Bioinformatics, 32(7), 2016, 1065–1073
doi: 10.1093/bioinformatics/btv734
Advance Access Publication Date: 14 December 2015
Original Paper

OXFORD

Data and text mining

Positive and negative forms of replicability in gene network analysis

W. Verleyen, S. Ballouz and J. Gillis*

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard Woodbury, NY 11797, USA

*To whom correspondence should be addressed.

Associate Editor: Igor Jurisica

Received on 13 July 2015; revised on 7 December 2015; accepted on 9 December 2015

Abstract

Motivation: Gene networks have become a central tool in the analysis of genomic data but are widely regarded as hard to interpret. This has motivated a great deal of comparative evaluation and research into best practices. We explore the possibility that this may lead to overfitting in the field as a whole.

Results: We construct a model of 'research communities' sampling from real gene network data and machine learning methods to characterize performance trends. Our analysis reveals an important principle limiting the value of replication, namely that targeting it directly causes 'easy' or uninformative replication to dominate analyses. We find that when sampling across network data and algorithms with similar variability, the relationship between replicability and accuracy is positive (Spearman's correlation, $r_s \sim 0.33$) but where no such constraint is imposed, the relationship becomes negative for a given gene function ($r_s \sim -0.13$). We predict factors driving replicability in some prior analyses of gene networks and show that they are unconnected with the correctness of the original result, instead reflecting replicable biases. Without these biases, the original results also vanish replicably. We show these effects can occur quite far upstream in network data and that there is a strong tendency within protein–protein interaction data for highly replicable interactions to be associated with poor quality control.

Availability and implementation: Algorithms, network data and a guide to the code available at: https://github.com/wimverleyen/AggregateGeneFunctionPrediction.

Contact: jgillis@cshl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Increasingly, biologists have turned to computational methods to sift through the vast array of pre-existing genomics data for validation that a gene has a molecular role in the phenotype of interest or to prioritize a candidate as disease causal (Moreau and Tranchevent, 2012; Wang and Marcotte, 2010). These computational methods usually fit under the rubric of 'machine learning' and use network data that represent the interaction of genes or their products. Many of these computational methods depend on a form of 'guilt by association', in which a gene is inferred to possess a particular function based on its similarity to other genes with that function (Oliver, 2000). The most common form of similarity used in these tasks is that of genomic sequence similarity which is easily implemented through supervised use of BLAST (Altschul *et al.*, 1990) and comparatively straightforward to interpret. While sequence-based analysis is essentially routine within biology, one of the promises of systems biology has been to extend the form of 'association' used to relate genes to potentially subtler relationships, such as protein–protein interaction (PPI), co-expression, genetic interaction or phylogenetic profiles. Systems-based prediction of gene function has found particular application in the interpretation of disease-causal variants due to the difficulty of finding overlaps in

known functions among candidate genes (Geschwind, 2008; Greene and Troyanskaya, 2012; Oellrich *et al.*, 2012). However, progress in the context of both data and methodology has been surprisingly uncertain (Pavlidis and Gillis, 2013).

The need for better assessment of methods in function inference and network analysis is widely recognized and has led to numerous field-wide evaluations, often called critical assessments (Bornigen *et al.*, 2012; Kryshtafovych *et al.*, 2014; Pena-Castillo *et al.*, 2008; Radivojac *et al.*, 2013). The two principal goals of critical assessments are (i) to make the performances of individual methods less prone to overfitting and (ii) for comparisons between methods to be within the same framework. Overfitting is minimized since participants are truly blind to the success of their method prior to assessment and thus cannot 'tailor' their solutions to the benchmarking metric. Gene networks possess unusually prominent consensus resources [e.g. the Gene Ontology (GO) (Ashburner *et al.*, 2000), BioGRID (Stark *et al.*, 2006)], making evaluation within a well-defined framework possible. By reducing overfitting and making methods directly comparable, critical assessments endeavor to make science more replicable; their outputs and comparative evaluations can be trusted to generalize.

The difficulty of characterizing the features in gene networks that drive successful uses has contributed to making replicability in their output, which can be more easily measured, particularly important to evaluation within their critical assessments [e.g. the DREAM challenges (Marbach *et al.*, 2012) and the Critical Assessment of protein Function Annotation algorithms, CAFA challenge (Radivojac *et al.*, 2013)]. In performing this evaluation, critical assessments are simply performing a more top–down version of the usual scientific process of refinement through replicability (Fisher, 1935). While this may be desirable in some ways, it creates a new potential for overfitting for the field in its entirety. We decided to explore this possibility by simulating multiple gene function prediction tasks and outcomes and hence the field of gene network analysis as a whole.

In our model of research in gene network analysis, each separate researcher is represented by an individually developed machine learning algorithm with access to particular data. The algorithms are both diverse and in common use for diverse bioinformatics problems and thus reasonably reflecting ordinary practice. The data resources (or 'library') given to these algorithms are similarly diverse and frequently used sources of human gene network information, varying from individual expression profiles to consensus pathway information. We refer to a specific combination of algorithm and data as a 'researcher' (Fig. 1). For example, a researcher may consist of the algorithm 'random walk with restarts' using specific co-expression data. The individual sampled resources do not represent partial data sets but rather ones which are at least as comprehensive as is typical of any given study. Because it is a central characteristic by which we judge results, our focus is on using these model researchers to understand replicability in gene network analysis. After deriving general principles through our simulations, we focus on two important applications affecting the interpretation of disease genes and PPI data in current research, with a focus on psychiatric genetics.

2 Methods

Our analysis occurs in two parts. In the first, we build a model of researchers assessing gene networks data, and in the second, we work through an application using PPI network (PPIN) data to characterize genes linked to autism and schizophrenia (SCZ). To build models of researchers, each using a network analysis method and data, we need to assemble these resources. In general, each of our network analysis methods is operating as a gene function prediction method. These methods consist of three components: functional annotations, biological data or network and a machine learning algorithm. We describe the functional annotations in Section 2.1, data resources in Section 2.2 and algorithms in Section 2.3. Because our model is concerned with the behavior of these methods in comparison to one another, we then describe methods for evaluating their aggregate accuracy (Section 2.4) and replicability (Section 2.5). We close our model analysis by evaluating variation at two different time points (corresponding to the start and close of the project, Section 2.6). We then move to an application of the principals derived from the model in the characterization of network properties of genes linked to autism and SCZ (Section 2.7). This suggests replicable interactions in PPI data may be problematic, which we directly evaluate using quality control data, with methods described in Section 2.8.

