutionary dynamics of the interactomes across the tree of life. Our protein interaction dataset contains a total of 8,762,166 physical interactions between 1,450,633 proteins from 1,840 species, encompassing all current protein interaction information at a cross-species scale (*SI Appendix*, section S1 and Table S4). We group these interactions by species and represent each species with a separate interactome network, in which nodes indicate a species' proteins and edges indicate experimentally documented physical interactions, including direct biophysical protein–protein interactions, regulatory protein–DNA interactions, metabolic pathway interactions, and kinase–substrate interactions measured in that species. We integrate into the dataset (22) the evolutionary history of species provided by the tree of life constructed from small subunit ribosomal RNA gene sequence information (12) (*SI Appendix*, section S2). Using network science, we study the network organization of each interactome, in particular its resilience to network failures, a critical factor determining the function of the interactome (23–26). We identify the relationship between the resilience of an interactome and evolution and use this resilience to uncover relationships with natural environments in which organisms live. Although the interactomes are incomplete and biased toward much-studied proteins and model species (*SI Appendix*, section S1 and Fig. S7), our analyses give results that are consistent across taxonomic groups, that are not sensitive to network data quality or network size change (*SI Appendix*, section S8 and Fig. S8), and indicate that our conclusions will still hold when more protein interaction data become available.

Significance

The interactome network of protein–protein interactions captures the structure of molecular machinery that underlies organismal complexity. The resilience to network failures is a critical property of the interactome as the breakdown of interactions may lead to cell death or disease. By studying interactomes from 1,840 species across the tree of life, we find that evolution leads to more resilient interactomes, providing evidence for a longstanding hypothesis that interactomes evolve favoring robustness against network failures. We find that a highly resilient interactome has a beneficial impact on the organism's survival in complex, variable, and competitive habitats. Our findings reveal how interactomes change through evolution and how these changes affect their response to environmental unpredictability.

Author contributions: M.Z., M.W.F., and J.L. designed research; M.Z., R.S., and J.L. performed research; M.Z., R.S., M.W.F., and J.L. analyzed data; and M.Z., R.S., M.W.F., and J.L. wrote the paper.

Reviewers: E.A., Harvard University; and A.B., Albert Einstein College Of Medicine.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: mfeldman@stanford.edu or jure@cs.stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1818013116/-/DCSupplemental.

Results

Modeling Resilience of the Interactome. Natural selection has influenced many features of living organisms, both at the level of individual genes (27) and at the level of whole organisms (13). To determine how natural selection influences the structure of interactomes, we study the resilience of interactomes to network failures (23, 25, 26). Resilience is a critical property of an interactome as the breakdown of proteins can fundamentally affect the exchange of any biological information between proteins in a cell (Fig. 1*A*). Network failure could occur through the removal of a protein (e.g., by a nonsense mutation) or the disruption of a protein–protein interaction (e.g., by environmental factors, such as availability of resources). The removal of even a small number of proteins can completely fragment the interactome and lead to cell death and disease (4, 5) (*SI Appendix*, section S5.1 and Table S3). Disruptions of interactions can thus affect the interactome to the extent that its connectivity can be completely lost and the interactome loses its biological function and increases the risk of disease (5).

We formally characterize the resilience of an interactome of a species by measuring how fragmented the interactome becomes when all interactions involving a fraction f of the proteins (nodes) are randomly removed from the interactome (Fig. 1*A*). The resulting isolated network components then determine the interactome fragmentation. A network component is a connected subnetwork of the interactome in which any two nodes can reach each other by a path of edges. The smaller the network component is, the fewer nodes can be reached from any given node in the component. To characterize how the interactome fragments into isolated components we use the Shannon diversity index (28–31), which we modify to ensure that the resilience of interactomes with different numbers of proteins can be compared (Fig. 1*B* and *SI Appendix*, section S5.2). In particular, when the interactome G is subjected to a network failure rate f, it is fragmented into a number of isolated components of varying sizes (*SI Appendix*, Fig. S1). We quantify connectivity of the resulting fragmented interactome G_f by calculating the modified Shannon diversity on the resulting set of isolated components. Let $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ be k isolated components in G_f. The modified Shannon diversity of G_f is then calculated as the entropy of $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ as

$$H_{\mathrm{msh}}(G_f) = -\frac{1}{\log N} \sum_{i=1}^{k} p_i \log p_i, \qquad [1]$$

where N is the number of proteins in the interactome and $p_i = |\mathcal{C}_i|/N$ is the proportion of proteins in the interactome G that are in component $\mathcal{C}_i$. We can interpret p_i as the probability of seeing a protein from component $\mathcal{C}_i$ when sampling one protein from the fragmented interactome G_f. That is, Eq. **1** quantifies the uncertainty in predicting the component identity of a protein that is taken at random from the interactome. Finally, we use the normalization factor $1/\log N$ because it corrects for differences in the numbers of proteins in the interactomes and ensures that interactomes of different species can be compared. The range of possible H_{msh} values is between 0 and 1, where these limits correspond, respectively, to a connected interactome in which any two proteins are connected by a path of edges and a completely fragmented interactome in which every protein is its own isolated component. If the fragmented interactome has one large component and only a few small broken-off components, then the modified Shannon diversity is low, providing evidence that the interactome has network structure that is resilient to network failures (23) (*SI Appendix*, Fig. S2). In contrast, if the interactome breaks into many small components, it becomes fragmented, and its modified Shannon diversity is high (*SI Appendix*, Fig. S2), indicating that the interactome is not resilient to network failures.

