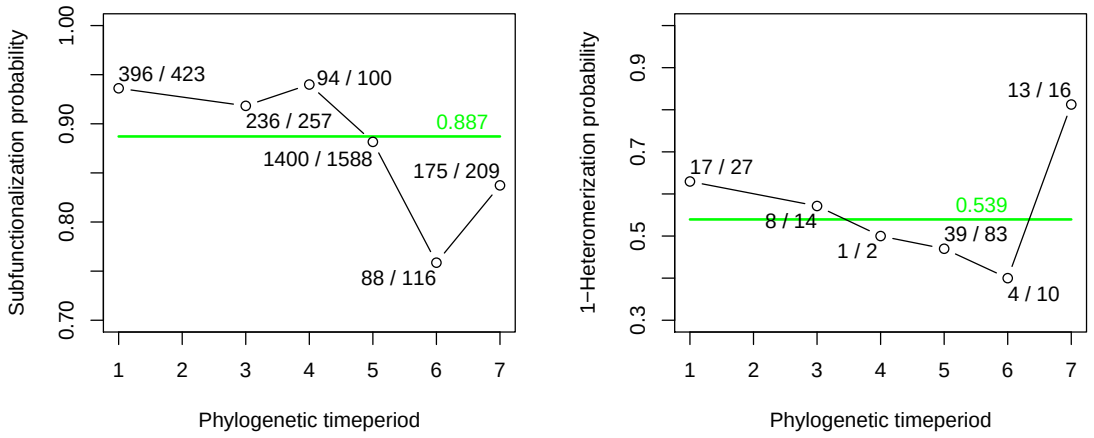
SUPPLEMENTARY MATERIAL
Improving evolutionary models of protein interaction networks

Todd A. Gibson and Debra S. Goldberg

Contents

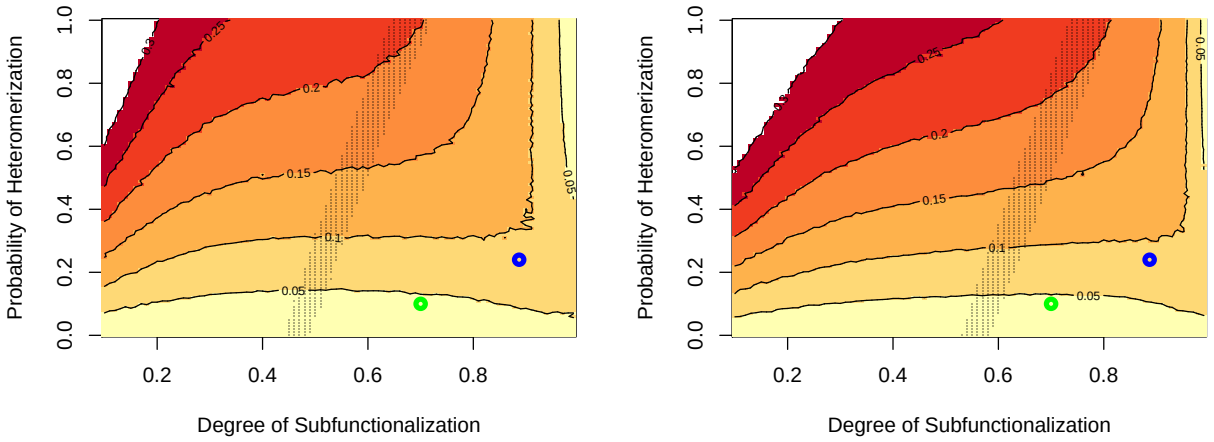
1 Supplementary Figures

These figures were specifically referenced in the main text. Additional figures found elsewhere in this supplementary document appear where they are referenced in the supplementary text.

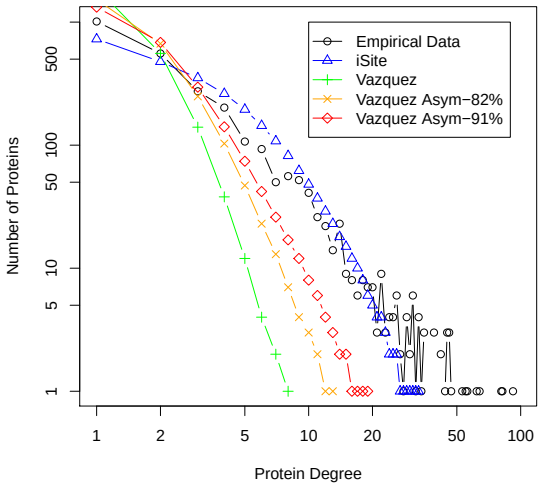


(A) **(B)**

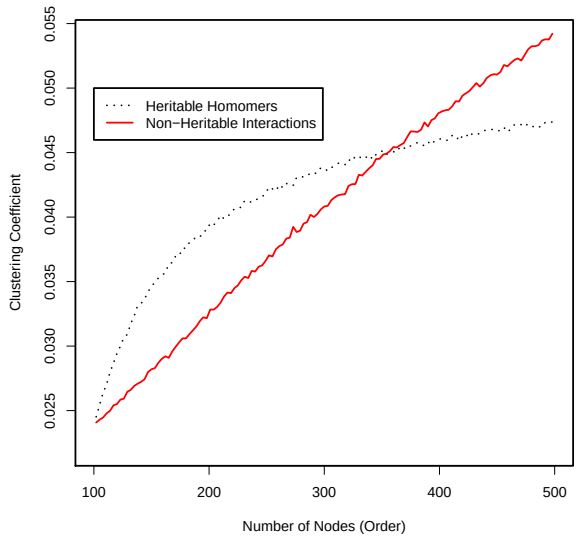
Supplementary Figure 1: Subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 293 duplication events across seven phylogenetic time periods for the empirical data set. The ratios adjacent to each data point are the number of interactions lost versus the total number of interactions added from duplication events. The large number of interactions shown in time period 5 is due to the whole-genome duplication event in the yeast lineage. The green line is the mean probability. Plots for alternative data sets are shown in in Supplementary Figures 8—13. (A) The loss rate of redundant interactions between a duplicate protein pair and their neighbors. (B) The complement of the heteromerization rate. That is, the loss rate of heteromeric interactions which form due to duplication of a self-interacting protein.



