Supplementary Material

Gene expression profile dataset

Gene expression data were downloaded from Array Express [1]. This is a public repository of gene expression profiles (GEPs) described in literature. GEPs are logically organized into experiments. An experiment is a collection of GEPs (usually performed in a single laboratory) together with their META-DATA to trace information such as the applied experimental protocol, type of sample (cell type or tissue), and all the other information required by the MIAME standard [2].

We collected 591 experiments in human for a total of 20,255 GEPs performed in human samples by hybridisation of total RNA to HG-U133A Affymetrix microarrays, as well as, 614 experiments for a total of 8,895 GEPs derived from mouse samples by hybridisation of total RNA to Mouse430A_2 Affymetrix microarrays. In Supplementary Table 1 we report the complete list of the experiments used. Since ArrayExpress stores for each experiment the normalized GEPs within the experiment, but not necessarily the unprocessed data (i.e. CEL files). Since our method can merge experiments processed with different algorithms (MAS5, RMA, GCRMA), we decided to use for our subsequent analysis normalized expression data only, thus maximising the number of 'usable' experiments.

Computation of pair-wise Mutual Information

GEPs derived from ArrayExpress are normalized within experiments, but not across experiments. Standard normalization techniques are difficult to apply to such large datasets and may not always be performed due to the lack of unprocessed data (CEL files). To overcome these limitations, we proceeded as follows: the expression values for each probe-set across the microarrays within one experiment are ranked and then discretized into a predetermined number of bins (three). Specifically, the three bins, each containing an equal number of values, are determined using the three non-parametric quantiles of the ranked normalized expression values as cut points. Each expression value is then replaced by an integer corresponding to the bin it falls into. Changing the number of bins does not considerably affects the results as shown in Supplementary Figure 4, where we compared the algorithm performance using 3, 5 or 7 bins and 100 or 1000, computer simulated GEPs obtained as described in [3]. Since the results are not changing significantly, and the computational complexity of the algorithm depends on the number of bins, we decided to use three bins as a good balance between these factors.

We then computed pair-wise Mutual Information (MI) for all the pairs of probe-sets present in the microarray model HG-U133A. MI is a pseudo distance between probability distributions; it measures the amount of information two random variables share. MI has been widely applied to reverse-engineer gene regulatory networks [4, 5]. Genes can be seen as random variables and their profiles as a random process. Once gene expression profiles were properly discretized into bins, we computed MI for each pair of probe-sets.

We considered two discrete random variables I and J assuming values in the set $\{1, ..., 3\}$ describing the discretized expression values of two probe-sets I and J. In this context, MI can be defined as:

$$MI_{IJ} = \sum_{i=1}^{3} \sum_{j=1}^{3} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}, \tag{1}$$

where π_{ij} represents the joint probability $P(I = i, J = j)$ with $(i, j) \in \{1, 2, 3\} \times \{1, 2, 3\}$ and $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are respectively their marginal probabilities $P(I = i)$ and $P(J = j)$.

In order to estimate the joint probability π_{ij}, we used a simple frequentist approach: let n^k_{ij} be the counts of the outcomes $(I = i, J = j)$ across the n^k GEPs of experiment k, then the frequency $\hat{\pi}_{ij}$ can be estimated jointly from the all the K experiments:

$$\hat{\pi}_{ij} = \frac{\sum_{k=1}^{K} n_{ij}^{k}}{\sum_{k=1}^{K} n^{k}} = \frac{n_{ij}}{n} \tag{2}$$

This leads to a point estimate of MI equal to

$$\widehat{MI}_{IJ} = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}}. \tag{3}$$

The computational complexity of our algorithm is $o(N^2 \cdot n)$ where N is the number of probe-sets and n is the total number of GEPs. In the human network N is 22,283 and n is 20,255, whereas in the mouse network N is 45,101 and n is 8,895. Due to the high computational cost, we decided to implement a parallel version of the algorithm to reduce computational time. The parallel algorithm distributes the gene expression profiles among several computing processes. Each process gets N/p probes where p is the number of processes available and N is the total number of probes. The processors are named P_0 to P_{p-1} and logically organized in a topological ring where i follows j if $i > j$ and P_0 follows P_{p-1}.

