BMC Bioinformatics

RESEARCH

Open Access

Integration of probabilistic functional networks without an external Gold Standard

Katherine James[1,2*], Aoesha Alsobhe[2,3], Simon J Cockell[4], Anil Wipat[2] and Matthew Pocock[2]

*Correspondence:
katherine.james@newcastle.ac.uk

[1] Department of Applied Sciences, Northumbria University, Sandyford Rd, Newcastle upon Tyne NE1 8ST, UK
[2] Interdisciplinary Computing and Complex BioSystems Group, Newcastle University, Science Square, Newcastle upon Tyne NE4 5TG, UK
[3] Saudi Electronic University, Abi Bakr As Siddiq Branch Rd, Riyadh 1332, Saudi Arabia
[4] School of Biomedical, Nutritional and Sports Science, Faculty of Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK

Abstract

Background: Probabilistic functional integrated networks (PFINs) are designed to aid our understanding of cellular biology and can be used to generate testable hypotheses about protein function. PFINs are generally created by scoring the quality of interaction datasets against a Gold Standard dataset, usually chosen from a separate high-quality data source, prior to their integration. Use of an external Gold Standard has several drawbacks, including data redundancy, data loss and the need for identifier mapping, which can complicate the network build and impact on PFIN performance. Additionally, there typically are no Gold Standard data for non-model organisms.

Results: We describe the development of an integration technique, ssNet, that scores and integrates both high-throughput and low-throughout data from a single source database in a consistent manner without the need for an external Gold Standard dataset. Using data from *Saccharomyces cerevisiae* we show that ssNet is easier and faster, overcoming the challenges of data redundancy, Gold Standard bias and ID mapping. In addition ssNet results in less loss of data and produces a more complete network.

Conclusions: The ssNet method allows PFINs to be built successfully from a single database, while producing comparable network performance to networks scored using an external Gold Standard source and with reduced data loss.

Keywords: Network integration, Bioinformatics, Gold Standards, Probabilistic functional integrated networks, Protein function prediction, Interactome

Background

Probabilistic functional integrated networks (PFINs) are an automated method that combines multiple sources of functional interaction data to produce a single network of interactions each annotated with confidence scores. PFINs allow false positive interactions to be identified by their low confidence scores, avoiding the need for work-intensive manual pre-integration data cleaning. These scores are often calculated by comparison with a high quality "Gold Standard" dataset [1–5]. The performance of any PFIN is dependent upon its component datasets and on the Gold Standard(s) chosen to evaluate them [6]. Integrated network theory relies on the principle that the whole network is greater than the sum of its dataset parts, so as more interaction data are produced, network performance should increase [7]; a PFIN integrated using data from

2010 is expected to perform better than one using data from 2006. However, our previous 2012 study, which investigated PFINs integrated using data from 2006 to 2010, demonstrated that the network performance fluctuated and that careful selection of data and Gold Standard is vital [8], particularly if the network is integrated to study a specific area of biology.

Since 2012 the volume of functional interaction data has increased considerably [9]. While high-throughput (HTP) data produced from largescale screenings have been estimated to have high false positive and false negative rates (up to 91% and 96%, respectively) [1, 10–15], the effect of this noise should be mitigated by an increase in less error-prone, targeted low-throughput (LTP) studies. Curated databases are constantly evolving in both content and structure to reflect current biological knowledge [16]. Manual curation has improved the quality of available data by editing or removal when errors and inconsistencies have been identified, and when data are found to be incorrect in light of new studies [17, 18].

Gold Standard data are chosen from a separate high-quality database, commonly manually-curated metabolic pathways [1, 19, 20], for example KEGG [21], or shared biological process [10, 22, 23], for example Gene Ontology [24]. The difference in database source between datasets and Gold Standard presents several challenges to scoring and integration. The Gold Standard may have poor overlap with the datasets to be scored since its focus differs, for instance a metabolic pathway database vs a protein-protein interaction (PPI) database; many PPI datasets may not score at all. Conversely, data redundancy can occur since separate databases may include the same studies, leading to overlap between Gold Standard and datasets, and biased scoring. Identifier format may also differ, requiring potentially error prone mapping between the two sources prior to scoring [25]. In Gold Standards derived from hierarchical annotation schemas, such as Gene Ontology, the choice of which annotation terms to include in the Gold Standard can vastly change the final scores and these schemas are known to have large biases [26].

With the increase in volume of high quality LTP data, for both model and non-model organisms, we hypothesise that these data can be used as a Gold Standard to score the larger HTP studies. Therefore, using functional interaction data from a single curated database, datasets can be scored and integrated without the requirement for an external Gold Standard dataset. Less data will be lost and the challenges of identifier mapping and data redundancy will be removed. Furthermore, since the Gold Standard is considered the highest-quality data, we propose that these data may be integrated into the final network using iterative scoring.

BioGRID[1] [27] is one of the most comprehensive manually-curated interaction databases, with excellent coverage of model organisms, including *Saccharomyces cerevisiae*. It benefits from a clear data format that facilitates integration of other interaction datasets into analysis pipelines. Here, we extend our 2012 study by analysing the changes in *Saccharomyces cerevisiae* functional interaction data contained in BioGRID and evaluating the performance of PFINs scored using an LTP Gold Standard. We find that the quantity of interaction data has increased to the extent whereby both scoring and

[1] https://thebiogrid.org.

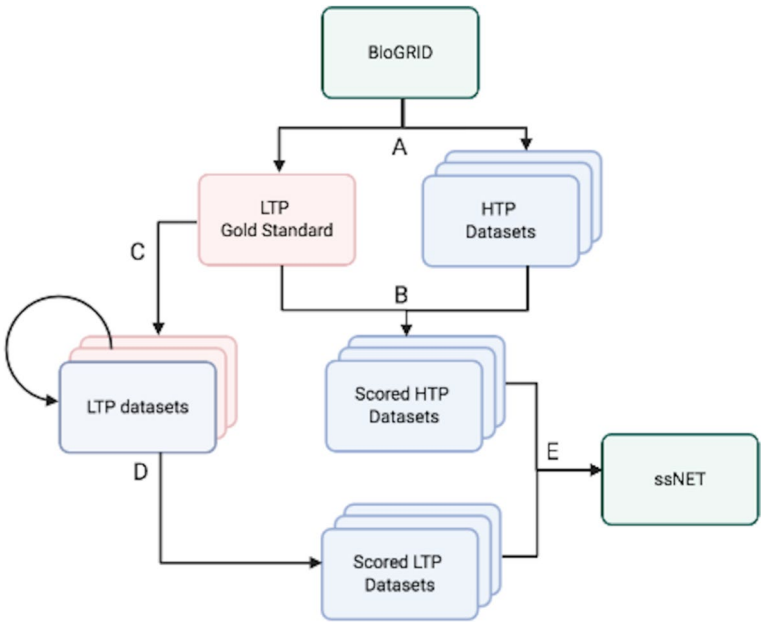

Fig. 1 The ssNet scoring and integration method. **A** BioGRID data are split into high throughput (HTP) and low throughput (LTP) data based on study size threshold. **B** The HTP datasets are scored using the LTP data as Gold Standard. **C** LTP data is then split into datasets by experimental type. **D** Each LTP data type is scored using the remaining LTP data types as Gold Standard. **E** The scored HTP and LTP datasets are then integrated into the final network

integration can be achieved from the data contained in BioGRID alone without the need for an external GoldStandard. We propose a new integration method, ssNET, that incorporates both LTP and HTP data in a consistent manner, and evaluate it using functional prediction.