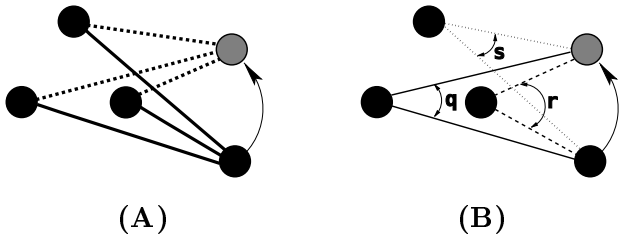
Supplementary Figure 2: The clustering landscape of Vázquez et al. Each data point is the mean clustering coefficient for 100 runs of the model at parameter values along the corresponding axes. Each network was constructed to 2647 nodes, the number of nodes in the empirical data (Table 1, main text). The green point identifies the parameter values derived by Vázquez et al. (2003) (see Supplementary Materials). The blue point identifies the parameter values derived from our analysis of the empirical data (Methods, main text). The gray band covers those parameter values which generate network sizes (i.e., number of interactions) within 10% of 5449, the size of the empirical network. (A) The published Vázquez model (i.e., with symmetric subfunctionalization). Reprinted here from the main text. (B) The Vázquez model at 82% asymmetric subfunctionalization, the asymmetry value derived from the empirical data.



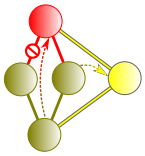
Supplementary Figure 3: The degree distribution for the empirical data set, the iSite model, and the Vázquez et al. model and its variants. The log scale is used on the both axes.



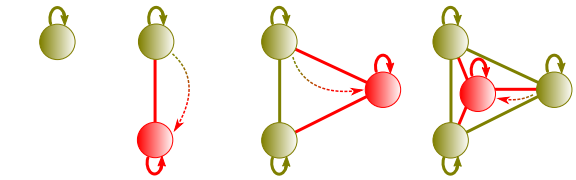
Supplementary Figure 4: Original Vázquez versus heritable homomers. Heritable homomers produce a larger clustering coefficient early on when many homomeric interactions produce many surviving paralogous interactions. As the network grows, the fixed rate heteromerization parameter of the original Vázquez model produces a larger number of paralogous interactions. The clustering coefficient of the original model overtakes the heritable model after approximately 250 duplications (i.e., at approximately 350 nodes). The data for the plot was generated by growing each model 10^4 times and measuring the clustering coefficient after each duplication. The values were averaged to generate the plot. Model parameters were assigned as specified elsewhere in these Supplementary Materials.



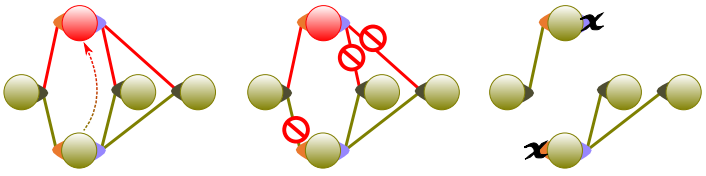
Supplementary Figure 5: Asymmetry in redundant interaction loss. (A) Asymmetric subfunctionalization in Solé et al. (2002). Lost redundant interactions are selected only from the progeny (the dashed lines). (B) Symmetric subfunctionalization in Vázquez, et al.(2003). Lost redundant interactions are equiprobably selected from either member of each pair q, r, and s.



Supplementary Figure 6: Determining the subfunctionalization rate. The subfunctionalization rate is affected by the order in which duplication events are selected within each phylogenetic node. Shown are two duplication events. Subfunctionalization has claimed one redundant interaction from the red duplication event. No redundant interactions have been lost from the yellow duplication. If the red duplication is measured first, one of two redundant edges are lost, accounting for a subfunctionalization rate of 0.5. If the yellow duplication is measured first, then the red duplication has lost one of three redundant interactions. The network subfunctionalization rate is calculated as 0.25 (1 of 4 new interactions lost), and avoids the order sensitivity intrinsic to the pairwise measure.



Supplementary Figure 7: Cliques of paralogous interactions. (A) A self-interacting protein. (B) After one duplication, the paralogs interact. (C) After a second duplication, a 3-protein clique has formed–every protein interacts with all of its paralogs. (D) After a third duplication, a 4-protein clique has formed.



Supplementary Figure 8: Multiple-iSite degeneration during subfunctionalization. (A) Duplication of a two-iSite protein. Each redundant iSite pair is grouped by color in the illustration. (B) One iSite from each redundant pair is selected to be subfunctionalized. Given the high rate of interaction loss, it is not uncommon for the chosen iSite to lose all interactions. All interactions of the first iSite in the progenitor may be silenced, and all interactions of the second iSite of the progeny may be silenced. (C) The formerly two-iSite protein has effectively been reduced to two, single-iSite offspring.

2 Supplementary Methods

2.1 Construction of clustering landscape

Each data point in the Vázquez et al. model landscape (Figure 2, main text) is the mean clustering coefficient from 100 runs of the model for the parameter values on the corresponding axes. The parameter values for the heteromerization probability, p, and the subfunctionalization (redundant edge loss) probability, q, were in the range of $p = [0.0, 1.0]$, and $q = [0.1, 1.0]$ at intervals of .01. Parameter values $q = [0.0, 0.1)$ were not used due to computation time and/or overflow errors. These subfunctionalization values are not biologically plausible and can therefore be safely omitted. Each of the networks were generated using the Vázquez et al. model from a 100-node Erdős-Renyí seed graph to 2647 nodes, the number of nodes in the empirical interaction set (shown in Table 1 in the main text).

2.2 Evolutionary framework initiation

The framework introduced in Gibson and Goldberg (2009b) was used to determine subfunctionalization and paralogous edge retention rates. The January, 2009 update of the Wapinski, et al. (Wapinski *et al.*, 2007) data was used as the phylogenetic gene tree input along with the interaction data described here. served as the protein interaction data input. Seven phylogenetic time periods were identified in the framework (Supplemental Figure 1). The alternative interaction data sets featured in the Supplementary Materials were used with the same tree data.