2.1 Ontology and annotations (GO)

The GO (revision 1.1363) (Ashburner *et al.*, 2000) and GO annotations (date: April 23, 2014) were used as gene annotations for gene function prediction. We first propagated the genes through the GO hierarchy and then filtered for GO terms with associated gene sizes ranging between 20 and 1000 genes and excluded associations with evidence codes from IEA (inferred from electronic annotation). Filtering GO terms within this range shows stable performance (Gillis and Pavlidis, 2011). A total of 2930 GO terms fit this criterion. To minimize selection biases, we considered the fixed set of genes with at least one GO annotation and which were present in our microarray expression data; this totaled 12 529 annotated genes. To make the GO analysis more tractable, we used a filtered subset of GO terms from GO slim (date: April 24, 2014), totaling 109 GO terms developed by the GO consortium. It shows similar performance when compared to the filtered subset of complete GO and was therefore appropriate for our analyses (Verleyen *et al.*, 2015).

2.2 Data resources and gene sets for network construction

We collected three different types of gene association/interaction data for our networks: (i) PPI, (ii) semantic similarity and (iii) co-expression. Each data resource was parsed into a network, described in more detail in the following sections. We converted all protein IDs and gene symbols into gene Entrez IDs using HUGO (White *et al.*, 1997). As above (Section 2.1), all the genes that had a gene association in the GO and that overlapped with the co-expression data, totaling 12 529 genes, were used as our basis gene set. We also restricted our genes to this same set once constructing networks based upon PPI and semantic similarity data.

2.2.1 Protein–protein interaction networks

We constructed PPINs from five different databases: (i) BioGRID (Chatr-aryamontri *et al.*, 2013), (ii) HIPPIE (Schaefer *et al.*, 2012), (iii) IntAct (Orchard *et al.*, 2014), (iv) I2D (Brown and Jurisica, 2007) and (v) GeneMANIA (Zuberi *et al.*, 2013). Each database has gene–gene or PPIs listed, which were used to create a binary network or a weighted network, depending on the available information. Data from the BioGRID database (version 3.2.111) were used to construct a binary network from all physical interactions and no further filtering on experimental type

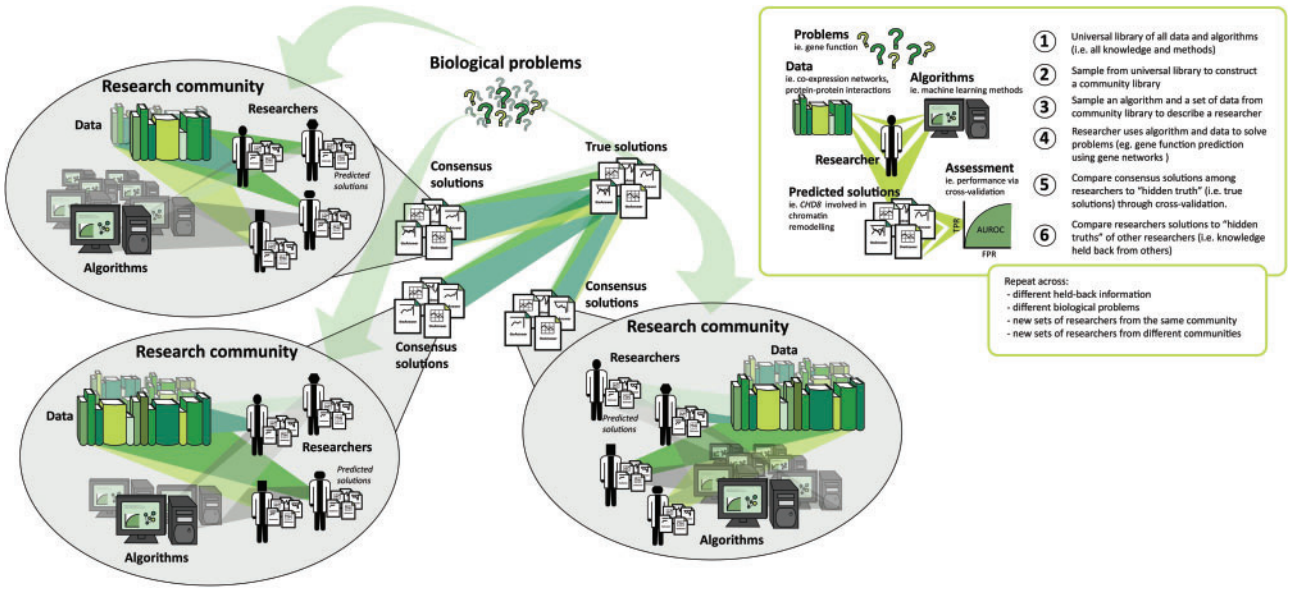

Fig. 1. Modelling replicability in gene network analysis. A library of algorithms and data is sampled from the complete library to be made available to a given research community. The set of researchers for that community is constructed by sampling from their available resources. Each instantiated researcher tries to predict gene function held back from that community based on prior knowledge. Replicability is assessed by comparing each researcher to the consensus of the remainder while performance of the total research community (consensus) is assessed with reference to the held back knowledge (cross-validation)

was applied. The HIPPIE database (version 1.6) was used to construct a weighted network, with no filtering on the data. The IntAct database (version 2.0; downloaded on April 22, 2014) was used to construct a weighted network, filtered on interactions from *Homo sapiens* (taxonomic ID 9606). The I2D database (version 2.3) was used to construct a binary network. The interactions are filtered from the following original databases in I2D: BioGRID, IntAct, HPRD, BIND, MINT and INNATEDB. The GeneMANIA data (version date October 24, 2013) was used to construct a binary network. For all networks, only the physical interactions from the databases were incorporated. Summary network properties of the networks are shown in Supplementary Table S1.