To fully characterize the interactome resilience of a species we measure fragmentation of the species' interactome across all possible network failure rates (*SI Appendix*, Fig. S3). Consider the interactome of the pathogenic bacterium *Haemophilus influenzae* and the interactome of humans, which have different resilience (Fig. 1*C*). In the *H. influenzae* interactome, on removing small fractions of all nodes many network components of varying sizes appear, producing a quickly increasing Shannon diversity. In contrast, the human interactome fragments into a few small components and one large component whose size slowly decreases as small components break off, resulting in Shannon diversity that increases linearly with the network failure rate (Fig. 1*C*). Thus, unlike the fragmentation of the *H. influenzae* interactome, the human interactome stays together as a large component for very high network failure rates, providing evidence for the topological stability of the interactome. In general, the calculation of modified Shannon diversity over all possible network failure rates f yields a monotonically increasing function that reaches its minimum value of 0 at $f = 0$ (i.e., a connected interactome) and its maximum value of 1 at $f = 1$ (i.e., a completely fragmented interactome) as the interactome becomes increasingly fragmented with increasing network failure rate f (*SI Appendix*, section S5.3 and Fig. S3). We therefore define resilience of interactome G as one minus the area under the curve defined by that function,

$$\mathrm{Resilience}(G) = 1 - \int_0^1 H_{\mathrm{msh}}(G_f)\,\mathrm{d}f, \qquad [2]$$

which takes values between 0 and 1; a higher value indicates a more resilient interactome.

Resilience of Interactomes Throughout Evolution. We characterize systematically the resilience of interactomes for all species in the dataset (Fig. 1*D* and *SI Appendix*, Table S5). We find that species display varying degrees of interactome resilience to network failures (Fig. 1*E*). At a global, cross-species scale, we find that a greater amount of genetic change is associated with a more resilient interactome structure [locally weighted scatterplot smoothing (LOWESS) fit; $R^2 = 0.36$; Fig. 1*F*], and this association remains strong even after statistical adjustments for the influence of many other variables (*SI Appendix*, section S8). The more genetic change a species has undergone, the more resilient is its interactome. The evolution of a species, which is represented by the total branch length from the root to the leaf taxon representing that species in the tree of life (12), thus predicts resilience of the species' interactome, providing empirical evidence that interactome resilience is an evolvable property of organisms (26). This finding also suggests that the structure of present-day interactomes reflects their history or that interactomes must have a certain structure because that structure is well suited to the network's biological function. From an evolutionary standpoint, this finding points in the direction of topologically stable interactomes, which suggests that evolutionary forces may shape protein interaction networks in such a way that their large-scale connectivity, i.e., the network's biological function, remains largely unaffected by small network failures as long the failures are random.

We also find that species from the same taxonomic domain have more similar interactome resilience than species from different domains ($P = 6 \cdot 10^{-11}$ for bacteria against eukaryotes; see *SI Appendix*, Fig. S10 for comparisons between other taxonomic groups). Furthermore, the degree of interactome resilience is significantly higher than expected by chance alone (*SI Appendix*, Fig. S9); that is, in a similar random network of identical size and degree distribution ($P = 5 \cdot 10^{-12}$), indicating that naturally occurring interactomes have higher