At the beginning of the computation each process calculates the discretization of its own gene expression profiles and computes the mutual information for each pair of its N/p probes. At the first communication step each process sends its probe to the following process and computes the MI for all the pairs of probes where the first is the local probe while the second is the received probe. At the i^{th} communication step process P_j receives probes from process $P_{(j-i)mod(p)}$. The algorithm completes in $\lfloor \frac{p}{2} \rfloor$ communication steps. The parallel algorithm has been implemented in C using the MPI standard. The code is available on request. The program has been executed on 105 processors of an HP XC6000 Cluster with Itanium 2 biprocessors nodes and a Quadrics ELAN 4 network. The total time for the execution of the algorithm on the human network was about 8 hours, while for the mouse network it took about 4 hours.

In order to select a significant threshold for the human and mouse networks, we fitted a Gamma distribution to the values of the MI across all the probes' pairs in human (mouse) using Maximum Likelihood estimation. The Gamma distribution has been shown to well describe the distribution of MI values under the null hypothesis of statistical independence among the two variables[6, 7]. We thus could assign a p-value to the MI of each gene pair and retained only those MI with a p-value < 0.01. This correspond to a MI of 0.04 for human and 0.025 for mouse.

We could have further pruned the networks by using one of the recently proposed schemes, such as the CLEAR method [8] and the Data Processing Inequality (DPI) method introduced by [5], but we decided against it, since we wanted to keep as many interactions as possible to have a broader overview of a gene function and regulation. In this work, we were not interested in identifying 'direct' interactions as done by [9], but we focused on the identification of "co-regulated" genes, i.e. both direct and indirect interactions.

Comparison with off-the-shelf method

We applied a state-of-the-art reverse engineering algorithm, ARACNe [5], on simulated in-silico gene expression data generated in [3]. Specifically, this in-silico dataset consists of 20 networks. For each network, 1000 gene expression profiles are simulated in-silico, and expression values are considered comparable across the 1000 expression profiles. In order to make this in silico dataset comparable to real gene expression profiles coming from different experiments, and therefore with values not comparable across experiments, we proceeded as follows: for each gene network, we subdivided the set of the 1000 in-silico expression profiles into 10 different subsets of 100 expression profiles each. We then applied ARACNe on each subset and the union of the gene network inferred was considered in further analysis. On the same dataset, we also applied our new reverse-engineering approach described previously. We observed that the average precision of across the 20 networks (True Positives/ (True Positive plus False Positive)) is 47%, whereas ARACNe reaches a precision of 21%. For comparison, when ARACNE is run on all the 1000 in-silico expression profiles together, its precision is 66% [3]. This means that ARACNE is very good at inferring gene networks but only if gene expression profiles are coming from homogeneous samples, where standard normalization procedures can be applied.

Construction of the Golden Standard Interactome

In order to validate the biological relevance of the predicted interactions from our MI approach, we built a golden standard (GoS) interaction network from the following interaction databases:
Reactome [10] is a curated knowledge-base of biological pathways; It represents a resource of core pathways and reactions in human biology. It collects a total of 32,821 interactions.
Cipher [11] is a tool to predict disease-genes on the bases of a computational framework that integrates human protein-protein interactions, disease phenotype similarities, and known gene-phonotype associations. Authors provide a protein interaction network assembled from the HPDR, BIND, MINT and OPHID PP interaction databases. It accounts for 40,649 interactions.
Tissue Specific Protein Interactions [12] is a global human interaction network obtained by integrating data from 21 different sources to define a network of a total of 67,200 physical interactions.

Results of the comparison between the inferred network and the GoS interactome are presented in Supplementary Figure 1a.

Relationship between gene degree, gene expression level and protein disorder

In order to relate gene degree to the average expression level of a gene, we proceeded as follows: since GEPs are normalized within experiments, and not comparable across experiments, we could not use the whole dataset to estimate the average expression of each gene. Therefore, we used 120 human GEPs (GSE5720) and 182 mouse GEPs (GSE10246) from GEO [13] measuring the expression of genes across normal tissues.

In each species, GEPs were normalized together by applying MAS5 algorithm [14]. Probe-set average expression levels were then compared with probe-set degrees, i.e. the number of connections of a probe-set in the network. We then divided probe-sets into groups with the same (or similar) degree, taking care that each group contained the same number of probe-sets (500 probe-sets each in human, and 1000 each in mouse). We then computed the average expression level and the average degree within each group, as shown in Supplementary Figure 2a,b.

