Analyzing and interpreting genome data at the network level with ConsensusPathDB

Ralf Herwig[1], Christopher Hardt[1], Matthias Lienhard[1] & Atanas Kamburov[2–4]

[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany. [2]Department of Pathology and Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Correspondence should be addressed to R.H. (herwig@molgen.mpg.de) or A.K. (kamburov@broadinstitute.org).

Published online 8 September 2016; doi:10.1038/nprot.2016.117

ConsensusPathDB consists of a comprehensive collection of human (as well as mouse and yeast) molecular interaction data integrated from 32 different public repositories and a web interface featuring a set of computational methods and visualization tools to explore these data. This protocol describes the use of ConsensusPathDB (http://consensuspathdb.org) with respect to the functional and network-based characterization of biomolecules (genes, proteins and metabolites) that are submitted to the system either as a priority list or together with associated experimental data such as RNA-seq. The tool reports interaction network modules, biochemical pathways and functional information that are significantly enriched by the user's input, applying computational methods for statistical over-representation, enrichment and graph analysis. The results of this protocol can be observed within a few minutes, even with genome-wide data. The resulting network associations can be used to interpret high-throughput data mechanistically, to characterize and prioritize biomarkers, to integrate different omics levels, to design follow-up functional assay experiments and to generate topology for kinetic models at different scales.

INTRODUCTION

Modern high-throughput experiments such as sequencing, microarray technology or mass spectrometry (MS) experiments generate large genome-wide data sets that provide deep insight into many different levels of molecular information—e.g., the transcriptome, proteome and metabolome, among others. Such information is used, for example, to characterize patient genomes using multiomics data[1], to describe developmental processes with temporal changes[2] or to derive predictive patterns for exogenous agents[3]. An emerging goal of data analysis is to reveal the underlying control mechanisms that govern the measured molecular phenotypes.

Typically, a key result of genome analysis is a list of statistically significant biomolecules (genes, proteins, metabolites) that contribute to the phenotypes of interest. A subsequent task then is to identify which biological functions can be associated with these molecules (over-representation analysis)[4]. This is done mainly by exploring whether predefined annotation sets—for example, specific signaling pathways—are enriched by the molecules under consideration. Independently, such enrichments can be inferred without statistical preselection of the molecules using the entirety of the experimental data ((gene set) enrichment analysis)[5]. Furthermore, data for all or a prioritized subset of molecules can be mapped onto interaction networks and analyzed with graph theoretic approaches. These methods identify subnetworks (network module analysis) that are likely to be responsive to the experiments under analysis[6]. All three approaches aim at enriching genome analysis with mechanistic network information, which enables an understanding of the underlying biological processes.

In ConsensusPathDB[7], we have implemented statistical methods for performing the above tasks by interrogating annotation sets based on molecular interaction information. We agglomerated the contents of 32 major public repositories for human molecular interactions of heterogeneous types, as well as biochemical pathways, resulting in one of the largest interactome collections available (**Table 1**). Furthermore, the database integrates the contents of 15 mouse and 14 yeast interaction repositories. In addition to gene ontology[8] (GO) and pathway annotations, ConsensusPathDB systematically explores the protein–protein interaction (PPI) network, as PPIs are key drivers of biological function[9]. However, only a minor fraction of the estimated ~650,000 human protein interactions have yet been experimentally measured[10]. Moreover, information on molecular interactions is scattered across >500 different databases worldwide[11], which necessitates the integration of as many resources as possible into meta-databases such as ConsensusPathDB (**Box 1**). Such interaction integration allows for better coverage of the interactome, which improves guidance in the functional interpretation of *omics* data.

ConsensusPathDB has been well adopted by the research community. Applications comprise over-representation analysis in order to characterize diverse sets of molecules[12–14], gene set enrichment analysis[15,16] and identification of upstream regulators[17] spanning various biological contexts. Furthermore, ConsensusPathDB is used as a database by other tools—for example, for enrichment analysis by Chipster[18] using web service connections or by Cytoscape[19] using a Java plugin for assessing interaction confidence of PPIs[20]. In addition to these analyses, the tool can be used as a resource for the generation of molecular interaction gene sets, which themselves can be used as predictive signatures. For example, it has been shown that network modules and pathways can be derived as predictive patterns in cancer diagnostics[21], as well as in tumor progression monitoring[22]. This enables biomarker analysis of entities ranging from single molecules to entire pathways.

Overview of the protocol

In this protocol, we review the contents and the different analysis scenarios enabled by ConsensusPathDB. All modules in this protocol aim to enable network-level interpretation and functional characterization of user-specified lists of molecules (genes, proteins and metabolites) and associated

TABLE 1 | Content of ConsensusPathDB.

Content type	Human	Mouse	Yeast
Integrated databases	32	15	14
Unique physical entities	158,523	31,679	17,672
Unique interactions	458,570	34,064	272,094
Gene regulations	17,098	2,196	316
Protein interactions	261,085	23,488	123,842
Genetic interactions	443	194	145,151
Biochemical reactions	21,070	8,186	2,785
Drug–target interactions	158,874	0	0
Pathway gene sets	4,593	2,173	1,101

high-throughput data. ConsensusPathDB helps users working with such data to do the following:

- Infer heterogeneous interaction networks for genes, proteins, metabolites, drugs and other biomolecules
- Compute over-represented pathways, PPI networks, protein complexes and GO annotations from a priority list of genes, proteins or metabolites
- Compute enriched pathways, PPI networks, protein complexes and GO annotations from genome-wide data such as RNA-seq or array technology
- Generate network modules that are over-represented by genes or proteins and thereby explore heterogeneous interactions such as PPI, drug–target, gene regulatory and genetic interactions.

Comparison with other tools

Several excellent tools, of which only some can be mentioned here, are available that perform either over-representation analysis (e.g., DAVID[23], IPA[24] and Enrichr[25]), gene or metabolite set enrichment analysis (e.g., GSEA[26] and MetaboAnalyst[27]) or network module analysis (e.g., Cytoscape[28] and Genes2Networks[29]). Although most of these tools are restricted to specific types of analysis and to a specific type of biomolecule, ConsensusPathDB offers a wider range of analysis functions and the option for gene/protein and metabolite analysis (**Table 2**). The statistical methods for over-representation analysis, enrichment analysis and network module analysis implemented by the individual tools differ, and thus results achieved with the same input can be fairly different. With respect to content, ConsensusPathDB has a focus on molecular interactions, and it provides deep exploration of the interactome network, protein complexes and pathway resources, whereas other tools incorporate additional annotation sets, for example, based on genomic locus enrichment, disease associations, experimental signatures or literature-derived sets. Huge collections of such annotation signatures are accessible through systems such as MSigDB[30]. Parallel attempts for sampling huge amounts of interaction data within a common framework are STRING[31] and PathwayCommons[32].

Limitations of the protocol

ConsensusPathDB currently supports only three organisms (human, mouse and yeast), and it is thereby missing widely used model organisms such as rat, fly and worm, among others, for which comprehensive interaction information has been collected and made available in the past. Moreover, ConsensusPathDB does not hold information on microorganisms—e.g., bacteria or fungi. In cases in which interaction information is available, the inclusion of more organisms is a key step in the future development of ConsensusPathDB.

Another limitation is the focus on annotation sets that are derived from molecular interactions and GO terminology. As stated in the previous section (**Table 2**), several tools incorporate additional information that allows for interpretation of data in alternative directions. However, a review of the literature shows that, by far, most applications use functional annotation sets defined by GO and pathway annotations, thus justifying the current focus of ConsensusPathDB.

With regard to the web server, ConsensusPathDB has some limitations with respect to visualization components. Presumably the most widely used tool available in this regard is Cytoscape, and thus we offer network downloads in a Cytoscape-compatible format, which enables easy transfer of computed network modules.

For some of its functionality, ConsensusPathDB already offers web services in order to allow the integration of analysis steps into automated workflows and stand-alone tools. However, not all steps described in this protocol are yet implemented as web services; their further development is a primary future task.

The response time of ConsensusPathDB depends on the size and complexity of the interaction network under investigation. For example, performing network analyses with many input nodes or many different types of interactions can lead to slow response and limited visualization performance.

Experimental design

Analysis paths. ConsensusPathDB contains predefined annotation sets that hold functional information such as pathways, GO categories, protein complexes and PPI network neighborhoods that were derived from the integrated resources.

Depending on the user's input, ConsensusPathDB allows the following analyses (**Fig. 1**):

- *Analysis path 1.* The interaction neighborhood of a single molecule can be inferred and a corresponding network can be generated; this can be done, for example, to reveal network-level information (i.e., interaction partners) for biomarkers of interest.
- *Analysis path 2.* Uploading a list of molecules (genes, proteins and metabolites) allows either performing over-representation analysis with predefined annotation sets or computing network associations between the molecules of interest through mining of the integrated interaction network.
- *Analysis path 3.* Inserting molecules and associated experimental data allows computing enrichment analysis of the annotation sets; this path uses a more unbiased analysis compared with analysis path 2, because it is not dependent on a predefined priority list of molecules.

To exemplify the procedures in this protocol, we use different data sets from various biological backgrounds, and measure using different high-throughput technologies. For analysis path 1,

Box 1 | Molecular interactions

Molecular interactions are key drivers of cellular function. In the times of omics technology, an ever-increasing number of molecular interactions are measured and cataloged. For example, huge amounts of PPIs have been measured by co-immunoprecipitation, tandem-affinity purification and yeast two-hybrid analysis, among others. ChIP-seq experiments allow the charting of protein–DNA interactions and histone modifications. Phosphoproteome measurements with MS such as ITRAQ and SILAC provide new insights into signaling networks. Metabolomics technologies such as NMR or gas chromatography–MS measure metabolites and fluxes through metabolic networks. These technologies gave rise to the development of multiple repositories that store and curate the experimental data along with previous literature annotation.

The ConsensusPathDB is a meta-database that currently consolidates human molecular interactions from 32 different databases, mouse molecular interactions from 15 different databases and yeast molecular interactions from 14 different databases.