2.3 Calculation of the subfunctionalization rate

For each time period, the framework in Gibson and Goldberg (2009b) identifies three states:

- **pre** before processing any duplications for the time period.
- **expanded** after processing all duplications, but before processing interaction losses.
- **post** after processing redundant interaction losses.

For the number of interactions, m, the subfunctionalization rate for time period, t, can be calculated as:

$$\frac{\text{lost interactions}}{\text{potential new interactions}} = \frac{m_{t_{\text{expanded}}} - m_{t_{\text{post}}}}{m_{t_{\text{expanded}}} - m_{t_{\text{pre}}}}$$

For the entire evolutionary history captured by the framework, the subfunctionalization rate is calculated as:

$$\text{subfunctionalization rate} = \frac{\sum_t \left(m_{t_{\text{expanded}}} - m_{t_{\text{post}}} \right)}{\sum_t \left(m_{t_{\text{expanded}}} - m_{t_{\text{pre}}} \right)}$$

2.4 Calculation of the heritable heteromerization rate

When a self-interacting protein's gene duplicates, the progeny and progenitor are self-interacting, and they interact with each other. We refer to the interaction between progeny and progenitor as heteromerization. If, within a time period, a self-interacting gene is duplicated more than once, the

heteromerizing interactions form a clique (Supplementary Figure 7). For a gene family, f, having d duplication events within a time period, the heteromerization rate for each time period t is:

$$\text{heteromerization rate}_t = 1 - \frac{\sum_f s_{tf_{\text{lost}}}}{\sum_f \frac{d_{tf}(d_{tf}+1)}{2}}$$

$s_{tf_{\text{lost}}}$ are the number of heteromerizing interactions lost after duplication.

As for the subfunctionalization rate, the network heteromerization rate is calculated by summing numerator for all time periods divided by a sum of the denominator for all time periods.

We avoided creating per-time period rates due to the high variance–some time periods have very few interactions from which to calculate a rate. We believe that despite the decreased resolution, aggregating the time period data leads to a more consistent rate.

2.5 Calculation of subfunctionalization asymmetry

Asymmetry in the empirical network can be calculated as follows. The number of distinct neighbors for the progenitor and progeny proteins are given as g_1 and g_2 respectively. The asymmetry is simply the larger of g_1 and g_2 divided by their sum. As with the model parameters, the evolutionary framework was used to calculate asymmetry in the empirical network. This produces an asymmetry of $1554/1905 = 82\%$.

2.6 Empirically-derived heteromerization rate, p, for the Vázquez et al. model and non-heritable homomers variants.

The empirically-calculated rate of $1 - 0.539 = 0.461$ is the paralogous interaction (survival) rate *among homomeric protein duplicates*. The Vázquez et al. model parameter is the paralogous interaction (add) rate *among all duplications*. Among 293 duplications in the evolutionary framework, 70 paralogous interactions survived. This corresponds to a $p = 0.24$ probability of a paralogous interaction per duplication used in the Vázquez et al. model and non-heritable homomer variants.

2.7 Seed graph construction

Each theoretical model evolves from a seed graph. All theoretical models in the main text were constructed from the same seed graph. The seed graph was constructed as a 100-node Erdős-Renyí random graph (with probability 0.025 of connecting each pair of nodes). For the Heritable Vázquez and iSite models, self loops were added to 83 of the nodes.

The number of self loops (83) was determined as follows. The empirical network consists of 2647 proteins, 412 of which self-interact. Random pairs of the 2647 nodes were selected and combined. If either of the pair were self-interacting then the combined node was made self-interacting. The process was repeated until the network was reduced to 100 nodes. A tally of the number of self-interacting nodes among the 100 was recorded. This process was repeated 10,000 times to calculate a mean 83 self-interacting nodes.

The Vázquez et al. model along with many other models propose starting with seed graphs consisting of only a few nodes. A 100-node seed graph was selected for this study due to the requirement of propagating heritable homomers. As explained above, 83 percent of the nodes are

homomeric in a 100-node seed graph. Homomeric interactions would saturate smaller seed graphs and make it impossible for the model to propagate them in large numbers through evolution using the estimated parameter values.

2.8 iSite model initialization

Each run of the iSite model evolves from a seed graph (described above) to 2647 nodes. Each protein starts with a number of iSites equal to the number of interacting neighbors.

2.9 Calculation of random equivalent networks

Table 1 in the main text includes a clustering coefficient of networks randomly-equivalent to each empirical or model network. Each random equivalent network was generated by rewiring interactions while preserving the degree distribution (Milo *et al.*, 2004). At each iteration a pair of edges were selected at random and one end from each edge was swapped. If the swap created a duplicate edge or a self-interaction the swap was aborted and the next iteration begun. The number of iterations performed was 100E where E is the number of edges in the network.

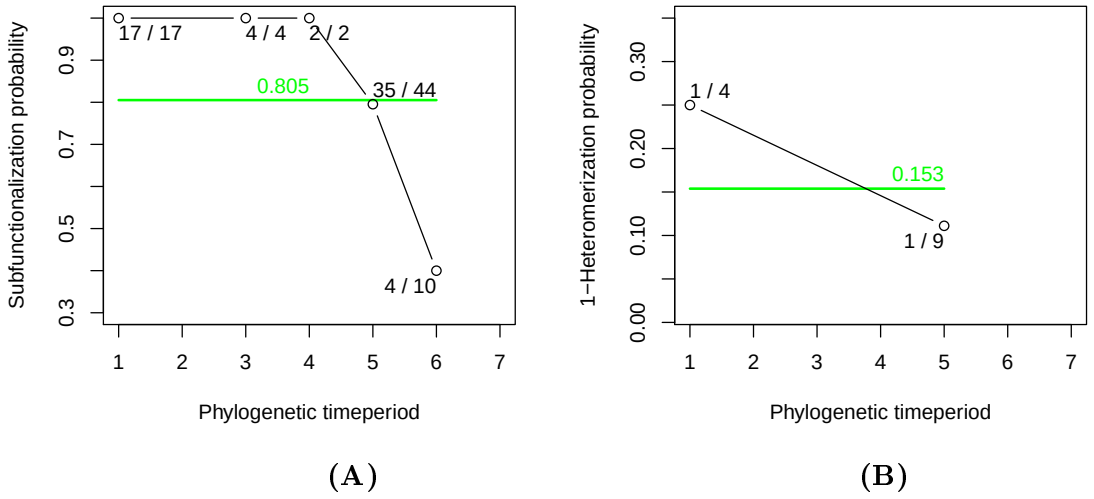
3 The evolutionary mechanics of Vázquez et al. (2003)