In order to relate gene degree to the average protein disorder of a gene, we proceeded as before, but this time we used GlobPlot [15] to compute the protein disorder for each gene corresponding to each probe-set. Results are shown in Supplementary Figure 3c,d.

Identification of Network Communities & Rich-clubs

We applied a classic hierarchical clustering algorithm with the average linkage algorithm on the matrix describing the network using as a distance the "*Jaccard*" metrics. This is defined as the ratio between the number of common interactions between two genes (i.e. two rows or two columns being the matrix symmetric) divided by the total number of interactions. The dendrogram, produced by the average linkage algorithm, was cut in order to maximize the number of clusters with more than 4 nodes. At the end of this procedure, we identified 393 (865) communities in the human (mouse) network with more than 4 nodes.

We then identified if these communities were enriched for genes sharing a common biological function. To this end we applied Gene Ontology Enrichment Analysis (GOEA) on the list of genes in each community. GOEA is a commonly used technique that allows the identification of statistically over-represented

Gene Ontology terms in a set of genes. Suppose to have a set of N of genes, m of which are annotated as associated to a Gene Ontology term of interest. Suppose to draw a subset of n genes from the complete list of N genes, then the probability of obtaining k genes all sharing the same Gene Ontology term of interest follows an hypergeometric probability distribution:

$$Pr(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{4}$$

from this, we can compute the cumulative distribution and hence the significance, or p-value, of the draw; i.e. the probability of having at least k genes sharing the same GO term of interest.

In order to identify rich-clubs we first defined a distance between two communities A and B: the Interaction Strength (IS_{AB}). Defining n_{AB} as the number of edges in the human (mouse) network across the K_A genes in community A and K_B genes in community B, then:

$$IS_{AB} = log(\frac{n_{AB}}{K_A \times K_B \times f}) \tag{5}$$

where f is the average frequency of edges across all the genes in the human (mouse) network.

We then constructed a community-wise network by creating an adjacency matrix whose element in row i and column j is the IS_{ij} between community i and community j, if the $IS_{ij} > 0$.

We the used this community-wise adjacency matrix to identify rich-clubs, i.e. communities of communities. To this end, we applied a novel message-passing clustering algorithm [16] which is able to return the number of clusters without any user-speficied parameter, using as inter-community distances the ISs. We thus obtained 58 human (227 mouse) clusters of communities, i.e. rich-clubs (as defined in network theory). For a complete list of genes within each community refer to Supplementary Table 6.

Guilty-by-association analysis

In order to predict gene function and/or localization from the human (mouse) network, we used a classic Gene Ontology Enrichment Analysis (GOEA) as follows: for each probe (i.e. gene) of the human (mouse) network, we selected the probes predicted to interact with it in the network (i.e. the gene's neighbors). For probes with more than 500 neighbors, only the 500 ones with the highest MI were retained; for nodes with less than 50 neighbors, we included also the gene's second neighbors (i.e. the neighbors of the neighbors) up to a maximum of 500.

Multiple probes may refer to the same gene and thus be highly co-regulated. To avoid bias in the GOEA computation, we removed from the neighbors, the probes associated to the same gene. GOEA was then performed on this subset of neighbors for each of the probe in the human network. In the testing phase, a prediction was claimed to be correct if the GO term that was enriched within a neighborhood was also the one associated to the probe itself according to the GO database [17].

Chromatin structure and gene expression

To investigate the relationship between physical contact of chromosomal loci and connections among genes in the network, we used a recent genome wide physical interaction map measured in 2 human cell lines (GM06690 and K562) via an innovative "HiC" chromosome capture technique [18]. Authors provide an intra-chromosomic [18] contact probability matrix at 100 Kilobases resolution, and a genome wide contact probability matrix at 1 Megabases resolution.

In [18], authors defined a correlation matrix C_p in which element (i, j) is the Pearson correlation between the i^{th} row and j^{th} column of M_p. This C_p matrix exhibits a strong 'plaid-pattern' as shown in Figure 4b.