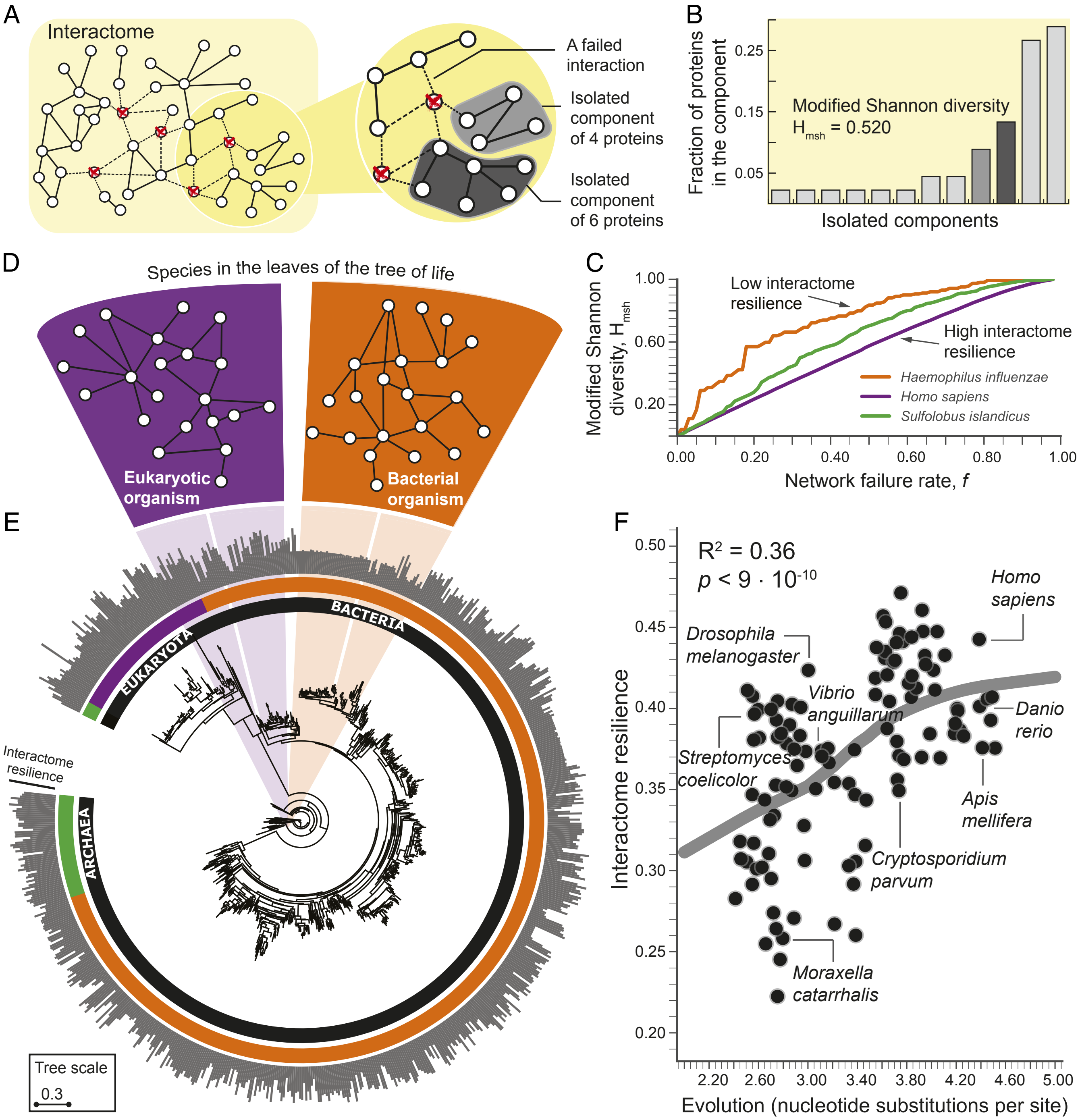

Fig. 1. Protein interaction data of 1,840 species consisting of 8,762,166 interactions by 1,450,633 proteins reveal the resilience of interactomes across vast evolutionary distances. (*A*) The interactome of an organism consists of all physical interactions between proteins in the organism. When interactions involving a certain fraction ($f = 5/45$ in this example) of the proteins are removed from the interactome, the interactome fragments into a number of isolated network components. (*B*) Modified Shannon diversity H_{msh} (*SI Appendix*, section S5) measures how the interactome fragments into isolated components at a given network failure rate f. (*C*) The resilience of the interactome integrates modified Shannon diversity H_{msh} across all possible failure rates f (*SI Appendix*, section S5). Resilience value 1 indicates the most resilient interactome, and resilience value 0 indicates a complete loss of the connectivity of the interactome (*SI Appendix*, Fig. S3). *Homo sapiens* (*H. influenzae*) has the most (least) resilient interactome (their resilience is 0.461 and 0.267, respectively) among the three selected organisms. (*D*) A small neighborhood of the interactome in a eukaryotic and a bacterial species. As ancestral species have gone extinct, older interactomes have been lost, and only interactomes of present-day species are available to us. (*E*) Phylogenetic tree showing 1,539 bacteria, 111 archaea, and 190 eukarya (12). Evolution of a species is represented as the total branch length (nucleotide substitutions per site) from the root to the corresponding leaf in the tree (*SI Appendix*, section S2). The outside circle of bars shows the interactome resilience of every species. Current protein–protein interaction data might be prone to notable selection and investigative biases (*SI Appendix*, section S1). (*F*) This plot shows the interactome resilience for 171 species with at least 1,000 publications in the NCBI PubMed (*SI Appendix*, Fig. S7). Across all species, evolution of a species predicts resilience of the species' interactome to network failures (LOWESS fit; $R^2 = 0.36$); more genetic change implies a more resilient interactome. Three species with the most nucleotide substitutions per site (far right on the x axis) have on average a 20.4% more resilient interactome than the three species with the least substitutions (far left on the x axis).

resilience than their random counterparts. These findings are independent of genomic attributes of the species, such as genome size and the number of protein-coding genes, and are not direct effects of network size, the number of interactions in each species, broad-tailed degree distributions (23), or the presence of hubs in the interactome networks (*SI Appendix*, Fig. S8 and Table S1). Furthermore, these findings are consistent across a variety of assays that are used to measure the interactome (*SI Appendix*, Table S2).

Relationship Between Interactome Resilience and Ecology. We next ask whether there is a relationship between species' interactome resilience and aspects of species' ecology (*SI Appendix*, section S4). We examine the relationship between interactome resilience and the fraction of regulatory genes and find that bacteria with more resilient interactomes have significantly more regulatory genes in their genomes ($R^2 = 0.32$; Fig. 2*A*). Bacteria with highly resilient interactomes can also survive in more diverse and competitive environments, as revealed by