Interaction databases and interaction types (human)

Interaction types include the following:

- Protein interactions (BIND, Biogrid, CORUM, DIP, DrugBank, HPRD, InnateDB, Intact, MINT, MIPS-MPPI, MatrixDB, NetPath, PDB, PDZBase, PIG, PINdb, PhosphoPOINT, Reactome and Spike)
- Signaling reactions (BioCarta, INOH, InnateDB, KEGG, NetPath, PID, PhosphoPOINT, PhosphoSitePlus, Reactome, Spike and Wikipathways)
- Metabolic reactions (EHMN, HumanCyc, INOH, KEGG, Reactome and Wikipathways)
- Gene regulatory interactions (BIND, BioCarta, InnateDB, PID and Spike)
- Genetic interactions (Biogrid)
- Drug–target interactions (Chembl, DrugBank, KEGG, PharmGKB, and TTD)
- Biochemical pathways (BioCarta, EHMN, HumanCyc, INOH, KEGG, NetPath, PID, PharmGKB, Reactome, SMDPB, Signalink and Wikipathways)

Interaction databases and interaction types (mouse)

Interaction types include the following:

- Protein interactions (BIND, Biogrid, DIP, InnateDB, Intact, MINT, MIPS-MPPI, MatrixDB, PDB, PDZBase and Reactome)
- Signaling reactions (InnateDB, KEGG, PhosphoSitePlus, Reactome and Wikipathways)
- Metabolic reactions (KEGG, MouseCyc, Reactome and Wikipathways)
- Gene regulatory interactions (BIND and InnateDB)
- Genetic interactions (Biogrid)
- Drug–target interactions (KEGG)
- Biochemical pathways (KEGG, MouseCyc, Reactome and Wikipathways)

Interaction databases and interaction types (yeast)

Interaction types include the following:

- Protein interactions (BIND, Biogrid, CYC2008, DIP, Intact, MINT, MIPS-MPACT, PDB, PINdb, PTM and Reactome)
- Signaling reactions (KEGG, Reactome and Wikipathways)
- Metabolic reactions (KEGG, PTM, Reactome, Wikipathways and YeastCyc)
- Gene regulatory interactions (BIND and PTM)
- Genetic interactions (Biogrid)
- Drug–target interactions (KEGG)
- Biochemical pathways (KEGG, Reactome, Wikipathways and YeastCyc)

we exemplify the protocol steps using the epidermal growth factor receptor (*EGFR*) gene that is a widely mutated gene in different types of cancer and also a primary target of cancer therapy[33]. For analysis path 2, we exemplify the over-representation analysis for genes using a list of 18 frequently mutated genes derived from whole-exome sequencing of a large lung adenocarcinoma cohort[1] (**Supplementary Data 1**). As a test case for over-representation analysis of metabolites, we use a list of 130 known uremic toxins that are associated with dysfunction of the kidney[34] (**Supplementary Data 2**). To demonstrate the network module analysis, we examined 691 targets of histone modification (H3K4me2) measured with ChIP-seq that are specific to T helper type 2 (T_H2) cells, as compared with naive T cells (**Supplementary Data 3**). The goal of the analysis is to recover potential gene regulatory networks controlling these genes, as was done in the original publication[17]. For analysis path 3, we use public expression data derived from different stages of human embryonic development that were generated with RNA-seq[2] (**Supplementary Data 4**). These data cover a wide range of genome analysis applications and diverse biological backgrounds. The corresponding gene lists vary in size from 18 (lung adenocarcinoma driver mutations) to ~16,000 (RNA-seq data set), demonstrating the scalability of the ConsensusPathDB analysis tools.

Identifier mapping. A recurrent problem when integrating data from different resources, or when analyzing high-throughput data by comparison with existing databases, is the nonuniformity of gene/protein/metabolite identifiers. In ConsensusPathDB, we have created comprehensive identifier maps by parsing the contents of 11 major genomic, proteomic and metabolite databases such as Ensembl, Uniprot and PubChem. These maps were used to match gene, protein and metabolite identifiers

TABLE 2 | Tool comparison.

Tool	Access	Analysis functions						Content types				
	Free	Upload of user-defined background	Gene-based ORA	GSEA	Network module analysis	Metabolite-based ORA	MSEA	Network neighbors	Protein complexes	Pathway resources	GOASs	Other
ConsensusPathDB	X	X	X	X	X	X	X	X	X	X	X	
DAVID	X	X	X						X	X	X	X
EnrichR	X		X	X				X	X	X	X	X
Cytoscape	X	X	X	X	X			X		X	X	X
IPA		X	X	X	X	X				X	X	X
GSEA/MSIGDB	X	X		X				X		X	X	X
Genes2Networks	X	X			X							
MetaboAnalyst	X	X				X	X			X		X
STRING	X		X					X	X	X	X	X
PathwayCommons	X							X	X	X		X

GOAS, gene ontology annotation set; GSEA, gene set enrichment analysis; MSEA, metabolite set enrichment analysis; ORA, over-representation analysis; other, annotation sets based on literature, experimental data, chromosomal location, disease associations, protein domains and so on.

originating from the 32 integrated sources of interaction and pathway information. Furthermore, they are used to map identifiers from the user input to these physical entities, and hence they allow great flexibility with respect to what identifier namespace is chosen by the user.

Annotation sets. ConsensusPathDB offers four types of predefined annotation sets: neighborhood-based entity sets (NESTs), protein complexes, pathways and GO terms.

- *NESTs.* These sets are derived from the integrated interaction network, which includes four types of biological interactions: protein–protein, biochemical, gene regulatory and genetic interactions. A NEST is defined as a central protein and its network neighbors. The size of the network neighborhood is determined by its radius. The user can choose between a radius equal to one and a radius equal to two. A radius equal to one adds only the direct neighbors to the center protein, whereas a radius equal to two adds, in addition, all direct neighbors of the direct neighbors. We recommend using a radius equal to one; otherwise, the neighborhoods grow too large and lose specificity. There are as many NESTs as proteins in the integrated network.
- *Protein complexes.* These sets are derived from specific databases that hold information on protein complexes. Note that most annotated protein complex sets are rather small (<5 members).
- *Pathways.* These sets comprise metabolic, signaling and gene regulatory pathways annotated by 12 source databases for human (4 each for mouse and yeast). Pathways range from very large biological processes—covering, for example, the complete metabolism and having $>1{,}000$ members—to very specific concepts that describe detailed processes.
- *GO terms.* ConsensusPathDB offers four levels of GO categories ranging from very general terms (level = 2) with $>1{,}000$ members to more specific terms (level = 5). In the analysis, the user can restrict the categories to specific level(s) or to the specific GO tree branches covering 'biological process', 'molecular function' and 'cellular compartment'.

Pathway annotation—specificity and redundancy. The pathway concept is essential for modern biology, and it usually describes a certain cellular process, for example 'apoptosis', in which the involved proteins or metabolites exert specific functions and are interconnected by molecular interactions of diverse types. ConsensusPathDB agglomerates 4,593 human pathway concepts (mouse: 2,173 and yeast: 1,101) originating from 12 different resources (**Table 1**). On one hand, these pathway concepts are partially redundant because they describe subpathways of a given pathway that are annotated by the same database. For example, the pathway 'apoptosis' might cover the subpathways 'extrinsic apoptosis' and 'intrinsic apoptosis' among others, with corresponding subsets of proteins. On the other hand, most generic pathways are annotated by several databases, leading to more than one annotation set referring to 'apoptosis'.

It is worth noting that pathway concepts from different resources might in fact involve different sets of molecules even when describing similar molecular processes. As a consequence,

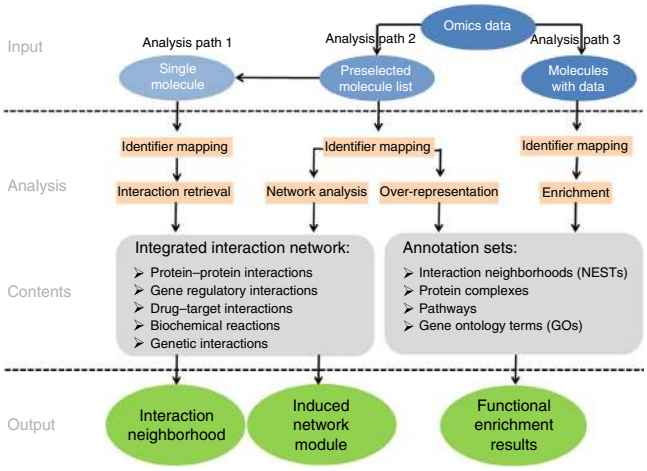

Figure 1 | Outline of the protocol. Three paths of analysis are described in the protocol that depend on the user's input. The content of the ConsensusPathDB (i.e., the integrated interaction graph and the predefined annotation sets) can be explored with single molecules (analysis path 1), with priority lists of molecules (genes, proteins and metabolites; analysis path 2) or with associated experimental data (analysis path 3). The Web server functionality includes over-representation analysis, enrichment analysis and network module analysis. The outputs are the generated tables and graphs that can be downloaded for further inspection.

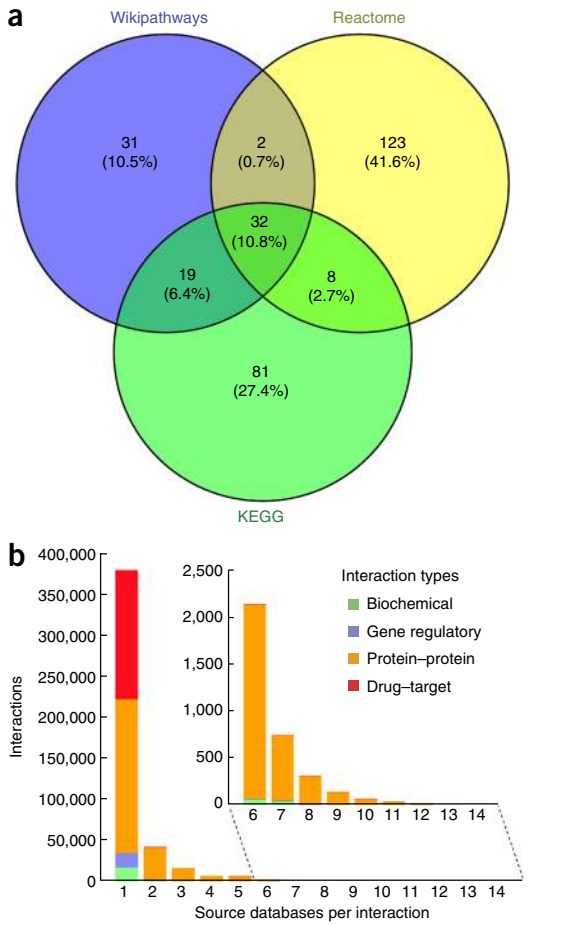

Figure 2 | Pathways and interactions in ConsensusPathDB. (**a**) Annotation specificity. Venn diagram generated with Venny 2.1 showing the proteins annotated for the apoptosis signaling pathway in three different databases (Wikipathways, WP254; Reactome, R-HSA-109581; and KEGG, hsa:04210). In total, 296 different proteins are annotated for apoptosis signaling, 84 in Wikipathways, 165 in Reactome (proteins with gene symbols) and 140 in KEGG. 61 of these proteins are common to all or at least two databases (20.6%), whereas the vast majority of proteins (235; 79.4%) are specific for a single database. (**b**) Histogram of the number of contributing databases per interaction (genetic interactions have been omitted in this figure, as their total number, n = 443, is comparatively too small to be visible).

this can lead to differences in functional enrichment analyses (analysis paths 2 and 3), because the different annotation sets might have deviating overlaps with the gene list submitted by the user. For example, comparison of gene sets for the 'apoptosis' pathway in the three widely used databases KEGG[35], Reactome[36] and WikiPathways[37] reveals that 79% of the annotated proteins are specific to a single database, as compared with the number of proteins that are shared by at least two of the three databases (**Fig. 2a**). The reason for this is that pathway boundaries are not clearly defined and that expert opinion on the extent of cross talk with other pathways is highly variable. In addition, pathway annotations are commonly focused on specific substructures or specific cellular contexts (e.g. tissues, diseases and organisms), which might result in variations of the assembled gene lists. Consequently, in ConsensusPathDB, such overlapping pathway concepts are not merged to generalized pathways; instead, the redundancy is kept and the annotated pathway set is always disclosed together with its source database.