In order to compare the physical contact probabilities of the chromosomal regions (C_p) to the human network, we first derived a "connection tendency matrix" (M_c) at 1 Mb resolution from the human network adjacency matrix. In the M_c matrix the element in position (i, j) reports the connection tendency between genes in the i^{th} Mb and genes in the j^{th} Mb. To generate the M_c matrix, we first subdivided the adjacency matrix by grouping together probes referring to genes which where within 1 Mb distance between each other, we then computed the IS as in Equation 5 genome-wide between the 1 Mb regions, with K_A equal to the number of genes in the A region (of 1 Mb), K_B the number of genes in the B region (of 1 Mb), n_{AB} the number of interactions between genes in region A and genes in region B, and f the average frequency of interactions across all the genes in the human network. We then derived a correlation matrix C_c as shown in Figure 4a, where the element in position (i, j) is the Pearson correlation between the i^{th} row and j^{th} column of the M_c matrix.

Chromosome-wise analysis and genome wide analysis was accomplished by computing the 2-dimensional Pearson correlation coefficient (PCC) between the C_p matrix, describing the physical contact probability, and the C_c matrix, describing the interaction probabilities among the genes in the human network. We computed the p-value following classic statistical theory for PCC. The p-value was then corrected following A Bonferroni False Discovery Rate (FDR) procedure. We deemed as significant those PCC with an $FDR < 0.05$.

Yeast-two-hybrid assays

The Yeast two Hybrid (Y2H) kit "ProQuest Two-Hybrid System" was obtained by Invitrogen. The kit includes the bait vector pDEST32 (containing the GAL4 transcription factor DNA binding domain, DBD), and the prey vector pDEST22 (containing the GAL4 transcription activation domain, AD) along with the strong positive interaction control of pEXP 32-Krev1/pEXP 22 RalGDS-wt pair, the weak positive interaction control pEXP 32-Krev1/pEXP 22 RalGDS-m1 pair, and the negative interaction control pEXP 32-Krev1/pEXP 22 RalGDS-m2. The "Ultimate ORF" of the genes of interest were purchased from INVITROGEN in order to generate prey and bait plasmids using the gateway technology. For each gene we created both the prey and the bait plasmid according to manufacturer instructions.

The Y2H recipient strain of the ProQuest Two-Hybrid System, S. cerevisiae MAV 203 has the genotypes of (MATα, leu2-3,112, trp1-901, his3Δ200, ade2-101, gal4Δ, gal80Δ, SPAL10::URA3, GAL1::lacZ, HIS3UAS GAL1::HIS3@LYS2, can1R, cyh2R). The culturing of MAV203 was performed in YPAD media using the standard protocol suggested by the manufacturer. Nutrient marker selective plates were made with a nitrogen base, a carbon source, and a "dropout" solution containing the appropriate amino acids (SIGMA). Synthetic medium lacking tryptophan, leucine, histidine, and uracil was used to select positive interaction. The protein-protein interaction assays were performed according to the instructions provided by Invitrogen.

Cell culture and transfection

For GRN functional assays, HeLa cells were cultured in DMEM supplemented with 10% FBS and treated for 96 hours in the presence of sucrose to a final concentration of 100mM with daily changes of medium. For immunofluorescence cells grown on glass coverlips were fixed with methanol for 10 minutes, washed with PBS, treated with 50 mM NH4Cl for 15 minutes, and permeabilized with PBS 0.1% Triton, blocked in blocking buffer (0.5% BSA, 50mM NH4Cl, 0.001% Triton in PBS pH 7.4), and incubated overnight at 4^{o}C with anti-LAMP2 antibody Santa Cruz and for one hour with Alexa-594 (Invitrogen). The levels of Granulin and Catepsin D (CTSD) were evaluated by Real-Time PCR (Roche). The amplification was performed using the following primers for *GAPDH*, "Fw: GAAGGTGAAGGTCGGAGTC and "Rev: GAAGATGGTGATGGGATTTC, for *GRN*, "Fw: TCCAGAGTAAGTGCCTCTCCA and "Rev: TCACCTCCATGTCACATTTCA, and for *CTSD*, "Fw: AACTGCTGGACATCGCTTGCT and "Rev: CATTCTTCACGTAGGTGCTGGA.