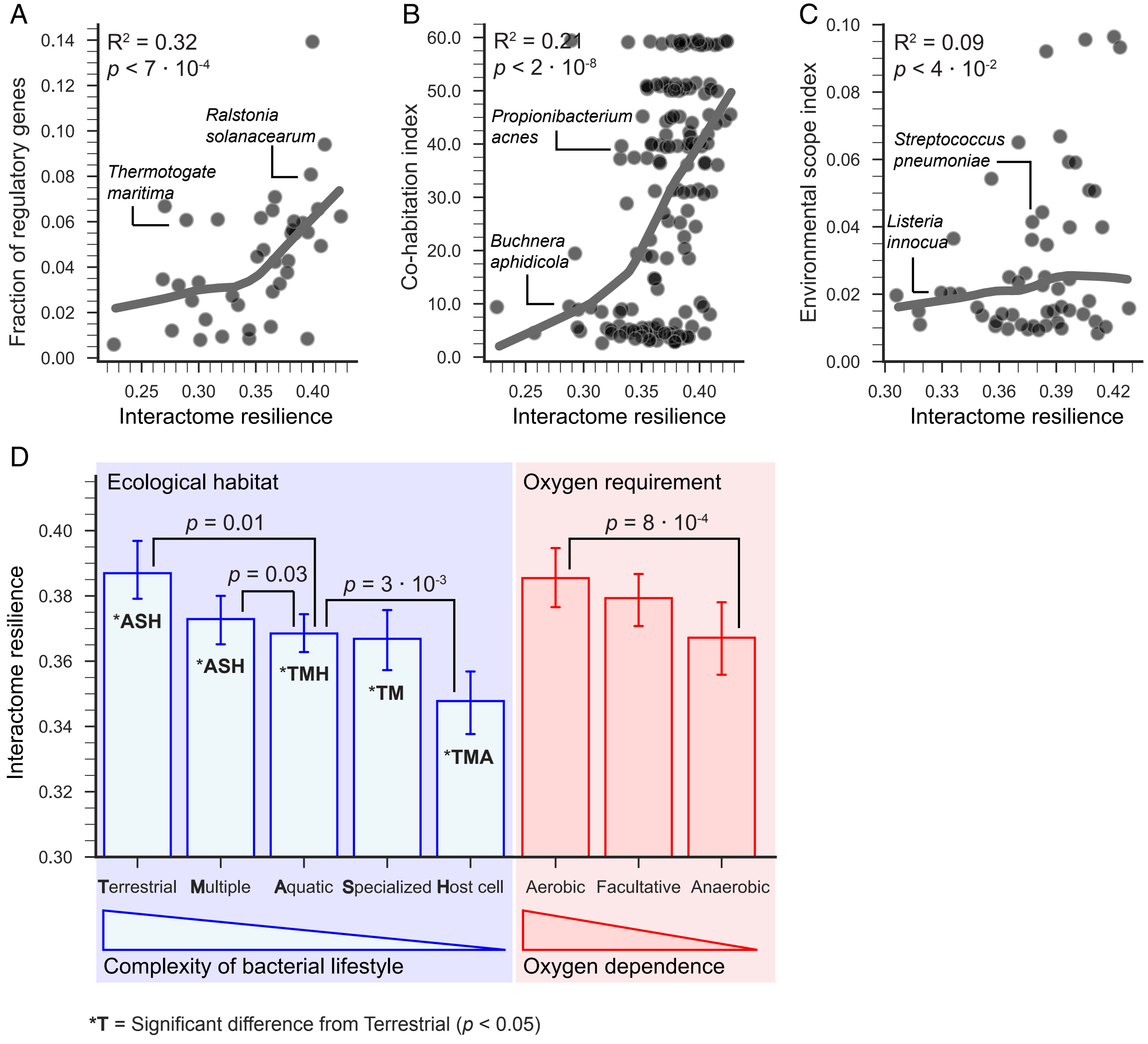

Fig. 2. Bacteria with more resilient interactomes survive in more complex, variable, and competitive environments. We use ecological information for 287 bacterial species (32) to examine the relationship between species' interactome resilience and their ecology (*SI Appendix*, section S4). (*A*) Interactome resilience positively correlates with the fraction of regulatory genes in bacteria, an established indicator of environmental variability of species' habitats (32) ($R^2 = 0.32$). (*B* and *C*) For environmental viability of a species, we use a cohabitation index that records how many organisms populate each environment in which the species is viable (i.e., the level of competition in each viable environment) and an environmental scope index that records a fraction of the environments in which the species is viable (i.e., species' environmental diversity) (32). The resilience of the interactome positively correlates with the level of cohabitation encountered by bacteria ($R^2 = 0.21$), and bacteria with resilient interactomes tend to thrive in highly diverse environments ($R^2 = 0.09$). (*D*) Terrestrial bacteria have the most resilient interactomes ($P = 7 \cdot 10^{-3}$), and host-associated bacteria have the least resilient interactomes ($P = 4 \cdot 10^{-6}$). In bacteria, interactome resilience is indicative of oxygen dependence. Aerobic bacteria have the most resilient interactomes ($P = 8 \cdot 10^{-4}$), followed by facultative and the anaerobic bacteria. Error bars indicate 95% bootstrap confidence interval; *P* values denote the significance of the difference of the means according to a Mann–Whitney *U* test.