Interaction retrieval for single biomolecules. ConsensusPathDB holds 158,523 unique physical entities (mouse: 31,679 and yeast: 17,672), and it offers the possibility of retrieving interaction information for these entities. The concept of an interaction in ConsensusPathDB is very general, so that proteins can have connections not only to other proteins but also to drugs, complexes or metabolites (**Box 1**). By selecting specific interactions, the user can generate fairly complex interaction networks.

The source database for each interaction is tracked by a color code, providing the user with the information on where the interaction originates. This allows for easy visualization of possible redundancy between databases, which might serve as an indicator for assessing confidence of the particular interaction. **Figure 2b** shows the distribution of the different interaction types and their origins. Most interactions are present for protein–protein and drug–target interaction types and are predominantly specific for a single or low number of databases.

Another level of confidence assessment is available for binary PPIs. Because a lot of PPI resources are integrated in ConsensusPathDB, control of false-positive interaction is of utmost importance. Therefore, binary PPIs have a quality score (range [0,1]) that is displayed with a color code. This score was computed as a meta-score integrating different methods for interaction confidence assessment, including graph-based topological criteria[38–40], literature evidence and pathway co-occurrence[41], and semantic similarity[42] using our IntScore[43] web tool (**Box 2**).

This section starts by defining the biomolecule of interest. Next, all interactions of that molecule are shown, which can be selected and visualized by the user based on prioritization or quality assessment. After generating the graph, the user can expand it at any given node and update the graph accordingly with further interactions.

Box 2 | Interaction confidence assessment

PPIs are derived from various different resources. It is well known that protein interactions are error prone and that PPI networks contain large numbers of false-positive interactions. This has given rise to attempts to control this error by judging interaction confidence. ConsensusPathDB PPIs were examined with six different methods:

Topology-based criteria

CAPPIC[37]—This method applies Markov clustering and evaluates the confidence that a PPI belongs to its local cluster of PPIs; score range is [0,1].

Common neighbors[38]—This method evaluates common network neighborhoods of interacting nodes; score range is [0,1].

Geometric embedding[39]—This method embeds the network in a geometric space and weights edges according to the distance between their incident nodes in that space; score range is [0,1].

Annotation-based criteria

Literature evidence[40] is the number of publications that report the interaction; score range is [0, ∞]

Pathway co-occurrence[40] indicates whether interaction partners are found in the same pathway; score range is {0,1}.

Semantic similarity[41] shows similarity of interaction partners computed from the GO tree; score range [0,1].

Interaction confidence computation was done with the IntScore tool (http://intscore.molgen.mpg.de)[42], which allows the assessment of additional user-defined PPIs.

In ConsensusPathDB, all binary PPIs have an aggregated confidence score, range [0,1], that was computed as a consensus score across the six methods described above. As a result, three classes of PPIs can be retrieved:

High-quality PPI interactions (score > 0.95):	81,736
Medium-quality PPI interactions (0.5 ≤ score ≤ 0.95):	52,514
Low-quality PPI interactions (score < 0.5):	70,620

All integrated PPIs can be downloaded from the ConsensusPathDB web server.

Over-representation analysis. This section of the protocol describes the interrogation of the annotation sets (pathways and others) with lists of genes, proteins or metabolites. This analysis requires prior data analysis by the user outside of ConsensusPathDB—for example, by applying a statistical test to the genome data and preselecting the most significant molecules, as is typically done for RNA-seq or microarray data. For computing the significance of the over-representation of the annotation sets with respect to user-input molecules, ConsensusPathDB applies Fisher's exact test. For each annotation set, the *P* value is calculated as follows:

$$P(x\,|\,n,m,N) = 1 - \sum_{i=0}^{x-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

where x is the number of entities uploaded by the user that overlap with the entities in the set, n is the number of entities in the set that are identifiable in the user-provided identifier namespace (e.g., if a pathway contains ten genes but only nine of them have an identifier of the type the user selects, then $n = 9$), m is the number of entities that the user uploads and that are part of at least one annotation set of the selected type and N is the number of entities present in the union of all annotation sets of the selected type and identifiable in the user-provided identifier namespace (background). As many annotation sets are tested, we correct for multiple hypothesis testing using the false discovery rate procedure within each type of annotation set[44]. This is a widely used method that is also applied by many other tools. After computation, annotation sets are ordered according to significance, and they can be downloaded in table format. In addition, specific sets and their overlaps can be visualized. Importantly, in contrast to many other tools, not only gene and protein lists—but also lists of metabolites—can be submitted to ConsensusPathDB; furthermore, the resulting functional categories can be visualized as overlap graphs, described below.

Enrichment analysis. This section describes the enrichment analysis using all experimental data in an unbiased way rather than prioritized molecule lists. It is worth noting that this analysis path is complementary to the above. With over-representation analysis, only the most significant changes are typically evaluated, whereas enrichment analysis is also sensitive to subtle but congruent changes of many molecules in the set. Therefore, in practice, a pathway might emerge as significantly enriched, although none of the genes in that pathway were in the priority list submitted for over-representation analysis. Furthermore, completely novel information can be retrieved—for example, involving interaction neighborhoods (NESTs) of proteins that were not even captured by the omics platform under consideration[45].

ConsensusPathDB assumes that the user submits case–control data, for example disease versus normal or treated versus untreated conditions, and that for each molecule both values are given. Alternatively, a single log-fold change value for each molecule can be submitted.

For computing enrichment *P* values for the annotation sets, ConsensusPathDB applies Wilcoxon's matched-pairs signed-rank test. This test is robust against experimental outliers because it is based on ranks rather than on experimental values. For each annotation set i with n molecules $m_{i1}, \ldots, m_{in}$ and experimental measurements $x_{i1}, \ldots, x_{in}$ and $y_{i1}, \ldots, y_{in}$, the ranks of the absolute differences, $|x_{ij} - y_{ij}|$, of the two experimental conditions are computed; next, all ranks from pairs with positive differences R^+ and negative differences R^- are summed. The test statistic is the minimum of both rank sums: $R = \min\{R^+, R^-\}$. Expectation,

$E(R)$, and variance of R, $\mathrm{Var}(R)$, can be derived, and the Z-score, Z, measures the significance of the observed outcome:

$$Z = \frac{R - E(R)}{\sqrt{Var(R)}}$$

with $E(R) = \frac{n(n+1)}{4}$ and $Var(R) = \frac{n(n+1)(2n+1)}{24}$, respectively[46].

If both conditions have similar values for $m_{i1}, \ldots, m_{in}$, then both rank sums are equal and the resulting test statistic is not significant, whereas otherwise the Z score becomes significant. The enrichment method was applied for the first time with array data monitoring blastocyst development[47] and since then has been applied multiple times.

As in the previous section, annotation sets are ordered according to significance, and they can be downloaded in table format. In addition, specific sets and their overlaps can also be visualized by overlaying color-coded experimental (e.g., gene expression) data.

Network module analysis. The idea behind this kind of analysis is to start with a priority list of genes/proteins (seed nodes) and to compute from the underlying interaction graph a subgraph that connects the seed nodes together through functional and physical links. In ConsensusPathDB, we re-implemented a previously published induced network approach[29]. The output subnetwork produced by this algorithm may optionally include nodes that are not originally part of the user's list but have an abundant number of connections to seed nodes. These nodes are called intermediate nodes. Intermediate nodes are potential upstream regulators and specific participants in pathways, protein complexes and modules involving the input seed list. For example, if a list of downregulated genes as measured through RNA-seq is uploaded, ConsensusPathDB may output a common transcription factor regulating those genes that might be mutated or otherwise dysregulated (but is not on the transcriptional level and hence not included in the seed list). To judge the significance of the intermediate node, a Z score value is computed using a binomial proportions test as follows:

$$Z = \frac{\left(\frac{a}{c} - \frac{b}{d}\right)}{\sqrt{\frac{\frac{b}{d}\left(1 - \frac{b}{d}\right)}{d}}}$$

Here, a equals the number of links from the intermediate node being examined to nodes from the input seed list, b equals the number of total links for the intermediate node in the consolidated background reference network, c is the number of total links in the output subnetwork and d is the number of total links in the consolidated background reference network. The threshold for the Z score can be set interactively by the user. It is clear that for large lists and/or low Z scores the outputted network can be fairly large, and thus we recommend either not using >100 genes/proteins as seed nodes or restricting the analysis to specific interaction types in the parameter setting.

MATERIALS

EQUIPMENT

- *Data files*. Example data files are supplied as **Supplementary Data 1–4**.
- Computer with Internet access. The link to ConsensusPathDB is http://consensuspathdb.org. For hardware and software requirements, see Equipment Setup.

EQUIPMENT SETUP

Hardware requirements ConsensusPathDB is accessed via the web interface. Data analysis is performed on a dedicated web server. Thus, even complex network analyses can be performed by the user with a standard computer. For this protocol, we ran all analyses on (i) a standard single-CPU 3.6 GHz Windows 7 PC with 8 GB RAM, (ii) a Linux workstation, (iii) a MacBook Pro 2.8 GHz computer and (iv) an Android (version 5.1.1) Samsung S6 smartphone.