The network produced in this study is available at https://figshare.com/projects/Yeast_ssNet/114366 and the code for ssNet network generation at https://github.com/kj-intbio/ssnet/.

Methods

Data sources

BioGRID archive datasets [27] were downloaded for a fifteen year period at six month intervals[2] from the first available version (V17, 2006) to the latest version (V186, 2020). Entrez ID lists for each species were downloaded from the NCBI database[3] [28] (*S. cerevisiae* 5th June 2020), BioSystems Gold Standard data was downloaded from the NCBI's FTP server[4] (version 20170421). Gene Ontology Gold Standard data and annotation information for evaluation were downloaded from the Gene Ontology Resource[5] [24, 29]: the Gene Ontology obo format (23rd March 2020) and annotation mapping files (*S. cerevisiae* 23rd March 2020) [30].

[2] https://downloads.thebiogrid.org/BioGRID/Release-Archive/.

[3] https://www.ncbi.nlm.nih.gov/.

[4] https://ftp.ncbi.nih.gov/pub/biosystems/.

[5] http://www.geneontology.org/.

Dataset scoring and integration

The ssNet method involves a five step scoring and integration (Fig. 1). BioGRID data were first split by PubMed ID into HTP and LTP data based on study size threshold (100 throughout unless otherwise stated) and HTP datasets were treated as separate studies. For consistency dataset names are standardised throughtout the results as [Author]. [PubMed ID]. These datasets were then scored using the LTP data as Gold Standard. LTP data were then split into datasets by data type based on the BioGRID experimental code[6]. Each LTP dataset was scored using the remaining LTP data types as Gold Standard. Finally, the HTP and LTP datasets were integrated.

For comparison both LTP and HTP datasets were scored against three Gold Standards:

- BioSystems molecular pathway-dervived Gold Standard in which proteins in the same pathway were considered Gold Standard pairs (MP_GS).
- A Gene Ontology biological process-derived Gold Standard in which proteins belonging to processes covering <10% of the genome were considered Gold Standard pairs (BP10_GS).
- A Gene Ontology biological process-derived Gold Standard in which proteins belonging to processes with <100 annotations in the yeast genome were considered Gold Standard pairs (BP100_GS).

Confidence scores were calculated using the methods developed by Lee and colleagues [1], which calculates a log-likelihood score for each dataset:

$$lls^{L}(E) = \ln\left(\frac{P(L|E)/\neg P(L|E)}{P(L)/\neg P(L)}\right) \tag{1}$$

where, $P(L|E)$ and $\neg P(L|E)$ represent the frequencies of linkages L observed in a dataset E between genes that are linked and not not linked in the Gold Standard, respectively, and, $P(L)$ and $\neg P(L)$ represent the prior expectation of linkages between genes that are linked and not not linked in the Gold Standard, respectively. A score greater than zero indicates that the dataset links pairs of genes present in the Gold Standard with higher scores indicating greater confidence in the data. Datasets that did not have a positive score or that did not score ($P(L|E) = 0$) were discarded. Datasets scoring infinity due to perfect overlap with the Gold Standard ($\neg P(L|E) = 0$) were given a score of $\lceil h \rceil + 1$, where h is the set of dataset scores.

Scores were then integrated using the Lee method [1]:

$$WS = \sum_{i=1}^{n} \frac{L_i}{D^{(i-1)}} \tag{2}$$

where L_1 is the highest confidence score and L_n the lowest confidence score of a set of n datasets. The D parameter provides a non-parametric weighting of datasets by their confidence score rankings. At $D = 1$, dataset are given equal weight regardless of their confidence scores. As D increases above 1, datasets with higher relative confidence are

[6] https://wiki.thebiogrid.org/doku.php/experimental_systems.

up-weighted. A *D*-value of 1 was chosen for our initial network builds, while higher *D*-values were chosen where stated to investigate the effect of *D*-value choice on network performance. In the resulting network, interactions with multiple lines of high quality evidence have higher scores, while those with little/low quality evidence have lower scores. In HTP-only networks only the scored HTP datasets were integrated.

Evaluation

Networks were clustered using the Markov Clustering Algorithm (MCL) version 14-137 with default inflation value [31]. Networks were visualised in Cytoscape Version 3.7.2 [32] and the Network Analyser plugin version 3.3.2 was used to calculate topological statistics [33]. Functional prediction was carried out using the Maximum Weight decision rule [34] in which annotations were propagated along the highest weighted edge surrounding a protein. Leave-one-out cross-validation of known annotations was carried out for all GO biological process with at least 100 and no more than 1000 annotations in the genome. Automated annotations with the code Inferred from Electronic Annotation (IEA) were excluded from evaluation. The performance of the networks was evaluated using the area under curve (AUC) of Receiver Operator Characteristic curves for each term [35] using the BioConductor ROC package (v 1.62.0) [36].

The error of the AUC was calculated using the standard error of the Wilcoxon statistic SE(W) [35, 37]:

$$SE(W) = \sqrt{\theta(1-\theta) + (C_p - 1)(Q_1 - \theta^2) + (C_n - 1)(Q_2 - \theta^2)/C_p C_n} \tag{3}$$

where, θ is the AUC, C_p is the number of positive examples, C_n is the number of negative examples and Q_1 and Q_2 are the probabilities of incorrect annotation assignment as defined by:

$$Q_1 = \frac{\theta}{2-\theta} \tag{4}$$

$$Q_2 = \frac{2\theta^2}{1+\theta} \tag{5}$$

Results

LTP data can be used as a Gold Standard to score HTP data quality

PFINs are traditionally integrated after datasets have been scored against an external Gold Standard dataset [1, 19, 20, 22]. Since the use of external data has several drawbacks, we investigated whether the less error-prone low-throughput (LTP) studies are suitable to be used as an internal Gold Standard. We first assessed the changes in BioGRID data for the yeast *Saccharomyces cerevisiae* from the first available version (V17, 2006) to the latest available version (V186, 2020). HTP interaction data (defined here as studies $\geq$ 100 interactions; see "Methods") have increased considerably since our earlier study [8] both in terms of unique proteins and unique interactions between them (Fig. 2A, B). The overlap between HTP and LTP interactions remains relatively low, increasing from 3382 (7.1% of total interactions) at V17 to 16724 interactions at V186