2.2.2 Semantic similarity networks

We constructed semantic similarity profiles based upon five databases: (i) KEGG (Ogata *et al.*, 1999), (ii) Reactome (Joshi-Tope *et al.*, 2005), (iii) Phenocarta (Portales-Casamar *et al.*, 2013), (iv) InterPro (Hunter *et al.*, 2012) and (v) Pfam (Finn *et al.*, 2014). The semantic similarity is defined as equal to the Jaccard index between two genes over the set membership they exhibit in each of the five listed databases (Mistry and Pavlidis, 2008). Weighted semantic similarity networks were constructed from the semantic similarity of all gene pairs, so that the weight of each gene–gene edge in the network is equal to the Jaccard index. The KEGG database was used to construct a similarity profile based upon 274 biological pathways. The Reactome database (version date May 7, 2014) was used to construct a similarity profile based upon 128 biological pathways. The Phenocarta database (version date May 1, 2014) was used to construct a similarity profile based upon 2923 terms. The InterPro database (version 46.0) was used to construct a similarity profile based upon 7137 shared protein domains. The Pfam database (version 27.0) was used to construct a similarity profile based upon 6142 shared protein domains. Summary network properties of the networks are shown in Supplementary Table S2.

2.2.3 Co-expression networks

We generated aggregate co-expression networks constructed from publically available expression data sets (Gene Expression Omnibus and Sequence Read Archive), using either RNA-seq or microarray-based studies as previously described (Ballouz *et al.*, 2015). For each aggregate network, 20 individual co-expression networks were selected. Each co-expression network was constructed by taking the Spearman correlation coefficient of the gene pairs as an edge weight. Each value was then ranked, and the aggregate was the sum of these ranked weights. The aggregate network was then thresholded for the top 1% connections. Four such aggregate networks were constructed, two from microarray datasets and two from RNA-seq data sets. The individual experiments for each aggregate are listed in Supplementary Table S3.

2.3 Machine learning algorithms

A set of six machine learning algorithms were chosen in our analyses. We selected algorithms that are well established for gene function prediction and those that belong to different machine learning categories. We picked three network inference algorithms based upon different mathematical formalizations: (i) neighbor voting, (ii) GeneMANIA and (iii) random walk with random restart (implemented for this project specifically, see our supporting information and its earlier use for general properties). These algorithms typically exploit topological characteristics of the network. We also implemented less specialized algorithms which interpret network data as sets of features (i.e. the feature data for a given gene is its connectivity profile with other genes). We selected (iv) logistic regression as it is perhaps the most fundamental machine learning algorithm for building classifiers as well as two online or lazy setting algorithms, (v) support vector machine with a stochastic gradient descent solver and (vi) the passive aggressive approach. All three of these more general methods were implemented using sckit-learn (Pedregosa *et al.*, 2011). The output of each gene function prediction task is, for a given function, a vector of ranked values (across all the genes)

indicating the probability of the gene belonging to the function. Performance is calculated using 3-fold cross-validation. We describe each algorithm in more detail in the Supplementary Information and have previously benchmarked them in yeast (Verleyen *et al.*, 2015). We repeat the benchmarking task on human data (see Supplementary Figs S1–S3). We calculate Spearman correlation coefficients (r_s) throughout using the scipy.stats.spearmanr function, which also calculates a p value from a two-sided test of non-correlation (http://www.scipy.org/).

2.4 Aggregation of methods

Aggregation is a fundamental approach to create more robust computational models and improve overall performance of these models under their given task. We have performed aggregation at the level of the output scores of a predictor. We define a predictor as the output from a gene function prediction method containing an algorithm and a parsed network. For algorithm aggregation, we used the same network and aggregated the output scores of each algorithm. For data aggregation, we used the same algorithm on different networks and combined their scores. Note that our results are robust to more sophisticated aggregation strategies, such as weighting (e.g. correlation-based feature selection).

2.5 Replicability

In our application of replicability in gene function prediction, we have variability stemming from the choice of machine learning algorithms and data resources. We calculated the replicability in our gene function prediction task by measuring the degree to which a given held-out predictor was 'validated' by the consensus (treated as a gold standard). More precisely, we held back knowledge of gene-function from all methods and ran all the possible predictors (i.e. algorithm and data resource combination). Leaving out the vector of scores of one predictor, we created a consensus solution by averaging the scores for each gene from the other predictors. Using this consensus solution, we created a new label vector with the top 10 genes labeled as positives and the other genes labeled as negatives. The area under the receiver operating characteristic curve (AUROC) is computed based upon these new labels and scores from the prediction of the held-out algorithm. In other words, the held-out algorithm is validated in its predictions of held out gene-function data not by reality, but by the consensus among other algorithms. This was iterated over all possible combination of predictors. The final measure of replicability was the averaged AUROC over all the iterations.

2.6 Temporal variation

To examine the degree to which the trends we observed might be changing with time or reflect a temporary snapshot of the data, we re-ran analyses after freezing all data on April 24, 2014 and then updating any relevant data to that available on August 20, 2015. This duration covered roughly the beginning of final analyses for the project (first freezing) to midway through review (updating). The updated resources are listed in Supplementary Table S4.

2.7 Analysis of psychiatric disorder studies

In our application, we analyze topological network characteristics from disease-associated genes and compare to that of randomly constructed distributions from null networks. The random distributions from Monte Carlo-based methods are typically based upon node permutation (i.e. shuffling the nodes of the network). An alternative is to shuffle across edges or links (e.g. in the list of gene pairs giving connections in the network, permute among all of the second of the pair to create random connectivities that preserve node degree). Using link permutations instead of node permutations allows us to test for a selection bias related to the node degree of the genes in the gene list. To demonstrate that these biases influence results and interpretations in real research problems, we selected gene lists from major studies on autism spectrum disorder (ASD) (O'Roak *et al.*, 2012) and SCZ (Gulsuner *et al.*, 2013) and performed the Monte Carlo experiment based upon link permutation alongside the original analyses.