Starting from a seed graph, the Vázquez et al. model iteratively follows a "duplication and divergence" process. An existing protein (node) in the network is randomly selected as a progenitor. A progeny node is created and an interaction (edge) is formed between the progenitor and progeny with probability p. This probability represents interactions formed via duplication of a homomeric protein and is independent of previous duplications–it is not a heritable trait.

Additionally, an interaction is added from the progeny to each of the interacting neighbors of the progenitor. The progenitor and progeny proteins now have identical interacting neighbors; the pair of interactions to each neighbor are putatively redundant.

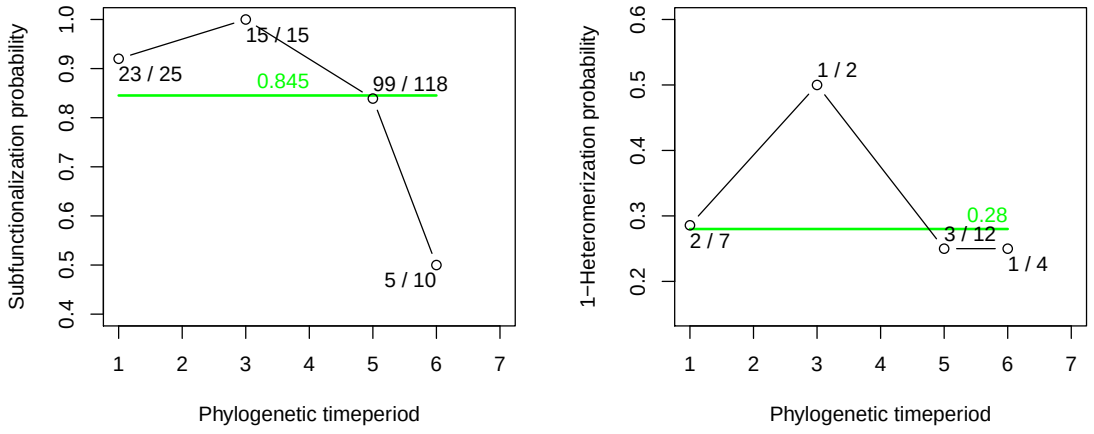
Following duplication is divergence. Each neighbor loses one of the pair of redundant interactions with probability q.

4 Additional parameter rate estimates



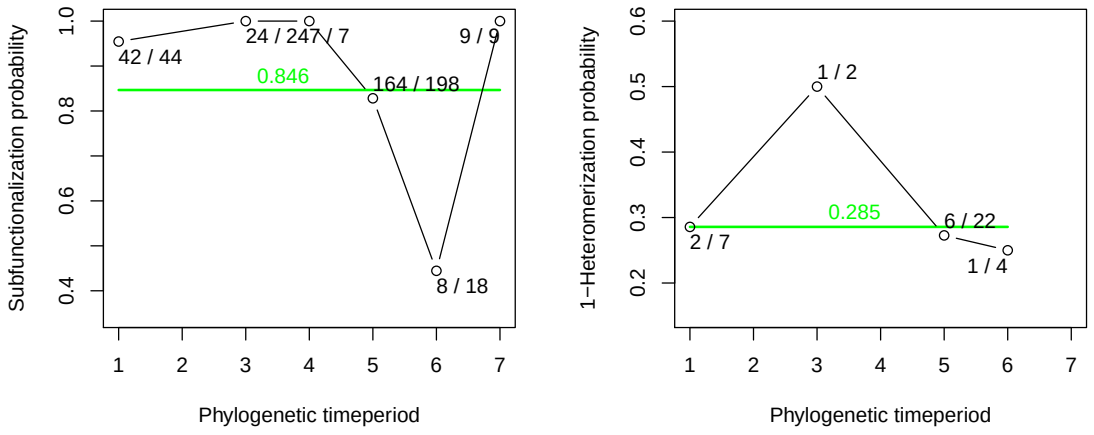
(A) **(B)**

Supplementary Figure 9: Uetz et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 31 duplication events across seven phylogenetic time periods. Interpretation of plots is as per Supplementary Figure 1.



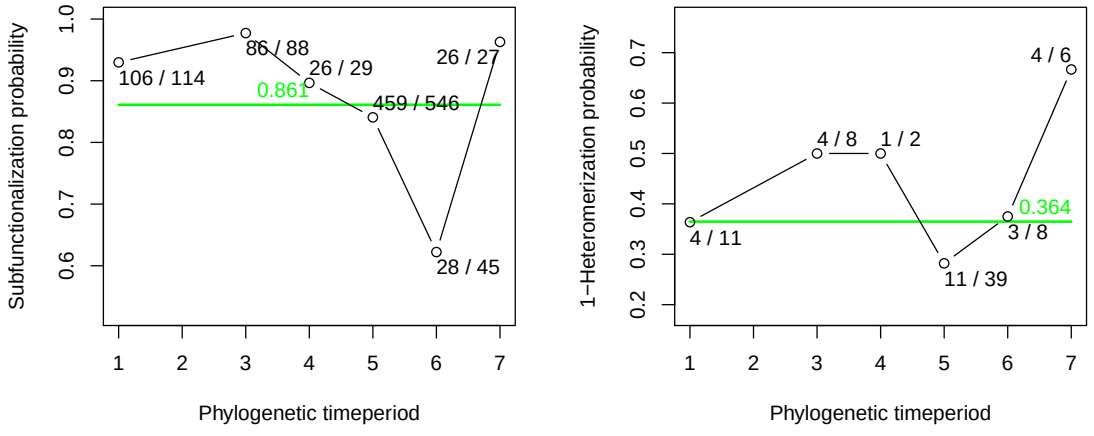
(A) **(B)**

Supplementary Figure 10: Ito et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 51 duplication events across seven phylogenetic time periods. Interpretation of plots is as per Supplementary Figure 1.