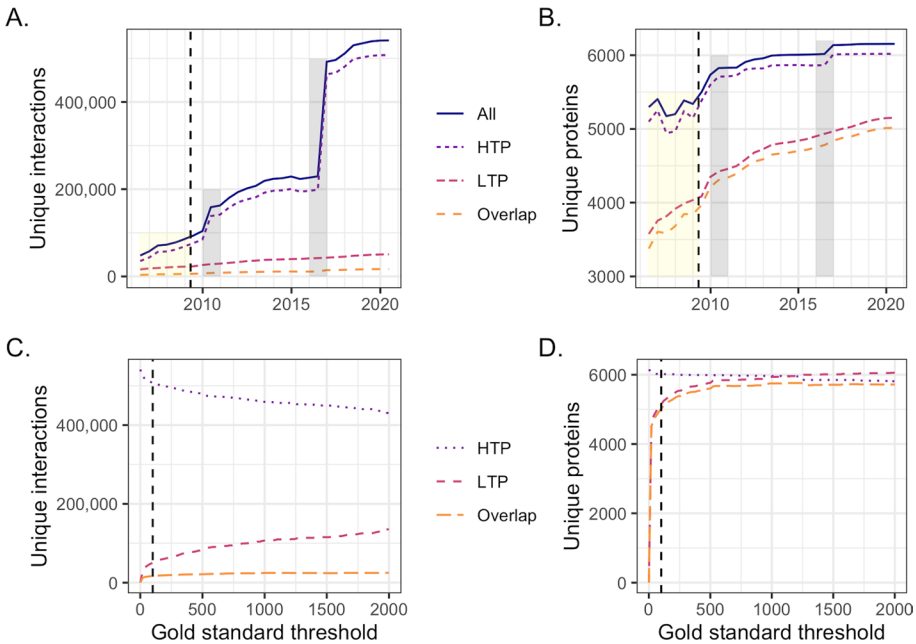

Fig. 2 The BioGRID *Saccharomyces cerevisiae* data increase. **A** High-throughput (HTP) interaction data has increased since our earlier study [8] (yellow box; dashed line), with two large increases in 2010 and 2016 (grey boxes). By contrast, low-throughput (LTP) data has increased to a lesser extent, with the overlap between HTP and LTP also increasing slightly. **B** The fluctuation in representation of unique proteins that was observed in our earlier study [8] (yellow box; dashed line) is not evident in the following years. Unique protein number has increased steadily for both HTP and LTP data and the large increases in interaction data (grey boxes) only affect this increase slightly. Version 186 LTP data contains ~1/3 of the total unique proteins in BioGRID, with LTP data almost completely overlapping with the HTP dataset. **C** The size and overlap of interactions in V186 HTP datasets and LTP Gold Standard data as the Gold Standard size threshold is increased. **D** The size and overlap of proteins in V186 HTP datasets and LTP Gold Standard data as the Gold Standard size threshold is increased. An LTP dataset size threshold of 100 interactions (dashed line) was chosen for subsequent analyses to maximise both the number of the LTP Gold Standard's unique proteins and unique interactions, while minimising its overall size

(3.1% of total interactions). While the number of unique proteins fluctuated in earlier versions of BioGRID, their numbers have steadily increased in later years. Protein overlap between HTP and LTP data has also increased from 3377 at V17 to 5016 at V186: 63.8% and 81.5% of total proteins, respectively. Increasing the size threshold for the LTP datasets from 100 interactions had little effect on the percentage overlap between HTP and LTP datasets (Fig. 2C, D).

We constructed an LTP-derived Gold Standard (LTP_GS) and three Gold Standards derived from the metabolic pathways of the BioSystems database [38] (MP_GS) and biological processes of the Gene Ontology [24] (BP10_GS and BP100_GS) for comparison (see "Methods"). An LTP dataset size threshold of < 100 interactions was chosen to maximise the number of unique proteins and unique interactions in the LTP_GS, while minimising the its overall size (Fig. 2C, D). The LTP_GS had greater overlap with HTP data in terms of individual proteins, with ~83% of the total proteins shared (Table 1), in comparison to ~42% for MP_GS, ~81% for BP10_GS and ~73% for BP100_GS. These results suggest that while overlap between LTP and HTP data is low, LTP data has better overlap than an external Gold Standard in terms of nodes, and therefore may provide an improved scoring. The LTP_GS had higher overlap with the HTP datasets than

Table 1 Gold Standard overlap. The overlap in terms of proteins and interactions of the high-throughput datasets (HTP) with the low-throughout Gold Standard (LTP_GS), the BioSystems-derived Gold Standard (MP_GS) and Gene Ontology biological process-derived Gold Standards (BP10_GS and BP100_GS) (highest overlaps are shown in bold).

		LTP_GS	MP_GS	BP10_GS	BP100_GS
Proteins	Gold Standard	5510	2551	5071	4551
(6,018 HTP)	Overlap	5016	2549	4868	4369
	Overlap %	**83.3**	42.4	80.9	72.6
Interactions	Gold Standard	50,491	9924	2,947,156	384,450
(507,352 HTP)	Overlap	16,724	1045	174,934	39,200
	Overlap %	3.3	0.2	**34.5**	7.73

Total numbers differ slightly from Fig. 2 since only those identifiers that could be mapped between BioSystems and BioGrid are included

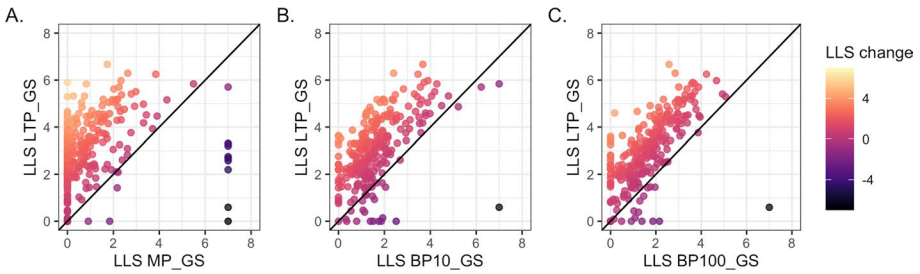

Fig. 3 Dataset scoring. The log likelihood scores (LLS) comparison for V186 BioGRID datasets scored against the low throughput (LTP_GS). Datasets that had perfect overlap with Gold Standard ($\neg P(L|E) = 0$) so are shown as score 7.0. Datasets shown as 0 were lost due scoring <0 or no score ($P(L|E) = 0$). **A** BioSystems-derived (MP_GS) Gold Standard (5 datasets score 0 for both Gold Standards). **B** Gene ontology biological process terms annotating <10% of the genome (BP10_GS) Gold Standard. **C** Gene ontology biological process terms annotating annotation <100 genes (BP100_GS) Gold Standard (2 datasets score 0 for both Gold Standards)

MP_GS in terms of interactions with 3.3% in comparison to 0.2%, but lower overlap than BP10_GS (34.5%) and BP100_GS (7.73%). While the largest Gold Standard, BP10_GS has the highest overlap with the HTP datasets, it's size—2,947,156 vs 9924 and 384,450 for MP_GS and BP100_GS, respectively—is likely to adversely affect scoring since the prior expectation ($P(L)$; see "Methods") will contain a large number of linkages between proteins with shared biological process that are less likely to directly interact.

When these Gold Standards were used in dataset scoring the majority of scores were higher using the LTP_GS standard in all three cases. The scores also had greater spread, indicating that the LTP_GS provides wider scope to score more diverse datasets (Fig. 3). Datasets may score infinity if they have perfect overlap with the Gold Standard ($\neg P(L|E) = 0$, see "Methods"). No datasets scored infinity using the LTP_GS, in comparison to 10 datasets using MP_GS, 2 using BP10_GS and 1 using BP100_GS (shown in Fig. 3 as a score of 7.0 on the right of the plot).

Datasets may also be lost during scoring if they do not have any overlap with the Gold Standard data ($P(L|E) = 0$, see "Methods"), or have negative scores. Fewer datasets were lost when scored against the LTP_GS (8 with LTP_GS, 96 with MP_GS, 18 with BP100_GS and 23 with BP10_GS; shown as score 0 in Fig. 3). Since the most recent BioSystems dataset available was April 2017, we also confirmed the difference between LTP_GS and

Table 2 Network topology. Network statistics for the low throughput (LTP_GS), BioSystems (MP_GS) and Gene Ontology (BP10_GS and BP100_GS) scored networks. [31]

	LTP_GS	MP_GS	BP10_GS	BP100_GS
Proteins	5960	5919	5960	5958
Interactions	501,299	458,778	498,385	495,007
Average degree	168	155	167	166
Connected components	1	1	1	1
Diameter	4	5	4	4
Characteristic path length	2.07	2.15	2.07	2.07
Centralization	0.57	0.33	0.58	0.57
Correlation Power law	0.41	0.46	0.41	0.43
R^2 Power law	0.74	0.75	0.74	0.74
Clusters	87	220	31	169

Statistics were calculated using the Cytoscape NetworkAnalyser [33] plugin and clustering was carried out using the Markov Clustering Algorithm (MCL)

MP_GS using equivalent BioGRID data from 2017 (V150, data not shown). The use of LTP data as for scoring therefore results in less loss of data than using the external Gold Standards.