2.8 Quality control for PPIN data

To further study the paradoxical effects of replicability, we looked at the relationship between recurrence, a common metric of replicability and quality control in the PPI databases previously described. We used data on protein contaminants from the CRAPome (Mellacheruvu *et al.*, 2013) (version 1.1, *Homo sapiens*, date: January 1, 2014). This database contains the spectral counts of 8473 proteins identified in controls across a collection of 411 affinity capture mass spectrometry (AC-MS) experiments. For each protein in the CRAPome, a quality control 'reliability' score was calculated as the average of spectral counts across all the experiments. For proteins missing from the CRAPome, they were given a score of 0. However, we also assessed missing proteins by giving them a score of NA; all results are robust to this choice. Then, for each PPI in the given PPI database, we calculated the quality of the interaction score as the rank of the sum of the reliability scores of the bait and the prey proteins. We then measured the recurrence of a PPI in the individual PPI databases as the count of how many individual studies the protein pair appeared in and compared that to the reliability score.

3 Results

Our approach to formalizing replicability is to treat it exactly parallel to how performance is conventionally assessed. In general, performance is measured by determining if a researcher can correctly predict some unknown (or held back) result; likewise, we measure replicability by measuring how well a researcher correctly predicts the consensus across other researchers. That is, there are conventional metrics for assessing whether a given researcher's answer is similar to the truth; in measuring replicability, we perform the identical assessment but treat the consensus output among other researchers as the truth against which a given researcher is evaluated. We assess this using the area under the receiver operating characteristic curve (AUROC) in both cases (see Section 2 for further details). Note that all of our analysis is readily reproducible, by which we simply mean that analyses can be re-done, which we differentiate from replicability involving independent analysis, the phenomena we are modelling.

3.1 Modelling replicability

3.1.1 Concurrence among either algorithms or data predicts improved accuracy

In our first set of model experiments, the algorithms are using various data to predict gene functions as annotated in GO. We first consider researchers sampling from different machine learning algorithms using a single aggregated co-expression network resource (summated across 80 independent transcriptomic experiments and totaling 5672 separate expression samples; see Fig. 2A). In using this data, the more replicable the researcher output, the likelier the joint

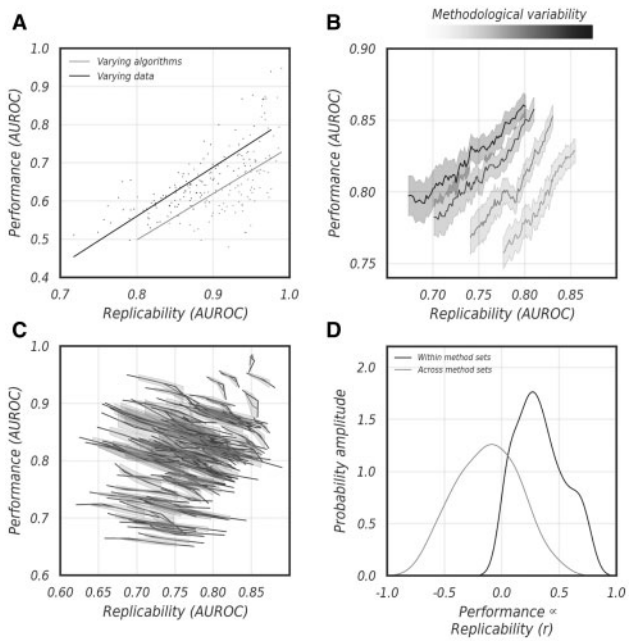

Fig. 2. Meta-analytic properties of predicted gene functions. Performance is assessed by conventional cross-validation against held-out true positives. Replicability is assessed by cross-validation against consensus predictions from held-out researchers. (**A**) Assessing replicability and performance in co-expression data for GO groups (points) and showing two research communities, one where the researchers vary only by algorithm (gray) used on consensus data and one where the researchers vary only by data using a consensus method (black). (**B**) Research communities are constructed by sampling from across data and algorithms with research communities drawing on resources of different degrees of variability (or independence). These research communities are grouped into quartiles by this variability and aggregate performance and replicability are plotted for each set of research communities. The mean (line) is plotted along with the standard deviation (shadow); window size 35. (**C**) The variability in performance and replicability for each gene function across the different research community quartiles. Smoothed with a window size of 2, so the principal independent observation is that the slopes are uniformly negative. (**D**) The relationship between performance and replicability within each research community is positive (black; $r_s = 0.333$), but for a given gene function, the performance is negative across research communities (gray; mean $r_s = -0.125$)

output is to be correct in predicting which genes possess a given function (Fig. 2A, gray line, Spearman's $r_s \sim 0.66$, $P \sim 4.7$ E-15). Alternatively, we can consider using a common algorithm for all researchers, with each researcher using a co-expression network derived from different data (Fig. 2a, black line). In this case, the correlation between replicability and truth is also quite high (Spearman's $r_s \sim 0.85$, $P \sim 1.77$ E-31). This positive relationship between replicability and performance is maintained in virtually every case, with researchers sampling from diverse data and algorithms (see Supplementary Figs S4–S7). Note, however, that the line describing researchers which vary by algorithm sits below the line in which researchers vary by data. That is, for a given replicability, alignment with truth is higher if the researchers used different data.

3.1.2 Concurrence within a fixed set of algorithms and data predicts improved accuracy

We now generalize from our previous case and construct communities of researchers sampling randomly from combinations of algorithms and data. Researchers may sample from co-expression, PPI and pathway data and use a variety of pre-existing algorithms to 'guess' gene function. Because we are modeling the behavior of these researchers, we vary the scenarios under which they operate. In our case, that takes the form of varying the degree of independence among the resources these researchers sample. So, for example, we simulate all the cases in which the field as a whole uses only a single algorithm on (a variety of) PPI data, or similarly, the researchers use either co-expression or PPI data across five algorithms, etc. We call each of these sets of simulations, which draw upon particular data or algorithms, 'research communities' (Fig. 1) and we can assess replicability and performance for any given research community. The research communities can be regarded as describing the state of the field as a whole given the parameters describing what algorithms and data resources are available (and how variable they are). In total, we analyzed 266 research communities, each running 10 researchers, repeated 10 times for a given parameter set, with each researcher making predictions for 109 GO functions across 12 529 genes (across 3-folds).

