Supplementary Information to:

GeNets: A unified web platform for network-based analyses of genomic data

Taibo Li\textsuperscript{1,2,3,4,13}, April Kim\textsuperscript{1,2,13}, Joseph Rosenbluh\textsuperscript{2,5}, Heiko Horn\textsuperscript{1,2}, Liraz Greenfeld\textsuperscript{1,2}, David An\textsuperscript{2}, Andrew Zimmer\textsuperscript{2}, Arthur Liberzon\textsuperscript{2}, Jon Bistline\textsuperscript{2}, Ted Natoli\textsuperscript{2}, Yang Li\textsuperscript{2,\textit{6,7}}, Aviad Tsherniak\textsuperscript{2}, Rajiv Narayan\textsuperscript{2}, Aravind Subramanian\textsuperscript{2}, Ted Liefield\textsuperscript{2}, Bang Wong\textsuperscript{2}, Dawn Thompson\textsuperscript{2}, Sarah Calvo\textsuperscript{2,6}, Steve Carr\textsuperscript{2}, Jesse Boehm\textsuperscript{2}, Jake Jaffe\textsuperscript{2}, Jill Mesirov\textsuperscript{2,8}, Nir Hacohen\textsuperscript{2,9,10}, Aviv Regev\textsuperscript{2,11}, Kasper Lage\textsuperscript{1,2,10,12}

Affiliations:
1. Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA
2. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
3. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
4. Johns Hopkins School of Medicine, Baltimore, MD, USA
5. Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA.
6. Howard Hughes Medical Institute and Department of Molecular Biology, MGH, Boston, Massachusetts, USA
7. Department of Statistics, Harvard University, Cambridge, Massachusetts, USA
8. University of California, San Diego, California, USA.
9. Center for Immunology and Inflammatory Diseases, Massachusetts General Hospital, Boston, Massachusetts, USA
10. Harvard Medical School, Boston, Massachusetts, USA.
11. Howard Hughes Medical Institute, Department of Biology, MIT, Cambridge, Massachusetts, USA.
12. Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark
13. These authors contributed equally

Correspondence should be addressed to Kasper Lage (lage.kasper@mgh.harvard.edu)

Table of contents:

- Tables S1 & S2 ............................................. Page 2-3
- Supplementary Note 1 ....................................... Page 4
- Supplementary Note 2 ....................................... Page 5-7
- Supplementary Note 3 ....................................... Page 8
- Supplementary Note 4 ....................................... Page 9
- Supplementary Note 5 ....................................... Page 10
- Supplementary Note 6 ....................................... Page 11
- Supplementary Note 7 ....................................... Page 12-13
- Supplementary Note 8 ....................................... Page 14
- Supp. References ........................................... Page 15-16
| Network     | # of nodes | # of edges |
|------------|------------|------------|
| InWeb      | 12,357     | 428,429    |
| AchillesNet| 9,219      | 500,000    |
| LINCSNet   | 6,721      | 500,000    |
| GEONet     | 12,390     | 500,000    |
| CLIMENet   | 8,279      | 500,000    |

Supplementary Table 1| Final Network Sizes.
1. Eigenvector Centrality in the Pathway
2. Weighted Degree in the Network
3. Weighted Degree in the Pathway
4. Ratio of the Weighted Degree in the Pathway to that of the Network
5. Ratio of the Pathway Degree to that of the Network
6. Eigenvector Centrality in the Network
7. Closeness Centrality in the Pathway
8. Ratio of the Pathway Eigenvector Centrality to that of the Network
9. Ratio of the Pathway Closeness Centrality to that of the Network
10. Local Clustering Coefficient in the Network
11. Degree in the Network
12. Betweenness Centrality in the Network
13. Closeness Centrality in the Network
14. Betweenness Centrality in the Pathway
15. Degree in the Pathway
16. Local Clustering Coefficient in the Pathway
17. Ratio of the Pathway Local Clustering Coefficient to that of the Network
18. Ratio of the Pathway Betweenness Centrality to that of the Network