Four networks were integrated from the V186 HTP datasets scored against the LTP_GS and the MP_GS, BP10_GS and BP100_GS Gold Standards for comparison. The networks were all formed of one connected component, with the LTP_GS network being larger than MP_GS and BP100_GS and having the largest average degree (168 compared to 155 for MP_GS, 167 for BP10_GS and 166 for BP100_GS) (Table 2). The networks all had moderate correlation with the power law [39], possibly a reflection of their content being solely derived from noisy HTP data. The high connectivity of the LTP_GS network was reflected in the clustering output with the network being split into 87 clusters, compared to 220 clusters for the MP_GS-scored network and 169 for BP100_GS. BP10_GS gave the fewest clusters with just 31.

We compared the performance of the networks in prediction of 398 Gene Ontology (GO) biological process terms using leave one out prediction of known annotations measured using area under curve (AUC) of receiver operator characteristic (ROC) plots. The LTP_GS network had improved performance for the majority of GO terms in comparison to MP_GS (Fig. 4A) with 327 terms improved compared to 71. Of the AUC changes 227 were statistically significant with 209 (92%) improved using LTP_GS. Performance was also improved over BP10_GS with 230 terms improved compared to 168 (Fig 4B). Of the changes 114 were significant with 64 (56%) improved using LTP_GS. LTP_GS and BP100_GS had the most similar performance with 105 terms improved for LTP_GS compared to 293 (Fig. 4C). However, only 85 changes were statistically significant and with just 3 (4%) were improved using LTP_GS.

Incorporation of Gold Standard data improves network performance

Since the LTP_GS data are considered the highest-quality, their inclusion in the final network is desirable. We therefore investigated how LTP data could be scored and incorporated into the final integrated network. Since LTP data can be further divided by experimental type [19], we employed an iterative LTP scoring method, ssNet, in which

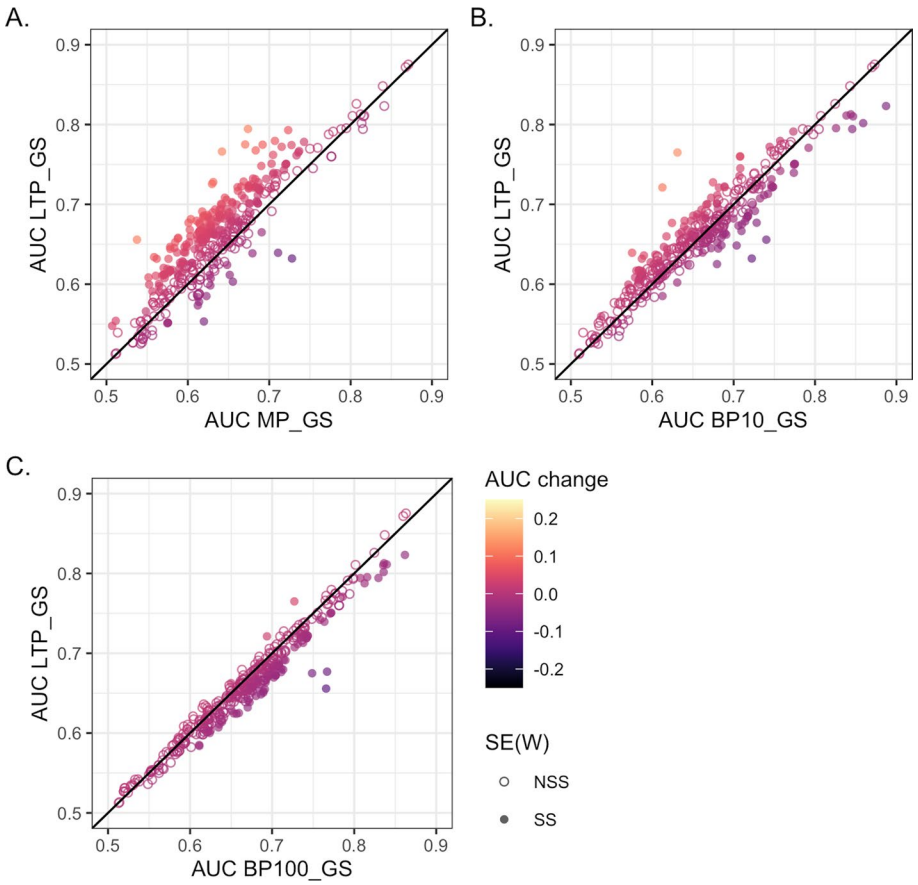

Fig. 4 Network evaluation. The functional prediction performance of the low throughput (LTP_GS), BioSystems and Gene Ontology-scored networks as measured by area under curve (AUC) of receiver operator characteristic plots. The error of the AUC was calculated using the standard error of the Wilcoxon statistic SE(W): not statistically significant (NSS); statistically significant (SS). A. BioSystems-derived (MP_GS) Gold Standard. Of 398 Gene Ontology biological processes, 327 had improved prediction using the LTP_GS network, and 71 using the MP_GS network. Of these changes 227 were statistically significant with 209 (92%) improved using LTP_GS. B. Gene ontology biological process terms annotating <10% of the genome (BP10_GS) Gold Standard. Of 398 Gene Ontology biological processes, 230 had improved prediction using the LTP_GS network, compared to 168 using BP10_GS. Of the changes 114 were significant with 64 (56%) improved using LTP_GS. C. Gene ontology biological process terms annotating annotation <100 genes (BP100_GS) Gold Standard. Of 398 Gene Ontology biological processes, 105 had improved prediction using the LTP_GS network, compared to 293 using BP100_GS. Of the changes just 85 were significant with 3 (4%) improved using LTP_GS

each LTP data type is scored as a single dataset using the remaining LTP types as Gold Standard, before integration of the LTP and HTP dataset scores (see "Methods"). Scores for the 28 LTP datasets were in a higher but overlapping range to those of the HTP datasets (Fig. 5A), while LTP scores using the MP_GS, BP10_GS and BP100_GS Gold Standards did not reflect the higher quality of these datasets (Fig. 5B–D). LTP ssNet scores where also higher in comparison to scores using a the other three Gold Standards (Fig. 6; five LTP datasets that did not score using MP_GS are shown as score 0). Overall, ssNET resulted in less loss of data than the use of external Gold Standards, particularly in terms of dataset loss (Table 3).

We integrated four networks of V186 LTP and HTP data scored using the ssNet, MP_GS , BP10_GS and BP100_GS Gold Standards, for comparison. The ssNet, BP10_GS and

Table 3 Data loss. Percentage loss of BioGRID proteins, interactions and datasets in the ssNet, BioSystems (MP_GS) and Gene Ontology (BP10_GS and BP100_GS) scored networks.