Our first analysis of these research communities is to determine whether the relationship between replicability and accuracy varies based on independence, i.e. the underlying variability of data resources in the research communities. We characterize the degree of researcher independence by the probability of the researchers within a research community of having sampled from data of the same modality (PPI data or semantic data). For example, when each researcher within the research community is sampling from the same data modality, it is considered to be a case of low methodological variability or low independence. At this stage, we divide the data into four groups (quartiles) depending on the fraction of data a research community uses which is of the same modality (see Supplementary Fig. S8). While we describe these as 'more dependent' researchers, they are all independent analyses as the term is typically used. Their variation in dependence is more like that between, for example, research groups interested in similar data types. The trend seen in the expression data remains true in this broader model: researchers always show a positive relationship between replicability and performance, but as they become more dependent, they sit further to the bottom-right of the performance-replicability space. Averaging the research communities into quartiles by independence, we can plot the average relationship between replicability and performance (Fig. 2B). In these quartile groupings, the relationship is strongly positive (Spearman's $r_s > 0.34$); however, the more variability in the 'library' for the research community, the better the performance for a given replicability.

3.1.3 Sampling from algorithms and data with improved joint replicability yields lower accuracy

Because each set of model researchers was considering the same set of scientific questions (i.e. which genes have a particular function), we can determine the correlation between replicability and scientific truth for each such scientific question across our quartiles. This is plotted in Figure 2C and is negative for any given scientific question. In other words, if we are asked a given scientific question, the more replicable the answer for a research community, the less likely it is to be true (across research communities). This is similar to asking what types of practices in science are 'good' ones which lead research communities to converge on accurate information. As a commonplace example of this effect in practice, we might suppose that removing genetic variation in model organisms through inbreeding makes replicability easier to achieve but should make results less meaningful for a given degree of replication (artifactual properties can now dominate replication). Generalization outside of the model

organism would be expected to become harder. This is true after grouping communities by variability; we next assess whether it is true across all communities without such grouping.

Our quartile plots show a very strong average trend, but the approximate effect is visible within virtually every research community. The correlation between replicability and truth is positive for a given research community (Fig. 2D, Spearman's $r_s \sim 0.33$), but the distribution of correlations is negative across research communities for a given scientific question (Fig. 2D, Spearman's $r_s \sim -0.13$). This result arises through a generalized version of the Yule-Simpson effect (Bickel *et al.*, 1975): it is possible for replication to be useful in assessing truth for every fixed level of dependence in experimental design but have negative value in assessing the truth of a given scientific question overall. In essence, some results will replicate more easily than others not because they are correct, but because methods or data have been more tightly controlled. The less diversity in methods and data, the less likely we are to converge on the truth in aggregate.

3.1.4. Temporal variation in replicability trends

We initially froze data on April 24, 2014, in our analyses and updated available resources to August 20, 2015, to test for variation in any of our reported results (Supplementary Tables S4–S6). This only affected our semantic and PPI network data, since the co-expression networks reflect particular experimental data and are not updated meta-analytic resources themselves. For this analysis, all aspects other than the updated data were held constant; i.e. set partitioning in cross-validation and the exact combination of data and methods each simulated researcher sampled in each case is identical. That is, we are not just holding the sampling distributions constant, but the actual 'random' selection.

We assessed each algorithm in each of the seven network resources which underwent updates in this interval. We characterize each combination by its average performance in cross-validation on the GO slim prediction used throughout, at the two time points. The correlation of performances between the two time points is quite high (Spearman's $r_s = 0.868$) and nearly follows the identity line (Supplementary Fig. S9). While this correlation in performance trends is high enough for our own modeling results to hold (see below), it is interesting to note that it implies that comparison between methods or data are potentially fragile with respect to subtle variations in time.

We next updated the results from panels B, C and D shown in Figure 2 to that using the newer data (Supplementary Fig. S9). The change in reported results is extremely modest, with the Spearman correlation between performance and replicability across gene functions within a research community falling from $r_s = 0.333$ to $r_s = 0.328$. The correlation between performance and replicability for a given GO functions across research communities falls from $r_s = -0.125$ to $r_s = -0.135$. These modest changes leave it an open question as to whether the non-independence of data is varying with time. For an analyses of the effect of temporal variation in GO and its annotations, readers are referred to our previous work (Gillis and Pavlidis, 2013).

3.2 assessing and predicting replicability

3.2.1. Autism de novo variant network convergence exhibits artifacts

To the extent our model is accurate, we should predict that false results will be likely to replicate precisely because they are false. That is, if replication is dominated by artifactual overlaps, then results which are purely due to those artifacts will replicate very well across data. We turn to an interesting natural experiment to demonstrate this effect. An important result in the analysis of candidate psychiatric genetic variants is that the disease genes cluster within network data (Parikshak *et al.*, 2013). In general, this helps us to believe that we are finding some point of functional convergence defining the disease. Among the most influential of such reports is provided in the analysis of autism *de novo* variants in PPI data by O'Roak *et al.* (2012) (Fig. 3A). However, the data used in this case were problematic. The authors report '1.5 million physical interactions' which is far too many, even after halving this number (to make it unique interactions). In fact, due to an interpretation error, their interaction set includes many tested pairs which were not actually positive results. For example, data derived from a study of the human autophagy system (Behrends *et al.*, 2010) adds nearly 200 000 interactions, which is approximately its list of tested pairs, rather than the ~700 interactions actually reported as positive results in that study. The erroneous parsing specifically contributes 49 interactions to the excess (out of ~200) observed by O'Roak among their disease set. An ordinary response to this issue would be to look to replicate the result in other datasets where these problems do not exist and for that replication to validate the original finding. We hypothesized that the result would, in fact, replicate but that this would be because most PPI data has overlapping artifacts. Furthermore, we supposed that if we could then determine what these artifacts were and control for them, the replicated result would vanish in data. In other words, decreased variance among potential replicating data has destroyed its value as an indicator of truth.