Software requirements The ConsensusPathDB can be accessed via a web server or via web services. Any modern HTML5-enabled browser can be used. The analysis performed for this protocol was tested with the current versions of Firefox (version 44.0.2) and Chrome (version 48.0.2564.109 m) on the Windows and Linux PCs, with Safari (version 8.0.3) on the MacBook computer and with a standard Android web browser. We recommend upgrading to the latest version of JAVA. As HTML5 support in Internet Explorer has historically been poor, we have not optimized the web server for this browser, and thus we cannot recommend its use.

Data files The data used in this protocol consist essentially of lists of genes and metabolites with or without experimental data. All data files are available in **Supplementary Data 1–4**.

Results download Results comprise tables in .txt file format and figures in .png file format. Furthermore, the networks can be downloaded in .sif format, as well as .owl BioPax 3 format, which is readable by other tools such as Cytoscape[28]. ConsensusPathDB also offers a download section comprising pathway annotation sets, as well as the integrated PPI network.

Web services Automated access to ConsensusPathDB's over-representation and enrichment analysis functionality is possible through a SOAP/WSDL interface. A WSDL file needed for connecting to the SOAP/WSDL interface is provided in the 'download / data access' section on the web page.

PROCEDURE

Selection of the organism of interest ● TIMING < 1 min

1| Choose the correct version of the database (human, mouse or yeast). This protocol refers to the human version, which is the default setting. The mouse and yeast versions of ConsensusPathDB can be accessed on the home page by clicking the respective organism at the top center position. The active version of the database is shown in red and underlined.

? TROUBLESHOOTING

Data analysis

2| Depending on the required analysis method, see the table below to select the appropriate option.

Option	Method	Description
A	Network neighborhoods of single biomolecules	Search and retrieve the complex interaction neighborhood of a biomolecule of interest; interactively expand the network or overlay experimental data
B	Over-representation analysis of a batch of genes/proteins	Characterize and visualize a predefined gene/protein list with respect to GO, pathway, complexes and interaction information
C	Over-representation analysis of a batch of metabolites	Characterize and visualize a predefined metabolite list with respect to pathway information
D	Induced network approach	Identify network relationships of a list of genes/proteins in order to retrieve upstream regulators, or drug or protein interactions
E	Enrichment analysis with high-throughput experiments	Characterize and visualize genes/proteins with associated experimental data with respect to GO, pathway, complexes and interaction information (whole-genome approach)

(A) Identification of network neighborhoods of individual biomolecules ● TIMING 1–10 min

(i) Submit a single entity (**Fig. 3a**). This is done by clicking on 'search' and 'interactions of molecules/pathways'. There are two ways to search for a molecule of interest: The user can enter the molecule of interest in the input box and specify whether the text represents a name—for example, 'epidermal growth factor receptor'—or the user can enter an identifier—for example 'EGFR' or 'ENSG00000146648'. A list of 138 valid identifier types is displayed when clicking on 'list of valid types'.

? TROUBLESHOOTING

(ii) Specify search entry. The next page displays all results that match the search entry. For each entry, it holds the available names, the corresponding types and the supporting databases. The user can then select the corresponding entry and click 'show interactions'.

▲ **CRITICAL STEP** For highly connected biomolecules such as EGFR, it may take 10–30 s until the full list of interactions is loaded, depending on the Internet connection. Wait until the full list is loaded before proceeding.

(iii) Review the interaction results page. The user can review the interactions retrieved from the integrated databases (**Fig. 3b**). Interactions cover PPI, gene regulatory, genetic, metabolic, signaling and drug–target interactions. The specific role of the molecule of interest is displayed by a one-letter code. Mouseover reveals the full text, for example, '*I*'=physical interactor and '*T*'=target. For PPIs, in addition, a quality criterion (range [0,1]) is given with a traffic light-code that represents the interaction confidence (green: >0.95, orange: [0.5, 0.95] and red: <0.5). The rightmost column displays the source database; original information about each interaction from the corresponding source database can be accessed by a mouse click. After selecting all or several interactions, the user can click on 'map and visualize interactions'.

▲ **CRITICAL STEP** Note that selecting all interactions for highly connected molecules (e.g., EGFR has 4,670 interactions, of which 2,095 are distinct) will generate a warning. It might be infeasible to display these networks. We thus recommend focusing on smaller subnetworks—e.g., by selecting interactions of interest from the interaction list rather than clicking 'select all'.

? TROUBLESHOOTING

(iv) Explore and modify the interaction network. The network is displayed as a bipartite graph in which interactions are displayed as circular nodes and biomolecules as square nodes (**Fig. 3c**). The colors and shapes represent the different types of interactions and biomolecules, respectively, as explained in the graph legend. The network is interactive and can be rearranged by the user. Furthermore, specific action items can be performed when clicking the nodes.

Feature	Description/options
Biomolecule node (square)	• Physical entity information: displays the different identifier and annotation
	• More interactions of this entity: allows selection of further interactions of this biomolecule and, thus, extending the network
	• Hide external interactions: hides all interactions that are not related to that specific node
	• Hide name: hides the name of that specific node

Interaction node (circular)	• Interaction information: displays details of the interaction
	• Hide interaction
	• Hide all except the current interaction
	• Hide secondary participants: physical entities with a secondary role (e.g., enzymes, modifiers) can be hidden, for example, when the user wishes to visualize only the mass flow in a biochemical pathway
Task bar	• Graph legend: displays the legend
	• Overlay values: allows submitting experimental data for the selected biomolecules (cf. Step 2E(v))
	• Graph settings: allows modification of the graphical output
	• Misc functions: displays a summary of network statistics and allows the network to be exported as a BioPAX level 3 .owl file or as plain graph dump

? TROUBLESHOOTING

(B) Over-representation analysis of a batch of genes/proteins ● TIMING 1–10 min

(i) Submit a list of genes/proteins (**Fig. 4a**). Submit a list of genes/proteins of interest and perform over-representation with predefined functional sets such as pathways and GO associations. This is done by clicking on 'gene set analysis' and 'over-representation analysis'. There are two options for insertion: copy the list of genes/proteins into the box, or upload a corresponding text file (.txt format). As an example in this protocol, the gene list from **Supplementary Data 1** can be used.

▲ CRITICAL STEP ConsensusPathDB recognizes multiple entries that are the same. It reduces the list to unique identifiers automatically. Thus, if the user's list contains duplicates, then the system reports fewer entries than were in the submitted list.

(ii) Specify the background for the genes/proteins. By default, the whole genome/proteome is used as a background, but the user can provide his or her own background sets instead (e.g., when only a subset of the genome has been measured). This influences the computation of the statistical significance.

(iii) Specify identifier type. Diverse identifier types (e.g., Ensembl, Entrez and Uniprot) are supported. Select the appropriate ones for your list and hit 'Proceed'. For the example gene list used here, select 'gene symbol (HGNC symbol)'.

▲ CRITICAL STEP The identifier type must match the names in the submitted list. Otherwise, the system reports that no identifier could be mapped.

? TROUBLESHOOTING

(iv) Select annotation sets for genes/proteins. The next page allows selection of the different types of annotation sets for over-representation analysis (**Fig. 4b**).

Option	Description
Neighborhood-based sets (NESTs)	• Protein interaction neighborhoods centered on a certain protein; NESTs can be of radius 1 (direct neighbors) or 2 (i.e., including direct neighbors of direct neighbors)
Pathway-based sets	• Interaction networks as defined by 12 different public databases, which can be specified by the user
Gene ontology (GO) categories	• This option allows filtering for GO levels (depth of terms in the GO-directed acyclic graph), as well as for the three GO domains (biological process, molecular function and cellular component)
Protein-complex-based gene sets	• Sets of genes whose products are found together in protein complexes

For each annotation set, different parameters specifying overlaps, *P* value cutoffs and minimal set sizes can be adjusted by the user. After setting all optional parameters, click 'Find enriched sets' to get to the results page.

▲ CRITICAL STEP The NEST analysis can take up to several minutes, depending on the settings. We recommend using NEST radius equal to two only exceptionally. Note that if the response time is too long, the session expires and the results are lost.