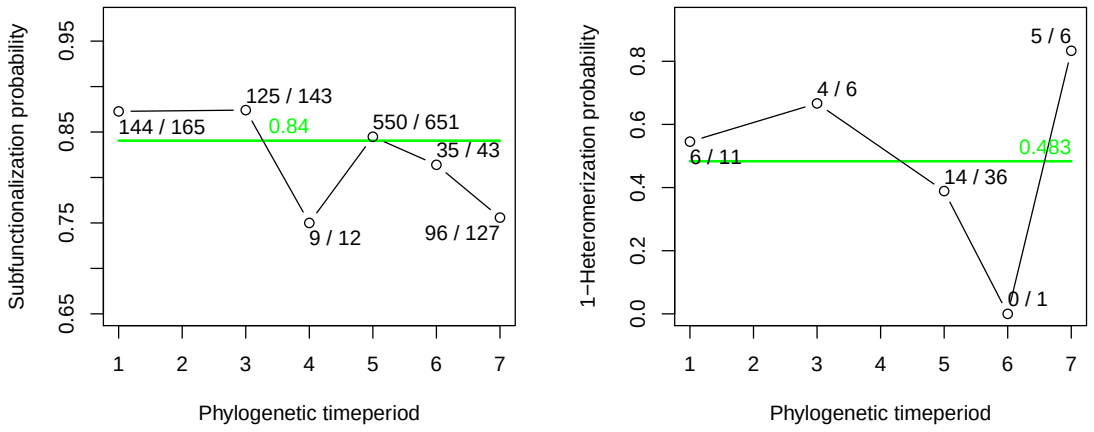
(A) **(B)**

Supplementary Figure 11: Combined Ito et al. and Uetz et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 85 duplication events across seven phylogenetic time periods. Note that the number of duplication events exceeds the sum of the individual data sets. This is due to single-member of paralogous families in individual data sets which form multi-protein families in the empirical data. Interpretation of plots is as per Supplementary Figure 1.



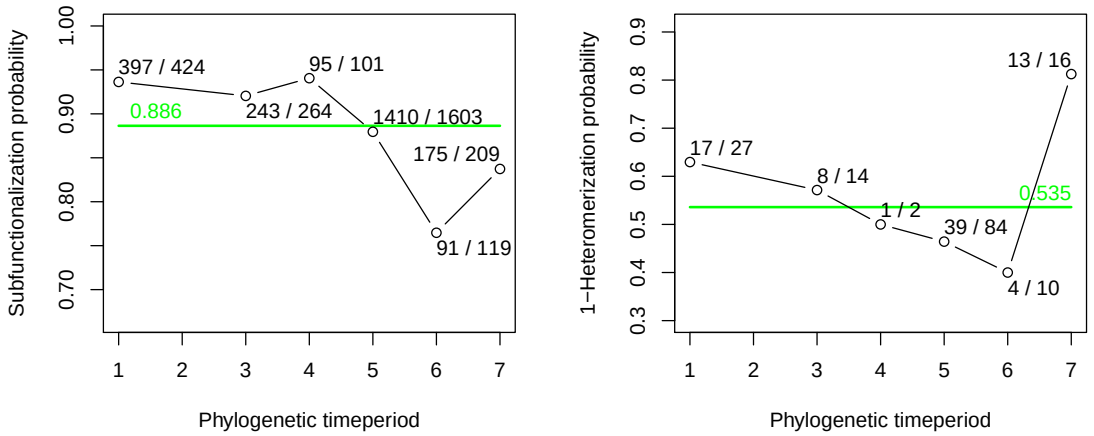
(A) **(B)**

Supplementary Figure 12: Yu et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 181 duplication events across seven phylogenetic time periods. Interpretation of plots is as per Supplementary Figure 1.



(A) **(B)**

Supplementary Figure 13: Tarassov et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 94 duplication events across seven phylogenetic time periods. Interpretation of plots is as per Supplementary Figure 1.



(A) **(B)**

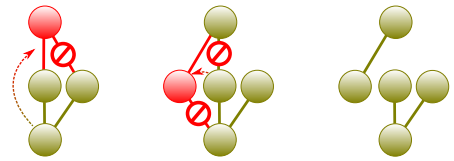
Supplementary Figure 14: Combined Ito et al., Uetz et al., Yu et al., Tarassov et al. subfunctionalization and heteromerization rates by phylogenetic time period. Shown are 298 duplication events across seven phylogenetic time periods. Interpretation of plots is as per Supplementary Figure 1.

5 Low clustered empirical data and Vázquez

The coverage of the data sets is an important component of the data used. The Ito et al. and Uetz et al. data sets are incorporated into the Yu et al. data set and feature low clustering coefficients. This lower clustering does not make it any more accessible to the original Vázquez et al. model. Using parameter values derived from these data sets in isolation (q.v., supplementary figures 7-12), the Vázquez et al. model produces a very high clustering coefficient compared to that of these empirical data sets. (Estimates of the clustering coefficient for the alternative parameter values can be obtained by cross-referencing the parameter values with the heatmap in Figure 2 in the main text). The low clustering of these data sets can be attributed in part to their relatively low coverage. The higher clustering coefficient obtained by combining complementary data sets is consistent with Friedel and Zimmer (2006) who concluded that a sampled (i.e., incomplete) biological data set would have a lower clustering coefficient than its corresponding complete data set.

6 Network components

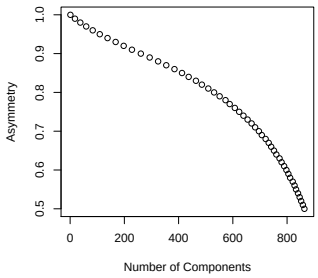
The different approaches to asymmetry in the Vázquez et al. and Solé et al. models affect the number of components in the networks they generate. Completely asymmetric subfunctionalization as in Solé et al. generates a network with only a single component. Any interactions retained by the progeny are connected to the single component. If the progeny retains no interactions, then the duplication is deemed to have failed and the neighborless node is removed from the network. By contrast, symmetric subfunctionalization allows multiple components to develop in the network (Supplementary Figure 15).