3.2.2. Network results can replicate despite artifacts driving the original report

We find that the result replicates drawing on data from other PPI collections which do not ostensibly suffer from the described problem (Fig. 3B). While these resources draw on similar data, they exhibit substantial differences depending on curation practices and assessment, the very factors at issue for this particular analysis and therefore in need of replication; the new resources have only 33 661 in their intersect out of 200 499 pairs in their union. Because the results hold in all of this other data, the quality control issues in the original analysis are pure happenstance not affecting the result.

This might reflect the strength of the finding and that even substantial noise added to the datasets does not affect the result. Alternatively, it may support the view that the dominant signal across resources is heavily influenced by some bias, even where accurately collected and that replicability is no guarantor of correctness. In this case, selection bias is a natural candidate for the shared confound precisely because of the contamination of 'tested' data as positives within the original analysis. One point suggesting the disease clustering is less significant than raw P values indicate is that performing the same analysis across GO groups reveals that the disease set, while very significant, is far less clustered than normal 'functional' sets of genes (Fig. 3C).

3.2.3. Network results replicate because of artifacts

In this case, assessing the mechanism underlying replicability is straightforward because we know the results replicate even in data heavily influenced by selection bias. We re-analyzed all of the data with an alternate control, permuting through interactions rather than nodes to calculate the null (Maslov and Sneppen, 2002). This entirely randomizes the network connectivity rather than just the labelling but retains the same number of connections associated

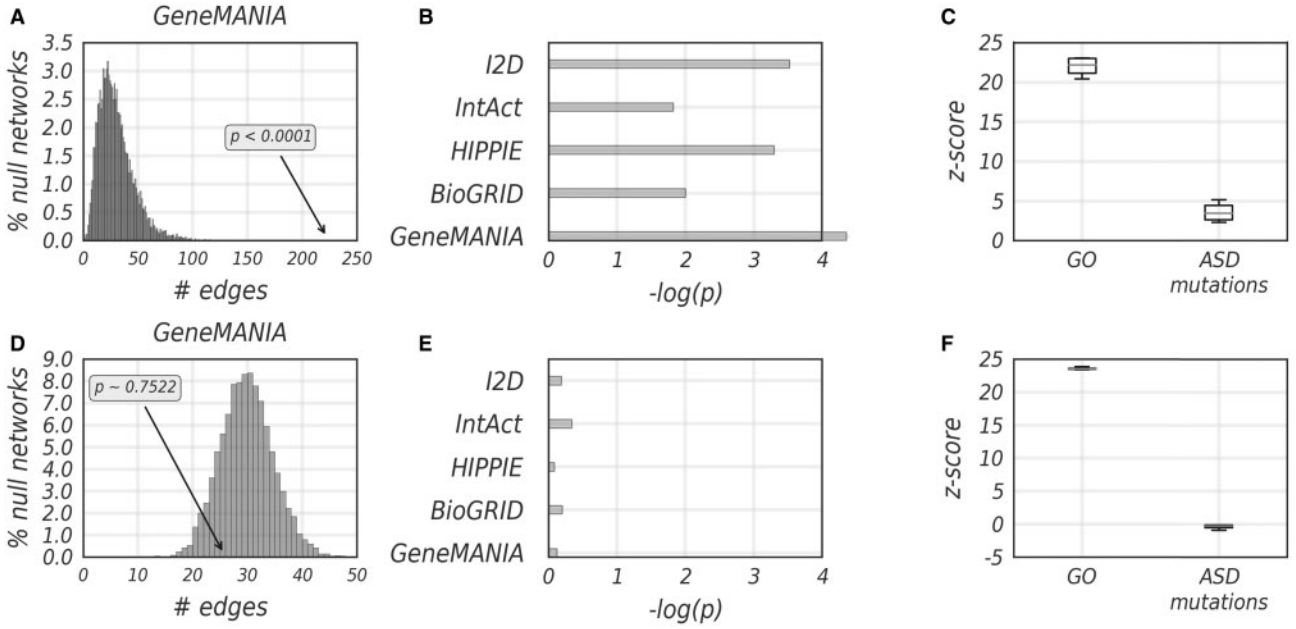

Fig. 3. Replicating network properties of de novo mutations in autism and edge permutation as a null removes all significance from the autism-derived gene list. (**A**) Replication of the Monte Carlo experiment performed in O'Roak *et al.* (2012) (see Supplementary Fig. S10). The number of edges between the genes in the autism (ASD) gene list in a network based upon GeneMANIA (version date August 3, 2011) physical interaction data is statistically different ($P < 0.0001$). The parsing of this network in the original analysis conflates 'tested' pairs with 'validated' pairs of interacting genes across much of the data. (**B**) Networks drawing on different network resources also show a statistically significant number of edges between the genes in the ASD gene list: BioGRID ($P \sim 0.007$), HIPPIE ($P \sim 0.0006$), IntAct ($P \sim 0.015$) and I2D ($P \sim 0.0004$). Because we ran 10 000 iterations to calculate P values, the maximum value on the graph is 4; the original GeneMANIA-based result is some point past this, indicated by its placement. (**C**) However, conducting the same analysis on gene lists derived from GO terms reveals that they are much more likely to exhibit significant linkage than the ASD gene list (GO: mean z-score = 21.96; ASD gene list = 4.41). (**D**) Using the corrected GeneMANIA data as well as holding node degree per each gene fixed in the null simulations removes all significant association between ASD genes. Note the null distribution of number of edges differs between this and (A); axis scale has also changed. (**E**) This property replicates across network data derived from multiple sources. (**F**) Functional sets of genes defined by GO remain learnable even after accounting for node degree in this way

with each gene, a proxy for selection bias predictive of functional properties (Gillis and Pavlidis, 2011). Aside from controlling for selection bias, significance should generally be easier to attain then the node permutation case since the null now has no structure. In this analysis, the significance vanishes from all of the real data (Fig. 3D and E), including the updated and corrected version of the original PPI data. Crucially, functional sets of genes as defined by GO retain their significant connectivity (Fig. 3F). Altogether, this indicates that replicability in the disease gene analysis indicated replicability of bias, a factor which attaches no more to this study than any other, except for some irrelevant bad luck in choosing data for which it would be difficult to obtain meaningful results. We perform a qualitatively identical analysis in another case (Gulsuner *et al.*, 2013) with similar results in the Supplementary Material (see Supplementary Figs S10–S14). While these cases involve methods accidentally exploiting selection bias, the problem they identify is essentially orthogonal. The meta-analytic confound we have identified is likely to be dominated by other biases in other data modalities.