? TROUBLESHOOTING

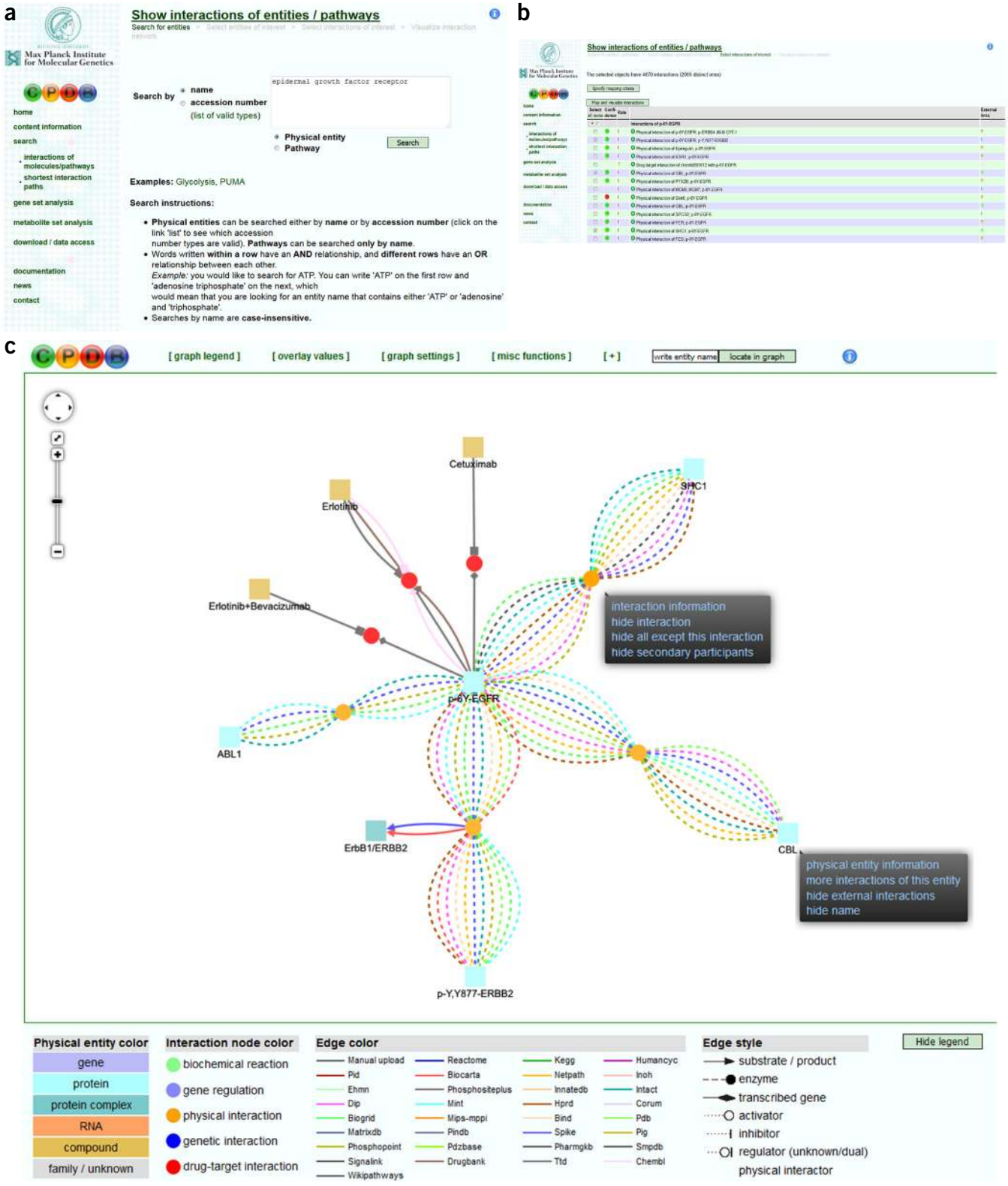

Figure 3 | Interaction neighborhood retrieval in ConsensusPathDB. (**a**) Search field. (**b**) Interactions found for the molecule of interest (here EGFR). (**c**) Complex interaction neighborhood comprising different interaction types (colored circles) and biomolecules (colored squares).

(v) Inspect over-representation analysis results. For each annotation set, a different table is generated listing the enriched sets by *P* value (**Fig. 4c**). The information comprises annotated names, source, set size, overlap, *P* value and *q* value. The individual annotation sets, here pathways, can be further inspected by clicking on the number in the

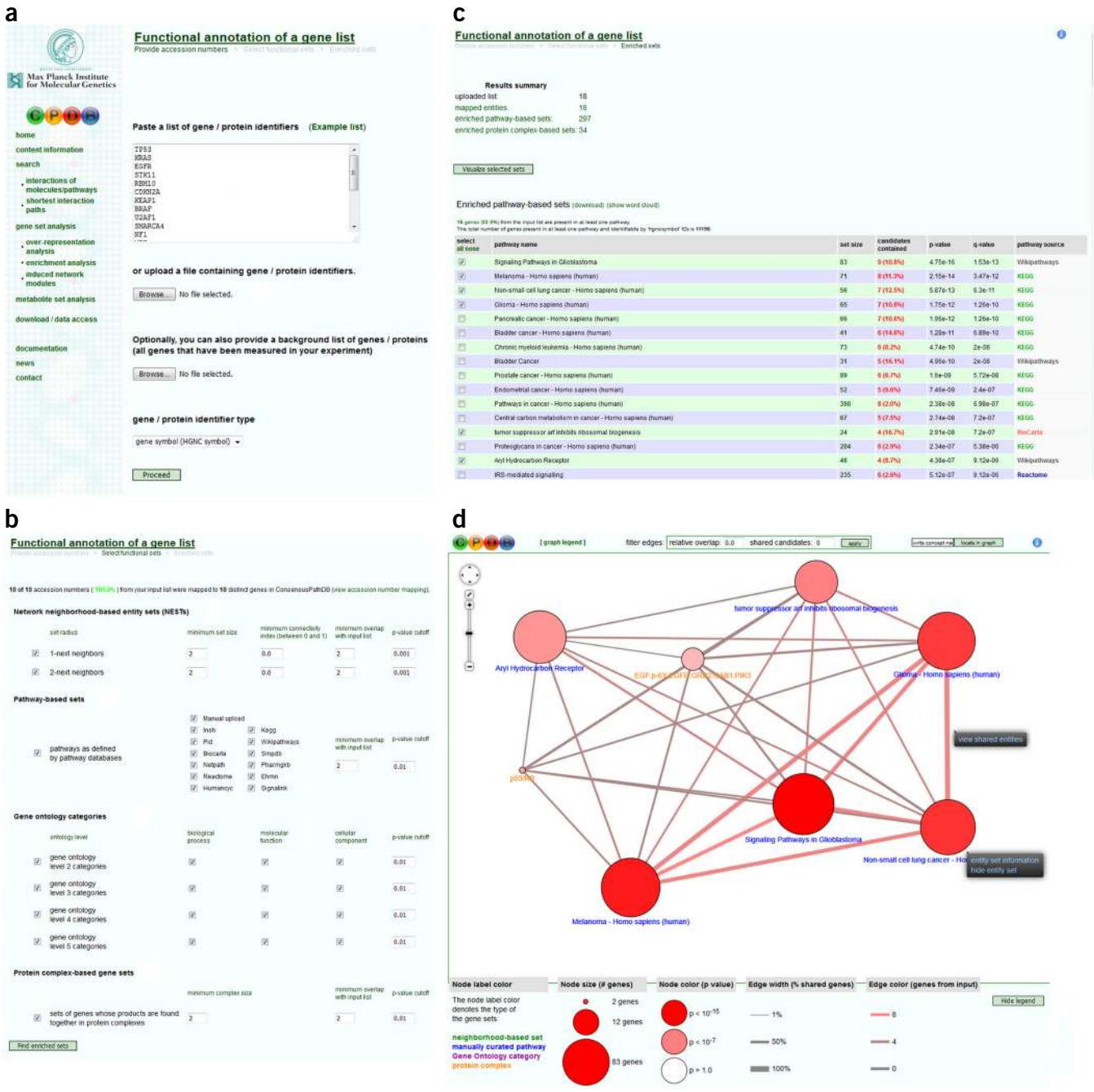

Figure 4 | Over-representation analysis with gene lists in ConsensusPathDB. (**a**) Submission page. (**b**) Selection of annotation sets. (**c**) Results page showing top-enriched pathways. (**d**) Visualization of overlap between enriched annotation sets.

column 'candidates contained'. Further details of the original annotation can be inferred by clicking on the source database link in the rightmost column. This lists the members of the annotation set and the overlap with the gene list. The enriched sets can be downloaded as a tab-delimited file or visualized as a word cloud. Annotation sets can be checked by tick boxes and visualized in graph format by clicking 'Visualize enriched sets'.

(vi) Visualize enriched annotation sets. The different sets and their overlap can be viewed in an interactive map (**Fig. 4d**). The size of nodes and edges is proportional to the number of members and the overlap of the annotation sets, as described in the graph legend. Further information for each annotation set can be reviewed by clicking on the node. Further information for each edge can be reviewed by clicking on the edge. The complexity of the graph can be reduced by filtering edges based on the overlap.

(C) Over-representation analysis of a batch of metabolites

● **TIMING 1–10 min**

(i) Submit the list of metabolites. As in the case of genes/proteins, submit a list of metabolites of interest and perform over-representation with predefined functional sets such as pathways and GO associations. This is done by clicking on 'metabolite set analysis' and 'over-representation analysis'. There are two options for insertion: copy the list of metabolites into the box or upload a corresponding text file (.txt format). As an example in this protocol, the metabolite list from **Supplementary Data 2** can be used.

(ii) Specify the background for the metabolites. By default, the whole metabolome is used as a background, but the user can provide his or her own background sets instead (e.g., when only a subset of the metabolome has been measured).

(iii) Specify the identifier type. Diverse identifier types (KEGG, ChEBI, PubChem, CAS and HMDB) are supported. Select the appropriate ones for your list (for this example, select 'KEGG') and hit 'Proceed'.
▲ **CRITICAL STEP** The identifier type must match the names in the submitted list. Otherwise, the system reports that no identifier could be mapped.
▲ **CRITICAL STEP** Metabolites are often referred to by nonstandardized names—for example, 'glyoxal'. To map these names to database entries, it is usually necessary to convert them into the corresponding identifiers; otherwise, the system cannot match the entry. This is a manual step, which should be done outside of ConsensusPathDB. In this example, we have converted all identifiers to KEGG IDs, resulting in 'C14448' instead of 'glyoxal'.
? TROUBLESHOOTING

(iv) Select annotation sets for metabolites. For metabolites, the analysis is restricted to pathway annotation sets (**Fig. 5a**). Select pathway database resources, set a minimum required overlap and define a *P* value cutoff. Click on 'Find enriched sets' to get to the results page.

(v) Inspect over-representation analysis results. The resulting page lists the pathways that are enriched among the input set, ranked by *P* value (**Fig. 5b**). Information is given on pathway name, size, overlap, *P* value, *q* value and source database. The individual pathways can be further inspected by clicking on the number in the column 'candidates contained'. This lists the members of the annotation set and the overlap with the gene list. The enriched sets can be downloaded as a tab-delimited file or visualized as a word cloud. By ticking boxes, annotation sets can be checked and visualized in graph format.

(vi) Visualize enriched annotation sets. This step is the same as Step 2B(vi) above (**Fig. 5c**).