Supplementary Figure 15: Symmetric subfunctionalization generates multiple components. (A) A gene duplication and redundant interaction loss. (B) A second duplication and symmetric loss of redundant interactions. (C) The resulting network has two components.

To evaluate how different degrees of asymmetry affect the number of components, we added an asymmetry parameter to the Vázquez et al. model. The average number of components decreases as the asymmetry increases in the model (Supplementary Figure 16).

With empirically-derived parameter values, the original Vázquez et al. model featuring symmetric subfunctionalization produces an average of 864 components. The number and distribution of components for the empirical data can be found in Supplementary Table 1. Of the 165 components in the empirical network, nearly all interactions are contained in the large component. In fact, the network's largest component contains all of the empirical network's triangles. Supplementary Table 1 shows the distribution of nodes across all of the components in the empirical data set.



Supplementary Figure 16: Number of components as a function of asymmetry. The Vázquez et al. model was run with the empirically-derived parameter values computed in Methods, main text and in Supplementary Methods ($p = 0.24, q = 0.887, \text{order} = 2647$). The asymmetry value along the vertical axis is defined as the probability of an interaction selected for loss being removed the progeny. At 0.50, the losses are symmetric as in the original Vázquez et al. model. At 1.00, the losses are completely asymmetric as in the original Solé et al. model. Each data point represents the mean value from 1000 runs of the Vázquez et al. model. Asymmetry values ranged from 0.5 to 1.0 in 0.01 increments.

component order	1	2	3	4	6	2345	total
num. of components	60	77	22	3	2	1	165

Supplementary Table 1: Component distribution of Combined data set. The *component order* refers to the number of nodes in the component. A single large component contains nearly all of the nodes. Single node components are nodes with a self-interaction. A single edge connects two-node components, possibly with one or both nodes self-interacting. Three-node components are all paths of length two (all triangles are contained in the largest component).

7 Additional iSite model notes

7.1 iSite asymmetry

The iSite model's tendency towards single iSite proteins underlies high interaction asymmetry even when iSites are subfunctionalized with no asymmetry. Each time a single-iSite protein is duplicated, silencing redundant iSite interactions is by definition a 100% asymmetric operation. That is, some, none, or all interactions will be silenced on the progeny's iSite, or the progenitor's iSite, but not both.

7.2 Enhancing subfunctionalization in the iSite model

Both the iSite and Vázquez et al. models fix the subfunctionalization parameter with a single probability. However, protein interaction change rates are not monolithic. Interactions among constituents of an evolving protein complex change more slowly than shorter interacting protein peptide regions which aren't part of a globular domain (Neduva and Russell, 2005). Semantically, the iSite model can accommodate additional interaction types by generalizing the concept of an

iSite to be any portion of a protein which mediates an interaction. To lend meaning to such an accommodation however also requires expanding the model's concept of subfunctionalization. Multiple iSite types could be identified in the model, each with its own subfunctionalization rate. Alternately, different evolutionary rates could be accommodated by associating each iSite's subfunctionalization parameter with a distribution rather than a single value. In either case, each protein's iSites would be preserved as heritable traits, passed on to protein progeny.

7.3 alternative iSite implementation

An alternative implementation of the iSite model treats all interactions of an iSite as a single unit for the sake of computing subfunctionalization. The Vázquez et al. model silences one interaction of a redundant pair with some probability. Similarly, we silence all interactions of a redundant iSite with some probability. This alternative implementation of the iSite model produces far too many edges and an unrealistically-high clustering coefficient. Using empirically-derived parameter values, 100 runs of the model produce a mean clustering coefficient of 0.99 and over 43,000 interactions.

The reason the alternative iSite model generates too many interactions is that with each duplication, an iSite can gain interactions but can never lose interactions. For example, Figure 2 (main text) shows that each iSite belonging to the paralogs' neighbors (light green and dark olive) has gained an interaction. Subfunctionalization may return some of those neighbor iSites back to their original number of interactions, but they will never lose interactions. As the model evolves, iSites become associated with ever greater numbers of interactions, leading to dense networks. Though the number of interactions in the Vázquez et al. model also grows as it evolves, the independence of the interactions results in a much slower rate of interaction growth.

8 Additional topological measures

The clustering coefficient is the primary topological measure we have used to assess the performance of the iSite model. The clustering coefficient is an attractive assessment measure because it is directly affected by the evolutionary mechanics featured in the iSite model, the Vázquez model, and its variants (Gibson and Goldberg, 2009a) Here we present additional topological measures for networks derived from various empirical data sets and evolutionary models.

In Supplementary Table 2 we present the Average Shortest Path Length and the Degree Assortativity. These measures provide additional insight into the overall topology produced by the iSite model.

There is a disparity between empirical values and model values, suggesting that either existing model mechanics can be modified or additional mechanics can be incorporated to improve the fit of these topological features. The average minimum path length appears to be tied to subfunctionalization asymmetry; the 91% asymmetric Vázquez model and the (similarly asymmetric) iSite model feature comparably large minimum path lengths.