For *TFEB* over-expression, HeLa cells were maintained at 37^{o}C in a 5% CO2-humidified incubator, and cultured in DMEM (GIBCO BRL) supplemented with 10% heat-inactivated fetal bovine serum

(FBS) (Invitrogen), 1% L-glutammine and 1% antibiotic/antimycotic solution (GIBCO BRL). 500,000 HEK293 cells were transfected with 4μg of DNA expressing the transcriptional factor EB (*TFEB*) using lipofectamine transfection reagent (Invitrogen) following the manufacturer protocol. After 48 hours from transfection cells were collected, the mRNA extracted and the levels of Granulin, Catepsin D (*CTSD*) and *TFEB* were evaluated by Real-Time PCR (Roche). The amplification of *TFEB* was performed using the following primers, "Fw: CCAGAAGCGAGAGCTCACAGAT and "Rev :TGTGATTGTCTTTCTTCT-GCCG.

For *GRN* over-expression 500,000 HeLa cells were transfected with 4μg of DNA expressing the human *GRN* (Invitrogen, MGC Full-length (IRAT) clone ID3457813) using lipofectamine transfection reagent (Invitrogen). As control we transfected 4μg of pCMV FLAG plasmid (Sigma-Aldrich) and 4μg of DNA expressing a green fluorescent protein (*GFP*).

After 48 hours from transfection cells were collected and analysed by immunofluorescence. Medium was concentrated on filters (Vivaspin Sartorius Stedim) and loaded on 10% SDS-PAGE. Transfer membranes were incubated with anti human granulin antibody (Invitrogen) at 1:50 dilution.

Identification of Binding sites in Granulin promoter region

We used the Position Weight Matrix of the Transcription Factor TFEB and run MATCH [19] to find its binding sites across -1000 +1000 base pairs flanking the UTR region of GRN. We found two binding sites on chromosome 17, plus strain, with location 42,422,444 to 42,422,457 and chromosome 17, plus strain, location 42,422,460 42,422,473 (Genome Browser, Human assembly Feb. 2009).

References

[1] Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., and Brazma, A. (January, 2007) ArrayExpress–a public database of microarray experiments and gene expression profiles.. *Nucleic Acids Res,* **35**(Database issue).

[2] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (December, 2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet,* **29**(4), 365–371.

[3] Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (February, 2007) How to infer gene networks from expression profiles.. *Molecular systems biology,* **3**.

[4] Butte, A. J. and Kohane, I. S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 418–429.

[5] Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics,* **7 Suppl 1**, S7.

[6] Hutter, M. (2004) Distribution of Mutual Information. *Advanced in Neuronal Information Processing Systems,* **18**, 339–406.

[7] Goebel, B., Dawy, Z., Hagenauer, J., and Mueller, J. (May, 2005) An approximation to the distribution of finite sample size mutual information estimates. *Communications, 2005. ICC 2005. 2005 IEEE International Conference on,* **2**, 1102–1106.

[8] Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (January, 2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology,* **5**(1), e8+.

[9] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (Apr, 2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet,* **37**(4), 382–390.

[10] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (January, 2005) Reactome: a knowledgebase of biological pathways.. *Nucleic Acids Res,* **33**(Database issue).

[11] Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (May, 2008) Network-based global inference of human disease genes. *Mol Syst Biol,* **4**.

[12] Bossi, A. and Lehner, B. (April, 2009) Tissue specificity and the human protein interaction network. *Mol Syst Biol,* **5**.

[13] Edgar, R., Domrachev, M., and Lash, A. E. (January, 2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.,* **30**(1), 207–210.

[14] Hubbell, E., Liu, W.-M., and Mei, R. (December, 2002) Robust estimators for expression analysis. *Bioinformatics,* **18**(12), 1585–1592.

[15] Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (July, 2003) GlobPlot: Exploring protein sequences for globularity and disorder.. *Nucleic acids research,* **31**(13), 3701–3708.

[16] Frey, B. J. J. and Dueck, D. (January, 2007) Clustering by Passing Messages Between Data Points.. *Science,* **315**.

[17] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (May, 2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.. *Nature genetics,* **25**(1), 25–29.

[18] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (October, 2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science,* **326**(5950), 289–293.

[19] Kel, A. E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (July, 2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences.. *Nucleic acids research,* **31**(13), 3576–3579.