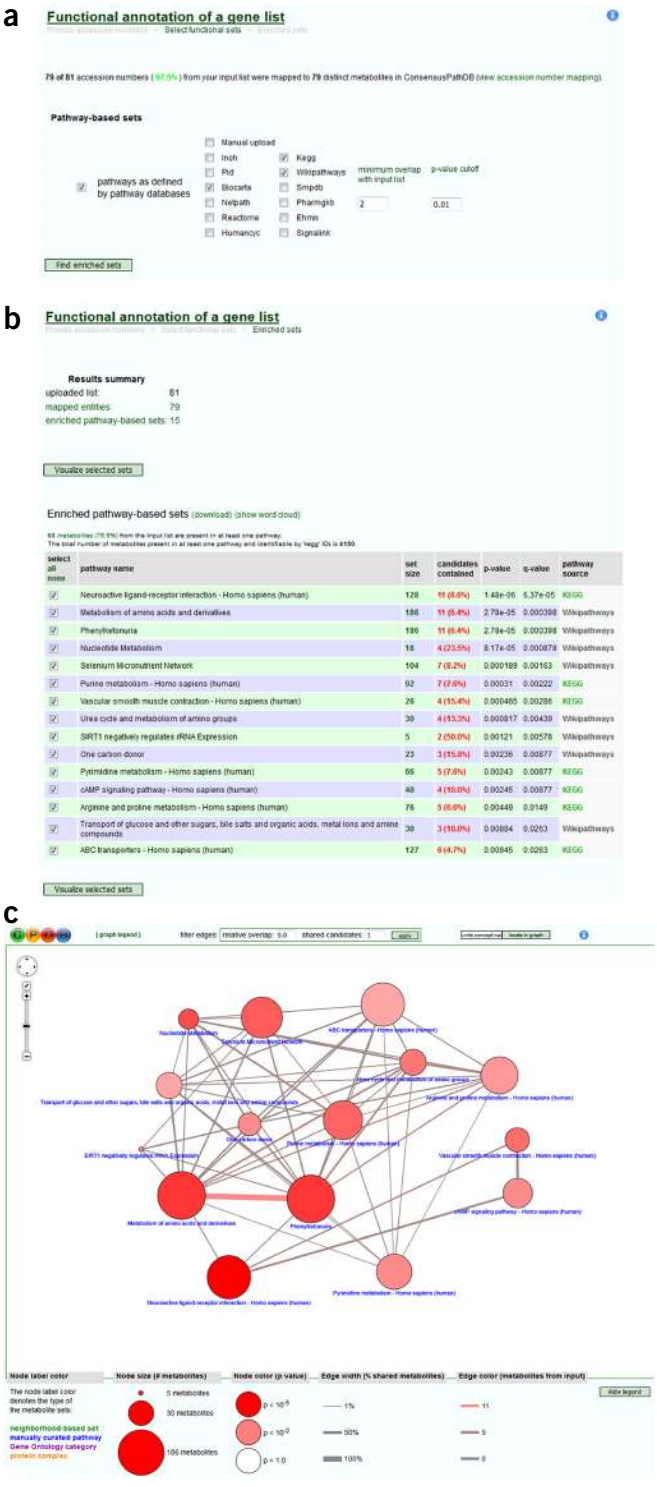

Figure 5 | Over-representation analysis with list of metabolites in ConsensusPathDB. (**a**) Selection of databases and parameters. (**b**) Results page showing the top-enriched pathways. (**c**) Visualization of annotation sets and their overlap.

(D) Induced network approach ● **TIMING 3-15 min**

(i) Submit seed genes. ConsensusPathDB allows the generation of a network that connects as many members of an input gene list (seed genes) as possible with intermediate nodes using the induced network graph algorithm. This is done by

clicking on 'gene set analysis' and 'induced network modules' (**Fig. 6a**). There are two options for insertion: copy the list of genes/proteins into the box or upload a corresponding text file (.txt format). As an example in this protocol, the gene list from **Supplementary Data 3** can be used.

(ii) Specify the identifier type. This is the same as Step 2B(iii) above. Select the appropriate identifiers for your list and hit 'Proceed'. For the example gene list used here, select 'gene symbol (HGNC symbol)'.

▲ **CRITICAL STEP** The identifier type must match the names in the submitted list. Otherwise, the system reports that no identifier could be mapped.

? TROUBLESHOOTING

(iii) Select interactions and database sources. At the top of the next page, there is a summary of the identifier mapping from the imported list to internal identifiers. To select the content to be included, the user has several options (**Fig. 6b**). After selecting options, click 'Find induced modules'.

Option	Description
Specify interactions to be included in the analysis	• Protein interactions with quality score (low, medium, high confidence according to the traffic signal criterion (cf. Step 2A(iii) above)
	• Genetic interactions
	• Biochemical reactions
	• Gene regulatory interactions
	• Drug–target interactions
Specify database sources to be included in the analysis	• Seed nodes are displayed either as seed node identifiers or as default display names
Specify display settings	• Show nonconnected seed nodes or not
	• Intermediate nodes: display or leave out. If this box is deselected, the created network will be based solely on interactions between seed nodes
	• Seed nodes are displayed either as seed node identifiers or as default display names
	• Show nonconnected seed nodes or not

▲ **CRITICAL STEP** The computation time, as well as the graph visualization, is heavily dependent on the number of nodes and edges. Thus, we recommend using only a small to medium number of genes as seed nodes (<200) or, for larger lists, to restrict the interactions to specific types—for example, gene regulatory interactions. Running a large list with all interactions selected might cause an error.

? TROUBLESHOOTING

(iv) Explore and modify the resulting network module. After the network analysis is completed, the user is directed to the visualization page (**Fig. 6c**). Biomolecules (genes, proteins and compounds) are depicted as squares, and interactions are depicted as lines connecting the interaction partners. Seed genes are labeled in black, whereas intermediate nodes are labeled in magenta. The colors of the lines and rectangles represent the interaction types and biomolecule classes. The user can customize the layout of the network directly on the webpage by moving and hiding nodes. This is achieved by clicking on and dragging the respective nodes and edges. The menu bar at the top of the page offers the following options:

Option	Description
Graph legend button	Click on the 'graph legend' button in the menu bar on top of the page for a detailed description
Slide bar	Using the slide bar, the user can control the threshold for the intermediate nodes: a more stringent threshold will reduce the number of nodes, whereas a lower threshold will increase this number
Export button	Using the 'export' button, the final network module can be downloaded in a text-based format, which can be imported into network visualization software such as Cytoscape[11]
Overlay values button	The 'overlay values' button in the menu bar allows uploading of a file with numerical values for each gene from additional experimental data; these will be displayed by color-coding of the nodes

▲ **CRITICAL STEP** The ConsensusPathDB offers basic visualization features. If the user wants to change the graph layout, we recommend exporting the graph as a .sif file and modifying the layout with appropriate tools—for example, Cytoscape[28].

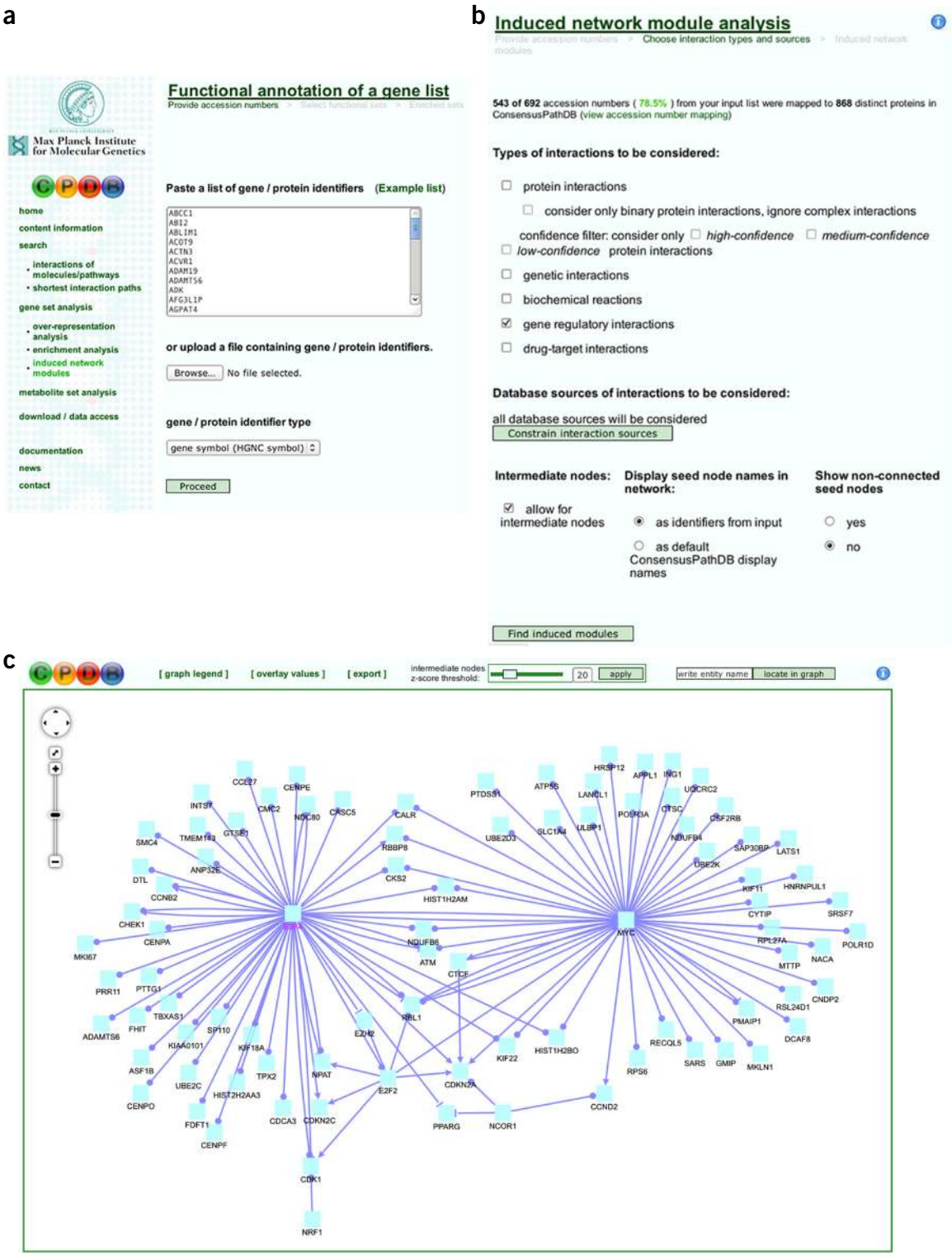

Figure 6 | Induced network module analysis in ConsensusPathDB. (**a**) List of seed nodes inserted by the user. (**b**) Setting of parameters that determine the underlying interaction graph. (**c**) Visualization of results. Seed node labels are shown in black, and intermediate node label, here E2F4, is shown in magenta.