Data set	Path Length	Assortativity
Uetz	3.40	-0.032
Ito	3.57	-0.075
Uetz/Ito	4.43	-0.046
Yu	4.69	-0.05
Tarassov	4.61	0.192
Yu/Tarassov	4.88	0.041
Uetz/Ito/Yu/Tarassov	4.88	0.042
Vázquez	1.13	-0.048
Asym-82% Vázquez	2.58	-0.133
Asym-91% Vázquez	8.61	-0.140
Heritable Vázquez	1.03	-0.072
iSite	8.22	-0.047
iSite (SF seed)	9.00	-0.051

Supplementary Table 2: Average shortest path length and degree assortativity for networks studied. Both measures indicate topological discrepancies between the empirical network and the iSite networks. The *SF seed* is the scale-free seed used in network construction

9 Erdős-Renyí and Preferential Attachment seed graphs

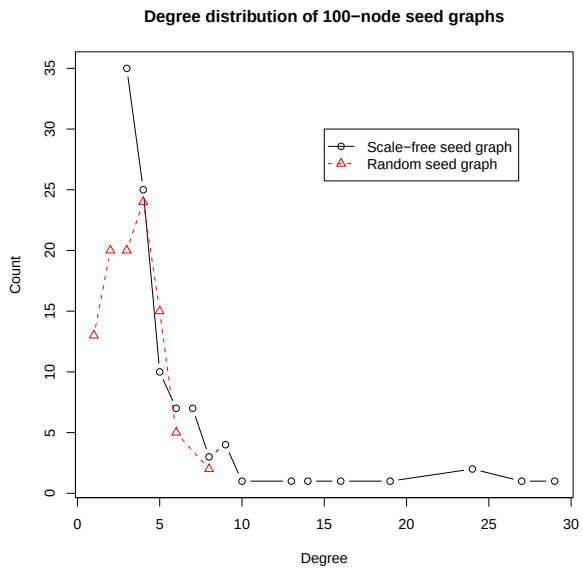
Seed graph	Order	Size	Avg. shortest path	Assortativity
Erdős-Renyí	100	165	3.97	-0.024
Preferential Attachment	100	291	2.58	-0.128

Supplementary Table 3: Topological statistics of seed graphs used.

All evolving graph models in the main text were built from a single 100-node Erdős-Renyí random graph as described in the Supplementary Methods, Section 2.7. In order to better exercise the iSite model, it was also tested using a scale-free seed graph. The scale-free seed graph was constructed using the preferential attachment model (Barabási and Albert, 1999). During seed graph construction, each new node was preferentially attached to 3 other nodes. After construction 83 self loops were randomly assigned to the nodes (Supplementary Methods, Section 2.7). Supplementary Figure 17 and Supplementary Table 3 feature the topological characteristics of the two seed graphs. Supplementary Table 4 shows shows the fidelity of the iSite model under different seed graphs.

Dataset	Order	Size	Triangles	Connected Triples	C	C of R.E.	Deg. Dist. Difference	Homomeric Proteins
Combined	2647	5449	3299	66799	0.148	1.7×10^{-2}	0	412
iSite (ER seed)	2647	5515 (101%)	1699 (52%)	33088 (50%)	0.154	4.8×10^{-3}	334	652 (158%)
iSite (SF seed)	2647	5547 (101%)	1708 (51%)	33367 (49%)	0.154	5.0×10^{-3}	331	642 (155%)

Supplementary Table 4: A comparison of topological measures between iSite model networks generated with different seed graphs. For easy reference, the empirical data set ("Combined") featured in Table 1 in the main text is reprinted.



Supplementary Figure 17: The degree distribution of the Erdős-Renyí and Preferential Attachment seed graphs.

10 Estimate of interaction change rate in Saccharomyces cerevisiae

Beltrao and Serrano (2007) calculated the interaction rate change of eukaryotic species to be on the order of 10^{-5} interactions per million years (My). For Saccharomyces cerevisiae specifically, they reported a rate of 2.4×10^{-5} changed interactions/My. The variables used in the calculation are the number of changed interactions over a time period, the number of possible protein pairings (i.e. potential interaction changes), and the evolutionary time period being measured.

Specifically, they measure the interaction change rate as:

$$\frac{x_{\text{changed}}}{x_{\text{potential}} \times t}$$

x_{changed} is the number of interactions lost (or gained) over a measured time period, t.

Additionally, Beltrao and Serrano derive the number of potential interactions as:

$$x_{\text{potential}} = P_{\text{new}} \times P_{\text{old}} + \frac{P_{\text{new}}(P_{\text{new}} - 1)}{2}$$

Where P_{new} is the number of new proteins (i.e., surviving duplicates) with interactions and P_{old} is the number interacting proteins present in the last common ancestor (LCA) prior to the gene duplications enumerated in P_{new}. From our data (see below), $P_{\text{new}} = 293$ and $P_{\text{old}} = 2242$.

The number of potential interactions is then:

$$293 \times 2242 + \frac{293(293-1)}{2} = 699684$$

P_{old} and P_{new} were calculated from counts of interacting proteins in the LCA and the extant Saccharomyces cerevisiae networks respectively. The extant number of interacting proteins in the network is simply the intersection between the proteins in the interaction data and the genes in the reconciled gene trees. Removing duplicated (i.e., paralogous) proteins (as identified in the gene tree data) produces the number of proteins in the LCA. A detailed treatment of deriving the LCA network can be found in (Gibson and Goldberg, 2009b).

According to our methodology, over the 300My period covering the duplication events (Gibson and Goldberg, 2009b), the number of changed interactions is $2389 + 82 + 117 = 2588$. The first two terms are the sum of the numerators from Figure 2A&B in the main text. The third number counts 117 lost self-interactions (which aren't shown in the figure). There are 699,684 potential interactions. The number of changed interactions is:

$$\frac{2588}{699684 \times 300My} = 1.2 \times 10^{-5} \text{ changed interactions/My}$$

Beltrao and Serrano used different interaction data and a completely different methodology to derive values assigned to the calculation. Despite this, our estimate of the interaction change rate in Saccharomyces cerevisiae matches Beltrao and Serrano's rate of 2.4×10^{-5} changed interactions/My quite well.

11 Saccharomyces cerevisiae interaction data

Two data sets were combined for use in this study. The Yu et al. (2008) is a "second-generation" data set comprised of high-confidence binary interactions identified in previous large-scale studies, and interactions identified independently more than once in the literature. Yu et al. further vetted the quality of the interactions by randomly selecting interactions and testing them with both the yeast two-hybrid assay and protein-fragment complementation assay (PCA). The Tarassov et al. (2008) data are high-quality interactions identified in vivo via PCA. They were included due to the complementary nature of the interactions. This high-throughput data avoids potential biases in relying solely on literature-curated data sets due to biologists favoring "interesting" genes (e.g. a preponderance of essential genes) (Reguly *et al.*, 2006).