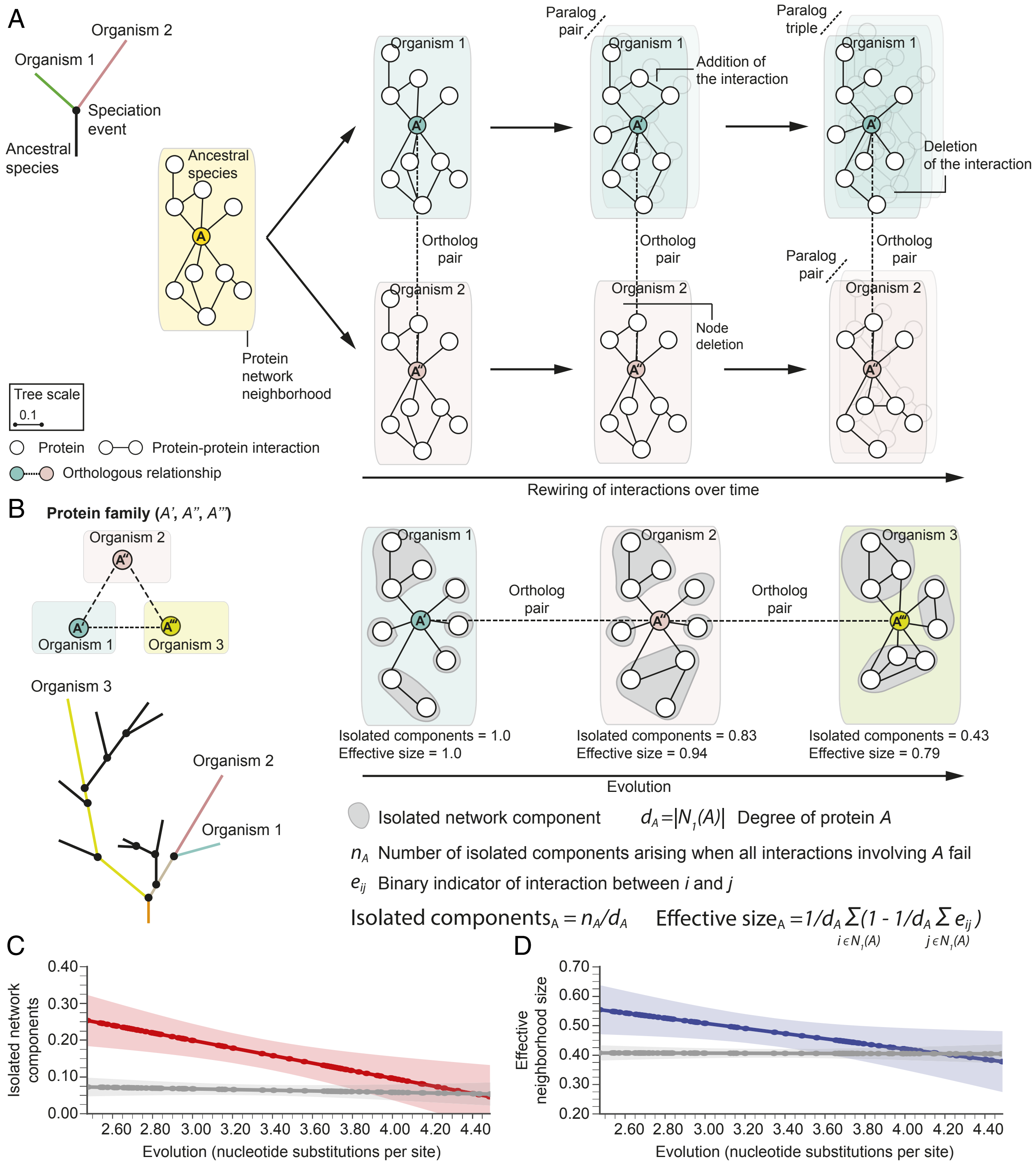
A
Organism 2
Organism 1
Speciation event
Ancestral species
Ancestral species
A
Protein network neighborhood
Tree scale
0.1
Protein
Protein-protein interaction
Orthologous relationship
Organism 1
A′
Ortholog pair
Organism 2
A″
Paralog pair
Addition of the interaction
Node deletion
Paralog triple
Deletion of the interaction
Paralog pair
Rewiring of interactions over time
B
Protein family (A′, A″, A‴)
Organism 2
Organism 1
Organism 3
Organism 3
Organism 2
Organism 1
Ortholog pair
Ortholog pair
Organism 3
Isolated components = 1.0
Effective size = 1.0
Isolated components = 0.83
Effective size = 0.94
Isolated components = 0.43
Effective size = 0.79
Evolution
Isolated network component
$d_A=|N_1(A)|$ Degree of protein A
n_A Number of isolated components arising when all interactions involving A fail
e_{ij} Binary indicator of interaction between i and j
Isolated components$_A = n_A/d_A$
Effective size$_A = 1/d_A \sum_{i\in N_1(A)} (1 - 1/d_A \sum_{j\in N_1(A)} e_{ij})$
C
Isolated network components
0.00
0.10
0.20
0.30
0.40
2.60
2.80
3.00
3.20
3.40
3.60
3.80
4.00
4.20
4.40
Evolution (nucleotide substitutions per site)
D
Effective neighborhood size
0.20
0.30
0.40
0.50
0.60
0.70
2.60
2.80
3.00
3.20
3.40
3.60
3.80
4.00
4.20
4.40
Evolution (nucleotide substitutions per site)

Fig. 3. Evolution mitigates local network structural changes in protein interactomes. (*A*) A hypothetical phylogenetic tree illustrates a speciation event that gives rise to two lineages according to the speciation-divergence model (11) and leads to present-day organisms "1" (green) and "2" (pink). In this example, a single ancestral protein *A* that was present in the ancestral species gives rise to proteins A' and A'' upon speciation; A' and A'' form an orthologous protein pair. As the two newly arising species diverge and protein sequences evolve, protein network neighborhoods (*SI Appendix*, Fig. S4) in their interactomes can rewire independently over time. Shown are also in-paralogs, proteins which arise through gene duplication events in species 1 and 2 after speciation. (*B*) A hypothetical protein family with three protein members (A', A'', A'''), each from a different organism. In the phylogenetic tree, organism 1 is located at the tip of the lineage with the shortest branch length, whereas organism 3 is in the lineage with the longest branch length in the tree. We represent the protein family by a sequence of orthologous proteins ordered by the branch length of proteins' originating species (*SI Appendix*, section S3). We then characterize the network neighborhood of each protein in the family by calculating two network metrics (*SI Appendix*, Fig. S5). Isolated components are given by the degree-adjusted number of connected components in the neighborhood that arise when the central protein is removed from the interactome (gray) (*SI Appendix*, section S6). The neighborhood size down-weighted by the redundancy of local interactions gives the effective size of the neighborhood (*SI Appendix*, section S6). (*C* and *D*) The number of isolated network components and the effective size of protein neighborhoods both decrease with evolution ($P = 3 \cdot 10^{-8}$ and $P = 0.03$, respectively; Spearman's ρ rank correlation), suggesting that local interaction neighborhoods rewire via a coordinated evolutionary mechanism. Lines in *C* and *D* show the LOWESS fit of median-aggregated network metric values for 81,673 proteins from 2,224 protein families; color bands indicate 95% confidence band for the LOWESS fit; gray lines show random expectation.

exceptionally strong associations between the resilience and the level of cohabitation and the environmental scope (Fig. 2 *B* and *C*). Furthermore, using a categorization of bacteria into five groups based on their natural environments [NCBI classification for bacterial lifestyle (terrestrial, multiple, host cell, aquatic, specialized) ordered by the complexity of each category (32)], we find that terrestrial bacteria living in the most complex ecological habitats have the most resilient interactomes ($P = 7 \cdot 10^{-3}$; Mann–Whitney *U* test) and that host-associated bacteria have the least resilient interactomes ($P = 4 \cdot 10^{-6}$; Fig. 2*D*). Our analysis further reveals that interactome resilience is indicative of oxygen dependence; the most resilient interactomes are those of aerobic bacteria ($P = 8 \cdot 10^{-4}$), followed by facultative and then anaerobic bacteria, which do not require oxygen for growth (Fig. 2*D*).