Supplementary Figures

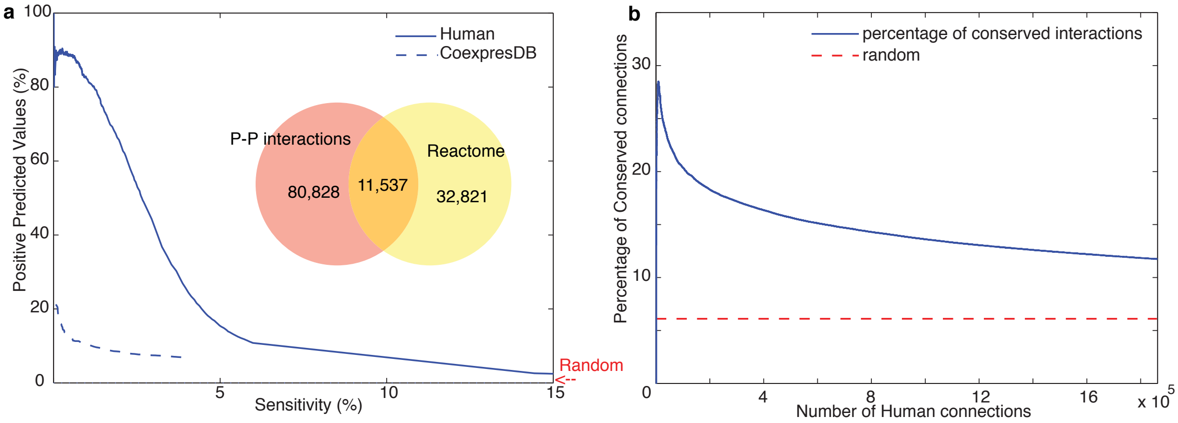

Figure 1: Validation of predicted connections and their conservation across species. (a) *In-silico* validation of the network compared with an experimental Golden Standard (GoS) interactome. The Venn diagram reports details about the composition of the GoS interactome. The inferred network (blue-line) significantly outperforms random performance or a correlation-based approach (CoexpresDB dashed line). (b) Percentage of connections conserved in the mouse. Connections are sorted according to their MI value. The curve peaks at 28%; the genes involved in these most conserved interactions are highly enriched for cell cycle phase $P = 1.0x10^{-15}$.