	Proteins	Interactions	Datasets
ssNet	**16.5**	**28.9**	**2.8**
MP_GS	17.6	34.6	34.9
BP10_GS	**16.5**	29.3	6.2
BP100_GS	**16.5**	29.7	8.0

The lowest data loss are shown in bold

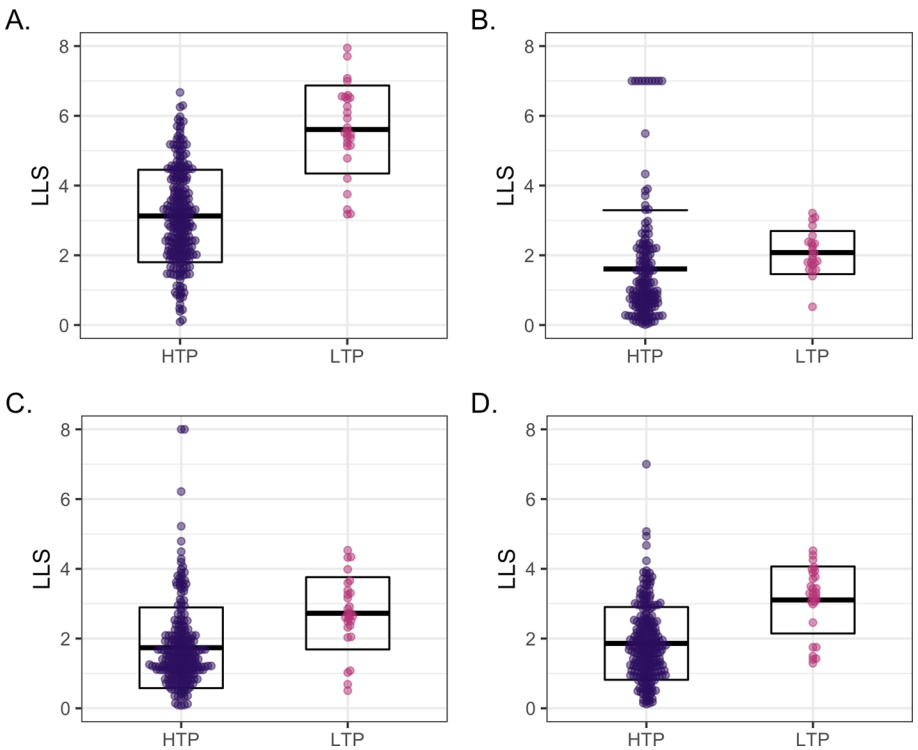

Fig. 5 LTP and HTP scores. **A**. Comparison of log-likelihood (LLS) score range for low-throughput (LTP) and high-throughput (HTP) datasets scored using ssNet. **B**. Comparison of log-likelihood (LLS) score range for low-throughput (LTP) and high-throughput (HTP) datasets scored using MP_GS. **C**. Comparison of log-likelihood (LLS) score range for low-throughput (LTP) and high-throughput (HTP) datasets scored using BP10_GS. **D**. Comparison of log-likelihood (LLS) score range for low-throughput (LTP) and high-throughput (HTP) datasets scored using BP100_GS

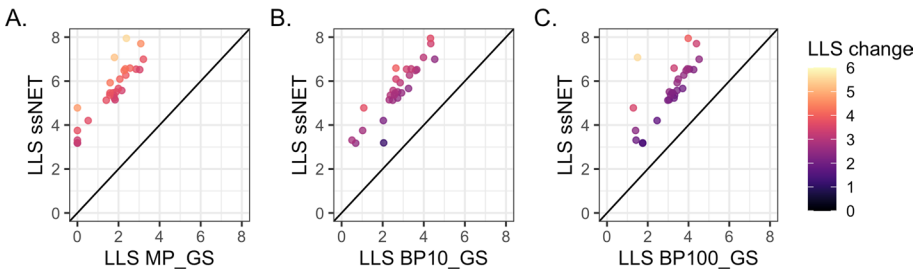

Fig. 6 ssNet scoring. **A**. Comparison of LTP dataset scores using ssNet and the BioSystems-derived Gold Standard (MP_GS). Datasets which were lost due scoring <0 or no score ($P(L|E) = 0$) are shown as 0. **B**. Comparison of LTP dataset scores using ssNet and the Gene Ontology Gold Standard BP10_GS. **C**. Comparison of LTP dataset scores using ssNet and the Gene Ontology Gold Standard BP100_GS

Table 4 Network statistics. Topological characteristics for the ssNet, BioSystems (MP_GS) and Gene Ontology (BP10_GS and BP100_GS) scored networks (including the LTP datasets).

	ssNet	MP_GS	BP10_GS	BP100_GS
Proteins	6119	6039	6119	6118
Interactions	535,183	492,424	532,349	529,036
Average degree	175	163	174	173
Connected components	1	2	1	1
Diameter	6	5	6	6
Characteristic path length	2.11	2.16	2.11	2.11
Centralization	0.56	0.33	0.56	0.56
Correlation Power law	0.61	0.58	0.60	0.61
R^2 Power law	0.74	0.76	0.74	0.74
Clusters	308	394	263	359

Statistics are calculated for the larger (6036 proteins) component of the BioSystems network

BP100_GS networks were formed of one connected component, while MP_GS had a second small component consisting of three proteins. The ssNet network was larger and more tightly connected (Table 4) with >535,000 interactions and all networks had higher correlation with the power law than those based on HTP data alone (Table 2).

We compared the performance of the networks in prediction of 398 Gene Ontology (GO) biological process terms as before. The results were similar to the HTP-only networks with ssNet having improved performance for the majority of GO terms in comparison to MP_GS (Fig. 7A): 325 terms improved compared to 73. Of the AUC changes 127 were statistically significant with 124 (98%) improved using LTP_GS. Performance was not improved over BP10_GS with 128 terms improved compared to 270 and of these changes just 20 were significant with 2 (10%) improved using LTP_GS (Fig. 7B). LTP_GS and BP100_GS had the most similar performance with 102 terms improved for LTP_GS compared to 296. However, only 2 changes were statistically significant both of which were improved using BP100_GS (Fig. 7C).

Finally, we investigated the ssNet network's performance when integrated using different *D*-values [see "Methods", Eq. (2)]. In general network performance is better at lower *D*-values, corresponding to lower up-weighting of higher-scoring datasets (Fig. 8A). However, individual Gene Ontology biological processes have variation in performance with differing optimal *D*-values (Fig. 8B).

Discussion

Since PFINS are generally created in order to support or derive new biological hypotheses, it is important to generate networks which are as complete and unbiased as possible [22]. Previous PFIN integration techniques have had several drawbacks due to their reliance on an external Gold Standard for dataset scoring, which may not be available for some species. Experimental datasets will also score differently depending upon the Gold Standard chosen [26, 40]. One popular Gold Standard is the Gene Ontology (GO) [10, 22, 23, 41–43]. The use of GO as a Gold Standard presents some problems due to the hierarchical directed acyclic graph (DAG) nature of the database; terms are connected to one another in a parent to child hierarchy, with annotation to a child term automatically implying annotation to all parent terms of