These relationships suggest that molecular mechanisms that render a species' interactome more resilient might also allow it to cope better with environmental challenges. In the network context, high interactome resilience suggests that proteins can interact with each other even in the face of high protein failure rate. High interactome resilience indicates that a species has a robust interactome, in which many mutations represent network failures that are neutral in a given environment, have no phenotypic effect on the network's function, and are thus invisible to natural selection (26). However, neutral mutations may not remain neutral indefinitely, and a once-neutral mutation may have phenotypic effects in a changed environment and be important for evolutionary innovation (25). Although a large number of mutations in a resilient interactome might not change the network's primary function, they might alter other network features, which can drive future adaptations as the environment of the species changes (33). Changes that are neutral concerning one aspect of the network's function could lead to innovation in other aspects, suggesting that a resilient interactome can harbor a vast reservoir of neutral mutations.

Structural Changes of Protein Network Neighborhoods. A resilient interactome may arise through changes in the network structure of individual proteins over time (Fig. 3*A*). To investigate such changes in local protein neighborhoods, we decompose species interactomes into local protein networks, using a 2-hop subnetwork centered around each protein in a given species as a local representation of a protein's direct and nearby interactions in the species' interactome (*SI Appendix*, section S6.1 and Fig. S4). We obtain 81,673 protein network neighborhoods and then use orthologous relationships between proteins to group them into 2,224 protein families, with an average of 38 protein neighborhoods originating from 12 species in each family (*SI Appendix*, section S3). Each family represents a group of orthologous proteins that share a common ancestral protein (Fig. 3*B*).

By examining protein families, we find that the number of isolated network components in protein network neighborhoods and the effective size of the neighborhoods (Fig. 3*B* and *SI Appendix*, section S6.2) both decrease with evolution ($P = 3 \cdot 10^{-8}$ and $P = 0.03$, respectively; Fig. 3 *C* and *D*), indicating that protein neighborhoods become more connected during

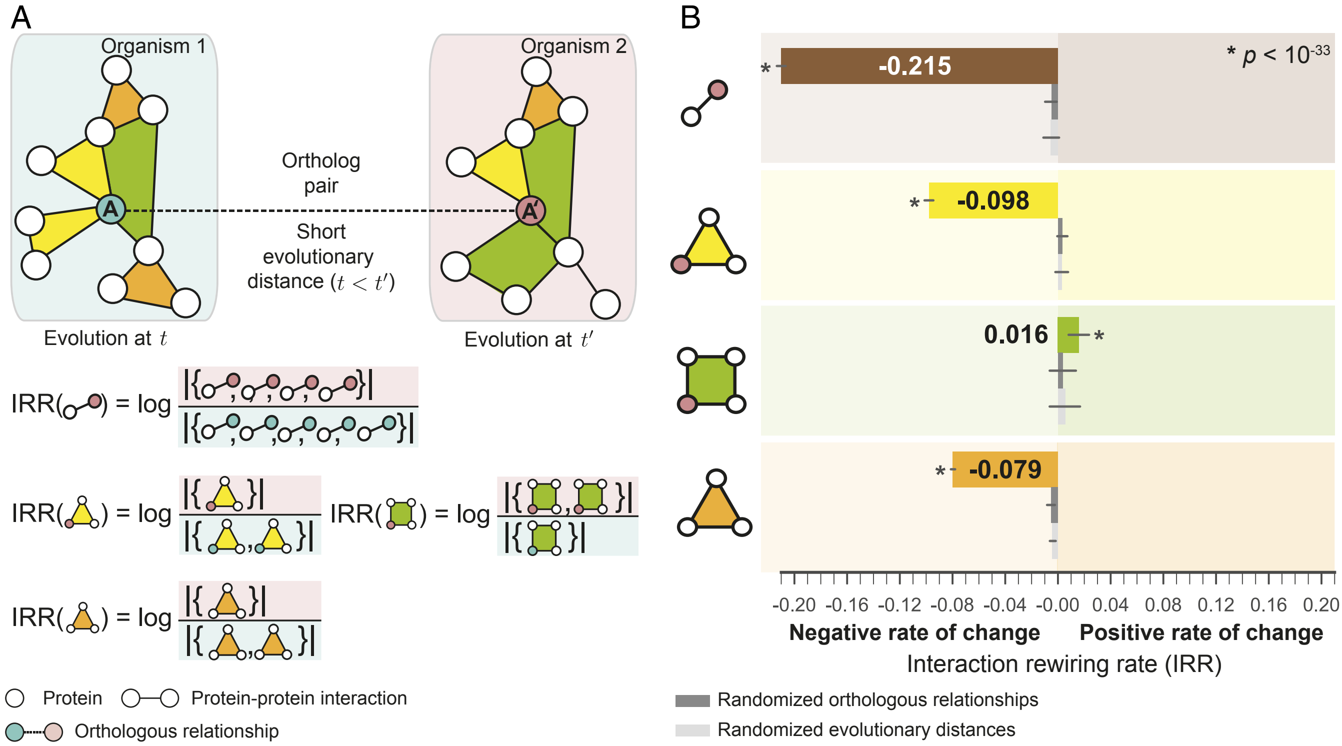

Fig. 4. The rewiring rate of interactions in local protein neighborhoods varies with the topology of network motifs. (*A*) Interaction rewiring rate (IRR) measures the fold change between the probability of observing a particular network motif in the network neighborhood of protein *A*′ and the probability of observing the same motif in the neighborhood of an evolutionarily younger orthologous protein *A*. A positive (negative) rate of change indicates the motif becomes more (less) common over time (*SI Appendix*, section S7). Shown are the rewiring rates for interactions (i.e., edges; the number of interactors of *A*′ vs. *A*), triangle motifs touching the orthologous protein (yellow), square motifs touching the orthologous protein (green), and triangle motifs in the protein network neighborhood (orange). (*B*) Square motifs become more common in protein neighborhoods during evolution ($P < 10^{-33}$), which is supported by a range of biological evidence (18, 37, 38). However, triangle motifs become less common over time ($P < 10^{-33}$ for both types of triangle motifs). Gray bars indicate random expectation (*SI Appendix*, section S7), either for random orthologous relationships (dark gray) or for random evolutionary distances (light gray); error bars indicate 95% bootstrap confidence interval; and *P* values denote the significance of the difference of IRR distributions using a two-sample Kolmogorov–Smirnov test.