Supplementary Table 2 | Full list of metrics used by Quack.
Supplementary Note 1 | Introduction to biological networks and motivation for creating the GeNets technology. Following the technological breakthroughs of the human genome project (1), large-scale methods to map genomic, transcriptomic, and proteomic data have become ubiquitous in natural and biomedical sciences. Integrating the resulting large data sets into functional genomics networks (where nodes represent genes and edges represent some functional association between gene pairs) is a powerful and convenient way of representing the data to decipher complex biological relationships that would not emerge without a holistic view of the functional associations at genome scale (reviewed in 2,3). This is in part because these networks describe biologically relevant gene-gene relationships that have not been elucidated by smaller scale experiments. Remarkably, diverse functional genomics networks share similar global design by being scale-free (meaning that most genes have a small number of connections and a few genes have a large number of connections (3, 4, 5)), small world (meaning that shortest path length between any gene pair in the networks is relatively small (6)), and modular (meaning that there are sub-networks of gene sets that connect more significantly to each other than to the rest of the network (2,7,8)). Analyses have further shown that these global rules reflect fundamental principles of molecular biology as most genes in the same subnetwork work in concert to execute specific cellular functions - i.e., as molecular machines, signaling circuits, enzymatic cascades or rigid topological structures.

Understanding the global network-based location and properties of genes have important implications for biology and phenotype-genotype relationships. For example, genome-wide association studies or exome-sequencing studies both in somatic cancers, Mendelian diseases, and common complex disorders have revealed that genes in close proximity of each other (and often in the same sub-network or module) in these networks often are affected by genetic variation linked to the same disease. These results illustrate that functional genomics networks can serve as a model to interpret large genomic data sets and suggest targeted and cost-efficient follow up experiments (9, 10, 11, 12, 13, 14 and reviewed in 15). Other analyses of global topologies have indicated that mutated genes central in the networks are more often embryonic lethal and have deleterious effects on gene function, and that genes involved in diseases are often located more in the periphery of functional genomics networks because mutations in these genes are tolerated from an embryonic developmental perspective, but can affect human health later in life (16).

However, as many complementary functional genomics network become available, and as they grow in size, their inherent complexity (both in terms of nodes and edges) makes it challenging to understand and quantitatively map how biological signal is organized within and between different types of networks. It is also a significant technological challenge to provide standardized and computationally efficient statistical methods with the needed flexibility to quantitatively compare the biological signal of different networks and to fine map differences in how pathways are organized within them to assess their individual strengths and weaknesses when addressing a particular biological problem. It is also challenging to design a standardized and unified way to integrate many different types of networks with large genomic datasets that are now being produced as part of the ongoing genomic revolution. Finally, it is important that a resource towards these aims are scalable while providing the opportunity to visualize, manage, save and share analyses with collaborators. GeNets is a unified web platform for analyzing, quantitatively comparing, interacting with and, sharing analyses from large heterogeneous functional genomics networks and for integrating these analyses with genetic data with the ultimate aim of informing cost-efficient and targeted follow up experiments.
Supplementary Note 2 | Networks used in this work. Briefly, we used i) Protein-protein interaction data from the literature with associated credibility extracted from the InWeb database (9); ii) Gene expression correlation data derived from the Gene Expression Omnibus (17) (details below); iii) Similar global transcriptional responses (due to perturbations) from the LINCS Connectivity Map (18, 19, 20) (details below); iv) Cancer codependency correlations from 216 cancer lines (21) (details below). v) Similarity in phylogenetic patterns based on the method ‘clustering by inferred models of evolution’ (22) (details below).

Details on Creating GEONet Gene Expression Network. A co-expression network was derived from the Gene Expression Omnibus (17). This expression matrix was derived using several metrics reported from Affymetrix arrays. Thresholding based on the distributions of the metrics was conducted such that extreme outliers would be excluded: First, the ratio of signal coming from the 3’ versus the 5’ end of the transcript of the gene beta-actin, a standard control gene. The reason for measuring this ratio is that transcription runs from the 3’ to the 5’ end of a gene. If mRNA samples are of good quality, one should observe equal amounts of signal from the 3’ and 5’ ends. However, if the 5’ end is underrepresented, this suggests that the mRNA was degraded or wasn't labeled completely. A filter of ratio <= 2 was used based on the empirical distribution. Second, the same filter was applied to the gene GAPDH with a rule of ratio <= 2 based on the empirical distribution. Third, the average signal intensity across all genes from a given sample with a filter for samples with value >= 150 based on the empirical distribution. Fourth, the fraction of genes that the Affymatrix platform has called as being reliably detected, or "present", was used to filter for samples with value >= 30. Finally, a power law regression is fit to 80 "invariant" genes in each sample. A goodness of fit for this regression was estimated and filtering was conducted for samples whose goodness of fit was at least 4 based on the empirical distribution. This process concluded in a matrix of 22,268 probes by 19,019 samples. Probes were collapsed to HUGO gene symbols by averaging and the resulting gene-by-gene correlation matrix was of size 12,716 by 12,716. Finally, we applied both global silencing (23) and network deconvolution (24) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) as outlined in Supplementary Note 3 which resulted in the selection of the deconvoluted matrix thresholded to the strongest positive 500K gene expression correlations.