(E) Enrichment analysis with high-throughput experiments ● TIMING 3–15 min

(i) Submit experimental data. In this section, we describe the enrichment analysis with respect to gene expression. Analysis of quantitative protein or metabolite data can be carried out analogously. Enrichment analysis is activated by clicking on 'gene set analysis' and then on 'enrichment analysis'. The procedure is similar to that for the over-representation

analysis described in Step 2B(i–vi), except that in this case the user's genes/proteins of interest have numerical values associated with them (**Fig. 7a**). To insert data, the user has two options: copy the list of genes in the box or upload a corresponding text file (in tab-delimited .txt format). As an example in this protocol, the list from

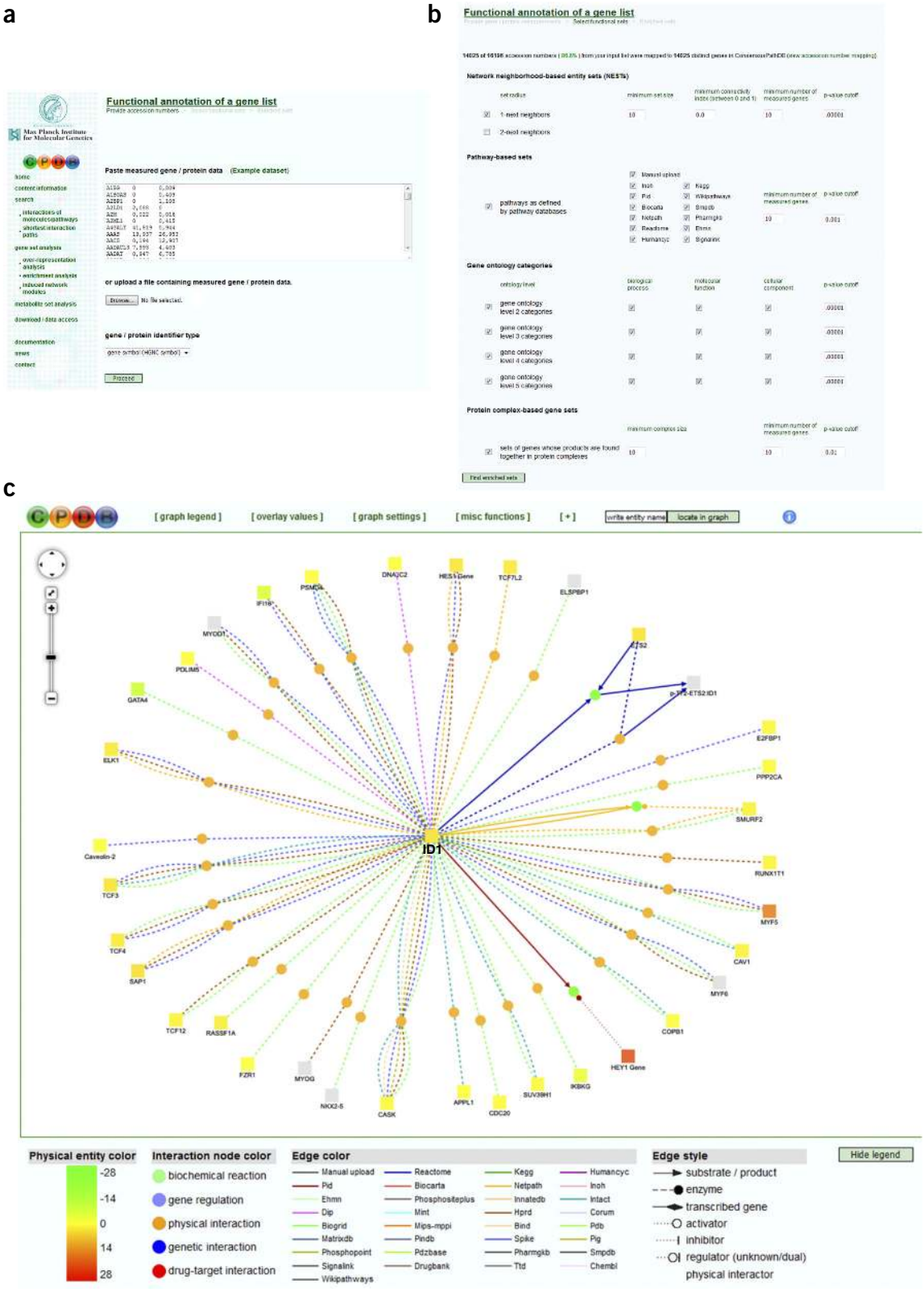

Figure 7 | Enrichment analysis with RNA-seq data in ConsensusPathDB. (**a**) Data input as a three-column file holding the RPKM values from two conditions per gene. (**b**) Specification of annotation sets and parameters. (**c**) Visualization of a significantly enriched NEST with ID1 as center protein with overlaid gene expression data (logFC).

Supplementary Data 4 can be used. ConsensusPathDB assumes that data come from case–control studies that compare two experimental conditions—for example, disease versus control, treated versus untreated, time point 1 versus time point 2 and so on. In both cases, the user thus has two options for submitting experimental data:

Option	Description
One value per gene/protein	In this case, the value will be interpreted as a (log-)fold change of case versus control expression; this setting is recommended if the user wants to overlay expression data on the networks, because the color-coding then allows better interpretability
Two values per gene/protein	In this case, values will be interpreted as expression values for case and control conditions, respectively

(ii) Specify the identifier type. This is the same as Step 2B(iii) above.
▲ **CRITICAL STEP** The identifier type must match the names in the submitted list. Otherwise, the system reports that no identifier could be mapped.
? TROUBLESHOOTING

(iii) Select annotation sets for genes/proteins. This is the same as Step 2B(iv) above (**Fig. 7b**). Finally, click on 'Find enriched sets'.
▲ **CRITICAL STEP** The NEST analysis can take up to several minutes, depending on the settings. We recommend using NEST radius equal to two only exceptionally. Note that if the response time is too long, the session expires and the results are lost. Furthermore, to ensure proper execution, the NEST analysis can be done separately from the other analyses by repeating Step 2E(i–iii) multiple times with different annotation sets.
? TROUBLESHOOTING

(iv) Inspect enrichment analysis results. This is the same as Step 2B(v) above.

(v) Visualize enriched annotation sets. This is the same as Step 2B(vi) above. In addition to the visualization of the annotation sets, for example, significant NESTs can be further inspected. For example, experimental data can be overlaid to significant NESTs, as is shown in **Figure 7c** by performing Step 2A(iv).

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 3**.

TABLE 3 | Troubleshooting table.

Step	Problem	Possible reason	Solution
1	No matching entity was found	User chose the wrong version of the database (e.g., submitted mouse identifiers to the human version of the database)	Go to the home page and click on the correct organism at the center top panel; the active version of the ConsensusPathDB is shown underlined in red
2A(i)	No matching entity was found	User set wrong settings (e.g., entry is not a name but an accession number or vice versa)	Check your settings; check whether your ID is supported by clicking on 'list of valid types'; otherwise, convert your IDs to a supported type
2A(iii)	Network neighborhood is not generated	The size of the network is too large, and the browser fails to show it	Restrict the number of interactions for visualization (recommended size <500)
2A(iv)	Overlaying experimental data does not work	The identifier type in the data list does not match the identifier type used for generating the network	Modify the data file accordingly
2B(iii), 2C(iii), 2D(ii), 2E(ii)	No identifier could be mapped	The identifier types in the submitted list and the specified identifier type do not match	Redo submission with matching identifier types; also cf. Step 1
2B(iv), 2E(iii)	No enrichment results are displayed; session expires	The number of annotation sets is too large; this is more of a problem for Step 2E(iii) (enrichment analysis) when using whole-experimental data	We recommend using NEST radius equal to two only exceptionally because it generates very large annotation sets. Perform Step 2E(iii) in an iterative way—i.e., activate only NEST annotations first, and then store the results, return to the page and activate the other annotation sets

(continued)

TABLE 3 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
2D(iii)	Induced network module graph is not generated	The number of seed nodes is too high Too many interactions are selected The session expired because it exceeded allowable run time	Restrict either the number of seed nodes or the number of interaction types
	Induced network module graph shows disconnected subsets of seed nodes	The user did not allow for showing intermediate nodes	Tick the corresponding check box

● TIMING

The time required to execute the above protocol is strongly dependent on the size of the analyzed data set, the load generated from the number of users on the local servers and in general on the network traffic.

Analysis steps for the different use cases of this protocol were performed with a standard Windows 7 PC with a single CPU (3.6 GHz) and 8 GB of memory, and had the following time requirements:

Step 1, selection of the organism of interest: <1 min

Step 2A, identification of network neighborhoods of single biomolecules: 1–10 min

Step 2B, over-representation analysis of a batch of genes/proteins: 1–10 min

Step 2C, over-representation analysis of a batch of metabolites: 1–10 min

Step 2D, induced network approach: 3–15 min

Step 2E, enrichment analysis with high-throughput experiments: 3–15 min

It should be noted that the achieved performance refers solely to the computation time. In between the different steps, the user has multiple options for inspecting results and for literature review of the identified interactions and proteins, as well as for modifications of visualization features. Thus, full analysis time is heavily dependent on the amount of user interaction.

ANTICIPATED RESULTS

Here we discuss plausibility of the results achieved with the protocol for the different analysis paths.

Generation of complex interaction networks for biomarkers (Step 2A)

As a use case, we have inspected interactions of *EGFR*. This is a well-investigated gene that is important for tumor progression and is a primary target for several therapies. *EGFR* has 4,670 interactions, of which 2,095 are actually distinct. ConsensusPathDB discovers diverse interactions for this gene, mostly with proteins such as SHC1[48] and drugs such as cetuximab[49]. When inspecting the individual interactions (**Fig. 3c**), the user can review further annotations and literature references. It should be noted that ConsensusPathDB's identifier matching is very important because in many publications this gene is described with aliases—e.g., *HER1* or *ERBB1*. The traffic light score for the individual interactions typically relates to the number of references found. For example, although the interaction between *EGFR* and *SHC1* has 56 references and a score of 0.9999 (green), the interaction of *EGFR* with *DOK6* has only one reference, corresponding to a score of 0.0067 (red). On the other hand, such interactions could be also of particular interest, as they might refer to rather novel findings.

Characterization of lists of genes/proteins with network-based information (Step 2B)

As a use case, we have uploaded the list of 18 driver mutation genes (**Supplementary Data 1**) that have been identified by the Cancer Genome Atlas Network to be frequently mutated in lung adenocarcinomas[1]. Not surprisingly, on inspecting the pathway results, gene sets involving cancer pathways are found to be most significantly enriched, the top three being those for glioblastoma, melanoma and lung cancer (**Fig. 4c**). Looking at GO terms reveals that many genes from the input list (12 out of 18) are involved in cell death, the evasion of which is a hallmark of cancer[50]. In addition to pathways and GO terms, which can also often be found with other online tools, ConsensusPathDB provides results for NESTs and protein complexes. This enables the user to find other genes/proteins (i.e., those not from the input list) in the form of NEST centers or members of protein complexes, which themselves may not have been regarded as interesting but can be potential targets or biomarkers as well because of their molecular connectivity to genes of interest. For example, in the original publication[1], the authors analyzed the effects of somatic mutations in 230 patients on specific signaling pathways and discovered MAPK, mTOR and AMPK signaling as major targets. Our analysis based solely on the 18 driver mutations can confirm this finding. MTOR has interactions with 4 of the driver genes (*RBM10*, *PIK3CA*, *EGFR* and *TP53*), which gives a significant result for the respective NEST ($Q = 8.49 \times 10^{-4}$). Furthermore, all pathways found to be substantially affected by somatic mutations in the

original study are also identified by the ConsensusPathDB over-representation analysis (AMPK signaling, $Q = 2.97 \times 10^{-4}$; MAPK signaling, $Q = 2.25 \times 10^{-5}$; mTOR signaling, $Q = 2.54 \times 10^{-4}$.