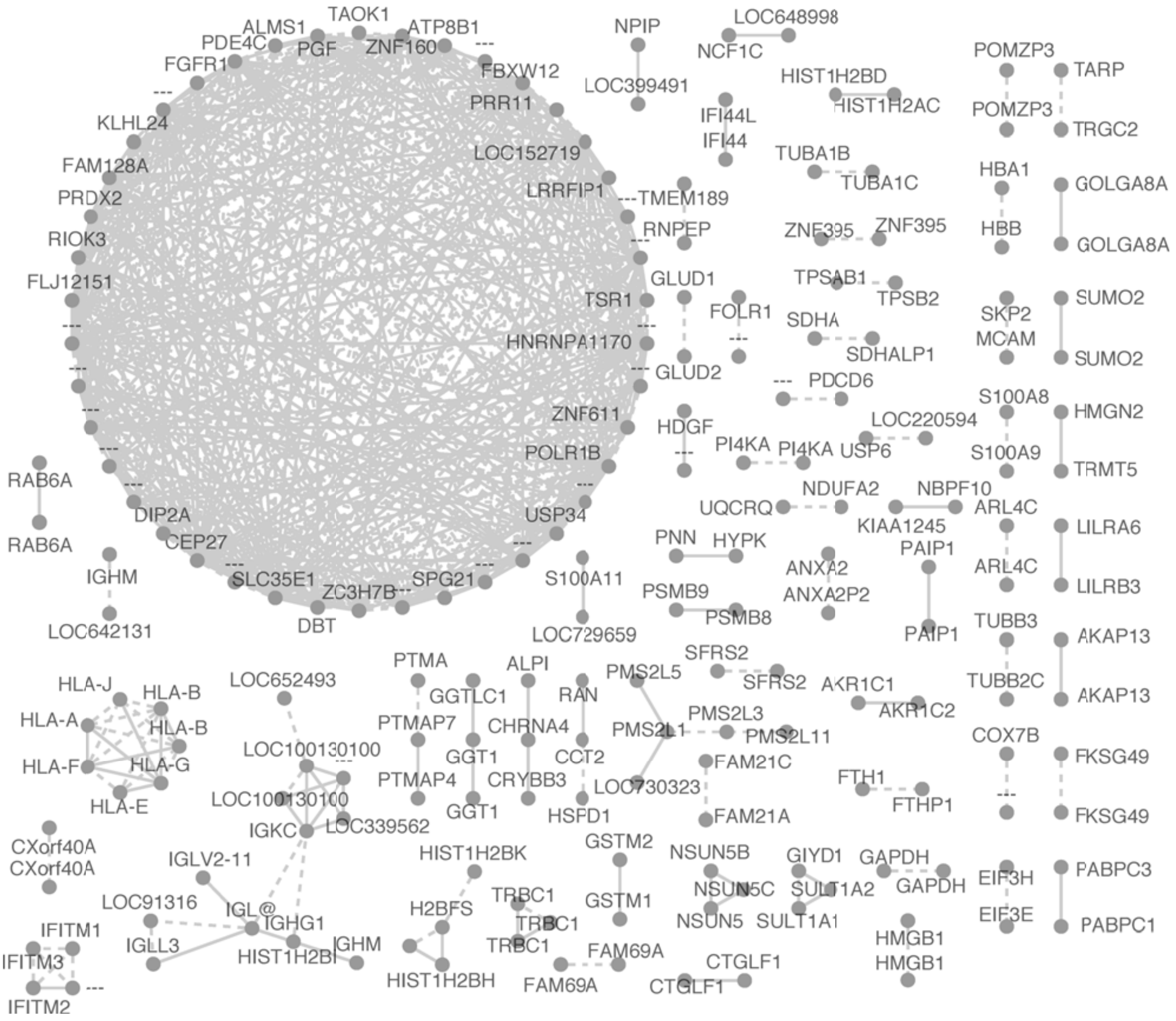

Figure 2: Subnetworks obtained by collecting the top 1000 connections with the highest MI within the human network. Since we considered the MI among probes and not genes, the same gene symbol may appear more than ones.