Details on Creating the LINCSNet Cell perturbation Profile Network. Here, we utilize L1000, a high-throughput, bead-based gene expression assay in which mRNA is extracted from cultured human cells treated with various chemical or genomic perturbagens (small molecules, gene knockdowns, or gene over-expression constructs) as previously described (19). This mRNA is reverse transcribed into first-strand cDNA. Gene specific probes containing barcodes and universal primer sites are annealed to the first strand cDNA. The probes are ligated to form a template for PCR. The template is PCR amplified with biotinylated-universal primers. The end products are biotinylated, fixed length, barcoded amplicons. The amplicons can then be mixed with Luminex beads that contain complementary barcodes to those encoded in each of the 1000 amplified landmark genes. These 1000 landmark genes were chosen as a reduced representation of the transcriptome and account for the majority of expression variation across many cellular contexts (Subramanian, et al., manuscript in preparation). These beads are then stained with fluorescent streptavidin-phycoerythrin (SAPE) and detected in 384-well plate format on a Luminex FlexMap flow cytometry-based scanner. The resulting readout is a measure of mean fluorescent intensity (MFI) for each landmark gene. The raw expression data are log2-scaled, quantile normalized, and z-
scored, such that a differential expression value is achieved for each gene in each well. In the standard L1000 protocol, each well corresponds to a different perturbagen and these differential expression values are collapsed across replicate wells to yield a differential expression signature for each perturbagen. The signatures of different perturbagens can then be compared to identify those that result in similar or dissimilar transcriptional responses as previously described (18, 20, 25). We computed the similarity between all pairwise combinations of the roughly 460,000 signatures in the CMap database. We then performed a summary of these query results to arrive at a more perturbagen-centric view of connectivity. To summarize, the query result is first grouped by cell line and perturbagen type (small molecule, gene knockdown, or overexpression). The connectivity scores are then normalized by dividing by the signed mean score of each group. The scores are converted to percentile ranks within each group. The perturbagens are then ranked according to the direction of connectivity. Positive connections are ranked highest, and negative connections ranked lowest. For each unique perturbagen, we considered the average percentile rank in the four cell lines for which the connection to the query was strongest. The matrix was then transformed into a symmetric matrix by averaging the (i,j) and (j,i) values. Next, we applied both global silencing (23) and network deconvolution (24) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) as outlined in Supplementary Note 3 which resulted in the selection of the original matrix (i.e., non-silenced) thresholded at the strongest 500K edges.

Details on Creating AchillesNet Cancer Codependency Network. A codependency network was derived from the Project Achilles dataset v2.4.3 (21). This dataset is from RNAi screens of 216 cancer cell lines, each one infected with a pool of over 54,000 shRNAs each designed to knockdown one gene, for a total of more than 17,000 genes. These data were processed by the ATARiS method to yield dependency scores representing the degree of dependence of each cell line on each gene. Pearson correlation coefficients were computed from the dependency profiles of all pairs of targeted genes, resulting in a 17Kx17K correlation matrix. Finally, we applied both global silencing (23) and network deconvolution (24) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) as outlined in Supplementary Note 4 which resulted in the selection of the deconvoluted matrix thresholded to the strongest 500K co-dependencies.