3.2.4. Replicability indicates poor data quality in PPIs

The evidence of high bias in the underlying PPI data suggested to us that replicability within the networks themselves might be dominated by artifacts. We also see some evidence for this within the model analysis, where research communities dominated by PPI data had replicability to performance correlations significantly lower than those seen in the co-expression data, which is less prone to selection bias since genome wide (Spearman's $r_s \sim 0.31$ versus 0.53). To assess these upstream effects in the underlying data, we focus on among the most commonly used network resource, BioGRID (Stark *et al.*, 2006). Awareness of the potential for confounds in aggregated PPI data has made some degree of quality control in using BioGRID commonplace. The most common approach is to threshold for replicability by requiring interactions to have been reported multiple times (Anastassiadis *et al.*, 2011). That is, replicability is the method by which correction for bias is generally attempted.

The individual reports determining this replicability should have quite strong variation in their degree of dependence (in our terms) since practices underlying data collection can vary enormously across what is, in essence, almost the entire field of proteomics. We would therefore hypothesize that there should be little value to replication. A recent comprehensive analysis of the quality of AC-MS data (Mellacheruvu *et al.*, 2013) allows us to evaluate this quantitatively, by determining whether replicable interactions are more likely to involve proteins for which results cannot be considered reliable. We find a strikingly strong relationship between the degree of replicability and the mean 'unreliability' score of the interactions (Supplementary Fig. S15, Spearman's $r_s \sim 0.99$, $P \sim 9.24$ E-6), suggesting that replicability has negative value in PPI data. It is not just that the PPI data is noisy but that how we most easily fix such problems is now incorrect.

4 Discussion

Concerns about replicability in science have been much discussed recently (Begley and Ellis, 2012; Ioannidis, 2005). It has not been clear whether this has emerged as a recent focus because our systems are

becoming more complex, our practices less precise or whether our evaluation of problems is simply clearer. Our analysis of gene network methods demonstrates an alternate possibility: that as methods and data are optimized to improve replicability, the independent value of replicability diminishes. Goodhart's law, that '[a]ny observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes' (Goodhart, 1975), is true even for scientific replication.

While dependencies underlying replicability have been touched on previously both in the context of human judgment and machine learning algorithms, this has principally been seen as a major factor to exploit in improving performance (Breiman, 1996; Mellers *et al.*, 2014). Our demonstration of a potentially negative value for replicability may sound incompatible with these findings or past scientific practice in general, but we suggest that this is not really the case. Normally, assessment of the value of a result depends on an understanding of the factors underlying it. It is only where we target replicability and are indifferent to how it is achieved that our report raises a central concern. Unfortunately, the increased focus and attention on overlapping reference data and consensus evaluation makes field-wide overfitting ('bad' replicability) a real possibility for gene network methods.

The problem of overfitting is central to machine learning, and approaches for avoiding it even when comparing methods, have been addressed within that literature (Demsar *et al.*, 2006). These mostly resemble comparisons of the sort we cover in sections 3.1.1 and so also exploit replicability. One standard approach, is 5 x 2 cross-validation (Dietterich, 1998), where 2-fold cross-validation is repeated multiple times to ensure comparative performances are robust/replicable. It might be tempting for us to target the problem of data orthogonality directly within this framework, but this merely raises Goodhart's law anew. Specifically targeting orthogonal data will make us sensitive to overfitting on, however, we characterize 'orthogonality'. However, heuristics for feature selection do already incorporate orthogonality as a desirable property (Hall, 2000), and it is natural to wonder whether refinement of this analysis or filtering at the data collection stage would minimize this problem. Our data collection was intended to be reflective of general use (e.g. Mousefunc included both Pfam and InterPro as separate resources) rather than whatever we might consider ideal and we certainly leave open the possibility that better assessment of data orthogonality could make the field insensitive to the problems we have identified, remembering that for any given field of research modeled, the relationship between replicability and accuracy was a positive one.

Indeed, it is important to recognize that replicability and meaningful comparative evaluation really are desirable properties. Just as high performance in an algorithm is a good property but not one we should enforce by fiat, the same can be said of replicability. We particularly note that the effect we describe is a problem with the way we attempt to fix problems in scientific practice or data and not just a scientific problem in itself. For example, it is typical to threshold by replicability in PPI data to avoid methodological or data issues. Likewise, consensus resources, data sets, methodological practices, animal models, etc., are often specified precisely to allow researchers to better obtain replicable results. While it may seem intuitive that replication should be a test of robustness and that it will lose value where this is not true, our observation is that most explicit focus on replication strongly diminishes its utility through the use of very tightly constrained systems, data and methods.

Our suggestion is that these difficulties arise particularly in genomics because our real problems are often poorly defined. Predicting 'gene function' is hard and so we swap in the better defined problem of predicting GO or otherwise alter evaluation to make results 'sensible'; however, closing this feedback loop so directly removes independence between method and assessment. We suggest this meta-analytic difficulty can be solved by targeting a real biological problem with diverse and well-powered data. This will differ from field to field and should be regarded as an important research effort in itself. Within transcriptomics, predicting the sex of the organism from which all public data were collected for some recent interval trained on the past would be a worthwhile and achievable task, before moving on to tissue, cell-type, etc. The critical point is to pick a problem in which the question is perfectly defined and the answer is perfectly knowable. In the meantime, we do not recommend deviating from the current dry-lab cross-validation practice on standardized data and instead advocate in favor of control experiments which reveal what factors affect performance, as in our examples.