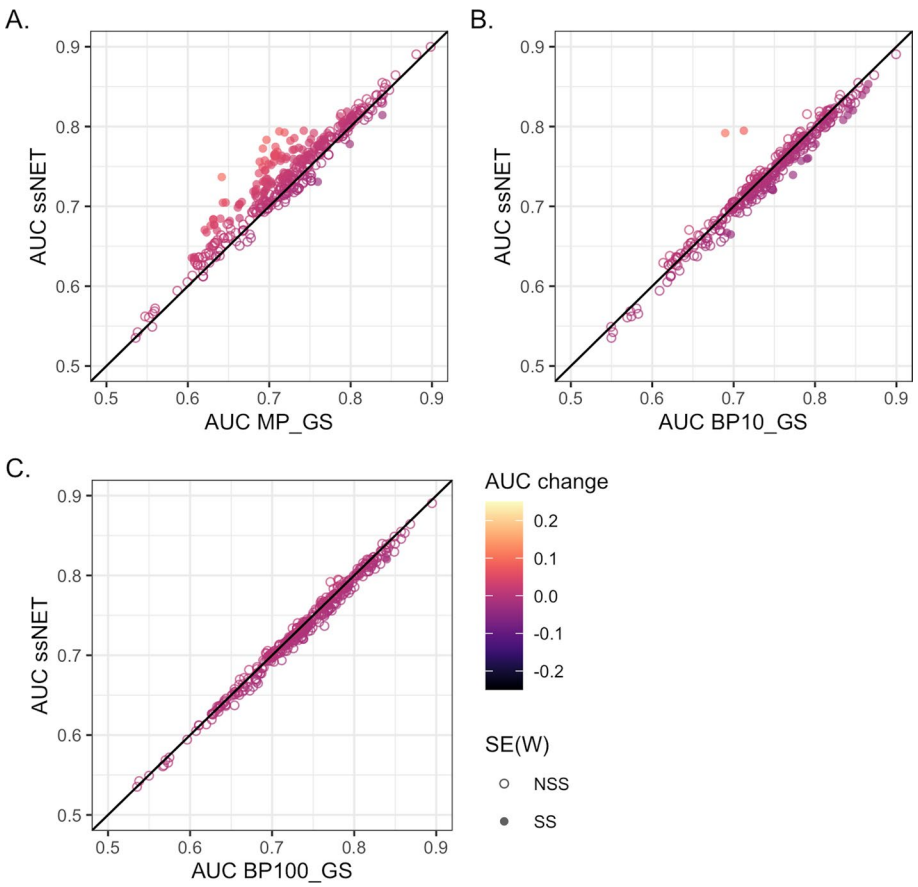

Fig. 7 ssNet evaluation. The functional prediction performance of ssNet, BioSystems and Gene Ontology-scored networks as measured by area under curve of AUC of receiver operator characteristic (ROC) plots. The error of the AUC was calculated using the standard error of the Wilcoxon statistic SE(W): not statistically significant (NSS); statistically significant (SS). **A**. BioSystems-derived (MP_GS) Gold Standard. Of 298 Gene Ontology biological processes, 325 had improved prediction using the ssNet network, and 73 using the MP_GS network. Of these changes 127 were statistically significant with 124 (98%) improved using LTP_GS. **B**. Gene ontology biological process terms annotating <10% of the genome (BP10_GS) Gold Standard. Of 398 Gene Ontology biological processes, 128 had improved prediction using the LTP_GS network, compared to 270 using BP10_GS. Of the changes just 20 were significant with 2 (10%) improved using LTP_GS. **C**. Gene ontology biological process terms annotating annotation <100 genes (BP100_GS) Gold Standard. Of 398 Gene Ontology biological processes, 102 had improved prediction using the LTP_GS network, compared to 296 using BP100_GS. Of the changes only 2 were significant, both of which were improved using BP100_GS

that term. Many high level terms, for instance metabolic process (GO:0008152), are too general to imply a realistic functional association [26, 44]. Assuming a functional link based on this term would add noise to the Gold Standard by linking many protein pairs which do not participate in the same cellular process. This problem may be overcome to some extent by ignoring the high-level terms [22, 45]. However, despite GO's hierarchical nature, the level of a term in the DAG is not necessarily indicative of a term's specificity [46, 47].

In addition, GO has a number of evidence types with some GO annotations considered more reliable than others [48]. Annotations with the IEA (inferred from electronic annotation) evidence code are considered the least reliable since they are not manually-curated. While the remaining GO annotations are manually-curated, the

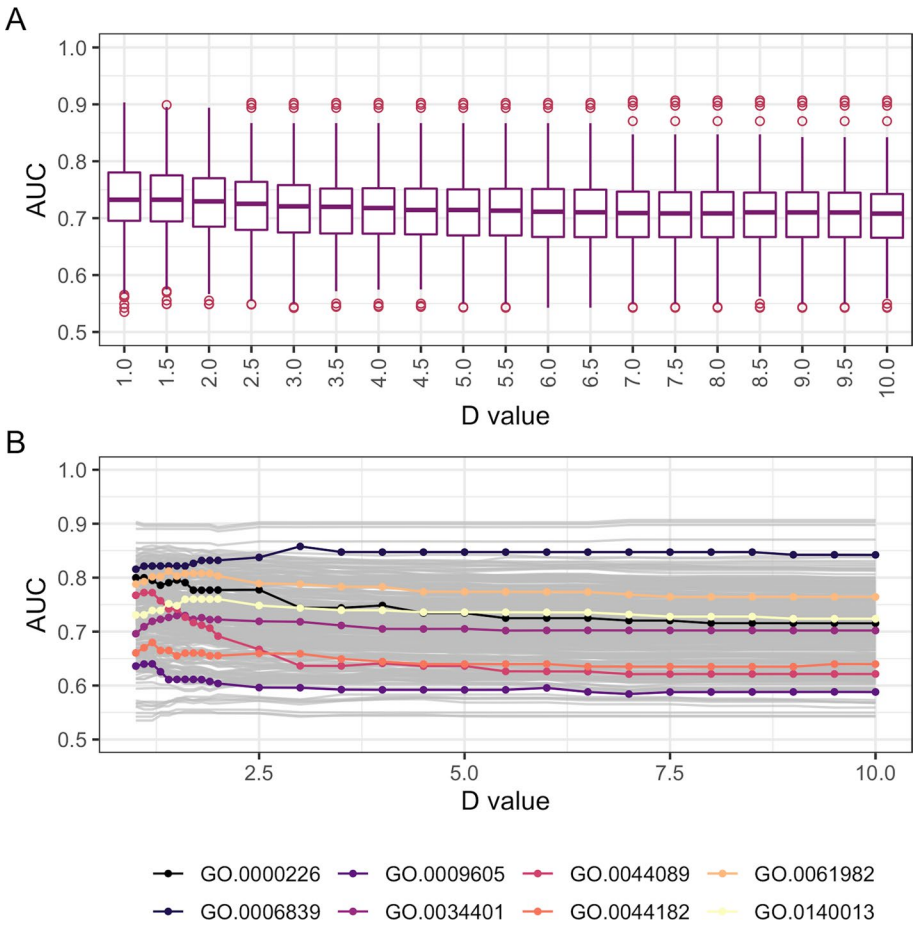

Fig. 8 The *D*-value integration parameter. **A** Area under curve (AUC) of receiver operator characteristic plots for 298 Gene Ontology biological processes at different *D*-values. **B** Variation of AUC for 298 Gene Ontology biological processes at different *D*-values. Eight individual processes are highlighted as exemplars

different evidence types are also thought to differ in their accuracy. In particular, computational evidence codes are generally considered to be less accurate than the experimental codes, and the codes ISS (inferred from sequence or structural similarity), IEP(inferred from expression pattern) and NAS (non-traceable author statement) are considered lower reliability than the other codes of this class [49]. Therefore, the choice of what data to include in a GO-based Gold Standard is highly complex and can affect the final network. In our study we chose two different thresholds based on annotation number which resulted in a similar number of proteins but vastly different number of interactions in the Gold Standard datasets. While these may not be the optimal datasets derived from GO, it is clear from our results that ssNet produces comparable performance to a GO-derived Gold Standard.