Details on creating CLIMENet Gene Coevolution Network. CLIMENet was created using the algorithm CLIME (clustering by inferred models of evolution) (22) applied to 1,025 curated human gene sets from GO and KEGG (downloaded from http://www.gene-clime.org). CLIME predicts functionally related genes based on three inputs: a functionally-related gene set G, a species tree S, and a phylogenetic matrix containing presence/absence vectors of all genes in a reference genome across all species in S. Briefly, CLIME first partitions G into disjoint evolutionarily conserved modules (ECMs), and then scores all other genes in the genome a log-likelihood ratio (LLR) score to quantify the possibility that each gene has arisen under the ECM’s inferred model of evolution compared to a background null model. The “expanded ECM” or ECM+ is the set of all other genes (not in G) with LLR>0. ECM+ genes share a similar evolutionary history with a subset of G and thus may be functionally related. To create CLIMENet, CLIME was applied to all 909 GO cellular components and 116 KEGG pathways (using a human centric phylogenetic matrix and a 138-species eukaryotic tree, see www.gene-clime.org) – resulting in a total of 13,307 ECMs. Of these 13,307 ECMs, 10,606 are singletons (containing only one gene within G, indicating no shared evolutionary history with other pathway members). The resulting
13,307 ECM+ expansion sets contain a total of 667,592 genes, many redundant. In CLIMENet, genes A and B are connected by an edge if A and B occur together in at least one ECM+, and the edge weight is assigned the mean LLR scores for A and B across all ECM+s that contain both genes (LLR>0).

For example, the GO “voltage-gated potassium channel complex” contains 70 genes, which CLIME partitions into 28 ECMs (including 10 singletons). ECM1 contains 9 known potassium channel genes and the ECM1+ expansion contains 45 predictions with LLR>0 (most sharing the pfam domain “IRK” – which causes these genes to have a similar phylogenetic profile across 138 eukaryotic species). All gene pairs in ECM1+ (45*44=1,980 gene pairs) will be connected in CLIMENet due to the shared phylogenetic history (in this case due to the IRK protein domain). Note that the original 9 potassium channel genes in ECM1 are not used in the creation of CLIMENet – only the ECM+ genes. This avoids any circularity in assessing CLIMENet performance on MSigDB C2 (which has substantial overlap with the 1025 curated sets used to create CLIMENet).

Finally, we applied both global silencing (23) and network deconvolution (24) to this matrix and compared performance of the original matrix and silenced solutions at various network sizes (i.e., edge thresholds) Supplementary Note 3 which resulted in the selection of the original matrix (i.e., non-silenced) thresholded at the strongest 500K edges.
Supplementary Note 3 | The effect of silencing indirect links in the networks. Matrices of functional genomics data contain both direct and indirect associations between gene pairs. To remove indirect effects and focus on direct functional relationships between pairs of genes, we applied both global silencing (23) and network deconvolution (24).

For each dense matrix (all functional genomics data previously listed with the exception of protein-protein interactions), we ranked the connections in descending order by the original connections weights, the deconvoluted weights, and the globally silenced weights. We then filtered to the top 500K, 750K, 1M, 1.25M, and 1.5M connections. To assess what size networks to use in the next step of pathway modeling, we conducted both conventional permutation testing on 853 canonical gene sets from MSigDB, and assessed the classification power (AUC) of our classifier for each of networks and sizes. For the permutation approach, for each gene set, we compute the global clustering coefficient (with disconnected genes set to zero) and then sampled 500 gene sets with similar overall degree distributions from the respective network. The empirical p-value for each pathway is determined by assessing the pathways clustering coefficient to that of null distribution, i.e. the % of randomly sampled gene sets with global clustering coefficient greater than or equal to the pathway under consideration. We then plot the % of pathways among the 1,300 that are significantly connected (using a significance level of 0.10) and assess the behavior of this metric as we relax the threshold for the top connections (i.e., increase the size of the network); these results are illustrated in Supplementary Figure 8. Next, we assessed the behavior of the AUC for our classifier as we relaxed the network size for each of these networks (Supplementary Figure 9) and observed that the AUC’s were consistently stable across network at 1M edges. Finally, we incorporated the learnings from both methods and selected the top 500K edges from the deconvoluted cancer codependency network, original phylogenetic pattern network, the original cell perturbation network, and the deconvoluted gene expression network.
Supplementary Note 4 | Details of the Quack algorithm.