evolution. These structural changes in the neighborhoods suggest a molecular network model of evolution (Fig. 3*B*): For orthologous proteins in two species, as the evolutionary distance between the species increases, the proteins' local network neighborhoods become increasingly different and the neighborhood becomes more interconnected in the species that has undergone more genetic change.

Network Rewiring of Protein–Protein Interactions. To study evolutionary mechanisms of structural changes in the interactomes, we investigate network motifs (34, 35). We first identify orthologous protein pairs from evolutionarily close species (*SI Appendix*, section S3), resulting in 2,485,564 protein pairs, which we then use to calculate interaction rewiring rates (IRRs) for selected network motifs (Fig. 4*A*). We calculate the number of times each motif appears in each protein neighborhood and derive the IRR by comparing the motif occurrences between the interactomes of the older and the younger species of each protein pair (*SI Appendix*, section S7.1). We find strong statistical evidence that network motifs rewire during evolution ($P < 10^{-33}$ for all network motifs; Fig. 4*B*), suggesting that rewiring of interactions is an important mechanism for the evolution of interactomes. For example, proteins in evolutionarily older species on average participate in a factor of 0.861 fewer protein–protein interactions compared with proteins in evolutionarily younger species ($\mathrm{IRR} = -0.215$; Fig. 4*B*). This significant negative correlation between a protein's number of interactions and the protein's evolutionary age confirms earlier studies of *Saccharomyces cerevisiae* (36). We also find that square motifs of interactions become more common in protein neighborhoods during evolution ($\mathrm{IRR} = 0.016$; Fig. 4*B*). A range of biological evidence (18, 37, 38) supports this positive rate of change in the number of square motifs: From a structural perspective (38), protein–protein interactions often require complementary interfaces; hence two proteins with similar interfaces share many of their neighbors. However, they might not interact directly with each other, which manifests in the interactome as a square motif of interactions (see *SI Appendix*, Fig. S6 for an illustration of interaction interfaces recognizing the binding sites in proteins). Evolutionary arguments following gene duplication (18) reach the same conclusion; proteins with multiple shared interaction partners are likely to share even more partners and thereby produce new square motifs of interactions. To test the predictive power of our motif-based model of structural network changes, we estimate the size of the whole human interactome by extrapolating the *S. cerevisiae* interactome, using IRRs from Fig. 4*B* (*SI Appendix*, section S7.3). Assuming one splice isoform per gene, we predict the number of interactions in humans to be $\sim$160,000. This prediction is in surprisingly good agreement with three previous estimates of the size of the human interactome, which range from 150,000 to 370,000 interactions (15, 17, 39) and have proved crucial in establishing the complexity of the human interactome (19).

Discussion

Our analyses reveal how protein–protein interaction networks change through evolution and how changes in these networks affect phenotypes and organismal response to environmental complexity. This systematic investigation of protein–protein interaction networks from an evolutionary perspective was enabled by a dataset of interactomes, consisting of protein–protein interaction networks from 1,840 species. To date, most evolutionary analyses of biological networks have focused on a small number of organisms with high-coverage protein–protein interaction data, such as *S. cerevisiae*, *Mus musculus*, and humans. This is because interactomes mapped by unbiased tests of all possible pairwise combinations of proteins on the same platform remain scarce, an important limitation of the present study. Furthermore, experimentally documented protein interactions are currently subject to a high number of false positives and negatives. As more protein interaction data are collected, and more genomes become available, the generalizability of our findings can be further evaluated. However, our results are consistent across both different subsets of protein interaction data (*SI Appendix*, Table S2) and different phylogenetic lineages (*SI Appendix*, Fig. S10) and are not explained by many possible genomic and network confounders (*SI Appendix*, section S8, Fig. S8, and Table S1), thus providing confidence that our key findings cannot be attributed to biases in the datasets.

Interactome resilience is an important aspect of our study. The resilience measures fragmentation of the interactome into isolated components and thus represents a global measure of the interactome's topological stability. Beyond fragmentation, there are other possible modifications of the interactome that could alter the network's biological function without necessarily disconnecting the network (40–42). As more detailed information about functions of individual proteins in the interactome (43), as well as dynamic protein-expression data (44), becomes available, our measure of interactome resilience could be adapted to give a more complex definition of resiliency, which might yield more detailed evolutionary predictions. Additionally, information on how protein–protein interactions change dynamically both in time and space (45–47) might reveal how topological stability of the interactome depends on large-scale interactome connectivity as well as on the interactome's dynamic properties (40).

Our study presents an additional paradigm for evolutionary studies by demonstrating that interactomes reveal fundamental structural principles of molecular networks. Our findings highlight evolution as an important predictor of structural network change and show that evolution of a species predicts resilience of the species' interactome to protein failures. The findings offer quantitative evidence for the biological proposition that an organism that has undergone more genetic change has a more resilient interactome, which, in turn, is associated with the greater ability of the organism to survive in a more complex, variable, or competitive environment. Our findings can also help clarify the mechanisms of how interactomes change during evolution, why currently observed network structures exist, and how they may change in the future and facilitate the extrapolation of functional information from experimentally characterized proteins to their orthologous proteins in poorly studied organisms.

Materials and Methods

Detailed description of data, statistical methodology, and additional analyses are provided in *SI Appendix*.