Gold Standards from metabolic pathway databases are also commonly used [1, 19, 20]. However, finding a suitable up-to-date metabolic dataset, that has no redundancy with the data to be scored, is hard. Our 2012 study was able to use monthly updates of the KEGG dataset for comparison with monthly updates of BioGRID [8]. However, KEGG is no longer freely-available and, while BioSystems represents a comprehensive collection of metabolic pathways that includes KEGG, it has not been updated since 2017.

While it was not possible to find an ideal metabolic Gold Standard for comparison, it is clear from our results that using a dataset with a different focus than the data to be scored, in this case primary metabolic interactions versus experimental interactions, can result in loss of data. Notably, the four LTP yeast data types that did not score using BioSystems (Fig. 6 B)—Positive Genetic, Protein-RNA, Synthetic Haploinsufficiency and Affinity Capture-RNA—were genetic/RNA data types. Similarly, many of the HTP datasets lost due to not scoring against BioSystems (Fig. 3) were genetic interaction types. Of those lost that were physical interaction data, the vast majority had a non-metabolic focus and four were large-scale >2000 interaction datasets [50–53].

If a network study's focus is non-metabolic, then a metabolic Gold Standard may not be suitable as relevant datasets may be lost. There were far fewer datasets with no score using ssNET and so fewer potentially-relevant datasets were discarded. Scoring and integration from a single data source in this way gives a far more complete network without bias towards specific data types. Importantly, since the data can be spilt into individual source studies, this method also ensures no redundancy between data and Gold Standard, which may bias the network.

A metabolic-focused Gold Standard may be desirable if metabolic pathways are the area of interest for downstream analyses, and this is reflected in some of the individual dataset scores. Several datasets scored higher using the metabolic Gold Standard [54–56]. However these were relatively small changes, and while some biological process terms also had better prediction using BioSystems, ssNet had comparable performance for all, in addition to the advantages of the computational ease of using a single source. Furthermore, since key external Gold Standards, such as BioSystems, are no longer being maintained, at some point they will no longer be sufficiently up-to-date for use as a Gold Standard.

The single source method also overcomes any need for potentially error prone identifier mapping [25]. Identifier mapping required significant effort during creation of our BioSystems scored networks; although mapping tools were used, some redundancy of identifiers remained that needed time-consuming manual curation. In this case we only counted entrez IDs that could be mapped to gene symbols, however, it should be noted that there were approximately 300 proteins and 1000 interactions that were in the BioSystems dataset and could not be mapped, the majority of which did not overlap with V186 BioGRID data when manually checked. Therefore if ID mapping could have been fully achieved, it would not have improved scoring results. This discrepancy is reflected in the lack of overlap between BioSystems and our source network data.

Yeast remains by far the most complete interaction dataset with data covering almost all of its ~6600 genes[7] [57]. Data coverage in yeast has levelled off since our initial study [8], in particular the fluctuations in protein coverage seen prior to 2010 are no longer observed, reflecting BioGRID's efforts to improve and standardise curation methods [9, 58]. Although the overlap in terms of interactions between HTP and LTP data remains relatively low, it still provides enough overlap to allow scoring. While several datasets scored infinity against the BioSystems and GO-derived Gold Standards, very few did so

[7] 79% verified; https://www.yeastgenome.org/genomesnapshot on 15th July 2020.

using the LTP data. Infinity scores are hard to deal with since they must be assigned an arbitrary high score (see "Methods"). However, this is not ideal as demonstrated by the scores of the nine datasets in Fig. 3, which have a range of scores using LTP data. By reducing or removing entirely both zero and infinity scores, an LTP-derived Gold Standard produces more accurate scoring and better coverage of the data. Moreover, as network-based analyses move beyond model organisms to non-models [59], which often lack traditional Gold Standard data, an LTP-derived dataset provides the means to intergrate PFINs in more diverse species.

The main disadvantage of many PFINs is that the data considered highest quality is used in scoring and not included in the final network [1, 22]. We used an iterative method to score the LTP data by type, before integration with the HTP datasets. While the LTP scores were generally higher than HTP (Fig. 5A), they are in an overlapping range indicating a difference in quality between the data types, and that they can be integrated without introducing significant bias towards the LTP data.

Interestingly, the functional prediction performance of the ssNet networks did not differ from that of the HTP-only networks to any great extent. However, assessing and comparing networks is not straightforward. Here, we measured the quality of networks by prediction of known annotations to produce a numeric measure of network performance for each biological process. While the network performance at functional prediction was similar, PFIN analyses are often performed in an exploratory manner. The increased connectivity provided by inclusion of the LTP data is likely to be beneficial for other network-based analysis, for instance in clustering and subnetwork analysis [60–62].

We chose thresholds based on our previous studies in yeast [8, 19] but it is possible that different values will be more appropriate in other species, in particular the LTP threshold of 100 interactions. There are also other ways to split the data than by PubMed ID; some studies contain both HTP and LTP data and/or more than one data type. For instance large-scale HTP screens may also include LTP confirmation of some interactions [63] and studies may combine different data types [64]. However, the focus of many studies is in itself something that can be harnessed during integration [19], and we believe keeping each study as a single dataset and assessing it as a whole is preferable. Importantly, due to BioGRIDs standardised format other data sources, including unpublished data, can easily be added to the integration as described on the ssNET repository[8].

While the ssNET integration method could be applied using any quality scoring metric, we chose that of Lee and colleagues [1] since it has been used extensively in multiple species [22, 65–71]. A *D*-value of 1.0 was used in our yeast network during integration (see "Methods", Equation 2), which produces a sum of the dataset confidence scores. Higher *D*-values increase the contribution of highest quality datasets to the network, and the optimal *D*-value is dependent on the area of biology being studied. Producing networks and evaluations at all possible values would produce a vast amount of data, and is future work for further study in yeast and other species. However, we have provided networks, dataset scores and AUCs at multiple *D*-values (Fig. 8), in order to allow

[8] https://github.com/kj-intbio/ssnet/.

selection of the most appropriate data, method and parameters depending on area of study.

Conclusions

The ssNet method provides a computationally amenable one-step PFIN integration method for functional interaction data which has a number of benefits:

- PFINs are generated from a single data source without using an external Gold Standard.
- The ssNet PFINs are of a quality that is comparable to that of the external Gold Standard PFINs, and exceeds it by some metrics.
- The ssNet PFINs incorporate more information from the data source into the final network than the Gold Standard PFINs, increasing over-all network coverage.
- Integration is easier and faster, overcoming the challenges of data redundancy, Gold Standard bias and ID mapping.
- The highest quality Gold Standard data can be consistently integrated into the network adding to its quality.

Acknowledgements
Not applicable.

Author contributions
KJ designed the study and conducted the analyses. KJ and MP wrote the network integration code. KJ, MP, AW, SJC and AA wrote the manuscript. All authors read and approved the final manuscript.

Funding
AA is sponsored by Saudi Electronic University.

Availability of data and materials
The networks produced in this study are available at https://figshare.com/projects/Yeast_ssNet/114366 and the code for ssNet network generation at https://github.com/kj-intbio/ssnet/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 November 2021 Accepted: 11 July 2022
Published online: 25 July 2022

References

1. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science. 2004;306:1555–8.
2. Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. J Comput Biol. 2004;11:463–75.
3. Xia K, Dong D, Han JDJ. IntNetDB v1. 0: an integrated protein-protein interaction network database generated by a probabilistic model. BMC Bioinform. 2006;7:508.
4. Shim H, Kim JH, Kim CY, Hwang S, Kim H, Yang S, et al. Function-driven discovery of disease genes in zebrafish using an integrated genomics big data resource. Nucleic Acids Res. 2016;44:9611–23.
5. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. Bioinformatics. 2007;23:2322–30.
6. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol. 2004;7:535–45.

7. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinform. 2007;8:236.
8. James K, Wipat A, Hallinan J. Is newer better?-evaluating the effects of data curation on integrated analyses in Saccharomyces cerevisiae. Integr Biol (Camb). 2012;4:715–27.
9. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Rcids Res. 2019;47:D529–41.
10. Huttenhower C, Troyanskaya OG. Assessing the functional structure of genomic data. Bioinformatics. 2008;24:i330–8.
11. Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. Nat Rev Genet. 2007;8:699–710.
12. Marcotte E, Date S. Exploiting big biology: integrating large-scale biological data for function inference. Brief Bioinform. 2001;2:363–74.
13. Huang H, Jedynak BM, Bader JS. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. PLoS Comput Biol. 2007;3: e214.
14. Mrowka R, Patzak A, Herzel H. Is there a bias in proteome research? Genome Res. 2001;11:1971–3.
15. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet. 2002;18:529–36.
16. Rigden DJ, Fernández XM. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Res. 2020;48:D1–8.
17. Wren JD, Bateman A. Databases, data tombs and dust in the wind. Bioinformatics. 2008;24:2127–8.
18. Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. Big data: the future of biocuration. Nature. 2008;455:47–50.
19. James K, Wipat A, Hallinan J. Integration of full-coverage probabilistic functional networks with relevance to specific biological processes. In: Paton NW, Missier P, Hedeler C, editors. Data Integration in the Life Sciences. DILS 2009. Lecture Notes in Computer Science. Springer, Berlin; 2009;31–46.
20. Yamanishi Y, Vert JP, Kanehisa M. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics. 2004;20:i363–70.
21. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.
22. Lee I, Li Z, Marcotte EM. An improved, bias-reduced probabilistic functional gene network of baker's yeast. Saccharomyces cerevisiae. PloS One. 2007;2: e988.
23. Lee I, Seo YS, Coltrane D, Hwang S, Oh T, Marcotte EM, et al. Genetic dissection of the biotic stress response using a genome-scale gene network for rice. Proc Natl Acad Sci USA. 2011;108:18548–53.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
25. Drăghici S, Sellamuthu S, Khatri P. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. Bioinformatics. 2006;22:2934–9.
26. Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG. Finding function: evaluation methods for functional genomic data. BMC Genomics. 2006;7:187.
27. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.
28. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019;47:D23–8.
29. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. 2004;32:D262–6.
30. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res. 2002;30:69–72.
31. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
33. Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 2008;24:282–4.
34. Linghu B, Snitkin ES, Holloway DT, Gustafson AM, Xia Y, Delisi C. High-precision high-coverage functional inference from integrated data sources. BMC Bioinform. 2008;9:119.
35. Henderson AR. Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann Clin Biochem. 1993;30:521–39.
36. Carey V, Redestig H. ROC: utilities for ROC, with microarray focus; 2019. R package version 1.62.0. Available from: http://www.bioconductor.org.
37. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
38. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, et al. The NCBI BioSystems database. Nucleic Acids Res. 2010;38:D492–6.
39. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.
40. Yu J, Finley RL. Combining multiple positive training sets to generate confidence scores for protein–protein interactions. Bioinformatics. 2009;25:105–11.
41. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci USA. 2003;100:8348–53.
42. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, et al. Discovery of biological networks from diverse functional genomic data. Genome Biol. 2005;6:R114.

43. Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. Nucleic Acids Res. 2004;32:6414–24.
44. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform. 2006;7:302.
45. Wu X, Zhu L, Guo J, Zhang DY, Lin K. Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res. 2006;34:2137–50.
46. Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics. 2010;26:1759–65.
47. Wang J, Zhou X, Zhu J, Zhou C, Guo Z. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinform. 2010;11:290.
48. du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology-a primer for bioinformaticians. Brief Bioinform. 2011;12:723–35.
49. Lee I, Marcotte EM. Effects of functional bias on supervised learning of a gene network model. Methods Mol Biol. 2009;541:463–75.
50. Miller JE, Zhang L, Jiang H, Li Y, Pugh BF, Reese JC. Genome-wide mapping of decay factor-mRNA interactions in yeast identifies nutrient-responsive transcripts as targets of the deadenylase Ccr4. G3 (Bethesda). 2018;8:315–30.
51. Jungfleisch J, Nedialkova DD, Dotu I, Sloan KE, Martinez-Bosch N, Brüning L, et al. A novel translational control mechanism involving RNA structures within coding sequences. Genome Res. 2017;27:95–106.
52. Wang Y, Zhang X, Zhang H, Lu Y, Huang H, Dong X, et al. Coiled-coil networking shapes cell molecular machinery. Mol Biol Cell. 2012;23:3911–22.
53. Batisse J, Batisse C, Budd A, Böttcher B, Hurt E. Purification of nuclear poly(A)-binding protein Nab2 reveals association with the yeast transcriptome and a messenger ribonucleoprotein core structure. J Biol Chem. 2009;284:34911–7.
54. Chymkowitch P, Nguéa PA, Aanes H, Robertson J, Klungland A, Enserink JM. TORC1-dependent sumoylation of Rpc82 promotes RNA polymerase III assembly and activity. Proc Natl Acad Sci USA. 2017;114:1039–44.
55. Buser R, Kellner V, Melnik A, Wilson-Zbinden C, Schellhaas R, Kastner L, et al. The replisome-coupled E3 ubiquitin ligase Rtt101^{Mms22} counteracts Mrc1 function to tolerate genotoxic stress. PLoS Genet. 2016;12: e1005843.
56. Szappanos B, Kovács K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet. 2011;43:656–62.
57. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274:546,563-546,567.
58. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Res. 2017;45:D369–79.
59. James K, Wipat A, Cockell SJ. Expanding interactome analyses beyond model eukaryotes. Brief Funct Genomics [in press]. 2022.
60. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci USA. 2002;99:7821–6.
61. Lancichinetti A, Kivelä M, Saramäki J, Fortunato S. Characterizing the community structure of complex networks. PloS One. 2010;5: e11976.
62. Voevodski K, Teng SH, Xia Y. Finding local communities in protein networks. BMC Bioinform. 2009;10:297.
63. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005;122:957–68.
64. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature. 2005;437:1173–8.
65. Kim H, Shim JE, Shin J, Lee I. EcoliNet: a database of cofunctional gene network for Escherichia coli. Database. 2015;2015:bav001.
66. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol. 2010;28:149–56.
67. Lee T, Oh T, Yang S, Shin J, Hwang S, Kim CY, et al. RiceNet v2: an improved network prioritization server for rice genes. Nuc Acids Res. 2015;43:W122–7.
68. Lee S, Lee T, Yang S, Lee I. BarleyNet: a network-based functional omics analysis server for cultivated Barley. Hordeum vulgare L. Front Plant Sci. 2020;11:98.
69. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genet. 2008;40:181–8.
70. Lee T, Lee S, Yang S, Lee I. MaizeNet: a co-functional network for network-assisted systems genetics in Zea mays. Plant J. 2019;99:571–82.
71. Kim E, Hwang S, Kim H, Shim H, Kang B, Yang S, et al. MouseNet v2: a database of gene networks for studying the laboratory mouse and eight other model vertebrates. Nuc Acids Res. 2016;44:D848–54.

Publisher's note