We formulated a Random Forest classifier to assess whether the topological features observed in Supplementary Figure 7 could be leveraged for two specific purposes. First, we wanted to test if it is possible to use the topological patterns of pathways to compare the global and local biological signal of user-defined networks. Second, we wanted to test if we could predict new candidates in any network based on a seed set of genes using this approach (see Main Text paragraphs, 2, 3 and 4 for more details about the aims of the Quack algorithm).

Specifically, using the InWeb network and 853 pathways (Supplementary Data 1), we repeat the following calculations: First, we compute all 18 topological properties for each gene in the pathway in question (Supplementary Figure 10a) and add these observations to the modeling dataset, where the gene is indexed as a row and the eighteen metrics as columns in that row. Second, we determine the context gene set for the pathway in question. For each gene in the context set, we compute the 18 topological properties under the assumption that the gene is a member of the pathway (Supplementary Figure 10a) and add a subset of these observations as one row per gene in the modeling dataset to ensure a more balanced representation (26) of pathway members to context genes. The pathway member and context gene observations from all 853 pathways are then combined to obtain a modeling dataset of 752,172 rows with each eighteen columns (Supplementary Figure 10a). We randomly sample 70% of the pathways as the training set and use the remaining 30% of the pathways for validation.

The Random Forest classifier was constructed using the training data to build an ensemble of 500 smaller classifiers (i.e., the forest). In each of these, a subset of the topological metrics is used to construct a tree that maximizes the segmentation of pathway members from context genes (Supplementary Figure 10b). To assign a probability that a gene in the validation dataset belongs to a specific pathway, the gene’s eighteen topological metrics (in relation to that specific pathway) are used as input to the forest, whereby each of the 500 trees each cast a vote. The probability that a gene belongs to the pathway in question is the proportion of trees that votes the gene as a pathway member (Supplementary Figure 10b). Because the method predicts whether a gene is part of a pathway based on the way it ‘walks and talks’ in terms of network topology we named the classifier Quack. When using this approach to differentiate between pathway and context genes in the validation data, the area under the receiver operating characteristics curves (AUCs) range from 0.59 to 0.93 (Main Text Figure 2B).

Finally, to ensure the probabilities assigned to genes by Quack have a straightforward and intuitive biological interpretation, we show that there is a direct correlation between Quack probabilities and true-positive rates when predicting pathway members in held-out pathways (Supplementary Figure 11). Overall, our results show that, despite fundamentally different topological signatures across networks, our approach effectively learns how to use network-specific topological properties of pathways to identify functional biological relationships.
Supplementary Note 5 | Differential pathway topologies in heterogeneous network data.

To quantify the relative importance of the various topological metrics in defining pathway relationships, we permute the values of each topological metric being evaluated within each tree and compare the average classification error across trees before and after permutation. This provides an estimate of the overall importance for each topological metric when determining pathway relationships in the network under investigation.

We applied this analysis to quantify the extent to which the diverging signatures (illustrated in Supplementary Figures 2-7) are mirrored in varying importance of topological metrics when Quack is applied to different networks and found that although there are metrics for which the predictive power is generally high or low, there is considerable variation across networks (Supplementary Figure 13). For example, the eigenvector centrality within a pathway is a consistently top-ranked metric across networks. Interestingly, most of the ratios derived from the respective pathway-to-network metrics were also ranked highly (four of six metrics based on ratios are ranked in the top 10). For example, the eighth most important metric (based on the average rank) is the ratio of the degree in the pathway of a gene to that of its degree in the overall network. In contrast, the degree in the pathway evaluated alone is unimportant across all networks (ranked 18th, i.e., the least important metric overall). This proves the intuition that - in any network - the number of connections a gene has within a given gene set (which is sometimes taken as naive support for a gene's role in a biological process of interest) can be highly misleading if not evaluated on the background of its total number of connections in the network in question.

Conversely, there is significant variation in the predictive power of the local clustering coefficient in the network, degree in the network, and the betweenness and eigenvector centrality in the network. Where betweenness centrality in network was the most important metric for gene networks based on co-expression across gene expression omnibus (GEONet) and protein-protein interactions (InWeb), it is one of the least important for those based on cancer synthetic lethalities (AchillesNet), phylogenetic relationships (CLIMENet), and cell perturbation profiles (LINCSNet). Similarly, the local clustering coefficient in network is the most important metric for CLIMENet, and the fourth most important for InWeb, while being relatively unimportant in AchillesNet, LINCSNet, and GEONet.