Code and Data Availability. Software implementation of statistical methodology is publicly available at snap.stanford.edu/tree-of-life. All data used in this paper, including the processed interactomes, are shared with the community and available from snap.stanford.edu/tree-of-life.

ACKNOWLEDGMENTS. For helpful discussion and comments on the manuscript, we thank Aaron Goodman, Hunter B. Fraser, Alice Y. Ting, Emma Pierson, Michael An, and Pratyaksh Sharma. M.Z., R.S., and J.L. were supported in part by NSF, NIH, Defense Advanced Research Projects Agency, Boeing, Stanford Data Science Initiative, and Chan Zuckerberg Biohub. M.W.F. was supported in part by the Stanford Center for Computational, Evolutionary, and Human Genomics and a grant from the John Templeton Foundation.

1. Castelle CJ, Banfield JF (2018) Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* 172:1181–1197.
2. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.
3. Hu JX, Thomas CE, Brunak S (2016) Network biology concepts in complex disease comorbidities. *Nat Rev Genet* 17:615–629.
4. Huttlin EL, et al. (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature* 545:505–509.

5. Chen S, et al. (2018) An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders. *Nat Genet* 50:1032–1040.
6. Britten RJ (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231:1393–1398.
7. Yang Z, Rannala B (2012) Molecular phylogenetics: Principles and practice. *Nat Rev Genet* 13:303–314.
8. Marsit S, et al. (2017) Evolutionary biology through the lens of budding yeast comparative genomics. *Nat Rev Genet* 18:581–598.
9. Studer RA, et al. (2016) Evolution of protein phosphorylation across 18 fungal species. *Science* 354:229–232.
10. Sorrells TR, Booth LN, Tuch BB, Johnson AD (2015) Intersecting transcription networks constrain gene regulatory evolution. *Nature* 523:361–365.
11. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
12. Hug LA, et al. (2016) A new view of the tree of life. *Nat Microbiol* 1:16048.
13. Parks DH, et al. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2:1533–1542.
14. Mukherjee S, et al. (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* 35:676–683.
15. Rual JF, et al. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437:1173–1178.
16. Roguev A, et al. (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 322:405–410.
17. Venkatesan K, et al. (2009) An empirical framework for binary interactome mapping. *Nat Methods* 6:83–90.
18. Arabidopsis Interactome Mapping Consortium, et al. (2011) Evidence for network evolution in an Arabidopsis interactome map. *Science* 333:601–607.
19. Rolland T, et al. (2014) A proteome-scale map of the human interactome network. *Cell* 159:1212–1226.
20. Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ed Grossman RL (ACM, New York), pp 177–187.
21. Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. *ACM Trans Knowl Discovery Data* 1:1–41.
22. Zitnik M, et al. (2019) Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf Fusion* 50:71–91.
23. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382.
24. Wagner A (2003) Does selection mold molecular networks? *Sci Signaling* 2003:pe41.
25. Wagner A (2005) Robustness, evolvability, and neutrality. *FEBS Lett* 579:1772–1778.
26. Wagner A (2013) *Robustness and Evolvability in Living Systems* (Princeton Univ Press, Princeton).
27. Vo TV, et al. (2016) A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell* 164:310–323.
28. Sheldon AL (1969) Equitability indices: Dependence on the species count. *Ecology* 50:466–467.
29. MaGuarran A (1988) *Ecological Diversity and Its Measurement* (Princeton Univ Press, Princeton).
30. Goodwin SB, Spielman L, Matuszak J, Bergeron S, Fry W (1992) Clonal diversity and genetic differentiation of Phytophthora infestans populations in northern and central Mexico. *Phytopathology* 82:955–961.
31. Baczkowski A, Joanes D, Shamia G (1997) Properties of a generalized diversity index. *J Theor Biol* 188:207–213.
32. Freilich S, et al. (2009) Metabolic-network-driven analysis of bacterial ecological strategies. *Genome Biol* 10:R61.
33. Barve A, Wagner A (2013) A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* 500:203–206.
34. Benson AR, Gleich DF, Leskovec J (2016) Higher-order organization of complex networks. *Science* 353:163–166.
35. Yin H, Benson AR, Leskovec J (2018) Higher-order clustering in networks. *Phys Rev E* 97:052306.
36. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296:750–752.
37. Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res* 33:3629–3635.
38. Keskin O, Tuncbag N, Gursoy A (2016) Predicting protein–protein interactions from the molecular to the proteome level. *Chem Rev* 116:4884–4909.
39. Hart GT, Ramani AK, Marcotte EM (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* 7:120.
40. Siegal ML, Promislow DE, Bergman A (2007) Functional and evolutionary inference in gene networks: Does topology matter? *Genetica* 129:83–103.
41. Leclerc RD (2008) Survival of the sparsest: Robust gene networks are parsimonious. *Mol Syst Biol* 4:213.
42. Masel J, Siegal ML (2009) Robustness: Mechanisms and consequences. *Trends Genet* 25:395–403.
43. Agrawal M, Zitnik M, Leskovec J (2018) Large-scale analysis of disease pathways in the human interactome. *Pacific Symposium on Biocomputing*, eds Altman RB, Dunker AK, Hunter L, Ritchie MD, Klein TE (World Scientific, Singapore), Vol 23, pp 111–122.
44. Deng Q, Ramsköld D, Reinius B, Sandberg R (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343:193–196.
45. Han JDJ, et al. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430:88–93.
46. Smith C, Puzio RS, Bergman A (2014) Hierarchical network structure promotes dynamical robustness. arXiv:1412.0709. Preprint, posted December 1, 2014.
47. Smith C, Pechuan X, Puzio RS, Biro D, Bergman A (2015) Potential unsatisfiability of cyclic constraints on stochastic biological networks biases selection towards hierarchical architectures. *J R Soc Interface* 12:20150179.