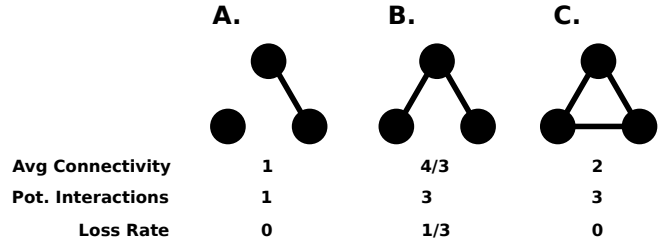
Dataset	Order	Size	Triangles	Connected Triples	C	C of R.E.	Deg. Dist. Difference	Homomeric Proteins
Uetz et al.	806	682	7	1139	0.018	0.0	-	38
Ito et al.	813	843	34	4085	0.025	5.1×10^{-3}	-	82
Yu et al.	2018	2930	212	26933	0.024	1.1×10^{-2}	-	225
Tarassov et al.	1084	2616	3017	34006	0.266	3.5×10^{-2}	-	221

Supplementary Table 5: Topological measures of the Yu et al. (2008), Tarassov et al. (2008), Uetz et al. (2000), and Ito et al. (2001) data sets

Data Set	Order	Size	Avg. Connect.	Paralogous Loss Rate	Subfunc. Rate
Uetz	806	682	1.69	0.153	0.805
Ito	813	843	2.07	0.280	0.845
Uetz/Ito	1301	1395	2.14	0.285	0.846
Yu	2018	2930	2.90	0.364	0.861
Tarassov	1084	2616	4.83	0.483	0.840
Yu/Tarassov	2647	5449	4.12	0.539	0.887
U/I/Y/T	2663	5476	4.11	0.535	0.886

Supplementary Table 6: Parameter estimation by data set. Network order, size, average connectivity, paralogous interaction loss rate, and subfunctionalization rate for various individual and combined data sets. All four data sets combined are shown as U/I/Y/T.

Supplementary Tables 5 and 6 present additional high-quality data sets and their effect on parameter estimation. The Ito et al. (2001) and Uetz et al. (2000) data sets are older, cover fewer interactions and proteins than the newer Yu et al. (2008) and Tarassov et al. (2008) data sets. In fact, the Ito and Uetz data sets are subsets of the Yu data set. As indicated in Supplementary Table 6, paralogous interaction loss rates are quite sensitive to different data sets. In fact with the exception of the highest average connectivity, the paralogous interaction loss rate increases as the average connectivity (i.e., the average number of interactions per protein) increases. This is counterintuitive to the notion that the loss rate would decrease monotonically as the average connectivity increases. Supplementary Figure 18 illustrates a plausible scenario consistent with the increasing connectivity and fluctuating paralogous interaction loss rates.



Supplementary Figure 18: Paralogous interaction loss rate versus average connectivity. Shown are the proteins for a paralogous 3-gene family. As the average connectivity of the network increases, the paralogous interaction loss rate increases. However once a high average connectivity is reached, the loss rate once again decreases. There are two phenomena responsible for this behavior. The first is that regardless of the size of a protein family, only those paralogous members which have degree greater than zero in the interaction network are included in the calculation. So in pane A, only two of the three paralogous proteins are used in the calculation of potential paralogous interactions. The second is that potential paralogous interactions are calculated as $n(n-1)/2$. The number of potential interactions increases with the square of the number of paralogous proteins in the network (for a given family).

The subfunctionalization rate is much more stable. The subfunctionalization rate only increases

by 0.082 as the average connectivity increases.

As a final observation, recall the evolutionary phenomena underlying the subfunctionalization and paralogous interaction loss rates are the same. Both are based on the principle that functions made redundant by initially identical duplicates are partitioned between the duplicates as their sequences mutate. The paralogous interaction formed of a duplicated homomeric interaction is redundant to the progenitor's homomeric interaction. It is interesting to note that as the empirical "sample" increases (Supplementary Table 6), the paralogous interaction loss rate approaches the subfunctionalization rate. Though speculative, one possibility to consider is that as empirical data sets grow, a single rate may ultimately be sufficient to capture both types of interaction loss.

12 Parameter values derived by Vázquez et al.(2003)

Vázquez, et al. (2003) selected parameter values by calculating the average degree of their empirical network, and then "tuning" the model parameters such that the model generated networks with the same average degree. Using this method they arrived at a subfunctionalization probability of $q = 0.7$, and a heteromerization probability of $p = 0.1$ (probability of an interaction between progenitor and progeny). These parameter values produced networks having "reasonable agreement" with the observed data. They further observed that the proportion of self-interacting proteins in their empirical network (0.04) was not far from their tuned value of $p = 0.1$. This latter observation is somewhat dubious since p represents the probability of an interaction between duplicates *surviving*, not the number of self-interactions in the network.

References

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Beltrao, P. and Serrano, L. (2007). Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol*, **3**(2), e25.

Friedel, C. C. and Zimmer, R. (2006). Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, **7**, 519.

Gibson, T. A. and Goldberg, D. S. (2009a). Questioning the ubiquity of neofunctionalization. *PLoS Comput Biol*, **5**(1), e1000252.

Gibson, T. A. and Goldberg, D. S. (2009b). Reverse engineering the evolution of protein interaction networks. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Pac Symp Biocomput*, pages 190–202.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**(8), 4569–4574.

Milo, R., Kashtan, N., Itzkovitz, S., Newman, M., and Alon, U. (2004). On the uniform generation of random graphs with prescribed degree sequences. http://aps.arxiv.org/abs/cond-mat/0312028/.

Neduva, V. and Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Lett*, **579**(15), 3342–3345.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., and Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, **5**(4), 11.

Solé, R. V., Pastor-Satorras, R., Smith, E., and Kepler, T. B. (2002). A model of large-scale proteome evolution. *Advances in Complex Systems*, **5**, 43.

Tarassov, K., Messier, V., Landry, C. R., Radinovic, S., Molina, M. M. S., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. *Science*, **320**, 1465–1470.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Vázquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Modeling of protein interaction networks. *ComPlexUs*, **1**, 38–44.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449**(7158), 54–61.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.