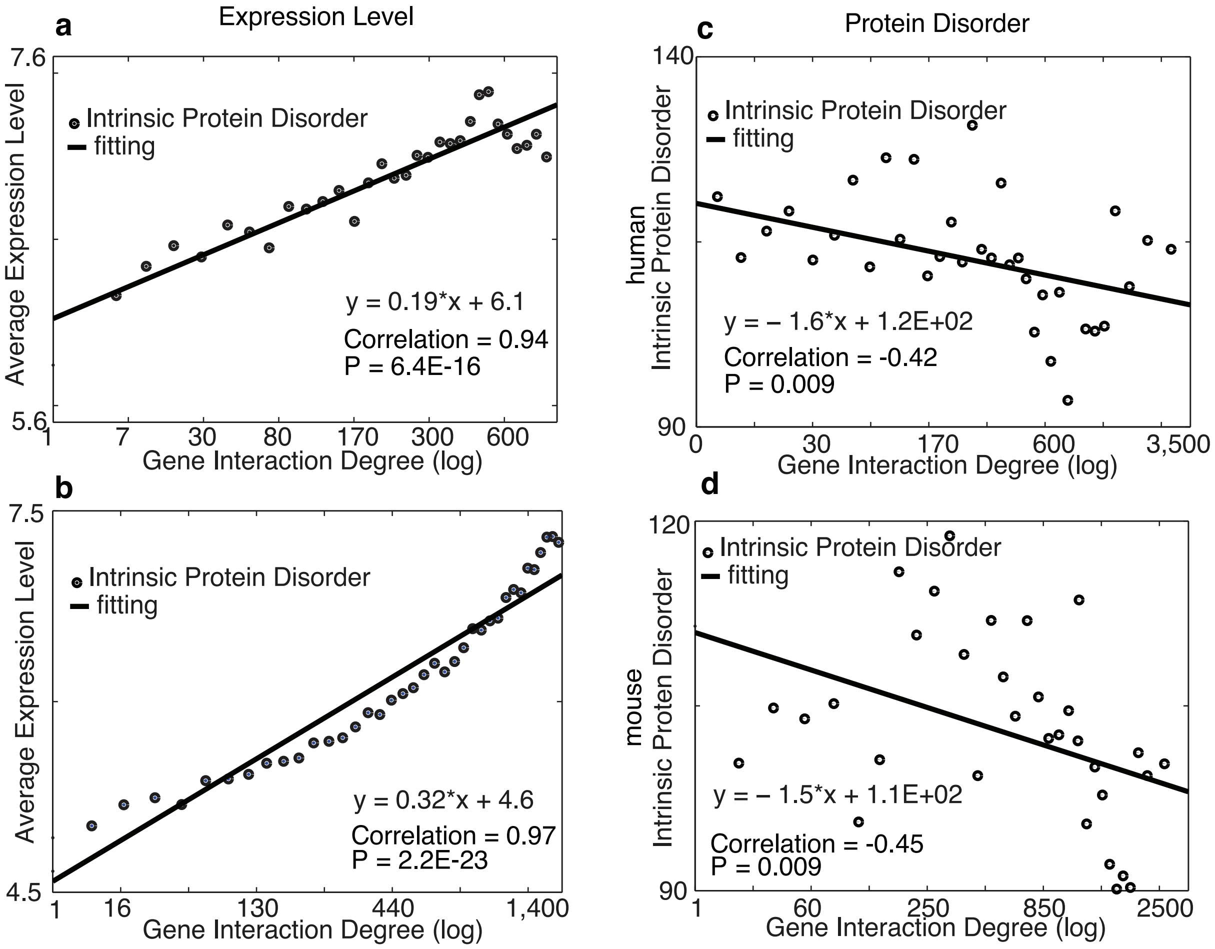

Figure 3: "Hub" genes tend to be more expressed and to have a lower protein disorder compared to other genes. (a-b) Gene degree (x-axis) versus intrinsic protein disorder in both human (a) and mouse (b) networks. Genes were grouped in equally sampled quantiles (500 genes each in human, 1000 genes each in mouse) and the average gene degree (x-axis) and protein disorder (y-axis) were computed. (c-d) Gene degree (x-axis) versus expression level. The average gene degree was computed as before, the average gene expression level was computed by computing the average expression of each gene across 79 tissues (GEO Accession number: GSE2003).

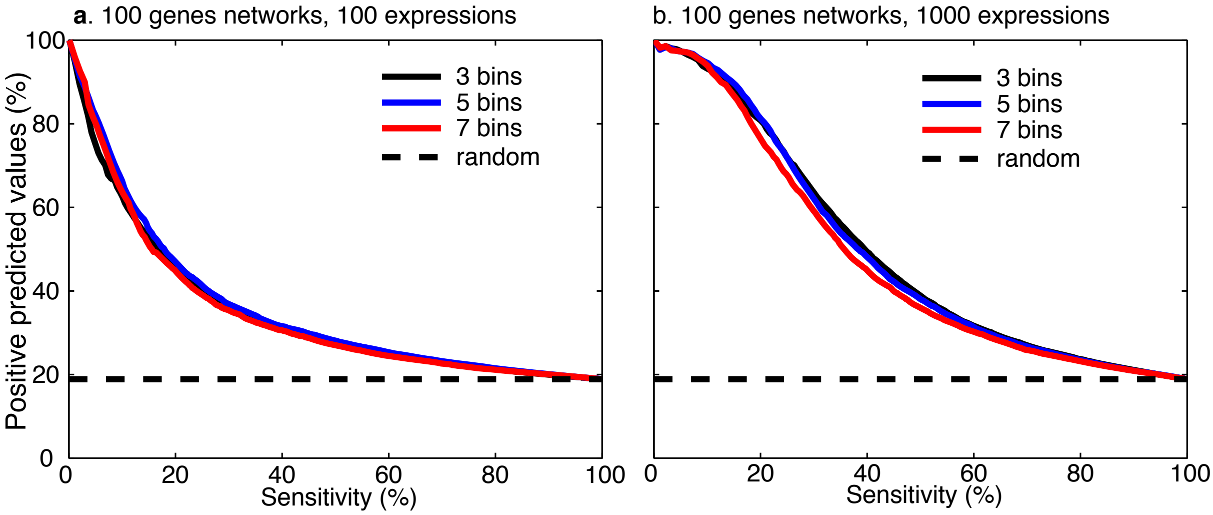

Figure 4: Comparison of the algorithm performance when changing the number of bins. Simulated networks of size 100 were used for the comparison. Data used in panel (a) are related to the inference from locally perturbed GEPs (100 experiments). Whereas, in panel (b) were used simulated GEPs related to the global perturbation (1000 experiments). The plots report the precision-recall curve of the performances.