Characterization of lists of metabolites with pathway information (Step 2C)

As a use case for metabolites, we have investigated the list of 130 uremic toxins provided by the EUTox work group, which is an international consortium for research on chronic kidney disease. As the metabolite names are given in nonstandardized format, we have manually converted them to KEGG identifiers in order to be able to map them to pathways (**Supplementary Data 2**). However, not all metabolites could be mapped, resulting in 79 unique KEGG identifiers. Uremic toxins, by definition, have an increased concentration in the blood due to renal insufficiency, and they can cause severe organ damage, atherosclerosis and vascular remodeling[34]. As a major result of the pathway over-representation analysis (**Fig. 5b**), pathways appear that are associated with transport processes. This result is in line with previous reports showing that high levels of uremic toxins can compete with transporter molecules for elimination and distribution of drugs and other compounds[51]. Furthermore, cardiovascular pathways are significantly enriched, emphasizing that chronic kidney disease is a major risk factor for cardiovascular disease. In particular, we find that the pathway 'vascular smooth muscle contraction' is affected by the metabolites, which confirms knowledge from the literature. Recently, it has been shown that a particular uremic toxin, *p*-Cresyl sulfate, causes vascular smooth muscle cell damage through oxidative stress[52].

Generation of interaction network module maps from lists of genes/proteins (Step 2D)

For the induced network module computation, we repeated the analysis from a published application of the ConsensusPathDB (**Supplementary Data 3**). The authors performed genome-wide histone modification analysis of H3K4me2 with ChIP-seq in T cells isolated from asthmatic and normal individuals in order to identify disease-specific enhancers. They identified *MYC*, *E2F2* and *E2F4* as major regulators for H3K4me2 methylation target genes (comparing T_H2 cells with naive T cells from asthmatic individuals)[17]. They performed analysis Step 2D(i–iv) using an input list of 691 different targets. In Step 2D(iii), they restricted analysis to gene regulatory interactions, which makes the analysis of such a large list computationally feasible. Next, the three master regulators appeared as central hubs of the induced network (**Fig. 6c**). Interestingly, *E2F4* was not a seed node—i.e., not among the target genes—and its role as a master regulator was inferred from the background interaction network of ConsensusPathDB.

Enrichment analysis with lists of genes/proteins (Step 2E)

As a use case, we analyzed whole-genome single-cell RNA-seq data derived from different stages of human development[2]. These data were provided as reads per kilobase of transcript per million mapped reads (RPKM)-normalized values for each gene across the different cells. For each developmental stage, we averaged the corresponding biological replicates (i.e., single cells) and used human epiblast (EPI) samples, as well as embryonic stem cells (hESCs) derived from these epiblasts, as a case–control study. The EPI is a layer of cells in the late blastocyst that differentiates from the inner cell mass, usually around day 9 of human embryonic development, and builds the precursor of all embryonic tissues[53]. One goal of the original study was to identify differentially expressed genes between EPI and hESC cells.

Data were uploaded in .txt file format (Step 2E(i)) as a three-column file containing the gene symbols and the two RPKM values per gene. Before data analysis, we excluded genes that had zero RPKM values in both conditions. Enrichment analysis found huge differences between EPI and hESC cells, in line with the result of the original work in which the authors identified 1,498 genes differentially expressed between the two stages. Among these genes were pluripotency-related genes, as well as Wnt signaling genes, underlying the prominent role of this pathway for development. Consistently, we find several Wnt-related signatures among the top enriched categories such as 'signaling by Wnt' ($Q = 6.69 \times 10^{-5}$), 'TCF=dependent signaling in response to WNT' ($Q = 2.82 \times 10^{-5}$) and 'Wnt signaling pathway and pluripotency' ($Q = 5.36 \times 10^{-3}$). *ID1*, a gene associated with pluripotency, was found to be significantly upregulated in hESC cells as compared with EPI cells in the original publication, and the NEST for *ID1* was found to be significantly enriched by ConsensusPathDB analysis. Inspecting its network neighborhood (Step 2A(iv)) also reveals that most of the interaction partners of *ID1* were upregulated (**Fig. 7c**).

In summary, the results generated with the ConsensusPathDB protocol show that the tool is a useful complement for genome analysis and that it delivers plausible functional and network-based interpretation for heterogeneous applications.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS We are grateful to all scientists who provided annotation of the original molecular interaction data and are allowing automated access to their databases. Integration of interaction data could be achieved only because the original data were provided in an excellently documented way.

This work was financed in part by the European Commission under its 7th Framework Programme (HeCaToS 602156 to R.H.) and the Max Planck Society.

AUTHOR CONTRIBUTIONS R.H. and A.K. conceived ConsensusPathDB, designed the protocol and wrote the manuscript; A.K. developed ConsensusPathDB; C.H. and M.L. conducted the procedure, performed data analysis and contributed to the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
2. Yan, L. *et al.* Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
3. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
4. Khatri, P., Sirota, M. & Butte, A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comp. Biol.* **8**, e1002375 (2012).
5. Maciejewski, H. Gene set analysis methods: statistical models and methodological differences. *Brief Bioinf.* **15**, 504–518 (2014).
6. Taylor, I.W. *et al.* Dynamic modularity on protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27**, 199–204 (2009).
7. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.* **37**, D623–D628 (2009).
8. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
9. Vidal, M., Cusick, M.E. & Barabasi, A.L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
10. Stumpf, M.P.H. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959–6964 (2008).
11. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34**, D504–D506 (2006).
12. Hoehe, M.R. *et al.* Multiple haplotype-resolved genomes reveal population patterns of gene and protein diplotypes. *Nat. Commun.* **5**, 5569 (2014).
13. Grossmann, A. *et al.* Phospho-tyrosine dependent protein-protein interaction network. *Mol. Syst. Biol.* **11**, 794 (2015).
14. Li, A.H. *et al.* Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat. Genet.* **47**, 640–642 (2015).
15. Timme, S. *et al.* STAT3 expression, activity and functional consequences of STAT3 inhibition in esophageal squamous cell carcinomas and Barrett's adenocarcinomas. *Oncogene* **33**, 3256–3266 (2014).
16. Sun, C. *et al.* High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* **48**, 323–330 (2016).
17. Seumois, G. *et al.* Epigenomic analysis of primary human T cells reveals enhancers associated with T_H2 memory cell differentiation and asthma susceptibility. *Nat. Immunol.* **15**, 777–788 (2014).
18. Kallio, M.A. *et al.* Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507 (2011).
19. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat. Methods* **9**, 1069–1076 (2012).
20. Pentchev, K., Ono, K., Herwig, R., Ideker, T. & Kamburov, A. Evidence mining and novelty assessment of protein-protein interactions with the ConsensusPathDB plugin for Cytoscape. *Bioinformatics* **26**, 2796–2797 (2010).
21. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
22. Yildirimman, Y. *et al.* Human embryonic stem cell derived hepatocyte-like cells as a tool for *in vitro* hazard assessment of chemical carcinogenicity. *Tox Sci.* **124**, 278–290 (2011).
23. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocol.* **4**, 44–56 (2009).
24. Krämer, A., Green, J., Pollard, J. Jr. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30**, 523–530 (2014).
25. Chen, E.Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.* **14**, 128 (2014).
26. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
27. Xia, J. & Wishart, D.S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protocol.* **6**, 743–760 (2011).
28. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protocol.* **2**, 2366–2382 (2007).
29. Berger, S.I., Posner, J.M. & Ma'ayan, A. Genes2Networks: connecting lists of gene symbols using mammalian protein interaction databases. *BMC Bioinf.* **8**, 372 (2007).
30. Liberzon, A. *et al.* Molecular signature database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
31. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
32. Cerami, E.G. *et al.* Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
33. Oda, K., Matsuoka, Y., Funahashi, A. & Kitano, H. A comprehensive pathway map of epidermal growth factor receptor signalling. *Mol. Syst. Biol.* **1**, 2005.0010 (2005).
34. Vanholder, R. *et al.* Review on uremic toxins: classification, concentration, and interindividual variability. *Kidney Int.* **63**, 1934–1943 (2003).
35. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
36. Fabregat, A. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016).
37. Kutmon, M. *et al.* Wikipathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **44**, D488–D494 (2016).
38. Kamburov, A., Grossmann, A., Herwig, R. & Stelzl, U. Cluster-based assessment of protein-protein interaction confidence. *BMC Bioinf.* **13**, 262 (2012).
39. Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA* **100**, 4372–4376 (2003).
40. Kuchaiev, O., Rasajski, M., Higham, D.J. & Przulj, N. Geometric de-noising of protein-protein interaction networks. *PLoS Comp. Biol.* **5**, e1000454 (2009).
41. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41**, D793–D800 (2013).
42. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
43. Kamburov, A., Stelzl, U. & Herwig, R. IntScore: a web tool for confidence scoring of biological interactions. *Nucleic Acids Res.* **40**, W140–W146 (2012).
44. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57**, 289–300 (1995).
45. Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.* **39**, D712–D717 (2011).
46. Lehmann, E. *Nonparametrics: Statistical Methods Based on Ranks* (San Francisco, California: Holden-Day, 1975).
47. Adjaye, J. *et al.* Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and trophectoderm cells. *Stem Cells* **23**, 1514–1525 (2005).
48. Zheng, Y. *et al.* Temporal regulation of EGF signaling networks by the scaffold protein Shc1. *Nature* **499**, 166–171 (2012).
49. Li, S. *et al.* Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. *Cancer Cell* **7**, 301–311 (2005).
50. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
51. Reyes, M. & Benet, L.Z. Effects of uremic toxins and metabolism of different biopharmaceutics drug disposition classification system xenobiotics. *J. Pharm. Sci.* **100**, 3831–3842 (2011).
52. Watanabe, H. *et al.* p-Cresyl sulfate, a uremic toxin, causes vascular endothelial and smooth muscle cell damages by inducing oxidative stress. *Pharmacol. Res. Perspect.* **3**, e00092 (2015).
53. Niakan, K.K., Han, J., Pedersen, R.A., Simon, C. & Pera, R.A.R. Human pre-implantation embryo development. *Development* **139**, 829–841 (2012).