To illustrate general differences in the organization of pathways across networks in the context of specific biology we visualized the eigenvector centralities (as it is consistently ranked high in Supplementary Figure 13) of genes in the context of PDGF, ERBB1, E2F pathways. Despite the general importance of this metric there is considerable network-specific divergence of its patterns and strengths. Nevertheless there is considerable network-specific divergence the patterns of strengths of this metric (Supplementary Figure 14). This visualization illustrates that the topological properties of pathways are fundamentally different across networks, but that Quack can effectively learn those properties and use them to identify biological signal across all network types.
Supplementary Note 6 | Discussions of potential biases and weaknesses of Quack. Through the systematic investigation of 18 topological properties across 853 pathways in gene networks of correlated mRNA expression, phylogenetic patterns, cancer codependency relationships, cell perturbation profiles, and protein-protein interactions, we have quantified and compared the network-specific organization of pathway information across heterogeneous networks that are widely used in the biomedical community.

We have shown that the organization of pathway signal is network-specific and harness this fundamental principle to formulate a machine learning approach, Quack, that can accurately assign probabilities that genes, in any gene set, form a pathway in any network. Given a set of genes, Quack can also predict likely new pathway genes amongst context genes. The complexity of biological networks and the pathway relationships they encode is illustrated by the fact that the median size of pathways is 48 genes, while the contexts of most pathways are several orders of magnitude higher [i.e., a median of 3,400 context genes and maximum greater than 11K context genes]. This inherent complexity of biology underscores the importance of developing scalable and efficient machine learning strategies for elucidating functional insights from genome-scale networks.

It could be claimed that Quack is biased by knowledge contamination (i.e., the idea that genes that are part of the same pathway are more studied in relation to each other and therefore have more connections in a network) when applied to the InWeb protein-protein interaction network. However, the four other networks used in this analysis (CLIMENet, AchillesNet, GEONet and LINCSNet) are built from systematically exploring gene expression datasets or the alignment of genomes and are therefore completely independent of this form of knowledge contamination. Therefore, it is all the more noteworthy that the biological principles we observe for InWeb are mirrored in analogous observations in CLIMENet, AchillesNet, GEONet and LINCSNet. Although the performance of Quack on the InWeb protein-protein interaction network is higher than most other networks, the approach is able to accurately predict pathway relationships in all five networks making it unlikely that potential study bias of proteins in InWeb has any major impact on our results or observations.

Correlation-based networks such as CLIMENet, AchillesNet, GEONet, and LINCSNet can be influenced by indirect relationships (i.e., the property that if node pairs A-B and B-C are strongly correlated in any of these networks independently of each other, this may create a link between A-C even if there is no direct biological information flow between A and C). We applied two algorithms that eliminate indirect effects to all of the networks and repeated our analyses (Supplementary Note 3 and Supplementary Figures 8 and 9). This analysis shows that indirect effects do not significantly influence our results and that Quack performs similarly with or without these preprocessing steps. Give that it can be computationally very demanding to eliminate indirect effects from large networks it is desirable from a technology standpoint that this pre-processing step is not necessary for the analytical framework we have developed. Our analysis of pathway topologies is based on a relatively comprehensive set of metrics that describe genes topological relationships in a pathway and in an entire network. This approach can also be extended to include information about the other types of information such as whether the pathway is metabolic, regulatory, or involved in signaling. We have shown that such a pathway stratification step can improve biological discovery. We further show that there is considerable divergence in how pathways are topologically oriented in different network types with the greatest divergence observed between physical interactions (InWeb) and phylogenetic patterns (CLIMENet).
**Supplementary Note 7 | Qualitative comparison of GeNets to other technologies.** Below we have qualitatively compared both Quack and GeNets to SANTA (27), SAFE (28), IMP 2.0 (29), NetPEA (30), and Cytoscape Apps (31). SANTA can associate networks with gene sets and phenotypes and is suitable for gene prioritization, but it does not readily predict new gene associations. SAFE is a method mainly designed for network comparison rather than annotating gene sets. Furthermore, SAFE requires functional information which cannot be easily obtained in humans. Similarly, NetPEA is a tool for pathway enrichment analysis, but does not provide gene predictions. IMP 2.0 has the potential to make custom predictions, but it integrates data from multiple sources (e.g. IP, co-expression, TF binding site, etc) and cannot be applied to user-defined networks.