Because the model we have constructed is quite general, we might expect to see this effect outside science and we suggest that this is, in fact, the case. The heuristic our model suggests is that high concurrence is a reason to disbelieve a claim if the methods whereby that concurrence arose are unknown. A careful reading of the substantial literature in the social sciences on persuasiveness suggest that this effect, while not previously recognized, may be responsible for some otherwise puzzling results. For example, people in diverse environments are more interested in opinion information (as opposed to factual) (Scheufele, 2014). This is explained as their preparing for argument, but our analysis suggests that people may simply be drawing the rational inference that opinion information is more valuable where it is diverse (and where concurrence will imply correctness). Similarly, experts become more convincing when their views are *more* divergent from pre-existing beliefs (Pornpitakpan, 2004), which may be puzzling in a naively Bayesian sense but is intuitive if the divergence of belief within the populace as a whole is being estimated and used to weight the value of opinion. We call this effect the 'talking points' heuristic, since accusations that concurrence must be artifactual simply because it is high are sometimes described in this way. Our findings also have clear implications for public science funding, which has increasingly focused on generating reference data as a matter of deliberate policy, sometimes specifically to target replicability.

In this article, we have emphasized some subtleties around evaluating meta-analytical properties of results. Results independently derived from different data resources do not replicate 'easily' and so are more meaningful where it occurs; similarly, they profit the most from aggregation or comparison. However, folding such aggregation directly into methodological construction makes replicable results easier to achieve and less meaningful. In these cases, we must seek more orthogonal validation or more carefully calibrated control experiments. While our work is the most comprehensive quantification of this problem, these ideas have already found purchase within machine learning based on more specific analyses: the perils of overfitting are often discussed. Our perception is that replicability within genomics, and gene network analysis particularly, is usually seen as somehow more fundamental than these concerns and so a universal good to be strived for. As our analysis quantifies, this is far from true and becomes ever less so the more replicability is enforced from the top down. Replicability as a form of validation is a finite resource even if data generation is not and more thoughtful stewardship by scientific organizers (of all types) is necessary.

Acknowledgements

We thank Paul Pavlidis and Shane McCarthy for helpful comments on a draft of the manuscript. We thank Quaid Morris for the GeneMANIA code.

Funding

JG, WV, and SB were supported by a grant from T. and V. Stanley.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Anastassiadis,T. *et al.* (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1039–1045.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Ballouz,S. *et al.* (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, **31**, 2123–2130.

Begley,C.G. and Ellis,L.M. (2012) Drug development: raise standards for preclinical cancer research. *Nature*, **483**, 531–533.

Behrends,C. *et al.* (2010) Network organization of the human autophagy system. *Nature*, **466**, 68–76.

Bickel,P.J. *et al.*, (1975) Sex bias in graduate admissions: data from Berkeley. *Science*, **187**, 398–404.

Bornigen,D. *et al.* (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics*, **28**, 3081–3088.

Breiman,L. (1996) Bagging predictors. *J. Mach. Learn. Res.*, **24**, 123–140.

Brown,K.V. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, **R95**.

Chatr-aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.

Demsar, J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

Dietterich,T.G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, **10**, 1895–1923.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Fisher,R.A. (1935) *The Design of Experiments*. Oliver and Boyde, Edinburgh, London.

Geschwind,D.H. (2008) Autism: many genes, common pathways? *Cell*, **135**, 391–395.

Gillis,J. and Pavlidis,P. (2011) The impact of multifunctional genes on "guilt by association" analysis. *PLoS One*, **6**, e17258.

Gillis,J. and Pavlidis,P. (2013) Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics*, **29**, 476–482.

Goodhart,C.A.E. (1975) *Problems of Monetary Management: The UK Experience*. Reserve Bank of Australia, Papers in Monetary Economics.

Greene,C.S. and Troyanskaya,O.G. (2012) Accurate evaluation and analysis of functional genomics data and methods. *Ann. N. Y. Acad. Sci.*, **1260**, 95–100.

Gulsuner,S. *et al.* (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, **154**, 518–529.

Hall, M.A. (2000) Correlation-based feature selection for discrete and numeric class machine learning. In: Langley,P. (ed), *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 359–366.

Hunter,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Ioannidis,J.P. (2005) Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, **294**, 218–228.

Joshi-Tope,G. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.

Kryshtafovych,A. *et al.* (2014) CASP10 results compared to those of previous CASP experiments. *Proteins*, **82** (**suppl.**), 164–174.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.

Mellacheruvu,D. *et al.* (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods*, **10**, 730–736.

Mellers,B. *et al.* (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci.*, **25**, 1106–1115.

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

Moreau,Y. and Tranchevent,L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.

O'Roak,B.J. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246–250.

Oellrich,A. *et al.* (2012) Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PLoS One*, **7**, e38937.

Ogata,H. *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.

Oliver,S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.

Orchard,S. *et al.* (2014) The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.

Parikshak,N.N. *et al.* (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, **155**, 1008–1021.

Pavlidis,P. and Gillis,J. (2013) Progress and challenges in the computational prediction of gene function using networks: 2012-2013 update. *F1000Res.*, **2**, 230.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Pena-Castillo,L. *et al.* (2008) A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.*, **9** (**suppl.**), S2.

Pornpitakpan,C. (2004) The persuasiveness of source credibility: a critical review of five decades' evidence. *J. Appl. Soc. Psychol.*, **34**, 243–281.

Portales-Casamar,E. *et al.* (2013) Neurocarta: aggregating and sharing disease-gene relations for the neurosciences. *BMC Genomics*, **14**, 129.

Radivojac,P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.

Schaefer,M.H. *et al.* (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.

Scheufele,D.A. (2014) Science communication as political communication. *Proc. Natl. Acad. Sci. USA*, **111** (**suppl.**), 13585–13592.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.

Verleyen,W. *et al.* (2015) Measuring the wisdom of the crowds in network-based gene function inference. *Bioinformatics*, **31**, 745–752.

Wang,P.I. and Marcotte,E.M. (2010) It's the machine that matters: predicting gene function and phenotype from protein networks. *J. Proteomics*, **73**, 2277–2289.

White,J.A. *et al.* (1997) Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics*, **45**, 468–471.

Zuberi,K. *et al.* (2013) GeneMANIA prediction server 2013 update. *Nucleic Acids Res.*, **41**, W115–W122.