**Comparison of Quack to other methods.**

| Method | Reasons why not comparable with Quack |
|--------|--------------------------------------|
| SANTA  | - Associates networks with gene sets and phenotypes; does not predict new gene associations  
- Requires gene weights which may not be readily available |
| SAFE   | - Annotates networks rather than gene sets  
- Requires functional information which can be challenging to obtain in humans |
| IMP 2.0| - Integrates data from multiple sources and thus cannot be run on user-defined networks |
| NetPEA | - Provides pathway enrichment analysis but not gene predictions |
Comparison of GeNets to other methods.

| Method      | Works with user-defined networks | Can train custom network models | Can quantitatively compare networks | Can predict new candidate genes based on seed set in any network | Does not require additional data when running analyses | Analyses can be stored online | Analyses can be interactively explored without significant computational expertise | Analyses can be visualized without significant computational expertise | Analyses can be shared by clicking hyperlinks |
|-------------|----------------------------------|---------------------------------|------------------------------------|---------------------------------------------------------------|------------------------------------------------|-----------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| GeNets      | YES                              | YES                             | YES                                | YES                                                           | YES                                           | YES                         | YES                                                                            | YES                                                                            | YES                                                                            |
| SANTA       | YES                              | NO                              | YES                                | NO                                                            | NO (gene weights required)                     | NO                          | NO (R program available)                                                      | NO (R program available)                                                      | NO (R program available)                                                      |
| SAFE        | YES                              | NO                              | YES                                | NO                                                            | NO (functional information required)           | NO                          | NO (MATLAB & Cytoscape available)                                             | NO (MATLAB & Cytoscape available)                                             | NO (MATLAB & Cytoscape available)                                             |
| IMP 2.0     | NO                               | NO                              | YES                                | NO                                                            | YES (but only for one month)                   | YES                         | YES                                                                            | YES                                                                            | NO                                                                            |
| NetPEA      | YES                              | NO                              | NO                                 | NO                                                            | YES                                           | NO                          | NO                                                              | NO                                                              | NO                                                              |
| Cytoscape Apps | YES                              | NO                              | NO                                 | YES (e.g. Diffusion)                                         | YES                                           | NO                          | YES                                                                            | YES                                                                            | NO                                                                            |
Supplementary Note 8 | Benchmarks comparing Quack to other technologies:

We tested Quack, SANTA (27), and GeneMANIA (32) on neuronal pathways in MSigDB (i.e., the case study presented in the Main Text Fig. 2). Specifically, for each of the 45 neuronal pathways which Quack scored in Fig 2a, we randomly partitioned pathway genes into 70% seed and 30% hold-out genes, and asked each method to predict the hold-out set. We then calculated AUCs based on method-specific scores and pathway membership using the pROC package (https://cran.r-project.org/web/packages/pROC/index.html, version 1.10.0).

To make the fairest comparison, we used the same network (InWeb) for Quack and SANTA. However, when testing GeneMANIA, the Cytoscape plug-in (33) unfortunately does not work. Since uploading own networks is only possible in the Cytoscape plug in, we used the default GeneMANIA network and executed all benchmarks with this method through the web portal [www.genemania.org] manually. Results are summarized in Supplementary Figure 1a. Next, for both Quack and SANTA, we expanded the analysis to all 853 pathways used in the Main Text Fig. 1. We did not perform this analysis for GeneMANIA because it is not logistically possible to run >800 pathways manually using the web interface. We used the same procedure as above to score each of the canonical pathways, and the result is summarized in Supplementary Figure 1b.

Overall, Quack’s performance is significantly better than SANTA (on the same network-pathway pairs) and GeneMANIA on the same pathways using the default GeneMANIA network. Considering the careful training of the Quack machine learning model to learn the specific topology of the neuronal pathways in the InWeb network, it is not surprising that the performance of Quack outperforms the other two other high quality methods.
Supplementary Information References

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. Nature, 409(February), 860–921. http://doi.org/10.1038/35057062
2. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. Nat. Rev. Genet. 12, 56–68 (2011).
3. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell’s functional organization. Nat. Rev. Genet. 5, 101–13 (2004).
4. Hein, O., Schwind, M., & König, W. (2006). Scale-free networks. Wirtschaftsinformatik, 48, 267–275. http://doi.org/10.1007/s11576-006-0058-2
5. Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. Science (New York, N.Y.), 325, 412–413. http://doi.org/10.1126/science.1173299
6. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. Nature 393, 440–2 (1998).
7. Mitra, K., Carvunis, A.-R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. Nature Reviews. Genetics, 14(10), 719–32. http://doi.org/10.1038/nrg3552
8. Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, a W. (1999). From molecular to modular cell biology. Nature, 402(December), C47–C52. http://doi.org/10.1038/35011540
9. Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N. et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25, 309–316.
10. Lage, K. et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc. Natl. Acad. Sci. U. S. A. 105, 20870–5 (2008).
11. Lundby, A. et al. Annotation of loci from genome-wide association studies using tissue-specific quantitative interaction proteomics. Nat. Methods 11, 868–74 (2014).
12. Neale, B. M., Kou, Y., Liu, L., Ma’ayan, A., Samocha, K. E., Sabo, A., … Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature, 485(7397), 242–245. http://doi.org/10.1038/nature11011
13. O’Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246–250 doi:10.1038/nature1098922495309
14. Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., … Daly, M. J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genetics, 7(1), e1001273. http://doi.org/10.1371/journal.pgen.1001273
15. Lage, K. (2014). Protein-protein interactions and genetic diseases: The interactome. Biochimica et Biophysica Acta - Molecular Basis of Disease, 1842(10), 1971–1980. http://doi.org/10.1016/j.bbadis.2014.05.028
16. Goh, K., Cusick, M. E., Valle, D., Childs, B., & Vidal, M. (2007). The human disease network. Proc. Natl. Acad. Sci. USA, 104, 8685–8690.
17. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10
18. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–35 (2006).
19. Peck, D., Crawford, E. D., Ross, K. N., Stegmaier, K., Golub, T. R., Lamb, J. (2006) A method for high throughput gene expression signature analysis. *Genome Biol.* **7**, R61
20. Lamb, J., (2007) The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* **7**, 54–60.
21. Cowley, G. S. et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci. Data* **1**, 140035 (2014).
22. Li, Y., Calvo, S. E., Gutman, R., Liu, J. S. & Mootha, V. K. Expansion of biological pathways based on evolutionary inference. *Cell* **158**, 213–25 (2014).
23. Barzel, B. & Barabási, A.-L. Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* **31**, 720–5 (2013).
24. Feizi, S., Marbach, D., Ménard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–33 (2013).
25. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci USA* 2005;102:15545-15550.
26. Japkowicz, N. (2000). Learning from Imbalanced Data Sets. *Papers from AAAI Workshop*, 21(9), 10–15. Retrieved from http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf
27. Cornish, A. J. & Markowetz, F. SANTA: quantifying the functional content of molecular networks. *PLoS Comput. Biol.* **10**, e1003808 (2014).
28. Baryshnikova, A. Systematic Functional Annotation and Visualization of Biological Networks. *Cell Syst.* **2**, 412–21 (2016).
29. Wong, A. K., Krishnan, A., Yao, V., Tadych, A. & Troyanskaya, O. G. IMP 2.0: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.* **43**, W128-33 (2015).
30. Liu, L. & Ruan, J. Network-based Pathway Enrichment Analysis. *Proceedings. (IEEE. Int. Conf. Bioinformatics Biomed).* 218–221 (2013). doi:10.1109/BIBM.2013.6732493
31. Lotia, S., Montojo, J., Dong, Y., Bader, G. D. & Pico, A. R. Cytoscape app store. *Bioinformatics* **29**, 1350–1 (2013).
32. Zuberi, K. et al. GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* **41**, W115-22 (2013).
33. Montojo, J., Zuberi, K., Rodriguez, H., Bader, G. D. & Morris, Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. *F1000Research* **3**, 153 (2014).